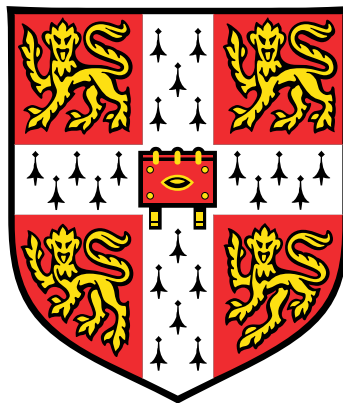


Resolving Cell Fate

Experimental, computational and mathematical methods
in single cell transcriptomic analysis



Wajid Bin Jawaid

Department of Haematology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

March 2019

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Acknowledgements and specified in the text.

The content is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in Acknowledgements and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in Acknowledgements and specified in the text.

This dissertation contains 57,624 words excluding references.

Wajid Bin Jawaid
March 2019

Resolving Cell Fate

Experimental, computational and mathematical methods in single cell transcriptomic analysis

Wajid Bin Jawaid

Abstract

Deeper understanding of the embryological origins of tissues and organs is likely to provide insights into novel clinically relevant preventative and therapeutic strategies. To do so effectively and at a large scale so as to have clinical significance requires an exhaustive and meticulously accurate knowledge of normal morphological development and the underlying molecular pathways.

A variety of lineage tracing and classical molecular biology techniques have led to key insights but emerging technologies now offer the possibility of more fine grained and precise measurements at the level of the single cell rather than ensembles of heterogeneous cells. In this study the rapidly developing technology of single cell RNA sequencing was combined with the development of state of the art computational methods to study murine gastrulation and early organogenesis at a hitherto unprecedented granularity.

A comprehensive analysis of FLK1⁺ cells harvested from gastrulating murine embryos was performed. In-silico reconstruction of pseudo-temporal and pseudo-spacial relations were demonstrated. Using prior knowledge of known driver genes deeper substructure was revealed in clusters that were initially defined by unsupervised algorithms, illustrating that careful implementation of supervised approaches can outperform naïve unsupervised methods. In this way multiple members of the leukotriene branch of the arachidonic acid pathway were found to be enriched within a subset of the endothelial cluster that has a molecular signature consistent with yolk sac derived definitive wave haematopoiesis. This was subsequently validated in an in-vitro embryonic stem cell differentiation colony assay.

By developing a novel adaptation to tSNE dimensionality reduction that now allows new data points to be mapped back to previously calculated embeddings, the FLK1⁺ data set became a reference to which cells from other experiments were mapped. The *Tal1*^{-/-} knockout mutant was characterised, reaffirming the known phenotype with complete failure of embryonic haematopoiesis. Analysing the *Tal1*^{-/-} endothelial cluster shows definitively in the in-vivo organism no activation of an alternative cardiac fate programme as was previously postulated from an in-vitro model system. Additionally tSNE mapped *Brachyury* cells into the void in the FLK1⁺ dataset and identified a node like population.

Loss of spacial context remains an Achilles' heel of single cell protocols. Generation of the single cell suspensions leads to disruption of cellular contacts and loss of any spacial information. A novel method is described which uses tSNE of bulk spacial data to pre-condition a tSNE map upon which single cells can then be computationally positioned reconstructing spacial context.

To model gene interactions and perturbations from single-cell data, a hybrid feed-forward deep neural network was trained on branching pseudo-temporally arranged single cell qPCR data of in-vivo wild-type murine developmental haematopoiesis. Strikingly despite the model never having 'seen' a mutant, in-silico gene perturbations in the deep neural network are able to faithfully reproduce the *Tal1*^{-/-} phenotype.

In summary the use of single cell transcriptomics to probe early murine embryology combined with development of new methods has uncovered a novel pathway in embryonic haematopoietic development and allowed in-silico reconstruction of a short period of early embryonic haematopoiesis. Critically these methods have broad application within the fields of developmental and stem cell biology.

Acknowledgements

I would like to thank my supervisor Berthold Göttgens for the opportunity to join the lab and for providing an environment conducive to learning and developing novel approaches to single cell analyses. I would like to extend my thanks to Jennifer Nichols my second supervisor for her support and collaboration. I am very grateful to the Wellcome Trust for funding my fellowship.

Special thanks goes to Yosuke Tanaka who taught me to dissect mouse embryos, to Fernando Calero-Nieto for generating 10x Genomics® Chromium™ libraries and help with optimising Smart-seq2 libraries, to Vasilis Ladopoulos who helped with setting up FACS protocols, choosing antibodies for the $T^{nEGFP-CreERT2/+}$ experiments and performing the in-vitro stem cell haematopoietic differentiation colony assays and to Sonia Nestorowa for helping to generate Smart-seq2 libraries for the $T^{nEGFP-CreERT2/+}$ experiments.

I am also grateful to Ben Steventon for help with confocal imaging of $T^{nEGFP-CreERT2/+}$ embryos and Alfonso Martinez-Arias for allowing use of his facilities and many interesting discussions. I would additionally like to thank all members of the Göttgens lab for stimulating conversations and friendly chat including from the wet lab Fernando Nieto, Vasilis Ladopoulos, Nicola Wilson, Moosa Qureshi, Winnie Lau, Sarah Kinston, Joakim Dahlin, Caroline Guibentif, Iwo Kucinski and from the bioinformatics office Chee Lim, Lila Diamanti, Xianon Wang, Rebecca Hannah and Fiona Hamey.

I would also like to thank Lorenz Wernisch and John Reid for their conversations and support when developing methods in pseudo-time inference specifically the idea of using auto-regressive models.

I would particularly like to thank Professor Paul Losty and Mr Edwin Jesudason at Alder Hey Children's Hospital were pivotal in their roles as academic supervisors in encouraging and supporting me in applying successfully for a research fellowship.

Finally and most importantly I would like to thank my family especially my parents for their unwavering support.

I would like to dedicate this thesis to my loving family.

Table of contents

List of figures	xv
List of tables	xxi
Nomenclature	xxv
Papers	xxix
Software	xxxii
1 Introduction & Aims	1
1.1 Motivation and Clinical relevance	1
1.1.1 Congenital Anomalies	1
1.1.2 Non-communicable diseases	3
1.1.3 Developmental Biology	3
1.1.4 Stem Cell Biology	5
1.2 Embryology - A brief overview	7
1.2.1 Pre-implantation	7
1.2.2 Post-implantation	8
1.2.3 Embryology of Cardiac Development	12
1.2.4 Developmental Haematology	13
1.3 Experimental approaches in embryology	15
1.3.1 Model organisms	15
1.3.2 Computational lineage reconstruction	16
1.4 Single cell methods	20
1.5 Aims	21
2 Methods	25
2.1 Mouse work	25

Table of contents

2.1.1	Embryo dissections and imaging	26
2.1.2	Genotyping	26
2.2	Processing embryos	27
2.2.1	Generating single cell suspensions	27
2.2.2	Antibody hybridisation	27
2.2.3	Fluorescent activated cell sorting	28
2.3	Library preparation	30
2.3.1	Smart-seq2	30
2.3.2	10x Genomics®	33
2.4	Sequencing	41
2.4.1	Smart-seq2 experiments	41
2.4.2	10x Genomics® Chromium™ single cell experiments	41
2.5	Pre-processing	41
2.5.1	Alignment and QC	42
2.5.2	Normalisation and feature selection	42
2.6	Data exploration	43
2.6.1	Clustering	43
2.6.2	Visualisations	43
2.6.3	Pseudotime inference	44
2.7	ESC to blood differentiation assay	44
3	Single-cell transcriptomic analysis of murine gastrulation	47
3.1	Background	47
3.2	Cell sampling and library preparation	50
3.3	Sequencing and Alignment	56
3.4	Quality Control	60
3.5	Normalisation	63
3.6	Feature selection	64
3.7	Identification of cell populations	67
3.7.1	Hierarchical clustering	67
3.7.2	Assigning identities to clusters	68
3.7.3	PCA	72
3.7.4	tSNE	77
3.8	E6.5 cluster is polarised	80
3.9	Spacial axis dichotomy - ‘Pseudospace’	83
3.10	Charting a developmental journey - ‘Pseudotime’	90
3.10.1	Diffusion Maps	90

3.10.2	Ontogenic reconstruction of embryonic blood development	92
3.10.3	Inferring gene dynamics during primitive-wave haematopoiesis . . .	98
3.11	Coarse temporal data corroborates findings	107
3.12	Finding substructure within clusters	111
3.12.1	Structure within the pharyngeal mesoderm	111
3.12.2	Discovery and validation of a novel pathway	114
3.13	Single-cell resolved characterisation of a mutant	122
3.14	Brachyury plugs the hiatus	132
3.14.1	Genetic <i>Brachyury</i> reporter	133
3.14.2	Mouse crosses, embryo collection and microscopy	135
3.14.3	FACS	136
3.14.4	Library preparation and sequencing	138
3.14.5	Alignment and QC	140
3.14.6	Data exploration	144
3.15	Conclusions	151
3.15.1	Technical challenges	151
3.15.2	Cell type assignment	153
3.15.3	Mutants and reporters	154
3.15.4	Novel pathway discovery and validation	154
3.15.5	Pseudotime and pseudospace	156
3.15.6	Summary	157
4	Single-cell census of mouse organogenesis	159
4.1	Background	159
4.2	Embryo collection and data generation	160
4.3	Defining clusters	169
4.4	Within cluster substructure	177
4.4.1	Definitive endoderm	177
4.4.2	Neural & mesodermal progenitors	182
4.5	Cardiac divergence convergence	187
4.6	Transcriptional termination variants	195
4.7	Conclusions	201
5	Computational tools development	203
5.1	Background	203
5.2	Reconstructing ordered ontogenic relationships	204
5.2.1	Overview	204

Table of contents

5.2.2	Algorithm	206
5.2.3	Usage and Examples	212
5.3	Projecting on tSNE	216
5.3.1	Background	216
5.3.2	Algorithm	216
5.3.3	Examples	219
5.4	Simulating developmental processes	227
5.4.1	Finding fate trajectories and dynamics of gene expression	228
5.4.2	Autoregressive model	230
5.4.3	Deep neural network	232
5.5	Conclusions	247
6	Discussion	251
6.1	Identifying cell populations	251
6.2	Tracing biologically plausible trajectories	253
6.3	Identifying novel molecular pathways	256
6.4	Computationally simulating developmental processes	257
6.5	Concluding remarks	259
	References	261
	Index	287

List of figures

1.1	Contemporary neonatal intensive supportive care	2
1.2	Mouse blastocyst hatching from zona pellucida	8
1.3	Preimplantation murine embryo	9
1.4	Schematics of mouse embryos	10
1.5	Putative cardiac development model	13
1.6	Common haematopoietic and cardiovascular progenitor.	14
1.7	The three distinct waves of haematopoiesis	15
1.8	Figure demonstrates the averaging effect of bulk sequencing as compared to single cell methods which can resolve the expression pattern of individual cells from which bulk data can be reconstructed. In contrast there is no practical way of inferring single cell expression patterns from bulk data. . .	21
1.9	Graphical summary of the overarching aims of this thesis	23
3.1	Embryo schematics	51
3.2	Experimental strategy	52
3.3	Initial experiment FACS	53
3.4	Overview of Smart-seq2 protocol	55
3.5	Bridge amplification	57
3.6	Illumina® sequencing	58
3.7	QC plots	62
3.8	Comparison of spike-in reads	65
3.9	Selecting highly variable genes	66
3.10	Dynamic tree cut clustering	69
3.11	Heatmap with Ward clustering	69
3.12	Cell population assignments by cell cycle	71
3.13	PCA pairs plots	73
3.14	Box plot of <i>PC1</i> coordinates by lane of sequencing flow cell.	74

List of figures

3.15	Relationship between <i>PC1</i> with QC parameters	75
3.16	PCA loadings plot for <i>PC2</i> and <i>PC3</i>	76
3.17	Published in-situ or immunofluorescence images	76
3.18	tSNE of whole dataset	77
3.19	Gene expression patterns by cluster	79
3.20	Brachyury expression domain from published literature	80
3.21	Capturing the transcriptome-wide gene expression programme of mouse gastrulation	81
3.22	Genes most highly associated with <i>Brachyury (T)</i> expression.	82
3.23	Nascent mesoderm is composed of two zones	83
3.24	Expression of selected transcripts on cropped tSNE	84
3.25	Literature curated in-situs support spacially distinct zones	85
3.26	Expected Hox gene patterns from literature	87
3.27	<i>Hoxa</i> gene patterns in the nascent mesoderm	88
3.28	<i>Hoxb</i> gene patterns in the nascent mesoderm	89
3.29	Diffusion map dimensionality reduction of embryonic blood development	93
3.30	Diffusion components (<i>DC1</i> and <i>DC3</i>) in blood development	94
3.31	Diffusion pseudotime and weight decay	97
3.32	Exclusion of outlier cells	99
3.33	Correlation between diffusion pseudotime and principal curve pseudotime	100
3.34	Gene expression patterns along pseudotime	103
3.35	Prototypic gene expression profiles from the fuzzy clusters	104
3.36	tSNE after removing E6.5 cells	108
3.36	tSNE after removing E6.5 cells (contd.)	109
3.37	Coarse temporal separation of cells. Separated by stage of embryo from which the cell was harvested, primitive streak (PS, purple), neural plate (NP, green) and head fold (HF, red) stage.	110
3.38	Figure showing <i>Nkx2-5</i> expression within a subset of pharyngeal mesoderm	112
3.39	Heatmap displaying \log_{10} normalised counts within the pharyngeal mesoderm cluster	113
3.40	Clustering of endothelial cluster using genes that are highly correlated with <i>Gfi1b</i> , <i>Itga2b</i> and/or <i>Igtb3</i>	114
3.41	117
3.41	Tracks of Chip-Seq data for haematopoietic transcription factors	118
3.42	DNAase I hypersensitivity tracks	119

3.43	Schematic diagram summarising the role of ALOX5, ALOX5AP and LTC4S in the leukotriene pathway	120
3.44	Overview of experimental protocol generating erythroid bodies (EBs) from murine embryonic stem cells	120
3.45	Bar plot showing the fold change in the number of colonies relative to carrier control after treatment with zileuton or LTC ₄	121
3.46	Mutant <i>Tal1</i> LacZ knock-in model generation	123
3.47	Sample of flow cytometry of WT and <i>Tal1</i> ^{-/-} embryos at head fold (HF) stage and at the four somite stage (E8.25)	124
3.48	PCR genotyping gels from <i>Tal1</i> ^{LacZ/+} crosses	124
3.49	QC plots generated for plate 4179 containing FLK1 sorted cells from a wild type <i>Tal1</i> ^{+/+} embryo	125
3.50	QC plots generated for plate 4178 containing FLK1 sorted cells from a null mutant <i>Tal1</i> ^{-/-}	126
3.51	Heatmap of mapped reads on the 96 well plates with null mutants <i>Tal1</i> ^{-/-} on the left column and <i>Tal1</i> ^{+/+} samples on the right	128
3.52	De-novo and projected <i>Tal1</i> data on PCA.	129
3.53	<i>Tal1</i> ^{-/-} cells projected on tSNE	131
3.54	Summary of the targeting strategy for knocking in the target vector containing the additional 2A-nucEGFP-2A-CreERT2 into embryonic stem cells using homologous recombination	134
3.55	Wild type to Brachyury-2A-nucEGFP-2A-CreERT2 [Imuta et al., 2013] cross is predicted to provide 50% GFP and 50% WT murine embryos.	135
3.56	Collage of selected confocal images from each embryo	137
3.57	Summary of sort layout	139
3.57	Summary of sort layout	140
3.58	Bionalyser trace from (a) cleaned cDNA product and (b) cleaned library.	141
3.59	Panel of QC filters as seen previously in figs. 3.15, 3.49 and 3.50. Noticeably a high number of reads are not aligned.	143
3.60	Recalculated de-novo PCA and projected PCA for <i>T</i> ^{nEGFP-CreERT2/+}	144
3.61	Mapping the new <i>T</i> ^{nEGFP-CreERT2/+} cells onto the reference tSNE	146
3.61	Mapping the new <i>T</i> ^{nEGFP-CreERT2/+} cells onto the reference tSNE	147
3.62	Gene expression profiles in <i>T</i> ^{nEGFP-CreERT2/+} and reference cells on the tSNE	148
3.62	Gene expression profiles in <i>T</i> ^{nEGFP-CreERT2/+} and reference cells on the tSNE	149
3.63	<i>T</i> expression and cell cycle allocation in the <i>T</i> ^{nEGFP-CreERT2/+} cells	149

List of figures

3.64	tSNE plots coloured by gene expression displaying genes highly expressed in the putative node/organiser population.	150
4.1	A male B6CBAF1/J and a female B6CBAF1/J were crossed to yield F2 embryos. B6CBAF1/J, an F1 cross, were generated by crossing a female C57BL/6J with a male CBA/J.	161
4.2	Summary of the 10x Genomics® workflow	162
4.3	Distributions of the numbers of cells with a barcode shared with another cell	163
4.4	Views of embryos and haemocytometer chamber for counting cells	165
4.5	Overview of 10x Genomics® Chromium™ library oligo-dT and final library construct	166
4.6	10x Genomics® Chromium™ experimental and computation workflow	167
4.7	10x Genomics® Chromium™ data filtering	167
4.8	Summary of cell counts, numbers of genes expressed in cells and the percentage reads mapped to mitochondrial genes	168
4.9	Selecting highly variable genes in 10x Genomics® Chromium™ data	168
4.10	tSNE dimensionality reduction of 6978 single cell expression profiles on 7585 genomic features	170
4.11	<i>Dppa3</i> is almost exclusively expressed in the PGC cluster	174
4.12	Histograms showing distributions of <i>Xist</i> expression in the 3 embryos	175
4.13	tSNE and DrL plots of definitive endoderm cells	178
4.14	Selection of genes and boxplots of their expression patterns across the Louvain clusters helped to assign putative cell type identities	179
4.14	Selection of genes and boxplots of their expression patterns across the Louvain clusters helped to assign putative cell type identities	180
4.14	Selection of genes and boxplots of their expression patterns across the Louvain clusters helped to assign putative cell type identities.	181
4.15	In-situ hybridisation to <i>Pyy</i> in a 4 somite pair embryo	182
4.16	tSNE and DrL side by side of neural and mesodermal progenitors	183
4.17	Combined gene expression of <i>Sox2</i> , <i>T</i> and <i>Cyp26a1</i> on a single plot	184
4.18	DrL with cells coloured by previously defined clusters, table 4.4 in the left pattern	188
4.19	Heatmap of genes expressed differentially between the putatively assigned <i>Irx4</i> ⁺ ventricular myocytes and <i>Irx4</i> ⁻ atrial/OFT cells.	189
4.20	Genes known to be highly expressed in cardiac cells	193
4.21	Representative bioanalyser traces of the post cDNA amplification reaction	197

4.22 Oligo-dT is hybridising at different regions so suggesting transcripts terminating at different genomic locations	198
4.22 Oligo-dT is hybridising at different regions so suggesting transcripts terminating at different genomic locations	199
4.22 Oligo-dT is hybridising at different regions so suggesting transcripts terminating at different genomic locations	200
5.1 Sampling the manifold sufficiently to allow accurate transitions to be inferred	205
5.2 Sorting strategy for adult haematopoiesis data, from Figure 1 in Nestorowa et al. [2016]	214
5.3 Diffusion maps and SPRING representation of the adult blood dataset . . .	215
5.4 tSNE representation of iris data with 3 species in different colours as indicated by the legend	221
5.5 Enabling tSNE to use prior information	222
5.6 Experimental protocol used by Peng et al. [2016]	223
5.7 Corn plot with single cell data from Peng et al. [2016] projected tSNE . . .	225
5.8 3D positioning of single cell data on a landscape preconditioned using bulk data	226
5.9 Experimental protocol for Moignard et al. [2015] data set	227
5.10 Diffusion maps reconstructed from Moignard et al. [2015]	228
5.11 Gene dynamics along trajectories identified in fig. 5.10c	229
5.12 Autoregressive model inferred on the blood gene expression dynamics . . .	233
5.13 Autoregressive model inferred on the endothelial gene expression dynamics	234
5.14 Abstraction of a network inference function that can predict the next cell along in a developmental trajectory	235
5.15 Graphical representation of a linear regression model	235
5.16 A more flexible model than linear regression as shown in fig. 5.15	236
5.17 Building on the neural network model	237
5.18 Architecture of the artificial neural net used to model gene expression along pseudotime	240
5.19 Two layers of a neural net to help demonstrate both forward and back propagation	242
5.20 Gene expression profiles generated by the neural network reconstructs the bifurcating developmental trajectories from early mesodermal progenitors to blood (a) and endothelium (b)	245
5.21 Neural net gene perturbation simulation results	246

List of figures

6.1	The Monocle 3 workflow. From http://cole-trapnell-lab.github.io/monocle-release/monocle3/	254
-----	--	-----

List of tables

2.1	Genotyping primers.	26
2.2	Genotyping PCR programme.	27
2.3	Staining protocol.	29
2.4	Annealing mix.	30
2.5	Smart-seq2 reverse transcription mixture.	31
2.6	Smart-seq2 reverse transcription thermocycler programme.	31
2.7	PCR mix	32
2.8	PCR mix thermocycler programme	32
2.9	Tagmentation mix	32
2.10	Nextera PCR indexing programme	33
2.11	10x Genomics® Chromium™ single cell 3' experiment cell suspension calculations. The column labelled additional volume provides the additional volume of nuclease-free H ₂ O to be added to each sample after generating the master mix with only 26.5 µl for each sample table 2.12.	34
2.12	10x Genomics® Chromium™ single cell 3' reverse transcription GEM generation Master Mix	34
2.13	10x Genomics® Chromium™ single cell 3' sample preparation in an 8-tube strip.	35
2.14	10x Genomics® Chromium™ single cell 3' GEM-RT	35
2.15	DynaBeads Cleanup Mix	36
2.16	Elution Solution I	36
2.17	Elution Solution II	37
2.18	10x Genomics® Chromium™ cDNA amplification reaction mix.	37
2.19	10x Genomics® Chromium™ single cell 3' cDNA amplification reaction programme.	37
2.20	End repair and A-tailing mix preparation.	39
2.21	End repair and A-tailing thermocycler programme.	39

List of tables

2.22	Adaptor ligation mix.	39
2.23	10x Genomics® sample index PCR mix.	40
2.24	10x Genomics® Chromium™ sample index PCR thermocycler protocol.	40
3.1	Primer sequences used in the Smart-seq2 protocol	54
3.2	Nextera® XT DNA sample preparation kit primers	56
3.3	Anatomy of <i>fastq</i> file	59
3.4	Description of QC parameters	61
3.5	Parameters set out in linkage equation from Adachi [2017]. n_i, n_j, n_k are numbers of items in each cluster.	68
3.6	Collation of evidence for assigning cell types to data-driven cluster identification. Table and references reproduced from that collated by Victoria Moignard.	70
3.7	Correlation of diffusion components with library size	94
3.8	Genes dynamic profiles clustered using dynamic time warping distances and fuzzy clustering	105
3.9	GO Biological process enrichment analysis using enrichR for endothelial subcluster	115
3.10	Summary of <i>Tal</i> ^{-/-} and <i>Tal</i> ^{+/+} experiment	127
3.11	Random forest assigned clusters to <i>Tal</i> dataset.	130
3.12	Overview of embryos harvested from <i>T</i> ^{nEGFP-CreERT2/+} and <i>T</i> ^{+/+} cross	136
3.13	Staining protocol including antibody, fluorochrome and the dilutions used. LP, long pass.	138
3.14	Embryo 2 FACS table indicating numbers of cells in each gate using the strategy shown in fig. 3.57 for a GFP -ve <i>T</i> ^{+/+} mouse embryo.	140
3.15	Embryo 10. FACS table indicating numbers of cells in each gate using the strategy shown in fig. 3.57	142
3.16	Summary of the numbers of cells from the different embryos in the <i>Brachyury</i> experiment	142
4.1	Approximate expected multiplet rates at different cell loading concentrations for the first generation 10x Genomics® Chromium™ protocol.	162
4.2	Cell concentrations from each embryo after re-suspending dissociated cells in 0.04% BSA	164
4.3	Sample indices used in this experiment and their corresponding 8 bp sequences	164
4.4	Summary of 10x Genomics® Chromium™ annotated cell clusters, final post-QC cell numbers and identified marker genes	171

4.5	Genes identified as differentially expressed between the 12 primordial germ cells and all other cells	175
4.6	Each cluster received equivalent contributions from the different embryos except for the blood and visceral endoderm clusters.	176
4.7	Table of the number of cells allocated to each of the endodermal sub-clusters and the colours used.	178
4.8	Top 50 genes more highly expressed in the NMP cluster relative to other mesodermal clusters	185
4.9	Differentially expressed transcription factors between the two cardiac clusters	190
4.10	Top 50 transcription factors most significant positively correlated to <i>Nkx2-5</i> , <i>Smarcd3</i> , <i>Gata4</i> , <i>Tbx5</i> and <i>Tbx20</i>	194
4.11	Constituent parts of the 10x Genomics® Chromium™ library molecule and their sequence lengths in base pairs (bp).	196
5.1	An excerpt from the iris data showing the features collected in the different species [Fisher, 1936].	220
5.2	Average (mean) features for the 3 species used as data points to project onto the preexisting dataset on fig. 5.4.	220
5.3	Table illustrating the essence of the autoregressive model	231

Nomenclature

Acronyms / Abbreviations

4SP 4 paired somites embryo stage

AGM Aorta, gonad, mesonephros

ASCII American Standard Code for Information Interchange

bcl Binary base call file

BSA Bovine serum albumin

CMP Common myeloid progenitors

CNV Copy number variant

CyTOF Cytometry by time of flight

DAPI 4',6-diamido-2-phenylindole

DGE Digital gene expression

DNA De-oxyribose nucleic acid

cDNA complementary DNA

dNTP Any of the four deoxy-nucleotides

dpt Diffusion pseudotime

DrL Also known as VxOrd - Directed graph layout

Nomenclature

dtw Dynamic time warping

ECMO Extra-corporeal membrane oxygenator

EGFP Enhanced green fluorescent protein

EMP Erythroid myeloid precursors

EMT Epithelial to mesenchyme transition

ERCC – 92 External RNA Controls Consortium, 92

FACS Fluorescence activated cell sorting

FBS Foetal bovine serum

FLP Flippase

FPKM Fragment per kilobase of transcript per million mapped reads

FRT Flippase recognition target

GMP Granulocyte-monocyte progenitors

GO Gene Ontology

GPU Graphics processing unit

GTF Gene Transfer Format

HF Head fold embryo stage

HSPC Haematopoietic stem and progenitor cells

iPS Induced pluripotent stem cells

kb Kilobases

Line – 1 Long interspersed element type 1, aka L1

LMPP Lymphoid multipotent progenitors

lncRNA Long non-coding RNAs

<i>LPM</i>	Lateral plate mesoderm
<i>LSTM</i>	Long short-term memory
<i>LT – HSC</i>	Long-term haematopoietic stem cells
<i>LTR</i>	Long terminal repeat
<i>MEP</i>	Megakaryocyte-erythroid progenitors
<i>MPP</i>	Multipotent progenitors
<i>NP</i>	Neural plate embryo stage
<i>OCT</i>	Optimal Cutting Temperature
<i>PBS</i>	Phosphate buffered saline
<i>PBS</i>	Phosphate buffered saline
<i>PCR</i>	Polymerase chain reaction
<i>qPCR</i>	Quantitative polymerase chain reaction
<i>PS</i>	Primitive streak embryo stage
<i>QC</i>	Quality control
<i>RNA</i>	Ribose nucleic acid
<i>mRNA</i>	messenger RNA
<i>RT</i>	Reverse transcriptase
<i>SBS</i>	Sequencing by synthesis
<i>SNV</i>	Single nucleotide variant
<i>SPRI</i>	Solid phase reversible immobilisation
<i>SSC</i>	Side scatter
<i>STR</i>	Short tandem repeat, aka microsatellite

Nomenclature

TPKM Transcripts per kilobase per million mapped reads - some papers refer to this as TPM

TPM Transcripts per million

TSS Transcription start site

UMI Unique Molecular Identifier

WHO World Health Organisation

WT wild type

Papers

* Indicates equal contribution

1: Scialdone A*, Tanaka Y*, Jawaid W*, Moignard V*, Wilson NK, Macaulay IC, Marioni JC, Göttgens B. **Resolving early mesoderm diversification through single-cell expression profiling.** Nature 2016 Jul 14;535(7611):289-293.

2: Ibarra-Soria X*, Jawaid W*, Pijuan-Sala B, Ladopoulos V, Scialdone A, Jörg DJ, Tyser RCV, Calero-Nieto FJ, Mulas C, Nichols J, Vallier L, Srinivas S, Simons BD, Göttgens B, Marioni JC. **Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation.** Nat. Cell Biol. 2018 Feb;20(2):127-134.

3: Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, Reik W, Srinivas S, Simons BD, Nichols J, Marioni JC, Göttgens B. **A single-cell molecular map of mouse gastrulation and early organogenesis.** Nature 2019 Feb;566(7745):490-495

4: Edri S, Hayward P, Jawaid W, Martinez-Arias A. **Emergence of a node-like population within an in vitro derived Neural Mesodermal Progenitors (NMPs) population.** Development 2019 Jun 24;146(12)

5: Sainz de Aja J, Menchero S, Rollan I, Barral A, Jawaid W, Cossio I, Alvarez A, Carreño-Tarragona G, Badia-Careaga C, Tiana M, Nichols J, Göttgens B, Isern J, Manzanares M. **The pluripotency factor NANOG controls primitive hematopoiesis and directly regulates Tal1.** EMBO J. 2019 Apr 01;38(7)

6: Belluschi S, Calderbank EF, Ciaurro V, Pijuan-Sala B, Santoro A, Mende N, Diamanti E, Sham KYC, Wang X, Lau WWY, Jawaid W, Göttgens B, Laurenti E. **Myelo-lymphoid lineage restriction occurs in the human haematopoietic stem cell compartment before lymphoid-primed multipotent progenitors.** Nat. Commun. 2018 Oct 5;9(1):4100.

Papers

7: Glass LL, Calero-Nieto FJ, Jawaid W, Larraufie P, Kay RG, Göttgens B, Reimann F, Gribble FM. **Single-cell RNA-sequencing reveals a distinct population of proglucagon-expressing cells specific to the mouse upper small intestine.** Mol. Metab. 2017 Oct;6(10):1296-1303.

8: Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa SI, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B. **Decoding the regulatory network of early blood development from single-cell gene expression measurements.** Nat. Biotechnol. 2015 Mar;33(3):269-276.

Software

All software is freely available.

Rtsne R implementation of tSNE adapted to allow new datapoints to be added to a pre-existing map. Available at <https://github.com/wjawaid/Rtsne>.

roots Reconstructing ordered ontogenic trajectories. Implementation of algorithms to allow construction and visualisation of ontogenic trajectories. Available through CRAN at <https://cran.r-project.org/package=roots> or GitHub at <https://github.com/wjawaid/roots>.

bglab Berthold Göttgens laboratory single cell analysis toolkit. Available at <https://github.com/wjawaid/bglab>.

gpr Gaussian process regression. Allows smoothing of gene expression profiles along selected ontogenic trajectories. Available at <https://github.com/wjawaid/gpr>.

gatepoints Gatepoints. Allows graphical user supervised selection of points of interest by drawing arbitrary complex shapes enclosing the points. Available through CRAN at <https://cran.r-project.org/package=gatepoints> or GitHub at <https://github.com/wjawaid/gatepoints>.

enrichR R wrapper for accessing the Enrichr database. Available through CRAN at <https://cran.r-project.org/package=enrichR> or GitHub at <https://github.com/wjawaid/enrichR>.

Chapter 1

Introduction & Aims

1.1 Motivation and Clinical relevance

1.1.1 Congenital Anomalies

Congenital anomalies contribute substantially to the global burden of neonatal morbidity and mortality. The World Health Organisation (WHO) estimates 276,000 neonatal deaths per annum from congenital lesions¹. Considering that many pregnancies may be non-viable without specialist antenatal care and even where such specialist care is available parents may elect for termination this is likely to be an underestimate having not accounted for 'hidden mortality' [Brownlee et al., 2009; Harrison et al., 1978]. Though arguably the aetiology of structural congenital anomalies for example oesophageal atresia, diaphragmatic hernia, chest or cardiac lesions and migrational anomalies such as Hirschsprung disease may not be wholly genetic and have environmental underpinnings, an understanding of normal biological development is likely to provide informative insights into prevention and treatment.

Recent decades have witnessed significant improvements in neonatal mortality from congenital defects but in severe cases, particularly where there is inadequate tissue for surgical reconstruction to achieve normal anatomy and physiology, morbidity is substantial [Jawaid et al., 2012, 2013]. For example the palliative Fontan circulation in children with physiologically a single ventricle allows for them to survive and to some extent have a relatively normal lifestyle but these children are chronically hypoxic and exercise tolerance and overall survival remain limited and sub-optimal. Patients with a variety of other conditions have similar sub-optimal outcomes including children with short bowel syndrome consequent to

¹<http://www.who.int/mediacentre/factsheets/fs370/en/>

Introduction & Aims

vanishing gastroschisis, intestinal atresias, malrotation with midgut volvulus or total intestinal aganglionosis. Current supportive strategies rely on mechanical devices such as ventilators, extra-corporeal membrane oxygenation devices (ECMO), heamofiltration/dialysis or parenteral alimentation to temporarily or in some cases to provide long-term support for inadequate organ function. These devices carry substantial risks, are often unwieldy (fig. 1.1) and require constant monitoring and adjusting. An example in point is the comparison between long-term dialysis and kidney transplantation. With the former patients may have to make regular weekly/bi-weekly visits to their local dialysis centre and be on a strict fluid restriction protocol, a transplant despite its risks and the need for life-long immunosuppression, can in such cases be life transforming. Transplantation cannot be offered to all patients in whom it is indicated because supply of good quality cadaveric organs particularly for paediatric use where the organ must survive for a lifetime, is severely limited.



Fig. 1.1 Contemporary neonatal intensive supportive care. Showing a neonate with pre-operative congenital diaphragmatic hernia and pulmonary hypertension requiring ventilatory support with inspired nitric oxide and venous-arterial ECMO for oxygenation and circulatory support. Various infusions are running for nutrition and to modulate pressures in the systemic and pulmonary circulations to maintain perfusion, achieve adequate oxygenation and prevent acidosis.

1.1.2 Non-communicable diseases

Noncommunicable diseases (NCDs) are estimated to account for 38 million deaths/annum and almost three quarters are reported to occur in middle or low income countries². NCDs include diseases such as obesity, diabetes, ischaemic heart disease and cancer. These diseases clearly have an environmental basis as reflected by the observation that higher prevalences of these conditions in developing countries have coincided with increasing affluence and adoption of high risk behaviours such as tobacco use, alcohol misuse, sedentary lifestyles and unhealthy diets. But accumulating evidence suggests that the intra-uterine environment can modulate an individual's risks for developing such conditions, possibly through durable epigenetic modifications [Battista et al., 2002; Dabelea et al., 2000, 2008; Power and Jefferis, 2002; Siklenka et al., 2015; Sun et al., 2013; Yura et al., 2005; Zohdi et al., 2014]. Furthermore evidence suggests that severely stressful events experienced by a population, such as famine, can modify risks for complex traits in second generation offspring, indicating that some epigenetic changes may endure, withstanding the genome wide chromatin 'resetting' imposed upon germ cells and the developing ovum [Kaati et al., 2002; Radford et al., 2014; Veenendaal et al., 2013].

1.1.3 Developmental Biology

Barring a few notable exceptions for example osteoclasts, giant multinucleated cells, histiocytes, germ, muscle and hepatic cells, in general cells have identical genomic (DNA) content in the same abundance. Germ cells are unique in this context in that meiosis allows genes to be shuffled between paternal and maternal chromosomes and results in a haploid cell. Other cells can also shuffle specific portions of their DNA for example B-cells of the immune system can recombine their so-called *V* and *J* regions in the case of light chains and *V*, *D* and *J* regions for the heavy chains to generate antibodies. *V(D)J* recombinase, the enzyme responsible for this is imprecise and can do so in several ways allowing a large repertoire of B-cell clones to be generated. The clones are subsequently selected for to prevent self-antigen recognition and autoimmunity, but can recognise a vast variety of invading organisms, viral antigens displayed on infected cells, cancerous cells and abnormal cells that may have the potential for malignant transformation.

In general, noting the occasional exception as described above and in the case of mosaicism, cells in an individual contain identical DNA, generated from a single fertilised ovum, that by cell division, symmetric and asymmetric establishes a multicellular organism containing in

²<http://www.who.int/mediacentre/factsheets/fs355/en/>

Introduction & Aims

many cases hundreds of different cell types precisely organised into functional units. The process by which multipotent progenitor cells gradually diversify is termed differentiation and has been described by Waddington as an abstract landscape, along which a cell travels altering its gene expression programme until it reaches a stable point, notionally synonymous to its terminally differentiated cell type.

A detailed and comprehensive understanding of developmental biology particularly in humans has the potential to provide clinically significant insights in to the aetiology and pathogenesis of not only congenital anomalies but also adult-onset diseases described in section 1.1.2. The recommendation for folic acid supplementation in women planning a pregnancy to reduce the risk of spina bifida stands as a paradigm for preventative strategies in birth defects.

Birth defects often co-occur as part of a syndrome or in well described *sequences* of associated anomalies. For example oesophageal atresia with or without tracheo-oesophageal fistula co-occurs with vertebral anomalies, anorectal malformations, cardiac defects, renal anomalies and limb deformities (VACTERL) more frequently than would be statistically expected. This is not unique in fact there are many other examples including the CHARGE sequence (Coloboma, heart defects, anorectal malformations, intellectual disability, genito-urinary anomalies and ear deformities), Pierre-Robin sequence, velocardial facial syndrome and various polyploidy syndromes e.g. Down's syndrome. Despite the cause of Down's syndrome (trisomy 21) being known, penetrance of the described defects is highly variable. A lack of understanding of normal development conceals the underlying mechanisms responsible for these associations. Some anomalies may not display an immediate clinically evident phenotype, particularly certain cardiac, renal or spinal cord anomalies therefore screening in selected groups is indicated as population wide screening is impractical for logistic and economic reasons. Such screening can enable early initiation of preventative strategies. An understanding of normal development beyond our current observational approach of described sequences can help with defining such populations and designing screening protocols.

Accurately defined clinical phenotypes combined with massive parallel sequencing genomic analysis can provide useful insights into critical genes and signalling pathways involved in prenatal development, causes of birth defects and reveal potential treatment strategies. Mosaicism appears to be an important feature in some birth defects, often with mutations that would be lethal in the non-mosaic state. A mosaic somatic non-synonymous single nucleotide mutation (c.548G→A, p.Arg183Gln) in GNAQ was found in 23 out of 26 patients with Sturge-Weber [Shirley et al., 2013], mosaicism with a somatic activating mutation (c.49G→A, p.Glu17Lys) in the oncogene AKT1 was found in 26 out of 29 patients with Proteus syndrome [Lindhurst et al., 2011] and all 6 patients with CLOVES sequence were

1.1 Motivation and Clinical relevance

found to be mosaic for activating mutations in PIK3CA [Kurek et al., 2012]. Levels of mosaicism varied between tissues but mutant allele frequency in affected tissues varied from 1 to 18% for GNAQ, 1 to 50% for AKT1 and 8 to 30% for PIK3CA. Lindhurst et al. [2011] found individuals usually had the same mutant allele in all affected tissues even those from different germ layers suggesting that the mutation occurred early in development to effect several cell lineages. Notably these findings suggest molecular pathways that may provide ‘drugable’ therapeutic targets.

A knowledge of normal development can clearly influence clinical decision but the examples above suggest also that the reverse is true: clinical findings can reveal potentially useful insights into normal human development. In fact for understanding normal human development we often have to work in reverse by describing a phenotype and finding a cause as opposed to perturbing normal development and then observing the resultant phenotype. New technologies such as massive parallel sequencing now offer powerful tools to work in either direction.

Normal development in complex organisms especially mammals is extremely challenging to understand due to the diversity of cell types, complexity and redundancies in signalling cascades and the complicated morphogenetic movements that can lead to a cell migrating and being subject to not only alternate paracrine morphogen gradients but also alterations in direct cell to cell signalling. Combining this with the difficulties in real-time observation of development in placental eutherians much work has been done on alternate model organisms such as *Drosophila melanogaster*, *Xenopus*, *Danio rerio* and *Caenorhabditis elegans*.

1.1.4 Stem Cell Biology

The diverse cell types in complex multicellular organisms arise not from differences in genomic material between the cell types nor do cells, in general, gradually lose genomic information critical to normal development as they specialise through differentiation, this was elegantly demonstrated in *Xenopus laevis* by Gurdon et al. [1975]. This is neither specific to the organism nor cell type as has been confirmed in more contemporary experiments where cloned mammals have been produced by nuclear transfer from a variety of differentiated cells into enucleated eggs [Campbell et al., 1996; Cibelli, 2007; Hochedlinger and Jaenisch, 2002; Wakayama et al., 1998].

These observations revealed that the nuclear chromatin of terminally differentiated cells has the plasticity to be re-organised so that it could re-run the developmental programme. Takahashi and Yamanaka [2006] went further and showed that by over-expressing the four

Introduction & Aims

transcription factors, *Oct3/4*, *Sox2*, *c-Myc* and *Klf4* in murine adult fibroblasts, they were able to generate induced pluripotent cells (iPS) capable of producing teratomas in immunodeficient mice and of contributing to all tissues of the developing mouse when injected into blastocysts. These cells behaved like embryonic stem cells and showed that it was not necessary to hijack the machinery and cytoplasm from an ovum to ‘reprogramme’ the nucleus of a terminally differentiated fibroblast.

In recent years it has become clear that ‘human embryonic stem cells’ and murine embryonic stem cells (ESCs) exist in divergent states of pluripotency [Brons et al., 2007; Nichols and Smith, 2011]. Human ESCs appear to behave more like murine epiblast stem cells (EpiSCs). It is difficult to expand them from a single clone and mouse EpiSCs, derived from the post-implantation epiblast using protocols similar to those used to generate human ESCs, do not readily contribute to chimaeras. In contrast, mouse ESCs can contribute to all cell lineages including the germ line and are therefore defined as being in a state of naïve pluripotency. An important feature of these naïve pluripotent cells is to contribute to tissues of all three germ layers and to form teratomas, tumours containing disordered tissues from all three germ layers i.e. ectoderm, endoderm and mesoderm. In fact sacrococcygeal teratomas, tumours with varying malignant potential that can be diagnosed on antenatal sonography are hypothesized to arise from persistent epiblast cells that failed to correctly migrate through the primitive streak at gastrulation [Fadler and Askin, 2008; Moore et al., 2015; Solari et al., 2011].

The discovery of iPS cells has given a significant boost to the field of regenerative medicine. The immune system will reject any implant detected as foreign leading to not only failure and necrosis of the implanted tissue but also in some cases a severe systemic response. Availability of iPS cells potentially opens the possibility of generating replacement tissues genetically identical to the host, abrogating the risk of rejection, the need for pharmaceutical immunosuppression, the risks of live organ donation and associated ethical concerns and the shortage of cadaveric organs as discussed earlier in section 1.1.1.

Other types of stem cells for example mesenchymal stem cells have been used in clinical practice. Macchiarini et al. [2008] seeded a de-cellularised cadaveric trachea with epithelial cells and mesenchymal stem cell derived chondrocytes and transplanted the trachea into a 30 year old patient with bronchomalacia. They reported the graft remained patent, well vascularised and completely epithelialised with normal ciliary function and mucous clearance at 12 months [Gonfiotti et al., 2014]. Since then, there has been concern with regard to the veracity of the claims and the lead author has been accused of scientific misconduct [Vogel, 2015].

More significantly a 70 year Japanese woman has had an autologous iPS cell derived sheet of retinal pigment epithelium transplanted into the eye for age-related macular degeneration [Cyranoski, 2014]. Despite the environment within the eye enjoying immune privilege, transplantation can activate microglia and the use of autologous cells may improve graft survival [Xian and Huang, 2015]. Potentially more practical and useful in the near future is the use of iPS cells to model disease and test effects both desirable and undesirable of pharmaceuticals, *in vitro*.

The application of these techniques requires a good working knowledge of normal development to produce the correct cell types, in the required proportions and with the desired structural and spatial configuration. Pluripotent cells (iPSCs and ESCs) are maintained and expanded under specific culture conditions to maintain the pluripotent state and prevent differentiation, the process by which cells become increasingly fate restricted and eventually adopt a mature stable functional state. Directing differentiation of iPS cells into mature functional cells of the required type efficiently, without contamination from other cell types is challenging but understanding normal development can potentially provide valuable insights.

1.2 Embryology - A brief overview

This overview of embryonic development focuses in most part on mouse embryology.

1.2.1 Pre-implantation

Fertilisation triggers a cascade of events that subject to error free progression, result in the birth of an offspring. The successful haploid spermatozoon will have undergone *capacitation* to pass through the *corona radiata*, *acrosomal activation* to penetrate the *zona pellucida* and and fusion with the oocyte to eventually release its pronucleus into the ova's cytoplasm. The ova responds by releasing cortical lysosomal enzymes making its cell membrane impenetrable to other sperm and altering the structure and composition of the *zona pellucida* to prevent multiple sperm binding and penetration. The ovum will resume the second meiotic division producing the second polar body and the definitive oocyte containing the female pronucleus. There is metabolic reactivation, both pronuclei duplicate their DNA and become organised on the spindle to undergo normal mitotic division resulting in two diploid cells with the requisite set of chromosomes, at about 1 day post conception (E1.5³).

³E1.5 is notation for embryonic day 1.5, 1.5 days since ovulation

Introduction & Aims

The two-cell zygote undergoes multiple rounds of mitosis confined within the zona pellucida, increasing in cell number but each cell reducing in size with every cleavage. These cells now called blastomeres, are loosely clumped until the eight-cell stage (3rd cleavage, E2.5) when they undergo *compaction* by maximising their contact with neighbours and forming tight junctions. Compaction segregates the cells into an inner group, the *inner cell mass*, communicating extensively by gap junctions and an outer group. The inner cells will go on to form the embryo proper while the outer cells generate the *trophoblast*, which forms the placenta. Around E3.0, the cells divide once again to form the 16-cell *morula* and cells of the trophoblast secrete fluid into intercellular spaces between the inner cells. During a process termed *cavitation* the fluid coalesces into a single cavity, the *blastocoele*, marking the transition of the morula to a *blastocyst* at about E4.0. The blastocyst hatches from the zona pellucida ready for implantation into the prepared *endometrium* lining the uterine cavity (fig. 1.2). The inner cell mass becomes organised into two readily identifiable epithelial layers, the epiblast dorsally and the hypoblast⁴ ventrally. Pre-implantation development of the embryo is summarised in fig. 1.3.



Fig. 1.2 Mouse blastocyst hatching from zona pellucida [Valley et al., 2010].

1.2.2 Post-implantation

Development continues post implantation as the polar trophectoderm, the epiblast and the primitive endoderm continue to proliferate. By E.5 a *proamniotic cavity* develops within the epiblast, a similar cavity develops within the extraembryonic ectoderm and the two fuse. At around the same time the proximal region of the extraembryonic ectoderm gives rise to the *ectoplacental cone*, which rapidly divides and contributes to the placenta.

The characteristic egg cylinder with the *inverted germ layers* is peculiar to rodent development. The early rodent embryo macroscopically appears to be drastically different from equivalent stage embryonic development in humans and other model organisms but crucially important structural relations are preserved.

⁴Also known as the primitive endoderm and will go on to form the *yolk sac*



Redacted

Fig. 1.3 Pre-implantation murine embryo. <http://stemcells.nih.gov/info/scireport/pages/appendixA.aspx>

Primitive endoderm segregates into the visceral endoderm lining the ventral surface of the epiblast and the parietal endoderm lining the opposing surface of the blastocele, leading to the formation of Reichert's membrane (equivalent to Heuser's membrane in humans) between the parietal endoderm and the trophoblast.

Despite minimal contribution to the embryo proper, the primitive endoderm and extraembryonic ectoderm are instrumental in orchestrating embryo patterning e.g. anterior-posterior (AP) axis specification in the epiblast. The *distal visceral endoderm* (DVE) characterised by *Lefty1*⁺ *Gata6*⁺ cells appears in the distal pole of the egg cylinder. The DVE migrates anteriorly and cells of the *anterior visceral endoderm*, distinct from cells of the DVE develop and move anteriorly along the midline [Takaoka et al., 2011]. Both the AVE and DVE produce *Lefty1* and *Cer1* to antagonise posteriorising signals mediated by *Nodal*, *Gdf3* and *Cripto*, induced in the epiblast by the convertase enzymes *Furin* and *Pcsk6* produced in the extraembryonic ectoderm. This inhibition from the AVE restricts the development of the *primitive streak* to the posterior of the embryo, which marks the onset of the mass migration of the epiblast in a process called *gastrulation* setting the blue print for further development.

Introduction & Aims

Gastrulation is a fundamental process in the post-implantation mammalian embryo that lays out the body plan and by which the early embryo is transformed from a bilaminar to a trilaminar structure. In the E6.5 mouse embryo, onset of gastrulation is marked by the appearance of the primitive streak at the posterior extreme of the epiblast juxtaposed to the embryonic-extraembryonic junction fig. 1.4. Generation of the primitive streak is induced by *Bmp4* secreted by the extraembryonic ectoderm activating *Wnt3* regulatory pathways which in turn activates expression of the canonical mesoderm marker *Brachyury(T)*. Epiblast cells adjacent to the nascent streak undergo *epithelial to mesenchymal transition* (EMT), some migrate between the epiblast and the visceral layer of the primitive endoderm, forming embryonic mesoderm and others intercalate displacing the primitive endoderm laterally and forming the definitive endoderm. Not only are all three canonical germ layers generated from the epiblast but it also contributes cells to extraembryonic mesoderm including the yolk sac and amniotic ectoderm.

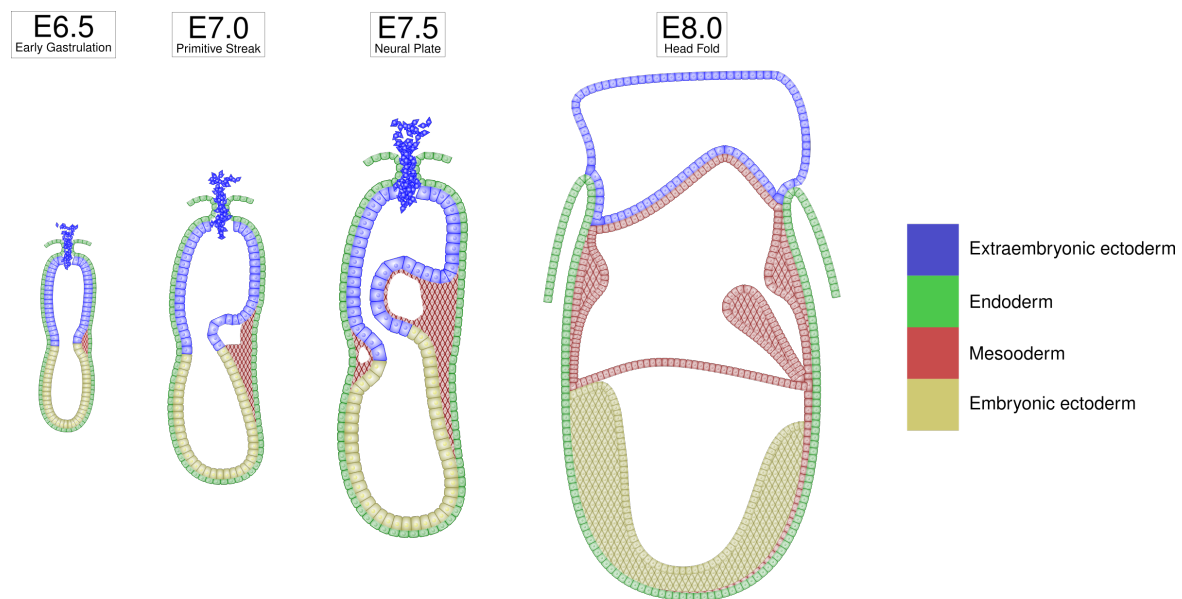


Fig. 1.4 Mid-sagittal sections through egg cylinders between E6.5 and E8.0. At E6.5 the anterior-posterior axis of the embryo has been determined. The primitive streak appears at the posterior extreme of the embryo at the junction with the extraembryonic ectoderm. The visceral endoderm is shown in green lining the ectoderm, the parietal endoderm and Reichert's membrane are not shown. The earliest cells to ingress through the primitive streak are extraembryonic mesoderm of the yolk sac. The amniotic cavity begins to form in the extraembryonic mesoderm forming an eccentric doughnut which finally fuses forming the amniotic cavity. The amniotic ectoderm is not shown. At the E8.0 stage the allantois is clearly visible at the posterior end of the embryo and blood islands can be seen in the yolk sac.

1.2 Embryology - A brief overview

EMT is the process by which cells at the streak lose cohesion with their epiblast neighbours and display a more fluid like behaviour to ingress into the potential space between the epiblast and hypoblast. *Fgf* signalling plays a central role in EMT, Ciruna and Rossant [2001] demonstrated that in *Fgfr1*^{-/-} mutants epiblast cells fail to undergo EMT. At gastrulation in the primitive streak *Fgfr1* coordinates EMT and mesoderm morphogenesis through up-regulation of *Snail* and down-regulation of *E-cad* a key component for maintaining cohesion in epithelia.

As gastrulation progresses the primitive streak extends anteriorly in the midline. Embryonic and extraembryonic mesoderm migrates around the ‘rim’ of the egg cylinder analagous to the annular movements observed in species where the germinal layers maintain a disc shape rather than a ‘cup’. As the extraembryonic mesodermal layer thickens greater in the posterior a doughnut shaped ring develops with an eccentric aperture and within the extraembryonic mesoderm the amniotic cavity develops as a ring, fig. 1.4. Eventually the aperture located eccentrically to the anterior closes forming the amniotic cavity, lined by the amniotic ectoderm derived from the epiblast. Concurrently the *allantois* develops in from the posterior mesoderm into the amniotic cavity and will eventually communicate directly with the trophoctoderm. The blood vessels will form the *umbilical vessels* and the proximal intra-embryonic part of the allantois communicates with the developing bladder. In humans the remnant of this is clearly visible on the interior surface of the anterior abdominal wall as the *median umbilical ligament*. On occasions this either does not completely regress and postnatally a *patent urachus* communicates with the anterior abdominal wall or it only partially regresses so that cystic remnants that communicate with the bladder remain and these *urachal cysts* have a propensity to get infected.

The intraembryonic mesoderm becomes organised, sandwiched between the developing ectoderm and endoderm. Centrally the axial mesoderm forms the *notochord* ventral to the developing *neural tube*. Immediately lateral is the *paraxial mesoderm*, further laterally the intermediate mesoderm and then the *lateral plate mesoderm* which is divided by the intraembryonic coelom into *splanchnic* and *somatic mesoderm*.

The mesoderm will also undergo segmentation in the anterior-posterior axis, forming anteriorly the pharyngeal arches (from the head mesoderm) and more caudally the *somites* will eventually form from the paraxial mesoderm. At E9.5 after *turning*, the process by which the inverted embryo ‘flips’ inside-out, the pharyngeal arches, ventral to the developing *hindbrain* and caudal to the *maxillary arch* and *stomodeum*, will be clearly evident. Each pharyngeal arch receives contributions from the *neural crest* and contains cartilaginous, neural and arterial

Introduction & Aims

components supported by a mesodermal core derived from paraxial and splanchnic lateral plate mesoderm.

This completes the discussion for the early period of embryological development concerned with this current project. Specifics of cardiovascular and haematopoietic ontogeny will now be described very briefly.

1.2.3 Embryology of Cardiac Development

The developing embryo soon outgrows the ability of *diffusion* alone to provide adequate oxygen and nutrient delivery. This remains highly problematic for those attempting to grow organs for the purposes of regenerative medicine but evolution has solved this restriction with the development of the cardiovascular and haematopoietic systems.

The heart develops from two separate mesodermal populations, the *first* (FHF) and *second heart fields* (SHF). FHF cells, derived from the lateral splanchnic mesoderm differentiate earlier than the SHF and predominantly populate the *left ventricle*. In contrast, SHF cells though contiguous with FHF cells are more medial and in fact are a subset of the *pharyngeal mesoderm*. Pharyngeal mesoderm itself is a subset of head mesoderm as described above and undergoes gastrulation prior to trunk mesoderm. Furthermore the paraxial and lateral plate mesoderm merge to produce the mesodermal core of the prospective *pharyngeal arches*. As the heart tube loops the SHF cells contribute to the *outflow tract*, *atria* and the *right ventricle*.

It has been suggested that both FHF and SHF originate from a common progenitor and *Mesp1* is the earliest known marker of cardiac progenitors [Lescroart et al., 2014; Saga et al., 1999]. The timing of lineage bifurcation is not known but Lescroart et al. [2014] have suggested a revised model of specification for cardiac progenitors. They propose that two temporally distinct populations of cells express *Mesp1* independently and are therefore lineage segregated prior to *Mesp1* expression. Moreover utilising a formal statistical analytic strategy combined with single cell labelling they were able to show that FHF progenitors are unipotent and that SHF progenitors in contrast may be unipotent or bipotent, fig. 1.5.

Epiblast cells adjacent to the primitive streak lose *E-cad* adhesion, undergo EMT and give rise to nascent *E-cad*⁻ mesoderm which is multipotent and can differentiate to generate blood, endothelium, smooth muscle cells and cardiomyocytes, fig. 1.6. The FHF and SHF clearly have a common progenitor but no marker of this population has been found to date.

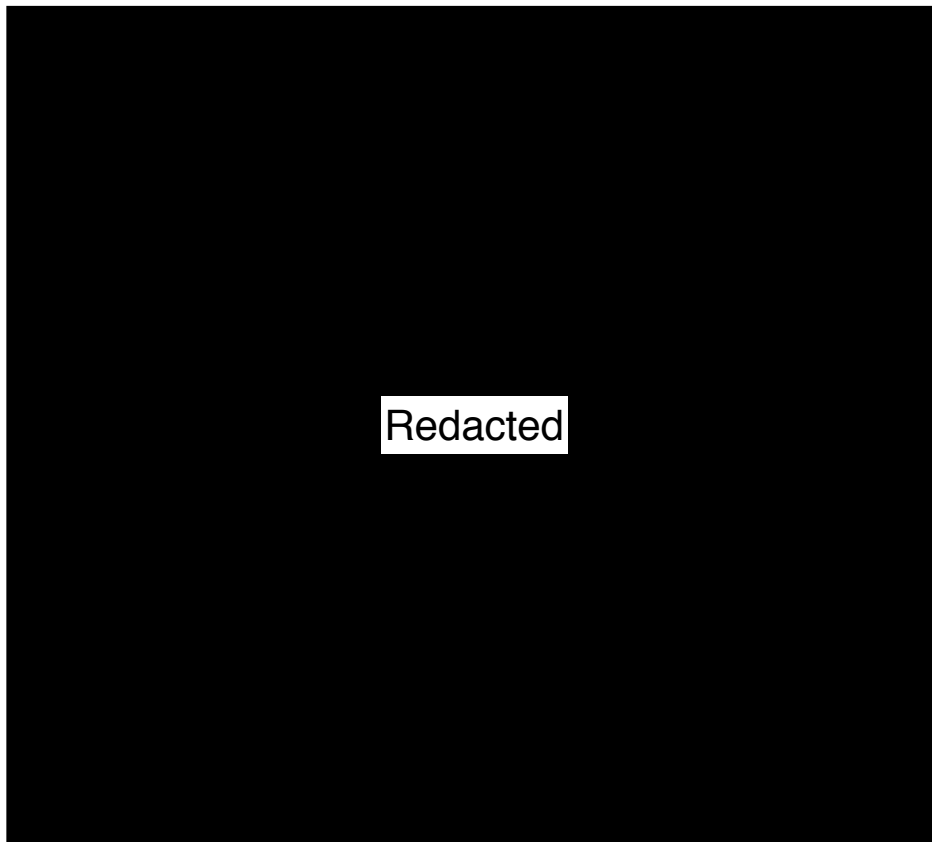


Fig. 1.5 From Lescroart et al. [2014]. Summarises their revised model. Unipotent FHF *Mesp1*⁺ cells can generate either CMs or ECs (red). SHF *Mesp1*⁺ cells in contrast can be unipotent or bipotent (green). CM - cardiomyocyte, EPDC - epicardial derived cell, EC - endothelial cell, SMC - Smooth muscle cell.

1.2.4 Developmental Haematology

Haematopoiesis is critical to the developing embryo and defects are lethal. The first blood cells originate from the blood islands of the yolk sac alongside organised endothelial structures at \sim E7.5, fig. 1.4. The primitive erythrocytes produced at this stage express an alternate β -chain of *haemaglobin* the β H1 with an altered oxygen binding curve compared to adult haemaglobin, to allow for oxygen transport at lower partial pressures in the embryo. Later the definitive yolk sac phase also produces *myeloid* cells and *Lyve1* has been proposed as a marker of multipotent progenitors in the yolk sac that can produce cells of the *myeloid* lineage.

Production subsequently moves to the aorta-gonadal-mesonephros region (AGM) and the haemogenic endothelium in the dorsal aorta from where definitive haematopoietic stem cells (HSCs) are generated and populate the foetal liver, spleen and eventually the late foetal and

Introduction & Aims

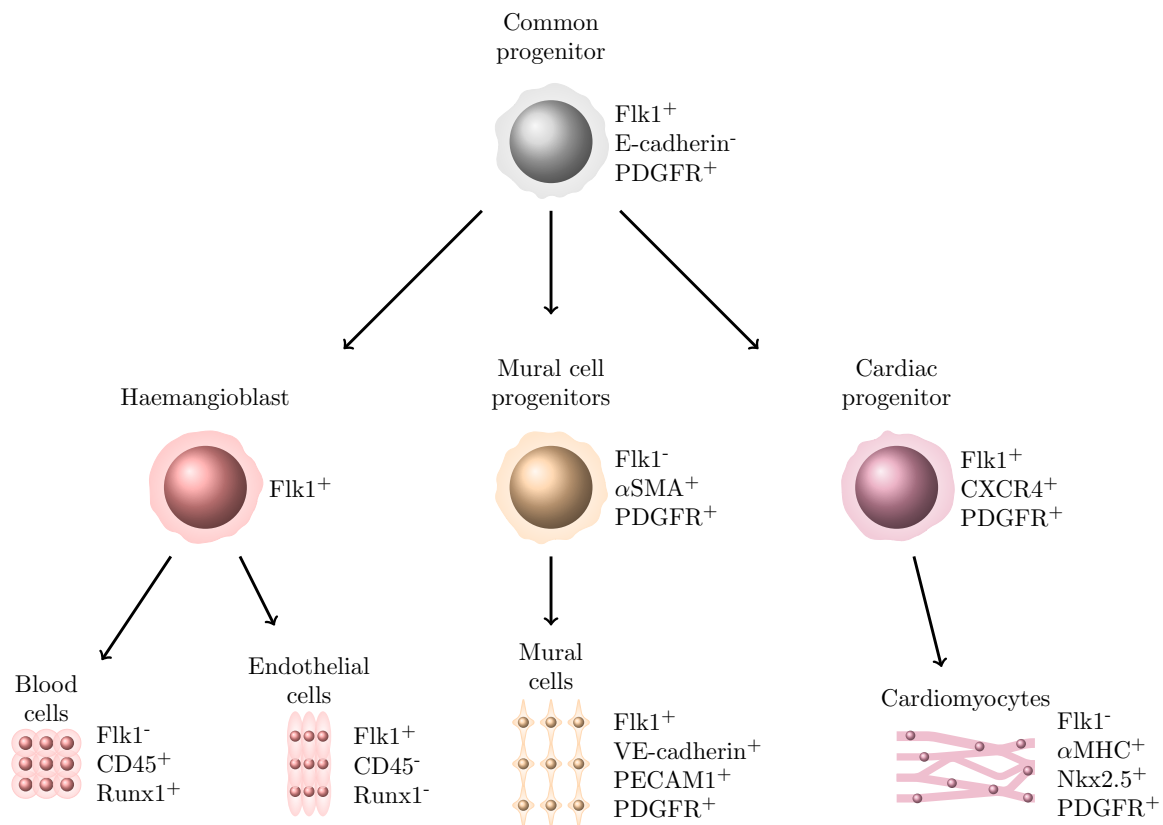


Fig. 1.6 Schematic of putative in-vivo bifurcation points in a developmental hierarchy of nascent mesoderm to haematopoietic, endothelial, smooth muscle and cardiac cells. The existence of an in-vivo haemangioblast as shown in this figure is controversial and the theory that blood cells are derived from haemogenic endothelium is now gaining traction.

postnatal bone marrow, fig. 1.7 [Ciau-Uitz et al., 2014; Dzierzak and Medvinsky, 1995; Medvinsky et al., 2011].

From E7.0 onwards in mouse all haematopoietic potential is confined to the $FLK1^+$ and $RUNX1^+$ mesoderm populations [Shalaby et al., 1997]. All haematogenic cells can therefore be separated from the remaining mesoderm using flow cytometry gating for $FLK1^+$ and $RUNX1^+$.

1.3 Experimental approaches in embryology

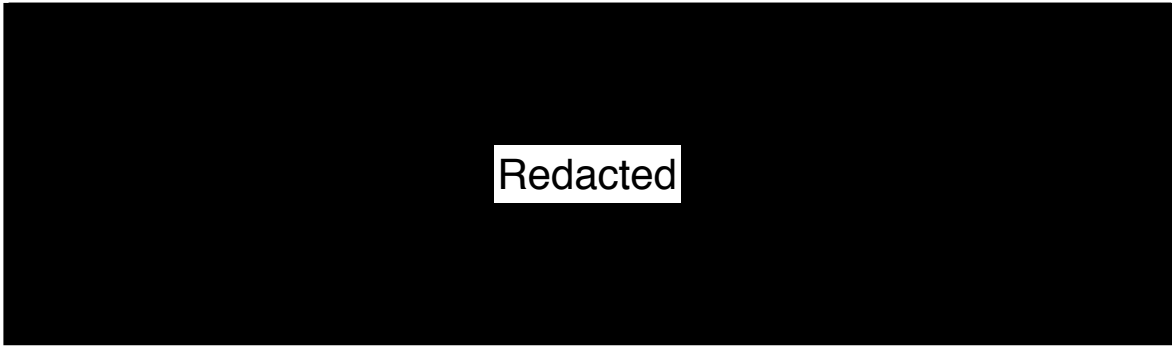


Fig. 1.7 The three distinct waves of haematopoiesis in mouse. YS, yolk sac; Ery, erythroid; Myel, myeloid; EMPs, erythromyeloid progenitors; FL, foetal liver; BM, bone marrow. Modified from Ciau-Uitz et al. [2014]

1.3 Experimental approaches in embryology

The development of the early microscope has been instrumental for scientific study in many fields of biology, in particular experimental embryology allowing it to move beyond its theological foundations. Charles Otis Whitman was amongst the earliest to generate a fate map by observing early leech embryos and describing the emergence of the germ layer from the 8-cell stage [Whitman, 1887]. Advances in imaging combined with novel methods of disrupting developmental processes and labelling individual cells and their progeny have led to progressively more detailed observations of normal and abnormal development in increasingly complex organisms. Clinical observations of human congenital abnormalities and correlation with known embryology provide additional insights into their aetiology and pathophysiology.

1.3.1 Model organisms

Model organisms have been adopted to reveal overarching themes and concepts common to many organisms for example gastrulation. This process by which the two primitive germ layers generate the three definitive germ layers has garnered much research attention due to its ubiquity albeit with some differences amongst complex multicellular organisms.

Research has often focused on species such as *Danio rerio*, *Caenorhabditis elegans*, *Xenopus laevis* and *Xenopus tropicalis* due to their small size and the transparent nature of their embryos allowing longitudinal time lapse imaging [Vogt, 1929]. Other organisms such as chick and quail though not transparent in their natural form are not confined to a uterus and simple experimental techniques can allow them to be cultured either ex-vivo and again imaged longitudinally for significant parts of development [Dieterlen-Lievre, 1975; Le Douarin, 1969, 1973; Rawles, 1948].

Introduction & Aims

The additional advantage that embryonic cells from quail (*Coturnix coturnix japonica*) can be transplanted into chick (*Gallus gallus*) embryos and distinguished from one another based on their chromatin patterns and Feulgen staining was recognised and exploited by Le Douarin [1969]. Later Le Douarin and Teillet [1973] using isotropic, isochronic transplant of quail neural tube and neural crest into chick host experiments show that neural crest between the 1st and 7th somites prior to the 13th somite embryo stage colonise the chick gut along its total length while cells caudal to the 28th somite, additionally provide some contribution to the post-umbilical gastrointestinal tract. These fate mapping experiments led to key observations on the development origins of ganglion cells which allow normal co-ordination of bowel peristalsis and their absence is pathognomic of Hirschsprung disease.

1.3.2 Reconstructing developmental lineage trees

Establishing lineage relationships between cells in an adult organism can uncover their ontogenic journeys possibly revealing key molecular pathways responsible for fate decisions. Evolutionary conservation of multicellular organism development allows insights gained from other organisms to be cautiously applied to human embryology. Experimental techniques include prospective and retrospective lineage tracing.

Prospective lineage tracing

Prospective methods in some cases also known as forward genetic, encompass techniques that apply some relatively indelible mark to a single cell or a clearly defined set of cells of interest at an early time point in an organism and then carefully examine their progeny at a later point in development.

The earliest methods involved solely detailed direct observation with an early example in leech [Whitman, 1887] and nearly a century later using a Nomarski microscope, the complete description of the lineage tree of *Caenorhabditis elegans* by Sulston et al. [1983]. Vital dyes, stains that mark cells without killing them, can either mark the cell membrane, cytoplasm or nucleus. Vogt [1929] an early pioneer implanted tiny agar chips of Nile Blue, a water-soluble dye, over cells of interest and labelled different regions of the 32 cell *Xenopus* blastula. Water soluble dyes though can diffuse and inadvertently mark nearby cells making it difficult to draw conclusions about lineages. Alternatives such as carbocyanine dyes and high molecular weight dextrans mitigate risks of diffusion but are technically more difficult to apply, requiring injection into cells of interest. The aforementioned techniques all suffer from the limitation that they are diluted at each cell division and can become undetectable in highly proliferative cells.

1.3 Experimental approaches in embryology

Transplantation of embryonic regions from pigmented strains into unpigmented strains of chicken confirmed that, as had already been shown in amphibians, pigmented cells were ontogenically derived from the neural crest [Rawles, 1948]. These transplantation experiments as with quail chick experiments described previously are technically highly demanding. Unlike lineage tracing using dyes there is no dilutional effect with transplantation but it does require the investigator to show to a certain degree that the behaviour of the engrafted cells is identical or at least similar to that of native cells. Cell behaviour when transplanted may be very different from that in the steady state and what is being assayed is cell fate potential within the transplanted tissue microenvironment rather than the fate of the cell within an otherwise undisturbed system [Hsu, 2015].

Genetic labelling of native cells as with transplantation leaves indelible marks that are not prone to dilution with proliferation and all progeny of a founder cell(s) will be marked. Recombinant DNA technologies using retroviral libraries with sparse infection strategies can incorporate a reporter gene such as β -galactosidase or green fluorescent protein along with barcodes and allow lineage tracing. These cells are otherwise genetically identical and though these exogenous sequences are integrated randomly and can potentially interfere with development they are much more similar and have greater fidelity than xenotransplants. These methods can be used in ex-vivo systems such as organotypic slice cultures or progenitor cultures [Woodworth et al., 2017]. Combining retroviral infections with transgenic lines expressing virus receptors only within specific cells can allow lineage tracing to be restricted to cells of interest.

An alternative method of inserting foreign genetic material into native cells is the use of transposons but they do not incorporate into the cells genome and again suffer from dilution at every generation. To overcome this limitation dual plasmids can be transfected, a donor plasmid containing the transgene of interest the researcher wishes to incorporate into the genome and a helper plasmid containing the transposase [Wu et al., 2006]. In cells that are transfected by both plasmids through either lipofection or electroporation, the transposase allows the transgene flanked by terminal repeats to be stably but randomly integrated into the genome. The transgene will then be propagated to all the cell's progeny while the helper plasmid with the transposase will be gradually diluted out.

Advances in genetic engineering technologies permit conditional genetic recombination in a cell specific or tissue specific manner. Two such technologies are the Cre-*loxP* and the FLP-*FRT* systems. Taking the Cre-*loxP* system as an example a mouse line expressing Cre recombinase under the control of some cell-specific promoter may be crossed with a reporter line that has a stop codon flanked on either side by *loxP* sites, followed by a reporter gene

Introduction & Aims

such as GFP knocked into a ubiquitously expressed locus such as the *Rosa26*. In the embryos from the cross all cells in which the particular promoter is active will express *Cre* that will excise the stop codon and the cells and their progeny will be marked by expression of the reporter gene.

Though specific to only cells that express a particular gene there is no temporal restriction as a gene may be expressed in the same lineage at different developmental time points. To achieve temporal control an inducible *Cre* such as *CreERT2* may be used that is expressed under the control of the promoter of interest but further requires administration of Tamoxifen (an oestrogen analogue) to allow it to translocate to the nucleus, remove the stop codon flanked by *loxP* sequences and constitutively express the reporter. A further advantage is that by administering low doses of Tamoxifen cells of interest may be labelled sparsely so that a less than specific promoter may be used where a specific promoter line has either not been generated or no promoter with the required specificity has been recognised.

Occasionally a cell type of interest will have the unique characteristic of two active promoters, while each promoter alone may be non-specific. In this case a dual system either adopting both the *Cre-loxP* and the *FLP-FRT* systems together or a split *Cre* may be used.

In the first instance the *Flippase* may be under the control of one promoter with a second promoter driving *Cre* expression but with an upstream stop codon flanked by *FRT*. The *Cre* would then remove a stop codon upstream of a reporter under control of a ubiquitously active promoter [Hsu, 2015]. In this way both promoters must be activated in the same cell or related cell concurrently or in sequence. A possible pitfall with this system is if a cell activates the first promoter but not the second and one of its progeny then later activates the second promoter then even with the first promoter inactivate the system will allow this cell and all its progeny to misleadingly express the reporter gene.

In the second instance each of the two promoters will control the expression of either the N-terminus of *Cre* (NH₂-*Cre*) or the C-terminus of *Cre* (*Cre*-COOH). Each fragment is inactive but spontaneous association of the two fragments generates the active *Cre*. This spontaneous association has low efficiency and an alternative version uses an intein peptide which enhances joining of the two fragments while is itself excised in a process that has been termed protein splicing [Anraku et al., 2008; Casanova et al., 2003; Wang et al., 2012].

The same *Cre-loxP* technology can be used with either incompatible *flox* sites or multiple fluorescent proteins to generate variable random recombinations. These allow cells to randomly express some combination of fluorescent proteins. Two such technologies include

1.3 Experimental approaches in embryology

the brainbow and confetti mice lineages and have been used to study neural and adult gastrointestinal lineages [Livet et al., 2007; Snippert et al., 2010].

Retrospective lineage tracing

Methods taking advantage of naturally occurring somatic mutations are a relatively recent innovation and depend on recognising the mutations that are shared between different cells in the developed organism and then generating a relatedness tree of the cells of interest based on maximal parsimony. There exist different genetic variants including retrotransposons, copy number variants (CNVs), single nucleotide variants (SNVs) and microsatellite (short tandem repeats, STRs) variants.

Retrotransposons can be grouped into two main categories long terminal repeats (LTR) and non-LTR retrotransposons. The LTR retrotransposons are remnants of ancient retroviruses that have incorporated into the human germline [Faulkner and Billon, 2018]. Non-LTR retrotransposons are sub-categorised into those that can mobilise autonomously and those that require other retrotransposons for activity. The long interspersed element 1 (LINE-1) are the only autonomous non-LTR retrotransposons recognised in humans and are considered to be the most active transposable elements. These have been particularly extensively studied in the brain and it has been proposed that they are de-repressed in somatic cells and remain active even in post-mitotic cells [Bodea et al., 2018; Faulkner and Billon, 2018].

Subchromosomal CNVs are a type of structural variant that are common in the human genome, they have been recognised to occur in somatic tissues and a small fraction are shared between cells indicating a common developmental ancestry and suggesting a further method of constructing lineage trees [Cai et al., 2014; Zarrei et al., 2015]. Several mechanisms have been postulated to give rise to CNVs and these mechanisms can be broadly classified into homologous or non-homologous recombinations. Homologous recombination events occur during meiosis due to mis-alignment of homologous chromosomes. Many non-homologous mechanisms have been suggested including polymerase slippage, non-homologous or micro-homology end-joining and break-fusion-bridge cycles.

SNVs are differences of individual bases on both strands at specific loci within the genome. Germ line SNVs are recognised to be major sources of disease causing variants and evolutionary mutations. SNVs can build up in cells during development so that a cell with an established SNV relative to the germ line genome will pass it on to all its progeny providing an almost indelible marker. Studies that have studied somatic SNVs describe differing frequencies from 100 to 1000 SNVs per cell [Behjati et al., 2014; Hazen et al., 2016; Lodato et al., 2015].

Introduction & Aims

Microsatellite variants are some of the most frequent somatic variants and are tandem repeats of short nucleotide sequences. Microsatellite loci represent regions of the genome that are highly prone to mutation in particular the number of tandem repeats. The extent of polymorphism between individuals at these loci is high and it has been postulated that even within a single organism somatic microsatellite variants can reconstruct its complete lineage tree [Frumkin et al., 2005; Takezaki and Nei, 2009].

Single cells contain only 6 pg to 7 pg of genomic DNA and whole genome amplification can allow genome wide identification of microsatellites (STRs) and CNVs [Vander Plaetsen et al., 2017]. Alternatively, more focused analysis specifically on a set of 81 microsatellite loci in mis-match repair deficient mice has allowed oocyte lineage to be resolved from bone marrow and ovarian cumulus cells [Reizel et al., 2012].

Lineage tracing experiments have been classified into prospective and retrospective but hybrid combination of these strategies exist that utilise in vivo gene editing combined with single cell genome sequencing. The CRISPR/Cas9 system has been exploited in several technologies including GESTALT [McKenna et al., 2016], scartrace [Junker et al., 2016] and others [Kester and Oudenaarden, 2018].

1.4 Single cell methods

Some of the methods described previously are single cell but lack high throughput in that they can only analyse at most a handful of genes within a confined region and these are mostly microscopy based. In simple organisms prospective methods can be used to visualise and document changes as they occur while in more complex organisms such as placentals only retrospective methods may be effectively utilised.

Higher throughput methods exist but are limited by resolution and have in the past been assays that require large numbers of cells to provide an adequate signal. These have included microarrays and bulk mRNA capture and sequencing methods. As techniques have become more refined assays have been developed that can use diminishing amounts of starting material. Earlier methods include qPCR in large numbers of single cells and measuring multiple genes but these have become quickly superseded by more advanced methods that are high throughput in cell number and being genome wide assays precluding the need to select genes of interest.

The idea of single cell resolved assays is highly powerful. As demonstrated in fig. 1.8 analysing cells in a bulk assay leads to an average measurement across multiple cells that is

impossible to deconvolute. An experimental intervention may make striking alterations to the signal from a rare cell type but this may never be discovered from a bulk assay due to the signal being swamped by lack of effect in the majority of cells sampled. The only way to overcome this would be to have some way of purifying the cells of interest and performing the intervention specifically on the selected pool but such prolonged processing carries risk of altering the cells RNA expression patterns. Single cell methods on the other hand can allow characterisation of individual cells and post-hoc assignment of an effect to specific cell types.

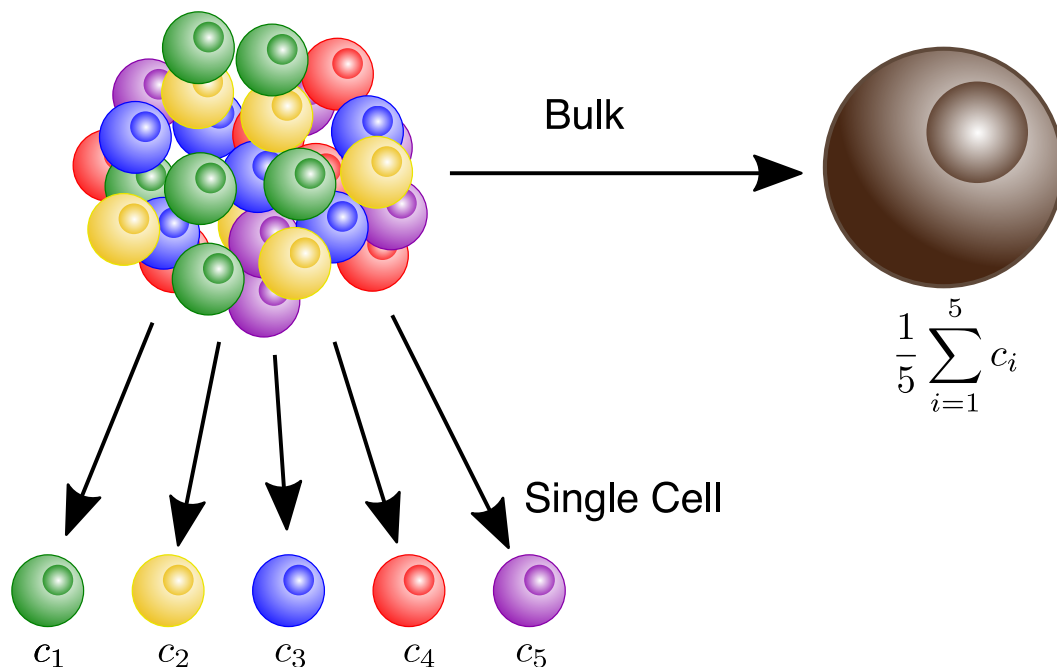


Fig. 1.8 Figure demonstrates the averaging effect of bulk sequencing as compared to single cell methods which can resolve the expression pattern of individual cells from which bulk data can be reconstructed. In contrast there is no practical way of inferring single cell expression patterns from bulk data.

1.5 Aims

The development of new technologies allowing application of '-omics' methods to single-cells paves the way to understanding cell biology at much higher resolutions than previously

Introduction & Aims

possible. One application is in elucidating the journey of a progenitor cell(s) as it becomes sequentially fate restricted until it finally reaches its destination cell type. Despite the advances in technology it is not possible to follow this journey across multiple genes/proteins over time in-vivo. Our technique sacrifices cells at various time points and we aim to computationally reconstruct this journey using tools currently available.

The overarching theme of this body of work is to develop methods to reconstruct ordered ontogenic trajectories through sequentially sampled cross-sectional data. The main focus here being using single-cell resolved, transcriptomic data collected during early mouse embryonic development. Where available this has been supplemented with limited, hand selected cell surface proteomic measurements.

The aims are summarised in fig. 1.9 and are:

1. Identify cell populations
2. Trace biologically plausible trajectories
3. Identify novel molecular pathways
4. Develop models that can faithfully simulate cell progression along trajectories

This method of lineage reconstruction best fits with retrospective lineage tracing. Dependent on the cell phenotype a subset of the exome will be sequenced but the sequence and the genome itself has not been used for lineage reconstruction. Instead lineages are traced based on the assumption that within a window of developmental time, cells with the most similar transcriptional signatures are related by lineage.

Pre-existing state of the art methods will be combined with novel computational methods and are described in the subsequent chapters. The first results chapter describes data from single cells harvested from mouse embryos at the gastrulation phase processed using the Smart-Seq2 protocol. The second chapter focuses on a later stage of development, early organogenesis, uses a higher throughput microfluidics commercial system 10x Genomics® Chromium™ to generate the dataset. Established and novel techniques are used to explore the data and uncover novel insights. The final results chapter describes and provides the mathematical and computational background for the novel methods that have been used in previous chapters and also focuses more on studying models that can simulate developmental trajectories.



Fig. 1.9 Graphical summary of the overarching aims of this thesis

Chapter 2

Methods

2.1 Mouse work

All procedures were performed in strict adherence to United Kingdom Home Office regulations (project licence 70/8406) [Scialdone et al., 2016].

For the gastrulation project (chapter 3) dissections were chiefly performed by Dr Yosuke Tanaka. Timed matings were set up between CD1 mice preferentially selected because they produce large litters. Embryos were staged according to the morphological criteria of Downs and Davies and classified broadly as early gastrulation, primitive streak, neural plate or head fold stage [Downs and Davies, 1993].

Mice were mated at the biomedical campus mouse facility and time of conception was assumed to be midnight on the day of plugging. Dams were sacrificed at the designated time points E6.5 (early gastrulation), E7.0 (primitive streak), E7.5 (neural plate) and E7.75 (head fold) and uteri harvested into ice cold PBS.

For the 10x Genomics® Chromium™ single cell experiments later stage embryos at E8.25 were harvested to span a period beyond gastrulation that incorporates early organogenesis. 2 matings (B6CBAF1/J x B6CBAF1/J) were setup but only one was productive, generating 7 E8.25 embryos. The B6CBAF1/J mice themselves are an F1 cross generated from crossing a C57BL/6J female with a CBA/J male. Mice were all ordered and supplied by Charles River Inc..

Methods

Table 2.1 Genotyping primers.

ERT2	5'-GGGCTCTACTTCATCGCATTCC-3'
tF	5'-TATCCCAGTCTCTGGTCTGTGA-3'
tR	5'-TAGGACCCTACCTAGCAAAGGA-3'

2.1.1 Embryo dissections and imaging

Mice were dissected in ice cold 3 % Gibco® FBS/PBS (ThermoFisher) under a bright-field dissecting microscope. The *Brachyury-2A-nucEGFP-2A-CreER^{T2}* embryos were imaged under the confocal microscope by Dr Ben Steventon from Prof. Alfonso Martinez-Arias' lab. Confocal images were captured on the LSM700 AxioObserver (Zeiss) using the Plan-Apochromat 20x/0.8 M27 lens. *z* stacks were acquired every 5.6 µm. Images were viewed using FIJI [Schindelin et al., 2012] and stacks of interest collated into a single image using Inkscape (www.inkscape.org). Embryos that appeared abnormal or that were of the wrong stage were discarded.

2.1.2 Genotyping

To maintain *Brachyury-2A-nucEGFP-2A-CreER^{T2}* mouse colonies ear notches were used for genotyping. Ear notches were received in 1.5 ml Eppendorf tubes.

DNA extraction was performed using the HotSHOT method [Montero-Pau et al., 2008]. Samples were fully immersed in 50 µl of the lysis reagent and incubated at 95 °C for 1 h, were allowed to cool and 50 µl of neutralising reagent was added and pipetted gently to dissociate. The samples were centrifuged at 400G for 5 min and the supernatant transferred to new 1.5 ml Eppendorfs. 4 µl of nuclease-free H₂O was added to labelled strips of tubes. 1 µl of sample was added to each tube.

PCR was performed using 3 primers as set out in table 2.1.

The primers were ordered from Sigma-Aldrich and diluted to 100 µmol⁻¹ in TE (Tris EDTA) buffer, as stock. A working batch was generated using a 1:100 dilution - 200 µl of each primer was made. A master PCR mix was made for each sample and additionally a positive and negative control (*HoxB8* mouse cell line). Samples were transferred to a PCR thermocycler with the following program, table 2.2.

A 1.2 % agarose gel was prepared by adding 1.2 g of agarose to 100 ml of TBE *Tris – Borate – EDTA*. It was heated at full power for 2 min until fully dissolved. 5 µl of Ethidium bromide was added to the gel and it was poured with 2 twelve well combs. After the gel set

Table 2.2 Genotyping PCR programme.

Temp	Purpose	Duration	Cycles
95 °C	Denature	2 min	
92 °C	Denature	10 sec	30
58 °C	Anneal	30 sec	30
68 °C	Elongate	2 min	30
72 °C	Final elongation	5 min	
12 °C			

it was transferred into the electrophoresis chamber and the left most wells were loaded with 5 µl of New England Biolab's 2-log DNA ladder (0.1-10kB). 10 µl of sample was added in order to each of the other wells. 100 V was applied to the gel for 45 min. The homozygous mutant is non-viable therefore two bands at 400bp and 550bp were expected in the viable heterozygous mutant mice while a single 550bp fragment was expected in the wild-type mice.

2.2 Processing embryos

2.2.1 Generating single cell suspensions

To generate single cell suspensions embryos were individually placed into 250 µl of Gibco® TrypLE™ Express (ThermoFisher®) in a 1.5 ml Eppendorf® tube. Embryos were then incubated at 37 °C for 3 min to 5 min, gently pipetted to dissociate and quenched in 1000 µl heat-inactivated serum. Cells were gently centrifuged for 5 min at 400G and washed in 3 % FBS/PBS with supernatant carefully aspirated to avoid cell loss, three times. Cells were then resuspended in 25 µl of FACS buffer (2 % FBS/PBS) with mouse F_c (CD16/CD32) block. For the 10x Genomics® experiments the cells were instead resuspended in 0.04 % BSA.

2.2.2 Antibody hybridisation

For the Smart-seq2 gastrulation wild-type experiment, *Tall*^{-/-} and *Brachyury-2A-nucEGFP-2A-CreER*^{T2} [Imuta et al., 2013] experiments cells were stained using a master mix of antibodies and a viability stain, DAPI. For the gastrulation wild-type and *Tall*^{-/-} experiments FLK1 and CD41 antibodies were hybridised to the single cell suspensions and FACS performed by Yosuke Tanaka.

For the *Brachyury-2A-nucEGFP-2A-CreER*^{T2} experiment single cell suspensions were stained for viability with DAPI and for 7 surface proteins as shown in table 2.3. Master mix

Methods

(2x concentration) was made in 250 μ l for the 9 embryos section 3.14.3 and fig. 3.56. Embryo 2 was sacrificed for unstained control and DAPI control. The samples were incubated with the master mix stain at 4 °C for 20 min. Samples were spun and resuspended in 300 μ l FACS buffer and 1:5000 DAPI except E-CAD-biotin which was secondarily stained with Streptavidin PCPCy5.5, 0.25 μ l in 50 μ l of FACS buffer for 20 min at 4 °C. The secondary stain was washed and the cells suspended in 300 μ l. The remaining sample was stored at 4 °C while the secondary staining was performed.

Alongside the 9 samples, an unstained control and DAPI control using Embryo 2, and single stain controls using OneComp eBeads (eBiosciences) were performed. One drop of beads was added to sorting compatible FACS tubes with 25 μ l of FACS buffer. 0.5 μ l of each antibody stain was added to the corresponding tube and incubated at 4 °C for 20 min. Beads were washed and resuspended in 300 μ l of FACS buffer except for the E-CAD single stain beads.

2.2.3 Fluorescent activated cell sorting

Six 96 well plates and a test plate containing lysis buffer were prepared according to a protocol used in the laboratory, itself adapted from Picelli et al. [2014]. The protocol was optimised by many including Dr Nicola Wilson and Dr Fernando Calero-Nieto.

1 μ l of RNase inhibitor was added to 19 μ l of 0.2 % (vol/vol) Triton X-100 solution. 2.3 μ l of the mixture was aliquoted into every well of a 96 well plate. Six full plates and one test plate with one column (8 wells) was prepared. Plates were stored at 4 °C.

Stained cells were sorted on the 5 laser machine (BD Influx) in single-cell mode with index sorting by the staff at the flow sorting facility at CIMR. Cells were sorted into lysis buffer in the six pre-prepared 96 well plates, two plates per embryo. A further 8 cells were sorted into a 'test plate'. The test plate was used to check optimal ERCC concentration and number of cycles to use. After sorting plates were spun down at 700G for 1 min and stored at -80 °C.

Table 2.3 Staining protocol.

Antibody	Fluorophore	Laser & Filter	Manufacturer	Cat Num.	Dilution	In 250 μ l of 2x stain
CKIT	BV421	405 450/50	Biologend	105827	1:100	5 μ l
ECDH	biotin		eBioscience	13-3249-82	1:100	5 μ l
Streptavidin	PCPCy5.5	488 695/40	eBioScience	45-4317-82	1:200	Not in master mix
CD31	PE	561 586/15	BD Pharmingen	553373	1:200	2.5 μ l
PDFGR α	PECF594	561 610/20	BD Horizon	562775	1:400	1.25 μ l
FLK1	PECy7	561 780/60	Biologend	136414	1:200	2.5 μ l
TIE2	APC	640 670/40	Biologend	124010	1:200	2.5 μ l
EPCAM/CD326	APCCy7	640 780/60	Biologend	118218	1:200	2.5 μ l
BRACHYURY	GFP	488 530/40				
Viability	DAPI				1:5000	

2.3 Library preparation

For the Smart-seq2 gastrulation wild-type experiment and *Tall*^{-/-} library preparation was performed by Dr Nicola Wilson [Scialdone et al., 2016]. For the *Brachyury*-2A-nucEGFP-2A-CreER^{T2} experiment the libraries were prepared by Sonia Nestorowa and myself with the help of Dr Fernando Calero-Nieto.

2.3.1 Smart-seq2

Complementary DNA generation

Smart-seq2 libraries were prepared as described by Picelli et al. [2014]. Plates were removed from -80°C , allowed to thaw on ice and spun down at 700G for 1 min. Plates were kept on ice throughout. The annealing mixture was prepared for 96 wells for sample plates or 8 wells for the test plate. A dilution of 1:3,000,000 was selected for ERCCs based on results from the test plate. 2 μl annealing mixture of the ERCCs, oligo-dT and dNTPs was incubated with each single cell sample in the 96 well plate at 72°C for 3 min and immediately placed on ice, see table 2.4.

Table 2.4 Annealing mix.

	1 \times 100 wells
ERCC 1:3 $\times 10^6$	10 μl
Oligo-dT 100 $\mu\text{mol dm}^{-3}$	10 μl
dNTP 10 mmol dm^{-3}	100 μl
dH ₂ O	80 μl
Total	200 μl

The reverse transcription mix was prepared containing Superscript II® reverse transcription enzyme (200 U/ μl), RNase inhibitor (20 U/ μl), Superscript first strand buffer, dithiothreitol, betaine, MgCl₂ and transcript switching oligomer, see table 2.5. 5.7 μl of the reverse transcription mix was added to each well and incubated at 42°C for 90 min and cycled 10 times to 50°C to allow secondary RNA structures to unfold as summarised in table 2.6.

cDNA was then generated by performing PCR. The PCR mix was prepared as shown in table 2.7. 15 μl of the mix is added to each well and PCR performed, table 2.8. For the Smart-seq2 gastrulation wild-type and *Tall*^{-/-} experiments were performed by Dr Nicola Wilson. For the *Brachyury*-2A-nucEGFP-2A-CreER^{T2} experiment the initial test plate experiment

2.3 Library preparation

Table 2.5 Smart-seq2 reverse transcription mixture.

	Per well	1 × 100 wells
Superscript II RT (200 U μ l)	0.5 μ l	50 μ l
RNase inhibitor (20 U μ l)	0.25 μ l	25 μ l
5 × superscript II first strand buffer	2 μ l	200 μ l
100 mmol dm ⁻³ DTT	0.5 μ l	50 μ l
5 mol dm ⁻³ Betaine	2 μ l	200 μ l
1 mol dm ⁻³ MgCl ₂	0.06 μ l	60 μ l
TSO (100 μ mol dm ⁻³)	0.1 μ l	10 μ l
dH ₂ O	0.29 μ l	29 μ l
Total	5.7 μ l	570 μ l

Table 2.6 Smart-seq2 reverse transcription thermocycler programme.

cycle	Temperature °C	Time min	Purpose
1	42	90	RT and template switching
10	50	2	Unfolding of RNA secondary structures
	42	2	Completion of RT and template switching
1	70	15	Enzyme inactivation
-	4	Hold	Safe storage

suggested the use of 21 PCR cycles was optimum for cDNA library generation and so this was selected for the sample plates, table 2.8.

The cDNA libraries were size selected using Ampure XP® beads at a DNA:bead ratios 1:0.6. Ampure Beads were added to each sample and the solution homogenised by pipetting and incubated at room temperature for 8 min. The 96 well culture plates were placed on a magnet for 5 min to allow the beads to settle, the beads with bound DNA were washed with freshly made 80 % ethanol twice and then eluted into 23 μ l elution buffer and pipetted. 20 μ l of cleaned samples were then transferred to a new sample plate. Samples of the cleaned cDNA were quality checked on the Agilent® high-sensitivity DNA chip bioanalyser. DNA content was quantified using the ThermoFisher Scientific® Quant-iT™ PicoGreen™ dsDNA assay kit.

Library generation

The cDNA libraries were thawed and dilutions made so that most samples contained 0.1 ng μ l⁻¹ to 0.15 ng μ l⁻¹ of total input DNA. Sequencing libraries were generated by

Methods

Table 2.7 PCR mix

	1 well (μl)	100 wells (μl)
KAPA HiFi Hotstart ReadyMix (2x)	12.5	1250
IS PCR primer ($10 \mu\text{mol dm}^{-3}$)	0.25	25
Nuclease-free dH ₂ O	2.25	225
Total	15	1500

Table 2.8 PCR mix thermocycler programme

Temp $^{\circ}\text{C}$	Purpose	Duration	Cycles
98	Denature	3 min	1
98	Denature	20 sec	21
67	Anneal	15 sec	
72	Extend	6 min	
72	Extend	5 min	1
4	Hold	-	1

tagmentation using the Illumina Nextera XT DNA kits as described by Picelli et al. [2014]. Tagmentation was performed by creating a master mix, pipetting 3.75 μl of the mix into a new Library 96 well plate and adding 1.25 μl from the corresponding well in the sample plate into the wells in the Library plate, table 2.9. Plates were then sealed and placed in the thermal cycler at 55 $^{\circ}\text{C}$ for 10 min followed by cooling to 10 $^{\circ}\text{C}$. The DNA content at this tagmentation step was critical to ensure the tagmented DNA fragments are of the required length.

Table 2.9 Tagmentation mix

Reagent	Volume per sample μl	X 96 well plate + 10 %
Tagmentaion DNA buffer	2.5	264
Amplicon Tagment Mix	1.25	132
Sample DNA	1.25	
Total	5	

1.25 μl NT buffer was added to the library plate so that each well now had a volume of 6.25 μl . Indices were now required to be added to the tagmented DNA fragments. 3.75 μl of Nextera PCR Master Mix was added to each sample to give a total volume of 10 μl . Unique index pairs were then added to each plate with 12 column indices and 8 row indices. 1.25 μl

2.3 Library preparation

of each index mix was added and now the total well volume was 12.5 μ l. Plates were sealed and placed in the thermal cycler with the programme shown in table 2.10.

Table 2.10 Nextera PCR indexing programme

Temperature °C	Time	Cycles
72	3 min	1
95	30 sec	1
95	10 sec	12
55	30 sec	
72	60 sec	
72	5 min	1
10	hold	

Since all molecules were now indexed by their cell of origin the libraries from each single 96 well plate were pooled. In this case 2 μ l from each sample was pooled into a 1.5 ml Eppendorf® tube for each plate. The pooled sequencing libraries were now cleaned sequentially at three different DNA:bead ratios 1:0.8, 1:0.9 and 1:0.7 in a manner similar to that previously described for the cDNA libraries though in this case a single clean for each pooled library was required. The pooled cleaned libraries was assessed using the Agilent® high-sensitivity DNA chip bioanalyser. The pooled libraries were quantified using the ThermoFisher Scientific® Quant-iT™ PicoGreen™ dsDNA assay kit and diluted to 10 nmol to 20 nmol.

2.3.2 10x Genomics® Chromium™ Single-cell Gene Expression

The 10x Genomics® experiment libraries presented in this thesis were prepared by Dr Fernando Calero-Nieto. Single cell suspensions from dissociated mice embryos in 0.04 % BSA were used as input. Version 1 of the 10x Genomics® Chromium™ Single Cell 3' Reagent Kit was used to prepare the libraries. After preparing the single cell suspensions a cell count was performed so the desired number of cells could be loaded on the 10x Genomics® single cell chip, table 2.11.

GEM generation and Barcoding

Gel bead in EMulsion (GEM) was generated by preparing the reverse transcription single cell master mix table 2.12. A target output of 1200 cells was used to dilute the cell suspensions and calculate volumes for the master mix which was made with the lowest amount of nuclease-free H₂O and additional was added to individual samples as required, table 2.11.

Methods

Table 2.11 10x Genomics® Chromium™ single cell 3' experiment cell suspension calculations. The column labelled additional volume provides the additional volume of nuclease-free H₂O to be added to each sample after generating the master mix with only 26.5 µl for each sample table 2.12.

Embryo	Concentration cells/µl	Cell suspension volume µl	Nuclease-free H ₂ O µl	Additional to be added µl
2	235	10.9	26.5	0.0
3	710	3.6	33.8	7.3
4	365	7.0	30.4	3.9

Table 2.12 10x Genomics® Chromium™ single cell 3' reverse transcription GEM generation Master Mix

Reagent	1X (µl)	8.8X (µl)
Nuclease-Free H ₂ O	26.5	233.2
RT Reagent Mix	50.0	440
RT Primer	4.0	35.2
RNase Inhibitor	1.5	13.2
Additive A	2.5	22.0
RT Enzyme	4.6	40.5
Total	89.1	784.1

89.1 µl of the master mix was added to each tube of an 8-tube strip. Additional Nuclease-free H₂O and the specified amount of cell suspension volume were added to the appropriate tubes of the 8-tube strip, tables 2.11 and 2.13.

A 10x Genomics® Chromium™ single cell 3' chip was placed in its holder and loaded. The two unused lanes were loaded with surrogate fluid (50 % volume/volume glycerol/water) with 90 µl in row 1, 40 µl in row 2 and 270 µl in row 3, table 2.13. 90 µl of the samples were now loaded from the 8-tube strip into row 1 of the 10x Genomics® Chromium™ single cell 3' chip. For the sample lanes 40 µl of 10x Genomics® Chromium™ 3' gel beads from the kit were loaded into row 2 without introducing bubbles and 270 µl of partitioning oil into row 3. The provided gasket was fitted and the 10x Genomics® Chromium™ single cell 3' chip loaded into the Chromium™ controller. The programme was run so that cells were incorporated into gel beads and collected in the recovery wells. 105 µl of samples were aspirated and placed into an Eppendorf® twin-tec 96 well PCR plate. The plate was sealed with a pierceable foil heat seal using a plate sealer plate block set at 185 °C for 6 min.

2.3 Library preparation

Table 2.13 10x Genomics® Chromium™ single cell 3' sample preparation in an 8-tube strip.

Lane	Embryo	Master mix volume (µl)	Nuclease-free H ₂ O (µl)	Sample cell suspension volume (µl)
1	2	89.1	0	10.9
2	2	89.1	0	10.9
3	3	89.1	7.3	3.6
4	3	89.1	7.3	3.6
5	4	89.1	3.9	7.0
6	4	89.1	3.9	7.0
7	Surrogate fluid			
8	Surrogate fluid			

Reverse transcription was now performed with the cells trapped in the GEMs with unique 10x Genomics® barcodes. Reverse transcription was performed in the Bio-Rad C1000 Touch™ with the programme set out in table 2.14, lid temperature set at 55 °C and reaction volume set at 25 µl.

Table 2.14 10x Genomics® Chromium™ single cell 3' GEM-RT

Temp °C	Time	Cycles
55	2 h	1
85	5 min	1
4	Hold	

Post GEM-RT cleanup & cDNA Amplification

After the GEM-RT step the mRNA moieties from cells had been captured on the barcoded primers and the first strand cDNA of the full length mRNA generated capped with a TSO on the 3' end. Since the individual molecules from each cell had been barcoded the emulsion was safely broken. The samples were transferred from the Eppendorf® twin-tec 96 well PCR plate to an 8-tube strip. The emulsion was separated by adding 125 µl of Recovery agent. The recovered mixture was now biphasic containing distinct layers, the clear aqueous phase and the pink recovery agent/partitioning fluid. 125 µl pink layer was aspirated and discarded.

A clean-up step using Silane DynaBeads ensued. The DynaBeads Cleanup Mix was prepared, table 2.15. 200 µl of the DynaBeads Cleanup Mix was added to each sample. The sample and mix were pipetted and incubated at room temperature for 10 min. After the first 5 min the mixture was re-pipetted.

