Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht-Karls-University
Heidelberg

Presented by
M.Sc. Jasper Panten
born in Kiel, Germany
Oral examination: 18.03.2024

Causes and consequences of context-specific allelic imbalance

Referees:
Prof. Oliver Stegle
Prof. Duncan Odom

Summary

Gene expression has to be regulated in a cell type-specific manner to ensure proper functionality of cell types and tissues. In a diploid organism, the two alleles of a gene can be regulated independently, causing differential contribution to mRNA levels and thus to cellular function. Allelic imbalance in gene expression has long been recognized as a contributor to cellular phenotypes, however, it is not well understood how and to which extent allelic imbalance is shaped by the regulatory environment of different cell types. Recent advances in single-cell technologies provide the opportunity to profile gene expression and its regulation in a cell type-specific manner at scale, extending our understanding of genome function.

In this thesis, I performed a comprehensive analysis of allele-specific expression (ASE) at single-cell resolution in interspecific mouse hybrids. I first analysed the differentiation-dependence of allelic imbalance caused by strain-specific genetic effects during murine spermatogenesis. This analysis shows that across cell types, variation in ASE is extremely pervasive. Using an F1 trio design, I further separated *cis*- and *trans*-contributions to gene expression divergence and showed that cell type-specific action of regulatory variants is mainly driven by the interaction of *cis*-effects with the cellular environment. Finally, I investigated the contribution of dynamic genetic effects to cell type-specific transcriptional evolution.

Next, I focussed on ASE caused by an epigenetic mechanism, namely X-chromosome inactivation (XCI). In female humans and mice, XCI causes mosaic haplotype-specific expression of X-linked genes, and escape from XCI can lead to increased gene dosage compared to males. Using single-cell genomics assays, I developed an analysis approach to distinguish active and inactive X-chromosomes in individual cells, which allowed me to identify cell type-specific escape. I further showed that T-cell expansion during ageing leads to globally impaired silencing of the inactive X which is associated with an exhaustion phenotype. These findings replicated on the level of chromatin accessibility, demonstrating that variation in escape is associated with an active chromatin state. Collectively, I showed that escape can vary at the cell type level and during organismal ageing.

While these results show that escape from XCI is plastic, they do not address how it might be regulated in different cell types. In the final chapter, I therefore explored whether the *Xist* long non-coding RNA can regulate escapee expression. Using allele-specific RNA-Seq data, I showed that increased *Xist*-levels lead to almost complete silencing of escapees in neural progenitor cells. Modelling of silencing trajectories showed substantial variability among genes in both their resistance to silencing as well as their reversibility, suggesting that escape is genomically encoded. Finally, I demonstrated that over-expression of *Xist* leads to escapee silencing in early embryogenesis. These results provide a potential mechanism that might drive variability in expression from the inactive X.

Taken together, this thesis delineates to which extent allelic imbalance is driven by cell type-specific regulatory environments and suggests analysis approaches for allele-resolved single-cell data. This provides the basis for a comprehensive survey of allelic usage *in vivo* and the molecular mechanisms causing its context-specificity.

Zusammenfassung

Um die Funktionalität von Zellen und Organen zu gewährleisten, muss Genexpression auf Zelltyp-spezifische Weise reguliert werden. In diploiden Organismen können die zwei Allele eines Gens unabhängig voneinander reguliert werden und so differenziell zu mRNA-Levels und zellulären Funktionen beitragen. Es ist bekannt, dass Allelspezifität in der Genexpression zelluläre Phänotypen beeinflussen kann, aber zu welchem Maße sie von Zelltyp-spezifischer Genregulation beeinflusst wird, ist nicht genau verstanden. Moderne einzelzell-aufgelöste Sequenzierungstechnologien bieten nun die Möglichkeit Genexpression und seine Regulation genomweit in verschieden Zelltypen zu untersuchen und ermöglichen so Einblicke in die Funktionsweise von Genomen.

In dieser Arbeit führe ich eine umfassende Analyse von allelspezifischer Expression (ASE) in individuellen Zellen inter-spezifischer Maushybride durch. Ich untersuche zuerst inwieweit ASE, die durch spezies-spezifische genetische Effekte verursacht wird, während der Spermatogenese variiert. Diese Experimente zeigen, dass Variabilität von ASE zwischen Zelltypen ein extrem häufiges Phänomen ist. Durch ein F1-Trio-Experiment kann ich zudem durch *cis*- und *trans*-Effekte verursachte Änderungen der Genexpression trennen und so zeigen, dass regulatorische Varianten, die Zelltyp-spezifisch aktiv sind, vor allem durch Interaktionen von *cis*-Effekten mit regulatorischen Netzwerken wirken. Zuletzt untersuche ich die Wirkung von dynamischen genetischen Effekten auf die Zelltyp-spezifische Evolution von Transkriptionsprofilen.

Als nächstes fokussiere ich mich auf ASE, die durch epigenetische Mechanismen hervorgerufen wird, speziell durch X-Chromosom-Inaktivierung (XCI). In weiblichen Menschen und Mäusen führt XCI zu mosaisch haplotyp-spezifischer Expression von X-chromosomalen Genen, aber manche Gene entgehen dieser Inaktivierung (escape), was zu einer erhöhten Expressionsdosis gegenüber männlichen Zellen führt. Unter Verwendung von einzelzell-aufgelösten Sequenzierungsmethoden entwickle ich ein Verfahren, um aktive und inaktive X-chromosomen in einzelnen Zellen zu identifizieren, und so Zelltyp-spezifische Escapees zu finden. Zudem zeige ich, dass die Expansion von T-Zellen im alten Immunsystem zu global beeinträchtigter XCI führt. Diese Ergebnisse bestätige ich durch eine Analyse der Chromatinoffenheit in den selben Zelltypen und zeige, dass escape mit einem aktiven Chromatinstatus einhergeht. Zusammenfassend zeige ich, dass escape zwischen Zelltypen und während des Alterns eines Organismus variieren kann.

Obwohl diese Ergebnisse eine gewisse Plastizität beim escape von der XCI zeigen, erklären sie nicht, wie dies in verschiedenen Zelltypen reguliert werden könnte. Im letzten Kapitel untersuche ich daher ob die nicht-kodierende RNA *Xist* die Expression von escapees reguliert. Unter Verwendung von allelspezifischen RNA-Sequenzierungsdaten zeige ich, dass erhöhte *Xist*-Levels escape in neuralen Progenitorzellen fast vollständig verhindern. Ein statistisches Modell des Inaktivierungsverlaufs zeigt, dass Gene variabel auf erhöhte *Xist*-Level reagieren, und dass diese Inaktivierung zu unterschiedlichen Graden reversibel ist. Zuletzt zeige ich, dass *Xist* escape auch *in vivo* während der frühen Embryonalentwicklung reduzieren kann. Diese Daten zeigen einen potentiellen Mechanismus, der Variabilität beim escape von der XCI erklärt.

Im Ganzen zeigt diese Arbeit, inwieweit ASE durch Zelltyp-spezifische Genregulation hervorgerufen werden kann und demonstriert Verfahren zur statistischen Analyse von Einzelzelldaten mit allelspezifischer Auflösung. Damit lege ich die Grundlagen für eine vollständige Analyse von ASE *in vivo* und den molekularen Mechanismen, die diese Kontextspezifität verursachen.

Acknowledgements

Nothing in science happens because of a single person and none of this work would have been possible if not for the efforts of many people.

My first thanks go to my supervisors Duncan Odom and Oliver Stegle. Duncan picked me up during the PhD interviews and placed me between two fantastic labs. I am grateful for the trust both of them put and continue to put in me, for their encouragement to follow my ideas and for always pushing me to try and do better. So much I learned during my PhD - from mouse genetics and evolution and statistical modelling to the peculiar art of scientific publishing, I owe to them.

I've been fortunate to work with amazing scientists during my PhD. I am deeply grateful to Stefania for letting me join her on the immune cell sex bias project, which has been the most fun science I did during my PhD, and for reminding me that you never know! I want to thank Paul for being an absolute scientific role model and for always being there to discuss and ask the tough questions. I've learned a ton from you and you're always there for everyone (although discussing ageing with you is getting a bit old... get it??).

I want to thank everyone who's made Odom lab 2.5 the great place it is, and especially Meike, Anja, Marie, Nina and Maja for endless help with mice, experiments and 10x libraries, to Fritjof and Raphael for weekly updates on metadata tables and to Amy, Mika, Lea, Perrine and Klaudija for discussions and coffee breaks. Thanks to James and Lauren for chatting science and the teamwork during the 4CM, and thanks to Lilla, Johanna and Sabine for helping to navigate DKFZ administrative quagmires. Another big thanks goes to Eleonore for the great work you did during your masters despite my inept teaching, which has taught me a lot as well. I'm grateful to the Steglians, especially Tobi and Florin for the work we've done together – it's been great to talk science (even though I mainly show up for retreats).

I am very grateful for my semi-adoption into the Heard Lab, and especially the great collaboration with Agnese and Antonia. The work we did was a great crash course into XCI and epigenetics and it's been great to be on a project where everything always works!! ☺ I'm thankful to Edith Heard for everything I've learned during our meetings.

I want to thank Judith Zaugg for being on my TAC alongside Duncan and Oli and for her enthusiasm and input whenever we meet. Further thanks go to Henrik Kaessmann and Justin Crocker for completing my defense comittee, I hope you will enjoy it. Thanks go to Stefania and Becky for their comments on this thesis.

I'm thankful to all the DKFZ core facilities and services which let us do the work in the first place, in particular the sequencing open lab for game-changing overnight sequencing and the animal caretakers. I thank everyone who is making the DKFZ a better working place and the DKFZ and BMBF for funding my work.

I would not be where I am today if I hadn't worked with Marc during my masters. His enthusiasm for science is infectious to the point of glueing yourself to the cell culture bench for eight hours a day (to be fair, I don't particularly miss that...) and I've learned so much during our time together. Too good to be true! Thanks also to everyone else I had the joy of working

with during my Masters, and to Simon and Vera for the great work we've continued to do.

I want to thank all my friends for fun times and being there for me during the last years and before. Thanks to Mika, Sylvain, Stef and Martino for after work beers and christmas market visists. Thanks to my great friend and previous colleague Pablo, as well as Isa, for dinners, escape rooms and evenings that end much later than expected. Thanks to Vera, Domi, Flo and everyone else from the HI-STEM years. Finally, thanks to the Mobis Jakob, Jens, Jonathan, Julia, Tanja and David, for years of friendship and for always managing to reunite.
I'm especially grateful to Simon, for 20 years of friendship and always having my back, through school, uni, living together and everything else. I owe you so much.

Finally, Beck, thank you so much for sticking with me through the whole four years, I couldn't imagine having done it without you. You're making me a better person every day and I'm infinitely grateful for your love, patience and positivity. I'm so looking forward to our next adventure!

Zuletzt, vielen Dank an meine Familie, und vor allem meine Eltern. Ihr habt mich immer bei allem unterstützt und ich wäre ohne euch niemals wo ich jetzt bin.

Contents

List of Figures

Acronyms

DNA	desoxyribonucleic acid
RNA	ribonucleic acid
H2A, H3	histone 2A, 3
H3K4me3	H3 Lysin 4 tri-methylation
H3K27me3	H3 Lysin 27 tri-methylation
H2Ak119Ub	H2A Lysin 119 ubiquitination
Pol II	RNA polymerase II
TAD	topologically assoiated domain
DNAme	DNA methylation
TF	transcription factor
TFBS	transcription factor binding site
SNV	single nucleotide variant
GWAS	genome wide association study
F1, F2	filial generations 1, 2
bp	basepairs
QTL, eQTL, pQTL, meQTL	quantitative trait locus (expression-/protein-/methylation-QTL)
ICR	imprinting control region
DMR	differentially methylated region
CRE	*cis*-regulatory element
XCI	X-chromosome inactivation
PAR	pseudo-autosomal region
RAME	random monoallelic expression
scRNA-Seq	single-cell RNA-sequencing
scATAC-Seq	single-cell ATAC-sequencing
FACS	fluorescence-assisted cell sorting
RAMA	random monoallelic accessibility
AI	allelic Imbalance
ASE	allele-specific expression
ATAC	assay for transposase-accessible chromatin
PCR	polymerase chain reaction
tRNA	transfer-RNA
FISH	fluorescence *in situ* hybridization
GP	gaussian process
RBF	radial basis function
LMM	linear mixed model
LM	linear model
gLM	generalized linear model
gLMM	generalized linear mixed model
scDALI	single-cell differential allelic imbalance
FDR	false-discovery rate
ANOVA	analysis of variance
SNP	single-nucleotide polymorphism
MNN	mutual nearest neighbours
UMAP	uniform manifold approximation and projection
smFISH	single-molecule FISH
cDNA	complementary DNA
ELBO	evidence lower bound

VGP	variational gaussian process
AUC	area under the curve
CHIP-Seq	chromatin immunoprecipitation followed by sequencing
BF	Bayes factor
GE	gene expression
CA	chromatin accessibility
PBMCs	peripheral blood mononuclear cells
tSNE	t-distributed stochastic neighbour embedding
OR	odds ratio
UMI	unique molecular identifier
BIC	bayesian information criterion
SC	spermatocyte
RS	spermatid
ES	elongating spermatid
SG	spermatogonium
LOESS	locally estimated scatterplot smoothing
H0/H1	null hypothesis / alternative hypothesis
MACS	magnetic cell separation
MZ	marginal zone
GC	germinal center
DC	dendritic cell
NPC	neural progenitor cell
RTTA	reverse-targeted tetracyclin transactivator
CRISPR	clustered regularly interspaced short palindromic repeats

Chapter 1

Introduction

1.1 Regulation of eukaryotic gene expression in *cis* and *trans*

The ”central dogma” of molecular biology states that the DNA sequence encodes instructions for the synthesis of proteins encoded through an RNA intermediate [Crick, 1970][1]. However, in multicellular organisms, all cells contain (approximately) the same genetic information, even though their protein composition can be vastly different. Since the formulation of this central dogma, molecular biology has uncovered a remarkable diversity of strategies at which the transcription of genomic DNA to RNA and the translation of mRNA into proteins is regulated, allowing cells to express the appropriate set of genes necessary to perform cell type-specific functions, respond to stressors, self-replicate and to assemble full multi-cellular organisms [ENCODE Project Consortium, 2012]. In the following I am going to outline the role of *cis*-regulatory elements within the DNA and *trans*-acting factors that act on them to ensure timely regulation of gene expression **(Figure 1.1)**.

