Ribosome profiling of porcine reproductive and respiratory syndrome virus reveals novel features of viral gene expression

Georgia May Cook

King’s College
Department of Pathology
University of Cambridge

Supervisors: Prof. Ian Brierley and Dr Andrew E. Firth

This thesis is submitted for the degree of Doctor of Philosophy

May 2021
Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee. I was granted special permission by the Faculty of Biology Degree Committee for an extension of the prescribed word limit to 80,000 words (excluding appendices, bibliography, footnotes and tables).

Published work to which I contributed as part of my PhD, but which is not included in the main body of this thesis (* indicates co-first authorship):

  - This paper can be found in Appendix section 6. I performed the ribosome profiling experiments, analysed the data, and contributed to manuscript preparation.

  - This paper can be found in Appendix section 7. I analysed the ribosome profiling data and contributed to manuscript preparation.
Abstract

Porcine reproductive and respiratory syndrome virus (PRRSV) is an arterivirus which causes significant economic losses to the swine industry worldwide. Here, the viral translatome was characterised using ribosome profiling (RiboSeq), a high-throughput sequencing-based technique that allows the positions of translating ribosomes to be mapped to a genome with sub-codon precision. This was used in parallel with RNA sequencing (RNASeq) to analyse changes in viral gene expression over a timecourse of infection in MARC-145 cells.

The PRRSV genome contains two programmed ribosomal frameshift (PRF) signals, at which conserved elements promote the backwards slippage of the ribosome by one or two nucleotides. PRRSV encodes a canonical −1 PRF site, at which frameshifting is directed by a downstream RNA pseudoknot structure, to facilitate expression of the ORF1b-encoded viral replicase. At a second PRF site, both −1 and −2 PRF are stimulated by a trans-acting complex of a viral (nsp1β) and cellular (PCBP) protein in a unique, non-canonical mechanism, generating alternative forms (nsp2N and nsp2TF) of a viral non-structural protein, nsp2. Frameshift efficiency at both sites was quantified and, at the non-canonical site, was found to increase with time, rendering this the second known example of temporally regulated PRF during infection. This novel aspect of viral gene expression regulation is likely facilitated by accumulation of the PRF-stimulatory viral protein, nsp1β, during the virus life cycle. Surprisingly, frameshift efficiency at the canonical ORF1ab PRF site was also found to increase over time, overturning the common assumption that RNA structure-directed sites operate at a fixed efficiency, with potential implications for the numerous other viruses which encode canonical PRF sites.

Several highly translated additional ORFs were discovered in the PRRSV genome, the translation of which is potentially facilitated by multiple novel viral transcripts. For example, a 125-codon ORF overlapping nsp12 was discovered, which is expressed as highly as nsp12 itself in the late stages of replication and likely translated from novel subgenomic (sg) RNA transcripts overlapping the 3′ end of ORF1b. Similar transcripts were discovered for both PRRSV-1 and PRRSV-2, suggesting a potential conserved mechanism for temporal regulation of expression of the 3′-proximal region of ORF1b. In addition, a highly translated, short upstream ORF (uORF) was identified in the 5′ UTR, the presence of which is highly conserved amongst PRRSV-2 isolates.

This work is the first application of RiboSeq to arterivirus-infected cells and reveals new features which add to our understanding of the complexity of gene expression programmes in this important family of nidoviruses.
To Gordon and Patricia Cook
Acknowledgements

So many people have helped me along the way to finishing this PhD, both academically and socially. I count myself truly lucky to have such a vibrant and supportive community of friends and mentors, of which I mention just a fraction here. My thanks go to:

- Prof. Ian Brierley, my primary supervisor, for his academic input and ability to see the bigger picture as I drown in the detail, but also for making the lab an enjoyable place to be and always providing a well-curated soundtrack (with a few exceptions!).
- Dr Andrew Firth, my co-supervisor, for agreeing to take me on to learn bioinformatics and for never complaining when my promise that it would be “hardly any extra work” for him turned out to be somewhat optimistic.
- Prof. Ying Fang and her whole lab (especially Dr Pengcheng Shang and Dr Yanhua Li), for hosting me in Kansas for six weeks and for performing all the virus work using the North American PRRSV isolate in this thesis.
- Dr Katherine Brown, who supervised me for the computational aspect of all my projects and taught me a huge amount about bioinformatics. Katy will always make time to help others, a quality of hers from which I benefited countless times during my PhD.
- Dr Sawsan Napthine, who supervised me for the experimental component of the TMEV project and who’s thoughtful and kind presence I have greatly missed since working from home.
- Dr Chris Hill, who supervised me for the beginnings of a structural biology project that was sadly put on the backburner to allow time to complete my other projects. His quick sense of humour and listening ear have also helped me through many a long day in the lab.
- Dr Nerea Irigoyen and Dr Liliana Echavarría-Consuegra, with whom I collaborated on the coronavirus project and whose company is always a pleasure.
- Dr Adam Dinan, who patiently spent three months at the beginning of my PhD teaching me bash and R from scratch.
- Charlotte Lefèvre, for being a much better student to me than I was to Adam.
- Everyone else in the Division of Virology, which really has been a wonderful place to work.
- The Wellcome Trust and the Department of Pathology for funding, and the Microbiology Society, King’s College Bedford Fund, and Cambridge Philosophical Society for additional travel funding.
- King’s College, for providing a wonderful community of people with which to enjoy the “extra-curricular” side of life.
- My ever-supportive friends, from all parts of my life.
• My partner, Tom Pryce, who is the kindest person I have ever met and does so much to support me – including working in the living room so I could have the study to myself while we both had to work from home for the past year.
• My family, who I also count among my friends and to whom I owe everything.
• And finally, Patricia and Gordon Cook, my granny and grampa, who have always looked out for my wellbeing and to whom I dedicate this thesis.
# Contents

- Declaration ........................................................................................................................ ii
- Abstract .............................................................................................................................. iii
- Acknowledgements .......................................................................................................... vi
- Contents .......................................................................................................................... viii
- List of Tables ..................................................................................................................... xiii
- List of Figures .................................................................................................................... xiv
- Table of abbreviations ...................................................................................................... xviii

## Chapter 1: Introduction ................................................................................................. 1

1.1 Mechanism of eukaryotic translation ........................................................................... 1

  1.1.1 Structure of the ribosome ................................................................................. 1

  1.1.2 Translation initiation ....................................................................................... 3

  1.1.3 Translation elongation .................................................................................... 7

  1.1.4 Translation termination and ribosome recycling ........................................... 12

  1.1.5 Translation quality control ........................................................................... 13

1.2 Programmed ribosomal frameshifting ..................................................................... 17

  1.2.1 Canonical mechanisms of $-1$ PRF ............................................................. 18

  1.2.2 Non-canonical mechanisms of trans-activated $-1/-2$ PRF ......................... 21

    1.2.2.1 Protein-stimulated $-1$ and $-2$ PRF in arteriviruses ............................ 22

1.3 Ribosome profiling ............................................................................................... 27

  1.3.1 Foundation of the technique ....................................................................... 27

    1.3.1.1 History of ribosome profiling ............................................................. 27

    1.3.1.2 Fundamental basis of the technique ................................................... 27

    1.3.1.3 Inference of ribosomal P site position and determination of RPF phase ........... 29

  1.3.2 Further development of the protocol .............................................................. 31

    1.3.2.1 Use of cycloheximide pre-treatment ............................................... 31

    1.3.2.2 Discovery of a second population of RPFs ...................................... 34

    1.3.2.3 Adaptations to capture different populations of ribosomes ............... 36

1.3.3 Sources of bias ............................................................................................... 36

    1.3.3.1 Technical biases .............................................................................. 37

      1.3.3.1.1 Nuclease bias (RiboSeq) .............................................................. 37

      1.3.3.1.2 Fragmentation bias (RNASeq) ..................................................... 37

      1.3.3.1.3 Ligation bias ........................................................................... 37
1.3.3.1.4 PCR bias.................................................................38
1.3.3.1.5 Other sources of technical bias.................................38
1.3.3.1.6 Effects of technical bias...........................................40
1.3.3.2 Biological factors..........................................................41
1.4 Porcine reproductive and respiratory syndrome virus..................43
  1.4.1 Disease........................................................................43
  1.4.2 Emergence and phylogeny.................................................43
  1.4.3 Tropism and culture systems............................................48
  1.4.4 Disease control................................................................51
  1.4.5 Replication cycle............................................................52
  1.4.6 Genome structure and protein functions.............................53
    1.4.6.1 Non-structural proteins.............................................54
      1.4.6.1.1 Expression strategy.............................................54
      1.4.6.1.2 Protein functions............................................56
    1.4.6.2 Structural proteins....................................................59
      1.4.6.2.1 Expression strategy.............................................59
      1.4.6.2.2 Protein functions............................................62
    1.4.6.3 UTRs......................................................................63

Chapter 2: Materials and Methods..................................................67
  2.1 Cells and viruses..................................................................67
    2.1.1 EU PRRSV (PRRSV-1) infections (performed by Prof. Ian Brierley and Dr Adrian Mockett)..................67
    2.1.2 NA PRRSV (PRRSV-2) infections (performed by Dr Pengcheng Shang, Dr Yanhua Li and Georgia Cook)........67
  2.2 Western blotting....................................................................69
  2.3 Ribosome profiling library preparation ................................70
    2.3.1 Preparation of RNA fragments.......................................72
      2.3.1.1 Isolation of RNA fragments for RNASeq..........................72
      2.3.1.2 Isolation of ribosome-protected fragments for RiboSeq.......72
    2.3.2 Gel purification of RNA fragments.....................................73
    2.3.3 Generation of amplicon libraries......................................73
    2.3.4 High-throughput sequencing..........................................74
    2.3.5 Details of protocol variations.........................................75
  2.4 Ribosome profiling data analysis............................................76
    2.4.1 Core pipeline..................................................................76
2.4.1.1 Mapping reads to viral and host genomes ........................................ 76
2.4.1.2 Quality control ................................................................................. 77
  2.4.1.2.1 Host read length distribution and phasing ................................. 77
  2.4.1.2.2 Distribution of reads relative to host mRNA start and stop codons 77
  2.4.1.2.3 Length distribution of reads mapping to host transcript CDSs compared to 3’ UTRs ................................................................. 77
  2.4.1.2.4 Virus read length distribution and phasing ............................... 78
  2.4.1.3 Plots of read distribution on virus genomes .................................... 78
  2.4.2 Frameshift efficiency calculations ...................................................... 79
    2.4.2.1 Nsp2 density normalisation-based method .................................... 79
    2.4.2.2 Nsp2 phasing-based method ...................................................... 80
    2.4.2.3 ORF1ab density-based method ................................................... 81
    2.4.2.4 Bootstrap resampling .................................................................. 81
  2.4.3 Chimeric read analysis for novel transcript discovery ........................ 82
  2.4.4 Sequence conservation plots .............................................................. 83
  2.4.5 Novel ORF analysis ........................................................................... 84
  2.4.6 Transcript abundance and translation efficiency analyses .................... 84
  2.4.7 Data availability ............................................................................... 86

Chapter 3: Analysis of programmed ribosomal frameshifting on the PRRSV genome 87

  3.1 Introduction ............................................................................................. 87
  3.2 Experimental set-up ................................................................................ 90
    3.2.1 Infections and library preparation .................................................. 90
    3.2.2 Aligning reads to host and viral genomes ....................................... 92
  3.3 Quality control analysis .......................................................................... 94
    3.3.1 Quality control analysis of the NA PRRSV noCHX dataset .............. 94
      3.3.1.1 Library composition ................................................................. 94
      3.3.1.2 Read length distribution ......................................................... 95
      3.3.1.3 Distribution of reads relative to start and stop codons ............... 97
      3.3.1.4 Phase composition ................................................................. 99
      3.3.1.5 Assessment of potential RNP contamination in RiboSeq libraries ................................................................. 100
    3.3.2 Quality control analysis of the EU PRRSV dataset ............................ 105
    3.3.3 Quality control analysis of the NA PRRSV CHX dataset ................. 107
  3.4 An overview of viral transcription and translation .................................... 110
  3.5 Investigation of PRF on the viral genome .............................................. 118
3.5.1 Visualising the occurrence of PRF .......................................................... 118
3.5.2 Calculation of PRF efficiency at the nsp2 site ...................................... 127
3.5.3 Calculation of PRF efficiency at the ORF1ab site ............................... 133
3.5.4 Investigating ribosomal pausing at the PRF sites ............................... 136
3.5.5 Investigating the potential for lengthened reads at the PRF sites ......... 148

3.6 Discussion and future work .................................................................. 152
  3.6.1 Discussion ......................................................................................... 152
  3.6.2 Future work ..................................................................................... 155
  3.6.3 Conclusion ......................................................................................... 157

Chapter 4: Analysis of PRRSV gene expression ........................................ 159
  4.1 Introduction .......................................................................................... 159
  4.2 Characterising the PRRSV transcriptome ........................................... 161
    4.2.1 Pipeline to detect chimeric viral transcripts ................................. 161
    4.2.2 Overview of results ...................................................................... 163
    4.2.3 TRS-spanning junctions ................................................................. 169
    4.2.4 Non-TRS-spanning junctions ......................................................... 171
    4.2.5 Novel transcripts that might artificially inflate ORF1ab PRF efficiency 173
  4.3 Characterising the PRRSV translatome .............................................. 177
    4.3.1 Detection of novel ORFs using PRICE ...................................... 177
    4.3.2 Novel viral ORFs .......................................................................... 177
  4.4 Quantification of PRRSV gene expression ......................................... 194
    4.4.1 Overview of results ...................................................................... 194
    4.4.2 Investigating the potential for ORF1b sgRNAs to regulate in-frame translation of ORF1b ............................................................... 209
    4.4.3 Investigating the effect of heteroclite sgRNAs on translation ........ 219
  4.5 Discussion and future work ................................................................. 221
    4.5.1 Discussion ..................................................................................... 221
    4.5.2 Future work .................................................................................. 224
    4.5.3 Conclusion .................................................................................... 227

Chapter 5: Plenary discussion .................................................................. 229
  5.1 PRRSV ............................................................................................... 229
  5.2 Programmed ribosomal frameshifting ................................................ 231
  5.3 Ribosome profiling ............................................................................. 233
  5.4 Conclusion ......................................................................................... 237
References ...................................................................................................................... 239
Appendices ................................................................................................................... 279

1 EU PRRSV genome sequence .............................................................................. 279
2 Oligonucleotide sequences .................................................................................... 283
3 Coordinates of regions of PRRSV genomes used for analyses ......................... 284
   3.1 General plots and analyses .............................................................................. 284
   3.2 Transcript abundance calculations ................................................................. 286
      3.2.1 For all calculations except investigation of the “heteroclite” region .......... 286
      3.2.2 For investigation of the “heteroclite” region ............................................. 287
   3.3 Translation level calculations .......................................................................... 288
      3.3.1 For all calculations except investigation of the “heteroclite” region ......... 288
      3.3.2 For investigation of the “heteroclite” region ............................................. 289
4 Composition of mapping databases ................................................................... 290
   4.1 Host databases .................................................................................................. 290
   4.2 Contaminants database .................................................................................... 290
5 Viral ORFs detected by PRICE ............................................................................ 292
   5.1 NA PRRSV ...................................................................................................... 292
   5.2 EU PRRSV ...................................................................................................... 296
6 Investigating molecular mechanisms of 2A-stimulated ribosomal pausing and   frameshifting in Theilovirus ................................................................................. 297
7 Manipulation of the unfolded protein response: A pharmacological strategy against coronavirus infection ................................................................. 352
List of Tables

Table 1 PRRSV proteins and their functions .......................................................... 65
Table 2 Datasets in this study ................................................................................. 71
Table 3 RNase added to Riboseq lysates ............................................................... 73
Table 4 WT and KO2 library pairings ................................................................. 79
List of Figures

Figure 1.1 Structure of the eukaryotic ribosome and tRNA ........................................... 2
Figure 1.2 Mechanism of canonical eukaryotic translation initiation ................................ 4
Figure 1.3 Mechanism of canonical eukaryotic translation elongation ................................. 9
Figure 1.4 Mechanism of canonical eukaryotic translation termination and ribosome recycling ................................................................. 12
Figure 1.5 Quality control pathways triggered by ribosome collisions .............................. 16
Figure 1.6 Mechanisms of canonical −1 PRF ................................................................. 19
Figure 1.7 PRRSV nsp2 PRF site .................................................................................. 24
Figure 1.8 Key experimental steps in preparation of RiboSeq libraries and parallel RNASeq libraries ........................................................................ 28
Figure 1.9 Hypothetical example of RiboSeq and RNASeq read distributions on an illustrative transcript species ................................................................. 29
Figure 1.10 Hypothetical example of RiboSeq read 5’ end distribution relative to initiation codons ......................................................................................... 30
Figure 1.11 The effect of CHX pre-treatment and stress on S. cerevisiae RiboSeq read distribution relative to initiation codons ......................................................... 32
Figure 1.12 The effect of CHX pre-treatment and stress on Mus musculus RiboSeq read distribution relative to initiation and termination codons ................................. 32
Figure 1.13 Wu et al. model for origins of the two populations of differently sized RPFs ..................................................................................................................... 35
Figure 1.14 Phylogenetic tree of the European (type 1) lineage of PRRSV ........................... 45
Figure 1.15 Phylogenetic tree of the North American (type 2) lineage of PRRSV .............. 47
Figure 1.16 Phylogenetic tree of the Arteriviridae family ................................................... 48
Figure 1.17 PRRSV replication cycle ................................................................................. 53
Figure 1.18 PRRSV genome and subgenomic transcripts .................................................... 54
Figure 1.19 PRRSV ORF1ab −1 PRF site ......................................................................... 55
Figure 1.20 Nsp2 and related proteins produced by frameshifting .................................... 55
Figure 1.21 Arterivirus replication and sgRNA production ................................................. 60
Figure 1.22 PRRSV virion architecture ............................................................................. 62
Figure 1.23 PRRSV 5’ UTR predicted structure ............................................................... 64
Figure 3.1 The PRF-defective KO2 mutant virus ............................................................... 88
Figure 3.2 Western blots of profiling lysates .................................................................... 91
Figure 3.3 Composition of libraries .................................................................................. 96
Figure 3.4 Length distribution of CDS-mapping reads ..................................................... 97
Figure 3.5 Average distribution of 5’ ends of host mRNA-mapping reads relative to start and stop codons .......................................................... 98
Figure 3.6 Phase composition of CDS-mapping reads........................................ 100
Figure 3.7 Length distribution of RPFs mapping to specified regions of the viral genome ................................................................................................. 101
Figure 3.8 Number of reads of each length that are attributed to each phase, for RPFs mapping to specified regions of the viral genome.................. 102
Figure 3.9 Assessment of potential RNP contamination of host-mapping RiboSeq reads .................................................................................................. 104
Figure 3.10 Quality control for EU PRRSV dataset ........................................... 106
Figure 3.11 Quality control for NA PRRSV CHX dataset ............................... 108
Figure 3.12 Overview of viral transcription and translation over a timecourse of NA PRRSV infection ................................................................. 112
Figure 3.13 Further replicates of a timecourse of NA PRRSV viral transcription..... 114
Figure 3.14 Further replicates of a timecourse of NA PRRSV viral translation ...... 115
Figure 3.15 Overview of EU PRRSV transcription and translation at 8 hpi and investigation of negative-sense RiboSeq reads.............................. 116
Figure 3.16 Distribution of reads in each phase across ORF1ab on the NA PRRSV genome .................................................................................. 119
Figure 3.17 Distribution of reads in each phase across ORF1ab on the NA PRRSV genome, further replicates .......................................................... 120
Figure 3.18 Percentage of RiboSeq reads in each phase across the ORF1a region of the NA PRRSV genome................................................................. 121
Figure 3.19 Bar charts of percentage of reads in each phase in the specified regions of the viral genome, for NA and EU PRRSV .................................. 122
Figure 3.20 Distribution of reads across the ORF1ab region of the EU PRRSV genome ............................................................................................... 123
Figure 3.21 RNASeq read densities on the ORF1ab polyprotein region of the NA PRRSV viral genome ........................................................................ 125
Figure 3.22 Ratio of RPF density at each position on the NA PRRSV WT genome relative to the KO2 genome.............................................................. 126
Figure 3.23 Percentage frameshift efficiency at the nsp2 site, calculated by two different methods................................................................. 130
Figure 3.24 Percentage frameshift efficiency at the ORF1ab site ....................... 134
Figure 3.25 RiboSeq read densities around the nsp2 frameshift site, all read lengths138
Figure 3.26 RiboSeq read densities around the nsp2 frameshift site, short read lengths ........................................................................................................ 140
Figure 3.27 RiboSeq read densities around the nsp2 frameshift site, classical read lengths................................................................. 141
Figure 3.28 Histogram of stAI of sense codons in M. mulatta……………………………………. 144
Figure 3.29 RiboSeq read densities around the ORF1ab frameshift site, all read lengths .................................................................................................................. 145
Figure 3.30 RiboSeq read densities around the ORF1ab frameshift site, short read lengths.................................................................................................................. 146
Figure 3.31 RiboSeq read densities around the ORF1ab frameshift site, classical read lengths.................................................................................................................. 147
Figure 3.32 Broad spectrum ribosome profiling at 9 hpi (replicate 2).................................150
Figure 3.33 Density of disome-protected fragments around A) the nsp2 frameshift site and B) the ORF1ab frameshift site ................................................................. 151
Figure 4.1 Schematic of a junction-spanning (chimeric) read.............................................162
Figure 4.2 Box plot of the widths of merged junctions .........................................................162
Figure 4.3 The NA PRRSV transcriptome at A) 3 and B) 6 hpi........................................164
Figure 4.4 The NA PRRSV transcriptome at 9 hpi .............................................................165
Figure 4.5 The NA PRRSV transcriptome at 12 hpi..........................................................166
Figure 4.6 The EU PRRSV transcriptome at 8 hpi...........................................................167
Figure 4.7 The NA PRRSV transcriptome at 9 hpi, replicate 1 only ...............................168
Figure 4.8 Alternative TRS bodies for the N sgRNA.......................................................169
Figure 4.9 Novel transcripts that could potentially inflate calculated ORF1ab frameshift efficiency for NA PRRSV at 12 hpi.................................................................174
Figure 4.10 ORF1ab −1 PRF efficiency calculated using the truncated downstream region ......................................................................................................................176
Figure 4.11 Overview of novel ORFs in the NA PRRSV translatome.............................178
Figure 4.12 Overview of novel ORFs in the EU PRRSV translatome............................180
Figure 4.13 A novel, highly expressed uORF in the NA PRRSV 5′ UTR........................182
Figure 4.14 A predicted novel uORF in the EU PRRSV 5′ UTR .................................183
Figure 4.15 (spans multiple pages). Plots of RPF distribution on regions of the NA PRRSV genome predicted to encode novel ORFs..............................................185
Figure 4.16 (spans multiple pages). Plots of RPF distribution on regions of the EU PRRSV genome predicted to encode novel ORFs..............................................191
Figure 4.17 Subgenomic mRNA transcripts and ORFs included in viral gene expression analysis of A) NA PRRSV and B) EU PRRSV .....................................................195
Figure 4.18 Viral transcript abundance and total translation of viral genes .....................196
Figure 4.19 Transcript abundance of positive-sense novel and canonical viral transcripts calculated using decumulation .................................................................198
Figure 4.20 Transcript abundance of positive-sense novel and canonical viral transcripts calculated using junction-spanning reads.................................................199
Figure 4.21 Transcript abundance of positive-sense novel and canonical viral transcripts calculated using junction-spanning reads for libraries with shorter purified RNA fragments................................................................. 200

Figure 4.22 Percentage of the viral transcriptome represented by each transcript as calculated using the decumulation method (left) and chimeric junction-spanning reads method (right) .................................................................................. 201

Figure 4.23 RiboSeq density attributed to each viral ORF ........................................... 203

Figure 4.24 Percentage of the viral translatome represented by each ORF............. 205

Figure 4.25 Translation efficiency of viral ORFs calculated using decumulated RNASeq density .................................................................................................................................................. 206

Figure 4.26 Translation efficiency of viral ORFs calculated using the junction-spanning RNASeq reads .................................................................................................................................................. 207

Figure 4.27 Comparison of translation efficiency for canonical viral ORFs and host genes with ten or more RiboSeq and RNASeq reads ........................................ 208

Figure 4.28 Distribution of RNASeq (left) and RiboSeq (right) reads mapping to the ORF1b region of the NA PRRSV genome ..................................................... 210

Figure 4.29 Distribution of RNASeq (upper) and RiboSeq (lower) reads mapping to the ORF1b region of the EU PRRSV genome ..................................................... 211

Figure 4.30 RiboSeq read density attributed to different regions of ORF1b .......... 212

Figure 4.31 (spans multiple pages). RPF distribution in regions downstream of ORF1b sgRNA TRS bodies and initiation context of potential start codons .......... 213

Figure 4.32 Expression of different regions of ORF1a ............................................. 220

Figure 4.33 Features of a uORF which correlate with the extent of main ORF (CDS) translational repression ......................................................................................... 224

Figure 5.1 Schematic summary of non-canonical mechanisms of PRRSV gene expression regulation supported by this study ........................................... 229
## Table of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>C. griseus</td>
<td><em>Cricetulus griseus</em></td>
</tr>
<tr>
<td>C. sabaeus</td>
<td><em>Chlorocebus sabaeus</em></td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>CHX</td>
<td>cycloheximide</td>
</tr>
<tr>
<td>CoV</td>
<td>coronavirus</td>
</tr>
<tr>
<td>DI</td>
<td>defective interfering</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco’s modified Eagle’s medium</td>
</tr>
<tr>
<td>DMV</td>
<td>double-membrane vesicle</td>
</tr>
<tr>
<td>DPF</td>
<td>disome-protected fragment</td>
</tr>
<tr>
<td>ds</td>
<td>double-stranded</td>
</tr>
<tr>
<td>DUB</td>
<td>deubiquitinase</td>
</tr>
<tr>
<td>E. coli</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>EAV</td>
<td>equine arteritis virus</td>
</tr>
<tr>
<td>eEF</td>
<td>eukaryotic elongation factor</td>
</tr>
<tr>
<td>eIF</td>
<td>eukaryotic initiation factor</td>
</tr>
<tr>
<td>EM</td>
<td>electron microscopy</td>
</tr>
<tr>
<td>EMSA</td>
<td>electrophoretic mobility shift assay</td>
</tr>
<tr>
<td>ER</td>
<td>endoplasmic reticulum</td>
</tr>
<tr>
<td>eRF</td>
<td>eukaryotic release factor</td>
</tr>
<tr>
<td>ES</td>
<td>expansion segment</td>
</tr>
<tr>
<td>EU</td>
<td>European (PRRSV-1)</td>
</tr>
<tr>
<td>FBS</td>
<td>Foetal bovine serum</td>
</tr>
<tr>
<td>FS</td>
<td>frameshift</td>
</tr>
<tr>
<td>fwd</td>
<td>forward sense</td>
</tr>
<tr>
<td>gDNA</td>
<td>genomic DNA</td>
</tr>
<tr>
<td>hnRNP</td>
<td>heterogeneous nuclear ribonucleoprotein</td>
</tr>
<tr>
<td>HOG</td>
<td>high-osmolarity glycerol</td>
</tr>
<tr>
<td>hpi</td>
<td>hours post-infection</td>
</tr>
<tr>
<td>HVR</td>
<td>hyper-variable region</td>
</tr>
<tr>
<td>IC</td>
<td>initiation complex</td>
</tr>
<tr>
<td>IFN</td>
<td>interferon</td>
</tr>
<tr>
<td>IL</td>
<td>interleukin</td>
</tr>
<tr>
<td>lRES</td>
<td>internal ribosome entry site</td>
</tr>
<tr>
<td>KH</td>
<td>K homology</td>
</tr>
<tr>
<td>LDV</td>
<td>Lactate dehydrogenase elevating virus</td>
</tr>
<tr>
<td>M. auratus</td>
<td><em>Mesocricetus auratus</em></td>
</tr>
<tr>
<td>M. mulatta</td>
<td><em>Macaca mulatta</em></td>
</tr>
<tr>
<td>M. musculus</td>
<td><em>Mus musculus</em></td>
</tr>
<tr>
<td>MERS</td>
<td>Middle East respiratory syndrome virus</td>
</tr>
<tr>
<td>MFC</td>
<td>multi-factor complex</td>
</tr>
<tr>
<td>MHV</td>
<td>murine hepatitis virus</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MNase</td>
<td>micrococcal S7 nuclease</td>
</tr>
<tr>
<td>MOI</td>
<td>Multiplicity of infection</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NA</td>
<td>North American (PRRSV-2)</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>NGD</td>
<td>no go decay</td>
</tr>
<tr>
<td>NiRAN</td>
<td>nidovirus RdRp-associated nucleotidyltransferase</td>
</tr>
<tr>
<td>nsp</td>
<td>non-structural protein</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotides</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>OTU</td>
<td>ovarian tumour domain</td>
</tr>
<tr>
<td>PABP</td>
<td>poly(A)-binding protein</td>
</tr>
<tr>
<td>PAM</td>
<td>porcine alveolar macrophage</td>
</tr>
<tr>
<td>PCBP</td>
<td>poly(C)-binding protein</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PIC</td>
<td>pre-initiation complex</td>
</tr>
<tr>
<td>PLP</td>
<td>papain-like protease</td>
</tr>
<tr>
<td>PNK</td>
<td>polynucleotide kinase</td>
</tr>
<tr>
<td>PRF</td>
<td>programmed ribosomal frameshifting</td>
</tr>
<tr>
<td>PRRS</td>
<td>porcine reproductive and respiratory syndrome</td>
</tr>
<tr>
<td>PRRSV</td>
<td>porcine reproductive and respiratory syndrome virus</td>
</tr>
<tr>
<td>RdRp</td>
<td>RNA-dependent RNA polymerase</td>
</tr>
<tr>
<td>rep</td>
<td>replicate</td>
</tr>
<tr>
<td>rev</td>
<td>reverse sense</td>
</tr>
<tr>
<td>RiboSeq</td>
<td>ribosome profiling</td>
</tr>
<tr>
<td>RNase</td>
<td>ribonuclease</td>
</tr>
<tr>
<td>RNP</td>
<td>ribonucleoprotein</td>
</tr>
<tr>
<td>RPF</td>
<td>ribosome-protected fragment</td>
</tr>
<tr>
<td>RPM</td>
<td>reads per kilobase per million mapped reads</td>
</tr>
<tr>
<td>RPKM</td>
<td>reads per million mapped reads</td>
</tr>
<tr>
<td>RQC</td>
<td>ribosome-associated quality control</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RT</td>
<td>reverse transcription</td>
</tr>
<tr>
<td>RTC</td>
<td>replication and transcription complex</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>Saccharomyces cerevisiae</td>
</tr>
<tr>
<td>SARS</td>
<td>severe acute respiratory syndrome virus</td>
</tr>
<tr>
<td>SD</td>
<td>Shine-Dalgarno</td>
</tr>
<tr>
<td>sg</td>
<td>subgenomic</td>
</tr>
<tr>
<td>sgORF</td>
<td>subgenomic ORF</td>
</tr>
<tr>
<td>sgRNA</td>
<td>subgenomic RNA</td>
</tr>
<tr>
<td>SHFV</td>
<td>simian haemorrhagic fever virus</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
</tr>
<tr>
<td>ss</td>
<td>single-stranded</td>
</tr>
<tr>
<td>SS</td>
<td>slippery sequence</td>
</tr>
<tr>
<td>stAI</td>
<td>species-specific tRNA adaptation index</td>
</tr>
<tr>
<td>tAI</td>
<td>tRNA adaptation index</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>TC</td>
<td>ternary complex</td>
</tr>
<tr>
<td>TM</td>
<td>transmembrane</td>
</tr>
<tr>
<td>TMEV</td>
<td>Theiler’s murine encephalomyelitis virus</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>TRS</td>
<td>transcription regulatory sequence</td>
</tr>
<tr>
<td>uORF</td>
<td>upstream ORF</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>vRNA</td>
<td>viral RNA</td>
</tr>
<tr>
<td>WT</td>
<td>wild type</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

1.1 Mechanism of eukaryotic translation

Translation is the mechanism by which the information encoded in messenger RNA (mRNA) is used to make polypeptides. It is performed by the ribosome, the largest complex of RNA and protein in the cell, and can be broken down into four phases: initiation, elongation, termination and recycling. Each of these involves different factors alongside the ribosome, and each is subject to a different complement of regulatory mechanisms.

1.1.1 Structure of the ribosome

Structural studies of the ribosome have greatly informed understanding of its function. While at first the prokaryotic ribosome was the subject of many studies due to its smaller size making it more amenable to structural characterisation\textsuperscript{1-7}, recent years have seen many findings confirmed and extended to eukaryotic ribosomes.

An atomic-resolution structure of the complete eukaryotic ribosome was determined first for \textit{Saccharomyces cerevisiae}, by X-ray crystallography\textsuperscript{8}, preceded by several crystallography and cryo-electron microscopy (cryo-EM) studies at lower resolution\textsuperscript{9-13}. This revealed that, in line with translation being a fundamental requirement for facilitating life, the core of the ribosome is highly conserved between all three taxonomic domains, with variation tending to occur in the outer regions\textsuperscript{8,14} (Figure 1.1A). The core of the ribosome is mostly (~$\frac{2}{3}$ by mass) RNA, and the catalysis of peptide bond formation is performed by ribosomal RNA (rRNA) as opposed to protein\textsuperscript{15}. Amongst eukaryotes, the composition and structure of the whole ribosome is generally well-conserved, varying mostly in the presence or absence of insertions in particular regions of rRNA called expansion segments\textsuperscript{8,14} (ESs).
Figure 1.1 Structure of the eukaryotic ribosome and tRNA

A) Crystal structure of the subunits of the *S. cerevisiae* ribosome. Figure reproduced from Yusupova and Yusupov, 2014 (reference 14). Left: small (40S) subunit. Right: large (60S) subunit. Top: view of both subunits observing the inter-subunit interface. Bottom: view of both subunits observing the opposite, solvent side. Selected features of the ribosomal architecture are labelled. The conserved core of the ribosome consists of rRNA (white) and proteins (light orange), with eukaryotic-specific moieties shown in red. The highly conserved ring around the peptide exit channel corresponds to the region which interacts with membranes during protein synthesis in both prokaryotes and eukaryotes. DC = decoding centre; CP = central protuberance; PTC = peptidyl transferase centre; ES = expansion segment. B) Schematic of the structure of the ribosome. Scales and shapes of components are not representative. Numbers in the mRNA in the small subunit decoding centre represent the first, second and third nucleotide within each codon, and correspond to numbers in panel C. C) The
structure of tRNA (in this case Phe-tRNA, Protein Data Bank code 1OB5). Figure reproduced from Lant et al., 2019 (reference 16). Left: The tertiary structure forms an L shape. AA = amino acid. Right: the secondary structure forms a clover-leaf. Bases in the tRNA anticodon (purple) pair with the mRNA codon, depicted underneath the secondary structure with bases numbered as in panel B.

The ribosome consists of two subunits, the small subunit (40S in eukaryotes) and the large subunit (60S in eukaryotes), which together form the 80S ribosome (Figure 1.1B). The interface between the two subunits forms three binding sites for transfer RNA (tRNA), known as the A (aminoacyl), P (peptidyl) and E (exit) sites. On the small subunit, these are part of the decoding centre: a pocket in the ribosome which facilitates the selection of the cognate aminoacyl-tRNA (or release factor) to the mRNA codon in the A site17. On the large subunit, they are part of the peptidyl-transferase centre: the core catalytic region of the ribosome, in which rRNA catalyses the formation of a peptide bond between the aminoacyl-tRNA in the A site and the nascent peptidyl-tRNA in the P site17. The passage of mRNA into and out of the decoding centre is through the mRNA entrance and exit tunnels, which are wide enough to allow passage of single-stranded (ss) RNA but not structured RNA12,18,19. As such, RNA must be unwound by the helicase centre in the ribosome before it passes through the entry tunnel20–22. Similarly, a conserved tunnel facilitates egress of the nascent peptide7,23. On the interior of the tunnel, conserved rRNA and protein residues make a hydrophilic environment which, with a few exceptions (see section 1.3.3.2), generally prevents the nascent peptide chain from binding to the tunnel and stalling on its way out of the ribosome7,24–28. Unlike the mRNA entry and exit tunnels, the peptide exit tunnel is wide enough to accommodate a structured substrate. It can accommodate polypeptide chains formed into an α-helix, in which case it is long enough for ~60 amino acids, or unstructured polypeptide, in which case it is long enough for ~30 amino acids24–28.

The structure of tRNAs is also highly conserved. Four stem loops form a cloverleaf secondary structure, which assumes an L-shaped tertiary arrangement16 (Figure 1.1C). At the 3’ end of the tRNA, aminoacyl-tRNA synthetases catalyse the covalent linkage of the cognate amino acid to the conserved CCA motif. In the anticodon stem loop, three nucleotides in the loop base pair with the three nucleotides in the mRNA codon to confer specificity16.

### 1.1.2 Translation initiation

Canonical eukaryotic translation initiation is a complex process involving at least 12 different eukaryotic initiation factors (eIFs), many of which are composed of several subunits29 (Figure 1.2). This affords the initiation process many opportunities for modulation and it is generally considered the stage of translation that is most highly regulated30.
Figure 1.2 Mechanism of canonical eukaryotic translation initiation

Figure reproduced from Hinnebusch 2014 (reference 29). GTP is represented by a green circle, GDP by a red circle. eIF = eukaryotic initiation factor; PIC = pre-initiation complex; MFC = multi-factor complex; TC = ternary complex; Met-tRNAi = methionyl initiator tRNA; P = phosphate; IC = initiation complex; PABP = poly(A)-binding protein.
Eukaryotic translation initiation usually begins with activation of the mRNA, in which proteins bind both ends of the mRNA, and each other, to form a closed-loop structure. The 5′ end of eukaryotic mRNAs have a “cap” consisting of a 7-methylguanosine (m^7G) residue covalently linked by a triphosphate bridge to the first nucleotide of the transcript\textsuperscript{31}. This is bound by the cap-binding complex eIF4F, comprising three factors: eIF4E, which binds the cap, eIF4A, which is an ATP-dependent RNA helicase and eIF4G, which is a scaffold protein\textsuperscript{29}. The helicase activity of eIF4A is thought to catalyse unwinding of secondary structures in the 5′ UTR and this is stimulated by eIF4G along with two additional factors: eIF4B and eIF4H\textsuperscript{32–34}. At the 3′ end of eukaryotic mRNAs is the poly(A) tail, which is bound by poly(A)-binding proteins (PABPs). PABP interacts with the scaffold protein eIF4G, facilitating circularisation\textsuperscript{29}.

At the same time, the 43S pre-initiation complex (PIC) is assembled by loading the 40S small ribosomal subunit with the initiator tRNA (Met-tRNA\textsubscript{i}). Met-tRNA\textsubscript{i} is presented in the form of the ternary complex (TC), which is formed of eIF2, a heterotrimer of eIF2\textsubscript{α}, eIF2\textsubscript{β} and eIF2\textsubscript{γ}, bound to Met-tRNA\textsubscript{i} and GTP\textsuperscript{29}. Binding of the TC to the 40S subunit is assisted by eIF1, eIF1A, and eIF3, which bind the small ribosomal subunit in addition to eIF5\textsuperscript{35–38}. An alternative pathway of complex formation involves all these initiation factors except eIF1A assembling with the TC to form the multi-factor complex (MFC) before binding the 40S ribosomal subunit\textsuperscript{39,40}. The complex of the small ribosomal subunit, TC and eIF1, -1A, -3 and -5 is called the 43S PIC.

The assembled 43S PIC binds to the 5′ end of the closed-loop mRNA, stimulated by interactions between PIC proteins and subunits of eIF4F, to form the 43S/mRNA complex. In mammalian cells, eIF4G directly interacts with eIF3e to help recruit the 43S PIC, while in yeast this interaction occurs through eIF5 as a bridge\textsuperscript{41–43}. The helicase activity of eIF4A, stimulated by eIF4G, -B and -H, is also thought to promote association of the 43S PIC by ensuring the mRNA is presented as a ssRNA binding platform\textsuperscript{44–47}.

Once assembled, the PIC scans the 5′ UTR in the 5′ to 3′ direction until it finds a start codon to match the Met-tRNA\textsubscript{i} anticodon in the P site of the decoding centre, at which point it is known as the 48S PIC. Scanning the UTR requires mRNA to pass through the 40S subunit in a single-stranded conformation and, as such, the process depends on helicase action\textsuperscript{46,48}. As well as eIF4A, other DEAD-box helicases, such as Dhx29 and Ddx3 in mammals, play a role in unwinding stable RNA structures to permit passage of the PIC\textsuperscript{49–51}. In addition to a single-stranded substrate, the maintenance of the PIC in an “open” conformation by eIF1 and eIF1A is thought to be essential for allowing the mRNA to thread through during scanning\textsuperscript{46,48}.

Recognition of the initiator AUG, and scanning past intervening AUGs or near-cognate codons, is crucial for determination of reading frame and translation of the correct open reading frame (ORF). Recognition is predominantly facilitated by base pairing of the Met-tRNA\textsubscript{i} anticodon with the
AUG$^{37,52,53}$, and selectivity is conferred by eIF1, eIF1A and eIF5 (references $^{54-57}$). The nucleotide sequence surrounding the AUG is also important. In higher eukaryotes, the optimum sequence context, known as the Kozak consensus sequence$^{58}$, is 5'-C(A/G)CC AUGG-3' (where the initiation codon is underlined and the A is designated as position +1)$^{54}$. Contributions of these surrounding nucleotides to AUG recognition are facilitated by interactions with a modified base in the Met-tRNA$_i$ and likely also by interactions with initiation factors$^{54-57,59,60}$.

Upon recognition of the start codon, a sequence of events is triggered which culminates in association of the large ribosomal subunit. The formation of the duplex between the start codon and the Met-tRNA$_i$ anticodon stabilises the interaction of the TC with the 40S subunit and stimulates conformational changes in the PIC that displace eIF1 (references $^{37,52,53}$). This marks the transition from the open, scanning-compatible conformation to the closed conformation. During scanning, the GTP in the ternary complex is hydrolysed, but eIF1 prevents the release of P$_i$ until it is displaced in the closed state, after which P$_i$ dissociation blocks reassociation of eIF1 (references $^{29,55}$). GDP-bound eIF2 has a lower affinity for Met-tRNA$_i$ than its GTP-bound form and, stimulated by eIF5B, eIF2-GDP dissociates in complex with eIF5 (references $^{59,61,62}$). The PIC is bound by eIF5B-GTP and this accelerates joining of the 60S subunit to make the 80S initiation complex (IC)$^{63,64}$. Hydrolysis of the GTP associated with eIF5B reduces its affinity for the 80S IC, triggering eIF5B release which, in turn, stimulates the dissociation of eIF1A$^{64-66}$. This marks the completion of initiation and the resultant complex moves on to the elongation stage of translation.

To promote future translation initiation events, eIF2-GDP must be recycled back to eIF2-GTP, a process which is highly sensitive to regulation. Its recycling is catalysed by the guanine nucleotide exchange factor eIF2B, which exchanges GDP for GTP$^{67}$. When eIF2 is phosphorylated on Ser-51 of the α subunit, which occurs as a response to cellular stresses such as starvation or viral infection, it tightly binds and inhibits eIF2B, preventing the recycling of eIF2-GDP to eIF2-GTP and translation initiation$^{68-70}$. Due to the greater abundance of eIF2 compared to eIF2B, phosphorylation of even a small proportion of the cellular pool of eIF2α can have a profoundly inhibitory effect on translation initiation$^{70}$.

Several variations on the canonical, cap-dependent mechanism described above have been documented. On the simple end of the spectrum, initiation can occur using a non-canonical initiation codon. In these cases, the mechanism is usually the same as that described above, including the use of Met-tRNA$_i$ as the first aminoacyl-tRNA$^{71}$. Initiation is less efficient at non-canonical AUG codons, although efficiency varies according to the particular codon and its context. CUG is usually considered the most efficient, reaching values of up to 70% under some circumstances (although typically efficiency is lower)$^{71-78}$. Other fairly efficient (up to 15% dependent on Kozak context) non-canonical initiation codons are ACG, GUG, AUA and AUU$^{71,74-78}$. 
On the more complex end of the spectrum of non-canonical initiation events, viruses employ a variety of non-canonical initiation mechanisms to facilitate complex polycistronic gene expression strategies and, often, continued translation in the face of host translational shut-off. Internal ribosome entry sites (IRESs) provide a well-studied example of this. These are highly structured regions of RNA which recruit the ribosome without requiring the m'G cap, and often also obviate the need for particular cellular initiation factors. Two other non-canonical initiation events which are commonly employed by RNA viruses are leaky scanning and re-initiation (reviewed in reference ). In leaky scanning, a proportion of ribosomes scan past the first start codon on the mRNA and initiate instead at a downstream start codon. Re-initiation can occur after translation of a short ORF, if the 40S subunit does not dissociate but instead resumes scanning, reinitiating downstream.

1.1.3 Translation elongation

After initiation has completed, the ribosome is positioned with Met-tRNA, in the P site and both the A and E sites empty, after which the cycle of translation elongation begins (Figure 1.3A) and repeats until termination. In contrast to the numerous factors required for translation initiation, the elongation phase requires only two eukaryotic elongation factors (eEFs): eEF1A (EF-Tu in prokaryotes) and eEF2 (EF-G in prokaryotes). Recently, eIF5A (EF-P in prokaryotes), has also been shown to assist with elongation (further information below), although it is not essential. In fungi, an additional factor, eEF3, is required, which promotes dissociation of the deacylated tRNA in the E site, but this factor has no counterpart in higher eukaryotes and will not be discussed further.
The elongation cycle begins with binding of the cognate aminoacyl-tRNA in the A site of the ribosome, delivered by eEF1A in the form of a ternary complex along with GTP. Several non-productive sampling events can occur, in which non-cognate aminoacyl-tRNAs enter the A site and briefly bind the codon within, but these complexes are labile and rapidly dissociate. When the cognate aminoacyl-tRNA binds, the complex is stabilised, not only by the greater energetic favourability of the codon-anticodon duplex itself, but also by interactions between this duplex and the ribosome, maintaining the fidelity of translation. Residues A1824, A1825 and G626 (in rabbit ribosomes) of the 18S rRNA interact with the minor groove of the codon-anticodon helix to stabilise it, monitoring the Watson-Crick base pairing at positions 1 and 2 in the codon and leaving the “wobble” base at position 3 free to make non-canonical base pairs. Further stabilisation may occur via a conserved His residue in the N-terminus of a eukaryote-specific ribosomal protein, eS30, which becomes ordered and extends into the decoding centre upon binding of the cognate aminoacyl-tRNA to the ribosome. In a second kinetic mechanism of ensuring translational fidelity, the formation of the cognate complex rapidly stimulates irreversible hydrolysis of the GTP bound to eEF1A. This causes conformational changes in the ribosome (known as accommodation) that allow the acceptor stem of the aminoacyl-tRNA to enter the A site of the PTC, and promotes dissociation of eEF1A-GDP. Recycling of eEF1A-GDP to eEF1A-GTP is catalysed by the guanine nucleotide exchange factor eEF1B. Accommodation is rapidly followed by peptide bond formation between the aminoacyl-tRNA in the A site and the peptidyl-tRNA in the P site, transferring the nascent peptide from the latter to the former. The structure of the PTC is highly conserved between prokaryotes and eukaryotes and functions to position the substrates favourably for peptide bond formation, which occurs by nucleophilic attack by the amino group of the aminoacyl-tRNA on the ester linkage of the peptidyl-tRNA. Interestingly, eIF5A (also known as eEF5), which was originally identified as a translation initiation factor, has recently been shown to bind the ribosomal E site and interact with the acceptor stem of the peptidyl-tRNA in the P site to assist with substrate positioning during
elongation\textsuperscript{102–104}. This interaction is mediated by the polyamine-derived amino acid hypusine, formed by post-translational modification of a conserved Lys residue in eIF5A - the only protein in higher eukaryotes known to use hypusine\textsuperscript{104–107}. The involvement of eIF5A is not essential for protein synthesis but speeds it up by a factor of two-three\textsuperscript{108}, and is particularly beneficial for synthesis of motifs with sub-optimal peptide transfer such as poly-proline motifs (due to the restricted geometry of the proline imino acid)\textsuperscript{84,104,106,109,110}. Despite not being strictly essential for translation, eIF5A is present in eukaryotic cells at more than twice the concentration of the ribosome itself, and binds the ribosome very strongly, likely contributing to most rounds of the elongation cycle\textsuperscript{83,84,111,112}.

After peptide bond formation, the complex of the two tRNAs and the mRNA (sometimes known as tRNA\textsubscript{2}-mRNA) is moved one codon onwards (5′-wards) through the ribosome in a multi-step process called translocation (Figure 1.3B). In order for reading frame to be maintained, the contacts between the mRNA codons and tRNA anticodons must be preserved during this series of ribosomal movements\textsuperscript{113}. The first step of translocation is a rotation of the small ribosomal subunit relative to the large subunit, generating a “rotated” or “hybrid” state, in which the acceptor stems of the A and P site tRNAs have progressed to the subsequent (P and E) sites in the PTC but the anticodon stem loops remain in place\textsuperscript{114–116}. This state, denoted as A/P P/E, is reached by proceeding through an intermediate A/A P/E state in which only the P site tRNA acceptor stem has moved from the classical position\textsuperscript{115,117}. The ribosome is “unlocked” from its classical state by deacylation of the P site tRNA during peptide bond formation and is then free to rotate reversibly between the hybrid and classical states without the requirement for elongation factors\textsuperscript{118,119}. In relation to translocation, both states are known as PRE states: PRE classical and PRE-2 hybrid. Upon eEF2 binding, the PRE-2 hybrid ribosome undergoes (at least) three sets of conformational changes through three translocation intermediates, POST-1, -2 and -3 (reference \textsuperscript{100}). In POST-1, the 40S body, which is considered fully rotated in the PRE-2 hybrid state, partially back-rotates, although the head domain remains highly swivelled. This generates a conformation in which the tRNAs occupy chimeric ap/P pe/E positions where the anticodon stem loops of the tRNAs occupy a unique binding site on the small ribosomal subunit which is part-way between the classical and hybrid positions\textsuperscript{100,120}. In the transition from POST-1 to POST-2, the 40S subunit rotates further backwards, almost all the way to the classical state position, although the 40S head domain only partially follows, resulting in an exaggerated swivel between the body and head domains of the 40S subunit and the maintenance of the chimeric ap/P pe/E positions of the tRNAs\textsuperscript{100}. POST-3 is almost indistinguishable from the fully completed POST state: the tRNAs assume the classical post-translocation P/P E/E positions, the 40S body completes back-rotation to the classical state, and the head back-swivels to its classical state but with a small residual tilt\textsuperscript{100}. Translocation is completed, and the full POST state achieved, when eEF2
dissociates from the ribosome, leaving the A site vacant and ready for the next round of elongation to commence with accommodation\textsuperscript{100}.

The role of the elongation factor in translocation has been described as that of a “doorstop”, preventing the back-rotation of the small ribosomal subunit from moving the tRNA\textsubscript{2}-mRNA complex backwards as well, thus allowing the ribosome to move onwards relative to the mRNA\textsuperscript{121,122}. The uncoupling of the ribosome from the tRNA\textsubscript{2}-mRNA complex is promoted by eEF2 entering the A site and disrupting the interaction between the codon-anticodon duplex and A1824 and A1825 of the 18S rRNA\textsuperscript{123–126}. In eukaryotes, the modification of a conserved His residue to diphthamide in the domain of eEF2 that binds the decoding centre has been suggested to play a role in this process, and in maintaining the fidelity of reading frame\textsuperscript{100,127,128}. Hydrolysis of the GTP bound to eEF2/EF-G appears to play a different role in eukaryotes compared to prokaryotes. In eukaryotes, the POST-3 state can be reached without GTP hydrolysis, and its main contribution seems to be promoting dissociation of eEF2 to form the POST complex\textsuperscript{100}. In bacteria, translocation can fully complete, although considerably more slowly, in the absence of GTP hydrolysis, the contribution of which is to trap the ribosome in the POST state and prevent reversion to previous translocation intermediate states\textsuperscript{100}. These differences may be related to E site occupancy\textsuperscript{100}, although the trigger (or lack thereof) for dissociation of the deacylated tRNA from the E site has been the subject of considerable debate\textsuperscript{100,129–134}. A popular theory was that it is coupled to the arrival of the ternary complex in the A site; however, for prokaryotes it appears this is a relatively minor pathway, relevant only for the first few cycles after initiation\textsuperscript{131}. It is thought that prokaryotic E site tRNA dissociation is generally rapid and spontaneous\textsuperscript{131–134}, and can occur before translocation reaches the full POST state\textsuperscript{100,132}, although kinetics are strongly dependent on factors such as tRNA identity and buffer composition\textsuperscript{131,133}. In contrast, dissociation of the E site deacyl-tRNA in eukaryotes is thought to be slower\textsuperscript{100}. Some evidence favours the A site allostery hypothesis\textsuperscript{129}, although another study found that deacyl-tRNA remained associated with the eukaryotic ribosome until the inter-subunit rotation of the following round of translocation\textsuperscript{117}.

Although the mechanisms are less well-studied than for initiation, elongation is also subject to regulation, by phosphorylation and consequent inactivation of eEF2. In metazoans, the calcium/calmodulin-dependent kinase eEF2K phosphorylates eEF2 on the conserved residue Thr-56, inhibiting its ability to stimulate elongation\textsuperscript{135}. This kinase represents a convergence point for several pathways to regulate translation elongation in response to various stimuli, such as nutrient starvation (via mTORC1 and/or AMPK)\textsuperscript{136,137}. Phosphorylation of eEF2 is similarly used to inhibit translation in response to multiple cellular stresses in yeast\textsuperscript{138–140}.

The elongation stage of translation can be subverted to promote non-canonical translation events known as “recoding”, in which the canonical decoding of triplet codons is broken and alternative polypeptides are synthesised. Due to the pressure to maintain small genomes, events such as these
are particularly common in viruses as they often facilitate complex gene expression strategies without requiring large stretches of untranslated regulatory sequences. Examples of this include programmed ribosomal frameshifting (PRF), discussed in section 1.2, and StopGo, in which a specific nascent peptide motif promotes failure of peptide bond formation and results in synthesis of two separate proteins without encountering a stop codon or undergoing canonical termination and re-initiation in between. These and further examples, such as stop codon readthrough, are reviewed in Firth & Brierley, 2012, Atkins et al., 2016, Dever et al., 2018 and Rodnina et al., 2020 (references 79,83,143,144).

1.1.4 Translation termination and ribosome recycling

In the final stages of translation, the nascent peptide is released (termination), and the ribosomal subunits separated from each other, ready to be reassembled for the next round (recycling) (Figure 1.4).

Figure 1.4 Mechanism of canonical eukaryotic translation termination and ribosome recycling

Figure adapted from Rodnina 2010 (reference 158).

Termination is triggered when the A site of the ribosome is occupied by a stop codon: UAA, UAG or UGA. These (in most eukaryotes) do not have cognate aminoacyl-tRNAs, and the A site is instead bound by eukaryotic release factor (eRF) 1. This protein has several conserved elements which specifically recognise termination codons, as well as a conserved GGQ motif that coordinates a water molecule in the PTC, positioning it to promote nucleophilic attack on the nascent peptidyl-RNA to facilitate nascent peptide release. Like tRNAs, eRF1 is presented to the ribosome in complex with a GTPase, in this case eRF3, which promotes termination by a mechanism dependent on GTP hydrolysis, and promotes dissociation of eRF1 to facilitate further rounds of termination.
Although eRF1 and eRF3 are the only core peptide release factors, termination can also be enhanced by other proteins. The elongation factor eIF5A has also been shown to promote termination, potentially in an analogous way to how it stimulates the closely related nucleophilic attack of peptide bond formation. ABCE1 (Rli1 in yeast), a key factor involved in ribosome recycling (discussed below) also stimulates termination by binding eRF1 to stabilise the active conformation and stimulate catalytic activity.

Termination is the final stage of translation, by which point most of the energy and metabolites involved in protein synthesis have already been expended. As such, it is not usually subject to regulatory pathways in response to stimuli; however, certain features do affect termination efficiency, with failure to terminate manifesting as readthrough, in which a near-cognate aminoacyl-tRNA is accommodated, or frameshifting, both resulting in continued translation.

The nucleotide sequence at the stop codon affects termination efficiency but, interestingly, this is not confined to the three bases of the stop codon itself and extends several bases downstream. The identity of the nucleotide directly after the stop codon (the +4 nucleotide) has a particularly strong impact on termination efficiency, likely conferred by its interaction with the 18S rRNA. Protein interactions also play a role. The initiation factor eIF3 and its associated factor/subunit eIF3J (Hcr1 in yeast) associates with the pre-termination complex to interfere with decoding of the wobble base at the stop codon, promoting amino acid misincorporation and stop codon readthrough. PABP interacts with eRF3 and promotes eRF1 and eRF3 recruitment to the ribosome and stimulates peptidyl hydrolysis, increasing termination efficiency.

After release of the nascent peptide chain, the 80S ribosome must be disassembled into its component subunits so it can be reused. This process is called ribosome recycling, and is primarily catalysed by ABCE1, which enters the A site and undergoes conformational changes, triggered by ATP binding and/or hydrolysis, that drive the ribosomal subunits apart. When in the A site, ABCE1 contacts eRF1, presenting a potential link between termination and recycling by stimulating eRF1 termination activity as described above. Interestingly, ABCE1 may present another link with a different part of the translation process, as it has been shown to interact with several initiation factors and has been found in a cryo-EM structure of the 48S pre-initiation complex (where it was originally mis-identified as eIF3i/g).

1.1.5 Translation quality control

There are several pathways that can be triggered to protect the cell from deleterious, aberrant proteins when problems are detected in translation or the transcript itself. This section focuses on pathways which resolve problems in translation, in particular ribosomal stalling within ORFs (internal stalls). There are three possible aims of pathways that resolve these stalls: dissociation (rescue) of the
ribosomal subunits so these can be recycled, degradation of the incomplete nascent peptide to avoid deleterious effects, and degradation of the mRNA upon which translation stalled in case it has an error\textsuperscript{177,178}.

Recently, it has come to light that stalled ribosomes are likely recognised by virtue of causing collisions, in which a “trailing” or “colliding” ribosome catches up with the “leading” stalled ribosome and is consequently itself stalled, forming a disome (di-ribosome)\textsuperscript{177}. Further ribosomes may collide with the trailing ribosome to form trisomes or higher order ribosome queues. The presence of collided ribosomes presumably helps distinguish slow but continuing translation from translation which is stalled potentially irrecoverably\textsuperscript{179}. Collided ribosomes present a unique interface which is thought to be specifically recognised by ZNF598 (Hel2 in yeast), an E3 ligase that ubiquitinates ribosomal proteins eS10 and uS10, facilitating recruitment of other factors to trigger downstream pathways\textsuperscript{180–185}. Interestingly, structural studies revealed differences in the collided ribosome interface in disomes resulting from slow ongoing translation compared to those formed at strong stall-inducing motifs, which may further contribute to preventing over-zealous quality control interfering with non-aberrant translation\textsuperscript{181,186}.

The recruitment of ZNF598 at collided ribosomes and the resultant ribosomal ubiquitination is the first stage of (at least) two distinct pathways, of which the most frequently triggered is the ASCC pathway (Figure 1.5, upper)\textsuperscript{179,183,187,188}. In mammalian cells, this is carried out by the ASC-1 complex (ASCC), composed of ASCC3, ASCC2 and TRIP4, while in yeast the homologous proteins Slh1, Cue3 and Rqt4 form the analogous ribosome QC trigger complex (RQT)\textsuperscript{179,183,187}. This complex performs its actions on intact mRNAs, rescuing the leading ribosome in a pair (or higher order queue) of collided ribosomes, allowing the trailing ribosome(s) to continue translation\textsuperscript{179,187}. ASCC3 is an RNA helicase, and the splitting of the ribosomal subunits depends on its ATPase activity\textsuperscript{179,187}. Once the ribosomal subunits have been split, the peptidyl-tRNA bound to the 60S subunit triggers ribosome-associated quality control (RQC) to degrade the potentially deleterious incomplete nascent peptide\textsuperscript{189–191}. NEMF (Rqc2 in yeast) recognises the exposed peptidyl-tRNA and recruits alanine- and threonine-charged tRNAs to the A site to catalyse elongation of the nascent peptide, independent of the mRNA or the small ribosomal subunit, generating a carboxy-terminal Ala and Thr tail (CAT tail)\textsuperscript{192}. This, in turn, causes the recruitment of Listerin (Ltn1 in yeast), an E3 ubiquitin ligase which ubiquitinates the nascent peptide, targeting it for degradation by the proteasome\textsuperscript{193}.

The second pathway triggered by ZNF598 recognition of collided ribosomes is no go decay (NGD), which functions to degrade the stall-inducing mRNA (Figure 1.5, lower). Like many translation quality control pathways, this has predominantly been characterised in yeast, in which Cue2 was found to cleave mRNA in the A site of internally stalled ribosomes, in a mechanism dependent on the yeast ZNF598 homologue, Hel2 (references \textsuperscript{181,194}). Tlr1, a polynucleotide kinase, then
phosphorylates the 5'-OH of the cleaved mRNA to facilitate further processing by exonucleases\textsuperscript{195}. The key players in this process in mammalian cells have not yet been confirmed, but N4BP2 is a likely candidate. It is a Cue2 homologue that contains an additional polynucleotide kinase domain thought to perform a similar function to Tlr1, and it has been shown to physically associate with colliding ribosomes\textsuperscript{196,197}. Based on comparisons with canonical mRNA turnover mechanisms, it is likely that the cleaved mRNA is degraded by exonucleases, after decapping to make the 5' end available. Although there has been some suggestion that, in yeast, Syh1 may recruit the decapping complex used in canonical mRNA decay (Dcp1/2), this is currently unconfirmed, as are the identities of any potential exonucleases involved, although evidence implicates Xrn1 (references \textsuperscript{177,198}).

NGD also triggers ribosome rescue by Pelota (Dom34 in yeast), Hbs1 and ABCE1. Pelota is a parologue of eRF1 and, analogously to ribosome recycling, works together with ABCE1 to split the ribosomal subunits in an ATP-dependent manner\textsuperscript{199,200}. This is stimulated by the eRF3 parologue Hbs1, although this protein is not essential\textsuperscript{199}. Interestingly, the triggering of this ribosome rescue pathway is dependent on the mRNA being shortened, providing a likely mechanistic link with the endonucleolytic cleavage that defines NGD\textsuperscript{199,201–204}. After subunit dissociation, the 60S subunit with associated peptidyl-tRNA is free to trigger RQC and nascent peptide degradation\textsuperscript{178}.

Finally, a pathway triggered by collided ribosomes that could act independently of ZNF598 was recently characterised, in which EDF1 was found to be an adaptor protein that binds to collided ribosomes and recruits GIGYF2. GIGYF2 recruits an alternative cap-binding protein 4EHP (also known as elf4E2), which competes with cap-binding initiation factors to inhibit translation initiation (in cis), decreasing the ribosomal load on the mRNA to prevent severe “traffic jams” and synthesis of incomplete proteins\textsuperscript{197,205–207}. Interestingly, evidence for a second, ZNF598-dependent method of GIGYF2/EDF1-recruitment was also found, which did not depend on the ubiquitinating activity of ZNF598 (reference \textsuperscript{205}). In yeast, Syh1 and Smy2 are homologues of GIGFY2, although they seem to play a slightly different role, predominantly triggering mRNA decay in a Hel2-dependent manner\textsuperscript{177,205}. Additionally, there is no known homologue of 4EHP in yeast so a mechanism of translation inhibition in this instance is not obvious\textsuperscript{177}. Although there was initially no evidence to suggest the GIGFY2/4EHP pathway leads to mRNA decay in mammals\textsuperscript{197,205,206}, later experiments suggest this does occur\textsuperscript{208}.

Despite the clear evidence of ZNF598 recognition triggering the above quality control pathways on problematic transcripts, it appears to be a pathway which is rarely triggered by endogenous substrates, with deletion of Hel2 in yeast affecting very few endogenous substrates\textsuperscript{187,194,209}. Recent evidence suggests a key role may be to recognise ribosomes stalled on endogenous transcripts encoding secretory proteins, to prevent continued synthesis (and resultant mistargeting) of proteins in which the signal recognition peptide has not been correctly recognised\textsuperscript{188}. 
Figure 1.5 Quality control pathways triggered by ribosome collisions

Figure adapted from D’Orazio et al., 2021 (reference 177). Proteins are labelled internally with yeast names and externally with mammalian names. **Upper**: ASCC (ASC-1 complex) pathway. **Lower**: No go decay (NGD). RQC = ribosome-associated quality control, RQT (yeast equivalent to ASCC) = RQC trigger complex.
1.2 Programmed ribosomal frameshifting

Programmed ribosomal frameshifting (PRF) is the slippage of the ribosome out of the 0 frame into an alternative reading frame, stimulated by conserved elements to fulfil a particular biological purpose, such as to produce a distinct protein. It was first discovered in Rous sarcoma virus, a retrovirus in which −1 PRF, where the ribosome slips backwards (towards the 5′ end of the RNA) by a single nucleotide, facilitates expression of the Gag-Pol fusion protein instead of just Gag\textsuperscript{210}. In this case, frameshifting is used to provide a means of controlling the stoichiometry of these two proteins, as PRF is approximately 5% efficient and leads to a 20:1 ratio of Gag:Gag-Pol\textsuperscript{210}. Subsequently, PRF was found to be commonly used by RNA viruses from many different taxa to control stoichiometry of viral proteins, often the replicase\textsuperscript{143}. The use of frameshifting in viruses is not restricted to −1 PRF – several viruses, such as influenza A virus, employ +1 PRF, in which the ribosome slips forwards by a single nucleotide\textsuperscript{211,212}.

PRF sites are also found in cellular genes from all domains of life. Although the majority of these are in mobile chromosomal elements, there are some known cases in stable cellular chromosomal genes. In Escherichia coli, the dnaX gene, which encodes two subunits of the DNA polymerase, is a well-studied example of cellular −1 PRF\textsuperscript{213-215}. Ribosomes which translate the full-length, 0-frame ORF synthesise the τ subunit, while ribosomes that frameshift at the −1 PRF site two thirds of the way through the ORF terminate translation one codon later and synthesise a truncated protein: the γ subunit. Frameshifting at this site is \~50% efficient, meaning a 1:1 ratio of these subunits is produced\textsuperscript{213-215}. In mammals, a small number of retrotransposon −1 PRF sites have become stably integrated into the genome and produce functionally important proteins, for example PEG10 (also known as Edr) is thought to be important for placental development\textsuperscript{216,217}.

Although +1 PRF is not utilised by any of the viruses studied in this thesis, and is discussed here only briefly, it is an interesting phenomenon. In an elegant example of the use of PRF as a cellular regulatory mechanism, +1 PRF is required for synthesis of the full-length protein ornithine decarboxylase antizyme (OAZ, often known just as antizyme) in many higher eukaryotes\textsuperscript{218-220}. In mammals, PRF at the antizyme +1 PRF site operates at low efficiency in the absence of extra stimulation, but its efficiency is greatly increased by an abundance of polyamines, providing a feedback mechanism to regulate the levels of antizyme, an inhibitor of polyamine synthesis and uptake\textsuperscript{218,219}. Another interesting example of +1, and +2, PRF is presented by the ciliate Euplotes crassus, which has over 1500 putative +1/+2 PRF sites in its transcriptome, with some transcripts predicted to contain more than one PRF site\textsuperscript{221}. In addition to its surprising prevalence, +1/+2 PRF in members of the Euplotes genus is thought to be \~100% efficient\textsuperscript{143,221}. The mechanism(s) of +1 PRF are distinct from that of frameshifting in the reverse direction and will not be discussed in detail, but in some cases it involves disrupted translation termination causing pausing over a stop codon,
Introduction

which is resolved by frameshifting and incorporation of an amino acid in the +1 frame (instead of termination in the 0 frame). In *Euplotes*, this is thought to be a consequence of reassignment of the UGA codon from a stop codon to a cysteine, accompanied by evolution of eRF1 to prevent UGA recognition, that also resulted in weakened recognition of the remaining UAA and UAG stop codons\(^{222,223}\).

1.2.1 Canonical mechanisms of −1 PRF

Despite much research, the mechanism of PRF is not well understood. The search for mechanistic understanding is partly impeded by the fact that different PRF sites likely operate by slightly different mechanisms, leading to ostensibly conflicting results. Most postulated mechanisms of −1 PRF broadly fall into two categories, depending on the stage of elongation during which the change in frame occurs\(^{144}\) (Figure 1.6). The mechanism thought to be used most often is slippage of both the A and P site tRNAs into the alternative frame during translocation, referred to herein as the “tandem slippage” model\(^{224–227}\). An alternative mechanism, often called “hungry codon” frameshifting, is characterised by slippage of a single tRNA in the P site while the A site is empty during the decoding stage of elongation\(^{228–230}\). Two features are central to both mechanisms. The sequence over which the change in frame occurs is known as a “slippery” sequence (SS) – a short sequence of mRNA which permits re-pairing of the tRNA anticodon(s) in the new reading frame without a large energetic penalty\(^{143,144}\). The second key feature is a frameshift-stimulatory element in the mRNA, which is usually a sequence that causes the ribosome to pause over the slippery sequence, stimulating the frameshift\(^{143,144}\).

PRF sites thought to operate using the tandem slippage mechanism have slippery sequences based on the consensus X\_XXY\_YYZ, where underscores represent boundaries between 0-frame codons, XXX represents any three identical nucleotides, YYY represents AAA or UUU, and Z represents A, C or U\(^{79,143}\). Upon slippage into the −1 frame, this sequence allows re-pairing with mismatches only at the less stringent wobble bases, meaning it is not highly thermodynamically and geometrically unfavourable\(^ {231}\). As the fidelity of the codon-anticodon duplex is not monitored as strictly in the P site as it is in the A site, deviations from the consensus motif are more common in the XXX positions\(^ {91,143}\). In canonical eukaryotic −1 PRF, the frameshift-stimulatory element at sites that use tandem slippage is a pause-inducing stable secondary RNA structure – usually a pseudoknot but sometimes a stem loop\(^ {143,232,233}\). This structural element is downstream of the slippery sequence, separated from it by a 5-9 nt “spacer” which is thought to position the structure such that it pauses the ribosome with the P and A sites over the XXY and YYZ of the slippery sequence, respectively\(^ {20,22,234,235}\). In some prokaryotic −1 PRF sites, such as *DnaX*, there is an additional pause-
inducing element upstream of the slippery sequence, which resembles the Shine-Dalgarno (SD) sequence that binds the prokaryotic ribosome during initiation. The specific mechanism by which the secondary RNA structure stimulates frameshifting is usually attributed to it presenting a topological problem for the ribosome, creating a “roadblock” as the ribosome struggles to unwind the structure. In many tandem slippage models, this is proposed to create tension in the mRNA which leads to disruption of the codon-anticodon duplex and facilitates the slippage into an alternative reading frame. Detailed studies of the kinetics of translation of frameshift cassettes (sequences containing both the slippery sequence and the frameshift-stimulatory element) have informed mechanistic understanding. Although there are some differences between the results of various studies, potentially due to the use of different frameshift cassettes and experimental conditions, some common themes emerge. PRF at tandem slippage sites is characterised by a lengthened pause during translocation, while the ribosome is in the rotated state or a non-canonical hyper-rotated state, during which time the

**Figure 1.6 Mechanisms of canonical −1 PRF**

Figure adapted from Rodnina *et al.*, 2020 (reference 144). **Upper:** tandem slippage mechanism. **Lower:** hungry codon mechanism.
translocase (EF-G/eEF2) binds (and GTP hydrolysis occurs) multiple times in unproductive attempts to complete translocation\textsuperscript{226,227,240,241}. In some models, the back-swivel of the head of the small ribosomal subunit is prevented by the secondary RNA structure\textsuperscript{22,225}. Slipping backwards into the $-1$ frame places the ribosome in a different orientation with respect to the secondary structure, and is thought to make resolution of the elongation blockade more favourable, allowing the ribosome to progress\textsuperscript{225,237}. A caveat of interpreting these detailed mechanistic explorations is that they are usually performed \textit{in vitro} using prokaryotic ribosomes. Although some eukaryotic PRF sites function in bacterial systems\textsuperscript{242–244}, there may be some differences in the mechanistic details of prokaryotic and eukaryotic frameshifting, given the slight variations in the mechanism of canonical translation elongation between the two systems (section 1.1.3). This is particularly pertinent to experiments performed using the \textit{DnaX} frameshift cassette, due to its additional SD-like frameshift-stimulatory element.

Hungry codon $-1$ PRF does not require a downstream secondary RNA structure, but instead the elongation blockade is presented by a “hungry” codon: a codon which takes a long time to decode\textsuperscript{228,230,245}. This is usually a “rare” codon, for which the aminoacyl-tRNA is not abundant in the cell, although stop codons are also thought to be able to fulfil this function\textsuperscript{226,230,245,246}. During frameshifting, the slow decoding of the hungry codon in the A site causes the ribosome to pause with an empty A site and the P site positioned over the slippery sequence, which is directly upstream\textsuperscript{228,229,245}. Slippage into the $-1$ frame changes the identity of the codon in the A site, obviating the need for the scarce aminoacyl-tRNA and allowing translation to continue. Due to the empty A site, only the P site tRNA is required to re-form the codon-anticodon duplex in the alternative frame and, as such, the slippery sequence motif is only four nucleotides long\textsuperscript{228,229,247,248}. Similarly to slippery sequences that facilitate tandem slippage, tetranucleotide sequences which permit favourable re-pairing of the tRNA (in this case only the P site tRNA) with the codon in the new reading frame promote hungry codon PRF, for example AAAA, AAG or UUU\textsuperscript{228,229,248,249}.

Due to the distinct mechanistic features required, canonical PRF sites which operate by tandem slippage should be relatively easily distinguished from those which operate by the hungry codon mechanism. For example, all the PRF sites examined in this thesis have a heptanucleotide slippery sequence, separated from a downstream translation blockade by a short spacer, indicating they likely operate by a tandem slippage mechanism. However, there is interesting evidence that, in some cases, a single PRF site can facilitate both mechanisms: under conditions of saturating aa-tRNA, tandem slippage is the predominant mechanism, and when the concentration of the A site aa-tRNA is limiting, hungry codon frameshifting is used (Figure 1.6). This has been shown to occur for HIV-1, \textit{DnaX} and Semliki forest virus $-1$ PRF sites and may be a strategy used by some viruses to ensure frameshifting can occur in a variety of conditions\textsuperscript{228,242,250–253}. 

- 20 -
Various factors can affect frameshift efficiency, although some are likely specific to particular PRF cassettes. Unsurprisingly, the energetics of base pairing between the codons and anticodons in the 0 and −1 frames plays a key role in determining PRF efficiency, as do the presence and location of the downstream secondary RNA structure at tandem slippage PRF sites. Frameshift efficiency correlates with the stability of the downstream secondary RNA structure, particularly in the region that is first unwound by the ribosome. Further, the precise geometry of the tertiary structure formed by PRF-stimulatory RNA is important, as mutations that alter tertiary geometry with minimal effect on secondary structure can have a significant effect on PRF efficiency. Although ribosomal pausing is considered by many to be an integral part of the frameshift mechanism, there is conflicting evidence on the correlation between pause duration and PRF efficiency, likely due to the contribution of other factors such as the phase of the elongation blockade relative to the ribosome. Some evidence has also suggested that the ability of RNA structures to sample different conformational states may play a role in stimulating PRF.

In some cases, translational features which are less intuitively linked to the mechanism of frameshifting have also been found to play a role. For example, in alphaviruses, co-translational folding and membrane integration of the nascent peptide has been shown to stimulate −1 PRF, in addition to the downstream stem loop. At the SARS-CoV-2 PRF site, mutations designed to disrupt the interaction between the nascent peptide and the peptide exit tunnel affect −1 PRF efficiency, as does the distance between the slippery sequence and the 0-frame stop codon downstream. The density of ribosomes on the transcript can also affect frameshift efficiency, thought to be related to ribosomal collisions and/or whether there is enough time for the frameshift-stimulatory structure to re-fold before the next ribosome passes through.

Recently, the first example of a cellular factor that inhibits −1 PRF was discovered: Shiftless, the product of a cellular interferon-stimulated gene, was found to inhibit −1 PRF at a variety of frameshift sites. Its mechanism of action is not completely characterised, although it was found to bind to both ribosomal subunits and the HIV-1 frameshift cassette (on which it inhibits PRF). It was proposed that the association of Shiftless with frameshifting ribosomes may cause extended pausing and premature termination mediated by eRF1 and eRF3 (reference). Shiftless (also previously known as C19orf66 and RyDEN) may have multiple modes of action, as it inhibits the replication of a range of viruses, including dengue virus, which does not encode a PRF site.

### 1.2.2 Non-canonical mechanisms of trans-activated −1/−2 PRF

In the previous section, examples were given in which additional elements of the transcript/nascent peptide, other than the canonical slippery sequence, secondary RNA structure, hungry codon or SD-like sequence can augment frameshifting. However, in some cases, trans-acting elements can
stimulate $-1/-2$ PRF. First shown in proof-of-concept studies on artificial PRF sites, some rare naturally occurring examples were later found.

PRF can be stimulated by binding of an antisense oligonucleotide downstream of the slippery sequence, in lieu of a secondary RNA structure. In vivo, miRNAs have been reported to play a regulatory role in $-1$ PRF in humans, by stabilising the pseudoknot that stimulates $-1$ PRF on the CCR5 transcript to increase PRF efficiency – although the miRNAs in this case augment PRF stimulated by an existing pseudoknot as opposed to acting in its stead. Additionally, concerns have been raised over whether the frameshifting observed at this site may have been artefactual.

An artificial construct based on the HIV-1 gag-pol frameshift site provided the first evidence that protein binding could stimulate $-1$ PRF. The frameshift-stimulatory pseudoknot was replaced with the iron-responsive element from ferritin mRNA, which is a stem loop that binds iron-regulatory proteins when cellular iron levels are low. Although the stem loop on its own was capable of stimulating $-1$ PRF, frameshift efficiency was somewhat enhanced by the binding of iron-regulatory proteins to the stem loop. In more recent studies, arteriviruses and cardioviruses were found to provide the only two known naturally occurring examples of protein-stimulated $-1$ PRF. The mechanism of arterivirus protein-stimulated frameshifting is described below, while the cardiovirus mechanism is described in Appendix section 6.

1.2.2.1 Protein-stimulated $-1$ and $-2$ PRF in arteriviruses

Arteriviruses are positive-sense, single-stranded RNA viruses in the Order Nidovirales. The arterivirus protein-stimulated PRF site was first characterised in porcine reproductive and respiratory syndrome virus (PRRSV), a virus that infects pigs and to which a full introduction is given in section 1.4. The key elements of the frameshift site are conserved in all other species of arterivirus except equine arteritis virus (EAV) and wobbly possum disease virus (WPDV) and it is likely that this novel mechanism of PRF is widely used within this family of viruses. The non-canonical PRF site is located within the nsp2 region of the ORF encoding the viral polyprotein. As such, it is referred to herein as the nsp2 PRF site, to differentiate it from the other PRF site on the arterivirus genome (the ORF1ab PRF site), which operates by a canonical, pseudoknot-stimulated mechanism.

The nsp2 PRF site (Figure 1.7A) is unusual for several reasons. First, it stimulates both $-1$ and $-2$ PRF, with frameshift efficiencies measured at 7% and 20%, respectively (at 24 hpi in MARC-145 cells infected with SD01-08 PRRSV at MOI 0.1), representing the first example of efficient use of $-2$ PRF in a eukaryotic system. This propensity to efficiently stimulate two different PRF events at the same site means that three proteins are produced: ribosomes which do not frameshift produce nsp2, while those which shift into the $-1$ or $-2$ frames produce nsp2N or nsp2TF, respectively. The functions of these proteins, and the importance of this PRF event for viral fitness, are discussed in
The slippery sequence of the nsp2 PRF site is G_GUU_UUU, which does not follow the X_XXY_YYZ consensus motif. This frameshift site facilitates perfect re-pairing in the A site in both the −2 and −1 frame, however re-pairing in the P site for both frames would entail one mismatch in the −1 frame and two or three mismatches in the −2 frame. The precise number of −2-frame mismatches depends on the identity of the base directly upstream of the heptanucleotide slippery sequence, which is usually either A or G. This is consistent with the more relaxed monitoring of fidelity in the P site compared to the A site, and with the observation that deviations from the consensus X_XXY_YYZ motif are more often tolerated at the P site codon.

Another interesting feature of this frameshift site is that, as described above, PRF is stimulated not by a secondary downstream RNA structure, but by binding of a trans-acting protein complex. The region downstream of the slippery sequence is predicted to be unstructured but contains a conserved C-rich region with consensus sequence CCCANCUCC, 11 nt downstream of the slippery sequence. This was found to bind a complex of the viral protein nsp1β and the cellular protein poly(C)-binding protein (PCBP) and is hypothesised to stimulate frameshifting by presenting an elongation blockade, in a way analogous to RNA structures in canonical −1 PRF.

The C-rich motif to which these proteins bind is not a canonical PCBP-binding motif and, indeed, it cannot stably bind PCBP without the presence of nsp1β, nor can nsp1β stably bind without PCBP. The relative importance of the bases in the C-rich motif for maintaining PRF efficiency has been determined, revealing the C residues as particularly important, along with the central A (Figure 1.7B).

The components of the PCBP/nsp1β elongation blockade have been reasonably well-characterised, although a high-resolution structure of the two proteins in complex with RNA is lacking. Ultra-centrifugation was used to characterise a complex of human PCBP2 with nsp1β and a 34 nt PRF cassette from European lineage PRRSV isolate SD01-08, and found a 1:1:1 ratio of nsp1β:PCBP:ssRNA. PCBP and nsp1β associate with each other in the absence of RNA, and the (proposed) RNA-binding residues of nsp1β and PCBP are not required to facilitate the protein-protein interaction, although they are required to facilitate stable binding of the complex to RNA. It should also be noted that there is likely some heterogeneity in the complexes formed by nsp1β/PCBP/RNA, as three classes of complex were found in the ultra-centrifugation experiment (although not all classes contained all three components). One abundant class of complex was proposed to represent monomeric nsp1β binding a single molecule of RNA, suggesting nsp1β may bind RNA in the absence of PCBP, although such complexes are not reliably detectable by electrophoretic mobility shift assays (EMSAs).
Introduction

Figure 1.7 PRRSV nsp2 PRF site

A) Sequence of the nsp2 PRF site from PRRSV isolate SD01-08 (European lineage). Figure reproduced from Li et al., 2014 (reference 281). The slippery sequence is printed in red, with P
and A site codons in each frame (0, −1 and −2) indicated by square brackets. In almost all sequenced PRRSV isolates there is a stop codon in the −1 frame overlapping the slippery sequence, which is underlined. The C-rich motif is printed in blue. B) The relative importance of bases in and around the C-rich motif for PRF efficiency. Figure reproduced from Napthine et al., 2016 (reference 282). The slippery sequence is printed in navy. Bases (indicated by subscript numbers) in and around the C-rich motif were mutated and the resultant inhibition of PRF was quantified, with larger font size indicating greater inhibition of PRF. C) Crystal structure of nsp1β from PRRSV isolate XH-GD (North American lineage). Figure adapted from Li et al., 2014 (reference 281), structure determined by Xue et al., 2010 (reference 288) (PDB ID: 3MTV). Left: cartoon representation of dimer structure. The N-terminal domain of each monomer is coloured purple, PLP1β domain in green, the C-terminal extension in red, and helix α4 (within the PLP1β domain), which contains the 123-GKYLQRRLQ-131 motif, in orange with basic residues represented as sticks. Right: electrostatic surface potential of the nsp1β dimer, with the positively charged (blue) region formed by the basic residues in the 123-GKYLQRRLQ-131 motif indicated on each monomer by an orange box. These have been suggested to form a potential surface for RNA binding. D) Amino acid alignment of nsp1β sequences from several arterivirus isolates. Figure reproduced from Li et al., 2014 (reference 281). Secondary structure elements from the crystal structure in panel C are indicated above the alignment in corresponding colours. Only a section of the protein sequence is depicted, and the active site cysteine and histidine of the PLP protease domain are indicated by asterisks. E) Domain organisation of PCBP family proteins. Figure reproduced from Napthine et al., 2016 (reference 282), in which it was made by adapting Figures from Makeyev et al., 2002 (reference 289) and Choi et al., 2009 (reference 290). Underneath the KH domain of hnRNP K the general domain architecture of KH domains is illustrated. The KI (K-protein interactive) domain is responsible for many interactions between hnRNP K and other proteins, and the KNS (hnRNP K-specific nuclear shuttling) domain is responsible for bidirectional shuttling of hnRNP K through the nuclear pore complex289.

