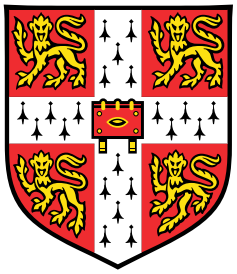
Quantifying expression variability in single-cell RNA sequencing data



Nils Eling

European Molecular Biology Laboratory,
European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Pembroke College August 2018

Ich widme diese Arbeit meiner Familie - Hildegard, Jörg und Laura - und bedanke mich für eure Unterstützung und Geduld während meiner Promotion.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the individual declarations at the beginning of each chapter. To further indicate the parts of the thesis for which I used data of others or for which others were involved in interpreting the results, I use the pronoun "we". For the parts of my thesis that are purely my own work, I use the pronoun "I".

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

Nils Eling
August 2018

Acknowledgements

First and foremost, I want to thank my supervisor John Marioni for offering me to work in his research group. John took the risk in supervising me as a PhD student from a wet-lab background and introduced me to statistical analyses. I'm thankful for all his enthusiasm and support which guided me to successfully finishing my PhD in the most enjoyable way. I also want to thank Catalina Vallejos for supervising me and turning me into a "Bayesian statistician". Not only did she support my work but also became a close friend.

This work builds on excellent collaborations. I want to thank Duncan Odom not only for allowing me to collaborate with his lab but also for guiding me through important career-related processes (e.g. paper writing). Christina Ernst supported large parts of this work with experiments, data interpretation and late nights of editing. I also want to thank Celia Martinez-Jimenez for large-scale experimental support and data interpretation which led to a successful publication. Maike de la Roche and Hung-Chang Chen also played a crucial role for publishing the final version of this paper. Furthermore, I thank Detlev Arendt and Kaia Achim for collaborating on an early project during my PhD. My thesis advisory committee members, Sarah Teichmann, Wolfgang Huber and Duncan Odom provided fruitful discussions that advanced my research over the last four years. I also thank the EMBL PhD programme for funding my research and the EMBL-European Bioinformatics Institute as well as the CRUK-Cambridge Institute for offering such a friendly work place.

All current and former lab members of the Marioni lab deserve a huge thanks for providing such a pleasant and stimulating work environment. I specifically thank: Aaron Lun for his statistical support, entertainment and being a BK lunch companion; Michael Morgan for discussing what "noise" really means; Arianne Richard for biological insights into the confusing world of T cells. I also thank Anna, Chris, Luis, Antonio, Konrad, Catalina, Liora, Jonny, Tom, Ximena, Christina, and Rebecca for all the fun times during the last four years.

A special thanks also goes to everyone who read and commented on this thesis: Christina Ernst, Jack Monahan, Hannah Meyer, Michael Morgan, Catalina Vallejos, and John Marioni.

I thank my examiners Chris Wallace and Ben Lehner for a truly enjoyable viva and the valuable discussions, which improved the presentation of this thesis.

Importantly, I want to thank my friends from Cambridge, Heidelberg and Essen for their moral support, all the memories and keeping me sane during stressful times. Omar, thank you for distracting me with pizza and comedy nights, endless trips to the Blue Moon and for being a good friend. Hannah, Jack, Christina, Catalina, Dani, Lara, Julia, it was always fun going to formals, punting or trying to discover as many pubs in Cambridge as possible.

Sebastian and Ruben, thank you for listening to my complaints and helping me whenever needed, and together with Jonathan, Felix, Georg, Fabian for years of friendship, and for the most exhausting but fun "Alumni Wochenenden".

I also want to thank Nico, Fabian, Josh, Martin, Marc, Alex, and Mike for always welcoming me back home. Even though this only happens few times per year, I truly value your friendship.

The Pembroke College Boat Club helped me to find a balance between sports and work and even during high-intensity phases I fully enjoyed the early mornings and the distraction that rowing offered.

Finally, I want to thank Bodi for her constant support and patience during the last four years. Our adventures, Skypes, holidays and weekends together gave me the energy to focus on my work.

Abstract

Transcriptional noise is an intrinsic feature of cell populations and plays a driving role in mammalian development, tissue homoeostasis and immune function. While expression heterogeneity, a phenotypic readout of transcriptional noise, has been broadly studied in prokaryotic model systems or by profiling individual genes, few whole-transcriptome studies in mammalian systems have been reported. The development of single-cell RNA sequencing technologies introduced powerful tools to investigate transcriptional differences between individual cells, therefore allowing the in-depth characterisation of expression variability. In this thesis, I computationally analysed single-cell RNA sequencing data to understand transcriptional variability and expanded a statistical model to avoid confounding effects when quantifying such variability. First, I profiled individual transcriptomes of $CD4^+$ T cells, identifying a global decrease in transcriptional variability upon immune activation. By extending this analysis across two sub-species of mice, I identified an evolutionarily conserved set of immune response genes for which transcriptional variability increases during ageing. I used a Bayesian modelling framework to quantify mean expression and transcriptional variability but due to a strong confounding effect between these two parameters, variability analysis was restricted to genes that are similarly expressed across the tested conditions. To address this problem, I extended the computational framework allowing the parallel assessment of changes in mean expression and variability. Within this Bayesian framework, I introduced a joint prior linking mean expression and variability parameters, which allowed a residual over-dispersion to be measured for each gene. This measure allowed me to statistically assess changes in variability even for genes with differences in mean expression between conditions. Finally, I applied the model to identify temporal changes in variability over the time-course of spermatogenesis. This unidirectional differentiation process involves several complex steps before mature sperm form from spermatogonial stem cells. When profiling changes in variability across this developmental time-course, peaks in variability are caused by rapid changes in gene expression along the differentiation trajectory. This thesis provides a deeper understanding of technical and biological factors that drive transcriptional variability and offers a basis for future research to characterise its role in health and disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

Ccl4	chemokine (C-C motif) ligand 4
Eif1	eukaryotic translation initiation factor 1
Fasl	Fas ligand
H2-Aa	H-2 class II histocompatibility antigen
Il2ra	interleukin 2 receptor alpha
PHO5	repressible acid phosphatase
Sir	sirtuin
TDH3	glyceraldehyde-3-phosphate dehydrogenase 3
2i	2 inhibitor
A	type A spermatogonia
a2i	alternative 2 inhibitor
Akap4	A-kinase anchoring protein 4
A. queenslandica	*Amphimedon queenslandica*
AML	acute myeloid leukemia
B. subtilis	*Bacillus subtilis*
B	type B spermatogonia
B6	*Mus musculus domesticus*
BASiCS	Bayesian Inference of Single-Cell Sequencing data
BCR	B cell receptor
C. elegans	*Caenorhabditis elegans*

CA	*Cornu Ammonis*
Cabs1	calcium binding protein, spermatid associated 1
Carm1	coactivator associated arginine methyltransferase 1
Cas9	CRISPR-associated protein 9
CAST	*Mus musculus castaneus*
CAVI	coordinate ascent mean-field variational inference
cDNA	complementary DNA
CGI	CpG islands
CNV	copy number variation
CpG	5'-cytosine–phosphate–guanine-3'
CPM	counts per million
CRISPR	clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
CUT&RUN	cleavage under targets & release using nuclease
CV^2	squared coefficient of variation
Cxcr5	C-X-C chemokine receptor type 5
Cypt1	cysteine-rich perinuclear theca 1
CyTOF	cytometry by time-of-flight
DamID	DNA adenine methyltransferase identification
DC	dendritic cells
Dmrt2	doublesex and mab-3 related transcription factor 2
D. melanogaster	*Drosophila melanogaster*
DR-Seq	DNA and mRNA sequencing
DroNc-Seq	massively parallel single-nuclei sequencing with droplet technology
DS	diplotene spermatocyte
DSB	double strand breaks
E. coli	*Escherichia coli*
E	embryonic day
EC	endothelial cell
EFDR	expected false discovery rate
ELBO	evidence lower bound

EM	effector memory
ERCC	External RNA Control Consortium
ERK	extracellular signal–regulated kinase
FACS	fluorescence-activated cell sorting
Fgf	fibroblast growth factor
FPKM	fragments per kilobase per million mapped reads
FPR	false positive rate
FSC	forward scatter
G&T-Seq	genome and transcriptome sequencing
GABA	gamma-aminobutyric acid
Gata	GATA binding protein
gDNA	genomic DNA
GEM	gel beads in emulsions
GESTALT	genome editing of synthetic target arrays for lineage tracing
GO	gene ontology
GRBF	Gaussian radial basis function
GRCm38	Genome Reference Consortium mouse build 38
H2afx	H2A histone family member X
H3K27me3	tri-methylation of lysine 27 of histone H3
H3K36me3	tri-methylation of lysine 36 of histone H3
H3K4me1	mono-methylation of lysine 4 of histone H3
H3K4me2	di-methylation of lysine 4 of histone H3
H3K4me3	tri-methylation of lysine 4 of histone H3
H3K9ac	acetylation of lysine 9 of histone H3
H3K9me3	tri-methylation of lysine 9 of histone H3
H3R26	histone H3 arginine-26
H4K16ac	acetylation of lysine 16 of histone H3
H4K20me3	tri-methylation of lysine 20 of histone H3
HiC	high-troughput chromosome confirmation capture
HIV	human immunodeficiency virus

ICM inner cell mass
IFC integrated fluidic circuit
Ifn interferon
Ikzf4 IKAROS Family Zinc Finger 4
IL immature Leydig
Il interleukin
Il7r interleukin 7 receptor
ILC innate lymphoid cell
In intermediate spermatogonia
IP immunoprecipitation
iPSC induced pluripotent stem cell
IVT *in vitro* transcription

KL Kullback-Leibler divergence
Klrg1 killer cell lectin-like receptor subfamily G member 1

LIF leukemia inhibitory factor
log2FC $\log_2$ fold change
LPS lipopolysaccharide
LS leptotene spermatocyte

M metaphase I and II
m6A N^6-methylation of adenosine
MACS magnetic-activated cell sorting
MALBAC multiple annealing and looping-based amplification cycles
MARS-Seq massively parallel RNA single-cell sequencing
MCMC Markov Chain Monte Carlo
MDA multiple displacement amplification
Meox2 mesenchyme homeobox 2
MERFISH multiplexed error-robust fluorescence *in situ* hybridization
mESC mouse embryonic stem cell
MHCI major histocompatibility complex class I

miRNA	micro RNA
MLP	multilayer perceptron
M. leidyi	*Mnemiopsis leidyi*
mRNA	messenger RNA
MSCI	meiotic sex chromosome inactivation
MTase	methyltransferase
NB	negative binomial
N. vectensis	*Nematostella vectensis*
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NFAT	nuclear factor of activated T cells
NK	natural killer
Nr	nuclear receptor
P	post-natal
PBMC	peripheral blood mononuclear cell
PCA	principal component analysis
PCR	polymerase chain reaction
PD-1	programmed cell death protein 1
PD-L1	programmed death-ligand 1
pgt	*pseudo* ground truth
P. dumerilii	*Platynereis dumerilii*
Pl	preleptotene spermatocyte
PRC	polycomb repressive complex
Prm	protamine
PS	pachytene spermatocyte
PSM	pre-somitic mesoderm
PTM	peritubular myoid
QC	quality control
qPCR	quantitative PCR
RA	retinoic acid

Rnf8	ring finger protein 8
RNAPII	RNA polymerase II
rRNA	ribosomal RNA
RT	reverse transcription
RT-PCR	real time PCR
RTase	reverse transcriptase
RTE	recent thymic emigrant
S	speramtid
SC	spermatocyte
Scml2	sex comb on midleg-like 2
sc5hmC-Seq	single-cell 5-hydroxymethylcytosine sequencing
scATAC-Seq	single-cell assay of transposase-accessible chromatin using sequencing
scBS-Seq	single cell bisulfite sequencing
scChIP-Seq	single-cell chromatin IP followed by sequencing
scDNA-Seq	single-cell whole genome sequencing
sci-Seq	single-cell combinatorial indexed sequencing
scLVM	single-cell latent variable model
scM&T-Seq	single-cell methylome and transcriptome sequencing
SCOMP	single-cell comparative genomic hybridization protocol
scRNA-Seq	single-cell RNA sequencing
scRRBS-Seq	single-cell reduced representation bisulfite sequencing
scVI	single-cell variational inference
SG	spermatogonia
sgRNA	single guide RNA
siRNA	small interfering RNA
SM	somitic mesoderm
smFISH	single molecule fluorescence *in situ* hybridization
SNN	shared nearest-neighbour
SNV	single nucleotide variant
SP	single positive
SPLiT-Seq	split-pool ligation-based transcriptome sequencing
Ssxb	synovial sarcoma, X member B, breakpoint

SSC	side scatter
Stra8	stimulated by retinoic acid 8
STORM	stochastic optical reconstruction microscopy
STRT	single-cell tagged reverse transcription
Tbx21	T-box 21
TCR	T cell receptor
TF	transcription factor
TFBS	transcription factor binding site
Tfh	T follicular helper
Tgf	transforming growth factor
Th	T-helper
Tigit	T cell immunoreceptor with Ig And ITIM domains
tMg	testicular macrophage
Tnf	tumour necrosis factor
Tnp	transition protein
TPE-OLD	telomere position-effect on long distance
TPR	true positive rate
T. adhaerens	*Trichoplax adhaerens*
TRAIL	TNF-related apoptosis-inducing ligand
Treg	regulatory T cell
Tsga8	testis specific gene A8
tSNE	t-distributed stochastic neighbour embedding
TSO	template-switch oligo
TSS	transcriptional start site
Tyk2	tyrosine kinase 2
UMI	unique molecular identifier
WGA	whole genome amplification
X:A	X chromosome:Autosome

Zfy zinc finger protein Y-linked
ZIFA zero-inflated factor analysis
ZINB zero-inflated negative binomial
ZS zygotene spermatocyte

1

Introduction

The intrinsic stochasticity of biochemical reactions introduces phenotypic heterogeneity in seemingly homogeneous populations of cells. This phenomenon has been widely studied in prokaryotic and eukaryotic systems and the functional role of phenotypic variation in development, health and disease is the subject of ongoing research. Biological noise, defined as stochasticity in biochemical reactions within individual cells, contributes to form molecular phenotypic variability in cell populations. Intrinsic noise summarises stochastic differences in transcription and translation between individual genes [1–3]. Extrinsic noise on the other hand arises due to fluctuations in cellular states (e.g. cell cycle, cell-to-cell signalling and metabolism) [4–6]. Recent technological advances allow the in-depth analysis of molecular phenotypic variability as proxy for biological noise in cell populations. Imaging methodologies [7] and single-cell "omics" techniques [8] permit the quantification of thousands of mRNA species, the genomic sequence, its epigenetic modification, and selected sets of proteins per cell. Moreover, the development of multi-omics technologies introduced the possibility to link cell-to-cell variation between multiple regulatory layers across individual cells [9]. With the emergence of single-cell RNA sequencing (scRNA-Seq) technologies, new computational strategies to quantify variability were introduced [10–15]. Applying high-throughput scRNA-Seq to mammalian systems characterised the functional role of molecular variability in healthy as well as diseased contexts. Recent studies have described changes in transcriptional variability at different stages during embryonic development, which hints at stochastic contributions to early cell fate decisions [16–18]. On the one hand, phenotypic variation in immune cells, potentially driven by transcriptional noise, increases cellular plasticity and facilitates the population response to pathogens [19, 20]. On the other hand, genetic and non-genetic heterogeneity within cell populations was described as driver for cancer development [21] and disrupts immune responses in aged animals [22]. Here, I introduce noise as an inherent feature of biological system and discuss its positive and negative consequences in cell populations. Furthermore, I outline recent developments of single-cell sequencing and imaging technologies and comment on robust approaches for quantifying molecular phenotypic variability. Finally, I will summarise Bayesian inference as a powerful statistical framework to model transcriptional variability from scRNA-Seq data. ■

1.1 Biology of expression noise

The intrinsic stochasticity of biochemical reactions contributes to a wide distribution of messenger RNAs (mRNAs) and proteins across a seemingly homogeneous populations of cells [1]. In the scientific litearature, this phenomenon is often referred to as “biological noise” (see **Box 1**). All cellular systems are exposed to varying levels of noise and employ strategies to make use of or cope with this source of variation. The sources and consequences of biological noise have been studied in an array of viral, prokaryotic and eukaryotic systems [23–25]. Across these systems the extent of its function remains unclear.

Box 1: Defining biological noise

Biological noise in cell populations is defined as stochastic effects on transcription and translation that propagates to form cell-to-cell phenotypic differences. To understand noise, one needs to distinguish between different sources of cell-to-cell variability in multiple measurable factors. On the broadest level, differences between single cells in a population can arise from structured and unstructured sources. When capturing cell populations that contain discrete cell states and/or cell types [26–28], measuring cell-specific features results in the detection of non-stochastic but rather correlated (structured) differences between individual cells. When the cell population structure is not driven by correlated features (unstructured variation), continuous processes (e.g. differentiation) can be the dominating source of cell-to-cell phenotypic variability [29]. Computational approaches allow the detection of these trajectories (e.g. via principal component analysis (PCA) or pseudotime inference [30, 31]). Therefore, my work and work of others [32, 33] focus on studying "molecular phenotypic variability", independent of measurement errors, in homogeneous cell populations as proxy for biological noise.

Classically and specifically in populations of bacteria [1], biological noise has broadly been classified into intrinsic and extrinsic noise. Intrinsic noise originates from stochastic biochemical effects that directly influence mRNA and protein expression gene-specifically (e.g. transcription factor (TF) binding dynamics, see [34]). Extrinsic noise on the other hand introduces co-variation across multiple genes (also in a pathway specific manner [35]) due to variations in cell-specific factors such as stress response, mitochondrial maintenance, amino-acid synthesis [36] or cell cycle [4]. Within a population of bacteria, intrinsic noise can therefore be measured as expression differences between co-regulated genes in one cell, while extrinsic noise is measured as co-regulated variance in gene sets across all cells. In multicellular systems however, the observed molecular phenotypic variability is a combination of stochastic (noise) and deterministic effects, which are difficult to delineate.

When discussing the role of noise in biological systems, it is crucial to differentiate between unicellular systems (prokaryotes, viruses, and yeast) and higher, multicellular eukaryotic systems that show complex signalling events. Furthermore, measuring the stochastic component of biological noise is difficult and requires time-resolved reporter gene read-outs in truly homogeneous cell populations [1]. Due to this, the majority of studies presented in this chapter use the observable molecular phenotypic variation in form of single-cell transcriptomic or proteomic read-outs as proxy for biological noise (see **Box 1**). This variation is confounded by unobserved deterministic processes (e.g subtle cell-cycle variation) and delineating the stochastic and deterministic component is challenging.

1.1.1 Bet-hedging in unicellular systems

Biological noise has been proposed to trigger the differential decision between latency and replication in viruses such as human immunodeficiency virus (HIV) and the λ-phage. In the case of the λ-phage, infected cells either reside in a lysogenic state where the genetic material of the virus is transmitted to daughter cells without inducing cell death, or a lytic state where the virus destroys the host cell **(Fig. 1.1)** [37]. Previous studies have shown that the lysis-lysogeny switch in λ-phage is driven by intrinsic and extrinsic noise [38, 39]. This idea has been extended by Zeng *el al.*, 2010 where the lysis-lysogeny switch does not depend on a single noise-driven decision but on the sum of all individual phages per cell [40]. In general, by summing across stochastic events or if the lysis-lysogeny decision can be predicted based on cellular volume, the switch does not occur as stochastically as initially anticipated. In the case of HIV, the virus either rapidly replicates or resides in a long-lived latent state from which the virus can switch to replication [41]. It has been shown that combining noise-enhancing and activating drugs shifts latent viruses into the active-replication state that can be targeted by anti-retroviral therapeutics [42]. Independent of the stochastic contribution to the latency-replication switch, this study present one of the first approaches to modulate phenotypic variability of a biological system to enhance therapeutic efficiency.

In unicellular organisms, biological noise has been linked to 'bet-hedging' strategies, where a sub-optimal fitness landscape is tolerated across a population of cells in order to facilitate an effective response to environmental changes. Here, phenotypic heterogeneity facilitates the commitment to alternative cell states in cases of stress (e.g. nutrient deprivation, temperature fluctuations). For example, *Bacillus subtilis* (*B. subtilis*) either commits to sporulation or competence upon starvation or DNA damage. Sporulation describes an irreversible process during which vegetative growth ends and the cell forms endospores that survive the altered

environment. Competent bacteria on the other hand take up DNA from endospores to repair DNA damage [43]. The probabilistic and transient activation of competence in a sub-population of *B. subtilis* cells is modulated by fluctuations in the competence regulators ComK and ComS. An excitable system of negative and positive feedback loops controls the number of cells that reversibly commit to competence while other cells irreversibly execute sporulation [44]. Variations in the process of transferring phoshporyl groups across a cascade of regulators maintains a constant probability for cells committing to sporulation under nutrient deprived conditions [45]. A similar phenomenon is observed in *Escherichia coli* (*E. coli*) populations exposed to antibiotics where pre-existing phenotypic heterogeneity allows some cells to resist antibiotic treatment. Once regrown, these cells remain sensitive to the antibiotic [46].

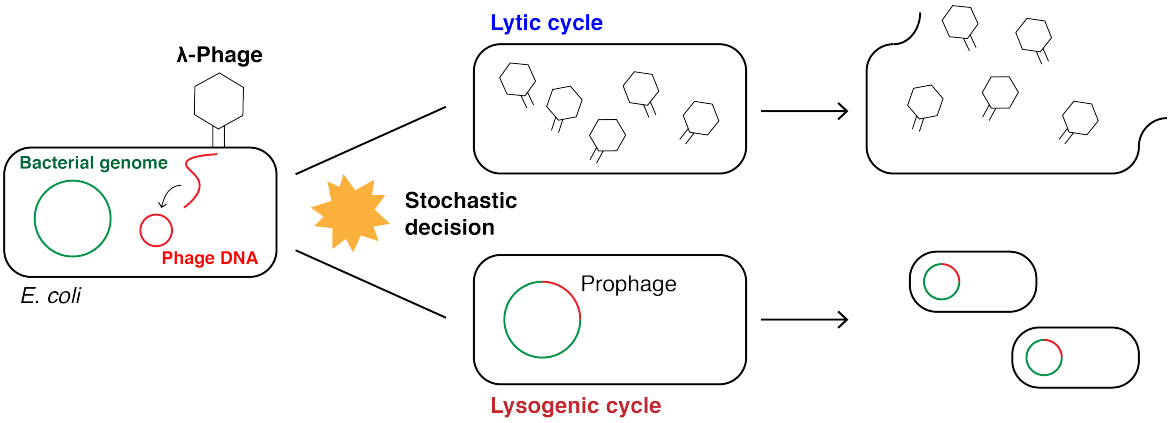


Fig. 1.1: Bet-hedging strategy of the λ-phage.
The linear genome of the λ-phage enters the *E. coli* host cell and circularises. A stochastic decision is made to enter the (i) lytic or (ii) lysogenic cycle where (i) the λ-phage genome replicates, the λ-phage particles assemble in the host cell and the cell is destroyed releasing the virions or (ii) the λ-phage is integrated into the host genome (prophage) and transferred to daughter cells during cell divisions. Under stress conditions, the λ-phage genome is excised from the host genome and enters the lytic cycle.

Similar to phenotypic heterogeneity in unicellular prokaryotes, transcriptional noise facilitates the switching between mating phenotypes in yeast upon exposure to pheromones [47]. Comparably, commitment to utilising galactose as a nutrient source is a cell fate transition, which is facilitated by stochastic gene expression [48].

In these systems, biological noise introduces variation in mRNA and proteins that increase plasticity for cells to adapt to changeing environments. However, to control and balance the number of cells that commit to a specific fate, noise needs to be buffered, for example, by the regulatory network of feedback loops controlling sporulation and competence.

1.1.2 Development and differentiation

Similar to bet-hedging strategies in unicellular organisms, noise can facilitate the switch between cell states and the probabilistic induction of differentiation processes [25, 49]. However, as mentioned above, measuring biological noise in differentiating multi-cellular organisms is challenging and the observed molecular phenotypic variability is a combination of stochastic and deterministic components. It has been shown that transcriptional variability increases throughout differentiation [50] and development [51]. Dissecting differentiation processes of haematopoietic progenitor cells revealed an increase in transcriptional variability directly before cell fate decisions are made [52, 15]. Once committed, differentiating cell populations collapse in variability and move towards a new attractor state. These studies highlight a possible contribution of molecular variability to cell fate decision event. However, the observed change in variability within differentiating cell populations is purely correlative and it is not possible, with these experiments, to differentiate between variability causing differentiation or differentiation causing variability.

Studies of recent years proposed that stochasticity in expression contributes to early (pre-implantation) embryonic development, and to gastrulation [53]. As early as the 4-cell stage embryo, targets of master pluripotency factors Oct4 and Sox2 are heterogeneously expressed **(Fig. 1.2, left panel)**. This is caused by heterogeneous methylation patterns of histone H3 arginine-26 (H3R26) induced by coactivator associated arginine methyltransferase 1 (Carm1), which in turn facilitates the binding of Oct4 and Sox2 to induce pluripotency. Cells with unmethylated H3R26 differentiate towards the extra-embryonic trophoectoderm while pluripotent cells form the inner cell mass [16]. Once the cells compact at embryonic day (E) 3.5, cells of the inner cell mass (ICM) stochastically express genes to initiate heterogeneity within the cell population **(Fig. 1.2, 2nd panel)**. Fgf4 driven signal reinforcement controls this heterogeneity to form a salt-and-pepper like cell state pattern at E3.5. Positional information and the establishment of gene regulatory networks facilitate the segregation of the epiblast and primitive endoderm lineage at E4.5 **(Fig. 1.2, 3rd panel)** [18]. In line with this, scRNA-Seq revealed high levels of transcriptional variability in the uncommitted inner cell mass at E3.5 (64-cell stage) in comparison to the E4.5 committed epiblast. Transcriptional variability increases again upon exit from pluripotency in the E6.5 epiblast while cells of the primitive streak at E6.5 synchronise their expression patterns and variability is reduced **(Fig. 1.2, right panel)** [17].

Besides the hypothesis of transcriptional variation contributing to embryonic development, a number of alternative drivers for cell fate decisions the mouse embryo exist [54]. For

example, in the 8-cell to 16-cell stage embryo, symmetry breaking could be achieved by an interaction between the cell's position and polarity, its cortical tension, and the orientation of cell division [54]. Maître *et al.* proposed a system where robust self-organization of 8- to 16-cell stage embryos is achieved by differences in contractility between polar and apolar cells, which leads to the internalization of the more contractile apolar cells [55]. Taken together, it is unclear to which fraction transcriptional variability plays a role in cell fate decision-making and if purely the occurrence of differentiating cells induces transcriptional variation.

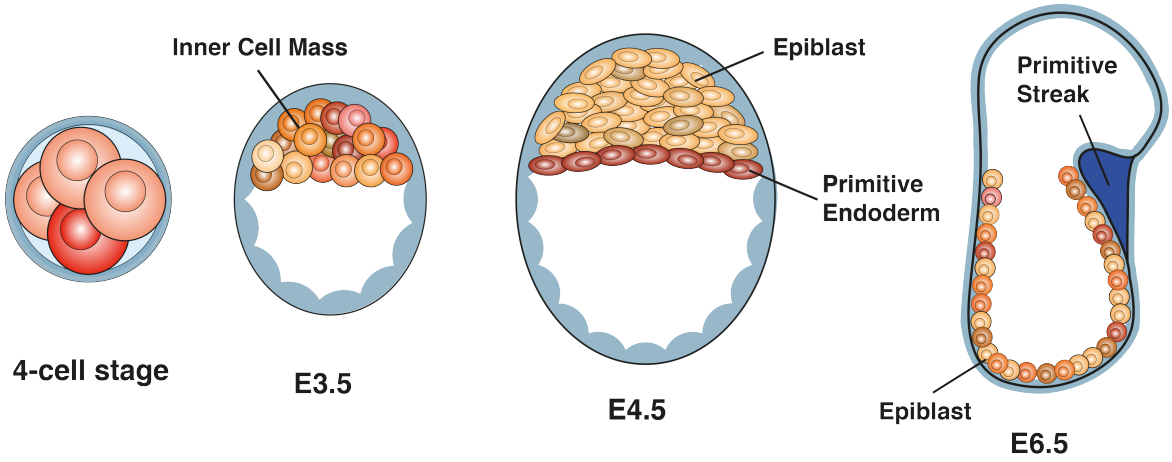


Fig. 1.2: Progression of transcriptional noise during embryonic development.
From left to right: schematic of mouse embryonic development from the 4-cell stage to early gastrulation at E6.5. Cell colours indicate gene expression strength. Variable expression at the 4-cell stage induce commitment to form extra-embryonic lineages or pluripotent cells. These pluripotent cells at E3.5 show high expression variability forming the inner cell mass (ICM). Cells rearrange to form the epiblast and primitive endoderm at E4.5 while noise levels increase in the epiblast at E6.5 compared to the primitive streak.

While pluripotent stem cells in the mouse embryo commit irreversibly to cell lineages during development, *in vitro* cultured mouse embryonic stem cells (mESCs) reside in a self-renewing, metastable state [56] and heterogenity within the cell population depends on the growth condition. Transcription factor heterogeneity, especially of the pluripotency regulator Nanog, is highest in leukemia inhibitory factor (LIF)/serum grown cells and allows the Nanog-negative cells to commit to differentiation [57, 58]. Heterogeneously expressed genes that show a bimodal distribution in expression counts correlate with each other indicative of the presence of distinct states in mESCs. These distinct states show differences in promoter methylation patterns, introducing the role of epigenetic modifications to maintain heterogeneity in mESCs [59]. In-depth analysis of mESCs grown in different media

(serum, 2 inhibitor (2i) and alternative 2 inhibitor (a2i)) shows the presence of three distinct cell states in the serum grown cells. mESCs grown in 2i media show less variability in pluripotency markers but higher heterogeneity in cell cycle related genes [12]. From the pluripotent ground state, mESCs can differentiate along somatic lineages via specific differentiation events or noise-induced transitions between attractor states. Mathematical modelling has shown that mESCs differentiate stochastically through distinct hidden cell (micro-)states within a defined (macro-)state coupled to an increase in variability [50]. In contrast to the beneficial features of noise in stem cell differentiation, stochastic events during induced pluripotent stem cell (iPSC) reprogramming limit the formation of single iPSCs [60, 61]. It has been shown that probabilistic events dominate in an early phase of reprogramming while the transcription of *Sox2* induces a later, more deterministic, phase [62].

These findings indicate an intrinsic heterogeneity of pluripotent cell populations. Extrinsic cues, such as growth medium or signalling networks in the embryo, are needed to control this heterogeneity. However, it is not clear if this seemingly random expression of pluripotent marker genes is truely stochastic or driven by unobserved regulatory mechanisms. Hoppe *et al.* challenged the idea of lineage choice by stochastic fluctuations of lineage-specific transcription factors and highlighted, using time-resolved measurements, that these transcription factors are solely reinforcing lineage choice [63]. Therefore, lineage choice can be initiated by unobserved cues that induce variation in genes expression.

1.1.3 Stochasticity in immune responses

Fast and flexible immune responses are only possible within cell populations that show high plasticity and react to a broad spectrum of stimuli. Stochasticity in cytokine expression can lead to phenotypic variability in the T-helper (Th) cell repertoire and increases the effectiveness to respond upon immune stimuli [64]. For example, fluctuating expression of the lineage defining cytokines interferon (Ifn)γ for Th1 and interleukin (Il) 4 for Th2 in small populations of cells drive the cell population towards a Th1 or Th2 cell fate while most cells co-express the lineage defining transcription factors GATA binding protein (Gata) 3 and T-box 21 (Tbx21) [65, 66].

Furthermore, Shalek *et al.*, 2014 have shown that upon lipopolysaccharide (LPS) stimulation a small subset of dendritic cells become activated much earlier than the rest of the cell population while expressing Ifnβ. These early responders support the activation of late responding cells via cell-to-cell communication (paracrine signalling) and self-stimulation via autocrine signalling **(Fig. 1.3)** [19]. Likewise, a bimodal (digital) expression of Il2 is

detected in Th cells after immunisation where the number of Il2 expressing cells scales with antigen level. Il2 expressing cells support the activation of surrounding cells via paracrine signalling [67]. Similarly, digital activation processes can be observed in the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) signalling pathway. The fraction of cells that activate this signalling pathway increase with LPS concentration to avoid strong immune activation at low concentrations of a stimulus [68].

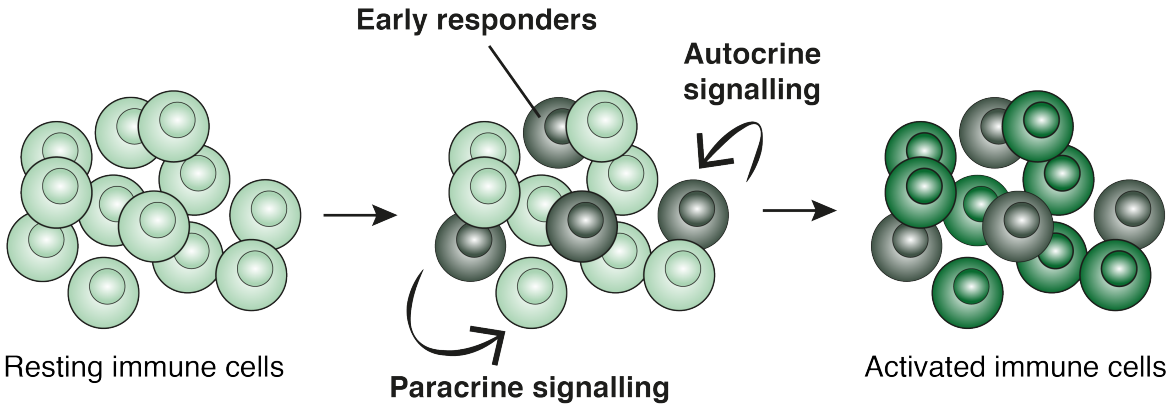


Fig. 1.3: Early responders are important for homogeneous immune activation.
Within a population of immune cells (e.g. dendritic cells (DC), Th cells), a sub-population either show higher response strength or induce the production of cytokines such as Il2 or Ifnβ. These early responders induce activation of surrounding cells via paracrine signalling and self-stimulation via autocrine signalling.

While the plasticity and reactivity of immune cell populations is finely tuned by introducing phenotypic heterogeneity, it is not understood how individual cells commit to each phenotype. In part, stochastic expression introduces molecular phenotypic variability that in turn is tightly controlled by external and internal signalling networks. It will therefore be crucial to study the behaviour of immune cells while incorporating their spatial location which might allow the prediction of each cell's phenotype [69].

1.1.4 Tissue development and homeostasis

Coping with the influence of biological noise is important for regulated tissue development and homeostasis. An early study showed that in order to minimise the effect of stochasticity in development, plants express heat-shock protein 90 to stabilise regulators of growth and development [70]. Furthermore, redundancy in the *Caenorhabditis elegans* (*C. elegans*) intestinal gene regulatory network buffers variability in the down-stream master regulator *elt-2*. Once highly connected regulators of this network are removed, phenotypic variation arises from bimodal expression of *elt-2* [23]. The cooperation of positive and negative

feedback loops in these highly connected regulatory networks ensure robust expression of key developmental genes [71]. Other models have been proposed in which noise helps to form sharp boundaries between neighbouring domains [72]. Contact based adhesion and repulsion between cells sharpens narrow transition regions by sorting cells within a tissue across small scales. Noise-driven cell state plasticity on the other hand allows cells to switch states and therefore helps narrow a wider transition region [73]. The plasticity to migrate within a population of cells also allow the correction of sensing errors. These errors are induced by either too strong or too weak responses of individual cells to a signalling gradient [74].

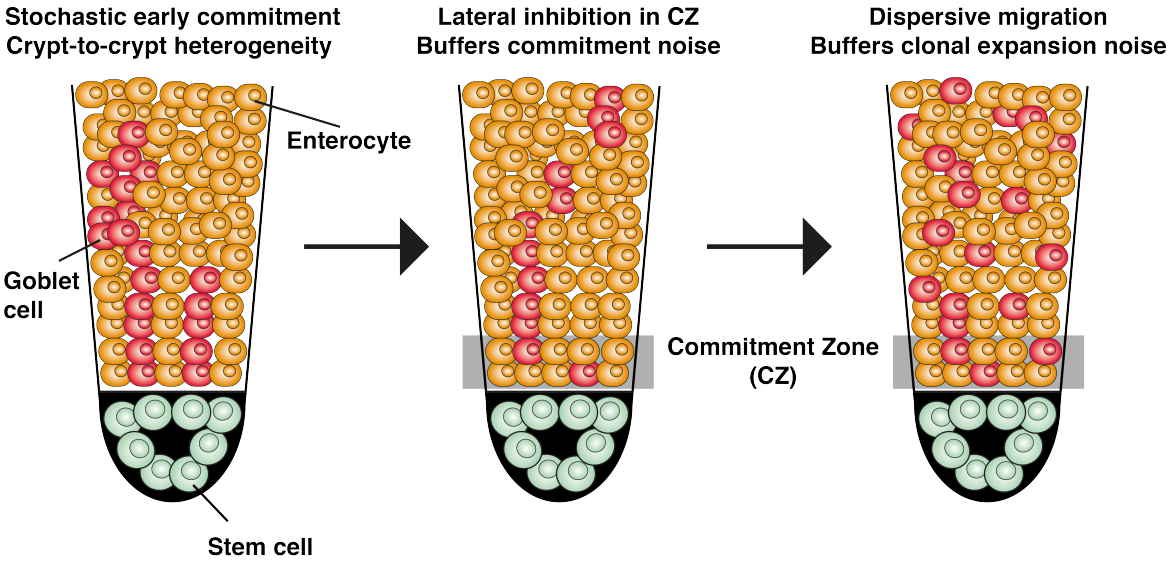


Fig. 1.4: Buffering of noise in the colonic crypt.
Each colonic crypt harbours 6-8 stem cells that divide to form stem cells and progenitors cells. Early commitment of progenitor cells to cell fates (e.g. goblet cells or enterocytes) leads to crypt-to-crypt heterogeneity due to commitment noise (left panel). Lateral inhibition within a restricted zone (commitment zone) allows cell fate switching and therefore buffers the crypt-to-crypt heterogeneity (middle panel). Migration of goblet cells after the commitment zone buffers the stochastic occurrence of goblet cell or enteroctye patches within the crypt and allows a constant ratio of 1:3 goblet cells to enterocytes in each crypt (right panel). Adapted from [75].

While the cell division rate within tissues is higher during development, tissue homeostasis is maintained by stochastic events that balance cell division and apoptosis [76]. The effect of noise on maintaining tissue homeostasis has been studied in a diverse set of organs. In fat tissue, a complex system of signalling feedback loops controls protein abundance noise to induce differentiation at a low rate but prevents stochastic de-differentiation [77]. To maintain coordination in liver function, longer mRNA lifetimes of bursty genes and polyploidy reduce noise in gene expression [78]. Another mechanism to achieve tissue-wide expression

responses involves spatial coordination of stochastically expressing cells in the pituitary gland [79]. Spatially constrained signalling events have also been demonstrated to play a role in maintaining colonic crypt cell type diversity. Per crypt, eight stem cells differentiate into a defined ratio of cell types. To reduce noise in this process, lateral inhibition within a commitment zone reduces the number of differentiated goblet cells and following slower dispersive migration as well as decreased division rates of goblet cells ensures a distinct 1:3 ratio to enterocytes across all crypts **(Fig. 1.4)** [75].

In sum, phenotypic heterogeneity in tissues can arise from stochastic expression driven by noise. To control for correct tissue responses, signalling networks are in place to modulate this variation. In most studies, individual signalling networks and few molecular read-outs were chosen to understand the variation observed within tissues. However, a combination of multiple regulatory signalling events control cell fate within tissues and disentangling the individual components has not been feasible.

1.1.5 Evolution

As discussed above, biological noise can be beneficial for cell fate commitment but needs to be controlled to allow coordinated expression in cell populations. During evolution, a trade-off between cellular plasticity, the expression responsiveness during environmental changes, and robust expression formed. Natural selection acts on genetically controlled expansions of quantitative phenotypes, which, in part, are derived from biological noise [25]. For example, variable expression of stress response genes allows a cell population to adapt to changing environments [80]. Specifically, the expression of genes controlled by TATA-box containing promoters shows strong divergence between species [81]. To control for robust expression levels once selection becomes stabilising, noise levels are reduced [80, 25, 82].

Lehner, 2008 discussed specifically evolutionary selection to minimise noise in genes that show harmful phenotypic effects upon alteration ("dosage-sensitive genes"). These genes show low expression variability to reduce the probability of altered expression and also lower expression divergence between species [83]. Furthermore, essential genes tend to cluster in the genome in regions with persistent open chromatin to reduce the effect of noise [84]. In line with this, the promoters of core cellular components show a decoupling between expression plasticity and expression variability, which indicates that responsiveness in expression is not a general attribute of high expression variability [85].

In unicellular populations, it has been proposed that the contribution of noise on molecular phenotypic variability evolutionarily increased as a form of rudimentary regulation [86]. As a consequence, phenotypic heterogeneity increases the adaption rate of cell populations to extreme environments [87]. Conversely, in multicellular organisms, collections of cells need to respond in a coordinated manner. It has therefore been proposed that nuclear compartmentalisation in higher organisms reduces noise by mRNA retention at the nuclear membrane [88, 89].

In most cases, cells in an unperturbed state have been profiled to decipher evolutionary selection acting on variability in gene expression. However, in fluctuating environments where the averaged protein abundance across a cell population is far from the optimum, variability in expression leads to some cells expressing protein levels closer to the optimum. By contrast, in stable environments, noise in gene expression can be deleterious by leading to suboptimal growth conditions for many cells [90, 91]. It is therefore crucial to discuss the fitness effect of changes in molecular variability in the context of fluctuating as well as stable environments.

1.1.6 Cancer

While biological noise can contribute to the adjustment of cells to new microenvironments, errors in the form of gene mutations induce transitions from healthy cells towards a cancer attractor state **(Fig. 1.5)** [21]. Non-genetic heterogeneity supports the phenotypic adaptation to the new attractor state [92]. The emergence of non-genetic heterogeneity in tumours is coupled to epigenetic dysregulation that allows the survival of cancer cells [93]. Furthermore, it has been proposed that genome wide intra-sample methylation heterogeneity is increased in chronic lymphoitic leukemia increasing cancer cell plasticity in the search for new attractor states [94]. Increased variability in expression can also be observed for more aggressive cancer sub-types across multiple patients [95].

An important consequence of increased phenotypic heterogeneity in cancer cells is the fractional killing of cell populations upon drug treatment **(Fig. 1.5)** [96]. Variability in proteins mediating TNF-related apoptosis-inducing ligand (TRAIL) induced apoptosis leads to the survival of small fractions of cells [97], which could consequently repopulate the tumour environment. Similarly, the stochastic acquisition of DNA damage upon cisplatin exposure introduces heterogeneity in the up-regulation of p53. Slow up-regulation leads to cell cycle arrest and inhibits apoptosis while only fast up-regulation leads to cell death [98]. In patient derived melanoma cells, sporadic expression of resistance markers forms a rare cell

population that grows into resistant colonies after treatment. While pre-resistant cells do not display epigenetic marks and are therefore close to the non-resistant ground state, treatment induces large epigenetic reprogramming, forming stable resistant cancer colonies [99]. To tackle this problem, combinatorial therapies have been proposed to reduce variability and fractional killing in cancer cell populations [98, 100].

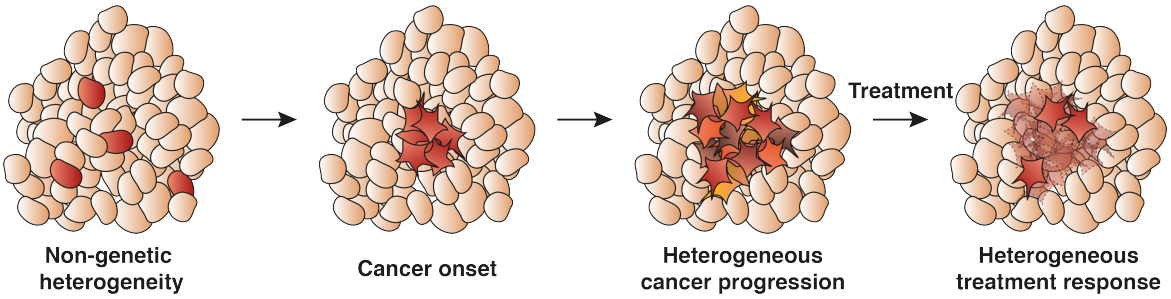


Fig. 1.5: Heterogeneous cell states and cell responses in cancer development.
Stochasticity in expression introduces non-genetic heterogeneity that supports the adaptation of cancerous cells. Cancer progresses to form a collection of cells with divergent expression patterns. This phenotypic heterogeneity leads to fractional killing during treatment and cancer recurrence.

These studies propose a contribution of non-genetic heterogeneity, potentially induced by the loss of noise control, to cancer onset and inefficient treatment response. However, cancer is a heterogeneous disease that develops in a multi-step process involving disregulation in various cellular systems [101]. Therefore, and similar to molecular phenotypic variability in embryonic development, the observed non-genetic heterogeneity can be a phenotypic consequence rather than a driver for cancer onset.

1.1.7 Ageing

Similarly to the onset of cancer, destructive roles of biological noise have been reported during organismal ageing. Previously, it has been debated whether expression noise changes during the lifespan of animals [102, 103]. While these initial studies only used small panels of genes, transcriptional profiling of single cells led to the discovery of a destabilised immune activation programme in $CD4^+$ T cells due to increased expression noise [22]. Similarly, transcriptional noise increases with age in human pancreas coupled to an increased stress signature and atypical hormone expression [104]. For further discussion of age-related effects on transcriptional noise, see **Chapter 2**.

Table 1.1: Positive and negative effects of biological noise on cellular systems.

System	Friend	Foe
Unicellular organism	Bet-hedging	
Development and differentiation	Probabilistic induction of cell differentiation	
Immune response	Plasticity in immune response Control of response strength	
Tissue development and homeostasis	Low cell differentiation rate	Non-uniform development Uncontrolled tissue response
Evolution	Adjustment to fluctuating environment	Non-uniform, stabilising expression Uncontrolled tissue responses
Cancer		Phenotypic adaption to cancer state Fractional killing of cancer cells
Ageing		Unsynchronised immune response Increased stress signatures

1.2 Sources of expression noise

Molecular phenotypic variability across homogeneous populations of cells can arise from intrinsic and extrinsic noise, and deterministic components (see **Box 1** on page 2). While intrinsic noise is promoter-specific and therefore induces uncoordinated variation in RNA or protein expression between individual genes, extrinsic noise globally influences gene expression across multiple cells and therefore leads to co-variation across larger sets of genes. Here, I give an overview on the different sources of intrinsic and extrinsic noise in a variety of biological systems.

1.2.1 Intrinsic noise

Intrinsic noise in cell populations arises from stochasticity in biochemical reactions that lead to the synthesis of mRNAs (transcription) and proteins (translation) within individual cells. Regulatory features on the genomic, epigenetic, transcriptional and translational level influence and control the strength of intrinsic noise (for an overview see **Fig. 1.6**).

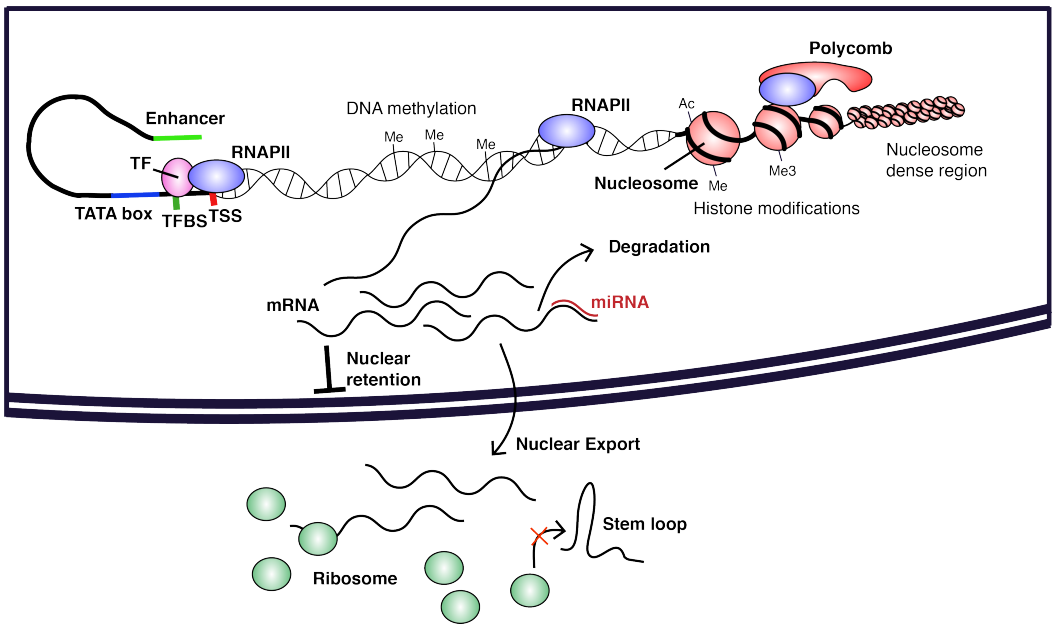


Fig. 1.6: Regulatory features that modulate expression noise.
Promoter sequence, number of transcription factor (TF) binding sites (TFBS), number of transcriptional start sites (TSS), enhancer elements, RNA polymerase II (RNAPII) loading, DNA methylation, nucleosome positioning, histone modifications, polycomb repressive complex binding, micro RNAs (miRNAs), nuclear export of mRNA, ribosome binding and blockage via stem loop formation are features that induce gene-specific intrinsic noise.

DNA features

One of the key regulatory steps prior to RNA synthesis is the binding of TFs to specific DNA sequences within the regulatory region (promoter) of a gene which then triggers the controlled production of primary RNA transcripts from the DNA of this gene [105]. Mutations in the DNA sequence such as single nucleotide variants (SNVs) can alter the binding affinity of TFs and therefore the rate at which a gene is expressed **(Fig. 1.7)**. A systematic study of the glyceraldehyde-3-phosphate dehydrogenase 3 (*TDH3*) gene expression in yeast found that mutations in known transcription factor binding sites (TFBSs) decrease mean expression and increase expression noise. Moreover, Metzger *et al.*, 2015 proposed that evolutionary selection removes mutations that increase expression noise and that SNVs with large effects on expression noise show the lowest frequency within sampled yeast strains [106]. However, the authors examined one promoter in stable environmental conditions. How selection on mutations that induce variability in expression works in more complex systems and across multiple promoters is still unexplored.

One of the most widely studied DNA motifs in relation to transcriptional noise is the TATA-box motif in promoters. Generally, TATA-box containing promoters show high levels of transcriptional noise **(Fig. 1.7)** [32], possibly due to a simple activation cycle containing one or few inactive states [107]. Moreover, TATA-box containing genes show an increased interspecies variability [81] and higher spontaneous mutational variation [108], indicating an increased evolvability of these particular genes. In an early study, Raser *et al.*, 2004 studied the noisy expression controlled by the budding yeast repressible acid phosphatase (*PHO5*) promoter. This promoter contains the TATA-box motif and it has been shown that transcriptional noise is reduced when a mutational modification decreases the TATA-box strength [2]. A more recent study confirmed this result and found mutations in yeast promoters that eliminate the TATA-box motif which lead to reduced noise levels for these genes [109]. The TATA-box is therefore one genomic feature that can differentiate between genes with variable and stable expression and are enriched amongst stress response genes, which support their role in early adjustment to changing environmental conditions [80]

However, a possible confounding factor for the increased noise of TATA-box containing promoters is the number of TFBSs. Tirosh *et al.*, 2006 detected a two-fold enrichment of TFBSs in TATA-box containing promoters [81]. A later study showed that transcriptional noise scales with increased numbers of TFBSs **(Fig. 1.7)** [110]. Furthermore, TATA-box containing genes lack enhancing histone marks and their increased variability in expression can therefore be explained by repressed chromatin [111] (see **Section 1.2.1**).

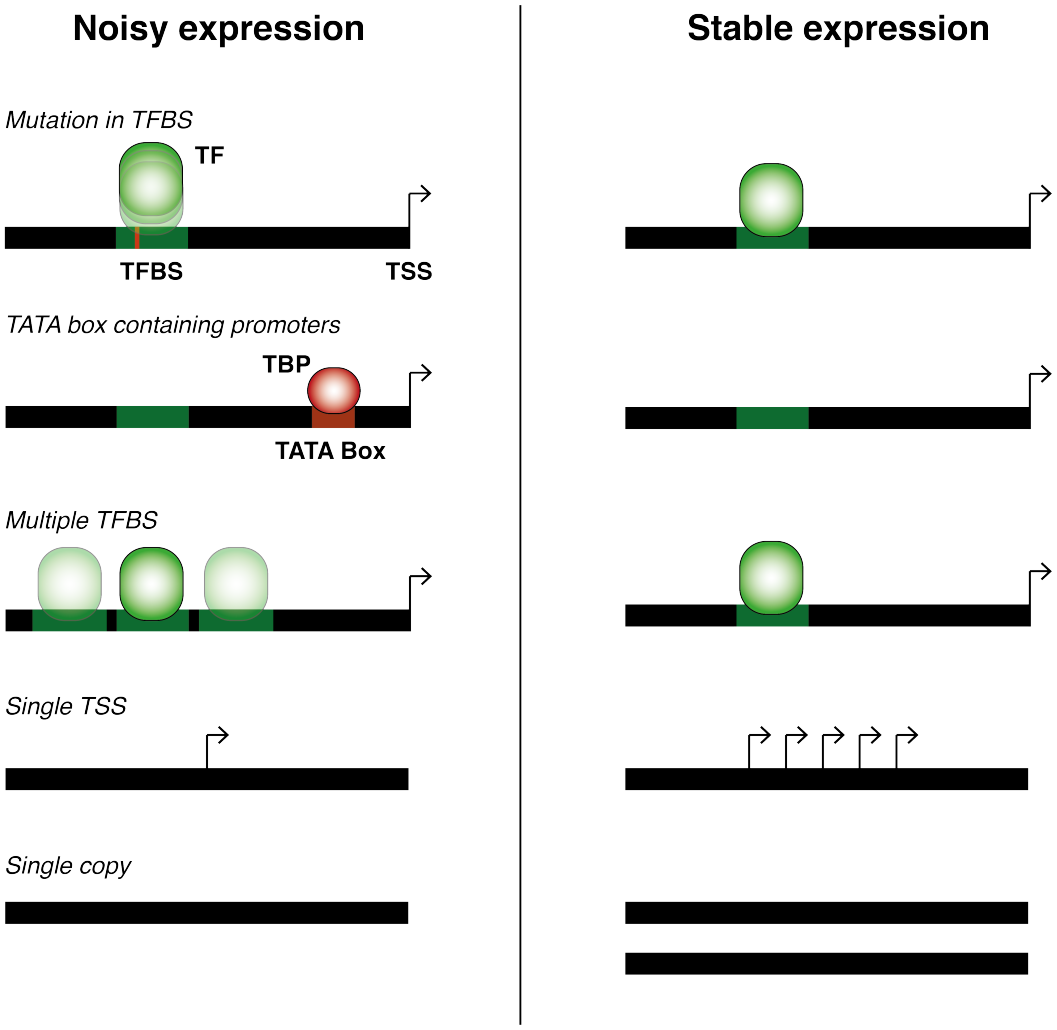


Fig. 1.7: Features of the DNA sequence induce expression noise.
Mutations of the transcription factor (TF) binding site (TFBS), the presence of a TATA box, increase number of TFBSs, reduced number of transcriptional start sites (TSSs) and reduced copy number of genes can induce transcriptional noise.

Promoters can be classified based on their shape as narrow, with few transcriptional start sites (TSSs) that predominantly control tissue-specific gene expression, and broad promoters with larger numbers of TSSs that control the expression of house keeping genes. Mutations that alter the shape of promoters increase transcriptional variability [112]. Furthermore, promoters with one or few TSS show higher levels of expression variability **(Fig. 1.7)** [32].

In addition to SNVs, copy number variations (CNVs) (usually defined as copy number variability of regions $\geq$ 1kb in comparison to a reference genome) in parts of the genome influence gene expression and contribute to, for example, schizophrenia and autism [113]. Combined analysis of DNA and RNA has shown that genes with low copy number tend to be more noisily expressed compared to genes encoded by multiple copies **(Fig. 1.7)** [114]. In the context of monoallelic expression, genes located on the X chromosome show increased mRNA half-life which in turn increases transcript stability and reduces noise to levels of autosomal genes [32].

In sum, these findings highlight that multiple correlated genomic features are associated with modulating noise. It is therefore challenging to disentangle the individual underlying sources of transcriptional variability.

Epigenetic factors

Epigenetic research is defined as "the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence" [115]. Epigenetic factors are generally described as DNA methylation at 5'-cytosine–phosphate–guanine-3' (CpG) dinucleotides, histone modifications and nucleosome positioning [116]. **Table 1.2** summarises the relationship between epigenetic features and variable gene expression.

CpG islands (CGI) are genomic sites of more than 200 bases with a GC content of more than 50% and are usually unmethylated. Methylation of CGIs in promoters is linked to gene silencing while DNA methylation in gene bodies facilitates transcription [116]. Recently, the presence of CGIs in gene bodies but also at the TSS and in promoter regions was linked to a reduction in transcriptional variability [32]. Morgan and Marioni, 2018 further distinguished between gene promoters associated with short and long CGIs. Similar to the presence of TATA-box motifs as described above, the length of CGIs in promoter regions controls how variably a gene is expressed. Genes associated with short CGIs tend to be more variably expressed and allow an early response to stimulation, exemplified by observations in mouse bone-marrow derived dendritic cells and human breast cancer cells [33]. However, it is not clear whether the length of CGIs is the sole driver for variable gene expression or how multiple genomic features work together to induce transcriptional variability.

Modifications of histones induce the opening or repression of chromatin and therefore indirectly modulate gene expression [117]. In an extensive study to link histone modifications

to transcriptional variability, Faure *et al.*, 2017 detected several histone modifications in promoter/core promoter motifs, at the TSS and in gene bodies that increase or decrease variability. The repressive tri-methylation of lysine 27 of histone H3 (H3K27me3) mark is linked to higher variability when present at the TSS, in promoters and in gene bodies. The enhancer related mono-methylation of lysine 4 of histone H3 (H3K4me1) mark only increases variability when present at the TSS and in the core promoter sequence while the repressive tri-methylation of lysine 9 of histone H3 (H3K9me3) mark increases variability when present in the promoter motif. The activating marks tri-methylation of lysine 4 of histone H3 (H3K4me3), acetylation of lysine 9 of histone H3 (H3K9ac) and tri-methylation of lysine 36 of histone H3 (H3K36me3) are linked to low levels of variability when present in gene bodies. In addition to these single features, bivalent promoters that carry the repressive H3K27me3 and enhancing H3K4me3 marks show high levels of transcriptional variability [32]. Here and in Morgan and Marioni, 2018, the authors profiled molecular phenotypic variability in "homogeneous" populations of mESCs as proxy for transcriptional noise. While the effect of fluctuation in cell-cycle stages was regressed out, unobserved variation in, for example, the differentiation potential of mESCs in serum grown medium [12] could still exist. It is therefore difficult to use scRNA-Seq data to study the true underlying effect of transcriptional noise on the overall observable phenotypic variability.

One suggestion why bivalent promoters show high transcriptional variability was brought forward by Kar *et al.*, 2017. Here, the authors studied the function of polycomb repressive complexes (PRCs) in mESCs. PRCs are epigenetic modifiers of histones that repress transcription of developmental genes [118] and they can bind together with active RNA polymerase II (RNAPII) to bivalent promoters. Switching between the repressed and active states introduces gene expression variability across a population of cells [119]. However, bulk measures were used to identify the bivalency of promoters. That leaves the possibility that in a fraction of cells the promoter resides in an open state while in other cells the promoter is repressed. This highlights the fact that bulk measures might not be suitable to obtain a correct measure of promoter states in cell populations that could contain unobserved cell state heterogeneity.

Chromatin is the packaged state of DNA within the nucleus and its central elements are nucleosomes. Nucleosomes are combinations of eight of the four histones (H3, H4, H2A, H2B) around which 147 bases of DNA twist. An array of histone modifying enzymes exist that regulate the opening or closing of the chromatin; termed heterochromatin and euchromatin, respectively [120]. Tirosh *et al.*, 2008 showed that promoters with high nucleosome

occupancy close to the TSS tend to display a high range of expression levels across varying conditions (transcriptional plasticity). Distant nucleosome-rich regions are on the other hand associated with low transcriptional variability [121]. Nucleosome covered promoters display shorter transcriptional rates, which in turn explains increased transcriptional variability for these promoters [114]. Single-cell measures indicate cell-to-cell variations in nucleosome positioning around the *PHO5* promoter upon stress induction. Even in the non-stressed state, a small fraction of cells exhibit nucleosome free regions at the promoter which explains low and possibly noisy expression of *PHO5* [122]. This observation again highlights the lack of resolution when using bulk measures to profile the promoter architectures in cell populations. However, current single-cell technologies to profile epigenetic marks lack throughput and are influenced by high levels of technical noise. The observed variations in nucleosome occupancy could therefore be driven by technical variation.

Boundaries between heterochromatin and euchromatin are controlled by boundary elements, such as the transcription factor CCCTC-binding factor (CTCF), that recruit chromatin modifying factors [120]. CTCF also regulates transcription by activating or repressing promoters and regulates distant chromatin interactions [123]. Recent studies suggest that long-range enhancer-promoter interactions modulate transcriptional noise. Interference of CTCF-mediated enhancer-promoter contact either by CTCF knock-out or CTCF-binding site deletion leads to increased expression variability in selected genes [124]. This study however only profiled protein abundance of few genes and did not correct for changes in mean expression that are highly correlated with changes in variance [10]. Enhancers are cis-regulatory elements of non-coding DNA containing TFBSs that regulate the expression of neighbouring genes [125]. Genes within super-enhancer loci, a region with multiple enhancers, control pluripotency master regulators and show high levels of variability in expression down-stream targets of these master regulators show similar co-variation across mESCs [32].

Table 1.2: Epigenetic control of transcriptional noise

	Feature	Variable	Stable
DNA methylation	CGIs		✓
	Short CGIs	✓	
	Gene body methylation		✓
Histone modification	H3K27me3 (TSS, promoter, gene body)	✓	
	H3K4me1 (TSS, promoter)	✓	
	H3K9me3 (promoter)	✓	
	H3K4me3 (gene bodies)		✓
	H3K9ac (gene bodies)		✓
	H3K36me3 (gene bodies)		✓
	H3K27me3 and H3K4me3	✓	
Nucleosome position	Nucleosome rich promoters	✓	
	Distant nucleosome rich regions		✓
	Deletion of nucleosome remodelling complexes	✓	
Genome architecture	CTCF knock-out	✓	
	CTCF binding site depletion	✓	
	Clustered genes		✓
	Nuclear-lamina associated genes	✓	

Moreover, the positioning of genes on the genome controls expression noise with densely clustered genes being less variably expressed in comparison to non-clustered genes [126]. Additionally, genes positioned next to “noisy” genes display higher levels of transcriptional variability compared to genes that are located in proximity to “stable” genes [119]. Expression variability is also increased for genes that are located in a repressed neighbourhood, namely active genes in constitutive nuclear lamina-associated domains [32]. This finding again highlights that genes associated with repressed chromatin display higher transcriptional variability compared to genes associated with open chromatin. Single-cell measures are needed to provide an insight into whether this effect is driven by so called "leaky" expression from closed promoters of if heterogeneous promoter states can be observed.

Transcriptional features

Transcription is initiated by TFs binding to specific regulatory DNA sequences followed by recruitment of RNAPII and RNA synthesis. As discussed above, promoter architecture, namely the location and accessibility of TFBS and RNAPII binding sites, dictates mean expression and transcriptional variability.

In bacteria, the intracellular physical distance between TF source and the promoter sequence influences expression variability. TF expression proximal to their target genes results in less variable expression compared to TF expression which occurs distant to the promoter sequence [127]. Once TFs bind to their target sequence, Carey *et al.*, 2013 showed that the mean expression to expression variability ratio is promoter dependent while in the majority of cases, variability negatively scales with mean expression [128].

Similar to TF binding dynamics, the assembly of RNAPII complexes modulates transcriptional noise. An early study identified the connection between paused RNAPII and synchronous expression of target genes. Genes without pre-loaded RNAPII show more stochastic activation patterns [129]. This finding has later been confirmed using scRNA-Seq data where increased variability was detected for genes with actively transcribing RNAPII across the full range of expression levels **(Fig. 1.8)** [130]. However, genes with pre-loaded RNAPII also have a higher CpG content and are depleted for TATA-box elements [130]. Once again, the correlation between genomic factors and their individual associations with variation creates a challenge for disentangling their specific effects on molecular phenotypic variation. Alternatively, this may also represent multiple regulatory layers that combine to modulate noise.

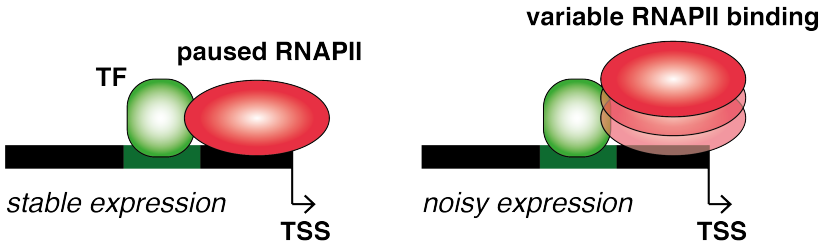


Fig. 1.8: RNAPII pausing reduces transcriptional noise.
Left: Pre-loaded RNA polymerase II (RNAPII) allows direct transcription upon transcription factor (TF) binding. Right: RNAPII recruitment induces gene expression variability.

Post-transcriptional and translational features

After synthesis, pre-RNAs are polyadenylated and spliced to form mRNA that relocates from the nucleus to the cytoplasm where translation occurs to synthesise proteins [131]. On the post-transcriptional and translational level, mRNA location, structure, degradation and translation have been shown to influence cell-to-cell variation in protein abundance.

Upon transcriptional activation, RNAs are produced in burst-like patterns where burst frequency modulates mean expression and noise, and burst size influences solely mean expression [109]. While bursty transcript synthesis introduces stochastic fluctuations in nuclei between cells, active export of mRNAs into the cytoplasm can dampen this source of variability **(Fig. 1.9)** [69]. Reduced cytoplasmic noise has also been shown for two nuclear-retained genes in the mammalian liver. Furthermore, this mode of noise control was proposed to be active across a range of metabolic tissues [132].

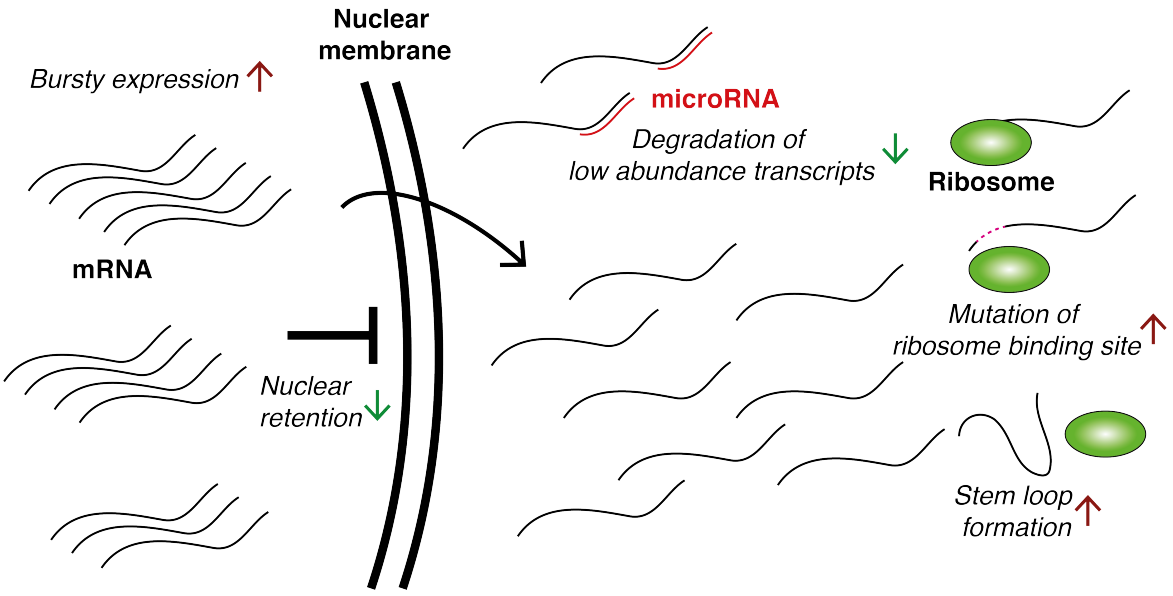


Fig. 1.9: Post-transcriptional regulation to control noisy expression.
Bursty expression introduces nuclear variation in transcript abundance that is buffered due to retention at the nuclear membrane. Within the cytoplasm, micro RNAs degrade lowly expressed genes to reduce expression noise. Deletion of the ribsosome binding site as well as stem loop formation increase variability in protein abundance across cells. Arrows indicate either increased (red) or decreased noise (green) depending on the regulatory mechanism.

Conversely, a recent study by Hansen *et al.*, 2018 proposed an amplification of transcriptional variability due to nuclear export [133]. However, the authors computed the Fano factor (variance divided by mean expression) as measure of variability and assumed that its value is not correlated with mean expression. This assumption arises from an underlying

Poissonian (Fano factor is equal to 1) or over-dispersed Poissonian (Fanor factor is larger than 1) distribution of transcript counts. In reality, the Fano factor is not constant across the range of mean expression. This has been discussed by Grün *et al.*, 2014, who showed scale differences in the coefficient of variation (CV) versus mean expression dependency for lowly and highly expressed genes [134]. Hansen *et al.* also showed this effect when plotting the Fano factor versus mean expression. Across all genes, the cytoplasmic transcript abundance as well as the Fano factor were larger compared to nuclear measures, which is to be expected by the model proposed by Grün *et al.*, 2014. It is therefore recommended to fit a non-parametric curve to variability measure versus mean expression and compare the regression residuals as explained in Chapter 3.

Within the cytoplasm, mRNAs are subject to translation or degradation. At this stage, variability induced by bursty gene expression is propagated to form variation in protein abundance. The availability of mRNAs for translation is not only dictated by their synthesis but also their degradation rate. mRNA degradation is accelerated by recognition of miRNAs. This process has been shown to preferentially reduce noise levels for lowly expressed genes in mESCs, possibly to retain cellular identity **(Fig. 1.9)** [135].

In addition to noise introduced by stochastic processes on the transcriptional level, the recognition and binding of ribosomes to mRNAs for translation initiation is a source for variation in protein abundance. Modulating translational efficiency by mutating the ribosome binding site and initiation codon showed an association between translation and variation in protein abundance **(Fig. 1.9)** [136]. Additionally, mRNA secondary structure formed by stem loops and poly(G) motifs affects translation initiation and increases variation in protein levels **(Fig. 1.9)** [137].

Together, these studies again highlight a multitude of factors that can modulate the observable variation in transcript and protein abundance. Moving forward, models that correct or account for variations introduced by different factors need to be developed to disentangle the individual sources of molecular variation.

1.2.2 Extrinsic noise

Classically, extrinsic noise has been described to arise from fluctuations in molecules that affect the global gene expression landscape of the cell [1]. Measuring extrinsic noise is only possible in bacterial populations that do not show fluctuations in cell states. Nowadays, extrinsic noise is often described to arises from cells being in different regulatory states. Here, differences in cellular components introduce variation in mRNA and protein abundance. The presence of cell states in otherwise homogeneous populations is characterised by differences in metabolism, cell cycle, cellular volume, cell-to-cell and environmental signalling as well as cell density. It has been shown that extrinsic noise forms a major contribution to variation in gene expression and that transcript distributions can be predicted from the cellular state, population context and microenvironment [69]. Being able to predict transcript abundance however indicates that extrinsic noise is not purely stochastic but might also contain deterministic components.

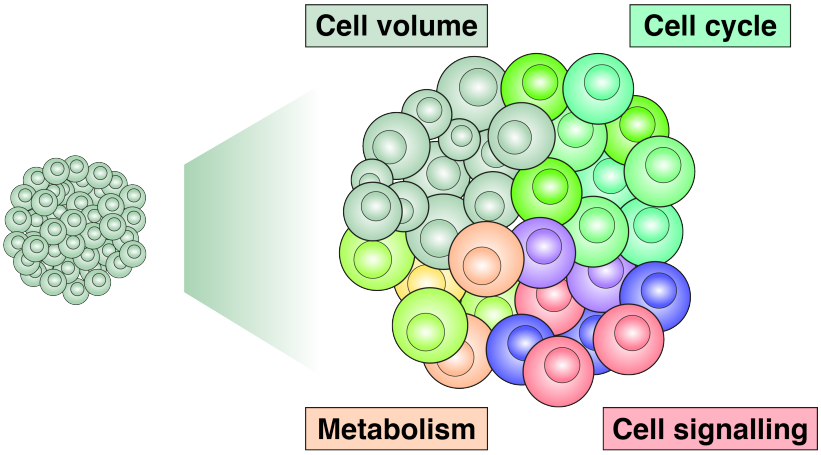


Fig. 1.10: Differences in cell states induce extrinsic noise.
Within a homogeneous population of cells (left), individual cells reside in different cellular states (e.g. cell cycle, cell signalling, metabolism) and show differences in cellular volume.

Cell cycle

Cell cycle has been widely discussed to form a major source of extrinsic noise [138, 139]. In yeast populations, differences in transcriptional activities between the G1 and S/G2/M phases of the cell cycle lead to large-scale transcriptional heterogeneity across cell populations **(Fig. 1.10)** [4]. Under nutrient-poor conditions, growth rate is reduced and transcriptional variability is elevated due to cells being in different cell cycle stages [140]. Even under optimal growth conditions for mESCs (2i media), cell cycle related genes show strong

heterogeneity in expression across the cell population [12]. However, it is possible that unobserved regulatory mechanisms are in place to induce proliferation, which otherwise appears to occur randomly across a population of cells. When quantifying cell-to-cell variation, cell cycle induced extrinsic noise is often seen as unwanted variation and can mask more subtle transcriptional heterogeneity. Computational methods have been developed to correct for this confounding effect to enhance the underlying signal [13, 141].

Cell volume

Cellular volume provides another explanation for global differences in mRNA content between individual cells introducing large-scale transcriptional variability **(Fig. 1.10)**. Even though cell volume changes during cell cycle progression, within each phase, cell volume can vary as much as across all phases. It has been shown that mRNA counts scale with cellular volume to maintain transcript concentrations within each cell [142–144]. Again, it is not fully understood how cell volume is controlled across a population of cells; especially within multicellular organisms. Therefore, the volume of a cell does not necessarily stochastically but rather deterministically contribute to the observed molecular phenotypic variability. To avoid this source of heterogeneity, normalisation approaches correct for differences in mRNA content between individual cells [145].

Metabolism

The effect of metabolic fluctuations has been studied in *E. coli* populations. Variations in biochemical reactions are induced by noise in the expression of their corresponding catalytic enzymes. Changes in metabolism are then coupled to varying growth rates of individual cells, which in turn introduce large-scale transcriptional heterogeneity in cell populations **(Fig. 1.10)** [6].

Expression capacity

Expression capacity is defined as the ability of a cell to express proteins from a gene by utilising the transcriptional and translational machinery. Fluctuations in the expression capacity of cells due to quantitative differences in RNAPII or ribosomes can induce global variability among the majority of proteins [138].

Cell signalling

A different source of extrinsic noise is the intra- or inter-cellular signalling state of individual cells. Fluctuations in membrane bound or cytoplasmic proteins lead to inconsistent transmission of signalling stimuli as exemplified by variability in TRAIL-induced apoptosis [97]. Similarly, variations of regulators in the extracellular signal–regulated kinase (ERK) signalling pathway introduce downstream variability in nuclear response **(Fig. 1.10)**. The degree to which nuclear ERK response varies depends on the position of the regulator in the topology of the signalling pathway [5]. In *C. elegans*, perturbation of the Wnt signalling pathway displayed different degrees of variability in expression of the key Hox gene for Q neuroblast migration, *mab-5*. It has been proposed that extrinsic noise, in this case the strength of the Wnt signal, modulates intrinsic variation in the expression of *mab-5* [71]. These examples describe a system where noise introduces variation in individual components. However, this variation is modulated by signalling networks, which allows the cells to precisely respond to external cues.

Physical constraints

Physical constraints on cell growth and the direct population context influence the state of individual cells [69]. Snijder *et al.*, 2009 performed detailed imaging based analysis of adherent human cells that were infected with different viruses. Clathrin mediated endocytosis was most variable with low cell density leading to inefficient mouse hepatits virus infection. Dengue virus preferentially infects edge cells while simian virus 40 infection decreased with large cell density [146]. These experiments indicate the importance of local cellular microenvironment and cell-cell contacts leading to heterogeneity in cell states **(Fig. 1.11)**.

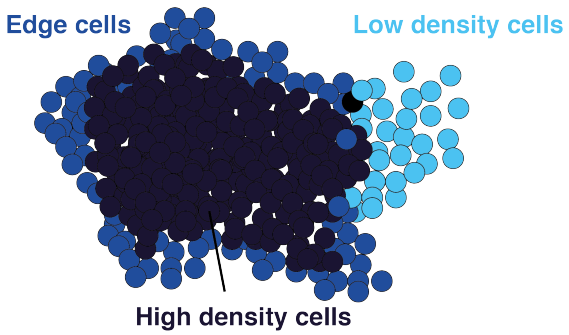


Fig. 1.11: Physical constraints induce heterogeneous expression patterns.
Cell density increases during the expansion of a homogeneous population of cell forming patches with high and low density, pushing cells to the edge of the population. Based on these physical constraints, cells change their transcriptional programme, inducing variability across the population.

1.3 Quantification of molecular variability

In the last ten years, the scale of single-cell assays increased from measuring few to hundreds and thousands of genomic, epigenetic, transcriptomic or proteomic features. These technologies can be used to measure molecular phenotypic variability, as well as gain an understanding of the regulatory features that modulate it. The ability to study noise using technologies that destroys the cell is formulated on the basis that a cross-section over a population of cells is representative of the time-resolve noise profile of any given cell [136]. In this section, I will discuss the applicability of single-cell sequencing and imaging technologies as a potential read-out for cell population heterogeneity induced by transcriptional noise.

1.3.1 Single-cell sequencing

Next generation sequencing approaches have been applied to individual cells to quantify variation in DNA sequence, mRNA expression, epigenetic marks and protein abundance within a cell population.

Single-cell whole genome sequencing

Single-cell whole genome sequencing (scDNA-Seq) has previously been used to identify CNVs and SNVs between single cells [147]. Based on these read-outs, tumour heterogeneity and evolution [148] as well as lineage relationships in the human brain were inferred [149]. To obtain enough genomic material, whole genome amplification (WGA) is performed on DNA from individual cells. The single-cell comparative genomic hybridization protocol (SCOMP) degrades DNA via restriction enzymes, includes a primary polymerase chain reaction (PCR) amplification step and a later re-amplification via comparative genomic hybridisation [150]. Multiple displacement amplification (MDA) is based on the random initiation of amplification via oligonucleotide primers with strand displacement [151]. Compared to MDA, multiple annealing and looping-based amplification cycles (MALBAC) achieves an initial quasi-linear amplification step by pre-amplification using primers with handle sequences. Full amplicons form hairpins that are exponentially amplified prior to sequencing [147].

The main limitation of single-cell DNA sequencing is the genomic coverage per cell. While the detection of SNVs requires deep sequencing of individual cells, CNVs can be detected with shallow sequencing therefore allowing the throughput to increase **(Fig. 1.12)** [152, 153]. Recently, Vitak *et al.*, 2017 introduced single-cell combinatorial indexed sequencing (sci-

Seq) which allows the generation of thousands of single-cell genomes for sequencing. In the first step, multiple nuclei are sorted into each well of a 96 well plate and the genomic DNA is labelled with barcodes by transposase tagging. In the second step, 15-25 tagged cells are sorted into individual wells of a PCR plate where the second round of barcoding is performed during amplification. In that way, CNVs of over 15,000 cells can be assessed [154]. While previously only bulk measures have been used to link mutations to changes in transcriptional variations [106], scDNA-Seq with high read-depth can be used to solve the question of whether the same mutation in each cell (germ line mutation) or a heterogeneous mutational pattern (somatic mutation) causes the observed increase/decrease in phenotypic variability.

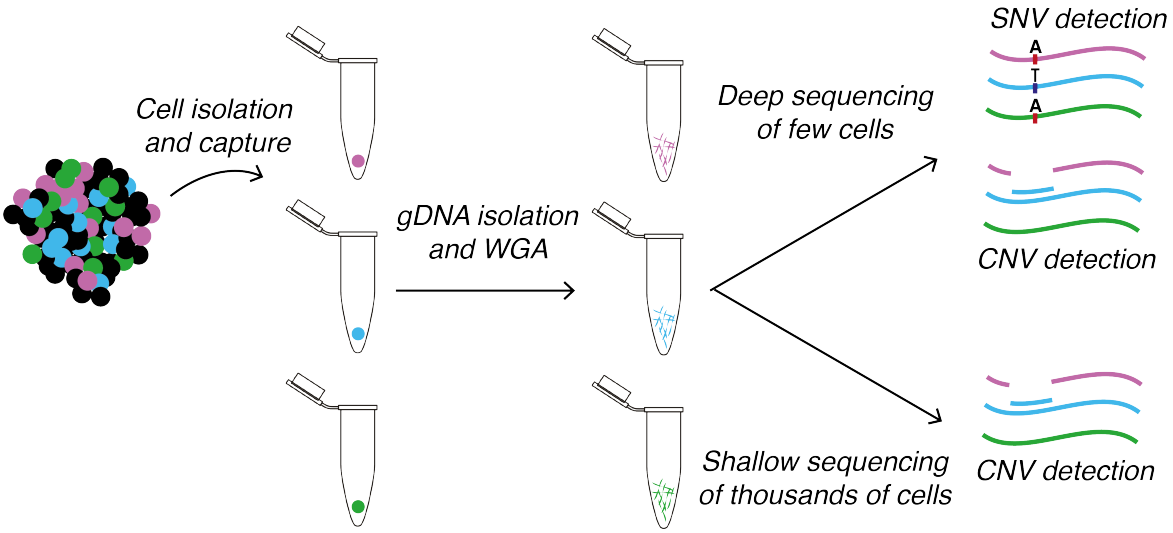


Fig. 1.12: ScDNA-Seq allows detection of SNVs and CNVs between individual cells.
Individual nuclei are captured in 96-well plates and directly lysed or fixated for multiplexing. Whole genome amplification (WGA) can be performed using MDA, MALBAC or SCOMP resulting in amplified genome segments. Depending on the biological question, whole genomes are either sequenced thoroughly to detect single nucleotide variants (SNVs) while shallow sequencing can be used to detect copy number variations (CNVs).

