

Predicting Response to Neoadjuvant Chemotherapy in Breast Cancer with Gene Expression and Computational Pathology



Wei Cope

St Edmund's College, University of Cambridge

Department of Oncology

This dissertation is submitted for the degree of

Doctor of Philosophy

April 2021

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. In each chapter a preface sets out my personal contribution to the work and the assistance I received from others:

This dissertation is within the word limit requirements set by the Clinical Medicine Degree Committee and the Board of Graduate Studies.

I confirm that no part of this dissertation has previously been submitted for any other qualification.

Acknowledgements

First and foremost, I would like to thank all the patients who gave their consent and took part in the trials and studies in this thesis - without them the work would have been impossible. I am also indebted to the Cambridge Cancer Centre for funding my research.

I would like to express my deepest gratitude to my supervisors, Professor Carlos Caldas and Dr Elena Provenzano. I am privileged to have been given the opportunity to spend four years with their guidance and support, especially through difficult times. I am extremely grateful to Dr Suet-Feung Chin and Dr Oscar Rueda, who taught me so much with great patience, and made the projects a great learning experience. I am also very grateful to Professor Paul Edwards for his invaluable advice and suggestions.

I thank Dr Ali Dariush and Dr Raza Ali for their collaboration. I also wish to thank Dr Stephen-John Sammut, Dr Kate Eason and Raquel Manzano Garcia for sharing their experience.

I extend my deepest gratitude to Helen Bardwell for her meticulous work behind the scenes. I am also grateful to the Histopathology core facility at CRUK for their assistance.

I wish to thank everyone in the Caldas group, for making it such a caring, collaborative and inspirational place; especially Marion Karniely, for looking after us all.

It is with my deepest appreciation that I thank my parents for all that they have done for me; my parents-in-law and sisters-in-law for their support over the years; and my daughters Eleanor and Catherine for the love and joy they bring. My special thanks to my husband, Thomas, for always being there with passion and love, at my best and worst.

Dedication

I dedicate this thesis to my daughters, Eleanor and Catherine, who bring sunshine wherever they go.

Abstract

Predicting Response to Neoadjuvant Chemotherapy in Breast Cancer with Gene Expression and Computational Pathology

Wei Cope

Many breast cancer patients are treated with chemotherapy before surgery for the removal of the tumour, which is known as neoadjuvant chemotherapy (NAT). It improves outcome for many patients, and their response to the treatment is prognostic for their overall outcome. However, not all patients who receive NAT respond to the treatment, and, as a result, many suffer unnecessarily from side effects and delays to surgery. In this thesis, I designed and evaluated two potentially independent and complementary strategies for predicting an individual patient's response to treatment based on gene expression and computational pathology.

Firstly, I attempted to develop a clinically practical approach for classifying breast tumours using RNA from routine formalin fixed paraffin embedded (FFPE) histopathological samples. Breast cancers can be classified into distinct groups, 'integrative clusters' (IntClusts), with different outcomes and potentially different response to NAT. The published classification is based on RNA expression from fresh frozen tissue, which is impractical in a clinical setting. Initially, I attempted to build an accurate classifier for IntClusts based on RNA expression from widely available FFPE tissue, using a user-friendly NanoString technique. Unfortunately, it was not possible to achieve reliable classification with this method using the gene probe set pre-selected based on fresh frozen tissue. Next, I sought to identify the genes whose expression was not affected by the fixative process by comparing paired FFPE and fresh frozen tissue with RNA sequencing, a method that allows the quantification of all genes. There was poor agreement in measured gene expression between the two types of tissue sample, FFPE versus frozen, and between the assessment methods on the same tissue type, RNA-sequencing versus Illumina microarray, resulting in unreliable classification of

tumours into integrative clusters. These findings represent a challenge to the adoption of the integrative clusters in real-world precision medicine.

Secondly, I developed quantitative computational methods for the assessment of digitized H&E slides, which are routinely produced clinically. I developed and validated two machine learning methods for cell classification, where cells on an image can be automatically detected and identified as tumour cells, lymphocytes, or stromal cells. I also show that my method can be effectively generalised to immunohistochemistry slides. In a novel dataset, using a new method, I replicate the previous finding that the presence of immune infiltrate in pre-treatment biopsies is predictive of NAT response. I then used this cell classification to demonstrate that the spatial profiles of tumour clusters and their relationship to immune cells are associated with treatment outcome. Specifically, I found that larger tumour clusters, more heterogeneous tumour clusters, and more lymphocytes in the region immediately bordering tumour clusters are all correlated with pathological complete response to NAT. The spatial features of the peritumour region were more predictive than the features of the tumour itself, suggesting a particularly important role for the interface between tumour clusters and their immune microenvironment.

To conclude, I review the large number of further studies spawned by the work I present in this thesis, and explain how they might improve our ability to understand tumour biology, and translate this understanding to the clinical setting.

In summary, I have explored the real-world applicability of gene expression profiling and computational pathology methods in widely available clinical samples. These approaches have the potential for translation into adjuncts to existing stratification methods, offering patients better care. This thesis provides a step towards the translation of the molecular classification of breast cancer, and computational methods for pathology image analysis, into real-world precision medicine, predicting response to neoadjuvant chemotherapy.

List of Abbreviations

| | |
|---------------|---|
| AACR | American Association for Cancer Research |
| AJCC | American Joint Committee on Cancer |
| AUC | area under the curve |
| BAM | binary alignment map |
| BC | before Christ |
| CE | Conformité Européenne |
| CNA | copy number aberration |
| CRUK | Cancer Research UK |
| CUH | Cambridge University Hospitals |
| DBSCAN | density-based spatial clustering of applications with noise |
| DCIS | ductal carcinoma in situ |
| DE | differential expression |
| DF | degrees of freedom |
| DFS | disease free survival |
| DNA | deoxyribonucleic acid |

| | |
|-----------------|--|
| ER | oestrogen receptor |
| EU | European Union |
| FDA | Food and Drug Administration |
| FDR | full data range |
| FF | fresh frozen |
| FFPE | formalin fixed paraffin embedded |
| FISH | fluorescence in situ hybridisation |
| GATK | genome analysis toolkit |
| GPU | graphics processing unit |
| HTA | Human Tissue Authority |
| IC | integrative clusters |
| IHC | immunohistochemistry/immunohistochemical |
| IntClust | integrative clusters |
| IQR | interquartile range |
| LN | lymph node(s) |

| | |
|-------------|---|
| NAT | neoadjuvant chemotherapy |
| NHS | National Health Service |
| NICE | national institute for health and care excellence |
| NST | no special type |
| PCA | principle component analysis |
| PCR | polymerase chain reaction |
| pCR | pathological complete response |
| PPV | positive predictive value |
| PR | progesterone receptor |
| RCB | residual cancer burden |
| RD | residual disease |
| RF | random forest |
| RFS | relapse free survival |
| RIN | RNA integrity number |
| RNA | ribonucleic acid |
| ROC | receiver operating characteristic |
| rpm | rotations per minute |
| SVM | support vector machine |

| | |
|-------------|--|
| TMA | tissue microarray |
| TME | tumour microenvironment |
| TNBC | triple negative breast cancer |
| TPM | transcripts per million |
| UICC | Union for International Cancer Control |
| WHO | World Health Organisation |
| WLE | wide local excision |

Table of Contents

| | |
|--|-----------|
| 1. INTRODUCTION | 1 |
| 1.1 BREAST CANCER | 1 |
| 1.2 TYPES OF BREAST CANCER | 2 |
| 1.3 HISTOLOGICAL CLASSIFICATION..... | 2 |
| 1.3.1 Histological tumour type | 3 |
| 1.3.2 Histological grade | 3 |
| 1.3.3 Receptor status..... | 4 |
| 1.3.4 Molecular classification..... | 5 |
| 1.3.5 Staging of breast cancer | 8 |
| 1.3.6 Personalised Approach..... | 9 |
| 1.4 TREATMENT OF BREAST CANCER..... | 10 |
| 1.4.1 The development of treatment | 10 |
| 1.4.2 Neoadjuvant therapy | 12 |
| 1.4.3 Response to neoadjuvant therapy | 13 |
| 1.4.4 Factors determining the response to neoadjuvant therapy | 14 |
| 1.5 THE IMPORTANCE OF FORMALIN FIXED PARAFFIN EMBEDDED (FFPE) TISSUE | 16 |
| 1.6 DIGITAL PATHOLOGY IMAGE ANALYSIS..... | 18 |
| 1.7 AIMS AND OBJECTIVES | 21 |
| 2. MATERIALS AND METHODS | 22 |

| | | |
|--------|---|----|
| 2.1 | PREFACE..... | 22 |
| 2.2 | NANOSTRING..... | 23 |
| 2.2.1 | RNA extraction..... | 23 |
| 2.2.2 | Nanodrop..... | 24 |
| 2.2.3 | RNA quality assessment..... | 25 |
| 2.2.4 | RNA hybridisation for NanoString | 26 |
| 2.2.5 | NanoString Platform | 27 |
| 2.3 | RNA SEQUENCING | 28 |
| 2.3.1 | Ribosomal RNA depletion | 28 |
| 2.3.2 | RNA fragmentation and priming | 31 |
| 2.3.3 | First Strand cDNA Synthesis | 32 |
| 2.3.4 | Second Strand cDNA Synthesis..... | 33 |
| 2.3.5 | Purification of double-stranded cDNA..... | 34 |
| 2.3.6 | End prep of cDNA library | 34 |
| 2.3.7 | Adaptor ligation..... | 35 |
| 2.3.8 | Purification of the ligation reaction..... | 35 |
| 2.3.9 | PCR enrichment..... | 36 |
| 2.3.10 | Purification of PCR reaction..... | 37 |
| 2.3.11 | DNA quality assessment and fragment size determination..... | 37 |
| 2.3.12 | Quantitative PCR (qPCR) | 38 |
| 2.3.13 | Preparation for sequencing..... | 38 |
| 2.4 | ANALYSIS OF RNA SEQUENCING DATA | 39 |

| | | |
|-----------|---|-----------|
| 2.4.1 | Quality control..... | 39 |
| 2.4.2 | Alignment..... | 39 |
| 2.4.3 | Pre-processing..... | 39 |
| 2.4.4 | Haplotype calling..... | 40 |
| 2.5 | SLIDE DIGITISATION..... | 41 |
| 2.5.1 | Slide provenance..... | 41 |
| 2.5.2 | Slide Annotation..... | 42 |
| 2.5.3 | Management of Information..... | 42 |
| 2.6 | CELL SEGMENTATION AND CLASSIFICATION..... | 43 |
| 2.6.1 | Validation of the classification methods..... | 47 |
| 2.6.2 | Lymphocyte density..... | 47 |
| 2.7 | CELL NEIGHBOURHOOD ANALYSIS..... | 48 |
| 2.7.1 | Identification of cell clusters..... | 49 |
| 2.7.2 | Defining the peri-cluster region..... | 51 |
| 2.7.3 | Distance between nuclei..... | 52 |
| 2.7.4 | Neighbourhood normalisation..... | 53 |
| 3. | NANOSTRING..... | 54 |
| 3.1 | PREFACE..... | 54 |
| 3.2 | NANOSTRING TECHNOLOGY..... | 55 |
| 3.3 | DESIGNING THE PROBE SET..... | 57 |
| 3.4 | SAMPLE SET..... | 58 |
| 3.5 | RNA EXTRACTION..... | 59 |

| | | |
|--------|---|----|
| 3.6 | RNA QUALITY CONTROL..... | 60 |
| 3.6.1 | RNA Purity | 60 |
| 3.6.2 | RNA Integrity | 61 |
| 3.7 | NANOSTRING RESULTS..... | 62 |
| 3.7.1 | RNA count | 62 |
| 3.7.2 | Initial validation on receptor status..... | 63 |
| 3.7.3 | Validation across genes of interest | 69 |
| 3.7.4 | Relationship to integrative clusters | 72 |
| 3.8 | BUILDING A CLASSIFIER BASED ON NANOSTRING DATA..... | 79 |
| 3.8.1 | Using the existing iC10 package | 79 |
| 3.8.2 | Decision tree and random forest classifier..... | 82 |
| 3.9 | IMPROVING IC10 CLASSIFICATION | 88 |
| 3.9.1 | Posterior probability thresholding..... | 89 |
| 3.9.2 | Combining integrative clusters 3 and 8..... | 90 |
| 3.10 | DISCUSSION | 91 |
| 3.10.1 | RNA quality | 91 |
| 3.10.2 | Probe positions | 92 |
| 3.10.3 | Probe selection..... | 93 |
| 3.10.4 | Gene expression correlation | 94 |
| 3.10.5 | Poorly correlated key genes..... | 95 |
| 3.10.6 | Effect of tissue type..... | 97 |
| 3.11 | CONCLUSION: | 99 |

| | | |
|-----------|---|------------|
| 4. | RNA SEQUENCING | 100 |
| 4.1 | PREFACE..... | 100 |
| 4.2 | BACKGROUND..... | 101 |
| 4.3 | PILOT STUDY..... | 103 |
| 4.4 | LARGER SCALE COMPARISON..... | 103 |
| 4.4.1 | RNA..... | 104 |
| 4.4.2 | cDNA..... | 105 |
| 4.4.3 | Total number of reads aligned..... | 106 |
| 4.4.4 | Correlation of expression within-patient across gene..... | 110 |
| 4.4.5 | Ubiquity of gene expression..... | 114 |
| 4.4.6 | Within-gene across-sample correlations..... | 117 |
| 4.4.7 | Receptor expression..... | 120 |
| 4.4.8 | Previously problematic key genes..... | 121 |
| 4.4.9 | Integrative Cluster genes..... | 122 |
| 4.4.10 | iC10 gene relationships..... | 124 |
| 4.4.11 | Differential Expression Analysis..... | 126 |
| 4.4.12 | iC10 classification using the existing classifier..... | 127 |
| 4.5 | FURTHER INVESTIGATIONS OF DATA QUALITY..... | 130 |
| 4.5.1 | Haplotyping..... | 131 |
| 4.5.2 | Comparison with Illumina Microarray..... | 135 |
| 4.6 | CONCLUSION..... | 137 |
| 5. | DIGITAL PATHOLOGY - CELL CLASSIFICATION | 139 |

| | | |
|-----------|--|------------|
| 5.1 | PREFACE..... | 139 |
| 5.2 | BACKGROUND..... | 141 |
| 5.3 | AIMS AND OBJECTIVES..... | 145 |
| 5.4 | MATERIALS: BREAST CANCER CASES AND THE TRIALS THEY WERE OBTAINED FROM..... | 146 |
| 5.5 | TRANSNEO PATIENT CHARACTERISTICS..... | 150 |
| 5.6 | RESULTS – ALGORITHM PERFORMANCE..... | 152 |
| 5.6.1 | Qualitative properties of the methods, visual inspection..... | 152 |
| 5.6.2 | Cross-validation within H&E training set..... | 155 |
| 5.6.3 | Test against histopathologist on new data..... | 160 |
| 5.6.4 | Immunohistochemistry (IHC)..... | 163 |
| 5.7 | RESULTS – SLIDE PROPERTY VERIFICATION..... | 166 |
| 5.7.1 | Cellular composition of the samples..... | 166 |
| 5.7.2 | Median lymphocyte density..... | 168 |
| 5.8 | DISCUSSION..... | 170 |
| 5.8.1 | Our algorithms..... | 172 |
| 5.8.2 | Challenges..... | 173 |
| 5.8.3 | Potential solutions..... | 176 |
| 5.9 | CONCLUSION..... | 177 |
| 6. | SPATIAL ANALYSIS..... | 178 |
| 6.1 | PREFACE..... | 178 |
| 6.2 | BACKGROUND..... | 179 |
| 6.2.1 | The interaction of tumour cells and immune cells..... | 180 |

| | | |
|------------|--|------------|
| 6.2.2 | The importance of spatial relationships between tumour and immune cells..... | 183 |
| 6.3 | AIMS AND OBJECTIVES | 185 |
| 6.4 | IDENTIFICATION OF TUMOUR CLUSTERS..... | 186 |
| 6.4.1 | Method overview | 187 |
| 6.4.2 | Results..... | 193 |
| 6.4.3 | Multivariate prediction of outcome..... | 197 |
| 6.5 | GRAPH ANALYSIS..... | 201 |
| 6.6 | TUMOUR CELL NEIGHBOURHOOD ANALYSIS..... | 209 |
| 6.7 | CONCLUSION | 213 |
| 7. | DISCUSSION..... | 216 |
| 7.1 | CLINICAL RELEVANCE OF MY THESIS..... | 216 |
| 7.2 | MOLECULAR CLASSIFICATION..... | 218 |
| 7.3 | DIGITAL PATHOLOGY | 220 |
| 7.4 | MULTI-MODAL ASSESSMENT | 223 |
| 8. | SUMMARY | 224 |
| 9. | REFERENCES | 226 |
| 10. | APPENDIX: PROBE SET FOR NANOSTRING | 242 |

List of Figures

| | |
|--|----|
| Figure 1: Proportion of annually published papers mentioning breast cancer and FFPE tissue in their Pubmed citation, 1980-2019 | 17 |
| Figure 2. Example of tiles at L3 level, after exclusion of blank space | 43 |
| Figure 3. Example of nuclear segmentation. Each green cross denotes the centre of a nucleus..... | 44 |
| Figure 4: Examples of nuclear images used in Random Forest classifier..... | 45 |
| Figure 5: Three example clusterings for tumour cells. based on minPts values of 3, 5 and 10. | 50 |
| Figure 6: Tumour clusters and their surrounding peri-cluster regions | 51 |
| Figure 7: Schematic illustration of the calculated inter-nuclear distances..... | 52 |
| Figure 8: NanoString® platform workflow, adapted from NanoString training material | 56 |
| Figure 9: Representation of integrative clusters across the 456 validation samples..... | 58 |
| Figure 10: Extracted RNA concentration for the samples that overlapped with the original study | 59 |
| Figure 11: Extracted RNA concentration for the samples that did not overlap with the original study | 59 |
| Figure 12: Purity of extracted RNA for the samples that overlapped with the original study | 60 |
| Figure 13: Purity of extracted RNA for the samples that did not overlap with the original study | 60 |
| Figure 14: Integrity of extracted RNA for the samples that overlapped with the original study | 61 |
| Figure 15: Integrity of extracted RNA for the samples that did not overlap with the original study | 61 |

| | |
|---|----|
| Figure 16: Total RNA count before and after the application of NanoStringNorm, for all samples | 62 |
| Figure 17: Relationship between ESR1 expression from the NanoString technique and ER status from Immunohistochemistry | 63 |
| Figure 18: Relationship between ERBB2 expression from the NanoString technique and HER2 status from Immunohistochemistry..... | 64 |
| Figure 19: Relationship between ESR1 expression from the NanoString technique and ER status from the Illumina Microarray..... | 65 |
| Figure 20: Relationship between ERBB2 expression from the NanoString technique and HER2 status from the Illumina Microarray | 66 |
| Figure 21: Strong correlation between GRB7 and ERBB7 (HER2) gene expressions as measured by NanoString on FFPE tissue, adjusted $r^2=0.865$ | 67 |
| Figure 22: Very weak correlation between ESR1 (ER) and ERBB2 (HER2) gene expressions, adjusted $r^2=0.013$ | 68 |
| Figure 23: Spearman's rank correlation between gene expression orders across NanoString and Illumina techniques for every sample individually..... | 69 |
| Figure 24: Spearman's rank correlation between sample expression orders across NanoString and Illumina techniques for every gene individually..... | 70 |
| Figure 25: Spearman's rank correlation between gene expression orders across NanoString and Illumina techniques for every sample, broken down by known integrative cluster membership | 72 |
| Figure 26: Cross-correlation matrix within sample across gene, ordered by integrative cluster membership | 73 |
| Figure 27: Heatmap of unsupervised clustering of using data from Illumina restricted to 207 genes Each column is a sample and each row is a gene..... | 74 |
| Figure 28: Heatmap of unsupervised Clustering of using data from NanoString restricted to 207 genes. Each row is a gene, and each column is a sample | 75 |

Figure 29: Heatmap of across-sample correlations for the expression of every gene for the Illumina data, arranged by known integrative cluster. 76

Figure 32: Heatmap of across-sample correlations for the expression of every gene for the NanoString data, arranged by known integrative cluster. 77

Figure 33: Differential gene expression by integrative cluster, normalised across all samples. 78

Figure 34. Gene expressions for the original iC10 training data (left), compared to our NanoString data (right) for the 139 genes present in both datasets..... 80

Figure 35: An example decision tree for classifying a single sample based on all other samples (leave-one-out cross-validation)..... 82

Figure 36: An example of iterative boosting for a single sample across 300 iterations. If the fit were improved then alpha would trend upwards and exceed 1. This was not observed. 85

Figure 37: Distance of the probe sequence from the 3' end of the mRNA for NanoString (y-axis) and Illumina (x-axis) platforms. 92

Figure 38: Correlation of gene expression across platforms for two samples with outlier RSF1 expression on the NanoString platform..... 95

Figure 39: Correlation between CLNS1A and RSF1 expression for Illumina and NanoString platforms. 96

Figure 40: Correlation of gene expression in the iC10 classifier across FFPE and Fresh Frozen tissues quantified using RNA-seq (log scales). 103

Figure 41: RNA concentration (in ng/ml) extracted from fresh frozen and FFPE tissue preparations. 104

Figure 42: cDNA concentration (in nM/ml) extracted from fresh frozen and FFPE tissue types..... 105

Figure 43: Total number of reads in fresh frozen and FFPE tissue types by alignment method. 106

| | |
|--|-----|
| Figure 44: Correlation of total number of reads obtained with STAR and Salmon alignment methods..... | 107 |
| Figure 45: No correlation between total number of reads and RNA concentration | 109 |
| Figure 46: No correlation between total number of reads and cDNA concentration.. | 109 |
| Figure 47: Spearman's rank correlation between gene expression orders for FFPE and fresh frozen tissue for every sample individually. | 111 |
| Figure 48: Correlation in gene expression profile against number of genes detected in both fresh frozen and FFPE samples..... | 112 |
| Figure 49: The number of gene expressions detected in each sample, by sample preparation method..... | 113 |
| Figure 50: Number of genes detected in common for a given number of samples..... | 114 |
| Figure 51: Spearman's rank correlation between gene expression orders for FFPE and fresh frozen tissue for every sample individually, restricted to the 8040 universally detected genes..... | 115 |
| Figure 52: Spearman's rank correlation between TPM sample expression orders for FFPE and fresh frozen tissue for every gene individually, for all genes present in at least 15 sample pairs..... | 116 |
| Figure 53: Spearman's rank correlation between sample expression orders for FFPE and fresh frozen tissue for every gene individually, restricted to the 8040 universally detected genes..... | 117 |
| Figure 54: Spearman's rank correlation between sample expression orders for FFPE and fresh frozen tissue for every gene individually, for all genes present in at least 15 sample pairs. | 118 |
| Figure 55: Spearman's rank correlation between TPM sample expression orders for FFPE and fresh frozen tissue for every gene individually, for the 8040 universally detected genes (left) and for all genes present in at least 15 sample pairs (right). | 119 |
| Figure 56: Receptor gene expression correlations across FFPE and fresh frozen tissues. Scales are TPM log ₁₀ | 120 |

| | |
|---|-----|
| Figure 57: RSF ₁ and CLNS _{1A} gene expression correlations across FFPE and fresh frozen tissues. Scales are TPM log ₁₀ | 121 |
| Figure 58: Number of samples where Integrative Cluster genes were not expressed. . | 122 |
| Figure 59: Total Integrative Cluster gene expressions not detected in each sample ... | 122 |
| Figure 60: Spearman and Pearson correlations within iC ₁₀ genes of interest across-samples. | 123 |
| Figure 61: Log gene expression count for iC ₁₀ genes by tissue type, coloured by sample identity, with a line of unity plotted. | 124 |
| Figure 62: Correlation between gene expression for FFPE vs fresh tissue for a 'good' and a 'bad' sample. | 125 |
| Figure 63: Autocorrelation of allelic variants in fresh frozen samples..... | 132 |
| Figure 64: Autocorrelation of allelic variants in FFPE samples..... | 132 |
| Figure 65: Concordance of allelic variants between FFPE and fresh frozen samples, with fresh samples as the template. | 133 |
| Figure 66: Two example problematic samples for haplotype calling. P _{44D8} (left) has abnormally high concordance with all samples and P _{62F2} (right) has abnormally low concordance with all other samples, including its FFPE equivalent..... | 133 |
| Figure 67: Concordance of allelic variants between fresh frozen and FFPE samples, with FFPE samples as the template. | 134 |
| Figure 68: Within-patient across gene correlation for Illumina Microarray and RNASeq expression..... | 135 |
| Figure 69: Within-gene within-patient correlation for Illumina Microarray and RNASeq expression..... | 136 |
| Figure 70: Examples of cell classification from H&E tissue | 153 |
| Figure 71. Higher magnification illustrative images..... | 154 |
| Figure 72: Manual validation of SVM and RF classifiers on H&E tissue, by cell type. . | 162 |

| | |
|--|-----|
| Figure 73: Manual validation of RF classifier on IHC, by cell type. PPV: Positive predictive value | 165 |
| Figure 74: Cellular composition differences between diagnostic biopsies and surgical specimens..... | 167 |
| Figure 75: Pre-treatment biopsy lymphocyte density by chemotherapy outcome. Data from the support vector machine are plotted, but those from the random forest technique were extremely similar at this scale..... | 168 |
| Figure 76: Pre-treatment biopsy lymphocyte density by receptor status and response to neo-adjuvant chemotherapy. | 169 |
| Figure 77: Examples of common segmentation errors..... | 174 |
| Figure 78: Problematic vesicular nuclei during segmentation. | 175 |
| Figure 79: Examples of patients with similar median lymphocyte density across the whole slide, but showing very different distributions of lymphocytes and different treatment outcomes..... | 182 |
| Figure 80: An example slide comprising multiple tissue cores, correctly identified as separate objects as indicated by the green boundaries..... | 187 |
| Figure 81: The spatial location of all cells identified by segmentation..... | 188 |
| Figure 82: All tumour clusters, identified by my implementation of the DBScan algorithm. | 188 |
| Figure 83: All tumour clusters, with their surrounding buffer zones. | 190 |
| Figure 84: Overlapping buffer zones for neighbouring tumour clusters before trimming. | 190 |
| Figure 85: Tumour buffer zones shaded in random colours, trimmed to exclude areas of overlap with neighbouring tumour clusters..... | 191 |
| Figure 86: Example distributions of lymphocyte density in peri-tumoural regions across three slides. | 192 |
| Figure 87: The number of tumour clusters identified per slide..... | 194 |
| Figure 88: The number of tumour cells in clusters of each size, log base ten..... | 194 |

| | |
|--|-----|
| Figure 89: The association between tumour cluster size and response to neoadjuvant chemotherapy | 195 |
| Figure 90: Median lymphocyte density in tumour clusters (left) and around tumour clusters (right), broken down by response to NAT | 196 |
| Figure 91: Scree plot from principal component analysis for derived spatial measures for peri-tumour regions (left), tumour clusters (middle), and both combined (right). | 198 |
| Figure 92: Factor loadings for the first two principal components in peri-tumour regions (top), tumour clusters (middle), and both combined (bottom). | 199 |
| Figure 93: An example of Louvain community detection for homotypic epithelial (tumour cell) interactions. | 203 |
| Figure 94 Pearson correlation between derived network measures within each environment..... | 204 |
| Figure 95 The effect of microenvironment on each measure..... | 206 |
| Figure 96 The eight measure and compartment combinations that differed significantly between patients who showed pathologically complete response and those who did not. | 208 |
| Figure 97: An example of the running likelihood ratio for tumour cell neighbours in a single core..... | 210 |
| Figure 98: Tumour cell neighbours..... | 212 |

List of Tables

| | |
|---|----|
| Table 1: Intrinsic subtype IHC features and grade | 5 |
| Table 2: Integrative cluster frequency, prognosis, distinguishing molecular features and receptor status..... | 7 |
| Table 3: DNase digestion components..... | 23 |
| Table 4: RNA Hybridisation components..... | 26 |
| Table 5: Probe hybridisation components..... | 29 |
| Table 6: Temperature settings for probe hybridisation | 29 |
| Table 7: RNase H digestion components..... | 30 |
| Table 8: DNase I digestion components..... | 30 |
| Table 9: RNA fragmentation and priming reaction components..... | 31 |
| Table 10: First strand synthesis reaction components | 32 |
| Table 11: Temperature settings for first strand cDNA synthesis..... | 32 |
| Table 12: Second strand synthesis reaction components | 33 |
| Table 13: Temperature settings for second strand cDNA synthesis | 33 |
| Table 14: End preparation reaction components..... | 34 |
| Table 15: Temperature settings for the end preparation reaction | 34 |
| Table 16: Adaptor ligation reaction components | 35 |
| Table 17: Polymerase chain reaction components..... | 36 |
| Table 18: Temperature setting for polymerase chain reaction | 36 |
| Table 19: Tapestation reaction components | 37 |
| Table 20: Quantitative polymerase chain reaction components..... | 38 |
| Table 21: Features of objects calculated by the cell segmentation algorithm..... | 46 |
| Table 22: The genes that displayed the strongest and weakest correlations | 71 |

| | |
|---|-----|
| Table 23: Pearson correlation in gene expression profiles for each integrative cluster between the training data for the iC10 algorithm and our NanoString data..... | 79 |
| Table 24: Confusion matrix for sample classification using the iC10 algorithm on NanoString data | 81 |
| Table 25: Confusion matrix for all samples classified with a decision tree..... | 83 |
| Table 26: Decision tree demonstrating perfect classification using boosting. | 84 |
| Table 27: Confusion matrix for all samples classified with a boosted decision tree. | 86 |
| Table 28: Confusion matrix for a random forest classifier, based on decision trees. | 87 |
| Table 29: Optimal cut-point for discarding uncertain samples to maximise kappa. | 89 |
| Table 30: Confusion matrix and summary statistics for the iC10 classifier, with posterior probability thresholding and the combination of integrative clusters 3 and 8. Absolute sample numbers rather than percentages are shown. | 90 |
| Table 31: Performance of the iC10 classifier with only genes that were highly correlated across platform..... | 94 |
| Table 32: Number of samples sequenced from each integrative cluster | 103 |
| Table 33: Confusion matrix for iC10 classifier based on RNA-seq gene expression from Fresh Frozen tissue samples..... | 128 |
| Table 34: Confusion matrix for iC10 classifier based on RNA-seq gene expression from FFPE tissue samples..... | 129 |
| Table 35: Characteristics of the trials and studies from which digital pathology slides were generated | 147 |
| Table 36: The number of slides annotated and scanned from each study and trial..... | 148 |
| Table 37: Breakdown of slides by stain and tissue type | 149 |
| Table 38: Trans-Neo study tumour characteristics | 151 |
| Table 39: K-fold cross validation results for support vector machine across 5 folds ... | 156 |
| Table 40: Confusion matrix for support vector machine classifier. | 157 |
| Table 41: K-fold cross validation results for random forest across 5 folds | 158 |

| | |
|--|-----|
| Table 42: Confusion matrix for random forest classifier..... | 158 |
| Table 43: H&E Test set object classification. | 160 |
| Table 44: Confusion matrix for support vector machine and random forest on H&E images against pathologist. | 161 |
| Table 45: IHC test set object classification..... | 164 |
| Table 46: Confusion matrix for random forest on test IHC images against pathologist classified cells. Rows represent classified identity, while columns represent true class | 165 |
| Table 48: Derived measures from which outcome prediction was attempted..... | 197 |
| Table 49 Principal component analysis of mean centred, variance normalised data from each microenvironment. First three components shown..... | 205 |
| Table 50 Repeated measures ANOVAs of the effect of microenvironment on each measure. DF: degrees of freedom..... | 207 |

1. Introduction

1.1 Breast cancer

Breast cancer is the uncontrolled growth of cells from breast tissue. It has been independently described all over the world for thousands of years. The Edwin Smith Papyrus, an ancient Egyptian manuscript written in about 3000 BC, described probable breast cancers that were removed by cauterisation, and that there was no cure. Hippocrates described not only the appearance of breast cancer in 460 BC, but also its progressive nature (Lakhtakia, 2014). He coined the term *karkinos*, which describes the sprawling “veins” that looks like a “crab” seen at the cut surface of the tumour (Moss, 2004). This is an early example of what a pathologist would call a macroscopic description. The term *karkinos* has eventually developed into the word we use today: carcinoma. Breast cancer was also documented in China, initially around 1500 BC, and by the 12th century it was noted that it more commonly affected women aged over 40 years old and that the typical survival was 3 years (Jushi). To this day, breast cancer remains one of the leading causes of death. In the UK, there were over 55,000 new breast cancer cases in 2017, and around 11,400 breast cancer deaths. This makes breast cancer the second most common cause of cancer death in women, and fourth most common across all genders in the UK (CancerResearchUK).

1.2 Types of breast cancer

Breast cancer is a heterogeneous entity. Breast tumours can show inter-tumoural and intra-tumoural heterogeneity, not only in histopathological appearance, but also at the genomic, epigenomic, transcriptomic, and phenotypic levels (Stingl and Caldas, 2007). As a result, numerous clinical trials have shown that different subtypes of breast cancer respond differently to treatment and have different prognoses. At present, a combination of evidence-based classification and risk stratification factors are clinically balanced to decide treatment approach. This decision is traditionally influenced by properties of the tumour, such as size, histological type and grade, oestrogen receptor and HER2 status, and stage; as well as characteristics of the patient such as age and comorbidities (NICE, 2018).

However, the traditional histological features used to determine treatment only partially reflect the underlying genetic diversity, and do not permit a truly individualised approach to treatment. As a result, some women are overtreated, and others develop metastatic disease and die despite therapy. Better methods of breast cancer classification are needed that reflect the molecular biology of the tumour, enabling improved prediction of risk and individual tailoring of therapy (Katz *et al.*, 2018).

1.3 Histological classification

Histological classification of breast cancer is determined by pathologists based on the architecture of the tumour and the morphology of the tumour cells. There are two components of traditional histological classification of breast cancer: histological tumour type and tumour grade. Both histological tumour type and grade are reflections of the degree of differentiation of the cancer, or the extent to which it resembles normal glandular breast tissue. Tumours that occur in the breast can be primary, which originate from tissue within the breast, or metastatic from another site in the body.

1.3.1 Histological tumour type

Almost all primary breast cancers are adenocarcinomas, which are cancers derived from glandular epithelium that show evidence of glandular differentiation, such as tubule formation or mucin production. *Invasive ductal carcinoma of no special type* (NST) is the most common type, which accounts for approximately 75% of all breast cancers (Ellis, 2003; Weigelt and Reis-Filho, 2009). This is a group of carcinomas that do not show any of the features of the special types of breast cancer, and yet are a diverse entity in terms of morphology, underlying molecular biology and behaviour that would particularly benefit from more individualised therapeutic approaches. This has been recognised in the growing importance of receptor status in defining treatment decisions, but this categorisation still results in sub-groups that display great heterogeneity in terms of their behaviour.

The other types of breast adenocarcinoma include *invasive lobular cancer* (10 – 15%), with a distinctively discohesive growth pattern, which is a manifestation of a lack of E-cadherin expression (de Leeuw *et al.*, 1997; Vos *et al.*, 1997; Rakha *et al.*, 2010a). Certain subtypes have prognostic value, for example *metaplastic carcinomas* confers poorer survival (Liao *et al.*, 2018). Other types include *medullary-like carcinoma*, a high grade tumour with a better prognosis that is often associated with BRCA1 mutation (Eisinger *et al.*, 1998), and normally displays a molecular expression profile known as *basal-like*. However, most other special types (such as mucinous carcinoma and tubular carcinoma) are not associated with specific mutations in single genes but rather show genetic and molecular diversity. These different subtypes sometimes behave differently and are thus treated differently (Lyons *et al.*, 2000; Yin *et al.*, 2016; NICE, 2018).

1.3.2 Histological grade

As well as histological type, breast cancers are also given a histological grade by pathologists. The Nottingham Grading System is recommended by professional bodies worldwide, including the World Health Organization (WHO), American Joint Committee on Cancer (AJCC), European Union (EU), and the Royal College of Pathologists in the UK (RCPATH) (Rakha *et al.*, 2010b; Provenzano *et al.*, 2015). The

grading system is based on the tumour's architecture, morphology and frequency of mitoses. Specifically, grade is a reflection of the degree of differentiation and is composed of three variables: tubule formation, nuclear pleomorphism and mitotic count. Each variable is given a score of 1-3 with 3 being the most poorly differentiated. They are added together and an overall grade of between 1 (low grade, scoring below 6) and 3 (high grade, scoring above 7) is allocated. Grade 3 tumours are more aggressive than grade 1 tumours.

Histological grade has been shown by various studies to be a good predictor of tumour behaviour and prognosis, especially in smaller or early tumours, where the tumour has not had time to develop and manifest its other characteristics. There is an association between grade and some of the special types of breast cancer, for example tubular cancers are always grade 1, however it has also been shown to be an independent prognostic factor in certain subgroups of breast cancer (Rakha *et al.*, 2010b). However, grading is a subjective assessment requiring a high degree of training and even in best practice settings is prone to inter-observer variance (Robbins *et al.*, 1995; Chowdhury *et al.*, 2006; Longacre *et al.*, 2006).

1.3.3 Receptor status

The tumour's hormone receptor status and HER2 (human epidermal growth factor receptor 2) status are also reported by pathologists, and are used to further stratify patients for treatment and prognosis (Thomas and Berner, 2000; Sopik *et al.*, 2017). Oestrogen receptor (ER) and HER2 are two of the key drivers of proliferation in breast cancer and there are specific drugs that inhibit their action, such as tamoxifen and trastuzumab.

Oestrogen receptor (ER) and progesterone receptor (PR) statuses are scored on immunohistochemistry (IHC) slides using a standardised scoring system, such as Allred, and are based on the proportion of tumour cells showing positivity and the intensity of the stain. Response to endocrine therapy is related to the degree of expression, and as few as 1% of cells showing positive staining indicates the possibility of a treatment benefit (Campbell *et al.*, 2016). HER2 receptor status is reported semi-quantitatively on

IHC slides, with any borderline tumours assessed further using fluorescent in situ hybridisation (FISH), as overexpression of HER2 protein is driven by amplification of the HER2 gene (Burstein, 2005; Gajria and Chandarlapaty, 2011).

1.3.4 Molecular classification

Whilst traditional histological classification provides some information on prognosis, it provides very limited information on the molecular genetic changes in the tumour. In the last two decades new systems have been developed that consider classification on the molecular and genetic level. Two examples of such classifications are known as *intrinsic subtypes* and *integrative clusters* (Russnes *et al.*, 2017).

Intrinsic subtypes were first described by Sørlie *et al.* (2001). Subtypes are classified initially from gene expression data using 456 cDNA clones, acquired using microarrays. Gene expression data showed that breast cancers are largely divided on the basis of expression of the ER and HER2 pathways, and on proliferation. The subtypes, on the whole, have different immunohistochemical (IHC) staining patterns, different tumour grade, and different prognoses (Table 1) (Callagy *et al.*, 2008; Cheang *et al.*, 2009; Parker *et al.*, 2009). A classifier based on this approach, but using fewer features, PAM50, has been developed by Prosigna[®] (Nielsen *et al.*, 2010) and validated as highly reproducible across multiple clinical testing laboratories on RNA extracted from formalin fixed paraffin embedded (FFPE) tissue (Nielsen *et al.*, 2014).

| Intrinsic subtype | IHC | Grade |
|---------------------|----------------------------------|-------|
| Luminal A | ER/PR +, HER2 -, Ki67 - | 1-2 |
| Luminal B | ER/PR +, HER2 +/-, Ki67 + | 2-3 |
| HER2 overexpression | ER/PR -, HER2 + | 2-3 |
| Basal-like | ER/PR -, HER2 -, Basal markers + | 3 |
| Normal-like | ER/PR +, HER2 -, Ki67 - | 1-3 |

Table 1: Intrinsic subtype IHC features and grade

A more fine-grained classification system, integrative clusters (iClust or iC) was proposed by Curtis *et al.* (2012), who performed an integrated analysis of copy number

and gene expression in 2000 breast tumours, and identified 10 subtypes of breast cancer that show distinct clinical outcomes (Table 2). The classifier incorporates copy number aberrations (CNA) and driver gene mutations, some of which are potential therapeutic targets (for example PIK3CA). The classifier (iC10) derived from this study uses 754 features from gene expression data acquired using microarray, but has so far only been validated on fresh frozen tissue, which is not always available in a clinical setting.

| iC | Frequency | Prognosis | Distinguishing Molecular Features | Histological Features | Intrinsic Subtype | ER/HER2 |
|----|-----------|--------------|---|--|---------------------|----------------------|
| 1 | 7% | Intermediate | 17q23 amplification <i>GATA3</i> mutations High genomic instability | High grade, NST | Luminal B | 85% ER+ 15% HER2+ |
| 2 | 4% | Poor | 11q13-14 amplification (<i>CCND1</i>) High genomic instability | Variable | Luminal A and B | 94% ER+ 8% HER2+ |
| 3 | 15% | Good | Low genomic instability with few copy number changes <i>PIK3CA</i> and <i>CDH1</i> mutations | Low grade, Tubular, Lobular, Mixed NST/special types | Luminal A | 94% ER+ 0% HER2+ |
| 4 | 17% | Good | Low genomic instability Up regulation immune response genes | Non-grade 3, lymphocytic infiltrate | Luminal A (mixed) | 70% ER+ 8% HER2+ |
| 5 | 10% | Poor | <i>ERBB2</i> amplification | Grade 3 | Her2E and Luminal B | 51% ER+ 85% HER2+ |
| 6 | 4% | Intermediate | 8p12 amplification (<i>ZNF703</i>) High genomic instability | Non-low grade | Luminal B | 98% ER+ 4% HER2+ |

| | | | | | | |
|----|-----|--------------|--|-------------------------|-------------------|-----------------------|
| 7 | 10% | Good | 16p gain, 16q loss, 8q amplification <i>MAP3K1</i> mutations | Non-grade 3 | Luminal A | 97% ER+ 0.5% HER2+ |
| 8 | 15% | Good | 1q gain, 16q loss PIK3CA and GATA3 mutations | Low grade | Luminal A | 96% ER+ 1% HER2+ |
| 9 | 7% | Intermediate | 8q gain, 20q amplification High genomic instability TP53 mutations | Grade 3 | Luminal B (mixed) | 87% ER+ 12% HER2+ |
| 10 | 11% | Poor | 5q loss, 8q, 10p and 12p gain Impaired DNA checkpoint regulation, TP53 mutations | Grade 3, Medullary like | Basal-like | 14% ER+ 3% HER2+ |

Table 2: Integrative cluster frequency, prognosis, distinguishing molecular features and receptor status

1.3.5 Staging of breast cancer

As with most cancers, breast cancer can be divided into different stages of disease, taking into account factors related to the tumour itself, lymph node metastasis, and distant metastasis. Staging is a reflection of how long the tumour has been there, how quickly it is growing and its ability to spread. The most commonly used staging system is the TNM staging system, devised originally by the Union for International Cancer Control in 1968 (UICC). In 1977, the American Joint Committee on Cancer (AJCC) published their first manual, and subsequent editions were synchronised and are adopted worldwide. For breast cancer, T denotes the primary tumour stage and is determined by the size of the tumour and the extent of its invasion into surrounding tissue. It ranges from T₁ – T₄, with an additional category of T_{is} denoting in situ disease only. N denotes the nodal category, and is based on the number of the lymph nodes involved, their location, as well as the sizes of the tumour deposits. T and N categories can be assessed by a pathologist if the patient has had surgery and are designated pathological (p) categories, or can be assessed clinically using imaging to give clinical (c) categories. M denotes the absence (M₀) or the presence (M₁) of metastatic breast cancer. The T, N and M categories are combined to give the overall Stage, which is defined as I-IV. I-III indicate early stage breast cancer confined to the breast and axilla, with stage IV indicating the presence of distant metastatic disease.

The staging system reflects the extent of the disease at the time of assessment, and is widely used clinically and by cancer registries. It has been consistently applied in clinical trials and forms the basis of many treatment decisions.

1.3.6 Personalised Approach

More recently, with development in genomic and transcriptomic technologies, finer details within subtypes can be identified. Therapies have toxic side effects, and many women are currently given chemotherapy unnecessarily for small benefits. Conversely, some women still do poorly despite chemotherapy and need alternative agents that target specific molecular abnormalities in their tumour. Personalised medicine involves getting a detailed molecular profile of the tumour and using that to tailor a unique treatment plan for the individual including not only chemotherapy and endocrine therapy but specific targeted agents.

Projects such as the Personalised Breast Cancer Programme (PBCP) in Cambridge, are aiming to develop methods that allow tailoring the treatment to each individual patient with information based on the analysis of their DNA and RNA. Patients enrolled in the programme have whole genome sequencing of their germline and tumour DNA, cataloguing both inherited and acquired mutations for clinical and research purposes. This approach draws from molecular classification systems, and can take into account the heterogeneity of the disease within each subgroup, allowing clinicians to focus on the characteristics of the tumour in each individual. It has the potential to incorporate multi-modal information, such as radiological imaging data and computational pathology data, to truly optimise the treatment plan for an individual.

1.4 Treatment of breast cancer

1.4.1 The development of treatment

The treatment of breast cancer has evolved over the years, as we increase our understanding of the underlying molecular biology of the disease and with the emergence of new treatment options. Treatment has moved from surgery only, to local therapy with surgery and radiotherapy, to a systemic approach that combines local therapy with either hormone therapy, chemotherapy or both. The development of new therapies has been more focused on targeted treatment aimed at specific molecular abnormalities in the cancer but with less impact on normal cells. The issue of overtreatment has received a lot of attention, and the pendulum has swung from giving all patients maximum therapy to more conservative approaches, not only in surgical extent but also systemic therapy. As a result, the accurate stratification of patients is more important than ever.

Surgery was the earliest treatment of breast cancer, and is still the definitive treatment today. Radical resection was the norm in the middle of 19th century, but with the rapid advances in medical radiation and chemotherapy, the surgical approach has started to be modified. However, it wasn't until the late 1980s before evidence emerged supporting limited surgery (Lakhtakia, 2014). With the introduction of screening, more early and small tumours are being detected, for which a wide local excision (WLE), has become common practice (NICE, 2018). Indiscriminate axillary lymph node clearance surgery is much less common since the development of sentinel node biopsy (Krag *et al.*, 2010), which has spared many patients from lymphoedema – an undesirable side effect. Despite all of our advances, there are still some tumours that are not operable due to their size and involvement of surrounding tissue.

Radiotherapy is often given after surgery, in accordance with National Institute of Clinical Excellence (NICE) guidelines (NICE, 2018). In those patients who have had WLE, radiation can eliminate residual microscopic disease that may remain in the breast even when negative resection margins are obtained (no tumour cells reaching the edge of the resected specimen). In both mastectomy and WLE cases, there is

significant reduction in local recurrence and mortality (Early Breast Cancer Trialists' Collaborative *et al.*, 2011; Moo *et al.*, 2018).

Endocrine therapy is used against the tumours that are classified as hormone receptor positive, and is often used in adjunct with other treatments. Oestrogen and progesterone receptors reside intranuclearly and drive proliferation of breast cancer (Daniel *et al.*, 2011; Xue *et al.*, 2019). If a tumour is ER (or sometimes PR) positive on immunohistochemistry, an inhibitor of the ER pathway is offered. These include selective oestrogen receptor modulators such as tamoxifen, selective oestrogen receptor downregulators such as fulvestrant, and aromatase inhibitors such as anastrozole.

Another receptor that is commonly considered alongside hormone receptors is HER-2. HER2 is a membrane bound tyrosine kinase receptor that is overexpressed by tumour cells due to gene amplification (Burstein, 2005). Monoclonal antibodies such as trastuzumab block HER2 receptor dimerisation and downstream signalling (Hudis, 2007), and have transformed HER2 positive breast cancer from having the worst prognosis to good prognosis for those that respond to this treatment.

Chemotherapy most commonly involves a combination of cytotoxic drugs, mostly targeting mitotically active cells, broadly by inducing DNA damage and thus cell death. Bonadonna *et al.* (1976) published one of the first studies that showed that a 12-week long chemotherapy regime including cyclophosphamide, methotrexate and 5-fluorouracil given after surgery could significantly reduce recurrence rate after mastectomy. In the next few years to decades, methotrexate was gradually replaced by anthracyclines, and taxanes were introduced. The choice of chemotherapy agents, dosing, the order of administration of the agents, and the duration of the treatment have been subjects of many clinical trials and are constantly developing with the continual emergence of new compounds.

1.4.2 Neoadjuvant therapy

Historically, chemotherapy was given after surgery for patients at high risk of developing local recurrence or distant metastatic disease; a process known as adjuvant chemotherapy. The rationale is that the chemotherapy will destroy microscopic foci of tumour that may have already spread around the body, preventing the development of distant metastases which is the major cause of cancer related mortality (Hagemeister Jr *et al.*, 1980; Lu *et al.*, 2009). Increasingly, hormone and/or chemotherapy is given prior to surgery in patients with high-risk breast cancers, tumours ≥ 2 cm, or with locally advanced disease (including disease initially thought ineligible for resection). This is known as 'neoadjuvant therapy' (NAT). Rubens *et al.* (1980) looked at tumour regression in locally advanced tumours that were inoperable, and found a good response rate with chemotherapy alone. Many agents that are used in post-operative chemotherapy were then trialled on locally advanced tumours. Fisher *et al.* (1998) published a large study that showed that NAT is as effective as adjuvant chemotherapy, and permits more limited surgery. It is now well known that neoadjuvant therapy (NAT) has the benefit of not only reducing the tumour size prior to surgery, thus reducing the extent of the surgery, but also of making previously inoperable tumours operable.

The standard of care for NAT was established in 2006 by Bear *et al.* (Bear *et al.*, 2006) to be anthracyclines and taxanes. The order of administration has been shown to be important to achieve better outcome by trials such as Neo-tAngo (Bines *et al.*, 2014; Earl *et al.*, 2014).

The tumour's response to NAT is also highly prognostic (Cortazar *et al.*, 2014). Patients that have a pathological complete response (pCR) have a good prognosis across all molecular subtypes (Thompson and Moulder-Thompson, 2012; Cortazar *et al.*, 2014), meaning that the intensity of future treatment can be reduced. pCR is also a surrogate marker of survival, which means that it can serve as the primary endpoint in neoadjuvant clinical trials; these trials are smaller, and generate faster results, allowing more cost-effective and timely approval of new agents. Conversely, tumours that do not respond have a poor prognosis, and recent data from the KATHERINE (Von Minckwitz *et al.*, 2019) and CREATE-X (Masuda *et al.*, 2017; Zujewski and Rubinstein, 2017) trials show that survival outcomes then improve with additional adjuvant therapy. This is

another way of personalising care, based on an individual's response to first line treatment, but methods to predict poor response in advance would reduce treatment burden and may render second line therapies more effective if given earlier.

1.4.3 Response to neoadjuvant therapy

Response to neoadjuvant therapy is assessed by pathologists using the surgical resection specimen. Symmans *et al.* (2007) developed a method for measuring residual cancer burden (RCB). The method takes into account both the residual invasive tumour as well as the lymph nodes and is calculated as:

$$RCB = 1.4(f_{inv}d_{prim})^{0.17} + [4(1 - 0.75^{LN})d_{met}]^{0.17}$$

where f_{inv} is the proportion of invasive carcinoma within the cross-sectional area of the primary tumour bed, corrected for the component of in situ carcinoma; d_{prim} is the geometric mean of bidimensional measurements of the primary tumour bed in millimetres; LN is the number of axillary lymph nodes containing metastatic carcinoma; and d_{met} is the diameter of the largest metastasis in an axillary lymph node.

RCB is a continuous scale starting from 0, where there is no residual disease, which is more commonly known as pCR. For the patients with residual disease ($RCB > 0$), two cut off points were identified at $RCB = 3.28$ and $RCB = 1.36$ for the three classes of RCB (RCB III, RCB II and RCB I). RCB classification is the preferred quantification method because of its reproducibility, and clinical validation with long-term follow-up data (Provenzano *et al.*, 2015).

RCB has been shown to have long term prognostic value for all subtypes of breast cancer (Symmans *et al.*, 2017). A particularly good outcome is observed in cases where patients reached pCR at surgery (Thompson and Moulder-Thompson, 2012; Cortazar *et al.*, 2014), and pCR predicts relapse-free survival (RFS) for some subtypes (Esserman *et al.*, 2012).

1.4.4 Factors determining the response to neoadjuvant therapy

It is clear that NAT is of benefit to many patients. However, there are a proportion of patients, whose disease would have been predicted to respond to the treatment, but who did not benefit from NAT. Their tumour may have remained the same or, in some cases, grown after the commencement of NAT. It is vital to be able to identify this subgroup of patients, so that they can receive alternative treatment promptly.

The reasons behind varied response to NAT have not yet been fully understood. Many studies have examined how different subgroups of breast cancer respond to NAT, and looked for new markers that would help clinicians to accurately predict response. An accurate prediction can allow surgeons to know which patients might benefit from early radical surgery (mastectomy), rather than attempting to “downstage” the tumour and opt for breast conservation (wide local excision).

Histological types and grades are shown by various studies to be predictive. A meta-analysis by Loibl *et al.* (2014) showed that patients with invasive lobular cancer have significantly lower rates of pCR. Fisher *et al.* (2002) noted that higher histological grade is predictive of pCR, at least in part because more rapidly dividing nuclei are more vulnerable to cytotoxic medications (Kim *et al.*, 2014).

The molecular profiles of tumours are being intensively investigated. Various studies have shown that the tumours with high expression of ER had much lower pCR rates compared to those with low or no expression of ER (Colleoni *et al.*, 2009; Liedtke *et al.*, 2009; Osako *et al.*, 2012). By classifying tumours into intrinsic subtypes, better response to NAT has been noted in multiple trials in tumours that are *basal-like* or HER2 positive (Colleoni *et al.*, 2004; Rouzier *et al.*, 2005; Asselain *et al.*, 2018). Classifying using integrative clusters, Ali *et al.* (2014) demonstrated differential pCR rates between clusters, and showed that integrative clusters are as accurate at predicting pCR as PAM50.

As well as factors intrinsic to the tumour cells, the tumour microenvironment (Montagna *et al.*, 2015) also plays a key role in influencing response to NAT. Zitvogel *et al.* (2008; 2011) showed that immune response might mediate the chemotherapy effect, and therefore may be an important determinant of treatment response. Karagiannis *et*

al. (2017) proposed that the interaction between tumour cells and its adjacent environment during NAT may lead to metastasis. Ali et al. (2016; 2017) demonstrated and validated that lymphocyte density is a predictor of response to chemotherapy. Epidemiological factors are also found to be predictive of pCR, perhaps also through tumour-environment interactions. Huober *et al.* (2010) noted that young age, and Fontanella *et al.* (2015) noted that a lower body mass index, are both predictive of pCR. Together, these studies suggest that we should not focus on tumour cells alone, but combine this with a consideration of their interaction with their local microenvironment and host.

1.5 The importance of formalin fixed paraffin embedded (FFPE) tissue

During the diagnosis and treatment of breast cancer, or indeed with most cancers, a tissue sample is taken from the patient for assessment by pathologists. For the majority of the patients, the samples are fixed in 10% formalin for between 4 and 24 hours. Formalin is a preservative that crosslinks proteins and preserves tissue morphology for histological assessment. However, it also causes crosslinks between nucleic acids and breaks the phosphodiester backbone, which leads to fragmentation, resulting in smaller fragments of DNA and RNA (Evers *et al.*, 2011; Jones *et al.*, 2019) that create difficulties for some genetic sequencing and analysis methods (Marczyk *et al.*, 2019).

For the minority of patients that have enrolled in a study or a trial, an additional small sample may also be taken and frozen in liquid nitrogen for subsequent analysis. The larger samples are then cut up and appropriate sections selected by a pathologist before further processing, while the smaller samples are processed directly. The processing steps include sequential dehydration from an aqueous environment to an alcohol environment (most often ethanol), subsequent replacement by xylene (or xylene substitute) in a process referred to as clearing, and replacement of the xylene with paraffin (impregnation). This produces a paraffin block that holds tissue in for further sectioning and staining, known as formalin fixed paraffin embedded (FFPE) tissue.

FFPE tissue is widely used worldwide as part of standard clinical practise. Once embedded in paraffin, the tissue can be stored easily, and RNA, DNA and protein can be extracted and analysed for at least 12 years (Kokkat *et al.*, 2013). In the UK, the clinical slides and FFPE blocks are archived when no longer required by clinicians for at least decades. Where consent is prospectively obtained or retrospectively granted, the archived tissue could be revisited with the development of new technology or breakthrough in understanding of the disease. Some rarer or smaller tumours are only available in FFPE blocks due to clinical needs, for example some small screen-detected early breast cancers are so tiny that it is not possible to process an additional frozen sample for biobanking or research. Being able to utilise FFPE tissue will allow these

tumours to be studied and provides an opportunity, in the setting of breast cancer, to study the early days of tumour development (Gaffney *et al.*, 2018).

FFPE tissue is used much less frequently than fresh frozen tissue in molecular studies, as the process of producing FFPE tissue may alter molecule integrity and structure. Chromosomes and DNA are relatively robust, but RNAs are degraded and broken down into smaller fragments. Even more problematically, there may be differential degradation of different mRNA molecules (Srinivasan *et al.*, 2002; Greytak *et al.*, 2015). Proteins appear much more difficult to extract (Gaffney *et al.*, 2018). Since 2005 there has been a dramatic increase in breast cancer research published using FFPE tissue (Figure 1), which reflects the development in technology aimed at overcoming these difficulties, and the importance of translating research work performed on frozen tissue to more abundant and practical tissue.

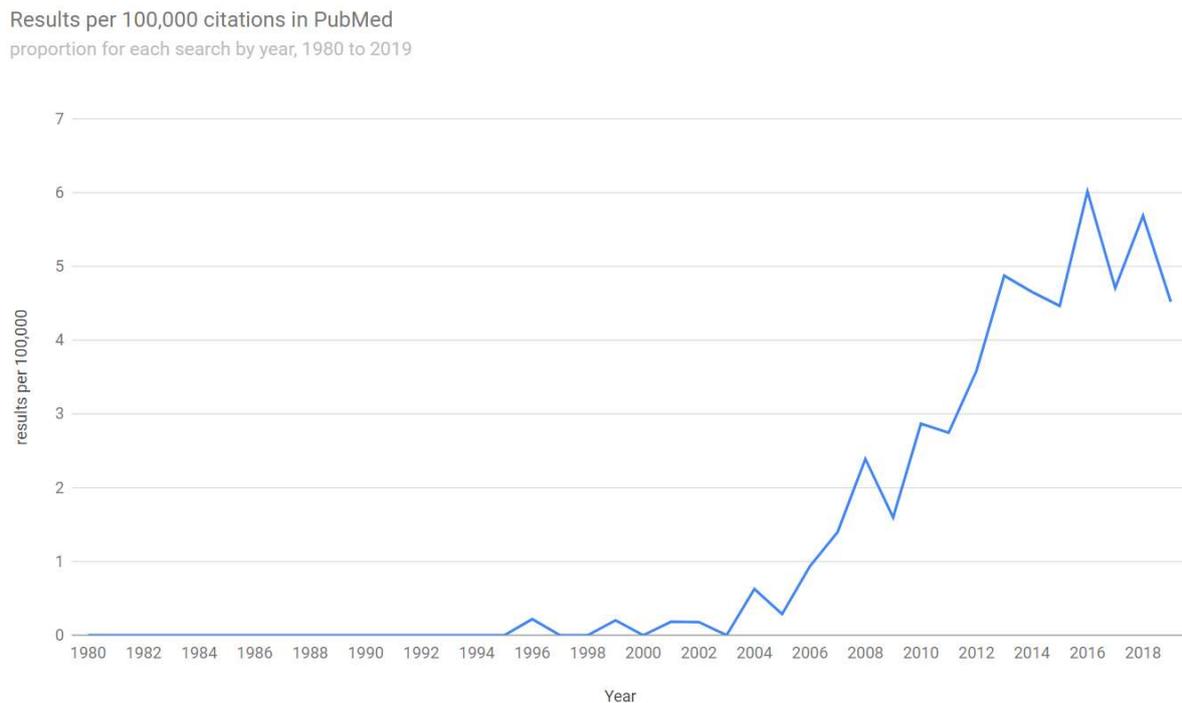


Figure 1: Proportion of annually published papers mentioning breast cancer and FFPE tissue in their Pubmed citation, 1980-2019

1.6 Digital pathology image analysis

Histopathology is an integral part of the diagnosis and management of breast cancer. It plays a vital role in directing treatment and providing prognostic information (Elston and Ellis, 1991). Many features, such as tumour grade and lymphovascular space invasion, are susceptible to inter-observer variation. However, modern technologies such as whole slide imaging and digital image analysis potentially allow many factors, such as lymphocyte density, to be rapidly and consistently quantified across the large cohorts recruited to clinical trials.

In combination with rapidly advancing computing technology, digital pathology is increasingly being utilised in research and routine diagnostic practice. Using standard Haematoxylin and Eosin (H&E) staining alone, digital pathology combined with computerised image analysis can generate a large quantity of data based on colour, contour and contrast, which are then used to automatically identify cellular and stromal components of the images. A training set is generated by histopathologists, in which clinically relevant components of the images are manually identified. Various machine learning methods are then used to identify features which are of clinical significance (Heindl *et al.*, 2015). Digital pathology and machine learning algorithms can also be developed and adapted to score immunohistochemical stains, such as for oestrogen and HER2 receptors (Ali *et al.*, 2013).

Yu *et al.* (2016) examined images of non-small cell lung cancer slides and used a program called CellProfiler (Carpenter *et al.*, 2006) to identify cell types, before extracting features including cell size, shape, and distribution of pixel intensity in the cells and nuclei, as well as texture of the cells and nuclei. These features were then tested for their ability to classify benign and malignant tissue using machine learning methods. Further, they correlated clinically relevant features with survival, and found that features such as the texture of the nuclei are meaningful predictors of outcome. These features are difficult for histopathologists to quantify without computerised methods.

Beck *et al.* (2011) have used tissue microarrays (TMA) of breast cancer to predict survival in their cohorts. They first separated images into epithelial regions and stromal regions, and then constructed higher-level contextual and relational features both within and

between these regions. Such features include relationships between morphologically regular and irregular nuclei, relationships of contiguous epithelial regions with underlying nuclear objects, characteristics of stromal nuclei and stromal matrix and relationships between epithelial and stromal objects. Using bootstrap analysis, relational features of stromal regions, such as the contiguousness of stroma and presence of inflammation and neighbouring nuclei morphology were found to be strong predictors of survival.

Dong et al. (2014) used whole slide images of biopsies of patients with intraductal proliferations. They used nuclear shapes and colour hues to perform segmentation and then computed nuclear features. By using logistic regression with Lasso regularisation, they identified nuclear features that can distinguish between ductal carcinoma in situ (DCIS) and usual hyperplasia.

Dekker et al. (2015) focused on stroma in biopsies before neoadjuvant therapy. They looked at H&E stain, AZAN trichrome stain and immunohistochemistry (IHC). They measured stromal fibre angles to calculate stromal orientation, which is then used to assess stromal organisation. By using IHC to assess for the presence of phosphorylated Smad2 (pS2) as a marker for TGF- β pathway activity, they suggested that stromal compartments of tumours with active signalling in the TGF- β pathway are likely to have more organised stroma. This results from altered synthesis of the collagen matrix, hypothetically influencing both tumour cell mobility and vulnerability to chemotherapy. They then used logistic regression to successfully predict pathological response to NAT.

Natrajan et al. (2016) used whole slide images of frozen sections from surgery with H&E stain. By measuring parameters that would define nuclear morphology, tumour cells and stromal cells were identified, and their spatial variability was calculated. This was then used to generate measures of environmental diversity and heterogeneity. By using unsupervised Gaussian mixture clustering, they identified subgroups with distinct outcomes and with genomic correlation.

In our group, in collaboration with Institute of Astronomy, Ali et al. (2016; 2017) have described and validated a method by which lymphocyte density can be assessed as a predictive factor for response to chemotherapy. The significance of tumour infiltrating lymphocytes has long been debated, and the measurement of this has been subjective

and qualitative in traditional clinical practice. With computational pathology, more standardised and objective measurements can be made to allow for more stringent statistical analysis. Using a computer algorithm, Ali *et al.* first identified the area of tissue that needs to be analysed and excluded the “white space” in the digital images. Cell nuclei were then identified and classified into lymphocytes, tumours or stromal cells based on a training set, using support vector machine methods. Based on these, descriptive cellular metrics were calculated. Then, lymphocyte density was calculated by working out the average distance (r) to the nearest 50 lymphocytes using K-nearest-neighbour methods. For each lymphocyte the density was then quantified, and for each region the median density of all the lymphocytes detected was calculated. Importantly, they demonstrated that image metrics reflected molecular subtypes of the primary tumour. They showed that lymphocyte density at baseline was an independent predictor of pCR, however an increase in lymphocyte density following exposure to chemotherapy was associated with poorer response.

Similarly, Yuan *et al.* (2012) investigated the correlation between image analysis and genomic data. They used whole slide images with H&E stain to quantify lymphocytes proportions and stromal spatial patterns. They showed that the proportion of lymphocytes in a tumour is complementary to gene expression signatures for lymphocyte infiltration, and that integrating both types of evidence improved prediction of patient survival. Stromal spatial patterns were shown to be an independent prognosticator. This study showed how digital pathology and genomic studies can complement each other by combining the genomic information from the tumour and the spatial information of tumour microenvironment.

These effects are not restricted to breast cancer. Gong *et al.* (2019) examined the spatial relationships between tumour cells and CD8+ T cells in colorectal cancer. They showed that patients with more high-density T cell clusters, which were found in both circular and elongated shapes in the regions immediately adjacent to the tumour mass, responded better to PD-1 blockade treatment designed to reverse immune tolerance to tumour cells. This highlighted the importance of not just the presence of immune cells on a whole slide, but also the spatial proximity of the immune population to the tumour cells.

1.7 Aims and Objectives

The unifying aim of this thesis is to improve our ability to accurately predict a patient's response to neoadjuvant chemotherapy using 'real-world' techniques that will be available in the majority of hospitals. I approach this task in two complementary ways, developing methods based on radically different modalities of data that are likely to be independently predictive of response. Specifically, I will be focusing on material that all hospitals will have – FFPE of tumour tissue and the H&E slides that have been made from them at the time of diagnosis.

First, I assess the feasibility and reproducibility of classifying tumours into integrative clusters (ICs) using RNA extracted from FFPE tissue. I will do this in cores of breast tumour taken from the METABRIC study, the original cohort on which ICs were developed. These samples have the advantage that their IC classification is already known and is deemed the "gold standard". I will develop a classifier using the NanoString, the platform that Prosigna[®] (PAM50) is based on and that has been approved by the FDA, CE marked, and NICE-endorsed in a subset of patients with breast cancer. The advantage of this platform is that it is fast and easy to operate, compared to other techniques for expression analysis that would be less practical to roll out into standard clinical environments.

Second, I will develop methods for computational pathology that allow the automated assessment of spatial features that are hard to quantify manually. I will do this in images from neoadjuvant trials and studies conducted locally at Cambridge University Hospital. I will build a well annotated image database to complement current machine learning image analysis algorithms, and also allow for the future development of neural network algorithms. As part of this, I will collaborate with multi-disciplinary image analysis and machine learning experts to improve existing cell classification algorithms as well as validate existing algorithms. I will perform clustering analyses, in order to identify tumour-stromal interfaces automatically, and thus quantitatively measure the spatial relationships between immune infiltrates and tumour cells.

2. Materials and Methods

2.1 Preface

This chapter sets out the general methods I employed in the experimental work I report in chapters 3 (NanoString, section 3), 4 (RNA sequencing, sections 2.3 and 2.4), 5 (Digital pathology, section 2.6) and 6 (Tumour microenvironment, section 2.7). I undertook all of the procedures myself, except where I have specifically stated otherwise in the text.

I would like to thank Professor Susan Ramus¹ for suggesting modifications that improved the efficiency of RNA extraction using Qiagen miRNeasy FFPE kits and allowing the extraction of DNA from the remaining tissue pellet post RNA extractions.

I am particularly grateful to Dr Suet-Feung Chin, who collaborated with me in validating the methods used to perform RNA sequencing using RNA from FFPE tissue, and to Raquel Manzano Garcia, Dr Oscar Rueda and Dr Stephen John Sammut for sharing their code and pipelines for RNA sequencing analysis.

I would also like to thank Ms Helen Bardwell from the Histopathology Core Facility at the CRUK Cambridge Institute, who meticulously matched the digitised histopathology images to the physical samples.

I collaborated with Dr A Dariush, Department of Astronomy, University of Cambridge, to refine the image processing methods in section 2.6.

¹ School of Women's and Children's Health, Faculty of Medicine, University of NSW Sydney, Sydney, Australia

2.2 NanoString

2.2.1 RNA extraction

FFPE sections stained with H&E were reviewed by pathologists to identify regions with high tumour content, and from which 1.5mm cores were taken. RNA was extracted from one core per patient, using Qiagen miRNeasy FFPE kits (Qiagen, Germany) according to the manufacturer's instructions, with modifications as detailed below.

To dewax, the FFPE cores were vortexed in 150ul PKD buffer and heated for 3 min at 55 °C. The cores were vortexed again before centrifugation at 20,000g for 2 min. For digestion, the tissue was incubated with 10 µl of proteinase K at 56°C for 60 minutes whilst shaking at 800 rpm on an Eppendorf® Thermomixer Compact, followed by shaking at 500 rpm at 80 °C for 15 min. The mixture was incubated on ice for 3 min before centrifugation for 15 min at 20,000 × g. To remove DNA contaminant, the supernatant was transferred to a new tube and incubated with DNase at room temperature for 15 min (Table 3). The remaining pellet was stored at -80°C for future DNA extractions.

| DNase digestion | |
|------------------------|----------------------|
| ~ 160 µl | Supernatant |
| 16 µl | DNase booster buffer |
| 10 µl | DNase I |

Table 3: DNase digestion components

Post DNase digestion, the supernatant was mixed thoroughly with 320 µl of RBC buffer and 1120 µl of 100% ethanol. 700 µl of the mixture, including any precipitate, was transferred to a RNeasy MinElute spin column placed in a 2 ml collection tube, and

centrifuged at $1000 \times g$ for 15 seconds. The flow-through was discarded. This was repeated until all the sample lysate has passed through the column.

The RNeasy MinElute spin column was washed twice with 500 μ l of RPE buffer and centrifuged at $10,000 \times g$ for 15 s and 2 min respectively. The column was dried by a further centrifugation at $20,000 \times g$ for 5 min. The RNA was eluted from the spin column with 30 μ l of RNase-free water added to the spin column membrane and centrifuged for 1 min at $20,000 \times g$ RNA before being stored at -80°C .

2.2.2 Nanodrop

All RNA samples were quantified using a Nanodrop spectrophotometer (ThermoFisher, USA), according to manufacturer's instructions. Blank measurement using nuclease-free water was made before each usage, and 2 μ l of sample was used each time.

The absorbances at 230nm, 260nm and 280nm were recorded, and the ratio of absorbances of 260nm and 280nm (260/280) was used in combination with the ratio of those of 260nm and 230nm (260/230) to assess purity (Patterson and Dackerman, 1952). According to the manufacturer's instruction, "pure" RNA would have a 260/280 of ~ 2.0 and 260/230 between 2.0-2.2.

The concentration measured using Nanodrop was used for subsequent experiments and analyses.

2.2.3 RNA quality assessment

The quality of RNA extracted was assessed using either Bioanalyser or Tapestation (both Agilent, USA), according to the manufacturer's instructions.

Bioanalyser

Samples were analysed using an RNA 6000 Nano kit on the Bioanalyser 2100. 65µl of filtered gel matrix was mixed with 1 µl of dye concentrate and centrifuged at 13000 × g for 10 min at room temperature. 9 µl of the gel-dye mix was loaded to the marked wells with the plunger set at the RNA position. The samples and ladder were denatured at 72 °C for 2 min. Marker, ladder and samples were then loaded into the marked wells, and the chip was vortexed for 1 min at 2400 rpm. The chip was then loaded onto the machine and processed using Eukaryote total RNA Nano default settings.

Tapestation

Samples were analysed using RNA ScreenTape on the Tapestation 4200. 1 µl of sample or ladder was mixed with 5 µl of sample buffer in each well or tube, and vortexed at 2000 rpm for 1 min. The sample and ladder mixtures were then denatured at 72 °C for 3 min, followed by incubation on ice for 2 min, before being briefly centrifuged and loaded onto the machine with the tape and run using using TS4200 Controller Software for Eukaryotic RNA.

2.2.4 RNA hybridisation for NanoString

RNA was hybridised with custom nCounter XT CodeSet Gene Expression Assays (NanoString, US), according to manufacturer's instructions. Hybridisation buffer was added to the tube containing the Reporter CodeSet, with 300ng RNA added to each reaction, before the Capture CodeSet was added. The mixture (Table 4) was then incubated at 65°C for 18 hours, followed by storing at 4°C.

RNA Hybridisation

| | |
|-----------|----------------------|
| 5 μ l | Hybridisation buffer |
| 3 μ l | Reporter CodeSet |
| 2 μ l | Capture CodeSet |
| 5 μ l | 300 ng of RNA |

Table 4: RNA Hybridisation components

2.2.5 NanoString Platform

Samples were kept away from light and at 4°C until being loaded onto the Prep Station, according to manufacturer's instruction. Briefly, two reagent plates were centrifuged at $2000 \times g$ for 2 minutes before loading. The cartridge for sample alignment and barcode reading was warmed from -20°C in storage to room temperature before use in order to avoid condensation. Each run could include a maximum of 12 samples. The Prep Station was run using the High Sensitivity protocol to maximise binding of molecules to the cartridge.

Once completed on Prep Station, up to six cartridges containing samples were transferred to the Digital Analyser. A *.rlf* file containing the gene names and their corresponding barcode was uploaded together with a *.cdf* file containing sample identifiers. The field of view count (*FOVCount*), which is the number of images to analyse per assay, was set to the maximum available value of 555 to maximise data collection.

2.3 RNA sequencing

The first stage of RNA sequencing is making the library (sections 2.3.1 to 2.3.8). This involves reverse transcribing messenger RNA (mRNA) into complementary DNA (cDNA), followed by the addition of adaptors. The library is then enriched by polymerase chain reaction (PCR) before sequencing.

The library was made using the NEBNext® rRNA Depletion Kit (Human/Mouse/Rat) and NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina®, both from New England BioLabs (USA). For RNA extracted from fresh frozen tissue, 100 ng of RNA was used, whereas for the RNA from FFPE samples 300 ng was used. This is to take into account the fragmented nature of RNA from FFPE that produces a lower yield of cDNA.

2.3.1 Ribosomal RNA depletion

Ribosomal RNA (rRNA) was removed to allow for the accurate analysis of messenger RNA (mRNA). The rRNA was bound by single stranded DNA (ssDNA) probes and then broken down by RNase H. The residual ssDNA probes were then removed by DNase I, leaving mRNA in the sample.

Single strand DNA Probe hybridisation to rRNA

The sample was prepared according to manufacturer's protocol. Briefly, RNA was diluted to 12 μ l final volume with nuclease-free water and kept on ice, which was then mixed with hybridisation buffer and rRNA depletion solution (Table 5), and hybridised in a thermal cycler (Table 6).

Probe Hybridisation

| | |
|------------|---------------------------------|
| 1 μ l | NEBNext rRNA Depletion Solution |
| 2 μ l | Probe hybridisation buffer |
| 12 μ l | RNA |

Table 5: Probe hybridisation components

Thermal Cycler Setting for Probe Hybridisation

| | |
|-------|------------------------|
| 105°C | lid |
| 95°C | 2 min |
| 22°C | Ramp down at 0.1°C/sec |
| 22°C | 5 min |

Table 6: Temperature settings for probe hybridisation

RNase H digestion

The sample was then mixed with RNase H master mix (Table 7) and incubated for 30 min at 37°C on a thermal cycler with lid temperature set to 40°C.

RNase H Master Mix (prepared on ice)

| | |
|-----------|---------------------------------|
| 2 μ l | NEBNext RNase H |
| 2 μ l | NEBNext RNase H Reaction Buffer |
| 1 μ l | Nuclease-free water |

Table 7: RNase H digestion components

DNase I Digestion

The DNase I master mix (Table 8) was immediately added to the sample and incubated for 30 min at 37°C on a thermal cycler with lid temperature set to 40°C.

DNase I Master Mix (prepared on ice)

| | |
|--------------|-------------------------|
| 5 μ l | DNase I Reaction Buffer |
| 2.5 μ l | DNase I (RNase-free) |
| 22.5 μ l | Nuclease-free water |

Table 8: DNase I digestion components

Clean up

Because of availability, Agencourt AMPure XP beads were used for the cleanup process instead of the default Agencourt RNAClean XP (both made by Beckman Coulter™, USA), and this substitution is allowed as stated on the New England BioLabs (NEB) website.

The sample was mixed with 110 μl (2.2 \times) beads and incubated for 15 min on ice. The mixture was then placed on a magnetic rack to isolate the beads. The beads, while on the magnetic rack, were washed twice with 200 μl of freshly prepared 80% ethanol for 30 seconds. The beads were then air dried for 2 ~ 5 min and RNA was eluted with 7 μl of nuclease-free water. The tubes were removed from the magnetic rack and incubated for 2 min at room temperature before being placed back on the magnetic rack until the solution was clear. 5 μl of the supernatant containing the purified RNA was removed and transferred to a nuclease-free tube.

2.3.2 RNA fragmentation and priming

The rRNA depleted RNA were fragmented to similar lengths, by mixing with random primers and reaction buffer (Table 9), and incubated for 7 minutes for the RNA with an RNA integrity number (RIN) less than 7, and 15 min for those with RIN greater than 7, at 94°C on a thermal cycler.

Fragmentation and Priming reaction (prepared on ice)

| | |
|-----------------|--|
| 5 μl | RNA sample |
| 4 μl | NEBNext First Strand Synthesis Reaction Buffer |
| 1 μl | Random Primers |

Table 9: RNA fragmentation and priming reaction components

2.3.3 First Strand cDNA Synthesis

The process of first strand cDNA synthesis involves reverse transcribing the mRNA. To do this the first strand synthesis enzyme mix was added to the sample and mixed thoroughly (Table 10), then incubated in a pre-heated thermal cycler (Table 11), according to manufacturer's instructions.

First strand synthesis reaction (prepared on ice)

| | |
|------------|---|
| 10 μ l | RNA sample, fragmented and primed |
| 2 μ l | NEBNext First Strand Synthesis Enzyme Mix |
| 8 μ l | Nuclease-free water |

Table 10: First strand synthesis reaction components

Thermal Cycler Setting for First Strand cDNA Synthesis

| | |
|------|--------|
| 80°C | Lid |
| 25°C | 10 min |
| 42°C | 15 min |
| 70°C | 15 min |
| 4°C | Hold |

Table 11: Temperature settings for first strand cDNA synthesis

2.3.4 Second Strand cDNA Synthesis

The second strand of cDNA is the complementary strand for the first strand of cDNA, making double stranded cDNA. The second strand synthesis was prepared (Table 12) and incubated according to the manufacturer's instructions (Table 13).

Second strand synthesis reaction (prepared on ice)

| | |
|------------|--|
| 20 μ l | Sample after first strand synthesis |
| 8 μ l | NEBNext Second Strand Synthesis Reaction Buffer (10 \times) |
| 4 μ l | NEBNext Second Strand Synthesis Enzyme Mix |
| 48 μ l | Nuclease-free water |

Table 12: Second strand synthesis reaction components

Thermal cycler setting for second strand cDNA synthesis

| | |
|------|------|
| 40°C | Lid |
| 16°C | 1 hr |

Table 13: Temperature settings for second strand cDNA synthesis

2.3.5 Purification of double-stranded cDNA

The resultant sample was mixed with 144 μl (1.8 \times) of resuspended AMPure beads and incubated for 5 min at room temperature. The mixture was then placed on a magnet to separate the beads from the supernatant. The beads were then washed twice with fresh 80% ethanol while remaining on the magnetic rack at room temperature for 30 sec each time. The beads were air dried and DNA eluted from the beads by adding 53 μl of 0.1 \times TE buffer and incubated for 2 min at room temperature off the magnetic rack. The tubes were placed back on the magnetic rack for approximately 2 mins, and 50 μl of the cleared supernatant containing the DNA was transferred to a new tube.

2.3.6 End prep of cDNA library

This step is to remove the overhangs of the cDNAs and ensure they are blunt-ended for subsequent reactions. This reaction was prepared (Table 14), mixed thoroughly, and then incubated (Table 15) according to the enzyme mix manufacturer's instructions.

End prep reaction (prepared on ice)

| | |
|------------------|---|
| 50 μl | Sample after second strand synthesis |
| 7 μl | NEBNext Ultra II End Prep Reaction Buffer |
| 3 μl | NEBNext Ultra II End Prep Enzyme Mix |

Table 14: End preparation reaction components

Thermal cycler setting for end prep reaction

| | |
|------|--------|
| 75°C | Lid |
| 20°C | 30 min |
| 65°C | 30 min |
| 4°C | Hold |

Table 15: Temperature settings for the end preparation reaction

2.3.7 Adaptor ligation

Adaptors are needed for subsequent PCR reactions. The NEBNext adaptor was diluted 5 fold according to the manufacturer's instructions. The reaction was then assembled and the mixture (Table 16) then incubated for 15 min at 20°C in a thermal cycler. 3 µl of USER (Uracil-Specific Excision Reagent) enzyme was then added to the mixture, thoroughly mixed and incubated at 35°C for 15 minutes with a heated lid at 45°C.

| Ligation reaction (prepared on ice) | |
|--|--------------------------------------|
| 60 µl | Sample after end prep reaction |
| 2.5 µl | Diluted adaptor |
| 1 µl | NEBNext Ligation Enhancer |
| 30 µl | NEBNext Ultra II Ligation Master Mix |

Table 16: Adaptor ligation reaction components

2.3.8 Purification of the ligation reaction

The sample was mixed with 87 µl Agencourt Ampure beads (0.9x), incubated for 5 min at room temperature, and placed on a magnetic rack for 5 min before discarding the supernatant. The beads were then washed twice with 200 µl of freshly prepared 80% ethanol by incubating at room temperature for 30 sec then discarding the supernatant. The beads were then air dried for 5 min before removal from the magnetic rack, and DNA was eluted with 17 µl of 0.1x TE buffer by incubating at room temperature for 2 min and placed back on the rack. 15 µl of the supernatant was removed for the next step.

2.3.9 PCR enrichment

The library was then enriched by PCR. Each sample also had a unique barcode added to allow multiplexing during sequencing. NEBNext Multiplex Oligos for Illumina (96 Index Primers) were used, and the PCR reaction mixture (Table 17) was prepared and incubated (Table 18) according to the manufacturer's instruction. The number of cycles chosen was based on the total RNA input, according to manufacturer's recommendation.

PCR reaction mixture

| | |
|------------|--------------------------------|
| 15 μ l | Sample after adaptor ligation |
| 25 μ l | NEBNext Ultra II Q5 Master Mix |
| 10 μ l | Primer |

Table 17: Polymerase chain reaction components

Thermal Cycler Setting for PCR

| | | |
|-------|--------|-------------|
| 105°C | Lid | |
| 98°C | 30 sec | Once |
| 98°C | 10 sec | } 11 cycles |
| 65°C | 75 sec | |
| 65°C | 5 min | Once |
| 4°C | Hold | |

Table 18: Temperature setting for polymerase chain reaction

2.3.10 Purification of PCR reaction

Agencourt Ampure beads were used for the purification process. 45 μl of beads (0.9 \times) was added to the sample, and incubated for 5 min at room temperature. The sample was then placed on the magnetic rack and the supernatant carefully removed. The beads were then washed with 200 μl of 80% fresh ethanol twice by incubating at room temperature for 30 sec. The beads were then dried and the DNA eluted using 23 μl of 0.1 \times TE buffer, then incubated at room temperature for 2 min. The sample was then placed on the magnetic rack and the supernatant including the DNA was removed and stored at -20°C .

2.3.11 DNA quality assessment and fragment size determination

DNA quality assessment was used to assess the library quality and fragment size using D1000 screentapes. Sample buffer was mixed with sample or ladder (Table 19) and vortexed at 2000 rpm for 1 min, before being centrifuged and analysed on the machine. The sample was diluted to ensure an equal volume of water before being used in the high sensitivity reactions.

| Tapestation Reaction – High Sensitivity | |
|--|------------------|
| 2 μl | Sample buffer |
| 2 μl | Sample or ladder |

Table 19: Tapestation reaction components

2.3.12 Quantitative PCR (qPCR)

qPCR is used to ascertain the concentration of each library by quantifying the amount of sequenceable fragments with adapters. qPCR was performed using QuantStudio (Applied Biosystems, Thermo Fisher Scientific, USA) according to the manufacturer's instructions.

Serial dilution of the sample was performed to achieve dilution factors of 1/5, 1/100, 1/1000, 1/10000 and 1/100000. The 1/100000 was used for subsequent quantification, mixed with primer and master mix (Table 20) according to the manufacturer's instructions. 6 DNA standards and 2 samples of water were used as controls.

qPCR reaction mixture

| | |
|-------------|---|
| 1.2 μ l | Illumina Primer Premix (10 \times) |
| 5.8 μ l | KAPA SYBR [®] FAST qPCR Master Mix (2 \times) |
| 3 μ l | Sample (1/100000 dilution) |

Table 20: Quantitative polymerase chain reaction components

2.3.13 Preparation for sequencing

Each sample was then normalised to 10 nM. 96 samples were divided into 2 pooled samples, one containing those samples of lower concentration of the library and the other containing those samples of higher concentration, in order to further reduce any effect of concentration. 2 μ l of each sample was taken and well mixed. 20 μ l of the pooled sample was submitted to be sequenced by the Genomic Core using HiSeq 4000, sequencing with single read of length 50, and PhiX at 1% (Manley *et al.*, 2016).

2.4 Analysis of RNA sequencing data

2.4.1 Quality control

FastQC is an automated quality control tool for high throughput sequencing data (Andrews, 2010) that combines a number of tests into a merged FASTQ file (Cock *et al.*, 2010). Our outputs of interest were sequence quality, GC content (which is a surrogate marker of contamination or subset bias), sequence length distribution, sequence duplication levels and overrepresented sequences. Any samples automatically flagged as low quality were manually checked and corrected or excluded from analysis.

2.4.2 Alignment

Alignment describes the process by which DNA fragments are arranged according to sequence similarity against a reference genome or transcriptome, for future quantification of expression levels of individual genes. Our samples were aligned using *Salmon* version 0.13.1 (Patro *et al.*, 2017) and *STAR* version 2.7.3a (Dobin *et al.*, 2013). Indexes were generated from the GRCh38 decoy assembly of the human genome, and a transcriptomic Gene Transfer Format (GTF) guide obtained from *Ensembl* Release 87 (Yates *et al.*, 2016).