The crystal structure of PRRSV nsp1β (isolate XH-GD from the North American lineage) is known and is a homo-dimer (Figure 1.7C), in contrast to the monomeric nsp1β suggested to form the nsp1β:PCBP:ssRNA in the ultra-centrifugation study288,291. A GKYLQRRLQ motif was identified, which is conserved (to some extent) across all arterviruses except EAV and WPDV which, as noted above, also lack the nsp2TF frameshift signal281,287 (Figure 1.7D). The motif forms an α-helix and the tyrosine residue and the first arginine residue (in bold) were found to be essential for stimulating PRF in both PRRSV and SHFV281,287. The positively charged residues within the GKYLQRRLQ motif have been suggested to form a surface for RNA binding281. A later study, which extended investigation of this motif a little further downstream, identified another basic residue critical for stimulation of PRF in PRRSV: a downstream arginine (in bold in the extended motif GKYLQRRLQVNGLR) which is part of a beta strand in the nsp1β crystal structure291. Although they are not essential for PRF, mutation of the other two basic residues (the lysine or the second arginine) in this motif reduced PRF efficiency, further supporting the importance of positively charged residues287,291.

As for the cellular protein involved in stimulating PRF, PCBPs are a family of proteins with five members in mammals: PCBP1-4, also known as α-CP or heterogeneous nuclear ribonucleoprotein (hnRNP) E proteins, and hnRNP K/J which is somewhat divergent from the other PCBPs289,290. The domain architecture of members of this family is characterised by a “beads on a string” arrangement.
Introduction

of three hnRNP K homology (KH) domains, the first two of which are separated by a short linker, followed by a longer linker between the second and third\(^{289,290}\) (Figure 1.7E). No full-length crystal structures of PCBPs have been characterised, although several structures of individual PCBP KH domains have been obtained, including some in complex with oligonucleotides\(^{291,293–295}\). PCBP1, 2 and 3 can stimulate PRF at the nsp2 site \textit{in vitro}, although this stimulatory effect has only been verified for PCBP1 and 2 in infected cells\(^{282}\). PCBP4 and hnRNP K/J are not able to stimulate PRF at this site\(^{282}\). The RNA-binding residues of PCBP1/2 KH1 and KH3 domains were shown to be crucial for facilitating PCBP/nsp1\(\beta\) binding to the PRF cassette and stimulating efficient frameshifting, although similar mutations in the KH2 domain had little effect\(^{282}\). However, other residues of the KH2 domain likely play a role in PRF, as the PCBP2 KH2 domain was shown to interact with the C-terminus of nsp1\(\beta\) (reference \(^{296}\)). Interestingly, some \textit{trans}-complementation appears to be possible for PCBP1, as providing both KH1-mutant and KH3-mutant PCBP1 restored PRF efficiency to \(\sim 20\%\) of WT levels (compared to the complete abrogation of PRF when only one mutant was present)\(^{282}\).

The PCBP/nsp1\(\beta\)/RNA complex was confirmed to represent an elongation blockade, as it stimulated ribosomal pausing over a defective slippery sequence \textit{in vitro}\(^{282}\). Like for canonical secondary RNA structure-directed PRF sites, the precise positioning of the blockade relative to the slippery sequence was shown to be important. Alteration of the spacer length revealed that efficient PRF only occurred on PRF cassettes in which the spacer was within the narrow range of 9-12 nt in wheat germ lysates\(^{282}\). This exceeds the canonical spacer length of 5-9 nt, potentially because the proteins occlude a greater region than just the specific bases of the C-rich motif to which they bind, making the “effective spacer length” shorter than the 10 nt value defined relative to the binding motif. Further evidence of the importance of the specific positioning of the blockade is provided by the differences observed between PCBP1 and PCBP2. The two proteins require different spacer lengths for maximum PRF efficiency, and it was observed that PCBP1 preferentially stimulated \(\sim 2\) PRF, while PCBP2 preferentially stimulated \(\sim 1\) PRF\(^{282}\). This suggests that, perhaps due to slight differences in size or conformation, the two proteins have slightly different optimal positions relative to the slippery sequence. Further, it suggests that \(\sim 1\) and \(\sim 2\) PRF are also preferentially stimulated by particular positionings of the elongation blockade, corroborated by the fact that in comparisons using the same PCBP, shorter spacer lengths favoured \(\sim 2\) PRF\(^{282}\). This positioning could potentially facilitate specific interactions with the ribosome to stimulate frameshifting, as both PCBP and nsp1\(\beta\) can interact with ribosomal components. PCBP1 interacts with RACK1, a ribosome-associated protein positioned near the mRNA entry channel\(^{297}\), while nsp1\(\beta\) interacts with rpS14, a ribosomal protein which neighbours rpS3, a protein implicated in the mRNA unwinding activity of the ribosome\(^{21,292}\). Structural characterisation of this elongation blockade and its orientation relative to the ribosome would likely be highly informative regarding the mechanism of this unique PRF site.
1.3 Ribosome profiling

1.3.1 Foundation of the technique

1.3.1.1 History of ribosome profiling

The ribosomal profiling (RiboSeq) technique, first published in 2009 by Ingolia and colleagues\textsuperscript{298}, is based on the knowledge that translating ribosomes protect a footprint of approximately 30 nt of mRNA from nuclease digestion\textsuperscript{299–302}. RiboSeq evolved from the earlier ribosome footprinting method developed by Wolin and Walter in 1988 (reference \textsuperscript{303}), in which ribosome-protected mRNA fragments (RPFs) are hybridised to an anti-sense cDNA of the relevant transcript and their position located by virtue of their ability to block extension of a radiolabelled DNA primer by T7 DNA polymerase. While a powerful and informative technique, ribosome footprinting was capable only of assessing ribosome positions on a single mRNA species at a time. Ingolia et al. further developed this practice of purifying and analysing RPFs and combined it with recently developed deep sequencing technology\textsuperscript{304} to develop ribosome profiling: a high throughput technique that generates a global snapshot of the positions of translating ribosomes in cells, at sub-codon resolution.

1.3.1.2 Fundamental basis of the technique

The following section is based on Ingolia et al. 2009 (reference \textsuperscript{298}), except where otherwise stated. The basic experimental framework has since been used in many publications, and in some cases adaptations were made which are discussed in subsequent sections.

Ribosome profiling is often carried out in parallel with RNA sequencing (RNASeq) of the same cell lysate to gather information on both the translatome and the transcriptome (Figure 1.8). A ribonuclease (RNase), usually RNase I, is added to the aliquot destined for RiboSeq to digest the unprotected regions of RNA. Ribosomes and their enclosed fragments of RNA are purified by high speed centrifugation through a sucrose cushion\textsuperscript{305} and the RPFs purified by phenol:chloroform extraction and ethanol precipitation. A size selection step is performed, in which gel electrophoresis is used to separate fragments by length, and a gel slice corresponding to the range of lengths protected by the ribosome is excised and the RPFs eluted and purified (typically 25-34 nt for cytosolic eukaryotic ribosomes). Adapters, which are necessary for deep sequencing, are ligated to the RPFs and the products are converted to DNA amplicons by reverse transcription. Finally, these are amplified by PCR, although the number of cycles is usually kept to a minimum to avoid compounding PCR-related biases (see section 1.3.3.1.4).
In contrast, in the aliquot used for RNASeq, the cellular RNA is separated from the ribosomes during the first stage of the protocol (without RNase treatment) and subsequently fragmented using alkaline hydrolysis. This generates RNA fragments of similar lengths, but for which the precise sequence is not determined by ribosome position. These RNA fragments are subjected to the same library preparation procedure to generate amplicons for deep sequencing.

Figure 1.8 Key experimental steps in preparation of RiboSeq libraries and parallel RNASeq libraries
Figure 1.9 Hypothetical example of RiboSeq and RNASeq read distributions on an illustrative transcript species

The sequences of the RNA fragments from the RiboSeq and RNASeq protocols are then aligned to the reference transcriptome or genome of interest (Figure 1.9). For RiboSeq, the read density at each position on the transcript is a proxy for the ribosome density at that position, which is an averaged result from all the mRNAs sampled in the deep sequencing. For RNASeq, the read density provides a measure of transcript abundance. The use of these two forms of sequencing in parallel adds power to ribosome profiling, for example by facilitating calculation of translation efficiency.

1.3.1.3 Inference of ribosomal P site position and determination of RPF phase

An approximate location of the P site of the ribosome on RiboSeq reads can be inferred by looking at the distribution of read 5’ ends relative to the position of the start codon for that CDS (Figure 1.10). As the UTRs are usually less translated, very few RPFs map to these regions. A jump in RPF density around the beginning of the CDS corresponds to ribosomes initiating translation at the start codon, and the predominant peak associated with this jump is termed the “initiation peak”. Initiation occurs with the initiator AUG in the ribosomal P site, so for reads of the initiation peak, both the read 5’ position and P site position are known. Therefore, the most common distance between RiboSeq read 5’ ends and the ribosomal P site can be calculated and extrapolated to infer the P site location for other RPFs. This calibration is usually performed using a metagene analysis in which many transcripts are aligned based on their annotated initiation codons and the combined read density from all the transcripts is plotted (often termed a “metagene profile”).

- 29 -
Figure 1.10 Hypothetical example of RiboSeq read 5′ end distribution relative to initiation codons

Annotated transcripts are aligned according to start codon position, and total read density across all these transcripts plotted. This example depicts a mammalian dataset.

Where the distance between RPF read 5′ ends and the ribosomal P site is reasonably homogeneous, RPF 5′ ends map to coding sequences with a triplet periodicity, with an unequal proportion mapping to the three positions within a codon\(^{208,305-308}\). This is known as “phasing” and reflects the fact that the ribosome moves in steps of three nucleotides at once. RPFs whose 5′ ends map to the first position within a codon are attributed to phase 0, with the second and third positions within the codon corresponding to phases +1/−2 and +2/−1, respectively. Various factors can affect the distance between the RPF 5′ end and the P site, and one of the major factors is the extent of nuclease digestion (“trimming”) during the first stage of the experimental protocol. Adjustment of the amount of RNase or the digestion conditions can improve trimming and phasing; however, other contributing factors, such as technical biases (discussed in section 1.3.3.1), are largely outside of the investigator’s control. Additionally, ribosomes from different systems can be of different sizes and protect different footprints, leading to different length distributions and phase compositions\(^{308,309}\).

A metagene analysis of CDS-mapping RPFs to determine their phase composition can help assess the heterogeneity of the distance between the 5′ end and the P site in a particular dataset. For well-trimmed mammalian datasets (such as the example in Figure 1.10), the distance between the RPF 5′
end and the P site is typically 12 nt. This means a large proportion of RPFs are attributed to phase 0. The greater this proportion is, the lower the overall heterogeneity of the RPF 5′ end to P site distance. However, care should be taken when extrapolating this to interpret inferred P site locations on any one specific sequence, because factors such as technical biases that are likely to average out in a metagene analysis may have a considerable effect at the individual codon/nucleotide level. Biological factors could also play a role, for example if some sequences of mRNA are more compact in the mRNA exit channel this could lead to a longer protected region 5′ of the P site. Nonetheless, the overall heterogeneity tends to be relatively minor in well-trimmed mammalian datasets, and when multiple codons are averaged together, the triplet periodicity and phasing can be reliably used to determine reading frame. The ability to determine the locations of ribosomes with sub-codon precision is an extremely powerful benefit of ribosome profiling. This has obvious advantages for studying viruses with complex genome arrangements involving overlapping reading frames and programmed ribosomal frameshift sites.

1.3.2 Further development of the protocol

1.3.2.1 Use of cycloheximide pre-treatment

Cycloheximide (CHX) is a translation elongation inhibitor which binds to the E site of the eukaryotic ribosome and, when the E site is occupied by tRNA, prevents translocation from reaching completion. In the original ribosome profiling protocol, a CHX pre-treatment step was initially included directly before harvesting cells for ribosome profiling. The rationale was to freeze ribosomes in position on the mRNA before exposing cells to the stress of harvesting, which might alter gene expression programmes. However, in Saccharomyces cerevisiae, metagene analyses similar to those in Figure 1.10 demonstrated that CHX pre-treatment exacerbates a 5′ “ramp” of excess density at the beginning of the CDS, decreasing over the first 100-200 codons to reach a plateau maintained for the remainder of the CDS (Figure 1.11). For S. cerevisiae harvested by rapid flash-freezing in the absence of CHX, this ramp was still detectable, although smaller than in CHX pre-treated cells, so it was not considered entirely artefactual. In mammalian cells, however, the ramp was only detectable after CHX pre-treatment. For both yeast and mammalian cells harvested with CHX pre-treatment, stress conditions caused the ramp to become more pronounced. This, along with the observed increase in upstream ORF (uORF) translation, was suggested by some to be a regulatory mechanism that formed part of the stress response. Later experiments revealed that, in S. cerevisiae, the stress-induced increase in both 5′ ramp and uORF translation were an artefact of CHX pre-treatment and did not occur in samples harvested without a drug pre-treatment step (Figure 1.11). Experiments in
mammalian cells yielded similar results (Figure 1.12; Dr Nerea Irigoyen, unpublished data), and inclusion of the CHX pre-treatment step became less common.

**Figure 1.11** The effect of CHX pre-treatment and stress on *S. cerevisiae* RiboSeq read distribution relative to initiation codons

Figure reproduced from Gerashchenko and Gladyshev, 2014 (reference 313). Peroxide was the reagent used to induce oxidative stress, and “drug” refers to CHX pre-treatment. AU = arbitrary units.

**Figure 1.12** The effect of CHX pre-treatment and stress on *Mus musculus* RiboSeq read distribution relative to initiation and termination codons

Figure produced using data from Irigoyen et al., 2016 (reference 307) and Dr Nerea Irigoyen’s unpublished data. All libraries were made by Dr Nerea Irigoyen. The stress condition is represented by 17 Cl-1 cells infected with murine coronavirus strain MHV-A59, harvested at 8 hpi. “CHX” and “No CHX” denote whether the CHX pre-treatment step was included in the harvesting protocol. The peak four codons downstream of the initiation peak and the slight trough four codons upstream of the termination peak are likely the compounded effects of technical biases (discussed in section 1.3.3.1) as reads at these positions all share 5’-AUG or U(A/G)(A/G)-3’ respectively.
In almost all ribosome profiling protocols, irrespective of whether a pre-treatment step is included, CHX is present in the lysis buffer, RNase digestion and ribosome purification solutions to inhibit elongation during sample processing. Despite this, a shallow trough is sometimes observed in place of the ramp in the distribution of reads at the beginning of CDSs (e.g. Figure 1.12). The reason for this is unclear – one possibility is that it represents “run-off” elongation in which translation elongation continues to occur during harvesting, but initiation is inhibited, and this depletes ribosomes from the beginning of the CDS\textsuperscript{298,305,317}.

In addition to the ramp, CHX pre-treatment can affect the size of the initiation and termination peaks. Typically, CHX pre-treated samples have a heightened initiation peak and a smaller termination peak, while samples harvested by flash-freezing without CHX pre-treatment have a large termination peak and a smaller initiation peak\textsuperscript{298,318}.

As well as these effects at the beginnings and ends of CDSs, CHX pre-treatment was found to cause even wider-reaching perturbations that affect the entire length of the CDS. The extent to which sense codons are adapted to the cellular tRNA pool is thought to have a major influence on the speed at which each codon is translated\textsuperscript{319–322}. This degree of adaptation is often quantified using the tRNA adaptation index (tAI), a measure which takes into account the genomic copy number of each tRNA that can decode a codon (as a proxy for intracellular tRNA abundance), and empirically determined weights that penalise wobble base pairs\textsuperscript{323}. Codons with high tAI are expected to be translated rapidly, and therefore codon-specific ribosome occupancy should correlate with 1/tAI; however, the extent to which this is observed varies dramatically between different ribosome profiling datasets\textsuperscript{317}.

The disruption of the correlation between ribosome occupancy and 1/tAI was attributed to CHX pre-treatment\textsuperscript{317}, as CHX pre-treated datasets displayed weak negative correlation\textsuperscript{324,325}, which increased as CHX concentration decreased\textsuperscript{313}, to reach a strong positive correlation in datasets harvested without pre-treatment\textsuperscript{326,327}.

This was proposed to be due to translation continuing during the CHX pre-treatment step, at rates which were still codon-specific but were altered compared to the codon-specific rates observed in the absence of CHX\textsuperscript{317}. This continued translation perturbs the observation of the untreated, \textit{in vivo} steady state translational landscape and leads to peaks moving downstream and spreading out (the “wave hypothesis”). The extent to which this peak movement occurred was found to depend on CHX concentration\textsuperscript{313,317}, indicating it may be related to incomplete inhibition of elongation, and found to increase as the CHX pre-treatment incubation time increased\textsuperscript{317,328}. However, even in some datasets harvested without CHX pre-treatment there was evidence of the “wave” occurring downstream of rare, putatively pause-inducing codons, leading the authors to suggest that continued, CHX-perturbed translation may also occur in lysates where CHX is present. This is tentatively supported by the considerable variation in calculated codon-specific occupancies for several non-pre-treated datasets, in some cases even from the same study\textsuperscript{317,329}. Notably, many of the biases and perturbations
discussed in this section may also be relevant to other inhibitors, for example harringtonine (see section 1.3.2.3) preferentially stalls the ribosome over Lys, Arg, or Tyr codons which may confound interpretation of experiments.\[330\]

\textbf{1.3.2.2 Discovery of a second population of RPFs}

In 2014, Lareau \textit{et al.}\[331\] serendipitously discovered that, when CHX was omitted from all steps of their \textit{S. cerevisiae} ribosome profiling library preparations, a second, shorter population of RPFs became apparent. This is corroborated by an earlier observation made by Wolin and Walter in wheat germ extracts\[303\]. Lareau \textit{et al.} verified that these shorter fragments represent genuinely translating ribosomes, and proposed the two populations of RPFs represented different conformations of the elongating ribosome: 20-22 nt RPFs protected by rotated (“hybrid”) state ribosomes and 28-30 nt RPFs protected by non-rotated (“classic”) state ribosomes. A comprehensive study by Wu \textit{et al.} in 2019 (reference \[140\]) further characterised these two populations of RPFs and came, instead, to the conclusion that short RPFs represent ribosomes with an unoccupied A site, while for the longer population the A site is occupied by a tRNA (Figure 1.13). Both studies were supported by evidence from an experiment in which tRNA-His was depleted, expected to induce ribosomal stalling during accommodation. Lareau \textit{et al.} found 28-30 nt RPFs enriched at His codons (in the A site), while Wu \textit{et al.} found 20-22 nt RPFs enriched – each result supporting the investigators’ opposing conclusions.

In addition, Wu \textit{et al.} carried out DMS probing of the ribosomal conformations leading to each RPF length, which supported their hypothesis that short RPFs represent ribosomes with an empty A site. The difference in conclusions drawn by the two studies, and the reason short RPFs were not noticed in many previous profiling studies, is likely due in part to the previously underestimated extent to which translation continues in yeast lysates\[140\]. In yeast, the population of short RPFs was almost completely absent if CHX was included in the lysate, as it is by default in the vast majority of profiling protocols\[331\]. CHX prevents ribosomes from completing translocation, and traps peptidyl-tRNA in the A site\[115, 312\]. Where translation elongation continues in the lysate, ribosomes that were undergoing accommodation (corresponding to short RPFs in the Wu model) in the cell can seemingly complete this process in the lysate, continuing elongation until they are stalled by CHX at the pre-translocation stage (corresponding to classical 28-30 nt RPFs)\[140\]. In order to prevent this collapse of the two ribosomal populations into one in the lysate, Wu \textit{et al.} added the accommodation inhibitor, tigecycline\[332\], to the lysate in addition to CHX, while Lareau \textit{et al.} chose to omit all elongation inhibitors from the lysate. In the Lareau \textit{et al.} study, it is possible this permitted further elongation in the lysate that may have confounded the tRNA-His depletion experiment.
Further, Wu et al. found that when anisomycin was used to inhibit peptidyl transfer\textsuperscript{333,334}, ribosomal A sites were occupied as expected when assayed in cells; however, when it was assayed subsequently in the lysate, A site occupancy decreased despite the maintenance of anisomycin. This suggests that unreacted A site tRNAs can dissociate from the ribosome during sample processing, further complicating the preservation of the two populations of RPFs in the lysate (Figure 1.13). The three distinct stages of translation elongation at which ribosomes can be captured by the above antibiotics (pre-accommodation, pre-peptidyl transfer and pre-translocation) are represented by only two length distributions of RPFs; therefore, one population of RPFs will always comprise two populations of ribosomes. The investigator’s choice of antibiotics to include in the lysate will determine which is the pure population; for example Wu et al. suggest using both CHX and tigecycline in order to capture a short RPF population composed purely of ribosomes at the pre-accommodation stage. Under these conditions, their analysis of the population of short RPFs revealed the highest correlation between observed codon-specific ribosome occupancy and 1/tAI of any ribosome profiling study up to that point\textsuperscript{140}.

The population of short RPFs was shown to be present in \textit{Caenorhabditis elegans} embryos and human cells, and they were verified as representing ribosomes with an empty A site, indicating this modified profiling protocol is widely applicable\textsuperscript{140}. However, unlike in yeast, treating HeLa cell lysates with CHX alone does not cause the collapse of the short RPF population, nor reduce the extent to which short RPFs accumulate at the expected codons when stimulated (in this case by Gln
starvation). This perhaps indicates a reduced capacity for continued translation in these lysates\(^{140}\). Further work is required to fully understand the relationship between the two populations of RPFs in mammalian cell lysates.

### 1.3.2.3 Adaptations to capture different populations of ribosomes

As well as the adaptations above to capture short RPFs, the ribosome profiling protocol has been modified in multiple different ways to capture other specific populations of ribosomes\(^{335}\).

In a similar vein, the lengths of fragments purified during size selection can be increased to capture ribosome queues. Experiments so far have captured footprints resulting from two or three ribosomes (disomes and trisomes, respectively) closely packed together, revealing hot-spots for ribosome collisions on the transcriptome\(^{186,202,336–338}\).

Pre-treatment of cells with initiation inhibitors (harringtonine or lactimidomycin) instead of elongation inhibitors can be used to identify sites of translation initiation\(^{305,339,340}\). Lactimidomycin is technically an elongation inhibitor but it preferentially binds the initiating ribosome, preventing it from undergoing elongation after initiation\(^{312,334}\). The pre-treatment period allows run-off elongation to occur while initiation is inhibited, leading to a clearly identifiable peak at the initiation codon. Further modification, to completely avoid capturing elongating ribosomes, increased the power to quantify initiation peaks, giving rise to quantitative translation initiation sequencing, or QTI-seq\(^ {339}\).

Treating lysates with a puromycin analogue, which selectively puromycylates actively translating ribosomes, facilitates separation of these from stalled ribosomes in order to observe only active translation\(^ {341}\). This technique is called RiboLace.

Normally, ribosome profiling only captures translating ribosomes, and does not provide information on scanning small ribosomal subunits. Translation complex profile sequencing (TCP-seq) is a modified protocol that uses formaldehyde cross-linking to capture scanning ribosomes, which are separated from translating ribosomes by sucrose gradient sedimentation\(^ {342,343}\).

Finally, tagged ribosomes can be specifically pulled down to isolate populations of interest. This can be used to profile ribosomes from specific tissues\(^ {339,344,345}\) or organelles\(^ {346–349}\), or those which come into proximity with specific proteins (modified to form a biotin ligase fusion protein)\(^ {350}\).

### 1.3.3 Sources of bias

As with all deep sequencing-based techniques, ribosome profiling can be affected by various factors at both the experimental and computational stage that can bias the results.
1.3.3.1 Technical biases

Technical biases can result from each stage of the library preparation procedure. Many of these technical biases are shared between RiboSeq and RNASeq, rendering parallel RNASeq libraries a useful control condition, but some (e.g. nuclease and fragmentation bias) are specific. The most commonly acknowledged biases result from sequence preferences in the enzymes used in the library preparation protocol.

1.3.3.1.1 Nuclease bias (RiboSeq)

Several nucleases can be used to digest the unprotected RNA in the first stage of the RiboSeq protocol: RNase I, micrococcal S7 nuclease (MNase), RNase A, RNase T1 and, recently, RelE\cite{351-353}. RNases A and T1 are generally avoided, due to severe sequence-based cutting preferences (cutting almost exclusively after C/U and G, respectively)\cite{351,354-357}. Both RNase I and MNase cut after all four bases, however RNase I is usually favoured due to an almost complete absence of sequence-specific cutting preference\cite{351,354}. MNase is generally used for bacterial profiling, as RNase I has been reported to be inactive in E. coli\cite{351,358}; however, MNase shows complex cutting preferences, that depend on the length of the RPF, but with a general preference for cleaving after A or U\cite{351-353}. Some nuclease biases may be more complex than a simple preference for particular nucleotides at fragment ends, for example RNase I has been reported to slightly bias datasets towards short genes, although the reason for this is unclear\cite{359}. Notably, nuclease selection can also impact RPF length, with implications for phasing and P site location determination (see section 1.3.1.3). In this regard RNase I performs well, producing the narrowest RPF length distribution and the strongest triplet periodicity of the first four nucleases mentioned above\cite{351}.

1.3.3.1.2 Fragmentation bias (RNASeq)

To generate RNA fragments in a comparable size range to RiboSeq, RNASeq samples are usually fragmented instead of digested with a nuclease. Although fragmentation bias is less well characterised than nuclease bias, and generally considered to be fairly minimal\cite{360,361}, it still has some effect. Enzymatic fragmentation methods were found to induce greater bias than heat fragmentation\cite{362}.

1.3.3.1.3 Ligation bias

The method used by Ingolia et al.\cite{298} involves circularisation of the amplicons, requiring just a single ligation event, catalysed by Rnl2. Later, the protocol was adapted to use a small-RNA library preparation method in which two successive adapter ligations, catalysed by Rnl1 and Rnl2, were used instead of amplicon circularisation\cite{306,308}. Both Rnl1 and Rnl2 have complex sequence preferences\cite{363}. Rnl1 prefers single-stranded substrates\cite{364,365}, while Rnl2 shows a preference for single-strandedness downstream of the ligation site but double-strandedness upstream\cite{365}. The
addition of randomised bases to the 5′ end of the 3′ adapter and the 3′ end of the 5′ adapter is thought to reduce ligation bias by generating a diverse population of conformations at the ends of the substrate for ligation. Addition of randomised bases to the 5′ end of the 3′ adapter and the 3′ end of the 5′ adapter is thought to reduce ligation bias by generating a diverse population of conformations at the ends of the substrate for ligation.

1.3.3.1.4 PCR bias

Not all fragments are amplified by PCR with equal efficiency, which can lead to bias in the amplified population. This problem is exacerbated in protocols where low input material means many PCR cycles are required. Fragments with a GC content of approximately 50% are preferentially amplified, which can lead to the relative depletion of GC- or AT-rich fragments. Additionally, some polymerases preferentially amplify shorter fragments. The inclusion of randomised bases (sometimes termed “unique molecular identifiers” or UMIs) in the ligated adapters (section 1.3.3.1.4) has the added benefit of providing a means of distinguishing between biological duplicates and duplicates resulting from PCR amplification. The latter can therefore be removed, which reduces the effect of PCR bias. However, while this removes over-represented sequences, those which are refractory to PCR may still be under-represented.

1.3.3.1.5 Other sources of technical bias

The four biases listed above are generally considered to be the main sources of technical bias in ribosome profiling/RNASeq data, along with the confounding effect of cycloheximide pre-treatment, if used (discussed in section 1.3.2.1). There are, however, several additional potential sources of bias within the work-flow.

At the RNA extraction stage, a miRNA study showed that highly structured or GC-poor miRNAs (notably a similar length to RPFs) were disproportionately poorly precipitated compared to other miRNAs when TRIzol precipitation was used and the total input RNA concentration was low.

Ribosomal RNA (rRNA) is extremely abundant, sometimes comprising over 90% of RNASeq reads. For RiboSeq, ribosomes are selectively purified which enriches the proportion of rRNA even further. Investigators are not usually interested in analysing rRNA and therefore this reduces the number of usable reads in their libraries. For both types of sequencing, methods are usually employed to deplete the rRNA during library preparation.

For RNASeq, poly(A) selection is a popular choice, in which oligo(dT) beads are used to bind and purify RNA species with poly(A) tails. However, this introduces several biases. Generally, poly(A) selection generates a bias towards 19-130% higher coverage at the 3′ end of transcripts, as cleavage and degradation can occur during sample handling and the 5′ portion is no longer captured. Cleavage is more likely for longer genes and therefore this may create a bias in which longer genes show reduced coverage. Also, several RNA species have short and/or highly heterogeneous poly(A) tail lengths, or no tail at all (e.g. histone mRNAs). This will severely reduce...
their selection efficiency – a particularly relevant consideration for the study of pathogens with diverse translation strategies that may not use poly(A) tails (e.g. many species of RNA viruses and negative-sense replication intermediates even from viruses which do use poly(A) tails).

Poly(A) selection is not applicable to RiboSeq samples, so alternative rRNA depletion strategies are used and, in many cases (such as this work), applied instead of poly(A) selection to the parallel RNASeq libraries. Subtractive hybridisation is common, in which specific complementary oligonucleotides hybridise to the rRNA and are used to remove it. There are several commercial kits available for popular organisms based on this method. Assessment of three kits for yeast RiboSeq libraries revealed that Ribo-Zero Gold performed best, introducing minimal bias into the mRNA population (Pearson $R^2$ of 0.98 compared to untreated library). However, off-target depletion of non-rRNA reads that hybridise with the probe is still a potential concern. For investigators studying non-model organisms, the sequence-specific nature of subtractive hybridisation can preclude the use of commercial kits. In these cases, custom oligonucleotides can be generated based on the known rRNA sequence. For RiboSeq, some fragments of rRNA are extremely abundant, presumably due to being available for RNase cleavage during the nuclease digestion step and being in the size range selected during fragment purification. Specific oligonucleotides can be used to target these and alleviate fears of these unusually abundant fragments saturating the hybridisation capacity of kits (such as Ribo-Zero) designed for depleting RNASeq data, which is expected to have fragments which uniformly cover the rRNA. Alternatively, duplex-specific nuclease (DSN) can be used to degrade rRNA after its conversion to cDNA, taking advantage of the more rapid hybridisation of highly abundant sequences (rRNA) relative to less abundant sequences (mRNA). Duplex-specific nuclease selectively degrades double-stranded DNA, without requiring prior knowledge of the rRNA sequence. Due to its great abundance, this method predominantly degrades rRNA, although it can slightly deplete highly abundant mRNAs, and introduces a slight bias in GC content. It is also expected to be incompatible with the use of randomised bases in adapters.

The process of deep sequencing itself can introduce biases and artefacts, although in some cases these are platform-specific. Low complexity libraries in which a large proportion of reads have the same base at the same position can induce errors in base calling by Illumina, a potential problem for RiboSeq libraries which have not had rRNA depleted. This can be alleviated by increasing library complexity, for example by spiking in a small amount of a high complexity library (typically the “PhiX” spike-in provided by Illumina). Several different samples are usually sequenced in one (multiplexed) run and identified through a unique index sequence, introduced using the adapters or primers during library preparation. Errors in sequencing and/or recognising the index region can lead to “index hopping” or “sample bleeding”, in which reads from one sample are incorrectly attributed to another. While the percentage of reads to which this applies tends to be small (< 1%) it can cause significant problems if it mixes reads between drastically different libraries, such
as infected cells from late and early timepoints. This can be avoided by dual indexing, selecting combinations of indices which are less prone to mis-identification and sequencing different types of libraries separately.

Finally, bias can be introduced at the computational stage. The quality of the genome annotation to which reads are mapped can influence analyses such as differential gene expression analysis\(^\text{380,381}\). Additionally, different strategies for dealing with multi-mapping reads, a particularly pertinent consideration for RiboSeq due to the short read length, can yield considerably different read coverages\(^\text{382}\). Some authors choose to discard reads which can align to multiple different locations and allow only uniquely mapping reads\(^\text{306}\). This stringency ensures the read is confidently aligned; however, it can mean discarding up to 30% of reads, potentially leading to low coverage, particularly of regions which are repetitive or homologous to other regions of the genome\(^\text{382}\). More commonly, reads are allowed to have multiple possible alignments and only one is chosen for reporting. In some cases the alignment reported is chosen at random\(^\text{383,384}\), while in others the alignment with the fewest mismatches is reported (choosing randomly between equally good matches)\(^\text{307,385,386}\). Both approaches can introduce bias, for example if a highly expressed gene shares a region of homology with a poorly expressed gene, the latter will likely have its read density artificially inflated.

At almost all stages of the RiboSeq work-flow, characteristics of the RNA may affect the efficiency with which particular fragments are carried over to the next step. Many of these are specifically addressed above but it is likely there are more that are not yet completely understood. For example particular dinucleotides were found to bias expression levels estimated by RNASeq, without a clear mechanistic explanation\(^\text{387}\).

1.3.3.1.6 Effects of technical bias

Although most protocols aim, as far as possible, to avoid introducing bias, it is generally thought that this is not critically important for applications such as estimation of gene expression\(^\text{382,388}\). The effects of technical biases are likely to average out over reasonably large regions such as CDSs. Further, for applications such as differential gene expression, the same technical biases will usually be experienced by the test and control samples under comparison. As such, datasets with modifications to ameliorate bias tend to agree well with control datasets. For example, a comparison of gene expression estimates from libraries prepared using RNase I, micrococcal S7 nuclease, RNase A or RNase T1 showed very good agreement, with Pearson correlation coefficients of above 0.95 in all comparisons\(^\text{351}\).

For applications, such as codon-specific ribosome occupancy calculations or analysis of ribosomal pausing, which require interpretation of ribosome positions with high spatial resolution, technical biases have a much greater impact on the results and should be carefully controlled\(^\text{317}\).
1.3.3.2 Biological factors

In addition to the technical biases discussed above, biological factors can play a role in generating bias. This is not necessarily undesirable, as these phenomena are representations of genuine biological occurrences, however they can confound analyses if not correctly recognised and accounted for.

As discussed previously (sections 1.3.2.2 and 1.3.2.3), ribosomes in different states or conformations can generate footprints of different sizes. Often, these are excluded from the library due to the inclusion of a size-selection step, and this may introduce bias. For example, if mammalian lysates harvested without CHX pre-treatment contain short RPFs but these are excluded during RNA purification, this may mean translation regulation at the decoding/accommodation stage is underestimated. Biological events which change the length of protected RNA may also be overlooked if this shifts RPFs outside the selected range, for example translation termination on truncated (stop codonless) mRNAs produces ~16 nt RPFs, and ribosomal interaction with Shine-Dalgarno sequences increases RPF length.

Ribonucleoprotein complexes (RNPs) can reach large enough molecular weights that they are inadvertently co-purified with ribosomes. If the footprints of RNA protected by these RNPs is within the length range of RPFs, RNP-derived reads can be included in the library and mis-assigned as RPFs. RNPs are not expected to display phasing, nor the sharp peak(s) in length distribution characteristic of RPFs. They can, therefore, be identified at the scale of a metagene analysis, but are difficult to differentiate from RPFs at the scale of individual reads. RNPs usually make up only a small proportion of RiboSeq libraries, but viruses often encode very highly expressed RNA-binding proteins that mean virus-infected samples can be severely affected by RNPs.

Many strategies for estimating translation levels assume a constant rate of translation across the region being measured, and there are multiple biological factors that can affect translation rates and therefore may cause bias in these measurements. In some cases, it is thought that organisms have evolved to take advantage of these for specific purposes, for example slowing down translation in inter-domain regions to allow time for protein domain folding.

As described above (section 1.3.2.1), a major contributor to determining translation speed is the level to which a codon or CDS is adapted to the cellular tRNA pool. This is often quantified using tAI. This metric, by definition, uses wobble pairing penalties calculated from empirical S. cerevisiae gene expression data, selecting penalty values which optimise the correlation between gene expression levels and tAI. However, the penalties for wobble pairing likely vary between species, so a species-specific tAI (stAI) was developed which uses wobble pairing penalties calculated for each organism independently.
Introduction

As well as the variation caused by codon-specific translation rates, specific motifs or features of the CDS can induce ribosomal pausing or stalling. Positively charged amino acids within the nascent peptide can interact with the negatively charged exit tunnel to slow down the ribosome at poly-basic motifs\textsuperscript{307,325,327,336}. Proline residues have restricted geometry that is known to be able to induce translational pausing at the peptidyl-transfer stage or within the exit tunnel\textsuperscript{305,374,395}. Particular pairs of codons (dicodons) can induce ribosome pausing, for example CGA–CGA and CGA–CCG in S. \textit{cerevisiae} do so by inducing unusual conformations of the ribosome active site\textsuperscript{396,397}. As mentioned above (section 1.2.1), stable secondary RNA structures can also pause the ribosome\textsuperscript{20,22}. Many other factors or motifs have been suggested to induce pausing\textsuperscript{316,398,399}, for example co-translational binding of proteins such as chaperones\textsuperscript{315}. 
1.4 Porcine reproductive and respiratory syndrome virus

1.4.1 Disease

Porcine reproductive and respiratory syndrome virus (PRRSV) is an enveloped, positive-sense, single-stranded RNA virus\(^{400,401}\), and the etiological agent of the disease from which it takes its name: porcine reproductive and respiratory syndrome (PRRS). Pigs are the only known natural host of PRRSV, and PRRS is the most economically devastating disease of swine, costing an estimated $664 million a year in the US alone\(^{402,403}\).

As the name suggests, PRRS is characterised by reproductive failure in adult sows and respiratory failure in young pigs. This reproductive failure leads to an increase in the number of stillborn, mummified or aborted foetuses and spawned the term “abortion storms”, referring to the spread of these symptoms throughout swine herds\(^{404}\). Cyanosis of the ear and vulva, due to decreased blood flow, can lead to blue or red discolouration, termed “blue ear”\(^{405,406}\). Respiratory symptoms are described as a mild to moderate pneumonia, and PRRS is often associated with lesions in the lung\(^{407}\). In piglets, it is also often accompanied by an increase in secondary infections\(^{402}\). The severity of PRRS symptoms varies from sub-clinical to lethal, thought to be dependent on the genetics of both the virus and the host\(^{404,407-409}\).

The great economic cost associated with PRRS is due, in part, to its ability to spread rapidly through swine populations. It is spread through direct contact between infected pigs, or through contact with contaminated objects, and transmission occurs by the respiratory and oral routes, percutaneously, or through mucosae\(^{410}\). It can be transmitted vertically between mother and foetus through the placenta, and can be spread through infected semen, in which it can persist for extended periods of time\(^{407,410-412}\). Artificial insemination is a very common husbandry practice, and this can facilitate transmission between distant farms if left undetected\(^{407,411,413}\). Airborne transmission of PRRSV occurs over relatively short distances (1-120 m), but for some strains transmission distances can increase to up to 9 km\(^{414-416}\). Most pigs clear PRRSV infection within 3-4 months; however, in some it can persist for several months, sometimes remaining in their tonsils long after it is cleared from other tissues, facilitating further spread within the herd\(^{411,417}\).

1.4.2 Emergence and phylogeny

The first outbreak of PRRS was reported in North America in 1987 and, just three years later, an outbreak was reported in Europe\(^{404}\). Despite the similar clinical symptoms, isolation and characterisation of the viruses responsible revealed drastically different serology and a pairwise nucleotide similarity of just 60%, indicating that these outbreaks represented two independent,
almost simultaneous emergence events\textsuperscript{404,418–421}. These two lineages were named the “North American” (Type 2) and “European” (Type 1) lineages, exemplified by the prototype viruses VR-2332 and Lelystad, respectively\textsuperscript{418,419}. A retrospective analysis of pig sera collected in Canada between 1978 and 1982 revealed the earliest sample seropositive for PRRSV antibodies was in 1979 (reference \textsuperscript{422}). Similar studies in the US, Europe and Asia also revealed the presence of seropositive samples before outbreaks were detected, with the rate of seropositivity rapidly increasing shortly before outbreak detection\textsuperscript{408,423}. Molecular clock analyses predicted that the two lineages diverged separately from a very distant common ancestor and estimates of the time of divergence varied from ten to > 150 years prior to the clinical recognition of PRRS\textsuperscript{403,424–426}. Collectively, this evidence suggests that non-pathogenic ancestors of both PRRSV lineages were circulating in swine (including wild boar) for several years before becoming pathogenic and emerging in the late 1980s\textsuperscript{423}.

Since their emergence, the two lineages have continued to evolve rapidly and spread geographically and have since been categorised into several subgroups\textsuperscript{412} (Figure 1.14 and Figure 1.15). The rapid rate of PRRSV evolution can largely be attributed to the lack of proofreading ability of the viral RNA-dependent RNA polymerase (RdRp) and frequent recombination events\textsuperscript{403}. New strains constantly emerge, often divergent enough to evade existing immunity, causing disease outbreaks even in vaccinated herds\textsuperscript{403,412,423}. For example, MN184, a highly virulent strain of North American lineage with > 14.5% nucleotide difference compared to VR-2332, emerged in 2001 (reference \textsuperscript{427}). In 2006, a highly pathogenic strain (now known as HP-PRRSV) of the North American lineage emerged in China\textsuperscript{428}. This is of particular concern as it quickly spread through most of Asia and has a very high mortality rate of 20-100\%\textsuperscript{428}. NADC30 and NADC30-like strains from the North American lineage were identified in the US in 2008 and soon spread across China, with mortality rates of 30-50\%\textsuperscript{429–431}. The emergence of the Lena strain in 2007 revealed that the European lineage was also capable of giving rise to highly pathogenic new strains\textsuperscript{432}. This strain had enhanced pathogenicity and 87% nucleotide similarity compared to the Lelystad prototype\textsuperscript{432}. Although the North American lineage is predominantly localised in North America and Asia, and the European lineage in Europe, cases do occur in the reverse locations, thought to be due partly to unforeseen natural transmission of live vaccine strains\textsuperscript{412,433}.
Figure 1.14 Phylogenetic tree of the European (type 1) lineage of PRRSV

Figure reproduced from Shi et al., 2010 (reference 434), who constructed a mid-point rooted tree using MrBayes, based on ORF5 sequences. To the right of each clade designation, the geographical distribution is shown, and the number of sequences is given in parentheses (some isolates were omitted from the tree for clarity).
Introduction
Figure 1.15 Phylogenetic tree of the North American (type 2) lineage of PRRSV

Figure reproduced from Shi et al., 2010 (reference 435), who constructed a mid-point rooted tree using MrBayes, based on ORF5 sequences. Note that both the European and North American lineages have diversified further since 2010, and these trees are intended to provide an overview of the diversity of the species, not an exhaustive current snapshot of the phylogeny.

PRRSV has been reclassified and renamed several times since its emergence. Originally named swine infertility and respiratory syndrome virus, it was placed in the Arterivirus genus in 1995 (reference 436). This genus was subsumed into the new order Nidovirales one year later437 – an order which PRRSV shares with coronaviruses (CoVs). In 2005 the virus was renamed to PRRSV, and in 2015 the European and North American lineages were split into separate species, named Porcine reproductive and respiratory syndrome virus 1 (PRRSV-1) and Porcine reproductive and respiratory syndrome virus 2 (PRRSV-2), respectively438. In 2018 these species were renamed to Betaarterivirus suid 1 and Betaarterivirus suid 2. They are currently allocated to the subgenera Eurpobartevirus and Ampobartevirus, respectively, and they are each the only species within their subgenus. Both subgenera are within the Betaarterivirus genus, Variarterivirinae subfamily, Arteriviridae family. The Arteriviridae family is diverse, particularly since the discovery of several new viruses related to simian haemorrhagic fever virus (SHFV) in non-human primates439 (Figure 1.16). Throughout this thesis, PRRSV-1 is referred to as EU (European) and PRRSV-2 as NA (North American) PRRSV.
Figure 1.16 Phylogenetic tree of the Arteriviridae family

Figure reproduced from Kuhn et al., 2016 (reference 439), who built this phylogenetic tree based on ORF1b amino acid sequences. Animals depicted represent the natural hosts of each virus, with question marks indicating uncertainty. KRTGV-1/2 = Kibale red-tailed guenon viruses 1 and 2; DeBMV-1 = DeBrazza’s monkey virus 1; PBJV = Pebjah virus; SHFV, simian haemorrhagic fever virus; KKCBV-1 = Kafue kinda-chacma baboon virus 1; MYBV-1 = Mikumi yellow baboon virus 1; SWBV-1 = Southwest baboon virus 1; KRCV-1/2 = Kibale red colobus viruses 1 and 2; SHEV = simian haemorrhagic encephalitis virus; DMVV-1 = Drakensberg Mountain vervet virus 1; LaDV-1/2 = lactate dehydrogenase-elevating viruses 1 and 2; APRAV-1 = African pouched rat virus 1; EAV = equine arteritis virus; WPDV = wobbly possum disease virus.

1.4.3 Tropism and culture systems

The tropism of PRRSV is highly selective. Pigs and wild boars are the only known natural hosts, and within the host it is restricted to cells of the monocyte-macrophage lineage440. Porcine alveolar
Porcine reproductive and respiratory syndrome virus

macrophages (PAMs) are the primary target, but PRRSV can also infect pulmonary intravascular macrophages and macrophages from other tissues: the spleen, liver, lymph nodes, thymus, tonsils, Peyer’s patch, blood and bone marrow. Dendritic cells and monocytes can also be infected, though to a lesser extent. The tissue tropism has been reported to vary for some strains, with some suggestion that this may result from use of different entry receptors.

Expression of cellular entry receptors governs PRRSV tropism. Many different cellular receptors are involved in PRRSV cell entry, which occurs through receptor-mediated endocytosis. Heparin sulfate is responsible for the initial attachment of PRRSV to the cell, although is not absolutely required for infectivity. Sialoadhesin (also known as CD169 or SIGLEC1) plays a role in both attachment and endocytosis of PRRSV, but not virus uncoating. However, PRRSV infection of sialoadhesin knockout pigs was not significantly reduced compared to WT pigs, indicating sialoadhesin is not essential for PRRSV cell entry in vivo. Vimentin binds the virion during attachment and may play a further role in transportation and replication of PRRSV inside the cell.

CD163 is a protein in the scavenger receptor cysteine-rich family. It is expressed on cells of the monocyte-macrophage lineage and is essential for virus entry, acting at the stage of viral uncoating. MYH9 binds CD163 and is essential for PRRSV internalisation in cell culture. DC-SIGN (also known as CD209) has been reported to enhance PRRSV infectivity in vitro. CD151 is also important for permissiveness in vitro, however the mechanism of action is uncharacterised.

Of all these proteins, the only receptor that has been shown to be indispensable for PRRSV replication in vivo is CD163, leading to it being considered the major entry receptor for PRRSV.

Partly due to the still-evolving literature on PRRSV cell entry receptors/mechanisms, only two cell types (PAMs and MA-104-derived cell lines) were known to be fully permissive to PRRSV until recently. Primary PAMs, which are the main target for infection in vivo, are permissive to infection in vitro after isolation. Surprisingly, the MA-104 cell line was found to be susceptible to PRRSV infection, despite not being of the monocyte-macrophage lineage but derived from immortalised Chlorocebus sabaeus (African green monkey) kidney. Further propagation of this, presumably heterogeneous, epithelial cell line to select for cells highly permissive to PRRSV infection led to the development of the MARC-145 cell line, which is very widely used to culture, propagate, and investigate PRRSV. Other cell lines derived from C. sabaeus kidney, such as CRL11171 and CRL2621a, were also found permissive to PRRSV infection, although these are less commonly used. Later characterisation revealed that MARC-145 cells express CD163, heparin sulfate, CD151 and vimentin, but not sialoadhesin. PAMs do express sialoadhesin (as well as all the receptors expressed by MARC-145 cells), indicating there may be some differences in the pathways of attachment and uncoating between the two systems.
Although PAMs are more physiologically relevant for culturing PRRSV, they have several disadvantages compared with MARC-145 cells. They cannot be easily maintained or stored for long periods and must be isolated from pigs by pulmonary lavage\textsuperscript{477}. This is time-consuming and expensive and yields a heterogeneous population of cells that may be subject to undetected contamination\textsuperscript{475}. PAMs are very poorly permissive to PRRSV infection when first isolated, and infection efficiency increases as the duration of culture increases (up to 96 hours). This indicates that PAMs undergo physiological changes (for example differentiation) in culture that mean they may no longer faithfully reflect the physiology of in vivo PAMs\textsuperscript{478}. This, and other factors such as genetic and physiological heterogeneity between pigs from which cells are isolated, contributes to the variability in infection efficiency of PRRSV in PAMs, reported to range from 15 to 60\%\textsuperscript{474,478,479}. MARC-145, on the other hand, is an immortalised cell line, clonally expanded from a single cell, which can be maintained relatively cheaply and is reliably highly permissive to the PRRSV strains commonly used in academic research\textsuperscript{475}. Notably, permissiveness of both MARC-145 and PAMs varies depending on the PRRSV isolate, and both are unable to support culture of certain field isolates\textsuperscript{480}. PAMs are the more frequently used cell line for isolating PRRSV from clinical samples\textsuperscript{480}.

In recent years, as the receptors important for PRRSV cell entry became better characterised, a greater range of cell lines permissive for infection were generated. Genetic modification of non-permissive cell lines to induce them to stably express CD163 renders various cell lines permissive, for example porcine kidney (PK 032495), feline kidney (NLFK), hamster ovary (CHO) or baby hamster kidney (BHK-21) cell lines\textsuperscript{472}. This has also been used to render previously non-permissive immortalised macrophage cell lines (generated from porcine kidney macrophages or PAMs) susceptible to PRRSV infection\textsuperscript{480,481}. ZMAC is an immortalised PAM cell line which is highly (\~80\%) permissive to PRRSV infection without the requirement for genetic modification to induce CD163 expression\textsuperscript{479,482}.

Reverse genetics systems for PRRSV have been available since 1998 and have greatly aided the study of virus replication, protein function, and vaccine design\textsuperscript{483,484}. The PRRSV genome is infectious and can be generated from a cDNA encoding the genome (termed virus rescue) in two ways: transfection of the cDNA itself into cells, or in vitro transcription of the cDNA and capping to generate viral RNA for chemical transfection or electroporation\textsuperscript{484}. For direct transfection of cDNA, the genome is put under the control of a eukaryotic polymerase II promoter (often the CMV promoter), to be transcribed into vRNA in the transfected cell\textsuperscript{485}. Sequences encoding ribozymes are introduced at either end of the viral genome cDNA in the transfected plasmid to catalyse post-transcriptional RNA cleavage to generate the authentic termini of the vRNA\textsuperscript{486,487}. In the in vitro transcription method, the genome is usually placed under the control of a bacteriophage T7 promoter for efficient cell-free transcription. A 21-109 nt poly(A) stretch is encoded in the cDNA directly.
downstream of the genome, followed by a restriction site used to linearise the plasmid prior to transcription and prevent transcription of non-viral nucleotides after the poly(A) tail \(^{483,485}\). MARC-145 cells are often used to rescue the virus, but BHK-21 cells, which can be transfected at higher efficiency, can be used to produce the first round of infectious progeny virus, although they are not permissive to infection themselves \(^{483,484,488-491}\).

### 1.4.4 Disease control

Several factors contribute to the current lack of success in controlling PRRS, in particular the highly divergent phylogeny, rapid viral evolution rate, and the multiple mechanisms employed to suppress the innate immune response.

Several vaccines are available, although these have had limited success. Most vaccines utilise modified live virus (MLV), for example Porcilis® PRRS (Merck) against EU PRRSV and Ingelvac PRRS MLV (Boehringer Ingelheim) against NA PRRSV. Although these elicit cellular and humoral immune responses that confer protection against parental strains, their effectiveness is drastically reduced against heterologous strains \(^{423,492}\). In addition, there are serious safety concerns. Spread of PRRSV vaccine strains between pigs has been frequently demonstrated \(^{433,493-496}\) and is thought to be responsible for the introduction of NA PRRSV into Denmark \(^{497}\). Combined with reports of recombination between MLV strains and virulent field isolates, this is a major concern \(^{498}\). Inactivated and subunit vaccines are available, which contribute to clearance in pigs that have been previously infected but are not very effective as a sole defence \(^{499,500}\).

A hallmark of PRRSV infection is a weak immune response. The innate immune response is severely suppressed by several viral proteins (Table 1), leading to a weak adaptive response. Non-neutralising antibodies against PRRSV proteins appear after 7-9 days, but it takes one month for neutralising antibodies to appear \(^{501}\). Glycan shielding is thought to contribute to the weak or delayed induction of neutralising antibodies \(^{488,502-504}\). Interleukin (IL)-10 is up-regulated during PRRSV infection, leading to an increase in Tregs which is thought to contribute to immune suppression and viral persistence \(^{505}\). Vaccine strains designed to combat some of these features, for example by using strains which induce the interferon response, may be a promising strategy \(^{506-509}\).

Alternative disease control strategies exist but are not currently widely used. Instead of vaccination, a potential alternative disease control strategy was presented by the finding that pigs with the major PRRSV entry receptor, CD163, knocked out were immune to infection \(^{470}\). This was further refined by a study in which only the SRCR5 domain, which is essential for the interaction with the PRRSV virion, was knocked out, potentially leading to fewer off-target effects \(^{471,510}\). Several anti-viral drugs, small RNAs, and nanobodies are known to inhibit PRRSV infection, although these are not currently deployed on a large scale \(^{505}\).
1.4.5 Replication cycle

The replication cycle of PRRSV generally takes approximately 12 hours\(^5\) (Figure 1.17). The virus enters cells through clathrin-mediated endocytosis\(^6\), with the processes of viral attachment, internalisation, endosome acidification and uncoating involving interaction with several host factors, as discussed in section 1.4.3. After uncoating, the viral genome is directly translated by host ribosomes to produce the polyproteins (pp) pp1a and pp1ab, which are cleaved by viral proteases into 16 non-structural proteins (section 1.4.6.1). Together, these proteins modify host ER membranes to generate double-membrane vesicles (DMVs) in the perinuclear region and assemble at this site to form replication and transcription complexes (RTCs). The RTC catalyses viral genome replication and subgenomic (sg)RNA production, both of which proceed via a negative-sense intermediate (section 1.4.6.2.1). The sgRNAs are translated into structural proteins (section 1.4.6.2.2). The N protein encapsidates the viral genome, forming a nucleocapsid which buds into the smooth endoplasmic reticulum (ER) or Golgi, the membrane of which contains the remaining structural proteins\(^7\). The envelope proteins are extensively post-translationally modified in the Golgi as the virion progresses through the exocytic pathway, after which it is released at the plasma membrane\(^8\). Recently, an additional mechanism of intercellular spread was characterised, in which PRRSV RNA and proteins spread through nanotubes between cells without the requirement for entry receptor involvement\(^9\).
1.4.6 Genome structure and protein functions

The PRRSV genome is large, varying from 14.9 to 15.5 kb (Figure 1.18), and is a positive-sense, single-stranded RNA that is 5’-capped and 3’-polyadenylated, and directly translated\(^ {403, 515, 517, 518} \). As well as the full length genome, a nested set of subgenomic transcripts is produced during infection, through a hallmark transcription strategy that gave the Nidovirales order its name (from the Latin \textit{nido}, meaning nested)\(^ {403, 515, 517} \). The non-structural proteins (nsp) are expressed from the genomic RNA (gRNA) and the structural proteins from the sgRNAs.
Introduction

Figure 1.18 PRRSV genome and subgenomic transcripts

Based on the NA PRRSV isolate used in this study (SD95-21). ORFs are coloured and offset on the y axis according to their frame relative to ORF1a (0: purple, no offset; +1/−2: blue, above axis; +2/−1: yellow, below axis). Subgenomic (sg) RNAs are shown beneath the full length genomic RNA (gRNA), with the region of 5′ UTR that is identical to the genomic 5′ UTR shown in grey. ORFs translated from each sgRNA are depicted as coloured boxes and named to the right.

1.4.6.1 Non-structural proteins

1.4.6.1.1 Expression strategy

The non-structural proteins are encoded by two long ORFs: ORF1a and ORF1b. Expression of these ORFs generates large polyproteins, which are cleaved by four viral proteases to generate the proteins depicted in Figure 1.18 (references 403,515,517,519). ORF1a encodes pp1a, which is cleaved to generate nsps 1α-8. Near the end of ORF1a, there is a canonical, pseudoknot-stimulated −1 PRF site (Figure 1.19). Ribosomes which do not frameshift translate the one remaining codon of ORF1a before encountering a stop codon and terminating translation. Only ribosomes which frameshift can access ORF1b, which is encoded in the −1 frame relative to ORF1a. PRF and continued translation in the −1 frame generates pp1ab, which is cleaved to generate nsps 1α-7β and nsps 9-12 (references 403,515,517,519). The first 44 amino acids of nsp9 are encoded in the 0 frame (purple in Figure 1.18) and are identical to the first 44 residues of nsp8. The remainder of nsp9 is encoded in the −1 frame (orange in Figure 1.18), and nsp9 is effectively an extended version of nsp8, although it lacks the cysteine residue found at the nsp8 C-terminus (the final 0-frame codon of ORF1a).

There is a second PRF site in the genome: the non-canonical, protein-stimulated PRF site described in section 1.2.2.1. This is located in the region of ORF1a that encodes nsp2, and it facilitates both −1 and −2 PRF. Ribosomes which frameshift into the −1 frame encounter a stop codon immediately and make a truncated polyprotein, pp1a-nsp2N, which is cleaved into nsp1α, nsp1β and nsp2N. Ribosomes which undergo −2 PRF translate 169 codons before termination, generating
Porcine reproductive and respiratory syndrome virus

pp1a-nsp2TF, which is cleaved into nsp1α, nsp1β and nsp2TF. This PRF site thus facilitates production of three variants of nsp2 (nsp2, nsp2TF and nsp2N; Figure 1.20).

![Figure 1.19 PRRSV ORF1ab −1 PRF site](image1)

**Figure 1.19 PRRSV ORF1ab −1 PRF site**

Figure reproduced from Li *et al.*, 2014 (reference 281) (sequence from the EU PRRSV isolate SD01-08). The stimulatory RNA in this case is a pseudoknot consisting of two stems (red and blue) joined by two single-stranded regions (pink and green).

![Figure 1.20 Nsp2 and related proteins produced by frameshifting](image2)

**Figure 1.20 Nsp2 and related proteins produced by frameshifting**

Protein names are indicated on the left, and the reading frame after the PRF site (red arrow) is indicated on the right-hand side. All three products share papain-like proteinase (PLP) domains, and a hyper-variable region (HVR). While nsp2N is truncated directly after the PRF site, nsp2 and nsp2TF each have a different transmembrane (TM) domain, containing multiple predicted membrane-spanning helices, after the PRF site. Nsp2 also has a cysteine-rich tail domain (C) of unknown function which is not present in either transframe product.

In addition to the functional ramifications of generating variants of nsp2, which are discussed in section 1.4.6.1.2.1, PRRSV is thought to utilise PRF to control the stoichiometry of the proteins produced from the gRNA. This is a strategy employed by many RNA viruses to reduce expression of the replicase compared to other proteins, as the replicase is required in smaller quantities. The
ORF1ab PRF site performs this function in PRRSV – ORF1b encodes the viral RdRp and helicase. Modulation of PRF efficiency from the natural level is often detrimental to viral fitness, both in cases where it is increased or decreased, supporting the idea that PRF is used to finely tune ratios of proteins to within an optimal window\(^{521-523}\). The frameshift efficiency at the PRRSV ORF1ab site has not been measured in infected cells. It is assumed to be 15-20% based on experiments with another member of Arteriviridae, equine arteritis virus (EAV), for which a reporter gene construct containing the ORF1ab PRF site was transfected into HeLa cells\(^{524}\). Another study using a reporter construct based on the PRRSV ORF1ab PRF signal similarly yielded 16% frameshift efficiency, however this experiment was performed in yeast\(^{525}\). Although these transfected reporter construct assays may not accurately reflect PRF efficiency in the context of infection in the natural host, they give an approximate percentage of ribosomes which are expected to translate ORF1b. The frameshift efficiency at the nsp2 site in MARC-145 cells infected with EU PRRSV was calculated as 20% for \(-2\) PRF and 7% for \(-1\) PRF\(^{280}\). Ribosomes which frameshift at this site terminate translation 1 or 169 codons downstream, adding a second “sink” point at which the number of continuing ribosomes can be reduced to alter the stoichiometry of upstream and downstream proteins. This diversion of ribosomes may be a consequence of PRF that is important for viral fitness, as well as the functions of the frameshift products themselves\(^{280}\).

1.4.6.1.2 Protein functions

A brief summary of the functions of each protein on the PRRSV genome is shown in Table 1. Note that some of these functions have been attributed to PRRSV proteins by analogy to other arteriviruses, particularly EAV. As PRF is a prominent focus of this thesis, the functions of nsp2 and the frameshift products are discussed below in more detail.

1.4.6.1.2.1 Nsp2, nsp2TF and nsp2N

Nsp2 is a large, multi-functional protein which is essential for viral replication\(^{526}\). Some functions are shared with nsp2TF and nsp2N while others are predicted to be absent in the frameshift products due to differences in the C-terminal region of the proteins (Figure 1.20).

All the variants share the N-terminal portion, which contains a papain-like protease (PLP) 2 domain. This protease domain belongs to the ovarian tumour domain (OTU) superfamily and possesses both cis- and trans-cleavage activities\(^{527,528}\). It is responsible for the cleavage of the PRRSV polyprotein at the nsp2/3 site\(^{527}\). In EAV, nsp2 has additionally been shown to act as a co-factor for the nsp4 serine protease\(^{529,530}\). PRRSV nsp2 has also been shown to act as a deubiquitinase (DUB) towards cellular ubiquitin conjugates, and it has deISGylation activity that removes conjugates of ISG15, an interferon-induced ubiquitin homologue\(^{527,531-535}\). This has an immune antagonistic effect and interferon (IFN)-β signalling inhibition, DUB activity and deISGylation activity has been demonstrated for all three variants of nsp2, most strongly for nsp2N\(^{536}\).
Downstream of the PLP2 domain, all three nsp2 variants share a long hyper-variable region (HVR). This varies in size between isolates, from ~300-500 amino acids, and is largely responsible for the variation in PRRSV genome size^{428,537,538}. The HVR is the fastest-evolving region of the PRRSV genome and is frequently the site of amino acid substitutions, insertions and deletions^{427,428,539}. Deletions in nsp2 are often characteristic of highly pathogenic strains^{427,428,537}, although they can inhibit as well as enhance replication^{540-542}. The exact function of the HVR is unknown, although it has been shown to contain immune epitopes which induce production of neutralising antibodies^{543}, potentially explained by the finding that nsp2 is incorporated into the virion^{544}. The HVR has also been suggested to play a role in PAM-specific tropism^{537,545}. All three nsp2-related proteins are hyperphosphorylated, and one phosphorylation site in particular (Ser918 in EU PRRSV isolate SD01-08) has been shown to be important for viral replication and sgRNA production^{546}.

After the frameshift site the three proteins differ. Nsp2 contains two highly conserved domains which are absent from the frameshift products: a putative multi-spanning transmembrane (TM) domain, predicted to contain several membrane-spanning helices, and a cysteine-rich tail domain. In vitro experiments using artificial membranes showed that nsp2 is an integral membrane protein^{547}. This attribute is likely imbued by the predicted transmembrane domain, although the precise membrane-spanning regions were not conclusively defined^{547}. In EAV infection, nsp2 interacts with nsp3 to collaboratively manipulate ER membranes to create DMVs – the site of arterivirus RNA synthesis^{548-550}. Further, the transmembrane regions of EAV nsp2, nsp3 and nsp5 are thought to anchor the viral RTC to intracellular membranes^{548,549}. By analogy, similar functions have been suggested for PRRSV nsp2 but not proven experimentally. The function of the tail domain was unknown, until a recent observation that both the transmembrane and tail domains contribute to both aggrephagy and autophagy, with the tail domain inducing autophagy by binding the host protein 14-3-3ε^{539}.

Nsp2TF contains an alternative putative multi-spanning transmembrane domain^{280,281,547}, while Nsp2N is a truncated form of nsp2, which lacks any predicted transmembrane helices and does not have any additional domains^{280,281}. Likely due to these differences in the presence or characteristics of their putative transmembrane domains, the cellular localisation of the variants of nsp2 differs. Nsp2N is predicted to be cytosolic due to the lack of any predicted transmembrane domain^{280,536}. Nsp2 localises to the RTC in the perinuclear region^{548,549,551}. In infected MARC-145 cells, nsp2TF also localises to the perinuclear region, but it does not colocalise with RTCs (as determined by staining for nsp4 as an RTC marker)^{280,552}. Instead, the labelled foci were found to overlap with exocytic pathway markers of the intermediate compartment and Golgi complex^{280,552}. The PRRSV structural proteins GP5 and M (section 1.4.6.2.2) also localise to the exocytic pathway, where they associate with nsp2TF^{552}. The DUB activity of nsp2TF has recently been shown to prevent proteasomal degradation of these two structural proteins by catalysing their deubiquitination, likely
Introduction

playing a role in facilitating virus assembly\textsuperscript{280,552}. Finally, nsp2TF has also been shown to cause down-regulation of swine leukocyte antigen class I expression, dependent on the C-terminal 68 amino acids of nsp2TF\textsuperscript{553}.

Other functions of nsp2 require further characterisation before it can be determined whether they are shared with the frameshift products. For example, PRRSV nsp2 has been shown to associate with cellular proteins such as Bip (binding immunoglobulin protein, also known as Hspa5) a chaperone involved in the unfolded protein response\textsuperscript{554} and microtubule-associated protein 1 light chain kinase (LC3), which plays a role in autophagy\textsuperscript{555}.

1.4.6.1.2.2 Functions of other non-structural proteins

Nsp2 is one of four proteases encoded by the PRRSV genome. The major protease is nsp4, which is a serine protease responsible for all cleavage events downstream of nsp3 (references \textsuperscript{515,556,557}). Nsp1\(\alpha\) and nsp1\(\beta\) both contain PLP domains, although these differ structurally from the PLP domain of nsp2. They each cleave the junction directly downstream to free themselves from the polyprotein, and this is thought to be the only cleavage event they catalyse as they retain their C-terminus within their substrate-binding pockets, presumably inhibiting further activity\textsuperscript{288,558,559}. Nsp1\(\alpha\) may also play a role in sgRNA synthesis, discussed in section 1.4.6.2.1.

Many of the PRRSV non-structural proteins are found in the RTC. Alongside nsp2, nsp3 and nsp5 contain putative transmembrane domains and, between them, these proteins are thought to modify intracellular membranes to create DMVs and then anchor the RTC to this sub-cellular location\textsuperscript{548–551,560}. Nsp9 and nsp10 are the viral RdRp and helicase, respectively\textsuperscript{524,561}. The RdRp catalyses genome replication and transcription of sgRNAs, and has an additional, nidovirus-specific N-terminal domain termed the nidovirus RdRp-associated nucleotidyltransferase (NiRAN)\textsuperscript{562}. The NiRAN domain catalyses the covalent addition of nucleotide monophosphates to protein substrates, although the function of this within the Nidovirus replication cycle is currently unclear\textsuperscript{562,563}. The nsp10 helicase is a superfamily 1 helicase with confirmed polynucleotide binding and NTPase activity and 5'–3' polarity\textsuperscript{564–566}. It has a zinc-binding domain at the N-terminus which is implicated in sgRNA synthesis\textsuperscript{565,567}.

The function of nsp11 has been somewhat enigmatic, although it has been better characterised in recent years. It is an endonuclease with a preference for cleaving after uridylate, earning it the name NendoU: nidovirus endoribonuclease specific for uridylate\textsuperscript{568,569}. \textit{In vitro}, it displays broad substrate specificity and cleaves single-stranded (ss)RNA and double-stranded (ds)RNA\textsuperscript{570}. In cells, it is highly toxic when expressed outside the context of infection\textsuperscript{571,572}. This led to the suggestion that the restriction of the sub-cellular localisation of nsp11 to the perinuclear region may be an important way to prevent its expression becoming “suicidal” for the virus\textsuperscript{515,573}. Despite this, nsp11 is an essential protein for PRRSV replication and it is thought to play a role in viral RNA synthesis and
Porcine reproductive and respiratory syndrome virus

innate immune suppression\textsuperscript{572–577}. Surprisingly, nsp11 was also shown to have DUB activity, specifically cleaving K48-linked polyubiquitin chains\textsuperscript{575}. This activity is responsible for its ability to inhibit NF-κB activation, one of several different ways in which nsp11 inhibits interferon production\textsuperscript{572,573,576–578}.

The function of nsp12 was, similarly, only recently discovered. It is an essential protein; however, its role is not in the production of negative-sense gRNA but is specific to sgRNA production (both positive and negative sense)\textsuperscript{579}. This is dependent on two cysteine residues (C35 and C79 in PRRSV HuN4). This activity is consistent with reports that nsp12 is a central component of the RTC, where it acts as an interaction hub interacting with many nsps\textsuperscript{580,581}.

Many of the non-structural proteins of PRRSV are immune antagonists: nsp1α, nsp1β, nsp2, nsp2TF, nsp2N, nsp4, and nsp11, as well as the structural protein N (Table 1). Nsp1α and nsp1β are considered the most potent. Although nsp1β localises to the perinuclear region early in infection, both nsp1α and nsp1β later translocate to the nucleus\textsuperscript{551,582}. Nsp1α shuttles between the nucleus and the cytoplasm\textsuperscript{583}, inhibiting NF-κB activation by suppressing IκB phosphorylation in the cytoplasm\textsuperscript{584}, and inhibiting IRF3 signalling by degrading CREB-binding protein in the nucleus\textsuperscript{585–587}. Nsp1β inhibits interferon synthesis by imprisoning host mRNA within the nucleus\textsuperscript{588–590}. It achieves this by directly binding Nup62, causing the nuclear pore complex to disintegrate\textsuperscript{590}. Interestingly this was found to depend on a SAP motif (124-KXLQXXLXXGL-135) which overlaps the highly conserved 123-GKYLQRRLQ-131 motif that contains residues essential for PRF stimulation\textsuperscript{588–592} (section 1.2.2.1).

1.4.6.2 Structural proteins

1.4.6.2.1 Expression strategy

The ORFs encoding the structural proteins of PRRSV are located in the 3’-proximal region of the genome and are expressed from a nested series of sgRNAs (Figure 1.18). Except for sgRNA7, each sgRNA encodes multiple ORFs, although only one (usually the 5’-most) is expressed from each transcript. For sgRNA2 and sgRNA5, use of alternative initiation sites permits a proportion of ribosomes to translate small overlapping ORFs (E and 5a) instead of the longer main 5’ ORFs (GP2a and GP5). Each sgRNA is produced from a negative-sense template made by discontinuous transcription involving complementarity between a hexanucleotide transcription regulatory sequence (TRS) in the 5’ UTR (the leader TRS) and the 3’-proximal region (the body TRS)\textsuperscript{593}. The conserved TRS leader sequence in PRRSV, UUAACC, is located at the end of the 5’ UTR\textsuperscript{403,594,595}, while the TRS body sequences are more variable in sequence but maintain some degree of complementarity to the TRS leader\textsuperscript{403,426,595–600}. 
**Introduction**

The mechanism of sgRNA synthesis begins in a similar way to genome replication, and there are several “decision points” thereafter that determine whether a full length negative-sense gRNA or a shorter negative-sense sgRNA is produced\(^\text{403,515,593}\) (Figure 1.21). The RdRp begins synthesising the negative-sense complement to the viral genome at the 3′ end and proceeds in the 5′ direction until it reaches a TRS body. This represents a decision point at which the polymerase can either continue, or pause transcription and move to the TRS leader before continuing to add the reverse complement of the 5′ UTR to the nascent RNA\(^\text{594,595,598,600}\). If the RdRp continues uninterrupted at all TRS bodies, a negative-sense copy of the full length gRNA is made – the template for genome replication\(^\text{403,515,593}\) (Figure 1.21, upper). If the RdRp jumps to the leader TRS, this discontinuous transcription produces a negative-sense subgenome-length transcript. This transcript comprises the reverse complement of the 3′-proximal region fused with that of the 5′ UTR, omitting the region of the genome that encodes ORF1ab. This forms the template for sgRNA transcription\(^\text{595,598,600}\) (Figure 1.21, lower). To produce longer species of sgRNA, the RdRp must proceed, without jumping, through a greater number of TRS bodies, and this is thought to contribute to regulation of the ratios of sgRNAs to each other\(^\text{595}\).

**Figure 1.21 Arterivirus replication and sgRNA production**

Figure reproduced from Snijder et al., 2013 (reference 515). The leader region (light blue) is shared by all sgRNAs and the gRNA. In PRRSV the entire 5′ UTR is the leader, but for other nidoviruses the leader is shorter. L = leader TRS, B = body TRS, +/- = positive/negative sense.

An essential requirement for polymerase jumping is base pairing between the TRS body and TRS leader\(^\text{598,600}\). The leader TRS is located in the loop of a conserved stem loop, the leader TRS hairpin,
Porcine reproductive and respiratory syndrome virus

formed by the 3′ end of the 5′ UTR and the beginning of the nsp1α coding sequence\textsuperscript{598,601}. The structural context of the TRS leader is essential for efficient sgRNA synthesis and likely plays a role in recognition of the correct donor TRS sequence by the acceptor TRS body\textsuperscript{598,601}.

The factors that govern preferential selection of some TRS bodies over other identical hexanucleotide sequences elsewhere in the genome are incompletely understood. Complementarity that extends beyond the “core” TRS sequences has an effect, with stronger duplexes between the extended regions of negative-sense TRS body and positive-sense TRS leader being more likely to generate jump points\textsuperscript{600,602}. A common theme in the proposed mechanism for sgRNA synthesis is the attenuation of transcription over the TRS body, providing an opportunity for the RdRp to jump\textsuperscript{593,597,599,603}. As such, factors expected to impede the RdRp are likely to contribute to selection preferences. The RNA structure context of the TRS body likely plays a role\textsuperscript{501}, for example Berne virus (a nidovirus in the Torovirus genus) directs polymerase jumping to generate the S sgRNA via a stem loop instead of a TRS body\textsuperscript{604}. Secondary structures surrounding TRS bodies in EAV have been suggested to contribute, though TRS bodies themselves tend to be predicted as single-stranded, presumably to facilitate base pairing with the TRS leader\textsuperscript{597,599}. Long range interactions likely also play a role in both attenuating transcription and bringing the TRS leader and TRS body sequences into proximity, for example RNA-RNA interactions (as observed for coronaviruses\textsuperscript{605,606}), protein-protein interactions, or protein-RNA interactions\textsuperscript{599,603}.

Interactions involving proteins provide a possible mechanism for a “switch” between genome replication and sgRNA production (Figure 1.21) which represents a possible function for PRRSV proteins specifically implicated in sgRNA synthesis (nsp1, nsp10, and nsp12). Nsp1 is often suggested as the protein that is responsible for this switch, based on the observation that its deletion in EAV abolished sgRNA production but didn’t affect gRNA production\textsuperscript{607,608}. Nsp1 in EAV is not cleaved into nsp1α and nsp1β as it is in PRRSV, as the first PLP domain is not catalytically active. The activity in EAV is dependent on the zinc finger at the nsp1 N-terminus, although both PLP domains also play a role\textsuperscript{607–609}. In PRRSV, nsp1α likely performs an analogous role, as this contains a homologous zinc finger domain, and abrogation of nsp1α protease activity had a similar phenotype to deletion of EAV nsp1 (reference \textsuperscript{610}). This is likely due to inhibited release of nsp1α from the polyprotein as opposed to a direct impact of protease activity on sgRNA formation, as EAV nsp1 with defective PLP actives sites could rescue sgRNA production in trans-complementation assays\textsuperscript{607}. A similar phenotype of abrogated sgRNA synthesis and unaffected gRNA synthesis has been observed upon mutation or deletion of nsp10 and, recently, nsp12 (references \textsuperscript{565,579,611}). Nsp1α was favoured as the sgRNA-inducing protein over nsp10 partly due to the observation that many nsp10 mutations abolish gRNA synthesis as well, indicating it plays a role in both replication and transcription\textsuperscript{566}. In the case of all three proteins, further study is required to confirm their participation and elucidate potential mechanisms involved.
1.4.6.2.2 Protein functions

All eight proteins expressed from the sgRNAs are incorporated into the virion, and all except 5a are essential for the generation of infectious progeny\(^{612-614}\). PRRSV virions are enveloped, roughly spherical, and vary in size from \(\sim 50-60 \text{ nm}^{615}\) (Figure 1.22). They are composed of an internal core, made up of viral RNA bound by the N protein, which is positioned in a hollow cavity within a membranous envelope embedded with the remaining structural proteins\(^{615}\).

![Figure 1.22 PRRSV virion architecture](image)

**Figure 1.22 PRRSV virion architecture**

Figure adapted from Snijder et al., 2013 (reference \(^{515}\)). **A)** Transmission electron microscopy image of negatively stained, purified PRRSV particles. Bar = 25 nm. Original figure from Spilman et al., 2009 (reference \(^{615}\)). **B)** Cryo-electron microscopy image of a PRRSV particle in vitreous ice. Original figure from Spilman et al., 2009 (reference \(^{615}\)). **C)** Tomographic reconstruction of a PRRSV virion based on cryo-electron microscopy. Left: cut-away view of internal core, coloured by radius from particle centre (red to blue). Middle: cut-away view of the middle of the core, coloured as for left panel. Right: 6.3 nm slab through the centre of one PRRSV particle, with N protein C-terminal domain crystal structure superimposed (coloured). Original figure from Spilman et al., 2009 (reference \(^{615}\)). **D)** Schematic of virion architecture (protein stoichiometry chosen arbitrarily).

The internal core is pleomorphic, with a loosely organised filamentous structure. N forms dimers which are thought to “sandwich” the genome, likely binding RNA via the positively charged residues in the N-terminal domain\(^{615,616}\). A specific encapsidation motif is yet to be identified, although a 35 nt stretch within nsp1α has been implicated in packaging\(^{617-619}\), and N has been shown to bind the PRRSV UTR\(^{620}\).

The remaining proteins are transmembrane proteins found in the virion envelope. They are divided into “major” envelope proteins (GP5 and M), which are essential for virus assembly, and “minor” envelope proteins (GP2a, GP3, GP4, E and 5a). The major envelope proteins form a disulfide-linked heterodimer\(^{621-623}\), which interacts with many of the cellular receptors involved in PRRSV cell entry.
Porcine reproductive and respiratory syndrome virus

Porcine reproductive and respiratory syndrome virus (section 1.4.3). Heparin sulfate interacts with the dimer, while the interaction with MYH9 has been mapped to GP5, and the interaction with sialoadhesin is via sialic acid residues on the surface of GP5 (references 447,462,465). The GP5 ectodomain is an important target for immune epitope generation. It elicits both neutralising and non-neutralising antibodies, and it contains a hyper-variable region that may contribute to immune evasion, in addition to multiple N-glycosylation sites.

Of the minor envelope proteins, the three which are glycosylated (GP2a, GP3 and GP4) form heterotrimers (GP2a-GP3-GP4) and heterodimers (GP2a-GP4) which are essential for infectivity but not particle assembly. The two non-glycosylated minor structural proteins, E and 5a, are small proteins which are expressed from the same transcripts as Gp2a and Gp5, respectively, with which their coding sequences overlap (Figure 1.18). E is a myristoylated protein which associates with the minor GP heterotrimer. It has the properties of an ion channel protein and has been suggested to oligomerise to form an ion channel in the viral envelope to facilitate uncoating. The recently discovered 5a is a hydrophobic protein of unknown function, which is not essential for infectivity but does promote viral fitness.

N plays other roles in addition to encapsidation of the viral genome. In addition to being found at the site of viral RNA synthesis, it contains a nuclear localisation signal. Its ability to shuttle between the nucleus, where it interacts with several host proteins, and the cytoplasm is important for virulence. N also suppresses interferon by antagonising IRF3 activation and by binding TRIM25, preventing it from interacting with RIG-I to trigger a signalling cascade resulting in expression of IFNβ. N also interacts with vimentin, one of the PRRSV entry receptors.

1.4.6.3 UTRs

The 5’ UTR is 217-222 nt long for EU PRRSV and 188-191 nt for NA PRRSV and is ~96% conserved at the nucleotide level within species. Despite only 50% inter-species conservation at the nucleotide level, the predicted structure of EU PRRSV and NA PRRSV 5’ UTRs is reasonably well conserved (Figure 1.23), and the 5’ UTR of NA PRRSV can be replaced by that of EU PRRSV without affecting infectivity. In addition to the TRS leader hairpin, several other sequence or structural motifs in both the 5’ and 3’ UTRs are thought to be important in viral replication and infectivity.
Figure 1.23 PRRSV 5’ UTR predicted structure

Figure reproduced from Gao et al., 2013 (reference 637), who predicted the structure of the consensus sequence for EU PRRSV (i; 5 sequences) and NA PRRSV (ii; 11 sequences) using mfold. The TRS leader, found within the leader TRS hairpin, is shaded in grey and the ORF1a initiation codon indicated by a box. SL = stem loop.
## Table 1 PRRSV proteins and their functions

References can be found within the main text. Table continues overleaf.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein</th>
<th>Function</th>
</tr>
</thead>
</table>
|       | nsp1α   | PLP domain: cleave nsp1α/nsp1β junction  
Interferon suppression by multiple mechanisms  
POTa role in sgRNA synthesis |
|       | nsp1β   | PLP domain: cleave nsp1β/nsp2 junction  
Stimulation of PRF at nsp2 site  
Interferon suppression by nuclear imprisonment of host mRNA |
|       | nsp2    | Co-factor for nsp4 serine protease; interacts with nsp3 to manipulate ER membranes to create DMVs  
PLP (OTU) domain: cleavage of nsp2/3 site; DUB and delSGylation activity → immune antagonist  
HVR: immune epitope  
TM: membrane anchor for RTC  
Tail: induces autophagy by binding 14-3-3ε |
| ORF1a | nsp2TF  | Likely shares some other functions of nsp2 PLP/HVR regions  
PLP (OTU) domain: DUB and delSGylation activity → immune antagonist, prevent degradation of GP5 and M  
Alternative TM domain: target protein to exocytic pathway?  
C-terminal 68 amino acids: down-regulation of Swine Leukocyte Antigen class I expression |
|       | nsp2N   | Likely shares some other functions of nsp2 PLP/HVR regions  
PLP (OTU) domain: DUB and delSGylation activity → immune antagonist |
|       | nsp3    | Generate DMVs and anchor RTC |
|       | nsp4    | Serine protease, cleaves all junctions downstream of nsp3  
Interferon suppression by multiple mechanisms |
|       | nsp5    | Anchor RTC to DMVs  
Degradation of STAT3 by ubiquitin-proteasomal pathway |
|       | nsp6    | Function unknown |
|       | nsp7α   | Function unknown |
|       | nsp7β   | Function unknown |
|       | nsp8    | Function unknown |
| ORF1b | nsp9    | RdRpt  
NiRAN domain: covalent addition of nucleotide monophosphates to protein substrate |
|       | nsp10   | Superfamily 1 helicase; NTPase; zinc-binding domain implicated in sgRNA synthesis |
|       | nsp11   | Uridylate-specific endoribonuclease; DUB activity against K48-linked polyubiquitin chains; interferon suppression by several mechanisms |
|       | nsp12   | sgRNA synthesis; interaction hub for RTC |
| ORF2a | GP2a    | Minor envelope protein; glycosylated; forms GP2a-GP3-GP4 heterotrimer and GP2a-GP4 heterodimer |
| ORF2b | E       | Minor envelope protein; myristoylated; potential formation of ion channel involved in uncoating |
| ORF3  | GP3     | Minor envelope protein; glycosylated; forms GP2a-GP3-GP4 heterotrimer |
| ORF4  | GP4     | Minor envelope protein; glycosylated; forms GP2a-GP3-GP4 heterotrimer and GP2a-GP4 heterodimer |
## Introduction

<table>
<thead>
<tr>
<th>ORF5</th>
<th>GP5</th>
<th>Major envelope protein (N-glycosylated); forms GP5-M heterodimer; immune epitope generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF5a</td>
<td>5a</td>
<td>Minor envelope protein; unknown function</td>
</tr>
<tr>
<td>ORF6</td>
<td>M</td>
<td>Major envelope protein; forms GP5-M heterodimer</td>
</tr>
<tr>
<td>ORF7</td>
<td>N</td>
<td>Forms internal core of virion, encapsidating viral genome; Interferon suppression by multiple mechanisms</td>
</tr>
</tbody>
</table>
Chapter 2: Materials and Methods

This chapter describes the experimental and computational methods used to generate and analyse all three datasets in this study (covering both results chapters). In all instances, Tris refers to a solution of Tris base buffered using HCl to the given pH. Unless otherwise specified, I performed all experiments described.

