Computational reconstruction of mouse development using single cell transcriptomics

Iván Imaz Rosshandler

Department of Haematology
University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Clare College
May 2021
DECLARATION

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the individual declarations at the beginning of each chapter.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation 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. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices and has fewer than 150 figures.

Iván Imaz Rosshandler
May 2021
Abstract

Computational reconstruction of mouse development using single cell transcriptomics
Iván Imaz Rosshandler

Single cell transcriptomics has significantly contributed to our understanding of cell types across species, organs and developmental processes. The rapid development of technologies, protocols and computational methods reflects a highly dynamic field over constant improvement. Particularly for Developmental Biology, being able to study at single-cell level and over time is required for a more detailed understanding of the underlying molecular mechanisms associated with cell differentiation and fate choice. Longitudinal cohort studies, ideal for capturing the dynamic nature of developmental mechanisms delineate the temporal relationship between cell type diversity and developmental dynamics. The efforts to build cell atlases from different model organisms already include extensive transcriptomic profiling of embryonic development, where detailed non-human models are essential to compensate the limitations arising from ethical reasons. In this context, the transcriptomic landscape of mouse development is perhaps the most complete among mammals. However, as cells are destroyed when measuring their transcriptomic profiles, only snapshots of the dynamical system are effectively captured. Thus, computational reconstructions of cell differentiation trajectories are essential even when lineage tracing experiments can be performed. The work presented in this dissertation uses a time course large-scale single-cell transcriptomic experiment of mouse embryos to present an effective mathematical and computational strategy that adequately describes the dynamics of gastrulation and early organogenesis on a variety of cell types during mouse embryonic development. The experimental data generated for this project, incorporates new time points into the mouse gastrulation and early organogenesis atlas already publicly available (Pijuan-Sala et al., 2018). That is, an extended version of the existing atlas. Results show a pipeline of sophisticated computational strategies to integrate the new time points. Then, to overcome the challenge of identifying and reconstructing the dynamics of all cell lineages, previously generated knowledge was combined with a variety of state-of-the-art computational methodologies. The analysis of differentiation trajectories is then taken to a deeper level in regards to the emergence of haemato-endothelial lineages, with emphasis on resolving the so-called Primitive and Definitive waves of blood production. Furthermore, perturbations to the system using mouse embryonic chimaera KO models are included. The analysis of haemato-endothelial lineages, presents a detailed reconstruction of the in vivo developmental process not reached by previous studies with single cell transcriptomics. Lastly, predicted cell fates are compared to experimental observations by leveraging lineage trace experiments using cell grafting. In summary, this work high-
lights the complexity associated to generating developmental atlases, how to overcome the corresponding computational challenges and leverage this resource in a variety of contexts.
ACKNOWLEDGEMENTS

Bertie Göttgens and John Marioni for the comprehensive support, kind guidance as well as bright opportunities and experiences provided during the entire course of my PhD.

My examiners Peter Rugg-Gunn and Lars Velten for their amazing job at challenging my thesis work.
Carolina Guibentif, for adopting me as a student and putting so much effort in our collaborations.

Melania Barile, for reading my entire thesis in two days as well as providing new perspectives of work and collaboration during the last stage of my PhD.

Christina Rode and Marella De Brujin, for such a friendly, interesting and supportive collaboration.

Ricard Argelaguet, for the opportunity of collaborating with him, as well as his support and advice over a variety of projects. Also, for the critical thinking provided.

The bioinformatics team at Göttgen’s group: Blanca Pijuan Sala, Samuel Watchman, Fiona Hamey, Xionan Wang, Rebbeca Hannah, Hugo Da Silva Bastos, Kendig Sham, Mai-Linh Ton, Mariana Quiroga-Londono, Shirom Chabra and Vijaya Mahalingam Shanmugiah for their general support and company during the PhD.

John Marioni’s lab members: Tom W. Hiscock, Aron Lun, Jonathan A. Griffiths, Shila Ghazanfar, Alsu Missarova, Andrian Yang, Michael Morgan, Adi Steif, Rebecca Barrens and Ximena Ibarra-Soria for a truly supportive and friendly experience as the sporadic member of the group I was.

Nicola Wilson, Fernando Calero-Nieto and Iwo Kucinski for their advice and constant willing to help.

Claudia Rangel Escareño, for going through the pain of reading another thesis written by me. More importantly, for all years of support.
Juan Edgar Castillo Fernández and Jordana Bell, for the opportunity of collaborating with them during the initial stage of my PhD.

Chip Stewart, Akinyemi Onjesina and Andrea Manica, for their mentorship and support.

My parents Ana Rosshandler Fernández and Jorge Arturo Imaz Lira, as somehow they took me where I am.

My grandmothers, Dolores Fernández Ramírez †, Berta Fernández Ramírez and Berta Lira Mendez †, for enlightening my life.

My thesis comrades, now Dr. Margherita Protasoni and Dr. Luz Alonso-Crisostomo, hard times indeed create close bonds.

Joe Thomson and Joe Halstead (a.k.a. The Joe(s)), for their encourage to chase my dreams.

Lena Morril and D.J. Kunz for their truly kind and caring friendship through my PhD.


My hood brothers and sisters that were always present, Daniel Alberto Estrada Santos †, Canek Malagamba Rosshandler, Raúl Emerico and Luis Rodríguez Martínez, Luis Enrique de la Toba, Angel Zavala Baute, Jorge Nicolas Peña Mora, Jesús Ramírez Arreguín, Sebastian Aguilar-Pierlé, Alfonso Santiago Romero Ruiz de Aguirre, Luis Pineda Orozco, Ricardo Rastorga Castañeda, Erick Cuevas García Figueroa, Ulises and Tania Hurwitz, Esther Toledano Bautista, Susana del Rocío Solorzano Rosales, Karina Pinto Savin, Viridiana Hernández Muñoz-Ledo, Rigel Nuñez and Alejandra Ahumada, and many more but there is no more space. For a bond that goes beyond death and life, in which you can always count.

The mice sacrificed for this experiment, for giving their life for the scientific progress.

The Wellcome Mathematical Genomics and Medicine PhD programme for such a fantastic opportunity and support.
Dedicated to Susana Solorzano, for a decade of pure love... Umuntu ngumuntu ngabantu
# Table of contents

List of figures  
List of tables  
Acronyms  

1 Introduction  
1.1 Large scale single cell transcriptomics  
1.2 Approaches for scRNA-seq  
  1.2.1 Smart-seq  
  1.2.2 MARS-seq  
  1.2.3 Drop-seq  
  1.2.4 10x Genomics  
  1.2.5 SPLiT-seq  
1.3 The role of scRNA-seq in developmental biology  
  1.3.1 Allele specific expression  
  1.3.2 Lineage tracing  
  1.3.3 Spatially resolved transcriptomics  
  1.3.4 Multi-omics  
1.4 An overview of Mouse development  
  1.4.1 Pre-implantation  
  1.4.2 Post-implantation  
1.5 Transcriptomic atlases of mouse development  
1.6 Computational analysis of scRNA-seq data  
  1.6.1 Low level analysis  
  1.6.2 Selection of highly variable genes  
  1.6.3 Dimensionality reduction and visualisation  
  1.6.4 Cell heterogeneity and clustering  
  1.6.5 Integration of multiple datasets  
  1.6.6 Developmental trajectories and pseudotime  
1.7 Incorporating time to infer developmental trajectories  
  1.7.1 RNA velocity in single cells
# Table of contents

1.7.2 Waddington-Optimal Transport ........................................... 29  
1.8 Aims ......................................................................................... 34  
1.8.1 Extending the mouse gastrulation atlas of gastrulation  
and early organogenesis to E9.5 ................................................. 34  
1.8.2 Reconstructing the haemato-endothelial trajectory landscape .... 35  
1.8.3 Analysis of Smart-seq2 transcriptomic profiles  
from Primitive Streak dissections at E7.5 ................................. 36  
1.9 Other contributions ................................................................. 38  

2 Extending the single cell atlas coverage of early organogenesis to E9.5 39  
2.1 Introduction ................................................................................ 41  
2.2 Experimental design .................................................................. 42  
2.3 A publicly available mouse gastrulation atlas ............................... 43  
2.4 Low level analysis of new time points ........................................ 43  
2.4.1 Sequencing reads processing ................................................. 43  
2.4.2 Detection and removal of barcode swapping ............................ 44  
2.4.3 Cell calling ............................................................................ 44  
2.4.4 QC ....................................................................................... 45  
2.4.5 Normalisation ......................................................................... 45  
2.4.6 Highly variable gene selection .............................................. 48  
2.4.7 Batch effect correction ............................................................ 49  
2.4.8 Doublet & stripped nucleus identification ............................... 51  
2.5 Creating an integrated landscape .............................................. 52  
2.5.1 Mapping stage and cell type annotations within the atlas ......... 58  
2.5.2 Estimating the descendants from cell populations at E8.5 ......... 58  
2.5.3 Computational validation of transport maps ......................... 59  
2.6 The annotation beyond cell descendants ................................... 62  
2.6.1 An overview of Gut tube cell heterogeneity at E8.75 ............... 63  
2.6.2 Exploring the mouse Brain development atlas ..................... 65  
2.6.3 A clustering based approach ................................................. 65  
2.6.4 Towards a complete annotation of cell types ....................... 68  
2.7 Discussion ................................................................................. 76  

3 Resolving Haemato-Endothelial diversification 79  
3.1 Introduction ................................................................................. 81  
3.2 The origin of haemato-endothelial lineages ................................. 83  
3.2.1 Waddington-OT trajectories in the interval E6.5-E8.5 ............. 83
### Table of contents

- 3.2.2 The role of master regulators in lineage diversification . . . . . . . 92
- 3.3 Perturbations in development: a Tal1 Chimaera . . . . . . . . . . . . . . . . 92
  - 3.3.1 Exploring the role of Tal1 in blood production at E7.5 . . . . . . . 94
- 3.4 The mRNA kinetics of Erythroid differentiation . . . . . . . . . . . . . . . 102
- 3.5 Extending the haemato-endothelial landscape . . . . . . . . . . . . . . . . 103
- 3.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114

- 4 Dissecting the primitive streak at E7.5 to explore cell fate bias 117
  - 4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
  - 4.2 Low level analysis of Smart-seq2 experiments . . . . . . . . . . . . . . . 121
  - 4.3 Mapping primitive streak cells against the atlas . . . . . . . . . . . . . . 121
  - 4.4 Clustering and Differential gene expression . . . . . . . . . . . . . . . . . 124
  - 4.5 Predicted vs Observed fates of Primitive Streak cells . . . . . . . . . . . . 130
  - 4.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134

- 5 Conclusion and future directions 137
  - 5.1 About fundamental questions . . . . . . . . . . . . . . . . . . . . . . . . 137
  - 5.2 The nightmare of batch effects . . . . . . . . . . . . . . . . . . . . . . . . 138
  - 5.3 Is RNA velocity useful? . . . . . . . . . . . . . . . . . . . . . . . . . . . . 139
  - 5.4 Lessons learned from Waddington-OT . . . . . . . . . . . . . . . . . . . . 139
  - 5.5 When will cell type annotation end? . . . . . . . . . . . . . . . . . . . . . 140
  - 5.6 Key observations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 141
  - 5.7 Ongoing and future work . . . . . . . . . . . . . . . . . . . . . . . . . . . 142

References 143

Appendix A Experimental methods 161
  - A.1 Embryo collection and processing . . . . . . . . . . . . . . . . . . . . . . 161
    - A.1.1 E8.5-E9.5 time-points . . . . . . . . . . . . . . . . . . . . . . . . . . 161
    - A.1.2 Tal1 Chimaera . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 162
    - A.1.3 Primitive streak dissected cells library preparation . . . . . . . . . 163
LIST OF FIGURES

1.1 Single-cell library preparation summary. ........................................... 3
1.2 10X Single-cell Chromium v3 library preparation summary. ............... 7
1.3 Schematic representation of mouse embryonic pre-implantation. ........ 14
1.4 Diagram illustrating gastrulation. ...................................................... 15
1.5 Diagram illustrating organogenesis. .................................................... 15
1.6 The concept of pseudotime .............................................................. 25
1.7 RNA velocity can be used to infer developmental trajectories .......... 28
1.8 Introduction to Waddington OT ....................................................... 29
1.9 The use of transport maps .............................................................. 30

2.1 Experimental design of the extended atlas of mouse development ...... 42
2.2 UMAP of mouse atlas of gastrulation and early organogenesis .......... 43
2.3 Swapped barcodes across all samples ................................................ 44
2.4 Library complexity. ................................................................. 46
2.5 Removal of cells with high fraction of mitochondrial UMIs. ............. 47
2.6 Distribution of cell libraries retained after QC across time points. ...... 48
2.7 Normalisation size factors ............................................................ 49
2.8 Highly variable genes selection ...................................................... 50
2.9 Example of doublet identification in Sample 9. ................................ 52
2.10 Median doublet score across samples. ............................................. 53
2.11 Distribution of UMIs of singlets and doublets. ................................ 54
2.12 tSNE displaying doublet identification. .......................................... 55
2.13 Identification of cells with stripped nuclei. ..................................... 56
2.14 Number of cells across time points covered by the extended atlas. .... 56
2.15 Overview of the extended atlas using a UMAP representation. ......... 57
2.16 Graphical visualisation of E8.5 cell descendant predictions using W-OT 60
2.17 Geodesic interpolation ............................................................... 62
2.18 Mapping cell types from Nowotschin et al., 2019, against the extended atlas 64
2.19 Mapping cell types from [1] against the extended atlas ................. 66
2.20 Mapping tissues and region dissections from La Manno et al., against the extended atlas. ............................................................. 67
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.21</td>
<td>Extended atlas UMAP displaying louvain clustering results</td>
<td>69</td>
</tr>
<tr>
<td>2.22</td>
<td>The Ectoderm-Endoderm and Mesoderm landscapes of lineage diversification.</td>
<td>70</td>
</tr>
<tr>
<td>2.23</td>
<td>Leiden clustering of different diversification landscapes.</td>
<td>71</td>
</tr>
<tr>
<td>2.24</td>
<td>Force layout alternatives for visualizing lineage diversification.</td>
<td>72</td>
</tr>
<tr>
<td>2.25</td>
<td>Force layouts clustering based annotation results.</td>
<td>73</td>
</tr>
<tr>
<td>2.26</td>
<td>UMAP layout containing single-cell transcriptomes belonging from E6.5 to E9.5</td>
<td>75</td>
</tr>
<tr>
<td>3.1</td>
<td>Cellular origin of primitive and definitive haematopoiesis.</td>
<td>82</td>
</tr>
<tr>
<td>3.2</td>
<td>Haematoendothelium (E6.5-E8.5).</td>
<td>84</td>
</tr>
<tr>
<td>3.3</td>
<td>Mesoderm landscape (E6.5-E8.5).</td>
<td>85</td>
</tr>
<tr>
<td>3.4</td>
<td>Ancestors mass distributions of mesodermal fates diverge across time (E6.5-E8.5) and cell types</td>
<td>86</td>
</tr>
<tr>
<td>3.5</td>
<td>Trajectory of Erythroid sub-cluster 4 (Ery4).</td>
<td>88</td>
</tr>
<tr>
<td>3.6</td>
<td>Log-likelihood ratios of mesodermal fates.</td>
<td>89</td>
</tr>
<tr>
<td>3.7</td>
<td>3D UMAP of erythroid differentiation.</td>
<td>90</td>
</tr>
<tr>
<td>3.8</td>
<td>Force directed layout of cells likely belonging to haemato-endothelial trajectories</td>
<td>91</td>
</tr>
<tr>
<td>3.9</td>
<td>Trajectory trends of master regulators.</td>
<td>93</td>
</tr>
<tr>
<td>3.10</td>
<td>UMAP of E7.5 <em>Tal1</em> chimaera experiment.</td>
<td>95</td>
</tr>
<tr>
<td>3.11</td>
<td>Integrating trajectory analysis with the E7.5 <em>Tal1</em> Chimaera experiment.</td>
<td>96</td>
</tr>
<tr>
<td>3.12</td>
<td>Chimaera cells projected into multiple atlas trajectories.</td>
<td>98</td>
</tr>
<tr>
<td>3.13</td>
<td>Inspection of <em>Tal1</em> expression patterns in early mouse development.</td>
<td>99</td>
</tr>
<tr>
<td>3.14</td>
<td>Multi-omics inspection of <em>Tal1</em> during gastrulation (E5.5-E7.5).</td>
<td>100</td>
</tr>
<tr>
<td>3.15</td>
<td><em>Tal1</em> splicing dynamics shows transcriptional burst in the Epiblast at E7.0.</td>
<td>101</td>
</tr>
<tr>
<td>3.16</td>
<td>RNA velocity landscape.</td>
<td>104</td>
</tr>
<tr>
<td>3.17</td>
<td>Identification of MURK genes along yolk-sac erythropoiesis.</td>
<td>105</td>
</tr>
<tr>
<td>3.18</td>
<td>Removal of MURK genes for correcting <em>scVelo</em> predictions.</td>
<td>106</td>
</tr>
<tr>
<td>3.19</td>
<td>Haemato-endothelial $log_{10}(odds)$ of fate probabilities across the mouse extended atlas landscape.</td>
<td>108</td>
</tr>
<tr>
<td>3.20</td>
<td>Confounding factors in W-OT trajectory reconstruction.</td>
<td>109</td>
</tr>
<tr>
<td>3.21</td>
<td>Force layouts displaying the reconstruction of Haemato-endothelial developmental trajectories.</td>
<td>112</td>
</tr>
<tr>
<td>3.22</td>
<td>Dissecting the origins of EMPs.</td>
<td>113</td>
</tr>
<tr>
<td>4.1</td>
<td>Primitive streak dissections at E7.5.</td>
<td>120</td>
</tr>
<tr>
<td>4.2</td>
<td>Low level analysis of Smart-seq2 primitive streak cells at E7.5</td>
<td>122</td>
</tr>
</tbody>
</table>
4.3 Batch correction of Smart-seq2 primitive streak cells at E7.5. . . . . . . 123
4.4 UMAP layout of primitive streak cell transcriptomes at E7.5. . . . . . . . 125
4.5 Projection of primitive streak cells at E7.5 into the mouse development atlas. 126
4.6 Primitive Streak portion D cells annotated as blood progenitors. . . . . . . 127
4.7 Unsupervised clustering of primitive streak cell transcriptomes at E7.5. . . 128
4.8 Heat-map of differentially expressed genes between unsupervised clusters. . 129
4.9 Fate plots of Primitive Streak portion dissected cells. . . . . . . . . . . . . 131
4.10 Association between Primitive Streak dissected portions and computational predictions of cell fates. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 132
4.11 Schematic representation of primitive streak cell grafting experiments. . . 133
4.12 Association between Primitive Streak dissected portions and experimental observations of cell fates. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133
4.13 Experimental design for molecular analysis of grafted cells followed by single cell transcriptomic profiling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 135
List of tables

2.1 Summary of methods for trajectory reconstruction. 61
ACRONYMS

AE  Allantois Endothelium
Allantois/EC Allantois endothelial cells
ASE  allele-specific expression
ATAC-seq assay for transposase accessible chromatin sequencing

BBKNN batch balanced KNN graph
BH  Benjamini–Hochberg
BP  Blood progenitors

CARLIN CRISPR array repair lineage tracing
CCA  Canonical Correlation Analysis
cDNA complementary DNA
CITE-seq Cellular indexing of transcriptomes and epitopes by sequencing
CRISPR/Cas9 clustered regularly interspaced short palindromic repeat/CRISPR associated protein 9
CUT&Tag cleavage under targets and tagmentatio

DR-seq gDNA-mRNA sequencing
DVE  Distal visceral endoderm

E  Embryonic day
EC  Endothelial cells
Em/EC Embryo proper endothelial cells
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP</td>
<td>Erythroid-myeloid-progenitors</td>
</tr>
<tr>
<td>EMT</td>
<td>epithelial-to-mesenchymal transition</td>
</tr>
<tr>
<td>EP-EMP</td>
<td>Embryo proper-Erythroid-myeloid-progenitors</td>
</tr>
<tr>
<td>EPDC</td>
<td>Epicardial derived cells</td>
</tr>
<tr>
<td>EPE</td>
<td>Embryo proper endothelium</td>
</tr>
<tr>
<td>EPI</td>
<td>Epiblast</td>
</tr>
<tr>
<td>ERCC</td>
<td>External RNA Control Consortium</td>
</tr>
<tr>
<td>ERY</td>
<td>Erythroid</td>
</tr>
<tr>
<td>ExE</td>
<td>Extraembryonic</td>
</tr>
<tr>
<td>ExE ectoderm</td>
<td>Extra-embryonic ectoderm</td>
</tr>
<tr>
<td>FACS</td>
<td>fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>FHF</td>
<td>First and second heart field</td>
</tr>
<tr>
<td>FISH</td>
<td>fluorescence in situ hybridization</td>
</tr>
<tr>
<td>FPKM</td>
<td>fragments per kilobase per million mapped reads</td>
</tr>
<tr>
<td>G&amp;T-seq</td>
<td>Genome and Transcriptome sequencing</td>
</tr>
<tr>
<td>GEMs</td>
<td>gel beads in emulsions</td>
</tr>
<tr>
<td>HEP</td>
<td>Haemato-endothelial progenitors</td>
</tr>
<tr>
<td>HSC</td>
<td>Haematopoietic Stem Cells</td>
</tr>
<tr>
<td>HVGs</td>
<td>highly variable genes</td>
</tr>
<tr>
<td>ICM</td>
<td>inner cell mass</td>
</tr>
<tr>
<td>iNMF</td>
<td>integrative non-negative matrix factorization</td>
</tr>
<tr>
<td>IVT</td>
<td><em>in vitro</em> transcription</td>
</tr>
<tr>
<td>KNN</td>
<td>K-nearest neighbours</td>
</tr>
<tr>
<td>KO</td>
<td>knockout</td>
</tr>
<tr>
<td>LPM</td>
<td>Lateral plate mesoderm</td>
</tr>
<tr>
<td>Acronyms</td>
<td>Definition</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>M&amp;T-seq</td>
<td>Methylome and Transcriptome sequencing</td>
</tr>
<tr>
<td>MAD</td>
<td>median absolute deviation</td>
</tr>
<tr>
<td>MARS-seq</td>
<td>Massively parallel RNA single-cell sequencing</td>
</tr>
<tr>
<td>MEP</td>
<td>Megakaryocyte-erythroid-progenitors</td>
</tr>
<tr>
<td>MERFISH</td>
<td>multiplex error-robust fluorescence in situ hybridization</td>
</tr>
<tr>
<td>mESC</td>
<td>mouse embryonic stem cells</td>
</tr>
<tr>
<td>Mk</td>
<td>Megakariocyte</td>
</tr>
<tr>
<td>MNN</td>
<td>mutual nearest neighbours</td>
</tr>
<tr>
<td>MOFA</td>
<td>multi-omics factor analysis</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>mtDNA</td>
<td>mitochondrial DNA</td>
</tr>
<tr>
<td>MURK</td>
<td>multiple rate Kinetics</td>
</tr>
<tr>
<td>My</td>
<td>Myeloid</td>
</tr>
<tr>
<td>NK</td>
<td>Natural killer</td>
</tr>
<tr>
<td>NMPs</td>
<td>Neuro-mesodermal progenitors</td>
</tr>
<tr>
<td>NMT-seq</td>
<td>Nucleosome, Methylation and Transcription sequencing</td>
</tr>
<tr>
<td>ODE</td>
<td>ordinary differential equation</td>
</tr>
<tr>
<td>OT</td>
<td>Optimal Transport</td>
</tr>
<tr>
<td>Paired-seq</td>
<td>parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PE</td>
<td>Primitive Endoderm</td>
</tr>
<tr>
<td>PGCs</td>
<td>Primordial germ cells</td>
</tr>
<tr>
<td>QC</td>
<td>quality control</td>
</tr>
<tr>
<td>qRT-PCR</td>
<td>quantitative real-time polymerase chain reaction</td>
</tr>
<tr>
<td>RAID</td>
<td>RNA and immunodetection</td>
</tr>
<tr>
<td>REAP-seq</td>
<td>RNA expression and protein sequencing assay</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>RLE</td>
<td>Relative Log Expression</td>
</tr>
<tr>
<td>SCDS</td>
<td>single-cell data science</td>
</tr>
<tr>
<td>scRNA-seq</td>
<td>single cell RNA-seq</td>
</tr>
<tr>
<td>seqFISH</td>
<td>sequential fluorescence <em>in situ</em> hybridization</td>
</tr>
<tr>
<td>SHF</td>
<td>Second heart field</td>
</tr>
<tr>
<td>SIDR</td>
<td>simultaneous isolation of genomic DNA and total RN</td>
</tr>
<tr>
<td>Smart-seq</td>
<td>Switch mechanism at the 5’ end of RNA templates sequencing</td>
</tr>
<tr>
<td>SNN</td>
<td>shared nearest neighbour</td>
</tr>
<tr>
<td>SPLiT-seq</td>
<td>Split-pool ligation-based transcriptome sequencing</td>
</tr>
<tr>
<td>SVD</td>
<td>singular value decomposition</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-distributed stochastic neighbour embedding</td>
</tr>
<tr>
<td>TE</td>
<td>Trophectoderm</td>
</tr>
<tr>
<td>TMM</td>
<td>trimmed mean of M-values</td>
</tr>
<tr>
<td>Trio-seq</td>
<td>Triple omics sequencing</td>
</tr>
<tr>
<td>UMAP</td>
<td>uniform manifold approximation and projection for dimension reduction</td>
</tr>
<tr>
<td>UMIs</td>
<td>unique molecular identifiers</td>
</tr>
<tr>
<td>UOT</td>
<td>Unbalance Optimal Transport</td>
</tr>
<tr>
<td>VE</td>
<td>Visceral endoderm</td>
</tr>
<tr>
<td>W-OT</td>
<td>Waddington-Optimal Transport</td>
</tr>
<tr>
<td>XCI</td>
<td>X chromosome inactivation</td>
</tr>
<tr>
<td>YS</td>
<td>Yolk-Sac</td>
</tr>
<tr>
<td>YS-EMP</td>
<td>Yolk-Sac Erythroid-myeloid-progenitors</td>
</tr>
<tr>
<td>YS-HEP</td>
<td>Yolk-Sac Haemato-endothelial progenitors</td>
</tr>
<tr>
<td>YS/EC</td>
<td>Yolk-Sac endothelial cells</td>
</tr>
<tr>
<td>YSE</td>
<td>Yolk Sac endothelium</td>
</tr>
</tbody>
</table>
Introduction

The era of high-throughput sequencing technologies provides an unprecedented opportunity for exploring the molecular mechanisms that underlie biology. It has now reached single cell resolution. Thus, the study of macromolecules such as RNA across thousands or even millions of cells in parallel has become a reality. The rapid development of technologies, protocols and computational methods is already showing great promise for research applications in life sciences. The analysis of single cell gene expression at large scale is impacting a number of main areas such as Cancer Biology, Immunology, Microbiology, Neurobiology and Developmental Biology. In all cases, it leads to the possibility of detecting rare events of relevance and confers an unbiased view of the heterogeneity of single cells within a complex population. Furthermore, since the vast majority of living cells in the environment cannot be cultivated in vitro, single-cell approaches also hold the promise of studying poorly understood cell types or regulatory processes of biotechnological or medical relevance.