2.4.3 Pre-processing

Using the binary alignment maps (BAM files) generated by *STAR*, duplicates were marked using Picard MarkDuplicates, and split according to known intron position by SplitNCigarReads (Van der Auwera *et al.*, 2013). The BaseRecalibrator and ApplyBQSR

functions were then used to calibrate base quality. The resulting file was used for subsequent analysis.

2.4.4 Haplotype calling

Haplotype calling allows the matching of samples from the same individual across tissue types (here, FFPE and fresh frozen), despite the expected single nucleotide variation arising from fixation technique (Parker *et al.*, 2019). This usually provides a safety check that the samples have been correctly labelled, although here also provided an additional quality control step for our gene expression data. Haplotype was determined using GATK Haplotypecaller (Poplin *et al.*, 2017).

2.5 Slide Digitisation

Slides were collected from the Tissue Bank at Addenbrooke's Hospital, Cambridge, which is licensed by the Human Tissue Authority (HTA). Slides were cleaned and anonymised prior to scanning.

Slides were scanned using an Aperio AT2 Digital Pathology Scanner, stored in .svs format, managed and annotated using Aperio eSlide Manager system, both by Leica Biosystems.

Scanned images were stored on a local server at Cancer Research UK Cambridge Institute, and transferred to Institute of Astronomy to be analysed using the pipeline described in section 2.6 on a purpose built cluster. The features extracted from each slide were then used in subsequent analysis.

2.5.1 Slide provenance

Patients receiving NAT in the Trans-NEO study in Addenbrooke's Hospital, Cambridge, are recruited to undertake comparative studies of breast cancer before, during and after chemotherapy, allowing for the longitudinal study of the changes within tumour cells (Montagna *et al.*, 2015), as well as the complex interaction between cancer cells and tumour microenvironment. There is ongoing genomic and radiological research on this cohort to study the responses of the tumour. In addition, Addenbrooke's Hospital is also involved in other large multi-centre neoadjuvant trials, such as Neo-tANGo, ARTEMIS and PARTNER. All of these sources were utilised to obtain the images used in this thesis.

2.5.2 Slide Annotation

I manually reviewed each scanned slide and annotated with the following information:

- The trial ID
- The block number that the slide was made from
- Stain used
- Tissue of origin
- Presence of tumour and tumour type
- Presence of *in-situ* tumour and tumour type
- Presence of lymphovascular space invasion

For lymph node sections, I also recorded:

- Presence of tumour metastasis
- Presence of fibrosis in response to treatment
- Total number of lymph nodes taken for the patient

2.5.3 Management of Information

Confidential data was communicated using secure hospital networks, and was anonymised prior to transfer to external networks.

2.6 Cell segmentation and classification

The scanned images were in .svs format, and comprised four independent layers of Lo, L1, L2 and L3. Layer Lo is the image of the highest resolution, and each subsequent layer is scaled down by constant factors along each axis. The segmentation and classification algorithm performed the following steps, implemented in Python 3 with the scikit-image (Van der Walt *et al.*, 2014), scikit-learn (Pedregosa *et al.*, 2011) and OpenCV (Howse, 2013) packages.

1. Tissue selection

Images were extracted at L3 (lowest resolution) level, transformed into grey-scale and inverted, such that stained areas were bright and unstained areas dark. They were then divided into identically sized tiles with slight overlaps to ensure no cells fell entirely on tile boundaries, and as a check to allow alignment if the coordinate system were ever disrupted. The size of the tile, when transformed to the high resolution (Lo) image, was constrained to fall within a range of $1 - 4 \times 10^6$ pixels squared. This sub-division was necessary because the computers available, even with the GPU based high performance cluster facility used for this study, had insufficient memory to analyse the whole slide at high resolution. Histograms of pixel intensity for each tile were thresholded to exclude those that do not contain tissue from future analysis (Figure 2).

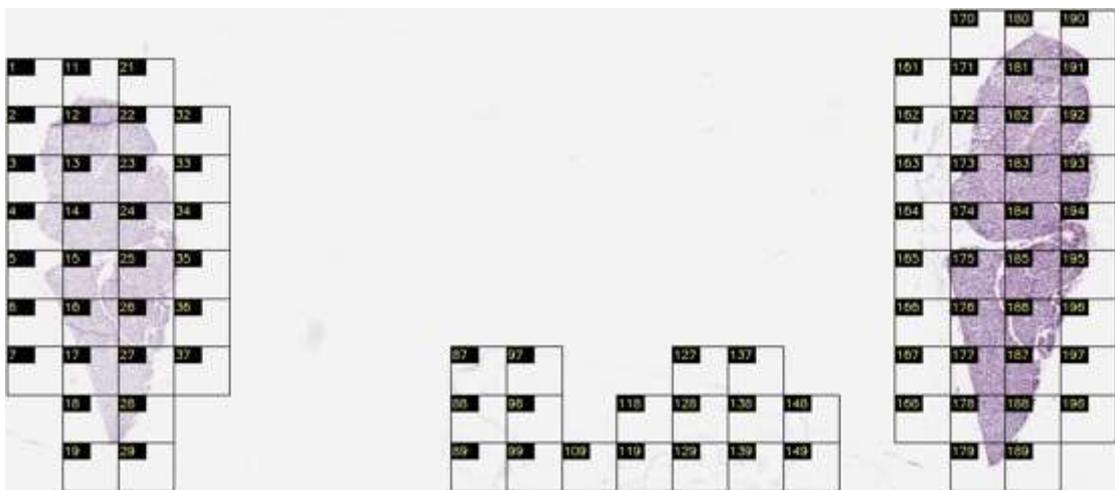


Figure 2. Example of tiles at L3 level, after exclusion of blank space

2. Nuclear segmentation

Tiles of interest from the Lo (highest resolution) layer were transformed into YCrCb colour space. Binary images were then created for each individual object by adaptive thresholding of the image global histogram to result in a set of objects that were predominantly nuclei. The location of the centre of each object was catalogued and overlaid onto the H&E image (Figure 3). Parameters of interest were then calculated and catalogued for each object separately, across all tiles of the image, such that the location of every property was recorded for future validation.

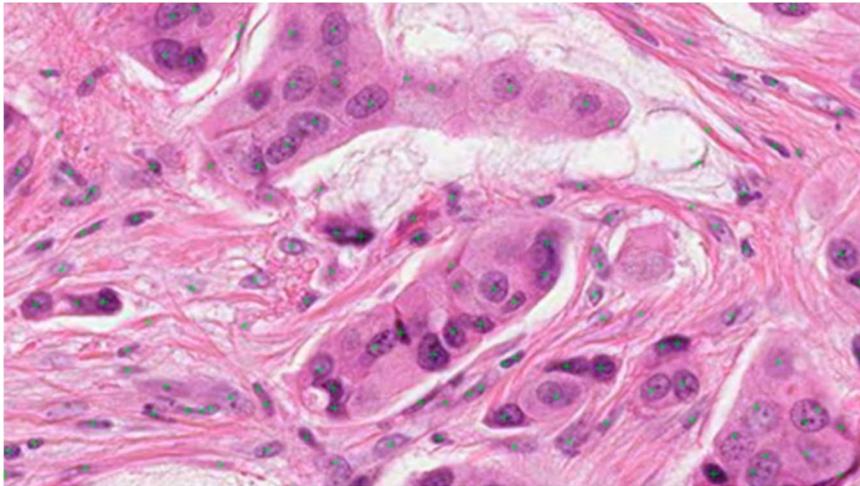


Figure 3. Example of nuclear segmentation. Each green cross denotes the centre of a nucleus.

3. Artefact exclusion

While the object catalogues comprised predominantly cell nuclei, artefacts such as pen marks or tissue folding can create false positive detection of objects. The algorithm excluded these at first pass using a non-linear support vector machine (SVM) method. I then manually checked each slide, identified harder to distinguish artefacts, and re-trained the SVM algorithm to refine the model.

4. Cell classification with SVM

The classifier was designed to classify objects into three categories, namely lymphocytes, stromal cells and tumour cells. Stromal cells include any non-epithelial and non-immune cells, such as fibroblasts, adipocytes and endothelial cells. I created a training set with approximately 1000 objects of each of category from randomly selected tiles from randomly selected images. I then overlaid the catalogue from the output of segmentation to the corresponding Lo image using an astronomical programme called *Aladin* (version 9.0). Cells were classified and the corresponding catalogue was created and saved using *TopCat* as *.fits* files. I then trained a linear SVM algorithm with this training set, and used this to classify objects from the test set into the cell categories.

5. Cell classification with Random Forests (alternative to step 4)

For this classifier, once an object was identified and artefacts excluded, a square image of the object (32 pixels x 32 pixels) was extracted (Figure 4), and used as training set for a random forest classification method in the same way as for the SVM classifier in step 4. Therefore, for every object, a classification with SVM and Random Forest was obtained for future validation (Chapter 5)

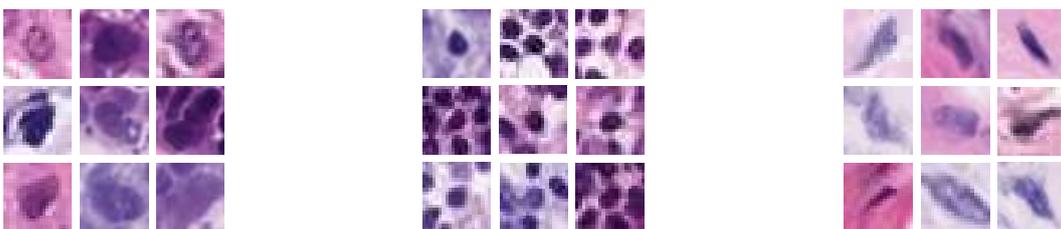


Figure 4: Examples of nuclear images used in Random Forest classifier.

Left: tumour; middle: lymphocyte; right: stromal cells

6. Final output

A final catalogue of objects is generated with the headings listed in Table 21 below, along with local image properties such hue and colour flux.

| Outputs | | Physical properties | |
|-------------------|--|----------------------------|---|
| X | X co-ordinate within tile | area_cnt | Area of the shape |
| Y | Y co-ordinate within tile | area_minCircle | Smallest circle that can contain object |
| X_global | X co-ordinate within slide | area_ellipse | Ellipse that best contains object |
| Y_global | Y co-ordinate within slide | perimeter | Object perimeter |
| overlap | Object in an overlap? (and hence duplicated) | eqDiameter | Object equatorial diameter |
| slide_mpp_x | Total slide x-dimension in pixels | extent | |
| slide_mpp_y | Total slide y-dimension in pixels | ell_angle | Ellipse angle |
| slide_mpp | Total slide resolution in pixels | ell_majorAxis | Ellipse size – major axis |
| p_corner_XY | Tile position within slide | ell_minorAxis | Ellipse size – minor axis |
| cell_WH | Tile size | ell_e | Ellipse eccentricity |
| sub_id | Subject ID | circularity | Circularity |
| slide_name | Slide name | roundness | Roundness |
| imgRefTrimmedMean | Mean pixel value | compactness | Compactness |
| object_type | Flagged as artefact? | AR | Object axis ratio |
| noise | Background noise | ell_AR | Ellipse axis ratio |
| noiseStdDev | Standard deviation of noise | solidity | Solidity |
| s2n | Signal to noise ratio | convexity | Convexity |
| cell_type | Cell classification according to SVM | flux | Optical colour property |
| cell_type_RF | Cell classification according to Random Forest | hue | Optical colour property |

Table 21: Features of objects calculated by the cell segmentation algorithm

2.6.1 Validation of the classification methods

I manually validated the accuracy of the machine learning algorithm. From a series of random images, random areas of between 250 – 500 pixels by 200 – 400 pixels were selected, to manually count and identify all cells. The image was annotated using the *Aladin* programme and exported using *TopCat*. This was used as the gold-standard to judge the automated classification of these same cells obtained using the support vector machine and random forest learning algorithms in chapter 5.

2.6.2 Lymphocyte density

Lymphocyte density was calculated using the k th nearest neighbour method (knn). The distance, r , was defined as the distance between a lymphocyte, l , and its k th nearest lymphocyte. This means that a circle with l as the centre and radius r , will have k lymphocytes within its perimeter. The density is then calculated as $k / (\pi * r^2)$.

2.7 Cell neighbourhood analysis

All calculations in this section were performed in MATLAB_R2018b (MathWorks, USA) using the Image Analysis and Bioinformatics toolboxes. My code is available at <https://github.com/copew/ImageAnalysis>.

Diagnostic biopsies often comprise several cores of tissue from the same tumour, placed close together before being sliced and stained. This may result in cells appearing artificially closer together if the cores are placed adjacent to each other. For this reason, before further analysis the H and E image was converted into a binary image, allowing for the identification of the boundary of each core using the *bwconncomp* and *bwtraceboundary* functions. The cells within each boundary were considered to be from the same core. Summary statistics for each slide were averaged across all cores.

2.7.1 Identification of cell clusters

Density-based spatial clustering of applications with noise (DBSCAN) is a clustering algorithm designed by Ester *et al.* (1996) It is a density-based clustering non-parametric algorithm to identify objects that are close together.

The identification of clusters using DBSCAN is based on two parameters:

epsilon: the radius of a neighbourhood with respect to some object

minPts: The minimum number of objects required to form a dense region

A cluster is defined as a group of at least *minPts* objects that is “connected” by having at least one other group member no greater than *epsilon* away.

There is no histological definition of a cell cluster other than subjective judgment. In order to determine *epsilon* and *minPts* robustly, taking into account variation in tumour cell sizes across samples, the following steps were used:

epsilon: this distance was calculated for each slide based on the distance between the centre of the nuclei of the cells of interest. The pair-wise Euclidean distance was calculated for all the cells of interest, and a density plot for the distances formed. The distance that corresponded to the knee of this curve was used as *epsilon* for subsequent calculations.

minPts: the value of *minPts* for tumour clustering was based on blinded rating against histopathologist visual inspection for a range of different values over a range of slides (Figure 5). The value that best delineated the tumour clusters in most cases is 5. Similarly, an optimal *minPts* value for lymphocyte clusters in most cases was 20.

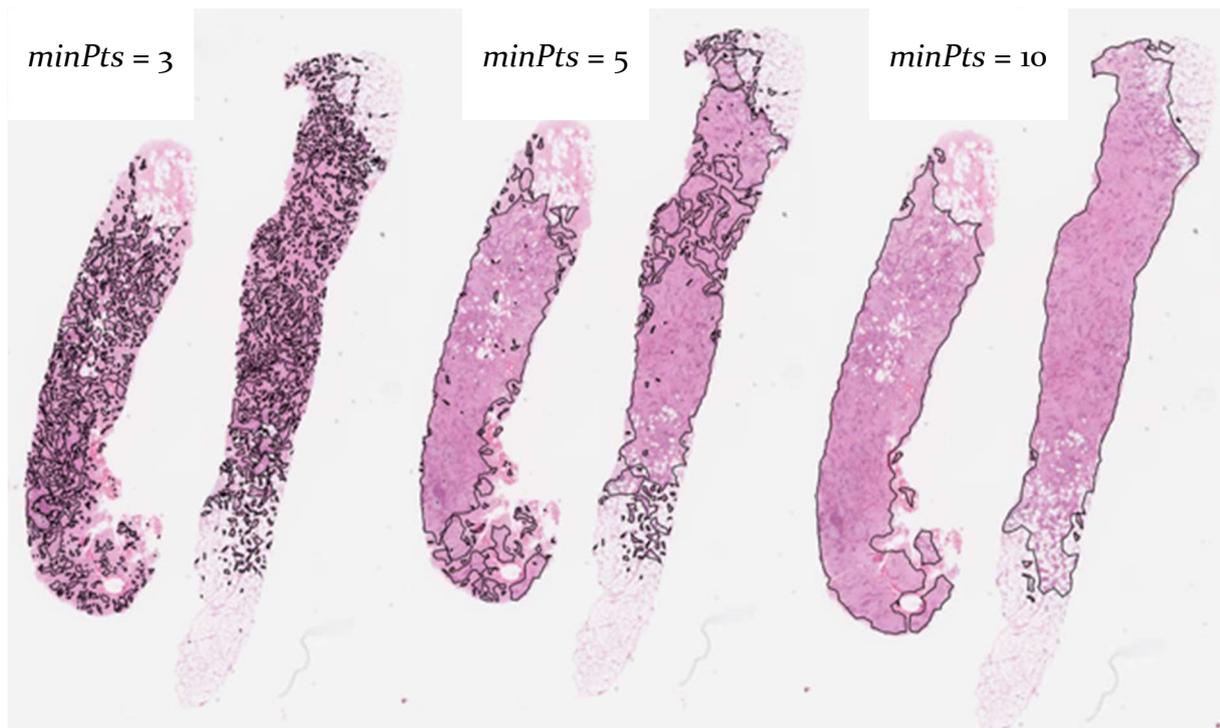


Figure 5: Three example clusterings for tumour cells. based on *minPts* values of 3, 5 and 10.

A value of 3 was too low, resulting in many small, disconnected islands within true clusters, while that of 10 was too high, resulting in the hole core being identified as tumour cluster. Blinded rating consistently identified an optimal value of 5 for tumour clusters and 20 for lymphocyte clusters.

2.7.2 Defining the peri-cluster region

The peri-cluster region of a tumour cluster was defined as the region within 50 μm (equivalent of 5 – 7 lymphocytes) of the outline of the cluster (Figure 6). These regions were then trimmed such that where there was overlap between a peri-cluster region and an adjacent tumour cluster, the overlapped region was excluded and considered to be of tumour cluster only. Where two peri-cluster regions overlapped with each other, they were not excluded and were considered to belong to both clusters.

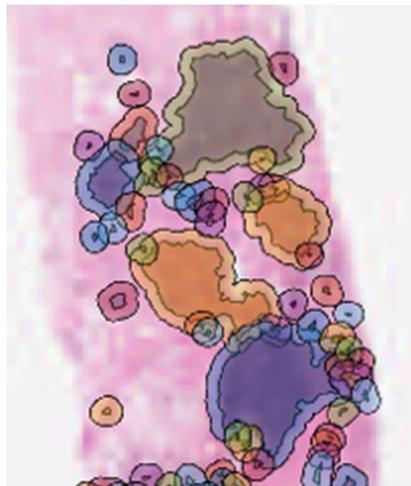


Figure 6: Tumour clusters and their surrounding peri-cluster regions

2.7.3 Distance between nuclei

Pair-wise distance was calculated from the coordinate of the centre of each cell nucleus to all other nuclei in the same core (Figure 7). The distances between a cell, C , and all other cells in the same tissue core were calculated, and sorted from the nearest to the furthest using the *pdist2* function. These distance measures are confounded by cell size, such that lymphocytes appear much closer their neighbours than tumour cells simply because they are smaller and have minimal cytoplasm, however this will be consistent across patients so will not bias inter-patient comparisons.

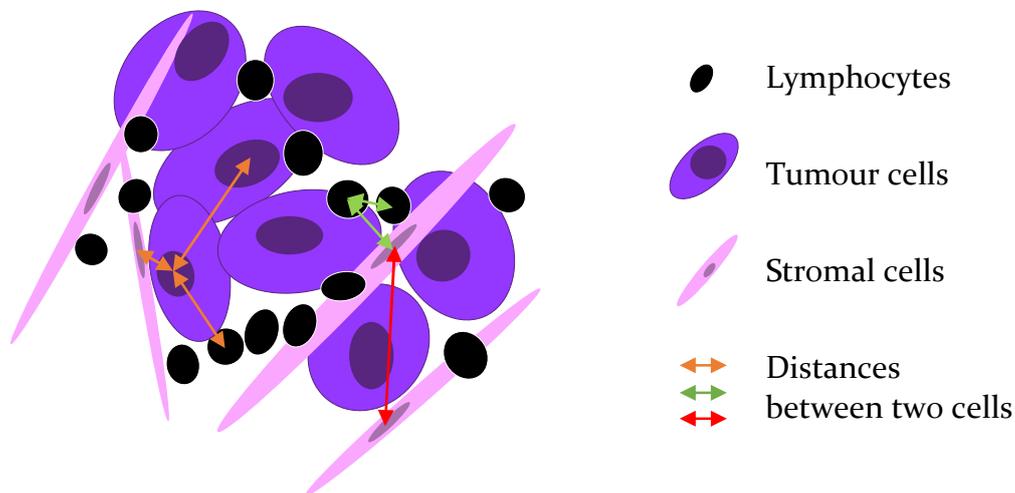


Figure 7: Schematic illustration of the calculated inter-nuclear distances

2.7.4 Neighbourhood normalisation

For each cell type, the average proportion of tumour cells, lymphocyte and stromal cell was calculated at each position, and normalised based on the overall proportion of each cell type in the slide. This results in a ratio of the observed neighbouring cell type frequency over expected neighbour cell type frequency.

Specifically, for each cell on the slide, the “closeness” of other cells can be approximated by sorting the Euclidean distances between the centre of nucleus of a given cell and all other cells in the same core. From this, the type of cells that is in the vicinity of a given cell can be identified and ordered by distance. This measure is, however, biased by the proportion of each cell type present on the slide. This effect is minimised by obtaining the ratio of the two measurements. In other words, calculating the relative closeness between any two cell types compared to chance.

For the closeness between cell types X and Y at the n^{th} closest cell from Y:

$$\text{Normalised ratio} = \frac{\textit{Proportion of cell type X that is at the } n^{\text{th}} \textit{ closest place from cell type Y}}{\textit{Proportion of cell type X on the slide.}}$$

A normalised ratio greater than 1 would suggest that these two cell types are closer to each other than expected, and vice versa. As the number of cells being included, n , increases, the normalised ratio asymptotes to 1.

3. NanoString

3.1 Preface

The first step towards our overall aim of translating precision medicine for breast cancer into a clinical setting is to develop a method to classify tumours into integrative clusters using widely available samples and simple processing techniques. This chapter describes our attempts to develop and validate a pipeline for the analysis of formalin fixed, paraffin embedded (FFPE) samples based on an FDA approved, CE marked, NICE-endorsed, commercially available platform that was previously developed for intrinsic subtypes. Overall, however, my investigation showed that measured gene expression was heavily dependent on tissue preparation method and analysis technique, and that there was consequently poor integrative cluster classification. This chapter therefore takes the form of a narrative investigation into the potential sources of discrepancy introduced by tissue preparation method and analysis technique.

I am grateful to Dr H Raza Ali and Marcus D. R. Klarqvist who designed the probe set for the NanoString experiment. I am grateful to Dr Kate Eason, who performed the computational analysis in section 3.9 Improving iC10 classification; we held discussions at every stage and jointly designed the workflow, but she implemented it *in silico*.

3.2 NanoString technology

The company NanoString was initially known for its product Prosigna[®], which is a breast cancer prognostic gene signature assay (Cesano, 2015; Eastel *et al.*, 2019). The Prosigna[®] assay is an in vitro diagnostic assay which is performed on the NanoString nCounter[®] Dx Analysis System using RNA extracted from FFPE breast tumour tissue from diagnosed breast carcinoma (Nielsen *et al.*, 2014). Prosigna[®] compares a tumour's gene expression profile to that of intrinsic subtypes (Sørlie *et al.*, 2001; Rouzier *et al.*, 2005; Dai *et al.*, 2015; Goto *et al.*, 2018), and combines this with a proliferation score, based on the expression of different genes, to produce an overall prognosis for clinicians (Wallden *et al.*, 2015). The advantage of the system is that it uses FFPE material that is readily available, requires little hands-on time, and has a fast turn-over time.

The technology relies on hybridising mRNAs with unique barcodes that are attached to complementary strands, thus quantifying mRNAs of interest. The workflow is shown in Figure 8. The number of user-defined mRNA targets that can be quantified at the same time is limited to 800 by the length of the barcode. Within each probe set there are internal positive and negative controls.

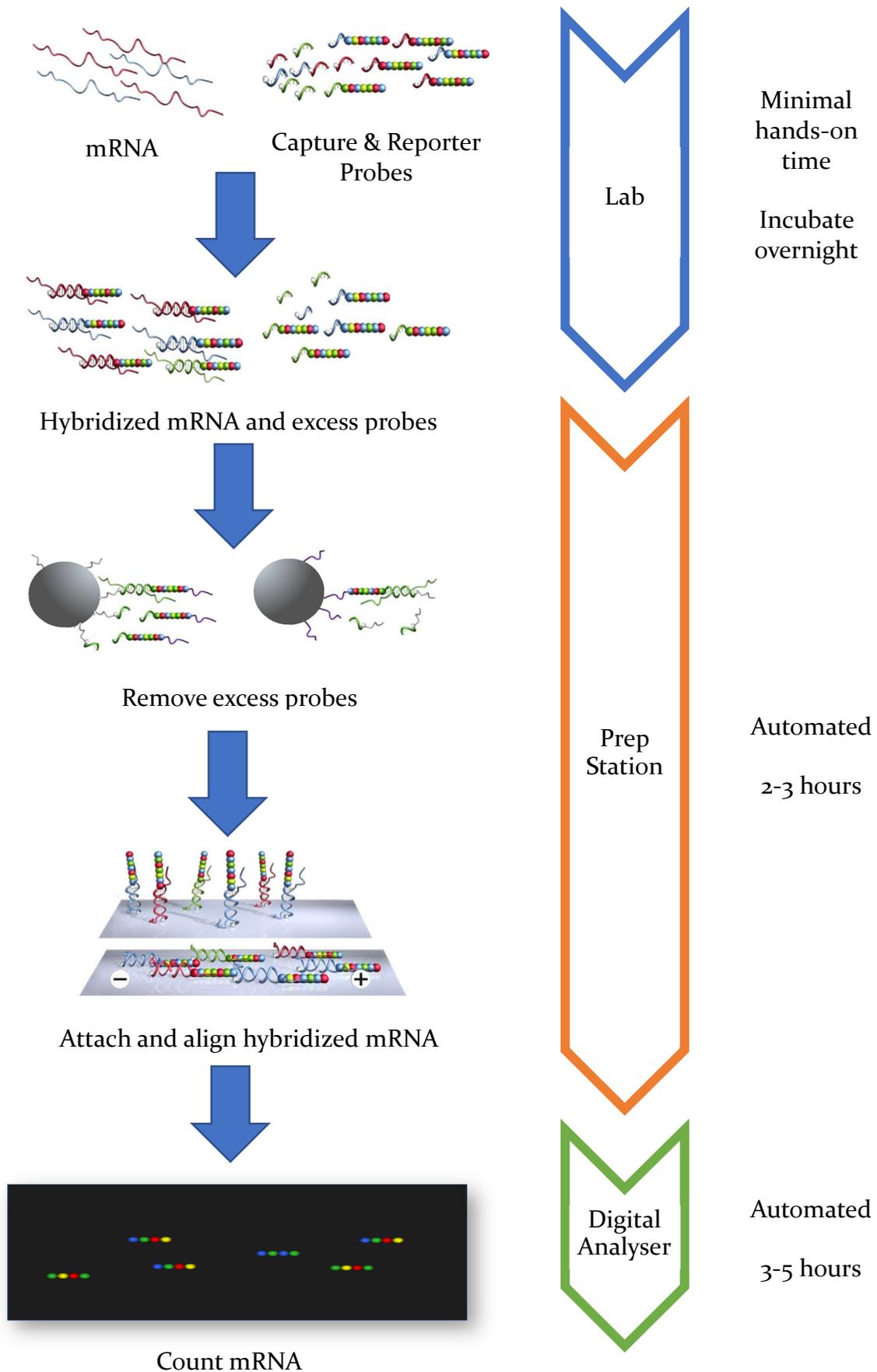


Figure 8: NanoString® platform workflow, adapted from NanoString training material

3.3 Designing the probe set

The aim of this chapter is to extend the NanoString platform from intrinsic subtypes to integrative clusters, based on widely available FFPE tissue. Integrative clusters were originally based on the expression of 754 genes from fresh frozen tissue (Curtis *et al.*, 2012). However, using all of these genes in a clinical assay would be prohibitively expensive because the price scales linearly with the number of probes. In fact, although it is theoretically possible to sequence 800 mRNAs in a research setting, standard quotations from the company for commercial use on large sample numbers only scale to a maximum of 300 genes. It was therefore decided to reduce the number of features to reduce the cost to that which would be achievable by an NHS organisation. Dr H Raza Ali and Marcus Klarqvist reduced the original 754 features to 207 for the NanoString platform. Probe selection was based on clustering and expression data from the original Illumina® microarray published by Curtis *et al.* (2012).

To reduce the cost of this experiment to the charitable funders (Cancer Research UK), we entered a data sharing agreement with NanoString to include an additional 566 genes of their specification on the same samples. These targets were primarily immune markers. Therefore, the custom probe set comprised 773 targets, of which 207 were for the purpose of classification of integrative clusters. The set also included 6 positive controls, 8 negative controls, and endogenous controls. The whole custom probe set is provided as an appendix to this thesis.

3.4 Sample set

I extracted RNA from 562 METABRIC samples, 18 TransNEo samples and 23 NeoTango samples. 456 of these samples, all from the METABRIC study, were part of the original study set published by Curtis *et al.* (2012), and therefore had “known” integrative clusters. These were the samples used for final analysis, with the primary validation measure being concordance between classification based on NanoString® data and the original Illumina® microarray using RNA from fresh frozen tissue. Across the 456 samples there was good representation from all integrative clusters (IC) (Figure 9). The remaining 147 samples were intended to be a prospective test set if the original integrative clusters could be accurately determined on the new platform.

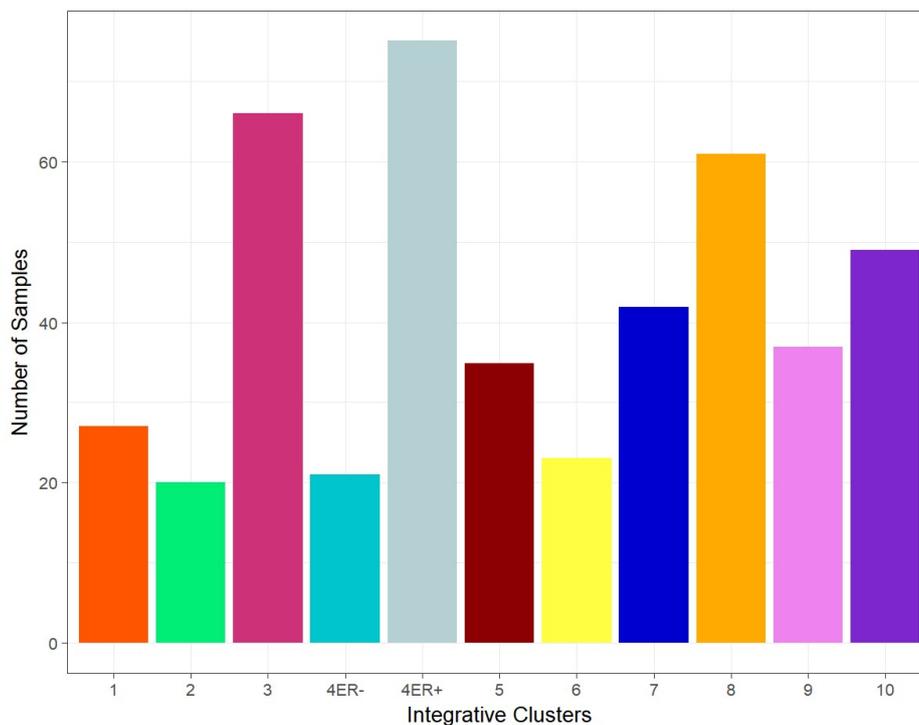


Figure 9: Representation of integrative clusters across the 456 validation samples.

3.5 RNA Extraction

The median concentration the RNA extracted from the original study samples was 122.5 ng/ μ l (IQR 64.5 – 202.6) (Figure 10). Each sample has 30 μ l of RNA. The median concentration for the samples that did not overlap with the original study was 116.4 ng/ μ l (IQR 68.4 – 169.8). (Figure 11)

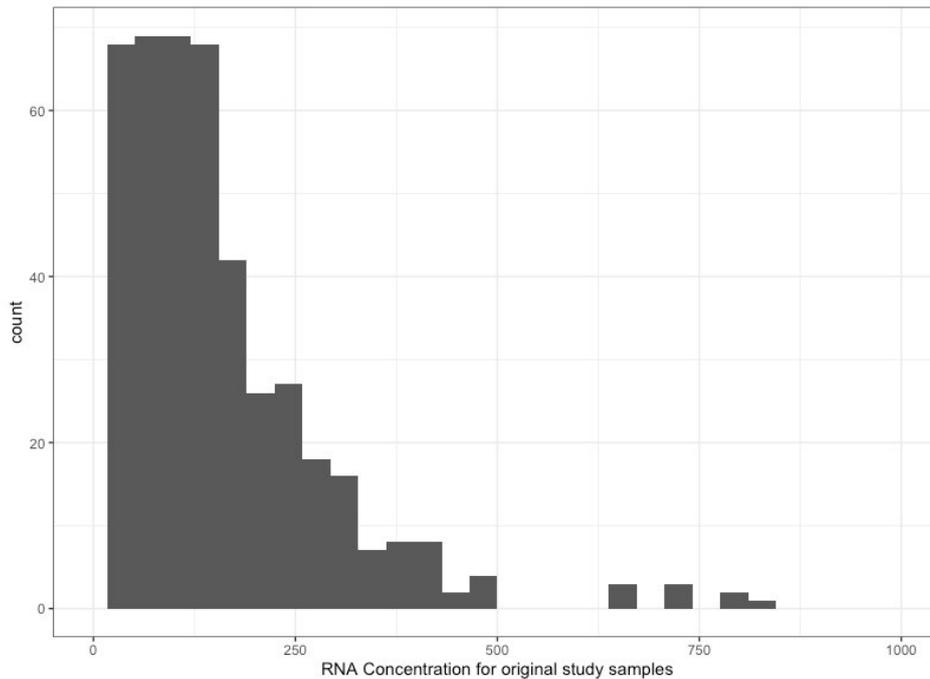


Figure 10: Extracted RNA concentration for the samples that overlapped with the original study

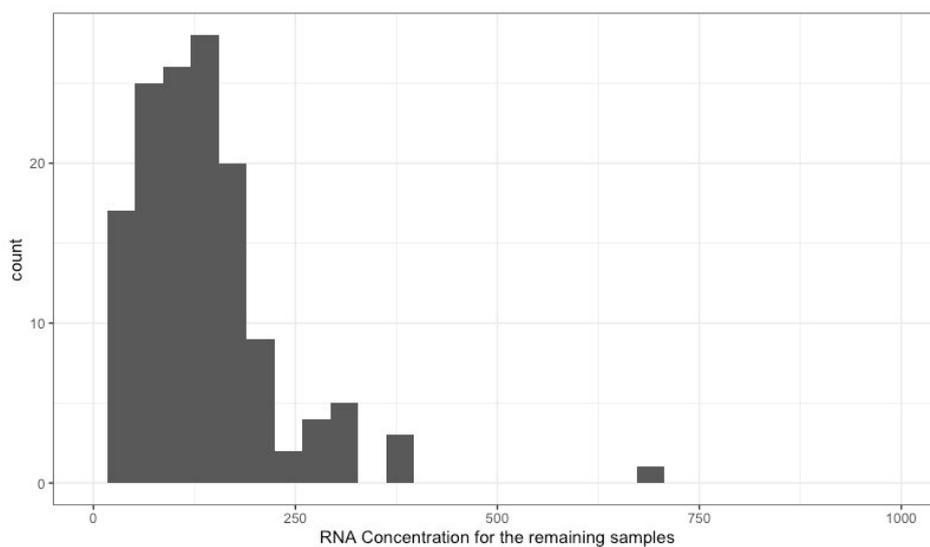


Figure 11: Extracted RNA concentration for the samples that did not overlap with the original study

3.6 RNA quality control

3.6.1 RNA Purity

The 260/280 ratio is a measurement of nucleic acid purity (Patterson and Dackerman, 1952), with ~2.0 being “pure”. Measured using Nanodrop (Desjardins and Conklin, 2010), the RNA extracted from the samples that overlapped with the original study had a median ratio of 1.95 (IQR 1.91– 1.98) (Figure 12). There was only one outlier, with a ratio of less than one. I did not exclude it at this stage, but marked it for closer review after NanoString analysis. Those that did not overlap with the original study had a median ratio of 1.94 (IQR 1.90 – 1.98) (Figure 13), with no outliers.

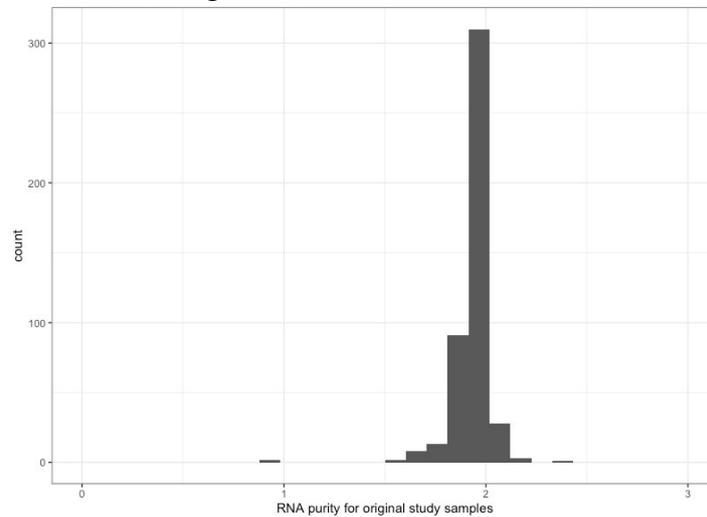


Figure 12: Purity of extracted RNA for the samples that overlapped with the original study

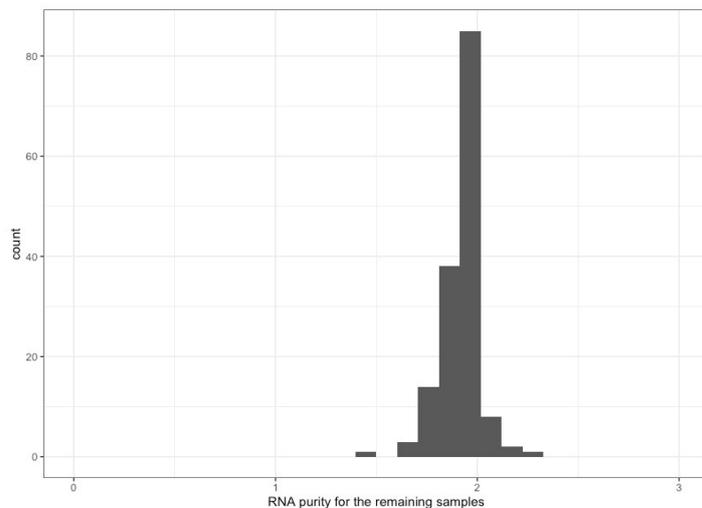


Figure 13: Purity of extracted RNA for the samples that did not overlap with the original study

3.6.2 RNA Integrity

RNA integrity number (RIN) is a widely used measurement of the quality of the RNA extracted based on the characteristics of electropherogram outputs (Mueller *et al.*, 2004; Fleige and Pfaffl, 2006; Schroeder *et al.*, 2006). The RNA extracted from the samples that overlapped with the original study had a median RIN of 2.50 (IQR 2.30– 2.65). Those that did not overlap with the original study had a median ratio of 2.50 (IQR 2.20 – 2.70). Overall, the RIN for these samples was comparable with other published RIN based on FFPE samples (Roberts *et al.*, 2009; Chen *et al.*, 2016), but significantly lower than that generally observed in fresh frozen tissue, where RIN values between 4 and 10 are the norm (Fleige and Pfaffl, 2006). Again, no samples were excluded based on their RIN value.

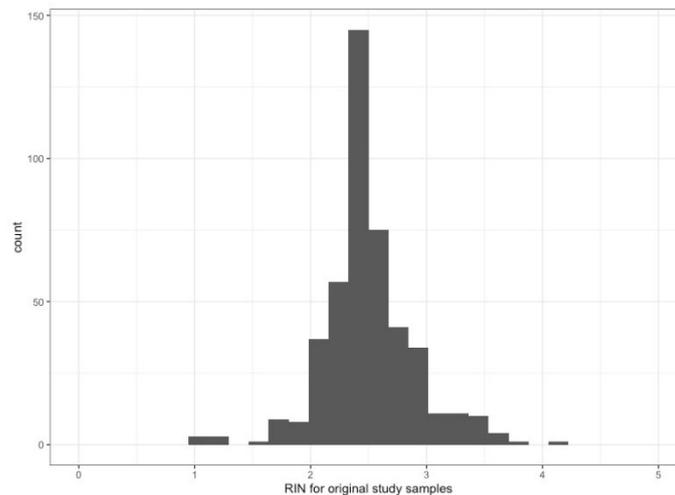


Figure 14: Integrity of extracted RNA for the samples that overlapped with the original study

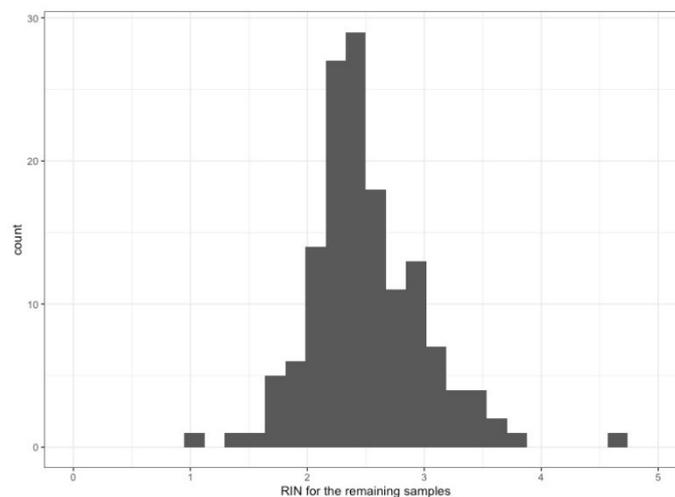


Figure 15: Integrity of extracted RNA for the samples that did not overlap with the original study

3.7 NanoString Results

All results in this section are for the 456 samples that were in the original Curtis *et al.* (2012) paper. Please note that several of the p-values in this section appear identical at $p < 2.2 \times 10^{-16}$, because this is the smallest threshold that can be computed by R version 3.6.2.

3.7.1 RNA count

The output was normalised using the NanoStringNorm package (Waggott *et al.*, 2012) in R prior to further analysis. This constitutes a multi-step normalisation process that accounts for technical artefacts resulting from variability in sample preservation and RNA extraction, as well as fluctuations in the analysis platform.

Before normalisation, the median total count across all samples was 409947 across a range of three orders of magnitude from 16682-1714877. The median normalised count per sample was 323070, and the distribution showed much greater kurtosis, and all samples were within a range of only one order of magnitude from 125426-997163 (Figure 16).

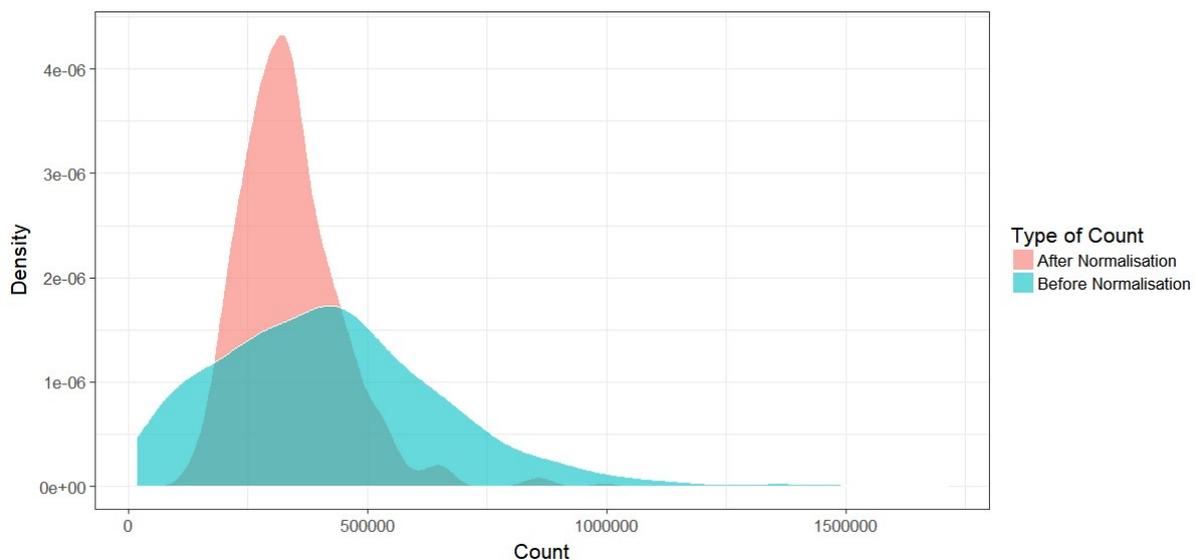


Figure 16: Total RNA count before and after the application of NanoStringNorm, for all samples

3.7.2 Initial validation on receptor status

One of the key determinants of breast cancer outcome and therapeutic choice is oestrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) status. Clinically, this is usually determined based on immunohistochemistry (IHC) and fluorescent in-situ hybridisation (FISH). These two methods have high concordance where immunohistochemistry is definitively negative or strongly positive, with FISH providing useful adjudication in borderline (“IHC 2+”) cases (Dowsett *et al.*, 2003).

As initial validation of our NanoString dataset I quantified the mRNA expressions of the *ESR1* (which codes for the ER protein product) and *ERBB2* (which codes for the HER2 protein product) genes and compared these to immunohistochemistry in our samples. Those samples that were classified as ER+ by IHC showed significantly higher *ESR1* expression using the NanoString technique (Wilcoxon $p < 2.2 \times 10^{-16}$), but there were some outlier and overlap samples from both groups (Figure 17).

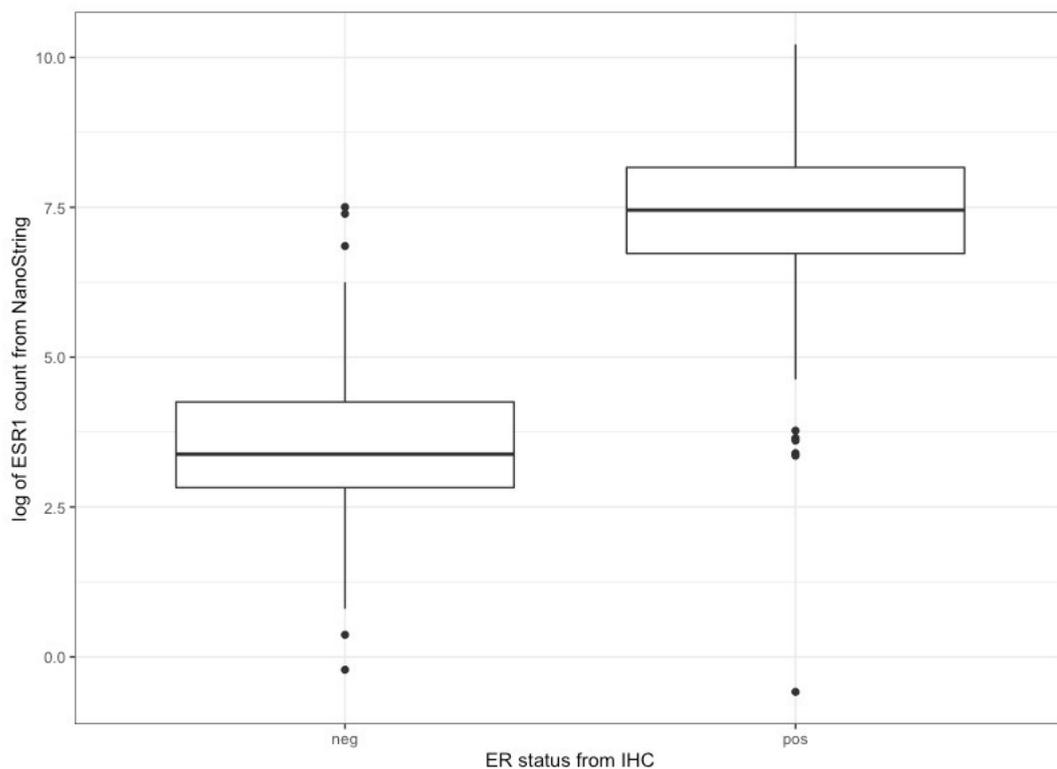


Figure 17: Relationship between *ESR1* expression from the NanoString technique and ER status from Immunohistochemistry

Those samples that were classified as strongly positive for HER2 by IHC (IHC 3) showed significantly higher ERBB2 expression using the NanoString technique than those classified as negative for HER2 (IHC 0/1) (Wilcoxon $p = 2.54 \times 10^{-12}$). However, those samples that were classified as borderline (IHC 2) showed low ERBB2 expression on NanoString, with no significant difference in expression between IHC 0/1 and IHC 2 (Wilcoxon $p = 0.662$), and IHC 2 being significantly lower than IHC 3 (Wilcoxon $p = 1.20 \times 10^{-3}$). There was only one exception sample, with high NanoString ERBB2 expression and IHC 2. I do not have FISH for our samples, but from the literature I would have expected approximately half of IHC 2 samples to be deemed HER2 positive from FISH (Dowsett *et al.*, 2003). It therefore does not seem that NanoString has the potential to replace FISH in adjudicating these samples. Again, there was some overlap between classes, indicating that NanoString does not perform perfectly at HER2 classification, even for differentiating IHC 0/1 from IHC 3.

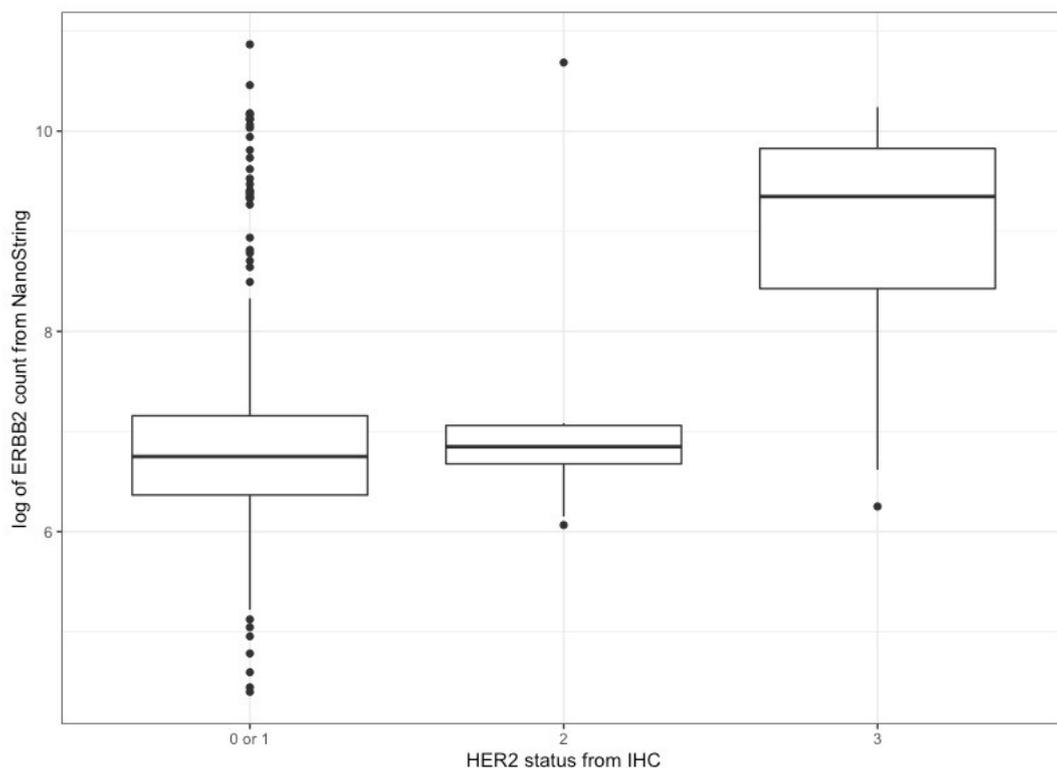


Figure 18: Relationship between ERBB2 expression from the NanoString technique and HER2 status from Immunohistochemistry

Next, I compared the expression of ESR1 and ERBB2 using our NanoString technique to the receptor status classification obtained from the original Illumina® microarray. Again, those samples that were classified as ER+ by Illumina showed significantly higher ESR1 expression using the NanoString technique (Wilcoxon $p < 2.2 \times 10^{-16}$), but there were again some outliers and overlap samples from both groups (Figure 19).

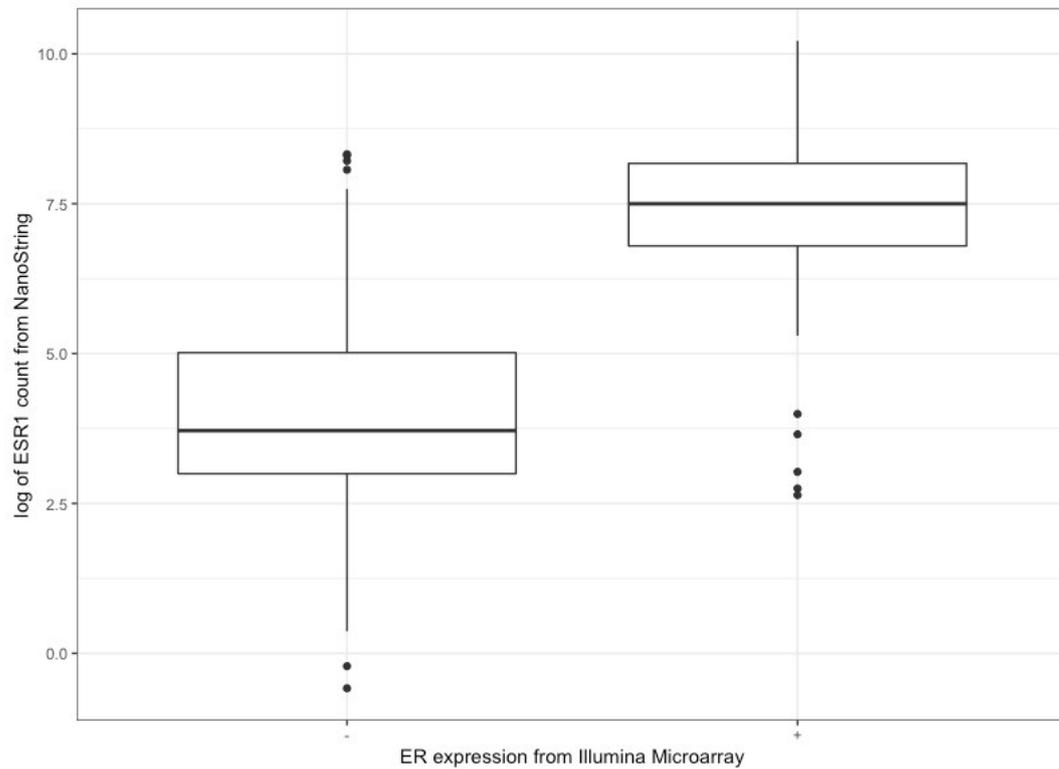


Figure 19: Relationship between ESR1 expression from the NanoString technique and ER status from the Illumina Microarray

Similarly, those samples that were classified as HER2+ by Illumina showed significantly higher ERBB2 expression using the NanoString technique (Wilcoxon $p < 2.2 \times 10^{-16}$), again with some outlier and overlap samples, predominantly from the HER2 negative group (Figure 20).

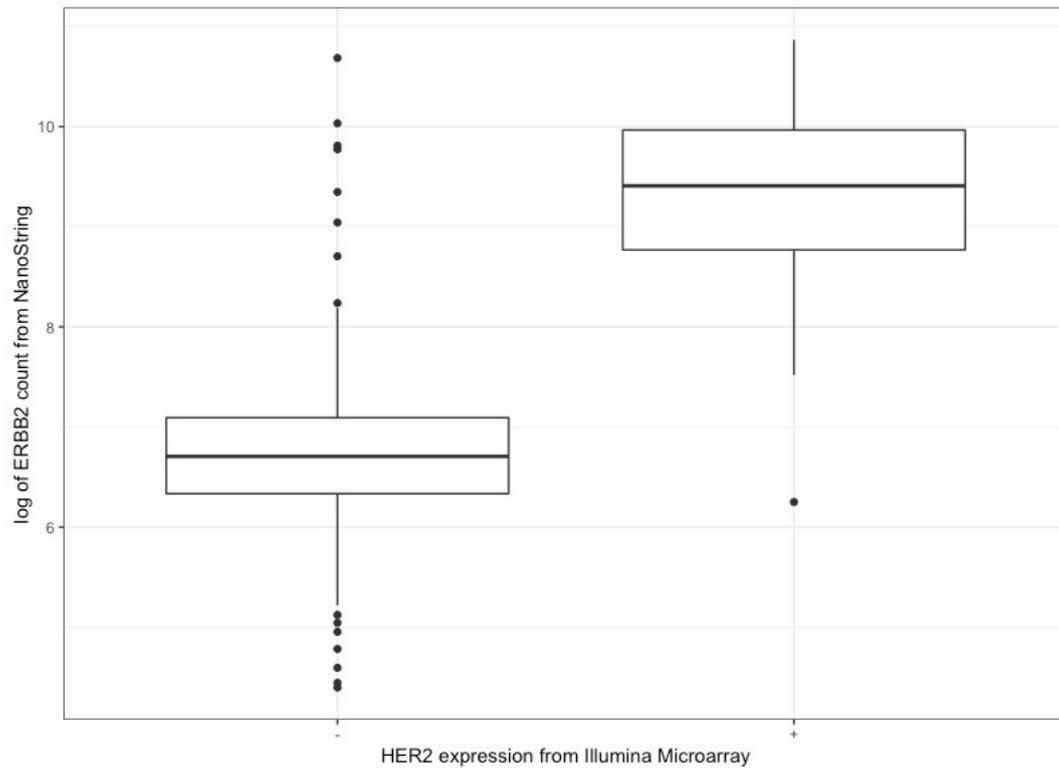


Figure 20: Relationship between ERBB2 expression from the NanoString technique and HER2 status from the Illumina Microarray

As further validation I quantified expression of the GRB7 gene. GRB7 is located in close proximity to the ERBB2 gene, and these genes tend to be over-expressed in unison (Walch *et al.*, 2004; Ramsey *et al.*, 2011; Bivin *et al.*, 2017), meaning that it can be used as a surrogate measure of HER2 status. In this NanoString dataset, expression of GRB7 and ERBB2 was very strongly correlated ($F(1,454) = 2919$, $p < 2.2 \times 10^{-16}$, adjusted $r^2=0.865$) (Figure 21).

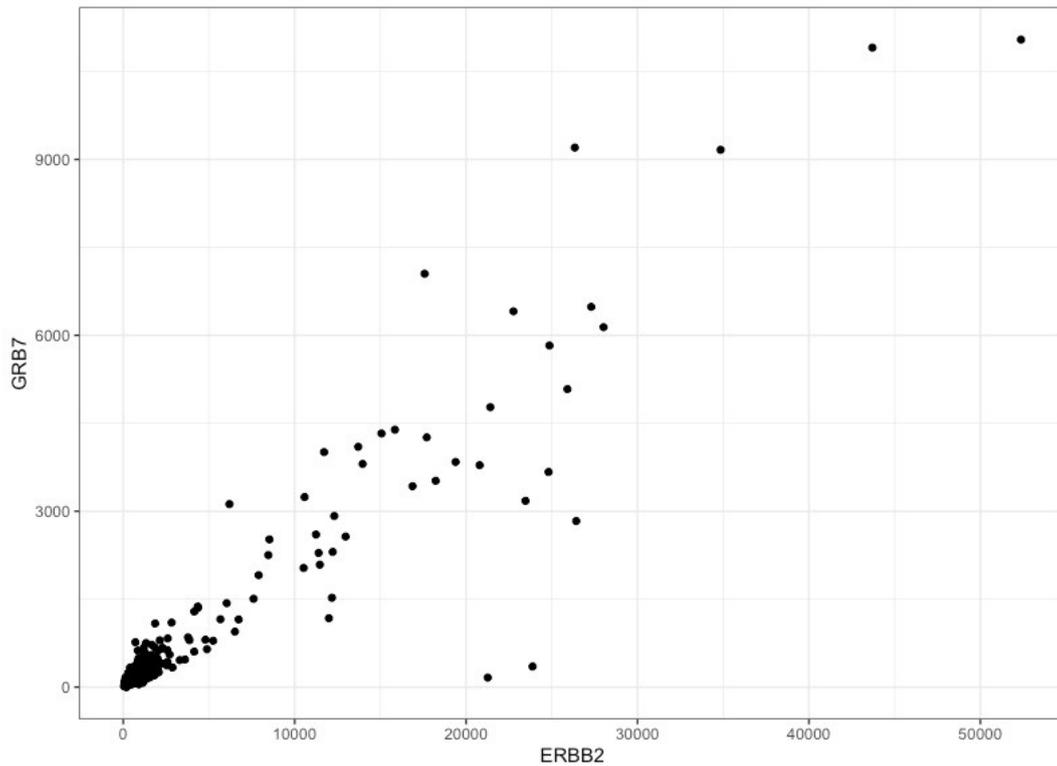


Figure 21: Strong correlation between GRB7 and ERBB2 (HER2) gene expressions as measured by NanoString on FFPE tissue, adjusted $r^2=0.865$

To check that this was not a non-specific effect relating to a failure of our normalisation procedures to fully account for total overall expression I examined the correlation of ESR1 and ERBB2 (Figure 22). ER and HER2 status are independent, and expression of these mRNAs should not be correlated. Reassuringly, in our dataset very little of the variance in ESR1 was accounted for by ERBB2 expression (adjusted $r^2=0.013$).

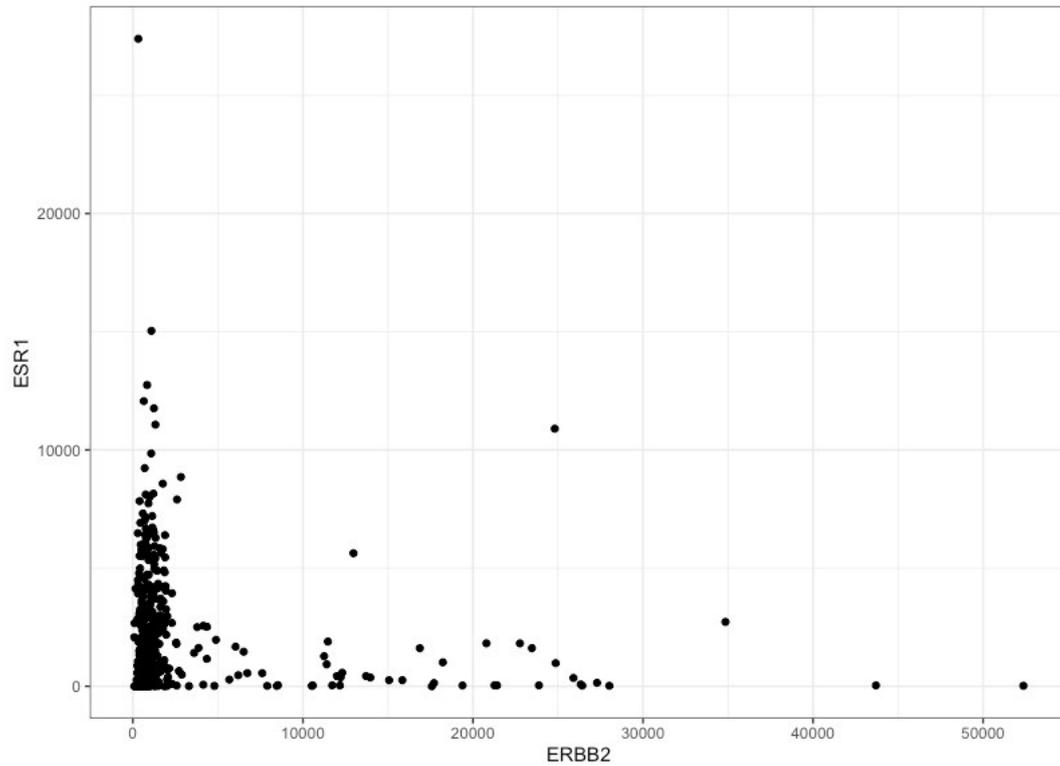


Figure 22: Very weak correlation between ESR1 (ER) and ERBB2 (HER2) gene expressions, adjusted $r^2=0.013$

3.7.3 Validation across genes of interest

The next step in validating my technique was to compare expression of the genes of interest from FFPE samples using our NanoString method to the expression value of the same genes from fresh frozen tissue quantified with Illumina as in the original classifier (Curtis *et al.*, 2012). I did this in two ways.

Firstly, for every sample I ranked the relative expression of every gene using both techniques to produce an individual expression profile. I then quantified the non-parametric correlation between the methods across all genes for each sample individually. This non-parametric approach looking at the expression profile as a whole is particularly powerful when the expression of any individual gene might be differentially affected by the tissue preparation technique. Similar approaches have been widely used in neuroimaging to compare the distribution of brain signals across modalities (Kriegeskorte *et al.*, 2008). Performing this within-patient across-gene correlation showed moderate agreement, with a median Spearman's rho of 0.660 (Figure 23). The two techniques correlated much better than chance for all samples (minimum rho (205) = 0.3484, $p < 0.00001$), but even in the best case showed less than complete agreement about gene expression order (maximum rho (205) = 0.7684).

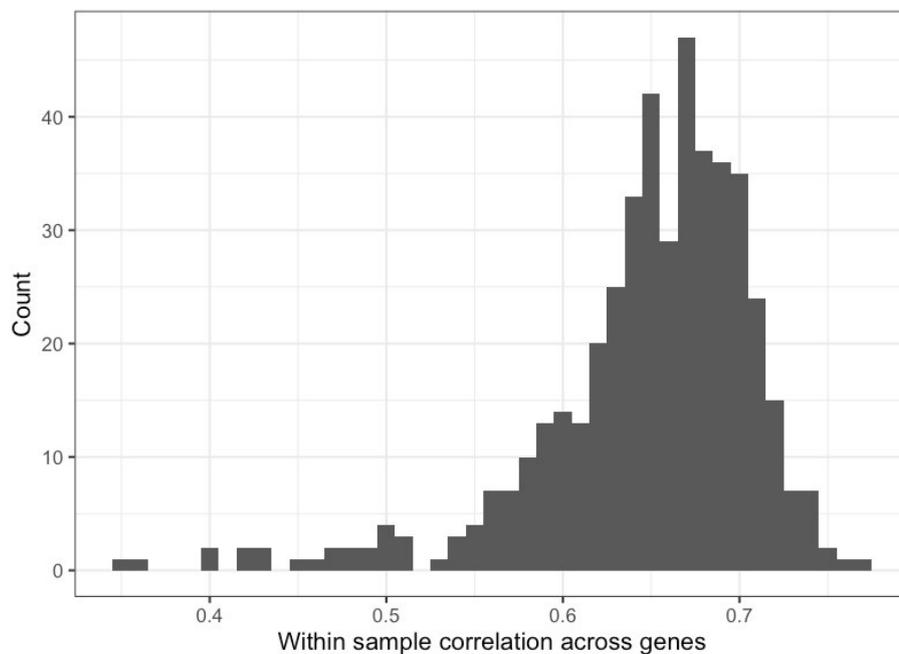


Figure 23: Spearman's rank correlation between gene expression orders across NanoString and Illumina techniques for every sample individually

This first technique tells us how well the assays agree about the overall expression order of genes, but is by design blind to large differences in individual genes. Therefore, secondly, I assessed the expression data of each gene individually by examining the correlation within-genes across-samples. In other words, I asked if expression of each gene was high in the same individual samples on Illumina and NanoString. This gives us an insight into which genes can be safely transferred between the platforms, and which are more vulnerable to differences in preparation technique. Overall, the expression data of all individual genes correlated significantly across samples, with a median rho of 0.56 (Figure 24).

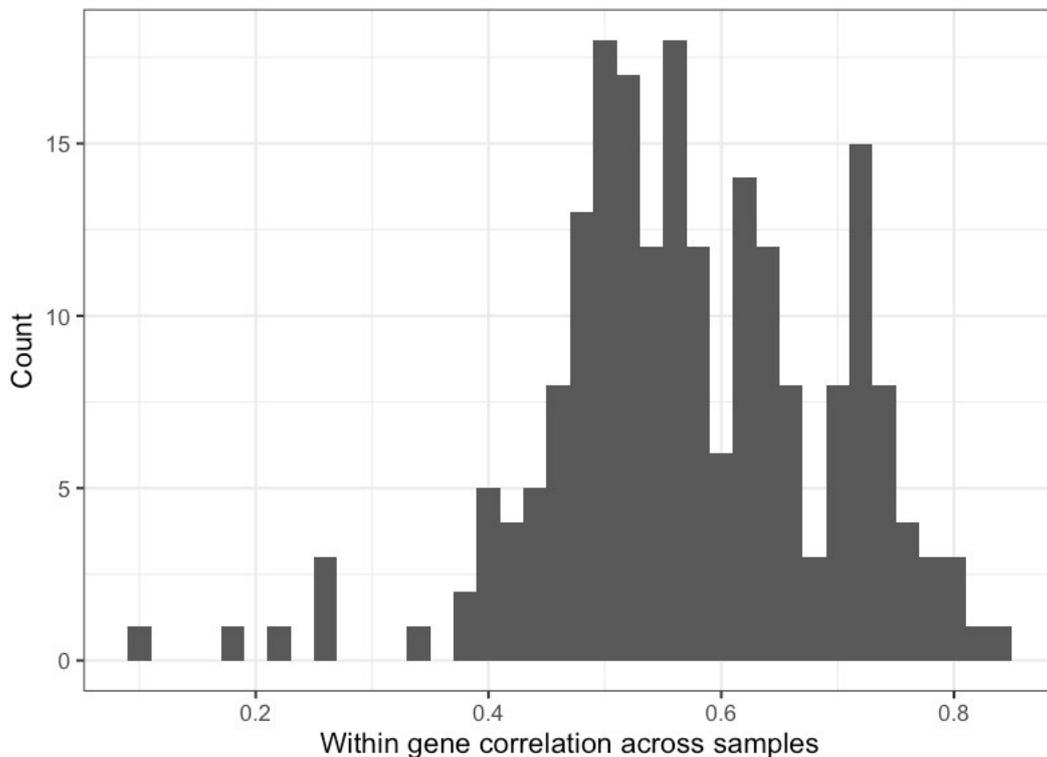


Figure 24: Spearman's rank correlation between sample expression orders across NanoString and Illumina techniques for every gene individually

However, for some genes very little of the across-sample variance was explained, and the correlation was only marginally significant (minimum rho (444) = 0.1047, $p = 0.027$ for gene CDCA3). The best correlation was observed for gene ESR1, with a spearman's rho of 0.83, confirming the excellent agreement between NanoString and Illumina for ER status already shown in Figure 19. ER status is a graded response with a wide

variation between tumour samples, so a high correlation in this gene is reassuring. The genes that correlated most ($\rho > 0.75$) and least ($\rho < 0.4$) well between the NanoString and Illumina methods are listed in Table 22.

| Strongest correlations | | Weakest correlations | |
|-------------------------------|-------------------------|-----------------------------|-------------------------|
| Gene Name | Correlation Coefficient | Gene Name | Correlation Coefficient |
| ESR1 | 0.83 | CDCA3 | 0.10 |
| SCUBE2 | 0.82 | JTB | 0.18 |
| DNALI1 | 0.79 | CHTOP | 0.22 |
| C1orf106 | 0.79 | RSF1 | 0.25 |
| AFF3 | 0.79 | KCTD21 | 0.26 |
| RAB11FIP1 | 0.78 | RFWD2 | 0.27 |
| FAM134B | 0.77 | CDK12 | 0.35 |
| NOSTRIN | 0.77 | OTUD6B | 0.38 |
| PPAPDC1B | 0.76 | FTSJ3 | 0.38 |
| CAPN8 | 0.76 | DEDD | 0.39 |
| TRIP13 | 0.75 | PIGM | 0.39 |
| LAD1 | 0.75 | | |

Table 22: The genes that displayed the strongest and weakest correlations

These are genes that displayed the strongest ($\rho > 0.75$) and weakest ($\rho < 0.4$) correlations in expression across samples

3.7.4 Relationship to integrative clusters

Here I break down first stage of the validation analysis in section 3.7.3 by integrative cluster membership. First, I assessed whether the within-sample across-gene correlations shown in Figure 23 varied by integrative cluster membership (Figure 25). There was a statistically significant difference in the degree of within-sample across-gene correlation between integrative clusters (ANOVA $F(9,446) = 8.66$, $p < 5.24 \times 10^{-12}$). Samples from integrative clusters 3 and 9 were particularly well correlated across genes, while those in integrative cluster 4 performed less well.

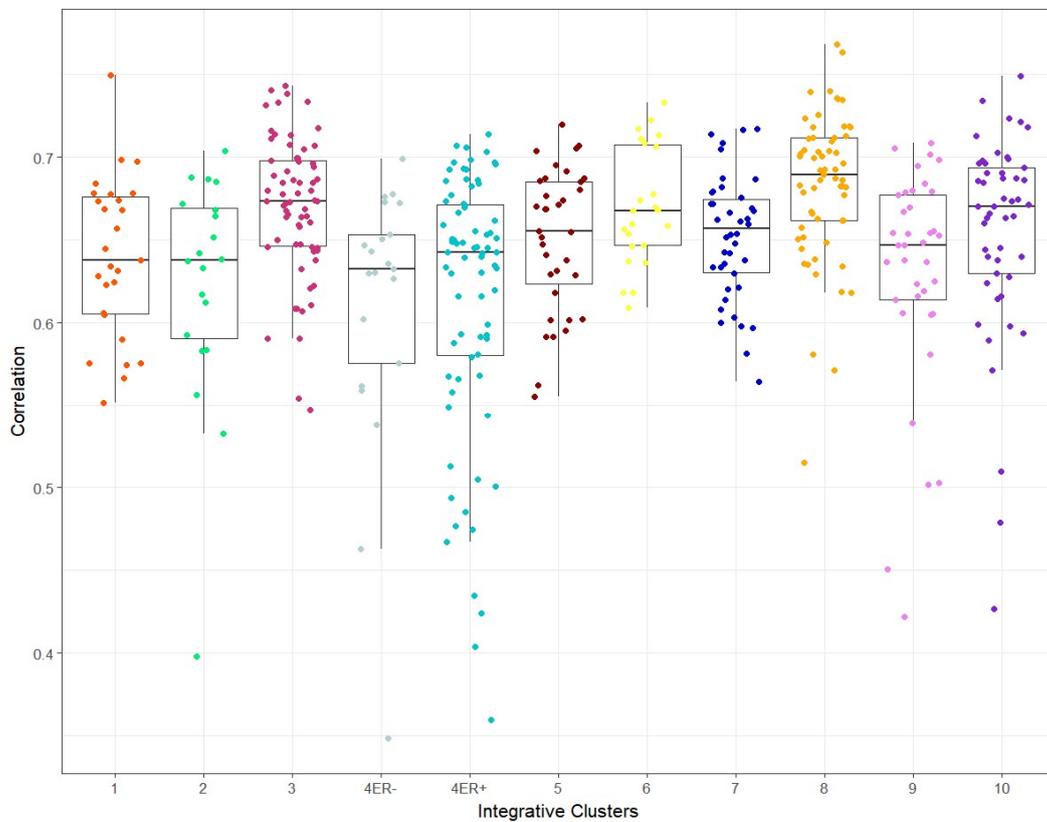


Figure 25: Spearman's rank correlation between gene expression orders across NanoString and Illumina techniques for every sample, broken down by known integrative cluster membership

This can be seen visually by plotting a cross-correlation matrix (“heatmap”), ordered by integrative cluster membership (Figure 26). In concordance with Figure 25, the strongest block-structure can be seen for integrative clusters 3 and 8, while integrative cluster 4 has a much looser correlation.

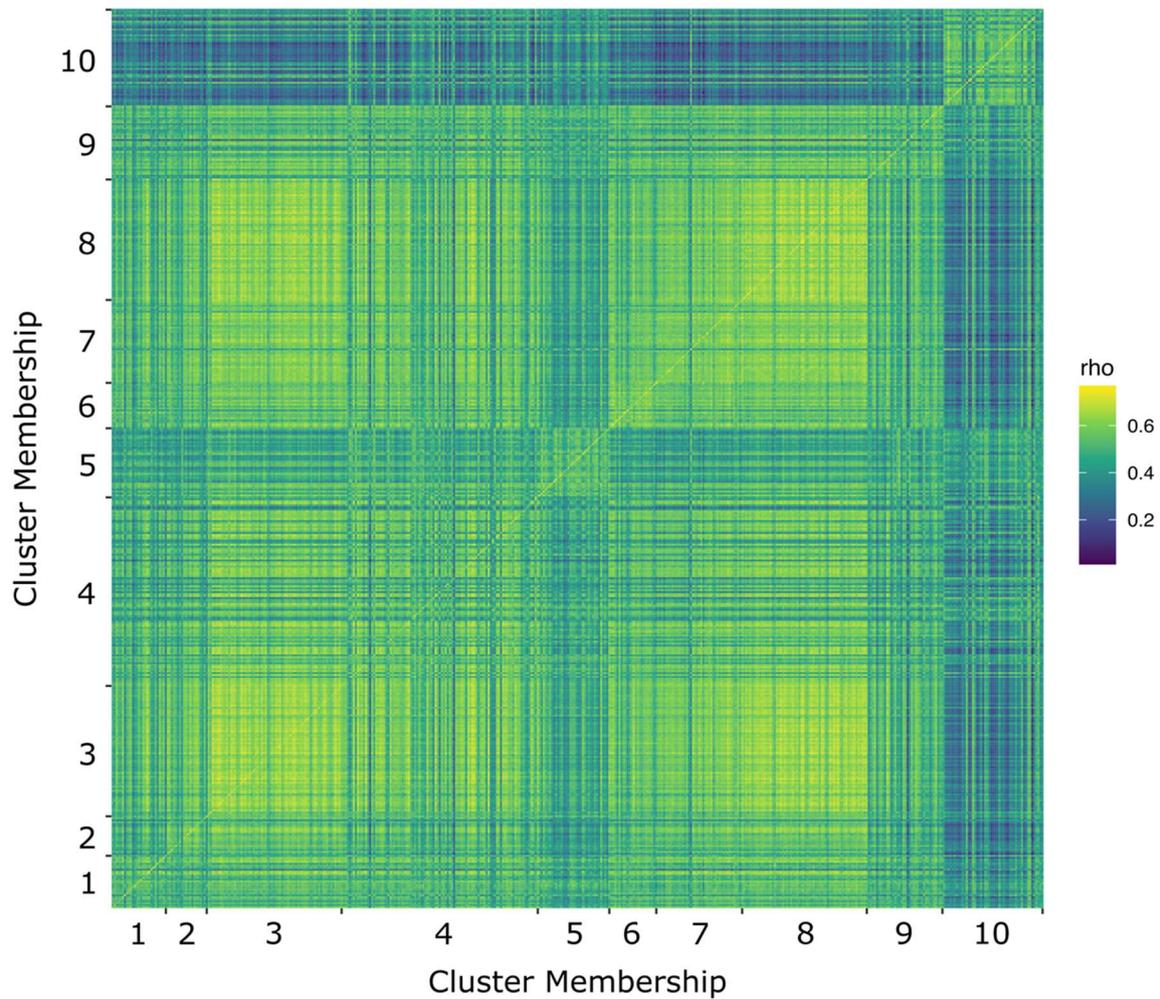


Figure 26: Cross-correlation matrix within sample across gene, ordered by integrative cluster membership

Next, I assessed how well our data were able to recapitulate the “known” integrative cluster membership of the 456 samples that were also part of the original cohort. I performed unsupervised clustering on the expression data from both Illumina and NanoString methods and compared the outcome to the originally published integrative clusters.

The Illumina data here was a sub-set of that used in to define “known” integrative cluster membership. In essence, I was assessing whether it was possible to automatically recapitulate integrative clusters based on only the 207 pre-specified genes rather than the 754 in the original paper. Using these data (Figure 27) some integrative clusters, for example clusters 5 and 10, remain very discretely classified, but others, such as clusters 3 and 8 are forming discrete sub-groups along distant cluster branches.

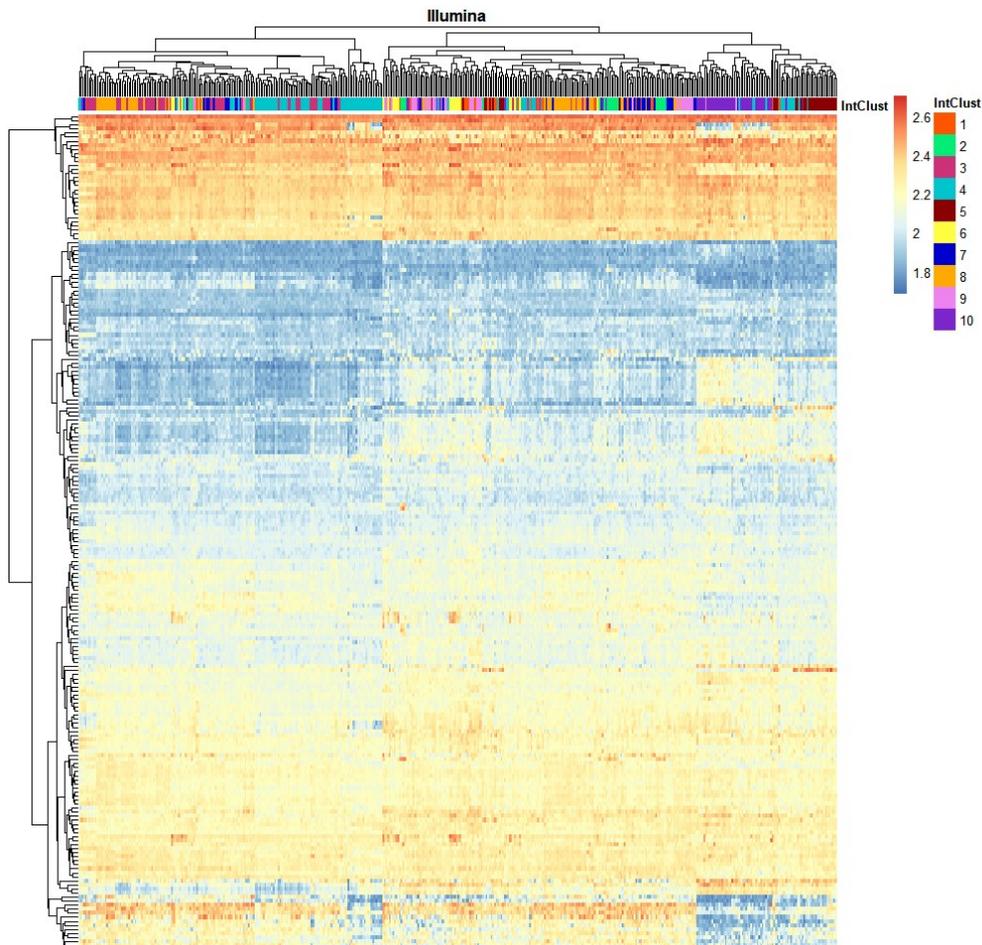


Figure 27: Heatmap of unsupervised clustering of using data from Illumina restricted to 207 genes Each column is a sample and each row is a gene.

The success of this method can be assessed by examining the coloured bar at the top of the chart for contiguous blocks.

Performing the same unsupervised clustering with NanoString data produced fewer discrete clusters that were in agreement with the published data, although there were still some clusters of high agreement, for example cluster 10 (Figure 28).

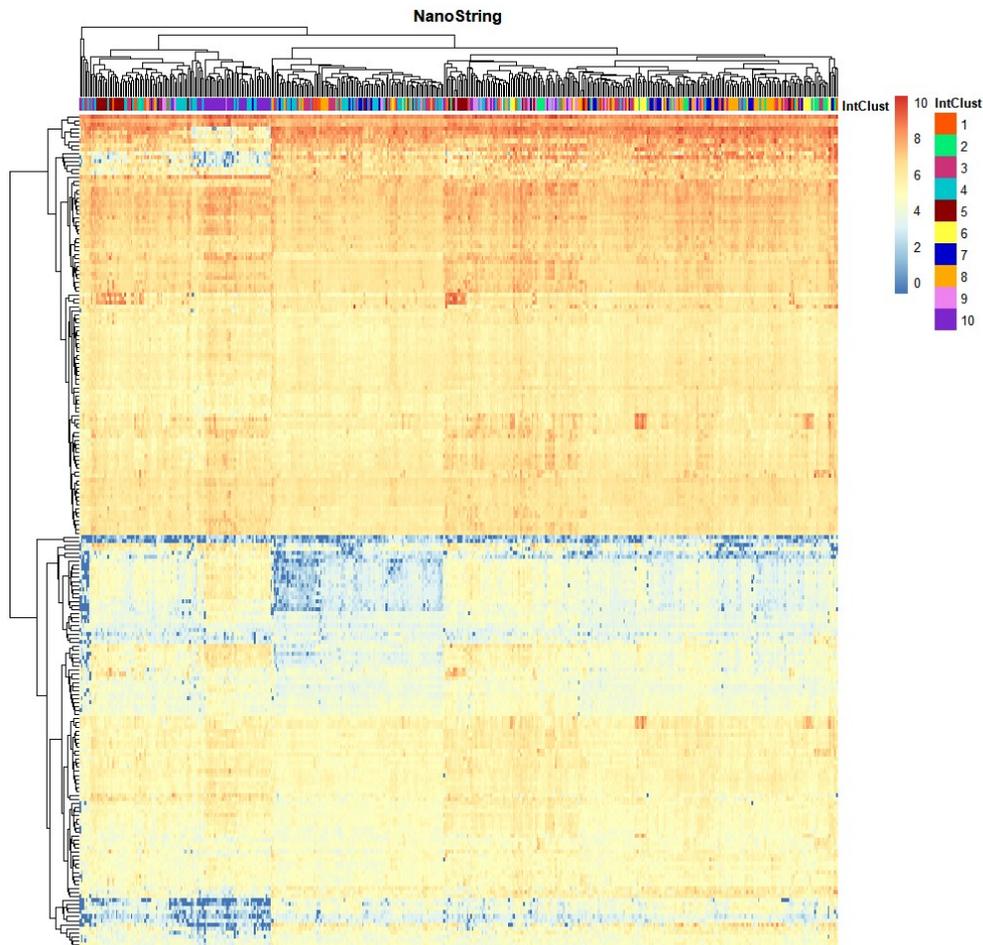


Figure 28: Heatmap of unsupervised Clustering of using data from NanoString restricted to 207 genes. Each row is a gene, and each column is a sample

Overall, this classification was disappointing. Unsupervised clustering weights all genes relatively equally, and I have a-priori reasons to believe that it might be the case that relatively few genes were responsible for cluster membership, for example integrative cluster 5 is largely determined by HER2 status. This is shown for the Illumina data in Figure 29, where samples have been arranged along the x-axis by known cluster membership, and unsupervised clustering has been performed only on the genes (arranged along the y-axis) for visualisation. Many genes display relatively uniform expression across integrative clusters, with some blocks of strong correlation for each group in only a few genes.