Methods

Table 2.15 DynaBeads Cleanup Mix

Reagent	1X (µl)	8.8X (µl)
Buffer for sample cleanup 1	182	1602
DynaBeads MyOne Silane	14	123
Additive A	4	35
Total	200	1760

Elution solution I was prepared as shown in table 2.16. After the 10 min incubation the 8 tube strip was placed in the 10x Genomics® Magnetic Separator in the high position until the solution was clear. The supernatant was carefully aspirated and discarded. 300 µl of freshly prepared 80 % ethanol was added to the pellet while still on the magnet and allowed to stand for 30 sec. This was carefully removed without disturbing the pellet and a further 200 µl of 80 % ethanol was added to the pellet and allowed to stand for 30 sec before being removed. The strip was centrifuged briefly and placed again in the 10x Genomics® Magnetic Separator this time in the low position. Any remaining ethanol was discarded and the samples allowed to dry for 2 min. The tube was removed from the magnet and 50.5 µl of Elution solution I added and mixed thoroughly by pipetting 15 times and then incubated at room temperature or 1 min. The solution was once again placed in the 10x Genomics® Magnetic Separator in the high position until the solution was clear. 50 µl of supernatant was aspirated from each sample and transferred to a new 8-tube strip.

Table 2.16 Elution Solution I

Reagent	1 reaction (µl)	10 reactions (µl)
Buffer EB	98	980
10 % Tween20	1	10
Additive A	1	10
Total	100	1000

In the next step excess primers and TSO were removed using Beckman Coulter SPRIselect magnetic beads. Elution solution II was prepared as shown in table 2.17. The SPRIselect reagent was vortexed until full resuspended, 30 µl was added to each of the 6 samples in the 8 tube strip and mixed by pipetting 15 times. The mixtures were incubated at room temperature for 5 min. The strip was placed in the 10x Genomics® Magnetic Separator in the high position until the solution was clear and the supernatant aspirated and discarded. 125 µl of freshly prepared 80 % ethanol was added to each of the sample pellets, allowed to stand for 30 sec and then discarded. This ethanol was repeated once. After the final

2.3 Library preparation

ethanol wash the tube was briefly centrifuged, placed in the magnet in the high position and any remaining ethanol was discarded. The sample/bead pellets were allowed to dry for 1 min, then the 8-tube strip was removed from the magnet and 35.5 μ l of Elution solution II added. The solution was mixed by pipetting 15 times and incubated at room temperature for 2 min. The strip was replaced on to the 10x Genomics® Magnetic Separator in the low position and 35 μ l of supernatant transferred to a new 8-tube strip.

Table 2.17 Elution Solution II

Reagent	1 reaction (μ l)	10 reactions (μ l)
Buffer EB	99	990
Additive A	1	10
Total	100	1000

The cDNA reaction mix was now prepared, table 2.18. 65 μ l of the cDNA amplification reaction mix was added to each tube containing 35 μ l of sample. Tubes were then capped and loaded into the Bio-Rad C1000 Touch™ thermocycler, with the lid temperature set at 105 °C and reaction volume set at 100 μ l. The PCR programme is given in table 2.19.

Table 2.18 10x Genomics® Chromium™ cDNA amplification reaction mix.

Reagent	1X (μ l)	8.8X (μ l)
Nuclease-free H ₂ O	8	70
Amplification Master Mix	50	440
cDNA Additive	5	44
cDNA Primer Mix	2	18
Total	65	572

Table 2.19 10x Genomics® Chromium™ single cell 3' cDNA amplification reaction programme.

Temperature (°C)	Time	Cycles
98	3 min	1
98	15 sec	14
67	20 sec	
72	1 min	
72	1 min	1
4	Hold	

Methods

A further clean-up step with SPRIselect leads followed. SPRIselect reagent was vortexed until fully resuspended and 60 µl added to each sample and mixed thoroughly by pipetting 15 times. The mixture was incubated at room temperature for 5 min and then placed in the 10x Genomics® Magnetic Separator in the high position until until the solution was clear. The supernatant was carefully aspirated and discarded. The sample/beads were washed with freshly prepared 80 % ethanol twice by adding 200 µl of ethanol, allowing to stand for 30 sec and discarding at each wash. The tube was then briefly centrifuged, replaced on the magnet in the low position and any remaining ethanol discarded. The pellet was allowed to air dry for 2 min, the tube was removed from the magnet and the beads resuspended in 55.5 µl of EB buffer (Qiagen®). The solution was thoroughly mixed by pipetting 15 times and incubated at room temperature for 2 min. The strip was placed in the 10x Genomics® Magnetic Separator in the high position and 55 µl of supernatant transferred to a new strip. 1 µl of each sample was diluted 1:2 in nuclease-free H₂O and run on the Agilent® high sensitivity DNA-chip bioanalyser.

Library construction

51 µl of each sample were sheared using the Covaris® S220 focused ultrasonicator™ to achieve a target peak size of 200bp for a standard DNA sample. Post shearing samples were centrifuged briefly and 50 µl transferred to a new 8-tube strip.

Post shearing a double ended size selection step using SPRIselect beads followed. SPRIselect reagent was vortexed until fully resuspended. 30 µl of SPRIselect beads was added to each 50 µl sample and thoroughly mixed by pipetting 15 times. The strip was incubated at room temperature for 5 min. The 8-tube strip was placed in the 10x Genomics® Magnetic Separator in the high position until the solution was clear. 75 µl of the supernatant was then transferred to a new 8-tube strip and the SPRIselect beads discarded. An additional 10 µl of SPRIselect beads were then added to the 75 µl sample and SPRIselect reagent mix. This was incubated at room temperature for 5 min and the 8-tube strip placed in the 10x Genomics® Magnetic Separator in the high position until the solution was clear. The supernatant was very carefully removed and discarded without disturbing the beads. A total of two washes were then performed with 125 µl of 80 % ethanol allowing the solution to stand for 30 sec before discarding. After the second wash the 8-tube strip was briefly centrifuged, replaced on the magnet in the low position and any remaining ethanol removed. The 8-tube strip was then removed from the magnet and resuspended in 50.5 µl of EB buffer thoroughly mixing by pipetting 15 times. The samples were incubated at room temperature for 2 min placed on

2.3 Library preparation

the 10x Genomics® Magnetic Separator until clear and 50 µl of the sample transferred to new 8-tube strip.

The end repair and A-tailing mix was prepared as shown in table 2.20. 10 µl of the end repair and A-tailing mix was added to each 50 µl sample and mixed thoroughly by pipetting 15 times. The samples are then placed in the Bio-Rad C1000 Touch™ thermocycler, with the lid temperature set at 85 °C, reaction volume set to 60 µl and the programme given in table 2.21.

Table 2.20 End repair and A-tailing mix preparation.

Reagent	1X (µl)	8.8X (µl)
End Repair and A-tailing Buffer	7	62
End Repair and A-tailing Enzyme	3	26
Total	10	88

Table 2.21 End repair and A-tailing thermocycler programme.

Temperature (°C)	Time	Cycle
20	30 min	1
65	30 min	1
4	Hold	

Adaptors were now ligated by preparing the Adaptor ligation mix as shown in table 2.22. 50 µl of the adaptor ligation mix was added to 60 µl samples after end repair and A-tailing. The solution was mixed thoroughly by pipetting 15 times and incubated in the Bio-Rad C1000 Touch™ thermocycler with lid temperature 30 °C and reaction volume 110 µl for 15 min at 20 °C.

Table 2.22 Adaptor ligation mix.

Reagent	1X (µl)	8.8X (µl)
Nuclease-free H ₂ O	7.5	66
Ligation buffer	30	264
DNA ligase	10	88
R1 Adaptor Mix	2.5	22
Total	50	440

Post adaptor ligation cleanup was performed twice. For the first cleanup 88 µl of SPRIselect beads were added, mixed thoroughly by pipetting 15 times and incubated at room temperature

Methods

for 5 min. The 8-tube strip was then placed in the 10x Genomics® Magnetic Separator in the high position until the solution was clear. The supernatant was discarded followed by two 30 sec, 200 µl 80 % ethanol washes allowing. The samples were briefly centrifuged, replaced in the 10x Genomics® Magnetic Separator, any excess ethanol discarded and allowed to dry for 2 min. The beads were removed from the magnet and resuspended in 30.5 µl EB buffer, thoroughly mixed by pipetting 15 times and incubated at room temperature for 2 min. The 8-tube strip was replaced in the 10x Genomics® Magnetic Separator until the solution was clear and the 30 µl supernatant transferred to a new 8-tube strip. The cleanup step was repeated with 50 µl of SPRIselect beads with 125 µl ethanol washes and again resuspending in 30.5 µl of EB buffer and transferring 30 µl to a new 8-tube strip.

Sample indices were selected from the supplied kit and were now ligated by preparing the sample index PCR mix, table 2.23. 60 µl of the sample index PCR mix and 10 µl of an individual single cell 3' sample index was added to each of the 6, 30 µl samples and mixed thoroughly by pipetting 15 times. The sample indices were used sequentially and recorded to allow later de-multiplexing. The samples were then placed in the Bio-Rad C1000 Touch™ thermocycler with lid temperature 105 °C and reaction volume 100 µl with the programme shown in table 2.24.

Table 2.23 10x Genomics® sample index PCR mix.

Reagent	1X (µl)	8.8X (µl)
Nuclease-free H ₂ O	8	70
Amplification master mix	50	440
SI-PCR Primer	2	18
Total	60	528

Table 2.24 10x Genomics® Chromium™ sample index PCR thermocycler protocol.

Temperature (°C)	Time (sec)	Cycles
98	45	1
98	20	10
60	30	
72	20	
72	60	1
4	Hold	

A further two SPRIselect bead cleanup steps were performed after the sample index PCR stage. The SPRIselect beads were vortexed until fully resuspended. The cleaning steps were

as previously described but 100 μ l of beads was added to the 100 μ l of sample and washed twice with 200 μ l of ethanol. Finally the sample was eluted off the beads in 50.5 μ l of EB buffer and 50 μ l was transferred to a new 8-tube strip. In the second wash 50 μ l of SPRIselect beads was added and two washes with 125 μ l 80 % ethanol were performed. After the final wash the sample was resuspended in 35.5 μ l and 35 μ l was transferred to a new 8-tube strip.

1 μ l of each of the final sample libraries was diluted 1:10 and assessed using the Agilent® high sensitivity DNA-chip bioanalyser. The individual sample libraries were quantified using the ThermoFisher Scientific® Quant-iT™ PicoGreen™ dsDNA assay kit. Since an estimate of cell counts had been made, DNA content per cell could be quantified and the libraries were pooled with relative volumes selected by aiming to achieve equal DNA content per cell in the pooled libraries.

2.4 Sequencing

2.4.1 Smart-seq2 experiments

The sequencing libraries for the wild-type experiments and *Tall*^{-/-} experiments were submitted by Dr Nicola Wilson. They were all sequenced on the Illumina® HiSeq2500 in high-throughput mode. The *Brachyury-2A-nucEGFP-2A-CreER*^{T2} libraries were submitted to the CRUK sequencing facility. For all Smart-seq2 prepared libraries 96 cells from each plate were sequenced individually in a separate lane of a Illumina® high-throughput sequencing flowcell.

2.4.2 10x Genomics® Chromium™ single cell experiments

The sample indices allowed the 10x Genomics® Chromium™ libraries to be pooled prior to submitting for sequencing and they were pooled aiming to achieve equivalent DNA content per cell. The libraries were submitted by Dr Fernando Calero-Nieto and sequenced on the Illumina® HiSeq2500 in high-throughput mode across 16 lanes on 2 flowcells. The libraries were sequenced with an additional two unrelated samples.

2.5 Pre-processing

The sequence data was processed and barcoded reads de-multiplexed by the CRUK sequencing facility resulting in 1 fastq file per cell. The fastq files were assessed using FastQC (version 0.11.3) available from <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

Methods

2.5.1 Alignment and QC

The Smart-seq2 data was aligned using GSNAP with default parameters (version 2017-11-15 available at <http://research-pub.gene.com/gmap/src/gmap-gsnap-2017-11-15.tar.gz>) to Ensembl's mouse genome build release 88 (<http://mar2017.archive.ensembl.org/index.html>). Gene level read counts were produced by providing the aligned sam files output by GSNAP to HTseq (version 0.8.0). QC was performed using the *bglab* package (version 2.0.6, available at <https://github.com/wjawaid/bglab>). For wild-type mice the thresholds used were greater than 200,000, greater than 20 % of reads mapping to endogenous genes and less than 20 % of mapped reads arising from mitochondrial genes.

For the 10x Genomics® Chromium™ single cell 3' experiment the pooled libraries required multiple layers of de-multiplexing were required. The sequencing facility de-multiplexed the samples from the sample indices with each sample index having four potential sequence pairs. Cell Ranger (version 2.1.1) a computational tool provided by 10x Genomics® was then used to perform first pass quality control, cell-level de-multiplexing and alignment for which it uses STAR aligner. It then generated gene-level count data for each cell. Ensembl's mouse genome build release 88 was used to generate a bespoke reference genome for Cell Ranger. Counts from the same embryos were aggregated together using Cell Ranger.

2.5.2 Normalisation and feature selection

Smart-seq2 data was normalised using *bglab* (version 2.0.6) which provides a wrapper that accesses the normalisation algorithm implemented by *scraper* (version 1.2.2) using default values. Feature selection was performed within the *bglab* package using the algorithm described by Brennecke et al. [2013] but without using ERCC spike-ins with the model only fitted on genes with mean counts greater than 10 rather than the quantile threshold originally described and an FDR of 0.1.

Scanpy (version 1.2.2) was used for the initial exploration and normalisation in the 10x Genomics® Chromium™ single cell 3' experiment. Cells with greater than 5 % of reads mapping to mitochondrial genes and cells with more than 6000 expressed genes were excluded from further analysis. Unexpressed genomic features were excluded from further analysis. Feature selection was performed by fitting a linear model to the log transformed mean, coefficient of variance relationship and all 7585 genes above the fitted line were included for downstream analysis.

2.6 Data exploration

2.6.1 Clustering

For the Smart-seq2 wild-type data clustering was performed by Dr Antonio Scialdone using the dynamic tree cut algorithm [Langfelder et al., 2008]. Parameter selection was performed by bootstrapping 100 samples and selecting the most robust parameters.

The Louvain method for community detection was used for clustering the 10x Genomics® Chromium™ single cell 3' dataset [Blondel et al., 2008]. A graph was first generated from the counts data using the *roots* package (<https://www.github.com/wjawaid/roots>), additionally it provided a convenient wrapper to *igraph* for performing Louvain clustering.

2.6.2 Visualisations

Throughout the thesis the figures were almost exclusively produced in R using either the *bglab* (version 2.0.6) or *roots* (version 1.0.4) packages.

Heatmaps were produced in *bglab* with hierarchical clustering, dissimilarity calculated using Spearman rank correlation, Ward's link agglomerative clustering and a fast leaf re-ordering algorithm [Bar-Joseph et al., 2001; Murtagh and Legendre, 2014].

PCA and PCA loading plots were produced in the *bglab* package. Areas of interest were interactively selected in plots generated in R using the *gatepoints* package (version 0.1.3, available from CRAN at <https://cran.r-project.org/package=gatepoints> or <https://www.github.com/wjawaid/gatepoints>). To present multiple PCA dimensions pairs plots were generated from R's base graphics package.

2-dimensional tSNE dimensionality reduction plots were produced by calculating a distance matrix based on the Spearman Rank correlation using the *bglab* package which itself relied on the modified *Rtsne* package (version 0.13) originally written by Jesse Krijthe (<https://github.com/jkrijthe/Rtsne>) and available through CRAN (<https://cran.r-project.org/package=Rtsne>). The modified version is available at <https://www.github.com/wjawaid/Rtsne>.

Diffusion maps were generated in *roots*. All dimensionality plots with gene expression were produced in *bglab*. Other scatter plots, boxplots, density plots were produced in base R. *gatepoints* was used to select points and mark regions of interest on plots. Cell cycle assignment was performed using the *cyclone* function within *scran*.

Methods

2.6.3 Pseudotime inference

Pseudotime inference was performed in two complementary ways. Diffusion pseudotime was calculated as described by Haghverdi et al. [2016] using the *roots* implementation.

Additionally a principal curve was fitted to the data and the distance along the curve from the start to the projection of the cell of interest taken as the pseudotime [Hastie and Stuetzle, 1989]. Principal curves were calculated using the package *princurve* version 1.1-12 available from CRAN at https://cran.r-project.org/src/contrib/Archive/princurve/princurve_1.1-12.tar.gz.

Smoothed gene expression profiles along inferred trajectories with 95 % confidence intervals were calculated using Gaussian process regression as implemented in the package *gpr* version 1.1 available at <https://www.github.com/wjawaid/gpr>.

2.7 ESC to blood differentiation assay

The ESC haematopoietic culture assays to assess the in-vitro effects of the leukotriene LTC₄ were performed by Dr Vasilis Ladopoulos. The methods described here are from Ibarra-Soria et al. [2018] and were written by Dr Vasilis Ladopoulos.

HM-1 murine ESCs (kindly provided by D. Melton) were grown in Knock-Out DMEM (Gibco) supplemented with 15 % serum batch tested for maintenance of pluripotency (HyClone), 1000 U/ml leukaemia inhibitory factor (LIF) (Millipore), 2 mmol dm⁻³ L-glutamine / 100 U/ml penicillin / 100 µg ml⁻¹ streptomycin (Gibco), 0.1 mmol dm⁻³ β-mercaptoethanol (Gibco) at 37 °C, 5 % CO₂, on gelatinized plates (Falcon, Corning) at a plating density of $\approx 2 \times 10^4$ cells/cm². Cells were split every 2–3 days as necessary. ESCs were validated by their ability to differentiate into derivatives of the three germ layers and tested negative for mycoplasma contamination.

ESCs were harvested and plated on gelatinized dishes at a density of 4×10^4 cells/cm² in standard ESC growth medium (described above). The cells were dissociated and plated on gelatinized dishes at a density of 4×10^4 cells/cm² 24 h later. The cells were then dissociated again 24 h later and washed once with PBS to remove all remaining ESC medium and LIF. The cells were resuspended in Iscove's modified Dulbecco's medium (IMDM)-based in vitro differentiation (IVD) medium containing 15 % serum batch tested for embryoid body differentiation (Gibco), 10 % protein-free hybridoma medium II (Gibco), 2 µmol dm⁻³ l-glutamine / 100 U/ml penicillin / 100 µg ml⁻¹ streptomycin, 0.15 µmol dm⁻³ monothioglycerol (MTG), 180 µg ml⁻¹ human transferrin (Roche) and 50 µg ml⁻¹ l-ascorbic acid (Sigma) at a density of 1×10^4 cells/ml. The cells were plated in Costar low-adherence 6-well plates (Corning) and

2.7 ESC to blood differentiation assay

incubated for 4 days at 37 °C, 5 % CO₂ to form embryoid bodies. Zileuton (Sigma), LTC₄ (Abcam) or carrier were added on day 3 at the indicated concentrations. The embryoid body suspension was harvested on day 4, transferred to appropriate tubes and the embryoid bodies were left to settle by gravity for 10 min. The medium was discarded, the embryoid bodies were washed with PBS and left to settle again by gravity. PBS was removed, and the embryoid bodies were completely dissociated by addition of 1 ml TrypLE and gentle pipetting. TrypLE was inactivated by adding 10 ml IMDM containing 20 % embryoid body serum. The cells were counted, centrifuged at 300g for 5 min at room temperature and resuspended in IVD medium. 4×10^4 cells were transferred in 4 ml of MethoCult GF M3434 (STEMCELL Technologies) supplemented with 100 U/ml penicillin / 100 µg ml⁻¹ streptomycin (Gibco). 1 ml aliquots were plated in triplicate in 35 ml low-adherence dishes (Corning). Colonies were counted on day 14, and differences in colony numbers were tested with a two-tailed Student's t-test.

To ensure that treatment with Zileuton or LTC₄ does not affect the proliferation of the mouse ESCs, 1×10^6 cells were harvested by centrifugation after dissociation of embryoid bodies on day 4 and washed in PBS. The cell pellet was resuspended in residual volume and fixed by dropwise addition of ice cold 70 % methanol. The cells were incubated at 4 °C for 1 h and then washed twice with PBS. The cells were resuspended in 300 µl propidium iodide-staining buffer (200 µg ml⁻¹ RNaseA, 20 µg ml⁻¹ propidium iodide, 0.1 % Triton X-100 in PBS) and stained at room temperature for 1 h. The cells were analysed on a BD Fortessa. Post-acquisition analysis was performed with the FlowLogic suite.

To ensure that treatment with Zileuton or LTC₄ does not affect the viability of the mouse ESCs, 1×10^6 cells were harvested by centrifugation after dissociation of embryoid bodies on day 4 and washed in PBS. The cells were resuspended in 100 µl Annexin binding buffer (10 mmol dm⁻³ HEPES, 150 mmol dm⁻³ NaCl, 5 mmol dm⁻³ KCl, 1 mmol dm⁻³ MgCl₂, 1.8 mmol dm⁻³ CaCl₂) containing 5 µl Annexin V APC (BD Biosciences; Cat. no. 550474; Lot. 16808) and 1 µg ml⁻¹ 4,6-diamidino-2-phenylindole (DAPI). The cells were diluted up to 400 µl with Annexin binding buffer and analysed on a BD Fortessa cytometer. Post-acquisition analysis was performed with the FlowLogic suite.

Chapter 3

Single-cell transcriptomic analysis of murine gastrulation

3.1 Background

Substantial progress has been made in understanding the molecular processes that govern embryological development since the landmark chick quail chimera experiments that utilised differences in appearance of the Feulgen stained interphase nuclei [Dieterlen-Lievre, 1975; Le Douarin, 1973; Le Douarin and Teillet, 1973]. At the macroscopic level the well orchestrated and highly stereotyped unidirectional process begins with a single-cell, the fertilised ovum that by cell division, symmetric and asymmetric, progressing through gastrulation forms the well recognised germ layers. The trilaminar disc establishes a multicellular organism containing in many cases hundreds of different cell types precisely organised into functional units.

Understanding the mechanisms underlying this progression in mammals remains challenging. Advances in molecular biology have built on the early avian chimera-based fate mapping experiments. In particular temporal, spatial and cell-type specific genetic manipulations that for example with the use of a reporter enzyme or fluorescent label allow lineage tracing of early cell types selectively marked by their expression of a particular gene at a chosen time [Livet et al., 2007; Utomo et al., 1999]. Transgenic mice can then be used to activate a permanent genetic change in a specific cell type at a specified developmental stage and all its progeny. This change may be the expression of a readily measurable marker e.g. a fluorescent protein or a reporter enzyme [Livet et al., 2007; Utomo et al., 1999]. But they rely on apriori identifying a specific and early genetic marker, possibly in cell subpopulations identified as

Single-cell transcriptomic analysis of murine gastrulation

label retaining in pulse-chase type experiments [Barker et al., 2007]. Additionally leaky and non-specific expression can make interpretation and reproducibility difficult.

Furthermore many of the downstream measurements of gene expression of cell subpopulations are averages over several hundreds of cells. Multiple and sequential time series experiments can elucidate gross trends but transient and rare cell subpopulations may be difficult or impossible to identify with such methods, particularly if the individual cells making independent fate decisions are out of developmental phase and each cell enters and leaves the transient state of interest asynchronously [Magwene et al., 2003]. In contrast, single-cell data can potentially quantify intercellular heterogeneity within populations and may allow for finer temporal resolution.

Reconstruction of a developmental chronology from single cell data during the critical gastrulation phase, which lays the foundation of future organogenesis, combined with gene expression data provides a means to infer gene expression changes along developmental trajectories. Mining this rich source of data for master regulators that direct differentiation will reveal critical molecular pathways and the temporal order at which these master regulators are required to form normal tissues. This deeper understanding will permit a more systematic and practical approach to designing protocols for the *in vitro* differentiation of naive pluripotent iPS or ES cells into mature tissues of interest.

The mesoderm provides a framework and support, structural and nutritional for the epithelial layers and critically has been implicated in directing epithelial growth and differentiation, yet is poorly understood and studied [Rankin et al., 2015]. Early multipotent mesodermal progenitors can form cardiac myocytes, pericytes, smooth muscle cells, endothelium and haematopoietic precursors, see fig. 1.4. Uncovering normal differentiation trajectories can provide not only a basis from which to design robust protocols for differentiating naïve cells to desired target cell types for use in generating disease models and cell-based regenerative therapies but also insights into normal embryological development.

In the early post-implantation murine embryo the epiblast undergoes a substantial transformation characterised by stereotyped morphogenetic movements. The primitive streak appears posteriorly at the epiblast-extraembryonic ectoderm border and extends anteriorly to the distal vertex of the cup, the prospective location of the future node. The process begins at approximately E6.5 and is essentially complete by E8. Prior to the onset of gastrulation the ventral dorsal axis is at least conceptually identified by the dorsal epiblast and ventral endoderm. The appearance of the primitive streak, though possibly secondary, defines the anterior-posterior axis thereby fixing the left right axis at a global level even if molecular differences are not yet discernible.

A key challenge in the development of multicellular organisms is the limit of diffusive capacity, when nutrient delivery and waste removal can be maintained no longer, as numbers of cells and the size of the organism increases. In preparation for surpassing this critical threshold, many organisms have evolved circulatory systems consisting of pumps, vessels and a nutrient and waste exchange fluid. The earliest cells to lose adhesion from the epiblast by undergoing epithelial to mesenchyme transition (EMT) are thus directed towards cardiac, endothelial and blood fates. These cells, barring cardiac, generated at the posterior primitive streak migrate proximally into extraembryonic tissues forming blood islands, where the first wave of blood is generated.

The rapid but continuous changes occurring throughout this period offer an opportunity to use single-cell transcriptomic profiling to interrogate some of the molecular processes governing this ‘metamorphosis’. To achieve this in a cost-effective manner fluorescent labelled antibodies and flow activated cell sorting can prove useful to sort cells prior to mRNA extraction, cDNA construction and preparation of sequencing libraries from specific populations at post-E6.5 embryonic stages. Two cell surface markers (FLK1, CD41) were used to sort the required cell populations prior to mRNA extraction and library preparation. A brief overview and introduction to each molecular marker is given below.

Fetal Liver Kinase (*Flk1*) also known as the kinase insert domain receptor (*KDR*) or vascular endothelial growth factor receptor 2 (*VEGFR-2*) is a cell surface receptor transcribed from the *KDR* locus. Shalaby et al. [1997] showed that *Flk1*^{-/-} mice are non-viable failing to develop mature epithelium and haematopoiesis. Additionally *Flk1*^{-/-} embryonic stem cells are able to contribute to neither endothelium nor embryonic or definitive waves of haematopoiesis in chimera experiments suggesting a cell autonomous role in these lineages [Shalaby et al., 1997]. FLK1 provides a useful cell surface marker to enrich for progenitors of several mesodermal derived tissues in this experiment, particularly for early embryonic haematopoietic precursors, see fig. 1.4 [Evseenko et al., 2010; Kabrun et al., 1997; Kattman et al., 2006; Nishikawa et al., 1998; Yang et al., 2008].

Cluster differentiation 41 (CD41) another cell surface marker, transcribed from the Integrin alpha-IIb (*Itga2b*) locus is present on blood cells but is expressed particularly highly on platelets [Dumon et al., 2012; Ferkowicz et al., 2003]. In platelets, *Itga2b* forms a heterodimer with *Itgb3* to form the cell surface glycoprotein IIb/IIIa receptor which binds fibrinogen playing a central role in platelet aggregation, adhesion and thrombus formation [Xiang et al., 2016]. In development, CD41 antigen is expressed not only by greater than 80% of embryoid body derived haemogenic precursors but also most yolk sac and AGM haematopoietic progenitors [Mitjavila-Garcia et al., 2002]. CD41 expression in yolk sac primitive haematopoiesis

Single-cell transcriptomic analysis of murine gastrulation

makes it a useful target to enrich for mesoderm derived cells destined toward a blood fate that may have formerly down-regulated FLK1 [Ferkowicz et al., 2003; McGrath et al., 2015]. Therefore combining FLK1 and CD41 markers at appropriate embryological stages allows capture of a more complete developmental trajectory.

Early development of the FLK1⁺ mesodermal sub-compartment has been previously studied by the Gottgens group. Single cell qPCR gene expression for 48 selected genes in nearly 4000 FLK1⁺ mouse embryo derived cells has been measured using the Fluidigm[®] Biomark[™] HD system [Moignard et al., 2015]. Dimensionality reduction techniques have been used to infer developmental trajectories and a boolean gene regulatory network has been synthesised from the state space graph. The data and analysis show the expected trajectories (endothelial vs. haematogenic) but given that genes were selected apriori this is unsurprising. A more general transcriptome wide analysis will therefore be more revealing.

This chapter describes a transcriptome-wide survey of single cells spanning gastrulation focusing particularly on the haematopoietic mesoderm using the surface markers FLK1⁺ and CD41⁺. To try and capture early gastrulation an unbiased sample of single cells from the E6.5 stage mouse embryo was processed in a similar fashion.

3.2 Cell sampling and library preparation

Single cell unbiased whole transcriptome RNAseq analysis was performed in mouse embryos at 4 timepoints spanning gastrulation: E6.5; E7.0; E7.5 and E7.75. In total 33 mouse embryos were processed supplying 1205 cells that passed quality control. 501 cells from seven E6.5 whole embryos were harvested by generating a single cell suspension from the distal end of the egg cylinder. Single cell suspensions generated from larger older embryos acquired at E7.0 (138 cells, 3 embryos), E7.5 (259 cells, 12 embryos) and E7.75 (307 cells, 11 embryos) were FACS sorted for FLK1 or CD41.

Embryos were dissected under magnification in ice-cold 2% FBS/PBS. The yolk-sac was preserved during dissection and the embryos transferred into TrypLE[™] and incubated at 37 °C for 5 to 10 minutes until the embryos were visually determined to be sufficiently dissociated. Larger embryos were dissected into smaller fragments to aid dissociation and avoid excessive periods in TrypLE[™].

Single cell suspensions generated from E6.0 embryos were cell sorted using a *viability* stain only (DAPI). All three E7.0 embryos, four E7.5 and three E7.75 embryos were sorted for Flk1⁺ and no *index* data was collected. Eight E7.5 embryos and eight E7.75 embryos were

3.2 Cell sampling and library preparation

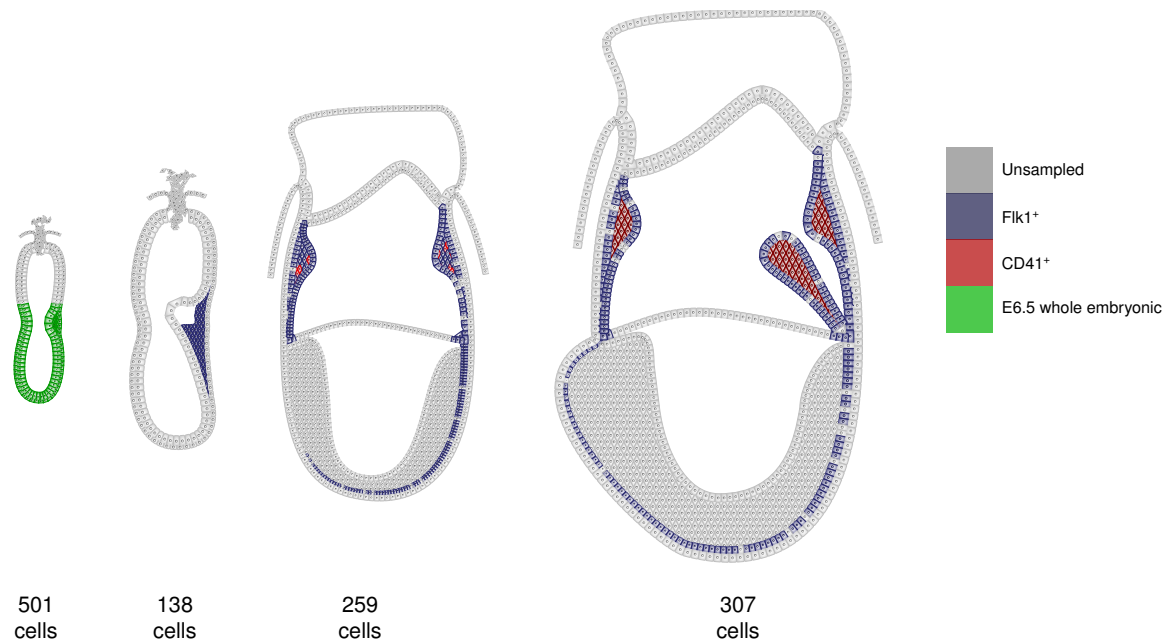


Fig. 3.1 Mid-sagittal sections through embryos at sampling stages E6.5-E7.75 similar to fig. 1.4. From left to right schematics of mouse embryos from early streak (E6.5); mid-streak (E7.0); neural plate (E7.5) and head fold (E7.75) stages are shown. In colour are shown the expected anatomical locations within the developing embryo of the cells sampled from the defined sorting strategies. Numbers of cells collected after quality control are shown below the images.

sorted on CD41 expression with collection of FLK1 index data, figs. 3.1 to 3.3. The embryo dissections and FACS was performed by Dr Yosuke Tanaka and Victoria Moignard compiled the FACS panels, fig. 3.3, for Scialdone et al. [2016]. This sorting strategy subset the cells that were sampled from the embryo into nascent mesoderm $FLK1^+/CD41^-$, progenitor populations $FLK1^+/CD41^+$ and committed cells $FLK1^-/CD41^+$. Presumed anatomical locations from which the cells with different markers were sampled at each stage are outlined in fig. 3.1.

In total fifteen 96-well plates were processed and sent for sequencing, corresponding to a maximum number of 1440 cells but four wells were reported as empty from FACS. The few number of empty wells reflects the protocol used at the time, which involved refilling wells marked as empty by the cell sorter. Later we found that one of the four wells marked as empty in fact produced reads and passed all quality control to be identified as a normal cell. This highlighted the risk of potentially introducing two cells into a well with a strategy of refilling and the protocol was accordingly modified by abstaining from refilling to prevent such doublets in future experiments.

Single-cell transcriptomic analysis of murine gastrulation

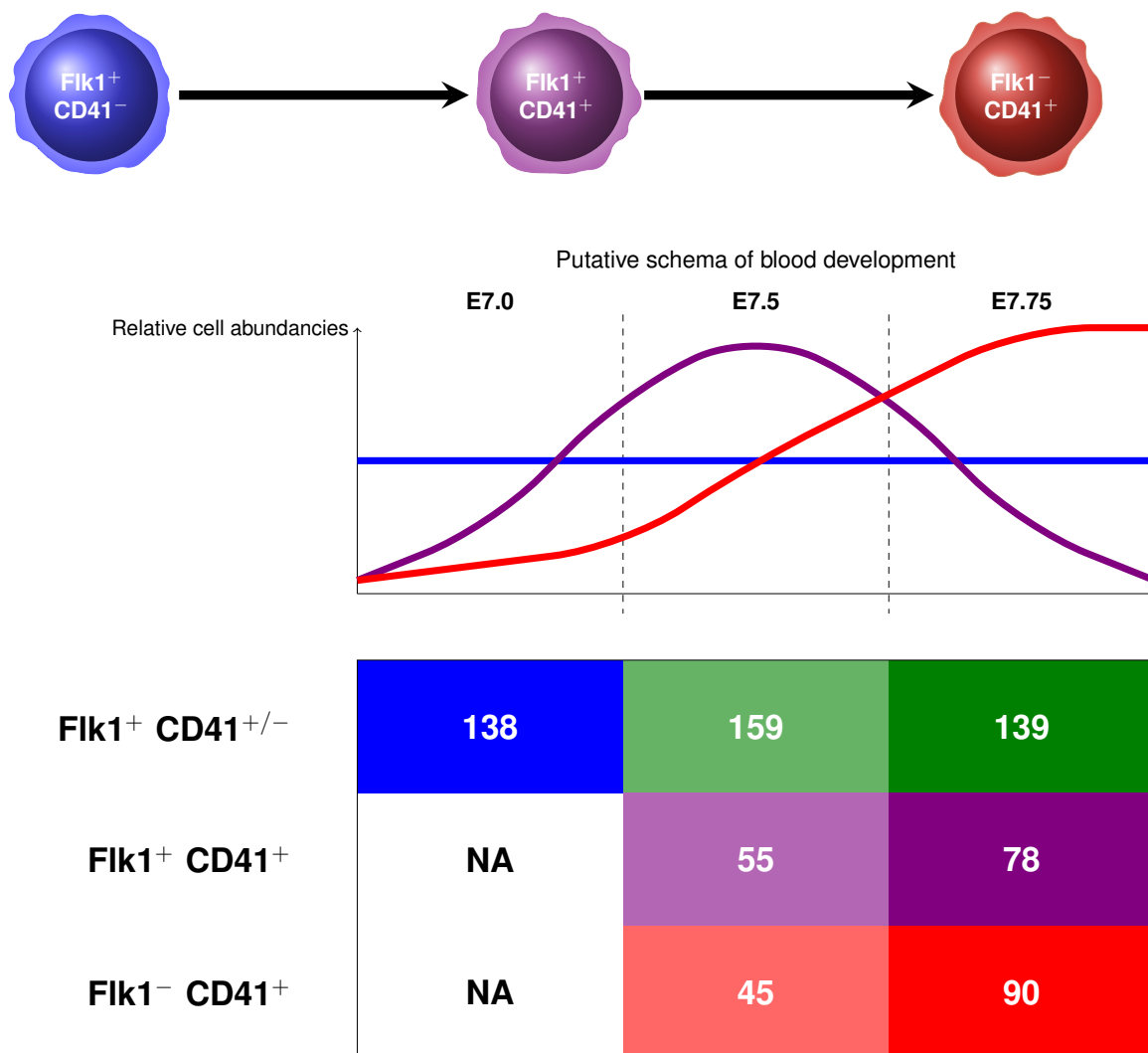


Fig. 3.2 Top: Simplified haematopoietic differentiation from FLK1⁺ CD41⁻ nascent mesoderm cells to double FLK1⁺ CD41⁺ primed and single positive committed FLK1⁻ CD41⁺ erythroid cells. Middle: Density plot of expected relative numbers of the different population of cells at the sampled time points. Bottom: Number of cells sorted with each strategy at each sampling stage. Flk1 sorting was performed without collecting any index data (First row). All other cells were sorted for CD41^{high} and FLK1 *index* data was collected. (Second and Third rows)

The cells were sorted into freshly made lysis buffer containing RNAase inhibitor and then stored at -80°C . Dr Ian MacCauley and Dr Nicola Wilson generated single-cell libraries using the Smart-Seq2 protocol as described by Picelli et al. [2014].

After lysis, pre-prepared ERCC-92 external polyadenylated RNA spike-ins were added prior to capturing native and spike-in polyadenylated mRNA molecules by oligo-dT primers with

3.2 Cell sampling and library preparation

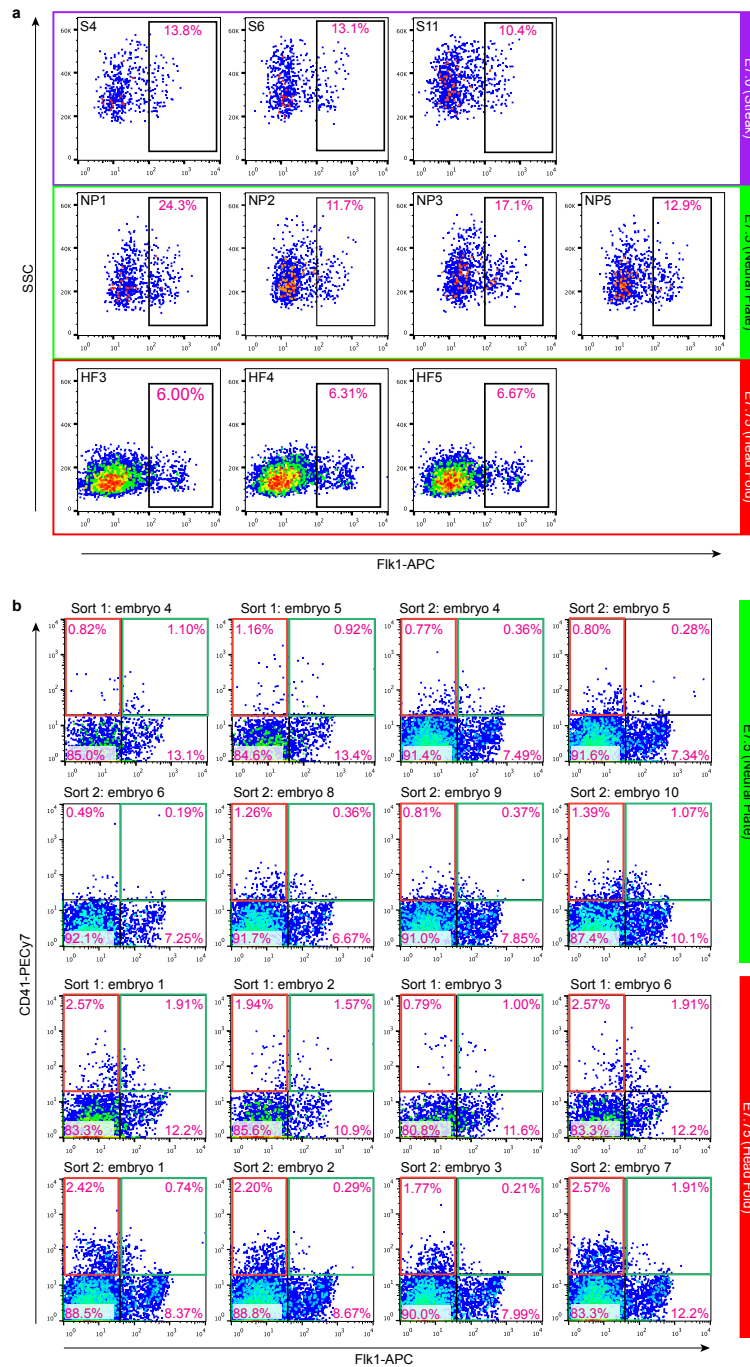


Fig. 3.3 **a**, FLK1⁺ cells were sorted from three E7.0; four E7.5 and three E7.75 stage embryos. No further index data was collected for them and they are shown plotted against SSC. *S*, *Streak*; *NP*; *Neural Plate*; *HF*; *Head fold*. **b**, Gating of CD41⁺/FLK1⁺ (Green) or CD41⁺/FLK1⁻ (Red) cells by embryo. The FACS plots presented here were produced by Victoria Moignard from sorting data collected by Yosuke Tanaka, as part of Scialdone et al. [2016].

Single-cell transcriptomic analysis of murine gastrulation

Table 3.1 Primer sequences used in the Smart-Seq2 protocol. Note the identical segments in the oligo-dT primer and TSO allowing for a single ISPCR primer to be used in the PCR reaction. **V**, Adenosine, Cytidine or Guanosine; **N**, any nucleotide; **TSO** Template switching oligomer; **rG** riboguanosine; **+G**, modified guanosine to produce a LNA [Picelli et al., 2014]. Note that the Oligo-DT primer and the TSO have the ISPCR at the 5' end.

Name	Sequence
Oligo-dT ₃₀ VN	5'-AAGCAGTGGTATCAACGCAGAGTACT ₃₀ VN- 3'
TSO	5'-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3'
ISPCR oligo	5'-AAGCAGTGGTATCAACGCAGAGT-3'

a universal anchor sequence, table 3.1. SuperscriptTM II a synthetic form of the Moloney murine leukaemia virus reverse transcriptase was used for first strand synthesis adding 2 to 5 additional non-templated cytosine residues to the 3' end of the new strand, the 5' end of the template fig. 3.4. A template switching oligomer (TSO) with two riboguanosines and a modified guanosine with a locking nucleic acid (LNA) at the 3' end, hybridises to the additional cytosine residues allowing the SuperscriptTM II to switch templates and synthesise a sequence complementary to the TSO, fig. 3.4 [Matz et al., 1999]. The stability of the RT enzyme allows essentially full-length product to be generated for transcripts up to 12.3 kilobases in length [Kotewicz et al., 1988]. The use of identical 5' and 3' regions allows for a single primer PCR reaction, table 3.1 and fig. 3.4.

Number of cycles of cDNA amplification PCR reaction are based on previous experiments particularly if a test plate was collected at the same time as the sample. Ideally the cells in the test plate would be identical to the sample plate. The purpose of the test plate being to ensure that cells were correctly placed in wells during sorting, to provide an indication of the number of cycles required to get sufficient product from the PCR reaction and to confirm the batch of reagents was good.

Solid phase reversible immobilisation (SPRI) magnetic bead size selection was performed after cDNA generation to remove excess primers, primer dimers and TSO concatemers. Libraries were now generated by tagmentation adding Tn5 transposome to an appropriate dilution of cDNA. It was critical to ensure the sample was not over concentrated so that libraries with insert lengths of the desired length were generated for optimal clustering on the sequencing flowcell.

The Tn5 transposome is engineered to randomly fragment and insert the predetermined Read 1 or Read 2 oligonucleotides in a single step, see table 3.2 and fig. 3.4. After Read 1 and Read 2 adaptors have been annealed, unique combinations of indexed P5 and P7 primers are

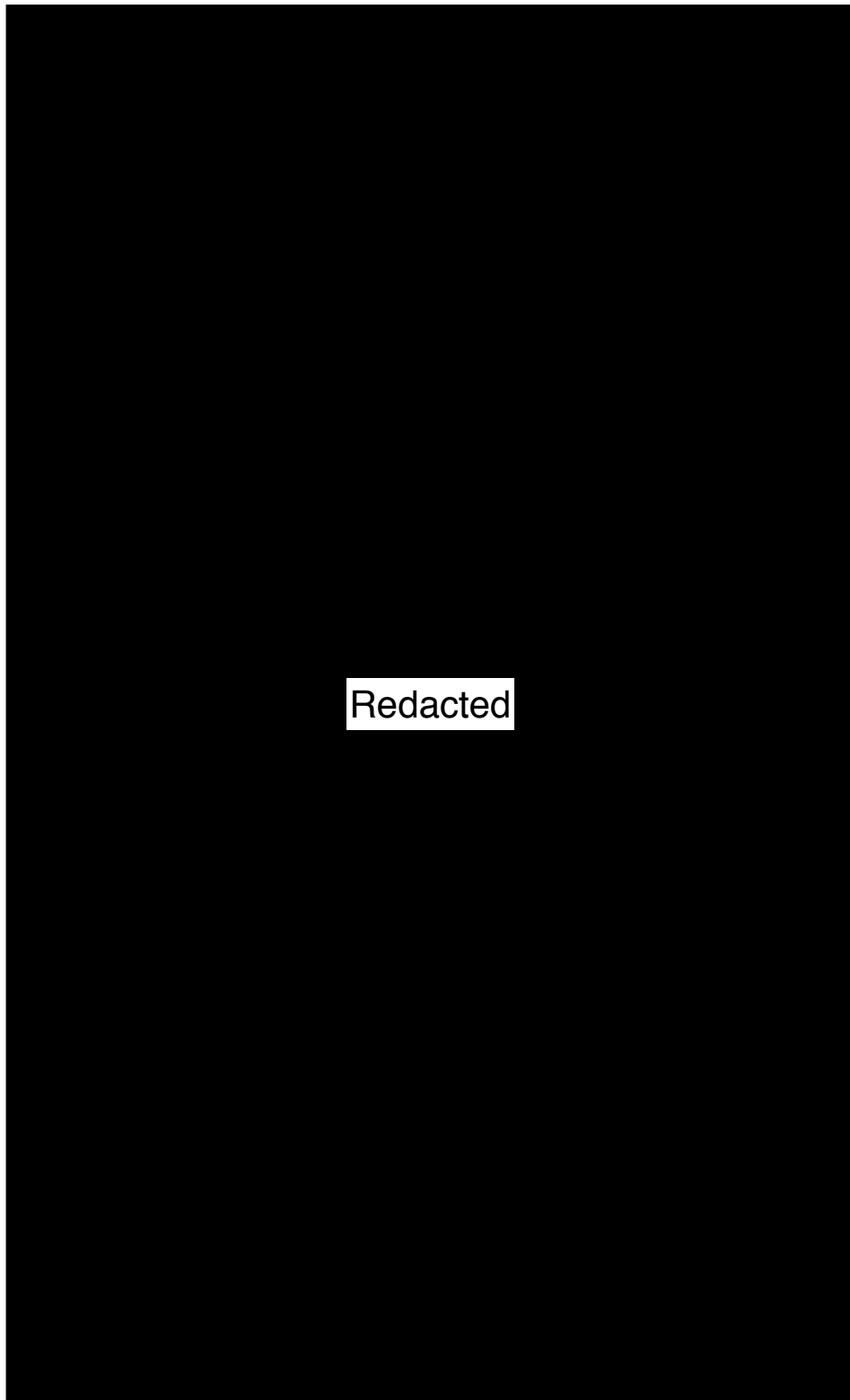


Fig. 3.4 Overview of Smart-seq2 protocol. Image from Picelli et al. [2014].

Single-cell transcriptomic analysis of murine gastrulation

Table 3.2 Nextera® XT DNA sample preparation kit primers. **i5 & i7** indicate the positions of the 8 base pair index tags. The underlined regions highlight the matching regions that the P5 or P7 primer will anneal to on the complementary strand, see fig. 3.4.

Name	Sequence
Read 1	5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3'
Read 2	5'- <u>GTCTCGTGGGCTCGG</u> AGATGTGTATAAGAGACAG-3'
P5 PCR primer	5'-AATGATACGGCGACCACCGAGATCTACAC[i5] <u>TCGTCGGCAGCGTC</u> -3'
P7 PCR primer	5'-CAAGCAGAAGACGGCATACGAGAT[i7] <u>GTCTCGTGGGCTCGG</u> -3'

added to each of the 96 wells for the enrichment PCR step producing dual-indexed libraries with Illumina sequencing adaptors, table 3.2 and fig. 3.4. There are 8 different indexed P5 (S502-S508,S517) adaptors and 12 different indexed P7 adaptors (N701-N712) thus yielding 96 unique dual-indices.

A further SPRI magnetic bead cleanup step yielded the final libraries that were quantified, diluted and pooled to achieve the target concentration of DNA required by the sequencing facility. This dilution step is critical for optimal sequencing by preventing under or overclustering on the sequencing flowcell as will be discussed in section 3.3.

3.3 Sequencing and Alignment

Sequencing was submitted by Dr Nicola Wilson to be performed by the sequencing service offered at the Cambridge Institute - Cancer Research UK. The 15 samples of 96 well plates were sequenced on the Illumina® HiSeq 2500 in high throughput mode with one sample per lane.

The libraries which were generated to incorporate the sequencing adaptor as described in section 3.2 were denatured and loaded on to separate lanes of the flowcells where the adaptor hybridised to a complementary surface bound sequence, see top panel of fig. 3.5. The bound molecules then underwent solid state bridge amplification to form clonal clusters of up to a thousand molecules, centre and bottom panels of fig. 3.5.

Loading libraries of the correct concentration was critical as having highly concentrated libraries loaded onto the flow cells would have resulted in a high density of clonal clusters making it difficult or impossible for the downstream image analysis software to differentiate between two different seeding molecules. In contrast, too low concentration libraries would have resulted in reduced efficiency and higher costs.

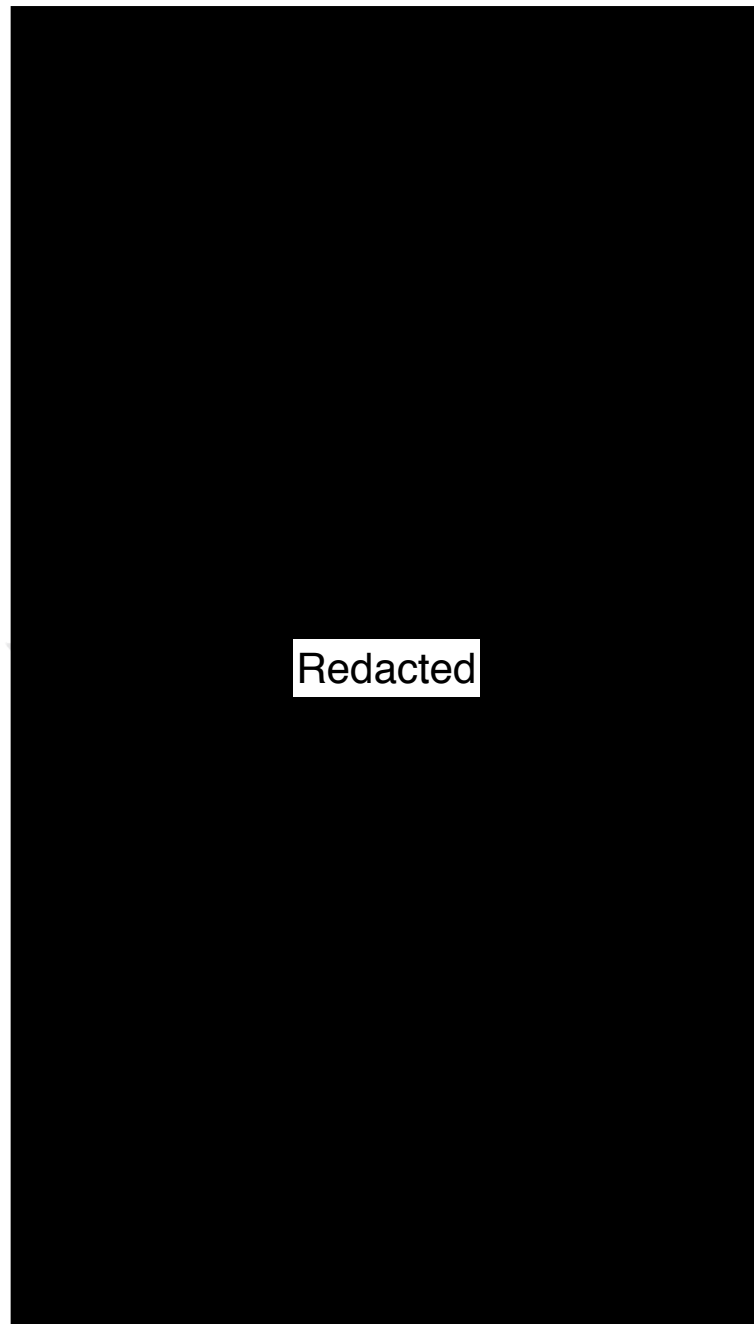


Fig. 3.5 Illumina® sequencing platform. **Upper:** Compatible libraries are loaded onto the flowcell and hybridise to a lawn of complementary sequences. **Middle and bottom:** Repeated cycles of bridge amplification generate large clonal clusters - neighbourhoods of identical DNA sequences. *Figures modified from https://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf and <https://www.cegat.de/en/services/next-generation-sequencing/> (accessed 15 October 2017).*

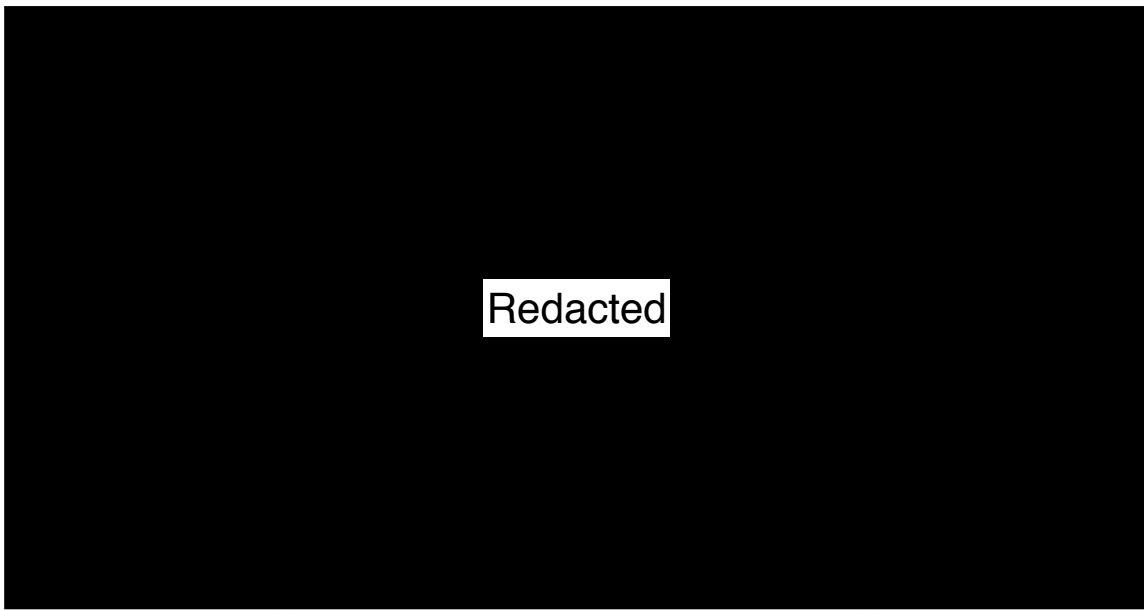


Fig. 3.6 **a**: At each cycle one of the four reversible fluorescent labelled terminator bound dNTPs are added by a DNA polymerase, images are taken, reagents washed off, fluorescent label and terminator cleaved and a new cycle initiated. This is done in parallel for all clusters making the process very high throughput. *Figure from https://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf. **b**: Fake colour composite image generated from individual channels. From http://www.dkfz.de/de/presse/pressemitteilungen/2012/images/120227_rippe_mertens_abbildung_pm.jpg (accessed 15 October 2017.)*

The Illumina platform chemistry performs sequencing by synthesis (SBS) using four reversible fluorescent labelled terminator bound dNTPs. At each cycle a DNA polymerase catalyses single nucleotide addition to the DNA template in parallel and fluorescent images are taken for each of the colour channels corresponding to the fluorescent label for each base, fig. 3.6. By doing this in parallel for millions of molecules per lane, very high-throughput is achieved. The reagents are washed off, the fluorescent label and terminator are cleaved and removed and the cycle repeated until the required length of read is generated.

Initially the P7 bound fragments are cleaved leaving only the P5 bound bridge amplified molecules. The primers for i5 index, read1 and i7 index are added and sequencing by synthesis performed in turn. For paired end reads a further phase of bridge amplification is performed and on this occasion the P5 bound molecules are cleaved leaving the P7 bound molecules, now the read2 primer is added and the second read sequenced.

The sequencers perform image analysis and base calling producing raw *bcl* files. These are then demultiplexed using the library barcodes into individual *fastq* files, one per cell. The

Single-cell transcriptomic analysis of murine gastrulation

aligned *bam* files generated were then passed to HTSeq to produce gene-level transcript counts for each cell [Anders et al., 2015].

Subsequent analyses were predominantly performed in the R language and environment for statistical computing [R Core Team, 2017]. For reproducibility all functions have been extensively documented and packaged into *bglab* which is freely available online at <https://www.github.com/wjawaid/bglab>.

3.4 Quality Control

1440 libraries were generated from the fifteen 96-well plates. Some of the wells may have no cell, a damaged cell or a highly quiescent cell resulting in poor libraries. Down stream analyses of data such as normalisation and selection of highly variable genes can be unduly affected by these poor quality libraries which effectively are treated as cells at this stage. To prevent this, poor quality cells are removed based on empirical cut-off values. To remain consistent, where possible these cutoff values should be maintained across all experiments.

Though highly quiescent cells can make normalisation and other downstream analyses challenging it may be necessary to reappraise their removal as they could have great biological significance. An example is long lived multipotent stem cells of mature adult systems maintaining homeostasis in a steady state.

Of the possible 1440 cells 1205 passed QC criterion, remarkably this included a cell corresponding to a well that was marked as empty during FACS. This highlights the risk of inadvertently making libraries from cell doublets, when employing a strategy of refilling wells reported by the sorter to be empty.

Box 3.2: Quality Control

Three QC criteria were used for each cell (library):

- > 200,000 mapped reads.
This appears to be a reasonable cut-off to remove libraries that may be generated from wells containing either damaged or fragmented cells.
- > 20% of mapped reads map to either nuclear or mitochondrial genes.
The remaining reads would either be unmapped or map to ERCCs. Most unmapped reads are often TSO concatemers or primer dimers.
- < 20% of mapped reads map to mitochondrial genes.
Damaged cells are found to have higher mitochondrial mapping genes.

For each of the 96 well plates (output from one lane of the sequencer) multiple plots were generated during QC, an example is given in fig. 3.7. In all the subplots each point represents a cell and the abscissa represents the total number of reads from the library generated from that cell. This includes all mapped and un-mapped genes and also those mapping to ERCC spike-ins. In the example shown in fig. 3.7 only 2 cells failed due predominantly to a low number of mapped reads; the red dashed lines show the thresholds as declared in box 3.2. These two samples likely represent severely damaged cells, cell fragments or small highly quiescent cells providing low input RNA for the assay, so that primer dimers and TSO concatemers are preferentially generated during cycles of PCR.

Table 3.4 Descriptions of QC parameters plotted in fig. 3.7

Feature	Description
Total reads	All reads mapped and unmapped
Mapped reads	Reads mapped to native mRNA or spike-ins
Nuclear reads	Reads mapped to nuclear annotated genes
Genes	Proportion of reads mapped to native mRNA relative to total
No feature reads	Reads not mapping to an annotated genomic region
Ambiguous reads	Reads that map to a single locus but overlap two annotated regions
Low quality reads	Reads with poor low quality
Unaligned reads	Reads not mapping to anything in the reference genome including spike-ins
Alignment not unique	Reads mapping to multiple locations
High coverage reads	Number of genes with reads greater than 10 reads per million for that cell
Mitochondrial reads	Mapping to mitochondrial reads
ERCC	Reads mapping to spike-ins

Single-cell transcriptomic analysis of murine gastrulation

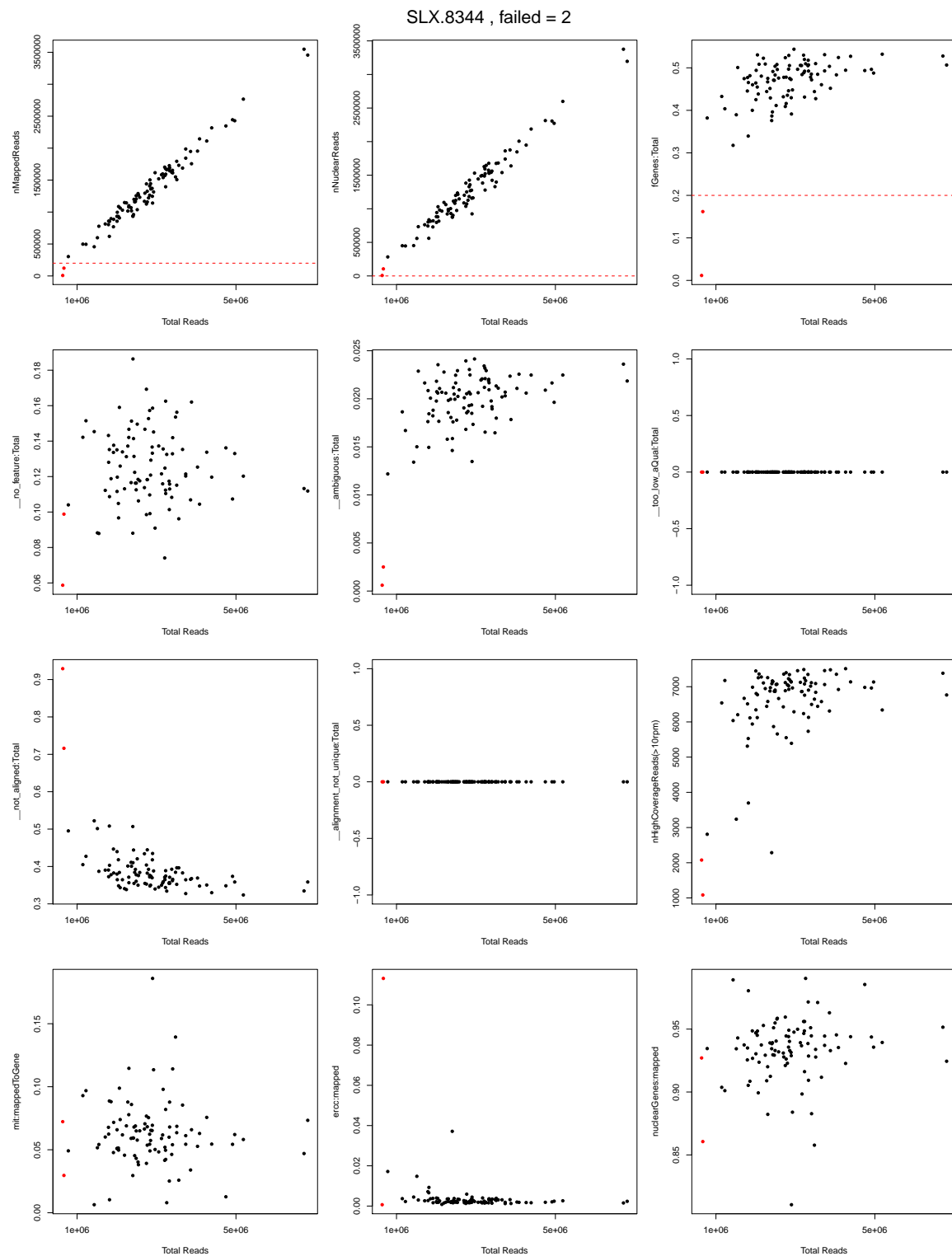


Fig. 3.7 Several plots of parameters used to assess quality, further details in table 3.4. Each point represents a cells and the red dashed line represents QC thresholds, box 3.2.

3.5 Normalisation

The libraries generated are only pooled at the final steps and libraries from different 96 well plates are processed completely separately. Technical variability in single cell RNA can be high due to the low amount of starting RNA and variations in capture efficiency. Furthermore capture of longer mRNA transcripts is known to be higher than for short transcripts one reason being there is a greater probability of two Tn5 transposomes attacking a long transcript and randomly annealing a P5 on one end and P7 on the other. This makes comparison of abundancies between genes uninformative.

An important part of downstream analysis is comparing transcript counts between cells but this can only be achieved if the libraries have been adequately normalised. Some methods adopted from bulk RNAseq include FPKM, TPM, TPKM and the DESeq size factor [Conesa et al., 2016; Love et al., 2014; Risso et al., 2014]. Newer methods have become available since, specifically targeting single cell analyses such as scran, SCDE and more advanced but computationally intensive algorithms such as SCnorm [Bacher et al., 2017; Kharchenko et al., 2014; Lun et al., 2016a].

In Scialdone et al. [2016], DESeq size factor normalisation was used and works well for Smart-Seq2 protocol single cell RNAseq generation where the *dropout* effect is not as conspicuous as for other protocols. For the purposes of this analysis scran an adaptation of the DESeq size factor algorithm that incorporates pooling in an attempt to overcome the effect of *dropout* was used. In the case of this dataset there is little difference between normalisation by DESeq size factor or scran.

Given an array of counts x for m genes G and for n cells C of dimensions $m \times n$ the geometric mean gm is calculated for every $g \in G$ as

$$gm_g = \left(\prod_{c=1}^n x_g^c \right)^{\frac{1}{n}}, \quad \text{for } c \in C.$$

Then the size factor f is calculated for every cell as

$$f_c = \sum_{j \in G^*} \frac{x_j^c}{gm_j}, \quad \text{for } G^* \subset G, \quad \forall j \in G^* : gm_j \neq 0.$$

The size factor therefore is only calculated using genes that are expressed in every single cell, any gene that has *dropped out* in even one cell will be excluded from the calculation.

Single-cell transcriptomic analysis of murine gastrulation

Where dropouts are common this becomes problematic so Lun et al. [2016a] have modified the DESeq size factor calculation method by repeated systematic pooling of similar cells based on their cosine distance and solving the resulting linear system of equations. These normalisation methods do not account for gene length nor is there any sort of gene wise normalisation, so counts between genes cannot be reliably compared. Furthermore the methods used or discussed here do not use spike-ins for normalisation, such methods may provide superior results [Bacher et al., 2017; Ding et al., 2016].

3.6 Feature selection

Data generated from single cell RNAseq is large and high-dimensional. The gene counts matrix is rather sparse and as such suffers from the *curse of dimensionality*. Despite the large numbers of cells, the feature space sampled diminishes to insignificance, in comparison to the vast volume of space spanned by the high number of annotated genomic features. This becomes highly problematic when calculating distances [Adachi, 2017]. Many of the genomic features may not carry any usable information and therefore simply add *noise* to calculations. To somewhat mitigate against this we planned to use a feature selection method that estimates technical noise from spike-ins and only selects genomic features that vary beyond the estimated technical variance [Brennecke et al., 2013].

Spike-ins were added to the samples for technical noise estimation but an underlying assumption is that equal amounts of ERCC spike-ins are titrated into each sample [Brennecke et al., 2013]. To assess this assumption, the ERCC to mapped read ratios were plotted for all cells by sequencing flowcell lane equivalent to each 96 well plate, fig. 3.8. From fig. 3.8 it is clearly evident that there is significant variation in the amount of ERCC spike-in for each 96 well plate and this assumption does not hold. Spike-ins were therefore not used to infer technical noise but a mean CV-squared¹ relationship was inferred on the data itself [Brennecke et al., 2013]. 1937 genes out of 41,388 annotated genomic features were found to vary beyond the technical limit of detection in the 1205 cells sampled, fig. 3.9 and were carried forward for further analysis. Though this was a stringent choice, it was only for purposes of dimensionality reduction and identifying clusters. This set of genes or other sets similarly calculated may be referred to as *highly variable genes* in the remainder of the text.

¹coefficient of variance squared = $\frac{\sigma^2}{\mu^2}$

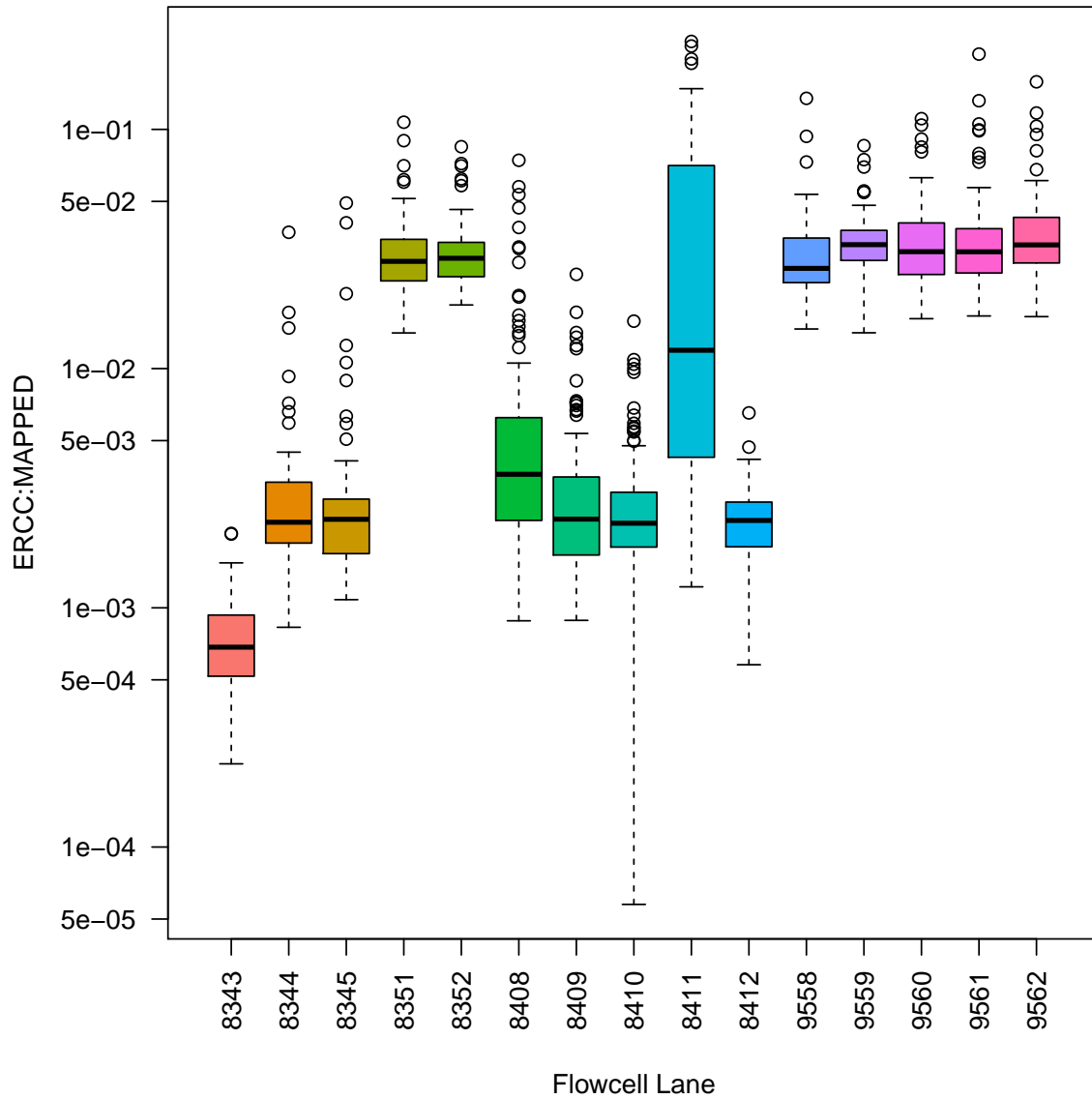


Fig. 3.8 Ratio of reads allocated to the ERCC-92 spike-ins to total mapped reads. Boxplots comparing ERCC ratios for each flowcell. Ordinate: Ratio ERCC to mapped reads on a log scale. Abscissa: Each flowcell is represented by a box plot. There is significant variability in the ratio of ERCC spike-in to biological material.

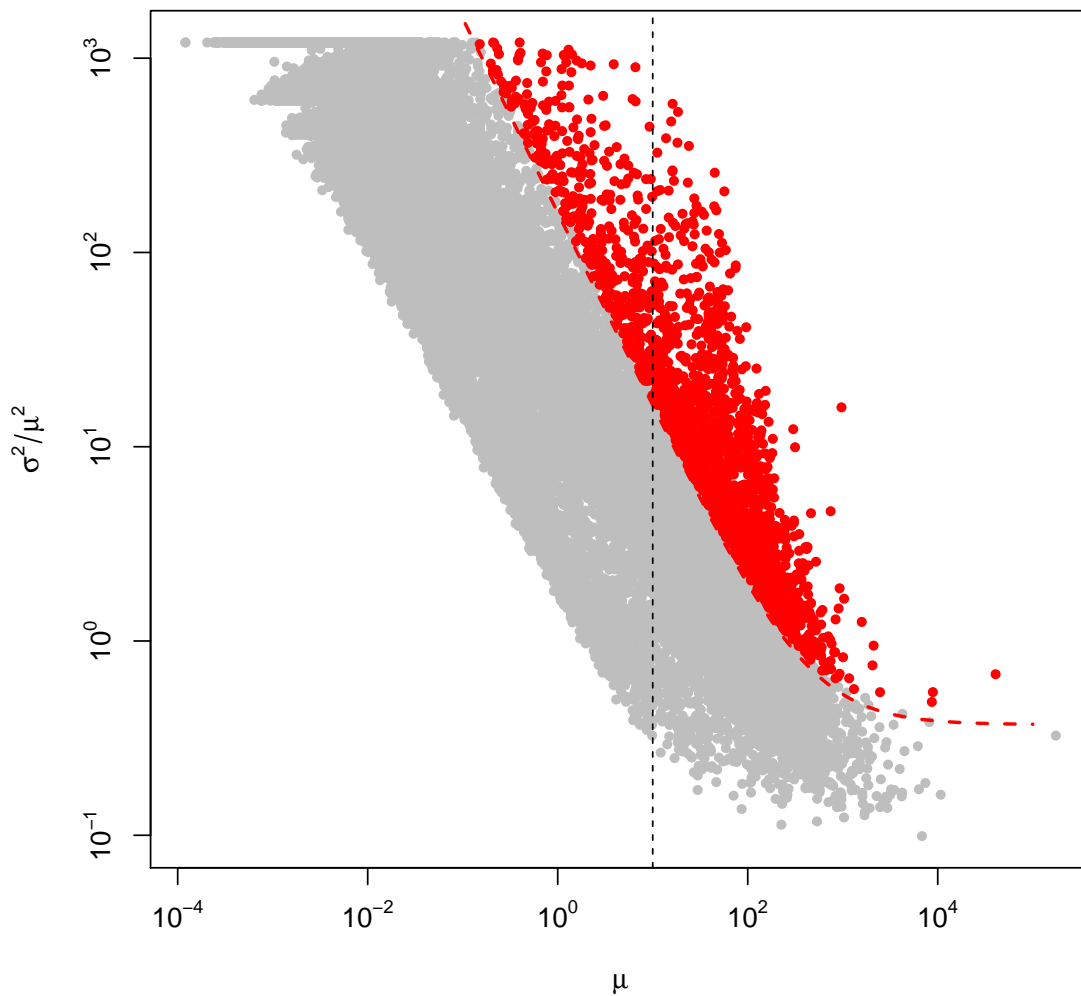


Fig. 3.9 Mean normalised count against CV^2 for each gene. Red coloured dots represent genes that vary beyond the expected technical detection limit of the experiment (10% FDR), grey coloured dots are not. Red dashed line shows threshold and data points to the right of the vertical black dotted line were used for parameter fitting.

3.7 Identification of cell populations

3.7.1 Hierarchical clustering

Identification of cell types within the sampled population requires some form of clustering or grouping of similar cells together separate from others. Hierarchical clustering involves building a clustered tree or dendrogram of cells so that those most similar are closest together nested within larger groups. Hierarchical clustering can be performed either ‘bottom-up’ also known as ‘agglomerative’, merging the most similar cells (clusters) together iteratively or ‘top-down’ also called ‘divisive’, separating a single group into two subgroups, again repeated iteratively. Clustering algorithms require some notion of similarity (or dissimilarity) between cells and most employ an agglomerative approach, box 3.3.

Box 3.3: Agglomerative hierarchical clustering

1. Each cell is assigned to its own cluster
2. The closest clusters (cells in the first step) are merged into a single cluster, so there is now one less cluster
3. New similarities (or dissimilarities) are calculated for the newly merged cluster.
4. The steps are repeated until all cells are in one cluster.

Various methods of recalculating dissimilarities from consolidated clusters in step 3 above, have been described including single linkage, complete linkage, group averaging, weighted averaging, centroid/median or Ward’s method. Single linkage calculates the new similarity as that between the closest cells (points in high dimensional space) in the clusters having the disadvantage that it can lead to clusters being merged prematurely. In contrast, complete linkage calculates the new dissimilarity as that between the furthest points in the clusters, now causing delayed merger of clusters. Adachi [2017] has presented a single parameterised equation that encompasses the calculation of all these linkage methods:

$$d(C_i \cup C_j, C_k) = \alpha_1 d(C_i, C_k) + \alpha_2 d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|,$$

where C_i and C_j are the clusters that form the new merged cluster $C_i \cup C_j$, C_k is either a cell or a previously merged cluster, with a function $d(a, b)$ that calculates dissimilarity between a and b . α_1 , α_2 , β and γ are parameters as set out in table 3.5.

Single-cell transcriptomic analysis of murine gastrulation

Table 3.5 Parameters set out in linkage equation from Adachi [2017]. n_i, n_j, n_k are numbers of items in each cluster.

method	α_1	α_2	β	γ
single	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
complete	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
group average	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
weighted average	$\frac{1}{2}$	$\frac{1}{2}$	0	0
centroid	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
median	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{4}$
Ward	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$\frac{n_j+n_k}{n_i+n_j+n_k}$	$\frac{-n_k}{n_i+n_j+n_k}$	0

Though hierarchical clustering generates a dendrogram of relatedness it does not generate discrete clusters per se; to achieve this the dendrogram must be suitably ‘cut’. Hierarchical clustering using group average linkage and a dynamic tree cut [Langfelder et al., 2008] were performed by Dr Antonio Scialdone and robust parameters were selected by testing on 100 bootstrapped samples [Scialdone et al., 2016]. The heatmap in fig. 3.10 shows the clusters with a handful of genes that are most highly uniquely differentially expressed.

Figure 3.10 shows the clusters and some selected genes but it is useful and re-assuring to take an overall view (all highly variably expressed genes) and directly assess the clustering dendrogram. This can be seen by repeating the clustering on a dissimilarity matrix calculated on the Spearman rank correlation using *R*’s *hclust* function with *Ward*’s linkage agglomerative clustering method [Murtagh and Legendre, 2014] combined with a fast leaf re-ordering algorithm [Bar-Joseph et al., 2001]. Adding an indicator of the clusters inferred from the dynamic tree cut algorithm shows consistency with the dendrogram calculated this way, fig. 3.11. It is re-assuring that this alternate strategy reproduces similar clusters and only cluster 5 is not so clearly defined but diffusely interposed between clusters 4 and 7.

3.7.2 Assigning identities to clusters

Cluster assignment was a complex task and was performed chiefly by Yosuke Tanaka and myself [Scialdone et al., 2016]. This was performed by identifying differentially expressed genes using a variety of methods including DESeq2 and by visualising the distribution of gene expression on the dimensionality reduction tSNE plot, see section 3.7.4. Putative cell

3.7 Identification of cell populations

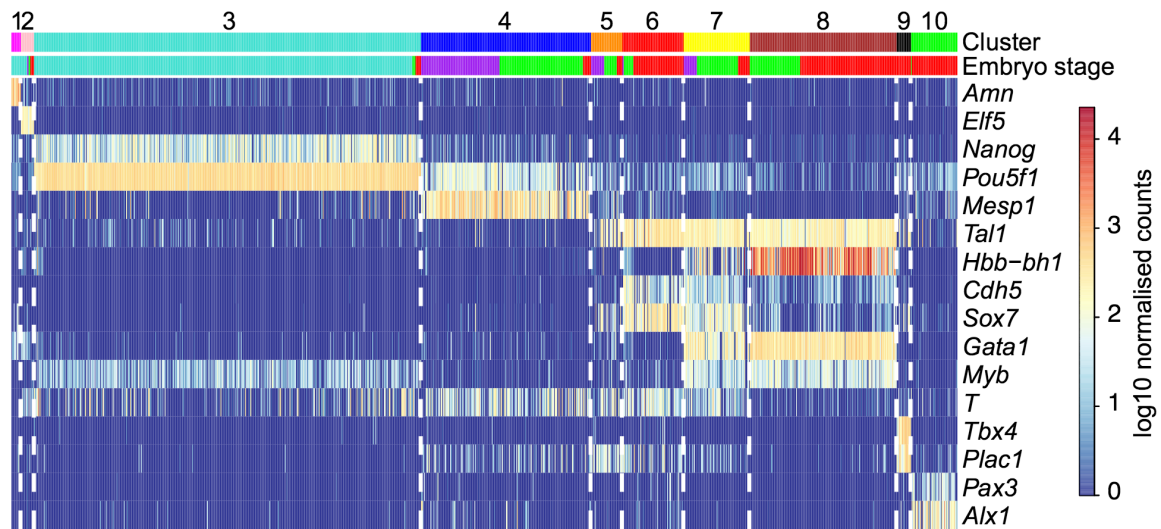


Fig. 3.10 Heatmap with clusters identified by a dynamic tree cut performed by AS on all 1205 cells and 1937 highly variable genes. Bootstrapping was used to select parameters. Upper horizontal bar defines cluster as found by dynamic tree cut and lower horizontal bar depicts the stage of the embryo from which the cell was harvested. Image was produced by Vicki Moignard.

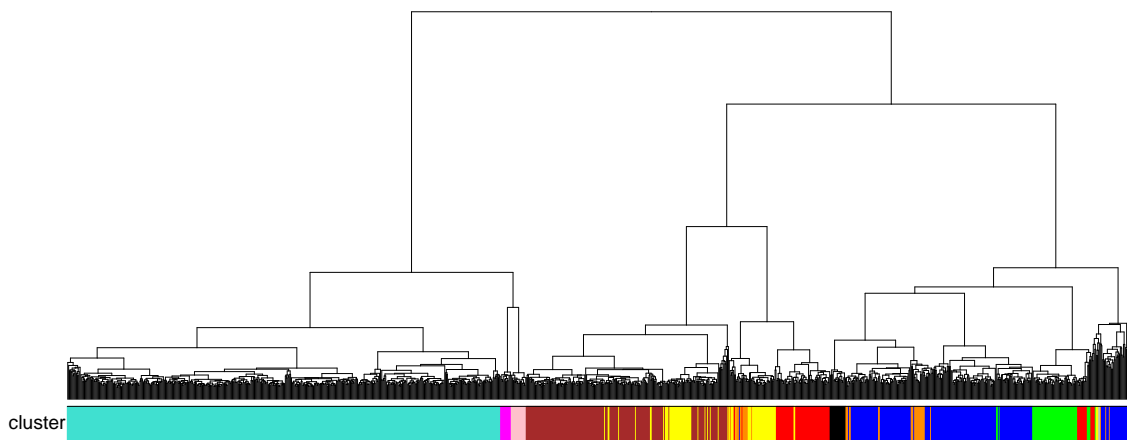


Fig. 3.11 Hierarchical clustering of all 1205 cells using all 1937 highly variable genes shows consistency with the dynamic tree cut. The horizontal bar corresponds to clusters in fig. 3.10

types were assigned and refined through comparison with available in-situ hybridisation and immunofluorescence experiments. The similarity between cell types also provided clues to the biological identity of the cells. A web interface was developed and is provided online so that gene expression of user selected genes can be viewed on a dimensionality reduction plot <http://gastrulation.stemcells.cam.ac.uk>.

Single-cell transcriptomic analysis of murine gastrulation

Cluster Number	Cluster colour	Main stage	Assigned ID	Key genes & references
1	Magenta	E6.5	Visceral endoderm	<i>Amn</i> , <i>Cubn</i> [Du et al., 2010]
2	Pink	E6.5	Extra-embryonic endoderm	<i>Essrb</i> [Mitsunaga et al., 2004], <i>Elf5</i> [Donnison et al., 2005]
3	Turquoise	E6.5	Epiblast and primitive streak	<i>Nanog</i> [Morkel et al., 2003], <i>Sox2</i> [Wood and Episkopou, 1999]
4	Blue	PS/NP	Mesodermal progenitors	<i>Mesp1</i> [Saga et al., 1999], <i>Mixl</i> [Pearce and Evans, 1999], <i>Wnt3</i> [Du et al., 2010]
5	Orange	NP	Posterior mesoderm	
6	Red	HF	Endothelium (Yolk sac)	<i>Sox18</i> [Pennisi et al., 2000], <i>Pecam1</i> [Baldwin et al., 1994], <i>Lyve1</i> [Gordon et al., 2008; Padrón-Barthe et al., 2014]
7	Yellow	NP	Blood progenitors	<i>Tal1</i> [Kallianpur et al., 1994; Robb et al., 1995], <i>Runx1</i> [North et al., 1999; Tanaka et al., 2012, 2014]
8	Brown	HF	Primitive erythroid cells	<i>Hbb-bhl</i> [Palis et al., 1995], <i>Klf5</i> [Drissen et al., 2005; Southwood et al., 1996], <i>Gata1</i> [Silver and Palis, 1997]
9	Black	HF	Allantois	<i>Pitx1</i> [Lanctôt et al., 1997], <i>Tbx4</i> [Naiche et al., 2011], <i>Coll1a1</i> [Tamplin et al., 2008]
10	Green	HF	Pharyngeal mesoderm	<i>Tbx1</i> [Lania et al., 2015], <i>Prdm1</i> [Vincent et al., 2014]

Table 3.6 Collation of evidence for assigning cell types to data-driven cluster identification. Table and references reproduced from that collated by Victoria Moignard.

This assignment of cell type was critical for attributing biological meaning to the single cell data and combined with dimensionality reduction hinted at tentative relationships between cell populations.

The approach taken here is manually intensive and highly supervised but if this turns out to be an accurate ‘*curation*’ of cell states it will be possible to assign cells/samples sequenced in future to clusters using an automated bioinformatic approach.

3.7 Identification of cell populations

Cell cycle is a fundamental biological process but we do not expect our clusters at this embryological stage to be solely driven by this process. The clustering of cells does not appear to be grossly cell cycle driven, fig. 3.12. There are some notable differences but the small numbers of cells in some clusters preclude any definitive differences about variations in cell cycle between different populations to be ascertained. Given that the algorithm used was developed on ESCs it was felt that this algorithm was not sufficiently robust to pursue this avenue of enquiry.

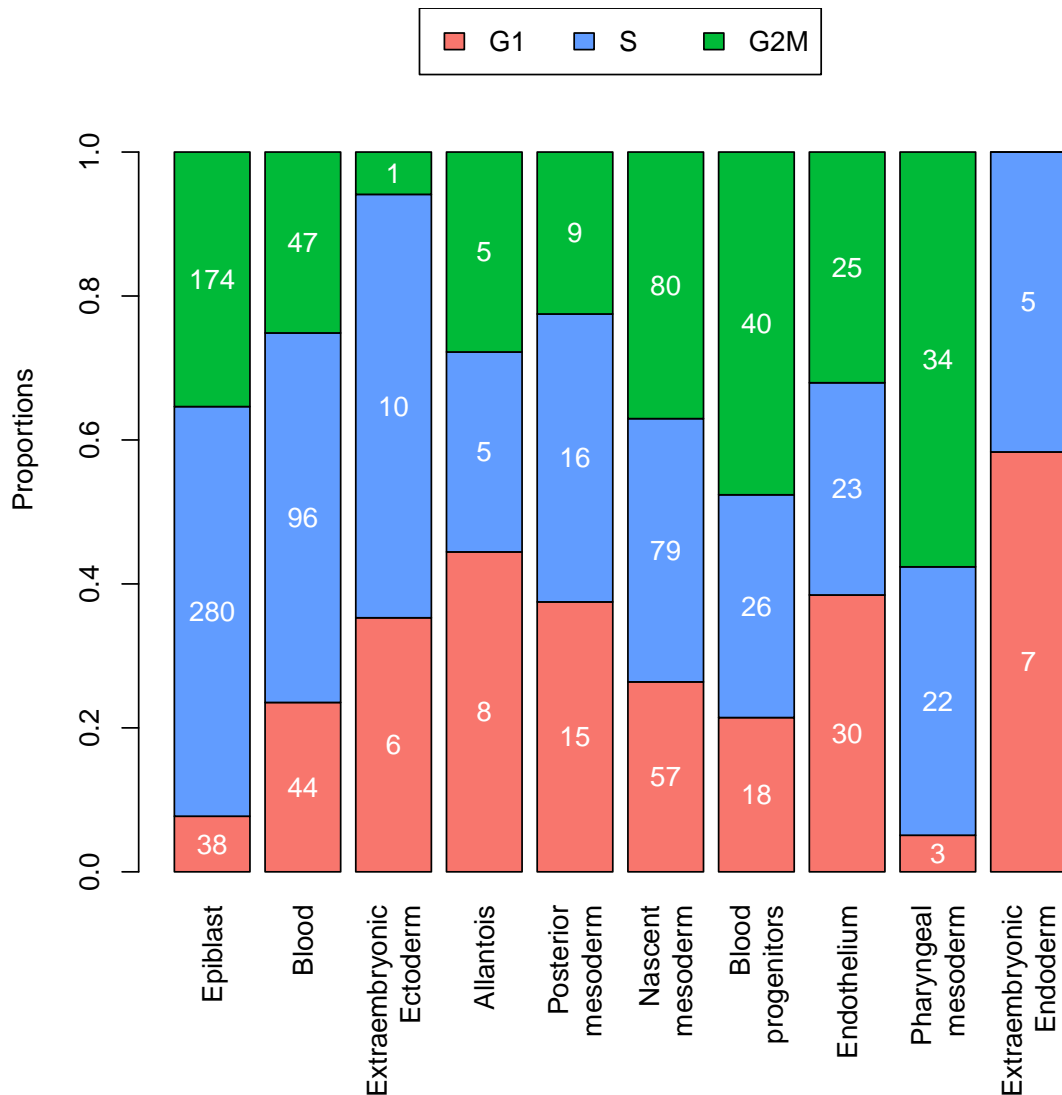


Fig. 3.12 Proportions of each annotated cell population assigned to different stages of the cell cycles. Numbers in bar give the actual number of cells. Cell cycle assigned by using the ‘cyclone’ function in the R package ‘scran’.

3.7.3 PCA

A useful and commonly deployed method of assessing the overall structure of the data is *principal component analysis* (PCA). Essentially the *scaled* and *centered* gene count data set is considered to *span* a high-dimensional space. *Eigenvectors* and *eigenvalues* are calculated on the covariance matrix to find a *linear transformation* of the data that maximises variance in the basis vectors. Dimensionality reduction to an n -dimensional representation is achieved by retaining eigenvectors corresponding to the largest n eigenvalues. Eigenvectors are unit vectors and the eigenvalues give an indication of the proportion of variability explained by the corresponding eigenvector.

The 2-dimensional PCA representations of the first 5 eigenvectors of the dataset are displayed in fig. 3.13 with cells coloured according to the previously defined clusters. The different components pick out disparate cell types. PC_1 does not appear to pick out any biologically meaningful aspect of the data nor is there any trend relating to either: the embryo, the stage of the embryos, forward scatter, side scatter, DAPI intensity or FLK1/CD41 fluorescence. Some lanes of the flow cell appear to have a trend towards more negative values for PC_1 , fig. 3.14, giving a clue that this may be related to some technical artefact during library preparation. To investigate further PC_1 was correlated against the number of mapped reads ($\rho = 0.862$, $p < 2.2 \times 10^{-16}$) and ratio of unaligned to total reads ($\rho = -0.606$, $p < 2.2 \times 10^{-16}$), fig. 3.15. The relationships are non-linear but monotonic therefore strength of the association was tested using Spearman's rank correlation. Both the number of mapped reads and the ratio of unaligned to total reads act as proxies for the amount of input RNA; so PC_1 appears to correlate most strongly with either the amount of RNA in the cell or at least the efficiency of the initial reverse transcriptase step of library production. This is additionally supported by the variation in ERCC:mapped ratio displayed in fig. 3.8. Where input RNA is low, mapped reads as a proportion of total reads are much lower than would be expected and most unaligned reads are generated from either TSO concatemers or PCR primers.

PC_2 segregates the epiblast cluster in particular, PC_4 the endothelium, PC_5 picks out both the extraembryonic endoderm and extraembryonic ectoderm clusters and PC_3 distinguishes the remaining clusters with the blood cluster having low values in PC_3 while the nascent mesoderm having high values. The consistent separation of clusters on PCA as defined by hierarchical clustering and dynamic tree cut provides additional re-assurance for the robustness of the analysis.

An advantage of PCA is that it is a linear transformation with weights attributed to each feature in this case genes for each component. These weights provide a means of visualising the genes driving the separation in the corresponding principal component in way of loading

3.7 Identification of cell populations

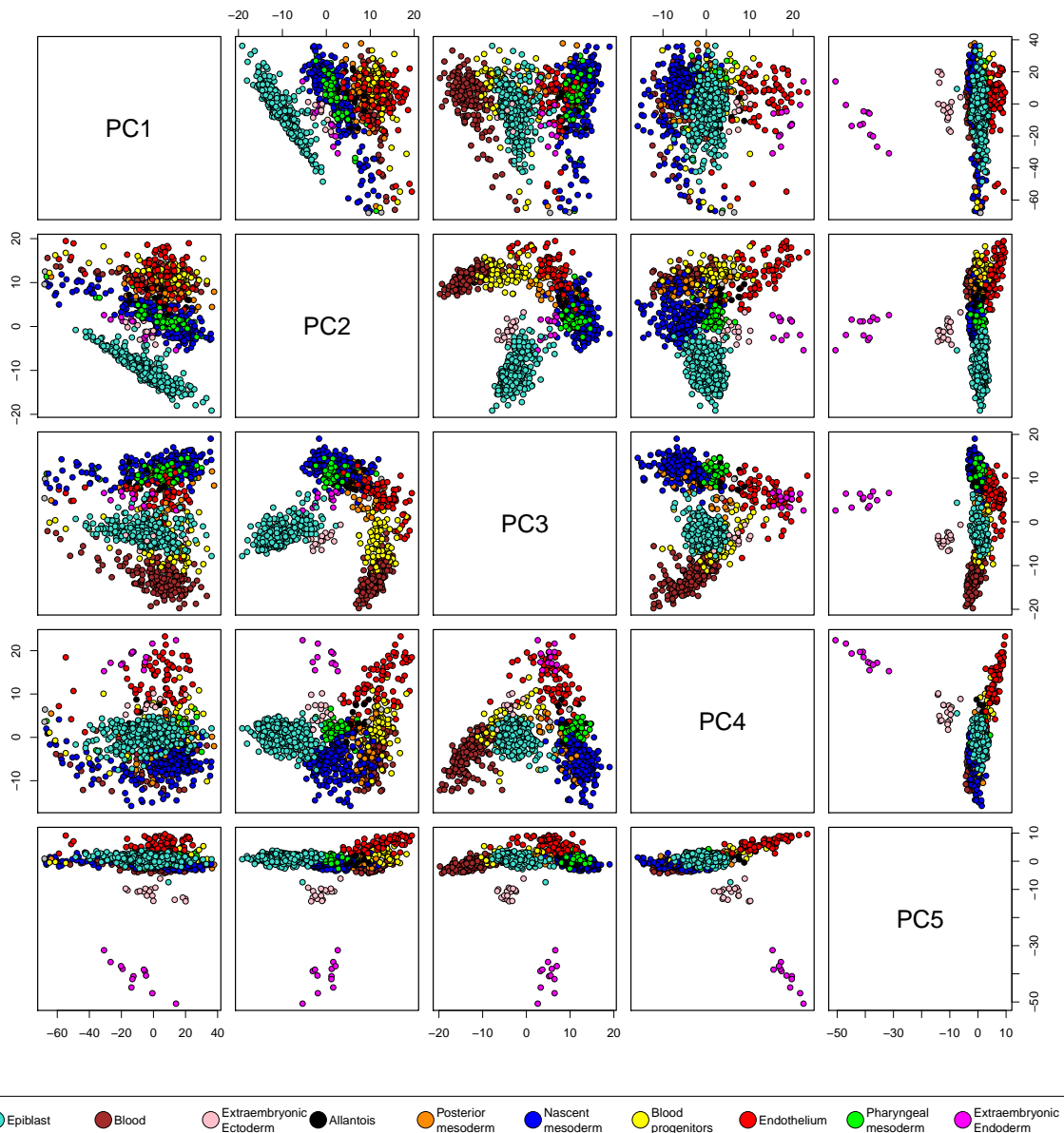


Fig. 3.13 Pairs plot of the first 5 principal components of 1205 cells using 1937 highly variable genes.

plots. Cells allocated to the putative E6.5 epiblast cluster have lower PC_2 coordinates than others, fig. 3.13. In fig. 3.16 the plot on the right focuses on the region highlighted in the plot on the left. This indicates that the genes driving PC_2 include *Cdh1* (*E-cad*), *Otx2*, *Tdgl1* (*Cripto*), *Pou5f1* (*Oct4*), *Trh*, *Fgf5* and *Dnmt3b* amongst others. *E-cad* is not only expressed in the epiblast but has been shown to play a critical role in the epiblast maintaining pluripotency and preventing the epiblast from deliquescing into a mesenchymal state [Burdal

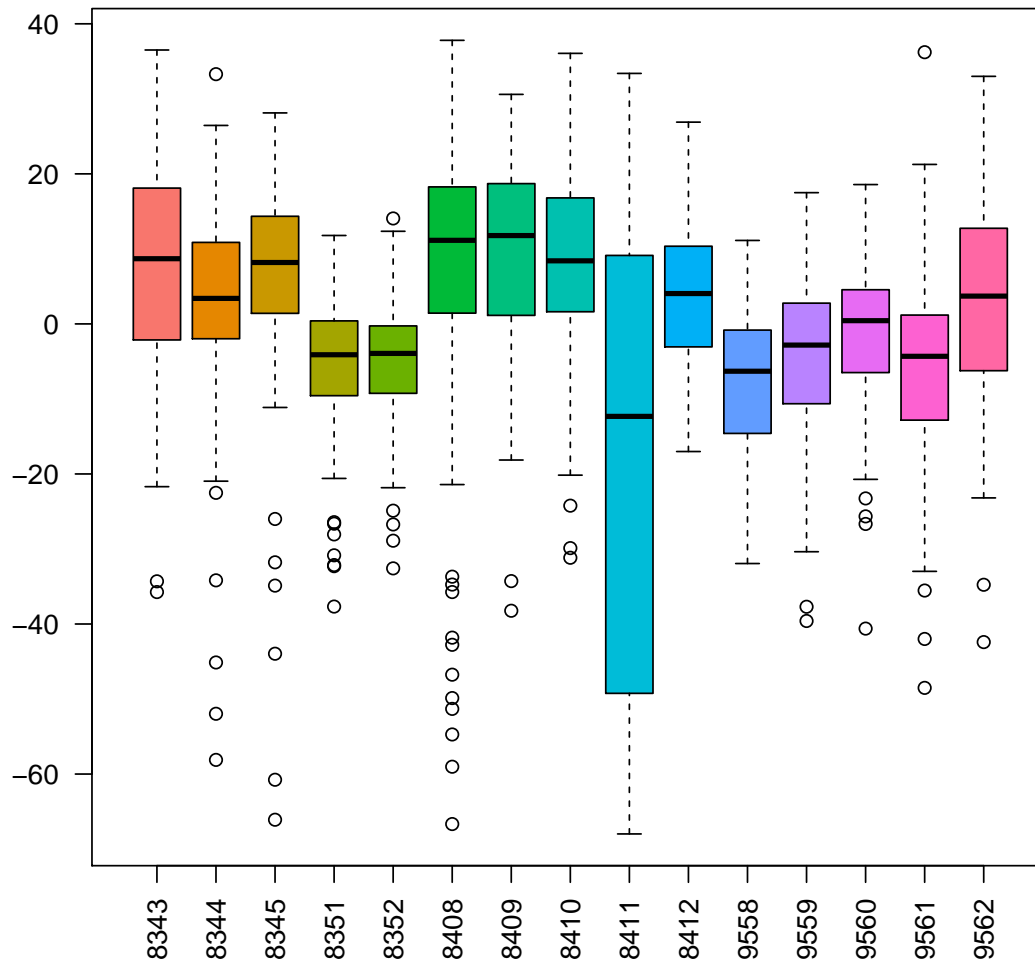


Fig. 3.14 Box plot of *PC1* coordinates by lane of sequencing flow cell.

et al., 1993; Götz et al., 2000; Ohtsuka et al., 2012; Sumi et al., 2013]. *Otx2* and *Cripto* are expressed in the early epiblast and have a role in translating the proximal-distal axis to an anterior-posterior axis [Boyl et al., 2001; Ding et al., 1998; Fiorenzano et al., 2016; Kimura et al., 2001]. *Oct4* a marker of pluripotency is highly expressed in the E6.5 epiblast later becoming restricted to primordial germ cells (PGCs) [Song et al., 2016]. Thyrotropin releasing hormone *Trh*, known most widely for its activity on the hypothalamic pituitary thyroid axis is dynamically expressed through development and is expressed in the E6.5 epiblast [McKnight et al., 2007]. *Fgf5* is a known marker of the post-implantation epiblast and a marker of epiblast stem cells [Khoa et al., 2016]. DNA methyltransferase 3B *Dnmt3b*

3.7 Identification of cell populations

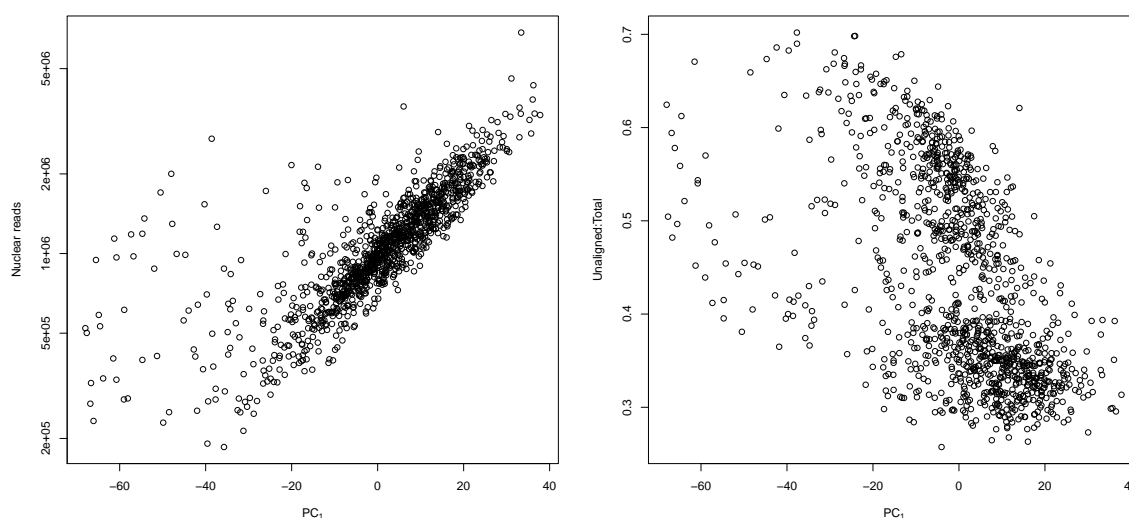


Fig. 3.15 Relationship between the first principal component coordinate of each cell with the number of mapped nuclear reads (left) or the ratio of unmapped to total reads (right). The nuclear mapped reads are plotted on a log scale.

encodes a key regulatory protein involved in *de novo* methylation of cytosine residues in DNA chiefly at CpG islands thereby modulating gene expression and mutations have been causally associated with human disease [Xie et al., 2006]. The post implantation embryo undergoes a wave of demethylation re-establishing a developmentally primed methylation pattern from the hypomethylated blastocyst state [Monk et al., 1987; Smith et al., 2017]. Expression of *Dnmt3b* correlates with increasing methylation patterns in the E6.5 epiblast [Auclair et al., 2014]. The preceding synopsis gives a brief introduction to how cell types were annotated by gradually collating supporting evidence from the literature and visualising gene expression in clusters primarily using tSNE dimensionality reduction section 3.7.4.