1.1.1 *Cis*-regulatory sequences and epigenomic state

Only 1.5% of the human and comparable mammalian genomes is coding for known proteins [Lander et al., 2001]. While a large fraction of the remainder is composed of non-functional repeat elements, it is thought that it also includes many regulatory elements that allow for genes to be transcribed in a context-specific manner [Khodosevich et al., 2002]. Sites of active transcription can generally be identified by a promoter upstream of the transcription start site (TSS) that allows for the primary binding of the transcriptional machinery centered around the RNA polymerase complex (protein coding genes are transcribed by RNA polymerase II, the separate polymerases RNA Pol I, III and IV transcribe the ribosomal, transfer RNAs and small interfering RNAs) [Haberle and Stark, 2018]. While promoters contain characteristic elements thought necessary to initiate transcription (e.g. TATA boxes), much of the context-specific regulation of genes relies on a second class of regulatory elements known as enhancers. These are distal regulators, that is, they regulate transcription of genes which are kilo- mega-bases away [Banerji et al., 1983]. While it is known that genes can have multiple enhancers [Heinz et al., 2015], their relative importance, which has to be carefully assessed by enhancer

[1] The way the central dogma is usually cited is somewhat divorced from its original meaning. It states primarily that information in DNA and RNA are *not* generated based on protein sequence. It is also clear that RNA can indeed code for DNA or itself, for example in viruses.

deletion assays, in regulating expression in a spatiotemporal manner is still largely elusive.

How do enhancers activate and support gene expression? Their presence is closely linked with what is known as an active regulatory state. In the nucleus, DNA is bound to histone proteins as chromatin, which physically allows for transcription only in its uncondensed, active form known as euchromatin. While non-active chromatin (heterochromatin) associates with the nuclear lamina, euchromatin is able to form higher order structures including loops that can bring enhancers into proximity of their target genes. The resulting topologically associating domains (TADs) form hubs that concentrate factors necessary for transcription [Jerkovic and Cavalli, 2021]. Active and inactive chromatin states are furthermore strongly associated with covalent modifications of both the DNA sequence and histone proteins. In the non-active genome and repeat elements, cytosines are marked with 5'-methyl modifications (also known as DNA methylation, DNAme), while its absence marks active regulatory elements [Bell et al., 2011]. Indeed, DNAme at promoters can abrogate transcription, while removal of DNAme ectopically activates repeats [Pappalardo and Barra, 2021, Smith et al., 2020]. CpG islands, stretches of DNA sequence with high GC content, are often found in promoters and were among the first sequence-based features to predict regulatory elements. While DNA methylation undoubtedly plays functional roles in regulation, it is fully absent in some eukaryotes (for example, in yeast and fruit flies) and cells can exist in hypomethylated states, for example embryonic stem cells [Lyko et al., 2000, Deshmukh et al., 2018, Leitch et al., 2013].

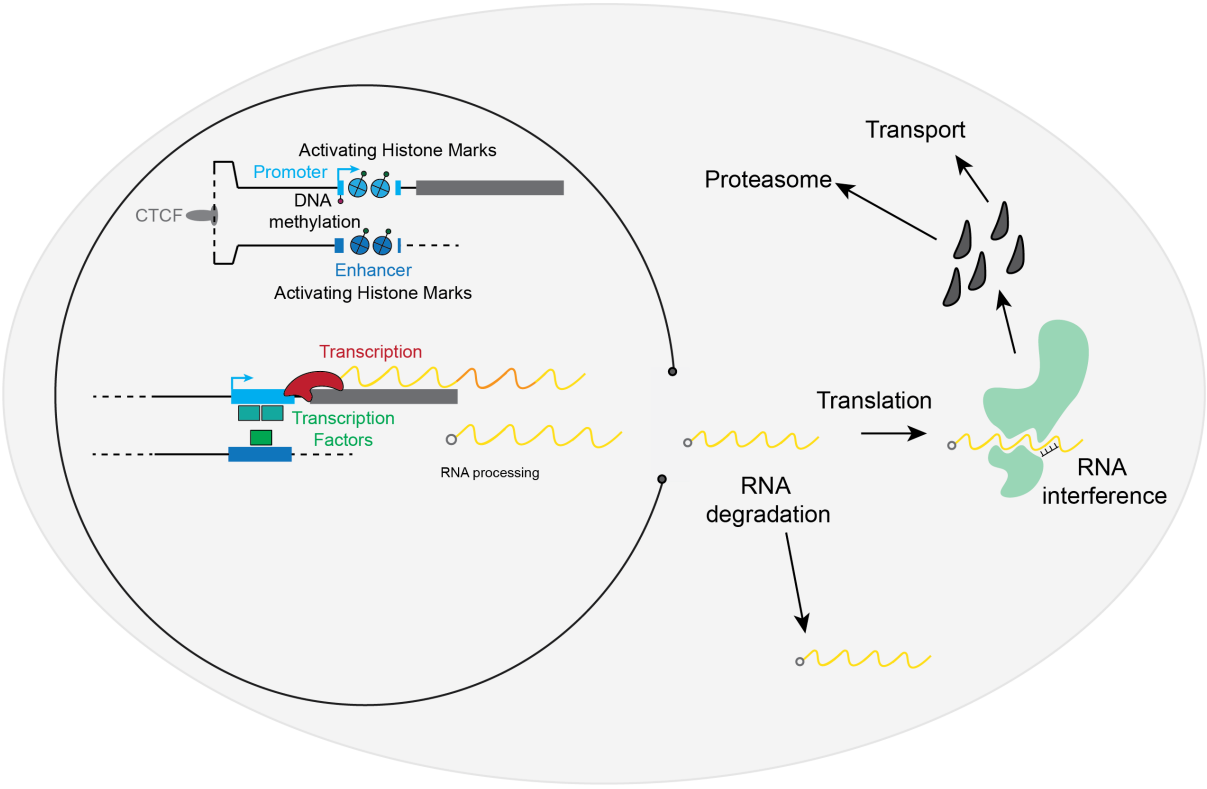

Figure 1.1: Gene regulatory mechanisms in eukaryotes. Broad overview over strategies to control gene expression from regulatory elements in the DNA sequence with epigenetic modifications and three dimensional chromatin structure, transcription-factor driving RNA polymerase activity, RNA processing through splicing, capping and polyA-tailing. Finally, RNA is exported from the nucleus and translated or degraded, and protein levels are further controlled by post-translational modifications, active degradation and intracellular transport.

While DNAme is by far the most common covalent modification of DNA, histones can be modified at many sites and the complex "histone code" marks various classes of genomic elements. Most modifications are present on lysine residues (K) of histone 3 (H3), in particular mono-, di- and tri-methylation of H3K4 and methylation and acetylation of H3K27. These partly antagonistic marks correlate well with the genomic activity of promoters and enhancers and, although the functional impact of some marks is debated, interference with their deposition affects gene expression [Blackledge and Klose, 2021, Lawrence et al., 2016, Berger, 2002]. For example, this is clearly the case for H3K27me3 and H2AK118ub/H2AK119ub which are modified by polycomb repressive complexes which confer gene silencing [Piunti and Shilatifard, 2021].

While the evidence for functionality of promoters and enhancers is clearly the strongest, other classes of regulatory elements in the genome have been described. *Insulators*, mainly associated with the DNA-binding protein CTCF, are thought to abrogate possible interactions between activating regulatory elements and their target genes [Doane and Elemento, 2017]. Similarly, *silencers* are thought to confer silencing by direct interaction with a target genes, but their functional relevance is less clear [Segert et al., 2021].

1.1.2 Regulation of transcription and mRNA processing

While the DNA sequence in regulatory elements encodes information on potential activity, they will only act in combination with *trans*-acting factors that interpret this information and transfer it into an effect on gene expression, so called *transcription factors* (TFs). A simple but surprisingly effective model of transcriptional regulation posits that TFs bind to TF binding sites (TFBS) in enhancers, a specific DNA sequence to which they have a high affinity, and act as bridges to allow for the first steps in the initiation of transcription [Haberle and Stark, 2018]. In this way, the DNA sequence is converted into a binary signal (presence / absence of TFBSs in a regulatory element) which is incomplete, but provides a powerful approximation to construct regulatory networks [Badia-i Mompel et al., 2023]. In reality, TF binding is driven by stereotypic DNA-binding domains, but their affinity is dependent on co-binding with other regulatory factors (cooperativity), DNA methylation, competition with nucleosomes and chromatin accessibility [Spitz and Furlong, 2012]. TFs are thought to orchestrate regulatory processes as diverse as immune reactions and embryonic patterning [Duboule, 2007].

Besides their DNA binding domains, TFs contain effector domains through which they exert their function. Fundamentally, (activating) TFs promote the recruitment of RNA polymerase II through direct interactions or via chromatin remodellers, histone modifying complexes, the mediator complex which conntects TFs to RNA Pol II, or other TFs [Lambert et al., 2018]. Once RNA Pol II is recruited at the promoter, it undergoes a series of tightly regulated steps to generate mRNA molecules (Initiation, pausing, elongation and termination) [Vervoort et al., 2022]. During the process of transcription, the pre-mRNA is spliced to remove intronic sequences and modified with poly-A tails and 5' capping nucleotides [Wilkinson et al., 2020]. Active regulation and dysregulation of these processes lead to different proteoforms through alternatively spliced mRNA and changes in transcript abundance via mRNA stability. Since translation takes place in the cytosol, nuclear retention and mRNA transport provide another opportunity for expression regulation [Das et al., 2021].

In the cytosol, mature mRNA complexes will be translated by ribosomes, where the rates of protein production depend on active regulation of the translation steps, the availability

of tRNAs and mRNA modifications. Ribosome assembly can be specifically inhibited by RNA interference through miRNAs, which also promote mRNA decay [Bartel, 2004]. Finally, protein stability, transport and degradation are likewise heavily modulated, which allows for RNA-independent control of expression [Gebauer and Hentze, 2004, Hershey et al., 2012]. Considering the wealth of regulatory processes, it is central to consider a quantitative description of gene regulation, as regulatory processes are likely to affect all genes, but their relative contribution to expression levels will on the one hand drastically differ and on the other be highly context-specific [Ay and Arnosti, 2011]. An important corollary from this is that mRNA levels, which are often taken as a proxy for gene expression will provide a useful first approximation for protein levels, but can not replace direct measurements [Liu et al., 2016, Carpenter et al., 2014].

1.1.3 Rewiring of regulatory landscapes during cell state transitions

While I was only able to provide a reduced overview of the existing knowledge regarding gene regulatory mechanisms, their general principles are most relevant to this work. Gene regulatory relationships are genetically encoded and interpreted by the *trans*-regulatory environment of the cell. This allows for cell type-specific interpretation of genomes into gene expression, and for cells to respond to exogenous stimuli or to undergo programmed cellular functions like the cell cycle, differentiation or apoptosis. Regulation is multi-layered and it is not always clear which step in the transcriptional process will determine the final expression levels. Also, these mechanisms are usually not on-off switches, but occur in a chemical equilibrium. For example, effective transcription factor binding depends on its affinity to the target sequence in combination with its abundance, and might compete with the binding of a repressor [Neikes et al., 2023]. In this complex dynamic system, the high number of inputs might confer both flexibility and robustness [Macneil and Walhout, 2011]. I am going to use these ideas when discussing gene regulation at the single-allele level, which is the topic of the next section and this thesis in general. Before, that, I am going to briefly discuss cell type transitions in mammalian biology which require rewiring of the regulatory landscape.

The arguably most remarkable changes in cell fate occur during embryonic development, where all cell types of the body derive from a single cell. The fertilized zygote develops into the blastocyst which contains of embryonic stem cells and precursors of extra-embryonic tissues. This represents the first lineage decision stem cells make in the developing embryo. During a step of symmetry breaking, the pluripotent ESCs then commit to either ectodermal, mesodermal or endodermal fate in a process called gastrulation. Patterning signals then instruct cell types to further differentiate into tissue-specific precursors and induce embryonic geometry until organogenesis is completed [Tam and Loebel, 2007]. The permissive chromatin state in ESCs represents their pluripotency and cellular differentiation processes are thought to represent a successive restriction of lineage potential, which accompanies progressive remodelling and inactivation of the regulatory landscape. Pluripotency is maintained by a network of transcription factors (including *Nanog, Pou5f1, Sox2*) which act on enhancers and promoters [Kinoshita and Smith, 2018, Shi et al., 2016]. Interestingly, differentiation-induced genes are though to be maintained in a *poised* chromatin state marked by bivalent domains of both active H3K4me3 and repressive H3K27me3, suggesting that pluripotency is defined by active repression of differentiation **(Fig. 1.2)** [Calo and Wysocka, 2013].

When cells acquire new fates, this is accompanied by a global reorganization of enhancer

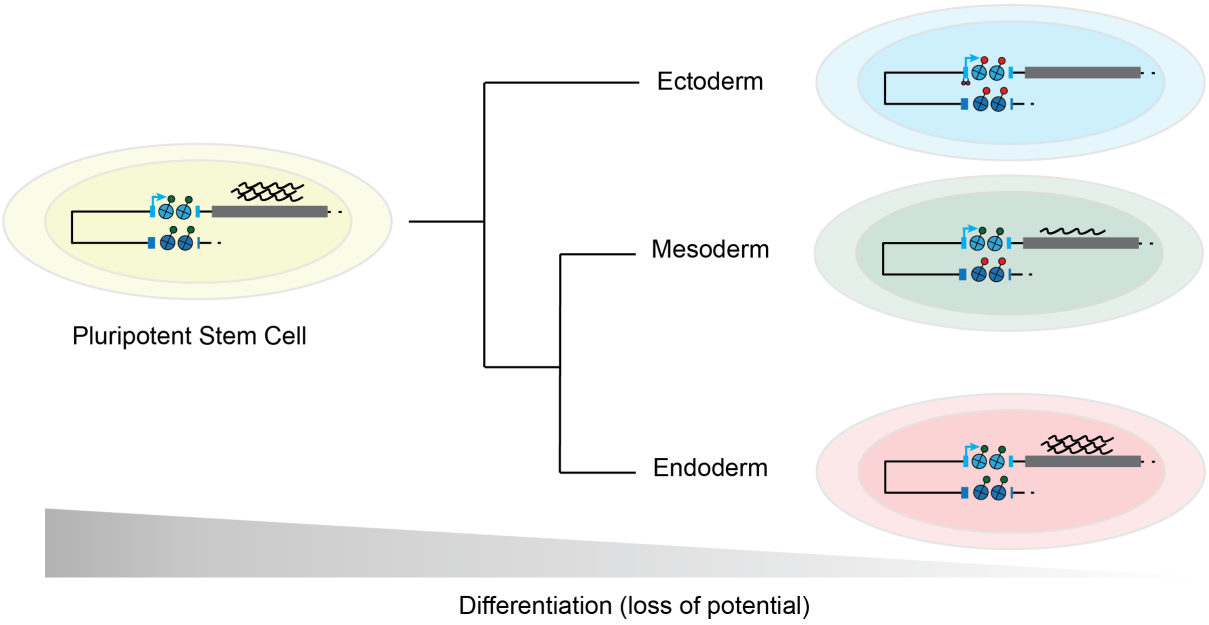

Figure 1.2: The changes in gene regulatory landscapes upon cellular differentiation during development. Pluripotent stem cells differentiate into mature progeny in a step-wise manner, forming first ectoderm and then mesoderm and endoderm. During this process histone modifications switch from being active (green) to inactive (red) at promoters and enhancers in some cell types, in parallel a gain in DNA methylation occurs. This leads to a loss of gene expression, and differentiation potential.

usage, epigenetic marks, three dimensional chromatin structure and expression. The regulation of axial patterning by *Hox* genes can be seen as an example of this rewiring, where multiple transcription factors interact with promoters and enhancers do drive expression of the appropriate variants [Duboule, 2007]. These regulatory hierarchies defined during development are paralleled by adult stem cell systems, where lineage-restricted stem cells differentiate to replenish tissues with fast turnover. While some of these are unidirectional, that is, they produce only one differentiated cell type through a series of intermediates (for example basal cells in the skin or spermatogenesis), others include branching points, where cells decide between alternative fates (hematopoietic cells, gut epithelium). These systems recapitulate many of the molecular features of embryonic development, but are distinct in that they are required to maintain a homeostatic tissue, whereas development moves towards the mature embryo [Crane et al., 2017, Gehart and Clevers, 2019].