In recent years, single cell transcriptomics has begun to uncover the molecular programs that drive developmental processes and provided the alluring prospect of cataloguing cell type diversity and developmental dynamics in a quantitative and comprehensive manner. It has matured to the point that it is possible to generate large single cell atlases of developing embryos, being the mouse model among the most advanced. The associated computational aspects gave rise to a new and dynamic research field of data science that faces unprecedented challenges. This thesis addresses the computational analysis associated with the generation and extension of the single cell transcriptomic atlas of mouse gastrulation and early organogenesis, and emphasizes in the reconstruction of developmental trajectories using innovative state-of-the-art methodologies.
1.1 Large scale single cell transcriptomics

Exploration of single cell messenger RNA (mRNA) profiles has been pioneered by a host of probe-dependent methods including reporter fusions to fluorescent proteins, fluorescence in situ hybridization (FISH), quantitative real-time polymerase chain reaction (qRT-PCR), and microarrays [2]. The first transcriptomes generated via single cell RNA-seq (scRNA-seq) were published in 2009 [3], two years after the first applications of RNA-seq to bulk populations of cells [4]. Since then, several technological improvements have been developed and it is still an area of active research. Initially, most studies used plate-based assays, where library preparation is performed manually on cells sorted into and lysed in individual wells of a microwell plate [5, 6]. The development of integrated fluidic circuits and the introduction of liquid handling robots into the process scaled cell numbers to several thousands, which then further increased with nanodroplet and picowell technologies [7]. The introduction of in situ barcoding escalated throughput to hundreds of thousands of cells with the latest developments in spatial methods integrating the spatial location of transcriptomic information within tissue sections [8]. These methods have been developed by academic groups [9–11] and commercially, by 10X Genomics [12, 13]. Despite considerable progress that has been made, important challenges remain, and scRNA-seq protocols continue to have substantially increased levels of nuisance variation relative to bulk RNA-seq. Capture efficiency (percentage of mRNA molecules in the cell lysate that are captured and amplified), amplification bias (non-uniform amplification of transcripts), and sequencing efficiency (rate at which complementary DNA (cDNA) fragments in a library are sequenced) are major contributors to technical variation.

In general, this technology allows us to address several types of biological questions such as identifying and describing cellular subpopulations, understanding stochasticity and the kinetics of gene expression as well as developing finer descriptions of gene-regulatory networks. Moreover, it can be used to reliably identify and explore the role of allele specific gene expression and variability at the transcript level among these phenomena [14]. Nonetheless, substantial computational challenges arise from these experiments and the subsequently generated data [15]. Not only existing methodologies used for analysis of bulk expression experiments need to be adapted to the requirements of scRNA-seq, but new custom methods have to be developed. Experimental design considerations and analysis pipelines are also affected by the various protocols and platforms available for this technology. Additionally, several new types of artefacts have to be taken into account. From a computational perspective, this is an opportunity to expand the field of data analysis by creating strategies that deal with the complexity of single cell high-throughput biology, such
that these aims have preceded the emergence of a new field named single-cell data science (SCDS) [16]. Ultimately, the goal is to provide the necessary tools that can successfully solve novel biological and medical problems. Thus, playing a crucial role on state-of-the-art multidisciplinary research.

1.2 Approaches for scRNA-seq

Over the past 5 years, scRNA-seq has become the most commonly used approach for assaying single-cell gene expression profiles. There are two broad sets of methods for applying scRNA-seq, “plate-based” and “droplet-based” (Figure 1.1). Each approach has its own advantages and disadvantages. Plate-based methods tend to provide higher-quality libraries at the cost of lower cellular throughput, processing hundreds or thousands of cells compared to the hundreds of thousands that droplet methods can process. Two popular protocols for each general approach are described below.

Fig. 1.1: Single-cell library preparation summary. Schematic description of the two primary methods for generating single cell transcriptomic profiles are droplet based and plate based methods. Reproduced from Jonathan A Griffiths et. al., 2018 [17].
1.2.1 **Smart-seq**

Switch mechanism at the 5’ end of RNA templates sequencing (Smart-seq) was developed as a single-cell sequencing protocol with improved read coverage across transcripts [18]. Complete coverage across the genome allows the detection of alternative transcript isoforms and SNPs. Three consecutive versions of Smart-seq have been developed: Smart-Seq, Smart-seq2 and Smart-seq3 [5, 19, 20]. Several improvements have been made to the protocol throughout these versions. The latest improvement combines full-length transcriptome coverage with a 5’ unique molecular identifier RNA counting strategy that enables *in silico* reconstruction of thousands of RNA molecules per cell. In general Smart-seq works as follows: cells are lysed, and the RNA is hybridized to an oligo(dT)-containing primer. The first strand of the cDNA is synthesized with the addition of a few untemplated C nucleotides. This poly-C overhang is added exclusively to full-length transcripts. An oligonucleotide primer is hybridized to the poly-C overhang and used to synthesize the second strand. Full-length cDNAs are PCR-amplified to obtain nanogram amounts of DNA. The polymerase chain reaction (PCR) products are purified for sequencing. The protocol can also be combined with fluorescence-activated cell sorting (FACS) which stains cells with fluorophore-conjugated antibodies in order to facilitate separation from a heterogeneous suspension. Thus, allowing to “index sort” using FACS to isolate individual cells with known characteristics (e.g., selected marker expression), and record their positional location within an assay plate.

1.2.2 **MARS-seq**

Massively parallel RNA single-cell sequencing (MARS-seq) was developed at the Weizmann Institute of Science and was introduced in 2014 as an alternative for performing single-cell RNA-sequencing experiments. It assembles an automated experimental platform that enables RNA profiling of cells sorted from tissues using flow cytometry [4]. In MARS-seq’s protocol, single cells are sorted into cell capture plates, containing cell lysis solution. Capture plates are prepared with an automated liquid handling platform. Sorting is then performed using a FACS, which is very fast and can be easily scaled up to hundreds of cells. The FACS instrument and flow cytometrist then need to be able to consistently and accurately sort cells into the center of the well, thus ensuring that each cell is immersed in the lysis buffer. Importantly, this also enables enrichment of particular cells of interest if they are fluorescently labeled. In brief, MARS-seq’s protocol for library preparation is based on Poly-A tailing plus second strand synthesis and *in vitro* transcription (IVT) pre-amplification reaction.
1.2 Approaches for scRNA-seq

1.2.3 **Drop-seq**

Drop-seq is a system that allows making thousands of scRNA-seq libraries in parallel inexpensive easy experiments. This technology uses droplets, which already have been used to study gene expression and copy number variation. Droplets provide a way of scaling molecular reactions in such way that it is possible to perform thousands of thousands of molecular reactions in a tiny micro-centrifuge tube. Scalability is increased using microfluidics technology. Drop-seq allows to do experiments in the range of 10,000 to 20,000 cells. The challenge with droplets is that they do not stay in one place and they do not have stable addresses. Then, a different DNA barcode is added to the RNA molecules of each cell. This is done using beads that deliver DNA barcodes into droplets. Oligos are synthesized directly on the beads. To make the barcodes, beads are randomly split into groups, then a different base of DNA is added to each group and beads are pooled again. Repeating this splitting and pooling process 12 times produces 16 million different barcoded beads. A microfluidic device delivers these beads and cells together into droplets [10, 11].

1.2.4 **10x Genomics**

The 10x Genomics single cell chromium technology can capture molecular readouts of cell activity in multiple modalities, including gene expression, cell surface proteins, immune clonotype, antigen specificity, and chromatin accessibility [12, 21, 22]. The workflow for generating a single cell gene expression library, builds on a technology that combines barcoded sequences with high resolution partitioning of the input into emulsions on a microfluidic chip. A single cell or nuclear suspension sample is prepared and potentially labelled with the featured barcoded oligo conjugator analytes of desire. The gel beads and cell mixed solutions are loaded into separate input wells of the microfluidic chips and oil is added to the recovery well. During the chromium run, cells and gel beads are combined into a single aqueous stream in a channel within the chip. This aqueous solution flows through the channel and passes the interface of the recovery well and the channel containing cells in gel beads. Droplets are formed at this interface, once formed the gel beads in emulsions (GEMs) are collected in the recovery well. Most GEMs will contain a 10X gel bead, while a subset will also contain a single cell nucleus. Upon the generation of GEMs, the gel beads are dissolved and the cells nuclei lysed realising the oligos and mRNAs into the solution. Within the GEMs a reverse transcription reaction occurs, which converts mRNA into 10X barcoded cDNA. Then, the barcode fragments are pooled together in a final library construction step that is performed in bulk to be ready for sequencing. Once sequenced, the 10X barcodes are
used to associate individual reads back to the individual partitions and to each cell or nucleus. The resulting output are scRNAseq profiles.

The Single cell 3’ gel beads contain many copies of the three types of oligos. The components of these oligos are a partial TruSeq read 1 sequence, a 10X barcode which identifies the particular cell or nucleus in which a transcript (or a feature) is found. There are approximately 3.6 million defined 10X barcode sequences. A unique molecular identifier which is used to identify the particular molecule that is present in your library. Essentially, these barcode is used to identify each unique cDNA there is synthesized from each poly-adenylated mRNA molecule. Lastly, a poly(dT) stretch that enables to capture poly-adenylated mRNA molecules. Figure 1.2 shows an schematic representation of all library preparation steps explained above. To enable the feature bar coding technology two additional oligos are present which contain capture sequences for features of interest.

### 1.2.5 SPLiT-seq

Split-pool ligation-based transcriptome sequencing (SPLiT-seq) is a method that enables profile single-cell transcriptomes without requiring the physical isolation of each cell by allowing labelling RNA molecules with cell-of-origin information using combinatorial indexing [23]. This method removes the need for any specific single-cell instrumentation, it uses basic laboratory equipment, and implements Drop-Seq library preparation after cell RNA is labelled. Across a set of split-pool rounds, fixed cells or nuclei are randomly distributed into wells, and transcripts are labeled with well-specific barcodes. Barcoded RT primers are used in the first round. Second and third-round barcodes are appended to cDNA through ligation. A fourth barcode is added to cDNA molecules by PCR during sequencing library preparation. In order to implement this methodology cells are required to be fixed (which can be challenging in some circumstances). Thus, opening the possibility of single cell transcriptome profiling from formalin-fixed, paraffin-embedded tissue.

### 1.3 The role of scRNA-seq in developmental biology

Cell fate decisions are influenced by both, internal and external factors, meaning that this process is not cell-to-cell independent. However, cell fate commitment is ultimately defined individually such that during embryonic development leads to an exponential diversification into highly heterogeneous populations. Thus, developmental biology has naturally leveraged the arise of scRNA-seq technologies. Significant progress has been made on identification of novel cell types, revision to established hierarchies and identification of novel regulators...
1.3 The role of scRNA-seq in developmental biology

Fig. 1.2: 10X Single-cell Chromium v3 library preparation summary. A. Gene expression polyadenylated mRNA is directly captured by the corresponding gel bead oligo by reverse transcription and template switching. B. cDNA amplification of poly-adenylated mRNA. C. Library construction of poly-adenylated mRNA. D. The completed library fragments generated from the 3' gene expression workflow each contain a P5 adaptor, a TruSeq read 1 sequence, a 10X barcode and unique molecular identifier, a poly(dT) sequence, a TruSeq read 2 sequence and an i7 sample index and the P7 adaptor.
of fate choice [24]. *In vitro* stem cell biology systems are studied in more detail as well as benchmarked against their *in vivo* counterparts [25]. From an evolutionary perspective, new insights are brought by cross-species single-cell transcriptomic analyses of developmental processes [26]. Furthermore, the efforts to create single cell genomic atlases from a wide variety of tissues have challenged paradigms of cell types and cell states [27]. Lastly, the integration of scRNA-seq with other technologies has expanded the scope towards spatio-temporal profiling, and multiple types of simultaneous measurements (multi-omics) are providing the means for an unprecedented reconstruction of developmental processes.

1.3.1 **Allele specific expression**

The concept of allele-specific expression (ASE) refers to the phenomenon that occurs in diploid or polyploid genomes, where two or more alleles of a gene has an imbalanced expression. To assay ASE at the single-cell level, most of these studies analysed scRNA-seq data generated on full-length transcript platforms from hundreds or thousands of cells (e.g., Smart-Seq2 [19]), with few exceptions that have even ventured into 10X Genomics droplet based protocols such as [28]. The first scRNA-seq study of ASE used early-stage mouse embryos (up to the blastocyst stage) and several adult tissues [29], observing a high rate of monoallelic expression (12-25%) for even highly-expressed autosomal genes. Cells in the same embryo expressed different genes monoallelically, implicating stochastic behaviour in deciding which alleles are expressed in individual cells.

Most genes display expression from a specific allele chosen apparently at random, such phenomena is referred as autosomal random monoallelic expression (aRME). The application of scRNAseq to clonal cell populations allowed the dissection of clonal and dynamic aRME, showing that less than 1% of genes showing aRME had conserved behaviour among clones, with most aRME observed dynamically [30].

ASE has also been a useful tool for studying X chromosome inactivation (XCI), the process by which the dosage of X chromosome genes is controlled between sexes in mammals. Experiments in both mouse[31] and human [32] showed that the process is asynchronous across cells, and that gene expression from the silenced X chromosome is gradually and uniformly reduced. A noticeable difference between the two is that *Xist* is biallelically expressed during XCI in humans, and monoallelically expressed in mice.

1.3.2 **Lineage tracing**

Beside the computational methods for differentiation trajectory reconstruction (addressed further in Section 1.6.6 and 1.7), engineered *in vivo* systems are emerging to characterize
the lineage history and transcriptomic states of tissues in different organisms. The advent of clustered regularly interspaced short palindromic repeat/CRISPR associated protein 9 (CRISPR/Cas9) (which enables researchers to edit parts of the genome by removing, adding or altering sections of the DNA sequence [33]) has led to the development of a variety of lineage tracing tools that use errors from non-homologous end-joining DNA repair (also known as scars) to generate a high diversity of unique and heritable DNA barcodes. Initially, CRISPR-based systems required embryonic injection of editing reagents [34]. Modified variants of this system that use expressed barcodes have allowed for the simultaneous measurements of single-cell gene expression levels and lineage tracing [35–37]. More recent protocols rely on the editing of a DNA-inserted barcode: endogenously expressed Cas9 (with an appropriate guide RNA) that progressively and randomly alters this barcode, leaving permanent sequence changes that are inherited by daughter cells. The cell may transcribe the barcode, amplifying its presence within the cell, from where the sequence can be read out by amplicon sequencing or RNA-seq [38, 39]. Both of these approaches use constitutively expressed Cas9 and use multiple target arrays (barcodes) to generate diversity. Nonetheless, new embryonic manipulations are required to generate mice every time, and the resulting mice are impractical for breeding given the high number of randomly inserted transgenes. Therefore, these models are unsuitable for lineage analysis of adult tissues, an approach that circumnavigates this challenge is CRISPR array repair lineage tracing (CARLIN) [40], an engineered mouse line that genomically encodes all the editing components, which can be activated in an inducible manner and thus, it works across adult mouse tissues. An alternative for lineage tracing without CRISPR/Cas9 technology, involves the use of somatic mutations in mitochondrial DNA (mtDNA) identified from either scRNA-seq or assay for transposase accessible chromatin sequencing (ATAC-seq). It has been described that somatic mtDNA mutations can be leveraged as natural genetic barcodes and be used as highly accurate clonal markers to infer cellular relationships [41]. In fact, also single cell whole genome sequencing has been leveraged to perform lineage tracing of differentiation processes [42].

### 1.3.3 Spatially resolved transcriptomics

The spatial context is necessary for understanding cell and tissue interactions during development. However, when using scRNA-seq this information is lost during cell dissociation. Likewise the dynamical components of development, the computational reconstruction of space from only the transcriptomic profiles is a complicated problem. There are notably few computational methods built with this aim [43–45]. Upon recent technological advances, high-throughput image-based single-cell transcriptomics became a reality.
In general, spatial techniques can be classified into those that involve gene expression analysis on microdissected tissues and those that involve in situ hybridization, in situ sequencing, in situ capturing, and computational reconstruction of spatial data [46]. Single-molecule fluorescence in situ hybridization (smFISH) was the pioneer of the hybridization-based approaches with spatial techniques [13], while the applications to fixed tissue sections were inspired by [47] which led to the Visium Spatial Gene Expression technology offered by 10x Genomics, whereby fixed, stained tissue is imaged and then permeabilized. The released mRNAs move and attach to an array beneath the tissue with barcoded oligonucleotides, which fixes them in the position they had in the tissue. After reverse transcription the tissue is enzymatically removed but spatially barcoded complementary DNA molecules attached to the oligo array remain. Then, after the sequencing of cDNAs the positional barcodes provide spatially resolved transcriptomic information at a resolution of tens of cells (100 $\mu$m).

Further on, the integration of scRNAseq and sequential fluorescence in situ hybridization (seqFISH), allowed to profile the expression of hundreds or thousands of genes within single cells whose spatial location is preserved [48–50, 45]. The transcripts are fixed in a tissue’s cells. With each round of hybridization, probes are removed for the next hybridization round with a different fluorophore. The sequence of fluorescence signals delivers spatially resolved transcriptomic data. The method uses sequential rounds of hybridization to barcode mRNAs with a set of FISH probes that carry a fluorophore. These protocols are now being applied to study embryonic development and new atlases that integrate information from those constructed using common scRNA-seq approaches have been generated [51]. For instance, existing cell type annotations can be transferred between datasets by mapping the corresponding gene expression profiles. Slide-seq, simplifies the procedure to more accessible experimental set ups by developing a protocol to transfer RNAs to beads, where the smaller the beads the better resolution achieved (approximately 10$\mu$m) [52]. Then, multiplex error-robust fluorescence in situ hybridization (MERFISH) incorporates barcodes to RNAs during several hybridization rounds. Non-fluorescent targeting probes bind the mRNA and fluorescent readout probes hybridize to the targeting probes. The readout sequences yield a barcode distinct to each mRNA. The sequential rounds of hybridization, the barcoding and error correction make it possible to capture many genes. The challenge with hybridization rounds, however, is the accumulation of errors [49]. Lastly, Seq-Scope achieves submicrometer resolution spatial barcoding [53]. Its technology relies on the solid-phase amplification of a random barcode molecule [54] and it has a resolution of of 0.5–0.8 $\mu$m, far superior to previous technologies and comparable to an optical microscope. These technologies hold the promise of uncovering knew knowledge in developmental biology, and has already challenged existing ideas of development. It was actually named the method of
1.3 The role of scRNA-seq in developmental biology

year: 2021 [55]. Furthermore, the computational methods required for the analysis of this data have are a crucial area of research.

1.3.4 Multi-omics

As useful as single cell transcriptomics has been for the study of embryonic development, it just represents one level of information among several others (e.g., protein levels, DNA mutations and epigenetic features). Hence, in addition to the spatio-temporal component a multi modal integrative (multi-omics) approach is necessary for achieving a comprehensive understanding of developmental processes. Multi-omics approaches at single cell resolution aim to measure two or more features simultaneously in a cell. There are several instances of these approaches: Triple omics sequencing (Trio-seq) [56], gDNA-mRNA sequencing (DR-seq) [57], Genome and Transcriptome sequencing (G&T-seq) [58], Methylome and Transcriptome sequencing (M&T-seq) [59], simultaneous isolation of genomic DNA and total RN (SIDR) [60], TARGET-seq [61] and even nucleosome, Nucleosome, Methylation and Transcription sequencing (NMT-seq) [62]. These approaches are limited by a relatively small number of cells. When increasing throughput capabilities, multi-omics protocols tend to sacrifice sequencing sensitivity and thus, the resulting data is sparser. Such as the case of SNARE-seq [63] or 10X Genomics recently introduced: multiome assay (Chromium Single Cell Multiome ATAC + Gene Expression) which allow profiling open chromatin and mRNAs from the same single nuclei using a microdroplet platform and barcoded beads. Another recent and promising method for ultra-high throughput profiling is Paired-Tag that allows simultaneous measurements of histone modifications and transcriptome in single cells to produce cell-type-resolved maps of chromatin state and transcriptome in complex tissues [64]. This, through the extension of parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing (Paired-seq) [65], for co-assay of open chromatin and gene expression in single cells, by adapting the cleavage under targets and tagmentatio (CUT&Tag) strategy [66]. The computational methods associated to these analyses are benefited from the large numbers of cells, but also challenged by much higher levels of noise and scalability issues.

Lastly, recent approaches to assay the proteome include Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) [67] which extends the concept of cell hashing (i.e., using oligo-tagged antibodies against highly expressed surface proteins to place a sample barcode on each single cell) to a panel of antibodies to cell surface proteins, allowing to measure both transcript and protein levels, via the frequency of the conjugated oligonucleotide, in the same cells. Also, RNA expression and protein sequencing assay (REAP-seq) [68] which is a similar technology with a different construct of the barcode conjugated to the bead. Unlike
the CITE-seq and REAP-seq methods, that are only able to target cell surface proteins, an alternative method, single-cell RNA and immunodetection (RAID), can detect intracellular proteins or phosphorylated proteins together with mRNAs [69]. A Multi-omic atlas of mouse development generated with scNMT is already available [70]. Furthermore, this atlas is also a time course experiment. It revealed that the initial exit from pluripotency coincides with the establishment of a global repressive epigenetic landscape, followed by the emergence of lineage-specific epigenetic patterns during gastrulation. It showed that the ectodermal epigenetic landscape is established as early as in the Epiblast, and that it is followed by the mesoderm and ectoderm in which cells undergo through coordinated epigenetic rearrangements at enhancer marks [70].
1.4 An overview of Mouse development

The mouse (in particular *Mus musculus*) has always been a good embryological model, easily (litters 8-20) and quickly (21d). The availability of a wide variety of experimental tools as well as high quality bioinformatic resources makes it a natural choice for several experimental designs. The study of gene knockout (KO) impact during development has been largely benefited from mouse models. Despite important differences between mouse and human development, it is still a good proxy for the exploration of a large range of molecular mechanisms in mammals. In mammals, the meeting of the oocyte and sperm, and subsequent fertilization, take place in the ampulla of the oviduct. During the following days, the embryo travels down the oviduct to the uterus, and prepares for implantation. In this context, cell fate decisions occurring during embryonic development can be divided into two broad events pre and post-implantation.

1.4.1 Pre-implantation

After fertilization, a diploid cell resulting from the combination of two haploid gametes is formed. This diploid cell, also known as the zygote undergoes a series of cleavage divisions. After approximately four days of this "cleavage stage" the embryo compacts, at which time it is referred to as a "morula" [71, 72]. A fluid cavity then forms inside the embryo, at which point it is called a "blastocyst". Cells in the blastocyst differentiate into either the inner cell mass (which will develop into the fetus) or the trophectoderm (which will go on to become the embryonic part of the placenta). The embryo then implants into the endometrium of the uterus. Pre-implantation embryo development occurs with almost no change in mass. Thus, all these early developmental events largely consist of a reorganization of the same material that was present in the unfertilized egg. Figure 1.3 shows an schematic representation of this process.

1.4.2 Post-implantation

Following implantation at embryonic day 4.5, the mouse embryo starts proliferating extensively and cells undergo through drastic morphogenetic changes. By E5.5, the embryo has adopted a cup-like shape, and its distinct lineages have diversified: the trophectoderm has given rise to the Extra-embryonic ectoderm (ExE ectoderm) and the ectoplacental cone, which are progenitors of the placenta; and the primitive endoderm has segregated into the parietal endoderm, which is in direct contact with the maternal tissue, and the Visceral endoderm (VE), which surrounds the embryo [73]. *Hex* expression remarks the emergence of
Fig. 1.3: Schematic representation of mouse embryonic pre-implantation across time (Embryonic day (E)). Following fertilisation, the zygote divides symmetrically until the 8-cell stage. The embryo then becomes compacted and forms the so-called "morula". During the 8-to-16 and the 16-to-32 cell stage transitions, some cells will divide symmetrically and will become specified into the outer cell layer, called Trophectoderm (TE); other cells will undergo asymmetric divisions, where one of the daughters will become TE and the other will become part of the inner cell mass (ICM), which will give rise to the embryo proper. Following this first lineage specification, the ICM undergoes a second fate choice and becomes segregated into the Epiblast (EPI) and the Primitive Endoderm (PE) at the 32-to-64 cell stage, the latter giving rise to the yolk sac later on. At the same time, the embryo also undergoes a process of cavitation, and the blastocoel will form. Finally, at the blastocyst stage, spatial cell rearrangement gives rise to Epiblast cells surrounded by outer primitive endoderm. Inspired by Pijuan-Sala, Blanca. Created with BioRender.com.

Distal visceral endoderm (DVE) marking the first antero-posterior asymmetry in the mouse embryo. The DVE migrates proximally to the junction between the prospective anterior epiblast and ExE Ectoderm, giving rise to the Anterior visceral endoderm (AVE) expressing antagonists of the Nodal signalling factor, Cerberus and Lefty1. The secretion of these factors allegedly contributes to the formation of a structure called the Primitive streak (PS) in the Epiblast of the posterior side of the embryo at E6.5 [74]. Defining thus, the beginning of gastrulation (Figure 1.4).

During gastrulation, Epiblast cells undergo intensive cell intermingling and migrate towards the embryonic proximo-posterior side [75, 76]. This migration, together with a potential contribution from the AVE, leads to the emergence of the PS at this side, which is marked by a high expression of the early mesodermal marker Brachyury (encoded in the T gene) and Wnt3 [77, 78]. Upon their arrival at the PS, Epiblast cells undergo an epithelial-to-mesenchymal transition (EMT) to ingress through the streak and then become either mesoderm or endoderm [79]. As development progresses, the PS extends anteriorly until reaching the most distal part of the embryo, where the node is formed [73]. While the first cells to egress through the proximo-posterior region of the PS will mainly become part of the extra-embryonic mesoderm, cells from the mid-primitive streak will form the lateral plate mesoderm as well as the paraxial and cardiac mesoderm. Finally, cells that egress at the anterior region of
the PS will form midline axial mesendodermal tissues, including definitive endoderm, the notochordal plate and the node[75]. Importantly, not all epiblast cells will egress through the PS; the remaining group of cells will, instead, adopt an ectodermal program and, consequently, develop into tissues such as the brain, the spinal cord and the skin [75, 80, 81]. As gastrulation proceeds, cells that have passed through the PS or that remain in the Anterior epiblast will subsequently start diversifying into several progenitors of the major organs of the mouse embryo, marking the start of organogenesis (Figure 1.5)

Fig. 1.4: Diagram illustrating gastrulation. Anterior-posterior axis patterning positions the primitive streak prior to gastrulation. After implantation, the embryo transitions from using the proximal-distal axis to the embryonic anterior-posterior axis. The establishment of this axis depends on a sequence of events including specification of the DVE at E5.5 and formation and migration of the AVE at E5.75. The primitive streak is restricted to the posterior pole by E6.5 and elongates distally and anteriorly until E7.5. Adapted from Pijuan-Sala, et al., 2019 [82]

Fig. 1.5: Diagram illustrating organogenesis describing the time line of of development following gastrulation. Created with BioRender.com
1.5 **Transcriptomic atlases of mouse development**

The efforts to build cell atlases from different model organisms already include extensive transcriptomic profiling of embryonic development, where detailed non-human models are essential to compensate the limitations arising from ethical reasons [83]. In this context, the transcriptomic landscape of mouse development is perhaps the most complete among mammals, and now covers a wide range of different embryonic stages and cell types. Time-course experiments of embryonic days E3.5-E6.5, E6.5-E7.5, E6.5-E8.5, E9.5-13.5 and E10.5-E15.0 have been respectively carried out in [84, 85, 82, 86, 87]. Other efforts have been conducted to allow finer reconstruction of specific organs such as the Brain[1], Gut tube [88, 89] and Limbs [87]. Moreover, some approaches have simultaneously measured transcription, methylation and chromatin accessibility at single cell resolution [70].
1.6 Computational analysis of scRNA-seq data

1.6.1 Low level analysis

Measuring transcripts expression

Published scRNA-seq protocols have reported limits of detection of mRNA molecules equivalent to a capture efficiency of around 10%, and all current methods use amplification, either by PCR or by IVT. To correct for amplification bias, molecules can be directly counted through the use of unique molecular identifiers (UMIs) [90, 91]. Thus, the vast majority of protocols have been now adapted for the use of UMIs. Additionally, UMIs provide a direct and quantitative measure of gene expression [90], as molecule counting not only corrects for amplification induced artefacts but also provides an absolute scale of measurement with a defined zero level. In contrast, standard RNA-seq uses relative measures such as fragments per kilobase per million mapped reads (FPKM), which mask differences in total mRNA content. For example, a gene may be up regulated in terms of FPKM and have a decrease in absolute expression levels if the total mRNA content also changes. Thus, an absolute scale of measurement is crucial for interpreting transcriptional dynamics in single cells. The vast majority of protocols with UMIs sequence only a fragment of each molecule (from either the 5’ or the 3’ end of the transcript). Hence, correcting for transcript length during normalization is unnecessary. Nevertheless, as a consequence of this protocol UMIs can hardly be used in studies of isoforms or allele-specific expression [15]. A few long read protocols aimed to address this problem have been developed (e.g., [92]), however, to say the least, these methods have still space for improvement.