Looking at genes driving PC_2 in the positive direction it is clear that haematopoietic genes play a major role with expression of *Tal1*, *Fli1*, *Lmo2* amongst many others. Another key set of genes that have not previously been implicated in developmental haematopoiesis but play key roles in T-cells include *Nfatc1*, *Skap1*, *Cd59a*. Their role in developmental haematopoiesis though beyond the scope of the current work, warrants further attention particularly due to their recognised role in inter-cellular signalling.

Single-cell transcriptomic analysis of murine gastrulation

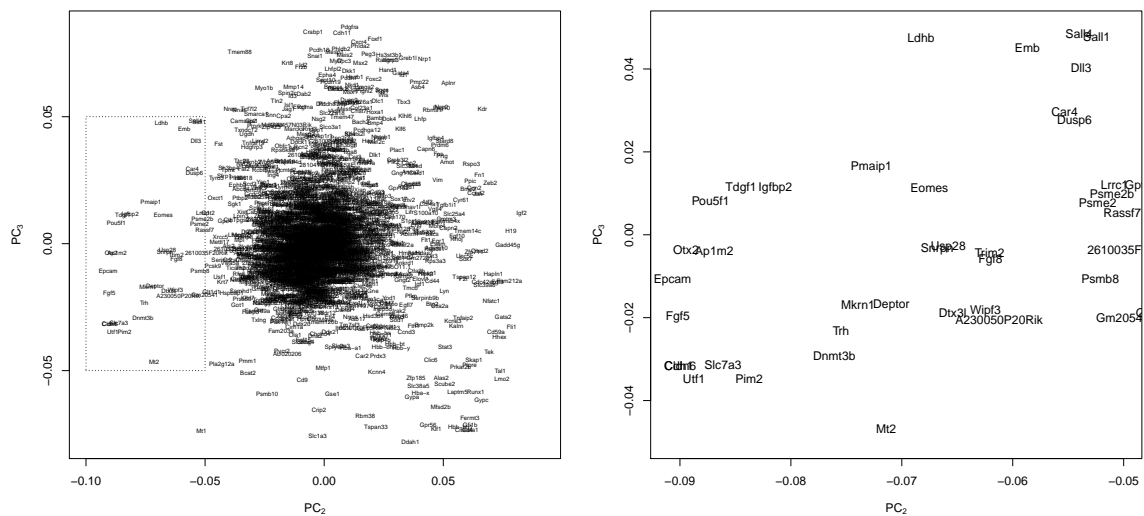


Fig. 3.16 PCA loadings plot for PC_2 and PC_3 . Left panel shows all highly variable genes. The dotted box indicates the zoomed in plot that is shown in the right panel. (*Cldnb* and *Cdh1* are overlapping at (-0.09,-0.03))

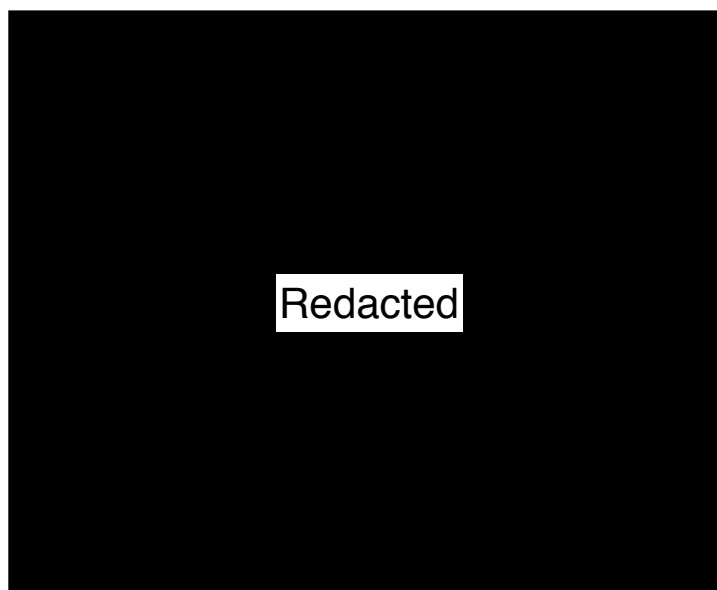


Fig. 3.17 Evidence collated from published literature showing the E6.5 epiblast localisation of gene/protein expression for genes identified by the PCA loadings plot fig. 3.16. All images are of E6.5 mouse embryos. RNA in-situ hybridisation images for *Oct4*, *Trh*, *Cripto*; immunohistochemistry for OTX2, E-CAD and immunofluorescence for DNMT3B and FGF5 [Auclair et al., 2014; Boyl et al., 2001; Burdsal et al., 1993; Ding et al., 1998; Fiorenzano et al., 2016; Götz et al., 2000; Khoa et al., 2016; Kimura et al., 2001; McKnight et al., 2007; Ohtsuka et al., 2012; Song et al., 2016; Sumi et al., 2013; Xie et al., 2006].

3.7.4 tSNE

PCA is a linear transformation and reveals some structure to the data fig. 3.13 but it cannot capture a non-linear manifold in the high dimensional space to which the data may be constrained. The non-linear mapping technique *t-distributed stochastic neighbourhood embedding* tSNE, reduces a high-dimensional dataset to a given number of lower dimensions maintaining as far as possible the same distance relations between samples in this case cells. Mathematical details of how this is performed will be discussed in section 5.3. In this case the data is reduced to a 2-dimensional representation, figs. 3.18a and 3.18b.

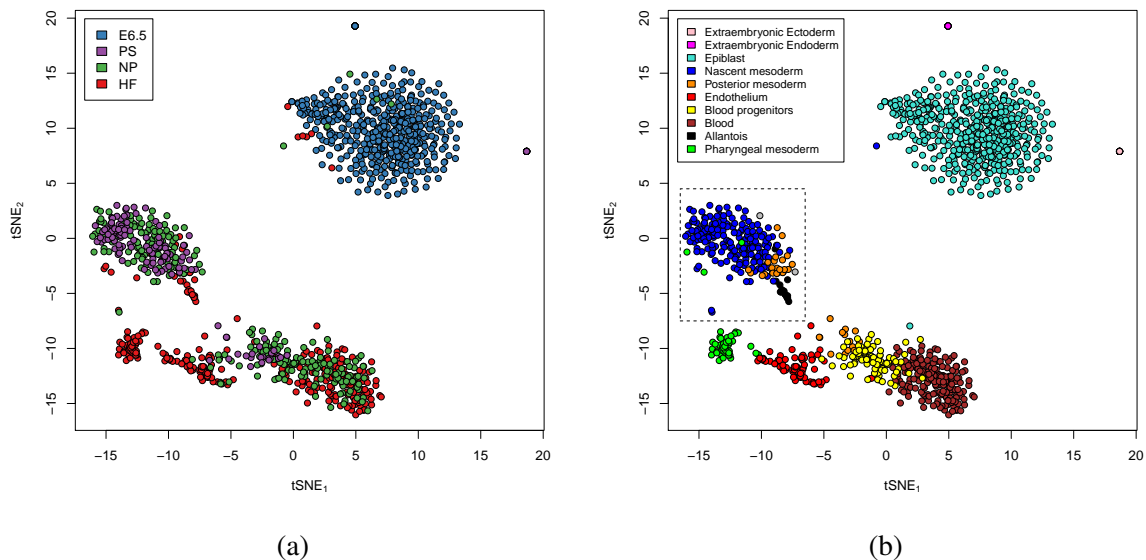


Fig. 3.18 tSNE dimensionality reduction of all cells and highly variable genes. Coloured by the stage of the embryo (a) or by previously defined clusters (b).

Data driven clusters identified from hierarchical clustering with dynamic tree cut are consistently co-localised in the tSNE even more so than with PCA, other than the posterior mesodermal cluster (orange). This is split in two between the nascent mesoderm in blue and the blood progenitors in yellow.

Not only do they separate well but their spatial arrangement imply biologically meaningful relationships. At the sampled point in development blood and endothelium are predominantly located in the yolk sac. The allantois is developing from the posterior of the embryo and will later also contribute to hematopoiesis and through vasculogenesis and angiogenesis develop the prospective umbilical vessels. Additionally the blood progenitor (yellow) and blood (brown) clusters are closely related suggestive of developmental progression, section 3.10.

Single-cell transcriptomic analysis of murine gastrulation

Though unsupervised dimensionality reduction techniques provide a powerful tool for visualising the data and allude to biologically meaningful relationships, prior biological knowledge is required to assert for example spatial or temporal relations. These observations advocate taking a more nuanced and cautious approach when assigning biological meaning to the data. Despite this and although cluster identity was described earlier in the text in table 3.6, tSNE dimensionality reduction was indispensable for ascribing identity to clusters. It was extensively used to plot gene expression gradients to quickly visually determine the regions of tSNE with high expression and so identifying associated clusters as seen in fig. 3.19. Plotting gene expression on tSNE permitted rapid comparison with published RNA in-situs, immunohistochemistry or immunofluorescence allowed gradual aggregation of supportive evidence.

Figure 3.19 summarises key marker genes, their localisation and more importantly their co-localisation on tSNE aiding cluster cluster assignment. Several known markers of expected tissues can be used to identify their location on the tSNE. Marker genes can be assessed individually or using different colour channels in combination.

Markers of extraembryonic endoderm include *Afp* and *Ttr* both being specific within the cell populations assayed in this current experiment. *Amn* is not specific but in combination with the others extraembryonic endoderm is clearly defined on the tSNE and though there appears to be a single point, several cells are condensed into the same point due to their relative similarity. The extraembryonic ectoderm cluster is similarly condensed onto what appears to be a single point on the tSNE. Known markers *Essrb*, *Elf5* and *Cdx2* are co-expressed only within this region as can be clearly seen on the combined three colour images. Such images are frequently used in immunofluorescence to appreciate co-expression of proteins in single cells and are readily interpretable.

The endothelial, pharyngeal mesoderm and allantoic clusters can be identified in a similar fashion. The utility of co-expression, with marker expression represented by the three primary colours, can be appreciated for these clusters where none of the markers individually are highly-specific but in combination are expressed more intensely in their expected regions.

3.7 Identification of cell populations

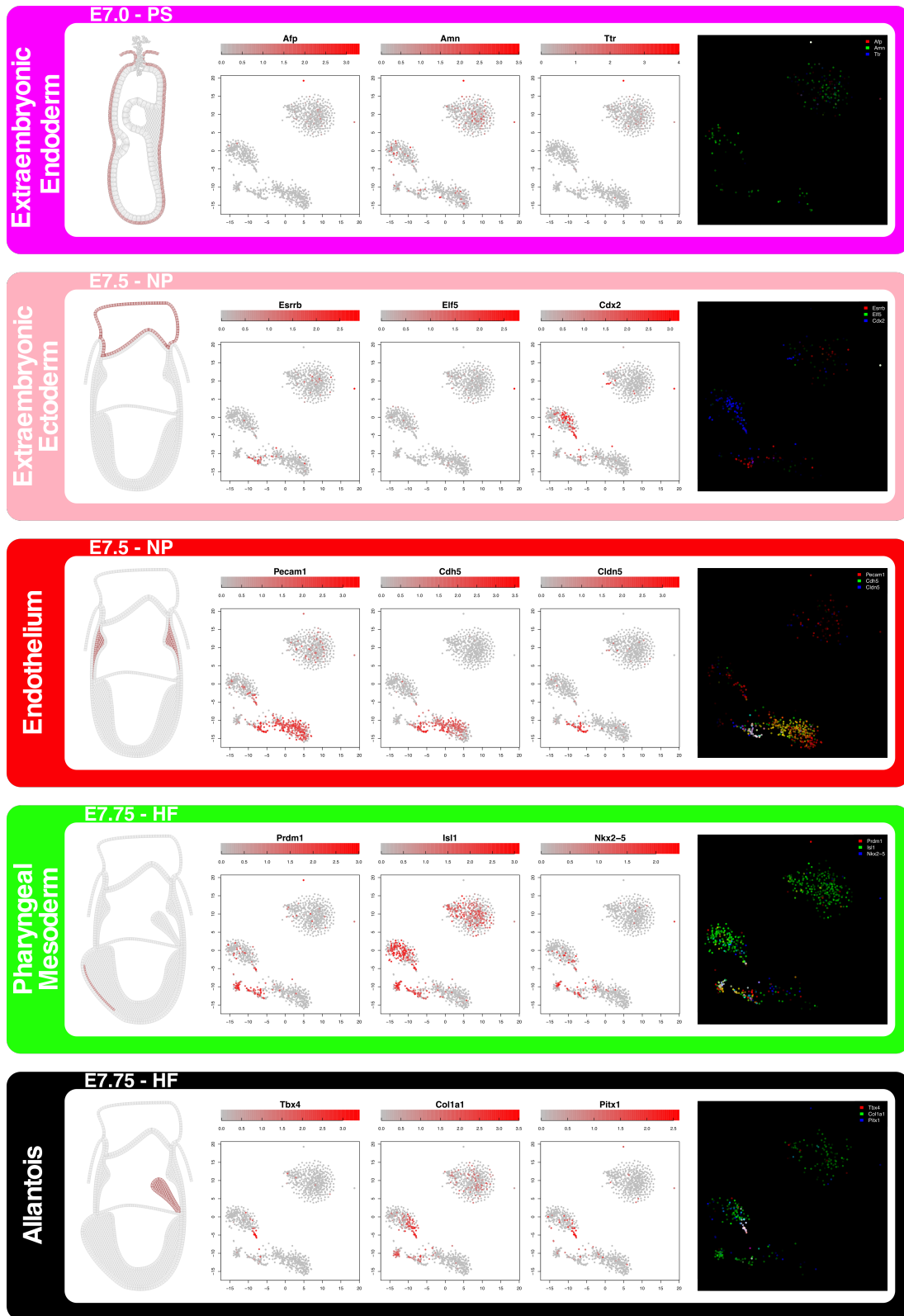


Fig. 3.19 Gene expression patterns in clusters for individual genes and combinations using the 3 primary colours in immunofluorescence-like tSNE maps. Schematics on left show anatomically expected locations.

3.8 E6.5 cluster is polarised

Gastrulation as discussed earlier is a critical component of embryonic development and at the very least it distributes cells to the appropriate region of the embryo but may also play a role in specification. *Brachyury* is a key player in gastrulation, highly expressed within the primitive streak at the streak stage. Figure 3.20 from Rivera-Pérez and Magnuson [2005] shows expression of *Brachyury* circumferentially in the extra-embryonic ectoderm juxtaposed to the epiblast and highly asymmetrically in the epiblast with the posterior epiblast having high expression. Plotting its expression on tSNE shows that the E6.5 cluster is polarised with the tapered end populated by the transforming cells undergoing ingression through the primitive streak, fig. 3.21a. To focus in on this data highly variable genes fig. 3.21b and tSNE embedding were recalculated on this restricted E6.5 epiblast population, fig. 3.21c.



Fig. 3.20 From Rivera-Pérez and Magnuson [2005] shows expression of *Brachyury* circumferentially in the extra-embryonic ectoderm and posteriorly in the epiblast of an E6.5 embryo

Having transcriptome wide gene expression at single-cell resolution it is possible to search for other genes showing similar or opposing patterns of expression to *Brachyury*, by performing a correlation analysis, fig. 3.22. This revealed genes known to be associated with gastrulation *Evx1* [Bell et al., 2016], *Fgf8* [Tam and Behringer, 1997], *Frzb*, *Wnt3* [Kemp et al., 2005; Lickert et al., 2005; Rivera-Pérez and Magnuson, 2005] and *Mixl1* but we also identify novel markers e.g. *5730457N03Rik* (known to be an antisense non-coding RNA to *Evx1*) and *Cxx1c*. GO gene set enrichment analysis of the top 100 highly *Brachyury* correlated genes using EnrichR (<https://www.github.com/wjawaid/enrichR> an R interface to a comprehensive gene set enrichment web server [Kuleshov et al., 2016]) showed significant enrichment of genes related to gastrulation ($p = 1.3 \times 10^{-3}$), angiogenesis ($p = 5.7 \times 10^{-3}$) and mesoderm formation ($p = 1.3 \times 10^{-3}$) after adjustment for multiple testing. Single cell RNAseq therefore can provide a powerful technique to identify novel genes related to a particular biological process. Further analysis may allow for generation of testable hypotheses regarding the involvement of these novel genes in gastrulation.

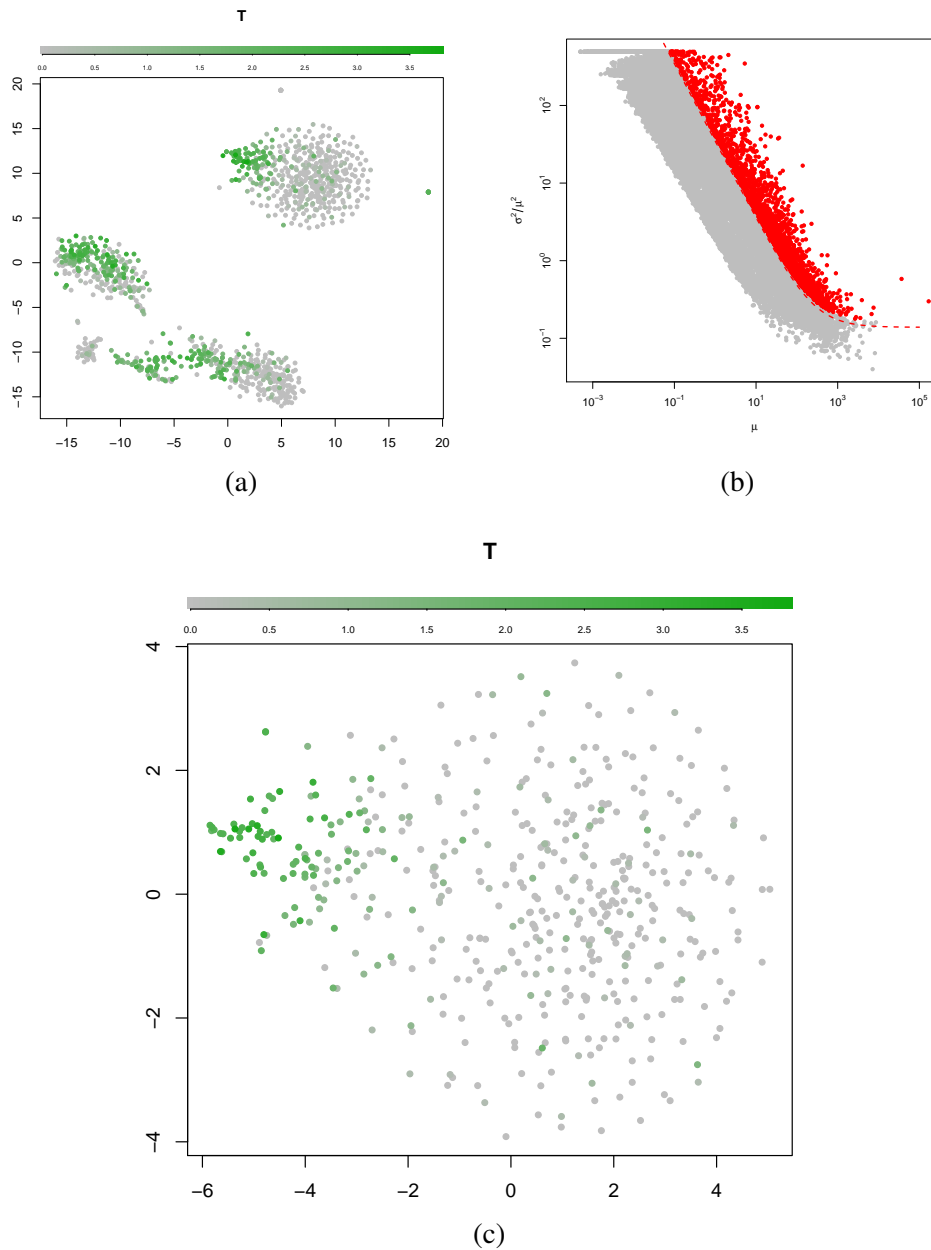


Fig. 3.21 Capturing the transcriptome-wide gene expression programme of mouse gastrulation. (a) Superimposes gene expression level of *Brachyury* (*T*) onto the tSNE representation of the dataset. (b) Highly variable genes are recalculated on the subset of E6.5 cells only. (c) tSNE recalculated on the E6.5 population with *T* expression.

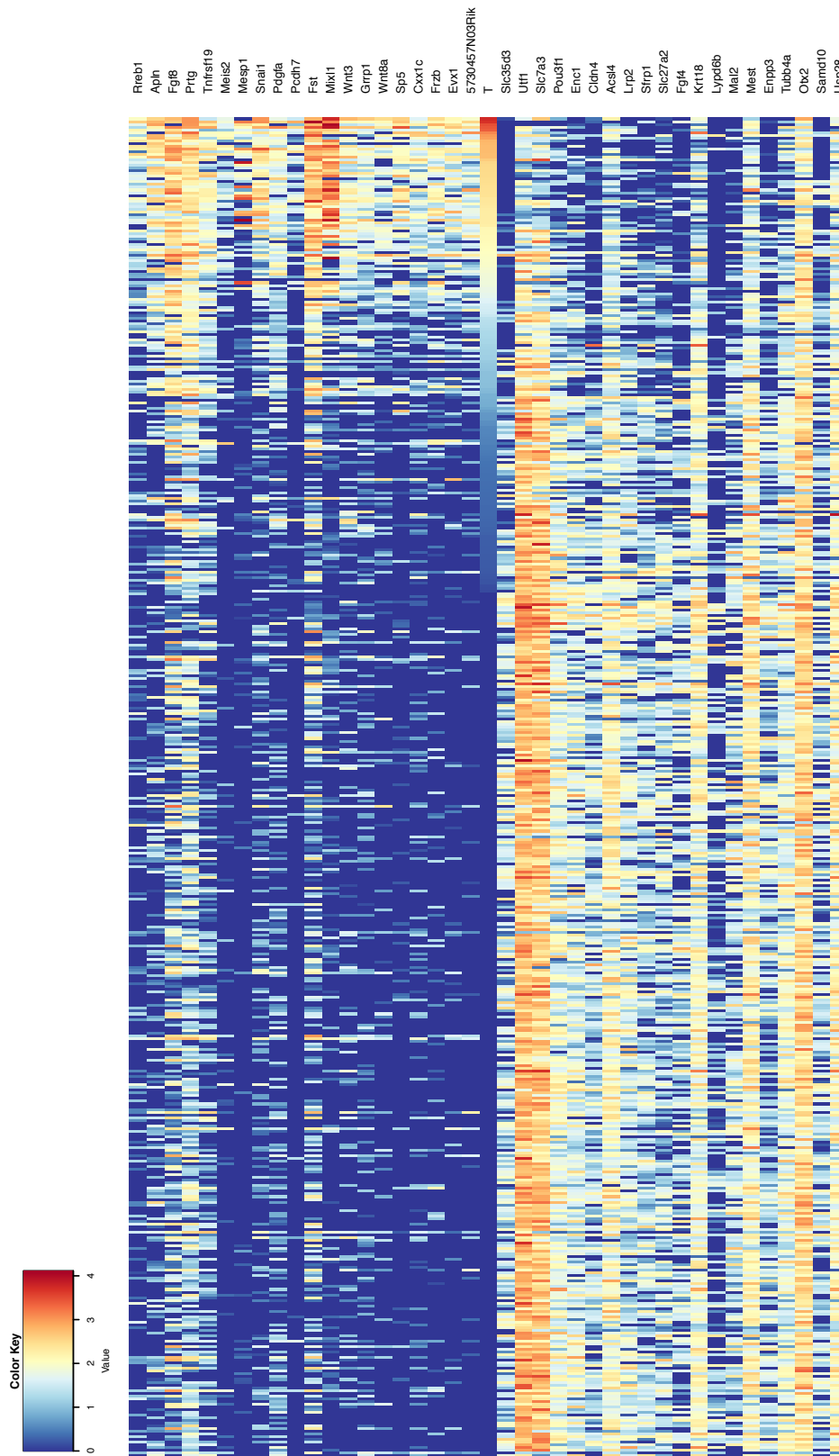


Fig. 3.22 Genes most highly correlated (upper half) or anticorrelated (lower half) with *Brachyury* (*T*) within the E6.5 dataset.

3.9 Spatial axis dichotomy - ‘Pseudospace’

Careful examination of the tSNE representation of the data and exploration of gene expression superimposed upon it, suggest cluster 4 representing *nascent mesoderm* may be divided in two. Superimposed partially transparent cell representation as shown in the left subplot of fig. 3.23 suggests that the nascent mesoderm has two centres of densities. Similarly the right subplot of fig. 3.23 displaying contours of cell density superimposed on the tSNE, is highly suggestive of two separate zones within the nascent mesodermal cluster.

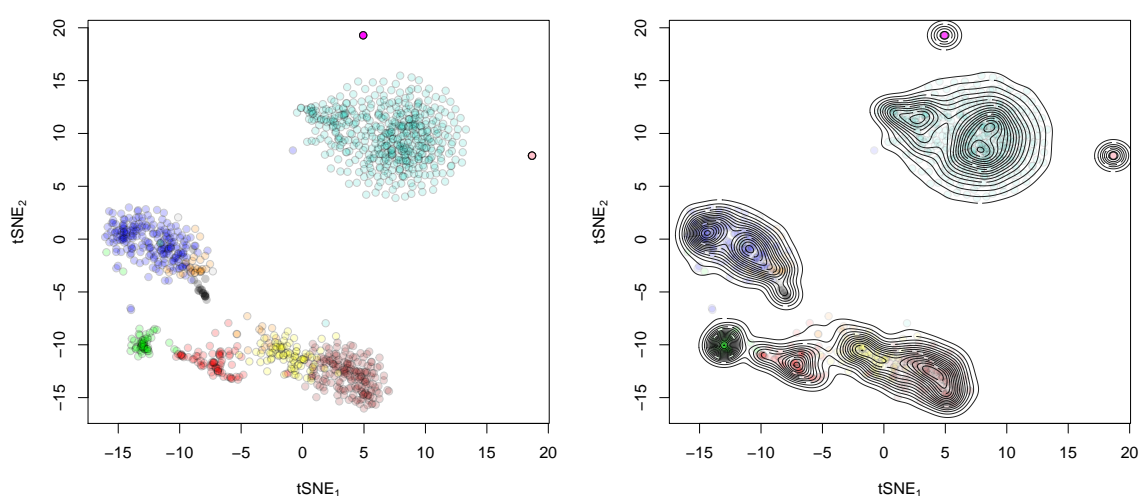


Fig. 3.23 Nascent mesoderm is composed of two zones. Left: A figure identical to fig. 3.18b with the same colour scheme but the cells are partially transparent allowing for some appreciation of cell density. This suggests two centres of densities or possibility of two separate zones within the nascent mesoderm cluster. Right: Similar to left with a contour plot of cell densities again highly suggestive of two separate populations.

Plotting expression of key marker genes suggests that these zones represent spatial regions within the embryonic mesoderm. More specifically a proximal zone expressing *Tbx2*, *Tbx3*, *EfnA1* and *Bmp4* [Du et al., 2010; McBride and Ruiz, 1998; Singh et al., 2005; Weidgang et al., 2013]. On whole mount in-situ these are expressed in the proximal region and mostly posterior but circumferentially extended around the cup into rostral mesodermal areas, see lower row of panels on fig. 3.25. To visualise these expression patterns tSNE gene expression plots focused on the nascent mesoderm cluster, fig. 3.24.

Single-cell transcriptomic analysis of murine gastrulation

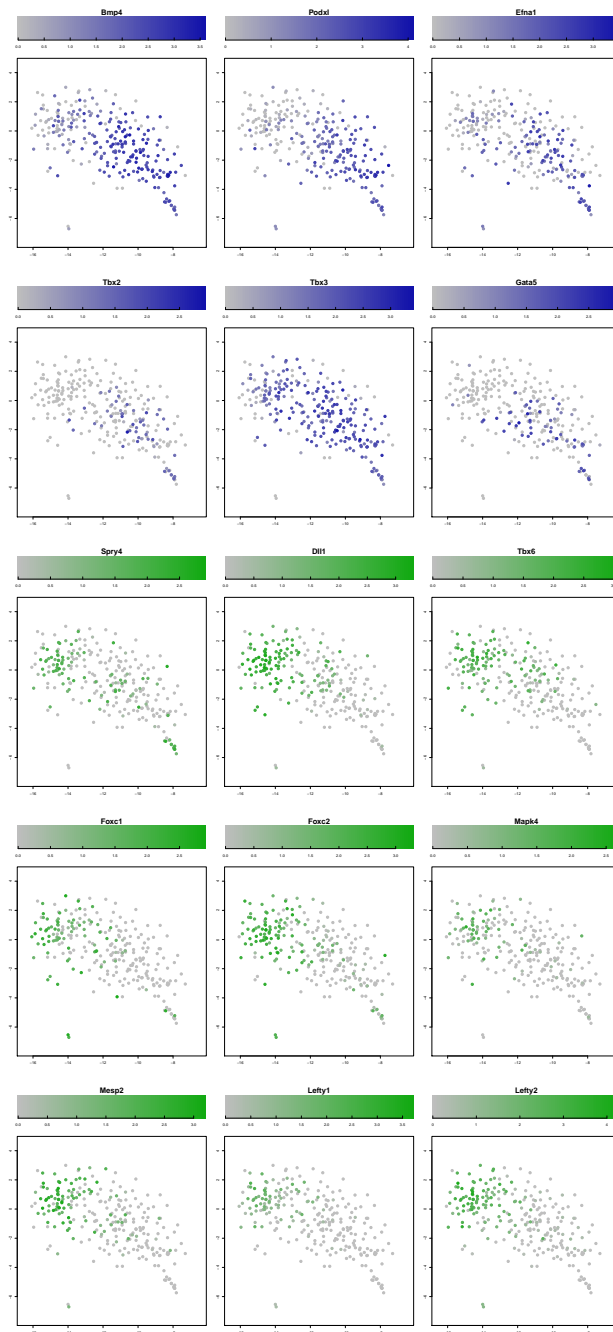


Fig. 3.24 Expression of selected transcripts on a cropped tSNE (fig. 3.23) differentially expressed between putative proximal, upper two rows of panels in blue and distal zones, lower three panels in green of the nascent mesoderm on tSNE.

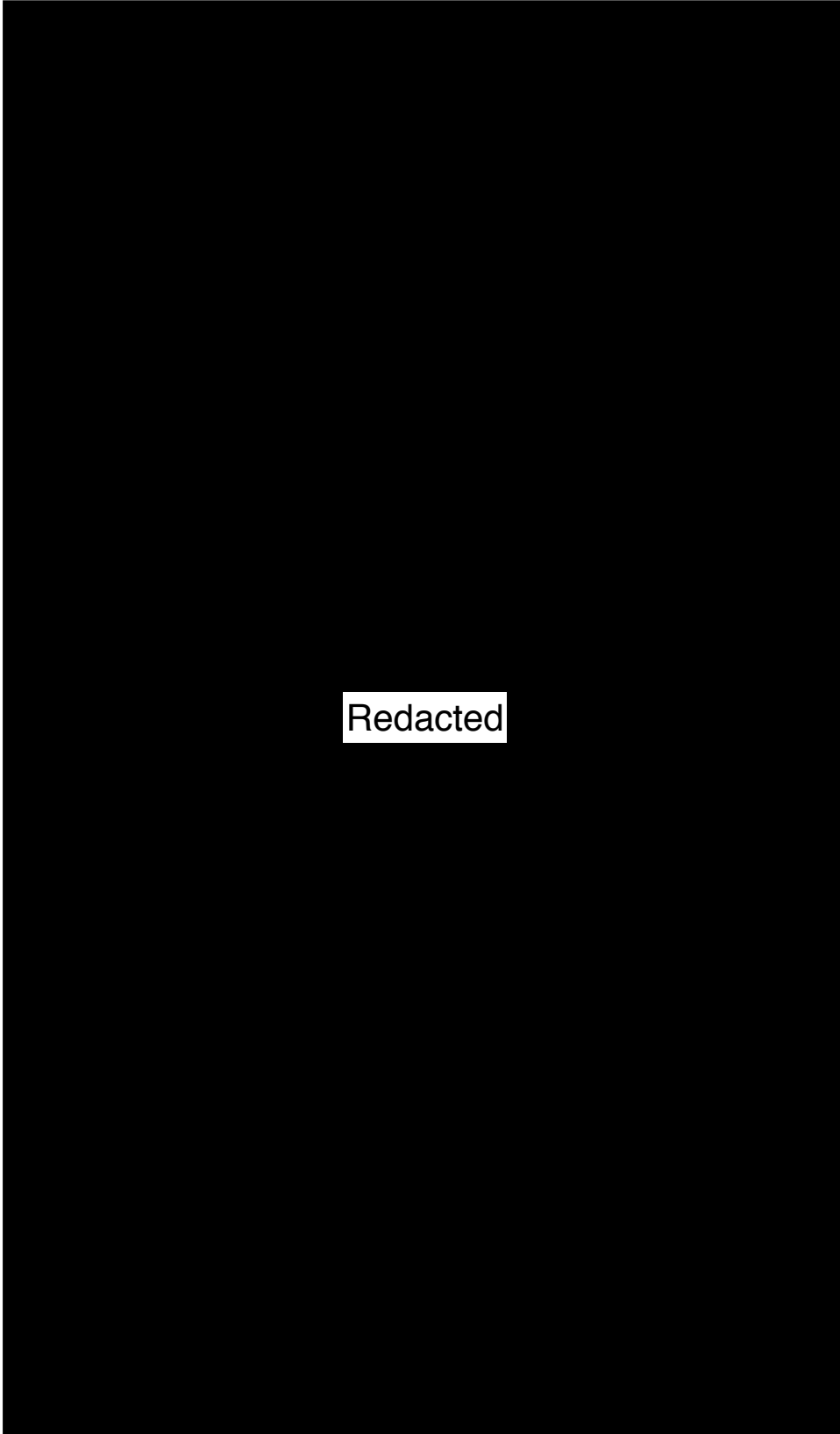


Fig. 3.25 In-situs from literature supporting spatially distinct zones within the nascent mesoderm [Du et al., 2010,?; Lickert et al., 2005; McBride and Ruiz, 1998; Minowada et al., 1999; Singh et al., 2005; Weidgang et al., 2013].

Single-cell transcriptomic analysis of murine gastrulation

Transcripts expressed more in the distal cup include epithelial to mesenchymal transition related genes, *Snail*, *Mesp1*, *Mesp2*, *Twist1*, *Lefty1*, *Lefty2*, *Foxc1*, *Foxc2*, Wnt related genes *Wnt3*, *Wnt5a*, *Frzb*, *Fzd7*, *Pou5f1*, Map-kinase related genes *Mapk4*, *Spry4*, Notch signalling genes *Dll1*, *Dll3* and *Tbx6* [Du et al., 2010; Lickert et al., 2005; Minowada et al., 1999]. In contrast to proximal zone transcripts that are expressed circumferentially in whole mount in-situ images, these are mostly restricted to the posterior half of the distal cup, see lower three rows of fig. 3.24 and the upper panels of fig. 3.25.

Cells making up both the anterior and posterior zones are almost exclusively from streak and neural plate stages of embryonic development (fig. 3.18a). Whole mount in-situ images have thus been selected at the appropriate developmental stage to support the proposal that the two zones represent spatially distinct but possibly overlapping regions of mesoderm, fig. 3.25. Additionally when interpreting in-situs one must remain cognizant that only FLK1 positive cells were sorted and analysed during these stages.

The *Hox* genes with evolutionary conserved homeodomains are colinearly ordered on chromosomal clusters and play key roles in morphogenesis. *Hox* gene temporal and spatial expression during embryogenesis mirrors their genomic order on the chromosome in an anterior to posterior axis and a 3' to 5' arrangement. *Hox* clusters in *Drosophila* and their 4 counterpart orthologous clusters in mouse are summarised in fig. 3.26. *Hox* genes may provide additional supporting evidence to the proximal and distal zone segregation within the nascent mesoderm.

To further evaluate the putative anterior posterior axis in the dataset, gene expression of the *HoxA* and *HoxB* cluster genes was plotted on the cropped tSNE, figs. 3.27 and 3.28. There is a clear predilection of the more cephalad expressed genes in fig. 3.26 to be expressed in the anterior region in comparison to the later more caudal genes. This is consistent between both *HoxA* and *HoxB* genes.

The *Hox* gene expression pattern in the more caudally and later expressed genes does not exactly match that which would be expected from fig. 3.26 partly because the cells captured in the dataset are restricted both temporally and by their surface marker identity so that only FLK1 or CD41 expressing cells have been assayed. Despite this caveat a spacial segregation can be appreciated on the tSNE representation of the data.

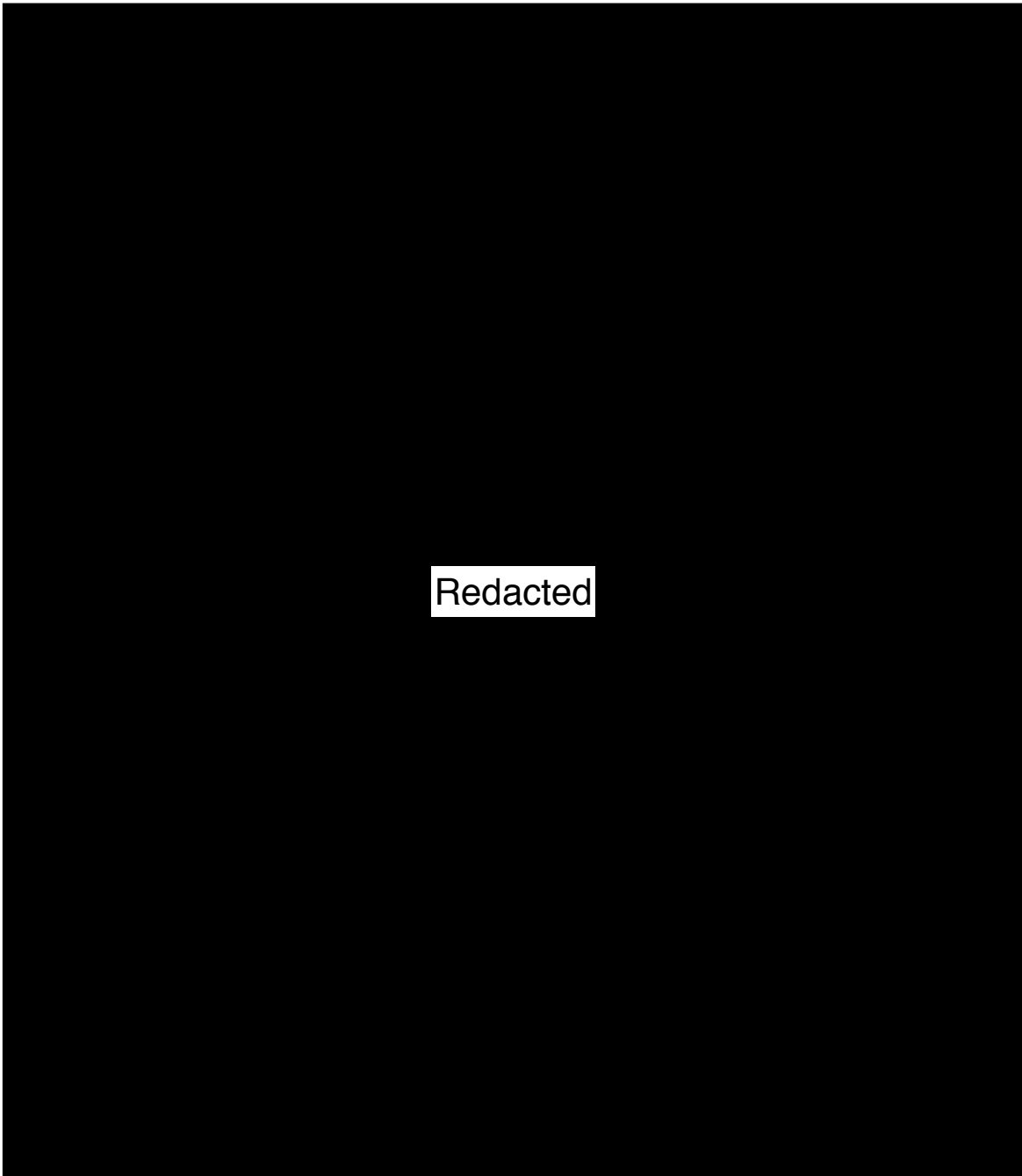


Fig. 3.26 *Hox* gene patterns are evolutionarily conserved across species and ordered 3' to 5'. Compare these expected *Hox* gene expression patterns with gene expression in the putative anterior and posterior regions in the single-cell dataset figs. 3.27 and 3.28. Image from <http://www.ehu.es/ehusfera/genetica/2012/12/16/un-modelo-matematico-aplicado-a-los-genes-hox-explica-la-formacion-de-los-dedos/> accessed January 2018.

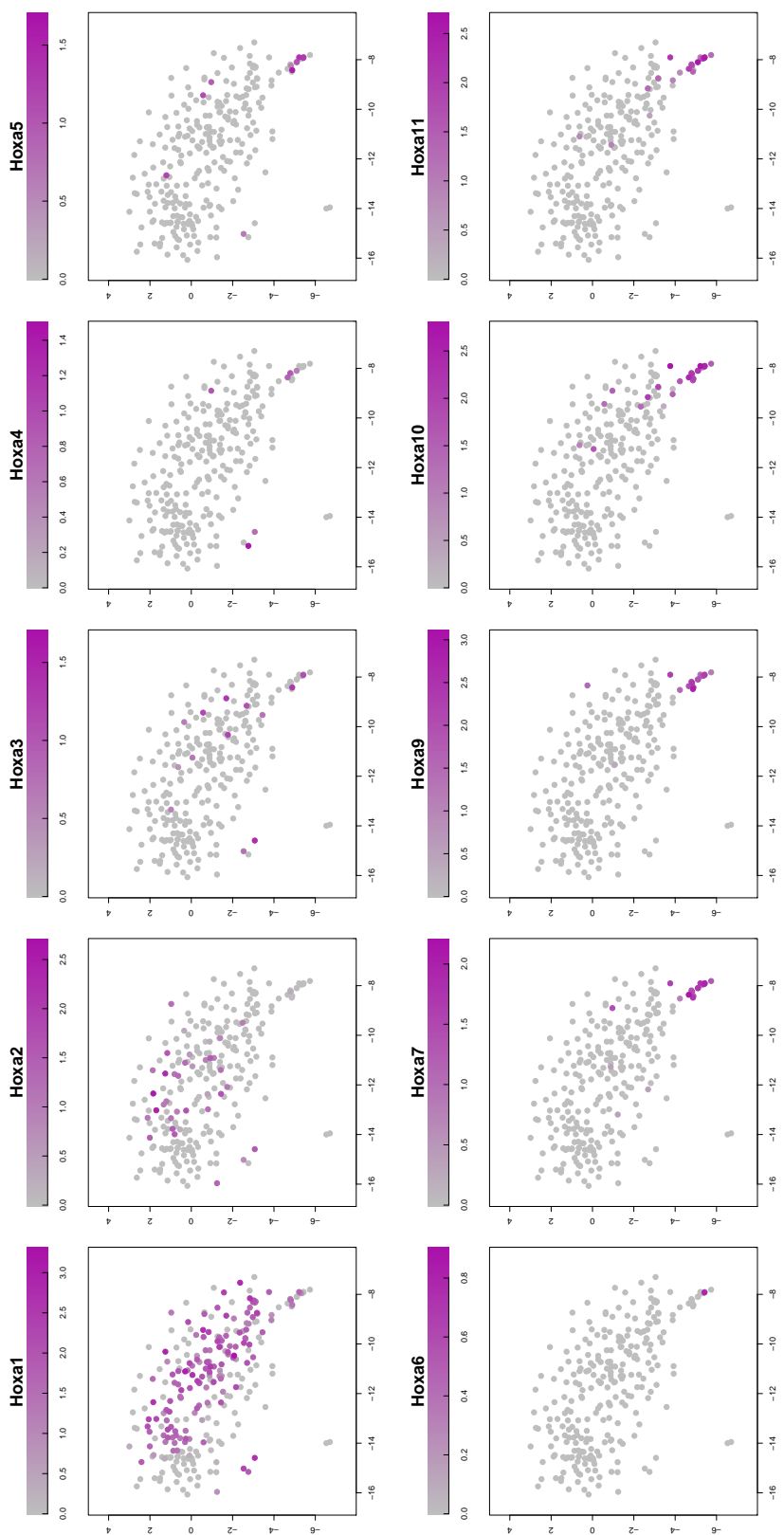


Fig. 3.27 Cropped tSNE of fig. 3.23 focusing on the nascent medoderm cluster with Hoxa gene expression plotted, providing support to the putative anterior vs posterior assignment. Compare with expected expression patterns fig. 3.26.

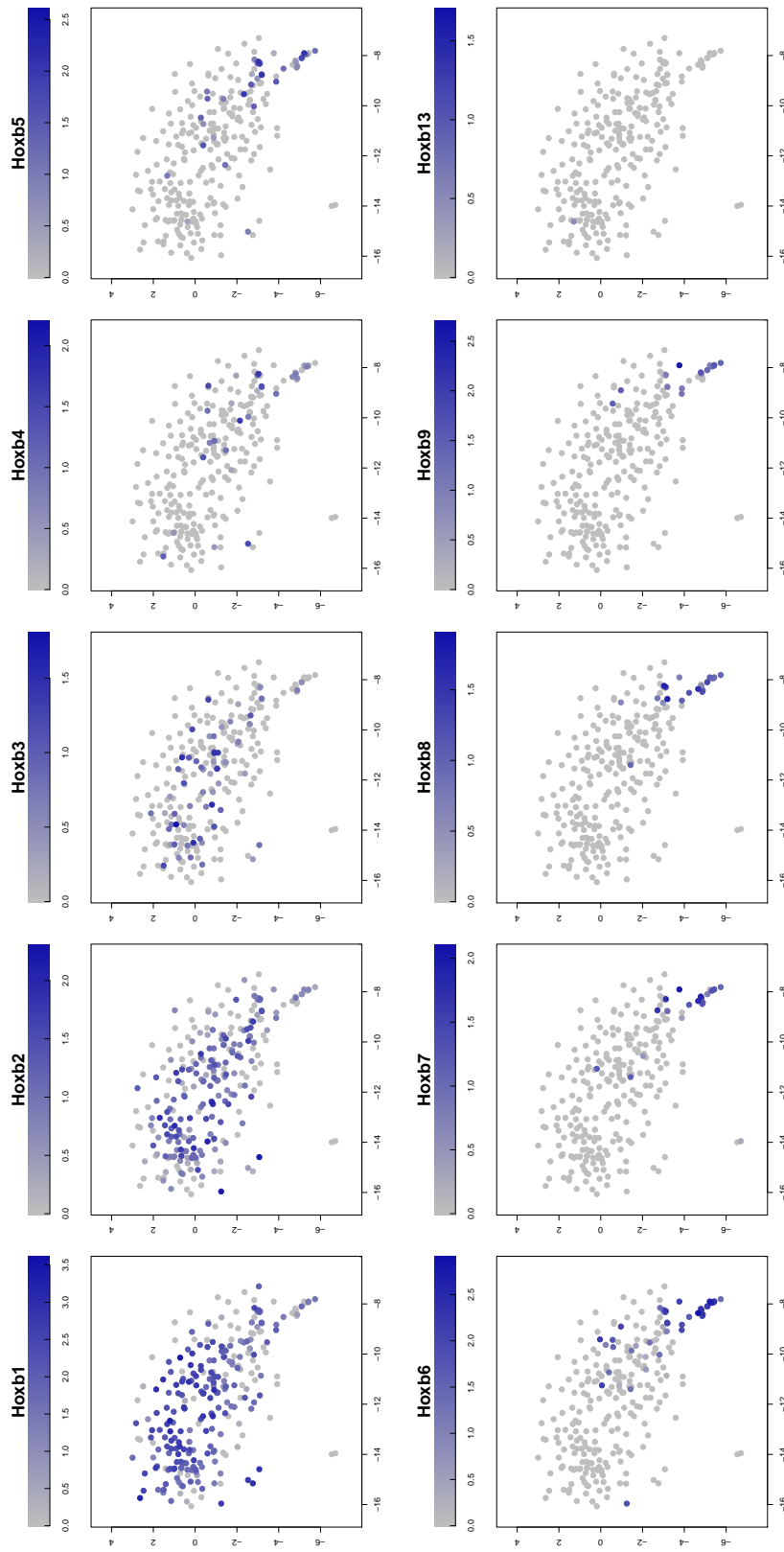


Fig. 3.28 Cropped tSNE of fig. 3.23 focusing on the nascent medoderm cluster with Hoxb gene expression plotted, providing support to the putative anterior vs posterior assignment. Compare with expected expression patterns fig. 3.26.

3.10 Charting a developmental journey - ‘Pseudotime’

Assignment of clusters 7 and 8 to *blood progenitor* and *erythroid* suggests a developmental relationship. Computationally ‘isolating’ these cells, recalculating highly variable genes allows for the cells to be ordered to provide a temporal ordering.

A tool that has been adopted for this purpose of temporal ordering in single-cell data is diffusion maps, where-by a Gaussian kernel is applied allowing for calculation of a first-order Markovian transition matrix with jump probabilities of transitioning from one cell to another in a random-walk. The spectral decomposition of this matrix then provides a means of calculating the diffusion distance. Due to spectral decay of ordered eigenvalues, the first few eigenvectors (excluding the first trivial solution) are sufficient to provide a good estimate of diffusion distance - thus forming the basis of diffusion map dimensionality reduction [Angerer et al., 2016; Coifman et al., 2005; Haghverdi et al., 2015; Nadler et al., 2006].

3.10.1 Diffusion Maps

Given an array of normalised expression values $X \in \mathbb{R}^{m \times n}$ of m genes G and n cells C , for Euclidean distance the isotropic positive semi-definite Gaussian kernel $L \in \mathbb{R}^{n \times n}$ is calculated:

$$L_{i,j} = \sqrt{\frac{2\sigma_i\sigma_j}{\sigma_i^2 + \sigma_j^2}} \exp\left(-\frac{\|x_i - x_j\|^2}{2(\sigma_i^2 + \sigma_j^2)}\right), \quad \text{for } i \neq j, \text{ where } x_i, x_j \in \mathbb{R}^m \text{ are column vectors of } X$$

In the case above a distance function d is defined:

$$d(x,y) = \|x - y\|, \quad x, y \in \mathbb{R}^m$$

For cell i an ordered set of distances to all other cell, from closest to furthest can be constructed $B^i \in \mathbb{R}^n$. σ_i can then be calculated as [Haghverdi et al., 2016]:

$$\sigma_i = \sqrt{\frac{B_{knn+1}^i}{2}}$$

where $knn \in \mathbb{Z}$ is the k -th nearest neighbour. In this case the $knn + 1$ cell is chosen as the closest will be the cell itself.

3.10 Charting a developmental journey - ‘Pseudotime’

A feature of Diffusion map is the ability to tune the influence of data point density and this can be parameterised with α :

$$L_{i,j}^{(\alpha)} = \frac{L_{i,j}}{(\sum_{k \in C} L_{i,k} \sum_{k \in C} L_{k,j})^\alpha}$$

In the remainder of the analysis we have set $\alpha = 1$ so approximating the Laplace-Beltrami operator. The density corrected kernel can be re-written:

$$L' = D^{-1}LD^{-1}$$

Where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $D_{i,i} = \sum_j L_{i,j}$. Finally the matrix must be row normalised to form the transition matrix of a Markov chain, M .

$$M = B^{-1}D^{-1}LD^{-1}$$

Where $B \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $B_{i,i} = \sum_j L'_{i,j} \neq 0$. The matrix M allows one to imagine a random walk with probability of jumping from cell i to cell j given by $p(c_j, t+1 | c_i, t) = M_{i,j}$. Jump probabilities at t steps is given by raising the transition matrix to the appropriate power $M_{i,j}^t$. The computational complexity of such naïve matrix multiplication can be prohibitive, $\mathcal{O}(n^3)$. The spectral decomposition or Jordan canonical form can provide a useful means of calculating probabilities at higher values of t . The singular value decomposition can allow such calculations:

$$M_{i,j}^t = \sum_l \lambda_l^t \phi_l(x_i) \psi_l(x_j)$$

Where λ_l is the l -th ordered eigenvalues, $\phi_l(x_i)$ the i -th component of the l -th left eigenvector and $\psi_l(x_j)$ the j -th component of the l -th right eigenvector.

The spectral decomposition of M , specifically the right eigenvectors now form the diffusion map but though the graph Laplacian L is positive semi-definite and will have positive real-valued eigenvalues and real valued eigenvectors, it is not clearly so for $B^{-1}D^{-1}LD^{-1}$:

$$B^{-1}D^{-1}LD^{-1}\Psi = \Psi\Lambda, \quad \Psi, \Lambda \in \mathbb{C}^{n \times n}$$

Single-cell transcriptomic analysis of murine gastrulation

Where $\Psi \in \mathbb{C}^{n \times n}$ is a matrix of ordered right eigenvectors and $\Lambda \in \mathbb{C}^{n \times n}$ is a diagonal matrix of correspondingly ordered eigenvalues. But since $B \in \mathbb{R}^{n \times n}$ is diagonal with all non-zero diagonal entries a numerically stable solution can be found and shown to be real valued by solving the related symmetric system:

$$\left(B^{-\frac{1}{2}}D^{-1}LD^{-1}B^{-\frac{1}{2}}\right)\Psi' = \Psi'\Lambda', \quad \Psi', \Lambda' \in \mathbb{R}^{n \times n}$$

Where $\Psi' \in \mathbb{R}^{n \times n}$ is a matrix with columns composed of ordered eigenvectors and $\Lambda' \in \mathbb{R}^{n \times n}$ is a diagonal matrix of ordered eigenvalues. Now multiplying both sides by $B^{-\frac{1}{2}}$ gives:

$$\left(B^{-1}D^{-1}LD^{-1}\right)\left(B^{-\frac{1}{2}}\Psi'\right) = \left(B^{-\frac{1}{2}}\Psi'\right)\Lambda'$$

A numerically stable and accelerated algorithm can now be implemented to calculate eigenvectors of the related symmetric system and Ψ calculated by simply applying the diagonal matrix $B^{-\frac{1}{2}}$:

$$\Psi = B^{-\frac{1}{2}}\Psi' \text{ and } \Lambda = \Lambda'$$

Furthermore since $B \in \mathbb{R}^{n \times n}$ with only diagonal entries and $\Psi' \in \mathbb{R}^{n \times n}$ we can now also say $\Psi, \Lambda \in \mathbb{R}^{n \times n}$ making them more easily interpretable. This method combined with either the Arnoldi or Lanczos algorithms [Arnoldi, 1951; Lanczos, 1950] provides a computationally efficient way of calculating the diffusion map dimensionality reduction and is implemented in the *roots* package available on CRAN and Github at <https://github.com/wjawaid/roots>.

3.10.2 Ontogenic reconstruction of embryonic blood development

The blood progenitor and blood clusters were selected on the tSNE, highly variable genes calculated and diffusion map dimensionality reduction performed. The first three diffusion map components are plotted in a pairs plot in fig. 3.29.

Diffusion component (DC) 1 segregates blood progenitors from more differentiated blood cells in a manner consistent with a pseudo-temporal developmental process. In contrast DC2 and DC3 do not appear to resolve any meaningful biological process. Additionally DC2 appears to be strongly correlated with library size as compared to DC1 and DC3, see table 3.7. Given this DC1 and DC3 were selected for future visualisation within this dataset. DC2 though does provide some evidence that certain cells may be outliers and so may need

3.10 Charting a developmental journey - 'Pseudotime'

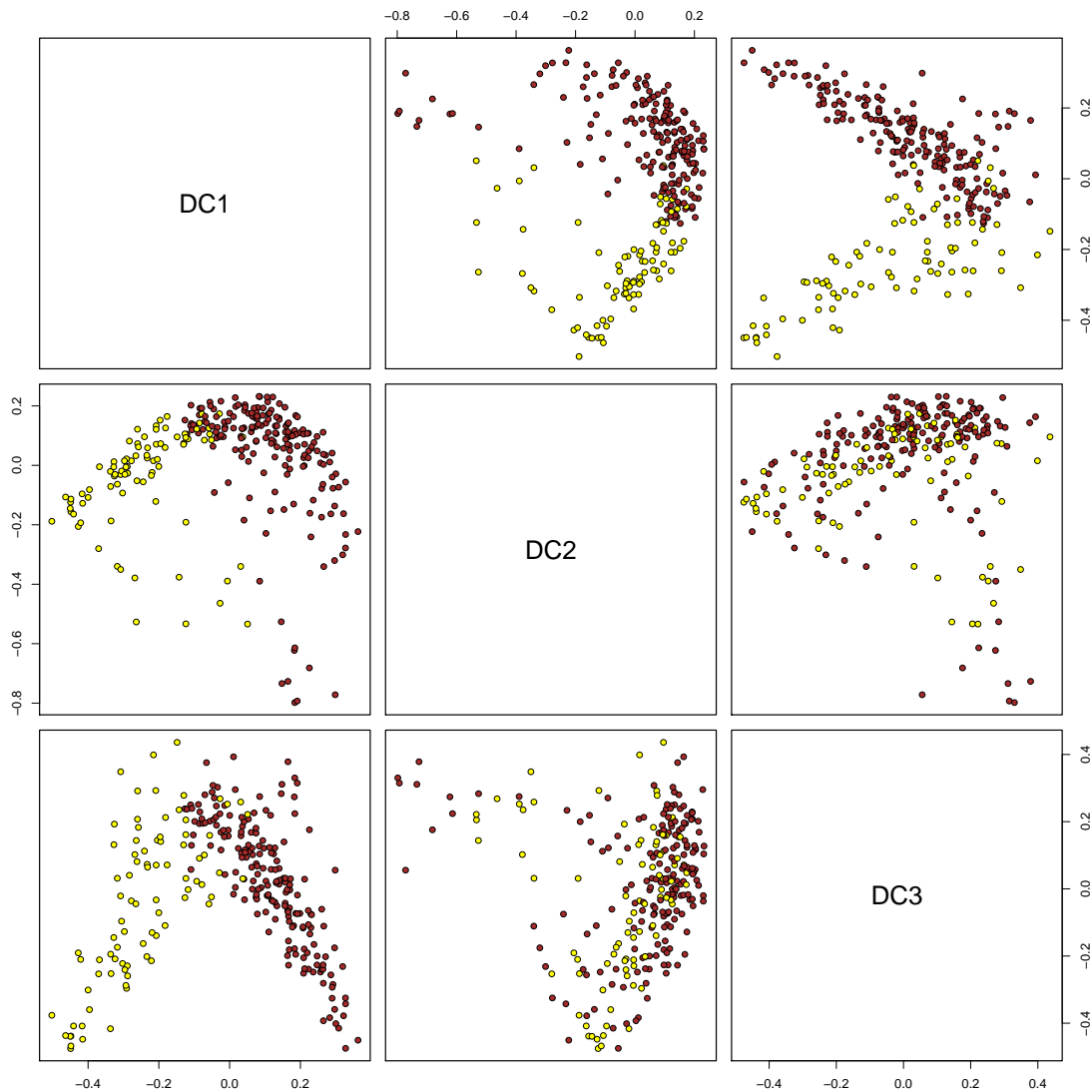


Fig. 3.29 Diffusion map dimensionality reduction. A pair plot showing the first 3 diffusion map components. Similar cells are closer in diffusion distance and are placed closer on the diffusion map so that related cells are adjacent. Yellow - Blood progenitors, Brown - Blood see fig. 3.18b. The geometric relationship is reminiscent of an ontogenic progression from blood progenitor towards blood as would be expected from the cell type assignment. This is further supported by plotting the stage of the embryo from which the cell was harvested fig. 3.30a showing a progression from blood progenitors to blood.

Single-cell transcriptomic analysis of murine gastrulation

Table 3.7 Correlation of diffusion components with library size shows strong correlation between DC2 and library size, so DC3 was favoured over DC2 for visualisation.

	DC_1	DC_2	DC_3
Pearson Correlation	-0.0625	0.2098	0.0148
p-value	0.3052	0.0005	0.8084

to be removed from downstream analysis. Plotting embryonic stage of the embryo from which the cell was harvested provides support to the temporal ontogenic progression from left to right on the diffusion map fig. 3.30a.

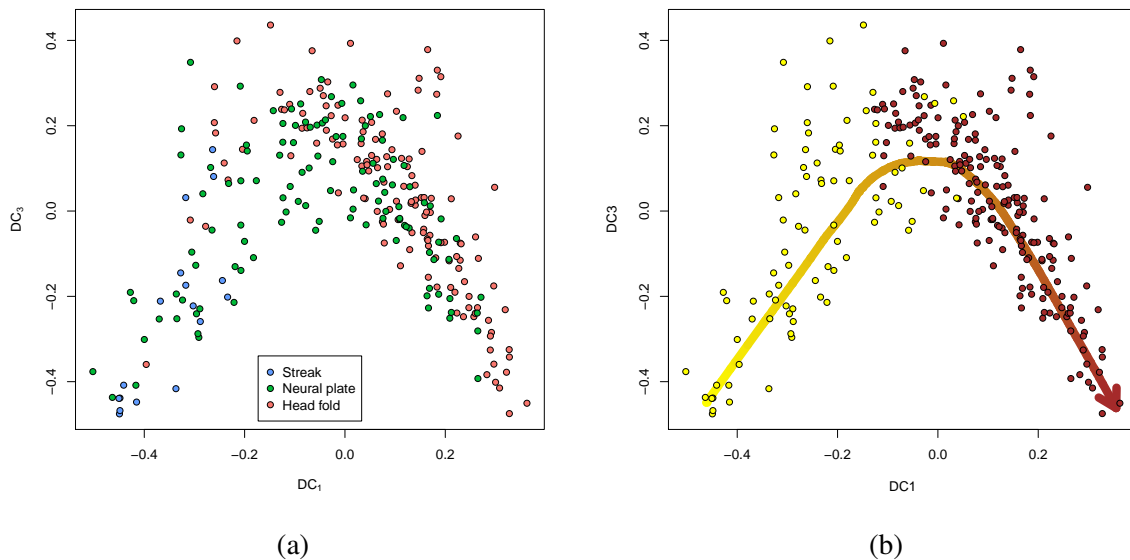


Fig. 3.30 Diffusion components (DC_1 and DC_3). In (a) each cell is coloured by the stage of the embryo from which it was harvested. In (b) cells are coloured by the assigned cluster see fig. 3.29. The path is a principal curve [Hastie and Stuetzle, 1989] and the arrow indicates the expected direction of developmental progression.

DC_1 may in this context be used as a surrogate for ontogenic progression, alternatively a principal curve [Hastie and Stuetzle, 1989] fitted and the cell projection along this used as developmental progression (fig. 3.30b) but Haghverdi et al. [2016] have proposed a formal method to calculate pseudotime based on the calculations used in diffusion maps but without necessitating any dimensionality reduction. This utilises the Markovian matrix and its spectral decomposition to evaluate the time evolution of a starting probability distribution $f(t) \in \mathbb{R}^n$:

3.10 Charting a developmental journey - ‘Pseudotime’

$$f(t) = M^t f(0)$$

Haghverdi et al. [2016] have argued that it is not important whether the asymmetric matrix, $M = B^{-1}D^{-1}LD^{-1}$ or the symmetric matrix $S = D^{-1}LD^{-1}$ are used, since diffusion maps reveal the underlying geometry of the manifold rather than directions on the manifold. Now the eigen decomposition of S can be used for calculating S^t :

$$S_{i,k}^t = \sum_{l=1}^n \lambda_l^t \psi_l(x_i) \psi_l(x_k), \quad S \in \mathbb{R}^{n \times n}$$

Where $\psi \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^n$ now represent the ordered eigenvectors and eigenvalues of S , respectively. Now rather than using a particular t and to account for asynchrony of differentiating cells, Haghverdi et al. [2016] suggest taking the infinite sum of the time evolution:

$$\sum_{t=1}^{\infty} f(t) = \sum_{t=1}^{\infty} S^t f(0)$$

Using the eigenvector decomposition this becomes:

$$\sum_{t=1}^{\infty} S_{i,k}^t = \sum_{t=1}^{\infty} \sum_{l=1}^n \lambda_l^t \psi_l(x_i) \psi_l(x_k)$$

Re-arranging summations:

$$\sum_{t=1}^{\infty} S_{i,k}^t = \sum_{l=1}^n \sum_{t=1}^{\infty} \lambda_l^t \psi_l(x_i) \psi_l(x_k)$$

This geometric sum will converge as long as all eigenvalues $\lambda_l < 1$. The first eigenvalue representing the stationary distribution with $\lambda_1 = 1$ and $\psi_1 \in \mathbb{R}^n$ with all entries of value $\frac{1}{\sqrt{n}}$. The stationary distribution only contains information of the cell sampling density and not about the temporal evolution [Haghverdi et al., 2016]. This can therefore be removed and we now call this new matrix T with entries $T(i, k)$:

$$T(i, k) = \sum_{l=2}^n \sum_{t=1}^{\infty} \lambda_l^t \psi_l(x_i) \psi_l(x_k)$$

Single-cell transcriptomic analysis of murine gastrulation

and so as $t \rightarrow \infty$ the geometric sum equates to:

$$T(i, k) = \sum_{l=2}^n \frac{\lambda_l}{1 - \lambda_l} \psi_l(x_i) \psi_l(x_k)$$

Haghverdi et al. [2016] now consider $T(i, \cdot)$, the i -th row of matrix T to be the feature representation for cell x_i . They then define the diffusion pseudotime distance measure $dpt(i, j)$ as:

$$\begin{aligned} dpt^2(i, j) &= \|T(i, \cdot) - T(j, \cdot)\|^2 \\ &= \sum_{k=1}^n (T(i, k) - T(j, k))^2 \\ &= \sum_{k=1}^n \left(\sum_{l=2}^n \psi_l(x_k) \left(\frac{\lambda_l}{1 - \lambda_l} \right) (\psi_l(x_i) - \psi_l(x_j)) \right)^2 \end{aligned}$$

and since the eigenvectors are chosen to form an orthonormal basis:

$$dpt^2(i, j) = \sum_{l=2}^n \left(\frac{\lambda_l}{1 - \lambda_l} \right)^2 (\psi_l(x_i) - \psi_l(x_j))^2$$

This now provides a method for calculating diffusion pseudotime and ordering cells in an ontogenic pseudotime. It is also evident that diffusion pseudotime is a weighted Euclidean distance in diffusion map space; each eigenvector distance weighted by a function of the corresponding eigenvalue.

The eigenvalues and so the weights decay quickly suggesting that a good approximation can be achieved by limiting the number of eigenvectors used in diffusion pseudotime calculation so as to reduce computational resource usage, fig. 3.31a. This method is implemented in the R package *roots* available on CRAN and Github, <https://github.com/wjawaid/roots>.

3.10 Charting a developmental journey - 'Pseudotime'

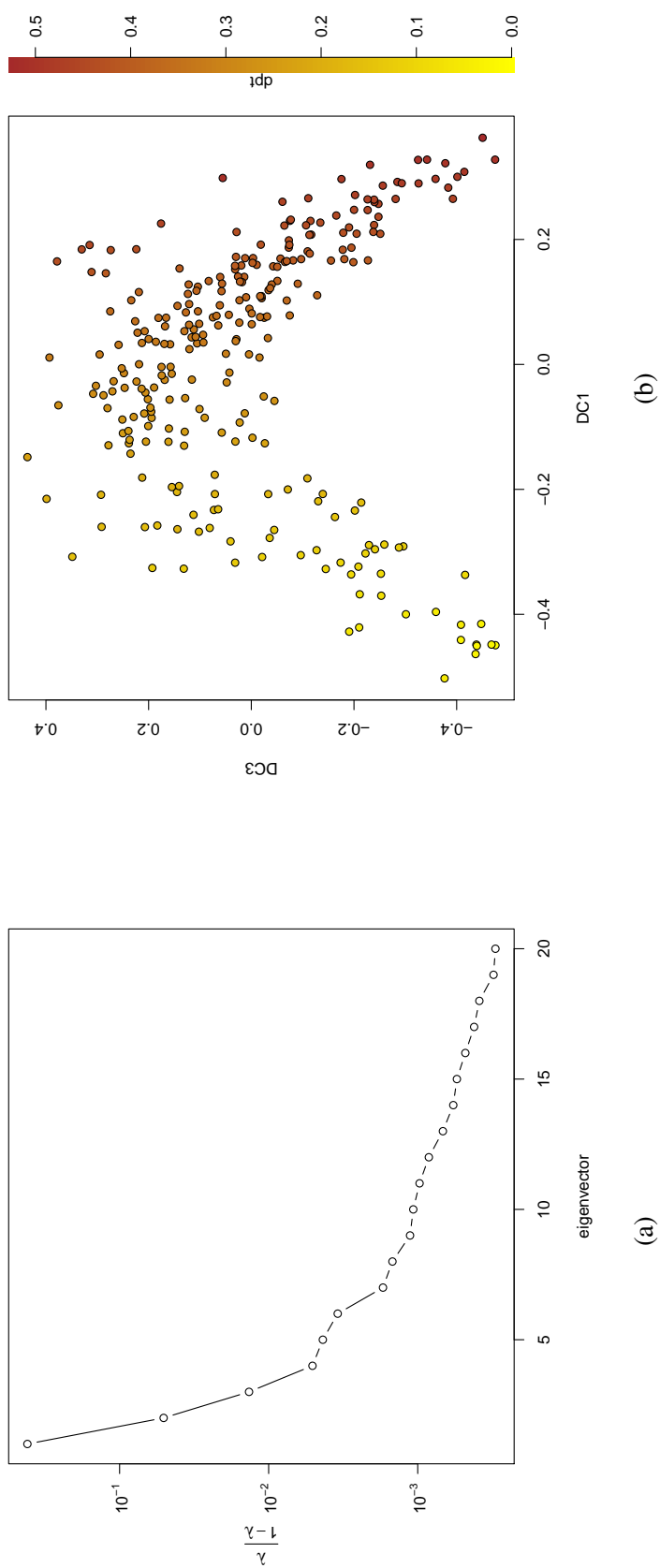


Fig. 3.31 (a) Value of the unsquared weights $\left(\frac{\lambda}{1-\lambda}\right)$ for each eigenvector showing substantial weight decay over the first 20 eigenvectors, ordinate is on a log-scale. (b) Diffusion pseudotime is plotted on the diffusion map components 1 and 3, showing progression from blood progenitors to blood.

Single-cell transcriptomic analysis of murine gastrulation

Applying this to the subset of cells forming the embryonic blood and progenitor populations reveals a developmental progression consistent with the diffusion map representation and the principal curve, figs. 3.30b and 3.31. Additionally it becomes even more apparent that some of the cells may be outliers and should be excluded when generating gene dynamics, fig. 3.32.

Diffusion pseudotime [Haghverdi et al., 2016] and the projection on to the principal curve [Hastie and Stuetzle, 1989] fitted to the diffusion map are consistent in their time ordering showing strong correlation pearson's correlation $\rho = 0.968$ and $p = 5.7 \times 10^{-146}$) fig. 3.33. Subsequent analyses have been limited to using diffusion pseudotime.

3.10.3 Inferring gene dynamics during primitive-wave haematopoiesis

Armed with transcriptome-wide single cell gene expression along this developmental trajectory it is possible to define the dynamics of gene expression as progenitor cells differentiate into committed erythroid cells. Faithful reconstruction of these gene expression dynamics may provide a means to infer gene interactions and therefore provide insights into the gene regulatory networks that govern differentiation within this subset of cells.

Gene expression profiles though are affected by noise from experimental measurement fidelity limits, which in single cell data includes drop-out. 'Drop-out' refers to the failure to capture an mRNA molecule from a cell despite moieties being present due to the random nature of the capture technique, fig. 3.4.

A supervised learning approach can be used to fit a model to the noisy data providing two main advantages:

1. Smooth out the noise in the data
2. Make predictions at unmeasured time points

A variety of methods exist including generalised linear regression where model complexity e.g. order of polynomial may be selected; the model fitted followed by subsequent model selection. Selecting models may be difficult with simple models not fitting data well and more complex models overfitting the training data. For the purposes of fitting gene expression dynamics we adopted an alternative approach of giving a prior probability to every possible function but higher probabilities assigned to functions considered to be more likely, for example smoother functions. Rasmussen and Williams [2005] describe how this can be done using a generalisation of the Gaussian probability distribution - the Gaussian process. This defines a fully stochastic model of gene expression so that the expectation (mean) and the

3.10 Charting a developmental journey - 'Pseudotime'

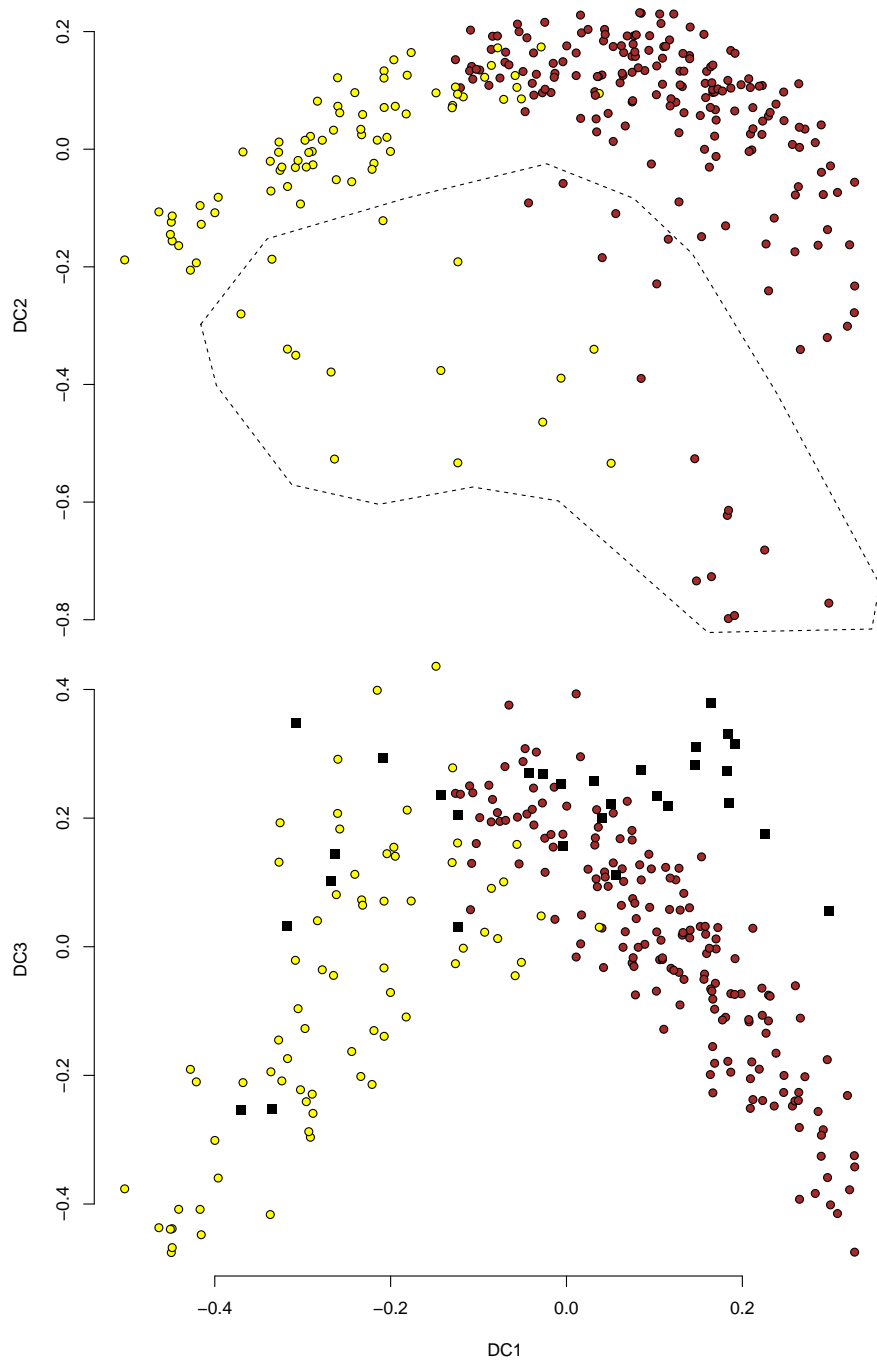


Fig. 3.32 Outlier cells were excluded based on DC2 (above). The corresponding cells are marked on the DC1/DC3 plot as black squares (below).

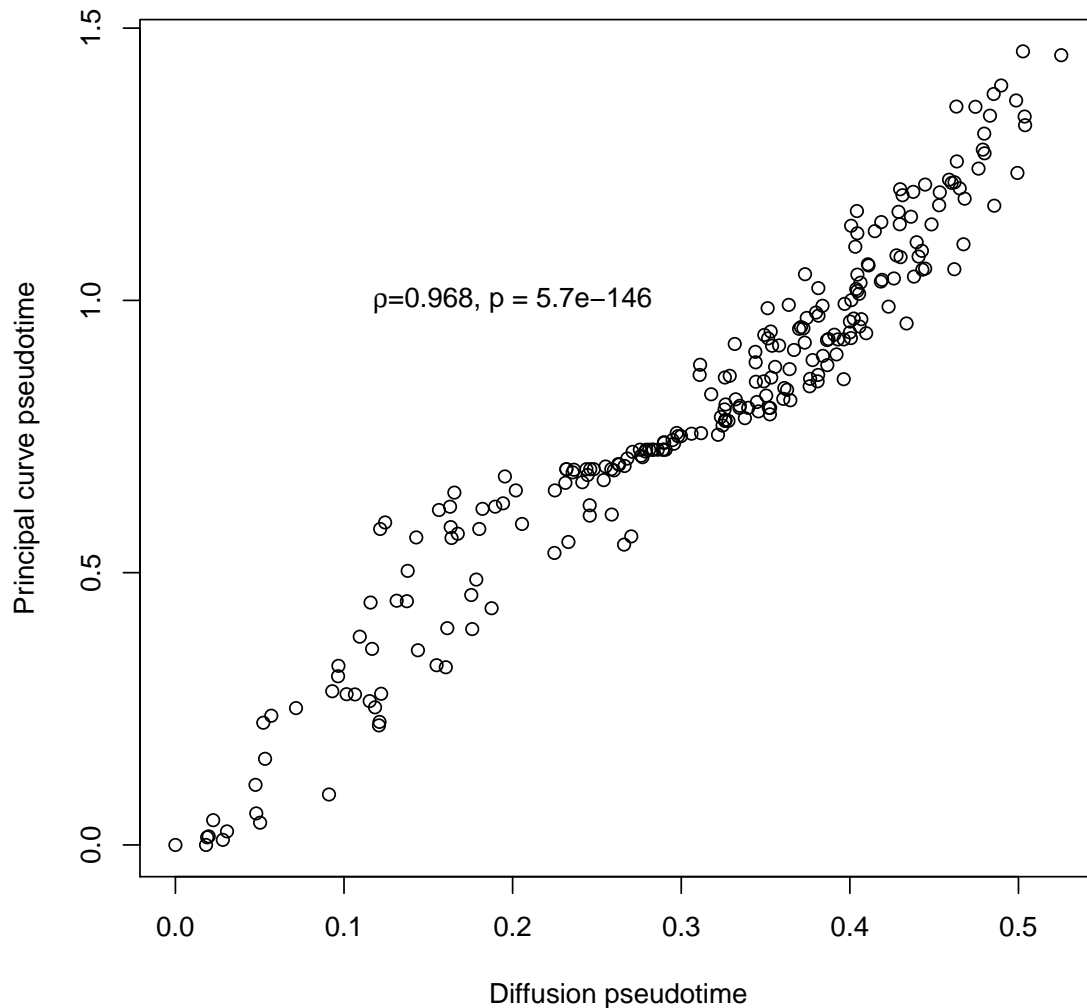


Fig. 3.33 Diffusion pseudotime and principal curve pseudotime are highly correlated $\rho = 0.968, p = 5.7 \times 10^{-146}$. This high correlation provides empirical support that diffusion pseudotime appears to be resolving a temporal progression mapped out by the principal curve on the diffusion map, see fig. 3.29.

variance may be calculated at any pseudotime, additionally a marginal likelihood can be calculated to allow formal model comparison.

Gaussian processes can be fully specified by a mean function $m(x)$ and a covariance function $k(x, x')$ in this case for a gene expression profile function $f(x)$, from Rasmussen and Williams [2005]:

3.10 Charting a developmental journey - 'Pseudotime'

$$m(x) = \mathbb{E}[f(x)],$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x')))]$$

The covariance function that will be used here is the parameterised squared exponential:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2l^2} |x - x'|^2\right) + \sigma_n^2 \delta_{xx'}$$

where the length-scale l , the signal variance σ_f^2 and the noise variance σ_n^2 assuming independent noise are hyperparameters [Rasmussen and Williams, 2005]. $\delta_{xx'}$ is the Kronecker delta:

$$\delta_{xx'} = \begin{cases} 1, & \text{if } x = x', \\ 0, & \text{if } x \neq x'. \end{cases}$$

Assuming that we have observations (\mathbf{x}, \mathbf{y}) where $\mathbf{x} \in \mathbb{R}^n$ in the current application is a vector of pseudotime of n cells and $\mathbf{y} \in \mathbb{R}^r$ is the value of expression of any particular gene corresponding to the pseudotime x . Now we wish to make predictions $(\mathbf{x}_*, \mathbf{f}_*)$ of the generating process under the multivariate normal distribution prior for the joint distribution

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_{x_*} \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I} & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

where the function K is a vectorised function, such that $A = K(\mathbf{p}, \mathbf{q})$, $\mathbf{p} \in \mathbb{R}^m$ and $\mathbf{q} \in \mathbb{R}^r$ returns $A \in \mathbb{R}^{m \times r}$ with $A_{i,j} = k(\mathbf{p}_i, \mathbf{q}_j)$. $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. $\boldsymbol{\mu}_x = m(\mathbf{x})$ and $\boldsymbol{\mu}_{x_*} = m(\mathbf{x}_*)$ and both $\boldsymbol{\mu}_x = \mathbf{0} = \boldsymbol{\mu}_{x_*}$ where $\mathbf{0} \in \mathbb{R}^n$ is the zero vector. Given such a joint distribution the marginal distribution of \mathbf{f}_* is simply given by

$$\mathbf{f}_* \sim N(\boldsymbol{\mu}_{x_*}, K(\mathbf{x}_*, \mathbf{x}_*))$$

and the conditional distribution of \mathbf{f}_* given \mathbf{y} is

$$\begin{aligned} \mathbf{f}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_* &\sim N(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)), \quad \text{where} \\ \bar{\mathbf{f}}_* &= \boldsymbol{\mu}_x + K(\mathbf{x}_*, \mathbf{x}) [K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I}]^{-1} (\mathbf{y} - \boldsymbol{\mu}_x), \quad \text{and} \\ \text{cov}(\mathbf{f}_*) &= K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{x}) [K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I}]^{-1} K(\mathbf{x}, \mathbf{x}_*) \end{aligned}$$

Inversion of the symmetric positive semi-definite covariance matrix $[K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I}]$ may be computationally challenging and is addressed by using the Cholesky decomposition, where L is a lower triangular matrix

$$[K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \mathbf{I}] = LL^T$$

Gaussian process regression using a naïve uninformative prior mean function $m(x) = 0$ and the squared exponential function with $\sigma_f = 1, l = 0.2, \sigma_n = 1$ hyperparameters, was applied to gene expression profiles along diffusion pseudotime to fit smoothed curves. This was implemented in the R package *gpr* that can be downloaded at <https://github.com/wjawaid/gpr>. This completely naïve uninformative prior led to a poor fit at the edges of the dynamic process due to the sparsity of the data. Choosing a better prior should circumvent this problem. There are several choices but a simplistic method is that given we are modelling gene g with mean expressions \bar{x}_p^g in the progenitor blood population and \bar{x}_b^g in the more differentiated blood population on the pseudotime interval $[0, T]$, we assume a linear mean prior:

$$m(t) = \frac{\bar{x}_b^g - \bar{x}_p^g}{T} t + \bar{x}_p^g$$

Expression profiles of a few key haematopoietic genes are shown in fig. 3.34. This demonstrates some of the expected dynamics with *Hbb-bhl* steadily increasing and *Kdr* (encoding the protein FLK1) steadily decreasing. The variance around the mean in Gaussian process regression shown in the shaded grey region depends on the sampling density around that pseudotime and on the amount of variance within the dataset itself.

Gene dynamics across the transcriptome along this differentiation process can now be compared by correlating the smoothed expression profiles along the pseudotime and clustering into the main gene dynamic profiles. Furthermore since noise is explicitly modelled the variance for the mean of each gene along pseudotime can be compared to evaluate genes that have similar noise patterns. This noise pattern comparison would be more useful in a

3.10 Charting a developmental journey - 'Pseudotime'

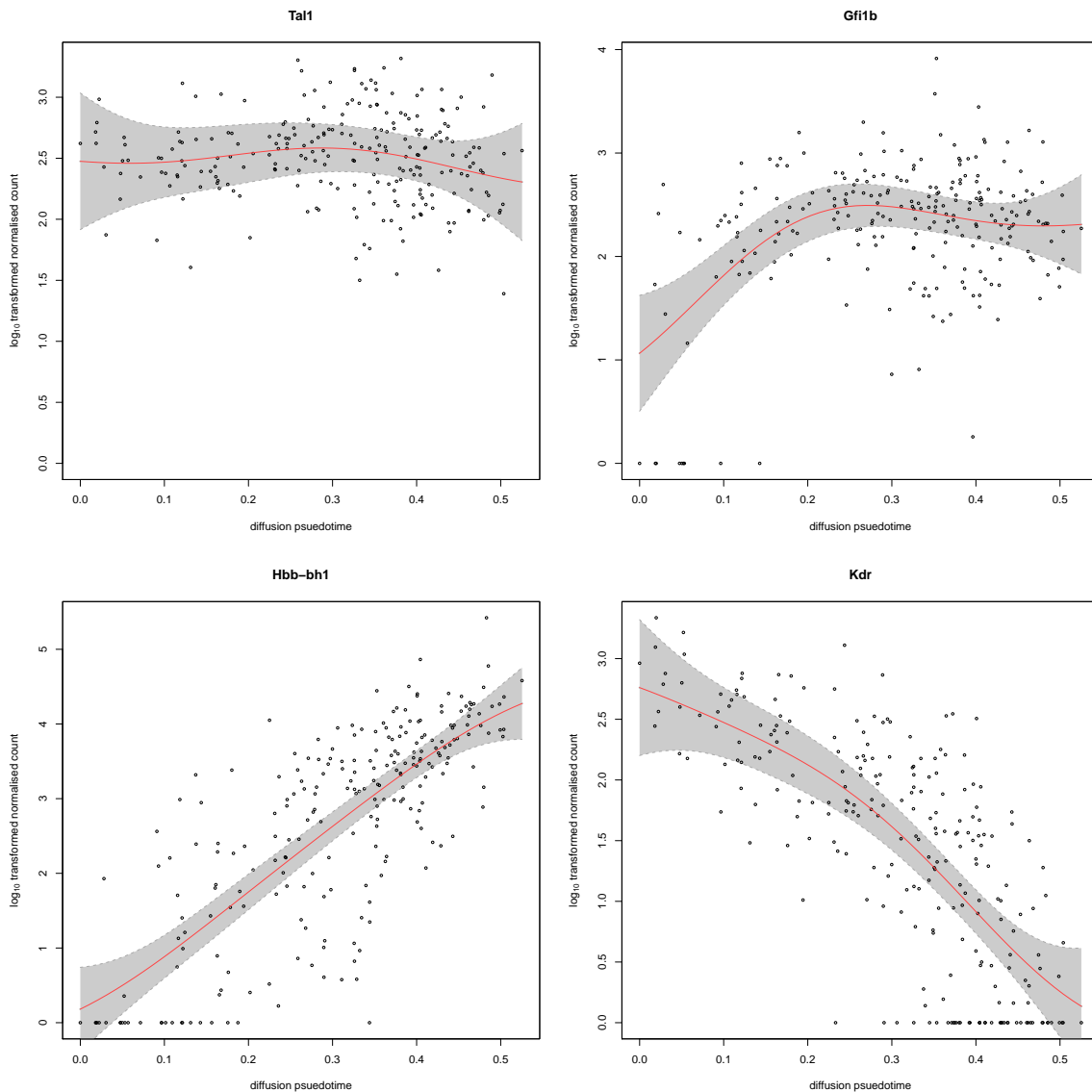


Fig. 3.34 Gene expression patterns showing *Tal1* a gene that is activated even at the earliest time, *Gfi1b* that becomes activated early then plateaus, *Hbb-bh1* that continues to increase and *Kdr* which starts with high expression and is no longer expressed toward the end of pseudotime. The red line represents the mean and the shaded region either side marks the limit of the 95% confidence interval. Each point represents a cell and marks the actual data i.e. the \log_{10} transformed normalised counts.

bifurcating process as a noisy system may resolve with expression profiles of cells converging around two different expression levels.

Gaussian process regression was performed on the 9286 genes annotated by ENSEMBL as protein-coding, expressed in at least 30% of cells within the haemogenic progenitor and

Single-cell transcriptomic analysis of murine gastrulation

embryonic blood clusters. These gene expression profiles typically form a time-series and as such methods designed for analysing time-series based on dynamic time warping (dtw) [Sakoe and Chiba, 1971, 1978] were used to characterise dissimilarity, to perform clustering and to extract per cluster prototypic gene expression profiles. Before progressing genes were further filtered to keep only those genes that have sufficient variation, so that only genes with posterior fitted means having a standard deviation across the whole dynamic process of greater than 0.2 were clustered. This yielded 1546 gene profiles that were analysed using the dtwclust package in R using fuzzy clustering (<https://cran.r-project.org/package=dtwclust>), see methods for further details. Six distinct clusters were identified and a prototypic gene profile from each cluster is displayed in fig. 3.35. Table 3.8 shows some genes that are most strongly linked with each of the fuzzy clusters.

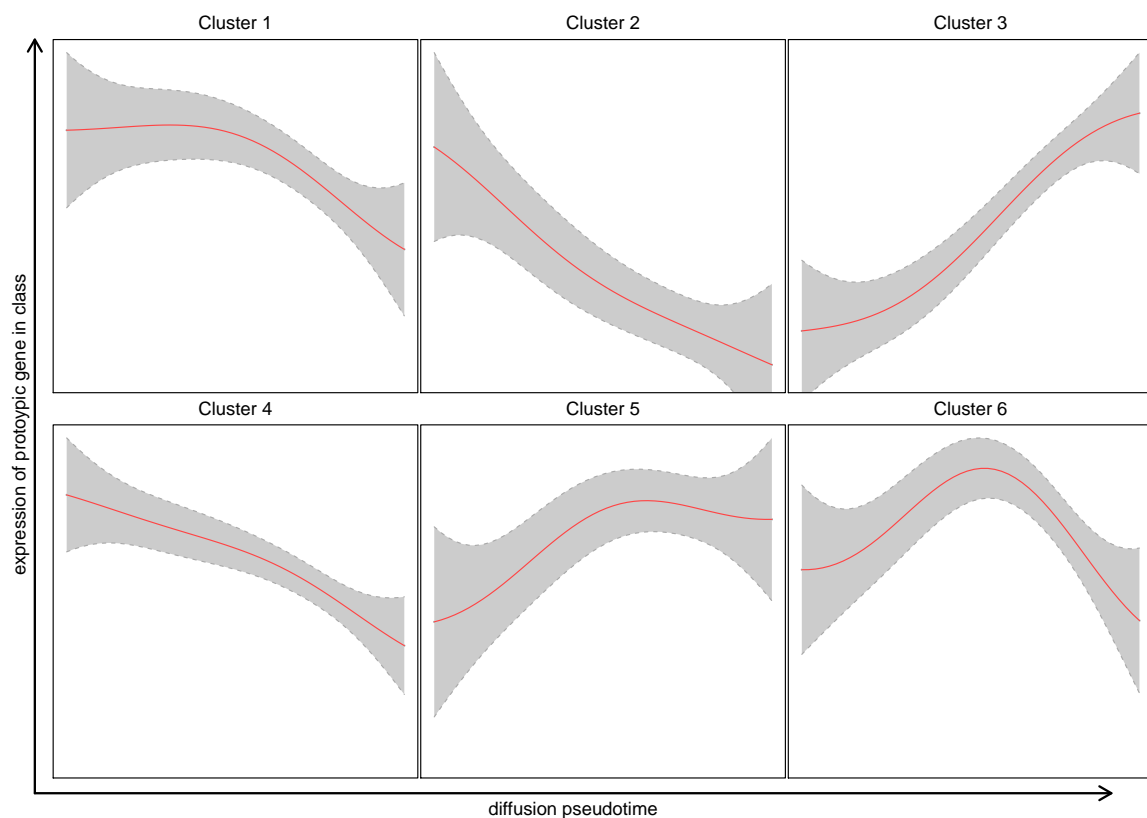


Fig. 3.35 Gene profiles of a prototypic gene from each of the fuzzy clusters. Some of the genes assigned to these clusters are shown in table 3.8

3.10 Charting a developmental journey - 'Pseudotime'

Table 3.8 Genes dynamic profiles clustered using dynamic time warping distances and fuzzy clustering. Genes cluster assignments for some key genes are shown. See fig. 3.35 for expected dynamics from each cluster.

Cluster	Genes
1	<i>Smad4, Cmtm7, Fn1, Bcor11, Rab31, Nek6, Fign, Golim4, Sepn1, Sertad1, Klhl2, Trim44, Gpr108, Sptbn1, Emilin1, Arhgef2, Rhbdd2, Vgll4, Tmbim1, Myo5a, Appl2, Zyx, Plcl2, Arfgef1, Cx3cl1, Dyrk2, Bend4, Amot, Fbxo18, Fam21, Pex6, Hspg2, Clic6, Sh2b3, Nfxl1, Hhex, Cmtm3, Pde4a, Wwc2, Map4k4, Hapln1, Sall2, Brd1, Smurf2, Igf2, Depdc1b, Smpd2, Elmo2, Thap3, Stx6</i>
2	<i>Pced1b, Fam84b, Tsku, Lef1, Tmem98, Phlda2, Gnai1, Dscr3, Sh3rf1, Prickle1, Phf6, Csrp2, Tmod2, D17H6S53E, Map3k1, Gpc3, Rnf103, Rgs19, Zfp260, Ap1s2, Tbx3, Wwtr1, Ptpkr, Plekha1, Plk2, Klf5, Dusp16, Rbp1, Nid2, Ank2, Saysd1, Basp1, Frmd6, Dst, Dcbl2, Adam23, Krt8, Zscan21, Ehd2, Grrp1, Serpinb6a, Tinf2, Ptpn13, Nr4a1, Galnt11, Map1b, Pcsk7, Rasl11b, Slc30a4</i>
3	<i>Chchd10, Grina, Pradc1, Sept4, Slc38a5, Epb4.2, Tuba4a, Sord, Kcnn4, Smox, Fads3, Smim1, Cpne7, Rgs10, Xk, Psmb10, Ube2l6, Uros, Agfg2, Cpeb4, Hpdl, Ppox, Gstm1, Gabra4, Nmrk1, Car2, Syng1, Slc25a37, Gm6665, Tmod1, Akap7, Blvrb, Angptl4, Slc14a1, Cercam, Hbb-bh1, Ndufaf7, Dhhrs11, Hebp1, Mt1, Slc16a10, Daam1, Alas2, Trim10, Rab3il1, Ell2, Hba-x, Mtg1, Spefl</i>
4	<i>Rbms1, Cep55, Dag1, Cnn3, Vcl, Ptpn9, Zfp521, Gng12, Gng2, H2afy2, Kif2c, Pcgf5, Rbm4b, Mpnd, Fnbp11, Zfp3611, Efcab14, Mapk3, Smtn, Kifc1, Wtip, Shc1, Fli1, Rbpms, Ndfip1, Elavl2, Ckap4, Sptlc2, Kif1b, Dock1, Pvrl2, Tmem185b, Podxl, Nde1, Zfp652, Kif13a, Bcl2l1, Rras2, Limd2, Fkbp7, Serpinh1, Dock11, Cfl2, Rasal2, Cenpf, Sox11, Itpkb, Rpp25, Kdr</i>

Continued on next page

Single-cell transcriptomic analysis of murine gastrulation

Table 3.8 – continued from previous page

Cluster	Genes
5	<i>Ddx26b, Prkca, Bid, Mpp2, Gucyl1a3, Slc25a12, Cep70, Atmin, Pfkml, Bsn, Ehd3, Micall2, Gamt, Lgals9, Epor, Itga2b, Ajap1, Ppargc1b, Spns2, Gja6, Nfe2, St3gal5, Was, Cited4, Gne, Mfsd2b, E2f8, Taok3, Tspan32, Ccnblip1, Rps2, Gfi1b, Hk1, Gsta4, Trem12, Homer2, Gata1, Matn1, Mbnl1, Fam110c, Cacna2d2, Hivep1, Sla2, Hes6, Ano2, Abcg1, Fam58b, Syk, Igsf3, Mxi1</i>
6	<i>Rnf125, Tie1, Plxnc1, Capn2, Afap111, Col18a1, Egr1, Oaf, Itga6, Lrp11, Src, Bcl11b, Ier5, Ptrf, Bcr, Dpysl3, Sdc3, Ogfod1, Dusp2, Unc45a, Dpy1911, Slc18a2, Plec, Fam160b2, Sipal, Ascl2, Pear1, Timp3, AI467606, Fam193b, Zswim6, Actr5, Slc9a3r2, Gna12, Atad2b, Ikzf2, Nrros, Kit, Mgat4b, Bcl9l, Tnfaip8, Tgm2, Pigl, Gpr97, Cmtm4, Inf2, Rcor2, Pank1, Cila2a, Hist3h2ba</i>

Clusters 1, 2 and 4 include genes that are predicted to decrease during the course of blood, fig. 3.35. Cluster 2 decreases during the earliest periods of the time-course while cluster 4 is decreasing from the beginning but has expression falls more rapidly during the second half of the time-course. Cluster 1 appears to stay constant over the first-half and then shows falling expression in the second-half of the pseudotime. Genes contributing to these clusters include *Kdr, Fli1, Sox11, Klf5, Tbx3, Igf2* and *Hhex* amongst others.

Clusters 3 and 5 are composed of genes that show an activating expression profile over the inferred pseudotime with cluster 5 showing early plateauing compared to cluster 3 fig. 3.35. These clusters include genes such as *Hbb-bh1, Hba-x, Blvrb, Angptl4, Cercam, Itga2b, Nfe2, Cited4, Gfi1b* and *Gata1*.

Cluster 6 interestingly shows dynamics of activation with subsequent deactivation during this inferred haemogenic development. It includes genes such as *Tie1, Bcr, Bcl11b, Dusp2, Ikzf2, Kit* and *Ctla2a*.

These lists contain genes known to be involved in embryonic blood development and their known dynamics support the pseudotime ordering but also genes previously not associated with embryonic haemogenesis are identified and they may provide promising avenues for further study revealing hitherto unrecognised targetable pathways.

3.11 Coarse temporal data corroborates findings

A useful feature of this experiment is the collection of metadata, this can provide supporting evidence and inform certain choices we make, for example the starting point of a *pseudotime*. The stage of the embryo is one of the features that was included in the metadata, it gives coarse temporal information and can provide useful insights into the underlying biology.

Taking the whole data set but excluding E6.5 the tSNE was recalculated, fig. 3.36. The reason for excluding E6.5 cells is that there is a substantial developmental gap between this time point and the next so that intermediate cells are likely to have been missed by our sampling strategy as is evidenced by the separation of the E6.5 epiblast cluster from the remainder of the timepoints which show reasonable mixing on the original tSNE fig. 3.18

To evaluate how cell populations evolve over developmental time, cells were plotted in separate panels by stage of embryo from which the cell was harvested, fig. 3.37. Care must be taken when interpreting plots split by embryo stage because of the pitfalls of not taking account of the different sorting strategies deployed so it is useful to keep fig. 3.2 in mind, as it provides an overview of the experimental protocol and therefore the cell population that was sampled. The number of FLK1⁺ sorted cells were about the same at all stages but no CD41⁺ were sorted at the primitive streak stage.

At the primitive stage there appear to be three groups, two not so discrete within the nascent mesoderm cluster (see fig. 3.36a) and one in the blood progenitor population that goes on to expand substantially in the neural plate and head fold time points bifurcating to form blood and possibly endothelium, lower arrow in fig. 3.37. This supports initiating a pseudotime trajectory from the blood progenitor population rather than the posterior subset of the nascent mesoderm or even the epiblast.

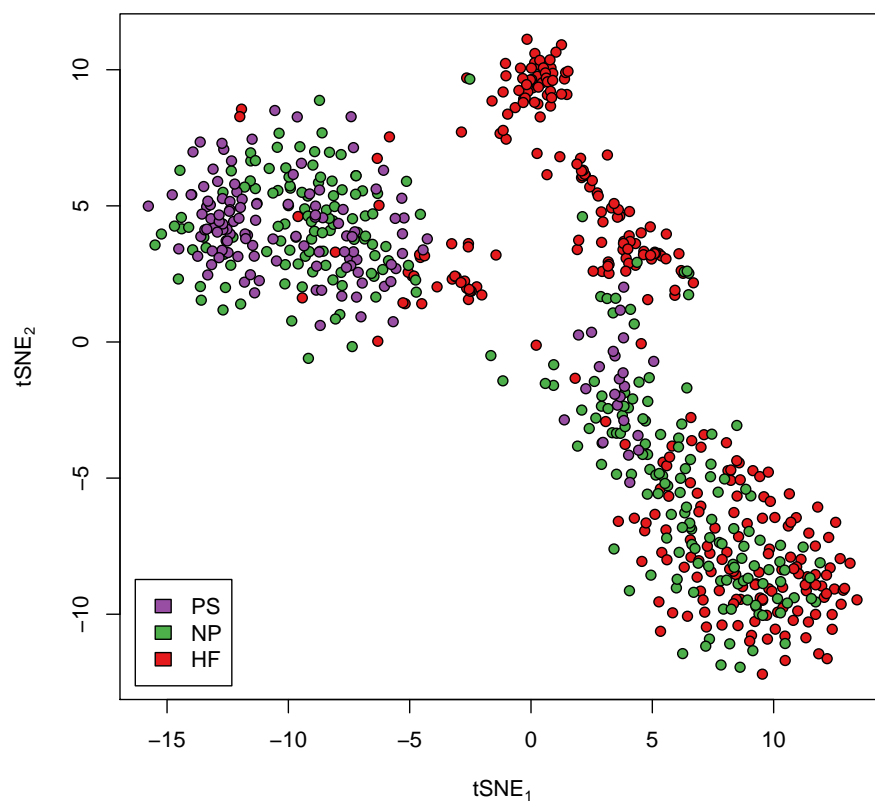
Moving on to the neural plate stage there appears to be significant proliferation of the group in the blood progenitor population but this may arise from enriching CD41⁺ by the sort strategy rather than an underlying biological effect. Another observation is that the anterior compartment of the nascent mesoderm has been depleted of cells as compared to the PS stage while the cell density in the posterior compartment has increased.