Stem cell differentiation represents a cell state transition that spans days or months and radically changes the morphology and function of cells. However, cell state changes can be induced over a range of time scales and with varying degrees of induced change. Examining T-cell differentiation and activation provides an example of different cell state changes, and how they differ from cell type transitions during differentiation **(Fig. 1.3)** [Kumar et al., 2018]. Lymphocytes are generated from hematopoietic stem cells through a series of bifurcating differentiation steps, similar to other mammalian stem cells systems [Cabezas-Wallscheid et al., 2014]. As in embryonic development, this is accompanied by a successive loss of potency, and requires large-scale rearrangement of gene regulation driven by master transcription factors and exemplifies transitions between *cell types*. In the thymus, T-cells then differentiate into sub-celltypes, such as cytotoxic and regulatory classes, which can then be further categorized based on function and marker gene expression

[Henning et al., 2018]. Meanwhile, the activation of T-cells by an antigen-induced stimulus changes their molecular phenotype similarly drastically from a resting *cell state* to proliferation, differentiation and cytokine release. This transition is however naturally reversible and activated T-cells are not usually considered a separate cell type, but rather a different cell state [Martinez-Jimenez et al., 2017]. Similarly, prolonged or repeated antigen contact leads ot T-cell exhaustion in which effector function is lost. Although this is an irreversible transition, it is not clear whether it should be considered a new cell type or a different state of a T-lymphocyte [Kumar et al., 2018]. These examples demonstrate that change in cellular function is complex, and that terms like cell type and state are not fully defined, although cell state changes like cell cycle phase can be considered independent of other transitions.

As illustrated at the example of T-cell exhaustion, cell state changes can also arise from accumulated exposure to stimuli over long periods of time. In immune cells, this can be a result of their tissue environment. For example, macrophages drastically differ in their gene expression programs based on the tissue they are found in, and some of them are considered different cell types (for example Kupffer cells in liver and microglia in the brain, which are possibly seeded at different times in development) [Lavin et al., 2014]. Similarly, accumulated expression changes can be the result of aberrant signalling in diseases such as cancer, or due to organismal ageing. While the pathways and mechanisms driving age-related expression changes are not well understood, it is thought that they result from prolonged exposure to signalling, epigenetic erosion and genomic instability [López-Otín et al., 2013].

In conclusion, cell state changes encompass a variety of fast and slow processes that require a reorganization of gene regulatory mechanisms, while also defining the set of *trans*-acting factors available in a cell. Single-cell genomics methods are generating systems-level data which allow us to move towards a generalizable definition, which I further discuss in **Section 1.3**.

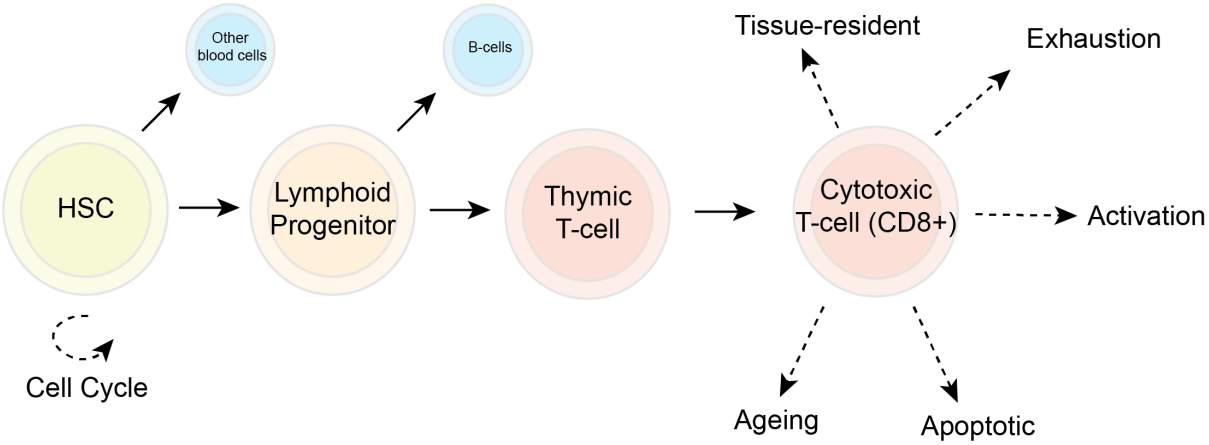

Figure 1.3: Cell type and state changes in T-cells. The T-cell lifecycle demonstrates cell type and state changes. Lymphocytes differentiate from hematopoietic stem cells (HSCs), as all other blood cells. After T-cell commitment in lymphoid cells, double negative T-cells move to the thymus, where they are selected based on their reactivity and become naive T-cells, for example with cytotoxic function. These cells can then acutely adopt different cell states through activation or apoptosis, or gradually through exhaustion, ageing or tissue-specific signalling.

1.2 Causes and consequences of allelic imbalance

In general, the term ”gene expression” is used to refer to a gene as a single unit, however, for most genes in most eukaryotic cells, two near-identical copies of the same DNA are present. This diploid state provides in principle two independent copies of each gene that can be independently regulated. Indeed, this is frequently the case and mainly the result of three causes. First, due to both germ line and somatic variation, the two allelic copies do not carry the exact same sequence **(Fig. 1.4)**. These changes will interact with the regulatory machinery at different levels and slightly (or strongly) alter expression. Second, specific epigenetic mechanisms lead to full or partial silencing of single alleles. This is exemplified by classic genomic imprinting and X-chromosome inactivation. Importantly, this leads to allelic bias in the absence of any sequence variation, although genetic effects will interact with these processes as well.

Both sequence changes as well as epigenetic modifications are inherited through cell division and embryonic development, so these allelic biases will therefore act in every cell of the organism. Third, allelic imbalance can be the result of stochastics in transcriptional regulation. As the two alleles are independent physical entities, regulatory mechanisms are not directly coordinated. At small timescales, transcription is not uniform, but is thought to occur in a bursting manner, which therefore leads to variable allelic bias over time [Cleary and Seoighe, 2021, Robles-Espinoza et al., 2021]. I will now discuss all of these processes in more detail, focussing on genetic and epigenetic heritable allelic variation, which are the foundation of the work in **Chapter 2** and **Chapter 3** and **4** respectively.

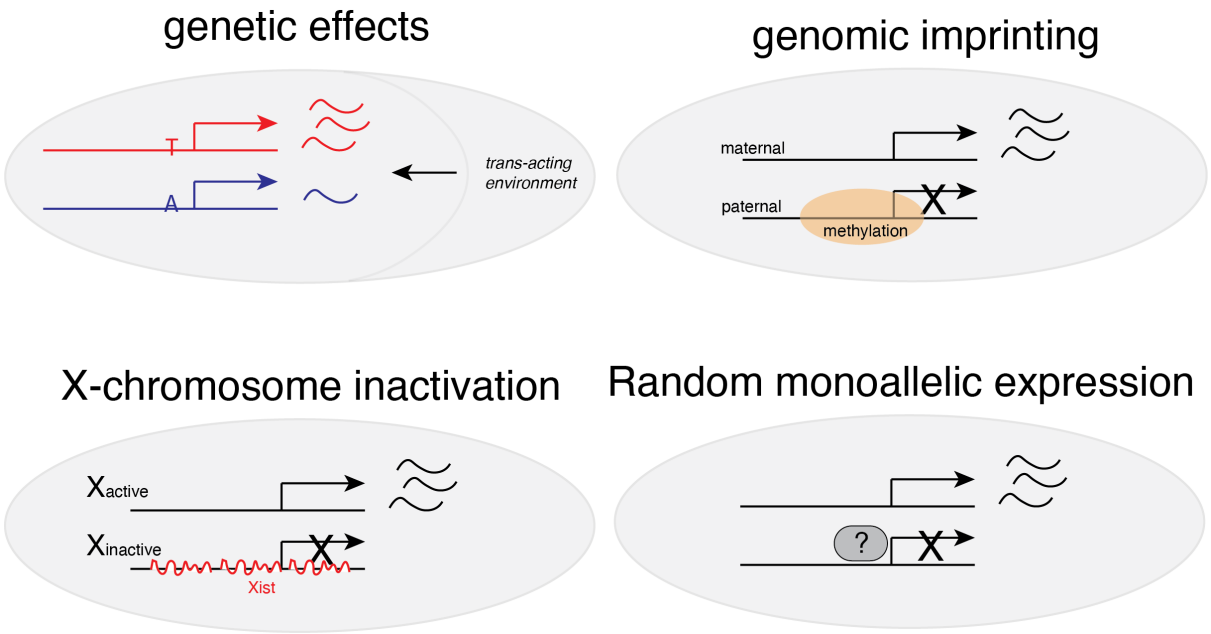

Figure 1.4: Causes of allelic imbalance in diploid cells. Allelic copies are independently expressed through three major mechanisms: First, the presence of regulatory variants that affect transcription or mRNA processing. Second, epigenetic mechanisms, mainly genomic imprinting, in which an allele is inactivated in a parent-of-origin manner, and random inactivation of one X-chromosome in female cells. Third, stochastic transcriptional dynamics lead to unequal allelic usage over time.

1.2.1 Genetic variation causing expression imbalance

A main contributor of allelic imbalance are sequence changes between the two haplotypes, mainly caused by single nucleotide variants (SNVs). Human genomes contain an average of 3 million SNVs (1 per kilobase), approximately half of which are expected to be heterozygous [Albert and Kruglyak, 2015]. As the majority of these are located in non-coding parts of the genome, it is thought that a main driver of phenotypic variation between individuals are the effects of these sequence changes on regulatory elements [Lappalainen and MacArthur, 2021]. The main hypothesis is that the binding affinity of transcription factors or other DNA- and RNA-interacting molecules can be perturbed by single base changes, which then leads to reduced enhancer activity, perturbed splicing or chromatin folding. However, it is generally difficult to predict how a sequence change will mechanistically affect the functionality of a regulatory element, especially when only single bases are affected[2]. Therefore, our understanding of the impact of non-coding variation on disease is still limited to individual variant-gene links (see, for example, [Chatterjee et al., 2016, Weedon et al., 2013, Benko et al., 2009, Sun et al., 2018a]).

A systematic approach to identifying regulatory variation are genome-wide association studies (GWAS). In GWAS, a phenotype of interest, for example a disease status, is measured in a large number of individuals alongside a genome-wide analysis of the genome. As many variants are common in populations and will be shared between multiple individuals statistical analysis, usually linear regression on the number of alternative alleles, can then be used to ask whether there is an association between a specific variant and the trait of interest. Measuring variants genome wide has been made possible through cost-effective whole-genome sequencing and genotyping arrays. In the last 20 years, GWASs with ever increasing sample sizes have been used to identify common and rare genetic risk factors for many diseases [Uffelmann et al., 2021]. Of note, the ability of a GWAS to find risk factors is intrinsically coupled to its sample size: First, GWAS explicitly makes use of inter-individual variation to identify causal variants across many individuals (if a correlation persists across the many ways in which phenotypes differ among individuals, it should be a direct relationship). Second, by computing correlations between potentially millions of variant-trait pairs, any statistical analysis suffers from a large multiple testing burden. Early studies were therefore intrinsically underpowered and the amount of identified links steadily rises with sample size. In the recent release of the UK Biobank datasets (more than 150.000 individuals analyzed) this increase seems to saturate, indicating that ”all”, at least all naturally occuring, genotype-trait associations can be identified for common genotypes given large enough sample sizes [Sudlow et al., 2015]. One major success of these studies is the ability to quantify the fraction of phenotypic variation that is explained by the genetic sequence, and not by their environment, life history or change. This measure is also called the *heritability* of a trait and quantifies the extent to which it is genetically predisposed [Zaitlen and Kraft, 2012].

Naturally, a similar analysis approach can be used to find associations between genotype and molecular phenotypes. These are known as *quantitative trait loci* (QTLs) and traditionally refer to any quantitative (rather than binary) trait, but have become increasingly synonymous with associations measured by ”omics” analysis (also known as molecular QTLs). The specific power of these studies is driven by novel technologies that perform genome-wide molecular

[2] As they are not the focus of this thesis, I am ignoring non-SNV mutations such as larger deletions or translocations here, which are known to have strong effects on allelic gene regulation and can even be oncogenic.

measurements of all genes, for example through transcriptome-wide RNA-Sequencing [GTEx Consortium et al., 2017]. In this way, one can in principle assess the impact of genetic variants on the expression of all genes and in this way, QTL mapping can help to close the gap in mechanistic understanding between a variant and an associated GWAS-trait. Although the most common molecular readout are gene expression levels, QTL studies have been performed using many molecular readouts, indicated by a prepended letter x for (x)QTL-studies, referring to expression (eQTL), protein (pQTL), meQTL (methylation) or others [Ye et al., 2020].