Single-cell RNA sequencing

Initial approaches to quantify mRNA abundance within single cells included targeted microfluidic-based single-cell real time PCR (RT-PCR) [155] and whole-transcriptome read-outs of hand-picked cells [156]. Methods for cell capture range from micromanipulation [157] and laser capture microdissection [158] as targeted methods with low throughput to fluorescence-activated cell sorting (FACS) [159–161], microfluidics [30, 162] and micro-droplets [163, 164] as high-throughput approaches [165] **(Fig. 1.13)**.

A variety of scRNA-Seq protocols have been published that utilise different methods for mRNA reverse transcription (RT), complementary DNA (cDNA) amplification and library preparation. All of these commonly used techniques for scRNA-Seq select and reverse transcribe mRNA (poly(A) tailing). The initial protocol introduced by Tang *et al*, 2009 [156] was improved by incorporating a template switching mechanism at the 5' end of the mRNA thus reducing the 3' sequencing bias present in previous methods [166] (see below and **Fig. 1.13**). This single-cell tagged reverse transcription (STRT) method shows 5' bias in read mapping and was later modified for full-length transcript detection (SmartSeq [167] and SmartSeq2 [168]). CEL-Seq [169] and CEL-Seq2 [170] use *in vitro* transcription (IVT) to linearly amplify cDNA prior to sequencing as opposed to exponential amplification in other techniques. Protocols for sequencing library preparation have been optimised for Illumina, SOLiD or PacBio sequencing [171].

During scRNA-Seq, minute amounts of mRNA are captured and amplified generating a high degree of technical noise, which distorts quantification of true biological variability. To account for this, a set of external RNAs developed by the External RNA Control Consortium (ERCC) [172] can be added to the cell lysate. Based on the reads mapped to ERCC spike-ins, technical noise can be removed from total expression variability [10, 11]. Another way to reduce noise derived from amplification biases in scRNA-Seq experiments is to tag each mRNA molecule with a unique molecular identifier (UMI) [173, 174].

One example of a commercially available platform that captures individual cells and performs lysis, reverse transcription and pre-amplification of cDNA is the Fluidigm® C1 system. Individual cells are loaded into integrated fluidic circuits (IFCs), also termed "chips", that allows capturing of 96 to 800 cells. Depending on the size of the cells, this system offers chips with different capture well sizes. Each well can be microscopically inspected to differentiate between empty capture sites and single cells [171]. The C1 system uses the SMARTer® chemistry to capture poly(A) mRNA with modified oligo(dT) primer. Next, the reverse transcriptase (RTase) reverse transcribes from the 3' to the 5' end of the mRNA and adds non-templated deoxycytidines to the 3' end of the cDNA. The template-switch primer contains guanosines at its 3' end that base-pair with the deoxycytidines on the cDNA to create an extended template. The RTase extends to the end of the template-switch primer. This produces single-stranded cDNA that contains the SMARTer tag sequence, the 3' end of the mRNA, the full-length transcript up to the 5' end of the mRNA, and the reverse complement of the SMARTer tag sequence. Amplification of this cDNA is performed by PCR on the chip [175]. After pre-amplification, the cDNA is collected and prepared for sequencing.

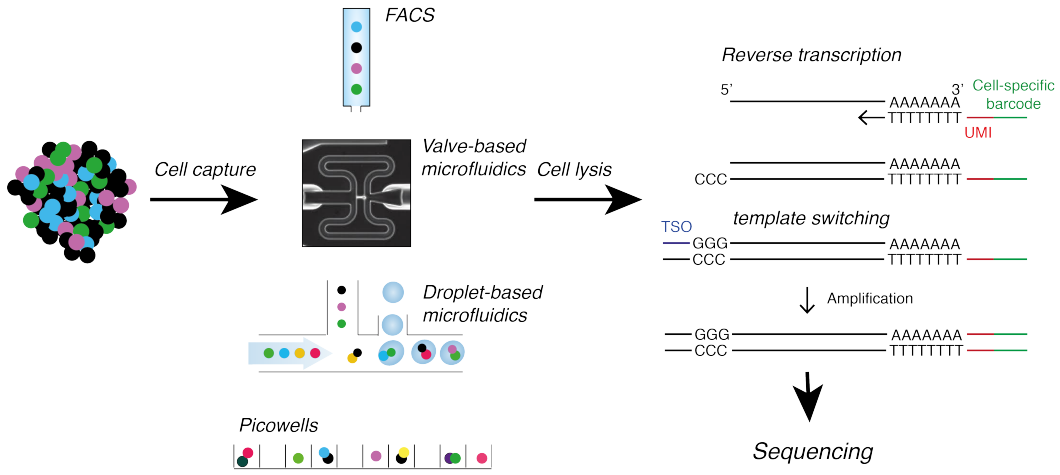


Fig. 1.13: Workflow for scRNA-Seq technologies.
Single cell suspensions are obtained by tissue dissection and dissociation. Commonly used cell capture technologies include fluorescence-activated cell sorting (FACS), valve-based microfluidics (Fluidigm® C1 system), droplet-based microfluidics (10X Genomics® system), or picowells. After cell capture and lysis, poly(dT) oligos capture mRNA prior to reverse transcription. In the case of droplet-based cell capture, poly(dT) oligos are tagged with a unique molecular identifier (UMI) and a cell-specific barcode. RT generates cDNA from the template RNA. One strategy for RT is the template-switching protocol where the reverse transcriptase adds three cytidines at the 5' end of the template. A template-switch oligo (TSO) binds to the cytidines and allows amplification from the 5' end. After cDNA amplification, libraries are prepared for sequencing. For this, transposase degrades full length transcripts and Illumina sequencing primers are added (C1 system). In the case of the 10X Genomics system, the first read has been added next to the cell-specific barcode while the second read is added after cDNA fragmentation. This protocol shows a 3' bias.

In parallel to extending scRNA-Seq protocols to robustly capture mRNA transcripts, efforts have been made to increase the throughput of this technology. Jaitin *et al.*, 2014 introduced massively parallel RNA single-cell sequencing (MARS-Seq) to sequence over 4000 cells of the mouse spleen. MARS-Seq captures cells in 384 well plates and labels transcripts of each cell with a combination of 2 random barcodes. This multiplexing strategy is performed using a liquid handling robot and cells can be pooled for sequencing library preparation which reduces costs and time effort [161]. The first large-scale technique that captured tens of thousands of cells was introduced by Fan *et al.* in 2015. Here, 10,000 cells were captured in a 100,000 microwell surface. Additionally, barcoded beads were loaded into the surface until saturation. This Cyto-Seq approach is similar to the more recent Seq-Well technology [176]. Each bead is coated with barcodes containing a unique sequence, a bead-specific barcode, a UMI and a oligo(dT) primer. After cell lysis and mRNA capture, beads are pooled and cDNA synthesis can be performed prior to sequencing [177]. In the same year, the inDrop and Drop-Seq technologies were introduced [163, 164]. Both technologies use microfluidic

platforms to merge droplets containing barcodes, lysis and reverse transcription reagents with droplets containing cells. Similar to the above described Cyto-Seq [177], after lysis and mRNA capture, droplets are pooled for sequencing. The main difference is that cell-specific barcodes in Drop-Seq are bound to beads and to a polyacrylamide mesh in inDrop.

10X Genomics™ has introduced a platform that uses these concepts to generate hundreds of thousands of gel beads in emulsions (GEM). Around 80% of generated oil droplets capture barcoded gel beads in 8 channels in parallel. Each barcode consists of a sequencing adapter and primer, a 14bp sequence from a pool of 750,000 barcodes, a 10bp UMI and a 30bp poly(dT) oligotide to capture poly(A) mRNA [178]. GEMs are fused with individual cells at a low concentration and cell lysis begins instantaneously. mRNA molecules are captured by the poly(dT) barcode and enzymes needed for RT are released from the gel beads. After RT, each cDNA contains a transcript-specific UMI and a GEM-specific barcode making demultiplexing possible **(Fig. 1.13)**. Barcoded cDNA is pooled for PCR amplification and library preparation [178].

Methods that even further increased the throughput of scRNA-Seq include split-pool ligation-based transcriptome sequencing (SPLiT-Seq) and sci-RNA-Seq. Similar to sci-DNA-Seq (see above) these technologies are based on combinatorial indexing of mRNA in fixed cells or nuclei. Sci-RNA-Seq tags transcripts during two rounds of indexing with UMIs and a combination of two cell specific barcodes [179]. SPLiT-Seq on the other hand performs transcript tagging during 4 cycles adding 4 barcodes [28]. In that way, around 1 million cells can be uniquely labelled. At this stage the limiting factor is the sequencing depth needed to obtain high-resolution whole transcriptomes of each cell. These approaches as well as the recently developed massively parallel single-nuclei sequencing with droplet technology (DroNc-Seq) also allow sequencing mRNA from nuclei which is the preferred method for clinical samples, archived materials, and tissues that cannot be readily dissociated [180].

Single-cell epigenomics

Single-cell epigenomic methods capture the chromatin state, histone modifications and DNA methylation state of individual cells and allow quantification of epigenetic variability across a population of cells **(Fig. 1.14)** [181]. To observe methylation states of CpG motifs, single cell bisulfite sequencing (scBS-Seq) involves the extraction of genomic DNA and cytosine to uracil bisulfite conversion prior to library preparation. 5-methylcytosine remains intact during conversion [182, 183]. The throughput of this approach was scaled up by combinatorial indexing of fixed nuclei similar to sci-DNA-Seq (i) prior to bisulfite conversion

and (ii) during PCR amplification (sci-MET-Seq) [184]. Single-cell reduced representation bisulfite sequencing (scRRBS-Seq) enzymatically digests genomic DNA prior to bisulfite conversion. CpG-rich fragments can be enriched and amplified via ligated adapters before high-throughput sequencing [185]. Extending the read-out of scBS-Seq and scRRBS-Seq, single-cell 5-hydroxymethylcytosine sequencing (sc5hmC-Seq) captures the first oxidative product of CpG sites towards de-methylation and therefore cellular variation of methylation dynamics. Instead of bisulfite conversion, 5hmC sites are glucosylated before enzymatic digestion and adapter ligation [186].

To measure histone modifications or transcription factor binding dynamics at the single-cell level, digested chromatin from individual cells is tagged with barcodes prior to immunoprecipitation (IP) during single-cell chromatin IP followed by sequencing (scChIP-Seq) **(Fig. 1.14)**. With this droplet-based method, variable chromatin signatures were detected across a population of ESCs based on the di-methylation of lysine 4 of histone H3 (H3K4me2) [187].
Other epigenomic approaches focus on estimating the patterns of open chromatin by measuring chromatin accessibility **(Fig. 1.14)**. Single-cell assay of transposase-accessible chromatin using sequencing (scATAC-Seq) captures individual cells on IFCs before inserting sequencing adapters into accessible regions via the prokaryotic Tn5 transposase and pre-amplifiation. After library collection, cell-specific barcodes are added via a second round of PCR prior to sequencing [188]. Capturing cells in IFCs before barcoding limits the throughput to around tens or hundreds of cells at one time. Combinatorial indexing by tagging cells with barcodes in a two step process increases throughput for scATAC-Seq to thousands of cells [189]. An alternative approach to measure open chromatin involves the digestion of DNA with DNase I (Pico-Seq). The resulting small fragments undergo end-repair, adaptor ligation and PCR amplification in the presence of circular carrier DNA to avoid the loss of the minute amount of fragments [190]. Similarly, nucleosome positioning can be detected by using the GpC-specific DNA methyltransferase (MTase) M.CviPI to methylate cytosines of GpC motifs in regions where DNA is accessible. Individual cells are isolated and their DNA digested prior to bisulfite conversion. Patterns of methylated and unmethylated GpCs indicate the positioning of nucleosomes along the DNA [122].

Single-cell technologies to study large-scale chromosome structure include DNA adenine methyltransferase identification (DamID) [191], a method to identify lamina-associated domains, and single-cell high-troughput chromosome confirmation capture (HiC) **(Fig. 1.14)** [192]. Similar to sci-DNA-Seq, sci-RNA-Seq, sci-ATAC-Seq and sci-MET-Seq, sci-Hi-C

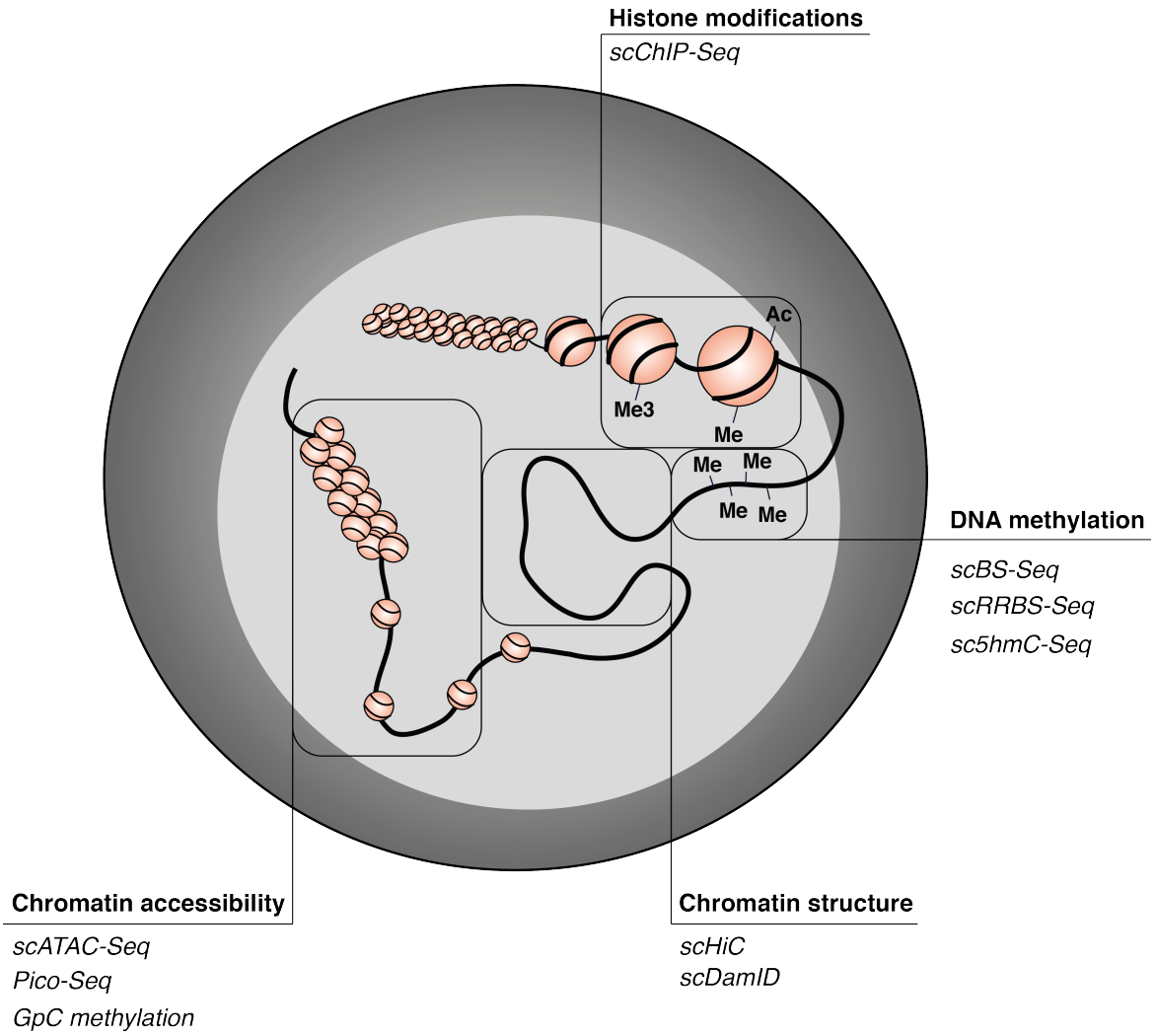


Fig. 1.14: Single-cell epigenomics to study chromatin structure and modifications.
Single-cell epigenomic technologies are used to study variation in DNA methylation, histone modifications, chromatin structure and nucleosome positioning across individual cells.

uses multiplexing of fixated nuclei after digestion to insert (i) a biotinylated bridge adapter and later on a second adapter after lysis [193]. This technology allows the demultiplexing of thousands of cells after bulk-HiC-like processing.

Multi-omics approaches

In recent years, some of the above described techniques were combined to measure transcriptomic, genomic, epigenomic and proteomic ("multi-omic") features of single cells in parallel [9]. The first approach for combinatorial DNA and mRNA sequencing (DR-Seq) from the same cell amplifies genomic DNA and cDNA derived from reverse transcribed mRNA in one reaction step to avoid losses. After initial amplification, the sample is split to further process genomic DNA (gDNA) and cDNA separately. PCR amplification increases the amount of gDNA while IVT amplifies cDNA prior to sequencing [114]. An alternative approach, genome and transcriptome sequencing (G&T-Seq), firstly separates gDNA and mRNA before whole-transcriptome and whole-genome amplification. Biotinylated oligo(dT) primers capture mRNA and are coupled to streptavidin coated beads. Once mRNA and gDNA is separated, the SmartSeq2 protocol is used to perform whole-transcriptome amplification while MDA or PicoPlex approaches can be used to amplify gDNA prior to sequencing **(Fig. 1.15)** [194].

Similarly, single-cell methylome and transcriptome sequencing (scM&T-Seq) initially separates genomic DNA from mRNA. The scBS-Seq protocol is applied to isolated gDNA and is used to identify methylated CpG positions while mRNA was amplified via the SmartSeq2 protocol [195]. The scM&T-Seq method has been extended to detect accessible chromatin regions in parallel to capturing methylated CpG sites and whole-transcriptome information. Prior to bisulfite conversion of gDNA, GpC sites are methylated by MTase in nucleosome sparse regions **(Fig. 1.15)** [196, 197].

Attempts have been made to capture $\sim$96 mRNAs in combination with proteins within individual cells. After cell lysis, samples are split to process mRNA and protein separately. mRNA is reverse transcribed and pre-amplified prior to quantitative PCR (qPCR) while oligonucleotide tagged antibodies bind to proteins. The free 3'-ends are complementary and can be extended by polymerisation to create a DNA reporter molecule. Similar to mRNAs, these molecules are detected using qPCR [198]. This method has been scaled up by integration of droplet digital PCR [199]. Alternatively, proximity ligation assay for RNA allows isotope tagging of RNA molecules, which are detected in parallel to proteins via mass cytometry **(Fig. 1.15)** [200].

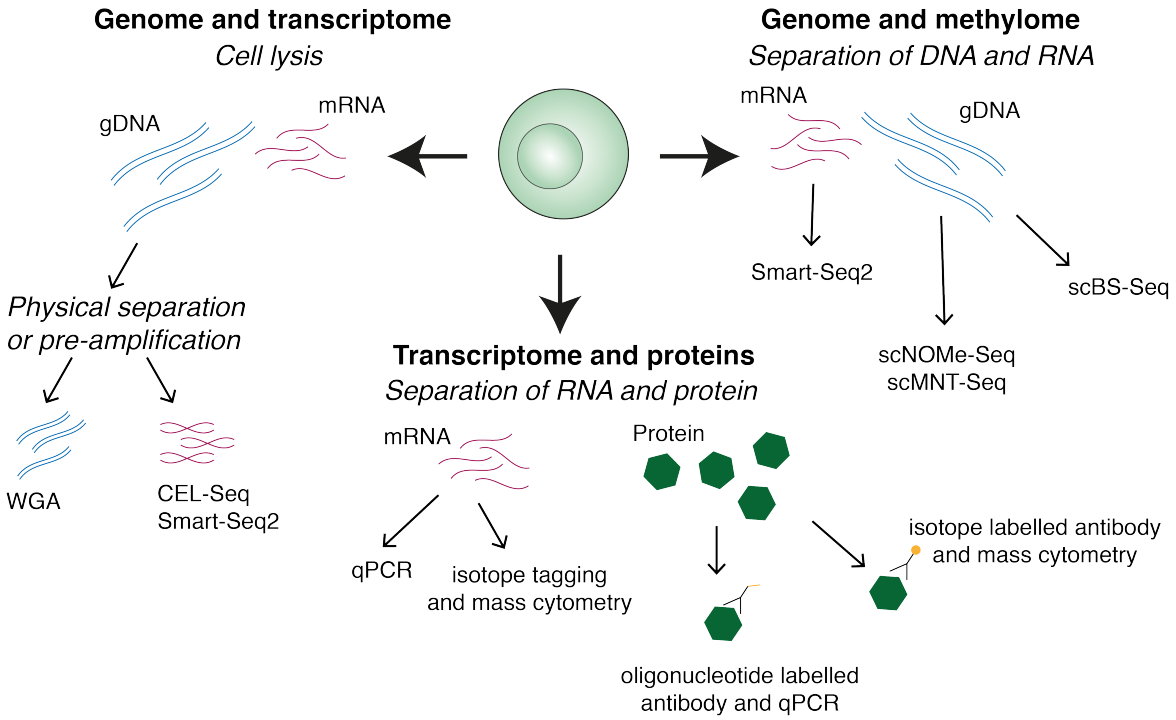


Fig. 1.15: Single-cell multi-omic approaches.
Single-cell DNA and RNA-Seq either directly separates RNA and DNA or pre-amplifies both prior to separation. Measuring RNA and protein abundance from individual cells is done after physical separation followed by either oligonucleotide tagging of proteins or isotope tagging of RNA molecules. For methylome and transcriptome sequencing, DNA and RNA are separated prior to RNA sequencing and bisulfite conversion.

1.3.2 Imaging approaches

Similar to single-cell sequencing, RNA or protein imaging approaches quantify noise in biological systems [201]. Initial studies that addressed the extent of biological noise in bacterial populations used the expression of fluorescent proteins controlled by promoters of interest (reporter assays) to quantify expression noise [1, 202]. Later on, single molecule fluorescence *in situ* hybridization (smFISH) was developed to capture variation in mRNA abundance across multiple cells [65, 203, 3] and in whole organs [204]. Furthermore, the combination of fluorescently labelled proteins and smFISH allows the detection of co-variation between protein and mRNA levels within individual cells [205]. High-throughput automated smFISH of target RNAs in thousands of wells [88] identified nuclear retention of RNAs as a mechanism to reduce cytoplasmic transcript variability [69]. Moreover, computerised image analysis and supervised machine learning extracts hundreds of cellular features from microscopy images and can therefore dissect variation of biological processes such as virus infection [146].

The development of super-resolution microscopy allows detection of fluorophores that are spaced less than 100nm apart [206]. By combining stochastic optical reconstruction microscopy (STORM) and combinatorial labelling of RNA inside the cell, multiple transcripts from different genes can be visualised [207]. This approach has been advanced to measure hundreds to thousands of RNA species per cell. Multiplexed error-robust fluorescence *in situ* hybridization (MERFISH) hybridises encoding probes to target RNAs prior to N rounds of combinatorial labelling using fluorescently labelled read-out probes **(Fig. 1.16)**. MERFISH uses an encoding scheme that corrects for individual read-out errors based on a certain hamming distance between possible N-bit codes. Therefore, with 16 rounds of combinatorial labelling and a hamming distance of 4, 140 RNA species can be detected [7]. A similar approach has been developed to profile spatial expression patterns in the mouse hippocampus (seqFISH) [208]. By replacing the photobleaching step between consecutive rounds of combinatorial labelling with chemical cleavage and using multi-color imaging, the throughput of MERFISH can be increased [209]. Background fluorescence in tissue sections can be reduced by matrix-embedding of labelled RNA and cellular digestion [210].

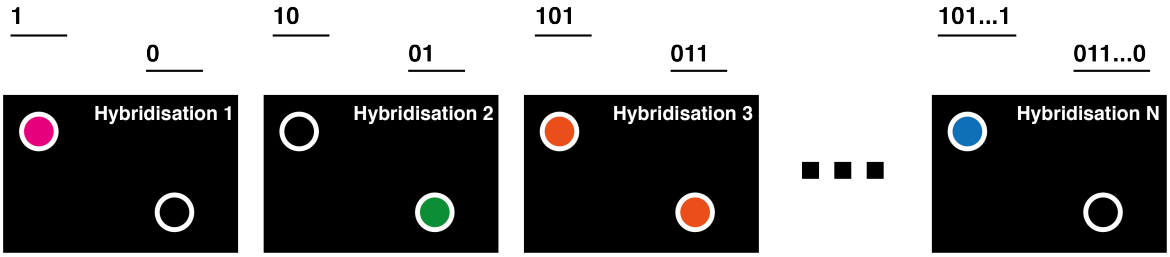


Fig. 1.16: MERFISH-type spatial transcriptomics.
Each transcript species is tagged with encoding probes that contain a sequence to recognise the RNA and multiple read-out sequences. During each hybridisation cycle, individual read-out probes hybridise with their specific sequences on the encoding probes. After multiple rounds of hybridisation and imaging, individual RNA transcripts can be decoded.

1.3.3 Computational modelling and quantification

Previous research focused on the derivation of mathematical frameworks to model expression dynamics in biological systems [211]. In the simplest case, the central dogma of molecular biology states that mRNAs are synthesised from DNA at rate k_m and proteins are translated from mRNAs at rate k_p. Furthermore, mRNAs are degraded at rate γ_m and proteins at rate γ_p. In a noise-free system, this dogma leads to the following **deterministic**, first-order differential equation describing the number of mRNAs (m) and proteins (p) over time:

$$\frac{dm}{dt} = k_m - \gamma_m m, \quad \frac{dp}{dt} = k_p m - \gamma_p p \tag{1.1}$$

Steady-state transcript counts in this simple, two-stage system are defined as $\langle m \rangle = \frac{k_m}{\gamma_m}$ and protein abundance as $\langle p \rangle = \frac{k_m k_p}{\gamma_m \gamma_p}$. The variance for transcript and protein distributions are defined as: $\sigma^2 = \langle m \rangle$ and $\sigma_p^2 = \langle p \rangle \left[\frac{k_p}{\gamma_p + \gamma_m} + 1 \right] = \langle p \rangle \left[\frac{b}{1+\eta} + 1 \right]$, where $b = k_p/\gamma_m$ is the average number of proteins produced per transcript and $\eta = \gamma_p/\gamma_m$ [211, 212]. mRNAs usually decay much faster than proteins. Therefore $\gamma_m \gg \gamma_p$ and $\sigma_p^2 \cong \langle p \rangle [b+1]$ [212]. For this system, the mean translational burst size can be described as the Fano factor $\frac{\sigma_p^2}{\langle p \rangle} \cong b+1 \approx b$ and burst frequency is captured by the inverse squared coefficient of variation $\frac{\langle p \rangle^2}{\sigma_p^2} \approx \frac{\langle p \rangle}{b} = \frac{k_m}{\gamma_p} = a$. The latter assumes that mRNAs are directly translated as soon as they are produced [213].

To account for stochasticity in this system, probabilistic expressions of the aforementioned equations have been described. The chemical master equation defines the time evolution of the probability of observing a system containing m mRNAs and p proteins at time point t:

$$\begin{aligned} \frac{\partial P_{m,p}}{\partial t} &= k_m \left[P_{m-1,p} - P_{m,p} \right] + \gamma_m \left[(m+1) P_{m+1,p} - m P_{m,p} \right] \\ &+ k_p m \left[P_{m,p-1} - P_{m,p} \right] + \gamma_p \left[(p+1) P_{m,p+1} - p P_{m,p} \right] \end{aligned} \tag{1.2}$$

The stationary probability distribution for this discrete representation of the master equation has the form of a negative-binomial distribution:

$$P_p = \frac{\Gamma(a+p)}{\Gamma(p+1)\Gamma(a)} \left(\frac{b}{1+b} \right)^p \left(1 - \frac{b}{1+b} \right)^a \tag{1.3}$$

where a represents the burst frequency, b the mean burst size and $\Gamma(n)$ the Gamma function [214, 213, 211]. Friedman *et al.*, 2006 derived a stationary probability distribution from a continuous form of the chemical master equation [213]. This solution takes the form of a Gamma distribution:

$$P_p = \frac{1}{b^a \Gamma(a)} p^{a-1} e^{-p/b} \tag{1.4}$$

This simple system has also been extended to incorporate the ON-OFF switching of promoters [215, 214]. Extensive modelling and quantification of mRNA and protein abundance in prokaryotic and eukaryotic cell populations confirmed this negative binomial (over-dispersed Poissonian) relationship between protein variance and abundance [136, 216]. The over-dispersion in protein abundance arises from biological noise (η_{tot}), which can be decomposed into intrinsic (η_{int}) and extrinsic (η_{ext}) contributions ($\eta_{tot} = \eta_{int} + \eta_{ext}$) [34, 217].

These components can be directly computed when using a two reporter system controlled by identical promoters [1].

Classic mathematical approaches to model transcriptional and translational dynamics use simplified assumptions for analytical tractability. Similar to the described translational bursting, transcriptional bursting as observed in eukaryotic cells [218] leads to an over-dispersion in mRNA transcripts. Furthermore, while most models focus on single promoter dynamics, cases in which multiple promoters and competitor sites dilute TF binding have only recently been addressed [219]. The assumption that translation from mRNA follows a first-order process was extended by using a hyperbolic Michaelis-Menten kinetic to model the translation process. This approach allows for continuous levels of ribosome occupancy on mRNAs [220].

While the models described above theoretically describe the expected distributions of proteins and mRNA across a population of cells, in practice, absolute measures (e.g. transcript counts or fluorescence intensity) have to be used to quantify variation across a population of cells. In an early approach to model promoter kinetics from scRNA-Seq data, Kim and Marioni, 2013 proposed a hierarchical Beta-Poisson model that relies on the switching dynamics of promoters between the "ON" and the "OFF" state (k_{ON}, k_{OFF}) as well as the transcription rate s and the decay rate d [221]. The model was formulated as follows:

$$X|s,p \sim \text{Poisson}(sp)$$
$$p|k_{ON},k_{OFF} \sim \text{Beta}(k_{ON},k_{OFF}),$$

where X is the transcript count per cell and p a random effect dictated by promoter switching. Gene-specific inference was implemented as a Bayesian framework using Gamma distributions as priors for the hyper-parameters and Gibbs sampling to derive the posterior distributions of model parameters. The model indicates that RNAPII binding as well as histone modifications modulate burst size and burst frequency [221].

As an alternative, a variety of heterogeneity point estimates were computed to quantify biological noise. The variance σ^2, either calculated across all cells or across all expressing cells [19], captures variability in RNA and protein abundance and scales linearly with mean expression μ [222]. The squared coefficient of variation (CV^2) or the Fano factor are more widely used to measure heterogeneous RNA expression [10, 215] and protein abundance [139]. Lowly expressed genes show higher levels of noise compared to highly expressed genes [10]. Therefore, the CV^2 decreases with mean expression. To compare variability

measures across different biological conditions where mean expression changes, regression approaches have been used to correct for the mean-variance relationship [12, 14]. Other approaches directly model biological variability as the excess in dispersion after removing technical noise [11]. Similar to the CV^2 [10] this over-dispersion measure decreases with increasing mean expression [11]. Moreover, heterogeneous expression can be captured by computing the Shannon entropy. Gene-specific entropy is defined as $H = -\sum_i p_i \log_2(p_i)$ where p_i is the probability for a given gene being expressed in bin i. Binning across the expression counts can be done by choosing a fixed width [15] or an adaptive width [50]. Additionally, average pairwise distances between cells can capture increasing or decreasing heterogeneity in cell populations [17].

In general, theoretical work has been done to model the transcription and translation process to predict the distribution of transcripts and proteins across a population of homogeneous cells. However, in reality, the measured point estimates do not perfectly fit the predicted distributions [134]. It is therefore crucial to use non-parametric regression approaches to correct for unobserved confounding factors [13] or the mean-variance trend [32]. This is further discussed in Chapter 3.

1.4 General applications of scRNA-Seq in biology

The following section outlines the applications of scRNA-Seq to a broad spectrum of biological systems. Whole transcriptomic read-outs of individual cells allowed the in-depth characterisation of embryonic development, haematopoiesis, immune responses, the detection of rare cell types and have led to new insights into disease progression including cancer development.

1.4.1 Atlas-type approaches

Until recently, scRNA-Seq technologies generated transcriptomes of less than one thousand cells to study cell type heterogeneity, allele-specific expression or pseudo-temporal trajectories in cellular systems [171]. With the increased scalability of scRNA-Seq technologies, cellular composition of whole tissues and organisms can be assayed. The largest of these so called "expression atlases" to date is the 10X Genomics® brain dataset comprising 1.3 million cells from embryonic mice. It was generated using 133 libraries sequenced on 11 Illumina HiSeq® 4000 flowcells [223]. This experiment has been performed to exemplify the applicability of the commercial 10X Genomics platform to generate more than 10 billion transcriptomes of individual cells across the human body as envisioned by the Human Cell Atlas Consortium [224].

So far, transcriptional atlases that comprise thousands of cells include the mouse cell atlas, a thymus organogenesis atlas, an ageing lung atlas and the full characterisation of cell types in *C. elegans*. For example, Microwell-Seq was developed to capture more than 400,000 cells covering all mouse organs. This analysis reveals rare cell types such as 2-cell-stage like mESCs and allows the construction of a cross-tissue correlation network [225]. Similarly, the *Tabula Muris* aimed at detecting all major cell type across 20 organs of the mouse. Here, the Tabula Muris Consortium used droplet-based 3'-end scRNA-Seq and FACS-based full length transcript analysis to generate (i) a broad atlas and (ii) an in-depth characterisation of each tissue [226]. Cao *et al.*, 2017 generated more than 40,000 cells from the L2 stage *C. elegans* using sci-RNA-Seq and identified nineteen distinct cell types and seven mixed cell types. Furthermore, this atlas allows the dissection of neuronal cell types that split across seven clusters [179]. To study thymus development, Kernfeld *et al.*, 2018 generated around 25,000 transcriptomes of individual cells from the embryonic thymus at E12.5, E13.5, E14.5, E15.5, E16.5, E17.5, E18.5, and P0. This experimental set-up resolves the temporal development

of immune cell types such as T cells, myeolid cells, natural killer cells, innate lymphoid cells, and $\gamma\delta$ T cells as well as thymic epithelial cells [227]. Finally, to study the effect of ageing on a whole tissue, Angelidis *et al.*, 2018 isolated 14,000 cells from lungs of young and old animals and found (i) an increase in transcriptional noise during ageing and (ii) altered transcriptional profiles of alveolar macrophages and type 2 pneumocytes [228].

The following paragraphs summarise scRNA-Seq applications that are aimed at more targeted analysis of regulatory processes.

1.4.2 Developmental biology

Over the last few years, the development of new scRNA-Seq technologies and algorithms to perform data analysis uncovered driving factors in development and cell fate decisions [229]. An early study of mouse embryonic development identified that transcriptional differences between the two cells in the 2-cell stage embryo increase from the zygote to late 2-cell stage embryos. This is caused by an initial partitioning "error" where transcripts are unevenly distributed between the daughter cells. Later on, these differences are strengthened by the onset of transcription coupled to transcriptional noise [230, 231]. A reproducible distribution of transcripts in the first cell division was also detected by Biase *et al.*, 2014 [232]. These biases between cells at the 2-cell stage propagate to form transcription biases at the 4-cell pushing cells towards forming the pluripotent inner cell mass or the extra-embryonic trophoectoderm [16, 231]. To obtain a more complete view on gastrulation in the mouse, Scialdone *et al.* captured cells from the epiblast at E6.5 and mesodermal cells at E7.0, E7.5 and E7.75. The authors also sampled cells from *Tal1* knock-out animals and showed that this transcription factor is a key regulator for blood develeopment [233].

This year, large-scale scRNA-Seq studies profiled organogenesis in the mouse and zebrafish. Ibarra-Soria *et al.*, 2018 sampled more than 20,000 cells from E8.25 embryos following gastrulation and identified 20 major cell types including different mesoderm lineages, neural progenitor cells, blood, gut and extra-embryonic cells. They further used this data to dissect gut formation and to find oscillating expression patterns during somitogenesis [27]. Similarly, inDrop and Drop-Seq approaches were used to generate $\sim$7000 cells from *Drosophila melanogaster* (*D. melanogaster*) embryos at the onset of gastrulation [234] or to generate more than 90,000 cells from the zebrafish embryo during the first day of development [235].

In the last two years, experimental procedures were developed to track cells across multiple divisions termed "lineages". For this, the genome editing tool: clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) [236], was used to introduce so called "scars" at specific DNA sequences. In bacteria, the CRISPR/Cas system is used to degrade invasive DNA which involves a CRISPR RNA that recognises the invading DNA and a Cas protein for degradation. For genome editing purposes, the CRISPR/Cas9 uses single guide RNAs (sgRNAs) to specific genomic sites and induces double strand breaks (DSB). Upon repair, insertions or deletion mutations are introduced that render a specific gene non-functional [237]. The first approach to use the CRISPR/Cas9 for scarring: genome editing of synthetic target arrays for lineage tracing (GESTALT), inserted an array of 10 CRISPR/Cas9 with variable specificity into the genome of individual cells. Upon the expression of the Cas9 protein and the sgRNA, random scars are introduced into the genomic array. After days of growth, genomic DNA was harvested and the array was sequenced to construct the relationship between individual cells **(Fig. 1.17)** [238]. This technology has been extended as a scRNA-Seq approach to capture the RNA together with the expressed CRISPR/Cas9 array that allows cell type assessment alongside lineage detection. This approach also includes a heat shock inducible system to start the scarring at later stages of development [239]. Similar approaches used multiplexed smFISH read-outs to infer lineage relationship between individual cells [240] or tranposase-based insertion of a random 20mer sequence into the genome [235].

One current challenge, especially in the field of developmental biology, is to obtain spatially-resolved whole-transcriptome read-outs of individual cells. The imaging technologies introduced above, MERFISH and SeqFISH, are capable of capturing single RNA molecules of thousands of genes across thousands of cells. Early approaches in the field of spatial transcriptomics employed spatial gene expression atlases to map isolated single cells back into the tissue of origin [241, 242]. A similar approach has recently been used to spatially locate cells isolated form the *D. melanogaster* embryo [234]. Moreover, Tomo-Seq was developed to sequence RNA extracted from slices of the zebrafish embryos. RNA was extracted from each slice into a tube and barcoded prior to sequencing. Matched histology and mathematical modelling was used to reconstruct the spatial expression patterns across the embryo [243].

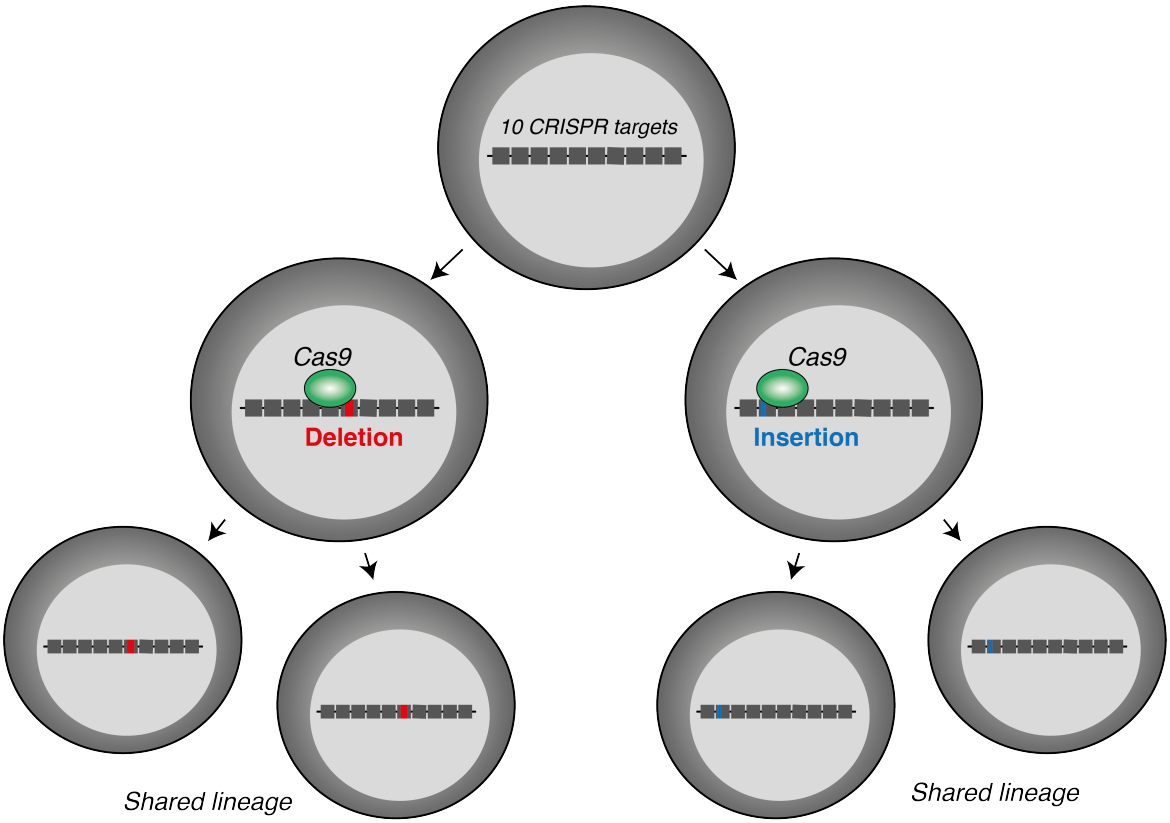


Fig. 1.17: CRISPR scarring for lineage tracing.
A cassette of 10 CRISPR targets is inserted into the genome. Upon Cas9 expression, random insertions and deletions are added to these cassettes. Single-cell DNA or RNA sequencing allows detecting the scars and therefore cell lineage reconstruction.

1.4.3 Cell type evolution

Evolutionary biology is a research field that less frequently uses scRNA-Seq approaches to understand the evolutionary origin of cell types. To do this, cell compositions of non-model organisms such as *Platynereis dumerilii* (*P. dumerilii*) (annelid), *Nematostella vectensis* (*N. vectensis*) (cniderian), *Amphimedon queenslandica* (*A. queenslandica*) (sponge), *Mnemiopsis leidyi* (*M. leidyi*) (ctenophore) and *Trichoplax adhaerens* (*T. adhaerens*) (placozoan) are compared. All of these organisms (except annelids) are non-bilaterians and therefore evolutionary older than mouse and humans. As an example and part of an early project, we used the developing larva of *P. dumerilii* to study diversification of cell types in early bilaterian evolution. We detected cells from the apical neuroectoderm, the midgut, striated musculatrue, ciliated cells and non-apical blastopore cells. By assessing the transcriptional distance between these cell types, we formulated a hypothesis of related cell type families

that originate from an ancesteral cell type and are conserved during evolution [244].

Sebé-Pedrós *et al.*, 2018 generated a single-cell expression atlas of adult and larval *N. vectensis* using MARS-Seq. This dataset allowed the dissection of neuronal diversification and transcription factor regulatory programmes in this early sister group of bilaterians [245]. The authors furthermore generated similar atlases of *A. queenslandica*, *M. leidyi* and *T. adhaerens* and performed cross-species gene module analysis after cell type identification. Co-regulation of cell type-specific gene modules strongly diverged between the species except for a few house keeping modules. Moreover, regulatory TF modules appear to be cell type and species-specific [246]. These studies highlight scRNA-Seq as a powerful tool to study inter-species relationships of cell types in non-model organisms.

1.4.4 Immunology

The immune system has been extensively studied using scRNA-Seq to detect activation responses and to dissect the transcriptional heterogenity among immune cells [247, 248]. White blood cells are broadly grouped into cells of the innate and adaptive immune system. Innate immune cells (DC, mast cells, macrophages, basophils, neutrophils, natural killer (NK) cells, eosinophils) are fast responders that represent the first line of defence upon infection. The adaptive immune system (B cells, $CD4^+$ and $CD8^+$ T cells) responds more slowly but installs an antigenic specificity and immune memory after infection. NK T cells and γδ T cells are cytotoxic lymphocytes that show both innate and adaptive characteristics **(Fig. 1.18)** [249].

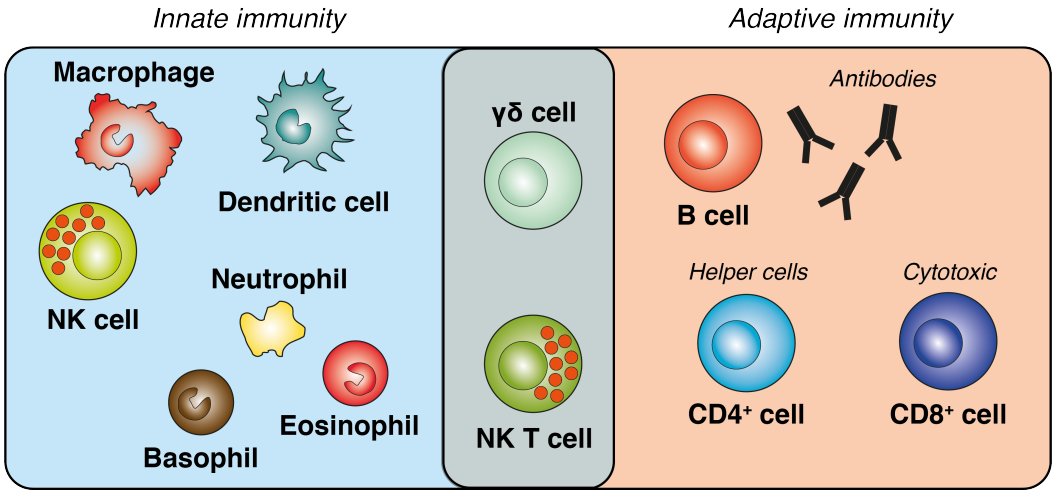


Fig. 1.18: Cell types of the adaptive and innate immune system.

An unbiased approach analysed ∼65,000 human peripheral blood mononuclear cells (PBMCs) and identified the major innate and adaptive immune cell types. The authors further used droplet-based scRNA-Seq to study bone marrow mononuclear cells after haematopoietic stem cell transplant in an acute myeloid leukemia (AML) patient. The technology allowed the distinction between host and donor cells and the detection of residual AML cells in the host [178].

More targeted approaches focused on either cells of the innate or adaptive immune system.

scRNA-Seq to study innate immunity

Villani *et al.*, 2017 used plate-based scRNA-Seq to dissect the DC and monocyte compartment of human PBMCs. In general, DCs can be subdivided into $CD11C^+$ conventional DCs ($CD141^+$ and $CD1C^+$) that activate $CD4^+$ and $CD8^+$ and interferon producing, plasmacytoid DCs. Monocytes were classically subdivided into $CD14^+$ and $CD16^+$ cells. After analysing more than 2,400 DCs and monocytes, the authors expanded DCs to consist of 6 groups and detected conventional DC progenitor cells. Furthermore, they detected two new groups of uncharacterised monocytes [250]. Shalek *et al.*, 2014 used the Fluidigm C1 system to generate transcriptomes of more than 1700 primary mouse bone marrow derived DCs to study their activation response during LPS stimulation. Within one hour of activation, early responding cells up-regulate Ifnβ and support the activation of surrounding cells via paracrine signalling. By isolating activated cells in individual chambers, the authors showed a decrease in the total number of activated cells after 4h stimulation with LPS. Furthermore, activated cells need autocrine stimulation to fully activate and Ifnβ secretion during the first hour of activation is the crucial trigger for homogeneous DC activation [19]. Björklund *et al.*, 2016 performed targeted scRNA-Seq of Lin^-CD127^+ innate lymphoid cells (ILCs) and NK cells from tonsil tissue of adult humans using the SmartSeq2 protocol. They firstly identified the three major lineages of ILCs (ILC1, ILC2 and ILC3) and further assessed heterogeneity within the ILC3 population. Dissecting this rare cell type allows a deeper understanding of immune regulation in humans [251].

scRNA-Seq to study adaptive immunity

The majority of scRNA-Seq studies focused on dissecting heterogeneity in the $CD4^+$ or $CD8^+$ T cell compartment to understand adaptive immunity. For example, Proserpio *et al.*, 2016 studied $CD4^+$ Th2 cell differentiation after *Nippostrongylus brasiliensis* infection.

5 days after infection, three major cell states can be detected: (i) activated cells, (ii) proliferating cells and (iii) Th2 cytokine expressing cells. Furthermore, more differentiated, cytokine-producing cells show higher proliferation which was validated by an *in vivo* model of Th1 differentiation upon malaria infection [252]. A follow-up study characterised the time course of malaria infection by sampling labelled $CD4^+$ T cells 2, 3, 4 and 7 days after infection. Gaussian process modelling of expression changes over the differentiation time course revealed a bifurcation between the Th1 and T follicular helper (Tfh) cell lineage. Identifying such branching dynamics during cell differentiation allows the characterisation of decision-making molecules and underlying regulatory mechanisms [253].

Prior to differentiation and to trigger activation responses, the T cell receptor (TCR) on T cells recognises peptides bound to major histocompatibility complex class I (MHCI) molecules on an antigen-presenting cell (e.g. DC). Upon peptide recognition and co-activator stimulation, cytoskeletal rearrangements, metabolic changes and transcription-factor activity lead to the proliferation and effector differentiation of T cells [254]. Richard *et al.*, 2018 studied the effects of reduced TCR-peptide affinity and discovered that $CD8^+$ cells show heterogeneous response patterns to stimulation with low-affinity peptides. Nevertheless, once activated, $CD8^+$ T cells reach similar activation stages compared to cells activated with high-affinity peptides [254]. Similar to the differentiation time course experiments using $CD4^+$ T cells, Kakaradov *et al.*, 2017 tracked expression changes over the differentiation course of $CD8^+$ T cells after lymphocytic choriomeningitis virus infection. The authors detected strong divergence during the first division of $CD8^+$ T cells after infection which is strengthened by epigenetic silencing of genes associated to memory formation [255].

The development of scRNA-Seq technologies has also driven the expansion of tools to analyse single-cell transcriptome data. Key algorithms for T cell and B cell analysis include the reconstruction of the TCR and B cell receptor (BCR). Constructing the TCR of individual cells allows the identification of clonal expansion of $CD4^+$ T cells upon *Salmonella* infection [256]. Similarly, the BCR can be reconstructed to detect clonal heterogeneity in B cell populations [257, 258]

1.4.5 Tissue function and disease

One example for scRNA-Seq applications to study tissue functions is the work by Sten Linnarsson's lab that aims at dissecting the complexity of the mammalian brain. In an early study, Zeisel *et al.*, 2015 sequenced $\sim$ 3000 cells from the mouse somatosensory cortex

and hippocampal *Cornu Ammonis* (CA)1 region and identified a variety of interneurons, S1 pyramidal neurons, CA1 pyramidal neurons, mural cells, endothelial cells, microglia cells, ependymal cells, astrocytes and oligodendrocytes [259]. More recently, Häring *et al.*, 2018 studied sensory neurons in the dorsal horn of mice and detected multiple sub-groups of gamma-aminobutyric acid (GABA)ergic and glutamatergic neurons [260]. A large-scale study of more than 500,000 cells of the mouse nervous system generated an atlas comprising all major neuronal and non-neuronal cell types. For this, cells from anatomical units of the brain as well as cells from the spinal cord, dorsal root ganglia, sympathetic ganglion and the enteric nervous system were sampled [261]. On a smaller scale, Davie *et al.*, 2018 generated an atlas of 57,000 cells from the *D. melanogaster* brain from young and old animals [262].

Bach *et al.*, 2017 used droplet-based scRNA-Seq to dissect the development of the adult mammary gland. By sampling cells from the mammary gland at different time points (8 weeks virgin, 14.5 days gestation, 6 days lactation and 11 days post involution), the authors were able to reconstruct differentiation processes including the development of hormone-sensing and secretory cells [263]. To study the spatial expression pattern of the liver, Halpern *et al.*, 2017 mapped scRNA-Seq profiles onto 9 layers around the central vein. This approach allows the detection of novel gene expression patterns that correlated with the distance of the cell layer to the central vein [264]. Young *et al.*, 2018 generated 70,000 single-cell transcriptomes from human renal tumours as well as healthy foetal, pediatric, and adult kidneys. With this data, the authors linked childhood Wilms tumour to abnormal foetal cells and identified precursor cells for adult tumours [265].

Extending this, scRNA-Seq was used to study malignant tumours where full characterisation is hindered by complex cellular heterogeneity. Early work by Patel *et al.*, 2014 identified strong heterogeneity within glioblastomas of five patients related to oncogenic signalling, proliferation, immune response, and hypoxia. Furthermore, within each tumour, cells are classified as different glioblastoma sub-types which potentially complicates treatment strategies [266]. More recently, scRNA-Seq has been performed to dissect the multicellular ecosystems of metastatic melanoma [267] and head and neck cancer [268]. In both studies, malignant cells clustered patient-dependently and non-malignant cells clustered based on their cell type. These studies highlight the extreme transcriptional heterogeneity within and between tumours which complicates standard therapies.

1.5 Bayesian approaches to model scRNA-Seq data

In the last decade, an array of tools to process and analyse scRNA-Seq data has been developed. These methods include tools for data acquisition (e.g. alignment, de-duplication, quantification), data filtering (e.g. quality control, normalisation, imputation), cell labelling (e.g. clustering, classification, ordering) and gene-level analysis (e.g. differential expression, detection of expression patterns) [269]. Extensive comparisons of these methods have been performed for each stage of the analysis pipeline [270, 271]. In this thesis, I will focus on the application and development of Bayesian statistical methodologies designed to characterise cellular heterogeneity using scRNA-Seq data. This section describes key concepts of Bayesian inference and related previous work that serve as a foundation for the results in later chapters.

1.5.1 The basics of Bayesian inference

The main difference between classical and Bayesian inference is the treatment of model parameters. While classical inference considers model parameters as fixed but unknown values, Bayesian approaches treat parameters as random variables for which probability distributions quantify uncertainty [272]. In this context, prior beliefs about the distribution of the model parameter ω are summarised in the form of a *prior distribution* $\pi(\omega)$. Once the data D is observed, the prior distribution is updated using the Bayes theorem [273] to form the posterior distribution $\pi^*(\omega|D)$:

$$\pi^*(\omega|D) = \frac{L(D|\omega)\pi(\omega)}{L(D)}, \text{ where } L(D) = \int_\omega L(D|\omega)\pi(\omega)d\omega \tag{1.5}$$

Here, $L(D|\omega)$ is the likelihood of observing the data given the parameter ω and $L(D)$ is the marginal likelihood after integrating out the parameter ω. As discussed in **Section 1.5.2**, $L(D)$ does not have a closed form, except for specific prior choices. Despite this, the numerical methods described in **Section 1.5.3** enable estimating the posterior distribution $\pi^*(\omega|D)$ without the need of calculating $L(D)$.

1.5.2 Prior distributions

The role of the prior distribution is to incorporate prior knowledge about the model parameters. Using the Bayes theorem, this is then combined with the data to form the posterior distribution. For this purpose, a prior distribution should ideally describe the experimenter's prior knowledge regarding the unknown parameters (e.g. based on previous experiments). For practical use, the prior distribution can be chosen to form an analytically tractable solution for the integral that is required to calculate $L(D)$. This is achieved by using conjugate prior distributions, for which the prior is of the same family as the posterior distribution. As such, conjugate prior distributions lead to a closed form posterior distribution, which facilitates posterior inference. A list of commonly used conjugate prior distributions can be seen in **Table 1.3** which was taken from Fink *et al.*, 1997 [274].

Table 1.3: Conjugate prior distributions for common likelihood functions.

Discrete		
Data generation process	**Prior**	**Posterior**
Bernoulli	Beta	Beta
Poisson	Gamma	Gamma
Negative Binomial	Beta	Beta
Continuous		
Data generation process	**Prior**	**Posterior**
Uniform	Pareto	Pareto
Normal (unknown mean)	Normal	Normal
Normal (unknown variance)	Inverse Gamma	Inverse Gamma
Gamma (unknown rate)	Gamma	Gamma
Exponential	Gamma	Gamma

When prior knowledge of the model parameters is not available, non-informative or objective priors may be used (e.g. the Jeffreys prior [275]). However, a detailed discussion regarding such priors is outside the scope of this thesis.

1.5.3 Posterior inference

Before the wide availability of computers, Bayesian research centred around finding pairs of likelihood functions and prior distributions that produce well-defined and tractable solutions for posterior distributions (conjugate priors). More recently, the increase in computing power

supported the development of numerical methods to approximate the integrals needed to form posterior distributions [274]. Numerical approximations are frequently needed when objective priors are used or for models with large complexity. Here, I will focus on Markov Chain Monte Carlo (MCMC) [276, 277], one of the most popular strategies to approximate posterior distributions. The idea behind MCMC is to generate a random sample of the posterior distribution $\pi^*(\omega|D)$ when this distribution cannot be obtained in closed form. For this, a Markov chain is simulated over n iterations whose equilibrium distribution is $\pi^*(\omega|D)$ [278]. Extensive research led to the development of algorithms that generate this equilibrium distribution [279–283]. The following two examples of MCMC are commonly used for a range of applications in Bayesian statistics [278].

Gibbs sampling

Consider a statistical model for which $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ represents a vector of unknown parameters. If the joint posterior $\pi^*(\boldsymbol{\theta}|D) = \pi^*(\theta_1, ..., \theta_K|D)$ is not tractable, it may be numerically approximated. For each model parameter θ_k, one defines the *full conditional* distribution as:

$$\pi^*(\theta_k|D, \theta_{k'}, k' \neq k), \quad k = 1, ..., K \tag{1.6}$$

This is the density of the individual component θ_k, given the data and specified values of all other components $\theta_{k'}$ [284] For each iteration t with $t = 1, ..., T$:

$$\begin{aligned}
&\text{draw} \quad \theta_1^{(t+1)} \quad \text{from} \quad \pi^*(\theta|D, \theta_2^{(t)}, ...\theta_K^{(t)}) \\
&\text{draw} \quad \theta_2^{(t+1)} \quad \text{from} \quad \pi^*(\theta|D, \theta_1^{(t+1)}, \theta_3^{(t)}, ...\theta_K^{(t)}) \\
&\qquad\qquad\qquad\qquad\quad . \\
&\qquad\qquad\qquad\qquad\quad . \\
&\qquad\qquad\qquad\qquad\quad . \\
&\text{draw} \quad \theta_K^{(t+1)} \quad \text{from} \quad \pi^*(\theta|D, \theta_1^{(t+1)}, ...\theta_{K-1}^{(t+1)})
\end{aligned}$$

For $t \to \infty$ the joint distribution of $(\theta_1^{(t)}, ..., \theta_k^{(t)})$ converge to the posterior distribution $\pi^*(\boldsymbol{\theta}|D)$ [285, 284]. The implementation of Gibbs sampling is straightforward when the full conditionals have a known form. Alternatively, stochastic simulation techniques can be used to generate draws from one or more full conditionals.

Metropolis-Hastings

One of these techniques is the *Metropolis-Hastings algorithm* [276, 277] which constructs a Markov chain $\theta_k^{(1)},...,\theta_k^{(T)}$ as follows:
For each iteration $t = 1,...,T$, let $q(\theta_k^{(t)},\theta_k')$ be a proposal distribution for which, given a current draw $\theta_k^{(t)}$, a candidate value θ_k' is proposed. Subsequently, with some probability $\alpha(\theta_k^{(t)},\theta_k'|\theta_{k'}^{(t)},k' \neq k)$ the proposed value θ_k' is accepted (i.e. $\theta_k^{(t+1)} = \theta_k'$) and otherwise rejected ($\theta_k^{(t+1)} = \theta_k^{(t)}$) [285, 277]. In practice, the update is performed as follows:

1. Sample $v \sim \text{Unif}(0,1)$ and a candidate θ_k' from $q(\theta_k^{(t)},\theta_k')$.

2. Define

$$\alpha(\theta_k^{(t)},\theta_k'|\theta_{k'}^{(t)},k' \neq k) = \min\left\{1, \frac{\pi^*(\theta_k'|D,\theta_{k'}^{(t)},k' \neq k)q(\theta_k',\theta_k^{(t)})}{\pi^*(\theta_k^{(t)}|D,\theta_{k'}^{(t)},k' \neq k)q(\theta_k^{(t)},\theta_k')}\right\} \tag{1.7}$$

3. If $v \leq \alpha(\theta_k^{(t)},\theta_k'|\theta_{k'}^{(t)},k' \neq k)$, set $\theta_k^{(t+1)} = \theta_k'$ otherwise set $\theta_k^{(t+1)} = \theta_k^{(t)}$

For this algorithm, the proposal distribution $q(\theta_k^{(t)},\theta_k')$ needs to be chosen. A common choice is a Normal distribution centred at $\theta_k^{(t)}$, where the variance needs to be selected to have some level of optimality in the performance of the algorithm [286]. An automated tuning process for the variance of the proposal distribution $q(\theta_k^{(t)},\theta_k')$ was introduced by Roberts and Rosenthal, 2009 [287]. This *adaptive Metropolis-Hastings* algorithm is often used in combination with Gibbs sampling (*adaptive Metropolis-within-Gibbs sampling*) to approximate the posterior distribution of model parameters for complex models [287].

Practical considerations

When an MCMC sampler is chosen, one needs to assess the convergence of the chain. For this purpose, the initial iterations of the MCMC algorithm are typically discarded (*burn-in*). Within the burn-in period, the autocorrelation of the chain is expected to decay to a negligible level [281]. After burn-in, one can compute the autocorrelation of the chain to assess how well the chain mixed. High autocorrelation suggest that subsequent MCMC draws are similar, and therefore the chain explores the parameter space slowly (i.e. it may take a long time before the chain converges). The standard deviation of the chain (as a measure of statistical uncertainty) is defined as:

$$\sigma = \frac{\sigma^*}{\sqrt{T}}\sqrt{\frac{1+\rho}{1-\rho}} \tag{1.8}$$

Here, σ^* is the posterior standard deviation of $\pi^*(\omega|D)$, T is the sample size and ρ the autocorrelation [288]. Therefore, σ decreases when ρ is small. This also supports finding an optimal run length until sufficient mixing is achieved. In practice, storing only every 10 or 100 samples (thinning) reduces autocorrelation of the chain and reduces storage requirements [281]. One formal way of assessing the convergence of the chain was introduced by Geweke, 1992. Here, the means of the first 10% and the last 50% of the samples are compared. If the means are different, this test suggests that the chain has not yet reached equilibrium [289]. Alternatively, graphical summaries, such as traceplots, can be used to formally assess convergence.

1.5.4 Variational Bayes

When datasets are large and models are complex, the above described MCMC sampling methods are slow to derive posterior distributions. Instead, variational inference can derive an approximate posterior distribution using optimisation, rather than sampling. The principle of variational inference is to select a member of a family of densities Q by minimising the Kullback-Leibler divergence (KL):

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in Q}{\operatorname{argmin}} \operatorname{KL}(q(\boldsymbol{\theta})||\pi^*(\boldsymbol{\theta}|D)) \tag{1.9}$$

A common choice in variational Bayesian approaches is to assume that model parameters are mutually independent so that the *mean-field variational family* of distributions can be chosen for $q(\boldsymbol{\theta})$:

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{k} q_j(\theta_j) \tag{1.10}$$

Here, each model parameter θ_j is governed by its own variational factor [290].

The posterior distribution is then approximated by $q^*(\boldsymbol{\theta})$[290]. In general, variational inference tends to be faster than MCMC, albeit MCMC allows producing exact samples from the target density [290]. Therefore, variational inferences may be preferred when datasets are large and exact samples are not needed. A common approach to minimise the KL is to maximise the evidence lower bound (ELBO) (see e.g. [291]) which is defined as:

$$\operatorname{ELBO}(q) = \mathbb{E}[\log(L(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta}))] - \mathbb{E}[\log(q(\boldsymbol{\theta}))] \tag{1.11}$$

One commonly used technique to maximise the ELBO is *coordinate ascent mean-field variational inference (CAVI)*. Similar to Gibbs sampling, CAVI maximises the ELBO for one parameter while keeping all other parameters constant. This is done iteratively until the ELBO converges against a local maximum [290].

1.5.5 Bayesian decision theory

Assume the data D have arisen under one of the models M_1 or M_0. The posterior probabilities for each model are denoted as $\pi^*(M_1|D)$ and $\pi^*(M_0|D)$. Similarly, prior probabilities for each model are denoted as $\pi(M_1)$ and $\pi(M_0)$. The Bayes factor B_{10} [292] is defined as the ratio of the posterior odds of M_1 to its prior odds:

$$B_{10} = \frac{\pi^*(M_1|D)}{\pi^*(M_0|D)} \Big/ \frac{\pi(M_1)}{\pi(M_0)} = \frac{L(D|M_1)}{L(D|M_0)} \tag{1.12}$$

When the models M_1 and M_0 are equally probable *a priori*, the Bayes factor is equal to the posterior odds in favour of M_1 [293]. To compute the Bayes factor, one needs to find the marginal likelihoods $L(D|M_1)$ and $L(D|M_0)$ (see equation (1.5)). Typically, these marginal likelihoods are intractable and this measure is difficult to compute.

Alternatively, *tail posterior probabilities* can be computed as a selection rule regarding M_1 and M_0 when these models are defined by restrictions in the parameter space (e.g. $M_1 : \theta \in \Theta$ versus $M_0 : \theta \notin \Theta$). For example, posterior tail probabilities were introduced to test the difference δ_g in log-expression of gene i between condition A and condition B [294]. Here, the posterior tail probability of δ_g being larger than a given threshold $\delta_g^{(\alpha)}$ is defined as:

$$\pi(\delta_g, \delta_g^{(\alpha)}) = P\left\{|\delta_g| > \delta_g^{(\alpha)}|D\right\} \tag{1.13}$$

In the case of testing changes in mean expression, the difference δ_g represents the log-fold change in mean expression ($\log(\frac{\mu^{(B)}}{\mu^{(A)}})$). In practice, for each iteration of the MCMC, this difference is computed and the tail posterior probability is the fraction of the absolute distance being larger than the threshold. If the tail posterior probability is larger than an evidence threshold (e.g. 80%) one would reject the null hypothesis $|\delta_g| \leq \delta_g^{(\alpha)}$ [295].

1.5.6 Modelling scRNA-Seq data

Several approaches have been proposed to estimate model parameters based on scRNA-Seq data. Commonly, the count data is modelled as negative binomial (NB) distributed [11, 296, 297]. The NB distribution is defined as:

$$f_{NB}(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)y!}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y} \tag{1.14}$$

Here, the dispersion of the NB is $\delta = \theta^{-1}$ [296]. In terms of a hierarchical model, the NB can be decomposed as a Poisson distribution with a Gamma random effect [11]:

$$\begin{aligned} y|\cdot &\sim \text{Poisson}(\nu\mu) \\ \nu|\alpha,\beta &\sim \text{Gamma}(\alpha,\beta) \end{aligned}$$

In some cases [296, 297] the NB is extended to account for dropout events in scRNA-Seq data [298]. The zero-inflated negative binomial (ZINB) takes the form:

$$f_{ZINB}(y;\mu,\theta,\pi) = \pi\delta_0(y) + (1-\pi)f_{NB}(y;\mu,\theta) \tag{1.15}$$

where $\delta_0(\cdot)$ is the Dirac function and $\pi \in [0,1]$ is the probability that 0 is observed instead of the count y [296]. Nevertheless, it has been shown that the zero inflation is not needed to capture expression dropouts as it can be modelled by the over-dispersion in the NB distribution [297]. In a hierarchical formulation this model writes as:

$$\begin{aligned} y|\cdot &= \begin{cases} x & \text{if } h = 0, \\ 0 & \text{otherwise} \end{cases} \\ x|\cdot &\sim \text{Poisson}(\nu\mu) \\ h &\sim \text{Bernoulli}(p) \\ \nu|\alpha,\beta &\sim \text{Gamma}(\alpha,\beta) \end{aligned}$$

Other approaches model scRNA-Seq counts as log-normally distributed [299, 300]. Zero-inflated factor analysis (ZIFA) assumes that the data $Y = [y_1,...,y_N]$, where N is the number of samples with D genes, are generated from an unobserved low-dimensional space $Z = [z_1,...,z_N]$ with dimension K, $K \ll D$ [300]. The generation process is a linear transformation with added Gaussian noise from the latent space ($N \times K$) into the latent high-dimensional gene expression space with dimension $N \times D$. Additionally, with some probability being a

function of the latent expression level of gene j in cell i x_{ij}: $p_0 = \exp(-\lambda x_{ij}^2)$, a dropout is observed [300]. The full model is defined as:

$$
\begin{aligned}
y_{ij} &= \begin{cases} x_{ij} & \text{if } h_{ij} = 0, \\ 0 & \text{otherwise} \end{cases} \\
\boldsymbol{x}_i | \boldsymbol{z}_i &\sim \text{Normal}(\boldsymbol{A}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{W}) \\
h_{ij} | x_{ij} &\sim \text{Bernoulli}(p_0) \\
\boldsymbol{z_i} &\sim \text{Normal}(0, \boldsymbol{I})
\end{aligned}
$$

Here, $\boldsymbol{A}$ denotes a $D \times K$ factor loadings matrix, $\boldsymbol{\mu}$ a $D \times 1$ mean vector, $\boldsymbol{W}$ a $D \times D$ covariance matrix and $\boldsymbol{I}$ the $K \times K$ identity matrix.

1.5.7 BASiCS: Bayesian Inference of Single-Cell Sequencing data

Throughout this thesis, I will use and extend the Bayesian Inference of Single-Cell Sequencing data (BASiCS) framework [11, 295]. BASiCS models scRNA-Seq data generated from seemingly homogeneous populations of cells (i.e. where no distinct sub-populations are found) to perform down-stream analysis (e.g. normalisation, differential expression testing).

In BASiCS, the expression count of gene i ($\in \{1, \ldots, q\}$) in cell j ($\in \{1, \ldots, n\}$) X_{ij} is treated as a random variable. Compared to bulk RNA-Seq, scRNA-Seq is inherently noisy due to low starting amounts of RNA [10]. To control for technical noise, BASiCS incorporates reads from synthetic RNA spike-ins [301]. Here, the first q_0 genes are biological and the remaining $q - q_0$ genes are technical.

$$
X_{ij} | \mu_i, \phi_j, \nu_j, \rho_{ij} \sim \begin{cases} \text{Poisson}(\phi_j \nu_j \mu_i \rho_{ij}), & i = 1, ..., q_0, j = 1, ...n; \\ \text{Poisson}(\nu_j \mu_i), & i = q_0 + 1, ..., q, j = 1, ..., n, \end{cases} \tag{1.16}
$$

In this model, two random effects were added to model the technical and biological part of the over-dispersion:

$$
\nu_j | s_j, \theta \sim \text{Gamma}\left(\frac{1}{\theta}, \frac{1}{s_j \theta}\right), \quad \rho_{ij} | \delta_i \sim \text{Gamma}\left(\frac{1}{\delta_i}, \frac{1}{\delta_i}\right) \tag{1.17}
$$

Note that this formulation relates to the hierarchical representation of the NB distribution, but this model is more general as two sets of random effects are incorporated. Here, ϕ_j represents a cell-specific normalisation parameter to correct for differences in mRNA content between cells and s_j models cell-specific scale differences affecting all biological and technical genes

(e.g. amplification biases or RNA capture biases). Moreover, the random effect ν_j captures unexplained technical noise that is not accounted for by the normalisation. The strength of this noise is then quantified by a global parameter θ (shared across all genes and cells). Heterogeneous gene expression across cells is captured by ρ_{ij}, whose strength is controlled by gene-specific over-dispersion parameters δ_i. These quantify the excess of variability that is observed with respect to Poisson sampling noise, after accounting for technical noise. Finally, gene-specific parameters μ_i represent average expression of a gene across cells **(Fig. 1.19A)**.

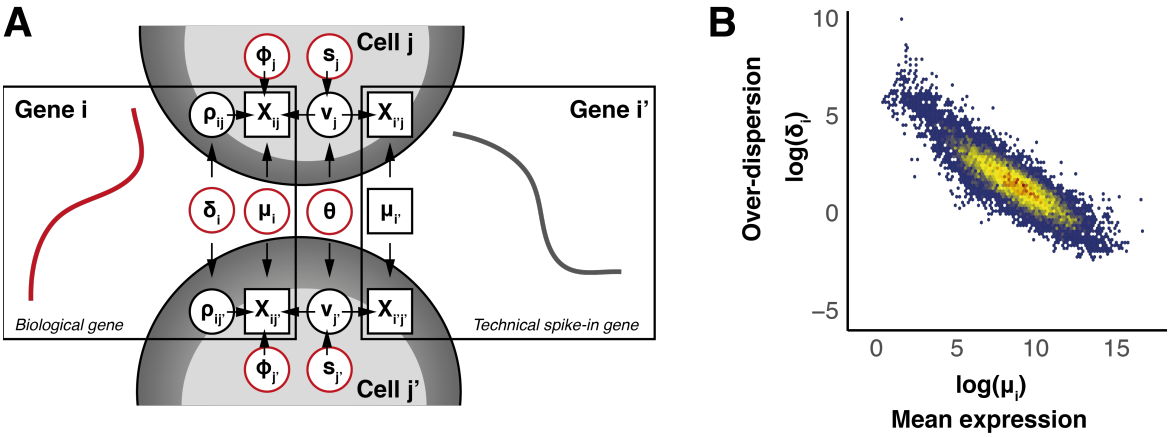


Fig. 1.19: The BASiCS model.
(A) Hierarchical formulation of the statistical model underlying BASiCS visualised for two cells (j and j') and two genes (i and i', gene i' represents a technical spike-in gene). Squared nodes indicate known quantities (observed expression counts and added number of spike-in molecules). Round nodes indicate unknown quantities. Red circles represent unknown model-parameters while black circles indicate the random effects that play intermediate roles effecting expression counts. Adapted from [11], **(B)** Illustration of the typical confounding effect that is observed between gene-specific estimates of over-dispersion parameters δ_i and mean expression parameters μ_i.

Prior specifications for the model parameters are chosen as follows:

$$\begin{aligned}
\mu_i &\sim \text{log-N}(0, a_\mu^2) \quad \text{for } i = 1, \ldots, q_0 \\
\delta_i &\sim \text{log-N}(0, a_\delta^2) \quad \text{for } i = 1, \ldots, q_0 \\
s_j &\sim \text{Gamma}(a_s, b_s), \quad j = 1, \ldots, n \\
\theta &\sim \text{Gamma}(a_\theta, b_\theta) \\
\Phi &\sim n\text{Dirichlet}(a_\Phi), \quad \Phi = (\phi_1, \ldots, \phi_n)
\end{aligned}$$

After integrating out the ρ_{ij} to enhance mixing of the MCMC algorithm that was adopted to perform posterior inference [11] the likelihood is defined as:

$$\mathscr{L} = \left[\prod_{i=1}^{q_0}\prod_{j=1}^{n}\frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})x_{ij}!}\left(\frac{\frac{1}{\delta_i}}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\phi_j\nu_j\mu_i}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right)^{x_{ij}}\right] \times\left[\prod_{i=q_0+1}^{q}\prod_{j=1}^{n}\frac{(\nu_j\mu_i)^{x_{ij}}}{x_{ij}!}\exp\{-\nu_j\mu_i\}\right] \tag{1.18}$$

Given the NB model, the expected biological counts take the form:

$$\mathbb{E}(X_{ij}|\mu_i,\delta_i,\phi_j,s_j,\theta) = \phi_j s_j \mu_i \tag{1.19}$$

This formulation can be used to obtain normalised counts. Furthermore, the coefficient of variation is defined as:

$$\mathrm{CV}^2(X_{ij}|\mu_i,\delta_i,\phi_j,s_j,\theta) = \frac{1}{\phi_j s_j \mu_i} + \theta + \delta_i(\theta+1) \tag{1.20}$$

As discussed in Vallejos *et al.*, 2016 [295], the CV^2 is inversely proportional to the mean expression μ_i. Furthermore, δ_i can be interpreted as the residual CV^2 after removing Poisson sampling and residual technical over-dispersion [11, 295]. We will therefore use δ_i as a proxy for the biological part of transcriptional variability when modelling scRNA-Seq data.

Posterior inference is implemented using adaptive Metropolis-within-Gibbs sampling (see **Section 1.5.3**) [11, 295]. Once posterior distributions are obtained, down-stream analyses can be performed. These include: normalisation of expression counts, variance decomposition into biological and technical noise, detection of highly and lowly variable genes and differential mean and differential over-dispersion testing. The latter is done by computing the tail posterior probabilities of the difference in mean expression or over-dispersion between two conditions (p and p') being larger than an evidence threshold τ_0 or ω_0 (see **Section 1.5.5**, and [294, 295]):

$$\pi_{ipp'}(\tau_0) \equiv P(|\log(\mu_i^{(p)}/\mu_i^{(p')})| > \tau_0|D) > \alpha_m$$
$$\pi_{ipp'}(\omega_0) \equiv P(|\log(\delta_i^{(p)}/\delta_i^{(p')})| > \omega_0|D) > \alpha_d$$

If the tail posterior probability is larger than a given propability threshold α_m or α_d, the gene is considered to be differentially expressed or differentially over-dispersed [295]. The evidence threshold is usually fixed *a priori* and the probability threshold is defined to control the expected false discovery rate (EFDR) to (e.g.) 5% [302, 295].

In this model [295], estimates of the over-dispersion parameters δ_i are negatively correlated to mean expression μ_i **(Fig. 1.19B)** indicating that in homogeneous populations of cells, highly expressed genes tend to be less noisy than lowly expressed genes. This effect confounds differential over-dispersion testing between two populations when mean expression changes. Therefore, when assessing changes in over-dispersion, only genes with no changes in mean expression are considered (see Vallejos *et al.*, 2016 [295] and **Section 2.3**).

1.6 Outline

The overarching topic of this thesis is the quantification and interpretation of transcriptional noise as measured by scRNA-Seq. **Chapter 2** presents an initial experiment to study how transcriptional noise effects the immune system. We see that transcriptional noise increases across multiple immune response genes during ageing which therefore could explain a disrupted immune response in older individuals. This finding has been published in the following paper:

> Celia P. Martinez-Jimenez*, Nils Eling*, Hung-Chang Chen, Catalina A. Vallejos, Aleksandra Kolodziejczyk, Frances Connor, Lovorka Stojic, Tim F. Rayner, Michael J. T. Stubbington, Sarah A. Teichmann, Maike de la Roche, John C. Marioni, Duncan T. Odom. Ageing increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 1436: 1433-1436, 2017,
> (* equal contributions) ■

The computational code associated to this project can be found at:
https://github.com/MarioniLab/ImmuneAging2017

Studying changes in variability between two conditions was restricted to genes that did not change in mean expression due to a strong confounding between variability and mean expression. In **Chapter 3**, I therefore extended the statistical framework from chapter 2 to correct for this confounding effect. This correction leads to (i) a stabilisation of model parameters, (ii) expansion of the gene set that can be tested for changes in variability and (iii) a novel way of interpreting transcriptional dynamics. This project has been published as:

> Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, Catalina A. Vallejos. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems*, 7: 284-294, 2018 ■

Scripts used for analysis are deposited at:
https://github.com/MarioniLab/RegressionBASiCS2017

The extended model offers the unique opportunity to study changes in variability across multiple cell types even when mean expression changes. In **Chapter 4**, I apply the newly developed model to test changes in variability over pseudo-time. For this, droplet-based scRNA-Seq data of mouse spermatogenesis was used to dissect the transcriptional dynamics during this developmental process. Parts of the study are available online as:

> Christina Ernst*, Nils Eling*, Celia P. Martinez-Jimenez, John C. Marioni, Duncan T. Odom. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *bioRxiv*, 2018, (* equal contributions) ■

Code for computational analysis can be found at:
https://github.com/MarioniLab/Spermatogenesis2018

Finally, I will discuss current challenges in modelling transcriptional noise from scRNA-Seq data and experimental strategies to modulate expression variability.

1.7 Other contributions

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

> Kaia Achim*, Nils Eling*, Hernando Martinez Vergara, Paola Yanina Bertucci, Jacob Musser, Pavel Vopalensky, Thibaut Brunet, Paul Collier, Vladimir Benes, John C. Marioni, Detlev Arendt. Whole-Body Single-Cell Sequencing Reveals Transcriptional Domains in the Annelid Larval Body. *Molecular Biology and Evolution*, 35: 1047-1062, 2018, (* equal contributions) ■

> Christina Ernst, Jeremy Pike, Sarah J. Aitken, Hannah K. Long, Nils Eling, Lovorka Stojic, Michelle C. Ward, Frances Connor, Timothy F. Rayner, Margus Lukk, Robert J. Klose, Claudia Kutter, Duncan T Odom. Successful transmission and transcriptional deployment of a human chromosome via mouse male meiosis. *eLife*, 5: e20235, 2016 ■

The github repository containing the source code of this thesis can be found here:
https://github.com/nilseling/Thesis

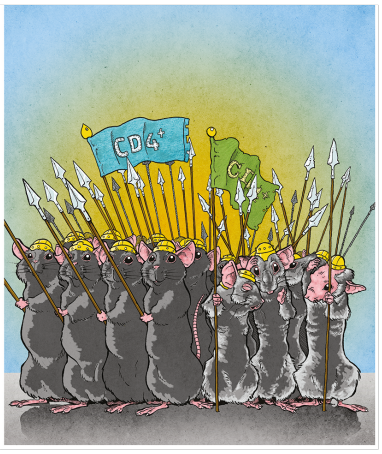
2

Ageing increases transcriptional noise in CD4$^+$ T cell activation

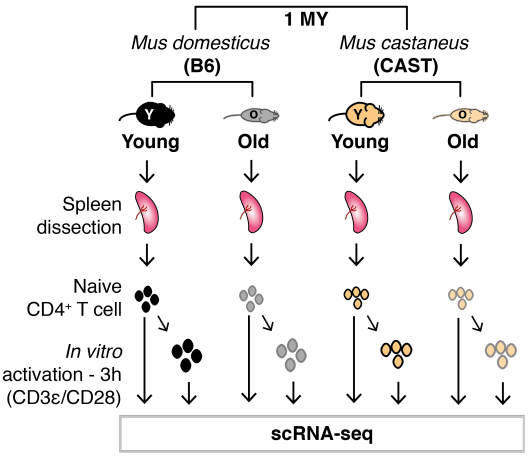
Ageing is characterised by progressive loss of physiological and cellular functions, but the molecular basis of this decline remains largely unexplored. Here, we study how ageing impacts transcriptional dynamics using single-cell RNA-sequencing of over a thousand unstimulated and stimulated naive and effector memory CD4$^+$ T cells from young and old mice. Furthermore, we sampled cells from two divergent strains of mice to assess the evolutionary conservation of the molecular ageing signature. In young animals, immunological activation drives a transcriptomic switch from variable to tightly regulated gene expression, characterised by a strong up-regulation of a core activation program, coupled to a decrease in cell-to-cell variability. The up-regulation of a set of immune response genes is conserved between the two mouse strains as is the decrease in expression variability upon immune activation. Ageing significantly perturbed the activation of the core immune response programme by increasing expression heterogeneity across different populations of CD4$^+$ T cells. This discovery adds transcriptional noise as an unexplored hallmark of ageing to the list of known phenotypic changes. ■

Declaration This work was a joint effort of the Marioni, Odom, de la Roche and Teichmann labs. Celia P. Martinez-Jimenez, Duncan T. Odom, John C. Marioni and Sarah Teichmann designed the study. Celia P. Martinez-Jimenez and Aleksandra A. Kolodziejczyk performed preliminary experiments. Celia P. Martinez-Jimenez performed all single-cell RNA sequencing experiments displayed in this chapter. Hung-Chang Chen and Maike de la Roche provided extensive support during the revision process. Hung-Chang Chen performed FACS experiments during the revision process. Lovorka Stojic and Frances Connor provided experimental support. Timothy F. Rayner provided technical support. Michael J. T. Stubbington performed the T cell receptor and clone analysis. Catalina A. Vallejos helped with the statistical analysis by providing additional explanations of the BASiCS model. Celia P. Martinez-Jimenez and I interpreted results. Celia P. Martinez-Jimenez, Duncan T. Odom, John C. Marioni and I wrote the manuscript. Duncan T. Odom and John C. Marioni supervised the study. I performed computational analysis of all data displayed in this chapter and generated all figures with the exception of the FACS analysis (Fig. 2.2 and Fig. 2.15A-C). The paper has been published as:

Celia P. Martinez-Jimenez*, Nils Eling*, Hung-Chang Chen, Catalina A. Vallejos, Aleksandra Kolodziejczyk, Frances Connor, Lovorka Stojic, Tim F. Rayner, Michael J. T. Stubbington, Sarah A. Teichmann, Maike de la Roche, John C. Marioni, Duncan T. Odom. Ageing increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 1436: 1433-1436, 2017, (* equal contributions)



Credit to Spencer Phillips, EBI



2.1 Introduction

Ageing is characterised by the progressive decline of physiological and cellular functions [303, 304]. Nine hallmarks of ageing have been described to determine the ageing phenotype: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, de-regulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication [303]. Ageing can have a complex and tissue-specific impact on gene expression levels [305], as seen by microarray expression analyses of collections of mouse $CD4^+$ and $CD8^+$ T cells [306], rat hepatocytes [307], mouse and human brain [308, 309], human muscle [310, 311], human kidney [312], human retina [313], and different species of Drosophila and Caenorhabditis [314]. For instance, ageing affects distinct functional pathways, even in closely related $CD4^+$ and $CD8^+$ T cells [306].