The computational process for producing measures of transcript expression works as follows: raw reads are de-multiplexed and quality control (QC) reports are produced (e.g. using the software FastQC [93]). Sequence adapters, barcodes, UMIs and poly-A sequences are removed before mapping against a reference genome. Read alignment is done using programs such as Bowtie2 [94] or STAR [95], as these are able to perform spliced alignment. All barcodes and UMIs are included in sequence names in order to keep track of them. Thus, UMIs can be assigned to the corresponding cells and collapsed in further steps. Then, UMI counts are designated to genes according to provided gene intervals (Genome annotation). Notably, barcode decomposition is not trivial-particularly for the random sequences of UMIs because sequencing errors can alter their observed sequences. Methods have been developed to account for this by predicting which barcodes have arisen by error and which truly existed within the sample [96]. The resulting file is an expression matrix of UMI counts with cells in columns and genes in rows.
**Cell calling**

Particularly for droplet methods, cell calling refers to identifying which of the hundreds of thousands of possible cell barcodes were actually associated with a viable cell, and which were either empty, although likely still containing cell-free RNA from the original pool of cells. A “standard” approach is included in 10X Genomics’ CellRanger software package, where a threshold on the number of molecules associated with a barcode is used to call cells. The threshold is determined dynamically for each sample, so that sequencing depth does not affect cell calling. However, the range of barcode UMI counts varies considerably, even among captured cells. Some droplets may, by chance, capture a large amount of cell-free RNA (or otherwise capture RNA more efficiently) and be called as cells, and some droplets that do contain real cells may only contain a small number of molecules, and therefore be rejected. This is especially problematic for samples where different cell-types may contain different amounts of RNA. An alternative approach to cell-calling that addresses these issues is emptyDrops [97]. This method considers a vector composed of the gene counts summed across barcodes associated with only very few UMIs (e.g., <100 UMIs), which is assumed to represent the pool of cell-free RNA. Any barcodes with a transcriptional profile that significantly deviates from this distribution may then be retained as a real cell, irrespective of its library size.

**Quality Control**

Proper detection and filtering out technical artefacts before proceeding with the analysis. Cell QC is commonly performed based on three QC covariates: the number of counts per barcode (library size), the number of genes per cell barcode (library complexity), and the fraction of counts from mitochondrial genes per cell barcode (mitochondrial count fraction). These outlier barcodes can correspond to dying cells, cells whose membranes are broken, or doublets (when two or more cells are mistakenly considered as a single cell). The distributions of these QC covariates are examined for outlier peaks that are filtered out by setting thresholds. Doublet identification is arguably the most challenging task, particularly for droplet-based methods. There is a collection of heuristic as well as sophisticated approaches commonly used to filter out potential doublets. These include rejecting cells expressing sets of biologically mutually exclusive markers (e.g., X-inactive specific transcript (Xist) and Y-chromosome genes) [98], and by identifying small clusters composed of cells with large library size whose expression profiles correlate strongly with at least two other clusters in the dataset [99]. Several algorithmic approaches are also available, by either identifying cells that resemble simulated doublets as scrublet and DoubletFinder [100, 101] or by considering potential
doublet clusters as convolutions of other singlet clusters (DoubletDecon [102]). In very large datasets, it is probably wise to combine different approaches.

**Confounding factors**

Like any other genomic sequencing technique, scRNA-seq is susceptible to the influence of confounding effects. For instance, systematic differences between experimental batches must be corrected before any downstream analysis in which relevant technical or biological interpretations are performed, emphasising the importance of good experimental design [103]. Given the logistics of how scRNA-seq experiments are performed, and the fact that this technology is being used to discover new cell types, batch effects are of particular concern. Batch-effect correction strategies that are commonly applied to bulk RNA-sequencing studies (e.g., ComBat-seq [104]) are often ineffective for scRNA-seq data due to compositional differences between different batches. Thus, specialised batch-correction methods have therefore been developed to address these effects. A large fraction of them are inspired on the concept of mutual nearest neighbours (MNN) [105], Seurat3 [106, 107], BBKNN [108] and Scanorama [109]), broadly explained in Section 2.4.7. Alternative approaches include Harmony which applies an iterative correction factor between centroids of batch specific and overall fuzzy clusters [110] or LIGER an algorithm that delineates shared and dataset-specific features of cell identity using an integrative non-negative matrix factorization (iNMF) to learn a low-dimensional space in which each cell is defined by one set of dataset-specific factors, and another set of shared factors [111]. Moreover, true biological differences may produce signals that could be considered confounding factors. In particular, cell size (as reflected by total RNA content) often manifests itself in the number of detected genes in each cell [112]. This measure of cell size may derive from true biology, from technical effects (e.g., variable efficiency of reverse transcription), or from both. It is during the crucial step of normalisation, where differences due to sequencing depth and total RNA content are corrected. While there is no such thing as a gold-standard method for normalisation of scRNA-seq data, there are frameworks and metrics that help assessing their performance (see scone [113]). The addition of precisely quantified exogenous RNA species (known as External RNA Control Consortium (ERCC) spike-in genes) to each cell’s lysate allows the estimation of absolute amounts of RNA [114], and the use of UMIs helps to reduce bias introduced by PCR or IVT amplification, as duplicate reads may be confidently collapsed down to a true origin molecule. When well-behaved and representative synthetic spike-ins or UMIs are available, further refinement is possible. However, the use of ERCC spike-in controls is not practical for droplet-based assays. Despite the existence of highly sophisticated methods for normalisation and batch-effect correction, a careful experimental design is a critical aspect of successful use
Introduction

of scRNA-seq. In fact, whether the biological effects of interest are completely confounded with unwanted technical effects, no statistical method will be able to extract meaningful signal from the data [112]. Other biological factors such as cell-cycle stage can also mask the signal of interest [115]. In this context, computational strategies such as scLVM or Seurat3 can be used for identification and removal of these effects [116, 117].

1.6.2 Selection of highly variable genes

Feature selection in gene expression space is a step with critical impact on downstream analysis. It refers to excluding uninformative genes such as those which exhibit no meaningful biological variation across samples [118]. The simplest approach to feature selection is to select the most variable genes based on their expression across the population by computing either their coefficient of variation, which is the mathematical definition of transcriptional noise [119]. More sophisticated approaches involve modelling the per-gene variance. That is, decomposing the total variance of each gene into its biological and technical components by fitting a trend to the endogenous variances [120]. The fitted value of the trend is used as an estimate of the technical component, and it is subtracted from the total variance to obtain the biological component for each gene. If spike-ins controls are available, these can be used to fit the trend instead. Furthermore, uninteresting factors can be blocked by performing the trend fitting and decomposition within each block before combining the statistics across blocks for output. Alternatively, this task may be accomplished by ranking genes using the deviance, which quantifies how well each gene fits a null model of constant expression across cells. Unlike other strategies, which are sensitive to normalization, ranking genes by deviance operates on raw UMI counts. An approximate multinomial deviance statistic can be computed in closed form [118].

1.6.3 Dimensionality reduction and visualisation

A collection of mRNA levels for a single cell is called an expression profile and is often represented mathematically by a vector in gene expression space. This is a vector space has dimension equal to the number of genes, with the value of the $ith$ coordinate of an expression profile vector representing the number of copies of mRNA for the $ith$ gene. As the number of cells profiled reaches large numbers, the analysis might face limitations in terms of computational resources. More importantly, another problem arises due to the high-dimensional nature of the data (as the number of genes is the number of variables); this effect is referred to as the “curse of dimensionality”. In such a situation, the variances among samples become large and sparse, statistical power gets compromised and clustering
analysis might lead to meaningless results (e.g. [121]). The large number of dimensions also makes visualisation and interpretation of results much more challenging. An effective solution to these problems is to summarise the data in a reduced number of dimensions. The most common approach is to compute a principal component analysis (PCA), and retain the top $N$ components (where components are ranked by variance explained/eigenvalue, and $N$ is defined by the researcher according to the scope of a dataset). The formal definition of PCA is formulated as the solution to an eigen-problem or, alternatively, from the singular value decomposition (SVD) of the (centred) data matrix [122]. PCA can be based on either the covariance matrix or the correlation matrix. Preserving as much variability as possible translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other. Under the assumption that the biological signal in the data is considerably stronger than any technical noise, the use of principal components (PCs) may also help to remove unwanted sources of variation that should accumulate in the lower-ranked PCs. In principle, these PCs may then be used for most downstream analyses such as clustering, as the distances between cells in the PCA space closely approximate the distances in the high-dimension gene expression space. There are many variations of PCA as well as a diverse range of sophisticated methods that have been developed for the purposes of dimensionality reduction. In particular for scRNA-seq experiments, diffusion maps have proven useful for summarising continuous variation, such as the case of developmental processes [123] (see Section 1.6.6). In this context, distances between two cells are reframed as probabilities of transitioning between them along a random walk via other cells. One can therefore consider these transition probabilities as distances along the manifold of cells in the high-dimensional gene expression space (or PCA space) rather than the shortest distance between two points (i.e., Euclidean distance).

The complexity of single cell transcriptomic landscape is hardly summarized in two or three dimensions even after the application of PCA or diffusion maps. Thus, human-readable visual representation requires strategies that are able to capture larger orders of complexity using few dimensions. Intuitively, a non-linear manifold would be a better approach for this. However, it is practically not possible to preserve accurate distances for all distance scales when projecting high-dimensional data into such a small subspace. Historically, scRNA-seq data analysis has favoured the preservation of local structure, that is to group neighbouring data points together to prioritize visualization of heterogeneity in the data. For instance, the $t$-distributed stochastic neighbour embedding ($t$-SNE) [124]. The $t$-SNE algorithm comprises two main stages. First, it constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects are assigned a higher probability while dissimilar
points are assigned a lower probability. Second, it defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map. This approach has been widely used in scRNA-seq. However, sacrificing global distances limits interpretation of cluster similarities when these fall apart from each other, and new methods have been developed to overcome some of this limitations. That is the case of uniform manifold approximation and projection for dimension reduction (UMAP) [125]. Both, t-SNE and UMAP use graph layout algorithms to arrange data in low-dimensional space. Following a similar concept as for t-SNE, UMAP constructs a high dimensional graph representation of the data and then optimizes a low-dimensional graph to be as structurally similar as possible. However, UMAP is not only faster and more scalable. It is also, much more sophisticated in terms of mathematical formality. In order to construct the initial high-dimensional graph, UMAP builds a topological space called "fuzzy simplicial complex". This is essentially a representation of a weighted graph, with edge weights representing the likelihood that two points are connected. To determine connectedness, UMAP extends a radius outwards from each point, connecting points when those radii overlap. The choice of this radius is critical, a very small value will lead to small, isolated clusters, while very a large value will connect everything together. UMAP overcomes this challenge by choosing a radius locally, based on the distance to each point’s $n$th nearest neighbour. It then makes the graph "fuzzy" by decreasing the likelihood of connection as the radius grows. Finally, by stipulating that each point must be connected to at least its closest neighbour, UMAP ensures that local structure is preserved in balance with global structure. The two most commonly used parameters are $n\_neighbors$ and $min\_dist$, which are effectively used to control the balance between local and global structure in the final projection. Another similar approach is referred as Force directed graphs, in particular the algorithm ForceAtlas2 [126]. This algorithm combines an attractive force between neighbouring pairs of points with a repulsive force between all points. A stronger attraction can better represent continuous manifold structures, while stronger repulsion can better represent discrete cluster structures. Thus, it is commonly used for visualizing developmental single-cell transcriptomic data in favour of continuous manifolds. It has been suggested that all these algorithms: t-SNE, UMAP and ForceAtlas2, can be placed onto this attraction-repulsion spectrum, such that UMAP embeddings correspond to t-SNE with increased attraction and ForceAtlas2 yields embeddings corresponding to t-SNE with the attraction increased even more [127].
1.6.4 Cell heterogeneity and clustering

Classification of cells using clusters to represent biologically meaningful groups such as cell types is a common aim in scRNA-seq data analysis. However, there are fundamental questions about this approach that still lack of clear answers such as, do clusters represent appropriately and equivalently all cell stages and cell types? or what is the optimal number of clusters that shall be used to describe a certain degree of cell heterogeneity? On the other hand, clustering is essential and complementary of several analyses. For instance, differential gene expression analysis requires the definition of groups that are not always known a priori. The majority of trajectory reconstruction methods are based on clusters and hierarchies between them. It is a valuable tool for identifying sequencing artefacts (e.g., doublets) and even helps the implementation of batch-effect correction methods as in the case of Harmony (See section 1.6.1). There are numerous methods available for clustering expression profiles, the choice of which in part depends on the scale of data that is to be analysed. For smaller datasets, hierarchical clustering is a popular approach, and provides users with a useful visualisation of distances between cells or clusters via a dendrogram. However, the method scales poorly with the number of cells, both in terms of computation time and memory requirements (e.g., the computation of a cell × cell distance matrix). More contemporary methods use graphical representations of the data, where each node is a cell, and edges connect cells that are sufficiently similar (e.g., K-nearest neighbours (KNN), or cells that share nearest neighbours [128]). This significantly reduces the computational cost as only the closest neighbours of cells need to be retained, rather than all distances to other cells. Also, further computations become more efficient whether sophisticated nearest-neighbor finding algorithms are used (e.g., cover-tree). Perhaps the most popular method at the moment is Louvain designed for large scale networks community detection [129]. However, it was recently claimed that the Louvain algorithm may yield arbitrarily badly connected communities and to avoid this pitfall a novel approach was proposed [130]. While research on clustering methods and applications to scRNA-seq continues, as well as the majority of cell types annotations rely on clustering analysis. Non-clustering based strategies have also been developed in a similar scope. For instance, cellassign is a probabilistic model that takes advantage of previously known gene markers for cell type classification together with the presence of batch effects [131].

1.6.5 Integration of multiple datasets

Integrating and annotating transcriptomic profiles in different contexts (eg. across species, technologies, experiments, time points, etc.) is an elementary task towards the construction
and leveraging of scRNA-seq data, for which a variety of computational approaches have been developed. Batch correction methods inspired on mutual nearest neighbours such as MNN [105], Seurat3 [106, 107], BBKNN [108] and Scanorama [109] can be used to integrate distinct datasets and subsequently transfer annotations from one dataset to another. Likewise, alternative batch correction approaches such as Harmony [110]. The choice of input for these methods is an important factor to consider. While a PCA encodes both shared and non-shared variance, Canonical Correlation Analysis (CCA) captures only shared variation. It has been noticed with concern that sometimes results considerably differ depending on the approach used.

In the context of developmental processes, all these methods face significant challenges mainly due to the difficulty of dealing with highly similar as well as transitional cell populations arising along the trajectories of differentiation. While at the same time, accounting for the presence of systematic confounding factors such as batch effects. Cross-species integration adds an evolutionary component to the problem, which in practice is carried out by matching of homologue genes with the disadvantage of potentially losing relevant information (e.g., due to poorly annotated species or missing homologies). Multi-omics integration is another major challenge as the feature space is not only completely different across orthogonal omics, but also associated with each other in complex ways. A novel approach is multi-omics factor analysis (MOFA) which is essentially a generalization of PCA designed to take as input and arbitrary number matrices and decomposes them into a matrix of hidden latent factors and weight matrices that can be used to assess the contribution to the variance from each latent factor in each data modality as well as visualization of cells in factor space [132, 133].

1.6.6 Developmental trajectories and pseudotime

Developmental processes are inherently dynamic, however, the vast majority of protocols require cell lysis before sequencing which implies that only static snapshots of cell state are measured and therefore, multiple dynamical processes could give rise to observed transcriptomics landscapes [134]. In general, this challenge has led to both innovative experimental and computational methods for trajectory reconstruction. Aside the protocols involving lineage tracing (See section 1.3.2), the main approach to overcome this limitation is to infer a pseudotemporal ordering of the cells according to their similarity in gene expression, such that each cell is assigned a one-dimensional coordinate that reflects its state in the developmental process. The concept of pseudotime was first introduced by [135]. It is a latent (unobserved) dimension which measures the cells’ progress through the transition from one progenitor to all downstream cells. Pseudotime is related to but not necessarily
1.6 Computational analysis of scRNA-seq data

the same as laboratory capture time, thus interpretation of results must be done carefully. It is assumed that this transcriptomic similarity is dominated by the process of interest, and that the more similar two cells are the closer is their stage of progression along this process [136]. First, a similarity measure is defined that quantifies the similarity between the gene expression profiles of two cells. Then, by ordering the cells according to this similarity measure, a hypothetical trajectory (i.e., pseudotime) of the cells’ development can be inferred. In other words, each cell is projected to one point along a one-dimensional process using the information available from its gene expression profile. Thus, despite the fundamental limitation arising form the destructive nature of scRNA-seq, the use of pseudotime makes possible the study of developmental processes such as cellular differentiation or maturation (Figure 1.6).

![Fig. 1.6: The high-dimensional data manifold can be reconstructed via low-dimensional embedding of the data to represent the data manifold in two dimensions (Dim 1 and Dim 2) (Right). Important regions (e.g. root, branching and tip regions; indicated by coloured rings) and pseudotime inference (indicated by arrows) can be identified and visualised on the low-dimensional embedding. Pseudotime direction depends on the choice of the root population. Adapted from [136].](image-url)
1.7 Incorporating time to infer developmental trajectories

There are several computational methods for reconstructing developmental trajectories and it is still a very active area of research. Furthermore, most of these methods are under constant improvement. An extensive benchmarking of trajectory reconstruction methods highlighted that lack of a gold-standard methodology, such that the choice of method should depend mostly on the dataset dimensions and trajectory topology. Also, keeping in mind that the input they require and output models they produce vary substantially [137]. However, the vast majority of trajectory methods rely on the concept of pseudotime (see Section 1.6.6) which is essentially a distance function from the progenitor to all downstream cells. While pseudotime is a useful and widely implemented approach for trajectory inference, incorporating a real notion of time into these methodologies is desired. In this context, there are two very different approaches that were not included in the benchmarking analysis previously mentioned. These are state-of-the-art inference of developmental trajectories are RNA velocity [138] and Waddington-Optimal Transport (W-OT) [139]. These two will be explained in larger detail as they were crucial for this thesis as the whole computational analysis is performed over a time course experiment.

1.7.1 RNA velocity in single cells

The concept of RNA velocity in single cells relies on a dynamical system model describing the mRNA kinetics of unspliced, spliced and degraded mRNA molecules [138]. In contrast to previous approaches, this strategy works under the rationale that the manifolds observed in scRNA-seq data could be used to resemble an explicit time-course of differentiation by extrapolating cell states which can be directly projected using a variety of methodologies (e.g., 1.7). The dynamical system of mRNA kinetics was first introduced in [140] but its application to single cell transcriptomics is relatively recent. The following differential equations describe how the expected number of unspliced mRNA molecules $u$, and spliced molecules $s$, evolve over time.

$$\frac{du(t)}{dt} = \alpha(t) - \beta(t)u(t) \quad (1.1)$$

$$\frac{ds(t)}{dt} = \beta(t)u(t) - \gamma(t)s(t) \quad (1.2)$$

Here, $\alpha(t)$ is the time-dependent rate of transcription, $\beta(t)$ is the rate of splicing, $\gamma(t)$ is the rate of degradation. Under an assumption of constant (time-independent) rates $\alpha(t) = \alpha$,
\( \gamma(t) = \gamma \) and setting \( \beta(t) = 1 \) (i.e. measuring all rates in units of the splicing rate), the rate equations simplify to:

\[
\frac{du(t)}{dt} = \alpha - u(t) \tag{1.3}
\]

\[
\frac{ds(t)}{dt} = u(t) - \gamma s(t) \tag{1.4}
\]

The complete solution to the system is given by:

\[
u(t) = \alpha(1 - e^{-t}) + u_0 e^{-t} \tag{1.5}\]

\[
s(t) = \frac{e^{-\gamma(1+\gamma)}[e^{(1+\gamma)\alpha(\gamma-1)} + e^{\gamma(u_0 - \alpha)\gamma} + e^{\gamma(\alpha - \gamma(s_0 + u_0 + s_0 \gamma))}]}{\gamma(\gamma-1)} \tag{1.6}\]

with the initial conditions \( u(0) = u_0 \) and \( s(0) = s_0 \). This solution can be used to extrapolate mRNA abundance \( s \) to a future timepoint \( t_1 \), under the assumption stated above, by entering the current state of the cell as \( u_0 \) and \( s_0 \), and then computing \( s(t_1) \).

The equations above hold for a single gene. Across all genes, the same equations hold under the same assumptions, but with gene-specific rate constants. Note that setting \( \beta(1) = 1 \) for all genes implies that we assume a common, constant rate of splicing. Moreover, in steady-state populations, where \( \frac{ds(t)}{dt} = 0 \), it is possible to estimate \( \gamma \) of a given gene as the ratio of unspliced to spliced mRNA molecules (again setting \( \beta(1) = 1 \)):

\[
\gamma = \frac{u}{s} \tag{1.7}
\]

\[
\alpha = u \tag{1.8}
\]

The steady-state assumption may be realistic for genes expressed in populations known to be terminally differentiated. However, for genes expressed transiently during development, or in cases where the terminal population was not sampled, the steady-state assumption will fail. Currently, two alternative computational frameworks have been developed based on RNA velocity to analyse scRNA-seq data[138, 141]. These are named veloctyo and scVelo respectively. Essentially, the former works under the the central assumptions of a common splicing rate and the observation of the full splicing dynamics with steady-state mRNA levels. The latter is an extension that overcomes these limitations by solving the full transcriptional dynamics of splicing kinetics using a likelihood-based dynamical model. This approach is more robust but also computationally more extensive.
**Cell trajectories**

RNA velocity can be used to infer the next cell state. Therefore, by correlating the velocity vector with other cells in the landscape it is possible to predict the path of differentiation. The following diagram summarizes the *scVelo* approach for estimating the vector fields that indicate the course of differentiation.

**mRNA splicing dynamics**

![Diagram of mRNA splicing dynamics](image)

**Fig. 1.7:** RNA velocity can be used to infer developmental trajectories as the mRNA splicing dynamics add an additional temporal dimension directly estimated from single cell transcriptomics data. The timescale of cellular development is comparable to the kinetics of the mRNA life cycle, from the transcription of precursor mRNA to the production of mRNA via splicing to mRNA degradation [138]. Adapted from Tritschler et al., 2019 [136].
1.7 Incorporating time to infer developmental trajectories

1.7.2 Waddington-Optimal Transport

Inspired by the Waddington landscape [142] (Figure 1.8.A), the so-called W-OT framework models a developmental process $P_t$ as a generalization of a stochastic process for cell populations evolving in gene expression space according to an unknown temporal landscape of static snapshots from consecutive time points $t_1, t_2, \ldots, t_T$ [139]. Thus, an empirical developmental process $\hat{P}_t$ is time varying distribution constructed from a developmental time course $S_1, \ldots, S_T$, supported on the data $x \in S_k$:

$$\hat{P}_{t_k} = \frac{1}{|S_k|} \sum_{x \in S_k} \delta_x$$  \hspace{1cm} (1.9)

To estimate the time-dependence structure $\gamma_{t, t+1}$ (i.e. coupling between distributions), the method uses optimal transport theory for measuring the total distance that mass must be transported to transform one distribution into another [143, 144]. The resulting coupling estimation also known as transport map $\pi_{t, t+1} \approx \gamma_{t, t+1}$, allows estimating the distribution of descendants and ancestors for a target population of cells. Furthermore, given that the classical formulation of optimal transport does not allow cells to grow (or die) during transportation, such formulation was modified to allow cells to grow at different rates by means of an Unbalance Optimal Transport (UOT) approach [145].

---

Fig. 1.8: Introduction to W-OT. A. Waddington’s landscape model. (From [142]). B. A temporal progression of a time-varying distribution $P_t$ (left) can be sampled to obtain finite empirical distributions of cells $\hat{P}_{t_j}$ at time points $t_1, t_2, t_3$ (right). Over short time scales, the unknown true coupling, $\gamma_{t_1, t_2}$, is assumed to be close to the optimal transport coupling, $\pi_{t_1, t_2}$, which can be approximated by $\hat{\pi}_{t_1, t_2}$ computed from the empirical distributions $\hat{P}_{t_1}$ and $\hat{P}_{t_2}$. Adapted from Schiebinger et al., 2019 [139].
**Transport maps**

The transport map indicates how many and what kind of descendants each cell from time \( t_i \) would have at time \( t_{i+1} \), if the measurement process had not destroyed the cell. Specifically, there is a row for each cell from time \( t_i \) and a column for each cell from time \( t_{i+1} \). Each row gives the descendant distribution of some cell \( x_i \) (Figure 1.8.B). The units of the transport map are 'descendant mass'. For instance a value of 0.1 in the \((x_i,y_j)\) entry means that cell \( x_i \) will have on average 0.1 descendants of type \( y_j \) at time \( t_2 \). Similarly, the maps can be used to estimate ancestors distribution by moving backwards on time. Thus, each column gives the ancestor distribution of some cell \( y_j \).

**Fig. 1.9:** The schematic landscape of a developmental process with sampling time information (left) highlights the application of transport maps for estimating the ancestors and descendants of a cell \( y_j \) (right). Adapted from https://broadinstitute.github.io/wot/tutorial/.

To compute the transport map \( \pi_{t_1,t_2} \) connecting cells \( x_1, \ldots, x_n \) at time \( t_1 \) to cells \( y_1, \ldots, y_m \) (which absorb mass \( 1/m \)) at time \( t_2 \), the method solves an optimization problem over all matrices \( \pi \) that obey certain row-sum and column-sum constraints. These constraints ensure that the total amount of mass flowing out of each cell \( x_i \) and into each cell \( y_j \) adds up the correct amount. The transport matrix with the lowest possible transport cost subject to these constraint is selected. This cost is measured by mass \( \times \) squared distance travelled. In brief, the aim is to solve the UOT problem as introduced in [145], where only the row-sum constraints is enforced (approximately) and adding entropy to the transport matrix:
\[ \hat{\pi}_{t_i,t_{i+1}} = \arg\min_{\pi} \sum_{x \in S_1} \sum_{y \in S_2} c(x,y)\pi(x,y)dxdy - \varepsilon \int \pi(x,y) \log \pi(x,y) dxdy \]
\[ + \lambda_1 \text{KL} \left( \sum_{x \in S_1} \pi(x,y) \bigg|\bigg| d\hat{\mathbb{P}} t_2(y) \right) \]
\[ + \lambda_2 \text{KL} \left( \sum_{y \in S_2} \pi(x,y) \bigg|\bigg| d\hat{\mathbb{Q}} t_1(x) \right), \] (1.10)

where
\[ \hat{\mathbb{P}} t_i = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \] (1.11)

is the empirical distribution of samples \( x_1, \ldots, x_n \) at time \( t_1 \), and \( \mathbb{Q}_x \) denotes the distribution obtained by rescaling \( \mathbb{P}_t \) by the relative growth function \( g(x) \) which encodes the growth rate of cell \( x \), and is used to specify the budget of descendant mass for each cell \( x_i \) at time \( t_1 \). That is
\[ \hat{\mathbb{Q}} t_1(x) = \hat{\mathbb{P}} t_1(x) \frac{\hat{g}(x)^{t_2-t_1}}{\int \hat{g}(x)^{t_2-t_1} \hat{\mathbb{P}} t_1(x) dx}. \] (1.12)

\( \text{KL}(\mathbb{P}||\mathbb{Q}) \) denotes the KL-divergence between distributions \( \mathbb{P} \) and \( \mathbb{Q} \). The function \( c(x,y) \) encodes the cost of transporting a unit mass from \( x \) to \( y \). The cost function \( c(x,y) \) is defined as the squared euclidean distance between cells in local PCA space. This PCA space is computed separately for each pair of consecutive time points. In fact, it is possible to substitute the gene expression matrix with a PCA (if local PCA computation is disabled).