By the head fold stage the region populated by nascent mesoderm is depleted and the FLK1⁺ cells are now endothelial or pharyngeal mesoderm. The pharyngeal mesoderm in particular is replete of cells, the origin of this cluster of cells though is unclear from this data set. It is also interesting that the density of cells in the blood progenitor and the subset of the posterior

Single-cell transcriptomic analysis of murine gastrulation

mesoderm cluster nearest to it appear subtly less populated while this is not apparent in region near the posterior mesoderm cluster interspersed with the nascent mesoderm.

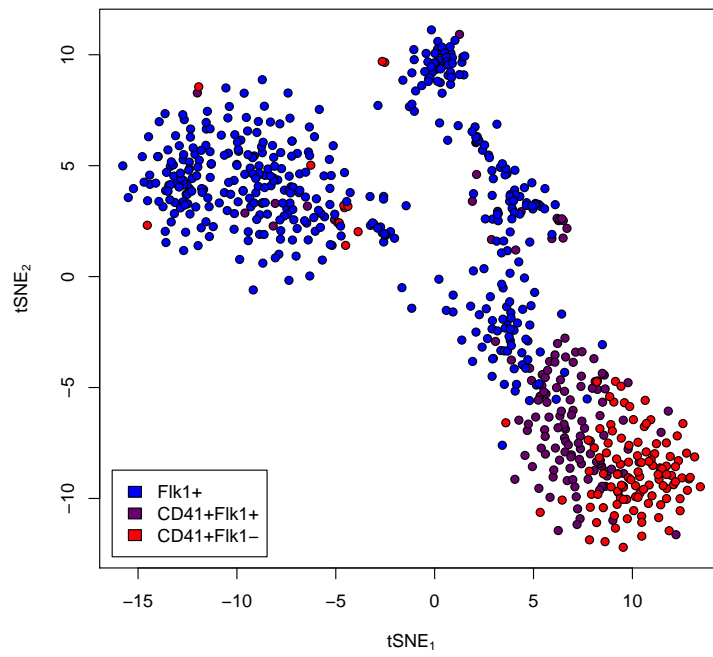
Another population that is almost exclusive to the head fold stage is the allantois and it is adjacent but non-overlapping with the posterior compartment of the nascent mesoderm and appears to be derived from the posterior and nascent mesodermal cells. Additionally the erythroid cluster particularly towards the apex has increased cell density at the HF stage as compared with the NP stage, possibly indicating the emergence of terminally differentiated embryonic blood.



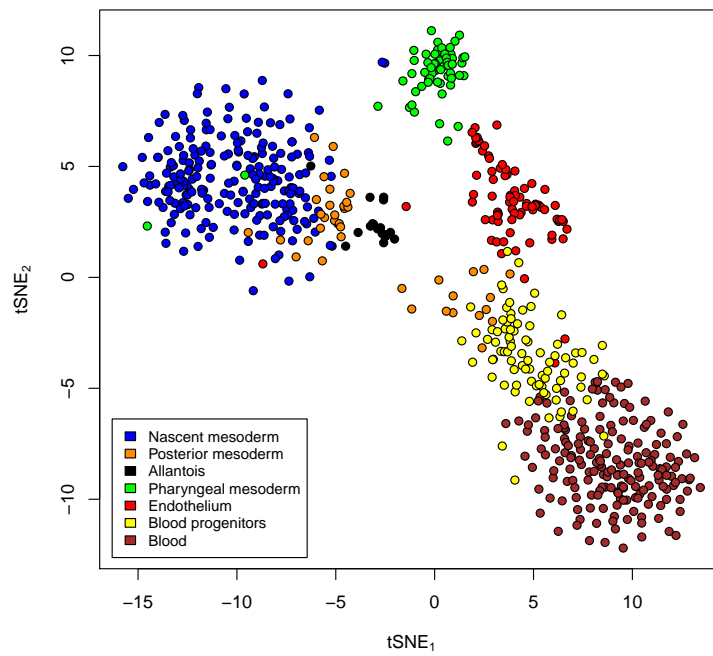
(a)

Fig. 3.36 tSNE performed after removing the E6.5 cells and recalculating highly variable genes. (a) Coloured by stage of embryo from which cell was harvested. (b) Coloured by sorting strategy. The CD41⁺ sorted cells were indexed for FLK1. (c) Coloured by cluster assignment.

3.11 Coarse temporal data corroborates findings



(b)



(c)

Fig. 3.36 tSNE after removing E6.5 cells (contd.)

Single-cell transcriptomic analysis of murine gastrulation

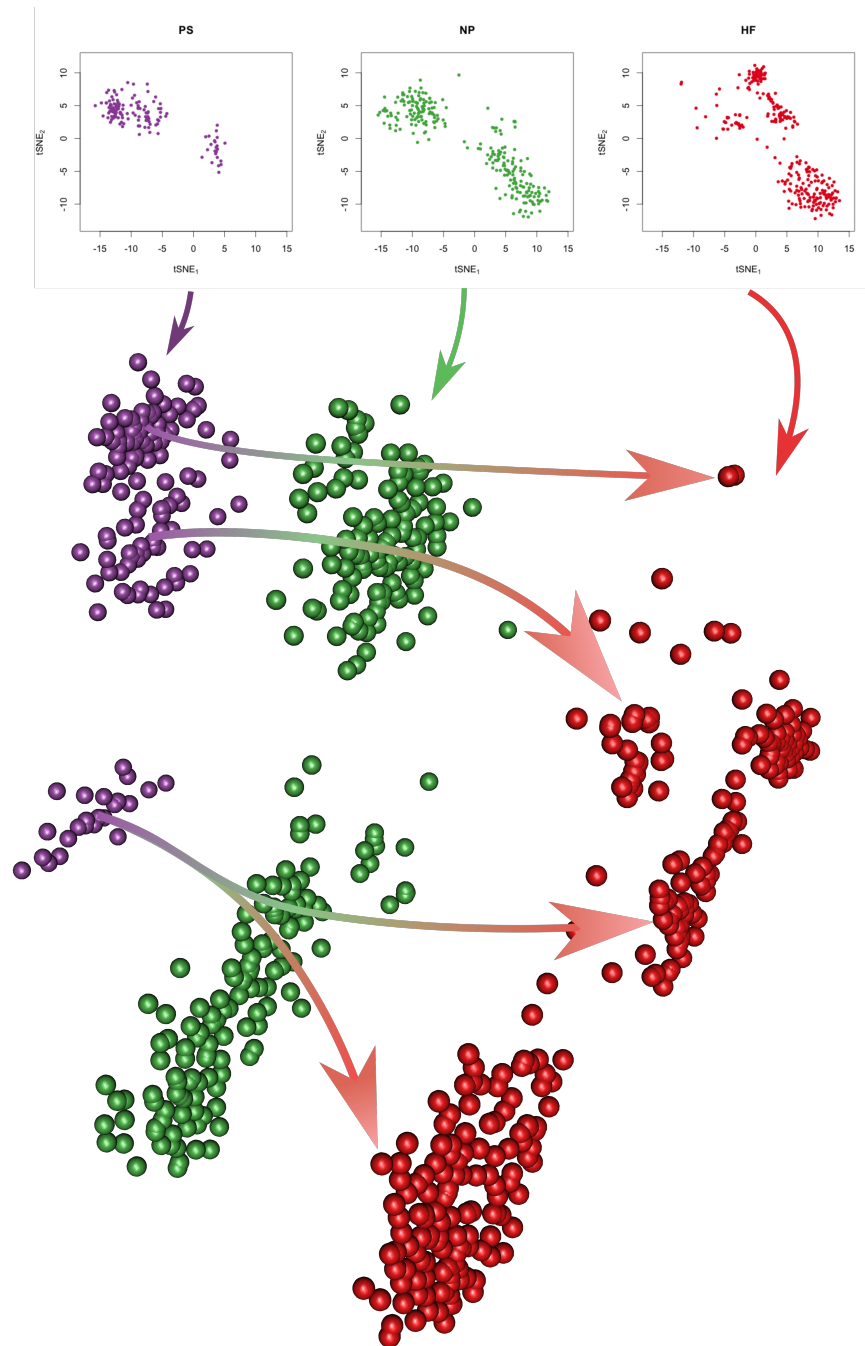


Fig. 3.37 Coarse temporal separation of cells. Separated by stage of embryo from which the cell was harvested, primitive streak (PS, purple), neural plate (NP, green) and head fold (HF, red) stage.

3.12 Finding substructure within clusters

Data driven cluster identification has robustly defined cell subpopulations but it is clear from the dendrogram in fig. 3.11 that there may be biologically meaningful substructure within these ‘macroclusters’. Polarisation of the epiblast and the anterior/posterior axis of the nascent mesoderm clusters demonstrate this point (see sections 3.8 and 3.9 for further details).

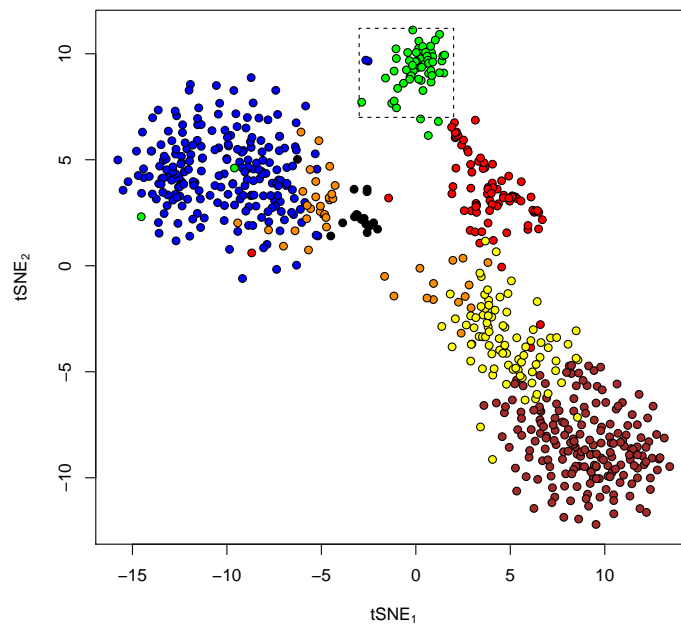
Further delineation of the clusters is possible by utilising even relatively limited prior knowledge of biological processes and developmental biology. This is exemplified within this dataset by studying the pharyngeal mesoderm and endothelial clusters more closely as described below.

3.12.1 Structure within the pharyngeal mesoderm

The second heart field (SHF) is known to arise from the pharyngeal mesoderm and express the molecular marker *Nkx2.5* early in development [Zhang et al., 2014]. The pharyngeal mesoderm cluster contains both *Nkx2.5*⁺ and *Nkx2.5*⁻ cells on the tSNE, fig. 3.38. The *Nkx2-5* expressing cells of the pharyngeal mesoderm likely constitute the second heart field.

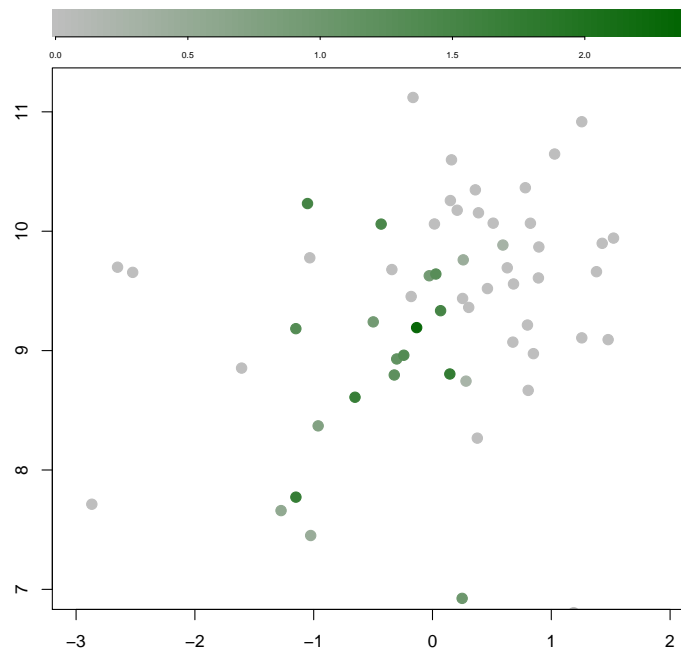
To identify genes related to early cardiac specification, cells were empirically split by *Nkx2.5* expression into two groups using a normalised log transformed gene count threshold of 0.5. Differential gene expression identified 14 genes (Mann Whitney U-test with multiple testing correction, 5% FDR), fig. 3.39. Several of these genes, *Mef2c* [Vincentz Joshua W. et al., 2008], *Gata5* [Bonachea et al., 2014; Hempel et al., 2017], *Cfc1* [Goldmuntz et al., 2002], *Isl1* [Colombo et al., 2018; Peng et al., 2013; Zhuang et al., 2013], *Hand2* [Laurent et al., 2017], *Gata4* [Neshati et al., 2018; Yu et al., 2018] and *Tcf21* [Yang et al., 2017] are known to be involved in cardiac development. In fact *Isl1* is considered to be a marker of the second heart field.

CRYPTIC (encoded by *Cfc1*), is a protein known to be associated with heterotaxia, particularly conotruncal cardiac malformations and isomerism, is expressed at the head fold stage in FLK1 positive cells with high *Nkx2.5* expression [Goldmuntz et al., 2002]. Notably we have also identified *Lrrn4*, a gene that is highly correlated with CRYPTIC within the pharyngeal mesoderm cluster as a novel gene that may be involved in early cardiac development, exemplifying how this rich dataset can be used generate new hypotheses.



(a)

Nkx2-5



(b)

Fig. 3.38 Figure showing *Nkx2-5* expression within a subset of pharyngeal mesoderm cells likely corresponding to the second heart field. (a) shows region of tSNE plot on which expression of *Nkx2-5* is displayed in (b).

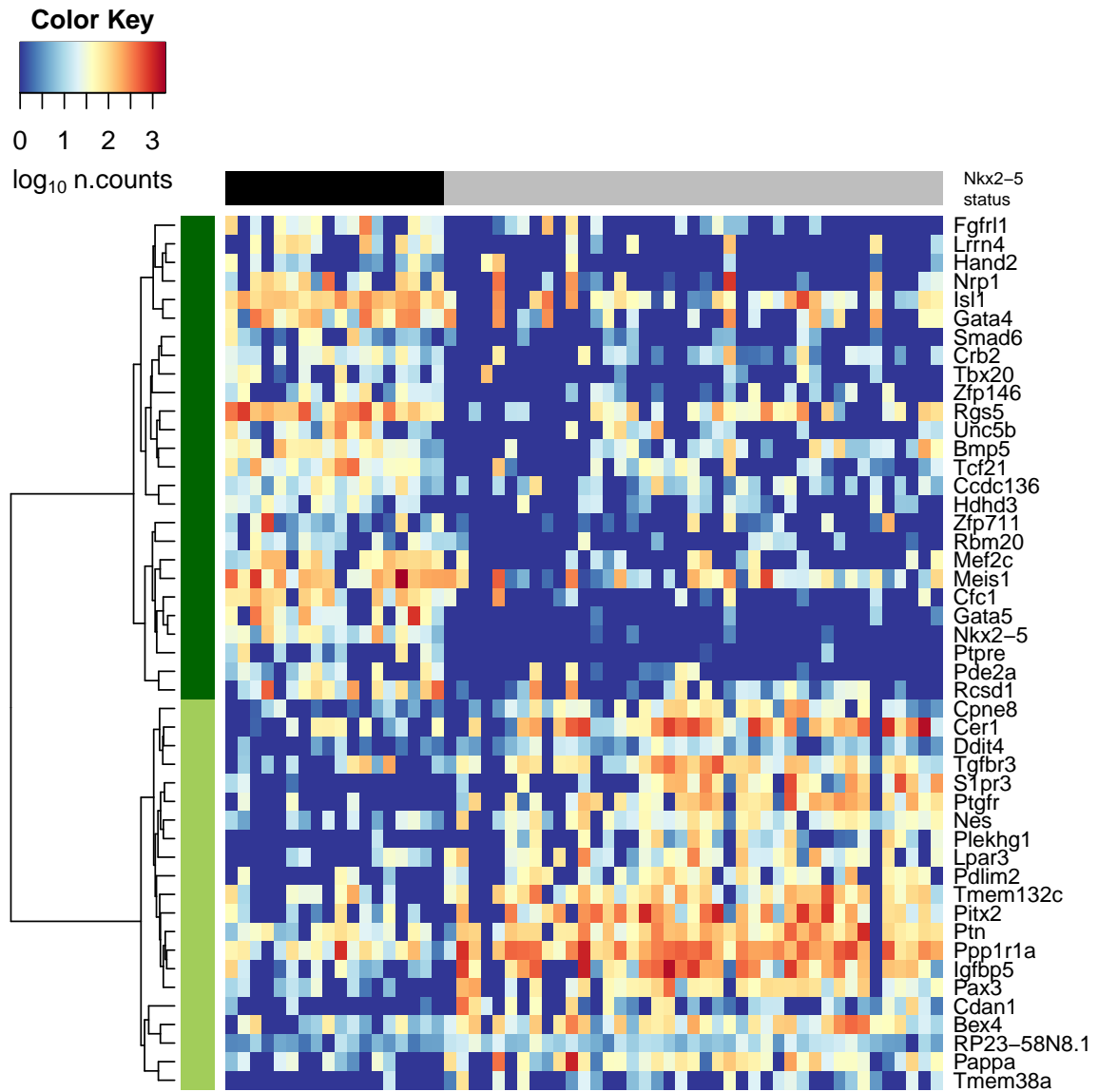


Fig. 3.39 Heatmap displaying \log_{10} normalised counts within the pharyngeal mesoderm cluster cells in columns and genes found to be differentially expressed between $Nkx2.5^+$ (Black in horizontal bar) and $Nkx2.5^-$ cells (Grey in horizontal bar) in rows. Rows are clustered, there is no clustering on the columns. Row side colours: Dark green - Up-regulated in $Nkx2.5^+$; Light green - Down-regulated in $Nkx2.5^-$.

3.12.2 Discovery and validation of a novel pathway

Taking the same approach as described for the pharyngeal mesoderm, a sub-cluster was identified within the endothelial cluster corresponding to a putative erythroid myeloid progenitor (EMP) subset. The EMP population was identified using prior knowledge of genes related to induction of the EMP programme in haemogenic endothelium, namely *Gfi1b*, *Itga2b* (CD41) and *Itgb3* (CD61). Analysis was limited to the endothelial cell population, the suggested source of EMPs.

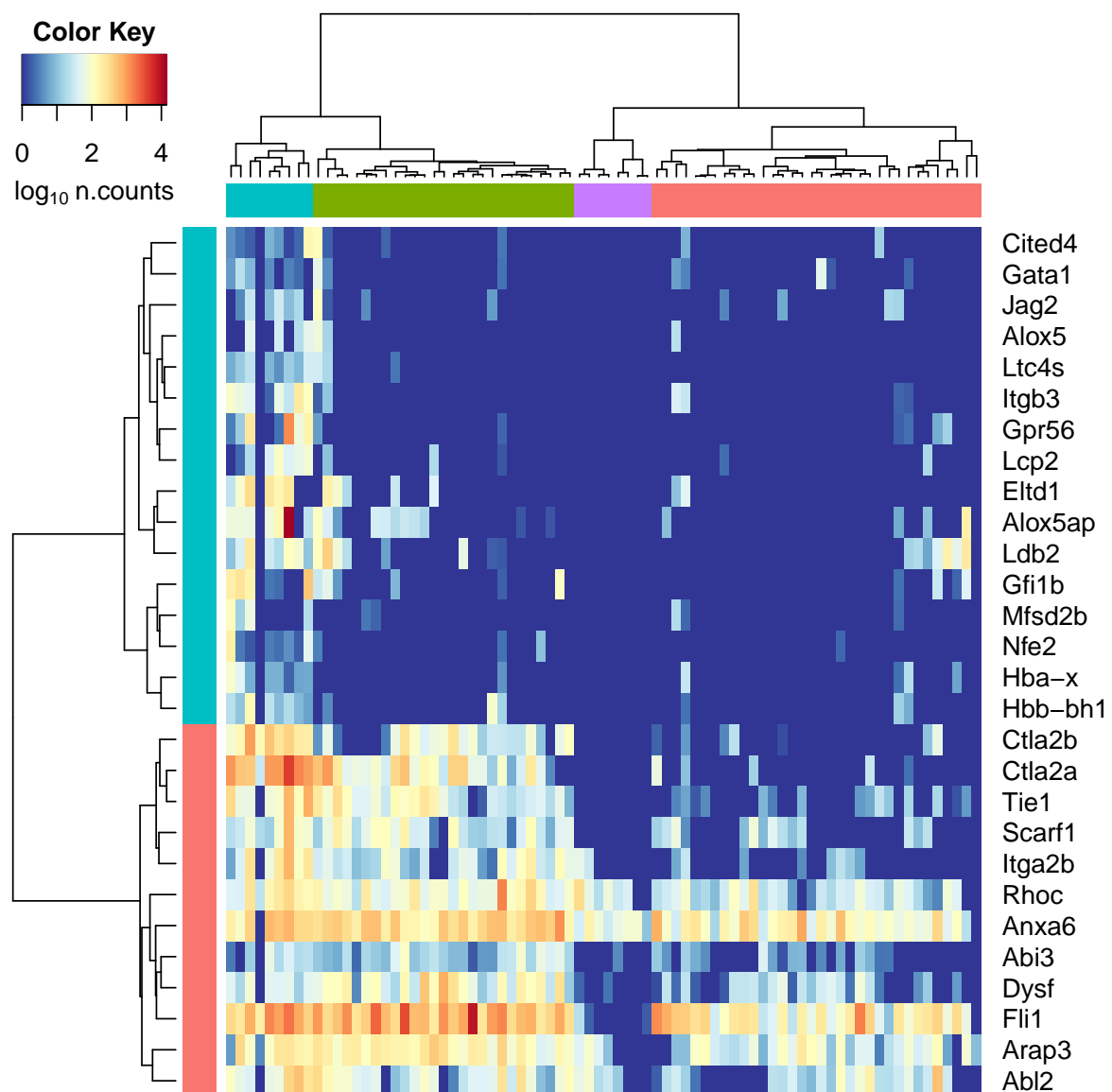


Fig. 3.40 Clustering of endothelial cluster using genes that are highly correlated with *Gfi1b*, *Itga2b* and/or *Itgb3*.

3.12 Finding substructure within clusters

Spearman's rank correlation was used to identify genes highly correlated with *Gfi1b*, *Itga2b*, *Itgb3*. Multiple testing correction was applied by weighting genes as described by Ignatiadis et al. [2016] and $p < 1 \times 10^4$ for the adjusted p-values was considered significant, yielding 28 highly correlated genes. Hierarchical clustering resulted in 2 major gene clusters and 4 cell sub-types, fig. 3.40. The cyan colour in the bar over the heatmap in fig. 3.40 indicates the cluster of cells that appears to have a gene expression signature consistent with an EMP progenitor. Gene ontology enrichment focusing on biological processes, for the 28 highly correlated genes strongly suggest a role for the leukotriene/lipoxin branch of the arachidonic acid pathway in yolk sac definitive-wave haematopoiesis, see table 3.9. Nearly all of the top 10 hits are related to the arachidonic acid pathway specifically to the genes *Alox5*, *Alox5ap* and *Ltc4s*.

Table 3.9 GO Biological process enrichment analysis using enrichR. Hits are ordered by adjusted p-value and top 10 are shown with p-values. Enrichment analysis performed using enrichR available from CRAN (<https://cran.r-project.org/package=enrichR>) or GitHub (<https://github.com/wjawaid/enrichR>). Database interrogated is GO_Biological_Process_2017.

GO Biological process term	adjusted p-values ($\times 10^{-3}$)
lipoxin biosynthetic process (GO:2001301)	3.8
leukotriene catabolic process (GO:0036100)	3.8
lipoxin metabolic process (GO:2001300)	3.8
lipoxin B4 metabolic process (GO:2001304)	3.8
lipoxin A4 metabolic process (GO:2001302)	3.8
green leaf volatile biosynthetic process (GO:0010597)	3.8
leukotriene D4 biosynthetic process (GO:1901750)	6.2
regulation of cell migration (GO:0030334)	6.2
leukotriene D4 metabolic process (GO:1901748)	6.2
leukotriene metabolic process (GO:0006691)	6.2

This leukotriene branch of the arachidonic acid pathway has not previously been implicated in embryonic haemogenesis while there is evidence of other arachidonic acid metabolic branches being involved in zebrafish embryonic haematopoiesis: prostaglandin E2 and Epoxyeicosatrienoic acids have been shown to play critical roles [Li et al., 2015; North et al., 2007].

We then took advantage of our recently published chromatin immuno-precipitation sequencing (ChIP-Seq) reference experiment looking at several key haematopoiesis related transcription factors in a mouse embryonic stem cell (ESC) differentiation model of haematopoiesis [Goode et al., 2016]. The data can easily be accessed through CODEX (<http://codex.stemcells>).

Single-cell transcriptomic analysis of murine gastrulation

cam.ac.uk/). First we looked for binding of key transcription factors TAL1/SCL, PU.1, FLI1, LMO2, GFI1 and GFI1B around gene loci of *Alox5*, *Alox5ap* and *Ltc4s*.

There were several consistent peaks found at the *Ltc4s* locus in haematopoietic progenitors, haemangioblasts and haemogenic endothelium. *Alox5* and *Alox5ap* loci also show consistent convincing binding peaks (not shown, can be looked up at <http://codex.stemcells.cam.ac.uk/>). This provides supportive evidence that *Ltc4s*, a key enzyme in the leukotriene pathway, is a direct target of several haematopoietic transcription factors.

Next we interrogated DNAase I hypersensitivity data from Goode et al. [2016]. This showed that the genomic region at the peak located $\approx 2kb$ from the TSS for *Ltc4s* is open in both haemogenic endothelium and haematopoietic progenitors but not in ontogenically earlier mesodermal progenitors or haemangioblasts nor in the more mature macrophages fig. 3.42.

Arachidonic acid is a poly-unsaturated fatty acid and an integral part of the phospholipid bilayer forming the cell membrane. Arachidonic acid is metabolised through several pathways forming active substances including prostaglandins, eicosanoids and leukotrienes all of which have been implicated in both physiological and pathological inflammatory responses.

Alox5, *Alox5ap* and *Ltc4s* form a complex on the outer nuclear envelope and generate leukotrienes in particular leukotriene C₄ (LTC₄). The pathway is summarised in fig. 3.43, produced by Dr Vasilis Ladopoulos. Arachidonic acid itself is generated by the action of phospholipases on phospholipids from the cell membrane phospholipid bilayer.

To further study the role of the leukotriene pathway in HE and EMP ontogeny, we used a frequently used in-vitro embryonic stem cell (ESC) differentiation assay [Coulombel, 2004; Pereira et al., 2007]. This assay is designed to assess the number of seeded cells that can go on to form haematopoietic colonies. Further it can be used to assess the type of colonies generated. Rather than liquid culture where cells are mobile in this semi-solid medium cells are plated at low density and then the number of colonies counted prior to them merging with cells from neighbouring colonies.

These experiments were all performed by Dr Vasilis Ladopoulos. Mouse ESCs maintained in serum/LIF were differentiated into embryoid bodies (EBs) and exposed to either Zileuton, a pharmacological *Alox5* inhibitor approved clinically for use in patients with asthma or to LTC₄, fig. 3.44. At Day 4, EBs were dissociated and the compounds washed out. Cells were counted and plated at equivalent numbers in methylcellulose to perform the short-term colony forming cell (CFC) assay [Coulombel, 2004; Pereira et al., 2007]. Total colonies were enumerated at day 14, colony morphology itself was not assessed.

3.12 Finding substructure within clusters



Fig. 3.41
117

Single-cell transcriptomic analysis of murine gastrulation



Fig. 3.41 (contd.) Tracks of Chip-Seq data (from UCSC genome browser using CODEX <http://codex.stemcells.cam.ac.uk/>) for transcription factors *Lmo2*, *Spi1*, *Runx1*, *Gfi1*, *Tal1/Scf* and *Fli1* and their genomic binding in haematopoietic progenitors (Figure previous page) and haemangioblasts and haemogenic endothelium (Above). A consistent peak in all populations for several transcription factors is found $\approx 2\text{kb}$ upstream of the transcription start site (TSS) of *Ltc4s*. The peak is located within an evolutionary conserved region of the genome (see conservation tracks bottom of figure previous page).

3.12 Finding substructure within clusters

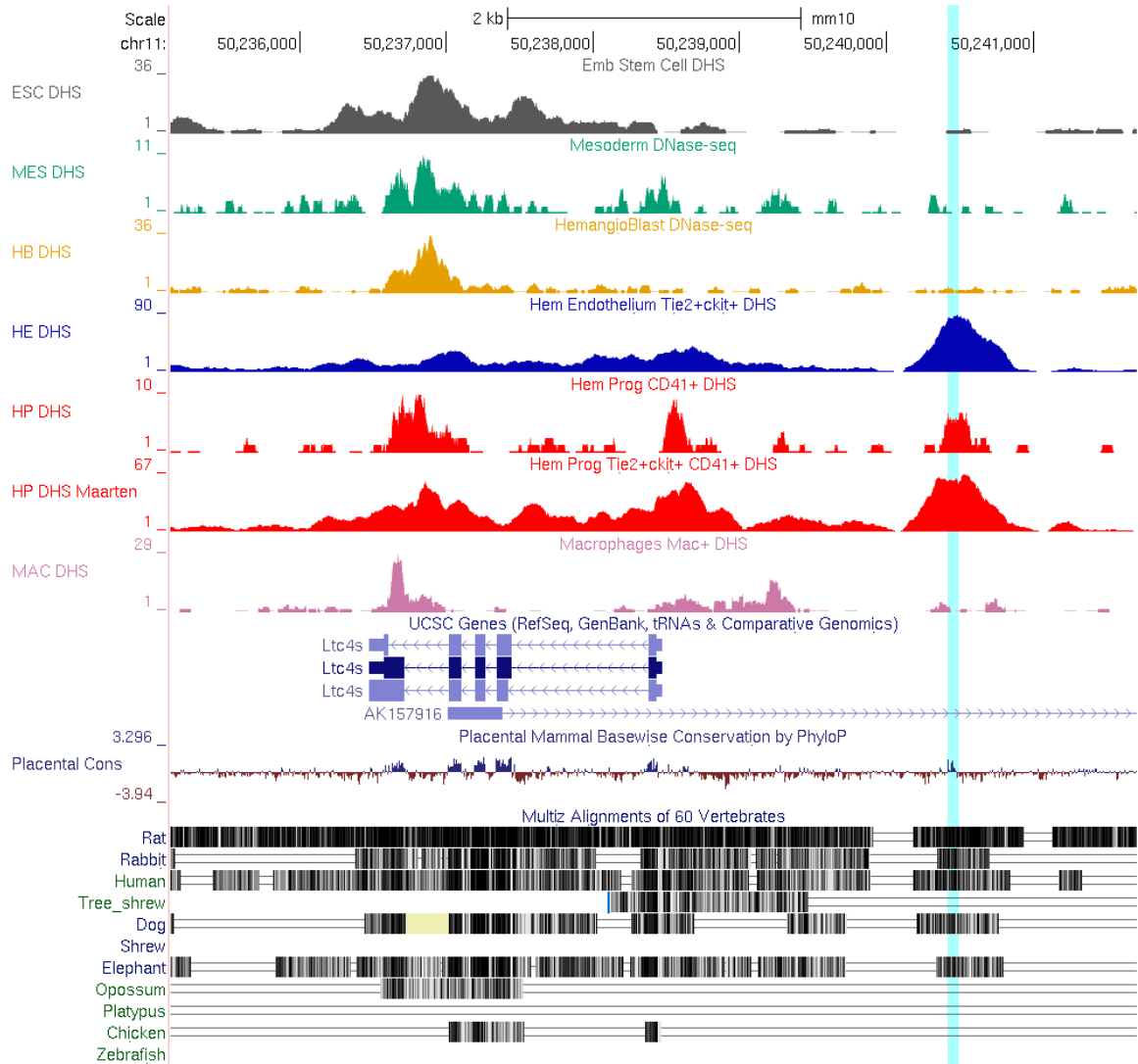


Fig. 3.42 DNAse I hypersensitivity tracks in embryonic stem cells (ESC), mesodermal cells (MES), haemangioblasts (HB), haemogenic endothelium (HE), haemogenic progenitors (HP) and macrophages (MAC). Confirming that the genomic locus previous noted for binding of haematopoietic transcription factors in fig. 3.41 has an open chromatin conformation in HE and HPs but not in ESCs, MES or MACs. The tracks at the bottom show evolutionary conserved sites as in fig. 3.41.

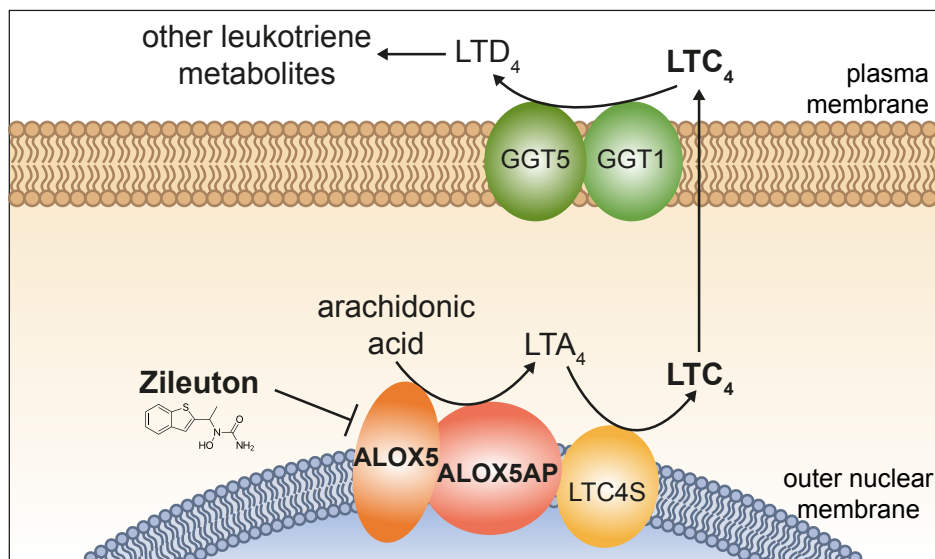


Fig. 3.43 Schematic diagram summarising the role of ALOX5, ALOX5AP and LTC4S in the leukotriene pathway - They form a complex on the outer nuclear envelope that converts arachidonic acid to LTA₄ and subsequently to LTC₄. LTC₄ can then be transported out of the cell and converted to downstream leukotrienes like LTD₄ by gamma-glutamyltransferases (GGT1 and GGT5). Zileuton is a synthetic drug that inhibits ALOX5. Arachidonic acid itself is produced by the action of phospholipases. Figure produced by Dr Vasilis Ladopoulos and for the paper Scialdone et al. [2016].

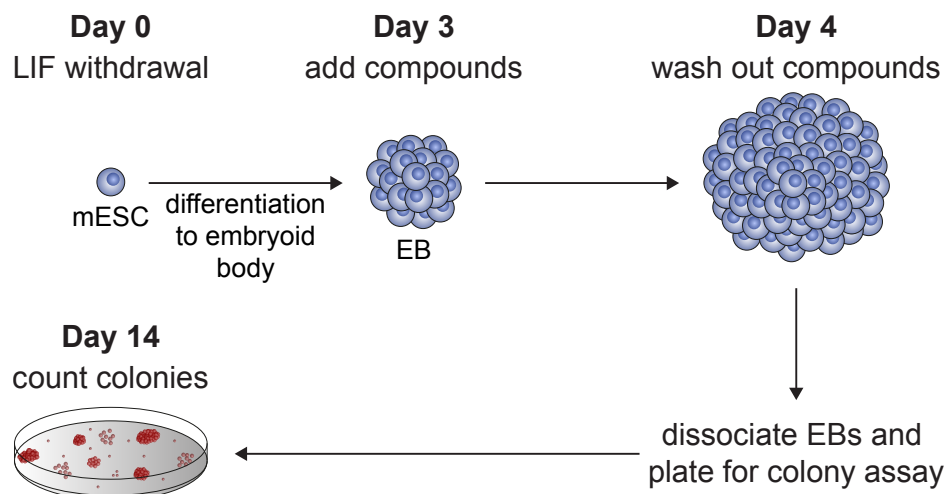


Fig. 3.44 Overview of experimental protocol generating erythroid bodies (EBs) from murine embryonic stem cells (mESCs) and subsequent colony assays after adding either Zileuton or LTC₄ at Day 3. Cells are cultured and colonies counted before reaching confluence at Day 14. LIF, Leukaemia Inhibitory Factor.

3.12 Finding substructure within clusters

Zileuton, LTC₄ or carrier were added during EB generation on day 3. Zileuton was added at 10 μ M, 50 μ M and 100 μ M and LTC₄ at 50 nM, 100 nM and 300 nM. All experiments were performed in triplicate.

Zileuton at 10 μ M and LTC₄ at 50 nM had no effect on colony counts, fig. 3.45. Zileuton at higher concentration significantly inhibited colony formation in a dose-dependent manner while addition of LTC₄ significantly stimulated colony numbers again in a dose-responsive fashion, fig. 3.45.

This provides strong supportive evidence from an in vitro assay that the leukotriene branch of the arachidonic acid pathway alters the number of haematopoietic progenitors present in EBs at Day 4.

Taken together with finding of activation of the leukotriene pathway (fig. 3.40 and table 3.9) in vivo suggests that this branch of the arachidonic acid pathway plays a key role in the yolk-sac embryonic erythroid myeloid progenitor haematopoiesis.

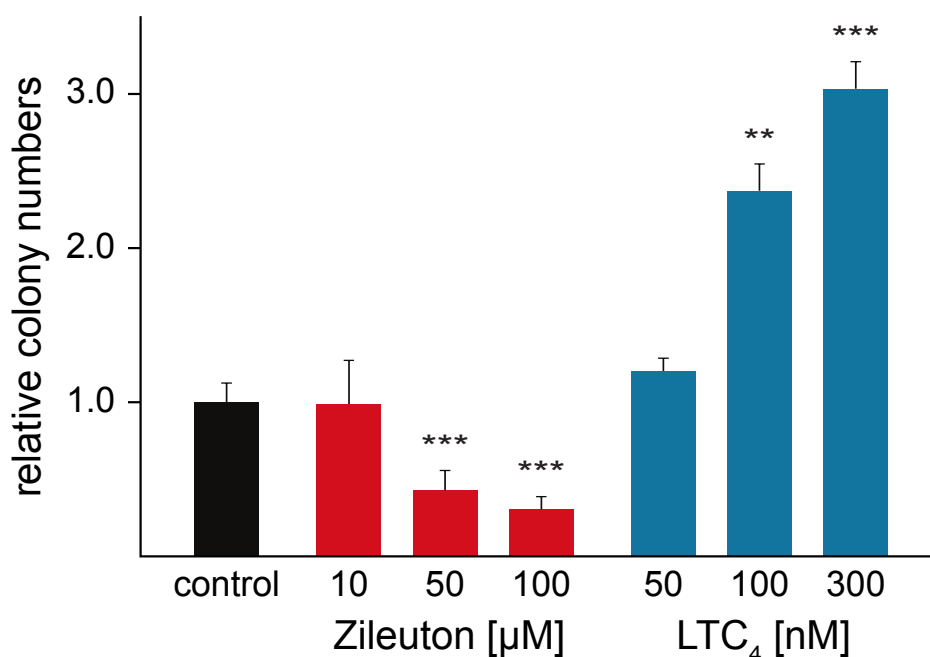


Fig. 3.45 Bar plot showing the fold change in the number of colonies relative to carrier control when embryoid bodies were treated with the indicated concentrations of Zileuton or LTC₄ for 24 hours. Error bars represent the mean plus standard deviation of $n = 3$ biological replicates. Significant changes compared to controls were tested with a one-tail Student's t-test ($p = 0.004$ for Zileuton (50 μ M); 0.002 for Zileuton (100 μ M); 0.027 for LTC₄ (100 nM); 0.007 for LTC₄ (300 nM)). Figure produced by Dr Vasilis Ladopoulos.

3.13 Single-cell resolved characterisation of a mutant

The T-cell acute lymphocytic leukaemia protein 1 encoded by *Tall* also known as stem cell leukaemia (*Scl*) is a basic helix-loop-helix transcription factor that is essential for haematopoiesis and as shown previously is expressed throughout the time-course from the blood progenitor to blood clusers, fig. 3.34.

Shivdasani et al. [1995] generated a *Tall* knock-out mouse line and mated heterozygous *Tall*^{+/-} mice, finding no reduction in the number of homozygote knock-out *Tall*^{-/-} embryos at E8.5 but by E10 all surviving embryos were either homozygous wild-type *Tall*^{+/+} or heterozygotes *Tall*^{+/-}, there were no surviving *Tall*^{-/-} embryos. They also found that by E9.5 *Tall*^{-/-} embryos were smaller, with profound pallor, had a dilated pericardial sac and that their yolk sacs and placentas were bloodless. Despite this other major developmental milestones were achieved normally, including chorio-allantoic fusion, rotation of the embryo, closure of the caudal neural pore and initiation of cardiac contractions. Other mesodermal tissues including somites, blood vessels and myocardium were present and foetal loss was attributed to the severe anaemia resulting from an early block of embryonic erythropoiesis.

Porcher et al. [1996] used haematopoietic in-vitro differentiation assays and chimeric mice to overcome the universal embryonic lethality of the homozygous *Tall*^{-/-}. They demonstrate an absolute requirement for *Tall* in the generation of all haematopoietic lineages. Lancrin et al. [2009] demonstrate that *Tall* is indispensable for the establishment of haemogenic endothelium. *Tall* has also been used in combination with other transcription factors to re-programme fibroblasts to generate haemogenic progenitors through a haemogenic endothelium intermediate [Batta et al., 2014].

Given the early and central role of *Tall* in early embryonic and definitive haematopoiesis we sought to study the in vivo differences in gene expression between wild type (WT) and *Tall*^{-/-} mice in the FLK1⁺ sorted population during during gastrulation.

The *Tall*^{lacZ/+} knock-in mouse was used, fig. 3.46 [Elefanty et al., 1998]. Since the background of these mice, generated from the ES cell line W9.5, an embryonic cell line derived from the 129/S mouse strain [Stevens, 1970; Szabo and Mann, 1994], was different from the CD1 mouse line used in the previous experiments it was deemed prudent to additionally sample some wild-type embryos on this background.

Dr Yosuke Tanaka and I performed the embryo dissections and generated the single cell suspension. During dissection some extra-embryonic tissue was set aside for genotyping to identify the homozygous embryos to be used for further study. The generated single

3.13 Single-cell resolved characterisation of a mutant

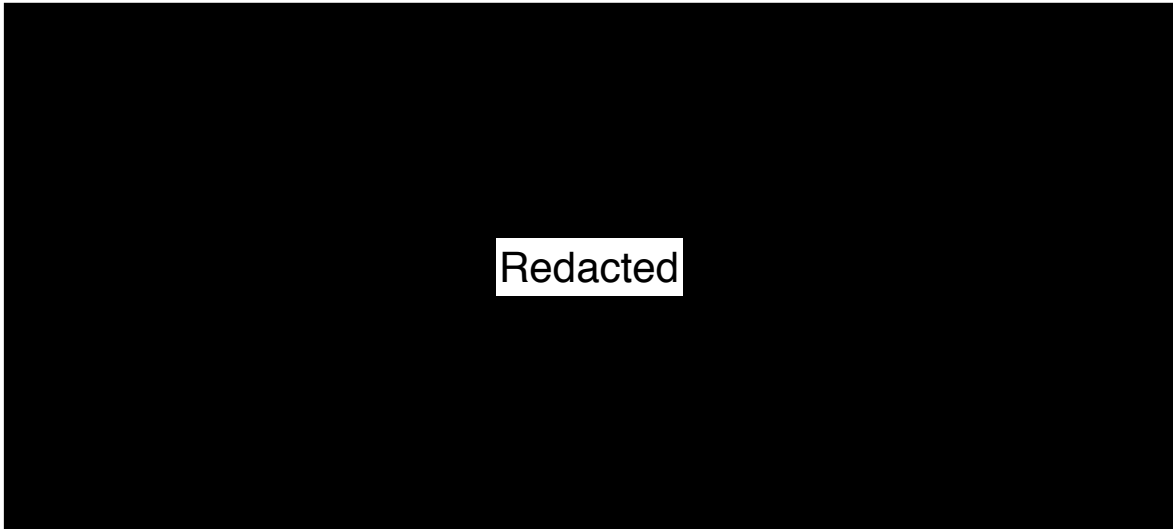


Fig. 3.46 The *Tall* (also known as *Scl*) LacZ knock-in mutant generation strategy (from Elefanty et al. [1998]). The SCL-lacZ targeting vector is shown above the *Tall* genomic locus with regions of expected homologous recombination shown by crosses. Exons IV and V are completely replaced and exons III and VI are partially replaced knocking out *Tall* transcription at the locus.

cell suspensions were single cell sorted into lysis buffer plates using FACS for FLK1 with indexing for CD41, fig. 3.47. This confirmed that the *Tall*^{-/-} embryos were almost completely devoid of CD41⁺ cells.

The sorted cells were stored at -80 °C awaiting genotype characterisation of each embryo. Genotyping was performed by Victoria Moignard and images of the genotyping gels are shown in fig. 3.48.

Selecting 8 homozygous embryos, marked by the red asterisks in fig. 3.48, the appropriate plates were thawed and sequencing libraries prepared by Victoria Moignard as previously described in section 3.3. There were 3 embryos at neural plate stage, 3 at head fold stage and 2 at the four somite stage (E8.25) table 3.10. The prepared libraries were submitted for sequencing and processed through the previously described pipeline of QC, normalisation and feature selection as described in sections 3.4 to 3.6. 482 cells passed QC and were used for downstream analysis. An example of QC performed on one of the 96-well plates from this dataset is shown in fig. 3.49 a comparison can be made with fig. 3.7. The ERCC to total mapped genes is significantly higher and this is suggestive of either less input mRNA or a higher amount of ERCC spike-in, the latter being more likely.

Additionally though both *Tall*^{+/+} and *Tall*^{-/-} samples were processed in parallel, exactly the same equipment could not be used simultaneously. This meant that the different genotypes

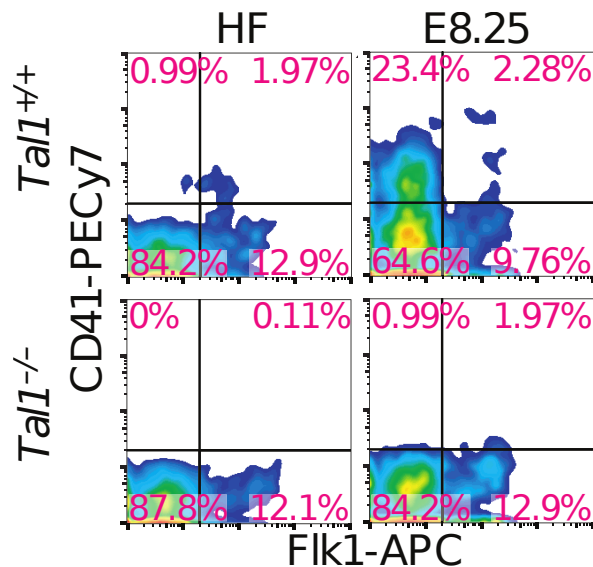


Fig. 3.47 Sample of flow cytometry of WT and *Tal1*^{-/-} embryos at head fold (HF) stage and at the four somite stage (E8.25). This shows a severe depletion of cells in the CD41⁺ gate in *Tal1*^{-/-} embryos compared with the *Tal1*^{+/+} embryos as would be expected from the known haematopoietic defect in *Tal1*-null mutants. Image produced from flow cytometry data by Victoria Moignard.

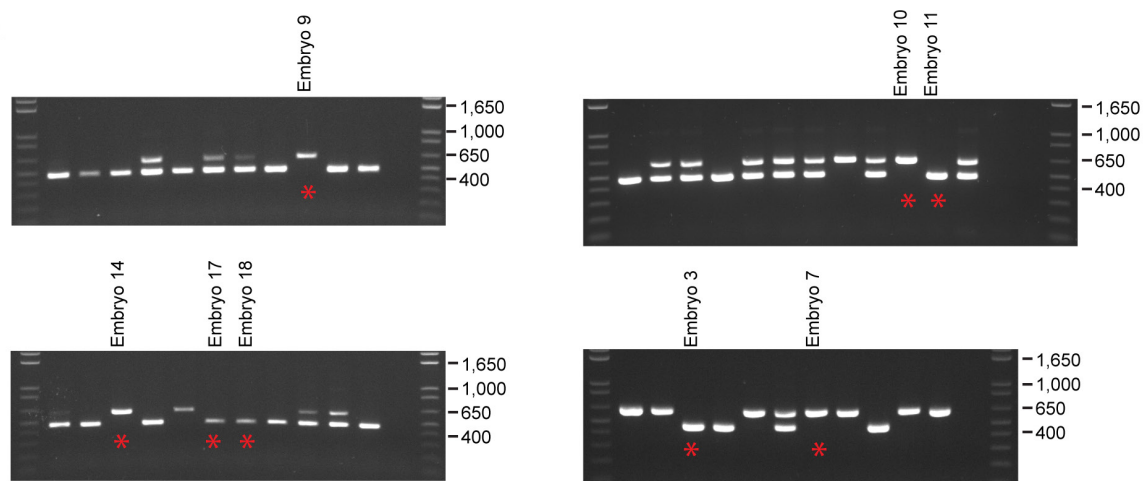


Fig. 3.48 PCR genotyping gels from *Tal1*^{LacZ/+} crosses. Lower band is the wild-type allele and upper band is the mutant allele carrying a neomycin knock in fig. 3.46. Presence of both bands indicates heterozygosity. Embryos from which sequencing data were obtained are indicated with a red star. The embryo numbers match those available in the metadata from Scialdone et al. [2016]. Genotyping and image produced by Victoria Moignard.

3.13 Single-cell resolved characterisation of a mutant

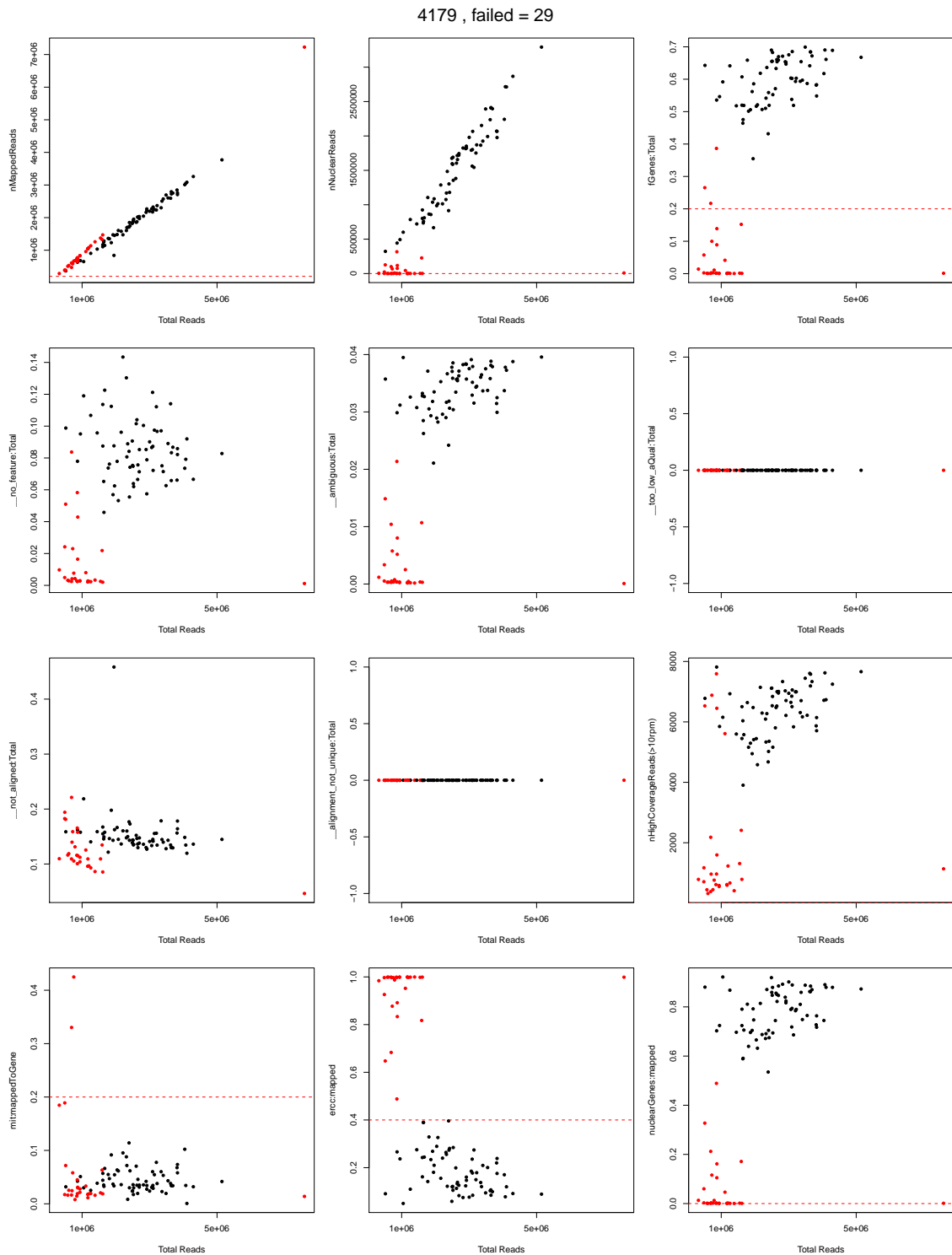


Fig. 3.49 QC plots generated for plate 4179 containing FLK1 sorted cells from a wild type *Tall*^{+/+} embryo. Comparison should be made to fig. 3.7 from the reference experiment and to cells sorted from a null mutant *Tall*^{-/-} fig. 3.50. The ratio of ERCC to total mapped genes is considerably higher in this sample as compared to the previous experiment suggesting higher input amounts of ERCC spike-ins.

Single-cell transcriptomic analysis of murine gastrulation

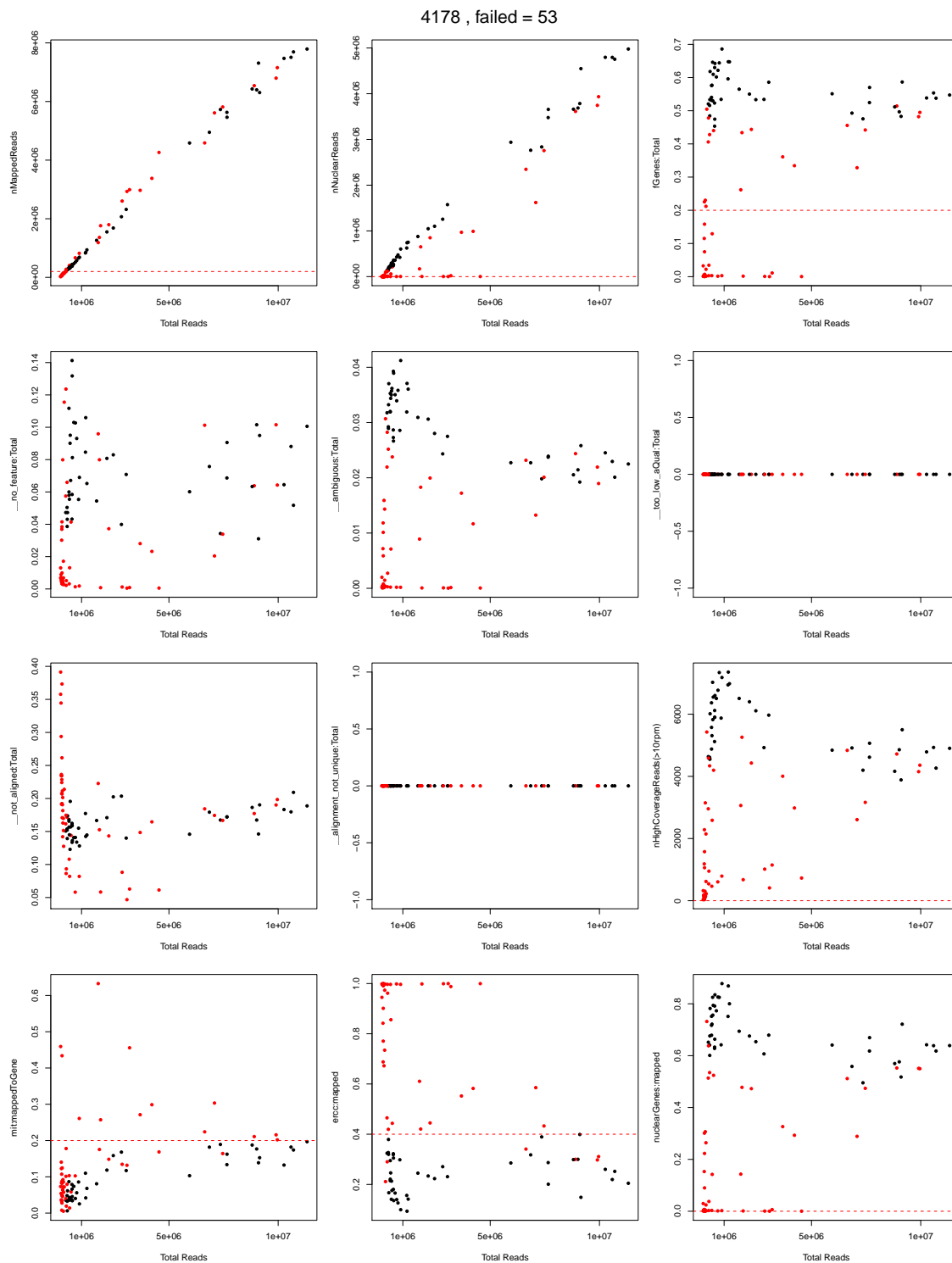


Fig. 3.50 QC plots generated for plate 4178 containing FLK1 sorted cells from a null mutant *Tall1*^{-/-}. Comparison should be made to fig. 3.49. The total number of counts appear to be much higher in some cells than others giving the impression of 2 clusters. The cause of for this is discussed in the text and summarised in fig. 3.51.

3.13 Single-cell resolved characterisation of a mutant

Table 3.10 Summary of *Tal^{-/-}* and *Tall^{+/+}* experiment. Embryonic stage, embryo number matching with fig. 3.48, Genotype with Null corresponding to *Tall^{-/-}* and WT with *Tall^{+/+}*, Lane showing the sequencing flow cell lane matching with fig. 3.51 and Cell number the number of cells that passed QC. NP - Neural plate stage, HF - Head fold stage and 4SP - 4 somite pairs stage.

embryoStage	embryo	Genotype	Lane	Cell number post initial QC	Cell number post final QC
NP	9	Null	4175	44	26
NP	17	WT	4179	67	57
NP	18	WT	4180	73	66
HF	10	Null	4176	56	36
HF	14	Null	4178	43	24
HF	11	WT	4177	62	60
4SP	7	Null	4174	57	35
4SP	3	WT	4173	80	73

were often processed simultaneously but consistently in different blocks of our thermocycler. Assessing the total number of reads mapped to nuclear genes and how the cells were arranged on the 96-well plates it becomes apparent that cells on the outskirts of the plate in the *Tall^{-/-}* experiments generated fewer reads compared to those located more centrally fig. 3.51. This is not so for the *Tall^{+/+}* plates. The most likely cause for this is drying of the peripherally located samples on each plate due to the thermocycler lid not fitting tightly. Each 96-well plate was sequenced in a separate lane of the sequencer and the total DNA concentration loaded on to each lane of the sequencing flow cell was titrated to that recommended by Illumina, this led to oversampling of the cells more centrally placed on the 96 well plate during sorting for the *Tall^{-/-}* plates.

A key concern with the over-sampled cells is duplicate reads. This is where the same mRNA molecule is counted twice. Post-tagmentation duplicates can be readily recognised in genes with low to moderate counts since reads are unlikely to start at exactly the same location due to the random nature of the tagmentation reaction, see fig. 3.4 for experiment protocol summary. Pre-tagmentation duplicates though cannot be easily recognised. Several of the outlier cells had greater than 80% duplicate read rate, all cells meeting this criteria were therefore excluded as a final step in the QC, table 3.10.

A popular method for visualising high-dimensional data that additionally allows projection on to a previously calculated representation is PCA. Performing PCA de-novo on the whole dataset after normalising together and re-calculating highly variable genes, highlights a

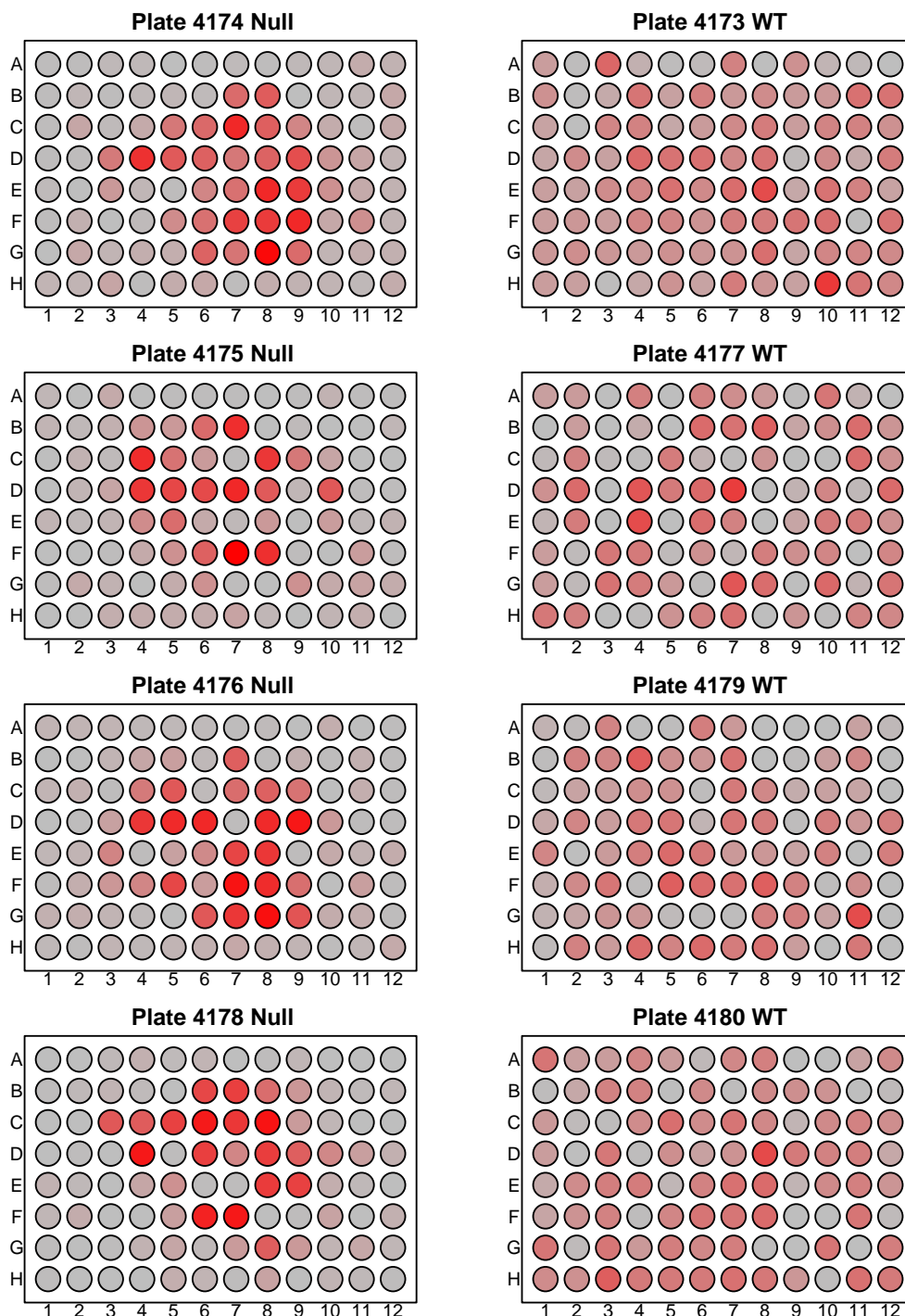


Fig. 3.51 Heatmap of mapped reads on the 96 well plates with null mutants $Tal1^{-/-}$ on the left column and $Tal1^{+/+}$ samples on the right. All samples of the same genotype were processed in the same block of the thermocycler but $Tal1^{-/-}$ plates were all processed on a different block from the $Tal1^{+/+}$ plates. In the $Tal1^{-/-}$ plates there is a clear preponderance for reads to be amassed by centrally located wells on the plate. In contrast reads are more evenly distributed across the plate in $Tal1^{+/+}$ plates.

3.13 Single-cell resolved characterisation of a mutant

significant number of outlier cells in the lower right fig. 3.52a. There is an additional set of outliers in the upper right that have not been specifically highlighted. In the lower right set there are 74 cells and 72 are in plates sorted from *Tall*^{-/-} mice. In fact they correspond very tightly with the highly sequenced cells shown in fig. 3.51. This suggests that the normalisation and feature selection methods alone in this case are not able to deal with experimental batch effects to recover signals of biological interest.

Focusing on the former de-novo PCA and not recognising that the outliers represent a batch effect, it may appear that these outlier cells represent an aberrant cell type produced in-vivo from the failure of normal haematopoiesis due to lack of the master regulator *Tall*. Differential gene expression though shows that nearly all genes are more highly expressed in this outlier set and it becomes apparent this has arisen due to the increased sequencing depth of some cells due to their position on the plate as discussed earlier.

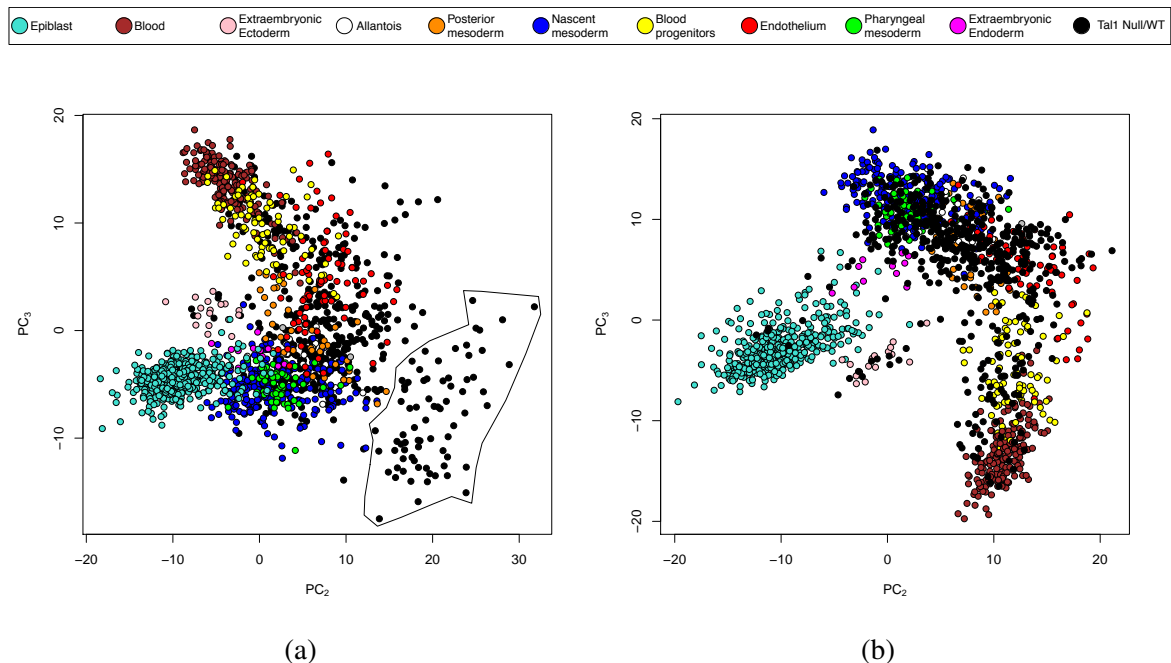


Fig. 3.52 (a) PCA calculated de-novo after normalising the combined experiments together and re-calculating highly variable genes. Polygon in the lower right outlines the cells from the *Tall* experiment that appear to be outliers. Each point is a cell and is coloured using the legend colours above note slightly different colour schema from fig. 3.13. (b) The new *Tall* dataset is projected onto the previously calculated PCA fitting neatly onto the previous populations and giving some indication of the cell types present.

An alternative to performing de-novo PCA is to use the same gene set as calculated previously for the reference dataset (see section 3.6) but sticking with the new normalised log-transformed counts and projecting onto the PCA calculated on the reference dataset, fig. 3.13.

Single-cell transcriptomic analysis of murine gastrulation

The cells from the new experiment now map neatly onto the old populations of cells fig. 3.52b. A caveat to keep in mind with such analysis is that a new cell state can possibly be missed if it cannot be identified with the previously selected feature set.

A random forest algorithm was trained on the reference dataset to predict cell type by Dr Antonio Scialdone as described in Scialdone et al. [2016]. The random forest algorithm was then used to assign cluster identity to the newly assayed 482 cells from the *Tall* dataset. Alternative cluster assignment using nearest medioids gave a very similar result indicating the robustness of the data. Cluster assignment reveals a clear lack of the haematopoietic progenitor and blood population in cells from *Tall*^{-/-} embryos, table 3.11. Enumeration of the remaining populations except the extra-embryonic ectoderm shows similar numbers in *Tall*^{+/+} and *Tall*^{-/-} embryos.

Table 3.11 Random forest assigned clusters to *Tall* dataset.

Assigned cluster	Genotype	
	<i>Tall</i> ^{-/-}	<i>Tall</i> ^{+/+}
Blood	0	26
Blood progenitors	0	46
Endothelial	51	68
Allantois	12	16
Nascent mesoderm	27	47
Posterior mesoderm	15	16
Pharyngeal mesoderm	12	20
Extra-embryonic ectoderm	1	8
Epiblast	3	9

Projection on to the previously calculated PCA compared to de-novo PCA calculation indicates the benefit of using a previous data-set to understand a newly collected dataset. But the visualisation method used previously to explore the data was mainly tSNE, which provided better spacial segregation of the previously assigned clusters. Unfortunately no ‘projection’ method existed for tSNE when calculated in this manner. For our paper Scialdone et al. [2016] we recalculated the complete tSNE de novo and used colour to identify the new locations of the previously defined clusters.

Developing a new method it is now possible to position or ‘project’ new points on a previously calculated tSNE. The method is described later in Chapter 5. Harnessing this approach and calculating positions for the new cells we can appreciate the new *Tall* data set on the previously calculated tSNE fig. 3.53. Figure 3.53a shows the *Tall* data with the reference data set faded in the background but coloured by cluster. Figure 3.53b shows the same data

3.13 Single-cell resolved characterisation of a mutant

set but now the reference data set is greyed out and the *Tall* data set coloured by the newly assigned clusters calculated according to the random forest algorithm implemented by Dr Antonio Scialdone [Scialdone et al., 2016]. The projected data confirms the absence of blood progenitor and blood populations from the *Tall*^{-/-} mice consistent with previous publications and the lack of CD41 cells on index sorting fig. 3.47. Additionally the random forest assigned cluster identities appear to conform well to their spatial localisation on the tSNE fig. 3.53b.

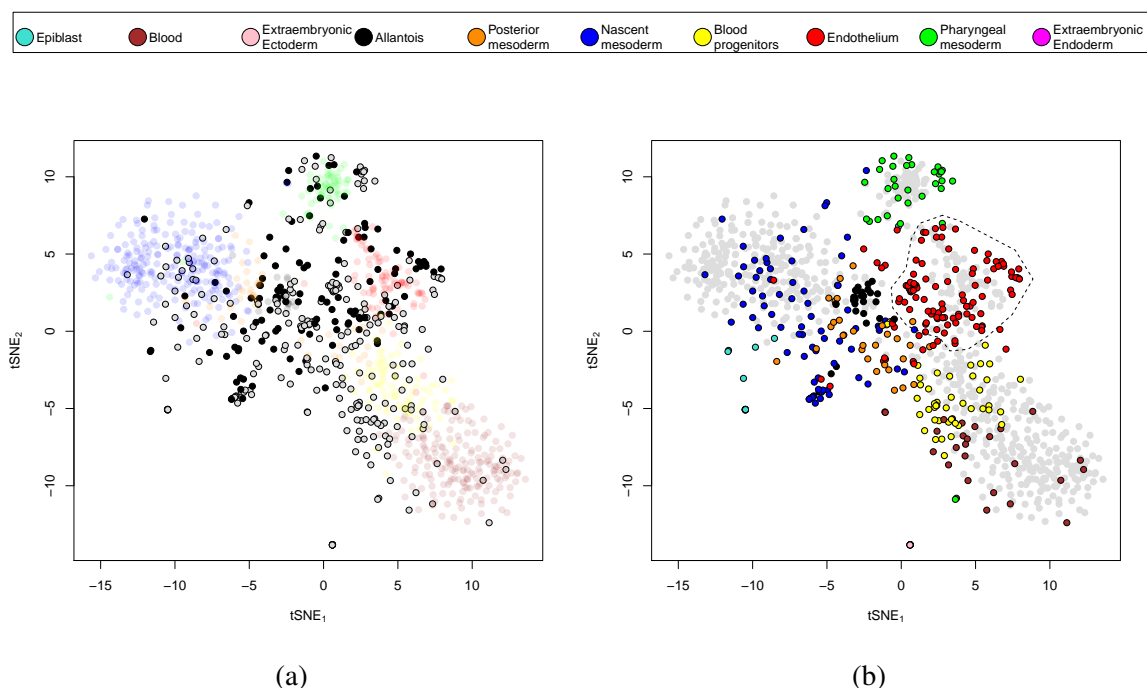


Fig. 3.53 (a) *Tall*^{+/+} light grey filled circles and *Tall*^{-/-} dark filled circles cells are projected on the previously calculated tSNE from the reference experiment fig. 3.36. The points representing the cells from the reference experiment are faded into the background to allow better visualisation of the *Tall* experimental data. (b) Cells generated from the *Tall* experiment coloured according to their assigned cluster as designated by the random forest algorithm calculated by Dr Antonio Scialdone [Scialdone et al., 2016]. This shows good correspondence between the random forest algorithm and the ‘projected’ tSNE. It also allows a more accurate cluster assignment by allowing selection of only those cells that are consistent on the tSNE for example in the case of the endothelial, only those red points that are within the dashed area in (b) are confirmed as endothelial.

To establish a more accurate and confident cluster assignment for the endothelial population the cells mapped to endothelium by the random forest algorithm was filtered to only include those that were consistent with the tSNE fig. 3.53b. Gene expression between *Tall*^{+/+} and *Tall*^{-/-} cells within this filtered set were compared to assess the affect of the *Tall* deletion in this population that is critical for definitive haematopoiesis. The filtering resulted in 43 *Tall*^{-/-}

Single-cell transcriptomic analysis of murine gastrulation

cells and 53 *tall*^{+/+} cells. These were compared using the DESeq algorithm revealing 205 genes that were significantly downregulated in the *Tall*^{-/-} compared to the *Tall*^{+/+} endothelial cell population. This included several known key regulators of early blood development including *Itga2b* [Dumon et al., 2012], *Lyl1* [Capron et al., 2006], *Fli1* [Zhao et al., 2018], *Ets2* [Stankiewicz and Crispino, 2009], *Sox7* [Lilly et al., 2017], *Gata2* [Kaimakis et al., 2016] and *Hoxb5* [Chen et al., 2016] consistent with the known function of *Tall* in specifying a haematopoietic fate in embryonic endothelial progenitor cells. *Hoxb5* notably has recently emerged as a powerful marker of definitive blood stem cells [Chen et al., 2016].

Strikingly in contrast to recent findings of cardiac fate specification after in vitro culture of yolk sac endothelial progenitors from *Tall*^{-/-} embryos [Org et al., 2015] we find no identifiable activation of any cardiac related pathways when looking at the differentially expressed genes. This led us to propose an alternate mechanism for fate specification in the early mesoderm whereby rather than acting as bipotential switch between haematopoietic and cardiac fates as suggested in Org et al. [2015], *Tall* is a master regulator of blood development as evidenced by its absolute requirement for in-vivo embryonic haematopoiesis but the loss of *Tall* is not sufficient by itself to switch cells towards an alternative cardiac fate [Scialdone et al., 2016].

3.14 Brachyury plugs the hiatus

BRACHYURY encoded by the *T* gene is the founder member of the T-box family of DNA binding transcription factors [Herrmann et al., 1990]. Heterozygous *T*^{+/-} mice have a short kinked tail while homozygous *T*^{-/-} mice are non-viable and die shortly after gastrulation around day 10 of gestation [Dobrovolskaïa-Zavadskaïa, N, 1927].

Studies in mutant embryos demonstrate that BRACHYURY plays a critical role in gastrulation and defects are described in several mesoderm tissues including extra-embryonic mesoderm where crucially there is failure of development of the allantois and subsequent chorioallantoic fusion [Gluecksohn-Schoenheimer, 1944]. Furthermore intra-embryonic mesodermal defects include an apparently absent notochord with abnormal somites and neural tube [Beddington et al., 1992].

Quantification of mesoderm:ectoderm ratios and chimeric experiments suggest that morphogenetic cell movements are abnormal supported by the observations of accumulation of *T*^{-/-} in the caudal region of the embryo and higher levels of chimerism in the ectoderm than the mesoderm [Beddington et al., 1992; Yanagisawa et al., 1981].

In humans mutations of the T-box family of genes have been associated with congenital diseases such as DiGeorge and Holt-Oram syndromes, characterised by defects in the mesodermally-derived cardiac tissues [Bruneau et al., 2001; Ryan and Chin, 2003].

In this chapter thus far, we have described utilising FK1⁺ to sort mesoderm specified cells with the advantage that it is expressed on the cell surface and FACS validated antibodies were readily available in the lab, so that this population of cells could be selected and sorted, for subsequent single-cell analysis. A more widely expressed protein within the mesodermal population is BRACHYURY but this is a transcription factor and so cannot be readily detected using standard FACS protocols. But given the large sampling gap between the epiblast population, a random sample of cells at E6.5, and the nascent mesodermal population collected by sorting on FLK1 and the high expression of *T* in both regions around the gap (see figs. 3.18 and 3.21a) we speculated that sorting *Brachyury*⁺ cells would help identify the missing population of cells bridging the gap and possibly provide insights into mechanism of early fate diversification.

3.14.1 Genetic *Brachyury* reporter

Taking advantage of a previously generated *Brachyury* knock-in reporter line that was available locally at the Gurdon Institute we set out to capture a wider range of mesodermal cells. The line described by Imuta et al. [2013] is summarised in fig. 3.54. This knocks in the *Brachyury*-2A-nucEGFP-2A-CreER^{T2} construct into the *Brachyury* locus in a TT2 ES cell line, which themselves are derived from a F1 embryo crossed between a C57BL/6 female and a CBA male [Yagi et al., 1993].

The construct was designed to express *Brachyury* from the knocked-in sequence but some heterozygous mice had variable degrees of tail defects including curled, short or no tail while the homozygous was non-viable [Imuta et al., 2013]. The *Brachyury*-2A-nucEGFP-2A-CreER^{T2} sequence includes the self cleaving 2A peptide between the protein coding sequences [Szymczak et al., 2004]. The nuclear localisation of the fluorescent protein is achieved by using histone fusion protein H2B-EGFP construct and there is the final protein coding sequence CreER^{T2} conditionally expressing *Cre* during induction with tamoxifen [Feil et al., 1997]. For this experiment we did not utilise the CreER^{T2} sequence for lineage-tracing but simply use this line to sort *Brachyury* expressing cells.

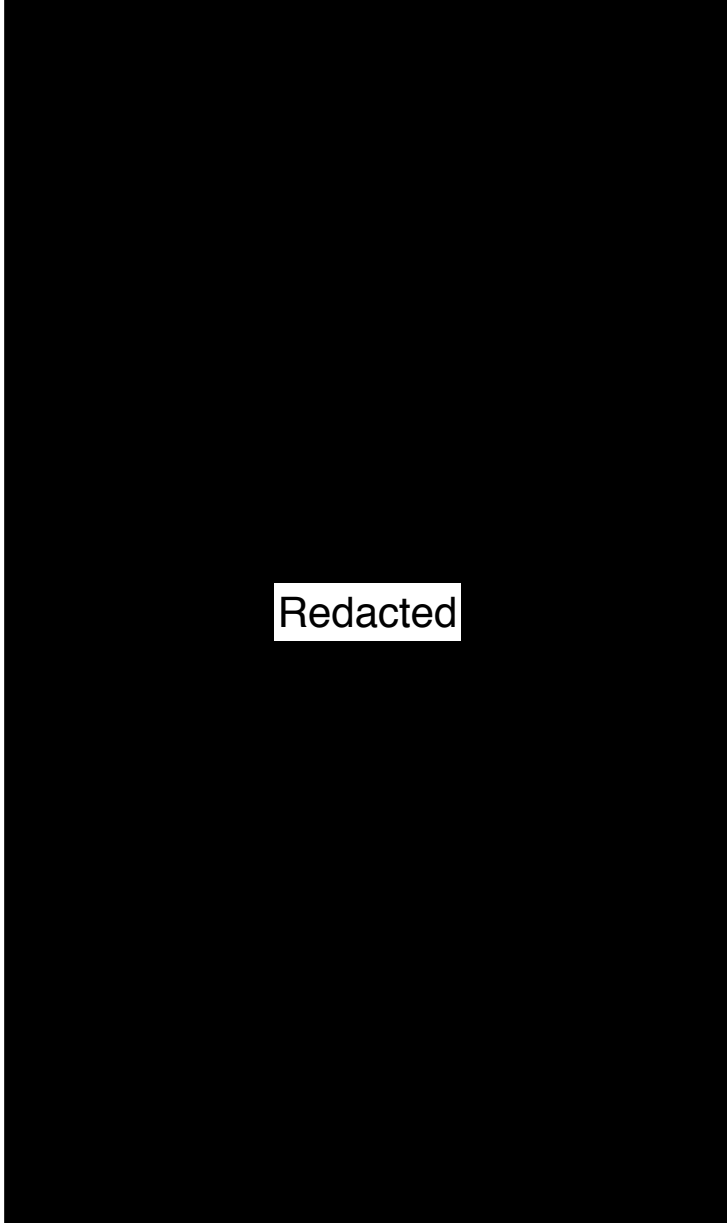


Fig. 3.54 Summary of the targeting strategy for knocking in the target vector containing the additional 2A-nucEGFP-2A-CreERT2 into embryonic stem cells using homologous recombination. The neomycin selection cassette ACN is self-excised in the germline [Bunting et al., 1999]. Confirmed ES cell clones were injected into 8-cell mouse embryos on the ICR background to form chimaeras. Chimaeric founders were crossed with C57BL/6J mice and were subsequently maintained on an ICR background. Figure taken from Imuta et al. [2013].

3.14.2 Mouse crosses, embryo collection and microscopy

Crosses were setup between $T^{\text{nEGFP-CreERT2/+}}$ on the ICR background and wild-type mice C57BL/6J and these would be expected to produce F1 embryos with 50% heterozygous $T^{\text{nEGFP-CreERT2/+}}$ and 50% wild-type mice $T^{+/+}$, fig. 3.55. This was performed at the Gurdon Institute with the help of Caroline Lee. 2 dams were sacrificed at day 7 post-plug by Caroline Lee and uteri harvested. The uteri were collected and embryos dissected out in the lab of Professor Alfonso Martinez Arias in the Department of Genetics. 12 embryos were collected and imaged using a confocal microscope by Dr Ben Steventon. 2 embryos were clearly pre-streak during dissection under the dissecting microscope and were therefore discarded and excluded from downstream processing.

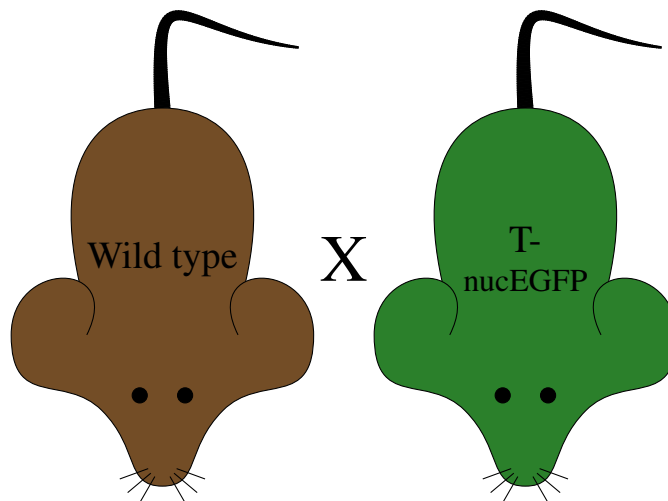


Fig. 3.55 Wild type to Brachyury-2A-nucEGFP-2A-CreERT2 [Imuta et al., 2013] cross is predicted to provide 50% GFP and 50% WT murine embryos.

The embryos, the litter from which they were derived and whether GFP was assessed to be present on confocal microscopy is reported in table 3.12. Microscopy of the embryos was used to in an attempt to identify heterozygous $T^{\text{nEGFP-CreERT2/+}}$ embryos so that fewer single cell suspensions would need to be generated to expedite single cell sorting into lysis buffer and reduce the time from harvesting to storage at -80°C . Unfortunately on initial assessment it appeared that all embryos except embryo 2, were positive for EGFP. This was highly improbable given the cross.

Subsequent FACS showed that only 3 of the 10 embryos were in fact the required heterozygous $T^{\text{nEGFP-CreERT2/+}}$ genotype. Later review of the stored images in fact confirmed concordance between the confocal images and the findings at FACS. The spurious expression seen in the other embryos is likely related to the recognised auto-fluorescence of extra-

Single-cell transcriptomic analysis of murine gastrulation

embryonic tissues and possibly rushed setting of laser intensity and gain settings [McGrath et al., 2003; Perea-Gomez et al., 2007]. Expression of EGFP was limited to embryos 6, 9, and 10 table 3.12.

Table 3.12 Overview of embryos harvested from $T^{nEGFP-CreERT2/+}$ and $T^{+/+}$ cross. Embryos were collected from 2 litters and litter mates are shown. The stage of the embryo as assessed during dissection under the bright field dissecting microscope - NP, neural plate; HF, head fold; MLS, mid to late streak. Presence or absence of EGFP on confocal microscopy and presence of EGFP⁺ cells during FACS.

Embryo	Litter	Stage	EGFP on confocal	EGFP on sort	Yolk sac
1	1	NP	No	No	Present
2	1	NP	No	No	Absent
3	1	HF	No	No	Present
4	1	MLS	No	No	Present
5	1	NP	No	No	Present
6	1	MLS	Yes	Yes	Present
7	1	EHF	No	No	Present
8	1	EHF	No	No	Absent
9	2	MLS	Yes	Yes	Present
10	2	MLS	Yes	Yes	Present
11	2	Pre-streak	————	Discarded	————
12	2	Pre-streak	————	Discarded	————

3.14.3 FACS

During single cell suspension generation and prior to FACS, multiple antibodies conjugated to selected fluorophores were incubated with the single cells to allow index sorting for FLK1, E-CAD, TIE2, EPCAM, CD31, PDGFR α and CKIT. The antibodies and associated fluorophores are summarised in table 3.13. The antibody/fluorophore sort panel was developed together with Dr Vasilis Ladopoulos.

Sorting was performed at the flow cytometry core facility in CIMR by Dr Chiara Cossetti on a 5 laser sorter. Embryo 2, the GFP -ve embryo on initial confocal observations was expected to be homozygous for the wild type alleles $T^{+/+}$ and was used as a negative control to set voltages including EGFP.

As there was a limited supply of cells, OneComp beads were used for the single stain controls to adjust compensations. The main population of cells was gated followed by singlet selection and exclusion of non-viable cells by DAPI fig. 3.57. A summary of the numbers

3.14 Brachyury plugs the hiatus



Fig. 3.56 Collage of selected confocal images from each embryo. Only embryos 6, 9 and 10 display EGFP signal within the epiblast. The remainder show signal due to auto-fluorescence of the extra-embryonic tissues as has been previously reported by McGrath et al. [2003]; Perea-Gomez et al. [2007].

Single-cell transcriptomic analysis of murine gastrulation

Table 3.13 Staining protocol including antibody, fluorochrome and the dilutions used. LP, long pass.

Antibody	Fluorophore	Laser wavelength nm & Filter	Dilution
CKIT	BV421	405 460/50	1:100
E-CAD	biotin		1:100
Streptavidin	PCPCy5.5	488 670/30	1:200
CD31/PECAM1	PE	561 585/29	1:200
PDGFR α	PECF594/Texas Red	561 610/20	1:400
FLK1	PECy7	561 750LP	1:200
TIE2	APC	640 670/30	1:200
EPCAM/CD326	APCCy7	640 750LP	1:200
BRACHYURY	nEGFP	488 530/40	
Viability	DAPI	355 460/50	1:5000

and percentages of cells in the gates for embryos 2 and 10 are summarised in tables 3.14 and 3.15.

There is a significant enrichment of cells in the GFP +ve selection gate P4 in embryo 10 as compared with embryo 2 19.72 % vs. 0.41 %. Additionally the GFP +ve cells additionally appear to be mostly PDGFR α +ve fig. 3.57.

The cells were sorted directly into freshly prepared chilled lysis buffer in 96 well plates, ensuring they were delivered into the lysis buffer in the centre of the well. Only the *Brachyury* expressing cells were sorted into 2 plates per embryo and stored for subsequent processing at -80°C .

3.14.4 Library preparation and sequencing

Libraries were generated as previously described using the Smart-Seq2 protocol fig. 3.4 and section 3.3 [Picelli et al., 2014] with the help of Dr Fernando Calero-Nieto and Sonia Nestorowa. After spiking in ERCCs and reverse transcription the cDNA product underwent PCR pre-amplification for 21 cycles. SPRI bead selection was used to remove primer dimers and TSO concatemers section 3.3.

DNA quantification and size distribution assessment was performed showing product in a number of wells an example of an assay is shown in fig. 3.58. Libraries were then generated by tagmentation, enrichment PCR and SPRI bead cleanup. The libraries were again checked for size distribution (fig. 3.58) then quantified and diluted ready for submission to the sequencing facility at CRUK.

3.14 Brachyury plugs the hiatus

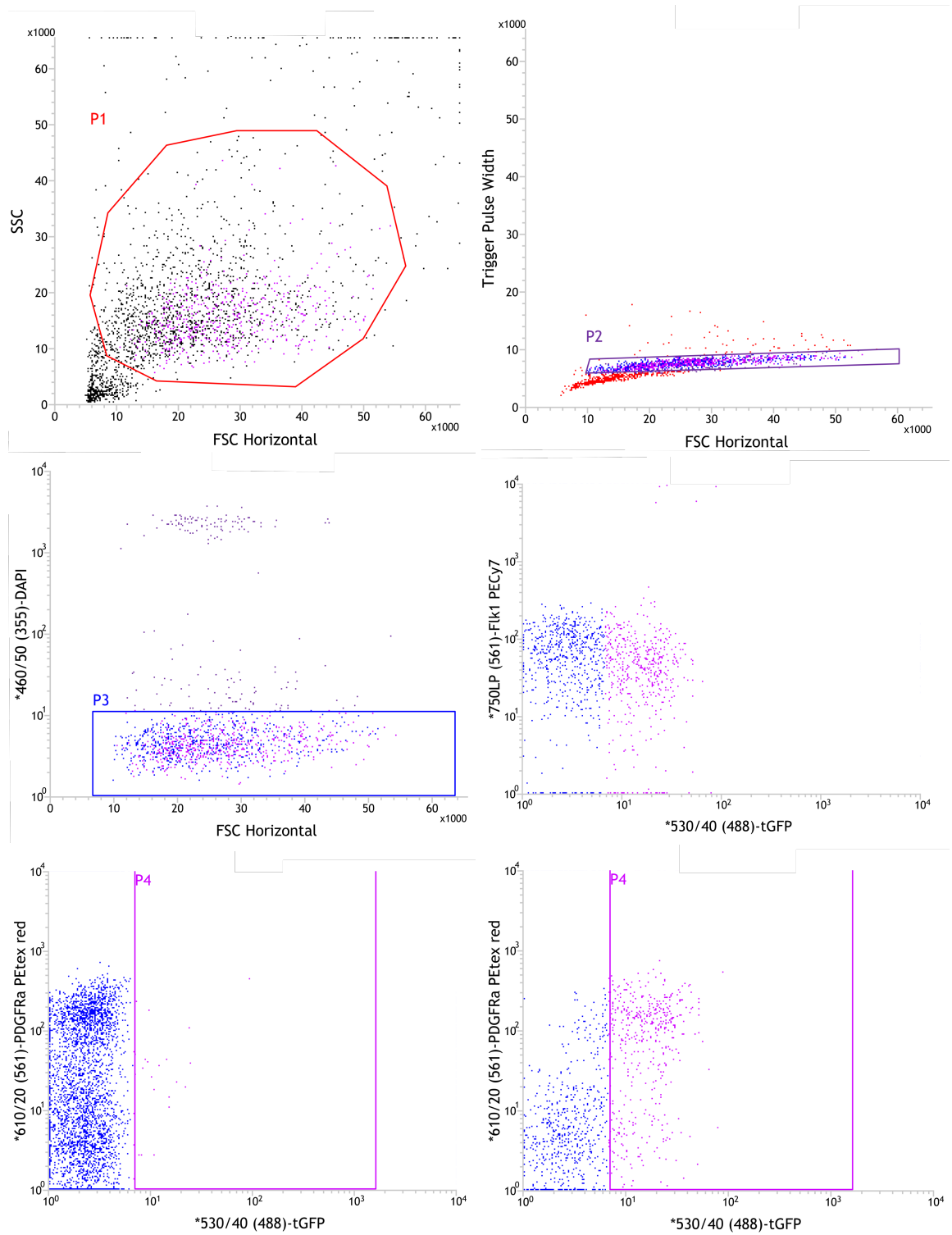


Fig. 3.57 Summary of sort layout (Continued on next page ...)

Single-cell transcriptomic analysis of murine gastrulation

Fig. 3.57 Summary of sort layout (contd.). Top left: Selection of main cellular population. Top right: Singlet enrichment. This is key in single cell sorting to ensure low doublet sorting. Middle left: Viability staining with DAPI. Cells with damaged membranes allow DAPI in to bind to DNA and become positive. Middle right: Abscissa represents GFP fluorescence and the ordinate represents FLK1 fluorescence, dots in blue are gated as GFP -ve while the purple dots represents cells gated as GFP +ve. Lower panels: In both abscessa represents GFP fluorescence and ordinate PDGFR α fluorescence. Lower left plot is of embryos 2 which is $T^{+/+}$ and so has no GFP +ve cells, while the lower right plot is of embryo 10 which is $T^{nEGFP-CreERT2/+}$ and the majority of the PDGFR α +ve cells are GFP +ve.

The cDNA product trace (fig. 3.58a) after cleaning still shows significant contamination from primers and possible TSO, suggesting there was little mRNA starting material. A possible reason for this was that at the time of collecting the cells we had in a bad batch of RNAase inhibitor which may have degraded the samples. This subsequently leads to the relatively small peak in the trace of the final library fig. 3.58b.

Libraries were submitted so that each 96 well plate was sequenced on a lane of a sequencing flow cell on the Illumina HiSeq2500 in high throughput mode.

Table 3.14 Embryo 2 FACS table indicating numbers of cells in each gate using the strategy shown in fig. 3.57 for a GFP -ve $T^{+/+}$ mouse embryo.

Populations	Events	% Total	% Parent	EGFP Mean	PDGFR α Mean
All Evenets	6369	100.00		6	67
P1	4876	76.56	76.56	3	65
P2	3912	61.42	80.23	3	62
P3	3467	54.44	88.62	2	65
P4	26	0.41	0.75	14	56

3.14.5 Alignment and QC

The raw reads from 2 plates for each of the 3 $T^{nEGFP-CreERT2/+}$ embryos were returned from the CRUK as zipped fastq files, see section 3.3, box 3.1, and table 3.3. Initial QC was performed using the fastQC tool [Andrews, 2010]. Due to a technical failure with the sequencer the initial run returned very few reads so the libraries were sequenced twice and the counts combined. The files were unzipped and aligned to a modified reference genome that included the EGFP sequence by Evangelina Diamanti using the splice aware aligner GSNAP [Wu and Nacu, 2010]. Gene counts were generated using HTSeq on the aligned

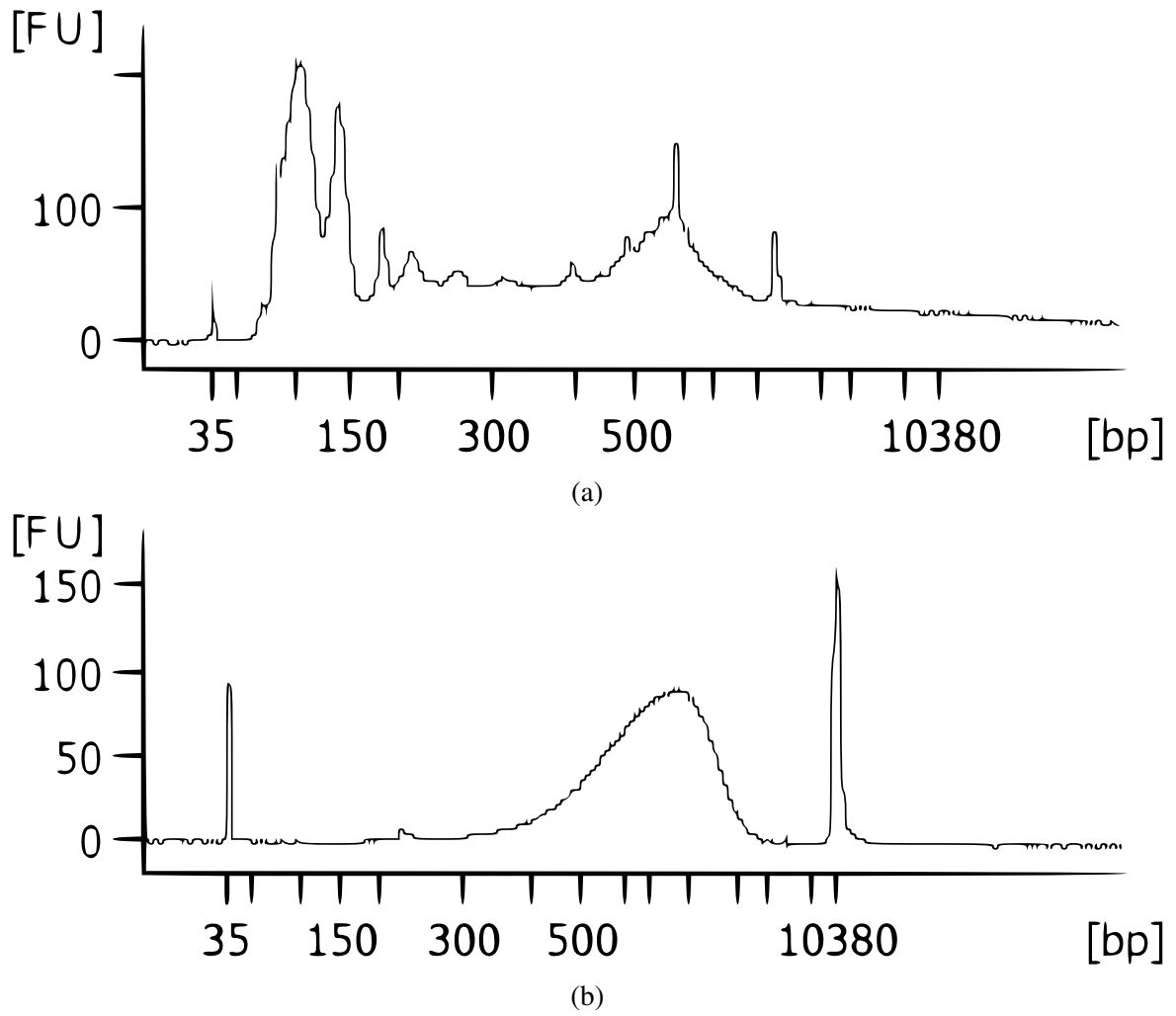


Fig. 3.58 Bionalyser trace from (a) cleaned cDNA product and (b) cleaned library.

Single-cell transcriptomic analysis of murine gastrulation

Table 3.15 Embryo 10. FACS table indicating numbers of cells in each gate using the strategy shown in fig. 3.57 for a GFP +ve $T^{nEGFP-CreERT2/+}$ mouse embryo. There is a much higher proportion of cells relative to both total and parent P3 population in P4 as compared to embryo 2, see table 3.14.

Populations	Events	% Total	% Parent	EGFP Mean	PDGFR α Mean
All Evenets	2469	100.00		11	64
P1	1745	70.68	70.68	9	64
P2	1265	51.24	72.49	10	61
P3	1087	44.03	85.93	10	64
P4	487	19.72	44.80	18	119

SAM files again using a modified Gene Transfer Format (GTF) file [Anders et al., 2015]. The combined counts were then used together with the reference experiment previously discussed. QC was now performed using the previously defined cut-offs. Due to the low mRNA content and high sequencing depth an additional QC filter was applied whereby all cells with greater than 80 % PCR duplicates were removed from further analysis, so that 334 $T^{nEGFP-CreERT2/+}$ cells remained for downstream analysis table 3.16. Gene expression was then normalised across both datasets together.

Table 3.16 Summary of the numbers of cells from the different embryos in the *Brachyury* experiment. The sample identifier corresponding to a lane on the sequencing flowcell, the embryo from which it was derived, the number of cells passing the initial QC and those that remained after removing cells with very high numbers of duplicates.

Sample identifier	Embryo	Passed Initial QC	Passed & has < 80% duplicates
11309	6	41	20
11310	6	76	67
11311	9	79	69
11313	9	83	70
11314	10	84	78
11471	10	57	30

Feature selection was performed on the remaining cells by looking for highly variable genes as described by Brennecke et al. [2013] resulting in 1534 highly variable genes.