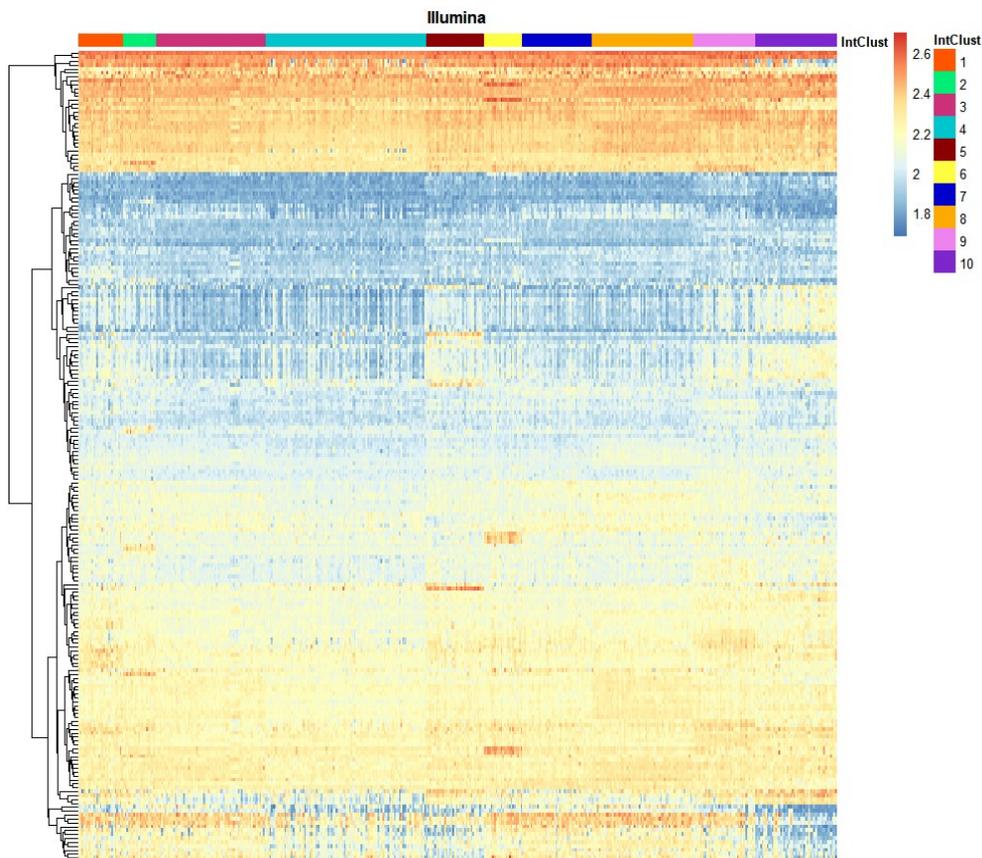


Figure 29: Heatmap of across-sample correlations for the expression of every gene for the Illumina data, arranged by known integrative cluster.

For the NanoString data (Figure 30) this was even more the case, with most genes displaying a very uniform correlation across integrative clusters. There was, however, still quite a lot of structure noticeable in the data. This was not restricted to integrative cluster 5, which showed the expected strong block of correlation in HER2 related genes, but can also be seen in integrative clusters 2 and 6. This gave me hope that a more targeted, machine-learning based algorithm may be superior to unsupervised clustering, as features of interest could be identified and weighted appropriately.

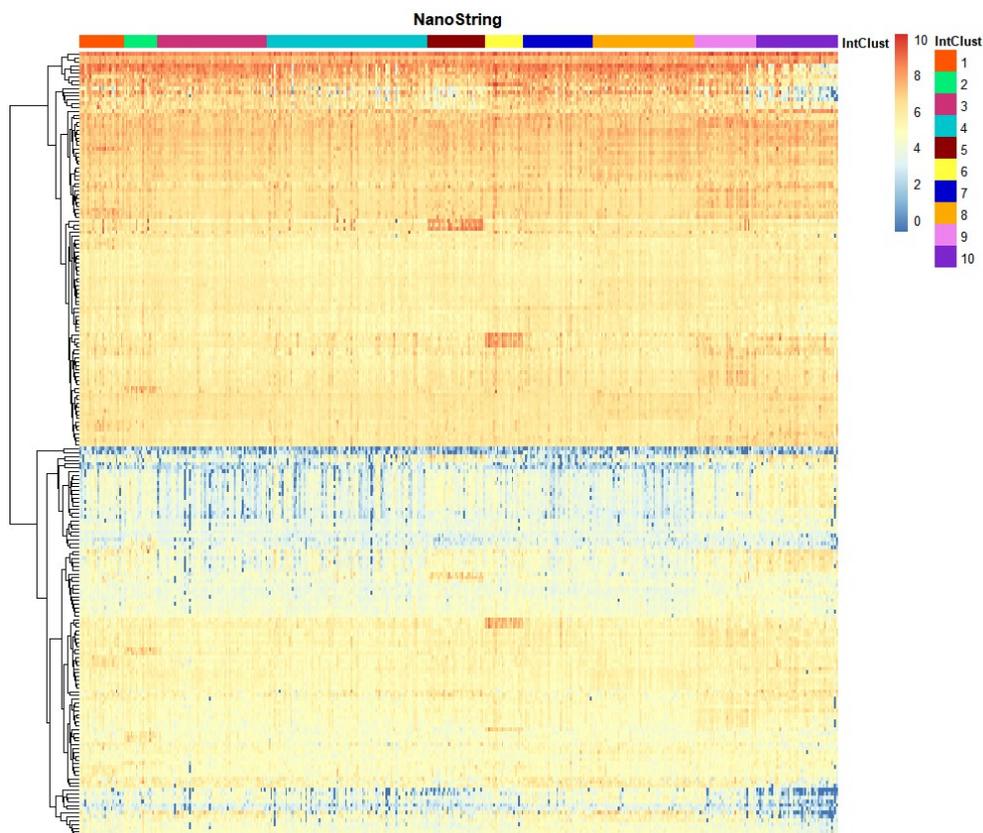


Figure 30: Heatmap of across-sample correlations for the expression of every gene for the NanoString data, arranged by known integrative cluster.

Differential expression like this can be examined parametrically by normalising expression across all samples and computing a z-score for each cluster individually, such that positive scores indicate increased expression in that cluster, and negative scores decreased expression (Figure 31). Genes with potential utility in distinguishing clusters will show z-scores that significantly differ from zero in only a few clusters. Clusters that can be potentially distinguished will show z scores that significantly differ from zero in at least one gene. Using the NanoString data, many genes display the desired properties. Some integrative clusters, for example clusters 2, 6 and 10, also display the desired properties, however others, such as clusters 4 and 7, display uniformly ‘normal’ gene expression across the whole panel, theoretically making them difficult to distinguish.

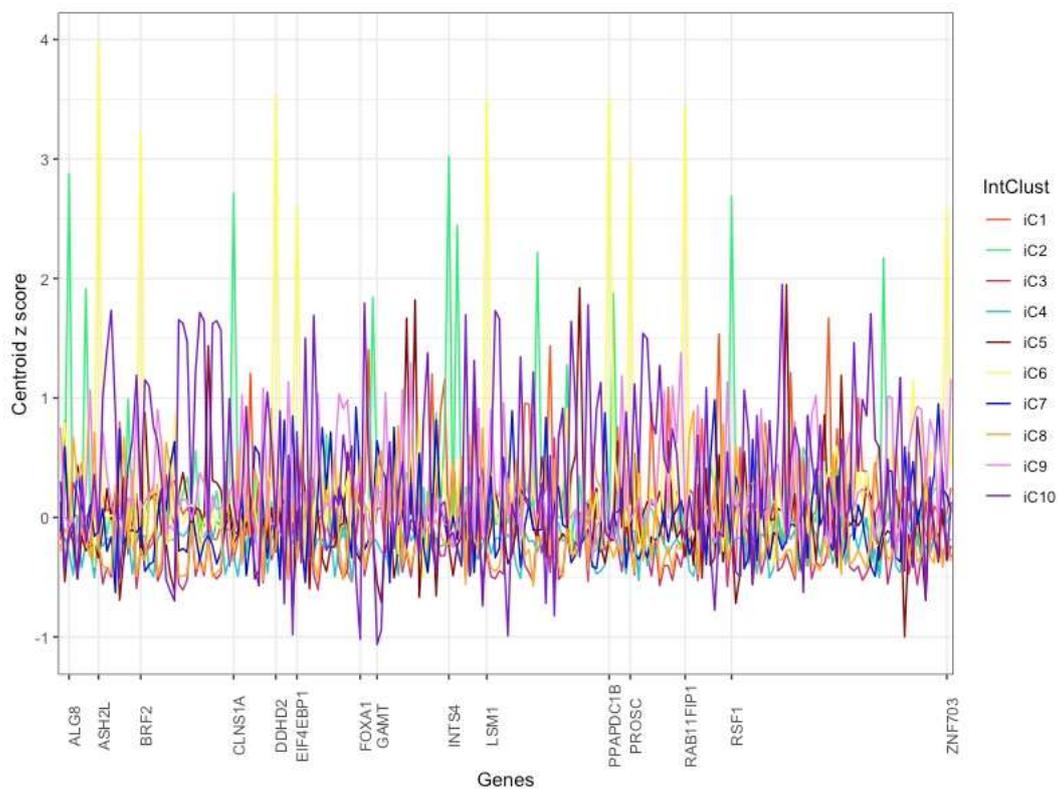


Figure 31: Differential gene expression by integrative cluster, normalised across all samples.

Genes are arranged alphabetically along the x-axis. For illustration, gene names are shown where a cluster displays a centroid z-score of >2.5 or <-1 , but any absolute value greater than the noise threshold of approximately 0.5 is likely to be useful to a classifying algorithm.

3.8 Building a classifier based on NanoString data

3.8.1 Using the existing iC10 package

The original integrative cluster classification method (Curtis *et al.*, 2012) has already been developed into an expression-based classifier package, iC10 (Ali *et al.*, 2014; Rueda, 2015). I first attempted to use iC10 for the expression data acquired through NanoString to classify our samples, with the aim of comparing the results with the original integrative cluster assignment from Curtis *et al.* (2012). The iC10 package is based on a subset of 612 genes from the 754 in the original paper. 139 of these genes were in my NanoString probe set. Across these genes, the expression profiles for each integrative cluster obtained by my NanoString method (the “test set”) was very similar to the originally published data (the “training set”) (Table 23) (Figure 32).

| Integrative Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Pearson Correlation | 0.857 | 0.950 | 0.935 | 0.835 | 0.931 | 0.952 | 0.915 | 0.962 | 0.961 | 0.915 |

Table 23: Pearson correlation in gene expression profiles for each integrative cluster between the training data for the iC10 algorithm and our NanoString data.

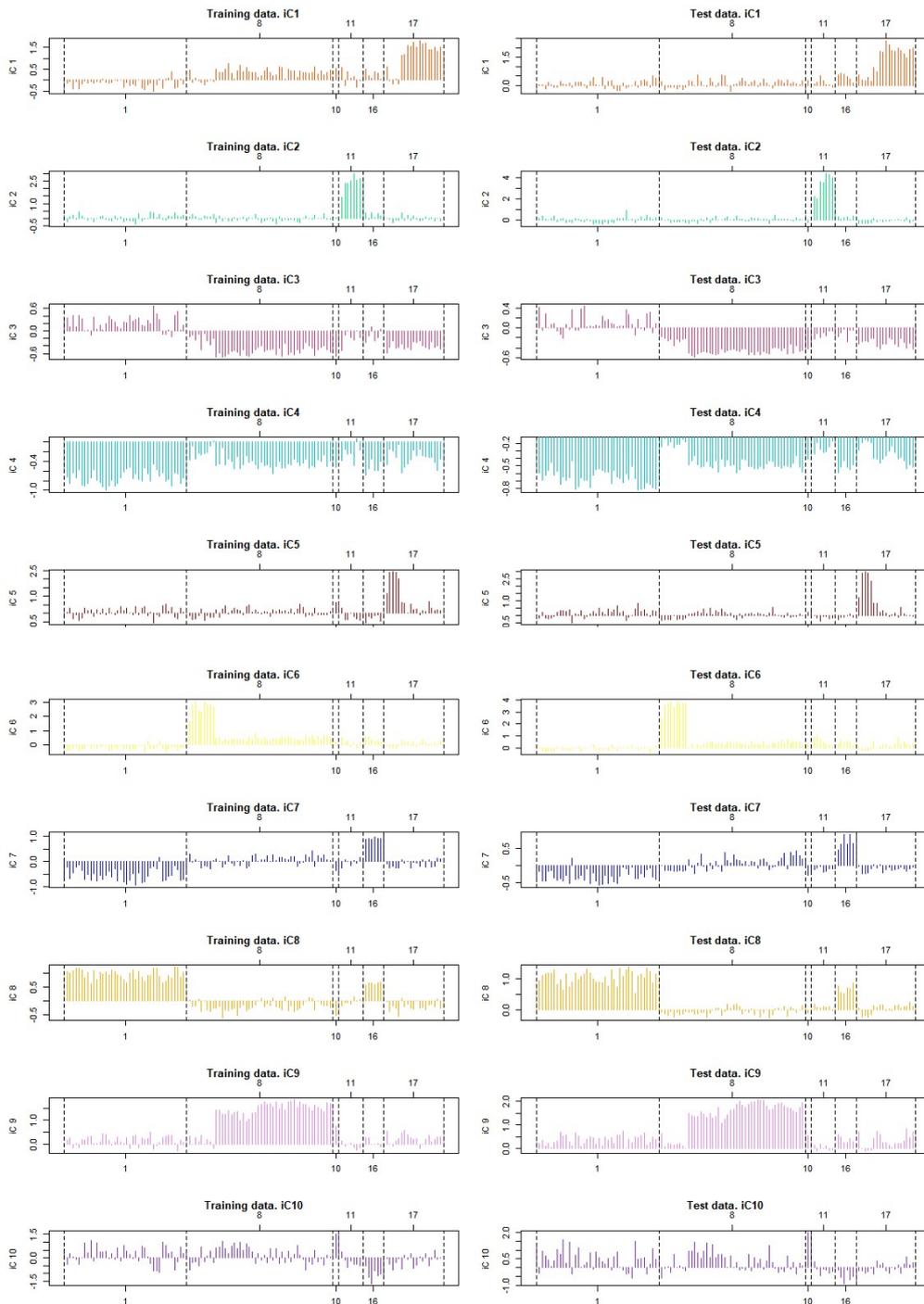


Figure 32. Gene expressions for the original iC_{10} training data (left), compared to our NanoString data (right) for the 139 genes present in both datasets.

The profiles are visually very similar, as might be expected from the high overall correlations between datasets (Table 23).

Despite this excellent correlation in gene expression, only 230 of the 456 samples (50.4%) were correctly classified by the iC₁₀ algorithm using NanoString data (Table 24). Encouragingly, samples were correctly classified into each integrative cluster at much better than chance. However, the algorithm was more likely to place samples in some clusters than others, for example a large number of samples were placed by the algorithm into integrative cluster 4. This resulted in positive predictive values that varied markedly, from 95% for integrative cluster 6 down to 21% for integrative cluster 4.

| Intergrative Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1 | 0.48 | 0.05 | 0.02 | 0.06 | 0.06 | 0.00 | 0.00 | 0.02 | 0.05 | 0.02 |
| 2 | 0.00 | 0.45 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 |
| 3 | 0.04 | 0.10 | 0.52 | 0.14 | 0.06 | 0.04 | 0.07 | 0.34 | 0.05 | 0.04 |
| 4 | 0.30 | 0.15 | 0.30 | 0.56 | 0.26 | 0.26 | 0.33 | 0.16 | 0.19 | 0.12 |
| 5 | 0.00 | 0.05 | 0.00 | 0.03 | 0.54 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 |
| 6 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.61 | 0.00 | 0.00 | 0.00 | 0.02 |
| 7 | 0.00 | 0.00 | 0.05 | 0.06 | 0.03 | 0.00 | 0.33 | 0.02 | 0.08 | 0.02 |
| 8 | 0.15 | 0.10 | 0.11 | 0.02 | 0.03 | 0.04 | 0.12 | 0.44 | 0.03 | 0.06 |
| 9 | 0.04 | 0.10 | 0.00 | 0.07 | 0.03 | 0.04 | 0.14 | 0.02 | 0.51 | 0.16 |
| 10 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.55 |

Table 24: Confusion matrix for sample classification using the iC₁₀ algorithm on NanoString data

Columns represent “true” class, while rows represent iC₁₀ classification. The diagonal, in bold, shows the proportion of each integrative cluster that was correctly classified.

3.8.2 Decision tree and random forest classifier

While there is obvious appeal in applying an independent and existing classification method like iC10 to our data, it is possible that a more bespoke method may produce a superior result. Given the differential gene expression between integrative clusters shown in Figure 30 and Figure 31, the data seem suitable for the application of a decision tree classifier based on random forests. This supervised classifier generates flowcharts that separate groups based on individual features, making binary decisions at each stage to separate data until a final classification is reached.

For every sample, the optimal decision tree was constructed based on all of the other samples, and then classification accuracy of the sample itself was assessed with leave-one-out cross-validation. An example decision tree is shown in Figure 33. This is relatively typical in that it is only based on a small number of the most discriminatory genes across the panel (in this case only 15 of 207).

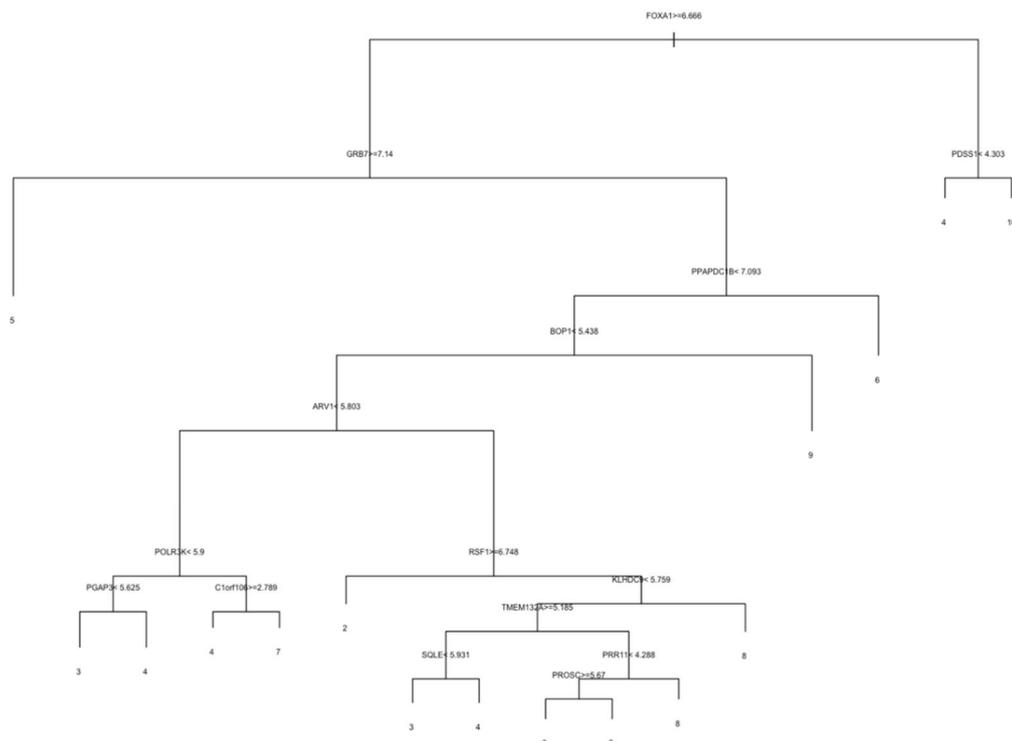


Figure 33: An example decision tree for classifying a single sample based on all other samples (leave-one-out cross-validation).

This demonstrates how, by starting at the “top” of this chart, depending on the expression of each gene, a sample can be classified according to the criteria at each “junction”, and arrive at a final integrative cluster (the numbers at the end of the “branches”)

Classification accuracies using this decision tree method were disappointingly poor, with only 38% of samples correctly classified (Table 25). Kappa is an overall metric of classification performance, corrected for imbalanced group sizes, with a value of zero representing random chance and a value of 1 representing perfect classification; for these data it was 0.284. Further training of the classifier using bootstrapping (a method of random sampling with replacement) had minimal effect, improving accuracy by 1% to 39%.

A large number of samples were again classified as belonging to integrative cluster number 4, which as noted in Figure 31 has no strong genetic profile. This occurs because the decision tree is based on yes or no decisions related to outlier-expression of particular genes (as shown in Figure 33), and if no strong features are detected the sample is classified into cluster 4. There was also a high degree of confusion between integrative clusters 3 and 8.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|----|---|----|----|----|----|----|----|----|----|
| 1 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 1 | 1 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 3 | 4 | 21 | 16 | 2 | 2 | 5 | 24 | 2 | 1 |
| 4 | 10 | 4 | 24 | 42 | 5 | 4 | 18 | 7 | 10 | 12 |
| 5 | 4 | 2 | 0 | 9 | 30 | 2 | 1 | 0 | 3 | 0 |
| 6 | 1 | 1 | 0 | 5 | 0 | 12 | 1 | 1 | 5 | 2 |
| 7 | 1 | 0 | 2 | 6 | 0 | 0 | 6 | 0 | 4 | 0 |
| 8 | 4 | 5 | 17 | 1 | 1 | 1 | 7 | 23 | 2 | 2 |
| 9 | 4 | 3 | 3 | 8 | 3 | 2 | 5 | 6 | 12 | 4 |
| 10 | 1 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 2 | 36 |

Table 25: Confusion matrix for all samples classified with a decision tree.

Columns represent “true” class, while rows represent classification.

One way to improve decision tree classification is a process called boosting. This iteratively re-weights features to ensure that it is theoretically possible to correctly classify previously inaccurately classified samples. With this method, we were able to obtain perfect classification before cross validation (Table 26).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|----|----|----|----|----|----|----|----|----|----|
| 1 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 67 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 96 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 41 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 43 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 63 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 41 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 57 |

Table 26: Decision tree demonstrating perfect classification using boosting.

Columns represent “true” class, while rows represent classification.

However, this method is extremely susceptible to over-fitting. One measure of this is known as alpha, which is a coefficient of the importance of each feature to the classifier based on errors. The larger the error, the smaller the alpha and vice versa. If boosting were robustly improving classification, one would expect to see alpha tending to increase with each iteration. This was not observed in our dataset, where over 300 iterations alpha remained below 1, which is relatively small.

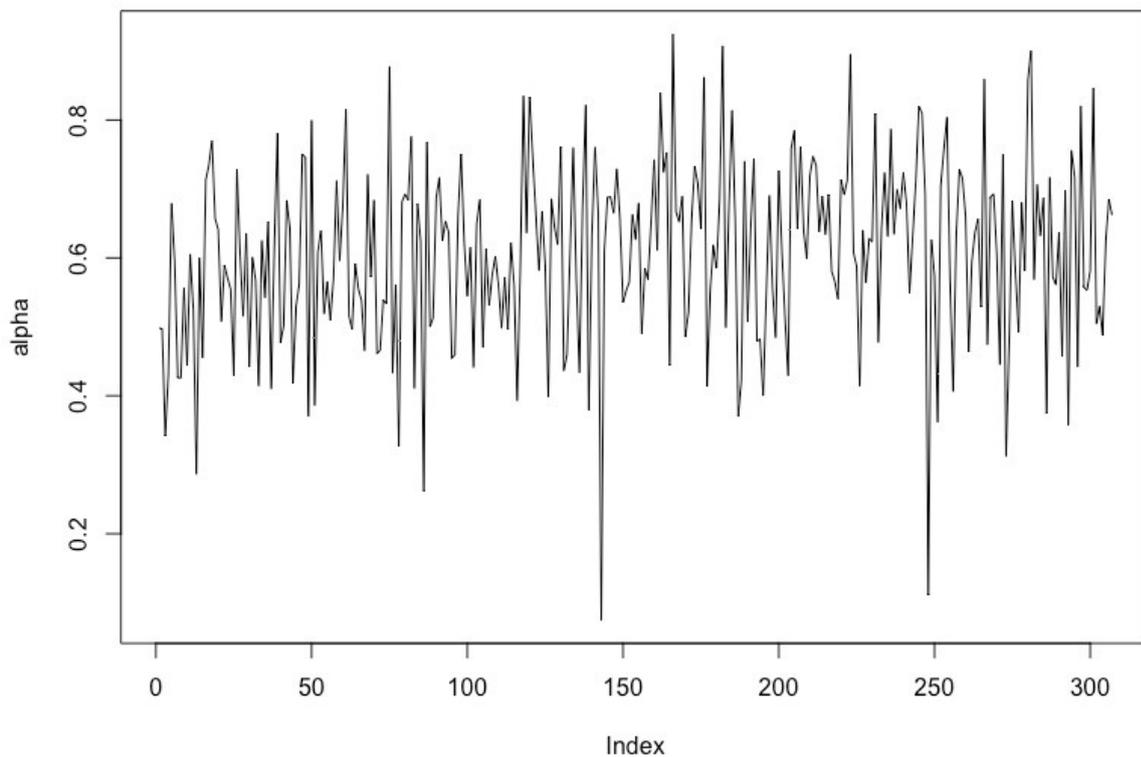


Figure 34: An example of iterative boosting for a single sample across 300 iterations. If the fit were improved then alpha would trend upwards and exceed 1. This was not observed.

Another way to assess the generalisability of the boosted model, giving an indirect measure of over-fitting, is by assessing its accuracy with leave-one-out cross validation. This confirmed that boosting resulted only in over-fitting, with overall cross-validated classification accuracy falling to 35% (Table 27).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|---|---|----|----|----|---|----|----|----|----|
| 1 | 7 | 2 | 3 | 5 | 2 | 1 | 3 | 3 | 0 | 5 |
| 2 | 1 | 3 | 1 | 3 | 0 | 1 | 0 | 1 | 2 | 1 |
| 3 | 2 | 3 | 20 | 16 | 0 | 0 | 5 | 16 | 4 | 3 |
| 4 | 7 | 2 | 21 | 33 | 12 | 6 | 12 | 13 | 8 | 13 |
| 5 | 0 | 1 | 0 | 6 | 22 | 1 | 1 | 0 | 1 | 0 |
| 6 | 0 | 0 | 2 | 3 | 0 | 9 | 2 | 1 | 2 | 2 |
| 7 | 4 | 5 | 5 | 7 | 2 | 2 | 9 | 3 | 7 | 1 |
| 8 | 3 | 1 | 10 | 8 | 2 | 0 | 3 | 23 | 2 | 1 |
| 9 | 3 | 2 | 5 | 2 | 1 | 3 | 7 | 1 | 12 | 3 |
| 10 | 2 | 1 | 0 | 13 | 0 | 0 | 1 | 2 | 3 | 28 |

Table 27: Confusion matrix for all samples classified with a boosted decision tree.

Columns represent “true” class, while rows represent classification.

Another way to improve decision tree classification that is less vulnerable to over-fitting is known as a random forest. This is a well established machine learning algorithm combines bootstrapped decision tree classification with majority voting. For each sample, many decision trees are constructed, each based on a sub-set of the remaining samples. A process called majority voting is then employed, with the sample classified a large number of times and the most frequent result outputted. This process resulted in an improvement in classification accuracy to 54.6% percent, with a kappa value of 0.475 (Table 28). Particular improvements were seen in integrative cluster 4, which now acted less like a magnet for other samples, but clusters 1 and 7 remained particularly poorly classified. Therefore this random forest classifier based entirely on our NanoString data was only marginally able to out-perform the previously published iC10 classifier, which had a classification accuracy of 50.4%.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | class.error |
|----|---|---|----|----|----|----|----|----|----|----|-------------|
| 1 | 1 | 1 | 1 | 11 | 3 | 0 | 3 | 3 | 2 | 2 | 0.9629630 |
| 2 | 1 | 5 | 3 | 4 | 2 | 1 | 2 | 0 | 2 | 0 | 0.7500000 |
| 3 | 0 | 1 | 28 | 18 | 0 | 0 | 1 | 17 | 1 | 0 | 0.5757576 |
| 4 | 1 | 0 | 12 | 56 | 7 | 3 | 5 | 1 | 4 | 7 | 0.4166667 |
| 5 | 0 | 0 | 0 | 5 | 28 | 0 | 0 | 1 | 1 | 0 | 0.2000000 |
| 6 | 0 | 0 | 0 | 4 | 0 | 17 | 0 | 1 | 1 | 0 | 0.2608696 |
| 7 | 0 | 0 | 2 | 18 | 0 | 1 | 12 | 4 | 4 | 1 | 0.7142857 |
| 8 | 1 | 0 | 14 | 5 | 0 | 0 | 0 | 38 | 3 | 0 | 0.3770492 |
| 9 | 0 | 1 | 2 | 8 | 2 | 1 | 0 | 1 | 21 | 1 | 0.4324324 |
| 10 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 1 | 2 | 36 | 0.2653061 |

Table 28: Confusion matrix for a random forest classifier, based on decision trees.

Rows represent “true” class, while columns represent classification.

3.9 Improving iC10 classification

Given that the best classifier based on our data using random forests was only marginally able to outperform the previously published iC10 classifier, we therefore revisited this and attempted to improve it. The confusion matrix displayed in Table 24 represents the baseline from which we are trying to improve, and is derived from the results of iC10 classification based on the expression of 139 genes assessed by NanoString with standard normalisation.

As a first step, automatically rejecting 43 of the 456 samples due to quality flags using NanoStringQCPro (which included those samples earlier flagged as containing poor quality RNA) improved classification accuracy from 50.4% to 57%. Subsequent analyses are based on these data.

3.9.1 Posterior probability thresholding

In iC10, a decision is made about which integrative cluster every sample belongs. However, it also outputs a posterior probability for group membership, which can be taken to be a surrogate measure of certainty. Here, as a first step, I assessed the impact of applying thresholds on the posterior probability of each sample belonging to a subtype, effectively removing samples where the algorithm felt that subtype assignment was uncertain and assessing accuracy only for those samples where the algorithm was ‘sure’.

To do this, I took 70% of the samples as a training set and determined the posterior probability cut-off point for each subtype that performed optimally in classifying samples. Optimising this against classification accuracy would bias the algorithm towards larger groups, so instead we optimised kappa. When the original iC10 classifier was applied, after rejecting samples that failed quality control, the accuracy was 57% and Kappa was 0.506.

Optimising the cut-off point to the thresholds in Table 29 resulted in iC10 classification accuracy improving from 57% to 68% and kappa improving from 0.51 to 0.64. However, this was at the significant cost of losing 52/118 (44%) samples, which were deemed uncertain.

| Integrative Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Optimal Cut-off point for Posterior Probability | 0.922 | 0.996 | 0.902 | 0.969 | 0.891 | 0.993 | 0.939 | 0.999 | 0.999 | 0.988 |

Table 29: Optimal cut-point for discarding uncertain samples to maximise kappa.

3.9.2 Combining integrative clusters 3 and 8.

I observed a high degree of classifier confusion in my data between integrative clusters 3 and 8 using both the iC₁₀ (Table 24) and random forest (Table 28) methods. Returning to the literature, I observed that their profile and prognostic value was also very similar in the original paper (Curtis *et al.*, 2012). I therefore decided to combine groups 3 and 8 and re-run the classification analysis, including the thresholding process. This improved the results very significantly, with classification accuracy up to 81%, kappa up to 0.76, and only 34/118 (28%) samples unclassified (Table 30).

| Confusion Matrix and Statistics | | | | | | | | | |
|---------------------------------|---|----|---|-----|---|---|---|---|---|
| Reference | | | | | | | | | |
| Prediction | 1 | 10 | 2 | 3/8 | 4 | 5 | 6 | 7 | 9 |
| 1 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3/8 | 0 | 0 | 0 | 24 | 0 | 0 | 0 | 1 | 0 |
| 4 | 1 | 3 | 0 | 2 | 5 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 1 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |

| Overall Statistics | | | | | | | | | |
|---------------------------------|--|--|--|--|--|--|--|--|--|
| Accuracy : 0.8082 | | | | | | | | | |
| 95% CI : (0.6992, 0.891) | | | | | | | | | |
| No Information Rate : 0.3699 | | | | | | | | | |
| P-Value [Acc > NIR] : 2.175e-14 | | | | | | | | | |
| Kappa : 0.7648 | | | | | | | | | |

Table 30: Confusion matrix and summary statistics for the iC₁₀ classifier, with posterior probability thresholding and the combination of integrative clusters 3 and 8. Absolute sample numbers rather than percentages are shown.

3.10 Discussion

While these results show proof of concept for a NanoString based semi-automated workflow using widely available FFPE tissue performing well above chance, they are not sufficient to be used as a clinical tool. Misclassification rates are high, even with data optimisation, and many samples are uncertain. This discussion will focus on identifying potential causes for this, and propose possibilities for future improvement.

3.10.1 RNA quality

The normalisation programmes (NanoStringNorm and NanoStringQCPro) both flagged up a number of samples that failed quality control. The quality control programme here is based on the positive and negative controls included in the probe set as well as housekeeping genes. Run quality, background reads, and variation between samples are all taken into account and normalised in subsequent analysis. There was no significant correlation between these samples and their concentration or RIN. However, RIN is not always the best measure of RNA quality for downstream analysis (Wimmer *et al.*, 2018). Formalin fixation can induce RNA crosslinking that can vary significantly between samples without affecting RIN (Srinivasan *et al.*, 2002). Formalin preferentially affects bases A and C in comparison to G and U, so it is possible that the fixation process affected the genes of interest disproportionately. Evidence for this comes from Table 22, which shows large differences in the correlation of gene expression for individual genes between NanoString (FFPE) and Illumina (Fresh Frozen) platforms.

Overall, these issues may account for the improvement in iC10 classification when flagged samples were excluded, despite them showing acceptable RIN and RNA purity in all but one case. More importantly, it may be that there are further gene expression imbalances induced by formalin fixation, to which our quality control methods at RNA and NanoString platform level are insensitive.

3.10.2 Probe positions

The position of an RNA probe can impact the measured concentration if RNA is degraded, as this preferentially occurs towards the 5' end (Opitz *et al.*, 2010). I therefore evaluated how the different positions of the NanoString and Illumina probes along transcripts could alter results.

Figure 35 shows the distance of the probe from the 3' end as a percentage of the transcript length for Illumina (x axis) and NanoString (y axis) probes. It shows that NanoString probes are spread all along the transcript, whereas Illumina probes are concentrated at the 3' end.

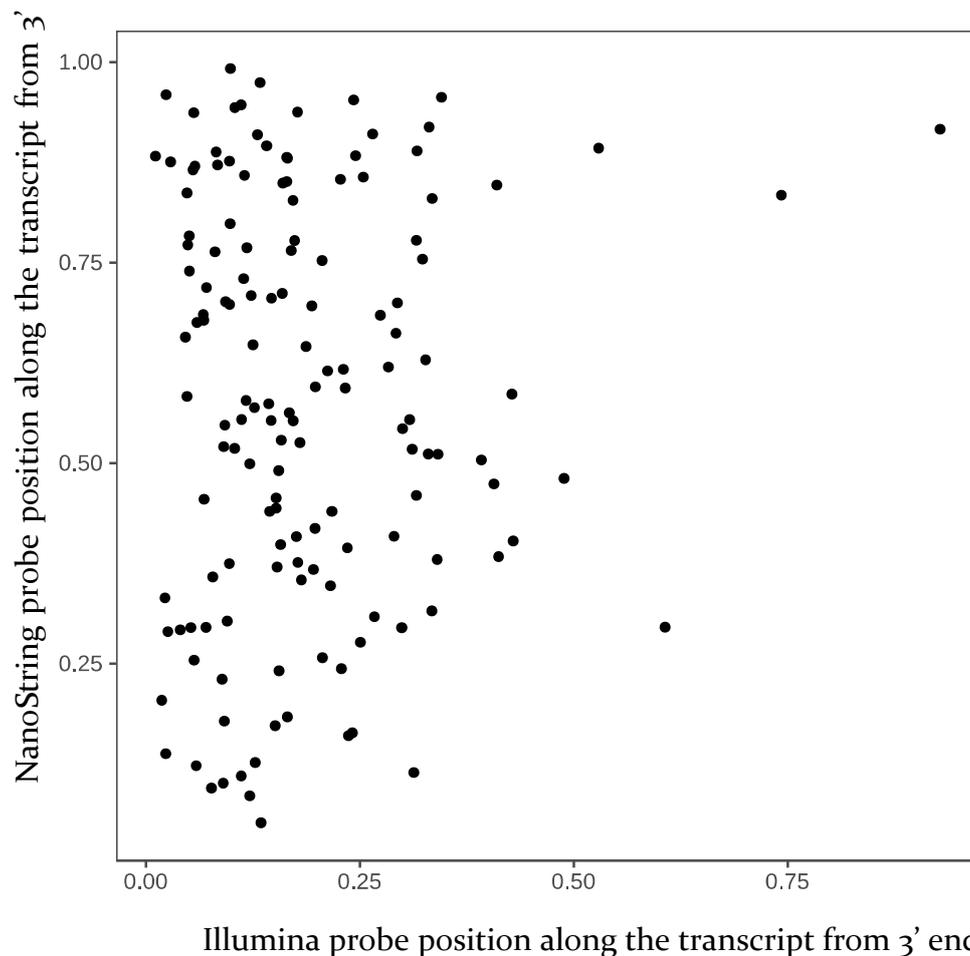


Figure 35: Distance of the probe sequence from the 3' end of the mRNA for NanoString (y-axis) and Illumina (x-axis) platforms.

Most of the Illumina probes were within the first third of the transcript, while NanoString probes were evenly distributed along the mRNA.

3.10.3 Probe selection

For reasons of cost and practicality, it was decided to perform the NanoString technique only on 207 pre-selected genes, rather than all 754 in the original integrative clusters paper. I examined whether this was the cause of the iC10 platform's poor performance, or whether it was instead due to differences in RNA expression between FFPE and fresh-frozen tissue types.

As a first step, I ran the iC10 classifier on the same samples using Illumina data from fresh frozen tissue, but only for the 139 genes that were present in the NanoString panel. Cross-validated classification accuracy was excellent, at 85.5%. However, these samples were all used for the original training of the classifier so this value is vulnerable to overfitting. I therefore repeated this procedure on Illumina data from all 1980 samples in the METABRIC dataset. Cross-validation classification accuracy was still very high, at 83.4%. Documented optimal classification using iC10 with all genes is 94.7% (Ali *et al.*, 2014). Therefore, for this tissue type and analysis platform, sub-selecting 139 of the 612 genes expected by the classifier imparted a penalty of approximately 10% in classification accuracy, thereby tripling the percentage of misclassified samples from 5.3% to 16.6%. However, it remained far superior to iC10 classification based on NanoString data, which returned classification accuracies of 50.4% or 57% depending on the stringency with which samples were rejected by quality control. Probe selection is therefore not sufficient as an explanation for the poor classification performance with my data.

3.10.4 Gene expression correlation

As shown in Figure 24 and Table 22, individual genes varied markedly in the degree to which their expression correlated between NanoString (FFPE) and Illumina (Fresh Frozen) platforms. I therefore examined whether further sub-selecting only those genes that were highly correlated across platforms would close the gap between iC10 classification performances across these tissue preparations and analysis techniques.

I re-ran the classifier using expression data from both platforms, using only the 25 or 50 most correlated genes across platform (Table 31).

| Number of genes | Illumina | NanoString |
|-----------------|----------|------------|
| 25 | 0.649 | 0.441 |
| 50 | 0.726 | 0.445 |

Table 31: Performance of the iC10 classifier with only genes that were highly correlated across platform.

The Illumina-based iC10 classification took a heavy accuracy penalty from sub-setting in this way, falling from 85.5% to 64.9% with only 25 genes, and 72.6% with 50 genes. The NanoString classifier showed a much smaller drop in performance from 50.4% with the whole dataset to 44.1% with 25 genes and 44.5% with 50 genes. It is particularly notable how little improvement there was in increasing from 25 to 50 genes. This suggests that those genes that were weakly correlated across platforms are not contributing much information to iC10 classification.

Next, I investigated whether the NanoString RSF1 probe was simply ineffective, by comparing the expression of RSF1 to CLNS1A, which is located very close on chromosome 11q and tends to be co-expressed. Expression of these two genes was in fact more strongly correlated on the NanoString than Illumina platform, indicating that the RSF1 probe is indeed quantifying something related to chromosome 11q and this cannot be the explanation for poor correlation across platforms.

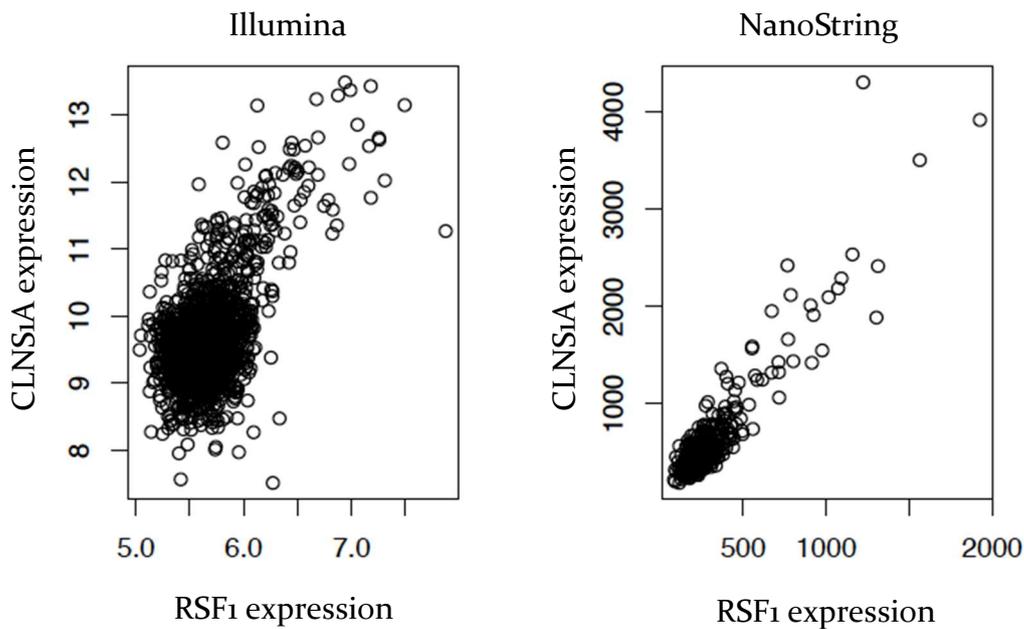


Figure 37: Correlation between CLNS1A and RSF1 expression for Illumina and NanoString platforms.

3.10.6 Effect of tissue type

The bulk of this analysis is comparing gene expression across workflows that differ in both tissue processing (FFPE vs fresh frozen) and platform (NanoString vs Illumina). To assess the impact of tissue processing in isolation, I chose 24 ‘difficult’ samples evenly distributed across integrative clusters, that were classified at chance (8.3% accuracy) by the iC10 algorithm on NanoString FFPE data, and for which RNA extracted from frozen tissue was also available. These samples were also less well classified than average by the iC10 algorithm on the Illumina data (79.2% accuracy).

For each of these samples I processed the RNA from fresh frozen tissue on the NanoString platform. Firstly, I examined the correlation in gene expression across platforms (Illumina vs NanoString) on the same (Fresh Frozen) tissue type. The median within-patient, across-gene correlation was 0.65 (range 0.54-0.74, IQR 0.61-0.68). At first glance this seems strikingly similar to the median correlation that was obtained between the NanoString (FFPE) and Illumina (Fresh Frozen) methods for all samples (median 0.66, Figure 23), however these ‘difficult’ samples showed a poorer correlation on that comparison (median rho 0.57).

The median within-gene, across-sample correlation comparing Illumina and NanoString on the same (frozen) tissue type was 0.60 (range 0.07-0.95, IQR 0.48-0.75), again very similar to that obtained across all samples between the NanoString (FFPE) and Illumina (Fresh Frozen) methods (median 0.56 Figure 24), but again superior to the same comparison for only these 24 samples (median rho 0.33).

Next, I examined the correlation in gene expression across tissues on the same (NanoString) platform. The median within-patient, across-gene correlation was 0.80

(range 0.62-0.93, IQR 0.73-0.84). The median within-gene, across-sample correlation was 0.47 (range -0.05-0.81, IQR 0.36-0.58).

Overall, therefore, tissue type is having a particularly large influence on the expression of certain genes, whereas the choice of platform is having a larger effect on the overall profile of gene expression for each individual.

Finally, I re-ran the iC10 classifier on the 24 difficult samples using the NanoString (fresh frozen) data. Classification accuracy improved from 8.3% to 33.3%, but was still well below performance based on the Illumina (fresh frozen) data for which the iC10 classifier was designed.

3.11 Conclusion:

Overall, I conclude that:

1. The NanoString platform based on FFPE samples showed good agreement with IHC and Illumina microarray in determining ER and HER2 status.
2. However, it was not possible to completely distinguish 10 integrative cluster groups based on the current NanoString data to a clinically useful degree using either the previously published iC10 algorithm or a bespoke random forest classifier.
3. After optimisation, iC10 was superior to the random forest classifier, despite it having been designed for a larger number of genes assessed on a different tissue type on a different platform.
4. Using iC10, a significant proportion of the samples were below the threshold for being determined confidently, and the groups that are similar in terms of gene expression (3 and 8) were particularly poorly distinguished by both iC10 and random forest methods.
5. Poor classification was not solely due to gene selection, as using the same genes based on Illumina data the iC10 package can predict cluster membership with high accuracy.
6. One of the primary problems may be poor correlation of the expression of key genes across tissue types and platforms, with RSF1 being a key example.
7. It is likely that tissue preparation method (FFPE vs fresh frozen) is the key driver of this poor correlation of particular genes, while the choice of platform had a larger effect on the overall profile of gene expressions for each individual.

To examine the effect of tissue preparation method more closely, independent of platform and without restriction to a pre-defined probe set, the next chapter will examine gene expression profiles using RNA sequencing on paired samples of FFPE and fresh frozen tissue from the same patients.

4. RNA Sequencing

4.1 Preface

Following on from our comparison of NanoString (FFPE) and Illumina (Fresh Frozen) based classifiers for integrative clusters, this chapter examines the effect of tissue preparation method more closely, independent of platform. I employed RNA sequencing (RNA-seq), both to avoid biasing my analyses towards either of these platforms, and also because it does not rely on specific probes. This potentially means that a more comprehensive analysis can be performed without pre-specification of genes of interest, which is important given the future goal of defining an optimal panel of genes for assessment of integrative clusters from FFPE tissue.

I am grateful to Dr Stephen John Sammut for writing the pipeline to run these data through the GATK Haplotypecaller algorithm to determine sample haplotype and therefore their identity to validate their pairings. I am also grateful to Raquel Manzano Garcia for aligning the transcriptome using *Salmon* (Patro *et al.*, 2017) and *STAR* (Dobin *et al.*, 2013). I undertook all of the other procedures and analyses myself, except where I have specifically stated otherwise in the text.

4.2 Background

The abnormal biology of tumour cells is ultimately driven in large part by differences in protein expression from healthy cells. In recent years, our ability to measure this with proteomics has expanded with the development of technologies such as mass cytometry, but it is well recognised that it is limited by the lack of a comprehensive database of normal expression (Dupree *et al.*, 2020). Wilhelm *et al.* (2014) showed that the primary determinant of the quantity of protein in a given cell is regulation of mRNA levels. Therefore, assessing the expression level of mRNA is a surrogate for protein levels. Using mRNA rather than DNA means that the effect of modulators such as promoters can be detected.

The advent of next-generation sequencing technologies in the mid 2000s allowed for the development of methods to sequence RNA (as well as DNA) at large scale, known as RNA-seq (Wang *et al.*, 2009). It is most often used for the analysis of differential gene expression (Costa-Silva *et al.*, 2017). In this method, RNA is isolated from tissue, and ribosomal RNA (rRNA) is then removed. The remaining RNA is then reverse transcribed to cDNA, which is then amplified, making up the “library” (Conesa *et al.*, 2016). The library is usually barcoded to allow for multiplexed sequencing (Craig *et al.*, 2008), which significantly streamlines the workflow by allowing multiple samples to be sequenced simultaneously, with explicit labelling. The result is then analysed in several computational steps, including aligning the sequencing reads to a reference transcriptome, quantifying reads, normalising between samples, and then identifying any statistically significant changes in the expression level of the transcripts of interest.

Before RNA-seq became an established and validated method, gene expression analysis was performed using hybridisation techniques such as Illumina microarray or NanoString, which relied on pre-defined probes for the genes of interest. As such it is not always able to detect novel changes in the transcripts including indels, gene fusions and single nucleotide variations (Maher *et al.*, 2009; Mantione *et al.*, 2014). RNA-seq is also better suited than hybridisation-based techniques to assessing genes with very low or very high expression level. Hybridisation-based techniques can lose low expression

signals to background noise, and may saturate at high expression levels. In contrast, RNA-seq can adjust its sensitivity by altering sequencing depth. The other, and here perhaps most important, advantage of RNA-seq over microarray and NanoString is that the result is not limited to genes of interest that are specified in advance by specific probes.

However, a traditional drawback of using RNA-seq for archival material is that, until the mid 2010s, good quality RNA was needed to ensure even gene coverage, reduce false positive differential expression rate, and avoid high duplication rate. Since then, library preparation methods have improved, and in particular several different methods have been developed to reduce the loss of degraded and fragmented mRNAs (Stark *et al.*, 2019). This has allowed RNA extracted from FFPE samples, which are of lower quality and generally degraded, to be analysed using RNA-seq.

The starting aim of this chapter was to use RNA-seq to find genes that were differentially expressed across the ten integrative clusters, but not between the FFPE and fresh-frozen preparations of the same chemotherapy-naïve resection specimens.

4.3 Pilot study

Before proceeding to sequence a larger dataset I assessed the suitability of the method on 3 samples from the METABRIC study on the basis of correct classification with NanoString and a reasonable RNA concentration in both FFPE and fresh frozen tissues. For each, RNA was extracted from both FFPE and fresh frozen tissue and analysed as a pair. As a first step, the genes that comprise the iC10 classifier were correlated. For every sample, after excluding those genes for which expression could not be detected, the remainder correlated very strongly within every pair (all $r(>450) \geq 0.88$, $p < 2.2 \times 10^{-16}$) (Figure 38).

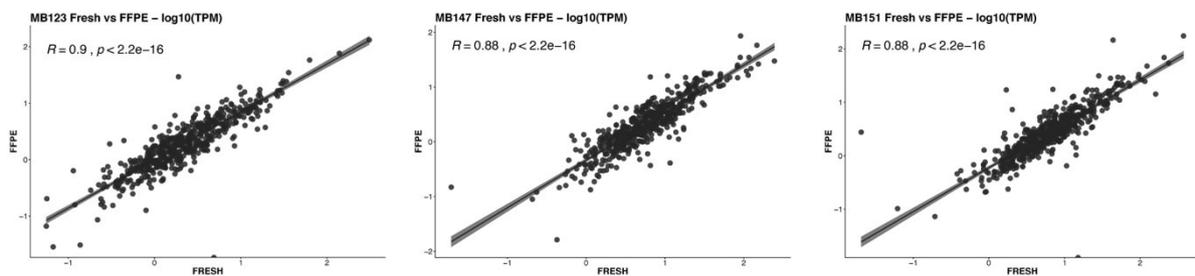


Figure 38: Correlation of gene expression in the iC10 classifier across FFPE and Fresh Frozen tissues quantified using RNA-seq (log scales).

4.4 Larger scale comparison

Following the success of the pilot samples, 48 further pairs of samples were chosen from the METABRIC study to include all ten integrative clusters, weighted towards those that were more difficult to classify using the NanoString (FFPE) technique (Table 32). 48 pairs of samples were chosen, as the maximum number of samples that could be run on our equipment was 96 and we could therefore avoid the confound of batch effects. Again, RNA was extracted from both FFPE and fresh frozen tissue for every sample.

| Integrative Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| Number of samples | 7 | 5 | 5 | 6 | 4 | 2 | 5 | 3 | 5 | 6 |

Table 32: Number of samples sequenced from each integrative cluster

4.4.1 RNA

As a baseline descriptive step, I assessed the quality of mRNA extracted from samples with each tissue preparation method, in terms of RNA concentration, mRNA/cDNA concentration, and total number of reads aligned.

The median concentration of RNA extracted from fresh frozen samples was 155.7 ng/ml (49 – 502), while that from FFPE tissue was 137.6 ng/ml (13.2 – 398.8) (Figure 39). There was no significant difference in RNA concentration between the two tissue preparation techniques (Wilcoxon rank-sum $W(48,48)=1288$, $p=0.321$).

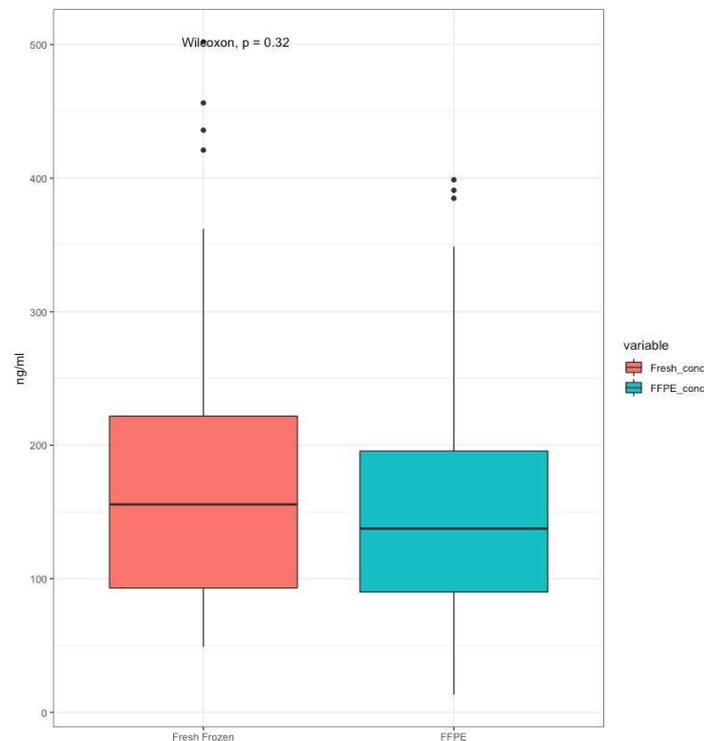


Figure 39: RNA concentration (in ng/ml) extracted from fresh frozen and FFPE tissue preparations.

RNA concentration was significantly higher from fresh frozen tissue (Wilcoxon rank-sum $W(48,48) = 1288$, $p = 0.321$).

4.4.2 cDNA

However, the concentration of cDNA constructed from this RNA differed very significantly between the fresh frozen and the FFPE samples (Figure 40), presumably because the RNA in the FFPE sample was much more degraded. This was despite using three times the quantity of RNA from FFPE (300ng vs 100ng) to compensate for degradation. cDNA extracted from fresh frozen tissue had a median concentration of 20.5 nM/ml (0 - 56.7), while FFPE samples had a median of 10.5 nM/ml (0 - 28.2) (Wilcoxon rank-sum $W(48,48) = 1653$, $p = 1.89 \times 10^{-4}$). One fresh frozen and three FFPE samples yielded extremely low cDNA concentrations, less than 0.001 ng/ml – these samples were excluded from future analysis. As the FFPE samples were all run in the same batch, this likely reflects extremely poor-quality RNA in those samples rather than an error in the processing pipeline.

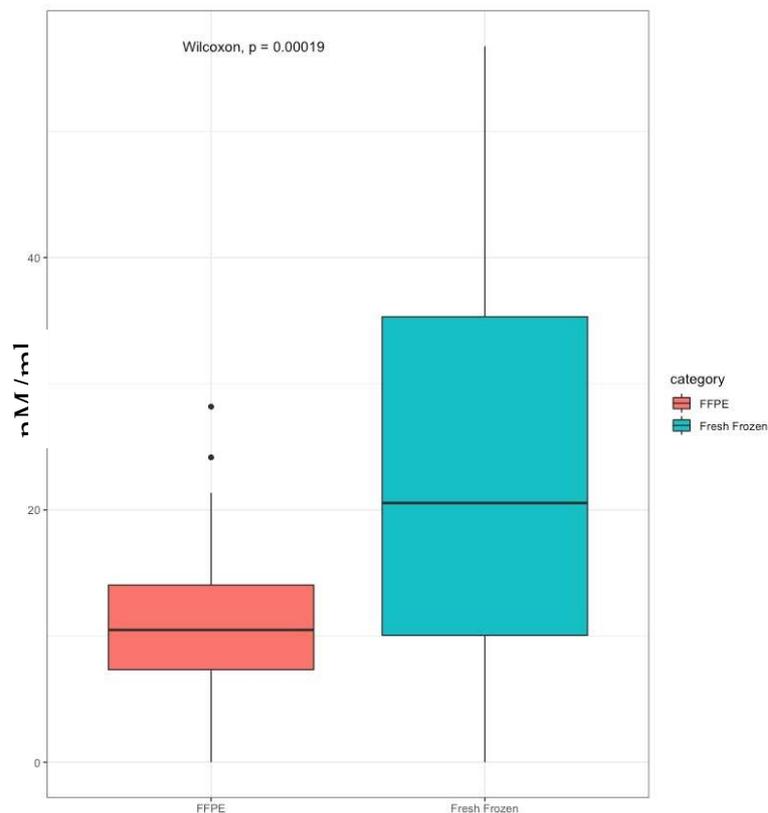


Figure 40: cDNA concentration (in nM/ml) extracted from fresh frozen and FFPE tissue types.

4.4.3 Total number of reads aligned

The total number of reads aligned is the number of transcripts that is obtained from sequencing the cDNA library, with a higher number reflecting a greater “depth” of read, normally suggesting better gene coverage.

The samples were aligned using *Salmon* version 0.13.1 (Patro *et al.*, 2017) and *STAR* version 2.7.3a (Dobin *et al.*, 2013), both of which output a total number of reads as their primary outcome measure.

Read number was significantly higher in fresh frozen samples than in FFPE samples with both alignment methods (Salmon Wilcoxon rank-sum $W(47,45) = 1739$, $p = 2.27 \times 10^{-8}$, STAR Wilcoxon rank-sum $W(47,45) = 1914$, $p = 1.93 \times 10^{-13}$) (Figure 41). The median number of reads from fresh frozen samples was 7.32×10^6 (range $2.38 \times 10^6 - 3.92 \times 10^7$), while from FFPE it was 3.72×10^6 (range $455 - 2.49 \times 10^7$).

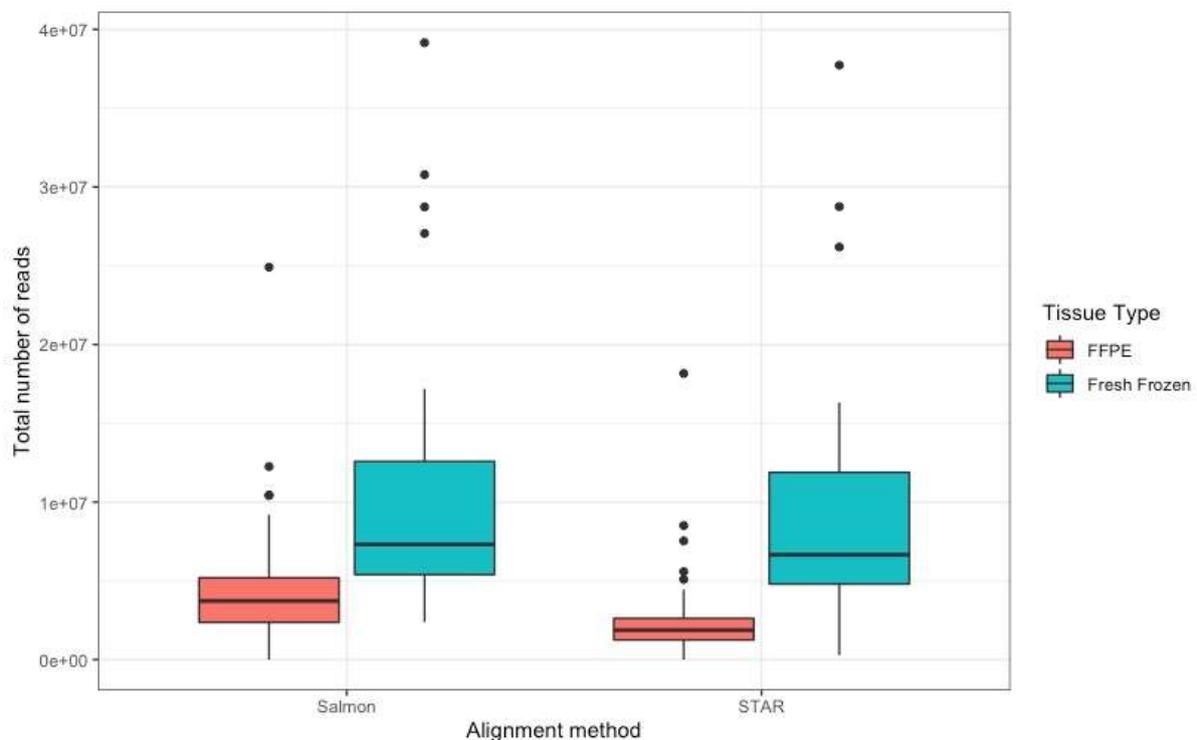


Figure 41: Total number of reads in fresh frozen and FFPE tissue types by alignment method.

Read number was significantly higher in fresh frozen samples than in FFPE samples with both alignment methods (Salmon Wilcoxon rank-sum $W(47,45) = 1739$, $p = 2.27 \times 10^{-8}$, STAR Wilcoxon rank-sum $W(47,45) = 1914$, $p = 1.93 \times 10^{-13}$).

The output of *STAR* was never higher than that of *Salmon* (Figure 42). More importantly, the output of these two methods appeared much more strongly correlated for fresh frozen tissue samples ($\rho_{(45)} = 0.985$, $p < 2.2 \times 10^{-16}$) than for FFPE samples ($\rho_{(43)} = 0.603$, $p = 1.76 \times 10^{-5}$). Using parametric linear regression modelling confirmed a difference in regression slope (interaction) on the basis of tissue type (main effect of read number $t(88) = 10.3$, $p < 2.2 \times 10^{-16}$, main effect of tissue type $t(88) = 0.613$, $p = 0.542$, interaction between read number and tissue type $t(88) = 4.18$, $p = 6.84 \times 10^{-5}$).

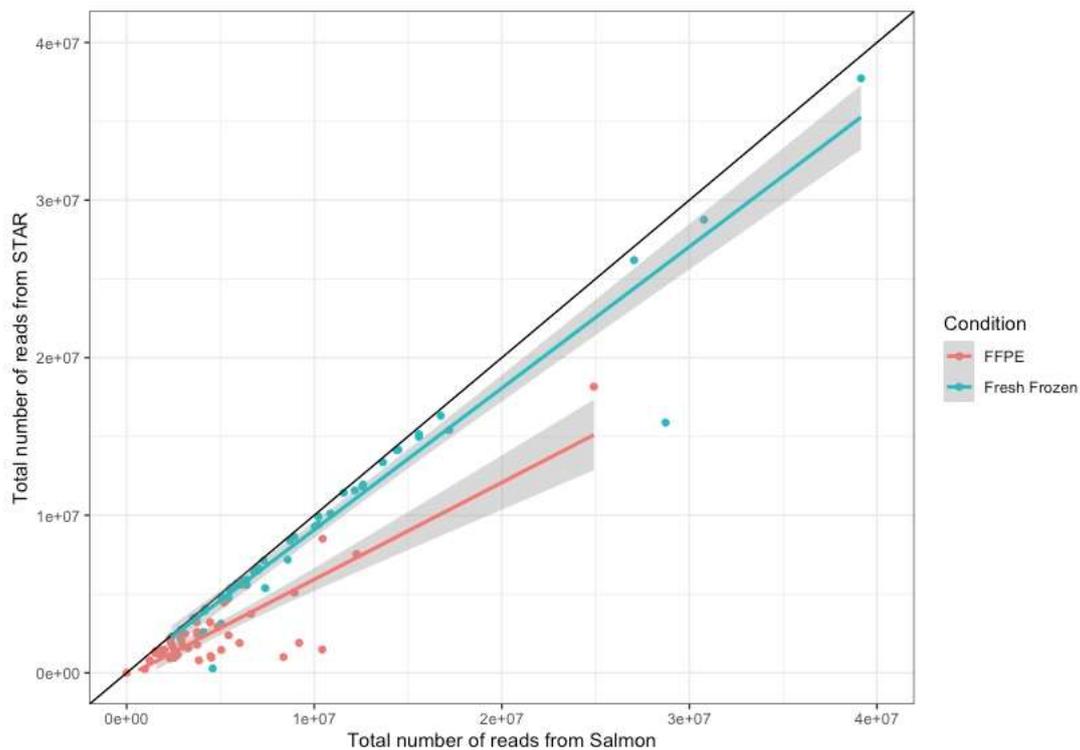


Figure 42: Correlation of total number of reads obtained with STAR and Salmon alignment methods.

The two methods were more strongly correlated for fresh frozen tissue samples ($\rho_{(45)} = 0.985$, $p < 2.2 \times 10^{-16}$) than for FFPE samples ($\rho_{(43)} = 0.603$, $p = 1.76 \times 10^{-5}$). Interaction between read number and tissue type $t(88) = 4.18$, $p = 6.84 \times 10^{-5}$.

To ensure the robustness of my normalisation procedures, I assessed the relationship between sequencing effectiveness and RNA/cDNA concentration. While RNA concentration differed between samples, I used the same total amount to make each library, so I would hope not to find relationships between these measures.

There was no correlation between the total number of reads and RNA concentration with all samples combined (Salmon $\rho(90) = 0.011$, $p=0.920$, STAR $\rho(90) = 0.143$, $p=0.174$) (Figure 43). Nor was there a correlation between the total number of reads and cDNA concentration (Salmon $\rho(90) = -0.119$, $p=0.260$, STAR $\rho(90) = 0.050$, $p=0.635$) (Figure 44). One sample with very low cDNA concentration also produced a negligible number of reads, but others produced an adequate or high number of reads.

Overall, this agrees with the prevailing view that RNA and cDNA concentrations are poor measures of sequencing effectiveness alone, and that cDNA concentration is not the primary driver of total number of reads (Marioni *et al.*, 2008).

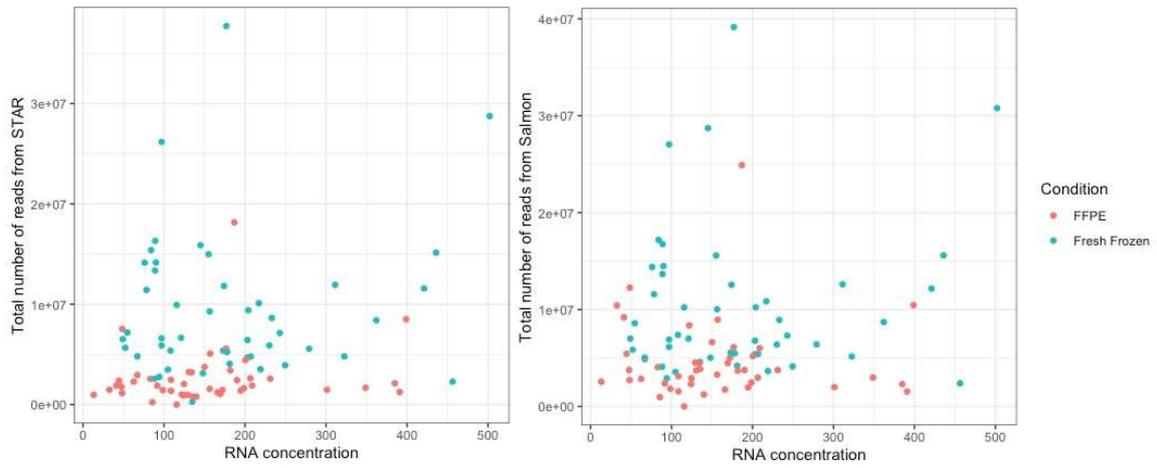


Figure 43: No correlation between total number of reads and RNA concentration

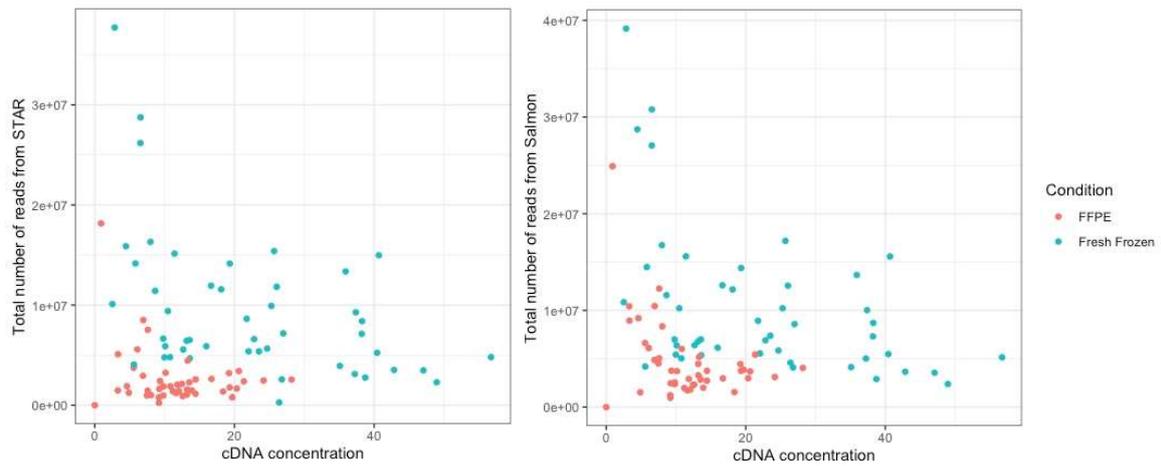


Figure 44: No correlation between total number of reads and cDNA concentration

4.4.4 Correlation of expression within-patient across gene

Samples with fewer than 1 million reads with either Salmon or STAR were removed as poor quality, along with their matched pairs. After removing the poor-quality samples, 44 pairs remained. For further analysis we decided to use the output of Salmon rather than STAR, because it reported a smaller difference in number of reads between FFPE and fresh frozen tissue (Figure 41), and generally gave higher read numbers, especially for FFPE (Figure 42). However, for haplotyping STAR was used because GATK Haplotypecaller (Poplin *et al.*, 2017) is incompatible with data output from Salmon.

As with the comparison of NanoString and Illumina, I began by examining how well gene expression correlated between fresh frozen and FFPE in each sample pair across all genes. Again, for every sample pair I ranked the relative expression of every gene in each tissue type to produce two expression profiles. Unlike in the NanoString method, where our panel contained only 207 genes, here we aligned RNA from 60105 transcripts. This is potentially problematic if large numbers of genes are not expressed in our tissue, as this will falsely inflate our measures of correlation. Therefore, any gene that was not detected in either paired sample was excluded from the analysis.

Order agreement for the number of reads for all genes in between tissue types was strong, with a median rho of 0.834 (IQR 0.770-0.873) (Figure 45). This was much stronger than observed between the NanoString and Illumina platforms (median rho 0.660), and very similar to that observed between tissue types with NanoString (median rho 0.80).

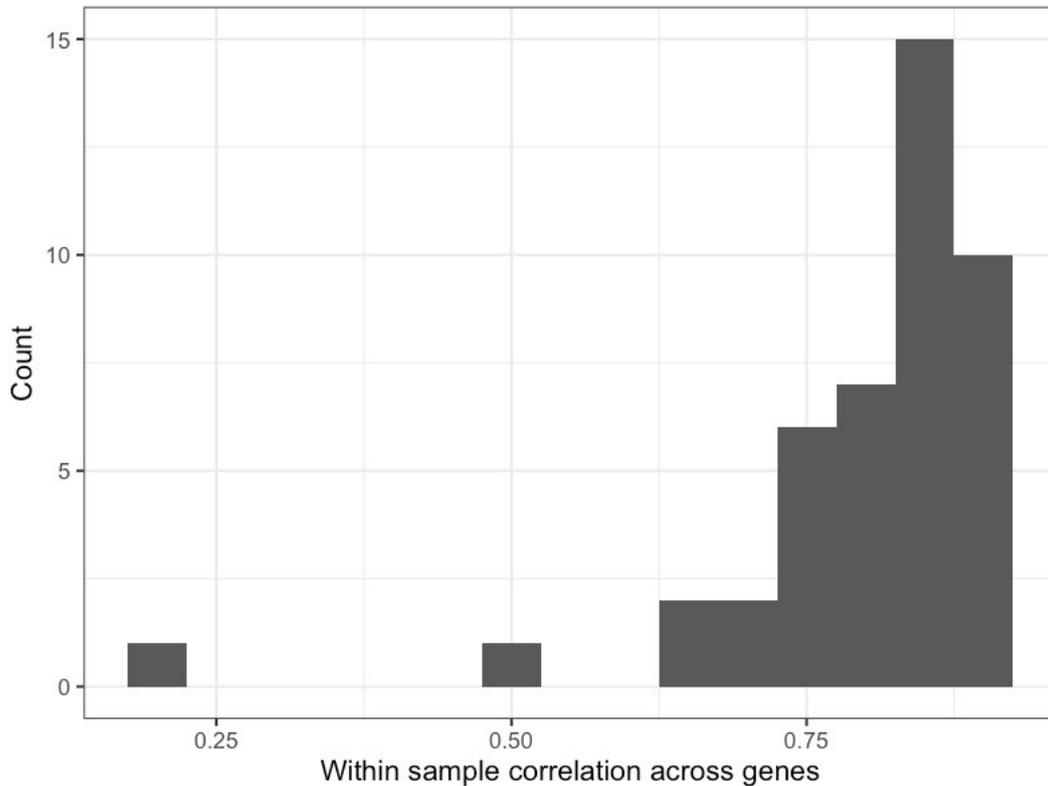


Figure 45: Spearman's rank correlation between gene expression orders for FFPE and fresh frozen tissue for every sample individually.

To assess whether RNA quality was a driver of the strength of agreement between the detected level of gene expression across sample preparation methods, I examined the relationship between the number of aligned transcripts and the gene expression correlation between sample pairs.

The best correlation was observed in a sample with 28437 genes detected in both tissue types ($\rho(28435) = 0.916$, $p < 2.2 \times 10^{-16}$). In contrast, the worst performing sample detected only 156 genes in common across the two samples, and correlated extremely poorly ($\rho(154) = 0.204$, $p = 0.01$). While this single poor sample was an outlier at the lower end, across all samples, there was a very strong correlation between the number of genes detected in both samples and the strength of correlation in expression profiles between FFPE and fresh frozen tissues ($\rho(42) = 0.817$, $p < 2.2 \times 10^{-16}$) (Figure 46). This is likely to mean that the primary driver of how well gene expression agrees in FFPE and fresh frozen tissues is RNA quality.

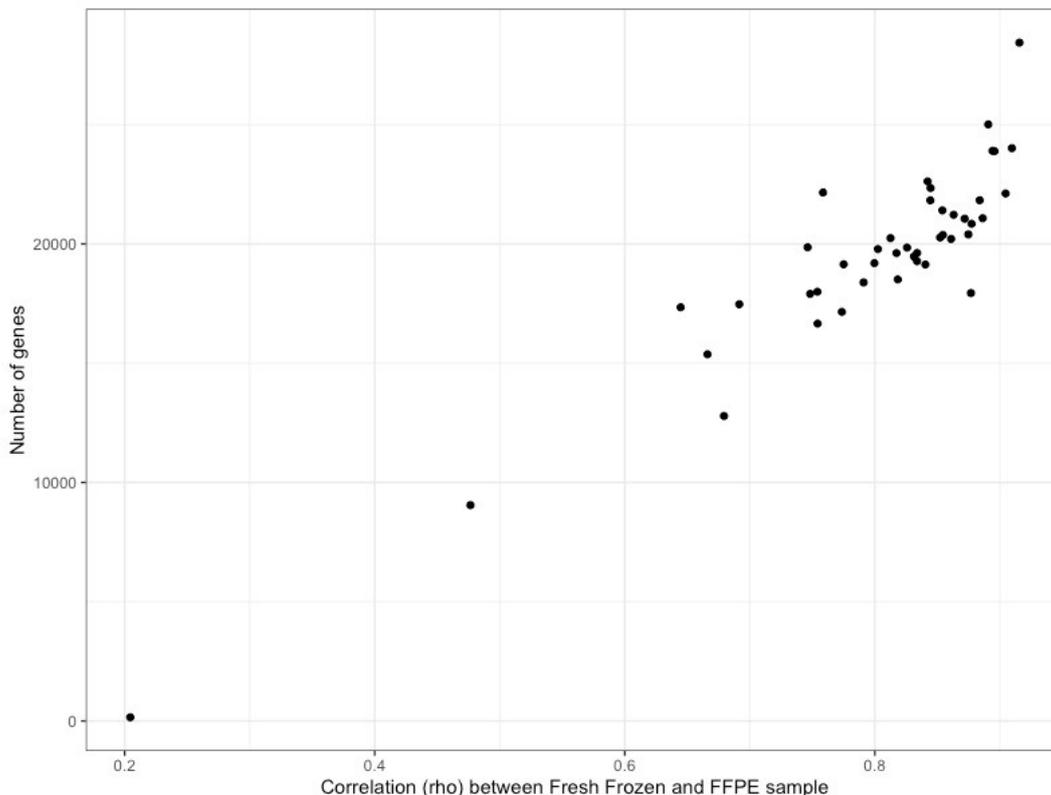


Figure 46: Correlation in gene expression profile against number of genes detected in both fresh frozen and FFPE samples.

Note that ρ , the non-parametric correlation between the level of gene expression in the samples, is not dependent on the number of comparisons being undertaken.

To assess whether this effect was primarily due to globally poor RNA quality in some biopsies, or whether it was specific to FFPE or fresh frozen tissue, I plotted the number of genes detected in each sample for both methods (Figure 47). There was no global correlation between the number of genes detected in a sample between tissue types (Spearman's $\rho(42) = 0.186$, $p = 0.225$), suggesting that the effect was not due to globally poor RNA quality in some biopsies. Instead, the centre of mass of the points was above the line of unity, with significantly more genes being detected in fresh frozen tissue than FFPE for matched samples (Wilcoxon rank sum $W(44,44) = 1659$ $p = 8.29 \times 10^{-9}$).

Overall, FFPE samples that had good quality RNA, where the expression of a large number of genes could be detected, correlated better with fresh frozen tissue than those where more gene expression products were lost.

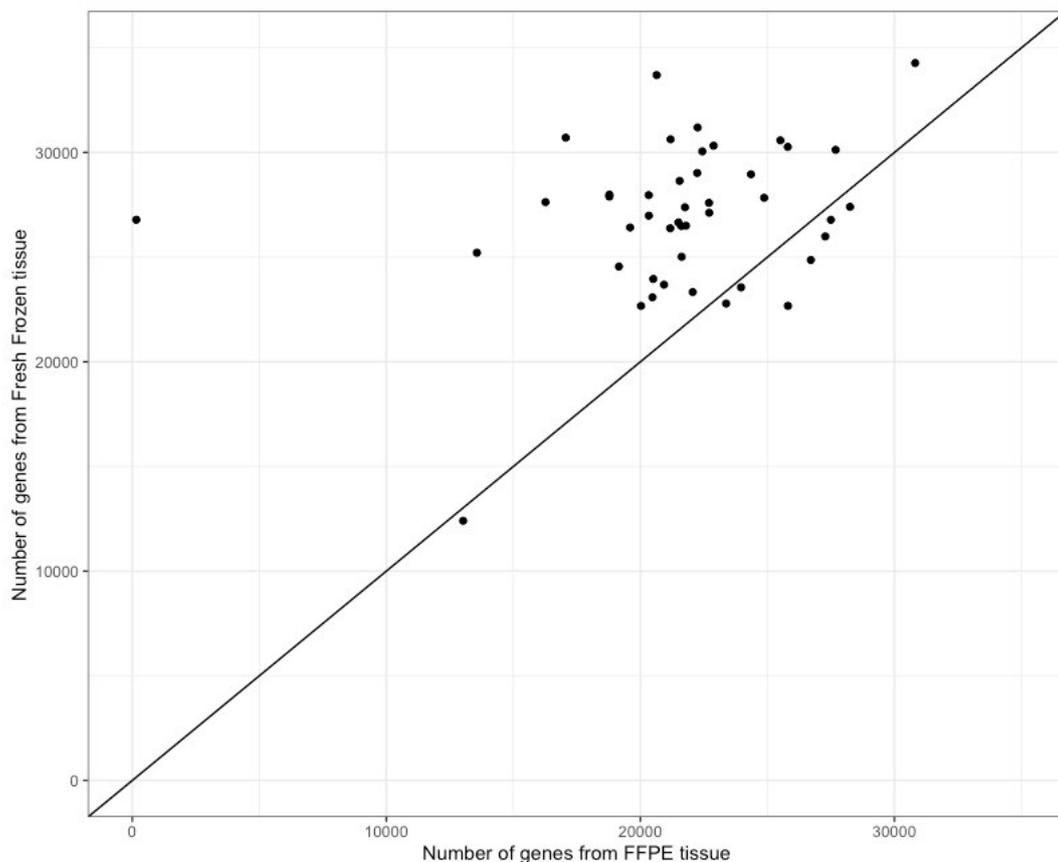


Figure 47: The number of gene expressions detected in each sample, by sample preparation method.

4.4.5 Ubiquity of gene expression

For further analyses, I first removed the two outliers with the smallest number of genes in common, which were also the two samples with the weakest correlation in gene expression profiles (both $\rho < 0.5$ and number of genes < 10000).

It is only possible to assess the correlation of expression across-samples for genes that are well represented in a sufficiently large number of samples. Therefore, as a first step, for the remaining 84 samples (42 pairs), I calculated and plotted how many genes were in common amongst at least a given number of samples (Figure 48). The majority of transcripts were detected in only one or two samples, and therefore cannot be assessed. However, more positively, the expression level of 8040 genes could be assessed in every sample.

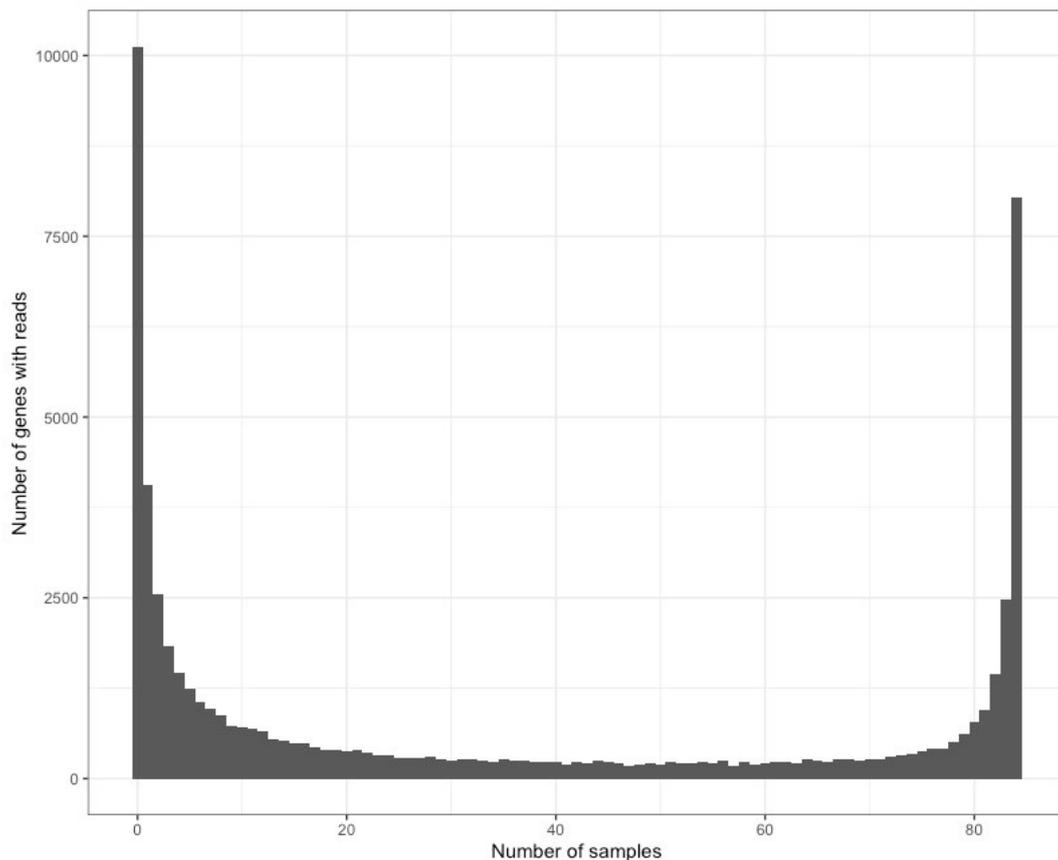


Figure 48: Number of genes detected in common for a given number of samples.

Restricting the within-patient, across gene-correlations previously plotted (Figure 45) to these 8040 genes that were ubiquitously detected in all samples slightly reduced correlation strength overall (median rho 0.757, IQR 0.687-0.788) (Figure 49). This indicates that, while these genes are robustly detected across all samples, they are not necessarily those that are most differentially expressed between tumours.

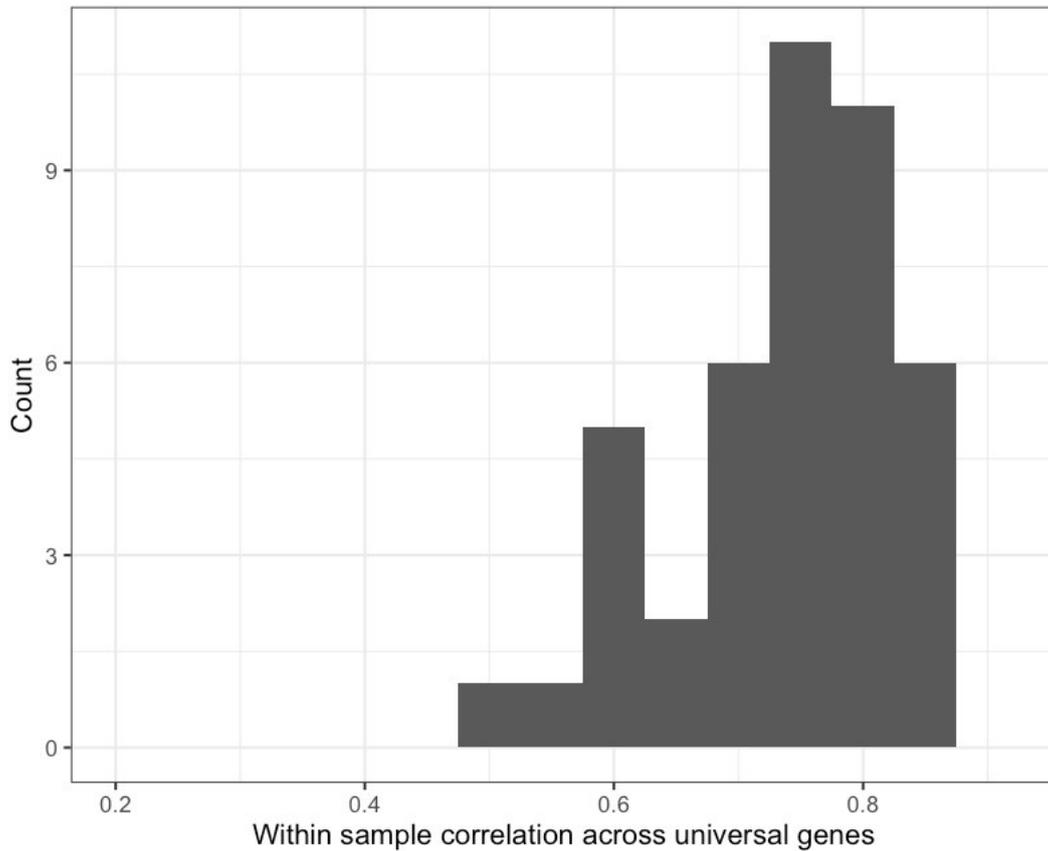


Figure 49: Spearman's rank correlation between gene expression orders for FFPE and fresh frozen tissue for every sample individually, restricted to the 8040 universally detected genes.