This is a convex optimization problem in the matrix variable \( \pi \in \mathbb{R}^{nxm} \), where \( n = |S_1| \) is the number of cells profiled at time \( t_1 \). Notice that by default the densities (on the discrete set \( S_1 \)) of the empirical distributions specified in equation (1.9) are simply \( d\hat{\mathbb{P}} t_1(x) = \frac{1}{n} \). Nonetheless, one could propose non-uniform empirical distributions (e.g. to include information about cell quality). The optimization problem has three regularization parameters: First, \( \varepsilon \) controls the degree of entropy in the transport map. Second, \( \lambda_1 \) controls the constraint on the row sums of \( \pi_{t_1,t_2} \), which depend on the growth rate function \( g(x) \). A smaller value of \( \lambda_1 \) enforces the constraints less strictly, which is useful when we do not have precise information about \( g(x) \). Third, \( \lambda_2 \) controls the constraint on the column sums of \( \pi_{t_1,t_2} \). To define the growth rate function \( g(x) \), an initial estimate of cellular growth rates based on gene signatures of proliferation and apoptosis is used. Then, this estimate is refined using unbalanced optimal transport as follows. Because the row-sum constraints are not enforced strictly, the actual row-sums of the optimal transport map \( \pi \) can be different than the initial input growth rate function \( g \). Hence, these new row-sums could be interpreted as a meaningful estimate of
growth rates, and used to form a new estimate of the growth function \( g^{(1)} \) which can be plugged back into the Optimal Transport (OT) optimization problem to compute a new transport map \( \pi^{(1)} \). It is therefore possible iterating back and forth between for learning growth rates \( g^{(it)} \) and transport maps \( \pi^{(it)} \) until the procedure eventually converges. In practice, just a few iterations are usually sufficient.

After having computed transport maps and used them to estimate the temporal couplings between adjacent time points, the next step is to infer transitions over a longer time interval \((t_k, t_l)\). Under the assumption that the developmental stochastic process is Markovian, long-range transitions can be estimated by composing transport maps as follows:

\[
\gamma_{t_k,t_l} = \gamma_{t_k,t_{k+1}} \circ \gamma_{t_{k+1},t_{k+2}} \circ \cdots \circ \gamma_{t_{l-1},t_l} \approx \pi_{t_k,t_{k+1}} \circ \pi_{t_{k+1},t_{k+2}} \circ \cdots \circ \pi_{t_{l-1},t_l} .
\]

(1.13)

Here \( \circ \) denotes matrix multiplication. The resulting temporal coupling \( \gamma(t_k, t_l) \) has a row for each cell at time \( t_k \) and a column for each cell at time \( t_l \). Likewise short-range coupling, the units are ‘transported mass’. So a value of \( \gamma_{t_k,t_l}(x,y) = 1.2 \) means that cell \( x \) will have on average 1.2 descendants with expression profile similar to \( y \) at time \( t_l \). The sum of the row shows the total number of descendants that a cell will have at time \( t_l \).

**Ancestors, Descendants and Trajectories**

There are several options for analysis and interpretation of transport maps. Having the aim of reconstructing developmental trajectories, which relies on the techniques for computing the ancestors and descendants distributions for a cell populations (i.e. a set of cells). Given a set \( C \) of cells at time \( t_k \), the descendant distribution at time \( t_{k+1} \) can be computed by pushing the cell set through the transport matrix. This push forward operation is implemented in a matrix multiplication. The first step is defining a the probability vector \( p_{t_k} \) to represent the cell set \( C \) as follows:

\[
p_{t_k}(x) = \begin{cases} 
\frac{1}{|C|} & x \in C \\
0 & \text{otherwise}
\end{cases} .
\]

(1.14)

Being a column vector, the push forward operation is performed by multiplying by the transport map on the right.

\[
p_{t_{k+1}}^T = p_{t_k}^T \pi_{t_k,t_{k+1}} .
\]

(1.15)

Subsequently, \( p_{t_{k+1}} \) is pushed forward again to compute the descendant distribution at the next time step. Continuing in this way, until any later time point \( t_\ell > t_k \). Note that this...
is equivalent to first forming the long-range coupling $\gamma_{t_k, t_\ell}$ and then pushing $p_{t_k}$ forward according to this coupling

$$p_{t_\ell}^T = p_{t_k}^T \gamma_{t_k, t_\ell}. \quad (1.16)$$

However, it is faster to compute the descendant distribution with successive push-forward operations, each of which involves a matrix-vector multiplication. In contrast, forming the long-range coupling involves matrix-matrix multiplies, which are significantly more expensive computationally. To compute the ancestors of $C$ at an earlier time point $t_k < t_\ell$, the cell set can be pulled back through the transport map. This pull-back operation is also implemented as a matrix-vector multiplication

$$p_{t_{\ell-1}} = \pi_{t_{\ell-1}, t_\ell} p_{t_\ell}. \quad (1.17)$$

The trajectory of a cell set $C$ refers to the sequence of ancestor distributions at earlier time points and descendant distributions at later time points.

**Fate matrices**

Another user application of transport maps are named fate matrices. Consider a pair of time points $t_k < t_\ell$ and a collection of cell sets $C_1, \ldots, C_N$ spanning all cells at the later time point, $t_\ell$ (That is, each cell $y$ from time $t_\ell$ is in some $C_L$). The probability that a cell $x$ from time $t_k$ will transition to a cell set $C_L$ at time $t_\ell$ is defined as the fate probability for cell $x$. Thus, a fate matrix $F_{t_k, t_\ell}$ is a matrix with a row containing the fate probabilities for each cell $x$ from time $t_k$. To compute the $F_{t_k, t_\ell}$, the columns of the long-range coupling $\gamma_{t_k, t_\ell}$ are aggregated according to the cell sets $C_1, \ldots, C_N$, which yields an un-normalized fate matrix $\tilde{F}_{t_k, t_\ell}$:

$$\tilde{F}_{t_k, t_\ell}(x, L) = \sum_{y \in C_L} \gamma_{t_k, t_\ell}(x, y). \quad (1.18)$$

which is then normalized in such way that each row adds up to 1:

$$F_{t_k, t_\ell}(x, L) = \frac{\tilde{F}_{t_k, t_\ell}(x, L)}{\sum_{L=1}^N \tilde{F}_{t_k, t_\ell}(x, L)}. \quad (1.19)$$
1.8 | **Aims**

1.8.1 **Extending the mouse gastrulation atlas of gastrulation and early organogenesis to E9.5**

The advent of single cell transcriptomics has been followed by a massive amount of data being constantly generated by a large number of scientific groups with a diverse range of interests. The vast majority of scRNA-seq data is available for public access in order to encourage scientific progress across several fields, and for the purposes of scientific reproducibility. Data from multiple independent experiments can be then integrated to complement new ones, allowing to explore different biological questions from those addressed in the corresponding publications, benchmarking of experimental and computational techniques as well as for improvement and development of data analysis strategies. However, the variety of scRNA-seq approaches and the speed at which these technologies are updated implies constant adaptation from the computational perspective. Furthermore, *in silico* integration of multiple experiments across technologies, species and diverse experimental contexts represents a major challenge for computational biologists working in the field. Lastly, providing informative and efficient access to such datasets is also a crucial aspect to be considered.

The so-called single cell transcriptomic atlases have proven to be a valuable resource for the scientific community where the challenges described above acquire critical relevance. In the field of developmental biology, temporal profiling within the atlases becomes increasingly important as such biological processes are inherently dynamic. Time course experiments are of special use when reconstructing cell differentiation trajectories is the aim. Chapter two of this dissertation describes the integration of a publicly available mouse development atlas (Pijuan-Sala et al., 2019 [82]) with newly generated 10X scRNA-seq data covering time points E8.5, E8.75, E9.0, E9.25 and E9.5. This, in order to extend the time course atlas of mouse gastrulation and early organogenesis (E6.5-E9.5). In summary, the strategy for generating an integrated landscape involved following the low level analysis used for the original publication. Every step includes a careful application of data processing techniques and strict quality control. The general tasks performed in this analysis are listed below:

1. Performing Low level analysis of newly generated 10X scRNA-seq data covering time points E8.5, E8.75, E9.0, E9.25 and E9.5, following the pipeline used or the publicly available mouse development atlas (Pijuan-Sala et al., 2019 [82]).

   (a) Alignment of raw sequences against the mouse genome reference.

   (b) Cell calling.
1.8 Aims

(c) Quality Control.
(d) Normalisation.
(e) Selecting Highly Variable genes selection.
(f) Batch effect correction.
(g) Doublet and stripped nucleus identification.

2. Integrating newly generated transcriptomic profiles with the existing mouse development atlas to create an extended version covering 13 time points (E6.5-E9.5).

(a) Performing batch correction using a mutual nearest neighbour approach (MNN).
(b) Transferring cell types annotations within overlapping time points.
(c) Computing transport maps using Waddington-Optimal Transport (W-OT).
(d) Predicting the descendant distributions from cell populations at E8.5.
(e) Integrating annotations from several publicly available datasets.
(f) Generating a clustered landscape based on the main lineages: Mesoderm, Ectoderm and Endoderm.
(g) Generating new cell type annotations based on multiple sources of evidence.

1.8.2 Reconstructing the haemato-endothelial trajectory landscape

There is a large number of computational tools for reconstructing developmental trajectories from single cell transcriptomics, a crucial task to leverage this technology in the field of Developmental Biology. The majority of those rely on pseudotime ordering of cells, while only a few methods incorporate sampling time into their frameworks. State-of-the-art examples of these methods include Waddington Optimal-Transport, a creative application of optimal transport theory to model developmental processes with stochastic modelling. An alternative approach that explicitly models time is RNA velocity. This method is based on the mRNA kinetics at the level of transcription, splicing and degradation; leading to inferences of future cell states via solving a dynamical system on gene expression space. Therefore, allowing reconstruction of differentiation trajectories. Importantly, these two methods do not rely on discrete topological structures and naturally incorporate the concept of cell fate probabilities.

In this context, the experimental design of the mouse atlas of gastrulation and early organogenesis fits perfectly, while at the same time presents a significant challenge due to the landscape’s size and complexity. The possibilities for exploration with this atlas are vast,
but one can select a specific cell fate to be deeply explored based on biological questions of particular interests as well as computational aspects. The haematopoietic system formation during development has been widely studied but the underlying molecular mechanisms defining its emergence are relatively poorly understood. From a general perspective, during embryonic development blood cells are produced in at least two consecutive waves; the first arising around E7.5 and the second from E8.25. The first wave (primitive) generates nucleated erythrocytes, which disappear shortly after birth. The second wave (yolk-sac-definitive) starts with the emergence of erythro-myeloid progenitors (EMPs) from yolk-sac haemogenic endothelium. However, there is evidence that suggests other tissues are also capable of generating blood cells. A computational reconstruction of In vivo haematopoietic development is an important step for dissecting the origins and diversification of haematopoietic cells. The aim of chapter three is to achieve the reconstruction of the entire haematopoietic landscape covering E6.5-E9.5 using the above mentioned State-of-the-art computational methods. Furthermore, to integrate perturbations experiments with trajectory reconstruction in order to explore the role of key transcription factors in blood development such as *Tal1*. The general tasks performed in this analysis are listed below:

1. Performing *W-OT* trajectory analysis of haematopoietic fates across the time points E6.5-E85.
2. Integrating trajectory analysis with a *Tal1* Chimaera model.
3. Exploring the mRNA kinetics of erythropoiesis using RNA velocity.
4. Extending the haematopoietic trajectory landscape towards E9.5.

### 1.8.3 Analysis of Smart-seq2 transcriptomic profiles from Primitive Streak dissections at E7.5

The primitive streak is a transient structure whose formation starts opposite to the anterior visceral endoderm. It is a region of the epiblast along which precursor cells of the mesoderm and the definitive endoderm ingress during gastrulation when they undergo an epithelial to mesenchymal transition. Cells are progressively recruited into the primitive streak as it elongates distally, giving rise to regionally distinct precursors, both along the mediolateral and anterior–posterior axes. A number of experimental efforts have highlighted the association cell fate bias across the primitive streak. However, the majority of them have been focused on the study of specific progenitors (e.g., cardiomyocytes, paraxial mesoderm and hemogenic endothelium). In this context, single cell transcriptomics provides the means for a wider exploration of the molecular mechanisms defining such distinctive set patterns.
1.8 Aims

To study the relation between cell fate decisions across different primitive streak regions a set of dissected portions (from proximal to distal) were profiled using single cell transcriptomics, having the goal of identifying expression profiles associated to them. Importantly, independent lineage tracing assays using grafting experiments were performed under the same experimental context. Furthermore, the application of Waddington OT developed in chapter 1 and 2, provides a full reconstruction of trajectories for all the cell types annotated in the original atlas [82]. Thus, opening the possibility for comparing computational predictions of cell fates with experimental observations from grafting experiments. The third chapter covers a standard scRNA-seq analysis aiming to explore cell heterogeneity in primitive streak cells. In addition, a comparison between computational predictions and experimental observations is carried out by taking advantage of previous atlas reconstruction performed in this dissertation, in order to assess the ability of W-OT to infer future cell states in such a complex scenario. The general tasks performed in this analysis are listed below:

1. Performing standard scRNAseq analysis (i.e., low level analysis, clustering and differential gene expression).

2. Mapping Primitive Streak dissected cells against the mouse development atlas.

3. Predicting potential fates of primitive streak cells by integrating W-OT trajectories.

4. Comparing predicted fates with experimental observations.
1.9 Other contributions

Contributions to papers that are not discussed in this Thesis are as follows:


Extending the single cell atlas coverage of early organogenesis to E9.5

The efforts to build cell atlases using single cell transcriptomics are at some extent analogous to the generation of reference genomes two decades ago. The challenges are vast, but thanks to the collective and dynamic effort of scientists in the field, the boundaries of technology as well as computational and experimental methods are under constant improvement. In the context of developmental biology, the spatio-temporal component becomes particularly crucial as developmental processes are inherently dynamic. The mouse atlas of gastrulation and early organogenesis resulting from a collaborative effort put in motion by John Marioni’s and Berthold Göttgen’s lab was conceived to explore the dynamics of lineage diversification across nine consecutive sampling time-points E6.5-E8.5, leading to the identification of 37 major cell populations. In this chapter, an extended version of this atlas was generated by integrating four new time points and approximately 300,000 transcriptomic profiles to the existing atlas. This dissertation was focused on the computational analysis associated to this project, emphasising on a careful and detailed pipeline of pre-processing (i.e., low level analysis), and the challenges to overcome in order to achieve high quality and reliable cell type annotation through a combination of different strategies, including the reconstruction of developmental trajectories.
Declaration This work is the result of joint effort put in motion in Berthold Gottgens’s and John Marioni’s laboratories. Both head researchers, designed the study and supervised the project. Tina Hamilton set up the timed matings and collected the uteri for embryo dissection. Carolina Guibentif performed then embryo dissection and Fernando Calero-Nieto carried out the 10X sequencing library preparations. I performed the entire computational analysis: all low level analyses on newly generated data as well as the processing of all publicly available datasets required to complement this study. Having generated an extended atlas of mouse development, I also established the reconstruction of all developmental trajectories and sorted out the computational tasks involved in cell type annotations (i.e. clustering, differential expression analysis and transferring annotations between atlases). The annotation of cell types for the extended atlas was heavily supported by literature review and manual inspections performed by Carolina Guibentif. She provided general guidance throughout the entire analysis. The manuscript is in preparation to be submitted for publication.
2.1 Introduction

Single cell transcriptomics has played an important role in the study of mouse development. The process of gastrulation and early organogenesis has been recently profiled across time, space and with multi-omic approaches (Section 1.5). These so-called atlases of mouse development are a valuable resource when unveiling the molecular mechanisms driving the process of embryogenesis, which impacts even further as researchers in the field use them as references to complement a wide variety of experiments (e.g., KO and disease models). The process of embryo development is inherently dynamic, but reconstructing the differentiation trajectories remains a major challenge (Section 1.6.6). Thus, from a computational perspective sampling time is an essential factor to consider in these experiments. In fact, the resolution of time intervals, representativeness of sampling and number of collected time points all define our ability to reconstruct the underlying dynamics in a reliable form. Furthermore, the computational integration of multiple independent datasets is not trivial. In contrast with other confounding factors (e.g. cell cycle) the presence of batch effects requires applying sophisticated methods to avoid misleading biological conclusions. The size of scRNA-seq atlases usually also implies new challenges at every single step of the analysis. Although sometimes overlooked, there are important limitations imposed by computational resources and computing time. The following sections will describe the entire pipeline of analysis performed to integrate a newly generated time course experiment to expand the existing atlas of gastrulation and early organogenesis [82]. The reasoning and methodologies behind each step as well as their results are described in detail, starting from low level analysis of newly generated transcriptomic profiles to a fully integrated developmental landscape with cell type annotations covering the stages E6.5-E9.5. The goal is then to use this atlas for computational reconstruction of fate specific landscapes of diversification in order to reveal biological insights about these processes and explore the underlying molecular mechanisms driving them.
### 2.2 Experimental design

The single cell atlas of mouse gastrulation and early organogenesis provided the transcriptomic landscape to study the origin and differentiation of early progenitors for a large number of tissues and organs across the embryonic stages E6.5-E8.5 [82]. The extension of this atlas to E9.5 is described in terms of the computational work performed, starting from low level analysis to a fully annotated integrated landscape. It is worth mentioning, however, that the mouse atlas previously generated (which will be referred as the original atlas) contains 116,312 cells distributed across 9 time points, while the new dataset (also referred as the extension atlas) has 314,027 cells distributed across one overlapping time point and four new ones (E8.5-E9.5). Thus, the resulting integration reaches up a total of 430,339 cells and 13 time points (E6.5-E9.5). Furthermore, these new embryos were dissected according to anatomical locations such as Yolk’Sac and Embryo proper, and sequenced independently to provide a notion of the embryonic region and facilitate identification of specific cell populations. The experimental design is summarized in Figure 2.1.

![Fig. 2.1: Schematic description of the experimental design used to generate an extended landscape of mouse development across embryonic stages E6.5-E9.5. A total of 314,027 single-cell transcriptomic profiles covering new time points were integrated to the existing atlas in an effort to expand the scope of cellular diversification in mouse development.](image-url)
2.3 A publicly available mouse gastrulation atlas

This time course scRNA-seq experiment contains 116,312 cells distributed across 9 time points sampled in six-hour intervals between E6.5-E8.5 [82]. Only two samples of this dataset (where a sample is a single lane of a 10x Chromium chip) had pooled embryos staged across several time points. Cells from these samples were referred as "Mixed" (see Figure 2.2.A). The corresponding metadata such as cell type annotations are also available (Figure 2.2.B). Despite processed data being at hand, this dataset was reprocessed starting from raw counts following the same strategy described in [82] to push forward the low-level analysis pipeline to ensure its reproducibility.

Fig. 2.2: Landscape of mouse atlas of gastrulation and early organogenesis represented with a UMAP. A. Cells are coloured by time point. B. Cells are coloured by cell type.

2.4 Low level analysis of new time points

This section describes the reasoning and results, step-by-step, of the low level analysis pipeline applied to newly generated 10X Genomics scRNA-seq profiles. Beside different versions of CellRanger software, the entire pipeline was performed in accordance to the analysis described in the original atlas [82] to ensure reproducibility with previous results, compatibility between datasets and facilitate their integration.

2.4.1 Sequencing reads processing

The first goal of the scRNA-seq data analysis is to obtain a matrix of features vs samples (genes vs cell barcodes) with the corresponding number of UMIs measured in the experiment. Here, alignment of sequencing reads against the mm10 genome, barcode assignment and
UMI collapsing were performed with 10X Genomics’ CellRanger software (v 3.1.0) and the GRCm38.92 genome annotation (Ensembl 92). The mouse reference sequence and gene annotation used followed the original mouse atlas having in mind further integration.

2.4.2 Detection and removal of barcode swapping

Molecule counts that were derived from barcode swapping were removed from all 10x samples using the function swappedDrops from 'DropletUtils'. This method identifies molecules that share UMI, cell barcode, and aligned gene between samples that were multiplexed for sequencing, which are fairly unlikely to have occurred [146]. In concordance with the expectations of sequencing on HiSeq 2500, few swapped reads were identified and excluded (Figure 2.3).

![Swapping molecules percentage]

**Fig. 2.3:** Histogram of percentage of swapped barcodes across all samples. Only one sample was above 1% within a total of 68 samples.

2.4.3 Cell calling

To perform cell calling, the standard CellRanger method retains barcodes with total UMI count of $\geq 10\%$ of the 99th percentile of the expected number of recovered cells, a number
that is heuristically defined by the user. Therefore, \textit{emptyDrops} function was used \cite{97} instead of the standard approach. This method considers a vector composed of the gene counts summed across barcodes associated with only very few UMIs (e.g., <100 UMIs), which is assumed to represent the pool of cell-free RNA. Any barcodes with a transcriptional profile that significantly deviates from this distribution may then be retained as a real cell, irrespective of its library size. In this case, barcodes of minimum 5,000 UMIs were considered, and the background RNA vector consisted of molecules from barcodes with fewer than 100 UMIs. A total number of 380,346 cell libraries were recovered across all samples.

\subsection*{2.4.4 QC}

Low complexity cell libraries were excluded from the analysis (fewer than 1,000 expressed genes). In general, library complexity in this dataset was higher than in the original atlas due to the use of different 10x Genomics Chromium systems (v.1 vs v.3 Chemistry respectively). The comparison between both datasets is shown in Figure 2.4.A. Moreover, only one library presented aberrant behaviour in the new dataset. It is worth to mention that Yolk-Sac dissected samples were associated with lower cell complexity as they have less cell type diversity compared to the embryo proper (since in a way it is a single "organ", while the embryo proper has many more). Next, cells with high mitochondrial UMI content were removed, as this is a known indicator of cell stress. The mitochondrial UMI fraction is a measurement associated with apoptotic, stressed and low-quality cells in scRNAseq experiments. In order to set a suitable threshold for this dataset, a normal distribution centred on the median, with variance estimated using the median absolute deviation (MAD) was used to build a null model for computing a p-value for each cell library. A Benjamini–Hochberg (BH) adjusted p-value < 0.05 was considered for rejection. This threshold was 4.68\% (Figure 2.5). One sample exhibited significantly larger numbers of cells with high mitochondrial fractions but also retained a considerable number of UMIs. The resulting distribution of UMIs was fairly homogeneous across time points (Figure 2.6) as well as sequencing rounds (data not shown). After this step, 351,872 cell libraries remained.

\subsection*{2.4.5 Normalisation}

Systematic differences in the total number of counts that each cell library owns must be corrected in order to prevent misleading biological conclusions. The naive approach of library-size normalisation is sensitive to composition effects. Where strongly differentially expressed genes are present, scaling by total RNA content may result in spurious differences for the remaining genes in that sample (the sampling pool of under-represented genes is decreased).
Fig. 2.4: Library complexity of both, the original atlas (after QC) and the new dataset (i.e., extension atlas).
Fig. 2.5: Removal of cells with high fraction of mitochondrial UMIs (top), and number of UMIs per sample highlighting the fraction of cells removed due to high mitochondrial content (bottom). The black arrow refers to the sample with the largest fraction of cells removed.
To address this, several improved methods have been developed for bulk RNAseq where size factors are computed using more sophisticated methods, such as weighted trimmed mean of M-values (TMM) [147] or Relative Log Expression (RLE) from DESeq [148]. Nonetheless, these methodologies were not designed to account for the high prevalence of zero-counts in single-cell data. In particular, zero-counts greatly impact DESeq due to its reliance on genes with non-zero counts in every cell, while TMM frequently overcorrects for scaling factor sizes [149]. Hence, library size factors of this dataset were estimated using scran [120, 150], which normalises many pools of combined single cell transcriptomes, before deconvolving individual size factors based on a system of linear equations established during the pooling. Clustering of cells before pooling aims to separate groups with very different library sizes or transcriptional profiles, which further assists in handling possible composition effects. Size factors were computed using default settings of `computeSumFactors`, having clustered the cells with the `quickCluster` function in scran (using the igraph clustering algorithm). Size factors are shown in Figure 2.7.

### 2.4.6 Highly variable gene selection

The process of highly variable genes (HVGs) selection aims to identify genes driving biological heterogeneity in scRNA-seq experiments. To decompose the variability coming from technical and biological sources, a mean-variance trend was fitted for further decomposition.
2.4 Low level analysis of new time points

based on the assumption that the majority of genes are not variably expressed. Therefore, the technical component dominates the total variance for most genes, such that the fitted trend can be treated as an estimate of the technical component. The scran function `trendVar`, was used to perform the fitting using a loess regression with a reduced span argument of 0.05. Then, genes were retained according to a significant deviation above the fitted trend (BH-corrected p<0.05, via scran function `decomposeVar`, $\chi^2$ test). These are highlighted in Figure 2.8.A. Genes with a mean-log expression value lower than 0.001 were previously excluded, to ease the strictness of the multiple testing correction. In fact, HVGs were selected skipping the false discovery rate (FDR) threshold at the extent of requiring larger computational resources for downstream calculations. This decision was taken following the pipeline from the original atlas, where it was thought for leveraging the extraction of sample-batch effect genes whose variation would hold in further applications of dimensionality reduction with PCA, in such way that the batch effect signal becomes easier to estimate when correcting with `fastMNN`. Figure 2.8.B shows the application of `ModelVar` function from scran which provides an updated more robust procedure for extracting genes driving the biological variability. When this method is used in downstream analysis, it is explicitly indicated. Otherwise the approach used in [82] was followed.

2.4.7 Batch effect correction

As mentioned above (Section 2.4.6), genes selected as highly variable also include those that are affected by batch effects, and such are preserved in any dimensionality reduction.

**Fig. 2.7:** Normalisation size factors with respect to library size (Total UMIs per cell).
To correct for batch-effects a mutual nearest neighbours (MNN) algorithm was applied [105] using `scran` function `fastMNN`. This specific implementation includes the prefix "fast" because it is designed for batch-correction over reduced spaces such as PCA, which is convenient when working with such a large dataset. Nevertheless, the MNN algorithm was first conceived to work in gene expression space. The approach works under three main assumptions: there is at least one cell population that is present in both batches, the batch effect is orthogonal to the biological subspace, and the batch-effect variation is much smaller than the biological-effect variation between different cell types.

The biological subspace refers to a set of basis vectors that represent a variety of biological processes, such that the real profile of each cell can be expressed as the linear sum of these vectors. Meanwhile, the batch effect is represented by a vector of length equal to the total number of features, which is added to the profile for each cell in the same batch. To obtain this vector the method identifies pairs of cells that are mutually nearest neighbours between batches (i.e., samples). For each MNN pair, a batch-correction vector is computed as the vector difference between the profiles of the paired cells. A smoothed cell-specific batch-correction vector is then calculated as a weighted average of these pair-specific vectors, as computed with a Gaussian kernel. Thus, enabling a locally linear batch correction. Cells that do not have nearby MNN pairings will effectively not be corrected, and differences in cell type frequency between samples should not affect correction. A batch effect that affects different cells in different ways will also be handled by this approach. The MNN method is
primarily based on batch pairs. Intuitively, to perform corrections over several batches the algorithm is applied sequentially. In the context of a developmental process, the ordering of this sequence heavily impacts on the algorithm capacity to drive a meaningful correction. Hence, the order of batch correction was first performed within each time point, moving from the samples with most cells to the samples with least (allowing the maximum number of discrete MNN pairs to be identified in the first corrections, such that a reasonably sized manifold can be robustly built up). Subsequently, the correction was applied from the latest time points to the earliest ones (with the mixed time point between E7.0 and E7.25). Similar results when correcting in the opposite direction of time were observed. In contrast, a random ordering resulted in significant lost of continuity within the manifold. The application of this method was ubiquitous in all downstream analysis.