3.14 Brachyury plugs the hiatus

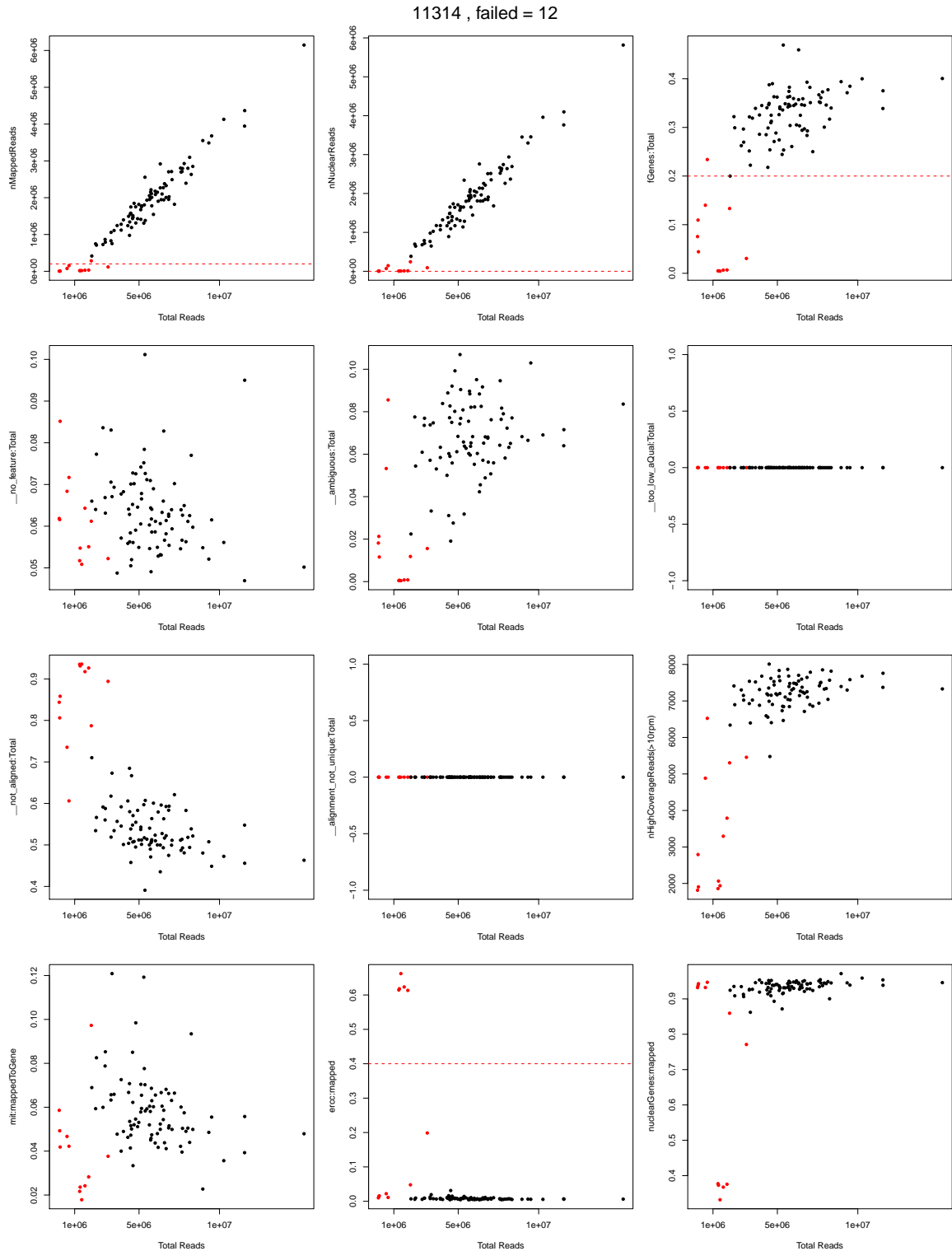


Fig. 3.59 Panel of QC filters as seen previously in figs. 3.15, 3.49 and 3.50. Noticeably a high number of reads are not aligned.

3.14.6 Data exploration

Performing de-novo PCA on the gene expression data having performed QC and selected the highly variable genes can provide an overview of how the new cells fit with the previous reference dataset fig. 3.60a. PC_1 does not provide any useful separation of the various cell populations and has not been displayed. Reassuringly the new dataset does not cluster completely separately from the reference dataset suggesting, that the QC and normalisation have been effective at reducing major batch effects. Importantly we must also keep in mind that the genetic background of these mice is different from the reference set and the $T^{nEGFP-CreERT2/+}$ phenotype is not completely benign and recognised to have a shorter tail.

Additionally the $T^{nEGFP-CreERT2/+}$ cells from mid to late streak stage appear to almost plug the gap between the unsorted E6.5 epiblast population and the later $Flik1^+$ sorted nascent mesodermal progenitors. Additionally though the PCA is different from that calculated solely on the reference data it appears to perform equally well in segregating the defined populations, cf. fig. 3.13. In conjunction with filling the void the $T^{nEGFP-CreERT2/+}$ embryos also generated BRACHYURY⁺ cells that are sparsely scattered through the endothelial and blood progenitor clusters.

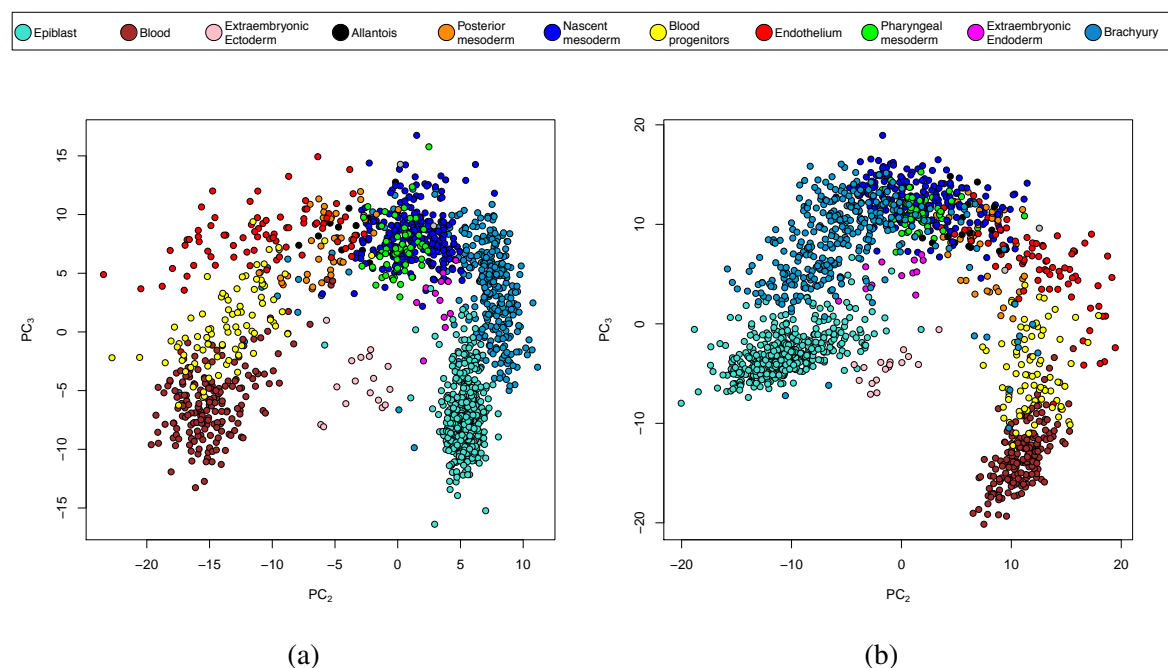


Fig. 3.60 (a) PCA recalculated on combined $T^{nEGFP-CreERT2/+}$ and reference data. Showing that the $T^{nEGFP-CreERT2/+}$ cells appear to bridge the previous hiatus of cells between the E6.5 epiblast and the nascent mesoderm. (b) $T^{nEGFP-CreERT2/+}$ cells projected onto reference dataset cf. fig. 3.13.

3.14 Brachyury plugs the hiatus

An advantage of PCA is that the new data set can be readily projected onto the reference using the gene weights that were calculated, in much the same way as was done for the *Tall*^{-/-} data fig. 3.52b. The *T*^{nEGFP-CreERT2/+} cells project directly between the E6.5 epiblast and the nascent mesoderm even more convincingly, completing what appears to be an arc of related cell types on the PCA shown in fig. 3.60b.

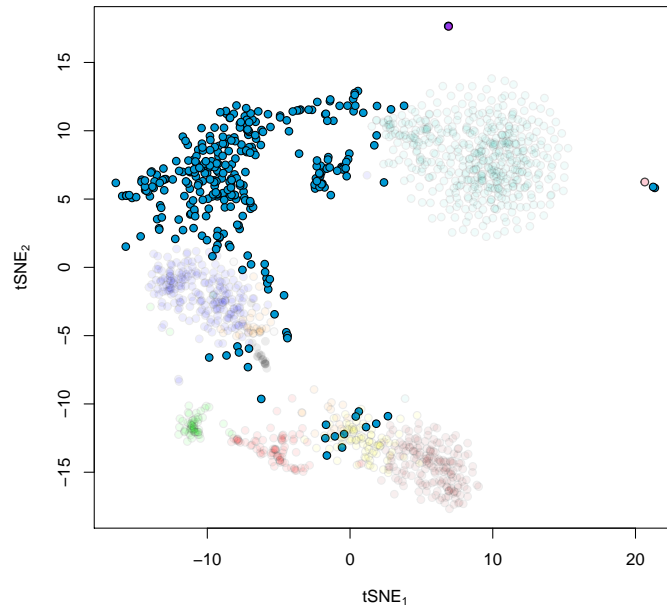
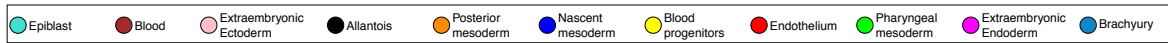
In this instance there was a subtle difficulty in that the *T*^{nEGFP-CreERT2/+} data set cell counts were aggregated using a more recent GTF file. This meant that 2 genes that had previously been annotated and used for the projection in the reference experiment were now not available. These 2 genes carried very little weight amongst the principal components of interest and so had little effect on the final positions of the points.

On the PCA it is difficult to identify the previously characterised polarisation of the epiblast cluster and the anterior and posterior domains within the nascent mesoderm. These were much better presented on the tSNE, furthermore during the comprehensive exploration and analysis of the data it was the tSNE to which we often referred and used as a reference map. It follows therefore that a more intuitive representation may be found if the new *Tall*^{nEGFP-CreERT2/+} cells could be mapped onto the tSNE calculated on the reference data fig. 3.61.

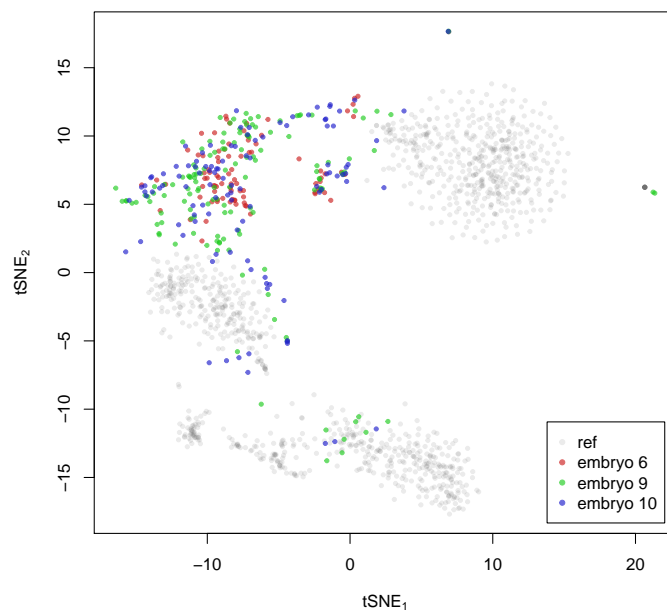
The new *T*^{nEGFP-CreERT2/+} cells map into the hiatus between the E6.5 unsorted epiblast and the nascent mesoderm as they did on the PCA. But additionally our familiarity with the substructure of the tSNE suggests that the cells appear to arise from the polarised tip of the E6.5 epiblast cluster as previously characterised in section 3.8. This is where the E6.5 epiblast cells appeared to be undergoing an epithelial to mesenchyme transition (EMT) as demonstrated by their gene expression profiles, consistent with cells delaminating and transiting through the primitive streak. From here the cells appear to stream towards the nascent mesoderm fanning out in an arc generating both the anterior and posterior domains fig. 3.61.

To assess the fidelity of the ‘projection’ on to the tSNE we looked at expression profiles of known genes just as we did for defining cluster identity in section 3.7.2. The key genes clearly in this case would be those related to pluripotency and gastrulation in particular EMT fig. 3.62. Using the genes previously identified as differentiating the anterior and posterior domains it is clear that few if any posterior domain cells have been captured by the *T*^{nEGFP-CreERT2/+} data. This though is not surprising on review of the reference dataset. Using the coarse time scale data based on embryos staged during microscopy, fig. 3.37 shows that while the anterior domain of the nascent mesoderm is replete with cells in the PS stage, the posterior domain appears to have a peak density around the neural plate stage and is relatively sparse at PS. Considering the *T*^{nEGFP-CreERT2/+} embryos were harvested from an

Single-cell transcriptomic analysis of murine gastrulation



(a)



(b)

Fig. 3.61 (Continued on next page)

3.14 Brachyury plugs the hiatus

Fig. 3.61 (Continued from previous page) (a) Mapping the new $T^{\text{nEGFP-CreERT2/+}}$ cells onto the reference tSNE, see fig. 3.18b. The cells map well into the space between the polarised tip of the E6.5 epiblast cluster and fan out towards the nascent mesoderm giving the impression of a branching ontogenic time-course. An additional population lies below the tip of the main cluster. Other cells are scattered amongst the extra-embryonic ectoderm and endoderm clusters, endothelial and blood progenitor populations. (b) Confirms that the cells from the different embryos do not cluster together but intermingle. Embryo 6 appears to have consisted of cells that were developmentally earlier than embryos 9 and 10, having little representation in the posterior nascent mesoderm, endothelial or blood progenitor clusters.

earlier developmental stage based on microscopy appearance it is consistent that they have generated relatively few cells mapping to the posterior nascent mesoderm. Additionally T expression in the posterior domain appears to be rapidly downregulated and is significantly lower than that of cells in the anterior domain fig. 3.63a suggesting the possibility that sorting on T at this stage of development may miss a significant proportion of this population.

The pattern generated on the tSNE displays an intriguing hour-glass figure, the E6.5 epiblast and the $T^{\text{nEGFP-CreERT2/+}}$ cells forming the two bulbs with a connecting stalk of cells. The epiblast cells therefore appear not only to converge on the primitive streak in spatio-temporal dimensions but they also seem to converge in terms of their gene expression profiles forming a narrow path between the source and destination populations. The low density of cells sampled that map to this region suggest that the ingressing cells either progress rapidly through this convergent state or that the cells become highly proliferative immediately afterwards.

Looking at cell cycle genes which are highly related to proliferation there is no clear signature of differences in proliferative state between the $T^{\text{nEGFP-CreERT2/+}}$ cells compared to the cells in the polarised region of the E6.5 epiblast or the convergent belt region fig. 3.63b. Cells associated with the region of interest are homogeneously in the G2/M or S phases of the cell cycle suggesting they are all proliferating. These observations lead us to speculate that cells transit rapidly between these two populations following a narrow trajectory in gene expression space with little or no ascertainable fate bias until entering the main body of the $T^{\text{nEGFP-CreERT2/+}}$ cells. This is consistent with experimental findings from epiblast cell transplant experiments suggesting high degrees of fate plasticity [Le Douarin and Teillet, 1973].

An alternative possibility is that our sampling strategy of selecting only $T^{\text{nEGFP-CreERT2/+}}$ or the limited range of embryo stages has led to the sparsity in this region. With regards to the former, it is unlikely that cells along this developmental journey would down regulate and then immediately up-regulate T sufficiently within such a narrow time-window so as

Single-cell transcriptomic analysis of murine gastrulation

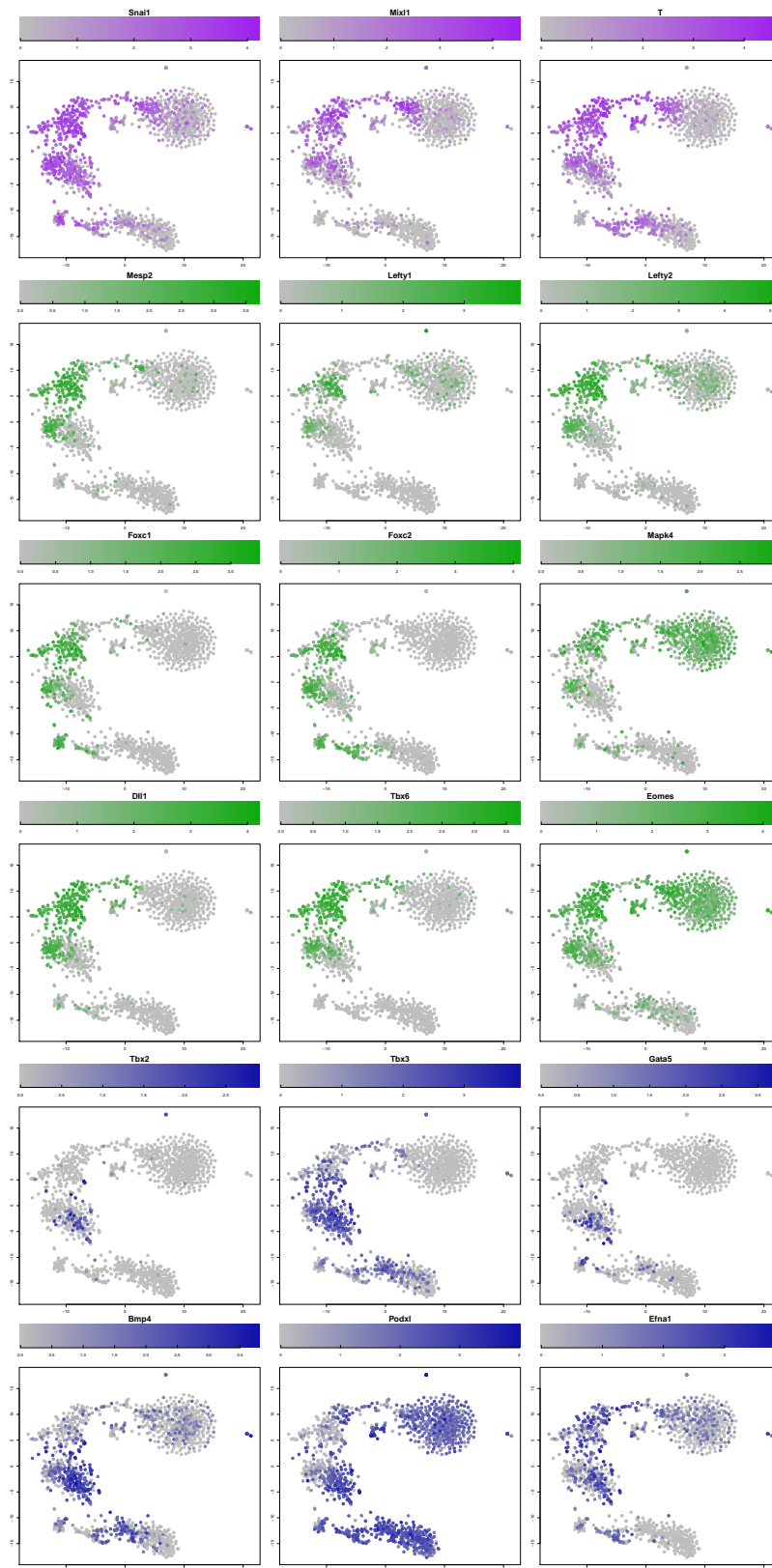


Fig. 3.62 Gene expression profiles in $T^{nEGFP-CreERT2/+}$ and reference cells on the tSNE (Continued on next page ...)

3.14 Brachyury plugs the hiatus

Fig. 3.62 Gene expression profiles in $T^{\text{nEGFP-CreERT2/+}}$ and reference cells on the tSNE (Contd ...). EMT genes *Mixl1* and *Snail* are expressed in the polarised tip of the E6.5 epiblast cluster and the $T^{\text{nEGFP-CreERT2/+}}$ cells (Upper panels in purple). Genes that were previously found to be expressed highly in the anterior domain of the nascent mesoderm (Middle panels in green) and those expressed highly in the posterior domain (Lower panels in blue).

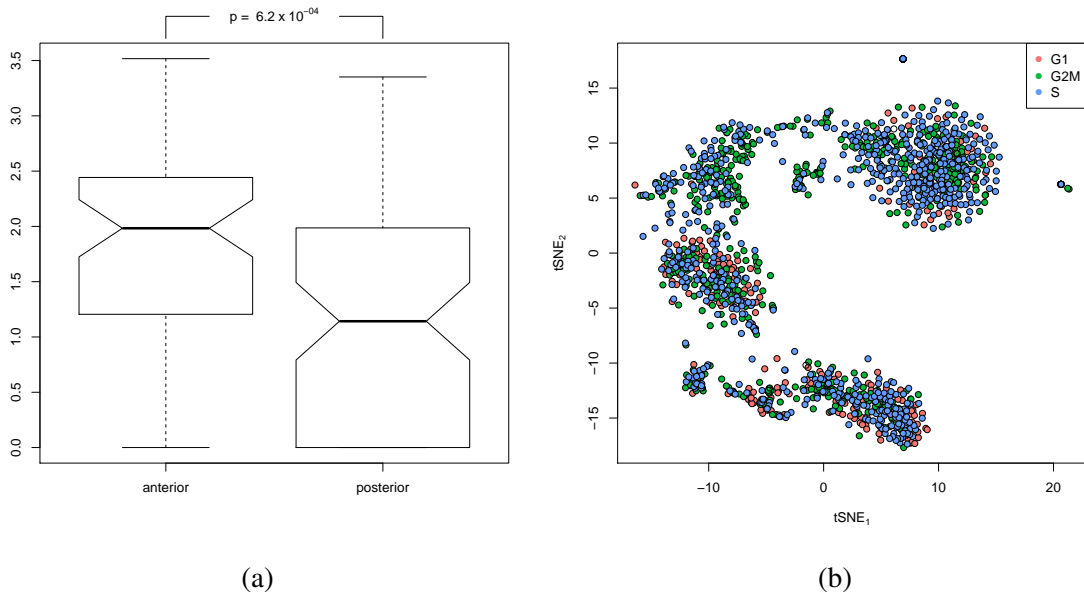


Fig. 3.63 (a) *T* expression is significantly lower in the posterior domain of the nascent mesoderm compared with the anterior domain. (b) Cell cycle stage was assigned to cells within the whole dataset. Between the tapered end of the E6.5 epiblast culture and the $T^{\text{nEGFP-CreERT2/+}}$ cells there is no appreciable difference in proportions in the different cycles of cell stage but nearly all cells are in either G2/M or S phase indicating active proliferation.

to lead to such a sampling bias. Since all the analysed embryos were at mid to late streak stage, the possibility that collecting earlier embryos would fill out this sparsity in the putative convergent region cannot be excluded. Focused dissection of the streak without selection may help to bridge this gap.

The outlier cluster lying between the main $T^{\text{nEGFP-CreERT2/+}}$ body and E6.5 epiblast clusters was of particular interest and has a gene signature at this developmental stage consistent with cells derived from the node or organiser and surrounding structures. The genes that are differentially expressed in this cluster relative to the remaining $T^{\text{nEGFP-CreERT2/+}}$ cells were identified.

Single-cell transcriptomic analysis of murine gastrulation

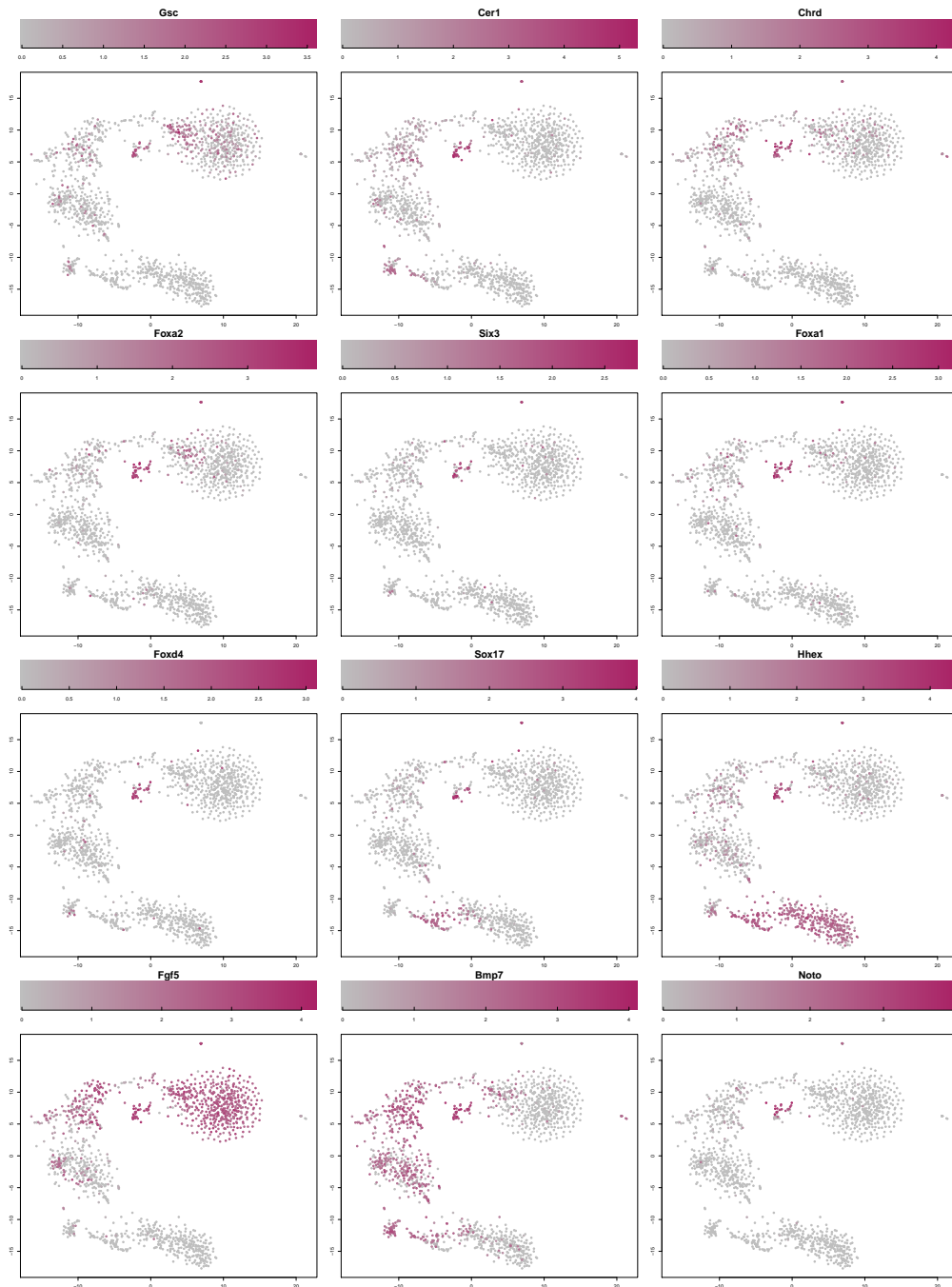


Fig. 3.64 tSNE plots coloured by gene expression displaying genes highly expressed in the putative node/organiser population.

3.15 Conclusions

This chapter has focused chiefly on developing a fine grained map of FLK1⁺ cells during gastrulation with a particular focus on haematopoiesis. Our lab has previously looked at a similar population of cells but using qPCR where genes were selected by hand [Moignard et al., 2015]. In this chapter with the use of Smart-Seq2 technology no apriori gene selection was required [Picelli et al., 2014]. Being a plate based technique and in combination with the use of a sorter allowed us to collect surface antigen data using index sorting and marry the data together with gene counts.

The advantage of single-cell techniques as is well illustrated in this chapter is that by assaying gene expression at a single cell level a pure population does not need to be isolated as is required for bulk sequencing. Furthermore cell identity can be assigned post-hoc and subsequent comparisons or alternative analysis performed using in-silico defined populations.

In-vivo experiments capture the true underlying biological processes that govern normal development while in-vitro platforms may phenocopy a certain biological phenomenon but do so for a limited time-course and lack fidelity. The limit with in-vivo strategies is that our knowledge of cell types in embryos even in early development may be incomplete and certainly we do not have sufficiently accurate markers to collect pure populations. This combined with the extremely low numbers precluded use of bulk RNAseq methods.

3.15.1 Technical challenges

Single-cell methods do pose certain challenges mostly concerned with the very low input mRNA. An individual cell may only contain 1 pM to 26 pM of RNA, most of which is ribosomal RNA, and during harvesting and dissection gene expression will begin to change [Ståhlberg and Bengtsson, 2010]. This is partly counteracted by performing all manipulations at 4 °C except tissue dissociation into single cells where the duration of incubation is minimised. In Smart-Seq2, mRNA enrichment is performed using poly-A enrichment so biasing against non-polyadenylated RNAs such as long non-coding RNAs (lncRNAs). Despite these limitations the numbers of genes that are assayed in multiplex is enormous with no user dependent apriori gene selection.

The rate of capture of these mRNAs should at this stage be solely dependent on presence or absence of a poly-A tail and not on the gene lengths. The amount of primers and TSO added at this stage must be optimised and can be problematic in where cellular RNA content is extremely low possibly due to poor tissue handling or because the cells themselves are in a relatively quiescent phase for example stem cells. Smart-Seq2 then uses transcript switching

Single-cell transcriptomic analysis of murine gastrulation

to generate full length cDNA product which after some cycles of pre-amplification provides the substrate for the Tn5 transposase used to generate libraries. At this stage the length of the transcripts biases towards higher rates of capturing longer transcripts. Additionally in the case of long transcripts there is a higher probability for the same molecule to be sampled twice at different locations, generating two reads from the same molecule - under such a circumstance there is no way of differentiating between the subsequent reads being generated from the same or different mRNA molecules and may be considered to be a non-identifiable duplicate. At the subsequent step these molecules are amplified and then sequenced. At the point of sequencing given that gene counts tend to be low in single-cell experiments two reads starting at same genomic locus are likely to represent PCR duplicates and these are to some extent identifiable and may be removed. Depending on the quality of the libraries generated, the number of cycles for the second amplification stage is adjusted and in poor quality samples can lead to excessive numbers of duplicate reads.

The indexing of the Tn5 transposase to generate the transposome allows unique barcodes to be added to the DNA fragments along with the obligatory sequencing adaptors. The use of two transposomes allows indexing at both ends of the captured molecule at the same time as adding the two different obligatory sequencing adaptors, the P5 and P7. In our experiment by combining 8 indexed P5 adaptors and 12 indexed P7 adaptors we were able to multiplex 96-cells for the final PCR amplification and downstream sequencing reducing batch effects at these stages and minimising sequencing costs. Multiplexed sequencing for single-cell is very useful as deep sequencing gives diminishing returns in terms of finding novel transcripts and often simply captures many more PCR duplicates.

A critical step within this tagmentation process is to ensure the correct ratio of transposase to cDNA. Too much cDNA or insufficient transposase and the library fragment length will be too long and too little cDNA or excessive transposase and the expected fragment length will be too small. Library fragment length being important for efficient binding to the sequencing flowcell and subsequent bridge amplification.

As alluded to earlier the amount of primer and TSO added at the initial reaction requires careful optimisation as this can lead to high concentrations of primer dimers and TSO concatemers which will sequester a large number of sequencing reads. In downstream analysis during QC this is often seen as a high number of unmapped reads figs. 3.7, 3.49, 3.50 and 3.59. At this stage re-analysing the raw fastq files will show an abundance of reads from primers dimers and TSO concatemers.

Normalisation of counts remains challenging especially when combining separate experiments where batch effects predominate. Throughout this chapter for normalisation we have

used the Bioconductor package ‘scran’ [Lun et al., 2016b]. This is an adaptation on earlier methods that use size factors to normalise the data [Love et al., 2014]. The limitation that ‘scran’ attempts to address is a problem these methods run into for single cell experiments where counts matrices are sparse. This is overcome by pooling what appear to be similar cells, see section 3.5. The problem here for ‘scran’ becomes of a cyclical nature in that cells are first clustered to normalise the data and then subsequent feature selection and formal clustering is proposed. Empirically it has worked well in the current data set but can clearly be problematic where batch effects predominate as they can overwhelm any clustering. Additionally some groups may well be too small to allow for adequate pooling and this can lead to them being excluded or the algorithm failing. Additionally since this is a fast moving field the package remains in development and version updates can lead to significant differences in outcome, which become particularly noticeable due to exclusion of cells.

3.15.2 Cell type assignment

A major challenge after generating the data, in producing the reference map was to assign cell identities. This was performed through an iterative process using prior knowledge of the cells likely to be present at each stage and known markers. For example the blood cells could be readily identified by their high expression of *Hbb-bhl*. Surface marker information from the index sort particularly CD41 confirmed and provided confidence in this assignment. The endothelial and blood progenitor populations were readily identified by expression of markers such as *Tie1*, *Tek* and *VE-Cad (Cdh5)*. The E6.5 cells populations, the epiblast (*Nanog*, *Pou5f1/Oct4*), extra-embryonic ectoderm (*Elf5*) and endoderm (*Ttr*, *Afp*) were also easily identified. Others for example the second heart field, nascent mesoderm, allantois and especially the posterior mesodermal clusters were more difficult to assign. The problem is that markers that were once thought to be highly specific turn out to be expressed in other tissues separated by embryo location and/or developmental time.

To aid with cell type assignment a web application was deployed using Rook and rApache (available at <https://github.com/jeffreyhorner/Rook.git> and <http://rapache.net/>). This displays expression levels of user defined genes on the tSNE and is available on the Cambridge Stem Cell Institute website <http://gastrulation.stemcells.cam.ac.uk/allData.html>.

Having assigned cell type we now have a set of single cell profiles that are with a particular cell state. This was then used to identify cell types in the *Tall^{-/-}* data set. This could be extended to provide an easy to access publicly available repository of previously sequenced cell states that could help inform cell type assignment in future experiments.

3.15.3 Mutants and reporters

This experiment was one of the first to evaluate a mutant mouse model at single cell resolution [Scialdone et al., 2016]. The reference cells were used to infer cell type in the *Tall*^{-/-} cells using a random forest algorithm implemented by Dr Antonio Scialdone. This produced almost identical results to an alternative algorithm, the nearest medioid algorithm supporting the cell type assignment. The mapping of cells on tSNE was also consistent with the assigned clusters providing added reassurance. When performing differential expression only those cells assigned endothelial identity and also mapping to the correct region of the tSNE were included. Due to batch effects between experiments only wild type *Tall*^{+/+} cells from the *Tall* experiment were used for comparison against *Tall*^{+/-} cells.

The *Brachyury* reporter mouse despite the limitations of having an abnormal phenotype still does produce viable heterozygous offspring and though development may not be completely normal the embryo still undergoes very similar morphological changes and most if not all cell types are produced in the desired quantities and anatomical locations. At our current level of mechanistic understanding any mechanisms we do uncover that govern cell differentiation or gross morphology will be maintained in even such a model organism and any models we do generate should be resilient to such differences - the ideal model would be able to recapitulate the complete development of the mouse with the abnormal tail in the heterozygous.

Specifically interrogating the early time-point by using this reporter we have been able to sample a population that was missed with the FLK1 sorting strategy. Despite the poor quality of the sequencing libraries from this sample it neatly slotted into the space between the unsorted E6.5 epiblast and the nascent mesoderm when mapping onto a tSNE calculated on the high quality reference data set. This gives some insight into how high quality data can be used to make sense of other data sets by using algorithms that can take advantage of prior information.

3.15.4 Novel pathway discovery and validation

Enriched expression of the genes *Alox5*, *Alox5ap* and *Ltc4s*, integral components of the lipoxin/leukotriene branch of the arachidonic acid pathway was discovered within a subset of endothelial cells. This enrichment though not apparent with unsupervised methods was hinted at through the structure of the endothelial cluster on tSNE. The pathway was discovered by hand selecting genes known to be associated with the first yolk-sac derived wave of definitive haematopoiesis that generates erythroid myeloid precursors (EMPs) and finding highly correlated genes. Since only a handful of genes were present within the clusters these

genes stood out on inspection fig. 3.40. Subsequent GO analysis for biological process terms showed strong enrichment for the lipoxin/leukotriene pathway table 3.9.

To validate the pathway we used the 5-lipoxygenase (ALOX5) inhibitor Zileuton which is licensed for prophylactic use in asthma. This showed clear dose-dependent inhibition of colony formation in an in-vitro culture system while addition of leukotriene C₄ led to the induction of increased colony formation fig. 3.45. In combination this provided strong evidence in support of the pathway in developmental haematopoiesis.

Interestingly Zileuton has been studied in the context of sickle-cell anaemia in a bid to achieve an increase in the fraction of foetal haemoglobin, HbF [Haynes et al., 2004]. Zileuton is a derivative of hydroxyurea and its role in sickle cell anaemia has been predicated on its structural resemblance to hydroxyurea [Haynes et al., 2004; Rohrman and Mazziotti, 2005], which has been used in sickle cell disease. Alternatives to hydroxyurea have been sought due to its bone marrow suppressive effects and Zileuton has been suggested as an alternative worthy of further study [Haynes et al., 2004; Rohrman and Mazziotti, 2005].

The leukotriene branch is not unique amongst the arachidonic acid pathways implicated as having a role in developmental haematopoiesis. In zebrafish, the epoxyeicosatrienoic and the prostaglandin branches have been proposed as important pathways in embryonic blood development [Li et al., 2015; North et al., 2007]. Using leukotriene C₄ promoter driven expression of EGFP, Strid et al. [2008] showed specific promoter activity in monocytic and myeloid leukaemia lines and confirmed expression of *Ltc₄s* transcript.

Finkensieper et al. [2010] demonstrate that both *Alox5* and *Alox5ap* are expressed at Day 6 of EB formation from ES cells and by day 10 ALOX5⁺/CD68⁺ (CD68 was used as a specific marker of macrophages) co-expressing cells and ALOX5⁺/CD45⁺ co-expressing cells were seen prominently on immunohistochemistry. Using high performance liquid chromatography they demonstrate synthesis of related leukotrienes LTB₄ and LTD₄ in EBs between days 6 to 10. They also show that leukotrienes play a role in vasculogenesis and that pharmacological inhibition of leukotriene receptors and small hairpin RNA silencing exert anti-angiogenic effects in differentiating ESCs.

Wang et al. [2015] found up-regulation of *Alox5* by real-time PCR in sorted human CML progenitors but not in patients in remission after treatment with tyrosine-kinase inhibitors. In an in-vitro model of leukaemia stem cells (LSCs) in Chronic myeloid leukaemia (CML) they cultured LSCs identified on flow cytometry as Lin⁻cKit⁺Sca1⁺ from CML patients with the tyrosine kinase resistance mutation T315I in *Bcr-Abl*. They show apoptosis after Zileuton treatment but not after Imatinib (a tyrosine kinase inhibitor). In a xenograft model after

Single-cell transcriptomic analysis of murine gastrulation

transferring human mutant *Bcr-Abl-T315I* CML cells into NOD/SCID mice they found that intra-peritoneal injection of Zileuton inhibited LSCs compared with normal saline controls and imatinib treated mice.

We have shown for the first time in-vivo that all three *Alox5*, *Alox5ap* and *Ltc4s* are co-expressed within a subset of FLK1⁺ endothelial cells in the gastrulating murine embryo. This now provides a basis for further mechanistic studies into the observations seen in in-vitro systems described above. Additionally we highlight an alternative mechanism by which not only Zileuton but also the widely used drug hydroxyurea may increase HbF fraction in sickle-cell patients.

Inspecting the heatmap fig. 3.40 the close relationship of other genes to leukotriene associated genes suggests they may have an important role to play in haematopoietic stem cell emergence. *Jag2* is a *Notch* related gene that is highly correlated with *Alox5*. Other genes such as *Eltf1* a pro-angiogenic adhesion G-coupled protein receptor and *Ldb2* the Lim-domain binding protein 2 suggested as a key driver of transendothelial migration of leucocytes are good candidates for further study [Masiero et al., 2013; Mylona et al., 2013; Shang et al., 2014]. The related gene *Ldb1* a transcriptional cofactor has already been proposed as novel player in early murine haematopoiesis. These very preliminary observations highlight how there continue to be many opportunities to gain further insights into processes governing early development within this dataset.

3.15.5 Pseudotime and pseudospace

It is appealing to consider that sampling embryos and generating single cell suspensions at different developmental stages can allow the reconstruction at molecular levels of processes governing early embryonic development, essentially pasting together cross-sectional snapshots into a continuity - a pseudotime. This idea is predicated on the logical assumption that cell differentiation is a smooth process taking place in gene expression space and that cells that are most similar to each other by gene expression will be most likely to be ontogenically related. Then by simply defining a starting point we can string together cells that are most closely related up to an end point.

This idea is not new to single-cell assays in fact it forms the basis of many scientific observations. For example gross embryology in eutherian placentals is based on specimens collected, fixed, stained and observed under the microscope. Embryo development has then been described according to what is seen on these cross-sectional snapshots on occasion consolidated together with longitudinal observations from model organisms. A problem

specific to our current method of generating single cells is that we lose almost all spatial context unless focused dissection is used.

It is not only the loss of spatial relationships that hampers computational reconstruction of a pseudotime. Evolution has led to the re-purposing of many biochemical and molecular pathways for alternate purposes activated in different tissues. This leads to disparate cell types sampled concurrently sharing a common gene expression set making assessment of similarity difficult. For example proliferating cells in the same phase of the cell cycle may have many genes highly expressed that are common. This signal may be sufficient to swamp any cell type signal so causing cells to be close to each other by cell cycle (or any other biological process with a common gene set) belying their developmental origins.

Additionally our assays have limited fidelity. Single cell RNAseq is prone to drop-out where genes that may be expressed at low-levels are highly likely to have zero counts, though randomly some counts will be captured in a few cells. Due to the large number of such genes a different subset may be randomly sampled from cells that are very similar again making assessing similarity difficult.

Despite these caveats thoughtful interpretation of visible spatial relationships in dimensionality reduction plots can help identify populations that are linked ontogenically or spatially to identify pseudo-temporal or pseudo-spacial relationships. Examples are given in this chapter where a clear pseudotime relationship exists between the blood progenitor and blood clusters. Another pseudo-temporal relationship is revealed by the *Brachyury* data between the polarised tip of the E6.5 epiblast cluster and the nascent mesoderm. The latter was not formally investigated partly because of batch effects between the datasets.

Clearly not all neighbouring populations are temporally related. Morphogen gradients can lead to spatial relationships between juxtaposed populations and this is seen within the nascent mesoderm cluster which can itself be sub-clustered into anterior and posterior domains. Making sense of such relationships, deciding on subsequent algorithms to be used and interpretation continue to heavily rely on human input.

3.15.6 Summary

Harnessing the advantage of single-cell sequencing we have generated a high quality map of *Flik1* cells during gastrulation. On this map we reveal both pseudo-temporal and pseudo-spacial relationships. Further we use the pseudo-temporal relation to identify patterns of gene expression. Having established cluster we further refine the clusters using previous knowledge of developmental biology and key genes to define a cardiac sub-cluster within the

Single-cell transcriptomic analysis of murine gastrulation

pharyngeal mesoderm and intriguingly activation of the leukotriene pathway during early yolk-sac definitive wave haematopoiesis.

Characterisation of the mutant *Tal1*^{-/-} using the previous map as a reference we confirm a failure of haematopoiesis with production of no blood progenitor or blood cells. Focusing on the putative endothelial cluster we also again re-affirm failure of activation of the blood programme but also show no evidence to support the suggestion that these cells having failed to progress towards a blood fate now adopt an alternative cardiac fate.

Finally we show that the reference data set combined with sophisticated computational techniques that can incorporate prior information can overcome experimental batch effects allowing the *Brachyury* data to be assimilated and revealing a nodal population on tSNE not apparent on PCA.

Chapter 4

Single-cell census of mouse organogenesis

4.1 Background

Early embryological development is a highly dynamic process, during which an organism undergoes major changes in macroscopic appearance underpinned by morphological cellular movements and molecular changes in epigenetic landscape, gene expression and proteome. Traditionally dynamics have been investigated by collecting cross-sectional samples at time intervals and inferring dynamics from serial static images and molecular signatures are garnered from bulk cellular samples either from heterogenous cell populations or from ESC differentiation models.

The previous chapter focused specifically on gastrulation the period during which coordinated cellular movements transform the bilaminar disc into a trilaminar structure. Post gastrulation there is a significant expansion in cell number and the redistributed cells now begin to generate the primordial organs with contributions from the relevant germ layers in particular the mesoderm - the source of vasculature that is required ubiquitously across all organ systems. As organogenesis progresses and so that the developing organism can achieve eventual autonomy, the diversity of functions and therefore cell types that perform them expand considerably.

Though there are multiple incarnations of single cell sequencing protocols including scRNA-seq, CEL-seq, MARS-seq, STRT-seq and Quartz-seq, Smart-seq2 was used for experiments described in the previous chapter as it has no requirements for specialist equipment and pri-

Single-cell census of mouse organogenesis

marily due to the familiarity with the protocol within the Göttgens laboratory [Hashimshony et al., 2012; Islam et al., 2011; Jaitin et al., 2014; Ramskold et al., 2012; Sasagawa et al., 2013; Tang et al., 2009]. Smart-seq2 using the appropriate indexes can be performed on either 96-well or 384-well plates [Picelli et al., 2014].

Some of the limitations of Smart-seq2 have been discussed, see section 3.15.1. They include lots of pipetting steps and requirement for use of multiple plates with increasing cell numbers, both can lead to variations within experiments resulting in more failures in some plates than others biasing cell types that are finally sampled and creating batch effects. This can be further confounded by sorting different populations by plate so making it difficult to ascertain the cause of any variation. Duplicate reads where multiple reads are produced from the same mRNA molecule cannot be unambiguously identified when using Smart-seq2 in contrast this problem is resolved in some protocols by the use of unique molecular identifiers (UMIs). These are a random sequence of nucleotides that are added to each molecule along with a cell identifier and form an integral part of some protocols. Post-sequencing the cell barcodes are demultiplexed and as each UMI corresponds to a single mRNA molecule these counts are often referred to as digital gene expression (DGE). Advantages of Smart-seq2 include it has reasonable scalability with automated clean up steps, no requirement for specialist equipment and sampling of full-length transcripts that may potentially be used for enumerating splice variants.

For sampling all cell types in a mouse embryo immediately after gastrulation during very early organogenesis with increasingly complex cell ensembles, Smart-seq2 was impractical and cost prohibitive. In this context emerging droplet based methods such as Drop-Seq and InDrop offer the ability for much greater scalability whilst concurrently reducing library preparation and sequencing costs and reducing batch effects by early pooling of barcoded cells and increased multiplexing [Klein et al., 2015; Macosko et al., 2015]. In this chapter we used a droplet based proprietary platform, 10x Genomics® Chromium™ to generate single-cell libraries from E8.25 mouse embryos to capture the earliest stages of organogenesis [Zheng et al., 2017].

4.2 Embryo collection and data generation

Embryos were harvested from two time-mated B6CBAF1/J dams crossed with B6CBAF1/J males at E8.25, fig. 4.1. One litter being resorbed was discarded, from the remaining litter, 7 embryos were harvested and single-cell suspensions generated. A critical step in this workflow was that the single cell suspensions had to be made up to an appropriate cell

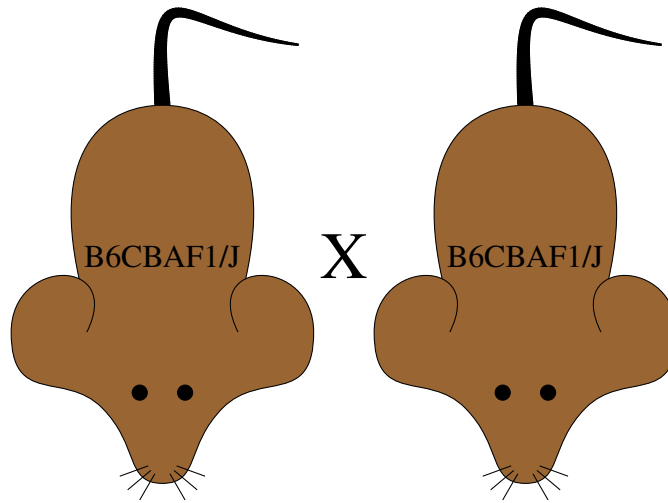


Fig. 4.1 A male B6CBAF1/J and a female B6CBAF1/J were crossed to yield F2 embryos. B6CBAF1/J, an F1 cross, were generated by crossing a female C57BL/6J with a male CBA/J.

concentration. Drop-seq based technologies such as the 10x Genomics® Chromium™ as summarised in fig. 4.2 use a microfluidics device to bring cells and enzymes together with barcoded beads into a nanolitre-scale Gel Bead-In-EMulsions (GEM). The 14 bp bead barcodes are randomly sampled from a pool of $\sim 750,000$, so each cell has a unique barcode. In the first generation technology used in this experiment but now discontinued up to 6000 cells could be sampled in a single experiment. Collisions or different cells receiving the same barcode can conceivably occur by two mechanisms. Firstly by random chance two captured cells can be captured in two separate GEMs but by chance both GEMs contain beads with identical cell barcodes, an index collision. Probability of no collision p_c can be calculated as:

$$p_c = \prod_{k=0}^{n-1} \frac{q-k}{q}, \quad n \leq q$$

Where n is the number of cells sampled and q is the pool of available indices. The probability of no collisions in a 1000 cell experiment is 0.51 but by the time an experiment has 6000 cells the probability of no collision is extremely small 3.55×10^{-11} , i.e. collisions are almost certain. The distribution of the number of collisions in experiments with varying numbers of cells were simulated and are shown in fig. 4.3. The number of collisions can be minimised by sampling fewer cells.

The second mechanism by which mRNA moieties from two different cells can receive the identical cell barcode is by becoming incorporated into the same GEM. This risk is mitigated by adding cells at a limiting dilution so that the majority of GEMS have no cells and the

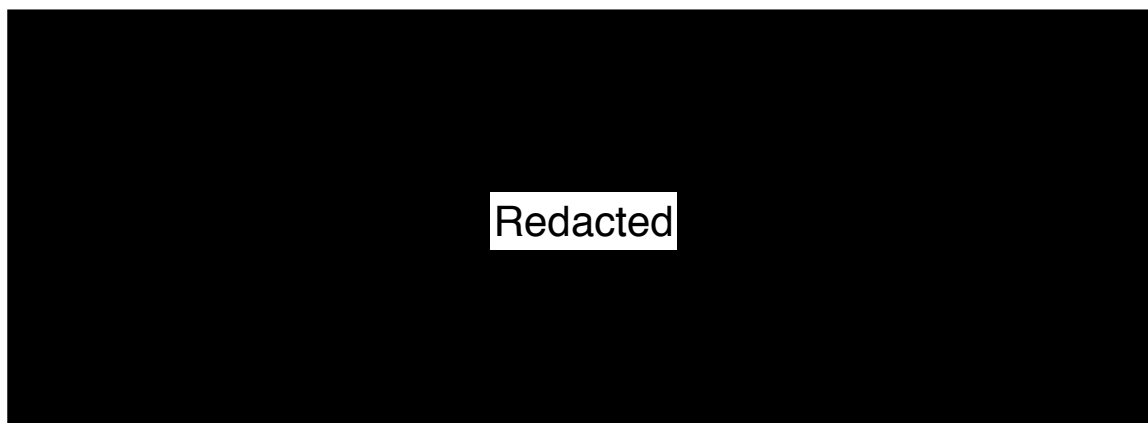


Fig. 4.2 Summary of the 10x Genomics® Chromium™ workflow. Cells are loaded at a limiting dilution so that most (~93 % to 99 %) GEMs generated will contain no cells with the aim that the remainder contain a single cell. Image taken from 10x Genomics® Chromium™ user manual.

remaining are most likely to only have one. This rate can again be controlled by keeping the number of cells per experiment low. 10x Genomics® Chromium™ helpfully provide a table of suggested cell dilutions and expected multiplet rates table 4.1. Such a table can be generated by estimating the overall multiplet rate using mixed species cells where a collision between cells of two different species can be easily identified and therefore the overall collision rate inferred.

Table 4.1 Approximate expected multiplet rates at different cell loading concentrations for the first generation 10x Genomics® Chromium™ protocol.

Multiplets rate (%)	Number of cells loaded	Expected number of cells recovered
1.1	2600	1200
1.8	4300	2000
2.7	6400	3000
3.6	8500	4000
4.5	10,700	5000
5.3	12,800	6000

After generating single-cell suspensions from the 7 litter-mate embryos cells were counted using a Neubauer chamber haemocytometer, providing an estimate for cell concentration and the assessing the effectiveness of the dissociation, as clumps of cells can clog channels on the microfluidic chip scuppering the experiment.

Embryos 2, 3 and 4 were selected for further processing owing to the higher cell yield. Images of the embryos are shown in fig. 4.4.

4.2 Embryo collection and data generation

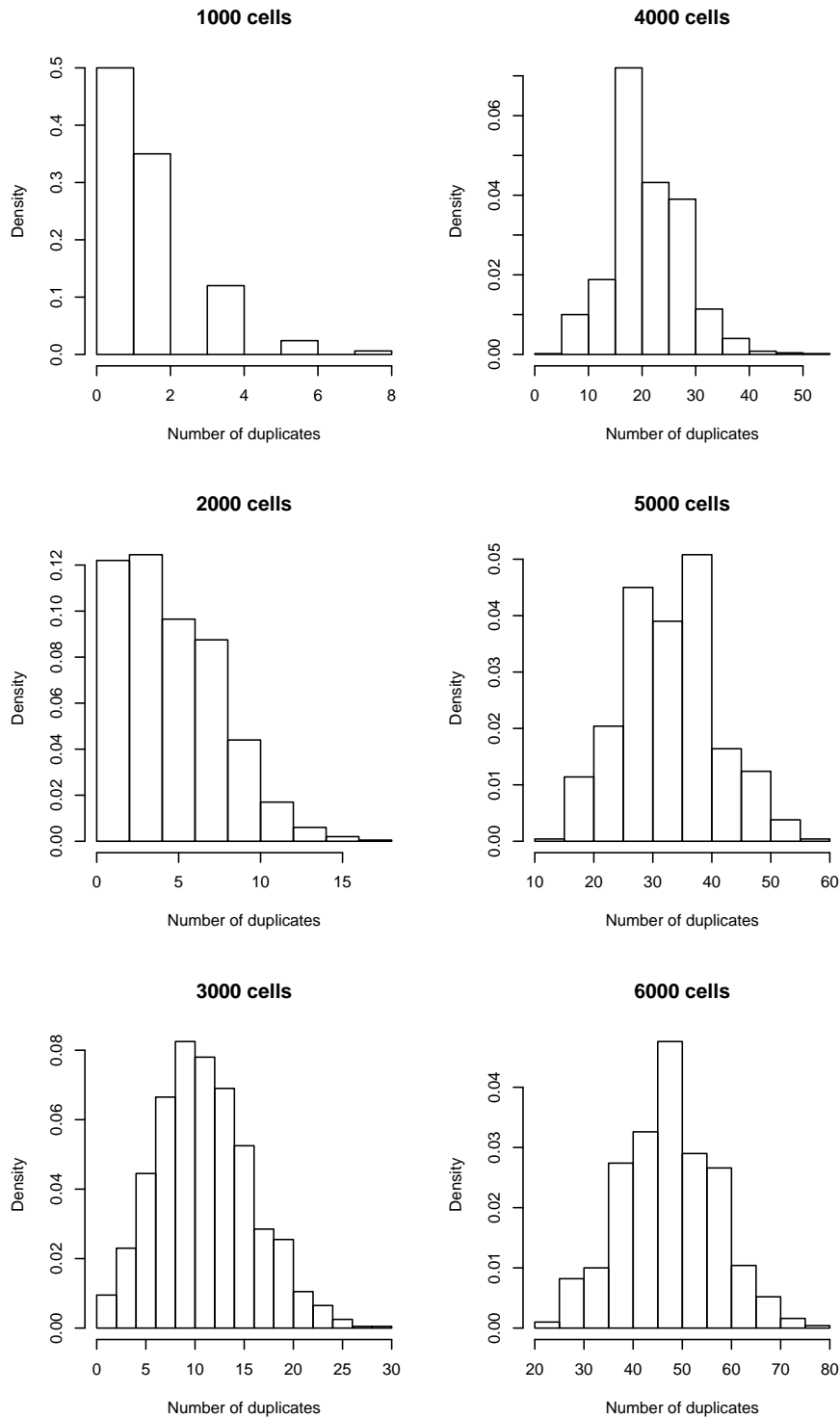


Fig. 4.3 Distributions of the numbers of cells with a barcode shared with another cell, when sampling the given numbers of cells from a library of $\sim 750,000$ unique cell barcodes. Distributions shown were calculated on 1000 simulations.

Single-cell census of mouse organogenesis

Table 4.2 Cell concentrations from each embryo after re-suspending dissociated cells in 0.04% BSA. Calculated after cells counted on Neubauer chamber haemocytometer.

Embryo	Cell concentration (cells/ μ l)
1	95
2	235
3	710
4	365
5	45
6	90
7	165

Each of the 3 embryo single cell suspension samples was split equally so as to be processed across two lanes of the microfluidics chip. The cells with reagents, GEM beads and the partitioning oil were loaded onto the chip and the cells were incorporated into GEMs by the 10x Genomics® Chromium™ instrument. The resulting emulsions with each cell trapped with a single bead in an oil droplet was then incubated for reverse transcription fig. 4.5a. Care was taken when transferring samples from the microfluidics chip to the PCR plate to ensure the emulsion was not disrupted.

After reverse transcription within each individual GEM the generated cDNA was uniquely barcoded so the emulsion could be disrupted and the cDNA pooled for subsequent steps. The cDNA was then amplified, sheared and the adaptors and sample indices were incorporated into the final library [Zheng et al., 2017]. The sample indices allowed sample pooling and massive multiplexing on the sequencer. There were 4 sample sequences allocated to each sample index so 24 sample sequences in this experiment table 4.3. The libraries were generated by Dr Fernando Calero-Nieto.

Table 4.3 Sample indices used in this experiment and their corresponding 8 bp sequences

Embryo	Sample Index	Sequence			
		1	2	3	4
2	SI-3A-A3	AAAGCATA	CTGCAGCC	GCCTTTAT	TGTAGCGG
2	SI-3A-A4	AGAACGCC	CATGGCAG	GTCTTTGA	TCGCAATT
3	SI-3A-A5	ATTGGGAA	CAGTCTGG	GGCATACT	TCACACTC
3	SI-3A-A6	ACGGGACT	CTTTCGAC	GAACATGA	TGCATCTG
4	SI-3A-A7	AGGTCATA	CTCATCAT	GCTGAGGG	TAACGTCC
4	SI-3A-A8	ATGATACG	CCACAGAA	GACTGTTC	TGTGCCGT

4.2 Embryo collection and data generation

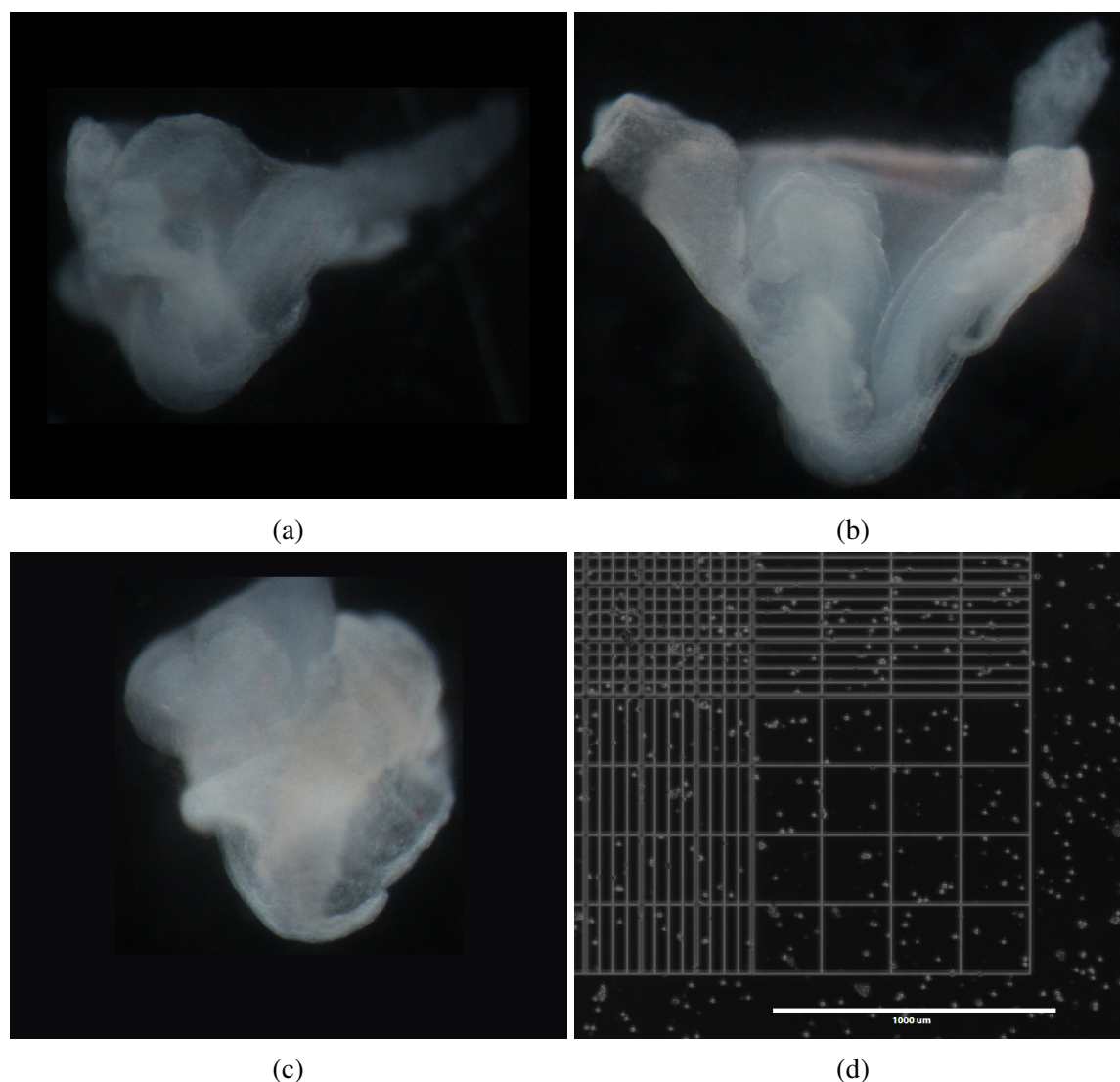


Fig. 4.4 Oblique anterior view of embryo 2 where head fold, cardiac crescent , anterior intestinal portal and allantois are clearly visible and the visceral endoderm has been mostly removed (a). (b) Embryo 3 where the visceral endoderm has been split to allow a lateral view of the embryo. Allantois and amnion can be seen. Embryo 4 is image postero-superiorly (c). (d) Image of dissociated cells in Neubauer chamber haemocytometer from a different experiment but representative of the current experiment.

The sample libraries were diluted and submitted to the sequencing core at Cancer Research UK Cambridge Institute by Dr Nicola Wilson. The libraries were sequenced over all 8 lanes of a single flowcell on the Illumina® HiSeq™ 2500 on high-throughput mode. The sequencing core returned the sequenced reads as zipped demultiplexed fastq files. The demultiplexed files were processed using Cell Ranger v2.1 which performed preliminary QC and produced aggregated count matrices for further analysis.



Redacted

Fig. 4.5 (a) Barcoded primers are released from beads after encapsulation of cells and into GEMs. These primers containing an anchored oligo-dT that captures cellular mRNAs ready for reverse transcription when a TSO is added to the 3' end of the cDNA molecule allowing subsequent efficient amplification, see section 3.2 for further details of transcript switching. The primers additional have a 10 bp randomer providing a UMI. (b) Construct of the final library. The read 1 primer binds at R1 and the cDNA sequence is read. The i7 primer binds to a region of R2 and reads the cell barcodes. The sample index is read from the grafted P5 oligomer and the read 2 primer binds at R2 after the complementary strand is generated to sequence the UMI. Image from [Zheng et al., 2017].

7106 cells were identified by Cell Ranger after aggregating all 6 samples from 3 embryos. Further filtering removed cells with greater than 5% counts mapping to mitochondrial genes or cells with more than 6000 expressed genes, see fig. 4.7. 6978 cells passing all QC parameters, were sequenced to a mean depth of $\approx 100,000$ reads per cell, with median expression of 10,223 (IQR 7447 to 13,132) UMIs and a median of 2828 (IQR 2352 to 3247) genes, see fig. 4.8. The 51,158 annotated genomic features were filtered with only 21,818 expressed in a minimum of 3 cells retained for downstream analysis.

Inspecting the counts plot shows two branches of cells suggesting some cells despite being sequenced to equivalent depths express different numbers of genes. The cause is immediately unclear but the effect is consistent across embryos and is not due to a batch effect from an outlying sample or embryo, right most panel in fig. 4.7. Postulating on the nature of these two populations it is conceivable that these are two different cell populations with differing amounts of mRNA possibly due to cell type, proliferative state or cell cycle stage. It is

4.2 Embryo collection and data generation

intriguing these two populations are so noticeable during pre-processing, such an early stage of the analysis.

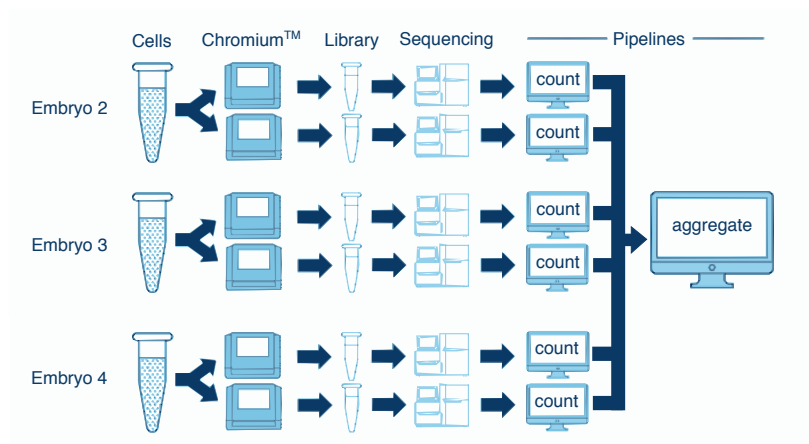


Fig. 4.6 Experimental and computational workflow. Each embryo was processed across two channels of the microfluidics chip generating technical replicates. The generated libraries were pooled and sequenced across 8 lines of a flowcell on the Illumina HiSeq 2500 on high-throughput mode. Resulting fastq files were processed separately using Cell Ranger's count command and then aggregated together using Cell Ranger's aggr command.

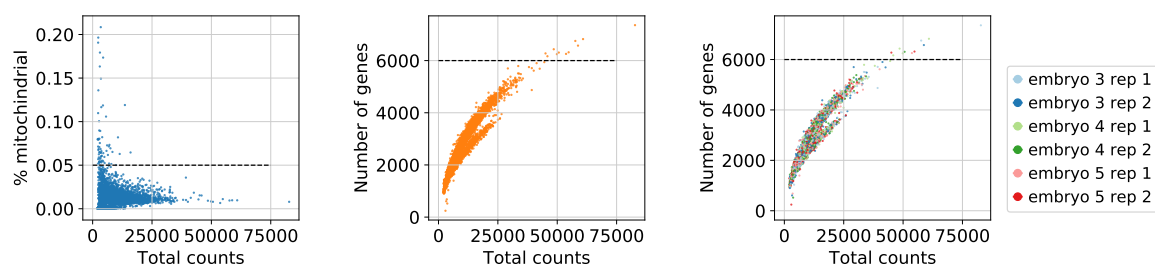


Fig. 4.7 Data filtered to remove cells with greater than 5% mitochondrial mapped reads or those with greater than 6000 different genes expressed. Thresholds are depicted by the horizontal dashed lines.

Normalisation in bulk RNAseq data often involves normalising for gene length but was not applied in the previous Smart-seq2 dataset with the understanding that comparisons were being made between cells and not between genes, so that the same gene in different cells will have the same bias. In the case of digital gene expression where due to the UMI tagging only the 3' end of the molecule is assayed, gene length has no effect on capture probability for transcripts beyond a minimum length which for the 10x Genomics® Chromium™ protocol is ≈ 400 bp. This residual transcript length limitation exists because of size selection used after shearing cDNA during library preparation.

Single-cell census of mouse organogenesis

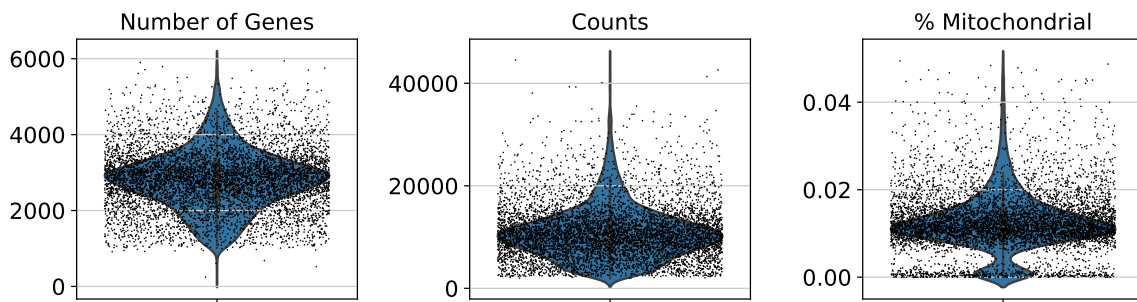


Fig. 4.8 Summary of cell counts, numbers of genes expressed in cells and the percentage reads mapped to mitochondrial genes

Data was normalised by calculating size factors using scran's clustering approach and further gene filtering was performed by finding highly variable genes. In this case the method by Brennecke et al. [2013] provided a poor fit to the mean, coefficient of variance relationship, fig. 4.9a. A much simpler linear model was therefore fitted to the log transformed values, fig. 4.9b and the 7585 genes with positive residuals were defined as highly variable and retained for further analysis.

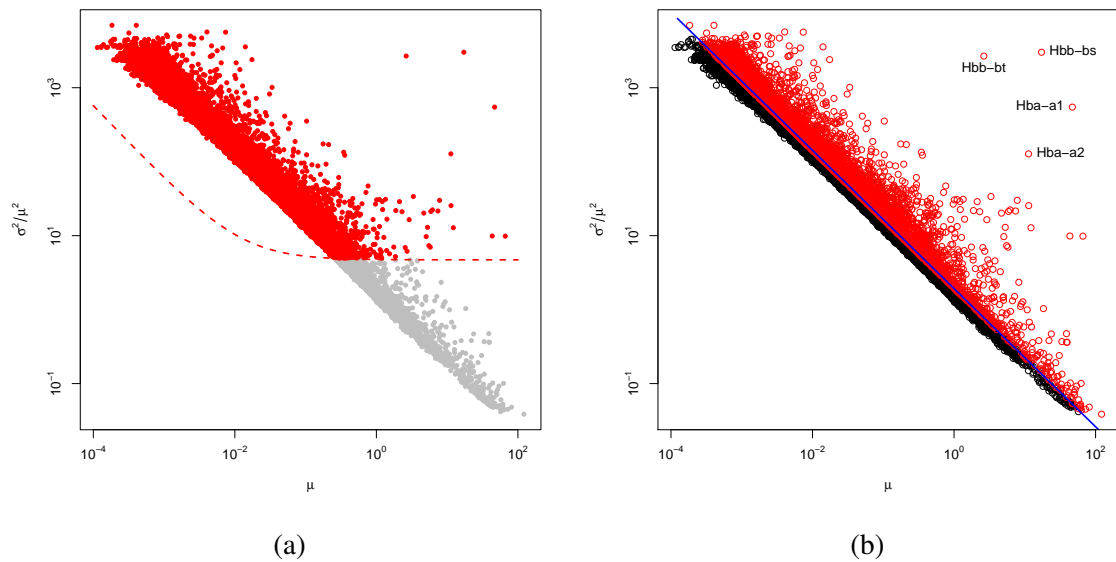


Fig. 4.9 (b) Log linear model fitted to mean coefficient of variance relationship is robust to the haemoglobin outliers. Model using algorithm described by Brennecke et al. [2013] does not fit the data correctly (a).

4.3 Defining clusters

A distance matrix, $D \in \mathbb{R}^{6978 \times 6978}$ was calculated from the resulting data matrix of 6978 cells and 7585 genomic features using:

$$D_{ij} = \frac{1 - \text{src}(\text{cell}_i, \text{cell}_j)}{2}, \quad \text{where } \text{src} \text{ is the Spearman Rank correlation}$$

The Spearman rank correlation provides a dissimilarity metric that is resilient to any cell-wise gene count normalisation as it is based on gene ranks but can still be effected by library sequencing depth due to the effect of drop-out. Lowly expressed transcripts will be randomly captured so that even in a simulation, identical cells will have a different set of genomic features with zero counts so generating a dissimilarity. Further different cell types may well have differing cardinalities for the lowly expressed transcript set so that dissimilarities between cells of the same type may be different between cell types. Additionally if these sets of lowly expressed transcripts have high cardinality, dissimilarity may be largely driven by them rather than more ‘informative’ genomic features.

A graph was generated from the dissimilarity matrix by applying a Gaussian kernel as described for diffusion maps by Haghverdi et al. [2015] with some modifications to remove spurious edges using *roots*. Louvain clustering was performed on the graph to identify 22 clusters from which 19 were co-localised on tSNE and could be annotated to identify cell type. Dimensionality reduction was performed on D using the tSNE algorithm, fig. 4.10.

Cluster identities were specified by reviewing differential gene expression between the set of cells in target cluster and its complement. This was done by first filtering for:

1. Genomic features with a minimum sum of normalised log transformed counts across all cells of 1.
2. A minimum fold change between the selected cluster and all remaining cells of 5.
3. Genomic features where fewer than 20% of the reference cells (i.e. those not in the cluster) express the gene above the mean in the target cluster population.

By applying this filter fewer genes were tested so increasing power when correcting for multiple testing.

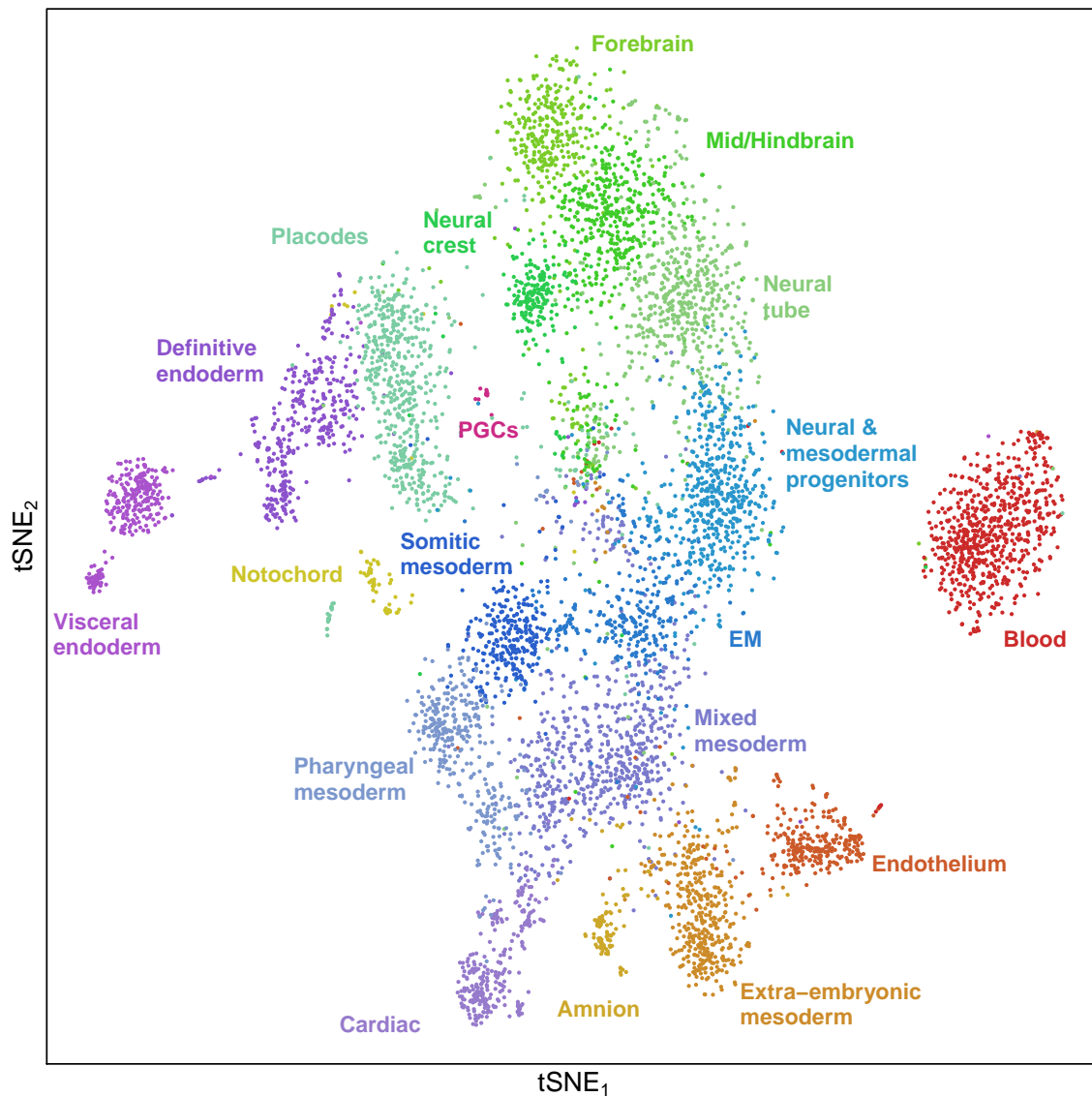







Fig. 4.10 tSNE dimensionality reduction of 6978 single cell expression profiles on 7585 genomic features. 19 clusters were defined using Louvain clustering on a graph calculated from a distance matrix based on the Spearman rank correlation related distance. Colours are grouped by the tissues germ layer of origin. Key genes used to identify cluster identity are summarised in table 4.4.

4.3 Defining clusters

Table 4.4 Summary of annotated cell clusters, final post-QC cell numbers and identified marker genes. Associated colours used in fig. 4.10 are shown in rectangles to the left; colours are grouped together by germ layers. Key genes were identified by filtering for only those with a fold change greater than 2 in the selected population and applying the student t-test with Welch's correction.

	Cluster	Cell number	Key genes
	Forebrain	361	<i>Fezf1, Fezf2, Six3, Pax6, Otx2, Hesx1, Lhx2, Lhx5, Rax, Six3</i>
	Mid and Hindbrain	454	<i>En1, Fgf8, Fgf15, Hoxb1, Vgll3, Pax5, Hes3</i>
	Neural crest	201	<i>Sox10, Foxd3, Wnt1, Tfap2a, Tfap2b, Cmtm5, Apod</i>
	Neural tube	596	<i>Fgfbp3, Hoxd4, Mafb, Arhgef28</i>
	Placodes	584	<i>Dlx5, Ripply3, Wnt6, Tfap2a, Gjb3, Foxi2, Foxg1, Sema3a</i>
	Neural & mesodermal progenitors	559	<i>T, Sox2, Nkx1-2, Fgf17, Hes7, Fgf8, Wnt3a</i>
	Early mesoderm	293	<i>Tbx6, Mesp2, Ripply2, Dll1, Dll3, Rspo3, Hes7</i>
	Somitic mesoderm	263	<i>Aldh1a2, geneTcf15, Meox1, Meox2, Uncx, Cd36, Cer1</i>

Continued on next page

Single-cell census of mouse organogenesis

Table 4.4 – continued from previous page

Cluster	Cell number	Key genes
Mixed mesoderm	680	<i>Foxf1, Osr1, Lhx1, Hoxd1, Hoxd9, Nxf3, Hand2</i>
Pharyngeal mesoderm	357	<i>Ebf1, Ebf2, Six2, Batf, Cped1, Ebf1, Tbx1</i>
Cardiac mesoderm	279	<i>Tnnt2, Hspb1, Mesp1, Tnni1, Myl4, Myl7, Acta2, Actc1, Csrp3, Tnni1, Ankrd1, Smarcd3, Hspb7, Asb2, Myh6, Pgam2</i>
Extra-embryonic mesoderm	406	<i>Plac1, Pitx1, Hoxa10, Hoxa11, Hoxc10, Tbx4</i>
Amnion	103	<i>Postn, Tdo2, Lum, Wisp1, Plac8</i>
Endothelium	290	<i>Kdr, Lmo2, Pecam1, Sox17, Hhex, Fli1, Gata2, Tek, Eng, Tall, Lmo2, Etv2, Gadd45g</i>
Blood	672	<i>Hbb-bh1, Hba-x, Hbb-bh0, Car2, Blvrb, Gypa, Gata1, Klf1</i>
Notochord	68	<i>Noto, T, Shh, Samd3, Nog, Defa30</i>
Primordial germ cells	12	<i>Dppa3, Nanog, Pou5f1, Kit</i>

Continued on next page

Table 4.4 – continued from previous page

Cluster	Cell number	Key genes
Definitive endoderm	384	<i>Apela, Cpm, Foxa1, Trh, Pyy, Nepn, Pax9, Gpx2</i>
Visceral endoderm	284	<i>Afp, Ttr, Amn, Apoa1, Apom, Apoa4, Lgals2</i>

Post filtering, differential gene expression was calculated using the student t-test with Welch's correction for differing group sizes within the filtered gene set. Multiple testing correction was performed using independent hypothesis weighting and $p < 0.05$ was considered significant [Ignatiadis et al., 2016]. Allocation of cellular identity to the clusters for some was very straightforward but not clear for others, table 4.4. The blood cluster could be immediately identified by high expression of haemoglobin genes *Hba-a1, Hbb-bh1, Gypa* amongst others. The endothelial cluster with high expression of *Pecam1 Sox17, Kdr* and *VE-Cadherin* was also readily recognised.

The primordial germ cell cluster despite consisting of only 12 cells (7 from embryo 2, 1 from embryo 3 and 4 from embryo 4) was identified using this clustering method and differential genes expression analysis allowed immediate allocation of cell type, table 4.5. This cluster expressed known markers such as *Dppa3(Stella), Nanog, Ifitm3* and *Pou5f1*. *Kit* is also expressed highly within the PGC cluster but the corrected p value is not significant.

Dppa3 is very highly and specifically expressed in this cluster with a fold change of 5350. Simply looking at genes highly correlated with *Dppa3* also identifies other known PGC related genes including *Prdm14* and *Sox15*. *Dnd1* has 53 times greater expression in PGCs than other cells has been shown to be required for PGC survival and suppression of germ cell tumours in mice [Yabuta et al., 2006; Yamaji et al., 2017].

Sex determination could not be phenotypically assessed under the microscope at this early stage but was evaluated by assessing expression of Y chromosome genomic features and *Xist* expression independently and both methods showed that embryos 2 and 3 were male and embryo 4 was female, fig. 4.12. The only Y chromosome genomic feature that remained

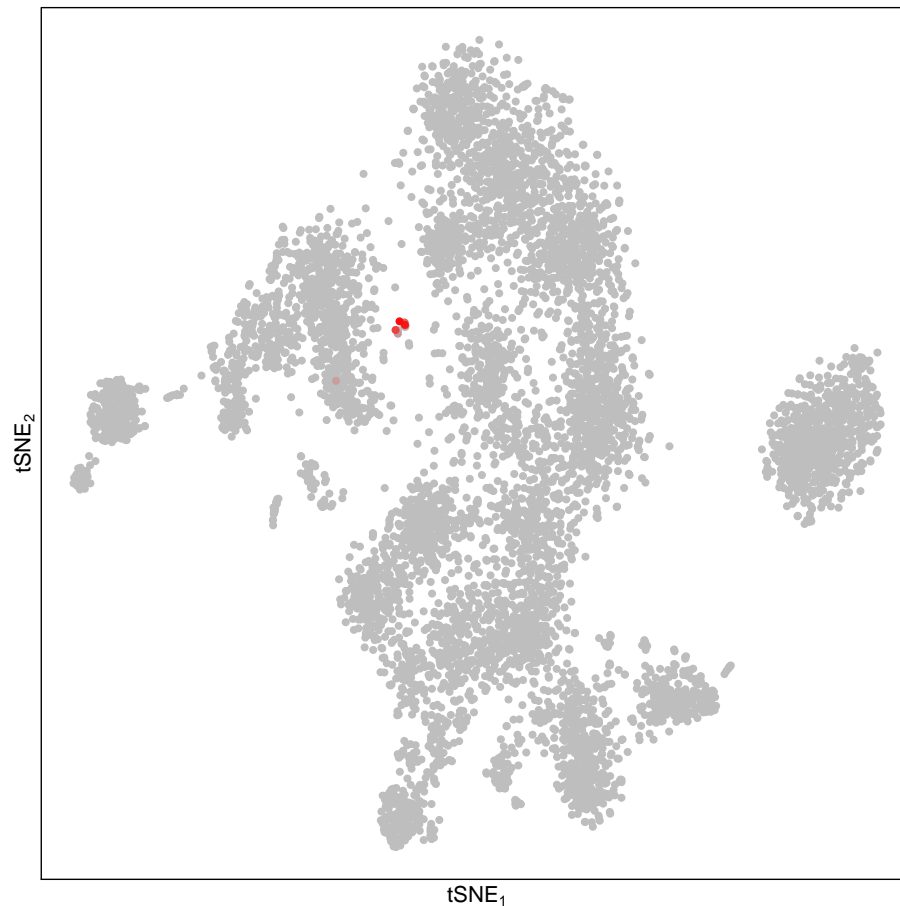


Fig. 4.11 *Dppa3* is almost exclusively expressed in the PGC cluster. Unlike other genes described *Dppa3* appears to be truly specific to the PGC cluster within the E8.25 stage embryo. Gene expression is plotted so that cells that have high expression are brought to the fore.

after early gene filtering was *Gm29650* and this was expressed in 17, 22 and 0 cells from embryos 2, 3 and 4 respectively, consistent with the *Xist* result.

Reassuringly contributions to the different clusters is consistent across the embryos except for embryo 2 having much lower contributions to the blood and visceral endoderm populations due to it being almost wholly excised during dissection, see table 4.6 and fig. 4.4. This highlights how dissection of different embryonic tissues can help validate cell type assignment using a miss one out protocol. In this way multiple embryos with overlapping tissues could be used to validate all cell types by their known spacial localisation.

4.3 Defining clusters

Table 4.5 Genes identified as differentially expressed between the 12 primordial germ cells and all other cells. Fold change, unadjusted p values and IHW adjusted p values are reported [Ignatiadis et al., 2016].

Gene	Fold change	Unadjusted p values	IHW adjusted p values
<i>Psmel</i>	5.42	2.10×10^{-6}	6.33×10^{-6}
<i>Pou5f1</i>	2.38×10^1	4.86×10^{-6}	1.38×10^{-5}
<i>Ifitm3</i>	1.08×10^1	5.00×10^{-6}	1.42×10^{-5}
<i>Klf5</i>	1.78×10^1	7.56×10^{-4}	1.78×10^{-3}
<i>Dnd1</i>	5.32×10^1	1.35×10^{-3}	3.17×10^{-3}
<i>Fam222a</i>	2.02×10^1	5.11×10^{-3}	1.03×10^{-2}
<i>Alpl</i>	9.53	7.44×10^{-3}	1.67×10^{-2}
<i>Bst2</i>	8.33	1.07×10^{-2}	1.96×10^{-2}
<i>Tfap2c</i>	1.01×10^1	1.32×10^{-2}	2.34×10^{-2}
<i>Ttc28</i>	5.48	1.32×10^{-2}	2.42×10^{-2}
<i>Ypel3</i>	5.15	1.31×10^{-2}	2.73×10^{-2}
<i>Tapbp</i>	1.98×10^1	1.58×10^{-2}	2.87×10^{-2}
<i>1700019D03Rik</i>	7.52	1.84×10^{-2}	3.20×10^{-2}
<i>Esrp1</i>	1.01×10^1	2.31×10^{-2}	4.05×10^{-2}
<i>Nanog</i>	2.60×10^2	2.58×10^{-2}	4.32×10^{-2}
<i>Dppa3</i>	5.35×10^3	2.60×10^{-2}	4.39×10^{-2}
<i>Pla2g16</i>	6.40	2.54×10^{-2}	4.44×10^{-2}
<i>Gjb3</i>	1.18×10^1	2.34×10^{-2}	4.69×10^{-2}

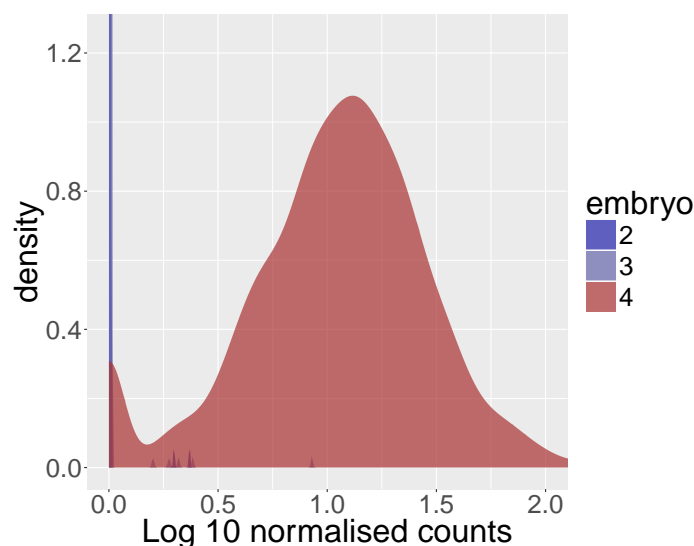


Fig. 4.12 Histograms showing distributions of *Xist* expression in the 3 embryos. Embryos 2 and 3 have no appreciable *Xist* expression while embryo 4 shows strong expression in most cells. The plot has been trimmed along the y-axis.

Single-cell census of mouse organogenesis

Table 4.6 Each cluster received equivalent contributions from the different embryos except for the blood and visceral endoderm clusters.

Cluster	Embryo		
	Male		Female
	2	3	4
Forebrain	123	124	114
Mid/hindbrain	153	143	158
Neural crest	52	78	71
Neural tube	204	198	194
Placodes	214	211	159
Neural & mesodermal progenitors	192	171	196
Early mesoderm	107	95	91
Somitic mesoderm	80	74	109
Pharyngeal mesoderm	152	84	121
Mixed mesoderm	234	267	179
Cardiac	92	90	97
Extraembryonic mesoderm	160	112	134
Amnion	25	44	34
Notochord	29	20	19
Endothelium	100	105	85
Blood	11	387	274
Definitive Endoderm	119	176	89
Visceral Endoderm	29	156	99
Primordial germ cells	7	1	4

4.4 Within cluster substructure

Fine grained further characterisation of major clusters was performed by iteratively clustering major clusters where substructure was evident on manual inspection.

4.4.1 Definitive endoderm

The lungs, gastrointestinal tract, thyroid, thymus, liver and pancreas amongst others will form from or receive significant contributions from the definitive endoderm. Cluster allocation was refined by focusing specifically on the definitive endoderm cluster, redefining highly variable genes and re-clustering. 7292 highly variable genes were identified amongst the 384 definitive endoderm cells.

7 sub-clusters were identified using Louvain clustering, one cluster formed of a single outlier cell was excluded. tSNE and directed graph layout (DrL) were used to visualise the cell relationships in 2 dimensional plot, fig. 4.13. Marker genes were identified by performing differential gene expression analysis and some key genes are summarised in fig. 4.14. The hindgut and midgut were readily recognised using marker genes.

Hou et al. [2007] performed a systematic search of genes in the endoderm focusing on the early definitive endoderm between 0 - 6 somite stage; foregut and hindgut at 8 - 12 somite stage. Here we have in-silico dissected out the definitive endoderm at the 4 paired somite stage. They revealed genes such as *Pyy* and *Nepn* (referred to as *5730521E12Rik*) expressed in the foregut endoderm. *Pyy* at the 4 somite stage has two expression domains on the lateral wings and centrally in the medial ventral anterior foregut on validation in-situ hybridisation experimnts [Hou et al., 2007]. The DrL also appears to split the *Pyy* expressing foregut into two domains one of which is very closely related to the pharyngeal endoderm, fig. 4.13 seemingly revealing the two *Pyy* expressing foregut domains. The embryos described here were at an earlier stage than [Hou et al., 2007] but even at this early time-point substructure consistent with later genes is identified furthermore this is consistent with in-situs from this earlier paper, fig. 4.15.

An interesting finding is the *Ttr* expressing cells within the definitive endoderm cluster. On the tSNE of all cells in fig. 4.10 a subset of this group of cells is midway between the visceral endoderm cluster and the definitive endoderm cluster. Kwon et al. [2008] have shown using genetic lineage tracing that cells from the visceral endoderm colonise the embryonic gut. Here we see a sub-cluster of *Ttr*⁺ *Afp*⁺ expressing cells, the same genes used as markers by Kwon et al. [2008], that were computationally clustered with the definitive endoderm.

Single-cell census of mouse organogenesis

Table 4.7 Table of the number of cells allocated to each of the endodermal sub-clusters and the colours used.

Colour	Cell type	Numbers
■	Pharyngeal 1	62
■	Pharyngeal 2	48
■	Foregut	67
■	Midgut	29
■	Hindgut	85
■	VE derived	25
■	Unclear	56

In-silico dissection by recursive clustering therefore captures this intriguing process of integration of visceral endoderm cells into the embryonic gut.

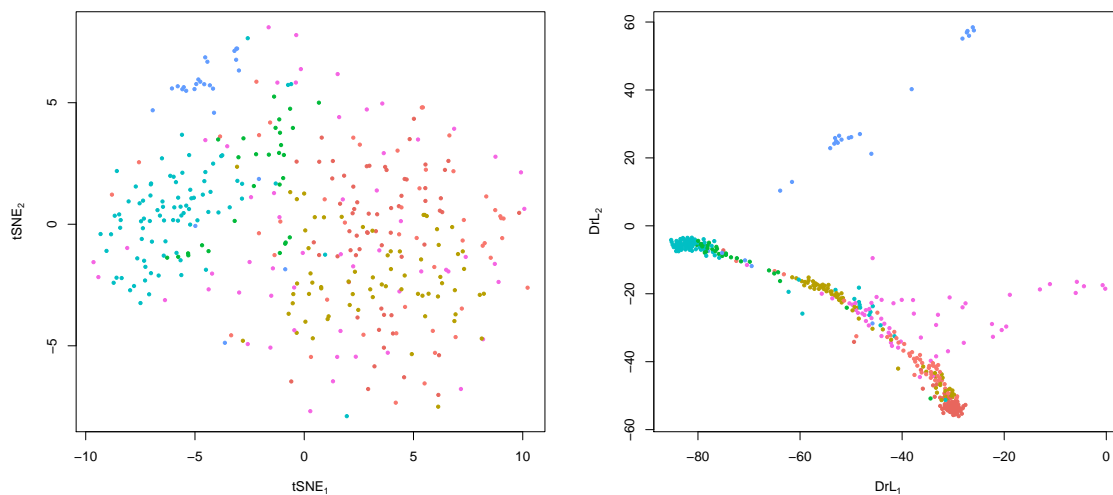


Fig. 4.13 tSNE dimensionality reduction on left and a directed recursive layout (DrL) on the right. Both separate out the major clusters but DrL resolves the Louvain clusters more clearly, clusters are better defined and relations between clusters are more apparent. The anterior endodermal clusters and the visceral endoderm component are clearly separated from the mid and hindgut. The final cluster labelled ‘Unclear’ is difficult to assign with only a single gene (*Pms2*) being differentially expressed between it and its complement fig. 4.14. Colours for the different clusters are defined in table 4.7.

2 Louvain clusters have been allocated as pharyngeal arch endoderm. Both clusters are very similar in terms of gene expression, both expressing *Six1*, *Pax1*, *Pax9*, *Irx3* and *Irx5*

4.4 Within cluster substructure

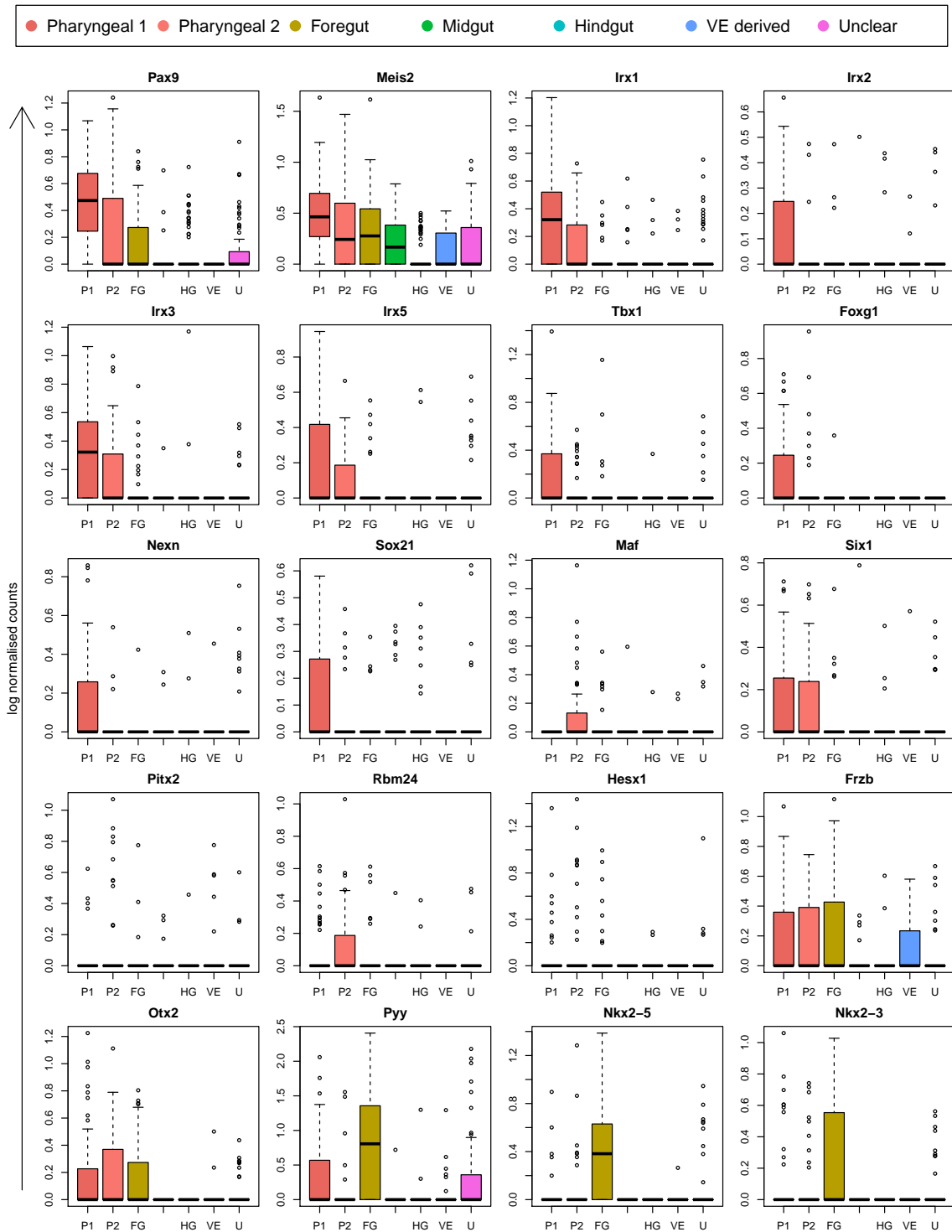


Fig. 4.14 (Continued on next page)

Single-cell census of mouse organogenesis

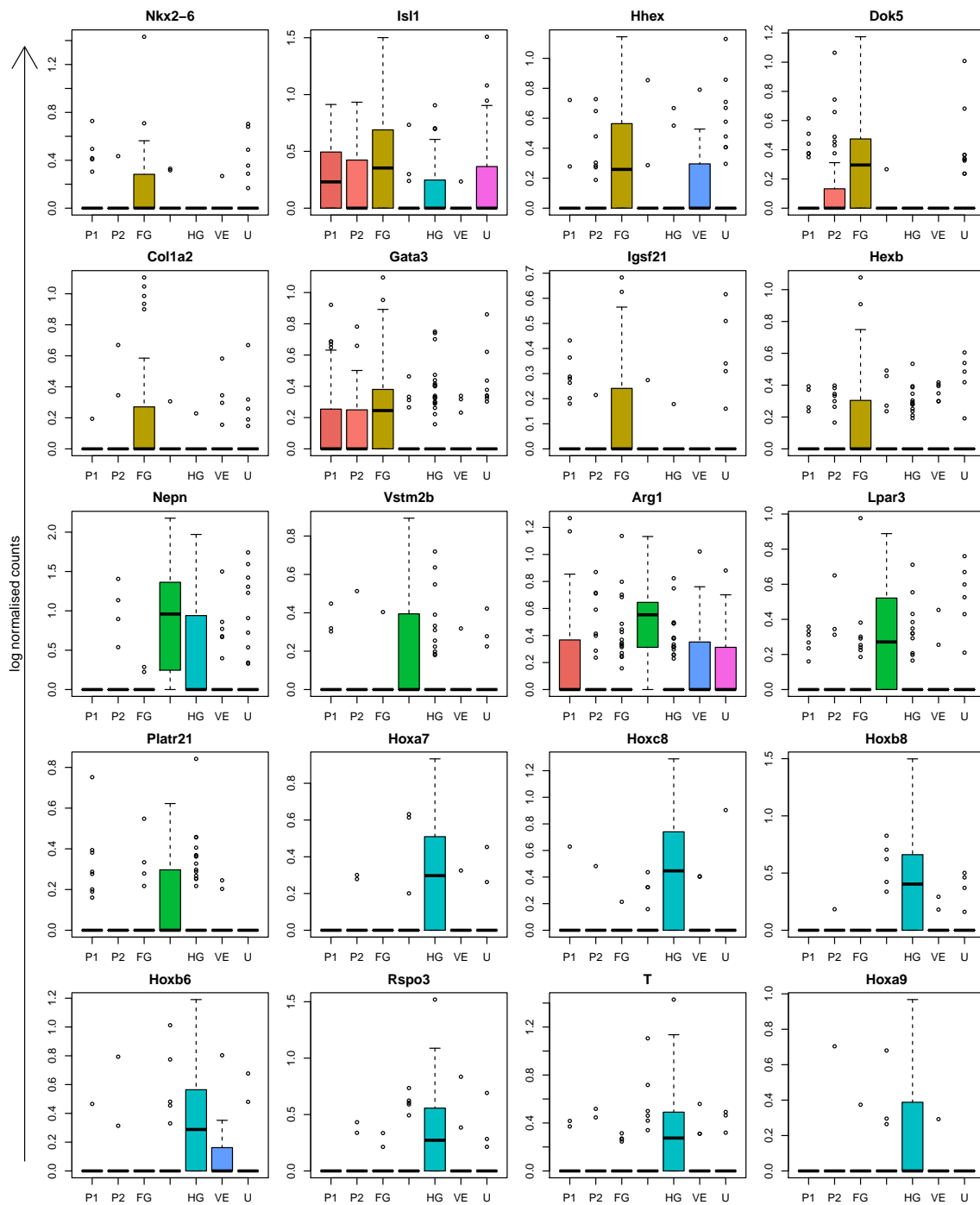


Fig. 4.14 (Continued on next page)

4.4 Within cluster substructure

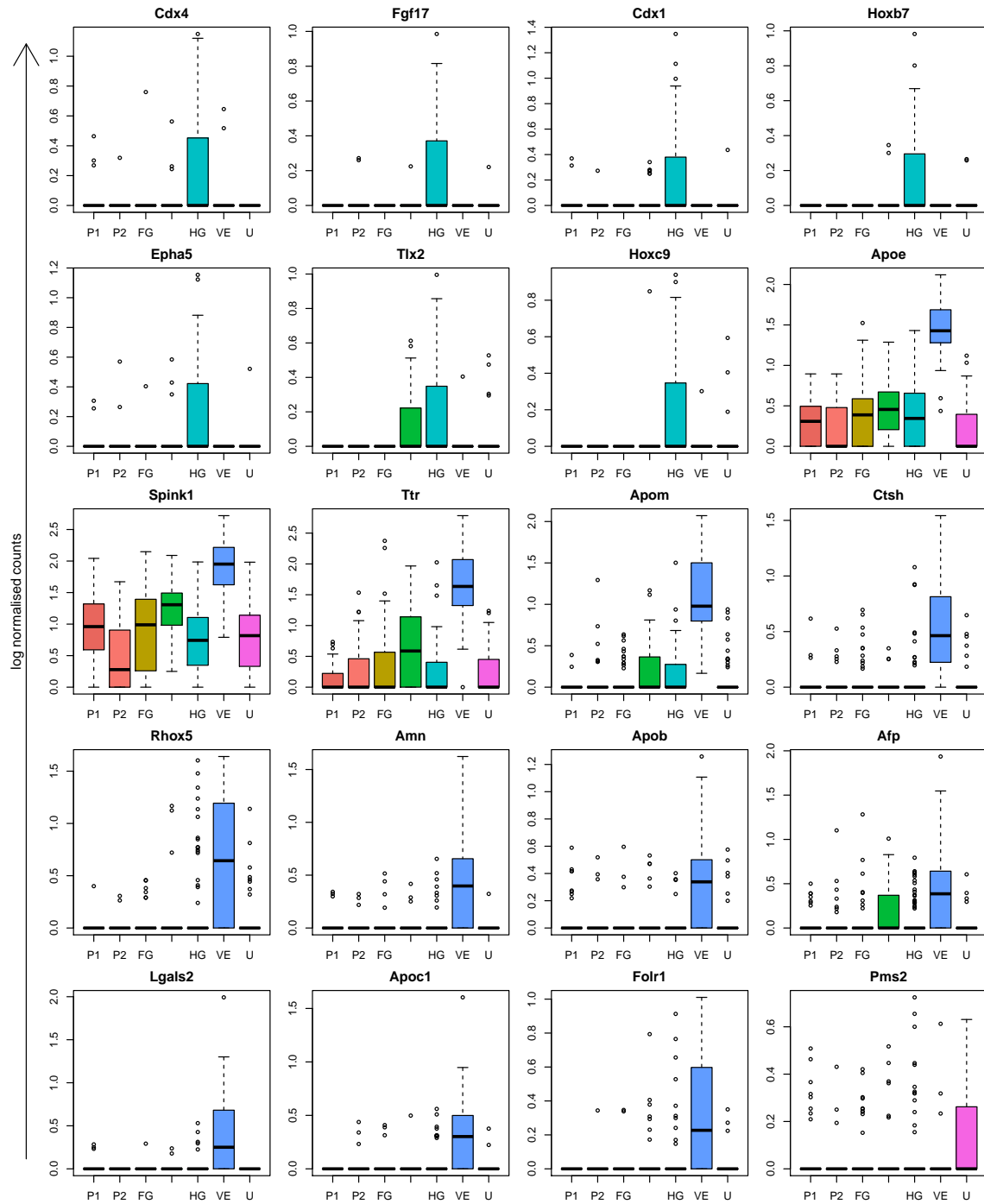


Fig. 4.14 Selection of genes and boxplots of their expression patterns across the Louvain clusters helped to assign putative cell type identities.