However, the number of reads aligned may not be the most appropriate measure for comparing across samples because it is influenced by sequencing depth (Wagner *et al.*, 2012). In other words, a lower number of reads in one sample than another (here in FFPE vs fresh frozen within any pair) may reflect a global reduction in the number of transcripts mapped in that sample, rather than a specific downregulation of the gene of interest. To account for this, the measure transcripts per million (TPM) has been proposed, which normalises the number of reads for each gene by dividing by the length of the gene (to get RPKM), then dividing again by the total RPKM for all genes in the sample, then multiplying by a million. There are pros and cons to this normalisation, and it may not work for all samples (Zhao *et al.*, 2020), and indeed it made little difference to the within-patient across-gene correlations shown in Figure 49., increasing median rho very slightly to 0.78 (IQR 0.75-0.82) (Figure 50).

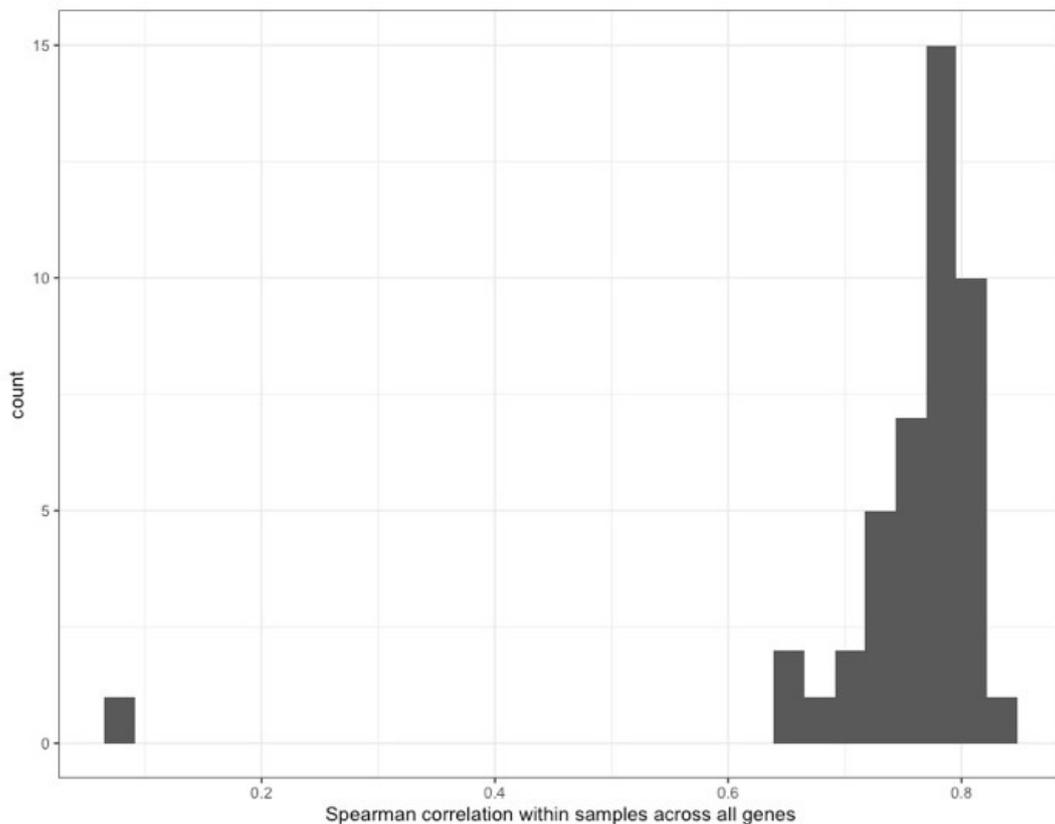


Figure 50: Spearman's rank correlation between TPM sample expression orders for FFPE and fresh frozen tissue for every gene individually, for all genes present in at least 15 sample pairs.

4.4.6 Within-gene across-sample correlations

I then assessed the within-gene, across-sample correlation for the 8040 universally detected genes (Figure 51). Correlations in gene expression across tissue preparation techniques obtained with RNAseq (median $\rho = 0.36$, range -0.13 - 0.80 , IQR 0.28 - 0.44), were poorer than those obtained with NanoString in the previous chapter (median $\rho = 0.47$, range -0.05 - 0.81 , IQR 0.36 - 0.58).

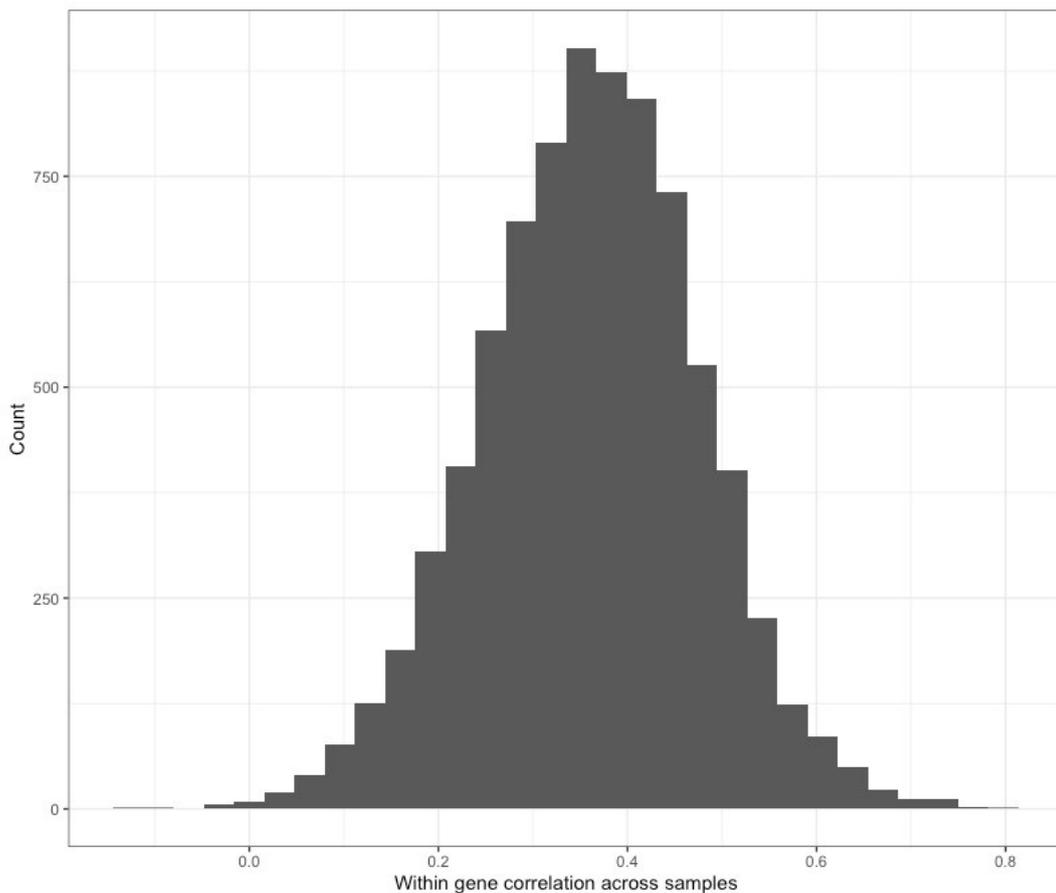


Figure 51: Spearman's rank correlation between sample expression orders for FFPE and fresh frozen tissue for every gene individually, restricted to the 8040 universally detected genes.

I considered whether this poor performance might be because I had biased my analysis towards universally expressed housekeeper genes by being too stringent in only considering those genes that were universally detected. I therefore repeated the analysis for all genes detected in at least 15 of the 42 pairs (Figure 52). This included key genes known to drive integrative cluster classification, such as *GATA3*, *RSF1* and *ZNF703*. Results were very similar, with median rho 0.33, range -0.32-0.85, IQR 0.23-0.43).

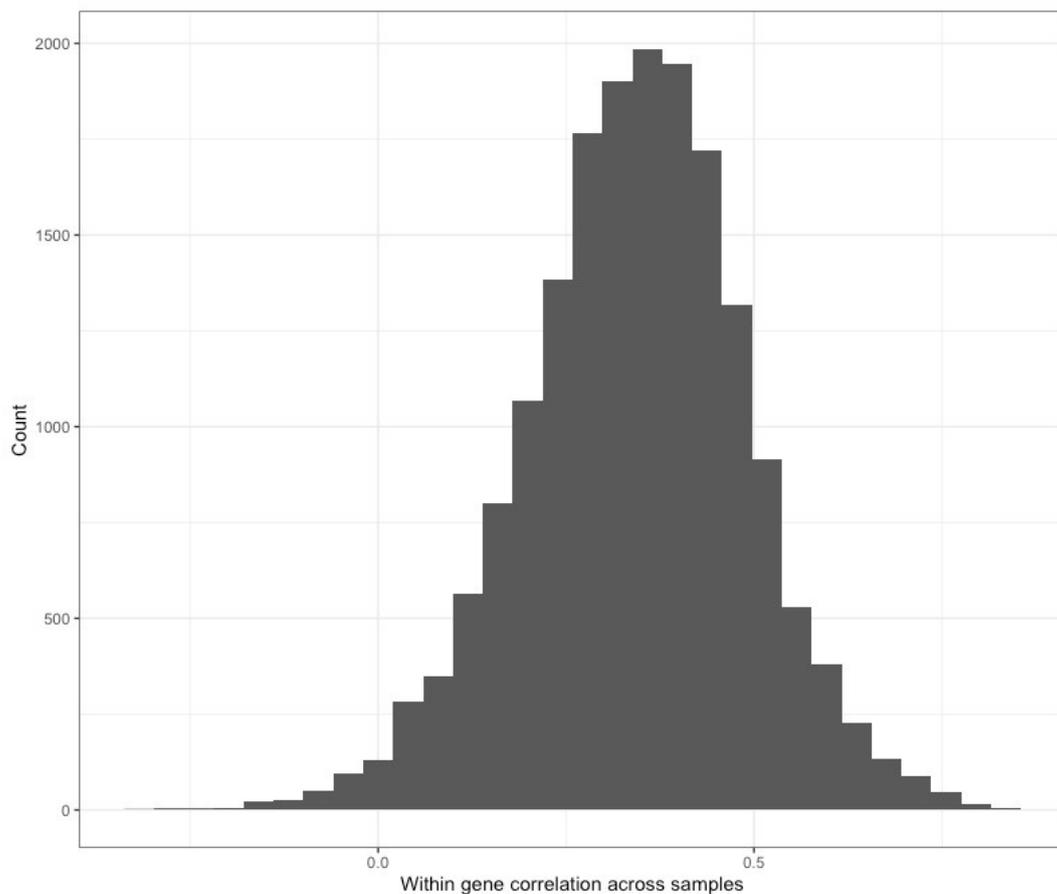


Figure 52: Spearman's rank correlation between sample expression orders for FFPE and fresh frozen tissue for every gene individually, for all genes present in at least 15 sample pairs.

Again, using samples normalised with TPM rather than number of reads made very little difference, with median rho for the 8040 universally detected genes being 0.31 (range -0.21-0.84, IQR 0.21-0.41) (Figure 53 left) and for those present in at least 15 sample pairs being 0.31 (range -0.41 – 0.91, IQR 0.19-0.42) (Figure 53 right). Nonetheless, from here forwards we will use TPM as our preferred measure of gene expression, because of the aforementioned theoretical advantages of this normalisation procedure.

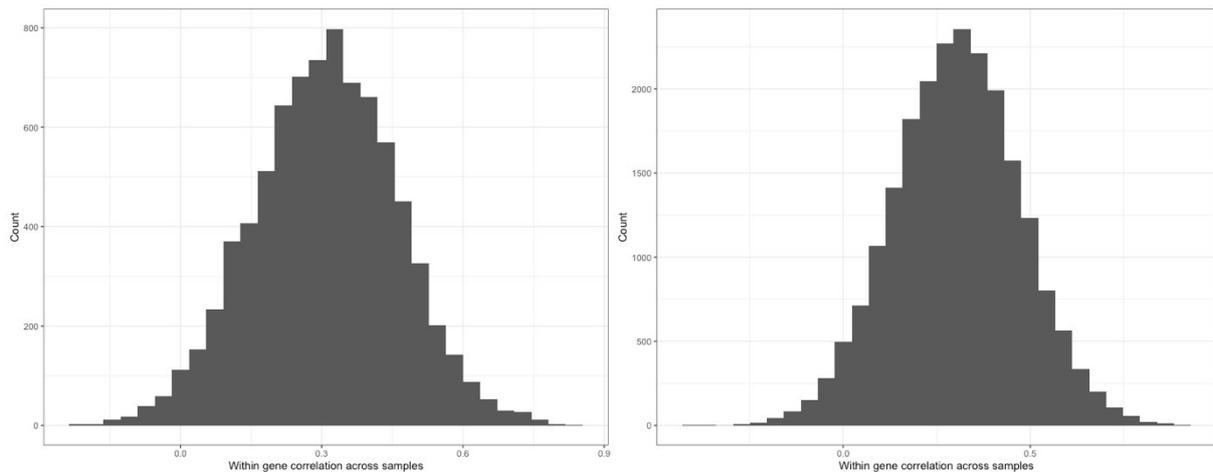


Figure 53: Spearman's rank correlation between TPM sample expression orders for FFPE and fresh frozen tissue for every gene individually, for the 8040 universally detected genes (left) and for all genes present in at least 15 sample pairs (right).

4.4.7 Receptor expression

Hormone receptor status is a robustly conserved feature across tumour types that defines treatment and prognosis (Figure 54). Plotting on a log scale, these demonstrated few outliers, and were strongly correlated across tissue types by both Spearman (ESR1 (ER) $\rho(40) = 0.72$, $p = 3.75 \times 10^{-7}$; ERBB2 (HER2) $\rho(40) = 0.44$, $p = 3.82 \times 10^{-3}$) and Pearson (ESR1 (ER) $r(40) = 0.74$, $p = 2.71 \times 10^{-8}$; ERBB2 (HER2) $r(40) = 0.55$, $p = 1.82 \times 10^{-4}$). While these correlations only explain about 54% of the shared variance in ER and 30% of shared variance in HER2, they are nonetheless somewhat reassuring, demonstrating that RNA-seq on FFPE is able to detect large changes in receptor expression relatively reliably.

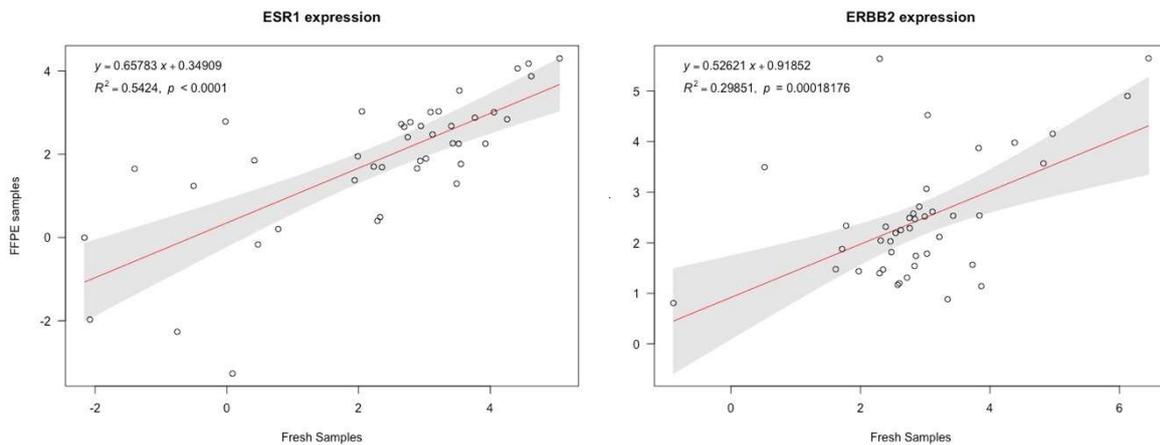


Figure 54: Receptor gene expression correlations across FFPE and fresh frozen tissues. Scales are TPM log₁₀.

4.4.8 Previously problematic key genes

Next, I examined RSF1 (and its co-expressed neighbour CLNS1A) which is particularly important for defining integrative cluster 2 (Curtis *et al.*, 2012), and showed very weak correlation in expression between NanoString (FFPE) and Illumina (Fresh Frozen) platforms (Spearman's rho 0.25; Chapter 3 Table 22).

RNA-seq displayed better across-preparation correlations (Figure 55). RSF1 was significantly correlated both non-parametrically ($\rho(40) = 0.34$, $p = 0.0305$) and parametrically ($r(40) = 0.51$, $p = 5.54 \times 10^{-4}$), and this was not driven by outliers. The distribution of CLNS1A expression levels was similar, and the correlation slightly better ($\rho(40) = 0.57$, $p = 1.27 \times 10^{-4}$; $r(40) = 0.69$, $p = 4.37 \times 10^{-7}$).

Overall, this shows that some genes that were difficult to compare using NanoString (FFPE) and Illumina (Fresh Frozen) may be more quantifiable using RNAseq, but still only around a third of the variance across tissue preparation techniques is explained. For samples with moderate expression, this level of precision may still be insufficient to support integrative cluster classification.

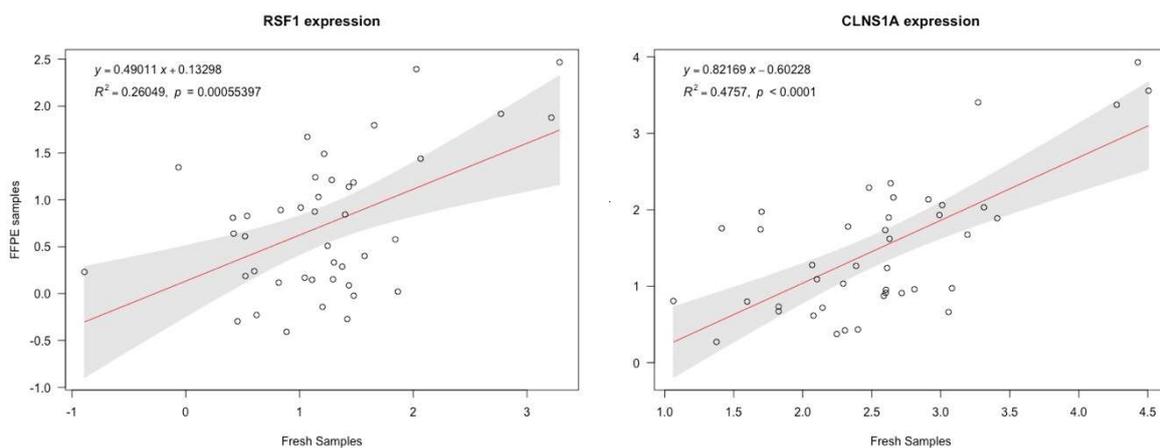


Figure 55: RSF1 and CLNS1A gene expression correlations across FFPE and fresh frozen tissues. Scales are TPM log10.

4.4.9 Integrative Cluster genes

With this in mind, I returned to the Curtis *et al.* (2012) paper to examine the genes that define Integrative Clusters. Of the 714 genes that comprised the 754 features used in this paper, 510 overlapped with the genes detected with RNAseq in at least one sample. However, some of these genes were missing in quite a large number of samples (Figure 56 left), with a clear elbow in the histogram for genes that were missing in more than six samples (Figure 56 right). Genes missing in more than six samples were excluded from further analysis. This resulted in the exclusion of 22 further genes.

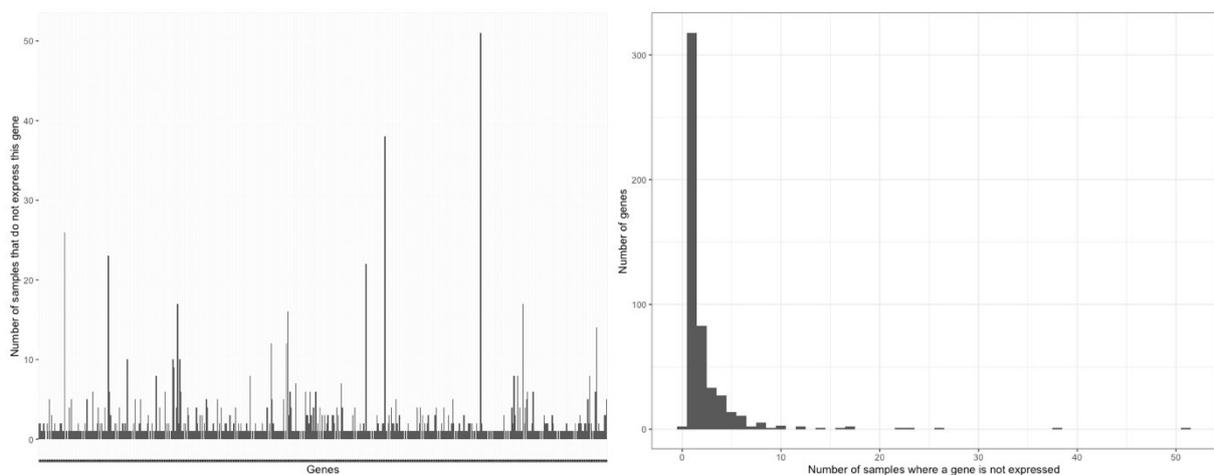


Figure 56: Number of samples where Integrative Cluster genes were not expressed.

A large majority of the rest of the genes were missing in one or two samples. This was primarily driven by a single sample that was missing almost all of the genes and another missing more than 100 (Figure 57). These two samples were excluded from further analysis.

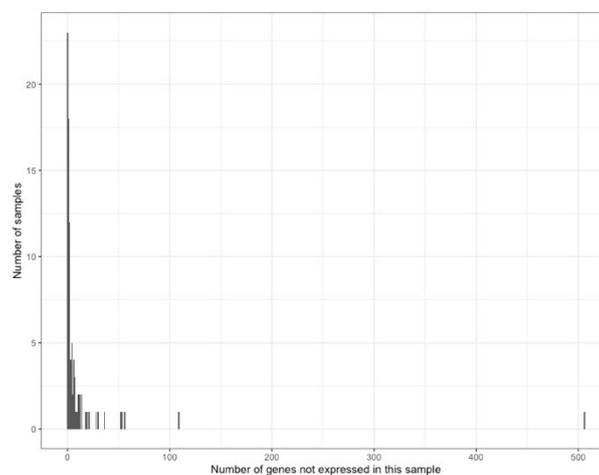


Figure 57: Total Integrative Cluster gene expressions not detected in each sample

Following this quality control, we had 488 genes that were detected in the vast majority of samples. Compared to the 8040 genes analysed earlier, these genes had very similar non-parametric correlation within-gene across-sample (median rho 0.35, range -0.02 – 0.68, IQR 0.26-0.42) (Figure 58 left), but a significantly lower and more similar parametric correlation within-gene across-sample (median r 0.55, range 0.11– 0.96, IQR 0.45-0.64) (Figure 58 right). While normally a lower correlation coefficient would be disappointing, here it is reassuring, as the similarity of the Spearman and Pearson coefficients suggests that the parametric correlations are no longer primarily driven by outliers, but rather that these genes display a consistent range of expression across our samples.

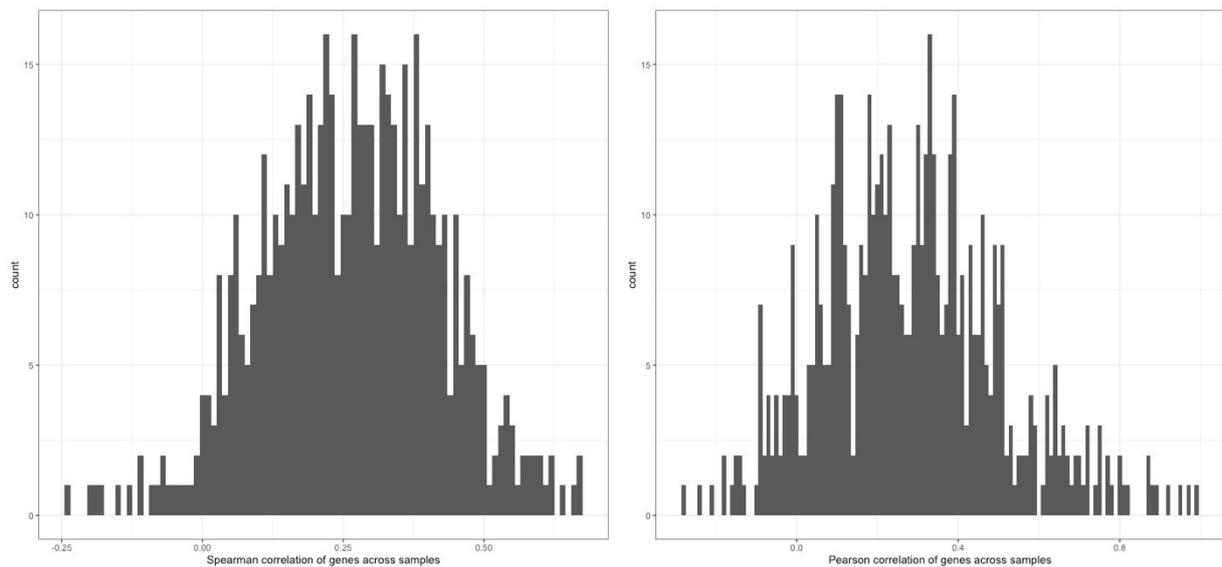


Figure 58: Spearman and Pearson correlations within iC10 genes of interest across-samples.

4.4.10 iC10 gene relationships

487 of the 488 genes that were detected in the vast majority of samples were also features of interest in the published iC10 classifier (Rueda, 2015). Overall, across all samples, the expression profiles of these genes correlated well between FFPE and fresh frozen tissue types with, as previously observed, a general tendency for mismatches to reflect higher expression levels in fresh than FFPE tissue (Figure 59).

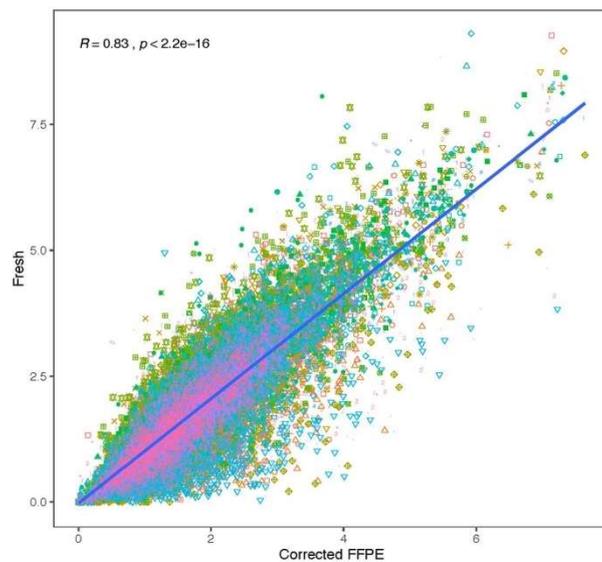


Figure 59: Log gene expression count for iC10 genes by tissue type, coloured by sample identity, with a line of unity plotted.

Measures were corrected by removing genes with zero detected expression in either preparation.

Some samples showed a tighter correlation in iC10 gene expression than others. Previously, in Figure 46, we observed that across all genes for all samples this was related to the number of genes detected. Even after our quality control steps, trimming the poorest quality samples and least consistent genes this remained the case, with the best correlating samples expressing almost all genes, and those with poorer correlations missing a significant number in either or both tissue types (Figure 6o). The best of these 'difficult' larger group samples approached the correlation of those 'good' samples used in the pilot study (Figure 38).

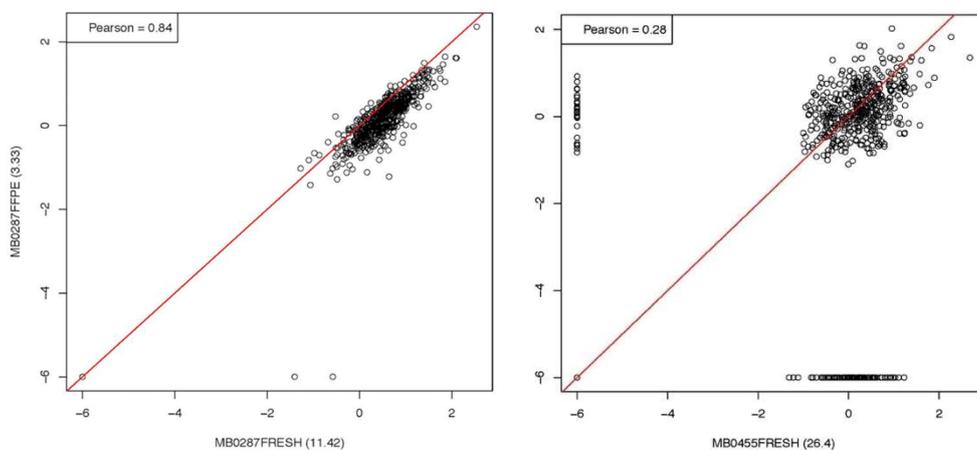


Figure 6o: Correlation between gene expression for FFPE vs fresh tissue for a 'good' and a 'bad' sample.

4.4.11 Differential Expression Analysis

In order to identify genes that were less affected by the tissue preparation process, I performed differential expression analysis between fresh frozen and FFPE samples on all genes using the edgeR package with voom transformation and DGEList functions. I then identified those that were not significantly differentially expressed between different tissue types.

I assessed a total of 58006 transcripts. 57315 were not significantly differentially expressed ($p > 0.05$), while 691 were differentially expressed. None of those that were significantly differentially expressed were in the iC10 classification genes.

However, the absence of an overall difference in expression between tissue preparation methods does not mean that expression levels are sufficiently correlated to support classification. I therefore performed within-gene Spearman correlations across paired samples for these non-differentially expressed genes. Only 639 displayed $\rho > 0.7$, which is generally considered a strong correlation in these software packages. Only eleven of the 540 genes in the iC10 panel were strongly correlated. This does not bode well for the classifier. However, some of the strongly correlated genes are known to be particularly important for cluster determination, for example CCND1, which had a rho of 0.819.

There were not enough samples to perform differential expression analysis between integrative clusters. In order to extend this analysis to identify genes that are differentially expressed between groups based on FFPE samples one would ideally require 12 samples per cluster (Schurch *et al.*, 2016).

4.4.12 iC10 classification using the existing classifier

Finally, I assessed how well I was able to recapitulate the integrative cluster membership from RNA-seq data for the 42 sample pairs that passed quality control. All of the sample pairs in this analysis were part of the original Illumina dataset (Curtis *et al.*, 2012), so for each the 'gold-standard' integrative cluster membership was known. The published iC10 classifier determines the integrative cluster of a sample based on the relative expression of key genes (Rueda, 2015). It was trained on Illumina microarray data from fresh frozen tissue, so here I have the ability to assess the impact of two variables. By comparing our fresh frozen samples, I can determine how well the existing classifier transfers from Illumina microarray data to RNA-seq data from the same samples, prepared in the same way. Then, by comparing the FFPE samples, I can examine the superadded effect of changing tissue preparation method.

For the fresh frozen samples, overall classification accuracy was only 33.3% (i.e., two thirds of samples were misclassified). Kappa is a measure of accuracy that accounts for differences in the number of samples in each category, and this was 0.249 (perfect classification would be result in a kappa of 1).

Examining the confusion matrix (Table 33) gives an insight into the drivers of this poor performance. The classifier assigned 31 of the 42 samples to integrative cluster 9, while only five of the samples actually belonged to this cluster. Integrative cluster 9 contains high grade tumours with p53 mutations that, importantly here, have very high genomic instability. It therefore seems likely that the classifier was assigning the RNA-seq expression profiles because they seemed to possess a high frequency of mutation, and lacked sufficient defining features for their true integrative clusters. A possible explanation for this is that RNA-seq is not limited to targeted genomic sequences, and with current read depth the signal to noise ratio for genes of interest may be poor.

However, the silver lining was that those samples that were not assigned to integrative cluster 9 were usually assigned to their true class. The positive predictive value of the classifier was 100% for integrative clusters 1 and 2, 80% for integrative cluster 5, and 67% for integrative cluster 6.

| Fresh | Classified Integrative Cluster | | | | | | | | | | |
|--------------------------|--------------------------------|---|---|---|---|---|---|---|---|---|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| True Integrative Cluster | 1 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 0 |
| | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| | 5 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| | 6 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| | 10 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 |

Table 33: Confusion matrix for iC10 classifier based on RNA-seq gene expression from Fresh Frozen tissue samples.

For FFPE tissue, integrative cluster classification was even worse, with an overall accuracy of 26.2% (i.e., almost three quarters of samples were misclassified) and a kappa of 0.164. The confusion matrix (Table 34) shows that again this time poor performance was not driven by misclassification to a single group. Again, some samples were misclassified to integrative cluster 9, but now a large number were assigned to integrative cluster 4, and some to integrative cluster 3. Integrative cluster 4 is “normal-like”, and does not have striking over-expression of any particular genes. It may be that the added noise introduced by the FFPE preparation method has reduced our ability to detect the over-expression of key genes, leading overall profiles to look more like the average to the classifier. Again, positive predictive value was good for some integrative clusters, especially 1, 2 and 7, but this was at the expense of very poor sensitivity.

| FFPE | Classified Integrative Cluster | | | | | | | | | | |
|--------------------------|--------------------------------|---|---|---|---|---|---|---|---|---|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| True Integrative Cluster | 1 | 1 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 3 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 4 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 5 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 0 |
| | 6 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 7 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 |
| | 8 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 9 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| | 10 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |

Table 34: Confusion matrix for iC10 classifier based on RNA-seq gene expression from FFPE tissue samples.

4.5 Further investigations of data quality

Machine learning classification methods can be highly influenced by non-specific differences in the profile of input data. It is therefore not surprising that the published iC10 classification method, trained on Illumina microarray data, performed poorly on RNA-seq data. My initial intention was to accrue a significantly larger dataset to train an entirely new classifier for Integrative Clusters based on RNA-seq. However, given the poor correlations in measured gene expression between FFPE and Fresh frozen tissue I have demonstrated above, this did not seem a good use of resource. Therefore, this seemed a sensible to pause, examine the possible reasons for poor agreement between methods, and plan future experiments accordingly.

4.5.1 Haplotyping

Haplotyping is a robust process that is usually used to verify that two samples are from the same patient. It relies on the matching of allelic variations, which form a ‘fingerprint’ that is relatively unique to an individual. Here, I used it as a basic check of data quality – could haplotyping match the RNA-seq data from FFPE and Fresh Frozen samples with a high degree of accuracy?

I used the GATK best-practice pipeline for variant calling using the joint calling method 2.0². I used variant calling files to compare the sequence in known variant regions to the reference sequence for identification. In our samples, I identified 554,285 variants in total. I removed variants with low coverage, defined as fewer than five reads in more than 75 samples. The primary outcome measure was concordance across paired samples, with a value of 1.0 being complete agreement between samples. As a first step I assessed concordance with self, to verify that the diagonal of the matrix was 1, and the off diagonals were lower.

For fresh frozen tissue (Figure 61) the results were reassuring, with uniformly high correlations along the identity matrix, and low correlations on the off-diagonals, indicating a good ability to haplotype match to self in fresh frozen tissue. However, for FFPE tissue (Figure 62) the results were much more mixed. There were still uniformly high correlations along the identity matrix, but now there were also high correlations in many of the off-diagonals. This is concerning, as it means that many samples cannot be reliably distinguished by haplotype calling from FFPE tissue, normally a robust technique.

² <https://gatk.broadinstitute.org/hc/en-us/articles/360035890431-The-logic-of-joint-calling-for-germline-short-variants> – retrieved 2nd February 2021

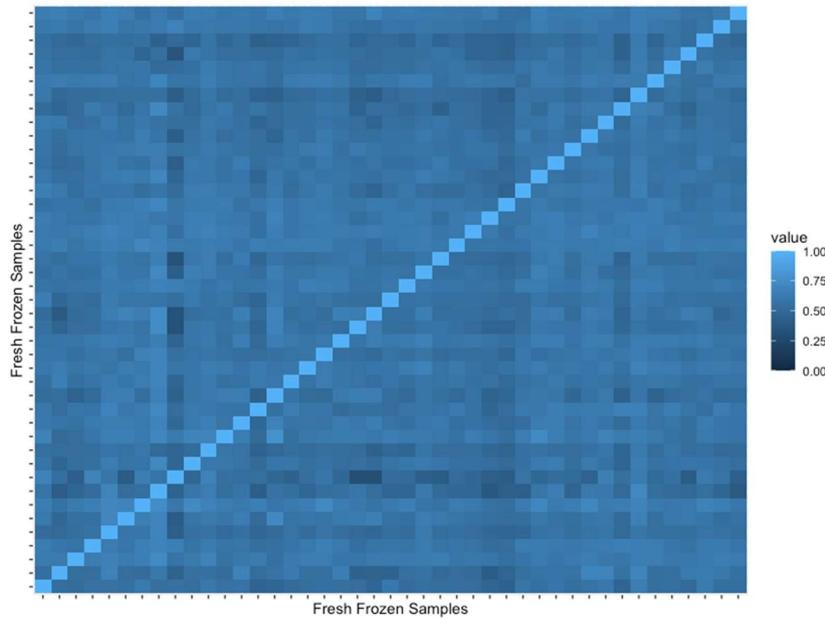


Figure 61: Autocorrelation of allelic variants in fresh frozen samples.

The template samples, which determine the allelic variants to be assessed, are on the x-axis, and the test samples are on the y-axis. Reassuringly, there are uniformly high correlations along the identity matrix, and low correlations on the off-diagonals, indicating a good ability to haplotype match to self in fresh frozen tissue.

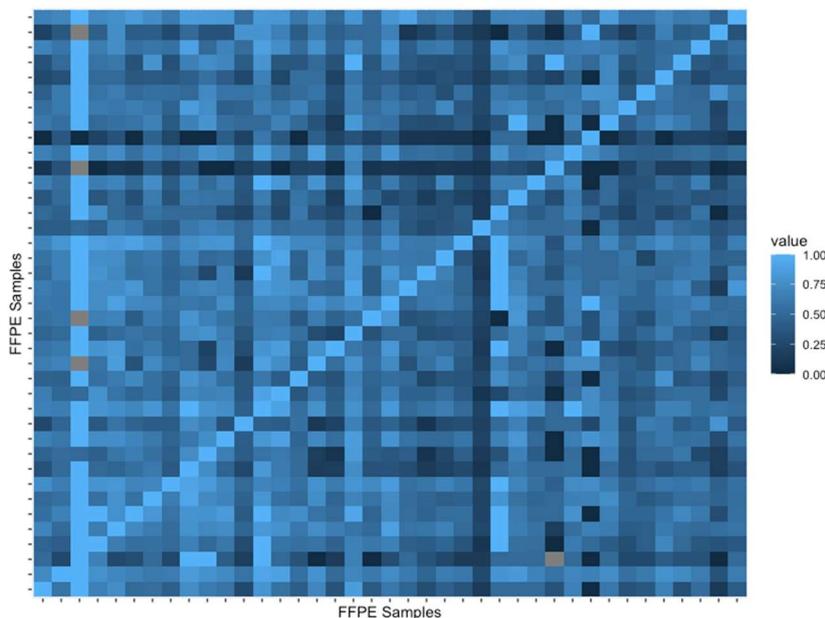


Figure 62: Autocorrelation of allelic variants in FFPE samples.

The template samples, which determine the allelic variants to be assessed, are on the x-axis, and the test samples are on the y-axis. Again, there are uniformly high correlations along the identity matrix, but now there are also high correlations in many of the off-diagonals. This is concerning, as it means that many samples cannot be reliably distinguished by haplotype calling from FFPE tissue, normally a robust technique.

Next, I examined the concordance between fresh and FFPE tissues, using fresh samples as the template to determine the allelic variants to be assessed (Figure 63). This is usual best practice, using the higher quality samples as the template. With this method, there were again many high correlations along the identity matrix, but this was not uniformly the case. Overall, eleven samples better correlated with at least one sample from another patient than with themselves. Sometimes this was because of samples displaying globally high concordance with non-self, (Figure 64 left), while sometimes it was due to globally low concordance, including with itself (Figure 64 right).

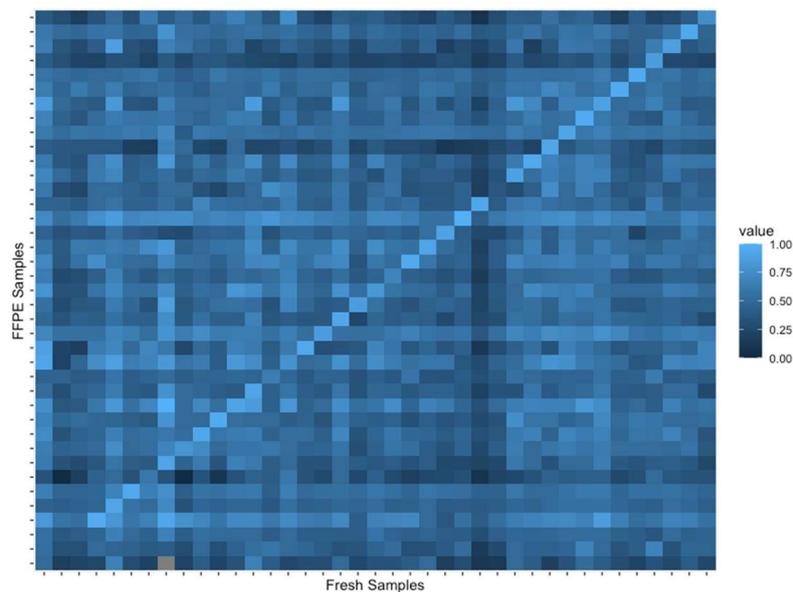


Figure 63: Concordance of allelic variants between FFPE and fresh frozen samples, with fresh samples as the template.

There were again many high correlations along the identity matrix, but this was not uniformly the case. For example, in the bottom left it can be seen that the first three samples have poor correlation with self. Overall, 11 of the 42 samples better correlated with at least one non-self sample than with self.

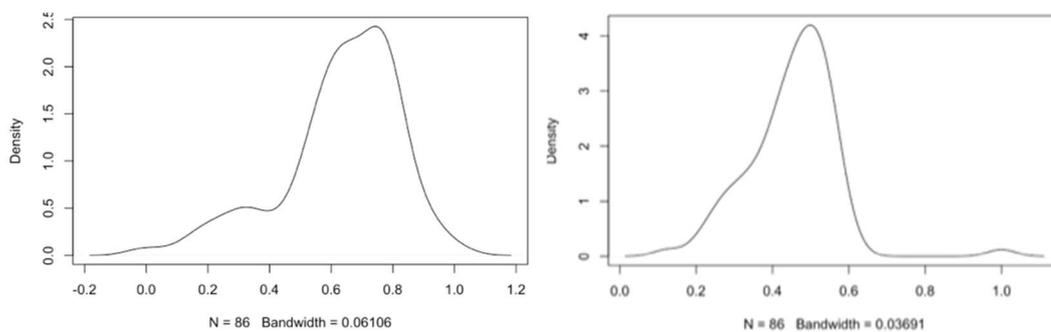


Figure 64: Two example problematic samples for haplotype calling. P44D8 (left) has abnormally high concordance with all samples and P62F2 (right) has abnormally low concordance with all other samples, including its FFPE equivalent

I also examined the same concordance using FFPE tissue as the template (Figure 65). With this method, nineteen samples were miscalled. As with the autocorrelation of FFPE samples, this seemed to be driven by a large number of samples displaying very high non-self correlations, as indicated by the overall brightness of the matrix, including the off-diagonals.

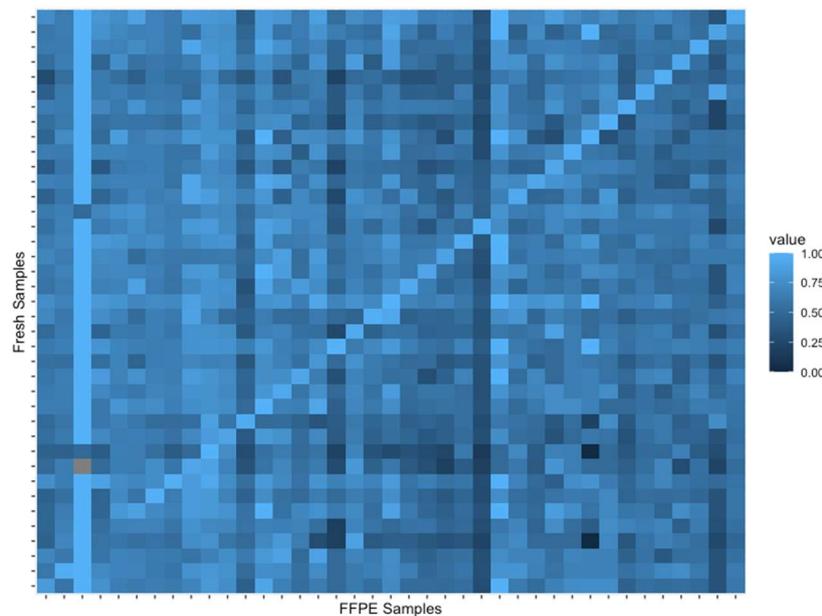


Figure 65: Concordance of allelic variants between fresh frozen and FFPE samples, with FFPE samples as the template.

There were again many high correlations along the identity matrix, but this was not uniformly the case. Now there were a large number of high off-diagonal correlations. This resulted, overall, in 19 of the 42 samples being mis-called.

Haplotype calling is generally a robust process, so the failure of this system in more than a quarter of our samples is of very significant concern. While it is, of course, possible that human error in sample labelling or at some other earlier stage might have rarely led to sample identities being misassigned, it is not realistic that this could apply to a quarter of my dataset. The poor performance seemed to be particularly driven by the FFPE tissue, which displayed high off-diagonal correlations even against self. This analysis therefore points to a significant data quality issue for RNA-seq data that will almost certainly have reduced the reliability of downstream analyses.

4.5.2 Comparison with Illumina Microarray

Given these concerns, I revisited the Illumina Microarray data from the previous chapter, in order to compare gene expression in fresh frozen tissue across different platforms.

I was able to obtain the Illumina Microarray data that was used in Curtis *et al.* (2012) for the same patients as the 42 pairs of samples. There were 16481 genes for which expression could be detected with both techniques in fresh frozen tissue. Correlation between techniques, within-patient across all genes, was relatively weak (median rho 0.687, IQR 0.677-0.696) (Figure 66), and in fact very similar to that previously seen between NanoString (FFPE) tissue and Illumina (Fresh) microarray (median rho = 0.660, Chapter 3, Figure 23 check before submission).

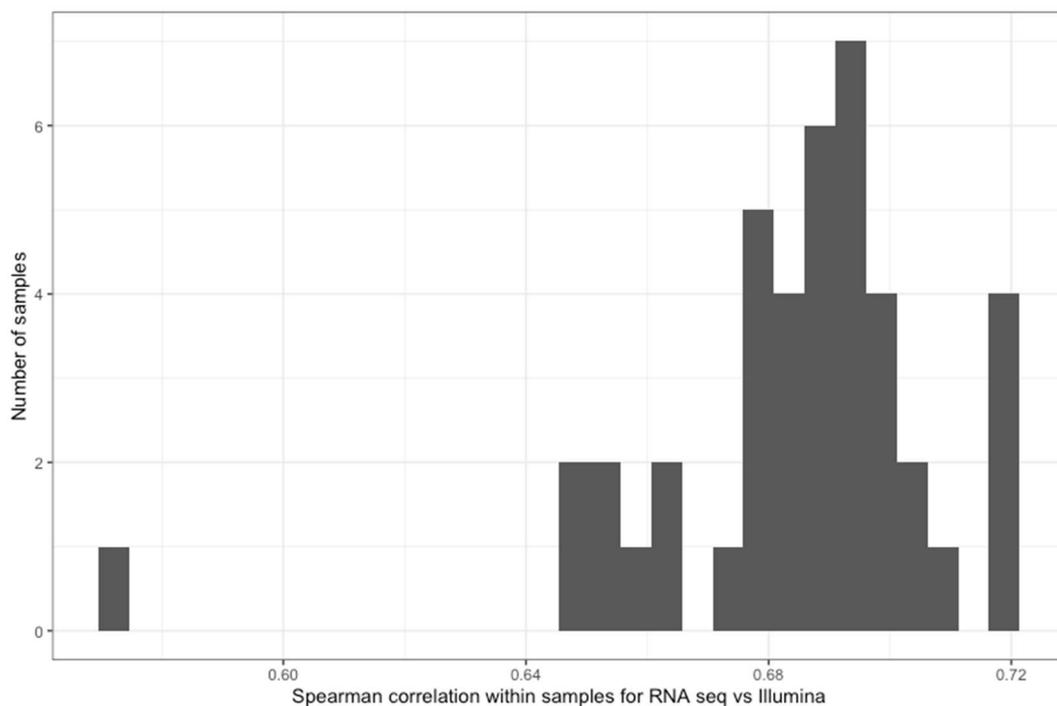


Figure 66: Within-patient across gene correlation for Illumina Microarray and RNASeq expression

I then assessed the within-gene, within-patient correlation for each of the 16481 genes individually. The correlation was poor (median 0.256, IQR -0.085 – 0.440, range 0.604 - 0.936, Figure 67).

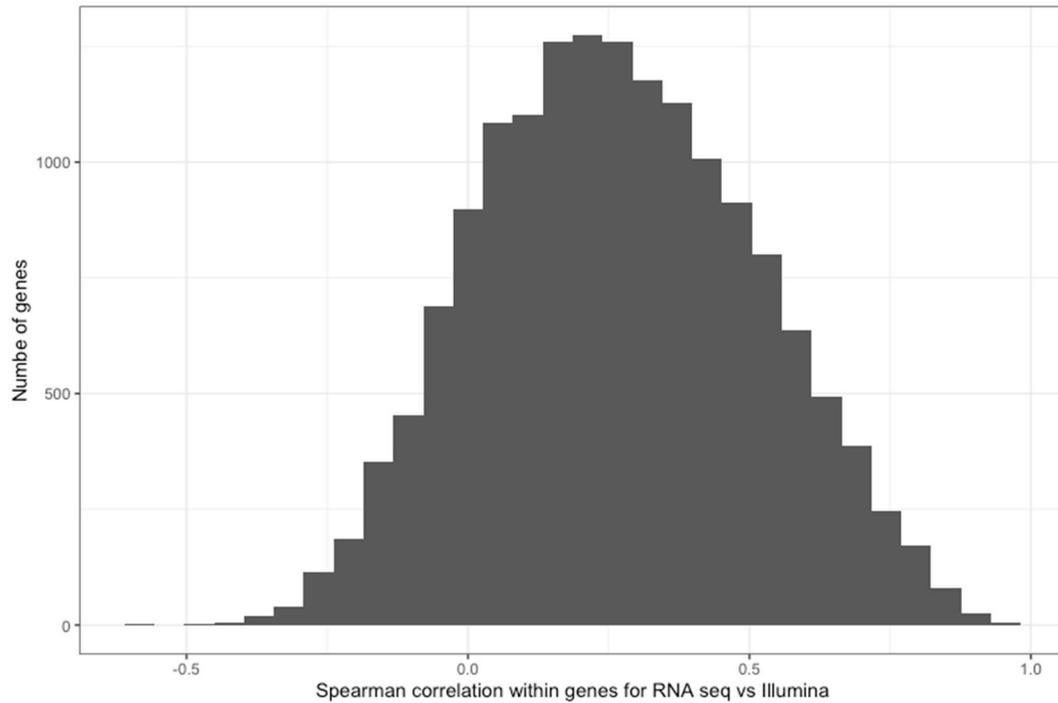


Figure 67: Within-gene within-patient correlation for Illumina Microarray and RNASeq expression

When sub-setted to only the 540 genes that were present in the iC10 package the within-gene correlation remained low (median rho 0.286, IQR 0.161 – 0.410, range -0.174 – 0.869).

4.6 Conclusion

Overall, I have raised fundamental questions about the ability of RNA-seq to recapitulate gene expression levels measured from the same patients using Illumina microarray. Further, there was relatively poor correlation between the gene expression levels measured with RNA-seq between fresh frozen and FFPE tissues. In these circumstances, it is unsurprising that the iC10 package was unable to use RNA-seq data to successfully classify two thirds of the samples from fresh frozen tissue and three quarters of the samples from FFPE tissue.

My initial pilot study of three samples assessed only within-patient across-gene correlations due to the small sample size and found very good agreement. In the larger scale study of 96 samples, some pairs showed similarly high agreement, but a significant number did not. It was not possible to predict which samples would show poor agreement prior to sequencing, as it was not correlated with RNA (Figure 43) or cDNA (Figure 44) concentration. Instead, the strength of within-patient correlations across tissue types was strongly correlated with the number of genes for which expression could be detected (Figure 46). This suggests that it is primarily driven by mRNA sequencing quality rather than quantity. It may be that performance could be improved by optimisation of the sample processing pipeline, such as increasing the read depth.

The next step would be to formally validate the use of RNA-seq data with the existing iC10 classifier, which was designed based on Illumina microarray data. This should initially involve the original cohort as it is the “gold standard”, and repeat the microarray experiment and RNA-seq with the same, stored fresh frozen RNA samples. Then both Illumina microarray and RNA-seq could be performed on a separate cohort, also with fresh frozen RNA samples, and the results compared. Classifying using these new data and comparing across samples would validate the use of iC10 classifier on data acquired from RNA-seq with fresh frozen RNA samples.

If the reliability of the technique can be improved with future work, the next step will be to establish a classifier that is robust to tissue preparation method and sequencing technique. This will require a larger dataset, containing at least 12 patients per

integrative cluster, with gene expression measured in as many ways as possible for every sample (Illumina microarray, NanoString, and RNA-seq) on both fresh frozen and FFPE tissue. In this way, a classifier could be trained to determine which genes best determine integrative cluster membership, and which should be treated with caution as heavily influenced by tissue preparation or analysis technique.

5. Digital pathology - cell classification

5.1 Preface

In the previous chapters I examined practical and deployable methods for the molecular classification of breast cancer. Another deployable resource that is almost universally available is the H&E stained pre-treatment diagnostic biopsy, and accompanying simple immunohistochemistry. In this chapter, I discuss the refinement of automated methods for nuclear segmentation and cellular classification, on both H&E and IHC images.

These methods have great potential for automatically quantifying metrics like tumour burden and lymphocyte density, but also open the door to more complicated analyses of the spatial relationships between tumours and their microenvironment, such as those that I will describe in chapter 6. Pathological profiling of tumour characteristics in clinical practice must be robust to inter-observer variation. In cancers such as melanoma, where the “briskness” of lymphocyte infiltration is widely acknowledged to be prognostically important, this remains a subjective judgment vulnerable to disagreement between pathologists. Automated quantification methods could potentially augment such judgments, allowing a more consistent assessment of established features and facilitating the translation of a wider range of tumour properties that have previously been confined to the research domain.

I am grateful to Dr A Dariush (an astrophysicist), who wrote the base code for cell segmentation and classification; he and Dr H Raza Ali had already designed the support vector machine (SVM) workflow before I started (Ali *et al.*, 2016). I created a larger training set for this technique, and together we discussed, developed and optimised the

Random Forest method, with me providing the biological theory and making the training set, and Dr A Dariush writing and implementing the Python code. I implemented all of the validation analyses on test materials, described in this chapter. I am also grateful to Ms Helen Bardwell, who assisted me during the process of digitising the slides. I undertook all other procedures described in this chapter myself, except where I have specifically stated otherwise in the text.

5.2 Background

Histopathology is a vital part of patient care, forming the basis of many treatment decisions, especially for breast cancer, but it requires a great deal of training, is labour intensive, and largely non-quantitative. There is a long-term ambition to complement expert clinical pathology with automated, digital methods. Images of part or all of a slide, taken with a camera attached to the microscope, have a long history in medical education. However, the digitisation of whole slides for diagnostic and research purposes is a much more recent development. In 1999, Wetzel and Gilbertson (Wetzel *et al.*, 1999) employed the Pittsburgh Supercomputer to perform the first automated geometrical abstractions of microscopic images, using prostate cancer biopsies, as proof of concept. But it took another fifteen years for mainstream computer hardware to evolve to a level that could support the large file sizes and significant graphical demands of storing and displaying high magnification images at sufficient quality (Ghaznavi *et al.*, 2013). Better internet connectivity is now enabling the examination of slides from a remote location in clinical practice. This is known colloquially as digital pathology, but it still entails manual assessment of the digital image by a human histopathologist.

In the research domain, computer algorithms have been developed to interpret the images, building on Wetzel and Gilbertson's first steps. Most of these algorithms employ a first step of segmenting (separating out) the basic features of a histopathology slide, such as cells or structures. The segmentation of individual cells is usually performed by the identification of cell nuclei (Irshad *et al.*, 2013), as histological stains usually make this the most prominent visual feature of a cell. These segmentation algorithms are relatively robustly developed (Thomas and John, 2017), but the classification of identified cells into their types is a less mature process (Vu *et al.*, 2019). This chapter will focus on the development of methods for cell classification in breast cancer, on the basis of nuclear morphology and the characteristics of their immediate surroundings, forming the basis for Chapter 6, which deals with the larger-scale task of

assessing clusters of tumour cells and their spatial relationship with their microenvironment.

As discussed in Chapter 1, our group has previously published on automated methods for cell classification, based on a support vector machine technique (Ali *et al.*, 2013; Ali *et al.*, 2016; Ali *et al.*, 2017). To do this, a pre-defined set of visual features is first extracted from the slide using computerised image processing techniques (see Chapter 2, methods). Each of these visual features forms a dimension in multi-dimensional space. A support vector machine then finds the best way to separate groups of cells by placing boundary planes known as hyperplanes in this multi-dimensional space. It places these planes such that the distance between the plane and the nearest points is maximised, i.e. a ‘machine’ determines the plane that is best ‘supported’ by ‘vectors’ from the nearest points in high dimensional space – a ‘support vector machine’, widely known as an SVM. Support vector machines are robust and widely used, and have the advantage that the visual features of interest are pre-defined and the basis on which classification is occurring can be studied. However, a disadvantage is that there may be important but unknown visual features that are not included in the pre-defined list and hence are not available to the classifier.

The random forest machine learning algorithm, on the other hand, is based on a process known as decision trees. A decision tree is a flow-chart like algorithm, where a series of “tests” is put to the object being classified and the direction of the “flow” depends on the “answer”. A machine builds a very large number of such decision trees completely at random, and is trained through a process of ‘majority voting’ to build a classifier that achieves the best result. The advantage of such a classifier is that the features of interest do not need to be pre-specified, and it can simply be provided with an image of a nucleus and its immediate surroundings.

Machine learning methods are many and varied. Other well-known techniques include Bayesian networks and neural networks. Bayesian networks are probabilistic graphical models, and are particularly useful to calculate the probability of various possible causes of a given outcome. For example, a model can be built to calculate the probability that a dead plant is caused by overwatering, underwatering or disease. Unlike support vector machines and random forests, which both assign a definitive class to every cell, Bayesian networks assign probabilities and certainties. This can be helpful for some purposes, and replicates optimal human decision making, but makes it very difficult to undertake the sorts of spatial analyses I propose in Chapter 6. Also, while the algorithm is fast to compute and performs well with a high number of input features, the major disadvantage for our application is that it assumes all features are equally important and are independent from each other, which is often not true in cell morphology.

Neural network replicate brain functioning, transforming input information through a series of “layers” by transforming the inputs depending on their learned “weight” (importance), sometimes multiple times, to calculate an output. This is very computationally complex and training has a large demand on the hardware quality of computers. Neural networks perform best when they have a very large number of training data points, which is why their most common application is through corporations such as Google undertaking picture identification and handwriting recognition. As such, while these methods have huge promise for digital pathology analysis, this work has to date been performed on large-scale tiles rather than individual cells, as it is much more practical to generate the large training sets they require (Barker *et al.*, 2016; Alom *et al.*, 2019). A whole slide image is divided into “tiles” of smaller images and are fed into neural network algorithms to predict outcome. While this produces good abstractions of some data types, for example classifying tumour types or measuring immune infiltrate in a way that may be useful for broader prognostication, it does not provide the opportunity for acquiring the additional spatial information that we wish to assess following classification (Chapter 6).

For my purposes, therefore, random forests and support vector machines are the more appropriate tools, and indeed are the most successful tools for disease prediction in a wide variety of contexts (Uddin *et al.*, 2019). Both have been applied to histopathological images previously, but it is not known which is superior for the classification of cells generally, and much less for breast cancer specifically (Komura and Ishikawa, 2018). Theoretically, support vector machines have the advantage of dealing well with high dimensional correlated input, such as the large number of image features we have identified. However, it can struggle when classes can overlap - for example, in our case, reactive fibroblasts (stroma) can be similar to some tumour cells, and low grade tumour cells have a large morphological overlap with benign breast epithelial cells. Random forest also performs well for high dimensional data, and is less prone to over-fitting, which potentially makes the algorithm more applicable when applied to tissue processed across multiple laboratories and trials.

5.3 Aims and Objectives

The overarching aim of this chapter is to assess the accuracy of machine learning methods for automated cell classification in digitised pathological slides, which will form the basis for the spatial analyses of tumour clusters and their microenvironment described in Chapter 6. Detailed methods for these techniques have been previously described in chapter 2, but I touch on them again here when it aids the understanding of the validation analyses.

I addressed this aim in a number of ways:

I began by applying an existing published method based on a support vector machine (SVM) algorithm (Ali *et al.*, 2016) to a completely new dataset. I created a training set based on the haematoxylin and eosin (H&E) slides described in section 5.4.

I then developed a new method based on random forest (RF) algorithms, which I trained on the same H&E data to allow for direct comparison of these methods head-to-head, both with automated cross-validation on the same data, and manually on a novel dataset from a different trial.

Next, I extended the random forest method from H&E slides to immunohistochemistry (IHC) slides stained for a variety of immune cells. I performed detailed evaluations of both a membranous staining pattern (CD8), which can identify cytotoxic T-cells, and a nuclear stain (FOXP3), which identifies regulatory T-cells.

Finally, I assessed whether my automated methods could replicate a previously reported association between lymphocyte density and response to neo-adjuvant chemotherapy.

5.4 Materials: Breast cancer cases and the trials they were obtained from

For the analyses I describe here, I collected and organised the slides from pre-treatment biopsy, mid-treatment biopsy (where available) and surgery from four clinical trials and studies, detailed in Table 35.

MONET is an historical trial (EudraCT 2005-001698-89), during which patients were given neo-adjuvant hormone therapy only.

The PERSEPHONE trial (Earl *et al.*, 2018; Earl *et al.*, 2019) has completed accrual and is in the follow up phase. All patients received Herceptin for either 6 or 12 months. I only included those patients (38 out of 95) who received neo-adjuvant chemotherapy.

TransNEO (Sammut, 2019) is an ongoing study based at Cambridge University Hospitals (CUH), where patients are given standard neoadjuvant chemotherapy dependent on their molecular tumour subtype, but consent to detailed genomic analysis of their tumour. In addition to the pre- and post-treatment research samples taken in most other neoadjuvant trials, many patients have additional research-only biopsies taken mid-way through chemotherapy. Some of the patients that were in PERSEPHONE or PARTNER trial were also enrolled in TransNEO, and I ensured that each patient was only counted once in the statistical analysis.

PARTNER (Earl *et al.*, 2017; Abraham *et al.*, 2018) is an ongoing trial comparing the combination of Olaparib and platinum-based chemotherapy against chemotherapy alone for triple negative breast cancer, or any breast cancer with a germline mutation in either *BRCA1* or *BRCA2* gene, irrespective of hormone receptor status. *BRCA1/2* genes are involved in the DNA damage response and repair pathways. Olaparib is an inhibitor of the poly-ADP ribose polymerase (PARP), which is involved in the DNA single-strand break repair pathway and is particularly promising for the treatment of *BRCA1/2* breast cancers, where DNA repair mechanisms are already compromised (Robson *et al.*, 2017).

5. Digital pathology - cell classification

| Trial or Study Name | Breast Cancer Subtype | Design | Research Ethics Committee Number | Trial Registration Number | Number of Patients |
|----------------------------|-------------------------------------|---|---|----------------------------------|---------------------------|
| MONET | Post-menopausal ER positive | Clinical trial of neoadjuvant tamoxifen vs exemestane | 06/Q108/105 | 2005-001698-89 | 26 |
| PERSEPHONE | HER2 Positive | Clinical trial of Herceptin treatment duration: 6 vs 12 months | 07/MRE08/35 | 2006-007018-39 | 95 |
| TransNEO | All Types | Neoadjuvant chemotherapy observational study – patients gave additional blood and tissue samples during treatment | 12/EE/0484 | n/a | 175 (ongoing) |
| PARTNER | Basal Type Triple Negative and BRCA | Clinical trial of platinum-based chemotherapy +/- Olaparib | 15/NW/0926 | 2015-002811-13 | 33 (ongoing) |

Table 35: Characteristics of the trials and studies from which digital pathology slides were generated

In total, I scanned and annotated 6955 slides, including mega slides, divided amongst the trials as in Table 36

| Trial/Study | Number of Patients | Number of Slides from Diagnostic Biopsy | Number of Slides from Surgical Resection | Total Number of Slides |
|--------------------|---------------------------|--|---|-------------------------------|
| MONET | 26 | 75 | 507 | 582 |
| PERSEPHONE | 38 | 61 | 680 | 741 |
| TransNEO | 178 | 1157 | 3514 | 4671 |
| PARTNER | 33 | 302 | 659 | 961 |

Table 36: The number of slides annotated and scanned from each study and trial

The majority of the slides were stained with H&E stain (Table 37), with the remainder being IHC. Trial and study images from fresh frozen material are not included in the table below – I scanned and annotated these, but the nuclear morphology was varied, and likely too different from fixed material to use the training set generated from FFPE tissue, and there were too few samples to create a new training set for this preparation method. Most of the slides were of the breast or lymph nodes including tumour cells and background tissue, with the remainder being samplings from surrounding areas such as axilla or fibro-fatty tissue as described in Table 37.

| Trial / Study | Number of Slides from Biopsy | | | | | | Number of slides from Surgery | | | | | |
|---------------|------------------------------|-----|-------|--------|------------|-------|-------------------------------|-----|-------|--------|------------|-------|
| | H&E | IHC | Total | Breast | Lymph Node | Total | H&E | IHC | Total | Breast | Lymph Node | Total |
| MONET | 25 | 50 | 75 | 25 | 50 | 75 | 490 | 17 | 507 | 296 | 211 | 507 |
| PERSEPHONE | 42 | 19 | 61 | 38 | 23 | 61 | 668 | 12 | 680 | 460 | 220 | 680 |
| TransNEO | 421 | 736 | 1157 | 290 | 867 | 1157 | 3322 | 191 | 3513 | 2158 | 1355 | 3513 |
| PARTNER | 43 | 259 | 302 | 30 | 272 | 302 | 629 | 30 | 659 | 444 | 215 | 659 |

Table 37: Breakdown of slides by stain and tissue type

I scanned and annotated all of the slides in this table. The machine learning methods were trained on the H&E TransNEO data from breast tissue, both at biopsy and surgery, with the other trials providing data for validation analysis. The IHC and Lymph Node slides provide a well characterised, annotated digital dataset that I have made available for future studies.

5.5 TransNEO patient characteristics

The TransNEO study is unpublished and makes up approximately two thirds of the slide collection, so I calculated the residual cancer burden (RCB) status for slides where it had not previously been done, annotated and catalogued the slides, and summarised the pre-treatment tumour statistics (Table 38). Residual cancer burden is measured by the size and cellularity of the residual tumour in the resection specimen after neoadjuvant chemotherapy, and the presence of lymph node metastasis. RCB was obtained for 169 of the 178 patients, the remaining 9 not having surgical specimens available. Residual cancer burden (RCB) status was double-read and verified by Dr Elena Provenzano.

| | Number | Percent |
|-------------|--------|---------|
| Tumour size | | |
| 0 | 45 | 26.6 |
| 1 - 20 mm | 75 | 44.4 |
| 21 - 50 mm | 28 | 16.6 |
| >50 mm | 21 | 12.4 |
| Total | 169 | 100.0 |
| | | |
| Node status | | |
| Negative | 95 | 56.2 |
| Positive | 74 | 43.8 |
| Total | 169 | 100.0 |
| | | |
| Grade | | |
| 1 | 0 | 0 |
| 2 | 62 | 36.7 |
| 3 | 98 | 58.0 |
| Not known | 9 | 5.3 |
| Total | 169 | 100.0 |

| | Number | Percent |
|-----------------|--------|---------|
| ER, HER2 status | | |
| ER-, HER2- | 36 | 21.3 |
| ER-, HER2+ | 14 | 8.3 |
| ER+, HER2- | 61 | 36.1 |
| ER+, HER2+ | 49 | 29.0 |
| Not known | 9 | 5.3 |
| Total | 169 | 100.0 |
| | | |
| RCB | | |
| 0 (pCR) | 45 | 26.6 |
| I | 27 | 16.0 |
| II | 68 | 40.2 |
| III | 29 | 17.2 |
| Total | 169 | 100.0 |
| | | |
| | | |
| | | |
| | | |

Table 38: Trans-Neo study tumour characteristics

5.6 Results – algorithm performance

5.6.1 Qualitative properties of the methods, visual inspection

I trained and assessed two different classification systems; an existing support vector machine classifier (Ali *et al.*, 2016) that I re-trained on my dataset, and a novel random forest classifier. Figure 68 shows examples of cell classification based on H&E images, which I have chosen to illustrate the qualitative differences in output between these methods.

Overall, both methods are able to recapitulate the tissue structure, and are accurate and consistent in identifying stromal cells (Figure 68a). However, where lymphocyte density is high the support vector machine appears to perform less well, mistaking many of these cells for epithelium (Figure 68b) or stromal cells (Figure 68c). The random forest method appears superior here, but makes different errors elsewhere, for example confusing stromal cells surrounding lymphovascular spaces for epithelium (Figure 68d).

These are all subjective observations, which seemed to appear as common themes. In the sections that follow, I assess and quantify these classification errors systematically.

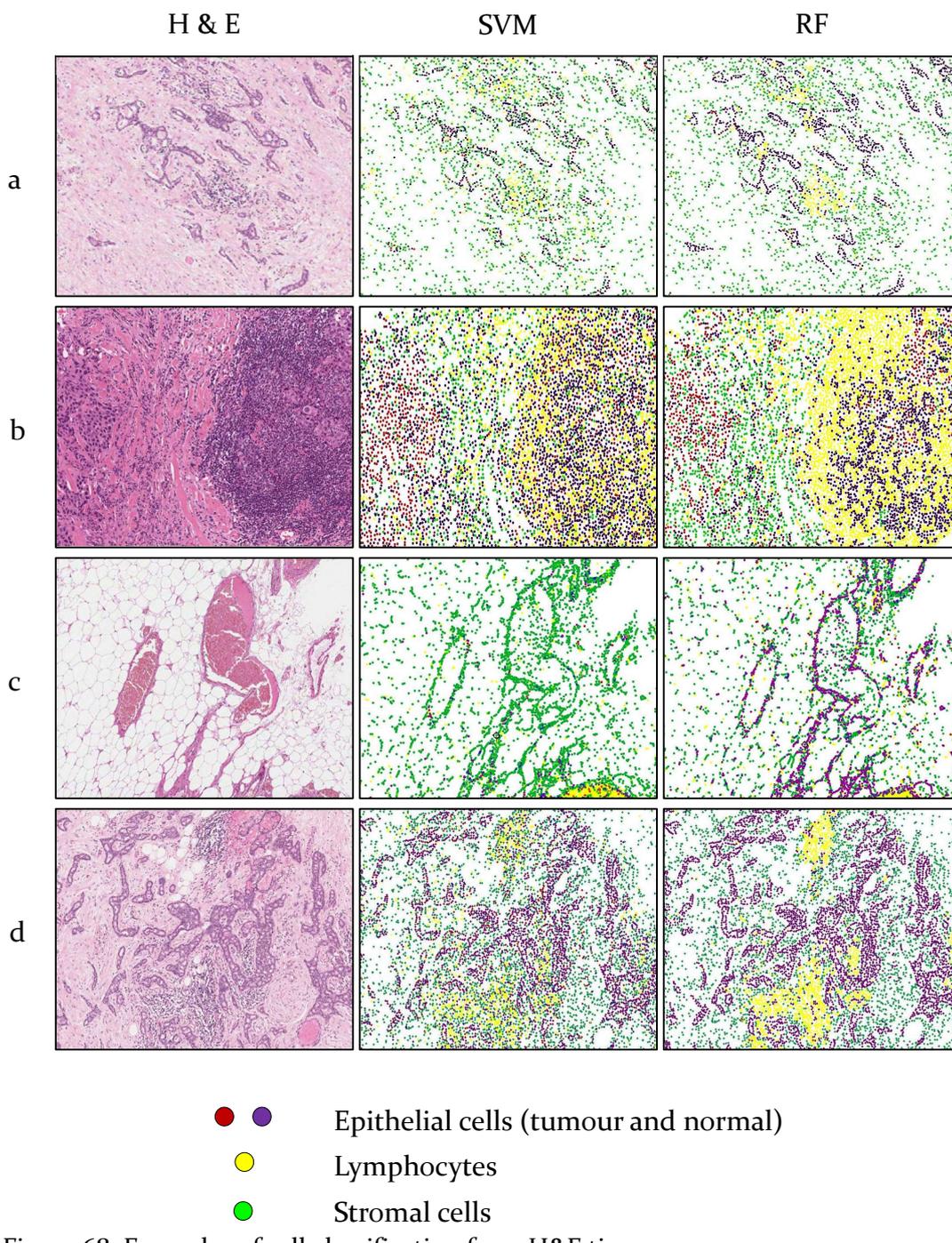


Figure 68: Examples of cell classification from H&E tissue

Images were chosen to illustrate the differences in support vector machine (SVM) and random forest (RF) classifications.

The left column shows the digital image of H&E stained slides, chosen for optimal illustration of the subjective properties of the methods. The middle and right columns show the location of the centre of segmented and classified cells by each method, colour coded by classified cell type. While tissue structures can be clearly appreciated in these panels, looking closely one can observe that they are in fact made up of dots, each of which is placed at the centre of a segmented nucleus.

Row a illustrates that both support vector machine and random forest methods appear accurate and consistent in identifying stromal cells (green).

Row b illustrates that in regions where the lymphocyte density is high, such as this germinal centre, support vector machine appears to classify more germinal centre cells as epithelial (red or purple) whereas random forest appears to correctly identify a higher proportion of the population as lymphocytes (yellow). Note that the yellow area is not shaded, but rather composed of individual tightly packed lymphocyte nuclei.

Row c shows the support vector machine correctly identifying the cells around the lymphovascular space as stromal cells (green), while random forest appears to be identifying them as epithelial cells (purple).

Row d shows that both methods can identify tumour clusters well, however the support vector machine appears to perform more poorly than random forest in areas with high lymphocyte density, here falsely identifying epithelial (purple) and stromal (green) cells intermixed with lymphocyte clusters.

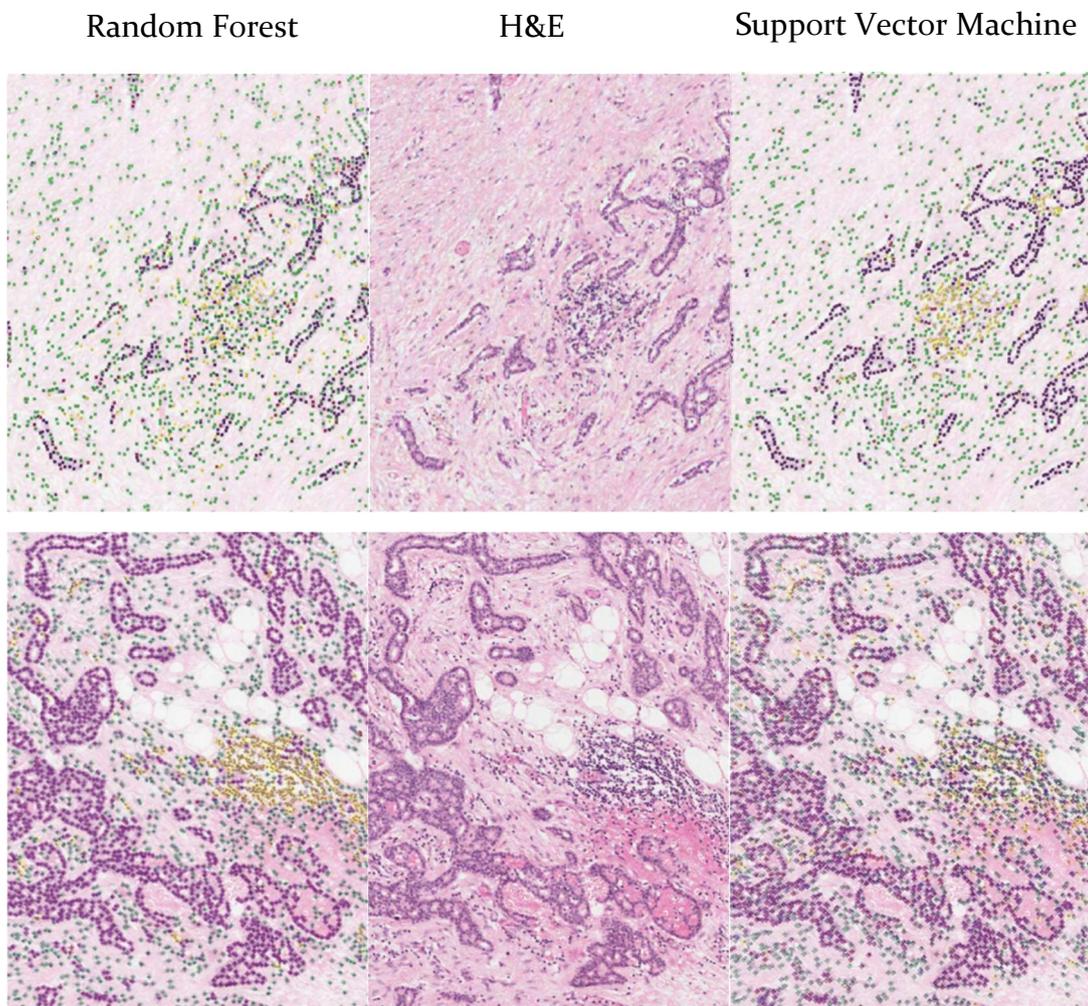


Figure 69. Higher magnification illustrative images.

The left panels are the cell classification by random forest overlaid on H&E images. The centre panels are the original H&E images. The right panels are the cell classification by support vector machine overlaid on H&E images.

5.6.2 Cross-validation within H&E training set

One way to assess classification performance is through cross-validation. In this procedure, a training set of many slides is created by a histopathologist, who manually identifies individual cells on a slide. In my case, I randomly selected 1000 cells per class across 40 slides from the TransNEO trial (an average of 25 cells of each class on each slide). The machine learning algorithm is then trained on a sub-set of these data, and tested on the remaining (withheld) data. This is known as a 'fold'. This fold is repeated a number of times, known conventionally as 'k'. Here I performed a k-fold cross-validation across 5 folds. This means that I trained each classifier five times on 80 percent of the data, and tested its accuracy on the remaining 20 percent, such that the accuracy of classification of every cell was assessed once. For each fold, overall classification performance can be assessed, meaning that both the average accuracy and also its variability over five repetitions can be quantified.

Here, I performed cross-validation after the removal of the "Mark" category (non-tissue slide components), and with breast epithelial cells divided into "Tumour" and "Normal".

Support vector machine

The overall mean accuracy of the support vector machine method was 0.884 (+/-0.008). It is possible to quantify several further performance metrics (Table 39). Positive predictive value (PPV) is the probability that if a cell is identified as being of a particular type, then it really is. In machine learning terminology, this measure is often called 'precision'. Sensitivity is the complementary measure, representing the proportion of cells of a given type that are identified correctly. Machine learning researchers sometimes refer to this as 'recall'. There is often a trade-off between these two measures, whereby increasing sensitivity leads to a reduction in positive predictive value and vice-versa. The F1-score is the harmonic mean of these measures. It is robust to differing group sizes, and is often targeted for optimisation when designing classification algorithms. All three measures were better for lymphocytes and stromal cells than they

were for tumour cells and normal cells (Table 39). These latter two cell types were often confused with each other, meaning that the support vector machine was less good at subclassifying epithelial cells than it was at differentiating them from other cell types (Table 40).

| | Positive Predictive Value | Sensitivity | F1-score | Number of cells |
|------------------------|----------------------------------|--------------------|-----------------|------------------------|
| Tumour | 0.88 | 0.77 | 0.82 | 393 |
| Lymphocytes | 0.94 | 0.95 | 0.95 | 373 |
| Stromal cells | 0.90 | 0.91 | 0.91 | 407 |
| Normal epithelial | 0.75 | 0.87 | 0.80 | 253 |
| Weighted Average/Total | 0.88 | 0.88 | 0.87 | 1426 |

Table 39: K-fold cross validation results for support vector machine across 5 folds

Positive predictive value (PPV) is the probability that if a cell is identified as being of a particular type, then it really is. Sensitivity represents the proportion of cells of a given type that are identified correctly. The F1-score is the harmonic mean of these measures.

| | Tumour | Lymphocytes | Stromal cells | Normal epithelial cells |
|--------------------------------|---------------|--------------------|----------------------|--------------------------------|
| Tumour | 301 | 12 | 22 | 58 |
| Lymphocytes | 6 | 356 | 4 | 7 |
| Stromal cells | 16 | 10 | 372 | 9 |
| Normal epithelial cells | 18 | 1 | 15 | 219 |

Table 40: Confusion matrix for support vector machine classifier.

Columns represent classified identity, while rows represent true class.

Random Forest

The mean accuracy of the random forest classification was 0.886 (+/-0.006). Validation broken down by cell type is in Table 41 and the confusion matrix in Table 42.

| | Positive Predictive Value | Sensitivity | F1-score | Number of cells |
|-------------------|----------------------------------|--------------------|-----------------|------------------------|
| Tumour | 0.90 | 0.79 | 0.84 | 411 |
| Lymphocytes | 0.93 | 0.95 | 0.94 | 372 |
| Stromal cells | 0.90 | 0.92 | 0.91 | 369 |
| Normal epithelial | 0.75 | 0.85 | 0.80 | 255 |
| Average/Total | 0.88 | 0.88 | 0.88 | 1407 |

Table 41: K-fold cross validation results for random forest across 5 folds

Positive predictive value (PPV) is the probability that if a cell is identified as being of a particular type, then it really is. Sensitivity represents the proportion of cells of a given type that are identified correctly. The F1-score is the harmonic mean of these measures.

| | Tumour | Lymphocytes | Stromal cells | Normal epithelial |
|-------------------|---------------|--------------------|----------------------|--------------------------|
| Tumour | 325 | 12 | 20 | 54 |
| Lymphocytes | 7 | 353 | 6 | 6 |
| Stromal cells | 6 | 13 | 338 | 12 |
| Normal epithelial | 24 | 3 | 12 | 216 |

Table 42: Confusion matrix for random forest classifier.

Columns represent classified identity, while rows represent true class.

Despite the subjective differences displayed in some tissue types, overall accuracy was not significantly different between the two methods; $t(8) = 0.25$, $p = 0.809$. Both methods were most likely to confuse normal and tumour breast epithelial cells, while correctly identifying lymphocytes and stromal cells.

Examining the misclassified cells visually, it appeared that neither algorithm was effective at distinguishing between normal breast epithelial cells and the less pleomorphic, lower grade tumour cells (i.e. these cells were falsely called normal), nor at distinguishing reactive normal epithelium from tumour (i.e. these cells were falsely called tumour). Training and cross-validation performs best when roughly equal numbers of each cell type are provided to the algorithm, and indeed that is what we have done here. However, the primary goal of this classification is to form the basis of a spatial analysis of pre-treatment biopsies (Chapter 6), where benign epithelium is very scarce, and of no interest to the questions being asked. Therefore, for subsequent analyses I took the approach of collapsing the “normal epithelial” category into the “tumour” category for biopsy slides. In this way, I ensured that almost all tumour cells were correctly identified, at the known but acceptable cost of also identifying a small number of healthy epithelial cells as abnormal.

5.6.3 Test against histopathologist on new data

Cross-validation tends to under-estimate the true error rate, as the data used to test the model come from the same slides as the data used to train and refine the model. This means that the algorithm is familiar with the non-specific characteristics of those slides, such as stain intensity, which might confuse it when applied to new slides that it has never encountered before. It is therefore a more rigorous and difficult test of an algorithm to compare it to gold-standard performance on slides it has never previously encountered. Here, I tested the ability of the models (which had been trained on data from the TransNEO trial) to classify cells on slides from the MONET trial.

Given the suboptimal performance of the algorithms at distinguishing tumour cells from benign epithelium in the cross-validation step, for H&E slides this test on new data was performed against the categories: tumour cells, lymphocytes, stromal cells and “Mark”. Marks are non-cellular objects such as apoptotic bodies and cellular debris that should be identified as such by the segmentation algorithms and excluded from subsequent analysis.

I randomly selected a total of 2552 objects across 9 H&E slides from the MONET trial and identified them manually for comparison against algorithmic classification (Table 43).

| | Mark | Tumour | Lymphocyte | Stroma | Total |
|---------------|------|--------|------------|--------|-------|
| Pathologist | 167 | 1126 | 696 | 563 | 2552 |
| SVM | 51 | 1436 | 371 | 664 | 2552 |
| Random Forest | 51 | 1546 | 470 | 485 | 2552 |

Table 43: H&E Test set object classification.

It is important to note that this table does not give any information about the accuracy of classification. For example, I manually identified 167 marks across 9 slides, while the segmentation algorithm identified 51. This does not, however, mean that all 51 of these identified marks were correct. In fact, as can be seen in Table 44, only ten were correct. 157 true marks were identified as cells by the algorithms, and 51 cells were identified as marks.

| | Pathologist | | | | | Total |
|------------------------|-------------|------|--------|------------|--------|-------|
| | | Mark | Tumour | Lymphocyte | Stroma | |
| Support Vector Machine | Mark | 10 | 10 | 10 | 21 | 51 |
| | Tumour | 44 | 931 | 223 | 238 | 1436 |
| | Lymphocyte | 13 | 9 | 347 | 2 | 371 |
| | Stroma | 100 | 176 | 116 | 302 | 694 |
| | Total | 167 | 1126 | 696 | 563 | 2552 |
| Random Forest | Mark | 10 | 10 | 10 | 21 | 51 |
| | Tumour | 17 | 1014 | 231 | 284 | 1546 |
| | Lymphocyte | 10 | 11 | 378 | 71 | 470 |
| | Stroma | 130 | 91 | 77 | 187 | 485 |
| | Total | 167 | 1126 | 696 | 563 | 2552 |

Table 44: Confusion matrix for support vector machine and random forest on H&E images against pathologist.