Transcriptional noise, RNA processing aberrations, impaired DNA repair, and chromosomal instability can be caused by epigenetic changes in DNA methylation, histone modifications and chromatin remodelling [303]. Global DNA methylation slightly decreases during ageing but increases in common disease-related genes over the lifespan of humans [315]. In mice, around 35% of assayed genes showed either increased or decreased DNA methylation over ageing, with substantial tissue-specificity [316]. Similarly, ageing introduces changes in histone modifications such as the increase of activating acetylation of lysine 16 of histone H3 (H4K16ac) and H3K4me3, and repressive tri-methylation of lysine 20 of histone H3 (H4K20me3). Furthermore, ageing decreases the repressive H3K9me3 and H3K27me3 marks [317, 318]. One well studied system that controls cellular function is the sirtuin (*Sir*)2 histone deacetlyase, which is encoded by seven homologs in mammals [319]. The chromatin-associated protein SIRT6 in mice has been shown to protect genomic stability by promoting resistance to DNA damage. Loss of this protein induces ageing-realted phenotypes within 4 weeks of murine lifespan [320]. Similar effects can be seen for SIRT1 [321].

While most studies focused on identifying age-associated gene expression profiles [322], the role of transcriptional noise during ageing has only been sporadically assessed. Analysis of fifteen genes in terminally differentiated cardiomyocytes suggested that ageing can lead to increased cell-to-cell transcriptional variability [102]. In contrast, single-cell analysis of the transcription of six genes in four different haematopoietic cell types showed few cell-to-cell changes between old and young animals, suggesting that transcriptional variability may not be a universal attribute of ageing [103]. Whether cell-to-cell gene expression variability in-

creases during ageing on a genome-wide basis, particularly for dynamic activation programs, remains largely unexplored.

Single-cell RNA sequencing presents a powerful technology to quantify transcriptional variability for thousands of genes across hundreds of cells simultaneously. For example, Kowalczyk *et al.* performed a high-resolution scRNA-seq analysis of haematopoietic stem cells in young and old mice. Here, cell cycle is the primary driver for cell-to-cell variability in gene expression, and ageing decreases the entry of long-term haematopoietic stem cells into G1 phase in a cell type-specific manner [323].

To evaluate the impact of ageing on gene expression levels and cell-to-cell transcriptional variability, we selected CD4$^+$ T cells as model system. As explained in **Box 1** (page 2), transcriptional noise is defined as cell-to-cell variability in expression within a homogeneous population of cells. Naive CD4$^+$ T cells are readily isolated as single, phenotypically homogeneous cells when purified from young and aged spleens and can easily be stimulated into a physiologically-relevant, activated transcriptional state *in vitro*. Furthermore, they are maintained in a quiescent state, but have the ability to respond to antigen stimulation with proliferation and effector differentiation, which is essential for life-long maintenance of adaptive immune function against infection and cancer [324, 325]. With this, they sit at the root of adaptive immunity and disruption of their transcriptional programme can lead to severe phenotypes during ageing.

Previously, comparing gene expression levels in matched tissues from different mammalian species was used as a tool for revealing conserved cell-type-specific regulatory programmes [326–329]. For instance, a conserved set of response genes has been identified by comparison of bulk gene expression between human and mouse CD4$^+$ T cells after immune activation [330]. So far, it is not known whether conservation of gene expression levels is also reflected in cell-to-cell variability.

Here, we dissected the activation dynamics of naive CD4$^+$ T cells at the single cell level during ageing in two sub-species of mice. With this, we assayed transcriptional dynamics during immune response and how ageing possibly perturbs this system. By comparing our findings across divergent strains of mice, we assessed the evolutionary conservation of the immune response and ageing phenotype. Furthermore, we isolated pure naive and effector memory CD4$^+$ T cells to profile age-related changes in different CD4$^+$ T cell subsets.

2.2 Single-cell RNA sequencing of murine CD4+ T cells

To assess transcriptional changes of the immune activation programme during ageing, we isolated CD4+ T cells from healthy individuals of two inbred mouse sub-species separated by 1 million years of divergence: the reference C57BL/6J, *Mus musculus domesticus* (B6) and CAST/EiJ, *Mus musculus castaneus* (CAST). Furthermore, we isolated CD4+ T cells from young (3 months) and old (21 months) individuals. To characterise their gene expression programme, we performed scRNA-Seq using the C1 Fluidigm system **(Fig. 2.1)**. The two sub-species have similar lifespans [331], and CAST mice showed the hallmarks of normal organismal ageing as observed in B6 mice [312]. All mice were healthy at the time of experiments. To assess different CD4+ T cell compartments, we assayed cell populations isolated with different levels of purity. First, we isolated all unstimulated CD4+ T cells from spleens of old and young animals. Secondly, we highly purified naive CD4+ T cells and effector memory (EM) CD4+ T cells. For simplification and clarity, purified unstimulated CD4+ T cells will be named naive; stimulated cells will be named activated. Highly purified naive CD4+ T cells will be named FACS-purified naive CD4+ T cells and highly purified EM CD4+ T cell will be referred to as FACS-purified EM CD4+ T cells. For each species/condition, scRNA-Seq experiments were independently performed using cells isolated from two individuals. Full experimental methods can be found in **Appendix A.1**.

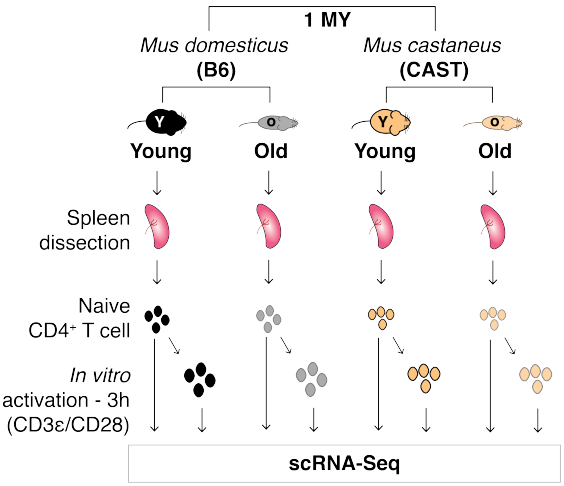


Fig. 2.1: scRNA-Seq of unstimulated and activated CD4+ T cells from young and old B6 and CAST animals.
Single cells were isolated from spleens of young (3 month) and old (21 month) individuals of two related mouse sub-species (*Mus musculus domesticus*, B6; *Mus musculus castaneus*, CAST). Isolated cells were subjected to single-cell mRNA sequencing (scRNA-Seq) before or after 3 hours of *in vitro* activation using anti-CD3ε and CD28 coated plates.

2.2.1 Experimental strategy

Unstimulated CD4+ T cells

Unstimulated CD4+ T cells were purified from dissociated mouse spleens using cell strainers, cell separation media and a magnetic-activated cell sorting (MACS) CD4+ CD62L+ T Cell Isolation Kit (see **Appendix A.1.2**).

Naive and effector memory CD4+ T cells

Naive and EM CD4+ T cells were purified from spleens of both young and old BL6 mice by FACS. Briefly, spleens were harvested from both young and old animals and single cell suspensions were obtained by meshing through a cell strainer. B cells were depleted from cell suspensions by MACS using CD19 microbeads and red blood cells were lysed with red blood cell lysis buffer. The enriched cell fraction was then stained with Fixable eFluor 780 viability dye following Fc receptor blocking with TruStain fcXTM and subsequent staining with a panel of fluorescence-conjugated antibodies against CD4, CD44, CD62L, CD24, Qa2, CD69 and PD-1. Stained cells were immediately sorted using FACS with the stringent gating strategy described in **Fig. 2.2** (see **Appendix A.1.3**).

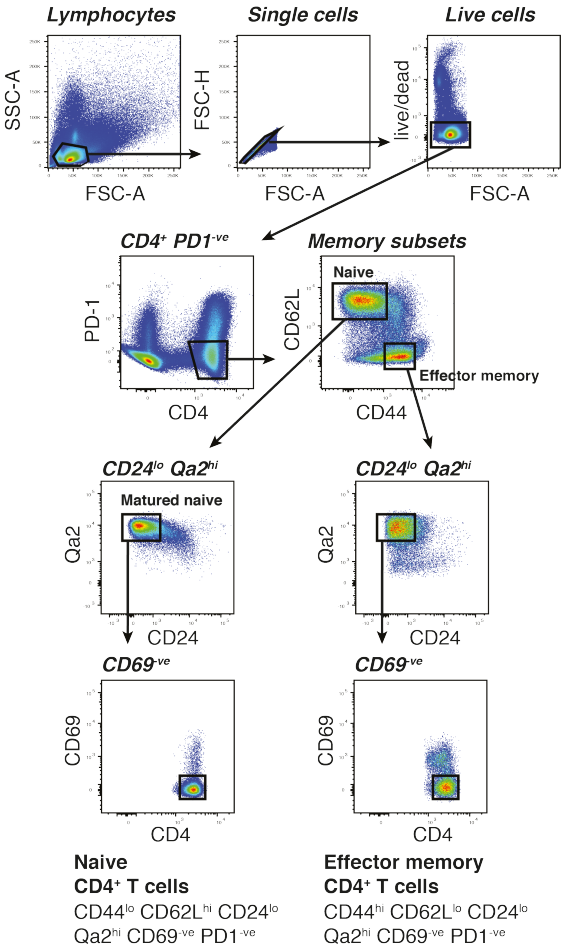


Fig. 2.2: FACS of naive and effector memory CD4+ T cells.
Gating Strategy: lymphocytes were gated by the use of forward scatter (FSC) and side scatter (SSC). Cell doublets were excluded according to area and height of FSC (FSC-A/FSC-H). Dead cells were removed using viability dye. Programmed cell death protein 1 (PD-1)+ CD4+ T cells were excluded and PD-1-ve CD4+ T cells were further separated into naive and EM CD4+ T cell subsets according to their CD44 and CD62L expression. Cells with a mature $CD24^{lo}$ $Qa2^{hi}$ phenotype were then gated from naive and EM subsets and $CD69^{+}$ cells were removed. From [22]. Reprinted with permission from AAAS.

Activation of CD4$^+$ T cells

96-well plates were coated with anti-CD3ε and anti-CD28 antibodies (see **Appendix A.1.2**). After this, naive and FACS-purified naive cells were seeded into these plates at a density of 80,000-120,000 cells/ml, and then cultured in a total volume of 100 µl media that did not contain cytokines or additional antibodies. With this strategy, CD4$^+$ T cells are purely activated but do not commit to a T helper cell fate [332, 333].

scRNA-Seq using the Fluidigm C1 system

Unstimulated and activated CD4$^+$ T cells were loaded on a 5–10 µm Auto Prep IFC to capture single cells using the C1 Single cell Auto Prep System (Fluidigm). All IFCs were visually inspected, and wells with multiple cells or cell debris were marked as low quality. Upon cell capture, reverse transcription and cDNA amplification were performed using the SMARTer PCR cDNA Synthesis Kit and the Advantage 2 PCR Kit. ERCC spike-in RNA (1 µL diluted at 1:50,000) was added to the C1 lysis mix. All capture sites were included for the RNA-Seq library preparation (see **Appendix A.1.4**).

2.2.2 Computational strategy

Read alignment to reference genomes

For all capture sites, read alignment to reference genomes was performed using *gsnap* with default parameters, while supplying splice-site positions [334]. Samples taken from B6 were mapped against the mouse reference Genome Reference Consortium mouse build 38 (GRCm38). CAST samples were aligned against the *Mus musculus castaneus de novo* genome assembly (ftp://ftp-mouse.sanger.ac.uk/REL-1509-Assembly/, now available on Ensembl ftp://ftp.ensembl.org/pub/release-92/fasta/mus_musculus_casteij/dna/), which was used under an advance access agreement (ftp://ftp-mouse.sanger.ac.uk/REL-1509-Assembly/README). Gene annotation for B6 was taken from the GRCm38 reference; gene annotation for CAST was taken from the newly constructed *Mus musculus castaneus* assembly (http://hgwdev.cse.ucsc.edu/~ifiddes/mouse_genomes_data/, version 0.2, now available at ftp://ftp.ensembl.org/pub/release-92/gtf/mus_musculus_casteij/). Additionally, since mitochondrial genes and certain immune genes (e.g., CD28) are absent from the *Mus musculus castaneus* annotation used, and since high mitochondrial gene expression is a well-established signature of low-quality single-cell transcriptome profiles [335], we also mapped CAST reads against GRCm38 and used the B6 annotation solely for the mitochondrial genes. Gene expression

counts were obtained using *HTSeq* with default options [336]. Only genes with orthologs in both species were considered for downstream analysis.

Quality control and filtering

We visually inspected the cell-capture sites in each C1 IFC using 40x magnification lensing to ensure precise capture of single cells **(Fig. 2.3A and B)**. Low-quality cells were computationally filtered using the following quality control criteria:
The percentage of reads mapping to annotated exonic regions was compared to the percentage of reads mapping to ERCC spike-ins. Cells with low genomic reads (< 20%) and/or high ERCC reads (> 50%) were excluded **(Fig. 2.3C)**. Additionally, cells with a low total number of mapped reads (< 1,000,000) were excluded. To exclude possible doublets and dying cells, capture sites with > 3000 or < 1250 detected genes were removed **(Fig. 2.3D and E)**. Next, cells with more than 10% or less than 0.5% of mitochondrial reads were excluded **(Fig. 2.3F)**. These quality filtered cells were tested for possible batch effects by computing a PCA on both replicates. We detect no batch effect since cells from the two individuals are overlapping (see **Fig. 2.3G** for an example of naive and activated CD4$^+$ T cells of young B6).

To control for biological contamination, known markers of lymphocytes were used to filter cells: CD19$^+$/H2-Aa$^+$ B cells as well as CD8$^+$ T cells were removed **(Fig. 2.3H)**. Finally, we visualised naive and activated CD4$^+$ T cells using t-distributed stochastic neighbour embedding (tSNE) and detect a strong grouping depending on the activation status. Here, non-activated T cells that were meant to be activated were removed from downstream analysis **(Fig. 2.3I)**. Read counts were normalised using the BASiCS package [337] incorporating spike-in reads for technical noise estimation. Prior to normalisation, genes not expressed in at least 3 cells (rpm > 20) were filtered out. Similarly, ERCC spike-ins were removed if not detected in the data set.

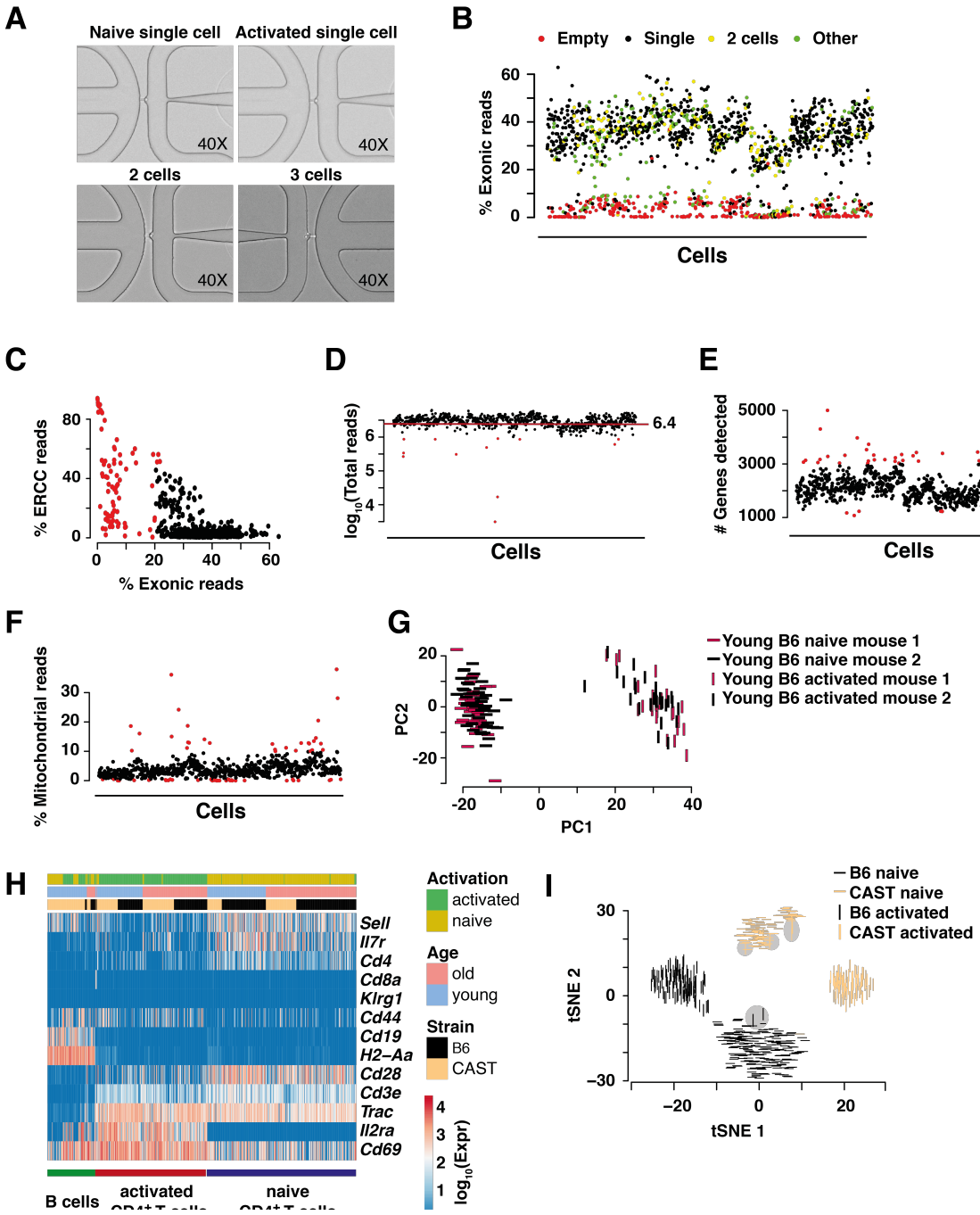


Fig. 2.3: Quality control of isolated CD4+ T cells (Full legend on next page).

Fig. 2.3: Quality control of isolated CD4$^+$ T cells (continued).
(A) Visual inspection of captured cells at 40x magnification in IFCs (C1, Fluidigm) allows manual removal of empty capture sites, and capture sites holding multiple cells or debris, **(B)** Percentage of reads mapping to exonic regions displayed for naive and activated CD4$^+$ T cells. Black dots: single cells, yellow dots: 2 cells, red dots: empty wells, green dots: debris, multiple cells, etc., **(C)** Removal of cells with less than 20% of mapped exonic reads and more than 50% of ERCC spike-in reads (red dots), **(D)** Cells with less than 1 million mapped reads were excluded from downstream analysis (red dots), **(E)** Cells with more than 3000 or less than 1250 genes were excluded in the analysis (red dots), **(F)** Cells with more than 10% or less than 0.5% of mitochondrial reads were excluded from downstream analysis (red dots), **(G)** Naive and activated cells isolated from young B6 animals (replicates) were coloured batch-specifically. 4 batches from 2 mice: naive and activated from mouse 1 (red bars), naive and activated from mouse 2 (black bars). Naive condition is represented in horizontal bars and activated condition in vertical bars, **(H)** Data set was filtered for immune markers to exclude B cell and CD8$^+$ T cell contamination. Cells in columns were labelled based on their activation state (naive in beige, activated in green), their age (old in red, young in blue), and the species of the animals (B6 in black, CAST in yellow). Cells were ordered based on their H-2 class II histocompatibility antigen (*H2-Aa*) and interleukin 2 receptor alpha (*Il2ra*) expression, **(I)** tSNE visualisation allowed the removal of not fully activated cells (indicated in grey circles). Cells were labelled based on their activation state (naive: horizontal bar, activated: vertical bar) and the species of the animals (B6 in black, CAST in yellow). From [22]. Reprinted with permission from AAAS.

BASiCS parameter estimation using transcriptomes of CD4$^+$ T cells

To quantify and assess changes in mean expression and expression variability, we used the Bayesian hierarchical framework BASiCS introduced in **Section 1.5.7** [11, 295]. The MCMC simulation was run on quality filtered transcriptomes of CD4$^+$ T cells condition-specifically for 40,000 iterations using 20,000 burn-in iterations and a thinning factor of 20. We used posterior medians of μ_i to capture mean expression and posterior medians of the over-dispersion parameter δ_i to quantify biological expression variability. Differential mean expression testing was performed using a probabilistic decision rule of $\log_2(\mu_i^{(A)}/\mu_i^{(B)}) > \tau_0$ being larger than a given probability threshold (e.g. 80%). The probability threshold was chosen to keep the EFDR at 5% [295]. Here, A and B indicate the different conditions and τ_0 is the chosen minimum tolerance threshold. The decision rule associated to differential variability testing is: $\log_2(\delta_i^{(A)}/\delta_i^{(B)}) > \omega_0$. Due to the strong confounding between mean expression μ_i and over-dispersion δ_i (as described in **Section 1.5.7**), we only consider genes with no changes in mean expression ($\tau_0 = 0$, EFDR = 5%) to assess changes in variability. Throughout this chapter, the decision rule is abbreviated with: $\log_2$ fold change (log2FC).

2.2.3 Characterisation of isolated CD4$^+$ T cells

To avoid biases during quantification of transcriptional variability in homogeneous populations (see **Box 1** on page 2), careful inspection to remove possible substructures in the isolated cells is needed. Possible drivers for cell-to-cell expression variability include: cell cycle, clonality, cell size, differences in activation state/exhaustion and T cell priming in form of lineage commitment. We therefore assessed these features in the MACS-purified naive CD4$^+$ T cells in their unstimulated and activated state.

Firstly, we perfomed computational analysis to determine cell cycle stage, clonality and cell size. We estimated the cell cycle stage of each cell using *cyclone* [338] implemented in the *scran* R package [339]. In contrast to haematopoeitic cells [323], even when activated, all CD4$^+$ T cells are in G1 phase of cell cycle **(Fig. 2.4A)**. We reconstructed the sequence of the T cell receptor for each cell [332] and did not detect clonal expansion in CD4$^+$ T cells from aged animals **(Fig. 2.4B)**. Similarly, we did not detect difference in cell size that could impact analysis of gene expression variability **(Fig. 2.4C)**.

Secondly, using flow cytometry analysis, we assessed the purity and activation state (Il2rα and Cd69) of CD4$^+$ T cells, confirming that 96.4% of the isolated CD4$^+$ T cells were naive in young B6 **(Fig. 2.4D)**. Old animals had a small population of CD4$^+$ T cells with slightly elevated CD44 levels, reduced CD62L expression, indicative of memory T cells, and attenuated activation dynamics **(Fig. 2.4E-G)**. We did not detect differences in the proportion of lymphocytes (interleukin 7 receptor (Il7r)) and natural killer cells (killer cell lectin-like receptor subfamily G member 1 (Klrg1)) between cells isolated from young and old animals.

Lastly, we determined if lineage commitment occurs in naive and activated CD4$^+$ T cells. In our data we do not detect any early differentiation in naive and activated CD4$^+$ T cell subsets. In accordance with the literature we found *Gata3* but not Th2 cytokines expressed in the majority of cells [340]. Interestingly, the Th1-related genes *Tbx21* and *Ifng* were up-regulated, in an uncoordinated manner, in a small population of activated CD4$^+$ T cells of old animals. This is consistent with a known Th1 bias in CD4$^+$ T cell responses in old mice [341] and humans [342] **(Fig. 2.4H)**.

Furthermore, we did not detect any difference in TCR components/signalling and importantly, detected no signs of T cell exhaustion [343], especially in cells isolated from old animals (**Fig. 2.4I**). We also ruled out species-specific differences in commitment towards T helper cell lineages (**Fig. 2.4J-K**).

A

	G1	S	G2/M
No. of cells	485	17	11

B

	No. of cells	No. of clones		
		1	2	>3
B6 young	146	140	1	0
B6 old	196	185	2	0
CAST young	70	68	0	0
CAST old	128	84	0	0

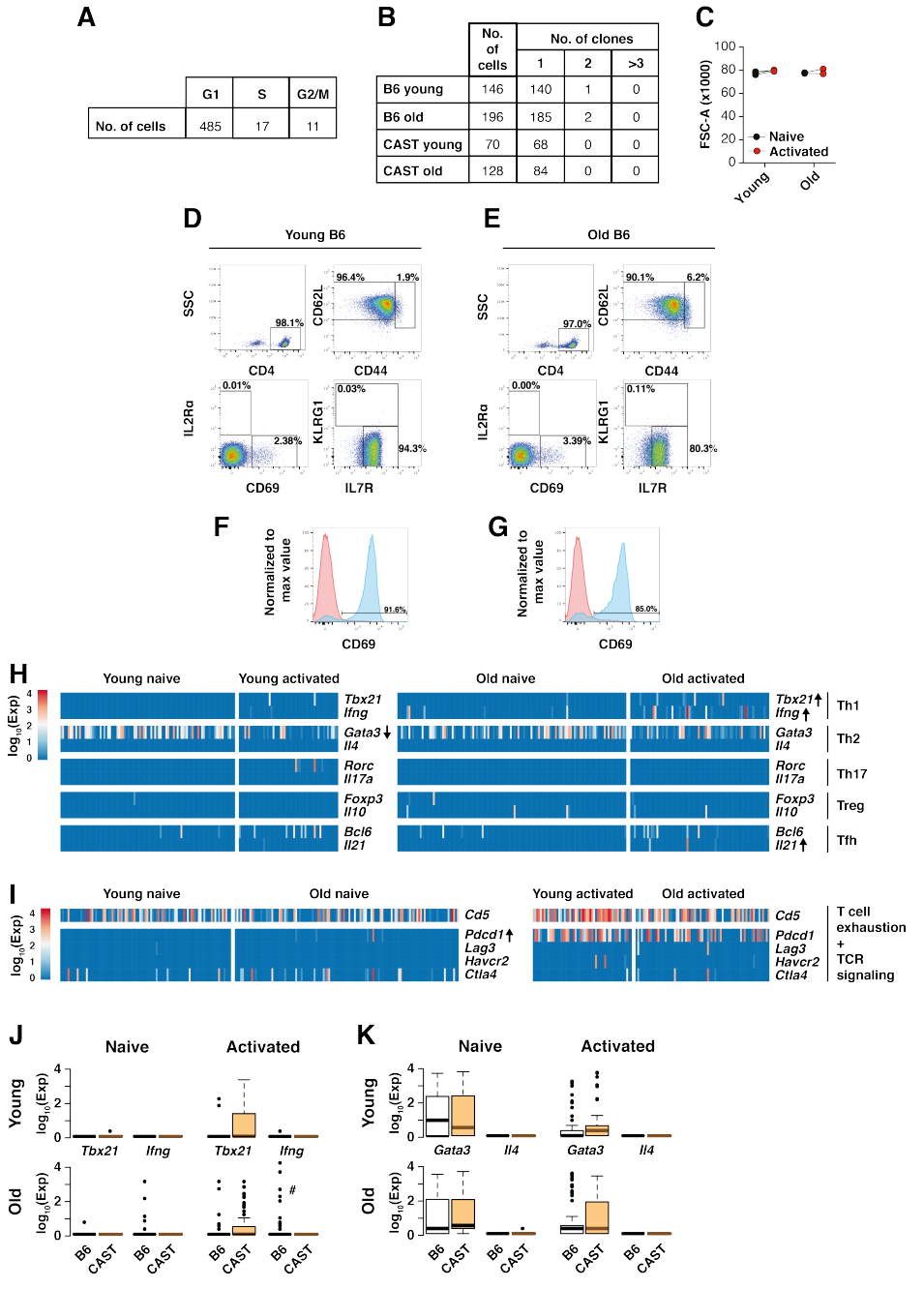


Fig. 2.4: Characterisation of isolated CD4+ T cells (Full legend on next page).

Fig. 2.4: Characterisation of isolated CD4$^+$ T cells (continued).
(A) *Cyclone* [338] was used to classify individual naive and activated CD4$^+$ T cells into the cell cycle phases G1, G2/M and S, **(B)** *TraCeR* [332] constructed T cell receptor sequences from scRNA-Seq data to analyze clonal diversity in naive and activated CD4$^+$ T cells, **(C)** Cell sizes were estimated for naive and activated CD4$^+$ T cells in young and old B6 animals measured by FSC using flow cytometry, **(D)-(E)** CD4$^+$ T cells were purified from spleens of young (D) and old (E) B6 animals and stained with antibodies against CD4, CD62L, CD44, CD69, IL2Rα (CD25), IL7R (CD127), and KLRG1 as well as viability dye. FACS plots shown are gated on single live cells (top left panel) and single live CD4$^+$ T cells (other panels), and percentages shown relate to total of gated cells, **(F)-(G)** Naive CD4$^+$ T cells were purified from spleens of five young (F) or two aged (G) B6 mice, and were either directly assayed or activated with plate-bound antibody against CD3ε and CD28 for 3 hours. Cells were stained with antibodies against CD4, CD69, and viability dye. Representative histograms for naive (red) and activated (blue) cells are shown, **(H)** Characterisation of possible differentiation processes leading to Th1, Th2, Th17, regulatory T cell (Treg) and Tfh cell lineages. For each lineage the major regulatory transcription factor (upper row) and an effector cytokine (lower row) is shown. Differential expression testing was performed between activated and naive cells from young B6 animals (left panel) and between activated and naive cells from old B6 animals (right panel). Upward arrow: up-regulation of expression (log2FC in $\mu_i > 2$, EFDR = 5%) after activation, Downward arrow: down-regulation of expression (log2FC in $\mu_i > 2$, EFDR = 5%) after activation, **(I)** Heatmap showing T cell exhaustion (*Pdcd1*, *Lag3*, *Havcr2*, *Ctla4*) and TCR activation markers (*Cd5*). Differential expression testing was performed in naive cells between young and old B6 animals (left panel) and in activated cells between young and old B6 animals (right panel). Upward arrow: up-regulation of expression (log2FC in $\mu_i > 2$, EFDR = 5%) during ageing, **(J)** Th1 lineage marker (*Tbx21*, *Ifng*) expression was compared between B6 and CAST in following conditions: naive cells from young animals (upper left panel), naive cells from old animals (lower left panel), activated cells from young animals (upper right panel), activated cells from old animals (lower right panel). #: statistically significant differential expression (log2FC in $\mu_i > 2$, EFDR = 5%), **(K)** Th2 lineage marker (*Gata3*, *Il4*) expression was compared between B6 and CAST in the following conditions: naive cells from young animals (upper left panel), naive cells from old animals (lower left panel), activated cells from young animals (upper right panel), activated cells from old animals (lower right panel). From [22]. Reprinted with permission from AAAS.

After the above analyses and the experimental characterisation, a total of 1514 high-quality CD4$^+$ T cell transcriptomes from young and old animals were analysed across all conditions (unstimulated and activated; naive, FACS-purified naive, FACS-purified EM) and species (B6 and CAST). An overview of all high-quality transcriptomes can be seen in **Fig. 2.5**. We detect that unstimulated and stimulated cells group together **(Fig. 2.5A)** while separation is also noticeable between species **(Fig. 2.5B)** and experimental methods **(Fig. 2.5C)**. As discussed below, cells from young and old animals do not separate when visualising the cells in form of a tSNE **(Fig. 2.5C)**.

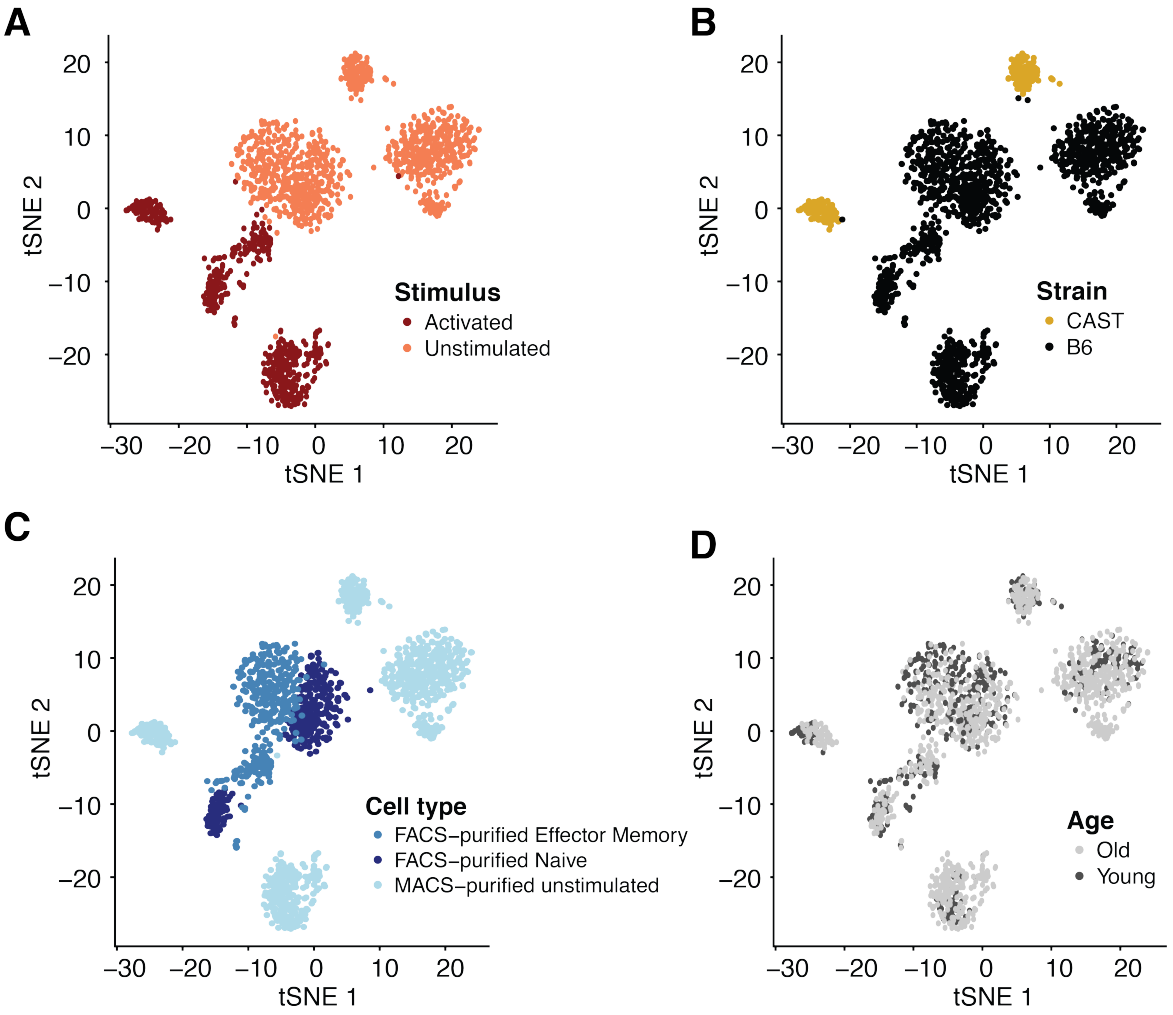


Fig. 2.5: Visualisation of all isolated CD4⁺ T cells.
tSNE dimensionality reduction of 1514 CD4⁺ T cells isolated from young and old mice of two related species. Cells were labelled based on **(A)** their activation state, **(B)** the mouse species, **(C)** experimental isolation approach and **(D)** age.

2.3 Species-specific gene expression in naive CD4+ T cells

To characterise the variation observed in **Fig. 2.5**, we first dissected differences in gene expression between the two mouse species using naive CD4+ T cells. We also assessed whether possible differences that are detected between the two species are driven by the assembly quality of the genome reference.

2.3.1 Avoiding transcript counting biases due to incorrect alignment

We used BASiCS [295] to detect differentially expressed genes as described in **Section 2.2.2**. In scRNA-Seq, technical noise is highest for lowly expressed genes [10] and we therefore excluded genes that had an average posterior mean expression < 50 in each population. Subsequently, we applied the differential expression test developed within BASiCS, using a threshold of log2FC in $\mu_i > 2$ with the EFDR controlled to 5%. We observed that 15% of expressed genes were differentially transcribed between CD4+ T cells of the two mouse species **(Fig. 2.6A)**.

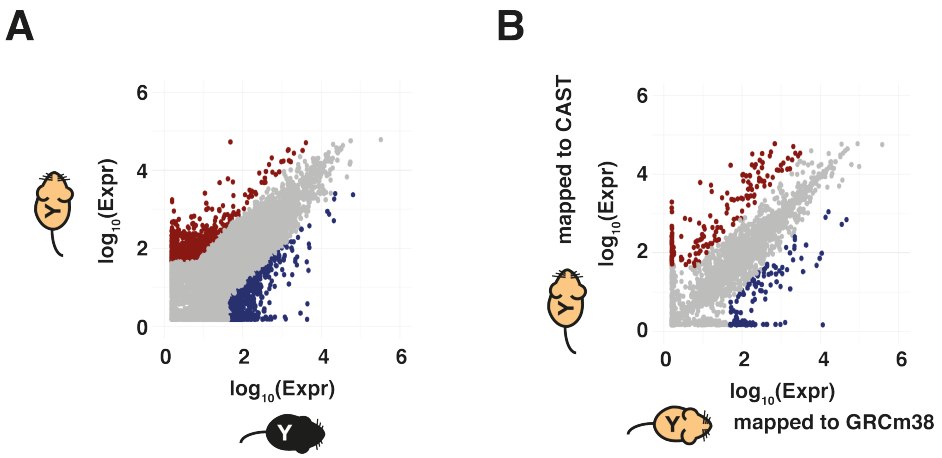


Fig. 2.6: Cross-mapping correction between divergent mouse species.
(A) Species-specific gene expression in B6 (blue) or CAST (red). Average gene expression using posterior estimation, threshold of means > 50, log2FC in $\mu_i > 2$, EFDR = 5%, **(B)** Mean, normalised transcript counts of mapped reads from CAST cells (young, naive) using the GRCm38 genome (x-axis) or the CAST genome (y-axis) as reference. Differentially mapped genes were removed from downstream analysis. Average gene expression using posterior estimation, threshold of means > 50, log2FC in $\mu_i > 2$, EFDR = 5%.

To rule out the possibility that these differences are driven by potential artefacts in the new *Mus musculus castaneus* genome assembly, we also mapped reads from young CAST samples to the GRCm38 genome. To estimate which differentially expressed genes may

arise due to errors in the new CAST genome assembly, we performed the same differential expression analysis on CAST samples by mapping these libraries onto both GRCm38 and CAST. Roughly 5% of all tested genes are detected as differentially expressed even though the samples being compared are identical and only mapped to different genomes **(Fig. 2.6B)**. Comparing this set of genes to the set of species-specific genes, we find that they make up 10% of differentially expressed genes between the two species. We performed a similar analysis for B6 samples. This approach allowed us to remove genes which show differences in expression from our analyses. This may be driven by the quality of the reference genome.

2.3.2 Transcriptional dynamics of species-specific genes

Similar for **Fig. 2.5**, we found that CD4$^+$ T cells cluster by species when only profiling naive cells **(Fig. 2.7A)**. As described above, theses differences are driven by the roughly 15% of differentially expressed genes. To assess the functional role of the species-specific genes that are not due to biases in read alignment, we qualitatively and quantitatively compared their expression across individual cells in both species. Firstly, species-specific genes were only expressed in subsets of the full population of naive CD4$^+$ T cells **(Fig. 2.7B)**. Furthermore, we used DAVID [344] to test for gene ontology (GO) enrichment in differentially expressed gene sets. In line with genes being only sporadically expressed across the full population of cells, we did not detect functional enrichment in either B6 or CAST specific genes **(Fig. 2.7C)**. Profiling individual cells using scRNA-Seq allows us to detect different patterns of expression for species-specific genes. Within the set of species-specifically expressed genes, we detect some that display low variability and some with high variability **(Fig. 2.7D)**. More quantitatively, when profiling expression variability, we detect species-specifically transcribed genes to be generally more variable on a cell-to-cell basis than genes expressed in both species **(Fig. 2.7E)**. These findings hint that the detected divergence in genes expression might be caused by neutral drift without functional support of the species-specific genes. We therefore argue that profiling transcriptional variability in a homogeneous population of cells is a measure for cell population function such as cell response to stimuli.

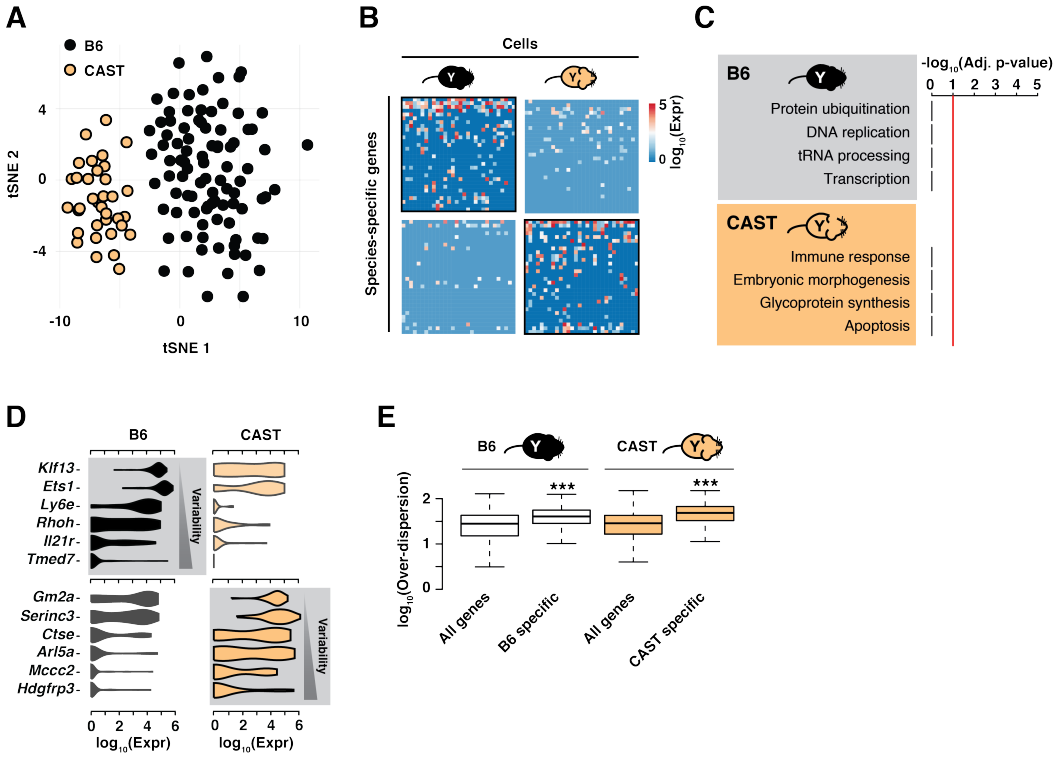


Fig. 2.7: Species-specific gene expression in naive CD4+ T cells.
(A) tSNE dimensionality reduction of scRNA-Seq data reveals species-specific clustering of naive CD4+ T cells from young animals, **(B)** Representative heatmap of 30 genes and 30 cells randomly selected from all species-specifically transcribed genes from young animals shows typical species-specific variations, **(C)** GO analysis of species-specific genes. Bonferroni corrected p-values (adjusted p-values) were used to visualise GO enrichment. The statistical significance threshold was set to adjusted p-value = 0.1 (red line), **(D)** Cell-to-cell variability in gene expression levels. Violin plots show distribution of single-cell expression of selected species-specifically transcribed genes (in grey background), ranked from lowest (top) to highest variability (bottom), **(E)** $\log_{10}$-transformed variability estimates of species-specific genes were compared to variability estimates of all genes expressed in B6 (left) and CAST (right). Mann-Whitney-Wilcoxon test; ***: $p<10^{-10}$

.

2.4 Expression dynamics during CD4$^+$ T cell activation

Functional CD4$^+$ T cell transcriptional responses start with an early, targeted activation of translational machinery and cytokine networks, followed by large-scale transcriptional changes associated with lineage commitment [330, 345]. To characterise the immediate early activation programme, we stimulated naive CD4$^+$ T cells of B6 animals for three hours with plate-bound anti-CD3ε/anti-CD28 antibodies, thus inducing a strong and uniform activation mimicking initial contact with an antigen-presenting cell. Importantly, we did not use additional cytokines to commit the naive CD4$^+$ T cells to specific helper cell lineages [333]. This was confirmed empirically by the analysis presented in **Section 2.2.3**.

2.4.1 Mean expression changes during immune activation

By visualising the dimensionality reduced transcriptional profiles of naive and activated CD4$^+$ T cells, we qualitatively observe strong transcriptional changes during immune activation **(Fig. 2.8A)**. Differential mean expression testing identifies thousands of genes as differentially expressed in CD4$^+$ T cells upon activation (2063 genes, log2FC in $\mu_i > 2$, EFDR = 5%) **(Fig. 2.8B)**. We sought to investigate the behaviour of up- and down-regulated genes across the population of naive or activated cells. Initially, we calculated, for each gene, the percentage of cells in which it was expressed (> 0 transcript counts). Genes whose expression is down-regulated after activation are expressed in a median of 18% naive CD4$^+$ T cells isolated from B6, while genes that are up-regulated are expressed in a median of 36% activated cells. Before activation, up-regulated genes are only expressed in a median of 5% naive cells while after activation, down-regulated genes are expressed in a median of 4% activated cells **(Fig. 2.8C)**. This analysis suggests that the immune response genes are similarly up-regulated across all cells, while genes that are down-regulated upon immune activation are more sporadically expressed in naive cells.

Furthermore, we performed GO analysis on genes either up- or down-regulated **(Fig. 2.8D)**. Among the down-regulated genes are components of the intra-cellular signalling network while genes that are up-regulated are mostly part of the translation machinery [346]. Furthermore, the transcriptional switch driven by TCR engagement and co-stimulation included classic markers of activation, including (*Il2ra*, **Fig. 2.8E**) and chemokine (C-C motif) ligand 4 (*Ccl4*) [345]. In contrast, we observed the coordinated suppression of *Sell* (*Cd62l*, **Fig. 2.8F**), as expected after activation [347].

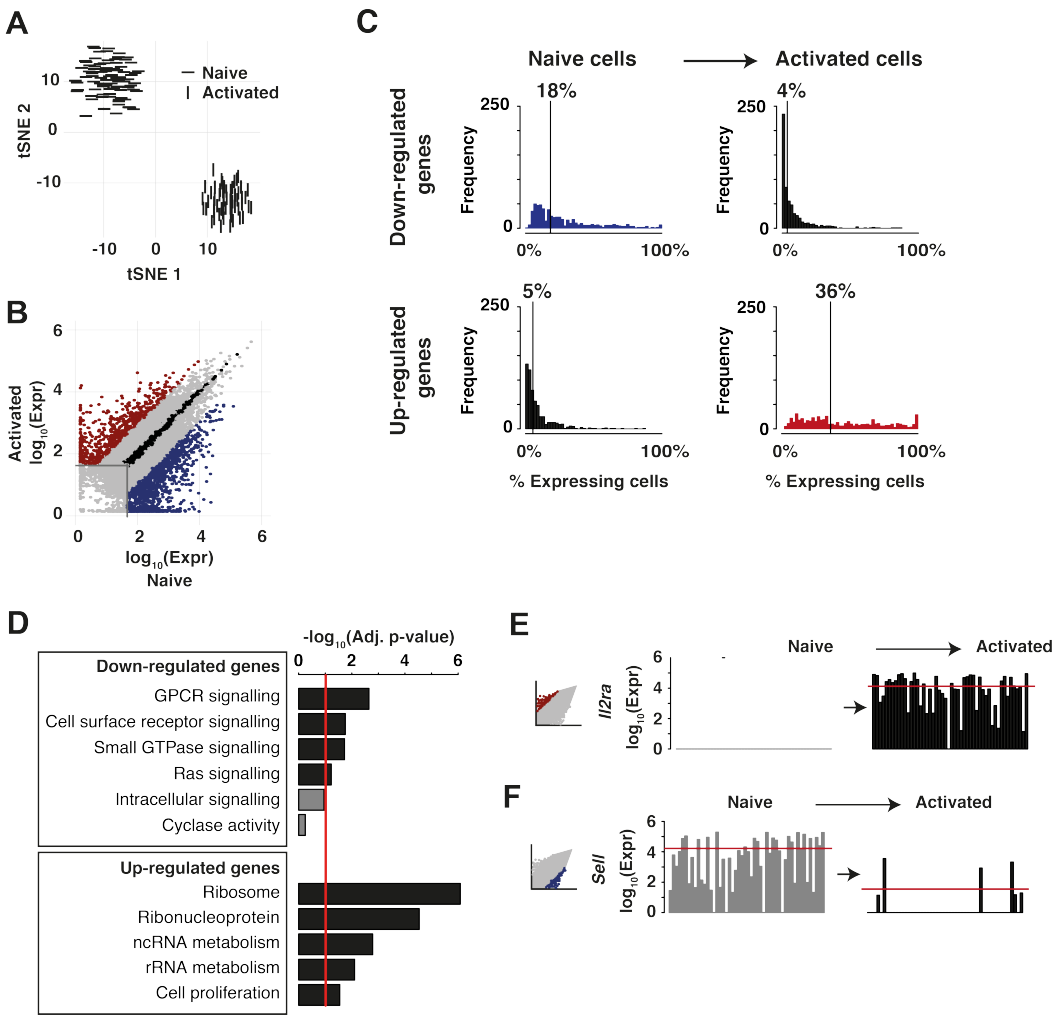


Fig. 2.8: Mean expression dynamics upon CD4+ T cell activation.
(A) Activation of CD4+ T cells from young B6 mice induces large-scale transcriptional changes which is visualised using tSNE dimensionality reduction, **(B)** Genes up-regulated (red) and down-regulated (blue) by immune stimulation in young B6 mice. Non-differentially expressed genes shown in black. Average gene expression using posterior estimation, threshold of means > 50, log2FC in $\mu_i > 2$, EFDR = 5%, **(C)** Fractions of cells in which down- or up-regulated genes are expressed (600 genes were randomly selected, histograms with 50 bins, median value is indicated). Upper panels: fraction of naive (left) and activated (right) cells in which down-regulated genes are expressed. Lower panels: fraction of naive (left) and activated (right) cells in which up-regulated genes are expressed, **(D)** Bar plots of functional gene categories enriched in up- and down-regulated genes during activation of CD4+ T cells in B6 (Bonferroni multiple testing corrected p-values, red line marking 0.1), **(E)** Example genes that represent transcriptional changes upon activation of CD4+ T cells: *Il2ra* (CD25) is highly and consistently up-regulated after stimulation, **(F)** *Cd62l* (*Sell*) is more stochastically expressed in naive CD4+ T cells, and is down-regulated upon activation.

2.4.2 Changes of expression variability during immune activation

We next profiled changes in expression variability upon immune activation. Due to the strong confounding effect observed between mean expression and variability estimates, we only profiled genes that show no changes in mean expression between the naive and activated state (see **Section 2.2.2**, indicated as black dots in **Fig. 2.9A**, log2FC in $\mu_i = 0$, EFDR = 5%). When comparing posterior medians of the over-dispersion parameter δ_i for genes that remain stable in mean expression, we observed a significant reduction in cell-to-cell transcriptional variability (Mann-Whitney-Wilcoxon test, $p<10^{-10}$) **(Fig. 2.9B)**. For example, the eukaryotic translation initiation factor 1 (*Eif1*), show a marked decrease in cell-to-cell transcriptional variability, consistent with increased regulatory coordination **(Fig. 2.9C)**.

We next investigated whether genes that are more variably expressed in the naive than the activated condition showed coordinated patterns of expression, which are potentially associated with cryptic substructure. For this, we identified genes with statistically higher expression variability in the naive population compared to activated cells (log2FC in $\delta_i >$ 0.4, EFDR = 5%, no change in mean expression). Genes with high variability and high pairwise correlation (Spearman's $\rho > 0.8$) in the naive population can be used to identify possible sub-populations of CD4$^+$ T cells. A hierarchical clustering analysis did not show any signs of such substructure **(Fig. 2.9D)**. This analysis therefore indicates that the collapse in variability is cause by a genuine shift from stochastic to regulated expression between two homogeneous populations of cells.

Finally, it has been observed that covariance between cells due to unobserved factors such as the cell cycle can mask potentially interesting biological signals [348, 13]. In our dataset, ribosomal biogenesis is the strongest mediator of CD4$^+$ T cell function upon activation which is strongly and homogeneously expressed across all cells **(Fig. 2.8D and Fig. 2.9C)**. We therefore regressed out this factor using the single-cell latent variable model (scLVM) [13]) to uncover underlying variance in activated CD4$^+$ T cells. Importantly, performing PCA on the uncorrected and corrected counts after regression did not reveal concealed cellular processes **(Fig. 2.9E)**.

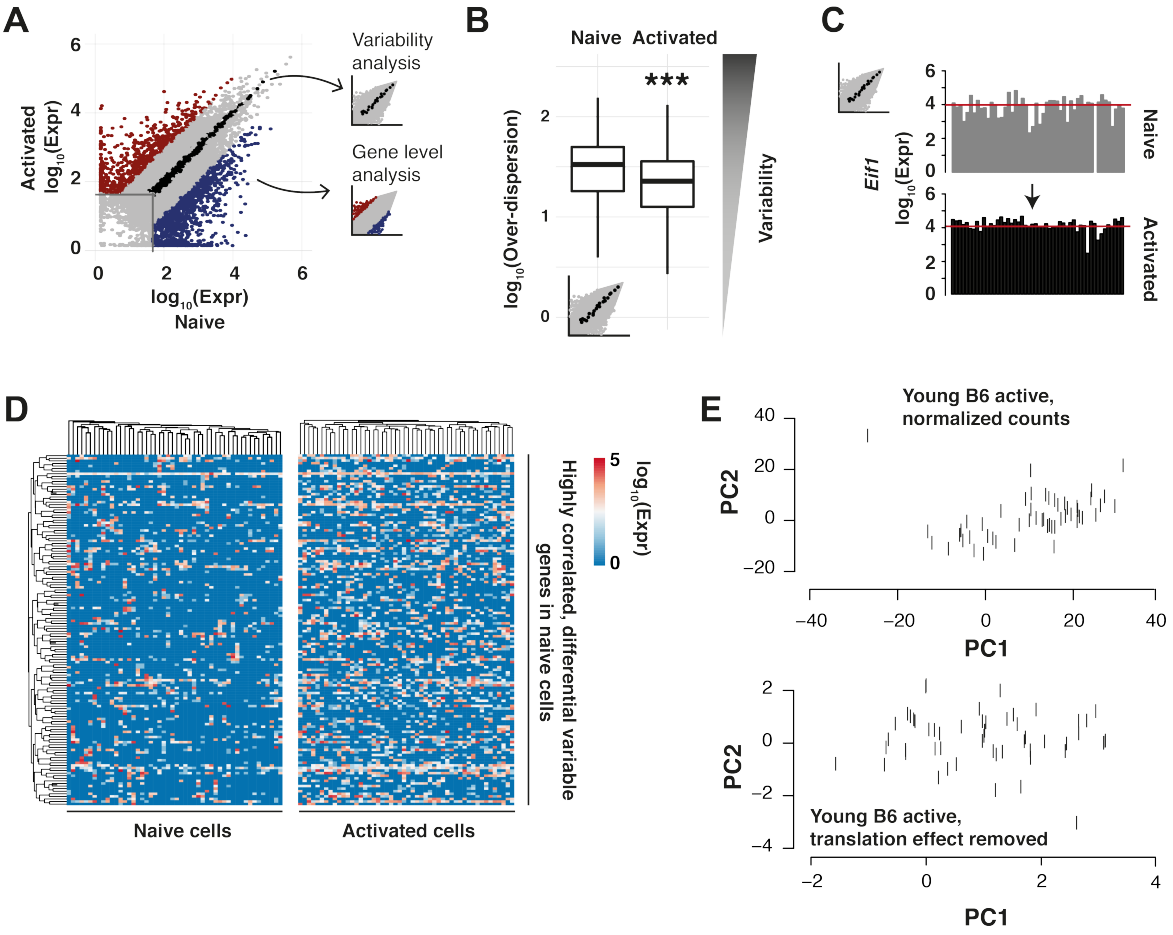


Fig. 2.9: Changes in transcriptional variability upon immune activation.
(A) Genes up-regulated (red) and down-regulated (blue) by immune stimulation in young B6 mice (log2FC in $\mu_i > 2$, EFDR = 5%). Non-differentially expressed genes shown in black (log2FC in $\mu_i = 0$, EFDR = 5%). Average gene expression using posterior estimation, threshold of means > 50, **(B)** Genes with no overall gene expression differences during activation (black dots in (A)) show decreased cell-to-cell variability in transcription (Mann-Whitney-Wilcoxon test, ***: $p<10^{-10}$), **(C)** *Eif1* is expressed in most cells in both conditions at similar levels, but shifts from stochastic to regulated expression, **(D)** To detect possible sub-populations in naive or activated CD4+ T cells, differentially variable genes (log2FC in $\delta_i > 0.4$, EFDR = 5%) in naive cells with high gene-to-gene correlation (Spearman's $\rho > 0.8$) were used for hierarchical cluster analysis, **(E)** Upper panel: PCA of normalised counts of activated CD4+ T cells from young B6 animals. Lower panel: PCA of activated CD4+ T cells of young B6 animals after removing differences in the translation programme as a confounding factor.

2.4.3 Response-related transcriptional dynamics in CAST

As described above, activation of CD4+ T cells drives a transcriptional switch that alters the global expression profile from a stochastic to a tightly regulated state. To test whether these transcriptional dynamics are evolutionarily conserved, we performed the same analysis for CD4+ T cells extracted from CAST. Similar to cells isolated from young B6, we detect (i) clustering based on activation state **(Fig. 2.10A)**, (ii) thousands of genes being differentially expressed **(Fig. 2.10B)**, (iii) a decrease in expression variability after immune activation for genes that are stable in mean expression **(Fig. 2.10C)** and (iv) the expression of up-regulated genes in a higher number of activated cells than down-regulated genes in naive cells **(Fig. 2.10D)**.

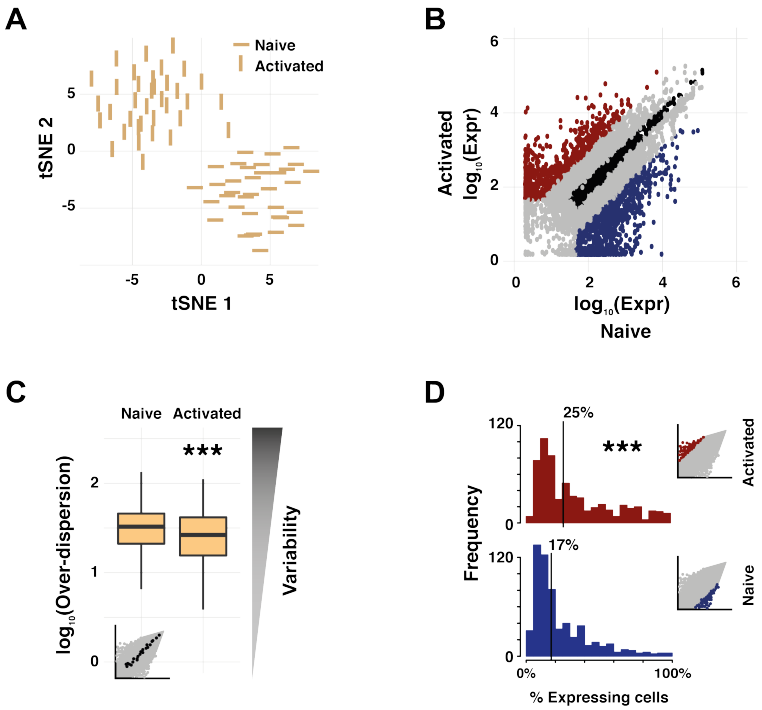


Fig. 2.10: Immune activation dynamics in young CAST animals.
(A) Activation of CD4+ T cells from young CAST mice in anti-CD3ε/CD28 coated plates induces large-scale transcriptional changes, visualised using tSNE dimensionality reduction, **(B)** Genes up-regulated (red) and down-regulated (blue) by immune stimulation in young CAST mice (log2FC in $\mu_i > 2$, EFDR = 5%). Non-differentially expressed genes used in (C) are shown in black (log2FC in $\mu_i = 2$, EFDR = 5%). Average gene expression using posterior estimation, threshold of means > 5, **(C)** Genes with no overall gene expression differences during activation show decreased cell-to-cell variability in transcription (Mann-Whitney-Wilcoxon test, ***: $p<10^{-10}$), **(D)** Up-regulated genes were expressed in a relatively large fraction of activated CD4+ T cells after stimulation (median 25%). Down-regulated genes were expressed in a smaller fraction of naive CD4+ T cells (median 17%). 600 genes of each condition were randomly selected. From [22]. Reprinted with permission from AAAS.

2.5 Conservation of the core activation process

As shown above, we detect similar activation patterns when comparing $CD4^+$ T cell activation between B6 and CAST. To further study the conservation of this immune response, we used the rapid divergence in gene expression between $CD4^+$ T cells from both species to refine the functional set of immune response genes activated upon immune stimulation [330]. Conserved functionality is assumed when genes are similarly up-regulated upon activation in both species. We hypothesise that such targets would be both conserved between species and expressed in most cells, whereas the genes activated species-specifically would be less likely to be functional targets and more likely to be sporadically expressed.

2.5.1 Detecting evolutionarily conserved response genes

As described for B6 above, we stimulated naive $CD4^+$ T cells isolated from young CAST males using anti-CD3ε/anti-CD28 antibodies followed by scRNA-Seq. As expected, cells clustered based on their activation state and species of origin **(Fig. 2.11A)**. To find genes that form the evolutionarily conserved, core activation programme, we test for differential expression between the naive and activated state separately in B6 and CAST (log2FC in μ_i > 2, EFDR = 5%). Genes that are up-regulated after activation similarly in B6 and CAST form the shared activation programme. Up-regulated genes that are differentially expressed in activated cells between the two species (log2FC in μ_i > 2, EFDR = 5%) represent the species-specific response genes.

We next ensured that the species-specific differences in transcriptional response are not caused by mapping artefacts between the different genome builds. For this, we quantified gene expression in activated CAST and B6 samples based on the CAST and GRCm38 genome as described above. Genes in the activated state that show differential mapping between the two genomes were excluded from the species-specific lists of response genes. After removing those, we find 1208 genes to be up-regulated across both species. Out of these, we detect 225 genes that are (i) strongly up-regulated upon activation and (ii) up-regulated similarly in B6 and CAST. The latter means that these genes are not detected as differentially expressed in activated cells between the two species. Out of all 1208 response genes, 171 are detected as differentially expressed between the two species forming the set of species-specific response genes **(Fig. 2.11B)**.

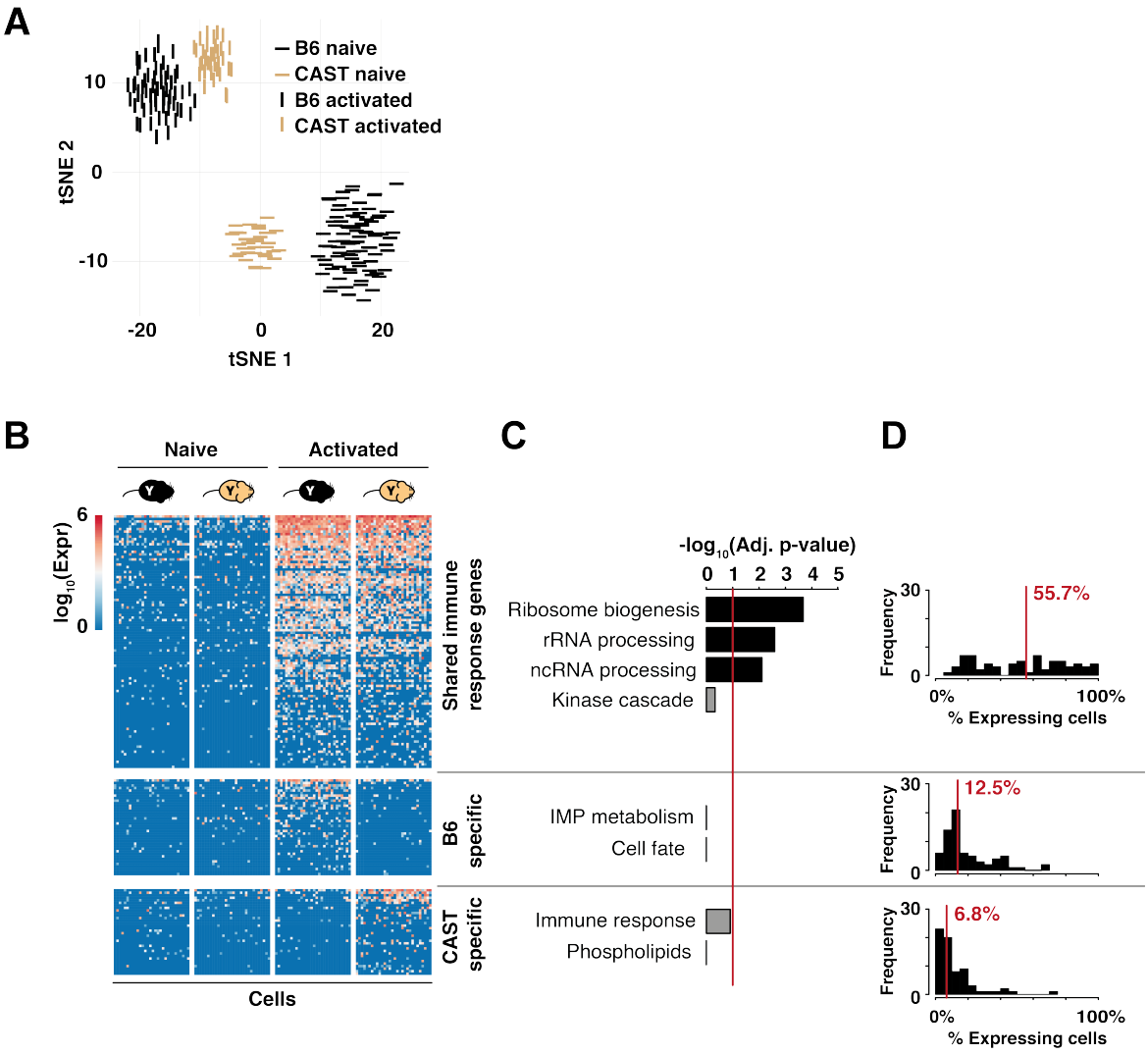


Fig. 2.11: Shared CD4+ T cell activation programme.
(A) CD4+ T cells isolated from B6 (black) and CAST (orange) show similar large-scale transcriptional changes upon immune stimulation, **(B)** Immune activation of CD4+ T cells triggers up-regulation of both conserved (upper panel), and species-specific (lower panels) transcriptional programmes. For visualisation purposes, genes in shared and species-specific categories were proportionately and randomly selected. 30 cells were randomly selected for each condition/species, **(C)** Genes up-regulated in both B6 and CAST highly enrich for known T cell functionality. Genes up-regulated in only B6 or CAST have no statistically significant functional enrichment (Bonferroni multiple testing corrected p-values, red line is 0.1), **(D)** Fractions of cells in which a gene is detected are displayed as histograms. 70 genes were randomly selected from each gene set. While most genes in the shared activation process are expressed in a high percentage of cells, only few cells express species-specific genes of the activation process. From [22]. Reprinted with permission from AAAS.

2.5.2 Functional assessment of the conserved response genes

We next estimated functionality of these 225 shared response genes by enrichment and variability analysis. Firstly, when performing GO analysis, the set of shared genes was strongly enriched for cellular processes known to be immediately activated by stimulation of $CD4^+$ T cells. This includes the core translational machinery (ribosome biogenesis and ribosomal RNAs (rRNAs), **Fig. 2.11C**) and key immune activation genes (such as *Il2ra* and *Tnfrsf9*, [345]). Other categories of immune response genes include cytokines, chemokines and their receptors (e.g.*Ccr8*, *Il2*, *Ccl3*, [349]), members of the nuclear receptor (Nr) superfamily (e.g. *Nr4a2*, *Nr4a3*, *Nr4a1*, [350]), components of NFκB signalling (e.g. *Nfkbid*, *Nfkb1*, *Nfkbie*, *Rel*, [351]) and tumour necrosis factor (Tnf) signalling (e.g. *Tnf*, *Tnfsf14*, *Tnfrsf4*, *Tnfrsf1b*, [352]). In contrast, species-specific genes (96 for B6, 75 for CAST) showed no enrichment for biological function **(Fig. 2.11C)**.

As described in **Section 2.3.2**, the degree of heterogeneity to which genes are expressed within a homogeneous population of cells can indicate the functional relevance of these genes for population responses. We therefore calculated the fraction of cells that express shared and species-specific immune response genes (> 0 counts). Shared immune response genes were expressed across most $CD4^+$ T cells in both B6 and CAST after activation. In contrast, species-specific response genes tend to be expressed in a smaller fraction of cells **(Fig. 2.11D)**.

Our interspecies comparison thus revealed that target genes involved in translational control and immune function represent the conserved signature within the early activation response. These genes are furthermore similarly up-regulated in most cells across the homogeneous population of activated $CD4^+$ T cells.

2.6 Destabilisation of CD4$^+$ T cell activation during ageing

Ageing can cause perturbation of cell cycle entry for haematopoietic stem cells, leading to a shift in the functional balance between self-renewal and differentiation [323]. We considered whether ageing might similarly perturb the transcriptional response of CD4$^+$ T cells to immune stimulation. For this, we performed differential expression and differential variability analysis between cells isolated from young and old animals. By comparing the activation responses between different sub-species of mice, we could also establish whether any observed impact of ageing is conserved. Lastly, we compare the effect of ageing across different subsets of CD4$^+$ T cells to assess how the effects of ageing differ across the immune system.

2.6.1 Ageing does not effect CD4$^+$ T cell transcription on a global level

We first asked whether the overall response of CD4$^+$ T cells is perturbed during ageing. For this we (i) performed PCA and (ii) compared mean expression of cells isolated from young and old mice. PCA revealed that the global expression profiles of naive or activated CD4$^+$ T cells not heavily effected by ageing **(Fig. 2.12A and B)**. Furthermore, we identified differentially expressed genes between young and old animals separately for naive and activated CD4$^+$ T cells **(Fig. 2.12C and D)**. This analysis was performed separately for each species. Only around 10% of all tested genes showed changes in mean expression between young and old animals. Additionally, these genes typically showed low expression in naive or activated cells taken from both young and old animals. To further quantify this, we computed the fraction of cells in which each differentially expressed gene was expressed. The distribution of these values was added as inlets to the plots in **Fig. 2.12C and D** (x-axis ranging from 0% to 100% of cells). We detected that differentially expressed genes are enriched for those that are only expressed in subsets of cells. This indicates that changes in mean expression during ageing only affect lowly expressed genes that are detected only in subsets of cells and therefore do not contain functionally relevant genes. These effects can also arise due to increased levels of noise for lowly expressed genes [10].

Nevertheless, to test whether these subtle effects during ageing are shared between the two species, we calculated the Jaccard index, which measures the overlap between sets of elements, separately for up- and down-regulated genes **(Fig. 2.12E and F)**. We only detect $\sim$3% of the differentially expressed genes as being shared between the two species either

for up-regulated genes or down-regulated genes during ageing. Therefore, we did not find a conserved ageing signature that affects expression levels in CD4+ T cells.

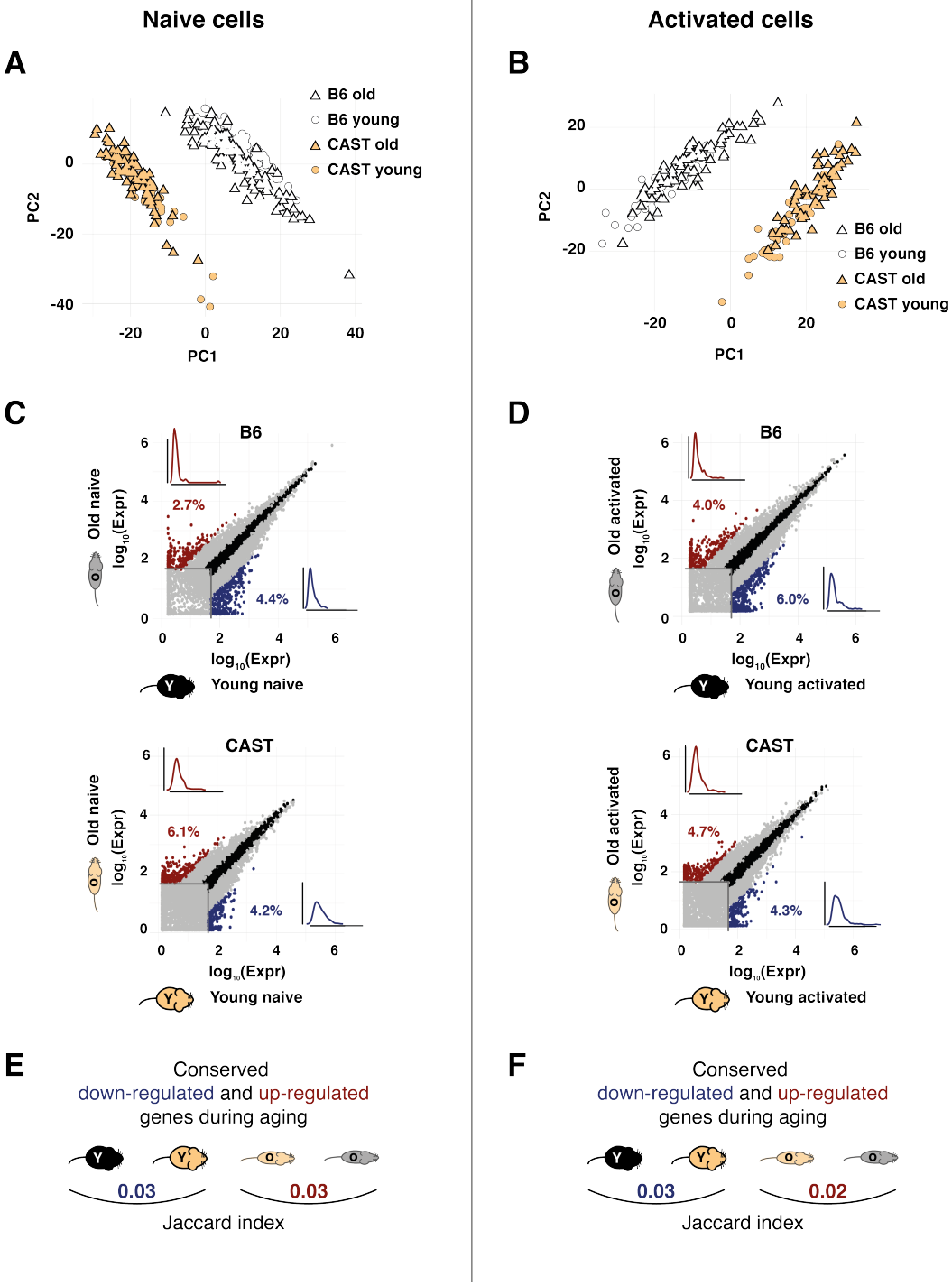


Fig. 2.12: Global immune response during ageing (full legend on next page).

Fig. 2.12: Global immune response during ageing (continued).
PCA reveals no separation between cells isolated from young or old animals in the naive **(A)** or activated **(B)** state, **(C)** 7.1% of all tested genes in B6 and 10.3% of genes in CAST are differentially expressed in naive CD4$^+$ T cells between old (red) and young (blue) animals. Average gene expression using posterior estimation, threshold of means > 50, log2FC in $\mu_i > 2$, EFDR = 5%. Insets show distributions of fraction of cells in which these genes are expressed. X-axis: 0% - 100% of cells, **(D)** 10% of all tested genes in B6 and 9% of genes in CAST are differentially expressed in activated CD4$^+$ T cells between old (red) and young (blue) animals. Average gene expression using posterior estimation, threshold of means > 50, log2FC in $\mu_i > 2$, EFDR = 5%. Insets show distributions of fraction of cells in which these genes are expressed. X-axis: 0% - 100% of cells, **(E)-(F)** The overlap of ageing-associated genes in **(E)** naive or **(F)** activated cells was calculated using the Jaccard index between gene sets. Genes highly expressed in old animals (red) or genes highly expressed in young animals (blue) show little overlap (2-3%) between B6 and CAST. From [22]. Reprinted with permission from AAAS.

2.6.2 Ageing increases transcriptional variability in response genes

To further dissect possible ageing effects on the core functionality of CD4$^+$ T cells, we next focused on the conserved activation programme. This analysis is more targeted and allows us to detect subtle changes caused by ageing. Qualitatively, and consistent with the findings in **Section 2.6.1**, the majority of genes in the core activation programme responded upon stimulation, irrespective of age **(Fig. 2.13A)**. We then profiled changes in mean expression of activated cells between young and old animals for both species. This analysis resulted in a subtle decrease in expression for aged individuals while the majority of genes showed similar expression between young and old, as expected **(Fig. 2.13B)**.

We next profiled changes in variability by considering genes with no changes in mean expression in activated cells between between young and old animals (see **Section 2.2.2**). By plotting the log2FC in δ_i for these genes, we observe an increase in cell-to-cell transcriptional variability of the core activation programme in older animals compared to young animals **(Fig. 2.13C)**. To identify the drivers of the increase in transcriptional variability of the immune response during ageing, we calculated the fraction of activated cells in which genes of the shared activation programme are expressed. By comparing these fractions between activated cells isolated from old or young animals in both species, we identify consistently fewer cells from aged animals that express the shared activation programme **(Fig. 2.13D)**.

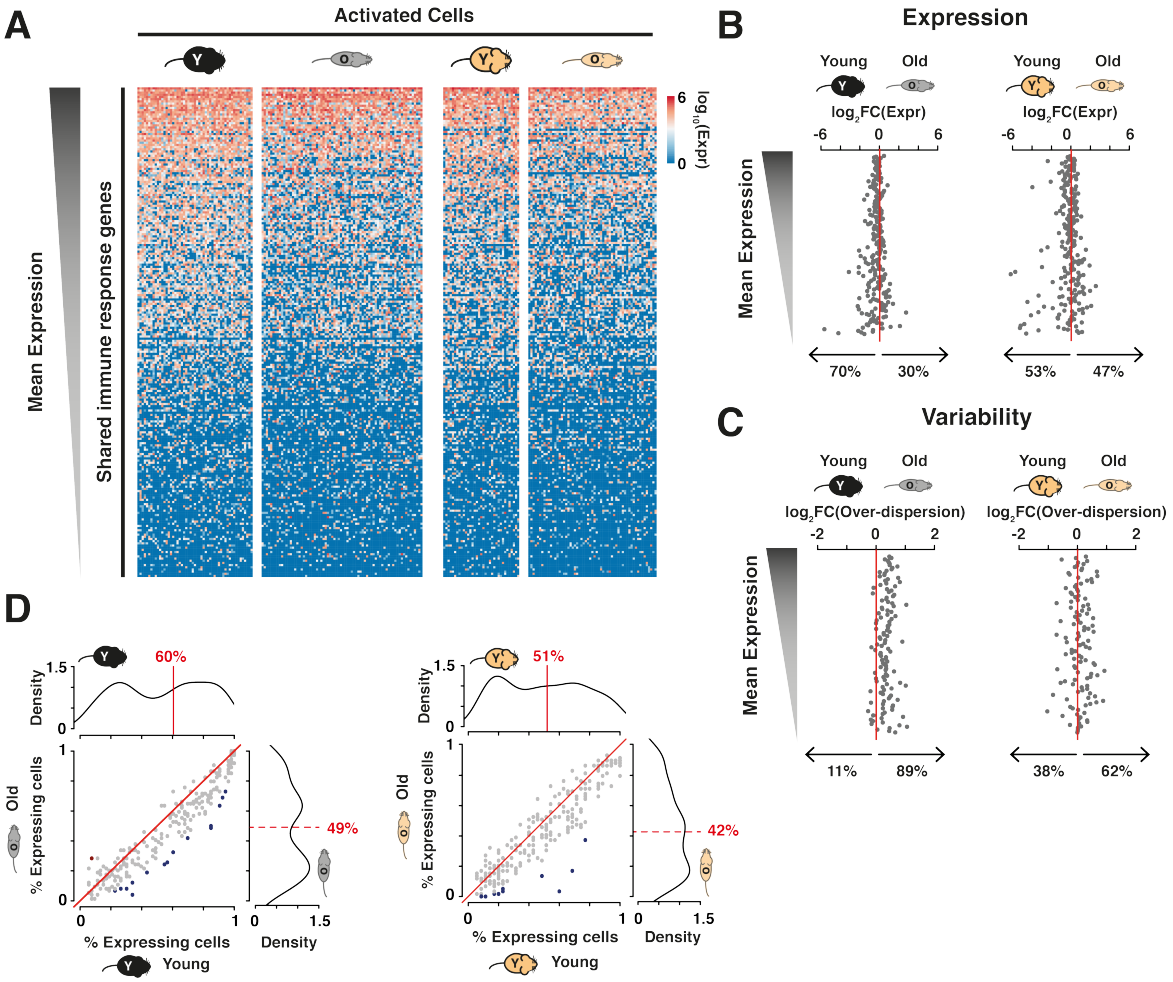


Fig. 2.13: Ageing destabilises the CD4+ T cell response.
(A) Full heatmap showing all 225 genes of the shared activation programme expressed in all activated cells from young and old CAST and B6. Genes were ordered based on their mean expression, **(B)** Fold changes in mean expression indicate a consistent trend in lower expression of the shared activation programme in cells from old animals (genes were ordered by mean expression, log2FC of posterior mean estimates), **(C)** Cells from old animals show higher transcriptional variability compared to young animals (genes were ordered by mean expression, log2FC of posterior over-dispersion estimates), **(D)** The fraction of cells in which genes of the shared activation process are expressed is reduced in activated CD4+ T cells from old animals. The distribution of fraction values is plotted on each corresponding axis (medians of fraction values are indicated in red); statistically significant changes in the percentage of cells expressing genes of the core activation process were assessed using a binomial test (blue points indicate bonferroni corrected p-values < 0.1). Gene expression in activated cells isolated from old animals was used as the Null-distribution.