2.1 Cells and viruses

2.1.1 EU PRRSV (PRRSV-1) infections (performed by Prof. Ian Brierley and Dr Adrian Mockett)

MA-104 cells, a cell line derived from foetal kidney tissue of the African green monkey species Chlorocebus sabaeus, were infected with a PRRSV strain derived from the Porcilis® vaccine strain (MSD Animal Health; GenBank accession KJ127878.1). The de novo assembled viral genome sequence from this study bears 148 substitutions relative to the GenBank accession entry and is given in Appendix section 1. Confluent 6 cm² dishes of MA-104 cells were infected at a multiplicity of infection (MOI) within the range 1-3. At 0 (mock), 8, 12, 21, and 25 hpi cycloheximide (Sigma-Aldrich) was added to the medium (final concentration 100 µg/ml) and incubated for 2 min. Cells were rinsed with 5 ml of ice cold PBS, placed on ice, and 400 µl lysis buffer (20 mM Tris pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 1 mM DTT, 1% Triton X-100, 100 µg/ml CHX and 25 U/ml TURBO™ DNase [Life Technologies]) was added drop-wise. Cells were scraped off the plate and triturated 10 times with a 26-G needle, before cell debris was removed by centrifugation (microfuge, 4°C, 20 min) and the supernatant was harvested and stored at −70°C.

2.1.2 NA PRRSV (PRRSV-2) infections (performed by Dr Pengcheng Shang, Dr Yanhua Li and Georgia Cook)

MARC-145 cells, a derivative of the MA-104 cell line, were maintained in Dulbecco’s modified Eagle’s medium (DMEM), supplemented with L-glutamine (1 mM), antibiotics, and foetal bovine serum (FBS) (7%), at 37°C and 5% CO₂. Cells were verified as mycoplasma-free by PCR (Universal Mycoplasma Detection Kit, ATCC) and deep sequencing. Cells were infected with NA PRRSV isolate SD95-21 (GenBank accession KC469618.1), or a mutant (KO2) based on this background. KO2 has been previously characterised²⁸⁰,²⁸¹,⁵³⁶, and has a modified nsp2 slippery sequence (G_GUU_UUU to G_GUA_UUC), a premature stop codon 5 nt downstream in the −2 frame, and several mutations in and around the conserved C-rich motif (U3886A, U3889C, C3895A, C3898G, C3902A, C3903G, C3904T, U3905G, U3906C, C3908A, G3910A, C3911G, U3912C, C3913A, C3914G, C3915A, C3917A, U3918C, U3919A, C3920G, C3921A, G3922A, C3923G, U3924C, G3925A, C3926G, C3927A).
U3899A, C3900G, C3901U, C3905U and C3907G), where all mutations are synonymous in the ORF1a reading frame.

In the generation of the NA PRRSV noCHX dataset, confluent 10 cm² dishes of MARC-145 cells were infected with WT or KO2 virus, or mock-infected. At 3, 6, 9, or 12 hpi, cells were washed with warm PBS and snap-frozen in liquid nitrogen. For the CHX pre-treated library (CHX-9hpi-WT, NA PRRSV CHX dataset), a confluent 10 cm² dish of MARC-145 cells was infected with WT virus. At 9 hpi, CHX was added to the medium directly prior to harvesting (final concentration 100 µg/ml), cells were incubated for 2 min, washed with ice cold PBS containing 100 µg/ml CHX, and snap-frozen.

Snap-frozen dishes were transferred to dry ice and 400 µl lysis buffer (composition as described in section 2.1.1) added. The dish was transferred to ice to defrost, and then cells were scraped and processed as described in section 2.1.1.
2.2 Western blotting

The lysates described in section 2.1 were resolved by 15% SDS-PAGE and transferred to 0.2 µm nitrocellulose membranes. Membranes were blocked with 5% Marvel milk powder (milk) dissolved in PBS (1 h, 25°C). Primary antibodies were diluted in 5% milk, PBS, 0.1% Tween-20 and incubated with membranes (1 h, 25°C). After three washes in PBS, 0.1% Tween-20, membranes were incubated with IRDye fluorescent antibodies in 5% milk, PBS, 0.1% Tween-20 (1 h, 25°C). Membranes were washed three times in PBS, 0.1% Tween-20 and rinsed in PBS prior to fluorescent imaging with an Odyssey CLx platform (LI-COR). Antibodies used were mouse monoclonal anti-NA-nsp1β (1/1000) and anti-EU-nsp1β (1/500) (kind gifts from Prof. Ying Fang), mouse IgM monoclonal anti-GAPDH (1/20,000, clone G8795, Sigma-Aldrich), goat anti-mouse IRDye 800 CW (1/10,000, LI-COR), and goat anti-mouse IgM (μ chain specific) IRDye 680RD (1/10,000, LI-COR). The EU PRRSV western blot was performed by Lior Soday.
2.3 Ribosome profiling library preparation

The experimental protocol for preparation of RiboSeq and RNASeq libraries is derived from Ingolia et al. (2009, 2012), Guo et al. (2010), Chung et al. (2015) and Irigoyen et al. (2016) (references 298,306-308,318). In section 2.3.1 and section 2.3.2 the core experimental protocol is described. For some libraries, slight variations on the protocol were performed, for example purification of a different range of fragment lengths. The specific adjustments made for each dataset are described in section 2.3.3 and summarised in Table 2. For the EU PRRSV dataset, library preparation was performed by Prof. Ian Brierley and Lior Soday.
### Table 2 Datasets in this study

<table>
<thead>
<tr>
<th>Virus</th>
<th>NA PRRSV noCHX</th>
<th>EU PRRSV</th>
<th>NA PRRSV CHX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virus</td>
<td>PRRSV SD95-21 (KC469618.1)</td>
<td>PRRSV MLV-DV (sequence in appendix)</td>
<td>PRRSV SD95-21 (KC469618.1)</td>
</tr>
<tr>
<td>Cells</td>
<td>MARC-145</td>
<td>MA-104</td>
<td>MARC-145</td>
</tr>
<tr>
<td>Timepoint(s) (hpi)</td>
<td>3, 6, 9, 12</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>CHX pre-treated?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fragment size purified (nt)</td>
<td><strong>9 hpi rep 1</strong>: 25-34 both RiboSeq and RNASeq</td>
<td>25-34 both RiboSeq and RNASeq</td>
<td>25-34</td>
</tr>
<tr>
<td></td>
<td><strong>9 hpi rep 2</strong>: 19-80 RiboSeq (&quot;broad spectrum&quot;), approx 50 RNASeq</td>
<td>All other libraries: 19-34 RiboSeq, approx 50 RNASeq</td>
<td></td>
</tr>
<tr>
<td>Adapters used to make amplicons</td>
<td>RA3N7p and RA5N7</td>
<td>RA3p and RA5 for replicate one, RA3N14p and RA5 for replicate two</td>
<td>RA3N7p and RA5N7</td>
</tr>
<tr>
<td>Infection and harvest performed by</td>
<td>Dr Pengcheng Shang, Dr Yanhua Li, Georgia Cook</td>
<td>Prof. Ian Brierley and Dr Adrian Mockett</td>
<td>Dr Pengcheng Shang and Dr Yanhua Li</td>
</tr>
<tr>
<td>Library preparation performed by</td>
<td>Georgia Cook</td>
<td>Prof. Ian Brierley and Lior Soday</td>
<td>Georgia Cook</td>
</tr>
</tbody>
</table>
2.3.1 Preparation of RNA fragments

For each lysate, 50-100 µl was taken and processed to generate RNASeq fragments, while 300 µl was processed to generate RPFs for RiboSeq.

2.3.1.1 Isolation of RNA fragments for RNASeq

The aliquot for RNASeq was incubated at 42°C for 30 min with 200 µl proteinase K solution (10 mM Tris pH 7.5, 10% SDS, 200 µg/ml proteinase K [NEB]), followed by sequential hot phenol:chloroform (Life Technologies) and chloroform extractions. RNA was purified by ethanol precipitation and resuspended in 10 mM Tris pH 7.5. Ribosomal RNA (rRNA) was removed using the Ribo-Zero Gold rRNA Removal Kit (Human, Mouse, Rat; Illumina) according to manufacturer’s instructions. RNA was purified by ethanol precipitation and fragmented by alkaline hydrolysis. 10 µl RNA sample (typically 5 µg) was incubated at 95°C for 15 min with 10 µl fragmentation buffer (2 mM EDTA, 10 mM Na₂CO₃, 90 mM NaHCO₃), 280 µl stop solution (300 mM NaOAc pH 5.5) was added and the products purified by ethanol precipitation and resuspended in H₂O.

2.3.1.2 Isolation of ribosome-protected fragments for RiboSeq

The aliquot for RiboSeq was thawed on ice and incubated with RNase I (100 U/µl, Life Technologies) at room temperature for 45 min on a rotating wheel. Total protein concentration in the sample was assayed by Pierce BCA (ThermoFisher Scientific), according to manufacturer’s instructions, as a proxy measure for cell density. This, along with analysis of the phase composition of libraries from earlier replicates within each dataset, was used to scale the volume of RNase I used (Table 3). SUPERase-In RNase inhibitor (20 U/ml, Life Technologies) was added at a ratio of 4:3 SUPERase-In:RNase to stop the RNase digestion, and the sample was centrifuged at 20,000 g for 2 min. Supernatant was applied to a 1 M sucrose cushion in polysome buffer (20 mM Tris pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 1 mM DTT, 100 µg/ml CHX) and ultracentrifuged for 16 hours at 28,000 rpm in a Beckman SW55Ti rotor at 4°C. The ribosome pellet was resuspended in 200 µl resuspension buffer (20 mM Tris pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 1 mM DTT) and incubated with 400 µl proteinase K solution (8.4 mM Tris pH 7.5, 1.5% SDS, 9.75 U/ml proteinase K [NEB]) at 42°C for 30 min. RNA (typically 1-2 µg) was purified and depleted of rRNA as described in 2.3.1.1.
Table 3 RNase added to RiboSeq lysates

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Libraries</th>
<th>Protein in lysate (mg/ml)</th>
<th>RNase I (U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA PRRSV noCHX 3 hpi</td>
<td>7.68</td>
<td>810</td>
<td></td>
</tr>
<tr>
<td>NA PRRSV noCHX 6 hpi</td>
<td>8.62</td>
<td>900</td>
<td></td>
</tr>
<tr>
<td>NA PRRSV noCHX 9 hpi rep 1</td>
<td>9.2</td>
<td>1500</td>
<td></td>
</tr>
<tr>
<td>NA PRRSV noCHX 9 hpi reps 2-4</td>
<td>9.2</td>
<td>1250</td>
<td></td>
</tr>
<tr>
<td>NA PRRSV noCHX 12 hpi</td>
<td>9.77</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>EU PRRSV all</td>
<td>not recorded</td>
<td>750</td>
<td></td>
</tr>
<tr>
<td>NA PRRSV CHX all</td>
<td>3.8</td>
<td>750</td>
<td></td>
</tr>
</tbody>
</table>

2.3.2 Gel purification of RNA fragments

RNA fragments were denatured at 80°C for 3 min in 60% v/v denaturing loading dye (Ambion) and loaded onto a 15% denaturing acrylamide gel, leaving an empty lane in between each sample. The gel was run at 15 W for 40 min and fragments of the required lengths excised and eluted in 600 µl gel extraction buffer (300 mM NaOAc pH 5.5, 1 mM EDTA, 0.25% SDS) overnight at 4°C on a rotating wheel. Eluted samples were ethanol precipitated and resuspended in 10 µl 10 mM Tris pH7.5.

2.3.3 Generation of amplicon libraries

The sequences of adapters and primers using during this protocol can be found in Appendix section 2.

The gel-purified fragments of RNA were denatured at 80°C for 2 min and the 3’ phosphate removed by incubation at 37°C for 2 h in a solution of 1 U/µl T4 PNK (NEB), 1 U/µl SUPERase-In, 1X PNK Buffer -A (NEB, no ATP), total volume 20 µl. After inactivation at 65°C for 10 min the RNA was ethanol precipitated and resuspended in 7 µl 10 mM Tris pH 7.5.

The 3’ adapter (RA3p, RA3N7p or RA3N14p) was pre-adenylated (15 µM adapter, 500 µM ATP, 15 µM Mth RNA ligase [NEB], 1X 5’ DNA Adenylation Reaction Buffer [NEB]) at 65°C for 1 h, inactivated at 85°C for 5 min and purified by phenol:chloroform extraction and ethanol precipitation. Pre-adenylated 3’ adapter (50 pmol) was added to the dephosphorylated sample and denatured at 80°C for 2 min, then incubated with a ligation solution (1X T4 Rnl2tr K227Q buffer [NEB, no ATP], 15% PEG8000, 1 U/µl SUPERase-In, 20 U/µl T4 Rnl2tr K227Q [NEB]; total reaction volume 20
µl) at 14°C overnight. The product was ethanol precipitated and gel purified (section 2.3.2), excising the region corresponding to the combined length of the original RNA fragment (section 2.3.1.1) and the relevant 3′ adapter.

The ligated products were re-phosphorylated at the 5′ end by incubation with a kinase mix (1 U/µl SUPERase-In, 1X PNK Buffer -A [no ATP], 1 mM ATP, 1 U/µl T4 PNK; total reaction volume 20 µl) at 37°C for 2 hours, followed by heat inactivation at 65°C for 10 min. RNA was purified by ethanol precipitation, resuspended in 10 µl 10 mM Tris pH 7.5, and incubated with 100 pmol 5′ RNA adapter (RA5 or RA5N7) in a ligation mix (1 U/µl SUPERase-In, 1X T4 Rnl buffer [Promega], 1 U/µl T4 RNA ligase [Promega; Rnl1]; total reaction volume 20 µl) at 14°C overnight. The product was ethanol precipitated and resuspended in 6 µl 10 mM Tris pH 7.5.

Half of the resuspended product was annealed to 50 pmol reverse transcription primer in 2.77 mM dNTP at 65°C for 5 mins, then incubated at 55°C for 50 min with a reverse transcription mix (1X first strand buffer [Invitrogen], 4.5 mM MgCl₂, 90 mM DTT, 1 U/µl SUPERase-In, 18 U/µl Superscript III [Invitrogen]; total reaction volume 11 µl) and heat inactivated at 85°C for 5 min.

PCR was performed using 2 µl of the reverse transcribed product, 20 pmol RP1 (a universal PCR primer) and 20 pmol RPI (an indexed PCR primer unique to each sample) in Phusion High-Fidelity PCR Master Mix (NEB), for 13 cycles of 10 s 98°C, 30 s 60°C, 15 s 72°C. The PCR product was purified using a Qiagen PCR clean-up kit, according to manufacturer’s instructions, ethanol precipitated, and resuspended in H₂O. Samples were gel purified on a 6% non-denaturing acrylamide gel and the region corresponding to the specific amplicon length excised. DNA was eluted in 600 µl DNA extraction buffer (300 mM NaCl, 10 mM Tris pH 8, 1 mM EDTA) at 4°C overnight on a rotating wheel, ethanol precipitated, and resuspended in 10 mM Tris pH 7.5.

### 2.3.4 High-throughput sequencing

The concentration of each amplicon library was measured using a high sensitivity dsDNA Qubit kit (ThermoFisher Scientific) prior to sample pooling. Pooled libraries were sequenced on single-end 75-cycle high output NextSeq 500 runs, with the exception of the NA PRRSV noCHX “9 hpi replicate two” RiboSeq libraries, which were sequenced on a paired-end run using a Mid Output v2 kit (150 cycles: 2×75), due to greater amplicon length. All pools were sequenced with an additional 1% PhiX spike-in to increase pool complexity.
2.3.5 Details of protocol variations

A summary of the specific details of library preparation for each dataset can be found in Table 2. Note that N7 and N14 in adapter names indicate the presence of the specified number of randomised bases (sometimes termed unique molecular identifiers) at the end destined for ligation. These are added to reduce the preference of enzymes for particular nucleotides causing bias, and to allow identification of PCR duplicates (as opposed to biological duplicates).

For the NA PRRSV noCHX dataset, all libraries were prepared using RA3N7p and RA5N7 adapters. For “9 hpi replicate one” libraries (both RiboSeq and RNASeq), RNA fragments from section 2.3.1.1 were gel purified to select 25-34 nt fragments. For all other RNASeq libraries, a gel slice centred at approximately 50 nt was purified. For “9 hpi replicate two” RiboSeq libraries, a single “broad spectrum” gel slice of 19-80 nt was purified. For all other RiboSeq libraries, a gel slice of 19-34 nt was purified.

For the EU PRRSV dataset, “replicate one” libraries were prepared using RA3p and RA5 adapters, and “replicate two” libraries were prepared using RA3N14p and RA5. For all libraries, RNA fragments from section 2.3.1.1 were gel purified to select 25-34 nt fragments.

For the NA PRRSV CHX dataset, the library was prepared using RA3N7p and RA5N7 adapters, and RNA fragments from section 2.3.1.1 were gel purified to select 25-34 nt fragments.
2.4 Ribosome profiling data analysis

Where read density is given as reads per million mapped reads (RPM) or reads per kilobase per million mapped reads (RPKM), the scaling factor to normalise for library size is the total number of positive-sense host mRNA-mapping reads plus positive-sense vRNA-mapping reads. In cases where only reads of certain lengths are selected for analysis, the scaling factor is also calculated based only on reads of the selected lengths. The genomic coordinates of regions of the PRRSV viral genome used for all analyses in both results chapters can be found in Appendix section 3.

2.4.1 Core pipeline

2.4.1.1 Mapping reads to viral and host genomes

For single-end libraries, fastx_clipper (FASTX Toolkit version 0.0.14, parameters: -Q33 -l 33 -c -n -v) was used to trim the universal adapter sequence from reads and to discard adapter-only reads, non-clipped reads, and reads with inferred original fragment lengths shorter than the minimum intended length experimentally purified. Adapter dimers were counted using grep and added to the adapter-only read count. For paired-end libraries, adapter trimming, read pair merging and removal of adapter-only reads was carried out using LeeHom (v.1.1.5) with the --ancient-dna option specified. This is designed for processing reads from fragments of ancient DNA, which have a similar length range to reads from these broad spectrum profiling libraries. Pairs of reads which LeeHom was unable to merge were put in the “non-clipped” category for the purposes of library composition analysis. Reads with inferred original fragment lengths shorter than the minimum intended length experimentally purified were removed using awk.

For libraries prepared using adapters with randomised bases, PCR duplicates were removed using awk, and seqtk (version 1.3) was used to trim the randomised bases from the reads. Bowtie (version 1.2.3) was used to map reads to host and viral genomes. Parameters “--n_n_mismatches --best” were used, where n_mismatches represents the maximum number of mismatches (one for RiboSeq, two for RNASeq) allowed between the read and the reference genome. These parameters mean bowtie reports a single alignment (the default) with the fewest mismatches. In cases where multiple alignments are equally good (for example alternative transcripts in the host mRNA database), bowtie randomly selects one to report. Reads can have any number of possible alignments – they are not discarded if they could possibly align to a large number of locations. Reads were mapped to each of the following databases in order and only reads that failed to align were mapped to the next database: ribosomal RNA (rRNA), virus RNA (vRNA), mRNA, non-coding RNA (ncRNA), genomic DNA (gDNA), contaminants. Note that for mapping to the contaminants database, the number of mismatches allowed was 0 for all libraries. Viral genome sequences used
for mapping are those specified in section 2.1 and were verified by de novo genome assembly using Trinity (version 2.9.1). The composition of the host and contaminants databases is given in Appendix section 4. The contaminants database is made up of potential contaminants that might be found in the laboratory or indicate undesirable co-infection, for example viruses worked on previously and mycoplasma sequences, and is a quality control measure to check that very few/no reads map to these contaminants. The position of vRNA within the database mapping order was altered to test and confirm that significant numbers of viral reads were not erroneously mapping to host databases or vice versa. Unless specifically stated (for example in the novel transcript analysis), all analyses were carried out using the reads mapped by bowtie as described above, meaning gaps or junctions within the alignment are not permitted.

2.4.1.2 Quality control

2.4.1.2.1 Host read length distribution and phasing
Positive-sense reads that map to the mRNA database in section 2.4.1.1 and with 5’ ends mapping between the first base of the start codon and 30 nt upstream of the stop codon were selected for inclusion in these analyses. This excludes initiating and terminating ribosomes. Phase was determined by the location of the read 5’ end relative to the annotated initiation codon on the transcript.

2.4.1.2.2 Distribution of reads relative to host mRNA start and stop codons
Positive-sense reads that map to the mRNA database in section 2.4.1.1 were filtered to select reads originating from transcripts with an annotated CDS of at least 150 codons, 5’ UTR of at least 60 nt, and 3’ UTR of at least 60 nt. For libraries from the NA PRRSV noCHX dataset, the required length of the 3’ UTR was increased to 90 nt to accommodate the greater read lengths of the RNASeq libraries. These plots are the only case in which read density is plotted on the genome/transcript at the 5’ end coordinate of the reads instead of the inferred P site of the ribosome that generated the reads. This is because these plots are used to verify the typical distance between the read 5’ end and the ribosomal P site (which is 12 nt for all datasets in this study).

2.4.1.2.3 Length distribution of reads mapping to host transcript CDSs compared to 3’ UTRs
Host mRNAs with a CDS of at least 150 codons and a 3’ UTR of at least 100 codons were included in the analysis. The length distributions of positive-sense reads with inferred P site positions in the CDS (codons −100 to −10 relative to the stop codon) and 3’ UTR (codons +10 to +100 relative to the stop codon) were determined and normalised by the total number of reads in each category to make the scales comparable. Some transcripts are likely to have a shorter 3’ UTR than is annotated, which would lead to a slight underestimation of 3’ UTR density.
2.4.1.2.4 Virus read length distribution and phasing

For plots of general viral read length distribution and phasing, positive-sense reads with 5′ ends mapping to viral coding sequences were used, with regions in which two ORFs overlap removed. For plots of reads mapping to specific regions of the viral genome, positive-sense reads with 5′ ends mapping to the regions specified in Appendix section 3 were selected for inclusion. Note that overlapping ORFs were permitted for the length distribution plots but not the phasing plots. For phasing plots, phase 0 was designated independently for each region, relative to the first nucleotide of that region’s ORF, to aid comparison (e.g. the dominant phase in ORF1b is designated 0 instead of -1). For the negative-sense read analysis, reads mapping to anywhere on the viral genome were used, and phase was determined using the 5′ end of the read (the 5′ end of the reverse complement reported by bowtie plus the read length).

2.4.1.3 Plots of read distribution on virus genomes

Read density is plotted at the inferred position of the ribosomal P site, defined for the datasets in this study as the 5′ end coordinate of the read + 12 nt. While this adjustment is only relevant for RiboSeq, it is applied to RNASeq as well for comparability. Where species-specific tRNA adaptation index (stAI) is displayed underneath genome sequences, stAI values were obtained from STADIUM^394 and heatmaps were constructed using the minimum and maximum values for codons from the relevant organism as the most extreme colours available in the gradient. No stAI databases were available for *C. sabaeus* so stAI values from the related species *Macaca mulatta* (downloaded 4th Oct 2020) were used.

For plots of read density ratios on one virus genome compared to another, sliding window sizes were tested to ensure the absence of any instances of zero in the denominator library after sliding window application. UTRs were excluded, and for many plots the region of analysis was further truncated to a specific region of the viral genome. The half-window was then excluded from either end of the region of interest and this formed the final region from which the data were taken to test sliding window sizes and perform the division (after application of the running mean filter). Only positive-sense reads were used for these plots.

For plots of the percentage of reads in each phase across regions of the viral genome, a similar strategy was employed to avoid mathematical errors or noise when attempting calculations for codons with no (or very few) attributed reads. Reads were separated according to phase, and a suitably sized running mean filter applied to avoid any instances of zero across the region of interest, as described above. From this, the percentage of reads in each phase at each codon in-frame with the feature of interest (i.e. PRRSV ORF1a) was calculated.
2.4.2 Frameshift efficiency calculations

Frameshift efficiency (%) was calculated by three different methods. The pairings of each WT library with the corresponding KO2 library were as indicated in Table 4.

Table 4 WT and KO2 library pairings

<table>
<thead>
<tr>
<th>WT library</th>
<th>KO2 library</th>
<th>Paired based on</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 hpi WT 1</td>
<td>3 hpi KO2 1</td>
<td>ranking of virus:host read ratios in RNASeq</td>
</tr>
<tr>
<td>3 hpi WT 2</td>
<td>3 hpi KO2 2</td>
<td>ranking of virus:host read ratios in RNASeq</td>
</tr>
<tr>
<td>6 hpi WT 1</td>
<td>6 hpi KO2 1</td>
<td>ranking of virus:host read ratios in RNASeq and RiboSeq (rankings agree)</td>
</tr>
<tr>
<td>6 hpi WT 2</td>
<td>6 hpi KO2 2</td>
<td>ranking of virus:host read ratios in RNASeq and RiboSeq (rankings agree)</td>
</tr>
<tr>
<td>9 hpi WT 1</td>
<td>9 hpi KO2 1</td>
<td>libraries were processed together and had different fragment lengths purified compared to other replicates</td>
</tr>
<tr>
<td>9 hpi WT 2</td>
<td>9 hpi KO2 2</td>
<td>libraries were processed together and RiboSeq libraries had different fragment lengths purified compared to other replicates</td>
</tr>
<tr>
<td>9 hpi WT 3</td>
<td>9 hpi KO2 4</td>
<td>ranking of virus:host read ratios in RNASeq and RiboSeq after excluding reps 1 and 2 (RiboSeq and RNASeq rankings agree)</td>
</tr>
<tr>
<td>9 hpi WT 4</td>
<td>9 hpi KO2 3</td>
<td>ranking of virus:host read ratios in RNASeq and RiboSeq after excluding reps 1 and 2 (RiboSeq and RNASeq rankings agree)</td>
</tr>
<tr>
<td>12 hpi WT 1</td>
<td>12 hpi KO2 1</td>
<td>ranking of virus:host read ratios in RNASeq and RiboSeq (rankings agree)</td>
</tr>
<tr>
<td>12 hpi WT 2</td>
<td>12 hpi KO2 2</td>
<td>ranking of virus:host read ratios in RNASeq and RiboSeq (rankings agree)</td>
</tr>
</tbody>
</table>

2.4.2.1 Nsp2 density normalisation-based method

For the nsp2 frameshift site, the density normalisation-based method calculates frameshift efficiency as $100 \times [1– (\text{downstream}/\text{upstream})]$, where downstream and upstream represent the reads per kilobase per million mapped reads (RPKM) values for the respective regions after normalisation of WT density by KO2. For RiboSeq libraries, read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used, while for RNASeq negative control libraries all read lengths were used. This method could not be applied to EU PRRSV libraries as no KO2 libraries were made.
2.4.2.2 Nsp2 phasing-based method

The phasing-based method of calculation of FS efficiency at the nsp2 site instead uses the proportion of reads in each phase in the upstream and transframe regions. This is represented here as: transframe_{-2} (the proportion of reads in the transframe region in the \(-2\) phase) and upstream_{0} and upstream_{-2} (the proportion of reads in the upstream region in the 0 and \(-2\) phases respectively), where in both regions phase is taken relative to the ORF1a reading frame. If we take the upstream region as providing the ‘input’ ribosomes, we can assume these are translating in the 0 frame. Observing the phase distribution in this region allows estimation of what proportion of 0-frame ribosomes generate reads attributed to the 0 phase (upstream_{0}), and what proportion generate reads attributed to the \(-2\) phase (upstream_{-2}). This can be extrapolated to determine what proportion of reads we expect in the \(-2\) phase in the transframe region (transframe_{-2}) in the absence of frameshifting (we expect this to be approximately the same proportion as in the upstream region). If a certain proportion (FS_proportion) of ribosomes undergo \(-2\) PRF, this will mean that, between the upstream and downstream region, FS_proportion of phase 0 reads will change from the 0 to the \(-2\) phase. It also means FS_proportion of \(-2\) phase reads will move to the \(-1\) phase (leaving 1 – FS_proportion in the \(-2\) phase). We can combine these concepts to make the equation:

\[
\text{transframe}_{-2} = (\text{FS}_\text{proportion} \times \text{upstream}_{0}) + (1 - \text{FS}_\text{proportion}) \times \text{upstream}_{-2}
\]

This can be rearranged to calculate percentage frameshift efficiency (which is FS_proportion expressed as a percentage):

\[
\text{FS}_\text{efficiency} = 100 \times \frac{\text{transframe}_{-2} - \text{upstream}_{-2}}{\text{upstream}_{0} - \text{upstream}_{-2}}
\]

This phasing-based method of calculating frameshift efficiency should theoretically be unaffected by RNP contamination, provided the RNP footprints are equally distributed between the three phases. Let R be the proportion of total reads that are RNPs, and let P_0 and P_{-2} be the proportion of total RNPs that are attributed to the 0 and \(-2\) phases, respectively. Reads originating from RNPs are not expected to change phase due to frameshifting. Therefore, the equation for calculating the fraction of reads that change from the 0 to \(-2\) phase becomes:

\[
\text{FS}_\text{proportion} \times (\text{upstream}_0 - \text{RP}_0)
\]

and the equation for calculating the fraction of reads that remain in the \(-2\) phase becomes:

\[
\text{RP}_{-2} + [(1 - \text{FS}_\text{proportion}) \times (\text{upstream}_{-2} - \text{RP}_{-2})]
\]

Combining these makes the equation:

\[
\text{transframe}_{-2} = 
\text{FS}_\text{proportion} \times [\text{upstream}_0 - \text{RP}_0] + \text{RP}_{-2} + [(1 - \text{FS}_\text{proportion}) \times (\text{upstream}_{-2} - \text{RP}_{-2})]
\]

This rearranges to:
FS_proportion = \frac{(\text{upstream}_{-2} - \text{transframe}_{-2})}{(\text{RP}_0 - \text{upstream}_0 - \text{RP}_{-2} + \text{upstream}_{-2})}

If P_0 = P_{-2} \text{ (for example if RNPs are equally distributed between all three phases then } P_0 = P_{-2} = \frac{1}{3}) \text{ then this causes both terms involving } R \text{ to cancel out of the equation. The remaining equation is the same as the original equation in which we did not consider RNPs, i.e. RNPs should not theoretically affect the result if distributed equally between the three phases. The same would hold for any other form of uniform non-phased contamination.}

For NA PRRSV libraries, read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used and, for EU PRRSV, read lengths with good phasing (indicated in Figure 3.10) were used.

### 2.4.2.3 ORF1ab density-based method

For the ORF1ab frameshift site, the density-based method calculates frameshift efficiency as 100 × (downstream/upstream), where downstream and upstream represent the reads per kilobase per million mapped reads (RPKM) values for the respective regions. The mutations on the KO2 genome should not affect frameshifting at the ORF1ab site so normalisation by a frameshift-defective mutant was not possible. KO2 libraries were treated the same as WT libraries for the calculations and the statistical tests. For NA PRRSV RiboSeq libraries, read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used, and for EU PRRSV RiboSeq libraries, read lengths with good phasing (indicated in Figure 3.10) were used. For all RNASeq negative control libraries, all read lengths were used. A two-tailed Mann-Whitney U test was employed to assess statistical significance of differences between groups of observed values.

### 2.4.2.4 Bootstrap resampling

In all cases, 100,000 randomised resamplings of codons in each respective region were performed. Each WT library was paired with its corresponding KO2 library (as indicated in Table 4), with matched codons selected for the two libraries in each resampling. Calculation of frameshift efficiency for each resampling was performed as described above, including selection of read lengths for inclusion. Regions and bounding coordinates used were also the same (Appendix section 3) except for the nsp2 phasing-based method, in which an additional 147-codon (the same length as the region of nsp2TF used) downstream region was added as a negative control (Appendix section 3). For all resamplings, n codons were sampled with replacement, where n is the total number of codons in the region undergoing resampling. Bootstrap resamplings were used to empirically determine p values. To determine whether PRF efficiency was significantly different between groups, the number of datapoints from the combined bootstrap distributions of all the libraries in one group were
Materials and Methods

compared to the mean average of the observed values (i.e. the values displayed in Figure 3.23A and Figure 3.24A) of the libraries in the second group. The number of bootstrap datapoints that were at least as extreme as the mean observed value was divided by the total number of bootstrap datapoints to generate the $p$ value.

2.4.3 Chimeric read analysis for novel transcript discovery

Reads which did not map to any of the host or viral databases (rRNA, vRNA, mRNA, ncRNA, gDNA, or contaminants) using bowtie (section 2.4.1.1) were used as input for mapping using the splice-aware mapper STAR\textsuperscript{646} version 2.7.3a. Reads mapped using STAR may have gaps and junctions within their alignment to the reference genome. To identify “chimeric” PRRSV reads (which map discontinuously to the viral genome) resulting from discontinuous transcription by the virus RdRp, mapping parameters were selected to switch off penalties for non-canonical splice junctions, based on parameters suggested in Kim \textit{et al.}, 2020 (reference \textsuperscript{647}). Where multiple (potentially discontinuous) alignments of the entire read were possible, reads with two reported alignments were permitted but reads with three or more potential alignments were discarded. For junctions, an overhang of at least 12 nt was required upstream of the “donor” site (5′-most on the genome) and downstream of the “acceptor” site (3′-most on the genome) to ensure specificity of mapping (\texttt{--alignSJoverhangMin}). Parameters used were:

\begin{verbatim}
--runMode alignReads --outSAMtype BAM SortedByCoordinate --outFilterType BySJout
--outFilterMultimapNmax 2 --alignSJoverhangMin 12 --outSJfilterOverhangMin 12 12 12 12
--outSJfilterCountUniqueMin 1 1 1 1 --outSJfilterCountTotalMin 1 1 1 1
--outSJfilterDistToOtherSJmin 0 0 0 0 --outFilterMismatchNmax 2 --scoreGapNoncan -4
--scoreGapATAC -4 --chimOutType Junctions --chimScoreJunctionNonGTAG 0
--alignEndsType EndToEnd --alignSJstitchMismatchNmax -1 -1 -1 -1 --alignIntronMin 20
--alignIntronMax 100000 --outSAMattributes NH HI AS nM jM jI
\end{verbatim}

In cases with stretches of near-identical bases at the donor and acceptor site of a junction, such as TRS leader-TRS body junctions, it is not possible for STAR to determine the exact position of the junction, and it can sometimes be assigned to the beginning of the stretch of identity, sometimes at the end. In order to merge these alternative assignments of junctions which potentially represent the same biological junction, neighbouring junctions were clustered and merged. First, low support junctions were removed by filtering late timepoint libraries (8 hpi onwards) to select only junctions with four or more reads supporting them. Reads which passed the filter were split into two categories, TRS-spanning and non-TRS-spanning, according to whether the donor site of the junction overlaps the TRS leader (genomic coordinates in Appendix section 3). Junctions were clustered by their acceptor site coordinate, with junctions being included in a cluster if the acceptor site coordinate
was within seven (for TRS-spanning junctions) or two (for non-TRS-spanning junctions) bases of the acceptor site of another junction in the cluster. Then, within each unique cluster of junctions grouped by acceptor site, junctions were clustered again in the same way but instead according to donor site. If the acceptor sites in each new cluster no longer formed a single cluster according to the definition specified above, the new cluster was subjected to another round of acceptor site and donor site clustering as described above. Finally, the junctions within each cluster were merged. The number of reads supporting the merged junction was defined as the sum of the supporting read counts for all the input junctions in the cluster. The positions of the donor and acceptor sites were defined as the midpoints of the ranges of coordinates observed for these designations in the input junctions of the cluster.

To identify junctions with high support, merged junctions were filtered so that only junctions which were present in more than one replicate (considering WT and KO2 as one group) pass the filter. Junctions were defined as matching if the ranges of the donor and acceptor coordinates for the junction in one library overlap with those of a junction in a second library. The matching junctions from all replicates in which they occur were merged as described above to make the final merged junction. TRS-spanning junctions were then designated as “known” junctions if they were the major junction responsible for one of the known canonical sgRNAs of PRRSV. For the NA PRRSV M junction there is a stretch of six bases that is identical upstream of the TRS leader and the TRS body, leading to separation of the two alternative junction position assignments by a distance greater than the seven bases required to combine TRS-spanning junctions into clusters. The two junction clusters that are assigned either side of this stretch of identical bases were specifically selected and merged into one at this stage. Non-TRS-spanning junctions were filtered according to whether they represent local (≤ 2000 nt) or distant (> 2000 nt) deletions.

The proportion of chimeric reads at donor and acceptor sites was calculated as chimeric/(chimeric + non-chimeric). The number of non-chimeric reads at the junction site was defined as the number of reads spanning at least the region 12 nt either side of the midpoint position of the donor or acceptor site. For all TRS-spanning junctions the donor midpoint was set according to the known TRS leader sequence (genomic coordinate 188 for NA PRRSV and 219 for EU PRRSV).

2.4.4 Sequence conservation plots

The NA PRRSV multiple sequence alignment was generated by Charlotte Tumescheit. Sequences were selected for inclusion based on requirement for ≥ 50% amino acid identity, ≥ 70% nucleotide identity and ≥ 95% coverage compared with the reference NA PRRSV sequence NC_001961 (resulting in 661 sequences). For some plots, these input sequences were clustered by Charlotte Tumescheit, based on the whole genome and with a nucleotide similarity threshold of 95%, and one
representative sequence from each cluster was selected to make a sequence alignment of 137 sequences (collection of genomes termed “representative of the diversity of the phylogeny”). The EU PRRSV multiple sequence alignment was generated by selecting sequences with ≥ 70% nucleotide identity and ≥ 95% coverage compared to the reference EU PRRSV sequence NC_043487 (resulting in 120 sequences). Logo plots and mini-alignment plots were generated using CIALign, with use of the --remove_insertions function where necessary to remove bases where over 50% of sequences have a gap. For the uORF analyses, genome sequences which began partway through the uORF(s) were excluded. Synplot results were provided by Dr Andrew Firth.

2.4.5 Novel ORF analysis

Novel ORF discovery was performed by visual inspection of the data and by using PRICE (version 1.0.3b). To run PRICE, a custom gtf file was made for each virus, with only the genomic RNA transcript and ORF1a annotated. Other known viral ORFs were not annotated but instead served as positive controls. These were each individually concatenated with the host gtf file (ChlSab1.1.101, downloaded from Ensembl) to make the input of known ORFs. The reference fasta files for the host and virus were similarly concatenated to make the input reference sequence. Reads were mapped to these combined references using bowtie, as described in the core pipeline but allowing one mismatch, and inputted into PRICE as one group per timepoint. WT and KO2 were treated as separate groups for the 9 hpi timepoint. Virus ORFs with \( p \leq 0.05 \) (no multiple testing correction) were selected for inclusion in the final output tables. For noCHX NA PRRSV libraries, read lengths with minimal RNP contamination were used, while for EU PRRSV libraries all read lengths were used.

2.4.6 Transcript abundance and translation efficiency analyses

Regions used for calculation of read density on each ORF or transcript can be found in Appendix section 3. For transcript abundance calculations, regions for sgRNA were defined as the region between the TRS body for the mRNA in question and the next major TRS body downstream. For RiboSeq density calculations, regions were defined as starting at the beginning of the ORF and ending at either the termination codon or, where relevant, just upstream of the next major TRS body. This avoids any potential confounding effects of translation of truncated forms of ORFs by ribosomes on downstream sgRNAs. For samples harvested without CHX pre-treatment (NA PRRSV libraries), termination codons were excluded, and for CHX pre-treated samples (EU PRRSV libraries), initiation codons were excluded. TRS bodies considered “major” are those that give rise to the transcripts depicted in Figure 4.17. In cases such as the N sgRNA, where two TRS bodies are
thought to give rise to two different forms of a transcript that express the same ORF(s), these were treated as a single transcript for the purposes of this analysis. In such cases, the beginning of the region of measurement for the heterogeneous transcript was defined as the downstream TRS, and the end of the region of measurement for the neighbouring sgRNA upstream was designated the upstream TRS. For assessment of transcript abundance using junction-spanning reads in these cases, read counts for both junctions were combined. Gp3-iORF and its putative associated sgRNA were excluded from the analysis. This was due to low abundance of the sgRNA most likely to facilitate its translation, and to avoid reducing the accuracy of density measurements for sgRNA2, which would have to be greatly truncated to permit measurement of the putative Gp3-iORF sgRNA.

For the decumulation analysis, input reads are those from the bowtie mapping pipeline (section 2.4.1.1), and alignment of each read to the reference genome is therefore without gaps or junctions. Decumulated RNASeq density (in RPKM) of each transcript is defined as the raw RNASeq density within that transcript’s region minus the density in the designated region directly upstream. For gRNA, no decumulation was performed and the raw value was used. The leader is present on the genomic RNA and all sgRNAs. As the raw RNASeq density in the N region is the combined result of the density of all viral transcripts, this was used as the RNASeq density for the leader in this method. The decumulation procedure can lead to negative transcript abundance values, likely due to a combined effect of noise and technical biases (most likely to be an issue for transcripts which are lowly expressed relative to their upstream transcript). These do not represent biologically relevant results and were excluded from all plots, mean calculations, and TE calculations. For 9 hpi WT replicate 4, the TE value for nsp10-iORF2 was a clear outlier, likely inflated due to a low calculated transcript abundance, and was excluded from the decumulated TE plot and mean calculations.

For the junction-spanning reads analysis, input reads are those from the STAR mapping pipeline (section 2.4.3), thus junctions within the alignment of each read to the reference genome are permitted. Abundance of the gRNA transcript is defined as the number of reads which span 12 nt either side of the midpoint of the TRS leader (genomic coordinates 176-199 for NA PRRSV, 207-230 for EU PRRSV). This is analogous to the 12 nt overhang required either side of a junction to qualify for mapping by STAR; however, these reads are not chimeric, and map specifically to gRNA. In the junction-spanning read analysis, leader abundance is defined as the total number of reads for all transcripts in the analysis combined, as the leader is present on all sgRNA species and the gRNA. For junction-spanning reads, normalisation by region length is not appropriate so read counts were normalised by library size only. For plots with logarithmic axes, data points with a value of zero were excluded from the plot, but not from mean calculations. These data points were also excluded from all TE calculations. For libraries with shorter read lengths (EU libraries and NA 9 hpi replicate 1 libraries) chimeric read counts are lower (and also subject to greater inaccuracies as a result of less dilution of possible read start- and end-point specific ligation biases) due to the requirement for a 12
Materials and Methods

nt overhang either side of the junction effectively representing a much larger proportion of the total read length. As such, these libraries are not directly comparable to the remaining NA PRRSV libraries and they were plotted separately and not included in NA mean calculations.

For all transcript abundance calculations, all read lengths were used, and reads mapping to any phase were counted. For the RiboSeq analysis, reads from the bowtie mapping pipeline (section 2.4.1.1) were used as input, and only reads mapping to the phase corresponding to the frame of the ORF were counted. Therefore, read lengths with good phasing and minimal RNP contamination were selected for inclusion in the RiboSeq analysis, as indicated in Figure 3.8 for NA PRRSV and Figure 3.10E for EU PRRSV.

To calculate host TE, the method was designed to be comparable to the decumulated viral TE calculations. Reads which mapped within the CDSs of host mRNAs were used, with initiating and terminating ribosomes excluded. The same read lengths as for the viral analysis were used, and for the RiboSeq analysis only reads mapping to the phase corresponding to the frame of the ORF were counted. Transcript variants of the same gene were grouped using the C. sabaeus RefSeq annotation GCF_015252025.1, genome build Vero_WHO_p1.0, gtf version 2.2. Readcounts for all transcript variants were combined, normalised by library size, and divided by the length of the longest annotated CDS of the component transcript variants to generate the read density. This will likely slightly underestimate transcript abundance and therefore slightly overestimate translation efficiency.

2.4.7 Data availability

All data presented in this thesis were deposited on ArrayExpress, to be publicly available after publication of the relevant papers. The PRRSV datasets were deposited under accession numbers E-MTAB-10621, E-MTAB-10622 and E-MTAB-10623.
Chapter 3: Analysis of programmed ribosomal frameshifting on the PRRSV genome

3.1 Introduction

Chapters 3 and 4 of this thesis describe experiments to investigate changes in viral gene expression over the course of PRRSV infection. A particular interest was taken in programmed ribosomal frameshifting, due to the unusual nature of the PRRSV nsp2 PRF site. As such, this first chapter predominantly focuses on programmed ribosomal frameshifting, aiming to address four questions:

1. What is the frameshift efficiency at the nsp2 and ORF1ab PRF sites in the context of infection?
2. Does frameshift efficiency change as infection progresses?
3. Does frameshift efficiency differ between the two species of PRRSV?
4. Is there any evidence of ribosomal pausing at the frameshift site, particularly at the non-canonical nsp2 site, which might have implications for our understanding of frameshift mechanisms?

Given the capacity of ribosome profiling to provide a global analysis of translating ribosomes at sub-codon resolution, this was deemed the most appropriate technique to address these aims. RNASeq was performed in parallel, both to provide a control for the shared technical biases in RiboSeq/RNASeq library preparation, and to supplement information on the translatome with information on the viral transcriptome. Both techniques were applied to cells harvested over a 12 hour timecourse of NA PRRSV infection, and to EU PRRSV libraries from a single timepoint (8 hpi).

In order to investigate frameshifting at the nsp2 PRF site, a frameshift-defective mutant virus, KO2, was employed in addition to WT virus in this study. Two PRF-defective mutant viruses have been previously generated and characterised. The KO1 mutant virus bears mutations that introduce stop codons into the −2 frame at codons 100 and 102 of the 169-codon nsp2TF transframe region. These lead to expression of a truncated form of nsp2TF, although the PRF site itself is left intact, so both −1 and −2 frameshifting still occur at WT levels, and nsp2N is still generated. The KO2 mutant virus instead bears mutations in the slippery sequence and PCBP-binding motif (Figure 3.1A). These mutations render the nsp2 PRF site unable to bind PCBP, induce −1 or −2 frameshifting, or permit production of nsp2N or nsp2TF. As an additional measure to ensure nsp2TF is not produced, two stop codons were introduced into the −2 frame shortly downstream of the PRF site (Figure 3.1A). For both mutants, all mutations are synonymous with respect to the 0-frame nsp2 ORF.
Analysis of programmed ribosomal frameshifting on the PRRSV genome

Figure 3.1 The PRF-defective KO2 mutant virus

A) Nucleotide sequence at the nsp2 PRF site of the North American lineage PRRSV viruses used in this study (SD95-21). Mutations made to disrupt PRF and/or expression of nsp2TF in the KO2 mutant virus are highlighted in blue. B) Plaque phenotype of WT and PRF-defective
viruses in MARC-145 cells. Figure adapted from Fang et al., 2012 (reference 280). This study used an isolate from the European lineage, SD01-08. Analogous mutations to those in panel A were made to generate vSD-KO2. The vSD-KO1 virus has a WT PRF site but has stop codons introduced at codons 100 and 102 of the 169-codon transframe ORF of nsp2TF. C) Growth characteristics of WT, KO1 and KO2 viruses on the background of the North American lineage isolate SD95-21. Figure adapted from Li et al., 2018 (reference 536). MARC-145 cells were infected at MOI 0.01. D) Growth characteristics of viruses from panel B (European lineage, SD01-08 isolate). Figure adapted from Fang et al., 2012 (reference 280). MARC-145 cells were infected at MOI 0.1. Note that the colours of KO1 and KO2 are reversed compared to panel C. E) Growth characteristics of viruses defective in PRF at the nsp2 site. Figure adapted from Cao et al., 2016 (reference 553). Mutations were made on the background of the pathogenic North American lineage isolate VR2385 to knock out PRF at the nsp2 site. All mutants contain the same nucleotide substitutions in the slippery sequence and PCBP-binding motif as the KO2 mutant depicted in panel A but differ in the number and location of −2-frame stop codons introduced (all mutations synonymous with respect to the 0-frame nsp2 ORF). They are all unable to stimulate frameshifting at the nsp2 PRF site. MARC-145 cells were infected at MOI 0.1.

Comparison of these two mutants gives an insight into the relative fitness contributions of nsp2TF compared to nsp2N and the potential translation regulation enacted by PRF itself. Both mutant viruses display a small plaque phenotype and impaired growth kinetics compared to WT, and KO2 is more impaired than KO1 (Figure 3.1B-D)280,536. In pigs, both KO1 and KO2 led to reduced viral load compared to WT virus, with KO2 again more impaired than KO1 (reference 650). This demonstrates that nsp2TF itself is important for viral fitness and suggests that there is an additional fitness advantage conferred by PRF and the production of nsp2N. Within the latter category, this experiment is not able to distinguish the relative contribution of nsp2N production, production of a full-length as opposed to truncated nsp2TF, or any potential ribosome sink function.

In this thesis, only the KO2 mutant was used. The growth kinetics of this mutant, or functionally equivalent mutants, have been assessed in three different background strains, between which the level of impairment varies280,536 (Figure 3.1C-E). This may indicate a variable dependence on PRF at this site for different strains or may be reflective of differences in assay conditions such as MOI. For the KO2 mutations on the SD95-21 background, which was used in this study, the viral titre was approximately 1-log lower than WT at 12 hpi, and peak titre was 0.5-log lower650 (Figure 3.1C).

Before the analysis of PRF is presented in this chapter, the datasets are introduced, including analysis of data quality (section 3.3). Next, an overview of the patterns of viral gene expression is given (section 3.4), to gain insight into the progress of infection at each timepoint and provide the context for interpretation of subsequent analyses. Finally, PRF at both sites on the viral genome is analysed, including quantification of PRF efficiency and analysis of putative ribosomal pauses at the frameshift sites (section 3.5). Further analysis of these datasets, to quantify the expression of individual viral transcripts and ORFs, is given in Chapter 4.
3.2 Experimental set-up

3.2.1 Infections and library preparation

To comprehensively investigate PRRSV gene expression, a representative genome from each species of PRRSV was chosen. Cells were infected with each of these viruses and subjected to RNASeq and parallel ribosome profiling to analyse the transcriptome and translatome.

For EU PRRSV, the virus stocks used for profiling were based on the strain used in the Porcilis® vaccine (MSD Animal Health; GenBank accession KJ127878.1). The de novo assembled viral genome sequence from this study bears 148 mutations relative to the GenBank accession and is given in Appendix section 1. This isolate is phylogenetically clustered with the prototype EU PRRSV strain, Lelystad virus (NC_043487.1). It is genetically similar, but harbours a 222 nt deletion in the hypervariable region of nsp2 (between genomic coordinates 2215 and 2216) relative to the Lelystad virus. For NA PRRSV, the isolate SD95-21 (GenBank accession KC469618.1) was used, which is phylogenetically clustered with the prototype NA PRRSV reference strain, VR-2332. For the KO2 mutant virus, mutations were made on the SD95-21 background as detailed in Figure 3.1A$^{280-282,536}$. These viruses are referred to as “EU” (European: PRRSV-1) and “NA” (North American: PRRSV-2) for ease of reference; however, as clarified in the introduction, these monikers reflect the historical epicentres of viral emergence and are not a comprehensive representation of the current geographies of the two species.

The interpretability of deep sequencing data obtained from virus-infected cells is improved by using genetically homogenous host cells and ensuring that a high proportion (> 90%) of cells are successfully infected. Additionally, for standard ribosome profiling protocols, a relatively large amount of starting material is required, in our case approximately $10^7$ cells per replicate. These restrictions rendered the use of PAMs impractical as they are time-consuming and costly to isolate, suffer from heterogeneity and batch-to-batch variation, and permissiveness to PRRSV infection varies from 15 to 60%$^{479}$. Therefore, MA-104 cells (Chlorocebus sabaeus) and the derivative MARC-145 cell line were used for profiling as these are widely used to study PRRSV and are highly permissive to infection$^{475}$. All infections were performed at high multiplicity of infection (MOI) to facilitate high infection efficiency: MOI 5 for NA PRRSV, MOI ~1-3 for EU PRRSV.

Three sets of infections were performed to generate the PRRSV datasets described in this study (Table 2). To generate the main dataset (referred to as the NA PRRSV noCHX dataset), MARC-145 cells were infected with WT or KO2 NA PRRSV, or mock-infected, and harvested at 3, 6, 9 and 12 hpi by flash-freezing without CHX pre-treatment. I went to Prof. Ying Fang’s laboratory at Kansas State University to learn the relevant techniques, grow up the virus and harvest the samples for ribosome profiling. However, the infections I carried out for this experiment did not reach high
enough infection efficiency (as judged by immunofluorescence) to use for ribosome profiling. Therefore, they were repeated by Dr Pengcheng Shang after my departure to generate the samples that were used to derive the sequencing libraries presented in this thesis. For the second PRRSV dataset (referred to as the EU PRRSV dataset), MA-104 cells were infected with WT EU PRRSV and harvested at 8 hpi using a protocol which included a CHX pre-treatment step and lysed the cells without flash-freezing. These experiments were performed by Prof. Ian Brierley and Dr Adrian Mockett before I joined the laboratory. For the third dataset (consisting of a single library, referred to as the NA PRRSV CHX dataset), MARC-145 cells were infected with WT NA PRRSV and harvested at 9 hpi after CHX pre-treatment by flash-freezing. This experiment was carried out by Dr Pengcheng Shang and Dr Yanhua Li (members of Prof. Ying Fang’s laboratory) before I began my PhD. Except for the discussion of data quality that follows in section 3.3.3, the results of this third dataset will not be discussed until section 4.3.2.

Timepoints were selected to cover a single round of viral replication. PRRSV non-structural proteins first become detectable by immunofluorescence at 6 hpi\(^{551}\), visible as a faint band in a western blot of nsp1β in the cell lysates used for NA PRRSV profiling (Figure 3.2A). Subsequently, levels increase substantially, and the nsp1β band is clearly visible at both 9 and 12 hpi ((Figure 3.2A). In line with results for NA PRRSV, EU PRRSV nsp1β is clearly visible by western blot from 8 hpi onwards ((Figure 3.2B, western blot performed by Lior Soday). The ribosome profiling timecourse ended at 12 hpi, when virion formation and release begin to occur\(^{551}\).

Libraries were prepared for deep sequencing as described in the Materials and Methods chapter. For both NA PRRSV datasets, libraries were prepared using adapters with seven randomised nucleotides at the end ligated to the biological material. This is expected to reduce both 5’ and 3’ ligation bias (based on previous studies\(^{327,359,363,366-368}\), and enables computational removal of reads which arise

![Figure 3.2 Western blots of profiling lysates](image_url)

**Figure 3.2 Western blots of profiling lysates**

**A)** Western blot of nsp1β expression in lysates used for NA PRRSV ribosome profiling (noCHX dataset, replicate one samples). Antibodies were against viral protein nsp1β (23 kDa) and cellular protein GAPDH (36kDa) as a loading control. **B)** Western blot (performed by Lior Soday) of nsp1β expression in MA-104 cells infected with EU PRRSV, harvested over a timecourse of 8-25 hpi. The 8 hpi timepoint was selected for EU PRRSV ribosome profiling.
from PCR duplication as opposed to the presence of multiple identical fragments of biological material (section 1.3.3.1). For the EU PRRSV dataset, replicate one libraries were prepared using adapters with no randomised nucleotides, and replicate two libraries were prepared using a 5’ adapter with no randomised nucleotides and a 3’ adapter with 14 randomised nucleotides at the 5’ end. It was, therefore, only possible to computationally remove PCR duplicates from replicate two libraries from the EU dataset.

There were some additional differences in the experimental procedures used to generate the libraries, in that RNA fragments of different lengths were purified during the size selection step (Table 2). For the EU dataset (library preparation performed by Prof. Ian Brierley and Lior Soday) and the NA PRRSV CHX-pre-treated dataset, 25-34 nt fragments were selected for all libraries (RiboSeq and RNASeq). For the NA PRRSV noCHX dataset, 9 hpi replicate one libraries (both RiboSeq and RNASeq) also had a 25-34 nt gel slice purified. Subsequently, the protocol for the remaining libraries was modified to facilitate inclusion of the shorter population of RPFs in RiboSeq libraries (section 1.3.2.2), and to generate longer reads for RNASeq libraries to aid accurate mapping. As such, a gel slice centred at around 50 nt was purified for all other NA PRRSV noCHX RNASeq libraries. For all other NA PRRSV noCHX RiboSeq libraries, a gel slice of 19-34 nt was purified, except for 9 hpi replicate two libraries, for which a 19-80 nt “broad spectrum” gel slice was purified. Fragments longer than 34 nt were selected for these libraries to investigate the potential presence of RNA fragments protected by species other than monosomes. These longer reads make up only a small proportion of these libraries, and are specifically selected for analysis and discussion in section 3.5.5, while the whole libraries are used for analyses in all other sections (unless specified).

3.2.2 Aligning reads to host and viral genomes

Universal adapter sequences were removed from reads and, where possible (all PRRSV libraries except EU PRRSV replicate one), PCR duplicates were computationally removed, before removal of any randomised bases originating from adapter sequences. Bowtie (version 1.2.3; reference 645) was used to map reads to the following databases, with reads that failed to align to one database used as input for mapping to the next: rRNA, vRNA, mRNA, ncRNA, gDNA, and contaminants. Bowtie version 1 is not a “splice-aware” mapper and does not permit gaps, deletions, insertions, or junctions in the alignment between the read and the reference. Unless stated (e.g. section 4.2) all plots and analyses in this chapter were carried out using reads aligned by bowtie, and therefore represent reads which continuously align to the reference. For all RiboSeq libraries, one mismatch between the read and the reference was permitted, while two mismatches were permitted for RNASeq libraries (except for mapping to the contaminants database, for which no mismatches were allowed for any libraries). The best alignment (defined by the number of mismatches) was reported and, where multiple alignments were equally good, one was selected at random.
To sequence the longer fragments of RNA in the NA PRRSV noCHX 9 hpi replicate two RiboSeq libraries, paired-end sequencing was used. Merging of the pairs of reads into a single sequence, and removal of the universal adapter, was performed using LeeHom\textsuperscript{644}, a tool designed for processing reads originating from fragments of ancient DNA which have a similar range of fragment lengths to these libraries. Reads were then processed using the pipeline described above, from PCR de-duplication onwards.
3.3 Quality control analysis

Several analyses were performed to verify that the quality of the data was high and that a large proportion of CDS-mapping RiboSeq reads represent genuine RPFs.

3.3.1 Quality control analysis of the NA PRRSV noCHX dataset

Note that some aspects of quality control can be affected by minor, unintended variations in library preparation, particularly when comparing libraries processed in different batches. For profiling of virus-infected cells, processing batches are organised to avoid potential cross-contamination. This is a particular concern because very small amounts of contamination, which would not be a problem for typical profiling experiments, can have a greater effect due to the orders of magnitude difference in the amount of vRNA between mock/early timepoint libraries and libraries from late timepoints of infection. Similarly, RNASeq can have a far greater proportion of vRNA than RiboSeq, so RiboSeq libraries were always processed separately from RNASeq libraries. Within these categories, all 6 hpi libraries were processed as one batch, as were all 12 hpi libraries. For 9 hpi libraries, replicates one and two were processed in separate batches, and replicates three and four were processed as one batch. As discussed in section 1.3.3.1.5, even if careful efforts are made to avoid contamination during library preparation, contamination is thought to be able to occur on the flow cell during sequencing. Additionally, the amount of RNase I added to each batch was refined to improve phasing in some cases (Materials and Methods, Table 3).

3.3.1.1 Library composition

Figure 3.3 shows the proportion of each library that maps to each database and the proportion discarded before mapping due to, for example, being too short. Despite the relatively poor genome annotation of C. sabaeus, the majority of reads that passed the initial filters aligned to the reference databases. RiboSeq libraries very frequently have a high proportion of rRNA, as observed here, despite including steps in the protocol to remove it. This is likely due to the Ribo-Zero rRNA depletion kit being optimised for removal of the roughly uniform distribution of rRNA fragments expected from RNASeq libraries, as opposed to the biased rRNA fragment composition arising from the RiboSeq protocol (see section 1.3.3.1.5). Correspondingly, the proportion of rRNA is much lower in the RNASeq libraries. The large variation in the proportion of the RiboSeq libraries that is designated “too-short” is likely due to variations in the lower boundary of the gel slice purified, particularly for libraries where the protocol was adjusted to purify shorter fragments as this was not previously common practice in the laboratory. Note that reads mapping to the gDNA database are not thought to arise from gDNA itself, as this is degraded by DNase in the lysis buffer,
but rather from unannotated transcripts. As these can originate from either strand, there is an approximately equal proportion of positive and negative-sense reads in the gDNA fraction of the libraries. Similarly, the relatively high number of negative-sense reads in the mRNA fraction of the RNASeq libraries likely arises from mis-annotated transcripts or mis-mapping reads, with the majority of these reads originating from just a small number of annotated transcripts. The 9 hpi RiboSeq mock replicate three library was not good quality and was discarded. As 9 hpi replicates three and four were processed concomitantly, 9 hpi mock replicate four provides the matched uninfected control for this batch.

3.3.1.2 Read length distribution

Translating ribosomes are known to protect two distinct lengths of mRNA from nuclease digestion: 28-30 nt, thought to originate from a ribosome with an occupied A site\textsuperscript{298,299,303}, and 20-22 nt, thought to originate from a ribosome with an empty A site\textsuperscript{140,331} (section 1.3.2.2). Metagene analysis of RiboSeq reads mapping to host and viral coding sequences revealed a read length distribution with two distinct peaks at the expected lengths for these two populations of RPFs (Figure 3.4, left). Interestingly, infected RiboSeq samples from late timepoints seem to have a slightly lower proportion of short reads than their mock-infected or early-timepoint counterparts. In most cases, the length distribution of the viral reads (green) closely matches that of the host reads (grey). For some libraries at 9 and 12 hpi the viral read length distribution is slightly flattened relative to the host, which could indicate a small proportion of viral reads originate from non-RPF sources, such as RNPs, which do not have the same bimodal length distribution. This is examined further below.

For the 3 hpi RiboSeq libraries the virus length distribution matches the host poorly in most cases and does not form the characteristic RPF-like length distribution. The length distribution of RiboSeq 3 hpi WT 2 is reasonably similar to the host length distribution, but this library (like the other RiboSeq 3 hpi libraries) suffers from low coverage, with only 205 reads mapping to the viral genome – most of these aligning to ORF1a. The viral reads in these libraries likely originate from a mixture of genuine RPFs and contaminants such as RNPs, but the combination of low viral read depth and poor signal to noise ratio would make robust analysis of these libraries difficult, and they are excluded from all analyses except for a select few in section 3.4 (where they provide an upper bound).

The length distribution of RNASeq reads (Figure 3.4, right) is not determined by the position of translating ribosomes but instead by alkaline hydrolysis and the range of the gel slice excised during RNA fragment purification. As such, the RNASeq libraries have much broader length distributions, and the viral distribution matches the host very closely. The shoulder on the length distribution of RNASeq 12 hpi WT 1 may be due to an irregularity in the shape of the gel slice.
Figure 3.3 Composition of libraries

RiboSeq reads were designated as “too-short” if the insert was shorter than 25 nt (9 hpi replicate one libraries) or 19 nt (all other libraries). RNASeq reads were designated as “too-short” if the insert was shorter than 25 nt (all libraries). For 9 hpi replicate two RiboSeq libraries, pairs of reads which LeeHom failed to merge were put in the “non-clipped” category for the purposes of this plot. RiboSeq 9 hpi mock replicate 3 was not a good quality library and was discarded.
Figure 3.4 Length distribution of CDS-mapping reads

Length distribution of positive-sense RiboSeq (left) and RNASeq (right) reads mapping within host (grey) or viral (green, mock excluded) CDSs. For 9 hpi replicate 1 samples (RiboSeq and RNASeq), fragments of 25-34 nt were size-selected during the library preparation; for all other samples the minimum length selected for was 19 nt for RiboSeq and ~45 nt for RNASeq.

3.3.1.3 Distribution of reads relative to start and stop codons

In mammalian cell lysates treated with enough nuclease that all unprotected regions of mRNA are fully digested (“trimmed”), there is a distance of ~12 nt between the 5’ end of the RPF and the first nucleotide in the P site of the ribosome\(^{298,305}\) (section 1.3.1.3). This is confirmed in the noCHX NA PRRSV dataset by a metagene analysis in which reads are aligned relative to the start and stop codons of their annotated transcripts, and the 5’ end positions are plotted (Figure 3.5, left). The magnified panel underlaid beneath the plot reveals a 12 nt distance between the RPF 5’ ends and the start codon (positioned in the P site), and a 15 nt distance between the RPF 5’ ends and the stop codon (positioned in the A site). This confirms that applying a +12 nt offset to RPF 5’ end coordinates is a reasonable approximation of P site position, and this will be used for this dataset throughout this study. Although inference of P site position is not applicable to RNASeq reads, the same +12 nt offset is applied in most cases for comparability.
Figure 3.5 Average distribution of 5’ ends of host mRNA-mapping reads relative to start and stop codons

Only transcripts with an annotated CDS of at least 150 codons, 5’ UTR of at least 60 nt, and 3’ UTR of at least 90 nt were included in the analysis. The total number of positive-sense reads from all these transcripts mapping to each position was plotted. Underneath the 12 hpi RiboSeq mock replicate two library, a magnified view of the 45 nt around the start and stop codons for that library is shown.
CDS-mapping reads display a clear triplet periodicity with most reads mapping to the first nt of the codon (phase 0). There is a short region of low density with density steadily increasing over the first ~100 nt of the CDS, that may indicate the occurrence of run-off translation occurring during harvesting\textsuperscript{298,305,313,331}. The prominent termination peaks in this meta-analysis are characteristic of samples harvested by snap-freezing without CHX pre-treatment\textsuperscript{298,313}. Approximately ten codons upstream of the termination peak, a small increase in density may indicate ribosomes queuing behind the terminating ribosome\textsuperscript{186,202,336,338}. There are few RPFs mapping to the UTRs, as these are usually untranslated. RPF density in the 5′ UTR is higher than that in the 3′ UTR due to the presence of uORFs\textsuperscript{298,305,313}, although these are not always in the same frame relative to the annotated CDS, so the phasing averages out.

The metagene analysis for RNASeq libraries (Figure 3.5, right) does not show significant triplet periodicity, because the 5′ end positions of RNASeq reads are not determined by the step-wise movement of the ribosome. The slight bias within the CDS towards phase 0 could indicate minor contamination with RiboSeq libraries, or could be due to the combined effect of technical biases (section 1.3.3.1) and a non-uniform nucleotide distribution resulting from codon usage biases, such as a preference for RNY codons\textsuperscript{307,652}. Despite the inclusion of randomised bases on the adapters, some compounded ligation bias is evident for reads with their 5′ end at the start codon, which all share the sequence 5′-AUG, or their 3′ end at the stop codon, which all share the sequence URR-3′. The slight drop in density in the 3′ UTR is likely due to transcripts which have truncated 3′ UTRs compared to their annotation.

### 3.3.1.4 Phase composition

Analysis of the proportion of host CDS-mapping reads attributed to each phase (Figure 3.6) confirms the phase preferences observed in Figure 3.5. For all RiboSeq libraries, the majority of RPFs map to phase 0, indicating good quality, welltrimmed data in which a large proportion of CDS-mapping reads represent genuine RPFs. The phase composition of virus CDS-mapping reads matches the host composition well at 6 hpi, however some 9 and 12 hpi libraries have a lower proportion of phase 0 reads in the virus fraction than the host. This may be indicative of RNP formation on the viral genome, discussed further below. The 3 hpi RiboSeq libraries are included here to provide a second method of assessing the proportion of virus-mapping reads that are genuine RPFs in these libraries. The phasing of virus-mapping RiboSeq reads at 3 hpi is generally poor, with the exception of WT replicate two. This supports the conclusions drawn based on the length distributions (Figure 3.4, left), confirming that a significant proportion of virus-mapping reads in the 3 hpi RiboSeq libraries likely originate from sources other than translating ribosomes.
Analysis of programmed ribosomal frameshifting on the PRRSV genome

![Graph showing phase composition of CDS-mapping reads](image)

**Figure 3.6 Phase composition of CDS-mapping reads**

**Upper:** Proportion of RiboSeq reads (all read lengths) attributed to each phase, from positive-sense reads mapping within host (top) or viral (bottom, mock excluded) CDSs. Overlapping regions of viral CDS were excluded. **Lower:** Analysis from upper panel carried out on RNASeq reads. These are not expected to show strong phasing as the 5' end position is determined by alkaline hydrolysis, not the position of a translating ribosome.

### 3.3.1.5 Assessment of potential RNP contamination in RiboSeq libraries

The length distribution and phasing of virus-mapping reads (Figure 3.4 and Figure 3.6) suggested there may be some contamination of this fraction of some RiboSeq libraries with reads originating from sources other than protection by translating ribosomes. There are various possible sources of this contamination; however, as viruses often produce large amounts of RNA-binding proteins, and the contamination disproportionately affects virus-mapping reads, viral RNPs which are heavy enough to co-sediment with ribosomes seem a likely source. Therefore, this is henceforth referred to as “RNP contamination”, although I acknowledge it may stem from other sources, and also note that reads originating from RNP complexes would represent genuine biological material from the infected cells as opposed to external “contamination” with foreign genetic material. The potential RNP contamination of the RiboSeq libraries was explored further by separating the viral genome into specific regions and assessing length distribution and phasing in each, as the potential background of RNP contamination should be more apparent in regions which are less translated. For
RNASeq libraries, RNP contamination is not a relevant concern, as protein is enzymatically digested before the RNA is first purified.

A flatter length distribution of virus-mapping RPFs compared to the host length distribution may indicate a noticeable proportion of reads in that region of the viral genome originate from RNPs (Figure 3.7). In addition, separating the reads of each length according to phase (Figure 3.8) can allow a more detailed characterisation of the library-specific patterns expected from genuine RPFs (host distribution) and, by comparison, allow identification of potential RNP contamination.

**Figure 3.7 Length distribution of RPFs mapping to specified regions of the viral genome**

Only positive-sense RiboSeq reads were used. Reads mapping to ORF1a are shown in purple, ORF1b in blue and sgRNA ORFs in red. The length distribution of host CDS-mapping RPFs from Figure 3.4 is reproduced in grey for comparison. Note that read counts for each length were normalised by the total number of reads in the region to make the scales comparable between regions. The coordinates used to define each region of the viral genome can be found in Appendix section 3.
Analysis of programmed ribosomal frameshifting on the PRRSV genome

Figure 3.8 Number of reads of each length that are attributed to each phase, for RPFs mapping to specified regions of the viral genome

Only positive-sense RiboSeq reads were used, and regions from Figure 3.7 were edited to exclude overlapping CDSs (coordinates given in Appendix section 3). Phase 0 is designated independently for each region, relative to the first nucleotide of that region’s ORF, to aid comparison (e.g. the dominant phase in ORF1b is designated 0 instead of −1). The host phasing information from Figure 3.6 was stratified by read length and reproduced for comparison. Note that y axis scaling was set for each library and region separately to facilitate comparison. Read lengths which show the best signal to noise ratio in the ORF1b region were determined for each library individually, according to similarity to that sample’s host phasing
Quality control analysis

distribution. Read lengths selected for inclusion in analyses of these libraries are indicated by square brackets.

At 6 hpi the difference in length distribution and phasing of sgORF-mapping RPFs and host RPFs is pronounced. However, this is unsurprising, as sgORF transcription and translation is minimal at this timepoint (discussed in section 3.4). As such, very few reads formed the input for the sgORF length distribution (156-293 reads) and N phase composition (34-49 reads) at 6 hpi, and these are likely highly subject to noise. At later timepoints, the sgORF plots closely match the host distributions, indicating that, relative to actual RPF levels, this region does not suffer significantly from RNP contamination. ORF1b is not highly translated, by virtue of being expressed only from the genomic RNA, and being positioned after two PRF sites which reduce the number of ribosomes that reach ORF1b. RNP contamination is more apparent in this region of the genome, evidenced by the flattening of the length distribution and the deterioration of the phasing patterns observed relative to the host fraction. ORF1a is more highly translated than ORF1b and the fraction of reads in this region which likely originate from RNPs generally appears to be lower. This distinction is less clear at 12 hpi, for which the length distributions of ORF1a- and ORF1b-mapping RPFs is similar. This may be partly the result of increased translation of ORF1b at this timepoint (discussed in sections 3.5.3 and 4.4.2) meaning the level of genuine translation in ORF1b is similar to that in ORF1a, making the ratios of RPF:RNP reads in the two regions more similar. Additionally, the deviation of both of these length distributions from that of the host at 12 hpi may indicate an increase in RNP contamination across ORF1ab, potentially resulting from genome packaging at this timepoint.

The phase composition plots (Figure 3.8) give an indication of which read lengths have the highest proportion of genuine RPFs for each library. The characteristic length ranges protected by the mammalian ribosome are known (20-22 nt and 28-30 nt\textsuperscript{140,305,331}). Inspection of length distribution plots Figure 3.4 further narrows down these length ranges and reveals that the most abundant read lengths tend to be 21, 28 and 29 nt for this dataset. These read lengths are expected to have the highest signal to noise ratio. The phase composition of these read lengths was assessed in the ORF1b region, the region with the most visible RNP contamination, and those for which the signal to noise ratio was deemed acceptable (based on comparison with host phasing) were selected for each library, indicated by square brackets. For all analyses of noCHX NA PRRSV RiboSeq data presented in section 3.4 and thereon, only the indicated read lengths were selected for inclusion in the analyses (unless specified).

RNP contamination caused by viral RNA-binding proteins (or host RNA-binding proteins stimulated by infection) often affects virus-mapping reads more than host reads\textsuperscript{307,375}, due to factors such as compartmentalisation of viral transcription and replication, or the presence of specific binding/nucleation sites on the viral genome. To test whether host RPFs are affected by RNP contamination, the length distribution and density of RiboSeq reads mapping to the CDSs was
compared to that of reads mapping to the 3’ UTRs (Figure 3.9). RNP contamination is expected to affect both the CDSs and the UTRs but would be more visible in the UTRs due to the lower density of genuine RPFs. In this analysis, 3’ UTR-mapping reads in libraries with no RNP contamination likely originate from alternative transcripts for which an identical region is annotated as CDS for one isoform and 3’ UTR for another, as multi-mapping reads are randomly assigned in such cases. Additionally, a small proportion of ribosomes are not captured by the ribosome recycling pathway at the termination codon, and they can enter the 3’ UTR\textsuperscript{202,359}. As such, the uncontaminated 3’ UTR-mapping read length distribution is expected to match the CDS-mapping read length distribution, although the read density in the 3’ UTR should be much lower. Where RNP contamination is present, this manifests as increased read density in the 3’ UTR, and a read length distribution that differs from that in the CDS.

![Graph](image)

**Figure 3.9 Assessment of potential RNP contamination of host-mapping RiboSeq reads**

Length distribution (line graph) and relative density (stacked bar chart) of RiboSeq reads mapping to host CDSs (grey) compared to host 3’ UTRs (red). Only host mRNAs with a CDS of at least 150 codons and a 3’ UTR of at least 100 codons were included in the analysis. Length distribution of positive-sense RPFs with inferred P sites in the CDS (codons −100 to −10 relative to the stop codon) and 3’ UTR (codons +10 to +100 relative to the stop codon) was determined and normalised by total number of reads in that category to make the scales comparable. To the right of each length distribution, a stacked bar chart indicates the relative density of reads in each region (note: 3’ UTR in red is a minor component in this chart).

For most libraries, the length distribution of 3’ UTR-mapping reads aligns reasonably well with the CDS distribution (Figure 3.9, line graphs). Differences between the two distributions, such as
reduced height of the 28-30 nt peak and slightly increased density between the two main peaks, are present in the matched mock libraries as well. This indicates these differences are unrelated to infection and perhaps represent formation of RNPs comprised of host proteins and mRNA. For 12 hpi WT 2 and 9 hpi WT 3 and KO2 4, however, the deviation of the 3’ UTR length distribution from the CDS distribution is greater than the deviation for the matched mock libraries, indicating there may be some contribution from viral proteins. For all libraries, the relative read density in the 3’ UTR is much lower than the CDS read density (Figure 3.9, stacked bar charts), indicating that any potential RNP contamination only represents a small proportion of CDS-mapping reads (except potentially for very lowly translated mRNAs). Therefore, RNP contamination is unlikely to significantly confound analysis of host RPFs.

Together, these analyses indicate that some RiboSeq libraries are affected by RNP contamination, however this is generally specific to the viral genome, and only noticeably affects regions which are poorly translated – particularly ORF1b. Read length stratification will likely ameliorate the effects of this in most cases, but for analyses focused on ORF1b it may affect the results of some libraries, particularly 9 hpi WT 3, 9 hpi KO2 2, 9 hpi KO2 4 and 12 hpi WT 2. For the 9 hpi libraries amongst these, 21 nt RPFs were excluded from the selected read lengths, so the population of short RPFs will not be represented in these libraries for most analyses.

3.3.2 Quality control analysis of the EU PRRSV dataset

The quality control analyses described above (section 3.3.1) were carried out for the EU PRRSV dataset (Figure 3.10).

The RiboSeq replicate one library has a large proportion of rRNA, and the replicate two library has a large proportion of reads designated as too short (Figure 3.10A). This reduces the fraction of the libraries that can be usefully analysed, however the total number of reads in each library was large, and enough reads mapped to the viral genome for further analysis (251,087 and 102,427 reads for replicates one and two, respectively).

The metagene analysis of read 5’ end positions relative to annotated host start and stop codons (Figure 3.10B) recapitulates the features described in section 3.3.1.3, except in this case the initiation peak is larger and the termination peak is smaller. This is characteristic of CHX pre-treated samples. This analysis confirms that adding a +12 nt offset to the 5’ end coordinate of RPFs is an appropriate approximation of P site position for this dataset. The trough in read density four codons upstream of the termination peak is less apparent for replicate two libraries, which were prepared using a 3’ adapter with 14 randomised bases at the 5’ end. This indicates the randomised bases have reduced the effect of ligation bias compounded by the common 3’-terminal URR shared by reads at this position.
Figure 3.10 Quality control for EU PRRSV dataset

A) Composition of libraries. Reads were designated as “too-short” if the insert was shorter than 25 nt. B) Average distribution of 5’ ends of host mRNA-mapping reads relative to start and stop codons. Plot constructed as in Figure 3.5. C) Length distribution of positive-sense host (grey) and virus (green) reads mapping within CDSs. Fragments of 25-34 nt were size-selected during the library preparation for all samples. D) Length distribution of positive-sense RiboSeq reads with 5′ ends mapping to specified regions of the viral genome: ORF1a (purple), ORF1b (blue), sgRNA ORFs (red). The length distribution of host CDS-mapping RPFs from C is reproduced for comparison (grey). The coordinates used to define each region of the viral genome for this and panel E can be found in Appendix section 3. E) Number of reads of each length that are attributed to each phase, for reads with 5′ ends mapping to specified regions of the viral genome. Plot constructed as in Figure 3.8. RNP contamination in these libraries is minimal, but to aid visualisation of phasing, read lengths for which at least ⅔ host CDS-
mapping reads are in phase 0 were selected for some (specified) plots and analyses. These read lengths are indicated by square brackets. F) Length distribution (line graph) and relative density (stacked bar chart) of reads mapping to host CDSs (grey) compared to host 3′ UTRs (red). Plot constructed as in Figure 3.9. G) Proportion of host and viral reads attributed to each phase. Plot constructed as in Figure 3.6.

The length distribution of host CDS-mapping RPFs demonstrates the characteristic sharp peak, although shifted slightly upwards, with a maximum at 30 nt (Figure 3.10C). The lower bound of the gel slice purified for these libraries was 25 nt, and the short RPF population is absent. The virus RPF length distribution is very similar to the host, a similarity which is maintained for all selected regions of the viral genome, indicating no significant contamination by viral RNPs (Figure 3.10D). This conclusion is supported by analysis of the phase composition of reads stratified by length, for which ORF1a-, ORF1b- and N-mapping reads closely match the host phase composition (Figure 3.10E). Comparison of RPFs mapping to host CDSs and 3′ UTRs reveals differences in the length distributions for longer reads (>32 nt) (Figure 3.10F). Although there are no mock libraries with which to compare, these differences are unlikely to be due to viral RNPs given the lack of visible contamination on the viral genome itself. Further, the read density in the 3′ UTR is much lower than the density in the CDS, indicating RNP contamination in host CDSs is minimal.

The overall phasing in the Riboseq libraries is good (Figure 3.10G). The slightly increased read length (Figure 3.10C) and the prominence of phase 2 in the phase composition of read lengths above 29 nt (Figure 3.10E, host panel) suggest that these libraries are marginally under-trimmed. To aid analyses involving visualisation of changes in phase or attributing reads to overlapping ORFs, read lengths with at least ⅔ of host CDS-mapping RPFs in phase 0 were selected for some analyses. These read lengths are indicated by square brackets in Figure 3.10E (host panel).

Overall, the quality of this dataset is high, and libraries are not significantly affected by RNP contamination.

### 3.3.3 Quality control analysis of the NA PRRSV CHX dataset

This dataset consists of a single Riboseq library, harvested after CHX pre-treatment at 9 hpi. The quality of the data is high (Figure 3.11). A large proportion of the library aligns to the host transcriptome or viral genome, giving good read depth for future analyses (Figure 3.11A). A large majority of reads map to the 0 phase (Figure 3.11B), and the length distribution has a sharp peak at 28-29 nt (Figure 3.11C), indicative of well-trimmed data (note that short RPFs were not included in the gel slice for this library). The metagene analysis of read 5′ end distribution relative to host start and stop codons gives the pattern characteristic of CHX pre-treated samples and confirms a +12 nt offset between RPF 5′ ends and the ribosomal P site (Figure 3.11D).
Figure 3.11 Quality control for NA PRRSV CHX dataset

A) Composition of libraries. Reads were designated as “too-short” if the insert was shorter than 25 nt. B) Proportion of host and viral reads attributed to each phase. Plot constructed as in Figure 3.6. C) Length distribution of positive-sense host (grey) and virus (green) reads mapping within CDSs. Fragments of 25-34 nt were size-selected during the library preparation. D) Average distribution of 5’ ends of host mRNA-mapping reads relative to start and stop codons. Plot constructed as in Figure 3.5. E) Length distribution of positive-sense RiboSeq reads with 5’ ends mapping to specified regions of the viral genome (coordinates in Appendix section 3). The length distribution of host CDS-mapping RPFs from C is reproduced for comparison (grey). F) Number of reads of each length that are attributed to each phase, for reads with 5’ ends mapping to specified regions of the viral genome. Plot constructed as in Figure 3.8. Read lengths selected as showing minimal effect of RNP contamination are indicated by square brackets, and are used in all plots and analyses for this library. G) Length...
distribution (line graph) and relative density (stacked bar chart) of reads mapping to host CDSs (grey) compared to host 3’ UTRs (red). Plot constructed as in Figure 3.9.