2.4.8 Doublet & stripped nucleus identification

Due to the scale of this dataset and the large numbers of cells captured in some samples, significant amount of doublets are expected to be present in the cell libraries. The function doubletCells from scran was applied to score cells for their doublet state in a sample-wise manner, because only cells from the same sample can form a doublet. This method simulates doublets by merging transcriptomes, and computes the density around each cell library of simulated doublet transcriptomes, normalised by the density of observed cell libraries. Thus, high scores are associated with high doublet probability. However, stochastic fluctuations in the pairs of cells that were sampled for doublet simulation could result in locally uneven doublet scoring. Therefore, doublet calling was performed in a sample-wise manner over groups of cells obtained using a Louvain clustering algorithm [151] (cluster_louvain; igraph package; default parameters) by building a shared nearest neighbour (SNN) with the top 50 principal components across the corresponding highly variable genes. The clustering procedure was repeated to break the data into smaller clusters to ensure that small regions of high doublet density were not clustered with large numbers of singlets. These clusters are denoted as "sub-clusters". An instance of this is shown in Figure 2.9. For each cluster, the median doublet score was used as summary statistic of the scores of its cells, as clusters with a high median score were likely to contain mostly doublets.

To perform doublet calling in each sample a null distribution for the scores was built using a median-centred MAD-variance normal distribution. The MAD estimate was calculated only on values above the median to avoid the effects of zero-truncation (as doublet scores cannot be less than zero). All cells in clusters with a median score at the extreme upper end of this distribution (BH-corrected p < 0.1) were labelled as doublets (Figure 2.10). Intuitively, cells in sub-clusters that were called as doublets presented larger library sizes than singlet-
called cells (Figure 2.11). Furthermore, co-expression of \textit{Xist} and Y-chromosome genes was assessed, which should be mutually exclusive based on sex. Lastly, a final refinement of doublet calling was applied aiming to account for the existence of clusters that contained both many singlet cells as well as doublets, such as it could be the case with red blood cells (this, due to the dominance of relatively large level of transcription occurred in haemoglobin genes). Therefore, a final clustering step across all samples mixed together was implemented, but this time without sub-clustering and with batch-corrected PCs (as performed in Section 2.4.7). To identify clusters that contained more doublets than expected, the fraction of cell libraries that were called as doublets in each sample was considered. Then, another null model with same assumptions described above was constructed but using this fractional measure instead of the median doublet score, which may not be comparable between cells from different samples. Note that doublets identified by this sharing of information across samples show similar patterns as the doublets called in individual manner (Figure 2.12). Finally, clusters containing cells with considerably reduced library size and mitochondrial read fraction were also excluded as they potentially contained nuclei that had been stripped of their cytosol; thus losing their mitochondria entirely (Figure 2.13). These cells would mainly carry nuclear RNA. Following the above QC steps, the final number of single-cell libraries that could be used for downstream analyses is 314,027.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{sample_9}
\caption{Sample 9 is used to illustrate the process of identifying clusters made of doublets. The tSNE located at top left corner shows the gradient of doublet scores across all cells in that sample. Further clustering and sub-clustering reveals the association of high doublet scores with louvain clusters.}
\end{figure}

\section{2.5 Creating an integrated landscape}

Having finished the entire low level analysis, log transformed normalised counts from the original atlas and those newly generated were integrated into an extended time course ex-
Fig. 2.10: The median value of doublet scores (log transformed) per sample is shown, highlighting in red those clusters at the extreme of the upper distribution that were labelled as doublet clusters.
experiment across mouse developmental stages E6.5-E9.5 (13 time points). Unfortunately, we noticed that E9.5 contained fewer cells than other newly generated time points (Figure 2.14) because fewer embryos were available for profiling in that specific time point. Moreover, these embryos were visually younger than expected. Their somite count actually overlapped with E9.25 embryos. Therefore, E9.25-E9.5 were collapsed when performing all trajectory inferences during downstream analysis. The expression matrix of the integrated atlas contained 23,972 genes and 430,339 cells. Selection of highly variable genes and batch correction with fastMNN was performed as described above (Section2.4.7), resulting in 5,665 genes and 75 batch-corrected PCs. Upon the previous procedure, a batch balanced KNN graph (BBKNN) [108] was built from the top 50 batch corrected PCs (as implemented in scanpy). Then, a UMAP was generated with the umap-learn algorithm [125]. The resulting integrated landscape (Figure 2.15) further on will be referred to as the extended atlas. Additionally to embryo stage and somite counts (both of them reflecting time), anatomical dissections were recorded for the new dataset. These annotations will be used for biological interpretations.

---

**Fig. 2.11:** Distribution of UMIs of singlets and doublets. Larger values of total UMIs per library are associated with clusters identified as made of doublets.
2.5 Creating an integrated landscape

Fig. 2.12: tSNE displaying doublet identification using two complementary strategies. The relation between clusters identified using a sample-wise approach (blue) and after refinement with all samples together (red) is highlighted. Cells retained as singlets are shown in grey colour.
Extending the single cell atlas coverage of early organogenesis to E9.5

**Fig. 2.13:** Identification of cells with stripped nuclei (red) associated with considerably reduced library size and mitochondrial read fraction.

**Fig. 2.14:** Number of cells across time points covered by the extended atlas.
Fig. 2.15: Overview of the extended atlas using a UMAP representation. A. Cells are coloured by atlas version, the original atlas (red) and the extension atlas (blue). B. Cells are coloured by anatomical dissections of the embryo. Those coloured in grey belong to earlier time points in which cells were pooled. C. Cells are coloured by time points displaying the entire time progression of the extended atlas. D. Cells are coloured by embryo somite number, another measure of developmental time.
2.5.1 Mapping stage and cell type annotations within the atlas

After generating an integrated atlas and in preparation for trajectory analysis, metadata annotations of embryonic developmental stages and cell types were expanded. The former annotations are required for any method that explicitly uses time point information, which are covered by all cells in the atlas but those labelled as "Mixed" gastrulation (see Section 2.3). The latter refers to E8.5 cell type annotations to be transferred towards cells from the overlapping time point in the new dataset. To transfer annotations, a fastMNN based approach was used. In brief, log transformed normalised counts from both, the reference (original atlas) and the query (extension atlas) dataset are merged together. Then, highly variable genes and top principal components are computed to subsequently use fastMNN for re-scaling the PCA manifold from both datasets. The annotations from the reference are assigned to the query metadata, by selecting the mode among a given number of nearest neighbours (KNN) between the query and the reference PCA subspaces (i.e. the most frequent cell type, stage or any other annotation among a number of neighbouring cells connecting the query with the reference manifold). The number of nearest neighbours is chosen depending on the resolution of transferred annotations. To extract the KNN cells between both PCA subspaces, queryX function from Biocneighbours was used. Cells labelled as "Mixed" gastrulation time points from the original mouse atlas were allocated to embryonic developmental stages using the 30 nearest neighbours queried from the top 50 batch-corrected principal components belonging to the E6.5-E8.5 original dataset. This, in order to increase the statistical power of further time course based trajectory reconstruction. New E8.5 cell type annotations were assigned with respect to 10 nearest neighbours of E8.5 cells from the original atlas (Figure 2.16.A). In this context, the E8.5 time point was not only leveraged for improving batch correction, but also for further prediction of descendant cell populations using trajectory analysis (Next Section 2.5.2).

2.5.2 Estimating the descendants from cell populations at E8.5

The main challenge behind reconstructing differentiation trajectories from scRNA-seq arises from the destructive nature of the data, as we are only able to capture snapshots of the dynamical process which are compatible with multiple dynamics [134]. There are several computational approaches for reconstructing these trajectories, the majority of them are based on the use of pseudotime that is essentially a distance function from the progenitor cell to all downstream cells [136]. However, in this particular case superimposing time into the reconstruction of trajectories represents an important advantage to be used for their improvement. Furthermore, an ideal method should be capable of modelling fate...
branching as a continuum, to avoid reliance on clusters of transitional populations, as well as calculating cell coupling probabilities along the time course of differentiation. In this context, the Waddington Optimal-Transport (W-OT) approach was conceived with the purpose of modelling time course experiments of developmental processes as a generalisation of a stochastic process using optimal transport theory (Section 1.7.2). Hence, it allows estimating the coupling probabilities between cells of consecutive time points, while taking into account cell growth and death [139]. The latter by means of unbalance transport (allowing to violate the law of mass conservation) [152]. Table 2.1 shows a summary of these characteristics for a collection of popular methods designed for trajectory reconstruction (notice that W-OT is the only one that ticks all the marks). Thus, this approach was selected and the transport maps of consecutive time points were constructed over the entire time course experiment with Waddington-OT (wot 1.0.8.post1) using three iterations for learning the cell growth rate and skipping the dimension-reduction step, and instead using the batch-corrected principal components as input. Embryonic stages E9.25 and E9.5 were collapsed into a single time point for computing the transport maps because their somite count overlapped (see section 2.5) and the number of cells profiled at E9.5 was considerably lower which could lead to potential confounding factors disrupting trajectory inference, mainly in terms of differentiation stages and growth rates estimation. These transport maps were used to estimate the full trajectories of every cell population present at E8.5. That is, for each cell population (i.e. Erythroid3) the coupling probabilities were used to reconstruct the sequences of ancestors and descendants distributions by pushing the cell set through the transport matrix backwards and forwards respectively. Cells belonging to E8.75-E9.5 were allocated as descendants from a given population at E8.5, by selecting those with maximum mass (i.e. probability) across all their potential origins. The resulting landscape is shown in Figure 2.16.A. Then, an additional refinement was added by leveraging previous work developed in the lab using the original atlas, over Early-Somites diversification [153], where the populations annotated as Paraxial and Somitic mesoderm where subdivided into more specific annotations. The trajectories for these subpopulations were also computed using W-OT and following the same strategy, E9.5 cells from Paraxial and Somitic mesoderm and their descendants were split into Cranial mesoderm, Sclerotome, Dermomyotome, Posterior somitic tissues and Pre-somitic mesoderm (See Figure 2.16.B).

2.5.3 Computational validation of transport maps

The resulting predictions of descendant populations were fairly satisfactory. However, it was important to assess whether temporal resolution for W-OT was fine enough. Thus, a stringent computational strategy based on geodesic interpolation was applied. Following
Fig. 2.16: Graphical visualisation of E8.5 cell descendant predictions using W-OT. A. UMAPs of the extended atlas showing the initial steps of cell type annotations. First, cell types from the original atlas are highlighted on top of cells that belong to the new dataset (coloured in grey). Then the overlapping E8.5 is used as anchor between both datasets and the annotations from original at this time point are transferred to new E8.5 cells. The rest of the atlas extension remains coloured with grey. Mapping was performed using the MNN algorithm above mentioned (Section 2.4.7). Last, E8.5 cell descendants are predicted based on these annotations and allocated in future time points according to highest probabilities of each cell across all estimated trajectories. B. The resolution of cell type diversification of Somitic and Paraxial mesoderm is expanded through a second application of descendant predictions using finer annotations obtained from previous work [153].
Table 2.1: Comparison of computational methods for trajectory reconstruction based on the characteristics desired for the mouse atlas dataset. Notice that some of these methods can be complementary, and that W-OT is the only one that ticks all the marks. An extensive benchmarking of trajectory inference methods can be found in [137]. However, it did not include W-OT.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type of branching</th>
<th>Models time</th>
<th>Models growth</th>
<th>Coupling (Fate probabilities)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPT [123]</td>
<td>Bifurcation</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>PBA [134]</td>
<td>Population in equilibrium</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PAGA [154]</td>
<td>Graph abstraction/Tree</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>FateID [155]</td>
<td>Continuum</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Palantir [156]</td>
<td>Continuum</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Monocle2 [135]</td>
<td>Tree</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Monocle3 [157]</td>
<td>Graph abstraction/Tree</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>velocyto [138]</td>
<td>Continuum</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>scVelo [141]</td>
<td>Continuum/PAGA</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>scDiff [158]</td>
<td>Tree</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>STITCH [159]</td>
<td>Continuum</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>W-OT [139]</strong></td>
<td>Continuum</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The mathematical notation provided for W-OT (Section 1.7.2), this approach works over the idea that a developmental process \( P_t \) traces a curve in the space of distributions, and that the time course experiment is represented as a noisy sample of this process \( \hat{P}_t \), which also includes batch effects. In this context, optimal transport allows connecting these noisy samples with straight lines named geodesics or shortest paths. Thus, the goodness of this fit provides a metric for testing the temporal resolution (i.e the fit improves as the resolution increases). However, given the level of noise in \( \hat{P}_t \), the fit will never be perfect. To quantify the baseline noise level, the distance between two independent replicates at the same time point is computed. When this noise level is comparable to the quality of interpolation, it is reassured that OT is working well. This procedure was applied first only using the original atlas, and then with the integrated atlas (Figure 2.17). In this process, all triplets of consecutive time points were taken \((t_i, t_{i+1}, t_{i+2})\) for which the data from time \( t_{i+1} \) was extracted to estimate the coupling \( \gamma_{t_it_{i+2}} \), connecting time \( t_1 \) to \( t_{i+2} \) for computing an interpolating distribution at time \( t_{i+1} \). Then, the interpolated and observed distributions (\( \hat{P}_t \)) are compared using Wasserstein distance. Given that each time point consist of multiple batches (e.g. \( \hat{P}_{t+1}^{(1)}, \hat{P}_{t+1}^{(2)} \)), these are compared with each other to estimate the baseline noise level. Furthermore, two types of null models were generated to be compared with the held-out data (with and without considering cell growth). The overall quality of the time course experiment seemed reasonable according to this procedure. The trends for distances between interpolated and real, as well as between
Extending the single cell atlas coverage of early organogenesis to E9.5

Fig. 2.17: Validation of transport maps using geodesic interpolation. A. Quality of interpolation at each triplet of consecutive time-points assessed in the original atlas (E6.5-E8.5). B. Extending the validation to the integrated atlas, E6.5 to E9.25-E9.5. The labels indicate the type of comparisons performed: interpolated vs real in blue, batch to batch in red (baseline noise level) and two different null models (orange and yellow).

batches move closely together with distances in general below the null models. It is at E7.25 where an important difference is observed, most likely due to rapid change undergoing over the molecular programs at this time point. The latest time points (E8.75-E9.5) showed significantly increased levels of noise level variation between batches (i.e., variance). This is related to the fact that the embryos from different dissected regions were sequenced in separate samples. Therefore, while some batches are relatively homogeneous between each other, others are coming from very different cell populations (eg. sample X vs sample Y, both from Embryo proper compared to sample Z from Yolk-Sac). The authors of W-OT noticed that this validation procedure only provides a guarantee when the temporal resolution is twice as fine as necessary. When the transport maps are computed to analyse the data, no time point is skipped. It is possible that $\pi_{t_1,t_2}$ is a good approximation to $\gamma_{t_1,t_2}$ even when $\pi_{t_1,t_2}$ cannot be validated as a good approximation to $\gamma_{t_1,t_2}$. Thus, it is advised to proceed (with caution) even without proving that the baseline noise level was achieved.

2.6 The annotation beyond cell descendants

The identification of cell descendants from E8.5 provided useful information from a lineage inference perspective but does not reveal itself any new population arising in the following
2.6 The annotation beyond cell descendants

time points. Currently, there is no such thing as a robust fully automated unsupervised method for cell type identification (as explained in section 1.6.4). Furthermore, the challenge increases when referring to developmental processes as discrete cell categories have to be assigned along a continuum defined by transitional states between a given pair of populations (i.e. from pluripotent cells in the Epiblast to Heart progenitors). The work performed in the previous section was only the basis for a deeper characterization of new and existing cell populations present in the extended atlas, an exhaustive annotation refinement process had yet to be applied. The strategy for refining the annotation relied on a combination of multiple complementary approaches such as clustering based differential expression analysis and manual review of gene markers extracted from literature. When possible, also guided by other publicly available references of mouse embryonic profiles at these stages. It is important to keep in mind, that these annotations are under constant improvement and therefore, might not be conclusive by the end of this project. The increasing complexity of detailed populations that can be reached across each of every major lineage was with few exceptions (e.g. blood production), out of the scope from this analysis. Lastly, these annotations are constantly consulted with a variety of experts and hence, sometimes are challenged ultimately leading to a more careful exploration.

2.6.1 An overview of Gut tube cell heterogeneity at E8.75

To provide guidance on the annotation process, additional publicly available data was leveraged. For instance, a landscape of mouse endodermal emergence which covers the stages E3.5–E8.75 [88]. The E8.75 time point from this dataset was extracted and integrated with the extended atlas using fastMNN. This subset contained 57,401 transcriptomic profiles. Thus, the integration was performed by splitting the E8.75 dataset into multiple chunks randomly sampled without replacement (to avoid computer memory overloading). Next, neighbouring cells between the two atlases were identified using function queryKNN, and the closest cells were used to project relevant annotations into the extended atlas. For instance the UMAP shown in Figure 2.18. The cell types annotations from [88] were generally consistent with the current knowledge of the extended atlas, indicating a successful integration. For example, the Haemato-endothelium, Notochord, Primordial germ cells (PGCs), and the extra-embryonic tissues. Gut tube annotations evidenced highly complex regions composed by several subpopulation with well defined patterns.
Fig. 2.18: Extended atlas UMAP displaying mapping E8.75 cell types from [88] against the extended atlas. Cells are coloured by cell types in Nowotschin et al., [88], displaying remarkable consistency with the current annotations of the extended atlas. The turquoise arrow is indicating a few visually inconsistent cells (e.g. mesoderm cells mapping to non-mesodermal regions).
2.6.2 Exploring the mouse Brain development atlas

The availability of a Brain focused mouse development atlas was a natural choice for supporting cell type annotations as it contains overlapping time points with relevant information for this purpose [1]. Thus, this atlas was downloaded and processed. The first five time points were extracted from the whole dataset covering stages E7-E10. Time point E10 does not overlap with the extended atlas. However, it still could be useful in further steps as an approximation to the latest time points. This subset contained 70,243 cells, which were mapped against the extended atlas using the same strategy described in the previous section (see section 2.6.1). The mapping of the overlapping landscape (E7-E10) showed relative consistency with the extended atlas. Nevertheless, there were a couple of notable exceptions (see Figure 2.19). Examples of contrasting annotations are few cells annotated as Endoderm mapping into the extended atlas endothelial region, as well as neural crest cells spreading around the mesoderm region. The finest annotations of the Brain atlas are not yet public, and from the pre-print manuscript is still hard to make solid statements about how these annotations were assigned. The tissue and dissected regions recorded in the metadata lead to questions about how the embryo dissections were performed (see Figure 2.20). For instance, it is unclear what was considered as head along the annotations from tissues and dissected regions.

Though with caution, those already available annotations from the Brain were used as guidance for identification of brain specific cell types such as neurons, which were validated using gene markers in further steps. Similarly, that was the case for immune system cells projected into haemato-endothelial landscape. Interestingly, a few cells annotated as mesoderm in both Brain and Gut development atlases mapped to a very localised section far out of the mesodermal landscape. This region is highlighted in Figure 2.19 among other observations that lead to more careful analyses.

2.6.3 A clustering based approach

Clustering of single cell transcriptomic profiles is arguably a necessary step for unsupervised cell type identification. The relationship between clusters and cell types has been widely exploited to characterize cells from scRNAseq experiments, but it is important to keep in mind that there is no assurance supporting this assumption (even if accounting for the presence of confounding factors like batch effects or cell cycle). Supervised approaches such as cellassign (see section 1.6.4) rely on a priori definition of gene markers which can be problematic in situations when de novo annotations are required. The limitations are similar for methods based on alignment between datasets, as only known annotations
Fig. 2.19: Extended atlas UMAP displaying mapping cell types from La Manno et al., 2020, [1] against the extended atlas. Though useful for identification of cell types such as neurons and immune cells. There were clear inconsistencies with the extended atlas. These are indicated by turquoise arrows.
Fig. 2.20: Extended atlas UMAP displaying mapped tissues and region dissections from La Manno et al., 2020, [1] against the extended atlas. Despite clear inconsistencies in the annotations with respect to brain cells mapping in the mesoderm, most likely due to misleading annotations of dissection regions (partially corrected at tissue level). This mapping allows to see the patterning of Hindbrain, Midbrain and Forebrain regions. **A.** Cells are coloured according to dissected regions in La Manno et al., 2020, [1]. **B.** Cells are coloured by tissue annotations from La Manno et al., 2020, [1].
can be transferred (e.g., Seurat [106, 107]). Thus, the next step was focused on creating a clustered landscape to be used as a basis for incorporating de novo annotations as well as for refining the current ones. Initially, the Louvain clustering algorithm described in Section 2.4.8 was applied (Figure 2.21). However, this algorithm does not provide a parameter for controlling the resolution of clustering (i.e., number of clusters), and though useful already for a superficial inspection, it was evident that increasing resolution was needed to capture the complexity aimed for the annotation. To simplify the problem, the extended atlas was divided into two subsets: one for inspecting the diversification of Endoderm, Ectoderm and Neuro-mesodermal progenitors (NMPs), another one focused only on the rest of the mesodermal landscape (Figure 2.22). Highly variable genes and batch corrected PCs were computed again for each subset. The resulting top 50 PCs were clustered independently (Figure 2.23). However, instead of Louvain’s iterative approach, this time the so-called Leiden algorithm was used due to recent evidence of Leiden outperforming Louvain in terms of the ability of clustering network communities that are guaranteed to be connected [130]; a desired property particularly when allocating cell type annotations. The resolution parameter of Leiden’s algorithm was set up to 5 for allowing finer cell populations to be identified. The large number of cell fates and transitional states as well as the relations between them, made visualisation significantly challenging. Different versions of force directed graphs (Figure 2.24) as well as 3D UMAPs (see https://github.com/rosshandler/3D_UMAPs_EmbryoTimecourse) were used for manual inspection. Even a virtual reality tool for visualisation of scRNA-seq data was used (CellexalVR, [160]). Upon generating clusters and visual representations, differential gene expression analysis (using scran function findMarkers), literature review and manual inspection of gene markers were performed. Supported by existing cell type annotations and the inference of population descendants, these analyses expanded the annotations towards identification of emerging cell types (Figure 2.25). Further details are provided in Section 2.6.4.

2.6.4 | Towards a complete annotation of cell types

The process of annotating clusters was an exhaustive effort that involved a variety of analyses that are beyond the computational scope of this thesis. However, a detailed and robust annotation is essential for interpretation of inferred trajectories. Having annotated the above mentioned clustered landscapes, there were still a couple of steps to be addressed. PGCs and their descendants (also PGCs) were not included in clustering analysis because a relatively large proportion of these cells were visually spread around other surrounding populations. Therefore, these cells were re-annotated using cellassign provided with known markers (e.g., Nanos3) as well as differentially expressed genes between PGCs and those populations. Cells
Fig. 2.21: Extended atlas UMAP displaying louvain clustering results. Major cell types are associated with the clusters. However, relatively few clusters are identified with previously applied louvain algorithm (Section 2.4.8). The *a priori* known complexity of the landscape requires increasing the resolution of clustering. Thus, an alternative strategy had to be constructed (Figures 2.22 and 2.23).
Extending the single cell atlas coverage of early organogenesis to E9.5

**Fig. 2.22:** The Ectoderm-Endoderm and Mesoderm landscapes of lineage diversification (E6.5-E9.5) represented using force-directed layouts in order to provide a finer perspective of potential differentiation trajectories. **A.** Force-directed layouts with cells coloured by time point. **B.** Force-directed layouts with cells coloured by embryo dissections. **C.** Force-directed layouts with cells coloured by cell type annotations up to E8.5.
2.6 The annotation beyond cell descendants

Fig. 2.23: Leiden clustering with resolution parameter 5 applied to different diversification landscapes. **A.** Ectoderm-Endoderm-NMPs landscape displaying their corresponding Leiden clusters, where a total of 70 clusters were assigned. **B.** Mesoderm landscape displaying their corresponding Leiden clusters, where a total of 71 clusters were assigned.
Fig. 2.24: Force layout alternatives for visualizing lineage diversification. A. Ectoderm-Endoderm-NMPs landscape. In the right ride, the force layout was initialized using a graph abstraction generated with PAGA and louvain. B. Mesoderm landscape also showing random initialization of the force layout compared with a graph abstraction generated with PAGA and louvain.
2.6 The annotation beyond cell descendants

Fig. 2.25: Force layouts displaying clustering based annotation results. Cells are coloured by cell populations. A. Force layout of the ectoderm-endoderm landscape. B. Force layout of the mesoderm landscape.
that were not allocated to PGCs by cellassign were corrected according to their 30 nearest neighbours in the integrated landscape. The resulting and preliminary final annotation led to 87 major cell populations (Figure 2.26). Furthermore, there were a few cases where previous annotations were corrected or renamed. For instance, cells annotated as ExE mesoderm in the original atlas were allocated to Lateral plate mesoderm after a discussion about gene markers with other experts on the field. To date, there are two cell populations whose presence and annotations have not been explained as these form isolated clusters while at the same time express either erythrocyte or endothelial gene markers. These populations are annotated as Erythroid and Mesothelium-Endothelium with the extra label "Masked" (not fully considered for the atlas until they are resolved). These cells do not show high doublet scores. While the Mesothelium-Endothelium/Masked population is composed by only Yolk Sac cells, the Erythroid/Masked population includes red blood cells from both Yolk-Sac and Embryo proper. These red blood cells highly express Hbb-y, a marker of embryonic erythrocytes. Thus, it is unlikely that these cells come from maternal contamination.
Fig. 2.26: UMAP layout containing single-cell transcriptomes belonging from E6.5 to E9.5. The integration of cell type annotations obtained from separated landscapes led to identification of 87 major cell populations. Primordial germ cells (PGC); Epicardial derived cells (EPDC); First and second heart field (FHF); Second heart field (SHF); Megakaryocyte-erythroid-progenitors (MEP); Erythroid-myeloid-progenitors (EMP); Yolk-Sac (YS); Neuro-mesodermal progenitors (NMPs); Extraembryonic (ExE).
2.7 Discussion

The work performed in this chapter resulted in the generation of an extended version of the previously published mouse atlas of gastrulation and early organogenesis [82]. The extended atlas was constructed from 430,339 single cell transcriptomic profiles measured across developmental stages E6.5-E9.5 covered by 13 time points distributed in six hour intervals. The extension is significantly incremental as it adds four new time points and 314,027 cells to the existing atlas. The integration of newly generated time points with the original atlas was a major challenge in many aspects that range from dealing with the implications of different 10X chemistry versions, the need of large computational resources for common tasks, and the emergence of several new cell types in latest time points. In this context, the experimental design was critical for a successful integration. For instance, the inclusion of an overlapping time point (E8.5) facilitated the crucial step of batch correction and recording embryo dissections in the metadata provided strong support for cell type identification and trajectory reconstruction. Integrating and expanding cell type annotations was achieved using a combination of several computational approaches as well as expert knowledge, manual inspection and literature review. Detailed and reliable cell type annotations are not only crucial for all further analyses performed in this dissertation, but also necessary for leveraging the extended atlas as public resource to be used by the scientific community. The number of major cell populations annotated in the atlas went from 37 to 87, and the process of annotation will continue. It was the most exhaustive and time consuming part of this project, highlighting the necessity for a more standardised and automated approach for this task. In fact, there is no automatic method for cell type identification suitable for developmental processes. A possible semi-automatic approach inspired in the work here performed could be established as a combination of W-OT trajectory inference and clustering analysis. An initial clustering step would provide the basis for selecting ending points in the landscape, such clusters would contain a larger fraction of cells belonging to later time points (other approaches could be also used to identify ending time points [161]) and the corresponding mass distributions (trajectories) could be computed for them. Next, a sub-clustering step could be used to compare whether the ancestor distributions between sub-clusters significantly change. Differences between sub-cluster distributions could be indicative of diverging lineages. The process would iterate until no significant changes within sub-cluster distributions are observed. This trajectory guided clustering landscape could be assessed through differential gene expression analysis, by comparing each cluster with its neighbours. The resulting gene lists would inevitably require to manual inspection and certainly, known gene markers missed when selecting top differentially expressed genes
would be tested. However, this would provide an idea of the resolution required for describing the differentiation trajectories towards targeted cell fates (ending points). Meaningful and efficient ways for using and exploring the extended atlas have to be provided in order for this resource to be widely used. Software applications to build website interfaces such as R shiny are commonly used. The original atlas shiny website can be taken as example for the extended atlas. However, the number of cells increased to a degree that demands significant computational resources for simple tasks such as inspecting gene expression levels in the landscape. The standard servers available for hosting the corresponding website could easily collapse only with a few number of simultaneous users. More sophisticated tools such as cellxgene [162] have been designed to deal with specially large datasets (at least one million cells). However, it is restricted to the software features and design made available by developers. While the software is open source implementing modifications might not be trivial as common users are not familiar with the code and even the programming languages used for it. In this context, there is a tool currently under development called scarf for which the publication has not been made available yet (developed by Parashar Dhapola from Göran Karlsson’s lab). This tool has incorporated data chunk reading and incremental learning techniques to perform several standard scRNA-seq data analyses (e.g, normalisation, highly variable gene selection, graph representations, UMAP visualisations, among other) in a much more efficient way with respect to computational resources. For example, while the 3 million dataset from Cao et. al., 2019 [86] required more than 700 GB to be analysed with scanpy (which is among the top computational frameworks for dealing with large scRNA-seq datasets) [163]), scarf was able to reduce the memory load to a maximum of 5000 MB. The main advantage is that this tool can be used through a programmatic environment where R shiny code can be incorporated. Thus, the website for the extended atlas will be created using shiny while at the same time leveraging scarf to provide quick and efficient data access for common operations such as displaying the gene expression levels in a UMAP layout.
Resolving Haemato-Endothelial diversification

The cellular and molecular processes underlying the initial generation of hematopoietic stem cells during ontogeny have been intensively studied. Indeed, the current interest in stem cell-based therapies has emphasized the importance of understanding how tissue-specific stem cells are specified in development. The hematopoietic system has been a paradigm for this. In the mouse, hematopoiesis occurs asynchronously at different sites in the embryo along the so-called primitive and definitive waves of blood production. The original and extended atlas of mouse gastrulation and early organogenesis introduced in the previous chapter provide a unique opportunity for reconstructing the origins of the haemato-endothelial landscape and the emergence of a variety of cell populations. Here, state-of-the-art computational methodologies such as RNA velocity and Waddington Optimal-Transport are applied using innovative strategies to accomplish a successful reconstruction of differentiation trajectories. The majority of results obtained in this study are concordant with well established knowledge and, also support relatively novel hypothesis about alternative origins for blood and endothelium. The most controversial observation suggests that cells in the Epiblast already hold a certain degree of fate bias towards the primitive wave of erythropoiesis. However, further work is still required. In general, the results emphasise on the computational strategies and discusses some limitations as well as potential extensions to these methodologies.
Declaration Following from previous chapter, this work is also the result of joint effort put in motion in Berthold Gottgens’s and John Marioni’s laboratories. I alone performed the entire computational analysis with two exceptions. First, the multi-omics analyses with scNMT-seq was performed in collaboration with Ricard Argelaguet. Specifically, he processed the whole dataset and I transferred the annotations from the scRNA-seq mouse atlas to the scNMT-seq landscape. Second, the analysis of mRNA kinetics in blood cells development was performed in collaboration with Melania Barile. Specifically, I processed the mouse atlas to obtain the exonic and intronic counts matrices for the analysis. I also generated the RNA velocity landscape of the mouse atlas. Followed by Melania’s work on the limitations of scVelo in terms of genes with multiple kinetics. She developed the strategy to address this problem. Carolina Guibentif performed chimaera generation, dissection, and library preparation. Also, she provided me with guidance and support throughout the analysis. To date, two publications resulted from data presented in this chapter.