These results indicate a destabilisation of the immune response programme during ageing. While most response genes are expressed at similar levels in activated cells of young and old animals, we detect a subset of cells where the expression of these genes is lost in aged mice. These dropouts in expression do not correlate across the population of cells and appear to be more stochastic.

2.6.3 Validation experiments to confirm changes in variability

We next asked whether the increase in variability is driven by (i) technical factors, (ii) biases in model parameter estimation or (iii) hidden sub-structure within the data. To address the first point, we generated independent replicates of naive and activated CD4$^+$ T cells from young and old B6 animals using Fluidigm C1 machines located at a different research institute. Profiling changes in variability using these biological replicates, we validated the increase in transcriptional variability during ageing **(Fig. 2.14A)**.

Secondly, quantification of transcriptional variability can be biased based on the number of cells present in each cell population. When assaying homogeneous cell populations with larger sample size, model parameter estimation is more precise compared to populations with smaller sample size (see **Section 3.4**). We therefore downsampled both young and old activated CD4$^+$ T cells to equal size and detected the same increase in variability during ageing **(Fig. 2.14B)** as previously when comparing the full set of CD4$^+$ T cells **(Fig. 2.13C)**.

Thirdly, in **Section 2.2.3** we observed that old animals have a small population of CD4$^+$ T cells with slightly elevated CD44 levels, reduced CD62L expression, and attenuated activation dynamics **(Fig. 2.4E-G)**. We therefore tested whether the global shift in variability is caused by different cell population structures between old and young B6 animals. To that end, cells expressing marker genes that were inconsistent with their activation state as well as cells with a possible Th1 differentiation bias (*Ifng* expressing) were removed. Based on the library size adjusted counts, we removed activated cells with low *Cd69* ($<$ 300 counts), high *Sell* ($>$ 10 counts), low *Trac* ($<$ 100 counts), low *Il2ra* ($<$ 100 counts) and *Ifng* ($>$ 0 counts) expression **(Fig. 2.14C)**. The remaining 37 and 26 activated cells in young and old B6 animals showed the same shift in transcriptional variability compared to the non-filtered data **(Fig. 2.14D)**.

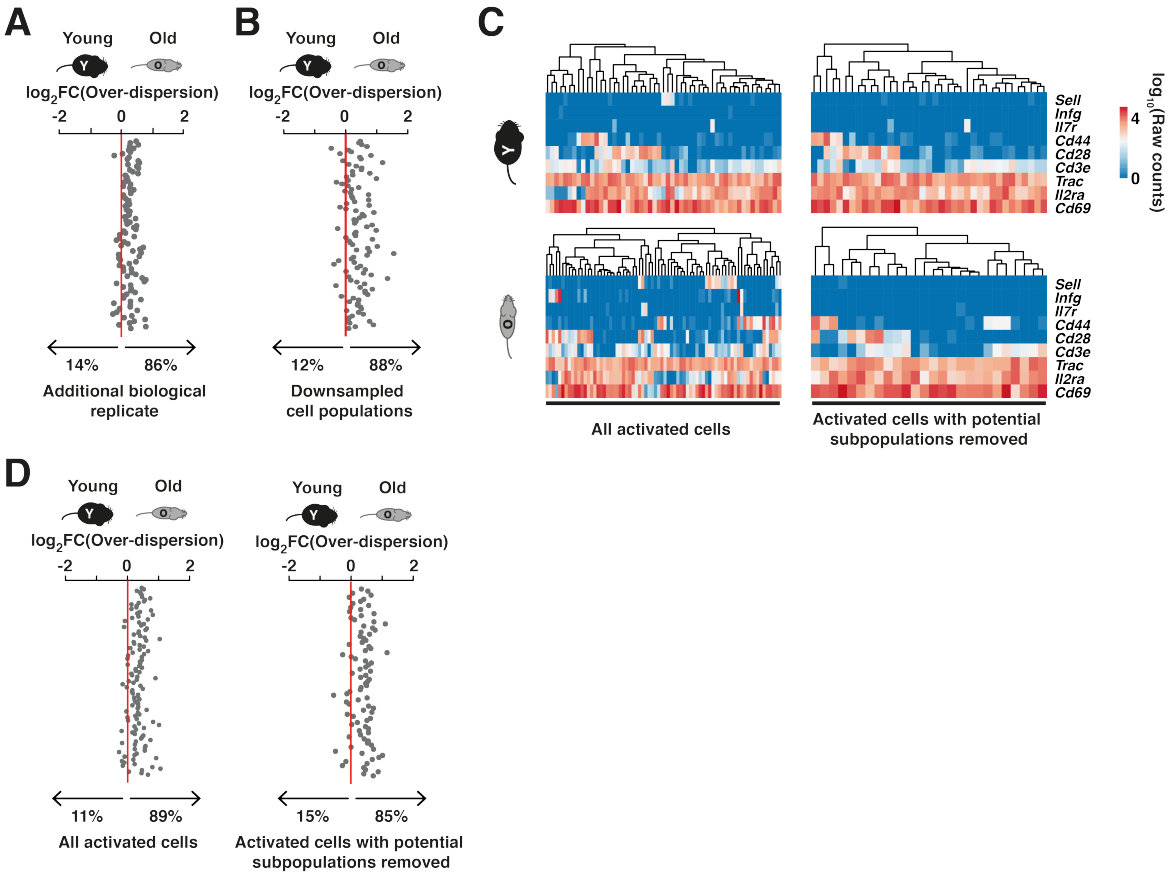


Fig. 2.14: Experimental validation of increased transcriptional variability during ageing. **(A)** A biological replicate of 115 activated CD4+ T cells from old B6 animals was generated and changes in gene expression variability were compared to activated CD4+ T cells from young B6 animals, **(B)** 30 out of all activated CD4+ T cells from young or old B6 mice were randomly selected and changes in gene expression variability were compared between the downsampled populations of cells, **(C)** Expression of CD4+ T cell activation markers in all activated cells from young and old B6 before (left) and after (right) filtering based on *Ifng*, *Sell*, *Trac*, *Il2ra* and *Cd69* expression, **(D)** Changes in gene expression variability of activated cells between young and old animals are displayed before (left) and after (right) filtering (see (C)).

2.6.4 Transcriptional variability in CD4$^+$ T cell subsets

It is well known that T cell type composition changes are driven by thymic involution which leads to a reduction of the thymic output in CD4$^+$ T cells over age. After infections, previously activated CD4$^+$ T cells survive and form central memory T cells [353].

We first address if the decrease in thymic output over age induces a reduction of recent thymic emigrants (RTEs) in the spleen and therefore biases expression variability to be higher in aged animals. Maturation of RTEs contributes significantly to the maintenance of the naive CD4$^+$ T cell pool in the periphery and is affected by ageing [354–356]. To estimate proportions of RTEs within the naive CD4$^+$ T cell pool, we characterised CD4 single positive (SP) thymocytes and splenic naive CD4$^+$ T cells by flow cytometry. RTEs can be identified by their CD24^{hi} Qa2^{lo} phenotype [354, 355]. We find the majority of CD4 SP thymocytes showed a CD24^{hi} Qa2^{lo} phenotype **(Fig. 2.15A)**. In contrast, we detected only a very small population of RTEs (2%) within splenic naive CD4$^+$ T cell pool similarly in young and old mice **(Fig. 2.15A and B)** [355]. This suggests minimal contamination of RTEs in naive CD4$^+$ T cells purified by MACS and therefore no bias when testing for changes in variability.

Next, we examine whether the age-mediated increase in cell-to-cell variability is conserved across different subsets of CD4$^+$ T cells. We therefore sorted naive and EM CD4$^+$ T cells as explained in **Fig. 2.2**. We detect the decline in naive CD4$^+$ T cells and an enrichment of EM CD4$^+$ T cells in old animals. Furthermore, and as expected, we identify a significantly higher proportion of EM CD4$^+$ T cells in old animals that express the activation markers CD69 and PD-1, indicating that a larger fraction of cells is already in an activated state **(Fig. 2.15C)**. To avoid this phenomenon, which might interfere with the quantification of transcriptional variability, cells stained positive for CD69 and/or PD-1 were excluded during the sort of EM CD4$^+$ T cells by FACS. After sorting, we compared transcriptional variability of the core set of immune response genes in activated cells between old and young animals. Critically, these genes showed an increase in variability in older animals in both FACS-purified naive and EM CD4$^+$ T cell subsets similar to MACS-purified cells **(Fig. 2.15D and E)**.

To conclude, ageing reduces the fraction of cells in which immune activation genes are up-regulated, thus increasing cell-to-cell heterogeneity and attenuating the response to stimulation across multiple CD4$^+$ T cell subsets.

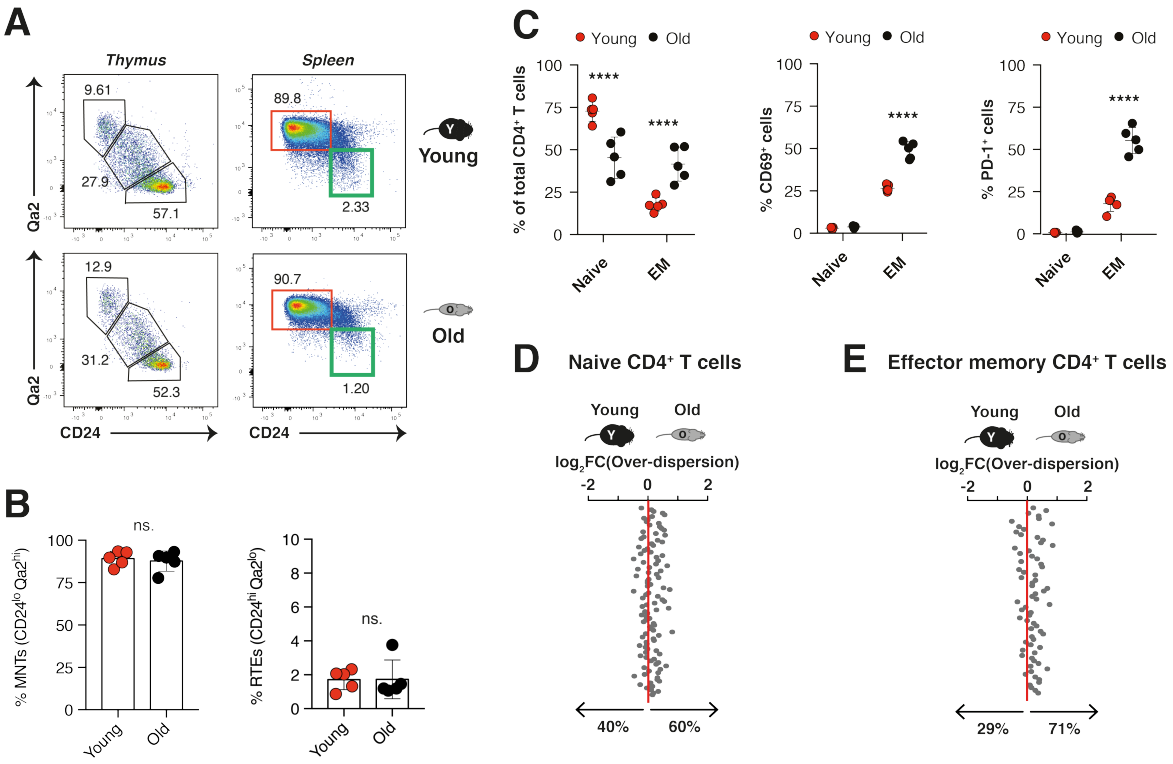


Fig. 2.15: Increased expression variability during ageing in different CD4+ T cell subsets. **(A)** Thymus and spleen were collected from young and old B6 mice, dissociated into single cell suspensions, stained with viability dye and antibodies against CD4, CD8, CD24 and Qa2 for thymocytes or antibodies against CD4, CD44, CD62L, CD24 and Qa2 for splenocytes. FACS plots shown are gated on single, live cells and either $CD4^+$ $CD8^-$ CD4 single positive (SP) thymocytes (thymus) or $CD44^{lo}$ $CD62L^{hi}$ naive $CD4^+$ T cells (spleen). Percentages relate to total gated cells. In the thymus the majority of CD4 SP express markers of recent thymic emigrants (RTEs, $CD24^{hi}$ $Qa2^{lo}$) while in the spleen naive $CD4^+$ T cells are comprised mainly of mature naive cells (red box), **(B)** Quantification of flow cytometry data from (A). Significance of difference was calculated by Mann-Whitney test (ns = not significant). MNT = mature naive T cell, **(C)** Spleens were collected from young and old B6 mice, dissociated into single cell suspensions, stained with viability dye and antibodies against CD4, CD44, CD62L, CD24, Qa2, CD69, and PD-1. FACS plots were gated on single, live, $CD4^+$ T cells and subsequently on either $CD44^{lo}$ $CD62L^{hi}$ (Naive) or $CD44^{hi}$ $CD62L^{lo}$ (EM) subsets. Expression of CD69 and PD-1 was analysed in these subsets. Results shown are pooled from 10 independent experiments with spleens harvested from 5 young and 5 old mice. Significance of difference was calculated by two-way ANOVA (****$p \leq 0.0001$), **(D)** Activated, FACS-purified naive $CD4^+$ T cells from old animals showed higher transcriptional variability compared to young animals, **(E)** Activated, FACS-purified EM $CD4^+$ T cells from old animals showed higher transcriptional variability compared to young animals.

2.7 Discussion

How cell-type-specific gene expression programmes change during organismal lifespan has long been debated [102, 103] but until the beginning of this project, few studies in mammals have quantified the cell-to-cell transcriptome-wide differences that accumulate during ageing [323]. Here, we systematically explored the effect of ageing on the dynamic activation programme of primary naive CD4$^+$ T cells. We analysed two sub-species of mice, which represents a powerful strategy to identify evolutionarily conserved gene expression programmes [330]. In contrast to humans, mice were housed in specific pathogen-free facilities that reduces transcriptional changes due to pathogen-induced immune activation [357]. In this chapter, we therefore profiled the intrinsic effect of ageing on transcriptional regulation in CD4$^+$ T cells.

By activating naive CD4$^+$ T cells and quantifying the transcriptional responses of hundreds of single-cells using scRNA-Seq, we confirmed that translation processes and immune response genes are rapidly up-regulated [345, 358, 349–352]. More interestingly, we discovered that transcriptional variability is reduced across thousands of transcripts that otherwise remain stable in mean expression levels. This indicates that immune activation rapidly reduces transcriptional heterogeneity across the population of CD4$^+$ T cells to up-regulate a specific response programme similarly in each individual cell. A similar programme has been identified in iPSC reprogramming where an early phase is characterised by probabilistic events while, later, the transcription of *Sox2* induces a more deterministic phase [62]. Previous studies assayed heterogeneity in immune responses by profiling individual cytokines such as Il2 and Ifnβ in immune cells. Early responding cells support the activation of surrounding cells by paracrine signalling [67, 19]. In contrast, by profiling thousands of genes, our approach identifies the global collapse of variability as a key event in immune activation.

Comparison of gene expression levels across species have been used as a means to identify transcription under strong selection in tissues [326, 328, 359–361], including bulk CD4$^+$ T cells from young mice and humans during immune stimulation [330]. By profiling two sub-species of mice, we identified a common set of activation genes, including well-characterised immune response genes such as *Il2ra*, that are similarly up-regulated across the two species. Furthermore, scRNA-Seq allowed us to determine the number of cells that express a certain gene. With this, we newly revealed that the vast majority

of cells within each species up-regulate the set of evolutionarily conserved genes upon immune stimulation. In contrast, we discovered that genes whose mean expression was up-regulated in a species-specific manner were often activated in only a small fraction of cells, suggesting weaker selection. Indeed, species-specific up-regulated genes showed no functional enrichment. This discovery suggests a novel defining feature of functional target genes: coherent transcriptional up-regulation across a population of cells.

Many attempts have been made to identify transcriptional signatures associated with ageing [322, 362, 363, 323]. On a genome-wide basis, we observed that ageing has minimal effects on mean expression levels in unstimulated and stimulated $CD4^+$ T cells. However, in the core set of activated genes, in both species and in distinct $CD4^+$ T cell subsets, we found a markedly more heterogeneous transcriptional response to stimulation in older mice. This increased heterogeneity was driven by ageing associated differences in the fraction of cells across the population that express these response genes. Instead of detecting structured heterogeneity, characterised by some cells not responding to the stimulus, we observed that all cells from old animals responded, but in contrast to young cells, failed to homogeneously up-regulate the response programme.

High numbers of $CD4^+$ T cells are needed to combat infection and cancer. The discovery that $CD4^+$ T cells from aged mice are unable to robustly up-regulate a core activation programme may in part explain the decrease of immune function observed in aged mammals [364, 365]. More generally, in the context of the current understanding of transcriptional dysregulation and chromatin destabilisation during ageing [304], increased cell-to-cell transcriptional variability is a major, and largely unexplored, intrinsic factor.

Following the publication of this study, several mechanisms for the increase in transcriptional variability during ageing have been proposed. In one study, the transcriptional noise increased in old compared to young human pancreatic β-cells. A possible mechanism for this increase in variability is the so called "fate drift" of β-cells to resemble α-cells. During ageing, β-cells that are defined by their expression of the hormone *insulin* increase expression of the hormone *glucagon*, the characteristic hormone of α-cells. This atypical hormone expression can result in increased transcriptional noise during ageing in the pancreas [104]. Another study by Deschênes and Chabot, 2017 proposed that the stochastic shortening of telomeres during ageing introduces variation in a process termed telomere position-effect on long distance (TPE-OLD). TPE-OLD regulates the expression of genes 10

Mb into the chromosome and its variation can lead to increased transcriptional heterogeneity during ageing [366]. Further, Cheung *et al.*, 2018 profiled a variety of epigenetic marks in different subsets of PBMCs in young (< 25 years) and old (> 65 years) humans at single-cell resolution [367]. Analysis of 40 chromatin marks in 20 cell types revealed a separation between young and old individuals and an enrichment in most chromatin marks during ageing. Furthermore, they found an increase in cell-to-cell variability for the majority of chromatin marks in aged individuals. The authors proposed a possible role for PRC in increasing epigenetic variation and showed that PRC-mediated H3K27me3 deposition explains the increase in transcriptional variability that we reported in this chapter [367].

While an increase in transcriptional noise has been shown to be associated with tissue ageing in pancreas and the immune system, a more complete view of whole-organism tissue ageing is missing. One study that begins to address this profiled changes in transcriptional noise in multiple cell types in young and old mice [228]. Not only did they confirm the increase in transcriptional noise during ageing in CD4$^+$ T cells but also observed this shift in a variety of cell types associated with the lung (e.g. NK cells, macrophages, dendritic cells, endothelial cells, smooth muscle cells and neutrophils) [228]. This analysis validates increased transcriptional noise as a major hallmark of ageing.

The major drawback in this chapter was the inability to profile all immune response genes for changes in variability due to the dependency of the over-dispersion parameter on the mean expression parameters (see **Section 1.5.7**). The simple approach to only profile genes with stable mean expression levels during immune activation excluded all immune-associated genes from analysis. These are generally the genes that define T cell phenotypes and functionality. In the next chapter, I will therefore describe an extension of the BASiCS framework to include genes that display changes in mean expression by regressing out this mean-variability dependency.

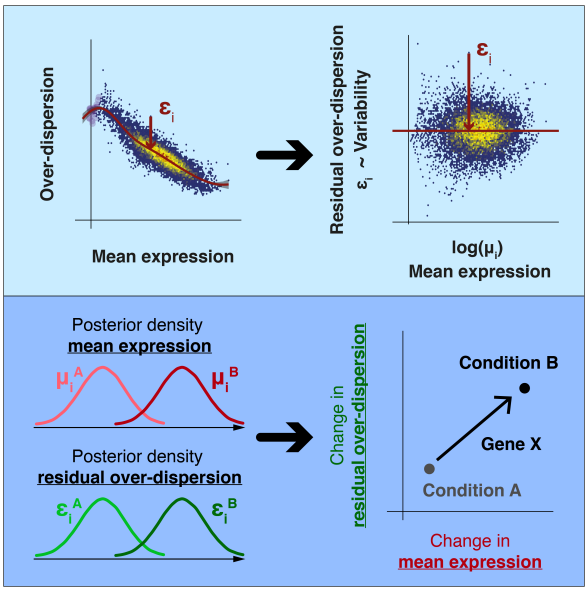
3

Addressing the mean-variability dependency in scRNA-Seq data

As shown above, cell-to-cell transcriptional variability in otherwise homogeneous cell populations plays an important role in immune activation and increases with age. Single-cell RNA sequencing can characterise this variability in a transcriptome-wide manner. However, the confounding between variability and mean expression estimates hinders meaningful comparison of expression variability between cell populations. To address this problem, we introduce a statistical approach that extends the BASiCS framework to derive a residual measure of variability that is not confounded by mean expression. This measure is used to test changes in variability in parallel to changes in mean expression on a gene-specific level. With this method, we assess changes in variability for genes responding to $CD4^+$ T cell activation and detect a synchronisation of biosynthetic machinery components. Furthermore, cytokines such as Il2 that support the activation of surrounding cells by paracrine signalling are heterogeneously up-regulated upon immune activation. When profiling more subtle transcriptional changes during $CD4^+$ T cell differentiation, we detect opposing patterns of changes in variability between *Tbx21* and *Cxcr5*, which are markers for Th1 and Tfh cells, indicating a delayed commitment process throughout differentiation. Finally, we confirmed the applicability of the newly extended BASiCS model to droplet-based scRNA-Seq data which is necessary for the subsequent chapter. ■

Declaration I worked on this project in close collaboration with Catalina Vallejos, Arianne Richard, Sylvia Richardson and John Marioni. In this project, I extended an existing Bayesian framework (BASiCS) that was developed to model scRNA-Seq data. Besides calculating the underlying math and programming the model, I helped re-write the BASiCS R package, which is now published on Bioconductor. Catalina Vallejos co-supervised me and introduced me to principles of Bayesian statistics. Arianne Richard helped with the interpretation of biological data and helped writing the manuscript. Sylvia Richardson provided statistical help on a part of the project that is not presented here. John Marioni co-supervised me and suggested analysis to be performed. Catalina Vallejos, John Marioni and I designed the study. Catalina Vallejos, John Marioni and I wrote the manuscript. I kindly thank Dominic Grün for providing the smFISH data matched to mESC scRNA-Seq data. The study has been published as:

Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, Catalina A. Vallejos. Robust expression variability testing reveals heterogeneous T cell responses. *Cell Systems*, In press, 2018



3.1 Introduction

Heterogeneity in gene expression within a population of single cells can arise from a variety of factors (see **Box 1** on page 2). In a seemingly homogeneous population of cells, unstructured expression heterogeneity can be linked to intrinsic or extrinsic noise [1]. Changes in physiological cell states (e.g. cell cycle) represent extrinsic noise, which has been found to influence expression variability within cell populations [140, 13, 6]. Intrinsic noise can be linked to epigenetic diversity [182], chromatin accessibility [188], and the genomic content of single genes [109].

Single-cell RNA sequencing generates transcriptional profiles of individual cells which allows the study cell-to-cell heterogeneity on a transcriptome-wide [134] and single gene level [16]. Consequently, this technique can be used to profile unstructured cell-to-cell variation in gene expression within and between homogeneous cell populations (i.e. where no distinct cell sub-types are present). As shown in **Section 2.4**, transcriptional noise decreases during immune activation. Ageing on the other hand destabilises the immune response, which manifests itself in the form of increased transcriptional noise. Furthermore, increasing evidence suggests that this heterogeneity plays an important role in development [49]. For instance, molecular noise was shown to increase before cells commit to lineages during differentiation [52], while the opposite is observed once an irreversible cell state is reached [15]. A similar pattern occurs during gastrulation, where expression noise is high in the uncommitted inner cell mass at E3.5 compared to the epiblast at E4.5 and where an increase in heterogeneity is observed when cells exit the pluripotent state and form the uncommitted epiblast at E6.5 (see **Fig. 1.2** and [17]).

Motivated by scRNA-Seq, recent studies have extended traditional differential expression analyses to explore more general patterns that characterise differences between cell populations [368]. As described in **Section 1.5.7** and **2.3**, BASiCS [11, 295] introduced a probabilistic tool to assess differences in cell-to-cell heterogeneity between two or more cell populations. To meaningfully assess changes in biological variability across the entire transcriptome, one strong confounding effect must be taken into account: differential variability between populations that is driven by changes in mean expression. This arises because biological noise is negatively correlated with protein abundance [216, 139, 205] or mean RNA expression (see **Section 1.5.7** and [10, 51]). To acknowledge the variance-mean relationship, the initial version of BASiCS restricted differential vari-

ability testing to those genes with equal mean expression across populations (see **Section 2.3**).

Previous studies derived measures of transcriptional variability that are independent of mean expression. These approaches ranged from a simple linear regression between the logarithm of the coefficient of variation $\log_2$(CV) and the $\log_2$(mean expression) [369] to more elaborate models as described by Grün *et al.*, 2014. Their model aims to capture (i) the Poissonian sampling noise for lowly expressed transcripts and (ii) differences in total transcript abundance between cells for highly expressed genes. The mixture of these effects introduces a non-linear relationship between mean expression and the CV. The model also captures technical noise by incorporating reads from technical spike-in RNA. In this case, the number of transcripts available for sequencing is Gamma distributed due to variation in capture efficiency. The sequencing process on the other hand is a Poisson process [370]. The combination of these distribution forms a negative binomial which models the expression counts of all biological genes best [134]. A non-parmateric strategy to model the mean-variance relationship was proposed by Kolodziejczyk *et al.*, 2015, where the mean-independent measure of variability is the distance between the CV^2 and a rolling median along mean expression [12].

In this chapter, we extend the statistical model in BASiCS by implementing a more general approach to account for this confounding effect. By incorporating a flexible, non-linear regression trend, we derive a residual measure of cell-to-cell transcriptional variability that is not confounded by mean expression. This is used to define a probabilistic rule to robustly highlight changes in variability, even for differentially expressed genes. Unlike previous approaches that derive point estimates of residual variability, our approach directly performs gene-specific statistical testing between two conditions using a readily available measure of uncertainty.

Using our approach, we identify a synchronisation of biosynthetic machinery components in $CD4^+$ T cells upon early immune activation as well as an increased variability in the expression of genes related to $CD4^+$ T cell immunological function. Furthermore, we detect evidence of early cell fate commitment of $CD4^+$ T cells during malaria infection characterised by a decrease in *Tbx21* expression heterogeneity and a rapid collapse of global transcriptional variability after infection. These results highlight biological insights into T cell activation and differentiation that are only revealed by jointly studying changes in mean expression and variability.

3.2 Extending the BASiCS model

Unlike bulk RNA sequencing, scRNA-Seq provides information about cell-to-cell expression heterogeneity within a population of cells. Past works have used a variety of measures to quantify this heterogeneity. Among others, this includes the CV^2 [10] and entropy measures [15]. The BASiCS model [11, 295], which was introduced in **Section 1.5.7**, focuses on biological *over-dispersion* as a proxy for transcriptional heterogeneity. This is defined as the excess of variability that is observed with respect to what would be predicted by Poisson sampling noise, after accounting for technical variation.

3.2.1 The BASiCS model

Let X_{ij} be a random variable representing the expression count of gene $i \in \{1,\ldots,q\}$ in cell $j \in \{1,\ldots,n\}$. To control for technical noise, we employ reads from synthetic RNA spike-ins (see [301]). We assume the first q_0 genes to be biological followed by the $q-q_0$ spike-in genes. BASiCS assumes a hierarchical Poisson formulation:

$$X_{ij}|\mu_i,\phi_j,\nu_j,\rho_{ij} \overset{\text{ind}}{\sim} \begin{cases} \text{Poisson}(\phi_j\nu_j\mu_i\rho_{ij}), & i=1,...,q_0, j=1,...n; \\ \text{Poisson}(\nu_j\mu_i), & i=q_0+1,...,q, j=1,...,n, \end{cases} \tag{3.1}$$

where, to account for technical (ν_j) and biological (ρ_{ij}) factors that affect the variance of the transcript counts, we incorporate two random effects:

$$\nu_j|s_j,\theta \overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{1}{\theta},\frac{1}{s_j\theta}\right), \quad \rho_{ij}|\delta_i \overset{\text{iid}}{\sim} \text{Gamma}\left(\frac{1}{\delta_i},\frac{1}{\delta_i}\right) \tag{3.2}$$

Here, ϕ_j represents a cell-specific normalisation parameter to correct for differences in mRNA content between cells. Gene-specific parameters μ_i represent average expression of a gene across cells. The strength of the technical noise ν_j is quantified by a global parameter θ (shared across all genes and cells). s_j models cell-specific differences in efficiency to capture RNA spike-in transcripts affecting all biological and technical genes. The strength of heterogeneous gene expression across cells ρ_{ij} is controlled by gene-specific over-dispersion parameters δ_i which we used as a proxy for biological expression variability in the previous chapter. As shown in the previous chapter, over-dispersion as a measure of variability can be used to identify genes whose transcriptional heterogeneity differs between groups of cells (defined by experimental conditions or cell types). However, the strong relationship that is typically observed between variability and mean estimates (see **Section 1.5.7** and [10]) can hinder the interpretation of these results.

3.2.2 Approaches to correct the mean-variability confounding effect

A simple solution to avoid the confounding effect of mean expression was used in **Chapter 2** by restricting the assessment of differential variability to genes with equal mean expression across populations **(Fig. 3.1A and Section 2.3)**. However, this is sub-optimal, particularly in the case of naive and activated $CD4^+$ T cells where large sets of genes are differentially regulated upon immune activation. With the current model, immune response genes (e.g. cytokines, nuclear receptors, transcription factors) are excluded from differential variability testing. An alternative approach is to directly adjust variability measures to remove this confounding. For example, Kolodziejczyk *et al.*, 2015 computed the empirical distance between the CV^2 to a rolling median along expression levels — referred to as the DM method [139, 12].

In line with this idea, our method extends the statistical model implemented in BASiCS [11, 295] to meaningfully assess changes in transcriptional heterogeneity when genes exhibit shifts in mean expression **(Fig. 3.1B)**. For this, we infer a regression trend between over-dispersion (δ_i) and gene-specific mean parameters (μ_i), by introducing a joint informative prior to capture the dependence between these parameters. A latent gene-specific *residual over-dispersion* parameter ε_i describes departures from this trend **(Fig 3.1C)**. The value of ε_i indicates whether a gene exhibits more (positive) or less (negative) variation than expected relative to genes with similar expression levels. Importantly, this measure is not confounded by mean expression **(Fig. 3.1D)**.

The hierarchical Bayesian approach infers full posterior distributions for the gene-specific latent residual over-dispersion parameters ε_i. As a result, we can directly use a probabilistic approach to identify genes with large absolute differences in residual over-dispersion between two groups of cells. When the posterior samples of ε_i in condition A are very different from posterior samples of ε_i in condition B, the majority of values for $|\varepsilon_i^A - \varepsilon_i^B|$ are larger than a given threshold $\psi_0 > 0$. In this case, the gene is found to be differentially variable between the two conditions **(Fig. 3.1E and Section 1.5.5)**. In contrast, mean-corrected point estimates for residual noise parameters (such as those obtained by the DM method) cannot be directly used to perform gene-specific statistical testing between two conditions as no measure of the uncertainty in the estimate is available.

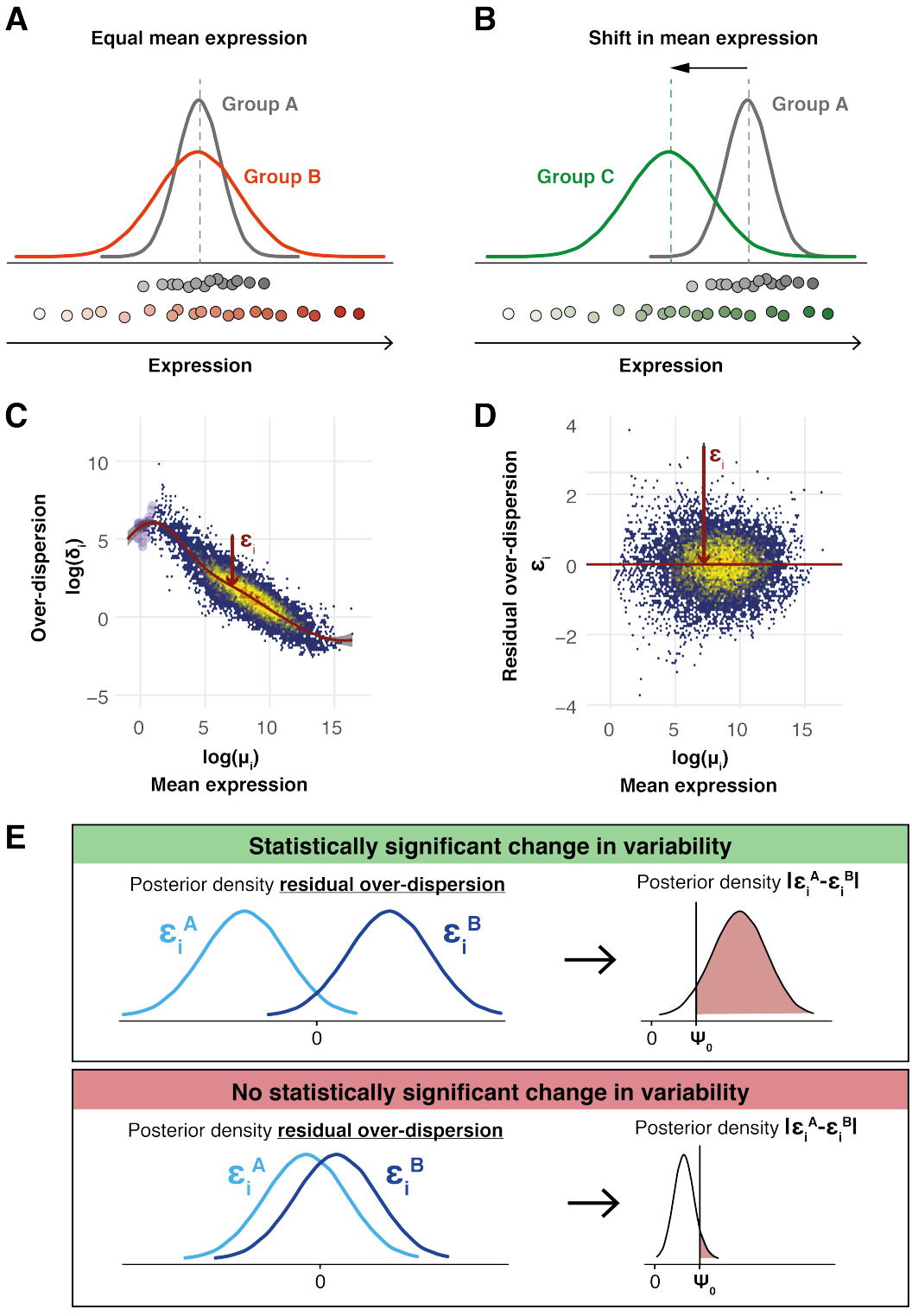


Fig. 3.1: Addressing the mean confounding effect in scRNA-Seq data (full legend on next page).

Fig. 3.1: Addressing the mean confounding effect in scRNA-Seq data (continued).
(A and B) Illustration of changes in expression variability for a single gene between two cell populations without (left) and with (right) changes in mean expression, **(C and D)** The extended BASiCS model infers a regression trend between gene-specific estimates of over-dispersion parameters δ_i and mean expression μ_i. Residual over-dispersion parameters ε_i are defined by departures from the regression trend (red arrow). The colour code within the scatterplots is used to represent areas with high (yellow/red) and low (blue) concentration of genes. For illustration purposes, the data introduced by Antolović *et al.*, 2017 [51] has been used, **(C)** Illustration of the typical confounding effect that is observed between gene-specific estimates of over-dispersion parameters δ_i and mean expression parameters μ_i. Genes that are not detected in at least 2 cells are indicated by purple points, **(D)** Gene-specific estimates of residual over-dispersion parameters ε_i are independent of mean expression parameters μ_i, **(E)** Illustration of how posterior uncertainty is used to highlight changes in residual over-dispersion. Two example genes with (upper panels) and without (lower panels) changes in residual over-dispersion are shown. Left panels illustrate the posterior density associated to residual over-dispersion parameters ε_i for a gene in two groups of cells (group A: light blue, group B: dark blue). The coloured area in the right panels represents the posterior probability of observing an absolute difference $|\varepsilon_i^A - \varepsilon_i^B|$ that is larger than the minimum tolerance threshold ψ_0.

3.2.3 Modelling the confounding between mean and over-dispersion

Here, we extend BASiCS to account for the confounding effect described above. In a Bayesian framework, the prior information captures the relationship between parameters. Therefore, we introduce the following joint prior distribution for $(\mu_i, \delta_i)'$:

$$\mu_i \sim \text{log-Normal}\left(0, s_\mu^2\right), \qquad \delta_i|\mu_i \sim \text{log-T}_\eta\left(\text{f}(\mu_i), \sigma^2\right). \tag{3.3}$$

The latter is equivalent to the following non-linear regression model:

$$\log(\delta_i) = \text{f}(\mu_i) + \varepsilon_i, \qquad \varepsilon_i \sim \text{T}_\eta(0, \sigma^2), \tag{3.4}$$

where $\text{f}(\mu_i)$ represents the over-dispersion (on the log-scale) that is predicted by the global trend (across all genes) for a given mean expression μ_i. Therefore, ε_i can be interpreted as a latent gene-specific *residual over-dispersion* parameter, capturing departures from the overall trend.

A similar approach was introduced by DESeq2 [371] in the context of bulk RNA sequencing. Whereas DESeq2 assumes normally distributed errors when estimating this trend, here we use Student-T distributed errors (with η degrees of freedom) as it leads to inference that is more robust to the presence of outlier genes [372].

Moreover, the parametric trend assumed by DESeq2 is replaced by a more flexible semi-parametric approach. This is defined by

$$\mathrm{f}(\mu_i) = \alpha_0 + \log(\mu_i)\alpha_1 + \sum_{l=1}^{L} \mathrm{g}_l(\log(\mu_i))\beta_l, \tag{3.5}$$

which is a linear combination of an intercept, a linear term $\log(\mu_i)$ and a set of L Gaussian radial basis function (GRBF) kernels $\mathrm{g}_1(\cdot), \ldots, \mathrm{g}_L(\cdot)$. As in Kapourani *et al.*, 2016 [373], these are defined as:

$$\mathrm{g}_l(\log(\mu_i)) = \exp\left\{-\frac{1}{2}\left(\frac{\log(\mu_i) - m_l}{h_l}\right)^2\right\}, \quad l = 1, \ldots, L, \tag{3.6}$$

where m_l and h_l represent location and scale hyper-parameters for GRBF kernels and $\alpha_0, \alpha_1, \beta_1, \ldots, \beta_L$ are regression coefficients.

In equation (3.5), the linear term captures the (typically negative) global correlation between δ_i and μ_i. Its addition also stabilises inference of GRBFs around mean expression values where only a few of genes are observed. In equation (3.6), the location and scale hyper-parameters (m_l, h_l) are assumed to be fixed *a priori* (see **Section 3.2.6**).

3.2.4 Implementation

Next, we will give a detailed explanation on how the model was built and how posterior sampling was performed.

Prior specification

For implementation purposes, the log-Student-T distribution in equation (3.3) is represented via a shape mixture of a log-Normal density with a Gamma density as in Vallejos *et al.*, 2015 [337]. This introduces an auxiliary set of parameters λ_i such that the full prior specifications of the extended BASiCS model are:

$$\begin{aligned}
\mu_i &\overset{\text{ind}}{\sim} \text{log-Normal}\left(0, s_\mu^2\right) \\
\delta_i|\mu_i,\beta,\sigma^2,\lambda_i,\eta &\overset{\text{ind}}{\sim} \text{log-N}\left(\text{f}(\mu_i), \frac{\sigma^2}{\lambda_i}\right) \\
\lambda_i|\eta &\overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{\eta}{2}, \frac{\eta}{2}\right) \\
\beta|\sigma^2 &\sim \text{Normal}(m_\beta, \sigma^2 V_\beta), \\
\sigma^2 &\sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}), \\
s_j &\overset{\text{iid}}{\sim} \text{Gamma}(a_s, b_s) \\
(\phi_1,\ldots,\phi_n)' &\sim n \times \text{Dirichlet}(a_\phi), \\
\theta &\sim \text{Gamma}(a_\theta, b_\theta)
\end{aligned}$$

Here, $s_\mu^2, m_\beta, V_\beta, a_{\sigma^2}, b_{\sigma^2}, a_s, b_s, a_\phi, a_\theta, b_\theta$ are hyper-parameters that are fixed *a priori*. Their initial values can be found in **Appendix B.1.2**. In principle, the degrees of freedom parameter η could also be estimated within a Bayesian framework. However, we observed that fixing this parameter *a priori* led to more stable results. A default choice for this parameter is described in **Section 3.2.6**.

Estimation of regression parameters

To simplify inference for the regression coefficients $\beta = (\alpha_0, \alpha_1, \beta_1, \ldots, \beta_L)'$ equation (3.5) can be rewritten as a linear regression model using

$$\mathrm{f}(\mu_i) = X\beta \tag{3.7}$$

Here, X is a $q_0 \times (L+2)$ model matrix given by

$$X = \begin{pmatrix} 1 & \log(\mu_1) & g_1(\log(\mu_1)) & \cdots & g_L(\log(\mu_1)) \\ 1 & \log(\mu_2) & g_1(\log(\mu_2)) & \cdots & g_L(\log(\mu_2)) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \log(\mu_{q_0}) & g_1(\log(\mu_{q_0})) & \cdots & g_L(\log(\mu_{q_0})) \end{pmatrix} \tag{3.8}$$

Each column contains either the intercept, the linear component or values of one of the L GRBF. This matrix is updated every 50 iterations during posterior sampling.

Posterior inference

Posterior inference for the model described above is implemented by extending the Adaptive Metropolis within Gibbs sampler [287] that was adopted by Vallejos *et al.*, 2016 [295]. To implement the sampler, the full conditionals for each model parameter need to be derived.

As in Vallejos *et al.*, 2015 [11], the random effect ρ_{ij} in 3.1 is integrated out, leading to the following count distributions:

$$X_{ij}|\mu_i, \delta_i, \phi_j, \nu_j \overset{\text{ind}}{\sim} \begin{cases} \text{Neg-Bin}\left(\frac{1}{\delta_i}, \frac{\phi_j \nu_j \mu_i}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right), & i = 1, ..., q_0, j = 1, ...n; \\ \text{Poisson}(\nu_j \mu_i), & i = q_0+1, ..., q, j = 1, ..., n \end{cases} \tag{3.9}$$

Based on equation (3.9), the likelihood function therefore takes the form

$$\begin{aligned} \mathscr{L} = & \left[\prod_{i=1}^{q_0}\prod_{j=1}^{n} \frac{\Gamma(x_{ij} + \frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})x_{ij}!} \left(\frac{\frac{1}{\delta_i}}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}} \left(\frac{\phi_j \nu_j \mu_i}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{x_{ij}}\right] \\ & \times \left[\prod_{i=q_0+1}^{q}\prod_{j=1}^{n} \frac{(\nu_j \mu_i)^{x_{ij}}}{x_{ij}!} \exp\{-\nu_j \mu_i\}\right] \times \left[\prod_{j=1}^{n} \frac{(s_j\theta)^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})} \nu_j^{\frac{1}{\theta}-1} \exp\left\{-\frac{\nu_j}{s_j\theta}\right\}\right] \end{aligned} \tag{3.10}$$

The full conditionals can now be derived by calculating the parameter-dependent part of the posterior distribution which is a product of the likelihood times the prior specifications $\pi^*(\cdot) \propto \mathscr{L} \times \pi(\cdot)$. Full conditionals for the model are as follows:

$$\pi^*(\mu_i|\cdot) \propto \frac{\mu_i^{\sum_{j=1}^n x_{ij}}}{\prod_{j=1}^n (\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i}+x_{ij}}} \times \exp\left(-\frac{(\log(\mu_i))^2}{2a_\mu^2} - \frac{\lambda_i(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2}\right)\frac{1}{\mu_i}$$

$$\pi^*(\delta_i|\cdot) \propto \left[\prod_{j=1}^n \frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})} \frac{(\frac{1}{\delta_i})^{\frac{1}{\delta_i}}}{(\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i}+x_{ij}}}\right] \times \exp\left\{-\frac{\lambda_i(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2}\right\}\frac{1}{\delta_i}$$

$$\pi^*(\beta|\cdot) \propto \text{Normal}(m_\beta^*, \sigma^2 V_\beta^*)$$

$$\pi^*(\lambda_i|\cdot) \propto \text{Gamma}(a_\lambda^*, b_\lambda^*)$$

$$\pi^*(\sigma^2|\cdot) \propto \text{Inv-Gamma}(a_{\sigma^2}^*, b_{\sigma^2}^*)$$

$$\pi^*(s_j|\cdot) \propto {s_j}^{a_s - \frac{1}{\theta}-1} \exp\{-\frac{\nu_j}{s_j\theta} - b_s s_j\}$$

$$\pi^*(\phi_j|\cdot) \propto \frac{\prod_{i=1}^{q_0} \phi_j^{\sum_{j=1}^n x_{ij}}}{\prod_{i=1}^{q_0}\prod_{j=1}^n (\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i}+x_{ij}}} \times \pi(\phi_j)$$

$$\pi^*(\nu_j|\cdot) \propto \left[\prod_{i=1}^{q_0}\left(\frac{1}{\phi_j\nu_j\mu_i + \frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\nu_j}{\phi_j\nu_j\mu_i + \frac{1}{\delta_i}}\right)^{x_{ij}}\right]\left[\prod_{i=q_0+1}^{q} {\nu_j}^{x_{ij}} \exp\{-\nu_j\mu_i\}\right] {\nu_j}^{\frac{1}{\theta}-1}\exp\{-\frac{\nu_j}{s_j\theta}\}$$

$$\pi^*(\theta|\cdot) \propto \frac{\left(\prod_{j=1}^n \frac{s_j}{\nu_j}\right)^{-\frac{1}{\theta}}}{\Gamma^n(\frac{1}{\theta})}\theta^{a_\theta - \frac{n}{\theta}-1}\exp\{-\frac{1}{\theta}\sum_{j=1}^n \frac{\nu_j}{s_j} - b_\theta\theta\}$$

Here, posteriors for $\beta, \lambda_i, \sigma^2$ take on closed form distributions. The posterior for s_j represents a Generalised Inverse Gaussian distribution. To sample all other posterior distributions adaptive Metropolis sampling was implemented as described in **Section 1.5.3**. The derivation of the full conditionals can be found in **Appendix B.1.4**. In practice, the MCMC sampler is run for 40,000 iterations with a 20,000 iteration burn in period. The chain was thinned by storing parameter samples every 20 iterations.

3.2.5 Probabilistic rule associated to the differential test

The residual over-dispersion parameter is calculated as $\varepsilon_i = \log(\delta_i) - \mathrm{f}(\mu_i)$. We can now implement a probabilistic approach to identify changes in residual over-disperison between groups of cells. Let δ_i^A and δ_i^B be the over-dispersion parameters associated to gene i in groups A and B. Following equation (3.4), the $\log_2$ fold change in over-dispersion between these groups can be decomposed as:

$$\log_2\left(\frac{\delta_i^A}{\delta_i^B}\right) = \log_2(e) \times \left[\underbrace{\mathrm{f}^A(\mu_i^A) - \mathrm{f}^B(\mu_i^B)}_{\text{Mean contribution}} + \underbrace{\varepsilon_i^A - \varepsilon_i^B}_{\text{Residual change}}\right] \tag{3.11}$$

where the first term captures the over-dispersion change that can be attributed to differences between μ_i^A and μ_i^B. The second term in equation (3.11) represents the change in residual over-dispersion that is not confounded by mean expression. Based on this observation, statistically significant differences in residual over-dispersion will be identified for those genes where the tail posterior probability of observing a large difference between ε_i^A and ε_i^B exceeds a certain threshold $\psi_0 > 0$:

$$\pi_i(\psi_0) = \mathrm{P}(\mid \varepsilon_i^A - \varepsilon_i^B \mid > \psi_0 \mid \mathrm{Data}) > \alpha_R \tag{3.12}$$

As a default choice for testing changes in over-dispersion we chose a 50% increase. This translates into $\psi_0 = \log_2(1.5)/\log_2(e) \approx 0.41$ as default threshold for testing changes in residual over-dispersion. In the limiting case when $\psi_0 = 0$, the probability in equation (3.12) is equal to 1 regardless of the information contained in the data. Therefore, as in Bochkina *et al.*, 2007 [294], our decision rule is based on the maximum of the posterior probabilities associated to the one-sided hypotheses $\varepsilon^A - \varepsilon_i^B > 0$ and $\varepsilon^A - \varepsilon_i^B < 0$:

$$2 \times \max\{\pi_i^+, 1 - \pi_i^+\} - 1 > \alpha_R, \;\; \text{with} \;\; \pi_i^+ = \mathrm{P}(\varepsilon_i^A - \varepsilon_i^B > 0 \mid \mathrm{Data}) \tag{3.13}$$

In both cases, the posterior probability threshold α_R is chosen to control the expected false discovery rate (EFDR) [302]. The default value for EFDR is set to 10%. The EFDR is defined as:

$$\mathrm{EFDR}_{\alpha_R}(\psi_0) = \frac{\sum_{i=1}^{q_0}(1 - \pi_i(\psi_0))\mathrm{I}(\pi_i(\psi_0) > \alpha_R)}{\sum_{i=1}^{q_0}\mathrm{I}(\pi_i(\psi_0) > \alpha_R)} \tag{3.14}$$

where I(A)=1 if the event A is true. As a default and to support interpretability of the results, we exclude genes that are not expressed in at least 2 cells per condition from differential variability testing.

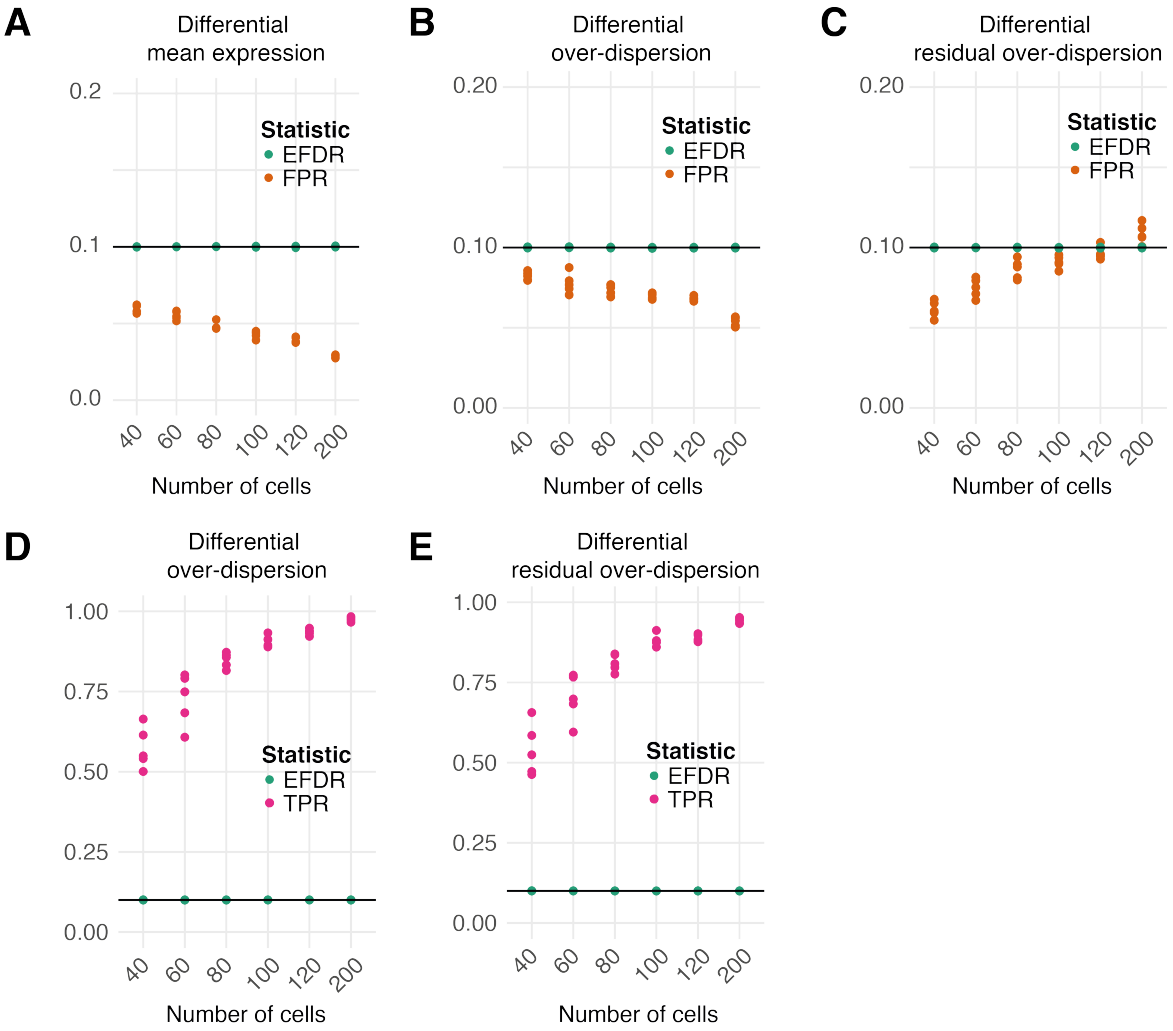


Fig. 3.2: EFDR, FPR and TPR estimation using simulated data (Full legend on next page). Data was simulated using the BASiCS model with model parameters set by empirical estimates based on 98 microglia cells [259] **(Table 3.1)**. Different samples sizes (40 - 200 cells) were simulated in replicates of 5. Differential testing was performed between 2 simulated datasets of equal size to calculate the false positive rate (FPR, number of detections divided by number of genes tested) and the true positive rate (TPR, number of true positive divided by number of all positives). Moreover, we report the expected false discovery rate (EFDR, [302]). For each test, the EFDR was controlled to 10% and the default minimum tolerance thresholds were used ($\tau_0 = \log_2(1.5)$, $\omega_0 = \log_2(1.5)$ and $\psi_0 = 0.41$), **(A)-(C)** Synthetic datasets generated using the null model (without changes in variability). FPR and EFDR for (A) differential mean expression, (B) differential over-dispersion and (C) differential residual over-dispersion testing using datasets with increasing samples sizes, **(D)-(E)** Synthetic datasets generated using the alternative model where 1000 genes were randomly selected and their associated over-dispersion parameters were increased or decreased by a $\log_2$ fold change of 5. TPR and EFDR for (D) differential over-dispersion testing and (E) differential residual over-dispersion testing using simulated datasets with increasing samples sizes.

To evaluate the performance of our test we generated synthetic data under a null model (without changes in variability) and an alternative model (with changes in variability). All datasets were generated following the BASiCS model, with parameter values set by empirical posterior estimates based on 98 microglia cells [259] (see **Table 3.1** in **Section 3.3**). To simulate data under an alternative model, 1000 genes were randomly selected and their associated δ_i's were increased or decreased by a $\log_2$ fold change of 5. Increasing numbers of cells were simulated to estimate the effect of sample size on differential testing. Differential testing was performed either between data simulated on the same set of parameters (null model) or between data simulated from the original parameters and the altered parameters (alternative model). We report the EFDR [302] as well as the false positive rate (FPR) for simulations under the null model **(Fig. 3.2A-C)** and the true positive rate (TPR) for simulations under the alternative model **(Fig. 3.2D and E)**.

As specified, the EFDR is controlled at 10%. Furthermore, the FPR for differential mean expression and differential over-dispersion is consistently smaller than 10% and is only slightly higher for differential residual over-dispersion testing. Since the data was simulated under the non-regression BASiCS model, the simulated expression variability cannot be controlled in terms of residual over-dispersion parameters ε_i leading to subtle differences between simulated cell populations. The TPR for differential over-dispersion and differential residual over-dispersion testing increases with increasing sample size and plateaus at 100% **(Fig. 3.2)**.

3.2.6 Choice of hyper-parameters

As discussed above, the degrees of freedom η, the number of GRBFs L as well as their hyper-parameters (m_l, h_l) are set *a priori*. Here, we explain the default values implemented in the extended BASiCS model. These were chosen to achieve a compromise between flexibility of the trend fit and the strength of shrinkage towards the estimated trend. Further discussion on the shrinkage can be found in **Section 3.4**.

Firstly, we observed that large values of L can lead to over-fitting but that small values of L can limit the flexibility to capture non-linear relations between $\log(\delta_i)$ and $\log(\mu_i)$ **(Fig. 3.3)**. Thus, as a parsimonious choice, we selected $L = 10$. Moreover, as in Kapourani *et al.*, 2016 [373], values for m_l were chosen to be equally spaced across the range of $\log(\mu_i)$:.

$$m_l = a + (l-1)\frac{b-a}{L-1}, \quad l = 1, \ldots, L, \tag{3.15}$$

where $a = \min_{i\in\{1,\ldots,q_0\}}\{\log(\mu_i)\}$ and $b = \max_{i\in\{1,\ldots,q_0\}}\{\log(\mu_i)\}$. As μ_i values are unknown *a priori* and change throughout the sampling procedure, a and b are updated every 50 MCMC iterations during the burn-in phase. Additionally, the scale hyper-parameters h_l control the width of the GRBFs and, consequently, the locality of the regression. As a default, we set these as $h_l = c \times \Delta m$, where c is a fixed proportionality constant and Δm is the distance between consecutive values of m_l. In practice, we observed that the choice of a particular value of c is not critical, as long as narrow kernels ($c < 0.5$) are avoided **(Fig. 3.3)**. As a default, $c = 1.2$ was chosen.

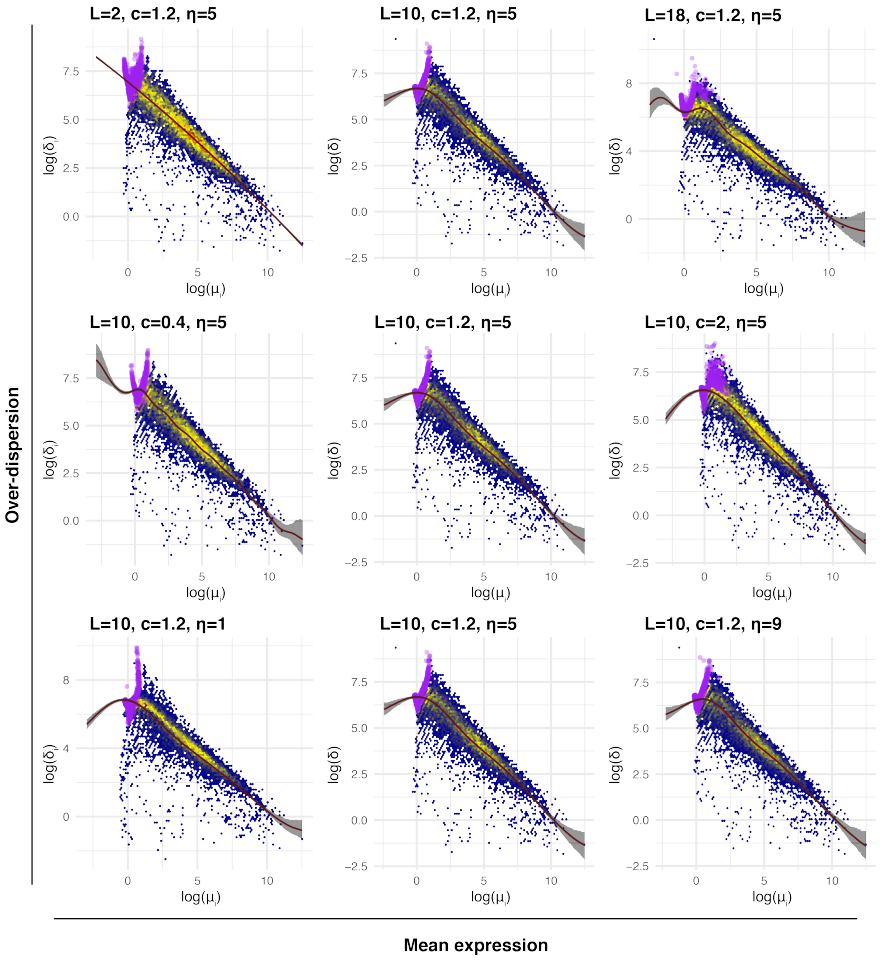


Fig. 3.3: Effect of regression hyper-parameters on trend fitting.
Posterior estimates of over-dispersion parameters δ_i are plotted versus posterior estimates of mean expression parameters μ_i on the log-log scale. The extended BASiCS model was used to estimate these parameters using naive CD4$^+$ T cells from the previous chapter. Different hyper-parameter combinations were used to fit the model. L: number of Gaussian Radial Basis Functions, c: constant multiplier of the scale parameter, η: degrees of freedom. Purple points indicate genes which are expressed in fewer than 2 cells.

The degrees of freedom η controls the tails of the distribution for the residual term in equation (3.4). This influences the shrinkage towards the global trend and the robustness against outlying observations **(Fig. 3.3)**. If $\eta \geq 30$, ε_i approximately follows a normal distribution for which posterior inference for β is known to be sensitive to outliers. Instead, small values of η introduce heavy-tails for ε_i, leading to more robust posterior inference. In principle, η could be estimated within a Bayesian framework. However, this is problematic as the likelihood function associated to equation (3.4) can be unbounded [372]. Here, we opt for a pragmatic approach where the value of η is fixed *a priori*. To select a reasonable default value, we ran the regression BASiCS model for a grid of possible values of η, using the datasets described in **Table 3.1** in **Section 3.3** (with L, m_l and h_l fixed as described above). In all cases, we calculated a Monte Carlo estimate for the log-likelihood associated to equation (3.1) as a proxy for goodness-of-fit **(Fig. 3.4A)**. We observed that log-likelihood estimates were consistently the smallest for $\eta = 1$ and that no substantial differences are observed across larger values of η. The *Dictyostelium* data show very similar log-likelihood estimates for all tested η. When visualising posterior estimates for the variance σ^2 of the distribution for the residual term depending on the degrees of freedom chosen, we observe a constant increase plateauing when the distribution reaches the normal distribution at $\eta = 30$ **(Fig. 3.4B)**. We chose $\eta = 5$ to be the default parameter as a compromise between shrinkage and sensitivity to outlying data points.

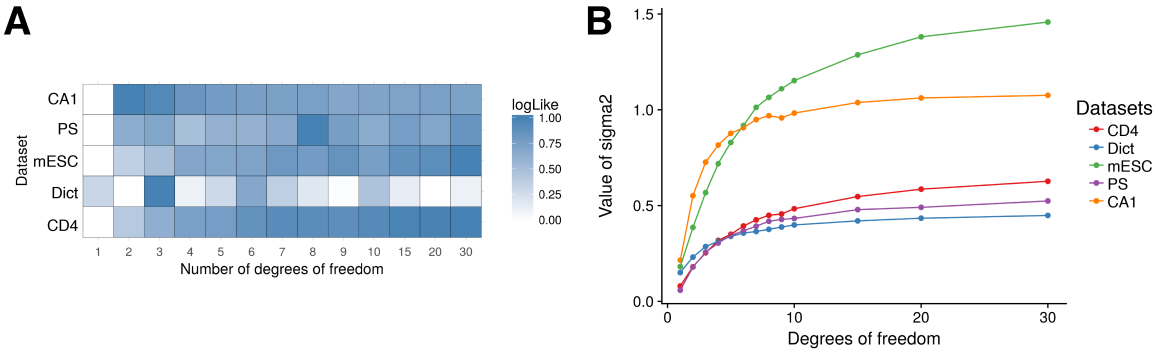


Fig. 3.4: Comparison of model fits for varying degrees of freedom.
The regression BASiCS model was fit to datasets listed in **Table 3.1**. These include CA1 pyramidal neurons (CA1, [259]), pool-and-split RNA 2i medium (PS, [134]), mouse embryonic stem cells 2i medium (mESC, [134]), *Dictyostelium* cells at day 0 of differentiation (Dict, [51]) and naive CD4$^+$ T cells (CD4, previous chapter). **(A)** The model was fit using varying degrees of freedom and the log-likelihood was calculated as stated in equation (3.10). The log-likelihood was scaled between the highest and lowest value for each dataset. **(B)** Posterior estimates of the variance parameter σ^2 depending on the number of degrees of freedom.

3.3 Pre-processing of scRNA-Seq data used in this chapter

We employed a range of different datasets to test the proposed methodology. These datasets were selected to cover different experimental techniques (with and without UMIs) and to encompass a variety of cell types. Moreover, key features of each dataset can be found in **Table 3.1**.

Table 3.1: Datasets used for model testing and analysis.
For each of the datasets analysed in this study: number of cells (2nd column), number of genes (biological + technical spike-ins, 3rd column), number of batches (4th column), type of data acquisition system (5th column), information on whether the data was generated using unique molecular identifiers (UMIs, 6th column) and the reference to the original study (7th column) are provided.

Dataset	# cells	# genes	# batches	Protocol	UMIs	Ref.
Young naive CD4$^+$ T cells	93	10,553	2	Fluidigm C1	No No	[22]
Young active CD4$^+$ T cells	53	10,553	2	Fluidigm C1	No	[22]
Microglia cells	98	10,687	1	Fluidigm C1	Yes	[259]
CA1 pyramidal neurons	948	10,687	1	Fluidigm C1	Yes	[259]
Malaria infected CD4$^+$ T cells day 2	89	7899	2	Fluidigm C1	No	[253]
Malaria infected CD4$^+$ T cells day 4	133	7899	2	Fluidigm C1	No	[253]
Malaria infected CD4$^+$ T cells day 7	64	7899	1	Fluidigm C1	No	[253]
Dictyostelium cells day 0	131	10,738	3	Fluidigm C1	No	[51]
Pool-split RNA 2i medium	76	8924	2	CEL-Seq	Yes	[134]
mESC 2i medium	74	8924	2	CEL-Seq	Yes	[134]
Pool-split RNA serum medium	56	8924	2	CEL-Seq	Yes	[134]
mECS serum medium	52	8924	2	CEL-Seq	Yes	[134]

3.3.1 *Dictyostelium* cells

Antolović *et al.*, 2017 studied changes in expression variability between 0 hours (undifferentiated), 3 hours and 6 hours of *Dictyostelium* differentiation [51]. Raw data is available by direct download (see Data S1 in [51]). Across all time points, 5 cells were removed due to low quality. Technical spike-in genes that were not detected and biological genes with an average expression (across all cells) smaller than 1 count were removed. In total, 433 cells (131 cells in 3 batches at 0h, 157 cells in 3 batches at 3h, and 145 cells in 3 batches at 6h) and 10,551 genes (88 technical and 10,650 biological genes) passed filtering. We used data from the 0h time point to test the functionality of our model.

3.3.2 Mouse brain cells

This dataset was composed of UMI scRNA-Seq data of cells isolated from the mouse somatosensory cortex and hippocampal CA1 region [259]. Raw data is available from Gene Expression Omnibus under accession code GSE60361. Prior to the analysis, we removed technical genes with 0 total counts and biological genes for which the average count across all 3007 cells was below 0.1. The groups comprising microglia cells and CA1 neurons were chosen for analysis to include cell populations comprising a small and large number of cells. For these groups, 98 cells (microglia), 939 cells (CA1 pyramidal neurons) and 10,744 genes (10,687 biological and 57 technical genes) passed filtering.

3.3.3 Pool-and-split RNAseq data

This UMI-based dataset provides a control experiment to assess changes in biological heterogeneity in a situation where mean expression remains unchanged across conditions. Pool-and-split samples were created by pooling 1 million mESCs grown in 2i or serum medium and splitting 20pg of RNA into aliquots. These libraries are compared against single-cell samples (mESCs) [134]. Raw data is available from Gene Expression Omnibus under accession code GSE54695.

As in Grün *et al.*, 2014 [134], some cells were removed from the analysis due to low expression of the stem cell marker *Oct4*. Technical genes with 0 total counts were also removed from the analysis. Additionally, lowly expressed biological genes with fewer than 0.5 counts (on average, across all samples) were excluded. This left 258 libraries (74 single mESCs grown in 2i medium, 52 single mESCs grown in serum medium, 76 pool-and-split

aliquots from cells grown in 2i medium and 56 pool-and-split aliquots from cells grown in serum medium) as well as 8924 genes (50 technical spike-ins and 8874 biological genes) for the analysis. Each condition contained 2 batches.

Matched smFISH data from mESCs grown in 2i and serum media were obtained from Dominic Grün (Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany) through personal communications. This smFISH experiment assayed 9 genes (*Gli1*, *Klf4*, *Notch1*, *Pcna*, *Pou5f1*, *Sohlh2*, *Sox2*, *Stag3*, *Tpx2*) in more than 70 cells per condition.

3.3.4 CD4$^+$ T cell activation

Non-UMI scRNA-Seq data of CD4$^+$ T cells represent data analysed in the previous chapter. Raw data is available from ArrayExpress under accession code E-MTAB-4888. To perform a variety of tests, naive and activated CD4$^+$ T cells from young *Mus musculus* (B6) mice were selected. Biological genes with an average count $<$ 1 and non-detected technical genes were removed from the analysis. In total, 146 cells (93 naive and 53 activated CD4$^+$ T cells) and 10,553 genes (10,495 biological and 58 technical genes) passed filtering. Each condition contains 2 replicates.

3.3.5 CD4$^+$ T cell differentiation

Non-UMI scRNA-Seq data were generated from CD4$^+$ T cells during differentiation towards Th1 and Tfh cell fates after *Plasmodium* infection [253]. Raw reads were downloaded from ArrayExpress [E-MTAB-4388] and mapped against the *Mus musculus* genome (GRCm38) using *gsnap* [334] with default settings. Read counting was performed using *HTSeq* [336] with default settings.

Quality control was performed by removing cells with fewer than 300,000 biological reads or fewer than 600,000 technical reads at day 2. At days 4 and 7, cells with fewer than 1,000,000 biological reads were excluded from downstream analysis. Additionally, we removed genes that did not show an average detection of more than 1 read at day 2, day 3, day 4 or day 7 after infection. After applying these criteria, 376 cells (Day 0: 16 cells, Day 2: 89, Day 3: 21, Day 4: 133, Day 7: 64, Day 7 non-infected: 53) and 7899 genes (7847 biological and 52 technical) remained for analysis. Note that, due to low sample sizes, we focused our analysis on data from day 2, day 4 and day 7 post-infection.

3.4 The informative prior stabilises parameter estimation

Our joint prior formulation induces a non-linear regression that captures the overall trend between gene-specific over-dispersion parameters δ_i and mean expression parameters μ_i. Thus, we also refer to the extended model induced by this prior as the *regression* BASiCS model. Accordingly, the model induced by the original independent prior specification [295] is referred to as the *non-regression* BASiCS model.

3.4.1 Dataset specificity of the regression trend

To study the performance of the regression BASiCS model, we applied both the regression and non-regression BASiCS model to a variety of scRNA-Seq datasets. Each dataset is unique in its composition, covering a range of different cell types and experimental protocols (see **Section 3.3** and **Table 3.1**). Qualitatively, we observe that the inferred regression trend varies substantially across different datasets **(Fig. 3.5)**, justifying the choice of a flexible semi-parametric approach (see **Section 3.2.3** and **Section 3.2.6**). Moreover, as expected, we observe that residual over-dispersion parameters ε_i are not confounded by mean expression. Additionally, we assessed whether the residual over-dispersion parameter is biased by the percentage of zero counts per gene **(Fig. 3.5, fourth column)**. This feature increases for lowly expressed genes due to technical expression drop-outs. Nevertheless, posterior estimates of gene-specific residual over-dispersion parameters are not confounded by the percentage of zero counts per gene **(Fig. 3.5)**.

Next, we observed that the regression BASiCS model shrinks the posterior estimates for μ_i and δ_i towards the regression trend. This is due to the joint prior specification on $(\mu_i, \delta_i)'$ and is consistent with the shrinkage observed in Love *et al.*, 2014 [371]. The strength of this shrinkage is dataset-specific, being more prominent in sparser datasets with a higher frequency of zero counts and for lowly-expressed genes where measurement error is greatest **(Fig. 3.5A)**.

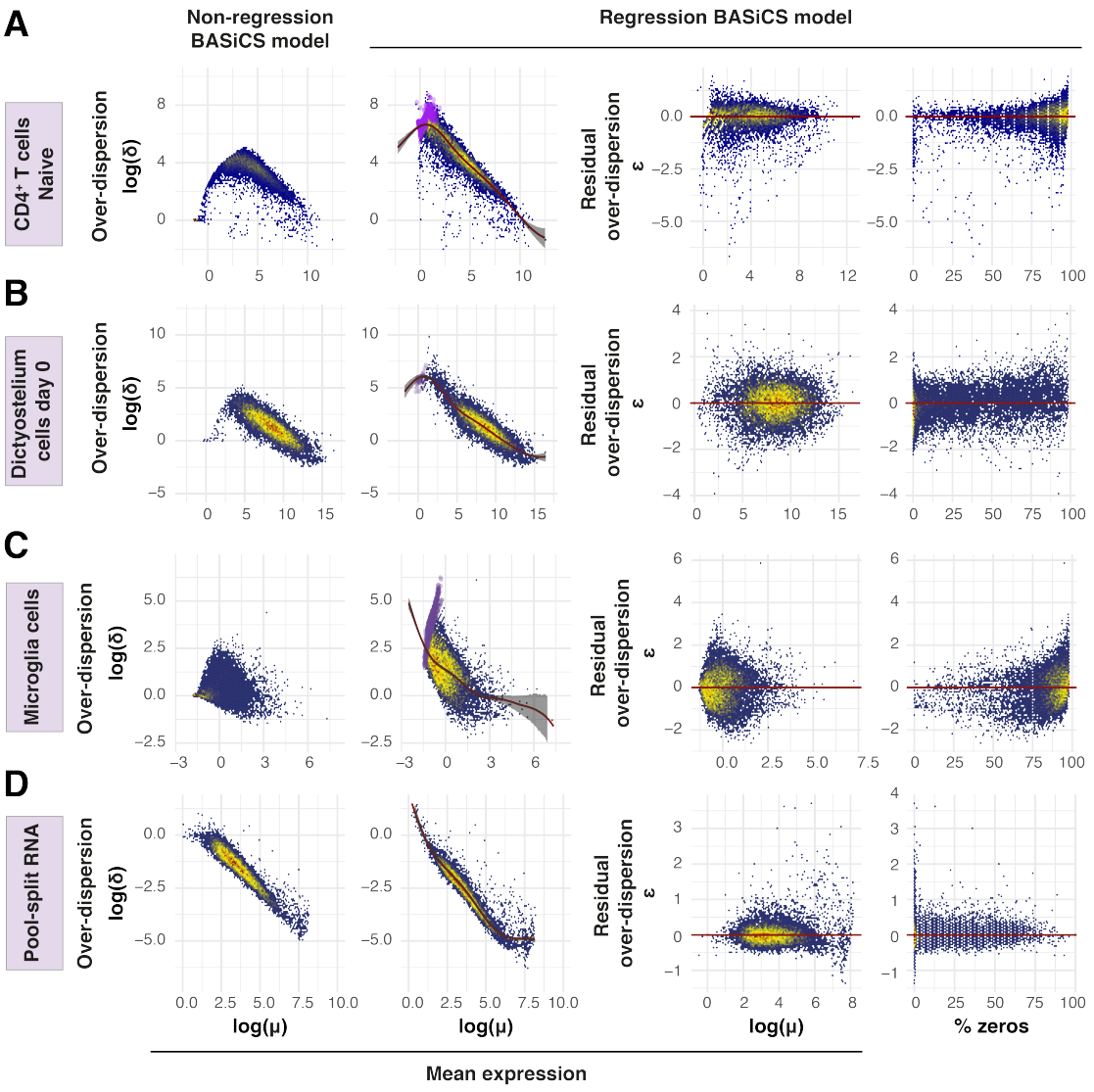


Fig. 3.5: Parameter estimation using a variety of scRNA-Seq datasets.
Model parameters were estimated using the regression and non-regression BASiCS models on **(A)** naive CD4$^+$ T cells [22], **(B)** *Dictyostelium* cells prior to differentiation (day 0) [51], **(C)** microglia cells [259] and **(D)** pool-and-split RNA [134]. These datasets were selected to highlight situations with different levels of sparsity (i.e. the proportion of zero counts, see fourth column). The colour code within the scatterplots is used to represent areas with high (yellow/red) and low (blue) density of genes. **First column:** gene-specific over-dispersion δ_i versus mean expression μ_i as estimated by the non-regression BASiCS model. **Second column:** gene-specific over-dispersion δ_i versus mean expression μ_i as estimated by the regression BASiCS model. The red line indicates the estimated regression trend. Purple dots indicate genes detected (i.e. with at least one count) in less than 2 cells. **Third column:** gene-specific residual over-dispersion ε_i versus mean expression μ_i as estimated by the regression BASiCS model. **Forth column:** gene-specific posterior estimates for residual over-dispersion ε_i parameters versus percentage of zero counts for each gene.

3.4.2 Stabilisation of posterior inference

Next, we asked whether or not the shrinkage introduced by the regression BASiCS model improves posterior inference. To assess this, we compared estimates for gene-specific parameters across (i) different sample sizes and (ii) different gene expression levels. Both the sample size and the level of expression influence posterior estimation of model parameters due to loss of power when few cells are used to estimate parameters for lowly expressed genes. More concretely, we used a large dataset containing 939 CA1 pyramidal neurons [259] (**Section 3.3.2**) to artificially generate smaller datasets by randomly sub-sampling 50-500 cells. For each sample size, parameter estimates were then obtained using both the regression and non-regression BASiCS models. Based on parameter estimates using the non-regression model, we split the genes into three sets: lowly expressed ($\mu_i < 1.89$), medium expressed ($1.89 < \mu_i < 5.37$) and highly expressed ($\mu_i > 5.37$). These cut-off values were chosen such that roughly a third of genes were assigned to each category. The distribution of these estimates is summarised in **Fig. 3.6**.

Firstly, we observe that both the regression and non-regression BASiCS models led to consistent and largely stable mean expression estimates μ_i across different sample sizes and expression levels **(Fig. 3.6A)**. Secondly, in line with the results in **Fig. 3.5**, the main differences between the methods arise when estimating the over-dispersion parameters δ_i **(Fig. 3.6B)**. In particular, we observe that the non-regression BASiCS model appears to underestimate δ_i for lowly expressed genes when the sample size is small (with respect to the parameter estimates obtained based on the full dataset of 939 cells). This is due to the original, non-informative prior: $\delta_i \sim \text{log-N}(0, a_\delta^2)$. In the case of lowly expressed genes, the data is not informative and the over-dispersion parameters are estimated as $\delta_i \approx 0$. In contrast, the shrinkage introduced by our regression BASiCS model aids parameter estimation, leading to robust estimates even for the smallest sample size. This is particularly important for rare cell populations where large sample sizes are difficult to obtain. A similar effect is observed for genes with medium and high expression levels, where the non-regression BASiCS model appears to slightly overestimate δ_i. We also observe that estimates of residual over-dispersion parameters ε_i are stable across sample sizes and expression levels. **Fig. 3.7A-C** summarises 10 replicates of the down-sampling experiment performed in **Fig. 3.6A-C**. We use parameters estimated from the full dataset as *pseudo* ground truth (pgt) values. For each sub-sampling experiment, sample size and gene set, we computed the median $\log_2$ fold change in μ_i and δ_i and the median difference for ε_i between the estimates and the pgt. The median and the

range of these values across 10 sub-sampling experiments is used for visualisation purposes (**Fig. 3.7A-C**).