The viral genome does, however, show evidence of low level RNP contamination in ORF1b. The length distribution of virus-mapping reads is slightly flattened relative to the host (Figure 3.11C), though this effect seems to be largely confined to the ORF1b region of the viral genome (Figure 3.11E). This is confirmed by the analysis of phase composition of reads stratified by length, in which ORF1a and N are similar to the host distribution (Figure 3.11F). Reads of 28 and 29 nt length were identified as having a high signal to noise ratio and were selected for inclusion in subsequent analyses of this library. There are some small differences between the length distribution of host CDS- and 3’ UTR-mapping reads (Figure 3.11G, line graph), but these are of a similar magnitude to those observed for mock-infected samples in the NA PRRSV noCHX dataset (Figure 3.9). Additionally, the read density in the 3’ UTR is very low compared to the CDS density (Figure 3.11G, stacked bar chart), so RNP contamination is unlikely to significantly affect host RPFs.
3.4 An overview of viral transcription and translation

The distribution of RNASeq and RiboSeq reads on the viral genome was analysed over the course of NA PRRSV infection to observe changes and trends in viral transcription and translation. In order to better observe trends in the data, some figures throughout this chapter present only one library from each condition (for which the quality control plots indicated the highest quality data), with the remaining replicates presented in a separate figure.

To visualise changes in transcription and translation, read density at each position on the viral genome was plotted (Figure 3.12A-C, Figure 3.13 and Figure 3.14). The progression of the virus through the replication cycle was assessed by calculating the ratio of vRNA-mapping reads to positive-sense host mRNA- and vRNA-mapping reads (Figure 3.12D and E). The relative production/translation of gRNA compared to sgRNAs was quantified using the ratio of read density in the sgRNA region of the viral genome to the density in the ORF1a region (Figure 3.12F).

At 3 hpi viral genome replication has likely not yet begun, due to a lack of the necessary viral proteins. The distribution of RNASeq reads across the genome is roughly uniform (Figure 3.12B, Figure 3.12F, Figure 3.13), and there are very few negative-sense reads, indicating these data represent the input genomes. The quality control data in Figure 3.4 and Figure 3.6 (sections 3.3.1.2 and 3.3.1.4) suggest that, at this timepoint, translation of the viral genome is just beginning to reach the level detectable by RiboSeq. This is consistent with the fact that time is required for attachment, internalisation, and uncoating to occur before the genome is released into the cytoplasm\textsuperscript{453,455,653}, and with observations that viral proteins do not generally accumulate to levels detectable by immunofluorescence until 6 hpi\textsuperscript{551,654}.

An early stage of viral genome replication, the production of a negative-sense replication intermediate, is evident at 6 hpi from the presence of negative-sense RNASeq reads mapping uniformly across the genome (Figure 3.12B, Figure 3.13). The total levels of positive-sense vRNA remain similar to those seen at 3 hpi, indicating the replication intermediate is not yet being converted into positive-sense progeny genomes to a noticeable level (Figure 3.12D and E). The production of sgRNA templates is also not yet evident, as the ratio of RNASeq density in the sgRNA region compared to ORF1a remains approximately one (Figure 3.12F). Translation of the viral genome is clearly detectable (Figure 3.12C, Figure 3.14), and most viral reads map to the ORF1ab polyprotein (Figure 3.12F; mean sgRNA:ORF1a ratio 0.08), consistent with the fact that gRNA is the predominant species of viral RNA at this timepoint.

At 9 hpi, massive production of both gRNA and sgRNA has occurred and viral translation represents a considerable proportion of the ongoing translation in the cell (Figure 3.12B-E, Figure 3.13, Figure 3.14). These results are highly consistent with Figure 3.2A, in which nsp1β is readily detectable by
An overview of viral transcription and translation

western blot at 9 hpi, and other studies which show strong staining of viral proteins, including RTCs, by immunofluorescence, and which observe replication of viral RNA at this timepoint\textsuperscript{512,551,654}. Both positive- and negative-sense viral transcripts are increased compared to 6 hpi and the ratio of sgRNA to gRNA production is increased to approximately 5:1 (Figure 3.12F). This is visible on the genome plot as the majority of RNASEq reads map to the 3′-proximal region of the genome or the 5′ UTR, which is present on all sgRNAs as well as the gRNA (Figure 3.12B). Translation of both ORF1ab and the sgORFs is evident, with approximately seven times as much ribosome density on the sgORFs as on the polyprotein (Figure 3.12F).

Positive-sense vRNA continues to accumulate between 9 and 12 hpi, although accumulation of the negative-sense counterpart appears to reach a plateau (Figure 3.12D and E). The ratio of positive-sense transcripts of sgRNA compared to gRNA remains roughly constant, but the translation ratio skews further in favour of sgRNA translation (Figure 3.12F). These changes likely represent a move towards packaging of the viral genome, for which the main components required are structural proteins, expressed from the sgRNAs, and positive-sense gRNA.

At all timepoints, a large RiboSeq peak in the 5′ UTR is seen (Figure 3.12C, Figure 3.14), which results from translation of novel uORF, discussed in section 4.3. There are no obvious differences between patterns of translation on the WT and KO2 genome, except for the occurrence of frameshifting at the nsp2 PRF site, which is evident on the WT genome at all timepoints but not the KO2 genome. PRF, and the accompanying ribosomal drop-off, is evident at the ORF1ab site for both viruses. Frameshifting at both sites will be discussed in detail in the next section.

The EU PRRSV-infected samples were harvested at 8 hpi, to capture a timepoint at which RTCs are established and viral transcription and translation is robustly measurable. The overall transcription and translation at this timepoint in the EU samples has a similar profile to the 9 hpi NA PRRSV libraries (Figure 3.15A-E).

In all RiboSeq libraries, a small proportion of the reads that map to the viral genome were in the negative sense (Figure 3.15D). There are no known translated ORFs encoded in the reverse sense on the PRRSV genome, and these reads do not display the characteristic length distribution or phase composition of RPFs (Figure 3.15G and H). Therefore, they likely originate from other sources such as RNP formation and are excluded from plots and analyses hereafter.

Taken together, these analyses provide an overview of the progress of viral infection throughout a 12 hour timecourse, highlighting a transition through several stages: translation of the viral genome beginning to reach the minimum threshold for detection by RiboSeq at ~3 hpi, production of negative-sense genome replication intermediates by 6 hpi, and extensive production and translation of viral gRNA and sgRNAs from 9 hpi onwards.
Figure 3.12 Overview of viral transcription and translation over a timecourse of NA PRRSV infection

A) Genome map of NA PRRSV (strain SD95-21, GenBank accession KC469618.1). ORFs are coloured and offset on the y axis according to their frame relative to ORF1a (0: purple, no offset; +1/−2: blue, above axis; +2/−1: yellow, below axis). Subgenomic RNAs are shown beneath the full length genomic RNA, with the region of 5′ UTR that is identical to the genomic 5′ UTR shown in grey. ORFs translated from each sgRNA are depicted as coloured boxes and named to the right.

B) RNASeq read densities in reads per million mapped reads (RPM) on the WT viral genome, after application of a 45 nt running mean filter, from cells harvested over a timecourse of 3-12 hpi. Positive-sense reads are plotted in green (above the horizontal axis), negative-sense in red (below the horizontal axis). Further replicates and KO2 libraries are shown in Figure 3.13.

C) RiboSeq read densities on the WT viral genome from the counterpart libraries to B. Reads were separated according to phase (0: purple, −2/+1: blue,
An overview of viral transcription and translation

-1/+2: yellow), and densities plotted after application of a 15 codon running mean filter. Only read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used to generate this plot. Further replicates and KO2 libraries are shown in Figure 3.14. **D)** Ratio of virus-mapping reads to [positive-sense host mRNA- and positive-sense vRNA-mapping reads]. Virus-mapping reads in the numerator were split into the following categories: positive-sense RNASeq (green), negative-sense RNASeq (red), and positive sense Riboseq (purple). All read lengths were used. The line graphs represent the mean ratios for each category, calculated from WT and KO2 data combined. The individual datapoints are also plotted, with WT (cross) offset to the left and KO2 (triangle) offset to the right to aid visualisation. The Riboseq (+) 3 hpi timepoint is plotted here to represent the upper limit of the virus fraction - quality control plots for this fraction of these libraries indicates they do not contain a high proportion of genuine RPFs, so the true ratio at this timepoint is likely lower. **E)** Data from D represented on a linear scale. Here, data from WT (solid bars) and KO2 (hatched bars) are plotted separately, and individual datapoints are plotted as black circles. **F)** Ratio of the density of sgRNA-mapping reads to ORF1a-mapping reads. All read lengths were used, and densities were calculated as RPKM of reads from each category in E. Note that the nsp2 frameshift site and downstream region were excluded from the ORF1a region (coordinates of both regions given in Appendix section 3). Riboseq 3 hpi libraries were excluded, and negative-sense RNASeq was omitted from the plot at 3 hpi due to the number of reads being insufficient for robust assessment of the ratio. Categories arranged and plot constructed as in E, with a grey line indicating a ratio of one.
Figure 3.13 Further replicates of a timecourse of NA PRRSV viral transcription

RNASeq read densities on the viral genome, for all libraries not shown in Figure 3.12B. Plots constructed as in Figure 3.12B.
An overview of viral transcription and translation

Figure 3.14 Further replicates of a timecourse of NA PRRSV viral translation

RiboSeq read densities on the viral genome, for all libraries not shown in Figure 3.12C. Plots constructed as in Figure 3.12C.
Figure 3.15 Overview of EU PRRSV transcription and translation at 8 hpi and investigation of negative-sense RiboSeq reads

A) Genome map of the EU PRRSV strain used in this study (full sequence given in Appendix section 1); similar to Porcilis® vaccine strain, GenBank accession KJ127878.1). Genome map constructed as in Figure 3.12A, with subgenomic RNAs omitted for space considerations. B) RNASeq read densities on the viral genome of EU PRRSV. Plot constructed as in Figure 3.12B. C) RiboSeq read densities on the viral genome of EU PRRSV. Plot constructed as in Figure 3.12C, except for the selection of read lengths to include — in this case, read lengths showing good phasing were selected for inclusion (indicated in Figure 3.10E). D) Ratio of virus-mapping reads to [positive-sense host mRNA- and positive-sense vRNA-mapping reads]. Mean values at each timepoint for NA PRRSV (lines) are reproduced from Figure 3.12D for comparison to EU PRRSV, which is plotted as two crosses of the corresponding colour representing the two samples. Plot constructed as in Figure 3.12D, but with the addition
of negative-sense RiboSeq reads (yellow line indicating mean values, black circles indicating individual NA PRRSV libraries). E) Data from D represented on a linear scale for EU samples and 9 hpi NA samples. Bar height represents the mean for each condition (NA sample mean was calculated using WT and KO2 samples combined), with individual datapoints plotted as black circles. F) Ratio of the density of sgRNA-mapping reads to ORF1a-mapping reads for EU samples and 9 hpi NA samples reproduced from Figure 3.12F for comparison. Ratio calculated as described for Figure 3.12F, using analogous regions of the EU PRRSV genome (coordinates in Appendix section 3). Bar height represents the mean for each condition (NA sample mean was calculated using WT and KO2 samples combined), with individual datapoints plotted as black circles. G) Length distribution of negative-sense RiboSeq reads mapping to the viral genome for NA (6 to 12 hpi) or EU (8 hpi) PRRSV. The length distribution of host CDS-mapping RPFs from Figure 3.4 and Figure 3.10C is reproduced for comparison (grey). H) Percentage of reads attributed to each phase for negative-sense RiboSeq reads mapping to the viral genome for NA (6 to 12 hpi) or EU (8 hpi) PRRSV.
3.5 Investigation of PRF on the viral genome

In this section, analysis of PRF provides a more detailed insight into changes in translation of the viral genome over the course of infection. PRF is an important mechanism used by PRRSV to control the production of viral proteins: at the canonical, ORF1ab PRF site it facilitates a reduction in stoichiometry of nsp9-12 compared to the upstream proteins\(^{426,520}\), and at the non-canonical nsp2 PRF site it facilitates production of three variants of nsp2 (references \(^{280-282,536}\)).

3.5.1 Visualising the occurrence of PRF

The occurrence of frameshifting at both sites is evident from changes in the proportion of reads in each phase after the PRF sites. In order to more clearly visualise this, plots of read density in each phase at each position on the viral genome were truncated to display only the polyprotein, to avoid the sgRNA region of the genome dominating the scaling at later timepoints (Figure 3.16 and Figure 3.17).

After the ORF1ab PRF site, a drop in ribosome density corresponds to the termination of ribosomes that do not frameshift, while those that continue translation into ORF1b shift into the \(-1\) frame, leading to a dominance of the \(-1\) phase in this region. At late timepoints, an increase in the proportion of reads in phases 0 and \(-2\) is seen towards the end of ORF1b, which results from translation of novel ORFs which will be discussed in section 4.3 (note that this region is therefore excluded from analyses of PRF).

Throughout the \(-2\) transframe region after the nsp2 frameshift site, there is a substantially increased proportion of reads in the \(-2\) phase, corresponding to ribosomes which have undergone \(-2\) frameshifting (Figure 3.16 and Figure 3.17). This is even more striking when the percentage of reads in each phase at each position in ORF1a is plotted (Figure 3.18). In these plots, a decrease in the proportion of reads in the 0 phase is seen over the transframe region, with a corresponding increase in reads in the \(-2\) phase. This is more pronounced at 9 and 12 hpi than at 6 hpi, suggesting a greater proportion of ribosomes may frameshift at the later timepoints.

Dividing the viral genome into four regions (the upstream, transframe and downstream regions relative to the nsp2 PRF site and the region downstream of the ORF1ab PRF site) further highlights the changes in phase composition due to frameshifting at both PRF sites (Figure 3.19). In WT libraries, a clear increase in the percentage of reads attributed to the \(-2\) phase is seen in the nsp2TF transframe region compared to the regions upstream and downstream. This is not observed for KO2 libraries, indicating that it is not the result of technical biases. Downstream of the ORF1ab PRF site, the predominant phase changes to \(-1\). Where the proportion of ‘out of phase’ reads here (i.e. reads in the 0 or \(-2\) phase) is greater than the proportion of ‘out of phase’ (\(-1\) or \(-2\) phase) reads in the
upstream region, this is likely due to some RNP contamination remaining even after read length stratification.

The same features of translation are evident on the EU PRRSV viral genome (Figure 3.19 and Figure 3.20), although the change in phase composition towards the −2 phase in the nsp2TF transframe region is less pronounced. This could be due to lower frameshift efficiency at this site in this strain/timepoint compared to the NA PRRSV strain, combined with differences in technical biases between both datasets.

**Figure 3.16 Distribution of reads in each phase across ORF1ab on the NA PRRSV genome**

RiboSeq read densities on the ORF1ab polyprotein region of the viral genome. Plot constructed as in Figure 3.12C. Regions defined as “upstream” and “downstream” in the frameshift efficiency calculations for the nsp2 and ORF1ab sites (sections 3.5.2 and 3.5.3) are annotated above the genome map. Only read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used to generate this plot. Further replicates are shown in Figure 3.17.
Figure 3.17 Distribution of reads in each phase across ORF1ab on the NA PRRSV genome, further replicates

RiboSeq read densities on the ORF1ab polyprotein region of the viral genome, for all libraries not shown in Figure 3.16. Plot constructed as in Figure 3.16.
Investigation of PRF on the viral genome

Figure 3.18 Percentage of RiboSeq reads in each phase across the ORF1a region of the NA PRRSV genome

Reads were separated according to phase, and a 183 codon running mean filter applied to avoid any instances of 0 across ORF1a (excluding the halfwindow at each end). From this, the percentage of reads in each phase at each codon was calculated (0: purple, +1/−2: blue, +2/−1: yellow). Only read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used to generate this plot.
Analysis of programmed ribosomal frameshifting on the PRRSV genome

**Figure 3.19** Bar charts of percentage of reads in each phase in the specified regions of the viral genome, for NA and EU PRRSV

Read lengths showing good phasing were selected for inclusion in this analysis, as indicated in Figure 3.8 for NA PRRSV libraries and Figure 3.10E for EU PRRSV libraries. The genome coordinates between which the 5’ end of RPFs must map to be counted as within each region can be found in Appendix section 3.
Figure 3.20 Distribution of reads across the ORF1ab region of the EU PRRSV genome

A) RiboSeq read densities separated according to phase. Read lengths selected for this analysis are those annotated in Figure 3.10E, and the plot was constructed as in Figure 3.16. B) Percentage of RiboSeq reads in each phase across the ORF1a region of the EU PRRSV genome. Read lengths selected for this analysis are those annotated in Figure 3.10E, and data were processed as described in Figure 3.18 (with the same 183 codon running mean filter) to generate percentages. C) RNASeq read densities on the ORF1ab polyprotein region of the EU PRRSV genome. Plot constructed as in Figure 3.12B, with a 45 nt running mean filter applied. All read lengths were used to generate this plot.

Although frameshifting at the nsp2 site is evident on both viral genomes from the analysis of RPF phasing, the accompanying drop-off in ribosome density, as ribosomes terminate at transframe stop codons, is not readily discernible like it is after the ORF1ab PRF site (Figure 3.16, Figure 3.17 and Figure 3.20). This may be due to a greater effect of technical bias in ORF1a or due to biological factors unrelated to frameshifting which affect the distribution of ribosomes. Interestingly, the pattern of ribosome distribution across ORF1a of NA PRRSV seems to change as infection progresses. At all timepoints, the RPF density in the first ~2000 nt is slightly greater than in the region directly downstream. This is potentially caused by the presence of non-canonical “heteroclite” sgRNAs, discussed in section 4.2 (note that this region is therefore excluded from analyses of PRF). After the first ~2000 nt, RPFs are reasonably evenly distributed across ORF1a at 6 hpi, while at 9 hpi, the skew in the RPF distribution changes to favour the region between the end of nsp2TF and nsp8. At 12 hpi the RPF distribution resembles the 9 hpi pattern, but with a heightened
peak shortly after the beginning of nsp2, corresponding to ribosomes with a proline codon in both the P and A sites (P site genomic coordinates 1583-1585). Prolines are well-known to be associated with ribosomal pausing, due to the restricted geometry of the amino acid, and this peak likely represents a ribosomal pause site\cite{305,374,395}. The extent of pausing increases over the course of infection, perhaps due to changes in ribosome density in the CDS or depletion of cellular factors such as prolyl-tRNA or eIF5A, an elongation factor which promotes peptidyl transfer on difficult motifs such as poly-proline\cite{84,104,106,109,110}. The reason for the observed changes in RPF distribution across the broader ORF1a region (downstream of the first ~2000 nt) is unclear, but is likely related to biological factors, as the effect of technical bias (which is largely determined by nucleotide sequence) is expected to remain constant over the timecourse. Non-canonical viral transcripts, such as transcripts with internal deletions, could contribute to such a phenotype. Although the RNASeq read density does not follow the same pattern as the RiboSeq (Figure 3.21), and no abundant non-canonical viral transcripts were found that would correlate with these changes in RPF density (section 4.2), low-abundance novel viral transcripts could contribute. Various other biological factors may also be at play, such as changes in the speed of translation, which may be affected by availability of cellular factors.

The KO2 mutant virus is expected to experience the same technical biases and potential confounding biological factors as the WT virus, and indeed displays the same changes in ribosome distribution over time. In order to control for these effects and visualise features of translation related to the presence of a functional PRF site and PCBP-binding motif, the read density at each position on the WT genome was normalised by the density in the corresponding position on the KO2 genome to produce a ratio plot (Figure 3.22). Note that, to try to manage factors such as batch effects and potential differences in rate of infection progression, WT and KO2 libraries were paired according to processing batches. Within each batch, libraries were paired according to the ratio of virus to host reads as a proxy for infection progression (a similar metric to that used in Figure 3.12D and E). Table 4 details the pairings and the rationale behind each of them.
Figure 3.21 RNASeq read densities on the ORF1ab polyprotein region of the NA PRRSV viral genome

Plot constructed as in Figure 3.20, with a 45 nt running mean filter applied. All read lengths were used to generate this plot.
Figure 3.22 Ratio of RPF density at each position on the NA PRRSV WT genome relative to the KO2 genome

The 5’ UTR and everything downstream of ORF1a was excluded, only positive-sense reads were used, and a 75 nt running mean filter was applied before the division. Only read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used to generate this plot. Regions defined as “upstream” and “downstream” in the nsp2 density-normalisation-based frameshift efficiency calculations are annotated above the genome map, with the average RPM ratio in these regions indicated by a grey line to the left of the plot, and the calculated combined −1 and −2 PRF efficiency annotated.
After normalisation, a decrease in ribosome density can be seen after the end of nsp2TF for most samples, corresponding to ribosomes which have undergone −2 PRF terminating translation at the end of the transframe region (Figure 3.22). There is no obvious drop in RPM ratio at the end of nsp2N indicating termination of ribosomes after −1 PRF. PRF efficiency at this site was previously determined by radiolabelling of MARC-145 cells infected with an EU PRRSV strain at 24 hpi, and revealed a −1 PRF efficiency of 7% (compared to a −2 PRF efficiency of 20%)\textsuperscript{280}. A frameshift efficiency of this relatively small magnitude may not be visible above the background variation in read density in this system. Alternatively, differences between the two reading frames in translation speed in the transframe region of nsp2TF could mask the potential drop-off. Encoding overlapping ORFs entails imposing certain restrictions on codon usage in order to maintain production of functional proteins in both frames, and this may lead to usage of sub-optimal codons or motifs which slow down translation. If this effect were more pronounced in the −2 frame this would lead to increased total (all phases combined) read density in the transframe region of the WT libraries compared to KO2. If the RPM ratio were increased in this way, it might mask ribosome drop-off at the end of nsp2N and delay its appearance until after the end of the nsp2TF transframe region, where the dual coding restrictions no longer apply. In fact, in the dual coding region, the −2 frame is better adapted to the cellular tRNA pool than the 0 frame according to their relative stAI values (0.433 and 0.415, respectively), which would theoretically lead to the opposite effect; however, stAI is far from a perfect correlate of translation speed, and many other factors, such as dicodon usage\textsuperscript{396,397}, likely contribute in this case.

The above analyses clearly demonstrate the occurrence of −1 PRF at the ORF1ab site and −2 PRF at the nsp2 site. The resultant changes in phasing are compelling, and are further supported by decreases in RPF density corresponding to ribosomal drop-off after both PRF sites.

### 3.5.2 Calculation of PRF efficiency at the nsp2 site

In addition to demonstrating the occurrence of PRF, ribosome profiling can be used to quantify its efficiency. Two different methods were used to assess PRF efficiency at the nsp2 site, and the results were compared. In addition, bootstrap resampling was performed for both methods to gain insight into the extent to which the results are affected by noise and technical biases and to gauge statistical significance.

The efficiency of PRF can be quantified using the observed drop-off in ribosome density after the frameshift site (referred to herein as the density normalisation-based method). In the case of the nsp2 PRF site, ribosomes which frameshift terminate translation either immediately (−1 PRF) or after 169 codons (−2 PRF). The ribosomes in the region downstream of nsp2TF are those which have not frameshifted, while the ribosomes upstream of the PRF site represent the total number of ribosomes
that pass through the PRF signal. Combined $-1$ and $-2$ PRF efficiency is therefore calculated as $1 -$ downstream/upstream, where downstream and upstream represent the read densities in the relevant regions (annotated in Figure 3.22; coordinates in Appendix section 3) on the WT genome, normalised by the density in the corresponding region of the KO2 genome to account for the potential confounding factors described above. Note that, due to the absence of a discernible drop in RPM ratio at the end of nsp2N, it was not possible to separate the calculation of $-1$ PRF efficiency from that of $-2$ PRF. The RNASeq libraries were used as a negative control, as the read density for these libraries is not expected to change between the upstream and downstream regions, so performing the same calculation should result in a value of 0%. Note that, due to the requirement for normalisation by KO2, this method was not performed for the EU PRRSV dataset, for which no KO2 libraries were made.

The efficiency of $-2$ PRF can alternatively be calculated based on the difference in phase composition of the region upstream of the PRF site and the $-2$ transframe region, in particular the increase in the proportion of reads in the $-2$ phase (Figure 3.19; see Materials and Methods for details of calculation). This method (referred to herein as the phasing-based method) cannot inform on $-1$ PRF, as ribosomes which shift into the $-1$ frame terminate immediately and do not contribute to the phase composition of the nsp2TF transframe region. This method has the advantage that, as it does not require a paired PRF-defective mutant library, it can be performed for the EU PRRSV dataset, and for NA PRRSV the KO2 libraries can be processed separately as a negative control.

Bootstrap resampling was performed to determine the robustness of the results of each method. The regions used in the PRF efficiency calculations (Appendix section 3) were divided into codons and $n$ codons were selected, with replacement, where $n$ is the length of the region (in codons). These codons were used as the input for both the PRF efficiency calculation methods described above, and this process was repeated 100,000 times (see Materials and Methods for further details). This introduces stochasticity into the system and avoids over-interpretation of results which depend heavily on the particular set of codons present in the data. Where the signal to noise ratio in the data is low, or where codon-to-codon variability in technical biases is high, bootstrap resampling is expected to result in a broader distribution of calculated values. To maximise the proportion of reads that are informative with regards to PRF efficiency, RiboSeq read lengths identified as having minimal RNP contamination (or good phasing in the case of the EU PRRSV libraries) were used as input for all PRF efficiency calculations and resampling using both methods. For RNASeq negative controls, all read lengths were used.

The density normalisation-based calculations give combined $-1$ and $-2$ PRF efficiency values of up to 38% (Figure 3.23A, filled bars; Figure 3.23B); however, the variation in calculated efficiency between replicates was considerable. At 6 hpi this was a particular problem, as the calculated PRF efficiencies of WT replicates one and two were 2.6% and 26% respectively, despite the fact that
efficient −2 PRF in the 6hpi-WT-1 library is evident from the change in phasing (Figure 3.17, Figure 3.18 and Figure 3.19). The read depth of the 6 hpi libraries is lower than at the later timepoints and, while the quotient involved in the density normalisation-based method is important for controlling bias, it may also amplify noise, particularly at low read depths. This is supported by the bootstrap results, in which the population of values obtained for the 6 hpi libraries (including the RNASeq libraries processed as a negative control) have broad distributions (Figure 3.23D).

For the density normalisation-based method, RNP contamination would artificially decrease the calculated efficiency of PRF, as the increase in read density caused by RNPs would have a greater relative effect in the region downstream of the frameshift site due to the lower density of bona fide RPFs in this region. Libraries for which the quality control plots (section 3.3.1.5) indicate this is a particular concern are 9hpi-WT-3 and 9hpi-KO2-4 (which are paired together), 9hpi-KO2-2 (paired with 9hpi-WT-2) and 12hpi-WT-2 (paired with 12hpi-KO2-2). There is not an obvious tendency for these libraries to generate lower frameshift efficiency values than other libraries (Figure 3.23A, filled bars; Figure 3.23D); however, the distributions of results from bootstrap resampling for these libraries are broader than other libraries at the same timepoints, indicating RNP contamination may be slightly increasing noise (Figure 3.23D).

The overall trend in the results of the density normalisation-based method suggests that PRF efficiency does not change between 9 and 12 hpi. The results from 6 hpi libraries were excluded from statistical testing due to the concerns discussed above. For the remaining libraries, statistical significance was assessed by determining how many datapoints from the bootstrap distributions of all 9 hpi populations combined (Figure 3.23D) were equal to or greater than the mean of the observed values at 12 hpi (Figure 3.23A, filled bars). This number was divided by the total number of datapoints in the combined 9 hpi populations to calculate a \( p \) value of 0.72, and the process repeated with the groups reversed \( (p = 0.43) \), indicating there is no significant change in frameshift efficiency between 9 and 12 hpi. However, the variability between replicates is a concern and these results should be interpreted with caution due to the caveats described above.

The phasing-based method reveals that −2 PRF occurs efficiently in both WT samples at 6 hpi and its efficiency increases at later timepoints (Figure 3.23A, dashed bar outlines). The variation between replicates is low, and the values calculated for the negative control KO2 libraries are close to 0% (mean: −3.0 ± 1.4%) with no significant changes over time (Figure 3.23C). For the bootstrap resampling, an additional negative control was added: a region of equal length to the nsp2TF −2 transframe region was selected, downstream of nsp2TF (coordinates in Appendix section 3), and the same calculations performed. This region does not contain overlapping ORFs and, as expected, generated PRF efficiency values of approximately 0% (Figure 3.23E).
Figure 3.23 Percentage frameshift efficiency at the nsp2 site, calculated by two different methods

A) Bar chart of observed values calculated by the density normalisation-based method (combined −1 and −2 PRF; filled bars) and the phasing-based method (−2 PRF only; unfilled bars with dashed outlines). For EU PRRSV samples no KO2 libraries were available so the normalisation-based method was not performed. Only read lengths identified as having minimal RNP contamination (NA PRRSV; Figure 3.8) or good phasing (EU PRRSV; Figure 3.10E) were used to perform these calculations. B) Application of the combined −1 and −2 PRF efficiency density normalisation-based calculation to the NA PRRSV RNASeq libraries as a negative control (expected value ~0%). All read lengths were used to perform these calculations. C) Percentage −2 PRF efficiency for the NA PRRSV KO2 negative control libraries, calculated using the phasing-based method, using read lengths with minimal RNP contamination (Figure 3.8). D) Distribution of results from calculation of frameshift efficiency by the nsp2 density normalisation-based method after 100,000 bootstrap resamples of the codons in the upstream and downstream regions. The same calculations were performed on
Investigation of PRF on the viral genome

If reads originating from RNPs are distributed evenly across the three phases, they should not influence values determined by the phasing-based method (see Materials and Methods for algebraic proof). In line with this, libraries for which RNP contamination is a potential concern do not yield −2 PRF efficiency values that differ obviously from the other replicates (Figure 3.23A, dashed bar outlines). These libraries do, however, give rise to a slightly broader distribution of values after bootstrap resampling (Figure 3.23E). This is because, although RNPs are not expected to create a systematic bias resulting in under- or over-estimation of −2 PRF efficiency by this method, they do reduce the fraction of the library that is representative of genuine translation, therefore lowering the signal to noise ratio.

A potential systematic bias could be introduced into frameshift efficiency values calculated using the phasing-based method if the two overlapping reading frames were translated at different speeds (discussed in the previous section). If translation were slower in the −2 frame in the overlapping region this would artificially inflate PRF efficiency (and vice versa). However, as this is an inherent property of the nucleotide sequence in the region, it is not expected to change over time, and therefore the trends observed using this method should be robust.

The strong reproducibility between replicates, including at 6 hpi and in two different negative controls, suggests that the phasing-based method of PRF efficiency calculation is not greatly affected by sources of noise in the data, and it is likely to be a more reliable measure of PRF efficiency than the density normalisation-based method.

The phasing-based method reveals an increase in −2 PRF efficiency over time. At 6 hpi, −2 PRF efficiency is 23% and this increases to 39% at 9 hpi (p value based on bootstrap populations = 3.5×10−5), at which point it reaches a plateau (Figure 3.23A, dashed bar outlines). This supports a model of increasing PRF efficiency as nsp1β, the viral protein responsible for stimulating PRF at this site, accumulates over the course of infection. This corresponds well to the observed changes in nsp1β protein levels over time, in which the protein is faintly detectable at 6 hpi, followed by a substantial increase at 9 hpi and a subsequent plateau (Figure 3.2A).

The 8 hpi EU PRRSV libraries, for which PRF efficiency calculations were only performed using the phasing-based method, had a mean −2 PRF efficiency of 26%. The fact that this is lower than
the efficiency at 9 hpi in the NA PRRSV libraries (p value based on bootstrap populations = 0.0014) is likely related to differences between the two viruses as opposed to the difference in timepoints. The analyses in section 3.4 indicated the 8 hpi EU PRRSV samples had progressed to a similar point in the replication cycle compared to the 9 hpi NA PRRSV libraries. In addition, nsp1β is clearly detectable by western blot in EU PRRSV-infected cells at 8 hpi and doesn’t substantially increase at later timepoints (Figure 3.2B), suggesting that a significant increase in PRF efficiency between 8 and 9 hpi is unlikely.

The only arterivirus for which PRF efficiency at the nsp2 site has previously been measured in the context of infection is the EU PRRSV isolate SD01-08, for which it was assayed by 35S-Met radiolabelling of MARC-145 cells at 24 hpi (MOI 0.1) [280]. In this system, −2 PRF was estimated to occur with 20% efficiency [280], which is similar to the 26% −2 PRF efficiency calculated for the EU PRRSV isolate used in this thesis, despite the much later timepoint used for the radiolabelling. This could be taken to suggest that −2 PRF efficiency for EU PRRSV does not change substantially after 8 hpi; however, biological differences between the two PRRSV isolates could have a considerable effect on PRF efficiency and it may be inappropriate to compare results from different timepoints in different systems.

Frameshift efficiency has not been previously quantified for the NA PRRSV nsp2 PRF site, neither in the context of infection nor in the context of a reporter construct transfected into cells. These ribosome profiling results reveal that, in infected MARC-145 cells, the −2 PRF efficiency for NA PRRSV is 39-41% at 9-12 hpi, which is significantly greater than the PRF efficiency for EU PRRSV at 8 hpi in this study, and is greater than the efficiency for SD01-08 PRRSV at 24 hpi [280]. This difference may reflect a different optimal abundance of the accessory factors nsp2TF and nsp2N between these two species of PRRSV, although it may be that within each species there are similar variations in PRF efficiency between different isolates. The −2 PRF efficiency calculated at 9-12 hpi for NA PRRSV is within the same range as values determined using reporter constructs based on the EU PRRSV SD01-08 PRF cassette. Transfection of these constructs into RK-13 (derived from rabbit kidney epithelium) or HEK-293T cells led to estimation of −2 PRF efficiencies of 50-56% and 38%, respectively [280,281,291]. Many factors can confound comparison of results from reporter constructs to those in the context of infection, for example PCBP1 and 2 were found to be partially redistributed upon infection, localising to viral RTCs instead of the uniform cytoplasmic distribution seen in mock-infected cells [292]. Differential compartmentalisation or cellular localisation of PCBP (or nsp1β) could affect PRF efficiency, and could contribute to the observed differences in PRF efficiency in NA/EU PRRSV-infected cells, as well as being relevant for comparing the contexts of transfection and infection. Despite the differences in both the cellular and sequence context, these SD01-08-based reporter construct studies confirm that, under the right conditions, the PRRSV nsp2
Investigation of PRF on the viral genome

site is capable of facilitating the highly efficient \( -2 \) PRF observed in these NA PRRSV ribosome profiling results.

The increase in \( -2 \) PRF efficiency observed over this timecourse of NA PRRSV infection represents a novel regulatory mechanism in arterivirus gene expression, providing only the second known example of temporally regulated PRF efficiency (alongside cardioviruses\(^{285}\)). A discussion of the potential evolutionary advantages of using this unique mechanism of protein-stimulated PRF for this purpose can be found in section 3.6.1.

### 3.5.3 Calculation of PRF efficiency at the ORF1ab site

To calculate \( -1 \) PRF efficiency at the ORF1ab site, the drop-off after the frameshift site was quantified. In this case, ribosomes in the region downstream of the PRF site represent ribosomes which have frameshifted, and PRF efficiency is simply defined as the RPF density in the downstream region divided by the upstream region (annotated in Figure 3.16; coordinates in Appendix section 3). Because frameshifting at this site facilitates expression of the RdRp, it is essential for viral replication and no knockout mutant was made. Therefore, read densities were used directly for these calculations, without normalisation. For NA PRRSV, read lengths identified as having minimal RNP contamination were used in this analysis and, to make the methods more comparable, read lengths with good phasing were used for EU PRRSV. After the PRF site, the region in which the 0-frame and the \( -1 \)-frame ORFs overlap is just a single codon in length, therefore quantification of PRF efficiency using changes in phasing was not attempted.

Surprisingly, this analysis revealed an increase in \( -1 \) PRF efficiency over the course of infection, apparently overturning the common assumption that PRF directed by secondary RNA structure operates at a fixed ratio (Figure 3.24A). As the mutations on the KO2 genome specifically target the nsp2 PRF site, they aren’t expected to affect PRF efficiency at the ORF1ab site. Therefore, WT and KO2 libraries were grouped together for statistical analysis, making the groups large enough to perform a two-tailed Mann-Whitney U test directly on the observed values. At 6 hpi, \( -1 \) PRF efficiency was 11%, increasing to 19% at 9 hpi \((p = 8.5\times10^{-3})\) and further increasing to 32% at 12 hpi \((p = 8.5\times10^{-3})\). The same trend was not observed in the RNASeq negative control libraries, which are expected to yield values of \(~100\%\). In fact, these data show the opposite trend, and RNASeq “PRF efficiency” decreases slightly between 6 and 12 hpi \((p = 0.030)\). This indicates that the increasing trend in the RiboSeq is not caused by technical biases shared between RNASeq and RiboSeq, or by an increase in non-canonical transcripts in the ORF1b region of the genome facilitating extra translation at later timepoints. Note that, in section 4.2, non-canonical viral transcripts are investigated in detail, and this informed the designation of the “ORF1ab downstream”
region used in these PRF calculations, which was 3’-truncated to avoid potentially confounding novel transcripts.

![Figure 3.24](image)

**Figure 3.24 Percentage frameshift efficiency at the ORF1ab site**

A) Percentage frameshift efficiency at the ORF1ab site, calculated using the read density on the viral genome upstream and downstream of the frameshift site. The bars represent the mean results for each timepoint, calculated from WT and KO2 data combined. The individual datapoints are overlaid, with WT (cross) offset to the left and KO2 (triangle) offset to the right to aid visualisation. The right-hand panel shows the results of applying these calculations to RNASeq reads as a control, for which the expected result is 100%. Only RiboSeq read lengths identified as having minimal RNP contamination (NA PRRSV; Figure 3.8) or, for comparability, good phasing (EU PRRSV; Figure 3.10E) were used to perform these calculations. For the RNASeq control all read lengths were used.

B) Distribution of results from calculation of frameshift efficiency by the ORF1ab density-based method after 100,000 bootstrap resamples of the codons in the upstream and downstream regions. The same calculations were performed on the RNASeq libraries as a negative control. For the violin plots, horizontal lines represent the median of the results distribution and the 95% confidence intervals. For both panels: * p < 0.05; ** p < 0.005; *** p < 0.0005.
Investigation of PRF on the viral genome

RNP contamination is a greater potential concern at the ORF1ab PRF site than the nsp2 site, as it has a greater effect in ORF1b than ORF1a due to the lower translational expression of the former. This could potentially cause artificial inflation of the calculated PRF efficiency. Although some of the samples for which this is a particular concern (9hpi-WT-3, 9hpi-KO2-2, 9hpi-KO2-4 and 12hpi-WT-2) have slightly broader populations of bootstrap resampling values (Figure 3.24B), the trend of increasing frameshift efficiency over time seems to be robust to the effect of RNPs. The aforementioned libraries do not show an obviously increased PRF efficiency compared to other replicates within their groups (Figure 3.24B) and the observed values of these libraries can be removed from their groups and the trend remains statistically significant (Mann-Whitney U p values of 0.020 and 0.037 for 6 to 9 hpi and 9 to 12 hpi comparisons, respectively).

The ORF1ab PRF efficiency on the EU PRRSV genome at 8 hpi was 23%, which is similar to the calculated efficiency for NA PRRSV at 9 hpi. Assessing statistical significance using the bootstrap populations (as described in the previous section), the EU PRRSV PRF efficiency is significantly lower than the value for NA PRRSV at 12 hpi ($p = 1.3 \times 10^{-4}$), and not significantly different from the NA PRRSV 9 hpi observed mean. This suggests that frameshifting at the ORF1ab site on the EU and NA PRRSV genomes occurs at similar efficiencies, which is to be expected since the stoichiometry of key replicase components such as the RdRp is likely to be conserved between the two species. The RNASeq control for the EU libraries had a mean value of 74% “PRF efficiency”, which is lower than the expected value of 100%. Inspection of the distribution of RNASeq reads across ORF1ab indicates there is some variation in read density between the upstream and downstream regions (Figure 3.20C). This suggests the EU library may be subject to technical biases which differ between these regions, as there were no non-canonical transcripts detected that would provide a biological explanation (section 4.2). If these technical biases are shared by the RiboSeq (for example ligation bias and PCR bias), this may cause an under-estimation of PRF efficiency at this site. In this case, normalisation of the RiboSeq data by the corresponding RNASeq data provides a potential method to control for these biases, but caution is advised as any non-overlapping biases (for example nuclease and alkaline hydrolysis biases) would be enhanced.

These results represent the first calculation of frameshift efficiency at the arterivirus ORF1ab −1 PRF site in infected cells. Previous studies have obtained estimates of PRF efficiency by transfecting reporter constructs into uninfected cells. In these studies, the PRRSV ORF1ab PRF site facilitated 16% −1 PRF efficiency in yeast$^{525}$, and the EAV ORF1ab PRF site facilitated 15-20% PRF efficiency in HeLa cells$^{524}$. Both sets of values fall within the 11-32% range of −1 PRF efficiency values observed in this thesis in the context of infection, suggesting a similar range of PRF efficiencies is required by diverse arteriviruses. Frameshift efficiency at the related coronavirus ORF1ab site has been determined in the context of infection, using the same method of ribosome profiling. For MHV, −1 PRF efficiency was found to be 48-70%$^{307}$, while for the newly emerged
Analysis of programmed ribosomal frameshifting on the PRRSV genome

SARS-CoV-2, efficiency was $57\% \pm 12\%^{375}$. These indicate PRF efficiency at the ORF1ab site is considerably higher for CoVs than for arteriviruses. This may reflect different nsp stoichiometry requirements, as the coronavirus genome is approximately twice the size of the arterivirus genome and encodes several nsps not present in arteriviruses$^{561}$, perhaps contributing to different selection pressures in the evolutionary optimisation of PRF efficiency. Whether the observed increase in PRRSV ORF1ab PRF efficiency over time also occurs in coronavirus-infected cells (or other arteriviruses) remains to be seen, as efficiency was only calculated at a single timepoint in these studies. Indeed, this was an unexpected result, for which the mechanism and potential significance remains to be elucidated (discussed in section 3.6).

3.5.4 Investigating ribosomal pausing at the PRF sites

Ribosome pausing over the slippery sequence has long been considered a mechanistically important feature of PRF$^{22,227,234,655}$. However, while observed to a small extent on WT shift sites in vitro$^{225,261,262,269,286}$, it has been elusive in ribosome profiling data$^{307,385}$. Mutating the slippery sequence but leaving the frameshift-stimulatory structure intact accentuates pausing at the frameshift site, which can aid its observation both in vitro and in profiling$^{22,282,285,286}$. A previous attempt to generate a virus (based on NA PRRSV strain SD95-21) with analogous mutations at the nsp2 PRF site, to render the slippery sequence defective but leave the PCBP/nsp1$\beta$ binding site intact, resulted in partial reversion (data not shown), so this dataset was discarded. However, the KO2 mutant virus, which has mutations rendering both the PCBP/nsp1$\beta$ binding site and the slippery sequence defective, was stable, presumably due to the greater number of bases that would need to revert to rescue the phenotype. As described above, this KO2 mutant provides a control for the technical biases inherent in ribosome profiling, which is particularly important when investigating pausing, as biases which may average out over larger regions can have a considerable effect at the level of individual nucleotides. By comparing the RPF density around the nsp2 frameshift site on the WT genome to that on the KO2 genome, many of these biases are accounted for and changes in dwell time that relate to PRF can be observed.

Figure 3.25A shows the density of RPFs at each position on the WT and KO2 viral genomes around the nsp2 frameshift site, plotted at the inferred ribosomal P site position according to a static +12 nt offset method of P site inference. As the length distributions of RPFs could be informative with regards to the conformation of ribosomes at the frameshift site, reads were not filtered by length to generate the plots in Figure 3.25.

Both frameshift sites on the PRRSV genome are thought to operate by tandem slippage frameshifting, based on their conserved heptanucleotide slippery sequences and PRF-stimulatory elements (section 1.2). Based on this model, the ribosome is expected to pause at the nsp2 frameshift
Investigation of PRF on the viral genome

site during the translocation event that begins with the GUU and UUU codons of the G_GUU_UUU slippery sequence in the P and A sites, respectively. At 9 hpi there is a small peak in the 0 phase (purple) on the WT genome corresponding to a ribosome with the GUU codon (red) in the P site, which could correspond to the beginning of this translocation event (Figure 3.25A). This is supported by the fact that this peak is predominantly composed of “classical” length RPFs (25-34 nt), which are thought to represent ribosomes trapped early in translocation, in the classical PRE state, based on evidence that CHX both stabilises this ribosomal conformation and enhances the proportion of classical RPFs.\textsuperscript{115,140,331} (Figure 3.25B). Although small, this peak is larger on the WT genome than the KO2 genome at 9 hpi. This could indicate slightly slower translocation over the slippery sequence when the frameshift cassette is functional; however, the 3' end of the classical-length reads in this peak fall within a region that is mutated in KO2, so differences in technical biases (such as ligation bias) affecting the 3' end could also contribute (Figure 3.25C).

One nucleotide downstream, a considerably larger peak is evident in the −2 phase at all timepoints on the WT genome, but not the KO2 genome (blue, marked with an asterisk; Figure 3.25A). Analysis of the distribution of read lengths that make up this peak (Figure 3.25D-E) reveals that for all WT libraries (except 12 hpi WT 2) the peak is predominantly composed of short reads, the majority of which are 21 nt long. As the bases mutated in KO2 are not adjacent to the ends of 21 nt reads at this position, technical biases are an unlikely cause of the differences between WT and KO2, and this peak is likely related to frameshifting (Figure 3.25C). Consistent with the fact that this peak predominantly consists of short reads, a much more prominent peak is seen at this position when the distribution of specifically short (19-24 nt) reads around the PRF site is plotted (Figure 3.26), although a smaller peak is also visible at 9 and 12 hpi in a plot of classical reads (Figure 3.27). The precise origin of this peak is unclear, although there are several possibilities, discussed below.
Figure 3.25 RiboSeq read densities around the nsp2 frameshift site, all read lengths

A) RiboSeq read densities around the nsp2 frameshift site, coloured according to the phase of RPF 5’ ends as in Figure 3.16. The genome sequence in this region is underlaid beneath the data in the top panel. Beneath this is underlaid the identity of the encoded amino acids, with the species-specific tRNA adaptation index (stAI) of each codon indicated as a heatmap (M. mulatta values were used as C. sabaeus values were unavailable). Codons in-frame with respect to nsp2 are displayed above those of the overlapping nsp2TF−2 frame. The slippery sequence is printed in red and the more prominent peak present at this position on the WT genome is indicated with an asterisk. All read lengths were used to make this plot. All libraries for a given timepoint are set to the same scale on the y axis, with no running mean filter applied. B) Length distribution of reads that make up the purple 0-phase peak corresponding to ribosomes with the GUU codon of the slippery sequence in the P site (one nucleotide upstream of the peak marked with an asterisk in panel A). Length distribution is plotted as the
number of reads (RPM) of each length, so libraries with a greater total number of reads at this position appear more prominent. Due to considerable differences in total read numbers at this base between WT (top) and KO2 (bottom) these libraries were plotted on separate scales. Note that the very low number of reads at this position on the KO2 genome means the length distributions in the KO2 panel are highly subject to noise. C) Positions of mutations in the KO2 viral genome (red) relative to the positions of reads comprising the majority of the peak marked with an asterisk in panel A (blue box) and the peak one nucleotide upstream (purple line). D) Length distribution of reads that make up the blue peak marked with an asterisk in panel A. Plot constructed as in panel B, again with the very low number of reads at this position on the KO2 genome meaning the length distributions in the KO2 panel are highly subject to noise. E) Data for selected libraries from D reproduced to show the number of reads of each length as a percentage of the total number of reads mapping to this base. All libraries are therefore displayed on the same scale regardless of total read counts. KO2 libraries were omitted, and 9 hpi WT replicate 1 was omitted due to the exclusion of reads below 25 nt in the preparation of this library.
Figure 3.26 RiboSeq read densities around the nsp2 frameshift site, short read lengths
Plot constructed as in Figure 3.25A, but using all read lengths within the range 19 to 24 nt.
Figure 3.27 RiboSeq read densities around the nsp2 frameshift site, classical read lengths
Plot constructed as in Figure 3.25A, but using all read lengths within the range 25 to 34 nt.
Although the translocation event during which frameshifting occurs begins with the GUU codon in the P site and the ribosome in the classical PRE state, the frameshift-associated pause is thought to occur at a late stage of translocation (section 1.2.1). The ribosome is thought to be in a rotated state, with some evidence suggesting a non-canonical and/or hyper-rotated conformation. The structure of this putative conformation is uncharacterised, and it may be that a different length of mRNA is protected from digestion by RNase I. This could lead to a slightly shorter distance between the 5’ end of the RPF and the ribosomal P site, displacing the pause peak in plots, such as these Figures, which use a static +12 nt offset to infer P site position. This could explain the appearance of the larger peak in the −2 phase instead of the 0 phase. As discussed above, translocating ribosomes are thought to protect fragments of RNA within the classical RPF length range, whereas this peak is composed predominantly of short RPFs, which are normally thought to originate from ribosomes without an aminoacyl-tRNA in the A site. As well as affecting the distance between the RPF 5’ end and the P site, a non-canonical conformation of the translocating ribosome could potentially affect the length of mRNA protected at the 3’ end of the RPF. For example, eEF2 is thought to bind the ribosome and dissociate multiple times during the pause in PRF, and it is unclear whether it would remain stably bound in the cell lysates used for profiling. If the ribosome pauses while the A site tRNA is partially translocated, and eEF2 dissociates in the lysate (or in the cell), there may be an opportunity for RNase I to gain access to the A site to a similar extent as it can when the A site is unoccupied during decoding, leading to shorter reads. While the above explanation is theoretically possible, it relies on several assumptions, and without further structural and/or biophysical characterisation of the translocation intermediates involved (and the regions of mRNA they protect) it remains purely speculative and realistically unlikely. Alternative explanations are explored below.

The nsp2 frameshift site likely operates by tandem slippage, and hungry codon frameshifting at this site is not supported by the observed ribosome profiles. Hungry codon frameshifting results in a long pause with the hungry codon (but no aminoacyl-tRNA) in the A site. In this case, this would be expected to result in a pronounced short-read peak with the P site in phase 0 over the GUU codon of the slippery sequence. However, the observed peak at this position, examined at the beginning of this section, is small and predominantly composed of classical length reads, arguing against hungry codon PRF.

It was previously suggested, based on the ability of the DnaX PRF site to stimulate two directly successive −1 frameshift events that result in a −2 frameshift, that the −2 PRF at the nsp2 site in PRRSV could potentially occur in a similar way. In the DnaX example, the second −1 PRF event was stimulated by the hungry codon mechanism, as the concentration of the relevant aminoacyl-tRNA was limiting. For the PRRSV nsp2 site, completion of a (presumably tandem slippage) −1 frameshift would position the −1-frame UGA stop codon in the A site. If this stop codon was slow
to decode (for example if release factor access to the A site was impeded by PCBP/nsp1β) this could stimulate a second −1 PRF event via the hungry codon mechanism\textsuperscript{228}. This would be expected to result in a pronounced short-read peak in the −1 phase (yellow), with the P site one nucleotide downstream of the blue peak observed here (Figure 3.26). No short-read peak at this alternative position is observed, and the ribosome is not known to adopt an unusual conformation during hungry codon PRF that might alter the distance between the RPF 5′ end and the P site to displace the peak on this plot. This suggests that a −1 PRF event stimulated by a UGA hungry codon is unlikely. Additionally, the more recently sequenced genomes of the arterivirus SHFV (and related viruses) reveal that the −1-frame stop codon is not positioned directly adjacent to the slippery sequence as it is in PRRSV, although −2 PRF is still utilised in SHFV\textsuperscript{287}. This supports the conclusion that two successive −1 PRF events, in which the second is stimulated by a hungry codon, is an unlikely mechanism for the −2 PRF event on the PRRSV genome, and suggests an alternative origin for this short-read peak.

The most parsimonious explanation is that the −2-phase peak (Figure 3.25A, Figure 3.26 and Figure 3.27; marked with an asterisk) does not result from a frameshift intermediate, but represents ribosomes pausing while decoding the first codon in the −2 frame after frameshifting (A site UUG), which would not be decoded on the PRF-defective KO2 genome. This would explain the fact that the peak occurs in the −2 phase as opposed to the 0 phase and would explain the preponderance of short RPFs if the ribosomes were paused with an empty A site during decoding. The UUG (Leu) codon that would occupy the A site under this hypothesis has a stAI value of 0.71, indicating it is not poorly adapted to the cellular tRNA pool (Figure 3.28) – in fact it is better-adapted than most of the surrounding codons in either the 0 or −2 frame (Figure 3.25A, Figure 3.26 and Figure 3.27; heatmap under 6 hpi WT replicate 2). This could argue against a decoding pause; however, it is possible that other factors, such as potentially incomplete or slow dissociation of the PCBP/nsp1β complex just downstream, might hinder decoding so soon after PRF and lead to a pause despite the relative abundance of the relevant aminoacyl-tRNA. Further, stAI is only one of many factors that correlate with translation speed, and there may be other features, not directly related to frameshifting, that could also play a role in this particular sequence context.
Analysis of programmed ribosomal frameshifting on the PRRSV genome

Figure 3.28 Histogram of stAI of sense codons in M. mulatta

The stAI for UUG (encoding leucine) is indicated by a black dashed line. Species-specific tAI values were not available for C. sabaeus, so M. mulatta was used as it is a closely related species.

At the ORF1ab frameshift site, analysis of potential frameshift-related changes in ribosome dwell times is less reliable, due to the lack of a frameshift-defective mutant to control for the effects of technical biases. Plots of the distribution of RiboSeq reads around the PRF site (Figure 3.29, Figure 3.30 and Figure 3.31) do not show evidence of a putative PRF translocation-intermediate-derived short-read peak at an analogous position to that seen at the nsp2 PRF site (which would be observed as a blue, −2-phase peak on the third U of the slippery sequence at the ORF1ab frameshift site). In some libraries (9 hpi WT replicate 1 and 12 hpi WT replicate 2) the peak corresponding to ribosomes with P sites over the UUA of the U_UUA_AAC slippery sequence is heightened compared to surrounding peaks (Figure 3.29 and Figure 3.31; peak indicated by asterisk). This could suggest a slight pause due to the downstream pseudoknot preventing translocation, consistent with the observation that this peak is predominantly composed of classical-length reads (Figure 3.29B). However, this peak is not heightened in most libraries, and without a PRF-defective mutant with which to compare it, it is unclear whether this putative pause is related to frameshifting.
Figure 3.29 RiboSeq read densities around the ORF1ab frameshift site, all read lengths

A) Plot of RPF density on the WT and KO2 viral genomes, constructed as in Figure 3.25A, using all read lengths. The peak (purple) corresponding to ribosomes with the slippery sequence (red) in the P and A sites is marked with an asterisk. B) Length distribution of reads that make up the purple 0-phase peak corresponding to ribosomes with the UUA codon of the slippery sequence in the P site (peak marked with an asterisk in panel A). Plot constructed as in Figure 3.25B. WT and KO2 libraries were plotted separately to aid visualisation of the large number of libraries, although in this case there are similar numbers of reads at this position on the two genomes.
Figure 3.30 RiboSeq read densities around the ORF1ab frameshift site, short read lengths
Plot constructed as in Figure 3.25A, using all read lengths within the range 19 to 24 nt.
Figure 3.31 RiboSeq read densities around the ORF1ab frameshift site, classical read lengths

Plot constructed as in Figure 3.25A, using all read lengths within the range 25 to 34 nt. The peak (purple) corresponding to ribosomes with the slippery sequence (red) in the P and A sites is marked with an asterisk.
Previous investigations of PRF by ribosome profiling did not capture the population of short RPFs\textsuperscript{285,307,385}, and it will be interesting to see whether future investigations which purify short reads, and have appropriate PRF-defective controls, reveal similar short-read peaks at other PRF sites. Through comparison of the specific positions of potential peaks, such studies may help discriminate between the potential origins of the short-read peak observed at the nsp2 PRF site here. Additionally, analysis of a mutant PRF cassette in which the slippery sequence, but not the frameshift-stimulatory elongation blockade, has been mutated would also be informative. This would be expected to accentuate the pause caused by the elongation blockade, which would help discriminate between a peak resulting from decoding in the alternative reading frame compared to a peak that might be mechanistically relevant for PRF. Under these motivations, an investigation of PRF on the TMEV genome, using profiling in which short RPFs were captured, was performed and is described in Appendix section 6. An analogous short-read peak was not discovered, although it is possible that the observed formation of disomes at the TMEV frameshift site could have obscured a short-read peak if one were present.

3.5.5 Investigating the potential for lengthened reads at the PRF sites

In addition to the 19-34 nt monosome-protected fragments, it is possible that longer protected fragments could be generated at or near frameshift sites which could shift the lengths of these RPFs outside the range that is purified in monosome profiling. There are two reasons this could be the case. Firstly, the elongation blockade in the PRF cassette may, itself, protect a fragment of mRNA, potentially extending the region protected by the ribosome. For example, at the nsp2 PRF site, the complex of PCBP and nsp1\(\beta\) binds the C-rich motif downstream of the slippery sequence. If this complex is stable enough to protect the mRNA from nuclease digestion, this could potentially lead to an extended RPF that is protected by a combination of the ribosome at the slippery sequence and the downstream PCBP/nsp1\(\beta\). Formation of the pseudoknot downstream of the ORF1\(\alpha\b) frameshift site could have a similar effect. If these putative 3’ extensions of the RPFs at the frameshift site increased their length above the ~34 nt threshold they would be excluded from the library during the size selection step in the RNA purification. The second possibility that could lead to longer protected fragments is the formation of disomes. For example, if ribosomes are paused at the frameshift site for long enough for the upstream ribosome to catch up, the leading and trailing ribosomes could collide and form a disome. Disomes are known to protect ~60 nt regions of mRNA and would therefore also be excluded from the library preparation\textsuperscript{303}.

To protect against unintentional exclusion of RPFs at the frameshift site for these reasons, a “broad spectrum” gel slice of 19-80 nt was purified for the 9 hpi replicate 2 libraries. The exact lengths of mRNA that would be protected by the elongation blockades at each site are not known, so reads of
all read lengths above 34 nt (referred to as broad spectrum reads) were used to investigate the potential presence of these fragments. Disome-protected fragments (DPFs), on the other hand, have been characterised in other ribosome profiling studies, and are expected to protect fragments of ~52 nt, ~54 nt and ~58-64 nt\textsuperscript{186,202,303,336}. Further, the 5′ end of DPFs is determined by ribosomal protection, just like monosome-protected fragments, so DPFs are expected to show phasing\textsuperscript{186}. Although the peaks in the length distribution of RiboSeq reads above 34 nt long are not clear above the level of background (Figure 3.32A), some read lengths display phasing skewed towards the 0 phase, suggesting an enriched fraction of DPFs in these read lengths (Figure 3.32B). Based on these observations of phasing, and the expected lengths protected by disomes, reads of lengths 51, 54 and 58-64 nt were selected for analysis as (putative) disome-protected fragments.

Analysis of the distribution of both broad spectrum reads and DPFs provided no evidence of protein binding, RNA secondary structure, or ribosome collisions causing formation of extended protected fragments of mRNA at either frameshift site (Figure 3.32C-D, Figure 3.33). However, several features of both the broad spectrum and disome fractions of this dataset indicate that there is a considerable level of non-RPF contamination. The proportion of negative-sense reads is high, and many prominent peaks in the positive sense are (approximately) mirrored in the negative sense, suggesting an origin other than ribosomal protection (Figure 3.32C-D). Given these features, and the observation that the distribution of reads on the viral genome is reasonably similar to the distribution observed at 9 hpi by RNASeq (greatest in the leader and sgRNA regions), a likely origin is RTCs, in which dsRNA (a replication/sgRNA transcription intermediate) may be protected by viral proteins, for example the RdRp. Furthermore, reads longer than 34 nt make up only a very small fraction of these libraries (< 6% of positive-sense vRNA- and mRNA-mapping reads), and coverage of putative DPFs on the viral genome is low (Figure 3.33). Overall, this analysis suggests that prominent formation of disomes or extended RPFs at the frameshift sites does not occur at levels high enough to observe above the background. However, it is possible that 3′-extended fragments would become apparent at the frameshift site if greater sequencing depth and a lower level of non-RPF background was obtained, therefore the results are inconclusive.
Figure 3.32 Broad spectrum ribosome profiling at 9 hpi (replicate 2)

A) Length distribution of reads mapping within host CDSs for broad spectrum libraries. B) Total number of reads attributed to each phase, for reads of each specified length, in the broad-spectrum libraries. Read lengths which were selected for inclusion in the “disome-protected fragment” plots (D and Figure 3.33) are indicated by square brackets. C) Density of broad spectrum reads on the WT and KO2 viral genomes, with all reads above 34 nt in length selected for inclusion. These reads likely originate not only from disome-protected fragments but also other sources such as fragments of RNA protected by bound protein. Read densities are plotted as RPM after application of a 15 nt running mean filter. Positive-sense reads are plotted in green (above the horizontal axis), negative-sense in red (below the horizontal axis). In all disome or broad spectrum plots, reads are plotted at the inferred P site position of the colliding ribosome of the disome (inferred by adding a 12 nt offset to the 5’ end of the read. D) Density of disome-protected fragments on the WT and KO2 viral genomes. Reads of lengths 51, 54, and 58-64 nt were selected for inclusion (as indicated in B), and the plot was constructed as in C.
Figure 3.33 Density of disome-protected fragments around A) the nsp2 frameshift site and B) the ORF1ab frameshift site

Density of disome-protected fragments (read lengths as indicated in Figure 3.32B), plotted at inferred P site positions of colliding ribosomes in potential disomes around the frameshift sites, coloured according to phase. Otherwise, the plots were constructed as in Figure 3.25A.
3.6 Discussion and future work

3.6.1 Discussion

To study viral gene expression throughout a timecourse of NA PRRSV infection, I generated 60 sequencing libraries, including both RiboSeq and RNASeq, harvested at four timepoints selected to cover a single replication cycle. To complement this timecourse and extend investigations to the highly diverged second PRRSV species, previously sequenced libraries generated from EU PRRSV-infected cells at a single timepoint were analysed. In these analyses, key phases of the viral replication cycle were observed, beginning with translation of the polyprotein, followed by viral replication and then a focus on sgRNA transcription and translation at later timepoints. This analysis provides an overview of viral transcription and translation throughout the replication cycle, with a more detailed analysis of expression levels of individual viral transcripts and ORFs provided in Chapter 4.

In infected libraries at late timepoints, the proportion of short RPFs in the host mRNA-mapping fraction of the library was lower than in mock-infected or early timepoint libraries. While this could simply be due to unintended variations in the harvesting procedure, it could represent a translational response to stress. To test whether this difference is significant, RiboSeq libraries were grouped into early (3 and 6 hpi) and late (9 and 12 hpi) timepoints, to provide enough replicates in each group to perform a two-tailed t test. Positive-sense RPFs mapping to host mRNA were used, and short reads were defined as 19-24 nt long, with the denominator formed by 19-34 nt long reads. In the late timepoint group, there was a significant difference in the percentage of short (19-24 nt) RPFs in infected cells compared to mock-infected cells ($p = 0.03$), while there was no difference in the early timepoint group ($p = 0.52$). Within each timepoint, libraries were processed concomitantly, so significant differences between mock and infected libraries are not expected to arise from harvesting and library preparation procedures, suggesting this result may represent a genuine biological effect on translation. In yeast, hyperosmotic or oxidative stresses lead to a reduced proportion of short RPFs. For hyperosmotic stress, this effect was found to be due to inhibition of translation elongation and was dependent on Rck2, an effector in the high-osmolarity glycerol (HOG) pathway, which phosphorylates the elongation factor eEF2 under such conditions. Oxidative stress also led to eEF2 phosphorylation and inhibition of translation elongation, thought to be responsible for the decrease in short RPFs. The pathway responsible for this effect was unknown, although it was distinct from the pathway activated by hyperosmotic stress, as it did not depend on Rck2. Rck2 is homologous to calmodulin kinases in higher eukaryotes and the components of the yeast HOG pathway are similar to the mammalian p38 MAPK (also known as SAPK2) pathway, which is activated by extracellular stimuli such as growth factors or inflammatory cytokines. The p38 MAPK pathway converges alongside several other signalling pathways, responding to a range of...
Discussion and future work

stimuli/stresses, to regulate eEF2K, the only known protein in mammals that phosphorylates eEF2 (references 136,657–659). These results raise the possibility that, during the late stages of PRRSV infection, pathways may be activated that globally regulate translation elongation – a facet of translational regulation which is often overlooked in favour of studying initiation. Further investigation to confirm these findings, and investigate the potential pathways involved, would be of interest.

In the analysis of frameshifting, the occurrence of −2 PRF at the nsp2 site and −1 PRF at the ORF1ab site were clearly evident for both species of PRRSV and the efficiency of these frameshift events was quantified. The occurrence of −1 PRF at the nsp2 site was not readily discernible in these datasets, likely due to a combination of its low efficiency and the lack of an extended ORF in the −1 frame, in which changes in phasing could be used to assess PRF.

At the nsp2 frameshift site, −2 PRF efficiency for EU PRRSV was 26% at 8 hpi, while for NA PRRSV it was shown to increase from 23% at 6 hpi to ~40% at 9 and 12 hpi. These data reveal a rare example of temporally regulated PRF efficiency. There are several reasons this may be beneficial for viral fitness, related to the functions of the proteins expressed as a result of frameshifting at this site. Nsp2, nsp2TF and nsp2N all share a PLP2 domain which acts as a DUB, removing cellular ubiquitin and ISG15 conjugates to suppress innate immune pathways and IFN-β signalling531–534,536. The innate immune suppression is strongest for nsp2N, followed by nsp2TF, with nsp2 being the least potent536. Increasing the ratio of the frameshift products compared to nsp2 may therefore be beneficial at the mid-late stage of the replication cycle, when viral replication and sgRNA production is rapidly occurring and dsRNA, a potent activator of innate immune signalling660, has accumulated. Additionally, nsp2TF has been shown to down-regulate expression of SLA-I, an MHC class I gene involved in antigen presentation to cytotoxic T cells553. The requirement to suppress antigen presentation may also increase later in infection as viral proteins accumulate, potentially making increased levels of nsp2TF beneficial at these timepoints. Recently, a newly discovered role was characterised for nsp2TF. Unlike nsp2, it localises to the exocytic pathway, where it associates with the PRRSV structural proteins GP5 and M280,552. By virtue of the DUB activity of its PLP2 domain, nsp2TF reverses the ubiquitination of GP5 and M, preventing their degradation by the proteasome552. This is thought to promote virus assembly, for which GP5 and M are essential552,612. M and GP5 are detectable by immunofluorescence from 8 hpi and 10 hpi onwards, respectively552, consistent with the hypothesis that the increase in −2 PRF efficiency observed by 9 hpi serves to increase nsp2TF levels in order to stabilise GP5 and M and promote virus assembly.

At early timepoints, it may be beneficial for viral fitness if most ribosomes continue translation without frameshifting at the nsp2 PRF site, as this enables them to translate the remainder of pp1a or pp1ab, which encode components of the RTC (section 1.4.6.1.2.2). Early in the replication cycle,
transcription of these RTC components (including nsp2 itself) is likely to be more important for establishing infection than translation of the accessory protein nsp2TF. Later in the replication cycle, once RTC formation has occurred and infection has been established, the requirement for RTC components may decrease. This coincides with an increased requirement for nsp2TF to prevent degradation of GP5 and M and promote virus assembly, and potentially with an increased requirement for suppression of host immune responses. Therefore, at these later timepoints, it is likely to be relatively more advantageous to the virus if ribosomes frameshift at the nsp2 site, producing nsp2TF (or nsp2N) instead of translating the remainder of the polyprotein.

Surprisingly, frameshift efficiency was also found to increase over time at the ORF1ab site, with −1 PRF efficiency for NA PRRSV increasing from 11% at 6 hpi, to 19% at 9 hpi and then 32% at 12 hpi, while the EU PRRSV efficiency was 23% at 8 hpi. Frameshift sites which operate by a canonical, secondary-structure-directed mechanism, such as the nidovirus ORF1ab PRF site, are usually assumed to operate at a fixed efficiency due to the absence of trans-acting regulatory factors. However, features of translation other than trans-acting, PRF-specific factors could affect frameshift efficiency, for example the putative global perturbation of translation described above, and/or availability of elongation factors or aminoacyl-tRNAs. Some studies suggest that ribosome density has an impact on PRF efficiency. Lopinski et al. (2000) (reference 269) used global translation initiation inhibition as a strategy to reduce ribosome density and found this led to increased −1 PRF efficiency on the genome of the virus L-A in yeast. The authors’ favoured explanation for this result was that, on transcripts with high ribosome density, frameshift-stimulatory structures, which are unfolded by the ribosome, do not always have time to re-fold before the next ribosome arrives, meaning not all ribosomes encounter the folded frameshift-stimulatory structure. When the spacing between ribosomes increases, re-folding of the pseudoknot more often has time to occur before passage of the next ribosome, meaning there are more opportunities for frameshift events. Smith et al. (2019) (reference 270) performed similar experiments, in which initiation efficiency was modulated on a PRF reporter construct (based on the pseudoknot-stimulated IS3 PRF site) in E coli, and found that −1 PRF efficiency increased as ribosome density decreased. In the NA PRRSV data, the translation efficiency (RPF density normalised by transcript abundance) of ORF1a decreases over the course of infection (discussed in section 4.4); however, estimates of translation efficiency for ORFs expressed from PRRSV genomic RNA are likely confounded by the inability to differentiate between translatable RNA and RNA destined for packaging. Changes in ribosome density provide a potential reason for the observed changes in frameshift efficiency, but the mechanistic understanding of this relationship is far from clear and the field would benefit from detailed systematic studies on this topic.

The potential benefit to viral fitness of an increase in PRF efficiency at the ORF1ab site is unclear, if indeed there is a benefit. The proteins encoded in the −1 frame downstream of the frameshift site
are nsp9-12, which include the RdRp and viral helicase – key components of the RTC (section 1.4.6.1.2.2). The function of frameshifting at the ORF1ab site is often described as determining the stoichiometry of these proteins relative to the proteins encoded by ORF1a, however the full composition and stoichiometry of the arterivirus RTC is unknown. The RdRp is expected to be required in smaller quantities later in infection, as packaging takes precedence over genome replication\textsuperscript{285}. This is seemingly in contrast to the observed increase in PRF efficiency; however, the main advantage of temporal stoichiometric regulation may be alteration of the ratio of structural to non-structural proteins, as opposed to the ratio of RdRp compared to other non-structural proteins (as represented by changes in ORF1ab PRF efficiency). Further discussion of the possible benefits of temporal changes in the stoichiometry of ORF1b-encoded proteins can be found in Chapter 4. It is also important to note that the increase in $-1$ PRF efficiency may be incidental, and this is not proposed to represent “regulation” of PRF efficiency, which implies an evolutionarily conserved mechanism subject to purifying selection (such as the \textit{trans}-acting proteins at the nsp2 site).

Nonetheless, this finding has potential ramifications for the many other viruses which encode structure-directed PRF sites in their genomes\textsuperscript{143}, as PRF efficiency is usually measured at a single timepoint and assumed to be constant throughout infection. Some tentative support for temporal changes in PRF efficiency in the coronavirus murine hepatitis virus (MHV) is provided by similar quantifications of ORF1ab $-1$ PRF efficiency by ribosome profiling\textsuperscript{307}. Two replicates from 5 hpi were quantified, revealing frameshift efficiencies of 48% and 70%; however, the higher value was derived from a replicate which was determined (by other analyses) to have advanced further in infection than its counterpart, despite being harvested at the same timepoint\textsuperscript{307}. As both these and the PRRSV results were determined by ribosome profiling, which could be affected by changes in ribosome distribution unrelated to PRF (for example attenuated initiation depleting ribosomes from ORF1a), confirmation of these findings by another assay (see next section) is important.

### 3.6.2 Future work

Due to the complex, closely packed and overlapping arrangement of ORFs in the PRRSV genome, specifically testing whether the temporal regulation of nsp2 PRF efficiency confers a fitness advantage to the virus would be challenging. One possible strategy would be to replace the nsp2 PRF site with a frameshift site that is thought to operate at a fixed efficiency, for example a canonical, secondary structure-directed frameshift site. However, examples of efficient structure-directed $-2$ PRF are rare, with the additional requirement for facilitating $-1$ PRF adding a further complication. To generate a structure-directed PRF site that facilitated $-1$ and $-2$ PRF with efficiency comparable to the PRRSV nsp2 site would likely require considerable optimisation \textit{in vitro}. Even if a suitable structure-directed PRF site were generated, making the relevant mutations
to incorporate the element into the viral genome would likely have other ramifications for viral fitness. Additionally, as discussed above, structure-directed PRF sites do not necessarily operate at a fixed efficiency throughout the course of infection and increases in PRF efficiency such as those observed for the ORF1ab site could confound the analysis. Another potential strategy would be to transfect an nsp1β expression plasmid into cells before infection, with the aim of saturating the system with nsp1β from the beginning of infection, obviating the need to wait for it to accumulate, and (potentially) resulting in high but constant PRF efficiency throughout the timecourse. A negative control could be provided by transfection of a plasmid expressing a previously characterised mutant nsp1β, with key residues in the putative RNA-binding domain mutated to knock out PRF-stimulatory activity. However, due to the fact that the 123-GKYLQRRLQVNGLR-136 motif responsible for stimulating PRF overlaps the SAP motif (124-KXLQXXLXXXGL-135) responsible for nuclear imprisonment of host mRNAs (and consequent innate immune suppression), experiments would be required to disentangle these effects.

For both frameshift sites, a valuable experiment would be to use an alternative technique to corroborate the finding that frameshift efficiency increases over time. An important requirement of such an assay is that it is deployable in the context of infection as opposed to using a reporter construct. Previously, frameshift efficiency at the nsp2 site has been measured in the context of infection by radiolabelling PRRSV-infected MARC-145 cells, immunoprecipitating nsp2, nsp2TF and nsp2N (using an antibody which binds their shared N-terminus), separating the proteins by gel electrophoresis, and quantifying their relative proportions. This would be a good option, although it is expected to be less accurate than RiboSeq, due to the poorer temporal resolution (measuring protein accumulation over the one hour radiolabelling period) and potential confounding effects of different turnover rates of the proteins assayed. The latter is a particularly pertinent concern as nsp2TF was shown to be turned over much more rapidly than nsp2 in radiolabelling pulse-chase experiments, with even greater instability observed for nsp2N.

For the nidovirus ORF1ab PRF site, quantification of PRF efficiency in the context of infection has not been performed using assays other than ribosome profiling. This is likely due to the fact that the complex proteolytic processing pathways lead to difficulty in interpreting results based on quantification of bands, not only in terms of reliable identification of products but also the potential for differential processing rates to confound results. Immunoprecipitation using an antibody against the shared N-terminus of nsp8 (0 frame) and nsp9 (−1 frame) would simplify the interpretation of radiolabelled gels, and would be necessary to facilitate identification of viral bands against the background of host proteins, although differences in processing and turnover are still a potential problem. Despite these potential limitations, this assay has the advantage that frameshift efficiency quantifications would not be affected by RNP formation or translation of non-canonical sgRNAs.
A simpler assay using a transfected PRF reporter construct based on the PRRSV ORF1ab PRF site would be suitable to test whether the observed increase in −1 PRF efficiency could potentially result from a decrease in ribosome density on the PRRSV genomic RNA over time. Mutations could be made in the Kozak sequence of the reporter to decrease initiation efficiency, similar to the strategy employed by Smith et al. (2019) (reference 270).

Although the majority of investigations into PRRSV biology are performed using MARC-145 cells, derived from African green monkey kidney, it would be beneficial to verify these findings in a more physiologically relevant cell line. As discussed in section 1.4.3 and section 3.2.1, porcine alveolar macrophages, the primary cell type infected in vivo, are not suitable for experiments requiring large numbers of cells or highly efficient infection (such as ribosome profiling). However, some additional physiologically relevant cell lines have become commercially available recently. ZMAC cells, an immortalised PAM cell line which is ~80% permissive for PRRSV infection, are a promising candidate if any of the ribosome profiling experiments were to be repeated. This comment applies to both this Chapter and Chapter 2; however, it is particularly relevant for the analysis of frameshifting at the nsp2 site, because the amount of PCBP1 and PCBP2 may differ between MARC-145 cells and porcine macrophages. The relative amounts of these two host proteins has been shown to affect the ratio of −1 and −2 PRF at the PRRSV nsp2 frameshift site in vitro, with PCBP1 favouring −2 PRF and PCBP2 favouring −1 PRF. With this in mind, all the experiments described above could be performed in ZMAC cells as opposed to MARC-145 cells.

3.6.3 Conclusion

These results represent the first investigation of arterivirus PRF efficiency over a timecourse of infection. At the nsp2 frameshift site, −2 PRF efficiency was found to increase and then plateau, consistent with the observed accumulation of the viral frameshift-stimulatory protein, nsp1β. This is the second known example of temporally regulated frameshifting, raising the possibility that temporal regulation of non-canonical translation events may be a more ubiquitous feature of viral gene expression than previously considered. Furthermore, an increase in −1 PRF efficiency was found at the ORF1ab site, suggesting that frameshift efficiency at RNA structure-directed sites may be sensitive to infection-induced changes in the intra-cellular translational landscape. This has potential ramifications for our understanding of gene expression in a range of viruses.
Chapter 4: Analysis of PRRSV gene expression

4.1 Introduction

In addition to the production of the canonical sgRNAs described in section 1.4.6.2.1, evidence of a much more complex and heterogeneous arterivirus transcriptome has accumulated in recent years. To start with, low throughput studies using techniques such as RT-PCR led to the identification of a handful of alternative transcripts, in which alternative TRS bodies were used to produce sgRNAs from which the canonical structural ORFs were thought to be expressed. These alternative transcripts were thought to provide a back-up in case the main TRS body loses functionality and to contribute to determination of viral protein stoichiometry. An example of this is the N transcript, which was found to have two abundant transcript variants in some NA PRRSV strains, with 5′ UTRs differing in length by over 100 nt.

Similar studies identified a new class of sgRNA, termed heteroclite sgRNAs, present in several different strains of NA PRRSV under various culture conditions. These share many characteristics with canonical sgRNAs, but the site of polymerase jumping is within ORF1a rather than at the TRS leader. Instead of the canonical TRS UUAACC sequence, RdRp jumping appears to be directed by short regions (2-12 nt) of sequence similarity between ORF1a and the 3′-proximal region of the genome that encodes structural proteins or the end of ORF1b. Other than the sequence similarity between the donor and acceptor sites of these junctions, no consensus sequence motif or sequence features were identified that rationalise the jump sites used at different heteroclite sgRNAs, although the involvement of secondary RNA structure was suggested. The function (if any) of these transcripts is unknown, but they are packaged into the virion, and are thought to express a short region of the ORF1a polyprotein, in many cases fused with a protein sequence encoded by the ORF1b or sgRNA region (all fusions characterised thus far occur between the region upstream of the nsp2 frameshift site and the region downstream of nsp9).

The popularisation of RNASeq led to much greater capacity and sensitivity in the search to characterise non-canonical viral transcripts. An application of this technique to SHFV increased the number of known utilised TRS bodies from ~10 to 96, including the identification of a small number of low-abundance sgRNAs for which the TRS body was in ORF1ab. This study, supported by numerous similar applications of RNASeq to other nidoviruses (especially CoVs), highlights that the PRRSV transcriptome may have unexplored complexity.
Analysis of PRRSV gene expression

Similarly, the advent of ribosome profiling has revealed hidden complexity in both host and viral translatomes, with many studies of virus-infected cells revealing translation of previously uncharacterised ORFs, some of which are important for viral fitness.\textsuperscript{307,375,385,386,666,671}

Given the possibility that RNASeq and RiboSeq could similarly reveal novel PRRSV transcripts and ORFs, the datasets from Chapter 3 were probed here to characterise the viral transcriptome and translatome, aiming to address the following questions:

1. Does PRRSV express transcripts other than the known canonical transcripts?
2. Does PRRSV express ORFs other than the known canonical ORFs?
3. What is the expression level of viral transcripts and ORFs over the course of PRRSV infection?
4. Is there any evidence for novel mechanisms regulating viral gene expression?
4.2 Characterising the PRRSV transcriptome

4.2.1 Pipeline to detect chimeric viral transcripts

A pipeline was written to detect viral transcripts, such as sgRNAs, which align discontinuously to the viral reference genome, here referred to as “chimeric” transcripts or reads. This pipeline (described in section 2.4.3) uses STAR, a splice-aware mapper, in contrast to the core pipeline (section 2.4.1.1) which uses bowtie and therefore does not permit discontinuous alignments.

Many junction sites occur in positions where there is similarity between the genome sequence at the donor and acceptor sites (Figure 4.1). As such, it is not usually possible to determine the exact position at which the jump occurred, and the location of the junction could be assigned either to one end of the stretch of similarity or the other. Furthermore, in sgRNA junctions where it was possible to discern, it was shown for many nidoviruses that the exact RdRp jump point within the TRS body was often heterogeneous. To account for this technical and biological heterogeneity, the junctions reported by STAR were clustered and junctions with both donor and acceptor sites in very close proximity to those of other junctions were merged together (see section 2.4.3 for further details). The aim of this was to group together junctions which were likely to be functionally equivalent to aid interpretation of the results; however, it should be noted that this leads to some instances where junctions with slightly different intron sizes are merged together and treated as equivalent. For TRS-spanning junctions (which generate sgRNAs) this is not expected to significantly affect interpretation of the results, as translation begins downstream of the junction. For non-TRS-spanning junctions (referred to herein as deletions), this strategy means that, in some cases, some junctions within a merged cluster cause a change in frame after the junction site while others do not (i.e. the merged junctions are not functionally equivalent). In cases where deletions such as these are discussed, this is clarified.

The above method was used to identify junction sites within each RNASeq library individually. Then, junctions were filtered to discard junctions which were unique to a single library within the timepoint, as these likely represent spurious transcripts. Junctions were defined as present in multiple replicates (where WT and KO2 libraries were treated as equivalent) if the donor clusters overlap and the acceptor clusters overlap. Matching junctions from all replicates within each timepoint were merged into one, resulting in a single set of junctions per timepoint. This final (merged, filtered and merged) result is presented in all plots and discussion in this section (4.2), unless specified. To ensure this filtering and merging strategy did not lead to loss of positional resolution due to clustering junctions from overly wide regions together, the widths of merged junction donor and acceptor sites were assessed (Figure 4.2), and the mean and median junction width for all analyses was found to be less than 3 nt.
Figure 4.1 Schematic of a junction-spanning (chimeric) read

The dotted portion of the chimeric read indicates a deletion in the read with respect to the genome (referred to in this work as an intron). A biological event fusing the beginnings of both regions of sequence similarity on the genome (red) would lead to an identical chimeric sequence compared to an event fusing the ends of the regions of similarity, i.e. the exact break point cannot be inferred from the chimeric read in such cases. To avoid ambiguity, herein I define the “donor” and “acceptor” sites of a junction as the 5′-most and 3′-most positions of the inferred intron (dotted line) respectively, where “5′-most” and “3′-most” refer to orientation with respect to the positive-sense genome, not with respect to the chimeric read (which differs according to sense). The terms “donor” and “acceptor” are used for convenience, although they are technically incorrect as (at least the majority of) nidovirus discontinuous transcription is thought to occur during synthesis of the negative strand, meaning the RdRp encounters first the “acceptor” and then the “donor” site.