Rows represent classified identity, while columns represent true class determined manually by the pathologist.

Overall, the accuracy of both methods was very similar, at around 75-80% for the three cell types (Figure 70). The random forest method was slightly more sensitive at detecting tumour calls and lymphocytes, at the cost of a small amount of specificity. Both methods used the same initial segmentation, which incorrectly classified most of the artefacts (“Marks”) on the slide as cells (Table 44), usually stromal for random forest, but often Tumour for support vector machine.

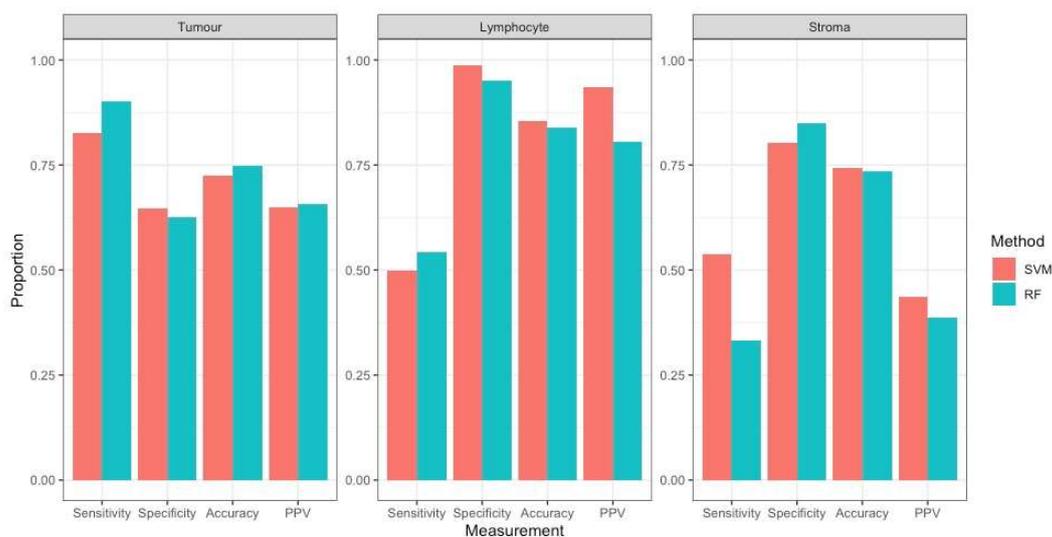


Figure 70: Manual validation of SVM and RF classifiers on H&E tissue, by cell type.

Sensitivity is the proportion of the positives that are correctly identified. Specificity is the proportion of the negatives that are correctly identified. Accuracy is the overall proportion correctly classified. PPV, positive predictive value, is the proportion of positive results that are true positive.

5.6.4 Immunohistochemistry (IHC)

Immunohistochemistry employs molecular staining to reveal cell phenotype that are not apparent from morphology alone. It is particularly useful in identifying subpopulations of lymphocytes, which may have prognostic value. For example, infiltrations of CD4⁺ helper T-cells (Gu-Trantien *et al.*, 2013), CD8⁺ cytotoxic T-cells (Liu *et al.*, 2012), and CD56⁺ natural killer cells (Rathore *et al.*, 2014) predict breast cancer survival in particular cancer sub-types. A particularly important modulatory role is proposed for FoxP3⁺ regulatory T-cells (Merlo *et al.*, 2009; Ladoire *et al.*, 2011; Lee *et al.*, 2013) and their interaction with the other cell types (Liu *et al.*, 2014).

Tissue microarrays allow for the analysis of these and other markers, but a robust automated method would significantly improve throughput. As a next step I therefore extended the application of my automated classification to IHC slides. Both algorithms performed similarly overall on H&E slides, but the random forest method was the more sensitive for both tumour cells and lymphocytes, so I therefore chose this algorithm to take forward to IHC. The aim was to quantify both the cells that were positive for the targeted antigen, but also the other lymphocytes on the same slide, to more accurately calculate the proportion of each sub population.

Based on the results of the H&E validation and test I created a slightly different training set for the random forest algorithm for IHC data. For example, “marks” were poorly identified on H&E slides through segmentation alone, and I anticipated that this would be even more difficult for IHC slides, where these marks are generally the same colour as positive cells. I therefore explicitly included marks in the training set, such that the random forest classifier was explicitly trained to identify them after segmentation. Additionally, I sub-divided lymphocytes into “brown”, if the cell expressed the antigen targeted, or “blue”, if the cell was negative. Tumour cells and stromal cells were trained as before, but benign epithelium was not trained given the previous difficulties in separating these two classes. The algorithm was trained on CD8 (membrane stain pattern) and FoxP3 (nuclear stain pattern), performed on tissue microarrays (TMA) from the B-CAST project (Barrdahl *et al.*, 2017). I also evaluated CD163 staining for macrophages, but the amount of non-specific staining made it difficult to manually

identify individual cells to allow the generation of a reliable training set so I did not train the random forest method on these data.

On H&E data, cross-validation overestimated performance compared to testing the algorithms against a pathologist on new data. I therefore proceeded straight to this more rigorous analysis and randomly selected and identified a total of 4521 untrained objects (Table 45) to test the random forest algorithm.

| | Mark | Tumour | Lymphocyte (blue) | Lymphocyte (brown) | Stroma | Total |
|---------------|------|--------|-------------------|--------------------|--------|-------|
| Pathologist | 748 | 1201 | 854 | 578 | 1140 | 4521 |
| Random Forest | 737 | 1110 | 1263 | 567 | 844 | 4521 |

Table 45: IHC test set object classification

Overall classification accuracy was very good (Figure 71), being above 80% for all cell types and significantly better than for H&E slides. The only large-scale confusion was that the algorithm had a tendency to misclassify stromal cells as “blue” (negative staining) lymphocytes (Table 46). Compared to the previous method, which excluded marks based on segmentation alone, training the algorithm dramatically increased its ability to correctly classify artefacts as not-cells, despite this subjectively being a more difficult task for IHC than H&E slides.

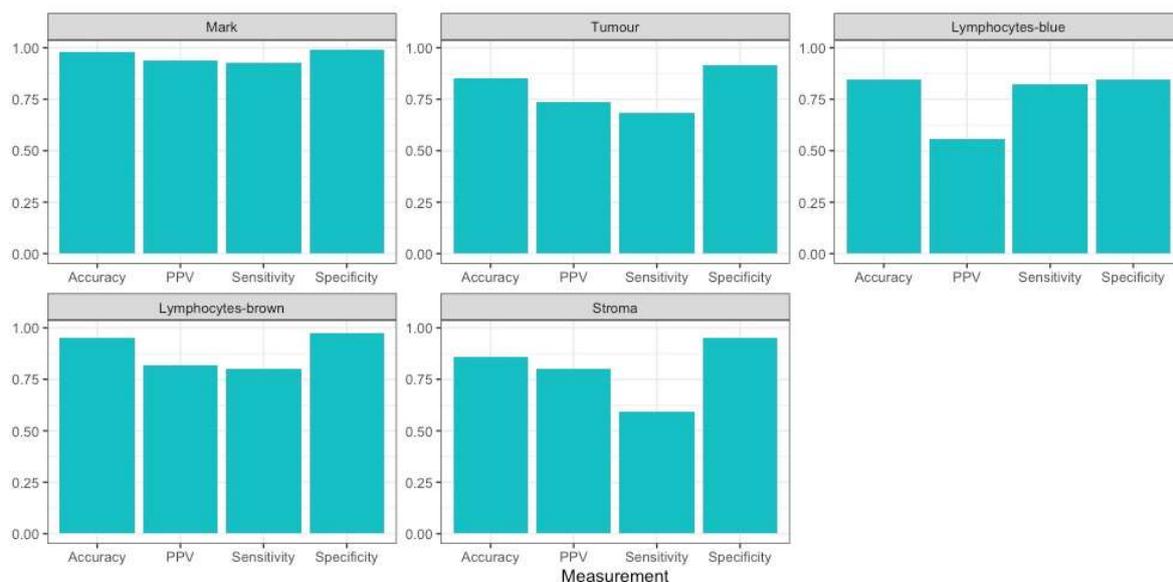


Figure 71: Manual validation of RF classifier on IHC, by cell type. PPV: Positive predictive value

| | | Pathologist | | | | | |
|---------------|--------------------|-------------|--------|-------------------|--------------------|--------|-------|
| | | Mark | Tumour | Lymphocyte (blue) | Lymphocyte (brown) | Stroma | Total |
| Random Forest | Mark | 692 | 11 | 1 | 19 | 14 | 737 |
| | Tumour | 16 | 819 | 60 | 1 | 214 | 1110 |
| | Lymphocyte (blue) | 7 | 242 | 705 | 86 | 223 | 1263 |
| | Lymphocyte (brown) | 30 | 15 | 47 | 463 | 12 | 567 |
| | Stroma | 3 | 114 | 41 | 9 | 677 | 844 |
| Total | | 748 | 1201 | 854 | 578 | 1140 | 4521 |

Table 46: Confusion matrix for random forest on test IHC images against pathologist classified cells. Rows represent classified identity, while columns represent true class

5.7 Results – slide property verification

5.7.1 Cellular composition of the samples

Having performed these validation analyses, I then applied the cell classifiers to the full dataset of H&E slides to ensure that they could perform overall counts and replicate simple associations.

The median number of cells segmented per slide was 218718 (range: 10 – 4219350). This is two or three orders of magnitude more than could feasibly be counted by a clinical histopathologist by hand. Slides were extremely variable in composition, with the proportion of tumour cells on a slide ranging between 0 and 90% (median 10%). Importantly, diagnostic specimens had a much larger number of cells classified as tumour than surgical specimens taken after NAT, showing that our algorithms were sensitive to treatment response (Figure 72). Breast tissue showed a much higher fraction of stromal tissue than lymphocytes, while the proportions were roughly equal in lymph nodes. The high fraction of stromal cells reflects increased sampling of background breast tissue, which often show a high proportion of stroma compared to epithelial cells. The surgical specimens from breast tissue showed far fewer lymphocytes than the diagnostic specimens, but in lymph nodes they had a slightly higher proportion of lymphocytes. This reflects both the normalisation of tissue post treatment, and also that sentinel lymph nodes are often surgically sampled, whereas only clinically and radiologically suspicious lymph node are sampled during diagnostic biopsies. It is important not to read too much into these data, as they reflect a difference in sampling methods between diagnostic biopsy and surgery, as well as true tissue changes as a result of NAT, but it is reassuring that they show changes in the expected direction.

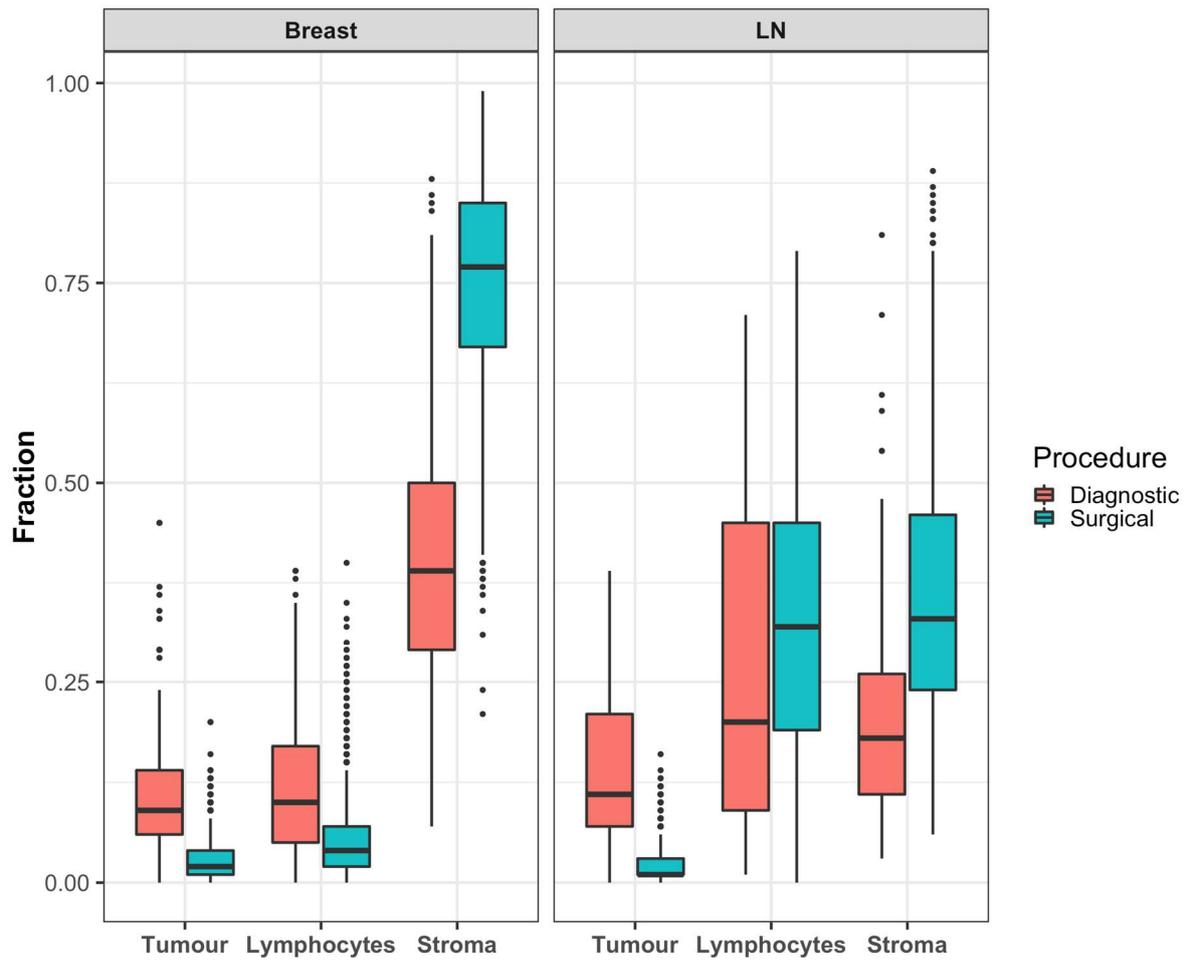


Figure 72: Cellular composition differences between diagnostic biopsies and surgical specimens.

Data from the support vector machine are plotted, but those from the random forest technique were extremely similar at this scale.

5.7.2 Median lymphocyte density

As discussed in the introduction (Chapter 1), median lymphocyte density is known to predict response to NAT. In this dataset I used the automated cell classifications to calculate the pre-treatment biopsy median lymphocyte density using the same k-nearest neighbour method used by Ali *et al.* (2016).

Assessed with my machine learning methods, the median lymphocyte density in the TransNEO trial was consistent with these previously published findings, replicating them in a novel dataset. Specifically, the median lymphocyte density of the biopsies from the patients who achieved pCR after NAT was significantly higher than that of those who had residual disease (Figure 73). There were no significant differences according to response in the density of any other cell type.

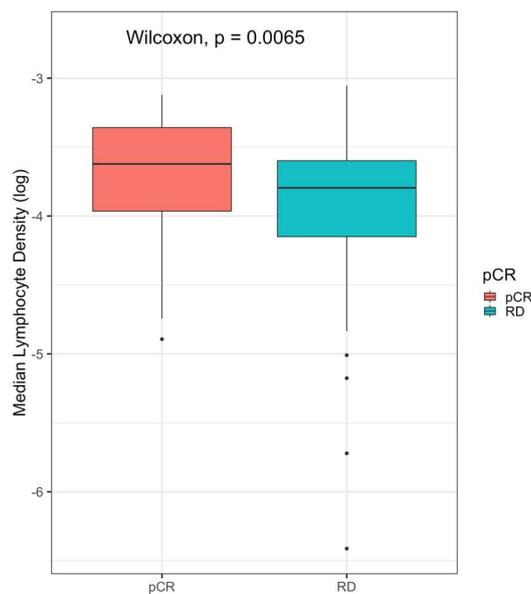


Figure 73: Pre-treatment biopsy lymphocyte density by chemotherapy outcome. Data from the support vector machine are plotted, but those from the random forest technique were extremely similar at this scale.

This effect seemed to be particularly pronounced for those samples that were HER2 positive (Figure 74)

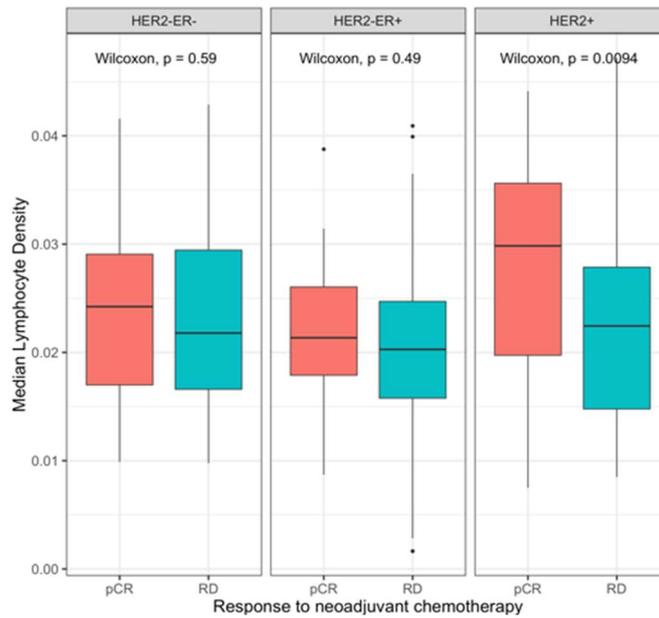


Figure 74: Pre-treatment biopsy lymphocyte density by receptor status and response to neo-adjuvant chemotherapy.

Data from the support vector machine are plotted, but those from the random forest technique were extremely similar at this scale.

5.8 Discussion

Accurately identifying cells or structures is the basis for any interpretation of the image. Many computational methods using complex mathematical models and image processing techniques have been developed to segment individual cells by accurately identifying the contrast between cell nuclei and the surrounding cytoplasm (Irshad *et al.*, 2013), but classification of these cells is a less mature field. The accuracy of the two methods we have used here, compared with pathologists' manual classification, is comparable with other published methods (Yuan *et al.*, 2012; Heindl *et al.*, 2018a).

Dundar *et al.* (2011) used a method called multiple-instance learning, using features extracted from cell nuclei including perimeter, the ratio of major to minor axis, and intensity. They attempted to distinguish *usual ductal hyperplasia* from atypical changes including *atypical ductal hyperplasia* and *ductal carcinoma in-situ*. For this binary classification, they trained with 66 cases and tested on 33 cases, and achieved an accuracy of 87.9%. Similarly, Beck *et al.* (2011) demonstrated 89% cross-validated binary classification accuracy in distinguishing breast epithelium and stroma.

Classification is more difficult as the number of tissue classes increases. Sirinukunwattana *et al.* (2016) used a Spatially Constrained Convolutional Neural Network to classify cells in colonic sections into four classes: *epithelial*, *fibroblasts*, *inflammatory* and *others*. They reported a F1-score (the harmonic mean of sensitivity and positive predictive value) of 0.784, using the same dataset in a 2 fold cross-validation. This is significantly inferior to our cross-validation performance, and comparable with our test on novel data.

However, some authors have recently reported superior results to those we present here based on neural networks. For example, models on brain and lung tissues often perform particularly well. Solorzano *et al.* (2021) used a fully connected neural network in glioma biopsies and achieved an accuracy of 83.5%. Wang *et al.* (2019) focused on adenocarcinoma of the lung, and classified cells into tumour, stroma and lymphocytes

using convolutional neural network. When using a true test set, the overall accuracy was 90.1%.

Some literature claims are somewhat difficult to interpret. For example, Lan et al report a support vector machine classifier with an accuracy of 97.1% for identifying ovarian cancer cells, and 89.1% for stroma (Lan et al., 2015). However, they discarded lymphocytes from their analysis as they were “close to other nuclei”, and it is unclear to me exactly how the evaluation was actually performed. The same group use the same method in melanomas, and report slightly inferior accuracies to our cross-validated measures, specifically 80.6% for stromal cells, 85.0% for cancer cells, 82.6% for lymphocytes, and 84.9% overall (Failmezger et al., 2020).

The advantages of automated methods are clear: for high throughput assessment of cellular composition of a slide, either H&E or IHC, they are fast and reproducible. Each image takes a computer between 1-3 minutes for the calculation of median lymphocyte density, saving hours of histopathologist time compared to manual methods for estimating these parameters by counting cells in randomly sampled fields. These methods also allow the calculation of spatial metrics that would otherwise be impossible to assess, techniques that form the basis of the next chapter.

Overall, the binary identification of only one type of cell, e.g. tumour cells vs all other cell types, is less challenging than the multiple class problem I have assessed here. This requires a more refined analysis of the features, as illustrated by our poor performance in separating normal and tumour epithelial cells. Our method was, however, sufficiently accurate to reproduce the known association between pre-treatment lymphocyte density and prognosis. Most importantly, it should be sufficient to support the spatial analyses I report in Chapter 6. Here, I look at the spatial relationships of clusters of cells, defining a tumour cluster as having a minimum size of 5 spatially proximate cells. Individually misclassified cells should not impair this analysis, as they are unlikely to occur in clusters of this size. Specifically, an 80% classification accuracy is unlikely to produce spurious clusters of five incorrectly identified tumour cells. My primary question relates to the spatial relationship between these clusters and lymphocytes, which are very accurately classified by our algorithm. However, it is

nonetheless important to identify the challenges preventing completely accurate classification, and these will be discussed below.

5.8.1 Our algorithms

For H&E images, the support vector machine and random forest algorithms show similar overall accuracy. The main difference between the two algorithms is that the support vector machine is feature-based and that all the features are based on cell nuclei, whereas random forest uses an image that is centred on cell nuclei, but also includes the area surrounding the nuclei, which includes cytoplasm for epithelial and stromal cells. The random forest algorithm therefore has the potential advantage of providing more information to the classifier. It does not, however, allow for transparent analysis of the relative importance of image features.

The algorithm we trained for immunohistochemistry has great potential. Many research projects or clinical assessments of tumour require the identification of not only the total amount of “brown” stain on the slide, but also confirmation that the stain is on a specific cell type, and even the proportion of that cell type that has been stained “brown”. Our algorithm has successfully demonstrated that it can be trained to adapt to different staining patterns, and that it can be very effective at identifying “negative” cells as well as “positive” cells. This will be very useful in future, not only to quantify the proportion of lymphocyte subtypes, but also potentially to assess receptor status such as ER positivity, and to automatically generate an Allred score with stain strength and proportion of cells stained (Phillips *et al.*, 2007). It can also be very useful in assessing for proliferative indices such as Ki-67 (Urruticoechea *et al.*, 2005; Yerushalmi *et al.*, 2010; Dowsett *et al.*, 2011), where often 1000 or more tumour cells need to be manually counted and the proportion of positive cells recorded, which is prone to human error and is time-consuming for pathologists. Our algorithm has been tested on tissue microarray, but it can just as easily be applied to whole slides.

5.8.2 Challenges

Before either classification algorithm can be applied, nuclei must be identified. This is a process known as segmentation. I did not develop or evaluate the segmentation method myself, using instead a previously published technique adapted to breast cancer histopathology from astronomical imaging (Ali *et al.*, 2013). Segmentation is generally performed by assessing the contrast between the nucleus and the surrounding region. This can result in other “dark” material being mistaken for a nucleus. Debris, pen-marks made by pathologists and formalin pigment deposits can usually be easily removed by algorithms on the basis of hue. However, some marks, for example nuclear debris or apoptotic bodies, can be identified as possible nuclei. This proved a problem for our H&E segmentation methods, which performed poorly in identifying these marks. In more than two thirds of cases they were misclassified as tumour or stromal cells. To address this, for the IHC dataset I generated a much larger training set for non-specific stains, which increased classification accuracy for these marks to 97.8%.

As well as cellular misclassification, other errors were occasionally observed. For example, nucleoli were sometimes identified as nuclei due to the colour intensity and contrast with their surroundings. Similarly, a single long and thin fibroblast nucleus was sometimes identified as two separate nuclei. Both of these errors can result in multiple nuclei being identified in a single nucleus, leading to double-counting and misclassification (Figure 75).

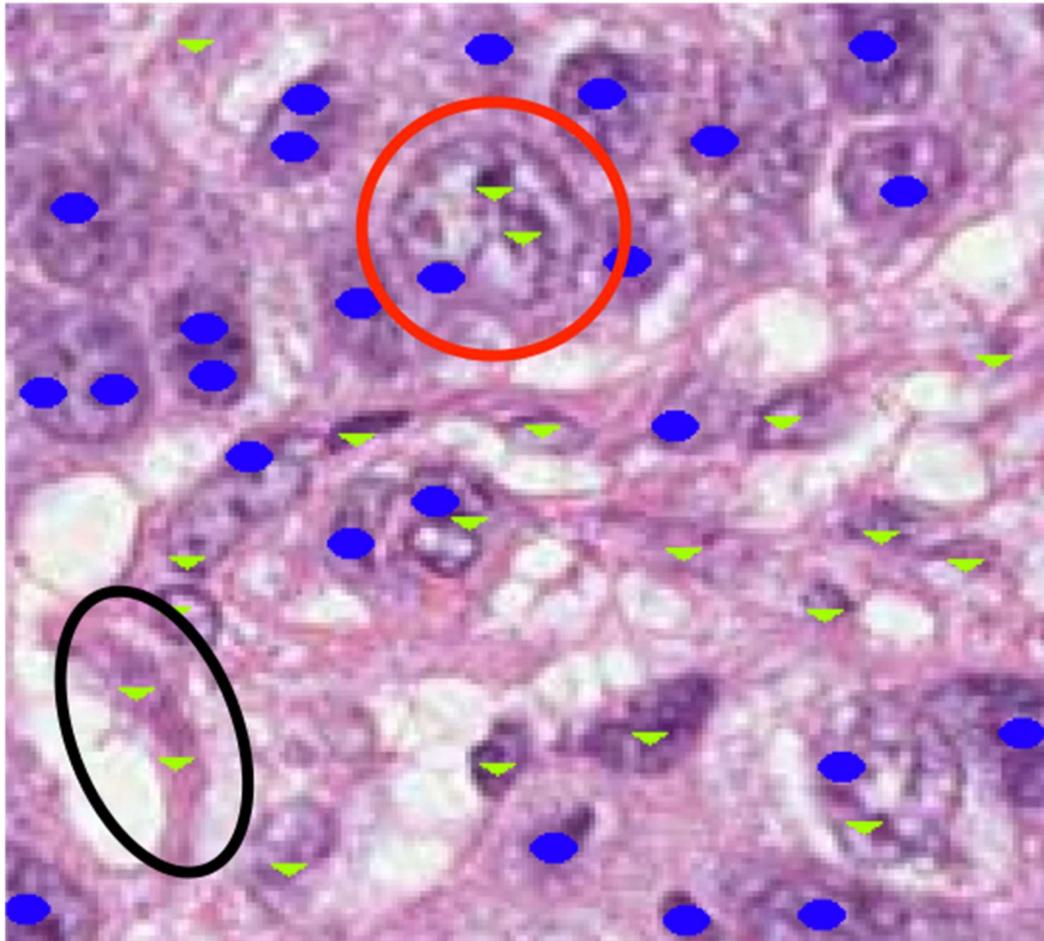


Figure 75: Examples of common segmentation errors.

This field was chosen for poor performance, with multiple nuclei double counted, and many consequently misclassified.

Blue dots were classified by the algorithm as tumour cells; green inverted triangles were classified as stromal cells

The red circle shows a single tumour cell nucleus identified as three due to the presence of nucleoli. Two of these objects have been classified as stromal cells and one as tumour, even though in reality they are all part of the same tumour cell.

The black ellipse shows a long fibroblast nucleus being identified as two and consequently double counted.

Another difficult case is cells, almost exclusively tumour cells, that are undergoing mitosis or apoptosis. Unlike the debris left by necrosis, mitosis and apoptosis often result in dark, dense nuclear material that are of comparable size and shape to that of a cell nucleus, confusing the segmentation algorithm.

Another difficulty for both algorithms is tumour cell nuclei. Firstly, as mentioned above, the presence of nucleoli often leads to a tumour nucleus being identified as multiple nuclei, and occasionally being mis-classified as a lymphocyte due to the size and shape of the nucleoli. Secondly, some tumour cells have vesicular nuclei, which means that the nucleus is light in colour, and does not have a solid area for object detection – this can result in it being missed entirely by both algorithms (Figure 76).



Figure 76: Problematic vesicular nuclei during segmentation.

Two nuclei circled in yellow were not identified. Note also that three other tumour nuclei to the left and below were double counted due to stipple chromatin.

A particularly difficult challenge for cell classification algorithms is the distinction between tumour cells and normal breast epithelial cells. Here, in the setting of targeted biopsies and TMA sections that are selected for tumour cells, the amount of normal epithelial cells is usually very small and does not affect overall analysis. However, this distinction is vital when analysing sections from resection specimens, or in the assessment of IHC where such a distinction is required, such as when assessing ER status or quantifying Ki-67 proportion.

5.8.3 Potential solutions

Work is ongoing to refine the current classification methods by increasing the number of cells included in the training set (Banko and Brill, 2001; Halevy *et al.*, 2009). This increases the diversity in each class and should narrow the gap in performance between the cross-validation and independent test as the model becomes more able to generalise to new slides by virtue of being trained on more examples.

However, adding more input features, or columns (to a fixed number of examples) may increase overfitting. Overfitting is where features that may be either irrelevant or redundant may also, by chance, improve classification of the examples at hand, but may not generalise well to stains performed in other laboratories.

For the problem caused by nuclear mimics, such as mitotic bodies and nucleoli, the random forest classification method has some theoretical advantages, in that it includes an area surrounding the centre of the object. By varying the size of the area of interest, more information of the surrounding can be included in the analysis. Training sets can thus be generated to potentially merge multiple identified nucleoli and mitotic bodies, and exclude apoptotic bodies. This will require extensive pathologist input to create different training sets for each of the size variables, and subsequent validation exercises. However, the dramatic improvement in “Mark” classification performance in IHC data provides proof of concept that such a two-stage approach to artefact identification has the potential to yield superior results.

The identification of vesicular nuclei is a particular difficulty for both algorithms. In future, it may be possible to use neural networks and deep learning techniques for this purpose, but these methods remain in their infancy for digital pathology, require very large training sets, and may lack transferability to novel data (Khosravi *et al.*, 2018; Mormont *et al.*, 2018; Tizhoosh and Pantanowitz, 2018; Alom *et al.*, 2019; Niazi *et al.*, 2019).

5.9 Conclusion

Here I have demonstrated that automated cell classification is possible on both H&E whole slide images and IHC tissue microarrays, using image analysis methods developed in collaboration with astrophysicists (Tedds *et al.*, 2008; Walton *et al.*, 2010; Ali *et al.*, 2013).

While not perfect, overall H&E cell classification accuracies for both algorithms was good, even on completely novel data from a different trial, at above 72% for tumour and stromal cells, and especially high at 84% for lymphocytes. This forms the necessary basis for the next chapter, which uses these classifications to derive quantitative spatial metrics to assess the tumour microenvironment (Whiteside, 2008; Anderson and Simon, 2020) in a way that would not be feasible for a histopathologist to do manually.

6. Spatial analysis

6.1 Preface

This chapter builds on the cellular classification developed and validated in the previous chapter, to examine the spatial relationships between tumour clusters and their microenvironment, especially in relation to tumour-lymphocyte interactions.

I am grateful to Dr Mireia Crispin Ortuzar, who provided her machine learning expertise to the analysis strategy for section 6.4.3, Multivariate prediction of outcome. I am also grateful to Dr H Raza Ali and Dr A Dariush for the theory and implementation of the graph theoretic measures reported in section 6.6, Tumour cell neighbourhood analysis, on which I performed the subsequent statistical analyses.

The analyses was implemented in MATLAB version 2018b with the Image Processing Toolbox 10.3, Bioinformatics Toolbox 4.11, and Statistics and Machine Learning Toolbox 11.4.

Code for all of the analyses reported here is available at:

<https://github.com/copew/ImageAnalysis>

6.2 Background

This chapter continues my effort to develop novel, practical methods to accurately predict an individual's response to neoadjuvant chemotherapy. Previously, in Chapters 3 and 4, I examined the gene expression profiles of tumours and their surrounding cells as a whole. This, however, does not allow the analysis of the spatial relationships that define and constrain the interactions of tumour cells with their environment and, in particular, immune cells.

Breast cancer is well known to be a heterogeneous disease and exhibit intra- and inter-tumoural heterogeneity (Ellsworth *et al.*, 2017), which makes the clinical management all the more challenging (Ng *et al.*, 2012; Zardavas *et al.*, 2015). Immune cell gene expression has been shown to have an important prognostic role (Ascierto *et al.*, 2012), which interacts with lymphocyte infiltration and hormone receptor status (Calabrò *et al.*, 2009). It is imperative, therefore, to study the spatial relationship between the immune cells and the tumour cells in breast cancer in more detail.

These spatial features are well represented on histology slides, providing the potential to complement the findings of gene expression. In Chapter 5, I laid the groundwork for these methods by describing and validating methods to identify individual cells and classify them into tumour, lymphocytes and stromal cells. The identity of each of these identified cells was stored at a precise set of Cartesian coordinates. This allows for the spatial relationships between tumour cells and immune cells to be examined robustly, objectively, and quantitatively in a much larger number of samples than would be feasible with manual methods.

6.2.1 The interaction of tumour cells and immune cells

Tumour infiltrating lymphocytes (TIL) have been recognised as an important prognostic factor in many solid tumours (Hendry *et al.*, 2017). For melanoma, the assessment of TILs form part of the routine histological report. Clark Jr *et al.* (1989) developed the system that is currently used in UK clinical practice, which suggests that a “brisk” lymphocyte infiltrate confers a better prognosis than a paucity of lymphocyte infiltration. Another commonly assessed tumour is colorectal carcinoma (CRC), where TILs have been shown to have prognostic value in all subtypes (Rozek *et al.*, 2016) and it has been proposed to add an immune score to the standard TNM staging (Galon *et al.*, 2014). Similar relationships with outcome and proposals for clinical measures have been made for lung cancer (Donnem *et al.*, 2016), and ovarian cancer (Zhang *et al.*, 2003).

However, in breast cancer the picture is more complex, and the prognostic value of TILs varies depends on the subtype of the tumour and the chemotherapeutic method (Salgado *et al.*, 2015b). In the adjuvant setting, the presence of TILs is good prognostic marker for *triple-negative* (TNBC) and *HER2*-positive breast cancer, where numerous large trials have suggested that a high presence of TILs is associated with better disease free survival in both of these subtypes (Adams *et al.*, 2014). TILs were also reported as an independent prognostic factor for disease free survival from early *HER2*+ breast cancer patients treated with adjuvant chemotherapy and trastuzumab in the randomized shortHER trial (Dieci *et al.*, 2019).

In the neoadjuvant setting, Denkert *et al.* (2018) found that increased lymphocytic infiltrate in the stroma, rather than inside the tumour, predicted response to neoadjuvant chemotherapy in all molecular subtypes assessed (TNBC, *HER2*-positive, and luminal type *HER2*-negative). They also confirmed that a high TIL was associated with a survival benefit in *HER2*-positive breast cancer and TNBC, but worse prognosis for luminal *HER2*-negative cancers. The predictive power of lymphocyte infiltration has been widely verified in TNBC (Herrero-Vicent *et al.*, 2017; Gao *et al.*, 2020a) and *HER2*-positive disease (Liu *et al.*, 2015; Salgado *et al.*, 2015a). However the importance of TILs in luminal subtypes remains the subject of debate; while some authors do show

relationships with pathological complete response (pCR) (Al-Saleh *et al.*, 2017), meta-analysis has not confirmed an association (Gao *et al.*, 2020b).

Ali *et al.* (2016) identified that median lymphocyte density across a slide correlated with pCR in a cohort with mixed molecular subtypes. In Chapter 5, I replicated this finding, and showed that the effect was particularly significant in *HER2*-positive tumours. When this analysis is performed on a biopsy slide, in many cases the tissue on the slides includes very little normal tissue. This means that the median lymphocyte density of the slide is very similar to the combined lymphocyte density of the tumour region and peri-tumoural region. However, this is not always the case, as in some biopsies there is significant amount of normal tissue present, and the median density of the whole slide may not be the most appropriate measure, as it may mask individual differences in the distribution of these immune cells (Figure 77).

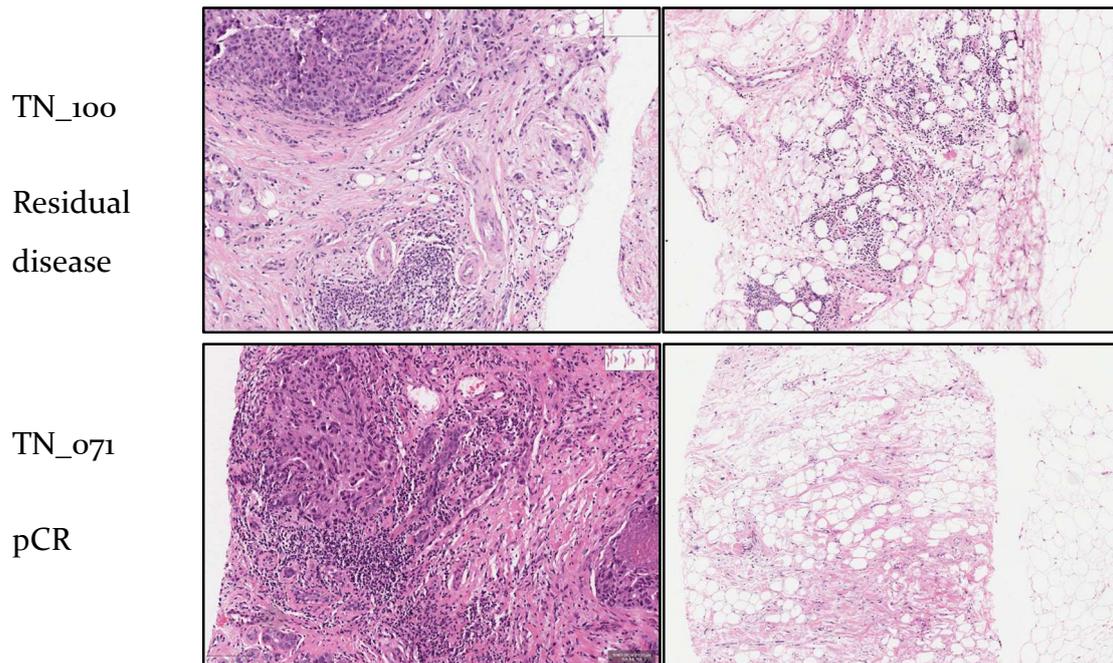


Figure 77: Examples of patients with similar median lymphocyte density across the whole slide, but showing very different distributions of lymphocytes and different treatment outcomes.

These are images from core biopsy (before neoadjuvant chemotherapy, NAT) from patients TN_100 (top row) and TN_071 (bottom row). The left two panels show tumour clusters, and the right two panels show normal tissue, which is mostly fat. In patient TN_100 there was minimal lymphocytic infiltrate within and around the tumour (top left), but there were a moderate number of lymphocytes infiltrating normal adipose tissue (top right). This patient had residual disease after NAT. In contrast, patient TN_071 showed lymphocytic infiltrate only around tumour cell clusters (bottom left), but not in normal tissue (bottom right). This patient had pathological complete response to NAT. These two patients had very similar median lymphocyte density across the whole biopsy, masking significant differences in spatial distribution.

6.2.2 The importance of spatial relationships between tumour and immune cells

The spatial context of a tumour in relation to immune cells is increasingly recognised as being prognostically important across a range of cancers (Galon *et al.*, 2006; Fridman *et al.*, 2012). In lung cancer, Enfield *et al.* (2019) showed that the spatial relationships of immune cells and tumour cells was more important than their density. CD8⁺ T-cells were particularly associated with disease free survival when they were completely surrounded by tumour cells, indicating infiltration.

Beyond TILs, the interface between tumour and normal tissue, often known as the invasive front, seems to be a particularly important area, and can reveal relationships that are obscured when the slide is considered as a whole. In colorectal cancer, Idos *et al.* (2020) noted a differential impact from the location of CD3⁺, CD8⁺ and Foxp3⁺ T-cell sub-populations. Better outcomes were observed when these sub-populations were within tumour clusters or at the invasive front, but not outside of these regions. The invasive front also seems to be important for treatment response. Gong *et al.* (2019) showed that, in colorectal tumours, the density of CD8⁺ T lymphocyte clusters in the tumour invasive front was significantly higher in patients who responded to PD-1 blockade treatment. This effect was not seen when the density of the lymphocyte clusters was compared across the slide as a whole. Their study highlighted the importance of spatial heterogeneity not only in the tumour, but also in the immediate environment adjacent to tumour.

In breast cancer, Heindl *et al.* (2018b) showed that the spatial arrangement of immune cells as associated with recurrence free survival after endocrine therapy for *ER*-positive tumours, even when their overall abundance was not. Mi *et al.* (2020) used five different immune markers to study the tumour microenvironment in TNBC. They identified that the invasive front of the tumour tended to have a higher density of immune cells and cell clusters, compared to either the centre of the tumour or normal tissue. Both intra- and inter-tumoural heterogeneity in the distribution of all five immune markers was

seen across all specimens, suggesting a universal role for the invasive front in determining tumour-immune system interactions.

These studies all highlight the importance of building on the whole-tumour or whole-slide relationships reported in Chapters 3-5, by looking specifically at the spatial interactions between tumour cells and immune cells. Immune cells that are in the tumour cluster, or immediately around them, appear to be particularly strongly associated with outcome and treatment response.

6.3 Aims and Objectives

The aim of this chapter is to use automated methods to determine whether the spatial profile of tumour cells, immune cells and stromal cells is correlated with response to NAT.

Motivated by the literature from a range of solid tumours, my overall hypothesis was that the presence of both tumour-infiltrating lymphocytes and dense clusters of lymphocytes in the peri-tumoural regions would associate with outcome from neoadjuvant chemotherapy. I assessed these hypotheses at a number of spatial scales.

First, I identified tumour clusters and demarcated regions of tumour, tumour front and non-tumoural tissue, from which I extracted and compared spatial features of potential interest.

Second, I used graph theoretic analysis to assess the interactions between communities of tumour cells, lymphocytes and stromal cells.

Finally, I assessed the specificity of my results for tumour clusters and the tumour interface zone, by examining how the spatial relationships of individual tumour cells influenced outcome.

6.4 Identification of tumour clusters

In order to identify intra-tumoural and peri-tumoural regions, I first identified clusters of tumour cells. Objects found to be within the resultant boundary of a cluster were then defined as intra-tumoural. An area outside, but close to the boundary of, a cluster was then defined as peri-tumoural (the “interface” or “invasive front”), and the area that is not included in either of these regions was defined as normal tissue. Detailed methods are given in Chapter 2, and all analysis code is available at <https://github.com/copew/ImageAnalysis>; here I briefly explain the process occurring at each step and the rationale behind it, with the aim of creating a coherent narrative. The overall aim was to define every cell on a slide as belonging to either a single tumour cluster, one or more peri-tumour zones, or normal tissue.

6.4.1 Method overview

To ensure uniformity of tissue sampling technique and patient treatment I performed clustering analysis only on the biopsy slides from TransNEO trial (the largest study in my dataset) that were all stained with H&E in the same lab, and that we know showed excellent cross-validated cell identification accuracies well in excess of 80%. To ensure that all patients were weighted equally in analysis, I selected one representative slide of breast biopsy from each patient.

Biopsy slides comprise multiple tissue cores, so the first analysis step is to delineate each core. This is important because it is not meaningful to look at the distance between a tumour cluster in one core and a lymphocyte in another. Once I had identified cores, I divided the cell catalogue into subsets so that spatial relationships were not mistakenly calculated across cores (Figure 78). To maximise data use and reflect sampling heterogeneity, all cores were analysed, and the results aggregated to give an overall measure from the slide for each patient.

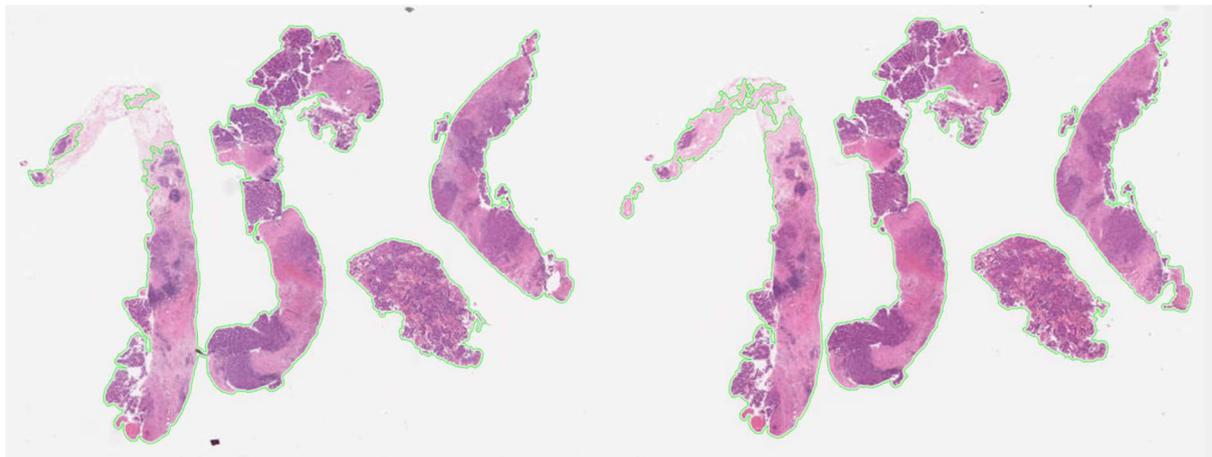


Figure 78: An example slide comprising multiple tissue cores, correctly identified as separate objects as indicated by the green boundaries.

Next, I used the co-ordinates of each cell in the segmented catalogues (Figure 79) to calculate the two-dimensional Euclidean distance between every cell pair in each core using the *pdist2* MATLAB function. With this information, I identified tumour clusters using my own implementation of the density-based spatial clustering of applications with noise algorithm (DBSCAN) (Ester *et al.*, 1996) (Figure 80). Tumour clusters were defined as comprising a minimum of five cells, each within an algorithmically defined maximum distance of at least one other cell in the cluster. In this way, my method was robust to occasional misclassification of tumour cells, as five cells in close proximity were required for tumour cluster membership.

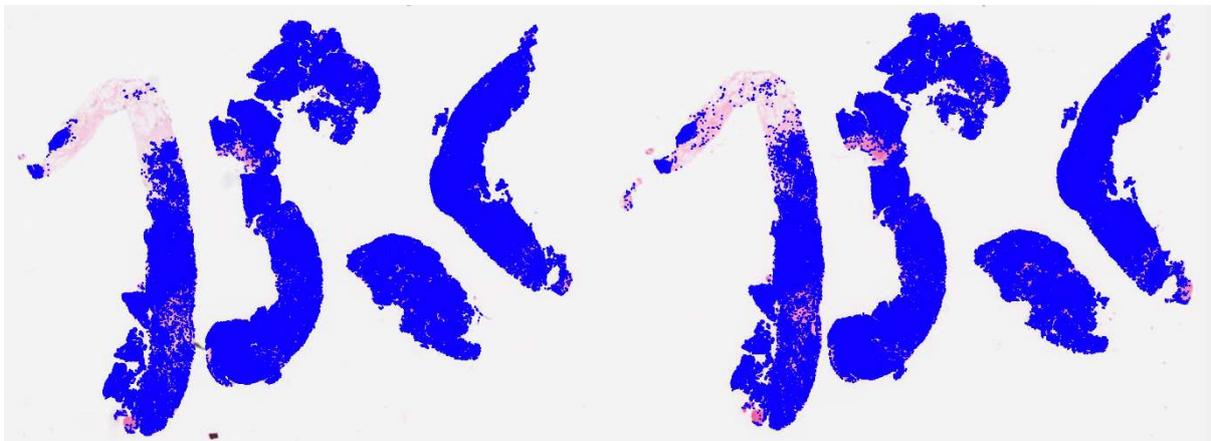


Figure 79: The spatial location of all cells identified by segmentation.

Overlaid as single blue dots on the low-magnification slide image as a visual check.

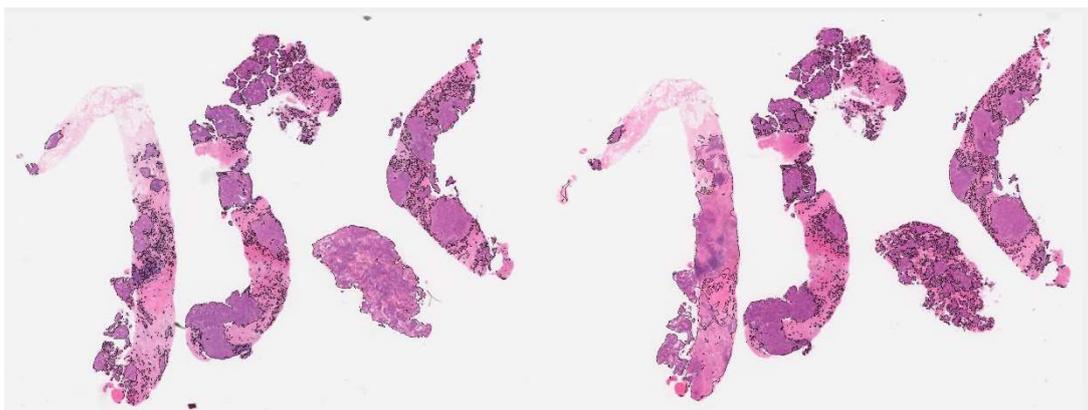


Figure 80: All tumour clusters, identified by my implementation of the DBScan algorithm.

Each tumour cluster is surrounded by a black line. Most tumour cells in this slide are within large, well-demarcated clusters, but there are some sub-sections of the image where there are a large number of small clusters. Although these small clusters appear almost as dots at this low magnification, each contains at least five tumour cells.

Next, I defined the “buffer zone”, or peri-tumoural region, of a cluster by forming a ring shaped polygon of fixed width from the edge of the cluster (Figure 81). Unfortunately, there is no consensus about the optimal size of this buffer zone in the literature. For example, Mlecnik *et al.* (2016) defined the peri-tumoural regions in colorectal cancer to be within 1mm of tumour clusters, without giving a particular reason. In our slides, 1mm would have been far too large a distance, and would have resulted in almost the entire slide being defined as either tumour cluster or peri-tumoural region, with significant overlap between the two.

We were primarily interested in close interactions between the tumour invading front and the host immune system, so wished to restrict our analysis to lymphocytes that were most likely to be directly interacting with tumour cells. From visual assessment across multiple slides, there was consensus amongst two histopathologists (myself and Dr Provenzano) that the optimal buffer width was 50 microns (100 pixels in our images), which is roughly a quarter of the diameter of a high-power field (x400), or the diameter of seven lymphocytes. I explored other distances, including 25 microns (50 pixels) and 100 microns (200 pixels). At 25 microns, some tumour clusters clearly had a “thicker” rim of surrounding lymphocytes that were not entirely captured within the buffer zone, which were not distinguished by the algorithm from those clusters with a much “thinner” rim of lymphocytes. It also created difficulties with the methods for calculating lymphocyte density. Two methods were implemented, one was total number of lymphocytes divided by area, while the other was using kth nearest neighbour (knn) method. The former gives an average density for the peri-tumoural zone but can be skewed by dense localised lymphocyte clusters. The knn method takes into account the heterogenous distribution of lymphocytes by calculating the distance from any lymphocyte to its 5th, 10th, 15th, 25th or 50th nearest neighbour. When the peri-tumoural zone, normally in the shape of a ring, was very thin, this often measured the distance to a lymphocyte on the opposite side of the “ring”, making the result meaningless as a measure of density. Conversely, for 100 microns (200 pixels), there were a large overlap between the peri-tumoural zones of adjacent clusters, and it was not possible to preferentially assign the overlapped area to any particular cluster. While this definition of the peri-tumoural zone seemed less optimal than 50 microns, I did run

a confirmatory analysis with arbitrary assignment of tumour cluster membership to avoid overlaps, and my results were not significantly altered.

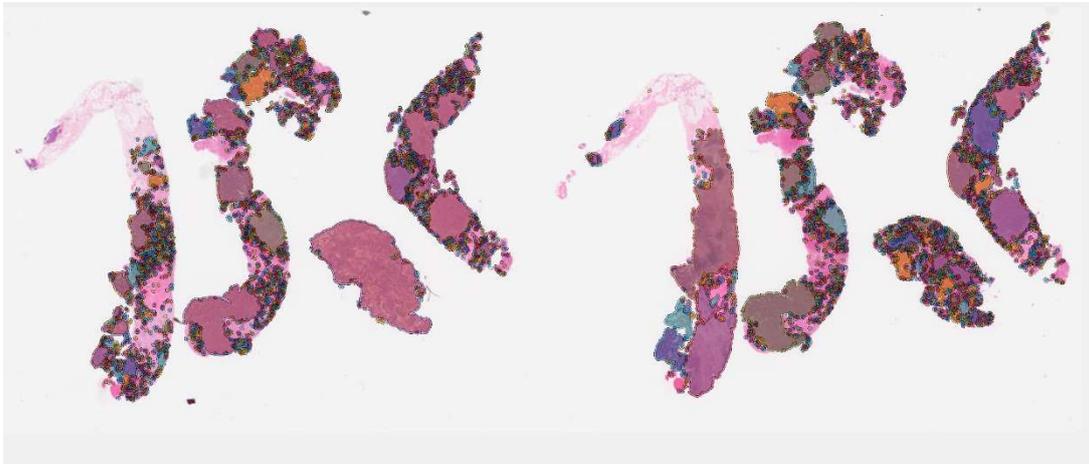


Figure 81: All tumour clusters, with their surrounding buffer zones.

For visualization, each cluster and buffer zone has been shaded with a different random colour. Areas of normal tissue are unshaded.

In some cases, the buffer zone of nearby tumour clusters overlapped. When this occurred, those regions were defined as belonging to the buffer zone of both tumour clusters (Figure 82).

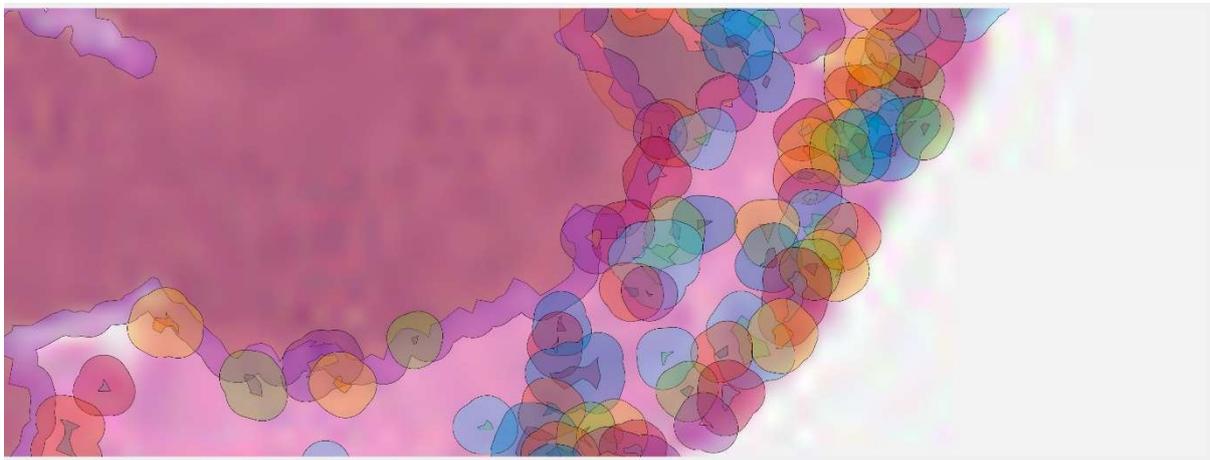


Figure 82: Overlapping buffer zones for neighbouring tumour clusters before trimming.

For visualization, each cluster and buffer zone has been shaded with a different random colour. Areas of normal tissue are unshaded. In the top left of the image, a large tumour cluster can be seen (shaded mauve), surrounded by its peri-tumoural region (shaded purple). In the bottom right of the image, a large number of small tumour clusters can be seen, along with their peri-tumoural regions, each shaded a different random colour. Areas could belong to the peri-tumoural region of more than one tumour cluster (indicated by multiple overlapping shades).

Although peri-tumoural regions were allowed to overlap, an area could not be defined as belonging to both a tumour cluster and to the peri-tumoural region of another cluster. In this situation, membership of a tumour cluster took precedence, and buffer zones were therefore trimmed to exclude areas of overlap with tumour (Figure 83).

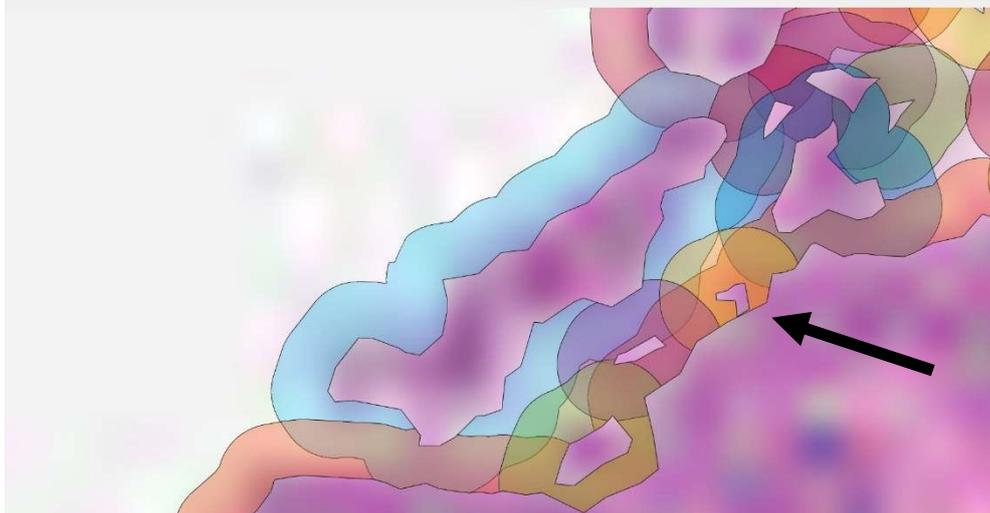


Figure 83: Tumour buffer zones shaded in random colours, trimmed to exclude areas of overlap with neighbouring tumour clusters.

Tumour clusters are now unshaded, and appear purple because of their dense haematoxylin staining. A large cluster can be seen in the bottom-right of the image and several small neighbouring clusters in the top right, each with a shaded buffer zone. Where the buffer zone would have overlapped with a tumour cluster it is clipped, for example, at the edge of the yellow shaded area indicated by the arrow.

All regions of the slide that were within a core but were not within either a tumour cluster or a buffer zone were defined as normal tissue.

Within each region I assessed the density and spatial distribution of tumour cells and lymphocytes. However, within-tumour heterogeneity also varies between individuals (Figure 84). Some patients show very consistent low (upper panel) or high (middle panel) peri-tumoral lymphocyte density, while others might have the same median density but show much more variability (lower panel). There are a large number of lymphocytes in peri-tumoural regions on each slide (median 9993, IQR 3861-26351), and taking the natural logarithm of the density of each approximates a normal distribution (Figure 84). Variability in this context can be quantified by the coefficient of variance, which is the standard deviation of these densities divided by their means.

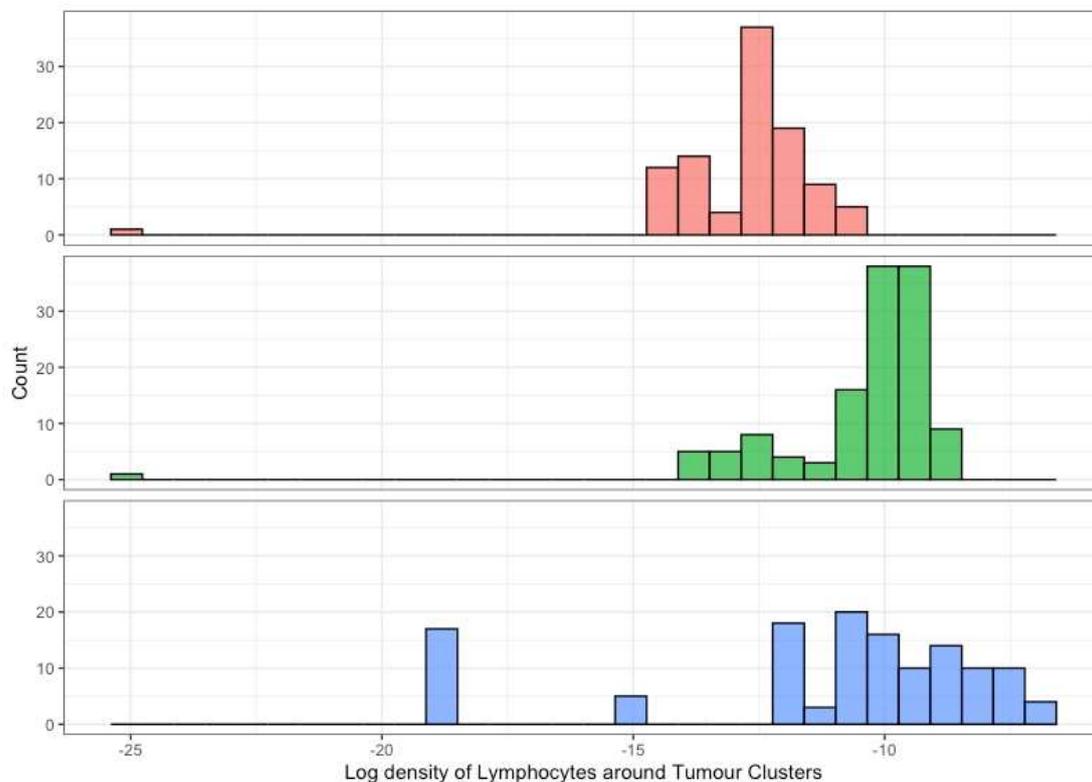


Figure 84: Example distributions of lymphocyte density in peri-tumoural regions across three slides.

The first slide differs from the other two slides in that it has a lower overall peri-tumoural lymphocyte density, while the third slide has the same mean as the second but differs in that it has a higher coefficient of variance.

6.4.2 Results

Distribution of tumour cells and cluster size

Although our main hypothesis relates to the relationship between tumour cells and lymphocytes, especially at the invasive front, we can begin by asking some simpler questions that might inform us about the underlying tumour biology, and later provide inputs to a multivariate prediction of outcome from neoadjuvant chemotherapy.

First, what were the characteristics of the tumour clusters themselves, and how did this relate to response to neo-adjuvant chemotherapy?

The number of tumour clusters per slide showed a wide distribution with median 715, IQR 460-1272 (Figure 85). These numbers were larger than might have been expected a-priori. Each slide contained 9 or fewer cores, meaning that each core contained dozens of tumour clusters. While this suggests that all slides had a large number of very small clusters, most tumour cells were nonetheless likely to be part of large clusters. The modal cluster size to which a cell was likely to belong being between 10000 and 16000 cells (Figure 86).

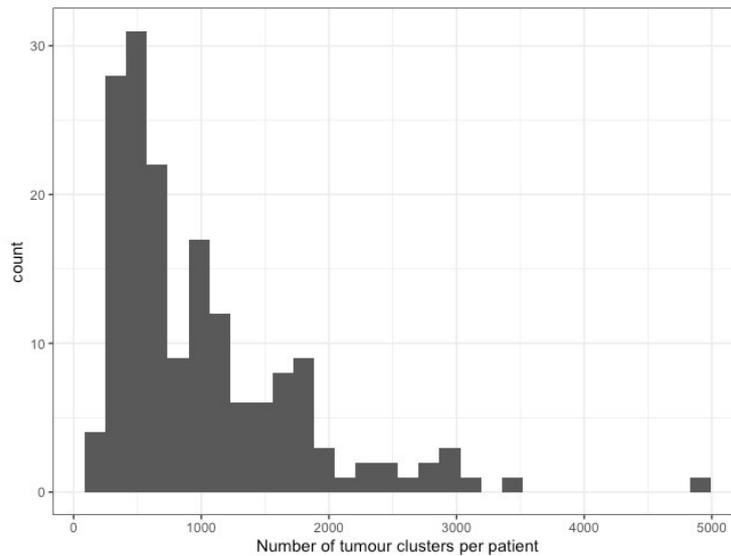


Figure 85: The number of tumour clusters identified per slide.

Each slide contained a large number of clusters, most of which were small.

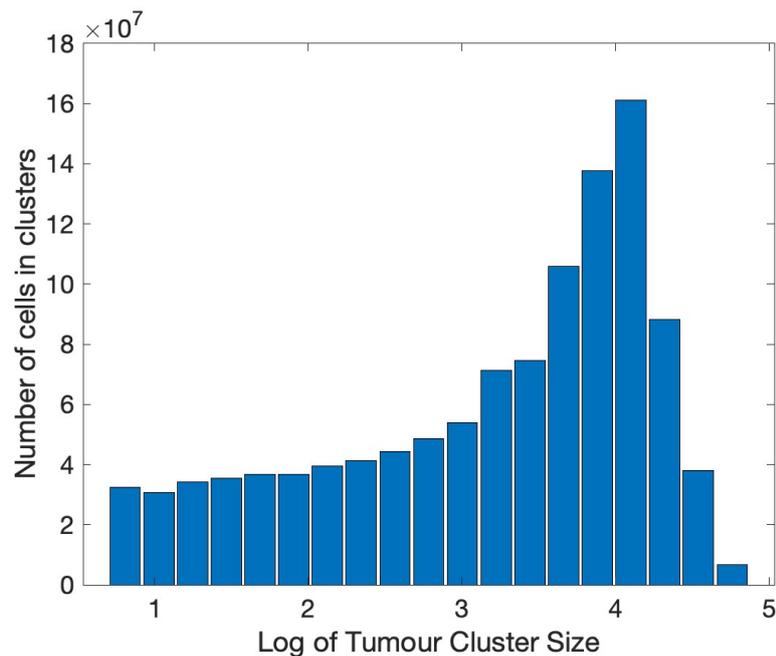


Figure 86: The number of tumour cells in clusters of each size, log base ten.

Although the majority of clusters were small, the majority of tumour cells belonged to very large clusters (modal cluster size between 10000 and 16000 cells with equally spaced histogram bins on a log scale).

In the patients who achieved pCR after NAT, the mean number of tumour cells per cluster was significantly higher (Wilcoxon $W(40,99)=2418$, $p = 0.042$) (Figure 87). This was not driven by trivial factors such as an overall difference in the number of tumour cells in the pre-treatment biopsy (Wilcoxon $W(40,99)=2086$, $p = 0.624$). There was no difference in the coefficient of variance (Wilcoxon $W(40,99)=2113$ $p = 0.538$). This would suggest that patients who achieved pCR had larger tumour clusters. This might seem counter-intuitive, but it may reflect that these tumours were more rapidly dividing, making them more susceptible to chemotherapy (Li *et al.*, 2016) – this possibility is considered in more detail in section 6.7.

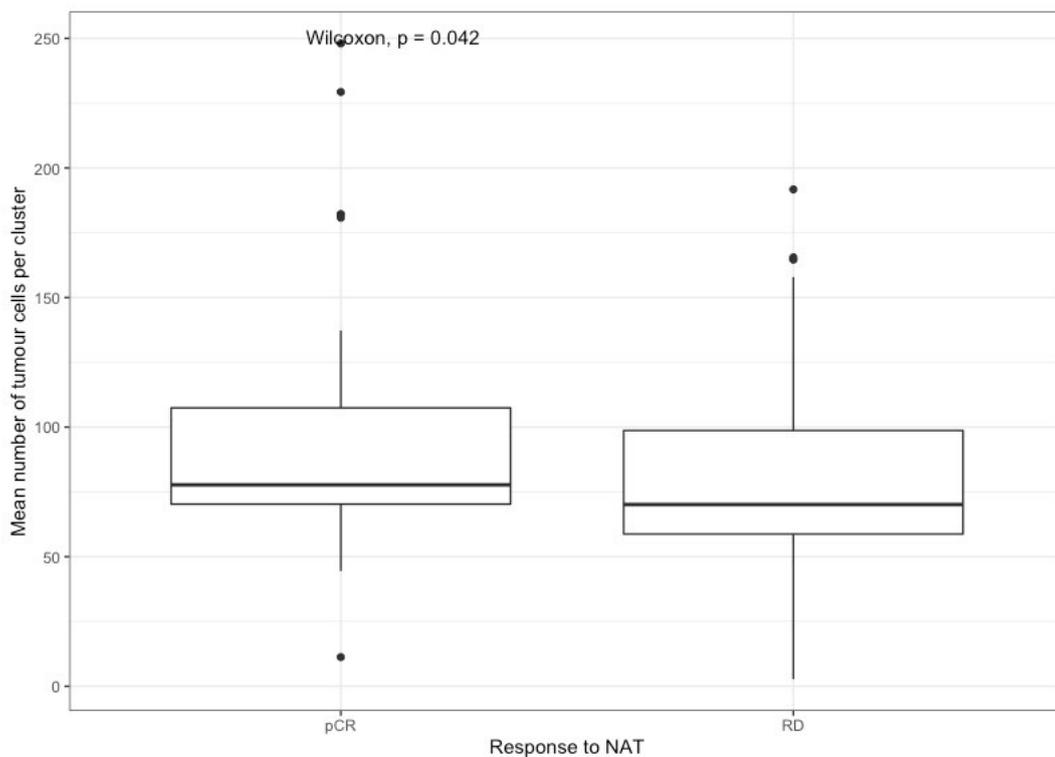


Figure 87: The association between tumour cluster size and response to neoadjuvant chemotherapy

Spatial distribution of immune cells

Our primary hypothesis was that the presence of dense clusters of lymphocytes in the peri-tumoural regions would associate with outcome to neoadjuvant chemotherapy.

To assess this hypothesis, I calculated the density of lymphocytes in tumour clusters and peri-tumoural regions with a “kth nearest neighbour” method, such that for every lymphocyte on the slide I calculated the distance between it and its k nearest lymphocyte neighbours that were within the same tumour cluster or buffer zone. I explored different values of k: the optimum was about 15 nearest neighbours. Fewer than this, and the measure was not representative, being biased towards small groups of lymphocytes rather than true clusters. Above this, visually identifiable lymphocyte clusters were excluded by the algorithm, often as they occurred in the interface zone of smaller tumour clusters.

The median lymphocyte density, assessed as the 15th nearest neighbour distance, was significantly higher in the patients who reached pCR than in those with residual disease after NAT both within the tumour clusters (Wilcoxon $W(41,110) = 2797$, $p = 0.0178$) and around the tumour clusters (Wilcoxon $W(41,110) = 2888$, $p = 0.00420$) (Figure 88). The variability of lymphocyte density, as quantified by the coefficient of variance, did not differ by NAT response either within (Wilcoxon $W(41,110) = 1948$, $p = 0.228$) or around tumour clusters (Wilcoxon $W(41,110) = 1909$, $p = 0.196$).

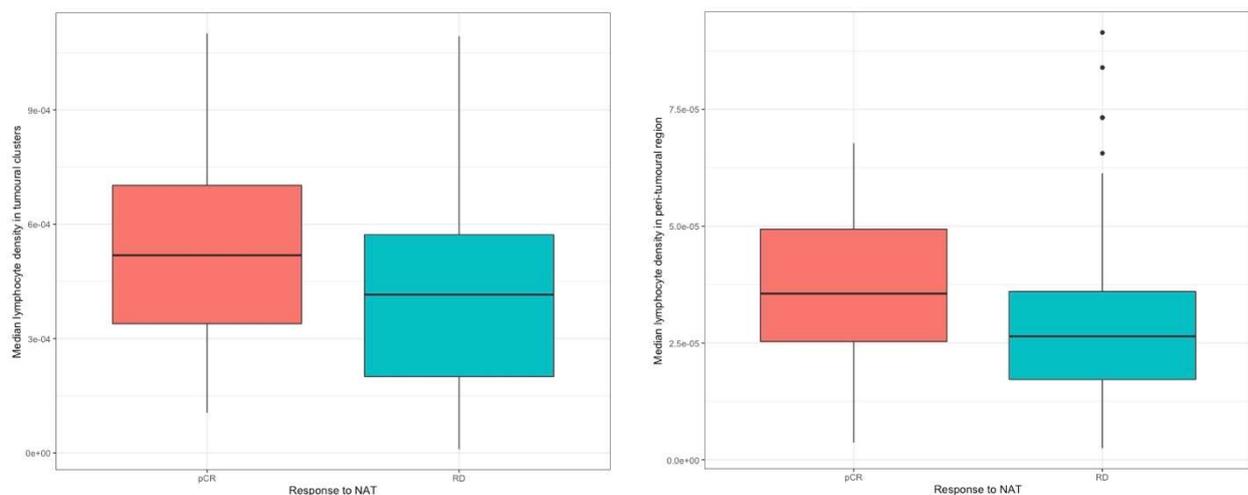


Figure 88: Median lymphocyte density in tumour clusters (left) and around tumour clusters (right), broken down by response to NAT

6.4.3 Multivariate prediction of outcome

Having verified our primary hypothesis we next explored our ability to use machine learning methods to predict outcome from a variety of spatial measures that could be derived from the digital images (Table 47), some of which we have already shown to be associated with response to neoadjuvant chemotherapy.

For each measure the median, mean, 10th percentile, 25th percentile, 75th, 90th percentile and standard deviation were calculated both within each tumour cluster and in the peri-tumoural region. Because we were particularly interested in median lymphocyte density, and the choice of 15 nearest neighbours to define a cluster was somewhat subjective, I also included median lymphocyte density calculated with knn method, with the values of k being 5, 10, 25 and 50. This resulted in a total of 38 highly correlated measures from within tumour clusters and another 38 from the peri-tumoural regions.

| Derived measure |
|------------------------------------|
| Tumour cell count |
| Lymphocyte cell count |
| Lymphocyte count per pixel squared |
| Lymphocyte density for knn, k=15 |
| Size in pixels squared |

Table 47: Derived measures from which outcome prediction was attempted

As a first step, I performed three principal component analyses (PCA) (in tumour clusters, in peri-tumoural regions, and combined), to reduce dimensionality while partitioning variance as independently as possible. Taking the elbow of the scree plot, four components survived as explaining more than 5% of the variance in all PCAs (Figure 89).

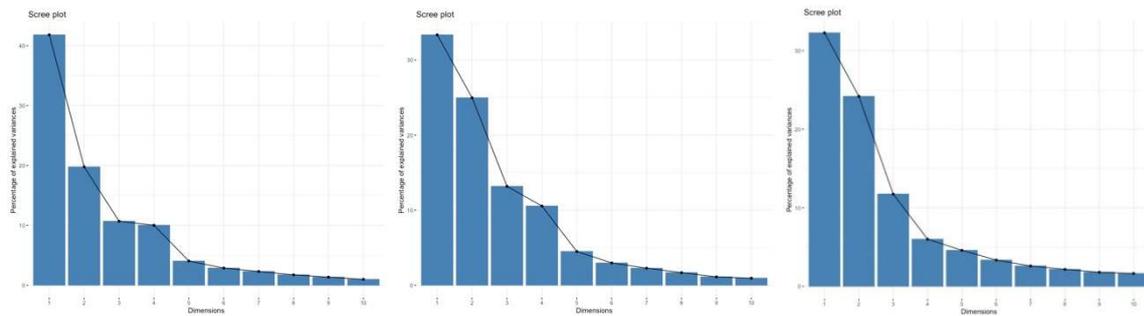


Figure 89: Scree plot from principal component analysis for derived spatial measures for peri-tumour regions (left), tumour clusters (middle), and both combined (right).

In peri-tumour regions, measures of lymphocyte count and density appeared almost orthogonal to measures of tumour cell count and peri-tumoural area in the first two components (Figure 90 top). By contrast, in the tumour itself, lymphocyte count was influenced more by tumour area, while lymphocyte density remained orthogonal to tumour cell count (Figure 90 middle). When tumour and peri-tumour regions were combined, there was much more spread in measures across components (Figure 90 bottom).

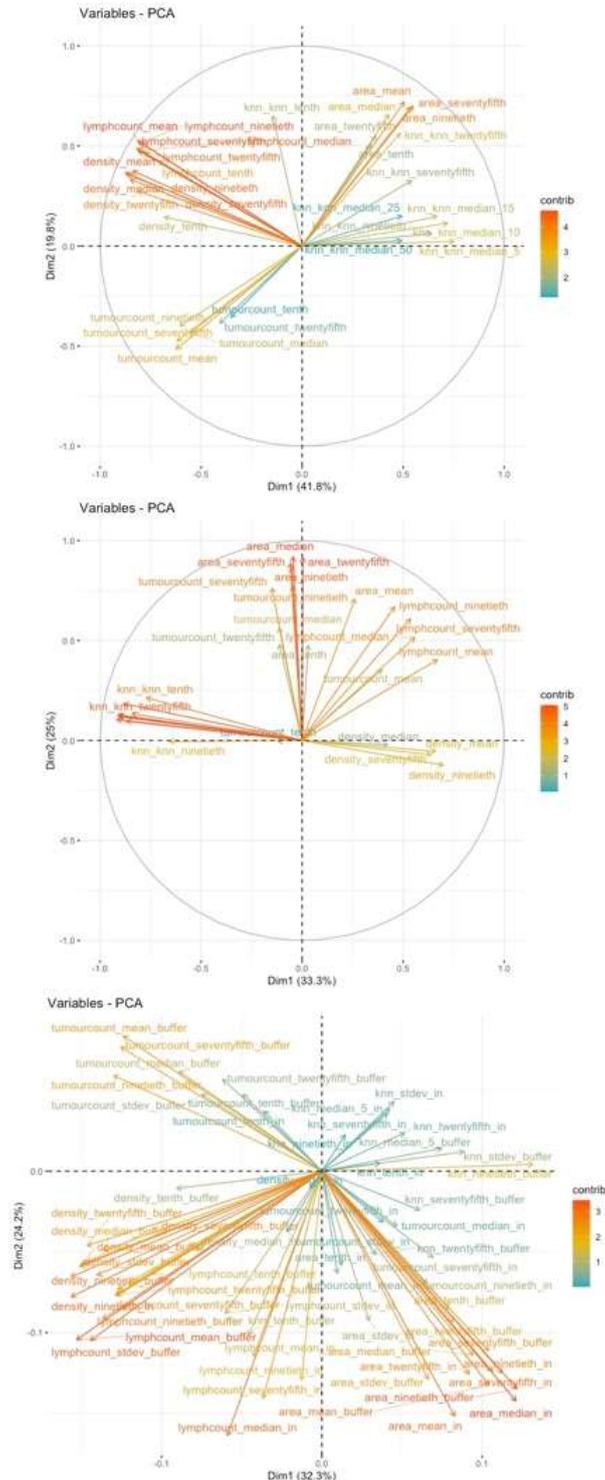


Figure 90: Factor loadings for the first two principal components in peri-tumour regions (top), tumour clusters (middle), and both combined (bottom).

In peri-tumour regions, measures of lymphocyte count and density appeared almost orthogonal to measures of tumour cell count and peri-tumoural area in the first two components (top). By contrast, in the tumour itself, lymphocyte count was influenced more by tumour area, while lymphocyte density remained orthogonal to tumour cell count (middle). When tumour and peri-tumour regions were combined, there was much more spread in measures across components (bottom).

Next, for each PCA I fitted a general linear model to predict NAT outcome as a binary response variable, pCR vs residual disease. I then performed ten repetitions of a ten-fold cross-validated analysis using 95 of the 151 cases as a training set and the remaining 56 as a test set, matched for NAT response.

One way of measuring the performance of such a prediction model is by calculating the cross-validated area under the curve (AUC) of its receiver operating characteristic (ROC) curve. This assesses the probability of a true positive against that of a false positive at various predictor thresholds. A perfect model would have an AUC of 1, producing only true positives. Chance-level performance is defined as an AUC of 0.5. Mean AUC for the peri-tumour region spatial features was 0.682, standard deviation 0.0314, for the within-tumour cluster spatial features was 0.629, standard deviation 0.0291, and for both combined was 0.669, standard deviation 0.0440. All regions were therefore able to predict NAT response better than chance.

A one-way ANOVA confirmed a difference in means between the peri-tumour AUC, within-tumour AUC, and combined AUC ($F(2,27)=6.05$, $p = 0.0068$). Post-hoc Tukey tests demonstrated that this was because the within-tumour cluster features provided less information than the peri-tumour region ($p = 0.0070$) or all features combined ($p = 0.043$). Combining all features did not improve AUC compared to using the peri-tumour region alone ($p = 0.723$).

Overall, therefore, while the spatial features of the tumour itself did predict NAT outcome better than chance, the peri-tumour region was more predictive. Combining spatial features from all of these regions did not provide additional information above the peri-tumour region alone.

6.5 Graph analysis

I next investigated whether an alternative way to measure cell interaction, graph analysis, would provide complementary insights into the spatial relationships associated with response to NAT.

Interaction between cells lies at the heart of tumour biology. Graph theory (West, 2001; Bollobás, 2013) is a branch of mathematics designed to quantify and analyse communication, which has becoming increasingly influential in biology over the last decade, especially in brain science where it can be applied to neuroimaging data to quantify large scale neuronal interactions (Stam *et al.*, 2007; Cope *et al.*, 2018). In cancer, it has been applied to cell-cell interactions to determine factors that influence metastasis in colorectal cancer and melanoma (Sirinukunwattana *et al.*, 2018; Failmezger *et al.*, 2020).

Graph theoretic measures rely on a matrix of connections, known as edges, between nodes, known as vertices. Here cells were vertices, and were defined as connected by edges if their nuclei were within 20 microns of each other, a distance of approximately 2-3 lymphocyte diameters, or one tumour cell diameter. This resulted in a relatively sparse graph that was not completely connected (i.e. some cells have no close neighbours with which they may interact). It is worth noting at the outset that absolute distance measures like this are perhaps not the optimal method of defining cells as being in communication, as they are confounded by cell size, but they represent a compromise approximation to make the problem computationally tractable. For example, in piloting, I attempted to implement an alternative method known as a Voronoi diagram (Kise *et al.*, 1998) where the slide is partitioned into tiny regions, each of which represents an approximation of the boundaries of a cell. Communication is then defined as occurring between cells that are in physical contact. However, this proved computationally intractable with currently available hardware and software – in a single day a single computer core was able to compute the diagram for less than 1% of a digital

slide image, meaning that even with efficient parallelisation on a supercomputer it would have taken months to compute the association matrices across all slides.

Five types of interaction, or more precisely spatial arrangement of cells, were examined, each of which I have termed a 'compartment'. The first three compartments were for the relationships between cells of the same type (homotypic interactions), specifically epithelial (which were largely tumour cells), lymphocyte and stromal cells. The fourth compartment was all cells within the buffer-zone around each tumour cluster, which is here termed the interface. Finally, the fifth compartment was the non-tumour cells that fell within the tumour cluster boundary, here termed the tumour microenvironment (TME).

Within each of the five compartments, for each of 166 slides separately, six graph theoretic measures were calculated in igraph (Csardi and Nepusz, 2006), resulting in a total of 4980 datapoints. The graph theoretic measures were:

1. Number of vertices: this is simply the number of cells in the compartment being examined, each acting as a vertex in the connectivity graph.
2. Number of edges: this is the number of connections that any vertex has, or that a graph has overall.
3. Transitivity: this measure is related to the clustering coefficient (Schank and Wagner, 2005), and measures the probability that any three vertices are connected. In other words, the probability that the neighbours of a node are neighbours of each other.
4. Diameter of each community: for this and the following parameters, Louvain community clustering was applied to each compartment of the graph to define cellular groups (De Meo *et al.*, 2011) (Figure 91). This is a hierarchical clustering method, which groups neighbouring cells into 'modules', and then neighbouring modules into larger communities. Its output differs most significantly from the DBSCAN clustering technique (section 1876.4.1) in that every cell is assigned to a community, even if it

is outside of a cluster, and large tumour clusters are sub-divided into smaller sub-communities.

5. Distance between the centroids of communities.
6. Density of communities.



Figure 91: An example of Louvain community detection for homotypic epithelial (tumour cell) interactions.

Cells within each community have been assigned a random colour.

Towards the top left of the image, several communities of densely packed cells can be seen, which would have been combined into a single tumour cluster with the DBSCAN method. Below this, in the middle section of the left-most core, cell communities are much more sparse – these cells would likely have been deemed either unclustered or comprising many small clusters by the DBSCAN method.

As a first step, I examined the independence of these graph theoretic measures. They were highly correlated within each microenvironment (Figure 92). The mean number of edges and vertices, community diameter, and inter-community distance all correlated significantly, with r -values uniformly greater than 0.8 between edges and vertices and between diameter and distance. Community density was strongly negatively correlated with these four measures. Transitivity showed a variable

relationship with the other measures, being positively correlated only in the homotypic lymphocyte compartment.

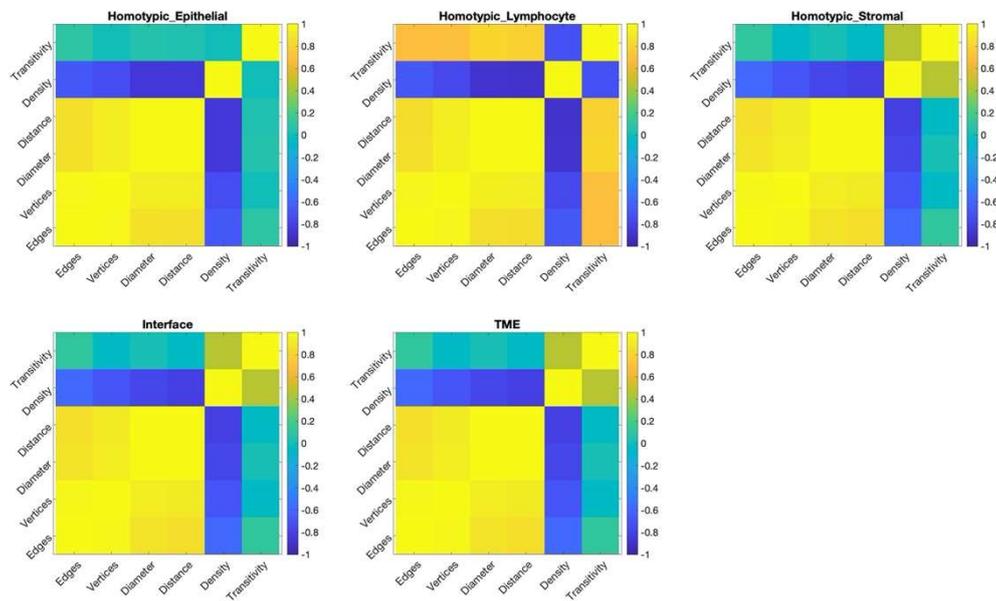


Figure 92 Pearson correlation between derived network measures within each environment.

Colour coded on a scale of -1 to 1.

Each panel corresponds to one compartment, and each row and column to a graph theoretic measure. The association matrices are symmetrical, with correlations of 1 along the diagonal.