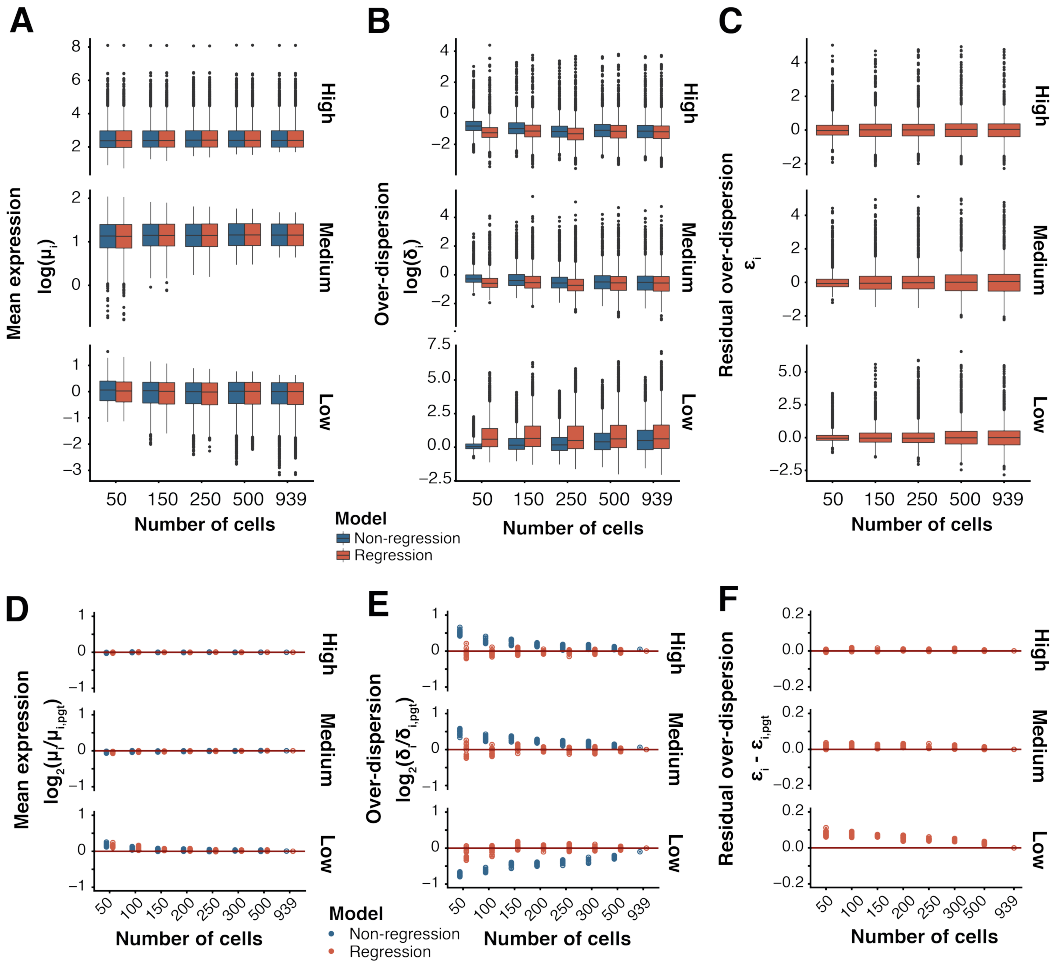


Fig. 3.6: Estimation of gene-specific model parameters for varying sample sizes.
The regression (orange) and non-regression (blue) BASiCS models were used to estimate gene-specific model parameters for lowly (lower panels), medium (mid panels) and highly (upper panels) expressed genes across populations with varying numbers of cells. These were generated by randomly sub-sampling cells from a population of 939 CA1 pyramidal neurons [259]. Extended results based on multiple downsampling experiments are displayed in **Fig. 3.7A-C**. **(A-C)** For a single sub-sampling experiment, boxplots summarise the distribution of gene-specific estimates for (A) mean expression parameters μ_i (log-scale), (B) over-dispersion parameters δ_i (log-scale) and (C) residual over-dispersion parameters ε_i. **(D-F)** For 10 sub-sampling experiments, parameter estimates were compared against a *pseudo* ground truth (pgt). The latter is defined as the parameter estimates obtained for the full population of 939 cells using the regression BASiCS model. For each sub-sampling experiment, gene-specific $\log_2$ fold changes ($\log_2(\mu_i/\mu_{i,pgt})$ and $\log_2(\delta_i/\delta_{i,pgt})$) and distances ($\varepsilon_i - \varepsilon_{i,pgt}$) between the estimates and the pgt were computed. For visualisation purposes, the medians across genes for each sub-sampling experiment are presented,

3.4.3 Validation of gene-specific posterior estimates by smFISH

As an external validation, we compared our posterior estimates of gene-specific model parameters obtained from scRNA-Seq data to empirical estimates from matched smFISH data of mouse embryonic stem cells grown in 2i and serum media [134]. Firstly, posterior estimates of mean-expression parameters μ_i exhibit high correlation to smFISH mean transcript counts **(Fig. 3.7D)**. Secondly, we also observe a strong correlation between posterior estimates for over-dispersion parameters δ_i and the empirical CV^2 values obtained from smFISH data **(Fig.3.7E)**. Finally, a similar behaviour is observed when comparing posterior estimates of residual over-dispersion parameters ε_i to a residual CV^2 **(Fig.3.7F)**. As in Brennecke *et al.*, 2013 [10], to obtain residual CV^2 values for the smFISH data, we fitted a gamma generalised linear model with identity link (*glmgam.fit* of the *statmod* package in R) between the CV^2 and the reciprocal log-transformed mean transcript counts.

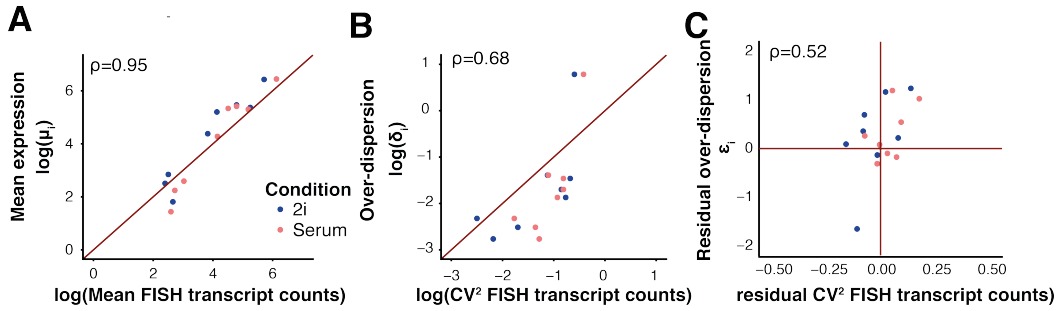


Fig. 3.7: Stability of posterior estimates for gene-specific parameters.
(A-C) Matched scRNA-Seq and smFISH data measured on mouse embryonic stem cells grown in 2i and serum media [134] was used to validate the performance of the regression BASiCS model. Gene-specific parameter estimates obtained by the regression BASiCS model were compared against empirical estimates calculated based on smFISH data. This comparison includes 8 genes, measured in both conditions. Pearson's correlation is indicated for each comparison, **(D)** Estimates of mean expression parameters μ_i (log-scale) are plotted against mean transcript count (smFISH), **(E)** Estimates of over-dispersion parameters δ_i (log-scale) are plotted against the squared coefficient of variation (CV^2) of transcript counts (smFISH), **(F)** Estimates for residual over-dispersion parameters ε_i are compared against residual estimates of variability estimated for the smFISH data.

3.5 Expression variability during immune responses

Here, we illustrate how the regression BASiCS model assesses changes in expression variability using CD4$^+$ T cell activation and differentiation. For all datasets, pre-processing steps are described in **Section 3.3**.

3.5.1 Testing variability changes upon immune activation

As described in the previous chapter, the non-regression BASiCS model only allows the assessment of changes in variability for genes that remain stable in gene expression across conditions. Here, we extend the previous analysis and test for changes in variability in parallel to changes in mean expression.

To identify gene expression changes during early T cell activation, we compared CD4$^+$ T cells before (naive) and after (active) 3 hours of stimulation with CD28 and CD3ε antibodies (see **Section 2.4** and [22]). For both conditions, we ran the regression BASiCS model independently and performed differential mean expression and differential variability testing using the residual over-dispersion parameters. Testing changes in variability through residual over-dispersion is performed across all genes, including the large set of genes that are up-regulated upon immune activation **(Fig. 2.8)**. The latter include immune-response genes and critical drivers for CD4$^+$ T cell functionality that had to be excluded from analysis in the previous chapter.

Comparison between the regression and non-regression BASiCS model

Firstly, we compared the results obtained by the regression BASiCS model to those presented in **Section 2.4**. To allow for a direct comparison of the results, the same inclusion criteria as in the previous chapter is adopted, i.e. we excluded genes with low mean expression ($\mu_i < 50$) in both conditions from testing. Moreover, the minimum tolerance thresholds were also adapted to match the choices in **Section 2.4**. To detect differentially expressed genes, a minimum tolerance threshold $\tau_0 = 2$ was used **(Fig. 3.8A)**. To compare the detection of differentially over-dispersed genes, we performed differential mean expression testing using a stringent minimum tolerance threshold $\tau_0 = 0$ for both models (this is to avoid the results being confounded by changes in mean, see upper panel in **Fig. 3.8B**). For the 463 genes that are detected as non-differentially expressed by both models for this threshold, a total of 111 genes are detected as differentially over-dispersed by either model (minimum tolerance $\log_2$

fold change threshold $\omega_0 = \log_2(1.5) = 0.58$). Out of this set, 93 genes (~83%) are detected as differentially over-dispersed by both models (see lower panel in **Fig. 3.8B**).

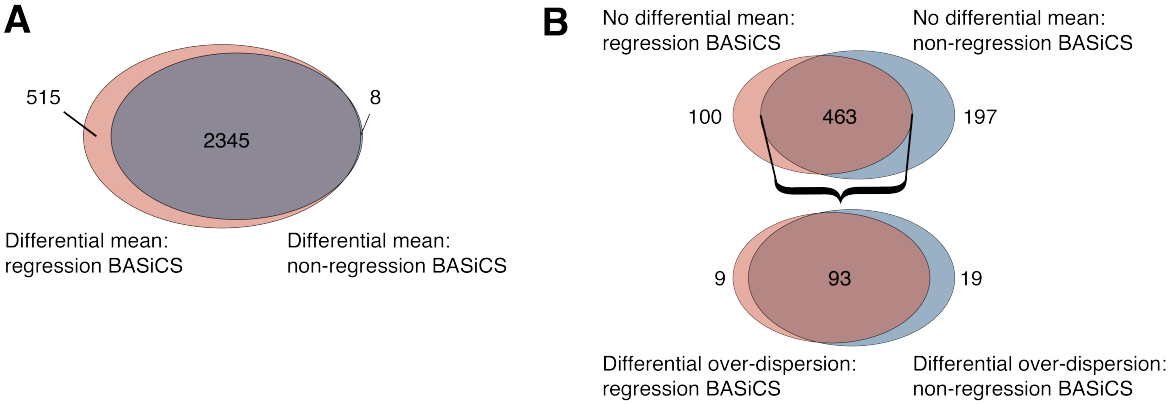


Fig. 3.8: Differential testing comparison between the regression and non-regression BASiCS model.
(A)-(B) Results of differential testing between naive and activated CD4$^+$ T cells were compared between the regression and non-regression BASiCS models. As in **Section 2.4**, genes with low mean expression ($\mu_i < 50$) in both conditions were excluded from testing. **(A)** Overlap of differentially expressed genes (mean) using a minimum tolerance threshold $\tau_0 = 2$ obtained by the regression and non-regression BASiCS models (EFDR = 10%), **(B)** Upper panel: overlap of genes detected as non-differentially expressed using a stringent minimum tolerance threshold $\tau_0 = 0$ obtained by the regression and non-regression BASiCS models (EFDR = 10%). Lower panel: overlap of differentially over-dispersed genes using a minimum tolerance threshold $\omega_0 = \log_2(1.5)$ obtained using the regression and non-regression BASiCS models for the 463 genes detected as non-differentially expressed by both models (EFDR = 10%).

Differential testing during immune activation

For further analyses in this chapter (and in contrast to the previous chapter), we exclude genes whose estimated mean expression parameter μ_i was below 1 from the differential testing. Furthermore, a $\log_2$ fold change threshold $\tau_0 = 1$ was adopted for mean expression testing. Unlike the more stringent threshold used in the previous chapter ($\tau_0 = 2$), this choice allows us to detect more subtle changes in mean expression. Moreover, the default threshold $\psi_0 = 0.41$ was used for differential variability testing. The EFDR was controlled to 10%. By using these thresholds, our model classifies genes into four categories based on their expression dynamics: down-regulated upon activation with (i) lower and (ii) higher variability, and up-regulated with (iii) lower and (iv) higher variability **(Fig. 3.9A)**.

Genes with up-regulated expression upon activation and decreased expression variability encode components of the splicing machinery (e.g. *Sf3a3*, *Plrg1*), RNA polymerase subunits

(e.g. *Polr2l*, *Polr1d*) as well as translation machinery components (e.g. *Ncl*, *Naf1*) (see **Fig. 3.9B**). These biosynthetic processes help naive T cells to rapidly enter a programme of proliferation and effector molecule synthesis [374, 375]. Therefore, rapid and uniform up-regulation of these transcripts would assist such processes. This observation also confirms our previous findings that the translational machinery is tightly regulated during early immune activation (see **Section 2.4**).

In contrast, genes with up-regulated expression and increased expression variability (see **Fig. 3.9C**) include the death-inducing and inhibitory transmembrane ligands Fas ligand (*Fasl*) and programmed death-ligand 1 (PD-L1) (gene symbol: *Cd274*), the regulatory transcription factor Smad3 (*Smad3*), and the TCR-induced transcription factor, Oct2 (*Pou2f2*). Additionally, we detect a heterogeneous up-regulation in the mRNA expression of the autocrine/paracrine growth factor Il2 (*Il2*) upon immune activation. This is in line with previous reports of binary Il2 expression within a population of activated T cells, which has been suggested to be necessary for a scalable antigen response [67]. Heterogeneity in expression of these genes suggests that, despite their uniform up-regulation of biosynthetic machinery, the T cells in this early activation culture represent a mixed population with varying degrees of activation and/or regulatory potential.

For each of these gene sets, functional annotation analysis was performed using all tested genes as background. The functional annotation clustering tool in DAVID [344] was used to cluster annotation categories based on similarity and sort them according to their enrichment score. Here, we list the top 3 functional annotation clusters per gene set and their corresponding enrichment score (ES):

- **Down-regulated with lower variability:** Pleckstrin homology domain (ES = 1.57), G protein signalling (ES = 1.51), glycosidase (ES = 1.49),
- **Down-regulated with higher variability:** Ankyrin repeat-containing domain (ES = 2.19), GTPase mediated signalling (ES = 1.51), steroid biosynthesis (ES = 0.89),
- **Up-regulated with lower variability:** RNA polymerase (ES = 1.6), RNA binding (ES = 1.53), splicing (ES = 1.41),
- **Up-regulated with higher variability:** Cytokine-cytokine receptor interaction (ES = 1.65), WD40 repeat (ES = 1.22), transcription (ES = 1.18).

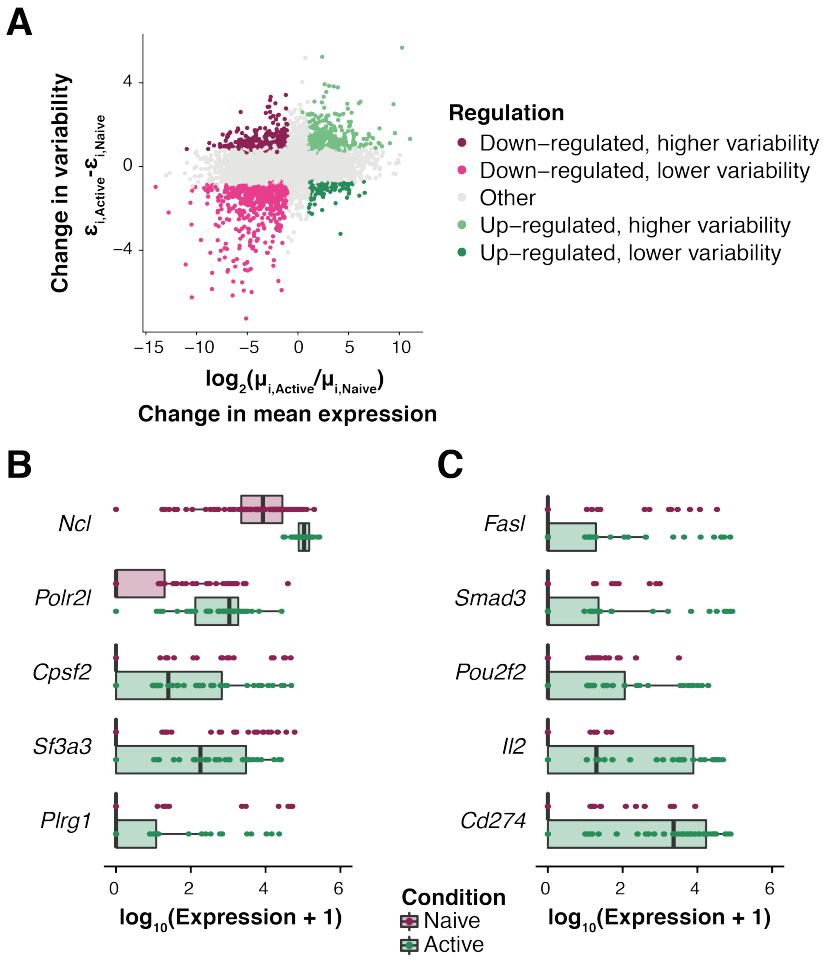


Fig. 3.9: Changes in expression patterns during early immune activation.
Differential testing (mean and residual over-dispersion) was performed between naive and activated murine CD4$^+$ T cells taken from the previous chapter. This analysis uses a minimum tolerance threshold of $\tau_0 = 1$ for changes in mean expression and a minimum tolerance threshold of $\psi_0 = 0.41$ for differential residual over-dispersion testing (expected false discovery rate is fixed at 10%). **(A)** For each gene, the difference in residual over-dispersion estimates (Active - Naive) is plotted versus the $\log_2$ fold change in mean expression (Active/Naive). Genes with statistically significant changes in mean expression and variability are coloured based on their regulation (up/down-regulated, higher/lower variability), **(B-C)** Normalised expression counts across the naive (purple) and active (green) CD4$^+$ T cell population are visualised for representative genes that (B) increase in mean expression and decrease in expression variability and (C) increase in mean expression as well as expression variability upon immune activation. Each dot represents a single cell.

Effect of expression outliers on changes in variability

We observe that for some genes (e.g. *Plrg1*), changes in variability are driven by a small number of outlier cells with high expression. The interpretation of these results is not trivial as it could reflect very subtle sub-structure or genuine changes in variability. To explore this, we performed the following synthetic experiment: We artificially created a mixed population of cells by combining 5 activated CD4$^+$ T cells with a population of 93 naive CD4$^+$ T cells therefore simulating expression outliers. Subsequently, we performed a differential testing (mean and residual over-dispersion) between this mixed population and a *pure* population of 93 naive CD4$^+$ T cells. As expected, this analysis shows an overall increase in variability in the mixed population. For example, among the genes that exhibit higher mean expression and higher residual over-dispersion in the mixed population, we found *Il2* which is up-regulated upon CD4$^+$ T cell activation **(Fig. 3.10A)**. Moreover, we observe that the genes in this category are enriched for those that are only expressed in the 5 activated CD4$^+$ T cells **(Fig. 3.10B)**. This result suggests that differential variability testing can potentially uncover markers for heterogeneous cell states or cell types that can provide important biological insights. However, changes in residual over-dispersion that are driven by outliers can also reflect unwanted contamination (e.g. mixed cell types), hence careful data filtering and clustering analysis should be performed prior to differential variability testing.

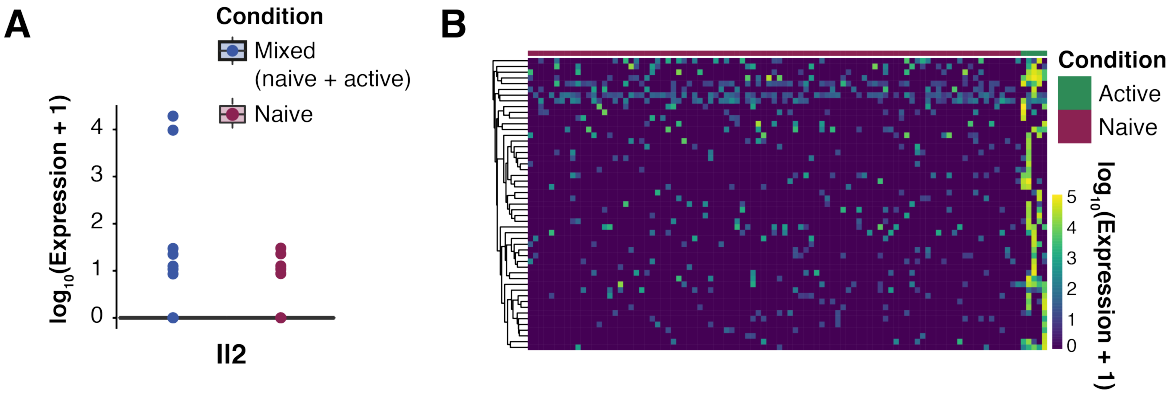


Fig. 3.10: Dissecting changes in variability driven by expression outliers.
5 activated CD4$^+$ T cells were combined with a population of 93 naive CD4$^+$ T cells. **(A)** Distribution of normalised expression counts for *Il2* in a population of naive CD4$^+$ T cells (red) and the mixture population representing a mix of 93 naive and 5 activated CD4$^+$ T cells (blue). Each dot represents a single cell, **(B)** Genes with increased mean expression and increased variability in the mixed population were detected. The heatmap shows normalised expression counts for these genes across the mixed population (93 naive and 5 activated CD4$^+$ T cells).

In summary, our approach allows us to extend the findings from the previous chapter, dissecting immune-response genes into two functional sets: (i) homogeneous up-regulation of biosynthetic machinery components and (ii) heterogeneous up-regulation of several immunoregulatory genes.

3.5.2 Expression dynamics during *in vivo* CD4$^+$ T cell differentiation

In contrast to the quick transcriptional switch that occurs within hours of naive T cell activation, transcriptional changes during cellular differentiation processes are more subtle and were found to be coupled with changes in variability prior to cell fate decisions [15, 52]. Here, we apply our method to study changes in expression variability during CD4$^+$ T cell differentiation after malaria infection using the dataset introduced by Lönnberg *et al.*, 2017 [253]. In particular, we focus on samples collected 2, 4 and 7 days post-malaria infection, for which more than 50 cells are available. The BASiCS model was run for 40,000 iterations independently for each condition.

Changes in variability over the differentiation time course

First, we studied global changes in over-dispersion along the differentiation time course by comparing posterior estimates for the gene-specific over-dispersion parameter δ_i, focusing on 126 genes for which mean expression does not change **(Fig. 3.11A)**. These genes were detected by testing changes in mean expression using a stringent threshold ($\tau_0 = 0$) between day 2 and day 4 as well as between day 4 and day 7. Genes that are not detected as differentially expressed in both tests were considered for variability analysis. We found that the expression of these genes is most tightly regulated at day 4, when cells are in a highly proliferative state. Moreover, between day 4 and day 7, the cell population becomes more heterogeneous. This is in line with the emergence of differentiated Th1 and Tfh cells that was observed by Lönnberg *et al.*, 2017.

Next, we exploited the residual over-dispersion parameters to identify changes in variability (irrespective of changes in mean expression) between consecutive time points. For this, we performed differential variability testing using the default threshold on changes in the residual over-dispersion parameter ($\psi_0 = 0.41$) between day 2 and day 4 as well as between day 4 and day 7. After testing, we excluded all genes that are expressed in fewer than 2 cells in at least one time point from down-stream analysis. Separating the remaining genes by whether their variability increases or decreases between time points revealed four different

patterns **(Fig. 3.11B)**. These include genes whose variability systematically increases (or decreases) as well as patterns where variability is highest (or lowest) at day 4.

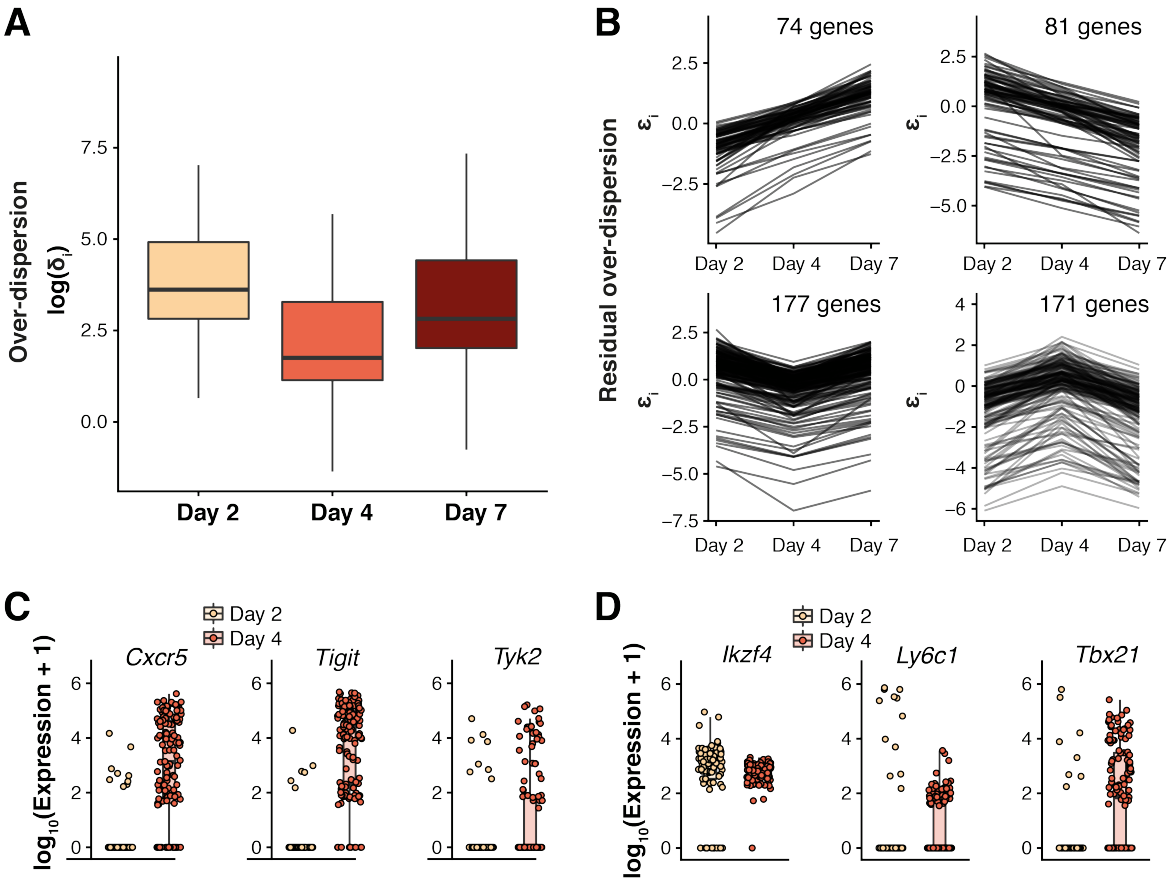


Fig. 3.11: Dynamics of expression variability throughout CD4⁺ T cell differentiation. Analysis was performed on CD4⁺ T cells assayed 2 days, 4 days and 7 days after *Plasmodium* infection. Changes in residual over-dispersion were tested using a minimum tolerance threshold of $\psi_0 = 0.41$ (EFDR is fixed at 10%), **(A)** Distribution of posterior estimates of over-dispersion parameters δ_i for genes that exhibit no changes in mean expression across the differentiation time course. Changes in mean expression were tested using a minimum tolerance threshold of $\tau_0 = 0$ (expected false discovery rate is fixed at 10%), **(B)** Posterior estimates for residual over-dispersion parameters ε_i, focusing on genes with statistically significant changes in expression variability between time points. Gene set size is indicated for each plot, **(C-D)** Normalised expression counts across cell populations at day 2 (yellow) and day 4 (red) post infection are visualised for representative genes that (C) increase or (D) decrease in variability during differentiation. Each dot represents a single cell.

Opposing expression dynamics of lineage-defining marker genes

The differential variability analysis between day 2 and day 4, revealed changes in expression variability for a set of immune-related genes **(Fig. 3.11C-D)**. For example, expression of

C-X-C chemokine receptor type 5 (*Cxcr5*) which encodes the chemokine receptor that directs Tfh cells to the B cell follicles [376], strongly increases in variability on day 4. This finding agrees with results from Lönnberg *et al.*, 2017 [253], where Tfh and Th1 differentiation was observed to be transcriptionally detectable at day 4 within a subset of activated cells. A similar behaviour was observed for tyrosine kinase 2 (*Tyk2*) and T cell immunoreceptor with Ig And ITIM domains (*Tigit*). The latter encodes a receptor that is expressed by a subset of Tfh cells and that was found to promote Tfh function [377]. In contrast, we observe a decrease in variability between day 2 and day 4 for IKAROS Family Zinc Finger 4 (*Ikzf4*) (Treg-associated gene), *Ly6c1* (expressed by effector T cells) and *Tbx21* (encoding the Th1 lineage-defining transcription factor Tbet).

To achieve a broader view on changes in variability within sets of genes that drive this differentiation process, we selected gene sets listed in Lönnberg *et al.*, 2017 to visualise their changes in mean expression and residual over-dispersion [253]. The first set of genes is taken from Figure 3E of the original publication, which filtered genes based on their association with the bifurcation of Th1 and Tfh differentiation. The second set of genes with sequential peak expression over pseudo-time is taken from Figure 5A of the original publication, which were selected based on immunological relevance from a list of dynamic genes during *in vivo* differentiation. For the genes that were detected to be lineage-associated, we detected a continuous increase in expression of Th1-associated genes but not Tfh-associated genes **(Fig. 3.12A)**, with the majority of changes in variability for these genes occurring between day 2 and day 4.

Finally, we examined immune-related genes (*Il2ra, Tbx21, Il2rb, Cxcr5, Selplg, Id2, Ifng, Icos, Ifngr1*) that were previously described as showing differences in their peak expression over the pseudo time course of differentiation [253] **(Fig. 3.12B)**. From this list, the lineage-associated genes *Tbx21* and *Cxcr5* are up-regulated between days 2 and 4. However, these genes exhibit opposite behaviours in terms of variability: *Cxcr5* increases and *Tbx21* decreases in variability between day 2 and day 4 **(Fig. 3.12C)**. The fact that variability of *Tbx21* (Tbet) expression was highest on day 2 suggests that Tbet is up-regulated very early in differentiation, as seen in [253] and similar to *in vitro* Th1 induction [378]. Moreover, this suggests that Th1 fate decisions (for at least a subset of cells) may be made even earlier than the differentiation bifurcation point identified on day 4 by the original study [253].

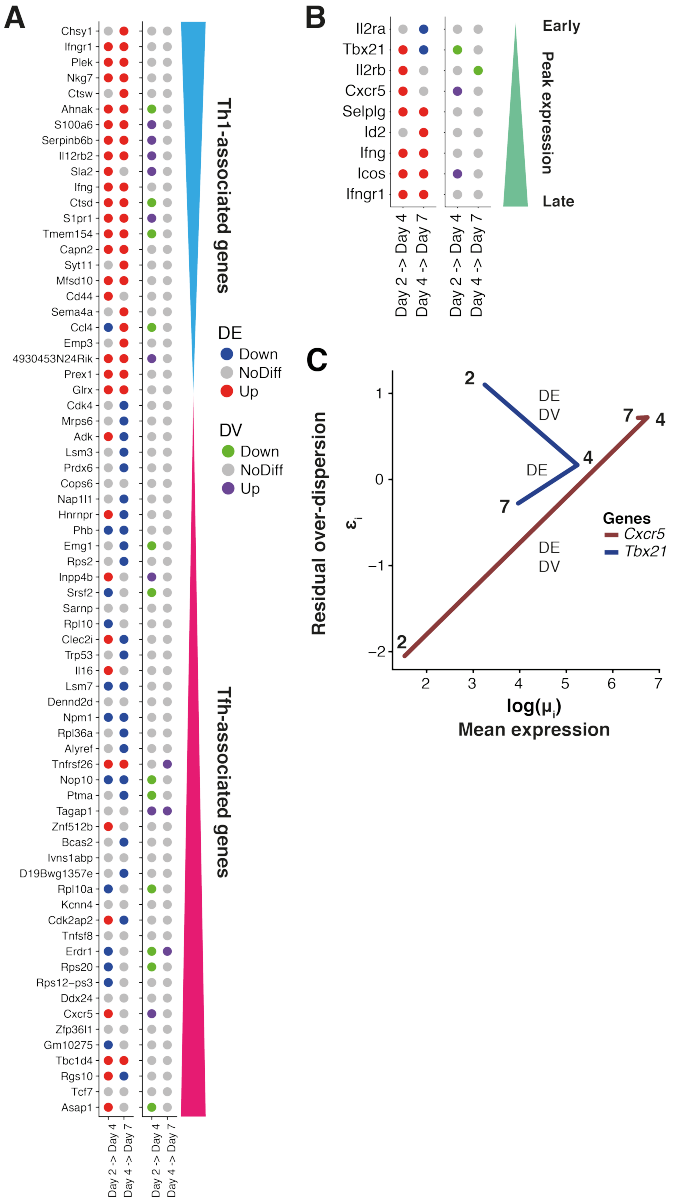


Fig. 3.12: Differential regulation of Th1- and Tfh-associated genes across the differentiation process.
Differential mean expression testing (minimum tolerance threshold $\tau_0 = 1$) and differential residual over-dispersion testing (minimum tolerance threshold $\psi_0 = 0.41$) was performed on cell populations between day 2 and day 4 as well as day 4 and day 7 controlling the EFDR to 10%. Genes that increase in expression over time are marked with a red dot while genes that decrease in expression over time are marked with a blue dot. Similarly, genes that increase in variability over time are marked in purple while genes that decrease in variability over time are marked in green. Only genes that pass filtering are visualised. **(A)** Differential testing results are visualised for Th1- and Tfh-associated genes taken from Figure 3E in Lönnberg *et al.*, 2017 [253]. Genes are ordered based on their correlation with the Th1 trend assignment (top to bottom) or their correlation to Tfh trend assignment (bottom to top), **(B)** Differential testing results are visualised for important genes during CD4$^+$ T cell differentiation taken from Figure 5A in [253]. Genes were ordered based on their peak expression point in pseudo-time as defined by Lönnberg *et al.*, 2017 [253], **(C)** *Tbx21* (blue) and *Cxcr5* (red) measured at day 2, day 4 and day 7 post-infection. Posterior estimates for residual over-dispersion parameters ε_i are plotted against posterior estimates for mean expression parameters μ_i. Statistically significant changes in mean expression (DE, minimum tolerance threshold of $\tau_0 = 1$) and variability (DV, minimum tolerance threshold of $\psi_0 = 0.41$) are indicated for each comparison

3.6 Application to droplet-based scRNA-Seq data

Declaration In the context of expanding the BASiCS framework to test changes in variability independent of mean expression, Catalina A. Vallejos (The Alan Turing Institute/MRC Institute of Human Genetics/University of Edinburgh) developed an approach where technical variation was quantified by borrowing information across multiple replicates. This avoids the use of technical spike-in genes when droplet-based scRNA-Seq data is analysed. This approach is part of the publication:

Nils Eling, Arianne C. Richard, Sylvia Richardson, John C. Marioni, Catalina A. Vallejos. Robust expression variability testing reveals heterogeneous T cell responses. *Cell Systems*, In press, 2018

Here, I will not describe this approach in detail as this has not been my own work. However in the context of this section, it integrates with my contribution to assess mean-independent changes in variability for droplet-based scRNA-Seq data.

With the development of droplet-based scRNA-Seq technologies, the number of cells that can be profiled per experiment strongly increased at the cost of lower sequencing depth per cell [164, 163, 178]. Furthermore, these technologies exclude the use of spike-in RNA to measure technical variation, which is essential for quantifying technical variation using the original BASiCS model [11, 295]. To ensure the broad applicability of the BASiCS model, both the regression and non-regression models have been expanded to handle datasets without spike-in genes. For this purpose, principles of measurement error models were exploited, where — in the absence of gold standard features — technical variation is quantified through *replication* [379]. This horizontal data integration approach is based on experimental designs where cells from a population are randomly allocated to multiple independent experimental replicates (batches). In such an experimental design, the no-spikes implementation of BASiCS assumes that biological effects are shared across batches and that technical variation will be reflected by spurious differences. As shown in the publication, posterior inference under the no-spikes BASiCS model closely matches the original implementation for datasets where spike-ins and batches are available. Technical details about the no-spikes implementation of BASiCS are discussed in the original publication (see **Declaration**).

3.6.1 Differential testing using somitic and pre-somitic mesoderm cells

To test the applicability of the regression BASiCS model to droplet-based scRNA-Seq data, I analysed cells isolated from mouse embryos at E8.25 [27]. Ibarra-Soria *et al.*, 2018 analysed more than 20,000 cells to identify the major cell types following gastrulation. Key findings included a spatial sub-structure within the foregut, detection of oscillating gene expression patterns during somitogenesis and the contribution of the leukotriene pathway to blood formation. I selected the cells identified as presomotic mesoderm and somitic mesoderm as test populations as they reside in contrasting differentiation stages. Somitogenesis is a rhythmic and sequential differentiation process from pre-somitic mesoderm (PSM) cells to mature somites, which later will give rise to bone, muscle and skin of the adult body. Throughout this process, oscillating gene expression patterns control the differentiation of PSM into somitic mesoderm (SM). Driving factors for this are Wnt and fibroblast growth factor (Fgf) signalling on the side of the PSM and retinoic acid signalling in somites [380]. The regression BASiCS model can now further dissect transcriptional regulation between these groups of cells.

Data processing

The raw counts data and assigned cluster labels can be obtained from ArrayExpress [E-MTAB-6153] [27]. I selected cells labelled as pre-somitic mesoderm and somitic mesoderm for further down-stream analysis and removed lowly expressed genes (< 0.1 reads on average). For visualisation purposes, I normalised the data using *scran* by pooling cells within each cell type. PCA computed on the log-transformed normalised counts shows a clear separation between these two cells types with an additional intermediate cell type labelled as 'presomiticmesoderm.b'. I excluded this small set of cells as well as outlying cells for which $\text{PC2} < -5$ **(Fig. 3.13A)**. After filtering, I obtained 791 pre-somitic mesoderm cells and 670 somitic mesoderm cells for further analysis. For each of the two populations, the MCMC was run for 20,000 iterations separately. After posterior inference was completed, I first confirmed that the model estimated the regression trend between the over-dispersion parameters δ_i and mean expression parameters μ_i correctly. For both conditions, the regression trend captures the full range of data points and differential testing can be performed **(Fig. 3.13B)**. For differential testing, I used a threshold of $\tau_0 = 1$ to assess changes in mean expression and the default threshold $\psi_0 \approx 0.41$ to test changes in residual over-dispersion.

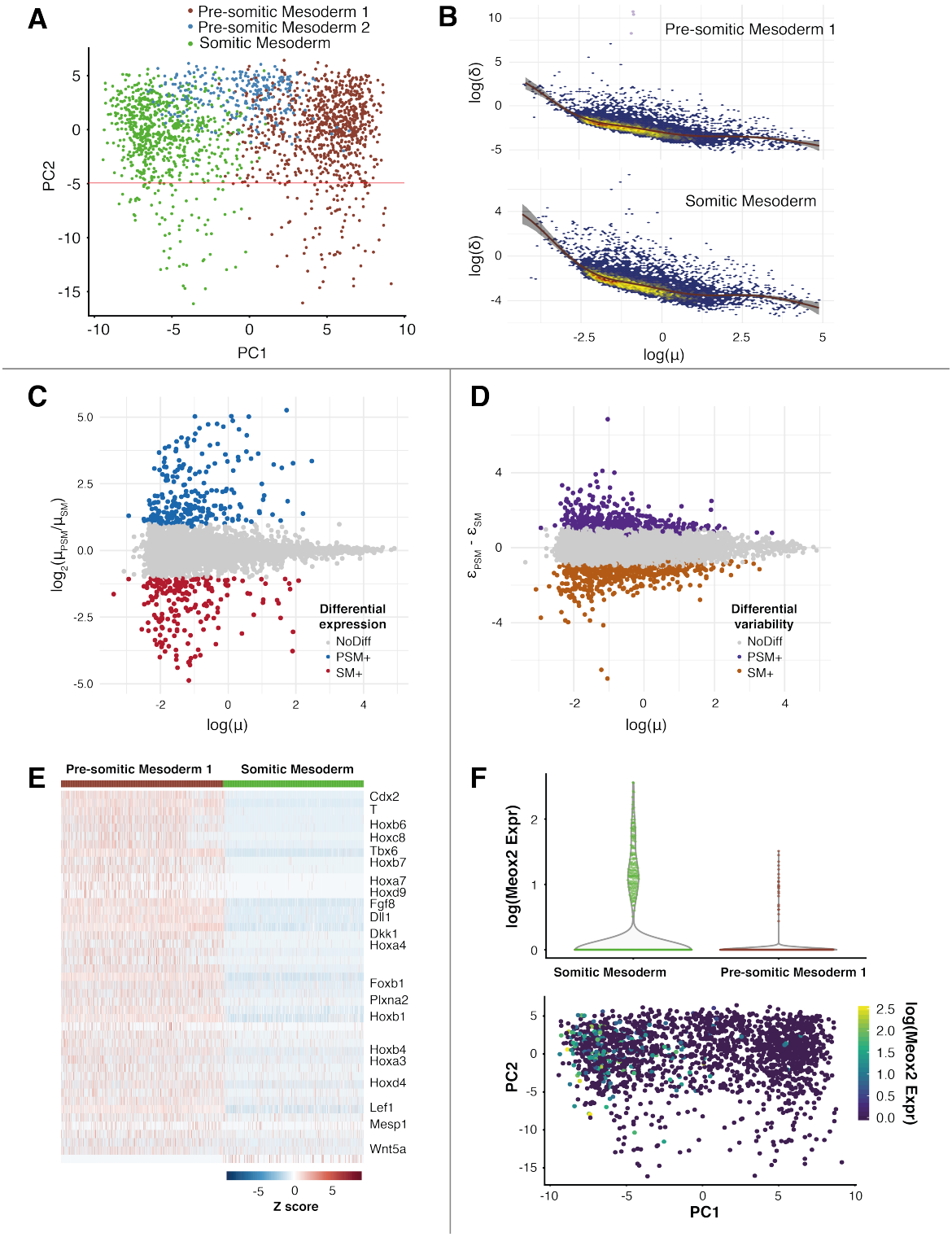


Fig. 3.13: Quantification of expression dynamics from droplet-based scRNA-Seq data (Full legend on next page).

Fig. 3.13: Quantification of expression dynamics from droplet-based scRNA-Seq data (continued).
(A) Somitic (SM) and pre-somitic mesoderm (PSM) cells from droplet-based scRNA-Seq data [27] were selected and visualised via a PCA. Colour labelling was done based on the cluster annotation taken from the original publication. For down-stream analysis cells with PC2 < -5 and marked as 'presomiticmesoderm.b'/Pre-somitic mesoderm 2 were removed, **(B)** For each condition, over-dispersion estimates δ_i were plotted against mean expression estimates μ_i. The regression trend is indicated as red line, **(C)** Posterior estimates for the $\log_2$ fold change in mean expression between PSM and SM were plotted against mean expression averaged across the two populations. Differentially expressed genes are coloured based on their regulation: blue: PSM-specific (PSM+), red: SM-specific (SM+), **(D)** Posterior estimates for differences in residual over-dispersion between PSM and SM were plotted against mean expression averaged across the two populations. Differentially variable genes are coloured based on their regulation: purple: PSM-specific (PSM+), brown: SM-specific (SM+), **(E)** Heatmap showing the Z score scaled gene expression of pre-somitic mesoderm-specific genes of the GO category GO:0009952: anterior/posterior pattern specification. Genes were ordered based on their $\log_2$ fold change in expression from highest to lowest, **(F)** Gene expression of *Meox2* in PSM and SM. This gene was detected to be heterogeneously up-regulated in SM. Upper panel: violin plots showing distribution of log-normalised expression counts. Lower panel: *Meox2* expression across the PCA from **(A)**.

Changes in mean expression during somitogenesis

I first tested changes in mean expression between SM and PSM and detected 203 SM-specific genes and 236 PSM-specific genes. Based on these cell type-specific gene lists, I performed GO analysis using the Bioconductor *goseq* package while correcting for gene length biases. As expected, top enriched categories for SM-specific genes include 'animal organ morphogenesis' and 'skeletal system development' which contain genes such as *Bmp7*, *Fgfr2*, *Gata6* and *Meox2*. Somites are rhythmically formed from the PSM and are embryonic precursors for vertebrae and skeletal muscles [381]. On the other hand, top categories for PSM-specific genes include 'pattern specification process', 'somitogenesis' and 'Wnt signaling pathway'. It is known that the posterior end of the PSM is high in Wnt and Fgf signalling that determines the oscillation dynamics of individual cells during somitogenesis [380]. **Fig. 3.13E** visualises the PSM-specific gene expression of the GO category: GO:0009952 - anterior/posterior pattern specification. This category includes contributors to embryonic patterning: *Fgf8* [382], Wnt signalling components (e.g. *Wnt5a*, *Dkk1*), Notch signalling components (e.g. *Dll1*) [381] and several members of the Hox gene family which control embryonic patterning [383].

Changes in variability during somitogenesis

Next, I tested changes in variability between SM and PSM and categorised genes based on their regulation as in **Section 3.5.1**. These categories include genes with higher expression and higher variability in SM, higher expression and lower variability in SM, higher expression and higher variability in PSM and higher expression and lower variability in PSM. Interestingly, I detected the mesenchyme homeobox 2 (*Meox2*) gene to be heterogeneously up-regulated in somitic mesoderm when compared to the precursor mesoderm **(Fig. 3.13F)**. *Meox2* has been shown to regulate somite morphogenesis, patterning and differentiation specifically in the sclerotome (forming the vertebrae and rib cartilage) alongside its paralog *Meox1*. While knocking out *Meox1* in mice only shows mild defects in vertebrate and rib bones, the knockout of *Meox2* induces defective differentiation and morphogenesis of the limb muscles [384]. The heterogeneous, but unstructured **(Fig. 3.13F)**, expression of *Meox2* might therefore indicate the identity of early progenitor cells that later on differentiate to form muscles of the limbs. Similarly, I find doublesex and mab-3 related transcription factor 2 (*Dmrt2*) heterogeneously up-regulated in SM compared to PSM. This transcription factor has been implicated in somite development with specific expression in the dermomyotome, the part of the mesoderm that forms skin. The homozygous loss of *Dmrt2* leads to severe somite patterning defects at E10.5 and mice die shortly after birth [385]. The differential variability analysis revealed an early and heterogeneous expression of *Dmrt2* in SM which leads to the possible identification of dermomyotome progenitor cells.

In sum, I confirmed that the regression BASiCS model can be applied to droplet-based scRNA-Seq data to assess changes in transcriptional variability between conditions. This is an important validation for the next chapter, where I apply the regression BASiCS model to continuous droplet-based scRNA-Seq data to study changes in variability during spermatogenesis.

3.7 Discussion

This chapter addressed a statistical problem that obstructed the analysis of the previous chapter, where changes in expression variability were confounded by changes in mean expression. Here, this problem is resolved by introducing a hierarchical Bayesian formulation that extends the BASiCS framework. In this formulation, an additional set of residual over-dispersion parameters ε_i that are not confounded by changes in mean expression are introduced. This extension ensures a broader applicability of the BASiCS software and allows statistical testing of changes in variability that are not confounded by mean expression.

The original implementation of BASiCS for testing changes in mean expression and expression variability aimed at borrowing information across all cells, all genes and both conditions. However, as cell-specific scaling factors and gene-specific mean expression parameters are jointly inferred by BASiCS, this led to an unidentifibility problem: a global offset (associated to global differences in mRNA content) between the two conditions is unknown a priori. This issue did not arise in most differential expression methods for bulk RNA-seq data (such as DESeq2 [371] and edgeR [386]), because they typically adopted a sequential approach: they first estimate cell-specific scaling factors that are subsequently used as fixed offsets in the downstream model for differential expression. Instead, BASiCS resolves the unidentifiability issue by performing inference on both conditions independently, performing a post-hoc offset correction.
This approach raises the question if parameter estimates are comparable between both conditions or if differences are cause due to over-fitting. We therefore validated the robustness of the model in multiple ways:

1. We ran the model on down-sampled cell populations and found that (i) the majority of model parameters show similar estimates independent of the sample size and expression levels, and (ii) the regression stabilises variability estimates for lowly expressed genes.
2. We simulated data and showed that (i) the false positive rate is less than 5% for mean and variability estimates and that (ii) the true positive rate reaches 100% for large sample size.
3. we compared BASiCS parameter estimates to absolute transcript counts using smFISH and observed a near perfect estimation of mean expression but lower correlation for variability and residual variability measures.

We are therefore confident that BASiCS correctly estimates mean expression and expression variability per condition and that this allows comparison of parameter estimates between conditions.

Other, dataset-specific factors that affect the robustness of posterior parameter estimation include sequencing depth, number of cells and heterogeneity within the population. BASiCS has been designed to be run on homogeneous populations of cells to avoid variation introduced by correlated sub-structure within the data. Furthermore, the use of UMIs for quantifying transcript counts can also improve variability estimation [134].

Our method enables the characterisation of the extent and nature of variable gene expression in $CD4^+$ T cell activation and differentiation. Firstly, we observe that during acute activation of naive T cells, genes of the biosynthetic machinery are homogeneously up-regulated, while specific immune-related genes are more heterogeneously up-regulated. In particular, increased variability in expression of the apoptosis-inducing Fas ligand [387] and the inhibitory ligand PD-L1 [388] suggests a mechanism by which newly activated cells might suppress re-activation of effector cells, thereby dynamically modulating the population response to activation. Likewise, more variable expression of *Smad3*, which translates inhibitory transforming growth factor (Tgf)β signals into transcriptional changes [389], may indicate increased diversity in cellular responses to this signal. Increased variability in *Pou2f2* (Oct2) expression after activation suggests heterogeneous activities of the NF-κB and/or nuclear factor of activated T cells (NFAT) signalling cascades that control its expression [390]. Moreover, we detect up-regulated and more variable *Il2* expression, suggesting heterogeneous Il2 protein expression, which is known to enable T cell population responses [67].
Additionally, we studied changes in gene expression variability during $CD4^+$ T cell differentiation towards a Th1 and Tfh cell state over a 7 day time course after *in vivo* malaria infection [253]. Our analysis provides several insights into this differentiation system. Firstly, we observe a tighter regulation in gene expression among genes that do not change in mean expression during differentiation at day 4. At this point, divergence of Th1 and Tfh differentiation was previously identified [253]. The decrease in variability on day 4 is potentially due to induction of a strong pan-lineage proliferation programme. However, we observe that not all genes follow this trend and uncover four different patterns of variability changes. Secondly, we observe that several Tfh and Th1 lineage-associated genes change in expression variability between days 2 and 4. For example, we noted a decrease in variability for one key Th1 regulator, *Tbx21* (encoding Tbet), which suggests that

a subset of cells may have already committed to the Th1 lineage at day 2. Three additional Th1 lineage-associated genes also followed this trend (*Ahnak*, *Ctsd*, *Tmem154*). These data suggest that differentiation fate decisions may arise as early as day 2 in subpopulations within this system, resulting in high gene expression variability. Such an effect is in accordance with the early commitment to effector T cell fates that was previously observed during viral infection [391]. As these results illustrate, diversity in the differentiation state within a population of T cells can drive our differential variability results. To further dissect these results, subsequent analyses such as the pseudo-time inference used in [253] could be used to characterise a continuous differentiation process.

Finally, I confirmed that the regression BASiCS model can be used to study heterogeneous differentiation processes in droplet-based scRNA-Seq data. While the sparsity of these data can lead to unstable posterior inference, the use of UMIs allows for robust measurement of transcript counts. Therefore, posterior estimates of model parameters can be used to detect changes in mean expression and in transcriptional variability. In the case of somitogenesis, I validated known Wnt and Fgf signalling pathways in pre-somitic mesoderm cells and newly discovered a possible priming for somitic mesoderm cells to later form limb muscles. Identifying these genes supports a deeper understanding of cell fate decisions in the developing embryo.

In sum, our model provides a robust tool for understanding the role of heterogeneity in gene expression during cell fate decisions. This tool can be widely applied to systems where strong expression changes occur and for sparse droplet-based scRNA-Seq datasets. It allows the statistical assessment of changes in variability due to available uncertainty estimates. A drawback of this framework is the restriction to relatively few (up to 1000) cells that can be modelled per run. In **Chapter 5**, I will discuss current and future extensions to the Bayesian inference of generative models that allows the estimation of model parameters using more than 1 million cells.

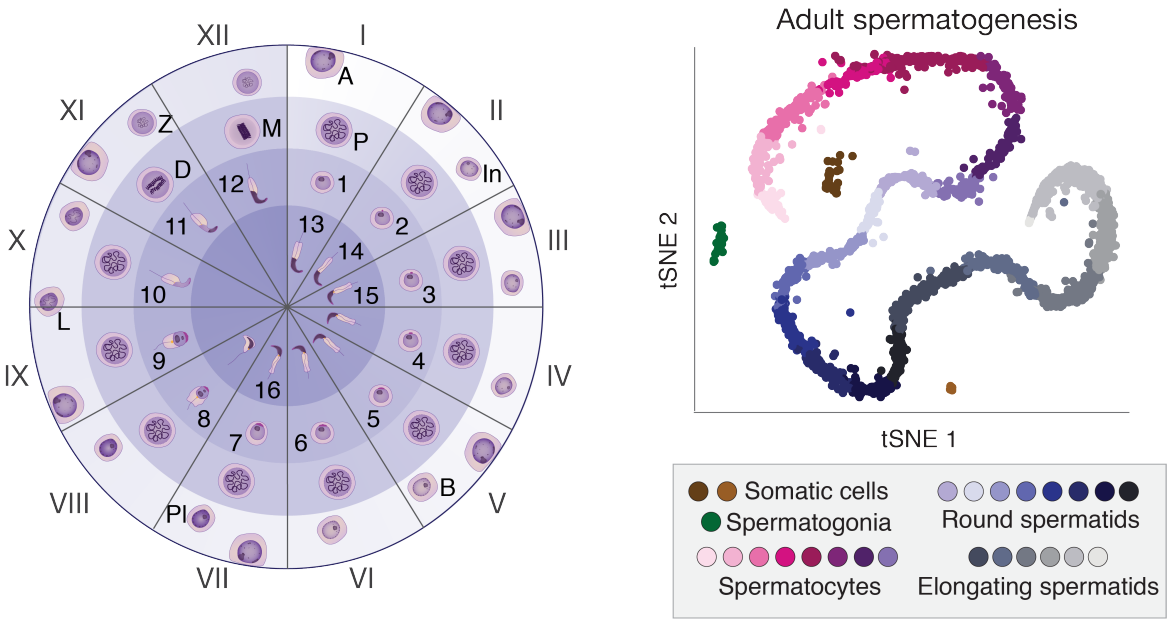
4

Transcriptional dynamics during spermatogenesis at single-cell resolution

Spermatogenesis is a recurring differentiation process that results in the production of male gametes within the testes. Consequently, its in-depth characterisation is needed to understand male fertility. During spermatogenesis, spermatogonial stem cells undergo a unidirectional differentiation process to form spermatocytes, round spermatids and lastly mature sperm. This process involves a complex sequence of developmental steps and is coupled to large-scale chromatin rearrangements, therefore making it difficult to profile. To address this, we thoroughly characterised spermatogenesis by profiling the transcriptomes of over 20,000 cells that were captured using droplet-based scRNA-Seq. To confidently connect transcriptional profiles to distinct developmental stages, we profiled multiple time points during the first wave of spermatogenesis. As juvenile animals progress through spermatogenesis for the first time, development has only progressed to a certain point, thus allowing the identification of the most mature cell type. With this precise labelling, we can dissect developmental processes such as spermatogonial differentiation, meiosis and spermiogenesis at a molecular level. Furthermore, our data captured the expression dynamics of the X chromosome, which is subject to meiotic silencing in spermatocytes, followed by a partial reactivation in spermatids. ScRNA-Seq reveals the distinct temporal expression dynamics present in the post-meiotic reactivation of the X chromosome. Profiling of the associated chromatin changes identified a set of genes specifically repressed by H3K9me3 in spermatocytes that later-on escape post-meiotic silencing in spermatids, demonstrating extensive chromatin remodelling on the X chromosome. After fully characterising spermatogenesis at a single-cell level, BASiCS was used to detect changes in transcriptional variability by estimating the residual over-dispersion measures for the different germ cell populations. In this analysis, the differentiation trajectory defines a new confounding factor that must be accounted for to accurately quantify stochastic variability in gene expression levels. ■

Declaration This project was done in collaboration with members of the Marioni and Odom lab. Christina Ernst performed all wet-lab experiments presented in this chapter. Celia P. Martinez-Jimenez performed preliminary experiments. John Marioni and Duncan Odom supervised the study. Christina Ernst, I, John Marioni and Duncan Odom designed the study and wrote the manuscript. I performed all computational analyses and produced all figures in this chapter (except the histology images and schematics). The last section of this chapter is purely my own contribution. The preprint has been made available online at:

Christina Ernst*, Nils Eling*, Celia P. Martinez-Jimenez, John C. Marioni, Duncan T. Odom. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *bioRxiv*, 2018, (* equal contributions)



4.1 Introduction

Gametogenesis describes the process that generates haploid gametes, which carry one copy of the individuals DNA. Sexual reproduction requires the fusion of two gametes, from each of the opposite sexes, to drive evolution and adaptation [392]. Spermatogenesis, the male version of gametogenesis, is a tightly regulated developmental process that ends in the generation of mature sperm. During spermatogenesis, spermatogonial stem cells undergo a unidirectional differentiation programme to form mature spermatozoa. This process occurs in the epithelium of seminiferous tubules in the testis **(Fig. 4.1B)** and is tightly coordinated to ensure the continuous and life-long production of mature sperm cells. In the mouse, the first step involves spermatogonial differentiation to form pre-leptotene spermatocytes [393–395]. Pre-leptotene spermatocytes then commit to meiosis, a cell division programme that consists of two consecutive cell divisions to produce haploid cells. To accommodate homologous recombination between sister chromatids and chromosome synapsis [396], prophase of meiosis I is extremely prolonged, lasting several days in males. It can be divided into four substages: leptonema, zygonema, pachynema and diplonema. Following the two consecutive cell divisions, haploid cells known as round spermatids undergo a complex differentiation programme called spermiogenesis to form mature spermatozoa [397].

Spermatogenesis takes place in a highly orchestrated fashion, with tubules periodically cycling through twelve epithelial stages defined by the combination of germ cells present (see **Fig. 4.1B** and [397]). The completion of one cycle takes 8.6 days in the mouse, and the overall differentiation process from spermatogonia to mature spermatozoa requires approximately 35 days [398]. Thus, three to four generations of germ cells are present within a tubule at any given time. In adults, each tubule resides in a different cycle stage meaning that at any given time point a continuum of germ cell types is present in the testis **(Fig. 4.1)**. The continuity of this differentiation process and the gradual transitions between spermatogenic cell types have made the isolation and thus the molecular characterisation of individual sub-stages during spermatogenesis difficult.

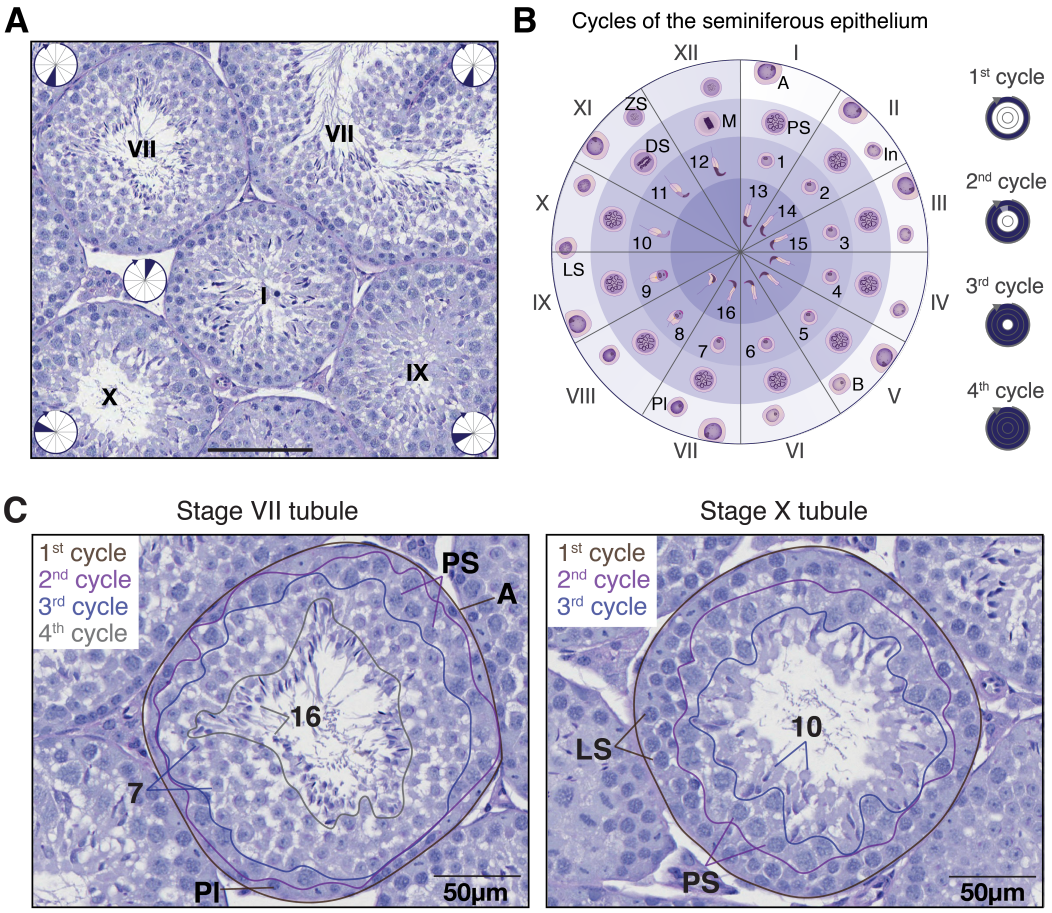


Fig. 4.1: Staging of the testicular seminiferous epithelium.
(A) Periodic Acid Schiff (PAS)-stained testis cross-section showing a number of seminiferous tubules at different epithelial stages (displayed as Roman numerals). Within each tubule, the inset circle refers to the corresponding section in (B). Scale bar represents 100 µm, **(B)** Schematic representation of the 12 stages of the seminiferous epithelium in mice. The colour gradient within the circle indicates the differentiation path of germ cells with the layers corresponding to individual cycles of the epithelium. The circle is divided into 12 section, each corresponding to one epithelial stage displaying the characteristic germ cells. Within each section, cells are positioned across the different layers according to their emergence during consecutive cycles, each being 8.6 days apart with more mature cells moving towards the centre, **(C)** Higher magnification of two tubules depicted in (A). The PAS-stained cross-sections show tubules in Stage VII and Stage X, with the different cell layers indicated by coloured lines. Stage VII tubules contain 4 different layers with germ cells from different generations that are approximately 8.6 days apart, whereas Stage X tubules only contain three layers. Cell types are labelled as: type A spermatogonia (A), intermediate spermatogonia (In), type B spermatogonia (B), preleptotene spermatocyte (Pl), leptotene spermatocyte (LS), zygotene spermatocyte (ZS), pachytene spermatocyte (PS), diplotene spermatocyte (DS), metaphase I and II (M), speramtid (S): round spermatids stages 1-8 and elongating spermatids stages 9-16, spermatogonia (SG), spermatocyte (SC).

To fully elucidate the molecular genetics of germ cell development, it is crucial to sample the full spectrum of germ cells present in testes of adult animals. For this purpose, we employed an unbiased droplet-based scRNA-Seq approach using the 10X Genomics™ platform. We used the transcriptomic profiles of thousands of single germ cells to characterise the complex transcriptional dynamics of spermatogenesis at a high-resolution. To confidently identify and label cell populations throughout the developmental trajectory, we profiled cells from juvenile testes during the first wave of spermatogenesis. In juveniles, spermatogenesis has only progressed to a defined developmental stage, which therefore allowed us to unambiguously identify the most mature cell type by comparison with adult. The correct labelling of cell types was then used to dissect differentiation processes such as meiosis and spermiogenesis. Furthermore, juvenile samples were enriched for spermatogonia, which allowed us to characterise spermatogonial differentiation. Another major developmental process during spermatogenesis is the inactivation and reactivation of the X chromosome, which is subject to transcriptional silencing as a consequence of asynapsis [399]. By combining bulk and single-cell RNA-Seq approaches with chromatin profiling, we identified that *de novo* activated X-linked genes carry distinct chromatin signatures with high levels of repressive H3K9me3 in spermatocytes.

Finally, after fully characterising the transcriptional changes during spermatogenesis, I used the regression model presented in the previous chapter to study changes in transcriptional variability over the differentiation time course. To this end, I generated *post hoc* posterior distributions of linear regression coefficients to statistically test whether individual genes increase or decrease in variability. Furthermore, the clustering of variability profiles showed that rapid transcriptional changes during differentiation can cause peaks in such variability profiles.

4.2 Data generation and processing strategies

To fully dissect mouse spermatogenesis, we performed three sets of experiments: 1. droplet-based scRNA-Seq of juvenile and adult animals, 2. bulk RNA-Seq of multiple time points during the first wave of spermatogenesis and 3. cleavage under targets & release using nuclease (CUT&RUN) to profile chromatin marks in juvenile mice. The following section will give an overview on the data generation and processing approaches. Detailed analysis steps are explained throughout the chapter. The full experimental set-up can be seen in **Fig. 4.2**.

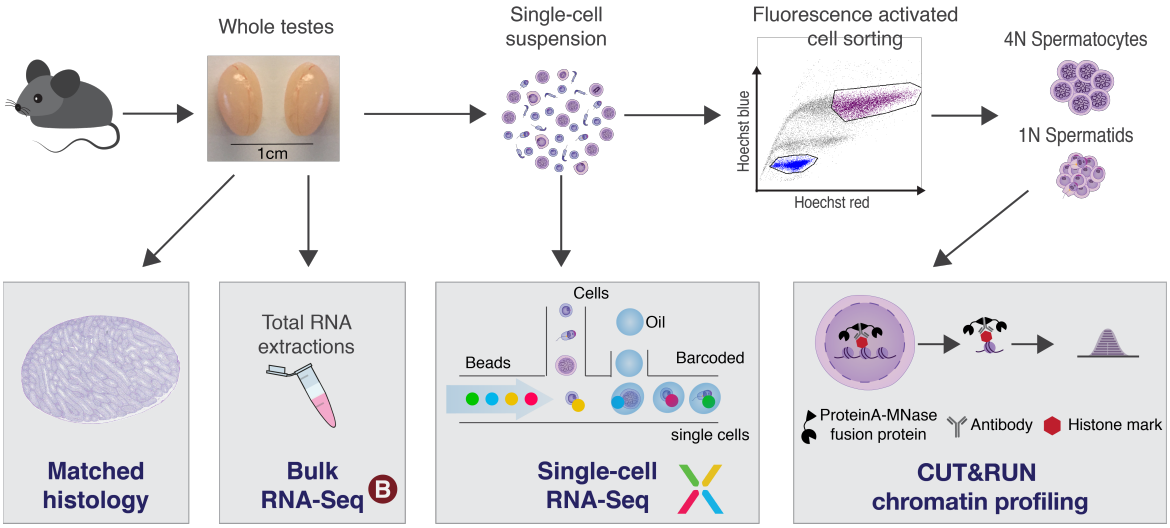


Fig. 4.2: Experimental design to dissect mouse spermatogenesis.
Overview of the experimental design yielding bulk RNA-Seq, droplet-based scRNA-Seq and chromatin profiling on FACS-purified cells using CUT&RUN from one testis while using the contralateral testis for matched histology.

4.2.1 scRNA-Seq using the 10X Genomics™ system

Droplet based scRNA-Seq was performed using the 10X Genomics™ technology [178]. For this, testes from specifically staged juvenile (between postnatal days 6 and 35) and adult (8-9 weeks) C57BL/6J (B6) mice were dissociated. Single-cell suspensions were loaded into one channel of the 10X Chromium™ Single Cell A Chip, aiming for a recovery of 4000-5000 high-quality cells. Further information on the experimental strategy can be found in **Appendix A.2** and **Table 4.1** summarises the cells captured per sample.

Table 4.1: Quality filtering of scRNA-Seq data.
Quality metrics of droplet-based scRNA-Seq. **Sample:** Stage information for all samples, **Library:** sample identifier, **CellRanger filter:** Number of retained cells after default thresholding using the CellRanger *counts* function, **After QC:** Number of cells obtained after quality control (QC), **Assigned Cell Type:** Number of cells that fall into annotated clusters (removing outlying cells), **EmptyDrops filter:** Number of cells retained after using the *emptyDrops* function controlling the FDR to 1%, **EmptyDrops QC:** Number of cells obtained after QC of the *emptyDrops* filtered cells.

Sample	Library	CellRanger filter	After QC	Assigned Cell Type	EmptyDrops filter	EmptyDrops QC
Adult	do17815	1157	1157	1123	4467	3400
Adult	do17816	2198	2198	2092	6145	4603
P10	do17821	3229	3213	3212	4976	4202
P15	do18195	4258	4258	4014	14050	13168
P20	do17824	1775	1775	1662	9400	7491
P25	do18196	4334	4334	4130	8038	6802
P30	do17825	2278	2278	2211	5393	4958
P35	do17827	3160	3160	3004	49002	10683

To process droplet-based scRNA-Seq data after sequencing, 10X Genomics™ developed a set of processing pipelines termed *Cell Ranger*. We obtained gene-specific transcript counts using the Cell Ranger *count* function with default settings. This pipeline aligns reads against the *Mus musculus* genome (GRCm38) and counts UMIs per transcript and sample. This software retains cells with similar UMI distributions [178]. We use this default threshold to obtain high-quality cells with large numbers of UMIs. For further quality control and after merging cells of all samples, we filtered out cells that express less than 1000 genes. Additionally, we exclude cells with more than 10% of reads mapping to the mitochondrial genome. The number of remaining cells per sample can be seen in **Table 4.1**.

The Cell Ranger *count* pipeline performs thresholding on the number of UMIs per cell to exclude empty droplets or droplets with low-quality cells. T his default threshold is unable to distinguish between smaller cells with lower transcriptional complexity from more heterogeneous milieu of background mRNAs. We therefore used the *emptyDrops* function provided in the *DropletUtils* Bioconductor package to statistically distinguish empty droplets from genuine cells (controlling the FDR to 1%) [400]. After applying *emptyDrops*, further quality control needs to be performed and after merging all remaining cells across all

samples, we filtered out cells with less than 500 genes expressed. Furthermore, we excluded cells with more than 10% of mitochondrial genes expressed **(Table 4.1)**.

The transcriptomes of quality filtered cells were normalised using the *scran* package [401]. For this, cells with similar transcriptomic complexity were pre-clustered using a graph-based approach (as implemented in the *quickCluster* function). Size factors were calculated within each cluster before being scaled between clusters using the *computeSumFactors* function. Throughout this paper, the $\log_2$-transformed, normalised counts (after adding one pseudocount) are displayed. For down-stream analysis, lowly expressed genes (averaged $\log_2$-transformed, normalised expression < 0.1) were excluded. After quality control and filtering, we retained more than 20,000 high-quality single cells and over 46,000 cells including those with lower transcriptional complexity **(Table 4.1)**. These cells were used to dissect molecular processes during spermatogenesis and to profile under-represented and transcriptionally inactive cell types in mouse testes.

4.2.2 Bulk RNA-Seq from juvenile animals

Additionally, we generated whole-tissue bulk RNA-Seq libraries from time points during the first wave of spermatogenesis (**Appendix A.2.3**). More specifically, we sampled (with biological replicates) testes from mice at post-natal (P)6 (2x), P8, P10 (2x), P12 (2x), P14 (2x), P16, P18 (2x), P20 (2x), P22 (2x), P24 (2x), P26 (2x), P28 (2x), P30 (2x), P32, P34, P35 and from adult animals (2x). Detailed experimental methods can be found in **Appendix A.2.3**.

Sequenced reads were aligned against the *Mus musculus* genome (GRCm38) using the *STAR* aligner with default settings [402]. Gene-level transcript counts were obtained using *HTSeq* [336] with the –s option set to “reverse” and using the GRCm38.88 genomic annotation file. We visualised several features of the aligned and counted data (number of intronic/exonic reads, number of multi-mapping reads, low-quality reads and total library size) and did not detect any low-quality RNA-Seq libraries. Next, we used the size factor normalisation approach implemented in DESeq2 [371] for data normalisation. For down-stream analysis and visualisation, lowly expressed genes (averaged counts < 10) were excluded. With this, we generated 30 bulk RNA-Seq libraries that will be used for developmental staging of cell types and to dissect X chromosome expression dynamics.

4.2.3 CUT&RUN from juvenile animals

To map chromatin states in purified cell populations we used CUT&RUN [403]. In brief, spermatocytes and spermatids were sorted as described in **Appendix A.2.2** and attached to concanavalin A–coated magnetic beads. After permeabilisation, anti-bodies against H3K9me3 and H3K4me3 were incubated with the cells. Inactive micrococcal nuclease linked to protein A was added to the mix and cooled to 0°C. Protein A binds to the antibodies and upon calcium-dependent activation the nuclease digests DNA next to the histones where the antibody bound. Cleaved DNA fragments diffused out of the nucleus and were prepared for sequencing. This technique allows targeted chromatin profiling for specific marks at a genome wide level and requires minimal cell input relative to conventional ChIP-Seq [403]. Detailed experimental methods can be found in **Appendix A.2.6**.

Sequenced reads were aligned to the *Mus musculus* genome (GRCm38) using *Bowtie2* with the following settings: --local --very-sensitive-local --no-unal -q --phred33. Paired end reads were counted in specified regions using the *regionCounts* function implemented in the *csaw* Bioconductor package [404]. For this, duplicated reads, reads that mapped more than 1000 bp apart and reads mapping to blacklisted regions (available at: http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz) were removed. Regions of interest included: promoters (obtained using the *promoters* function of the *GenomicFeatures* package), 1000 bp windows across the chromosome (using the *windowCounts* function of *csaw*) and whole chromosomes. Counts per region were normalised based on library size by computing counts per million (CPM) for promoter regions and 1000 bp windows. Additionally, when considering entire chromosomes, the length of the chromosome was accounted for by computing the fragments per kilobase per million mapped reads (FPKM).

4.2.4 Identification of germ cell types across all scRNA-Seq samples

After data generation and pre-processing steps, we next characterised the detectable cell types across all scRNA-Seq samples. We assume that cell types sampled from juvenile animals are also found among the cell types sampled from adult animals. To detect cell types consistently across all scRNA-Seq samples, we first performed batch correction. To remove batch-specific effects that arise when samples are prepared and sequenced in different experiments **(Tables 4.1)**, we used the *mnnCorrect* function implemented in the *scran* package [405]. We used the top 1000 genes with highest biological variation across all samples as informative genes for batch correction. The *mnnCorrect* function takes transcriptional profiles of cells isolated from adult B6 mice as the first input and uses this dataset as a reference for cell mapping **(Fig. 4.3A)**.

To identify cell types across all samples, batch corrected transcriptomes were clustered using a graph-based approach. A shared nearest-neighbour (SNN) graph [406] was constructed considering 3 shared nearest neighbours using the *buildSNNGraph* function in *scran*. In the next step, a multi-level modularity optimisation algorithm was used to find community structure in the graph [407] as implemented in the *igraph* R package. Cells in small clusters that had ambiguous identities were excluded from down-stream analysis. In total, we identified 28 clusters. To correctly label cell clusters based on their cell type, we identified marker genes for all germ cell types in the adult B6 samples. To this end, we performed differential expression testing using multiple pairwise comparisons. To detect cluster-specific marker genes, the *findMarkers* function implemented in *scran* was applied to the $\log_2$-transformed normalised counts while providing the cluster labels **(Fig. 4.3B)**.

By visualising the expression of detected marker genes, we identified the following cell types: spermatogonia (SG, based on *Dmrt1* expression, [408]), spermatocytes (SC, *Piwil1*, [409]), round and elongating spermatids (S, *Tex21* and *Tnp1*, respectively, [410]), as well as the main somatic cell types of the testis, Sertoli (*Cldn11*, [411]) and Leydig cells (*Fabp3*, [412]) **(Fig. 4.3B)**. Using a dimensionality reduction technique for visualisation (tSNE, **Fig. 4.3C**), the germ cell types from spermatocytes to elongating spermatids formed a continuum, which recapitulated the known developmental trajectory.

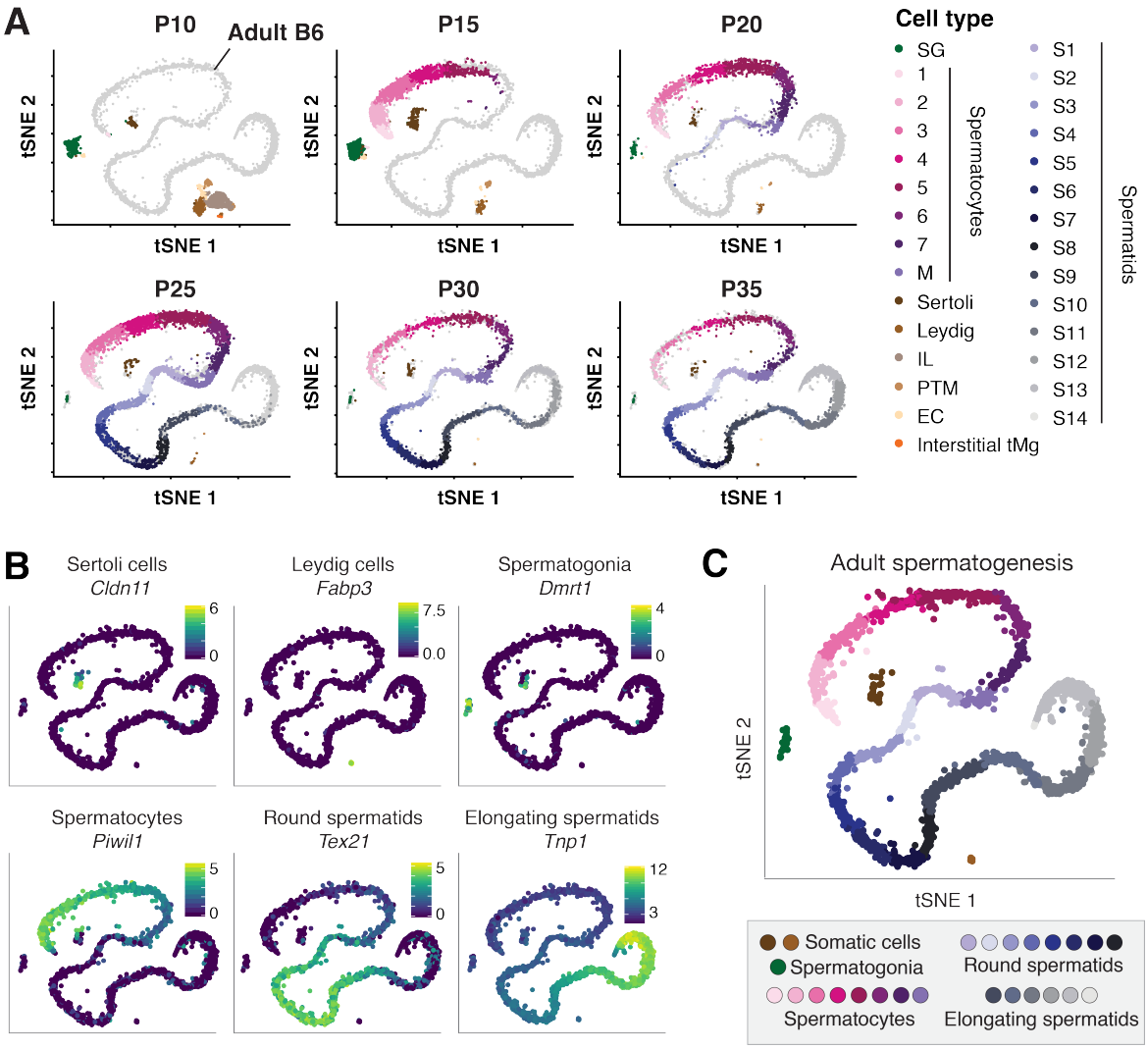


Fig. 4.3: Droplet based scRNA-Seq of juvenile and adult mouse spermatogenesis.
(A) tSNE representation of juvenile cells that were mapped to cells isolated from adult mice. Grey dots indicate all cells from adult animals that were used as a reference for cell mapping. Coloured dots represent cells isolated at each sampled stage during the first wave of spermatogenesis. Clustering has been perfomed across all cells after cell mapping. SG: Spermatogonia, M: Metaphase, IL: Imature Leydig, PTM: Peritubular Myoid Cells, EC: Endothelial Cells, tMg: testicular Marcophages, **(B)** tSNE representation of scRNA-Seq data from adult B6 mice with the colour gradient representing the expression of known marker genes for two somatic cell types and the main germ cell types. The x- and y-axis represent the first and second dimension of tSNE respectively. The colour legend shows $\log_2$-transformed, normalised expression counts, **(C)** Graph-based clustering identifies different sub-stages within major germ cell populations form adult B6 animals.

4.3 Developmental staging of mouse spermatogenesis

Historically, sub-staging of the major cell types within the testis was based on changes in nuclear or cellular morphology [397, 398]. Previous attempts to complement morphology with molecular signatures have been limited to FACS-based and sedimentation assays. Their resolution was not sufficient to differentiate between sub-cell types [413–418]. While a mixture of all spermatogenic cell types co-exists in the adult, the first wave of spermatogenesis during juvenile development is more synchronised. Starting around mouse postnatal day P4, spermatogonia begin to differentiate, forming the first generation of spermatocytes as early as P10, round spermatids by P20, and completing the first wave of spermatogenesis with the production of mature spermatozoa between P30 and P35 **(Fig. 4.1B and 4.4A)** [419–421]. In this section, we define well-known developmental transitions during spermatogenesis by (i) mapping cells sampled from defined epithelial stages during the first wave of spermatogenesis to cells sampled from adult testes and (ii) classifying the cell types identified above using bulk RNA-Seq sampled from juvenile testes every two days during the first wave of spermatogenesis.

4.3.1 Cell type characterisation using the first wave of spermatogenesis

We exploited the synchronised development of cell types throughout the first wave of spermatogenesis to define major and morphologically described check-points of the differentiation process. For this, we sampled multiple time points from juvenile animals to identify the most mature (differentiated) cell types. At any given time point during the first wave of spermatogenesis, there exist a defined number of known cell types in juvenile testis depending on the timing of the developmental cycle **(Fig. 4.1B)**. Based on known sperm developmental transitions, we chose six time points between P10 and P35 for generation of single-cell RNA-Seq libraries **(Fig. 4.4A and Table 4.1)**. As described above, we mapped the transcriptomes of juvenile samples onto the adult B6 sample **(Fig. 4.3A)**. For each juvenile experiment, we found that the population of developing germ cells was strongly enriched at the expected developmental stage, as quantified by the percentage of cells in each cell type cluster **(Fig. 4.4C)**. By associating the known cell types from juvenile animals with the corresponding cell types in the adult trajectory, we unambiguously assigned molecular and histological signatures to cells during adult spermatogenesis.