How do genetic variants affect molecular phenotypes? Although sequence changes will perturb gene regulation in many ways, some general principles can be established. Mutations in a coding sequence can lead to amino acid substitutions or frameshift errors, which can render the gene product non-functional. These variants will have the strongest effects, but will be comparatively rare, since only a small fraction of the genome is coding and strong deleterious changes will be selected against [Lappalainen and MacArthur, 2021]. Therefore, the majority of variants are so-called regulatory variation, affecting promoters and enhancers. On a molecular level, this can be the result of changes to the binding sites of transcription factors, proteins that change chromosome conformation, DNA methylation changes and others [Albert and Kruglyak, 2015]. In genes, sequences directing splicing, RNA editing and interactions with the ribosome impact have been shown to affect expression levels **(Fig. 1.5)**.

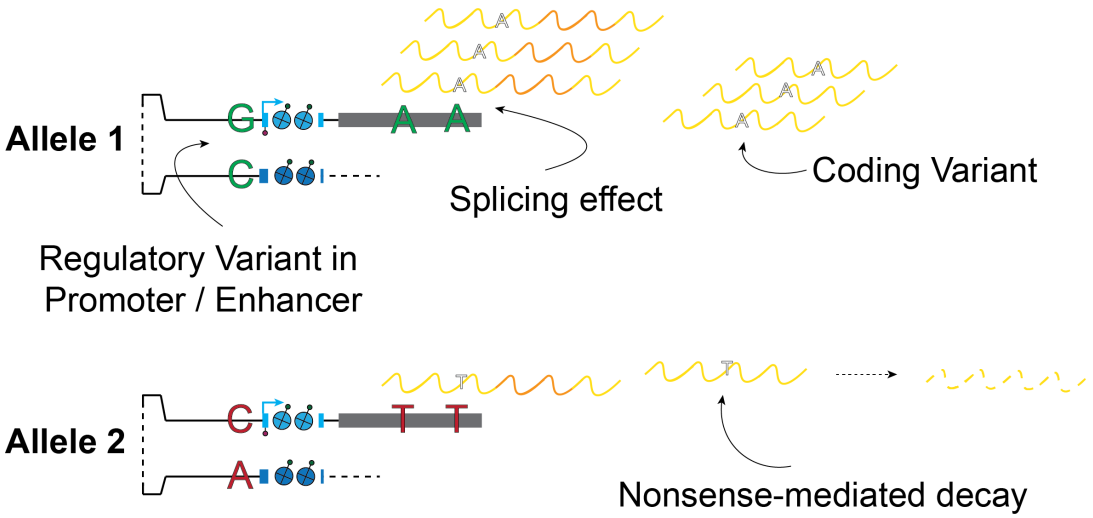

Figure 1.5: Classes of genetic variation affecting gene expression. Schematic providing an overview of the mechanisms genetic changes act through. Coding variants do not necessarily change expression levels, but will impact protein function. In contrast, variants in *cis*-regulatory elements can change transcription directly, regulate splicing or reduce mRNA levels through nonsense-mediated decay.

Although in principle modern eQTL mapping allows to test for associations between all measured genes and variants, the analysis is usually heavily focussed on a gene-centric view, considering only the surrounding 10^6 bp for a gene. This is due to the fact that we assume most genetic variation to be acting locally by affecting regulatory elements within or close to the gene body [Pai et al., 2015]. These variants are also referred to as *cis*-eQTLs, *cis* referring to action within the same chromosomes and with a relatively short distance. The definition of *cis*-eQTLs is analogous to *cis*-regulatory elements such as enhancers and promoters affecting gene expression nearby, and the former are often thought to be the result of mutations in the latter. For eQTL mapping, often variants within a one megabase window around the gene of

interest are considered. While the majority of known variant-expression associations are acting in *cis*, it can also act in textittrans in a potentially genome-wide manner [Võsa et al., 2021]. *Trans*-eQTLs are the result of a variant affecting the protein structure or the expression level of a *trans*-acting factor, which then result in differential expression of the factors target gene, potentially due to regulation of its gene expression, splicing, RNA stability or translation. I note that the analysed variant potentially exerts a *cis*-eQTL on the *trans*-factor, which then drives differential expression of target genes, *trans*-effects therefore inherently aim for a mechanistic understanding of a QTL. Practically, eQTL-testing requires massive sample sizes as the number of association tests for all variants V against all genes G is $V \times G$, scaling quadratically. *Trans*-eQTL testing can be constrained on prior biological knowledge, for example on known target genes of transcription factors and pathway information.

The distinction between *cis*- and *trans*-effects connects the discussion of genetic variation to allele-specific expression [Võsa et al., 2021, Bonder et al., 2021, Signor and Nuzhdin, 2018]. If a variant is heterozygous and *cis*-acting, it can change the allelic balance of expression within an individual. When affecting regulatory elements, this is often mediated through allelic imbalance in histone modification levels, DNA methylation or chromatin accessibility, and these effects can occur without a change in gene expression. Of note, allelic imbalance is a better estimator of *cis*-variant effects, because it measures the variants effect directly and not between individuals, where technical and biological differences might affect the measurement. In particular, allelic imbalance isolates the effect of *cis*-regulatory variants and removes any *trans*-effects, which will be further discussed in **Chapter 2** [Mohammadi et al., 2017].

GWAS, QTL and allelic imbalance analyses are not limited to human populations, but have also been used in model organisms. While analysis in humans has an obvious medical interest, model organisms can provide insight into the (evolutionary) mechanisms of how genetic changes shape phenotypic traits. While analysis in wild populations have been performed, using model organisms allows to control the genetic diversity that is analyzed by using inbred strains. Multi-generation inbreeding leads to organisms that are near-homozygous in the entire genome, therefore an F1 cross of two different inbred strains will be fully heterozygous in all strain-specific variants. In an F2 cross the alleles will be shuffled through meiotic recombination and the offspring will be a combination of homo- and heterozygous alleles. In contrast to mapping in wild populations, the shuffled genetic diversity will be the only determinant of gene expression differences, drastically improving the power to detect genetic effects [Zheng et al., 2011, Schadt et al., 2003, Orozco et al., 2012].

F1 hybrids provide a second, orthogonal opportunity to study genetic effects in *cis* and *trans*. When measuring gene expression, or other molecular traits between two inbred strains, one can keep environmental variables constant, so that as a first approximation, any measured difference should be due to the varying genetics only. In this way, comparing the founder strains ”integrates” over the set of genetic variants affecting the expression of a specific gene and shows the sum of genetic effects on gene expression, without explicitly determining which variants exert these effects. However, in an F1 hybrid, the two alleles are placed in the same *trans*-regulatory environment: This means that any strain-specific *trans*-acting factor will act on both alleles, so any residual difference will only be driven by *cis*-effects[3]. This is a powerful approach to quantitatively dissect total *cis*- and *trans*-contributions as it does not require to explicitely

[3]Of note, the absolute magnitude of both alleles in the F1 might be different to the F0 founders since the *trans*-factors are replaced by a mixture of both parents. It is only the ratio that will be unaffected by *trans*-actors.

nominate interacting variants which might be missed due to insufficient power. Here, only the sequencing depth controls the number of detected effects, which is easier to increase than the number of considered samples. From a QTL analysis point of view, do *cis*- or *trans*-effects play a larger role in driving gene expression differences and disease heritability? As all analysis so far lacks the power to show that it is comprehensive, that is, it has enough statistical power to detect all effects with a relevant effet sizes, this question remains open. *Trans*-effects can impact targets genome-wide, affecting for example all target genes of a transcription factor. In contrast*cis*-effects appear to have stronger effect sizes on their single target genes, potentially because the set of *trans*-regulators is much larger and therefore more redundant. Additionally, the set variants we can observe in standing populations are the result of selective pressure eliminating disadvantageous and lethal variants from the genome. Due to their higher pleiotropy, mutations affecting *trans*-actors are more likely to be deleterious and therefore selected against [Goncalves et al., 2012, Shen et al., 2014, Wittkopp et al., 2004, Halow et al., 2021].

1.2.2 Genomic imprinting

Genomic imprinting is the classical example of allelic imbalance in the absence of genetic variation and of its importance for embryonic development. Imprinting silences genes in a parent-of-origin specific manner, that is, for a given gene, either the maternally or paternally inherited chromosome will be expressed [Ferguson-Smith and Bourc'his, 2018]. This process is driven by the deposition of epigenetic marks, especially DNA methylation in differentially methylated regions (DMRs) during germ cell development, and therefore differs between the two sexes [Li and Sasaki, 2011, Reik and Walter, 2001]. During embryonic development, allelic differences in imprinted regions are erased and genes are either expressed or silenced fully, such that the combination of both alleles in the zygote will represent haploid expression [Ferguson-Smith, 2011]. While controversial for a time, it is now generally accepted that the number of affected loci is likely small, encompassing 100-200 genes within imprinting control regions (ICR), most famously the Callipyge locus (*Meg3*, only expressed from the maternal haplotype) and the *Igf2r* locus (*Slc22a3, Igf2, Mas1*) [Ferguson-Smith and Bourc'his, 2018]. Strikingly, imprinting is required to complete embryonic development, providing an example of non-genetic allelic imbalance that is essential for organismal viability. Furthermore, mutations in imprinted loci inherited from the non-imprinted parental allele will exert dominant effects, as the non-mutated allele is silenced. For example, paternal deletions in the human ICR of 15q11-13 causes Prader-Willi-syndrome, Beckwith-Wiedemann is caused by a loss of imprinting of *Igf2* and loss of maternal *Ube3a* causes Angelman syndrome [Monk et al., 2019]. In summary, genomic imprinting can drive an all-or-nothing allelic imbalance for many genes that is dictated by the parent of origin and can have drastic phenotypic consequences. While the classic imprinted loci are fully biased towards one allele, a large number is under partially imprinted control, although the epigenetic basis for this is still largely unclear [Goncalves et al., 2012].

As the imprints are established in the earliest stages of development and are passed on through mitosis, it is generally thought that imprinting varies little across somatic cells [Baran et al., 2015, Latham, 1995]. Anecdotal examples for tissue-specific imprinting exist especially in the brain where, likely due to neuron-specific effects [Laukoter et al., 2020], imprinting has important functions throughout organismal life [Wilkinson et al., 2007]. For example, *Ube3a* is a brain-specific imprinted gene [Yamasaki et al., 2003], while *Dlk1* loses its mono-allelic expression to support neural stem cell function [Ferrón et al., 2011]. Other exam-

ples of context-specificity include *Kcnq1* in the pancreas development [Travers et al., 2013]. Of note, tissue-specific genomic imprinting is pervasive in the developing trophoblast and in the placenta, and is required for its function supporting fetal growth and might be linked to its particular expression of endogenous retroviruses [Hanna, 2020]. Meanwhile, a comprehensive assessment of celltype-specific imprinting and its heterogeneity across single cells is still lacking. Preliminary work has demonstrated the possibility of using scRNA-Seq to identify imprinted genes [Santoni et al., 2017, Laukoter et al., 2020, Andergassen et al., 2017]. These studies highlight the independence of imprinting on cell types, but suggest variability across cell types. Also, it has been shown that beyond DNA methylation, polycomb repressive complexes, and insulation via CTCF is important for genomic imprinting, but it remains an open question whether these could contribute to context-specificity. In particular, non-canonical imprinting which might be driven by polycomb complexes is a candidate for context-specificity [Hanna, 2020]. Similarly, secondary DMRs are candidates for context-specific imprinting, as they can be established later during development, for example at the *Cdkn1c* gene [Wood et al., 2010, Fan et al., 2005].

1.2.3 X-chromosome inactivation

The second canonical class of allele-specific expression due to epigenetic effects arises from X-chromosome inactivation in female mammals[4], a process that acts as a dosage compensation mechanism to equalize gene expression levels between sexes with one or two X chromosomes, and is thought to be essential for cellular function [Loda et al., 2022]. Remarkably, the necessity for such a mechanism was first proposed without much knowledge about the molecular underpinnings, but has been fully confirmed experimentally later [Lyon, 1961]. Classical X-inactivation is agnostic to the parental haplotype of the X chromosome, and therefore randomly inactivates either copy during early development and just before gastrulation. Subsequently, the X-inactivation status is maintained through cell division and creates a mosaic adult individual. Mechanistically, XCI is triggered by a long non-coding RNA called *Xist* [Willard, 1996, Plath et al., 2002]. In the epiblast cells of the blastocyst, both X-chromosomes are active. In coordination with the exit from pluripotency, both chromosomes upregulate *Xist*, but a reciprocal feedback mechanism involving the antisense lncRNA *Tsix* represses a single allele at the expense of the other, eventually leading to full mono-allelic *Xist* expression within individual cells (**Fig. 1.6**). *Xist* then spreads, exploiting the 3-dimensional structure of the chromosome, in *cis* over the X-chromosome from the X-inactivation center where the *Xist* gene is located [Furlan and Galupa, 2022, Fang et al., 2019]. Silencing of the X-chromosome is initiated by Xist-mediated recruitment of repressive factors and complexes, including SPEN, polycomb and others. Establishment of the fully silenced X finally requires repressive histone marks like H3K27me3, DNA methylation and massive changes in the three dimensional chromosome structure [Giorgetti et al., 2016]. *In vivo* as well as in cell culture systems based on mouse embryonic stem cells, this process is fully completed after around 6 days and textbook knowledge predicts only minor changes to the genome structure of the Xi after that [Wutz et al., 2002].

As a mechanism causing allelic bias, XCI is expected to lead to full monoallelic expres-

[4]We are using male and female strictly as referring to (chromosomal) sexes defined by XY and XX karyotypes, not in relation to gender identity. Also, we are not considering rarer karyotypes such as XXY (Klinefelter syndrome), which undergo X-chromosome inactivation but have many characteristics of the male sex [Kinjo et al., 2020].