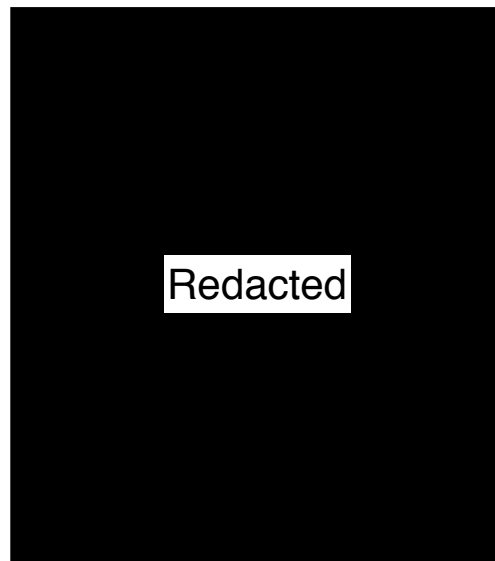


Fig. 4.15 In-situ hybridisation to *Pyy* in a 4 somite pair embryo from Hou et al. [2007] showing 2 domains of expression, laterally marked by arrow and median anterior endoderm marked by arrowhead.

[Becker et al., 2001; Bonnard et al., 2012; Chuang et al., 2018; Kaltenbach et al., 2009; Zou et al., 2006]. One gene that is significantly higher in the Pharyngeal 2 cluster is *Ttr* with a fold change of 2.45 ($p = 0.032$) possibly indicating that this represents a visceral endoderm component to the pharyngeal endoderm.

The final cluster is very difficult to define, remains unassigned and has been labelled Unclear. The only gene specifically expressed in this sub-cluster is *Pms2* a gene involved in apoptosis [Marinovic-Terzic et al., 2008; Shimodaira et al., 2003; Zeng et al., 2000]. Assaying a larger number of cells within these subsets may allow improved clustering and more powerful differential expression analysis.

4.4.2 Neural & mesodermal progenitors

A similar approach to that applied to the definitive endoderm can be applied to the neural and mesodermal progenitor cluster which likely consists of unipotent neural progenitors, unipotent mesodermal progenitors but also bipotent neuromesodermal precursors (NMP). A characteristic signature of these cells is co-expression of both *Brachyury/T* and *Sox2*. To focus in and identify relationships within the mesodermal populations the neural and mesodermal progenitor, neural tube, early mesoderm, mixed mesoderm and somitic mesoderm cell populations were selected and highly variable genes recalculated so identifying 2391 cells and 7713 genomic features.

4.4 Within cluster substructure

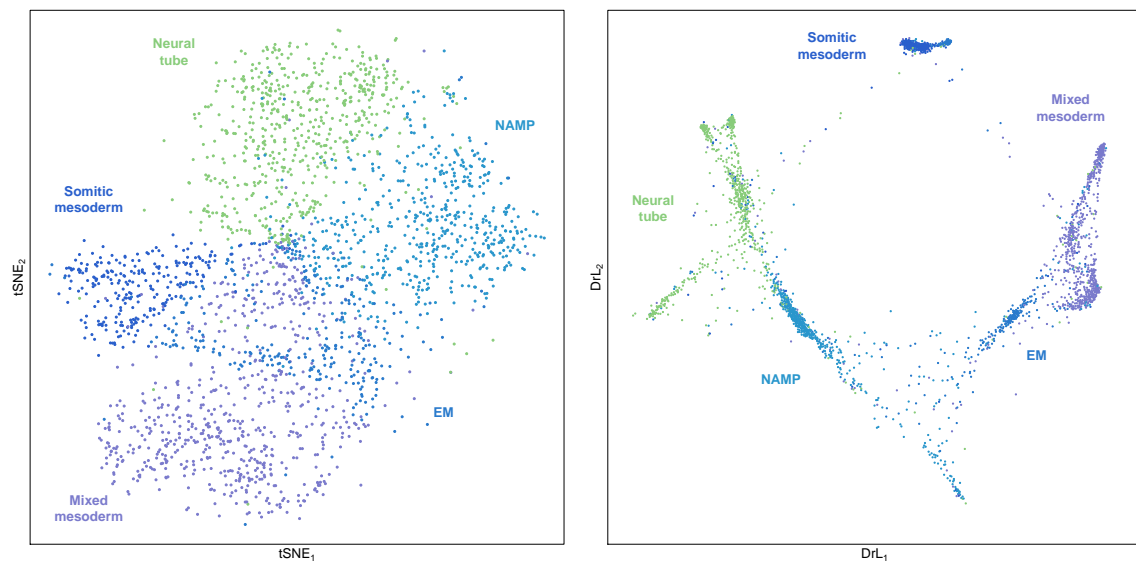


Fig. 4.16 tSNE and DrL side by side of neural and mesodermal progenitors (NAMP), early mesoderm (EM), neural tube, mixed mesoderm and somitic mesoderm populations coloured as shown in table 4.4. The new tSNE calculated on this subset of 2391 cells and 7713 genomic features is similar to the tSNE with all cells. The DrL displays the cells as almost bipartite with a potential bifurcation.

The tSNE for the mesodermal and neural tube populations fig. 4.16, as seen previously with the tSNE calculated for the definitive endoderm fig. 4.13 has a more diffuse appearance with clusters less tightly defined. Part of the reason for this may be that the populations are more homogenous and the variable gene calculation is therefore more prone to selecting biologically irrelevant features i.e. those that show variation between embryos or batches producing noise.

Despite the described difficulties with feature selection, the DrL, the right panel in fig. 4.16, separates the clusters and suggests relationships between the clusters that were not apparent on the tSNE. Applying Louvain clustering identified a *Sox2 T Cyp26a1* co-expressing cluster within the neural and mesodermal precursor population, fig. 4.17.

Having defined an NMP-like cluster, genes highly expressed in this cluster as compared to the other mesodermal clusters were identified. The top 50 genes that are significantly more highly expressed are summarised in table 4.8. This includes genes in pathways recognised to be activated in NMPs including *Brachyury/T, Nkx1-2, Cdx1, Cdx2, Cdx4, Wnt5b*. Multiple Hox genes are also expressed differentially including *Hoxa7, Hoxa9, Hoxb8, Hoxb9, Hoxc4, Hoxc6, Hoxc8, Hoxc9* and the long non-coding *Hoxaas3* and *Hoxb5os* [Diez del Corral and

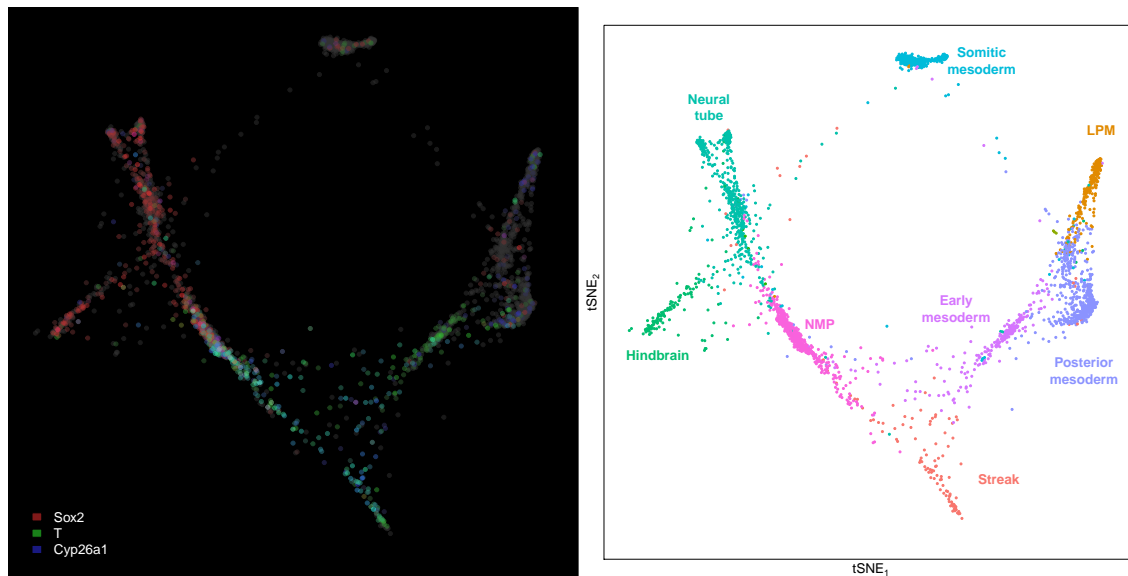


Fig. 4.17 Combined gene expression of *Sox2*, *T* and *Cyp26a1* on a single plot similar to a fluorescent microscopy image. A sub-population of NAMP cells is marked by concomitant high expression of these marker genes suggesting that these represent NMP or NMP-like cells. The Louvain clustering in the right panel splits that NAMP population into two, one of which corresponds to this NMP-like population. Compare clustering with previous coarse clustering fig. 4.13. LPM - Lateral plate mesoderm.

Morales, 2017; Henrique et al., 2015; Turner et al., 2014]. Table 4.8 now provides a list of new candidate genes that may be useful in identifying NMPs.

Another interesting feature of both tSNE of all cells in fig. 4.10 and the tSNE of the mesodermal and neural tube subset of cells fig. 4.17 is that there is an outlier subset of neural tube and in fig. 4.10 this includes cells from the mid/hindbrain cluster. This subset likely corresponds to a set of cells derived from NMPs which has been proposed to contribute to not only posterior spinal cord and presomitic mesoderm but also anterior spinal cord and hindbrain [Henrique et al., 2015]. Midbrain and hindbrain precursors produced from in-vitro mESC differentiation assays have been shown to express *En1* and *Otx2* [Turner et al., 2014].

On fig. 4.17 the early root cells are not clear. They may be at the *Sox2*, *T*, *Cyp26a1* cells but alternatively there may be a more primitive progenitor in the salmon coloured cluster at the bottom of the plot. This expresses genes similar to the NMP cluster such as *T* and *Cyp26a1* but does not express *Sox2*. Additionally genes such as *Mixl1* and *Snail* are expressed highly with several Claudins *Cldn6*, *Cldn7* and *Cldn9*, suggesting these cells may be undergoing EMT so possibly representing remaining streak cells at the node-streak border.

4.4 Within cluster substructure

Louvain clustering has additionally split the mixed mesoderm cluster into a posterior mesoderm and a lateral plate mesoderm. On the DrL the posterior mesoderm appears to be further segregated into a postero-lateral mesoderm and possibly paraxial mesoderm. It has not been possible to define any intermediate mesoderm though its progenitors at least must be present. The lateral plate mesoderm (LPM) also consists of two leaves the splanchnopleure or the somatopleure LPM but this separation has not been established with current approaches.

Table 4.8 Top 50 genes more highly expressed in the NMP cluster relative to other mesodermal clusters. Unadjusted p-values and IHW adjusted p-values are given [Ignatiadis et al., 2016].

Gene	Fold Change	p-value	
		Unadjusted	IHW adjusted
<i>Hoxaas3</i>	2.78	1.08×10^{-125}	6.20×10^{-122}
<i>Slc2a3</i>	2.46	1.46×10^{-82}	2.38×10^{-79}
<i>Cdx2</i>	3.21	1.78×10^{-69}	2.57×10^{-66}
<i>Grsf1</i>	2.39	6.04×10^{-68}	6.91×10^{-65}
<i>Hoxc8</i>	4.09	8.77×10^{-68}	8.44×10^{-65}
<i>Cdx1</i>	2.86	2.65×10^{-66}	1.90×10^{-63}
<i>Nkx1-2</i>	7.28	1.53×10^{-65}	9.76×10^{-63}
<i>Hoxb9</i>	3.43	2.00×10^{-63}	2.01×10^{-60}
<i>Hoxb8</i>	2.91	6.42×10^{-62}	4.95×10^{-59}
<i>Epha5</i>	6.00	6.67×10^{-58}	2.14×10^{-55}
<i>Cdx4</i>	3.10	3.69×10^{-57}	1.11×10^{-54}
<i>Hes3</i>	5.80	8.30×10^{-56}	2.38×10^{-53}
<i>Hoxb5os</i>	2.44	5.05×10^{-53}	1.31×10^{-50}
<i>Cystm1</i>	5.98	6.21×10^{-51}	1.55×10^{-48}
<i>Fgf8</i>	4.61	2.26×10^{-46}	5.17×10^{-44}
<i>Etv4</i>	5.06	5.25×10^{-46}	1.16×10^{-43}
<i>Stmn2</i>	2.67	1.25×10^{-44}	2.48×10^{-42}
<i>Greb1</i>	2.44	4.29×10^{-36}	6.31×10^{-34}
<i>Acot7</i>	3.21	2.71×10^{-36}	6.78×10^{-34}
<i>Pdgfa</i>	3.00	8.76×10^{-35}	1.23×10^{-32}
<i>Hoxa7</i>	3.21	9.27×10^{-31}	9.46×10^{-29}
<i>Fgf17</i>	4.73	8.08×10^{-29}	7.76×10^{-27}
<i>Lix1</i>	3.55	1.59×10^{-28}	1.47×10^{-26}

Continued on next page

Single-cell census of mouse organogenesis

Table 4.8 – continued from previous page

Gene	Fold Change	p-value	
		Unadjusted	IHW adjusted
<i>Pmaip1</i>	8.44	1.92×10^{-28}	2.03×10^{-26}
<i>Gap43</i>	3.13	2.32×10^{-28}	4.14×10^{-26}
<i>Mgst1</i>	2.32	1.58×10^{-27}	1.56×10^{-25}
<i>Evx1os</i>	4.02	1.54×10^{-25}	1.30×10^{-23}
<i>Hoxa9</i>	3.75	2.64×10^{-25}	2.21×10^{-23}
<i>Hoxc6</i>	4.67	5.14×10^{-25}	4.14×10^{-23}
<i>Lhpp</i>	2.63	9.34×10^{-24}	7.09×10^{-22}
<i>T</i>	3.26	5.79×10^{-23}	4.34×10^{-21}
<i>Smim3</i>	3.01	4.88×10^{-22}	3.42×10^{-20}
<i>Hes7</i>	2.71	5.42×10^{-22}	3.73×10^{-20}
<i>Rasgrp2</i>	4.96	6.00×10^{-22}	4.09×10^{-20}
<i>Sox2</i>	2.80	2.43×10^{-21}	2.76×10^{-19}
<i>Hoxc4</i>	2.86	6.94×10^{-21}	4.46×10^{-19}
<i>Foxb1</i>	2.49	3.64×10^{-20}	2.27×10^{-18}
<i>Cyp26a1</i>	3.36	7.19×10^{-20}	4.30×10^{-18}
<i>Sp8</i>	3.92	9.70×10^{-20}	4.90×10^{-18}
<i>Hoxc9</i>	3.39	2.79×10^{-19}	1.63×10^{-17}
<i>Scg5</i>	3.55	2.91×10^{-18}	1.59×10^{-16}
<i>Tcea3</i>	3.96	5.14×10^{-18}	2.33×10^{-16}
<i>Wnt5b</i>	2.58	1.11×10^{-17}	5.62×10^{-16}
<i>Tpd52</i>	2.50	4.86×10^{-17}	2.07×10^{-15}
<i>Spint2</i>	2.40	6.08×10^{-17}	2.96×10^{-15}
<i>Ezr</i>	2.17	3.88×10^{-17}	3.26×10^{-15}
<i>Ccnjl</i>	2.29	1.54×10^{-16}	6.14×10^{-15}
<i>Fgfbp3</i>	2.60	1.15×10^{-15}	4.26×10^{-14}
<i>Epha1</i>	5.70	1.02×10^{-15}	7.54×10^{-14}
<i>Car2</i>	2.12	2.42×10^{-15}	8.67×10^{-14}

4.5 Cardiac divergence convergence

The cardiac pump must be one of the earliest organs to begin functioning so that the growing organism can meet demands of nutrient delivery and waste disposal as diffusion alone becomes inadequate.

The developing heart receives cell contributions from both the first and second heart fields. The first heart field is derived from the lateral plate mesoderm while the second heart field from the pharyngeal mesoderm. The tSNE in fig. 4.10 reconstructs these relationships simply from distances calculated by their transcriptome wide expression patterns. It is quite surprising that a naïve analysis can reveal this convergence back towards a cardiac fate from cells that have previously diverged.

To study this in more detail the mixed mesoderm, pharyngeal mesoderm and cardiac clusters were analysed separately from the remainder of the cells. There were 1316 cells within these groups and 7453 genomic features were found to be variably expressed. Dimensionality reduction was performed using DrL and separated the cardiac cells into 2 clear clusters, fig. 4.18. Louvain clustering clearly differentiates the two and on DrL it is apparent they are related to both the pharyngeal mesoderm, the source of the second heart field or the mixed mesoderm which includes the lateral plate mesoderm, the source of the primary heart field. *Nkx2-5*, *Mef2c*, *Smarcd3/Baf60c*, *Gata4*, *Tbx5* and *Tbx20* are known cardiac associated genes and are expressed in the expected region of the DrL, fig. 4.20 [Colombo et al., 2018; Kelly et al., 2014; Neshati et al., 2018; Ryan and Chin, 2003; Singh et al., 2005; Takeuchi and Bruneau, 2009; Vincentz Joshua W. et al., 2008].

It is tempting to consider the two clusters may represent the first and second heart fields but gene expression suggests otherwise. Particularly expression of *Irx4* suggests that one of the clusters consists of *Irx4*⁺ cells fated to become ventricular myocardium while the cells populating the other cardiac cluster which are almost exclusively *Irx4*⁻ form either atria or outflow tract [Bruneau et al., 2000; Nelson et al., 2014].

Having identified these 2 cardiac populations computationally we can in-silico compare gene expression between the ventricular fated *Irx4*⁺ and alternate fated *Irx4*⁻ cells. After multiple testing correction using IHW this identified 1087 differentially expressed genomic features, 22 are exclusively expressed in a subset of the putative ventricular cluster and 63 are transcription factors table 4.9. The top 20 up and down regulated genes are clustered on the heatmap in fig. 4.19. These genes represent possible candidates for differentiating between ventricular and other cardiac precursors as early as the 4 somite stage. In addition to *Irx4* which was not in the top 20 up-regulated genes, *Myl2* the ventricular isoform of

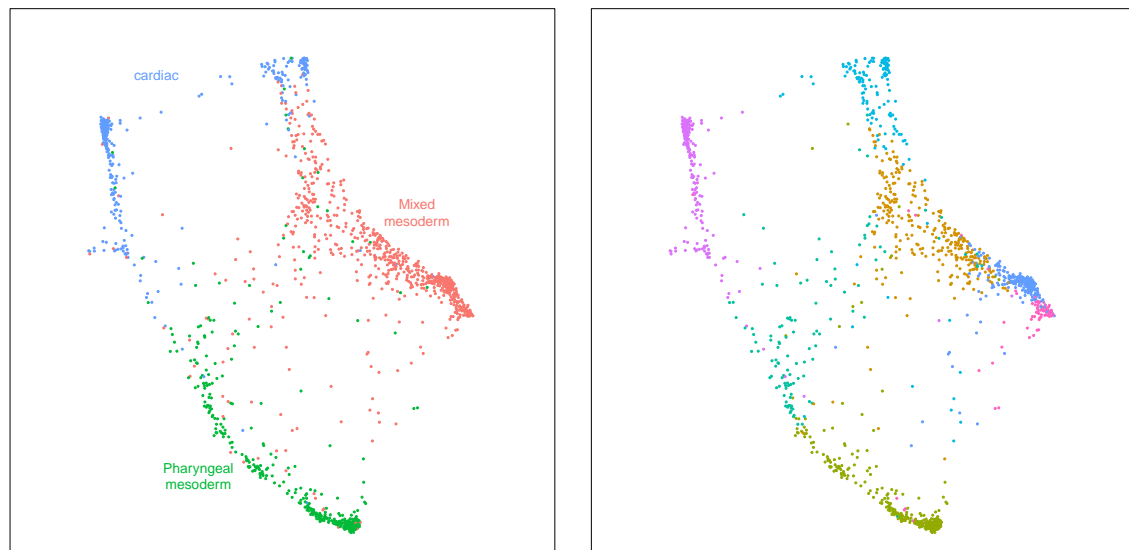


Fig. 4.18 DrL with cells coloured by previously defined clusters, table 4.4 in the left pattern. Cells coloured by newly calculated clustering in right panels shows cardiac cluster is split in to two distinct clusters.

myosin light chain was expressed 118 times higher (IHW adjusted $p = 1.2 \times 10^{-28}$) in the putative ventricular cluster as compared to the atrial/OFT cluster [Sheikh et al., 2015] further adding support that this cluster be assigned ventricular myocardium. Other Iroquois-related homeobox genes *Irx1* (fold change 34.9 IHW adjusted $p = 1.65 \times 10^{-9}$) and *Irx2* (fold change 21.2 IHW adjusted $p = 2.32 \times 10^{-6}$) were also significantly highly expressed in the ventricular myocardial cluster as has been previously described [Christoffels et al., 2000].

Bertrand et al. [2011] have shown that *Hoxb1* is down-stream of retinoic acid and marks a sub-domain of the second heart field contributing to the atria and the inferior wall of the OFT. Other genomic features differentially expressed between the putative ventricular and atria/OFT clusters for example the apparent differential expression of *Wnt*-related genes (*Sfrp5* and *Fzd7*) may suggest promising target pathways for directing differentiation towards different cardiac cellular components [Abu-Elmagd et al., 2017; Fujii et al., 2017]. The genes *Nr2f1* (*COUP-TFI*) and *Nr2f2* (*COUP-TFII*) are also known to be required for atrial induction [Devalla et al., 2015; Pereira et al., 1999].

Of the 22 genes exclusively expressed in the ventricular cluster *Klk1* has been linked with cardiac hypertrophy [Chao et al., 2010; Moreau et al., 2005], *Cacna1c* has been shown to be down-regulated in rat atria compared to ventricles and mutations have been associated with ventricular tachy-arrhythmias in humans [Bai et al., 2016; Hatano et al., 2006], and *Mitf* has been suggested to regulate cardiac growth and hypertrophy [Tshori et al., 2006].

4.5 Cardiac divergence convergence

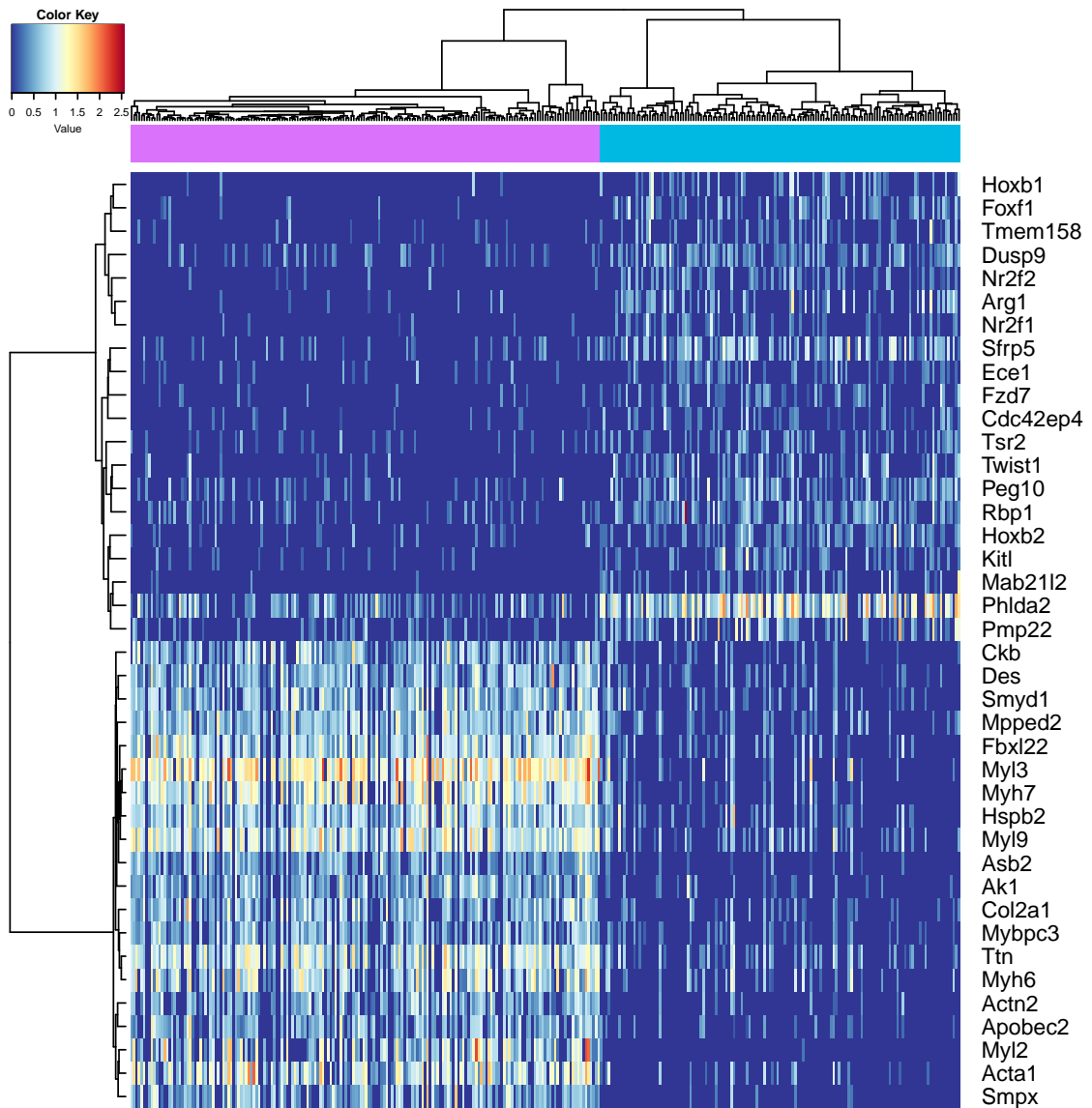


Fig. 4.19 Heatmap of genes expressed differentially between the putatively assigned *Irx4*⁺ ventricular myocytes and *Irx4*⁻ atrial/OFT cells.

Single-cell census of mouse organogenesis

Table 4.9 63 transcription factors identified to be differentially expressed between the two cardiac clusters identified using Louvain clustering fig. 4.18. Fold change is relative to the atria/OFT cluster. Unadjusted p-values and IHW adjusted p-values are given [Ignatiadis et al., 2016].

Gene	Fold Change	p-value	
		Unadjusted	IHW adjusted
<i>Hopx</i>	6.30	1.44×10^{-13}	1.52×10^{-12}
<i>Foxf1</i>	0.07	5.23×10^{-13}	5.12×10^{-12}
<i>Nr2f2</i>	0.08	1.50×10^{-11}	1.23×10^{-10}
<i>Hoxb2</i>	0.22	3.02×10^{-11}	2.41×10^{-10}
<i>Irx1</i>	34.87	2.50×10^{-10}	1.65×10^{-9}
<i>Irx4</i>	7.43	7.13×10^{-10}	4.43×10^{-9}
<i>Hoxb1</i>	0.06	3.31×10^{-9}	1.88×10^{-8}
<i>Twist1</i>	0.15	9.58×10^{-9}	4.92×10^{-8}
<i>Nr2f1</i>	0.09	1.54×10^{-8}	7.84×10^{-8}
<i>Etv5</i>	11.02	1.85×10^{-8}	9.14×10^{-8}
<i>Irx2</i>	21.23	5.89×10^{-7}	2.32×10^{-6}
<i>Hoxb4</i>	0.04	1.04×10^{-6}	3.85×10^{-6}
<i>Hoxa1</i>	0.14	1.10×10^{-6}	4.04×10^{-6}
<i>Tgif1</i>	0.31	2.79×10^{-6}	9.08×10^{-6}
<i>Rora</i>	7.78	4.29×10^{-6}	1.32×10^{-5}
<i>Zfp608</i>	4.70	1.42×10^{-5}	4.02×10^{-5}
<i>Hoxd1</i>	0.07	2.02×10^{-5}	5.57×10^{-5}
<i>Gli3</i>	0.26	2.13×10^{-5}	5.84×10^{-5}
<i>Alx1</i>	0.16	2.55×10^{-5}	6.91×10^{-5}
<i>Twist2</i>	0.16	2.70×10^{-5}	7.19×10^{-5}
<i>Tcf21</i>	0.11	2.71×10^{-5}	7.20×10^{-5}
<i>Snai2</i>	0.13	2.76×10^{-5}	7.22×10^{-5}
<i>Prox1</i>	5.19	2.83×10^{-5}	7.39×10^{-5}
<i>Stat3</i>	3.31	6.01×10^{-5}	1.46×10^{-4}
<i>Mitf</i>	∞	1.16×10^{-4}	2.62×10^{-4}
<i>Nfya</i>	0.18	3.57×10^{-4}	7.09×10^{-4}
<i>Hoxb5</i>	0.05	4.53×10^{-4}	8.56×10^{-4}
<i>Hey2</i>	3.81	6.85×10^{-4}	1.23×10^{-3}

Continued on next page

4.5 Cardiac divergence convergence

Table 4.9 – continued from previous page

Gene	Fold Change	p-value	
		Unadjusted	IHW adjusted
<i>Zfp9</i>	0.08	7.16×10^{-4}	1.28×10^{-3}
<i>L3mbtl3</i>	0.27	7.88×10^{-4}	1.38×10^{-3}
<i>Tfap4</i>	0.29	9.84×10^{-4}	1.66×10^{-3}
<i>Ets1</i>	0.04	1.10×10^{-3}	1.84×10^{-3}
<i>Zbtb44</i>	6.75	1.11×10^{-3}	1.86×10^{-3}
<i>Etv4</i>	5.43	1.63×10^{-3}	2.64×10^{-3}
<i>Tbx18</i>	0.12	1.89×10^{-3}	2.98×10^{-3}
<i>Bach2</i>	3.24	1.94×10^{-3}	3.04×10^{-3}
<i>Fosl2</i>	13.71	1.95×10^{-3}	3.05×10^{-3}
<i>Tox3</i>	0.22	2.04×10^{-3}	3.16×10^{-3}
<i>Rarb</i>	0.23	2.46×10^{-3}	3.71×10^{-3}
<i>Gli1</i>	0.10	2.52×10^{-3}	3.77×10^{-3}
<i>Capn15</i>	0.18	2.71×10^{-3}	4.04×10^{-3}
<i>Hoxa5</i>	0.16	3.76×10^{-3}	5.32×10^{-3}
<i>Zfp654</i>	5.98	3.87×10^{-3}	5.47×10^{-3}
<i>Snai1</i>	0.22	4.02×10^{-3}	5.62×10^{-3}
<i>Tshz1</i>	0.32	5.11×10^{-3}	6.93×10^{-3}
<i>Zfp105</i>	0.28	5.23×10^{-3}	7.07×10^{-3}
<i>Zbtb2</i>	0.20	7.24×10^{-3}	9.30×10^{-3}
<i>Sox6</i>	4.37	9.64×10^{-3}	1.20×10^{-2}
<i>Myc</i>	0.33	1.04×10^{-2}	1.27×10^{-2}
<i>Barx1</i>	4.95	1.08×10^{-2}	1.31×10^{-2}
<i>Ncoal</i>	4.37	1.40×10^{-2}	1.65×10^{-2}
<i>Zfp959</i>	0.27	1.69×10^{-2}	1.95×10^{-2}
<i>Hes7</i>	5.43	2.20×10^{-2}	2.46×10^{-2}
<i>Zfp618</i>	0.23	2.24×10^{-2}	2.48×10^{-2}
<i>Zbtb37</i>	0.30	2.98×10^{-2}	3.22×10^{-2}
<i>Creb3l1</i>	0.30	3.01×10^{-2}	3.24×10^{-2}
<i>Stat6</i>	3.52	3.36×10^{-2}	3.57×10^{-2}
<i>Tfeb</i>	3.35	3.40×10^{-2}	3.61×10^{-2}
<i>Sox13</i>	0.30	3.66×10^{-2}	3.86×10^{-2}

Continued on next page

Single-cell census of mouse organogenesis

Table 4.9 – continued from previous page

Gene	Fold Change	p-value	
		Unadjusted	IHW adjusted
<i>Zbtb46</i>	3.94	3.70×10^{-2}	3.88×10^{-2}
<i>Pou6f2</i>	3.61	3.75×10^{-2}	3.92×10^{-2}
<i>Zbtb26</i>	3.34	3.97×10^{-2}	4.13×10^{-2}
<i>Nr1dl</i>	3.27	4.01×10^{-2}	4.15×10^{-2}

At the E8.25, 4 somite pair stage, the cardiac crescent has formed from the first heart field and additional cells from the second heart field are gradually recruited. This suggests a possible continuity of cells from the second heart field subset of pharyngeal mesoderm towards the heart while those from the first heart field have likely already been specified at this stage. Assuming the DrL at least partly captures a differentiation process from the pharyngeal mesoderm toward the second heart field some of the genes appear to be expressed in a step-wise fashion towards both *Irx4*⁺ ventricular and *Irx4*⁻ cells with *Mef2c* and *Nkx2-5* expressed early while *Tbx5* and *Tbx20* are expressed later, fig. 4.20. Differentiation protocols that incorporate such ordered gene activation may therefore more faithfully recapitulate developmental processes by following such a step-wise approach rather than activating all genes concurrently [Takeuchi and Bruneau, 2009].

Using these known genes other genes with similar expression profiles may be identified using a guilty by association approach so that highly correlated genes may also be involved in cardiomyogenesis. This identifies multiple genes some previously known to be associated with cardiogenesis but others not previously implicated. Filtering the genes only for transcription factors with $p < 0.01$ identified 126 transcription factors positively correlated to the known cardiac genes and 144 transcription factors negatively correlated within this subset of cells in the 4 somite pair stage mouse embryo. The top 50 most highly positively correlated transcription factors are summarised in table 4.10. This table includes several genes known to play a critical role in cardiogenesis including *Mef2c*, *Cas2l*, *Gata5*, *Hopx*, *Irx4* and other Iroquois homeobox factors [Bonachea et al., 2014; Bruneau et al., 2000; Chen et al., 2002; Christoffels et al., 2000; Hempel et al., 2017; Huang et al., 2016; Nelson et al., 2014; Schneider et al., 2015; Vincentz Joshua W. et al., 2008].

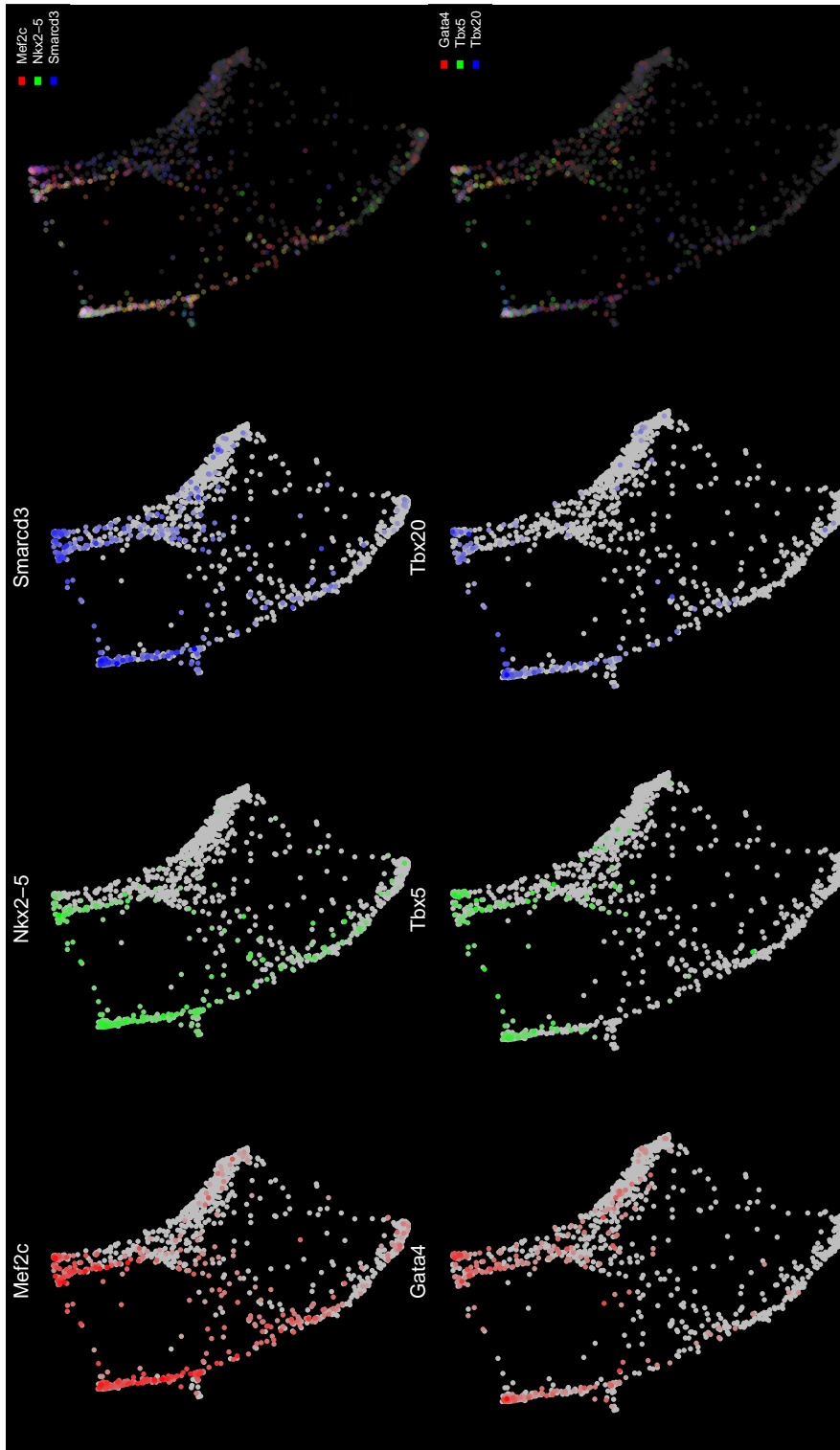


Fig. 4.20 Genes that are known to be highly expressed in cardiac cells. On the right composite of 3 genes to the left highlights patterns of co-expression for a maximum of 3 genes at a time.

Single-cell census of mouse organogenesis

Table 4.10 Top 50 transcription factors most significant positively correlated to *Nkx2-5*, *Smarcd3*, *Gata4*, *Tbx5* and *Tbx20* within the selected subset of cells fig. 4.18. Spearman's rank correlation was performed. Unadjusted p-values and IHW adjusted p-values are given [Ignatiadis et al., 2016].

Gene	Correlation	p-value	
		Unadjusted	IHW adjusted
<i>Mef2c</i>	0.69	7.97×10^{-185}	1.74×10^{-181}
<i>Casz1</i>	0.48	1.52×10^{-77}	7.28×10^{-75}
<i>Gata5</i>	0.41	6.47×10^{-54}	1.63×10^{-51}
<i>Hopx</i>	0.38	3.39×10^{-47}	6.38×10^{-45}
<i>Irx4</i>	0.38	8.89×10^{-46}	3.30×10^{-43}
<i>Gata6</i>	0.37	6.00×10^{-43}	8.12×10^{-41}
<i>Zbtb20</i>	0.36	5.29×10^{-42}	8.77×10^{-40}
<i>Plagl1</i>	0.32	3.97×10^{-32}	4.59×10^{-30}
<i>Tsc22d1</i>	0.29	9.93×10^{-28}	6.82×10^{-26}
<i>Fosb</i>	0.29	1.17×10^{-26}	1.07×10^{-24}
<i>Fos</i>	0.29	2.12×10^{-26}	1.14×10^{-24}
<i>Prox1</i>	0.29	1.11×10^{-26}	1.72×10^{-24}
<i>Esrrg</i>	0.28	3.77×10^{-25}	5.86×10^{-23}
<i>Prdm6</i>	0.24	1.64×10^{-18}	8.33×10^{-17}
<i>Smad6</i>	0.23	1.73×10^{-17}	9.91×10^{-16}
<i>Id2</i>	0.22	3.71×10^{-16}	1.27×10^{-14}
<i>Mef2d</i>	0.22	2.60×10^{-16}	2.18×10^{-14}
<i>Prrx2</i>	0.22	1.09×10^{-15}	3.45×10^{-14}
<i>Tead1</i>	0.20	7.59×10^{-14}	2.41×10^{-12}
<i>Hey2</i>	0.20	2.51×10^{-13}	1.62×10^{-11}
<i>Tbx2</i>	0.20	5.14×10^{-13}	4.40×10^{-11}
<i>Nr2f1</i>	0.19	2.36×10^{-12}	8.07×10^{-11}
<i>Stat3</i>	0.19	2.36×10^{-12}	8.29×10^{-11}
<i>Hes6</i>	0.19	3.33×10^{-12}	1.15×10^{-10}
<i>Barx1</i>	0.19	2.22×10^{-12}	2.02×10^{-10}
<i>Klf2</i>	0.19	9.68×10^{-12}	2.52×10^{-10}
<i>Creb3l2</i>	0.18	2.10×10^{-11}	6.61×10^{-10}
<i>Peg3</i>	0.18	4.66×10^{-11}	9.85×10^{-10}

Continued on next page

4.6 Transcriptional termination variants

Table 4.10 – continued from previous page

Gene	Correlation	p-value	
		Unadjusted	IHW adjusted
<i>Lrrfip1</i>	0.18	6.68×10^{-11}	1.55×10^{-9}
<i>Klf4</i>	0.18	7.27×10^{-11}	2.17×10^{-9}
<i>Cdip1</i>	0.17	3.09×10^{-10}	6.66×10^{-9}
<i>Jund</i>	0.17	5.02×10^{-10}	7.60×10^{-9}
<i>Arid3b</i>	0.17	4.75×10^{-10}	8.78×10^{-9}
<i>Pou6f2</i>	0.17	3.27×10^{-10}	2.02×10^{-8}
<i>Nfe2l2</i>	0.17	1.44×10^{-9}	3.60×10^{-8}
<i>Grhl1</i>	0.17	1.09×10^{-9}	6.84×10^{-8}
<i>Klf6</i>	0.16	6.31×10^{-9}	9.72×10^{-8}
<i>Tfeb</i>	0.17	1.69×10^{-9}	1.03×10^{-7}
<i>Irx2</i>	0.16	9.19×10^{-9}	2.06×10^{-7}
<i>Zim1</i>	0.16	8.30×10^{-9}	2.67×10^{-7}
<i>Zfpml</i>	0.16	1.21×10^{-8}	3.48×10^{-7}
<i>Hif1a</i>	0.15	4.17×10^{-8}	4.85×10^{-7}
<i>Hand1</i>	0.15	3.05×10^{-8}	4.95×10^{-7}
<i>Bhlhe40</i>	0.16	1.17×10^{-8}	5.78×10^{-7}
<i>Nr4a1</i>	0.15	5.53×10^{-8}	8.31×10^{-7}
<i>Esrra</i>	0.15	4.61×10^{-8}	1.21×10^{-6}
<i>Mafk</i>	0.15	6.50×10^{-8}	1.24×10^{-6}
<i>Pbx3</i>	0.15	7.03×10^{-8}	1.34×10^{-6}
<i>Rit1</i>	0.15	8.54×10^{-8}	1.61×10^{-6}
<i>Creb3</i>	0.14	1.41×10^{-7}	2.25×10^{-6}

4.6 Transcriptional termination variants

The single cell transcriptomic analysis has been performed on gene level data that is to say all gene isoforms are assumed to count towards the same genomic feature. Yet different gene isoforms can have differing, even opposing functions [Hakre et al., 2006]. Due to the low depth of single cell data even that produced by protocols such as Smart-seq2 discussed in section 3.2, isoform level data is difficult to infer accurately though this has been performed

Single-cell census of mouse organogenesis

for bulk RNAseq [Qiu et al., 2017a; Shen et al., 2014]. Additionally the 10x Genomics® Chromium™ protocol and most methods that incorporate UMIs to avoid bias from PCR duplicates can only capture one end of the mRNA molecule, usually the 3' end. This therefore precludes a complete splice variant analysis.

In the case of the current 10x Genomics® Chromium™ protocol, the mRNAs are captured on barcoded anchored oligo-dT at the 3' end of the molecule, so that the terminal part of the mRNA molecule is captured and so theoretically any variation of termination site may be determined. Figure 4.5a summarises the 10x Genomics® Chromium™ protocol and the sequences used have been previously given in table 3.2 and fig. 3.4; together these demonstrate where the cDNA insert fragment is located in the final library construct. The reads from the sequencer correspond to the 98bp along this fragment starting from the Read 2 end. The length of the library without including the insert length can be calculated as 183bp plus the insert length, table 4.11. From the bioanalyser trace of the final library we can calculate the expected length which is between 300 to 700 bp fig. 4.21. We therefore can deduce that the fragment will be 117 to 517 bp in length with only the first 98 bp being sequenced due to the sequencing protocol used. The library size distribution, keeping in mind that the abscissa is log-scaled, fig. 4.21, will to some extent determine the distribution of read pile-up and will be expected to be centered at ≈ 250 bp upstream of the annotated transcription termination site.

Table 4.11 Constituent parts of the 10x Genomics® Chromium™ library molecule and their sequence lengths in base pairs (bp).

Name	Sequence	Length (bp)
P7	CAAGCAGAAGACGGCATACGAGAT	25
10X cell barcode	NNNNNNNNNNNNNN	14
Read 2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	34
UMI	NNNNNNNNNN	10
Oligo-dT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	30
Insert	-	Variable
Read 1	CTGTCTCTTATACACATCTGACGCTGCCGACGA	33
Sample index	NNNNNNNN	8
P5	GTGTAGATCTCGGTGGTCGCCGTATCATT	29
Total		183 + Insert

A useful finding would be if for the same genomic feature multiple peaks were identified, suggesting variant transcription termination sites which may then be used as additional features and exploited in defining disparate cell types. Figure 4.22 shows that reads are

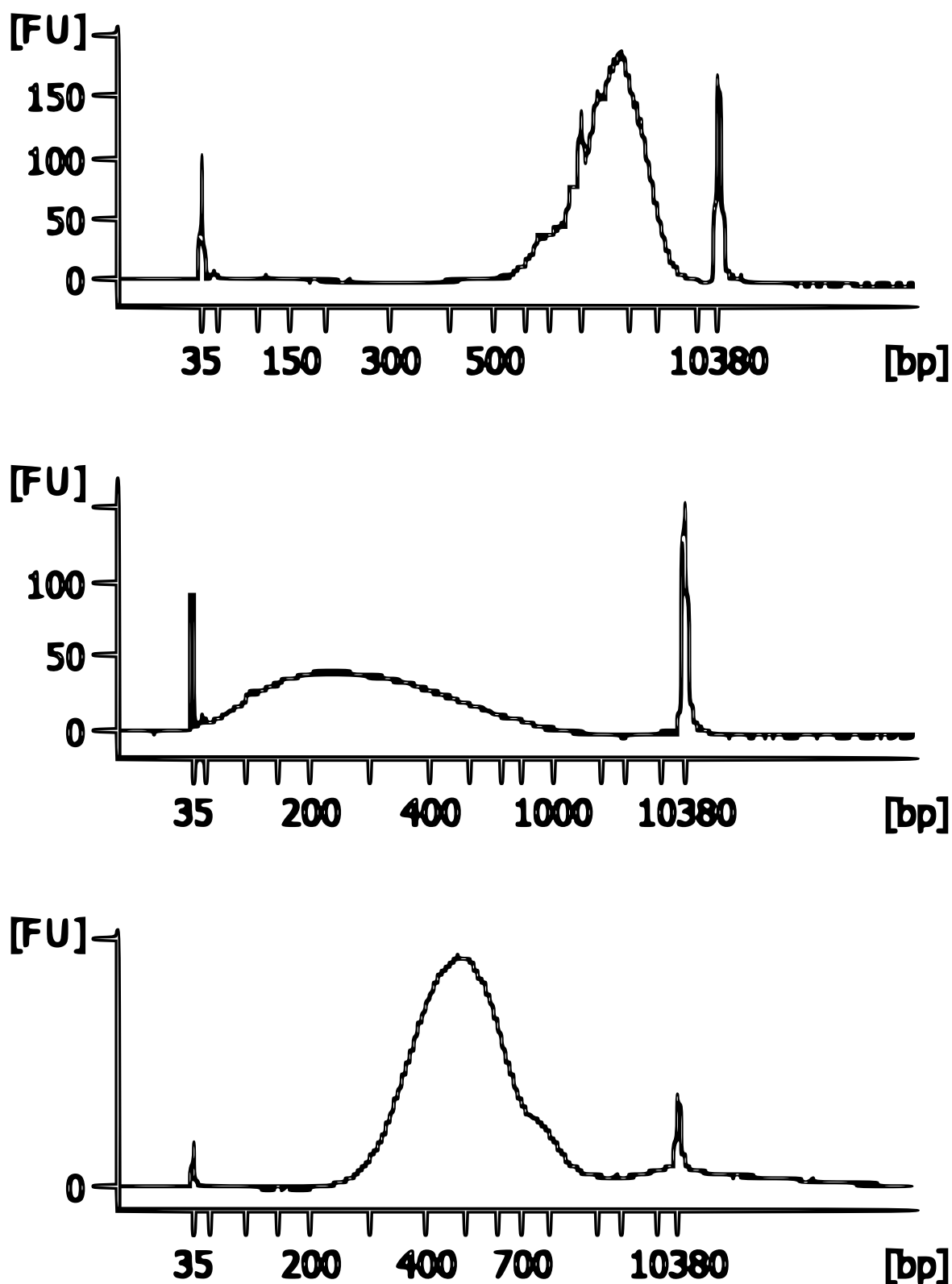


Fig. 4.21 Representative bioanalyser traces of the post cDNA amplification reaction QC (top), the post shearing fragments (middle) and of the final library (bottom) from the 10x Genomics® Chromium™ single cell protocol, demonstrating the final library size is between 300 to 700 base pairs (bp).

Single-cell census of mouse organogenesis

mapped not only to a single location but there are multiple peaks for each of the example genes shown. Each peak is ≈ 400 bp as was expected. Some reads are clearly distributed over splice sites see *Hoxc8* for example fig. 4.22.

This is not an exhaustive set. A systematic approach here would be useful to identify all such genomic features and then to re-annotate the GTF file so that they can be differentiated to see if they correspond to any biological effect. This has not been pursued here but is part of continuing work. Additionally it can be seen that peaks for some genes for example *Cdx4* and *Nkx-2* lie beyond the end of the annotation so these reads will have been incorrectly discarded when assigning reads to genes.

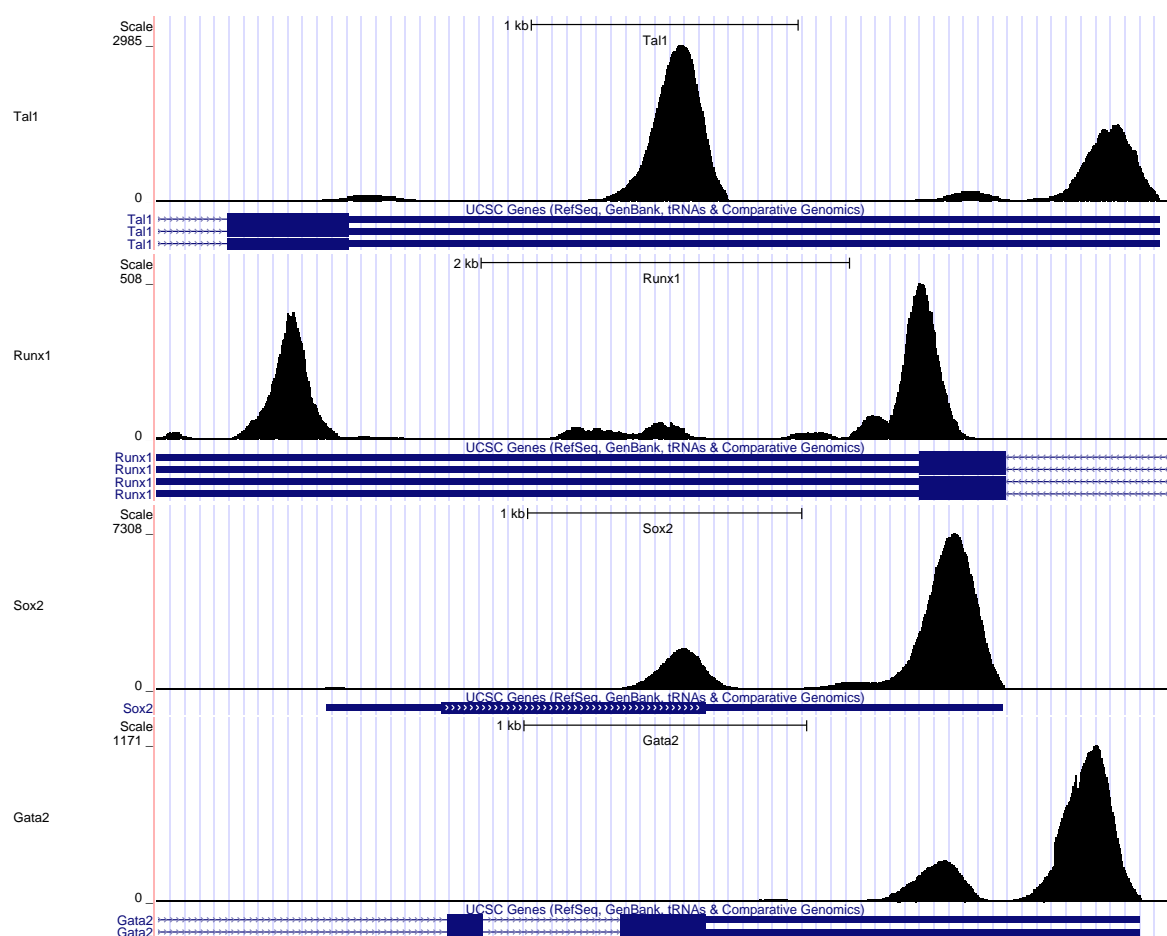


Fig. 4.22 Continued on next page ...

4.6 Transcriptional termination variants

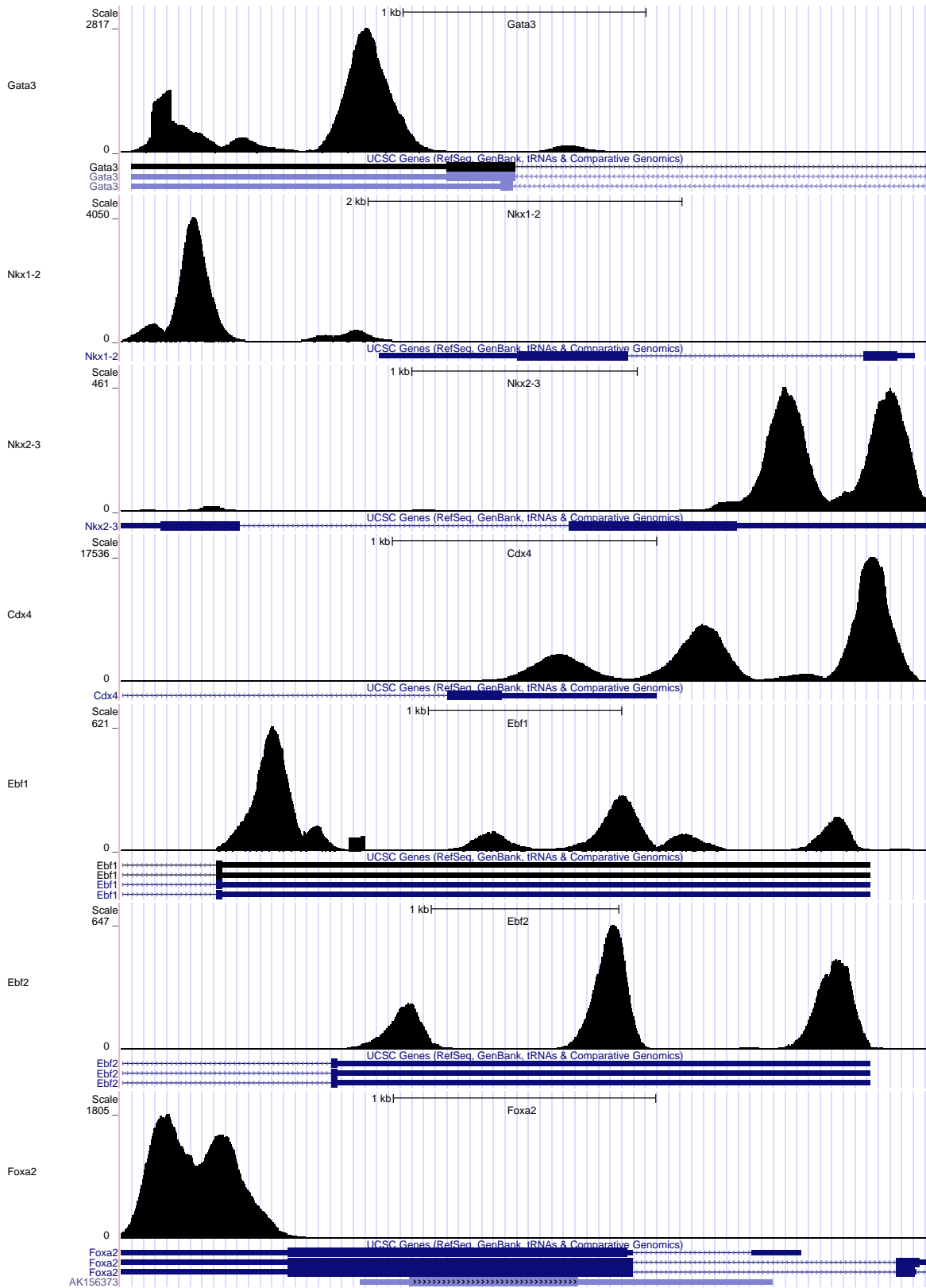


Fig. 4.22 Continued on next page ...

Single-cell census of mouse organogenesis

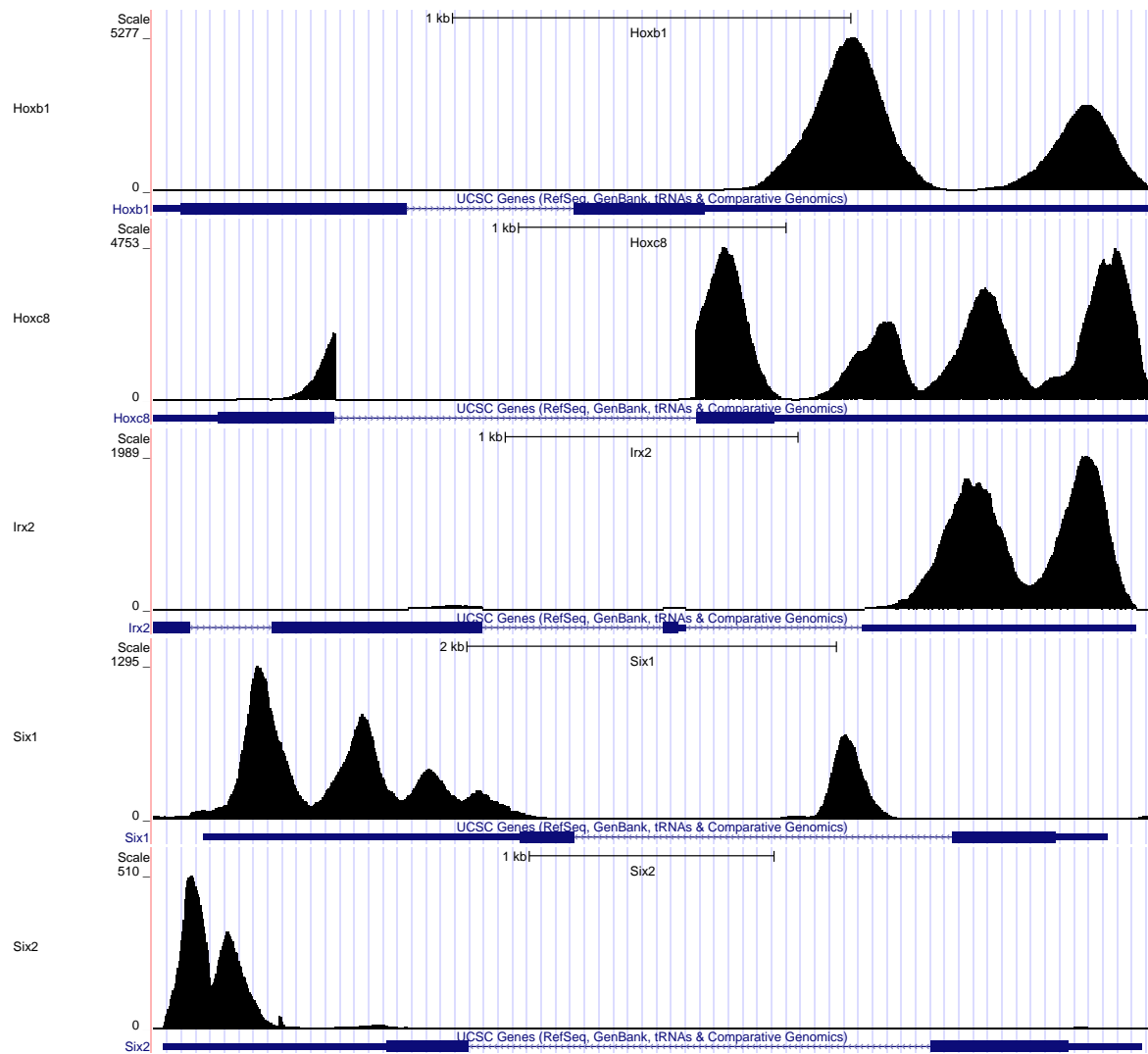


Fig. 4.22 Multiple peaks are found on the 3' untranslated regions of multiple genomic features. This suggests that the oligo-dT is hybridizing at different regions, so suggesting transcripts terminating at different genomic locations. Arrows on the gene marks indicate 5' to 3' direction.

4.7 Conclusions

The 10x Genomics® Chromium™ platform allows for many more cells to be assayed though sequencing depth is reduced and drop-outs become more severe of a problem. Despite these drawbacks the large number of cells appear to overcome the reduced signal to noise ratio per cell. The variety of cell types and large numbers of each type allowed for robust and confident cell assignment for many clusters in the 4 somite stage murine embryo, fig. 4.10.

The large dataset poses additional problems in terms of computational resource utilisation. The large datasets require large amounts of long-term storage but additionally create problems with memory as many calculations. Taking advantage of sparsity can help reduce the storage and short-term memory requirements. But even then calculation of a distance matrix is $\mathcal{O}(n^2)$ in memory and time for n cells. Again changing tact by using a similarity matrix and retaining only nearest neighbours can help reduce memory utilisation. Initial data analysis and normalisation was in this case performed in Python rather than R due to its favourable memory utilisation and better performance.

Feature selection is a key early step and the Brennecke et al. [2013] method has been used in the previous chapter. On the current dataset despite altering various parameters it was unable to give any reasonable fit to the data and so a linear fit was used and all genes with variation above the fitted line were included for downstream analyses. This concept of finding a relationship between gene expression and variability and then selecting genes that vary beyond a certain amount based on their mean is rather simplistic and rather naïve. Apparent gene variability can have many causes but biologically the most relevant genes are likely to be those that can discriminate between cell types.

A large part of this chapter has focused on clustering and sub-clustering cells. Many examples are given but in this large dataset many avenues of investigation remain unpursued. The *Pyy* expressing foregut domains identified by sub-clustering the definitive endoderm are clearly split on the DrL and on in-situ hybridisations at the corresponding embryonic stage figs. 4.13 and 4.15. Though not performed here, the in-silico recovery of these two sub-domains provides an opportunity to perform differential expression to identify novel genes that can differentiate between the two sub-domains.

Different aspects of the dataset can be appreciated using alternative analytic approaches. For example the original tSNE allows appreciation of convergence of previously divergent cell types back on to a cardiac myocyte cell type. Sub-setting the data, recalculating variable genes and plotting using DrL as implemented in the *roots* package presents a different appreciation of the data, now separating atrial and ventricular fated cardiomyocytes. Looking

Single-cell census of mouse organogenesis

for a single analytic technique that can concurrently display all biologically useful properties of a dataset given their complexities is likely to be futile and the more pragmatic approach of using a few approaches and then manually exploring the data with some background knowledge can often be quite revealing.

In summary this chapter has focused on a 10x Genomics® Chromium™ generated dataset and using multiple techniques have demonstrated how cell type identity may be assigned using an iterative approach. This method has been successfully used on a diverse set of cell types from the definitive endoderm to cardiac mesoderm. At the end of the chapter transcriptional termination variants are briefly introduced and it is demonstrated that the 3' 10x Genomics® Chromium™ protocol appears to capture this additional information that may be useful if collected systematically.

Chapter 5

Computational tools development

5.1 Background

High throughput and massively multiplexed single cell transcriptomic methods such as Smart-seq2 and droplet technologies have developed rapidly and as discussed in previous chapters pose several analytic challenges around the central problem of large sparse datasets with noisy measurements. Zero counts are a particular problem and can have multiple origins, examples include:

1. Gene is not expressed in the cell and no such mRNA moieties are present.
2. Gene is expressed
 - but no mRNA moieties are captured by hybridisation to the oligo-dT.
 - mRNA moieties are captured and amplified
 - but not sheared (either physically or by use of a transposome) to the desired length
 - and are sheared to the desired length
 - * but not captured on the flowcell to form an adequate cluster

Other than the first in this list the remainder produce false negatives. The probabilities of the events that cause these false negatives are clearly related to and dependent upon the underlying true expression level i.e. the number of mRNA moieties in the cell that may be captured at the initial step, assuming the remainder of the experimental conditions are optimised. This apparent increase in the number of zeros in datasets goes by several different

Computational tools development

names but in single cells is often called zero-inflation or drop-out. Several methods for inferring drop-outs have been adopted and applied to single-cell data but despite these inference methods zero-inflation effectively adds additional noise to the data.

Visualisation techniques commonly used to explore single-cell data include PCA, tSNE, multi-dimensional scaling (MDS), diffusion maps and SPRING [Andrews and Hemberg, 2018; Becht et al., 2018; Ding et al., 2018; Rostom et al., 2017; Taskesen and Reinders, 2016]. Applying even the same method on a given dataset but selecting different features can expose different aspects of the data. But as described in the previous chapter applying different dimensionality reduction techniques can help appreciate different aspects of the same dataset. In this chapter with view to simulating a continuous process of ontogenic progression an alternative visualisation technique has been developed and published as an R package. tSNE is commonly and successfully deployed in exploring single-cell data but currently no method exists to allow the user to add new data points to a pre-existing embedding and the latest Rtsne package is tweaked to allow this. Finally a novel method is developed to simulate developmental progression by training a deep neural network on plausible single cell transitions.

5.2 Reconstructing ordered ontogenic relationships

The R package *roots* was developed with view to reconstructing ordered ontogenic relationships from single cells data. Below follows a description, the algorithm and some examples of its use. This has been used in previous chapters but is discussed here in further detail.

5.2.1 Overview

Abstractly cells may be considered to be resident in a high-dimensional gene expression space. This simplistic view excludes other key properties for example their spacial positioning, cellular neighbourhood and cell to cell signalling, surrounding molecular signalling milieu and their epigenetic and proteomic landscapes. Distance between cells may then be considered to be some dissimilarity between their gene expression vectors. This allows a cell to cell similarity to be calculated for all cells and a graph to be generated. Since a distance and thus a similarity may be calculated between all cells a fully connected graph is created and can be used to generate a transition matrix whereby a cell may theoretically transition from one state to another with a probability inversely related to the distance.

5.2 Reconstructing ordered ontogenic relationships

Such graphs give rise to appreciable probabilities on non-plausible transitions. An example in 3-dimensional space may be that of travelling on Earth, represented as a circle in fig. 5.1. The problem can be seen in fig. 5.1a where only the starting and end points are sampled leaving the underlying surface or manifold unidentified. Any transition inferred between these two points by calculating a distance in Euclidean space would therefore be implausible and the true distance would be very different. Alternatively sampling more heavily in the region of interest would reveal the surface of the planet and so plausible interim states. This arc in a high-dimensional space would correspond to the underlying manifold and using these intermediate states would uncover a plausible and accurate journey without unrealistic short-cuts.

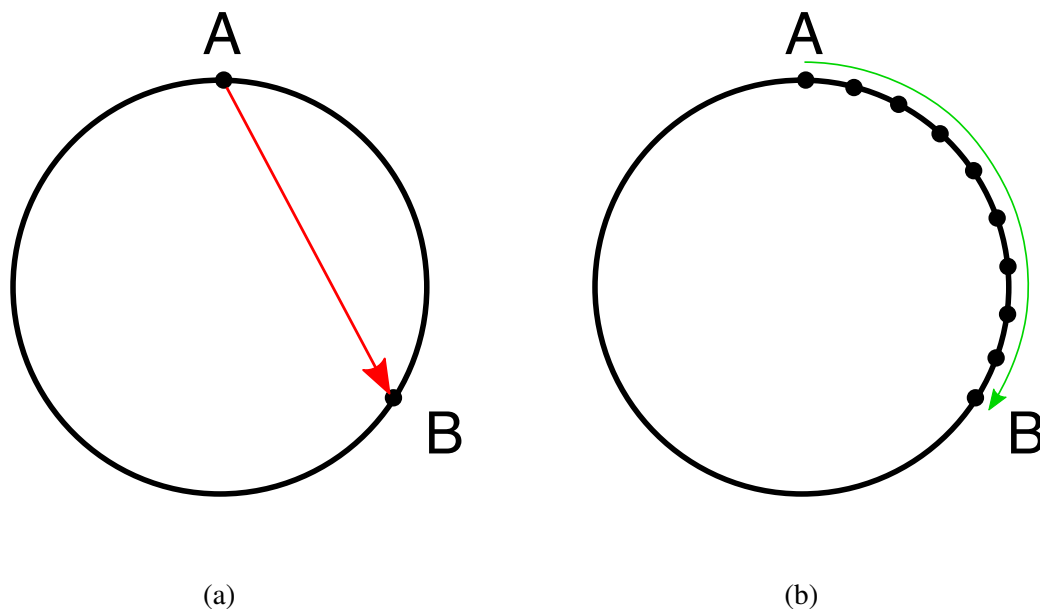


Fig. 5.1 (a) Low or minimal sampling strategy with only the initial and final cell states sampled does not provide sufficient information to infer the manifold and suggests a transition takes a cell through an implausible path, red arrow. (b) Increasing sampling density evenly across the unknown manifold and allowing only short transitions determines the underlying manifold and reveals a plausible journey, green arrow.

This partly highlights the importance of heavy sampling density to avoid implausible transitions. Sampling density is even more important when we consider that our measurements are noisy. Clearly where more fine grained ascertainment is required or dimensionality of the landscape and complexity of the underlying manifold increases the sampling density must be increased substantially. The minimum sampling density though is difficult to determine and is likely to be different at different parts of the manifold. Furthermore, simply extending this

Computational tools development

thought experiment it can be appreciated that sufficient sampling density will also depend on the signal to noise ratio of the measurements.

5.2.2 Algorithm

Normalisation

A data matrix $A \in \mathbb{R}^{n \times m}$ of n cells, C and m genes G is re-normalised using background gene expression. For each cell $c \in C$ a set of background genes, G'_c is defined as those that have counts less than a user-defined proportion α of the library size.

$$A_i^c < \alpha \sum_{g \in G} A_g^c, \quad \forall i \in G'_c \subset G$$

The background gene set G' is then the intersection between each cell's background gene set across all cells.

$$G' = \bigcap_{c \in C} G'_c$$

The background gene set G' is then considered to be a house-keeper gene heuristic. Cell counts are then normalised using this set G' and re-scaled by the mean library size to give the normalised counts matrix \bar{A} .

$$\bar{A}_g^c = \frac{A_g^c \sum_{g \in G, c \in C} A_g^c}{\sum_{i \in G'} A_i^c n}$$

Feature selection

Having normalised the data the genes are filtered by using hard lower bounds for mean gene expression μ for each gene across the whole dataset and the Fano factor $\frac{\sigma^2}{\mu}$, where σ^2 is the variance.

5.2 Reconstructing ordered ontogenic relationships

$$\mu_g = \frac{1}{n} \sum_{c \in C} \bar{A}_g^c$$

$$\sigma_g^2 = \frac{1}{n-1} \sum_{c \in C} (\bar{A}_g^c - \mu_g)^2$$

The hard bounds select a subset of genes $\hat{G} \subset G$ for downstream analysis, such that:

$$\mu_i > \beta \quad \text{and} \quad \frac{\sigma_i^2}{\mu_i} > \delta, \quad \forall i \in \hat{G} \subset G$$

No mean-Fano factor relationship is fitted as performed by Brennecke et al. [2013] but simple hard thresholds were used and selected manually by inspection resulting in a filtered normalised matrix $\hat{A}_g^c \in \mathbb{R}^{n \times m^*}$.

Orthogonalisation and distance calculation

Each feature or gene expression is mean centred and scaled.

$$\tilde{\mu}_g = \frac{1}{m^*} \sum_{c \in C} \hat{A}_g^c, \quad \tilde{\mu} \in \mathbb{R}^{m^*}$$

$$\tilde{\sigma}_g^2 = \frac{1}{m^* - 1} \sum_{c \in C} (\hat{A}_g^c - \tilde{\mu}_g)^2, \quad \tilde{\sigma} \in \mathbb{R}^{m^*}$$

$$\tilde{A} = \Sigma^{-\frac{1}{2}} (\hat{A} - h \cdot \tilde{\mu}_g^T), \quad \tilde{A} \in \mathbb{R}^{n \times m^*}, h \in \mathbb{R}^n$$

where $\Sigma_{i=j} = \tilde{\sigma}_i^2$ and $\Sigma_{i \neq j} = 0$, $\Sigma \in \mathbb{R}^{m^* \times m^*}$

and $h_c = 1$, $\forall c \in C$

PCA dimensionality reduction was performed retaining a user-defined l number of eigenvectors. To speed this up the large sparse matrix Arnoldi algorithm as implemented in rARPACK was utilised [Arnoldi, 1951]. Each cell was now represented by a vector of length l . Euclidean distances between all cell pairs were calculated in PCA space, generating a fully populated dense distance matrix $D \in \mathbb{R}^{n \times n}$.

Computational tools development

Gaussian kernel width estimation

A Gaussian kernel, L is calculated as in section 3.10.1 using the distance matrix, D . It is parameterised by σ_i . These are calculated as described by Haghverdi et al. [2016] and require a user selected k -nearest neighbour. The kernel width for each cell σ_i is then calculated accordingly.

$$\sigma_i = \sqrt[2]{\frac{D_{i,k}}{2}}$$

where $D_{i,k}$ represents the distance between cell i and its k -th nearest neighbour. The entries of L are then calculated as previously shown in section 3.10.1

$$L_{i,j} = \sqrt[2]{\frac{2\sigma_i\sigma_j}{\sigma_i^2 + \sigma_j^2}} \exp\left(-\frac{\|c_i - c_j\|^2}{2(\sigma_i^2 + \sigma_j^2)}\right), \quad \text{for } i \neq j,$$

and $L_{i,j} = 0,$ for $i = j$

where $c_i, c_j \in \mathbb{R}^n$ are column/row vectors of D . Now the entries of the matrix L may be considered un-normalised transition probabilities.

Transition matrix calculation

This is now normalised to give the Markov matrix, $\bar{L} \in \mathbb{R}^{n \times n}$ with entries:

$$\bar{L}_{i,j} = \frac{L_{i,j}}{\sum_j L_{i,j}}$$

All off diagonal entries in \bar{L} will be non-zero so giving a finite probability to some transitions between very different cells that are biologically implausible. To reduce this effect all transition probabilities below a threshold are set to zero. One method of setting such a threshold, ϵ , was to find the highest threshold that ensures all cells have at least one non-zero transition probability. Thus giving a sparse Markov matrix $\hat{L} \in \mathbb{R}^{n \times n}$.

5.2 Reconstructing ordered ontogenic relationships

$$\begin{aligned}
 \text{so that } \hat{L}_{i,j} &= 0, & \forall \bar{L}_{i,j} < \varepsilon \\
 \text{and } \hat{L}_{i,j} &= \bar{L}_{i,j}, & \forall \bar{L}_{i,j} \geq \varepsilon \\
 \varepsilon &= \min_i \left(\max_j (\bar{L}_{i,j}) \right)
 \end{aligned}$$

Another method could be to assume that one can reasonably select a maximum number of cells that a single cell within the sampled population should be able to transition to with equivalent probability, for example in our dataset this may be taken to be 10% of the population so that a threshold transition probability of $\approx \frac{1}{700}$ may be set. Such thresholds may be set iteratively after defining clusters so that cells from different clusters have different thresholds. This approach was not followed here.

Despite making the matrix sparse some spurious edges will remain and these are removed by assuming that if cell i can transition to cell j directly there must exist at least one second order transition between cells i and j . So that any transition $i \rightarrow j$ with no second order transition is regarded spurious and set to zero. To reduce computational complexity all second order transitions, $S \in \mathbb{R}^{n \times n}$ are calculated using the adjacency $A \in \mathbb{R}^{n \times n}$.

$$\begin{aligned}
 A_{i,j} &= 1, & \forall L_{i,j} > 0 \\
 A_{i,j} &= 0, & \forall L_{i,j} = 0 \\
 S &= A^2
 \end{aligned}$$

Now edges are pruned from the Markov transition matrix, $\hat{L}_{i,j}^{pruned} \in \mathbb{R}^{n \times n}$.

$$\begin{aligned}
 \hat{L}_{i,j}^{pruned} &= \hat{L}_{i,j}, & \forall S_{i,j} > 0 \\
 \hat{L}_{i,j}^{pruned} &= 0, & \forall S_{i,j} = 0
 \end{aligned}$$

Computational tools development

By pruning the transition matrix from all the cells C some will have no transitions and will form end points from which it is impossible to transition out. These cells were removed to leave a set $C' \in C$ where all cells have at least one potential transition possibility.

$$\sum_j \hat{L}_{i',j}^{pruned} > 0, \quad \forall i' \in C' \subset C$$

The probabilities will require re-normalisation to account for the transitions that have been disallowed in the previous two steps to give a new matrix $\tilde{L} \in \mathbb{R}^{|C'| \times |C'|}$.

$$\tilde{L}_{i',j'} = \frac{\hat{L}_{i',j'}^{pruned}}{\sum_{j'} \hat{L}_{i',j'}^{pruned}}, \quad \forall i', j' \in C' \subset C.$$

This matrix $\tilde{L} \in \mathbb{R}^{|C'| \times |C'|}$ is now taken as a representation of the underlying manifold of the cell expression space sampled.

Some downstream clustering and visualising algorithms though require a symmetric matrix as input. The Markov matrix was therefore symmetrised:

$$\tilde{L}_{i,j}^{sym} = \tilde{L}_{j,i}^{sym} = \max(\tilde{L}_{i,j}, \tilde{L}_{j,i})$$

Clustering

Clustering can now be performed by looking for modules within the graph representation of the cells. In this case Louvain modularity was used due to its advantages as described in Blondel et al. [2008]:

- Unsupervised requiring no selection of the number of clusters, intuitive and easy to implement.
- Extremely fast and has been used on hundreds of millions of nodes.
- Somewhat overcomes the resolution limit problem of modularity by initially starting with all cells separately and iteratively combining nodes so that intermediate results may be meaningful.

5.2 Reconstructing ordered ontogenic relationships

Louvain clustering is based on optimising modularity, Q defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[\tilde{L}_{i,j}^{sym} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \quad \text{where } i, j \in C'$$

$$m = \frac{1}{2} \sum_{i,j} \tilde{L}_{i,j}^{sym}$$

$$k_i = \sum_j \tilde{L}_{i,j}^{sym}$$

$$\delta(c_i, c_j) = \begin{cases} 1, & \text{if } c_i = c_j \\ 0, & \text{if } c_i \neq c_j \end{cases}$$

and c_i is the community to which cell i belongs. The algorithm as implemented in igraph was used [Csardi and Nepusz, 2006]. If instead of using the symmetric \tilde{L}^{sym} the normalised Markovian matrix \tilde{L} was used, k_i would be defined as 1 and $2m$ would be equivalent to the number of cells, this may allow for further optimisation as it will not need to be recalculated at each iteration.

Data visualisation

Data visualisation is a key step and allows exploration of the data and can allow the user to discriminate relationships between cells and clusters. Graph based visualisations provide a natural way of visualising the manifold \tilde{L} . The OpenOrd algorithm provides a means where all nodes repel each other and edges allow for nodes to be attracted to each other and is based on the Fruchterman-Reingold algorithm [Fruchterman and Reingold, 1991; Martin et al., 2011].

A 2-dimensional representation where each cell c_1, \dots, c_n is represented by a 2-tuple vector $(x_1^1, x_1^2), \dots, (x_n^1, x_n^2)$ where the superscript denotes the tuple index and the subscript represents the vector/cell index. The complete vector may then be denoted g and the position of each cell indexed using g_i . The optimum cell position in the reduced dimensionality space, g^* is then identified by solving the following:

$$g^* = \underset{g}{\operatorname{argmin}} \sum_i \left(\sum_j \left(\tilde{L}_{i,j}^{sym} d(g_i, g_j)^2 \right) + D_{x_i} \right)$$

$$d(g_i, g_j) = \|g_i - g_j\|^2$$

Computational tools development

where D_{x_i} is the density of points around x_i [Martin et al., 2011]. OpenOrd uses a greedy algorithm with a simulated annealing schedule. Additionally Martin et al. [2011] employ:

- Edge-cutting, to remove highly stressed edges,
- Parallelisation to speed-up graph generation and
- Multi-level graph layout, adopted from Hendrickson and Leland [1995]; Walshaw [2003] where a sequence of graphs is produced utilising a coarsening step and average-link clustering.

The OpenOrd implementation supplied with igraph where it is called ‘DrL’ was used [Csardi and Nepusz, 2006].

Code availability

The *roots* algorithm is implemented in R and freely available through CRAN (<https://cran.r-project.org/package=roots>) or GitHub (<https://github.com/wjawaid/roots>).

5.2.3 Usage and Examples

This algorithm has been used extensively to generate the plots produced in chapter 4, where Louvain clustering was used for the clustering shown on the tSNE in fig. 4.10 and several plots were generated with the *roots* in figs. 4.13, 4.16 to 4.18 and 4.20. Comparing DrL to tSNE when subclustering the definitive endoderm, fig. 4.13 shows clusters are much better defined and segregated in the DrL compared to tSNE. Though the clusters can be somewhat appreciated in tSNE, the result on DrL is much more apparent and aesthetically pleasing.

The definitive endoderm above is not the only example. The same is also true for the neural and mesodermal progenitors in fig. 4.16. Again the different clusters can be appreciated on the tSNE but are much more condensed and relationships more apparent on the neighbouring DrL. These spacial co-localisations allow correlation with known biological relationships supporting cell type assignment and suggesting ontogenic origins. Furthermore Louvain clustering on these population subsets after re-calculating variable genes appears to reveal deeper substructure.

Re-analysing the cardiac assigned cells using the new ‘*roots*’ algorithm, in section 4.5 revealed a different aspect of the data separating atrial and ventricular components. This was not apparent previously on the tSNE but by focusing in on this specific subset of cells a different aspect of the data could be appreciated resolving atrial and ventricular fated cell types.

5.2 Reconstructing ordered ontogenic relationships

To look beyond the current datasets, the *roots* algorithm was also applied to the adult murine haematopoietic system and compared to diffusion maps and a more recent and related visualisation technique called SPRING [Weinreb et al., 2018]. Bone marrow harvesting, cell sorting by FACS and library preparation were performed by Sonia Nestorowa and Dr Nicola Wilson; sequencing data was processed by Evangelia Diamanti and downstream analysis including generation of the SPRING map which was not part of the original publication was performed by Fiona Hamey.

Harvested bone marrow cells were sorted on lineage negative Lin⁻ and C-KIT⁺ and index sorted for FLK2, CD34, SCA1 and CD16/32 to allow retrospective population assignment [Nestorowa et al., 2016]. 3 main populations were sorted, the long-term HSCs (LT-HSC), haematopoietic stem and progenitor cells (HSPCs) and haematopoietic progenitors (Prog) as shown in fig. 5.2.

Data visualisation in Nestorowa et al. [2016] was predominantly performed using a diffusion map representation as shown in fig. 5.3. This displays the major subtypes with LT-HSCs positioned towards the upper apex, MPP and LMPP cells towards the rightward apex, granulocytes at the lower apex and erythroid cells and megakaryocytes positioned along the tail to the left fig. 5.3.

The SPRING layout was calculated by Fiona Hamey and is displayed in the lower left panel in fig. 5.3. This representation more clearly separates specific cell types into finger like projections from a central cloud, displaying better separation of the terminal cell types.

In comparison the *roots* representation calculated using the new algorithm provides an alternative more tree like representation ridding the layout of a disorganised central cloud. Furthermore the *roots* layout separates out the megakaryocyte and erythroid cells from the remainder. This large separation or gap between the cell types may be indicative of the sorting strategy omitting intervening cells between the progenitor SCA1⁻ and HSPC SCA1⁺ gates. The groups here were identified on the *roots* layout and were further characterised and assigned identity by Dr Joakim Dahlin. Not only does the *roots* representation appear to better segregate disparate cell types more effectively in addition it appears to better preserve ontogenic relationships with megakaryocytes not appearing to originate from some central cloud but there appears to be a common megakaryocyte erythroid precursor from which both arise, lower right panel of fig. 5.3.

An additional cluster of cells appears to condense just below the monocyte cluster with high and very specific expression of *Aif1* and *Spp1* indicating a monocyte/macrophage or possibly even a dendritic cell type fig. 5.3 [Elizondo et al., 2017].

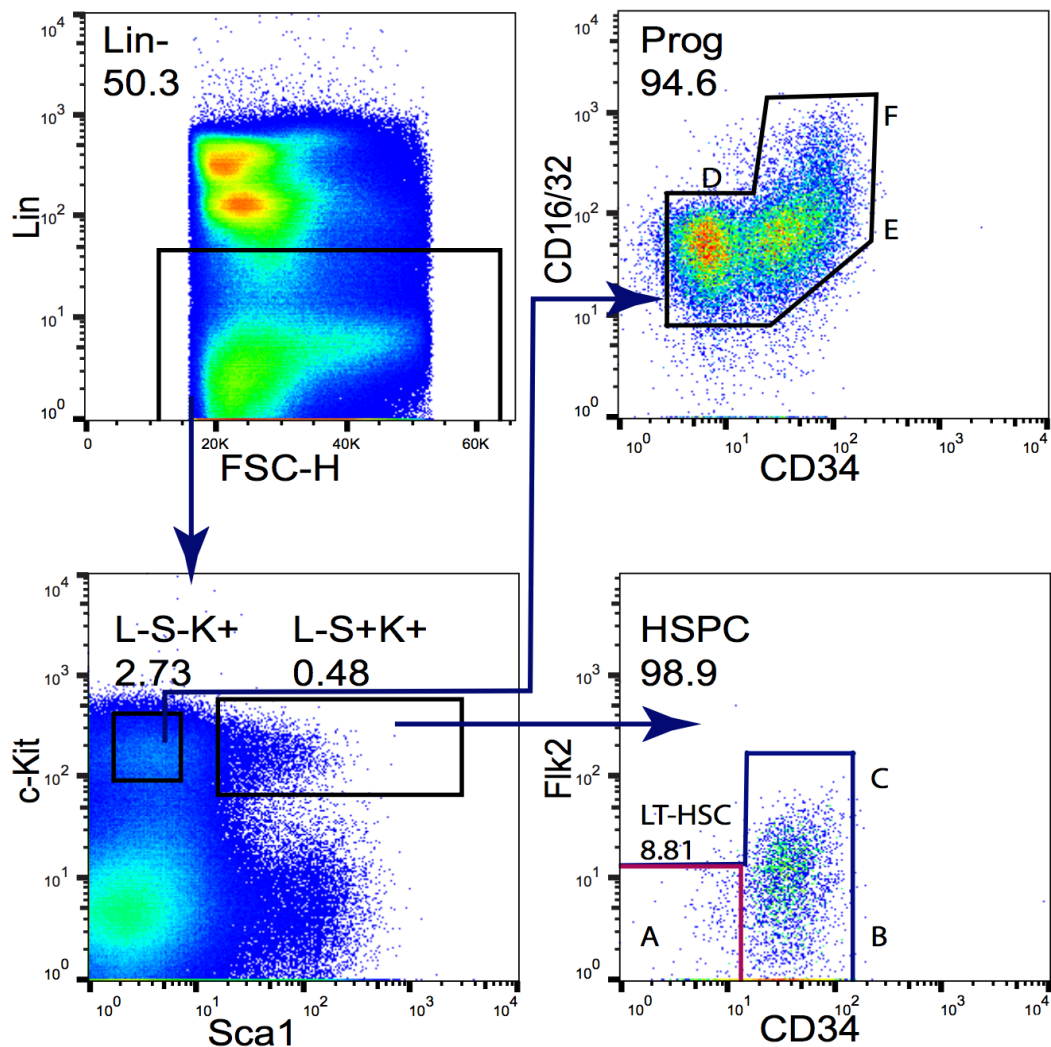


Fig. 5.2 Sorting strategy for adult haematopoiesis data, from Figure 1 in Nestorowa et al. [2016]. Post selecting cells and removing multiplets, the upper left panel displays cells were sorted selecting for Lin⁻ cells. These cells were subsequently sorted using 2 gates to select C-KIT⁺ cells that were either SCA1⁺ or SCA1⁻. Notably the SCA1 gates were separated so that intermediate cells remained unsampled. The Lin⁻SCA1⁻C-KIT⁺ progenitor cells were further indexed by CD16/32 and CD34 markers and Lin⁻SCA1⁺C-KIT⁺ were further indexed by CD34 and FLK2. To increase sampling of the rare LT-HSC population more cells were sorted from this gate, see Nestorowa et al. [2016] for further details. A, LT-HSCs; B, MPP; C, LMPP; D, MEP; E, CMP; F, GMP

5.2 Reconstructing ordered ontogenic relationships

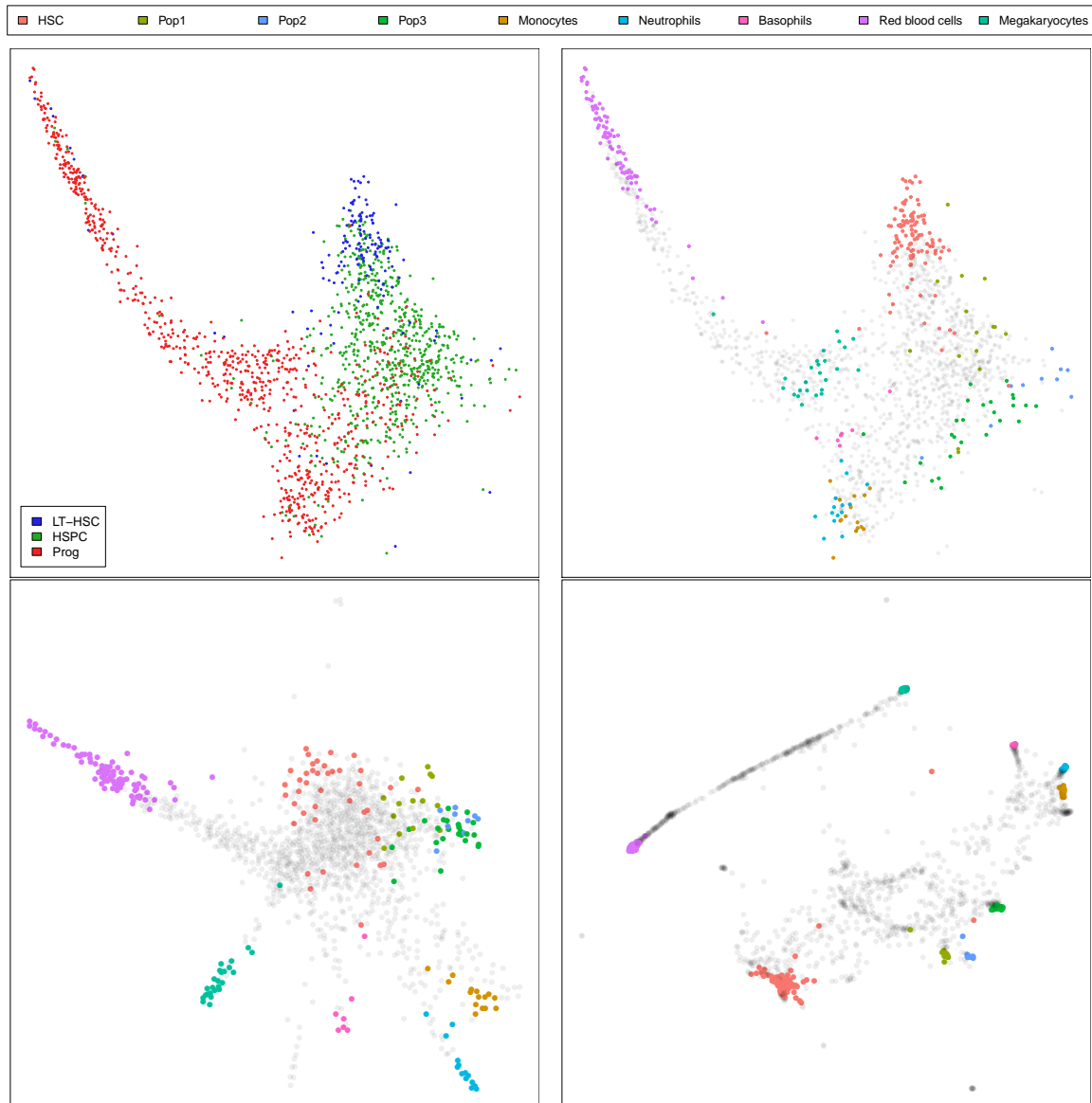


Fig. 5.3 Upper panels are a diffusion map representation of the data recalculated from data downloaded from the website of Nestorowa et al. [2016] (compare with figure 2 in Nestorowa et al. [2016]). The same populations are marked with the same colour as in the legend at the top in the remainder of the plots. The upper left figure is coloured according to the cell sorting strategy used, see fig. 5.2. The diffusion map to the right is coloured according to more specific cell characterisation performed retrospectively by Dr Joakim Dahlin as shown in legend panel above figures. Lower left is the SPRING representation of the same data calculated with parameter optimisation by Fiona Hamey. The lower right panel displays the *roots* visualisation showing a hierarchy that is more consistent with our current understanding of haematopoietic differentiation. Pop 1 to Pop 3 are lymphoid populations retrospectively identified by Dr Joakim Dahlin.

5.3 Projecting on tSNE

5.3.1 Background

Maaten and Hinton [2008] developed tSNE as a machine learning tool for dimensionality reduction with wide application but it has been widely adopted and become the de facto go to dimensionality reduction technique in single cell RNAseq [Becht et al., 2018; Briggs et al., 2017; Cao et al., 2017; Grün et al., 2015; Grün and van Oudenaarden, 2015; Habib et al., 2016; Ibarra-Soria et al., 2018; Jaitin et al., 2016; Lai et al., 2017; Scialdone et al., 2016; Taskesen and Reinders, 2016]. Given its popularity, several computational tools provide integrated use of tSNE including Seurat, Scanpy, Scater and a proprietary tool from 10x Genomics® called Cellranger [Butler et al., 2018; McCarthy et al., 2017; Wolf et al., 2018].

A particular disadvantage of tSNE as compared with PCA or diffusion maps is that there is no way of positioning new data points on a pre-constructed map. This ability if it existed would be very useful when trying to interpret new data given that a user may already have previous related data and may be familiar with a previous representation. Re-performing tSNE on the combined dataset depending on the number of new data points can drastically change the original layout.

To address this the tSNE algorithm has been adapted so that new points may be projected onto a pre-calculated embedding.

5.3.2 Algorithm

To explain the adaptation to the algorithm the tSNE algorithm as first described by Maaten and Hinton [2008] will be initially summarised.

Stochastic Neighbour Embedding (SNE)

Given a dataset of n cells C and m genomic features G , a matrix of log-transformed normalised gene expression values is given by $\mathcal{X} \in \mathbb{R}^{n \times m}$. The expression vector $x_i \in \mathbb{R}^m$ represents the gene expression vector for cell i so that $\mathcal{X}^T = \{x_1, x_2, \dots, x_n\}$. Similar to diffusion maps (see section 3.10.1) and the calculation of the first order Markov matrix in *roots* (see section 5.2.2), the conditional probability $p_{j|i}$, the probability that cell x_i chooses cell x_j as its neighbour is calculated by applying Gaussian probability density centered at x_i with variance σ_i^2 . The conditional probability is then given by:

$$p_{j|i} = \frac{\exp\left(-\|x_i^2 - x_j^2\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i^2 - x_k^2\|^2 / 2\sigma_i^2\right)}$$

A low-dimensional embedding $\mathcal{Y} \in \mathbb{R}^{n \times d}$ where in the remainder of this work $d = 2$ is then sought so that the conditional probabilities $q_{j|i}$ approximate $p_{j|i}$. In a similar fashion as for \mathcal{X} , each cell i is represented by the d -dimensional vector $y_i \in \mathbb{R}^d$. In the SNE algorithm originally proposed by Hinton and Roweis [2002] $q_{j|i}^*$ in the low dimensionality space is calculated in the same manner as $p_{j|i}$

$$q_{j|i}^* = \frac{\exp\left(-\|y_i^2 - y_j^2\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i^2 - y_k^2\|^2\right)}$$

The cost function C_o is then defined the Kullback-Leibner divergence between the two conditional probability distributions and the gradient $\frac{\partial C}{\partial y_i}$:

$$C_o = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad \text{and}$$

$$\frac{\partial C_o}{\partial y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}) (y_i - y_j)$$

Selection of σ_i is clearly critical in adjusting probabilities $p_{j|i}$ inducing a distribution P_i over all cells from cell i . Maaten and Hinton [2008] recognise that the variance must be set individually for each data point. This is achieved by a user selected perplexity $Perp(P_i)$ based on Shannon's entropy H , where:

$$Perp(P_i) = 2^{H(P_i)}, \text{ and}$$

$$H(P_i) = - \sum_j p_{j|i} \log_2 p_{j|i}$$

Computational tools development

A σ_i is then found that produces the user selected constant perplexity which can be interpreted as a smooth measure of the effective number of neighbours [Maaten and Hinton, 2008]. Adjusting perplexity can therefore affect the final visualisation with smaller perplexities breaking apart clusters into smaller units and higher perplexities leading to a more homogeneous crowded map.

t-Distributed Stochastic Neighbour Embedding (tSNE)

In tSNE, Maaten and Hinton [2008] make 2 changes to the algorithm: first they use a symmetrised probability distribution in both high and low-dimensional spaces and second they address the crowding problem by using Student's t-distribution with one degree of freedom, taking advantage of its thick tails in the low-dimensional space.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2},$$

$$q_{ij} = \frac{1 + \left(\|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} 1 + \left(\|y_k - y_l\|^2\right)^{-1}},$$

$$C_o = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

and the gradient is now given by:

$$\frac{\partial C_o}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij}) (y_i - y_j) \left(1 + \|y_i - y_j\|^2\right)^{-1}$$

Maaten and Hinton [2008] then use gradient descent with momentum to optimise the cost function. They have shown by using the thicker tailed t-distribution in the low dimensional space rather than a Gaussian allows for a more separated cluster structure.

Modification allowing projection on previously calculated embedding

Now a simple tweak can help project n^* new data points in a matrix $\mathcal{X}^* \in \mathbb{R}^{n^* \times m}$ and $\mathcal{X}^{*T} = \{x_1, x_2, \dots, x_{n^*}\}$ onto a pre-existing embedding $\mathcal{Y} \in \mathbb{R}^{n \times d}$ originally derived from a reference dataset $\mathcal{X} \in \mathbb{R}^{n \times m}$. Now for all cells $c \in C$ in the reference dataset forming the original embedding:

$$\frac{\partial \mathcal{C}_o}{\partial y_{c \in C}} \triangleq 0$$

Positions for the new set of cells, C^* , must now be calculated. For all cells indexed $i \in C^*$ and $j \in C \cup C^*$ the gradient is calculated in the usual method as described above. This ensures that at each update only the newly added data points are allowed to move on the low-dimensionality embedding and optimised

$$\frac{\partial \mathcal{C}_o}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij}) (y_i - y_j) \left(1 + \|y_i - y_j\|^2\right)^{-1}, \quad \forall i \in C^*, j \in C \cup C^*$$

This allows the newly added cells to somewhat influence the landscape when other cells are projected as the Kullback-Leibner divergence is calculated against $j \in C \cup C^*$. If instead the Kullback-Leibner divergence was calculated solely against $j \in C$ then the new cells will have no influence and the algorithm will be more consistent with a projection as available for PCA and diffusion maps.

Implementation

The modified algorithm was implemented by adapting the most up-to-date Rtsne Git repository from GitHub (<https://github.com/jkrijthe/Rtsne>) [Krijthe, 2015]. The implementation by [Krijthe, 2015] uses the Barnes-Hut algorithm for optimisation [Maaten, 2014]. The Barnes-Hut algorithm is not used when projecting. The modified algorithm is available through GitHub at <https://github.com/wjawaid/Rtsne>.

5.3.3 Examples

The algorithm was initially tested for convenience using the ‘iris’ dataset within R, introduced by Ronald Fisher in 1936 containing 3 plant species and four features measured for each sample, an excerpt is given in table 5.1 [Fisher, 1936].

tSNE dimensionality reduction was performed on the iris dataset and is shown in fig. 5.4. This shows clear separation of setosa species from versicolor and virginica. But the 3 species are also clearly separated from each other. Means were calculated for all features in each species and these were used as new data points and projected onto the pre-existing embedding. The new data points outlined in black are also presented on fig. 5.4. It is immediately evident

Computational tools development

Table 5.1 An excerpt from the iris data showing the features collected in the different species [Fisher, 1936].

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica

that the original data points are in exactly the same position as before. Additionally it is easy to appreciate that the newly calculated means project centrally within their corresponding species on the tSNE, giving confidence that the algorithm works well at least for this small data set.

Table 5.2 Average (mean) features for the 3 species used as data points to project onto the preexisting dataset on fig. 5.4.

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
setosa	5.01	3.43	1.46	0.25
versicolor	5.94	2.77	4.26	1.33
virginica	6.60	2.98	5.56	2.03

The algorithm was applied to a single cell dataset discussed in an earlier chapter. In fig. 3.61 *Brachyury* cells from an intermediate time point mapped between the E6.5 cluster and the E7.25 nascent mesoderm cluster. This allowed identification of node or organiser-like cells within the *Brachyury* dataset. Additionally the projection does not appear to be affected by cell cycle nor embryo batch figs. 3.61 and 3.63b.