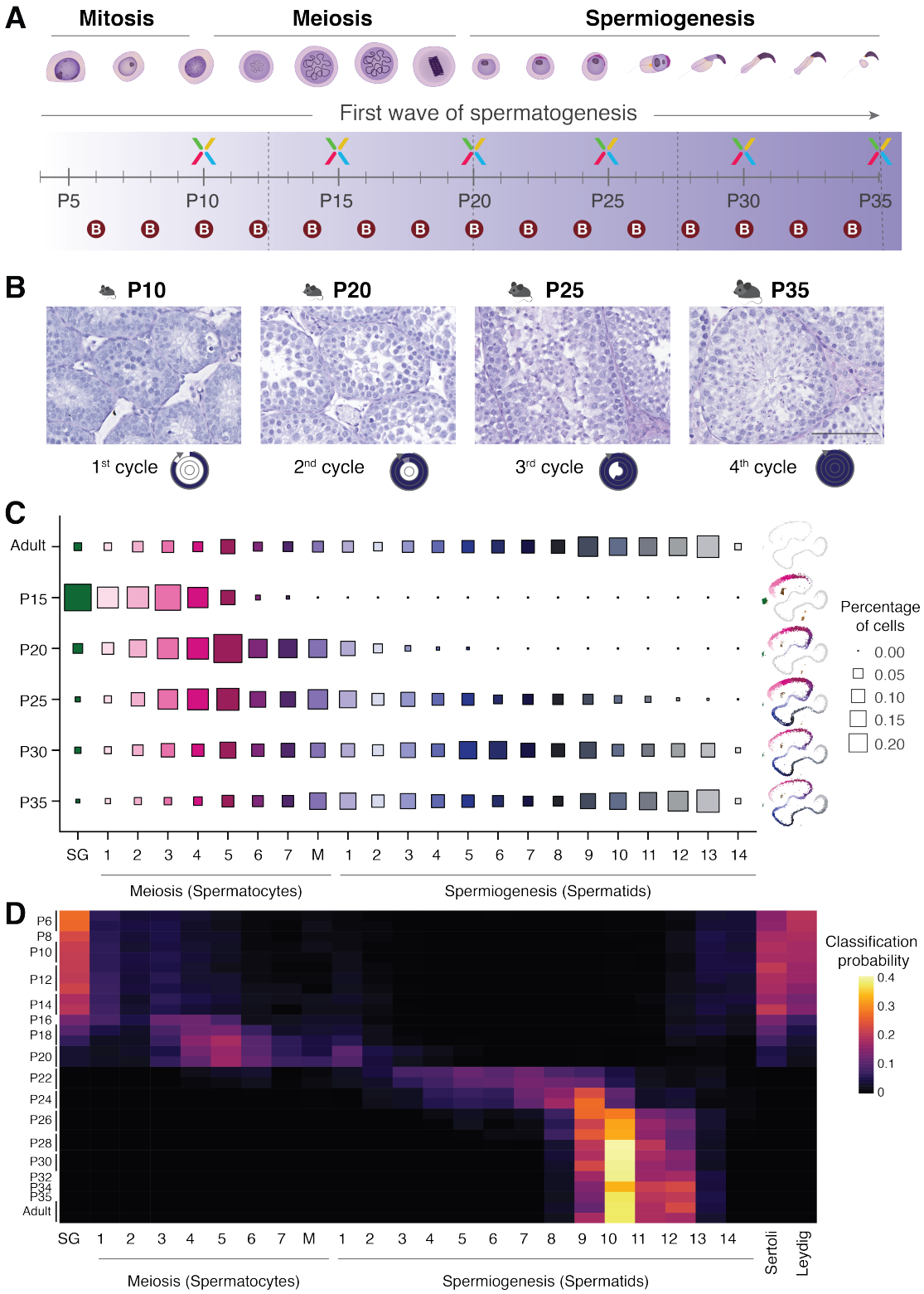


Fig. 4.4: Staging of cell types during mouse spermatogenesis (full legend on next page).

Fig. 4.4: Staging of cell types during mouse spermatogenesis (continued).
(A) Schematic representation of the major germ cell types and their corresponding developmental processes. Spermatogonia differentiate undergoing mitotic cell divisions before forming spermatocytes that divide by meiotic division. Following meiosis, spermatids differentiate throughout spermiogenesis to form mature sperm. The timeline in the lower panel indicates at which point during the first wave of spermatogenesis samples were harvested for the generation of scRNA-Seq (X) or bulk RNA-Seq (B) data, **(B)** Representative images of seminiferous tubules from animals harvested at different postnatal (P) time points during the first wave of spermatogenesis. The approximate timing of the stage and cycle of the tubule is illustrated below in the form of a circle (see **Fig. 4.1B**), **(C)** After cell mapping and clustering, the percentage of cells in each cluster can be calculated for each sample. The size of squares corresponds to this percentage and the colours indicate the cluster labels depicted in **Fig. 4.3C**. tSNEs on the right-hand side of each panel (juvenile samples only) illustrate progress through spermatogenesis. SG: Spermatogonia, M: Metaphase, **(D)** Probabilistic mapping of bulk RNA-Seq libraries to the cell clusters identified in the adult scRNA-Seq data using a random forest approach. The colour gradient indicates the probability with which a bulk sample can be assigned to the specific cell cluster.

In adults, we detect a homogeneous distribution of cells across the germ cell types ranging from spermatogonia to S14 spermatids **(Fig. 4.4C)**. To characterise germ cell types, we focused on samples taken from P15-P35 animals since at P10, the majority of cell types do not show germ cell properties **(Fig. 4.3A)**. In earlier stages, cells are enriched for the most mature cell type in each cycle. For instance, at P15 the majority of cells are spermatogonia and spermatocytes progressing through the mid-pachytene stage [422]. Interestingly, less mature cell types that exist prior to the mid-pachytene stage are also present at this (and later) time points. This supports recent reports that the first wave of spermatogenesis is less synchronised than previously anticipated [423]. At P20, we detect an enrichment for spermatocytes, cells undergoing meiotic cell division, and a small group of early round spermatids. This population structure is in line with matched histology, which shows a large number of tubules in late stages IX-XII and the first occurrence of early round spermatids [419]. It has been shown that spermatids first reach the elongating state, which occurs from S10 spermatids onwards, between P24 and P26 [420]. At P25, we observed that cells mapped to our first ten clusters of spermatids, which we then labelled according to morphologically-defined spermatid substages S1 – S10 **(Fig. 4.4C)**. At P30 and P35, we observed a relatively even distribution of cells across all groups, closely resembling the adult distribution up to S14, indicating that the first wave of spermatogenesis is complete. With this computational mapping of cells collected at different developmental time points, we linked transcriptional profiles of single cells to morphologically defined transitions during germ cell development.

4.3.2 Classification of cell types based on bulk RNA-Seq data

While the analysis performed above determines crucial developmental transitions during spermatogenesis, we did not achieve the high-resolution required for the mapping of defined developmental cell types to the clusters identified above. To further validate the identity of the cell clusters, we used bulk RNA-Seq from testis collected during the first wave of spermatogenesis. These samples were harvested every two days between P6 and P34 and allowed us to refine the mapping analysis performed above **(Fig. 4.4A)**. The batch-correction approach used above was developed to match hundreds of single cells across samples and is not suitable to map the 30 bulk RNA-Seq samples onto the adult trajectory. To classify each bulk RNA-Seq sample to one or multiple clusters identified in the scRNA-Seq data, we used a regression approach that performs probabilistic classification. Using the top 50 cluster-specific marker genes for spermatogonia, all spermatocyte groups, all spermatid groups, sertoli and leydig cells, we trained a random forest classifier (implemented in the *randomForest* R package [424]) on 2000 cells isolated from adult B6 testes. Model testing was performed on the remaining 1215 cells isolated from adult B6 testes. Prior to training and testing, $\log_2$-transformed, normalised counts were scaled by computing the Z score for each gene. Probabilistic prediction was performed using the Z score of $\log_2$-transformed, normalised bulk RNA-Seq reads of the input genes. The output of this analysis is the classification probability for each bulk RNA-Seq sample to belong to each scRNA-Seq cluster.

This classification confirmed that between P6 – P14 spermatogonia and somatic cells contributed most to the transcriptomic profile **(Fig. 4.4D)**. Between P16 and P20 we observed the emergence of spermatocyte-specific gene expression signatures, after which spermatids become the transcriptionally dominant cell type. By P26, spermatids reach the elongating state where transcription is uniformly shut-down due to the beginning of the histone-to-protamine transition [425]. Following this, changes in RNA content are mostly due to degradation. Bulk transcriptional profiles can only be classified up to S10 because transcription is largely inactive thereafter and no new cluster-specific marker genes emerge.

4.4 Under-represented cell types in spermatogenesis

The analysis of early stages of juvenile mice in the previous section showed an enrichment for cell types that are relatively under-represented in later stages and in adults (e.g. spermatogonia and early spermatocytes at P15, **Fig. 4.4C**). Additionally, we detected the absence of germ cells in the P10 sample which leads to a relative enrichment of somatic cell types **(Fig. 4.3A)**. This relative enrichment allows us to dissect somatic cell types and spermatogonial differentiation at higher resolution compared to cells isolated from adult samples.

4.4.1 Somatic cell types in juvenile testes

To study heterogeneity within the somatic cell population, we focused on the P10 stage, where somatic cells are relatively more frequent **(Fig. 4.5)**. As expected, we identify substantial numbers of Sertoli and Leydig cells, which are the main somatic cell types in adult. Leydig cells are the primary producers of steroid hormones such as testosterone that regulate sexual differentiation and development of secondary sex characteristics [426, 427]. Sertoli cells on the other hand reside inside the tubules and provide the structural features required for testis development. They further support differentiating germ cells by providing growth factors and nutrients [428]. In addition, we newly identified a large population of immature Leydig (IL) cells, based on *Dlk1* expression [429]. ILs shape the embryonic development of testes and rapidly decline in numbers after birth [430].

Furthermore, we detected the cells that form the basal lamina. These include peritubular myoid (PTM) cells (*Acta2*, [431]), vascular endothelial cells (ECs) (*Tm4sf1*, [432]), and testicular macrophages (tMgs) (*Cd14*, [433]) **(Fig. 4.5A and B)**. PTM are contractile cells that form the wall of seminiferous tubules and induce the release of the testicular fluid which contains mature spermatozoa [434]. In addition, endothelial cells play a major role in forming the testis cord during development and together with PTMs surround Sertoli cells in the adult testes [435]. Testicular macrophages are the most abundant immune cell type in the organ and play a crucial role in testes development by expressing anti-inflammatory cytokines that induce a tolerogenic environment in testes. This is needed since germ cell specific antigens are only expressed after puberty when the immune system has already matured [436]. By performing differential expression analysis (using the *findMarkers* function in *scran*), we identified novel markers for these cell populations that are relatively under-represented in adult testes. **Fig. 4.5B** visualises the top marker genes for the enriched

somatic cell types detected at P10.

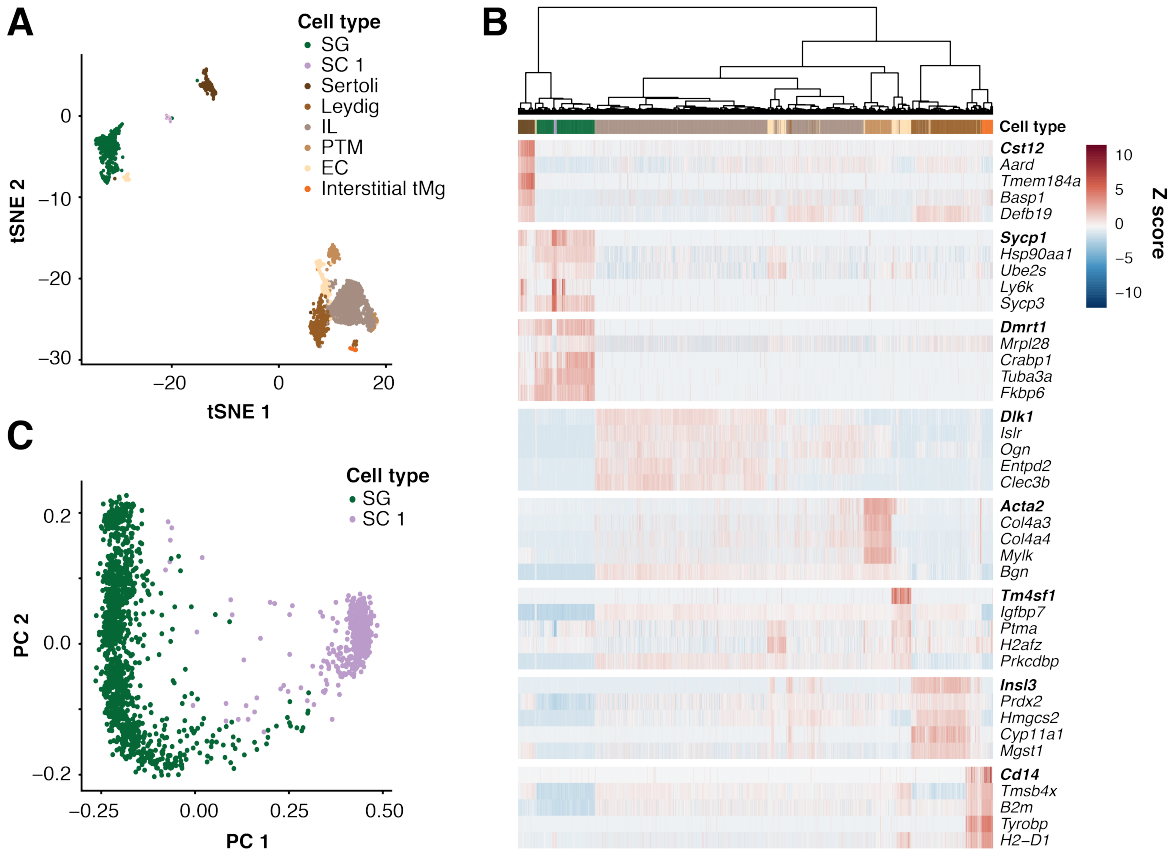


Fig. 4.5: Enrichment of under-represented somatic cell types in juvenile samples.
(A) tSNE representation of cells isolated from P10 animals that were mapped to cells from adult mice. Cell types were identified by unbiased, graph-based clustering and annotated after marker gene extraction. SG: Spermatogonia, SC: Spermatocytes, IL: Immature Leydig, PTM: Peritubular Myoid Cells, EC: Endothelial Cells, tMg: testicular Macrophages, **(B)** Heatmap representation of cell type-specific marker genes. Gene in bold are previously described markers of the following cell type: Sertoli cells (*Cst12*), early spermatocytes (*Sycp1*), spermatogonia (*Dmrt1*), immature Leydig cells (*Dlk1*), endothelial cells (*Acta2*), peritubular myoid cells (*Tm4sf1*), Leydig cells (Insl3), testicular macrophages (*Cd14*), **(C)** PCA of spermatogonia (SG) and early spermatocytes (SC 1) from P10 and P15 animals.

Furthermore, we detect a relative enrichment of spermatogonia compared to other germ cell types at P10 and P15 **(Fig. 4.5C)**. Using this large amount of stem cell like cells sampled from different time points during development allows us to dissect its differentiation programme.

4.4.2 Spermatogonial differentiation

In the mouse, spermatogenesis is initiated with the division of a type A spermatogonia, also termed A_{single}, to form first a pair, and then a connected chain of undifferentiated spermatogonia (A_{paired} and $A_{aligned}$) [393, 394]. These cells have competency to undergo spermatogonial differentiation, which involves six transit-amplifying mitotic divisions generating A_{1-4}, Intermediate (In), and B spermatogonia, which then give rise to pre-leptotene spermatocytes (Pl) [395] **(Fig. 4.1C)**. Given this, we expect a high level of heterogeneity within the spermatogonia population but identifying spermatogonial sub-populations in adult testes is greatly complicated by their rarity relative to other germ cell types [437]. However, as shown above, during early juvenile development spermatogonia are relatively enriched, which we exploited to further characterise their heterogeneity **(Fig. 4.6A)**.

By combining cells from P10 and P15, we obtained 1,186 transcriptional profiles that capture sub-populations during spermatogonial differentiation **(Fig. 4.6B)**. To jointly analyse transcriptomes of P10 and P15 samples, we performed batch correction between these samples as described above and clustered batch corrected data using a graph-based approach. In order to label the cell types corresponding to the different clusters, we performed marker gene detection using the *findMarker* function in *scran*. By visualising the individual marker genes, we detect two clusters corresponding to undifferentiated spermatogonia (A_{undiff}) based on their expression of *Nanos3* and *Zbtb16* **(Fig. 4.6B and C)** [438, 439]). These cells comprise A_s, A_{paired}, and $A_{aligned}$ spermatogonia that decrease in stemness as they divide and gain competency to differentiate [440]. Additionally, these cells express a number of marker genes also detected in undifferentiated human spermatogonial stem cells, such as *Gfra1*, *Bcl6* and *Id4* [441]. Based on the expression of stimulated by retinoic acid 8 (*Stra8*), we can map the point at which spermatogonial differentiation is induced ($A_{aligned}$-to-A_1 transition), thus marking the beginning of differentiating spermatogonia (A_{diff}) [442] **(Fig. 4.6B)**. A_{diff} are marked by the expression of *Sohlh1* [443] and are highly proliferative, generating $A_{1\text{-}4}$, Intermediate and B spermatogonia. Late differentiating spermatocytes express *Dmrtb1*, which mediates the mitosis-to-meiosis transition and quickly disappears in pre-leptotene spermatocytes **(Fig. 4.6B)**. This latter population shows a second increase in *Stra8* expression levels, which is necessary for initiation of meiosis **(Fig. 4.6B and C)** [444, 442, 341].

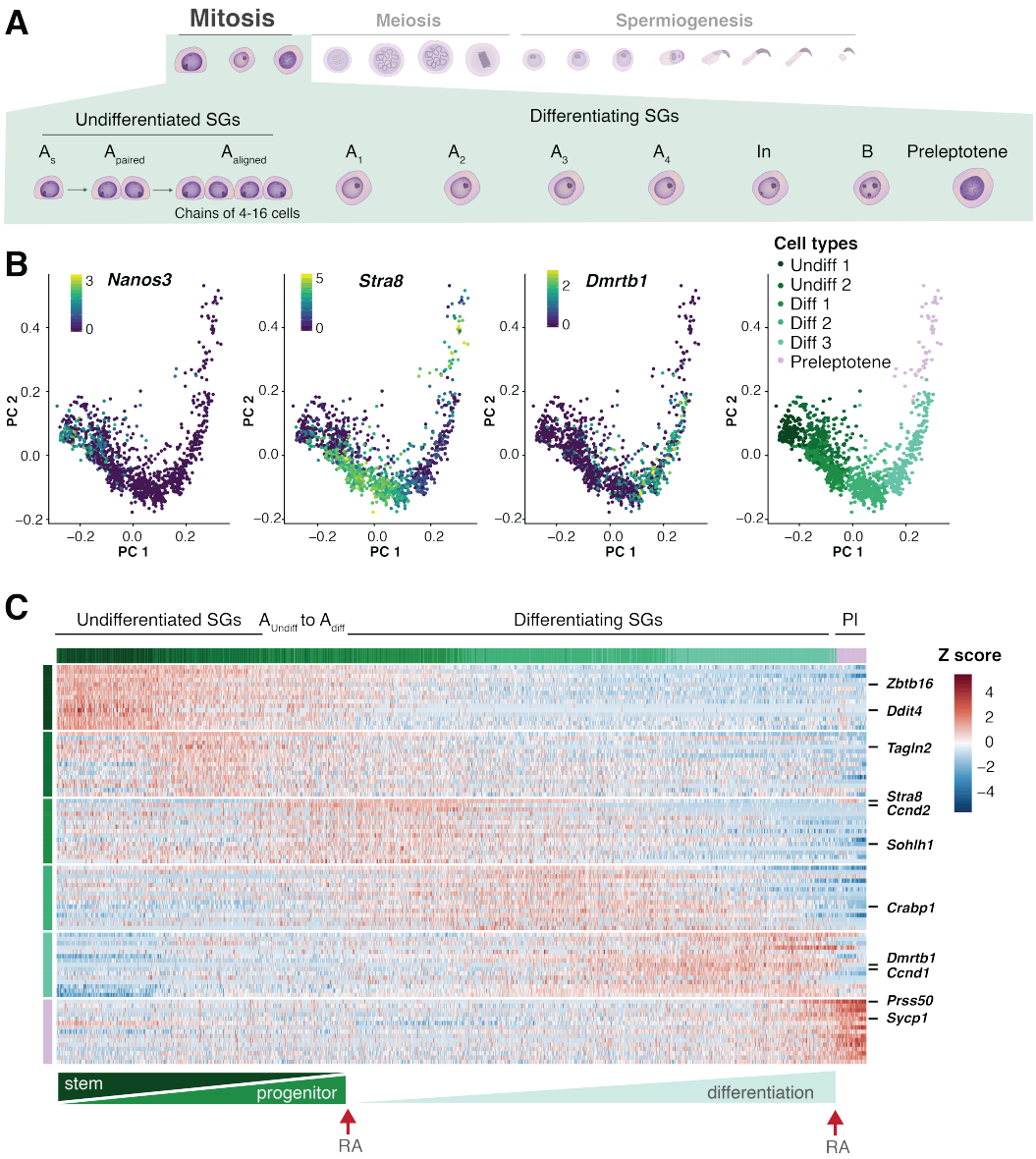


Fig. 4.6: Cellular heterogeneity during spermatogonial differentiation.
(A) Schematic representation of spermatogonial differentiation including sub-stages of undifferentiated (A_s, A_{paired}, $A_{aligned}$) and differentiating (A_1, A_2, A_3, A_4, In, B) spermatogonia (SGs) as well as pre-leptotene spermatocytes (Pl), **(B)** Sub-structure detection in spermatogonia isolated from P10 and P15 animals. PCA was computed on transcriptomes after batch correction between P10 and P15 samples. The first three panels represent expression of known marker genes for undifferentiated (Undiff, *Nanos3*) and differentiating (Diff, *Stra8* and *Dmrtb1*) spermatogonia. The colour scale shows $\log_2$-transformed, normalised counts. The last panel overlays cluster identity by sub-clustering batch-corrected transcriptomes of spermatogonia, **(C)** Z score of normalised expression counts of the top 15 marker genes per cell cluster. Column and row labels represent the cell clusters identified in the last panel of (B). The lower bar indicates the gradual differentiation from undifferentiated spermatogonia to pre-leptotene cells driven by two retinoic acid (RA) signals.

4.4.3 Leptotene and zygotene spermatocytes

The transition between differentiating spermatogonia and spermatocytes is a gradual process that occurs in stage VIII tubules when B spermatogonia divide and form pre-leptotene spermatocytes [444, 445]. When visualising the first two components of a PCA, we did not observe a continuous differentiation trajectory bridging spermatogonia to spermatocytes **(Fig. 4.5C)** which indicates a possible loss of cells that characterise the transition between these two cell types. One possible explanation is that leptotene and zygotene spermatocytes have decreased transcriptional activity [446, 447], and are thus likely to be classified as empty droplets by the 10X CellRanger pipeline.

To capture these transcriptionally quiescent cells, we used the *emptyDrops* function from the *DropletUtils* R package to distinguish between droplets capturing genuine cells with low transcriptional complexity and empty droplets containing only ambient mRNA [400]. Applying this approach increased the number of early spermatocytes in all samples and, in particular, identified a population of cells connecting spermatogonia and spermatocytes at the predicted position in the cell trajectory **(Fig. 4.7A-C)**. We strongly enrich for leptotene and zygotene spermatocytes, especially in the P15 sample, after including these smaller cells in the analysis. Due to low transcriptional complexity, these two cell types cluster together which makes it hard to detect a clear mitosis-to-meiosis transition **(Fig. 4.7B and C)**. As expected for leptotene and zygotene spermatocytes, these cells show high mRNA levels for genes involved in synaptonemal complex formation, chromosome synapsis and DNA double-strand break (DSB) formation such as *Sycp1*, *H2afx* and *Hormad1* [448–450] **(Fig. 4.7D)**.

In addition to early spermatocytes, droplets with lower transcriptional complexity also captured late condensing spermatids. As mentioned above, these late stages of spermiogenesis are characterised by continuous degradation of RNA after transcriptional shut-down at the round-to-elongating transition [425] **(Fig. 4.7E)**. Nevertheless, including droplets with low transcriptional complexity increases the risk of including low-quality cells and debris. In our case, the large cluster of unidentified cells in **Fig. 4.7A** could represent membrane vesicles containing RNA at the end of spermiogenesis that form during a process termed "cytoplasmic extrusion" [451].

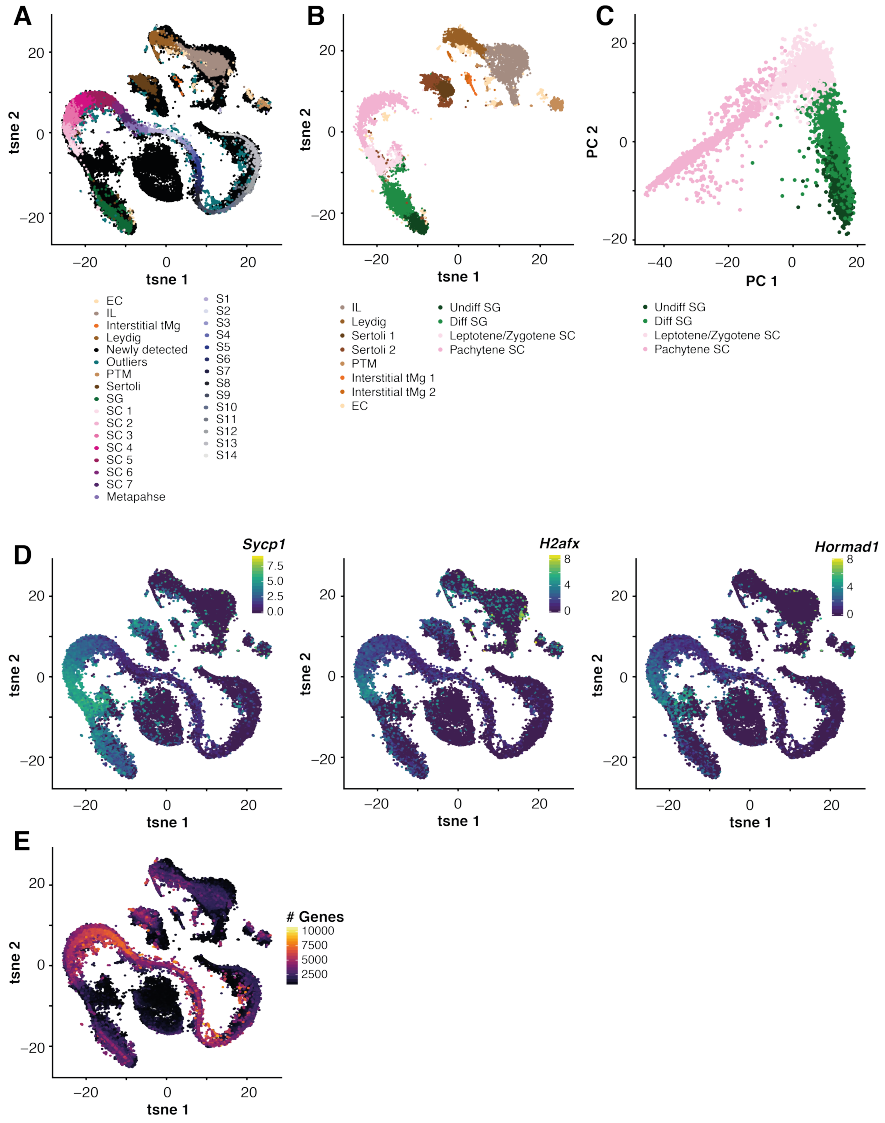


Fig. 4.7: Transcriptionally silent cell types in spermatogenesis.
(A) tSNE representation of cells selected by the *emptyDrops* filtering strategy. Coloured dots represent annotated cell types detected using the default *CellRanger* filtering pipeline while black dots represent cells detected by the *emptyDrops* filtering. SG: Spermatogonia, SC: Spermatocytes, IL: Immature Leydig, PTM: Peritubular Myoid Cells, EC: Endothelial Cells, tMg: testicular Macrophages, S: Spermatids, **(B)** tSNE representation of emptyDrops filtered cells from the P15 sample. Cell colouring corresponds to clustering performed on this sample. Undiff SG: undifferentiated spermatogonia, Diff SG: differentiating spermatogonia, **(C)** PCA representation of spermatogonia and spermatocytes detected in the P15 sample after *emptyDrops* filtering. Labelling corresponds to the clusters shown in (B), **(D)** Leptotene and zygotene spermatocyte marker gene expression. The colour scale represents $\log_2$-transformed, normalised counts, **(E)** Visualisation of the number of genes expressed (> 0 counts) per cell.

4.5 Characterisation of male meiosis

After characterising the major germ and somatic cell types, we next profiled the transcriptional programmes of known developmental processes during spermatogenesis. These include firstly meiosis and later on spermiogenesis, which will be analysed and discussed in the next section.

The mitotic expansion of spermatogonia produces large numbers of spermatocytes, which then undergo male meiosis where two consecutive cell divisions give rise to four haploid spermatids. In contrast to mitotic cell divisions, prophase of meiosis I is extremely prolonged, lasting up to 10 days in male mice [452]. Furthermore, meiosis includes programmed DSB formation, homologous recombination, and chromosome synapsis [396], which represent molecular processes to induce genetic variation between offspring. Most meiotic processes have been histologically described, but a full transcriptional characterisation of spermatocytes undergoing meiosis is lacking.

The continuum of sampled cell types allows us to perform in-depth characterisation of transcriptional changes that occur during meiosis. For this, we ordered spermatocytes along their differentiation trajectory by fitting a principal curve [453] to the first 3 principal components using the *principal.curve* function implemented in the *princurve* R package. This approach allows us to order cells along the developmental trajectory. The directionality of the trajectory was inferred using prior information based on the cluster annotation. Here, the ordering of cell types is as follows: leptotene spermatocytes (SCs, not present in CellRanger filtered data), zygotene SCs, pachytene SCs, diplotene SCs and finally cells in metaphase **(Fig. 4.8A)**.

To detect molecular processes that occur during meiosis, we first profiled the overall transcriptional rate before dissecting changes in expression on a gene-specific level. As shown before [454], we identified a strong increase in the number of genes expressed as spermatocytes progress through prophase, with the highest number being expressed immediately before the cells divide **(Fig. 4.8A)**. Using this as a proxy for active transcription, we identified diplotene spermatocytes, which are the latest cell type in prophase I in which RNA synthesis is occurring [447].

We used the increase and later decrease in transcription as a guide for the progressive changes in transcription throughout meiosis. Therefore, to detect functional genes that influence this process, we correlated each gene's normalised expression level to the number of genes expressed. For this, we used the *correlatePairs* function implemented in *scran* [339]. First, we constructed an empirical null distribution using the *correlateNull* function implemented in *scran*. Next, we tested the observed Spearman's ρ for each gene against this null distribution. Genes with ρ < -0.3 and a Benjamini-Hochberg corrected empirical p-value < 0.1 were considered as negatively correlated and genes with ρ > 0.3 and a Benjamini-Hochberg corrected empirical p-value < 0.1 were considered as positively correlated.

As expected, previously known marker genes for early meiotic processes such as *Hormad1* and *Sycp3* decreased in expression during Prophase I, whereas *Pou5f2* and *Tcte2*, a male meiosis-specific gene [455] increased in expression **(Fig. 4.8B)**. Supporting our identification of diplotene spermatocytes, *Pou5f2* has previously been shown to be specifically expressed during a 36- to 48-hour period preceding the meiotic cell division [456].

In the next step, we performed a targeted analysis and detected marker genes for each of the spermatocyte sub-cell types. Despite the overall increase in transcription, we observed distinct temporal expression patterns when visualising these specific marker genes for individual spermatocyte populations. Even within pachytene spermatocytes at different stages in their developmental progression, there exists substantial heterogeneity **(Fig. 4.8C)**. As expected, early spermatocyte markers (SC 1 and SC 2) were enriched for genes with known functions in male or female fertility such as *Piwil1* (*Miwi*), *Cks2*, *Sycp1*, reflecting a history of intensive investigation [409, 457, 450]. We performed literature search and used the database www.mousephenotype.org to annotate genes regarding their sterility phenotype.

In sum, we dissected the transcriptional heterogeneity within spermatocytes undergoing meiosis and identified a set of genes that form potential drivers for this process. Genes which show a high expression in early stages during meiosis have been validated to induce sterility once removed from the system.

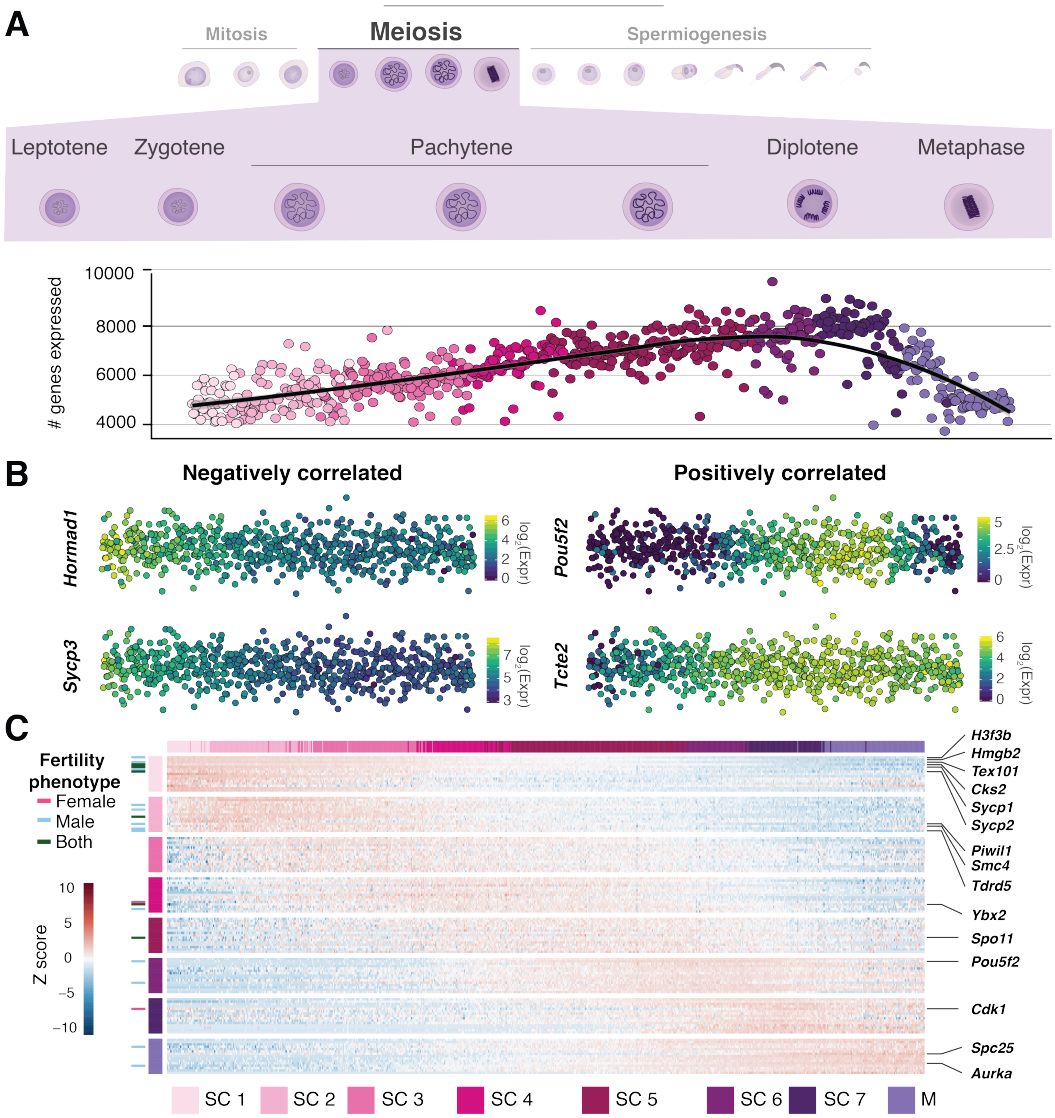


Fig. 4.8: Gene expression dynamics during male meiosis.
(A) Number of genes expressed per spermatocyte. Cells are ordered by their developmental progression during meiotic prophase until metaphase, **(B)** Expression of genes that are negatively or positively correlated with the number of genes expressed during meiotic prophase (negatively correlated: ρ < -0.3, Benjamini-Hochberg corrected empirical p-value < 0.1; positively correlated: ρ > 0.3, Benjamini-Hochberg corrected empirical p-value < 0.1). Per category, two genes are visualised. The colour gradient represents $\log_2$-transformed, normalised counts, **(C)** Heatmap visualising the Z score scaled expression of the top 15 marker genes per cell type. Row and column labels correspond to the different populations of spermatocytes (SC). M: Metaphase. Genes are labelled based on their fertility phenotype: pink – infertile or sub-fertile in females, light blue - infertile or sub-fertile in males, dark green - infertile or sub-fertile in both males and females. The sterility phenotype was annotated using www.mousephenotype.org.

4.6 Transcriptional dynamics during spermiogenesis

Once the meiotic divisions result in the production of four haploid cells, round spermatids progress to form first elongating and finally mature sperm during a process termed "spermiogenesis" **(Fig. 4.9A)**. A key event during spermiogenesis is chromatin condensation, which is required to package the haploid genome into the confined space of the sperm nucleus. Our data allowed us to dissect at high-resolution the transcriptional regulation needed for gradual chromatin remodelling during spermatid differentiation, involving the replacement of canonical histones by histone variants followed by transition proteins and eventually protamines [458, 459]. This chromatin remodelling later induces a transcriptional shut-down where changes in RNA content are purely driven by degradation [425].

4.6.1 Expression of chromatin components during spermiogenesis

We first explored how expression of histone variants changed throughout early spermatid maturation **(Fig. 4.9A)**. Similar to the developmental ordering presented in the previous section, we ordered cells by fitting a principal curve to the first three principal components calculated on S1-S14 spermatids. Annotations for histone variants and canonical histones were taken from El Kennani *et al.*, 2017 [459]. Multiple variants of H3 and H2A are expressed in spermatocytes [460, 449, 461], and our data showed that many of these histones are highly expressed in early round spermatids. For instance, Histone H3.3 is a histone variant consisting of two genomic copies (*H3f3a* and *H3f3b*). Across spermatogenesis, we observed distinct expression patterns for the two genes, with *H3f3a* being consistently highly expressed until the transcriptional shut-down at spermatid stage S10. In contrast, *H3f3b* showed a much more dynamic expression profile, starting at a high level in spermatocytes, dropping throughout meiotic prophase, followed by up-regulation in round spermatids **(Fig. 4.9B)**. Although both genes have been implicated in male fertility, the phenotypes associated with perturbations of the more dynamically regulated paralog *H3f3b* are much more severe [461, 462].

When profiling the expression of canonical histones, we detected increased expression of *Hist1h2bp* and *Hist1h4a*, with a distinct up-regulation during early and mid-spermiogenesis **(Fig. 4.9C)**. Canonical histones are typically transcribed in a replication-dependent manner during S phase [463], thus the atypical expression during spermiogenesis could suggest important roles as replacement histones during chromatin remodelling. Nevertheless, canon-

ical histones appeared to be the set of annotated histones that is least correlated to the developmental trajectory **(Fig. 4.9A)**.

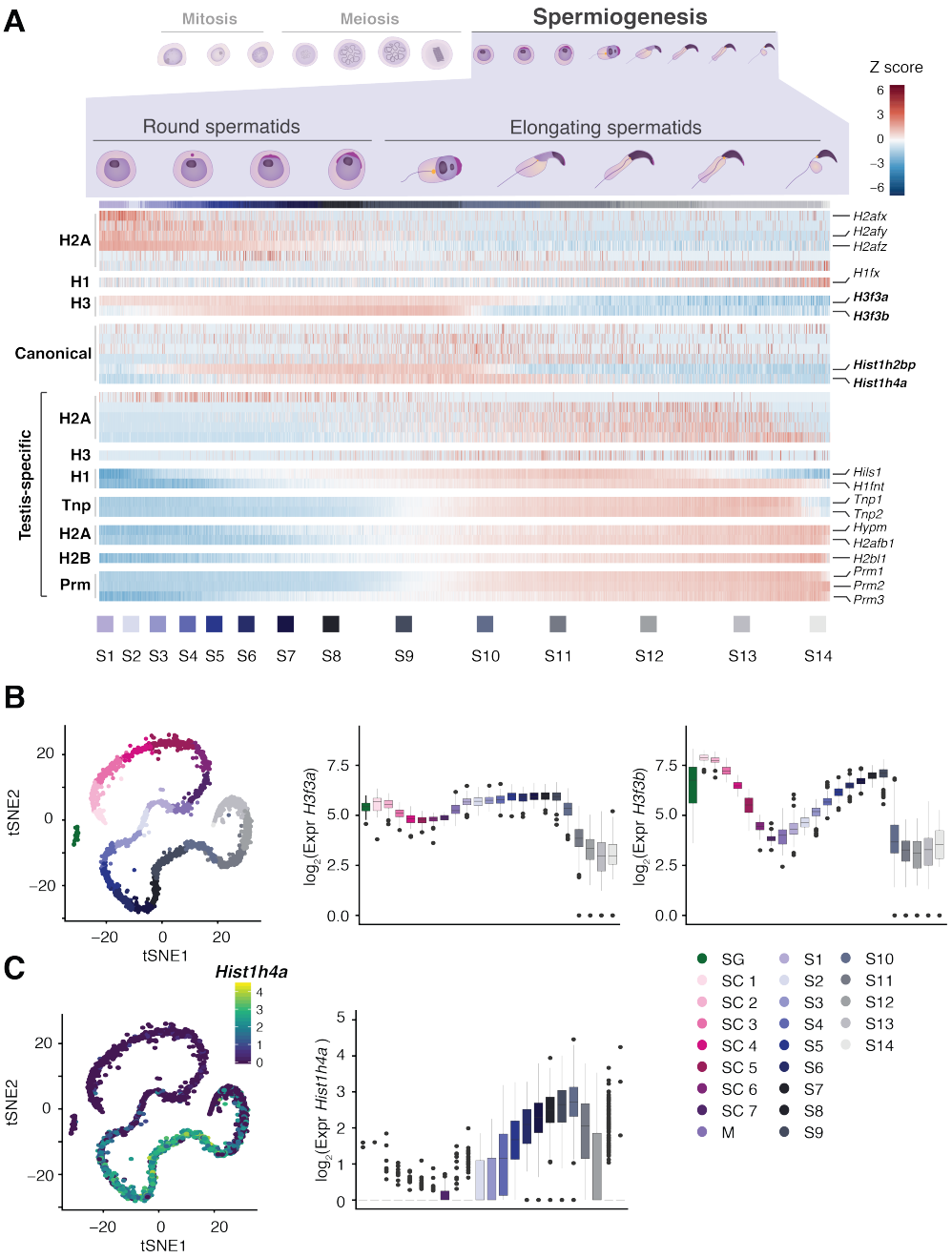


Fig. 4.9: Transcriptional dynamics and chromatin remodelling during spermiogenesis.
(A) Z score scaled, normalised expression of histone variants (H1, H2A, H2B, H3), canonical histones, transition proteins (Tnps) and protamines (Prms) during spermiogenesis. Cells were ordered based on their developmental trajectory ranging from round spermatids (S1-S8) to elongating spermatids (S9-S14), **(B)** Expression of *H3f3a* (middle panel) and *H3f3b* (right panel) across the different germ cell populations, **(C)** Similar visualisation as in (B) for *Hist1h4a* expression across germ cells.

We next profiled the transcriptional dynamics of testis-specific histone variants. They showed highest expression in elongating spermatids, with most variants increasing strongly in expression from S5 onwards. While some variants had a consistently high expression level, *Hils1* and *H1fnt* decreased in expression towards the late stages, similarly to *Tnp1* and *Tnp2* [464]. Both histone variants are important for male fertility, and *Hils1* has previously been shown to interact with *Tnp1* [465]. In contrast, three testis-specific histone variants *Hypm*, *H2afb1* and *H2bl1* (*1700024p04rik*) showed consistently high expression until the end of differentiation similar to protamines, suggesting these variants may contribute to the final genome condensation.

4.6.2 Identifying the point of transcriptional shut-down

As a consequence of chromatin condensation, transcription ceases in spermatids at the round to elongating switch, consistent with the lack of active RNA Pol II at S10 and later stages [466]. By fitting a smooth regression (loess) to the number of genes expressed per cell along the differentiation trajectory, we easily identified the point of transcriptional shut-down. The number of expressed genes is stable until approximately S9 before gradually declining by roughly 50% **(Fig. 4.10A)**. In the 8 days following transcriptional shut-down, spermatids still need to undergo drastic morphological changes, including the assembly of sperm-specific structures such as the flagellum, before mature testicular sperm can be released into the lumen [467]. To achieve this in the absence of active transcription, spermatids store large amounts of mRNAs in a perinuclear RNA granule termed the chromatoid body [468]. RNA stored in the chromatoid body is then released for translation, suggesting that these molecules may play vital roles during late stages of spermiogenesis. However, identifying the RNAs that are stored has been hindered by difficulties in purifying late spermatids.

By correlating normalised gene expression against the number of genes expressed, we identified a large number of genes that gradually decrease in relative expression after transcriptional shut-down. We reasoned that transcripts for which the relative expression after transcriptional shut-down appeared to increase are likely protected from degradation **(Fig. 4.10B)**. This included genes with well-known spermiogenesis-specific functions. Among those that relatively increase in expression, we find transition proteins and protamines that are involved in chromatin condensation. Furthermore, we detect genes that are involved in the development of sperm motility such as A-kinase anchoring protein 4 (*Akap4*) and calcium binding protein, spermatid associated 1 (*Cabs1*) [469, 470].

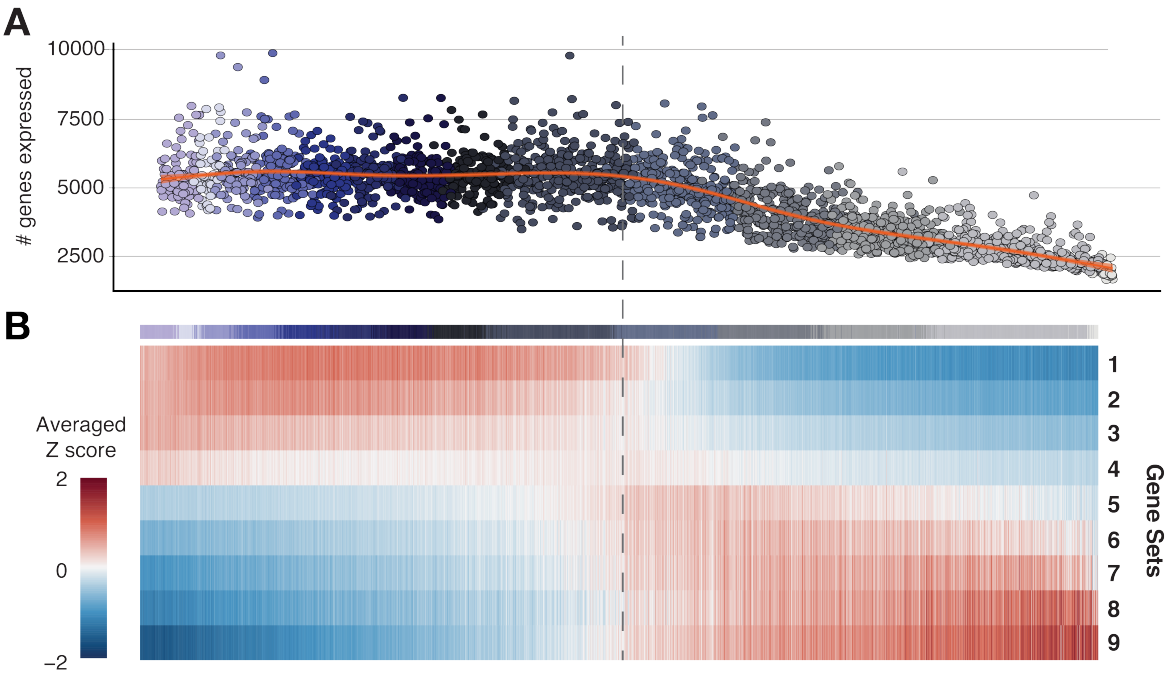


Fig. 4.10: Transcriptional shut-down during spermiogenesis.
(A) Number of genes expressed per spermatid. Cells were ordered based on their developmental trajectory. Red line indicates a smooth regression (loess) fit, **(B)** For each gene, its normalised expression per cell was correlated with the number of genes expressed per cell. Genes were ordered based on the correlation coefficient and grouped into 9 sets. Z score scaled expression was averaged across genes within each gene set. Vertical dashed line indicates transcriptional shut-down between S9 and S10.

With this analysis, we explored transcriptional processes occurring throughout the process of spermiogenesis that (i) regulate the expression of chromatin components and (ii) lead to the degradation of unneeded transcripts.

4.7 Meiotic silencing dynamics of sex chromosomes

A male-specific feature of meiosis is the transcriptional silencing of sex chromosomes, followed by a partial reactivation in post-meiotic spermatids. This process is termed meiotic sex chromosome inactivation (MSCI), and is caused by asynapsis of the sex chromosomes, leading to accumulation of phosphorylated H2A histone family member X (H2afx) and the formation of the sex body [471] **(Fig. 4.11A)**. We next profiled transcriptional changes mediated by the inactivation and reactivation of the sex chromosomes in single-cell and bulk RNA-Seq data.

To assess overall transcriptional dynamics of the sex chromosomes, we computed the ratio of expression from the X and Y chromosome and chromosome 9 to all autosomes. For this, we selected genes that were expressed in more than 30% of spermatogonia or 30% of spermatids, the cell types with detectable sex chromosome expression. For each cell, the mean expression across these genes per chromosome was calculated. Mean expression of the sex chromosomes and chromosome 9 was divided by mean expression across all autosomes. By plotting the ratio of gene expression from the X or Y chromosomes compared to all autosomes, the inactivation and re-activation status of the sex chromosomes can be inferred **(Fig. 4.11B)**.

The X chromosome is partially up-regulated in spermatogonia as described by Sangrithi *et al.*, 2017 (X chromosome:Autosome (X:A) ratio < 1) [472]. This is followed by transcriptional silencing in spermatocytes. Throughout spermiogenesis, expression from the X chromosome gradually increases, reaching X:A ratios comparable to spermatogonia, therefore suggesting a substantial reactivation of the X chromosome in post-meiotic spermatids. We detect similar behaviour for the Y chromosome but due to the small number of expressed genes, the signal is noisier **(Fig. 4.11B)**. In comparison, chromosome 9 shows consistent expression across all cell types throughout spermatogenesis (9:A $\approx$ 1).

Transcriptional silencing was originally thought to persist throughout post-meiotic development [460, 473]. However, several genes have been shown to be re- or *de novo* activated in spermatids, some of which are dependent on ring finger protein 8 (*Rnf8*) and/or sex comb on midleg-like 2 (*Scml2*) [474–476]. The precise timing and order of the transcriptional reactivation of *de novo* escape genes during spermiogenesis has not been explored. We therefore first classified *de novo* activated escape genes using bulk RNA-Seq

data and profiled their temporal expression directly following meiosis.

Profiling whole-testis transcriptomes of juvenile mice sampled every two days during the first wave of spermatogenesis allowed the sensitive detection of spermatid-specific escape genes **(Fig. 4.1A)**. Due to the gradual emergence of germ cell types during the first spermatogenic wave, differential expression analysis between early ($\leq$ P20) and late (> P20) time points revealed genes exclusively expressed in spermatids and which are thus *de novo* activated escape genes (n = 128) **(Fig. 4.11C)**. We used *edgeR* to identify differentially expressed genes between these conditions [386]. Spermatid-specific genes are identified with a $\log_2$-fold change > 5 in samples after day 20 compared to samples before day 20 (controlling the FDR to 10%).

Within the set of *de novo* activated escape genes we find many of the previously annotated escape genes such as *Cypt1*, *Cycl1*, and *Akap4*. Interestingly, this set of genes show an enrichment for targets of H3K27 acetylation which is mediated by *Rnf8* or *Scml2* (Fisher's Exact Test: *Rnf8*-targets, p-value < $5\text{x}10^{-12}$; *Scml2*-targets, p-value < $2\text{x}10^{-9}$) **(Fig. 4.11C)**. This chromatin mark represents active enhancers necessary for the reactivation of gene expression in spermatids [477].

While the bulk RNA-Seq data is ideal for identifying spermatid-specific, *de novo* activated genes, it lacks the temporal resolution to differentiate between early and late reactivated genes. We therefore ordered the 128 *de novo* activated genes based on their peak in expression using the scRNA-Seq data **(Fig. 4.11D)**. The *de novo* activated genes across our single cell RNA-Seq dataset showed a broad range of temporal expression patterns. The earliest expression, directly following meiosis and lasting until stages S4-S5 was observed for three members of the synovial sarcoma, X member B, breakpoint (*Ssxb*) multi-copy gene family (*Ssxb1*, *Ssxb2*, *Ssxb3*). Multi-copy genes have previously been described to have spermatid-specific expression [478], and their ampliconic structure has been speculated to play a role in escaping meiotic silencing via self-pairing [479].

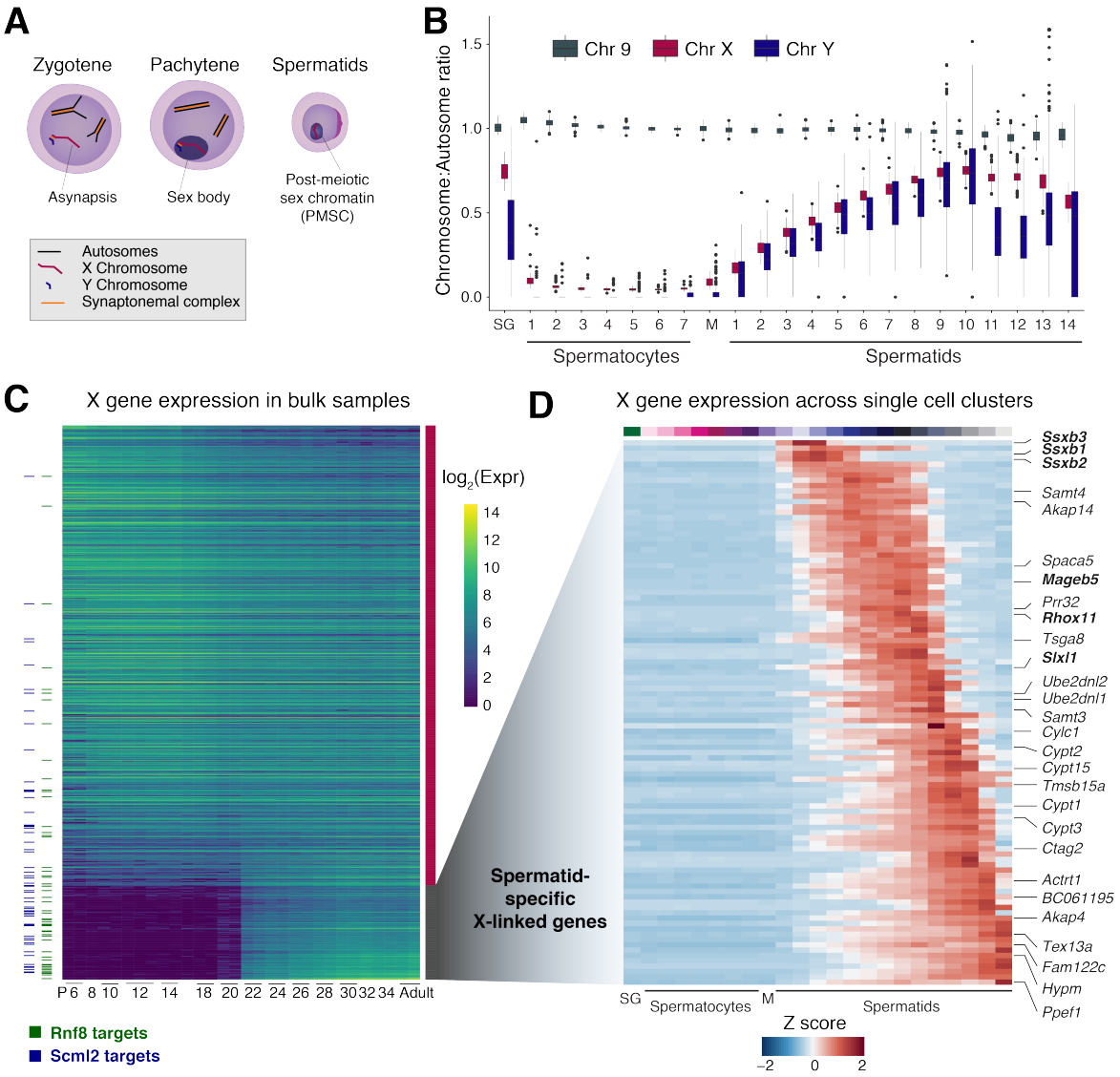


Fig. 4.11: X chromosome dynamics during spermatogenesis.
(A) Schematic of sex chromosome sub-nuclear localisation through spermatogenesis, **(B)** For each cell, the ratio of mean expression of genes on Chr 9, Chr X and Chr Y to the mean expression of genes across all autosomes is represented as a boxplot for cells allocated to each developmental stage. SG: Spermatogonia, M: Metaphase, **(C)** Expression of all X chromosome genes (> 10 average counts) in bulk RNA-Seq data across the juvenile time course. Columns correspond to developmental stage and rows are ordered by the $\log_2$ fold change between spermatocytes (stages before and including postnatal day (P) 20) and spermatids (stages after P20). Horizontal dashes indicate genes that are targets of *Rnf8* (green) and *Scml2* (blue) [477], **(D)** Average expression of spermatid-specific genes (panel (C)) per germ cell type. Columns are ordered by developmental stage and rows are ordered by peak gene expression through development. Multi-copy genes are highlighted in bold.

4.8 Epigenetic mechanisms of X chromosome reactivation

After identifying *de novo* activated escape genes, we next profiled the epigenetic basis that might underpin such transcriptional dynamics. For this, we profiled the chromatin landscape in spermatocytes and spermatids using the newly developed CUT&RUN protocol for low cell numbers **(Appendix A.2)** [403].

4.8.1 CUT&RUN to profile H3K4me3 and H3K9me3 marks

In brief, from two individuals, we sorted spermatocytes and spermatids at P26 during the first wave of spermatogenesis **(Fig. 4.12A)**. At this stage, tubules contain spermatocytes close to the meiotic cell divisions and elongating spermatids. We assayed H3K4me3 as a proxy for promoter activity, as well as repressive H3K9me3 mark. By profiling the enrichment of H3K9me3 across all chromosomes, we confirmed that the X chromosome has high levels of H3K9me3 in spermatids which has been previously shown [480, 460, 481]. In addition, we now show that H3K9me3 accumulation begins earlier in meiosis, and indeed spermatocytes show enrichment of this repressive mark on the X chromosome **(Fig. 4.12B)**.

On autosomes, H3K9me3 is enriched in pericentromeric regions of constitutive heterochromatin [482]. To assay the distribution of read pairs across whole chromosomes, we binned reads in 1kb windows across the chromosome. Next, we calculated the cumulative sum across 10,000 randomly sampled bins starting at windows with the highest H3K9me3 enrichment. This measure indicates whether each window contains equal enrichment (slope is similar across the curve) or if some windows are enriched for the mark (slope decreases across the curve). As seen in **Fig. 4.12D**, the H3K9me3 enrichment appears to be homogeneously distributed across the X chromosome while, for example, the enrichment of the H3K9me3 mark on chromosome 9 is a lot more heterogeneous.

Nevertheless, when merging the 1000 windows with highest H3K9me3 enrichment, we detected broad regions showing particularly high levels of H3K9me3 scattered across the X chromosome **(Fig. 4.12C)**. Among the merged regions with highest H3K9me3 enrichment, we detect the promoter of *Akap4*, a well-known escape gene. This discovery prompted us to profile the chromatin dynamics of active and repressive marks at promoters of *de novo* escape genes (*spermatid-specific genes*) versus the promoters of all other expressed X-chromosome genes (*non-spermatid specific genes*) **(Fig. 4.11C)**.

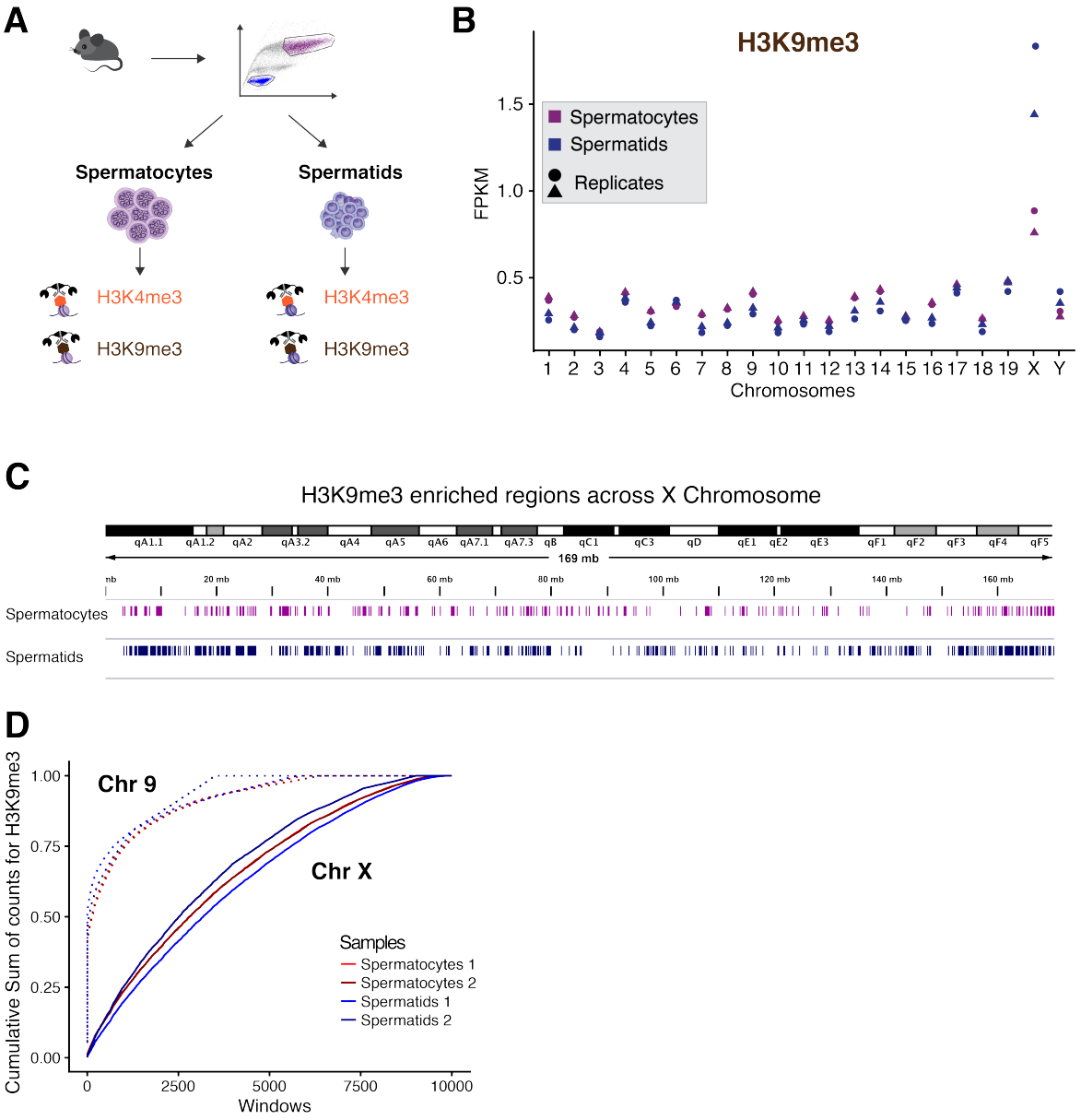


Fig. 4.12: Chromatin profiling in spermatocytes and spermatids.
(A) Spermatocytes and spermatids were isolated from the same individual using FACS and profiled using H3K4me3 (active mark) and H3K9me3 (repressive mark) using CUT&RUN, **(B)** Number of H3K9me3 Fragments Per Kilobase per Million (FPKM) for each chromosome. Pink: spermatocytes, blue: spermatids. Shape corresponds to biological replicate, **(C)** The top 1000 windows with highest H3K9me3 signal (1000 bp width, CPM) were merged using a tolerance of 1500 bp. Representative tracks of one replicate in spermatocytes and one replicate in spermatids are shown, **(D)** Cumulative summed counts per million across 10000 randomly sampled windows (1000 bp width) visualising the distribution of the H3K9me3 signal across chromosome 9 (dashed line) and chromosome X (solid line).

4.8.2 Targeted silencing of spermatid-specific escape genes

Here, we profiled the enrichment for H3K4me3 and H3K9me3 marks at promoters of spermatid-specific escape genes and all other X-linked genes in spermatids and spermatocytes. As a measure for enrichment, we calculated the CPM for paired reads per promoter. In spermatocytes, spermatid-specific genes showed lower enrichment in H3K4me3 than non-spermatid specific genes (Wilcoxon-Mann-Whitney: p-value < $2.2\text{x}10^{-16}$) **(Fig. 4.13A, left panel)**. In contrast, spermatid-specific genes have on average elevated H3K4me3 in spermatids, as expected based on their increased expression level compared to spermatocytes **(Fig. 4.13A, right panel)**.

When examining the deposition of H3K9me3 on the promoters of X-linked genes, we detected a strong enrichment in spermatid-specific escape genes in spermatocytes (Wilcoxon-Mann-Whitney: p-value < $3.7\text{x}10^{-11}$) **(Fig. 4.13B, left panel)**. This pattern indicates that spermatid-specific genes are more strongly repressed in spermatocytes. Due to the strong enrichment of H3K9me3 on the post-meiotic X chromosome, we detect similar H3K9me3 enrichment in promoters for both spermatid-specific and non-specific X-linked genes **(Fig. 4.13B, right panel)**.

Our results describe the precise epigenetic changes associated with escape gene activation in post-meiotic cells. These dynamics are exemplified by the chromatin remodelling that occurs around *Akap4* and cysteine-rich perinuclear theca 1 (*Cypt1*), both of which are well-studied spermatid-specific genes **(Fig. 4.13C)**. The promoters of these genes have high levels of H3K9me3 in spermatocytes, which decreases in spermatids, while H3K4me3 levels are strongly increased. This targeted repression of a subset of X-linked escape genes could indicate a mechanism to repress otherwise lethal genes in spermatocytes that are later on needed in spermatid development. Examples of spermatocyte-lethal genes involved in spermatid development are two Y chromosome encoded genes: zinc finger protein Y-linked (*Zfy*) 1 and 2 [483].

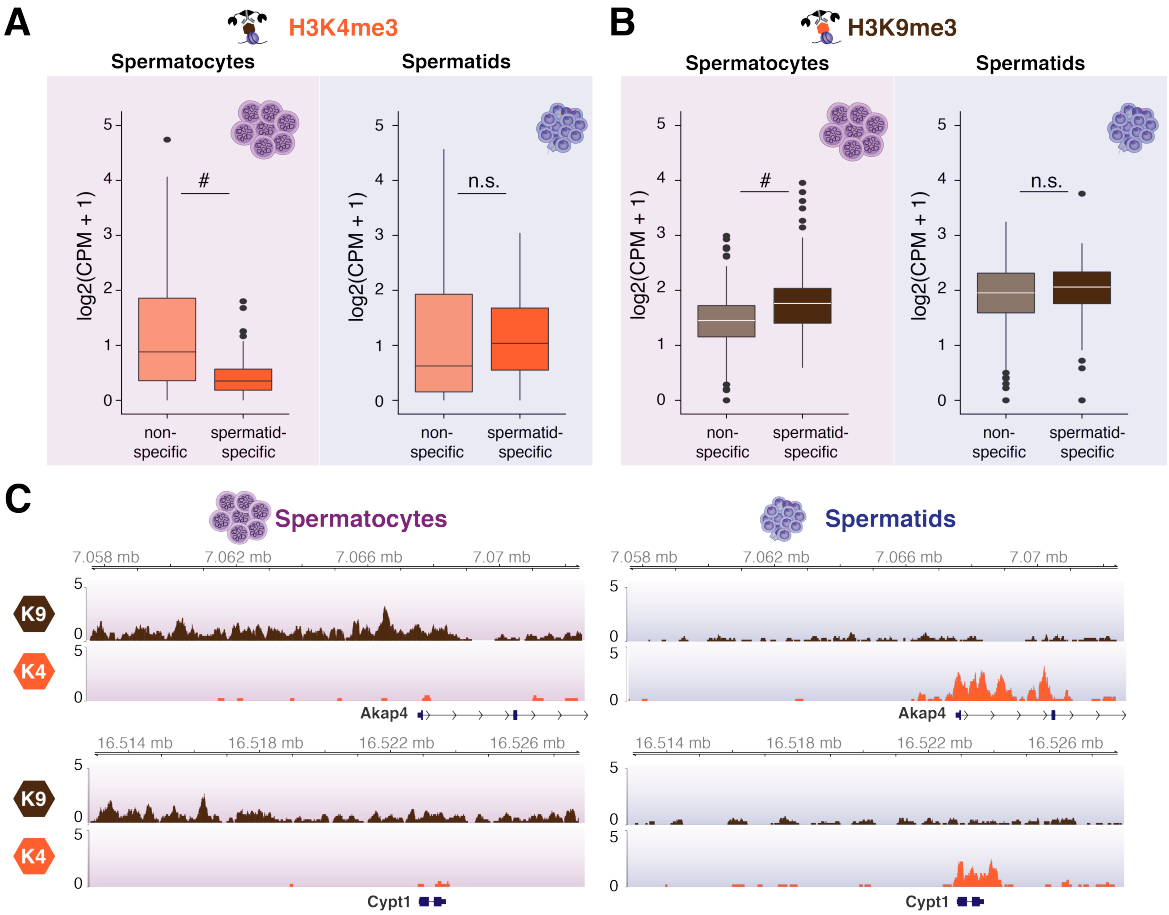


Fig. 4.13: Targeted repression of spermatid-specifc escape genes in spermatocytes.
(A) and **(B)** Boxplot of H3K4me3 (A) and H3K9me3 (B) Counts Per Million (CPM) in promoter regions of spermatid specific (n=127) and non-spermatid specific (n=617) genes for spermatocytes (left) and spermatids (right). # indicates statistical significance (Wilcoxon-Mann-Whitney: p-value < $1x10^{-10}$), n.s. – not significant, **(C)** Genome tracks of H3K4me3 and H3K9me3 for two representative spermatid-specific genes (*Akap4* and *Cypt1*) for spermatocytes (left) and spermatids (right). Reads were scaled by library size. The genomic location of these genes is indicated below the tracks where exons are labelled as blocks and the directionality of transcription is shown by arrows.

4.9 Measuring changes in variability over pseudo-time

As described above, spermatogenesis is a unidirectional and continuous differentiation process coupled to a complex system of developmental steps. I next asked whether this differentiation process is coupled to changes in transcriptional variability. In mouse haematopoietic cell differentiation, cell-to-cell diversity increases at critical state transitions where cell fate decisions are made [52]. A similar effect was detected in chicken erythroid progenitor cells where the Shannon entropy is highest directly at the point of fate commitment and declines upon the irreversible commitment to differentiation [15]. In the previous chapter, we have demonstrated that transcriptional variability shows dynamic changes during $CD4^+$ T cell differentiation with high variability being observed at a possible early commitment point and a decrease in variability upon proliferation. In this section, I applied the regression BASiCS model, which was developed in the previous chapter, to study changes in transcriptional variability over the time-course of spermatogenesis. More specifically, I profiled changes in variability for individual genes during spermiogenesis, the differentiation process that directly follows meiosis (see **Section 4.6**). As described above, spermiogenesis is a differentiation process that involves an extensive remodelling of the chromatin with transcriptional shut-down occurring at around spermatid stage S10. Modelling changes in expression over a differentiation time-course is done by ordering transcriptional profiles of individual cells along their so called *pseudo-time*. Different methods have been proposed to perform this ordering based on minimum spanning trees [30] and nearest-neighbour graphs [484], Gaussian Processes [485, 486] and diffusion maps [487]. Once the pseudo-temporal ordering is determined, genes that change in expression over pseudo-time can be found by fitting a generalised linear model to the expression counts and performing a likelihood ratio test against a null model with no pseudo-time dependence [30]. Profiling changes in variability is more complicated as single-cell measures of variability are not available.

4.9.1 Using BASiCS on continuous data

Here, I use BASiCS to estimate residual over-dispersion parameters for homogeneous cell populations along the differentiation time-course. Different approaches of identifying homogeneous populations exist. First, ordered cells can be split into populations of equal size (e.g. 200 cells per group). This approach produces heterogeneous cell populations when cell state transitions occur within the population. I therefore rely on the clustering performed in **Section 4.2.4** which splits the full cell population along the differentiation trajectory. For

each cluster from S1 to S14, the regression BASiCS model was run for 40,000 iterations with 20,000 iterations of burn-in and a thinning value of 20.

For each gene in each of the 14 spermatid populations, BASiCS generates a posterior distribution estimating the residual over-dispersion parameter in form of an MCMC chain **(Fig. 4.14A)**. These measures are independent of mean expression (see previous chapter) and can therefore be used to study changes in variability which are not confounded by changes in mean expression throughout the differentiation of sperm. I chose two approaches to profile and test temporal changes of transcriptional variability during spermiogenesis.

First, I used the iterative fitting of a linear regression model between the residual over-dispersion parameters and the progression of spermiogenesis to find linear changes in variability. For this, I selected spermatids from stages S1 to S9 prior to transcriptional shut-down. Transcriptional changes after S10 are only due to degradation of mRNA and I assume that linear changes in variability occur before S10. In more detail, for each MCMC iteration, I fit a linear regression model between the current samples of ε_i against the cluster label **(Fig. 4.14B)**. This fitting is performed for each gene individually and generates a *post hoc* distribution of the intercept and the slope regression coefficient that captures uncertainty in the regression fit. Focusing on the slope coefficient, I can compute the posterior tail probability of the slope coefficient being different from 0. If the posterior tail probability is larger than a threshold (e.g. 80%), I consider the transcriptional variability of this gene to be either positively or negatively associated with temporal ordering depending on the sign of the median slope coefficient **(Fig. 4.14B)**. Similar to differential testing described in the previous chapter, the probability threshold is determined by fixing the expected false discovery rate to 10%. A similar testing can be done for the slope coefficient when fitting a linear model between the group wise mean expression parameter $\log(\mu_i)$ and the group labels.

Secondly, to detect non-linear patterns of changes in transcriptional variability, I perform clustering on the gene-specific variability profiles across spermatid populations S1-S14. Similar approaches have been used to find patterns of genes expression across pseudo-time. Common patterns for changes in expression levels include immediate, transient and gradual up- or down-regulation [30]. When profiling changes in variability over the time-course of differentiation these clustered profiles can indicate similarly strong or weak transcriptional regulation or similar expression rates **(Fig. 4.14C)**.

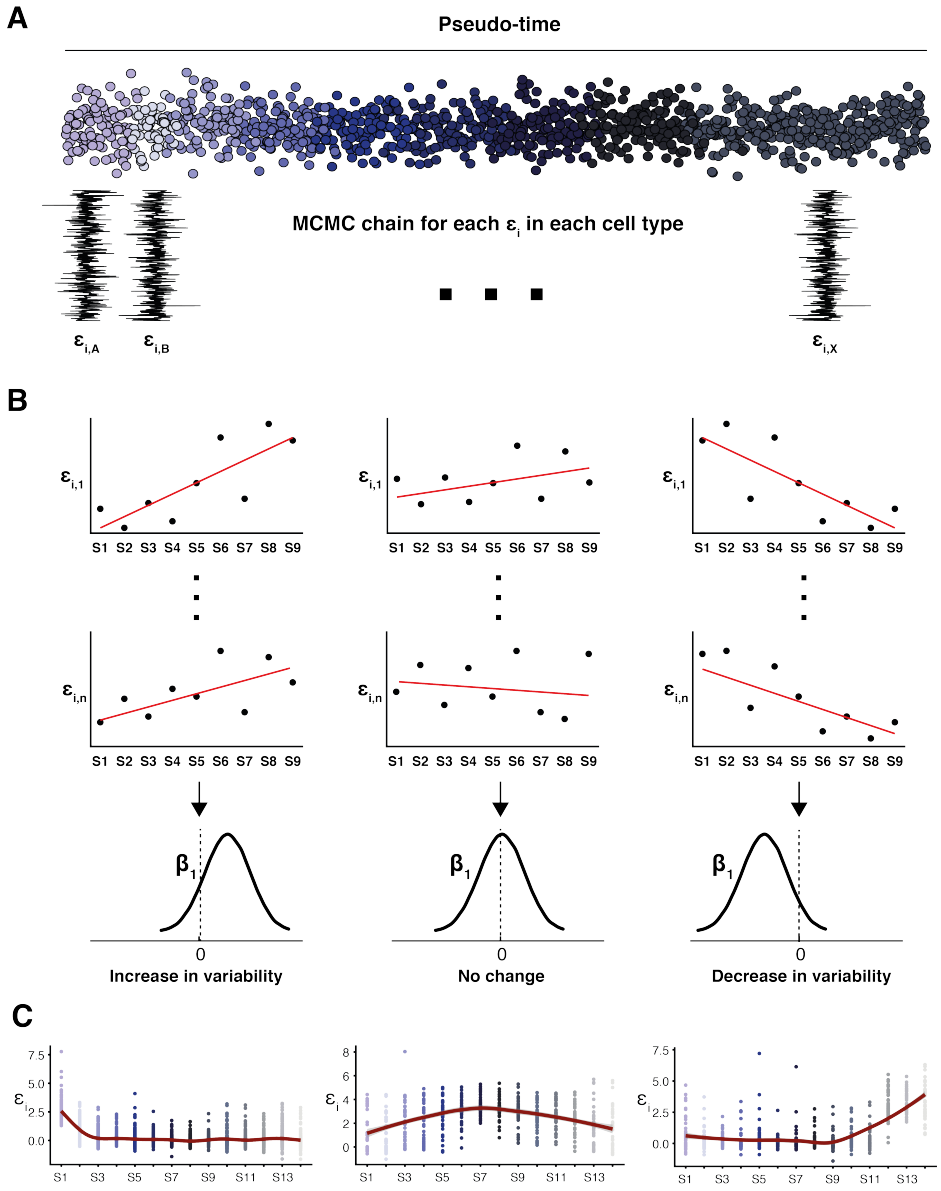


Fig. 4.14: Detecting changes in variability over pseudo-time.
(A) For each group of spermatids, BASiCS generates a posterior distribution of residual over-dispersion parameters ε_i. Cell groups can be ordered based on their pseudo-time (upper panel). Lower panels indicate the MCMC chain for gene-specific ε_i per group (A, B, ..., X), **(B)** For each iteration of the MCMC (1,...,n), a linear regression was fit between the current samples of ε_i against the group labels for spermatids (S) 1-9. This approach generates a *post hoc* distribution of the slope coefficient β_1 (lower panels). The distribution is used to calculate the posterior probability of observing $\beta_1 \neq 0$, **(C)** Clustering was performed on variability profiles across spermatid populations S1 to S14. A smooth regression (loess) was fit to the median ε_i's of the genes within each cluster. Genes that quickly decrease in variability (left panel), increase then decrease in variability (middle panel) or quickly increase in variability (right panel) can be identified.

4.9.2 Finding continuous changes in variability by linear model fitting

To detect single genes that continuously increase or decrease in variability, I fit a linear regression model to each iteration of the MCMCs sampling ε_i or μ_i *versus* the group labels **(Fig. 4.14B)**. The posterior distributions of the slope coefficient were used to categorise genes based on their transcription dynamics along the differentiation time-course (middle panel in **Fig. 4.15**). These categories include:

- Increase in mean expression, no change in variability
- Increase in mean expression, increase in variability
- Increase in mean expression, decrease in variability
- Decrease in mean expression, no change in variability
- Decrease in mean expression, increase in variability
- Decrease in mean expression, decrease in variability
- No change in mean expression, no change in variability
- No change in mean expression, increase in variability
- No change in mean expression, decrease in variability

This approach leads to the detection of few genes that significantly change in variability over the differentiation time-course in a linear fashion while the majority of genes change only in mean expression. One hypothesis is that sperm maturation is a tightly regulated progress where the majority of genes follow a clear transcriptional pattern. Such a process contrasts with other differentiation programmes such as haematopoiesis where branching events occur and the whole cell population expands in transcriptional variability to find new attractor states [52]. To visualise changes in transcriptional variability, I selected representative genes from four categories: (i) Increase in mean expression, increase in variability, (ii) Increase in mean expression, decrease in variability, (iii) Decrease in mean expression, increase in variability, (iv) Decrease in mean expression, decrease in variability (see insets in **Fig. 4.15**). Interestingly, testis specific gene A8 (*Tsga8*), one of the most rapidly evolving X-linked genes, shows a strong increase in expression and a clear decrease in transcriptional variability. *Tsga8* has been reported to be involved in hybrid sterility where F_1 crosses of mice form different strains are unable to reproduce. This effect might be due to the strong divergence of the *Tsga8* sequence between species [488]. A tight regulation of its expression during spermiogenesis can therefore further control the phenotypic effect in F_1 animals.