Publications:

3.1 Introduction

During embryogenesis, the hematopoietic system is established in successive waves that are temporally and spatially restricted, with each giving rise to specific blood progenitors [164]. Primary, red blood cells originate from the primitive streak and are formed in at least two consecutive waves in the Yolk-Sac; the first arising around E7.5 and the second from E8.25 [165]. The first wave (primitive) generates nucleated erythrocytes with unique characteristics that are only found during early embryogenesis and the role of which is to deliver oxygen rapidly to the quickly expanding embryo [166, 167]. These cells disappear shortly after birth [168]. Macrophage and megakaryocyte progenitors are also generated during this earliest wave of blood emergence [169]. The second wave (yolk-sac-definitive) starts with the emergence of erythro-myeloid progenitors (EMPs) from Yolk-Sac haemogenic endothelium [170]. These latter migrate to the fetal liver and generate definitive erythrocytes [165]. By E9.0 to E9.5, lymphoid progenitors are generated in the yolk sac and probably also in the intraembryonic para-aortic splanchnopleura region [171, 172]. Figure 3.1 shows a schematic representation of embryonic haematopoiesis. In this context, the extended mouse atlas described in the previous chapter covers these so-called waves of blood emergence (E6.5-E9.5) providing a unique opportunity for reconstructing the corresponding developmental trajectories to further study the dynamics of diversification resulting from the underlying molecular programs that give rise to these populations.
Fig. 3.1: Cellular origin of primitive and definitive haematopoiesis. Shown is a schematic representation of hemogenic angioblast or hemogenic endothelium emergence and the lineage contribution of primitive and definitive haematopoiesis. Inspired from Georges Lacauda and Valerie Kouskoff, 2016 [164]. Created with BioRender.com.
3.2 The origin of haemato-endothelial lineages

The analysis of haemato-endothelial diversification performed in the original atlas publication [82] created a basis for a deeper exploration of differentiation trajectories during stages E6.5-E8.5. As above mentioned, the formation of blood progenitors in the embryo has been described as a sequence of two waves clearly distinguished by time, the primitive and the Yolk Sac definitive [165]. It is well established that the Yolk Sac-Endothelium has haemogenic potential, and it has been recently suggested the possibility that other types of Endothelium also hold this potential [173–176]. Then, reconstructing this entire landscape not only provides new insights from the biological perspective (as the role and mechanisms of many genes in these processes are still relatively poorly understood), but also represents a significant challenge from the computational side. Thus, the Haematoendothelium became the main focus for trajectory reconstruction, having the advantage of previous work in the original atlas and counting on the lab expertise in haematopoiesis (Göttgens lab). The E6.5-E8.5 haemato-endothelial landscape includes finer cell type annotations relative to other populations allowing a more detailed exploration of these trajectories (Figure 3.2).

3.2.1 Waddington-OT trajectories in the interval E6.5-E8.5

These analyses were performed before the generation of the extended atlas, such that the transport maps described in Section 2.5.2 were first computed for the original atlas (E6.5-E8.5) using the same wot setting. The computational validation of these transport maps through geodesic interpolation was addressed in Section 2.5.3, where the difference between the transport maps of the original and the extended atlas was discussed. Initially, this analysis was restricted to the mesodermal landscape (Figure 3.3), followed by the generation of wot ancestors distributions (i.e., trajectories) for red blood cells (Erythroid3) and endothelial cells, as well as Cardiomyocytes (Figure 3.4). The latter was included to be contrasted with the haemato-endothelial trajectories in order to highlight similarities and dissimilarities with respect to other mesodermal cell fates. Though as overwhelming as it is displaying a separate histogram for each time point within each fate, it becomes clear that ancestors mass distributions significantly differ from each other across time as well as cell fates. Interestingly, the Kullback–Leibler divergence $D_{KL}$ between these three cell fates showed relatively large distances with respect to the Erythroid trajectory during an interval of time previous to blood emergence (E6.5-E7.25), suggesting important differences of fate bias at very early stages between red blood cells and the other two mesodermal fates.

As a consequence of such an heterogeneous collection of distributions, defining which cells belong to a trajectory is not a trivial problem. One alternative is to select cells above a given
Fig. 3.2: Force-directed graph layout of cells associated with the blood lineage across E6.5-E8.5 (reproduced from [82]), coloured by cell type (top) and cell type sub-cluster (bottom). The landscape contains 15,875 cells. Blood progenitor (BP); Endothelial cell (EC); Erythrocyte (Ery); Haematoendothelial progenitor (Haem); Mesodermal cell (Mes); Megakaryocyte (Mk); Myeloid cell (My).
3.2 The origin of haemato-endothelial lineages

Fig. 3.3: Third-dimensional UMAP of mesodermal diversification (E6.5-E8.5). Cells are coloured by time points (top) by cell type (bottom). This 3D UMAP was generated with the umap-learn algorithm [125], using the top 50 fastMNN batch-corrected PCs as input (the details of this procedure are explained in Section 2.4.7).
Fig. 3.4: The ancestors distributions of mesodermal fates diverge across time (E6.5-E8.5) and cell types. A. Ancestors mass distributions of red blood cells (Erythroid3) at each time point (histograms of $\log_{10}$ transformed mass units, $\log_{10}(mass + p)$ where p is a fixed pseudocount to avoid undefined operations). B. Ancestors mass distributions of endothelial cells. C. Ancestors mass distributions of cardiac progenitors. D. Kullback–Leibler divergence between mass distributions for every pair of cell fates across time (Endothelium vs Cardiomyocytes, Erythroid vs Endothelium, Erythroid vs Cardiomyocytes).
3.2 The origin of haemato-endothelial lineages

quantile at each time point (e.g., 95th) as likely belonging to a trajectory. For instance, the red blood cells trajectory shown in Figure 3.5. Notice that the transition from the Epiblast towards Blood progenitors seems to occur rapidly and thus, relatively few cells are actually allocated to multipotent mesodermal progenitors. This quantile approach works decently when the sample size is fairly balanced across time and the target cell population is not too small. However, one should be careful during interpretation and always keep in mind the actual magnitude of mass units relative to other cell fates. For example, the Erythroid ancestors distribution at E6.5 clearly shows heavier upper tails than the other two fates (Endothelium and Cardiomyocytes). This observation makes sense as the primitive wave is among the fastest differentiation trajectories in the embryo, because during development all other arising tissues need to be provided with oxygen.

To visually compare trajectories, one can use the log-likelihood ratios of two fates (e.g., Figure 3.6). In brief, the descendant distribution of any cell can be computed by extracting a row of the coupling matrix (for two adjacent time points this is the transport map, for longer time intervals transport maps are multiplied together). Hence, the probability of obtaining a fate is the fraction of descendants in that fate. Thus, the likelihood ratio is obtained by summing up the descendant mass in fate 1 and in fate 2, and dividing those.

Importantly, the trajectories of red blood cells were not affected by whether the transport maps included or not, the ectodermal and endodermal fates. Figure 3.7 shows a perspective of the Erythroid trajectory in the whole embryo 3D UMAP landscape (only Extraembryonic ectoderm and Extraembryonic endoderm cells were removed, as they derive from early lineage branching events that are not covered in this dataset). A movie of this rotating manifold can be found in https://github.com/rosshandler/3D_UMAPs_EmbryoTimecourse. To emphasize on the visualisation of one or more trajectories, one can extract the cells that likely belong to a number of trajectories and recompute highly variable genes, perform batch correction and generate a new two or three dimensional landscape. For the haematoendothelium (defined by the trajectories towards Erythroid3 and Endothelium), the resulting force directed layout is shown in Figure 3.8. Myeloid and megakaryocyte progenitors (My and Mk respectively) are placed at later times (E8.25-E8.5) than the vast majority of blood progenitors. These cells also seem connected with the endothelial population according the log-likelihood ratio between Erythroid3 and Endothelium trajectories, thus suggesting the initial stage of the yolk sac definitive wave. Interestingly, there seems to be multiple paths for making the different types of endothelial progenitors. The follow-up of these observations is addressed further in Section 3.5, where the haemato-endothelial landscape is expanded to E9.5.
Fig. 3.5: UMAP of mesodermal diversification showing the Erythroid sub-cluster 4 (Ery4) trajectory. Ery4 contains the most mature red blood cells in the landscape (see Figure 3.2). Cells above the 95th quantile across each distribution are highlighted as likely belonging to the trajectory. These cells are coloured by time point (left) as well as cell types (right) in correspondence with the legends of Figure 3.3.
3.2 The origin of haemato-endothelial lineages

Fig. 3.6: Log-likelihood ratios of mesodermal populations Erythroid3, Endothelium and Cardiomyocytes. Cells above the 92th quantile across each distribution were selected as likely belonging to a trajectory. The union of these two sets is highlighted using the Log-likelihood ratio colour scale. For instance, the comparisons involving the cell fate ‘Erythroid3’ indicate higher likelihood towards its trajectory for those cells approaching the red colour, and vice versa, cells approaching the blue colour indicate higher likelihoods towards either endothelial cells or cardiac progenitors.
Fig. 3.7: 3D UMAP of erythroid differentiation. To follow the atlas landscape cells are coloured by cell types (left) and by time point whether cells likely belong to the erythroid trajectory (right). Cells above the 95th quantile were selected as part of the erythroid trajectory.
3.2 The origin of haemato-endothelial lineages

Fig. 3.8: Force directed layout of cells likely belonging to haemato-endothelial trajectories. Cells above the 92th quantile across each distribution were selected as likely belonging to a trajectory. These cells were then used for recomputing highly variable genes and generating a new batch-corrected PCA to be used for building the KNN network used for generating the force directed layout. **A.** Cells are coloured by cell type. **B.** Cells are coloured by cell type sub-cluster. Endothelial cells (EC), Endothelial Cells, Myeloid (My), Megakaryocyte (Mk). **C.** Cells are coloured by time point. **D.** Cells are coloured according to the log-likelihood ratio interval of Endothelium/Erythroid3.
3.2.2 | The role of master regulators in lineage diversification

Master regulators are usually transcription factors at the top of a gene regulation hierarchy, particularly in regulatory pathways related to cell fate and differentiation. Canonical examples of master regulators include *Pou5f1*, *Sox2*, and *Nanog*, all involved in maintaining pluripotency in stem cells [177]. In heamato-endothelial diversification, a collection of transcription factors have been identified for playing the role of master regulators, such as the cases of *Tal1* [178], *Gata1* [179] and *Etv2* [180]. The reconstruction of developmental trajectories using W-OT allows the inspection of gene expression trends across time in relation to a target set of cells (i.e., cell fate). To extract gene trajectory trends, the expression values are weighted according to the ancestors mass distributions at each time point. Following this procedure, the above mentioned haemato-endothelial master regulators were inspected together with other transcription factors of none or little relevance for this specific cell fate, but significant importance for other mesodermal fates (Figure 3.9). These trends showed that the haemato-endothelial master regulators are significantly expressed before and during the emergence of blood and endothelial lineages (see the fraction of cell types across time in [82]-Figure 1).

Intuitively, the approach of trajectory trends can be implemented to perform differential expression analysis between cell fates across time. This could be particularly useful for identification of early markers. Mathematical modelling of gene-gene dependency can also be exploited for the study of master regulators and for building gene regulatory networks. These branches were also explored during the project. However, such work still requires to be reviewed as well as supported by experimental validation. Thus, it was left out of the scope in this thesis.

3.3 | Perturbations in development: a *Tal1* Chimaera

Gene knockout (KO) models are widely used to study the function of genes, including their role in development. Previous work has described the critical role of the basic helix–loop–helix (bHLH) transcription factor TAL1 (also known as SCL) in haematopoiesis; in these experiments, *Tal1*--/ mouse embryos died of severe anaemia at around E9.5 [181]. To overcome these difficulties, chimeric mouse embryos have been generated in which *Tal1*--/tdTomato+ mouse embryonic stem cells (mESC) were injected into wild-type blastocysts. In the resulting chimaeras, wild type cells still produce blood cells, and this allows the specific effects of TAL1 depletion to be studied in an otherwise healthy embryo [182]. The original publication of the single cell atlas of mouse gastrulation and early organogenesis...
3.3 Perturbations in development: a *Tal1* Chimaera

**Fig. 3.9:** Mean expression levels weighted by the ancestral distribution probabilities from all mesodermal fates involved in this landscape, which are indicated by the line colours (see Figure 3.3). The role of haemato-endothelial transcription factors: *Tal1*, *Gata1*, and *Etv2* is highlighted (top). The remaining transcription factors are included for comparison purposes.
showed how atlases of development could be used as reference for the study of perturbations using chimaera models [82]. To determine whether *Tal1* mutant cells were associated with abnormalities in specific lineages, tdTomato- (wild type) and tdTomato+ (*Tal1*/*-*) cells from chimeric embryos at E8.5 were profiled using scRNA-seq. The mutant cells showed complete absence of red blood cells compared to a normal wild type experiment [82].

### 3.3.1 Exploring the role of *Tal1* in blood production at E7.5

The observation of complete absence of red blood cells due to the lack of *Tal1* at E8.5 reflects its major role during blood development. Nonetheless, it reveals relatively little about the molecular mechanisms behind it. While the consequences of the KO in blood production are clearly observed, the underlying cause originated at earlier stages. The gene expression trend of *Tal1* in the atlas (Figure 3.9) showed a sharp increase from E6.5 to E6.75 for the haemato-endothelial fates, being Erythroid3 above Endothelium at E6.75. Notably, both gene trends reach their peak at E7.25, when only few cells annotated as either blood or haemato-endothelial progenitors are identified in the atlas. This analysis motivated the generation of a new *Tal1* chimaera experiment, this time at E7.5. This 10X Genomics dataset was processed following the same low level analysis pipeline described in Section 2.4. The UMAP of this chimaera after annotation transferring (methodologically described in Section 2.5.1) showed overall only few cells belonging to the either Blood or Haemato-endothelial progenitors (Figure 3.10). A wild type negative control (only with tdTomato+ and tdTomato-) experiment was also processed, in which a larger proportion of these progenitors types was observed (data not shown) suggesting that even with presence of non-mutant cells, haemato-endothelial development is affected. Then, in order to identify the landscape region and time point where blood progenitors are blocked due to the *Tal1* KO, the chimaera data was integrated with the analysis of the Erythroid3 trajectory (Figure 3.7). That is, the chimaera data was projected against their closest cells in the atlas and only those mapping against a cell in the blood trajectory were retained (Figure 3.11). The resulting landscape revealed a strong separation between *Tal1*/*-* tdTomato+ and tdTomato- cells along the blood trajectory. The mutant cells are not able to differentiate into blood progenitors. Moreover, the larger proportion of these cells mapped to the Epiblast (recall the UMAP in Figure 2.2). Despite the Chimaera was collected at E7.5, all mutant cells mapped at time points equal or earlier than E7.25. In contrast, non mutant cells in the chimaera as well as the independent negative control wild type experiment showed more advanced time points across mapping proportions. As suggested above, there may be also a systematic effect in blood development in the embryo as consequence of *Tal1*/*-* tdTomato- cells.
3.3 Perturbations in development: a *Tal1* Chimaera

**Fig. 3.10:** UMAP of E7.5 *Tal1* chimaera experiment. Cells are coloured by transferred cell type annotations from the mouse atlas (top) and by mutation status (bottom) indicated by tdTomato+ (*Tal1*+/−) and tdTomato- (wild type).
Fig. 3.11: UMAP plot showing all the cells of the atlas (116,312 cells) (recall the UMAP in Figure 2.2) highlighting cells from Erythroid3 trajectory (cells above the 92th quantile in the ancestors distributions) that were identified as closest cells in euclidean distance with respect to the E7.5 Tal1 chimaera experiment. Mutant cells (tdTomato+) are coloured in red and non-mutant cells (tdTomato-) in blue. The bar plots at the bottom show the proportion of chimaera cells mapping time points in the atlas along the Erythroid3 trajectory. There are two bars for each chimaera experiment with the corresponding tdTomato- and tdTomato+ cells: the Tal1-/- and the wild-type negative control (WT).
The fact that E7.5 mutant cells mapped almost exclusively into the Epiblast in a context of blood differentiation is unexpected, as the role of \textit{Tal1} in such fate is thought to be played at further stages of development, such as during the diversification of mesodermal progenitors. However, these findings are consistent with previous observations in the analysis over the striking differences between the ancestors distributions of blood and other fates (Endothelium and Cardiomyocytes) at time points E6.5-E7.25 (Figure 3.4), as well as the \textit{Tal1} mean expression trends along the haemato-endothelial trajectories displaying an early burst in expression levels in the interval E6.5-E7.5 (Figure 3.9). Moreover, no other fates showed similarly striking patterns on the chimaera cells mapping to their trajectories (e.g., Figure 3.12). Not even My and Mk identified as Myeloid and Megakaryocyte progenitors respectively. An immediate step before planning any further downstream analysis was to visually inspect \textit{Tal1} expression levels in the atlas, and due to 10X Genomics large number of dropouts, a higher coverage (Smart-seq2) publicly available mouse embryonic dataset \cite{85} was added (Figure 3.13). Having observed low but existing \textit{Tal1} signals in Epiblast cells, the following step was to leverage publicly available multi-omics (scNMT-seq) of embryonic development profiled at overlapping time points E6.5 and E7.5 \cite{70}. The goal was to look for epigenetic marks that indicated \textit{Tal1} activity. This multi-omics dataset was annotated using cell types transferred from the atlas (as described in Section 2.5.1), and the transcriptomes are essentially obtained following a Smart-seq2 approach. Despite having a relatively low number of cells (1,105 cells for the whole experiment), it was ideal to inspect the correlation of \textit{Tal1} expression with chromatin accessibility and methylation rates on its own promoter, as \textit{Tal1} actually binds to its own promoter \cite{183} (Figure 3.14). Interestingly, at E6.5 \textit{Tal1} locus chromatin accessibility and methylation rates were positively and negatively correlated respectively with \textit{Tal1} expression in the Epiblast, as one would expect from an active gene. Furthermore, the signal is not detected at E5.5 and seems to disappear by E7.5. All this evidence together could be explained by a burst of \textit{Tal1} expression in the interval E6.5-E7.5 pushing cells in a very fast differentiation process. Then, following from the atlas time course experiment it was natural to explore the dynamic of unsplicing (simplified with the phase portrait of \(u\) and \(s\), where \(s\) means spliced mRNAs and \(u\) unspliced mRNAs as explained in Section 1.7.1). The resulting relations of \(u\) and \(s\) agreed with the hypothesis of a burst of \textit{Tal1} expression starting and finishing along this interval (Figure 3.15). The largest values of \(u\) are reached at E7.0 (and quickly decrease afterwards), that is the previous time point where the peaks of the haemato-endothelial trajectories is reached (see Figure 3.9). To provide a notion of the mRNA kinetics mathematical modelling using the RNA velocity equation can be performed, but rather than constant, considering time-dependent parameters \(\alpha(t)\), \(\beta(t)\) and \(\gamma(t)\). This is still ongoing work.
Resolving Haemato-Endothelial diversification

Fig. 3.12: Chimaera cells projected into multiple atlas trajectories. Cells above the 95th quantile of the ancestors distributions were selected for reconstructing each trajectory. Endothelial cell (EC); Megakaryocyte (Mk); Myeloid (My). Recall from Figure 3.2.
Fig. 3.13: Inspection of *Tal1* expression patterns in early mouse development. A. *Tal1* expression patterns across the atlas (recall from Figure 2.2). B. *Tal1* expression patterns of cells from Scialdone et al. 2016 [85] after projection into the atlas.
Fig. 3.14: Multi-omics inspection of *Tal1* during gastrulation (E5.5-E7.5). A. *Tal1* expression levels vs chromatin accessibility rates at its own promoter. A positive correlation is detected at E6.5. Almost no *Tal1* expression was observed in cells at E7.5. B. *Tal1* expression levels vs methylation rates at its own promoter. A negative correlation is detected at E6.5, indicating potential activity of *Tal1* in the epiblast. This, in concordance with the observations of a positive correlation of *Tal1* expression and chromatin accessibility. C. Distribution of chromatin accessibility rates for cells that either express *Tal1* (TRUE) or not (FALSE) across stages E5.5-E7.5. The biggest difference is observed at E6.5, where larger accessibility rates are associated with expression. D. Distribution of methylation rates for cells that either express *Tal1* (TRUE) or not (FALSE) across stages E5.5-E7.5. In correspondence with the observation of accessibility rates at E6.5, methylation rates are lower in cells expressing *Tal1*. 
3.3 Perturbations in development: a *Tal1* Chimaera

*Fig. 3.15: Tal1* splicing dynamics shows transcriptional burst in the Epiblast at E7.0. Scatter plots of normalised unspliced and spliced mRNA counts. Each plot highlights (in black) cells belonging to a given time point in the interval E6.5-E7.5. The last plot shows all cells together coloured by time points. It is at E7.0 where a significant amount of unspliced counts is detected, suggesting a burst of expression that quickly disappears afterwards.
3.4 | The mRNA kinetics of Erythroid differentiation

The recently introduced concept of RNA velocity in single cells provided an elegant strategy for inferring the future state of a cell [138], which then can be leveraged to predict differentiation trajectories (refer to Section 1.7.1 for a broader explanation). The most advanced computational framework of RNA velocity is implemented in scVelo [141]. This software includes three alternatives for fitting the gene splicing dynamics across cells. These are named 'Deterministic', 'Stochastic' and 'Dynamical'. Whereas the former strategy is equivalent to the original implementation (velocyto, [138]), the 'Stochastic’ and ‘Dynamical’ implementations generalize this concept by relaxing previously made assumptions of a common splicing rate across genes and the observation of the full splicing dynamics with steady-state mRNA levels. The ‘Stochastic’ version extends the deterministic ordinary differential equation (ODE) model to account for stochasticity by treating transcription, splicing and degradation as probabilistic events. This implementation also works under the rationale of a steady-state assumption but shows larger capabilities for capturing the full dynamical model. The ‘Dynamical’ version is a likelihood-based model that solves the full gene-wise transcriptional dynamics of splicing kinetics, which is governed by two sets of parameters: reaction rates of transcription, splicing and degradation, and cell-specific latent variables of transcriptional state and time. The parameters are inferred iteratively via Expectation-Maximization (EM).

For a given estimate of reaction rate parameters, latent time points are assigned to each cell by minimizing its distance to the current phase trajectory (unspliced/spliced phase portraits). The transcriptional states are assigned by associating a likelihood to respective segments on the trajectory (i.e., induction, repression and active and inactive steady state as previously illustrated in Figure 1.7). The overall likelihood is then optimized by updating the model parameters of reaction rates.

At the extent of more expensive computations, the 'Dynamical’ implementation was applied to the atlas of mouse development across the stages E6.5-E8.5 (Figure 3.16). Intriguingly, RNA velocity correctly identified Epiblast cells as the starting point, but several trajectory predictions at later stages were inconsistent with both real time ordering and existing knowledge (Figure 3.16). The most striking discrepancy concerned red blood cell maturation, with velocity-inferred trajectories opposing the true differentiation path. Such patterns of discrepancy have been observed in several other independent experiments of erythrocyte differentiation (also in humans) [184]. These observations motivated a much deeper analysis focused on the genes displaying pronounced dynamic behaviour, which are systematically detected via their characterization by high likelihoods in the dynamic model. It was found that many of these genes display a steep increase of unspliced counts in the Erythroid3
population, leading to a reverse velocity prediction, progressing from Erythroid3 to earlier populations (Figure 3.17.A). After fitting a linear regression through each population and each gene for testing whether the inferred slopes reflected the expected order based on biological knowledge (Figure 3.17.B), 73 genes were selected and termed as Multiple Rate Kinetics or MURK genes (a few examples are shown in Figure 3.17.C). Unsupervised gene ontology analysis confirmed that biological functions essential for red blood cells were highly enriched, including "erythroid maturation" "gas transport" "iron ion homeostasis" and "heme biosynthetic process". Then, scVelo calculations were repeated without MURK genes and the discrepancies were corrected (Figure 3.18.A). Notably, as highlighted in Figure 3.18.B the latent time estimated after correction is also consistent to what seems to be the emergence of the yolk sac definitive blood wave (as suggested above 3.8).

3.5 Extending the haemato-endothelial landscape

The extended atlas introduced in Section 2.5 includes four additional time points covering early organogenesis (E8.75-E9.5) with respect to the original atlas (E6.5-E8.5) that up to this stage has been used for studying the transcriptional landscape of blood development. The integrated time course experiment (E6.5-E9.5) then, covers the two above mentioned waves of blood production: the Primitive and Yolk-Sac definitive. Thus, providing a unique opportunity for a broader exploration of the underlying differentiation trajectories. Nonetheless, trajectory reconstruction gets particularly challenging due to multiple potential origins of blood cells as well as the fact that blood cell populations become much more heterogeneous as consequence of several emerging cell types. This, in addition to highly unbalanced population sizes at latest time-points implied that the quantile threshold approach used in previous analyses is far from suitable and therefore, a more sophisticated and quantitative measure for selecting cells in a trajectory was required.