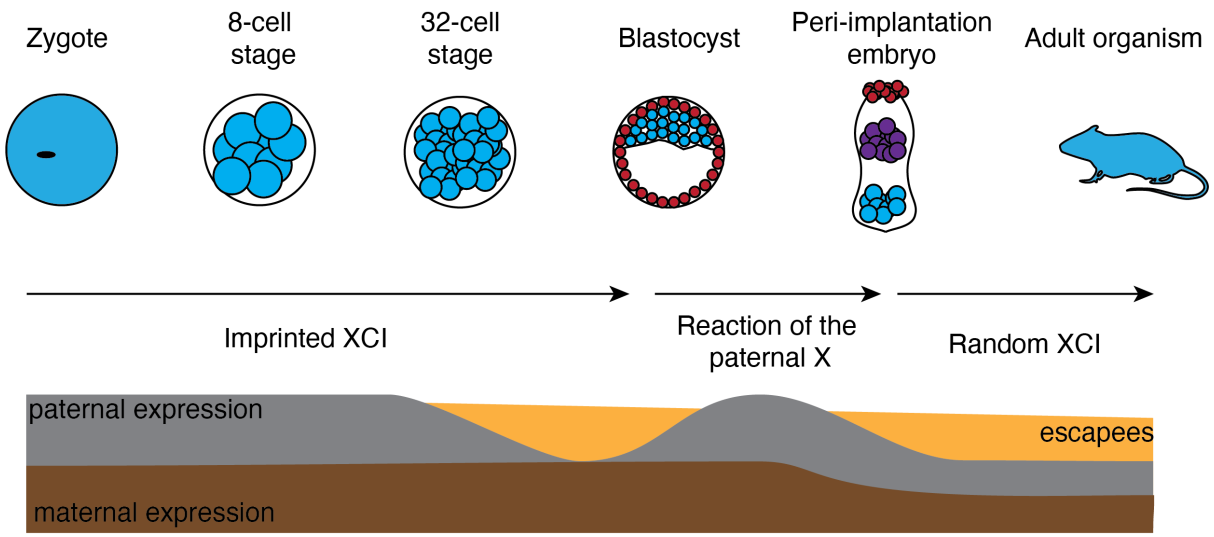

Figure 1.6: X-chromosome inactivation in the mouse. Adapted from [Loda et al., 2022]. XCI proceeds during embryonic development in two steps. First, imprinted XCI specifically inavtivates the paternal X-chromosome due to a non-canonical imprint on the maternal *Xist* gene. This process is finished in the blastocyst, and reversed shortly after. Then, random XCI takes place after implantation and affects the paternal haplotypes randomly. Both XCI waves half the expression output from X-linked genes, unless these genes escape XCI.

sion of X-linked genes at a randomly chosen haplotype at the single-cell level. However, not all genes are subject to full silencing and are known as genes that ”escape” XCI. For these escapees, XCI might introduce a range of allelic balance levels or not lead to any silencing at all. One focus on this thesis is to explore the extent of XCI quantitatively in different cell types, and to which extent the inactive X can be reactivated [Carrel and Willard, 2005, Balaton et al., 2015, Berletch et al., 2015, Marks et al., 2015]. Most escape from X-inactivation is assumed to be necessitated due to dosage compensation. While the majority of genes on the X-chromosome are dosage compensated with respect to the male karyotype by XCI, a subset of genes is present on both the X and Y as orthologous gene pairs (gametologs) [Zhou et al., 2023]. By escaping silencing, these genes maintain dosage compensation through XCI. In humans, these genes are mainly located on the pseudo-autosomal regions (PARs) of the sex chromosomes, which originate from the autosomal ancestor chromosome before their divergence and deterioration of the Y. In mice, where PARs are greatly reduced, gametologs are found spread across the chromosome. Notably, many of the gene pairs escaping XCI are shared between mice, humans and other therian mammals, for example *Km5c/Kdm5d*, *Kdm6a/Uty* and *Ddx3x/Ddx3y* This suggests an evolutionary origin for their escape from XCI.

Because gametologs and genes with similar escape patterns are biallelically expressed across almost all cell types and contexts so far surveyed, including imprinted XCI in pre-implantation embryos and extraembryonic tissues, they are sometimes summarized under the umbrella term ”constitutive” escapees **(Figure 1.7)**. This suggests that they show intrinsic resistance to X-inactivation, regardless of the cellular context. However, in mice and humans biallelic expression of non-constitutive escapees has also been observed sporadically and inconsistently across cell types. These genes are sometimes known as ”facultative” escapees, as they are not necessarily found in all contexts. Facultative in this context has been used interchangeably for variation in escape across cell types, tissues, developmental stages, individuals and individual

cells. There is a need for both comprehensive survey of escape across these contexts, and clearer terminology regarding potentially very different mechanisms causing variation in allelic balance [Wainer Katsir and Linial, 2019, Tukiainen et al., 2017, Tomofuji et al., 2023, Garieri et al., 2018, Peeters et al., 2014, Petropoulos et al., 2016].

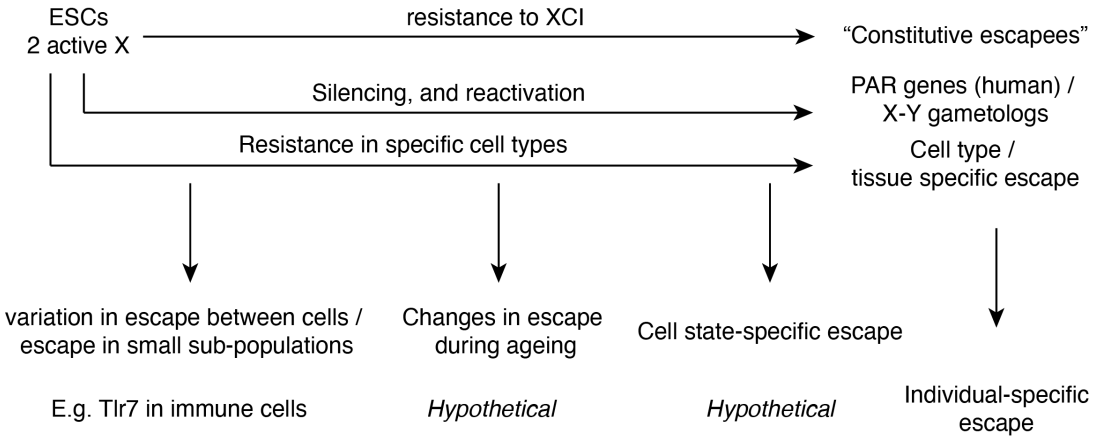

Figure 1.7: Different modes and sources of escape from XCI. Potential mechanisms driving expression from the Xi. Possible courses are a resistance to inactivation, especially in constitutive escapees or PAR genes. Other potential paths is reactivation in the adult, or resistance in specific cells, which lead to context-specific escape.

These molecular mechanisms, both regarding facultative and constitutive escape, remain largely elusive. It is known that escapees exhibit an active regulatory environment, including accessible chromatin, low DNA methylation at promoters and active histone marks [Berletch et al., 2015, Qu et al., 2015]. Interestingly, the three dimensional structure of the X chromosome places escapee genes outside of heterochromatic super-domains, which points towards CTCF-mediated TAD structures as a factor in allowing escape [Giorgetti et al., 2016]. The inactive X seems largely devoid of intergenic accessible chromatin, suggesting that proximal or distal enhancers are not a major driver of escapee expression [Giorgetti et al., 2016]. These results seem to suggest that at large, escapees on the active and inactive X are regulated equivalently. However, the possibility that CREs or transcription factors regulate expression specificically on the Xi remains open. It is likewise largely unclear whether in general, expression from the Xi results from of a lack of inactivation, or a specific re-activation. Many escapees appear to fully or partially escape silencing during random XCI in the blastocyst [Loda et al., 2022], but it is unclear whether the level of escape remains stable throughout development. Some escapees, especially tissue-specific ones, are likely re-activated. It is known that loss of *Xist* can lead to mild reactivation of the Xi in tissues and tumors, but the vast majority of genes does not show changes in escape even upon complete removal [Csankovszki et al., 2001, Splinter et al., 2011, Adrianse et al., 2018]. Similarly, reprogramming of differentiated cells in to a pluripotent state leads to re-activation of the Xi [Janiszewski et al., 2019].

The main cell biological relevance of escape from XCI is that it provides a possible explanation for sex-biased gene expression. In the absence of other regulatory effects, escapees will be expressed at a higher combined dosage in females compared to males, and sex-biased gene expression is thought to underlie sex-specific phenotypes [Carrel and Brown, 2017]. In disease, sex-bias is well documented, but which factors cause differential incidence between

sexes remains an open question [Xing et al., 2022]. It is known that sex-bias originates from differential environments and lifestyle choices between men and women, which are independent of biological differences. For example, lung cancer incidence tracks with smoking behaviour in both sexes [Ragavan and Patel, 2022]. Secondly, sex hormones are known to drive disease risk, which is mechanistically independent of the direct action of karyotypic differences [Moulton, 2018]. Effects directly through the chromosome complement are the third option, which are mostly due to Y-linked gene expression and escape from XCI. Which factors contribute most to disease risk likely depends on the condition in question and the affected individuals. Unravelling the different mechanisms underlying sex-biased disease risk will pave the way towards targeted therapies specific for women and men. One example with strong evidence for a contribution of escape to disease risk are autoimmune diseases, which will be discussed further in **Chapter 3**.

Escape has been studied extensively in human samples, but mouse and murine cell culture models remain an important workhorses for the study of escape from XCI. On the one hand, we rely on mice to study escape in controlled *in vivo* systems and under genomic perturbations. Furthermore, the limited density of heterozygous variants in human samples can preclude direct analysis of escape (see also the **Discussion**). To study human escapees, many articles therefore resolve to only assess sex-biased expression as a proxy, which can also result from regulatory differences on the active X [Tukiainen et al., 2017]. It is therefore important to assess the ability of model organisms such as mice to model human escape. On the one hand, escape from XCI is a phenomenon clearly conserved across therian mammals. Furthermore, the set of escapees seems to be partially conserved across many species, especially but not limited to gametologs [Peeters et al., 2014, Carrel and Willard, 2005]. However, escape seems to be much more common in humans than in mice. This is likely due to the small size of the PAR on the acrocentric X-chromosome in mice, whereas in humans, the entire pre-centromeric region is pseudo-autosomal. This enlarged PAR contains many escapees, and the centromere has been suggested to be a barrier to *Xist*-spreading [Berletch et al., 2010]. Previous studies have estimated the set of escapees on the mouse X to be 5-10% of expressed genes, as opposed to 20% on the human X [Loda et al., 2022]. However, comprehensive studies across developmental stages, tissues and cell types are still lacking, so a final quantification remains outstanding. In any case, in the study of escape it remains particularly important to validate findings in model organisms in humans.

Elucidating the full set of escapees across species and in different tissue-, cell type- and age-contexts will pave the way for a mechanistic understanding of this process and its consequences for sex-biased biology. A genome-wide survey has identified genes with tissue-specific escape in humans [Tukiainen et al., 2017]. Similarly, mouse tissues show varying escape to some degree, although expression levels from the Xi are relatively low in this study [Berletch et al., 2015]. Further work has implicated CTCF as a determinant of tissue-specific escape [Fang et al., 2023]. Variability is also seen between human individuals, although it is unclear whether this is genetically encoded or due to a stochastic epigenetic mechanism [Carrel and Willard, 2005]. For development-, Individual-, tissue- and even cell line-specific escape the umbrella term ”facultative escapee” has been used, in particular to contrast them to ”constitutive escapees” (which might also be a heterogeneous group). However, these processes are likely mechanistically variable. We will address in particular cell type- and age-specific escape in **Chapter 3** of this thesis.

1.2.4 Random mono-allelic expression

Finally, stable non-genetic allelic differences can be driven by other biological mechanisms that affect specific sets of genes [Gendrel et al., 2016]. As the direct cause of these biases is not known, they are commonly known as genes with ”random monoallelic expression” (RAME). It is known that immunoglobulin choice in B-cells selects a single allele of a specific Ig-gene [Nutt et al., 1999]. Similarly, olfactory receptors only express one allele of a single *Olfr* gene out of the 100s-1000s of possibilities, which is driven by *trans*-interactions between the two chromosomes [Magklara and Lomvardas, 2013, Monahan et al., 2019]. Other genes with such expression patterns include Interleukins in immune cells and [Chess, 2005, Guo et al., 2005]. However, these phenomena are comparatively rare, only described in specific gene sets and cell types, and mechanistic explanations or their necessity remain unexplored.

Whether mitotically stable inactivation of genes is a more widespread phenomenon, and whether it contributes to phenotypic differences has received further interest in the recent years. To distinguish RAME from genetic effects, an array of isogenic clonal populations has to be used. If genes appear varyingly mono-allelically expressed between both haplotypes in the same genetic background, it can be concluded that this is not a function of sequence variability. As such, persistent RAME that does not vary between cells or clonal propagates thereof can not be distinguished from direct genetic action, if it is always biased towards the same paternal haplotype. Early studies used these approaches and proposed that around 10% of genes might show RAME [Gimelbrant et al., 2007], although that number is debated [Gendrel et al., 2014, Li et al., 2012]. RAME has been shown to affect genes important for embryonic development, including the transcription factors *Six1* and *Eya1/2*, [Gendrel et al., 2014]. There is therefore substantial interest in determining the extent of RAME and its relevance for disease.

An open question is the extent of RAME *in vivo*, as it is inherently difficult to detect in population-based measurements. Direct identification requires either clonal or temporal tracing of the same cells, to prove that allelic variation is passed through mitosis. Recently, a study has used expression data from many tissues to identify genes whose allelic ratio varies across measurements within the same individual [Kravitz et al., 2023]. The results suggest that non-genetic variation in allelic ratios is common, but it is difficult to distinguish this from tissue-specific action on genetic effects.

1.2.5 Stochastic allelic dynamics

The last cause of ASE arises from stochastic fluctuations in allelic usage. Although it is sometimes discussed in similar terms, this class of allelic effects is fundamentally different to random mono-allelic expression [Gendrel et al., 2016, Raj and van Oudenaarden, 2008] (**Figure 1.7**). While RAME is mitotically stable and propagated clonally, allelic dynamics can show expression of either haplotype within the same cell and potentially in short time-spans. It is tightly connected to the bursting model of transcription, where it is assumed that mRNA is produced during short periods of gene activity, as opposed to continuous production of mRNA [Fukaya et al., 2016, Tunnacliffe and Chubb, 2020]. As both alleles are to a first approximation acting independently, bursting is the result of two stochastic processes and allelic usage will vary between full mono-allelic bias and bi-allelic expression over time. Bursts can be described in terms of their size (the amount of RNA molecules produced,

usually considered proportional to the burst length) and their frequency (how often they occur), which decomposes "expression level" into two distinct modes of regulation **(Fig. 1.8)**. Measuring bursting necessarily requires single-cell resolution. This can be achieved by single-molecule live-imaging, where the temporal resolution provides a direct readout of allelic usage over time. These studies have revealed, for example, how enhancers control expression by modulating parameters of bursting [Donovan et al., 2019, Chubb et al., 2006, Patel et al., 2023, Bartman et al., 2016, Rodriguez et al., 2019].