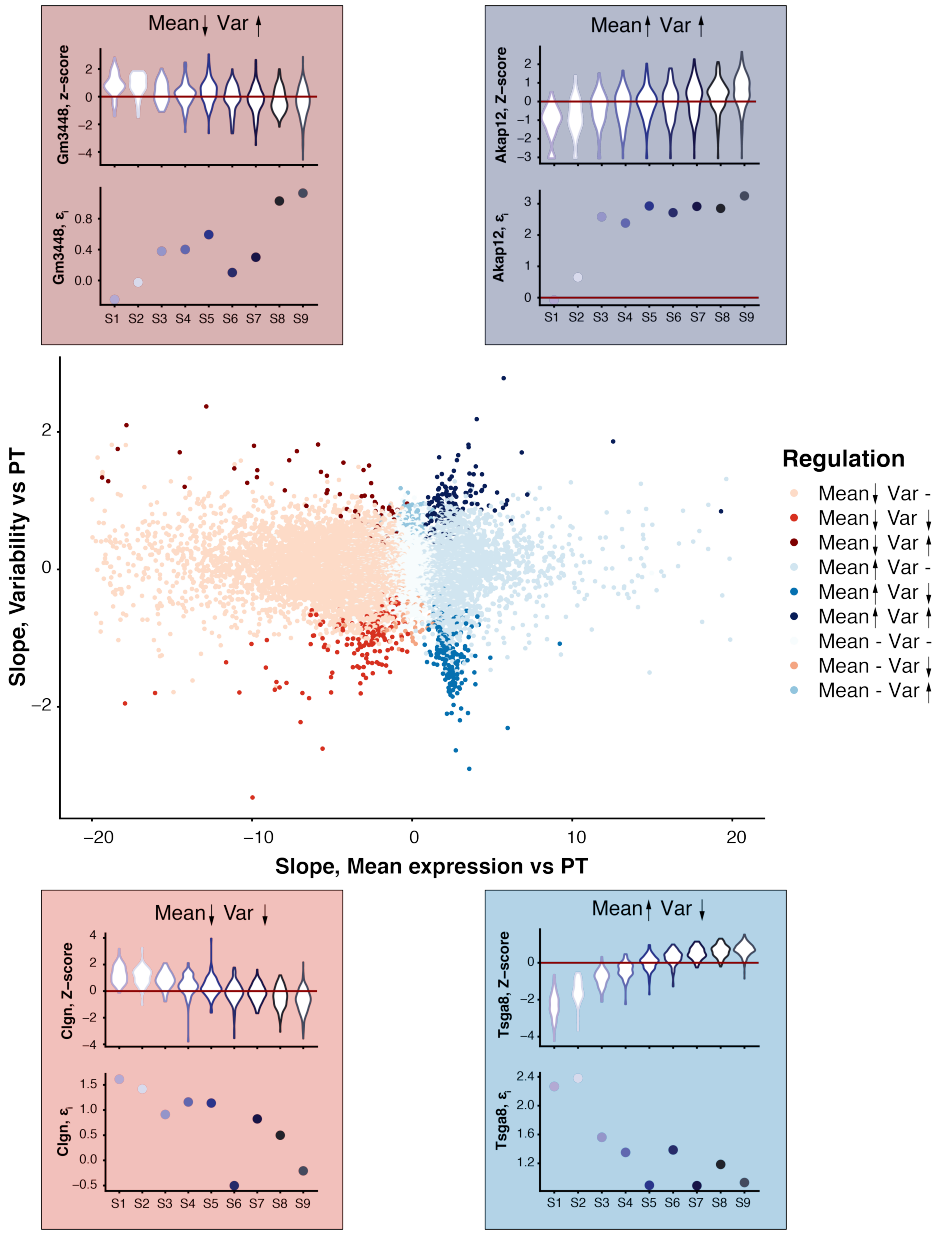


Fig. 4.15: Linear changes in variability over spermiogenesis.
Linear models were fit between the residual over-dispersion parameter ε_i or the mean expression parameter μ_i and the groups labels (S) 1-9 for each iteration of the MCMC. Median posterior estimates of the slope parameter of the variability fit were plotted against the slope parameter of the mean expression fit. Each dot represents a single gene. G enes are coloured based on their regulation (legend). Plot insets indicated the Z score scaled normalised expression (upper panel) and the median group-wise residual over-dispersion estimates ε_i (lower panels) of representative genes for four categories.

4.9.3 Clustering of variability profiles

To identify non-linear patterns across all genes, I first ordered variability profiles based on their peak variability **(Fig. 4.16A)**. Here, variability profiles are represented by the median residual over-dispersion parameter ε_i ordered from S1 to S14. Most variability profiles showed highest variability in one group, albeit other patterns of variability are also detectable.

To identify the major patterns of variability across the full range of spermiogenesis, I performed k-means clustering across all variability profiles. In this case, it was required to select the expected number of clusters. Due to the fact that most genes showed peak variability in exactly one group, I selected $k = 20$ to detect patterns other than peaks in single groups. After clustering, I detect a variety of variability patterns ranging from high variability in early spermiogenesis to high variability at later stages **(Fig. 4.16B)**. Interestingly, I observed patterns that show gradual increase in variability until around spermatid stage S9 and decrease afterwards **(Fig. 4.16B, middle panel)**.

The group with peak variability at around S9 contains all transition proteins (*Tnp1*, *Tnp2*) and protamins (*Prm1*, *Prm2*, *Prm3*). When visualising the expression patterns of *Prm1*, I detect a rapid shift in expression for cells from S9 **(Fig. 4.16C, middle panel)**. Similarly, genes that show the highest variability at later stages of spermiogenesis **(Fig. 4.16B, second to last panel)** show a quick transcriptional decline after transcriptional shut-down (e.g. *Tekt4*, **Fig. 4.16B**).

These results indicate that the changes in expression associated with the trajectory of pseudo-time are additional confounding factors when quantifying transcriptional variability. Similar to removing the confounding between mean expression and variability, a regression approach can be used to correct variability measures based on the correlation between expression and pseudo-time.

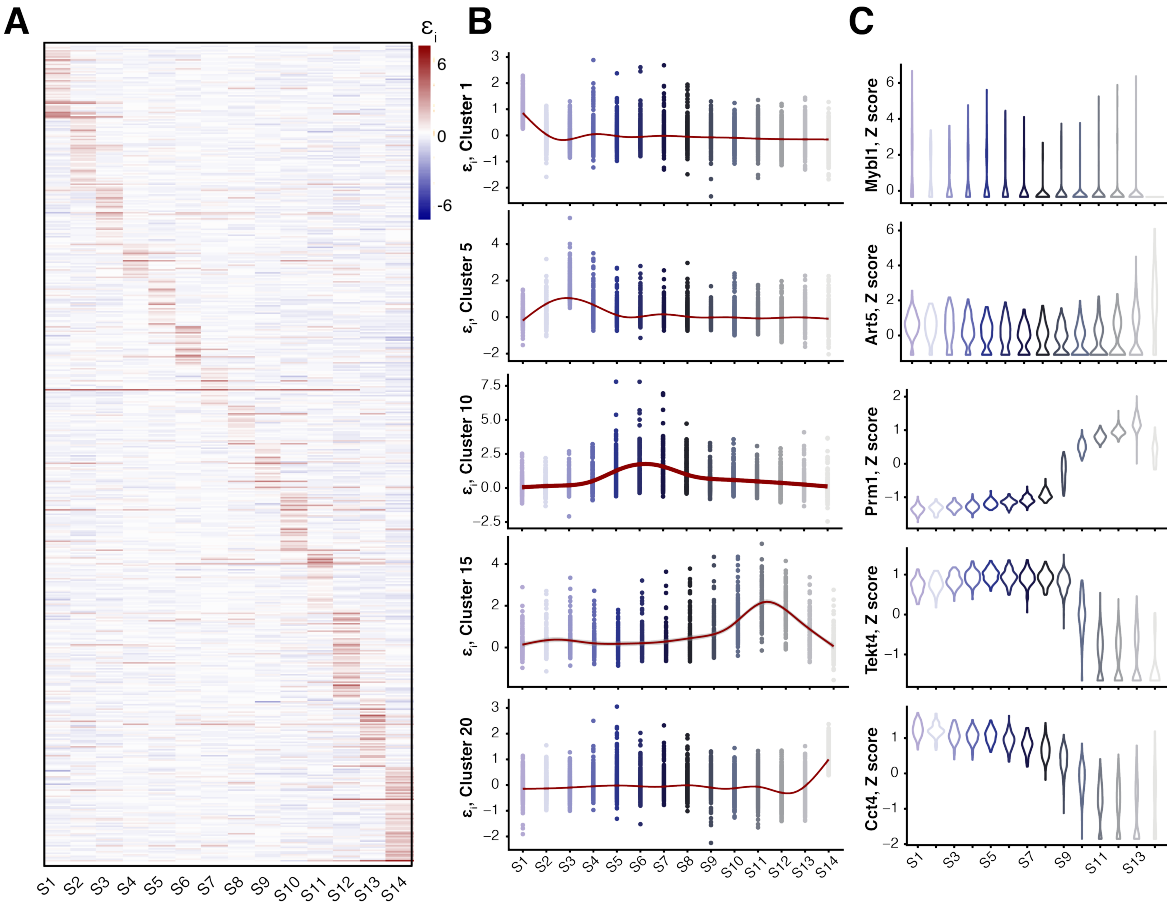


Fig. 4.16: Clustering of variability profiles.
(A) Variability profiles (median of the ε_i estimates ordered by developmental progression) were ordered based on their maximum ε_i starting in S1 spermatids, **(B)** Variability profiles were clustered using k-means with k = 20. 5 representative patterns of variability are displayed ranging from highest variability in round spermatids to highest variability in elongating spermatids, **(C)** Z score scaled, normalised expression of example genes per variability pattern taken from (B) are displayed in the form of boxplots.

4.10 Discussion

The testes are among the most proliferative tissues in the adult body and ensure fertility via the continuous production of millions of sperm per day. Most developmental differentiation processes require the profiling of cellular populations at several time points [227, 233, 235]. One of the exemptions is blood formation where commitment to different lineages can be profiled at once [29]. Similarly, spermatogenesis occurs in continuous waves throughout the reproductive life span of animals. At any given time point, all intermediate cell types that arise across the 35 day differentiation program are present in adult testes. This provided a powerful opportunity to capture and profile an entire differentiation process by profiling the transcriptomes of thousands of single-cells at a single time point.

We exploited the natural synchronisation of the first wave of spermatogenesis to identify key developmental transitions within the differentiation trajectory. In contrast, Chen *et al.*, 2018 sorted synchronised spermatocyte and spermatid populations after blocking spermatogenesis with WIN 18,446. This allowed a strict enrichment for cells in specific stages during spermatogenesis but lost the natural trajectory of this continuous differentiation process [489]. Profiling spermatogenesis in juvenile animals also naturally enriched for rare cell types that are under-represented in adults. In the case of haematopoiesis, cells need to be sorted to capture otherwise under-represented cell types [29]. Among these rare cell types, spermatogonia are of particular interest as these cells not only sustain male fertility, but are also the origin of the vast majority of testicular neoplasms [490]. We obtained more than 1,100 transcriptional profiles for spermatogonia, allowing the identification of specific cell clusters within this heterogeneous cell population thus greatly improving the resolution over previous studies that only studied adult testes [437]. Furthermore, our approach also enriched for and facilitated characterisation of the complexity within testicular somatic cell types. Among those are characteristic immune cells and precursor cells that only exist until a few days after birth.

Droplet-based scRNA-Seq can profile large number of cells simultaneously [163, 164, 178], but often captures cells with a wide range of transcriptional complexity. Consequently, droplet-based assays present a major computational challenge in distinguishing between (i) droplets contain transcriptionally inactive cells versus (ii) empty droplets that contain (background) ambient RNA. By using a stringent default threshold, we identified the majority of somatic and germ cell types in testes, similar to recent single-cell expression

studies in mouse and human [437, 454, 489]. In addition, we applied a statistical method to identify cells from droplet-based data by comparing the ambient RNA profiles [400], and were able to identify transcriptionally inactive leptotene/zygotene spermatocytes. This allowed us to bridge the developmental transition between spermatogonia and spermatocytes, thus providing a more complete view of the continuum of germ cell differentiation.

After the in-depth characterisation of germ and somatic cell types in adult testes, we profiled major developmental processes that occur during mouse spermatogenesis. During meiosis, we detect the expression of hundreds of genes associated with the developmental trajectory. Some of these genes show a sterility phenotype when perturbed and we reason that this is also the case for the majority of genes that follow the developmental trajectory in expression. Spermiogenesis is characterised by wide-scale chromatin rearrangements and we detect a clear increase in testis-specific histone variants, transition proteins and protamines during the late stages of sperm maturation. Again, genes that follow this trend could be important regulators that would cause sterility upon misexpression.

The transcriptional silencing of the sex chromosomes during meiosis and their subsequent partial re-activation post-meiosis is essential for male fertility [491]. Failure of MSCI results in the expression of spermatocyte-lethal genes, as demonstrated for two Y chromosome encoded genes: *Zfy* 1 and 2 [483]. Our discovery that H3K9me3 is enriched during meiosis at spermatid-specific genes suggests a stronger, targeted repression in spermatocytes for a key subset of X-linked genes. The deposition of H3K9me3 is specific to MSCI in males, and is not observed during general meiotic silencing of unpaired chromosomes [492–494]. Our finding that spermatid-specific genes are particularly enriched for H3K9me3 in spermatocytes suggests that their repression may be necessary for male fertility.

When profiling changes in variability over the differentiation trajectory, I detected a strong confounding effect between the variability measure and the correlation between expression and pseudo-time. Therefore, new measures of variability need to be derived to account for this dependency. For example, graph-based measures can assign a variability measure for each cell when comparing expression across a local neighbourhood. Next, fitting a generalized linear model between these variability estimates and the ordering of cells along pseudotime can be used to detect changes in variability. Nevertheless, confounding effects such as the expression level can obstruct such analysis.

5

Conclusion and future directions

My work focused on the statistical quantification of transcriptional noise in biological systems such as the activation response of $CD4^+$ T cells. Firstly, in collaboration with Celia P. Martinez-Jimenez, we used scRNA-Seq data of $CD4^+$ T cells to identify an age-related increase in transcriptional noise within a set of immune response genes (see **Chapter 2**). Assessment of changes in transcriptional variability was restricted to genes that show similar expression levels in naive and active cells or young and old animals. I therefore extended the BASiCS statistical framework to correct for the confounding effect between mean expression μ_i and over-dispersion δ_i by introducing a joint prior that captures the dependence of δ_i on μ_i. The derivation of residual over-dispersion parameters ε_i allowed me to robustly test for changes in expression variability even when genes display changes in mean expression (see **Chapter 3**). Finally, in collaboration with Christina Ernst, we dissected the transcriptional programme underlying mouse spermatogenesis and characterised developmental processes such as spermatogonia differentiation, meiosis and spermiogenesis. We further identified a set of X-linked, spermatid-specifically expressed genes that show high enrichment of the repressive H3K9me3 mark in their promoter regions. After full characterisation of this differentiation process, I used the extended BASiCS model to identify changes in variability along spermatogenesis. Abrupt changes in mean expression display a confounding effect when measuring transcriptional variability along this time course (see **Chapter 4**).

While technological and computational advances of the recent years facilitate the quantification of biological noise across a range of cell types and tissues, major challenges remain regarding robust measurement, mathematical modelling and experimental validation. Here, I discuss the results of my work in light of current challenges in the field of scRNA-Seq when measuring biological variation across individual cells.

5.1 Technologies to study the biological role of noise

The results of **Chapter 2** indicate two different settings where changes in variability are either related to the synchronisation of a dynamic cellular system or the disruption of such a system. We explored the effect of ageing on transcriptional noise during immune activation using scRNA-Seq data. Early immune activation induces a transcriptional switch from stochastic to regulated gene expression coupled with a reduction in transcriptional variability. These dynamics and, more importantly, a set of immune-related response genes are conserved during evolution. While ageing only shows subtle effects on the overall transcriptomic profiles of individual cells, we observe a strong increase in expression variability in the core set of immune response genes during ageing. Therefore, transcriptional variability is a largely unexplored factor of organismal ageing. This finding has been validated by several studies [104, 228, 367] adding the increase in transcriptional noise to the list of ageing-associated physiological effects.

Our study uncovered transcriptional noise as a factor that disturbs the dynamic response of an otherwise tightly regulated system. The systematic analysis of how transcriptional noise globally influences other cellular systems such as the developing embryo or disease onset is still lacking. Examples of studies that identified a link between cell fate commitment and heterogeneous gene expression using scRNA-Seq data include the development of the 4-cell stage embryo towards extraembryonic and pluripotent cell lineages [16]. Furthermore, Mohammed *et al.*, 2017 identified global changes in transcriptional noise during early mouse embryo development that correlate with the plasticity of cell populations. Pluripotent cells tend to display noisier gene expression compared to committed cells [17]. Nevertheless, technical limitations restricted the analysis to few hundreds of cells and specific tissues per embryo. With the newly developed combinatorial indexing approaches, hundreds of thousands of transcriptomes can be generated in parallel [179]. This allows an unbiased detection of all major cell types during (e.g.) embryonic development, which in turn offers a great resource to perform systematic comparisons of transcriptional noise between tissues and time points. Major drawbacks of this approach would be the reduced sequencing depth and the inability to validate the global change in variability as discussed below.

In **Section 3.6**, I tested for changes in expression variability between the pre-somitic and the somitic mesoderm of the developing mouse embryo. Interestingly, this analysis revealed heterogeneous up-regulation of lineage-associated genes that are later on expressed in

defined tissues. This shows that testing for changes in expression variability can lead to the identification of uncharacterised, early commitment processes during embryogenesis. Nevertheless, scRNA-Seq data does not directly allow identification of the underlying transcriptional regulation that induces heterogeneous expression of these genes. It is therefore impossible to predict whether heterogeneity in expression is induced by molecular noise or driven by deterministic processes.

So far, quantification of expression noise on a genome wide scale is only possible by scRNA-Seq. This raises the question if noise that is detected on the mRNA level propagates to form fluctuations in proteins which are the final driver for phenotypic variations between individual cells. Reports have been published that show a reduction of transcriptional noise during nuclear export of mRNAs [69, 132] indicating the possibility that studying biological noise on the mRNA level is further buffered in the cytoplasm by mechanisms such as miRNA-based degradation [135]. In recent years, technologies have been developed to measure protein abundance in single cells in high-throughput and high-content based approaches. Cytometry by time-of-flight (CyTOF) has been introduced as a single-cell technology to measure multiple proteins within hundreds of thousands of cells. For this, antibodies against membrane bound and intracellular proteins are labelled with transition element isotopes and quantified via mass spectroscopy. So far, the main application of CyTOF has been to identify immune cell dynamics [495]. To add the spatial component to mass cytometry, Giesen *et al.*, 2014 developed imaging mass cytometry to obtain spatial distributions of 32 proteins in breast cancer samples [496]. A similar approach has been introduced by Gut *et al.*, 2018 where off-the shelf antibodies are used to spatially resolve protein expression. During 20 rounds of primary and fluorescently-labelled secondary antibody staining, multiplexed read-outs of protein positions can be obtained from individual cells [497]. The spatial detection of proteins has been extended by simultaneously measuring mRNA transcripts by isotope tagging [498]. These approaches allow (i) quantification of protein expression noise, (ii) spatially-resolved inter- and intra-cellular variations of protein abundance and (iii) the assessment of noise propagation from the mRNA to protein level. To further enhance the connection between mRNA and protein noise, and chromatin state and mRNA noise, multi-omics technologies need to advance in precision and scalability.

5.2 Confounding effects when measuring noise

We used the BASiCS framework to quantify and compare measures of transcriptional noise in the immune response of $CD4^+$ T cells. By incorporating reads of synthetic RNA spike-in molecules, BASiCS quantifies and removes technical noise from the total transcriptional variation. Throughout this thesis, we used the over-dispersion parameter δ_i to capture biological variability in expression after removal of unwanted technical variation. Furthermore, to account for experimental designs where cells were captured in multiple replicates, BASiCS scales technical noise batch-specifically [11]. We described a genes' mean expression as an additional factor that confounds testing changes in over-dispersion. Therefore, we extended the BASiCS framework to derive residual over-dispersion estimates that show no correlation to mean expression (see **Chapter 3**).

By applying this model to capture changes in variability over the differentiation time course of spermatogenesis, I observed that the strength of transcriptional changes over time introduce an additional confounding factor that, so far, has not been accounted for. I will therefore discuss a variety of confounding factors that influence the quantification of transcriptional noise, grouping these into experimental and technical effects.

5.2.1 Experimental confounding factors

Transcriptional noise as defined in **Box 1** can only be measured in truly homogeneous populations of cells. Previous studies that quantified transcriptional variability from scRNA-Seq data either sequenced mESCs (e.g. [12]), primary chicken erythroid progenitor cells [15], a murine multipotent hematopoietic precursor cell line [52] or $CD4^+$ T cells [22], all of which reside in a homogeneous ground state prior to activation/differentiation. With the development of technologies that capture thousands of cells in an unbiased way, structured heterogeneity presents the major source of cell-to-cell variation in expression. As shown in **Section 3.6**, one relies on clustering approaches to identify homogeneous populations of cells that can be compared when testing for changes in transcriptional variability. It is therefore also crucial to understand the underlying biology that causes structured heterogeneity to avoid including low quality cells in the analysis. For example, Ibarra-Soria *et al.*, 2018 identified a small intermediate population between pre-somitic and somitic mesoderm with unknown identity [27]. It is recommended to remove such cells from analysis to avoid any unknown biological heterogeneity that confounds biological noise.

As shown in **Section 4.9**, quantification of transcriptional noise is also heavily influenced by the underlying differentiation programmes of otherwise homogeneous cell populations, as exemplified by the differentiation process of spermatogenesis. After extensive quality control and clustering, the remaining variation in germ cell populations is dictated by genes that strongly and abruptly change their expression levels (e.g. *Prm1*). This observation is in line with previous reports on how the cell cycle state of each cell masks underlying population structure [13]. For each gene i, the scLVM captures (e.g.) the cell cycle associated component $\hat{y}_i$ and allows the derivation of corrected counts y^* by substracting this effect from the observed count y_i: $y^* = y_i - \hat{y}_i$. This correction can therefore be seen as a regression approach to correct for a specific confounding effect (e.g. cell cycle). To incorporate this idea into the BASiCS framework, it is possible to introduce a flexible regression that accounts for any given confounding effect. In addition to correcting the mean expression effect, the model can be extended to perform a semi-parametric regression between the over-dispersion parameter and a measure of association to differentiation. This measure in the simplest case can be parameters of a regression fit between each cells' expression level and the ordering of cells along the differentiation time course.

5.2.2 Technical confounding factors

ScRNA-Seq is prone to high technical noise due to the low starting amounts of RNA transcripts that are first captured, reverse transcribed, pre-amplified, prepared for sequencing and sequenced. Only around 10%-20% of all transcripts are captured in each individual cell leading to high levels of technical noise. Furthermore, amplification biases exponentially enhance noise introduced by variation in capture efficiency. These biases are minimised by the introduction of UMIs that allow the direct quantification of transcript abundance [174]. In preliminary analyses to study parameter robustness as displayed in **Section 3.4.2**, we observed that UMI data [259] resulted in generally more robust estimates compared to non-UMI data (e.g. $CD4^+$T cells, [22]).

The incorporation of UMIs into droplet-based scRNA-Seq technologies facilitates a robust estimation of transcriptional variability. On the other hand, these high-throughput methods come at the price of reduced sequencing depth, the inability to quantify technical noise via RNA spike-ins and often reduced replication. A recent study addressed the question of how to allocate a given sequencing budget to scRNA-Seq experiments [499]. One can either choose to sequence more cells at lower depth or to deeply sequence few cells. More reads in fewer cells reduce technical noise when estimating the cellular transcription state while

more cells capture the full variance observed in the cell population. The authors propose that the optimal trade-off between number off cells and sequencing depth considering a fixed sequencing budget is an average ∼1 UMI per cell detected for the biologically relevant genes [499]. This trade-off was found by simulations and sub-sampling experiment similar to the ones displayed in **Chapter 3**. Further to the results of Zhang *et al.*, 2018, replications of droplet-based scRNA-Seq experiments are important to robustly quantify and validate measures of transcriptional variability.

5.3 Experimental validation and manipulation of noise

While the results throughout this thesis indicated the functional role of transcriptional variability in dynamic biological systems, one of the main experimental challenges is to alter transcriptional noise to validate the hypothesised role. Classically, unicellular systems were employed to study the sources of transcriptional noise. In these systems, genetic alterations allowed the modulation of transcriptional and translational variability [2, 35, 136, 109]. Specifically, changing promoter architecture strongly alters expression noise [215, 110]. These simple approaches are not feasible in multicellular organisms, for example, to alter transcriptional noise during embryogenesis. While several regulatory factors on the genomic, epigenetic, transcriptional and translational level influence transcriptional noise, it is difficult to introduce a targeted alteration of certain regulatory factors while simultaneously avoiding down-stream effects other than alterations of transcriptional noise of one or few genes. Dueck *et al.*, 2016 proposed *in vitro* experimental designs to perturb expression variability in cellular systems. Generally, these approaches can be grouped into targeted and general perturbations of transcriptional noise [500].

5.3.1 General perturbation of transcriptional noise

Dueck *et al.*, 2016 introduced the concept of increasing global transcriptional noise by utilising the off-target effects of small interfering RNAs (siRNAs). Similar to miRNAs, siRNAs are designed to complementarly bind target RNAs and induce their degradation. While most siRNAs lead to the cleavage of cognate RNA, due to partial complementary sequences in off-target RNAs, levels of off-target proteins are also perturbed [501]. By designing a system where siRNAs with primarily off-target effects are expressed under a regulated promoter such as the tetracycline-controlled transcriptional activation system [502], global changes in transcriptional variability can be induced [500]. In a similar fashion, the

controlled expression of a CRISPR/Cas9 system containing sgRNAs with random targets can introduce random deletions or insertions genome wide and therefore increase transcriptional noise. These settings do not control for changes in cellular states due to spontaneous up-or down-regulation of key regulatory components.

5.3.2 Targeted perturbation of transcriptional noise

To alter the variation in transcript abundance for specific RNAs, Dueck *et al.* proposed to (i) transfect a selected set of RNAs into specific cells, (ii) transfect the RNAs encoding specific TFs into cells or to (iii) over-express certain miRNAs that target multiple RNAs [500]. The first two approaches only increase RNA abundance for specific genes in specific cells. The third approach offers a more intriguing method to modulate protein abundance at the post-transcriptional level as demonstrated by Schmiedel *et al.*, 2015 and 2017 [135, 503]. The proposed role of miRNAs to reduce noise levels in protein abundance offers an experimental setting where depletion of certain miRNAs by targeted CRISPR/Cas9 interference could potentially increase noise in a set of miRNA targeted genes. This system could, for example, validate an ageing phenotype in the activation response of $CD4^+$ T cells as presented in **Chapter 2**.

The identification of miRNA-driven modulation of transcriptional noise further opens the question whether transcriptional noise can be modulated by other factors that affect mRNA stability, possibly by altering one out of more than 100 described RNA modifications [504]. For example, miRNAs regulate the N^6-methylation of adenosine (m6A) on RNAs by recruiting the methyltransferase METTL3 which affects the reprogramming of mouse embryonic fibroblasts [505]. Inducing alterations in the machinery that deposits or recognises such modifications of the RNA could lead to targeted increase or decrease in transcriptional noise.

5.4 Future approaches to model scRNA-Seq data

I introduced BASiCS as a Bayesian framework to quantify transcriptional noise from scRNA-Seq data with the main benefit of propagating statistical uncertainty from the data to down-stream differential variability testing. With the development of droplet-based scRNA-Seq approaches [164, 163] and large-scale microwell techniques [225], the amount of cells that can be assayed in one experiment scaled from hundreds to hundreds of thousands [506]. To learn model parameters of a generative model such as BASiCS across all cells and all genes became computationally challenging when considering a full Bayesian MCMC-based approach. To address this problem, a model framework called single-cell variational inference (scVI) has been developed that uses stochastic optimisation within a variational autoencoder network to approximate posterior distributions of model parameters and latent factors [297]. In scVI transcriptomes of each cell are encoded through a non-linear transformation into a low-dimensional latent vector of normal random variables.

The latent representation is non-linearly transformed to generate a posterior distribution of model parameters based on a ZINB. For this, the transcript count $x_{n,g}$ of gene g in cell n is modelled as:

$$
\begin{aligned}
x_{n,g} &= \begin{cases} y_{n,g} & \text{if } h_{n,g} = 0, \\ 0 & \text{otherwise} \end{cases} \\
h_{n,g} &\sim \text{Bernoulli}(f_h^g(z_n, s_n)) \\
y_{n,g} &\sim \text{Poisson}(l_n w_{n.g}) \\
w_{n,g} &\sim \text{Gamma}(\rho_n^g, \theta) \\
\rho_n &= f_w(z_n, s_n) \\
l_n &\sim \text{log-Normal}(l_\mu, l_\sigma^2) \\
z_n &\sim \text{Normal}(0, I)
\end{aligned}
$$

In this model, the NB distribution is realised as a hierarchical formulation of $y_{n,g}$ being Poisson distributed around the latent random variable l_n with an additional random effect $w_{n,g}$. Additionally, the zero-inflation of the model is controlled by the latent variable $h_{n,g}$. l_n is a random variable that represents nuisance variation due to differences in capture efficiency and sequencing depth and correlates with log-library size. l_n is log-normal distributed parametrised by $l_\mu, l_\sigma \in \mathbb{R}_+^B$ which are empirical mean and variance estimates of the log-library size per batch in B and which are therefore constants in the model **(Fig. 5.1A)**.

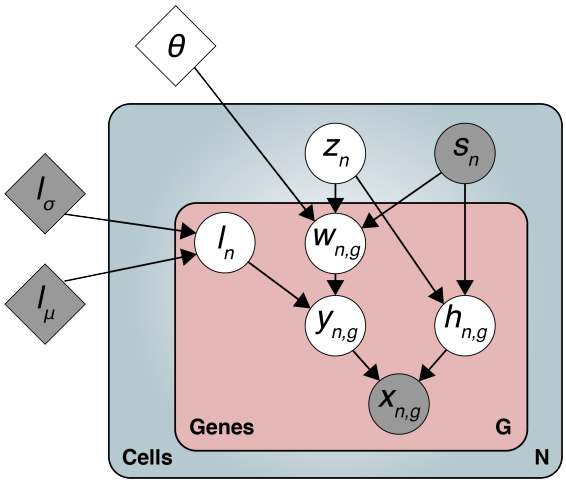


Fig. 5.1: The scVI model.
Hierarchical representation of the scVI model. Shaded nodes indicate observed quantities. White nodes indicated latent random variables. Shaded diamonds represent constants which were set *a priori*. White diamonds indicate variables shared across all genes and all cells. Edges show conditional dependency. Adapted from [297].

$w_{n,g}$ is Gamma distributed with the shape parameter ρ_n^g and the scale parameter θ. ρ_g represents an intermediate matrix that relates the observations $x_{n,g}$ to the latent variables z_n. It provides a batch-corrected, normalised estimate of the percentage of transcripts in each cell n from each gene g. θ is a global inverse-dispersion variable shared across all genes and all cells. The latent variable z_n captures a latent representation of the data reflecting biological variation between the cells. f_w and f_h are neural networks mapping the latent space and batch annotation back to the full dimension of all genes: $\mathbb{R}^d \times \{0,1\}^B \to \mathbb{R}^G$.

Fast inference of this model is implemented via stochastic optimisation. First, the latent variables $w_{n,g}$, $h_{n,g}$ and $y_{n,g}$ are integrated out by controlling that $p(x_{n,g}|z_n, l_n, s_n)$ has a closed form density and is ZINB (see Appendix A in [297]). In this formulation, the distribution of $x_{n,g}$ is only conditioned on the latent variables z_n and l_n. The posterior distributions has therefore the following form: $p(z_n, l_n|x_{n,g}, s_n)$. Mean-field variational inference is used to parametrised the posterior as:

$$p(z_n, l_n|x_{n,g}, s_n) = p(z_n|x_{n,g}, s_n)p(l_n|x_{n,g}, s_n) \tag{5.1}$$

The variational distribution $q(z_n|\cdot)$ is chosen to be Gaussian with diagonal covariance matrix and mean and covariance are learned by a multilayer perceptron (MLP) network similar to Kingma *et al.*, 2013 [507]. Similarly, $q(l_n|\cdot)$ is chosen to be log-normal where the scalar mean and variance are learned by a MLP [507]. The authors used reparametrisation to solve the variational lower bound of this system [297, 507]. Furthermore, scVI uses stochastic optimisation by sampling 128 cells for optimising the objective function. This approach is therefore fast (5 hours for > 1 million cells and 750 genes and 10 hours for > 1 million cells and 10,000 genes) and memory efficient.

The authors concluded that: 1. scRNA-Seq data is better fitted with a ZINB than log-Normal or zero-inflated log-Normal; 2. Zero-inflation is not needed as part of the model since the zeros in dataset can be explained by NB distribution; 3. When the number of cells is smaller than number of genes, scVI underfits the data [297]. The clear strength of the model is the fast estimation of model parameters that can be used for down-stream analysis (e.g. visualisation, normalisation, differential expression testing). The draw-back of this model is the inability to obtain gene-specific variability estimates but rather global variability measures calculated on the latent space.

As a future direction, generative models need to allow fast inference while providing interpretable model parameters that capture gene-specific measures of transcriptional noise. These measures should ideally be independent of technical noise, mean expression and possibly flexible enough to adjust for further confounding factors such as expression changes over a differentiation time course.

References

[1] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[2] Jonathan M. Raser and Erin K. O'Shea. Control of Stochasticity in Eukaryotic Gene Expression. *Science*, 304(5678):1811–1814, 2004.

[3] Alvaro Sanchez and Ido Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342:1188–1193, 2013.

[4] Cristopher J. Zopf, Katie Quinn, Joshua Zeidman, and Narendra Maheshri. Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. *PLoS Computational Biology*, 9(7):1–12, 2013.

[5] Kazunari Iwamoto, Yuki Shindo, and Koichi Takahashi. Modeling Cellular Noise Underlying Heterogeneous Cell Responses in the Epidermal Growth Factor Signaling Pathway. *PLoS Computational Biology*, 12(11):1–18, 2016.

[6] Daniel J. Kiviet, Philippe Nghe, Noreen Walker, Sarah Boulineau, Vanda Sunderlikova, and Sander J. Tans. Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522):376–379, 2014.

[7] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 1363:1–21, 2015.

[8] Christoph Bock, Matthias Farlik, and Nathan C. Sheffield. Multi-Omics of Single Cells: Strategies and Applications. *Trends in Biotechnology*, 34(8):605–608, 2016.

[9] Iain C. Macaulay, Chris P. Ponting, and Thierry Voet. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics*, 33(2):155–168, 2017.

[10] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah a Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, 2013.

[11] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis of single-cell sequencing data. *PLOS Computational Biology*, 11(6):e1004333, 2015.

[12] Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Jason C.H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C. Marioni, and Sarah A. Teichmann. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17:471–485, 2015.

[13] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah a Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

[14] Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E. Kaeser, Yun C. Yung, Joseph L. Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, and Peter V. Kharchenko. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, 13(3):241–244, 2016.

[15] Angelique Richard, Lois Boullu, Ulysse Herbach, Arnaud Bonnafoux, Valerie Morin, Elodie Vallin, Anissa Guillemin, Nan Papili Gao, Rudiyanto Gunawan, Jeremie Cosette, Ophelie Arnaud, Jean Jacques Kupiec, Thibault Espinasse, Sandrine Gonin-Giraud, and Olivier Gandrillon. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biology*, 14(12):1–35, 2016.

[16] Mubeen Goolam, Antonio Scialdone, Sarah J L Graham, Iain C. MacAulay, Agnieszka Jedrusik, Anna Hupalowska, Thierry Voet, John C. Marioni, and Magdalena Zernicka-Goetz. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell*, 165(1):61–74, 2016.

[17] Hisham Mohammed, Irene Hernando-Herraez, Aurora Savino, Jennifer Nichols, John C Marioni, Wolf Reik, Antonio Scialdone, Iain Macaulay, Carla Mulas, Tamir Chandra, Thierry Voet, and Wendy Dean. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Reports*, 20(5):1215–1228, 2017.

[18] Yusuke Ohnishi, Wolfgang Huber, Akiko Tsumura, Minjung Kang, Panagiotis Xenopoulos, Kazuki Kurimoto, Andrzej K Oleś, Marcos J Araúzo-Bravo, Mitinori Saitou, Anna-Katerina Hadjantonakis, and Takashi Hiiragi. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nature Cell Biology*, 16(1):27–37, 2014.

[19] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P May, and Aviv Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):263–269, 2014.

[20] Ryan A. Kellogg and Sava Tay. Noise facilitates transcriptional control under dynamic inputs. *Cell*, 160(3):381–392, 2015.

[21] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: A looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012.

[22] Celia P. Martinez-Jimenez, Nils Eling, Hung-Chang Chen, Catalina A Vallejos, Aleksandra A Kolodziejczyk, Frances Connor, Lovorka Stojic, Timothy F Rayner, Michael J T Stubbington, Sarah A Teichmann, Maike de la Roche, John C Marioni, and Duncan T Odom. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 1436:1433–1436, 2017.

[23] Arjun Raj, Scott A. Rifkin, Erik Andersen, and Alexander van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–918, 2010.

[24] Gábor Balázsi, Alexander Van Oudenaarden, and James J. Collins. Cellular decision making and biological noise: From microbes to mammals. *Cell*, 144(6):910–925, 2011.

[25] Avigdor Eldar and Michael B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.

[26] Franziska Paul, Ya'Ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, 163(7):1663–1677, 2015.

[27] Ximena Ibarra-Soria, Wajid Jawaid, Blanca Pijuan-Sala, Vasileios Ladopoulos, Antonio Scialdone, David J. Jörg, Richard C.V. Tyser, Fernando J. Calero-Nieto, Carla Mulas, Jennifer Nichols, Ludovic Vallier, Shankar Srinivas, Benjamin D. Simons, Berthold Göttgens, and John C. Marioni. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nature Cell Biology*, 20(2):127–134, 2018.

[28] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360:1–7, 2018.

[29] Joakim S. Dahlin, Fiona K. Hamey, Blanca Pijuan-Sala, Mairi Shepherd, Winnie W.Y. Lau, Sonia Nestorowa, Caleb Weinreb, Samuel Wolock, Rebecca Hannah, Evangelia Diamanti, David G. Kent, Berthold Göttgens, and Nicola K. Wilson. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood*, 131(21):1–11, 2018.

[30] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–6, 2014.

[31] Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J Theis, Carsten Marr, and Florian Buettner. Destiny - diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–3, 2015.

[32] Andre J. Faure, Jörn M. Schmiedel, and Ben Lehner. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. *Cell Systems*, 5(5):471–484, 2017.

[33] Michael D. Morgan and John C. Marioni. CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. *Genome Biology*, 19(1):1–13, 2018.

[34] Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[35] Jonathan M Raser and Erin K. O'Shea. Noise in Gene Expression: Origins, Consequences, and Control. *Science*, 309:2010–2014, 2005.

[36] Jacob Stewart-Ornstein, Jonathan S. Weissman, and Hana El-Samad. Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae. *Molecular Cell*, 45(4):483–493, 2012.

[37] Margaret Lieb. The establishment of lysogenicity in Escherichia coli. *Journal of bacteriology*, 65(6):642–651, 1953.

[38] Adam Arkin, John Ross, and Harley H McAdams. Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage lambda-Infected Escherichia coli Cells. *Genetics*, 149:1633–1648, 1998.

[39] Francois St-Pierre and Drew Endy. Determination of cell fate selection during phage. *Proceedings of the National Academy of Sciences*, 105(52):20705–20710, 2008.

[40] Lanying Zeng, Samuel O Skinner, Chenghang Zong, Jean Sippy, Michael Feiss, and Ido Golding. Decision Making at a Subcellular Level Determines the Outcome of Bacteriophage Infection. *Cell*, 141(4):682–691, 2010.

[41] Leor S. Weinberger. A minimal fate-selection switch. *Current Opinion in Cell Biology*, 37:111–118, 2015.

[42] Roy D. Dar, Nina N. Hosmane, Michelle R. Arkin, Robert F. Siliciano, and Leor S. Weinberger. Screening for noise in gene expression identifies drug synergies. *Science*, 344(6190):1392–1396, 2014.

[43] Daniel Schultz, Peter G. Wolynes, Eshel Ben Jacob, and José N. Onuchic. Deciding fate in adverse times: Sporulation and competence in Bacillus subtilis. *Proceedings of the National Academy of Sciences*, 106(50):21027–21034, 2009.

[44] Gürol M. Süel, Jordi Garcia-Ojalvo, Louisa M. Liberman, and Michael B. Elowitz. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440(7083):545–550, 2006.

[45] Jonathan R Russell, Matthew T Cabeen, Paul A Wiggins, Johan Paulsson, and Richard Losick. Noise in a phosphorelay drives stochastic entry into sporulation in Bacillus subtilis. *The EMBO Journal*, 36(19):e201796988, 2017.

[46] Nathalie Q Balaban, Jack Merrin, Remy Chait, Lukasz Kowalik, and Stanislas Leibler. Bacterial Persistence as a Phenotypic Switch. *Science*, 305:1622–1626, 2004.

[47] Saurabh Paliwal, Pablo A. Iglesias, Kyle Campbell, Zoe Hilioti, Alex Groisman, and Andre Levchenko. MAPK-mediated bimodal gene expression and adaptive gradient sensing in yeast. *Nature*, 446(7131):46–51, 2007.

[48] Murat Acar, Jerome T. Mettetal, and Alexander Van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, 2008.

[49] Hannah H. Chang, Martin Hemberg, Mauricio Barahona, Donald E. Ingber, and Sui Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–547, 2008.

[50] Patrick S. Stumpf, Rosanna C.G. Smith, Michael Lenz, Andreas Schuppert, Franz Josef Müller, Ann Babtie, Thalia E. Chan, Michael P.H. Stumpf, Colin P. Please, Sam D. Howison, Fumio Arai, and Ben D. MacArthur. Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Systems*, 5(3):268–282, 2017.

[51] Vlatka Antolović, Agnes Miermont, Adam M. Corrigan, and Jonathan R. Chubb. Generation of Single-Cell Transcript Variability by Repression. *Current Biology*, 27:1811–1817, 2017.

[52] Mitra Mojtahedi, Alexander Skupin, Joseph Zhou, Ivan G. Castano, Rebecca Y Y Leong-Quong, Hannah Chang, Kalliopi Trachana, Alessandro Giuliani, and Sui Huang. Cell fate decision as high-dimensional critical state transition. *PLoS Biology*, 14(12):1–28, 2016.

[53] Jens-Erik Dietrich and Takashi Hiiragi. Stochastic patterning in the mouse preimplantation embryo. *Development*, 134(23):4219–4231, 2007.

[54] Hui Ting Zhang and Takashi Hiiragi. Symmetry Breaking in the Mammalian Embryo. *Annual Review of Cell and Developmental Biology*, 34(1):405–426, 2018.

[55] Jean León Maître, Hervé Turlier, Rukshala Illukkumbura, Björn Eismann, Ritsuya Niwayama, François Nédélec, and Takashi Hiiragi. Asymmetric division of contractile domains couples cell positioning and fate specification. *Nature*, 536(7616):344–348, 2016.

[56] Katsuhiko Hayashi, Susana M Chuva, De Sousa Lopes, Fuchou Tang, and Kaiqin Lao. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell*, 3(4):391–401, 2008.

[57] Vijay Chickarmane, Victor Olariu, and Carsten Peterson. Probing the role of stochasticity in a model of the embryonic stem cell – heterogeneous gene expression and reprogramming efficiency. *BMC Systems Biology*, 6(1):98, 2012.

[58] Maria-Elena Torres-Padilla and Ian Chambers. Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. *Development*, 141(11):2173–2181, 2014.

[59] Zakary S. Singer, John Yong, Julia Tischler, Jamie A. Hackett, Alphan Altinok, M. Azim Surani, Long Cai, and Michael B. Elowitz. Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. *Molecular Cell*, 55(2):319–331, 2014.

[60] Jacob Hanna, Krishanu Saha, Bernardo Pando, Jeroen van Zon, Christopher J. Lengner, Menno P. Creyghton, Alexander van Oudenaarden, and Rudolf Jaenisch. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462(7273):595–601, 2009.

[61] Shinya Yamanaka. Elite and stochastic models for induced pluripotent stem cell generation. *Nature*, 460(7251):49–52, 2009.

[62] Yosef Buganim, Dina A. Faddah, Albert W. Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L. Klemm, Alexander Van Oudenaarden, and Rudolf Jaenisch. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222, 2012.

[63] Philipp S. Hoppe, Michael Schwarzfischer, Dirk Loeffler, Konstantinos D. Kokkaliaris, Oliver Hilsenbeck, Nadine Moritz, Max Endele, Adam Filipczyk, Adriana Gambardella, Nouraiz Ahmed, Martin Etzrodt, Daniel L. Coutu, Michael A. Rieger, Carsten Marr, Michael K. Strasser, Bernhard Schauberger, Ingo Burtscher, Olga Ermakova, Antje Bürger, Heiko Lickert, Claus Nerlov, Fabian J. Theis, and Timm Schroeder. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature*, 535(7611):299–302, 2016.

[64] Edward C Schrom and Andrea L Graham. Instructed subsets or agile swarms: how T-helper cells may adaptively counter uncertainty with variability and plasticity. *Current Opinion in Genetics & Development*, 47:75–82, 2017.

[65] Miaoqing Fang, Huangming Xie, Stephanie K. Dougan, Hidde Ploegh, and Alexander van Oudenaarden. Stochastic Cytokine Expression Induces Mixed T Helper Cell States. *PLoS Biology*, 11(7), 2013.

[66] Yaron E. Antebi, Shlomit Reich-Zeliger, Yuval Hart, Avi Mayo, Inbal Eizenberg, Jacob Rimer, Prabhakar Putheti, Dana Pe'er, and Nir Friedman. Mapping Differentiation under Mixed Culture Conditions Reveals a Tunable Continuum of T Cell Fates. *PLoS Biology*, 11(7):e1001616, 2013.

[67] Franziska Fuhrmann, Timo Lischke, Fridolin Gross, Tobias Scheel, Laura Bauer, Khalid Wasim Kalim, Andreas Radbruch, Hanspeter Herzel, Andreas Hutloff, and Ria Baumgrass. Adequate immune response ensured by binary IL-2 and graded CD25 expression in a murine transfer model. *eLife*, 5:1–17, 2016.

[68] Ryan A Kellogg, Chengzhe Tian, Tomasz Lipniacki, and Stephen R Quake. Digital signaling decouples activation probability and population heterogeneity. *eLife*, 4(e08931):1–26, 2015.

[69] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7):1596–1610, 2015.

[70] Christine Queitsch, Todd A. Sangster, and Susan Lindquist. Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889):618–624, 2002.

[71] Ni Ji, Teije C. Middelkoop, Remco A. Mentink, Marco C. Betist, Satto Tonegawa, Dylan Mooijman, Hendrik C. Korswagen, and Alexander Van Oudenaarden. Feedback control of gene expression variability in the caenorhabditis elegans wnt pathway. *Cell*, 155(4):869–880, 2013.

[72] Lei Zhang, Kelly Radtke, Likun Zheng, Anna Q. Cai, Thomas F. Schilling, and Qing Nie. Noise drives sharpening of gene expression boundaries in the zebrafish hindbrain. *Molecular Systems Biology*, 8(613):1–12, 2012.

[73] Qixuan Wang, William R. Holmes, Julian Sosnik, Thomas Schilling, and Qing Nie. Cell Sorting and Noise-Induced Cell Plasticity Coordinate to Sharpen Boundaries between Gene Expression Domains. *PLoS Computational Biology*, 13(1):1–23, 2017.

[74] Brian A. Camley and Wouter-Jan Rappel. Cell-to-cell variation sets a tissue-rheology-dependent bound on collective gradient sensing. *Proceedings of the National Academy of Sciences*, 114(47):E10074–E10082, 2017.

[75] Beáta Tóth, Shani Ben-Moshe, Avishai Gavish, Naama Barkai, and Shalev Itzkovitz. Early commitment and robust differentiation in colonic crypts. *Molecular Systems Biology*, 13(1):902, 2017.

[76] Jonas Ranft, Markus Basan, Jens Elgeti, Jean-Francois Joanny, Jacques Prost, and Frank Julicher. Fluidization of tissues by cell division and apoptosis. *Proceedings of the National Academy of Sciences*, 107(49):20863–20868, 2010.

[77] Robert Ahrends, Asuka Ota, Kyle M. Kovary, Takamasa Kudo, Byung O. Park, and Mary N. Teruel. Controlling low rates of cell differentiation through noise and ultrahigh feedback. *Science*, 344(6190):1384–1389, 2014.

[78] Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. *Molecular Cell*, 58(1):147–156, 2015.

[79] Karen Featherstone, Kirsty Hey, Hiroshi Momiji, Anne V. McNamara, Amanda L. Patist, Joanna Woodburn, David G. Spiller, Helen C. Christian, Alan S. McNeilly, John J. Mullins, Barbel F. Finkenstadt, David A. Rand, Michael R.H. White, and Julian R.E. Davis. Spatially coordinated dynamic gene transcription in living pituitary tissue. *eLife*, 5:1–25, 2016.

[80] Luis López-Maury, Samuel Marguerat, and Jürg Bähler. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 10(1):68–68, 2009.

[81] Itay Tirosh, Adina Weinberger, Miri Carmi, and Naama Barkai. A genetic signature of interspecies variations in gene expression. *Nature Genetics*, 38(7):830–834, 2006.

[82] J. Chris Pires and Gavin C. Conant. Robust Yet Fragile: Expression Noise, Protein Misfolding, and Gene Dosage in the Evolution of Genomes. *Annual Review of Genetics*, 50(1):113–131, 2016.

[83] Ben Lehner. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular Systems Biology*, 4(170), 2008.

[84] Nizar N Batada and Laurence D Hurst. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics*, 39(8):945–949, 2007.

[85] Ben Lehner. Conflict between noise and plasticity in yeast. *PLoS Genetics*, 6(11), 2010.

[86] Luise Wolf, Olin K. Silander, and Erik van Nimwegen. Expression noise facilitates the evolution of gene regulation. *eLife*, 4:1–48, 2015.

[87] Zoltán Bódi, Zoltán Farkas, Dmitry Nevozhay, Dorottya Kalapis, Viktória Lázár, Bálint Csörgő, Ákos Nyerges, Béla Szamecz, Gergely Fekete, Balázs Papp, Hugo Araújo, José L. Oliveira, Gabriela Moura, Manuel A.S. Santos, Tamás Székely, Gábor Balázsi, and Csaba Pál. Phenotypic heterogeneity promotes adaptive evolution. *PLoS Biology*, 15(5):1–26, 2017.

[88] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods*, 10(11):1127–1133, 2013.

[89] Thomas Stoeger, Nico Battich, and Lucas Pelkmans. Passive Noise Filtering by Cellular Compartmentalization. *Cell*, 164(6):1151–1161, 2016.

[90] Jörn M Schmiedel, Lucas B Carey, and Ben Lehner. Empirical noise-mean fitness landscapes and the evolution of gene expression. *bioRxiv*, pages 1–45, 2018.

[91] Fabien Duveau, Andrea Hodgins-Davis, Brian P.H. Metzger, Bing Yang, Stephen Tryban, Elizabeth A. Walker, Tricia Lybrook, and Patricia J Wittkopp. Fitness effects of altering gene expression noise in Saccharomyces cerevisiae. *eLife*, 7:e37272, 2018.

[92] Dongya Jia, Mohit Kumar Jolly, Prakash Kulkarni, and Herbert Levine. Phenotypic plasticity and cell fate decisions in cancer: Insights from dynamical systems theory. *Cancers*, 9(7):1–19, 2017.

[93] Winston Timp and Andrew P. Feinberg. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nature Reviews Cancer*, 13(7):497–510, 2013.

[94] Dan A. Landau, Kendell Clement, Michael J. Ziller, Patrick Boyle, Jean Fan, Hongcang Gu, Kristen Stevenson, Carrie Sougnez, Lili Wang, Shuqiang Li, Dylan Kotliar, Wandi Zhang, Mahmoud Ghandi, Levi Garraway, Stacey M. Fernandes, Kenneth J. Livak, Stacey Gabriel, Andreas Gnirke, Eric S. Lander, Jennifer R. Brown, Donna Neuberg, Peter V. Kharchenko, Nir Hacohen, Gad Getz, Alexander Meissner, and Catherine J. Wu. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*, 26(6):813–825, 2014.

[95] Simone Ecker, Vera Pancaldi, Daniel Rico, and Alfonso Valencia. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. *Genome Medicine*, 7(1):1–12, 2015.

[96] Deborah A. Flusberg and Peter K. Sorger. Surviving apoptosis: Life-death signaling in single cells. *Trends in Cell Biology*, 25(8):446–458, 2015.

[97] Sabrina L. Spencer, Suzanne Gaudet, John G. Albeck, John M. Burke, and Peter K. Sorger. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245):428–432, 2009.

[98] Andrew L. Paek, Julia C. Liu, Alexander Loewer, William C. Forrester, and Galit Lahav. Cell-to-Cell Variation in p53 Dynamics Leads to Fractional Killing. *Cell*, 165(3):631–642, 2016.

[99] Sydney M. Shaffer, Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A. Brafford, Min Xiao, Elliott Eggan, Ioannis N. Anastopoulos, Cesar A. Vargas-Garcia, Abhyudai Singh, Katherine L. Nathanson, Meenhard Herlyn, and Arjun Raj. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435, 2017.

[100] Jérémie Roux, Marc Hafner, Samuel Bandara, Joshua J Sims, Hannah Hudson, Diana Chai, and Peter K Sorger. Fractional killing arises from cell-to-cell variability in overcoming a caspase activity threshold. *Molecular Systems Biology*, 11(5):803, 2015.

[101] Douglas Hanahan and Robert a Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144:646–674, mar 2011.

[102] Rumana Bahar, Claudia H. Hartmann, Karl A. Rodriguez, Ashley D. Denny, Rita A. Busuttil, Martijn E. T. Dollé, R. Brent Calder, Gary B. Chisholm, Brad H. Pollock, Christoph A. Klein, and Jan Vijg. Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, 441(7096):1011–1014, 2006.

[103] Luigi A. Warren, Derrick J. Rossi, Geoffrey R. Schiebinger, Irving L. Weissman, Stuart K. Kim, and Stephen R. Quake. Transcriptional instability is not a universal attribute of aging. *Aging Cell*, 6(6):775–782, 2007.

[104] Martin Enge, H. Efsun Arda, Marco Mignardi, John Beausang, Rita Bottino, Seung K Kim, and Stephen R Quake. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell*, 171:1–10, 2017.

[105] D. S. Latchman. Transcription factors: An overview. *International Journal of Biochemistry and Cell Biology*, 29(12):1305–1312, 1997.

[106] Brian P. H. Metzger, David C. Yuan, Jonathan D. Gruber, Fabien Duveau, and Patricia J. Wittkopp. Selection on noise constrains variation in a eukaryotic promoter. *Nature*, 521(7552):344–347, 2015.

[107] Benjamin Zoller, Damien Nicolas, Nacho Molina, and Felix Naef. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular Systems Biology*, 11(7):823–823, 2015.

[108] Christian R. Landry, Bernardo Lemos, Scott A. Rifkin, W. J. Dickinson, and Daniel L. Hartl. Genetic Properties Influencing the Evolvability of Gene Expression. *Science*, 317:118–122, 2007.

[109] Gil Hornung, Raz Bar-ziv, Dalia Rosin, Nobuhiko Tokuriki, Dan S Tawfik, Moshe Oren, and Naama Barkai. Noise-mean relationship in mutated promoters. *Genome Research*, 22:2409–2417, 2012.

[110] Eilon Sharon, David Van Dijk, Yael Kalma, Leeat Keren, Ohad Manor, Zohar Yakhini, and Eran Segal. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Research*, 24(10):1698–1706, 2014.

[111] Jung Kyoon Choi and Young-Joon Kim. Epigenetic regulation and the variability of gene expression. *Nature Genetics*, 40(2):141–7, 2008.

[112] Ignacio E Schor, Jacob F Degner, Dermot Harnett, Enrico Cannavò, Francesco P Casale, Heejung Shim, David A Garfield, Ewan Birney, Matthew Stephens, Oliver Stegle, and Eileen E M Furlong. Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Genetics*, 49(4):550–558, 2017.

[113] Eric R. Gamazon and Barbara E. Stranger. The impact of human copy number variation on gene expression. *Briefings in Functional Genomics*, 14(5):352–357, 2015.

[114] Siddharth S Dey, Lennart Kester, Bastiaan Spanjaard, Magda Bienko, and Alexander van Oudenaarden. Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33:285–289, 2015.

[115] C. T. Wu and J. R. Morris. Genes, genetics, and epigenetics: A correspondence. *Science*, 293(5532):1103–1105, 2001.

[116] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, 2010.

[117] Tamaki Suganuma and Jerry L. Workman. Signals and Combinatorial Functions of Histone Modifications. *Annual Review of Biochemistry*, 80(1):473–499, 2011.

[118] Emily C. Chittock, Sebastian Latwiel, Thomas C.R. Miller, and Christoph W. Müller. Molecular architecture of polycomb repressive complexes. *Biochemical Society Transactions*, 45(1):193–205, 2017.

[119] Gozde Kar, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Kedar Nath Natarajan, Elena Torlai Triglia, Borbala Mifsud, Sarah Elderkin, John C. Marioni, Ana Pombo, and Sarah A. Teichmann. Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. *Nature Communications*, 8(1):36, 2017.

[120] Tony Kouzarides. Chromatin Modifications and Their Function. *Cell*, 128(4):693–705, 2007.

[121] Itay Tirosh and Naama Barkai. Two strategies for gene regulation by promoter nucleosomes. *Genome Research*, 18(7):1084–1091, 2008.

[122] Eliza C. Small, Liqun Xi, Ji-Ping Wang, Jonathan Widom, and Jonathan D. Licht. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences*, 111(24):E2462–E2471, 2014.

[123] Somi Kim, Nam Kyung Yu, and Bong Kiun Kaang. CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & molecular medicine*, 47(6):e166, 2015.

[124] Gang Ren, Wenfei Jin, Kairong Cui, Joseph Rodrigez, Gangqing Hu, Zhiying Zhang, Daniel R. Larson, and Keji Zhao. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell*, 67(6):1049–1058, 2017.

[125] Elizabeth M. Blackwood and James T. Kadonaga. Going the distance: A current view of enhancer action. *Science*, 281(5373):60–63, 1998.

[126] Georg Kustatscher, Piotr Grabowski, and Juri Rappsilber. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Molecular Systems Biology*, 13(927):1–14, 2017.

[127] Ángel Goñi-Moreno, Ilaria Benedetti, Juhyun Kim, and Víctor De Lorenzo. Deconvolution of Gene Expression Noise into Spatial Dynamics of Transcription Factor-Promoter Interplay. *ACS Synthetic Biology*, 6(7):1359–1369, 2017.

[128] Lucas B. Carey, David van Dijk, Peter M.A. Sloot, Jaap A. Kaandorp, and Eran Segal. Promoter Sequence Determines the Relationship between Expression Level and Noise. *PLoS Biology*, 11(4), 2013.

[129] Alistair N Boettiger and Michael Levine. Synchronous and Stochastic Drosophila Embryo. *Science*, 325(22):23–25, 2009.

[130] Daniel S. Day, Bing Zhang, Sean M. Stevens, Francesco Ferrari, Erica N. Larschan, Peter J. Park, and William T. Pu. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biology*, 17(1):1–17, 2016.

[131] Tina Glisovic, Jennifer L. Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14):1977–1986, 2008.

[132] Keren Bahar Halpern, Inbal Caspi, Doron Lemze, Maayan Levy, Shanie Landen, Eran Elinav, Igor Ulitsky, and Shalev Itzkovitz. Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports*, 13(12):2653–2662, 2015.

[133] Maike M.K. Hansen, Ravi V. Desai, Michael L. Simpson, and Leor S. Weinberger. Cytoplasmic Amplification of Transcriptional Noise Generates Substantial Cell-to-Cell Variability. *Cell Systems*, pages 1–14, 2018.

[134] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–40, 2014.

[135] Jörn M. Schmiedel, Sandy L. Klemm, Yannan Zheng, Apratim Sahay, Nils Blüthgen, Debora S. Marks, and Alexander van Oudenaarden. MicroRNA control of protein expression noise. *Science*, 348(6230):128–131, 2015.

[136] Ertugrul M. Ozbudak, Mukund Thattai, Iren Kurtser, Alan D. Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, 2002.

[137] Estelle Dacheux, Naglis Malys, Meng Xiang, Vinoy Ramachandran, Pedro Mendes, and John E.G. McCarthy. Translation initiation events on structured eukaryotic mRNAs generate gene expression noise. *Nucleic Acids Research*, 45(11):6981–6992, 2017.

[138] Alejandro Colman-Lerner, Andrew Gordon, Eduard Serra, Tina Chin, Orna Resnekov, Drew Endy, C. Gustavo Pesce, and Roger Brent. Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, 437(7059):699–706, 2005.

[139] John R. S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[140] Leeat Keren, David Van Dijk, Shira Weingarten-Gabbay, Dan Davidi, Ghil Jona, Adina Weinberger, Ron Milo, and Eran Segal. Noise in gene expression is coupled to growth rate. *Genome Research*, 25(12):1893–1902, 2015.

[141] Florian Buettner, Naruemon Pratanwanich, Davis J. McCarthy, John C. Marioni, and Oliver Stegle. f-scLVM: Scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, 18(1):1–13, 2017.

[142] Hermannus Kempe, Aanne Schwabe, Frederic Cremazy, Pernette J. Verschure, and Frank J. Bruggeman. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Molecular Biology of the Cell*, 26(4):797–804, 2015.

[143] Olivia Padovan-Merhar, Gautham P. Nair, Andrew G. Biaesch, Andreas Mayer, Steven Scarfone, Shawn W. Foley, Angela R. Wu, L. Stirling Churchman, Abhyudai Singh, and Arjun Raj. Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Molecular Cell*, 58(2):339–352, 2015.

[144] Jacob Zhurinsky, Klaus Leonhard, Stephen Watt, Samuel Marguerat, Jürg Bähler, and Paul Nurse. A coordinated global control over cellular transcription. *Current Biology*, 20(22):2010–2015, 2010.

[145] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571, 2017.

[146] Berend Snijder, Raphael Sacher, Pauli Rämö, Eva-Maria Damm, Prisca Liberali, and Lucas Pelkmans. Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263):520–523, 2009.

[147] Chenghang Zong, Sijia Lu, Alec R. Chapman, and X. Sunney Xie. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science*, 338:1622–1627, 2012.

[148] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[149] Gilad D Evrony, Eunjung Lee, Peter J Park, Christopher A Walsh, Gilad D Evrony, Eunjung Lee, Bhaven K Mehta, Yuval Benjamini, Robert M Johnson, Xuyu Cai, Lixing Yang, Psalm Haseley, Hillel S. Lehmann, Peter J. Park, and Christopher A. Walsh. Cell Lineage Analysis in Human Brain Using NeuroResource Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. *Neuron*, 85(1):49–59, 2015.

[150] Christoph A. Klein, Oleg Schmidt-Kittler, Julian A. Schardt, Klaus Pantel, Michael R. Speicher, and Gert Riethmueller. Comparative genomic hybridization , loss of heterozygosity , and DNA sequence analysis of single cells. *Proceedings of the National Academy of Sciences*, 96(April):4494–4499, 1999.

[151] Frank B Dean, Seiyu Hosono, Linhua Fang, Xiaohong Wu, A Fawad Faruqi, Patricia Bray-ward, Zhenyu Sun, Qiuling Zong, Yuefen Du, Jing Du, Mark Driscoll, Wanmin Song, Stephen F Kingsmore, Michael Egholm, and Roger S Lasken. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences*, 99(8):5261–5266, 2002.

[152] Kristin A Knouse, Jie Wu, and Angelika Amon. Assessment of megabase-scale somatic copy number variation using single cell sequencing. *Genome Research*, 26:376–384, 2016.

[153] Timour Baslan, Jude Kendall, Brian Ward, Hilary Cox, Anthony Leotta, Linda Rodgers, Michael Riggs, Sean D'Italia, Guoli Sun, Mao Yong, Kristy Miskimen, Hannah Gilmore, Michael Saborowski, Nevenka Dimitrova, Alexander Krasnitz, Lyndsay Harris, Michael Wigler, and James Hicks. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Research*, 125(5):714–724, 2015.

[154] Sarah A. Vitak, Kristof A. Torkenczy, Jimi L. Rosenkrantz, Andrew J. Fields, Lena Christiansen, Melissa H. Wong, Lucia Carbone, Frank J. Steemers, and Andrew Adey. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, 14(3):302–308, 2017.

[155] Luigi Warren, David Bryder, Irving L Weissman, and Stephen R Quake. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences*, 103(47):17807–17812, 2006.

[156] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–82, 2009.

[157] Rashel V. Grindberg, Michael J. McConnell Joyclyn L. Yee-Greenbaum, Mark Novotny, Georgina M. O'Shaughnessy Andy L. Lambert, Marcos J. Araúzo-Bravod, Jun Lee, Max Fishman, Gillian E. Robbins, Xiaoying Lin, Pratap Venepally, Jonathan H. Badger, David W. Galbraith, Fred H. Gage, and Roger S. Lasken. RNA-sequencing From Single Nuclei. *Proceedings of the National Academy of Sciences*, 110:19802–19807, 2014.

[158] Dan Frumkin, Adam Wasserstrom, Shalev Itzkovitz, Alon Harmelin, Gideon Rechavi, and Ehud Shapiro. Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. *BMC Biotechnology*, 8:1–16, 2008.

[159] Tetsutaro Hayashi, Norito Shibata, Ryo Okumura, Tomomi Kudome, Osamu Nishimura, Hiroshi Tarui, and Kiyokazu Agata. Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. *Development Growth and Differentiation*, 52(1):131–144, 2010.

[160] Piero Dalerba, Tomer Kalisky, Debashis Sahoo, Pradeep S Rajendran, Michael E Rothenberg, Anne a Leyrat, Sopheak Sim, Jennifer Okamoto, Darius M Johnston, Dalong Qian, Maider Zabala, Janet Bueno, Norma F Neff, Jianbin Wang, Andrew a Shelton, Brendan Visser, Shigeo Hisamori, Yohei Shimono, Marc van de Wetering, Hans Clevers, Michael F Clarke, and Stephen R Quake. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127, 2011.

[161] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[162] Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark a Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–5, 2014.

[163] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[164] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[165] Sanjay M. Prakadan, Alex K. Shalek, and David A. Weitz. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nature Reviews Genetics*, 18(6):345–361, 2017.

[166] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21:1160–1167, 2011.

[167] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory a Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, 2012.

[168] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–8, 2013.

[169] Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673, 2012.

[170] Tamar Hashimshony, Naftalie Senderovich, Gal Avital, Agnes Klochendler, Yaron de Leeuw, Leon Anavy, Dave Gennert, Shuqiang Li, Kenneth J. Livak, Orit Rozenblatt-Rosen, Yuval Dor, Aviv Regev, and Itai Yanai. CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):1–7, 2016.

[171] Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4):610–620, 2015.

[172] External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics*, 6:150, 2005.

[173] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, 2011.

[174] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–6, feb 2014.

[175] Fluidigm. Using C1 to Generate Single-Cell cDNA Libraries for mRNA Sequencing. Technical report, Fluidigm, 2015.

[176] Todd M. Gierahn, Marc H. Wadsworth, Travis K. Hughes, Bryan D. Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J. Christopher Love, and Alex K. Shalek. Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods*, 14(4):395–398, 2017.

[177] H. Christina Fan, Glenn K. Fu, and Stephen P. A. Fodor. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222):1–8, 2015.

[178] Grace X.Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:1–12, 2017.

[179] Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *Science*, 357:661–667, 2017.

[180] Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R. Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, David A. Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, 14(10):955–958, 2017.

[181] Stephen J. Clark, Heather J. Lee, Sébastien A. Smallwood, Gavin Kelsey, and Wolf Reik. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biology*, 17(1):72, 2016.

[182] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–20, 2014.

[183] Matthias Farlik, Nathan C Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, Johanna Klughammer, and Christoph Bock. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Reports*, 10(8):1386–97, 2015.

[184] Ryan M. Mulqueen, Dmitry Pokholok, Steve Norberg, Andrew J. Fields, Duanchen Sun, Kristof A. Torkenczy, Jay Shendure, Cole Trapnell, Brian J. O'Roak, Zheng Xia, Frank J. Steemers, and Andrew C. Adey. Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing. *bioRxiv*, 2017.

[185] Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23:2126–2135, 2013.

[186] Dylan Mooijman, Siddharth S Dey, Jean-Charles Boisset, Nicola Crosetto, and Alexander van Oudenaarden. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature Biotechnology*, 34:852–856, 2016.

[187] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–72, 2015.

[188] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–90, 2015.

[189] Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

[190] Wenfei Jin, Qingsong Tang, Mimi Wan, Kairong Cui, Yi Zhang, Gang Ren, Bing Ni, Jeffrey Sklar, Teresa M. Przytycka, Richard Childs, David Levens, and Keji Zhao. Genome-wide Detection of DNase I Hypersensitive Sites in Single Cells and FFPE Samples. *Nature*, 528(7580):142–146, 2015.

[191] Jop Kind, Ludo Pagie, Sandra S. De Vries, Leila Nahidiazar, Siddharth S. Dey, Magda Bienko, Ye Zhan, Bryan Lajoie, Carolyn A. De Graaf, Mario Amendola, Geoffrey Fudenberg, Maxim Imakaev, Leonid A. Mirny, Kees Jalink, Job Dekker, Alexander Van Oudenaarden, and Bas Van Steensel. Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell*, 163(1):134–147, 2015.

[192] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[193] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L. Gunderson, Frank J. Steemers, Christine M. Disteche, William S. Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, 2017.

[194] Iain C. Macaulay, Wilfried Haerty, Parveen Kumar, Yang I. Li, Tim Xiaoming Hu, Mabel J. Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M. Shirley, Miriam Smith, Niels Van Der Aa, Ruby Banerjee, Peter D. Ellis, Michael A. Quail, Harold P. Swerdlow, Magdalena Zernicka-Goetz, Frederick J. Livesey, Chris P. Ponting, and Thierry Voet. G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522, 2015.

[195] Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle, and Wolf Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–32, 2016.

[196] Sebastian Pott. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife*, 6:1–19, 2017.

[197] Stephen J Clark, Ricard Argelaguet, Chantriolnt-andreas Kapourani, M Thomas, Heather J Lee, Celia Alda-catalinas, Felix Krueger, and Guido Sanguinetti. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(781):1–9, 2018.

[198] Spyros Darmanis, Caroline Julie Gallant, Voichita Dana Marinescu, Mia Niklasson, Anna Segerman, Georgios Flamourakis, Simon Fredriksson, Erika Assarsson, Martin Lundberg, Sven Nelander, Bengt Westermark, and Ulf Landegren. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. *Cell Reports*, 14(2):380–389, 2016.

[199] Cem Albayrak, Christian A Jordi, Christoph Zechner, Jing Lin, Colette A Bichsel, Mustafa Khammash, and Savaş Tay. Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Molecular Cell*, 61(6):914–24, 2016.

[200] Andreas P Frei, Felice-Alessio Bava, Eli R Zunder, Elena W Y Hsieh, Shih-Yu Chen, Garry P Nolan, and Pier Federico Gherardini. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nature Methods*, 13(3):269–275, 2016.

[201] Marie D. Harton and Eric Batchelor. Determining the Limitations and Benefits of Noise in Gene Regulation and Signal Transduction through Single Cell, Microscopy-Based Analysis. *Journal of Molecular Biology*, 429(8):1143–1154, 2017.

[202] William J Blake, Mads KAErn, Charles R Cantor, and J J Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[203] Anna Lyubimova, Shalev Itzkovitz, Jan Philipp Junker, Zi Peng Fan, Xuebing Wu, and Alexander Van Oudenaarden. Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols*, 8(9):1743–1758, 2013.

[204] Bin Yang, Jennifer B. Treweek, Rajan P. Kulkarni, Benjamin E. Deverman, Chun Kan Chen, Eric Lubeck, Sheel Shah, Long Cai, and Viviana Gradinaru. Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell*, 158(4):945–958, 2014.

[205] Yuichi Taniguchi, Paul J Choi, Gene-wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329:533–539, 2010.

[206] Andrew M. Sydor, Kirk J. Czymmek, Elias M. Puchner, and Vito Mennella. Super-Resolution Microscopy: From Single Molecules to Supramolecular Assemblies. *Trends in Cell Biology*, 25(12):730–748, 2015.

[207] Eric Lubeck and Long Cai. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods*, 9(7):743–748, 2012.

[208] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron*, 92(2):342–357, 2016.

[209] Jeffrey R. Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P. Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, 113(39):11046–11051, 2016.

[210] Jeffrey R. Moffitt, Junjie Hao, Dhananjay Bambah-Mukku, Tian Lu, Catherine Dulac, and Xiaowei Zhuang. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proceedings of the National Academy of Sciences*, 113(50):14456–14461, 2016.

[211] Lev S. Tsimring. Noise in biology. *Reports on Progress in Physics*, 77:026601, 2014.

[212] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[213] Nir Friedman, Long Cai, and X. Sunney Xie. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Physical Review Letters*, 97(16):1–4, 2006.

[214] Vahid Shahrezaei and Peter S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 106(45):17256–17261, 2008.

[215] Daniel L. Jones, Robert C. Brewster, and Rob Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536, 2014.

[216] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, 2006.

[217] Audrey Qiuyan Fu and Lior Pachter. Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. *Statistical Applications in Genetics and Molecular Biology*, 15(6):447–471, 2016.

[218] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology*, 4(10):e309, 2006.

[219] Dipjyoti Das, Supravat Dey, Robert Brewster, and Sandeep Choubey. Effects of transcription factor titration on gene expression noise. *PLoS Computational Biology*, 13(2):e1005491., 2015.

[220] J. David Van Dyken. Propagation and control of gene expression noise with non-linear translation kinetics. *Journal of Theoretical Biology*, 430:185–194, 2017.

[221] Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14(1):1–12, 2013.

[222] Siddharth S Dey, Jonathan E Foley, Prajit Limsirichai, David V Schaffer, and Adam P Arkin. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Molecular Systems Biology*, 11(5):806–806, 2015.

[223] 10X Genomics. Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium Single Cell 3 ' Solution. Technical report, 10X Genomics, 2017.

[224] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife*, 6:e27041, 2017.

[225] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Zimin Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, Yibin Wang, Rui Yue, Tiefeng Li, He Huang, Stuart H. Orkin, Guo-Cheng Yuan, Ming Chen, and Guoji Guo. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17, 2018.

[226] The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562:367–372, 2018.

[227] Eric M. Kernfeld, Ryan M.J. Genga, Kashfia Neherin, Margaret E. Magaletta, Ping Xu, and René Maehr. A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. *Immunity*, 48:1–13, 2018.

[228] Ilias Angelidis, Lukas M Simon, Isis E Fernandez, Maximilian Strunz, Christoph H Mayr, Flavia R Greiffo, George Tsitsiridis, Elisabeth Graf, Tim M Strom, Oliver Eickelberg, Matthias Mann, Fabian J Theis, and Herbert B Schiller. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *bioRxiv*, 2018.

[229] Jonathan A Griffiths, Antonio Scialdone, and John C Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, 14(4):e8046, 2018.

[230] Vincent Piras, Masaru Tomita, and Kumar Selvarajoo. Transcriptome-wide variability in single embryonic development cells. *Scientific Reports*, 4(7137):1–9, 2014.

[231] Junchao Shi, Qi Chen, Xin Li, Xiudeng Zheng, Ying Zhang, Jie Qiao, Fuchou Tang, Yi Tao, Qi Zhou, and Enkui Duan. Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell RNA-seq. *Development*, 142(20):3468–3477, 2015.

[232] Fernando H. Biase, Xiaoyi Cao, and Sheng Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Research*, 24(11):1787–1796, 2014.

[233] Antonio Scialdone, Yosuke Tanaka, Wajid Jawaid, Victoria Moignard, Nicola K Wilson, Iain C Macaulay, John C Marioni, and Berthold Göttgens. Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535(7611):4–6, 2016.

[234] Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P. Zinzen. The Drosophila embryo at single-cell transcriptome resolution. *Science*, 358(6360):194–199, 2017.

[235] Daniel E. Wagner, Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, and Allon M. Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 4362:1–12, 2018.

[236] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337:816–822, 2012.

[237] Feng Zhang, Yan Wen, and Xiong Guo. CRISPR/Cas9 for genome editing: Progress, implications and challenges. *Human Molecular Genetics*, 23:40–46, 2014.

[238] Aaron McKenna, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298), 2016.

[239] Bushra Raj, Daniel E. Wagner, Aaron McKenna, Shristi Pandey, Allon M. Klein, Jay Shendure, James A. Gagnon, and Alexander F. Schier. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, 36(5):442–450, 2018.

[240] Kirsten L. Frieda, James M. Linton, Sahand Hormoz, Joonhyuk Choi, Ke Huan K. Chow, Zakary S. Singer, Mark W. Budde, Michael B. Elowitz, and Long Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, 2017.

[241] Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt, and John C Marioni. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, 33:503–509, 2015.

[242] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33:495–502, 2015.

[243] Jan Philipp Junker, Emily S. Noel, Victor Guryev, Kevin A. Peterson, Gopi Shah, Jan Huisken, Andrew P. McMahon, Eugene Berezikov, Jeroen Bakkers, and Alexander van Oudenaarden. Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell*, 159(3):662–675, 2014.

[244] Kaia Achim, Nils Eling, Hernando Martinez Vergara, Paola Yanina Bertucci, Jacob Musser, Pavel Vopalensky, Thibaut Brunet, Paul Collier, Vladimir Benes, John C Marioni, and Detlev Arendt. Whole-body single-cell sequencing reveals transcriptional

domains in the annelid larval body. *Molecular Biology and Evolution*, 35:1047–1062, 2018.

[245] Arnau Sebé-Pedrós, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie Pierre Mailhé, Justine Renno, Yann Loe-Mie, Aviezer Lifshitz, Zohar Mukamel, Sandrine Schmutz, Sophie Novault, Patrick R.H. Steinmetz, François Spitz, Amos Tanay, and Heather Marlow. Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell*, 173(6):1520–1534, 2018.

[246] Arnau Sebé-Pedrós, Elad Chomsky, Kevin Pang, David Lara-Astiaso, Federico Gaiti, Zohar Mukamel, Ido Amit, Andreas Hejnol, Bernard M. Degnan, and Amos Tanay. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nature Ecology and Evolution*, 2(7):1176–1188, 2018.

[247] Valentina Proserpio and Bidesh Mahata. Single-cell technologies to study the immune system. *Immunology*, 147:133–140, 2015.

[248] Rahul Satija and Alex K. Shalek. Heterogeneity in immune responses: From populations to single cells. *Trends in Immunology*, 35(5):219–229, 2014.

[249] Glenn Dranoff. Cytokines in cancer pathogenesis and cancer therapy. *Nature Reviews Cancer*, 4(1):11–22, 2004.

[250] Alexandra Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L. De Jager, Orit Rozenblatt-Rosen, Andrew A. Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), 2017.

[251] Åsa K Björklund, Marianne Forkel, Simone Picelli, Viktoria Konya, Jakob Theorell, Danielle Friberg, Rickard Sandberg, and Jenny Mjösberg. The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. *Nature Immunology*, 17:451–460, 2016.

[252] Valentina Proserpio, Andrea Piccolo, Liora Haim-Vilmovsky, Gozde Kar, Tapio Lönnberg, Valentine Svensson, Jhuma Pramanik, Kedar Nath Natarajan, Weichao Zhai, Xiuwei Zhang, Giacomo Donati, Melis Kayikci, Jurij Kotar, Andrew N. J. McKenzie, Ruddy Montandon, Oliver Billker, Steven Woodhouse, Pietro Cicuta, Mario Nicodemi, and Sarah A. Teichmann. Single-cell analysis of CD4+ T-cell differentiation reveals three major cell states and progressive acceleration of proliferation. *Genome Biology*, 17(103):1–15, 2016.

[253] Tapio Lönnberg, Valentine Svensson, Kylie R. James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan S. F. Soon, Lily G. Fogg, Arya Sheela Nair, Urijah N. Liligeto, Michael J. T. Stubbington, Lam-Ha Ly, Frederik Otzen Bagger, Max Zwiessele, Neil D. Lawrence, Fernando Souza-Fonseca-Guimaraes, Patrick T. Bunn, Christian R. Engwerda, William R. Heath, Oliver Billker, Oliver Stegle, Ashraful Haque, and Sarah A. Teichmann. Single-cell RNA-seq and computational analysis

using temporal mixture modeling resolves Th1/Tfh fate bifurcation in malaria. *Science Immunology*, 2:1–11, 2017.

[254] Arianne C. Richard, Aaron T.L. Lun, Winnie W.Y. Lau, Berthold Göttgens, John C. Marioni, and Gillian M. Griffiths. T cell cytolytic capacity is independent of initial stimulation strength. *Nature Immunology*, 19(August):1–10, 2018.

[255] Boyko Kakaradov, Janilyn Arsenio, Christella E Widjaja, Zhaoren He, Stefan Aigner, Patrick J Metz, Bingfei Yu, Ellen J Wehrens, Justine Lopez, Stephanie H Kim, Elina I Zuniga, Ananda W Goldrath, John T Chang, and Gene W Yeo. Early transcriptional and epigenetic regulation of CD8+ T cell differentiation revealed by single-cell RNA sequencing. *Nature Immunology*, 18(4):422–432, 2017.

[256] Michael J.T. Stubbington, Tapio Lönnberg, Valentina Proserpio, Simon Clare, Anneliese O. Speak, Gordon Dougan, and Sarah A. Teichmann. T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods*, 13(4):329–332, 2016.

[257] Stefan Canzar, Karlynn E. Neu, Qingming Tang, Patrick C. Wilson, and Aly A. Khan. BASIC: BCR assembly from single cells. *Bioinformatics*, 33(3):425–427, 2017.

[258] Ida Lindeman, Guy Emerton, Lira Mamanova, Omri Snir, Krzysztof Polanski, Shuo-Wang Qiao, Ludvig M. Sollid, Sarah A. Teichmann, and Michael J. T. Stubbington. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq To. *Nature Methods*, 15:563–565, 2018.

[259] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, HermanyMunguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

[260] Martin Häring, Amit Zeisel, Hannah Hochgerner, Puneet Rinwa, Jon E.T. Jakobsson, Peter Lönnerberg, Gioele La Manno, Nilesh Sharma, Lotta Borgius, Ole Kiehn, Malin C. Lagerström, Sten Linnarsson, and Patrik Ernfors. Neuronal atlas of the dorsal horn defines its architecture and links sensory input to transcriptional cell types. *Nature Neuroscience*, 21(6):869–880, 2018.

[261] Amit Zeisel, Hannah Hochgerner, Peter Lonnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Haring, Emelie Braun, Lars Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Nathan Skene, Kenneth D Harris, Jens Hjerling Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular architecture of the mouse nervous system. *bioRxiv*, 2018.

[262] Kristofer Davie, Jasper Janssens, Duygu Koldere, Maxime De Waegeneer, Uli Pech, Łukasz Kreft, Sara Aibar, Samira Makhzami, Valerie Christiaens, Carmen Bravo González-Blas, Suresh Poovathingal, Gert Hulselmans, Katina I. Spanier, Thomas Moerman, Bram Vanspauwen, Sarah Geurs, Thierry Voet, Jeroen Lammertyn, Bernard Thienpont, Sha Liu, Nikos Konstantinides, Mark Fiers, Patrik Verstreken, and Stein Aerts. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell*, 174:982–998, 2018.

[263] Karsten Bach, Sara Pensa, Marta Grzelak, James Hadfield, David J. Adams, John C. Marioni, and Walid T. Khaled. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications*, 8(2128):1–11, 2017.