The new strategy was based on the so-called W-OT fate matrix. In brief, a W-OT fate matrix is a transition probability matrix from all cells to a number of target cell sets at a given time point, commonly the ending point of the time course experiment (see Section 1.7.2 for a the mathematical definition). To fully cover the haemato-endothelium, a W-OT fates matrix was computed for all blood and endothelial cell populations in the landscape (Erythroid, Megakaryocyte, Megakaryocyte-erythroid progenitors (MEP), Erythroid-myeloid progenitors (EMP), Blood progenitors, Haemato-endothelial progenitors, as well as the endothelial populations named Yolk-Sac, Embryo proper, Allantois, Venous and Endocardium) leaving every other cell in the extended atlas grouped as “Other fate”. To further focus on the endothelial landscape a different fates matrix was computed where only the endothelial fates
Fig. 3.16: UMAP layout containing single-cell transcriptomes belonging from E6.5 to E8.5, coloured by sampled time point (left) and by cell-type (right). The overlaying arrows result from applying the scVelo pipeline to the whole embryonic dataset and represent inferred developmental trajectories. Arrowheads highlight the erythroid branch, displaying scVelo trajectory predictions that are inconsistent with real-time sampling.
Fig. 3.17: Identification of multiple rate Kinetics (MURK) genes along yolk-sac erythropoiesis. 
A. Phase plots of representative scVelo driver genes, with scVelo model prediction overlayed and highlighting the limitations of the model when dealing with MURK genes (right). B. Strategy followed to select MURK genes (i.e., fitting a linear regression through each population and each gene for testing whether the inferred slopes reflected the expected order based on biological knowledge). C. Examples of MURK genes identified in yolk sac erythropoiesis.
Fig. 3.18: Removal of MURK genes for correcting seVelo predictions. A. The discrepancies with time and existing knowledge of RNA velocity vector fields (arrows) are corrected after MURK genes removal. B. Latent time estimated with seVelo not only matches expected patterns but also recapitulates the emergence of the second wave of blood production (yolk-sac definitive).
were explicitly included. These two fates matrices contained the transition probabilities from all cells towards each of the above mentioned collection of cell sets (the rows of this matrix add up to 1).

To identify cells that form a differentiation trajectory towards any of the above mentioned cell fates, for every cell in the landscape, the likelihood probabilities associated with these fates and the remaining cells at E9.25-E9.5 (i.e., cells grouped as "Other fate") were compared using the log odds or ratio of probabilities. The log odds is defined as the logarithm of the ratio of probabilities of two different categoric and mutually exclusive outcomes $\log(p/1 - p)$.

Following this concept, three main comparisons were performed:

1. The Primitive and Yolk Sac (YS) definitive landscape:

$$\log_{10}[P(\text{All blood fates} + \text{YS endothelium})/P(\text{All remaining cells})].$$

2. The full haemato-endothelial landscape:

$$\log_{10}[P(\text{Full haematoendothelium})/P(\text{All remaining cells})].$$

3. The endothelial landscape:

$$\log_{10}[P(\text{All endothelial fates})/P(\text{All remaining cells})].$$

The resulting log odds were visualised on the whole atlas UMAP (Figure 3.19). These ratios provided a quantitative setting for more interpretable thresholds when selecting cells that potentially belong to a cell fate trajectory. Then, as previously shown in Figure 3.8 one can take advantage of such cells to generate new manifolds including only potentially fated cells towards blood and endothelium. That is, selecting pluripotent and mesodermal cells above a reasonable log odds threshold, recomputing highly variable genes, correcting for batch effects and generating new force directed layouts from inferred trajectories, as those indicated by the arrows in Figure 3.19.

Notably, red blood cells displayed relatively large $\log_{10}(\text{odds})$ values on the ExE Ectoderm as consequence of two confounding factors: the ancestors of ExE Ectoderm (as well as ExE Endoderm and Parietal Endoderm) are not included in this dataset, and there are not ExE Ectoderm cells detected at late time points because these cells contribute to the embryonic portion of the placenta, and they are not sampled at later time points. The Erythroid fate is probably the earliest and fastest mesodermal trajectory, in the absence of ExE exctoderm ancestors as well as cells among the rest of fates at E9.5 (grouped as "Other fate"), the mass allocated to red blood cells is superior to the average of other fates. When computing fate
Fig. 3.19: Haemato-endothelial log₁₀(odds) of fate probabilities across the mouse extended atlas landscape (UMAP). Potentially fated cells can be extracted by selecting cells above a given log₁₀(odds) threshold with respect to the fate of interest. The pseudocount 0.1 is added to both terms of the log odds in order to avoid undefined operations. A. UMAP layout containing single-cell transcriptomes belonging from E6.5 to E9.5, also referred as the extended atlas (left). Cells are coloured based on the log₁₀(odds) interval of blood and Yolk-Sac endothelium fate probabilities at E9.25-E.95. Then, cells with a log₁₀(odds) above -0.5 were selected as potentially fated towards the canonical two waves of blood production (i.e., Primitive and Yolk-Sac definitive) and a new force directed layout was generated to visualise the resulting trajectories (right). B. Following previous strategy for the generation of a new layout, here cells are coloured based on the log₁₀(odds) interval of Haemato-endothelium fate probabilities at E9.25-E.95. Cells with a log₁₀(odds) above -1 were selected as potentially fated towards the full haemato-endothelial landscape. C. Likewise, cells are coloured based on the log₁₀(odds) interval of Endothelium fate probabilities at E9.25-E.95 and cells with a log₁₀(odds) above -1 were selected as potentially fated towards the endothelial landscape.
Fig. 3.20: Confounding factors in W-OT trajectory reconstruction. UMAP layout containing single-cell transcriptomes belonging from E6.5 to E9.5. Cells are coloured by time point (left). ExE ectoderm is highlighted to emphasize that the vast majority of these cells belong to early time points of the landscape (< E8.5). Both ancestors and descendants of ExE ectoderm are not included in this dataset. An artificial signal towards blood cells is observed as consequence. When computing fate matrices at earlier time points (where ExE Ectoderm cells are still present) this artefact tends to be corrected (right).
matrices at earlier time points (where ExE Ectoderm cells are still present) this artefact tends to be corrected (Figure 3.20).

The resulting three landscapes aim to capture all the trajectories involved in the corresponding fates. The proof of concept for this approach refers to the well described Primitive and Yolk Sac definitive blood waves (Figure 3.21.A). Both, existing knowledge and sampling time matched the landscape generated. Interestingly, it suggests that the divergence of Erythroid and Myeloid progenitors is followed by the Erythroid and Megakaryocyte progenitors across the Yolk Sac definitive trajectory. However, this could be just an effect of visualisation. Notably, the landscape included non Yolk Sac endothelial cells as potential part of the trajectories (e.g., Epicardium and Venous endothelium). This, despite the fact that such fates were explicitly included when computing the fates matrix. Intrigued by these observations, and in order to explore hemogenic potential in the context of the entire Haematoendothelium. The landscape was extended to include all blood and endothelial fates (Figure 3.21.B). The complexity of this new landscape was significantly higher. As a first sight, it revealed a number of different origins of endothelial cells potentially related with blood fates. That is, in addition to the Yolk-Sac origin there were at least two different paths towards Endothelium arising from Embryo proper tissue, and one through the Allantois. To further improve visualisation over the multiple origins of endothelial progenitors another landscape was generated in which blood fates were excluded (Figure 3.21.C).

In general, the most surprising observation came from the full Haematoendothelium landscape where a relatively small subset of cells apparently descending from Embryo proper endothelium and annotated as Cranial mesoderm, converged with the Yolk Sac definitive EMP population (Turquoise arrowhead in Figure 3.21.B). Furthermore, there were several non Yolk Sac (Embryo proper) cells also annotated as EMPs. In fact, cells from every embryo dissection were present in the trajectory associated with EMPs (Figure 3.22.A). When clustering this landscape (using louvain), these cells were grouped together (Cluster 20 in Figure 3.22.B) indicating common molecular programs among these cells. However, visual inspection of gene markers and sub-clustering analysis of the EMP population had previously revealed a highly heterogeneous structure (data not shown). For example, cluster 9 is a relatively small set from the EMP population that showed very specific patterns of lymphoid and Natural killer (NK) cells. Then, to identify differences and similarities between the above mentioned cell populations from cluster 20 (Yolk-Sac Erythroid-myeloid-progenitors (YS-EMP), Embryo proper-Erythroid-myeloid-progenitors (EP-EMP), Embryo proper endothelium (EPE) and Cranial mesoderm), a broad collection of gene markers for blood, endothelium and haematopoietic stem cells was manually constructed. This collection included the MoLO genes [185], as they are associated with stemness in adult hemaetopoiesis.
Then, gene expression values within these populations were averaged and row normalised (i.e., transformed into a Z-score) to generate heat map (Figure 3.22.C). As expected, this analysis revealed both, distinct as well as shared patterns between these populations. Both EMPs, Yolk Sac and Embryo proper derived have expression of genes that are well known markers for the Yolk Sac definitive wave (e.g., *Tal1*, *Runx1*, *Gata1*, *Was* and *Klf1*). Genes associated with erythropoiesis showed in general higher expression values in YS-EMPs, a few of them were exclusively expressed in this population (e.g., *Hba-x* and *Hbb-y*). In contrast, a subset of gene markers associated to endothelial, Haematopoietic Stem Cells (HSC) and other blood progenitors (such as microglial or T-cells) were enriched in EPE-EMPs (*Vwf*, *Gata3*, *Elane*, *Gfi1*, *Ly86*, *P2ry12*, *Adgre1*). The set of cells annotated as Embryo proper endothelium (EPE) that reached cluster 20 showed low expression of some endothelial markers (such as *Cdh5* and *Klf4*) and variable expression across MoLO genes (e.g., *Neil2*, *Rnf168*, *Ccne2* and *Cenph* compared to *Procr*, *Vwf*, *Kif4* and *Ifitm1*). Interestingly, the population annotated as Cranial mesoderm showed clearer patterns associated with HSCs precursors. That is, a larger proportion of MoLO genes with high expression values (having the exception of *Procr*) as well as other dominating expression patterns that included for instance *Etv2* and *Hlf*. The latter reported to be an specific marker of HSCs precursors and not EMPs [186], while also the possibility of HSCs emergence in head mesoderm has been reported [174]. Other relevant genes such as *Tal1* were also expressed but at lower levels than in EMPs. In general, this analysis supports the hypothesis of non Yolk Sac hemogenic endothelium, and reveals a level of complexity in blood development not previously addressed using scRNAseq where multiple research questions can be explored.
Resolving Haemato-Endothelial diversification

Fig. 3.21: Force layouts displaying the reconstruction of Haemato-endothelial developmental trajectories. **A.** Primitive and Yolk Sac haematopoiesis reconstruction. From left to right, cells are coloured by time point, embryo dissection, and cell type annotations. The latter highlighting the Primitive and Yolk Sac definitive blood waves in red and blue respectively. Haemato-endothelial progenitors (HEP); Blood progenitors (BP); Erythroid (ERY); Yolk Sac endothelium (YSE); Erythroid-Myeloid progenitors (EMP); Megakaryocyte-Erythroid progenitors (MEP). **B.** Haemato-endothelial diversification full reconstruction (i.e., also including non Yolk Sac endothelial cells). Likewise, cells are coloured by time point, embryo dissection, and cell type annotations. The latter highlighting the presence of Embryo proper derived cells grouped with Yolk Sac EMPs, both named EP-EMP and YS-EMP respectively. **C.** Multiple origins of endothelial cells are highlighted with arrows. Yolk-Sac Haemato-endothelial progenitors (YS-HEP); Allantois Endothelium (AE); Embryo proper endothelium (EPE).
Fig. 3.22: Dissecting the origins of EMPs. **A.** Force layout of the haemato-endothelial trajectory landscape where cells are displayed separately for each embryo dissection to highlight (with arrows) the presence of cells from Yolk Sac and Embryo proper dissections towards differentiation of EMPs. Notice that the landscape labeled as ‘pooled cells’ contains Yolk Sac cells. **B.** Force layout haemato-endothelial trajectory landscape. Cells are coloured by louvain clusters. Cluster 20 and 9 are highlighted for containing cells fated towards the EMPs fate. Also, heterogeneous cell groups within cluster 20 are highlighted (Ys-EMP, EP-EMP, EPE and Cranial mesoderm). **C.** Heatmap showing row normalised expression values of genes associated with haemato-endothelial development and haemotopoietic stemness.
3.6 | Discussion

The reconstruction of haemato-endothelial developmental trajectories is a crucial step for addressing unanswered questions about how these lineages are specified at the molecular level. The experimental design of the extended atlas provides a unique opportunity for in vivo dissecting the molecular origin and dynamics of multiple blood lineages at single cell resolution, covering the emergence of so-called Primitive and Definitive blood waves. The computational inferences here performed are the backbone of more detailed work over specific lineages. Nevertheless, the trajectory analysis here performed already contributes to the discussion regarding important questions in the field. For instance, it is still debated whether blood and endothelium share a common origin. At earlier stages (E7.0-E7.5) the population identified as haemato-endothelial progenitors clearly bifurcates from a mesodermal cell into the primitive wave and endothelial cells, suggesting a common progenitor. However, primitive erythropoiesis differentiation seems initially defined by a much faster process according to the trajectory distributions, the W-OT KLD between the cardiac, endothelial and red blood cells lineages at earlier stages, and the gene expression trends of master regulators *Tal1* and *Etv2*. Erythropoiesis is apparently triggered by a burst of *Tal1* expression which then fluctuates after E7.5, matching with a more similar distribution of differentiating progenitors with respect to other mesodermal fates. By default WO-T initialises the growth rate $\hat{g}(x) = 1$ (i.e, all cells grow at the same rate) and then, the output row-sums of the optimal transport map $\hat{\pi}_{t_{i+1}}$ is used as a new estimate for the relative growth rate function, meaning that the optimisation problem is solved iteratively. Assuming $\lambda_1$ is well balanced in terms of constraint flexibility on the row sums, the cost function (in this case the euclidean distance) also plays a critical role on mass allocation and thus, the mass transported is heavily influenced by transcriptional similarity between time points, and not only by the size of a target population in further time points. In any case, one has to be careful when drawing conclusions from comparing trajectory distributions between population of significantly unbalanced sizes (for instance Cardiac and Erythroid) as proliferation and differentiation rates will together influence their shape. In fact, the shape of the distribution reflects cell fate commitment when it becomes bimodal. Ongoing population analysis performed by Melania Barile inspired by the concept of pseudodynamics [187] showed proliferation rates of blood progenitors at early stages are consistent with these observations. While controversial, this analysis suggests that cells at early stages as the Epiblast are potentially biased for red blood differentiation.

Subsequently from the primitive wave, *Tal1* levels increase once again around E8.0-E8.25, in particular over the Yolk-Sac hemogenic endothelium, which is where the Definitive
3.6 Discussion

The blood wave originates according to this landscape. Another interesting observation from the landscape is that the emergence of either Erythroid-Myeloid-Progenitors (EMPs) and Megacaryocytes-Erythroid-Progenitors (MEPs), occurs mostly but not exclusively during the Definitive blood wave. Furthermore, this analysis suggests that blood cells can be generated from Embryo proper endothelium in addition to Yolk-Sac endothelium. Common gene markers are shared between these populations, but the erythroid fate is under-represented in embryo proper hemogenic endothelium, in contrast with Myeloid cells. Interestingly, several alternative paths for making endothelial cells are observed. The most surprising finding was that cells descending from Cranial mesoderm are likely able to give rise to certain blood cells that do not share canonical markers such as Runx1 but express several other genes that have been associated with Haematopoietic Stem Cells (HSCs).

From a computational perspective the incorporation of explicit time courses into trajectory inferences provides a critical advantage to reconstruct complex trajectories such as the haemato-endothelial landscape. Pseudotime methods in this context are susceptible to dynamical reconstructions that do not correspond to the real observations. Furthermore, it is increasingly difficult to dissect multiple converging paths into a single lineage. The methods applied in this analysis are powerful computational frameworks with relatively sophisticated theoretical basis. The methods were chosen because both of them work under the rationale of a continuous differentiation process. They do not rely on clusters and unidimensional topological structures such as trees or graph abstraction and they are essentially probabilistic methods when it comes to trajectory inference. Moreover, there are ongoing extensions on these methodologies. For example, the implementation of RNA velocity will soon be able to model time-dependent rates of transcription, splicing and degradation. One will be able to include protein translation by adding another equation in the dynamical system, while simultaneous measurements of transcripts and proteins can be obtained using technologies such as CITE-seq. Alternatively, epigenetic features can also be incorporated into the model. Likewise, W-OT can be extended for multi-omics approaches, more sophisticated cost functions and modelling of gene regulatory networks.

Both methods are elegant examples of mathematical applications in biology. However, these methods still suffer from relevant limitations. The low detection rate of unspliced mRNAs in common scRNA-seq applications implies even nosier and sparse expression matrices. The pre-processing of unspliced counts often requires smoothing or imputation which might lead to artificial signals. The normalisation and batch effect correction do not necessarily follow the same assumptions. More complex scenarios such as the presence of genes with multiple kinetics across several cell diverging populations (i.e., differential kinetics) have to be considered. W-OT is notably a robust and mathematical formal method. While the
use of optimal transport to model developmental processes is not unique and significant improvements have been claimed through the use of autoencoders [188, 189]. However, the implementation of these methods remains challenging and wider benchmarking analyses are required.

In W-OT, the incorporation of unbalanced transport allows consideration of cell death and growth. However, one has to be careful as the sampling of each time point might not be representative of the population dynamics in the system. That is the case of the extended atlas, where the exponential increase of cells is not necessarily reflected across the time course experiment in terms of the absolute number of cells. On the other hand, the stringent approach for computational validation using geodesic interpolation provides certain confidence that the computed transport maps are not heavily affected by this factor. It is also important to point out that eliminating cell types where ancestors and descendants are not sampled might be desirable to avoid artificial associations (e.g., as the case of Extraembryonic Ectoderm and Erythrocytes). Fundamentally, W-OT uses the euclidean distance as cost function. Thus, it is also susceptible to geometric interpretation of this similarity measure. In both cases, scVelo and W-OT, processing the extended atlas required considerable computational resources. Computing the transport maps for the entire dataset reached approximately 700 GB of RAM and took more than 24 hours using 30 threads. The dynamical recovering with scvelo used 120 GB and ran for over one week. While subsampling is an option, it is clear that the size of modern datasets are challenging for these methods and more efficient techniques for memory use have to be developed and used by common users of scRNAseq datasets.
Dissecting the primitive streak at E7.5 to explore cell fate bias

During gastrulation, the Primitive streak extends anteriorly until reaching the most distal part of the embryo, where the node is formed. While the first cells to egress through the proximo-posterior region of the PS will mainly become part of the extra-embryonic mesoderm, cells from the mid-primitive streak will form the lateral plate mesoderm as well as the paraxial and cardiac mesoderm. It has been described that the proximal regions in the Primitive Streak also favour Haematopoietic progenitors. Nonetheless, the underlying molecular mechanisms behind cell fate bias still require deeper studies to be understood. This context motivated a lineage tracing experiment using grafted cells associated with specific Primitive streak portions, followed by scRNAseq of cells from individual embryos over four dissected portions from the Primitive streak ranging from distal to proximal regions. This experiment was then integrated with the publicly available mouse atlas of gastrulation and early organogenesis, including the reconstruction of trajectories carried out in the previous chapter. Standard scRNAseq analysis revealed heterogeneous gene expression patterns enriched for biological process associated with development of multiple fates. By taking advantage of W-OT trajectories, cells from the primitive streak were projected against the atlas for then predicting potential fate bias. Having lineage tracing results obtained from grafting experiments, it was possible to perform a comparison between computational predictions and experimental observations. The level of concordance is high, which highlights the potential of W-OT for reconstructing incredibly complex landscapes such as the entire mouse development atlas.
**Declaration** This work is the result of joint effort put in motion in Berthold Gottgens’s and Marella de Bruijn’s laboratories. I alone performed the entire computational analysis. Both head researchers, as well as Senior researcher Christina Rode designed the study and supervised the project. Christina Rode performed all the experiments involved in this project such as embryo dissections and library preparations. Two Smart-seq2 experiments were sequenced: one at Bertie’s lab and the other one at the sequencing facility from University of Oxford. John Marioni provided guidance on the computational side of the project.
4.1 Introduction

Perhaps one of the main questions that can be approached by the trajectory reconstruction analysis is to unveil the timing at which cells start to commit towards either one or a set of specific cell fates. While time course experiments such as the mouse development atlas provide already a notion, it is important to keep in mind that any conclusion drawn from analysis relies on the assumption that these inferences are correct. The level of trust provided to such predictions usually depends on the robustness of the computational method(s) applied, as well as their consistency with well established biological knowledge. However, to confirm such predictions experimental support is necessary. In this context, in vivo experiments have shown that, in the early gastrulating embryo, nascent mesoderm emerging from the proximal region of the Primitive streak migrates to the Yolk-Sac and differentiates into primitive erythrocytes and endothelium [190]. Also, phenotypic analysis based on cell grafting experiments of early mouse development performed at Marella’s lab with focus on the Yolk Sac and intra-embryonic haemato-endothelial origins, revealed that distinct fate patterns are associated with the different primitive streak portions where cells are located. Furthermore, as a tdTomato line was used to label all cells coming from a particular primitive portion it became possible to effectively trace several of the other lineages. These observations motivated a Smart-seq2 scRNAseq experiment designed to explore at what extent the gene expression signatures of cells from different primitive streak portions of mouse embryos at E7.5 reflect the potential fate bias observed with grafting experiments. Two independent Smart-seq2 experiments were performed, one with a single embryo and another one with two. These E7.5 embryos were dissected in order to extract consecutive primitive streak portions (from the most proximal to the most distal regions) for further sequencing. Figure 4.1 shows a schematic explanation of the primitive dissections performed for this project. In order to assess the consistency between computational predictions of fate bias and the experimental observations associated with the primitive streak portions, the computational approach used here took advantage of the W-OT trajectory reconstruction analysis across the interval E6.5-E8.5 performed with the original atlas, as described in the previous chapter (Section 3.2.1). Thus, highlighting the possibility of leveraging the atlas as a complementary data resource for exploring a wide variety of questions beyond more common designs (e.g., effects of perturbations, as well as In vitro versus In vitro and cross-species comparisons). While this is still an ongoing project, it has already provided useful insights about the molecular programs associated with the different regions of primitive streak dissections as well as the consistency between computational predictions and experimental observations.
**Fig. 4.1:** Primitive streak dissections at E7.5. Work with ‘goose-neck’ side lights: 1. Harvest uterine horns from pregnant dams. 2. Cut in-between each embryo ‘bead’ with spring scissors. If necessary cut away any excess uterine tissue. 3. Carefully grab edge of uterine tissue left around the decidua and peel/rip away with other forceps. Peel away uterine tissue from decidua and discard. 4. Decidua has a slight triangular elliptical shape with a thick red band running around the thicker end. Using spring scissors cut through or roughly below the middle of that red band, exposing the ectoplacental cone. Discard the cut decidua portion. 5. Slide one side of the forceps along the conceptus into the rest of decidua. Close forceps and that should have made a cut through one side of the decidua wall. Using the other forceps carefully peel away the decidua walls to expose the conceptus. 6. Slide forceps all around the conceptus to separate it from the decidua. The conceptus can be picked out by carefully grabbing the ectoplacental cone, where necessary. Discard rest of decidua. 7. Transfer concepti into a clean dish and using 2 pairs of syringe needles cut away the reddish Reichert’s membrane from the conceptus, without damaging the embryo. If necessary leave the ectoplacental cone intact. 8. Stage the conceptus and separate embryos based on stages. Using syringe needles cut conceptus in half along the embryonic/extra-embryonic boundary. 9. Determine the distal-most point of the embryonic portion and cut through. Continue cutting away the posterior portion of the embryo. It is possible to make a straight cut from the distal-most point straight up, then unfold the posterior half and cut away the ‘flanks’ to narrow the tissue to the primitive streak. 10. Flip the posterior portion, so you view it from inside the epiblast. Narrow the flanks if necessary. A. Determine the midpoint of the posterior portion and cut straight across. B. Determine the midpoint of the posterior-most half and cut straight across. C. Determine the midpoint of the distal-most half and cut straight across. One should end up with 4 primitive streak portions. Take care not to disrupt the right PS order before pipetting them into tubes. Use an Eppendorf pipette to carefully aspirate individual primitive streak portions into individual tubes (for sorts etc.) or buffer bubbles (for grafts).
4.2 Low level analysis of Smart-seq2 experiments

Sequencing reads were aligned against the mm10 genome following the GRCm38.95 genome annotation (Ensembl 95) using STAR version 2.5 [95] with --intronMotif and 2-pass option mapping for each sample separately (i.e., --twopassMode basic) to improve detection of spliced reads mapping to novel junctions. Sequence alignment map (SAM) files were then processed with samtools-1.6 in order to generate the corresponding count matrix with htsq-count from HTSeq, version 0.12.4. Cells with total counts lower than 10,000 reads were filtered out from downstream analysis based on observed library size distributions (Figure 4.2.A). Both count matrices, each from an independent sequencing batch were merged into a single matrix of dimension: 28,647 genes and 713 cells. The mitochondrial fraction of reads per cell and library complexity were computed as part of the QC (Figure 4.2.B-C). Only two cells showed a mitochondrial fraction larger than 4% and the lowest number of genes detected was 3,113. Thus, no cells were excluded during QC. Size factors were computed with scran and log transformed normalised counts were obtained with logNormCounts function from scater using size factors centred at unity prior to calculation of normalized expression values (Figure 4.2.D). Then, 181 highly variable genes were extracted using ModelVar function from scran, according to a significant deviation above the mean-variance fitted trend (BH-corrected p<0.05). These genes were selected for computing a PCA. The top 50 PCs were retained for batch correction using reducedMNN, which is an updated version of fastMNN (see Section 2.4.7). Figure 4.3 shows the two-dimensional PCA, before and after batch correction. The resulting batch-corrected PCs were used for clustering and visualisation in downstream analysis.

4.3 Mapping primitive streak cells against the atlas

Following the strategy to transfer annotations from the atlas to other datasets previously described in Section 2.5.1, cell type annotations were assigned to primitive streak cells according to the 15 nearest neighbours between the two manifolds (query data and atlas reference). The annotation transfer was performed separately for each batch to avoid introducing an extra confounding factor when applying MNN batch correction. Also, previous to mapping against the atlas, genes with total counts lower than 100 were filtered out from the Smart-seq2 datasets due to the large differences in dropouts. This, to decrease noise levels when merging Smart-seq2 and 10X count matrices. The resulting annotations were visualised using a UMAP (Figure 4.4). A large fraction of these cells mapped to pluripotent cell types such as Epiblast or Primitive Streak. However, cells mapping to mesoderm, ectoderm and
Fig. 4.2: Low level analysis of Smart-seq2 primitive streak cells at E7.5. **A.** Histogram of library size per batch highlighting the threshold of $10^4$ for filtering out low total count cells. **B.** Histogram of mitochondrial count fraction. **C.** Library complexity (i.e., number of detected genes) vs library size. **D.** Size factors as estimated with *scran* vs library size.
**Fig. 4.3:** Batch correction of Smart-seq2 primitive streak cells at E7.5. **A.** PCA of primitive streak cells highlighting strong batch effects. Cells are coloured by batch (left) and primitive streak portion (right). **B.** Batch-corrected PCA of primitive streak cells. Likewise, cells are coloured by batch (left) and primitive streak portion (right).
endoendoderm lineages were detected in considerable numbers, indicating a certain degree of fate bias already present in the Primitive Streak portions. The projection of primitive streak cells into their closest neighbours in the atlas revealed clear differences across Primitive Streak portions (Figure 4.5), mainly between A and D, the most proximal and distal portions respectively. The patterns observed were concordant with previous work in Marella’s lab with one exception, there were a few cells from portion D annotated as blood progenitors (Figure 4.6). In contrast, such fates supposed to be associated mainly with portion A. That is, according to experimental observations cells from portion D were unlikely to contribute in blood production. Further inspection of blood gene markers such as T\textit{al1} and \textit{Gata1} showed that these cells were correctly classified (Figure 4.6.A-B). However, such cells were only observed in one out of three embryos. While maternal contamination was discarded due to expression of \textit{Hbb-y} which is a marker of yolk sac blood islands, fetal liver, and embryonic erythrocytes, other type of contamination was still possible. Interestingly, these cells belong to the only female embryo. Such cells were retained for downstream analysis as they were notably few, and due the lack of a strong argument to discard them.