However, transcriptional kinetics can also be derived from static measurements of independent cells, as bursting will introduce an overdispersed signal and under certain assumptions, these kinetics can be recovered from repeated measurements of independent and identical cells from a homogenous population. ScRNA-Seq has provided the first genome-wide estimates of transcriptional bursting, which can yield important mechanistic information on gene regulation [Larsson et al., 2019, Deng et al., 2014a, Reinius et al., 2016]. These studies have confirmed that, at the single-cell level, allelic bias due to transcriptional bursting is common. They have also made first steps towards genome-wide mechanistic explanations of bursting by showing and confirming that both promoter and enhancer features contribute to bursting. However, it is important to concede that it is difficult to obtain precise overdispersion estimates which are critical for bursting kinetics inference. In particular technical and biological variation, such as PCR bias and inter-cell heterogeneity will increase the observed overdispersion regardless of bursting [Larsson et al., 2019, Grima and Esmenjaud, 2023].

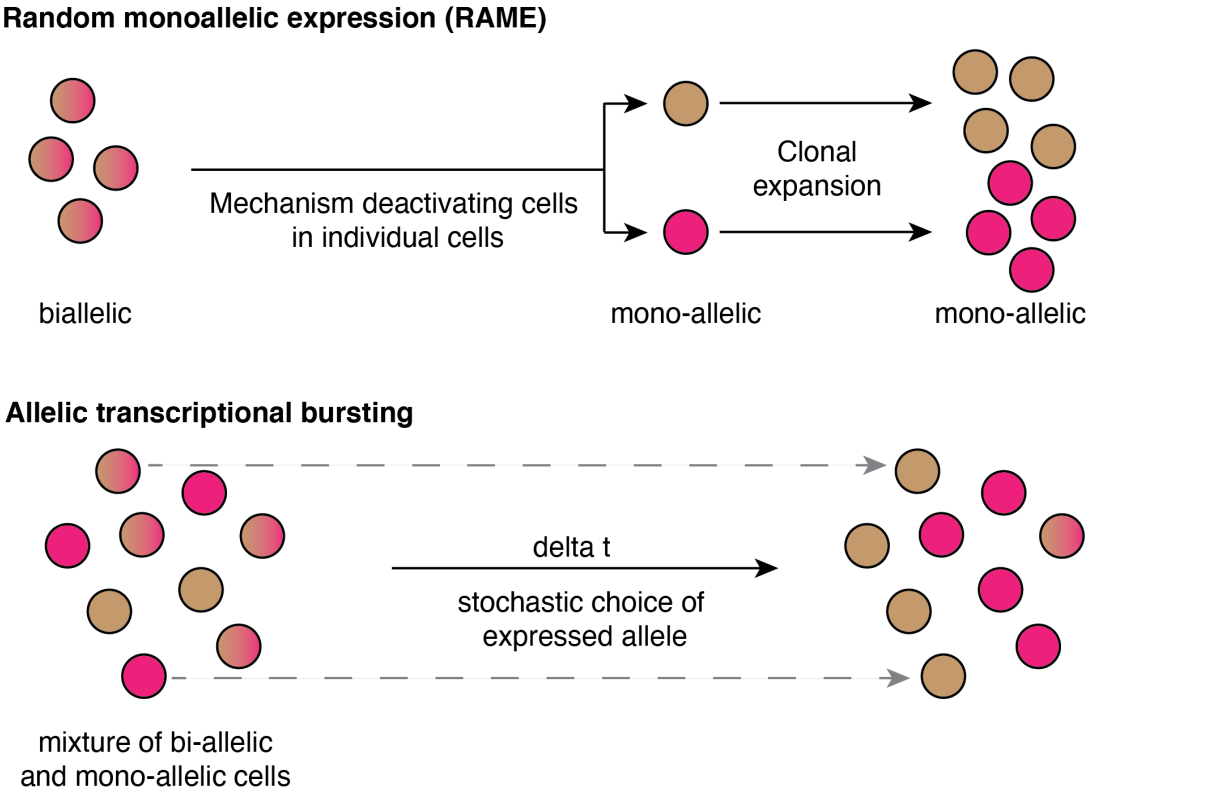

Figure 1.8: Differences between RAME and stochastic allelic usage. Illustrating the conceptual distinction of RAME to allelic bias due to transcriptional bursting. In RAME, individual cells inactivate one specific allele. This information is passed through mitosis into daughter cells, leading to a mosaic population, similar to that resulting from X-inactivation. Transcriptional bursting of two independent alleles leads to mosaic allelic usage, but is not stable in time or through mitosis.

1.3 Probing allele-specific expression in context using single-cell readouts

1.3.1 The resolution revolution

With the advent of cost-effective high-throughput sequencing and the availability of high-quality genomes, the last 20 years have seen a revolution in biologists ability to quantitatively measure molecular phenotypes. By converting the entire content of biomolecules in samples into sequencing libraries, one gets a genome-wide view of cells ('omics'-view), as opposed to classical assays that typically measure only one or few specific genes (or proteins and metabolites in non-sequencing-based assays). The first full map generated was of the human genome [Lander et al., 2001], but soon after, sequencing methods for the quantification of mRNA (via cDNA libraries), DNA modifications (in particular, DNA methylation), post-translational histone modifications, genome structure and RNA-protein interactions were developed [Stark et al., 2019, Moore et al., 2020, Park, 2009, Bienko, 2023, Hafner et al., 2021]. By now, novel "-Seq" methodologies are being developed yearly and allow for the measurement of various molecular processes.

While these assays generate genome-wide information, they only provide aggregate measurements across many cells and therefore average over cell-to-cell heterogeneity, which poses a difficulty in data analysis when working in complex samples. This short-coming has been addressed in the last decade by single-cell methods, which provide quantifications at the "biological unit" of a sample - the cell [Tang et al., 2009]. Especially mRNA-Seq of single cells has developed considerably, and processing thousands of cells is now a routine experiment using commercial solutions. Furthermore, single-cell level assays of chromatin accessibility, DNA methylation and histone modifications are now also widely available [Stuart and Satija, 2019]. Naturally, all sequencing-based methods also provide information about the sequenced allele if there is genetic variation distinguishing them.

1.3.2 Single-cell sequencing methods

Of the single-cell methods, sc-mRNA-Sequencing is the most mature [Hwang et al., 2018, Papalexi and Satija, 2017, Potter, 2018]. This is in part by the fact that a major limitation of single-cell based methods is the low amount of input material, and while cells only contain two copies of each gene, each mRNA can be present in potentially many copies. All scRNA-Seq methods function by first converting RNA into cDNA libraries, using either priming from the poly-A tail for mRNA-Sequencing or more elaborate strategies to also capture non-polyadenylated RNAs. To achieve single-cell resolution, different methods to compartmentalize individual cells have been used (**Fig. 1.9**). First proof of concept studies manually isolated single cells for sequencing, which was soon replaced by automated cell sorting, scaling up throughput to 100s and 1000s of cells [Picelli et al., 2014]. Microfluidic separation of cells for reverse transcription and successive pooling has increased throughput by an order of magnitude and underlies the currently most commonly used technologies [Zheng et al., 2017]. Even higher throughput can be achieved by so-called combinatorial indexing protocols, where the cells (or nuclei) themselves provide the reaction compartment to introduce barcodes during reverse transcription and ligation steps. The key inside here is that by repeatedly mixing and barcoding sub-pools of the sample, a large enough barcode space makes it stochastically unlikely to yield two cells with the same label, effectively achieving

single-cell resolution without physical separation of individual cells [Cao et al., 2017]. Besides their throughput in number of cells, these methods differ in their detection sensitivity, which is likely determined by the capture efficiency of mRNA molecules during reverse transcription [Ziegenhain et al., 2017]. While plate-based methods might perform better in this metric, although they more likely suffer from PCR bias if not corrected using unique molecular identifiers. In most applications, it is arguably more efficient to measure many cells with fewer detected mRNA-molecules, than to sequence few cells with the aim of capturing many, although measurements of cell-to-cell variability might rely on the second strategy. High-throughput methods furthermore allow higher sensitivity in detecting rare cell populations, and therefore might be preferred when for example complex tissues are analyzed. In general, the purpose of the application should dictate the method used.

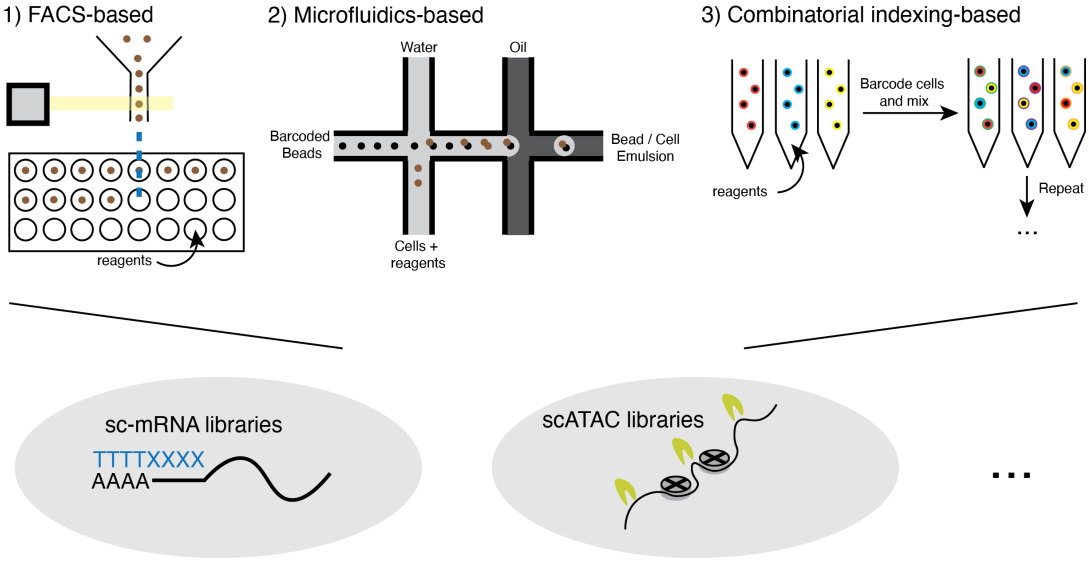

Figure 1.9: Different technologies to generate single-cell libraries. Individual cells need to be isolated for all technologies and barcoded before library preparation. This can be achieved by fluorescence-assisted cell sorting (FACS) into individual wells, in which the library is generated (1). Alternatively, microfluidics with beads can transfer barcodes, or combinatorial indexing can generate individual barcodes per cell.

A second breakthrough of single-cell methods has been the establishment of scATAC-seq to measure chromatin accessibility (*Assay for transposase-accessible chromatin*, [Buenrostro et al., 2015]). Here, the ability of a Tn5 transposome to cleave and barcode native chromatin is measured, which correlates with the regions accessibility to regulatory factors and RNA polymerase. Although accessibility can be similarly measured using DNAse and methyl-transferases, transposase-based assays have been easiest to incorporate into plate-, microfluidic- and combinatorial indexing and are most widely used. By using antibody-based targeting of the Tn5 transposome to specific chromatin marks, single-cell Cut-and-Tag methods extend scATAC-Seq to profile histone modifications in single cells, revealing heterogeneity in chromatin states [Bartosovic et al., 2021]. Also bisulfite-based DNA methylation has been analyzed at the single-cell level in both low- and high-throughput assays [Smallwood et al., 2014]. For all of these epigenomic single-cell methods, the interpretation of their quantitative signal differs from RNA-Seq. While mRNA-levels can be assumed to be the result from a genes activity, cells only contain two copies of their genome and can therefore only be in three different epigenetic states. With perfect measurement sensitivity,

the number of reads would represent the probability of accessibility (or presence of a histone modification or DNA methylation) across a population of cells, as is the case in a bulk assay. Single-cell epigenomic profiling is therefore best understood as linking these probabilities between genomic loci. However, we have to expect these assays to give highly sparse signal, and the absence of a read at a genomic locus does not correspond to the absence of the assayed feature.

Finally, sequencing methods are being extended to many other molecular readouts: At the RNA-level, there are single-cell methods for non-polyadenylated RNAs (such as miRNAs, tRNAs, transposable elements and others) and RNA bound to ribosomes [Salmen et al., 2022, VanInsberghe et al., 2021]. At the DNA-level, assays are being extended to measure replication timing and [Miura et al., 2019]. Although these are not topic of this thesis, all of these assays can be or have been applied to study allele-specific signals.

1.3.3 Tracing cell types and states with scRNA-Seq data

The primary purpose of single-cell resolution is to deconvolve expression and its variation into the cell types these measurements correspond to. On the one hand, this improves the sensitivity of differential expression analysis, as changes in specific, possibly rare cell types can be directly measured instead of being masked by a tissue-aggregated measurement. In this sense, single-cell readouts simply increase the resolution of bulk sequencing assays. However, measuring the totality of gene expression levels also provides a novel way of defining cell types in the first place. Cell types have traditionally been defined by their function or morphology [Fleck et al., 2023]. However, both of these are a result of their gene expression, which in turn stems from their cell type-specific transcriptional regulation. Measuring gene expression levels therefore assesses cell type identity at its source. Furthermore, different cellular phenotypes are encoded by the expression of independent gene sets. By assaying all genes simultaneously, scRNA-Seq can quantify the activity of orthogonal cellular programs to refine the definition of cell types or states.

Cells can change their cell type through cellular differentiation, which is directly necessitated by the fact that all cells originate from the same fertilized zygote. Both during embryonic development and in adult differentiation systems, pluri- or multipotent stem cells give rise to varying individually different cell types. Differentiation per se is uni-directional and irreversible as mature cells can not re-enter a multipotent state, although cell types can be changed artificially through cellular reprogramming [Takahashi and Yamanaka, 2006, Joung et al., 2023] or in regenerating species [Reddy et al., 2019]. However, in a given cell type, a systematic transcriptional shift is not necessarily creating a new one. For example, an activated T-cell shows drastically changed transcriptional profile, but is still considered a T-cell, and similarly changes in cell gene expression due to cell cycle phases, senescence or apoptosis is not necessarily a cell type transition. For this reason, these kind of regulatory shifts are often known as different *cell states* [Trapnell, 2015].

To identify cell types and states in scRNA-Seq data, one needs to identify populations of cells that are substantially different from each other. Starting with a matrix of expression levels across genes and cells, the most common approach is to first reduce the dimensionality of this space by excluding genes with constant expression and to compress the set of genes into a smaller amount of independent factors, for example using principal component or factor

analysis. Assuming that cell types represent sets of cells with similar expression profiles, but different from all other cells, clustering analysis is then usually use to group cells, and these clusters can be assigned to known or unknown cell types based prior knowledge about expression patterns. Of note, there is no obvious way to distinguish whether a shift in expression profile corresponds to a different cell types or state [Ianevski et al., 2022].