Next, I performed principal component analysis within each environment to confirm that these correlations resulted from the measures explaining similar data features. In all environments, a first principal component made up of edges, vertices, diameter, distance, and the inverse of density explained at least 65% of the variance in the data (Table 48). Similarly, in all environments a second principal component made up of transitivity explained more than half of the remaining variance. The third and subsequent principal components all explained less than 10% of the variance and were made up of variable metric combinations.

| | Epithelial | Lymphocyte | Stromal | Interface | TME |
|------------------------------|---------------|---------------|---------------|---------------|---------------|
| Principal Component 1 | | | | | |
| Edges | 0.911 | 0.827 | 0.911 | 0.862 | 0.817 |
| Vertices | 0.949 | 0.875 | 0.959 | 0.989 | 0.938 |
| Diameter | 0.968 | 0.907 | 0.973 | 1.000 | 1.000 |
| Distance | 0.965 | 0.899 | 0.973 | 0.982 | 0.995 |
| Density | -0.856 | -0.815 | -0.826 | -0.815 | -0.856 |
| Transitivity | 0.056 | 0.769 | -0.100 | 0.163 | -0.253 |
| Percentage Variance | 74.668 | 85.894 | 72.881 | 65.427 | 67.452 |
| Principal Component 2 | | | | | |
| Edges | 0.062 | 1.000 | 0.276 | 0.451 | 0.605 |
| Vertices | -0.044 | 0.746 | 0.105 | 0.239 | 0.348 |
| Diameter | 0.003 | -0.086 | 0.104 | -0.162 | -0.043 |
| Distance | -0.019 | -0.109 | 0.036 | -0.213 | -0.111 |
| Density | 0.066 | 0.764 | 0.469 | 0.511 | 0.484 |
| Transitivity | 1.000 | -0.885 | 1.000 | 1.000 | 1.000 |
| Percentage Variance | 16.756 | 7.352 | 20.754 | 21.638 | 22.520 |
| Principal Component 3 | | | | | |
| Edges | 0.767 | 0.319 | 1.000 | 1.000 | -0.694 |
| Vertices | 0.554 | -0.045 | 0.880 | 0.630 | -0.694 |
| Diameter | -0.174 | -0.164 | -0.533 | -0.546 | 0.511 |
| Distance | -0.202 | -0.280 | -0.559 | -0.614 | 0.499 |
| Density | 1.000 | 0.727 | 0.924 | 0.225 | -0.541 |
| Transitivity | -0.092 | 1.000 | -0.726 | -0.935 | 1.000 |
| Percentage Variance | 6.780 | 5.204 | 4.739 | 8.048 | 7.267 |

Table 48 Principal component analysis of mean centred, variance normalised data from each microenvironment. First three components shown.

All measures differed significantly according to the compartment in which they were measured (Figure 93). Some of these results are trivial, for example there were generally more tumour cells than lymphocytes, resulting in more edges and vertices in these compartments. However, some are potentially more interesting, such as the tendency for homotypic lymphocyte communities to be more compact (smaller diameter, lower distance) than any other compartment.

Because the graph theoretic measures were highly collinear, separate repeated-measures ANOVAs with Greenhouse-Geisser non-sphericity correction were performed for each measure independently. These all resulted in p-values below 2×10^{-28} , far surpassing any possible correction for multiple comparisons. Post-hoc Tukey-Kramer tests confirmed that homotypic epithelial environments (tumour-tumour interactions) differed from all other environments in all measures.

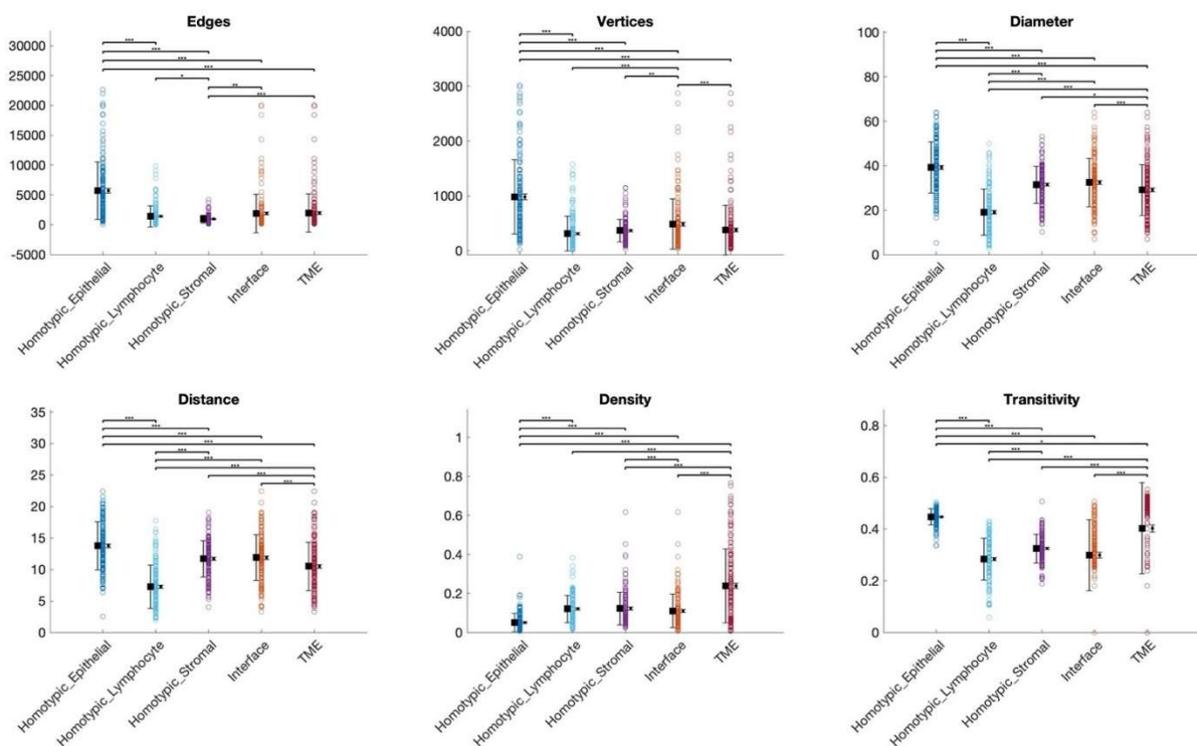


Figure 93 The effect of microenvironment on each measure.

All raw datapoints are shown, along with mean and standard deviation (to the left of the data points), and mean and standard error (to the right). Significantly differing microenvironments by post-hoc Tukey-Kramer tests are indicated by bars with $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

| Measure | DF | F | p-value | p-value (Greenhouse-Geisser) |
|--------------|--------|-----|---------------------|---------------------------------|
| Edges | 4, 660 | 94 | 3×10^{-63} | 3×10^{-37} |
| Vertices | 4, 660 | 104 | 2×10^{-68} | 5×10^{-45} |
| Diameter | 4, 660 | 159 | 3×10^{-95} | 1×10^{-75} |
| Distance | 4, 660 | 164 | 2×10^{-97} | 5×10^{-78} |
| Density | 4, 660 | 132 | 1×10^{-82} | 1×10^{-31} |
| Transitivity | 4, 660 | 81 | 5×10^{-56} | 2×10^{-28} |

Table 49 Repeated measures ANOVAs of the effect of microenvironment on each measure. DF: degrees of freedom.

I then looked to see whether any of these measures were associated with response to chemotherapy³. Eight measures differed between patients who had pathologically complete response and those that did not (Figure 94), using t-tests corrected for multiple comparisons with a Benjamini and Hochberg (1995) full data range (FDR) method. These were the number of edges and vertices in the homotypic epithelial (tumour-tumour) environment, and all six measures in the homotypic lymphocyte environment. No measures in any other environment significantly differed between pCR and RCB I-III, and no measures differed when combining pCR and RCB I disease and comparing this against RCB II-III disease.

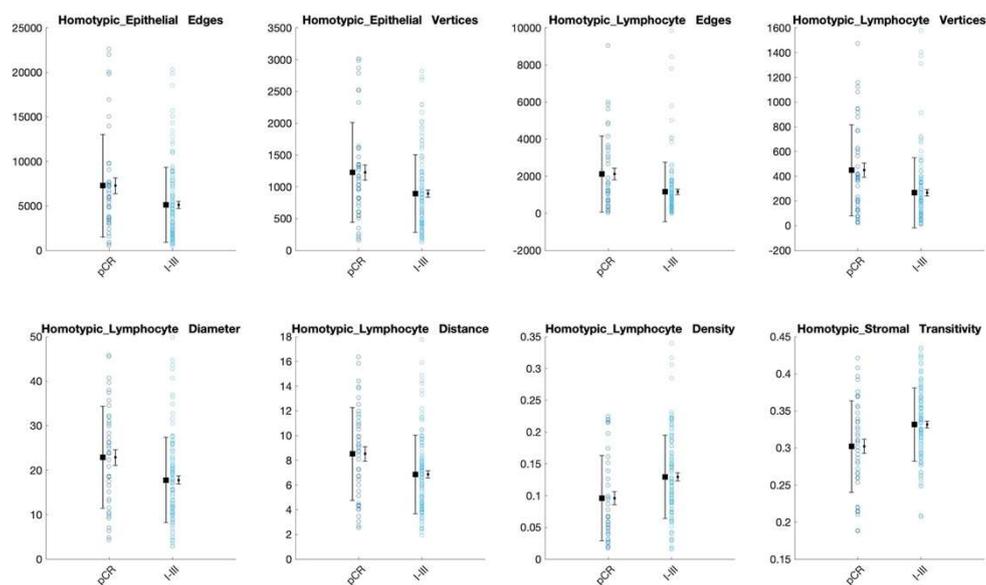


Figure 94 The eight measure and compartment combinations that differed significantly between patients who showed pathologically complete response and those who did not.

As in Figure 93, all raw datapoints are shown, with mean and standard deviation (to the left of the data points), and mean and standard error (to the right).

³ 42 patients (27%) had pathologically complete response (pCR), 25 (16%) had grade I disease, 65 (42%) had grade II disease, and 24 (15%) had grade III disease. The outcome for ten slides was unknown or not applicable (for example because the patient declined surgery).

6.6 Tumour cell neighbourhood analysis

So far, I have concentrated my analysis on the spatial relationships between tumour cells at the edge of clusters and lymphocytes in the interface zone. To examine the specificity of my results for this region, as a control analysis I next assessed the relationship between tumour cells and their local neighbourhood, regardless of their location within a cluster, or whether they belonged to a cluster at all.

To assess the relationship between a cell and its local neighbourhood, I first calculated the distance between a cell, C , and all other cells in the same core. I then sorted these distances from the closest to the furthest. For each cell type, I calculated the average proportion of tumour cells, lymphocyte and stromal cell in each position, and normalised this based on the proportion of each cell type overall in the slide. I then took a cumulative mean, such that I obtained a running ratio of observed neighbouring cell type frequency over expected neighbour cell type frequency that converged on one by the time all neighbours had been analysed (Figure 95).

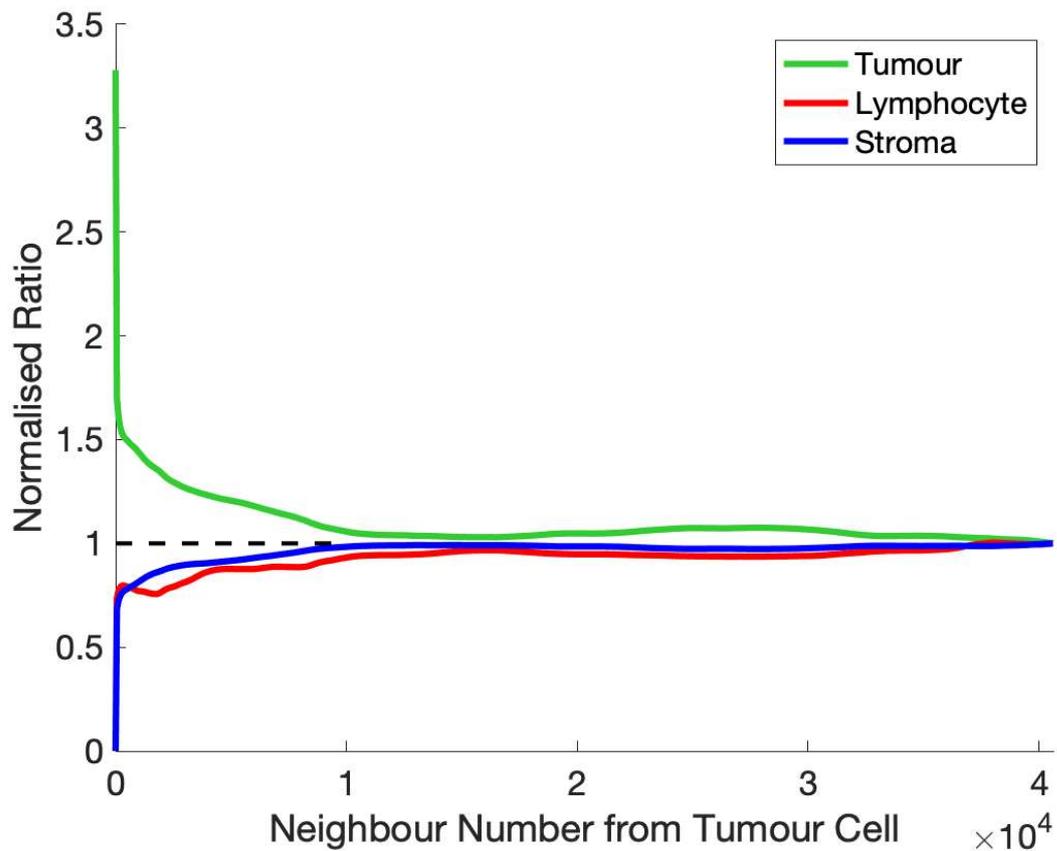


Figure 95: An example of the running likelihood ratio for tumour cell neighbours in a single core.

In this example, any tumour cell was most likely to have other tumour cells as its nearest neighbours, and indeed this was the case for all slides and cores overall. This is a somewhat trivial analysis, as we already know that tumour cells are most likely to exist as part of concentrated clusters, but it shows us that the strongest neighbourhood relationships are those closest to a cell, converging thereafter towards the average.

I analysed this ratio for the nearest 15 neighbours of tumour cells, as those most likely to be involved in cell-cell interactions. As shown in Figure 96, in all patients it was much more likely that tumour cells would have other tumour cells as close neighbours than stromal cells or lymphocytes. In those patients who had pCR to NAT, this effect was reduced, less likely to be close to tumour cells than in patients with residual disease. In other words, in those patients where a higher proportion of tumour cells were in close contact with other tumour cells, chemotherapy response was poorer. This might at first glance seem contradictory to the finding in (Figure 87), that larger tumour clusters were associated with pCR, however these measures are quite different. Note that the measure here is controlled for the proportion of each cell type overall in the slide, and is therefore controlled for differences in overall tumour volume. Additionally, my use of a non-parametric ordering method means that it is also controlled for differences in overall cell density. It is possible for a cluster to be large but heterogenous, or small and homogenous. And even in those slides displaying RCB after NAT the mean number of tumour cells in a cluster was more than 70, while here we are assessing the homogeneity of the nearest 15 neighbours.

Most tellingly for our purposes, the differences observed in tumour-lymphocyte neighbourhoods were not significant when the slide was considered as a whole, and the variance on these measures was wide, suggesting large variability in the likelihood that a tumour cell would have a lymphocyte as a near neighbour (Figure 96 middle). This stands in contrast to the prognostic effects we observed in median lymphocyte density of the whole slide in the previous chapter, and the spatial effects of lymphocyte density in peri-tumoural regions and graph theoretic measures of lymphocyte-lymphocyte interactions. Together, this indicates that dense groups of lymphocytes interacting with tumour cells in the interface zone of tumour clusters increase the likelihood of pCR more than lymphocytes scattered amongst tumour cells.

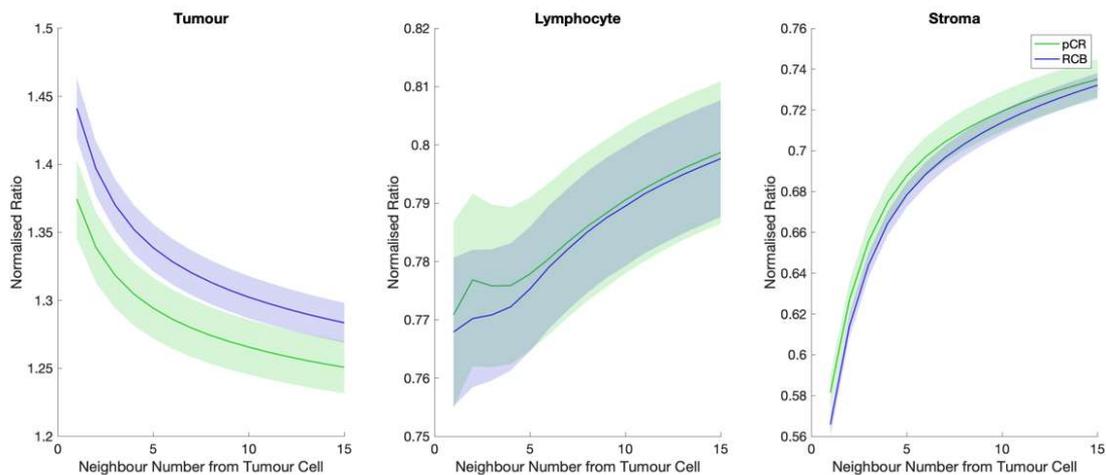


Figure 96: Tumour cell neighbours.

Separately plotted as a ratio of observed over expected tumour-tumour (left), tumour-lymphocyte (middle), and tumour-stroma (right) neighbours according to response to neoadjuvant chemotherapy.

The heavy line represents the mean across all slides and cores, comprising 39,626,516 cells in total, with the shaded area representing the standard error across slides. The normalised ratio can be interpreted as the ratio of the probability that any tumour cell will have that cell type as a neighbour in the position indicated on the x-axis.

Note on the left panel that tumour cells are more likely to have other tumour cells as neighbours than would be expected by chance, because they occur in clusters. This ratio is higher in those patients who had RCB after NAT than in those who displayed pCR.

In contrast, the probability of a lymphocyte neighbouring a randomly selected tumour cell does not significantly differ according to NAT outcome. There may be some difference at the level of first or second neighbour but because the probability is so variable when then the whole slide is considered this does not reach statistical significance and would have little predictive power.

6.7 Conclusion

In this chapter I evaluated a large number of novel spatial metrics, which are difficult for a histopathologist to quantify objectively, and which complement our finding in the previous chapter of better response to NAT in slides with higher median lymphocyte density. The methods I have employed are varied, each examining a different type of spatial relationship, but each revealed some statistically significant spatial determinants of response to neoadjuvant chemotherapy. These analyses are complementary, and together they reveal that there are two primary spatial determinants of NAT response.

The first determinant is the homotypic relationship between tumour cells themselves. Specifically, a patient is more likely to achieve pCR if their pre-treatment biopsy displays:

- 1) More tumour cells within each tumour cluster (Figure 87)
- 2) Larger, more connected tumour clusters (Figure 94)
- 3) Greater cellular heterogeneity within these large clusters (Figure 96)

This growth pattern can often be seen in high grade breast carcinomas, such as those with medullary features. In high grade breast cancer, the tumour cells typically grow in large, cohesive groups that display the three features above; large clusters containing many closely packed cells. This growth pattern is thought to be due to the high mitotic rate seen in the tumour cells. This is in contrast to low grade breast tumours, such as tubular carcinoma, which would display none of these features. Small groups of tumour cells forming individual tubules are often seen infiltrating the stroma, resulting in small tumour clusters, loosely connected in graph theoretic terms. Tumour grade is known to be an independent predictor of response to NAT. High grade tumours typically respond better than low grade tumours (Huober *et al.*, 2010), probably as a result of their higher mitotic rate making them more amenable to chemotherapy. Overall, therefore, this first determinant is likely identifying features of high grade tumours, but without explicitly assessing the same features as a pathologist (tubule formation, nuclear pleomorphism and mitotic rate). It is, instead, automatically abstracting the

associated spatial and architectural arrangements of the tumour cells, and their relationship to the surrounding stroma.

The second determinant is the relationship between tumour clusters and their surrounding lymphocytes, especially at the tumour front (interface zone). pCR was associated with higher lymphocyte density (Figure 88), and by larger, more densely packed lymphocyte communities (Figure 94). The association between lymphocyte density and NAT response was particularly strong in the interface zone, and was present to a lesser extent within tumour clusters (Figure 88).

My results verify the prognostic significance of tumour infiltrating lymphocytes (TIL) in the neoadjuvant setting, but emphasise the importance of the tumour invasive front, or interface (Mi *et al.*, 2020). My cohort was mixed in terms of receptor status and intrinsic sub-type, and it was too small to break down the results along these lines. It may be that the tumour invasive front is of universal importance, while TILs vary in their importance for predicting NAT outcome according to *HER2* (Liu *et al.*, 2015; Salgado *et al.*, 2015a) or *ER* status (Herrero-Vicent *et al.*, 2017; Gao *et al.*, 2020a), or be particularly relevant in only some intrinsic subtypes (Gao *et al.*, 2020b).

Using multivariate methods across all spatial metrics, I was better able to predict response to NAT from spatial metrics derived from the peri-tumoural region than from tumour clusters themselves. The AUC of these prediction was between 65% and 71%, which is insufficient for clinical applications, but provides proof of concept for the utility of spatial metrics, as well as giving mechanistic insight into previously observed trends in outcome related to lymphocyte density. The accuracy of the spatial metrics I have proposed here is limited by the accuracy of the cell classification described in chapter 5. While there are certain mitigations to reduce the effects of misclassification, such as the definition of a tumour cluster as comprising at least five cells, the metrics

could nonetheless potentially benefit from future improvements in cell classification. Perhaps more importantly, spatial relationships provide only one piece of the puzzle, and a multivariate predictor using only spatial metrics is inherently limited.

Ultimately, spatial metrics have the potential to be most powerful in combination with a wider variety of clinical, demographic and molecular features to enable cross-modal precision medicine. In the next chapter, I conclude the thesis by discussing these and other future directions.

7. Discussion

7.1 Clinical relevance of my thesis

Breast cancer is the most common type of cancer in women in the UK. Increasingly patients are being offered neoadjuvant chemotherapy (NAT) before surgery. Not everyone will respond to NAT, and being able to accurately identify the patients that will not benefit would not only avoid unnecessary side effects from the treatment, but also potentially allow them to be offered alternative treatments sooner. It would also allow better long-term prognostication, as those who achieve pathological complete response (pCR) have better prognosis than those who do not. All of my chapters are focussed towards the development of methods for predicting individual response to NAT that could be widely applicable in the real-world, facilitating a precision-medicine approach.

Three primary modalities have been shown to predict the response of a breast cancer to NAT; radiological imaging (Marinovich *et al.*, 2012), molecular profiling, and histopathology. Of these, only pathological assessment is in widespread clinical use, and even this is limited to conventional human judgments, rather than quantitative computational approaches. My thesis provides a first step towards facilitating the wider application of these techniques, by assessing the feasibility of molecular profiling and computational pathology on widely available samples.

As part of routine NHS care, tissue samples from breast cancer biopsies will be stored in FFPE blocks, and stained with H&E, and sometimes subjected to IHC. Increasingly, these slides are being routinely digitised for remote viewing.

As pathologists, we already provide information to aid, or sometimes even guide, clinical management. We are in the unique position to directly assess the characters of tumour, and the relationship between tumour and its surrounding. We already can acquire some characters of the tumour visually, such as tumour grade. With the help

of computational pathology and molecular classification, we hope to offer more accurate predictions of the benefits of treatment for each individual patient. Overall in my thesis I have shown that these methods are not yet sufficiently robust to be useful in clinical practice. However, the computational methods have provided novel insights into the spatial determinants of tumour biology, and show promise for inclusion in future multi-modality classifiers. Consequently, my thesis has spawned a large number of follow-up projects, which I will describe here.

7.2 Molecular classification

The stratification of breast cancers using gene expression from FFPE samples has proven to be very difficult. In this thesis I have showed that, using NanoString with the pre-selected probe set, I could not build a robust and accurate classifier for the integrative clusters. Using the existing classifier iC10 gave similar results. One potential reason for this was that the NanoString probe set that was selected was not well tailored to FFPE tissue. I therefore attempted to use RNA sequencing to re-design a probe set that was suitable for RNA from FFPE tissue. This approach suffered from poor data quality, which will require significant future work to resolve.

However, while the overall agreement in gene expression across tissue preparation methods and sequencing techniques was poor, it is important to remember that my ultimate aim was to find a clinically applicable method for prognostication. Rueda *et al.* (2019) have shown that integrative clusters 3, 7, 8 and 4 ER+ have a good prognosis. The classifiers I have attempted have all struggled to distinguish integrative clusters 3 and 8, but it is arguable that for the purpose of prognosis these two groups could be combined. They also identified 1, 2, 6, and 9 are late relapsers, hence these are key groups to be identified. RNA-seq data from both tissue types were able to specifically but not sensitively identify membership of iC1 and 2, but a very large number of samples were erroneously classified as belonging to iC9.

If these methods can be improved, and a novel and robust integrative cluster classifier designed for FFPE tissue, the next step will be to test it on biopsy tissue in our neoadjuvant cohort and to assess its ability to predict response to chemotherapy.

Despite all of the difficulties I have encountered, I still feel that utilising RNA from FFPE tissue is an important next step in precision medicine for breast cancer. With the clinical setting in mind, the large-scale use of RNA-seq requires a significant amount of bioinformatics input and it is time-consuming. A fast, high-throughput platform such as NanoString with a defined, validated probe set is much more feasible to provide a timely outcome prediction for patients who have just been diagnosed with breast cancer and are deciding whether to have NAT. Similarly, FFPE tissue can be transported

quickly and safely from local hospitals to a central testing centre without specialist equipment for transport.

Overall, therefore, the focus of next stage work should be on designing a new NanoString probe set for FFPE tissue, which can be validated and applied clinically. RNA-seq, despite its advantage of sequencing all genes, may not be the best way to approach this given the poor correlation in detected gene expression, even for the same tissue type.

7.3 Digital pathology

Digital pathology and its derivatives have benefited greatly from the advances in computing memory and processor power. New methods are constantly being developed, and existing methods are constantly being improved. In this thesis I have both improved and extended existing methods for cell classification, and developed completely novel algorithms for spatial analysis. I have produced a nuanced assessment of these methods, identifying their weaknesses as well as strengths, and proposing novel solutions for improvement.

With my methods, I have confirmed in a novel dataset the previous finding that median lymphocyte density is different in those with pCR compared to those without (Ali *et al.*, 2016). I have also identified spatial features that are associated with better response and found two primary determinants: the size of tumour clusters and their spatial relationship to immune cells, especially at the tumour interface zone. The primary advantage of these features over manual pathology assessment is that they can be reproduced robustly and quantified objectively. The natural next step is to apply these techniques to larger cohorts of patients to determine their associations across different tumour subtypes, for example using intrinsic subtypes, or based on hormone and *HER2* receptor status. This will not only provide further insight to the biology of the tumour and its interaction with its microenvironment, but also allow for the methods to be combined with other clinically applicable or in-use techniques, to improve the stratification of patients.

There are several next steps to this line of enquiry. The first is to improve the accuracy of cell classification. To this end, I have already trained an MPhil student how to create a larger training set for the random forest classifier, and to continue its optimisation in collaboration with Dr R Ali and Dr A Dariush.

With improving cell classification in mind, I looked at other ways to tackle especially problematic areas. In the last couple of years, some groups have extended neural networks from tile-based analysis into cell classification cell classification, with good results. It is worth considering moving onto a neural network approach eventually, if

the expansion of training set does not improve the accuracy of our machine learning methods. One project, in collaboration with Dr Mireia Crispin Ortuzar and Mira Valkonen, is designed specifically to use such neural networks to improve the identification of vesicular nuclei.

These vesicular nuclei may be particularly biologically and clinically relevant, as they can be a feature of an actively proliferating or actively synthesizing cell (Gartner and Hiatt, 2006). Given the potential difficulty of classifying these nuclei, even with neural networks, I have also helped to design and create a training set for a tile-based neural network to assess the relationship between the number of vesicular nuclei and response to neoadjuvant chemotherapy. In this approach, each biopsy core is split into small tiles, each containing a small number of cells. The overall properties of the tile are assessed semi-quantitatively according to its nuclear properties – vesicular, not vesicular, mixed, or no-tumour-cells. Once each tile has been labelled, an overall score will be calculated for each slide, and this will be correlated with the patient’s outcome from NAT. This can potentially function as an independent assessment tool for each image, without having to identify vesicular nuclei with great precision, a task that has proven to be difficult.

In this thesis I have also demonstrated the importance of spatial relationships between tumour cells and their microenvironment, specifically lymphocytes in the interface zone and infiltrating tumour clusters. The assessment of the other important components of the microenvironment, however, cannot be accurately assessed using cell segmentation methods. Two larger-scale structural components of the tumour microenvironment are vasculature and stroma. The former is a structure that cannot be identified by cell morphology alone, while the latter often shows a subjective difference in extracellular matrix structure that is not possible to assess based on nuclear morphology. I am collaborating in two projects using neural networks for these more “abstract” features.

In the first, Dr Tristan Whitmash and I are looking at blood vessels. These are rarely present in biopsy specimens, so we are using surgical samples from a number of different trials. The vasculature is vital to tumour development, and to the transport of

immune cells and drugs to the site of tumour. Based on nuclear morphology alone, our existing cell classification method cannot distinguish vessel endothelial cells, or cells in the vessel tunica media, from fibroblasts. We are therefore using CD31 IHC (a vessel marker) to highlight the vasculature within the tissue, and have trained a neural network to identify it accurately. Lymphocytes and tumour cells are also identified using nuclear morphology. Our preliminary results show that lymphocyte density is directly related to their distance from the blood vessels. We showed that there are more, but smaller vessels, closer to the tumour. This is a common subjective observation in tumour microscopy, which we can now quantify and potentially relate to outcome and prognosis.

Another key element of the microenvironment is stroma. It is difficult to assess stroma in biopsy specimens due to its paucity. The stroma in the resection specimen after NAT, however, is abundant and often shows interesting and varied features. Some appear more fibrotic, while others more myxoid, and still others have numerous macrophages. Often there is a mixture of different patterns. This has not yet been quantified and correlated with outcome measures, such as relapse and survival. In collaboration with Dr R Ali, Dr A Dariush and Dr E Provenzano, I have designed a tile-based training set for stromal characteristics. This method does not rely on nuclear morphology, which, in a post-NAT setting, is often distorted and misleading. Stromal texture in resection specimens may be especially important in those with pCR, where no tumour cell is left for assessment.

Finally, we have also started using neural networks to tackle tasks that are more “abstract” compared to nuclear segmentation. In collaboration with Dr A Dariush and Dr R Ali, I have started to assess the quality of tissue microarray core, before it is used for further analysis. This will improve the step where subjective assessment, often by non-specialists, needs to be made about whether a sample is suitable for further analysis.

It is clear from the work I have undertaken in this thesis that computational analysis of pathology images is a “gold-mine” for spatial information about the tumour and its microenvironment. It is potentially as easily applied to single patients and large datasets. It is much more clinically feasible to apply these methods to H&E and IHC

slides, than to apply existing spatial methods such as imaging mass cytometry, which offer a detailed spatial analysis of the molecular properties of cells, but require expensive and specialist equipment. With machine learning methods, and potentially in future with neural networks, I have shown that it is possible to use widely available samples to quantitatively assess spatial relationships between cells, cell groups, structures and even stromal textures.

7.4 Multi-modal assessment

In this thesis I have demonstrated two potentially complementary approaches to using clinically available material to predict the response to NAT in breast cancer. The attempt at using molecular classification on RNA extracted from FFPE tissue highlighted the potential difficulty in translating high quality bench research into clinically viable methods. I have presented potential next steps to improving this translation. The methods developed for computational assessment of histopathology slides were more successful. They demonstrated that response to NAT was associated with intrinsic tumour properties as well as its relationship with the microenvironment, especially lymphocyte density in the interface region at the edge of tumour clusters.

Moving forwards, it is vital to note that neither method should act alone in forming a predictive model. In future, a successful model is most likely to include information from multiple modalities, optimising the treatment choice for patients and improving overall outcome.

8. Summary

Neoadjuvant chemotherapy (NAT) has been shown to benefit many patients with breast cancer, and their response to the treatment is prognostic for their overall outcome. Chapter 1 introduces the approach to breast cancer treatment and the known determinants of breast cancer outcome. It also outlines a significant problem, specifically that not all patients who receive neoadjuvant chemotherapy respond to the treatment, and as a result suffer from potentially unnecessary side effects and delays to other treatment options. Being able to accurately predict an individual patient's response to treatment is crucial for optimising their care.

Recent advances have shown that gene expression can classify breast cancers into distinct integrative clusters with different outcomes and potentially different response to NAT. However, this research is based on RNA from fresh frozen tissue, which is difficult to manage in a clinical setting. Chapter 3 describes my attempt to build an accurate classifier for integrative clusters based on RNA expression, assessed from widely available FFPE tissue, using a user-friendly NanoString technique. Unfortunately, it was not possible to achieve reliable classification with this method, perhaps because the gene probe set was selected based on fresh frozen tissue. In Chapter 4 I therefore move on to compare gene expression across FFPE and fresh frozen tissue with RNA sequencing, which allows the quantification of all genes, without the need to pre-specify probes. However, there was poor agreement in measured gene expression between tissue preparations (FFPE vs Frozen), and between assessment methods (Illumina microarray vs RNA-seq), resulting in unreliable classification of tumours into integrative clusters. Overall, these findings represent a challenge to the immediate adoption of the important scientific advance of integrative clusters in real-world precision medicine approaches, so my chapters take a narrative approach to displaying a large number of quality control analyses that shed some light on where

discrepancies may have arisen. I speculate on where it might be best to focus future work to improve this approach.

In chapter 5, I turn to computational pathology. I begin by validating and improving cell classification methods using whole slide digital images from H&E diagnostic pre-treatment biopsies. In a novel dataset, using a new method, I validate the previous finding that the presence of immune infiltrate in pre-treatment biopsies is predictive of NAT response. In chapter 6, I use this cell classification to demonstrate that spatial profiles of tumour clusters and their relationship to immune cells are associated with treatment outcome. I identify that larger tumour clusters, more heterogeneous tumour clusters, and more lymphocytes in the region immediately bordering tumour clusters are all correlated with pathological complete response to NAT. This has not been previously shown in the context of breast cancer and response to NAT, and I plan to submit this part of work to a peer-reviewed journal for publication.

Both gene expression profiling and computational pathology have the potential for translation into adjuncts to existing stratification methods, offering patients better care. In this thesis I have explored the real-world applicability of both methods using widely available clinical samples, namely formalin fixed paraffin embedded (FFPE) tissue and H&E and IHC diagnostic slides. In chapter 7, I review the large number of further studies spawned by the work I present in this thesis, and explain how they might improve our ability to understand tumour biology, and translate this understanding to the clinical setting.

Overall, this thesis has provided a step towards the translation of the molecular classification of breast cancer, and computational methods for pathology image analysis, into real-world precision medicine, predicting response to neoadjuvant chemotherapy.

9. References

Abraham J, Vallier A-L, Qian W, Machin A, Grybowicz L, Thomas S, *et al.* PARTNER: Randomised, phase II/III trial to evaluate the safety and efficacy of the addition of olaparib to platinum-based neoadjuvant chemotherapy in triple negative and/or germline BRCA mutated breast cancer patients. American Society of Clinical Oncology; 2018.

Adams S, Gray RJ, Demaria S, Goldstein L, Perez EA, Shulman LN, *et al.* Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. *Journal of clinical oncology* 2014; 32(27): 2959.

Al-Saleh K, El-Aziz A, Ali A, Abozeed W, El-Warith A, Ibraheem A, *et al.* Predictive and prognostic significance of CD8+ tumor-infiltrating lymphocytes in patients with luminal B/HER 2 negative breast cancer treated with neoadjuvant chemotherapy. *Oncology letters* 2017; 14(1): 337-44.

Ali HR, Dariush A, Provenzano E, Bardwell H, Abraham JE, Iddawela M, *et al.* Computational pathology of pre-treatment biopsies identifies lymphocyte density as a predictor of response to neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res* 2016; 18(1): 21.

Ali HR, Dariush A, Thomas J, Provenzano E, Dunn J, Hiller L, *et al.* Lymphocyte density determined by computational pathology validated as a predictor of response to neoadjuvant chemotherapy in breast cancer: secondary analysis of the ARTemis trial. *Annals of Oncology* 2017; 28(8): 1832-5.

Ali HR, Irwin M, Morris L, Dawson S, Blows F, Provenzano E, *et al.* Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. *British journal of cancer* 2013; 108(3): 602-12.

Ali HR, Rueda OM, Chin S-F, Curtis C, Dunning MJ, Aparicio SA, *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology* 2014; 15(8): 431.

Alom MZ, Aspiras T, Taha TM, Asari VK, Bowen T, Billiter D, *et al.* Advanced deep convolutional neural network approaches for digital pathology image analysis: A comprehensive evaluation with different use cases. *arXiv preprint arXiv:190409075* 2019.

Anderson NM, Simon MC. The tumor microenvironment. *Current Biology* 2020; 30(16): R921-R5.

Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

Ascierto ML, Kmiecik M, Idowu MO, Manjili R, Zhao Y, Grimes M, *et al.* A signature of immune function genes associated with recurrence-free survival in breast cancer patients. *Breast Cancer Research and Treatment* 2012; 131(3): 871-80.

- Asselain B, Barlow W, Bartlett J, Bergh J, Bergsten-Nordström E, Bliss J, *et al.* Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. *The Lancet Oncology* 2018; 19(1): 27-39.
- Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Toulouse, France: Association for Computational Linguistics; 2001. p. 26-33.
- Barker J, Hoogi A, Depeursinge A, Rubin DL. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical image analysis* 2016; 30: 60-71.
- Barrdahl M, Rudolph A, Hopper JL, Southey MC, Broeks A, Fasching PA, *et al.* Gene-environment interactions involving functional variants: results from the Breast Cancer Association Consortium. *International journal of cancer* 2017; 141(9): 1830-40.
- Bear HD, Anderson S, Smith RE, Geyer CE, Jr., Mamounas EP, Fisher B, *et al.* Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer: National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2006; 24(13): 2019-27.
- Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, *et al.* Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine* 2011; 3(108): 108ra13-ra13.
- Bines J, Earl H, Buzaid AC, Saad ED. Anthracyclines and taxanes in the neo/adjuvant treatment of breast cancer: does the sequence matter? *Ann Oncol* 2014; 25(6): 1079-85.
- Bivin WW, Yergiyev O, Bunker ML, Silverman JF, Krishnamurti U. GRB7 expression and correlation with HER2 amplification in invasive breast carcinoma. *Applied Immunohistochemistry & Molecular Morphology* 2017; 25(8): 553-8.
- Bollobás B. *Modern graph theory*: Springer Science & Business Media; 2013.
- Bonadonna G, Brusamolino E, Valagussa P, Rossi A, Brugnatelli L, Brambilla C, *et al.* Combination Chemotherapy as an Adjuvant Treatment in Operable Breast Cancer. *New England Journal of Medicine* 1976; 294(8): 405-10.
- Burstein HJ. The distinctive nature of HER2-positive breast cancers. *New England Journal of Medicine* 2005; 353(16): 1652.
- Calabrò A, Beissbarth T, Kuner R, Stojanov M, Benner A, Asslaber M, *et al.* Effects of infiltrating lymphocytes and estrogen receptor on gene expression and prognosis in breast cancer. *Breast Cancer Research and Treatment* 2009; 116(1): 69-77.
- Callagy GM, Webber MJ, Pharoah PD, Caldas C. Meta-analysis confirms BCL2 is an independent prognostic marker in breast cancer. *BMC cancer* 2008; 8(1): 1-10.
- Campbell EJ, Tesson M, Doogan F, Mohammed ZM, Mallon E, Edwards J. The combined endocrine receptor in breast cancer, a novel approach to traditional hormone receptor interpretation and a better discriminator of outcome than ER and PR alone. *British journal of cancer* 2016; 115(8): 967-73.

- CancerResearchUK. Breast Cancer Statistics. [cited 2017 22 Feb]; Available from: <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive>
- Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* 2006; 7(10): R100.
- Cesano A. nCounter® PanCancer immune profiling panel (NanoString technologies, Inc., Seattle, WA). *Journal for immunotherapy of cancer* 2015; 3(1): 1-3.
- Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *JNCI: Journal of the National Cancer Institute* 2009; 101(10): 736-50.
- Chen X, Deane NG, Lewis KB, Li J, Zhu J, Washington MK, *et al.* Comparison of nanostring nCounter® data on FFPE colon cancer samples and affymetrix microarray data on matched frozen tissues. *PloS one* 2016; 11(5): e0153784.
- Chowdhury N, Pai MR, Lobo FD, Kini H, Varghese R. Interobserver variation in breast cancer grading: a statistical modeling approach. *Analytical and quantitative cytology and histology* 2006; 28(4): 213-8.
- Clark Jr WH, Elder DE, Guerry IV D, Braitman LE, Trock BJ, Schultz D, *et al.* Model predicting survival in stage I melanoma based on tumor progression. *JNCI: Journal of the National Cancer Institute* 1989; 81(24): 1893-904.
- Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* 2010; 38(6): 1767-71.
- Colleoni M, Bagnardi V, Rotmensz N, Gelber RD, Viale G, Pruneri G, *et al.* Increasing steroid hormone receptors expression defines breast cancer subtypes non responsive to preoperative chemotherapy. *Breast Cancer Res Treat* 2009; 116(2): 359-69.
- Colleoni M, Viale G, Zahrieh D, Pruneri G, Gentilini O, Veronesi P, *et al.* Chemotherapy Is More Effective in Patients with Breast Cancer Not Expressing Steroid Hormone Receptors. *Clinical Cancer Research* 2004; 10(19): 6622.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, *et al.* A survey of best practices for RNA-seq data analysis. *Genome biology* 2016; 17(1): 1-19.
- Cope TE, Rittman T, Borchert RJ, Jones PS, Vatansever D, Allinson K, *et al.* Tau burden and the functional connectome in Alzheimer's disease and progressive supranuclear palsy. *Brain* 2018; 141(2): 550-67.
- Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, *et al.* Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet* 2014; 384(9938): 164-72.
- Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one* 2017; 12(12): e0190152.
- Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nature methods* 2008; 5(10): 887-93.

- Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, complex systems* 2006; 1695(5): 1-9.
- Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; 486(7403): 346-52.
- Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research* 2015; 5(10): 2929-43.
- Daniel AR, Hagan CR, Lange CA. Progesterone receptor action: defining a role in breast cancer. *Expert review of endocrinology & metabolism* 2011; 6(3): 359-69.
- de Leeuw WJ, Berx G, Vos CB, Peterse JL, Van de Vijver MJ, Litvinov S, *et al.* Simultaneous loss of E-cadherin and catenins in invasive lobular breast cancer and lobular carcinoma in situ. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 1997; 183(4): 404-11.
- De Meo P, Ferrara E, Fiumara G, Provetti A. Generalized louvain method for community detection in large networks. 2011 11th international conference on intelligent systems design and applications; 2011: IEEE; 2011. p. 88-93.
- Dekker TJA, Charehbili A, Smit VTHBM, ten Dijke P, Meershoek-Klein Kranenburg E, van de Velde CJH, *et al.* Disorganised stroma determined on pre-treatment breast cancer biopsies is associated with poor response to neoadjuvant chemotherapy: Results from the NEOZOTAC trial. *Molecular Oncology* 2015; 9(6): 1120-8.
- Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, *et al.* Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *The lancet oncology* 2018; 19(1): 40-50.
- Desjardins P, Conklin D. NanoDrop microvolume quantitation of nucleic acids. *JoVE (Journal of Visualized Experiments)* 2010(45): e2565.
- Dieci M, Conte P, Bisagni G, Brandes A, Frassoldati A, Cavanna L, *et al.* Association of tumor-infiltrating lymphocytes with distant disease-free survival in the ShortHER randomized adjuvant trial for patients with early HER2+ breast cancer. *Annals of Oncology* 2019; 30(3): 418-23.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29(1): 15-21.
- Dong F, Irshad H, Oh EY, Lerwill MF, Brachtel EF, Jones NC, *et al.* Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. *PloS one* 2014; 9(12): e114885.
- Donnem T, Kilvaer T, Andersen S, Richardsen E, Paulsen E, Hald S, *et al.* Strategies for clinical implementation of TNM-Immunoscore in resected nonsmall-cell lung cancer. *Annals of Oncology* 2016; 27(2): 225-32.
- Dowsett M, Bartlett J, Ellis I, Salter J, Hills M, Mallon E, *et al.* Correlation between immunohistochemistry (HercepTest) and fluorescence in situ hybridization (FISH) for

- HER-2 in 426 breast carcinomas from 37 centres. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 2003; 199(4): 418-23.
- Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, *et al.* Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *Journal of the National Cancer Institute* 2011; 103(22): 1656-64.
- Dundar MM, Badve S, Bilgin G, Raykar V, Jain R, Sertel O, *et al.* Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering* 2011; 58(7): 1977-84.
- Dupree EJ, Jayathirtha M, Yorkey H, Mihasan M, Petre BA, Darie CC. A critical review of bottom-up proteomics: The good, the bad, and the future of this field. *Proteomes* 2020; 8(3): 14.
- Earl HM, Hiller L, Vallier A-L, Loi S, Howe D, Higgins HB, *et al.* PERSEPHONE: 6 versus 12 months (m) of adjuvant trastuzumab in patients (pts) with HER2 positive (+) early breast cancer (EBC): Randomised phase 3 non-inferiority trial with definitive 4-year (yr) disease-free survival (DFS) results. *American Society of Clinical Oncology*; 2018.
- Earl HM, Hiller L, Vallier A-L, Loi S, McAdam K, Hughes-Davies L, *et al.* 6 versus 12 months of adjuvant trastuzumab for HER2-positive early breast cancer (PERSEPHONE): 4-year disease-free survival results of a randomised phase 3 non-inferiority trial. *The Lancet* 2019; 393(10191): 2599-612.
- Earl HM, Vallier A-L, Qian W, Grybowicz L, Thomas S, Mahmud S, *et al.* PARTNER: Randomised, phase II/III trial to evaluate the safety and efficacy of the addition of olaparib to platinum-based neoadjuvant chemotherapy in triple negative and/or germline BRCA mutated breast cancer patients. *American Society of Clinical Oncology*; 2017.
- Earl HM, Vallier AL, Hiller L, Fenwick N, Young J, Iddawela M, *et al.* Effects of the addition of gemcitabine, and paclitaxel-first sequencing, in neoadjuvant sequential epirubicin, cyclophosphamide, and paclitaxel for women with high-risk early breast cancer (Neo-tAnGo): an open-label, 2x2 factorial randomised phase 3 trial. *Lancet Oncol* 2014; 15(2): 201-12.
- Early Breast Cancer Trialists' Collaborative G, Darby S, McGale P, Correa C, Taylor C, Arriagada R, *et al.* Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. *Lancet (London, England)* 2011; 378(9804): 1707-16.
- Eastel JM, Lam KW, Lee NL, Lok WY, Tsang AHF, Pei XM, *et al.* Application of NanoString technologies in companion diagnostic development. *Expert review of molecular diagnostics* 2019; 19(7): 591-8.
- Eisinger F, Jacquemier J, Charpin C, Stoppa-Lyonnet D, Bressac-de Paillerets B, Peyrat J-P, *et al.* Mutations at BRCA1: the medullary breast carcinoma revisited. *Cancer research* 1998; 58(8): 1588-92.
- Ellis P. *WHO Classification of Tumours. Pathology and Genetics of Tumours of the Breast and Female Genital Organs.*: Lyon Press, Lyon; 2003.

- Ellsworth RE, Blackburn HL, Shriver CD, Soon-Shiong P, Ellsworth DL. Molecular heterogeneity in breast cancer: State of the science and implications for patient care. *Semin Cell Dev Biol* 2017; 64: 65-72.
- Elston CW, Ellis IO. pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991; 19(5): 403-10.
- Enfield KS, Martin SD, Marshall EA, Kung SH, Gallagher P, Milne K, *et al.* Hyperspectral cell sociology reveals spatial tumor-immune cell interactions associated with lung cancer recurrence. *Journal for immunotherapy of cancer* 2019; 7(1): 1-13.
- Esserman LJ, Berry DA, DeMichele A, Carey L, Davis SE, Buxton M, *et al.* Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL--CALGB 150007/150012, ACRIN 6657. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2012; 30(26): 3242-9.
- Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Simoudis E, Han J, Fayyad UM, editors.: AAAI Press; 1996. p. 226-31.
- Evers DL, Fowler CB, Cunningham BR, Mason JT, O'Leary TJ. The effect of formaldehyde fixation on RNA: optimization of formaldehyde adduct removal. *The Journal of Molecular Diagnostics* 2011; 13(3): 282-8.
- Failmezger H, Muralidhar S, Rullan A, de Andrea CE, Sahai E, Yuan Y. Topological Tumor Graphs: a graph-based spatial model to infer stromal recruitment for immunosuppression in melanoma histology. *Cancer research* 2020; 80(5): 1199-209.
- Fisher B, Bryant J, Wolmark N, Mamounas E, Brown A, Fisher ER, *et al.* Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *Journal of Clinical Oncology* 1998; 16(8): 2672-85.
- Fisher ER, Wang J, Bryant J, Fisher B, Mamounas E, Wolmark N. Pathobiology of preoperative chemotherapy. *Cancer* 2002; 95(4): 681-95.
- Fleige S, Pfaffl MW. RNA integrity and the effect on the real-time qRT-PCR performance. *Molecular aspects of medicine* 2006; 27(2-3): 126-39.
- Fontanella C, Lederer B, Gade S, Vanoppen M, Blohmer JU, Costa SD, *et al.* Impact of body mass index on neoadjuvant treatment outcome: a pooled analysis of eight prospective neoadjuvant breast cancer trials. *Breast Cancer Res Treat* 2015; 150(1): 127-39.
- Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer* 2012; 12(4): 298-306.
- Gaffney EF, Riegman PH, Grizzle WE, Watson PH. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotechnic & Histochemistry* 2018; 93(5): 373-86.
- Gajria D, Chandarlapaty S. HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. *Expert review of anticancer therapy* 2011; 11(2): 263-75.

- Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, *et al.* Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science* 2006; 313(5795): 1960.
- Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C, *et al.* Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *The Journal of pathology* 2014; 232(2): 199-209.
- Gao G, Wang Z, Qu X, Zhang Z. Prognostic value of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer: a systematic review and meta-analysis. *BMC cancer* 2020a; 20(1): 1-15.
- Gao Z-h, Li C-x, Liu M, Jiang J-y. Predictive and prognostic role of tumour-infiltrating lymphocytes in breast cancer patients with different molecular subtypes: a meta-analysis. *BMC cancer* 2020b; 20(1): 1-14.
- Gartner LP, Hiatt JL. *Color textbook of histology e-book*: Elsevier Health Sciences; 2006.
- Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease* 2013; 8: 331-59.
- Gong C, Anders RA, Zhu Q, Taube JM, Green B, Cheng W, *et al.* Quantitative Characterization of CD8+ T Cell Clustering and Spatial Heterogeneity in Solid Tumors. *Front Oncol* 2019; 8(649).
- Goto W, Kashiwagi S, Takada K, Asano Y, Takahashi K, Fujita H, *et al.* Significance of intrinsic breast cancer subtypes on the long-term prognosis after neoadjuvant chemotherapy. *Journal of Translational Medicine* 2018; 16(1): 307.
- Greytak SR, Engel KB, Bass BP, Moore HM. Accuracy of Molecular Data Generated with FFPE Biospecimens: Lessons from the Literature. *Cancer research* 2015; 75(8): 1541-7.
- Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, de Wind A, *et al.* CD4(+) follicular helper T cell infiltration predicts breast cancer survival. *J Clin Invest* 2013; 123(7): 2873-92.
- Hagemeister Jr FB, Buzdar AU, Luna MA, Blumenschein GR. Causes of death in breast cancer a clinicopathologic study. *Cancer* 1980; 46(1): 162-7.
- Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 2009; 24(2): 8-12.
- Heindl A, Khan AM, Rodrigues DN, Eason K, Sadanandam A, Orbegoso C, *et al.* Microenvironmental niche divergence shapes BRCA1-dysregulated ovarian cancer morphological plasticity. *Nature Communications* 2018a; 9(1): 3917.
- Heindl A, Nawaz S, Yuan Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Lab Invest* 2015; 95(4): 377-84.
- Heindl A, Sestak I, Naidoo K, Cuzick J, Dowsett M, Yuan Y. Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. *JNCI: Journal of the National Cancer Institute* 2018b; 110(2): 166-75.
- Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, *et al.* Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal

- for a standardized method from the International Immuno-Oncology Biomarkers Working Group: part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Advances in anatomic pathology* 2017; 24(6): 311.
- Herrero-Vicent C, Guerrero A, Gavilá J, Gozalbo F, Hernández A, Sandiego S, *et al.* Predictive and prognostic impact of tumour-infiltrating lymphocytes in triple-negative breast cancer treated with neoadjuvant chemotherapy. *ecancermedicalscience* 2017; 11.
- Howse J. *OpenCV computer vision with python*: Packt Publishing Ltd; 2013.
- Hudis CA. Trastuzumab—mechanism of action and use in clinical practice. *New England journal of medicine* 2007; 357(1): 39-51.
- Huober J, von Minckwitz G, Denkert C, Tesch H, Weiss E, Zahm DM, *et al.* Effect of neoadjuvant anthracycline-taxane-based chemotherapy in different biological breast cancer phenotypes: overall results from the GeparTrio study. *Breast Cancer Res Treat* 2010; 124(1): 133-40.
- Idos GE, Kwok J, Bonthala N, Kysh L, Gruber SB, Qu C. The prognostic implications of tumor infiltrating lymphocytes in colorectal cancer: a systematic review and meta-analysis. *Scientific reports* 2020; 10(1): 1-14.
- Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE reviews in biomedical engineering* 2013; 7: 97-114.
- Jones W, Greytak S, Odeh H, Guan P, Powers J, Bavarva J, *et al.* Deleterious effects of formalin-fixation and delays to fixation on RNA and miRNA-Seq profiles. *Scientific reports* 2019; 9(1): 1-10.
- Jushi DX. Wei Ji Bao Shu.
- Karagiannis GS, Pastoriza JM, Wang Y, Harney AS, Entenberg D, Pignatelli J, *et al.* Neoadjuvant chemotherapy induces breast cancer metastasis through a TMEM-mediated mechanism. *Science Translational Medicine* 2017; 9(397).
- Katz SJ, Jagsi R, Morrow M. Reducing overtreatment of cancer with precision medicine: just what the doctor ordered. *Jama* 2018; 319(11): 1091-2.
- Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine* 2018; 27: 317-28.
- Kim KI, Lee KH, Kim TR, Chun YS, Lee TH, Park HK. Ki-67 as a predictor of response to neoadjuvant chemotherapy in breast cancer patients. *Journal of breast cancer* 2014; 17(1): 40.
- Kise K, Sato A, Iwata M. Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding* 1998; 70(3): 370-82.
- Kokkat TJ, Patel MS, McGarvey D, LiVolsi VA, Baloch ZW. Archived formalin-fixed paraffin-embedded (FFPE) blocks: A valuable underexploited resource for extraction of DNA, RNA, and protein. *Biopreserv Biobank* 2013; 11(2): 101-6.

- Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal* 2018; 16: 34-42.
- Krag DN, Anderson SJ, Julian TB, Brown AM, Harlow SP, Costantino JP, *et al.* Sentinel-lymph-node resection compared with conventional axillary-lymph-node dissection in clinically node-negative patients with breast cancer: overall survival findings from the NSABP B-32 randomised phase 3 trial. *Lancet Oncol* 2010; 11(10): 927-33.
- Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* 2008; 2: 4.
- Ladoire S, Arnould L, Mignot G, Coudert B, Rébé C, Chalmin F, *et al.* Presence of Foxp3 expression in tumor cells predicts better survival in HER2-overexpressing breast cancer patients treated with neoadjuvant chemotherapy. *Breast cancer research and treatment* 2011; 125(1): 65-72.
- Lakhtakia R. A Brief History of Breast Cancer: Part I: Surgical domination reinvented. *Sultan Qaboos Univ Med J* 2014; 14(2): e166-e9.
- Lan C, Heindl A, Huang X, Xi S, Banerjee S, Liu J, *et al.* Quantitative histology analysis of the ovarian tumour microenvironment. *Scientific reports* 2015; 5(1): 1-12.
- Lee S, Cho EY, Park YH, Ahn JS, Im Y-H. Prognostic impact of FOXP3 expression in triple-negative breast cancer. *Acta Oncologica* 2013; 52(1): 73-81.
- Li XB, Krishnamurti U, Bhattarai S, Klimov S, Reid MD, O'Regan R, *et al.* Biomarkers Predicting Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer. *Am J Clin Pathol* 2016; 145(6): 871-8.
- Liao H-Y, Zhang W-W, Sun J-Y, Li F-Y, He Z-Y, Wu S-G. The clinicopathological features and survival outcomes of different histological subtypes in triple-negative breast cancer. *Journal of Cancer* 2018; 9(2): 296.
- Liedtke C, Hatzis C, Symmans WF, Desmedt C, Haibe-Kains B, Valero V, *et al.* Genomic grade index is associated with response to chemotherapy in patients with breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2009; 27(19): 3185-91.
- Liu S, Duan X, Xu L, Xin L, Cheng Y, Liu Q, *et al.* Optimal threshold for stromal tumor-infiltrating lymphocytes: its predictive and prognostic value in HER2-positive breast cancer treated with trastuzumab-based neoadjuvant chemotherapy. *Breast cancer research and treatment* 2015; 154(2): 239-49.
- Liu S, Foulkes WD, Leung S, Gao D, Lau S, Kos Z, *et al.* Prognostic significance of FOXP3+ tumor-infiltrating lymphocytes in breast cancer depends on estrogen receptor and human epidermal growth factor receptor-2 expression status and concurrent cytotoxic T-cell infiltration. *Breast Cancer Research* 2014; 16(5): 1-12.
- Liu S, Lachapelle J, Leung S, Gao D, Foulkes WD, Nielsen TO. CD8+ lymphocyte infiltration is an independent favorable prognostic indicator in basal-like breast cancer. *Breast Cancer Res* 2012; 14(2): R48.
- Loibl S, Volz C, Mau C, Blohmer JU, Costa SD, Eidtmann H, *et al.* Response and prognosis after neoadjuvant chemotherapy in 1,051 patients with infiltrating lobular breast carcinoma. *Breast Cancer Res Treat* 2014; 144(1): 153-62.

- Longacre TA, Ennis M, Quenneville LA, Bane AL, Bleiweiss IJ, Carter BA, *et al.* Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. *Modern pathology* 2006; 19(2): 195-207.
- Lu J, Steeg PS, Price JE, Krishnamurthy S, Mani SA, Reuben J, *et al.* Breast cancer metastasis: challenges and opportunities. *AACR*; 2009.
- Lyons JA, Myles J, Pohlman B, Macklis RM, Crowe J, Crownover RL. Treatment of prognosis of primary breast lymphoma: a review of 13 cases. *Am J Clin Oncol* 2000; 23(4): 334-6.
- Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; 458(7234): 97-101.
- Manley LJ, Ma D, Levine SS. Monitoring error rates in Illumina sequencing. *Journal of biomolecular techniques: JBT* 2016; 27(4): 125.
- Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, *et al.* Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical science monitor basic research* 2014; 20: 138.
- Marczyk M, Fu C, Lau R, Du L, Trevarton AJ, Sinn BV, *et al.* The impact of RNA extraction method on accurate RNA sequencing from formalin-fixed paraffin-embedded tissues. *BMC Cancer* 2019; 19(1): 1189.
- Marinovich M, Sardanelli F, Ciatto S, Mamounas E, Brennan M, Macaskill P, *et al.* Early prediction of pathologic response to neoadjuvant therapy in breast cancer: systematic review of the accuracy of MRI. *The Breast* 2012; 21(5): 669-77.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 2008; 18(9): 1509-17.
- Masuda N, Lee S-J, Ohtani S, Im Y-H, Lee E-S, Yokota I, *et al.* Adjuvant capecitabine for breast cancer after preoperative chemotherapy. *New England Journal of Medicine* 2017; 376(22): 2147-59.
- Merlo A, Casalini P, Carcangiu ML, Malventano C, Triulzi T, Mènard S, *et al.* FOXP3 expression and overall survival in breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2009; 27(11): 1746-52.
- Mi H, Gong C, Sulam J, Fertig EJ, Szalay AS, Jaffee EM, *et al.* Digital Pathology Analysis Quantifies Spatial Heterogeneity of CD3, CD4, CD8, CD20, and FoxP3 Immune Markers in Triple-Negative Breast Cancer. *Frontiers in Physiology* 2020; 11(1325).
- Mlecnik B, Bindea G, Kirilovsky A, Angell HK, Obenauf AC, Tosolini M, *et al.* The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. *Science translational medicine* 2016; 8(327): 327ra26-ra26.
- Montagna E, Bagnardi V, Viale G, Rotmensz N, Sporchia A, Cancellato G, *et al.* Changes in PgR and Ki-67 in residual tumour and outcome of breast cancer patients treated with neoadjuvant chemotherapy. *Ann Oncol* 2015; 26(2): 307-13.

- Moo T-A, Sanford R, Dang C, Morrow M. Overview of Breast Cancer Therapy. *PET clinics* 2018; 13(3): 339-54.
- Mormont R, Geurts P, Marée R. Comparison of deep transfer learning strategies for digital pathology. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*; 2018; 2018. p. 2262-71.
- Moss RW. Galen on Cancer. *Cancer Decisions* 2004.
- Mueller O, Lightfoot S, Schroeder A. RNA integrity number (RIN)–standardization of RNA quality control. *Agilent application note, publication* 2004; 1: 1-8.
- Natrajan R, Sailem H, Mardakheh FK, Arias Garcia M, Tape CJ, Dowsett M, *et al.* Microenvironmental Heterogeneity Parallels Breast Cancer Progression: A Histology–Genomic Integration Analysis. *PLOS Medicine* 2016; 13(2): e1001961.
- Ng CK, Pemberton HN, Reis-Filho JS. Breast cancer intratumor genetic heterogeneity: causes and implications. *Expert Rev Anticancer Ther* 2012; 12(8): 1021-32.
- Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *The lancet oncology* 2019; 20(5): e253-e61.
- NICE. Early and locally advanced breast cancer: diagnosis and management. *NICE guidelines* 2018; NG101.
- Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, *et al.* Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC cancer* 2014; 14(1): 1-14.
- Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, *et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. *Clinical cancer research* 2010; 16(21): 5222-32.
- Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, *et al.* Impact of RNA degradation on gene expression profiling. *BMC medical genomics* 2010; 3(1): 1-14.
- Osako T, Nishimura R, Okumura Y, Toyozumi Y, Arima N. Predictive significance of the proportion of ER-positive or PgR-positive tumor cells in response to neoadjuvant chemotherapy for operable HER2-negative breast cancer. *Exp Ther Med* 2012; 3(1): 66-71.
- Parker JD, Yap SQ, Starks E, Slind J, Swanson L, Docking TR, *et al.* Fixation Effects on Variant Calling in a Clinical Resequencing Panel. *The Journal of Molecular Diagnostics* 2019; 21(4): 705-17.
- Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology* 2009; 27(8): 1160.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* 2017; 14(4): 417-9.
- Patterson EK, Dackerman ME. Nucleic acid content in relation to cell size in the mature larval salivary gland of *Drosophila melanogaster*. *Archives of biochemistry and biophysics* 1952; 36(1): 97-113.

- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011; 12: 2825-30.
- Phillips T, Murray G, Wakamiya K, Askaa J, Huang D, Welcher R, *et al.* Development of standard estrogen and progesterone receptor immunohistochemical assays for selection of patients for antihormonal therapy. *Applied Immunohistochemistry & Molecular Morphology* 2007; 15(3): 325-31.
- Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* 2017: 201178.
- Provenzano E, Bossuyt V, Viale G, Cameron D, Badve S, Denkert C, *et al.* Standardization of pathologic evaluation and reporting of postneoadjuvant specimens in clinical trials of breast cancer: recommendations from an international working group. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2015; 28(9): 1185-201.
- Rakha EA, Patel A, Powe DG, Benhasouna A, Green AR, Lambros MB, *et al.* Clinical and biological significance of E-cadherin protein expression in invasive lobular carcinoma of the breast. *The American journal of surgical pathology* 2010a; 34(10): 1472-9.
- Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, *et al.* Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research* 2010b; 12(4): 207.
- Ramsey B, Bai T, Newell AH, Troxell M, Park B, Olson S, *et al.* GRB7 protein over-expression and clinical outcome in breast cancer. *Breast cancer research and treatment* 2011; 127(3): 659-69.
- Rathore AS, Goel MM, Makker A, Kumar S, Srivastava AN. Is the tumor infiltrating natural killer cell (NK-TILs) count in infiltrating ductal carcinoma of breast prognostically significant? *Asian Pac J Cancer Prev* 2014; 15(8): 3757-61.
- Robbins P, Pinder S, De Klerk N, Dawkins H, Harvey J, Sterrett G, *et al.* Histological grading of breast carcinomas: a study of interobserver agreement. *Human pathology* 1995; 26(8): 873-9.
- Roberts L, Bowers J, Sensinger K, Lisowski A, Getts R, Anderson MG. Identification of methods for use of formalin-fixed, paraffin-embedded tissue samples in RNA expression profiling. *Genomics* 2009; 94(5): 341-8.
- Robson M, Im S-A, Senkus E, Xu B, Domchek SM, Masuda N, *et al.* Olaparib for metastatic breast cancer in patients with a germline BRCA mutation. *New England Journal of Medicine* 2017; 377(6): 523-33.
- Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res* 2005; 11(16): 5678-85.
- Rozek LS, Schmit SL, Greenson JK, Tomsho LP, Rennert HS, Rennert G, *et al.* Tumor-infiltrating lymphocytes, Crohn's-like lymphoid reaction, and survival from colorectal cancer. *JNCI: Journal of the National Cancer Institute* 2016; 108(8).

- Rubens RD, Sexton S, Tong D, Winter PJ, Knight RK, Hayward JL. Combined chemotherapy and radiotherapy for locally advanced breast cancer. *Eur J Cancer* 1980; 16(3): 351-6.
- Rueda OM. iC10: A Copy Number and Expression-Based Classifier for Breast Tumours. 2015.
- Rueda OM, Sammut S-J, Seoane JA, Chin S-F, Caswell-Jin JL, Callari M, *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* 2019; 567(7748): 399-404.
- Russnes HG, Lingjaerde OC, Borresen-Dale AL, Caldas C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am J Pathol* 2017; 187(10): 2152-62.
- Salgado R, Denkert C, Campbell C, Savas P, Nuciforo P, Aura C, *et al.* Tumor-infiltrating lymphocytes and associations with pathological complete response and event-free survival in HER2-positive early-stage breast cancer treated with lapatinib and trastuzumab: a secondary analysis of the NeoALTTO trial. *JAMA oncology* 2015a; 1(4): 448-55.
- Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, *et al.* The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Annals of oncology* 2015b; 26(2): 259-71.
- Sammut SJ. The clonal architecture and tumour microenvironment of breast cancers are shaped by neoadjuvant chemotherapy: University of Cambridge; 2019.
- Schank T, Wagner D. Approximating clustering coefficient and transitivity. *Journal of Graph Algorithms and Applications* 2005; 9(2): 265-75.
- Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology* 2006; 7(1): 1-14.
- Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna* 2016; 22(6): 839-51.
- Sirinukunwattana K, Raza SEA, Tsang Y-W, Snead DR, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging* 2016; 35(5): 1196-206.
- Sirinukunwattana K, Snead D, Epstein D, Aftab Z, Mujeeb I, Tsang YW, *et al.* Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Scientific reports* 2018; 8(1): 1-13.
- Solorzano L, Wik L, Bontell TO, Wang Y, Klemm AH, Öfverstedt J, *et al.* Machine learning for cell classification and neighborhood analysis in glioma tissue. *bioRxiv* 2021.
- Sopik V, Sun P, Narod SA. The prognostic effect of estrogen receptor status differs for younger versus older breast cancer patients. *Breast Cancer Res Treat* 2017; 165(2): 391-402.

- Sørli T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 2001; 98(19): 10869-74.
- Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *The American journal of pathology* 2002; 161(6): 1961-71.
- Stam CJ, Jones B, Nolte G, Breakspear M, Scheltens P. Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral cortex* 2007; 17(1): 92-9.
- Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nature Reviews Genetics* 2019; 20(11): 631-56.
- Stingl J, Caldas C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat Rev Cancer* 2007; 7(10): 791-9.
- Symmans WF, Peintinger F, Hatzis C, Rajan R, Kuerer H, Valero V, *et al.* Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007; 25(28): 4414-22.
- Symmans WF, Wei C, Gould R, Yu X, Zhang Y, Liu M, *et al.* Long-Term Prognostic Risk After Neoadjuvant Chemotherapy Associated With Residual Cancer Burden and Breast Cancer Subtype. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2017; 35(10): 1049-60.
- Tedds JA, Winstanley N, Lawrence A, Walton N, Auden E, Dalla S. VOExplorer: Visualising data discovery in the virtual observatory. *Astronomical Data Analysis Software and Systems XVII*; 2008; 2008. p. 159.
- Thomas E, Berner G. Prognostic and predictive implications of HER2 status for breast cancer patients. *Eur J Oncol Nurs* 2000; 4(Sa): 10-7.
- Thomas RM, John J. A review on cell detection and segmentation in microscopic images. 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT); 2017: IEEE; 2017. p. 1-5.
- Thompson AM, Moulder-Thompson SL. Neoadjuvant treatment of breast cancer. *Annals of Oncology* 2012; 23(suppl_10): x231-x6.
- Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics* 2018; 9.
- Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making* 2019; 19(1): 1-16.
- Urruticoechea A, Smith IE, Dowsett M. Proliferation marker Ki-67 in early breast cancer. *Journal of clinical oncology* 2005; 23(28): 7212-20.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics* 2013; 43(1): 11.0. 1-.0. 33.
- Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, *et al.* scikit-image: image processing in Python. *PeerJ* 2014; 2: e453.

- Von Minckwitz G, Huang C-S, Mano MS, Loibl S, Mamounas EP, Untch M, *et al.* Trastuzumab emtansine for residual invasive HER2-positive breast cancer. *New England Journal of Medicine* 2019; 380(7): 617-28.
- Vos C, Cleton-Jansen A, Berx G, De Leeuw W, Ter Haar N, Van Roy F, *et al.* E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis. *British journal of cancer* 1997; 76(9): 1131-3.
- Vu QD, Graham S, Kurc T, To MNN, Shaban M, Qaiser T, *et al.* Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology* 2019; 7: 53.
- Waggott D, Chu K, Yin S, Wouters BG, Liu F-F, Boutros PC. NanoStringNorm: an extensible R package for the pre-processing of NanoString mRNA and miRNA data. *Bioinformatics* 2012; 28(11): 1546-8.
- Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences* 2012; 131(4): 281-5.
- Walch A, Specht K, Braselmann H, Stein H, Siewert JR, Hopt U, *et al.* Coamplification and coexpression of GRB7 and ERBB2 is found in high grade intraepithelial neoplasia and in invasive Barrett's carcinoma. *International journal of cancer* 2004; 112(5): 747-53.
- Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics* 2015; 8: 54.
- Walton N, Brenton J, Caldas C, Irwin M, Akram A, Gonzalez-Solares E, *et al.* PathGrid: a service-orientated architecture for microscopy image analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2010; 368(1925): 3937-52.
- Wang S, Wang T, Yang L, Yang DM, Fujimoto J, Yi F, *et al.* ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by a convolutional neural network. *EBioMedicine* 2019; 50: 103-10.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* 2009; 10(1): 57-63.
- Weigelt B, Reis-Filho JS. Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nature reviews Clinical oncology* 2009; 6(12): 718.
- West DB. *Introduction to graph theory*: Prentice hall Upper Saddle River; 2001.
- Wetzel AW, Crowley R, Kim S, Dawson R, Zheng L, Joo Y, *et al.* Evaluation of prostate tumor grades by content-based image retrieval. 27th AIPR Workshop: Advances in Computer-Assisted Recognition; 1999: International Society for Optics and Photonics; 1999. p. 244-52.
- Whiteside T. The tumor microenvironment and its role in promoting tumor growth. *Oncogene* 2008; 27(45): 5904-12.
- Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 2014; 509(7502): 582-7.

- Wimmer I, Tröscher AR, Brunner F, Rubino SJ, Bien CG, Weiner HL, *et al.* Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. *Scientific reports* 2018; 8(1): 1-17.
- Xue M, Zhang K, Mu K, Xu J, Yang H, Liu Y, *et al.* Regulation of estrogen signaling and breast cancer proliferation by an ubiquitin ligase TRIM56. *Oncogenesis* 2019; 8(5): 1-14.
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, *et al.* Ensembl 2016. *Nucleic acids research* 2016; 44(D1): D710-D6.
- Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. *The lancet oncology* 2010; 11(2): 174-83.
- Yin M, Mackley HB, Drabick JJ, Harvey HA. Primary female breast sarcoma: clinicopathological features, treatment and prognosis. *Scientific reports* 2016; 6: 31497-.
- Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* 2016; 7: 12474.
- Yuan Y, Failmezger H, Rueda OM, Ali HR, Graf S, Chin SF, *et al.* Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med* 2012; 4(157): 157ra43.
- Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. *Nat Rev Clin Oncol* 2015; 12(7): 381-94.
- Zhang L, Conejo-Garcia JR, Katsaros D, Gimotty PA, Massobrio M, Regnani G, *et al.* Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *New England journal of medicine* 2003; 348(3): 203-13.
- Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *Rna* 2020; 26(8): 903-9.
- Zitvogel L, Apetoh L, Ghiringhelli F, Kroemer G. Immunological aspects of cancer chemotherapy. *Nat Rev Immunol* 2008; 8(1): 59-73.
- Zitvogel L, Kepp O, Kroemer G. Immune parameters affecting the efficacy of chemotherapeutic regimens. *Nat Rev Clin Oncol* 2011; 8(3): 151-60.
- Zujewski JA, Rubinstein L. CREATE-X a role for capecitabine in early-stage breast cancer: an analysis of available data. *NPJ Breast Cancer* 2017; 3(1): 1-5.