Figure 4.2 Box plot of the widths of merged junctions

Box plots represent donor and acceptor site widths of junctions in the specified group of libraries, after merging junctions within each replicate and then merging matching junctions across all replicates from that timepoint. For NA PRRSV, 12 hpi was chosen because the most chimeric reads were detected at this timepoint and therefore the largest clusters were expected. Note that, for NA PRRSV, the widest merged TRS junction is the junction for the canonical M sgRNA, which was generated by a custom merge, as a special case due to having a stretch of 6 nt upstream of the TRS body that is identical to the stretch upstream of the TRS leader (see section 2.4.3 for further details).
4.2.2 Overview of results

The NA PRRSV transcriptome was characterised at all timepoints (Figure 4.3, Figure 4.4 and Figure 4.5), with considerably more chimeric reads found at later timepoints. The EU PRRSV transcriptome was characterised at 8 hpi (Figure 4.6). The EU libraries were prepared by purifying RNA fragments of lengths 25-34 nt, compared to a ~50 nt fragment length for all NA PRRSV libraries except 9 hpi replicate one libraries. This is an important consideration for the chimeric read analysis, because a 12 nt region of alignment is required either side of a junction for a read to pass the mapping filters, meaning a smaller proportion of chimeric reads are expected from libraries (such as the EU PRRSV libraries) with these shorter read lengths. In addition, as the 5’ and 3’ ends of junction-spanning reads must fall within a narrower window to qualify for mapping, the effect of technical biases such as ligation bias is expected to be greater for shorter reads than for longer reads (over which this effect is diluted by averaging out). To aid comparison between EU and NA PRRSV, the pipeline described above was repeated using only the NA PRRSV 9 hpi replicate one libraries, which also had 25-34 nt fragments purified (Figure 4.7). This is in addition to the inclusion of these libraries in the full 9 hpi NA PRRSV analysis in Figure 4.4.

Reads were categorised as TRS-spanning (sgRNAs) or non-TRS-spanning (deletions), according to whether the donor site of the junction overlapped the TRS leader. Within the TRS-spanning category, transcripts were classified as canonical if they utilised the known, major TRS body for expression of the structural protein (for the N transcript, the previously characterised alternative TRS body for VR-2332 was denoted as canonical in addition to the major TRS body). Within the non-TRS-spanning category, deletions were classified as local or distant according to intron size (local $\leq$ 2000 nt, distant $>$ 2000 nt). The dataset resulting from this analysis contains over 300 junction sites, making it impractical to further investigate all of them, and the discussion below focuses on junctions which have the potential to significantly affect overall gene expression profiles.
Figure 4.3 The NA PRRSV transcriptome at A) 3 and B) 6 hpi

Sashimi plot of junctions in the NA PRRSV dataset at early timepoints during infection. The number of reads spanning each junction is indicated by the highest point of its arc (note the logarithmic scale of the y axis) and represents the total number of reads spanning the junction in all libraries from the specified timepoint combined. Only junctions for which this number is $\geq 2$ are plotted. Beneath the sashimi plot is an inverted bar chart (black) of the proportion of reads at each donor and acceptor site that span the junction of interest, calculated as junction-spanning / (junction-spanning + continuously aligned to reference genome). At both timepoints, no junctions where the donor site overlapped the TRS leader passed the minimum read count threshold for plotting. Internal deletions, in which the donor site does not overlap the leader TRS, are coloured according to whether the deletion is distant (> 2000 nt deleted, grey) or local (≤ 2000 nt deleted, red). At 6 hpi, all junctions plotted were classified as local deletions.
Figure 4.4 The NA PRRSV transcriptome at 9 hpi

A) Sashimi plot of junctions for which the donor site overlaps the leader TRS. Plot constructed as in Figure 4.3 but with the major junction for each canonical sgRNA shown in blue (including both N-long and N-short for the N sgRNA), and other junctions (“non-canonical”) shown in orange. The threshold for inclusion of junctions on this plot is ≥100 junction-spanning reads in total from all 9 hpi libraries (note that eight libraries were analysed at this timepoint compared to four at other timepoints). B) Sashimi plot of junctions representing internal deletions in the NA PRRSV genome at 9 hpi. Plot constructed as in Figure 4.3, except that the threshold for inclusion of junctions in the plot was adjusted to ≥100 supporting reads.
Figure 4.5 The NA PRRSV transcriptome at 12 hpi

A) Sashimi plot of junctions for which the donor site overlaps the leader TRS. Plot constructed as in Figure 4.4A except that the threshold for inclusion of junctions was adjusted to ≥ 50 supporting reads. B) Sashimi plot of junctions representing internal deletions, in which the donor site does not overlap the leader TRS. Plot constructed as in Figure 4.3, except that the threshold for inclusion of junctions in the plot was adjusted to ≥ 50 supporting reads.
Figure 4.6 The EU PRRSV transcriptome at 8 hpi

A) Sashimi plot of junctions for which the donor site overlaps the leader TRS. Plot constructed as in Figure 4.4A except that the threshold for inclusion of junctions was adjusted to ≥ 10 supporting reads. B) Sashimi plot of junctions representing internal deletions, in which the donor site does not overlap the leader TRS. Plot constructed as in Figure 4.3, except that the threshold for inclusion of junctions in the plot was adjusted to ≥ 10 supporting reads.
Figure 4.7 The NA PRRSV transcriptome at 9 hpi, replicate 1 only

A) Sashimi plot of junctions for which the donor site overlaps the leader TRS. Plot constructed as in Figure 4.4A except that the threshold for inclusion of junctions was adjusted to ≥ 10 supporting reads in total from the two libraries (WT and KO2). B) Sashimi plot of junctions representing internal deletions, in which the donor site does not overlap the leader TRS. Plot constructed as in Figure 4.3, except that the threshold for inclusion of junctions in the plot was adjusted to ≥ 10 supporting reads.
4.2.3 TRS-spanning junctions

A large proportion of the junctions returned by the above pipeline are TRS-spanning junctions with acceptor sites within the region of the genome that encodes structural proteins, which are expected to provide alternative transcripts for the canonical sgRNAs.

Many of these transcripts likely differ from their canonical counterparts mostly in the length of the 5’ UTR, as opposed to the identity of the protein expressed, and may contribute to refining the overall stoichiometry of structural proteins\textsuperscript{663}. The N transcript provides an example: for the NA PRRSV strain VR-2332, two highly abundant N transcript variants were previously characterised\textsuperscript{426}, which were also the two most abundant N transcripts in the NA PRRSV dataset (Figure 4.4A, Figure 4.5A and Figure 4.7A). In both studies, the major transcript variant, termed N-long in this thesis, was formed using a TRS body with sequence AUUACC (beginning at 14,761 on the SD95-21 genome), while the secondary transcript (N-short) was formed using a TRS body downstream (beginning at 14,875) with sequence UAAACC (Figure 4.8). These result in transcripts with 5’ UTRs of lengths 313 nt and 199 nt, respectively (where at least the first 184 nt of the UTR is shared by all viral transcripts). The major TRS body, N-long, is almost completely conserved amongst NA PRRSV isolates (differing in 3/661 sequences), while the N-short TRS body shows some sequence variation at the second base, suggesting N-long may be the more widely used TRS body across different NA PRRSV isolates (Figure 4.8A). This is supported by a study on a second NA PRRSV isolate (tw91) for which, like VR-2332 and SD95-21, N-long was the major transcript, but for which the second most abundant N transcript was formed not by using N-short, but by using a TRS body upstream of N-long\textsuperscript{426,662}. This upstream TRS body utilised by tw91 is present and utilised in the NA PRRSV dataset, but at very low efficiency compared to N-long and N-short (201 chimeric reads at 12 hpi compared to 42,039 for N-long and 7,136 for N-short).

**Figure 4.8 Alternative TRS bodies for the N sgRNA**

A) Consensus sequences of the N-long and N-short TRS bodies in NA and EU PRRSV, based on 661 and 120 genome sequences, respectively. B) The ratio of junction-spanning reads mapping specifically to N-long compared to N-short. Bars represent the mean, with individual data points plotted as crosses. At 6 hpi (NA PRRSV) no reads map to the short form so this timepoint was omitted.
The 114 nt difference in UTR length between N-long and N-short presents an opportunity for translation regulation, for example if certain features of the region unique to N-long promote or inhibit translation. The ratio of N-long to N-short remains approximately constant between 9 and 12 hpi in NA PRRSV (Figure 4.8B), in accordance with the finding that the abundance of different SHFV sgRNAs relative to each other remained constant during infection\(^{663}\). This indicates that the benefit of producing these two transcripts is not in temporally regulating their relative abundance, suggesting other possibilities. The extended region of the N-long 5' UTR may have additional regulatory elements not present in N-short, or there may be potential benefits in varying the distance between the N initiation codon and shared elements of the co-terminal 5' UTR such as RNA secondary structures or uORFs (see section 4.3). Interestingly, EU PRRSV differs considerably from NA PRRSV in terms of N transcript variation. In EU PRRSV isolates, N-short is universally conserved, and matches the UUAACC TRS leader sequence exactly, while N-long is an imperfect match to the TRS leader and is not as well conserved (Figure 4.8A). Consistent with this, N-short is by far the most abundant N transcript for EU PRRSV, with N-long only having 40 junction-spanning reads compared to 2,733 for N-short (Figure 4.8B). Only one other highly supported N transcript variant was detected in the EU PRRSV dataset (Figure 4.6A), which has a TRS body in between N-long and N-short, and this is similarly very lowly expressed (46 junction-spanning reads). This suggests that, if there is a fitness advantage to varying the length of the N 5' UTR, the mechanism may be specific to NA PRRSV and, indeed, specific to particular isolates.

As well as generating alternative transcripts for expression of the canonical structural proteins, many of the TRS-spanning junctions revealed by this analysis likely lead to expression of truncated forms of these proteins. One example of this is a previously characterised transcript variant for GP5, expressed by VR-2332 PRRSV (NA lineage) and termed “5-1” by Nelsen et al. (1999) (reference \(^{426}\)). This is predicted to express the C-terminal 69 amino acids of GP5, initiating at Met132 in the SD95-21 GP5 sequence. The same TRS body (UUAGCC, beginning at position 14,065) was utilised in the NA PRRSV dataset (junction donor-acceptor: 183-14062), with 370 junction-spanning reads at 12 hpi, which is approximately 1.7% of the number spanning the primary GP5 transcript. Although this particular 5-1 TRS body does not have a counterpart in the EU PRRSV dataset, almost all the structural protein ORFs in both datasets have novel junction acceptor sites downstream of the primary sgRNA junction, suggesting many of the PRRSV structural proteins may have truncated isoforms (panel A of Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7). This is corroborated by findings from SHFV, in which several GP5 transcript variants were found which were expected to express the same truncated form of GP5 as 5-1 (reference \(^{663}\)). Mutation of the putative initiator methionine reduced virus titre\(^{663}\), suggesting that expression of truncated forms of arterivirus structural proteins may have a fitness advantage, although the effect of this amino acid substitution on full-length GP5 could also be responsible for this phenotype.
In addition to the junctions which produce alternative transcripts of structural proteins, a small number of novel junctions were found which have acceptor sites within ORF1b (panel A of Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7). These are of particular interest because expression of ORF1ab is very low compared to the structural ORFs at late timepoints (section 3.4). Due to the massive abundance of the canonical sgRNAs, the transcript variants discussed in the previous paragraphs likely only contribute a small proportion of the total production of the encoded structural protein, despite the fact that the transcript variants are also reasonably abundant. ORF1ab, on the other hand, is (canonically) expressed only from the genomic RNA, which is much less abundant than sgRNAs at late timepoints and is thought to be translated inefficiently (section 4.4). As such, non-canonical sgRNAs that facilitate expression of parts of the polyprotein could have a much bigger proportional impact on protein production than alternative structural ORF transcripts, even if the novel ORF1b transcripts are present at relatively low abundance. These transcripts are henceforth referred to as ORF1b sgRNAs, and were identified in both NA and EU PRRSV, with (at least) eight and two highly supported junctions, respectively (panel A of Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7). These novel transcripts are investigated and discussed further throughout the rest of this chapter.

4.2.4 Non-TRS-spanning junctions

Junctions which do not span the TRS leader tend to have many fewer junction-spanning reads than TRS-spanning junctions, so they are expected to have a smaller effect on viral gene expression. This likely reflects the different mechanisms by which these two categories of junction are formed. While TRS-spanning junctions result from a highly active discontinuous transcription mechanism which is essential for virus viability, deletions likely arise from alternative, less efficient mechanisms.

Infection with RNA viruses is commonly found to promote recombination between viral genomes when viruses are passaged at high MOI, resulting in a deletion in the recombined genome \(665,672-675\). The resultant RNA can be a functional viral genome, a dead-end product which is not replicated and/or packaged, or a defective interfering (DI) RNA. DI RNAs are viral genomes with deletions that prevent independent replication, but still maintain all the necessary \(cis\)-acting elements for genome replication and packaging\(675\). When a “helper” virus (such as the parental, WT virus) is present, these DI RNAs can be replicated and packed into virions, often interfering with viral fitness by competing with WT genomes for replication proteins and reducing the average fitness of the total pool of virions \(665,673,675\). Many of the deletions observed in the NA and EU PRRSV datasets likely represent DI deletions.

Heteroclite sgRNAs (section 3.1) represent a class of deletion originally thought to be DI RNAs, but reclassified as heteroclite (“deviating from common forms or rules”) sgRNAs \(619,664\). They are
reported to differ from DI genomes in that they do not interfere with gRNA or canonical sgRNA production, and they are present in a range of conditions, including low MOI passage and virus isolates directly taken from the field\textsuperscript{619,664}. The formation of heteroclite sgRNAs is thought to be directed by sequence similarity between the donor and acceptor sites\textsuperscript{619,664,665}. Many of the long-range deletions observed at 12 hpi in the NA PRRSV dataset could represent heteroclite sgRNAs (Figure 4.5B). The most abundant deletion at this timepoint is 1747-13894, which generates a heteroclite sgRNA that was previously characterised for VR-2332, in which it was called “S-2”\textsuperscript{619,664}. There is a 6 nt region of sequence similarity (GCUCCC) between the donor and acceptor sites\textsuperscript{619,664}. Many of the long-range deletions observed at 12 hpi in the NA PRRSV dataset could represent heteroclite sgRNAs (Figure 4.5B). The most abundant deletion at this timepoint is 1747-13894, which generates a heteroclite sgRNA that was previously characterised for VR-2332, in which it was called “S-2”\textsuperscript{619,664}. There is a 6 nt region of sequence similarity (GCUCCC) between the donor and acceptor sites\textsuperscript{619,664}. Many of the long-range deletions observed at 12 hpi in the NA PRRSV dataset could represent heteroclite sgRNAs (Figure 4.5B). The most abundant deletion at this timepoint is 1747-13894, which generates a heteroclite sgRNA that was previously characterised for VR-2332, in which it was called “S-2”\textsuperscript{619,664}. There is a 6 nt region of sequence similarity (GCUCCC) between the donor and acceptor sites\textsuperscript{619,664}. Many of the long-range deletions observed at 12 hpi in the NA PRRSV dataset could represent heteroclite sgRNAs (Figure 4.5B). At 12 hpi, this junction is fairly abundant, with 312 junction-spanning reads (Figure 4.5B). To infer how much junctions are likely to affect overall gene expression levels, the proportion of donor- or acceptor-site-spanning reads which span the junction of interest was quantified, as a proxy for determining the proportion of transcripts likely to contain the junction. Note that, in these calculations, the total number of reads spanning each site was defined as reads spanning the junction of interest plus reads mapping continuously to the reference genome in that region (so in the sgRNA region the denominator will usually include reads originating from other sgRNAs). For the S-2 heteroclite sgRNA, 2.3% of reads at the donor site were junction-spanning reads mapping to 1747-13894, indicative of the approximate relative proportion of this heteroclite sgRNA to gRNA (Figure 4.5B). This is a relatively small proportion of reads, but it could be enough to have a noticeable effect on the total level of nsp1\(\alpha\) and nsp1\(\beta\). For comparison, S-2 transcript abundance is only an order of magnitude lower than that of the least abundant canonical sgRNA, Gp4 (3,784 junction-spanning reads, representing 11.6% of donor-site-spanning reads). Further, if gRNA is translated less efficiently than sgRNAs (section 4.4) the effect of transcript abundance on translation could be magnified. Consistent with the hypothesis that S-2 significantly contributes to translation, the RPF density in the region upstream of the 1747-13894 junction (donor site approximately 160 nt downstream of the large peak in nsp2) is considerably greater than the RPF density downstream (Figure 3.16 and Figure 3.17).

A similar profile of increased density in the region upstream of the S-2 heteroclite junction can be seen at earlier timepoints (Figure 3.16 and Figure 3.17). Although it is below the threshold for inclusion on the plot in Figure 4.4, the 1747-13894 junction was detected at 9 hpi, with 28 junction-spanning reads in total from three replicates. It was not detected in any libraries at 6 hpi (even disregarding the filter for presence in multiple replicates), but it was detected in both KO2 libraries at 3 hpi, with two junction-spanning reads in each library (Figure 4.3). This very early detection is consistent with the observation that this (and other heteroclite sgRNAs) are packaged into the virion\textsuperscript{619,664}, and it is likely that the transcript is similarly present in the samples harvested at 6 hpi,
but was not sampled in the population of sequenced reads due to its low abundance. Additionally, many other long-range deletions have donor and acceptor sites in a similar region and are promising candidates for heteroclite sgRNAs (Figure 4.5B). These junctions were not all systematically assessed, but taking the second most abundant long-range deletion as another example (junction 1068-13949 which has 208 junction-spanning reads), it was found to have a 7 nt region of sequence similarity (UUGGCUA) between the donor and acceptor sites, indicating it also represents a heteroclite sgRNA. This, and the other long-range deletions, most of which have donor sites at or upstream of the S-2 donor site (Figure 4.5B), could similarly contribute to increased translation of the 5′-proximal region of ORF1a. The changes in RPF density observed in Figure 3.16 and Figure 3.17 represent the first evidence that PRRSV heteroclite sgRNAs are translated in the context of infection and suggest that this makes a significant contribution to production of the proteins encoded by the N-terminal region of ORF1a. This raises the possibility that this putative stoichiometry modulation is important for viral fitness. However, this is currently untested, and in vitro expression of the protein encoded by an example heteroclite sgRNA (not S-2) revealed that cleavage at the PCP1 site to free nsp1α was inefficient, highlighting that the precise functionality of the proteins encoded by heteroclite sgRNAs is currently unclear.

Heteroclite sgRNAs have only been characterised for NA PRRSV, with no studies aiming to determine if they are present in EU PRRSV infection. No suitable junctions were found in the EU PRRSV dataset (Figure 4.6B). The only distant deletion that passed the filters (71-6657) has a donor site within the 5′ UTR and is therefore unlikely to be packaged, as the packaging signal is thought to be a 35 nt stem loop located within the region encoding nsp1α (genomic coordinates 372-406 in the EU PRRSV strain used in this thesis). However, putative heteroclite sgRNAs become more abundant at later timepoints for NA PRRSV, and they may be detectable for EU PRRSV if a later timepoint were assayed, and/or if longer reads were purified, which are more amenable to chimeric read analysis.

4.2.5 Novel transcripts that might artificially inflate ORF1ab PRF efficiency

The chimeric read analysis described above revealed multiple novel transcripts that potentially have the capacity to result in translation of incomplete parts of the ORF1ab polyprotein. This is potentially relevant to the calculation of −1 PRF efficiency at the ORF1ab site (section 3.5.3), which assumes that all ribosomes in the “downstream” region initiated at the ORF1a initiation codon, passed through the “upstream” window of measurement, then passed through the ORF1ab frameshift site to reach ORF1b. Non-canonical transcripts which subvert this assumption could alter the calculated PRF efficiency, which is a particular concern if the abundance of such transcripts relative to gRNA were to increase over time, as is the case for canonical sgRNAs (section 3.4).
Figure 4.9 Novel transcripts that could potentially inflate calculated ORF1ab frameshift efficiency for NA PRRSV at 12 hpi

A) Novel junctions that could potentially inflate ORF1ab frameshift (FS) efficiency values from calculations using the ORF1ab downstream region depicted (coordinates in Appendix section 3), ending near the end of nsp9 (results from these calculations described in section 3.5.3). Plot constructed as in Figure 4.3, with junctions which span the leader TRS plotted in green, and those which do not plotted in pink. The “Proportion chimeric reads” panel is plotted on the same scale as for all other junction plots to facilitate comparison, however the junctions displayed here have very low proportions of chimeric reads and therefore the bars are not clearly visible. Junctions were identified as having the potential to inflate ORF1ab FS efficiency calculations if the acceptor midpoint was within the genomic coordinates 7703-9159 (after the end of nsp8, up to the end of the “downstream” region). For non TRS-spanning junctions the donor midpoint was also required to be upstream of genomic coordinate 7699.
(the end of nsp8). **B)** Novel junctions that could potentially inflate ORF1ab frameshift efficiency values from calculations using the truncated (trunc) ORF1ab downstream region depicted (coordinates in Appendix section 3). Plot constructed as in A but with inclusion criteria adjusted so the junction acceptor midpoint was required to be within the genomic coordinates 7703-8482. Junctions with at least 10 supporting reads from all libraries combined are annotated with the intron coordinates (donor-acceptor, coordinates indicate deleted bases) and an assessment of their potential to inflate frameshift efficiency calculation values.

ORF1b sgRNAs are an example of a junction that could artificially inflate PRF efficiency and, as such, the “downstream” region was designated to avoid abundant ORF1b sgRNAs (> 90 junction-spanning reads in total at 12 hpi) in the analysis presented in Chapter 3. However, to avoid excessive truncation of the downstream region, some junctions were allowed to remain. Figure 4.9A shows novel junctions at 12 hpi that have the potential to inflate the calculated ORF1ab frameshift efficiency by enabling ribosomes to access the downstream region without passing through the frameshift site. These junctions are all present at low abundance; however, there are still some ORF1b sgRNAs included in the downstream region, which may be temporally regulated. As such, further truncation of the region was performed to assess whether the trend of increasing PRF efficiency was still significant when the downstream region was truncated to avoid the ORF1b sgRNA acceptor sites.

Of the deletions with acceptor sites remaining in the truncated region (Figure 4.9B), those with at least ten junction-spanning reads at 12 hpi were investigated for their potential to inflate ORF1ab frameshift efficiency (Figure 4.9B, annotations). Of these four junctions, two were deemed likely to inflate PRF efficiency, with 42 junction-spanning reads in total. This indicates the effect of novel transcripts on ORF1ab −1 PRF efficiency calculations in the truncated analysis is likely to be very minor, and when the analysis was repeated the trend was still significant Figure 4.10, indicating it is not caused by novel viral transcripts.
Figure 4.10 ORF1ab –1 PRF efficiency calculated using the truncated downstream region

Percentage frameshift efficiency at the ORF1ab site, calculated using truncated ORF1ab downstream region depicted in Figure 4.9B. Apart from the difference in downstream region, frameshift efficiency calculations were performed as described in section 3.5.3. Bars represent the mean results for each timepoint, calculated from WT and KO2 data combined. The individual datapoints are also plotted, with WT (cross) offset to the left and KO2 (triangle) offset to the right to aid visualisation. The right-hand panel shows the results of applying these calculations to RNASeq reads as a control, for which the expected result is 100% frameshift efficiency. * p < 0.05
4.3 Characterising the PRRSV translatome

4.3.1 Detection of novel ORFs using PRICE

PRICE (Probabilistic inference of codon activities by an EM algorithm) uses an annotated reference genome to infer the characteristics of RPFs in the input dataset, such as the heterogeneity in the distances between the 5′ end of the read and the ribosomal P site\(^ {311} \). The calculated values of these characteristics provide input parameters for a model of the relationship between a translated ORF and the resultant population of reads. This model is then applied to putative novel ORFs and an expectation maximisation (EM) algorithm is used to determine the codon (frame) assignments that have the highest likelihood of leading to the observed population of reads, which is designated as the candidate ORF\(^ {311} \). PRICE is capable of processing multiple libraries simultaneously but, due to the considerable memory requirements, RiboSeq libraries in the noCHX NA PRRSV dataset were run as sets of four (6 hpi, 9 hpi WT, 9 hpi KO2 and 12 hpi). For the EU dataset (in which RNP contamination is not evident), all read lengths were used as input for PRICE, while for the noCHX NA PRRSV dataset, read lengths with minimal RNP contamination were used. Viral ORFs with \( p \leq 0.05 \) (not corrected for multiple testing) were selected for inclusion in the output tables (Appendix section 5) and further analyses. Note that novel ORF names were designated for NA and EU PRRSV separately and ORFs which have the same names are not necessarily equivalent.

4.3.2 Novel viral ORFs

For NA PRRSV, 19 novel ORFs were detected (Figure 4.11), and nine were detected for EU PRRSV (Figure 4.12). The majority of these (ten for NA PRRSV and six for EU PRRSV) overlap ORF1b and many are likely expressed off the novel sgRNAs detected in section 4.2, with TRS bodies shortly upstream of the novel ORFs (Figure 4.11B-C, Figure 4.12B). Although this combination of novel ORFs and transcripts presents an opportunity for PRRSV to evolve overlapping ORFs encoding functional proteins, none of the ORFs in question are subject to purifying selection (Figure 4.11A and Figure 4.12A, \( p \) value plots), indicating that they are unlikely to be functional. Despite this, some of them are highly translated, with nsp12-iORF2 translated to a similar extent to nsp12 at 12 hpi (Figure 4.11B-C). This is a particularly interesting example as the overlapping ORF is surprisingly long, at 125 codons in the NA PRRSV strain used in this study. However, the ORF is not conserved in other isolates of NA PRRSV – it is often truncated and has poor conservation of the AUG initiation codon (Figure 4.11D). A hypothesis for why PRRSV tolerates such high levels of translation of presumably spurious ORFs is that these are side effects of production of the ORF1b sgRNAs, whose main function may be to regulate expression of the in-frame products of ORF1b. This is explored further throughout the rest of this chapter.
Figure 4.11 Overview of novel ORFs in the NA PRRSV translatome

A) Locations of novel ORFs on the NA PRRSV viral genome. The genome map of canonical ORFs is reproduced, without labels, in colours and offsets on the y axis corresponding to ORF frame relative to ORF1a. Novel ORFs discovered in this study are overlaid in red, offset on the y axis according to frame relative to ORF1a. Below the genome map, an analysis of synonymous site conservation in canonical protein-coding regions from 137 NA PRRSV genomes representative of the diversity of the phylogeny (section 2.4.4) is shown (performed by Dr Andrew Firth using SYNPLOT2 (reference 649)). The green line represents the probability (over a 25 codon sliding window) that the observed conservation could occur under a null model of neutral evolution at synonymous sites. Note that the UTRs are excluded because this analysis requires an annotated ORF to overlap the feature of interest. B) Translation of novel ORFs overlapping ORF1b. Reads mapping to the NA PRRSV viral genome between nsp9-iORF1 and the end of ORF1b are shown, separated according to phase.
and plotted after application of a 15 codon running mean filter. Only read lengths identified as having minimal RNP contamination (indicated in Figure 3.8) were used to generate this plot. Novel ORFs in this region are indicated on the genome map, coloured according to frame relative to ORF1a. Positions of moderately frequently used TRS bodies (at least 44 junction-spanning reads) that likely facilitate translation of these ORFs are indicated by red bars at the bottom of the plot. The height of each red bar is scaled according to the number of junction-spanning reads for that TRS body (in all 12 hpi RNASeq libraries combined), relative to the canonical sgRNA with the fewest junction-spanning reads (Gp4). “abund.” = abundance (number of junction-spanning reads relative to Gp4). C) Translation of NA PRRSV novel ORFs overlapping ORF1b, for libraries not shown in panel B. D) Conservation of nsp12-iORF2. Main: Predicted amino acid sequences of nsp12-iORF2 from 137 NA PRRSV genomes representative of the diversity of the phylogeny. The nsp12-iORF2 CDS nucleotide sequence was extracted from a multiple sequence alignment and frame 0 was translated. Each row represents one sequence, with each coloured rectangle representing an amino acid. Gaps indicate translation is predicted to have terminated due to a stop codon. Inset: The conservation of the context of the predicted nsp12-iORF2 initiation codon, based on 661 available sequences for full NA PRRSV genomes. The initiator AUG is indicated by a black box.
Figure 4.12 Overview of novel ORFs in the EU PRRSV translatome

A) Locations of novel ORFs on the EU PRRSV viral genome. Plot constructed as in Figure 4.11A, with synonymous site conservation analysed for 120 available EU PRRSV sequences (performed by Dr Andrew Firth).

B) Translation of EU PRRSV novel ORFs overlapping ORF1b. The region from nsp10-iORF1 to the end of ORF1b is shown, and the plot is constructed as in Figure 4.11B, with TRS bodies with over 10 junction-supporting chimeric reads included, and TRS body bar heights representing junction-spanning reads (for both libraries combined) scaled relative to the canonical sgRNA with the fewest junction-spanning reads (Gp3). Only read lengths with good phasing (indicated in Figure 3.10E) were used to generate this plot.

Overlapping ORFs were also found in the canonical sgRNA region and in ORF1ab upstream of the ORF1b sgRNAs. For those in the sgRNA region (only detected for NA PRRSV), there are several sgRNAs with TRS bodies downstream of each canonical ORF that could potentially facilitate expression of most of the small overlapping ORFs. For the ORF1ab-overlapping ORFs which begin upstream of the 5′-most ORF1b TRS body (detected in both NA and EU PRRSV), the potential mechanism of expression is unclear. These ORFs, which tend to be short and poorly expressed, may be erroneous PRICE annotations, or it may be that deletions or novel sgRNAs which are below the limit/threshold for detection by the chimeric read analysis pipeline (section 4.2) facilitate their expression. Similarly to the ORFs described above, none of the ORFs overlapping the canonical structural ORFs or ORF1ab are subject to purifying selection (Figure 4.11A and Figure 4.12A, \( p \) value plots).
For both NA and EU PRRSV, at least one uORF was detected in the 5’ UTR. In NA PRRSV, three uORFs were detected, with uORF1 being by far the most convincing and highly expressed. This uORF is very highly expressed at all timepoints (Figure 3.12C and Figure 3.14; blue peak in 5’ UTR), with RPF densities in the same range as that of the most highly expressed canonical ORF at each timepoint (ORF1a at 6 hpi, N at 9 and 12 hpi). To verify that expression of this uORF is genuine, the RPF distribution in the region of the 5’ UTR encoding uORF1 was plotted for the NA PRRSV noCHX 12hpi-WT-1 library and for the library in the NA PRRSV CHX dataset, CHX-9hpi-WT (Figure 4.13A). While samples harvested without CHX pre-treatment typically have heightened termination peaks, CHX pre-treated samples tend to have heightened initiation peaks (section 1.3.2.1), thus these two datasets serve to emphasise these two important features of putative novel ORFs (as exemplified here for uORF1). Initiation site profiling (using selective initiation inhibitors such as harringtonine) is a more robust method for determining the location of initiation codons; however, given that CHX pre-treated lysates were already available I opted to prepare the CHX NA PRRSV library to provide tentative support for the designation of initiation codons, instead of generating new, harringtonine-treated lysates. Note that CHX pre-treatment can increase expression of uORFs313, so inferences about relative levels of uORF translation should not be made using the CHX pre-treated library (which was also excluded from the PRICE ORF detection).

At only ten amino acids, the peptide predicted to be expressed from uORF1 is very short. Although there is a growing range of very short peptides shown to have conserved functions676–679, it is unlikely that uORF1 produces a functional protein, and the ORF is truncated in several isolates (Figure 4.13B). However, the presence of the uORF is highly conserved amongst NA PRRSV isolates (Figure 4.13C and D), suggesting it may have other advantages for viral fitness, for example modulating translation of other ORFs. There is a canonical (AUG) initiation codon in 558/564 sequences and sequences without an AUG have a non-canonical initiation codon with only a single mismatch compared to AUG (three have GUG, two have AUA and one has ACG), suggesting fairly efficient initiation is still possible. The stop codon is similarly well-conserved, with 576/598 sequences having UGA at the same position as SD95-21 PRRSV. Of the 22 sequences that do not have UGA at this position, six have an alternative stop codon (UAA) and the remaining 16 have CGA. Twelve of the sequences with CGA at this position have a stop codon upstream in the uORF, truncating the ORF to two codons. The remaining sequences have extended ORFs with stop codons 16 (three sequences) or 28 (one sequence) codons downstream of the codon which is UGA in SD95-21, meaning the uORF ends upstream of the initiation codon for ORF1a. Although the conservation of uORF1 suggests it may be beneficial for viral fitness, it may be conserved as an incidental consequence of selection pressure to maintain other features of the 5’ UTR, such as secondary RNA structural elements.
Figure 4.13 A novel, highly expressed uORF in the NA PRRSV 5’ UTR

A) Distribution of reads mapping to the region of the NA PRRSV 5’ UTR encoding uORF1. Plot constructed as in Figure 3.25, with the uORF1 amino acid sequence and stAI underlaid. Note that parts of the predicted ORFs uORF2 and uORF3 overlap the region depicted but reads mapping to these putative ORFs are not easily discernible at this scale. Only read lengths identified as having minimal RNP contamination (indicated in Figure 3.11F for CHX-9hpi-WT and Figure 3.10E for noCHX-12hpi-WT-1) were used to generate this plot. B) Predicted amino acid sequences of uORF1 from 98 NA PRRSV genomes representative of the diversity of the phylogeny (after sequences which begin after the uORF1 initiation codon or have a likely sequencing artefact [KT257963] were removed). Plot constructed as in Figure 4.11D (main). The translated sequence did not include the stop codon and one sequence out of these...
98 (KY348852) has a 28 codon extension to the ORF which is not depicted. **C and D)** Conservation of **C)** the initiation context and **D)** the stop codon for NA PRRSV uORF1, based on 661 input NA PRRSV sequences. Sequences were filtered to take only those spanning the entire feature of interest with no gaps, leaving **C)** 564 and **D)** 598 sequences in the alignment used for the logo plots. The initiator AUG is indicated by a black box. The initiation context of this ORF is weak, as defined by the absence of a G at position +4 or a A/G at position –3 relative to the A of the AUG, but the sequence is highly conserved.

**Figure 4.14 A predicted novel uORF in the EU PRRSV 5′ UTR**

**A)** Distribution of reads mapping to the region of the EU PRRSV 5′ UTR encoding the uORF. Plot constructed as in Figure 4.13A, using only read lengths with good phasing (indicated in Figure 3.10E). **B)** Predicted amino acid sequences of the EU PRRSV uORF from 98 EU PRRSV genomes (after sequences which begin after the uORF initiation codon or have a likely sequencing artefact [KY434183] were removed). Plot constructed as in Figure 4.13B. **C)** Conservation of the initiation context for the EU PRRSV uORF, based on the 98 EU PRRSV genomes used in panel B. Plot constructed as in Figure 4.13C. **D)** Conservation of the termination codon for the EU PRRSV uORF, based on 100 EU PRRSV genome sequences (after sequences which begin after the uORF initiation codon or have a likely sequencing artefact [KY434183] were removed). Note that the stop codon overlaps the TRS leader.
There is one predicted uORF in the EU PRRSV 5′ UTR, although this is not equivalent to the NA PRRSV uORF1 in position or length, and it is not very highly conserved (Figure 4.14). The predicted start site is a non-canonical GUG initiation codon with a weak initiation context, and is not highly conserved, with 33/98 sequences containing GCG at this position, making initiation unlikely in one third of isolates (Figure 4.14C). The stop codon is completely conserved (Figure 4.14D); however, it overlaps the TRS leader (UUAACC, stop codon underlined) so conservation is likely the result of pressure to maintain the TRS leader as opposed to the uORF. Protein binding sites and RNA secondary structures are common in viral UTRs, for example in PRRSV the TRS leader is located in the loop of a conserved stem loop601,680. This, combined with the sub-optimal initiation context and poor phasing in the putative uORF (Figure 4.14A), raises suspicions that the RiboSeq read density in this region (and consequent prediction of the ORF) may result from factors other than translating ribosomes.

Plots of the RPF distribution in the regions predicted to encode the remaining novel ORFs were made and, where novel transcripts from the chimeric read analysis provided potential mechanisms to facilitate expression, TRS body positions were indicated (Figure 4.15 and Figure 4.16). For EU PRRSV these plots do not constitute highly compelling evidence of translation of the novel ORFs; however, this may be due to the relatively low read counts used for these plots, as read lengths with good phasing were selected (whereas all read lengths were used as input for PRICE). Furthermore, if novel ORFs are translated from novel sgRNAs, these would be expected to become clearer at later timepoints, when sgRNA transcription and translation has increased. For NA PRRSV, translation is clearly evident for nsp10-iORF3, nsp11-iORF1, nsp11-iORF2, nsp11-iORF3, nsp12-iORF2, Gp2a-iORF and Gp4-iORF. The RPF distribution plots are consistent with translation of uORF2, uORF3, nsp9-iORF2, nsp10-iORF1 and Gp3-iORF, but do not provide unequivocal support for these ORFs. For nsp2-iORF, nsp3-4-iORF, nsp9-iORF1, nsp12-iORF1 and M-iORF, support for translation is not evident, and these ORFs may represent erroneous designations by PRICE, or may be translated at such a low level this is not easily visible on the RPF distribution plots (suggesting they are likely to be spurious).
Figure 4.15 (spans multiple pages). Plots of RPF distribution on regions of the NA PRRSV genome predicted to encode novel ORFs

Where space permits, the nucleotide sequence is underlaid beneath the read density. The amino acid sequence (three-letter code where space permits, single-letter code for longer ORFs) and stAl of the canonical ORF codons are underlaid beneath the nucleotide sequence, and the novel ORF amino acids beneath that. Where TRS bodies with $\geq 50$ supporting chimeric reads at 12 hpi are present (and lower-support TRS bodies in some indicated cases), their position is indicated by a red bar at the top of the genome map. Where TRS bodies that are likely responsible for translation of an ORF are a considerable distance upstream they are not included on the plot (for example Gp2a-iORF). Where PRICE detected the novel ORF in the 12 hpi group of libraries, the noCHX-12hpi-WT-1 library was selected to represent the noCHX data. Where PRICE did not detect the novel ORF in the 12 hpi group of libraries, a representative from the group of libraries it was detected in was used (Appendix section 5).

A) uORF2. Note that, to avoid the scaling being dominated by the highly expressed uORF1, peaks are truncated at 500 RPM on the y axis. B) uORF3. C) Nsp2-iORF. For the 12 hpi library, peaks are truncated at 30 RPM on the y axis to avoid the scaling being dominated by the large peak at the Gly shortly upstream of the predicted nsp2-iORF start codon. D) Nsp3-4-iORF. E) Nsp9-iORF1. F) Nsp9-iORF2. G) Nsp10-iORF1. H) Nsp10-iORF2. I) Nsp10-iORF3. J) Nsp11-iORF1. K) Nsp11-iORF2. L) Nsp11-iORF3. An additional TRS body is
shown despite having just under 50 supporting chimeric reads, as the more frequently used TRS body depicted would only permit production of a truncated form of the ORF. **M)** Nsp12-iORF1. **N)** Nsp12-iORF2. **O)** Gp2a-iORF. **P)** Gp3-iORF. **Q)** Gp4-iORF. **R)** M-iORF. For both libraries, peaks are truncated at 300 RPM on the y axis to avoid particularly large peaks from translation of the canonical ORFs dominating the scaling.
Characterising the PRRSV translatome
Analysis of PRRSV gene expression
Figure 4.16 (spans multiple pages). Plots of RPF distribution on regions of the EU PRRSV genome predicted to encode novel ORFs

Plots constructed as described for Figure 4.15, except the minimum requirement for plotting positions of TRS bodies was ≥10 supporting chimeric reads. A) Nsp2-iORF. B) Nsp5-iORF. C) Nsp9-iORF1. D) Nsp9-iORF2. E) Nsp10-iORF1. Peaks are truncated at 40 RPM to prevent the scaling being dominated by a large peak with P site at genomic coordinate 9248. F) Nsp10-iORF2. G) Nsp10-iORF3. H) Nsp11-iORF.
Analysis of PRRSV gene expression
4.4 Quantification of PRRSV gene expression

4.4.1 Overview of results

The novel PRRSV ORFs detected in section 4.3, and the known canonical ORFs, were paired with the transcripts from which they are most likely to be expressed (Figure 4.17), including novel transcripts characterised in section 4.2. Transcript abundance was calculated according to two different methods, based either on decumulated RNASeq read density (RPKM) or number of chimeric junction-spanning reads (RPM) (details given in Figure 4.17, Figure 4.18 and section 2.4.6). Note that the leader is not a separate viral transcript but is designated as a “transcript” in these analyses to provide the context for interpreting uORF expression. Total translation was calculated as RiboSeq read density mapping to each ORF, using only read lengths with good phasing and taking those that are in phase with the frame of the ORF (Figure 4.18 and section 2.4.6). All analyses were performed using positive-sense reads only, unless otherwise specified (Figure 4.18B and Figure 4.21).

The abundance of almost all viral transcripts increases over time (Figure 4.18, Figure 4.19, Figure 4.20 and Figure 4.21). Consistent with previous studies on SHFV and MHV307,663, the relative abundance of sgRNAs compared to each other remains reasonably stable between 9 and 12 hpi (Figure 4.22), while at 6 hpi, sgRNAs are not abundant enough for accurate comparison. Similar to findings for MHV and SARS-CoV-2 (references307,670), the decumulation method indicates that the abundance of canonical sgRNAs (except for M, discussed later) decreases as the distance between the 3′ end of the genome and the TRS body increases (Figure 4.18, Figure 4.19 and Figure 4.22). This trend is only somewhat true for EU PRRSV, and is not observed in the junction-spanning read counts for either virus (Figure 4.18, Figure 4.20, Figure 4.21 and Figure 4.22); however, this may be due to the fact that only the major, known TRS bodies were used for the canonical sgRNAs in this analysis, whereas the decumulation method captures RNASeq density originating from all transcripts in the designated regions. Despite this potential advantage, the decumulation method is hampered by susceptibility to noise, as evidenced by the fact that transcript abundance estimates were often negative values. This is a particular problem for transcripts which are not highly abundant compared to the (major) transcript with its TRS body next upstream. For example, novel sgRNA abundance estimates were frequently negative, suggesting this method is inappropriate for analysing non-canonical transcripts such as these. However, negative values were also found in some cases for the much more abundant canonical sgRNAs, for example sgRNA2 and sgRNA5 for both EU PRRSV libraries (and individual libraries for NA PRRSV sgRNAs 1, 2 and 5). For both methods, the relative abundances of the canonical negative-sense transcripts broadly mirror the trends seen in the positive-sense reads, except that there is a higher relative abundance of gRNA in the analysis of negative-sense reads (Figure 4.18).
Quantification of PRRSV gene expression

Figure 4.17 Subgenomic mRNA transcripts and ORFs included in viral gene expression analysis of A) NA PRRSV and B) EU PRRSV

Canonical transcripts and ORFs are labelled in black, novel ones in red. Consensus sequences of the TRS bodies used to generate the sgRNAs are plotted to the right of each transcript, based on A) 661 NA PRRSV and B) 120 EU PRRSV genome sequences. The genome map from nsp5 onwards is reproduced above for comparison. The leader (grey) is treated as a separate transcript for the purposes of these analyses, and the uORFs putatively expressed from it were omitted from these plots for clarity. Where two 5’ UTRs are depicted for some mRNAs this indicates that two TRS bodies were detected that likely give rise to transcripts from which the same ORF(s) are translated. The consensus sequences for these are arranged with the upstream TRS body on the left. In these cases, the two alternative transcripts were considered as one species in the gene expression analysis, and junction-spanning read counts for both junctions were combined. For canonical sgRNAs, non-canonical transcripts that could potentially contribute to canonical ORF expression were not assessed and only known TRS bodies were included. In addition to these sgRNAs and ORFs, ORF1a and the novel ORFs overlapping ORF1a (and EU PRRSV nsp9-iORF1 and nsp9-iORF2) are not depicted but were included in the analysis and designated as expressed from the gRNA transcript. Gp3-iORF (NA PRRSV) and its corresponding putative sgRNA were excluded to avoid excessive truncation of the region of measurement for sgRNA2.
Figure 4.18 Viral transcript abundance and total translation of viral genes

A) Timecourse of RNA synthesis and translation for canonical PRRSV genes, using only positive-sense reads. Mean values for NA PRRSV are indicated by the line graph, with
individual data points plotted as crosses. Mean values for EU PRRSV are plotted as filled circles, with individual data points omitted for clarity and some markers offset to the left to avoid overlap with other markers. Upper left: raw (total) RNASeq density. Upper right: decumulated (decum.) RNASeq density. Lower left: junction-spanning read abundance. Note that, due to the shorter RNA fragment lengths selected for EU libraries and for NA 9 hpi replicate one libraries, the chimeric read counts are not comparable to the remaining NA PRRSV libraries. For all junction-spanning read analyses in which averages were taken, replicate one libraries were excluded from the “9 hpi” group and analysed separately (in this case see Figure 4.21). Lower right: RiboSeq density. For the upper two panels, genome regions over which read density was measured were defined by the region between the TRS body for a given mRNA and the next major TRS body downstream (Figure 4.17, Appendix section 3). For RiboSeq density calculations the regions were defined by the region between the beginning of the ORF and the next major TRS body downstream, read lengths with good phasing (and minimal RNP contamination) were used and only reads mapping in-phase with the frame of the ORF were counted. ORF1b is omitted from this and several other plots and investigated separately in Figure 4.30. For this and all analyses henceforth, except those in Figure 4.32, “ORF1a” refers to the region from the beginning of ORF1a up to the nsp2 frameshift site, with “gRNA” transcript abundance (decumulated and chimeric) and “ORF1a” RPF density calculated ignoring the existence of putative heteroclite sgRNAs. To generate the decumulated RNASeq density value for a transcript, the raw read density for the neighbouring region upstream was subtracted. In some cases (for example EU PRRSV sgRNA5) this led to negative values, which are not biologically relevant results. These are likely reflective of the compounded effect of bias and/or noise in the data and were neither plotted nor included in the calculation of mean values. B) Timecourse of negative-sense RNA synthesis for canonical PRRSV transcripts. Plot constructed as in panel A, with RiboSeq panel omitted. This and Figure 4.21 are the only panels showing analyses of negative-sense reads.
Figure 4.19 Transcript abundance of positive-sense novel and canonical viral transcripts calculated using decumulation

Data for canonical transcripts is reproduced from Figure 4.18A, and the same decumulation calculations were performed for novel transcripts. Bars represent the mean, with individual data points represented as black circles with white outlines. Grey x axis labels indicate that all values were negative (and therefore excluded) for this group.
Figure 4.20 Transcript abundance of positive-sense novel and canonical viral transcripts calculated using junction-spanning reads

Data for canonical transcripts are reproduced from Figure 4.18A, and the same junction-spanning read abundance calculations were performed for novel transcripts. Plot constructed as in Figure 4.19.
Figure 4.21 Transcript abundance of positive-sense novel and canonical viral transcripts calculated using junction-spanning reads for libraries with shorter purified RNA fragments

Junction-spanning read abundance for each viral transcript, for NA PRRSV 9 hpi replicate 1 and for EU libraries. Plot constructed as in Figure 4.19.
Figure 4.22 Percentage of the viral transcriptome represented by each transcript as calculated using the decumulation method (left) and chimeric junction-spanning reads method (right)

RNASeq transcript abundance RPKM (decumulated) or RPM (chimeric) values for each transcript within a library were converted to a percentage of the total abundance of all transcripts plotted for that library, and bars plotted from left to right according to the position of their designated region on the genome. Note that the leader is not a separate viral transcript but is expected to be included on all other viral transcripts, meaning its percentage density in this analysis should be 50%. For the junction-spanning read analysis, the method of calculating leader “transcript abundance” pre-defines this as being the case (see section 2.4.6).

The M transcript (sgRNA5) is an outlier to the trend of shorter sgRNAs being more abundant in the decumulated analysis for NA PRRSV (Figure 4.18, Figure 4.19 and Figure 4.22). In both this and the junction-spanning read analyses (Figure 4.20 and Figure 4.21), transcript abundance is lower for sgRNA5 than for sgRNA4 or sgRNA6. In another deviation from the trend, the decumulated analysis...
indicates that the abundance of sgRNA5 decreases at 12 hpi, both in the forward and reverse sense reads (Figure 4.18), although for the positive-sense reads only one 12 hpi library had a positive value for transcript abundance. Scrutinising the results for each library individually reveals some variation between 9 hpi replicate one libraries, for which shorter (25-34 nt) RNA fragments were purified, and the remaining 9 hpi libraries (Figure 4.22). For the junction-spanning read method, the relative abundance of sgRNA5 is considerably lower for the libraries with shorter reads. This method is expected to be quite sensitive to differences in read length, due to the requirement for a 12 nt region of alignment either side of the junction site, which restricts the number of valid alignments more severely for shorter reads. Additionally, the M TRS body is preceded by 6 nt which are identical to those upstream of the TRS leader, meaning shorter reads are less likely to pass the filter that discards reads with more than two potential alignments (section 2.4.6). The decumulation method, however, is expected to be less sensitive to differences in read length and it is unclear why libraries with shorter read lengths have a greater proportion of reads attributed to sgRNA5 (Figure 4.22).

Similarly to the observed increases in viral transcript abundance, translation of NA PRRSV ORFs increases as infection progresses (Figure 4.18 and Figure 4.23). An exception to this is ORF1a, for which translation decreases between 9 and 12 hpi despite transcript abundance increasing, perhaps reflecting preferential utilisation of gRNA for replication instead of translation at the later timepoint. The same is true for nsp2-iORF, nsp3-4-iORF and nsp9-iORF1; however, translation of these ORFs was not convincingly evident from the RPF distribution (Figure 4.15C-E) and potential mechanisms for expression of nsp2-iORF and nsp3-4-iORF are not obvious. These lowly expressed putative ORFs overlapping ORF1ab potentially represent false positive results, and the observed decrease in their expression more likely reflects a decrease in RPFs which originate from ORF1ab translation but are mis-attributed to the overlapping ORFs due to heterogeneous phasing. This highlights an important caveat of this method of estimating ribosome density, and caution is advised when interpreting results from novel ORFs with low support (Figure 4.15 and Figure 4.16) that overlap much more highly expressed canonical ORFs.
Quantification of PRRSV gene expression

Figure 4.23 RiboSeq density attributed to each viral ORF

Data for canonical ORFs is reproduced from Figure 4.18A, and the same calculations were performed for novel ORFs. Plot constructed as in Figure 4.19, with ORFs coloured according to their paired transcript (Figure 4.17).

A comparison of the translation of viral ORFs relative to each other (represented by relative RPF density) confirms that the majority of translation at 6 hpi is of gRNA, with a marked shift towards sgRNA translation at 9 hpi and a similar profile at 12 hpi (Figure 4.24). NA PRRSV uORF1 is very highly expressed at all timepoints (Figure 4.23 and Figure 4.24), only surpassed by a select few ORFs (ORF1a at 6 hpi; Gp5, M and N at 9 and 12 hpi). The high expression at 6 hpi indicates uORF1 is expressed from gRNA, as this is by far the most abundant viral transcript at this timepoint (Figure 4.22, chimeric panel). The increase in the ratio of uORF1 compared to ORF1a translation at later...
timepoints, when sgRNAs become abundant, suggests that uORF1 is also translated from the sgRNAs (Figure 4.23 and Figure 4.24). Although assessments of RPF density in regions as small as uORF1 (ten codons) are subject to noise, this is a striking result and it is clear that uORF1 is very highly expressed.

The relative proportion of translation attributed to each ORF is similar between EU PRRSV (8 hpi) and NA PRRSV at 9 and 12 hpi, with the main noticeable difference being that E is more highly translated relative to the other ORFs in EU PRRSV (Figure 4.24). The overall similarity in translational profiles is not unexpected, as the majority of ongoing translation at these timepoints is of structural proteins, which are likely required at similar stoichiometries for both species of PRRSV\textsuperscript{627}. Variation in expression levels (and therefore stoichiometry) of E may be more tolerable, as this is a minor structural protein, essential for infectivity but not virion formation\textsuperscript{612,629–631,681}. Further, E is thought to be expressed from leaky scanning of the Gp2a transcript\textsuperscript{681}, which may mean that initiation efficiency is more sensitive to slight differences in sequence (or cellular) contexts between the two viruses. It would be interesting to test whether the presence of the highly expressed uORF1 in NA PRRSV (compared to the lack of convincing uORF expression in EU PRRSV) has any effect on initiation efficiency at this site.

Translation efficiency was calculated as the RPF density (RPKM) in each ORF divided by transcript abundance from both the decumulation (Figure 4.25) and junction-spanning read methods (Figure 4.26). Although ORFs can be compared within each method, they cannot be compared between methods as the denominators represent different measurements (RPKM and RPM, respectively). Furthermore, “translation efficiency” here is an imperfect proxy, as ribosome occupancy can also be affected by translation speed, although this effect is expected to average out over large areas such as some of the longer viral ORFs\textsuperscript{365}. The agreement in trends between the two methods of calculating TE is relatively poor, likely due to the quotient amplifying the existing differences in transcript abundance estimates (discussed above). However, in both methods, TE of uORF1 is low compared to other ORFs, highlighting that its very high translation is facilitated by massive transcript abundance (as it is present on all viral transcripts) as opposed to efficient initiation. This is consistent with the weak initiation context of the uORF and suggests that similarly high levels of uORF1 expression may be achieved in other NA PRRSV isolates, despite conservation of the weak initiation context (Figure 4.13C). The translation efficiency of M is high at late timepoints for NA PRRSV according to the decumulation method, consistent with the low decumulated transcript abundance, while the junction-spanning read analysis suggests a more moderate TE. In contrast, for EU PRRSV, the translation efficiency of M according to the chimeric method (the only method for which the calculation was possible) is amongst the lowest of the viral ORFs. Finally, TE of the canonical viral ORFs (according to the decumulation method) was compared with that of host ORFs (Figure 4.27).
This revealed that, like for MHV and SARS-CoV-2 (references 307,375), PRRSV ORFs are highly expressed due to great abundance of viral transcripts as opposed to preferential translation.

Figure 4.24 Percentage of the viral translatome represented by each ORF

Plot constructed as in Figure 4.22. Note that the novel ORFs detected on the EU PRRSV genome were named according to the same convention as for NA PRRSV novel ORFs so many are plotted in the same colour and labelled with the same name (in some cases without an appended “1”) but this does not indicate they are equivalent ORFs.
Figure 4.25 Translation efficiency of viral ORFs calculated using decumulated RNASeq density

ORFs are coloured according to their paired transcript. In cases where the decumulated RNASeq density value was a negative number, TE was not calculated. Note that the TE value for nsp10-iORF2 in 9 hpi WT replicate 4 was a clear outlier, likely inflated due to a low calculated transcript abundance, and was excluded from the mean calculation and from the scatter plot. Plot constructed as in Figure 4.19, using a linear scale.
Figure 4.26 Translation efficiency of viral ORFs calculated using the junction-spanning RNASEq reads

Plot constructed as in Figure 4.25. Where the number of junction-spanning reads for a library was zero, TE was not calculated.
Figure 4.27 Comparison of translation efficiency for canonical viral ORFs and host genes with ten or more RiboSeq and RNASeq reads

Transcript abundance, calculated for viral transcripts using the decumulation method, is plotted on the x axis, with TE on the y axis. The median host TE value is indicated by a blue line.
4.4.2 Investigating the potential for ORF1b sgRNAs to regulate in-frame translation of ORF1b

As described in section 4.3, the novel overlapping ORFs expressed from the ORF1b sgRNAs are not subject to purifying selection and are therefore unlikely to encode functional proteins. Despite this, the TRS body sequences for ORF1b-sgRNAs 2, 5 and 7 are the only NA PRRSV non-canonical TRS bodies analysed (except the sgRNA-Gp4-iORF TRS body) that have only a single mismatch compared to the UUAACC TRS leader sequence (Figure 4.17). Furthermore, these TRS bodies are well conserved, particularly the final two Cs, which have been identified as the most highly conserved part of the TRS body consensus for two NA PRRSV strains\textsuperscript{426,662}, and are conserved in all canonical sgRNA TRS bodies for both PRRSV isolates analysed in this study (Figure 4.17). This suggests these ORF1b sgRNAs may be produced in other isolates of NA PRRSV, although factors other than the TRS body sequence are important in determining TRS body usage efficiency (as described in section 1.4.6.2.1).

For EU PRRSV, the ORF1b sgRNA TRS bodies are not as close a match to the TRS leader and are not as well conserved as those in NA PRRSV (Figure 4.17). Despite this, they are reasonably efficiently utilised even at 8 hpi, and are two of only a handful of junctions which passed the thresholds for detection in EU PRRSV (Figure 4.6), indicating that expression of ORF1b sgRNAs is a shared feature of the transcriptome between the two species. Further supporting this, the TRS body for EU PRRSV ORF1b-sgRNA2 is positioned in exactly the same genomic locus as the NA PRRSV ORF1b-sgRNA5 TRS body. Although these TRS bodies differ in sequence, both have a conserved maintenance of a single mismatch compared to the TRS leader (Figure 4.17) despite this not being a requirement for maintaining the amino acid identities at this position. The position of the EU PRRSV ORF1b-sgRNA1 TRS body is similar to that of the ORF1b-sgRNA2 TRS body on the NA PRRSV genome, but slightly downstream (121 nt when genome sequences are aligned by amino acids).

These similarities between the ORF1b sgRNAs produced in NA and EU PRRSV suggest there may be a selective advantage in their production, or at least that it is well tolerated. Given that the encoded ORF1b-overlapping proteins are likely spurious, the possibility arises that these ORF1b sgRNAs could function to modulate expression of ORF1b itself. To explore this hypothesis, the RNASeq and RiboSeq read density mapping to ORF1b was plotted to examine whether any step changes in read density correspond to the locations of ORF1b sgRNA TRS bodies (Figure 4.28 and Figure 4.29). Additionally, ORF1b was divided into the regions between each ORF1b sgRNA TRS body and the density of RPFs in phase with ORF1b in each section was determined (Figure 4.30).
Figure 4.28 Distribution of RNASeq (left) and RiboSeq (right) reads mapping to the ORF1b region of the NA PRRSV genome

Plots constructed as in Figure 3.12B and C, respectively, with the application of a 213 nt running mean filter, and dotted lines indicating TRS body positions. The relative abundance of junction-spanning reads supporting use of each TRS body (at 12 hpi) is reproduced from Figure 4.11B, with TRS body logo plots reproduced from Figure 4.17. For RNASeq, all read lengths were used, and for RiboSeq, read lengths identified as having minimal RNP contamination were used. For the RiboSeq panel, the purple density at the beginning of the plot originates from ORF1a and is visible in this ORF1b plot due to the large sliding window. A similar effect occurs at the end of nsp12, although Gp2a is in the same frame as ORF1b so this is less easily discernible.
Figure 4.29 Distribution of RNASeq (upper) and RiboSeq (lower) reads mapping to the ORF1b region of the EU PRRSV genome

For RNASeq, all read lengths were used, and for RiboSeq, read lengths with good phasing were used. Plot constructed as in Figure 4.28, with the relative abundance of junction-spanning reads supporting use of each TRS body reproduced from Figure 4.12B. The TRS body annotated at the end of nsp12, was not included in the expression analysis but had 27 supporting junction-spanning reads in total.

For NA PRRSV at 6 hpi, when ORF1b sgRNAs are not detectable (Figure 4.3B), both RNASeq and RiboSeq read density are reasonably constant across ORF1b (Figure 4.28 and Figure 4.30). At 9 hpi, RPF density increases with proximity to the 3’ end of ORF1b, and the end of nsp12 is translated more highly than ORF1a (Figure 4.28 and Figure 4.30). At 12 hpi, this trend becomes very striking, with ORF1a expression exceeded by ORF1b expression from the ORF1b-sgRNA5 TRS body onwards. Furthermore, two of the greatest increases in ORF1b RiboSeq density are seen after ORF1b-sgRNAs 2 and 5 (alongside ORF1b-sgRNA8) (Figure 4.30), the two TRS bodies with counterparts in EU PRRSV. Despite this clear trend of increasing translation, changes in RNASeq density are less evident (Figure 4.28). This could be because jumps in density are masked by technical biases, although it could also result from the low “translation efficiency” of ORF1ab. If only a small proportion of gRNA is available for translation (for example due to genome packaging), a relatively small increase in abundance of ORF1b sgRNA (which is not expected to undergo packaging) could have a noticeable effect on translation without being visible against the background of untranslated gRNA by RNASeq.
Figure 4.30 RiboSeq read density attributed to different regions of ORF1b

ORF1b was divided into regions based on the positions of the sgRNAs included in the expression analysis, and RiboSeq density of reads in phase with ORF1b was calculated. All sgRNA numbers in the x axis labels refer to ORF1b sgRNAs. RiboSeq density in ORF1a (upstream of the nsp2 frameshift site) is included for comparison and the mean value is indicated by a solid grey line. Plot constructed as in Figure 4.25.

For EU PRRSV, clear increases in both RNASeq and RiboSeq read density are visible after both ORF1b sgRNA TRS bodies (Figure 4.29 and Figure 4.30), and translation downstream of ORF1b-sgRNA2 is already approximately equivalent to that of ORF1a at this timepoint. A further step change in RiboSeq density is apparent between the TRS bodies for ORF1b-sgRNAs 2 and 3, the reason for which is unclear as no abundant junctions (≥ 10 junction-spanning reads) are detectable.
at this position and intervening AUG codons would be expected to inhibit pre-initiation ribosomes from scanning that far on ORF1b-sgRNA2. However, the 25-34 nt RNA fragments purified for the EU PRRSV libraries are not ideal for performing chimeric read analyses and it is possible that there may be junctions here that are not detected in this analysis.

The increasing ORF1b-phase RPF density towards the end of this ORF supports the hypothesis that the ORF1b sgRNAs may serve to modulate the stoichiometry of nsps 10-12 later in infection. To further examine this, plots of the RPF distribution after the TRS bodies of the conserved ORF1b sgRNAs (2, 5 and 8 for NA PRRSV; 1 and 2 for EU PRRSV) were made and the initiation context and conservation of putative downstream initiation codons was assessed (Figure 4.31). Scores for the strength of the initiation context of these putative start codons were extracted from Noderer et al. for AUG start codons and De Arce et al. for non-AUG start codons. In some cases (such as EU PRRSV ORF1b-sgRNA1), conserved AUG codons downstream present likely potential candidates for initiation in frame with ORF1b; however, in other cases there are numerous intervening AUG codons in alternative reading frames, including those used to express the novel ORFs detected in section 4.3. The mechanism, if any, by which ORF1b sgRNAs facilitate the observed increase in ORF1b expression at late timepoints therefore requires further investigation.

Figure 4.31 (spans multiple pages). RPF distribution in regions downstream of ORF1b sgRNA TRS bodies and initiation context of potential start codons

Figure begins overleaf. Plots constructed as in Figure 4.15. Initiation context is shown for all AUG codons in any frame, for initiation codons of putative novel ORFs described in section 4.3, and for codons with prominent RiboSeq peaks and only one mismatch compared to AUG. Putative initiation codons are indicated by black boxes, and scores for the predicted initiation efficiency (extracted from references and ) are given above, based on the genome sequences of the isolates in this study. For AUG start codons these scores are based on nt positions to +5, while for non-AUG start codons they are based on nt positions to +4. All scores are the result of normalisation such that GCCACCAUGGC or CACCAUGG (generally thought to be the optimal initiation context; start codons underlined) have values of 100, meaning AUG and non-AUG scores are comparable. A) NA PRRSV ORF1b-sgRNA2. B) NA PRRSV ORF1b-sgRNA5. The top panel shows the region up to the end of nsp10 and the bottom panel shows the region from the beginning of nsp11. C) NA PRRSV ORF1b-sgRNA7. D) EU PRRSV ORF1b-sgRNA1. E) EU PRRSV ORF1b-sgRNA2.
Analysis of PRRSV gene expression
Quantification of PRRSV gene expression
Analysis of PRRSV gene expression
Quantification of PRRSV gene expression
Analysis of PRRSV gene expression
4.4.3 Investigating the effect of heteroclite sgRNAs on translation

The analyses described in sections 4.4.1 and 4.4.2 were performed using designated regions which ignored the existence of the heteroclite sgRNAs described in section 4.2. The region over which ORF1a translation was measured spanned from the beginning of the ORF to the nsp2 frameshift site, meaning translation from the heteroclite sgRNAs and the gRNA was included. Similarly, the gRNA transcript abundance estimates (for both methods) did not exclude reads originating from the putative heteroclite sgRNAs. To examine the effect of heteroclite sgRNAs on ORF1a translation, ORF1a was divided into three regions: “heteroclite”, between the start of ORF1a and the major (S-2) heteroclite junction (section 4.2.4), “before FS”, the region between the heteroclite junction and the nsp2 frameshift site, and “after FS”, the region between the nsp2 and ORF1ab frameshift sites. Similar regions were used to perform a decumulation-style transcript abundance analysis, where gRNA density was defined as the density between the major heteroclite junction and the ORF1ab frameshift site, and this was subtracted from the density in the heteroclite region to determine heteroclite sgRNA abundance. Note that the several other putative heteroclite sgRNAs with junctions within the heteroclite region (section 4.2.4) likely contribute to this density but these were not specifically analysed.

Although the chimeric read analysis did not detect the presence of any heteroclite sgRNAs in EU PRRSV at 8 hpi (Figure 4.6), there is an increased density in the analogous “heteroclite” region on the EU PRRSV genome (Figure 4.32, left), suggesting these sgRNAs may be present below the threshold of detection for the chimeric read analysis pipeline. Similarly, for both NA and EU PRRSV, the heteroclite region of ORF1a is more highly translated than the region directly downstream, with the ratio of heteroclite to before FS being the highest at 12 hpi (Figure 4.32, middle). Consistent with the observed changes in ribosome profile across ORF1a (Figure 3.16 and Figure 3.17), the density in the region after the frameshift site increases considerably at 9 hpi and remains high thereafter (Figure 4.32, middle), although the reason for this is unclear.

To determine translation efficiency, it was assumed that ORF1a is translated from the heteroclite sgRNAs and gRNA, and these transcript abundances were combined to form the denominator for the heteroclite region (Figure 4.32, right). TE of both the heteroclite and the before-FS regions decreases over time, similar to observations from SARS-CoV-2 (reference 683). This is consistent with the hypothesis that ORF1ab frameshift efficiency might increase in response to decreased ribosome load (section 3.6.1); however, TE of the region directly upstream of the ORF1ab frameshift site (“after FS” in this analysis) is not consistent with this trend. Further, an important caveat is that these “TE” calculations do not distinguish between RNA in the available translation pool compared to RNA which is not available for translation due to undergoing packaging or replication, which is
expected to be a considerable confounding factor for gRNA calculations. The potential significance of the heteroclite sgRNAs in PRRSV infection is discussed in the following section.

Figure 4.32 Expression of different regions of ORF1a

Transcript abundance for the heteroclite sgRNAs was calculated by decumulation, by subtracting the gRNA RNASeq RPKM (where the gRNA region was designated as beginning downstream of the major [S-2] heteroclite sgRNA junction) from the RPKM in the region between the TRS leader and the major heteroclite sgRNA junction (heteroclite). RiboSeq read density was calculated in the region between the beginning of ORF1a and the major heteroclite sgRNA junction (heteroclite), the region between this junction and the nsp2 PRF site (before FS), and the region between the nsp2 and ORF1ab frameshift sites (after FS). Although no junctions were detected for putative heteroclite sgRNAs in the EU dataset, regions were designated analogously to NA PRRSV, for comparison. For region coordinates see Appendix section 3. For TE of the “heteroclite” region, the denominator was both gRNA and (decumulated) heteroclite sgRNA combined. Plot constructed as in Figure 4.10.
4.5 Discussion and future work

4.5.1 Discussion

Characterisation of the PRRSV transcriptome, using a publicly available tool (STAR) and a custom pipeline, revealed an immensely complex transcriptional landscape, with numerous non-canonical transcripts. An unbiased broad screen further led to the discovery of over 20 novel ORFs between the two species of PRRSV. Although the expression of many of the overlapping novel ORFs can be attributed to the presence of the novel viral transcripts, these ORFs are not likely to be functional, suggesting that some of the non-canonical transcripts may instead function to modulate expression of canonical PRRSV proteins.

One example of such transcripts is the ORF1b sgRNAs, non-canonical sgRNAs for which the donor site of the junction spans the TRS leader and the acceptor site falls within ORF1b. ORF1b is canonically thought to be expressed only from gRNA, with an ORF1a:ORF1b ratio that remains constant throughout infection. As ORF1b sgRNA abundance increases over time, the discovery of these non-canonical sgRNAs implicated in facilitating ORF1b expression provides a second potential mechanism by which ORF1b expression might be temporally regulated (alongside the increasing ORF1ab frameshift efficiency described in Chapter 3). Given the low translation efficiency of gRNA compared to sgRNAs, these novel transcripts could have a significant effect on translation despite their relatively low abundance, a hypothesis supported by the increase in ORF1b-phase RPF density observed after the acceptor sites of these non-canonical junctions. Similar transcripts have been observed for SHFV, and quantitative mass spectrometry revealed peptides from nsp11 and nsp12 were 1.2- and 3.1-fold more abundant (respectively) than those from ORF1a-encoded nsp8 (reference 663). Although these mass spectrometry experiments could be confounded by differences in protein stability, the results are highly consistent with the increased RPF density in these regions of the genome for PRRSV at late timepoints, suggesting this may be a common feature of arterivirus gene expression. Supporting this hypothesis, an ORF1b sgRNA predicted to express the C-terminal 200 amino acids of ORF1b was identified in LDV684. ORF1b sgRNAs have also been detected for other members of the order Nidovirales: SARS-CoV-2, HCoV-229E and equine torovirus375,647,666,667. For SARS-CoV-2 and equine torovirus, it would be interesting to determine whether the publicly available ribosome profiling datasets contain evidence of increased translation after these non-canonical ORF1b TRS bodies375,666.

The ORF1b sgRNAs may simply represent a tolerable but unintended consequence of the discontinuous transcription mechanism required to make the canonical sgRNAs; however, the putative regulation of expression of the 3′-proximal region of ORF1b could be advantageous for viral fitness. The RdRp is the first protein encoded by ORF1b, and the majority of nidovirus ORF1b sgRNAs have junction acceptor sites located after or near the end of this region375,647,663,666,667,684.
suggesting they may function to increase the stoichiometry of the downstream proteins (including the helicase) relative to the RdRp late in infection. This could be important for determining the composition of the viral RTC. The stoichiometry of the arterivirus RTC is unknown, although the “core” coronavirus RdRp complex (also known as the holo-RdRp) has been characterised structurally, revealing a ratio of 1:2:1 nsp7:nsp8:nsp12 (references 685–687). Coronavirus nsp12 encodes the RdRp, while nsp7 and nsp8 (encoded by ORF1a) are cofactors essential for full RdRp activity688. Although the arterivirus RdRp is known (nsp9), cofactors playing the role of CoV nsp7 and nsp8 have not been identified, making comparisons between the two families challenging. Furthermore, additional proteins (nsp10, nsp13, nsp14 and nsp16) are required to make the full coronavirus holoenzyme complex for complete activity562,685,689,690, and the stoichiometry of the full complex is uncharacterised, with the added complication that some of the required proteins do not have known arterivirus counterparts562,691. Nonetheless, data from SARS-CoV-2 suggests that the helicase (nsp13 in CoVs, nsp10 in arteriviruses) may be required in greater abundance than the RdRp686. Chen et al. (2020) characterised the complex of the CoV holo-RdRp complex, RNA and the helicase, revealing that the stoichiometry varied, with either one or two copies of the helicase for each copy of the holo-RdRp686. This highlights the possibility that the composition of the RTC could change over time, with addition of an extra copy of the helicase (thought to be involved in promoting sgRNA transcription565,567,686) potentially regulating viral replication, for example by altering the ratio of gRNA to sgRNA production. PRRSV nsp12, the protein encoded at the 3′ end of ORF1b and which does not have a counterpart in CoVs692, could perform a similar function. Nsp12 is thought to be involved specifically with sgRNA synthesis (as opposed to replication)580,581, and up-regulation of nsp12 expression by ORF1b sgRNAs could represent a positive feedback loop that promotes the observed switch from gRNA to sgRNA synthesis at late timepoints.

Finally, nsp11 is an innate immune antagonist572,573,575–577 and is thought to play a role in synthesis of both gRNA and sgRNA (with some evidence for a specific role in sgRNA synthesis), although the mechanism is incompletely characterised573,574. It is an endoribonuclease with broad substrate specificity, the expression of which is highly toxic outside the context of infection, leading to suggestions that it requires compartmentalisation (for example in virally induced DMVs) to avoid its expression becoming “suicidal” for the virus515,568–573. As such, it may be beneficial to maintain nsp11 expression at low levels early in infection, affording time for viral manipulation of host membranes to create the optimal microenvironment for appropriate nsp11 localisation before expression is potentially increased by expression from ORF1b sgRNAs.

Another class of non-canonical sgRNAs that have the potential to modulate the stoichiometry of the non-structural proteins is the heteroclite sgRNAs, found in the NA PRRSV dataset in this study and in several other isolates of NA PRRSV619,664. Although no highly supported putative heteroclite sgRNA junctions were found in the chimeric read analysis of the EU PRRSV dataset, there was
some evidence for increased transcript abundance and translation in the 5′-proximal region of ORF1a, suggesting these transcripts might be present below the limit of detection by this pipeline. This is supported by the finding that similar transcripts were observed in several studies of CoVs. Although in some cases there was no evidence to suggest their formation was similarity-directed (as heteroclite sgRNAs are in PRRSV), there appear to be hot-spots for the donor sites of these non-canonical CoV sgRNAs, which are located in an analogous region of the genome to those in PRRSV, and similarly produce transcripts predicted to express nsp1 or nsp1-nsp2 (references 647,667,669,670). For PRRSV, these heteroclite sgRNAs are packaged into the virion, both in addition to gRNA and alone. They may represent DI-like RNAs which are neutral or detrimental to viral fitness (to levels not detectable under the conditions in Yuan et al.), who found they did not interfere with viral replication, or may represent a mechanism to increase the expression of nsp1α/β from very early timepoints onwards. This is consistent with the roles of nsp1α and nsp1β as the most potent of the innate immune suppressors encoded by PRRSV, which is likely a beneficial function at all stages of infection. Increased expression of nsp1β could also help promote frameshifting at the nsp2 site (section 1.2.2.1). The translation of these heteroclite sgRNAs is supported by the observed increase in RPF density in the 5′-proximal region of ORF1a, and the finding that similar transcripts in SARS-CoV-2 show evidence of translation in Riboseq and mass spectrometry datasets. However, it is important to verify that the heteroclite sgRNAs produce functional proteins in the context of infection (see next section), as Yuan et al. (2004) tested one example heteroclite sgRNA in vitro and only detected the uncleaved partial polyprotein, not cleaved nsp1α.

Of the novel PRRSV ORFs discovered in this study, the one most likely to have an effect on viral fitness is NA PRRSV uORF1. The presence of one or more uORFs in some cellular transcripts is known to confer resistance to translational inhibition mediated by eIF2α phosphorylation. This presents a potential role for uORF1, as PRRSV is known to activate the PERK branch of the unfolded protein response, which culminates in eIF2α phosphorylation. However, this uORF-conferred “translational resistance” is defined by comparing translation under stress to normal conditions, and the proposed mechanisms are generally based on uORF translation under normal conditions repressing translation of the main ORF (which is relieved when eIF2α is phosphorylated). While it is clear how utilisation of this de-repression strategy to differentially regulate select cellular transcripts is advantageous, uORF1 is present on all viral transcripts, and the benefit of potentially down-regulating translation of all viral ORFs in the early stage of infection (before UPR activation) is less obvious. The extent to which uORFs repress translation of the main ORF varies depending on several factors (Figure 4.33). As the uORF sequence is expected to be identical on all viral transcripts, many of these factors are also the same; however, the distance between the uORF stop codon and the main ORF start codon varies. For gRNA
this distance is 134 nt, which is extended by an additional 4-123 nt for the canonical sgRNAs depending on the distance between the TRS body and the main ORF start codon. These variations in UTR length might lead to a differential effect of uORF1 on translation of the different structural ORFs, which might function to refine overall expression levels. The N transcript is a good example, as the distance between the uORF stop codon and the main ORF start codon is either 143 or 257 nt depending on whether the N-short or N-long TRS body is used. It is interesting to note that EU PRRSV, which did not show convincing evidence of a highly translated uORF, was not found to produce two abundant N transcript variants.

**Figure 4.33 Features of a uORF which correlate with the extent of main ORF (CDS) translational repression**

Figure reproduced from Zhang et al., 2019 (reference 695). In parentheses are names of tools that can be used to assess these features. uAUG = uORF AUG, cAUG = CDS (main ORF) AUG.

### 4.5.2 Future work

Several things could be done using the datasets from this chapter to complement and/or improve the analyses described. To improve the quantification of viral gene expression, all non-canonical transcripts could be assessed and those which are expected to facilitate expression of full-length canonical structural proteins could be incorporated into the junction-spanning read method of quantifying transcript abundance. This would be expected to improve the accuracy of this method, allowing more reliable quantification of transcript abundance and TE, and meaning these structural ORF transcripts can be more accurately compared to the non-canonical ORF1ab sgRNAs (the abundance of which cannot be accurately assessed by the decumulation method). Further, the accuracy of estimation of translation levels would be improved by applying a more sophisticated method to account for the heterogeneity in distance between the 5′ end of RPFs and the P site of the ribosome, which causes problems for lowly expressed ORFs overlapping more highly expressed ORFs. The approach that was used to calculate −2 PRF efficiency at the nsp2 frameshift site could be applied, or programmes (such as PRICE) which use models accounting for this heterogeneity could be used to estimate translation levels.
To complement the analysis of the novel PRRSV transcriptome, a systematic assessment of the features of junction donor and acceptor sites could be performed to investigate whether particular features promote the formation of the different classes of junction. Although these have been assessed to some extent in PRRSV, it has been by low-throughput techniques ill-suited to capture the less abundant, non-canonical transcripts\textsuperscript{426,619,662,664,665}. Prediction of TRS body usage in other PRRSV isolates could potentially be improved by determination of the optimal TRS body consensus in the NA and EU isolates used in this study, for example by utilising the transcript abundance estimates to give TRS bodies for more abundant transcripts a greater weighting in generating the consensus. A similar approach could be applied to putative heteroclite sgRNAs to further elucidate potential mechanisms by which these are formed. Although slightly more challenging to assess in a high-throughput pipeline, the potential involvement of RNA structure or protein-binding sites could be assessed by using tools and databases such as mfold\textsuperscript{706} or the RNA-binding protein database\textsuperscript{707}.

Analysis of new datasets could also verify and further explore the findings of this study. For example, nanopore sequencing\textsuperscript{708,709} could be used to directly sequence viral RNA, obviating the need for amplification (and associated PCR bias) and allowing full-length viral transcripts to be sequenced in a single read. This would circumvent the potential alignment problems associated with using short reads for chimeric analyses and would allow characterisation of transcripts with multiple junctions. Although direct (nanopore) sequencing is less accurate than sequencing by synthesis (such as Illumina sequencing), a combination of the two techniques is becoming a popular strategy for viral transcriptomics, ensuring both accuracy and haplotype information are retained\textsuperscript{647,667,670}. EU PRRSV would be prioritised for these experiments, as there are some instances where the RiboSeq or RNASeq results are consistent with the presence of a junction (representing heteroclite sgRNAs or an extra ORF1b sgRNA TRS body between ORF1b-sgRNA2 and ORF1b-sgRNA3) but no junctions were detected by the chimeric read analysis pipeline. Where possible, these experiments should be performed on an isolate taken directly from the field, to avoid potential adaptations to tissue culture altering the transcriptome (for example if this promotes formation of DI genomes), and a physiologically relevant cell type (or \textit{in vivo} tissue samples) should be used. However, all 96 TRS leader-body junctions identified by RNASeq of SHFV in MA-104 (\textit{C. sabaeus} kidney) cells were verified in the more physiological model of primary macaque macrophages, suggesting that arterivirus sgRNA transcription is not exquisitely sensitive to cell type\textsuperscript{663}. Before further experiments were performed, a preferable option would be to subject publicly available PRRSV datasets to the chimeric read analysis pipeline. Several RNASeq datasets are available\textsuperscript{710–717} (including one nanopore dataset\textsuperscript{718}), mostly generated by authors aiming to characterise the host transcriptional response to PRRSV infection, in which the viral transcriptome was left uncharacterised and may be informative if high enough coverage of the viral genome was obtained.
Several experiments could help investigate the potential significance of the ORF1b sgRNAs. First, reverse transcriptase quantitative PCR could be performed using primers that bridge the junction sites, to verify the presence and abundance of these transcripts by an alternative method. Quantitative translation initiation sequencing (QTI-seq), a modified form of ribosome profiling which determines the location and approximate usage efficiency of initiation codons, could be used to assess whether initiation occurs in frame with ORF1b after prominent ORF1b sgRNA TRS bodies, and quantify any contribution this makes to overall ORF1b translation. Similarly to RNaseq, several proteomics datasets of PRRSV-infected cells are publicly available from studies which quantify changes in the host proteome in response to infection. The viral proteome could be analysed in these datasets to confirm expression of ORF1b-overlapping ORFs and to determine whether the abundance of (ORF1b-frame) peptides from the 3′-proximal regions of ORF1b is greater than those at the 5′ end. To determine whether this putative regulation of ORF1b translation is important for viral fitness, the TRS bodies of the more prominent ORF1b sgRNAs could be mutated to reduce their efficiency (without changing the amino acid sequence of ORF1b) and effects on plaque size and viral titre could be assayed. However, it would be important to verify that these knockouts were successful, and that potential nearby alternative TRS bodies were not utilised instead, as has been observed in TRS-knockout experiments for some nidoviruses.

Although more challenging due to the lack of a specific consensus motif, mutations could be made to reduce the sequence similarity at major NA PRRSV heteroclite sgRNA donor and acceptor sites, to attempt to reduce the formation of heteroclite sgRNAs and determine whether their presence is beneficial for the virus. To verify any potential effect on viral fitness, and confirm that these transcripts are translated in the context of infection, nsp1α and nsp1β PLP domains could be mutated on the gRNA. Mutating the PLP domain of nsp1α ablates sgRNA synthesis but not gRNA synthesis, while mutation of the PLP domain in nsp1β ablates both (although it is unclear whether this is due to nsp2 no longer being freed from the polyprotein). Similar to the work of Kroese et al., BHK-21 cells could be transfected with the mutant genomes and a first-round analysis of gRNA and sgRNA production performed to determine whether co-transfection of heteroclite sgRNAs could rescue these phenotypes (or virus viability) by trans-complementation.

To explore the potential function of NA PRRSV uORF1, a knockout mutant could be made to determine whether there is any detrimental effect on viral fitness. If so, the potential mechanism could be explored through transfecting cells with plasmids encoding a reporter gene (such as luciferase) and varying the 5′ UTR to mimic that of the viral gRNA or particular sgRNAs of interest, for example N-long, N-short and Gp2a/E. In this context, uORF1 could be knocked out and the effect on translation of the reporter could be assayed. As the effect of the uORF on translation of the main ORF may vary depending on the phosphorylation status of eIF2α, the reporter would ideally be transfected into cells prior to infection, and reporter expression assayed at multiple timepoints.
Discussion and future work

post-infection to capture different stages of UPR induction. Transfection of uninfected cells in the
presence or absence of a pharmacological UPR inducer, such as tunicamycin, is a potential
alternative which might prove simpler to interpret.