Having projected onto a pre-existing embedding, another utility for this technique may be to use prior knowledge to inform an embedding. One way to test this using the Iris data is to posi-

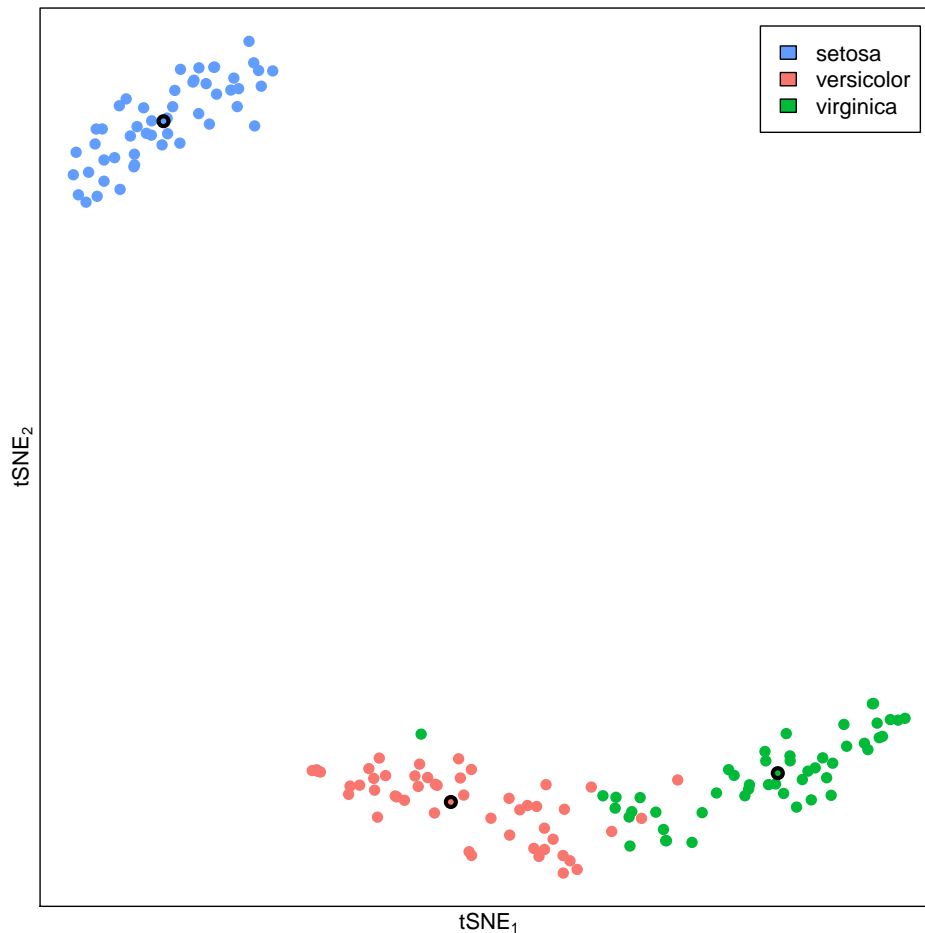


Fig. 5.4 tSNE representation of iris data with 3 species in different colours as indicated by the legend. The averages as shown in table 5.2 are then projected on to the pre-existing embedding and have been highlighted with a black outline. The average data points appear to be well centred over the corresponding species, suggesting the algorithm is performing well on this relatively small dataset.

tion the 3 average values as a user may like to position them. For example in this case all three average data points may be placed on the x -axis at co-ordinates $(-15, 0)$, $(0, 0)$ and $(15, 0)$ for setosa, versicolor and virginica respectively. Then given this layout, to position the individual data points on to this preconditioned landscape using the projection adaptation to the tSNE, fig. 5.5. This shows that this preconditioning of the landscape can help arrange the new data in a way that may be more meaningful to the user.

In fig. 5.5, it is an average of iris sepal features for each species that was used to precondition the tSNE landscape but for single cell RNAseq this could be for example bulk data with known spacial localisation. Building on this, data from Peng et al. [2016] was downloaded from the GEO repository.

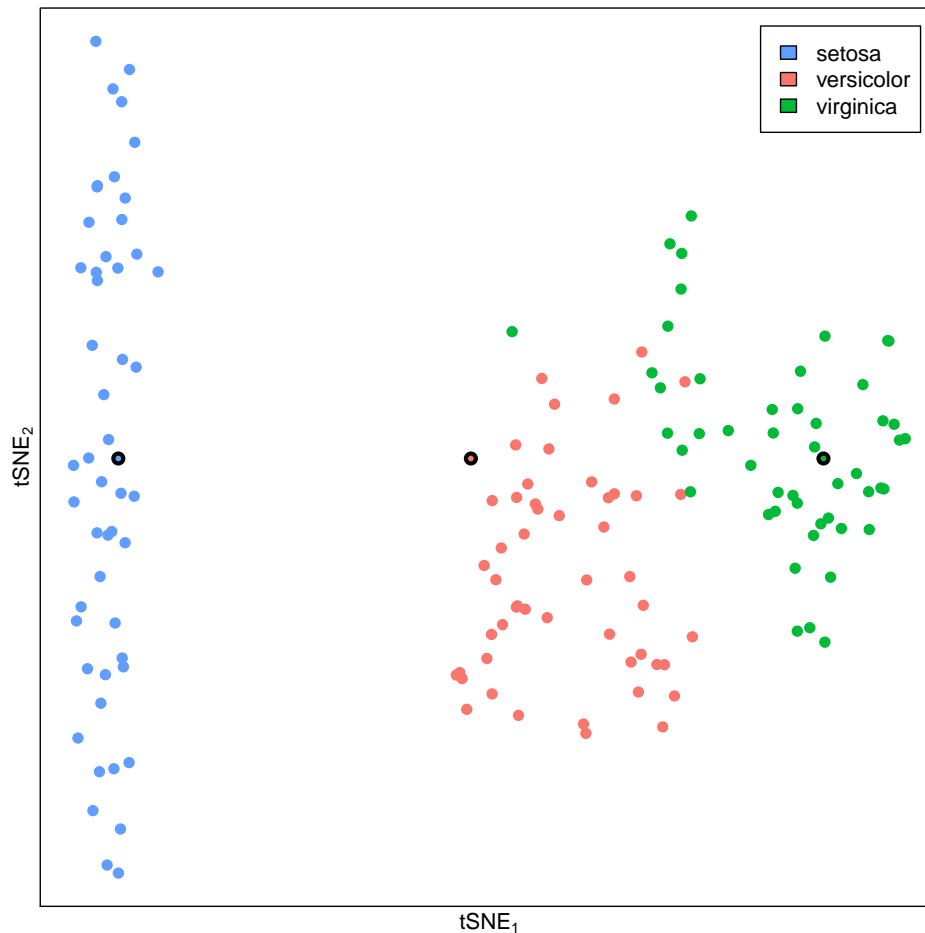


Fig. 5.5 Generating a prior information based pre-existing embedding can help drive the tSNE dimensionality reduction in an informed manner. This allows prior information to be incorporated into the tSNE.

Peng et al. [2016] aimed to identify spacially resolved transcriptomes for mouse embryos at mid-gastrulation. They harvested 3 mouse embryos at E7.0 and embedded each in Optimal Cutting Temperature (OCT) in a known orientation. They then sectioned the embryos precisely using a cryostat microtome into 22 sections with alternate sections being used for either imaging or RNAseq analysis. Laser capture micro-dissection was used to capture ≈ 20 cells from 4 quadrants in 10 of the more proximal sections and from 2 segments in the most distal section, fig. 5.6. During the data collection they managed to get samples from 41 of the potential 42 segments for the reference embryo, Embryo 1 in their dataset. These segments were subsequently processed using a low input RNAseq protocol. The missing segment from the anterior in slice 3 was inferred by using the mean between the anterior segments of slices 2 and 4. They present their data using corn plots where each dot represents a segment and colour represents expression level of a selected gene fig. 5.6.

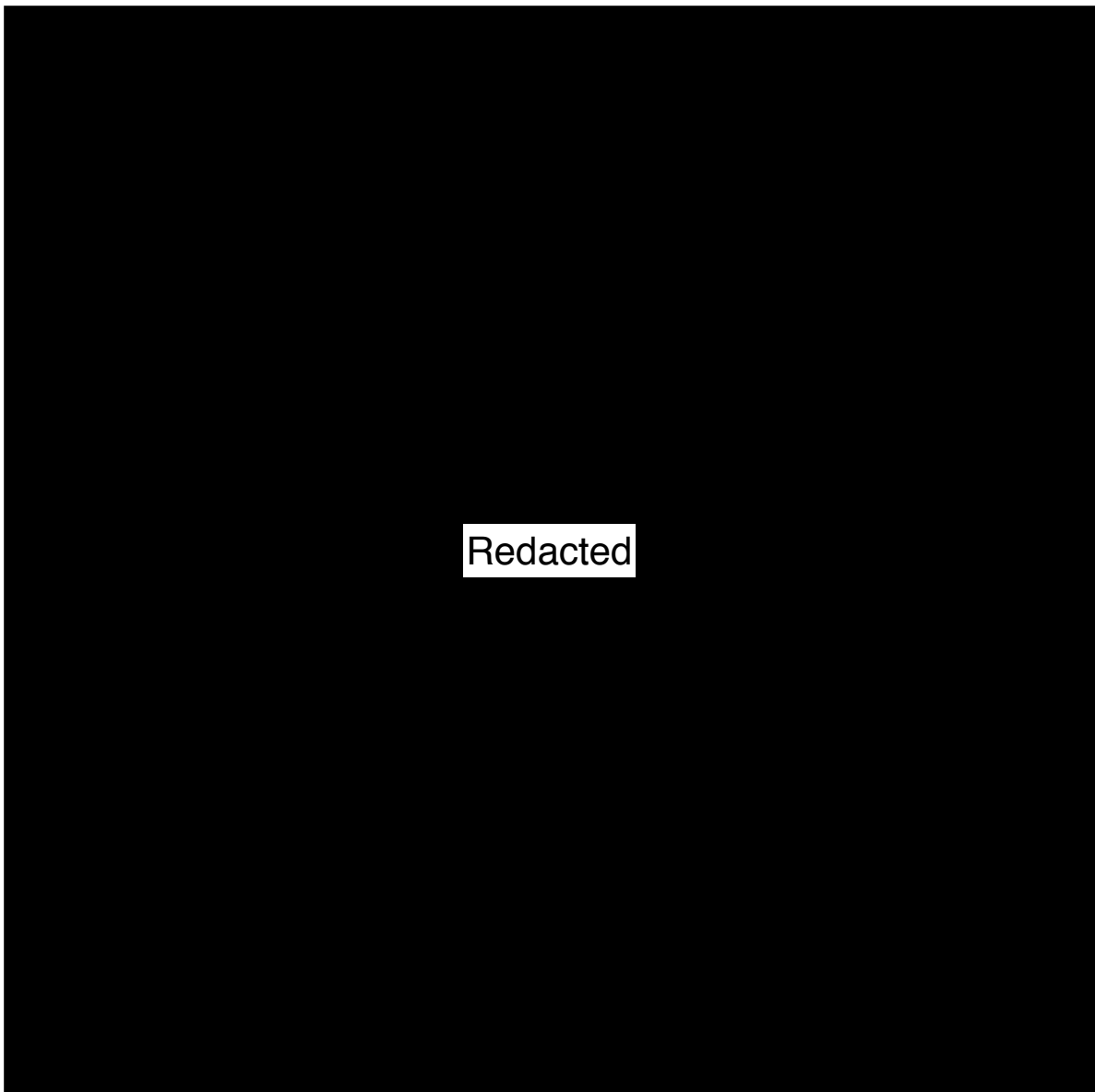


Fig. 5.6 Experimental protocol used by Peng et al. [2016]. Mid-gastrulation (E7.0) mouse embryos were harvested, embedded in OCT and sectioned as displayed in the figure. Alternate sections were used for sampling and RNAseq analysis. Laser capture micro-dissection was used to harvest cells from four quadrants from each slice except the most distal: anterior, A; posterior, P; right, R and left, L. For the most distal slice (S1) only anterior and posterior segments were captured. Each segment consisted of ≈ 20 cells. Figure reproduced from Peng et al. [2016].

Additionally Peng et al. [2016] harvested 70 individual cells from either the posterior or anterior aspect of the embryo. This now provides a potential test data set for the modified Rtsne algorithm to try and computationally post-hoc reconstruct spatial localisation of the

Computational tools development

single cells. Peng et al. [2016] present gene expression data in a corn plots as shown in the lower right panel in fig. 5.6.

In the first instance, the new single cell data points were projected onto the corn plot representation of bulk data, fig. 5.7. This was performed by pre-conditioning the tSNE landscape with the spacially resolved transcriptomic data and then using the new tSNE projection algorithm from section 5.3.2 to map the single cells onto the pre-conditioned landscape. The red and black points have a preponderance towards either the posterior or anterior part of the embryo, respectively. Though there exists this preponderance the spatial reconstruction seems poor with many cells located between the left and right segments. This is unsurprising as the corn plot representation of the data was to allow visualisation of 3D data on a 2D page and so produces non-metric distances for which this tSNE algorithm cannot account.

There is an alternative modification of tSNE that attempts to manage non-metric distances by producing multiple maps with different weightings for data points but this is unnecessarily complicated and counter-intuitive for the current purpose [Cook et al., 2007; van der Maaten and Hinton, 2012]. A more intuitive approach is to simply expand the low-dimensionality space to 3 dimensions so that the right to left, anterior to posterior and proximal to distal axis are all orthogonal to one another fig. 5.8. Now the red and black points are more clearly separated and the Euclidean distance used in the tSNE algorithm for measuring relative distances between the single cells, in the low dimensionality space better representative of the spatial data.

Peng et al. [2016] also perform a spatial reconstruction but to do so they first go through several steps to define a set of ‘zip-code’ genes. This gene set is used to calculate rank correlation coefficients and a smoothing process using rank correlations of adjoining regions. In the tSNE projection proposed here only those genes that are unexpressed have been filtered and the remaining genes used to calculate the projection in an unbiased way. The tSNE projection may therefore be further improved by selecting only informative genes.

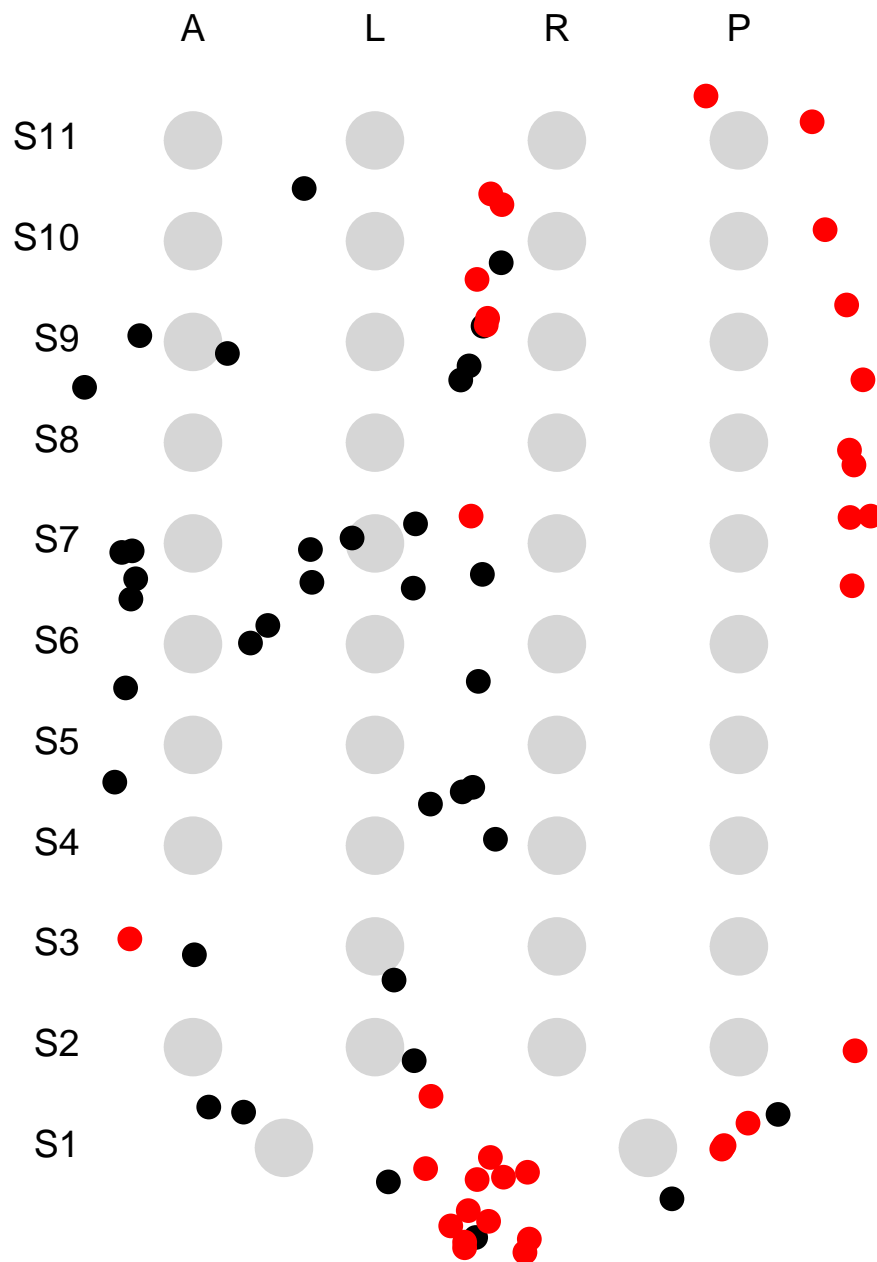


Fig. 5.7 Corn plot with single cell data from Peng et al. [2016] projected on tSNE. The large grey dots represent the regions of the embryo with spatially resolved oligocellular transcriptomic data that were used to pre-condition the tSNE landscape. Red dots represent cells harvested from the posterior and black points from the anterior of mid-gastrulation mouse embryos. This corn plot has non-metric distances in the 2D representation of the 3D data with right and left locations sandwiched between anterior and posterior. This makes it impossible to accurately position a cell based on mutual distances. Above the plot labels represent A, anterior; P, posterior; R, right and L, left.

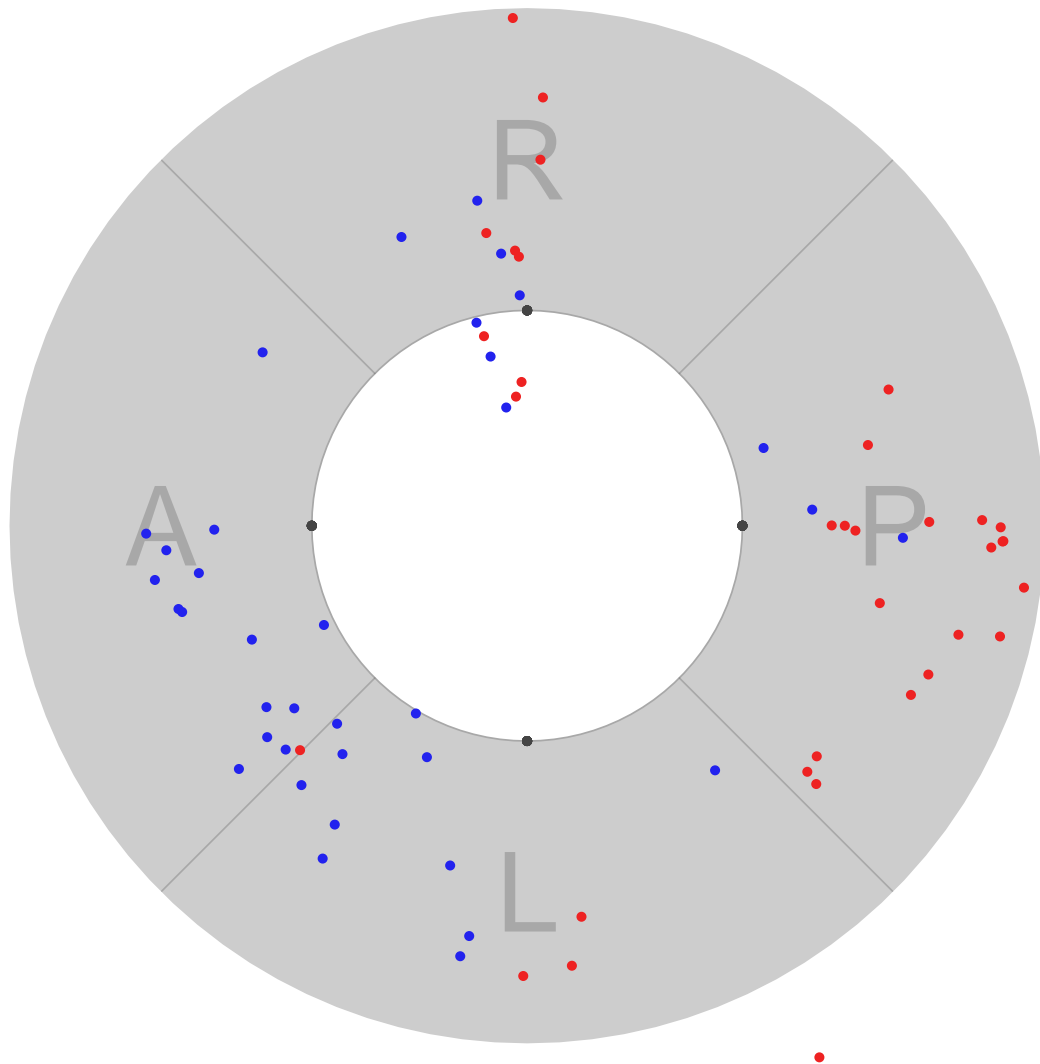


Fig. 5.8 3D positioning of single cell data using the novel tSNE algorithm on a landscape preconditioned using bulk oligocellular data. The 3D representation gives rise to distances in the low dimensional space that are congruent with the known spacial localisation of the bulk data in a single mapping. Red dots are gene expression profiles from cells harvested from the posterior and blue points profiles of cells from the anterior of mid-gastrulation mouse embryos The grey points represent the reference bulk data.

5.4 Simulating developmental processes

A chief motivation for reconstructing ontogenic trajectories is so that gene expression changes can be determined to allow gene interactions to be inferred. This may then allow identification of key driver genes that play central roles in directing cell fate. These observations may then be incorporated into differentiation protocols to achieve more faithful production of cells for potential cellular therapies.

To achieve this a reliable developmental path must be inferred for subsequent analysis before trying to simulate along such a pathway. A previously generated data set of early in-vivo haematopoiesis with a bifurcation event towards blood and endothelium has been exploited. Moignard et al. [2015] used qPCR to characterise single-cells sorted on RUNX1 and FLK1 in mouse embryos at four time points between E7 and E8.25, fig. 5.9. The data is similar to that introduced in section 3.2 except here only 42 manually selected gene features were assessed.

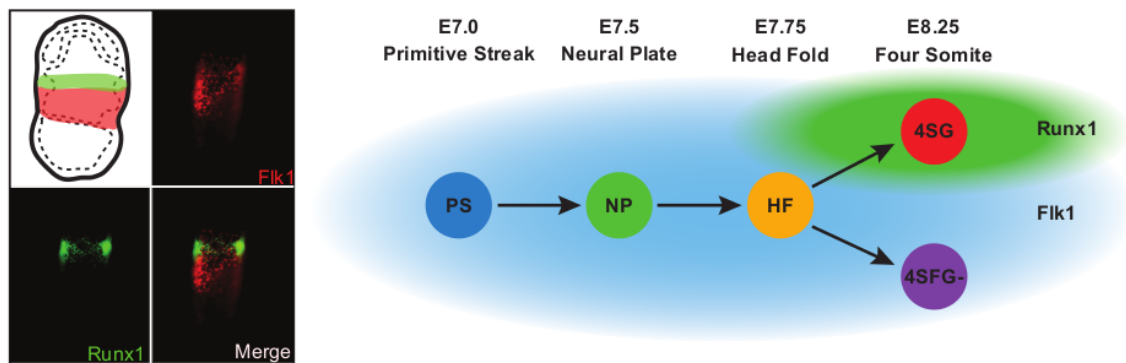


Fig. 5.9 Experimental protocol for Moignard et al. [2015] data set. Single cells were sorted from four time points between E7 and E8.25 from mouse embryos for FLK1 which was used to capture the early multipotent cells and RUNX1 to capture later blood fated cells which down-regulate FLK1. A bifurcation between endothelial and blood fated cells should be captured using this protocol.

The gene features were manually selecting for genes that are known or postulated to be involved in yolk sac haematopoiesis and endothelial development. This data was collected and processed by Dr Victoria Moignard, see Moignard et al. [2015] for further details.

Moignard et al. [2015] visualise the qPCR data using diffusion components 1, 2 and 4, see fig. 5.10a. This shows computational reconstruction of a developmental trajectory from early FLK1 primitive streak cells towards endothelial and blood fates.

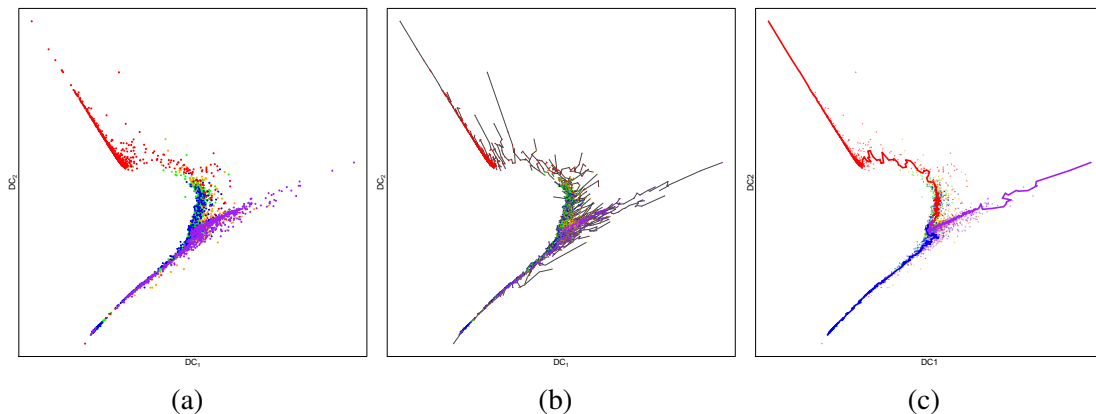


Fig. 5.10 Diffusion maps reconstructed from Moignard et al. [2015]. Colours are consistent with those in the experimental protocol overview fig. 5.9. (a) An overview diffusion map very similar to that published in the original paper fig. 5.9. (b) Minimum spanning tree on the diffusion map. (c) Shortest paths from the root at the lower left towards endothelial and blood end points - Common path in blue, blood trajectory in red and endothelial trajectory in purple.

5.4.1 Finding fate trajectories and dynamics of gene expression

The diffusion map here produces a representation which shows a bifurcating process representative of fate diversification. A minimum spanning tree (MST) on the diffusion map may therefore reproduce this bifurcation process and provide a good representation of the ontogenic processes. A MST was generated and diameter paths from the root to both endothelial and blood fates calculated fig. 5.10b. This allowed a common stem and the two blood and endothelial segments to be defined. The MST was cut to generate two trajectories for each of the blood and endothelial fate, fig. 5.10.

A principal curve was fitted independently to the two trajectories as was previously described in section 3.10. The arc length from the root cell of the cell projection on to the principal curve was taken to be a surrogate measure of developmental time i.e. pseudotime. Rescaling was performed to ensure that both processes had the same pseudotime at bifurcation and at the end. Gaussian process regression models were fitted to these rescaled pseudotimes to compare gene dynamics between the blood and endothelial fates, fig. 5.11. The genes for qPCR were selected manually to be informative of this bifurcating process and the gene dynamic profile comparisons between blood and endothelial clearly show differences in most genes for example *Cdh5*, *Egfl7*, *Ets1*, *Gfi1b* and *HbbbH1*.

A limitation with synthesising gene regulatory networks from static data is that although relationships between genes may be identified for example using correlation it can be difficult

5.4 Simulating developmental processes

or impossible to infer cause and effect i.e. the direction of edges between genes. Now given the dynamics that have been reconstructed using pseudotime it may be possible to direct edges between genes since any causal event must be temporally earlier than the induced effect.

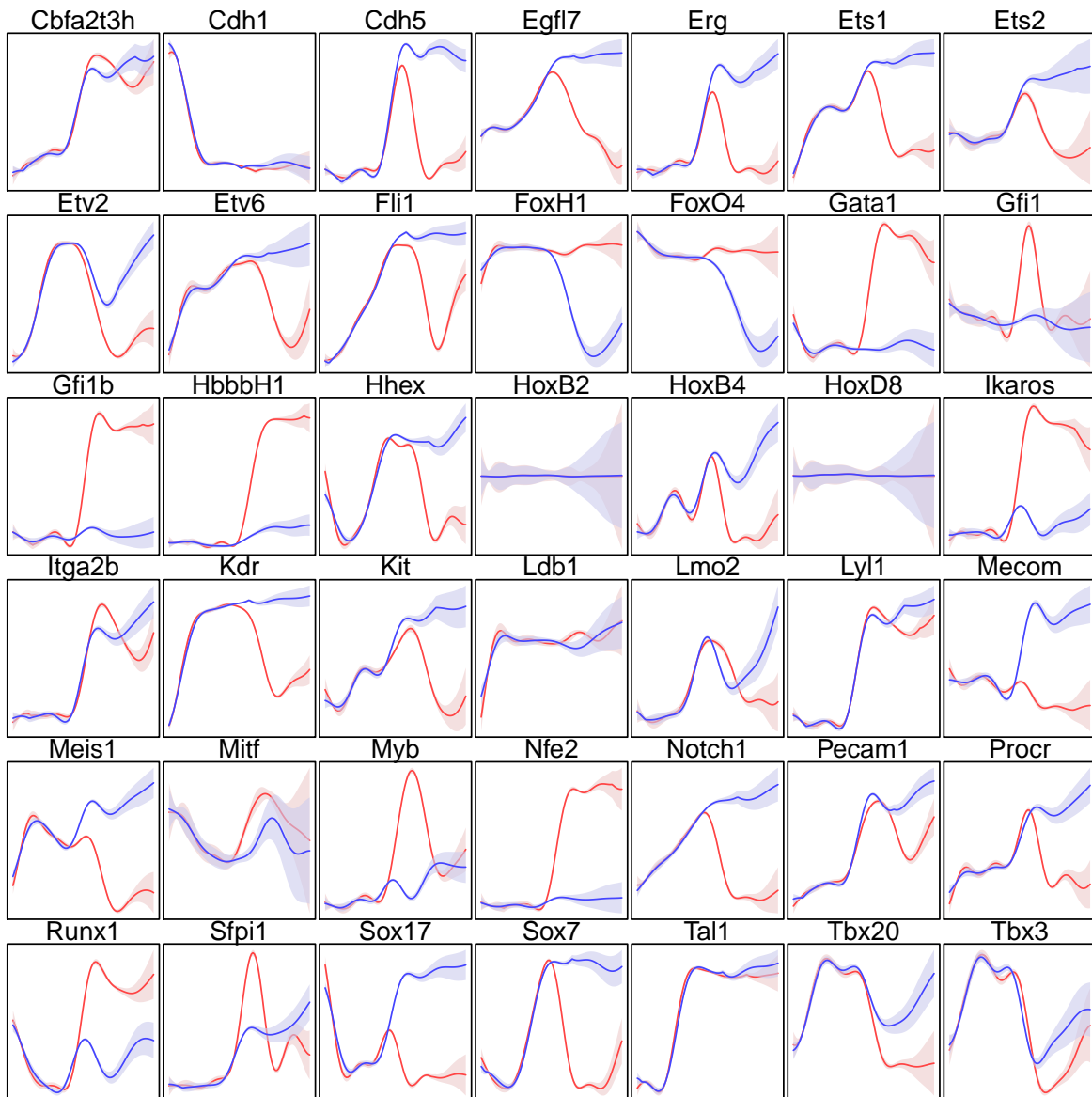


Fig. 5.11 Gene dynamics along trajectories identified in fig. 5.10c. Cells were mapped onto a principal curve fitted to these trajectories and the corresponding arc length used as a surrogate for pseudotime. The arc lengths along the two different trajectories were appropriately scaled so that bifurcation and termination were assigned the same pseudotime. Gene expression values were then smoothed using gaussian process regression and the inferred expression profile along pseudotime is plotted as a line with the shaded region representing 95 % confidence intervals. Red - haematopoietic fated path and blue - endothelial fated path.

5.4.2 Autoregressive model

Decoding gene interaction networks can provide deeper insights into molecular mechanisms and potentially allow researchers to target specific pathways to direct differentiation towards desired fates and redirect away from undesired cell types. One way to achieve this is to explicitly model associations between genes and if temporal dynamics are available this may even allow for causative relationships to be identified.

The smoothed gene expression profiles along pseudotime can be used to regress gene expression at a particular time point against gene expression at an arbitrarily earlier time. Given that gene expression of a cell at time $t \in \mathbb{Z}$ in the interval $[0, \tau]$, is given by $G(t) \in \mathbb{R}^m$ and for the expression of gene indexed i , $G_i(t)$ then the expected value given by $\hat{G}_i(t)$ can be regressed on the expression of all genes at an earlier point in pseudotime by finding all the parameters $\alpha_i \in \mathbb{R}^m$ which are essentially regression coefficients.

$$\begin{aligned}\hat{G}_i(t) &= \alpha_{i,1}G_1(t - \delta) + \alpha_{i,2}G_2(t - \delta) + \alpha_{i,3}G_3(t - \delta) + \dots + \alpha_{i,m}G_m(t - \delta) \\ &= \sum_{j=1}^m \alpha_{i,j}G_j(t - \delta) \\ \alpha_i &= \operatorname{argmin}_{\alpha_i} \sum_{t=\delta}^{\tau} |\hat{G}_i(t) - G_i(t)|\end{aligned}$$

This is pictorially represented in table 5.3. Cells, represented by columns, are ordered in pseudotime. Here ‘Gene 1’ for cell 2, which represents pseudotime 2, is considered to have gene expression that is some linear combination of all gene expression values at pseudotime 0 or cell 0 - all coloured in red. The same is true for ‘Gene 1’ for cell 3. This same approach can be used for all genes across all the dataset to infer all the α parameter values given in the equation above.

A naïve implementation of this approach will generate some non-zero value for all α . In such a case a hard threshold can be used so that all sufficiently small values are set to zero. An alternative is to set a constraint known as the least absolute shrinkage and selection operator (lasso), whereby the α values are constrained such that

$$\sum_{i=1}^m |\alpha_i| \leq s$$

5.4 Simulating developmental processes

Table 5.3 Table illustrating the essence of the autoregressive model. The model is trained so that ‘Gene 1’ in all cells is regressed against the value of the other genes at a fixed earlier pseudotime. The examples shown is how ‘Gene 1’ in cell 2 depends on all other genes at cell 0 and ‘Gene 1’ in cell 3 depends on all genes at cell 1.

	Cell 0	Cell 1	Cell 2	Cell 3	...	Cell τ
Gene 1	0.311	0.432	0.359	0.251	...	0.297
Gene 2	0.746	0.952	1.219	1.116	...	0.853
Gene 3	1.827	0.027	0.948	0.912	...	0.846
Gene 4	1.978	2.814	1.668	1.866	...	1.377
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
Gene m	0.537	0.577	1.015	0.756	0.911	1.640

where s is a user set constraint. The lasso though can be unstable for edge selection and a more consistent bootstrapped lasso using subsets of the data was adopted from the ‘mht’ package implemented in R.

This same approach can be extended so that expression $\hat{G}_i(t)$ depends on several discrete time points $\delta_1, \dots, \delta_\zeta$. This results in many more parameters that need to be estimated, so this approach was not used.

$$\begin{aligned}
 \hat{G}_i(t) &= (\alpha_{i,1}^{t-\delta_1} G_1(t-\delta_1) + \alpha_{i,1}^{t-\delta_2} G_1(t-\delta_2) + \dots + \alpha_{i,1}^{t-\delta_\zeta} G_1(t-\delta_\zeta)) + \\
 &\quad (\alpha_{i,2}^{t-\delta_1} G_2(t-\delta_1) + \alpha_{i,2}^{t-\delta_2} G_2(t-\delta_2) + \dots + \alpha_{i,2}^{t-\delta_\zeta} G_2(t-\delta_\zeta)) + \\
 &\quad \dots + \\
 &\quad (\alpha_{i,m}^{t-\delta_1} G_m(t-\delta_1) + \alpha_{i,m}^{t-\delta_2} G_m(t-\delta_2) + \dots + \alpha_{i,m}^{t-\delta_\zeta} G_m(t-\delta_\zeta)) \\
 &= \sum_{j=1}^m \sum_{k=1}^{\zeta} \alpha_{i,j}^{t-\delta_k} G_j(t-\delta_k)
 \end{aligned}$$

Transcriptomic assays like qPCR assess the level of gene expression and there is an assumption that this mirrors protein content but there will be some definite lag between production of the mRNA moiety to the production of the mature active transcription factor, enzyme or structural protein. It therefore seems reasonable to build in some temporal lag as it reflects contemporary understanding of biology. Further higher concentrations of transcription factors may be related to the level of activation of a particular genomic locus and so it may be more accurate to model the rate of change of the downstream gene’s counts rather than the counts themselves.

$$\begin{aligned}\Delta\hat{G}_i(t) &= \alpha_{i,1}G_1(t-\delta) + \alpha_{i,2}G_2(t-\delta) + \alpha_{i,3}G_3(t-\delta) + \dots + \alpha_{i,m}G_m(t-\delta) \\ &= \sum_{j=1}^m \alpha_{i,j}G_j(t-\delta)\end{aligned}$$

This was fitted separately for the smoothed gene profiles along both the blood and endothelial fates. Genes i and j were defined as interacting where $\alpha_{i,j} \neq 0$, activating when $\alpha_{i,j} > 0$ and repressing when $\alpha_{i,j} < 0$. The networks so inferred are shown in figs. 5.12 and 5.13. Genes are arranged in a circle and edges are directed. Blue edges indicate a positive interaction and red edges a negative interaction. Given the gene expression dynamics in fig. 5.11, the two networks appear consistent for example the *HbbbH1* gene is included in the blood network while it is completely disconnected in the endothelial network. Unfortunately these are linear and very simplistic networks and since each network is inferred separately a bifurcating process cannot be predicted by these networks. The utility of such models becomes limited as they cannot directly provide any information on how gene perturbations may influence fate bias, they can neither model complex non-linear relationships nor can they model bifurcating processes.

The linear model shown here is simple and extremely limited unless higher order terms and additional time points are included but increasing to higher order terms will additionally increase the number of parameters that must be estimated. A problem for all models though is that there exist genes with very similar profiles for example *Gfl1b* and *HbbbH1* and this collinearity can cause difficulties with inference.

5.4.3 Deep neural network

The concept of gene network inference can be abstracted to a model which takes a gene expression vector and outputs a gene expression vector for a cell a single step further along its fated journey as shown in fig. 5.14. The function takes in the gene vector of a cell at an arbitrary time point t and outputs the gene expression of a cell one step further along $t + 1$.

The previously described linear regression model may be considered graphically as shown in fig. 5.15. There is essentially a single layer of linear functions between the input gene vector and the output gene vector.

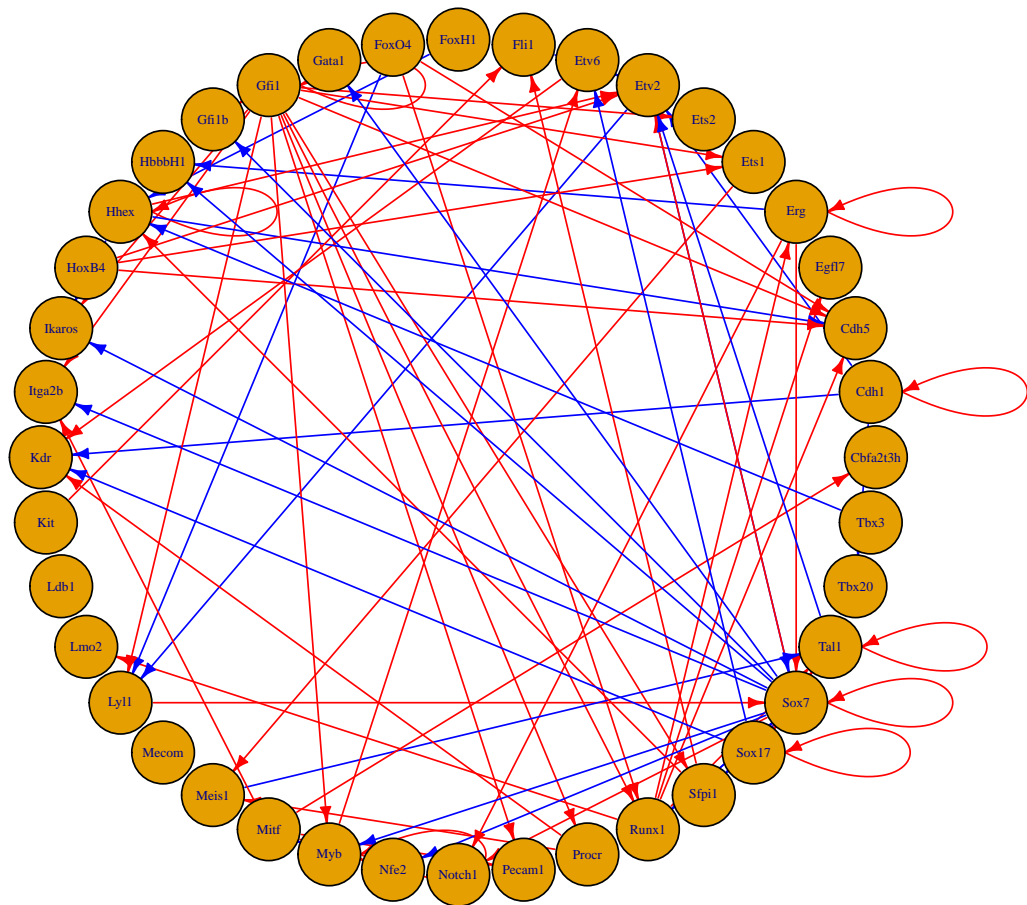


Fig. 5.12 Autoregressive model inferred on the blood gene expression dynamics, see fig. 5.11 and compare with fig. 5.13. Blue edges represent positive activating interactions and red edges represent negative inhibitory interactions.

Given this abstraction and following this line of thought a more complex non-linear function may be conceived, consisting of multiple latent or hidden variables, fig. 5.16. This model is consistent with a deep learning paradigm, rapidly gaining popularity and with novel applications in the fields of artificial intelligence and more generally in machine learning. Methods to train this type of model are widely available and one such method is back propagation.

Despite the ability of this model to generate highly complex non-linear functions it remains completely deterministic and given a single input will always produce exactly the same output. A bifurcating system cannot therefore be faithfully modelled.

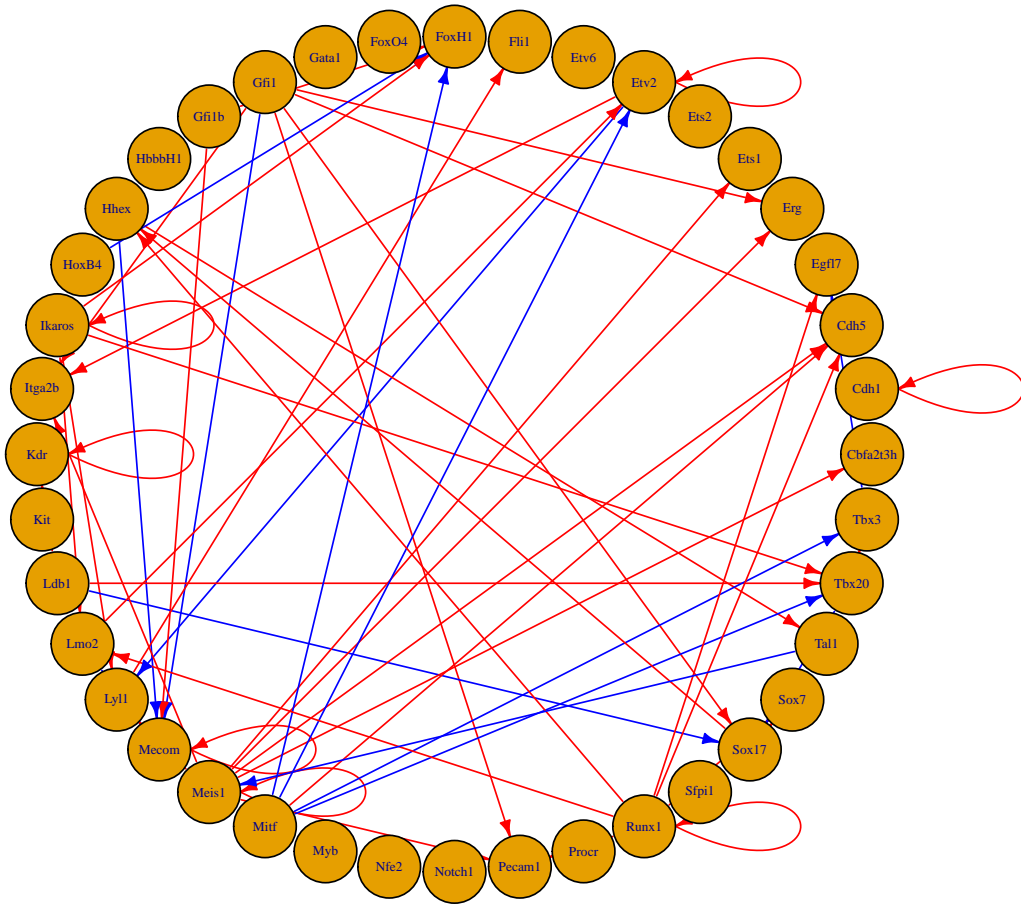


Fig. 5.13 Autoregressive model inferred on the endothelial gene expression dynamics, see fig. 5.11 and compare with fig. 5.12. Blue edges represent positive activating interactions and red edges represent negative inhibitory interactions.

To enable modelling of a bifurcating system, the output of the neural network can be considered to be the parameters of some random variable and a realisation of this random variable then be the output gene expression vector, as graphically demonstrated in fig. 5.17. Concretely using the gene expression example, the neural network will output the means $\mu \in \mathbb{R}^m$ and the gene output would be a realisation generated by sampling the random variable, y :

$$y \sim N(\mu, \Sigma), \quad \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m}.$$

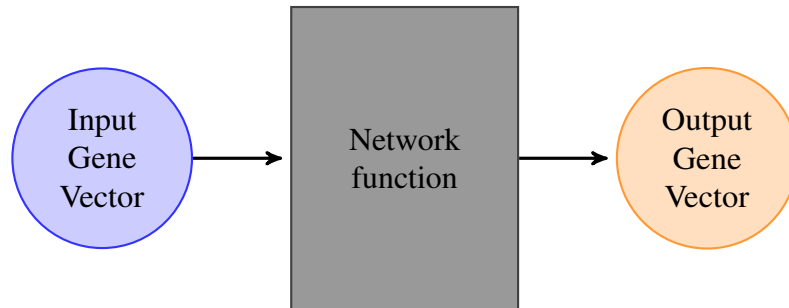


Fig. 5.14 Abstraction of a network inference function that can predict the next cell along in a developmental trajectory. Such a function would enable the user to make predictions of gene expression changes due to perturbations. The output can then be returned into the network function so that a complete trajectory can be realised.

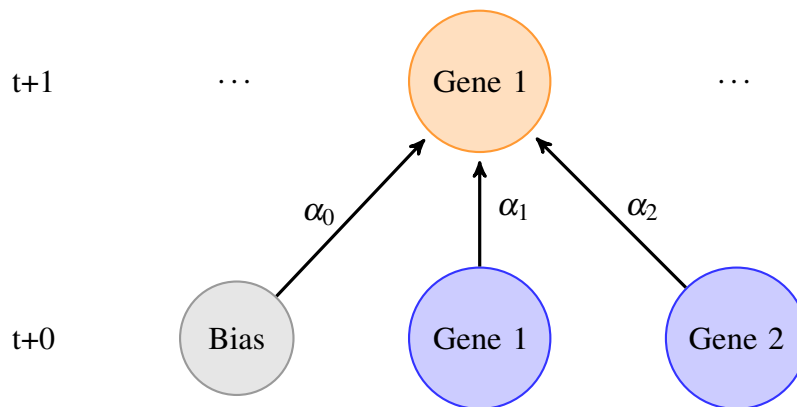


Fig. 5.15 Graphical representation of a linear regression model where gene expression at a time $t + 1$ is determined using a linear model from gene expression of a known cell at time $t + 0$. This can be used iteratively to determine a whole gene trajectory, see figs. 5.12 to 5.14. This representation can be extended to a more complex model, see fig. 5.16.

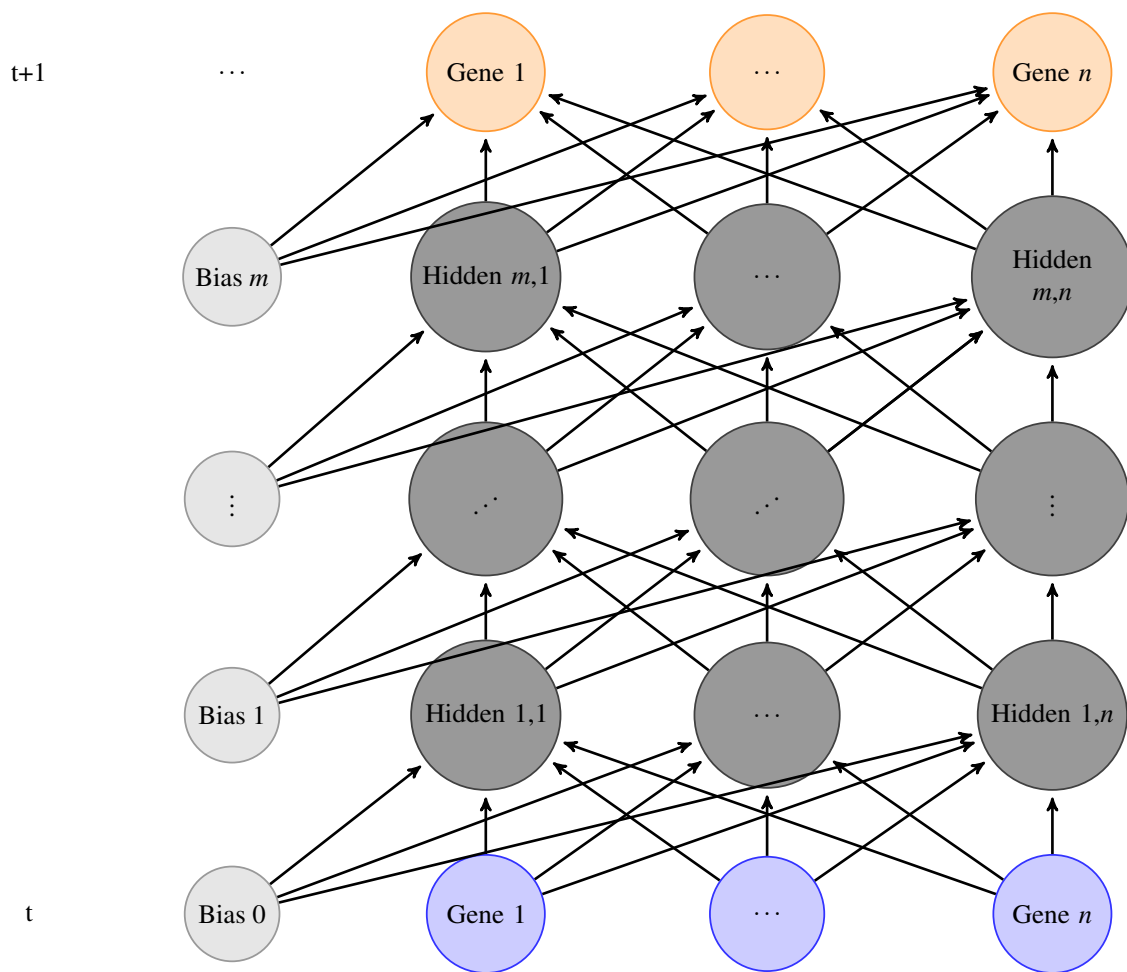


Fig. 5.16 A more flexible model than linear regression as shown in fig. 5.15. Now several hidden layers are present allowing for non-linear complex functions, a generalisation more fitting of the network abstraction depicted in fig. 5.14. Though not annotated for clarity, each interaction represented by an edge is parameterised in the same way as in fig. 5.15. A key feature is that the network is trained so that given an input gene vector at time t , in blue, it produces the expected gene vector at a time $t + 1$, in orange. Further the new output can now be returned to the input phase of the network. Such an architecture represents a recurrent neural network.

A training dataset was generated from the qPCR gene expression data. A transition matrix L with entries $L_{i,j}$, was calculated from the normalised qPCR data $X \in \mathbb{R}^{m \times n}$ of m genes G and n cells C , in a manner similar to the diffusion map calculation as previously described in section 3.10.1. To remain consistent with Moignard et al. [2015] the Gaussian kernel width was set $\sigma = 12$ and only top 20% of the nearest neighbours were retained.

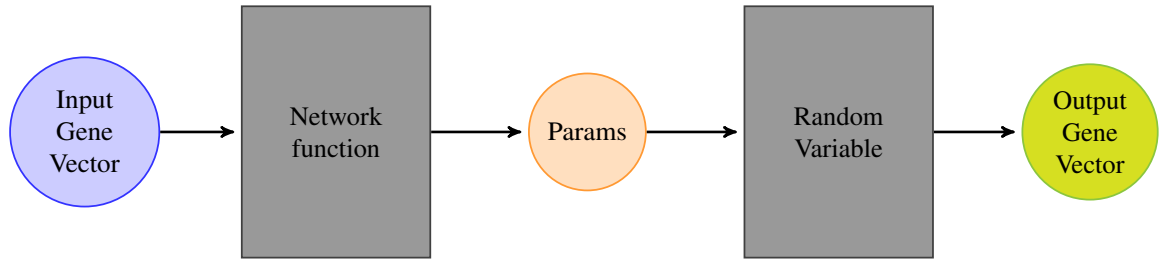


Fig. 5.17 Building on the neural network model. The deterministic output from the neural network outputs a mean vector $\mu \in \mathbb{R}^m$ which is used as the parameter for a random variable $y \sim N(\mu, \Sigma)$. The realisation of the random variable y would be used as the input vector at the next iteration. This built in stochasticity now may allow bifurcating trajectories.

$$d_i^j = \|x_i - x_j\|^2$$

and β_i is a permutation of d_i

$$\beta_i = \{\beta_i^1, \beta_i^2, \dots, \beta_i^n\}$$

such that,

$$\beta_i^{k+1} \geq \beta_i^k, \quad \forall k$$

and 20 % nearest neighbours were retained by setting,

$$\alpha_i = \beta_i^{\|n/5\|}$$

Now $L_{i,j}$ can be defined,

$$L_{i,j} = \begin{cases} 0, & \text{if } i = j \text{ or } \|x_i - x_j\|^2 > \alpha_i \\ \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), & \text{otherwise} \end{cases}$$

$x_i, x_j \in \mathbb{R}^m$ are columns of the matrix X representing qPCR measured gene expression of cells indexed i and j , respectively. The transition matrix establishes a connectivity and allows a random walk between cells to be implemented but overall generates no flux of cells from an

Computational tools development

early to late developmental time point. For faithful and biologically meaningful trajectories only pseudotime consistent and branch honouring transitions are allowed. Specifically there were 2 further constraints, first a pseudotime constraint so that any transitions that would lead to a transition that went backwards in pseudotime were disallowed. Second a branching constraint so that branch swapping was only allowed between contiguous branches. if either constraint was unmet the transition was disallowed. This is summarise below by $\hat{F}_{i,j}$.

The pseudotime was calculated using principal curve and branch assignment based on the MST as described in section 5.4.1. Each cell was assigned to either set B_r , the root branch in blue; set B_b , the blood branch in red or set B_e , the endothelial branch in purple, see fig. 5.10c. Additionally each cell c indexed by i was assigned a rescaled principal curve pseudotime pct_i . A directed transition matrix, F representing a flux through the biological system was therefore calculated honouring the branch structure and pseudotime.

$$\hat{F}_{i,j} = \begin{cases} 0, & \text{if } c_j \in B_b | c_i \in B_e \\ 0, & \text{if } c_j \in B_e | c_i \in B_b \\ 0, & \text{if } c_j \in B_r | c_i \in B_b \\ 0, & \text{if } c_j \in B_r | c_i \in B_e \\ 0, & \text{if } pct_j < pct_i \\ L_{i,j}, & \text{otherwise} \end{cases}$$

To avoid having cells with no outward transitions

$$F'_{i,j} = \begin{cases} 1, & \text{if } i = j, \sum_k \hat{F}_{i,k} = 0 \\ \hat{F}_{i,j}, & \text{otherwise} \end{cases}$$

and

$$F_{i,j} = \frac{F'_{i,j}}{\sum_i F'_{i,j}}$$

The training data is generating by randomly selecting 10,000 starting cells from a uniform distribution between $\{1, 2, 3, \dots, n\}$ with replacement. 10,000 terminating cells are then selected for each starting cell using a multinomial distribution parameterised by the row vector $F_{\{i,\cdot\}}$.

5.4 Simulating developmental processes

The qPCR gene expression data matrix, X was then mean centered and scaled by the standard deviation. Finally the mean centered standard deviation scaled data was normalised to be in the range $[0, 1]$.

$$\begin{aligned}\mu_i &= \frac{\sum_j X_{i,j}}{m} \\ \sigma_i &= \frac{\sum_j (X_{i,j} - \mu_i)^2}{m - 1} \\ X'_{i,j} &= \frac{X_{i,j} - \mu_i}{\sigma_i} \\ \hat{X}_{i,j} &= \frac{X'_{i,j} - \min_j (X'_{i,j})}{\max_j (X'_{i,j}) - \min_j (X'_{i,j})}\end{aligned}$$

This pre-processing has prepared the training data honouring the calculated pseudotime and the branching structure. The artificial neural network architecture was selected with input and output layers of 42 nodes with 3 intervening hidden layers of 100 artificial neurons fig. 5.18.

All nodes in all layers used the sigmoid activation function. For any single input-output training set pair, activation of the n -th neuron in layer l is given by a_n^l . The edge weight between node $a_m^{(l-1)}$ and node a_n^l is given by $\omega_{m,n}^{(l-1)}$. The number of neurons in layer $l - 1$ is r .

$$\begin{aligned}z_n^l &= \sum_{i=1}^r \omega_{i,n}^{(l-1)} a_i^{(l-1)} \\ a_n^l &= \frac{1}{1 + \exp(-z_n^l)}\end{aligned}$$

The artificial neural net may be considered to output a probability since the domain of the sigmoid function is $[0, 1]$, more specifically it outputs the probability that $y_i = 1$. Though in the specific case here the output is not a class but continuous gene expression the same method has been adopted. Training the artificial neural network involves adjusting all the weights to minimise the final error.

Training is performed using back propagation whereby the output of the final layer is compared to the expected output from the training dataset and then this used to adjust the

Computational tools development

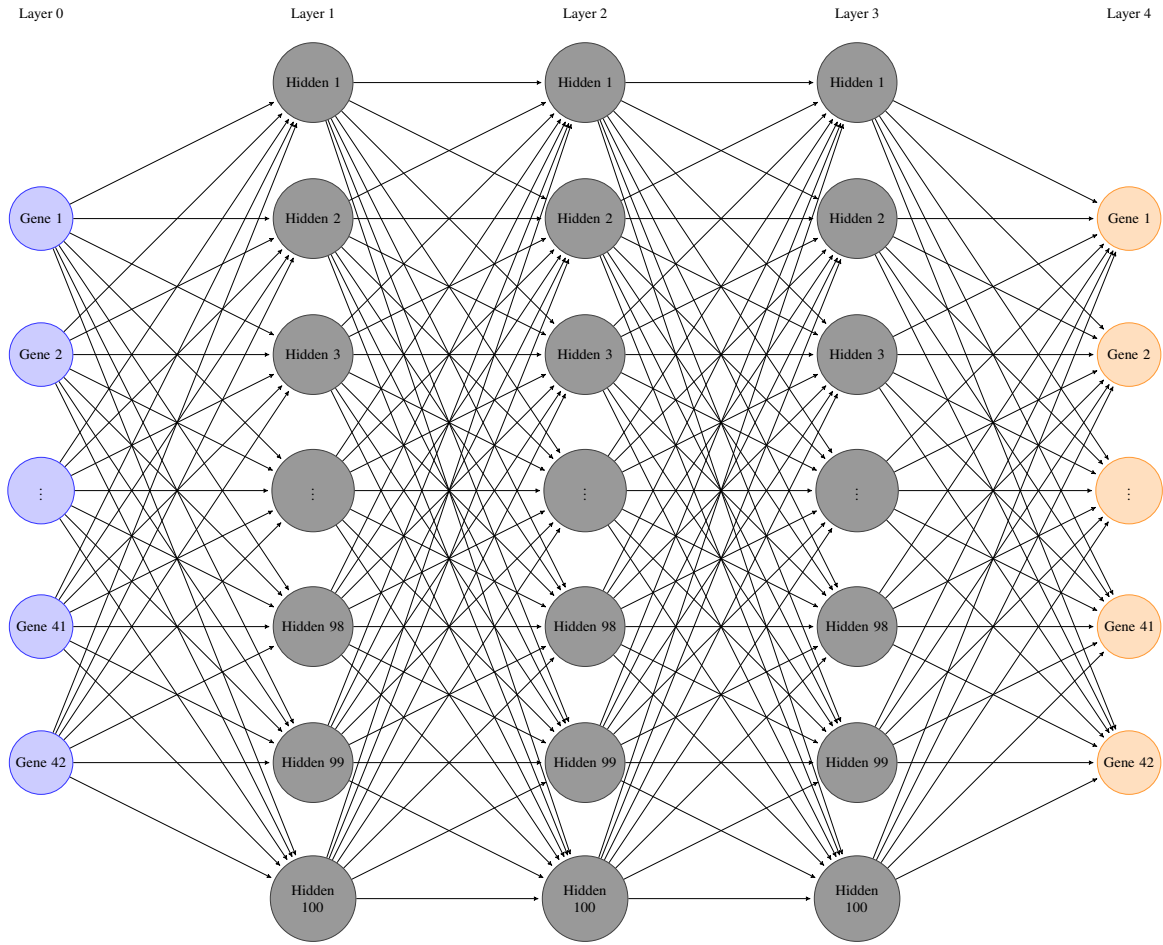


Fig. 5.18 Architecture of the artificial neural net used to model gene expression along pseudotime. There were 42 input and output genes and 3 hidden layers with 100 neurons each. Bias neurons have not been shown.

weights on the interaction between the penultimate and the final layer. The error is then back propagated through all the layers iteratively, adjusting all relevant weights to improve the predictive ability of the network.

In classification problems, given an individual training data point with $x_i \in \mathbb{R}^m$ input gene expression profile of gene set G , $y_i \in \mathbb{R}^m$ output gene expression profile and the output of a single layer of the artificial neural net $h_\omega(x_i) \in \mathbb{R}^m$, the probability of the expression of gene expression $L(\omega)$ for a single data point is calculated:

$$p(y_i|x_i, \omega) = \prod_{j \in G} h_\omega(x_i^j)^{y_i^j} (1 - h_\omega(x_i^j))^{(1-y_i^j)}$$

since

$$p(y_i = 1 | x_i, \omega) = \prod_{j \in G} h_{\omega}(x_i^j)^{y_i^j} \text{ and}$$

$$p(y_i = 0 | x_i, \omega) = \prod_{j \in G} (1 - h_{\omega}(x_i^j))^{(1-y_i^j)}$$

so the probability across all data points N :

$$p(y | x, \omega) = \prod_{i \in N} \prod_{j \in G} (h_{\omega}(x)_i^j)^{y_i^j} (1 - h_{\omega}(x)_i^j)^{(1-y_i^j)}$$

and the log-probability

$$L(\omega) = \sum_{i \in N} \sum_{j \in G} (y_i^j) \log(h_{\omega}(x)_i^j) + (1 - y_i^j) \log(1 - h_{\omega}(x)_i^j)$$

The cost function $J(\omega)$ is given by the average of the negative log probability:

$$J(\omega) = -\frac{1}{|N|} L(\omega)$$

The negative log probability is now used as the cost function $J(\omega)$ and the parameters ω are found that minimise this cost function using a conjugate gradient descent algorithm. Back propagation can now be used to calculate the errors starting at the output level and then propagate the error to lower layers iteratively terminating at the input level.

Taking the simple case of two layers of neurons with sigmoid activation functions, the m neurons in the first layer indexed by i and the n neurons in the second output layer indexed by j , see fig. 5.19, partial differentials for the predicted output $h_{\omega}(x)_j$ can be calculated.

The sigmoid function can be considered a compound function, broken into its constituents and the chain rule applied as follows,

$$z = \omega x, \quad z \in \mathbb{R}^n, \omega \in \mathbb{R}^{n \times m}, x \in \mathbb{R}^m$$

$$r = 1 + \exp(-z)$$

$$h_{\omega}(x) = \frac{1}{r}$$

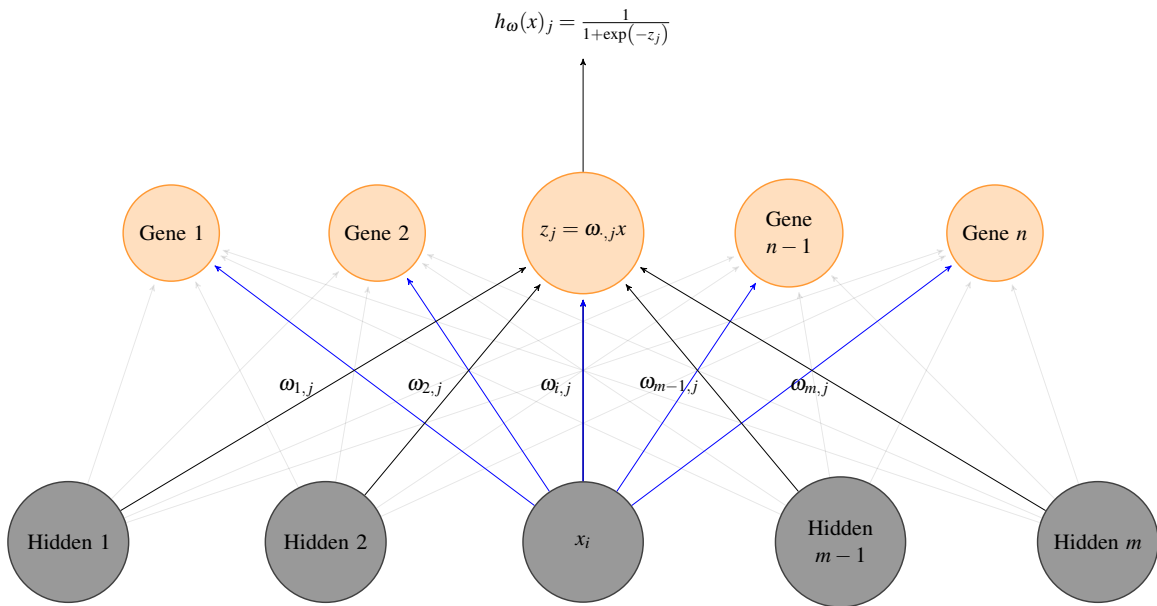


Fig. 5.19 Two layers of a neural net to help demonstrate both forward and back propagation. The activation of the hidden layer are given by x_i and the weights on the edges into a single output node indexed j are labelled $\omega_{.j}$. The activation of the j -th node in the upper output layer is labelled $h_{\omega}(x)_j$

so,

$$\begin{aligned}\frac{\partial z_j}{\partial \omega_{ij}} &= x_i \\ \frac{\partial r_j}{\partial z_j} &= -\exp(z_j) = 1 - r_j \\ \frac{\partial}{\partial r_j} h_{\omega}(x)_j &= -\frac{1}{r_j^2}\end{aligned}$$

therefore,

$$\begin{aligned}\frac{\partial}{\partial \omega_{ij}} h_{\omega}(x)_j &= \frac{\partial h_{\omega}(x)_j}{\partial r_j} \frac{\partial r_j}{\partial z_j} \frac{\partial z_j}{\partial \omega_{ij}} \\ &= x_i \frac{r_j - 1}{r_j^2} \\ &= x_i h_{\omega}(x)_j (1 - h_{\omega}(x)_j)\end{aligned}$$

and

$$\frac{\partial}{\partial z_j} h_{\omega}(x)_j = h_{\omega}(x)_j(1 - h_{\omega}(x)_j)$$

Having established the partial differential of the j -th output with respect to the weights $\frac{\partial}{\partial \omega_{i,j}} h_{\omega}(x)$, the partial differentials of the cost function $J(w)$ with respect to the weights may be calculated,

$$\begin{aligned} \frac{\partial}{\partial h_{\omega}(x)_j} J(\omega) &= -\frac{1}{|N|} \left(\frac{y_j}{h_{\omega}(x)_j} + \frac{y_j - 1}{1 - h_{\omega}(x)_j} \right) \\ &= \frac{1}{|N|} \left(\frac{h_{\omega}(x)_j - y_j}{(h_{\omega}(x)_j)(1 - h_{\omega}(x)_j)} \right) \end{aligned}$$

additionally

$$\begin{aligned} \frac{\partial}{\partial z_j} J(\omega) &= \frac{\partial h_{\omega}(x)_j}{\partial z_j} \frac{\partial J(\omega)}{\partial h_{\omega}(x)_j} \\ &= \frac{1}{|N|} (h_{\omega}(x)_j - y_j) \end{aligned}$$

and finally,

$$\begin{aligned} \frac{\partial}{\partial \omega_{i,j}} J(\omega) &= \frac{\partial z_j}{\partial \omega_{i,j}} \frac{\partial J(\omega)}{\partial z_j} \\ &= \frac{1}{|N|} x_i (h_{\omega}(x)_j - y_j) \end{aligned}$$

This now allows the final stratum weights to be adjusted to reduce error for a single training case but for back propagation to work the error or gradient must be propagated down to the activity of the neurons in the immediately adjacent layer, for demonstrative purposes a single neuron in that layer x_i is considered. Since the activity of this lower level has effects on multiple output level neurons as demonstrated by the blue interactions in fig. 5.19, all these partial derivatives must be summed together,

$$\begin{aligned}\frac{\partial}{\partial x_i} J(\omega) &= \sum_j \frac{\partial z_j}{\partial x_i} \frac{\partial J(\omega)}{\partial z_j} \\ &= \frac{1}{|N|} \sum_j \omega_{i,j} (h_{\omega}(x)_j - y_j)\end{aligned}$$

Having calculated the partial differentials of the cost function with respect to the activation of the lower level neurons the algorithm can be applied iteratively to update all weights of the network, layer by layer.

The neural network model has $m \times n$ different weight parameters and to counter against over-fitting and allow generalisation to new data, L_2 regularisation was used which is essentially similar to weight-decay. The regularisation term was parameterised using λ . Now the cost function becomes,

$$\begin{aligned}J(\omega) &= -\frac{1}{|N|} \left(\sum_{i \in N} \sum_{j \in G} (y_i^j) \log(h_w(x)_i^j) + (1 - y_i^j) \log(1 - h_w(x)_i^j) \right) \\ &\quad + \frac{\lambda}{2|N|} \left(\sum_l \sum_m \sum_n \omega_{m,n}^2 \right) \quad \lambda \in \mathbb{R}^+, l \in L, m \in k^l, n \in (k)^{l+1}\end{aligned}$$

L represents all layers of the neural net, and k^l all neurons on layer l . Now the gradient of the cost function with respect to the weight $\omega_{i,j}$ is given by

$$\frac{\partial}{\partial \omega_{i,j}} J(\omega) = \frac{1}{|N|} x_i (h_{\omega}(x)_j - y_j) + \frac{\lambda}{|N|} \omega_{i,j}$$

All components of the neural network have been described, data normalisation, scaling and centering, production of input and output gene expression vectors as the training dataset, the architecture of the network including activation functions, the cost function with regularisation and the back propagation algorithm. With all these in place a neural network training was implemented in Octave using the conjugate gradient descent function *fmincg* which was written by Carl Rasmussen [Eaton et al., 2015; Rasmussen, 2002].

5.4 Simulating developmental processes

Having trained the network it was used to make de-novo predictions, taking as input a given gene expression vector and outputting the gene expression vector of the next cell along the pseudotime. The initial input of a starting cell selected from within a region of the developmentally earlier cells was given to the network and the output used as parameters of a Gaussian for each gene. This realisation was then used as the new input to the neural network as summarised in fig. 5.17. The output was plotted on a diffusion map with early points coloured blue and the final points in red, see fig. 5.20. De-novo gene expression vectors were visualised by projecting onto the pre-constructed diffusion map. The network was able to generate gene expression profiles that projected meaningfully on the diffusion map and strikingly reproduce the bifurcating process on which it was trained. It was able to generate both blood as displayed in fig. 5.20a and endothelium, fig. 5.20b.

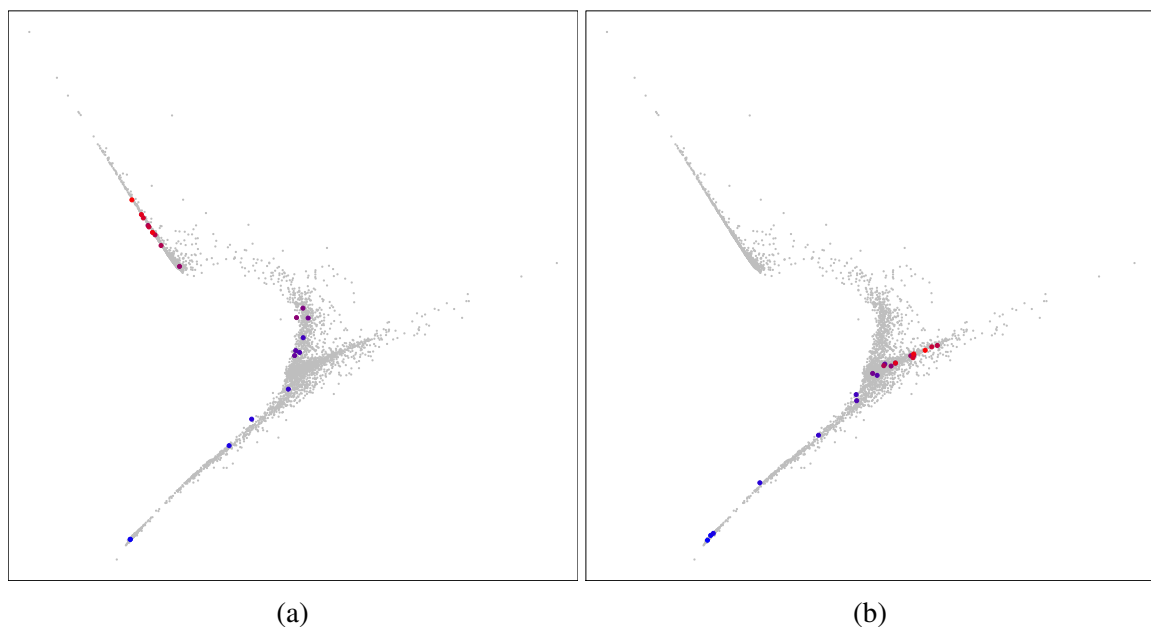


Fig. 5.20 Gene expression profiles generated by the neural network reconstructs the bifurcating developmental trajectories from early mesodermal progenitors to blood (a) and endothelium (b)

Gene perturbations can be simulated in the network by simply resetting the expression of a gene to an off or on level. For example *Tall* may be knocked out by setting the gene to off and 50 trajectories simulated, the resulting final gene expression profiles can be plotted on the diffusion map, fig. 5.21. The neural network predicts that knocking out *Tall* or *Gfi1b* blocks embryonic blood production while knocking out *Cdh5/VE-Cadherin* blocks an endothelial cell fate in keeping with known biology [Anderson et al., 2015; Moore et al., 2018; Robb et al., 1995].

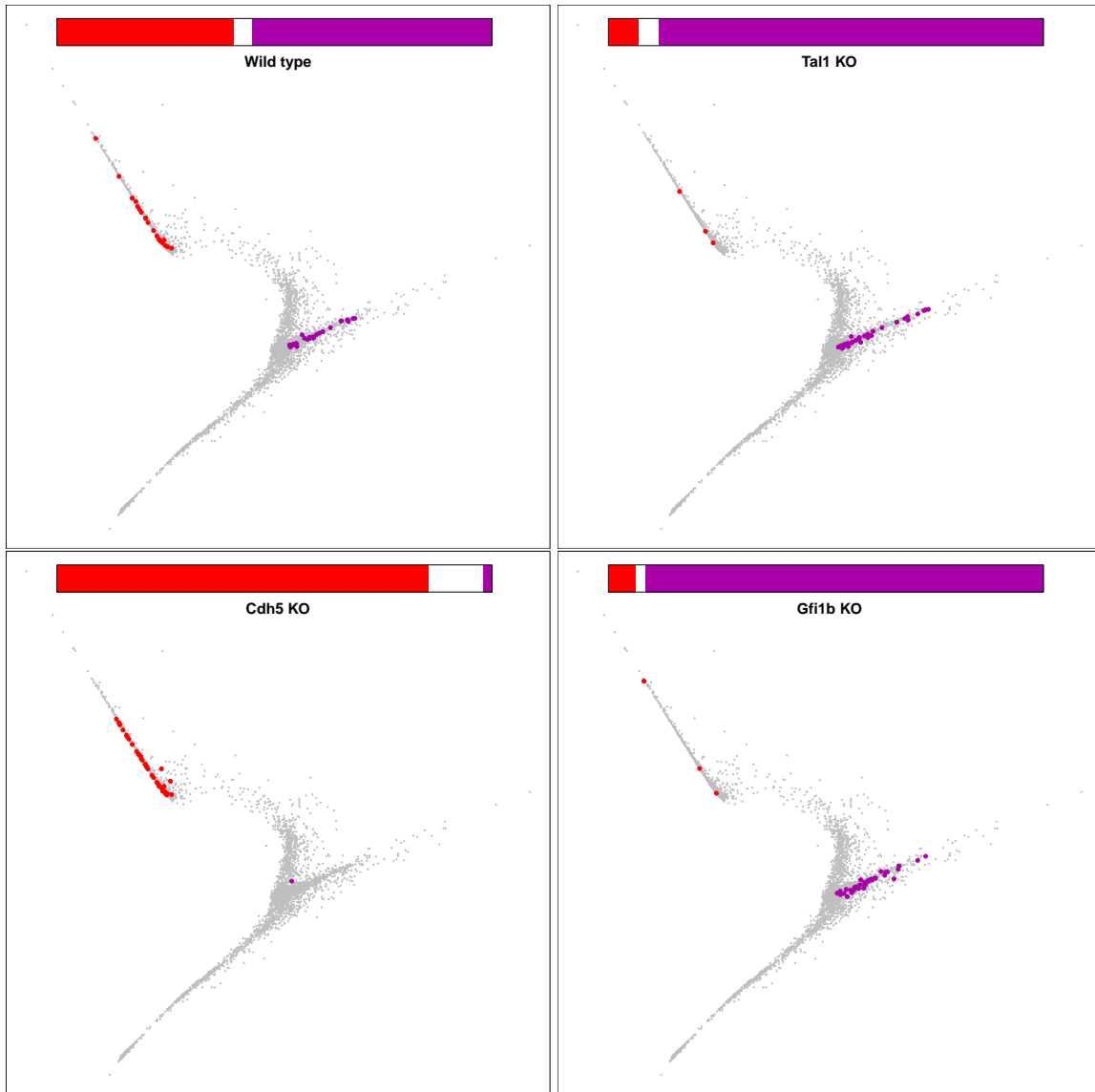


Fig. 5.21 Each figure demonstrates end points of 50 simulations with the perturbations in the title. Top left shows wild-type with balanced blood and endothelial fate. Top right shows that the neural network predicts that *Tal1* knock-out results in a failure of embryonic blood production [Robb et al., 1995]. In a similar way bottom right shows *Gfi1b* knock-out blocks embryonic blood development [Moore et al., 2018] and bottom left shows neural net prediction after *Cdh5* knock-out blocks endothelial production but still generates blood [Anderson et al., 2015]. The bars above demonstrate relative ratios of blood (red), endothelium (purple) or intermediate cell types (white).

5.5 Conclusions

This chapter, in contrast to previous chapters, has focused on computational and mathematical methods rather than generating biological data. Initially a novel visualisation tool *roots* was described. The development of this method was influenced by existing methods which used the concept of probabilistic transitions from one data point to another such as tSNE and diffusion maps.

A key concept focused on in *roots* though is the idea of ‘plausible transitions’. As described earlier this requires adequate sampling otherwise implausible short cuts are allowed between two or more points on the high dimensional gene expression manifold. After assuming adequate sampling this is achieved three fold. First by adjusting the Gaussian kernel as has been described by Haghverdi et al. [2015].

Second by setting an empirical threshold on the number of other cells a single cell is equally likely to transition to. So taking an overly simplistic example in a dataset of 5000 cells where one expects at least 5 equally sized disparate cell types one may consider that an individual cell is unlikely to be equally connected to more than 1000 cells so that any transition probability under $\frac{1}{1000}$ may be considered to be unrealistic. Though this is a hard threshold it is relatively easy and intuitive for a user with some knowledge of their experimental system to estimate.

Third despite having made the matrix sparse there will remain spurious edges connecting cells from distant regions of the graph. To handle these spurious edges, edges are considered authentic only if there exists a common neighbour, see section 5.2.2 for further details. This makes the transition matrix further sparse and the sparsity gives an additional benefit particularly useful in huge datasets as it reduces computational memory requirements and permits the use of pre-existing implementations of functions designed to take advantage of sparse matrices.

The sparse transition matrix was symmetrised and used as input for Louvain clustering providing fast clustering and giving very reasonable clusters even identifying the small primordial germ cell cluster in the 10X Genomics dataset, section 4.3. This population was not identified by the clustering method used in the original publication [Ibarra-Soria et al., 2018].

An alternative dataset generated from adult haematopoiesis was introduced and the *roots* visualisation compared to current state of the art methods including tSNE, diffusion maps and SPRING. The *roots* visualisation generates a plot that most closely fits our current knowledge

Computational tools development

of adult haematopoietic stem cell differentiation demonstrating the known relationship between erythrocytes and megakaryocytes. Additionally the purposefully built-in sparsity and removal of spurious edge algorithms allow *roots* to clearly demonstrate the gap between the megakaryocyte-erythroid lineage and the remainder, consistent with the gap in the sorting strategy.

Having described a novel visualisation technique the tSNE algorithm was then extended. tSNE is widely deployed for single cell analysis and forms an integral part of several software suites because of its performance and ability to separate different cell populations. It is a probabilistic algorithm, based on calculating transition matrices in both gene expression space and low-dimensionality space and minimising the difference between the two probability distributions as assessed by the Kullback-Leibner divergence, see section 5.3. One limitation of the tSNE algorithm as it stood was that there was no way of projecting onto a pre-existing map though an alternative method of training a neural network to minimise the Kullback-Leibner divergence and using this complex non-linear function to project new data points has been proposed as a parametric version of tSNE [Maaten, 2009]. But this requires regenerating the original tSNE using the new algorithm, altering the visual layout and necessitating a re-interpretation.

As an alternative, in this chapter the current widely used tSNE algorithm was modified so that pre-existing data points were fixed and had no influence on the newly defined cost function still based on the Kullback-Leibner divergence. This allows new data points to arrange around the previous points so fitting into the pre-existing embedding. This idea was further advanced by considering that the user may have some preconceived idea about the gross structure of their experimental data from previous work and so may want to influence the landscape. An example is given with spacially resolved bulk sequencing data and then projecting single cells on this pre-conditioned landscape. This modified tSNE algorithm therefore opens up many more opportunities when wanting to visualise new data sets. The new tSNE algorithm was used to interpret gene expression profiles of cells from the $T^{nEGFP-CreERT2/+}$ mouse line section 3.14.6.

A method that attempts to explicitly model gene interactions, the auto-regressive model, was used as the basis for the development of a more flexible and powerful simulation model. The auto-regressive model in a similar vein to other recent efforts attempts to model biological molecular mechanisms by breaking the problem into pair-wise gene interactions and using data to infer these explicitly. The neural network model in contrast is predicting gene expression patterns along developmental trajectories and only implicitly modelling complex gene interactions within the hidden layers.

Gene interactions are not directly modelled in the neural network like in an autoregressive model, rather the network is trained on gene expression patterns over a developmental progression. Though superficially this may seem to be a disadvantage, it does allow one to model complex gene dynamics without getting into the specific detail of how individual genes are interacting and in no way detracts from the ability to investigate gene interaction within the network. The complex functions a neural network is able to learn can allow it to potentially model a different interaction between two genes depending on context i.e. the cell's gene expression state. Remarkably the network when combined with random Gaussian noise was able to reproduce a bifurcating trajectory and given the limitation of the numbers of genes and cell states present in the system was able to model genetic perturbations in-silico. This network was implemented without any of the currently available tools available for deep learning and was trained simply on a CPU rather than a GPU. Neural networks thrive on large datasets and with increasing cell numbers and genomic features networks modelled in this way may reveal previously unrecognised insights.

In conclusion this chapter summarises the mathematical and computational concepts that have been formulated and implemented over the course of this PhD.

Chapter 6

Discussion

This chapter is laid out according to the 4 aims set out in the introduction. Techniques developed and used often help to address multiple aims, so pertinent features of the same technique may be commented upon in the different subsections. Their development, use and limitations are discussed along with potential opportunities for further development in future work.

6.1 Identifying cell populations

A key promise of single cell transcriptomic sequencing, even for what are apriori believed to be homogenous populations, is the potential to discover and define hitherto unrecognised cell types. Some transcriptional heterogeneity identified may simply be related to the cell cycle or other process unrelated to cellular identity but this in itself can be interesting, particularly if cells of the same type in different stages of cell cycle favour alternative cell fates.

Key to identifying populations is selection of useful genomic features. In this body of work feature selection has depended on choosing genes based on methods that have relied heavily on the statistical concept of variability, in particular variability beyond a certain threshold for a particular level of gene expression [Brennecke et al., 2013; Svensson et al., 2017]. Though useful this denovo strategy fails to harness any previously acquired knowledge. As a second step genes known to be important in a process, could have been hand picked and manually added but an automated method here would be appealing. New strategies for gene selection continue to be identified and there remains much room for improvement at this fundamental step. It may be argued that selecting genes at an early step biases downstream analyses and though this may be a potential hindrance, not selecting and using the whole dataset

Discussion

often leads to poor visualisations during dimensionality reduction and makes interpretation difficult, as true biologically important signals are swamped in noise.

Use of spike-ins to differentiate technical variability from potentially interesting biological variability introduces its own challenges, for example ensuring that sequencer reads are not overly allocated to spike-ins rather than cellular reads. Variations in spike-ins from experiment to experiment can render them unusable [Svensson et al., 2017]. Additionally spike-in capture efficiency may not mirror that of endogenous mRNAs due to sequence differences, shorter poly-A tails and lack of mRNA binding proteins [Svensson et al., 2017].

Having selected genes of interest, cell types have been identified using a variety of clustering algorithms that have been described in detail. Clustering though depends on calculating either similarities or dissimilarities. A variety of methods can be used but using absolute values such as normalised counts can cause problems with data from different experiments, in most part due to normalisation difficulties. Throughout this work similarities have been calculated using methods that use ranks, for example rather than using Pearson correlation the Spearman rank correlation has been widely used. Though this overcomes difficulties of normalisation it is problematic for the many lowly expressed genes due to the effect of dropout. Depending on the effective depth of sequencing of two cells, the number of dropouts may vary affecting the similarity score, whether calculated on the absolute values or rank.

Similarities though can be calculated readily and used to generate graphs with edge weights corresponding to similarities. These graphs are heavily connected as similarities based on correlation methods will generate non-zero values for almost all pairwise comparisons. Methods are described here to make the adjacency matrices on which the graphs are based sparse and then further to remove spurious edges. Using the same datasets this has revealed novel populations not identified by standard methods for example the germ cell population in fig. 4.10 and table 4.4. Additionally it allowed the NAMP population to be subdivided into the NMP and streak-like population that were not previously evident, fig. 4.17.

Pre-processing involves converting reads to counts which is performed by aligning reads to an annotated genome. The 10x Genomics® Chromium™ data by virtue of the anchored 3' oligo-DT barcoded primers offers an opportunity to identify and enumerate transcription termination variants (TTV), as defined in section 4.6. These variants are clearly visible in the combined reads data as shown in fig. 4.22 and demonstrated by the read pile-ups of about 400 bp width. Further work is required to assess whether these TTVs can provide any additional information when segregating cell types. The combined data can be used to annotate new potential features on the genome. Counts can then be generated according to the integrated curated gene data and the new features to evaluate their utility. These TTV features may

potentially allow for increased precision when defining clusters and sub-clusters one of the primary premises of single cell techniques.

To generate single cell sequences, a single cell suspension must be generated from the biological specimen of interest and neighbouring cells become spatially randomised in an Eppendorf® tube. The loss of spatial context remains problematic and though methods are reported for spatial transcriptomics there is a balance to strike between cell resolution and mRNA capture efficiency. Here a method is introduced that can potentially reconstruct spatial context by using less than single cell resolved transcriptional profiles as prior information to pre-condition a tSNE landscape. In this way it may be possible to apply hybrid methods combining techniques with good spatial resolution and good capture efficiency but limited cellular resolution and single cell experiments with no spatial context to generate a spatially resolved single cell map. The ability to position additional points on tSNE also allows data to be added to a tSNE that has been previously extensively interpreted and annotated. Such methods to our knowledge are novel and have not been previously described in the literature in relation to single cell applications.

6.2 Tracing biologically plausible trajectories

The potential to use data on similarity between cells or cell populations, from snapshots at different developmental time points to trace ontogenic relationships in a pseudo-time is very appealing. Cells though are not simply defined by their transcriptional state, chromatin state including structure, DNA modifications, histone modifications, proteomic state, pathway activation and neighbouring cell signalling are just an example of additional signals driving cell state and fate decisions. A key assumption in this work is that at some point all will converge and influence transcriptional profile to achieve a change in cell state and that our assays have the fidelity to capture this.

In this vein several algorithms and software implementations attempting to reconstruct developmental journeys have been published. Some methods even date before single cell resolved transcriptional assays and relate to microarray data [Magwene et al., 2003]. The concept though is not new, morphological changes from static sections of embryos in placentals combined with a variety of marking strategies have historically led to many of our current insights into developmental biology.

Trapnell et al. [2014] described one of the earliest pseudotime construction methods designed specifically for single cell transcriptomic data, called Monocle. Following feature selection Monocle performs dimensionality reduction using independent component analysis (ICA)

Discussion

and a minimum spanning tree (MST) is fitted. The longest path through the MST is taken as the main journey and a starting cell can then be defined. The same first author has since presented a new trajectory inference algorithm named Monocle 2 [Qiu et al., 2017b]. This uses reverse graph embedding (RGE) to reconstruct trajectories in an unsupervised way. Feature selection remains critical and is performed by first using PCA dimensionality reduction followed by tSNE upon which clustering is performed based on density peaks. Differential expression between the clusters is used to identify key genes to be included for downstream analysis highlighting the utility of an iterative approach. The high dimensional gene expression data from the selected genes is used for dimensionality reduction which can be PCA, diffusion maps or an alternative. RGE is performed by automatically selecting centroids, generating a spanning tree and repositioning cells iteratively until the tree and cells converge. Qiu et al. [2017b] compare their technique to other state of the art single cell trajectory methods and perform favourably across multiple datasets including simulated, real wild-type neural and haematopoietic and perturbation data. Monocle 3 is the latest incarnation and is an optimised implementation of Monocle 2 with the workflow shown in fig. 6.1.

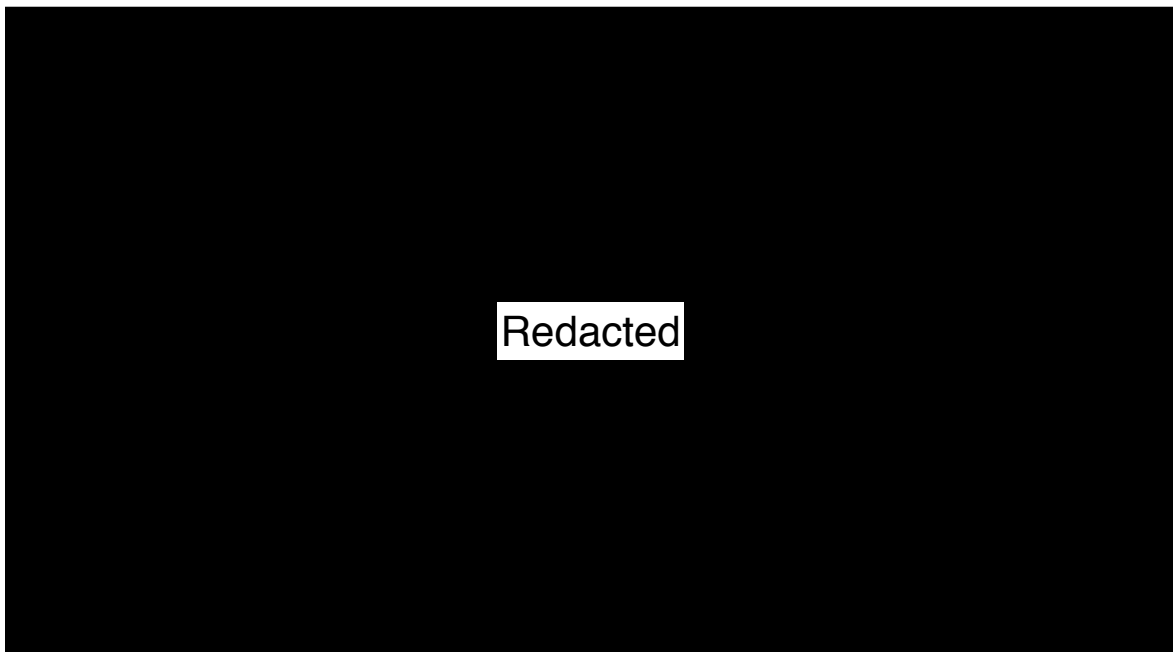


Fig. 6.1 The Monocle 3 workflow. From <http://cole-trapnell-lab.github.io/monocle-release/monocle3/>

Not long after Monocle was introduced, Bendall et al. [2014] described Wanderlust, a method designed for use with CyTOF mass spectrometry data that models non-bifurcating trajectories. Though Wanderlust pseudotime reconstruction can be adapted to transcriptomic

6.2 Tracing biologically plausible trajectories

data its inability to model multiple fates severely limits its utility [Ji and Ji, 2016]. Wanderlust generates an ensemble of graphs and uses waypoints to generate robust journeys [Bendall et al., 2014]. Using the Wanderlust pseudotime algorithm Bendall et al. [2014] were able to computationally reconstruct known hallmarks of human B lymphocyte development.

DeLorean a more recent algorithm uses a Gaussian process latent variable model with an apriori structure on the latent space selected to represent a pseudotime along developmental progression [Reid and Wernisch, 2016]. In this way repeated cross-sectional data can be re-ordered along a developmental time-course. DeLorean explicitly models gene expression changes along developmental progression as compared to other methods which infer cell orderings. A limitation of this method is the computational time which is prohibitive for large datasets.

Haghverdi et al. [2016] and Angerer et al. [2016] introduce Destiny an implementation using diffusion pseudotime to order cells. This builds on a single-cell dimensionality reduction and visualisation technique they introduced earlier [Haghverdi et al., 2015]. The underlying methods have been described in detail earlier in sections 3.10.1 and 3.10.2. Diffusion pseudotime is based on random walks on a graph, the same graph which forms the basis of diffusion maps visualisations.

An alternative to diffusion pseudotime that has been used in sections 3.10.2 and 3.10.3 is to use diffusion maps or alternative dimensionality reduction visualisation technique that captures a time-course process and use principal curves. For early embryonic blood development both diffusion pseudotime and principal curve fitting were consistent with one another fig. 3.33.

Though naïve unsupervised methods may seem attractive at first glance, alone they have limited usefulness. The individual components of trajectory and pseudotime estimation may be unsupervised but the data has been collected and experiment designed in a highly supervised manner in fact experiment design requires prior information and at the outset requires heavy user supervision to design experiments to collect the data. Even reported unsupervised methods focus on specific parts of the developmental process and are validated based on prior information or further experiments that require biological knowledge and expertise. Therefore in many respects naïve unsupervised methods have limitations on their utility and can only be interpreted within a supervised setting.

The package *roots* potentially allows faithful computational reconstruction of relationships between different cell populations as demonstrated in a hematopoietic dataset. This identifies a common megakaryocyte-erythroid precursor and supports a branching structure consistent

Discussion

with the current state of knowledge on haematopoiesis, fig. 5.3. This relationship was not clearly identified in other state of the art pseudotime methods utilised to analyse the same haematopoietic datasets. *roots* currently does not explicitly assign either a pseudotime or branch to cells but this can be achieved using principal curves after manually selecting branches.

One potential unexploited utility of ordering cells in time and inferring smoothed gene expression profiles along journeys is to identify a robust relationship between gene expression and drop-out, that can be used as a prior when inferring real expression from count data. Such relationships can be visualised in fig. 3.34. This may allow for improved models for inferring gene expression helping to reduce the difficulties described around drop-outs.

More recent methods have attempted to include isoform level data and shown that this improves accuracy when identifying developmentally regulated genes [Qiu et al., 2017a]. Another advance has been the use of unspliced transcripts to infer an RNA velocity that can help direct cells in trajectories [Manno et al., 2018]. Single-cell transcriptomics in this way can identify not only the topology and trajectories but also in a data-driven manner, it can infer the direction along trajectories that cells are likely to be travelling, potentially allowing quantification of flux.

6.3 Identifying novel molecular pathways

In addition to identifying novel sub-populations of cells, single cell transcriptomics potentially allows the identification of relationships between these populations. From the Smart-seq2 data of early gastrulation, unsupervised clustering revealed a coarse pattern of cell identities. Prior biological knowledge allowed for the recognition of a subpopulation of cells within the endothelial cluster that may indicate the emergence of the definitive wave of embryonic blood. To achieve this, pioneer genes known to be involved in this wave of haematopoiesis were identified and a set of genes highly correlated with these pioneer genes, within the endothelial population, were used to cluster the cells.

The combination of single-cell data with a knowledge driven approach, allowed the identification of evidence of activation of the leukotriene branch of the arachidonic acid pathway, fig. 3.40. Three genes in particular stood out including *Ltc4s*, *Alox5* and *Alox5ap* (FLAP) all involved in the leukotriene biosynthesis branch of the arachidonic acid metabolic pathway, table 3.9. Based on the in-silico identification of the leukotriene pathway, we postulated that it played a role in early haematopoiesis and this was subsequently validated in an in-vitro differentiation assay, using the leukotriene inhibitor Zileuton and LTC₄. Interestingly hydrox-

6.4 Computationally simulating developmental processes

urea, widely used in a clinical haematology setting and known to induce fetal haemoglobin is chemically related to Zileuton. This may uncover a new mechanism by which hydroxyurea and related compounds function.

Previous work involving large screens in zebrafish had suggested the involvement of the epoxyeicosatrienoic acid and prostaglandin branches of the arachidonic acid metabolic pathway in blood development but this is the first time the leukotriene branch has been implicated [Li et al., 2015; North et al., 2007].

The data set therefore provides a rich source of information for post-hoc in-silico experimentation. Identifying subgroups of cells using either unsupervised or supervised methods and then performing deeper analysis can identify novel pathways. This can help filter hypotheses and subsequently allow design of more resource intensive wet lab validation experiments.

6.4 Computationally simulating developmental processes

Having traced biologically plausible journeys, a more fundamental understanding of the molecular processes that regulate cell state transitions along these trajectories could reveal potential methods to direct in-vitro stem cell differentiation assays. Comparing inferred regulatory processes between normal and mutant models with known phenotypes may reveal deeper appreciation of how molecular pathways interact under normal and pathological conditions. The purpose of temporal alignment now becomes evident. Not only can associative relationships be inferred but causation can be established, based on the highly intuitive assumption that later events in time cannot cause earlier events but vice versa, the earlier event may either directly or indirectly cause a future event.

A commonly deployed method is the use of differential equations or for the case of discrete state transitions, difference equations. Though conceptually easy to grasp, vast numbers of features and large numbers of interactions at differing orders can create a complexity that can be difficult to solve, even with the huge numbers of cells that can be processed with the latest single cell technologies.

Reducing gene activity to a binary state of being either on or off allows classical Boolean logic rules to be inferred. These rules simply output whether the gene is on or off and cannot classically model different levels of gene activation [Lim et al., 2016; Müssel et al., 2010; Woodhouse et al., 2018]. This may not fit well with the contemporary paradigm in molecular biology of gene activity controlled through promoters and multiple enhancers.

Discussion

Correlation and partial correlation related methods also present intuitive alternative methods of generating networks. This method is introduced in the work and combined with a bootstrapped lasso to generate robust and sparse connectivity. Though frequently used and easy to interpret it is difficult to see how this simple model may faithfully reconstruct development gene profiles. But despite this it forms the basis of thinking about a more sophisticated model and this is described in detail in sections 5.4.2 and 5.5.

Correlation based methods of generating networks may be examined using an alternative viewpoint, as a layer of inputs fully connected to a layer of outputs with linear activation functions. These linear activation functions in a single layer are limited in the number and types of interactions that they can potentially simulate. Deeper neural networks with non-linear activation functions can simulate much more complex interactions and therefore potentially more faithfully reconstruct biological trajectories.

Neural networks provide a conceptually alternative approach, primarily modelling behaviour along trajectories rather than gene interactions themselves. In models based on differential equations, difference equations, Boolean rules and correlation networks the gene to gene interactions are explicitly parameterised and these parameters have an intuitive interpretation. In contrast the neural network used in this work attempts to faithfully reconstruct branching gene expression profiles along a given developmental process. Combining such a neural net with a stochastic output can enable faithful reconstruction of a branching developmental process, additionally allowing in-silico perturbations. In theory these in-silico experiments can include multi-gene perturbations timed at specific developmental stages along the trajectory and simulated in seconds, in contrast similar molecular perturbations would be impractical experimentally. Though these networks do not formally model interactions, in-silico experiments on the trained neural network can provide insights into the complex functions through simulation.

The neural network used was a feed-forward neural network with input generated from the previous stages output and built from the ground up. Though not a true recurrent network it bears considerable resemblance. Platforms now exist for much quicker implementations that harness the power of modern GPUs to allow highly parallelised training. Several frameworks such as Caffe, Google's Tensorflow and Theano exist that can allow higher level implementation of neural networks [Abadi et al., 2015; Al-Rfou et al., 2016; Jia et al., 2014]. These can be used to easily and quickly implement sophisticated recurrent networks which may include for example long short-term memory (LSTM) modules. LSTM modules allow the network to remember a previous state and allow historic events to influence later decisions [Hochreiter and Schmidhuber, 1997]. New ideas and implementations including

attentional interfaces and neural Turing machines are more sophisticated neural network architectures for recurrent networks that may perform even better than conventional LSTM recurrent networks [Graves et al., 2014; Xu et al., 2015]. These frameworks combined with the latest GPUs can likely allow training on a full RNAseq dataset rather than the qPCR data that was used here.

Ideally the work here would have been able to use these networks to identify novel molecular pathways. Contemporary pseudotime inference algorithms, the data generating experiments and the algorithms described herein are not sufficiently advanced to be able to predict and identify molecular pathways with sufficient fidelity. Despite these limitations single cell data combined with prior knowledge may reveal novel pathways as described in section 6.3.

6.5 Concluding remarks

This work has focused on using single cell resolved transcriptomic data to shed light on normal developmental time-courses, describe the expression profiles and try to identify the molecular signals that govern fate bifurcations. In summary the key advances reported in this work have been

1. Identifying and validating a novel pathway, the leukotriene branch of arachidonic acid metabolism, in early definitive wave embryonic blood development.
2. Revealing TTVs as a potential source of additional information for identifying cell types, aiding clustering and potentially helping with pseudotime reconstruction.
3. Developing a graph based method, *roots* for reconstructing ordered ontogenic trajectories from single cell transcriptomic data that was able to identify a common Megakaryocyte-erythroid progenitor that was not evident using current state of the art methods.
4. Adding functionality to the tSNE algorithm that allows new appropriately normalised data to be added to a previously constructed embedding. This has been taken further by showing that coarse bulk transcriptomic data may be used to reconstruct the sacrificed spatial context from single cell experiments. Analyses of cells from complex organisms all processed together can cause cells that are spatially distinct but with related functionality to be positioned very close or even amongst one another. Having prior knowledge and the appropriate data of spatial relations can incentivise this novel algorithm to position cells within a spacial context.