10. Appendix: Probe set for NanoString

| GeneName | TargetSeq |
|---------------|--|
| bindspike1.1 | NA |
| bindspike2.1 | NA |
| bindspike3.1 | NA |
| purespikes1.1 | NA |
| purespikes2.1 | NA |
| purespikes3.1 | NA |
| POS_A(128) | TCAGGCCTTGCCCTTACTAATGGCGCGTTGTAACGGGCCTTGAGGGAATGTCACT ATTGAGGCACCCGTTTCGACCCTCAGAGATATACCATTCCGCCTAT |
| POS_B(32) | ATGAAAGCGCTGCTACTATGATAAGAGTACACGTACAGGTCTCGCCCGATTGGA TTATGGCGAGCTGCCGCATTGACGGACATACCTTTGAACGTAATCG |
| POS_C(8) | GCCGCTTTCGCTCGGGTCTGCGGGTTATAGCTTTTCAGTCTCGACGGGCTAGCAC ACATCTGGTTGACTAGGCGCATAGTCGCCATTCACAGATTTGCTC |
| POS_D(2) | TACCTGGCATTGTTGGGCACTTCTTGCCTTTAAGCGGGAAAGATCGCGAGGGCCCG CTATTTGCGATACTTCCCATGTCGGTGCCGTCGCCTCTATGTACTC |
| POS_E(0.5) | GGTTGAATTTGAGCGGATGGGCTCAACTGCGTCGTAACCGGTAGATACAGGGCA TACGAGCCTCCCTATTTAACGGCATCATCCCGCGTAGTGCTGGTCA |
| POS_F(0.125) | ACGAAGCTGTTTCGGCCGCACGCAAGTACCTCCCCTTAGAAAGCGAATAACCCA ACGACCGTGTTCAACCCTGGCCGTCTCTCAACCAGGTATGCAATCA |
| NEG_A(0) | AACCGCCGCATACGGCCGATTGTCGCAGCCCGGGTCGATTATAACAACGGTGCA ATCTCAGCTAAACCGACGCAGTTTTGCTCCTTGGATTCTGAGCCCG |
| NEG_B(0) | GCACTGGCATTGGTCGTTTCAGGAGGCCATACAGAACTGGTTTATATGAAGGAA CATGGATCATTGAAGTCATTGGGGAAACCCTTGATGATGCGGCAG |
| NEG_C(0) | GTACAGGCTGCTGGCTCATGTTTCCTTCTACGCTGCACTTGCGGGCATAGAGGTC GGTTGCGATCTATATTCGGAGATAACTATTCACCCAGCGCCACTC |

| | |
|----------|--|
| NEG_D(o) | AGAGATCACGTGGACCAAAGCTGATTGATTACGGGACTGGCCGTAAGTGCTGCC CGCGAGTAGATCGTCTAGATCCGGCTAAAATTCCCTGCGGTGCCTT |
| NEG_E(o) | CCCAGATGACCTTCTCCCTCATAATCACTTAATCTGAGCGCAGGAGGCAGGCTGT ATTAATTCCGGCCTCCAACCGGACCGTGGAACGACGCGACCAAGT |
| NEG_F(o) | GGGTCCTTACCGGCTGTAAGCTCACTACAATCCAGGTACAGAGTGCGTTAACCG GCCATTAGAGGGCCGCTACACCCGTCAGAATTTAAACGTATGGGCG |
| NEG_G(o) | ACCCGTATGAACTGTTGCCGGCTCGGAAATGTTAAGGCTCTGCGCACGCACTTTA TCATTGCGAGCCTGTTCTGTCAGCGGGTCAGCCTAGGTTACGGTG |
| NEG_H(o) | TTGGTCCGAGGAGGCATATAGGAAACGATGGGCACGCGCTATTCAGACGTTATT TGGTATGGAGTAAGAGGCCGAAACTGGGCTCGATTGATGGATACT |
| ITGA6 | CTCATGCGAGCCTTCATTGATGTGACTGCTGCTGCCGAAAATATCAGGCTGCCAA ATGCAGGCACTCAGGTTGAGTGACTGTGTTTTCCCTCAAAGACTG |
| TP53 | GGGGAGCAGGGCTCACTCCAGCCACCTGAAGTCCAAAAGGGTCAGTCTACCTC CCGCCATAAAAAACTCATGTTCAAGACAGAAGGGCCTGACTCAGAC |
| IGF2R | TGCCGGGCTGCAACCGATATGCATCGGCTTGCCAGATGAAGTATGAAAAGATC AGGGCTCCTTCACTGAAGTGGTTTTCCATCAGTAACTTGGGAATGGC |
| PTGS2 | GCTACAAAAGCTGGGAAGCCTTCTAACCTCTCCTATTATACTAGAGCCCTTCC TCCTGTGCCTGATGATTGCCCGACTCCCTTGGGTGTCAAAGGTAA |
| ERBB2 | ACAGACACGTTTGAGTCCATGCCCAATCCCAGGGCCGGTATACATTGGGCGCC AGCTGTGTGACTGCCTGTCCCTACAACCTTTCTACGGACGTGG |
| CCNA2 | CGGGACAAAGCTGGCCTGAATCATTAAACGAAAGACTGGATATACCCTGGAAA GTCTTAAGCCTTGTCTCATGGACCTTACCAGACCTACCTCAAAGC |
| IL13RA1 | TCTGCACTGGAAGAAGTACGACATCTATGAGAAGCAAACCAAGGAGGAAACCGA CTCTGTAGTGCTGATAGAAAACCTGAAGAAAGCCTCTCAGTGATGG |
| IRAK2 | GTGTTGGCCGAGGTCTCACGGGCATCCCTGCAATGGATAACAACCGAAGCCCG GTTTACCTGAAGGACTTACTCCTCAGTGATATTCCAAGCAGCACCG |
| ACTG1 | AACCAGTTGTTTTGTCTTGCGGGTCTGTCAGGGTTGGAAAGTCCAAGCCGTAGG AACCAGTTTCCTTTCTTAGCTGATGTCTTTGGCCAGAACACCGTGG |

| | |
|--------|--|
| BCL6 | GTTGTGGACACTTGCCGGAAGTTTATTAAGGCCAGTGAAGCAGAGATGGTTTCT GCCATCAAGCCTCCTCGTGAAGAGTTCCTCAACAGCCGGATGCTGA |
| IRF1 | CTGTGCGAGTGTACCGGATGCTTCCACCTCTACCAAGAACCAGAGAAAAGAAA GAAAGTCGAAGTCCAGCCGAGATGCTAAGAGCAAGGCCAAGAGGAA |
| MYD88 | ACGTTTTTCTAGGTACAGCTCCCAGGAACAGCTAGGTGGGAAAAGTCCCATCACT GAGGGAGCCTAACCATGTCCCTGAACAAAATTGGGCACTCATCTA |
| STAT6 | AGAACATCCAGCCATTCTCTGCCAAAGACCTGTCCATTCGCTCACTGGGGGACCG AATCCGGGATCTTGCTCAGCTCAAAAATCTCTATCCCAAGAAGCC |
| TFRC | CAGTTTCCACCATCTCGGTCATCAGGATTGCCTAATATACTGTCCAGACAATCT CCAGAGCTGCTGCAGAAAAGCTGTTTGGGAATATGGAAGGAGACT |
| CDC45 | AGTCCCAGGATGCTGCACAACCATTTTGACCTCTCAGTAATTGAGCTGAAAGCTG AGGATCGGAGCAAGTTTCTGGACGCACCTATTTCCCTCCTGTCCT |
| CD83 | CTGTTCTTGAAGCAGTAGCCTAACACACTCCAAGATATGGACACACGGGAGCCG CTGGCAGAAGGGACTTCACGAAGTGTTCATGGATGTTTTAGCCAT |
| CCL20 | ATCTGTTCTTTGAGCTAAAAACCATGTGCTGTACCAAGAGTTTGCTCCTGGCTGC TTTGATGTCAGTGCTGCTACTCCACCTCTGCGGCGAATCAGAAGC |
| OAS3 | GAGTGCCTTAGACAGCCTGACTCTCCACAAACCACTGTAAAACTTACCTGCTAG GAATGCTAGATTGAATGGGATGGGAAGAGCCTTCCCTCATTATTG |
| NFATC2 | GACGGACATTGGAAGAAAGAACACGCGGGTGAGACTGGTTTTCCGAGTTCACAT CCCAGAGTCCAGTGGCAGAATCGTCTCTTACAGACTGCATCTAAC |
| BLNK | ACACCACTGAAGACAACCTCCAGTTGCCTCTCAACAGAATGCTTCAAGTGTGTG AAGAAAACCTATACCTGCTGAACGCCACCGAGGGTCAAGTCACA |
| MRPS23 | AAATCCGAACACTTGAGTGTGACACCACAGACTGCGTTGGAAGAAAACGAGACT CAGAAAGAAGTTCACAGGACCAGCATTGGAGGCACCTGCAGACC |
| POLR3K | TGCTTGCCTTGTCCTCGGGTAGATGCTTAGCTGGCAGTATGAGTTGTGTGTCC TGAGGGTCTTTGCTAGTGTGGTGGAAAGATAAACCTTTTGAGGTG |
| BRF2 | TACCTTCAGCGACGAGGGCAATCTCCGAGAGGTAACATATCCCGAAGCACAGG GGAAAACGAACAAGTTAGTCGCAGCCAGCAACGAGGTCTCCGGCGA |

| | |
|----------|---|
| MCM10 | AATAACTTCTTGACGCGGGAAAATGGCGAGCCCGACGCATTTGATGAGCTCTTT GATGCCGACGGCGACGGTGAATCTTATACAGAAGAGGCTGATGATG |
| ITGB2 | CATCGACCTGTACTATCTGATGGACCTCTCCTACTCCATGCTTGATGACCTCAGG AATGTCAAGAAGCTAGGTGGCGACCTGCTCCGGGCCCTCAACGAG |
| TNFRSF1B | CCCAGCTGAAGGGAGCACTGGCGACTTCGCTCTTCCAGTTGGACTGATTGTGGG TGTGACAGCCTTGGGTCTACTAATAATAGGAGTGGTGAACCTGTGTC |
| CTSG | AGTCCAGCAGGTCAGAGCAGATGTGGAGGGTTCCTGGTGCGAGAAGACTTTGTG CTGACAGCAGCTCATTGCTGGGGAAGCAATATAAATGTCACCCTGG |
| FPR1 | GCCATGGGAGGACATTGGCCTTTCGGCTGGTTCCTGTGCAAATTCGTCTTTACCA TAGTGGACATCAACTTGTTTCGGAAGTGTCTTCCTGATCGCCCTCA |
| S100A8 | CTGATAAAGGGGAATTTCCATGCCGTCTACAGGGATGACCTGAAGAAATTGCTA GAGACCGAGTGTCTCAGTATATCAGGAAAAGGGTGCAGACGTCT |
| IL1B | GGGACCAAAGGCGGCCAGGATATAACTGACTTCACCATGCAATTTGTGTCTTCTT AAAGAGAGCTGTACCCAGAGAGTCCTGTGCTGAATGTGGACTCAA |
| CXCL8 | ACAGCAGAGCACACAAGCTTCTAGGACAAGAGCCAGGAAGAAACCACCGGAAGG AACCATCTCACTGTGTGTAACATGACTTCCAAGCTGGCCGTGGCT |
| CSF2 | AGATGAGGCTGGCCAAGCCGGGAGCTGCTCTCTCATGAAACAAGAGCTAGAAA CTCAGGATGGTCATCTTGGAGGGACCAAGGGGTGGGCCACAGCCAT |
| HSD11B1 | GCCTACTACTATTCTGCAAACGAGGAATTCAGACCAGAGATGCTCCAAGGA AAGAAAGTGATTGTACAGGGGCCAGCAAAGGGATCGGAAGAGAGA |
| BTK | TGATCTGGTTCAGAAATATCACCCCTTGCTTCTGGATCGATGGGCAGTATCTCTGC TGCTCTCAGACAGCCAAAAATGCTATGGGCTGCCAAATTTTGGAG |
| CD40LG | GCATTTGATTTATCAGTGAAGATGCAGAAGGGAAATGGGGAGCCTCAGCTCACA TTCAGTTATGGTTGACTCTGGGTTTCTATGGCCTTGTTGGAGGGGG |
| CDKN1A | CATGTGTCCTGGTTCCTGTTTCTCCACCTAGACTGTAAACCTCTCGAGGGCAGGG ACCACACCCTGTAAGTGTCTGTGCTTTTACAGCTCCTCCACAA |
| IL10 | AAGGATCAGCTGGACAACCTGTTGTTAAAGGAGTCCTTGCTGGAGGACTTTAAG GGTTACCTGGGTTGCCAAGCCTTGCTGAGATGATCCAGTTTTTACC |

| | |
|---------|---|
| IL1A | ACTCCATGAAGGCTGCATGGATCAATCTGTGTCTCTGAGTATCTCTGAAACCTCT AAAACATCCAAGCTTACCTTCAAGGAGAGCATGGTGGTAGTAGCA |
| SPP1 | CGCCTTCTGATTGGGACAGCCGTGGGAAGGACAGTTATGAAACGAGTCAGCTGG ATGACCAGAGTGCTGAAACCCACAGCCACAAGCAGTCCAGATTATA |
| TNF | AGCAACAAGACCACCACTTCGAAACCTGGGATTCAGGAATGTGTGGCCTGCACA GTGAAGTGCTGGCAACCACTAAGAATTCAAACCTGGGGCCTCCAGAA |
| LTA | CTGATCAAGTCACCGGAGCTTTCAAAGAAGGAATTCTAGGCATCCCAGGGGACC ACACCTCCCTGAACCATCCCTGATGTCTGTCTGGCTGAGGATTTC |
| FASLG | TCCATGCCTCTGGAATGGGAAGACACCTATGGAATTGTCCTGCTTTCTGGAGTGA AGTATAAGAAGGGTGGCCTTGTGATCAATGAAACTGGGCTGTACT |
| TGFB1 | TATATGTTCTTCAACACATCAGAGCTCCGAGAAGCGGTACCTGAACCCGTGTTGC TCTCCCGGGCAGAGCTGCGTCTGCTGAGGCTCAAGTTAAAAGTGG |
| GSTP1 | TTTTGAGACCCTGCTGTCCCAGAACCAGGGAGGCAAGACCTTCATTGTGGGAGA CCAGATCTCCTTCGCTGACTACAACCTGCTGGACTTGTCTGCTGATC |
| BIRC5 | CCATTCTAAGTCATTGGGGAAACGGGGTGAACCTTCAGGTGGATGAGGAGACAGA ATAGAGTGATAGGAAGCGTCTGGCAGATACTCCTTTTGCCACTGCT |
| TNFRSF8 | GAAACCGCTCAGATGTTTTGGGGAAAGTTGGAGAAGCCGTGGCCTTGCAGAGAGG TGGTTACACCAGAACCTGGACATTGGCCAGAAGAAGCTTAAGTGGG |
| CD70 | CCTATGGGTGCGTCCTGCGGGCTGCTTTGGTCCCATTGGTTCGCGGGCTTGGTGAT CTGCCTCGTGGTGTGCATCCAGCGCTTCGCACAGGCTCAGCAGCA |
| TNFRSF9 | AGATTTGCAGTCCCTGTCTCCAAATAGTTTCTCCAGCGCAGGTGGACAAAGGAC CTGTGACATATGCAGGCAGTGTAAGGTGTTTTTCAGGACCAGGAA |
| MX1 | GCCTTTAATCAGGACATCACTGCTCTCATGCAAGGAGAGGAAACTGTAGGGGAG GAAGACATTCGGCTGTTTACCAGACTCCGACACGAGTTCCACAAAT |
| NFKB2 | ATCTCCGGGGGCATCAAACCTGAAGATTTCTCGAATGGACAAGACAGCAGGCTC TGTGCGGGGTGGAGATGAAGTTTATCTGCTTTGTGACAAGGTGCAG |
| NME3 | TTGGACGGGCTGCTGAACATCCACCTGTCTGGACGTTGCATGGAGGGTGGCGCA GGCCTCTCCAATCCCTGGCGTACAGGGTTTCCTGCCCGAGGAGCTG |

| | |
|---------|---|
| YWHAZ | CTTGGTGGCCATGTACTTTGGAAAAAGGCCGCATGATCTTTCTGGCTCCACTCAGT GTCTAAGGCACCCTGCTTCTTTGCTTGCATCCCACAGACTATTT |
| IKBK1 | TTCTGCGGGAGCGCTGCGAGGAGCTTCTGCATTTCCAAGCCAGCCAGAGGGAGG AGAAGGAGTTCCTCATGTGCAAGTTCCAGGAGGCCAGGAAACTGGT |
| TNFSF10 | GGGGGACCCAGCCTGGGACAGACCTGCGTGCTGATCGTGATCTTCACAGTGCT CCTGCAGTCTCTGTGTGGCTGTAACCTTACGTGTACTTTACCAAC |
| FADD | TGAGACTGCTAAGTAGGGGCAGTGATGGTTGCCAGGACGAATTGAGATAATATC TGTGAGGTGCTGATGAGTGATTGACACACAGCACTCTCTAAATCTT |
| BCL10 | TGAAAATACCATCTTCTTTCAACTACACTTCCCAGACCTGGGGACCCAGGGGCT CCTCCTTTGCCACCAGATCTACAGTTAGAAGAAGAAGGAACCTTGT |
| NFKB1 | AGGGTATAGCTTCCCACACTATGGATTTCTACTTATGGTGGGATTACTTTCCAT CCTGGAACACTAAATCTAATGCTGGGATGAAGCATGGAACCATG |
| CASP5 | TGCAATACAAAGTTTGATCACCTGCCTGCAAGGAATGGGGCTCACTATGACATC GTGGGGATGAAAAGGCTGCTTCAAGGCCTGGGCTACACTGTGGTTG |
| ABL1 | CTGCGTGAGCTATGTGGATTCCATCCAGCAAATGAGGAACAAGTTTGCCTTCCGA GAGGCCATCAACAACTGGAGAATAATCTCCGGGAGCTTCAGATC |
| LCK | ATTAAGTGGACAGCGCCAGAAGCCATTAACCTACGGGACATTCACCATCAAGTCA GATGTGTGGTCTTTTGGGATCCTGCTGACGGAAATTGTCACCCACG |
| IFI16 | ACGACTGAACACAATCAACTGTGAGGAAGGAGATAAACTGAAACTCACCAGCTT TGAATTGGCACCGAAAAGTGGGAATACCGGGGAGTTGAGATCTGTA |
| TNFAIP3 | CAAAGCCCTCATCGACAGAAACATCCAGGCCACCCTGGAAAGCCAGAAGAACT CAACTGGTGTGAGAAAGTCCGGAAGCTTGTGGCGCTGAAAACGAAC |
| TBK1 | ACCAGTCTTCAGGATATCGACAGCAGATTATCTCCAGGTGGATCACTGGCAGAC GCATGGGCACATCAAGAAGGCACTCATCCGAAAGACAGAAATGTAG |
| CD5 | CCAGAAGAAGCAGCGCCAGTGGATTGGCCCAACGGGAATGAACCAAAACATGTC TTTCCATCGCAACCACACGGCAACCGTCCGATCCCATGCTGAGAAC |
| TRAF2 | GTGGCCCTTCAACCAGAAGGTGACCTTAATGCTGCTCGACCAGAATAACCGGGA GCACGTGATTGACGCCTTCAGGCCCCGACGTGACTTCATCCTCTTTT |

| | |
|---------------------|---|
| SOCS ₁ | TTAACTGTATCTGGAGCCAGGACCTGAACTCGCACCTCCTACCTCTTCATGTTTACATATAACCCAGTATCTTTGCACAAACCAGGGGTTGGGGGAGGGTC |
| BIRC ₂ | TGGGATCCACCTCTAAGAATACGTCTCCAATGAGAAACAGTTTTGCACATTCATTATCTCCCACCTTGGAACATAGTAGCTTGTTTCAGTGGTTCTTACTC |
| IFNG | ATACTATCCAGTACTGCCGGTTTGAAAATATGCCTGCAATCTGAGCCAGTGCTTTAATGGCATGTCAGACAGAACTTGAATGTGTCAGGTGACCCTGAT |
| IL ₄ | GACACTCGCTGCCTGGGTGCGACTGCACAGCAGTTCCACAGGCACAAGCAGCTGATCCGATTCTGAAACGGCTCGACAGGAACCTCTGGGGCCTGGCGG |
| TNFSF ₁₁ | TACCTGATTCATGTAGGAGAATTAACAGGCCTTTCAAGGAGCTGTGCAAAAGGAATTACAACATATCGTTGGATCACAGCACATCAGAGCAGAGAAAGC |
| IFNGR ₁ | CCCGGGCAGCCATCTGACTCCAATAGAGAGAGAGAGTTCTTCACCTTTAAGTAGTAACCAGTCTGAACCTGGCAGCATCGCTTTAAACTCGTATCACTCC |
| TICAM ₁ | GGGGGAACCTTCAGGATGAGGCCCGAAACCGGTGTGGGTGGGACATTGCTGGGGA TCCAGGGAGCATCCGGACGCTCCAGTCCAATCTGGGCTGCCTCCCA |
| LTB | AGGAACAGGCGTTTCTGACGAGCGGGACGCAGTTCTCGGACGCCGAGGGGCTGGCGCTCCCGCAGGACGGCCTCTATTACCTCTACTGTCTCGTCGGCTA |
| IFNB ₁ | ACAGACTTACAGGTTACCTCCGAAACTGAAGATCTCCTAGCCTGTGCCTCTGGGACTGGACAATTGCTTCAAGCATTCTTCAACCAGCAGATGCTGTTTA |
| IL ₂ | AGGATGCAACTCCTGTCTTGCAATTGCACTAAGTCTTGCACTTGTCAAAACAGTGCACCTACTTCAAGTTCTACAAAGAAAACACAGCTACAACCTGGAGC |
| IL ₂ RA | CTTGTAAGAAGCCGGGAACAGACAACAGAAGTCATGAAGCCCAAGTGAAATCAAAGGTGCTAAATGGTGCAGGAGACATCCGTTGTGCTTGCCTGC |
| IL ₁₁ | TGAGACAGAGAACAGGGAATTAATGTGTCATACATATCCACTTGAGGGCGATTGTCTGAGAGCTGGGGCTGGATGCTTGGGTAACCTGGGGCAGGGCAG |
| ABCB ₁ | TATAGCACTAAAGTAGGAGACAAAGGAACTCAGCTCTCTGGTGGCCAGAAACAA CGCATTGCCATAGCTCGTGCCTTGTAGACAGCCTCATATTTTGC |
| CR ₂ | GGTGTCAAGCAAATAATATGTGGGGGCCGACACGACTACCAACCTGTGTAAGTGT TTTTCCCTCTCGAGTGTCCAGCACTTCTATGATCCACAATGGACA |

| | |
|--------|--|
| CHUK | TAGAACCCATGGAAAACCTGGCTACAGTTGATGTTGAATTGGGACCCTCAGCAGA GAGGAGGACCTGTTGACCTTACTTTGAAGCAGCCAAGATGTTTTGT |
| IKBKB | GTGATCTATACGCAGCTCAGTAAAACCTGTGGTTTTGCAAGCAGAAGGCGCTGGAA CTGTTGCCCAAGGTGGAAGAGGTGGTGAGCTTAATGAATGAGGATG |
| CCND3 | GGCCAGCCATGTCTGCATTTCCGGTGGCTAGTCAAGCTCCTCCTCCCTGCATCTGA CCAGCAGCGCCTTTCCCAACTCTAGCTGGGGGTGGGCCAGGCTGA |
| IL17A | TACTACAACCGATCCACCTCACCTTGAATCTCCACCGCAATGAGGACCCTGAGA GATATCCCTCTGTGATCTGGGAGGCAAAGTGCCGCCACTTGGGCT |
| CXCL9 | CACCATCTCCCATGAAGAAAGGGAACGGTGAAGTACTAAGCGCTAGAGGAAGCA GCCAAGTCGGTTAGTGGAAGCATGATTGGTGCCAGTTAGCCTCTG |
| IRF4 | GGGCACTGTTTAAAGGAAAGTTCCGAGAAGGCATCGACAAGCCGGACCCTCCCA CCTGGAAGACGCGCCTGCGGTGCGCTTTGAACAAGAGCAATGACTT |
| PDGFRB | CGTGGGCTTCCTCCCTAATGATGCCGAGGAACTATTCATCTTTCTCACGGAAATA ACTGAGATCACCATTCCATGCCGAGTAACAGACCCACAGCTGGTG |
| PSMB9 | TCAGGTATATGGAACCCTGGGAGGAATGCTGACTCGACAGCCTTTTGCCATTGGT GGCTCCGGCAGCACCTTTATCTATGGTTATGTGGATGCAGCATAT |
| CCL5 | AGTGTGTGCCAACCCAGAGAAGAAATGGGTTCCGGGAGTACATCAACTCTTTGGA GATGAGCTAGGATGGAGAGTCCTTGAACCTGAACTTACACAAATTT |
| TLR3 | CAGGTACCCGATGATCTACCCACAAACATAACAGTGTTGAACCTTACCCATAATC AACTCAGAAGATTACCAGCCGCCAACTTCAACAAGGTATAGCCAGC |
| PRKCD | ACCATAAACTGGACTCTGCTGGAAAAGCGGAGGTTGGAGCCACCTTTCAGGCCC AAAGTGAAGTCACCCAGAGACTACAGTAACTTTGACCAGGAGTTCC |
| DDX58 | CTGGCATATTGACTGGACGTGGCAAAACAAATCAGAACACAGGAATGACCCTCC CGGCACAGAAGTGTATATTGGATGCATTCAAAGCCAGTGGAGATCA |
| TLR7 | TGTGGGCACCACACAGGTGGTTGCTGCTTCAGTGCTTCCTGCTCTTTTTCTTGG GCCTGCTTCTGGGTTCCATAGGGAAACAGTAAGAAAGAAAGACAC |
| TLR9 | ACCTTCTTGGCTGTGCCACCCTGGAAGAGCTAAACCTGAGCTACAACAACATCA TGACTGTGCCTGCGCTGCCCAAATCCCTCATATCCCTGTCCCTCA |

| | |
|--------|---|
| NFKBIA | GGATGAGGAGAGCTATGACACAGAGTCAGAGTTCACGGAGTTCACAGAGGACGA GCTGCCCTATGATGACTGTGTGTTTGGAGGCCAGCGTCTGACGTTA |
| CCR5 | TAGGAACATACTTCAGCTCACACATGAGATCTAGGTGAGGATTGATTACCTAGTA GTCATTTTCATGGGTTGTTGGGAGGATTCTATGAGGCAACCACAGG |
| CXCL11 | CAGAATTCCACTGCCCAAAGGAGTCCAGCAATTAATGGATTTCTAGGAAAAGC TACCTTAAGAAAGGCTGGTTACCATCGGAGTTTACAAAGTGCTTTC |
| VEGFA | GAGTCCAACATCACCATGCAGATTATGCGGATCAAACCTCACCAAGGCCAGCAC ATAGGAGAGATGAGCTTCTACAGCACAACAAATGTGAATGCAGAC |
| CXCL10 | GCAGAGGAACCTCCAGTCTCAGCACCATGAATCAAACCTGCGATTCTGATTTGCTG CCTTATCTTTCTGACTCTAAGTGGCATTCAAGGAGTACCTCTCTC |
| IDO1 | CTATTATAAGATGCTCTGAAAACCTTTCAGACACTGAGGGGCACCAGAGGAGCA GACTACAAGAATGGCACACGCTATGGAAAACCTCTGGACAATCAGT |
| AURKB | AGATGCTCTAATGTACTGCCATGGGAAGAAGGTGATTCACAGAGACATAAAGCC AGAAAATCTGCTCTTAGGGCTCAAGGGAGAGCTGAAGATTGCTGAC |
| TBX21 | ACACAGGAGCGCACTGGATGCGCCAGGAAGTTTCATTTGGGAAACTAAAGCTCA CAAACAACAAGGGGGCGTCCAACAATGTGACCCAGATGATTGTGCT |
| BID | GCTTAGCTTTAGAAACAGTGCAACACTGGTCTGCTGTTCCAGTGGTAAGCTATGT CCCAGGAATCAGTTTAAAAGCACGACAGTGGATGCTGGGTCCATA |
| CCR4 | GGTCCTTCTTAGCATCGTGCTTCTGAGCAAGCCTGGCATTGCCTCACAGACCTT CCTCAGAGCCGCTTTCAGAAAAGCAAGCTGCTTCTGGTTGGGCCC |
| RORC | CTCATCAATGCCCATCGGCCAGGGCTCCAAGAGAAAAGGAAAGTAGAACAGCTG CAGTACAATCTGGAGCTGGCCTTTCATCATCATCTCTGCAAGACTC |
| IL17F | GCCCCGCTGTGCCAGGAGGTAGTATGAAGCTTGACATTGGCATCATCAATGAAA ACCAGCGGTTTCCATGTCACGTAACATCGAGAGCCGCTCCACCTC |
| STAT3 | AGACTTGGGCTTACCATTGGGTTTAAATCATAGGGACCTAGGGCGAGGGTTCAG GGCTTCTCTGGAGCAGATATTGTCAAGTTCATGGCCTTAGGTAGCA |
| STAT1 | ACAGTGGTTAGAAAAGCAAGACTGGGAGCACGCTGCCAATGATGTTTCATTTGC CACCATCCGTTTTTCATGACCTCCTGTACAGCTGGATGATCAATAT |

| | |
|-----------|---|
| TNFRSF13B | TGCAAAACCATTTGCAACCATCAGAGCCAGCGCACCTGTGCAGCCTTCTGCAGGT CACTCAGCTGCCGCAAGGAGCAAGGCAAGTTCTATGACCATCTCC |
| SMAD3 | TTAAAGGACAGTTGAAAAGGGCAAGAGGAAACCAGGGCAGTTCTAGAGGAGTGC TGGTGACTGGATAGCAGTTTTTAAGTGGCGTTCACCTAGTCAACACG |
| ITGA4 | GCCCACTGCCAACTGGCTCGCCAACGCTTCAGTGATCAATCCCGGGGCGATTTAC AGATGCAGGATCGGAAAGAATCCCGGCCAGACGTGCGAACAGCTC |
| CSF1R | CATACTGGTACTGCTGTAATGAGCCAAGTGGCAGCTAAAAGTTGGGGGTGTTCT GCCCAGTCCCGTCATTCTGGGCTAGAAGGCAGGGGACCTTGGCATG |
| GPLY | CAGGAGCTGGGCCGTGACTACAGGACCTGTCTGACGATAGTCCAAAACTGAAG AAGATGGTGGATAAGCCCACCCAGAGAAGTGTTC AATGCTGCGA |
| S100A12 | CAAGATGAACAGGTCGACTTTCAAGAATTCATATCCCTGGTAGCCATTGCGCTGA AGGCTGCCATTACCACACCCACAAAGAGTAGGTAGCTCTCTGAA |
| CD59 | GACTTGAAGTAGATTGCATGCTTCCTCCTTTGCTCTTGGGAAGACCAGCTTTGCA GTGACAGCTTGAGTGGGTTCTCTGCAGCCCTCAGATTATTTTTCC |
| PLAUR | GAGAAGACCAACAGGACCCTGAGCTATCGGACTGGCTTGAAGATCACCAGCCTT ACCGAGGTTGTGTGTGGGTTAGACTTGTGCAACCAGGGCAACTCTG |
| BCL2L1 | ATCTTGGCTTTGGATCTTAGAAGAGAATCACTAACCAGAGACGAGACTCAGTGA GTGAGCAGGTGTTTTGGACAATGGACTGGTTGAGCCCATCCCTATT |
| TNFSF4 | GAAGGTCAGGTCTGTCAACTCCTTGATGGTGGCCTCTCTGACTTACAAAGACAAA GTCTACTTGAATGTGACCACTGACAATACCTCCCTGGATGACTTC |
| CD69 | AGGACATGAACTTTCTAAAACGATACGCAGGTAGAGAGGAACACTGGGTTGGAC TGAAAAAGGAACCTGGTCACCCATGGAAGTGGTCAAATGGCAAAGA |
| GZMB | ACACTACAAGAGGTGAAGATGACAGTGCAGGAAGATCGAAAGTGCGAATCTGAC TTACGCCATTATTACGACAGTACCATTGAGTTGTGCGTGGGGGACC |
| PRF1 | ACTGTTTTTCAGGGAGGTGGCTGGGTTTACACGCTAATCCCGATTACCCTGTCC AAACTGCCTAAGCCCTCCGCCATTCTCAAGCCCTGCAGTCACAGC |
| CTLA4 | AGTCTGTGCGGCAACCTACATGATGGGGAATGAGTTGACCTTCTAGATGATTCC ATCTGCACGGGCACCTCCAGTGGAAATCAAGTGAACCTCACTATC |

| | |
|---------------------------------|--|
| ICOS | AACTCTGGCACCCAGGCATGAAGCACGTTGGCCAGTTTTCTCAACTTGAAGTGC AAGATTCTCTTATTTCCGGGACCACGGAGAGTCTGACTTAACTAC |
| TRAT ₁ | ACAGAGGACACAGAAGGACTTGGCAGCAGGGTGATGACCTGATCATTTGTTGAT GGGATGGTGGCTTACCTCTTATTACAGCTTACACTTATGCATGCC |
| BTLA | ACACTCCATCTTAGCAGGAGATCCCTTTGAACTAGAATGCCCTGTGAAATACTGT GCTAACAGGCCTCATGTGACTTGGTGCAAGCTCAATGGAACAACA |
| CD ₂₄₇ | TGGCAGGACAGGAAAAACCCGTCAATGTACTAGGATACTGCTGCGTCATTACAG GGCACAGGCCATGGATGGAAAACGCTCTCTGCTCTGCTTTTTTTCT |
| KLRB ₁ | TGAGTTAAACTTACCCACAGACTCAGGCCAGAAAGTTCTTCACCTTCATCTCTT CCTCGGGATGTCTGTCAGGGTTCACCTTGGCATCAATTTGCCCTG |
| KLRC ₁ | ACCTATCACTGCAAAGATTTACCATCAGCTCCAGAGAAGCTCATTGTTGGGATCC TGGGAATTATCTGTCTTATCTTAATGGCCTCTGTGGTAACGATAG |
| PDCD ₁ | CTTCTTCCCAGCCCTGCTCGTGGTGACCGAAGGGGACAACGCCACCTTCACCTGC AGCTTCTCCAACACATCGGAGAGCTTCGTGCTAAACTGGTACCGC |
| KLRK ₁ | GGACCAGGATTTACTTAAACTGGTGAAGTCATATCATTGGATGGGACTAGTACAC ATTCCAACAAATGGATCTTGGCAGTGGGAAGATGGCTCCATTCTC |
| CXCR ₁ | GCAGCCACCAGTCCATTGGGCAGGCAGATGTTCTAATAAAGCTTCTGTTCCGTG CTTGTCCTGTGGAAGTATCTTGGTTGTGACAGAGTCAAGGGTGT |
| CCR ₁ | CATCATTTGGGCCCTGGCCATCTTGGCTTCCATGCCAGGCTTATACTTTTCCAAG ACCCAATGGGAATTCACCTACCACACCTGCAGCCTTCACTTTCCT |
| CX ₃ CR ₁ | GGGCGCTCAGTCCACGTTGATTTCTCCTCATCTGAATCACAAAGGAGCAGGCATG GAAGTGTTCTGAGCAGCAATTTTACTTACCACACGAGTGATGGAG |
| CXCR ₃ | GTGAGTGACCACCAAGTGCTAAATGACGCCGAGGTTGCCGCCCTCCTGGAGAAC TTCAGCTCTTCTATGACTATGGAGAAAACGAGAGTGAAGTCTGCT |
| CXCR ₂ | AGGAGAAACTGGAAGTCTCGAGCGTTGCTGGGGGGGATTGTAATAATGGTGTGAC CACTGCAGAAGACAGTATGGCAGCTTTCCTCAAACTTCAGACATA |
| CCR ₇ | TTCCGAAAACCAGGCCTTATCTCCAAGACCAGAGATAGTGGGGAGACTTCTTGG CTTGGTGAGGAAAAGCGGACATCAGCTGGTCAAACAACTCTCTGA |

| | |
|---------------------|---|
| CXCR ₄ | ATTGATGTGTGTCTAGGCAGGACCTGTGGCCAAGTTCTTAGTTGCTGTATGTCTC GTGGTAGGACTGTAGAAAAGGGAAGTGAACATTCAGAGCGTGTA |
| CMKLR ₁ | CAACGTCTTCCTCCCAATCCATATCACCTATGCCGCCATGGACTACCACTGGGTT TTCGGGACAGCCATGTGCAAGATCAGCAACTTCCTTCTCATCCAC |
| XCR ₁ | GTCCTCCATCCTCGACACCATCTTCCACAAGGTGCTTTCTTCGGGCTGTGATTAT TCCGAATCACGTGGTACCTCACCTCCGTCTACCAGCACAACTC |
| CXCR ₆ | TTACCATGAAGACTATGGGTTCCAGCAGTTTCAATGACAGCAGCCAGGAGGAGCA TCAAGACTTCCTGCAGTTCAGCAAGGTCTTTCTGCCCTGCATGTAC |
| CCR ₆ | CTTTAACTGCGGGATGCTGCTCCTGACTTGCATTAGCATGGACCGGTACATCGCC ATTGTACAGGCGACTAAGTCATTCCGGCTCCGATCCAGAACACTA |
| ADGRE ₅ | GAACCTGCATTCCAAGAAGCAAGCCGAAGTGGAGGAGATATATGAAAGCAGCAT CCGTGGTGTCCAACCTCAGACGCCTCTCTGCCGTCAACTCCATCTTT |
| SLC _{52A2} | GGTCGCTTCAGCTGCTGCCTTCCAGGGTCTTCTGCTGCTGTTGCCGCCACCACCA TCTGTACCCACAGGGGAGTTAGGATCAGGCCTCCAGGTGGGAGCC |
| CCRL ₂ | GAAATACAAGTGTGCATTTAGCAGAACTCCCTTCTGCCAGCTGATGAGACATTC TGGAAGCATTTTCTGACTTTAAAAATGAACATTTCCGGTTCTTGTC |
| CEP ₅₅ | GTACTIONCCGATTGCTTGAACAGCTGGAAGAGACAACGAGAGAAGGAGAAAGGA GGGAGCAGGTGTTGAAAGCCTTATCTGAAGAGAAAGACGTATTGAA |
| GRB ₇ | GCCGATCTGGCCTCTATTACTCCACCAAGGGCACCTCTAAGGATCCGAGGCACCT GCAGTACGTGGCAGATGTGAACGAGTCCAACGTGTACGTGGTGAC |
| ESR ₁ | AGGAACCAGGGAAAATGTGTAGAGGGCATGGTGGAGATCTTCGACATGCTGCTG GCTACATCATCTCGGTTCCGCATGATGAATCTGCAGGGAGAGGAGT |
| EXO ₁ | TGGCCCAAAAGTAATTAAGCTGCCCGTCTCAGGGGGTAGATTGCCTCGTGG CTCCCTATGAAGCTGATGCGCAGTTGGCCTATCTTAACAAAGCGGG |
| SF _{3A1} | GATGATGAGGTGTACGCACCAGGTCTGGATATTGAGAGCAGCTTGAAGCAGTTG GCTGAGCGGCGTACTGACATCTTCGGTGTAGAGGAAACAGCCATTG |
| PSMC ₄ | CATCGGACAATTTCTGGAGGCTGTGGATCAGAATACAGCCATCGTGGGCTCTAC CACAGGCTCCAACCTATTATGTGCGCATCTGAGCACCATCGATCGG |

| | |
|----------------------|---|
| JAK ₃ | GTGCTGCTGAAGGTCATGGATGCCAAGCACAAGAACTGCATGGAGTCATTCCTG GAAGCAGCGAGCTTGATGAGCCAAGTGTCGTACCGGCATCTCGTGC |
| ZAP70 | GGAGCTCAAGGACAAGAAGCTCTTCCTGAAGCGCGATAACCTCCTCATAGCTGA CATTGAACTTGGCTGCGGCAACTTTGGCTCAGTGCGCCAGGGCGTG |
| IRAK ₁ | CACAGCCGTGGAAGGACTGGCCCTTGGCAGCTCTGCATCATCGTCGTCAGAGCC ACCGCAGATTATCATCAACCCTGCCCGACAGAAGATGGTCCAGAAG |
| BLK | AGCTTCTTGCTCCAATCAACAAGGCCGGCTCCTTTCTTATCAGAGAGAGTGAAAC CAACAAAGGTGCCTTCTCCCTGTCTGTGAAGGATGTCACCACCCA |
| FYN | GTCTTTGGAGGTGTGAACTCTTCGTCTCATAACGGGGACCTTGCGTACGAGAGGA GGAACAGGAGTGACACTCTTTGTGGCCCTTTATGACTATGAAGCAC |
| JAK ₁ | GAGAACACCAAGCTCTGGTATGCTCCAAATCGCACCATCACCGTTGATGACAAG ATGTCCCTCCGGCTCCACTACCGGATGAGGTTCTATTTACCAATT |
| PAK ₁ | TGAACCACTTCTGTCACTCCAACCTCGGGACGTGGCTACATCTCCATTTACCT ACTGAAAATAACACCACTCCACCAGATGCTTTGACCCGGAATACT |
| SRPK ₁ | GGGCCTTTTTGAGGTTCTAGTGGAGAAGTATGAGTGGTCGCAGGAAGAGGCAGC TGGCTTACAGATTTCTTACTGCCCATGTTGGAGCTGATCCCTGAG |
| CSK | GTGCGTGAGCTGCGACGGCAAGGTGGAGCACTACCGCATCATGTACCATGCCAG CAAGCTCAGCATCGACGAGGAGGTGTACTTTGAGAACCTCATGCAG |
| MAP4K ₂ | GAAGGCATCTACACACTCAACCTGCATGAACTGCATGAGGATACGCTGGAGAAG CTGATTTACATCGCTGCTCCTGGCTCTACTGCGTGAACAACGTGC |
| MAPKAPK ₂ | ACGGTGGAGAACTCTTTAGCCGAATCCAGGATCGAGGAGACCAGGCATTCACAG AAAGAGAAGCATCCGAAATCATGAAGAGCATCGGTGAGGCCATCCA |
| JAK ₂ | CTCCTCCCGCGACGGCAAATGTTCTGAAAAAGACTCTGCATGGGAATGGCCTGC CTTACGATGACAGAAATGGAGGGAACATCCACCTCTTCTATATATC |
| PLK ₁ | GCTTGGCTGCCAGTACCTGCACCGAAACCGAGTTATTCATCGAGACCTCAAGCTG GGCAACCTTTTCTGAATGAAGATCTGGAGGTGAAAATAGGGGAT |
| TLK ₂ | ATTCGACGATGCTTGGCCTACCGAAAGGAGGACCGCATTGATGTCCAGCAGCTG GCCTGTGATCCCTACTTGTTCCTCACATCCGAAAGTCAGTCTCTA |

| | |
|--------------------------------|---|
| IKBKE | CCGCATCATCGAACGGCTAAATAGAGTCCCAGCACCTCCTGATGTCTGAGCTCCA TGGGGCACATGAGGCATCCTGAAGCATTAGAATGATTCCAACACT |
| ADCK ₅ | GCTGAGCAGATATTTTACACCGGCTTCATCCACTCGGACCCACATCCTGGCAACG TTCTGGTGCGGAAAGGCCCGGACGGGAAAGCGGAGCTGGTGCTGC |
| CEBPB | CAACCGCACATGCAGATGGGGCTCCCGCCCGTGGTGTTATTTAAAGAAGAAACG TCTATGTGTACAGATGAATGATAAACTCTCTGCTTCTCCCTCTGCC |
| CD ₃ E | AAGTAACAGTCCCATGAAACAAAGATGCAGTCGGGCACTCACTGGAGAGTTCTG GGCCTCTGCCTCTTATCAGTTGGCGTTTGGGGCAAGATGGTAATG |
| EOMES | ATCCCATGCCCTGGGGTATTACCCAGACCCAACCTTTCCTGCAATGGCAGGGTGG GGAGGTCGAGGTTCTTACCAGAGGAAGATGGCAGCTGGACTACCA |
| GZMA | AGACCCTACATGGTCCTACTTAGTCTTGACAGAAAAACCATCTGTGCTGGGGCTT TGATTGCAAAAGACTGGGTGTTGACTGCAGCTCACTGTAACCTGA |
| MS ₄ A ₁ | CTTCTGATGATCCCAGCAGGGATCTATGCACCCATCTGTGTGACTGTGTGGTACC CTCTCTGGGGAGGCATTATGTATATTATTTCCGGATCACTCCTGG |
| THY ₁ | CCTGCCTAGTGGACCAGAGCCTTCGTCTGGACTGCCGCCATGAGAATACCAGCA GTTACCCATCCAGTACGAGTTCAGCCTGACCCGTGAGACAAAGAA |
| TMEM ₁₇₃ | CTGGCATGGTCATATTACATCGGATATCTGCGGCTGATCCTGCCAGAGCTCCAGG CCCGGATTTCGAACTTACAATCAGCATTACAACAACCTGCTACGGG |
| PUM ₁ | CTGGGGAACATCAGATCATTAGTTTCCCAGCCAATCATGGTGCAGAGAAGACC TGGTCAGAGTTTCCATGTGAACAGTGAGGTCAATTCTGTACTGTCC |
| CD68 | ACCGGTCCATCTTGCTGCCTCTCATCATCGGCCTGATCCTTCTTGGCCTCCTCGC CCTGGTGCTTATTGCTTCTGCATCATCCGGAGACGCCCATCCGC |
| SCUBE ₂ | CGTAAAGCCATCCGCACGCTCAGAAAGGCCGTCCACAGGGAGCAGTTTCACCTC CAGCTCTCAGGCATGAACCTCGACGTGGCTAAAAAGCCTCCAGAA |
| CCNB ₂ | AGGTTGATGTTGAACAGCACACTTTAGCCAAGTATTTGATGGAGCTGACTCTCAT CGACTATGATATGGTGCATTATCATCCTTCTAAGGTAGCAGCAGC |
| JTB | AGCTTGACGGTGAATCCTCGGGTTTTAAGGAGAGAGCAAAGTCCTGAGAGGGCG ACGTATTGTCCCTGCTCACCTAGCCCAGAATGAACAAACACGCGCC |

| | |
|----------------------|---|
| TLR ₄ | ACTCAGAAAAGCCCTGCTGGATGGTAAATCATGGAATCCAGAAGGAACAGTGGG TACAGGATGCAATTGGCAGGAAGCAACATCTATCTGAAGAGGAAAA |
| CASP ₁ | TGGAGACATCCCACAATGGGCTCTGTTTTTATTGGAAGACTCATTGAACATATGC AAGAATATGCCTGTTCTGTGATGTGGAGGAAATTTTCCGCAAGG |
| NCAM ₁ | GGTATTTGCCTATCCCAGTGCCACGATCTCATGGTTTCGGGATGGCCAGCTGCTG CCAAGCTCCAATTACAGCAATATCAAGATCTACAACACCCCCTCT |
| PTDSS ₁ | ACTGAGTCCCATGTGAGGTGCTGGTGAGTATTACCTTTCATCTGTGCCATGCTCT AGAACCTTGACCTTGATAGTTCACCACGTCTGATGGATCCCTGTT |
| PSAT ₁ | TGGGGTCTCCTGGGTTTACGCGGCTGTTGATTCAAGGTCAACATTGACCATTGGAG GAGTGGTTTAAAGAGTGCCAGGCGAAGGGCAAACCTGTAGATCGATC |
| GAMT | GCCATCGCAGCGTCAAAGGTGCAGGAGGCGCCATTGATGAGCATTGGATCATC GAGTGCAATGACGGCGTCTTCCAGCGGCTCCGGGACTGGGCCCCAC |
| ANGEL ₂ | AGGTGATATTAAGCTGACGCAATTGGCAATGCTACTGGCAGAGATTTCCAGTGT GCCCACCAGAAAGATGGCAGCTTCTGCCCTATTGTTATGTGTGGT |
| MME | GGATTGTAGGTGCAAGCTGTCCAGAGAAAAGAGTCCTTGTTCCAGCCCTATTCTG CCACTCCTGACAGGGTGACCTTGGGTATTTGCAATATTCCTTTGG |
| PAX ₅ | CTCCAAGAGGAGCACACTTTGGGGAGATGTCTGTTTCCTGCCTCCATTTCTCT GGGACCGATGCAGTATCAGCAGCTCTTTTCCAGATCAAAGAACTC |
| ITGA _{2B} | AGTTACCGCCCAGGCATCCTTTTGTGGCACGTGTCTCCCAGAGCCTCTCCTTTG ACTCCAGCAACCCAGAGTACTTTCGACGGCTACTGGGGGTACTCGG |
| AFF ₃ | GGCTAAATGAGGCCATTACCTAGATTGTCCCTTCAGTAGGAGCCTGAATCCTTCG CTTACCAGGCTGAGAGCTATTTGCCATGATAATGCTGATTCTCTC |
| TNFRSF ₁₇ | TCTGACCATTGCTTTCCACTCCCAGCTATGGAGGAAGGCGCAACCATTCTTGCA CCACGAAAACGAATGACTATTGCAAGAGCCTGCCAGCTGCTTTGA |
| CTSW | TGCACCGAGGGAGCAATACCTGTGGCATCACCAAGTTCCCGCTCACTGCCCCGTG TGCAGAAACCGGATATGAAGCCCCGAGTCTCCTGCCCTCCCTGAAC |
| IFIT ₂ | TGCATCCCATAGAGGTTAGTCCTGCATAGCCAGTAATGTGCTAAGTTCATCCAAA AGCTGGCGGACCAAAGTCTAAATAGGGCTCAGTATCCCCCATCGC |

| | |
|---------|--|
| CD8A | GCTCAGGGCTCTTTCTCCACACCATTTCAGGTCTTTCTTTCCGAGGCCCTGTCT CAGGGTGAGGTGCTTGAGTCTCCAACGGCAAGGGAACAAGTACTT |
| CD79A | AACGAGAAGCTCGGGTTGGATGCCGGGGATGAATATGAAGATGAAAACCTTTAT GAAGGCCTGAACCTGGACGACTGCTCCATGTATGAGGACATCTCCC |
| LAG3 | CTTTTGGTGACTGGAGCCTTTGGCTTTACCTTTGGAGAAGACAGTGGCGACCAA GACGATTTTCTGCCTTAGAGCAAGGGATTCACCCTCCGCAGGCTC |
| LTF | TTCCCCAACCTGTGTGCGCTGTGTGCGGGGACAGGGGAAAACAAATGTGCCTTC TCCTCCCAGGAACCGTACTTCAGCTACTCTGGTGCCTTCAAGTGTC |
| PPBP | GGAAAGGAACCCATTGCAACCAAGTCGAAGTGATAGCCACACTGAAGGATGGGA GGAAAATCTGCCTGGACCCAGATGCTCCCAGAATCAAGAAAATTGT |
| SIGLEC5 | CTTCCTTGAATAGAAAGGTCCTGCTGGCAAGTTCTCTCAAGGCTGGGGATGACCA GGCACAAAAAACAGGGCAGCAATATGTTGGTGTCACTCCCCTTCC |
| CAMP | CATCATTGCCAGGTCCTCAGCTACAAGGAAGCTGTGCTTCGTGCTATAGATGGC ATCAACCAGCGGTCCTCGGATGCTAACCTCTACCGCCTCCTGGAC |
| IFI35 | TGCCCTCTGCTTGCGGGCTCTGCTCTGATCACCTTTGATGACCCCAAAGTGGCTG AGCAGGTGCTGCAACAAAAGGAGCACACGATCAACATGGAGGAGT |
| EBI3 | ATCTTCTCACTGAAGTACTGGATCCGTTACAAGCGTCAGGGAGCTGCGCGCTTCC ACCGGGTGGGGCCCATGAAGCCACGTCCTTCATCCTCAGGGCTG |
| GZMH | AAAAAAGGGACACCTCCAGGAGTCTACATCAAGGTCTCACACTTCTGCCCTGG ATAAAGAGAACAATGAAGCGCCTCTAACAGCAGGCATGAGACTAAC |
| CD58 | GTGCTTGAGTCTCTCCATCTCCCACACTAACTTGTGCATTGACTAATGGAAGCA TTGAAGTCCAATGCATGATACCAGAGCATTACAACAGCCATCGAG |
| CEACAM6 | CTGGCAGATTGGACCAGACCCTGAATTCTTCTAGCTCCTCCAATCCCATTTTATC CCATGGAACCACTAAAAACAAGGTCTGCTCTGCTCCTGAAGCCCT |
| NKG7 | CTGTGGCGGTCCCCGTCTGGCTATGAAACCTTGTGAGCAGAAGGCAAGAGCGG CAAGATGAGTTTTGAGCGTTGTATTCCAAAGGCCTCATCTGGAGCC |
| CCL8 | AAGGAGAGATGGGTCAGGGATTCCATGAAGCATCTGGACCAAATATTTCAAAT CTGAAGCCATGAGCCTTCATACATGGACTGAGAGTCAGAGCTTGAA |

| | |
|---------|---|
| CEACAM8 | ATTTCCCCTTCAGACACCTATTACCATGCAGGGGTAAATCTCAACCTCTCCTGCC ATGCGGCCTCTAATCCACCTCACAGTATTCTTGGTCTGTCAATG |
| LY96 | AGAAGTTATTTGCCGAGGATCTGATGACGATTACTCTTTTTGCAGAGCTCTGAAG GGAGAGACTGTGAATACAACAATATCATTCTCCTTCAAGGGAATA |
| KIR3DL1 | TCTCCATTTCACTTGACCCCTGCCACCTCTCCAACCTAACTGGCTTACTTCTCTAG TCTACTTGAGGCTGCAATCACACTGAGGAACTCACAATTCCAAA |
| CD36 | AGCCAAGGAAAATGTAACCCAGGACGCTGAGGACAACACAGTCTCTTTCCTGCA GCCCAATGGTGCCATCTTCGAACCTTCACTATCAGTTGGAACAGAG |
| AIRE | CAGCGCCACCTCTTGTGCTGCTCGGCTGTAAACAGCTCTGTGTTTCTGGGGACA CCAGCCATCATGTGCCTGGAAATTAACCCTGCCCCACTTCTCTA |
| EGR2 | GGTGGAGCTAGCACTGCCCCCTTCCACCTAGAAGCAGGTTCTTCTAAACTTA GCCCATTTAGTCTCTCTTAGGTGAGTTGACTATCAACCCAAGGC |
| CD55 | CCCTACTCCACCCGTCTTGTGTTGTCCCACCCTTGGTGACGCAGAGCCCCAGCCCA GACCCCGCCCAAAGCACTCATTAACTGGTATTGCGGAGCCACGA |
| TAP1 | GTGGCTGCAGTGGGACAAGAGCCACAGGTATTTGGAAGAAGTCTTCAAGAAAAT ATTGCCTATGGCCTGACCCAGAAGCCAACCTATGGAGGAAATCACAG |
| CD74 | TTCAGCCCCCAGCCCCTCCCCATCTCCCACCCTGTACCTCATCCCATGAGACCC TGGTGCCTGGCTCTTTCGTCACCCTTGGACAAGACAAACCAAGTC |
| NLRP3 | AGTGGGGTTCAGATAATGCACGTGTTTCGAATCCCACTGTGATATGCCAGGAAG ACAGCATTGAAGAGGAGTGGATGGGTTTACTGGAGTACCTTTTCGAG |
| CFB | TAAGCTGAAATATGGCCAGACTATCAGGCCCATTTGTCTCCCCTGCACCGAGGGA ACAACCTCGAGCTTTGAGGCTTCTCCAACCTACCACTTGCCAGCAA |
| CD1D | GGCGTGCTTGCTGTTCTCCTCATTGTGGGCTTTACCTCCCGGTTTAAGAGGCAA ACTTCTATCAGGGCGTCTGTGACTCGCCTTGCCACATCTGTGT |
| HLA-DMB | CCCGTGAGCTGGAAGGAACAGATTTAATATCTAGGGGCTGGGTATCCCCACATC ACTCATTTGGGGGGTCAAGGGACCCGGGCAATATAGTATTCTGCTC |
| ICAM3 | AGCGTCCAGCTGCGAGTCCTGTATGGTCCCAAAATTGACCGAGCCACATGCCCC CAGCACTTGAAATGGAAAGATAAAACGAGACACGTCCTGCAGTGCC |

| | |
|--------|--|
| PLAU | TTCATTGATTACCCAAAGAAGGAGGACTACATCGTCTACCTGGGTGCTCAAGGC TAAACTCCAACACGCAAGGGGAGATGAAGTTTGAGGTGGAAAACC |
| STAT4 | AGACAATGGATCAGAGTGACAAGAATAGTGCCATGGTGAATCAGGAAGTTTTGA CACTGCAGGAAATGCTTAACAGCCTCGATTTCAAGAGAAAGGAGGC |
| CCL7 | CAGCCCCCAGGGGCTTGCTCAGCCAGTTGGGATTAATACTTCAACTACCTGCTGC TACAGATTTATCAATAAGAAAATCCCTAAGCAGAGGCTGGAGAGC |
| CCL23 | AGGGGCGACGTTTCTGTGCCAACCCAGTGATAAGCAAGTTCAGGTTTGCGTGA GAATGCTGAAGCTGGACACACGGATCAAGACCAGGAAGAATTGAAC |
| CLEC4A | ATTTCTACTGAATCAGCATCTTGGAAGACAGTGAGAAGGACTGTGCTAGAATG GAGGCTCACCTGCTGGTGATAAACTCAAGAAGAGCAGGATTTCA |
| TRAF6 | CACCCGCTTTGACATGGGTAGCCTTCGGAGGGAGGGTTTTTCAGCCACGAAGTAC TGATGCAGGGGTATAGCTTGCCCTCACTTGCTCAAAAACAACCTACC |
| HRAS | AGTACATGCGCACCGGGGAGGGCTTCCTGTGTGTGTTTGCCATCAACAACACCA AGTCTTTTGAGGACATCCACCAGTACAGGGAGCAGATCAAACGGGT |
| PSMB4 | CAGGACAGTTTTTACCGCATTCCGTCCACTCCCGATTCTTCATGGATCCGGCGTC TGCACTTTACAGAGGTCCAATCACGCGGACCCAGAACCCCATGGT |
| ATG12 | CCGCTCCTCTCCCGAGGTCTGTAGTCGCGGAGAAACACATGTTGCGTTACTAACG TTCAGAGGTCTGCGACAGCTTCGATTTGAATGACTAGCCGGGAAC |
| CERS2 | GACCGAGATGGACGTGTCTACGCCAAAGCCTCAGATCTCTATATCACGCTGCCCC TGGCCTTGCTCTTCCTCATCGTTTCGATACTTCTTTGAGCTGTACG |
| C6 | ACTCTCCTGTTTGGGCATGTCTTATTCAGTTCAGCTCATGACGCCCTGTAGCAT ACCCCTAGGTACCAACTTCCACAGCAGTCTCGTAAATTCTCCTGT |
| EEF1A1 | ACAAGCCCTTGCGCCTGCCTCTCCAGGATGTCTACAAAATTGGTGGTATTGGTAC TGTTCCCTGTTGGCCGAGTGGAGACTGGTGTCTCAAACCCGGTAT |
| IFITM1 | CCTGTTACTGGTATTCCGGCTCTGTGACAGTCTACCATATTATGTTACAGATAATA CAGGAAAAACGGGGTTACTAGTAGCCGCCATAGCCTGCAACCTT |
| IFNAR2 | AAATACCACAAGATCATTTTTGTGACCTCACAGATGAGTGGAGAAGCACACACGA GGCCTATGTCACCGTCCTAGAAGGATTCAGCGGGAACACAACGTTG |

| | |
|----------|--|
| PSMB8 | ACTCACAGAGACAGCTATTCTGGAGGCGTTGTCAATATGTACCACATGAAGGAA GATGGTTGGGTGAAAGTAGAAAGTACAGATGTCAGTGACCTGCTGC |
| DEFB1 | CTTATAAATACAGTGACGCTCCAGCCTCTGGAAGCCTCTGTCAGCTCAGCCTCCA AAGGAGCCAGCGTCTCCCCAGTTCCTGAAATCCTGGGTGTTGCCT |
| MAVS | ACCCACTGTTGGGGAGATTATCTACAATAACACCAGAAACACATTGGGGTGGAT TGGGGGTATCCTTATGGGTTCTTTTCAGGGAACCATTGCTGGACAA |
| IL1RN | GAGTCTGCCGCTGCCCCGTTGGTTCCTCTGCACAGCGATGGAAGCTGACCAG CCCCTCAGCCTACCAATATGCCTGACGAAGGCGTCATGGTCACCA |
| TIMELESS | AGAACTGTTACAACCGGCTCATGGGATCAGTAAAGGATCACCTGCTTCGGGAGA AAGCTCAGCAGCATGATGAGACCTATTATATGTGGGCCTTGGCTTT |
| CD8B | CAGCTGAGTGTGGTTGATTTCCCTCCCAACTGCCAGCCACCAAGAAGTCCA CCCTCAAGAAGAGAGTGTGCCGGTTACCCAGGCCAGAGACCCAGA |
| HLA-DRA | GGCCAACATAGCTGTGGACAAAGCCAACCTGGAAATCATGACAAAGCGCTCCAA CTATACTCCGATCACCAATGTACCTCCAGAGGTAAGTGTGCTCAG |
| CD3D | TATCTACTGGATGAGTTCCGCTGGGAGATGGAACATAGCACGTTTCTCTCTGGCC TGGTACTGGCTACCCTTCTCTCGCAAGTGAGCCCCTTCAAGATAC |
| ITGAM | GCCCTCCGAGGGTGTCTCAAGAGGATAGTGACATTGCCTTCTTGATTGATGGCT CTGGTAGCATCATCCCACATGACTTTCGGCGGATGAAGGAGTTTG |
| CD1A | CCTGTTTTAGATATCCCTTACTCCAGAGGGCCTTCCCTGACTTACAAGTGGGAAG CAGTCTCTTCTGGTCTGAACTCCCGCCACATTTTAGCCGTA |
| CD9 | ATATTCGCCATTGAAATAGCTGCGGCCATCTGGGGATATTCCACAAGGATGAG GTGATTAAGGAAGTCCAGGAGTTTTACAAGGACACCTACAACAAGC |
| CD7 | CCTACACCTGCCAGGCCATCACGGAGGTCAATGTCTACGGCTCCGGCACCCCTGG TCCTGGTGACAGAGGAACAGTCCAAGGATGGCACAGATGCTCGGA |
| CD22 | TTTTCCAGAAGATGAGGGGATTCATTACTCAGAGCTGATCCAGTTTGGGGTCCG GGAGCGGCCTCAGGCACAAGAAAATGTGGACTATGTGATCCTCAA |
| DPP4 | CAGCAGTCAGCTCAGATCTCCAAAGCCCTGGTCGATGTTGGAGTGGATTTCCAG GCAATGTGGTATACTGATGAAGACCATGGAATAGCTAGCAGCACAG |

| | |
|----------|---|
| ITGAX | CCCCTCAGCCTGTTGGCTTCTGTTACCAGCTGCAAGGGTTTACATACACGGCCA CCGCCATCCAAAATGTCGTGCACCGATTGTTCCATGCCTCATATG |
| PTPRC | AGGAGGAAATTGTTCCCTCGTCTGATAAGACAACAGTGGAGAAAGGACGCATGCT GTTTCTTAGGGACACGGCTGACTTCCAGATATGACCATGTATTTGT |
| CD48 | AATTTAAAGGCAGGGTCAGACTTGATCCTCAGAGTGGCGCACTGTACATCTCTAA GGTCCAGAAAGAGGACAACAGCACCTACATCATGAGGGTGTGAA |
| SLAMF1 | GTGTCTCTTGATCCATCCGAAGCAGGCCCTCCACGTTATCTAGGAGATCGCTACA AGTTTTATCTGGAGAATCTCACCTGGGGATACGGGAAAGCAGGA |
| HLA-DRB3 | GGTCTGAATCTGCACAGAGCAAGATGCTGAGTGGAGTCGGGGGCTTTGTGCTGG GCCTGCTCTTCCTGGGGCCGGGCTGTTTCATCTACTTCAGGAATCA |
| C1R | CAAGTTCCTGGAGCCTTTTGATATTGATGACCACCAGCAAGTACACTGCCCTAT GACCAGCTACAGATCTATGCCAACGGGAAGAACATTGGCGAGTTC |
| C1S | ACTGCACTGATTGGGGAGATTGCAAGTCCCAATTATCCCAAACCATATCCAGAGA ACTCAAGGTGTGAATACCAGATCCGGTTGGAGAAAGGGTTCCAAG |
| CD163 | CATCTGTGATTCGGACTTCTCTCTGGAAGCTGCCAGCGTTCTATGCAGGGAATTA CAGTGTGGCACAGTTGTCTCTATCCTGGGGGGAGCTCACTTTGGA |
| DNAL1 | GGTAAGCTGGAGGACTTCATCCTCAGTTAGGCTGCACAAGTAACATTACCTAAA AGGCACTAACATGCTCAGGTTCCCCAGAAAGAGGCGTAAGAAGGG |
| ITGA5 | AGAAGACTTTGTTGCTGGTGTGCCCAAAGGGAACCTCACTTACGGCTATGTCACC ATCCTTAATGGCTCAGACATTCGATCCCTCTACAACCTTCTCAGGG |
| NME1 | CTTGTGGTTTCACCCTGAGGAACTGGTAGATTACACGAGCTGTGCTCAGAACTGG ATCTATGAATGACAGGAGGGCAGACCACATTGCTTTTCACATCCA |
| RND2 | TGGGACTGTTTCATCCTAGTTAATGAAGTGGGCAATTCTCAGGCCATTAGGGGGTT TTAGAGCAGACCGACATATAATTAGTCAGCATTCTCAGCCCAGC |
| SERPING1 | GACAGAGGCGAAGGGAAGGTCGCAACAACAGTTATCTCCAAGATGCTATTCGTT GAACCCATCCTGGAGGTTTCCAGCTTGCCGACAACCAACTCAACAA |
| CD14 | GCCCAAGCACACTCGCCTGCCTTTTCTGCGAACAGGTTTCGCGCCTTCCCGGCC TTACCAGCCTAGACCTGTCTGACAATCCTGGACTGGGCGAACGCG |

| | |
|----------------------|---|
| HLA-DRB ₄ | TGTGAGTGTCAATTCCTCAATGGGACGGAGCGAGTGTGGAACCTGATCAGATAC ATCTATAACCAAGAGGAGTACGCGCGCTACAACAGTGACCTGGGGG |
| HLA-DRB ₅ | GAGTGTCAATTTCTTCAACGGGACGGAGCGGGTGC GGTTCTTGCACAGAGACATC TATAACCAAGAGGAGGACTTGC GCTTCGACAGCGACGTGGGGGAGT |
| FOXP ₃ | GGGCCATCCTGGAGGCTCCAGAGAAGCAGCGGACACTCAATGAGATCTACCACT GGTTCACACGCATGTTTGCCTTCTTCAGAAACCATCCTGCCACCTG |
| IL18 | GACAGTCAGCAAGGAATTGTCTCCCAGTGCATTTTGCCTCCTGGCTGCCAACTC TGGCTGCTAAAGCGGCTGCCACCTGCTGCAGTCTACACAGCTTCG |
| IL2RG | CCACAGCTGGACTGAACAATCAGTGGATTATAGACATAAGTTCTCCTTGCCTAGT GTGGATGGGCAGAAACGCTACACGTTTCGTGTTCCGAGCCGCTTT |
| IL10RA | TGCCCAGCCCTCCGTCTGTGTGGTTTGAAGCAGAATTTTCCACCACATCCTCCA CTGGACACCCATCCCAAATCAGTCTGAAAGTACCTGCTATGAAGT |
| IL21R | CGTGTTTGTGGTCAACAGATGACAACAGCCGTCCTCCCTCCTAGGGTCTTGTGTT GCAAGTTGGTCCACAGCATCTCCGGGGCTTTGTGGGATCAGGGCA |
| IL2RB | GTCCTGCTGCCCGAGCCAGGAAGTGTGTGTGTTGCAGGGGGGCAGTAACTCCCC AACTCCCTCGTTAATCACAGGATCCCACGAATTTAGGCTCAGAAGC |
| MIEN ₁ | CCTATGAGAAAGATCTCATTGAGGCCATCCGAAGAGCCAGTAATGGAGAAACCC TAGAAAAGATCACCAACAGCCGTCCTCCCTGCGTCATCCTGTGACT |
| CCDC ₂₄ | CAGCTCCCGCCCCATCTCTGACCCCTTTCTCTTCTGGCACCACCGCCTCTCCTA AAGGACCTCTTGCGCCAGGAGCTCC |
| CEACAM ₁ | AGAGGGAGGGGTTATAGCTTCAGGAGGGAACCAGCTTCTGATAAACACAATCTG CTAGGAACTTGGGAAAGGAATCAGAGAGCTGCCCTTCAGCGATTAT |
| CLNS _{1A} | ACATCGCTGAGAGCCGCCTGTCTTGGTTAGATGGCTCTGGATTAGGATTCTCACT GGAATACCCACCATTAGTTTACATGCATTATCCAGGGACCGAAG |
| CYB ₅₆₁ | TCCGCTGCTATAGACCAGTTCATTGTGTGTGGCTCCCGTGTCTCTGTTGCCCCCT TCAGTGCAGAAGGCTTTGGGTAGGACTTCGGGTGTTCCGGTCTG |
| GNPTG | TCACCCGTGTCTGGACCCGTGCATCTCTTCCGACTCTCGGGCAAGTGCTTCAGCC TGGTGGAGTCCACGTACAAGTATGAGTTCTGCCCGTTCCACAACG |

| | |
|--------|---|
| KIF2C | CCTGCTCTAACGGGGCGCTGATTCCAGGCAATTTATCCAAGGAAGAGGAGGAAC TGTCTTCCCAGATGTCCAGCTTTAACGAAGCCATGACTCAGATCAG |
| TAF2 | ACTTCACATGACTGGAGGTTACGGTGTGGTGTGGACTTGTACTTCACACTTT TTGGCCTCAGTAGACCTTCTGTTTACCCTTGCCAGAGCTTGGGT |
| TBC1D9 | ACCAATAAAGACAGCACACTGCCTCCCATTCTCACCTCCACTCCTTGCTCAGCG ATGATGTGGAACCTTACCCTGAGGTAGACATCTTTAGACTCATCA |
| VPS45 | CAGAACCCCAAAGTGACAGAGTTTGATGCTGCCCCGCTGGTGTATGCTTTATGCTT TACATTATGAGCGACACAGCAGCAATAGCCTGCCAGGACTAATGA |
| EGR1 | GAGGCATACCAAGATCCACTTGC GGCAGAAGGACAAGAAAGCAGACAAAAGTGT TGTGGCCTCTTCGGCCACCTCCTCTCTCTCTTCTACCCGTCCCCG |
| FCGR2B | AGGCTGACAAAGTTGGGGCTGAGAACACAATCACCTATTCACTTCTCATGCACCC GGATGCTCTGGAAGAGCCTGATGACCAGAACCGTATTTAGTCTCC |
| MUC1 | CCCTAGCAGTACCGATCGTAGCCCTATGAGAAGGTTTCTGCAGGTAATGGTGG CAGCAGCCTCTCTTACACAAACCCAGCAGTGGCAGCCACTTCTGCC |
| TAL1 | ACAGCATCTGTAGTCAGCCGACA ACTATTTTCGGCCTTTTGGGGGTGGGTCTGGCC GTACTTGTGATTTTCGATGGTACGTGACCCTCTGCTGAAGACTTGC |
| RECQL4 | GCAAGCAGGCATGGAAGCAGAAGTGGCGGAAGAAAGGGGAGTGTTTTGGGGGT GGTGGTGCCACAGTCACAACCAAGGAGTCTTGTTTCCTGAACGAGCA |
| BCL9 | GCAACCCAGGAAACATGATGTTTTAAGCTGCTAAGATGGGATGTGCCGATCCTT GTCAAAATGAGATTCCAGGTCCTGAGAGCTGCTTTGAGGCAGTTCC |
| PRCC | CCATTAGCTGTGTGTAGTTGCCCGGGACTAGGAGCTTAAGTGAAGAGGTACGCC TTGTTTCGGTGGAATCAGCCGTAGCCATGAGTTTCTGCCGGGGCTA |
| CXCL1 | TATGTTAATATTTCTGAGGAGCCTGCAACATGCCAGCCACTGTGATAGAGGCTGG CGGATCCAAGCAAATGGCCAATGAGATCATTGTGAAGGCAGGGGA |
| CXCL2 | ATCACATGTCAGCCACTGTGATAGAGGCTGAGGAATCCAAGAAAATGGCCAGTG AGATCAATGTGACGGCAGGGAAATGTATGTGTGTCTATTTTGTAAAC |
| CSF1 | TTTCTATGAGACACCTCTCCAGTTGCTGGAGAAGGTCAAGAATGTCTTTAATGAA ACAAAGAATCTCCTTGACAAGGACTGGAATATTTTCAGCAAGAAC |

| | |
|--------|--|
| IL6R | CTTTCTACATAGTGCCATGTGCGTCGCCAGTAGTGTCGGGAGCAAGTTCAGCAA AACTCAAACCTTTCAGGGTTGTGGAATCTTGCAGCCTGATCCGCC |
| PECAM1 | ATCTGCACTGCAGGTATTGACAAAGTGGTCAAGAAAAGCAACACAGTCCAGATA GTCGTATGTGAAATGCTCTCCCAGCCCAGGATTTCTTATGATGCC |
| IL12A | CTTTCTAGATCAAAACATGCTGGCAGTTATTGATGAGCTGATGCAGGCCCTGAAT TTCAACAGTGAGACTGTGCCACAAAAATCCTCCCTTGAAGAACCG |
| CDH5 | TCTCCCCTTCTCTGCCTCACCTGGTCGCCAATCCATGCTCTCTTTCTTTTCTCTGT CTACTCCTTATCCCTTGGTTTAGAGGAACCCAAGATGTGGCCTT |
| IL12B | GCAAGGCTGCAAGTACATCAGTTTTATGACAATCAGGAAGAATGCAGTGTTCTG ATACCAGTGCCATCATACTTGTGATGGATGGGAACGCAAGAGAT |
| TGFB2 | AAGCCAGAGTGCCTGAACAACGGATTGAGCTATATCAGATTCTCAAGTCCAAAG ATTTAACATCTCCAACCCAGCGCTACATCGACAGCAAAGTTGTGAA |
| FOXA1 | TGATACATTCTCAAGAGTTGCTTGACCGAAAGTTACAAGGACCCCAACCCCTTGG TCCTCTTACCCACAGATGGCCCTGGGAATCAATTCCTCAGGAAT |
| CCL11 | TGGGTGCAGGATTCCATGAAGTATCTGGACCAAAAATCTCCAACCTCAAAGCCAT AAATAATCACCATTTTTGAAACCAAACCAGAGCCTGAGTGTTGCC |
| RUNX1 | CAGCCATGAAGAACCAGTTGCAAGATTTAATGACCTCAGGTTTGTGCGTTCGAA GTGGAAGAGGGAAAAGCTTCACTCTGACCATCACTGTCTTCACAAA |
| C1QB | AACTCACTACTGGGCATGGAGGGTGCCAACAGCATCTTTTCCGGGTTCTGCTCT TTCCAGATATGGAGGCCTGACCTGTGGGCTGCTTCACATCCACCC |
| NT5E | ATTCGGGTTTTGAAATGGATAAACTCATCGCTCAGAAAGTGAGGGGTGTGGACG TCGTGGTGGGAGGACACTCCAACACATTTCTTTACACAGGCAATCC |
| MARCO | GCTGAAGTTTACTACAGTGGTACCTGGGGGACAATTTGCGATGACGAGTGGCAA AATTCTGATGCCATTGTCTTCTGCCGCATGCTGGGTTACTCCAAAG |
| TPSAB1 | GCAGGTGAAGGTCCCATAATGGAAAACCACATTTGTGACGCAAAATACCACCT TGGCGCCTACACGGGAGACGACGTCGCGCATCGTCCGTGACGACATG |
| CCL18 | CCCCTTTCCCTTCAACTCTTCGTACATTCAATGCATGGATCAATCAGTGTGATTA GCTTTCTCAGCAGACATTGTGCCATATGTATCAAATGACAAATCT |

| | |
|------------------------|---|
| CD1C | AGGCTGTGGAAGTTTGGCCCAAAGTGTCTGTCATCTACTCAATCATCAGTATGAA GGCGTCACAGAAACAGTGTATAATCTCATAAGAAGCACTTGCCCC |
| PTRH ₂ | AAGCTCCTGATGAAGAAACCCTGATTGCATTATTGGCCCATGCAAAAATGCTGG GACTGACTGTAAGTTTAATTCAAGATGCTGGACGTACTIONCAGATTGC |
| CDCA8 | TCCCTGTTTACTGAAGACCAAATACTGGTTTGGAGACAACCTCCATGTCTTGCTC TTCTACCTCCCTAGTTAGTGGAATTTGGATAAGGGAACCTGTAGG |
| GSTM ₂ | CACAACCTGTGCGGGGAATCAGAAAAGGAGCAGATTCGCGAAGACATTTTGGAG AACCAGTTTATGGACAGCCGTATGCAGCTGGCCAAACTCTGCTATG |
| UBE ₂ C | CTTTTAAGAAGTACCTGCAAGAAACCTACTCAAAGCAGGTCACCAGCCAGGAGC CCTGACCCAGGCTGCCAGCCTGTCCTTGTGTCGTCTTTTTAATTT |
| C8G | CTTCCTGCTTCAAGCCCGAGACGCCCAGGGGCTGTGCACGTGGTTGTCGCTGA GACCGACTACCAGAGTTTCGCTGTCCTGTACCTGGAGCGGGCGGGG |
| FCER ₁ A | GAATCCCCTACTCTACTGTGTGTAGCCTTACTGTTCTTCGCTCCAGATGGCGTGT TAGCAGTCCCTCAGAAACCTAAGGTCTCCTTGAACCTCCATGGA |
| FCER ₁ G | AGTGGTCTTGCTCTTACTCCTTTTGGTTGAACAAGCAGCGGCCCTGGGAGAGCCT CAGCTCTGCTATATCCTGGATGCCATCCTGTTTCTGTATGGAATT |
| TNFRSF ₁₃ C | TCAGACCTCACCATCTTTGACAGCCCTTGAAGGTGGTAGCCAGCTCCTGTTCTCT GTGCCTTCAAAGGCTGGGGCACTATGAGTAAAAGACCGCTTTTA |
| CCL ₂₄ | ATAGTAACCAGCCTTCTGTTCTTGGTGTCTGTGCCACCACATCATCCCTACGG GCTCTGTGGTCATCCCCTCTCCCTGCTGCATGTTCTTTGTTTCCA |
| IFN ₁ | ATCCCTCTCTTTATCAACAACTTGCAAGAAAGATTAAGGAGGAAGGAATAACAT CTGGTCCAACATGAAAACAATTCTTATTGACTCATAACCAGGTC |
| EPRS | TTGGGAAATTTGTTGAGCTTCCAGGTGCGGAGATGGGAAAGGTTACCGTCAGAT TTCCTCCAGAGGCCAGTGGTTACTTACACATTGGGCATGCAAAAGC |
| RRS ₁ | ACTAAATGTTAAGTTCTAGGCAATTATACGGGGACTCAGAAGGACCTGGCCGCT GCCTTCATTGAGTTTAAAGGGACAGGATTGCCCTTCCGTCAAGAAA |
| TLR ₁ | TCAACCAGGAATTGGAATACTTGGATTTGTCCCAACAAGTTGGTGAAGATTTTC TTGCCACCCTACTGTGAACCTCAAGCACTTGGACCTGTCATTTAA |

| | |
|----------|---|
| IRF7 | CGCAGCGTGAGGGTGTGTCTTCCCTGGATAGCAGCAGCCTCAGCCTCTGCCTGTC CAGCGCCAACAGCCTCTATGACGACATCGAGTGCTTCCTTATGGA |
| IRAK4 | GCTAGACAGACTCTCTTGCTTGGATGGTACTCCACCCTTTCTTGGCACATGAGA TGCAAGATTGCTCAGGGTGCAGCTAATGGCATCAATTTTCTACAT |
| C9 | TTTGACAATGAGTTCTACAATGGACTCTGTAACCGGGATCGGGATGGAAACACT CTGACATACTACCGAAGACCTTGGAACGTGGCTTCTTTGATCTATG |
| MSR1 | AGAGAACTGAAGTCCTTCAAAGCTGCACTGATTGCCCTTACCTCCTCGTGTTT GCAGTTCTCATCCCTCTCATTGGAATAGTGGCAGCTCAACTCCTG |
| C3 | CATCTACCTGGACAAGGTCTCACACTCTGAGGATGACTGTCTAGCTTTCAAAGTT CACCAATACTTTAATGTAGAGCTTATCCAGCCTGGAGCAGTCAAG |
| CYBB | TTTGAAGCATGAAAAAAGAGGGTTGGAGGTGGAGAATTAACCTCCTGCCATGAC TCTGGCTCATCTAGTCCTGCTCCTTGTGCTATAAAATAAATGCAGA |
| IRF5 | GCCCTGATTTCCCTGGTTTGGAGACTCACTTCCTCATCTCCCTGTCCTCTGAGATA ATATGAGTGAGCACTTAGGTATCATATCAGATGCTCAAGGCTGGC |
| IL13 | TTTCTTTCTGATGTCAAAAATGTCTTGGGTAGGCGGGAAGGAGGGTTAGGGAGG GGTAAAATTCCTTAGCTTAGACCTCAGCCTGTGCTGCCCGTCTTCA |
| C1orf106 | TTTTCTAATGTGCTCTGTGATGCACACACCAAGTGGTAGGTCAAAGGTCAGTATA TCCCGGTGGTGTATTGTCTTGCTAGACCCTGCTATTTTCCTGACC |
| MBL2 | TGGACTTGTCTTTTGGTGGACATGGTGCCTAATTTCACTACCTATCCAGGAGTG GAACTGGTAGAGGATGAGGAAAGCATGTATTCAGCTTTAGTAGAT |
| HLA-DPB1 | TCCAAATTGGATACTGCTGCCAAGAAGTTGCTCTGAAGTCAGTTTCTATCATTCT GCTCTTTGATTCAAAGCACTGTTTCTCTCACTGGGCCTCCAACCA |
| IL16 | GGCATCTCCAACATCATCATCCAACGAAGACTCAGCTGCAAATGGTTCTGCTGAA ACATCTGCCTTGGACACAGGGTTCTCGCTCAACCTTTCAGAGCTG |
| TOLLIP | GGAATAGAGCTGTTGATTTAAGGCACACACAATCCCTCACACTGTGGGTTTTTTTT TAGAACTTCCCAGACGAAAACCTCACGCCCTTGGCCCTAACGCGCTT |
| IL22 | CTATCTGATGAAGCAGGTGCTGAACTTCACCCTTGAAGAAGTGCTGTTCCCTCAA TCTGATAGGTTCCAGCCTTATATGCAGGAGGTGGTGCCCTTCCTG |

| | |
|----------------------|---|
| HLA-DPA ₁ | GGAGAGATCTGAACTCCAGCTGCCCTACAACTCCATCTCAGCTTTTCTTCTCAC TTCATGTGAAAACACTCCAGTGGCTGACTGAATTGCTGACCCTT |
| FN ₁ | GGGAATGGACATGCATTGCCTACTCGCAGCTTCGAGATCAGTGCATTGTTGATG ACATCACTTACAATGTGAACGACACATTCCACAAGCGTCATGAAGA |
| HLA-C | AGCTGGGAGCCATCTTCCCAGCCCACCATCCCCATCATGGGCATCGTTGCTGGCC TGGCTGTCTGGTTGTCTAGCTGTCTTGGAGCTGTGGTCACCG |
| CCL ₄ | TTCTGCAGCCTCACCTCTGAGAAAACCTCTTTGCCACCAATACCATGAAGCTCTG CGTGAAGTGTCTCTCTCCTCATGCTAGTAGCTGCCTTCTGCTC |
| FCAR | TGCTGAGATTATAGGCATGAGCCACCACGCCTGGCCAGATGCATGTTCAAACCA ATCAAATGGTGTCTTTCTTATGCAGGACTGATCGATTTGCACCCACC |
| CASP ₄ | CACAGAAAAAAGCCACTTAAGGTGTTGGAATCCCTGGGCAAAGATTTCTCACT GGTGTCTTTGGATAACTTGGTGAACAAAATGTACTGAACTGGAAGG |
| CD ₄₇ | GCCATATTGGTTATTCAGGTGATAGCCTATATCCTCGCTGTGGTTGGACTGAGTC TCTGTATTGCGGCGTGTATACCAATGCATGGCCCTCTTCTGATTT |
| IFNL ₁ | AGCTAGCGAGCTTCAAGAAGGCCAGGGACGCCTTGGAAGAGTCACTCAAGCTGA AAAACCTGGAGTTGCAGCTCTCCTGTCTTCCCCGGGAATTGGGACCT |
| PYCARD | ATGCGGAAGCTCTTCAGTTTCACACCAGCCTGGAAGTGGACCTGCAAGGACTTG CTCCTCCAGGCCCTAAGGGAGTCCCAGTCCCTACCTGGTGGAGGACC |
| PVT ₁ | GATGGCTGTGCCTGTCAGCTGCATGGAGCTTCGTTCAAGTATTTTCTGAGCCTGA TGGATTTACAGTGATCTTCAGTGGTCTGGGGAATAACGCTGGTGG |
| IL _{23A} | CAGGGACAACAGTCAGTTCTGCTTGCAAAGGATCCACCAGGGTCTGATTTTTTAT GAGAAGCTGCTAGGATCGGATATTTTACAGGGGAGCCTTCTCTG |
| CCL ₁₉ | GACCTCAGCCAAGATGAAGCGCCGAGCAGTTAACCTATGACCGTGCAGAGGGA GCCCCGAGTCCGAGTCAAGCATTGTGAATTATTACCTAACCTGGGG |
| IL ₂₇ | CAGGAGCTGCGGAGGGAGTTCACAGTCAGCCTGCATCTCGCCAGGAAGCTGCTC TCCGAGGTTCCGGGGCCAGGCCACCCTTTGCGGAATCTCACCTGC |
| MS _{4A2} | TTCTCACCATTCTGGGACTTGGTAGTGCTGTGTCACTCACAATCTGTGGAGCTGG GGAAGAACTCAAAGGAAACAAGGTTCCAGAGGATCGTGTCTTATGA |

| | |
|----------------------------------|---|
| CPA ₃ | ACCCACCACGTAGCTGCTAATATGATGGTGGATTTCCGAGTTAGTGAGAAGGAA TCCCAAGCCATCCAGTCTGCCTTGGATCAAAATAAAATGCACTATG |
| CCL ₂₂ | CTCGCCCAAGCAGCTGGTAATTCATTTTCATGTATTAGATGTCCCCTGGCCCTCT GTCCCCTCTTAATAACCCTAGTCACAGTCTCCGCAGATTCTTGGG |
| CCL ₁₆ | TTGTCCACGGTTAAAATTATTACAGCAAAGAATGGTCAACCCCAGCTCCTCAACT CCCAGTGATGACCAGGCTTTAGTGGAAAGCCCTTGTTTACAGAAGA |
| IL ₂₅ | CAGCTGCTGCTTAGGGCCGCCGGAAGCTGGTGTCTGTCATTTTCTCTCAGGAAA GGTTTTCAAAGTTCTGCCATTTCTGGAGGCCACCACTCCTGTCT |
| CCL ₂₆ | GGAGTGACATATCCAAGACCTGCTGCTTCCAATACAGCCACAAGCCCCTTCCCTG GACCTGGGTGCGAAGCTATGAATTCACCAGTAACAGCTGCTCCCA |
| CCL ₁₃ | CCAGAATTATATGAAACACCTGGGCCGGAAGCTCACACCCTGAAGACTTGAAC TCTGCTACCCCTACTGAAATCAAGCTGGAGTACGTGAAATGACTTT |
| C8B | GTTTTGAGGGCCCAGTTCTTGATCACAGGTATTATGCAGGTGGATGCTCCCCGCA TTACATCCTGAACACGAGGTTTAGGAAGCCCTACAATGTGGAAAG |
| C8A | GAGCTTCGATATGACTCCACCTGTGAACGTCTCTACTATGGAGATGATGAGAAAT ACTTTCGGAAACCCTACAACTTTCTGAAGTACCCTTTGAAGCCC |
| C ₇ | ATGCTTTTGAAACACAGTCTGTGAACCTACAAGAGGATGTCCAACAGAGGAGG GATGTGGAGAGCGTTTCAGGTGCTTTTCAGGTCAGTGCATCAGCAA |
| IL ₃ | GCCCTTGAAGACAAGCTGGGTAACTGCTCTAACATGATCGATGAAATTATAACA CACTTAAAGCAGCCACCTTTGCCTTTGCTGGACTTCAACAACCTC |
| IL ₁₂ RB ₂ | CCTCCGTGGGACATTAGAATCAAATTTCAAAGGCTTCTGTGAGCAGATGTACCC TTTATTGGAGAGATGAGGGACTGGTACTGCTTAATCGACTCAGAT |
| PPM _{1D} | CCCACTCTTGACCCTCAGAAGCACAAGTATATTATATTGGGGAGTGATGGACTT TGGAATATGATTCCACCACAAGATGCCATCTCAATGTGCCAGGAC |
| IL ₂₆ | CTGTCTCTTGCCATTGCCAAGCACAAGCAATCTTCCTTCACCAAAGTTGTTACC CAAGGGGAACATTGTCCAAGCTGTTGACGCTCTCTATATCAAAG |
| IL ₂₁ | CATGGAGAGGATTGTCATCTGTCTGATGGTCATCTTCTTGGGGACACTGGTCCAC AAATCAAGCTCCCAAGGTCAAGATCGCCACATGATTAGAATGCGT |

| | |
|----------------------------------|--|
| MASP ₂ | GACACTTTCTACTCGCTGGGCTCCAGCCTGGACATTACCTTCCGCTCCGACTACT CCAACGAGAAGCCGTTACGGGGTTCGAGGCCTTCTATGCAGCCG |
| IL ₂₂ RA ₂ | CACTTGCAACCATGATGCCTAAACATTGCTTTCTAGGCTTCCTCATCAGTTTCTT CCTTACTGGTGTAGCAGGAACTCAGTCAACGCATGAGTCTCTGAA |
| IL ₉ | AAGTACTAAAGAACAACAAGTGTCCATATTTTTCTGTGAACAGCCATGCAACCA AACCACGGCAGGCAACGCGCTGACATTTCTGAAGAGTCTTCTGGA |
| CFD | CTGGTTGGTCTTTATTGAGCACCTACTATATGCAGAAGGGGAGGCCGAGGTGGG AGGATCATTGGATCTCAGGAGTTCGAGATCAGCATGGGCCACGTAG |
| PTPN6 | TGGTGCAGACGGAGGCGCAGTACAAGTTCATCTACGTGGCCATCGCCCAGTTCA TTGAAACCACTAAGAAGAAGCTGGAGTCTGCAGTCGCAGAAGGG |
| TNFSF ₁₂ | GCGCCTTTCTGAACCGACTAGTTCGGCCTCGCAGAAGTGCACCTAAAGGCCGG AAAACACGGGCTCGAAGAGCGATCGCAGCCCATTATGAAGTTCATC |
| CENPA | CACTTTGAGCAGTTGCCTGGAAGGCTGGGCATTTCCATCATATAGACCTCTGCCC TTCAGAGTAGCCTCACCATTAGTGGCAGCATCATGTAAGTCTGAGTG |
| NFATC ₃ | GTCCTTGAAGTTCCTCCATATCATAACCCAGCAGTTACAGCTGCAGTGCAGGTGC ACTTTTATCTTTGCAATGGCAAGAGGAAAAAAGCCAGTCTCAAC |
| CR ₁ | CGCGGAGCACAATGATTGGTCACTCCTATTTTCGCTGAGCTTTTCTCTTATTTT AGTTTTCTTCGAGATCAAATCTGGTTTGTAGATGTGCTTGGGGAG |
| CTSS | ATGACAACGGCTTTCCAGTACATCATTGATAACAAGGGCATCGACTCAGACGCTT CCTATCCCTACAAAGCCATGGATCAGAAATGTCAATATGACTCAA |
| SPIB | CTTTGTCATGTACAGACTCCCTGGGATCCTCATGTTTTGGGTGACAGGACCTATG GACCACTATACTCGGGGAGGCAGGGTAGCAGTTCTTCCAGAATCC |
| IL ₃₄ | CGGCCCAAAGCCCTGCTGGACAACCTGCTTCCGGGTCATGGAGCTGCTGTACTGC TCCTGCTGTAAACAAAGCTCCGTCCTAAACTGGCAGGACTGTGAGG |
| CCL ₁₅ | TGCTAACACCTCCTGGTTGGAACCTACAGGAATAGAACTGGAAAGGGAAAAAAGG CAGCATTCACCACATCCCAATCCTGAATCCAAGAGTCTAAGATAGT |
| RAB ₂₅ | AAGATTATAACTTTGTCTTCAAGGTGGTGCTGATCGGCGAATCAGGTGTGGGGA AGACCAATCTACTCTCCCGATTACGCGCAATGAGTTCAGCCACGA |

| | |
|----------|---|
| CD276 | ACATTTCTTAGGGACACAGTACACTGACCACATCACCACCCTCTTCTTCCAGTGC TGCCTGGACCATCTGGCTGCCTTTTTTCTCCAAAAGATGCAATAT |
| CD86 | CCAGCTCTGCTCCGTATGCCAAGAGGAGACTTTAATTCTCTTACTGCTTCTTTTC ACTTCAGAGCACACTTATGGGCCAAGCCCAGCTTAATGGCTCATG |
| PDCD1LG2 | TGTGGAGCTGTGGCAAGTCCTCATATCAAATACAGAACATGATCTTCCTCCTGCT AATGTTGAGCCTGGAATTGCAGCTTCACCAGATAGCAGCTTTATT |
| PTK2 | ATCCTGTCTCCAGTCTACAGATTTGATAAGGAATGCTTCAAGTGTGCTCTTGTT CAAGCTGGATTATTTAGTGGAAGTGGCAATCGGCCAGAAGAAG |
| IL5RA | CTCCTGCACAGTTGCTCTGTACAAATCCTCCTCCATATTTGCTTAGAGAAAACGT GTTGCCATCCCATCATGAAGGAAGCTGCCTGAGAGTTTTTAACCA |
| KIT | GGTCCTATGGGATTTTTCTTTGGGAGCTGTTCTCTTTAGGAAGCAGCCCCTATCC TGGAATGCCGGTCGATTCTAAGTTCTACAAGATGATCAAGGAAGG |
| CD27 | CCAGATGTGTGAGCCAGGAACATTCCTCGTGAAGGACTGTGACCAGCATAGAAA GGCTGCTCAGTGTGATCCTTGATACCGGGGGTCTCCTTCTCTCCT |
| LILRA4 | AGACTGGGGCAGCAGTTGGGGAAGTGTCTGCTGAGAATATCAAGGGGAAGAAGC ATGGGTCAGGTGCAGGAAGATGTCTGGGTGTCTGTAGAAGATGCTT |
| PPFIA1 | GGAAAGAGATCTTCTGATGGTTCTTTAAGCCACGAGGAAGACCTTGCTAAAGTA ATTGAGCTCCAAGAAATCATAAGTAAGCAGTCAAGGGAACAGAGCC |
| SHARPIN | TCTCAGAGCTCGTTTTCCCGCCAGCCGTGCAACGCTGGGTGATCGGACGGTGCC TGTGTGTGCCTGAGCGCAGCCTTGCCCTTACGGGGTTCGGCAGGA |
| TIGIT | TGGATCTTAGAAGACTTTTATCCTTCCACCATCTCTCTCAGAGGAATGAGCGGGG AGGTTGGATTTACTGGTGACTGATTTTTCTTTTATGGGCCAAGGAA |
| DEGS2 | CGCAACCGCTGGCTGGCCGTGTTGCGCAACCTGCCCGTGGGTGTGCCCTACGCC GCCTCCTTCAAGAAGTACCACGTGGACCACCACCGCTACCTGGGCG |
| KLRD1 | CAATTTTACTGGATTGGACTCTTTACAGTGAGGAGCACACCGCCTGGTTGTGGG AGAATGGCTCTGCACTCTCCAGTATCTATTTCCATCATTTGAAA |
| ZNF34 | TTCAGTGATGGCTCAATCCTTATCCGACATCGTCGGACTCACACCGGAGAGAAGC CATTTGAGTGCAAGGAATGTGGCAAAGGCTTTACACAAAGTTCTA |

| | |
|---------------------|--|
| SNAP ₄₇ | GAAGACGTGGACGACATCAAGGTCCACTCACCTTACGAAATTAGCATCCGCCAG CGGTTTTATTGGAAAGCCAGACATGGCCTATCGTTTGATATCTGCCA |
| MRPL ₁₉ | GGAAGTATTCTTCGTGTTACTACAGCTGACCCATATGCCAGTGGAAAAATCAGCC AGTTTCTGGGGATTTGCATTCAGAGATCAGGAAGAGGACTTGGAG |
| FAHD ₁ | CATAACCTTGGGAAGAAGGAGATATTATCTTGACTGGGACGCCAAAGGGAGTTGG ACCGGTTAAAGAAAACGATGAGATCGAGGCTGGCATAACACGGGCTG |
| KIR _{3DL2} | TGCCACCCACGGAGGGACCTACAGATGCTTCGGCTCTTCCGTGCCCTGCCCTGC GTGTGGTCAAACCTCAAGTGACCCACTGCTTGTCTGTACAGGA |
| ICAM ₂ | ACCTCTCTAGATAAGATTCTGCTGGACGAACAGGCTCAGTGGAAACATTACTTGG TCTCAAACATCTCCCATGACACGGTCTCCAATGCCACTTCACCT |
| ARG ₂ | ATCAGAAAATCAAGCACGTGTGAGAATTTAGGAGACACTGTGCACTGACATGTT TCACAACAGGCATTCCAGAATTATGAGGCATTGAGGGGATAGATGA |
| TBCE | ATAATGAGTGACACTTTGACAGCGGATGTCATTGGTCGAAGAGTTGAAGTTAAT GGAGAACATGCAACAGTACGTTTTGCTGGTGTGTCCCTCCCGTGG |
| AURKA | AGCTCCAGTTGGAGGTCCAAAACGTGTTCTCGTGACTIONCAGCAATTTCCCTTGTCAG AATCCATTACCTGTAAATAGTGGCCAGGCTCAGCGGGTCTTGTGT |
| TRIP ₁₃ | AAGAGACAGAAAACATAATTGCAGCAAATCACTGGGTTCTACCTGCAGCTGAAT TCCATGGGCTTTGGGACAGCTTGGTATACGATGTGGAAGTCAAATC |
| COPA | ATGGTGGTATGATTGTGTTTAAGCTGGAACGGGAACGGCCAGCCTATGCTGTTC ATGGCAATATGCTACACTATGTCAAGGACCGATTCTTACGACAGCT |
| SCAMP ₃ | GCTAGGAATTGTCATGCTGAAACGGATCCACTCCTTATACCGCCGCACAGGTGCC AGCTTTCAGAAGGCCAGCAAGAATTTGCTGCTGGTGTCTTCTCC |
| UCK ₂ | CAGTACATTACGTTTCGTCAAGCCTGCCTTTGAGGAATTCTGCTTGCCAACAAAGA AGTATGCTGATGTGATCATCCCTAGAGGTGCAGATAATCTGGTGG |
| DUSP ₄ | GCACCGTAGCATGCAGATGTCAAGGCAGTTAGGAAGTAAATGGTGTCTTGTAGA TATGTGCAAGGTAGCATGATGAGCAACTTGAGTTTGTGGCCACTGA |
| GAB ₂ | CAGGGAGTCAAGGACCAGCAAACCAAAGTGGATAATGGACTTTTTTCATTCTGT TTTCTTGGCAGGAGAGAAGCAAGGCCACTAAAAGAGGAGATGGTG |