4.5.3 Conclusion

These results represent the first application of ribosome profiling to an arterivirus, and the first
characterisation of the non-canonical PRRSV transcriptome using an unbiased, high-throughput
technique. These analyses revealed enormous complexity in both the viral transcriptome and
translatome and provide support for two potential non-canonical mechanisms to regulate
stoichiometry of the proteins contained within the viral polyprotein. Further, a short but highly
translated uORF was discovered in the NA PRRSV 5′ UTR, the presence of which is very highly
conserved. This presents another opportunity for regulation of viral translation, potentially allowing
adaptation in response to infection-induced cellular stress. Performing ribosome profiling on other
arteriviruses would be an important step towards corroborating tentative evidence that some of these
putative regulatory mechanisms are conserved among members of the Arteriviridae family.
Chapter 5: Plenary discussion

5.1 PRRSV

The experiments presented in this thesis reveal a highly complex complement of PRRSV gene expression strategies (Figure 5.1). At the level of translational control, PRRSV was found to display a rare example of temporally regulated frameshifting, and a highly expressed uORF was discovered, with ramifications for initiation of downstream canonical ORFs. In terms of the viral transcriptome, numerous non-canonical transcripts were characterised with different temporal abundance profiles, which may differentially contribute to polyprotein translation during the course of infection. Taken together, these mechanisms permit temporally regulated modulation of the stoichiometry of proteins encoded by the polyprotein, that may act to balance the changing requirements for these proteins as infection progresses. However, deciphering which aspects of this intricate network are functionally important and which are merely incidental will require carefully controlled experimental investigation (discussed in sections 3.6.2 and 4.5.2).

![Figure 5.1 Schematic summary of non-canonical mechanisms of PRRSV gene expression regulation supported by this study](image)

The diversity of the viral transcriptome that results from the hallmark discontinuous transcription mechanism suggests that nidoviruses may be readily able to evolve novel ORFs or gene expression regulatory mechanisms. This is highlighted here for PRRSV in the discovery of numerous ORFs overlapping ORF1b, the translation of which is likely facilitated by the ORF1b sgRNAs. The surprisingly large number of these ORFs, in many cases combined with their high translation efficiency and relatively large predicted products, is a testament to the flexibility of the PRRSV translatome that is afforded by the inherent heterogeneity in the transcriptome. Although the ORFs
observed in this study have not been shown to produce functional proteins, this lability in the translatome potentially paves the way for similar ORFs to adapt to gain functions and become fixed in the viral population. If the peptides produced from the non-functional overlapping ORFs contain immune epitopes (for example reference 727), there may also be advantages in generating potential decoy epitopes from a population of spurious proteins which can evolve more rapidly to change epitope sequences than the conserved functional proteins.

As well as creating an environment for the evolution of novel ORFs, the diverse non-canonical transcriptome presents opportunities for adaptation of gene expression regulation, by using alternative transcripts to express select regions of the polyprotein. The potential presence of such mechanisms in PRRSV is implied by the changes in RPF density corresponding to ORF1b sgRNA TRS bodies; indeed, the recent discovery of planarian secretory cell nidovirus (PSCNV)728 lends credence to this as a viable mechanism of nidovirus gene expression regulation. PSCNV has the largest known RNA virus genome (41.1 kb), which is organised as a single, extremely long ORF728. Expression of the predicted structural proteins, which are encoded towards the 3′ end, is thought to be up-regulated at late timepoints by formation of a long sgRNA, expected to facilitate translation initiation at an internal methionine in the polyprotein728. Phylogenetic analyses suggest that this genome structure evolved from a common nidovirus ancestor with a multi-ORF arrangement728, highlighting the possibility that multi-ORF nidoviruses such as PRRSV have the capacity to modulate polyprotein stoichiometry through the evolution of novel sgRNAs.
5.2 Programmed ribosomal frameshifting

PRRSV represents an excellent model system in which to study PRF, as the viral genome contains not only a canonical, pseudoknot-directed frameshift site but also a second frameshift site which operates by a unique, non-canonical mechanism. The finding that −2 PRF efficiency at the nsp2 site increases over time, likely caused by accumulation of nsp1β, marks arteriviruses as only the second known example of temporally regulated frameshifting, alongside cardioviruses. Through investigating the possible occurrence of mechanistically relevant ribosomal pausing at this site, a peak in the density of short RPFs at the frameshift site was revealed on the WT genome, but not on the genome of the frameshift-defective control virus. The significance of this peak is unclear, but it appears to be specific to this frameshift site, as a counterpart was not found at the canonical ORF1ab frameshift site in PRRSV, nor at the protein-stimulated −1 PRF site in TMEV (see Appendix section 6).

Surprisingly, frameshift efficiency was also found to vary at the canonical, RNA structure-directed PRF site. While the trans-stimulated nature of the non-canonical frameshift site presents a clear potential mechanism for temporal regulation, no such mechanism is immediately obvious for the ORF1ab frameshift site, suggesting it may be caused by global changes in the cellular translational landscape. This raises the possibility that other RNA structure-directed frameshift sites may be more sensitive to cellular conditions than previously thought. This is supported by recent studies of another recoding event, stop codon readthrough, the efficiency of which was found (in some cases) to be highly dependent on cell type despite the lack of known involvement of site-specific trans-acting factors. To investigate the possibility that other viruses which encode canonical frameshift sites may also exhibit variable frameshift efficiency, and corroborate the findings from PRRSV, other viruses could be subjected to ribosome profiling over a timecourse of infection. To avoid the potential confounding effect of non-canonical transcripts contributing to polyprotein translation, a virus with a less heterogeneous transcriptome would be a good choice, for example an astrovirus (for which ribosome profiling of a single timepoint has already been performed). Further, concerns over RNP contamination could be alleviated by using RiboLace, a modified form of ribosome profiling which involves selective purification of actively translating ribosomes.

A key unanswered question is what mechanism could drive such an increase in frameshift efficiency? As discussed in section 3.6, changes in ribosome density are a possibility, and if a technique was used that was capable of discerning translatable gRNA from gRNA engaged in packaging, assessment of changes in ribosome loading over time would be highly informative as to whether this could cause the results observed for PRRSV. More detailed studies of how ribosome loading affects frameshift efficiency would be valuable not only for interpretation of these results, but would also be relevant to the wider field. For example, it would be important to determine the
relative timescales of pseudoknot refolding, post-frameshift translation of the RNA which forms the pseudoknot, and movement of the trailing ribosome onto the slippery sequence, to assess whether the interplay between these events is physiologically relevant. Further, it would be relevant to investigate the effect of different cellular stresses on frameshift efficiency, for example UPR activation causing eIF2α phosphorylation (and translation initiation inhibition), and stress pathways that result in eEF2 phosphorylation, as discussed in section 3.6.

In terms of the understanding of frameshift mechanisms in general, an important gap in the field is detailed studies using eukaryotic ribosomes, which are lacking due to the greater experimental difficulty associated with using these compared to their smaller prokaryotic counterparts. Some progress has been made towards structural characterisation of a eukaryotic ribosome stalled at a PRF site, although studies at high resolution which capture the ribosome in an intermediate state during frameshifting are still lacking. Similarly, popular techniques for studying the kinetics of prokaryotic frameshifting (such as single-molecule and/or ensemble fluorescence) would be highly valuable if optimised for use with eukaryotic ribosomes.
5.3 Ribosome profiling

Ribosome profiling has rapidly become an indispensable technique for the field of studying translation, and many of the findings in this study are highly unlikely to have been detected by other techniques. For example, determining frameshift efficiency during early stages of infection (when levels of viral proteins are low) is challenging with less sensitive techniques such as western blotting. Further, the characterisation of novel ORFs is a highly lauded strength of ribosome profiling and it has facilitated the detection of numerous novel viral ORFs in this work.

There are, however, limitations to the applications of ribosome profiling. Novel ORF detection, for example, typically requires high quality phasing, which can be challenging to achieve in some models (such as prokaryotic profiling), and its detection limit is dependent on ORF length and expression levels. This is particularly true for short overlapping ORFs. Due to the very short read length, haplotype information is lost in ribosome profiling, and in cases where a transcript has multiple isoforms it is not possible to determine if RPFs mapping to shared sequences originate from one isoform or another. Adaptation of the technique such that haplotype information is retained in RPFs would be a huge advancement for the field. A potential strategy would be to employ a ribonuclease such as RelE, which selectively cleaves RNA within the A site of the ribosome, leaving the transcript upstream of the ribosome intact731, and providing a potential substrate for nanopore sequencing. This would have many potential applications, for example determining tissue-specific translation profiles for different transcript variants, and in this study would allow quantification of the potential contribution of ORF1b sgRNAs to translation of the 3′-proximal region of ORF1b.

Some of the limitations of ribosome profiling could be managed by standardisation of the required quality control analyses. For profiling of virus-infected cells, for example, RNP contamination can be a major issue, and analyses to determine whether this affects the libraries should be mandatory for all such studies. As found in this work, undetected non-canonical transcripts can also be a major confounding factor in interpretation of changes in RPF density. This is a particular issue for profiling of cells infected with viruses which have a heterogeneous transcriptome, and for nidovirus profiling, one should consider unbiased determination of the viral transcriptome to be an essential control for interpretation of RPF density on the viral genome. Notably, some changes which were not clearly discernible in the total RNASeq density profile appeared to have significant effects on RPF density. Therefore, RNASeq libraries should be prepared such that chimeric read analyses can be robustly performed, for example sufficiently long RNA fragments (ideally greater than 25-34 nt) should be purified, or direct RNA sequencing of unfragmented RNA should be performed.

Finally, a feature of vertebrate cell profiling which remains incompletely understood is the differences in average ribosome density in the first few codons of host CDSs compared to the average density across the ORF, as revealed by metagene analyses similar to those in Figure 3.5. As
outlined in section 1.3.2.1, a 5′ “ramp” is seen in yeast cells, in which the beginning of the metagene CDS has an excess of ribosomes, slowly decreasing over the first 100-200 codons of the CDS to plateau at the density for the rest of the CDS\textsuperscript{298,313}. Although CHX pre-treatment exacerbates this ramp, particularly under stress conditions, it is thought by many to be a genuine feature of translation in \textit{S. cerevisiae}\textsuperscript{298,327}. Some suggested putative functions of this slower initial translation are to present a translational bottleneck early in the ORF to prevent ribosomal collisions downstream, or to allow time for nascent polypeptide chains to engage with chaperones\textsuperscript{298,313-316,322}. Taking a running mean average of tAI across the metagene CDS, Tuller \textit{et al.} (reference \textsuperscript{322}) showed that many organisms, including yeast, have codons with low tAI towards the beginning of the CDS, with tAI increasing over the first few codons (varying in length for different organisms). This is an approximate ‘mirror image’ of the observed translational ramp, and could be a partial explanation, as low tAI is expected to correspond to slow translation\textsuperscript{319-321}. However, other studies found that the observed skew in tAI can only account for a portion of the ramping seen, suggesting that other mechanisms, such as slow dissociation of initiation factors, may also play a role\textsuperscript{327,732}.

In unstressed vertebrate cells which have not been pre-treated with CHX, this ramp does not occur and, in some datasets (including the noCHX NA PRRSV dataset in this thesis), the inverse is seen and there is a trough in the metagene analysis in place of the ramp\textsuperscript{305,317}. The trend of poor tAI increasing over the beginning region of CDSs was more variable in the vertebrate species examined in Tuller \textit{et al.} than in some other species, and some vertebrates (such as chickens) maintained approximately constant tAI throughout the CDS. Nonetheless, there was no evidence for the inverse pattern in the vertebrate species examined therein, and if this trough is a conserved feature of translational regulation, the mechanism is unclear. Potentially, it is an artefact of the protocol used to harvest samples for ribosome profiling – if translation initiation is more severely inhibited than elongation by the harvesting process, this trough could be formed by run-off elongation. However, the exact stage at which this may occur is unclear, making it difficult to fully interpret its potential effects or reduce its occurrence.

Studies in which the ribosome profiling method was developed and refined in yeast found similar troughs when no CHX was included at any stage of the experiment, although it is not usually observed when CHX is present in the lysis buffer, suggesting continued elongation during lysis may contribute\textsuperscript{298,331}. This trough was alleviated by rapid flash-freezing of the samples before lysis\textsuperscript{298,331}, supporting the conclusion that delays between cell harvesting and freezing lead to this run-off elongation. Hussman \textit{et al.} (reference \textsuperscript{317}) noted that the “wave” of downstream spreading of ribosomal pause peaks, thought to be caused by CHX-perturbed translation, was present in some yeast samples harvested by flash-freezing without CHX pre-treatment (section 1.3.2.1). As CHX was not present in previous stages of the harvesting protocol, this was interpreted as evidence that continued elongation occurs in the lysate, although it is possible that changes in conditions other
than the presence of CHX may similarly perturb translation. The observation that the population of short RPFs is absent from yeast libraries harvested using lysis buffer containing CHX (even in the absence of a pre-treatment step) supports the idea that translation continues in yeast lysates, as CHX traps translating ribosomes in the conformation that protects the classical RPF length\textsuperscript{140,331} (section 1.3.2.2).

However, the above studies were carried out in yeast, and there are considerable differences between yeast and vertebrate cells, not only in terms of biology but also in terms of the requirements of harvesting cells for profiling. Yeast cultures usually require filtration before flash-freezing and lysis, which may cause cellular stress that triggers translational regulation or perturbation, as observed for filtered bacterial cultures\textsuperscript{733}. Vertebrate cell cultures, on the other hand, are often adherent and can be flash-frozen directly by submersion in liquid nitrogen.

In mammalian cells the population of short RPFs does not disappear when CHX is included in the lysis buffer, suggesting that translation is not prone to continue in mammalian lysates\textsuperscript{140}. While this may simply be an effect of the extent to which translation factors are diluted by lysis in these particular experiments, it highlights the concept that the collapse of the short-read peak is potentially a sensitive indicator of ongoing translation in the lysate. As such, if the trough observed in the metagene analysis results from translation in the lysate it would be expected that libraries displaying pronounced troughs have a very small proportion of short RPFs. However, this is not observed in the datasets in this thesis, in which the NA PRRSV noCHX dataset had a relatively pronounced trough (Figure 3.5) but had several libraries with over 50% short reads (Figure 3.4), while the TMEV dataset (harvested by flash-freezing without CHX pre-treatment) had a much lower proportion of short reads despite not having a pronounced trough (Appendix section 6). This suggests that the run-off elongation occurs before the samples are lysed, implicating either the short time spent at room temperature awaiting processing (after removal from the 37° incubator), or the warm PBS wash directly prior to flash-freezing. Alternatively, CHX may interact differently with yeast ribosomes compared to mammalian ribosomes and may not cause the collapse of the short-read peak in mammalian cell lysates even if translation were ongoing. This is supported by a study in human cells that found the population of short RPFs was present even after a long pre-treatment step in which CHX was used at high concentration\textsuperscript{734}.

A better understanding of how the different steps of the harvesting protocol affect profiling datasets would greatly aid interpretation and/or critique of potential global translation phenomena such as the 5′ ramp, as well as facilitating deeper interpretation of the quality of individual datasets. For example, in this thesis, ribosome profiling is used to investigate ribosomal pausing at high resolution, for which it is useful to know whether run-off elongation could be a potential confounding factor. If run-off results from solely perturbation of translation initiation (as opposed to elongation), then ribosome occupancy of codons a greater distance downstream of the initiation site than the distance
covered by the metagene trough should be relatively unaffected. An example of such a response could be inhibition of initiation in response to nutrient starvation after media removal\textsuperscript{15}. The frameshift sites examined in this study are all at least 3000 nt downstream of the corresponding initiation sites and are unlikely to be significantly affected by initiation inhibition of the durations suggested by their respective troughs. On the other hand, if the trough represents continued translation in the lysate, that is a condition under which it is more likely that elongation is also perturbed (for example by the presence of CHX\textsuperscript{17}), which is of greater concern when interpreting putative pauses in elongation such as those in Chapter 3. A systematic study in representative vertebrate systems of how harvesting conditions affect the distribution of RPFs across the CDS, and the proportion of short RPFs, would be extremely beneficial for the field of ribosome profiling.
5.4 Conclusion

This study highlights both the immense value of using ribosome profiling to study viral translation, and the great care required to ensure appropriate interpretation of the results. It represents the most comprehensive analysis of arterivirus gene expression to date and reveals a highly complex interplay between non-canonical translation mechanisms and an array of novel viral transcripts, providing numerous opportunities for temporal regulation. The finding that frameshift efficiency varies, not only at the protein-directed site but also at the canonical site, presents a new paradigm for understanding changes in viral translation during infection, and highlights arteriviruses as a potential model for studying gene expression mechanisms relevant in other contexts.
References


20. Qu, X. *et al.* The ribosome uses two active mechanisms to unwind messenger RNA during
References


57. Loughran, G., Firth, A. E., Atkins, J. F. & Ivanov, I. P. Translational autoregulation of BZW1


75. De Arce, A. J. D., Noderer, W. L. & Wang, C. L. Complete motif analysis of sequence


111. Duncan, R. F. & Hershey, J. W. B. Changes in eIF-4D hypusine modification or abundance are not correlated with translational repression in HeLa cells. *J. Biol. Chem.* 261, 12903–12906 (1986).


149. Frolova, L. Y. et al. Mutations in the highly conserved GGQ motif of class I polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. RNA 5, 1014–1020 (1999).


172. Dong, J. *et al.* The essential ATP-binding cassette protein RLI1 functions in translation by


212. Firth, A. E. et al. Ribosomal frameshifting used in influenza A virus expression occurs within the sequence UCC-UUU-CGU and is in the +1 direction. Open Biol. 2, (2012).


References


References


References


310. Michel, A. M. *et al.* Observation of dually decoded regions of the human genome using


344. Williams, C. C., Jan, C. H. & Weissman, J. S. Targeting and plasticity of mitochondrial


370. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium*...


389. O’Connor, P. B. F., Li, G.-W., Weissman, J. S., Atkins, J. F. & Baranov, P. V. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome


462. Gao, J. et al. MYH9 is an Essential Factor for Porcine Reproductive and Respiratory


References


Snijder, E. J., Van Tol, H., Roos, N. & Pedersen, K. W. Non-structural proteins 2 and 3


Sun, Y. *et al.* Nonstructural protein 11 of porcine reproductive and respiratory syndrome virus suppresses both MAVS and RIG-I expression as one of the mechanisms to antagonize Type I interferon production. *PloS One* **11**, e0168314 (2016).


References


596. Van Marle, G. *et al.* Arterivirus discontinuous mRNA transcription is guided by base pairing


References


635. Sagong, M. & Lee, C. Porcine reproductive and respiratory syndrome virus nucleocapsid protein modulates interferon-β production by inhibiting IRF3 activation in immortalized


692. Lehmann, K. C. *et al.* Arterivirus NSP12 versus the coronavirus NSP16 29-O-methyltransferase: Comparison of the C-terminal cleavage products of two nidovirus pp1ab
References

712. Zeng, N. *et al.* Transcriptome Analysis Reveals Dynamic Gene Expression Profiles in


731. Hwang, J. Y. & Buskirk, A. R. A ribosome profiling study of mRNA cleavage by the


EU PRRSV genome sequence

Appendices
1 EU PRRSV genome sequence
>PRRSV_EU_based_on_KJ127878
ATGATGTGTAGGGTATTCCCCCTACATACACGACACTTCTAGTGTTTGTGTACCTTGGA
GGCGTGGGTACAGCCCCGCCCCACCCCTTGGCCCCTGTTCTAGCCCAACAGGTATCCT
TCTCTCTCGGGGCGAGTGCGCCGCCTGCTGCTCCCTTGCAGCGGGAAGGACCTCCCGA
GTATTTCCGGAGAGCACCTGCTTTACGGGATCTCCACCCTTTAACCATGTCTGGGACG
TTCTCCCGGTGCATGTGCACCCCGGCTGCCCGGGTATTTTGGAACGCCGGCCAAGTCT
TTTGCACACGGTGTCTCAGTGCGCGGTCTCTTCTCTCTCCAGAGCTTCAGGACACTGAC
CTCGGCGCAGTTGGCTTGTTTTACAAGCCTAGGGACAAGCTTCACTGGAAAGTCCCTA
TCGGCATCCCTCAGGTGGAATGTACTCCATCCGGGTGCTGTTGGCTCTCAGCTGTTTTC
CCTTTGGCGCGTATGACCTCCGGCAATCACAACTTCCTCCAACGACTTGTGAAGGTTG
CTGATGTTTTGTACCGTGACGGTTGCTTGGCACCTCGACACCTTCGTGAACTCCAAGTT
TACGAGCGCGGCTGCAACTGGTACCCGATCACGGGACCCGTGCCCGGGATGGGTTTGT
TTGCGAACTCCATGCACGTATCCGACCAGCCGTTCCCTGGTGCCACCCATGTGTTGAC
TAACTCGCCTTTGCCTCAACAGGCTTGTCGGCAGCCGTTCTGTCCATTTGAGGAGGCTC
ATTCTAGCGTGTACAGGTGGAAGAAATTTGTGGTTTTCACGGACTCCTCCCTCAACGG
TCGATCTCGCATGATGTGGACGCCGGAATCCGATGATTCAGCCGCCCTGGAGGTGCTA
CCGCCTGAGTTAGAACGTCAGGTTGAAATCCTCATTCGGAGTTTTCCTGCTCATCACTC
TGTCGACCTGGCCGACTGGGAGCTCACTGAGTCCCCTGAGAACGGTTTTTCCTTCAAC
ACGTCTCATTCTTGCGGTCACCTTGTCCAAAACCCCGACGTGTTTGATGGCAAGTGCT
GGCTCTCCTGCTTTTTGGGCCAGTCGGCCGAAGTGCGCTGCCATGAGGAACATCTAGC
TGACGCCTTCGGTTACCAAACCAAGTGGGGCGTGCATGGTAAGTACCTCCAGCGCAG
GCTTCAAGTTCACGGCATTCGTGCTGTAGTCGATCCTGACGGTCCCATTCACGTTGAA
GCGCTGTCTTGCCCCCAGTCTTGGATCAGGCACCTGACTCTGAATGATGATGTCACCC
CAGGATTCGTTCGCCTGACATCCCTTCGCATTGTGCCGAACACAGAGCCTACCACTTC
CCGGATCTTTCGGTTTGGAGCGCATAAGTGGTATGGCGCTGCCGGCAAACGGGCTCGT
GCTAAGCGTGCCGCTAAAAGTGAGAAGGATTCGGCTCCCACCCCCAAGGTTGCCCTGC
CTGTCCCCACCTGTGGAATTACCACCTACTCTCCACCGACAGACGGGTCTTGTGGTTG
GCATGCCCTTGCCGCCATAATGAACCGGATGATAAATGGTGACTTCACGTCCCCTCTG
ACTCAGTACAACAGACCAGAGGATGATTGGGCTTCTGATTATGATCTTGTTCAGGCGA
TTCAATGTCTACAACTGCCTGCTACCGTGGTTCGGAATCGCGCCTGTCCTAACGCCAA
GTACCTTATAAAACTTAACGGAGTTCACTGGGAGGTAGAGGTGAGGTCTGGAATGGC
TCCTCGCTCCCTTTCTCGTGAATGTGTGGTTGGCGTTTGCTCTGAAGGCTGTGTCGCAC
CGCCTTATCCAGCAGACGGGCTACCTAAACGTGCACTCGAGGCCTTGGCGTCTGCTTA
CAGACTACCCTCCGATTGTGTTAGCTCTGGTATTGCTGACTTTCTTGCTAATCCACCTC
CTCAGGAATTCTGGACCCTCGACAAAATGTTGACCTCCCCGTCACCAGAGCGGTCCGG
CTTCTCTAGTTTGTATAAATTACTATTAGAGGTTGTTCCGCAAAAATGCGGTGCCACG
GAAGGGGCTTTCATCTATGCTGTTGAGAGGATGTTGAAGGATTGTCCGAGCTCCAAAC
AGGCCATGGCCCTTCTGGCAAAAATTAAAGTTCCATCCTCAAAGGCCCCGTCTGTGTC
CCTGGACGGGTGTTTCCCTACGGATGTTCCAGCCGACTTCGAGCCAGCATCTCCGGAA
AGGCCAGCTGGTCTAATTAACCTGGTAGGCGGGAATTTGTCCCCCTCAGACTCCATGA
AAGAAAACATGCTCAATAGCCGGGAAGACGAACCACTGGATTTGTCCCAACCAGCAC
CAGCTGCCACAACGACCCTTGTGAGAGAGCAAACACCCGACAACCCAGGTTCTGATG
CCGGTGCCCTCCCCGTCACCGTTCGAGAATTTGTCCCGACGGGGCCTATACTCCGTCA
TGTTGAGCACTGCGGCACGGAGTCGGGCGACAGCAGTTCGCCTTTGGACCAGTCTGAT
GCGCAAACCCTGGACCAGCCTTTAAATCTATCCCTGGCCGCTTGGCCAGTGAGGACCA
CCGCGTCTGACCCTGGCTGGGTCCACGGTAGGCGCGAGCCTGTTTTTGTAAAGCCTCG
- 279 -


EU PRRSV genome sequence

TGTTTTCTTTTGACACTTTTGTGCTTGCAATGGGCACCCCTGGTGCTTGACAGGTGTGG
TGATTAGACTCTTCAGCAGCTTCCTCAACCCCAACACTTTCTCTGGCAGTCTACGC
ACTCGGGGTTGTCTTGCTGTTTGCGTTCGCTGGCAATCTGGGACTTTTTGCTGCTGAGATTGTCT
GAAATTGCTCAAGCTCTTTGCACTAATGCTTCTTTCATCTAGGGTCCTTGCAAGGACAG
TTGTTGGCTCCATCCATACATCGTGACTCTCATACCCCTCCGCTGGATTCTGTGTTAT
TCAAAATACCGGTTGCTCGGACTTTGTTTCTGCTTGCTGATGGGAGAGTTCTCAACGCCTT
CTTCTACAGTATTGTGAAGGGAATTCGAAATCTCAAGGCTCTTGTTGGGCTAAG
ATGAAAGCTGGTCCTACTAACGGGCTGTCTTCTAGGCTGTCAGGT
ATTTTTGGTTCAGGCTTAACTCTGTAGCTTTTGCAATTGTCAGCTAGGCTACCTTG
AGTTTTGGTGCTGAGCTTTTCTGTTCTTGAGCAAAAATCTGCGTCCGGCAAAATCAC
ACGTGACACGCGCAATCCATCAACCAACAGGGATGGTTCACAGGCTATAGGAGGCTG
TGAACAGGCTTAAACTGTTAGCCGCCAGCGGCTTGACCCGCTGTGGCCGCGGAACATTGAAACGG
CCATCCTAGTCCGCAATTGGCCTGCAGTGGTGTACCCAGGCATTCATGAAGAAGGCTTGGAAGTCCCCAATTGCCTTGGGGAAAAACAAATTCAAGGAGCTGCATTGCACTGTCGCCGCAAGATGTCTTGAGGCCGACTTGGCCTCC
TGGTGCACCGCAGCTTGGGGTCCTACGCCTAGTCGCACTCGTAGTGGTGAAGGACTTGGCAGGAAG
ATTCAATGCAATCCTTGTAGCTTCTCTGACCTGAGGCTCAATAGGGCGGCTCTTCA
TTCTCTTTGGAGCAATGGGCTAAGGTTGATGTAGGAGCCTC
TGGGTTTCGAGATTTGGAAGACTTGGCAGGACAGGGCTATGAGGACCTGCCC
AATCCTGATGG

- 281 -
2 Oligonucleotide sequences

Primers are based on those in the TruSeq Small RNA Library Preparation Kit (Illumina)
Coordinates of regions of PRRSV genomes used for analyses

All coordinates denote the region within which the 5’ end of the reads must map to be included.

3.1 General plots and analyses

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Virus</th>
<th>Region name</th>
<th>Coordinates on viral genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length distribution on specific regions of viral genome</td>
<td>NA</td>
<td>ORF1a</td>
<td>179-7669</td>
</tr>
<tr>
<td>Length distribution on specific regions of viral genome</td>
<td>NA</td>
<td>ORF1b</td>
<td>7696-11595</td>
</tr>
<tr>
<td>Length distribution on specific regions of viral genome</td>
<td>NA</td>
<td>sgORFs</td>
<td>12062-15246</td>
</tr>
<tr>
<td>Phasing on specific regions of viral genome</td>
<td>NA</td>
<td>ORF1a</td>
<td>179-3862 and 4391-7669 combined</td>
</tr>
<tr>
<td>Phasing on specific regions of viral genome</td>
<td>NA</td>
<td>ORF1b</td>
<td>7696-10407</td>
</tr>
<tr>
<td>Phasing on specific regions of viral genome</td>
<td>NA</td>
<td>N</td>
<td>14890-15246</td>
</tr>
<tr>
<td>sgRNA:ORF1a read density ratio</td>
<td>NA</td>
<td>ORF1a</td>
<td>179-3862</td>
</tr>
<tr>
<td>sgRNA:ORF1a read density ratio</td>
<td>NA</td>
<td>sgRNA</td>
<td>12062-15246</td>
</tr>
<tr>
<td>Length distribution on specific regions of viral genome</td>
<td>EU</td>
<td>ORF1a</td>
<td>210-7157</td>
</tr>
<tr>
<td>Length distribution on specific regions of viral genome</td>
<td>EU</td>
<td>ORF1b</td>
<td>7184-11548</td>
</tr>
<tr>
<td>Length distribution on specific regions of viral genome</td>
<td>EU</td>
<td>sgORFs</td>
<td>11562-14747</td>
</tr>
<tr>
<td>Phasing on specific regions of viral genome</td>
<td>EU</td>
<td>ORF1a</td>
<td>210-3320 and 3849-7157 combined</td>
</tr>
<tr>
<td>Phasing on specific regions of viral genome</td>
<td>EU</td>
<td>ORF1b</td>
<td>7184-11548</td>
</tr>
<tr>
<td>Phasing on specific regions of viral genome</td>
<td>EU</td>
<td>N</td>
<td>14364-14747</td>
</tr>
<tr>
<td>sgRNA:ORF1a read density ratio</td>
<td>EU</td>
<td>ORF1a</td>
<td>210-3320</td>
</tr>
<tr>
<td>sgRNA:ORF1a read density ratio</td>
<td>EU</td>
<td>sgRNA</td>
<td>11562-14747</td>
</tr>
<tr>
<td>Analysis</td>
<td>Region</td>
<td>Location</td>
<td>Coordinates</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>--------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>nsp2 density normalisation-based frameshift efficiency calculation</td>
<td>NA</td>
<td>upstream</td>
<td>1922-3691</td>
</tr>
<tr>
<td>nsp2 density normalisation-based frameshift efficiency calculation</td>
<td>NA</td>
<td>downstream</td>
<td>4568-7495</td>
</tr>
<tr>
<td>ORF1ab frameshift efficiency calculation</td>
<td>NA</td>
<td>upstream</td>
<td>4568-7495</td>
</tr>
<tr>
<td>ORF1ab frameshift efficiency calculation</td>
<td>NA</td>
<td>downstream</td>
<td>7723-9147</td>
</tr>
<tr>
<td>ORF1ab frameshift efficiency calculation</td>
<td>EU</td>
<td>upstream</td>
<td>4026-6983</td>
</tr>
<tr>
<td>ORF1ab frameshift efficiency calculation</td>
<td>EU</td>
<td>downstream</td>
<td>7211-8755</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites (and for phasing-based nsp2 FS efficiency calculations)</td>
<td>NA</td>
<td>nsp2 upstream</td>
<td>1922-3691</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites (and for phasing-based nsp2 FS efficiency calculations)</td>
<td>NA</td>
<td>nsp2 transframe</td>
<td>3912-4352</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites</td>
<td>NA</td>
<td>ORF1ab downstream</td>
<td>7723-9147</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites (and for phasing-based nsp2 FS efficiency calculations)</td>
<td>EU</td>
<td>nsp2 upstream</td>
<td>1926-3149</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites (and for phasing-based nsp2 FS efficiency calculations)</td>
<td>EU</td>
<td>nsp2 transframe</td>
<td>3370-3810</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites</td>
<td>EU</td>
<td>nsp2 downstream/ORF1ab upstream</td>
<td>4026-6983</td>
</tr>
<tr>
<td>Phasing in regions upstream, transframe, or downstream relative to FS sites</td>
<td>EU</td>
<td>ORF1ab downstream</td>
<td>7211-8755</td>
</tr>
<tr>
<td>downstream region for nsp2 phasing-based frameshift efficiency bootstrap resampling negative control</td>
<td>NA</td>
<td>ctrl</td>
<td>4569-5009</td>
</tr>
<tr>
<td>downstream region for nsp2 phasing-based frameshift efficiency bootstrap resampling negative control</td>
<td>EU</td>
<td>ctrl</td>
<td>4027-4467</td>
</tr>
<tr>
<td>chimeric read analysis for novel transcript discovery</td>
<td>NA</td>
<td>TRS leader</td>
<td>179-193</td>
</tr>
<tr>
<td>chimeric read analysis for novel transcript discovery</td>
<td>EU</td>
<td>TRS leader</td>
<td>213-221</td>
</tr>
<tr>
<td>ORF1ab frameshift efficiency calculation using cut-down downstream region</td>
<td>NA</td>
<td>upstream</td>
<td>4568-7495</td>
</tr>
<tr>
<td>ORF1ab frameshift efficiency calculation using cut-down downstream region</td>
<td>NA</td>
<td>downstream</td>
<td>7723-8440</td>
</tr>
</tbody>
</table>
### 3.2 Transcript abundance calculations

#### 3.2.1 For all calculations except investigation of the “heteroclite” region

<table>
<thead>
<tr>
<th>Virus</th>
<th>Transcript</th>
<th>Region start (decum.)</th>
<th>Region end (decum.)</th>
<th>Input junctions counted (chimeric reads method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>leader</td>
<td>16</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>gRNA</td>
<td>206</td>
<td>8479</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA1</td>
<td>8510</td>
<td>9176</td>
<td>184-8493</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA2</td>
<td>9207</td>
<td>10410</td>
<td>184-9189, 184-9190</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA3</td>
<td>10441</td>
<td>10638</td>
<td>183-10422, 182-10422</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA4</td>
<td>10669</td>
<td>10771</td>
<td>184-10652, 183-10650</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA5</td>
<td>10802</td>
<td>11368</td>
<td>181-10782, 183-10783, 185-10786, 179-10780</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA6</td>
<td>11452</td>
<td>11581</td>
<td>184-11375, 184-11435</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA7</td>
<td>11635</td>
<td>11773</td>
<td>184-11599, 182-11593, 184-11618, 184-11619, 184-11617</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1b-sgRNA8</td>
<td>11804</td>
<td>12032</td>
<td>185-11788, 183-11782, 184-11787</td>
</tr>
<tr>
<td>NA</td>
<td>sgRNA-Gp2a-iORF</td>
<td>12188</td>
<td>12592</td>
<td>184-12170, 184-12171, 184-12173, 183-12170, 184-12172</td>
</tr>
<tr>
<td>NA</td>
<td>sgRNA2</td>
<td>12623</td>
<td>13216</td>
<td>184-12606, 183-12605, 184-12605, 186-12608, 185-12608, 184-12604, 182-12604</td>
</tr>
<tr>
<td>NA</td>
<td>sgRNA3</td>
<td>13247</td>
<td>13616</td>
<td>184-13230, 184-13231, 183-13229, 188-13234, 184-13228, 184-13229</td>
</tr>
<tr>
<td>NA</td>
<td>sgRNA-Gp4-iORF</td>
<td>13647</td>
<td>13727</td>
<td>193-13639, 184-13629, 184-13631</td>
</tr>
<tr>
<td>NA</td>
<td>sgRNA5</td>
<td>14368</td>
<td>14745</td>
<td>183-14348, 193-14362, 184-14353, 193-14360, 179-14346, 181-14347, 185-14353</td>
</tr>
</tbody>
</table>
### Coordinates of regions of PRRSV genomes used for analyses

<table>
<thead>
<tr>
<th></th>
<th>Transcript</th>
<th>Region start (decum.)</th>
<th>Region end (decum.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>sgRNA6</td>
<td>14890</td>
<td>15362</td>
</tr>
<tr>
<td>EU</td>
<td>leader</td>
<td>16</td>
<td>176</td>
</tr>
<tr>
<td>EU</td>
<td>gRNA</td>
<td>237</td>
<td>8785</td>
</tr>
<tr>
<td>EU</td>
<td>ORF1b-sgRNA1</td>
<td>8816</td>
<td>10259</td>
</tr>
<tr>
<td>EU</td>
<td>ORF1b-sgRNA2</td>
<td>10290</td>
<td>11411</td>
</tr>
<tr>
<td>EU</td>
<td>sgRNA2</td>
<td>12180</td>
<td>12618</td>
</tr>
<tr>
<td>EU</td>
<td>sgRNA3</td>
<td>12649</td>
<td>13218</td>
</tr>
<tr>
<td>EU</td>
<td>sgRNA4</td>
<td>13249</td>
<td>13819</td>
</tr>
</tbody>
</table>

#### 3.2.2 For investigation of the “heteroclite” region

<table>
<thead>
<tr>
<th>Virus</th>
<th>Transcript</th>
<th>Region start (decum.)</th>
<th>Region end (decum.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>gRNA</td>
<td>1762</td>
<td>8479</td>
</tr>
<tr>
<td>NA</td>
<td>hetero</td>
<td>206</td>
<td>1731</td>
</tr>
<tr>
<td>EU</td>
<td>gRNA</td>
<td>1765</td>
<td>8785</td>
</tr>
<tr>
<td>EU</td>
<td>hetero</td>
<td>237</td>
<td>1734</td>
</tr>
</tbody>
</table>
### 3.3 Translation level calculations

#### 3.3.1 For all calculations except investigation of the “heteroclite” region

<table>
<thead>
<tr>
<th>Virus</th>
<th>ORF</th>
<th>Region start</th>
<th>Region end</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>uORF1</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>NA</td>
<td>uORF2</td>
<td>29</td>
<td>85</td>
</tr>
<tr>
<td>NA</td>
<td>uORF3</td>
<td>48</td>
<td>170</td>
</tr>
<tr>
<td>NA</td>
<td>ORF1a</td>
<td>179</td>
<td>3862</td>
</tr>
<tr>
<td>NA</td>
<td>nsp2-iORF</td>
<td>2455</td>
<td>2520</td>
</tr>
<tr>
<td>NA</td>
<td>nsp3-4-iORF</td>
<td>5589</td>
<td>5630</td>
</tr>
<tr>
<td>NA</td>
<td>nsp9-genomic</td>
<td>7696</td>
<td>8475</td>
</tr>
<tr>
<td>NA</td>
<td>nsp9-iORF1</td>
<td>8513</td>
<td>8551</td>
</tr>
<tr>
<td>NA</td>
<td>nsp9-iORF2</td>
<td>9083</td>
<td>9151</td>
</tr>
<tr>
<td>NA</td>
<td>nsp9-sgRNA1</td>
<td>8491</td>
<td>9171</td>
</tr>
<tr>
<td>NA</td>
<td>nsp9-10-sgRNA2</td>
<td>9190</td>
<td>10404</td>
</tr>
<tr>
<td>NA</td>
<td>nsp10-sgRNA3</td>
<td>10426</td>
<td>10629</td>
</tr>
<tr>
<td>NA</td>
<td>nsp10-iORF1</td>
<td>10460</td>
<td>10513</td>
</tr>
<tr>
<td>NA</td>
<td>nsp10-iORF2</td>
<td>10640</td>
<td>10765</td>
</tr>
<tr>
<td>NA</td>
<td>nsp10-sgRNA4</td>
<td>10651</td>
<td>10764</td>
</tr>
<tr>
<td>NA</td>
<td>nsp10-11-sgRNA5</td>
<td>10789</td>
<td>11355</td>
</tr>
<tr>
<td>NA</td>
<td>nsp10-iORF3</td>
<td>10833</td>
<td>10862</td>
</tr>
<tr>
<td>NA</td>
<td>nsp11-iORF1</td>
<td>11057</td>
<td>11155</td>
</tr>
<tr>
<td>NA</td>
<td>nsp11-iORF2</td>
<td>11273</td>
<td>11353</td>
</tr>
<tr>
<td>NA</td>
<td>nsp11-iORF3</td>
<td>11433</td>
<td>11483</td>
</tr>
<tr>
<td>NA</td>
<td>nsp11-sgRNA6</td>
<td>11437</td>
<td>11574</td>
</tr>
<tr>
<td>NA</td>
<td>nsp12-sgRNA7</td>
<td>11620</td>
<td>11766</td>
</tr>
<tr>
<td>NA</td>
<td>nsp12-iORF1</td>
<td>11616</td>
<td>11735</td>
</tr>
<tr>
<td>NA</td>
<td>nsp12-iORF2</td>
<td>11633</td>
<td>11767</td>
</tr>
<tr>
<td>NA</td>
<td>nsp12-sgRNA8</td>
<td>11791</td>
<td>12024</td>
</tr>
<tr>
<td>NA</td>
<td>Gp2a</td>
<td>12062</td>
<td>12151</td>
</tr>
<tr>
<td>NA</td>
<td>E</td>
<td>12067</td>
<td>12150</td>
</tr>
<tr>
<td>NA</td>
<td>Gp2aiORF</td>
<td>12285</td>
<td>12374</td>
</tr>
<tr>
<td>NA</td>
<td>Gp3</td>
<td>12685</td>
<td>13209</td>
</tr>
<tr>
<td>NA</td>
<td>Gp4</td>
<td>13230</td>
<td>13610</td>
</tr>
<tr>
<td>NA</td>
<td>Gp4-iORF</td>
<td>13630</td>
<td>13695</td>
</tr>
<tr>
<td>NA</td>
<td>Gp5</td>
<td>13777</td>
<td>14331</td>
</tr>
<tr>
<td>NA</td>
<td>5a</td>
<td>13767</td>
<td>13916</td>
</tr>
<tr>
<td>NA</td>
<td>M</td>
<td>14364</td>
<td>14738</td>
</tr>
<tr>
<td>NA</td>
<td>M-iORF</td>
<td>14380</td>
<td>14445</td>
</tr>
<tr>
<td>NA</td>
<td>N</td>
<td>14878</td>
<td>15243</td>
</tr>
<tr>
<td>EU</td>
<td>uORF</td>
<td>124</td>
<td>204</td>
</tr>
<tr>
<td>EU</td>
<td>ORF1a</td>
<td>213</td>
<td>3320</td>
</tr>
<tr>
<td>EU</td>
<td>nsp2-iORF</td>
<td>2153</td>
<td>2215</td>
</tr>
</tbody>
</table>
### Coordinates of regions of PRRSV genomes used for analyses

<table>
<thead>
<tr>
<th>EU</th>
<th>nsp5-iORF</th>
<th>5993</th>
<th>6025</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU</td>
<td>nsp9-genomic</td>
<td>7187</td>
<td>8783</td>
</tr>
<tr>
<td>EU</td>
<td>nsp9-iORF1</td>
<td>7201</td>
<td>7224</td>
</tr>
<tr>
<td>EU</td>
<td>nsp9-iORF2</td>
<td>8392</td>
<td>8412</td>
</tr>
<tr>
<td>EU</td>
<td>nsp9-10-sgRNA1</td>
<td>8804</td>
<td>10255</td>
</tr>
<tr>
<td>EU</td>
<td>nsp10-iORF1</td>
<td>9192</td>
<td>9209</td>
</tr>
<tr>
<td>EU</td>
<td>nsp10-iORF2</td>
<td>9783</td>
<td>9911</td>
</tr>
<tr>
<td>EU</td>
<td>nsp10-iORF3</td>
<td>10251</td>
<td>10295</td>
</tr>
<tr>
<td>EU</td>
<td>nsp10-11-sgRNA2</td>
<td>10277</td>
<td>11407</td>
</tr>
<tr>
<td>EU</td>
<td>nsp11-iORF</td>
<td>10951</td>
<td>10998</td>
</tr>
<tr>
<td>EU</td>
<td>Gp2a</td>
<td>11565</td>
<td>12146</td>
</tr>
<tr>
<td>EU</td>
<td>E</td>
<td>11570</td>
<td>11776</td>
</tr>
<tr>
<td>EU</td>
<td>Gp3</td>
<td>12173</td>
<td>12613</td>
</tr>
<tr>
<td>EU</td>
<td>Gp4</td>
<td>12715</td>
<td>13215</td>
</tr>
<tr>
<td>EU</td>
<td>Gp5</td>
<td>13263</td>
<td>13814</td>
</tr>
<tr>
<td>EU</td>
<td>5a</td>
<td>13268</td>
<td>13393</td>
</tr>
<tr>
<td>EU</td>
<td>M</td>
<td>13856</td>
<td>14227</td>
</tr>
<tr>
<td>EU</td>
<td>N</td>
<td>14367</td>
<td>14747</td>
</tr>
</tbody>
</table>

### 3.3.2 For investigation of the “heteroclite” region

<table>
<thead>
<tr>
<th>Virus</th>
<th>ORF</th>
<th>Region start</th>
<th>Region end</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>heteroclite</td>
<td>179</td>
<td>1729</td>
</tr>
<tr>
<td>NA</td>
<td>before FS</td>
<td>1742</td>
<td>3862</td>
</tr>
<tr>
<td>NA</td>
<td>after FS</td>
<td>4391</td>
<td>7669</td>
</tr>
<tr>
<td>EU</td>
<td>heteroclite</td>
<td>213</td>
<td>1736</td>
</tr>
<tr>
<td>EU</td>
<td>before FS</td>
<td>1749</td>
<td>3320</td>
</tr>
<tr>
<td>EU</td>
<td>after FS</td>
<td>3852</td>
<td>7157</td>
</tr>
</tbody>
</table>
4 Composition of mapping databases

4.1 Host databases

<table>
<thead>
<tr>
<th>Species</th>
<th>Database</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. sabaeus</td>
<td>rRNA</td>
<td>NR_003287.2, NR_023379.1, NR_003285.2, NR_003286.2, AY603036.1, AF420058.1, AF420040.1, AY633510.1, AF352382.1, L35185.1, DQ983926.1, KJ193255.1, M30951.1, M30950.1, M30952.1, KJ193272.1, KJ193259.1, KJ193258.1, KJ193256.1, KJ193255.1, KJ193045.1, KJ193042.1, KJ193044.1, KJ193041.1, KJ193019.1, KJ193018.1, KJ193017.1 and AF420040.1</td>
</tr>
<tr>
<td>C. sabaeus</td>
<td>mRNA</td>
<td>Database was compiled by Dr Andrew Firth from available C. sabaeus NCBI RefSeq mRNAs, after removing transcripts with annotated changes in frame</td>
</tr>
<tr>
<td>C. sabaeus</td>
<td>ncRNA</td>
<td>Ensembl Chlorocebus_sabaeus.ChlSab1.1.ncrna.fa release</td>
</tr>
<tr>
<td>C. sabaeus</td>
<td>gDNA</td>
<td>Genome assembly ChlSab1.1</td>
</tr>
</tbody>
</table>

4.2 Contaminants database

<table>
<thead>
<tr>
<th>Potential contaminant</th>
<th>Accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli strain BL21-AI (protein expression strain used in lab)</td>
<td>CP047231.1</td>
</tr>
<tr>
<td>Genomes of viruses investigated in the lab (redacted for confidentiality)</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

Appendices
### Composition of mapping databases

<table>
<thead>
<tr>
<th>Mycoplasma</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_000908, NC_000912, NC_002771, NC_004432, NC_004829, NC_005364, NC_006360, NC_006908, NC_007294, NC_007295, NC_007332, NC_007633, NC_009497, NC_011025, NC_012806, NC_013511, NC_013948, NC_014014, NC_014448, NC_014552, NC_014751, NC_014760, NC_014921, NC_014970, NC_015153, NC_015155, NC_015431, NC_015725, NC_015946, NC_016638, NC_016807, NC_016829, NC_017502, NC_017503, NC_017504, NC_017509, NC_017519, NC_017520, NC_017521, NC_018077, NC_018149, NC_018406, NC_018407, NC_018408, NC_018409, NC_018410, NC_018411, NC_018412, NC_018413, NC_018495, NC_018496, NC_018497, NC_018498, NC_019552, NC_019949, NC_020076, NC_021002, NC_021007, NC_021025, NC_021083, NC_021283, NC_021831, NC_022575, NC_022807, NC_023062, NC_000912, NC_002771, NC_004432, NC_006908, NC_007294, NC_007295, NC_007332, NC_007633, NC_009497, NC_011025, NC_012806, NC_013511, NC_013948, NC_014014, NC_014448, NC_014552, NC_014751, NC_014760, NC_014921, NC_014970, NC_015153, NC_015155, NC_015431, NC_015725, NC_015946, NC_016638, NC_016807, NC_016829, NC_017502, NC_017503, NC_017504, NC_017509, NC_017519, NC_017520, NC_017521, NC_018077, NC_018149, NC_018406, NC_018407, NC_018408, NC_018409, NC_018410, NC_018411, NC_018412, NC_018413, NC_018495, NC_018496, NC_018497, NC_018498, NC_019552, NC_019949, NC_020076</td>
</tr>
</tbody>
</table>
5 Viral ORFs detected by PRICE

5.1 NA PRRSV

Output from running PRICE on four separate groups of libraries, collated for presentation in one table (blank rows left to aid visibility). “Lib 1”, “Lib 2”, “Lib 3” and “Lib 4” are WT-1, WT-2, KO2-1 and KO2-2 for the 6hpi and 12hpi groups, while for the 9hpi_WT and 9hpi_KO2 groups they are replicates one, two, three and four. Table spans multiple pages, novel ORFs discovered in this study are written in bold, and p values are not corrected for multiple testing.

<table>
<thead>
<tr>
<th>Name</th>
<th>Timepoint</th>
<th>Location</th>
<th>Candidate Location</th>
<th>p value</th>
<th>Lib 1 reads</th>
<th>Lib 2 reads</th>
<th>Lib 3 reads</th>
<th>Lib 4 reads</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>uORF1</td>
<td>6hpi</td>
<td>KC469618.1+:23-56</td>
<td>KC469618.1+:23-56</td>
<td>0.00079</td>
<td>21.7</td>
<td>37.5</td>
<td>10</td>
<td>13.4</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td>9hpi_WT</td>
<td>KC469618.1+:23-56</td>
<td>KC469618.1+:23-56</td>
<td>1.1E-05</td>
<td>5641.2</td>
<td>3986.4</td>
<td>612.8</td>
<td>4252.5</td>
<td>14492.9</td>
</tr>
<tr>
<td></td>
<td>9hpi_KO2</td>
<td>KC469618.1+:23-56</td>
<td>KC469618.1+:23-56</td>
<td>0.00024</td>
<td>7927.5</td>
<td>920.4</td>
<td>1595.1</td>
<td>396.2</td>
<td>10839.2</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1+:23-56</td>
<td>KC469618.1+:23-56</td>
<td>0.00203</td>
<td>38119.9</td>
<td>2972.8</td>
<td>18121.5</td>
<td>17117.3</td>
<td>76331.5</td>
</tr>
<tr>
<td>uORF2</td>
<td>9hpi_WT</td>
<td>KC469618.1+:40-103</td>
<td>KC469618.1+:28-103</td>
<td>0.0331</td>
<td>1134.8</td>
<td>570</td>
<td>50.4</td>
<td>452.5</td>
<td>2207.8</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1+:40-103</td>
<td>KC469618.1+:28-103</td>
<td>0.02152</td>
<td>4515.2</td>
<td>334.1</td>
<td>2725.6</td>
<td>1693.3</td>
<td>9268.2</td>
</tr>
<tr>
<td>uORF3</td>
<td>9hpi_WT</td>
<td>KC469618.1+:59-188</td>
<td>KC469618.1+:59-188</td>
<td>0.00611</td>
<td>1299.9</td>
<td>588.1</td>
<td>131.9</td>
<td>608.8</td>
<td>2628.7</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1+:59-188</td>
<td>KC469618.1+:59-188</td>
<td>0.00611</td>
<td>3612.8</td>
<td>330.8</td>
<td>2014.8</td>
<td>1378.1</td>
<td>7336.6</td>
</tr>
<tr>
<td>ORF1a</td>
<td>6hpi</td>
<td>KC469618.1+:190-7702</td>
<td>KC469618.1+:190-7702</td>
<td>0</td>
<td>6421.8</td>
<td>4893.4</td>
<td>2715.6</td>
<td>2970.5</td>
<td>17001.3</td>
</tr>
<tr>
<td></td>
<td>9hpi_WT</td>
<td>KC469618.1+:190-7702</td>
<td>KC469618.1+:190-7702</td>
<td>0</td>
<td>208013.1</td>
<td>77775.7</td>
<td>11123.6</td>
<td>47026.7</td>
<td>343939.2</td>
</tr>
<tr>
<td></td>
<td>9hpi_KO2</td>
<td>KC469618.1+:190-7702</td>
<td>KC469618.1+:190-7702</td>
<td>0</td>
<td>246385.1</td>
<td>24666.5</td>
<td>32872.7</td>
<td>7344.6</td>
<td>311269</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1+:1744-7702</td>
<td>KC469618.1+:190-7702</td>
<td>0</td>
<td>36401.3</td>
<td>12105.2</td>
<td>36330.8</td>
<td>30734.5</td>
<td>115571.8</td>
</tr>
<tr>
<td>nsp2-iORF</td>
<td>12hpi</td>
<td>KC469618.1+:2466-2538</td>
<td>KC469618.1+:2388-2538</td>
<td>0.01679</td>
<td>30.3</td>
<td>48.3</td>
<td>36.3</td>
<td>79.4</td>
<td>194.4</td>
</tr>
<tr>
<td>nsp2TF</td>
<td>6hpi</td>
<td>KC469618.1:+3908-4397</td>
<td>KC469618.1:+3908-4397</td>
<td>2.8E-15</td>
<td>83</td>
<td>81</td>
<td>5.3</td>
<td>10.8</td>
<td>180.2</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>---------</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>9hpi_WT</td>
<td>KC469618.1:+3968-4397</td>
<td>0</td>
<td>5192.6</td>
<td>2014</td>
<td>268.1</td>
<td>1574.2</td>
<td>9048.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12hpi</td>
<td>KC469618.1:+3881-4397</td>
<td>0</td>
<td>1423.6</td>
<td>466.8</td>
<td>227.8</td>
<td>214.8</td>
<td>2333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nsp3-4-iORF</td>
<td>6hpi</td>
<td>KC469618.1:+5600-5648</td>
<td>KC469618.1:+5600-5648</td>
<td>0.01219</td>
<td>4.8</td>
<td>2</td>
<td>2.8</td>
<td>5.8</td>
<td>15.4</td>
</tr>
<tr>
<td>ORF1b</td>
<td>6hpi</td>
<td>KC469618.1:+7695-12072</td>
<td>KC469618.1:+7695-12072</td>
<td>0</td>
<td>443</td>
<td>311.3</td>
<td>202.2</td>
<td>223.4</td>
<td>1179.9</td>
</tr>
<tr>
<td>9hpi_WT</td>
<td>KC469618.1:+7686-12072</td>
<td>0</td>
<td>32455.9</td>
<td>14958.7</td>
<td>2579.5</td>
<td>8450.5</td>
<td>58444.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9hpi_KO2</td>
<td>KC469618.1:+7764-12072</td>
<td>0</td>
<td>37034.5</td>
<td>4555.2</td>
<td>6332</td>
<td>1309.9</td>
<td>49231.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12hpi</td>
<td>KC469618.1:+8328-12072</td>
<td>0</td>
<td>23626.4</td>
<td>6279</td>
<td>20435.5</td>
<td>16388.4</td>
<td>66729.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nsp9-iORF1</td>
<td>12hpi</td>
<td>KC469618.1:+8524-8569</td>
<td>KC469618.1:+8524-8569</td>
<td>0.01885</td>
<td>15.3</td>
<td>6.5</td>
<td>11.4</td>
<td>11.2</td>
<td>44.4</td>
</tr>
<tr>
<td>nsp9-iORF2</td>
<td>12hpi</td>
<td>KC469618.1:+9094-9169</td>
<td>KC469618.1:+9094-9169</td>
<td>0.00348</td>
<td>58.8</td>
<td>10.5</td>
<td>49.8</td>
<td>37.6</td>
<td>156.7</td>
</tr>
<tr>
<td>nsp10-iORF1</td>
<td>12hpi</td>
<td>KC469618.1:+10471-10531</td>
<td>KC469618.1:+10471-10531</td>
<td>0.04644</td>
<td>208.3</td>
<td>26.7</td>
<td>162.2</td>
<td>141.9</td>
<td>539.1</td>
</tr>
<tr>
<td>nsp10-iORF2</td>
<td>12hpi</td>
<td>KC469618.1:+10651-10819</td>
<td>KC469618.1:+10534-10819</td>
<td>1E-05</td>
<td>564.3</td>
<td>131.8</td>
<td>403.8</td>
<td>413.6</td>
<td>1513.5</td>
</tr>
<tr>
<td>nsp10-iORF3</td>
<td>12hpi</td>
<td>KC469618.1:+10844-10880</td>
<td>KC469618.1:+10844-10880</td>
<td>0.00053</td>
<td>76.9</td>
<td>26.4</td>
<td>1.2</td>
<td>11.4</td>
<td>116</td>
</tr>
<tr>
<td>9hpi_KO2</td>
<td>KC469618.1:+10844-10880</td>
<td>0.0001</td>
<td>72.8</td>
<td>11.9</td>
<td>18.8</td>
<td>1.1</td>
<td>104.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12hpi</td>
<td>KC469618.1:+10844-10880</td>
<td>0.00421</td>
<td>467.9</td>
<td>30</td>
<td>251.5</td>
<td>219.3</td>
<td>968.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nsp11-iORF1</td>
<td>9hpi_WT</td>
<td>KC469618.1:+11068-11173</td>
<td>KC469618.1:+11068-11173</td>
<td>0.00787</td>
<td>424.9</td>
<td>131</td>
<td>42.8</td>
<td>147.2</td>
<td>746</td>
</tr>
<tr>
<td>9hpi_KO2</td>
<td>KC469618.1:+11113-11173</td>
<td>0.04827</td>
<td>362</td>
<td>38.1</td>
<td>67.9</td>
<td>14.7</td>
<td>482.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12hpi</td>
<td>KC469618.1:+11068-11173</td>
<td>2.8E-06</td>
<td>588.2</td>
<td>205</td>
<td>670.4</td>
<td>535.2</td>
<td>1998.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nsp11-iORF2</td>
<td>9hpi_WT</td>
<td>KC469618.1:+11284-11428</td>
<td>KC469618.1:+11284-11428</td>
<td>0.04495</td>
<td>239.7</td>
<td>136.8</td>
<td>13.8</td>
<td>79.3</td>
<td>469.6</td>
</tr>
<tr>
<td>9hpi_KO2</td>
<td>KC469618.1:+11341-11428</td>
<td>0.01816</td>
<td>175.2</td>
<td>53.4</td>
<td>49</td>
<td>12.6</td>
<td>290.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12hpi</td>
<td>KC469618.1:+11284-11428</td>
<td>0.00061</td>
<td>560.3</td>
<td>134.7</td>
<td>398.8</td>
<td>364.5</td>
<td>1458.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nsp11-iORF3</td>
<td>12hpi</td>
<td>KC469618.1:+11414-11501</td>
<td>KC469618.1:+11387-11501</td>
<td>7.2E-12</td>
<td>432.9</td>
<td>106</td>
<td>321.9</td>
<td>331.1</td>
<td>1191.9</td>
</tr>
<tr>
<td>Protein</td>
<td>Time</td>
<td>chr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>nsp12-iORF1</td>
<td>9hpi</td>
<td>KC469618.1:+11624-11753</td>
<td>KC469618.1:+11624-11753</td>
<td>0.03551</td>
<td>145</td>
<td>90.7</td>
<td>6.1</td>
<td>47.1</td>
<td>288.8</td>
</tr>
<tr>
<td>nsp12-iORF2</td>
<td>9hpi</td>
<td>KC469618.1:+11770-12019</td>
<td>KC469618.1:+11635-12019</td>
<td>3E-12</td>
<td>1648</td>
<td>656.4</td>
<td>114.7</td>
<td>573.5</td>
<td>2992.6</td>
</tr>
<tr>
<td></td>
<td>9hpi KO2</td>
<td>KC469618.1:+11776-12019</td>
<td>KC469618.1:+11596-12019</td>
<td>0</td>
<td>2629.5</td>
<td>321.4</td>
<td>454.1</td>
<td>83.2</td>
<td>3488.3</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1:+11644-12019</td>
<td>KC469618.1:+11644-12019</td>
<td>0</td>
<td>4163.7</td>
<td>1418</td>
<td>4032.3</td>
<td>3302.2</td>
<td>12916.1</td>
</tr>
<tr>
<td>E</td>
<td>9hpi</td>
<td>KC469618.1:+12108-12300</td>
<td>KC469618.1:+12108-12300</td>
<td>0</td>
<td>2992.2</td>
<td>1625</td>
<td>176.8</td>
<td>1567.6</td>
<td>6361.6</td>
</tr>
<tr>
<td></td>
<td>9hpi KO2</td>
<td>KC469618.1:+12108-12300</td>
<td>KC469618.1:+12108-12300</td>
<td>0</td>
<td>5355.8</td>
<td>743</td>
<td>1056.4</td>
<td>204</td>
<td>7359.2</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1:+12108-12300</td>
<td>KC469618.1:+12078-12300</td>
<td>5.6E-17</td>
<td>9258.7</td>
<td>1826.9</td>
<td>9885.3</td>
<td>8218.8</td>
<td>29189.7</td>
</tr>
<tr>
<td>Gp2a</td>
<td>9hpi</td>
<td>KC469618.1:+12109-12844</td>
<td>KC469618.1:+12109-12844</td>
<td>0</td>
<td>21506.6</td>
<td>13054.3</td>
<td>1931.1</td>
<td>10387.5</td>
<td>46879.4</td>
</tr>
<tr>
<td></td>
<td>9hpi KO2</td>
<td>KC469618.1:+12109-12844</td>
<td>KC469618.1:+12109-12844</td>
<td>0</td>
<td>32262.9</td>
<td>5340</td>
<td>6946.9</td>
<td>1316.6</td>
<td>45866.4</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1:+12109-12844</td>
<td>KC469618.1:+12073-12844</td>
<td>0</td>
<td>36138</td>
<td>7638.3</td>
<td>27008.4</td>
<td>22744.5</td>
<td>93529.3</td>
</tr>
<tr>
<td>Gp2a-iORF</td>
<td>9hpi</td>
<td>KC469618.1:+12320-12392</td>
<td>KC469618.1:+12281-12392</td>
<td>0.04157</td>
<td>611.7</td>
<td>215.6</td>
<td>31</td>
<td>248.4</td>
<td>1106.8</td>
</tr>
<tr>
<td></td>
<td>9hpi KO2</td>
<td>KC469618.1:+12296-12392</td>
<td>KC469618.1:+12296-12392</td>
<td>0.00575</td>
<td>1127.2</td>
<td>90.1</td>
<td>180.5</td>
<td>29.9</td>
<td>1427.7</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1:+12320-12392</td>
<td>KC469618.1:+12296-12392</td>
<td>0.03137</td>
<td>1458.8</td>
<td>520.9</td>
<td>1451.1</td>
<td>1194.9</td>
<td>4625.6</td>
</tr>
<tr>
<td>Gp3</td>
<td>9hpi</td>
<td>KC469618.1:+12723-13461</td>
<td>KC469618.1:+12723-13461</td>
<td>0</td>
<td>25431.4</td>
<td>13695.8</td>
<td>1806.2</td>
<td>9936</td>
<td>50869.4</td>
</tr>
<tr>
<td></td>
<td>9hpi KO2</td>
<td>KC469618.1:+12723-13461</td>
<td>KC469618.1:+12723-13461</td>
<td>0</td>
<td>41122</td>
<td>6233.9</td>
<td>8337.1</td>
<td>1377.8</td>
<td>57070.7</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1:+12696-13461</td>
<td>KC469618.1:+12696-13461</td>
<td>0</td>
<td>46101.6</td>
<td>12923.3</td>
<td>35407</td>
<td>29324</td>
<td>123755.9</td>
</tr>
<tr>
<td>Gp3-iORF</td>
<td>9hpi</td>
<td>KC469618.1:+12943-12967</td>
<td>KC469618.1:+12943-12967</td>
<td>0.04444</td>
<td>479.5</td>
<td>319.2</td>
<td>30.4</td>
<td>225.3</td>
<td>1054.4</td>
</tr>
<tr>
<td>Gp4</td>
<td>9hpi</td>
<td>KC469618.1:+13265-13778</td>
<td>KC469618.1:+13265-13778</td>
<td>0</td>
<td>14298.6</td>
<td>7495</td>
<td>1076.4</td>
<td>6778.7</td>
<td>29648.7</td>
</tr>
<tr>
<td></td>
<td>9hpi KO2</td>
<td>KC469618.1:+13265-13778</td>
<td>KC469618.1:+13265-13778</td>
<td>8.2E-16</td>
<td>28247.2</td>
<td>3512.5</td>
<td>6257.2</td>
<td>891.4</td>
<td>38908.3</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>KC469618.1:+13265-13778</td>
<td>KC469618.1:+13265-13778</td>
<td>1.5E-15</td>
<td>32936.1</td>
<td>12906.8</td>
<td>29165.7</td>
<td>25836.5</td>
<td>100845</td>
</tr>
<tr>
<td>Gp4-iORF</td>
<td>9hpi</td>
<td>KC469618.1:+13641-13713</td>
<td>KC469618.1:+13569-13713</td>
<td>0.02904</td>
<td>304.9</td>
<td>168.8</td>
<td>21.6</td>
<td>106.3</td>
<td>601.6</td>
</tr>
</tbody>
</table>

**Note:** The table provides a summary of expression levels across different time points and conditions for various proteins.
<table>
<thead>
<tr>
<th></th>
<th>9hpi_WT</th>
<th>9hpi_KO2</th>
<th>12hpi</th>
<th>5a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KC469618.1:+13778-13934</td>
<td>KC469618.1:+13778-13934</td>
<td>KC469618.1:+13778-13934</td>
<td>KC469618.1:+13778-13934</td>
</tr>
<tr>
<td></td>
<td>4.8E-14</td>
<td>6035.7</td>
<td>2699.8</td>
<td>278.4</td>
</tr>
<tr>
<td></td>
<td>2134.1</td>
<td>11148.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>12hpi</td>
<td>13778-13934</td>
<td>4.8E-14</td>
<td>6035.7</td>
</tr>
<tr>
<td></td>
<td>2699.8</td>
<td>11148.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>278.4</td>
<td>-</td>
<td>12hpi</td>
<td>2134.1</td>
</tr>
<tr>
<td></td>
<td>41833</td>
<td>13778-13934</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>9hpi_WT</th>
<th>9hpi_KO2</th>
<th>12hpi</th>
<th>Gp5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6035.7</td>
<td>2699.8</td>
<td>19738.8</td>
<td>4086.4</td>
</tr>
<tr>
<td></td>
<td>11148.1</td>
<td>-</td>
<td>153338.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2134.1</td>
<td>-</td>
<td>12hpi</td>
<td>11148.1</td>
</tr>
<tr>
<td></td>
<td>10068.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>41833</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>9hpi_WT</th>
<th>9hpi_KO2</th>
<th>12hpi</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15067.5</td>
<td>28385.3</td>
<td>14579.9</td>
<td>1579.9</td>
</tr>
<tr>
<td></td>
<td>25955.5</td>
<td>124752.2</td>
<td>3553.2</td>
<td>153538.8</td>
</tr>
<tr>
<td></td>
<td>124752.2</td>
<td>-</td>
<td>12hpi</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>153538.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>366840.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>9hpi_KO2</th>
<th>12hpi</th>
<th>M-ORF</th>
<th>6hpi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19738.8</td>
<td>102.8</td>
<td>9.0308</td>
<td>3294</td>
</tr>
<tr>
<td></td>
<td>4086.4</td>
<td>11.7</td>
<td>636.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>153338.9</td>
<td>636.8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>9hpi_WT</th>
<th>9hpi_KO2</th>
<th>12hpi</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14579.9</td>
<td>14579.9</td>
<td>14579.9</td>
<td>14579.9</td>
</tr>
<tr>
<td></td>
<td>153538.8</td>
<td>153538.8</td>
<td>153538.8</td>
<td>153538.8</td>
</tr>
<tr>
<td></td>
<td>366840.2</td>
<td>366840.2</td>
<td>366840.2</td>
<td>366840.2</td>
</tr>
</tbody>
</table>
## 5.2 EU PRRSV

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Candidate Location</th>
<th>p value</th>
<th>rep 1</th>
<th>rep 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>uORF</td>
<td>PRRSV_EU_vaccine_strain_Lior+:132-219</td>
<td>PRRSV_EU_vaccine_strain_Lior+:132-219</td>
<td>0.01088</td>
<td>932.1</td>
<td>296.6</td>
<td>1228.7</td>
</tr>
<tr>
<td>ORF1a</td>
<td>PRRSV_EU_vaccine_strain_Lior+:1226-7190</td>
<td>PRRSV_EU_vaccine_strain_Lior+:86-7190</td>
<td>0.00078</td>
<td>22.9</td>
<td>7.3</td>
<td>30.2</td>
</tr>
<tr>
<td>nsp2-iORF</td>
<td>PRRSV_EU_vaccine_strain_Lior+:2161-2230</td>
<td>PRRSV_EU_vaccine_strain_Lior+:2161-2230</td>
<td>6.4E-05</td>
<td>36.9</td>
<td>37</td>
<td>73.9</td>
</tr>
<tr>
<td>nsp2TF transframe</td>
<td>PRRSV_EU_vaccine_strain_Lior+:3414-3855</td>
<td>PRRSV_EU_vaccine_strain_Lior+:3345-3855</td>
<td>2.2E-07</td>
<td>778.2</td>
<td>330.1</td>
<td>1108.3</td>
</tr>
<tr>
<td>nsp5-iORF</td>
<td>PRRSV_EU_vaccine_strain_Lior+:6001-6040</td>
<td>PRRSV_EU_vaccine_strain_Lior+:6001-6040</td>
<td>0.01923</td>
<td>60.3</td>
<td>19.3</td>
<td>79.6</td>
</tr>
<tr>
<td>nsp5-trunc</td>
<td>PRRSV_EU_vaccine_strain_Lior+:8044-11563</td>
<td>PRRSV_EU_vaccine_strain_Lior+:8044-11563</td>
<td>0 17338.3</td>
<td>6569.2</td>
<td>23907.5</td>
<td></td>
</tr>
<tr>
<td>nsp10-iORF1</td>
<td>PRRSV_EU_vaccine_strain_Lior+:3414-3855</td>
<td>PRRSV_EU_vaccine_strain_Lior+:3345-3855</td>
<td>0.01516</td>
<td>22.8</td>
<td>11.8</td>
<td>34.6</td>
</tr>
<tr>
<td>nsp11-iORF1</td>
<td>PRRSV_EU_vaccine_strain_Lior+:7209-7239</td>
<td>PRRSV_EU_vaccine_strain_Lior+:7209-7239</td>
<td>0.01516</td>
<td>22.8</td>
<td>11.8</td>
<td>34.6</td>
</tr>
<tr>
<td>nsp10-iORF2</td>
<td>PRRSV_EU_vaccine_strain_Lior+:2161-2230</td>
<td>PRRSV_EU_vaccine_strain_Lior+:2161-2230</td>
<td>6.4E-05</td>
<td>36.9</td>
<td>37</td>
<td>73.9</td>
</tr>
<tr>
<td>nsp11-iORF2</td>
<td>PRRSV_EU_vaccine_strain_Lior+:3414-3855</td>
<td>PRRSV_EU_vaccine_strain_Lior+:3345-3855</td>
<td>2.2E-07</td>
<td>778.2</td>
<td>330.1</td>
<td>1108.3</td>
</tr>
<tr>
<td>nsp10-iORF3</td>
<td>PRRSV_EU_vaccine_strain_Lior+:10259-10310</td>
<td>PRRSV_EU_vaccine_strain_Lior+:10259-10310</td>
<td>0.0003</td>
<td>61.6</td>
<td>44</td>
<td>105.7</td>
</tr>
<tr>
<td>nsp11-iORF1</td>
<td>PRRSV_EU_vaccine_strain_Lior+:10595-11013</td>
<td>PRRSV_EU_vaccine_strain_Lior+:10595-11013</td>
<td>0.02745</td>
<td>188.1</td>
<td>48.5</td>
<td>236.6</td>
</tr>
<tr>
<td>Gp2</td>
<td>PRRSV_EU_vaccine_strain_Lior+:11570-12323</td>
<td>PRRSV_EU_vaccine_strain_Lior+:11570-12323</td>
<td>0 9306.3</td>
<td>3770.8</td>
<td>13077.1</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>PRRSV_EU_vaccine_strain_Lior+:11608-11791</td>
<td>PRRSV_EU_vaccine_strain_Lior+:11608-11791</td>
<td>1.6E-15</td>
<td>4330.6</td>
<td>1369.3</td>
<td>5699.8</td>
</tr>
<tr>
<td>Gp3</td>
<td>PRRSV_EU_vaccine_strain_Lior+:12607-12979</td>
<td>PRRSV_EU_vaccine_strain_Lior+:12292-12979</td>
<td>3.9E-12</td>
<td>6924.7</td>
<td>2490.9</td>
<td>9415.6</td>
</tr>
<tr>
<td>Gp4</td>
<td>PRRSV_EU_vaccine_strain_Lior+:12723-13275</td>
<td>PRRSV_EU_vaccine_strain_Lior+:12723-13275</td>
<td>0 14456</td>
<td>5822.6</td>
<td>20278.6</td>
<td></td>
</tr>
<tr>
<td>Gp5</td>
<td>PRRSV_EU_vaccine_strain_Lior+:13301-13877</td>
<td>PRRSV_EU_vaccine_strain_Lior+:13301-13877</td>
<td>0 28164.5</td>
<td>11654.1</td>
<td>39818.5</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>PRRSV_EU_vaccine_strain_Lior+:13966-14386</td>
<td>PRRSV_EU_vaccine_strain_Lior+:13966-14386</td>
<td>0 25079.6</td>
<td>10887</td>
<td>35966.5</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>PRRSV_EU_vaccine_strain_Lior+:14405-14762</td>
<td>PRRSV_EU_vaccine_strain_Lior+:14405-14762</td>
<td>0 36470.1</td>
<td>16749.1</td>
<td>53219.2</td>
<td></td>
</tr>
</tbody>
</table>
6 Investigating molecular mechanisms of 2A-stimulated ribosomal pausing and frameshifting in Theilovirus

The text and Figures of the manuscript on bioRxiv in May 2021 are appended, beginning on the next page (with size scaled down to facilitate binding, and page numbers given with respect to the manuscript).
Investigating molecular mechanisms of 2A-stimulated ribosomal pausing and frameshifting in Theilovirus

Chris H. Hill*, Georgia M. Cook†1, Sawsan Naphine†1, Anuja Kibe2, Katherine Brown1, Neva Caliskan2,3, Andrew E. Firth*, Stephen C. Graham*1 and Ian Brierley*1

* corresponding authors
† authors contributed equally to this work

1 Division of Virology, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, UK. CB2 1QP
2 Helmholtz Institute for RNA-based Infection Research (HiRI), Josef-Schneider-Straße 2/D15, 97080 Würzburg, Germany
3 Medical Faculty, Julius-Maximilians University Würzburg, 97074, Würzburg, Germany

Abstract

The 2A protein of Theiler’s murine encephalomyelitis virus (TMEV) acts as a switch to stimulate programmed −1 ribosomal frameshifting (PRF) during infection. Here we present the X-ray crystal structure of TMEV 2A and define how it recognises the stimulatory RNA element. We demonstrate a critical role for bases upstream of the originally predicted stem-loop, providing evidence for a pseudoknot-like conformation and suggesting that the recognition of this pseudoknot by beta-shell proteins is a conserved feature in cardioviruses. Through examination of PRF in TMEV-infected cells by ribosome profiling, we identify a series of ribosomal pauses around the site of PRF induced by the 2A-pseudoknot complex. Careful normalisation of ribosomal profiling data with a 2A knockout virus facilitated the identification, through disome analysis, of ribosome stacking at the TMEV frameshifting signal. These experiments provide unparalleled detail of the molecular mechanisms underpinning Theilovirus protein-stimulated frameshifting.

Introduction

Cardioviruses are a diverse group of picornaviruses that cause encephalitis, myocarditis and enteric disease in a variety of mammalian hosts including rodents, swine and humans (1). The Cardiovirus B or Theilovirus species comprises several isolates including Sihote-Alin virus (2), rat theilovirus and Theiler’s murine encephalomyelitis virus (TMEV), all of which are genetically distinct from the Cardiovirus A species encephalomyocarditis virus (EMCV) (3). Within the Cardiovirus B species, TMEV has been extensively characterised and serves as a mouse model for virus-induced demyelination and multiple sclerosis (4). Like all
picornaviruses, TMEV replication is cytoplasmic and begins with the translation of its single-stranded ~8 kb positive-sense RNA genome. The resultant polyprotein (L-1ABCD-2ABC-3ABCD) is subsequently processed, mainly by the virally encoded 3C protease (5,6).

Several "non-canonical" translation events occur during the production of the TMEV polyprotein. First, initiation is directed by a type II internal ribosome entry site (IRES) in the 5′ untranslated region (UTR) (7,8). Secondly, a co-translational StopGo or ribosome “skipping” event occurs at the junction between the 2A and 2B gene products (9,10). In this process, the peptidyl-transferase reaction fails between the glycine and proline in a conserved D(V/I)ExNPG|P motif, releasing the upstream L-1ABCD-2A product as the ribosome continues translating the downstream 2BC-3ABCD region. Note that in TMEV, however, the presence of a 3C protease cleavage site near the start of the 2B protein appears to make the StopGo reaction functionally redundant (11,12). Thirdly, programmed −1 ribosomal frameshifting (PRF) diverts a proportion of ribosomes out of the polyprotein reading frame and into a short overlapping ORF, termed 2B*, near the start of the 2B protein. In TMEV this ORF is only eight codons in length and the resulting transframe product 2B* has just 14 amino acids (~ 1.4 kDa), with no established functional role (11). Thus it has been hypothesised that in TMEV, the main function of PRF may simply be to downregulate translation of the enzymatic proteins encoded downstream of the frameshift site, particularly in the late stages of infection (11).

PRF is common amongst RNA viruses, where it is used as a translational control strategy to express gene products in optimal ratios for efficient virus replication. Additionally, the utilisation of overlapping ORFs permits more information to be encoded by a small genome. Our mechanistic understanding of PRF has been informed by the study of examples in hundreds of viruses (reviewed in ref. 13-15). Generally, PRF involves two elements within the viral messenger RNA (mRNA). A heptameric shift site or “slippery sequence” of the form X_XXY_YYZ (where XXX represents any three identical nucleotides or certain other triplets such as GGU, YYY represents AAA or UUU, and Z is any nucleotide except G (16) is located 5 – 9 nucleotides upstream of a structured RNA “stimulatory element” (usually a stem-loop or pseudoknot) that impedes the progress of the elongating ribosome, such that the ribosome pauses with the shift site in the decoding centre (17-19). This can facilitate a change of reading-frame if the codon-anticodon interactions of the P-site and A-site tRNAs slip and recouple in the −1 frame (XXY → XXX and YYY → YYY, respectively) during resolution of the stimulatory element. For any given system, such PRF generates a fixed ratio of products set by parameters that include the energetics of codon-anticodon pairing at the shift site, the conformational flexibility of the stimulatory element and the resistance of this element to unwinding by the ribosomal helicase (20,21). In cardioviruses, however, PRF exhibits some
intriguing mechanistic exceptions. First, the conserved G_GUU_UUU shift site is located 13 – 14 nucleotides upstream of the stimulatory stem-loop, seemingly too far away to position the shift site in the decoding centre during a pause. Secondly, the viral 2A protein is required as a trans-activator of frameshifting in cells and in vitro (11,22-24) which permits temporally controlled, variable-efficiency frameshifting related to the amount of 2A protein in the cell during infection. To date, cardioviruses present one of only two known examples of protein-stimulated frameshifting, along with porcine reproductive and respiratory syndrome virus (PRRSV) and other arteriviruses (family Arteriviridae), where a complex of viral nsp1β and host poly(C) binding protein stimulates PRF (25-27).

Our previous investigations of protein-stimulated frameshifting in EMCV and TMEV have revealed that the viral 2A protein acts in complex with a stem-loop downstream of the slippery sequence, and that this interaction can be disabled by mutating either a conserved cytosine triplet in the loop or a pair of conserved arginine residues in the protein (11,22,23). More recently, we solved the structure of EMCV 2A, revealing a novel protein fold that permits binding to both the stem-loop and to ribosomal RNA with high affinity (28). However, 2A protein sequences are highly divergent within cardioviruses, and the TMEV protein shares only ~27% pairwise amino acid sequence identity with its EMCV orthologue. Additionally, the stem-loop that comprises the stimulatory element in TMEV is significantly more compact than the equivalent structure in EMCV. Here we present the X-ray crystal structure of TMEV 2A and investigate the interaction with its cognate RNA using a variety of biochemical and biophysical techniques. We define a minimal TMEV stimulatory element necessary for 2A binding and show that the protein forms a 1:1 RNA-protein complex with nanomolar affinity. We provide evidence that the alternative pseudoknot-like conformation recently described for the EMCV stimulatory element (28) is also likely to be present in TMEV and other cardioviruses. Finally, we use metabolic labelling and ribosome profiling to study 2A-stimulated frameshifting and translation of the TMEV genome at sub-codon resolution in infected cells. Together, this body of work defines the molecular basis of Theilovirus protein-stimulated frameshifting, one of the most efficient frameshifting events known in nature.

Results and Discussion

Structure of TMEV 2A reveals a beta-shell protein with a divergent RNA-binding surface

TMEV 2A protein was purified following recombinant expression as a GST fusion in E. coli (Figure 1A). After cleavage of the GST tag and removal of contaminating nucleic acids by heparin affinity chromatography, high-salt conditions were required to prevent aggregation
and maintain solubility. Under optimised conditions, size-exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) revealed the protein to be predominantly monomeric (Figure 1B, peak 2) with an observed $M_w$ of 16,683 Da compared to a theoretical mass of 15,941 Da, as calculated from the primary sequence. A small proportion of trimers was also present (Figure 1B, peak 1: observed mass of 46,867 Da vs. theoretical mass of 47,823 Da). We crystallised the protein but were unable to solve the structure by molecular replacement using the crystal structure of the closest homologue, EMCV 2A (28), as a search model. Instead, we obtained experimental phases via bromine derivatisation and determined the structure by single-wavelength anomalous dispersion analysis. The asymmetric unit of the cubic cell contained a single copy of 2A, which was refined to 1.9 Å resolution (Table S1, Figure 1C). Interpretable electron density was visible for 2A residues 1-126, with a short G-$\Delta$P-L-G-S$_N$ N-terminal extension that resulted from 3C protease cleavage of the GST tag. In the crystalline lattice these residues mediate contacts between symmetry-related molecules, consistent with recombinant 2A forming trimers at high concentrations (Figure S1). These residues project away from the globular TMEV domain and lack regular secondary structure, suggesting that they will be flexible at lower concentrations of 2A in solution where monomers predominate (Figure 1B). Since tag-derived residues mediate the inter-subunit contacts, this 2A trimerisation is unlikely to be physiologically relevant.