Mathematically formulated, this concept considers a cell as a point in an n-dimensional space defined by the number of detected genes. In this space, different cell types are neighbourhoods around an average expression profile of that cell type and in which individual genes can vary in an accepted range. This variability can encompass sub-cell types, different cell states and the natural biological variability that a cell type can exist in. By this definition of sub-celltypes, the concept is hierarchical. For example, immune cells can be considered a (broad) cell type, which encompasses the myeloid and lymphoid cells as sub-cell types, which in turn encompass the known cell types of the hematopoietic tree. In this hierarchy, defined before the advent of transcriptomic profiling, cells with similar functions tend to have similar gene expression, and transcription tends to predict relatedness in hematopoiesis [Velten et al., 2017]. While the cell type concept necessitates a discrete classification, changes between cell types and states are naturally smooth transitions. One key contribution of single-cell methods is that by sampling many cells across such a transition, cell types can be assessed in a continuous manner. Indeed, describing cellular differentiation as pseudo-temporally ordered processes has been one of the main focus points of the first single-cell transcriptomics studies [Trapnell et al., 2014].

As this definition follows from the notion that cell types can be defined by their gene regulatory programs, profiling of chromatin states, protein content or even translation rates can also be used to define cell types [Pierce et al., 2021, VanInsberghe et al., 2021]. It is an open question which of these, in the absence of technical noise, carry the most information on cell state changes, if there is any difference at all [Fleck et al., 2023]. For example, it is conceivable that some cell state changes, for example ageing, might track more on the DNA methylation level and it is known that first responses to stimuli are often driven by post-transcriptional mechanisms [Teixeira and Lehmann, 2019].

There are limitations to this regulatory definition of a cell type. In particular, it largely ignores physical location and cell division history. While the context that a cell is placed in is likely to induce gene expression changes, for some cell type such as neurons, their location and connectivity is going to be vastly more informative than their expression profile [Zeng, 2022]. Similarly, different lymphocyte clones will be more well defined by the antigen they are reactive to, than the potentially small transcriptional variability between them. It is therefore necessary to consider the appropriate features when classifying cell states and how they relate to their function.

Finally, cell types have to be defined in relation to the evolutionary history they appear in [Arendt et al., 2016]. By comparative analysis, homologous cell types between different species can be defined by their molecular makeup [Sarropoulos et al., 2021, Behm et al., 2023]. At the same time, the transcriptomic similarity relates to the rate at which these cell types in different species diverge from their common ancestor [Murat et al., 2022]. [Arendt et al., 2016] define cell types as 'a set of cells in an organism that change in evolution together, partially independent of other cells, and are evolutionarily more closely related to each other than to other cells'. In this model novel cell types can emerge from ancestral ones through

the emergence of novel specifying gene regulatory networks. Also, ostensibly similar cell types can have different evolutionary histories even though they have simlar expression profiles.

In summary, single-cell methods provide the data necessary to formalize the powerful concept of cell types defined by their regulatory state. This has lead to broad ”cell atlas” initiatives that aim to map the complete possible space of expression states [Osumi-Sutherland et al., 2021].The conceptual importance for this thesis is that by adding allelically resolved data to whole transcriptome information, we can relate cell state to allele-specific regulatory mechanisms.

1.3.4 Multimodal assays and spatio-temporal analysis

The most recent developments in single-cell methods have aimed at two tasks: The first is to combine multiple single-cell measurements from the same individual cell. Doing so allows to decipher the inter-dependencies between transcriptome and regulatory layers [Baysoy et al., 2023]. Indeed, the first successful multi-omics approaches combined RNA-sequencing with either DNA methylation, chromatin accessibility, or both. Multimodal data can also be achieved by combining readouts that destroy the cell such as single-cell RNA-Sequencing with measurements that do not, for example antibody-based profiling of membrane protein expression by FACS sorting or using oligonucleotide-conjugated antibodies. As with unimodal readouts, multimodal assays vary in throughput, sensitivity and experimental complexity, and in the future, researchers will likely explore further combinations of existing sequencing assays and increase their technical capabilities [Stoeckius et al., 2017, Argelaguet et al., 2019]. Multimodal measurements provide two key opportunities: On the one hand, global patterns in different datasets can be directly compared to identify the modalities explaining specific sources of variability in the dataset. On the other hand, correlation-based approaches can be used to link measurements at specific loci, for example enhancer activities to gene expression, and thereby nominate mechanistic relationships [Argelaguet et al., 2018, Granja et al., 2021].

The second frontier has been to place cells back into their spatio-temporal context. For tissues, single-cell analysis usually starts with a digestion protocol that yields a single-cell solution, information of the cells position is therefore lost. Spatial transcriptomics methods have started to assay expression levels in native tissues, first by aggregating multiple cells, but increasingly at the cellular and sub-cellular resolution [Chen et al., 2015, Rodriguez and Larson, 2020]. These methods have been extended to also assay chromatin accessibility, histone modifications and combinations thereof and are an active field of research [Deng et al., 2022]. As an alternative to slide-based spatial transcriptomics, cells can be barcoded based on their position and then subjected to varying single-cell sequencing methods [Srivatsan et al., 2021]. This opens the potential to apply the vast array of single-cell methods directly to spatially resolved samples. A somewhat overlooked dimension is longitudinal analysis of the same cells over time. This is due to the fact that most readouts require lysis of the cell, but also because through longitudinal sampling or pseudotemporal analysis of dynamic processes, similar cells have been successfully used as proxys instead of repeatedly measuring the actual same cell. However, both repeated sampling of mRNA and an assay that ”records” the activity of DNA binding proteins for later analysis have been demonstrated and provide an exciting view of future possibilities [Chen et al., 2022, Frieda et al., 2017].

1.3.5 Quantifying allelic imbalance in expression levels measured by scRNA-Seq

When two alleles differ in their DNA sequence, it is straightforward to distinguish from which allele a sequencing reads was derived. In the context of RNA-Seq, this idea allows for the independent quantification of allelic gene expression levels. Individual reads mapped to a reference genome, and then checked for whether they contain a heterozygous variant differing between the two alleles. In practice, it is advisable that these variants are known a priori and not inferred from the measurement dataset, as variant calling for example from RNA-Seq data is prone to yield false positives [Quinones-Valdez et al., 2022]. A second distinction is whether the set of variants is *phased* across the genome, that is, whether for each pair of loci on the same chromosome, the different alleles can be assigned to one or the other chromosome. When information from the parents is available, the different chromosomes can be further phased into full maternal or paternal haplotypes, and allelic measurements can be fully assigned to them. For human samples, phasing can be achieved by sequencing trios of a donor and both parents, by using artificial separation of the two genomes or by imputation algorithms [Porubsky et al., 2020, Browning and Browning, 2011].

The second difficulty arises from potential errors in either mapping or allelic assignments. Erroneous genotyping will yield fully biased estimates of allelic usage, so it is important to ensure accurate input genotypes, also because false negatives pose less of a problem as they only reduce the amount of assignable reads [Castel et al., 2015]. However, mapping errors can be more challenging to identify and eliminate. As the DNA sequence differs between the two haplotypes their mapping this poses a fundamental bias, which makes allelic assignments especially difficult when the to be mapped regions are repetitive. Importantly, not only observed, but also unobserved variants bias this analysis, further highlighting the need for complete genotype information. In general, there will be a ”reference bias”, that is, the haplotype that is closer to the reference genome will be favoured due to these unobserved variants. One attempt to reduce mapping bias is to remove the tagging variants used for allelic assignment from the alignment procedure (N-masking).To account for this approaches like WASP propose a two-step procedure by which reads are mapped first as observed, and then simulated to contain the alternative variant. Only the reads mapping to the same locus in both scenarios are retained for analysis, which has been shown to improve specificity compared to N-masked mapping [van de Geijn et al., 2015]. However, neither approach accounts for scenarios in which the alternative genotype show unobserved variation. If possible, it is therefore advisable to include controls where the allelic bias is known a priori to identify problematic areas in the genome. Of note, these errors introduce bias that will be systematic across an experiment. When analysing variation in allelic bias between tissues or cells within the same experiment, these technical factors are largely going to introduce the same bias in every case.

Specific to single-cell readouts, the inherent sparsity of these methods poses an additional problem, which is further exacerbated by the necessity to cover a SNV to assign reads to a haplotype. The fraction of assignable reads depends on the SNP-density in the analyzed sample and the read length utilized. As all current single-cell methods rely on short-read sequencing, this parameter is inherently capped at the maximum read length of around 200 base pairs, potentially twice that when paired end sequencing is available. The variant density is similarly a fixed parameter - in humans it is commonly estimated as one variant every per kilobase, although the

density is non-uniform and coding regions are generally depleted for variants. Finally, for gene expression analysis, common methods often only measure reads derived from the 5' or 3'-end of the transcript, which further reduces the likelihood of hitting variants compared to full-length protocols, which are therefore commonly advertised as preferable for allelic mapping [Glinos et al., 2022, Cuomo et al., 2020]. The observed allele-specific signal is therefore likely substantially lower than the already sparse total signal, and this reduction is only partially uniform across the genome. This is particularly relevant for single-cell chromatin readouts, where the expected number of reads only covers a fraction of the accessible genome.

1.3.6 Exploring variable allelic imbalance using single-cell methods

Allelic bias and variability thereof have been part of single-cell sequencing analysis since the inception of early methods. One of the first large-scale studies measured allele-specific expression during the earliest steps of mouse development and were able to visualize zygotic genome activation, imprinted XCI and suggested widespread allelic bias [Deng et al., 2014b]. In particular X-inactivation has been extensively analyzed with single-cell methods, as it allows for the deconvolution of random inactivation. These studies have contributed to the quantification of silencing dynamics and X-chromosome upregulation [Pacini et al., 2021, Lentini et al., 2022, Borensztein et al., 2017]. Other work has addressed the question to which extent escape from XCI varies across cell types [Tukiainen et al., 2017, Tomofuji et al., 2023, Garieri et al., 2018, Wainer Katsir and Linial, 2019]. Single-cell analysis has also been used to detect mosaic imprinting, demonstrating heritable allelic variation [Santoni et al., 2017, Ginart et al., 2016].

ASE has also been a critical tool to explore context specific genetic effects in human samples. As described previously, ASE isolates the impact of *cis*-genetic effects on gene expression. Importantly, by using the ratio between two alleles, multiplicative biases on gene expression are internally normalized for. Single-cell level measurements therefore quantify the strength of an allelic effect at the single-cell level, without the necessity to quantify across large cohorts. This approach has been used in multiple single-cell genetics studies ([Qi et al., 2023, Heinen et al., 2022, Fan et al., 2021, Cuomo et al., 2020]).

Single-cell methods have finally been used to quantify random mono-allelic expression and transcriptional bursting. An scRNA-Seq based study has helped to address the question how much expression is mono-allelic [Reinius et al., 2016, Gimelbrant et al., 2007]. In particular these studies have helped to clarify the distinction between random bursting-derived mono-allelic expression and clonally inherited RAME, and addressed technical biases in these quantifications [Choi et al., 2019, Reinius and Sandberg, 2015, Kim and Marioni, 2013].

1.3.7 Non-sequencing based methods to measure allelic imbalance

All so far described methods rely on the presence of known heterozygous variants between the paternal haplotypes, which require genetically heterogeneous populations or bespoke animal models. There are currently no methods that provide a generalizable approach to map alleles in the absence of these variants. If the homologous chromosomes are genetically indistinguishable, the only other discerning factor is their nuclear position. This allows for imaging-based methods to be used to measure allele-specific signals, for example by RNA-fluorescence in situ hybridization (FISH) or labelling in transgenic lines

[Herzing et al., 2002, Raj and van Oudenaarden, 2008, Ginart et al., 2016]. This approach is a standard readout when characterizing the active and inactive X-chromosome, where the inactive X can additionally be tagged by the lncRNA *Xist*, although these approaches are rarely quantitative [Yue et al., 2014]. However, single-molecule FISH can be used to count transcripts with essentially full sensitivity, and by measuring chromatin-associated RNA, these can be assigned to the chromosome of origin [Chen et al., 2015]. Similarly, chromatin accessibility can be visualized at individual chromosomes [Chen et al., 2016]. However, these approaches do not currently scale in a genome-wide manner.

1.4 Statistical analysis of allele-specific single-cell omics data

The analysis of allele-specific sequencing data parallels the count based statistical models that have been used for the analysis of total read counts - but using a binomial model for fractional count data (k out of n) instead of poisson-like model for total counts. In this chapter, I am going to discuss the foundations underlying modelling of count-based data. Secondly, analysis of sequencing data greatly relies on generalized linear modelling to test for phenotypes of interest, while accounting for various sources of covariation and the count-based nature of the data. Generalized linear models can be used to perform such analysis for non-gaussian data, which I am also going to discuss. I am going to largely focus on RNA-Seq methods, but both total and allele-specific analysis of other sequencing modalities can, to a first approximation, be done equivalently.

1.4.1 From poisson to binomial statistics

While it is usually assumed that gene expression represents a continuous phenotype, sequencing methods measure mRNA abundance by counting a finite number of individual sequencing reads. As such, measurement models for discrete data can be used, in particular the Poisson-distribution with probability mass function to observe a count of k (k reads).

$$P_\lambda(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1.1}$$

The single parameter λ is simultaneously mean and variance of the distribution and, in the context of sequencing data, can be understood as a latent gene expression level. For large λ, a normal distribution $\mathcal{N}(\lambda, \sqrt{\lambda})$ is a good approximation of the Poisson distribution. For high read counts, explicit handling of the measurement noise can therefore be omitted and the counts can be considered approximately normal. This assumption, together with variance stabilization methods underlies much of the standard pre-processing of sequencing data [Luecken and Theis, 2019]. However especially for low read counts obtained in single-cell sequencing methods, this is critical [Svensson, 2020].