[264] Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beáta Tóth, Doron Lemze, Matan Golan, Efi E. Massasa, Shaked Baydatch, Shanie Landen, Andreas E. Moor, Alexander Brandis, Amir Giladi, Avigail Stokar-Avihail, Eyal David, Ido Amit, and Shalev Itzkovitz. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, 542(7641):352–356, 2017.

[265] Matthew D. Young, Thomas J. Mitchell, Felipe A. Vieira Braga, Maxine G. B. Tran, Benjamin J. Stewart, John R. Ferdinand, Grace Collord, Rachel A. Botting, Dorin-Mirel Popescu, Kevin W. Loudon, Roser Vento-Tormo, Emily Stephenson, Alex Cagan, Sarah J. Farndon, Martin Del Castillo Velasco-Herrera, Charlotte Guzzo, Nathan Richoz, Lira Mamanova, Tevita Aho, James N. Armitage, Antony C. P. Riddick, Imran Mushtaq, Stephen Farrell, Dyanne Rampling, James Nicholson, Andrew Filby, Johanna Burge, Steven Lisgo, Patrick H. Maxwell, Susan Lindsay, Anne Y. Warren, Grant D. Stewart, Neil Sebire, Nicholas Coleman, Muzlifah Haniffa, Sarah A. Teichmann, Menna Clatworthy, and Sam Behjati. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science*, 361(6402):594–599, 2018.

[266] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 334:1396–1400, 2014.

[267] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, M. H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-regester, J.-R. Jia-ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, A.-C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jane-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, L. A. Garraway, Marc H Wadsworth Ii, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-regester, J.-R. Jia-ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, and Kellie E. Kolb. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016.

[268] Sidharth V. Puram, Itay Tirosh, Anuraag S. Parikh, Anoop P. Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L. Luo, Edmund A. Mroz, Kevin S. Emerick, Daniel G. Deschler, Mark A. Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James W. Rocco, William C. Faquin, Derrick T. Lin, Aviv Regev, and Bradley E. Bernstein. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*, 171(7):1611–1624, 2017.

[269] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Computational Biology*, 14:e1006245, 2018.

[270] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*, 2018.

[271] Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018.

[272] José M Bernardo. Bayesian statistics. In *Encyclopedia of Life Support Systems (EOLSS)*, volume Prob. Stat, pages 1–46. UNSECO, 2003.

[273] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418, 1763.

[274] Daniel Fink. A Compendium of Conjugate Priors. Technical report, Environmental Statistics Group, Department of Biology, Montana State Univeristy, Bozeman, MT 59717, 1997.

[275] H. Jeffreys. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186(1007):453–461, 1946.

[276] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[277] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[278] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Ltd, 2000.

[279] George Casella and Edward I George. Explaining the Gibbs Sampler Stable. *The American Statistician*, 46(3):167–174, 1992.

[280] Alan E. Gelfand and Adrian F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[281] Charles J. Greyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

[282] J Besag and Pj Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B ((Methodologica)*, 55(1):25–37, 1993.

[283] Andrew Gelman and Dinald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992.

[284] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[285] Gareth O. Roberts and Adrian F.M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2):207–216, 1994.

[286] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[287] Gareth O. Roberts and Jeffrey S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[288] Luke Tierney. Exploring posterior distributions using Markov chains. Technical report, School of Statistics, University of Minnesota, Minneapolis, MN 55455, 1991.

[289] John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, pages 169–193, 1992.

[290] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[291] Matthew J Beal and Zoubin Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian statistics 7: proceedings of the seventh Valencia International Meeting, June 2-6, 2002*, page 453, 2003.

[292] Harold Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 3rd edition, 1961.

[293] Robert Kass and Adrian Raftery. Bayes factors. *J ournal of the American Statistical Association*, 90(430):773 – 95, 1995.

[294] Natalia Bochkina and Sylvia Richardson. Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics*, 63(4):1117–1125, 2007.

[295] Catalina A Vallejos, Sylvia Richardson, and John C Marioni. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biology*, 17(70), 2016.

[296] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17, 2018.

[297] Romain Lopez, Jeffrey Regier, Michael Cole, Michael Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018.

[298] Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2015.

[299] Elham Azizi, Sandhya Prabhakaran, Ambrose Carr, and Dana Pe'er. Bayesian Inference for Single-cell Clustering and Imputing. *Genomics and Computational Biology*, 3(1):46, 2017.

[300] Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

[301] Lichun Jiang, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21:1543–1551, 2011.

[302] Michael A. Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[303] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6), 2013.

[304] Lauren N. Booth and Anne Brunet. The Aging Epigenome. *Molecular Cell*, 62(5):728–744, 2016.

[305] Jacob M Zahn, Suresh Poosala, Art B Owen, Donald K Ingram, Ana Lustig, Arnell Carter, Ashani T Weeraratna, Dennis D Taub, Myriam Gorospe, Krystyna Mazan-mamczarz, Edward G Lakatta, Kenneth R Boheler, Xiangru Xu, Mark P Mattson, Geppino Falco, Minoru S H Ko, David Schlessinger, Jeffrey Firman, Sarah K Kummerfeld, William H Wood Iii, Alan B Zonderman, Stuart K Kim, and Kevin G Becker. AGEMAP : A Gene Expression Database for Aging in Mice. *PLoS Genetics*, 3(11):2326–2337, 2007.

[306] Noweeda Mirza, Kevin Pollock, Dominique B. Hoelzinger, Ana Lucia Dominguez, and Joseph Lustgarten. Comparative kinetic analyses of gene profiles of naïve CD4+ and CD8+ T cells from young and old animals reveal novel age-related alterations. *Aging Cell*, 10(5):853–867, 2011.

[307] Petra Tollet-Egnell, Amilcar Flores-Morales, Jacob Odeberg, Joakim Lundeberg, and Gunnar Norstedt. Differential Cloning of Growth Hormone-Regulated Hepatic Transcripts in the Aged Rat. *Endocrinology*, 141(3):910–921, 2000.

[308] Tao Lu, Ying Pan, Shyan-yuan Kao, Cheng Li, and Isaac Kohane. Gene regulation and DNA damage in the ageing human brain. *Nature Genetics*, 429:883–890, 2004.

[309] Cheol-koo Lee, Richard Weindruch, and Tomas A Prolla. Gene-expression profile of the ageing brain in mice. *Nature*, 25:294–297, 2000.

[310] Stephen Welle, Andrew I Brooks, Joseph M Delehanty, Nancy Needler, Charles A Thornton, Andrew I Brooks, Joseph M Dele, Nancy Needler, and Charles A Thornton Gene. Gene expression profile of aging in human muscle. *Physiological Genomics*, 14:149–159, 2003.

[311] Jacob M Zahn, Rebecca Sonu, Hannes Vogel, Emily Crane, Krystyna Mazan-mamczarz, Ralph Rabkin, Ronald W Davis, Kevin G Becker, Art B Owen, and Stuart K Kim. Transcriptional Profiling of Aging in Human Muscle Reveals a Common Aging Signature. *PLoS Genetics*, 2(7):1058–1069, 2006.

[312] Graham E J Rodwell, Rebecca Sonu, Jacob M Zahn, James Lund, Julie Wilhelmy, Lingli Wang, Wenzhong Xiao, Michael Mindrinos, Emily Crane, Eran Segal, Bryan D Myers, James D Brooks, Ronald W Davis, John Higgins, Art B Owen, and Stuart K Kim. A Transcriptional Profile of Aging in the Human Kidney. *Plos Biology*, 2(12):2191–2201, 2004.

[313] Shigeo Yoshida, Beverly M Yashar, Suja Hiriyanna, and Anand Swaroop. Microarray Analysis of Gene Expression in the Aging Human Retina. *IOVS*, 43(8):2554–2560, 2002.

[314] Steven A Mccarroll, Coleen T Murphy, Sige Zou, Scott D Pletcher, Chen-shan Chin, Yuh Nung Jan, Cynthia Kenyon, Cornelia I Bargmann, and Hao Li. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics*, 36(2):5–7, 2004.

[315] Rudolf P. Talens, Kaare Christensen, Hein Putter, Gonneke Willemsen, Lene Christiansen, Dennis Kremer, H. Eka D. Suchiman, P. Eline Slagboom, Dorret I. Boomsma, and Bastiaan T. Heijmans. Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell*, 11(4):694–703, 2012.

[316] Shinji Maegawa, George Hinkal, Hyun Soo Kim, Lanlan Shen, Li Zhang, Jiexin Zhang, Nianxiang Zhang, Shoudan Liang, Lawrence A Donehower, and Jean-pierre J Issa. Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Research*, 20:332–340, 2010.

[317] Shuo Han and Anne Brunet. Histone methylation makes its mark on longevity. *Trends in Cell Biology*, 22(1):42–49, 2012.

[318] Mario F. Fraga and Manel Esteller. Epigenetics and aging: the targets and the marks. *Trends in Genetics*, 23(8):413–418, 2007.

[319] Riekelt H Houtkooper, Eija Pirinen, and Johan Auwerx. Sirtuins as regulators of metabolism and healthspan. *Nature Reviews Molecular Cell Biology*, 13(4):225–238, 2016.

[320] Raul Mostoslavsky, Katrin F. Chua, David B. Lombard, Wendy W. Pang, Miriam R. Fischer, Lionel Gellon, Pingfang Liu, Gustavo Mostoslavsky, Sonia Franco, Michael M. Murphy, Kevin D. Mills, Parin Patel, Joyce T. Hsu, Andrew L. Hong, Ethan Ford, Hwei Ling Cheng, Caitlin Kennedy, Nomeli Nunez, Roderick Bronson, David Frendewey, Wojtek Auerbach, David Valenzuela, Margaret Karow, Michael O. Hottiger, Stephen Hursting, J. Carl Barrett, Leonard Guarente, Richard Mulligan, Bruce Demple, George D. Yancopoulos, and Frederick W. Alt. Genomic instability and aging-like phenotype in the absence of mammalian SIRT6. *Cell*, 124(2):315–329, 2006.

[321] Philipp Oberdoerffer, Shaday Michan, Michael McVay, Raul Mostoslavsky, James Vann, Sang Kyu Park, Andrea Hartlerode, Judith Stegmuller, Angela Hafner, Patrick Loerch, Sarah M. Wright, Kevin D. Mills, Azad Bonni, Bruce A. Yankner, Ralph Scully, Tomas A. Prolla, Frederick W. Alt, and David A. Sinclair. SIRT1 Redistribution on Chromatin Promotes Genomic Stability but Alters Gene Expression during Aging. *Cell*, 135(5):907–918, 2008.

[322] João Pedro de Magalhães, João Curado, and George M. Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7):875–881, 2009.

[323] Monika S Kowalczyk, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. Single cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research*, 25:gr.192237.115, 2015.

[324] Susan L Swain, K Kai Mckinstry, and Tara M Strutt. Expanding roles for CD4+ T cells in immunity to viruses. *Nature Reviews Immunology*, 12(2):136–148, 2012.

[325] Hye-jung Kim and Harvey Cantor. CD4 T-cell Subsets and Tumor Immunity: The Helpful and the Not-so-Helpful. *Cancer Immunology Research*, 2(2):91–99, 2014.

[326] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K. Konkel, Ankit Malhotra, Adrian M. Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J. P. Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y. K. Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard a. Gibbs, Gabor Marth, Christopher E. Mason, Androniki Menelaou, Donna M. Muzny, Bradley J. Nelson, Amina Noor, Nicholas F. Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey a. Shabalin, Andreas Untergasser, Jerilyn a. Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark a. Batzer, Steven a. McCarroll, Ryan E. Mills, Mark B. Gerstein, Ali Bashir, Oliver Stegle, Scott E. Devine, Charles Lee, Evan E. Eichler, and Jan O. Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81, 2015.

[327] Findley R Finseth, Eliana Bondra, and Richard G Harrison. Selective Constraint Dominates the Evolution of Genes Expressed in a Novel Reproductive Gland. *Molecular Biology and Evolution*, 31(12):3266–3281, 2014.

[328] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gabor Csardi, Patrick Harrigan, Manuela Weier, Angelica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank, Grutzner, Sven Bergmann, Rasmus Nielsen, Svante Paabo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478:343–348, 2011.

[329] Martin F Flajnik and Masanori Kasahara. Origin and evolution of the adaptive immune system : genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, 2009.

[330] Tal Shay, Vladimir Jojic, Or Zuk, Katherine Rothamel, David Puyraimond-Zemmour, Ting Feng, Ei Wakamatsu, Christophe Benoist, Daphne Koller, and Aviv Regev. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proceedings of the National Academy of Sciences*, 110(8):2946–51, 2013.

[331] Rong Yuan, Luanne L. Peters, and Beverly Paigen. Mice as a Mammalian Model for Research on the Genetics of Aging. *ILAR Journal*, 52(1):4–15, 2011.

[332] Michael J T Stubbington, Valentina Proserpio, Anneliese O Speak, and Gordon Dougan. Simultaneously inferring T cell fate and clonality from single cell transcriptomes. *Nature Methods*, 13:329–332, 2016.

[333] Jinfang Zhu, Hidehiro Yamane, and William E. Paul. Differentiation of Effector CD4 T Cell Populations. *Annual Review of Immunology*, 28(1):445–489, 2010.

[334] Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010.

[335] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James Mccarthy, John C Marioni, and Sarah A Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(29):1–15, 2016.

[336] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2014.

[337] Catalina A Vallejos and Mark F J Steel. Objective bayesian survival analysis using shape mixtures of log-normal distributions. *Journal of the American Statistical Association*, 110(510):697–710, 2015.

[338] Antonio Scialdone, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah a. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

[339] Aaron T. L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for basic analyses of single-cell RNA-seq data. *F1000Research*, 5(2122), 2016.

[340] I-Cheng Ho, Tzong-Shyuan Tai, and Sung-Yun Pai. GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. *Nature Reviews Immunology*, 9(2):125–135, 2009.

[341] Feifan Zhang, Abhishek Bhattacharya, Jessica C Nelson, Namiko Abe, Patricia Gordon, Carla Lloret-Fernandez, Miren Maicas, Nuria Flames, Richard S Mann, Daniel a Colón-Ramos, and Oliver Hobert. The LIM and POU homeobox genes ttx-3 and unc-86 act as terminal selectors in distinct cholinergic and serotonergic neuron types. *Development*, 141(2):422–35, 2014.

[342] S Sakata-Kaneko, Y Wakatsuki, Y Matsunaga, and T Usui. Altered Th1 / Th2 commitment in human CD4 1 T cells with ageing. *Clinical Experimental Immunology*, 120:267–273, 2000.

[343] E. John Wherry. T cell exhaustion. *Nature Immunology*, 12(6):492–499, 2011.

[344] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Lane, and Richard A Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(9):R60, 2003.

[345] Mohammed Asmal, John Colgan, Felix Naef, Bin Yu, Youngnam Lee, Marcelo Magnasco, and Jeremy Luban. Production of ribosome components in effector CD4+ T cells is accelerated by TCR stimulation and coordinated by ERK-MAPK. *Immunity*, 19(4):535–548, 2003.

[346] Eva Bjur, Ola Larsson, Ekaterina Yurchenko, Lei Zheng, Valentina Gandin, Ivan Topisirovic, Shui Li, Carston R. Wagner, Nahum Sonenberg, and Ciriaco A. Piccirillo. Distinct Translational Control in CD4+ T Cell Subsets. *PLoS Genetics*, 9(5), 2013.

[347] Heon Park, Zhaoxia Li, Xuexian O. Yang, Seon Hee Chang, Roza Nurieva, Yi Hong Wang, Ying Wang, Leroy Hood, Zhou Zhu, Qiang Tian, and Chen Dong. A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunology*, 6(11):1133–1141, 2005.

[348] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[349] Mark D. Turner, Belinda Nedjai, Tara Hurst, and Daniel J. Pennington. Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1843(11):2563–2582, 2014.

[350] Christopher K. Glass and Kaoru Saijo. Nuclear receptor transrepression pathways that regulate inflammation in macrophages and T cells. *Nature Reviews Immunology*, 10(5):365–376, 2010.

[351] Steve Gerondakis and Ulrich Siebenlist. Roles of the NF-kappaB pathway in lymphocyte development and function. *Cold Spring Harbor perspectives in biology*, 2(5):1–29, 2010.

[352] Michael Croft. The role of TNF superfamily members in T-cell function and diseases. *Nature Reviews Immunology*, 9(4):271–285, 2009.

[353] Marco Antonio Moro-García, Rebeca Alonso-Arias, and Carlos López-Larrea. When aging reaches CD4+ T-cells: Phenotypic and functional changes. *Frontiers in Immunology*, 4:1–12, 2013.

[354] Tamar E. Boursalian, Jonathan Golob, David M. Soper, Cristine J. Cooper, and Pamela J. Fink. Continued maturation of thymic emigrants in the periphery. *Nature Immunology*, 5(4):418–425, 2004.

[355] J. Scott Hale, Tamar E. Boursalian, Gail L. Turk, and Pamela J. Fink. Thymic output in aged mice. *Proceedings of the National Academy of Sciences*, 103(22):8447–8452, 2006.

[356] Pamela J. Fink. The Biology of Recent Thymic Emigrants. *Annual Review of Immunology*, 31(1):31–50, 2013.

[357] Lalit K. Beura, Sara E. Hamilton, Kevin Bi, Jason M. Schenkel, Oludare A. Odumade, Kerry A. Casey, Emily A. Thompson, Kathryn A. Fraser, Pamela C. Rosato, Ali Filali-Mouhim, Rafick P. Sekaly, Marc K. Jenkins, Vaiva Vezys, W. Nicholas Haining, Stephen C. Jameson, and David Masopust. Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature*, 532(7600):512–516, 2016.

[358] Rafik Neme and Diethard Tautz. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, 5:1–20, 2016.

[359] Irene Gallego Romero, Ilya Ruvinsky, and Yoav Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516, 2012.

[360] Nuno L. Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y. Xiong, Serge Gueroussov, Leo J. Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Çolak, Tae Hyung Kim, Christine M. Misquitta-Ali, Michael D. Wilson, Philip M. Kim, Duncan T. Odom, Brendan J. Frey, and Benjamin J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.

[361] George H Perry, Páll Melsted, John C Marioni, Ying Wang, Russell Bainer, Joseph K Pickrell, Katelyn Michelini, Sarah Zehr, Anne D Yoder, Matthew Stephens, Jonathan K Pritchard, and Yoav Gilad. Comparative RNA sequencing reveals substantial genetic variation in endangered primates Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*, 22:602–610, 2012.

[362] João Pedro de Magalhães, João Curado, and George M. Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7):875–881, 2009.

[363] Guobing Chen, Ana Lustig, and Nan Ping Weng. T cell aging: A review of the transcriptional changes determined from genome-wide analysis. *Frontiers in Immunology*, 4:1–11, 2013.

[364] Jörg J. Goronzy and Cornelia M. Weyand. Understanding immunosenescence to improve responses to vaccines. *Nature Immunology*, 14(5):428–436, 2013.

[365] Janko Nikolich-Žugich. The twilight of immunity: Emerging concepts in aging of the immune system. *Nature Immunology*, 19(1):10–19, 2018.

[366] Mathieu Deschênes and Benoit Chabot. The emerging role of alternative splicing in senescence and aging. *Aging Cell*, 16(5):918–933, 2017.

[367] Peggie Cheung, Francesco Vallania, Hayley C. Warsinske, Michele Donato, Steven Schaffert, Sarah E. Chang, Mai Dvorak, Cornelia L. Dekker, Mark M. Davis, Paul J. Utz, Purvesh Khatri, and Alex J. Kuo. Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell*, 173(6):1385–1397, 2018.

[368] Keegan D. Korthauer, Li-Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, 2016.

[369] Shaohuan Wu, Ke Li, Yingshu Li, Tong Zhao, Ting Li, and Yu-fei Yang. Independent regulation of gene expression level and noise by histone modifications. *PLoS Biology*, 13(6):e1005585, 2017.

[370] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18:1509–1517, 2008.

[371] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15:1–21, 2014.

[372] Carmen Fernandez and Mark F J Steel. Multivariate Student-t Regression Models: Pitfalls and Inference. *Biometrika*, 86(1):153–167, 1999.

[373] Chantriolnt Andreas Kapourani and Guido Sanguinetti. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412, 2016.

[374] Thomas C. J. Tan, John Knight, Thomas Sbarrato, Kate Dudek, Anne E. Willis, and Rose Zamoyska. Suboptimal T-cell receptor signaling compromises protein translation, ribosome biogenesis, and proliferation of mouse CD8 T cells. *Proceedings of the National Academy of Sciences*, 114(30):E6117–E6126, 2017.

[375] Koichi Araki, Masahiro Morita, Annelise G. Bederman, Bogumila T. Konieczny, Haydn T. Kissick, Nahum Sonenberg, and Rafi Ahmed. Translation is actively regulated during the differentiation of CD8 + effector T cells. *Nature Immunology*, 18(9):1046–1057, 2017.

[376] Shane Crotty. T Follicular Helper Cell Differentiation, Function, and Roles in Disease. *Immunity*, 41(4):529–542, 2014.

[377] Emmanuelle Godefroy, Hui Zhong, Petra Pham, David Friedman, and Karina Yazdanbakhsh. TIGIT-positive circulating follicular helper T cells display robust B-cell help functions: Potential role in sickle cell alloimmunization. *Haematologica*, 100(11):1415–1425, 2015.

[378] Susanne J Szabo, Sean T Kim, Gina L Costa, Xiankui Zhang, C.Garrison Fathman, and Laurie H Glimcher. A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell*, 100(6):655–669, 2000.

[379] Raymond J. Carroll. *Measurement Error in Epidemiologic Studies*. John Wiley & Sons, Ltd, Chichester, UK, dec 1998.

[380] A. C. Oates, L. G. Morelli, and S. Ares. Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. *Development*, 139(4):625–639, 2012.

[381] Mary L. Dequéant and Olivier Pourquié. Segmental patterning of the vertebrate embryonic axis. *Nature Reviews Genetics*, 9(5):370–382, 2008.

[382] Julien Dubrulle and Olivier Pourquié. fgf8 mRNA decay establishes a gradient that couples axial elongation to patterning in the vertebrate embryo. *Nature*, 427(6973):419–422, 2004.

[383] Joseph C. Pearson, Derek Lemons, and William McGinnis. Modulating Hox gene functions during animal body patterning. *Nature Reviews Genetics*, 6(12):893–904, 2005.

[384] Baljinder S. Mankoo, Susan Skuntz, Ian Harrigan, Elena Grigorieva, Al Candia, Christopher V. E. Wright, Heinz Arnheiter, and Vassilis Pachnis. The concerted action of Meox homeobox genes is required upstream of genetic pathways essential for the formation, patterning and differentiation of somites. *Development*, 130(19):4655–4664, 2003.

[385] Kwang Won Seo, Yingdi Wang, Hiroki Kokubo, Jae R. Kettlewell, David A. Zarkower, and Randy L. Johnson. Targeted disruption of the DM domain containing transcription factor Dmrt2 reveals an essential role in somite patterning. *Developmental Biology*, 290(1):200–210, 2006.

[386] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[387] Andreas Strasser, Philipp J. Jost, and Shigekazu Nagata. The many roles of FAS receptor signaling in the immune system. *Immunity*, 30(2):180–192, 2009.

[388] Shunsuke Chikuma. Basics of PD-1 in self-tolerance, infection, and cancer immunity. *International Journal of Clinical Oncology*, 21(3):448–455, jun 2016.

[389] J-S Delisle, M Giroux, G Boucher, J-R Landry, M-P Hardy, S Lemieux, R G Jones, B T Wilhelm, and C Perreault. The TGF-β-Smad3 pathway inhibits CD28-dependent cell growth and proliferation of CD4 T cells. *Genes and Immunity*, 14(2):115–126, 2013.

[390] Kerstin Mueller, Jasmin Quandt, Ralf B. Marienfeld, Petra Weihrich, Katja Fiedler, Melina Claussnitzer, Helmut Laumen, Martin Vaeth, Friederike Berberich-Siebelt, Edgar Serfling, Thomas Wirth, and Cornelia Brunner. Octamer-dependent transcription in T cells is mediated by NFAT and NF-iB. *Nucleic Acids Research*, 41(4):2138–2154, 2013.

[391] Youn Soo Choi, Robin Kageyama, Danelle Eto, Tania C. Escobar, Robert J. Johnston, Laurel Monticelli, Christopher Lao, and Shane Crotty. ICOS Receptor Instructs T Follicular Helper Cell versus Effector Cell Differentiation via Induction of the Transcriptional Repressor Bcl6. *Immunity*, 34(6):932–946, 2011.

[392] Michael J. McDonald, Daniel P. Rice, and Michael M. Desai. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature*, 531(7593):233–236, 2016.

[393] Eugene F. Oakberg. Spermatogonial stem-cell renewal in the mouse. *The Anatomical Record*, 169(3):515–531, 1971.

[394] Dirk G. De Rooij. Spermatogonial stem cell renewal in the mouse. *Cell Tissue Kinetics*, 6:281–287, 1973.

[395] Dirk G. De Rooij and Lonnie D. Russell. All you wanted to know about spermatogonia but were afraid to ask. *Journal of Andrology*, 6512:776–798, 2000.

[396] Adèle L. Marston and Angelika Amon. Meiosis: Cell-cycle controls shuffle and deal. *Nature Reviews Molecular Cell Biology*, 5(12):983–997, 2004.

[397] Eugene F. Oakberg. A description of spermiogenesis in the mouse and its use in analysis of the cycle of the seminiferous epithelium and germ cell renewal. *American Journal of Anatomy*, 99(3):391–413, 1956.

[398] Eugene F. Oakberg. Duration of spermatogenesis in the mouse and timing of stages of the cycle of the seminiferous epithelium. *American Journal of Anatomy*, 99(3):507–516, 1956.

[399] James M. A. Turner. Meiotic sex chromosome inactivation. *Development*, 134(10):1823–1831, 2007.

[400] Aaron T. L. Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John Marioni. Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *bioRxiv*, 2018.

[401] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75, 2016.

[402] Alexander Dobin, Carrie a. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[403] Peter J. Skene, Jorja G. Henikoff, and Steven Henikoff. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nature Protocols*, 13(5):1006–1019, 2018.

[404] Aaron T. L. Lun and Gordon K. Smyth. Csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*, 44(5), 2015.

[405] Laleh Haghverdi, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[406] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015.

[407] Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):1–12, 2008.

[408] Clinton K. Matson, Mark W. Murphy, Michael D. Griswold, Shosei Yoshida, Vivian J. Bardwell, and David Zarkower. The mammalian doublesex homolog DMRT1 is a transcriptional gatekeeper that controls the mitosis versus meiosis decision in male germ cells. *Developmental Cell*, 19(4):612–624, 2010.

[409] Wei Deng and Haifan Lin. Miwi, a Murine Homolog of Piwi, Encodes a Cytoplasmic Protein Essential for Spermatogenesis. *Developmental Cell*, 2(6):819–830, 2002.

[410] Takayuki Fujii, Kazutoshi Tamura, Kumiko Masai, Hiromitsu Tanaka, Yoshitake Nishimune, and Hiroshi Nojima. Use of stepwise subtraction to comprehensively isolate mouse genes whose transcription is up-regulated during spermiogenesis. *EMBO Reports*, 3(4):367–372, 2002.

[411] S. Mazaud-Guittot, E. Meugnier, S. Pesenti, X. Wu, H. Vidal, A. Gow, and B. Le Magueresse-Battistoni. Claudin 11 Deficiency in Mice Results in Loss of the Sertoli Cell Epithelial Phenotype in the Testis1. *Biology of Reproduction*, 82(1):202–213, 2010.

[412] Gerardo M. Oresti, Jesús García-López, Marta I. Aveldanõ, and Jesús Del Mazo. Cell-type-specific regulation of genes involved in testicular lipid metabolism: Fatty acid-binding proteins, diacylglycerol acyltransferases, and perilipin 2. *Reproduction*, 146(5):471–480, 2013.

[413] Henri Bastos, Bruno Lassalle, Alexandra Chicheportiche, Lydia Riou, Jacques Testart, Isabelle Allemand, and Pierre Fouchet. Flow cytometric characterization of viable meiotic and postmeiotic cells by Hoechst 33342 in mouse spermatogenesis. *Cytometry Part A*, 65(1):40–49, 2005.

[414] Valeriya Gaysinskaya, Ina Y. Soh, Godfried W. van der Heijden, and Alex Bortvin. Optimized flow cytometry isolation of murine spermatocytes. *Cytometry Part A*, 85(6):556–565, 2014.

[415] D M Lam, R Furrer, and W R Bruce. The separation, physical characterization, and differentiation kinetics of spermatogonial cells of the mouse. *Proceedings of the National Academy of Sciences*, 65(1):192–9, 1970.

[416] Marvin L Meistrich. Separation of Spermatogeneic Cells and Nuclei from Rodent Testes. In *Methods in Cell Biology*, pages 15–54. Academic Press, New York, 1977.

[417] L. J. Romrell, Anthony R. Bellvé, and Don W. Fawcett. Separation of mouse spermatogenic cells by sedimentation velocity. A morphological characterization. *Developmental Biology*, 49(1):119–131, 1976.

[418] Magali Soumillon, Anamaria Necsulea, Manuela Weier, David Brawand, Xiaolan Zhang, Hongcang Gu, Pauline Barthès, Maria Kokkinaki, Serge Nef, Andreas Gnirke, Martin Dym, Bernard DeMassy, Tarjei S. Mikkelsen, and Henrik Kaessmann. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*, 3(6):2179–2190, 2013.

[419] Anthony R Bellve, J. C. Cavicchia, Clarke F. Millette, Deborah A. O'Brien, Y. M. Bhatnagar, and Martin Dym. Spermatogenic cells of the prepubertal mouse. Isolation and morphological characterization. *Journal of Cell Biology*, 74(1):68–85, 1977.

[420] Frank C. Janca, Lorna K. Jost, and Donald P. Evenson. Mouse testicular and sperm cell development characterized from birth to adulthood by dual parameter flow cytometry. *Biology of Reproduction*, 34(4):613–23, 1986.

[421] Bernard R. Nebel, Anthony P. Amarose, and Elizabeth M. Hackett. Calendar of gametogenic development in the prepubertal male mouse. *Science*, 134:832–833, 1961.

[422] James M.A. Turner, Shantha K. Mahadevaiah, Oscar Fernandez-Capetillo, André Nussenzweig, Xiaoling Xu, Chu Xia Deng, and Paul S. Burgoyne. Silencing of unsynapsed meiotic chromosomes in the mouse. *Nature Genetics*, 37(1):41–47, 2004.

[423] Elizabeth M. Snyder, Christopher Small, and Michael D. Griswold. Retinoic Acid Availability Drives the Asynchronous Initiation of Spermatogonial Differentiation in the Mouse1. *Biology of Reproduction*, 83(5):783–790, 2010.

[424] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[425] Klaus Steger. Transcriptional and translational regulation of gene expression in haploid spermatids. *Anatomy and embryology*, 199(6):471–487, 1999.

[426] K. Svechnikov, L. Landreh, J. Weisser, G. Izzo, E. Colón, I. Svechnikov, and O. Söder. Origin, development and regulation of human leydig cells. *Hormone Research in Paediatrics*, 73(2):93–101, 2010.

[427] Syed G. Haider. Cell Biology of Leydig Cells in the Testis. *International Review of Cytology*, 233:181–241, 2004.

[428] Michael D. Griswold. The central role of Sertoli cells in spermatogenesis. *Seminars in Cell and Developmental Biology*, 9(4):411–416, 1998.

[429] G. Lottrup, J. E. Nielsen, L. L. Maroun, L. M.A. Møller, M. Yassin, H. Leffers, N. E. Skakkebæk, and E. Rajpert-De Meyts. Expression patterns of DLK1 and INSL3 identify stages of Leydig cell differentiation during normal development and in testicular pathologies, including testicular cancer and Klinefelter syndrome. *Human Reproduction*, 29(8):1637–1650, 2014.

[430] S.L. Griswold and R.R. Behringer. Fetal Leydig Cell Origin and Development. *Sexual Development*, 3(1):1–15, 2009.

[431] J. Cool, F. D. Carmona, J. C. Szucsik, and B. Capel. Peritubular myoid cells are not the migrating population required for testis cord formation in the XY gonad. *Sexual Development*, 2(3):128–133, 2008.

[432] Shou Ching Shih, Andrew Zukauskas, Dan Li, Guanmei Liu, Lay Hong Ang, Janice A. Nagy, Lawrence F. Brown, and Harold F. Dvorak. The l6 protein TM4SF1 is critical for endothelial cell function and tumor angiogenesis. *Cancer Research*, 69(8):3272–3277, 2009.

[433] Richard L. Kitchens. Role of CD14 in cellular recognition of bacterial lipopolysaccharides. *Chemical immunology*, 74:61–82, 2000.

[434] A. Díez-Torre, U. Silván, P. Moreno, J. Gumucio, and J. Aréchaga. Peritubular myoid cell-derived factors and its potential role in the progression of testicular germ cell tumours. *International Journal of Andrology*, 34(4 PART 2), 2011.

[435] Alexander N. Combes, Dagmar Wilhelm, Tara Davidson, Elisabetta Dejana, Vincent Harley, Andrew Sinclair, and Peter Koopman. Endothelial cell migration directs testis cord formation. *Developmental Biology*, 326(1):112–120, 2009.

[436] Monika Fijak and Andreas Meinhardt. The testis in immune privilege. *Immunological Reviews*, 213(1):66–81, 2006.

[437] S. Lukassen, E. Bosch, A. B. Ekici, and A. Winterpacht. Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Scientific Reports*, 8(1):8–14, 2018.

[438] F. William Buaas, Andrew L. Kirsh, Manju Sharma, Derek J. McLean, Jamie L. Morris, Michael D. Griswold, Dirk G. De Rooij, and Robert E. Braun. Plzf is required in adult male germ cells for stem cell self-renewal. *Nature Genetics*, 36(6):647–652, 2004.

[439] Francesca Lolicato, Rita Marino, Maria Paola Paronetto, Manuela Pellegrini, Susanna Dolci, Raffaele Geremia, and Paola Grimaldi. Potential role of Nanos3 in maintaining the undifferentiated spermatogonia population. *Developmental Biology*, 313(2):725–738, 2008.

[440] Hitomi Suzuki, Hyo Won Ahn, Tianjiao Chu, Wayne Bowden, Kathrin Gassei, Kyle Orwig, and Aleksandar Rajkovic. SOHLH1 and SOHLH2 coordinate spermatogonial differentiation. *Developmental Biology*, 361(2):301–312, 2012.

[441] Jingtao Guo, Edward J. Grow, Chongil Yi, Hana Mlcochova, Geoffrey J. Maher, Cecilia Lindskog, Patrick J. Murphy, Candice L. Wike, Douglas T. Carrell, Anne Goriely, James M. Hotaling, and Bradley R. Cairns. Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. *Cell Stem Cell*, 21(4):533–546.e6, 2017.

[442] Tsutomu Endo, Katherine A. Romer, Ericka L. Anderson, Andrew E. Baltus, Dirk G. de Rooij, and David C. Page. Periodic retinoic acid–STRA8 signaling intersects with periodic germ-cell competencies to regulate spermatogenesis. *Proceedings of the National Academy of Sciences*, 112(18):E2347–E2356, 2015.

[443] D. Ballow, M. L. Meistrich, M. Matzuk, and A. Rajkovic. Sohlh1 is essential for spermatogonial differentiation. *Developmental Biology*, 294(1):161–167, 2006.

[444] Ericka L. Anderson, Andrew E. Baltus, Hermien L. Roepers-Gajadien, Terry J. Hassold, Dirk G. de Rooij, Ans M. M. van Pelt, and David C. Page. Stra8 and its inducer, retinoic acid, regulate meiotic initiation in both spermatogenesis and oogenesis in mice. *Proceedings of the National Academy of Sciences*, 105(39):14976–14980, 2008.

[445] Andrew E. Baltus, Douglas B. Menke, Yueh Chiang Hu, Mary L. Goodheart, Anne E. Carpenter, Dirk G. De Rooij, and David C. Page. In germ cells of mouse embryonic ovaries, the decision to enter meiosis precedes premeiotic DNA replication. *Nature Genetics*, 38(12):1430–1434, 2006.

[446] A. L. Kierszenbaum and Laura L. Tres. Nucleolar and perichromosomal RNA synthesis during meiotic prophase in the mouse testis. *Journal of Cell Biology*, 60(1):39–53, 1974.

[447] Valerio Monesi. Differential rate of riobonucleis acid synthesis in the autosomes and sex chromosomes during male meiosis in the mouse. *Chromosoma*, 17:11–21, 1965.

[448] Katrin Daniel, Julian Lange, Khaled Hached, Jun Fu, Konstantinos Anastassiadis, Ignasi Roig, Howard J. Cooke, A. Francis Stewart, Katja Wassmann, Maria Jasin, Scott Keeney, and Attila Tóth. Meiotic homologue alignment and its quality surveillance are controlled by mouse HORMAD1. *Nature Cell Biology*, 13(5):599–610, 2011.

[449] Shantha K. Mahadevaiah, James M.A. Turner, Frédéric Baudat, Emmy P. Rogakou, Peter De Boer, Josefa Blanco-Rodríguez, Maria Jasin, Scott Keeney, William M. Bonner, and Paul S. Burgoyne. Recombinational DNA double-strand breaks in mice precede synapsis. *Nature Genetics*, 27(3):271–276, 2001.

[450] Femke A T De Vries, Esther De Boer, Mike Van Den Bosch, Willy M Baarends, Marja Ooms, Li Yuan, Jian-guo Liu, Albert a Van Zeeland, Christa Heyting, and Albert Pastink. Mouse Sycp1 functions in synaptonemal complex assembly, meiotic recombination, and XY body formation. *Genes & Development*, pages 1376–1389, 2005.

[451] Anil K. Rengan, Ashok Agarwal, Michelle van der Linde, and Stefan S. du Plessis. An investigation of excess residual cytoplasm in human spermatozoa and its distinction from the cytoplasmic droplet. *Reproductive Biology and Endocrinology*, 10(1):1, 2012.

[452] Y. Q.Shirleen Soh, Maria M. Mikedis, Mina Kojima, Alexander K. Godfrey, Dirk G. de Rooij, and David C. Page. Meioc maintains an extended meiotic prophase I in mice. *PLoS Genetics*, 13(4):1–33, 2017.

[453] Trevor Hastie and Werner Stuetzle. Principal Curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[454] Bo Xia, Maayan Baron, Yun Yan, Florian Wagner, Sang Y. Kim, David L. Keefe, Joseph P. Alukal, Jef D. Boeke, and Itai Yanai. Widespread transcriptional scanning in testes modulates gene evolution rates. *bioRxiv*, 2018.

[455] Giovanna Braidotti and Denise P. Barlow. Identification of a male meiosis-specific gene, Tcte2, which is differentially spliced in species that form sterile hybrids with laboratory mice and deleted in t chromosomes showing meiotic drive. *Developmental Biology*, 186(1):85–99, 1997.

[456] Bogi Andersen, Richard V Pearse, Pearse N Schlegel, Zbigniew Cichon, Marcus D Schonemann, C Wayne Bardin, and Michael G Rosenfeld. Sperm 1: a POU-domain gene transiently expressed immediately before meiosis I in the male germ cell. *Proceedings of the National Academy of Sciences of the United States of America*, 90(23):11084–8, 1993.

[457] Charles H. Spruck, Maria P. De Miguel, Adrian P.L. Smith, Aimee Ryan, Paula Stein, Richard M. Schultz, A. Jeannine Lincoln, Peter J. Donovan, and Steven I. Reed. Requirement of Cks2 for the first metaphase/anaphase transition of mammalian meiosis. *Science*, 300(5619):647–650, 2003.

[458] Rod Balhorn. The protamine family of sperm nuclear proteins. *Genome Biology*, 8(9), 2007.

[459] Sara El Kennani, Annie Adrait, Alexey K. Shaytan, Saadi Khochbin, Christophe Bruley, Anna R. Panchenko, David Landsman, Delphine Pflieger, and Jérôme Govin. MS-HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones. *Epigenetics and Chromatin*, 10(1):1–18, 2017.

[460] Ian K. Greaves, Danny Rangasamy, Michael Devoy, Jennifer A. Marshall Graves, and David J. Tremethick. The X and Y Chromosomes Assemble into H2A.Z, Containing Facultative Heterochromatin, following Meiosis. *Molecular and Cellular Biology*, 26(14):5394–5405, 2006.

[461] Michelle C.W. Tang, Shelley A. Jacobs, Deidre M. Mattiske, Yu May Soh, Alison N. Graham, An Tran, Shu Ly Lim, Damien F. Hudson, Paul Kalitsis, Moira K. O'Bryan, Lee H. Wong, and Jeffrey R. Mann. Contribution of the Two Genes Encoding Histone Variant H3.3 to Viability and Fertility in Mice. *PLoS Genetics*, 11(2):1–23, 2015.

[462] Benjamin T. K. Yuen, Kelly M. Bush, Bonnie L. Barrilleaux, Rebecca Cotterman, and Paul S. Knoepfler. Histone H3.3 regulates dynamic chromatin states during spermatogenesis. *Development*, 141(18):3483–3494, 2014.

[463] William F Marzluff, Preetam Gongidi, Keith R Woods, Jianping Jin, and Lois J Maltais. The human and mouse replication-dependent histone genes. *Genomics*, 80(5):487–98, 2002.

[464] Ming Zhao, Cynthia R. Shirley, Suzanne Mounsey, and Marvin L. Meistrich. Nucleoprotein Transitions During Spermiogenesis in Mice with Transition Nuclear Protein Tnp1 and Tnp2 Mutations. *Biology of Reproduction*, 71(3):1016–1025, 2004.

[465] Hiromitsu Tanaka, Naoko Iguchi, Ayako Isotani, Kouichi Kitamura, Yoshiro Toyama, Yasuhiro Matsuoka, Masayoshi Onishi, Kumiko Masai, Mamiko Maekawa, Kiyotaka Toshimori, Masaru Okabe, and Yoshitake Nishimune. HANP1/H1T2, a novel histone H1-like protein involved in nuclear formation and sperm fertility. *Molecular and Cellular Biology*, 25(16):7107–7119, 2005.

[466] C. Dottermusch-Heidel, E. S. Klaus, N. H. Gonzalez, S. Bhushan, A. Meinhardt, M. Bergmann, R. Renkawitz-Pohl, C. Rathke, and K. Steger. H3K79 methylation directly precedes the histone-to-protamine transition in mammalian spermatids and is sensitive to bacterial infections. *Andrology*, 2(5):655–665, 2014.

[467] Liza O'Donnell. Mechanisms of spermiogenesis and spermiation and how they are disturbed. *Spermatogenesis*, 4(2):e979623, 2014.

[468] Noora Kotaja and Paolo Sassone-corsi. The chromatoid body: a germ-cell- specific RNA-processing centre. *Nature Reviews Molecular Cell Biology*, 8:85–90, 2007.

[469] Akihiro Kawashima, Boran A.H. Osman, Minoru Takashima, Akihiko Kikuchi, Sae Kohchi, Emiko Satoh, Michiko Tamba, Manabu Matsuda, and Naomichi Okamura. CABS1 Is a Novel Calcium-Binding Protein Specifically Expressed in Elongate Spermatids of Mice1. *Biology of Reproduction*, 80(6):1293–1304, 2009.

[470] Kiyoshi Miki, William D. Willis, Paula R. Brown, Eugenia H. Goulding, Kerry D. Fulcher, and Edward M. Eddy. Targeted disruption of the Akap4 gene causes defects in sperm flagellum and motility. *Developmental Biology*, 248(2):331–342, 2002.

[471] Geert Hamer, Hermien L. Roepers-Gajadien, Annemarie van Duyn-Goedhart, Iris S. Gademan, Henk B. Kal, Paul P.W. van Buul, and Dirk G. de Rooij. DNA Double-Strand Breaks and γ-H2AX Signaling in the Testis1. *Biology of Reproduction*, 68(2):628–634, 2003.

[472] Mahesh N. Sangrithi, Helene Royo, Shantha K. Mahadevaiah, Obah Ojarikre, Leena Bhaw, Abdul Sesay, Antoine H.F.M. Peters, Michael Stadler, and James M.A. Turner. Non-Canonical and Sexually Dimorphic X Dosage Compensation States in the Mouse and Human Germline. *Developmental Cell*, 40(3):289–301.e3, 2017.

[473] James M A Turner, Shantha K. Mahadevaiah, P. J I Ellis, Michael J. Mitchell, and Paul S. Burgoyne. Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids. *Developmental Cell*, 10(4):521–529, 2006.

[474] Kazuteru Hasegawa, Ho Su Sin, So Maezawa, Tyler J. Broering, Andrey V. Kartashov, Kris G. Alavattam, Yosuke Ichijima, Fan Zhang, W. Clark Bacon, Kenneth D. Greis, Paul R. Andreassen, Artem Barski, and Satoshi H. Namekawa. SCML2 Establishes the Male Germline Epigenome through Regulation of Histone H2A Ubiquitination. *Developmental Cell*, 32(5):574–588, 2015.

[475] Ho Su Sin, Artem Barski, Fan Zhang, Andrey V. Kartashov, Andre Nussenzweig, Junjie Chen, Paul R. Andreassen, and Satoshi H. Namekawa. RNF8 regulates active epigenetic modifications and escape gene activation from inactive sex chromosomes in post-meiotic spermatids. *Genes and Development*, 26(24):2737–2748, 2012.

[476] Ho Su Sin, Andrey V. Kartashov, Kazuteru Hasegawa, Artem Barski, and Satoshi H. Namekawa. Poised chromatin and bivalent domains facilitate the mitosis-to-meiosis transition in the male germline. *BMC Biology*, 13(1):1–15, 2015.

[477] Shannel R. Adams, So Maezawa, Kris G. Alavattam, Hironori Abe, Akihiko Sakashita, Megan Shroder, Tyler J. Broering, Julie Sroga Rios, Michael A. Thomas, Xinhua Lin, Carolyn M. Price, Artem Barski, Paul R. Andreassen, and Satoshi H. Namekawa. RNF8 and SCML2 cooperate to regulate ubiquitination and H3K27 acetylation for escape gene activation on the sex chromosomes. *PLoS Genetics*, 14(2), 2018.

[478] Jacob L. Mueller, Shantha K. Mahadevaiah, Peter J. Park, Peter E. Warburton, David C. Page, and James M.A. Turner. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nature Genetics*, 40(6):794–799, 2008.

[479] Christine M. Disteche. The not-so-silent X. *Nature Genetics*, 40(6):689–690, 2008.

[480] Charlotte Moretti, Daniel Vaiman, Frederic Tores, and Julie Cocquet. Expression and epigenomic landscape of the sex chromosomes in mouse post-meiotic male germ cells. *Epigenetics and Chromatin*, 9(1):1–18, 2016.

[481] Makoto Tachibana, Masami Nozaki, Naoki Takeda, and Yoichi Shinkai. Functional dynamics of H3K9 methylation during meiotic prophase progression. *EMBO Journal*, 26(14):3346–3359, 2007.

[482] Antoine H.F.M. Peters, Dónal O'Carroll, Harry Scherthan, Karl Mechtler, Stephan Sauer, Christian Schöfer, Klara Weipoltshammer, Michaela Pagani, Monika Lachner, Alexander Kohlmaier, Susanne Opravil, Michael Doyle, Maria Sibilia, and Thomas Jenuwein. Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell*, 107(3):323–337, 2001.

[483] Hélne Royo, Grzegorz Polikiewicz, Shantha K. Mahadevaiah, Haydn Prosser, Mike Mitchell, Allan Bradley, Dirk G. De Rooij, Paul S. Burgoyne, and James M.A. Turner. Evidence that meiotic sex chromosome inactivation is essential for male fertility. *Current Biology*, 20(23):2117–2123, 2010.

[484] Manu Setty, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(April):1–14, 2016.

[485] John E. Reid and Lorenz Wernisch. Pseudotime Estimation: Deconfounding Single Cell Time Series. *Bioinformatics*, page btw372, 2016.

[486] Kieran R. Campbell and Christopher Yau. Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. *PLOS Computational Biology*, 12(11):e1005212, 2016.

[487] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016.

[488] Jeffrey M. Good, Dan Vanderpool, Kimberly L. Smith, and Michael W. Nachman. Extraordinary sequence divergence at Tsga8, an X-linked gene involved in mouse spermiogenesis. *Molecular Biology and Evolution*, 28(5):1675–1686, 2011.

[489] Yao Chen, Yuxuan Zheng, Yun Gao, Zhen Lin, Suming Yang, Tongtong Wang, Qiu Wang, Nannan Xie, and Rong Hua. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Research*, 0(July):1–18, 2018.

[490] George J Bosl and Robert J. Motzer. Testicular Germ Cell Cancer. *The New England Journal of Medicine*, 337(4):242 – 253, 1997.

[491] Shantha K. Mahadevaiah, Déborah Bourc'his, Dirk G. De Rooij, Timothy H. Bestor, James M.A. Turner, and Paul S. Burgoyne. Extensive meiotic asynapsis in mice antagonises meiotic silencing of unsynapsed chromatin and consequently disrupts meiotic sex chromosome inactivation. *Journal of Cell Biology*, 182(2):263–276, 2008.

[492] J. M. Cloutier, S. K. Mahadevaiah, E. ElInati, A. Tóth, and James Turner. Mammalian meiotic silencing exhibits sexually dimorphic features. *Chromosoma*, 125(2):215–226, 2016.

[493] Teruko Taketo and Anna K. Naumova. Oocyte heterogeneity with respect to the meiotic silencing of unsynapsed X chromosomes in the XY female mouse. *Chromosoma*, 122(5):337–349, 2013.

[494] James M.A. Turner, Olga Aprelikova, Xiaoling Xu, Ruihong Wang, Sangsoo Kim, Gadisetti V.R. Chandramouli, J. Carl Barrett, Paul S. Burgoyne, and Chu-Xia Deng. BRCA1, Histone H2AX Phosphorylation, and Male Meiotic Sex Chromosome Inactivation. *Current Biology*, 14:2135–2142, 2004.

[495] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El Ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Pe'er, Scott D. Tanner, and Garry P. Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

[496] Charlotte Giesen, Hao A.O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, Daniel Grolimund, Joachim M. Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J. Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, 2014.

[497] Gabriele Gut, Markus D Herrmann, and Lucas Pelkmans. Multiplexed protein maps link subcellular organization to cellular states. *Science*, 7042(361), 2018.

[498] Daniel Schulz, Vito Riccardo Tomaso Zanotelli, Jana Raja Fischer, Denis Schapiro, Stefanie Engler, Xiao Kang Lun, Hartland Warren Jackson, and Bernd Bodenmiller. Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. *Cell Systems*, 6(4):531, 2018.

[499] Martin J. Zhang, Vasilis Ntranos, and David Tse. One read per cell per gene is optimal for single-cell RNA-Seq. *bioRxiv*, 2018.

[500] Hannah Dueck, James Eberwine, and Junhyong Kim. Variation is function: Are single cell differences functionally important?: Testing the hypothesis that single cell variation is required for aggregate function. *BioEssays*, 38(2):172–180, 2016.

[501] Peter C. Scacheri, Orit Rozenblatt-Rosen, Natasha J. Caplen, Tyra G. Wolfsberg, Lowell Umayam, Jeffrey C. Lee, Christina M. Hughes, Kalai Selvi Shanmugam, Arindam Bhattacharjee, Matthew Meyerson, and Francis S. Collins. Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells. *Proceedings of the National Academy of Sciences*, 101(7):1892–1897, 2004.

[502] Manfred Gossen, Sabine Freundlieb, Gabriele Bender, Gerhard Mueller, Wolfgang Hillen, and Hermann Bujard. Transcriptional Activation By Tetracyclines in Mammalian-Cells. *Science*, 268(5218):1766–1769, 1995.

[503] Jorn Schmiedel, Debora S. Marks, Ben Lehner, and Nils Bluthgen. Noise control is a primary function of microRNAs and post-transcriptional regulation. *bioRxiv*, 2017.

[504] William A. Cantara, Pamela F. Crain, Jef Rozenski, James A. McCloskey, Kimberly A. Harris, Xiaonong Zhang, Franck A.P. Vendeix, Daniele Fabris, and Paul F. Agris. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Research*, 39:195–201, 2011.

[505] Tong Chen, Ya Juan Hao, Ying Zhang, Miao Miao Li, Meng Wang, Weifang Han, Yongsheng Wu, Ying Lv, Jie Hao, Libin Wang, Ang Li, Ying Yang, Kang Xuan Jin, Xu Zhao, Yuhuan Li, Xiao Li Ping, Wei Yi Lai, Li Gang Wu, Guibin Jiang, Hai Lin Wang, Lisi Sang, Xiu Jie Wang, Yun Gui Yang, and Qi Zhou. M6A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell*, 16(3):289–301, 2015.

[506] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the last decade. *Nature protocols*, 13(4):599–604, 2018.

[507] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv*, 2014.

[508] Christina Ernst, Jeremy Pike, Sarah J Aitken, Hannah K Long, Nils Eling, Lovorka Stojic, Michelle C Ward, Frances Connor, Timothy F Rayner, Margus Lukk, Robert J Klose, Claudia Kutter, and Duncan T Odom. Successful transmission and transcriptional deployment of a human chromosome via mouse male meiosis. *eLife*, 5:e20235, 2016.

Appendix A

Experimental methods

A.1 Ageing increases transcriptional noise in CD4$^+$ T cell activation

A.1.1 Mouse material

CAST/EiJ male mice were maintained under specific pathogen-free conditions at the University of Cambridge, CRUK – Cambridge Institute under the auspices of a UK Home Office license. Inbred wild-type C57BL/6 mice were purchased from Charles River UK Ltd (Margate, United Kingdom). Animals were euthanized in accordance with Schedule 1 of the Animals (Scientific Procedures) Act 1986. Each animal used was macroscopically examined. Animals with lesions or phenotypic alterations in their internal organs were discarded.

A.1.2 CD4$^+$ T cell isolation

Unstimulated CD4$^+$ T cells were purified from dissociated mouse spleens using EASY cell strainer (30 µm, Greiner BioOne), cell separation media (lympholyte, #CL5035) and the CD4$^+$ CD62L$^+$ T Cell Isolation Kit II (Miltenyi Biotec, #130-093-227). Flow cytometry confirmed that 96.4% of the isolated CD4$^+$ T cells were naive in young B6 **(Fig. 2.4D)**. Naive CD4$^+$ T cells formed a single, high-purity population in young animals. Old animals had a small population of CD4$^+$ T cells with slightly elevated CD44 levels, reduced CD62L expression, and attenuated activation dynamics **(Fig. 2.4E-G)**; their removal did not impact the results presented in **Chapter 2 (Fig. 2.14D)**.

Purified unstimulated CD4$^+$ T cells were cultured in IMDM medium (GIBCO, #21980-032) supplemented with 10% Fetal Bovine Serum (Life Technology, #10500064), 1 µg/mL Penicillin/Streptomicin (Life Technology, #15070063), and 50 µM 2-mercaptoethanol (Gibco, #31350-010). Cells were seeded into 96-well plates coated for 1h at 37°C with anti-CD3ε (1 µg/ml, clone: 145-2C11, eBioscience, #16-0031-82) and anti-CD28 (3 µg/ml, clone: 37.51, eBioscience, #16-0281-82) at a density of 80,000-120,000 cells/ml, and then cultured in a total volume of 100 µl media that did not contain cytokines or additional antibodies.

All cells were cultured in a humidified incubator at 37°C, with 5% CO2. Unstimulated and activated CD4$^+$ T cells were then immediately collected and loaded on a 5–10µm Auto Prep Integrated Fluidic Circuit (IFC; Fluidigm, San Francisco, CA) to capture single cells using the C1 single-cell Auto Prep System (Fluidigm). All IFCs were visually inspected, and wells with multiple cells or cell debris were identified per instructions of the manufacturer (PN 101-2711 A1 White Paper). Upon cell capture, reverse transcription and cDNA amplification were performed using the SMARTer PCR cDNA Synthesis Kit (Clontech) and the Advantage 2 PCR Kit (Clontech). ERCC spike-in RNA (Ambion) (1 µL diluted at 1:50,000) was added to the C1 lysis mix. All the capture sites were included for the RNA-seq library preparation, and wells identified above as multiple cells or containing debris were removed during computational analysis.

A.1.3 Flow cytometry

Unstimulated CD4$^+$ T cells were purified from spleens of young and old C57BL/6 mice (see above). Isolated cells were, directly or after 3h activation in vitro (see above), incubated with TruStainfcX (anti-mouse CD16/32, clone:93, BioLegend) before staining with immunofluorescence conjugated antibodies against murine CD4 (clone: RM4-5, BioLegend), CD44 (clone: IM7, BioLegend), CD62L (clone: MEL-14, BioLegend), CD25 (clone: 3C7, BioLegend), CD69 (clone: H1.2F3, BioLegend), CD127 (clone: A7R34, BioLegend), and KLRG1 (clone: 2F1, BD Biosciences). Cell viability was determined using Fixable eFluor 780 viability dye (eBioscience). Data were acquired on a 5-laser Aria IIu SORP instrument (BD Biosciences) and data analysis was performed using FlowJo software (Tree Star).

Naive and effector memory CD4$^+$ T cells were purified from spleens of both young and old C57BL/6 mice by FACS. Briefly, spleens were harvested from both young and old animals and single cell suspensions were obtained by meshing through a cell strainer (70 µm). B cells were depleted from cell suspensions by MACS using CD19 microbeads (Miltenyi

Biotec, #130-052-201) and red blood cells were lysed with RBC lysis buffer (Biolegend, #B205551). The enriched cell fraction was then stained with Fixable eFluor 780 viability dye (eBioscience) following by Fc receptor blocking with TruStain fcXTM (clone: 39, Biolegend) and subsequent staining with a panel of fluorescence-conjugated antibodies against CD4 (clone: RM4-5, BioLegend), CD44 (clone: IM7, BioLegend), CD62L (clone: MEL-14, BioLegend), CD24 (clone: M1/69, BioLegend), Qa2 (clone: 695H1-9-9, BioLegend), CD69 (clone: H1.2F3, BioLegend) and PD-1 (clone: RMP1-30, BioLegend). Stained cells were immediately sorted using a 5-laser Aria IIu SORP instrument (BD Biosciences) with the stringent gating strategy described in **Fig. 2.2**.

A.1.4 ScRNA-Seq library preparation and sequencing

ScRNA-Seq libraries were prepared using standard Fluidigm protocol (# PN 100-7168 K1) based on SMARTer chemistry and Illumina Nextera XT (Illumina) using paired-end 125bp sequencing on Illumina HiSeq2500. Each RNA-seq library was sequenced to a typical depth of 1.3 million reads on average. To account for potential batch effects, for each experimental condition, two biological replicates were prepared using independent C1 IFCs.

A.2 Transcriptional dynamics during spermatogenesis at single-cell resolution

A.2.1 Mouse material

All animals were housed in the Biological Resources Unit (BRU) in the Cancer Research UK – Cambridge Institute under Home Office Licences PPL 70/7535 until February 2018 and PPL P9855D13B from March 2018. C57BL/6 animals were purchased from Charles River UK Ltd (Margate, United Kingdom).

A.2.2 FACS of spermatogenic cell populations

Spermatogenic cell populations were isolated from adult mouse testes as described in Ernst *et al.*, 2016 [508]. In brief, the albuginea was removed and tissue was incubated in dissociation buffer containing 25 mg/ml Collagenase A, 25 mg/ml Dispase II and 2.5 mg/ml DNase I for 30 minutes at 37°C. Enzymatic digestion was quenched with Dulbecco's Modified Eagle Medium (DMEM, Gibco) supplemented with 10% Fetal calf serum (FCS, 10270106, Gibco). Cells were resuspended at a concentration of 1 million cells per ml and stained with Hoechst 33342 (H3570, ThermoFisher Scientific) at a final concentration of 5 µg/ml for 45 minutes at 37°C. Cells were resuspended in PBS containing 1% FCS and 2 mM EDTA and propidium iodide was added to a final concentration of 1 µg/ml prior to sorting.
Cells were sorted on an Aria IIu cell sorter (Becton Dickinson) using a 100 µm nozzle. Hoechst was excited with a UV laser at 355nm and fluorescence was recorded with a 450/50 filter (Hoechst blue) and 635LP filter (Hoechst red). Primary spermatocytes (4N) and round spermatids (1N) were sorted and collected in PBS containing 1% FCS and 2 mM EDTA.

A.2.3 Total RNA-Seq from bulk samples

Testes from prepubertal mice ranging between postnatal day 6 and 35 were flash frozen or directly used for RNA extraction using Trizol (Thermo Fisher, 15596026) following manufacturer's instructions. Purified RNA was DNase-treated using the TURBO DNA-free Kit according to manufacturer's instructions (Thermo Fisher, AM1907) and RNA quality was assessed using the Agilent Tapestation RNA Screentape. 800 ng of DNA-depleted RNA were used for RNA-Seq library preparation using the TruSeq Stranded Total RNA Library Kit with Ribo-Zero Gold for cytoplasmic and mitochondrial ribosomal RNA removal according

to manufacturer's instructions (Illumina, RS-122-2303). Libraries were then sequenced on Illumina HiSeq2500 using a paired-end 125bp run.

A.2.4 10X Genomics single-cell RNA-Seq

Mouse testes were enzymatically dissociated as described above and 34 µl of single-cell suspension at a concentration of 297,000 cells/ml was loaded into one channel of the ChromiumTM Single Cell A Chip (10X Genomics®), aiming for a recovery of 4000-5000 cells. The Chromium Single Cell 3' Library & Gel Bead Kit v2 (10X Genomics®, 120237) was used for single-cell barcoding, cDNA synthesis and library preparation, following manufacturer's instructions according to the Single Cell 3' Reagent Kits User Guide Version 2, Revision D. Libraries were sequenced on Illumina HiSeq2500 using a paired-end run sequencing 26bp on read 1 and 98bp on read 2.

A.2.5 Histology

Testes were fixed in neutral buffered formalin (NBF) for 24 hours, transferred to 70% ethanol, machine processed and paraffin embedded. Formalin-fixed paraffin-embedded (FFPE) sections of 3µm thickness were used for all histological stains and immunohistochemistry (IHC). For Periodic Acid Schiff (PAS) stainings slides were dewaxed, washed in water and placed in 0.5% Periodic Acid (Sigma P0430) for 5 minutes. After three washes in ultra-pure water, slides were placed in Schiff reagent (Thermo Fisher Scientific, J/7300/PB08) for 15-30 minutes in a closed container and washed again three times in ultra-pure water. Counterstain was performed using Mayers Haematoxylin (Thermo Fisher Scientific, LAMB/170-D) for 40 seconds followed by rinsing in tap water, dehydration and mounting. IHC was performed on FFPE sections using the Bond™ Polymer Refine Kit (DS9800, Leica Microsystems) on the automated Bond Platform. Anti-phospho-Histone H3 (Ser10) (pH3) antibody (Upstate, 06-570, 1:200 dilution) was used with DAB Enhancer (Leica Microsystems, AR9432) and heat-induced epitope retrieval was performed for 10 minutes at 100°C on the Bond platform with sodium citrate. All slides were scanned using Aperio XT (Leica Biosystems) and PH3 intensities were quantified using the Aperio eSlide Manager (Leica Biosystems).

A.2.6 Low cell number chromatin profiling using CUT&RUN

In situ chromatin profiling of FACS-purified spermatogenic cell populations was performed according to Skene *et al.*, 2018 [403]. In brief, spermatocytes and spermatids were sorted as described above and collected in PBS. Cells were spun down at 600g for 3 minutes in swinging-bucket rotor and washed twice with 1.5 ml Wash buffer (20 mM HEPES-KOH (pH 7.5), 150 mM NaCl, 0.5 mM Spermidine and 1X cOmplete™ EDTA-free protease inhibitor cocktail (04693159001, Roche)). During the cell washes, concanavalin A-coated magnetic beads (Bangs Laboratories, cat. No BP531) (10 µl per condition) were washed twice in 1.5 mL Binding Buffer (20 mM HEPES-KOH (pH 7.5), 10 mM KCl, 1mM CaCl, 1mM MnCl2) and resuspended in 10 µl Binding Buffer per condition. Cells were then mixed with beads and rotated for 10 minutes at room temperature (RT) and samples were split into aliquots according to number of antibodies profiled per cell type. We used 20,000-30,000 spermatocytes and 40,000-60,000 spermatids per chromatin mark.

Cells were then collected on magnetic beads and re-suspended in 50 µl Antibody Buffer (Wash buffer with 0.05% Digitonin and 2 mM EDTA) containing one of the following antibodies in 1:100 dilution: H3K4me3 (Millipore 05-1339 CMA304, Lot2780484) and H3K9me3 (Abcam, ab8898, Lot GR306402-1). Cells were incubated with antibodies for 10 minutes at RT and then washed once with 1 ml Digitonin buffer (Wash buffer with 0.05% Digitonin). For the mouse anti-H3K4me3 antibody, samples were incubated with a 1:100 dilution in Digitonin buffer of secondary rabbit anti-mouse antibody (Invitrogen, A27033, Lot RG240909) for 10 minutes at RT and then washed once with 1 mL Digitonin buffer. Samples were then incubated with 700 ng/ml ProteinA-MNase fusion protein (kindly provided by Steven Henikoff) for 10 minutes at room temperature followed by two washes with 1 ml Digitonin buffer. Cells were then resuspended in 100 µl Digitonin buffer and cooled down to 4°C before addition of $CaCl_2$ to a final concentration of 2 mM. Targeted digestion was performed for 30 minutes on ice until 100 µl of 2X STOP buffer (340 mM NaCl, 20 mM EDTA, 4 mM EGTA, 0.02% Digitonin, 250 mg RNase A, 250 µg Glycogen, 15 pg/ml yeast spike-in DNA (kindly provided by Steven Henikoff)) were added. Cells were then incubated at 37°C for 10 minutes to release cleaved chromatin fragments, spun down for 5 minutes at 16,000 g at 4°C and collected on magnet. Supernatant containing the cleaved chromatin fragments was then transferred and cleaned up using the Zymo Clean & Concentrator Kit. Library preparation was performed using the ThruPLEX® DNA-Seq Library Preparation Kit (R400407, Rubicon Genomics) with a modified Library Amplification programme: Extension

and cleavage for 3 minutes at 72°C followed by 2 minutes at 85°C, denaturation for 2 minutes at 98°C followed by four cycles of 20 seconds at 98°C, 20 seconds at 67°C and 40 seconds at 72°C for the addition of indexes. Amplification was then performed for 12-14 cycles of 20 seconds at 98°C and 15 seconds at 72°C. Average library size was tested on Agilent 4200 Tapestation using a DNA1000 High Sensitivity Screentape and quantification was performed using the KAPA Library Quantification Kit (Kapa Biosystems). CUT&RUN libraries were sequenced on a HiSeq2500 using a paired-end 125bp run.

Appendix B

Computational methods

B.1 Addressing the mean confounding effect for differential variability testing

B.1.1 Prior specifications of the extended BASiCS model

$$
\begin{aligned}
\mu_i &\overset{\text{ind}}{\sim} \text{log-Normal}\left(0, s_\mu^2\right) \\
\delta_i|\mu_i, \beta, \sigma^2, \lambda_i, \eta &\overset{\text{ind}}{\sim} \text{log-Normal}\left(\text{f}(\mu_i), \frac{\sigma^2}{\lambda_i}\right) \\
\lambda_i|\eta &\overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{\eta}{2}, \frac{\eta}{2}\right) \\
\beta|\sigma^2 &\sim \text{Normal}(m_\beta, \sigma^2 V_\beta), \\
\sigma^2 &\sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}), \\
s_j &\overset{\text{iid}}{\sim} \text{Gamma}(a_s, b_s) \\
(\phi_1, \ldots, \phi_n)' &\sim n \times \text{Dirichlet}(a_\phi), \\
\theta &\sim \text{Gamma}(a_\theta, b_\theta)
\end{aligned}
$$

B.1.2 Starting values for hyper-parameters

$$
\begin{aligned}
m_\beta &= \mathbf{0}_L \text{ (an } L\text{-dimensional vector of zeroes)} \\
V_\beta &= \mathbf{I}_L \text{ (an } L\text{-dimensional identity matrix)} \\
a_{\sigma^2} &= 2 \\
b_{\sigma^2} &= 2 \\
s^2_\mu &= 0.5 \\
a_s &= 1 \\
b_s &= 1 \\
a_\phi &= \mathbf{1}_n \\
a_\theta &= 1 \\
b_\theta &= 1
\end{aligned}
$$

B.1.3 Likelihood of the extended BASiCS model

The likelihood function of the extended BASiCS model takes the form:

$$
\begin{aligned}
\mathscr{L} = &\left[\prod_{i=1}^{q_0}\prod_{j=1}^{n} \frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})x_{ij}!}\left(\frac{\frac{1}{\delta_i}}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\phi_j\nu_j\mu_i}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right)^{x_{ij}}\right] \\
&\times\left[\prod_{i=q_0+1}^{q}\prod_{j=1}^{n}\frac{(\nu_j\mu_i)^{x_{ij}}}{x_{ij}!}\exp\{-\nu_j\mu_i\}\right]\times\left[\prod_{j=1}^{n}\frac{(s_j\theta)^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})}\nu_j^{\frac{1}{\theta}-1}\exp\left\{-\frac{\nu_j}{s_j\theta}\right\}\right]. \quad \text{(B.1)}
\end{aligned}
$$

B.1.4 Derivation of full conditionals for the extended BASiCS model

To calculate the full conditionals ($\pi^*(\cdot)$) for Gibbs sampling, the likelihood ($\mathscr{L}_j$ for cell-specific likelihood, $\mathscr{L}_i$ for gene-specific likelihood) is multiplied by the relevant prior specifications ($\pi(\cdot)$). q_0 indicates the number of biological genes while q is the number of biological and spike-in genes. L is the number of Gaussian Radial Basis Functions. Λ is a diagonal matrix with elements $(\lambda_1,\ldots,\lambda_{q_0})$ and $Y=(\log(\delta_1),\ldots,\log(\delta_{q_0}))'$.

Helper functions

For simplicity, the distributions of the joint prior specification, the product of this distribution across all biological genes q_0 and the multivariate Normal distribution for the prior on β take the form:

$$\text{log-Normal}(\text{f}(\mu_i), \frac{\sigma^2}{\lambda_i}) \propto (\frac{\lambda_i}{\sigma^2})^{\frac{1}{2}} \exp\{-\frac{\lambda_i}{2\sigma^2}(\log(\delta_i) - x_{i,*}^T\beta)^2\}$$

$$\prod_{i=1}^{q_o} \text{log-Normal}(\text{f}(\mu_i), \frac{\sigma^2}{\lambda_i}) \propto (\frac{1}{\sigma^2})^{\frac{q_0}{2}} (\prod_{i=1}^{q_0} \lambda_i)^{\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}[(Y - X\beta)^T \Lambda (Y - X\beta)$$

$$\text{Normal}(m_\beta, \sigma^2 V_\beta) \propto (\frac{1}{\sigma^2})^{\frac{L+2}{2}} \exp\{-\frac{1}{2\sigma^2}(\beta - m_\beta)^T V_\beta^{-1}(\beta - m_\beta)\}$$

Here, $x_{i,*}^T$ is the transposed vector of the ith row in the model matrix X.

Full conditionals

Full conditional for μ_i across all cells:

$$\begin{aligned}
\pi^*(\mu_i|\cdot) &\propto \mathscr{L}_i \times \pi(\mu_i) \times \pi(\delta_i|\mu_i, \beta, \sigma^2, \eta) \\
&\propto \left[\prod_{j=1}^{n} \left(\frac{1}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}} \left(\frac{\mu_i}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{x_{ij}}\right] \times \exp\left(-\frac{(\log(\mu_i) - 0)^2}{2a_\mu^2}\right) \frac{1}{\mu_i} \\
&\times \exp\left\{-\frac{\lambda_i}{2\sigma^2}(\log(\delta_i) - f(\mu_i))^2\right\} \\
&\propto \left[\prod_{j=1}^{n} \frac{(\mu_i)^{x_{ij}}}{(\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i} + x_{ij}}}\right] \times \exp\left(-\frac{(\log(\mu_i))^2}{2a_\mu^2} - \frac{\lambda_i(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2}\right) \frac{1}{\mu_i} \\
&\propto \frac{\mu_i^{\sum_{j=1}^{n} x_{ij}}}{\prod_{j=1}^{n} (\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i} + x_{ij}}} \times \exp\left(-\frac{(\log(\mu_i))^2}{2a_\mu^2} - \frac{\lambda_i(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2}\right) \frac{1}{\mu_i}
\end{aligned}$$

Full conditional for δ_i across all cells:

$$\begin{aligned}
\pi^*(\delta_i|\cdot) &\propto \mathscr{L}_i \times \pi(\delta_i|\mu_i,\beta,\sigma^2,\eta) \\
&\propto \left[\prod_{j=1}^{n} \frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})} \left(\frac{\frac{1}{\delta_i}}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}} \left(\frac{1}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{x_{ij}}\right] \\
&\times \exp\left\{-\frac{\lambda_i(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2}\right\} \frac{1}{\delta_i} \\
&\propto \left[\prod_{j=1}^{n} \frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})} \frac{(\frac{1}{\delta_i})^{\frac{1}{\delta_i}}}{(\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i}+x_{ij}}}\right] \times \exp\left\{-\frac{\lambda_i(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2}\right\} \frac{1}{\delta_i}
\end{aligned}$$

Full conditional for β across all cells and genes:

$$\begin{aligned}
\pi^*(\beta|\cdot) &\propto \mathscr{L} \times \pi(\delta|\mu,\beta,\sigma^2,\eta) \times \pi(\beta) \\
&\propto \exp\{-\frac{1}{2\sigma^2}[(Y - X\beta)'\Lambda(Y - X\beta)\} \times \\
&\exp\{-\frac{1}{2\sigma^2}(\beta - m_\beta)' V_\beta^{-1}(\beta - m_\beta)\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}[(Y - X\beta)'\Lambda(Y - X\beta) + (\beta - m_\beta)' V_\beta^{-1}(\beta - m_\beta)]\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}[Y'\Lambda Y - 2(X\beta)'\Lambda Y + (X\beta)'\Lambda X\beta \\
&+ \beta' V_\beta^{-1}\beta - 2m_\beta' V_\beta^{-1}\beta + m_\beta V_\beta^{-1} m_\beta]\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}[\beta' X'\Lambda X\beta + \beta' V_\beta^{-1}\beta - 2X'\Lambda Y\beta - 2V_\beta^{-1} m_\beta \beta]\} \\
&\propto \exp\{-\frac{1}{2\sigma^2}[\beta'(X'\Lambda X + V_\beta^{-1})\beta - 2(X'\Lambda Y + V_\beta^{-1} m_\beta)\beta]\} \\
&\propto N(m_\beta^*, \sigma^2 V_\beta^*)
\end{aligned}$$

With

$$\begin{aligned}
V_\beta^* &= (X'\Lambda X + V_\beta^{-1})^{-1} \\
m_\beta^* &= (X'\Lambda X + V_\beta^{-1})^{-1}(X'\Lambda Y + V_\beta^{-1} m_\beta)
\end{aligned}$$

Full conditional for λ_i across all cells:

$$\begin{aligned}
\pi^*(\lambda_i|\cdot) &\propto \mathscr{L}_i \times \pi(\delta_i|\mu,\beta,\sigma^2,\eta) \times \pi(\lambda_i) \\
&\propto \lambda_i^{1/2} \exp\{-\frac{\lambda_i}{2\sigma^2}(\log(\delta_i) - f(\mu_i))^2\} \cdot \lambda_i^{\frac{\eta}{2}-1} \exp(-\lambda_i \frac{\eta}{2}) \\
&\propto \lambda_i^{\frac{\eta+1}{2}-1} \exp\{-\frac{\lambda_i}{2}(\eta + \frac{1}{\sigma^2}(\log(\delta_i)_i - f(\mu_i)))\} \\
&\propto \text{Gamma}(a^*_\lambda, b^*_\lambda)
\end{aligned}$$

With

$$\begin{aligned}
a^*_\lambda &= \frac{\eta+1}{2} \\
b^*_\lambda &= \frac{1}{2}\left[\frac{1}{\sigma^2}(\log(\delta_i) - f(\mu_i))^2 + \eta\right]
\end{aligned}$$

Full conditional for σ^2 across all cells and genes:

$$\begin{aligned}
\pi^*(\sigma^2|\cdot) &\propto \mathscr{L} \times \pi(\delta|\mu,\beta,\sigma^2,\eta) \times \pi(\sigma^2) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{q_0}{2}} \exp\{-\frac{1}{2\sigma^2}(Y - X\beta)'\Lambda(Y - X\beta)\} \\
&\cdot \left(\frac{1}{\sigma^2}\right)^{L+2} \exp\{-\frac{1}{2\sigma^2}(\beta - m_\beta)'V_\beta^{-1}(\beta - m_\beta)\} \\
&\cdot \left(\frac{1}{\sigma^2}\right)^{a_{\sigma^2}+1} \exp\{-\frac{b_{\sigma^2}}{\sigma^2}\} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{q_0+L+2}{2}+a_{\sigma^2}+1} \exp\{-\frac{1}{\sigma^2}[b_{\sigma^2} + \frac{1}{2}[(Y - X\beta)'\Lambda(Y - X\beta) \\
&+ (\beta - m_\beta)'V_\beta^{-1}(\beta - m_\beta)]]\}
\end{aligned}$$

After completing the squares

$$\begin{aligned}
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{q_0+L+2}{2}+a_{\sigma^2}+1} \exp\{-\frac{1}{\sigma^2}[b_{\sigma^2} + \frac{1}{2}(Y'\Lambda Y + m_\beta{}'V_\beta{}^{-1}m_\beta \\
&+ (\beta - m^*_\beta)'(V^*_\beta)^{-1}(\beta - m^*_\beta) - (m^*_\beta)'(V^*_\beta)^{-1}m^*_\beta)]\} \\
&\propto (\frac{1}{\sigma^2})^{a_{n,\sigma^2}+1} \exp(-\frac{b^*_{\sigma^2}}{\sigma^2}) \\
&\propto \text{Inv-Gamma}(a^*_{\sigma^2}, b^*_{\sigma^2})
\end{aligned}$$

With

$$a^*_{\sigma^2} = \frac{q_0 + L + 2}{2} + a_{\sigma^2}$$

$$b^*_{\sigma^2} = b_{\sigma^2} + \frac{1}{2}(Y'\Lambda Y + m'_\beta V_\beta^{-1} m_\beta + (\beta - m^*_\beta)'(V^*_\beta)^{-1}(\beta - m^*_\beta) - (m^*_\beta)'(V^*_\beta)^{-1} m^*_\beta)$$

$$\equiv b_{\sigma^2} + \frac{1}{2}(Y'\Lambda Y + m'_\beta V_\beta^{-1} m_\beta + \beta'(V^*_\beta)^{-1}\beta - 2\beta'(V^*_\beta)^{-1} m^*_\beta)$$

Full conditional for s_j across all genes:

$$\begin{aligned}\pi^*(s_j|\cdot) &\propto \mathscr{L}_j \times \pi(s_j) \\ &\propto {s_j}^{a_s - 1} \exp\{-b_s s_j\} {s_j}^{-\frac{1}{\theta}} \exp\{-\frac{\nu_j}{s_j \theta}\} \\ &\propto {s_j}^{a_s - \frac{1}{\theta} - 1} \exp\{-\frac{\nu_j}{s_j \theta} - b_s s_j\}\end{aligned}$$

Full conditional for ϕ across all genes and cells:

$$\begin{aligned}\pi^*(\phi_j|\cdot) &\propto \mathscr{L}_j \times \pi(\phi_j) \\ &\propto \prod_{i=1}^{q_0}\prod_{j=1}^{n}\left(\frac{1}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\phi_j}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{x_{ij}} \times \pi(\phi_j) \\ &\propto \frac{\prod_{i=1}^{q_0}\prod_{j=1}^{n} {\phi_j}^{x_{ij}}}{\prod_{i=1}^{q_0}\prod_{j=1}^{n}(\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i} + x_{ij}}} \times \pi(\phi_j) \\ &\propto \frac{\prod_{i=1}^{q_0} {\phi_j}^{\sum_{j=1}^{n} x_{ij}}}{\prod_{i=1}^{q_0}\prod_{j=1}^{n}(\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i} + x_{ij}}} \times \pi(\phi_j)\end{aligned}$$

Full conditional for ν_j across all genes:

$$\begin{aligned}\pi^*(\nu_j|\cdot) &\propto \mathscr{L}_j \times \pi(\nu_j) \\ &\propto \left[\prod_{i=1}^{q_0}\left(\frac{1}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\nu_j}{\phi_j \nu_j \mu_i + \frac{1}{\delta_i}}\right)^{x_{ij}}\right]\left[\prod_{i=q_0+1}^{q} {\nu_j}^{x_{ij}} \exp\{-\nu_j \mu_i\}\right] \\ &\times {\nu_j}^{\frac{1}{\theta} - 1} \exp\{-\frac{\nu_j}{s_j \theta}\}\end{aligned}$$

Full conditional for θ across all genes and cells:

$$
\begin{aligned}
\pi^*(\theta|\cdot) &\propto \mathscr{L} \times \pi(\theta) \\
&\propto \left[\prod_{j=1}^{n} \frac{(s_j\theta)^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})} v_j^{\frac{1}{\theta}-1} \exp\{-\frac{v_j}{s_j\theta}\}\right] \times \theta^{a_\theta-1}\exp\{-b_\theta\theta\} \\
&\propto \left[\prod_{j=1}^{n} \frac{(s_j\theta)^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})} \frac{1}{v_j}^{-\frac{1}{\theta}} \exp\{-\frac{v_j}{s_j\theta}\}\right] \times \theta^{a_\theta-1}\exp\{-b_\theta\theta\} \\
&\propto \left[\prod_{j=1}^{n} \frac{\frac{s_j}{v_j}^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})} \theta^{-\frac{1}{\theta}} \exp\{-\frac{v_j}{s_j\theta}\}\right] \times \theta^{a_\theta-1}\exp\{-b_\theta\theta\} \\
&\propto \frac{\left(\prod_{j=1}^{n} \frac{s_j}{v_j}\right)^{-\frac{1}{\theta}}}{\Gamma^n(\frac{1}{\theta})} \theta^{-\frac{n}{\theta}} \exp\{-\frac{1}{\theta}\sum_{j=1}^{n}\frac{v_j}{s_j}\}\theta^{a_\theta-1}\exp\{-b_\theta\theta\} \\
&\propto \frac{\left(\prod_{j=1}^{n} \frac{s_j}{v_j}\right)^{-\frac{1}{\theta}}}{\Gamma^n(\frac{1}{\theta})} \theta^{a_\theta-\frac{n}{\theta}-1} \exp\{-\frac{1}{\theta}\sum_{j=1}^{n}\frac{v_j}{s_j} - b_\theta\theta\}
\end{aligned}
$$