The structure of TMEV 2A reveals a globular $\beta_3\alpha\beta_3\alpha$ fold, with a similar “beta-shell” architecture to EMCV 2A ($\alpha$-backbone RMSD of 2.65 Å over 130 residues) (28). An extensive curved antiparallel six-stranded beta sheet packs against two alpha helices on the concave surface of the sheet, whilst the loops between adjacent beta strands project from the opposite convex face. The protein is highly basic (pI ~ 9.4) and most of the contributing lysine, arginine and histidine residues are solvent-exposed. At physiological pH the protein will thus have a positive electrostatic surface potential across the convex face of the beta sheet, surrounding loops and the N-terminal end of helix $\alpha_1$, suggesting a putative RNA-binding surface (Figure 1C). This is supported by a previous biochemical analysis of TMEV 2A function, in which we made several point mutations of conserved basic residues and assessed their ability to stimulate PRF (23) (Figure 1D). The M3 mutant (R85A / R87A, termed ‘2A-mut’ in Naphthine et al., 2019) was found to completely inhibit frameshifting, and the M1 mutant (K24A / R28A) was found to reduce it by approximately four-fold. Both M3 and M1 are in surface loops on either side of this large, positively charged beta sheet, consistent with a role for this face of the protein in forming electrostatic interactions with the ribose phosphate backbone of the PRF stimulatory element. In contrast, M2 (R45A) had no effect – indeed most of this residue is buried (only approximately 11.6% of the residue’s surface area being solvent-accessible) and it likely plays a structural role in stabilising packing of helix $\alpha_1$ against the underside of the
central sheet. Perhaps surprisingly, M4 (K90A / K91A) also had no effect, despite these residues being located in the same positively charged loop as the essential R85 and R87. This demonstrates that precise spatial positioning of charge within this loop is key to the specificity of RNA-binding.

Despite low sequence identity, the overall fold of TMEV 2A is similar to EMCV 2A (Figure 1E). The most functionally important part of EMCV 2A is the “arginine loop” between strands β5 and β6 – necessary for PRF (22), RNA binding (28) and nuclear localisation (29). The two arginines in this loop (R95 / R97 in EMCV 2A; R85 / R87 in TMEV 2A) are amongst the only surface-exposed residues that are completely conserved across both species of cardiovirus. (Figure S2A and B). Structurally, this loop adopts an almost identical conformation, consistent with mutagenesis data suggesting that they are functionally equivalent (22,23). However, there are also some key differences between the two 2A orthologues (Figure 1E). The loop region between the end of α1 and the start of β4 is much longer in EMCV 2A. This loop, and the N-terminal end of α1, are two of three regions that contribute to the RNA-binding surface in EMCV 2A (28). These two regions are not conserved in TMEV 2A (Figure S2C): the backbone geometry is different, and, with the exception of K63 and R45, there are no chemically equivalent side chains in the vicinity that could form similar interactions. To investigate this further, we prepared variants of TMEV 2A containing point mutations (R85A / R87A, K63A, K83A and K53A / H56A; Figure S3A) and tested their ability to bind stimulatory element RNA (Figure S3B). Only the R85A / R87A protein showed a significant defect compared to wild-type protein, confirming the unique importance of this conserved arginine loop. Beyond this, the RNA binding surface in TMEV 2A likely differs from EMCV 2A, perhaps involving additional residues on the surface of the beta sheet that are only conserved amongst Theilovirus isolates (e.g. R7, D9, K24; Figure S2A-C). This is supported by a reduction in PRF seen with the M1 (K24A) mutant (23).

The conservation at the C-terminus of the protein (Figure S2A and B) is concentrated in the D(V/I)ExNPG motif required for StopGo peptide release between 2A and 2B gene products in other cardioviruses (9,10). As expected, this motif is unstructured in both proteins, and if we consider only the ordered amino acids, we observe some of the largest structural differences (Figure 1E). In EMCV 2A, the shorter α2 helix leads to a more pronounced curvature of the central beta sheet, and the C-terminus forms a short β7 strand that packs against β3. Conversely, in TMEV 2A, the α2 helix continues with a pronounced 90° kink until the end of the protein. This has consequences for the structure of the putative YxxxxLΦ motif at the C-terminus that has previously been reported to sequester eIF4E in a manner analogous to 4E-BP, thereby disabling cap-dependent translation of host mRNA (29). In TMEV 2A, the first
tyrosine residue of this motif (Y119) is located on the buried side of the α2 helix (with only approximately 4.0% of the residue’s surface area being solvent accessible) and is therefore not available to interact with eIF4E. A second tyrosine (Y120) is more exposed (approximately ~26.7% of the residue’s surface area being solvent accessible), but given the local secondary structure, it is unlikely that L124 and I125 would be able to interact with eIF4E in the same way as 4E-BP without a significant conformational change (Figure S2D). In EMCV 2A, the putative YxxxxLΦ motif is not helical, instead forming a more extended backbone (28). Superposition of this region reveals that this motif is not structurally conserved between 2A orthologues (Figure S2E). Therefore, the interaction with eIF4E must either involve other parts of the protein or must be accompanied by significant structural rearrangement of the α2 helix of TMEV.

**A minimal 37 nt pseudoknot in the viral RNA is required for 2A binding**

The PRF stimulatory element in TMEV consists of a stem-loop (seven base-pair stem and 10 nucleotide loop) located 14 nucleotides downstream from the shift site (Figure 2A). This is more compact than the equivalent element in EMCV, which has a 21 nt loop that may contain an additional stem (22,28). However, in both viruses, three conserved cytosines in the loop are essential for 2A binding (22,23). To determine the minimal RNA element necessary for interaction with TMEV 2A, we prepared a series of synthetic RNA constructs (Figure 2B) and assessed 2A binding by electrophoretic mobility shift assays (EMSA; Figure 2C) and microscale thermophoresis (MST; Figure 2D).

Binding of 2A was generally high affinity, with dissociation constants in the sub-micromolar range. (Figure 2D). Truncation of either the shift-site (TMEV 1, $K_D = 484 \pm 90 \text{nM}$), the 3’ extension (TMEV 3, $K_D = 495 \pm 110 \text{nM}$) or both (TMEV 6, $K_D = 280 \pm 40 \text{nM}$) had little effect on 2A binding. Shortening the loop by three nucleotides (TMEV 5, $K_D = 615 \pm 89 \text{nM}$) slightly reduced the affinity, however removal of the 5’ extension (TMEV 2, TMEV 4) completely abolished 2A binding, even though the stem-loop in TMEV 4 is predicted to be intact, and the three essential cytosines are present in the loop. To validate that these small RNAs are adopting conformations relevant to PRF, we performed competition experiments. A dual-luciferase-based reporter mRNA containing the TMEV shift site and stimulatory element was prepared and designed such that 0 frame and −1 frame products would be easily distinguishable by SDS-PAGE following in vitro translation in rabbit reticulocyte lysates (23). In the presence of 2A, PRF occurred efficiently (~57%; Figure 2E), but inclusion of a molar excess of the small RNA would be predicted to reduce PRF efficiency if the competitor RNA were to sequester 2A from the reporter mRNA. In line with the EMSA assays, TMEV 3 and
TMEV 6 both efficiently competed with the reporter mRNA, reducing PRF to 38% and 21%, respectively. RNAs lacking the 5’ extension (TMEV 2, TMEV 4) were unable to compete (Figure 2E).

We have recently shown that the 5’ extension is also important for 2A binding in EMCV, where it likely forms the second stem of an RNA pseudoknot via interaction with the CCC loop motif (28). However, given that the equivalent loop sequence is 11 nt shorter in TMEV, it was unclear whether this alternate conformation would be topologically possible for the more compact RNA element. Nevertheless, an alignment of cardiovirus RNA sequences that direct PRF shows that, in addition to the cytosine triplet, there are several nucleotides in the 5’ extension that are completely conserved in all isolates (Figure 3A, Figure S4). To investigate this in more detail, we truncated the 5’ extension one nucleotide at a time (Figure 3B) and assessed 2A binding by EMSA (Figure 3C). Removal of the first six nucleotides of the 5’ extension (CAGCCA; TMEV 7–9) had no dramatic effect on 2A binding, however removal of nucleotides from the conserved CAAGG motif progressively reduced binding until it was undetectable (TMEV 12).

To explore this further, we made point mutations and assessed their effects in a frameshift reporter assay (Figure 3D). As expected, loop mutations C36G or C38G abolished PRF. However, frameshifting in the C38G background was restored by introducing a G17C mutation in the 5’ extension, demonstrating the necessity for a base pair between positions 17 and 38. The importance of this alternate conformation for 2A recognition was verified by EMSA analysis (Figure 3E). In RNA TMEV 9, which comprises the minimal functional element as defined by the deletion analysis, individual mutation of either C38 or G17 completely abrogated 2A binding. An unrestrained RNA folding simulation (30) identified a pseudoknot-like conformation consistent with the biochemical data, including a base pair between G17 and C38 (Figure 3F). This is consistent with our evidence for a pseudoknot-like conformation in the EMCV sequence that is selectively recognised by 2A (28). Whilst the double C38G+G17C mutant could rescue PRF in our in vitro reporter assay, it did not restore 2A binding by EMSA, either in the background of TMEV 9 (Figure 3E) or the longer TMEV 1 RNA (data not shown). It is possible that the double mutant forms a topologically equivalent conformer that is either less stable and/or binds 2A with lower affinity. Thus, whilst this RNA is still able to stimulate PRF, the association does not persist for the necessary timescales to permit observation in a dissociative technique such as EMSA.

To assess the binding affinity and stoichiometry in solution with unlabelled RNA, we carried out isothermal titration calorimetry (ITC) experiments with TMEV 6 and 2A protein. To prevent protein aggregation in the ITC cell, it was necessary to perform the titration at 400 mM NaCl. Increasing the salt concentration had only a slight effect on 2A binding as judged by EMSA.
(Figure 4A) and under these conditions we observed a $K_D$ of $67 \pm 7$ nM and a ~ 1:1 molar ratio of protein to RNA (Figures 4B and C). The large contribution of $\Delta H$ ($-10.7 \pm 0.13$ kcal/mol) term to the overall free energy of binding ($-9.8$ kcal/mol) is consistent with an electrostatic interaction mechanism. To confirm the stoichiometry in solution, we performed a SEC-MALS analysis of the mixture retrieved from the ITC cell (Figure 4D). As expected, the two peaks correspond to an approximate 1:1 RNA-protein complex (early; $M_w$ 33,570 Da observed vs. 27,105 Da expected) and excess free RNA (late; $M_w$ 10,980 Da observed vs 11,164 Da expected).

**Snapshots of translation in TMEV-infected cells by ribosome profiling**

We extended our examination of 2A-stimulated PRF by analysing TMEV-infected cells using ribosome profiling, a deep sequencing-based technique that gives a global snapshot of the positions of translating ribosomes at sub-codon resolution (31,32). In addition to WT virus, two mutant viruses were employed. Virus SS has two synonymous mutations in the slippery sequence (G_GUU_UUU to A_GUG_UUU) that prevent frameshifting (23,24). Virus M3 has the WT slippery sequence, but contains the M3 mutations in the 2A protein described above (Figure 1D), rendering it unable to bind to the PRF-stimulatory RNA stem-loop (23).

BSR cells, a single cell clone of the *Mesocricetus auratus* BHK-21 cell line, were selected for the ribosome profiling due to their relative genetic homogeneity. Thus, we first verified in BSR cells several features of TMEV infection observed in BHK-21 cells (11,23). The small-plaque phenotype previously observed for the SS and M3 mutant viruses (22-24) was confirmed (Figure S5A), and expression of 2A was detectable from eight hours post-infection (hpi) onwards. Levels increased up to 12 hpi, at which point cytopathic effects become fairly extensive (Figure S5B).

Metabolic labelling experiments were carried out to verify the occurrence of highly efficient frameshifting in TMEV-infected BSR cells. Viral proteins downstream of the frameshift site are only translated by ribosomes which have not undergone frameshifting, as those that frameshift encounter a stop codon in the −1 frame eight codons downstream of the shift site. Frameshift efficiency was estimated from the ratio of downstream to upstream products (Figure 5A, Figure S5C), normalised by the frameshift-defective M3 mutant as a control (detailed in Methods, and previously described in refs. 11,33). The mean WT frameshift efficiency was found to increase from 63% at 10 hpi to 78% at 12 hpi (Figure 5B, blue bars) (two-tailed Welch’s $t$-test: $t = -2.92$, $df = 2.80$, $p = 0.067$) similar to the 74 – 82% range previously calculated by metabolic labelling of infected BHK-21 cells (11,23). The frameshift efficiency of
the SS mutant virus was low at both timepoints (5% and 1% at 10 and 12 hpi respectively), as previously seen for TMEV (23) and EMCV (24).

The 10 hpi timepoint was selected for the ribosome profiling. BSR cells were infected with TMEV WT, SS, or M3, or mock-infected, and harvested by snap-freezing in liquid nitrogen. Cells were lysed and the presence of 2A verified by western blotting (Figure S6A). Lysates were treated with RNase I and ribosome-protected mRNA fragments (RPFs) were harvested by pelleting the ribosomes through sucrose with subsequent phenol extraction (Figure 5C). RPFs were ligated to adapters, cloned, and deep sequenced. Reads were mapped to the host (M. auratus) and viral (based on NC_001366) genomes (Table S2) to determine precisely the locations of translating ribosomes. Quality control analysis, carried out as previously described (34), indicated that the datasets were of high quality (Figure S6). RPFs mapping to coding sequences exhibited the characteristic length distribution with peaks at 20 – 22 nt and 28 – 30 nt (Figure S6B), corresponding to the two distinct lengths of mRNA protected from nuclease digestion by the translating ribosome (31,35-38). In mammalian cell lysates treated with sufficient nuclease so that all unprotected regions of mRNA are fully digested (“trimmed”), there is a distance of ~ 12 nt between the 5′ end of the RPF and the first nucleotide in the P site of the ribosome (31). This can be used to infer the frame of translation. In our libraries, the majority of CDS-mapping RPF 5′ ends map to the first nucleotide position of codons (herein termed phase 0) (Figure S6C), indicating successful nuclease trimming. RPFs map to coding sequences with a triplet periodicity reflective of the length of a codon and, as expected, few RPFs map to the UTRs (Figure S6D).

2A-stimulated frameshifting occurs with 85% efficiency in infected cells

To analyse translation of the viral RNA, we plotted RPF distribution on the viral genome (Figure 5D-G, Figure S7). There is a clear dominance of the 0 phase throughout the genome (Figure 5E), including within the +1-frame L* ORF which overlaps L and VP0, indicating that the L* ORF is not highly translated. Ribosomes that undergo frameshifting at the 2B* shift site translate only eight codons in the −1 frame before encountering a termination codon. This results in a striking drop-off in ribosome density on the WT genome (Figure 5D). No such decrease in density is seen after the frameshift site on the SS virus genome, and there is actually a slight increase in read density in this region on the M3 genome. This could be due to differences between the two regions in mean translation rate or the extent of various biasing effects inherent in the ribosome profiling procedure, such as ligation, PCR or nuclease biases (39). In order to control for these effects and highlight translational features related to the presence of a functional 2A and/or shift site, the read densities on the WT and SS genomes
were divided by those in the corresponding position on the M3 genome (Figure 5F, Figure S7B). Frameshift efficiency was calculated from the ratio of M3-normalised RPF density in the regions downstream and upstream of the frameshift site, revealing a mean WT frameshift efficiency of 85%, which to our knowledge is the highest −1 PRF efficiency thus far recorded in any natural system (11,22,23) (Figure 5B, green bars). This is significantly greater than the 63% measured by metabolic labelling at the same timepoint (two-tailed Welch's t-test: $t = -10.2, df = 2.03, p = 0.0090$). The profiling assay (with M3 normalisation) is expected to be substantially more accurate than the metabolic labelling approach, which suffers from lower sensitivity (densitometry of a few protein bands versus high throughput sequencing) and lower temporal resolution (1 hour labelling vs snap-freezing translating ribosomes).

The occurrence of a highly efficient frameshift in WT TMEV was verified by the observation of a marked shift in the dominant RPF phase from 0 in the upstream region, to −1/+2 in the 2B* transframe region (Figure S8A). It should be noted that in frameshift efficiency calculations, neither the profiling nor the metabolic labelling experiments would reliably distinguish between ribosome drop-off due to termination at the 2B* stop codon post-frameshifting and ribosome drop-off at the StopGo site located just five codons upstream. However, StopGo is generally very efficient in cultured cells, with little drop-off (40,41), and this is evident here, as no obvious decrease in RPF density occurs after the StopGo motif in the M3 mutant (Figure 5D, Figure S7A). Surprisingly, the mean frameshift efficiency of the SS mutant was high, at 15%. Similar residual frameshift activity has been observed at other highly efficient frameshift signals with analogous mutations designed to knock out frameshifting (42,43). This may be reflective of the very strong frameshift-stimulatory activity of the 2A-RNA complex facilitating frameshifting even despite unfavourable codon-anticodon repairing in the −1 frame. However, prolonged pausing of ribosomes over the mutant slippery sequence (discussed below) could potentially influence mechanisms other than frameshifting that may contribute to this observed drop-off, such as ribosome rescue pathways (44).

**Profiling reveals multiple 2A-related ribosomal pausing events**

Ribosome pausing over the slippery sequence has long been considered a mechanistically important feature of PRF (16-18,21). However, while observed to a small extent on WT shift sites in vitro (23,42,45-47), it has been elusive in ribosome profiling data (34,48). However, if the slippery sequence is mutated to prevent frameshifting, a measurable pause is seen both in vitro (18,23,27) and in profiling experiments (22), perhaps reflecting a reduced ability of non-frameshifting ribosomes to resolve the topological problem posed by the downstream frameshift-stimulatory element. Ribosome profiling can allow the identification of ribosome
pauses with single-nucleotide precision, but at the level of individual nucleotides, the profiles can be strongly affected by nuclease, ligation and potentially, other biases introduced during library preparation. However, by comparing the WT and SS genome ribosome profiles with the M3 genome ribosome profile, many of these biases are normalised and allow changes in dwell time to be identified that are likely to arise as a result of 2A binding.

Apart from ribosome drop-off at the 2B* stop codon, the region of greatest difference in a comparison of the WT and M3 profiling datasets was surprisingly not at the frameshift site, but in the middle of the 2A ORF (Figure 5F, Figure S7B). This pause likely corresponds to ribosomes translating the RNA-binding arginine residues (R85 and R87), which are mutated to alanine in the M3 genome. The pause on the M3-normalised plot likely reflects increased decoding time for the arginine-encoding CGC codons in the WT, which are relatively poorly adapted to the cellular tRNA pool compared to the alanine-encoding GCC codons in the M3 mutant (49,50) (Figure S8B and C). This peak is present to a lesser extent on the SS genome, potentially due to the slightly slower replication kinetics of this mutant virus (11) meaning there are fewer copies of the viral genome to deplete the cellular supply of the relevant aminoacylated tRNA.

Looking specifically at the frameshift site, a single-nucleotide resolution plot of reads mapping to this region reveals a peak on the SS mutant genome corresponding to a ribosome paused with the GUG codon of the mutated slippery sequence (corresponding to WT GUU) in the P site (Figure 5G, Figure S7C). This putative pause is present to a lesser extent on the WT genome (marked 0 in purple), but not the M3 genome, indicating it is related to the presence of functional 2A. The UUU codon of the slippery sequence also appears to have a much larger phase 0 peak in SS than M3, however this may be enhanced by potential “run-on” effects in which a fraction of ribosome pauses may be able to resolve during cell harvesting and ribosomes translocate to the next codon (50). On both WT and SS genomes, a noticeable peak is present two nucleotides downstream of the main slippery sequence pause, corresponding to ribosomes which have frameshifted and then translocated one codon, and further peaks suggestive of −1-frame translation throughout the 2B* ORF, especially on the WT template. Closer inspection of the whole-genome (Figure 5D, Figure S7A) and M3-normalised plots (Figure 5F, Figure S7B) reveals that the pause at the frameshift site actually extends a little further upstream, ending just upstream of the 2A-2B junction formed by the StopGo motif. Further investigation of this region (Figure S8B) reveals prominent pauses in the SS mutant, and to a lesser extent the WT, with ribosomal P sites corresponding to the glutamic acid and methionine residues of the D(V/I)ExNPGP StopGo motif (pause sites highlighted in bold, where x is methionine in TMEV). These pauses are larger than the pause.
over the slippery sequence itself. Ribosomal pausing over StopGo motifs has been observed in vitro (41) but have been shown to occur with the ribosomal P site corresponding to the conserved glycine residue directly before the separation site (41,51), suggesting that the pauses we see have another origin.

**Disome profiling provides evidence for ribosome queuing at the frameshift and StopGo sites**

Noticing that the main pause over the StopGo motif was nine codons upstream of the pause over the slippery sequence, we wondered whether this might be consistent with the transient formation of disomes, in which the leading ribosome is paused over the slippery sequence. Disomes are routinely excluded during preparation of ribosome profiling libraries by the inclusion of a size-selection step (in this study, 19-34 nt) which selects for monosome-protected fragments (31). For mock, WT- and M3-infected lysates from replicate 3 we carried out two parallel size selection steps, in which the 19-34 nt “monosome” fraction and a 35-65 nt “broad spectrum” fraction were isolated from the same lysate. The length distribution of broad spectrum reads demonstrated local peaks at read lengths of around 51, 54 and 59-63 nt, consistent with expected lengths of RNA protected by disomes (36,52,53) (Figure 6A). Reads of lengths 51-52, 54-55 and 57-64 nt showed a bias in phase composition towards phase 0, indicating a portion of genuine ribosome footprints (Figure 6B). These read lengths were selected for analysis as potential “disome-protected fragments”, and their density plotted on the viral genome at the inferred P site position of the upstream, colliding ribosome (Figure 6C and D).

A very prominent peak is visible on the WT genome over the StopGo motif (Figure 6C), and closer inspection reveals that one of the highest peaks in this region is over the valine of the StopGo motif, ten codons upstream of the slippery sequence pause (Figure 6D). This is the expected distance between the P sites of ribosomes involved in a disome (52-54), and would be consistent with disome formation due to a ribosome translating the StopGo motif colliding with a ribosome paused over the slippery sequence. Further, this approximate ten-codon periodicity in distances between peaks extends upstream, consistent with potential formation of ribosome queues up to six ribosomes long (53,55) (Figure 6D, grey oblongs). An additional peak is evident over the arginine codon ten codons upstream of the StopGo release site, which would correspond to a disome in which the leading ribosome was paused over the conserved glycine codon of the StopGo motif (Figure 6D, brown oblongs). This is evidently a feature of the StopGo site itself and unrelated to the binding of 2A downstream, as it occurs in the M3 dataset as well as the WT. It should be noted that ribosome profiling generates an averaged
result of ribosome positions over multiple copies of the viral genome, and formation of the two putative disomes proposed here could not occur simultaneously on a single RNA.

The potential formation of ribosome queues behind the frameshift/StopGo site is supported by the monosome data, in which RPF density gradually increases throughout 2A, reaching a maximum over the StopGo motif. This is particularly apparent in the data from the SS mutant (Figure 5D and F), consistent with the idea that greater pausing over the mutant slippery sequence may be increasing disome formation, pushing the lengths of protected fragments into the disome fraction, and reducing their visibility in the monosome dataset. This is unusual, as no prominent disome peaks were detected near the frameshift site by broad spectrum ribosome profiling of murine coronavirus (34), nor at the −1 PRF site of L-A virus in yeast (56), although we note that −1 PRF is less efficient in both these systems when compared with TMEV. Indeed, the potential formation of disomes at the frameshift site in TMEV represents the first evidence of ribosomes forming queues at a −1 PRF signal in a eukaryotic system. Disome formation has recently been mechanistically implicated in +1 PRF in yeast and +1 and −1 PRF in bacteria (56-59). It may be that the impact of ribosomes colliding at the TMEV PRF site contributes to the complex energetic and conformational landscape required to overcome the translation blockade and break the triplet codon periodicity.

Conclusion

Our combination of structural, biophysical, biochemical and deep sequencing approaches affords a view of unparalleled detail into the molecular mechanisms underpinning Thiovirus protein-stimulated frameshifting. Despite highly divergent sequences within cardioviruses, the TMEV 2A protein adopts the second known occurrence of the beta-shell fold. Whilst the distribution of positively charged residues comprising the putative RNA-binding surface is different from that found in EMCV, it nevertheless recapitulates the same exquisite conformational selectivity for binding to its viral RNA target. Strikingly, we demonstrate that this is a pseudoknot that is likely to exist in equilibrium with the stem-loop previously suggested by structure probing experiments. Despite being 11 nt shorter, the stimulatory RNA element is able to adopt a conformation topologically equivalent to that previously seen in EMCV, involving interactions between the loop and the 5′ extension. In infected cells, stabilisation of this conformation by high-affinity 2A binding represents the ‘switch’ that controls frameshifting and thereby reprogramming of viral gene expression. Our ribosome profiling data reveals that, when invoked, this is up to 85% efficient, representing the highest known −1 PRF efficiency to date. Whilst frameshift-associated pausing is not normally detectable at WT shift sites in profiling data, we show that it is detectable here by analysing RPFs corresponding to disomes.
not only at the PRF site but also the adjacent StopGo motif. This is consistent with the relatively long pauses accompanying TMEV frameshifting in vitro, and suggests that ribosome collisions are more common than previously thought during translation of the TMEV genome. Taken together, these results suggest that there is a fine balance between necessary ribosome pausing associated with recoding events, and the detrimental effect on viral fitness that may result from these pauses lasting long enough to trigger ribosome quality control pathways, and the degradation of the viral RNA. In future, structural characterisation of the 2A-RNA complex will yield further insights into the molecular mechanisms underlying the potency of this elongation blockade.

**Materials and Methods**

*Protein expression and purification*

TMEV 2A cDNA was amplified by PCR from plasmid pGEX6P2-based constructs (23) (F 5′ AATTCATAGATCCGCTTCTCCTCTACCGC 3′; R 5′ AATTGGATCCCTATTAGCCTGGTCATTCTACATC 3′) and cloned into pOPT3G (60) to introduce a 3C protease-cleavable N-terminal GST tag. Recombinant protein was produced in *E. coli* BL21(DE3) pLysS cells. Cultures were grown in 2xTY broth supplemented with 100 μg/mL ampicillin (37 °C, 210 rpm). Expression was induced at A<sub>600</sub> of ~ 1 with 1.0 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and continued overnight (210 rpm, 21 °C, 16h). Bacteria were pelleted (4,000 × g, 4 °C, 20 min), washed in cold PBS and stored at -20 °C. Cells from four litres of culture were thawed in 200 mL lysis buffer (50 mM Tris (HCl) pH 7.4, 500 mM NaCl, 0.5 mM MgCl<sub>2</sub>, 5.0 mM DTT, 0.05% w/v Tween-20 supplemented with 50 μg/mL DNase I and EDTA-free protease inhibitors) and lysed using a cell disruptor (24 kPSI, 4 °C). The insoluble fraction was pelleted by centrifugation (39,000 × g, 40 min, 4 °C) and discarded. Supernatant was incubated (2 h, 4 °C) with 4.0 mL of Glutathione Sepharose 4B resin (GE Healthcare) that had been pre-equilibrated in the same buffer. Recombinant protein was washed three times by centrifugation (600 × g, 10 min, 4 °C) and re-suspension in 150 mL wash buffer (50 mM Tris (HCl) pH 7.4, 500 mM NaCl, 5.0 mM DTT). Washed resin was washed with 20 mL wash buffer supplemented with 25 mM reduced glutathione (1 h, 4 °C) and re-suspension in 150 mL wash buffer (50 mM Tris (HCl) pH 7.4, 500 mM NaCl, 5.0 mM DTT). Washed resin was transferred to a gravity column and protein was eluted in batch with 20 mL wash buffer supplemented with 25 mM reduced glutathione (1 h, 4 °C). GST-tagged 3C protease was added to the eluate (10 μg/mL) and the mixture was dialysed (3K molecular weight cut-off (MWCO), 4 °C, 16 h) against 2 L wash buffer to remove the glutathione. Dialysed proteins were then re-incubated with Glutathione Sepharose 4B resin (as above; 2 h, 4 °C) to remove the cleaved GST and GST-3C protease. The flow-through was subjected to heparin-affinity chromatography to remove nucleic acids. Samples were loaded on a 10 mL HiTrap Heparin column (GE Healthcare) at 2.0 mL/min, washed with two column volumes of buffer A (50 mM
Tris (HCl) pH 7.4, 500 mM NaCl, 1.0 mM DTT and eluted with a 0% → 100% gradient of buffer B (50 mM Tris (HCl) pH 8.0, 1.0 M NaCl, 1.0 mM DTT) over 20 column volumes. After removal of nucleic acids, protein became aggregation-prone and precipitated at low temperatures, therefore all subsequent steps were performed at 20 °C. Fractions corresponding to the 2A peak were pooled and concentrated using an Amicon® Ultra centrifugal filter unit (10K MWCO, 4,000 × g) prior to size exclusion chromatography (Superdex 75 16/600 column; 10 mM HEPES pH 7.9, 1.0 M NaCl, 1.0 mM DTT). Purity was assessed by 4-20% gradient SDS-PAGE, and protein identity verified by mass spectrometry. Purified protein was used immediately for crystallisation trials or was concentrated (~ 4.4 mg/mL, 282 μM), snap-frozen in liquid nitrogen and stored at -80 °C.

Size-exclusion chromatography coupled to multi-angle laser scattering (SEC-MALS)

For studies of the TMEV 2A protein in isolation, a Superdex 75 increase 10/300 GL column (GE Healthcare) was equilibrated with 20 mM Tris (HCl) pH 7.5, 1.0 M NaCl (0.4 mL/min flow, 25 °C). Per experiment, 100 μL of protein was injected at concentrations of 3.1, 0.8 and 0.38 mg/mL (molar concentrations of 200, 51.5 and 24.5 μM, respectively). The static light scattering, differential refractive index, and the UV absorbance at 280 nm were measured by DAWN 8+ (Wyatt Technology), Optilab T-rEX (Wyatt Technology), and Agilent 1260 UV (Agilent Technologies) detectors. The corresponding molar mass from each elution peak was calculated using ASTRA 6 software (Wyatt Technology) using the differential refractive index and a dn/dc value of 0.186 to calculate the protein concentration. For studies of 2A-RNA complexes, samples were recovered directly from the ITC cell after confirmation of binding and concentrated to an A_{280} of ~ 3.1 prior to injection of 100 μL onto a Superdex 75 increase 10/300 GL column pre-equilibrated in 50 mM Tris (HCl) pH 7.4, 200 mM NaCl (0.4 mL/min flow, 25 °C). Data were recorded as above, and to estimate the relative contributions of protein and RNA to molar mass, a protein conjugate analysis was performed within ASTRA 6, using a protein dn/dc value of 0.186 and an RNA (modifier) dn/dc value of 0.168. Prior to this analysis, extinction coefficients (at 280 nm) were determined experimentally from protein-only and RNA-only peaks using the “UV extinction from RI peak” method in ASTRA 6.

Protein crystallization

Purified TMEV 2A was concentrated to 4.38 mg/mL in 50 mM Tris (HCl) pH 7.4, 1.1 M NaCl, 1.0 mM DTT. Sitting-drop vapor diffusion experiments were set up in 96-well MRC plates with 80 μL reservoir solution, 200 nL protein and 200 nL crystallization buffer. Diffraction-quality crystals grew in 0.2 M KBr, 0.2 M potassium thiocyanate, 0.1 M sodium cacodylate pH 6.5, 3% w/v poly-γ-glutamic acid 200-400 and 10% w/v PEG-MME 2000. Crystals were harvested
in nylon loops and cryo-protected by removal from the mother liquor through a 0.5 μL layer of crystallization buffer that had been supplemented with 20% v/v glycerol, prior to flash-cooling by plunging into liquid nitrogen.

X-ray data collection, structure determination, refinement and analysis

Datasets of 900 images (Table S1) were recorded from two crystals at beamline I04-1, Diamond Light Source on a Pilatus 6M detector (Dectris), using 47% transmission, an oscillation range of 0.2 ° and an exposure time of 0.2 s per image. Data were collected at a wavelength of 0.9159 Å and a temperature of 100 K. Reflections were indexed and integrated with DIALS (61) (highest resolution crystal) or XDS (62) (structure determination crystal) and data were scaled and merged with AIMLESS (63) within the Xia2 data reduction pipeline (64). Resolution cut-off was decided by a CC₁/₂ value ≥ 0.5 and an I/σ(I) ≥ 2.0 in the highest resolution shell (65). The structure was solved by single-wavelength anomalous dispersion (SAD) analysis of the structure determination crystal, using anomalous signal from bromine present in the crystallisation buffer. SAD phasing was performed using autoSHARP (66), implementing SHELXD for substructure determination (67), SHARP for density modification (66) and ARP/wARP (68) for automated model building. This placed 123 residues out of 133 (92%) in the single chain that comprised the asymmetric unit of the cubic P2₁3 cell. This preliminary model was subsequently used as a molecular replacement search model to solve a higher resolution dataset (highest resolution crystal) using Phaser (69). This model was then subjected to several rounds of manual adjustment using COOT (70) and refinement with phenix.refine (71). Upon completion of model building, ISOLDE (72) was used to improve model geometry and resolve clashes prior to a final round of refinement using phenix.refine. MolProbity (73) was used to assess model geometry including Ramachandran outliers, bad rotamers and mainchain geometry deviations throughout the refinement process. For the electrostatic potential calculations, PDB2PQR (74) and PROPKA were used to estimate protein pKa values and assign partial charges. Electrostatic surface potentials were calculated using APBS (75). Relative solvent-accessible surface areas per residue were calculated using GetArea (76) and crystallographic interaction interfaces were assessed using PDBePISA (77). Structural figures depicting crystallographic data (cartoon, stick and surface representations) were rendered in PyMOL (Schrödinger LLC). The representation of surface conservation was generated using ConSurf (78).

Nucleotide and protein sequence alignments

The Logo plot of nucleotide sequence conservation at the PRF region was generated from a selection of divergent isolates (Cardiovirus A – M8161, M22457, KP892662, LC585221,
KC310737, JX257003; Cardiovirus B – EU542581, M20301, MF172923, MF352420; Cardiovirus D – EU376934; Cardiovirus E, F and unassigned – KY432928, KY432930, KY855434, KF823815) with WebLogo 2.8.2 (79) using the default parameters. For 2A protein sequence alignments, the match > align tool in UCSF Chimera (80) was first used to generate a seed alignment based on superposed structures of EMCV and TMEV 2A, prior to the subsequent alignment of other selected divergent TMEV-like (MF352420, M20301, MF172923, EU542581, EU376394, KY432930, KY432928, KF823815, KY855434) and EMCV-like (LC585221, KP892662, M81861, M22457, JX257003, KC310737) 2A sequences. JalView (81) was used to visualise the alignment, calculate the consensus sequence and generate the associated Logo plot.

**Electrophoretic Mobility Shift Assay (EMSA)**

RNA oligonucleotides (IDT) were reconstituted in distilled water. A 5′ Cy5 fluorescent label was incorporated using the 5′ EndTag kit (Vector Labs) according to the manufacturer’s instructions, prior to phenol:chloroform extraction, ethanol precipitation and aqueous resuspension. Binding reactions of 10 μL contained 1.0 μL 500 nM Cy5-RNA, 1.0 μL TMEV 2A at concentrations of 280, 140, 70, 35, 17.5, 8.7, 4.4, and 2.2 μM in 10 mM HEPES pH 7.9, 1.0 M NaCl, 5.0 μL 2 × buffer (20 mM Tris (HCl) pH 7.4, 80 mM NaCl, 4.0 mM magnesium acetate, 2.0 mM DTT, 10% v/v glycerol, 0.02% w/v bromophenol blue, 200 μg/mL porcine liver tRNA and 800 U/mL SUPERase-In [Invitrogen]) and 3.0 μL distilled water. Final concentrations in the binding reactions were therefore 50 nM RNA, 1 × buffer, ~ 140 mM NaCl and TMEV 2A at 28.0, 14.0, 7.0, 3.5, 1.75, 0.87, 0.44 and 0.22 μM. All binding reactions were prepared on ice, and samples were incubated at 37 °C for 20 min before analysis by non-denaturing 10% acrylamide/TBE PAGE (25 min, 200 V constant). Gels were imaged with a Typhoon FLA-7000 (GE) using the 635 nm laser / R670 filter.

**Microscale Thermophoresis (MST)**

Synthetic RNA oligonucleotides (IDT) were labelled at the 5′ end with Cy5 as described above, prior to purification using Clean and Concentrator kit (Zymo). TMEV 2A in 10 mM HEPES pH 7.9, 1.0 M NaCl was diluted in 2 × buffer (20 mM Tris (HCl) pH 7.4, 80 mM NaCl, 4.0 mM magnesium acetate, 2.0 mM DTT) to a final concentration of 20 μM. For the measurement, a series of 16 1:1 dilutions was prepared and each ligand dilution was mixed with one volume of labeled TMEV RNA. Final concentrations in the binding reactions were therefore 5.0 nM RNA, 1 × buffer, ~ 140 mM NaCl and TMEV 2A ranging from 0.0006 to 20 μM. The reaction was mixed and the samples loaded into Monolith NT.115 Premium Capillaries (NanoTemper Technologies). Measurements were performed using a Monolith NT.115Pico instrument.
(NanoTemper Technologies) at an ambient temperature of 25 °C. Instrument parameters were adjusted to 5% LED power and medium MST power. Data of two independently pipetted measurements were analysed for fraction bound using initial fluorescence (MO.Affinity Analysis software, NanoTemper Technologies). The non-binding RNAs (namely TMEV 2 and 4) were normalized using Prism 8.0.2 (GraphPad). Data was plotted using Prism 8.0.2 (GraphPad) software.

Isothermal Titration Calorimetry (ITC)

ITC analyses were carried out at 25 °C using a MicroCal PEAQ-ITC (Malvern Panalytical). RNAs and proteins were dialysed (24 h, 25 °C) into buffer (50 mM Tris (HCl) pH 7.4, 400 mM NaCl) before performing experiments. Final concentrations of protein and RNA after dialysis were determined by spectrophotometry (A$_{280}$ and A$_{260}$, respectively), using theoretical extinction coefficients based on the primary sequence of each component. RNA (60 μM) was titrated into protein (5 μM) with 1 × 0.4 μL injection followed by 12 × 3.0 μL injections. Control titrations of RNA into buffer, buffer into protein and buffer into buffer were also performed. Data were analyzed using the MicroCal PEAQ-ITC analysis software (Malvern Panalytical) and binding constants determined by fitting a single-site binding model.

In vitro transcription

For in vitro frameshifting assays, a 105 nt portion of the TMEV genome containing the GGUUUUUU shift site flanked by 6 nt upstream and 92 nt downstream was cloned into the dual luciferase plasmid pDluc at the XhoI/BglII sites (82). The sequence was inserted between the Renilla and firefly luciferase genes so that firefly luciferase expression is dependent on −1 PRF. Wild-type or derivative frameshift reporter plasmids were linearized with FspI and capped run-off transcripts generated using T7 RNA polymerase as described (83). Messenger RNAs were recovered by phenol/chloroform extraction (1:1 v/v), desalted by centrifugation through a NucAway Spin Column (Ambion) and concentrated by ethanol precipitation. The mRNA was resuspended in water, checked for integrity by agarose gel electrophoresis, and quantified by spectrophotometry.

In vitro translation

Messenger RNAs were translated in nuclease-treated rabbit reticulocyte lysate (RRL) (Promega). Typical reactions were composed of 90% (v/v) RRL, 20 μM amino acids (lacking methionine) and 0.2 MBq [$^{35}$S]-methionine and programmed with ~50 μg/ml template mRNA. Reactions were incubated for 1 h at 30 °C. Samples were mixed with 10 volumes of 2× Laemmlli's sample buffer, boiled for 3 min and resolved by SDS-PAGE. Dried gels were
exposed to a Storage Phosphor Screen (PerkinElmer), the screen was then scanned in a Typhoon FLA7000 using the phosphor autoradiography mode. Bands were quantified using ImageQuant™ TL software. The calculations of frameshifting efficiency (% FS) took into account the differential methionine content of the various products and % FS was calculated as % −1FS = 100 × (IFS/MetFS) / [(IS/MetS) + (IFS/MetFS)]. In the formula, the number of methionines in the stop, − are denoted by MetS, MetFS respectively; while the densitometry values for the same products are denoted by IS and IFS respectively. All frameshift assays were carried out a minimum of three times.

**Cells and viruses**

BSR (single cell clone of BHK-21 cells, species *Mesocricetus auratus*, provided by Polly Roy, LSHTM, UK) were maintained in Dulbecco’s modified Eagle’s medium (DMEM), high glucose, supplemented with L-glutamine (1 mM), antibiotics, and fetal bovine serum (FBS) (5%), at 37 °C and 5% CO₂. Cells were verified as mycoplasma-free by PCR (e-Myco plus Mycoplasma PCR Detection Kit, iNtRON Biotechnology). Cells were seeded to achieve 80% confluence on the day they were infected with virus stocks listed in (23), based on GDVII isolate NC_001366 with three nucleotide differences present in WT and mutant viruses (G2241A, A2390G and G4437A; nt coordinates with respect to NC_001366). All infections, except for plaque assays, were carried out at a MOI of 3 in serum-free media, or media only for the mock-infected samples. After incubation at 37 °C for 1 h, inoculum was replaced with serum-free media supplemented with FBS (2%), and infected cells were incubated at 37 °C until harvesting or further processing.

**Plaque assays**

BSR cells in 6-well plates at 90% confluence were inoculated with serial dilutions of virus stocks for 1 h then overlaid with 2 ml DMEM supplemented with L-glutamine (1.0 mM), antibiotics, FBS (2%), and carboxymethyl cellulose (0.6%). Plates were incubated at 37 °C for 48 h then fixed with formal saline and stained with 0.1% toluidine blue.

**Western blots of 2A expression**

BSR cells in 35 mm dishes were inoculated for 1 h at a MOI of 3, and incubated at 37 °C until 2, 4, 6, 8, 10, or 12 hpi. Cells were washed with cold PBS and lysed in 200 µl 1X radioimmunoprecipitation assay (RIPA) buffer (50 mM Tris (HCl) pH 8, 150 mM sodium chloride, 1% NP-40 substitute, 0.5% sodium deoxycholate, 0.1% SDS) supplemented with Halt™ Protease Inhibitor Cocktail (1X). Samples were resolved by 4-20% gradient SDS-PAGE and transferred to 0.2 µm nitrocellulose membranes. Membranes were blocked with 5% w/v
milk dissolved in PBS (1 h, 25 °C). Primary antibodies were diluted in 5% w/v milk, PBS, 0.1% v/v Tween-20 and incubated with membranes (1 h, 25 °C). After three washes in PBS, 0.1% v/v Tween-20, membranes were incubated with IRDye fluorescent antibodies in 5% w/v milk, PBS, 0.1% v/v Tween-20 (1 h, 25 °C). Membranes were washed three times in PBS, 0.1% v/v Tween-20 and rinsed in PBS prior to fluorescent imaging with an Odyssey CLx platform (LI-COR). Antibodies used were rabbit polyclonal anti-2A (22) (1/1000); mouse monoclonal anti-GAPDH (1/20,000, clone G8795, Sigma-Aldrich), goat anti-rabbit IRDye 800 CW (1/10,000, LI-COR), and goat anti-mouse IgM (μ chain specific) IRDye 680RD (1/10,000, Li-Cor).

**Metabolic labelling of infected cells**

BSR cells in 24-well plates were infected at a MOI of 3 in a volume of 150 μl. After 1 h, the inoculum was replaced with DMEM containing 2% FBS (1 ml) and cells were incubated at 37 °C for 5, 7, or 9 h delineating the 8, 10, or 12 hpi - timepoints respectively. At the set timepoint post-infection, cells were incubated for 1 h in methionine- and serum-free DMEM, then radiolabelled for 1 h with [³⁵S]-methionine at 100 μCi ml⁻¹ (~1,100 Ci mmol⁻¹) in methionine-free medium. Cells were harvested and washed twice by resuspension in 1 ml of ice-cold phosphate-buffered saline (PBS) and pelleted at 13,000 g for 2 min. Cell pellets were lysed in 40 μl 4 x SDS-PAGE sample buffer and boiled for 5 min before analysis by SDS-PAGE. Dried gels were exposed to X-ray films or to phosphorimager storage screens. Image analysis was carried out using ImageQuantTL 7.0 to quantify the radioactivity in virus-specific products. Bands which were quantifiable for all three viruses (Table S3) were carried forward for normalisation by methionine content and then by an average of the quantifiable proteins upstream of the frameshift site as a loading control. Bands in the WT and SS lanes were normalised by the counterpart bands in the M3 lane to control for differences in protein turnover. Frameshift efficiency (%) is given by the equation 100 × [1-(downstream/upstream)], where downstream and upstream represent the average of the fully normalised intensity values for proteins downstream and upstream of the frameshift site, respectively.

**Ribosome profiling library preparation**

BSR cells in 9 cm² dishes were inoculated in triplicate for 1 h at 37 °C at a MOI of 3, or media only for the mock-infected samples. After inoculation, cells were incubated in serum-free media supplemented with FBS (2%) at 37 °C for 9 h. Cells were harvested and ribosome-protected fragments (RPFs) purified and prepared for next-generation sequencing as described (84), based on (32,85,86), with the following modifications. The cycloheximide pretreatment was omitted from the harvesting protocol and cells were instead washed with warm PBS before snap-freezing in liquid nitrogen. RNase I (Ambion) was added to final
concentration 0.5 U/µl (replicate 1) or 2.5 U/µl (replicate 2 and 3 – higher concentration to ameliorate incomplete trimming noticed in replicate 1), and digestion inhibited by adding SUPERase-In RNase inhibitor (Invitrogen) to final concentration 0.13 U/µl and 0.67 U/µl, respectively. The range of fragment sizes selected during polyacrylamide gel purification before ligation to adapters was increased to 19-34 nt, and size ranges for post-ligation gels adjusted accordingly. For replicate 3 samples, two gel slices were excised per library, to purify monosome-protected (19-34 nt) and broad spectrum (35-65 nt) fragments from the same lysate. Depletion of ribosomal RNA was carried out solely by use of the RiboZero Gold Human/Mouse/Rat kit (Illumina). Adapter sequences were based on the TruSeq small RNA sequences (Illumina), with an additional seven random nucleotides at the 5′-end of the 3′-adapter and the 3′-end of the 5′-adapter, to reduce the effects of ligation bias and enable identification of PCR duplicates. For the broad-spectrum libraries only the 3′-adapter contained the 7 random nucleotides. Libraries were deep sequenced on the NextSeq 500 platform (Illumina), and data made publicly available on ArrayExpress under accession numbers E-MTAB-9438 and E-MTAB-9437.

Ribosome profiling data analysis

Adapter sequences were removed and reads resulting from RNA fragments shorter than 19 nt discarded using fastx_clipper (version 0.0.14, parameters: -Q33 -l 33 -c -n -v). Sequences with no adapters and those which consisted only of adapters were discarded. PCR duplicates were removed using awk, and the seven random nucleotides originating from the adapters were trimmed from each end of the reads using seqtk trimfq (version 1.3). Reads were aligned to reference databases using bowtie1, allowing one mismatch and reporting only the best alignment (version 1.2.3, parameters: -v 1 --best), in the following order: rRNA, virus genome (vRNA), mRNA, ncRNA, mtDNA, and gDNA. Quality control analysis indicated some contamination of replicate 2 libraries with *E. coli* RNA (BL21). These reads do not exhibit the expected features of genuine RPFs, however, indicating that the contamination occurred after lysates were harvested and thus they do not affect our conclusions. Reads were re-mapped with the addition of a BL21 reference database (CP047231.1) before vRNA, to remove these reads. The rRNA database comprised GenBank accession numbers NR_003287.2, NR_023379.1, NR_003285.2 and NR_003286.2. The mRNA database was compiled from the 36827 *M. auratus* GenBank RefSeq mRNAs available on 17th Nov 2017, after removing transcripts with annotated changes in frame. The ncRNA and gDNA databases are from *M. auratus* Ensembl release 90 (genome assembly 1.0). Viral genome sequences were verified by *de novo* assembly with Trinity (version 2.9.1), and reversion rates for mutated bases were verified as below 0.5%. For the broad-spectrum dataset, reads with no detected adapter were
not discarded by fastx_clipper, PCR duplicates were not removed, and two mismatches were allowed during bowtie1 mapping, except to BL21 *E. coli*, for which one mismatch was allowed.

To visualise RPF distribution on the viral genome, the number of reads with 5′-ends at each position was counted and divided by the total number of positive-sense vRNA and host mRNA reads for that library, to normalise for library size and calculate reads per million mapped reads (RPM). For plots covering the entire virus genome, a sliding window running mean filter of 15 nt was applied. To generate the plot of WT and SS data normalised by M3 data, UTRs plus a small buffer (7 nt at the 5′-end of the CDS and 27 nt at the 3′-end) were excluded and the result of the 15 nt running mean at each position on the WT or SS genome was divided by the 15 nt running mean at the corresponding position on the M3 genome. This avoided any instances of division by zero. Only positive-sense reads were used, and library pairs for normalisation were allocated according to replicate number. For the broad-spectrum dataset, guided by the length distribution and phasing plots, reads of lengths 51-52, 54-55 and 57-64 nt were selected for analysis as potential “disome-protected fragments”, and the denominator for normalisation of disome-protected read densities to RPM was calculated using only reads of these lengths. For all plots of read distribution on the viral genome, a +12 nt offset was added to the 5′-end coordinate of the read before plotting, to reflect the inferred position of the ribosomal P site. For disome-protected fragments, this represents the position of the P site of the colliding ribosome. For plots in the main text showing data from only one replicate, replicate 3 was used. Relative adaptiveness values for sense codons to the cellular tRNA pool were downloaded from the Species-Specific tRNA Adaptive Index Compendium (49), for *Cricetulus griseus* as a proxy for *M. auratus*, on 3rd Aug 2020. For the phasing and length distribution quality control plots, only reads which map completely within the CDSs of host mRNAs were included.

Frameshift efficiency (%) is given by the equation \(100 \times [1-(\text{downstream}/\text{upstream})]\), where downstream and upstream represent the reads per kilobase per million mapped reads (RPKM) values for the respective regions (genomic coordinates: upstream 1368-3943; downstream 4577-7679) after normalisation of WT and SS densities by M3. The percentage of reads in each phase in the upstream, transframe, and downstream regions relative to the 2B* frameshift site were determined using reads with inferred P site positions in the following ranges: upstream 1098-4211, transframe 4247-4273, downstream 4308-7946. Phases were designated relative to the polyprotein reading frame.
Data Availability

The datasets produced in this study are available in the following databases:

- Structure factors and coordinates (X-ray crystallography): world-wide Protein Data Bank (wwPDB; https://www.ebi.ac.uk/pdbe/; PDB ID 7NBV)
- Deep sequencing data (Ribosome profiling): ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) accession numbers E-MTAB-9438 and E-MTAB-9437.

Acknowledgements

We thank the staff of Diamond Light Source beamline I04-1 for assistance with crystal screening and data collection. We thank Janet Deane for assistance with SEC-MALS experiments. We thank Tatyana Koch for expert technical assistance. This research was funded in whole, or in part, by the Wellcome Trust 202797/Z/16/Z. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. C.H.H. and S.N. are funded by a Wellcome Trust Investigator Award (202797/Z/16/Z) to I.B.. G.M.C. is funded by a Wellcome Trust PhD Studentship (203864/Z/16/Z). S.C.G. is funded by a Sir Henry Dale fellowship (098406/Z/12/B) funded by the Wellcome Trust and the Royal Society. A.E.F. and K.B. are supported by Wellcome Trust (106207) and European Research Council (646891) grants to A.E.F.. N.C. is funded by the Helmholtz Association and European Research Council StG (948636).

Author contributions

C.H.H. and S.N. cloned, expressed and purified proteins and performed all biochemical experiments. C.H.H. and S.C.G performed crystallography experiments. A.K. and N.C. performed MST experiments. G.C., K.B. and A.E.F performed and analysed ribosomal profiling data. All authors contributed to preparing figures and writing the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.
Figure 1. TMEV 2A adopts the beta-shell fold.

A. SDS-PAGE analysis of TMEV 2A after Ni-NTA, heparin affinity and size-exclusion chromatography. The gel was stained with Coomassie blue.

B. SEC-MALS analysis of 2A in high-salt buffer at 3.1 mg/mL. The differential refractive index curve (blue) is shown and the weight-averaged molar masses for indicated peaks are listed.

C. Structural representation of the beta-shell fold of TMEV 2A.

D. Comparison of TMEV 2A and EMCV 2A structures, highlighting the cloning extension.

E. Close-up view of the arginine loop region in TMEV 2A.
C. X-ray crystal structure of TMEV 2A in two orthogonal views. N- and C- termini are indicated, and amino acids introduced as a result of the cloning strategy are labelled in purple. <Inset> Electrostatic surface potential calculated at pH 7.4, coloured between +3 (blue) and −3 kT/e− (red).

D. Locations of mutations are colour-coded and shown as sticks: M1 (K24A / R28A, green); M2 (R45A, blue); M3 (R85A / R95A, red); M4 (K90A / K91A, pink).

E. Superposition of TMEV 2A (blue) with EMCV 2A (orange). <Insets> Close-up views of regions of interest including the divergent structure at the C-terminus, the longer α1-β4 loop present in EMCV 2A and the conserved arginine loop.
Figure 2. 2A recognises a minimal 37 nt stimulatory element in the viral RNA.

A, B. Sequences and schematic diagrams of the TMEV 1-6 constructs used to assay 2A binding.
C. EMSA analysis conducted with 50 nM Cy5-labelled RNAs and 2A concentrations between 0 and 28 μM. Following non-denaturing electrophoresis, fluorescence was imaged using a Typhoon scanner.

D. MST binding curves and apparent $K_D$ values using the same constructs.

E. Experiment showing the effects of titrating excess short RNAs (TMEV 1-6) as competitors into an in vitro frameshift reporter assay. The concentrations of the reporter mRNA and 2A were kept constant in the RRL and short RNAs were added in 10- and 100- fold molar excess relative to the reporter mRNA, as indicated. Translation products were visualised by using autoradiography, and % frameshifting was calculated following densitometry and correction for the number of methionines present in 0 frame and −1 frame products.
Figure 3. A conserved AAGG motif present in the 5′ extension is required for 2A binding.

A. Logo plot showing conservation of aligned cardiovirus RNA sequences (see Figure S4) spanning the experimentally defined minimal element. Asterisks indicate invariant nucleotides.

B. Schematic diagrams of the TMEV 7-12 sequences used to assay 2A binding.
C. EMSA analysis conducted with 50 nM Cy5-labelled RNAs and 2A concentrations between zero and 28 μM. Following non-denaturing electrophoresis, fluorescence was imaged using a Typhoon scanner.

D. Frameshift reporter assays showing that mutation of the loop CCC motif inhibits frameshifting, but complementary mutations in the 5’ AAGG motif that allow base pairing between positions 17 and 38 restore frameshifting.

E. EMSA analysis showing attenuation of 2A binding to the minimal TMEV 9 RNA by introducing point mutations to either the 5’ AAGG motif or loop CCC motif.

F. Hypothetical equilibrium between predicted stem-loop and alternate pseudoknot-like conformations, colour-coded as in A. The pseudoknot-like conformation involves a base-pair between G17 and G18 in the 5’ extension, and C38 and C37 in the loop (shown as sticks).
Figure 4. 2A binds to the RNA stimulatory element with equimolar stoichiometry and nanomolar affinity.

A. EMSA analysis showing the effects of salt concentration (indicated) on RNA binding. Experiments were conducted with 50 nM Cy5-labelled TMEV 6 and 2A concentrations between zero and 28 μM. Following non-denaturing electrophoresis, fluorescence was imaged using a Typhoon scanner.

B. ITC isotherm for a titration of TMEV 6 RNA into 2A protein in the presence of 400 mM NaCl.

C. Binding curve from titration in B, showing approximately 1:1 molar ratio and nanomolar affinity. Histogram showing relative contributions of ΔH and TΔS terms to the overall exergonic interaction.

D. SEC-MALS analysis of 2A-TMEV 6 RNA complex in buffer containing 200 mM NaCl. The 280 nm UV absorbance trace is shown (green). Weight-averaged molar masses across the indicated peaks are listed, along with mass contributions from the protein (blue) and RNA (pink) components, following a protein conjugate analysis. The two peaks correspond to the RNA-protein complex (peak 1; Mw 33.6 kDa) and the excess RNA (peak 2; Mw 11.0 kDa).
Figure 5. 2A stimulates -1 PRF with 85% efficiency and leads to ribosomal pausing in infected cells.

A. Metabolic labelling of BSR cells infected with WT, SS, or M3 TMEV and harvested at 10 hpi. Positions of TMEV proteins are indicated, with those downstream of the frameshift site written in red.
B. Frameshift (FS) efficiency in infected cells, calculated from metabolic labelling (blue bars) and ribosome profiling (green bars). Bars represent the mean of three replicates, with values for each replicate indicated by crosses.

C. Schematic of ribosome profiling methodology. Cells are flash frozen before lysis. RNase I is added to digest regions of unprotected RNA, then ribosomes and enclosed ribosome-protected fragments (RPFs) of RNA are purified. RPFs are released and prepared for high-throughput sequencing to determine the positions of translating ribosomes at the time of cell lysis.

D. RPF densities in reads per million mapped reads (RPM) on WT, SS and M3 viral genomes (top panel), after application of a 15 nt running mean filter, from cells harvested at 10 hpi. Positive sense reads are plotted in green (above the horizontal axis), negative sense in red (below the horizontal axis; negligible amounts). In all plots, RPF densities from replicate 3 are shown, plotted at the inferred position of the ribosomal P site.

E. Positive-sense RPF densities from D, coloured according to phase (purple, blue, and yellow represent RPFs whose 5′ ends map to, respectively, the 1st, 2nd or 3rd nucleotides of polyprotein-frame codons, defined as phases 0, +1/−2, and +2/−1), after application of a 15 codon running mean filter. Frames for each viral ORF are designated with respect to the polyprotein (set to 0) and indicated on the genome map by colour and by an offset on the y axis in the corresponding direction.

F. Ratio of WT or SS RPF density at each position on the genome relative to the RPF density on the M3 mutant genome. UTRs were excluded, only positive sense reads were used, and a 15 nt running mean filter was applied before the division. Regions defined as “upstream” and “downstream” in the ribosome profiling frameshift efficiency calculations are annotated below, and average densities for these regions displayed to the left of the plots with grey lines.

G. Inferred positions of ribosomal P sites at the slippery sequence and in the 2B* region, coloured according to the phase of RPF 5′ ends as in E. The genome sequence in this region is underlaid beneath the data in the top panel, and all libraries are set to the same scale on the y axis, with no running mean filter applied.
Figure 6. Profiling of disome-protected fragments reveals ribosome stacking at the frameshift site.

A. Length distribution of reads mapping within host CDSs for broad spectrum libraries.

B. Total number of reads attributed to each phase, for reads of each specified length, in the broad-spectrum libraries. Read lengths which were selected for inclusion in the “disome-protected fragment” plots (C and D) are indicated by square brackets, and correspond to peaks in A.

C. Density of disome-protected fragments on the WT and M3 viral genomes, derived from the same lysate as the monosome-protected fragment data plotted in Figure 5 D-G. Reads of lengths 51-52, 54-55 and 57-64 nt were selected for inclusion, and read densities are plotted as RPM after application of a 15 nt running mean filter. Positive sense reads are plotted in green (above the horizontal axis), negative sense in red (below the horizontal axis). In all disome plots, reads are plotted at the inferred P site position of the colliding ribosome.
D. Density of disome-protected fragments, plotted at inferred P site positions of colliding ribosomes involved in disomes upstream of the 2B* frameshift site, coloured according to phase as in B. The encoded amino acid sequence in this region is underlaid beneath the data in the top panel, and both libraries are set to the same scale on the y axis, with no running mean filter applied. Codons on which leading ribosomes would be expected to pause due to StopGo and frameshifting (FS) are indicated in brown and grey respectively. Positions of ribosomes potentially involved in queue formation behind these pause sites are indicated (FS: above genome map; StopGo: below genome map), with P site positions annotated as dashed vertical lines, in the corresponding colours.
Supplementary Figures and Legends

Figure S1. TMEV 2A trimers observed in the crystalline lattice.
View of the trimeric protein assembly at the three-fold crystallographic symmetry axis. TMEV 2A molecules are shown as cartoons (cyan) in two orthogonal views, centred around the symmetry axis (black triangle and dotted line). This arrangement may correspond to trimeric assemblies predicted by PDBePISA and observed as a minor species in solution by SEC-MALS (Figure 1B). The interface between subunits is predominantly formed by the N-terminal cloning tag extension (G_{4}P_{3}L_{2}G_{1}S_{0}; red). The trimeric assembly is therefore not likely to be physiologically relevant.
Figure S2 (previous page). Comparison of TMEV 2A structure with orthologue from EMCV reveals a divergent RNA-binding surface and C-terminal region.

A. Amino acid sequence alignment of selected divergent TMEV-like (upper) and EMCV-like (lower) cardiovirus 2A protein sequences, guided by the structural alignment of EMCV and TMEV 2A proteins. Known secondary structures are indicated above the corresponding sequence for EMCV and TMEV proteins. N- and C- terminal residues not observed in the structures are greyed out. Conservation is highlighted in blue. Local motifs of functional significance are highlighted and annotated.

B. Surface of TMEV 2A, coloured by conservation from highly conserved (purple) to variable (teal). Highly conserved, surface-accessible residues are shown as sticks.

C. Comparison of RNA-binding residues in EMCV 2A (28) (beige) with equivalent surface in TMEV 2A (pale blue). Residues involved in RNA binding come from three regions of the EMCV protein as indicated in A (shown as red, crimson and brown sticks). TMEV residues that may be functionally-equivalent are shown as sticks (cyan).

D. Comparison of the YxxxxLΦ binding motif in 4E-BP1 and putative motif in TMEV 2A. The crystal structure of the complex between eIF4E and 4E-BP1 is shown (green and blue, respectively) with 2A (pale blue) docked via least-squares superposition of the YxxxxLΦ motif. A local surface cutaway, removing a section of β3 (dashed lines), shows the buried location of Y119. <Inset> Contrast between the two different helical conformations of the putative 2A YxxxxLΦ motif and the 4E-BP1 YxxxxLΦ motif, in a compact α-helical conformation.

E. Conformational differences between the C-terminal putative YxxxxLΦ motif in TMEV 2A (pale blue) and EMCV 2A (wheat).
Figure S3. Mutagenesis of putative RNA-binding residues demonstrate the importance of the conserved arginine loop.

A. Mutagenesis of TMEV residues equivalent to those observed at the RNA binding surface in the EMCV 2A-70S\textsubscript{IC} structure - see also Figure S2C. The locations of mutations R85A / R87A, K63A, K83A and K53A / H56A are shown as sticks.

B. EMSA analyses showing effects of the above mutations on stimulatory element RNA binding, compared to a wild-type control. Experiments were conducted with 50 nM Cy5-labelled TMEV 6 RNA and 2A concentrations as indicated between zero and 5.0 \( \mu \text{M} \). Following non-denaturing electrophoresis, fluorescence was imaged using a Typhoon scanner.
**Cardiovirus A / encephalomyocarditis virus**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1861</td>
<td>AAA CAG GUU UUC CAG ACC CAA GGA CAC GCA UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC GAC CUG</td>
</tr>
<tr>
<td>M2457</td>
<td>AAG CAG GUU UUC CAG ACC CAG GGA CAC GCA UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC GAC CUG</td>
</tr>
<tr>
<td>KF890662</td>
<td>AAG CAG GUU UUC CAG ACA CAG GCA CGU UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC GAC CUG</td>
</tr>
<tr>
<td>LC598221</td>
<td>AAA CAG GUU UUC CAG CCA CAA GGA GGA UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC GAC CUG</td>
</tr>
<tr>
<td>K0310737</td>
<td>AAA CAG GUU UUC CAG ACA CAA GGA GGA GGA UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC GAC CUG</td>
</tr>
<tr>
<td>JX257903</td>
<td>AAA CAG GUU UUC CAG ACC CAA GGA GGA UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC GAC CUG</td>
</tr>
</tbody>
</table>

**Cardiovirus B / TMEV, RTV, etc.**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED542381</td>
<td>CAG UCC GUU UUC CAG CCA CAA GGU GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
<tr>
<td>MG0361</td>
<td>CAG UCC GUU UUC CAG CCA CAA GGU GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
<tr>
<td>MF172923</td>
<td>CAG UCC GUU UUC CAG CCA CAA GGU GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
<tr>
<td>MF352420</td>
<td>CAG UCC GUU UUC CAG CCA CAA GGU GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
</tbody>
</table>

**Cardiovirus D / Saffold virus**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>E0376394</td>
<td>CAG UCC GUU UUC CAG UUG CAA GGU GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
</tbody>
</table>

**Cardiovirus E, F and unassigned**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>KY432928</td>
<td>AUG GGU UUC GUU CAG CCA CAA GGU GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
<tr>
<td>KY432930</td>
<td>GCU CAG GUU UUC CAG GAA CAA GGC GCG GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
<tr>
<td>KY855434</td>
<td>AUG CGU CUC GUU CAG GAA CAA GGC GCG GGU UGU UCA AUG UCA GGA AUA CAG AAC</td>
</tr>
<tr>
<td>KF823815</td>
<td>AAC CAG GUU UUC CAG CCA CAA GGA GGA UGU UCA AUG GCU AAA ACC CUA GUG CUG AAC</td>
</tr>
</tbody>
</table>

**Figure S4. Conservation at the frameshift stimulatory site in cardioviruses.**

Sequence of the PRF region from representative cardiovirus isolates, showing the highly conserved shift site (light blue) and 3’ RNA stem-loop structure (green). Nucleotide variations in the shift site are indicated in pale pink. Paired substitutions that preserve the predicted structure are highlighted in crimson, purple or orange; single substitutions that – via G:U base-pairing are compatible with the predicted structure – are highlighted in cyan. The conserved CCC loop triplet is shown in blue.
Figure S5. WT and mutant 2A in the context of infection.

A. Plaque assays of BSR cells infected with WT, SS, or M3 TMEV and fixed at 48 hpi.

B. Western blot of BSR cells infected with WT TMEV and harvested over a timecourse up to 12 hpi, or mock-infected (M) and harvested immediately. The SS mutant was assayed in parallel and produced similar results (data not shown).

C. Metabolic labelling of BSR cells infected with WT, SS, or M3 TMEV and harvested over a timecourse of 8-12 hpi. Positions of TMEV proteins are indicated, with those downstream of the frameshift site written in red. At 8 hpi, reliable quantification of viral proteins above the background of host translation was not possible. However, at 10 and 12 hpi, a large proportion of ongoing translation is viral, likely due to virus-induced shut-off of host gene expression, and viral proteins were clearly visible and quantifiable.
Figure S6. Ribosome profiling quality control analysis indicates high quality data.

A. Western blot of lysates used for ribosome profiling. Note that the mutant M3 2A migrates slightly faster during gel electrophoresis than the WT protein.

B. Length distribution of RPFs mapping within host (green) and viral (pink, mock excluded) CDSs in each library.

C. Percentage of reads (all read lengths) attributed to each phase, from reads mapping within host CDSs. Noting that approximately one third of the reads in replicate 1 libraries were attributed to the −1/+2 phase, we increased the amount of RNase I added to further replicates to improve trimming, yielding the very high proportion of phase 0 reads seen in replicates 2 and 3. Phase compositions of virus CDS-mapping reads closely matched those of host-mapping reads (Figure S8A, upstream and downstream panels).
D. Distribution of inferred P site positions of host mRNA-mapping reads relative to start and stop codons. Only transcripts with an annotated CDS of at least 150 codons and UTRs of at least 60 nt were included in the analysis, and the total number of reads from all these transcripts mapping to each position was plotted. RPFs map to coding sequences with a triplet periodicity and few RPFs map to the UTRs, particularly the 3′UTR. Typically heightened RPF peaks corresponding to the sites of translation initiation and termination are also observed (32,87).
Figure S7. Further replicates of ribosome profiling RPF distribution plots.

A. RPF distribution on the viral genome, as in main text Figure 6D, for replicates 1 and 2.

B. Ratio of RPF density on the WT or SS genome normalised by M3, as in main text Figure 6F, for replicates 1 and 2.

C. RPF distribution at the frameshift site, as in main text Figure 6G, for replicates 1 and 2.
Figure S8. Investigation of the region around the frameshift site by ribosome profiling.

A. Percentage of reads (all read lengths) attributed to each phase in the regions upstream and downstream of 2B* and in the short 2B/2B* overlap region (“transframe”).

B. Density of RPFs in the 200 nt up to and including the frameshift site, plotted at inferred P site positions of the ribosomes, and coloured according to RPF phase as in D. The encoded amino acid sequence in this region is underlaid beneath the data in the top panel, and all libraries are set to the same scale on the y axis, with no running mean filter applied. The last amino acid of 2A – after which StopGo-mediated peptide release occurs – is annotated in brown, the P-site of frameshifting (FS) in grey, and the RNA-binding arginines mutated to alanine in M3 in black.

C. Histogram of relative adaptiveness values of each sense codon to the cellular tRNA pool, defined based on a combination of intracellular tRNA abundance and the strength of the
codon-anticodon interaction. Values for the Chinese hamster *Cricetulus griseus* were taken from the Species-Specific tRNA Adaptive Index Compendium (49) as *M. auratus* values were not available. The adaptiveness values of the CGC arginine codons encoding R85 and R87, and the GCC alanine codons to which they are mutated in M3, are indicated by dashed lines.
Supplementary Tables

Table S1 – Crystallographic data collection and refinement.

<table>
<thead>
<tr>
<th></th>
<th>Highest resolution†</th>
<th>Structure determination†</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9159</td>
<td>0.9159</td>
</tr>
<tr>
<td>Space group</td>
<td>(P 2_1 3)</td>
<td>(P 2_1 3)</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a = b = c) (Å)</td>
<td>73.77</td>
<td>74.24</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>43.59–1.87 (1.92–1.87)</td>
<td>52.48–2.11 (2.15–2.11)</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>11,355 (577)</td>
<td>8100 (389)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>100.0 (100.0)</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Anomalous</td>
<td>-</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Multiplicity</td>
<td>19.0 (19.8)</td>
<td>19.4 (20.3)</td>
</tr>
<tr>
<td>Anomalous</td>
<td>-</td>
<td>10.3 (10.6)</td>
</tr>
<tr>
<td>(R_{merge})</td>
<td>0.113 (1.191)</td>
<td>0.078 (1.956)</td>
</tr>
<tr>
<td>(R_{pim})</td>
<td>0.026 (0.273)</td>
<td>0.025 (0.629)</td>
</tr>
<tr>
<td>(CC_{1/2})</td>
<td>0.999 (0.644)</td>
<td>1.00 (0.715)</td>
</tr>
<tr>
<td>(CC_{anom})</td>
<td>-</td>
<td>0.504 (0.002)</td>
</tr>
<tr>
<td>Mean I/σ(I)</td>
<td>16.8 (2.8)</td>
<td>26.9 (1.7)</td>
</tr>
<tr>
<td>Wilson B (Å²)</td>
<td>26.06</td>
<td>46.96</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>42.61–1.87 (1.93–1.87)</td>
<td></td>
</tr>
<tr>
<td>Reflections ‡</td>
<td>21,277 (2095)</td>
<td></td>
</tr>
<tr>
<td>Working set</td>
<td>20,162 (1961)</td>
<td></td>
</tr>
<tr>
<td>Test set</td>
<td>1115 (134)</td>
<td></td>
</tr>
<tr>
<td>(R_{work})</td>
<td>0.1847</td>
<td></td>
</tr>
<tr>
<td>(R_{free})</td>
<td>0.2302</td>
<td></td>
</tr>
<tr>
<td>No. of atoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>1112</td>
<td></td>
</tr>
<tr>
<td>Solvent</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Ions</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Root mean square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.890</td>
<td></td>
</tr>
<tr>
<td>Ramachandran</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>favoured (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ramachandran outliers (%)</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Poor rotamers (%)</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Mean B value (Å²)</td>
<td>31.6</td>
<td></td>
</tr>
</tbody>
</table>

*Values in parentheses are for the highest resolution shell
† Both crystals were harvested from the same crystallisation drop
‡ I+ and I- were treated as separate reflections in phenix.refine
Table S2. Number of ribosome profiling reads assigned to each category. The "_broad" samples are the broad-spectrum libraries, for which 35-65 nt fragments were purified. Reads under 19 nt long (for monosome libraries) or under 35 nt long (for broad-spectrum libraries) were defined as too short. Some categories are not applicable to the broad-spectrum libraries due to minor differences in computational processing (detailed in Methods).

<table>
<thead>
<tr>
<th>Virus</th>
<th>Replicate</th>
<th>Total reads</th>
<th>Too short</th>
<th>Adapter-only</th>
<th>No adapter</th>
<th>PCR duplicates</th>
<th>rRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>1</td>
<td>1790088</td>
<td>61648</td>
<td>876</td>
<td>12076</td>
<td>28721</td>
<td>586963</td>
</tr>
<tr>
<td>WT</td>
<td>2</td>
<td>26408581</td>
<td>2340239</td>
<td>297018</td>
<td>116064</td>
<td>1834921</td>
<td>2926112</td>
</tr>
<tr>
<td>WT</td>
<td>3</td>
<td>28688961</td>
<td>2756571</td>
<td>280096</td>
<td>406493</td>
<td>2627172</td>
<td>4719318</td>
</tr>
<tr>
<td>SS</td>
<td>1</td>
<td>2392994</td>
<td>34429</td>
<td>1859</td>
<td>9565</td>
<td>61705</td>
<td>505571</td>
</tr>
<tr>
<td>SS</td>
<td>2</td>
<td>25027728</td>
<td>2854707</td>
<td>353732</td>
<td>126430</td>
<td>1456929</td>
<td>2342432</td>
</tr>
<tr>
<td>SS</td>
<td>3</td>
<td>21788351</td>
<td>1927596</td>
<td>320336</td>
<td>503529</td>
<td>2258152</td>
<td>3634096</td>
</tr>
<tr>
<td>M3</td>
<td>1</td>
<td>2290607</td>
<td>149115</td>
<td>6052</td>
<td>18977</td>
<td>61127</td>
<td>619693</td>
</tr>
<tr>
<td>M3</td>
<td>2</td>
<td>27308973</td>
<td>1702862</td>
<td>157048</td>
<td>138108</td>
<td>1654696</td>
<td>2986823</td>
</tr>
<tr>
<td>M3</td>
<td>3</td>
<td>30504465</td>
<td>2009396</td>
<td>515817</td>
<td>438248</td>
<td>3248459</td>
<td>5122599</td>
</tr>
<tr>
<td>Mock</td>
<td>1</td>
<td>1914259</td>
<td>171017</td>
<td>1448</td>
<td>27268</td>
<td>34259</td>
<td>495928</td>
</tr>
<tr>
<td>Mock</td>
<td>2</td>
<td>26398293</td>
<td>2345528</td>
<td>117083</td>
<td>120712</td>
<td>1286218</td>
<td>2852963</td>
</tr>
<tr>
<td>Mock</td>
<td>3</td>
<td>27548554</td>
<td>1532098</td>
<td>145612</td>
<td>608503</td>
<td>2112106</td>
<td>4820304</td>
</tr>
<tr>
<td>WT_broad</td>
<td>3</td>
<td>10624342</td>
<td>2528401</td>
<td>1155</td>
<td></td>
<td>2546503</td>
<td></td>
</tr>
<tr>
<td>M3_broad</td>
<td>3</td>
<td>7230428</td>
<td>1686167</td>
<td>1182</td>
<td></td>
<td>2090553</td>
<td></td>
</tr>
<tr>
<td>mock_broad</td>
<td>3</td>
<td>9730736</td>
<td>1814114</td>
<td>950</td>
<td></td>
<td>3247839</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virus</th>
<th>Replicate</th>
<th>vRNA</th>
<th>mRNA</th>
<th>ncRNA</th>
<th>mtDNA</th>
<th>gDNA</th>
<th>BL21 E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>1</td>
<td>60322</td>
<td>371302</td>
<td>71085</td>
<td>1909</td>
<td>81527</td>
<td>2196</td>
</tr>
<tr>
<td>WT</td>
<td>2</td>
<td>799308</td>
<td>3476556</td>
<td>1470387</td>
<td>119371</td>
<td>930513</td>
<td>8738367</td>
</tr>
<tr>
<td>WT</td>
<td>3</td>
<td>1299982</td>
<td>5671747</td>
<td>2271869</td>
<td>61429</td>
<td>1890651</td>
<td>914704</td>
</tr>
<tr>
<td>SS</td>
<td>1</td>
<td>198472</td>
<td>700308</td>
<td>189411</td>
<td>7642</td>
<td>139138</td>
<td>6991</td>
</tr>
<tr>
<td>SS</td>
<td>2</td>
<td>769385</td>
<td>2293504</td>
<td>980217</td>
<td>49001</td>
<td>527344</td>
<td>11085126</td>
</tr>
<tr>
<td>SS</td>
<td>3</td>
<td>1136138</td>
<td>4143979</td>
<td>1661076</td>
<td>64024</td>
<td>1486102</td>
<td>650083</td>
</tr>
<tr>
<td>M3</td>
<td>1</td>
<td>102620</td>
<td>485732</td>
<td>111611</td>
<td>2746</td>
<td>129914</td>
<td>5070</td>
</tr>
<tr>
<td>M3</td>
<td>2</td>
<td>946929</td>
<td>3136082</td>
<td>1210389</td>
<td>56677</td>
<td>837885</td>
<td>12118144</td>
</tr>
<tr>
<td>M3</td>
<td>3</td>
<td>1669867</td>
<td>6825341</td>
<td>2494147</td>
<td>105091</td>
<td>2404980</td>
<td>586000</td>
</tr>
<tr>
<td>Mock</td>
<td>1</td>
<td>161</td>
<td>481985</td>
<td>88888</td>
<td>3701</td>
<td>105409</td>
<td>4205</td>
</tr>
<tr>
<td>Mock</td>
<td>2</td>
<td>426</td>
<td>4817616</td>
<td>904360</td>
<td>26897</td>
<td>1216025</td>
<td>10404660</td>
</tr>
<tr>
<td>Mock</td>
<td>3</td>
<td>742</td>
<td>9839798</td>
<td>1651028</td>
<td>67215</td>
<td>1765457</td>
<td>427314</td>
</tr>
<tr>
<td>WT_broad</td>
<td>3</td>
<td>97581</td>
<td>534229</td>
<td>1322626</td>
<td>271988</td>
<td>413111</td>
<td>485534</td>
</tr>
<tr>
<td>M3_broad</td>
<td>3</td>
<td>20310</td>
<td>327007</td>
<td>783510</td>
<td>268166</td>
<td>270980</td>
<td>110028</td>
</tr>
<tr>
<td>mock_broad</td>
<td>3</td>
<td>31</td>
<td>349897</td>
<td>1038588</td>
<td>294835</td>
<td>295848</td>
<td>163860</td>
</tr>
</tbody>
</table>
Table S3. Phosphorimager signals for viral-specific bands on metabolic labelling gels.

<table>
<thead>
<tr>
<th>Protein</th>
<th>WT</th>
<th>SS</th>
<th>M3</th>
<th>Carried forward for calculations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>2395.08</td>
<td>2470.9</td>
<td>6697.62</td>
<td>Yes</td>
</tr>
<tr>
<td>VP0</td>
<td>3099.31</td>
<td></td>
<td></td>
<td>No - band too faint for reliable quantification in mutant lanes</td>
</tr>
<tr>
<td>2C</td>
<td>2741.89</td>
<td>5435.87</td>
<td>10800.7</td>
<td>Yes</td>
</tr>
<tr>
<td>VP1</td>
<td>2976.72</td>
<td>1511.32</td>
<td>3126.03</td>
<td>Yes</td>
</tr>
<tr>
<td>3C</td>
<td>861.87</td>
<td>1954.94</td>
<td>No - band too faint for reliable quantification in mutant lanes</td>
<td></td>
</tr>
<tr>
<td>VP3</td>
<td>2703.24</td>
<td>1402.52</td>
<td>3178.16</td>
<td>Yes</td>
</tr>
<tr>
<td>2A</td>
<td>2325.6</td>
<td>1083.08</td>
<td>1736.97</td>
<td>No - methionine-normalised results suggest difference in turnover rate between WT and mutant 2A</td>
</tr>
<tr>
<td>2B</td>
<td>853.34</td>
<td>1141.5</td>
<td>2323.76</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>WT</th>
<th>SS</th>
<th>M3</th>
<th>Carried forward for calculations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>28715.83</td>
<td>26525.91</td>
<td>67503.74</td>
<td>Yes</td>
</tr>
<tr>
<td>VP0</td>
<td>32548</td>
<td></td>
<td></td>
<td>No - band too faint for reliable quantification in mutant lanes</td>
</tr>
<tr>
<td>2C</td>
<td>25909.88</td>
<td>39483.35</td>
<td>90913.71</td>
<td>Yes</td>
</tr>
<tr>
<td>VP1</td>
<td>28175.9</td>
<td>13058.42</td>
<td>31910.11</td>
<td>Yes</td>
</tr>
<tr>
<td>3C</td>
<td>9944.51</td>
<td>20698.75</td>
<td>No - band too faint for reliable quantification in WT lane</td>
<td></td>
</tr>
<tr>
<td>VP3</td>
<td>23789.64</td>
<td>12568.31</td>
<td>28983.54</td>
<td>Yes</td>
</tr>
<tr>
<td>2A</td>
<td>15459.04</td>
<td>8312.66</td>
<td>13068.13</td>
<td>No - methionine-normalised results suggest difference in turnover rate between WT and mutant 2A</td>
</tr>
<tr>
<td>2B</td>
<td>9735.73</td>
<td>16639.4</td>
<td></td>
<td>No - band too faint for reliable quantification in WT lane</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>WT</th>
<th>SS</th>
<th>M3</th>
<th>Carried forward for calculations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>1272.31</td>
<td>1287.21</td>
<td>2652.87</td>
<td>Yes</td>
</tr>
<tr>
<td>VP0</td>
<td>1577.01</td>
<td></td>
<td></td>
<td>No - band too faint for reliable quantification in mutant lanes</td>
</tr>
<tr>
<td>2C</td>
<td>1116.87</td>
<td>2546.23</td>
<td>3469.65</td>
<td>Yes</td>
</tr>
<tr>
<td>VP1</td>
<td>1155.83</td>
<td>729.01</td>
<td>889.83</td>
<td>Yes</td>
</tr>
<tr>
<td>3C</td>
<td>600.71</td>
<td></td>
<td></td>
<td>No - band too faint for reliable quantification in WT lane</td>
</tr>
<tr>
<td>VP3</td>
<td>1089.86</td>
<td>608.08</td>
<td>1074.88</td>
<td>Yes</td>
</tr>
<tr>
<td>2A</td>
<td></td>
<td></td>
<td></td>
<td>No - band not resolvable from dye front</td>
</tr>
<tr>
<td>2B</td>
<td></td>
<td></td>
<td></td>
<td>No - band not resolvable from dye front</td>
</tr>
<tr>
<td>12 hpi replicate 1</td>
<td>Protein</td>
<td>WT</td>
<td>SS</td>
<td>M3</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>3D</td>
<td>14521.16</td>
<td>21433.11</td>
<td>43338.38</td>
<td>Yes</td>
</tr>
<tr>
<td>VP0</td>
<td>31017.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2C</td>
<td>19216.2</td>
<td>33218.68</td>
<td>63307.42</td>
<td></td>
</tr>
<tr>
<td>VP1</td>
<td>28838.43</td>
<td>11491.95</td>
<td>20683.21</td>
<td>Yes</td>
</tr>
<tr>
<td>3C</td>
<td>9132.88</td>
<td>16079.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP3</td>
<td>23912.68</td>
<td>11430.74</td>
<td>20631.82</td>
<td>Yes</td>
</tr>
<tr>
<td>2A</td>
<td>15548.96</td>
<td>7338.38</td>
<td>8650.82</td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td>5747.61</td>
<td>7299.89</td>
<td>11206.57</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12 hpi replicate 2</th>
<th>Protein</th>
<th>WT</th>
<th>SS</th>
<th>M3</th>
<th>Carried forward for calculations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>945.31</td>
<td>2963.3</td>
<td>7539.19</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VP0</td>
<td>2397.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2C</td>
<td>1470.08</td>
<td>5552.5</td>
<td>10498.07</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VP1</td>
<td>2248.87</td>
<td>1656.4</td>
<td>3280.92</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>3C</td>
<td>963.09</td>
<td>1771.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP3</td>
<td>2434.33</td>
<td>1869.35</td>
<td>3682.06</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td>1847.17</td>
<td>1285.68</td>
<td>1844.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td>476.88</td>
<td>1350.26</td>
<td>2322.62</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12 hpi replicate 3</th>
<th>Protein</th>
<th>WT</th>
<th>SS</th>
<th>M3</th>
<th>Carried forward for calculations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D</td>
<td>664.88</td>
<td>7645.55</td>
<td>8779.04</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VP0</td>
<td>1760.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2C</td>
<td>934.12</td>
<td>7949.59</td>
<td>10576.14</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VP1</td>
<td>1515.47</td>
<td>2134.25</td>
<td>2391.4</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>3C</td>
<td>952.4</td>
<td>621.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP3</td>
<td>1619.9</td>
<td>2328.56</td>
<td>2969.14</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


43. Tsuchihashi, Z. and Brown, P.O. (1992) Sequence requirements for efficient translational frameshifting in the Escherichia coli dnaX gene and the role of an unstable interaction between tRNA(Lys) and an AAG lysine codon. Genes & development, 6, 511-519.