| | |
|----------|---|
| CDCA7 | GACGGTGTGCGACTGGGGTCCTTGTGTATTTAGCCAAATATCATGGCTTTGGGAA TGTGCATGCCTACTTGAAAAGCCTGAAACAGGAATTTGAAATGCA |
| FOXM1 | CAATTCGCCATCAACAGCACTGAGAGGAAGCGCATGACTTTGAAAGACATCTAT ACGTGGATTGAGGACCACCTTCCCTACTTTAAGCACATTGCCAAGC |
| EIF4EBP1 | CTGCGCAATAGCCCAGAAGATAAGCGGGCGGGCGGTGAAGAGTCACAGTTTGAG ATGGACATTTAAAGCACCAGCCATCGTGTGGAGCACTACCAAGGGG |
| MAPK1 | ACTGCCAGAGAACCCTGAGGGAGATAAAAATCTTACTGCGCTTCAGACATGAGA ACATCATTGGAATCAATGACATTATTGAGCACCAACCATCGAGCA |
| PPARG | CAGATCCAGTGGTTGCAGATTACAAGTATGACCTGAAACTTCAAGAGTACCAAA GTGCAATCAAAGTGGAGCCTGCATCTCCACCTTATTATTCTGAGAA |
| CASP10 | TTTCAGGCAATTTCCCTGAGAACCGTTTACTTCCAGAAGATTGGTGGAGCTTGAT CTGAAGGCTGGCCATGAAATCTCAAGGTCAACATTGGTATTCCAG |
| CASP3 | ACTCCACAGCACCTGGTTATTATTCTTGGCGAAATTCAAAGGATGGCTCCTGGTT CATCCAGTCGCTTTGTGCCATGCTGAAACAGTATGCCGACAAGCT |
| ICAM4 | ACAGCTTTGGCCTCCGGTTCCATCGCTGCCCTTGTAGGGATCCTCCTCACTGTGG GCGCTGCGTACCTATGCAAGTGCCTAGCTATGAAGTCCCAGGCGT |
| HLA-B | CCCTGAGATGGGAGCCGTCTTCCAGTCCACCGTCCCCATCGTGGGCATTGTTGC TGGCCTGGCTGTCTAGCAGTTGTGGTCATCGGAGCTGTGGTCCG |
| FAM134B | TTGACCAGTCAGAGCTGGATCAAATTGAGAGTGAATTGGGACTTACACAAGACC AGGAAGCAGAAGCACAGCAAAATAAGAAGTCTTCAGGTTTCCTTTC |
| LAD1 | ACAGCAGCACCCCTTTCCTCTCATTGTCCCTGTTCCCTTTTTGCCTGTGGATCTGTT TGGCCAGGGTCCCTGGGGTCAGGAATATTTGCAAGACTCAGCCA |
| HLA-DOB | ACGGGACAGAAAAGGTGCAGTTTGTGGTCAGATTCATCTTTAACTTGGAGGAGT ATGTACGTTTCGACAGTGATGTGGGGATGTTTGTGGCATTGACCAA |
| PPAPDC1B | CCGGAGGAGATGTGGCTCTACCGGAACCCCTACGTGGAGGCGGAGTATTTCCCC ACCAAGCCGATGTTTGTATTGCATTTCTCTCTCCACTGTCTCTGA |
| TSEN54 | TCAAGCGGTTGTCTTACCAGAGTGGGGATGTCCCTCTGATCTTTGCCCTGGTGG TCATGGTGACATCTCCTTCTACAGCTTCAGGGACTTCACGTTGCC |

| | |
|---------|--|
| IL5 | AGTGAGAATGAGGGCCAAGAAAGAGTCAGGCCTTAATTTTCAGTATAATTTAAC TTCAGAGGGAAAGTAAATATTTTCAGGCATACTGACACTTTGCCAGA |
| ATG16L1 | GTCCTTCCAAAGTCGGTCTGGCCTAACGCATGTCCCAACACCTTGGGTTCATTT GCCCGGTGAACTCACTTTAAGCATTGGATTAACGGAAACTCCCGA |
| PUF60 | ACCTCTCTGCGTGACAGTGGTCCCTCTCCCCGACTTGCACCTGTTTCCTTGTTTC CTCTGGGTTTTATAGTGATACAGTGGTGTCCCCGGGGCCAGGCGC |
| CYBRD1 | CTGGGAGACAATGATTTCACTACTAGCGGGAAGCAGTCCTAAAAGTTTAAAATC CGATAAGGAATATCTGGGACAGGGTTTAGATCATGACTCTACACAG |
| ITGB1 | TGGGTGGTGACAAATTCAACATTTTTTACAGGAAGGAATGCCTACTTCTGCACGA TGTGATGATTTAGAAGCCTTAAAAAAGAAGGGTTGCCCTCCAGAT |
| FBXL6 | GTCAAGGCGGAGAAGAAGCTCCTTGCTTCCCTGGAGTGGCTTATGCCCAATCGG TTTTACAGCTCCAGAGGCTGACCCTCATCCACTGGAAGTCTCAGG |
| PDSS1 | AGCGCGCCATAGCCTTAATTGCAGAAATGATCCACACTGCTAGTCTGGTTCACGA TGACGTTATTGACGATGCAAGTTCTCGAAGAGGAAAACACACAGT |
| BOP1 | ACCGGCAGCGATTCTGGCGTCTCCGACAGCGAGGAGAGTGTGTTCTCAGGCCTG GAAGATTCCGGCAGTGACAGCAGTGAGGATGATGACGAAGGCGACG |
| EPN3 | ACCCTCGTTCTCAGCTCTACCAAGTGGACTTTTTGCGGGGTGTGGCGGCCGGGT CTCGACCACAGCGTGGATCACCGGCTGTTTAGGAAACTGCAGCTG |
| RERG | GCTTTGTGCTGGTCTACGACATTACTGACCGAGGAAGTTTTGAGGAAGTGCTGCC ACTTAAGAACATCCTAGATGAGATCAAAAAGCCCAAGAATGTGAC |
| TACO1 | AGTGCCAAGCAGACATTAGACATATCCTGAATAAGAATGGAGGAGTGATGGCTG TAGGAGCTCGTCACTCTTTTGACAAAAGGGGGTGATTGTGGTTGA |
| APP | TAAAGCATTTGAGCATGTGCGCATGGTGGATCCCAAGAAAGCCGCTCAGATCC GGTCCCAGGTTATGACACACCTCCGTGTGATTTATGAGCGCATGAA |
| C8orf76 | GACACTTTGCTGTTGATAGCTGAGGTTATGGGAGAAGATATCCAGAAAAATA AAAGATGAAGTTCACCCAGAGGTGAAGTGTGTTGGCTCCGTAGCCC |
| IFNA2 | GAAAGTTTAAAGAAGTAAGGAATGAAAAGTGGTTCAACATGGAAATGATTTTCAT TGATTCGTATGCCAGCTCACCTTTTTTATGATCTGCCATTTCAAAGA |

| | |
|---------|--|
| DEDD | TACCTCGATGCATTCTGGCGTGACTACATCAATGGCTCTTTATTAGAGGCACTTA AAGGTGTCTTCATCACAGACTCCCTCAAGCAAGCTGTGGGCCATG |
| NOTCH1 | AGGCAAAGCTGGCTCACCTTCCGCACGCGGATTAATTTGCATCTGAAATAGGAA ACAAGTGAAAGCATATGGGTTAGATGTTGCCATGTGTTTTAGATGG |
| CLEC5A | GGCGTTGGATCAACAACCTCTGTGTTCAATGGCAATGTTACCAATCAGAATCAGAA TTTCAACTGTGCGACCATTGGCCTAACAAAGACATTTGATGCTGC |
| FCGR2C | AAAATTCCTGAGCAAACAAAACCACCTGGCCCTTAGAAATAGCTTTAACTTTGCT TAAACTACAAACACAAGCAAACTTCACGGGGTCATACTACATAC |
| CYC1 | CAGCTCCCGCTACGGACACCTCAGGCAGTGGCCTTGTCGTCGAAGTCTGGCCTTT CCCGAGGCCGGAAAGTGATGCTGTCAGCGCTGGGCATGCTGGCGG |
| LILRB3 | AAAAAAACAAAAACAAAAACAGACGTAAAGGCCGGGTGTGGTACTCAGGAGGC TGAGTGGGGAGGATTCCTTGAACACAAGAAGTTAAGGCTGCTGAGG |
| UFC1 | GTGGAGAACAACAAGAATGCTGACAACGATTGGTTCGACTGGAGTCCAACAAG GAAGGAACTCGGTGGTTTGGAAAATGCTGGTATATCCATGACCTCC |
| CD1B | TGCTCCTTTTGCTATGCCTTGCATTATGGTATATGAGGGCGCCGGTCATATCAGAA TATCCCATGAGCCATCATCATGTCTCCTCTCCCATTTCGCAATAAG |
| TLR8 | TTTAACTGATAGCCTATCTGACTTTACATCTTCCCTTCGGACACTGCTGCTGAGT CATAACAGGATTTCCACCTACCCTCTGGCTTTCTTTCTGAAGTC |
| CLEC10A | TGTCCCTGGGCCTCGGCCTCCTGCTGCTGGTCATCATCTGTGTGGTTGGATTCCA AAATTCCAAATTCAGAGGGACCTGGTGACCCTGAGAACAGATTT |
| KIR3DL3 | GGGTTCCCAGGTCAACTATTCCATGGGTCCCATGACACCTGCCCTTGCAGGGACC TACAGATGCTTTGGTTCTGTCACTCACTTACCCTATGAGTTGTCG |
| KIR2DS1 | CTTCACCCACTGAACCAAGCTCCGAAACCGGTAACCCAGACACCTACATGTTCT GATTGGGACCTCAGTGGTCAAATCCCTTTCACCATCCTCCTCTT |
| KLRC2 | TATGTGAGTCAGCTTATAGGAAGTACCAAGAACAGTCAAACCCATGGAGACAGA AAGTAGAATAGTGGTTGCCAATGTCTCAGGGAGGTTGAAATAGGAG |
| MIF | TCCTACAGCAAGCTGCTGTGCGGCCTGCTGGCCGAGCGCCTGCGCATCAGCCCG GACAGGGTCTACATCAACTATTACGACATGAACGCGGCCAATGTGG |

| | |
|--------------------|---|
| PFKP | GTCAAACCTCTCGGAGAACCGTGCCCGGAAAAAAGGCTGAATATTATTATTGTG GCTGAAGGAGCAATTGATACCCAAAATAAACCCATCACCTCTGAGA |
| HAMP | CTTGCCTCCTGCTCCTCCTCCTCGCCAGCCTGACCAGTGGCTCTGTTTTCCC ACAACAGACGGGACAACCTTGCAGAGCTGCAACCCCAGGACAGAGC |
| C ₄ BPA | CTCTCCCACAATGTGAAATTGTCAAGTGTAAAGCCTCCTCCAGACATCAGGAATGG AAGGCACAGCGGTGAAGAAAATTTCTACGCATACGGCTTTTCTGT |
| CD ₅₃ | CAGTAGTCCTGTGGTGAAGAGACTTGTTTCATCTCCGGAAATGCAAAACCATTTA TAGCATGAAGCCCTACATGATCACTGCAGGATGATCCTCCTCCA |
| COG ₂ | GCCCTCAACCAGCTTTCTGTGCCTTTGGGACAATTACGAGAAGAGGTTCTGAGCC TTAGATCGTCTGTCACTGAAGGAATTCGGGCAGTTGATGAACGAA |
| IL17B | ACCAGGTGCCACTGGACCTGGTGTACGGATGAAACCGTATGCCCGCATGGAGG AGTATGAGAGGAACATCGAGGAGATGGTGGCCAGCTGAGGAACAG |
| IL1RL1 | CCTCTTGAGTGGTTTAAGAATTGTCAGGCTTTCAAGGATCAAGGTACAGGGCG CACAAGTCATTTTTGGTCATTGATAATGTGATGACTGAGGACGCAG |
| SIGIRR | GTCCTGGGGGTCAACGTGACCAGCACTGAAGTCTATGGGGCCTTCACCTGCTCC ATCCAGAACATCAGCTTCTCCTCCTTCACTCTTCAGAGAGCTGGCC |
| MS4A4A | GTTTGAGGCCACCAAAAGATCAACAGACAAATGCTCCAGAAATCTATGCTGACT GTGACACAAGAGCCTCACATGAGAAATTACCAGTATCCAACCTTCGA |
| HAVCR ₂ | TATATGAAGTGGAGGAGCCCAATGAGTATTATTGCTATGTCAGCAGCAGGCAGC AACCCCTACAACCTTTGGGTTGTCGCTTTGCAATGCCATAGATCCA |
| CD ₄₆ | TATACCTCCTCTTGCCACCCATACTATTTGTGATCGGAATCATAATGGCTACCT GTCTCAGATGACGCCTGTTATAGAGAAACATGTCCATATATACGG |
| ILF ₃ | TGTAACAGAAGACAAGTACGAAATACTGCAATCTGTGACGATGCTGCGATTGT GATAAAAAACACAAAAGAGCCTCCATTGTCCCTGACCATCCACCTG |
| EXOSC ₄ | TACAGGCAGATGGTGGGACCTATGCAGCTTGTGTGAATGCAGCCACGCTGGCAG TGCTGGATGCCGGGATACCCATGAGAGACTTTGTGTGTGCGTGCTC |
| ROGDI | CTGGCTCAACGACGCCCTGGTCTACTTACCCTGCTCCCTGCAGCTCTGCCAGCAG CTCAAGGACAAGATCTCCGTGTTCTCCAGCTACTGGAGCTACAGA |

| | |
|----------------------------------|--|
| KIR ₂ DL ₃ | CTCCGAAACCGGTAACCCAGACACCTGCATGTTCTGATTGGGACCTCAGTGGTC ATCATCCTCTTCATCCTCCTCCTCTTCTTTCTCCTTCATCGCTGG |
| ATF ₃ | CTGGGTGGTACCCAGGCTTTAGCATTATTGGATGTCAATAGCATTGTTTTTGTCA TGTAGCTGTTTTAAGAAATCTGGCCCAGGGTGTTCGAGCTGTGA |
| ICAM ₁ | AAATACTGAACTTGCTGCCTATTGGGTATGCTGAGGCCCCACAGACTTACAGAA GAAGTGGCCCTCCATAGACATGTGTAGCATCAAAACACAAAGGCC |
| POU ₂ F ₂ | GACGCAAGAAGAGGACCAGCATCGAGACAAACGTCCGCTTCGCCTTAGAGAAGA GTTTTCTAGCGAACCAGAAGCCTACCTCAGAGGAGATCCTGCTGAT |
| TCF ₇ | CCCCTCAGGGAAGCAGGAGCTGCAGCCCTTCGACCGCAACCTGAAGACACAAGC AGAGTCCAAGGCAGAGAAGGAGGCCAAGAAGCCAACCATCAAGAAG |
| C _{10orf35} | GGGATGGCCCCTGGCTTGGCCTGCGAAGGTGAACCTGCCAGATTTATCAGTAG AGGCTGGACTCCCTCTGTGTCCTGCCCATGGTTGCAGCAGCCATGG |
| SLAMF ₆ | TTTTTCCAGGGCAACTGCCCTTGACAATGTCGTGTAAGTTGCTGAAAGGCCTCAG AGGAATTCGGGAATGACACGTCTTCTGATCCCATGAGACAGAACA |
| GUSB | CCGATTTTCATGACTGAACAGTCACCGACGAGAGTGCTGGGGAATAAAAAGGGGA TCTTCACTCGGCAGAGACAACCAAAAAGTGCAGCGTTCCTTTTTGCG |
| CD _{1E} | ATATCTGAATTATTAGGGCAGGTGTCCTGCCAAGGAATCCCTCCTTTAACAGAGC TTCAATGCTGCTCCTGTTCTCCTCTTCGAGGGTCTCTGCTGTCC |
| CD ₃ EAP | CAGCCTCAGAGTCCCCTCGTTTCTCCTTGGAGGCGCTGACGGGTCCAGATACGG AGCTGTGGCTTATTCAGGCCCTGCAGACTTTGCCCCAGAATGCTT |
| PYCR _L | CTGAGGGGCCCAAGAGATGGCGTCTTGGTCATTTGCCCGCATGGTTGGGCAGTT GGTTGAGGCCATGAACAGAACTTACGGTAACAGGCACGGCTGGCCC |
| IRGM | CCTGGAGAACTACCTGATGGAAATGCAGTTCAACCGGTATGACTTCATCATGGTT GCATCTGCACAATTCAGCATGAATCATGTGATGCTTGCCAAAACC |
| MELK | TATGAGAGGAAAATATGATGTTCCCAAGTGGCTCTCTCCAGTAGCATTCTGCTT CTTCAACAAATGCTGCAGGTGGACCCAAAGAAACGGATTTCTATG |
| LRRC ₁₄ | TACAGATCACTTCTCCGTCCGGTCTCTGAGAAAGCACCTGCTCCTTAAGTCTTCCT GCAACAAGTGCCACTGTTTTTAGGAACCTGGGCGTCCACATAGAC |

| | |
|----------------------------------|---|
| MRC ₁ | CTATGGAACCACAGACAATCTGTGCTCCAGAGGTTATGAAGCCATGTATACGCTA CTAGGCAATGCCAATGGAGCAACCTGTGCATTCCCGTTCAAGTTT |
| BST ₁ | GAAAAGGGCATCCATCCAGTATTCCAAGGATAGTTCTGGGGTGATCCACGTCAT GCTGAATGGTTCAGAGCCAACAGGAGCCTATCCCATCAAAGGTTTT |
| ZNF ₇₀₃ | GGCAGCCCCGGGTCGCTGTCCTTGCGGAATCCACACACTTTGGGCCTAAGCCGG TACCACCCCTATGGCAAGAGCCACTTATCCACAGCGGGGGGCTGG |
| GZMK | CTGTAAAGGTGTCTTCCACGCTATAGTCTCTGGAGGTCATGAATGTGGTGTGGCC ACAAAGCCTGGAATCTACACCCTGTAAACCAAGAAATACCAGACT |
| IL ₁₂ RB ₁ | GTGACCCTGCAGCTCTACAACCTCAGTTAAATATGAGCCTCCTCTGGGAGACATCA AGGTGTCCAAGTTGGCCGGCAGCTGCGTATGGAGTGGGAGACCC |
| CD ₄ | TGGCAGGCGGAGAGGGCTTCTCCTCCAAGTCTTGGATCACCTTTGACCTGAAG AACAAGGAAGTGTCTGTAAAACGGGTTACCCAGGACCCTAAGCTCC |
| NOSTRIN | GACCTTTTGAAGCGAACTCCTACAACTGTCATCAATGTTAGCAGAACTTGAGC AAAGACCTCAACCCAGCCATCCTTGTAGTAATTCCATCTTCAGGT |
| PSMB ₁₀ | ACCATCGCGGGCCTGGTGTTCGAAGACGGGGTCATTCTGGGCGCCGATACGCGA GCCACTAACGATTCCGGTCGTGGCGGACAAGAGCTGCGAGAAGATCC |
| SH ₂ D _{1A} | GCTGTATCACGGTTACATTTATACATACCGAGTGTCCCAGACAGAAACAGGTTCT TGGAGTGCTGAGACAGCACCTGGGGTACATAAAAAGATATTTCCGG |
| HLA-DMA | TTATTTGACAAAGAGTTCTGCGAGTGGATGATCCAGCAAATAGGGCCAAAACCT GATGGGAAAATCCCGGTGTCCAGAGGGTTTCTATCGCTGAAGTGT |
| GATA ₃ | GTGCATGACTCACTGGAGGACTTCCCCAAGAACAGCTCGTTTAACCCGGCCGCC CTCTCCAGACACATGTCTCCCTGAGCCACATCTCGCCCTTCAGCC |
| LGALS ₃ | CACGGTGAAGCCCAATGCAAACAGAATTGCTTTAGATTTCCAAAGAGGGAATGA TGTTGCCTTCCACTTTAACCACGCTTCAATGAGAACAACAGGAGA |
| SELL | CTAACTCCAGTGAAGTAATGGGGTCTGCTCAAGTTGAAAGAGTCCTATTTGCAC TGTAGCCTCGCCGTCTGTGAATTGGACCATCCTATTTAACTGGCT |
| PSMB ₇ | GTTACATTGGTGCAGCCCTAGTTTTAGGGGGAGTAGATGTTACTGGACCTCACCT CTACAGCATCTATCCTCATGGATCAACTGATAAGTTGCCTTATGT |

| | |
|---------------------|--|
| GRINA | TGCATCTTCATCCGGAACCGCATCCTGGAGATCGTGTACGCCTCACTGGGCGCTC TGCTCTTCACCTGCTTCCTCGCAGTGGACACCCAGCTGCTGCTGG |
| DAP ₃ | CACAGATGCAGTTGGAATTGTGCTGAAAGAGCTAAAGAGGCAAAGTTCTTTGGG TATGTTTCACCTCCTAGTGGCCGTGGATGGAATCAATGCTCTTTGG |
| ETS ₁ | GGTTTTACAGCATTAACTGCCTAACCTTCATGGTGAGAAATACACCATCTCTCT TCTAGTCATGCTGTGCATGCCGCTTACTCTGTTGGGGTCTATATA |
| BAX | TTTTTCCGAGTGGCAGCTGACATGTTTTCTGACGGCAACTTCAACTGGGGCCGGG TTGTGCCCCTTTTCTACTTTGCCAGCAAAGTGGTGCTCAAGGCC |
| IRF ₃ | TCATGGCCCCAGGACCAGCCGTGGACCAAGAGGCTCGTGATGGTCAAGGTTGTG CCCACGTGCCTCAGGGCCTTGGTAGAAATGGCCCCGGTAGGGGGTG |
| CASP8 | AGATGGACTTCAGCAGAAATCTTTATGATATTGGGGAACAAGTGGACAGTGAAG ATCTGGCCTCCCTCAAGTTCCTGAGCCTGGACTACATTCCGCAAAG |
| SELE | AATTCACCTACAAGTCCTCTTGTGCCTTCAGCTGTGAGGAGGGATTTGAATTACA TGGATCAACTCAACTTGAGTGCACATCTCAGGGACAATGGACAGA |
| TIRAP | ACCCACACATGCGAGTGACAGTGGCAGTAGTCGCTGGAGCAAAGACTATGACGT CTGCGTGTGCCACAGTGAGGAAGACCTGGTGGCCGCCAGGACCTG |
| C ₁ QBP | ATTATACACTCAACACAGATTCCTTGGACTGGGCCTTATATGACCACCTAATGGA TTTCCTTGCCGACCGAGGGGTGGACAACACTTTTGCAGATGAGCT |
| C8orf8 ₂ | CGTGGACCACCAGGGCCAGCTTTTCCTGGATGATTCCAAAATGAAGAATTTTCATC ACCTGCTTCAAAGACCCGCAGTTCCTGGTCACCTTCTTCTCCCGC |
| GSDMB | GATGGATGCTGGAGGGGATATGATTGCCGTTAGAAGCCTTGTTGATGCTGATAG ATCCGCTGCTTCCATCTGGTGGGGGAGAAGAGAACTTTCTTTGGA |
| ATG ₇ | TATGATCCCTGTAACCTTAGCCCAGTACCCTGGATGGCCTTTGAGGAATTTTTTGG TCCTAGCAGCCCACAGATGGAGTAGCAGTTTCCAGTCTGTTGAAG |
| STARD ₃ | CGCCGCACCTTCTGTCTCTTCGTCACCTTCGACCTGCTCTTCATCTCCCTGCTCTG GATCATCGAACTGAATACCAACACAGGCATCCGTAAGAACTTGG |
| NSMCE ₂ | GGACGCCATTGTTTCGCATGATTGAGTCCAGGCAAAAAGCGGAAGAAAAAGGCCTA TTGCCCTCAAATTGGCTGTAGCCACACGGATATAAGAAAGTCAGAT |

| | |
|-------------------|---|
| CHTOP | ACAGTCAGCGCCGAAAAGTTGTGCTAAAAAGCACCACCAAGATGTCTCTAAATGA GCGCTTTACTAATATGCTGAAGAACAAACAGCCGACGCCAGTGAAT |
| IFNA8 | TGGTACAACACGGAAATGATTCTTATAGACTAATACAGCAGCTCACACTTCGACA AGTTGTGCTCTTTCAAAGACCCTTGTTTCTGCCAAAACCATGCTA |
| IDO ₂ | GTTCTGCCTGGGATCATCCAGGAAGGATCTCAGCCCTATTCATGTTTCTGCTCTA CAGAGCACTATATTCTCCTTGTTGAGAGCTGTTGGCTTCACAAAG |
| CD44 | ACACCATGGACAAGTTTTGGTGGCACGCAGCCTGGGGACTCTGCCTCGTGCCGC TGAGCCTGGCGCAGATCGATTTGAATATAACCTGCCGCTTTCAGG |
| HLA-A | CCTTGTGTGGGACTGAGAGGCAAGAGTTGTTTCTGCCCTTCCCTTTGTGACTTGA AGAACCCTGACTTTGTTTCTGCAAAGGCACCTGCATGTGTCTGTG |
| TAP ₂ | GGCTTCCTTTAAATGCCAATGTGCTCTTGCGAAGCCTGGTGAAAGTGGTGGGGCT GTATGGCTTCATGCTCAGCATATCGCCTCGACTCACCTCCTTTC |
| IL19 | CCACAGACATGCACCATATAGAAGAGAGTTTCCAAGAAATCAAAAGAGCCATCC AAGCTAAGGACACCTTCCCAAATGTCACTATCCTGTCCACATTGGA |
| CD24 | ATAGACACTCCCCGAAGTCTTTTTGTTTCGCATGGTCCACACTGATGCTTAGATGT TCCAGTAATCTAATATGGCCACAGTAGTCTTGATGACCAAAGTCC |
| NARS ₂ | ATGCCAAGGATTTCCCATCAAATATAAAGAGAGGCATCCTCTGGAGTATCTGCG ACAATATCCTCACTTTAGGTGTAGGACTAACGTTCTGGGTTCTAT |
| BCL ₂ | AGTTCGGTGGGGTCATGTGTGTGGAGAGCGTCAACCGGGAGATGTGCCCCCTGG TGGACAACATCGCCCTGTGGATGACTGAGTACCTGAACCGGCACCT |
| C ₅ | ATTCGTTGAATGACGACTTGAAGCCAGCCAAAAGAGAACTGTCTTAACTTTTCAT AGATCCTGAAGGATCAGAAGTTGACATGGTAGAAGAAATTGATCA |
| CARD ₉ | GAGGCCCGACGCCTCCGGTGCATGGAGGAGAAGGAGATGTTTCGAGCTGCAGTGC CTGGCACTACGTAAGGACTCCAAGATGTACAAGGACCGCATCGAGG |
| ADA | TCCAAGAAGACCATGATCTCAATAGTCAGTACTGATGCTCCTGAACCCTATGTG TCCATTTCTGCACACACGTATACCTCGGCATGGCCGCGTCACTTC |
| RAG ₁ | CAGTCTACATTTGTA CTCTTTGTGATGCCACCCGTCTGGAAGCCTCTCAAAATCT TGTCTTCCACTCTATAACCAGAAGCCATGCTGAGAACCTGGAACG |

| | |
|--------|---|
| PSMD7 | CAACCTCGAAAACATTTGAACACGTGACCAGTGAAATTGGAGCAGAGGAAGCTG AGGAAGTTGGAGTTGAACACTTGTTACGAGATATCAAAGACACGAC |
| RSPH1 | AGGAGTTCCGCTATGACATGGATGAGGGAAACATTAATTCTGAAGAAGAAGAAA CTAGACAGTCAGACCTCCAGGACTAAGATGAAGTGAGCCGAGAGGA |
| CAPN8 | AGATGAGGACAGCCCTCAGGAAGGCAGGTTTCACCCTCAACAGCCAGGTGCAGC AGACCATTGCCCTGCGGTATGCGTGCAGCAAGCTCGGCATCAACTT |
| ATG10 | TCAACTATATCACATCATGGCTGAGCATTGTAGGGCCAGTTGTTGGGCTGAATCT ACCTCTGAGTTATGCCAAAGCAACGTCTCAGGATGAACGAAATGT |
| FBXO28 | AGCTCTTCTCCAAGCAAATCCTTCAAGACAAGAGGTTACCAAACCTCCAGCAGCA GGTTAAAACAAATGGTGCTGGCGTGACTGTTCTCAGGCGTGAAAT |
| NOL11 | TTGTCAGCAGAAGTATATAGGATACTTTCAGTGCAAGGGACAGAACCCTTGGTG CTCTTCAAGGAAGGTGCTGTTCTGTTTAGAGGCCTTGCTTGCAG |
| SMG8 | ATGAGATCTGCGTTGTGGGAATCTTCGGCAAGACGGCTCTACGCCTGAATTCCG AGAAGTTCTCTTGTGAATACGGTGTGCGACCGACAGGTCTTTCC |
| PRR11 | CACAGAGTTTAGATGAAAAGAGGAAGCTTATAACCATCGCCGAAAGCACGGAATC CACTAGTTACCGTCTCTGACTTGCAGCATGTTACCCTGAAACCTAA |
| INTS2 | CAGAAGCCACAAATCAGCCAGTCACAGAACAGGAGATACTCAATATTTTCCAAG GAGTCATTGGGGGTGACAACATCCGCCTTAATCAGCGTTTCAGTAT |
| HEATR6 | CTTTTATTCCTGATACGCCTGAACTTGGCAGCCCACAGTCAGTGTCTTGTATGAC TCTTACATTGAAAGACCCTTCTCCAAAGACACGTGCCTGTGCTCT |
| CENPN | TGAACTGACAACAATCCTGAAGGCCTGGGATTTTTTGTCTGAAAATCAACTGCAG ACTGTAAATTTCCGACAGAGAAAGGAATCTGTAGTTCAGCACTTG |
| RFWD2 | CGGCTGTGCAGATCACTGTGTCCACTACTATGATCTTCGTAACACTAAACAGCCA ATCATGGTATTCAAAGGACACCGTAAAGCAGTCTCTTATGCAAAG |
| DGUOK | CATTCCAGACATTTTCCTTTTTGAGCCGCCTGAAAGTACAGCTGGAGCCCTTCCC TGAGAAACTCTTACAGGCCAGGAAGCCAGTACAGATCTTTGAGAG |
| VCAM1 | CAGACTTCCCTGAATGTATTGAACTTGGAAAGAAATGCCCATCTATGTCCCTTGC TGTGAGCAAGAAGTCAAAGTAAAACCTTGCTGCCTGAAGAACAGTA |

| | |
|----------------------|---|
| FCRL ₂ | ACCTCATGACAGCTGGAGTTCTCTGGGGACTGTTTGGTGTCTTGGTTTCACTGG TGTTGCTTTGCTGTTGTATGCCTTGTTCACAAGATATCAGGAGA |
| CD ₃ G | GAACTAAATGCAGCCACCATATCTGGCTTTCTCTTTGCTGAAATCGTCAGCATTT TCGTCCTTGCTGTTGGGGTCTACTTCATTGCTGGACAGGATGGAG |
| CSF ₂ RB | GCCCAGGAGATGTGTCATTCCCTGCCAGAGTTTTGTCGTCCTGACTGACGTTGACTAC TTCTCATTCCAACCAGACAGGCCTCTGGGCACCCGGCTCACCGTC |
| IL6ST | TTCAATATAGGACCAAAGATGCCTCAACTTGGAGCCAGATTCTCCTGAAGACAC AGCATCCACCCGATCTTCATTCACTGTCCAAGACCTTAAACCTTT |
| SELPLG | CCTGCTTGCCCGGGACCGGAGACAGGCCACCGAATATGAGTACCTAGATTATGA TTTCTGCCAGAAACGGAGCCTCCAGAAATGCTGAGGAACAGCACT |
| LTBR | GAACCCGGGGAGCAGAGCCAGGTGGCCACGGTACCAATGGCATTTCATGTCACC GGCGGGTCTATGACTATCACTGGCAACATCTACATCTACAATGGAC |
| ICT ₁ | CGGAATCTGGCAGATTGCCTGCAGAAAATTCGAGACATGATCACTGAGGCCAGC CAGACACCGAAGGAGCCAACAAAAGAAGATGTAAACTTCATAGAA |
| LAMP ₃ | CAGCCATCGTCAGTCAAGACTGGAATTTATCAGGTTCTAAACGGAAGCAGACTCT GTATAAAGCAGAGATGGGGATACAGCTGATTGTTCAAGACAAGG |
| MRPS ₇ | ATTCAAGGATCCCTTGATTGACAAGGAATATTATCGCAAGCCAGTGGAGGAGCT AACTGAGGAGGAGAAATATGTTCCGGGAGCTCAAGAAGACTCAGCTC |
| CD ₉₉ | CCCAACCACCCTAGTTCTCCGGTAGCTTTTCAGATGCTGACCTTGCGGATGGCG TTTCAGGTGGAGAAGGAAAAGGAGGCAGTGATGGTGGAGGCAGCC |
| RCOR ₃ | ACAACATCGCCATCATTCTCAGCGTTCTAAGTGCCGTCCACCTAAGGGCATGTAT TTAACCCAGGAAGATGTGGTAGCAGTTTCTGTAGTCCCAATGCA |
| HLA-DQB ₁ | TGGTGAGTGCTGTGTAATAAGCATGGTAGAATTGTTTGGAGACATATATAGTG ATCCTTGGTCACTGGTGTTCAAACATTCTGGAAAGTCACATCGAT |
| SYK | TCCTACGCCCTGTGCCTGCTGCACGAAGGGAAGGTGCTGCACTATCGCATCGAC AAAGACAAGACAGGGAAGCTCTCCATCCCCGAGGGAAGAAGTTTCG |
| MAPK ₁₄ | GGTGACCCATCTCATGGGGGCAGATCTGAACAACATTGTGAAATGTCAGAAGCT TACAGATGACCATGTTCACTTCCTTATCTACCAAATTCTCCGAGGT |

| | |
|------------------------------------|---|
| ATM | GAAACACTCCCAGCTTCTCAAGGACAGTGATTTTAGTTTTTCAGGAGCCTATCATG GCTCTACGCACAGTCATTTTGGAGATCCTGATGGAAAAGGAAATG |
| ARHGAP ₃₉ | CTGCCGGAGTCGAGGATTCCAGGGTCGAACACTCGGTTGGAGTGGGTGGAGATC ATCGAACC GCGCACCCGCGAGCGCATGTACGCCAACCTGGTCACCG |
| HDC | ACCTTGGCCTGGTGGTTTTTCGTCTAAAGGGTCCTAATTGTCTCACAGAAAATGT GTTAAAGGAAATAGCTAAAGCTGGCCGTCTCTTCCTCATCCCGGC |
| RAB ₁₁ FIP ₁ | GCTGCTGGGAAAAAATTCAGTTTATAGCATTCCCTGCACCTCCCAAAGTAGATAAC CTGGAGGTCATTCAAGTTAACAACCTGTCCCTGAGGACTCAGTTTTG |
| DDHD ₂ | TCACAACAGGAGCAGTTGTCCCAGTCAGATCCATCTCCGTCACCAAACCTCATGTA GTTCCTTTGAGCTAATAGACATGGATGCTGGCAGCTTGTATGAAC |
| PROSC | TTGGCGAGAACTACGTTTCAGGAAGCTAGAAAAAGCATCAAATCCCAAATTC TGTCTTTGTGTCCTGAGATCAAATGGCACTTCATTGGCCACCTACA |
| AAMDC | TGGTGTGCAGCCTGCAGATGTGAAGGAAGTTGTTGAGAAGGGTGTACAGACTCT TGTGATTGGCCGAGGGATGAGTGAGGCCTTGAAGGTGCCTTCATCA |
| IL ₁₅ | AATCTATGCATATTGATGCTACTTTATATACGGAAAGTGATGTTACCCCAGTTG CAAAGTAACAGCAATGAAGTGCTTTCTCTTGGAGTTACAAGTTAT |
| TRAF ₃ | GTTGTGCAGAGCAGTTAATGCTGGGACATCTGCTGGTGCATTTAAAAAATGATTG CCATTTTGAAGAACTTCCATGTGTGCGTCCTGACTGCAAAGAAAA |
| B ₄ GALT ₃ | CTTCCGAAGTCTCAGTGCCCTATTTGGCCGAGATCAGGGACCGACATTTGACTAT TCTCACCCCTCGTGATGTCTACAGTAACCTCAGTCACCTGCCTGGG |
| USP ₂₁ | GCACCCTGACTCCTGCGAAGCTGTGAATCCTACTCGATTCCGAGCTGTCTTCCAG AAATATGTTCCCTCCTTCTCTGGATACAGCCAGCAGGATGCCCAA |
| PIGM | CATCCCAGAGAGCCATGGGTTTTCTAGACCAGAGAATTTAGAGGGAGATTGTGG AACTGAGGCTTAGGTGGTCAGATCGTTCCCTTATCACTGTAATAT |
| KLHDC ₉ | CTTTGGAATGATCAGCTTTACCTGGTTGGGGTTTTGGTGAGGATGGCAGGACA GCCAGTCCACAGTTTTGCATCCTGGACTTTATCTAAATAGTGCCAA |
| KCTD ₃ | ACTCAGCCAGGTTCTACTCCTTTAGCGTCATTCAAGATACTATCCCTGGAGGAGA CAGAAAGTCATGGTAGCTATTCCTCTGGAAATGACATAGGACCTT |

| | |
|----------|---|
| CD164 | GATTAATTCGAAAATAGGCAGAATTCCATTCCTCCCAAGGTGGCAAAAATTAGCT ATACTGATGTAATTGTCATTTACCTGGGTATGAATTCCCTGACAC |
| VPS72 | TGCGCCAGAAAATTGTCATTAAATGAAGAGATGTCTAGTCCTCAGAACTTCTTT CCTGCCCTGATTGGGGCTCTTGCTGTTCCGTTTCTTCTCCCTGCT |
| FAM20B | CACCCGTGTATTCTTTGACCAGTAAATGACCACACCTCAATGATGGTAAAACAGC ATCATCAGTAAGCTATCTTATATGCCTCATCCTGTGAGTTTGAGC |
| PTPN11 | TGTCAAATACTGGCCTGATGAGTATGCTCTAAAAGAATATGGCGTCATGCGTGTT AGGAACGTCAAAGAAAGCGCCGCTCATGACTATACGCTAAGAGAA |
| RSF1 | GCAAAAAATGTGGCCTTCCAAACCATCCTGAGCTAATTCTTCTGTGTGACTCTTG CGATAGTGGATAACCATACTGCCTGCCTTCGCCCTCCTCTGATGAT |
| RELA | GAAGCATTAACCTTCTCTGGAAAGGGGGGAGCTGGGGAAACTCAAACCTTTTCCCC TGTCCTGATGGTCAGCTCCCTTCTCTGTAGGGAACCTCTGGGGTCCC |
| IL24 | CGCAAGAAAATGAGATGTTTTCCATCAGAGACAGTGCACACAGGCGTTTTCTGC TATTCCGGAGAGCATTCAAACAGTTGGACGTAGAAGCAGCTCTGAC |
| TAPBP | GGTCACCCTGGAGGTAGCAGGTCTTTCAGGGCCCTCCCTTGAGGACAGCGTAGG CCTTTTCTGTCTGCCTTTCTTCTGCTTGGGCTCTTCAAGGCACTG |
| CCL2 | CATTCCCAAGGGCTCGCTCAGCCAGATGCAATCAATGCCCCAGTCACCTGCTGT TATAACTTCACCAATAGGAAGATCTCAGTGCAGAGGCTCGCGAGC |
| C10orf54 | CTGGGACACTTCTGAGTATGAAGCGGGATGCTATTA AAAACTACATGGGGAAAC AGGTGCAAACCCTGGAGATGGATTGTAAGAGCCAGTTTAAATCTGC |
| CD28 | GCTGCTCCTGTACCTTGGAGGTCCATTACATGGGAAAGTATTTTGGAAATGTGTC TTTTGAAGAGAGCATCAGAGTTCTTAAGGACTGGGTAAGGCCTG |
| CCL3 | CAGTTCTCTGCATCACTTGTGCTGACACGCCGACCGCCTGCTGCTTCAGCTACA CCTCCCGGCAGATTCCACAGAATTTATAGCTGACTACTTTGAGA |
| RUNDC1 | AGACAAGCTCCTTGTTTCACTAGGGATAGATGTGCTAGTCTTCTAGCATAGGAGC AAGGAATCAGAGGTGGGGTTAAAGGCATTTTCCAGACCCTGCTC |
| UBE2S | CGGCTGAGCTGGGCATCCGACACGTA CTGCTGACCATCAAGTGCCTGCTGATCC ACCCTAACCCCGAGTCTGCACTCAACGAGGAGGCGGGCCGCTGCT |

| | |
|---------------------|--|
| KLRF ₁ | TGCAATTA AATGCCAAAATCTCTTCTCCCTTCTCCCTCCATCATCGACTGGTC TAGCCTCAGAGTAACCCCTGTTAACAACTAAAATGTACTTCA |
| TCL _{1A} | TTCACCTGAGCAGGACCTCTGAAACCCTGGACCAGTGGTCTCACATGGTGCTACG CCTGCATGTAAACACGCCTGCAAACGCTGCCTGCCGGTAAACACG |
| LSM ₁ | GAAGCAGAGAAGTTGAAAGTGCAGGCCCTGAAGGACCGAGGTCTTTCCATTCT CGAGCAGATACTCTTGATGAGTACTAATCTTTTGCCCAGAGGCTGT |
| MS _{4A6A} | GAACAGACTTGCCTAACACAGGAACTTGTATGTCTCGAAGTGGCAATTCACA CATAAGGCTCCATGACTCCTGAACTCTCACAAATATTAGTTGGCTC |
| TCEB ₁ | AAGGTTTCGCTACACTAACAGCTCCACCGAGATTCTGAATTCCAATTGCACCTG AAATTGCACTGGAAGTCTGATGGCTGCGAACTTCTTAGATTGTT |
| LILR ₅ | CACCCTCTCAGCCCTGCCAGTCTGTGGTGACCTCAGGAGAGAACGTGACCCTC CAGTGTGGCTCACGGCTGAGATTGACAGGTTTATTCTGACTGAG |
| MLST ₈ | GGCTATCCACATCTGGGACTTGAAAACAGACCACAACGAGCAGCTGATCCCTGA GCCCAGGTCTCCATCACGTCCGCCACATCGATCCCGACGCCAGC |
| SQLE | TAAGCTTAAAGGGGAACCATTTGTGAATGAATATTTGGAAGTTACCAAGTCCTAA GAGACTTTTGGAAAGAGGATATATATAGCATAGTACCATAACCACTT |
| TLR ₅ | TGCCAATATCCAGGATGCCATCTGGAACAGTAGAAAGATCGTTTGTCTTGTGAGC AGACACTTCCCTAGAGATGGCTGGTGCCTTGAAGCCTTCAGTTAT |
| SLAMF ₇ | ACCAACACAGAGCTCACCATCTCTTATACTTAAGTGAAAACATGGGGAAGGGG AAAGGGGAATGGCTGCTTTTGATATGTTCCCTGACACATATCTTGA |
| TMCO ₁ | TCATTTACATAAGTATTTTCTGTGGGACCGACTCTCAAGGCACTGTGTATGCCCT GCAAGTTGGCTGTCTATGAGCATTTAGAGATTTAGAAGAAAATT |
| TMEM ₁₀₁ | ACAGGGGCTGAGCACTACAGAAGTCACATGGGTTCTCAGGGTATGCCAGGGGCA GAAACAGTACCGGCTCTCTGTCACTCACCTTGAGAGTAGAGCAGAC |
| IFN _{A17} | TGAGATGATCCAGCAGACCTTCAATCTCTTCAGCACAGAGGACTCATCTGCTGCT TGGGAACAGAGCCTCTAGAAAAATTTTCCACTGAACTTTACCAG |
| NFATC ₁ | CGAATTCTCTGGTGGTTGAGATCCCGCCATTTGGAATCAGAGGATAACCAGCCC CGTTCACGTCAGTTTCTACGTCTGCAACGGGAAGAGAAAGCGAAG |

| | |
|----------------------|---|
| FAM8 ₃ H | GTGGGGGGATTACAGGCGTGAGCCTTGCACCCCGCTAAGTCCCCTATCCTCTTGC AAGGGTCTCGCCTCTGTGCCTCAATTCCTCATTCTCTGGGCCCTT |
| CH ₂₅ H | ACGCAGTATATGAGCGTCTGGGAACTGTTTTCTTTGGGCTTCTTCGACATGATGA ACGTCACACTGCTCGGGTGCCACCCGCTCACCACCCTGACCTTCC |
| COMMD ₅ | GTGGGGGCTGCAACTCCATACCTGCATCATCCTGGTGATAGTCACAGTGGCCGA GTGAGTTTCTTGGGGGCCAGCTTCTCCAGAGGTGGCAGCAATGG |
| ZNF ₇₀₇ | TCAGAGCCTGGGTCATGCACCTGGAGTTGGGAGATCAAGTTGGGTCTCAGGGCA GTGAGGTGGCCATATCCACCACATCGCATTTCGTGGGGGAAGAGGT |
| TRAPPC ₉ | ACACAAAATGAAAAGCTTGCTGGGTCAGAACGTGTCAACCAAAAGTCCTTTCATC TATTCACCAATTATCGCACACAACCGTGGAGAAGAGCGGAACAAG |
| GNPAT | ATCCATGTGTACTTTGGAGATCCTGTGTCACTTCGATCTTTGGCAGCTGGGAGGA TGAGTCGGAGCTCATATAACTTGGTTCCAAGATACATTCCTCAGA |
| TMEM ₉ | TCTTGGGAATGTTGTTACCCCTGGAAGATAAAGCTGGGTCTTCAGGAACTCAGTG TCTGGGAGGAAAGCATGGCCCAGCATTTCAGCATGTGTTCTTTCT |
| HEATR ₁ | GAGGGTGCTCTTGGCTTTCTATGCTTCTACCATAGTGTGCGGCGCTGGTAGCTGCA GAGGACGTATCAGACAATATCATCGCCAAACTATTTCCCTATATC |
| ARV ₁ | AGTGACCCTAAACATCAACCGTAAGCTCTCCTTCTTGGCCGTGTTGAGTGGCTTA CTGCTGGAAAGCATCATGGTCTACTTCTTCCAGAGTATGGAATGG |
| UTP ₂₃ | GAAAGAAAGTATCAAACATCTCAAAGAGGAACAGGGTTTAGTGAAAAACACTGA ACAGAGTAGAAGAAAAAAGCGCAAGAAAATAAGTGGTCCCAATCCT |
| TIGD ₅ | CCCCTGGGGTGGCCACCGCATGGGTACAGGGGGTTCCAGGAATCCAAATCCAG CATGGCTTGGAGGAGCTCTGTTGGTGAGAGGTCGCCCTGCCTCACT |
| INTS ₄ | TGCTTTGCAATACTTGCTCCAGTTTGCCAGGAAGCCTGTGCGAGGCGGAAAGCGT AGAGGGAGTAGTCAGGATTCTCTTGAACATTATTACAAGGAGAAT |
| C1orf ₁₃₁ | CGGTTTGGTATCACGGGTATGGAAAAGGAAAGGAGAGAATCCTGGAACAGGAA CGTGCCATTATGCTGGGCGCTAAGCCTCCTAAAAAGAGTTATGTGA |
| IFNAR ₁ | TGGTGTCTATAGTCCAGTACATTGTATAAAGACCACAGTTGAAAATGAACTACCT CCACCAGAAAATATAGAAGTCAGTGTCCAAAATCAGAACTATGTT |

| | |
|----------------------|---|
| RRNAD ₁ | GTCAGGCCCAAGAAGCAGCATGAGATCCGGAGGCTGGGAGAGTTGGTGAAGAA GCTGAGTGATTTACAGGCTGCACCCAGGTTGTAGACGTGGGCTCAG |
| TNFSF8 | CCCTCAAAGGAGGAAAATTGCTCAGAAGACCTCTTATGTATCCTGAAAAGGGCTC CATTCAAGAAGTCATGGGCCTACCTCCAAGTGGCAAAGCATCTAAA |
| CD ₄₀ | GTGCCAGCCAGGACAGAACTGGTGAGTGA CTGCACAGAGTTCACTGAAACGGA ATGCCTTCCTTGCGGTGAAAGCGAATTCCTAGACACCTGGAACAGA |
| LRRC ₄₈ | GAAGCACTTGTCGAGTTTAAGTGCCATTCGAGAGGAGTTGGA ACTGCCAACAT TGAGAAGATGATCCTAGAATGCAGTGCTGACATCAGTGAGTTGTTC |
| C ₂ | CCAACCCAGGCATTTCACTGGGCGCAGTGCGGACAGGCTTCCGCTTTGGTCATG GGGACAAGGTCCGCTATCGCTGCTCCTCGAATCTTGTGCTCACGGG |
| TMEM _{132A} | GTACGCCCTGCTGGGAGTCTTCTGCGTGGCCATCTTCATCTTCTTGGTCAATGGT GTGGTCTTCGTCCTGCGCTATCAGCGCAAAGAACCTCCCGACAGT |
| ILiRAP | AAAGAACCAGGAGAGGAGCTACTCATTCCCTGTACGGTCTATTTTAGTTTTCTGA TGGATTCTCGCAATGAGGTTTGGTGGACCATTGATGGAAAAAAC |
| CD ₁₆₀ | AACAAAGACAACACCTTGAGTTCAGCCATAATGAAGGCACTCTCAGTTCAGGCTT CCTACAAGAAAAGGTCTGGGTAATGCTGGTCACCAGCCTTGTGGC |
| LIF | GGGATGGAAGGCTGTCTTCTTTTGAGGATGATCAGAGA ACTTGGGCATAGGAAC AATCTGGCAGAAGTTTCCAGAAGGAGTCACTTGGCATT CAGGCTC |
| FCGR _{1A} | GAGGATGGAATGTCCTTAAGCGCAGCCCTGAGTTGGAGCTTCAAGTGCTTGGC CTCCAGTTACCAACTCCTGTCTGGTTTCATGTCTTTTCTATCTGG |
| CXCL ₁₂ | CCGCCCCGCCGCCCCGCCATGAACGCCAAGGTCGTGGTCGTGCTGGTCC TCGTGCTGACCGCGCTCTGCCTCAGCGACGGGAAGCCCGTCAGCCT |
| NCF ₄ | GGGACACCAGCAAAAACCTT CAGCTCTCAGAGGAGATTGGGACCAGGAAAACCT GGGAGGATGGGCAGACTTCTGTCTTTGAGGCTAATGGACCCGTGG |
| NCR ₁ | CGATGTTTTGGCTCCTATAACAACCATGCCTGGTCTTTCCCAGTGAGCCAGTGA AGCTCCTGGTCACAGGCGACATTGAGAACCAGCCTTGCACCTG |
| KLRG ₁ | CAGAATGACTATGGACCACAGCAAAAATCTTCTCTTCCAGGCCTTCTTGTCTT GCCTTGTGGCAATAGCTTTGGGGCTTCTGACTGCAGTTCTTCTGA |

| | |
|-----------|---|
| FCGR2A | TGGAGACCCAAATGTCTCAGAATGTATGTCCCAGAAACCTGTGGCTGCTTCAACC ATTGACAGTTTTGCTGCTGCTGGCTTCTGCAGACAGTCAAGCTGC |
| ITGAL | TTCTGGACACATTTGAGAAGCTGAAAGATCTATTTCACTGAGCTGCAGAAGAAGA TCTATGTCATTGAGGGCACAAGCAAACAGGACCTGACTTCCTTCAA |
| CCR2 | TCTGATCTGCTTTTTCTTATTACTCTCCCATTGTGGGCTCACTCTGCTGCAAATGA GTGGGTCTTTGGGAATGCAATGTGCAAATTATTCACAGGGCTGT |
| FAS | TGTTCGAAAGAATGGTGTCAATGAAGCCAAAATAGATGAGATCAAGAATGACAA TGTCCAAGACACAGCAGAACAGAAAGTTCAACTGCTTCGTAATTGG |
| CLEC7A | TGTTAAACTCCGGTAAGTACCTAGCCCACATGATTTGACTCAGAGATTCTCTTTT GTCCACAGACAGTCATCTCAGGAGCAGAAAGAAAAGAGCTCCCAA |
| IL32 | TTCAAAGAGGGCTACCTGGAGACAGTGGCGGCTTATTATGAGGAGCAGCACCCA GAGCTCACTCCTCTACTTGAAAAAGAAAGAGATGGATTACGGTGCC |
| PSMD12 | CAGCCTCCATTTTACAGGAGTTACAGGTGGAACCTACGGGTCAATGGAAAAGA AAGAGCGAGTGAATTTATTTTGGAGCAAATGAGGCTCTGCCTAGC |
| SF3B4 | CGGGACCCTGACACAGGCAACTCCAAAGTTATGCCTTTATTAATTTTGCTTCAT TTGATGCTTCGGATGCAGCAATTGAAGCCATGAATGGGCAGTACC |
| TNFRSF10C | TGGCCCGGATCCCCAAGACCCTAAAGTTCGTCGTCGTCATCGTCGCGGTCTCTGCT GCCAGTCTTAGCTTACTCTGCCACCACTGCCCGGCAGGAGGAAGT |
| APOA1BP | CCCCAAAAATCTGCAACCCAGTTTACCGGTGCTACCATTACCTGGGGGGTCTGT TTTGTGCCACCTGCTCTGGAGAAGAAGTACCAGCTGAACCTGCCA |
| ITGAE | GTCCAGAACATCACTCAAGTGGGGAGTGTACCAAGACTGCCTCAGCCATGCAA CACGTCTTAGACAGCATCTTCACCTCAAGCCACGGCTCCAGGAGAA |
| CD79B | TTCCTCCAAGGAGCCTCGGACGTTGTACGGGTTTGGGGTCGGGGACAGAGCGG TGACCATGGCCAGGCTGGCGTTGTCTCCTGTGCCAGCCACTGGAT |
| IL1RL2 | ACCTTCTTGGAAGTAAAAATGGAAGATTATGGCCTTCCTTTCATGTGCCACGCTG GAGTGTCCACAGCATAATTATATTACAGCTCCCAGCTCCGGATT |
| IL23R | CATGCTTTTGGAAAATGATTCACCCAGTGAAACTATTCCAGAACAGACCCTGCTT CCTGATGAATTTGTCTCCTGTTTGGGGATCGTGAATGAGGAGTTG |

| | |
|----------------------|---|
| PTGER ₄ | CACTACGTGGACAAGCGATTGGCGGGCCTCACGCTCTTTGCAGTCTATGCGTCCA ACGTGCTCTTTTGC GCGCTGCCCAACATGGGTCTCGGTAGCTCGC |
| MR ₁ | GAGGTTCTCAGAAGGGACCTGTCAGTTTTTGGTTAAAAGAACCCGGAAAGAGAA GGACTATGGGGGAACTGATGGCGTTCCTGTTACCTCTCATCATTGT |
| TNFSF _{13B} | CCCAACCTTCAAAGTTCAAGTAGTGATATGGATGACTCCACAGAAAGGGAGCAG TCACGCCTTACTTCTTGCCTTAAGAAAAGAGAAGAAATGAAACTGA |
| WWP ₁ | TGCCGATGACACTGTTAATGGAGAATCATCCTCATTTCACCAACTGATAATGCG TCTGTCACGGGTACTCCAGTAGTGTCTGAAGAAAATGCCTTGTCT |
| CDCA ₅ | ACCCAGTGCGGCTGCAGTCAGAAAGCCCATCGTCTTAAAGAGGATCGTGGCCCA TGCTGTAGAGGTCCCAGCTGTCCAATCACCTCGCAGGAGCCCTAGG |
| CXCL ₁₃ | AGACGCTTCATTGATCGAATTCAAATCTTGCCCCGTGGGAATGGTTGTCCAAGAA AAGAAATCATAGTCTGGAAGAAGAACAAGTCAATTGTGTGTGTGG |
| CLU | CTTCTGACTCGGACGTTCCCTCCGGTGTCACTGAGGTGGTCGTGAAGCTCTTTGA CTCTGATCCCATCACTGTGACGGTCCCTGTAGAAGTCTCCAGGAA |
| IL6 | GGCACTGGCAGAAAACAACCTGAACCTTCCAAAGATGGCTGAAAAGATGGATG CTTCCAATCTGGATTCAATGAGGAGACTTGCCTGGTGAAAATCATC |
| CFP | CTGGGGGGTGGTGTGACGCTGGAAGACTGCTGTCTCAACTGCCTTTGCCTAC CAGAAACGTAGTGGTGGGCTCTGTCAGCCTTGACAGTCCCCACGAT |
| CD8 ₁ | CAGTGCTCAAGAACAATTTGTGTCCCTCGGGCAGCAACATCATCAGCAACCTCTT CAAGGAGGACTGCCACCAGAAGATCGATGACCTCTTCTCCGGGAA |
| C1QA | GGTGACCAGGTCTGGGTTGAAAAAGACCCCAAAAAGGGTCACATTTACCAGGGC TCTGAGGCCGACAGCGTCTTACAGCGGCTTCCTCATCTTCCCATCTG |
| BATF | CCTGGCAAACAGGACTCATCTGATGATGTGAGAAGAGTTCAGAGGAGGGAGAAA AATCGTATTGCCGCCAGAAAGAGCCGACAGAGGCAGACACAGAAGG |
| BST ₂ | GAAGCTGGCACATCTTGAAGGTCCGTCCTGCTCGGCTTTTCGCTTGAACATTCC CTTGATCTCATCAGTTCTGAGCGGGTCATGGGGCAACACGGTTAG |
| RELB | AGCAGGGACAGATGCGCCGGATGGATCCTGTGCTTTCCGAGCCCGTCTATGACA AGAAATCCACAAACACATCAGAGCTGCGGATTTGCCGAATTAACAA |

| | |
|---------|---|
| CD209 | TAGAGCTTGTTTTTCTGGCCCATCCTTGGAGCTTATGAGTGAGCTGGTGTGGGA TGCCTTTGGGGGTGGACTTGTGTTCCAAGAATCCACTCTCTCTTC |
| TNFSF15 | GGTGACCCTTGGCCCGTATTATAAATGCTTCCTATCCTGGGAGACCTCATGGATG AGTCTGAGAGGAAATTTGGCACCAAAATCACTCTCACTCTGGTTT |
| RPL41 | ACTGCATGCTACTGTCTAGAGCTTGTCTCAATGGATCTAGAACTTCATCGCCCTC TGATCGCCGATCACCTCTGAGACCCACCTTGCTCATAAACAAAAT |
| ZNF7 | TGCCCTGGAAGGGTCCACCTTTGTGAGCCGTAAAAAGGTTAATACTATAAAGAA ACTGCATCAGTGTGAAGACTGTGAGAAGATATTTAGGTGGCGTTCA |
| FTSJ3 | AGCCCATTTGACACTGATGGCTCTACGTTTGGCTTGTGACTTTTTGGCCCGTGGT GGCAGCTTCATCACAAAGGTTTTCCGTTCTCGTGAATATCAGCCT |
| TLR2 | CAATGATGCTGCCATTCTCATTCTTCTGGAGCCATTGAGAAAAAGCCATTCCC CAGCGCTTCTGCAAGCTGCGGAAGATAATGAACACCAAGACCTAC |
| BAD | CAGCACAGCGCTATGGCCGCGAGCTCCGGAGGATGAGTGACGAGTTTGTGGACT CCTTTAAGAAGGGACTTCTCGCCCGAAGAGCGCGGGCACAGCAAC |
| IRF8 | CCGCCGCCAGACCAGGTCTTCCGGATGTTTCCAGATATTTGTGCCTCACACCAGA GATCATTTTTTCAGAGAAAACCAACAGATCACCGTCTAAGTGCCTC |
| PRUNE | ACAGCTCTGTTCCATGTAAGTTGCCAACAGTTTCACTGAACAGTGGGGTATGTGA TGGTTTTGGCATGACATCTTCAGTATGAGGGGGACAGTTTGA |
| ASH2L | TATTCTTGGCGGAGCAAAAAGGGAACCAAGTTCCACCAGTCCATTGGCAAACAC TACTCTTCTGGCTATGGACAGGGAGACGTCTGGGATTTTATATTA |
| CLEC6A | TGGGGATGTTGCCAGCTTCTTGAAGTCATTTGGTTCCAGTTGCTACTTCATTT CCAGTGAAGAGAAGGTTTGGTCTAAGAGTGAGCAGAAGTGTGTTG |
| TLCD1 | GGCCATGGGTGCCTTCTTCTCCGGCATCTTTTGGAGCAGCTTTGTCTGGTGGGGGT GTCTTAACACTACTGGTGAAGTCAGCAACATCTTCTCACCATT |
| FCGR3B | AAGGTTTGGCAGTGTCAACCATCTCATCTCTCCACCTGGGTACCAAGTCTC TTTCTGCTTGGTGTGTTACTCCTTTTTGTCAGTGGACACAGGACT |
| MAF | TAATGACTTCGATCTGATGAAGTTTGAAGTGAAAAAGGAACCGGTGGAGACCGA CCGCATCATCAGCCAGTGCGGCCGTCTCATCGCCGGGGGCTCGCTG |

| | |
|------------------------------------|---|
| LILRB ₁ | TGTGTCAGTCACAGGGATGGATGCAAACCTTCCTTCTGACCAAGGAGGGGGCAG CTGATGACCCATGGCGTCTAAGATCAACGTACCAATCTCAAAAATA |
| CEACAM ₃ | GCCTTCCTGTGGGGGCGTCGCCGGCATCGTGACCGGGGTCTGGTCGGAGTGG CGCTGGTGGCCGCGCTGGTGTGTTTCTGCTCCTTGCCAAAACCTGG |
| GPI | CACCGAGGGTCGAGCCGTGCTGCACGTGGCTCTGCGGAACCGGTCAAACACACC CATCCTGGTAGACGGCAAGGATGTGATGCCAGAGGTCAACAAGGTT |
| STAT _{5B} | CCACTTCAGGAATATGTCCCTGAAACGAATTAAGAGGTCAGACCGTCGTGGGGC AGAGTCGGTGACAGAAGAAAATTTACAATCCTGTTTGAATCCCAG |
| DCAF ₇ | TGCTGATGGCTCGGTGCGGATGTTTGACCTCCGCCATCTAGAACACAGCACCATC ATTTACGAAGACCCACAGCATCACCCACTGCTTCGCCTCTGCTGG |
| ALG8 | TACCATTCCACAGATTTTGAAGTACACCGAAACTGGCTTGCTATCACTCACAGTT TGCCAATATCACAGTGGTATTATGAGGCAACTTCAGAGTGGACGT |
| RHPN ₁ -AS ₁ | TGACTGTCAGCAAGCATGTGTGACTGAAATGGTATCCAGTTCTTACGTATTTGTT TTGCAGGCTTCTTGTTTCAGCAGACAGAGGCCATTGGGAAGTTAA |
| MRPL ₂₄ | CGGGAAGCAGGGCAAAGTGGTTCAAGTTATCCGGCAGCGAAACTGGGTGGTCGT GGGAGGGCTGAACACACATTACCGCTACATTGGCAAGACCATGGAT |
| ZNF ₅₀₀ | CAGCCTCGCCCTTCCTTTCCGGCCTGGTCCCAGGTGCCCGTGAACCTGGAGGACGT GGCTGTATACCTTTCTGGGGAGGAGCCAAGATGCATGGACCCAGC |
| ZC _{3H3} | GCCGGTGTGCTCCTACTTCCTGAAGGGCATCTGCAGCAACAGCAACTGTCCCTAT AGCCACGTGTACGTGTCCCGCAAGGCCGAGGTCTGCAGCGACTTC |
| GOLPH _{3L} | TCCTGAGGAAGCTGATGTGTTATTCCTTCTCTGCATCGAAGGATCAGGAAGTTTG TGCTCTCTGCGTGGCTAAGTTTTTTCACCTACTAGGACGGGGGTGG |
| CDK ₁₂ | GATGGAAAGGAGTCCAAGGGTTCACCTGTATTTTTGCCTAGAAAAGAGAACAGT TCAGTAGAGGCTAAGGATTCAGGTTTGGAGTCTAAAAAGTTACCCA |
| CASC ₁ | TCCTTGAAGACAATGTGGTGGATTTATGCCAGTTCACAACCTCTGGGTGGAGTATA CCACTTGGATATTTTGGAGCTTCCTCCACAGTGTAACCAGTGAA |
| MAF ₁ | ATCCCTTCGGGGAGGATGGTAGCCTCTGGTCCTTCAACTACTTCTTCTACAACAA GCGGCTCAAGCGAATCGTCTTCTTTAGCTGCCGTTCCATCAGTGG |

| | |
|---------------------|---|
| SNX ₂₇ | TCAAAGAGGACAGCCTTTCCTCCCTCACCTTCTCCAAATCTAGGTGAAATCACAG AGTACAAAACGTGAGAATGCTGAATGTGTAAAGTTGCAGAGGGAT |
| USP ₃₅ | AGGGTACACAGAGATTCTCTCAGATATGGAAGTAAGACCTAAGTCCCTTTCATTG GGGATCAGTCCCATTAAAACCTTACACCCAAGTGTCTTGTTAAC |
| PMF ₁ | GGCTACCTCTGAGAACGGCTGAAATGGTGCCAGTCCATCAGCAGTGATGGAAT TTGCTGGAGGACTAGGCCAGAGCAAGCCTCACTGCCACTGTGCCTT |
| POGK | TAATGATAGGGCAGGATTCGTATGCAAGCTCTTGTTTCTCAGGCTGCCTGCAGA AGAAGTCGCTATAAATTATCTGTTGTCTACATGGTACAAGGCCCA |
| ZNF ₂₅₀ | AAGCAGCCTTCAGCCCAAGAGAAGTCTCTGTAACTCTATAGGAAGCTTTTCTTT GGCGATTCAGTGTACAAAATAACTCCAGAAAGAAGCACTTAGCG |
| AQP ₁₁ | TCTGCTCCTTCTCTTCCACAGCGCTCTGCTGCACTTCCAGGAAGTCCGAACCAA GCTTCGTATCCACCTGCTGGCTGCACTCATCACCTTTTTGGTCTA |
| GPAA ₁ | TATTGGGCCAAAGATATCGTCTTCCCTGGTAACAGAACATGACCTTCTGGGCACTG AGGCTTGGCTTGAAGCCTACCACGATGTCAATGTCACTGGCATGC |
| OTUD _{6B} | ACAAGAAGAGGAGGAAGCAACTCACCGAAGATGTGGCCAAGTTGGAAAAAGAA ATGGAACAGAAACATAGAGAGGAAGTGGAGCAATTGAAGCTGACTAC |
| MRPL ₉ | CAGAAGAGCCTATCACACGGTGGGGCGAGTATTGGTGTGAGGTGACGGTAAATG GGCTTGATACTGTGAGAGTGCCTATGTCTGTCTGTAACCTTGAGAA |
| FAM _{173A} | GCGGCAGGTGGAGCACGTGTTGTCGCTGCTGCGAGGACGCCCCGAAAAACGGT GGATCTGGGCTCTGGCGACGGCAGGATCGTGTGGCGGCCACAGG |
| FAM _{91A1} | GTGCAGGCTGGCTATATCACAGAAGATGACATCAAGATATGCACTTTCCTGAG AAATGCGCTGTTGATAAGATCATCGATTTCAGGCCCTCAACTCTCTG |
| THEM ₆ | TATGCTGGGGCTGCTGGTGGCGTTGCTGGCCCTGGGGCTCGCTGTCTTTGCGCTG CTGGACGTCTGGTACCTGGTGCCTTCCGTGCGCCGTGCTGCGC |
| TARS ₂ | GTACAACAGATCAGCTGGAAGCAGAGATCCAAAGCTGTCTTGATTCCTCCGTTT CGTCTATGCCGTTCTTGGCTTCTCCTTCCGCCTGGCACTGTCCAC |
| MTERF ₃ | TTGAACTTAGTGTGAAGAAGACTAGAGATCTGGTAGTTCGTCTCCAAGGCTGCT AACTGGAAGTCTGGAACCCGTGAAAGAAAATATGAAGGTTTATCG |

| | |
|-----------|--|
| GLI4 | GCGGCCAGTGCGGCCGCGCCTTCAGCCACAGCTCGCACTTCACGCAGCACCTGC GCATCCACAACGGCGAGAAGCCCTACAAGTGCGGCGAGTGCGGCCA |
| VPS28 | CCGCTGCATCGCAGACGTGGTCTCGCTCTTCATCACGGTCATGGACAAGCTGCGT CTGGAGATCCGCGCCATGGATGAGATCCAGCCCGACCTGCGAGAG |
| FLYWCH2 | TGCTCTGTGGCGCCCGCAAGTCCCTGTAACCTTGACAACAGGCGCATCCTCCCA GGCCACCAACCCAGCCATAGGCTCTTCTCTGTCCGCAGGGCTTCT |
| KCTD21 | CTCTCCTCTTCCAGCATGGAGGTCTTCAACGCCAACATCTTCAGCACCTCCTGCC TCTTCCTCAAGCTCCTTGGCTCTAAGCTCTTCTACTGCTCCAATG |
| COA3 | CCTTCACCTTCCAGGGACGCAGTTGTTACGAGGTTAGACGTGGCAGCTCTGTGCA GTGTTTGAGCCTACAGTGGGATACATAGGGTCAAATTGAGAATAA |
| ZMYND10 | GCGCTGCTCACGATGCCAGAATGAGTGGTATTGCTGCAGGGAGTGCCAAGTCAA GCACTGGGAAAAGCATGGAAAGACTTGTGTCTTGGCAGCCCAGGGT |
| DERL1 | GAGTAGTTGGGTTGCTTTGTGTTAGGAGGATCCAGATCATGTTGGCTACAGGGA GATGCTCTCTTTGAGAGGCTCCTGGGCATTGATTCCATTTCAATCT |
| TNFRSF11A | AGTACATGTCTTCTAAATGCACTACTACCTCTGACAGTGTATGTCTGCCCTGTGG CCCGGATGAATACTTGGATAGCTGGAATGAAGAAGATAAATGCTT |
| KIAA0125 | ACAGTTCTGAAGTCAAAGGCTGATGTCCTGTTTCTCTTTCCCTCTGTGACCGACT CCCTTCCAGTGGAACAAGTACCCACAGCTTGGTTTGAATTTCT |
| IL11RA | GATTCCACCTATAATTCTGTCTTGCTGGTGTGGATAGAAACCAGGCAGGACAGTA GATCCCTATGGTTGGATCTCAGCTGGAAGTTCTGTTTGGAGCCCA |
| CSF3R | AGGCCCTTTCAGCTCTATGAGATCATCGTGACTIONCCTTGTACCAGGACACCATGG GACCCTCCCAGCATGTCTATGCCTACTCTCAAGAAATGGCTCCCT |
| CD84 | TCTGCTAGAACAGTGCCGTGCTTTTCCACAGAAGGTTAGACCCTGAAAGAGATG GCTCAGCACCACCTATGGATCTTGCTCCTTTGCCTGCAAACCTGGC |
| HLA-DQB2 | GGTGTGGCCAGATAACATCTATAACCGCGAGGAGTACGGGCGCTTCGACAGCGAC GTTGGGGAGTTCCAGGCGGTGACCGAGCTGGGGCGGAGCATCGAGG |
| CD6 | AACCCTGGACACTGCATTACAGACCCGCCATCCCTGGGCCCTCAGTATCACCCGA GGAGCAACAGTGAGTCGAGCACCTCTTCAGGGGAGGATTACTGCA |

| | |
|----------|---|
| CYB5D2 | GTCAGCCGCTGAGATGCTGACACTTCACAATTGGCTTTCATTCTATGAGAAGAAT TATGTGTGTGTTGGGAGGGTGACAGGACGGTCTACGGAGAGGAT |
| IL1R2 | TCAGTCTCCACTTCCCGTGTCTCTGGAAGTTGTCAGGAGCAATGTTGCGCTTG TACGTGTTGGTAATGGGAGTTTCTGCCTTCACCCTTCAGCCTGCG |
| LILRB2 | GATACGACCAGAGCTTGTGAAGAACGGCCAGTTCCACATCCCATCCATCACCTG GGAACACACAGGGCGATATGGCTGTCAGTATTACAGCCGCGCTCGG |
| FLT3LG | CGCTGGATGGAGCGGCTCAAGACTGTCGCTGGGTCCAAGATGCAAGGCTTGCTG GAGCGCGTGAACACGGAGATACTTTGTACCAAATGTGCCTTTC |
| HLA-DQA1 | TGAGGTTCTGAGGTCACAGTGTTCCTCAAGTCTCCCGTGACACTGGGTCAGCCC AACACCCTCATTGTCTTGTGGACAACATCTTCCTCCTGTGGTC |
| CX3CL1 | ATTGTGGGAAGGGGAGATAAGGGTATCTGGTGACTTTCCTCTTTGGTCTACACTG TGCTGAGTCTGAAGGCTGGGTTCTGATCCTAGTTCCACCATCAAG |
| IL18R1 | CAGTGAATTAGGAAAAACGTAAGGCTCAACTGCTCTGCTTTGCTGAATGAAG AGGATGTAATTTATTGGATGTTCCGGGAAGAAAATGGATCGGATCC |
| ATG5 | AGAATGACAGATGACAAAGATGTGCTTCGAGATGTGTGGTTTGGACGAATTCCA ACTTGTTCACGCTATATCAGGATGAGATAACTGAAAGGGAAGCAG |
| S100A11 | GATGAAGAACTGGACACCAACAGTGATGGTCAGCTAGATTTCTCAGAATTTCTT AATCTGATTGGTGGCCTAGCTATGGCTTGCCATGACTCCTCCTC |
| NOD1 | GTAAGTCAGGAGACTTTCCTTCGGTTTCTGCCTTTGATGGCAAGAGGTGGAGATT GTGGCGGCGATTACAGAAAACATCTGGGAAGACAAGTTGCTGTTT |
| AICDA | ACACTCTGGACACCACTATGGACAGCCTCTTGATGAACCGGAGGAAGTTTCTTTA CCAATTCAAAAATGTCCGCTGGGCTAAGGGTCGGCGTGAGACCTA |
| GPATCH3 | CCGACCCAGCCGCTGAGGGCCAGCTTCTCTCTCAGACTTCGGCCACCGATGTCCG GCCTCTCTCCACTCGAGACTCTACTCCAATCCAGACCCGCACCTG |
| WDR19 | CATTTGATAGAAAGCGAAATCTTGGATGCTCAAGAAGAACGTGAGACTCGGCTT TTCCCAGCAGTGGATGATAAGTGCCGTATCTTATGCCATGCCTTAA |
| IFNA7 | TCAGATTCCCAGAGGAGGAGTTTGTGATGGCCACCAGTTCCAGAAGACTCAAGCCA TCTCTGTCTCCATGAGATGATCCAGCAGACCTTCAATCTCTTCAG |

| | |
|--------------------------------|--|
| IL ₇ R | AGAACTGGATGACTACTCATTCTCATGCTATAGCCAGTTGGAAGTGAATGGATCG CAGCACTCACTGACCTGTGCTTTTGGAGACCCAGATGTCAACATC |
| CDCA ₃ | TCCGGCTCCGGGACCCTTAGAGTCCATTCGCGTGCCTTCTATTGAAGAGCAGGAA CTGACGGTTGTTTTAGGCCTCTCAATTCCAGAAAGTTTCTGGGG |
| MAPK ₁₁ | CACCCTGATGGGCGCCGACCTGAACAACATCGTCAAGTGCCAGGCGCTGAGCGA CGAGCACGTTCAATTCCTGGTTTACCAGCTGCTGCGCGGGCTGAAG |
| MASP ₁ | TATTTCTTCAAAGACCAAGTGCTCGTCAGCTGTGACACAGGCTACAAAGTGCTGA AGGATAATGTGGAGATGGACACATTCCAGATTGAGTGTCTGAAGG |
| LILRA ₁ | CCCAGCCCTGTGGTGACCTCAGGAGGGAACGTGACCCTCCATTGTGTCTCACAG GTGGCATTGTTGGCAGCTTCATTCTGTGTAAGGAAGGAGAAGATGAAC |
| TYK ₂ | GGCTGCTTGCTGAGGGCCGGGGATGACTGCTTCTCTCTGCGTCGCTGTTGCCTGC CCCAACCAGGAGAAACCTCCAATCTCATCATCATGCGGGGGGCTC |
| IL ₄ R | AGAGGAGAATGGGGGCTTTTGCCAGCAGGACATGGGGGAGTCATGCCTTCTTCC ACCTTCGGGAAGTACGAGTGCTCACATGCCCTGGGATGAGTTCCCA |
| CD ₁₉ | CACAGCTCAAGACGCTGGAAAGTATTATTGTCACCGTGGCAACCTGACCATGTCA TTCCACCTGGAGATCACTGCTCGGCCAGTACTATGGCACTGGCTG |
| PGAP ₃ | AGGAATAGAATGGAGGGAGCTCCAGAACTTTCCATCCCAAAGGCAGTCTCCGT GGTTGAAGCAGACTGGATTTTTGCTCTGCCCTGACCCCTTGTTCC |
| CFI | GTGGGGGAATTTATATTGGTGGCTGTTGGATTCTGACTGCTGCACATTGTCTCAG AGCCAGTAAAACTCATCGTTACCAAATATGGACAACAGTAGTAGA |
| IL ₁ R ₁ | TGCTAAGGTGGAGGATTCAGGACATTACTATTGCGTGGTAAGAAATTCATCTTAC TGCCTCAGAATTTAAATAAGTGCAAAATTTGTGGAGAATGAGCCT |
| TNFSF ₁₄ | GGCGTGTCAGCCCTGCTCCAGACACCTTGGGCATGGAGGAGAGTGTCGTACGGC CCTCAGTGTTTGTGGTGGATGGACAGACCGACATCCATTACAGAG |
| CD ₃₄ | GACAACCTTGAAGCCTAGCCTGTCACCTGGAATGTTTCAGACCTTTCAACCACT AGCACTAGCCTTGAACATCTCCCACTAAACCCTATACATCATCT |
| APH ₁ B | CCTTCATTGCCTTCGGGCCTGCGCTCGCCCTTTATGTCTTACCATCGCCACCGA GCCGTTGCGTATCATCTTCTCATCGCCGGAGCTTTCTTCTGGTT |

| | |
|----------------------|---|
| CD8 _o | GATATCACTAATAACCTCTCCATTGTGATCCTGGCTCTGCGCCCATCTGACGAGG GCACATACGAGTGTGTTGTTCTGAAGTATGAAAAAGACGCTTTCA |
| IL7 | AGGGTCCTGGGAGTGA CTATGGGCGGTGAGAGCTTGCTCCTGCTCCAGTTGCGG TCATCATGACTACGCCCGCTCCCGCAGACCATGTTCCATGTTTCT |
| CD244 | CTCCAGGCGCTGGGGCTTTCTCAGTGGCCTTGTCAGCTCACAGCAGGCGTTAACA GCCTCTAATTGAGGAAACTGTGGCTGGACAGGTTGCAAGGCAGTT |
| ENTPD ₁ | CCCATATGAAACCAATAATCAGGAAACCTTTGGAGCTTTGGACCTTGGGGGAGC CTCTACACAAGTCACTTTTGTACCCCAAACCAGACTATCGAGTCC |
| CD274 | AGCTTCCCGAGGCTCCGCACCAGCCGCGCTTCTGTCCGCCTGCAGGGCATTCCA GAAAGATGAGGATATTTGCTGTCTTTATATTCATGACCTACTGGCA |
| CD2 | TCCATGAGGTGTTTTCTGTGTGCAGAACATTGTCACCTCCTGAGGCTGTGGGCCA CAGCCACCTCTGCATCTTCGAACTCAGCCATGTGGTCAACATCTG |
| TNFRSF ₄ | CAACTCTGCACCGTTCTAGGTGCCGATGGCTGCCTCCGGCTCTCTGCTTACGTAT GCCATGCATACCTCCTGCCCCGCGGGACCACAATAAAAACCTTGG |
| RPS6KB ₁ | TTACTTGGCAGAAATCTCCATGGCTTTGGGGCATTACATCAAAGGGGATCATC TACAGAGACCTGAAGCCGGAGAATATCATGCTTAATCACCAAGGT |
| HLA-DQA ₂ | GCAGTTGCCTATGTTTAGCAAATTTATAAGTTTTGACCCGCAGAGTGCCTGAGA AATATGGCTGTGGGAAAACACACCTTGAATTTCATGATGAGACAG |
| STAT ₂ | GATTTGGGACTTTGGTTACCTGACTCTGGTGGAGCAACGTTTCAGGTGGTTCAGG AAAGGGCAGCAATAAGGGGCCACTAGGTGTGACAGAGGAACTGCAC |
| IFIH ₁ | AAGTGGAAGAGCAACTTCTTTCAACCACAGTTCAGCCAAATCTGGAGAAGGAGG TCTGGGGCATGGAGAATAACTCATCAGAATCATCTTTTGCAGATTC |
| NOD ₂ | ATGCTGGACCTGGCATGGGAGCGGGTTTTCGTCAGCCAGTATGAATGTGATGAA ATCAGGTTGCCGATCTTCACACCGTCCAGAGGGCAAGAAGGCTGC |
| CD96 | CAAGGGTACATAGTAATAAACCAGCCCAATCAGACAACCTTGACCATTTGGTGTAT GGCTCTGTCTCCAGTCCCAGGAAATAAAGTGTGGAACATCTCATC |
| PGR | AGCCAGCCAGAGCCCACAATACAGCTTCGAGTCATTACCTCAGAAGATTTGTTTA ATCTGTGGGGATGAAGCATCAGGCTGTGATTATGGTGTCTTACC |

| | |
|----------------------|--|
| MTBP | AGTAATAGCAGGGAATCATTATCCTTGGCTGATCTCTATGAAGAAGCTGCAGAA AATTTGCATCAGCTGTCAGACAAGCTTCCTGCTCCTGGTAGAGCAA |
| ILi8RAP | TACTCAGTCGGATACTGTGAGTTCGTGGACAGTCAGAGCTGTTGTTCAAGTGA GAACCATTGTGGGAGACACTAAACTCAAACCAGATATTCTGGATC |
| MLPH | GACAGGACAGAGAGACAGAGCAGCCCTGCACTGTTTTCCCTCCACCACAGCCAT CCTGTCCCTCATTGGCTCTGTGCTTTCCACTATACACAGTCACCGT |
| ATAD ₂ | TGAAGCTAAGAGAACAGCACCAAGTATAGTGTATGTTCCCTCATATCCACGTGTGG TGGGAAATAGTTGGACCGACACTTAAAGCCACATTTACCACATTA |
| CTPS ₁ | ATATGATCGCTTGTGCTGGAGACCTGCTCTATTGCCCTTGTGGGCAAATACACGAAG TTCTCAGACTCCTATGCCTCTGTCATTAAGGCTCTGGAGCATTCT |
| HLA-DRB ₁ | AGCACGGTCTGAATCTGCACAGAGCAAGATGCTGAGTGGAGTCGGGGGCTTTGT GCTGGGCCTGCTCTTCCTTGGGGCCGGGCTGTTTCATCTACTTCAGG |
| POLR ₃ C | GCTGGGGAGCCCAAGGCCAAGAGACCAAAATATACTACAGATAACAAGGAGCCC ATTCCAGATGATGGGATTTATTGGCAGGCCAACCTTGACAGATTCC |
| ICOSLG | TGAGCTCACCTTCACGTGTACATCCATAAACGGCTACCCCAGGCCCAACGTGTAC TGGATCAATAAGACGGACAACAGCCTGCTGGACCAGGCTCTGCAG |
| TNFRSF ₁₄ | CTGCAGCCCAGGCCACTTCTGCATCGTCCAGGACGGGGACCACTGCGCCGCGTG CCGCGCTTACGCCACCTCCAGCCCAGGGCCAGAGGGTGCAGAAGGGA |
| HEXIM ₂ | GGTGTCACTAGTTCAGGCGTCTGCTGAAAGATTTGGAACAGAAGATGATGGCC ACTCCGAACCAGACCGCTGTAATGCAGAGTCACCAGTGGCCCTGG |
| WDYHV ₁ | AGCTCTGTGAATACATCAAAAACCATGACCAGTATCCTTTAGAAGAATGTTATGC TGTCTTCATATCTAATGAGAGGAAGATGATACCTATCTGGAAACA |
| TRBV ₁₂₋₃ | ATGTAAACCAATTTTCAGGCCACAACCTCCCTTTTCTGGTACAGACAGACCATGATG CGGGGACTGGAGTTGCTCATTTACTTTAACAACAACGTTCCGATA |
| TRDC | TTGAAGTGAAGACAGATTCTACAGATCACGTAACCAAGGAAACTGAAAACA CAAAGCAACCTTCAAAGAGCTGCCATAAACCAAGCCATAGTTCA |
| TRGC ₁ | TTTAATTGGATGACATCAAAATTGAACATCCAAGGTAAGAAACAGCATGGCAATT GGGCTGTGGAATTCTGTATTGGTTGTAAGAATGGTCCAACACCCCA |

| | |
|-------------------|--|
| TRBC ₁ | TCCGCTGTCAAGTCCAGTTCTACGGGCTCTCGGAGAATGACGAGTGGACCCAGG ATAGGGCCAAACCCGTCACCCAGATCGTCAGCGCCGAGGCCTGGGG |
| TRAC | AGCATTATTCCAGAAGACACCTTCTTCCCCAGCCCAGAAAGTTCCTGTGATGTCA AGCTGGTCGAGAAAAGCTTTGAAACAGATACGAACCTAAACTTTC |
| BOLA ₁ | TCCATACATACTCTCCGAAGATAGCAACTTGCTTCAGGTCAAAGTGAACCCGAGA AAAGAGAAGAATCACTCACTACTGCTCTTGCCCTGGACTATTAG |
| MFSD ₃ | ACGCTGGCACAATCTTGAGAGGGTCAGCCTTGCTGAGCCTATGTCTGCAGCACTT CTTGGGAGGCCTGGTCACCACAGTCACCTTCACTGGGATGATGCG |
| PPP1R16A | GCTCCTGAAGCAGGTCCTCTTCCCTCCCAGTGTTGTCCTTCTGGAGGCCGCTGCC CGAAATGACCTGGAAGAAGTCCGCCAGTTCCTTGGGAGTGGGGTC |
| PSMD ₄ | GTTTGGCTGATGCTCTCATCAGTTCTCCGATTTTGGCTGGTGAAGGTGGTGCCAT GCTGGGTCTTGGTGCCAGTGACTTTGAATTTGGAGTAGATCCCAG |
| XCL ₂ | TGAAGTCTCACATAGGAGGACCTGTGTGAGCCTCACTACCCAGCGACTGCCAGT TAGCAGAATCAAGACCTACACCATCACGGAAGGCTCCTTGAGAGCA |
| ARG ₁ | TCAATGACTGAAGTGGACAGACTAGGAATTGGCAAGGTGATGGAAGAAACACTC AGCTATCTACTAGGAAGAAAGAAAAGGCCAATTCATCTAAGTTTTG |