4.4 Clustering and Differential gene expression

To explore cell heterogeneity within primitive streak dissections, unsupervised clustering was performed using \textit{igraph} louvain’s algorithm (Figure 4.7). Followed by differential expression analysis for between each cluster and the rest. This was performed using \textit{scran findMarkers} function, and the comparisons were focused over up-regulated genes in each cluster (i.e., setting the parameter \textit{direction} = up). According to \textit{findMarkers} ranking, the top 15 differentially expressed genes of each comparison were selected (all with BH-adjusted p-values < 0.05). Then, genes with an average log$_2$ Fold Change $>$ 3 were used to generate a heat-map of Z-score transformed expression values (Figure 4.8). Despite these cells were collected from the Primitive Streak, they already show heterogeneous gene expression patterns that correspond to mapped cell types. Gene set enrichment analysis performed with GO [191] showed a variety of biological processes statistically associated (Fisher exact FDR-adjusted p-values < 0.05) with embryonic development that include multiple cell fates such as cardiac, haemato-endothelial, mesenchymal, somitogenic, brain and gut development (among others). One can also explore correlations between Primitive Streak and cell fate characteristic genes. The following section addresses this as well as the comparisons between computational inferences and experimental observations.
4.4 Clustering and Differential gene expression

**Fig. 4.4:** UMAP layout of primitive streak cell transcriptomes at E7.5. This layout the *umap-learn* algorithm provided as input with batch-corrected PCs. **A.** Cells are coloured by primitive streak portion. **B.** Cells are coloured by mapped cell type.
Dissecting the primitive streak at E7.5 to explore cell fate bias

**Fig. 4.5:** Projection of primitive streak cells at E7.5 into the mouse development atlas. A. UMAP layout containing single-cell transcriptomes belonging from E6.5 to E8.5 with cells coloured by cell type. A. The closest cells between Primitive Streak and the atlas are highlighted according to the corresponding dissected portions.
Fig. 4.6: Primitive Streak portion D cells annotated as blood progenitors. A. UMAP layout of primitive streak cell transcriptomes at E7.5, highlighting dissected portion D cells annotated as blood progenitors (with a brown arrowhead). B. UMAP layout of primitive streak cell transcriptomes at E7.5, displaying gene expression levels of blood development markers (*Tal1*, *Gata1*, *Etv2* and *Hbb-y*).
**Fig. 4.7:** UMAP layout of primitive streak cell transcriptomes at E7.5. Cells are coloured by transferred cell type annotation.
Fig. 4.8: Heat-map of differentially expressed genes between unsupervised clusters of Primitive Streak (PS) cells.
Dissecting the primitive streak at E7.5 to explore cell fate bias

4.5 | Predicted vs Observed fates of Primitive Streak cells

To explore potential cell fates among Primitive Streak dissections beyond mapped cell types, W-OT trajectories inferred in the atlas (see Section 3.2.1) were leveraged. It was reasoned that closest neighbouring cells in the atlas could be taken as reference for potential trajectories of Primitive Streak portion dissected cells. By selecting a set of target cell fates from the atlas, the allocated mass resulting from trajectory analysis associated with the closest neighbouring cells in the atlas can be assessed in quantitative terms. Then, a set of diverse cell fates was selected for this analysis and a special type of visualisation was designed, here referred as fate plots (Figure 4.9). However, even after following the explanation of these plots provided in the figure description, it remains challenging to draw conclusions from them as they contain a lot of information. Thus, to summarize the relation between cell fates and Primitive Streak portions, the number of cells above a fate probability threshold \( \log_{10}(mass) > -5 \) was extracted in such way that only 'Likely' fates across the different portions were considered (Figure 4.10.A). The association between mapped cell types and primitive streak portions was summarized following the same strategy (Figure 4.10.B). Together, these two plots showed a clear association between some cell fates and the corresponding portions of the Primitive Streak. Cells from the most proximal portion (A) were primary associated with some mesodermal fates such as allantois, cardiomyocytes, mesenchyme and haemato-endothelial progenitors. For the latter, keeping in mind the unexpected case of portion D cells displaying a blood phenotype. In contrast, cell fates such as Notochord and Def. endoderm are clearly favoured towards the most distant portion (D).

In this context, a similar approach was followed for the lineage tracing experiments performed with grafted cells (see Figure 4.11), such that another bar plot was generated from grafting experimental observations (see Figure 4.12). The cell fates show relatively concordant patterns across fate probabilities, mapped cell types and experimental observations. In summary, this work provided transcriptomic evidence at single cell resolution of the relation between developmental fate and primitive streak regions.
### Fate probabilities across Primitive Streak cells

**Fig. 4.9:** Fate plots of Primitive Streak portion dissected cells. Bar plots show along the x axis, the proportion of mapped cell types per Primitive Streak portion (A, B, C and D). Within the bars and along the y axis, dots represent cell fate probabilities associated the closest neighbouring cells in the atlas with respect to Primitive Streak cells. The range of values goes from $\log_{10}(mass)$ values of 0 to -15. This range was divided in three intervals (highlighted by dotted lines within the bar plots) to define three potential fate states: Likely (i.e., $[0, -5]$), Undetermined (i.e., $[-5, 10]$), and Unlikely (i.e., $[-10, 15]$).
Dissecting the primitive streak at E7.5 to explore cell fate bias

**Fig. 4.10:** Association between Primitive Streak dissected portions and computational predictions of cell fates. 

**A.** Bar plot displaying the number of cells allocated as 'Likely' giving rise to specific cell fates. That is, Primitive Streak cells reflecting a closest neighbour in the atlas with $\log_{10}(mass) > -5$.

**B.** Bar plot displaying the number of Primitive Streak cells annotated as specific cell types through mapping against the atlas.
4.5 Predicted vs Observed fates of Primitive Streak cells

**Fig. 4.11:** Schematic representation of primitive streak cell grafting experiments. Primitive streak (PS) portions are highlighted within an embryo donor (left). Orthotopic transplantation of E7.5 PS grafts is performed (right), followed by molecular profiling of grafted cells at E8.0 and post 24 hrs culture to establish fate relationships between computational predictions and experimental observations.

**Fig. 4.12:** Association between Primitive Streak dissected portions and experimental observations of cell fates. Bar plot displaying the number of Primitive Streak cells that gave rise to specific cell types by means of grafting experiments. Yolk-Sac endothelial cells (YS/EC), Allantois endothelial cells (Allantois/EC), Embryo proper endothelial cells (Em/EC), Lateral plate mesoderm (LPM).
4.6 Discussion

This experiment presented an innovative approach for studying cell fate commitment by incorporating single cell transcriptomics to assess the molecular mechanisms associated with fate bias across different proximal-distal regions of the Primitive streak in the E7.5 mouse embryo. That is, by linking the transcriptional profiles obtained from dissected cells across contiguous primitive streak portions with the time course guided trajectory reconstruction performed with W-OT for the mouse atlas of gastrulation and early organogenesis. Furthermore, by means of cell grafting experiments, the inferences made about fate bias within cells profiled with scRNAseq were contrasted with actual lineage tracing observations. Thus providing experimental support concerning the level of trust a researcher could have from W-OT computational predictions. The transcriptomic profiles resulting from unsupervised clustering and differential gene expression analysis revealed a heterogeneous collection of cells displaying multiple biological processes associated with a variety of cell fates. In general, the whole analysis supports a certain level of fate bias already present in the Primitive Streak. According to these results, there is an association between fate bias and the different dissected regions of the Primitive Streak (see Section 4.5). Both experimental observations and computational predictions showed that mesodermal populations are favoured in proximal regions of the Primitive streak (with the notable exception of few cells from the most distant primitive streak portion presenting blood phenotypes, a matter that remains under discussion in this project). In contrast, nodal cell types were found to be almost exclusive of distal portions.

Current work in this project involves increasing the number of profiled embryos using mT-mGx23GFP grafting experiments. Subsequently grafted cells with mTmG will be sequenced after culture for scRNAseq analysis (see Figure 4.13). 23GFP was used as the project has particular interest in the Haematoendothelium, which the 23GFP transgene marks [192]. This experiment will provide clearer molecular signatures associated with observed fates that might lead to functional experiments. Moreover, the accuracy of cell type annotation transferring using the atlas will be directly comparable with observed fates in the culture. The mass allocated to mapping cells in the atlas using W-OT trajectories can be then traced back in time using the ancestor distributions to determine the contribution Primitive streak cells towards observed fates, and perhaps, can be used to evaluate the possibility of predicting the primitive streak portion of origin across Primitive streak cells in the atlas.
Molecular analysis of grafted cells and comparison to fate trajectories

**Fig. 4.13:** Experimental design for molecular analysis of grafted cells followed by single cell transcriptomic profiling. Grafted cells with \( \text{mTmGx} \text{23GFP} \) from pre-defined Primitive streak regions ranging from proximal to distal are transferred from an embryo donor to a wild-type recipient. The embryo is cultured for lineage tracing and molecular profiling. Subsequently, scRNA-seq is performed.
Conclusion and future directions

5.1 About fundamental questions

Throughout this dissertation, there are fundamental questions about the nature of developmental processes that remain unanswered and require further discussion as they imply strong assumptions that determine our approaches and interpretations of the system. For instance, the formal definition of cell type and cell states (i.e., what is a cell type?), whether differentiation dynamics at a molecular as well as a population level is a discrete or a continuous process, and whether cell fate commitment is probabilistic or deterministic. These are essential concepts that have great impact on the overall computational analysis here presented and more generally, on the way developmental atlases should be built and leveraged. The tantalizing prospect of a systematic definition of cell types and cell states during development that mimics the structure of the periodic table of elements [27] can only be achieved after resolving or reaching a consensus for such elementary questions.

The way these questions are approached by computational and mathematical models determines a large fraction of advantages and limitations of these strategies to resemble a developmental process. For example, discrete models are heavily impacted by the definition of sub-populations in the system (i.e., cell type clusters), as the number of sub-populations increases, more parameters have to be considered likely leading to significant changes over the global paradigm and the interpretation outcome. In contrast, continuous models are able to incorporate internal variables within the modelled cell that constrain or alter the probabilities of different outcomes (e.g., haematopoiesis under the steady state vs stress, or differences in HSCs due to young and old donors). Moreover, the vast majority of models based on discrete populations have been unidirectional, but there is increasing evidence that reversibility is a reality (e.g., dedifferentiation, transdifferentiation and reprogramming [193]).
Incorporating reversibility comes natural when considering continuous models. Also, there is no need for constraining the models to certain topologies. On the other hand, discrete models usually gain interpretability over the continuum, they are computationally and mathematically simpler as well as do not require the inclusion of spatial boundary conditions. Moreover, it is still an active debate between whether the rates of cell differentiation at a population level should be represented as a step-wise progression or as a continuous sigmoid function, even under the assumption that transcriptomic level follows a continuum progression on the multidimensional gene expression space. Innovative experimental designs using scRNAseq have been performed in this context [194]. In 1665, the microscope made this possible for Robert Hooke to discover and define the term cell. In recent times, it has been mainly single cell transcriptomics, the technology that has allowed scientists to push forward this concept.

5.2 The nightmare of batch effects

Along this dissertation batch-effect correction is widely discussed and carefully applied because it represents a major concern in the analysis. Batch-effect confounding factors could easily lead to misleading results that further propagate in downstream analyses or even across multiple experiments where the atlas is used as reference. While strong batch effects in the atlas are effectively corrected by the time-sequential application of MNN over the PCA space for most purposes (using either fastMNN or reducedMNN). More subtle effects that are fully confounded with biological factors (such as within time points) are practically impossible to fully disentangle compromising interpretation of finer levels of relevant biological variation. The challenges increases particularly, for the case of the extended atlas, where different 10X chemistry kits were used and the library complexity of cells in newly generated data is significantly larger than in the original atlas. The differences in the amount of drop-outs (cells that are truly expressing a certain gene show an observed count of zero) between datasets could potentially be imputed from similar expression patterns using sophisticated approaches as Magic [195]. However, the risk of generating artificial signals has to be noticed and remains controversial [196]. Moreover, subtle differences of relevance within cell populations could be potentially removed. A potential approach for a more effective batch-effect correction in gene expression space where large differences in dropouts are observed could benefit from the recently introduced concept of MetaCell [197] in combination with existing methods such as MNN. "MetaCells" constitute highly homogeneous local building blocks for clustering and quantitative analysis of gene expression, that avoid enforcing any global structure on the data, thereby maintaining statistical control and minimizing biases.
5.3 Is RNA velocity useful?

The concept of RNA velocity is particularly attractive due to the need of systematic understanding of developmental process as the majority of computational strategies are limited to descriptive understanding from observed landscapes. The reality of its implementation is however, complicated. First, as the most common experimental protocols are not designed to quantify and distinguish between unspliced and spliced transcripts. The measurements obtained are then incredibly noisy and require more careful consideration when processing. Second, a more sophisticated implementation is required. RNA velocity was initially conceived as a deterministic model. However, incorporating stochasticity (as in scVelo) showed important improvements. Further extensions generalise the model to overcome key assumptions such as solving the equation in the steady state and fixing the splicing rate to be constant across genes. The application of scVelo in erythropoiesis in Section 3.4 highlighted the need for incorporating time dependent rate parameters of transcription, splicing and degradation. This one and other extensions to scVelo are already underdevelopment. The scVelo framework provides metrics to assess the quality of inferences (speed and coherence), but when results are clearly contradicting established knowledge, it becomes challenging to unveil the cause of it, and most of times these analyses are simply discarded. In conclusion, careful interpretations should be made with RNA velocity as it is subject to confirmation bias. However, it is definitely an idea that will hold importance for the reconstruction and understanding of developmental processes.

5.4 Lessons learned from Waddington-OT

Waddington-OT can be seen as the base of this dissertation. It was the chosen method for computational reconstructions not only because the experimental design of the mouse atlas fits perfectly, it also provides a powerful framework that is able to deal with the complexity of the atlas. The method incorporates sampling time, accounts for cell growth and death and models development as a continuum allowing for the computation of fate probabilities. The latter implies several advantages. Two main examples can be given: gene expression trends across time associated with specific trajectories can be recovered without the use of pseudotime methods by pondering gene expression using the fate probabilities. This capability was leveraged to select gene targets and time points for other chimaera experiments not mentioned in this work. Classifying cells as part of one trajectory purely by computational reconstruction will always bring doubts about how certain someone can be about such inferences. The use of fate probabilities provides an alternative perspective to this,
Conclusion and future directions

based on relative comparison between two or multiple cell fates leading to the possibility of identifying subtle bifurcation events across the landscape. Relative measures as the log odds used to reconstruct the extended haemato-endothelial landscape have proven to be really useful defining potential trajectories, as far as one keeps in mind the experimental context and established knowledge for avoiding over-interpretation of results. The majority of observations regarding the haemato-endothelial landscape are consistent with existing knowledge and could probably be recovered using alternative trajectory reconstruction methods. Nonetheless, the case of multiple convergent trajectories into a single fate (e.g., the convergence of primitive and definitive erythropoiesis) remains challenging for most of them. Furthermore, a pseudotime method would hardly identify the Cranial mesoderm subset of cells with inferred potential as EMPs or HSC precursors, as shown in Section 3.5. Of course, this finding is still subject to validation. Extensions to Waddington-OT, and more generally the applications of OT for modelling developmental process are an active area of research. The incorporation of autoencoders, among other advantages allows consideration of multi-omics measurements. Alternative definitions of cost function could provide insights about the impact of distance metrics commonly used to assess transcriptional similarity along differentiation. W-OT framework for learning regulatory models assumes that trajectories are cell autonomous, but might be extended to incorporate intercellular interactions, such as paracrine signalling, by using optimal transport for interacting particles [198, 199]. Lineage tracing experiments introducing barcodes can be used to recognize cells that descend from a recent common ancestor cell but do not currently directly reveal the full gene-expression state of the ancestral cell. However, they might be incorporated into the optimal-transport framework to better estimate temporal couplings.

5.5 When will cell type annotation end?

The annotation of cell types is crucial, and aims to benefit researchers interested on very different lineages. Beside the exhaustive effort required to accomplish this task already mentioned, the fact that expert knowledge is necessary for annotating practically every present lineage makes this process even more challenging. Even when the level of cell type resolution is kept relatively low, a reliable annotation requires the team to achieve new expertise that sometimes can only come from years of studying a particular system. Starting from a "final annotation" to be included in the manuscript in preparation, the strategy to follow for the extended atlas aims to provide an online web interface where the user can not only inspect the data but also suggest and/or challenge current annotations. Thus, decreasing
individual efforts but increasing the quality of annotation through the collective participation of the community benefiting from this resource.

5.6 Key observations

In summary, the following conclusions can be drawn from the analyses performed across the different chapters from this study:

1. The experimental design is crucial for a successful computational analysis capable of revealing biological findings of interest. In this case, relatively short sampling intervals and the inclusion of overlapping time-points between newly generated data and the existing atlas were fundamental factors for having reliable applications of trajectory reconstruction.

2. The importance of low level analysis is sometimes overlooked as consequence of the enthusiasm for exploring the data. However, this step also determines whether all downstream analyses lead to correct and informative results. The importance of good practices for low level analysis increases when the data generated aims to be a resource widely used by the community.

3. The estimation of descendants from all E8.5 cell populations in the atlas, as an initial step to guide the annotation process showed the power of W-OT to resolve such a complex landscape as a whole.

4. Transferring annotations between datasets is a useful alternative when populations are shared. However, when *de novo* annotations are required, a clustering strategy is difficult to avoid because in order to define gene markers from an specific cell type, differential gene expression between cell types remains necessary. A cluster free approach for finding gene markers could be an interesting method to develop.

5. The reconstruction of primitive erythropoiesis revealed a fast differentiation process in concordance to its goal of providing oxygen for the embryo. The exploration of *Tal1* role’s in early blood development using W-OT trajectory analysis, complemented with chimaeras and multi-omics assays suggests an early priming for red blood cells apparently governed by a burst of this transcription factor (*Tal1*) that initiates the primitive wave of erythropoiesis.

6. The reconstruction of the extended landscape of haemato-endothelial diversification (E6.5-E9.5) effectively captured established trajectories for both, primitive and defini-
tive waves of blood production, in concordance with existing knowledge. This landscape also supports the presence of non-hemogenic endothelium as well as potential blood emergence from non-endothelial cells. The level of resolution for the \textit{in vivo} process here achieved has no precedent among scRNAseq datasets. However, further analysis is still required to present a complete landscape of all the blood cell types emerging across this time-course experiment. Ultimately, also functional experiments need to be carried out.

7. The work performed with the Smart-seq2 experiments corroborated the observations of fate bias across dissected portions from the most proximal-distal regions along the Primitive Streak, and provided insights about the degree of cell heterogeneity at the transcriptomic level within this population at E7.5, highlighting biological processes enriched for a diverse set of developmental paths. Moreover, the availability of a complete reconstruction of atlas trajectories allowed the comparison between computational predictions of multiple cell fates and experimental observations from grafting experiments. Both approaches indicated that mesodermal tissues are favoured in the proximal regions while for example, nodal cell types are almost exclusive of distal portions.

### 5.7 Ongoing and future work

Regarding the extended atlas of mouse development, current work is focused on defining a preliminary final annotation to make the dataset public and the construction of the online application for accessing and exploring the data, as well as performing basic tasks with the extended atlas. Further refinement of haemato-endothelial trajectories is needed to present a complete diversification landscape. Similar analyses will be performed regarding heart formation, focusing on the emergence of first and second heart fold progenitors. Exploration of gene markers defining relevant differentiation trajectories will be followed by functional assays. The follow-up on the role of \textit{Tal1} also requires experimental support as consequence of the provocative proposition of a significant degree of bias towards red blood cells as early as in the Epiblast. Lastly, the project involving scRNAseq of Primitive streak dissected portions will integrate newly generated data from post-grafting experiments and potentially complemented with functional assays. From a computational method development perspective, a new approach inspired by W-OT for multi-omic spatio-temporal measurements currently seems as an exciting extension to pursue.
References


References


Appendix A

Experimental methods

A.1 | Embryo collection and processing

All procedures were performed in strict accordance to the UK Home Office regulations for animal research.

A.1.1 | E8.5-E9.5 time-points

Mouse embryos were collected under the project licence number PPL 70/8406. Animals used in this study were 6-10 week-old females, maintained on a lighting regime of 14 hours light and 10 hours darkness with food and water supplied ad libitum. Following wildtype C57BL/6 matings, females were sacrificed using cervical dislocation at E8.5, E8.75, E9.0, E9.25 and E9.5. The uteri were collected into PBS with 2% heat-inactivated FCS on ice and the embryos were immediately dissected and processed for scRNA-seq. For each time-point, four embryos were selected based on morphology and somite counts, in order to span the range expected for the given time-point according to Theiler, K., 1989, [200] and processed individually. The exception is the E9.5 time-point, where embryos were smaller than expected and only two embryos were collected with lower somite numbers. The Yolk-sac was systematically separated from the rest of the embryo and processed as a separate sample. Of the four selected embryos in each time-point, two were partitioned in defined anterior-posterior sections dissociated as separate samples, and two were dissociated as bulk and the suspension then divided into separate 10X samples. For the E8.5 partitioned embryos, they were divided in two with the section being made at the level of somite 4. The anterior portion (including headfolds, branchial bars and heart rudiment) and the posterior portion (including allantois, hindgut, primitive streak) were processed as individual samples. For E8.5 bulk-dissociated embryos, the single-cell suspension was divided into two 10X samples.
For E8.75-E9.5 partitioned embryos, they were divided into 3 segments, with sections being made below the otic pit and below the heart (at the level of somites 10 to 12). Anterior-most section (includes brain structures anterior to rhombomere 6 and branchial bars), mid-section (includes the heart and remains of viteline vessels) and posterior section (includes allantoic structures, hindgut and posterior-most somites) were then singularized and further processed as separate samples. Single-cell suspensions were prepared by incubating the samples with TrypLE Express dissociation reagent (Life Technologies) at 37 °C for 7 min under agitation, and quenching in PBS with 10% heat-inactivated serum. The resulting single-cell suspension was washed and resuspended in PBS with 0.4% BSA, and filtered through a Flowmi Tip Strainer with 40 µm porosity (ThermoFisher Scientific, # 136800040). Cell counts were then assessed with a haemocytometer. Single-cell RNA-seq libraries were generated using the 10X Genomics Chromium system (version 3 chemistry), and samples were sequenced according to the manufacturer’s instructions on the Illumina NovaSeq 6000 platform.

A.1.2 **Tal1 Chimaera**

TdTomato-expressing mouse embryonic stem cells (mESC) were derived as previously described [201] from E3.5 blastocysts obtained by crossing a male ROSA26tdTomato (Jax Labs - 007905) with a wild-type C57BL/6 female. The cells were negative for mycoplasma contamination. The cells were expanded under the 2i+LIF conditions [202] and transiently transfected with a Cre-IRES-GFP plasmid [203] using Lipofectamine 3000 Transfection Reagent (ThermoFisher Scientific, #L3000008) according to manufacturer’s instructions. Single GFP+ cells were sorted 48h post-transfection into 96-well plates. Individual clones were allowed to grow and were manually picked for expansion. A tdTomato-positive, male, karyotypically normal line, competent for chimaera generation as assessed using morula aggregation assay, was selected for targeting *Tal1*. Two guides targeting exon 4 were designed using the http://crispr.mit.edu tool (guide 1: GAACCCACTATGGAAAGAGA; guide 2: GAGGCCCTCCCCATATGAGA) and were cloned into the pX458 plasmid (Addgene, #48138) as previously described [56]. The resulting plasmids were then used to transfect the cells as detailed above. Single transfected clones were expanded and assessed for Cas9-induced mutations. Genomic DNA was isolated by incubating cell pellets in 0.1 mg/ml of Proteinase K (Sigma, #03115828001) in TE buffer at 50 °C for 2 hrs, followed by 5 min at 99 °C. The sequence flanking the guide-targeted sites was amplified from the genomic DNA by polymerase chain reaction (PCR) in a Biometra T3000 Thermocycler (30 sec at 98 °C; 30 cycles of 10 sec at 98 °C, 20 sec at 58 °C, 20 sec at 72 °C; and elongation for 7 min at 72 °C) using the Phusion High-Fidelity DNA Polymerase (NEB, #M0530S) according to the manufacturer’s instructions. Primers including Nextera overhangs were used (F-
GTCTCGTGCTCGGAGATGTGTATAAGAGACAGTTGCCCTCCATTATGTA GTCTCGGCACGTACGATGTGTATAAGAGACAGTATTCCAGCCAGCATTIT, allowing library preparation with the Nextera XT Kit (Illumina, #15052163), and sequencing was performed using the Illumina MiSeq system according to manufacturer’s instructions. An mESC clone showing a 77 base-pair deletion in exon 4 inactivating *Tal1* gene expression was then injected into C57BL/6 E3.5 blastocysts. Chimaeric embryos were subsequently transferred into recipient females at 0.5 days of pseudopregnancy following mating with vasectomised males, as described previously [204].

Chimaeric *Tal1* and wild-type embryos were harvested at E7.5 (4 and 16 embryos respectively), dissected, and single-cell suspensions were generated from pooled embryos as described above. Given the low detection rate of the tdTomato transcript, single-cell suspensions were sorted into tdTomato+ and tdTomato- samples using a BD Influx sorter with DAPI at 1µg/ml (Sigma) as a viability stain for subsequent 10X scRNA-seq library preparation (version 2 chemistry) and sequencing on an Illumina HiSeq 4000 platform, which resulted in 27,817 tdTomato- and 28,305 tdTomato+ cells that passed quality control. Flow cytometry of chimaeric embryos was performed in parallel using a BD Fortessa cytometer. Cells were stained with the conjugated antibodies CD45-APC-Cy7 (1:200; BD Pharmingen, cat# 557659, clone 30-F11, lot# 6126662), CD41-BV421 (1:200; Biolegend, cat# 133911, clone MWReg30, lot# B216311), Ter119-PerCP-Cy5 (1:200; Biolegend, cat# 116227, clone TER-119, lot# B169767) and CD71-FITC (1:400; BD Pharmingen, cat# 553266, clone C2, lot# 2307673) with Fc block CD16/32 (1:100; eBioscience, cat# 14–0161-85, clone 93, lot# 4316103), and DAPI at 1µg/ml (Sigma) as a viability stain. For the wildtype-into-wildtype experiment, a parental tdTomato+ Tal1+I- line was injected into C57BL/6 E3.5 blastocysts and processed as for the Tal1-/- samples. Four pooled embryos were used for scRNA-seq of Tal1-/- samples, and 785 tdTomato- and 449 tdTomato+ cells passed quality control. While sixteen pooled embryos were used for the wildtype-into-wildtype scRNA-seq, and 4,418 tdTomato- and 5,350 tdTomato+ cells passed quality control. Chimaera sample sizes were dependent on the number of viable embryos that did not show excessive global biases towards host or injected cells (i.e., very low or high fluorescence).

**A.1.3 Primitive streak dissected cells library preparation**

Following the dissection protocol described in Section 4.1, three single E7.5 embryos were collected and processed as two batches for further sequencing of cells from all four primitive streak portions. Single-cell suspensions were prepared by incubating the embryos with TrypLE Express dissociation reagent (Life Technologies at for 37 °C seven min and quenching with heat-inactivated serum. Single cells were subsequently stained with 4’,6-
diamidino-2-phenylindole (DAPI) as viability stain (1 µg/ml; Sigma). Live cells were isolated by fluorescence-activated cell sorting (FACS) using a BD Influx sorter into individual wells of a 96-well plate containing lysis buffer (0.2% (v/v) Triton X-100 and 2 U/µl SUPERase-In (Invitrogen, AM2696)) and stored at -80 °C. Plates were processed following the Smart-seq2 protocol as previously described [19] and libraries were generated using the Illumina Nextera XT DNA preparation kit. Libraries were pooled and sequenced on an Illumina HiSeq 4000.