7 Manipulation of the unfolded protein response: A pharmacological strategy against coronavirus infection

The manuscript published in PLoS Pathogens is appended, beginning on the next page (with size scaled down to facilitate binding, and page numbers given with respect to the manuscript).
Manipulation of the unfolded protein response: A pharmacological strategy against coronavirus infection

Liliana Echavarría-Consuegra, Georgia M. Cook, Idoia Busnadiego, Charlotte Lefèvre, Sarah Keep, Katherine Brown, Nicole Doyle, Giulia Dowgier, Krzysztof Franaszek, Nathan A. Moore, Stuart G. Siddell, Erica Bickerton, Benjamin G. Hale, Andrew E. Firth, Ian Brierley, Nerea Irigoyen*

1 Division of Virology, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, United Kingdom, 2 Institute of Medical Virology, University of Zurich, Zurich, Switzerland, 3 The Pirbright Institute, Woking, Surrey, United Kingdom, 4 Department of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom

☯ These authors contributed equally to this work.
¤ Current address: Basingstoke and North Hampshire Hospital, Hampshire Hospitals, NHS Foundation Trust, Basingstoke, United Kingdom.
*

Abstract

Coronavirus infection induces the unfolded protein response (UPR), a cellular signalling pathway composed of three branches, triggered by unfolded proteins in the endoplasmic reticulum (ER) due to high ER load. We have used RNA sequencing and ribosome profiling to investigate holistically the transcriptional and translational response to cellular infection by murine hepatitis virus (MHV), often used as a model for the Betacoronavirus genus to which the recently emerged SARS-CoV-2 also belongs. We found the UPR to be amongst the most significantly up-regulated pathways in response to MHV infection. To confirm and extend these observations, we show experimentally the induction of all three branches of the UPR in both MHV- and SARS-CoV-2-infected cells. Over-expression of the SARS-CoV-2 ORF8 or S proteins alone is itself sufficient to induce the UPR. Remarkably, pharmacological inhibition of the UPR greatly reduced the replication of both MHV and SARS-CoV-2, revealing the importance of this pathway for successful coronavirus replication. This was particularly striking when both IRE1α and ATF6 branches of the UPR were inhibited, reducing SARS-CoV-2 virion release (~1,000-fold). Together, these data highlight the UPR as a promising antiviral target to combat coronavirus infection.

Author summary

SARS-CoV-2 is the novel coronavirus responsible for the COVID-19 pandemic which has resulted in over 150 million cases since the end of 2019. Most people infected with the virus will experience mild to moderate respiratory illness and recover without any special treatment. However, older people, and those with underlying medical problems like...
chronic respiratory disease are more likely to develop a serious illness. So far, more than 3 million people have died of COVID-19. Unfortunately, there is no specific medication for this viral disease.

In order to produce viral proteins and to replicate their genetic information, all coronaviruses use a cellular structure known as the endoplasmic reticulum or ER. However, the massive production and modification of viral proteins stresses the ER and this activates a compensatory cellular response that tries to reduce ER protein levels. This is termed the unfolded protein response or UPR. We believe that coronaviruses take advantage of the activation of the UPR to enhance their replication.

The UPR is also activated in some types of cancer and neurodegenerative disorders and UPR inhibitor drugs have been developed to tackle these diseases. Here, we show also that these compounds can significantly reduce SARS-CoV-2 replication in human lung cells.

Introduction

The Coronaviridae are a family of enveloped viruses with positive-sense, non-segmented, single-stranded RNA genomes. Coronavirus (CoVs) cause a broad range of diseases in animals and humans. SARS-CoV, MERS-CoV and SARS-CoV-2, members of the genus Beta-coronavirus, are three CoVs of particular medical importance due to high mortality rates and pandemic capacity [1–3]. SARS-CoV-2 is the causative agent of the current COVID-19 pandemic, which has resulted in over 150 million cases and more than 3 million deaths since the end of 2019. Although up to 15% of the cases develop a severe pathology [4,5], no specific therapeutic treatment for COVID-19 has been approved to date, highlighting the urgent need to identify new antiviral strategies to combat SARS-CoV-2, besides future CoV zoonoses.

During CoV replication, the massive production and modification of viral proteins, as well as virion budding-related endoplasmic reticulum (ER) membrane depletion, can lead to over-loading of the folding capacity of the ER and consequently, ER stress [6]. This activates the unfolded protein response (UPR) which is controlled by three ER-resident transmembrane sensors: inositol-requiring enzyme-1 α (IRE1α), activating transcription factor-6 (ATF6), and PKR-like ER kinase (PERK), each triggering a different branch of the UPR (Fig 1A). Activation of these pathways leads to decreased protein synthesis and increased ER folding capacity, returning the cell to homeostasis [7].

Here, we characterise global changes in the host translome and transcriptome during murine coronavirus (MHV) infection using RNA sequencing (RNASeq) and ribosome profiling (RiboSeq). MHV is a member of the Betacoronavirus genus and is widely used as a model to study the replication and biology of members of the genus. In this analysis, the UPR is one of the most significantly enriched pathways. We further confirm the activation of all three branches of the UPR in MHV-infected cells. Extending our investigation to SARS-CoV-2, we find that infection with this novel CoV also activating all three UPR pathways. Moreover, we demonstrate that individual over-expression of SARS-CoV-2 ORF8 and 5 proteins is sufficient to induce the UPR. Remarkably, pharmacological inhibition of the UPR had a dramatic negative effect on MHV and SARS-CoV-2 replication, suggesting that CoVs may subvert the UPR to their own advantage. These results reveal that pharmacological manipulation of the UPR can be used as a therapeutic strategy against coronavirus infection.
Fig 1. Ribosome profiling reveals the unfolded protein response as a key pathway in the host response to MHV-A59 infection. (A) Schematic of the three branches of the UPR (IRE1α, ATF6, and PERK). ERAD = ER-associated protein degradation. (B) Top five most significantly enriched Reactome pathways [1] associated with the lists of transcriptionally up-regulated genes (orange triangles pointing upwards) and transcriptionally down-regulated genes (blue triangles pointing downwards), plotted according to the false discovery rate (FDR)-corrected p-value of the enrichment. Full results, including pathway IDs, are in S3 Table. (C) Volcano plot showing the relative change in abundance of cellular transcripts and the FDR-corrected p-value for differential expression between the mock and infected samples (n = 2 biological replicates). Grey vertical lines indicate a transcript abundance fold change of 2. Genes which have fold changes
greater than this threshold and a \( p \leq 0.05 \) value of less than 0.05 are considered significantly differentially expressed and coloured orange if up-regulated and blue if down-regulated. Selected genes are annotated in red. (D) Volcano plot showing the relative change in translation efficiency of cellular transcripts, and the FDR-corrected \( p \) value, between the mock and infected samples (\( n = 2 \) biological replicates). Colours and fold change and \( p \) value thresholds as in C. (E) Analysis of RPFs mapping to Atf4 (NCBI RefSeq mRNA NM_009716). Cells were infected with MHV-A59 or mock-infected and harvested at 5 h.p.i. (libraries from replicate 2) or 8 h.p.i. RPFs are plotted at the inferred position of the ribosomal P site and coloured according to phase (which position within the codon the 5' end of the read maps to: pink for 0, blue for +1, yellow for +2). The main ORF (0 frame) is shown at the top in pink, with start and stop codons in all three frames marked by green and red bars (respectively) in the three panels below. The two yellow rectangles in the +2 frame indicate the known Atf4 uORFs (the first of which is only three codons in length). Dotted lines serve as markers for the start and end of the features in their matching colour. Note that read densities are plotted as reads per million host-mRNA-mapping reads, and that bar widths were increased to 12 nt to aid visibility, and therefore overlap, and were plotted sequentially starting from the 5' end of the transcript. (F) Plot of log(fold changes) of translation efficiency (TE) vs transcript abundance for all genes included in both analyses. Grey lines indicate fold changes of 2. Fold changes are plotted without filtering for significant \( p \) values. Selected genes are marked: genes up-regulated predominantly by one of either transcription or TE are marked in orange (upper middle and right middle sections), and Chop, which is up-regulated at the level of both transcription and TE, is marked in green (top right section).

https://doi.org/10.1371/journal.ppat.1009644.g001

Results

Differential gene expression analysis of murine cells infected with MHV-A59

To survey genome-wide changes in host transcription and translation during CoV infection, murine 17 clone 1 cells (17 Cl-1) were infected with recombinant MHV-A59 at a multiplicity of infection (MOI) of 10, or mock-infected, in duplicate and harvested at 5 h post-infection (h.p.i.). Lysates were subjected to RNASeq and parallel Riboseq [8,9], which allows global monitoring of cellular translation by mapping the positions and abundance of translating ribosomes on the transcriptome with sub-codon precision. Quality control analysis confirmed the libraries were of high quality (S1 Fig and S1 Table).

To assess the effects of MHV infection on cellular transcript abundance, differential expression analysis was performed at 5 h.p.i. with DESeq2 [10] (Fig 1B and 1C and S2 and S3 Tables). At this timepoint, viral RNA synthesis approaches a maximum, but it precedes the onset of cytopathic effects such as syncytium formation [9]. Between infected and mock-infected conditions, genes with a fold change \( \geq 2 \) and a false discovery rate (FDR)-corrected \( p \) value of \( \leq 0.05 \) were considered to be significantly differentially transcribed (S2 Table). To determine the biological pathways involved in the response to infection, we carried out Reactome pathway enrichment analysis [11] on the lists of significantly differentially transcribed genes (Fig 1B and S3 Table). The most significantly enriched pathway associated with transcriptionally up-regulated genes was “Unfolded Protein Response” (R-HSA-381119, \( p = 1.1 \times 10^{-15} \)), and pathways denoting the three branches of the UPR (ATF6 branch: R-HSA-381183, PERK branch: R-HSA-380994, IRE1alpha branch: R-HSA-381070) were also significantly enriched (S3 Table). Consistent with this, gene ontology (GO) term enrichment analysis of the transcriptionally up-regulated gene list revealed that UPR-related GO terms, such as “response to unfolded protein” (GO:0006936), were significantly enriched (S3 Table). Many of the enriched pathways and GO terms associated with transcriptionally down-regulated genes are related to protein synthesis, again highlighting this as a key theme of the host response.

We provide the full database of differentially expressed genes and enriched pathways/GO terms for further exploration (S2 and S3 Tables) but in this manuscript we will focus predominantly on the UPR, which has been recognised as a host response to several CoVs due to the extensive dependence of CoV replication on the ER [6]. Accordingly, some of the most differentially transcribed genes are involved in the UPR, such as Herp (also known as Herpald), Chac1, Bip (also known as Grp78 or Hspa5), Chop (also known as Ddit3 or Gadd153) and Grp94 (also known as Hspa9) (Fig 1C).

To evaluate differences at the level of translation, we calculated relative translation efficiencies (TE; defined herein as the ratio of ribosome-protected-fragment [RPF] to total RNA
density in the CDS of a given gene) at 5 h p.i. using Xtail [12], applying the same fold change and p-value thresholds as for the transcription analysis. As shown in Fig 1D, several of the translationally up-regulated genes encode key proteins involved in activation of the UPR, for example ATF4, ATF5 and CHOP, which are effecter transcription factors [13–18], GADD34 (also known as MYD116/PPP1R15A), a protein that acts as a negative regulator to diminish eIF2α phosphorylation leading to the induction of its target genes.

Given that UPR activation can lead to eIF2α phosphorylation and host translational shut-off, we investigated whether the list of mRNAs found to be preferentially translated during MHV infection was enriched for genes resistant to translational repression by phosphorylated eIF2α (p-eIF2α) (Materials and Methods and S4 Table). We found a 9.15-fold enrichment of p-eIF2α resistant genes (p = 1.42×10^{-19}, Fisher Exact Test). Resistance to the effects of p-eIF2α has been linked to the presence of efficiently translated upstream open reading frames (uORFs) in the 5' UTR [13–18,21]. To investigate this in our dataset, we analysed ribosome occupancy of the main ORF compared to the uORFs on Atf4, a well-studied example [14] (Fig 1E). Translation of the short (three codon) uORF1 was observed under all conditions. In mock-infected samples, uORF2 was efficiently translated, largely precluding translation of the main ORF (pink). In contrast, in MHV-infected cells, a large proportion of ribosomes scan past uORF2 to translate the main ORF. This is consistent with previous studies on Atf4 translation under conditions of eIF2α phosphorylation, in which many ribosomes cannot reassemble a competent initiation complex before reaching uORF2 [13,14]. This facilitates increased production of ATF4 even when translation of most mRNAs is inhibited.

Comparison of the fold changes at the transcriptional and translational level for individual cellular mRNAs provides insight into the overall effect on gene expression (Fig 1F). Genes regulated in opposing directions transcriptionally and translationally likely result in a small overall change in expression, whereas genes regulated only in one direction likely result in a greater overall change. Many UPR genes fall into the latter category (orange points, top-centre and right-centre), reflecting published knowledge about the induction of these genes specifically at the transcriptional [22–25] or translational level [14–18,21]. Chop (green point, upper-right) is a rare example of a gene that is significantly up-regulated both transcriptionally and translationally during MHV infection. This reflects the fact that it is transcriptionally induced by ATF4 during UPR activation and translationally p-eIF2α-resistant [26,27].

Together, the ribosome profiling results highlight the UPR as a key pathway in the host response to MHV infection, with many of the greatest expression changes observed for UPR-related genes.

MHV infection and activation of the unfolded protein response

To further explore the extent of UPR activation during MHV infection, we investigated each of the three branches individually (Fig 1A), building on the work of several groups [28–32].

Monitoring the PERK-eIF2α-ATF4 branch. Upon ER stress, PERK oligomersises and auto-phosphorylates [33]. Activated PERK phosphorylates the α-subunit of eIF2 which in turn impairs recycling of inactive eIF2-GDP to active eIF2-GTP, resulting in a general shutdown of protein synthesis [34]. However, translation of ATF4 is increased in this situation [13,14,35] leading to the induction of its target genes Chop and Gadd34 (Fig 1A, right). To assay PERK activation, we monitored expression of PERK, CHOP, ATF4 and p-eIF2α, by qRT-PCR and western blotting. 17 CI-1 cells were infected with MHV-A59 or incubated with tunicamycin and harvested at 2.5, 5, 8 and 10 h. Tunicamycin, used as a positive control, is a pharmacological inducer of ER stress which activates all UPR signalling pathways. From 5 h p.i. onwards in MHV-infected cells, and at all timepoints in tunicamycin-treated cells, ATF4 and p-eIF2α
were detected and multiple bands were observed for PERK (Fig 2A) corresponding to the auto-phosphorylated species, indicative of activation of this kinase upon ER stress. In addition, as shown in Fig 2B, Chop and Gadd34 mRNA levels in MHV-infected cells (blue squares) increased from 2.5 to 8 h p.i., similarly to tunicamycin-treated cells (red circles), indicating their induction by the transcription factor ATF4. At 10 h p.i., the level of p-eIF2α decreased slightly compared to its peak at 8 h p.i., consistent with GADD34 stimulating its dephosphorylation.

Virus-induced inhibition of translation as a consequence of eIF2α phosphorylation was confirmed by analytical polysome profiling in 17 CI-1 cells (Fig 2C, upper panel), revealing the accumulation of monososomes (80S) in MHV-infected cells at 5 h p.i. In higher salt profiles (400 mM KCl; Fig 2C, lower panel), where 80S ribosomes lacking mRNA dissociate into constituent subunits, a large reduction in 80S ribosomes was seen. These data are highly consistent with inhibition of translation initiation and show that the vast majority of 80S ribosomes accumulating at this time point are not mRNA-associated. These data support the view that MHV infection leads to translational shut-off via inhibited initiation, consistent with the effects of eIF2α phosphorylation.

**Monitoring the IRE1α-XBP1 branch.** Activated IRE1α (Fig 1A, left) removes a 26-nt intron from unspliced Xbp1 (Xbp1-u) mRNA leading to a translational reading frame shift and a longer protein [25,36]. The product of spliced Xbp1 mRNA (Xbp1-s) is an active transcription factor that up-regulates the expression of ER-associated degradation (ERAD) components and ER chaperones. To study this, we analysed Xbp1-u and Xbp1-s mRNAs by reverse transcriptase PCR (RT-PCR), using specific primers flanking the splice site (Fig 2D). At all time-points, Xbp1-u was the predominant form in mock-infected cells whereas Xbp1-s was the major species in tunicamycin-treated cells. In virus-infected cells, Xbp1-s became predominant at 5 h p.i. This was corroborated at the translational level in the ribosome profiling datasets, in which infected samples showed increased translation of the extended ORF (yellow) generated by splicing (S2 Fig). An increase in active XBP1-s transcription factor was further supported by the finding that two of its target genes are transcriptionally up-regulated in infected cells (ERδtα–2.44-fold increase $p = 6.63 \times 10^{-5}$; and P58ipk–1.94-fold increase $p = 3.97 \times 10^{-11}$) (S2 Table). These data indicate that the IRE1α-Xbp1 pathway is activated by MHV infection.

**Monitoring the ATF6 branch.** The ATF6 branch is activated when ATF6 translocates from the ER to the Golgi apparatus, where it is cleaved [37]. After cleavage, the amino-terminus of ATF6 (ATF6-Nt) translocates to the nucleus to up-regulate ER chaperones (Fig 1A, middle). To monitor this pathway, 17 CI-1 cells were infected with MHV-A59 or incubated with tunicamycin and analysed by western blotting (to detect ATF6 cleavage) or by immunofluorescence (to detect ATF6 nuclear translocation) (S3A, S3B and S3C Fig). However, we were unable to detect the trimmed version of ATF6 nor a clear nuclear translocation. As ATF6-Nt was also not visible in the positive control tunicamycin-treated cells, it is likely that the antibodies used do not efficiently recognise mouse ATF6-Nt in this context.

As an alternative approach, we monitored the induction of Bip, Grp94 and Calreticulin, transcriptionally up-regulated genes in the Reactome category “ATF6 (ATF6-alpha) activates chaperone genes” (S3 Table) and known to be induced by ATF6-Nt [38,39]. Bip mRNA or protein levels are often used as a proxy for activation of the ATF6 pathway; however, its transcription can eventually be regulated by other UPR factors such as XBP1 [40] and ATF4 [41], so it can also be used as a general readout of ER stress [28,38]. Cells were harvested at 2.5, 5 and 8 h p.i. and analysed by qRT-PCR (Fig 2E). An increase in Bip transcription was observed in tunicamycin-treated (red circles) and to a lesser extent in MHV-infected cells (blue squares) from 2.5 to 8 h p.i., whereas mock-infected cells (green triangles) showed no induction. Despite the transcriptional up-regulation and a noticeable increase in RiboSeq reads mapping...
Fig 2. MHV infection and activation of the unfolded protein response. 17 Cl-1 cells were incubated in the presence of tunicamycin (2 μg/ml) or infected with MHV-A59 (MOI 5) and harvested at 2.5, 5, 8 and 10 h.p.i. (A) Western blot analysis of ATF4, p-eIF2α, eIF2α, PERK and MHV N proteins. GAPDH and BiP were used as loading controls. Molecular masses (kDa) are indicated on the left and the p-eIF2α band is indicated by a red asterisk. Protein band quantifications for p-eIF2α, normalised by eIF2α and given relative to the timepoint-matched mock value, are provided below the immunoblot. (B) RT-qPCR of Chop and Gadd45 mRNA for three biological replicates of a timecourse of MHV infection or tunicamycin treatment. Data are normalised as in B. (C) RT-PCR analysis of Xbp1s and Xbp1u mRNAs. Rp19 RT-PCR product was used as a loading control. Molecular size markers (nt) are indicated on the left. Xbp1 splicing was quantified as the ratio Xbp1s/Xbp1u, and the extent of splicing relative to the timepoint-matched mock is shown below each lane. The band that migrates above Xbp1u is thought to represent a duplex of Xbp1s and Xbp1u, known as the “hybrid” band (Xbp1-h) [116]. (D) RT-qPCR of Bip, Calreticulin and Grp94 mRNA for three biological replicates of a timecourse of MHV infection or tunicamycin treatment. Data are normalised as in B. (D) Cell lysates were analysed by 12% SDS-PAGE and immunoblotted using anti-BiP and anti-N antibodies. GAPDH was used as a loading control. Protein band quantifications for BiP were normalised to GAPDH.

Immunoblots and agarose gels are representative of three biological replicates.

https://doi.org/10.1371/journal.ppat.1009644.g002

to BiP (S3D Fig), the protein was not detectable by western blot in MHV-infected cells (Fig 2F). It is not yet clear why this is the case, although down-regulation of BiP at the protein level has previously been observed during infection with other members of the order Nidovirales [32,42]. Alternatively, it is possible that BiP might accumulate to levels detectable by western blot at later timepoints, but the extensive cell death that occurs shortly after 10 h p.i. makes this difficult to test. Nevertheless, an increase in Calreticulin and Grp94 transcription (Fig 2E) was observed in tunicamycin-treated cells (red circles) and to a greater extent in MHV-infected cells (blue squares) especially at 8 h p.i. This indicates that the ATF6 pathway is highly up-regulated during MHV-infection. Together with our studies of PERK-eIF2α-ATF4 and IRE1α-XBP1 above, these data confirm that MHV infection induces all three branches of the UPR.

Effect of UPR inhibitors on MHV replication

Based on the strong UPR activation brought about by MHV infection, we hypothesised that pharmacological manipulation of this pathway could be used to modulate viral replication. First, we determined cell viability in 17 Cl-1 cells after drug treatment using CellTiter-Blue and trypan blue exclusion assays (S4 Fig). Subsequently, we evaluated the inhibitory effect of four different UPR inhibitors (UPRi) on each one of the UPR branches in cells infected with MHV for 8 h at MOI 5 (S5 Fig).

GSK-2606414 (henceforth referred to as PERKi) is a specific inhibitor of PERK [43,44]. As expected, PERKi treatment prevented autophosphorylation of PERK and reduced phosphorylation of its substrate, eIF2α (S5A Fig), effectively blocking this branch of the UPR. Pulse labeling of infected cells for one hour at 5 h p.i. revealed a modest increase of both viral and host protein synthesis, with no effect on mock-infected cells (S5B Fig). Analytical polysome profiling of MHV-infected cells treated with 5 μM PERKi for 5 h (S5C Fig) revealed a decrease in the accumulation of monosomes (80S) compared to MHV-infected cells at 5 h p.i. (Fig 2C, upper right panel), indicating a relief of translation inhibition.

Integrated stress response inhibitor (ISIRI) acts downstream of eIF2α in the PERK pathway by preventing p-eIF2α from binding and inhibiting eIF2B [45]. Therefore, eIF2B can recycle eIF2-GDP to active eIF2-GTP, and translation initiation can still occur, despite the levels of p-eIF2α remaining unchanged. Inhibition of the PERK pathway downstream of eIF2α is evident from the decrease in Chop transcription in MHV-infected cells treated with 2 μM ISIRI (S3D Fig).

STF-083010 (henceforth referred to as IREi) is a specific IRE1α endonuclease inhibitor that does not affect its kinase activity [46]. In MHV-infected cells treated with IREi at 60 μM (8 h p. i.), the protein was not detectable by western blot in MHV-infected cells (Fig 2F). It is not yet clear why this is the case, although down-regulation of BiP at the protein level has previously been observed during infection with other members of the order Nidovirales [32,42]. Alternatively, it is possible that BiP might accumulate to levels detectable by western blot at later timepoints, but the extensive cell death that occurs shortly after 10 h p.i. makes this difficult to test. Nevertheless, an increase in Calreticulin and Grp94 transcription (Fig 2E) was observed in tunicamycin-treated cells (red circles) and to a greater extent in MHV-infected cells (blue squares) especially at 8 h p.i. This indicates that the ATF6 pathway is highly up-regulated during MHV-infection. Together with our studies of PERK-eIF2α-ATF4 and IRE1α-XBP1 above, these data confirm that MHV infection induces all three branches of the UPR.
Fig

the unspliced form of Xbp1 was more prominent compared to the untreated MHV-infected cells, indicating a reduction in the endonuclease activity of this enzyme.

AEBSF, a serine protease inhibitor, prevents ER stress-induced cleavage of ATF6 resulting in inhibition of transcriptional induction of ATF6 target genes [47]. We investigated the induction of ATF6 target genes in MHV-infected cells treated with 100 μM AEBSF as previously described. As anticipated, Calreticulin and Grp94 transcription was greatly reduced in AEBSF-treated cells (S5F).

Having shown these compounds effectively inhibit the UPR in the context of infection, we moved on to assess whether this could lead to an inhibition of viral replication. Cells were infected with MHV at MOI 5 and treated with the UPRi. At 8 h p.i., tissue culture supernatant was harvested and released progeny quantified by plaque assay. We found modest but significant reductions in virus titres for all UPRi treatments in comparison to control cells, with fold reductions of between ~two-fold (IREi) and ~six-fold (ISRIB) (Fig 3A). This supports our hypothesis that modulation of the UPR can have antiviral effects.

Next, we investigated whether using the UPRi in combination would have a cumulative effect on virus release. We confirmed that combination treatment conditions led to reversal of the three branches of the UPR, assayed as described above (S6 Fig). Fig 3B displays virus titres from infected cells (8 h p.i.) at MOI 1 (blue) and MOI 5 (red), treated with different UPRi combinations. Reductions in virus titre ranged from ~four-fold, in cells incubated with PERKi and ISRIB (both targeting the PERK-eIF2α-ATF4 branch), to ~40- and ~100-fold (MOI 5 and 1 respectively), in cells treated with IREi and AEBSF (targeting the IRE1α and the ATF6 pathways). This was confirmed by western blotting, demonstrating a striking decrease in N protein levels for treatment combinations where virus titres were lowest (Fig 3C). We note that some treatment conditions largely involving ISRIB or PERKi, led to an increase in N protein levels compared to mock-treated cells. We hypothesise that this is due to alleviation of p-eIF2α-mediated translation inhibition, facilitating increased translation of viral proteins, although this does not result in increased virus release. In addition, cell monolayers infected with MHV in the presence of IREi and AEBSF showed delayed cytopathic effect, as indicated by reduced syncytium formation, likely due to lower virus production (Fig 3D).

Mechanistic analysis of the UPR activation by SARS-CoV-2 proteins

Having established the use of UPRi as a potential antiviral strategy, we moved on to study UPR activation by SARS-CoV-2, initially assaying the cellular response to individual virus proteins in the context of transfection.

UPR activation associated with individual proteins from other CoVs in the Betacoronavirus genus has been investigated previously [30,48–50]. The response to the spike (S) protein has been the most thoroughly characterised, with SARS-CoV and human CoV HKU1 S activating the PERK-eIF2α-ATF4 branch [48,51] and MHV S activating the IRE1α-XBP1 pathway [30]. Additionally, SARS-CoV ORF3a, ORF6, ORF7a, ORF8ab or ORF8b proteins were able to activate the UPR [48–50,52,53], while the SARS-CoV E protein had an inhibitory effect [54]. Of these UPR-stimulatory accessory proteins, the response was characterised in detail for ORF3a, which activates the PERK-eIF2α-ATF4 branch [49], and ORF8ab, which activates the ATF6 branch [50]. To investigate potential UPR-stimulatory proteins in SARS-CoV-2, we selected the three proteins whose activation of the UPR had been best-characterised in other members of the genus: S, ORF3a and ORF8. These proteins are divergent from their SARS-CoV counterparts (76%, 72% and 26% amino acid identities, respectively), so there may be differences in UPR activation that would add to our understanding of the relationship between CoVs and the UPR. To investigate the response to these proteins, we expressed C-terminally-tagged S
(S-HA), ORF3a (ORF3a-FLAG) and ORF8 (ORF8-FLAG) proteins in human embryonic kidney cells (HEK-293T cells). N, a structural protein which is not documented as activating the UPR, was over-expressed as a negative control (N-FLAG). 

https://doi.org/10.1371/journal.ppat.1009644.g003

Fig 3. Effect of UPR inhibitors on MHV replication. (A) MHV-infected cells (MOI 5) were treated with UPR inhibitors (5 μM PERKi, 2 μM ISRIB, 60 μM IREi, or 100 μM AEBSF). The inhibitors were added to the cells immediately after the virus adsorption period and maintained in the medium until cells were harvested 8 h later. Plaque assays were performed with serial dilutions of the supernatant containing released virions from 17 C3-1 cells infected with MHV-A59 in the presence or absence of the UPR inhibitors. Values show the mean averages of the titration of three biological replicates. Error bars represent standard errors. (B-D) MHV-infected cells (MOI 1 and MOI 5) were treated with dual combinations of the UPR inhibitors. The inhibitors were added to the cells immediately after the virus adsorption period and maintained in the medium until cells were harvested 8 h later. (B) Released virions were quantified as described in A. (C) Western blot analysis of MHV N protein. GAPDH was used as a loading control. Protein band quantifications for N protein, normalised by GAPDH and given relative to untreated/infected cells, are provided below. Immunoblots are representative of three biological replicates. (D) Representative images of mock- and MHV-infected cells at 8 h p.i. under no-drug or IREi 60 μM/AEBSF 100μM treatment conditions. All t-tests were two-tailed and did not assume equal variance for the two populations being compared (* p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001). All p-values are from comparisons with the respective untreated control at the same MOI.
ER stress, assessed by the induction of HERP and BiP, was induced by SARS-CoV-2 S but not N (S7A Fig). The PERK-eIF2α-ATF4 branch was activated from 24 h.p.t onwards, as indicated by the phosphorylation of eIF2α and the detection of ATF4 (S7A Fig), although phosphorylation of PERK was not clearly evident. The activation of this pathway was further confirmed by the increase in CHOP transcription compared to mock-transfected cells (S7B Fig). The amino terminus of ATF6 (ATF6-Nt) was detected in S-transfected cells from 24 h.p.t onwards (S7A Fig), indicating activation of the ATF6 branch. Activation of the IRE1α pathway is also evident from an increase in the spliced form of XBP1 in S protein-transfected cells (S7A Fig). Contrary to previous findings for SARS-CoV, this indicates that the expression of the SARS-CoV-2 S protein is sufficient to induce all three major signalling pathways of the UPR.

In the case of SARS-CoV-2 ORF8 transfection, IRE1α-XBP1 and ATF6 were the main pathways induced (S7C Fig), again contrasting with findings for SARS-CoV [50]. Although a slight activation of ATF4 was observed in ORF8-transfected cells at 36 h.p.t. (S7C Fig), this was not accompanied by PERK nor eIF2α phosphorylation, and induction of CHOP transcription was lower than in S protein-transfected cells (S7B Fig). SARS-CoV-2 ORF3a transfection did not induce any of the branches of the UPR (S7C Fig).

We then asked whether the UPR induction caused by SARS-CoV-2 S and ORF8 overexpression could be reversed by treatment with UPRi. This was assessed for each inhibitor individually (S8 Fig). Additionally, we tested this using a combination treatment condition (Fig 4), for which we selected IRE1i/AEBSF as this gave the most promising reduction in viral titre during MHV infection (Fig 3B). Treatment of SARS-CoV-2 S- and ORF8-transfected cells with IRE1i/AEBSF reduced expression of HERP and BiP to levels comparable to mock-transfected cells (Fig 4A, 36 h.p.t.). This indicates the treatment successfully reversed the UPR activation by the two viral proteins. PERK pathway inhibition was evident in treated cells from the reduction in PERK and eIF2α phosphorylation (Fig 4A); however, ATF4 levels appeared to be slightly increased under these conditions, as was induction of its target gene CHOP (S8C Fig). ATF4 induction in the presence of IRE1i has been previously described [55]. Inhibition of the ATF6 and the IRE1α-XBP1 pathways was also evident, as very little ATF6-Nt and XBP1-s were present in IRE1i/AEBSF treated cells (Fig 4A and 4B).

In summary, over-expression of the S and the ORF8 proteins of SARS-CoV-2 is sufficient to activate the three branches of the UPR, and this can be reversed by UPRi treatment. It should be noted that, due to potential differences such as protein sub-cellular localisation, expression levels or available interaction partners, there may be differences in the UPR-stimulatory roles of these proteins in this system compared to infection. Future experiments using knockout mutant viruses could extend observations on the role of ORF8 to the context of infection, however S is essential for infectivity, making such experiments challenging.

**Induction of the UPR in SARS-CoV-2-infected cells**

We went on to study UPR activation in the context of SARS-CoV-2 infection. Vero CCL81 cells were infected at MOI 1 and harvested at 24 and 48 h.p.i, representing the exponential phase of viral replication without the cell viability being notably compromised [56]. Lysates were analysed as above. As shown in Fig 5A, the PERK-eIF2α-ATF4 branch was activated at 48 h.p.i as indicated by increased phosphorylation of PERK and eIF2α. This was further confirmed by the induction of CHOP in infected cells (S9A Fig). Detection of ATF6-Nt (Fig 5A) demonstrates that the ATF6 pathway is also activated during the course of infection. In addition, activation of the IRE1α pathway was evident from an increase in the spliced form of XBP1 in SARS-CoV-2-infected cells (Fig 5A). These findings were verified in a more
physiologically relevant context by assaying SARS-CoV-2-infected Calu3 cells (a human lung cell line) at 24 h p.i. (MOI 1), as a model for the primary site of SARS-CoV-2 infection [56,57] (Fig 5B). We conclude that SARS-CoV-2 infection induces all three branches of the UPR.

Fig 4. Mechanistic analysis of UPR activation by SARS-CoV-2 proteins. HEK-293T cells were transfected with plasmids encoding SARS-CoV-2 S (S-HA) or ORF8 (ORF8-FLAG), mock-transfected, or treated with tunicamycin (Tn). At 8 h p.t., cells were treated with 60 μM IREi and 100 μM AEBSF and then harvested at 24 and 36 h p.t. (A) Western blot analysis of ORF8-FLAG, S-HA, HERP, BiP, PERK, ATF4, p-eIF2α and ATF6 proteins. The specific p-eIF2α and ATF6-Nt bands are indicated with a red asterisk. Protein band quantifications for HERP, BiP, p-eIF2α and ATF6-Nt, normalised by eIF2α as a loading control and given relative to the mock, are provided below the respective immunoblots. (B) RT-PCR analysis of XBP1-u and XBP1-s mRNAs, performed as described in Fig 2D. Immunoblots and agarose gels are representative of three biological replicates. “h p.t.” = hours post-transfection.

https://doi.org/10.1371/journal.ppat.1009644.g004
Effect of the IREi/AEBSF combination on SARS-CoV-2 infection

Next, we investigated whether the previously described UPRi combinations could also be used as potential antiviral drugs against SARS-CoV-2. The gastrointestinal tract is known to be a site of SARS-CoV-2 infection in vivo [38] so we used Caco2 cells, human intestinal cells shown to be permissive for SARS-CoV-2 infection [56,57]. The UPRi compounds were very well-tolerated by Caco2 cells (S9B Fig). Cells were infected with SARS-CoV-2 at MOI 0.01 and treated with the different UPRi combinations. Supernatants were harvested at 48 h p.i. and released virions quantified by TCID50 assay (Fig 5C). Reductions in virus titre were observed and these were generally much greater than those seen for MHV (MOI 1 and 5, Fig 3B), with both the PERKi/IREi and IREi/AEBSF combinations reducing virus titres to below the limit of detection.

As the IREi/AEBSF combination had the greatest inhibitory activity against both MHV and SARS-CoV-2, we tested whether this combination could inhibit SARS-CoV-2 infection at a higher MOI in Vero CCL81 and Calu3 cells. The cytotoxicity profile of these compounds in both cell lines was assayed (S9C and S9D Fig). Cells were infected at MOI 1 or MOI 5 and virus titres assessed by plaque assays at 24 h p.i. Incubation of Vero cells with IREi/AEBSF led to a statistically significant (p = 0.0241 for MOI 1 and p = 0.0033 for MOI 5) ~100-fold reduction in virus titre (Fig 5D, left). In Calu3 cells, IREi/AEBSF treatment had an even greater antiviral effect, reducing released virions by ~1,000-fold (p = 0.0017) to at or around the limit of detection (Fig 5D, right).

Detailed analysis of the activation of the three UPR pathways under the IREi/AEBSF treatment condition was performed in SARS-CoV-2-infected Vero CCL81 cells at 24 and 48 h.p.i. (Figs 5A and S9A). Interestingly, in both SARS-CoV-2- and MHV-infected cells, the IREi/AEBSF combination was not only able to prevent activation of the IREi and ATF6 pathways, but also the PERK-eIF2α-ATF4 branch, as indicated by reduced phosphorylation of PERK and eIF2α (Figs 5A and S6) and reduced transcription of CHOP (S6C and S9A Figs). This may be due to the inhibition of viral replication leading to a reduced ER load, as opposed to specific inhibition of the PERK pathway. This is supported by the observation of a striking decrease in viral protein levels in infected cells treated with IREi/AEBSF (Figs 3C and 5A), consistent with reduced viral replication. This reversal of CoV-induced UPR activation by the UPRi suggests that the antiviral activity of these compounds can be attributed, at least in part, to specific inhibition of the UPR, a pathway which is evidently required for efficient viral replication.

In addition to its role in UPR inhibition, AEBSF is a relatively promiscuous arylsulfonylchloride serine protease inhibitor. It has been reported to inhibit TMPRSS2 [59,60], a host serine protease essential for SARS-CoV-2 cell entry [57]. To test whether AEBSF treatment inhibits SARS-CoV-2 cell entry, we transfected HEK-293T cells with TMPRSS2 and ACE2, the...
SARS-CoV-2 cell entry receptor [61] and incubated them with lentiviral particles pseudotyped with the SARS-CoV-2 S protein (S9E Fig). No significant inhibition of viral entry was observed upon treatment with 100μM AEBSF for 4 hours, suggesting that the antiviral activity of AEBSF is predominantly due to its inhibition of the UPR.

The effect of alternative UPR inhibitors on CoV infection

To confirm that the inhibition of virus production caused by IREi/AEBSF treatment is due to specific inhibition of the UPR, we employed two alternative compounds to target the same pathways. Ceapin-A7 selectively inhibits trafficking of ATF6 from the ER to the Golgi, thereby preventing its cleavage and activation [62–64]. KIRA8 (kinase-inhibiting RNase attenuator 8) specifically inhibits oligomerisation of IRE1α preventing activation of its RNase activity [65–67]. Ceapin-A7 at 15 μM and KIRA8 at 10 μM were very well-tolerated in 17 CI-1, Vero CCL81 and Calu3 cells (S10 Fig), with improved metabolic and viability profiles compared to IREi/AEBSF.

In the context of MHV infection, we verified that Ceapin-A7 and KIRA8 specifically inhibit their target branches of the UPR, both individually (S11A Fig), and when used in combination (S11B Fig). Further, both compounds significantly reduced the titre of virions released from MHV-infected 17 CI-1 cells at 8 h p.i. (S11C Fig), consistent with the observed reductions in N protein (S11A and S11B Fig) and viral RNA (vRNA) abundance (S11B Fig). These results confirm that specific inhibition of the ATF6 or IRE1α branch of the UPR is sufficient to inhibit MHV replication, with combination treatment producing the greatest effect.

Next, we tested Ceapin-A7 and KIRA8 in the context of SARS-CoV-2 infection. Vero cells were infected with SARS-CoV-2 at MOI 5, treated with the inhibitors individually or in combination, and harvested at 24 and 48 h p.i. (Fig 6). Successful inhibition of the UPR was confirmed (Fig 6A) and the effect on virion release was assayed (Fig 6B). Both Ceapin-A7 and KIRA8 significantly reduced the titre of released virions. KIRA8 caused a greater reduction than Ceapin-A7, while the combination of both inhibitors was the most effective, reducing released virions by ~60-fold and ~500-fold at 24 and 48 h p.i., respectively. Striking reductions in abundance of viral RNA (Fig 6C) and proteins (Fig 6A) upon Ceapin-A7/KIRA8 treatment provides further evidence of the strong inhibitory effect of these UPRi on SARS-CoV-2 replication. Ceapin-A7/KIRA8 treatment was ~4-fold more effective at reducing released virions than IREi/AEBSF at 24 h p.i. (p = 0.0039). At 48 h p.i., this increased to a ~135-fold difference in effectiveness (p < 0.0001). Further experiments in Calu3 cells (MOI 5) reveal that Ceapin-A7/KIRA8 treatment significantly reduced virion release at 24 h p.i., to the same extent as IREi/AEBSF treatment (S11D Fig). The similar, or even greater, reductions in virion release caused by Ceapin-A7/KIRA8 compared to the more promiscuous inhibitor combination, IREi/AEBSF, demonstrate that the antiviral activity of these compounds is due to specific inhibition of the UPR.

Taken together, these results show that highly specific inhibition of the ATF6 and IRE1α branches of the UPR significantly inhibits CoV replication, making the UPR a promising antiviral target.

Discussion

This study reveals that all three branches of the UPR are activated upon MHV and SARS-CoV-2 infection, and highlights this as a very prominent pathway in the host response. The UPR was the most significantly enriched Reactome pathway associated with genes transcriptionally up-regulated during MHV infection and, consistent with previous studies, we show activation of all three branches of the UPR by MHV [28,30]. Confirming the importance of
in SARS-CoV-2 infection, ER-related GO/KEGG terms are enriched in the differentially expressed genes lists of several proteomics/transcriptomics studies on SARS-CoV-2-infected cells [32, 68–71]. This is also a very prominent theme in proteomics studies identifying host interaction partners of SARS-CoV-2 proteins, in which ER proteins are reproducibly found [70, 72, 73]. In one such study, “response to endoplasmic reticulum stress” was the most highly enriched biological process GO annotation associated with the host interaction partners [73]. This suggests that SARS-CoV-2, like other CoVs [74–76], enacts a finely tuned modulation of the UPR that may involve direct interactions with its components. Despite this, the activation of the three branches of the UPR by SARS-CoV-2 has not been previously described, although it has been characterized for other CoVs [6, 32, 51, 74–77–80] including the closely related SARS-CoV [30, 48–50, 52–54, 81, 82]. Here we show that, like MHV, SARS-CoV-2 infection induces all three branches of the UPR, in contrast to results from SARS-CoV infection, in which only the PERK branch was activated [30, 54, 81].

Fig 6. Effect of specific inhibition of IRE1α and ATF6 pathways on SARS-CoV-2 replication. Vero CCL81 cells were incubated in the presence of tunicamycin (2 μg/ml) or infected with SARS-CoV-2 (MOI 5) and treated with 15 μM Caspin-A7 and 10 μM KIRA8 as individual treatments or in combination. The inhibitors were added to the cells immediately after the virus adsorption period and maintained in the medium for 24 or 48 h. (A) Western blot analysis (upper) of SARS-CoV-2 S, PERK, ATF4, p-eIF2α and ATF6. The specific p-eIF2α and ATF6-Nt bands are indicated by red asterisks. Protein band quantifications, normalised by eIF2α as a loading control and given relative to the mock, are provided below the respective immunoblots. RT-PCR analysis of XBP1-u and XBP1-s mRNAs (lower), performed as described in Fig 2D. Immunoblots and agarose gels are representative of three biological replicates. (B) Plaque assays were performed with serial dilutions of the supernatant containing released virions at 24 and 48 h p.i. Values show the mean averages of the titration of four biological replicates. Error bars represent standard errors. All p-values are from comparisons with the respective untreated control, with *p < 0.05, **p < 0.01, ***p < 0.001 and ****p < 0.0001. (C) RT-qPCR of vRNA from two biological replicates of Vero CCL81 cells infected and treated as described above. Data are normalised as described in Fig 2B.

https://doi.org/10.1371/journal.ppat.1009644.g006
Over-expression of the individual SARS-CoV-2 S or ORF8 proteins initiates UPR signalling. S protein was found to induce all three branches of the UPR in contrast to the counterpart protein of SARS-CoV, which appears to induce exclusively the PERK pathway [48]. Similarly, we identify ORF8 of SARS-CoV-2 as an inducer of both the IRE1α and ATF6 branches of the UPR, whereas the SARS-CoV equivalent has been shown to activate only ATF6 [50]. These differences can partly be explained by sequence divergence between the two viruses [83]. SARS--CoV--2 ORF8 maintains only 26% amino acid identity with SARS-CoV ORF8ab [84], so the two proteins may have very different relationships with the UPR. For example, SARS-CoV-2 ORF8 lacks the VLVVL motif that causes SARS-CoV ORF8 (specifically ORF8b) to aggregate and trigger intracellular stress pathways [53]. SARS-CoV ORF8ab was shown to mediate activation of the ATF6 pathway through a direct interaction with the ATF6 ER-lumenal domain [50], although it is uncertain whether the corresponding interaction occurs with SARS--CoV--2 ORF8. Recent proteomics-based interactome studies have identified interactions between SARS-CoV-2 ORF8 and several ER quality control proteins [70,72], which could contribute to the ORF8-induced UPR induction observed in our study. Alterations to this key UPR modulator have important ramifications: mutation or deletion of ORF8 in naturally occurring strains of SARS-CoV and SARS-CoV-2 correlate with milder disease and, in the latter case, lower incidence of hypoxia [85–88].

Here, we also demonstrate the importance of UPR activation to CoV infection by showing that pharmacological inhibition of the UPR leads to significant reductions in titres of virions released from MHV- and SARS-CoV-2-infected cells. Simultaneous inhibition of the IRE1α and ATF6 pathways was particularly effective, reducing virus titres by up to ~500–1,000-fold. Some of the compounds we used to target these pathways (i.e. STF-083010 and KIRA8 against IRE1α and AEBSF against ATF6) have been extensively used in preclinical studies for neurodegenerative diseases, autoimmunity diabetes, cancer and pulmonary fibrosis [46,65,89–93]. Thus, simultaneous inhibition of IRE1α and ATF6 pathways using these drugs represents a promising antiviral strategy that could rapidly progress into a clinical trial.

To date, the development of antivirals against SARS-CoV-2 has focused on drugs targeting virus replication, such as remdesivir. However, these antiviral therapies do not take into account that the pathophysiology associated with COVID-19 is mostly related to an aberrant cellular response. In some clinical manifestations of COVID-19, an exacerbated UPR could play a key role [94–96]. For example, activation of ER stress and the UPR is one of the major triggers of endothelial dysfunction [97,98], which is associated with acute respiratory distress syndrome (ARDS) [99], a diffuse inflammatory lung injury present in 20–67% of hospitalised patients [100,101]. Other clinical manifestations of COVID-19 such as thromboembolism, cerebro- and cardiovascular diseases and neurological complications, are also associated with endothelial dysfunction [102]. Furthermore, a recognised sequel of COVID-19 is pulmonary fibrosis [103], which can develop in up to 17% of COVID-19 patients [104]. Pulmonary fibrosis is a severe form of interstitial lung disease characterised by progressive dyspnea, hypoxemia, and respiratory failure due to the presence of patchy areas of fibrotic tissue. ER stress and UPR activation are known to be involved in the development and progression of this fibrotic disease [105]. This suggests that UPR activation in response to SARS-CoV-2 infection contributes to the lung pathophysiology associated with COVID-19. Therefore, the UPR inhibitors used in this study could have a dual therapeutic effect, not only contributing to the reduction of viral burden in patients, but also diminishing the pathophysiology associated with COVID-19. In addition, the idea of targeting an exaggerated cellular response instead of the virus itself substantially reduces the chances of generating virus escape mutants.
Materials and methods

Cells and viruses

Murine 17 clone 1 (17 Cl-1), Calu3 (ATCC, HTB-55, a kind gift from Prof Frank Kirchhoff, Institute of Molecular Virology, Ulm University Medical Center) and Vero (ATCC, CCL81) cells were maintained in Dulbecco’s modification of Eagle’s medium supplemented with 10% (vol/vol) fetal calf serum (FCS). HEK-293T cells (ATCC, CRL-11268) were cultured in DMEM supplemented with 5% FCS. Caco2 cells were a kind gift from Dr Valeria Lulla and were maintained in DMEM supplemented with 20% FCS. All cell lines were cultured in medium containing 100 U/ml penicillin, 100 μg/ml streptomycin and 1 mM L-glutamine. Cells were incubated at 37˚C in the presence of 5% CO₂.

Recombinant MHV strain A59 (MHV-A59) was derived as described previously [106]. Upon reaching 70–80% confluence, 17 Cl-1 cells were infected with MHV-A59 at MOI 5 as described [9]. Vero CCL81 and Calu3 cells were infected with SARS-CoV-2 (SARS-CoV-2/human/Switzerland/ZH-UZH-IMV5/2020) at two MOIs (1 and 5) for 24 or 48 h as previously described [107,108]. Caco2 cells were infected with SARS-CoV-2 (isolate hCoV-19/Edinburgh/2/2020, a kind gift from Dr Christine Tait-Burkhard and Dr Juergen Haas) at MOI 0.01 and incubated for 48 h in MEM containing 1% L-glutamine, 1% non-essential amino acids, 1% penicillin/streptomycin and supplemented with 2% FBS.

Ribosomal profiling and RNASeq data

17 Cl-1 cells were grown on 100-mm dishes to 90% confluence and infected with MHV-A59 at MOI 10. At the indicated time-points, cells were rinsed with 5 ml of ice-cold PBS, flash frozen in a dry ice/ethanol bath and lysed with 400 μl of lysis buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM MgCl₂, 1 mM DTT, 1% Triton X-100, 100 μg/ml cycloheximide and 25 U/ml TURBO DNase (Life Technologies)). The cells were scraped extensively to ensure lysis, collected and triturated ten times with a 26-G needle. Cell lysates were clarified by centrifugation at 13,000 g for 20 min at 4˚C. Lysates were subjected to RiboSeq and RNASeq based on previously reported protocols [9,109]. Ribosomal RNA was removed using Ribo-Zero Gold rRNA removal kit (Illumina) and library amplicons were constructed using a small RNA cloning strategy adapted to Illumina smallRNA v2 to allow multiplexing. Amplicon libraries were deep sequenced using an Illumina NextSeq500 platform. Due to the very large amounts of vRNA produced during infection, mock samples were processed separately from infected samples to avoid contamination. RiboSeq and RNASeq sequencing data have been deposited in the ArrayExpress database under the accession numbers E-MTAB-8650 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8650/) and E-MTAB-8651 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8651/).

Computational analysis of RiboSeq and RNASeq data

Reads were trimmed for adaptor sequences, filtered for length ≥ 25 nt, and reads mapping to Mus musculus rRNA (downloaded from the SILVA database [110] or MHV-A59 viral RNA (GenBank accession AT700211.1) (with up to 2 mismatches) removed, as previously described [9]. The remaining reads were aligned directly to the mouse genome (FASTA and GTF genome release M20, GRCh38, primary assembly) (with up to 2 mismatches) using STAR (parameters:--outFilterIntronMotifs RemoveNoncanonicalUnannotated --outMultimapper-Order Random) [111]. Reads on protein-coding genes were tabulated using htseq-count (version 0.9.1), covering the whole gene for differential transcription analysis (parameters: -a 0 -m union -s yes -t gene) and just the CDS for the translation efficiency analysis (parameters: -a 0
Differential transcription analysis was performed using DESeq2 (version 1.18.1) \[10\] and translation efficiency analysis with Xtail (version 1.1.5) \[12\]. For each analysis, low count genes (with fewer than ten counts from all samples combined) were discarded, following which read counts were normalised by the total number of reads mapping to host mRNA for that library, using standard DESeq2 normalisation. This minimises the effect of the large amount of vRNA present in infected samples. Shrinkage of the transcriptional fold changes to reduce noise in lowly-expressed genes was applied using lfcShrink (parameter: type = ’normal’).

A given gene was considered to be differentially expressed if the FDR-corrected \(p\) value was less than 0.05 and the fold change between the means of infected and mock replicates was greater than two. Volcano plots and transcription versus TE comparison plots were generated using R and FDR-corrected \(p\) values and log\(_2\) (fold change) values from the DESeq2 and Xtail analyses. All reported \(p\) values are corrected for multiple testing by the Benjamini-Hochberg method. Fold changes plotted in the transcription vs TE comparison are not filtered for significant \(p\) values before plotting.

To plot RNASeq and RPF profiles for specific transcripts, reads were mapped to the specified transcript from the NCBI genome assembly using bowtie \[113\] allowing two mismatches (parameters: -v 2,--best). Coordinates for known uORFs were taken from the literature \[13,14,25\] and the positions of start and stop codons in all frames determined. Read density (normalised by total reads mapping to host mRNA for each library, to give reads per million mapped reads) was calculated at each nucleotide on the transcript and plotted, coloured according to phase. Read positions were offset by +12 nt so that plotted data represent the inferred position of the ribosomal P site. Bar widths were increased to 12 nt (Fig 1E) or 4 nt (S2 Fig) to aid visibility and were plotted sequentially starting from the 5’ end of the transcript.

Gene ontology and Reactome pathway enrichment analyses

Lists of gene IDs of significantly differentially expressed genes (S2 Table) were used for GO term enrichment analysis by the PANTHER web server under the default conditions (release 20190606, GO database released 2019-02-02) \[114\], against a background list of all the genes that passed the threshold for inclusion in that expression analysis. For Reactome pathway enrichment (version 69) \[11\], the same differentially expressed gene lists were converted to their human orthologues and analysed, both using the reactome.org web server, to determine which pathways are significantly over-represented (FDR-corrected \(p\) value \(\leq 0.05\)).

Enrichment analysis for eIF2\(\alpha\)-phosphorylation-resistant genes

Resistance to translational repression by p-eIF2\(\alpha\) is not an existing GO term, so a list of genes reported to be p-eIF2\(\alpha\)-resistant was constructed based on Andreev et al., 2015 \[18\] and references within (excluding those from IRESite, which were not found to be p-eIF2\(\alpha\)-resistant in their study). Mouse homologues of these genes were identified using the NCBI homologene database (S4 Table). Enrichment of genes categorised as p-eIF2\(\alpha\)-resistant amongst the genes with significantly increased translational efficiency, compared to a background of all Mus musculus genes included in the TE analysis with any GO annotation, was calculated using a Fisher Exact test.

Chemicals

GSK-2606414 was a kind gift from Dr Edward Emmott and Prof Ian Goodfellow. AEBSF, STF-083010, ISRIB, Ceapin-A7 and tunicamycin were purchased from Sigma-Aldrich. KIRA8
(AMG-18) was obtained from MedChemExpress. GSK-2606414, STF-083010, ISRIB, Ceapin-A7, KIRA8 and tunicamycin were dissolved in DMSO, whereas AEBSF was dissolved in water, to the required concentrations. In all experiments, the final concentration of DMSO did not exceed 0.4% and no differences in viability were observed between untreated cells and DMSO-treated samples. Cytotoxicity after treatment with single and combined UPR inhibitors was measured using the CellTiter-Blue kit (Promega) following manufacturer’s instructions and trypan blue (Sigma) exclusion assay.

**Antibodies**

The following primary antibodies were used: mouse monoclonal antibodies against MHV N and S proteins (kind gifts of Dr Helmut Wege, University of Würzburg), rabbit polyclonal anti-SARS-CoV-2 spike glycoprotein antibody (ab272504, Abcam) mouse anti-GAPDH (IgM specific, G8795, Sigma–Aldrich), mouse anti-Flag (F3165, Sigma–Aldrich), rabbit anti-HA (3724, Cell Signaling Technology), rabbit anti-PERK (ab229912, Abcam), rabbit anti-HERPUD1 (ab150424, Abcam), rabbit anti-GRP78 (BIP, ab108613, Abcam), rabbit anti-eIF2α (9722, Cell Signaling Technology), rabbit anti-phospho-eIF2α (Ser51, 9721, Cell Signaling Technology), rabbit anti-ATF4 (10835-1-AP, Proteintech), rabbit anti-ATF6 (ab203119 and ab37149, Abcam), mouse anti-S6 (2317, Cell Signaling Technology) and rabbit RPL10a (ab174318, Abcam). Secondary antibodies used for western blotting were purchased from Licor: IRDye 800CW Donkey Anti-Mouse IgG (H+L), IRDye 800CW Donkey Anti-Rabbit IgG (H+L), IRDye 680RD Goat Anti-Mouse IgG (H+L) and IRDye 680RD Goat Anti-Mouse IgM (μ chain specific).

**Plasmids and transfections**

HEK-293T cells were transiently transfected with pcDNA3.1-SARS-CoV-2-S-HA (kind gift of Dr Jerome Cattin and Prof Sean Munro, MRC-LMB, Cambridge, UK), pcDNA6-SARS-CoV-2-N-FLAG, pcDNA6-SARS-CoV-2-ORF3a-FLAG and pcDNA6-SARS-CoV-2-ORF8-FLAG plasmids (kind gifts of Prof Peihui Wang, Shandong University, China) using a commercial liposome method (TransIT-LT1, Mirus). Transfection mixtures containing plasmid DNA, serum-free medium (Opti-MEM; Gibco-BRL) and liposomes were set up as recommended by the manufacturer and added dropwise to the tissue culture growth medium. Cells were harvested at 24 and 36 h post-transfection.

**Immunoblotting**

Cells were lysed in 1X Laemmli’s sample buffer. After denaturation at 98°C for 5 minutes, proteins were separated by 12% SDS-PAGE and transferred to nitrocellulose membranes. These were blocked (5% non-fat milk powder or bovine serum albumin in PBST [137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 1.5 mM KH2PO4, pH 6.7, and 0.1% Tween 20]) for 30 min at room temperature and probed with specific primary antibodies at 4°C overnight. Membranes were incubated in the dark with IRDye-conjugated secondary antibodies diluted to the recommended concentrations in PBST for 1 h at room temperature. Blots were scanned using an Odyssey Infrared Imaging System (Liorc). Quantification of protein bands (densitometry) was performed using the ‘area under the curve’ method in ImageJ 1.X software [115]. Relative protein expression was calculated by sequentially normalising against the loading controls (either GAPDH or eIF2α) and then against the timepoint-matched mock for cellular proteins or the timepoint-matched infected but untreated cells for viral proteins (summary of the quantifications in S6 Table).
Analysis of Xbp1 splicing by RT-PCR

Total RNA was isolated from infected or transfected cells as described previously [9], and cDNA synthesised from 500 ng total RNA using M-MLV Reverse Transcriptase (Promega). Mouse or human Xbp1 and Rpl19 were amplified using specific primers (S5 Table). Following PCR reactions, the resulting amplicons were subjected to electrophoresis in 3% agarose gels. Quantification of Xbp1-s and Xbp1-u PCR bands was performed using the 'area under the curve' method in ImageJ 1.X software [115]. The extent of Xbp1 splicing was quantified as Xbp1-s / (Xbp1-s + Xbp1-u), which represents the ratio of Xbp1-s to the sum of Xbp1-s and Xbp1-u, normalised against the same ratio for the timepoint-matched mock (summary of the quantifications in S6 Table).

Quantitative real-time PCR assays

Relative levels of mouse or human Bip, Chop, Gadd34, Calreticulin, Grp94, MHV N and SARS-CoV-2 N in cDNA samples were determined by quantitative real-time PCR (qPCR) using a Rotor-Gene 3000 (Corbett Research). Reactions were performed in a final volume of 20 μl containing Hot Start Taq (1 U, QIAGEN), 3.5 mM MgCl₂, 2.5 mM deoxynucleotides, 450 nM SYBR Green dye, 300 nM relevant forward and reverse primers (S5 Table) and 1 μl of cDNA. vRNA in MHV and SARS-CoV-2 infected cells is quantified with oligonucleotides whose primer binding site is within the N gene (S5 Table). This will detect all canonical positive sense vRNA (i.e. genomic and all subgenomic RNAs). No-template controls were included for each primer pair, and each qPCR reaction was carried out in duplicate. Fold changes in gene expression relative to the mock were calculated by the delta delta-cycle threshold (ΔΔCt) method, and Rpl19 was used as a normalising housekeeping gene.

Polysome profiling

17 Cl-1 cells were infected as described above. 10 min prior to harvesting, cells were treated with cycloheximide (100 μg/ml), washed with PBS and lysed in a buffer containing 20 mM Tris HCl pH 7.5, 100 mM KCl, 5 mM MgOAc, 0.375 mM CHX, 1 mM DTT, 0.1 mM PMSF, 2 U/μl DNase I, 0.5% NP-40, supplemented with protease and phosphatase inhibitors (ThermoFisher Scientific). Following trituration with a 26-G needle (ten passes), lysates were cleared (13,000 g at 4˚C for 20 min) and the supernatants layered onto 12 ml sucrose density gradients (10–50% sucrose in TMK buffer: 20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM MgCl₂) prepared in Beckman SW41 polypropylene tubes using a Gradient Master (Biocomp). Following centrifugation (200,000 g for 90 min at 4˚C), fractions were prepared using an ISCO fractionator monitoring absorbance at 254 nm. Proteins were concentrated from fractions using methanol-chloroform extraction and subjected to immunoblotting analysis. Polysome profiling in higher salt conditions was carried out as described above except that the lysis buffer and sucrose density gradient contained 400 mM KCl.

Virus plaque assays

To determine MHV-A59 titres by plaque assay, 17 Cl-1 cells in 6-well plates were infected with 400 μl of 10-fold serial dilutions of sample in infection medium (Hank’s balanced salt solution containing 50 μg/ml DEAE-dextran and 0.2% bovine serum albumin—BSA). After 45 min at 37˚C with regular rocking, the inoculum was removed and replaced with a 1:1 mixture of 2.4% Avicel and MEM 2X medium (20% MEM 10X, 2% non-essential aminoacids, 200 U/ml penicillin, 200 μg/ml streptomycin, 2 mM L-glutamine, 40 mM HEPES pH 6.8, 10% tryptose phosphate broth, 10% FCS and 0.01% sodium bicarbonate). Plates were incubated at 37˚C for 48 h
prior to fixing with 3.7% formaldehyde in PBS. Cell monolayers were stained with 0.1% toluidine blue to visualise plaques. SARS-CoV-2 plaque assays were performed as previously described [107]. Experiments were conducted using three biological repeats.

**TCID\textsubscript{50} assays**

SARS-CoV-2 replication was assessed using a 50% tissue culture infective dose (TCID\textsubscript{50}) assay in Vero E6 cells. Supernatant derived from infected Caco2 cells was subjected to 10-fold serial dilutions. At 72 h p.i., cells were fixed and stained as previously indicated. Wells showing any sign of cytopathic effect (CPE) were scored as positive.

**Statistical analysis of virus titre results**

Data were analysed in GraphPad Prism 9.0 (GraphPad software, San Diego, CA, USA). Values represent mean ± standard deviation. Statistical significance was evaluated using two-tailed t-tests on log\textsubscript{10}(virus titre) data, which did not assume equal variances for the two populations being compared, to calculate the p-values. Differences as compared to the control with p value ≤ 0.05 were considered as statistically significant, with ‘p < 0.05, “p < 0.01, ‘’p < 0.001 and ‘’’p < 0.0001.

**Supporting information**

S1 Fig. Quality control indicates sequencing data are of high quality. (A) Length distribution of positive-sense reads mapping within host CDSs. For RiboSeq libraries (pink) the characteristic sharp peak at 28–29 nt is observed, reflective of the length of mRNA protected from RNase I digestion. Read lengths of RNASeq libraries (green) are determined by alkaline hydrolysis and gel purification size selection (25–34 nt), leading to a broader length distribution. (B) Percentage of reads (all read lengths) attributed to each phase, for positive-sense reads mapping within host CDSs. Phases correspond to which position within the codon the 5' end of the read maps to (0: purple, 1: blue, 2: yellow). The 5' end coordinate of RiboSeq reads is influenced by the position of the translating ribosome, leading to a clear dominance of the 0 phase. For RNASeq reads the 5' end coordinate is determined by alkaline hydrolysis so does not result in a dominant phase. (C) Distribution of host mRNA-mapping reads relative to start and stop codons. Only transcripts with an annotated CDS of at least 150 codons, 5' UTR of at least 60 nt, and 3' UTR of at least 60 nt were included in the analysis. The total number of positive-sense reads from all these transcripts mapping to each position was plotted, with an offset of +12 relative to the 5' end coordinate to represent the inferred ribosomal P site. Although ribosomal P site position is not relevant to RNASeq reads, these were also plotted with a +12 nt offset to facilitate comparison. Data are coloured according to phase as in B. For RiboSeq libraries there is clear triplet periodicity visible across the CDS, reflective of the length of a codon, and a large peak corresponding to terminating ribosomes—characteristic of samples harvested without drug pre-treatment. Very few RiboSeq reads map to the UTRs (and particularly the 3' UTR), indicating very little contamination of the mRNA fraction with non-ribosome-protected-fragment reads. As expected, for RNASeq libraries the coverage does not differ greatly between the CDS and UTRs. (TIF)

S2 Fig. Distribution of reads mapping to specific host genes of interest. Analysis of RPFs (mock and MHV-infected samples plus tunicamycin-treated sample) and RNASeq reads (mock and MHV-infected samples) mapping to Xbp1-\textsubscript{u} (NCBI RefSeq mRNA NM_013842). Cells were infected with MHV-A59 or mock-infected and harvested at 5 h p.i. or 8 h p.i.
libraries from Fig 1D and 1E). One sample was treated with 2 μg/ml tunicamycin, a pharmacological inducer of all three branches of the UPR, as a positive control. Reads are plotted at the inferred position of the ribosomal P site and coloured according to phase: pink for 0, blue for +1, yellow for +2. The 5′ end position of RNASeq reads is not determined by ribosome position and therefore should not show a dominant phase. The main ORF (0 frame) is shown at the top in pink, with start and stop codons in all three frames marked by green and red bars (respectively) in the three panels below. The yellow rectangle in the +2 frame indicates the extended ORF that results from splicing by IRE1. Reads resulting mainly from translation of the spliced Xbp1-s isoform can be seen in yellow (+2 phase), downstream of the main ORF annotated stop codon. Dotted lines serve as markers for the start and end of the features in their matching colour. Read densities are plotted as reads per million host-mRNA-mapping reads. Bar widths were increased to 4 nt to aid visibility, and therefore overlap, and were plotted sequentially starting from the 5′ end of the transcript.

(TIF)

S3 Fig. ATF6 pathway activation in MHV-infected cells. 17 Cl-1 cells were incubated in the presence of tunicamycin (2 μg/ml) or infected with MHV-A59 (MOI 5) and harvested at 2.5, 5 and 8 h. (A) Cell lysates were separated by 12% SDS-PAGE and immunoblotted using anti-ATF6 (1:1000, Abcam ab203119, upper), anti-ATF6 (1:1000, Abcam ab37149, middle). GAPDH was used as loading control. Representative images of fixed and permeabilised cells treated with tunicamycin for 6 h (B) or infected with MHV for 8 h (C) and incubated with anti-ATF6 (1:300, Abcam ab37149, red) and anti-S protein (green). Nuclei are counterstained with DAPI (blue). Images were taken in an Evos FLII microscope at 60X magnification. Scale bar: 100 μm. (D) Analysis of RPFs and RNASeq reads mapping to Bip (NM_022310). Plot constructed as described in S2 Fig. Note that in order to properly visualise RPFs across the ORF, the y-axis has been truncated at 400 reads per million host-mRNA-mapping reads for the RiboSeq samples, leaving some RPF counts for tunicamycin-treated cells and MHV-infected cells off-scale.

(TIF)

S4 Fig. Cytotoxicity assays of UPR inhibitors in 17 Cl-1 cells. 17 Cl-1 cells were treated with the different UPR inhibitors at the indicated concentrations. Experiments were performed in triplicate using CellTiter-Blue Cell Viability Assay to assess metabolic capacity (dashed line represents 70% threshold) (A) and in duplicate using trypan blue exclusion assay to assess cell proliferation and viability (B) in treatment conditions involving IRE1 or AEBSF. Cell viability in all cases tested was greater than 85% (dotted line). Percentages are given relative to untreated cells. Error bars represent standard errors.

(TIF)

S5 Fig. Effect of UPR inhibitors on activation of the UPR during MHV infection. MHV-infected cells (MOI 5) were treated with UPR inhibitors (5 μM PERKi, 2 μM ISRIB, 60 μM IRE1, or 100 μM AEBSF). The inhibitors were added to the cells immediately after the virus adsorption period and maintained in the medium until cells were harvested 8 h later. (A) Western blot analysis of MHV N protein, PERK, ATF4, BiP, p-eIF2α and eIF2α as loading control. Protein band quantifications (performed as described for Fig 2) are provided for MHV N protein and p-eIF2α and immunoblots are representative of three biological replicates. (B) 17 Cl-1 cells infected with MHV-A59 and treated with 0, 2.5 or 5 μM of PERKi were metabolically pulse-labelled with [35S]Met for 1 h at 5 h p.i. Cells were lysed just after pulse and subjected to 10% SDS-PAGE followed by autoradiography. (C) Polysome profiling as described in Fig 2C of MHV-infected cells at 5 h p.i. treated with 5 μM of PERKi. (D) RT-qPCR (performed as described in Fig 2B) of Bip and Chop mRNA from two biological...
replicates of MHV-infected cells treated with UPR inhibitors as described in Fig 2B.  

(F) RT-qPCR performed as described in Fig 2B of Calreticulin and Grp94 mRNA from two biological replicates of MHV-infected cells treated with UPR inhibitors as described in Fig 2B.  

(E) RT-PCR analysis of Xbp1-u and Xbp1-s mRNAs, performed as described in Fig 2D.  

(TIF) S6 Fig. Effect of dual combinations of UPR inhibitors on activation of the three branches of the UPR during MHV infection. MHV-infected cells were treated with the combinations of the UPR inhibitors shown in Fig 3B. The inhibitors were added to the cells immediately after the virus adsorption period and maintained in the medium until cells were harvested 8 h later. Western blot analysis (upper) of MHV N, PERK, and p-eIF2α proteins from cell lysates of MHV-infected cells at MOI 1 (A) and at MOI 5 (B). The N protein panels have been duplicated from Fig 3C to facilitate comparison. Protein band quantifications, normalised by eIF2α as a loading control and given relative to the mock, are provided for p-eIF2α. RT-PCR analysis of Xbp1-u and Xbp1-s mRNAs (lower panels), performed as described in Fig 2D. Immunoblots and agarose gels are representative of three biological replicates.  

(TIF) S7 Fig. Mechanistic analysis of UPR activation by SARS-CoV-2 proteins. HEK-293T cells were treated with tunicamycin (Tn) or transfected with plasmids encoding SARS-CoV-2 S (S-HA), N (N-FLAG), ORF3 (ORF3-FLAG), ORF8 (ORF8-FLAG), pcDNA.3 as an empty vector (EV) control or mock-transfected as indicated. Cells were harvested at 24 and 36 h p.t. Western blot analysis (upper) of HERP, BiP, PERK, ATF4, p-eIF2α, eIF2α, ATF6, N-FLAG and either S-HA (A) or ORF3-FLAG and ORF8-FLAG (C). The specific p-eIF2α and ATF6-Nt bands are indicated with red asterisks. Protein band quantifications, normalised by eIF2α as a loading control and given relative to the mock, are provided for HERP, BiP, p-eIF2α and ATF6-Nt below the respective immunoblots. RT-PCR analysis of XBP1-u and XBP1-s mRNAs (lower), performed as described in Fig 2D. Immunoblots and agarose gels are representative of three biological replicates. (B) RT-qPCR of BIP and CHOP mRNA from two biological replicates of HEK-293T cells transfected with SARS-CoV-2 N, S or ORF8, or pcDNA.3 as an empty vector, and harvested at 24 and 36 h p.t.  

(H) = hours post-transfection.  

(TIF) S8 Fig. UPRi treatment reverses the activation of the UPR by SARS-CoV-2 proteins. (A) HEK-293T cells were transfected with a plasmid encoding SARS-CoV-2 S (S-HA), mock-transfected, or treated with tunicamycin (Tn). At 8 h p.t., cells were treated with UPR inhibitors (5 μM PERKi, 2 μM IREi, 60 μM IREi or 100 μM AEBSF) and then harvested at 24 and 36 h p.t. Western blot analysis (upper) of S-HA, HERP, BiP, PERK, ATF4, p-eIF2α and ATF6. GAPDH and eIF2α are used as loading controls. The specific p-eIF2α and ATF6-Nt bands are indicated with red asterisks. Protein band quantifications, normalised to the loading controls and given relative to the mock, are provided for HERP, BiP, p-eIF2α and ATF6-Nt below the respective immunoblots. RT-PCR analysis using primers flanking the XBP1 splice site (middle panel), performed as described in Fig 2D. RT-qPCR (lower) of BIP and CHOP mRNA from two biological replicates of HEK-293T cells transfected and treated as described in panel A. Data are normalised as described in Fig 2B. (B) HEK-293T cells were transfected with a plasmid encoding SARS-CoV-2 ORF8 (ORF8-FLAG). At 8 h p.t., cells were treated with 60 μM...
IREi or 100 μM AEBSF and then harvested at 24 and 36 h p.t. Western blotting (upper), RT-PCR (middle) and RT-qPCR (lower) were performed as described in panel A. Immunoblots and agarose gels are representative of three biological replicates. RT-qPCR data of SARS-CoV-2 N, S, ORF8 and empty vector in panels A and B are reproduced from S7B Fig for comparison to the treated conditions, which were performed as part of the same experiment. (C) RT-qPCR of BiP and CHOP mRNA from two biological replicates of HEK-293T cells transfected with SARS-CoV-2 S and ORF8 harvested at 24 and 36 h p.t. Cells were treated with 60 μM IREi and 100 μM AEBSF or a no-drug treatment control. Data are normalised as described in Fig 2B. “h p.t.” = hours post-transfection.

S9 Fig. Effect of the UPR inhibitors on SARS-CoV-2 infection. (A) RT-qPCR of BiP and CHOP mRNA from three biological replicates of Vero CCL81 cells infected and treated as described in Fig 5A. Data are normalised as described in Fig 2B. Caco2 (B), Vero CCL81 (C) or Calu3 cells (D) were treated with the different UPR inhibitors at the indicated concentrations for 24 h. Experiments were performed in triplicate using CellTiter-Blue Cell Viability Assay to assess metabolic activity (B, C left panel and D) and in duplicate using trypan blue exclusion assay to assess cell proliferation and viability (C middle and right panels) in selected conditions. Percentages are given relative to untreated cells. (E) Infectivity of lentiviral particles pseudotyped with SARS-CoV-2 S protein. HEK-293T cells were transfected with ACE2 and TMPRSS2 and treated with 100 μM AEBSF between 0 and 4 h. Lentiviral particles engineered to contain a firefly luciferase reporter were pseudotyped with SARS-CoV-2 S, or vesicular stomatitis virus glycoprotein (VSV-G) as a positive control (+, red) and empty pcDNA 3.1 vector as a negative control (-, orange). Infectivity was measured as firefly luciferase units. Values show the mean averages of three biological replicates. Error bars represent standard errors. All t-tests are two-tailed and do not assume equal variance for the two populations being compared. “ns” = not significant.

S10 Fig. Cytotoxicity assays of UPR inhibitors KIRA8 and Ceapin-A7. 17 CI-1 (A), Vero CCL81 (B) or Calu3 cells (C) were treated with 10 μM KIRA8 or 15 μM Ceapin-A7 as individual treatments or in combination for 8 h (A) or 24 h (B-C). Experiments were performed in triplicate using CellTiter-Blue Cell Viability Assay to assess metabolic activity and in duplicate using trypan blue exclusion assay to assess cell proliferation and viability. Percentages are given relative to untreated cells and dotted lines represent 80% of the mock value.

S11 Fig. Effect of specific inhibition of IRE1α and ATF6 pathways on MHV and SARS–CoV-2 infected cells. Western blot analysis (A and B upper left panels) of MHV N, PERK, ATF4 and p-eIF2α from cell lysates of MHV-infected cells (MOI 5) treated with 10 μM KIRA8 or 15 μM Ceapin-A7 as individual treatments (A) or in combination (B). The inhibitors were added to the cells immediately after the virus adsorption period and maintained in the medium until cells were harvested 8 h later. Protein band quantifications, normalised by eIF2α as a loading control and given relative to the mock, are provided for MHV N and p-eIF2α below the respective immunoblots. RT-PCR analysis of Xbp1-u and Xbp1-s mRNAs (A and B lower left panels) was performed as described in Fig 2D. RT-qPCR of Gpr94 mRNA (A) or MHV vRNA, Gpr94, BiP and Chop (B), performed for three biological replicates and normalised as described in Fig 2B. Immunoblots and agarose gels are representative of three biological replicates. (C) Plaque assays were performed with serial dilutions of the supernatant containing released virions from (A) and (B). (D) Viral titres of Calu3 cells infected with SARS-CoV-
2 at MOI 5 and treated with the different UPR inhibitors as described in Fig 6. Plaque assays were performed with serial dilutions of the supernatant containing released virions harvested at 24 h p.i. In all cases, values show the mean averages of the titration of three biological replicates. Error bars represent standard errors. All p-values are from comparisons with the respective untreated control, with ‘p < 0.05, ‘‘p < 0.01, ‘’’p < 0.001 and ‘’’’p < 0.0001.

(TIF)

S1 Table. Riboseq and RNASeq library composition. Number of reads assigned to each category. Reads under 25 nt long were designated “too short”.

(XLSX)

S2 Table. Differential gene expression results. Sheets 1–4: Ranked lists of genes which passed the thresholds of log₂(fold change) greater than or equal to 1 (corresponding to a fold change of 2) and p value less than or equal to 0.05 after false discovery rate (FDR) adjustment for multiple testing. Sheets are as follows: TS_up, TS_down–genes which are significantly more (up) or less (down) transcribed in infected samples compared to mock, TE_up, TE_down–genes which are significantly more (up) or less (down) efficiently translated in infected samples compared to mock. Read counts for each sample, normalised by the total number of host-mRNA-mapping reads for that sample, are given in the right-most columns. Sheets 5 and 6: Full lists of genes which pass the threshold for inclusion in the analyses (requires ten reads mapping to this gene between all samples). TS = transcription, TE = translation efficiency.

(XLSX)

S3 Table. Reactome pathway and GO term enrichment analysis results. Sheets 1–4: Enriched Reactome pathways. Lists of mouse gene names of significantly differentially expressed genes (S2 Table) were used for Reactome pathway enrichment [11], in which they were converted to their human orthologues and analysed to determine which pathways are significantly over-represented. Input gene lists are indicated in the sheet name, for example ‘Reactome_TS_up’ shows the Reactome enrichment results generated using the TS_up list from S2 Table as input. Sheets 5–8: Enriched GO terms. The same differentially expressed mouse gene lists were used for GO term enrichment analysis by PANTHER [114], against a background list of all the genes which passed the threshold for inclusion in that expression analysis. Column labels are as described in both Reactome and PANTHER user guides. All results with significant p values (≤ 0.05) are shown.

(XLSX)

S4 Table. List of genes classified as translationally resistant to eIF2α phosphorylation, based on Andreev et al [18]. The first column shows the human genes classified as p-eIF2α-resistant by Andreev et al (excluding those from IRESite, which were not found to be p-eIF2α-resistant in their study). The list of genes used for the enrichment analysis, displayed in the second column, was generated by identifying mouse homologues of the human genes using the NCBI Homologene database [Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 44, (2016)]. Genes which were significantly more efficiently translated during MHV infection compared to mock infection are highlighted in bold.

(XLSX)

S5 Table. List of oligonucleotides.

(DOCX)

S6 Table. Raw data analysis.

(XLSX)
SI Text. Supplementary Materials and Methods.

(PDF)

Acknowledgments
SK, ND, GD and EB would like to acknowledge Dr Holly Shelton, Dr Isabelle Dietrich and Dr Christine Reitmayer for their supervision in the CL3 suite, and Dr Christine Tait-Burkhard and Dr Juergen Haas for the SARS-CoV-2 isolate. BGH would like to thank Prof Frank Kirchhoff for the Calu3 cells. NI would like to thank Dr James Edgar for providing SARS-CoV-2 plasmids. NI would like to thank Dr Luke Meredith and Prof Ian Goodfellow for providing pcDNA3.1-ACE2, pcDNA3.1-TMPRSS2, pMD2-VSV-G and pNL4.3-Luc plasmids for pseudotyped virions.

Author Contributions
Conceptualization: Nerea Irigoyen.
Data curation: Georgia M. Cook, Katherine Brown, Andrew E. Firth.
Formal analysis: Liliana Echavarría-Consuegra, Georgia M. Cook, Idoia Busnadiego, Andrew E. Firth, Nerea Irigoyen.
Funding acquisition: Erica Bickerton, Benjamin G. Hale, Andrew E. Firth, Ian Brierley, Nerea Irigoyen.
Investigation: Liliana Echavarría-Consuegra, Georgia M. Cook, Idoia Busnadiego, Charlotte Lefèvre, Sarah Keep, Nicole Doyle, Giulia Dowgier, Nerea Irigoyen.
Methodology: Liliana Echavarría-Consuegra, Georgia M. Cook, Idoia Busnadiego, Katherine Brown, Stuart G. Siddell, Nerea Irigoyen.
Project administration: Nerea Irigoyen.
Resources: Katherine Brown, Stuart G. Siddell, Erica Bickerton, Benjamin G. Hale, Andrew E. Firth, Ian Brierley.
Software: Georgia M. Cook, Katherine Brown, Krzysztof Franaszek, Nathan A. Moore, Andrew E. Firth.
Supervision: Stuart G. Siddell, Erica Bickerton, Benjamin G. Hale, Andrew E. Firth, Ian Brierley, Nerea Irigoyen.
Validation: Liliana Echavarría-Consuegra, Idoia Busnadiego, Charlotte Lefèvre, Nerea Irigoyen.
Visualization: Liliana Echavarría-Consuegra, Georgia M. Cook, Nerea Irigoyen.
Writing – original draft: Liliana Echavarría-Consuegra, Georgia M. Cook, Ian Brierley, Nerea Irigoyen.
Writing – review & editing: Liliana Echavarría-Consuegra, Georgia M. Cook, Idoia Busnadiego, Charlotte Lefèvre, Sarah Keep, Katherine Brown, Stuart G. Siddell, Erica Bickerton, Benjamin G. Hale, Andrew E. Firth, Ian Brierley, Nerea Irigoyen.

References


22. Muniqru IX, Pagnan J, Kohannim O, Gargalovic PS, Lusis AJ. CHAC1/MGC4504 Is a Novel Proapoptotic Component of the Unfolded Protein Response, Downstream of the ATF4-ATF3-CHOP Cascade.
Targeting the unfolded protein response in coronavirus infection

The Journal of Immunology. 2009 Jan 1; 182(1). https://doi.org/10.4049/jimmunol.182.1.466 PMID: 19109178


25. Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K. XBP1 mRNA Is Induced by ATF6 and Spliced by IRE1 in Response to ER Stress to Produce a Highly Active Transcription Factor. Cell. 2001 Dec; 107 (7). https://doi.org/10.1016/s0092-8674(01)00611-0 PMID: 11779464


Azouz NP, Klingler AM, Rothenberg ME. Alpha 1 antitrypsin is an inhibitor of the SARS-CoV-2 priming protease TMPRSS2. bioRxiv. 2020;