Discussion

5. Introducing a novel method of modelling dynamic systems by using pseudotime ordered transcriptomic data to train a neural network so that it can inherently learn the gene interactions allowing simulation of perturbation experiments.

Though this work has focused primarily on transcriptional assays some cell surface proteomic data was collected. Single cell multi-omics is a rapidly advancing field and novel technologies may soon be able to systematically collect a wide range of omics data including transcriptomics, genomic, epigenomic and proteomic data from the same cell. This may provide important further clues when inferring trajectories and developing in-silico models. As described with the finding of TTVs, even analysing the currently available transcriptomic data in more detail may provide rewarding novel insights.

Single cell data collection, development of novel algorithms and computational analyses to be believable and robust need to be reproducible. Reproducibility and replicability are not a modern problem in science but inability to reproduce analyses is a worrying problem. To aid with reproducibility the work here has been, where possible curated into R packages and submitted to the comprehensive R archiving network CRAN or is available through GitHub repositories as indicated in the text. Furthermore Docker containers are available for each chapter to reproduce all figures in this thesis.

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Abu-Elmagd, M., Mulvaney, J., and Wheeler, G. N. (2017). Frizzled-7 is required for *Xenopus* heart development. *Biol. Open*, 6(12):1861–8.
- Adachi, S. (2017). Rigid geometry solves “curse of dimensionality” effects in clustering methods: An application to omics data. *bioRxiv*, page 094391.
- Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Bleicher Snyder, J., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., de Brébisson, A., Breuleux, O., Carrier, P.-L., Cho, K., Chorowski, J., Christiano, P., Cooijmans, T., Côté, M.-A., Côté, M., Courville, A., Dauphin, Y. N., Delalleau, O., Demouth, J., Desjardins, G., Dieleman, S., Dinh, L., Ducoffe, M., Dumoulin, V., Ebrahimi Kahou, S., Erhan, D., Fan, Z., Firat, O., Germain, M., Glorot, X., Goodfellow, I., Graham, M., Gulcehre, C., Hamel, P., Harlouchet, I., Heng, J.-P., Hidas, B., Honari, S., Jain, A., Jean, S., Jia, K., Korobov, M., Kulkarni, V., Lamb, A., Lamblin, P., Larsen, E., Laurent, C., Lee, S., Lefrançois, S., Lemieux, S., Léonard, N., Lin, Z., Livezey, J. A., Lorenz, C., Lowin, J., Ma, Q., Manzagol, P.-A., Mastropietro, O., McGibbon, R. T., Memisevic, R., van Merriënboer, B., Michalski, V., Mirza, M., Orlandi, A., Pal, C., Pascanu, R., Pezeshki, M., Raffel, C., Renshaw, D., Rocklin, M., Romero, A., Roth, M., Sadowski, P., Salvatier, J., Savard, F., Schlüter, J., Schulman, J., Schwartz, G., Serban, I. V., Serdyuk, D., Shabaniyan, S., Simon, E., Spieckermann, S., Subramanyam, S. R., Sygnowski, J., Tanguay, J., van Tulder, G., Turian, J., Urban, S., Vincent, P., Visin, F., de Vries, H., Warde-Farley, D., Webb, D. J., Willson, M., Xu, K., Xue, L., Yao, L., Zhang, S., and Zhang, Y. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.
- Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.
- Anderson, H., Patch, T. C., Reddy, P. N. G., Hagedorn, E. J., Kim, P. G., Soltis, K. A., Chen, M. J., Tamplin, O. J., Frye, M., MacLean, G. A., Hübner, K., Bauer, D. E., Kanki, J. P.,

References

- Vogin, G., Huston, N. C., Nguyen, M., Fujiwara, Y., Paw, B. H., Vestweber, D., Zon, L. I., Orkin, S. H., Daley, G. Q., and Shah, D. I. (2015). Hematopoietic stem cells develop in the absence of endothelial *cadherin 5* expression. *Blood*, 126(26):2811–20.
- Andrews, S. (2010). FastQC a quality control tool for high throughput sequence data.
- Andrews, T. S. and Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Mol. Aspects Med.*, 59:114–22.
- Angerer, P., Haghverdi, L., Büttner, M., Theis, F. J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32:1241–3.
- Anraku, Y., Mizutani, R., and Satow, Y. (2008). Protein Splicing: Its Discovery and Structural Insight into Novel Chemical Mechanisms. *IUBMB Life*, 57(8):563–74.
- Arnoldi, W. E. (1951). The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Q Appl Math*, 9(1):17–29.
- Auclair, G., Guibert, S., Bender, A., and Weber, M. (2014). Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol.*, 15:545.
- Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M., and Kendzioriski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods*, 14(6):584–586.
- Bai, J., Wang, K., Li, Q., Yuan, Y., and Zhang, H. (2016). Pro-arrhythmogenic effects of *CACNA1C* G1911r mutation in human ventricular tachycardia: insights from cardiac multi-scale models. *Sci. Rep.*, 6:31262.
- Baldwin, H. S., Shen, H. M., Yan, H. C., DeLisser, H. M., Chung, A., Mickanin, C., Trask, T., Kirschbaum, N. E., Newman, P. J., and Albelda, S. M. (1994). Platelet endothelial cell adhesion molecule-1 (PECAM-1/CD31): alternatively spliced, functionally distinct isoforms expressed during mammalian cardiovascular development. *Development*, 120(9):2539–2553.
- Bar-Joseph, Z., Gifford, D. K., and Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl 1):S22–S29.
- Barker, N., van Es, J. H., Kuipers, J., Kujala, P., van den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P. J., and Clevers, H. (2007). Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature*, 449:1003–7.
- Batta, K., Florkowska, M., Kouskoff, V., and Lacaud, G. (2014). Direct Reprogramming of Murine Fibroblasts to Hematopoietic Progenitor Cells. *Cell Reports*, 9(5):1871–1884.
- Battista, M. C., Oligny, L. L., St-Louis, J., and Brochu, M. (2002). Intrauterine growth restriction in rats is associated with hypertension and renal dysfunction in adulthood. *Am. J. Physiol. Endocrinol. Metab.*, 283:E124–31.
- Becht, E., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2018). Evaluation of UMAP as an alternative to t-SNE for single-cell data. *bioRxiv*, page 298430.
- Becker, M.-B., Zülch, A., Bosse, A., and Gruss, P. (2001). *Irx1* and *irx2* expression in early lung development. *Mech. Dev.*, 106(1):155–158.

- Beddington, R. S. P., Rashbass, P., and Wilson, V. (1992). Brachyury-a gene affecting mouse gastrulation and early organogenesis. *Development*, 116(Supplement):157–165.
- Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D. C., Tamuri, A. U., Martincorena, I., Petljak, M., Alexandrov, L. B., Gundem, G., Tarpey, P. S., Roerink, S., Blokker, J., Maddison, M., Mudie, L., Robinson, B., Nik-Zainal, S., Campbell, P., Goldman, N., van de Wetering, M., Cuppen, E., Clevers, H., and Stratton, M. R. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–5.
- Bell, C. C., Amaral, P. P., Kalsbeek, A., Magor, G. W., Gillinder, K. R., Tangermann, P., Lisio, L. d., Cheetham, S. W., Gruhl, F., Frith, J., Tallack, M. R., Ru, K.-L., Crawford, J., Mattick, J. S., Dinger, M. E., and Perkins, A. C. (2016). The *Evx1/Evx1as* gene locus regulates anterior-posterior patterning during gastrulation. *Sci. Rep.*, 6:26657.
- Bendall, S. C., Davis, K. L., Amir, e. l. . A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157:714–25.
- Bertrand, N., Roux, M., Ryckebusch, L., Niederreither, K., Dollé, P., Moon, A., Capecchi, M., and Zaffran, S. (2011). Hox genes define distinct progenitor sub-domains within the second heart field. *Dev. Biol.*, 353(2):266–74.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10):P10008.
- Bodea, G. O., McKelvey, E. G. Z., and Faulkner, G. J. (2018). Retrotransposon-induced mosaicism in the neural genome. *Open Biol.*, 8(7):180074.
- Bonachea, E. M., Chang, S.-W., Zender, G., LaHaye, S., Fitzgerald-Butt, S., McBride, K. L., and Garg, V. (2014). Rare *GATA5* sequence variants identified in individuals with bicuspid aortic valve. *Pediatr. Res.*, 76(2):211–216.
- Bonnard, C., Strobl, A. C., Shboul, M., Lee, H., Merriman, B., Nelson, S. F., Ababneh, O. H., Uz, E., Güran, T., Kayserili, H., Hamamy, H., and Reversade, B. (2012). Mutations in *IRX5* impair craniofacial development and germ cell migration via SDF1. *Nat. Genet.*, 44(6):709–13.
- Boyl, P. P., Signore, M., Acampora, D., Martinez-Barbera, J. P., Ilengo, C., Annino, A., Corte, G., and Simeone, A. (2001). Forebrain and midbrain development requires epiblast-restricted *Otx2* translational control mediated by its 3'UTR. *Development*, 128(15):2989–3000.
- Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10:1093–5.
- Briggs, J. A., Li, V. C., Lee, S., Woolf, C. J., Klein, A., and Kirschner, M. W. (2017). Mouse embryonic stem cells can differentiate via multiple paths to the same state. *bioRxiv*, page 124594.

References

- Brons, I. G. M., Smithers, L. E., Trotter, M. W. B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S. M., Howlett, S. K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R. A., and Vallier, L. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, 448(7150):191–5.
- Brownlee, E. M., Howatson, A. G., Davis, C. F., and Sabharwal, A. J. (2009). The hidden mortality of congenital diaphragmatic hernia: a 20-year review. *J. Pediatr. Surg.*, 44:317–20.
- Bruneau, B. G., Bao, Z.-Z., Tanaka, M., Schott, J.-J., Izumo, S., Cepko, C. L., Seidman, J. G., and Seidman, C. E. (2000). Cardiac Expression of the Ventricle-Specific Homeobox Gene *Irx4* Is Modulated by *Nkx2-5* and *dHand*. *Dev. Biol.*, 217(2):266–277.
- Bruneau, B. G., Nemer, G., Schmitt, J. P., Charron, F., Robitaille, L., Caron, S., Conner, D. A., Gessler, M., Nemer, M., Seidman, C. E., and Seidman, J. G. (2001). A Murine Model of Holt-Oram Syndrome Defines Roles of the T-Box Transcription Factor *Tbx5* in Cardiogenesis and Disease. *Cell*, 106(6):709–721.
- Bunting, M., Bernstein, K. E., Greer, J. M., Capecchi, M. R., and Thomas, K. R. (1999). Targeting genes for self-excision in the germ line. *Genes Dev.*, 13(12):1524–1528.
- Burdsal, C. A., Damsky, C. H., and Pedersen, R. A. (1993). The role of E-cadherin and integrins in mesoderm differentiation and migration at the mammalian primitive streak. *Development*, 118(3):829–844.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36(5):411–20.
- Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., and Walsh, C. A. (2014). Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Rep.*, 8(5):1280–9.
- Campbell, K. H., McWhir, J., Ritchie, W. A., and Wilmut, I. (1996). Sheep cloned by nuclear transfer from a cultured cell line. *Nature*, 380:64–6.
- Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–7.
- Capron, C., Lécluse, Y., Kaushik, A. L., Foudi, A., Lacout, C., Sekkai, D., Godin, I., Albagli, O., Poullion, I., Svinartchouk, F., Schanze, E., Vainchenker, W., Sablitzky, F., Bennaceur-Griscelli, A., and Duménil, D. (2006). The SCL relative *LYL-1* is required for fetal and adult hematopoietic stem cell function and B-cell differentiation. *Blood*, 107(12):4678–4686.
- Casanova, E., Lemberger, T., Fehsenfeld, S., Mantamadiotis, T., and Schütz, G. (2003). α Complementation in the Cre recombinase enzyme. *Genesis*, 37(1):25–9.
- Chao, J., Shen, B., Gao, L., Xia, C.-F., Bledsoe, G., and Chao, L. (2010). Tissue kallikrein in cardiovascular, cerebrovascular and renal diseases and skin wound healing. *Biol. Chem.*, 391(4):345–355.

- Chen, F., Kook, H., Milewski, R., Gitler, A. D., Lu, M. M., Li, J., Nazarian, R., Schnepf, R., Jen, K., Biben, C., Runke, G., Mackay, J. P., Novotny, J., Schwartz, R. J., Harvey, R. P., Mullins, M. C., and Epstein, J. A. (2002). Hop Is an Unusual Homeobox Gene that Modulates Cardiac Development. *Cell*, 110(6):713–23.
- Chen, J. Y., Miyanishi, M., Wang, S. K., Yamazaki, S., Sinha, R., Kao, K. S., Seita, J., Sahoo, D., Nakauchi, H., and Weissman, I. L. (2016). Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature*, 530(7589):223–227.
- Christoffels, V. M., Keijser, A. G. M., Houweling, A. C., Clout, D. E. W., and Moorman, A. F. M. (2000). Patterning the Embryonic Heart: Identification of Five Mouse Iroquois Homeobox Genes in the Developing Heart. *Dev. Biol.*, 224(2):263–74.
- Chuang, H.-N., Cheng, H.-Y., Hsiao, K.-M., Lin, C.-W., Lin, M.-L., and Pan, H. (2018). The zebrafish homeobox gene *irx11* is required for brain and pharyngeal arch morphogenesis. *Dev. Dyn.*, 239(2):639–50.
- Ciau-Uitz, A., Monteiro, R., Kirmizitas, A., and Patient, R. (2014). Developmental hematopoiesis: ontogeny, genetic programming and conservation. *Exp. Hematol.*, 42:669–83.
- Cibelli, J. (2007). Developmental biology. A decade of cloning mystique. *Science*, 316:990–2.
- Ciruna, B. and Rossant, J. (2001). FGF signaling regulates mesoderm cell fate specification and morphogenetic movement at the primitive streak. *Dev. Cell*, 1:37–49.
- Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, 38(6):1767–1771.
- Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.*, 102(21):7426–7431.
- Colombo, S., Sena-Tomás, C. d., George, V., Werdich, A. A., Kapur, S., MacRae, C. A., and Targoff, K. L. (2018). Nkx genes establish second heart field cardiomyocyte progenitors at the arterial pole and pattern the venous pole through *Isl1* repression. *Development*, 145(3):dev161497.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.*, 17.
- Cook, J., Sutskever, I., Mnih, A., and Hinton, G. (2007). Visualizing similarity data with a mixture of maps. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 67–74, San Juan, Puerto Rico. PMLR.
- Coulombel, L. (2004). Identification of hematopoietic stem/progenitor cells: strength and drawbacks of functional assays. *Oncogene*, 23(43):7210–7222.
- Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

References

- Cyranoski, D. (2014). Japanese woman is first recipient of next-generation stem cells. *Nature*.
- Dabelea, D., Hanson, R. L., Lindsay, R. S., Pettitt, D. J., Imperatore, G., Gabir, M. M., Roumain, J., Bennett, P. H., and Knowler, W. C. (2000). Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: a study of discordant sibships. *Diabetes*, 49:2208–11.
- Dabelea, D., Mayer-Davis, E. J., Lamichhane, A. P., D’Agostino, R. B., Liese, A. D., Vehik, K. S., Narayan, K. M., Zeitler, P., and Hamman, R. F. (2008). Association of intrauterine exposure to maternal diabetes and obesity with type 2 diabetes in youth: the SEARCH Case-Control Study. *Diabetes Care*, 31:1422–6.
- Devalla, H. D., Schwach, V., Ford, J. W., Milnes, J. T., El-Haou, S., Jackson, C., Gkatzis, K., Elliott, D. A., Lopes, S. M. C. d. S., Mummery, C. L., Verkerk, A. O., and Passier, R. (2015). Atrial-like cardiomyocytes from human pluripotent stem cells are a robust preclinical model for assessing atrial-selective pharmacology. *EMBO Mol. Med.*, 7(4):394–410.
- Dieterlen-Lievre, F. (1975). On the origin of haemopoietic stem cells in the avian embryo: an experimental approach. *J Embryol Exp Morphol*, 33:607–19.
- Diez del Corral, R. and Morales, A. V. (2017). The Multiple Roles of FGF Signaling in the Developing Spinal Cord. *Front. Cell Dev. Biol.*, 5.
- Ding, B., Zheng, L., and Wang, W. (2016). Assessment of single cell RNA-seq normalization methods. *bioRxiv*, page 064329.
- Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.*, 9(1):2002.
- Ding, J., Yang, L., Yan, Y.-T., Chen, A., Desai, N., Wynshaw-Boris, A., and Shen, M. M. (1998). *Cripto* is required for correct orientation of the anterior–posterior axis in the mouse embryo. *Nature*, 395(6703):27215.
- Dobrovolskaïa-Zavadskaïa, N (1927). Sur la mortification spontanée de la queue chez la souris nouveau-née et sur l’existence d’un caractère hereditaire “non-viable”. *C. R. Soc. Biol. Paris*, 97:114–6.
- Donnison, M., Beaton, A., Davey, H. W., Broadhurst, R., L’Huillier, P., and Pfeffer, P. L. (2005). Loss of the extraembryonic ectoderm in *Elf5* mutants leads to defects in embryonic patterning. *Development*, 132:2299–308.
- Downs, K. M. and Davies, T. (1993). Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development*, 118(4):1255–66.
- Drissen, R., von Lindern, M., Kolbus, A., Driegen, S., Steinlein, P., Beug, H., Grosveld, F., and Philipsen, S. (2005). The erythroid phenotype of *EKLF*-null mice: defects in hemoglobin metabolism and membrane stability. *Mol. Cell. Biol.*, 25(12):5205–5214.
- Du, J., Takeuchi, H., Leonhard-Melief, C., Shroyer, K. R., Dlugosz, M., Haltiwanger, R. S., and Holdener, B. C. (2010). O-fucosylation of thrombospondin type 1 repeats restricts epithelial to mesenchymal transition (EMT) and maintains epiblast pluripotency during mouse gastrulation. *Dev. Biol.*, 346(1):25–38.

- Dumon, S., Walton, D. S., Volpe, G., Wilson, N., Dassé, E., Pozzo, W. D., Landry, J.-R., Turner, B., O'Neill, L. P., Göttgens, B., and Frampton, J. (2012). Itga2b Regulation at the Onset of Definitive Hematopoiesis and Commitment to Differentiation. *PLoS One*, 7(8):e43300.
- Dzierzak, E. and Medvinsky, A. (1995). Mouse embryonic hematopoiesis. *Trends Genet.*, 11:359–66.
- Eaton, J. W., Bateman, D., Hauberg, S., and Wehbring, R. (2015). *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*. GNU Octave.
- Elefanty, A. G., Begley, C. G., Metcalf, D., Barnett, L., Köntgen, F., and Robb, L. (1998). Characterization of hematopoietic progenitor cells that express the transcription factor SCL, using a lacZ “knock-in” strategy. *Proc. Natl. Acad. Sci. U.S.A.*, 95(20):11897–11902.
- Elizondo, D. M., Andargie, T. E., Lee, C. M., Anderson, W. A., and Lipscomb, M. W. (2017). AIF1 expression regulates differentiation of dendritic cells from hematopoietic stem cell multipotent progenitor subsets and regulates antigen presentation capacity. *J. Immunol.*, 198(1 Supplement):202.3–202.3.
- Evseenko, D., Zhu, Y., Schenke-Layland, K., Kuo, J., Latour, B., Ge, S., Scholes, J., Dravid, G., Li, X., MacLellan, W. R., and Crooks, G. M. (2010). Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proc Natl Acad Sci U S A*, 107(31):13742–13747.
- Fadler, K. M. and Askin, D. F. (2008). Sacrococcygeal teratoma in the newborn: a case study of prenatal management and clinical intervention. *Neonatal Netw*, 27:185–91.
- Faulkner, G. J. and Billon, V. (2018). L1 retrotransposition in the soma: a field jumping ahead. *Mobile DNA*, 9(1):22.
- Feil, R., Wagner, J., Metzger, D., and Chambon, P. (1997). Regulation of Cre Recombinase Activity by Mutated Estrogen Receptor Ligand-Binding Domains. *Biochem. Biophys. Res. Commun.*, 237(3):752–757.
- Ferkowicz, M. J., Starr, M., Xie, X., Li, W., Johnson, S. A., Shelley, W. C., Morrison, P. R., and Yoder, M. C. (2003). CD41 expression defines the onset of primitive and definitive hematopoiesis in the murine embryo. *Development*, 130(18):4393–4403.
- Finkensieper, A., Kieser, S., Bekhite, M. M., Richter, M., Mueller, J. P., Graebner, R., Figulla, H.-R., Sauer, H., and Wartenberg, M. (2010). The 5-lipoxygenase pathway regulates vasculogenesis in differentiating mouse embryonic stem cells. *Cardiovasc. Res.*, 86(1):37–44.
- Fiorenzano, A., Pascale, E., D'Aniello, C., Acampora, D., Bassalart, C., Russo, F., Andolfi, G., Biffoni, M., Francescangeli, F., Zeuner, A., Angelini, C., Chazaud, C., Patriarca, E. J., Fico, A., and Minchiotti, G. (2016). Cripto is essential to capture mouse epiblast stem cell and human embryonic stem cell pluripotency. *Nat. Commun.*, 7:12589.
- Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugenics*, 7(2):179–88.
- Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software Pract. Exper*, 21(11):1129–64.

References

- Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U., and Shapiro, E. (2005). Genomic Variability within an Organism Exposes Its Cell Lineage Tree. *PLoS Comput. Biol.*, 1(5):e50.
- Fujii, M., Sakaguchi, A., Kamata, R., Nagao, M., Kikuchi, Y., Evans, S. M., Yoshizumi, M., Shimono, A., Saga, Y., and Kokubo, H. (2017). *Sfrp5* identifies murine cardiac progenitors for all myocardial structures except for the right ventricle. *Nat. Commun.*, 8:14664.
- Gluecksohn-Schoenheimer, S. (1944). The Development of Normal and Homozygous Brachy (T/T) Mouse Embryos in the Extraembryonic Coelom of the Chick. *Proc. Natl. Acad. Sci. U.S.A.*, 30:134–40.
- Goldmuntz, E., Bamford, R., Karkera, J. D., dela Cruz, J., Roessler, E., and Muenke, M. (2002). CFC1 Mutations in Patients with Transposition of the Great Arteries and Double-Outlet Right Ventricle. *Am. J. Hum. Genet.*, 70(3):776–780.
- Gonfiotti, A., Jaus, M. O., Barale, D., Baiguera, S., Comin, C., Lavorini, F., Fontana, G., Sibila, O., Rombolà, G., Jungebluth, P., and Macchiarini, P. (2014). The first tissue-engineered airway transplantation: 5-year follow-up results. *Lancet*, 383:238–44.
- Goode, D. K., Obier, N., Vijayabaskar, M. S., Lie-A-Ling, M., Lilly, A. J., Hannah, R., Lichtinger, M., Batta, K., Florkowska, M., Patel, R., Challinor, M., Wallace, K., Gilmour, J., Assi, S. A., Cauchy, P., Hoogenkamp, M., Westhead, D. R., Lacaud, G., Kouskoff, V., Göttgens, B., and Bonifer, C. (2016). Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Dev. Cell*, 36(5):572–587.
- Gordon, E. J., Gale, N. W., and Harvey, N. L. (2008). Expression of the hyaluronan receptor LYVE-1 is not restricted to the lymphatic vasculature; LYVE-1 is also expressed on embryonic blood vessels. *Dev. Dyn.*, 237(7):1901–1909.
- Graves, A., Wayne, G., and Danihelka, I. (2014). Neural Turing Machines. *arXiv:1410.5401 [cs]*. arXiv: 1410.5401.
- Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525:251–5.
- Grün, D. and van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. *Cell*, 163(4):799–810.
- Gurdon, J. B., Laskey, R. A., and Reeves, O. R. (1975). The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs. *J Embryol Exp Morphol*, 34:93–112.
- Götz, J., Probst, A., Mistl, C., Nitsch, R. M., and Ehler, E. (2000). Distinct role of protein phosphatase 2a subunit C α in the regulation of E-cadherin and β -catenin during development. *Mech. Dev.*, 93(1):83–93.
- Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J. J., Hession, C., Zhang, F., and Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*, 353:925–8.
- Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998.

- Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods*, 13(10):845–848.
- Hakre, S., Tussie-Luna, M. I., Ashworth, T., Novina, C. D., Settleman, J., Sharp, P. A., and Roy, A. L. (2006). Opposing functions of TFII-I spliced isoforms in growth factor-induced gene expression. *Mol. Cell*, 24(2):301–8.
- Harrison, M. R., Bjordal, R. I., Langmark, F., and Knutrud, O. (1978). Congenital diaphragmatic hernia: the hidden mortality. *J. Pediatr. Surg.*, 13:227–30.
- Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673.
- Hastie, T. and Stuetzle, W. (1989). Principal Curves. *J. Am. Stat. Assoc.*, 84(406):502.
- Hatano, S., Yamashita, T., Sekiguchi, A., Iwasaki, Y., Nakazawa, K., Sagara, K., Inuma, H., Aizawa, T., and Fu, L.-T. (2006). Molecular and Electrophysiological Differences in the L-Type Ca²⁺ Channel of the Atrium and Ventricle of Rat Hearts. *Circ. J.*, 70(5):610–4.
- Haynes, J., Baliga, B. S., Obiako, B., Ofori-Acquah, S., and Pace, B. (2004). Zileuton induces hemoglobin F synthesis in erythroid progenitors: role of the L-arginine–nitric oxide signaling pathway. *Blood*, 103(10):3945–3950.
- Hazen, J. L., Faust, G. G., Rodriguez, A. R., Ferguson, W. C., Shumilina, S., Clark, R. A., Boland, M. J., Martin, G., Chubukov, P., Tsunemoto, R. K., Torkamani, A., Kupriyanov, S., Hall, I. M., and Baldwin, K. K. (2016). The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning. *Neuron*, 89(6):1223–36.
- Hempel, M., Casar Tena, T., Diehl, T., Burczyk, M. S., Strom, T. M., Kubisch, C., Philipp, M., and Lessel, D. (2017). Compound heterozygous GATA5 mutations in a girl with hydrops fetalis, congenital heart defects and genital anomalies. *Hum. Genet.*, 136:339–346.
- Hendrickson, B. and Leland, R. (1995). A Multilevel Algorithm for Partitioning Graphs. In *Proceedings of the 1995 ACM/IEEE Conference on Supercomputing*, Supercomputing '95, New York, NY, USA. ACM.
- Henrique, D., Abranches, E., Verrier, L., and Storey, K. G. (2015). Neuromesodermal progenitors and the making of the spinal cord. *Development*, 142(17):2864–75.
- Herrmann, B. G., Labeit, S., Poustka, A., King, T. R., and Lehrach, H. (1990). Cloning of the T gene required in mesoderm formation in the mouse. *Nature*, 343(6259):617–622.
- Hinton, G. and Roweis, S. (2002). Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pages 857–64, Cambridge, MA, USA. MIT Press.
- Hochedlinger, K. and Jaenisch, R. (2002). Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature*, 415:1035–8.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

References

- Hou, J., Charters, A. M., Lee, S. C., Zhao, Y., Wu, M. K., Jones, S. J., Marra, M. A., and Hoodless, P. A. (2007). A systematic screen for genes expressed in definitive endoderm by Serial Analysis of Gene Expression (SAGE). *BMC Dev. Biol.*, 7:92.
- Hsu, Y.-C. (2015). Theory and Practice of Lineage Tracing: Theory and Practice of Lineage Tracing. *Stem Cells*, 33(11):3197–204.
- Huang, R.-T., Xue, S., Wang, J., Gu, J.-Y., Xu, J.-H., Li, Y.-J., Li, N., Yang, X.-X., Liu, H., Zhang, X.-D., Qu, X.-K., Xu, Y.-J., Qiu, X.-B., Li, R.-G., and Yang, Y.-Q. (2016). CASZ1 loss-of-function mutation associated with congenital heart disease. *Gene*, 595(1):62–8.
- Ibarra-Soria, X., Jawaid, W., Pijuan-Sala, B., Ladopoulos, V., Scialdone, A., Jörg, D. J., Tyser, R. C. V., Calero-Nieto, F. J., Mulas, C., Nichols, J., Vallier, L., Srinivas, S., Simons, B. D., Göttgens, B., and Marioni, J. C. (2018). Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.*, 20(2):127–134.
- Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods*, 13:577–80.
- Imuta, Y., Kiyonari, H., Jang, C.-W., Behringer, R. R., and Sasaki, H. (2013). Generation of knock-in mice that express nuclear enhanced green fluorescent protein and tamoxifen-inducible Cre recombinase in the notochord from *Foxa2* and *T* loci: Mouse Lines Expressing EGFP and Inducible CRE. *Genesis*, 51(3):210–218.
- Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lonnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167.
- Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(6172):776–779.
- Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7):1883–1896.e15.
- Jawaid, W., Chan, B., and Jesudason, E. C. (2012). Subspecialization may improve an esophageal atresia service but has not addressed declining trainee experience. *J. Pediatr. Surg.*, 47:1363–8.
- Jawaid, W. B., Qasem, E., Jones, M. O., Shaw, N. J., and Losty, P. D. (2013). Outcomes following prosthetic patch repair in newborns with congenital diaphragmatic hernia. *Br J Surg*, 100:1833–7.
- Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, 44(13):e117.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

- Junker, J. P., Spanjaard, B., Peterson-Maduro, J., Alemany, A., Hu, B., Florescu, M., and Oudenaarden, A. v. (2016). Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars. *bioRxiv*, page 056499.
- Kaati, G., Bygren, L. O., and Edvinsson, S. (2002). Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *Eur. J. Hum. Genet.*, 10:682–8.
- Kabrun, N., Buhring, H. J., Choi, K., Ullrich, A., Risau, W., and Keller, G. (1997). Flk-1 expression defines a population of early embryonic hematopoietic precursors. *Development*, 124(10):2039–2048.
- Kaimakis, P., de Pater, E., Eich, C., Solaimani Kartalaei, P., Kauts, M.-L., Vink, C. S., van der Linden, R., Jaegle, M., Yokomizo, T., Meijer, D., and Dzierzak, E. (2016). Functional and molecular characterization of mouse Gata2-independent hematopoietic progenitors. *Blood*, 127(11):1426–1437.
- Kallianpur, A. R., Jordan, J. E., and Brandt, S. J. (1994). The SCL/TAL-1 gene is expressed in progenitors of both the hematopoietic and vascular systems during embryogenesis. *Blood*, 83(5):1200–1208.
- Kaltenbach, S. L., Holland, L. Z., Holland, N. D., and Koop, D. (2009). Developmental expression of the three iroquois genes of amphioxus (BfIrxA, BfIrxB, and BfIrxC) with special attention to the gastrula organizer and anteroposterior boundaries in the central nervous system. *Gene Expr. Patterns*, 9(5):329–334.
- Kattman, S. J., Huber, T. L., and Keller, G. M. (2006). Multipotent flk-1+ cardiovascular progenitor cells give rise to the cardiomyocyte, endothelial, and vascular smooth muscle lineages. *Dev. Cell*, 11:723–32.
- Kelly, R. G., Buckingham, M. E., and Moorman, A. F. (2014). Heart Fields and Cardiac Morphogenesis. *Cold Spring Harb. Perspect. Med.*, 4(10).
- Kemp, C., Willems, E., Abdo, S., Lambiv, L., and Leyns, L. (2005). Expression of all Wnt genes and their secreted antagonists during mouse blastocyst and postimplantation development. *Dev. Dyn.*, 233(3):1064–1075.
- Kester, L. and Oudenaarden, A. v. (2018). Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*, 23(2):166–79.
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, 11(7):740–742.
- Khoa, L. T. P., Azami, T., Tsukiyama, T., Matsushita, J., Tsukiyama-Fujii, S., Takahashi, S., and Ema, M. (2016). Visualization of the Epiblast and Visceral Endodermal Cells Using Fgf5-P2a-Venus BAC Transgenic Mice and Epiblast Stem Cells. *PLoS One*, 11(7):e0159246.
- Kimura, C., Shen, M. M., Takeda, N., Aizawa, S., and Matsuo, I. (2001). Complementary Functions of Otx2 and Cripto in Initial Patterning of Mouse Epiblast. *Dev. Biol.*, 235(1):12–32.

References

- Klein, A., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D., and Kirschner, M. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201.
- Kotewicz, M. L., Sampson, C. M., D’Alessio, J. M., and Gerard, G. F. (1988). Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Res.*, 16(1):265–277.
- Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*.
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma’ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 44(Web Server issue):W90–W97.
- Kurek, K. C., Luks, V. L., Ayturk, U. M., Alomari, A. I., Fishman, S. J., Spencer, S. A., Mulliken, J. B., Bowen, M. E., Yamamoto, G. L., Kozakewich, H. P., and Warman, M. L. (2012). Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. *Am. J. Hum. Genet.*, 90:1108–15.
- Kwon, G. S., Viotti, M., and Hadjantonakis, A.-K. (2008). The Endoderm of the Mouse Embryo Arises by Dynamic Widespread Intercalation of Embryonic and Extraembryonic Lineages. *Dev. Cell*, 15(4):509–520.
- Lai, S., Xu, Y., Huang, W., Jiang, M., Chen, H., Ye, F., Wang, R., Qiu, Y., Jiang, X., Huang, D., Mao, J., Li, Y., Lu, Y., Xie, J., Fang, Q., Li, T., Huang, H., Han, X., and Guo, G. (2017). Mapping Human Hematopoietic Hierarchy At Single Cell Resolution By Microwell-seq. *bioRxiv*, page 127217.
- Lancrin, C., Sroczynska, P., Stephenson, C., Allen, T., Kouskoff, V., and Lacaud, G. (2009). The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature*, 457(7231):892–895.
- Lancôt, C., Lamolet, B., and Drouin, J. (1997). The bicoid-related homeoprotein Ptx1 defines the most anterior domain of the embryo and differentiates posterior from anterior lateral mesoderm. *Development*, 124(14):2807–2817.
- Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Stand.*, 45(4):255.
- Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24:719–20.
- Lania, G., Ferrentino, R., and Baldini, A. (2015). TBX1 Represses Vegfr2 Gene Expression and Enhances the Cardiac Fate of VEGFR2+ Cells. *PLoS One*, 10(9):e0138525.
- Laurent, F., Girdziusaitė, A., Gamart, J., Barozzi, I., Osterwalder, M., Akiyama, J. A., Lincoln, J., Lopez-Rios, J., Visel, A., Zuniga, A., and Zeller, R. (2017). HAND2 Target Gene Regulatory Networks Control Atrioventricular Canal and Cardiac Valve Development. *Cell Reports*, 19(8):1602–1613.
- Le Douarin, N. (1969). [Details of the interphase nucleus in Japanese quail (*Coturnix coturnix japonica*)]. *Bull. Biol. Fr. Belg.*, 103(3):435–52.

- Le Douarin, N. M. (1973). A Feulgen-positive nucleolus. *Exp. Cell Res.*, 77:459–68.
- Le Douarin, N. M. and Teillet, M. A. (1973). The migration of neural crest cells to the wall of the digestive tract in avian embryo. *J Embryol Exp Morphol*, 30:31–48.
- Lescroart, F., Chabab, S., Lin, X., Rulands, S., Paulissen, C., Rodolosse, A., Auer, H., Achouri, Y., Dubois, C., Bondue, A., Simons, B. D., and Blanpain, C. (2014). Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development. *Nat. Cell Biol.*, 16:829–40.
- Li, P., Lahvic, J. L., Binder, V., Pugach, E. K., Riley, E. B., Tamplin, O. J., Panigrahy, D., Bowman, T. V., Barrett, F. G., Heffner, G. C., McKinney-Freeman, S., Schlaeger, T. M., Daley, G. Q., Zeldin, D. C., and Zon, L. I. (2015). Epoxyeicosatrienoic acids enhance embryonic haematopoiesis and adult marrow engraftment. *Nature*, 523(7561):468–471.
- Lickert, H., Cox, B., Wehrle, C., Taketo, M. M., Kemler, R., and Rossant, J. (2005). Dissecting Wnt/ β -catenin signaling during gastrulation using RNA interference in mouse embryos. *Development*, 132(11):2599–2609.
- Lilly, A. J., Mazan, A., Scott, D. A., Lacaud, G., and Kouskoff, V. (2017). SOX7 expression is critically required in FLK1-expressing cells for vasculogenesis and angiogenesis during mouse embryonic development. *Mech. Dev.*, 146:31–41.
- Lim, C. Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., and Götting, B. (2016). BTR: training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics*, 17(1):355.
- Lindhurst, M. J., Sapp, J. C., Teer, J. K., Johnston, J. J., Finn, E. M., Peters, K., Turner, J., Cannons, J. L., Bick, D., Blakemore, L., Blumhorst, C., Brockmann, K., Calder, P., Cherman, N., Deardorff, M. A., Everman, D. B., Golas, G., Greenstein, R. M., Kato, B. M., Keppler-Noreuil, K. M., Kuznetsov, S. A., Miyamoto, R. T., Newman, K., Ng, D., O'Brien, K., Rothenberg, S., Schwartzenruber, D. J., Singhal, V., Tirabosco, R., Upton, J., Wientroub, S., Zackai, E. H., Hoag, K., Whitewood-Neal, T., Robey, P. G., Schwartzberg, P. L., Darling, T. N., Tosi, L. L., Mullikin, J. C., and Biesecker, L. G. (2011). A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N. Engl. J. Med.*, 365:611–9.
- Livet, J., Weissman, T. A., Kang, H., Draft, R. W., Lu, J., Bennis, R. A., Sanes, J. R., and Lichtman, J. W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450:56–62.
- Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D’Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J., and Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–8.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15:550.
- Lun, A. T., Bach, K., and Marioni, J. C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75.

References

- Lun, A. T., McCarthy, D. J., and Marioni, J. C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Res.*, 5:2122.
- Maaten, L. v. d. (2009). Learning a Parametric Embedding by Preserving Local Structure. In *Artificial Intelligence and Statistics*, pages 384–91.
- Maaten, L. v. d. (2014). Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*, 15:3221–45.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing Data using t-SNE. *J Mach. Learn. Res.*, 9(Nov):2579–605.
- Macchiarini, P., Jungebluth, P., Go, T., Asnaghi, M. A., Rees, L. E., Cogan, T. A., Dodson, A., Martorell, J., Bellini, S., Parnigotto, P. P., Dickinson, S. C., Hollander, A. P., Mantero, S., Conconi, M. T., and Birchall, M. A. (2008). Clinical transplantation of a tissue-engineered airway. *Lancet*, 372:2023–30.
- Macosko, E., Basu, A., Satija, R., Nemes, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A., Kamitaki, N., Martersteck, E., Trombetta, J., Weitz, D., Sanes, J., Shalek, A., Regev, A., and McCarroll, S. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.
- Magwene, P. M., Lizardi, P., and Kim, J. (2003). Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, 19:842–50.
- Manno, G. L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriiti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., Bruggen, D. v., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719):494.
- Marinovic-Terzic, I., Yoshioka-Yamashita, A., Shimodaira, H., Avdievich, E., Hunton, I. C., Kolodner, R. D., Edelmann, W., and Wang, J. Y. J. (2008). Apoptotic function of human PMS2 compromised by the nonsynonymous single-nucleotide polymorphic variant R20Q. *Proc. Natl. Acad. Sci. U.S.A.*, 105(37):13993–8.
- Martin, S., Brown, W. M., Klavans, R., and Boyack, K. W. (2011). OpenOrd: an open-source toolbox for large graph layout. In *Visualisation and Data Analysis*, volume 7868, pages 7868–06, San Francisco Airport, California, USA. SPIE.
- Masiero, M., Simões, F. C., Han, H. D., Snell, C., Peterkin, T., Bridges, E., Mangala, L. S., Wu, S. Y.-Y., Pradeep, S., Li, D., Han, C., Dalton, H., Lopez-Berestein, G., Tuynman, J. B., Mortensen, N., Li, J.-L., Patient, R., Sood, A. K., Banham, A. H., Harris, A. L., and Buffa, F. M. (2013). A Core Human Primary Tumor Angiogenesis Signature Identifies the Endothelial Orphan Receptor ELTD1 as a Key Regulator of Angiogenesis. *Cancer Cell*, 24(2):229–241.
- Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L., and Chenchik, A. (1999). Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.*, 27(6):1558–1560.
- McBride, J. L. and Ruiz, J. C. (1998). Ephrin-A1 is expressed at sites of vascular development in the mouse. *Mech. Dev.*, 77(2):201–204.

- McCarthy, D. J., Campbell, K. R., Lun, A. T. L., Wills, Q. F., and Hofacker, I. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186.
- McGrath, J., Somlo, S., Makova, S., Tian, X., and Brueckner, M. (2003). Two Populations of Node Monocilia Initiate Left-Right Asymmetry in the Mouse. *Cell*, 114(1):61–73.
- McGrath, K. E., Frame, J. M., Fegan, K. H., Bowen, J. R., Conway, S. J., Catherman, S. C., Kingsley, P. D., Koniski, A. D., and Palis, J. (2015). Distinct Sources of Hematopoietic Progenitors Emerge before HSCs and Provide Functional Blood Cells in the Mammalian Embryo. *Cell. Rep.*, 11(12):1892–1904.
- McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907.
- McKnight, K. D., Hou, J., and Hoodless, P. A. (2007). Dynamic expression of Thyrotropin-releasing hormone in the mouse definitive endoderm. *Dev. Dyn.*, 236(10):2909–2917.
- Medvinsky, A., Rybtsov, S., and Taoudi, S. (2011). Embryonic origin of the adult hematopoietic system: advances and questions. *Development*, 138:1017–31.
- Minowada, G., Jarvis, L. A., Chi, C. L., Neubuser, A., Sun, X., Hacoheh, N., Krasnow, M. A., and Martin, G. R. (1999). Vertebrate Sprouty genes are induced by FGF signaling and can cause chondrodysplasia when overexpressed. *Development*, 126(20):4465–4475.
- Mitjavila-Garcia, M. T., Cailleret, M., Godin, I., Nogueira, M. M., Cohen-Solal, K., Schiavon, V., Lecluse, Y., Pesteur, F. L., Lagrue, A. H., and Vainchenker, W. (2002). Expression of CD41 on hematopoietic progenitors derived from embryonic hematopoietic cells. *Development*, 129(8):2003–2013.
- Mitsunaga, K., Araki, K., Mizusaki, H., Morohashi, K.-I., Haruna, K., Nakagata, N., Giguère, V., Yamamura, K.-I., and Abe, K. (2004). Loss of PGC-specific expression of the orphan nuclear receptor ERR-beta results in reduction of germ cell number in mouse embryos. *Mech. Dev.*, 121(3):237–246.
- Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S. I., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J., and Göttgens, B. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, 33:269–276.
- Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development*, 99(3):371–382.
- Montero-Pau, J., Gómez, A., and Muñoz, J. (2008). Application of an inexpensive and high-throughput genomic DNA extraction method for the molecular ecology of zooplanktonic diapausing eggs. *Limnol Oceanogr Methods*, 6(6):218–22.
- Moore, C., Richens, J. L., Hough, Y., Ucanok, D., Malla, S., Sang, F., Chen, Y., Elworthy, S., Wilkinson, R. N., and Gering, M. (2018). Gfi1a and Gfi1b set the pace for primitive

References

- erythroblast differentiation from hemangioblasts in the zebrafish embryo. *Blood Adv.*, 2(20):2589–606.
- Moore, K. L., Persaud, T. V. N., and Torchia, M. G. (2015). *The Developing Human: Clinically Oriented Embryology*. Elsevier Health Sciences.
- Moreau, M. E., Garbacki, N., Molinaro, G., Brown, N. J., Marceau, F., and Adam, A. (2005). The Kallikrein-Kinin System: Current and Future Pharmacological Targets. *J. Pharmacol. Sci.*, 99(1):6–38.
- Morkel, M., Huelsken, J., Wakamiya, M., Ding, J., van de Wetering, M., Clevers, H., Taketo, M. M., Behringer, R. R., Shen, M. M., and Birchmeier, W. (2003). Beta-catenin regulates Cripto- and Wnt3-dependent gene expression programs in mouse axis and mesoderm formation. *Development*, 130(25):6283–6294.
- Murtagh, F. and Legendre, P. (2014). Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? *J. Classif.*, 31(3):274–295.
- Mylona, A., Andrieu-Soler, C., Thongjuea, S., Martella, A., Soler, E., Jorna, R., Hou, J., Kockx, C., van Ijcken, W., Lenhard, B., and Grosveld, F. (2013). Genome-wide analysis shows that Ldb1 controls essential hematopoietic genes/pathways in mouse early development and reveals novel players in hematopoiesis. *Blood*, 121(15):2902–2913.
- Müssel, C., Hopfensitz, M., and Kestler, H. A. (2010). BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*, 26(10):1378–80.
- Nadler, B., Lafon, S., Kevrekidis, I., and Coifman, R. R. (2006). Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In *Advances in neural information processing systems*, pages 955–962.
- Naiche, L. A., Arora, R., Kania, A., Lewandoski, M., and Papaioannou, V. E. (2011). Identity and fate of Tbx4-expressing cells reveal developmental cell fate decisions in the allantois, limb, and external genitalia. *Dev. Dyn.*, 240(10):2290–2300.
- Nelson, D. O., Jin, D. X., Downs, K. M., Kamp, T. J., and Lyons, G. E. (2014). Irx4 identifies a chamber-specific cell population that contributes to ventricular myocardium development. *Dev. Dyn.*, 243(3):381–392.
- Neshati, V., Mollazadeh, S., Bazzaz, B. S. F., Vries, A. A. d., Mojarrad, M., Naderi-Meshkin, H., Neshati, Z., and Kerachian, M. A. (2018). Cardiomyogenic differentiation of human adipose-derived mesenchymal stem cells transduced with Tbx20-encoding lentiviral vectors. *J. Cell. Biochem.*, 0(0).
- Nestorowa, S., Hamey, F. K., Sala, B. P., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N. K., Kent, D. G., and Göttgens, B. (2016). A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. *Blood*, pages blood–2016–05–716480.
- Nichols, J. and Smith, A. (2011). The origin and identity of embryonic stem cells. *Development*, 138(1):3–8.
- Nishikawa, S.-I., Nishikawa, S., Hirashima, M., Matsuyoshi, N., and Kodama, H. (1998). Progressive lineage analysis by cell sorting and culture identifies FLK1+ VE-cadherin+ cells

- at a diverging point of endothelial and hemopoietic lineages. *Development*, 125(9):1747–1757.
- North, T., Gu, T. L., Stacy, T., Wang, Q., Howard, L., Binder, M., Marín-Padilla, M., and Speck, N. A. (1999). *Cbfa2* is required for the formation of intra-aortic hematopoietic clusters. *Development*, 126(11):2563–2575.
- North, T. E., Goessling, W., Walkley, C. R., Lengerke, C., Kopani, K. R., Lord, A. M., Weber, G. J., Bowman, T. V., Jang, I.-H., Grosser, T., FitzGerald, G. A., Daley, G. Q., Orkin, S. H., and Zon, L. I. (2007). Prostaglandin E2 regulates vertebrate haematopoietic stem cell homeostasis. *Nature*, 447(7147):1007–1011.
- Ohtsuka, S., Nishikawa-Torikai, S., and Niwa, H. (2012). E-Cadherin Promotes Incorporation of Mouse Epiblast Stem Cells into Normal Development. *PLoS One*, 7(9):e45220.
- Org, T., Duan, D., Ferrari, R., Montel-Hagen, A., Handel, B. V., Kerényi, M. A., Sasidharan, R., Rubbi, L., Fujiwara, Y., Pellegrini, M., Orkin, S. H., Kurdistani, S. K., and Mikkola, H. K. (2015). *Scl* binds to primed enhancers in mesoderm to regulate hematopoietic and cardiac fate divergence. *EMBO J.*, 34(6):759–777.
- Padrón-Barthe, L., Temiño, S., Villa del Campo, C., Carramolino, L., Isern, J., and Torres, M. (2014). Clonal analysis identifies hemogenic endothelium as the source of the blood-endothelial common lineage in the mouse embryo. *Blood*, 124(16):2523–2532.
- Palis, J., McGrath, K. E., and Kingsley, P. D. (1995). Initiation of hematopoiesis and vasculogenesis in murine yolk sac explants. *Blood*, 86(1):156–163.
- Pearce, J. J. and Evans, M. J. (1999). *Mml*, a mouse Mix-like gene expressed in the primitive streak. *Mech. Dev.*, 87(1-2):189–192.
- Peng, G., Suo, S., Chen, J., Chen, W., Liu, C., Yu, F., Wang, R., Chen, S., Sun, N., Cui, G., Song, L., Tam, P., Han, J.-D., and Jing, N. (2016). Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev. Cell*, 36(6):681–97.
- Peng, T., Tian, Y., Boogerd, C. J., Lu, M. M., Kadzik, R. S., Stewart, K. M., Evans, S. M., and Morrisey, E. E. (2013). Coordination of heart and lung co-development by a multipotent cardiopulmonary progenitor. *Nature*, 500:589–92.
- Pennisi, D., Gardner, J., Chambers, D., Hosking, B., Peters, J., Muscat, G., Abbott, C., and Koopman, P. (2000). Mutations in *Sox18* underlie cardiovascular and hair follicle defects in ragged mice. *Nat. Genet.*, 24(4):434–437.
- Perea-Gomez, A., Meilhac, S. M., Piotrowska-Nitsche, K., Gray, D., Collignon, J., and Zernicka-Goetz, M. (2007). Regionalisation of the mouse visceral endoderm as the blastocyst transforms into the egg cylinder. *BMC Dev. Biol.*, 7:96.
- Pereira, C., Clarke, E., and Damen, J. (2007). Hematopoietic Colony-Forming Cell Assays. In *Stem Cell Assays, Methods in Molecular Biology™*, pages 177–208. Humana Press.
- Pereira, F. A., Qiu, Y., Zhou, G., Tsai, M.-J., and Tsai, S. Y. (1999). The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. *Genes Dev.*, 13(8):1037–1049.

References

- Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*, 9:171–81.
- Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F. W., and Orkin, S. H. (1996). The T Cell Leukemia Oncoprotein SCL/tal-1 Is Essential for Development of All Hematopoietic Lineages. *Cell*, 86(1):47–57.
- Power, C. and Jefferis, B. J. (2002). Fetal environment and subsequent obesity: a study of maternal smoking. *Int J Epidemiol*, 31:413–9.
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017a). Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*, 14(3):309–15.
- Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, 14:979–82.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Radford, E. J., Ito, M., Shi, H., Corish, J. A., Yamazawa, K., Isganaitis, E., Seisenberger, S., Hore, T. A., Reik, W., Erkek, S., Peters, A. H. F. M., Patti, M. E., and Ferguson-Smith, A. C. (2014). In utero effects. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. *Science*, 345:1255903.
- Ramskold, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtkova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., and Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, 30(8):777–782.
- Rankin, S. A., Thi Tran, H., Wlizla, M., Mancini, P., Shifley, E. T., Bloor, S. D., Han, L., Vleminckx, K., Wert, S. E., and Zorn, A. M. (2015). A Molecular atlas of Xenopus respiratory system development. *Dev. Dyn.*, 244:69–85.
- Rasmussen, C. E. (2002). Logistic Regression with regularization used to classify hand written digits - File Exchange - MATLAB Central.
- Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.
- Rawles, M. E. (1948). Origin of melanophores and their rôle in development of color patterns in vertebrates. *Physiol. Rev.*, 28(4):383–408.
- Reid, J. E. and Wernisch, L. (2016). Pseudotime estimation: deconfounding single cell time series. *Bioinformatics*, 32(19):2973–80.
- Reizel, Y., Itzkovitz, S., Adar, R., Elbaz, J., Jinich, A., Chapal-Ilani, N., Maruvka, Y. E., Nevo, N., Marx, Z., Horovitz, I., Wasserstrom, A., Mayo, A., Shur, I., Benayahu, D., Skorecki, K., Segal, E., Dekel, N., and Shapiro, E. (2012). Cell Lineage Analysis of the Mammalian Female Germline. *PLoS Genet.*, 8(2):e1002477.
- Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32:896–902.

- Rivera-Pérez, J. A. and Magnuson, T. (2005). Primitive streak formation in mice is preceded by localized activation of Brachyury and Wnt3. *Dev. Biol.*, 288(2):363–371.
- Robb, L., Lyons, I., Li, R., Hartley, L., Köntgen, F., Harvey, R. P., Metcalf, D., and Begley, C. G. (1995). Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proc. Natl. Acad. Sci. U.S.A.*, 92(15):7075–7079.
- Rohrman, B. A. and Mazziotti, D. A. (2005). Quantum Chemical Design of Hydroxyurea Derivatives for the Treatment of Sickle-Cell Anemia. *J. Phys. Chem. B*, 109(27):13392–96.
- Rostom, R., Svensson, V., Teichmann, S. A., and Kar, G. (2017). Computational approaches for interpreting scRNA-seq data. *FEBS Lett.*, 591(15):2213–2225.
- Ryan, K. and Chin, A. J. (2003). T-box genes and cardiac development. *Birth Defects Res. C Embryo Today*, 69(1):25–37.
- Saga, Y., Miyagawa-Tomita, S., Takagi, A., Kitajima, S., Miyazaki, J. i., and Inoue, T. (1999). MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. *Development*, 126:3437–47.
- Sakoe, H. and Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics, Budapest*, volume 3, pages 65–69, Budapest. Akadémiai Kiadó.
- Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acous.*, 26(1):43–49.
- Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., and Ueda, H. R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.*, 14(4).
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, 9(7):676–82.
- Schneider, M. D., Baker, A. H., and Riley, P. (2015). Hopx and the Cardiomyocyte Parentage. *Mol. Ther.*, 23(9):1420–22.
- Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535(7611):289–293.
- Shalaby, F., Ho, J., Stanford, W. L., Fischer, K. D., Schuh, A. C., Schwartz, L., Bernstein, A., and Rossant, J. (1997). A requirement for Flk1 in primitive and definitive hematopoiesis and vasculogenesis. *Cell*, 89:981–90.
- Shang, M.-M., Talukdar, H. A., Hofmann, J. J., Niaudet, C., Asl, H. F., Jain, R. K., Rossignoli, A., Cedergren, C., Silveira, A., Gigante, B., Leander, K., de Faire, U., Hamsten, A., Ruusalepp, A., Melander, O., Ivert, T., Michoel, T., Schadt, E. E., Betsholtz, C., Skogsberg, J., and Björkegren, J. L. M. (2014). Lim domain binding 2: a key driver of transendothelial migration of leukocytes and atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.*, 34(9):2068–2077.

References

- Sheikh, F., Lyon, R. C., and Chen, J. (2015). Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease. *Gene*, 569(1):14–20.
- Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, 111(51):E5593–601.
- Shimodaira, H., Yoshioka-Yamashita, A., Kolodner, R. D., and Wang, J. Y. J. (2003). Interaction of mismatch repair protein PMS2 and the p53-related transcription factor p73 in apoptosis response to cisplatin. *Proc. Natl. Acad. Sci. U.S.A.*, 100(5):2420–25.
- Shirley, M. D., Tang, H., Gallione, C. J., Baugher, J. D., Frelin, L. P., Cohen, B., North, P. E., Marchuk, D. A., Comi, A. M., and Pevsner, J. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. *N. Engl. J. Med.*, 368:1971–9.
- Shivdasani, R. A., Mayer, E. L., and Orkin, S. H. (1995). Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*, 373(6513):432–434.
- Siklenka, K., Erkek, S., Godmann, M., Lambrot, R., McGraw, S., Lafleur, C., Cohen, T., Xia, J., Suderman, M., Hallett, M., Trasler, J., Peters, A. H., and Kimmins, S. (2015). Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science*, 350:aab2006.
- Silver, L. and Palis, J. (1997). Initiation of murine embryonic erythropoiesis: a spatial analysis. *Blood*, 89(4):1154–1164.
- Singh, M. K., Christoffels, V. M., Dias, J. M., Trowe, M.-O., Petry, M., Schuster-Gossler, K., Bürger, A., Ericson, J., and Kispert, A. (2005). Tbx20 is essential for cardiac chamber differentiation and repression of Tbx2. *Development*, 132(12):2697–2707.
- Smith, Z. D., Shi, J., Gu, H., Donaghey, J., Clement, K., Cacchiarelli, D., Gnirke, A., Michor, F., and Meissner, A. (2017). Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer. *Nature*, 549(7673):nature23891.
- Snippert, H. J., Flier, L. G. v. d., Sato, T., Es, J. H. v., Born, M. v. d., Kroon-Veenboer, C., Barker, N., Klein, A. M., Rhee, J. v., Simons, B. D., and Clevers, H. (2010). Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. *Cell*, 143(1):134–44.
- Solari, V., Jawaid, W., and Jesudason, E. C. (2011). Enhancing safety of laparoscopic vascular control for neonatal sacrococcygeal teratoma. *J. Pediatr. Surg.*, 46:e5–7.
- Song, L., Chen, J., Peng, G., Tang, K., and Jing, N. (2016). Dynamic Heterogeneity of Brachyury in Mouse Epiblast Stem Cells Mediates Distinct Response to Extrinsic Bone Morphogenetic Protein (BMP) Signaling. *J. Biol. Chem.*, 291(29):15212–15225.
- Southwood, C. M., Downs, K. M., and Bieker, J. J. (1996). Erythroid Krüppel-like factor exhibits an early and sequentially localized pattern of expression during mammalian erythroid ontogeny. *Dev. Dyn.*, 206(3):248–259.
- Stankiewicz, M. J. and Crispino, J. D. (2009). ETS2 and ERG promote megakaryopoiesis and synergize with alterations in GATA-1 to immortalize hematopoietic progenitor cells. *Blood*, 113(14):3337–3347.

- Stevens, L. C. (1970). Experimental production of testicular teratomas in mice of strains 129, A/He, and their F1 hybrids. *J. Natl. Cancer Inst.*, 44(4):923–929.
- Strid, T., Söderström, M., and Hammarström, S. (2008). Leukotriene C4 synthase promoter driven expression of GFP reveals cell specificity. *Biochem. Biophys. Res. Commun.*, 366(1):80–5.
- Ståhlberg, A. and Bengtsson, M. (2010). Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *Methods*, 50(4):282–288.
- Sulston, J., Schierenberg, E., White, J., and Thomson, J. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.*, 100(1):64–119.
- Sumi, T., Oki, S., Kitajima, K., and Meno, C. (2013). Epiblast Ground State Is Controlled by Canonical Wnt/ β -Catenin Signaling in the Postimplantation Mouse Embryo and Epiblast Stem Cells. *PLoS One*, 8(5):e63378.
- Sun, C., Burgner, D. P., Ponsonby, A. L., Saffery, R., Huang, R. C., Vuillermin, P. J., Cheung, M., and Craig, J. M. (2013). Effects of early-life environment and epigenetics on cardiovascular disease risk in children: highlighting the role of twin studies. *Pediatr. Res.*, 73:523–30.
- Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, 14(4):381–7.
- Szabo, P. and Mann, J. R. (1994). Expression and methylation of imprinted genes during in vitro differentiation of mouse parthenogenetic and androgenetic embryonic stem cell lines. *Development*, 120(6):1651–1660.
- Szymczak, A. L., Workman, C. J., Wang, Y., Vignali, K. M., Dilioglou, S., Vanin, E. F., and Vignali, D. A. A. (2004). Correction of multi-gene deficiency *in vivo* using a single 'self-cleaving' 2a peptide-based retroviral vector. *Nat. Biotechnol.*, 22(5):589–594.
- Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126:663–76.
- Takaoka, K., Yamamoto, M., and Hamada, H. (2011). Origin and role of distal visceral endoderm, a group of cells that determines anterior-posterior polarity of the mouse embryo. *Nat. Cell Biol.*, 13:743–52.
- Takeuchi, J. K. and Bruneau, B. G. (2009). Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors. *Nature*, 459(7247):708–711.
- Takezaki, N. and Nei, M. (2009). Genomic Drift and Evolution of Microsatellite DNAs in Human Populations. *Mol. Biol. Evol.*, 26(8):1835–40.
- Tam, P. P. L. and Behringer, R. R. (1997). Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.*, 68(1):3–25.
- Tamplin, O. J., Kinzel, D., Cox, B. J., Bell, C. E., Rossant, J., and Lickert, H. (2008). Microarray analysis of *Foxa2* mutant mouse embryos reveals novel gene expression and inductive roles for the gastrula organizer and its derivatives. *BMC Genomics*, 9:511.

References

- Tanaka, Y., Hayashi, M., Kubota, Y., Nagai, H., Sheng, G., Nishikawa, S.-I., and Samokhvalov, I. M. (2012). Early ontogenic origin of the hematopoietic stem cell lineage. *Proc. Natl. Acad. Sci. U.S.A.*, 109(12):4515–4520.
- Tanaka, Y., Sanchez, V., Takata, N., Yokomizo, T., Yamanaka, Y., Kataoka, H., Hoppe, P. S., Schroeder, T., and Nishikawa, S.-I. (2014). Circulation-independent differentiation pathway from extraembryonic mesoderm toward hematopoietic stem cells via hemogenic angioblasts. *Cell Rep*, 8(1):31–39.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–U86.
- Taskesen, E. and Reinders, M. J. T. (2016). 2d Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. *PLoS One*, 11(2).
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32:381–386.
- Tshori, S., Gilon, D., Beeri, R., Nechushtan, H., Kaluzhny, D., Pikarsky, E., and Razin, E. (2006). Transcription factor MITF regulates cardiac growth and hypertrophy. *J Clin. Invest.*, 116(10):2673–81.
- Turner, D. A., Hayward, P. C., Baillie-Johnson, P., Rué, P., Broome, R., Faunes, F., and Arias, A. M. (2014). Wnt/ β -catenin and FGF signalling direct the specification and maintenance of a neuromesodermal axial progenitor in ensembles of mouse embryonic stem cells. *Development*, 141(22):4243–53.
- Utomo, A. R., Nikitin, A. Y., and Lee, W. H. (1999). Temporal, spatial, and cell type-specific control of Cre-mediated DNA recombination in transgenic mice. *Nat. Biotechnol.*, 17:1091–6.
- Valley, J. K., Swinton, P., Boscardin, W. J., Lue, T. F., Rinaudo, P. F., Wu, M. C., and Garcia, M. M. (2010). Preimplantation mouse embryo selection guided by light-induced dielectrophoresis. *PLoS One*, 5:e10160.
- Vander Plaetsen, A.-S., Deleye, L., Cornelis, S., Tilleman, L., Van Nieuwerburgh, F., and Deforce, D. (2017). STR profiling and Copy Number Variation analysis on single, preserved cells using current Whole Genome Amplification methods. *Sci. Rep.*, 7.
- van der Maaten, L. and Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Mach. Learn.*, 87(1):33–55.
- Veenendaal, M. V., Painter, R. C., de Rooij, S. R., Bossuyt, P. M., van der Post, J. A., Gluckman, P. D., Hanson, M. A., and Roseboom, T. J. (2013). Transgenerational effects of prenatal exposure to the 1944–45 Dutch famine. *BJOG*, 120:548–53.
- Vincent, S. D., Mayeuf-Louchart, A., Watanabe, Y., Brzezinski, J. A., Miyagawa-Tomita, S., Kelly, R. G., and Buckingham, M. (2014). Prdm1 functions in the mesoderm of the second heart field, where it interacts genetically with Tbx1, during outflow tract morphogenesis in the mouse embryo. *Hum. Mol. Genet.*, 23(19):5087–5101.

- Vincentz Joshua W., Barnes Ralston M., Firulli Beth A., Conway Simon J., and Firulli Anthony B. (2008). Cooperative interaction of Nkx2.5 and Mef2c transcription factors during heart development. *Dev. Dyn.*, 237(12):3809–3819.
- Vogel, G. (2015). Regenerative medicine. Report finds misconduct by surgeon. *Science*, 348:954–5.
- Vogt, W. (1929). Gestaltungsanalyse am Amphibienkeim mit Örtlicher Vitalfärbung. *W. Roux' Archiv f. Entwicklungsmechanik*, 120(1):384–706.
- Wakayama, T., Perry, A. C., Zuccotti, M., Johnson, K. R., and Yanagimachi, R. (1998). Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature*, 394:369–74.
- Walshaw, C. (2003). A multilevel algorithm for force-directed graph-drawing. *J. Graph Algorithms Appl.*, 7(3):253–285.
- Wang, J., Ma, D., Wang, P., Wu, W., Cao, L., Lu, T., Zhao, J., and Fang, Q. (2015). Alox-5 As a Potent Therapeutic Target on Overcoming TKI-Resistance in Chronic Myeloid Leukemia with T315i Mutation in Bcr-Abl. *Blood*, 126(23):4835.
- Wang, P., Chen, T., Sakurai, K., Han, B.-X., He, Z., Feng, G., and Wang, F. (2012). Intersectional Cre Driver Lines Generated Using Split-Intein Mediated Split-Cre Reconstitution. *Sci. Rep.*, 2:497.
- Weidgang, C., Russell, R., Tata, P., Kühl, S., Illing, A., Müller, M., Lin, Q., Brunner, C., Boeckers, T., Bauer, K., Kartikasari, A., Guo, Y., Radenz, M., Bernemann, C., Weiß, M., Seufferlein, T., Zenke, M., Iacovino, M., Kyba, M., Schöler, H., Kühl, M., Liebau, S., and Kleger, A. (2013). TBX3 Directs Cell-Fate Decision toward Mesendoderm. *Stem Cell Reports*, 1(3):248–265.
- Weinreb, C., Wolock, S., and Klein, A. M. (2018). SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–8.
- Whitman, C. O. (1887). A contribution to the history of the germ-layers in Clepsine. *J. Morphol.*, 1(1):105–82.
- Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15.
- Wood, H. B. and Episkopou, V. (1999). Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. *Mech. Dev.*, 86(1-2):197–201.
- Woodhouse, S., Piterman, N., Wintersteiger, C. M., Göttgens, B., and Fisher, J. (2018). SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC Syst. Biol.*, 12(1):59.
- Woodworth, M. B., Girskis, K. M., and Walsh, C. A. (2017). Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.*, 18(4):230–44.
- Wu, S. C.-Y., Meir, Y.-J. J., Coates, C. J., Handler, A. M., Pelczar, P., Moisyadi, S., and Kaminski, J. M. (2006). piggyBac is a flexible and highly active transposon as compared to Sleeping Beauty, Tol2, and Mos1 in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, 103(41):15008–13.

References

- Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881.
- Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875.
- Xian, B. and Huang, B. (2015). The immune response of stem cells in subretinal transplantation. *Stem Cell Res Ther*, 6:161.
- Xiang, Q., Ji, S.-D., Zhang, Z., Zhao, X., and Cui, Y.-M. (2016). Identification of itga2b and itgb3 single-nucleotide polymorphisms and their influences on the platelet function. *Biomed Res Int*, 2016:1–11.
- Xie, Z.-H., Huang, Y.-N., Chen, Z.-X., Riggs, A. D., Ding, J.-P., Gowher, H., Jeltsch, A., Sasaki, H., Hata, K., and Xu, G.-L. (2006). Mutations in DNA methyltransferase DNMT3b in ICF syndrome affect its regulation by DNMT3l. *Hum. Mol. Genet.*, 15(9):1375–1385.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv:1502.03044 [cs]*. arXiv: 1502.03044.
- Yabuta, Y., Kurimoto, K., Ohinata, Y., Seki, Y., and Saitou, M. (2006). Gene Expression Dynamics During Germline Specification in Mice Identified by Quantitative Single-Cell Gene Expression Profiling. *Biol. Reprod.*, 75(5):705–716.
- Yagi, T., Tokunaga, T., Furuta, Y., Nada, S., Yoshida, M., Tsukada, T., Saga, Y., Takeda, N., Ikawa, Y., and Aizawa, S. (1993). A Novel ES Cell Line, TT2, with High Germline-Differentiating Potency. *Anal. Biochem.*, 214(1):70–6.
- Yamaji, M., Jishage, M., Meyer, C., Suryawanshi, H., Der, E., Yamaji, M., Garzia, A., Morozov, P., Manickavel, S., McFarland, H. L., Roeder, R. G., Hafner, M., and Tuschl, T. (2017). DND1 maintains germline stem cells via recruitment of the CCR4–NOT complex to target mRNAs. *Nature*, 543(7646):568–572.
- Yanagisawa, K. O., Fujimoto, H., and Urushihara, H. (1981). Effects of the Brachyury (T) mutation on morphogenetic movement in the mouse embryo. *Dev. Biol.*, 87(2):242–248.
- Yang, L., Gao, X., Luo, H., Huang, Q., Su, D., Tan, X., and Lu, C. (2017). TCF21 rs12190287 Polymorphisms Are Associated with Ventricular Septal Defects in a Chinese Population. *Genet. Test Mol. Biomarkers*, 21(5):312–315.
- Yang, L., Soonpaa, M. H., Adler, E. D., Roepke, T. K., Kattman, S. J., Kennedy, M., Henckaerts, E., Bonham, K., Abbott, G. W., Linden, R. M., Field, L. J., and Keller, G. M. (2008). Human cardiovascular progenitor cells develop from a KDR+ embryonic-stem-cell-derived population. *Nature*, 453(7194):524–528.
- Yu, Y., Lei, W., Yang, J., Wei, Y.-C., Zhao, Z.-L., Zhao, Z.-A., and Hu, S. (2018). Functional mutant GATA4 identification and potential application in preimplantation diagnosis of congenital heart diseases. *Gene*, 641:349–354.
- Yura, S., Itoh, H., Sagawa, N., Yamamoto, H., Masuzaki, H., Nakao, K., Kawamura, M., Takemura, M., Kakui, K., Ogawa, Y., and Fujii, S. (2005). Role of premature leptin surge in obesity resulting from intrauterine undernutrition. *Cell Metab.*, 1:371–8.

- Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nat. Rev. Genet.*, 16(3):172–83.
- Zeng, M., Narayanan, L., Xu, X. S., Prolla, T. A., Liskay, R. M., and Glazer, P. M. (2000). Ionizing Radiation-induced Apoptosis via Separate Pms2- and p53-dependent Pathways. *Cancer Res.*, 60(17):4889–93.
- Zhang, L., Nomura-Kitabayashi, A., Sultana, N., Cai, W., Cai, X., Moon, A. M., and Cai, C.-L. (2014). Mesodermal Nkx2.5 is necessary and sufficient for early second heart field development. *Dev. Biol.*, 390(1):68–79.
- Zhao, H., Zhao, Y., Li, Z., Ouyang, Q., Sun, Y., Zhou, D., Xie, P., Zeng, S., Dong, L., Wen, H., Lu, G., Lin, G., and Hu, L. (2018). FLI1 and PKC co-activation promote highly efficient differentiation of human embryonic stem cells into endothelial-like cells. *Cell Death Dis.*, 9(2):131.
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049.
- Zhuang, S., Zhang, Q., Zhuang, T., Evans, S. M., Liang, X., and Sun, Y. (2013). Expression of Isl1 during mouse development. *Gene Expr. Patterns*, 13:407–12.
- Zohdi, V., Lim, K., Pearson, J. T., and Black, M. J. (2014). Developmental programming of cardiovascular disease following intrauterine growth restriction: findings utilising a rat model of maternal protein restriction. *Nutrients*, 7:119–52.
- Zou, D., Silvius, D., Davenport, J., Grifone, R., Maire, P., and Xu, P.-X. (2006). Patterning of the third pharyngeal pouch into thymus/parathyroid by Six and Eya1. *Dev. Biol.*, 293(2):499–512.

Index

- V(D)J* recombination, 3
10x Genomics®, ix, 23, 25, 27, 33–43, 160–
162, 164, 167, 195–197, 201, 202,
216, 252
2A peptide, 133
acidosis, 2
acrosomal activation, 7
age-related macular degeneration, 6
agglomerative clustering, 67
AGM, 49
allantois, 10, 11, 77, 78, 108, 132, 165
amnion, 165, 176
amniotic cavity, 10, 11
amniotic ectoderm, 10, 11
amniotic endoderm, 10
amphibians, 16
angiogenesis, 80
anorectal malformations, 4
antenatal sonography, 6
antibody, 27, 29
antibody hybridisation, 27
antisense, 80
aorta-gonadal mesonephros, 12
apoptosis, 182
arachidonic acid, 115, 119–121, 154, 155,
257
asthma, 155
atria, 12, 187, 189
atrium, 189, 201
attentional interfaces, 258
autoimmunity, 3
autologous cells, 6
B-cell, 3
bcl files, 58
bglab, 42, 43, 59
bifurcating process, 103
bilaminar disc, 159
bladder, 11
blastocoele, 8
blastocyst, 8, 74
blastomeres, 7
blastula, 16
blood, 176
blood islands, 10, 12, 49
blood progenitor, 90
bone marrow, 13
cadaveric organs, 6
Caenorhabditis elegans, 5
Caffe, 258
cancer, 3
capacitation, 7
carbocyanine dye, 16
cardiac, 13, 14, 176, 186, 187, 189, 191, 201
cardiac crescent, 165
cardiac defects, 4
cardiomyocyte, 13

Index

- cardiomyogenesis, 192
- caudal neural pore, 122
- cavitation, 8
- CEL-seq, 159
- cell cycle, 43, 70, 147, 251
- cell to cell signalling, 5
- CHARGE, 4
- Charles Otis Whitman, 15
- chick, 15, 16, 47
- chimera, 47, 49, 122, 132
- chondrocytes, 6
- chorioallantoic fusion, 122, 132
- chronic myeloid leukaemia, 155
- cleavage, 7
- CLOVES sequence, 4
- cluster assignment, 68
- Coloboma, 4
- compaction, 7
- conditional genetic recombination, 17
- confetti mouse, 18
- confocal microscopy, 26
- congenital anomalies, 1
- congenital diaphragmatic hernia, 1
- conotruncal cardiac malformations, 111
- copy number variant, 18, 19
- corona radiata, 7
- cosine distance, 64
- covariance function, 100
- CpG island, 74
- Cre, 18
- Cre-*loxP*, 17, 18
- CRISPR/Cas9, 19
- curse of dimensionality, 64
- CytoTOF, 254
- Danio rerio, 5
- DAPI, 27, 28, 50, 140
- definitive endoderm, 10, 176, 177, 182
- definitive yolk sac, 12
- DeLorean, 255
- Destiny, 255
- diabetes, 3
- dialysis, 2
- diffusion, 11
- diffusion maps, 90–93, 95, 97, 98, 100
- diffusion pseudotime, 94, 96, 98, 100
- DiGeorge syndrome, 132
- digital gene expression, 160
- divisive clustering, 67
- dorsal aorta, 12
- Down's syndrome, 4
- DrL, 177, 178, 183, 184, 187, 189, 192, 194, 201
- dropout, 63, 98, 157, 169, 204, 252
- Drosophila melanogaster, 5
- duplicate reads, 124
- duplicates, 152, 160
- dynamic time warping, 104, 105
- ear deformities, 4
- ECMO, 2
- ectoderm, 10, 11, 48, 72, 80, 132, 147
- ectoplacental cone, 9
- EGFP, 140
- egg cylinder, 9, 10
- eicosanoids, 119
- eigenvalues, 71, 90–92, 95, 96
- eigenvectors, 71, 72, 90–92, 95, 96
- electroporation, 17
- embryoid body, 49, 155
- EMP, 111, 114, 119
- EMT, 10, 12, 49, 86, 145, 149, 184
- endoderm, 11, 48, 78, 147, 165, 178–182
- endometrium, 8

- endothelium, 12–14, 49, 72, 107, 118, 122, 130, 147, 154, 158, 176
- epiblast, 8–12, 48, 72–74, 80, 107, 132, 144, 145, 147, 154, 157
- epigenetic, 3, 159
- EpiSC, 6
- epithelial to mesenchymal transition, 9
- epoxyeicosatrienoic acids, 115, 155, 257
- ERCC, 28, 51, 59–61, 64, 65, 72, 123–125, 138, 252
- erythroid, 90
- erythroid myeloid precursors, 111
- erythropoiesis, 122
- ESC, 5–7, 44, 45, 48, 49, 115, 118–120, 122, 159
- extraembryonic, 9
- extraembryonic ectoderm, 9, 10
- extraembryonic mesoderm, 10, 11, 176
- FACS, 27, 28, 45, 49–51, 53, 60, 123, 132, 136
- famine, 3
- fastq, 58, 59, 142, 152
- fate map, 15
- fate mapping, 47
- feature selection, 64
- fertilisation, 7
- feulgen, 15, 47
- filter, 29
- first heart field, 12, 13, 186, 187, 191, 192
- Flippase, 18
- flow cytometry, 13
- FLP-*FRT*, 17
- Fluidigm, 50
- Fluorophore, 29
- foetal liver, 13
- folic acid, 4
- Fontan, 1
- forebrain, 176
- foregut, 177
- foregut endoderm, 177
- FRT*, 18
- fuzzy clustering, 104, 105
- ganglion cell, 15
- gap junctions, 7
- gastrointestinal tract, 177
- gastroschisis, 1
- gastrulation, 6, 9, 10, 12, 15, 25, 27, 47, 48, 50, 80, 81, 122, 132, 145, 151, 157, 159, 256
- gatepoints, 43
- Gaussian kernel, 90, 169
- Gaussian process regression, 98, 102, 103
- gene dynamics, 98, 102
- genes
- 1700019D03Rik*, 175
 - 5730457N03Rik*, 80
 - 5730521E12Rik*, 177
 - AI467606*, 106
 - Abcg1*, 106
 - Acot7*, 185
 - Acta2*, 172
 - Actc1*, 172
 - Actr5*, 106
 - Adam23*, 105
 - Afap111*, 106
 - Afp*, 78, 153, 173, 177
 - Agfg2*, 105
 - Aif1*, 213
 - Ajap1*, 106
 - Akap7*, 105
 - Alas2*, 105
 - Aldh1a2*, 171

Index

- Alox5ap*, 115, 119, 154–156, 256
Alox5, 115, 119, 120, 154–156, 256
Alpl, 175
Alx1, 190
Amn, 78, 173
Amot, 105
Angptl4, 105, 106
Ank2, 105
Ankrd1, 172
Ano2, 106
Ap1s2, 105
Apela, 173
Apoa1, 173
Apoa4, 173
Apod, 171
Apom, 173
Appl2, 105
Arfgef1, 105
Arhgef28, 171
Arhgef2, 105
Arid3b, 195
Asb2, 172
Ascl2, 106
Atad2b, 106
Atmin, 106
Bach2, 190
Baf60c, 187
Barx1, 191, 194
Baspl, 105
Batf, 172
Bcl11b, 106
Bcl2l1, 105
Bcl9l, 106
Bcor1l, 105
Bcr-Abl-T315I, 155
Bcr-Abl, 155
Bcr, 106
Bend4, 105
Bhlhe40, 195
Bid, 106
Blvrb, 105, 106, 172
Bmp4, 9, 83
Brachyury, viii, 9, 26, 27, 30, 41, 80, 82, 132, 133, 138, 142, 154, 157, 158, 182, 183, 220
Brd1, 105
Bsn, 106
Bst2, 175
COUP-TFII, 189
COUP-TFI, 189
Cacna1c, 189
Cacna2d2, 106
Capn15, 191
Capn2, 106
Car2, 105, 172, 186
Casz1, 192, 194
Ccnblip1, 106
Ccnjl, 186
Cd36, 171
Cd59a, 76
Cdh1, 72, 75
Cdh5, 153, 228, 245, 246
Cdip1, 195
Cdx1, 183, 185
Cdx2, 78, 183, 185
Cdx4, 183, 185, 198
Cenpf, 105
Cep55, 105
Cep70, 106
Cer1, 9, 171
Cercam, 105, 106
Cfc1, 111

- Cfl2*, 105
Chchd10, 105
Cited4, 106
Ckap4, 105
Cldn6, 75, 184
Cldn7, 184
Cldn9, 184
Clic6, 105
Cmtm3, 105
Cmtm4, 106
Cmtm5, 171
Cmtm7, 105
Cnn3, 105
Col18a1, 106
Cpeb4, 105
Cped1, 172
Cpm, 173
Cpne7, 105
Creb3l1, 191
Creb3l2, 194
Creb3, 195
Cre, 133
Cripto, 9, 72, 73, 76
Csrp2, 105
Csrp3, 172
Ctla2a, 106
Cx3cl1, 105
Cxx1c, 80
Cyp26a1, 183, 184, 186
Cystm1, 185
D17H6S53E, 105
Daam1, 105
Dag1, 105
Dcbld2, 105
Ddx26b, 106
Defa30, 172
Depdc1b, 105
Dhrs11, 105
Dll1, 86, 171
Dll3, 86, 171
Dlx5, 171
Dnd1, 173, 175
Dnmt3b, 72, 74
Dock11, 105
Dock1, 105
Dppa3, 172–175
Dpy19l1, 106
Dpysl3, 106
Dscr3, 105
Dst, 105
Dusp16, 105
Dusp2, 106
Dyrk2, 105
E-cad, 10, 12, 72
E2f8, 106
Ebf1, 172
Ebf2, 172
Efcab14, 105
Efna1, 83
Egfl7, 228
Egr1, 106
Ehd2, 105
Ehd3, 106
Elavl2, 105
Elf5, 78, 153
Ell2, 105
Elmo2, 105
Eltl1, 156
Emilin1, 105
En1, 171, 184
Eng, 172
Epb4.2, 105

Index

- Epha1*, 186
Epha5, 185
Epor, 106
Esrp1, 175
Esrra, 195
Esrrg, 194
Essrb, 78
Ets1, 190, 228
Ets2, 130
Etv2, 172
Etv4, 185, 190
Etv5, 190
Evx1os, 185
Evx1, 80
Ezr, 186
Fads3, 105
Fam110c, 106
Fam160b2, 106
Fam193b, 106
Fam21, 105
Fam222a, 175
Fam58b, 106
Fam84b, 105
Fbxo18, 105
Fezf1, 171
Fezf2, 171
Fgf15, 171
Fgf17, 171, 185
Fgf5, 72, 74
Fgf8, 80, 171, 185
Fgfbp3, 171, 186
Fgfr1, 10
Fgf, 10
Fig, 105
Fkbp7, 105
Fli1, 76, 105, 106, 117, 130, 172
Flk1, 49, 144, 157
Fn1, 105
Fnbp11, 105
Fosb, 194
Fosl2, 191
Fos, 194
Foxa1, 173
Foxb1, 186
Foxc1, 86
Foxc2, 86
Foxd3, 171
Foxf1, 172, 189
Foxg1, 171
Foxi2, 171
Frmd6, 105
Frzb, 80, 86
Furin, 9
Fzd7, 86, 189
Gabra4, 105
Gadd45g, 172
Galnt11, 105
Gamt, 106
Gap43, 185
Gata1, 106, 172
Gata2, 130, 172
Gata4, 111, 187, 194
Gata5, 111, 192, 194
Gata6, 9, 194
Gdf3, 9
Gfi1b, 103, 106, 111, 114, 228, 232, 245,
246
Gfi1, 117
Gja6, 106
Gjb3, 171, 175
Gli1, 191
Gli3, 190

- Gm29650*, 174
Gm6665, 105
Gna12, 106
Gnail, 105
Gne, 106
Gng12, 105
Gng2, 105
Golim4, 105
Gpc3, 105
Gpr108, 105
Gpr97, 106
Gpx2, 173
Greb1, 185
Grhl1, 195
Grina, 105
Grrp1, 105
Grsf1, 185
Gsta4, 106
Gstm1, 105
Gucyl1a3, 106
Gypa, 172, 173
H2afy2, 105
Hand1, 195
Hand2, 111, 172
Hapln1, 105
Hba-a1, 173
Hba-x, 105, 106, 172
Hbb-bh0, 172
Hbb-bh1, 102, 103, 105, 106, 153, 172, 173
HbbbH1, 228, 232
Hebp1, 105
Hes3, 171, 185
Hes6, 106, 194
Hes7, 171, 186, 191
Hesx1, 171
Hey2, 190, 194
Hhex, 105, 106, 172
Hif1a, 195
Hist3h2ba, 106
Hivep1, 106
Hkl, 106
Homer2, 106
Hopx, 189, 192, 194
HoxA, 86
HoxB8, 26
HoxB, 86
Hoxa10, 172
Hoxa11, 172
Hoxa1, 190
Hoxa5, 191
Hoxa7, 183, 185
Hoxa9, 183, 185
Hoxaas3, 183, 185
Hoxb1, 171, 189, 190
Hoxb2, 189
Hoxb4, 190
Hoxb5os, 183, 185
Hoxb5, 130, 131, 190
Hoxb8, 183, 185
Hoxb9, 183, 185
Hoxc10, 172
Hoxc4, 183, 186
Hoxc6, 183, 185
Hoxc8, 183, 185, 196
Hoxc9, 183, 186
Hoxd1, 172, 190
Hoxd4, 171
Hoxd9, 172
Hox, 86, 87
Hpdl, 105
Hspbl, 172

Index

Hspb7, 172
Hspg2, 105
Id2, 194
Ier5, 106
Ifitm3, 173, 175
Igf2, 105, 106
Igsf3, 106
Igtb3, 114
Ikzf2, 106
Inf2, 106
Irx1, 189
Irx2, 189, 190, 195
Irx3, 178
Irx4, 187–190, 192, 194
Irx5, 178
Isl1, 111
Itga2b, 49, 106, 111, 114, 130
Itga6, 106
Itgb3, 49, 114
Itpkb, 105
Jag2, 156
Jund, 195
KDR, 49
Kcnn4, 105
Kdr, 102, 103, 105, 106, 172, 173
Kif13a, 105
Kif1b, 105
Kif2c, 105
Kifc1, 105
Kit, 106, 172, 173
Klf1, 172
Klf2, 194
Klf4, 5, 195
Klf5, 105, 106, 175
Klf6, 195
Klhl2, 105
Klk1, 189
Krt8, 105
L3mbil3, 190
Ldb1, 156
Ldb2, 156
Lef1, 105
Lefty1, 9, 86
Lefty2, 86
Lgals2, 173
Lgals9, 106
Lhpp, 185
Lhx1, 172
Lhx2, 171
Lhx5, 171
Limd2, 105
Lix1, 185
Lmo2, 76, 117, 172
Lrp11, 106
Lrrfip1, 194
Lrrn4, 111
Ltc4s, 115, 117, 119, 154, 156, 256
Ltc4s, 155
Lum, 172
Lyll1, 130
Lyve1, 12
Mafb, 171
Mapk, 195
Map1b, 105
Map3k1, 105
Map4k4, 105
Mapk3, 105
Mapk4, 86
Matn1, 106
Mbnl1, 106
Mef2c, 111, 187, 192, 194
Mef2d, 194

- Meox1*, 171
Meox2, 171
Mesp1, 12, 13, 86, 172
Mesp2, 86, 171
Mfsd2b, 106
Mgat4b, 106
Mgst1, 185
Micall2, 106
Mitf, 189, 190
Mixl1, 80, 149, 184
Mpnd, 105
Mpp2, 106
Mt1, 105
Mtg1, 105
Mxil, 106
Myc, 191
Myh6, 172
Myl2, 189
Myl4, 172
Myl7, 172
Myo5a, 105
Nanog, 153, 172, 173, 175
Ncoal, 191
Nde1, 105
Ndfip1, 105
Ndufaf7, 105
Nek6, 105
Nepn, 173, 177
Nfatc1, 76
Nfe2l2, 195
Nfe2, 106
Nfxl1, 105
Nfya, 190
Nid2, 105
Nkx-2, 198
Nkx1-2, 171, 183
Nkx1, 185
Nkx2-5, 111, 112, 187, 192, 194
Nkx2.5⁺, 113
Nkx2.5⁻, 113
Nkx2.5, 111
Nmrk1, 105
Nodal, 9
Nog, 172
Notch, 156
Noto, 172
Nr1d1, 191
Nr2f1, 189, 190, 194
Nr2f2, 189
Nr4a1, 105, 195
Nrros, 106
Nxf3, 172
Oaf, 106
Oct3/4, 5
Oct4, 72, 73, 76, 153
Ogfod1, 106
Osr1, 172
Otx2, 72, 73, 171, 184
Pank1, 106
Pax1, 178
Pax5, 171
Pax6, 171
Pax9, 173, 178
Pbx3, 195
Pced1b, 105
Pcgf5, 105
Pcsk6, 9
Pcsk7, 105
Pde4a, 105
Pdgfa, 185
Pear1, 106
Pecam1, 172, 173

Index

Peg3, 194
Pex6, 105
Pfkm, 106
Pgam2, 172
Phf6, 105
Phlda2, 105
Pigl, 106
Pitx1, 172
Pla2g16, 175
Plac1, 172
Plac8, 172
Plagl1, 194
Plcl2, 105
Plec, 106
Plekha1, 105
Plk2, 105
Plxnc1, 106
Pmaip1, 185
Pms2, 178, 182
Podxl, 105
Postn, 172
Pou5f1, 72, 86, 153, 172, 173, 175
Pou6f2, 191, 195
Ppargc1b, 106
Ppox, 105
Pradc1, 105
Prdm14, 173
Prdm6, 194
Prickle1, 105
Prkca, 106
Prox1, 190, 194
Prrx2, 194
Psmb10, 105
Psme1, 175
Ptpn13, 105
Ptpn9, 105
Ptprk, 105
Ptrf, 106
Pvrl2, 105
Pyy, 173, 177, 182, 201
Rab31, 105
Rab3il1, 105
Rarb, 191
Rasal2, 105
Rasgrp2, 186
Rasl11b, 105
Rax, 171
Rbm4b, 105
Rbms1, 105
Rbp1, 105
Rbpms, 105
Rcor2, 106
Rgs10, 105
Rgs19, 105
Rhbdd2, 105
Ripply2, 171
Ripply3, 171
Rit1, 195
Rnf103, 105
Rnf125, 106
Rora, 190
Rosa26, 17
Rpp25, 105
Rps2, 106
Rras2, 105
Rspo3, 171
Runx1, 117
Sall2, 105
Samd3, 172
Saysd1, 105
Scg5, 186
Scl, 117, 121, 122

- Sdc3*, 106
Sema3a, 171
Sepn1, 105
Sept4, 105
Serpinh6a, 105
Serpinh1, 105
Sertad1, 105
Sfrp5, 189
Sh2b3, 105
Sh3rf1, 105
Shc1, 105
Shh, 172
Sipa1, 106
Six1, 178
Six2, 172
Six3, 171
Skap1, 76
Sla2, 106
Slc14a1, 105
Slc16a10, 105
Slc18a2, 106
Slc25a12, 106
Slc25a37, 105
Slc2a3, 185
Slc30a4, 105
Slc38a5, 105
Slc9a3r2, 106
Smad4, 105
Smad6, 194
Smarcd3, 172, 187, 194
Smim1, 105
Smim3, 186
Smox, 105
Smpd2, 105
Smtn, 105
Smurf2, 105
Snai1, 10, 86, 149, 184, 191
Snai2, 190
Sord, 105
Sox10, 171
Sox11, 105, 106
Sox13, 191
Sox15, 173
Sox17, 172, 173
Sox2, 5, 171, 182–184, 186
Sox6, 191
Sox7, 130
Sp8, 186
Spefl, 105
Spi1, 117
Spint2, 186
Spns2, 106
Spp1, 213
Spry4, 86
Sptbn1, 105
Sptlc2, 105
Src, 106
St3gal5, 106
Stat3, 190, 194
Stat6, 191
Stella, 173
Stmn2, 185
Stx6, 105
Syk, 106
Syng1, 105
Tal1, viii, 27, 30, 41, 76, 103, 117, 121–131, 144, 145, 153, 154, 158, 172, 245, 246
Tal, 127
Taok3, 106
Tapbp, 175
Tbx18, 190

Index

- Tbx1*, 172
Tbx20, 187, 192, 194
Tbx2, 83, 194
Tbx3, 83, 105, 106
Tbx4, 172
Tbx5, 187, 192, 194
Tbx6, 86, 171
Tcea3, 186
Tcf21, 111, 190
Tdgl1, 72
Tdo2, 172
Tead1, 194
Tek, 153, 172
Tfap2a, 171
Tfap2b, 171
Tfap2c, 175
Tfap4, 190
Tfeb, 191, 195
Tgif1, 190
Tgm2, 106
Thap3, 105
Tie1, 106, 153
Timp3, 106
Tinf2, 105
Tmbim1, 105
Tmem185b, 105
Tmem98, 105
Tmod1, 105
Tmod2, 105
Tnfaip8, 106
Tnnc1, 172
Tnni1, 172
Tnnt2, 172
Tox3, 191
Tpd52, 186
Trem12, 106
Trh, 72, 73, 76, 173
Trim10, 105
Trim44, 105
Tsc22d1, 194
Tshz1, 191
Tsku, 105
Tspan32, 106
Ttc28, 175
Ttr, 78, 153, 173, 177, 178
Tuba4a, 105
Twist1, 86, 190
Twist2, 190
T, ix, 9, 81, 82, 132, 135, 136, 140, 142, 144, 145, 147–149, 171, 172, 182–184, 186, 248
Ube2l6, 105
Unc45a, 106
Uncx, 171
Uros, 105
VE-Cadherin, 173, 245
VE-Cad, 153
VEGFR-2, 49
Vcl, 105
Vgll3, 171
Vgll4, 105
Was, 106
Wisp1, 172
Wnt1, 171
Wnt3a, 171
Wnt3, 9, 80, 86
Wnt5a, 86
Wnt5b, 183, 186
Wnt6, 171
Wnt, 189
Wtip, 105
Wwc2, 105

- Wwtr1*, 105
Xist, 173–175
Xk, 105
Ypel3, 175
Zbtb20, 194
Zbtb26, 191
Zbtb2, 191
Zbtb37, 191
Zbtb44, 190
Zbtb46, 191
Zfp105, 191
Zfp260, 105
Zfp3611, 105
Zfp521, 105
Zfp608, 190
Zfp618, 191
Zfp652, 105
Zfp654, 191
Zfp959, 191
Zfp9, 190
Zfpm1, 195
Zim1, 195
Zscan21, 105
Zswim6, 106
Zyx, 105
c-Myc, 5
tal1, 130
genito-urinary anomalies, 4
genotyping, 26
germ cell tumours, 173
germ cells, 3
GESTALT, 19
GFP, 17, 140
giant multinucleated cells, 3
gpr, 44, 102
GPU, 258
GTF file, 142, 145, 198
gut, 15
haemangioblast, 14, 115, 117–119
haematopoietic, 13, 14
haemogenic endothelium, 12, 14, 115
haploid, 3
Hbb-bh1, 228, 232
head fold, 25
heart tube, 12
hepatic cells, 3
heterotaxia, 111
hidden mortality, 1
hierarchical clustering, 67
high molecular weight dextrans, 16
hindbrain, 11, 176, 184
hindgut, 177, 178
Hirschsprung disease, 1, 15
histiocytes, 3
Holt-Oram syndrome, 132
HSC, 12
hydroxyurea, 155, 156, 256
hypoblast, 8, 10
hypothalamus, 74
Imatinib, 155, 156
immunofluorescence, 77
immunosuppression, 2
in-silico dissection, 177
inner cell mass, 7, 8
intein, 18
intellectual disability, 4
intermediate mesoderm, 11, 184
interphase, 47
intestinal atresia, 2
intraembryonic coelom, 11
intraembryonic mesoderm, 11

Index

- iPS, 6, 7, 48
- ischaemic heart disease, 3
- isoform, 195
- isomerism, 111
- ISPCR, 54

- laser, 29
- lateral plate mesoderm, 11, 12, 184, 186, 187
- leech, 15, 16
- left ventricle, 12
- leukaemia, 155
- leukaemic stem cells, 155
- leukotriene, 115, 119–121, 154, 155, 157, 257
- limb deformities, 4
- lineage tree, 19
- lipofection, 17
- lipoxin, 115, 154
- liver, 177
- long interspersed element 1, 18
- long short-term memory, 258
- long terminal repeats, 18
- Louvain, 43
- Louvain modularity, 169, 170, 177–181, 183, 184, 189
- LTA₄, 119
- LTB₄, 155
- LTC₄, 119–121
- LTC₄, 44, 45, 119, 120, 155, 256
- LTD₄, 119
- LTD₄, 155
- lung, 177
- lysis buffer, 28

- macrophage, 118, 119
- malrotation, 2
- Markov chain, 91

- MARS-seq, 159
- maxillary arch, 11
- median umbilical ligament, 11
- meiosis, 3
- meosderm, 183
- mesenchymal, 73
- mesenchymal stem cells, 6
- mesoderm, 10–13, 48–50, 52, 72, 80, 83, 86, 108, 118, 119, 122, 132, 176, 182–184, 187
- mesodermal core, 12
- microarray, 20, 253
- microfluidics, 167
- microglia, 6
- microsatellite, 18, 19
- midbrain, 176, 184
- midgut, 178
- midgut volvulus, 2
- minimum spanning tree, 254
- mis-match repair, 19
- model organisms, 15
- Monocle, 253, 254
- morphogen gradient, 5
- morula, 7, 8
- mosaicism, 3, 4
- mRNA, 20, 151
- multivariate normal distribution, 101
- mycoplasma, 44
- myeloid, 12
- myocytes, 3

- NAMP, 183, 184
- nascent mesoderm, 12, 14, 83–86, 107, 108, 111, 144, 145, 147, 149, 154, 157
- Neubauer chamber haemocytometer, 165
- neural crest, 11, 15, 16, 176
- neural network, 258, 259

- neural plate, 25
 neural tube, 11, 15, 132, 176, 182, 183
 neural Turing machines, 258
 Nile Blue, 16
 NMP, 182–185
 NOD/SCID mice, 155
 Nomarski microscope, 16
 non-communicable disease, 3
 normalisation, 63, 166
 notochord, 11, 132, 176

 obesity, 3
 oesophageal atresia, 1, 4
 oocyte, 7
 organogenesis, 25
 osteoclasts, 3
 outflow tract, 12, 187, 189

 pancreas, 177
 paracrine, 5
 paraxial mesoderm, 11, 12, 184
 parenteral nutrition, 2
 parietal endoderm, 9, 10
 patent urachus, 11
 PCA, 71–73, 75, 76, 127, 129, 142, 144, 145, 158
 PCR, 152
 Pearson correlation, 252
 pericardial sac, 121
 PGC, 73, 173–176
 pharyngeal mesoderm, 187
 pharyngeal arch, 11, 12
 pharyngeal endoderm, 177, 178
 pharyngeal mesoderm, 12, 78, 107, 111, 157, 176, 187, 191, 192
 phospholipase, 119
 phospholipid bilayer, 119
 phospholipids, 119
 Phred quality score, 58, 59
 Pierre-Robin sequence, 4
 pituitary, 74
 placenta, 9, 121
 placental eutherians, 5
 placodes, 176
 plasmid, 17
 platelet, 49
 pluripotency, 73
 pluripotent, 73
 polar body, 7
 polymerase slippage, 19
 polymorphism, 19
 polyploidy, 4
 primer dimers, 54
 primitive endoderm, 8–10
 primitive erythrocytes, 12
 primitive mesoderm, 145
 primitive streak, 6, 9, 10, 12, 25, 48, 49, 80, 145, 147
 primitive endoderm, 9
 primitive streak, 9
 principal curve, 94, 98, 100
 princurve, 44
 proamniotic cavity, 9
 promoter, 18
 pronucleus, 7
 prospective lineage tracing, 16
 prostaglandin, 115, 119, 155, 257
 proteins
 ALOX5AP, 119
 ALOX5, 119, 155
 BRACHYURY, 29, 132, 138, 144
 CD16/32, 213, 214
 CD16, 27

Index

- CD31*, 29, 136, 138
CD326, 29, 138
CD32, 27
CD34, 213, 214
CD41, 27, 49–53, 72, 86, 107, 108, 111, 123, 130, 153
CD45, 155
CD61, 114
CD68, 155
CD41, 123
CD68, 155
CRYPTIC, 111
DNMT3B, 76
E-CAD, 28
E-CAD, 28, 76, 136, 138
ECDH, 29
EPCAM, 29, 136, 138
FLAP, 256
FLK1, vii, viii, 27
FGF5, 76
FK1, 132
FLII, 115
FLK1, viii, 13, 29, 49–53, 72, 86, 102, 107, 108, 111, 122, 123, 125, 126, 132, 136, 138, 140, 151, 154, 156, 227
FLK2, 213, 214
GGT1, 119
GGT5, 119
GFIIB, 115
GFII, 115
LMO2, 115
LTC4S, 119
OTX2, 76
PDGFR α , 29, 138
PDGFR, 136, 138, 140
PU.1, 115
PECAMI, 138
RUNX1⁺, 13
RUNX1, 13, 227
SCAI, 213, 214
SCL, 115
TALI, 115
TIE2, 29, 136, 138
C-KIT, 213, 214
CKIT, 29, 136, 138
proteome, 159
Proteus syndrome, 4
pseudospace, 83, 156
pseudotime, 44, 90, 101–103, 106, 107, 156, 157
qPCR, 20, 151
quail, 15, 16, 47
Quartz-seq, 159
recombinant DNA, 16
recombination, 19
recurrent network, 258
regenerative medicine, 12
Reichert's membrane, 10
rejection, 6
renal anomalies, 4
reporter gene, 18
retinal pigment epithelium, 6
retinoic acid, 189
retrotransposons, 18
retroviral libraries, 16
retrovirus, 18
reverse graph embedding, 254
right ventricle, 12
roots, 43, 44, 92, 96, 169, 201, 204, 212, 213, 215, 216, 247, 248, 255, 259

- Rtsne, 43
- sacrococcygeal teratomas, 6
- SAM files, 142
- scran, 42, 43
- screening, 4
- scRNA-seq, 159
- second heart field, 12, 13, 111, 186, 187, 189, 191, 192
- self-antigen, 3
- sequence, 4
- short bowel syndrome, 1
- short tandem repeats, 18
- sickle-cell anaemia, 155
- single nucleotide variant, 18, 19
- Smart-seq2, 23, 27, 30, 31, 41–43, 51, 54, 55, 63, 138, 151, 159, 160, 166, 195, 256
- smooth muscle, 14
- somatic mesoderm, 11
- somatic mutations, 18
- somatic variant, 19
- somatopleure, 184
- somite, 11, 15, 122, 132, 177
- somitic mesoderm, 176, 182, 183
- Spearman's rank correlation, 72, 114, 169, 170, 194, 252
- spina bifida, 4
- spinal cord, 184
- splanchnic, 11
- splanchnic mesoderm, 12
- splanchnopleure, 184
- spleen, 13
- splice variant, 196
- SPRI, 54
- stomodeum, 11
- stop codon, 17
- Streptavidin, 29
- STRT-seq, 159
- structural variant, 19
- Sturge-Weber syndrome, 4
- Superscript™ II, 52
- tagmentation, 152
- Tamoxifen, 17, 133
- tandem repeat, 19
- Tensorflow, 258
- terminal repeats, 17
- Theano, 258
- thymus, 177
- thyroid, 74, 177
- tight junctions, 7
- time-series, 104
- Tn5 transposase, 152
- Tn5 transposome, 54, 63
- Total intestinal aganglionosis, 2
- tracheo-oesophageal fistula, 4
- transcription termination variants, 195, 196, 198–200, 202, 252, 259
- transfection, 17
- transgene, 17
- transition matrix, 91
- transplant, 15, 16, 147
- transplantat, 2
- transposase, 17, 152
- transposome, 152
- transposon, 17
- trophectoderm, 9, 11
- trophoblast, 7
- trunk mesoderm, 12
- tSNE, 76–78, 80, 81, 83, 84, 86, 88, 107, 108, 111, 112, 129–131, 145, 147, 148, 154, 169, 170, 177, 178, 182, 183, 186, 253, 254, 259

Index

TSO, 54, 60, 72, 138, 151, 152, 166

tyrosine kinase inhibitor, 155

umbilical vessels, 11

UMI, 160, 166, 167, 196

urachal cysts, 11

VACTERL, 4

velocardial facial syndrome, 4

ventricle, 187, 189, 201

ventricular myocardium, 187

vertebral anomalies, 4

viral, 3

visceral endoderm, 9, 10, 174, 176–178, 182

vital dye, 16

Waddington's landscape, 4

Wanderlust, 254

WHO, 1

xenograft, 155

Xenopus, 5

xenotransplant, 17

yolk sac, 8, 10, 12, 49, 77, 115, 121, 131, 157

zero-inflation, 204

Zileuton, 45, 119–121, 155, 156, 256

zona pellucida, 7, 8