When assaying allele-specific expression, one obtains two counts per measurement, either $k_{Allele1}, k_{Allele2}$ or $k = k_{Allele1}, n = k_{Allele1} + k_{Allele2}$, the more common consideration with one arbitrarily chosen read count and the total counts. While it would be possible to simply model the two allelic expression levels independently, one will usually be interested their ratio, the so called *allelic ratio*, *allelic imbalance* (or *allelic balance*). Modelling the direct relationships between the alleles has a number of advantages, mainly that technical and biological factors affecting the expression measurement will usually affect both measurements equally. Indeed, assuming a multiplicative bias θ, the allelic ratio a given latent expression levels x_1, x_2 will be given as

$$a = \frac{x_1\theta}{x_2\theta} = \frac{x_1}{x_2} \tag{1.2}$$

and is therefore internally normalized. This additionally implies that sequencing depth is not a factor that needs to be accounted for, as opposed to Poisson-distributed data. The natural way to model proportions of counts is the binomial distribution with the probability mass function

$$P_{n,p}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.3}$$

Like the poisson model, this is a one-parameter model with mean np and variance $p(1-p)n$, where p can be understood as a latent ratio between the two read counts. Like the poisson, the binomial approaches a normal distribution with mean μ when $n \to \infty$ and $pn \to \mu$. In allelic analysis, this means that for high read counts, the allelic ratios p can be modelled as normally distributed. However, modelling binomial probabilities respects uncertainty due to differences in total read counts, which would be lost if one would work with allelic ratios and is critical for sparse single cell data.

1.4.2 Accounting for overdispersion

When modelling multiple samples as draws from the same Binomial (or Poisson)-distribution, one has to assume each sample represents a draw from the same underlying rate p (or level λ), which is usually not given in practice, where the measurement is additionally distorted by both biological and technical effects. Equivalently, it can be said that observed data is *overdispersed* with regards to these underlying distributions, where the variance is fully determined by the mean. To account for this, a common approach is to assume that p is drawn from a Beta-distribution with parameters a, b, leading to the hierarchical model

$$k \sim Bin(n, p) \tag{1.4}$$
$$p \sim Beta(a, b) \tag{1.5}$$

where $Beta(x, y) = cx^{a-1}(1-x)^{b-1}$. By marginalizing over p, it can be shown that the joint distribution has the form

$$P_{x,n,a,b} = \binom{n}{x} \frac{B(x+x, n-x+b)}{B(a,b)} \tag{1.6}$$

Where B is the Beta-function $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$. This distribution is called the *Beta-Binomial* and has mean and variance

$$\mu = n\frac{a}{a+b} \tag{1.7}$$
$$\sigma^2 = \frac{nab(a+b+n)}{(a+b)^2(a+b+1)} \tag{1.8}$$
$$= n\mu(1-\mu)\frac{a+b+n}{a+b+1} \tag{1.9}$$
$$= n\mu(1-\mu)(\theta(n-1)+1) \tag{1.10}$$

In this way, the Beta-Binomial can be re-parametrized based on μ and θ, where μ is equivalent to the Binomial rate p and θ controls the amount of excess variance compared to binomial counting noise (**Fig. 1.10**). For Poisson-distributed data, there is an equivalent approach by assuming a gamma-distributed $\lambda \sim Ga(\mu, \theta)$, where the compound distribution is called a *negative binomial*, which underlies most approaches detecting differentially expressed genes [Love et al., 2014a, Robinson et al., 2010].

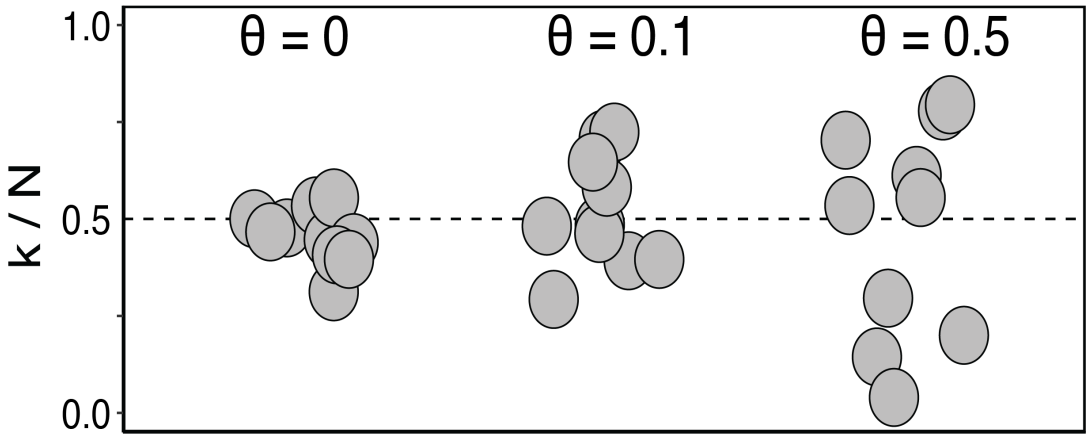

Figure 1.10: Demonstration of overdispersed binomial data. For different overdispersion parameters, draws with $p = 0.5$ and $n = 50$ are shown.

1.4.3 Statistical testing in gLMs

While one can use the model introduced in the previous section to describe the observed distribution of reads given some expression level (which can also be motivated by mechanistic models of transcription) usually one will be interested in differences in expression or allelic ratio over some condition. Given a set of measurements $\mathbf{y}$ and a covariate indicator matrix $\mathbf{X}$ for n covariates, the simplest model is to assume a *linear* (additive) relationship such that for an individual sample i it is:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_n x_{i,N} + \psi_i = \mathbf{X}\boldsymbol{b} + \boldsymbol{\psi} \tag{1.11}$$

with introduction of an intercept term β_0 that captures the mean of $\mathbf{y}$ and use matrix notation to aggregate across all samples. By allowing for normally distributed error $\boldsymbol{\psi} = (\psi_1, ..., \psi_n) \sim \mathcal{N}(0, \sigma_n^2)$, to acknowledge that the measurement will be imprecise. In matrix notation, one can write this as a probability distribution:

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{b}, \sigma_n^2 \boldsymbol{I_N}) \tag{1.12}$$

To identify the optimal set of parameters $\boldsymbol{\beta}*$ one can maximize the likelihood of this distribution or, equivalently, minimize the sum of the residuals $\boldsymbol{y} - \mathbf{X}\boldsymbol{b}$, which through analytical optimization can be shown to be the ordinary least squares solution:

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{1.13}$$

The power of this linear model lies in its ability to encode arbitrary designs in $\boldsymbol{X}$ that are not limited to continuous covariates. For example, in a two-group design $\boldsymbol{X} = (0, ..., 0, 1, ..., 1)$, $\boldsymbol{\beta}$ will measure the estimated group difference and statistical testing in this model is equivalent to the t-test (analogously for multi-group designs and ANOVA). Also, non-additive interactions between covariates can be encoded using interaction effect designs.

However, this model is appropriate only for continuous data with normally distributed residuals. To make it applicable to count-based data, two modifications have to be made: First of all, one no longer directly observes the dependent variable in the linear model, but replaces the observations $\boldsymbol{y}$ with a latent variable that will be connected to the rate parameter of the noise distribution. Finally, when the rate is only defined on a subset of the real numbers, an invertible

link function has to be introduced that maps $\boldsymbol{y}$ to the rate. The resulting model is called a *generalized* linear model, and for Poisson-noise is given by

$$\boldsymbol{k} \sim Poi(\boldsymbol{\lambda}) \tag{1.14}$$
$$\phi(\boldsymbol{\lambda}) = \boldsymbol{\mu} \tag{1.15}$$
$$\boldsymbol{\mu} = \sum_{n=1}^{C} \beta_i \boldsymbol{x}_i \tag{1.16}$$

where the link function $\phi(x) = log(x)$ applied element-wise transforms μ to the positive real numbers. A similar model that uses a negative binomial noise model underlies most methods for differential gene expression (*edgeR* and *DESeq2*). These models use empirical Bayes estimation to stabilize the over-dispersion parameter estimates across many genes, since the number of replicates in most experiments is too low for reliable inference [Love et al., 2014b]. For binomial data, the rate parameter p is restricted to $(0, 1)$, so the logit transformation $\phi(x) = log(\frac{x}{1-x})$ is usually used. This leads to the model

$$\boldsymbol{k} \sim Bin(\boldsymbol{n}, \boldsymbol{p}) \tag{1.17}$$
$$\phi(\boldsymbol{p}) = \boldsymbol{\mu} \tag{1.18}$$
$$\boldsymbol{\mu} = \sum_{n=1}^{C} \beta_i \boldsymbol{x}_i \tag{1.19}$$

This is also known as the *logistic regression* model, with $k \in 0, 1$. In this model, p is the success probability of the binomial, and the link function converts it into log odds $log(\frac{p}{1-p})$. The linear model coefficients can be interpreted as a fold change in log odds, since for a one-parameter model

$$\beta_1 = (\beta_1 + \beta_0) - \beta_0 \tag{1.20}$$
$$= \mu_2 - \mu_1 \tag{1.21}$$
$$= log(\frac{p_2}{1-p_2}) - log(\frac{p_1}{1-p_1}) \tag{1.22}$$
$$= log(\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}) \tag{1.23}$$

When accounting for overdispersion, one can replace the binomial with a beta-binomial and introduce an overdispersion parameter vector $\boldsymbol{\theta}$ we get

$$\boldsymbol{k} \sim BetaBin(\boldsymbol{n}, \boldsymbol{p}, \boldsymbol{\theta}) \tag{1.24}$$
$$\phi(\boldsymbol{p}) = \boldsymbol{\mu} \tag{1.25}$$
$$\boldsymbol{\mu} = \sum_{n=1}^{C} \beta_i \boldsymbol{x}_i \tag{1.26}$$

which will be the core model to describe allelic read counts in this thesis.

To test for significant associations between covariates and read counts, one can use hypothesis testing. The null hypothesis is that a parameter of interest β is equal to zero, while the alternative is that it is not:

$$H_0 : \beta = 0 \tag{1.27}$$
$$H_1 : \beta \neq 0 \tag{1.28}$$

In gLMs, statistical testing uses the likelihood function of the model and evaluates its difference between the null and alternative model. This difference can be quantified by different test statistics S, whose distribution under the null is known. One can then evaluate the probability to observe a test statistic as extreme as the observed one under the null and define a level at which to make the decision to be confident enough to reject the null. This probability is known as the p-value and is often arbitrarily set at 0.05. There are three commonly used test statistic in gLMs:

The Wald test finds the maximum likelihood solution $\boldsymbol{\beta_{MLE}}$ under the alternative model and compares it to a null estimate $\boldsymbol{\beta_0}$:

$$S = (\boldsymbol{\beta_{MLE}} - \boldsymbol{\beta_0})^T (var(\boldsymbol{\beta^*})^{-1})(\boldsymbol{\beta_{MLE}} - \boldsymbol{\beta_0}) \tag{1.29}$$

Intuitively, this is similar to a t-test, where the mean difference to the null is scaled by the variance estimate. It can be shown under some assumptions that $S \sim \chi^2_d$, where d is the number of parameters in $\boldsymbol{\beta}$ and corresponds to the degrees of freedom of the χ^2 distribution [Love et al., 2014b].
More generally, a likelihood ratio test can be used. Here, one directly compares the likelihood of null to the alternative model given the data with using the test statistic

$$S = log(\mathcal{L}_{alt}/\mathcal{L}_{null}) \tag{1.30}$$

is used after maximum likelihood estimation. It is likewise distributed as $2S \sim \chi^2_d$.
Finally, a score test can be used. The score statistic is given by calculating the derivative of the likelihood function under the null:

$$S = \sigma(\beta_0)^T [var(\beta_0)]^{-1} \sigma(\beta_0) \tag{1.31}$$

where $\sigma(\beta) = \frac{\delta \mathcal{L}}{\delta \beta}$. The score test has the advantage that it does not require the MLE of the alternative model to be computed, which can be more complex than the null. Again the score is asymptotically distributed as $S \sim \chi^{2,d}$ where d is the number of parameters in $\boldsymbol{b}$ [Heinen et al., 2021]. All of these statistical tests provide a *p-value* which represents the probability under the null model to observe a test statistic that is at least as extreme as the observed one. If this is unlikely, the most common threshold is $p < 0.05$, one assumes that H0 is wrong. If the null model is correct, p will be uniformly distributed, which means that if many tests are performed at a p-value cut-off of 0.05, 5% of these will yield a significant result even when the null is true in every case (Type II error). This is called the multiple testing problem, and requires taking stricter p-values to retain confidence in their validity (the

changed p-value is also called an adjusted p-value p_{adj}). One approach is to multiply each p-value p by the number of performed tests. This means that a cut-off $p < 0.05$ now controls the probability of rejecting the null wrongly once among all test. This *Bonferroni-procedure* leads to extremely conservative p-values. In sequencing data analysis, the more common approach is to control the false-discovery rate (FDR) directly using the *Benjamini-Hochberg* method [Benjamini and Hochberg, 1995]. Here, instead of multiplying by the number of tests, each p-value is adjusted as $p_i^{adj} = \frac{Np_i}{r_i}$ where N is the number of tests and r_i is the rank of p_i in a list that orders the p-values from lowest to highest. By choosing the largest adjusted p-value $p_j^{adj} < 0.05$, this approach ensures at most 5% of type II errors, which is acceptable in many cases. It can be immediately seen that every adjusted p-value except p_1 is lower than the Bonferron-corrected one, which makes this procedure less prone to type I errors (wrong rejection of H1).

1.4.4 Allelic fold change models

One recently introduced alternative approach shows how allelic ratios can be modelled using standard gaussian or poisson statistics by introducing a linear model with an allele-specific covariate [Mohammadi et al., 2017]. When working with gaussianized expression data, this can be written as

$$log(\mathbf{y}_i) = \beta_0 + \beta_A \mathbf{x}_{i,A} \tag{1.32}$$

β_A captures the fold change difference in expression from the two alleles, which can be extended to joint modelling of multiple alleles. By using interaction terms, this approach can immediately be used to assess differences in allelic usage, for example through

$$log(\mathbf{y}_i) = \beta_0 + \beta_A \mathbf{x}_{i,A} + \beta_C \mathbf{x}_{i,C} + \beta_{AC} \mathbf{x}_{i,AC} \tag{1.33}$$

Here, β_C will denote the expression differences between covariate c and β_{AC} will capture its differential allelic usage. This approach should be advantageous especially when the allelic ratio is close to being fully biased, which places the binomial rate close to its parameter boundaries. It also provides a somewhat more intuitive interpretation of the allelic difference as an expression fold change. However, the power of this approach has not been directly compared to the commonly used binomial model.

1.4.5 Bayesian reasoning in statistical models

Before continuing to discuss extensions of gLMs, I will briefly cover Bayesian statistical models. One can consider the Bayesian approach as an alternative to the (frequentist) reasoning used in hypothesis-based testing, where the focus is on assessing the significance of individual parameters in some hypothesis test. In contrast, Bayesian analysis focusses on the models itself and uses a full probabilistic description, including all parameters. This is done by replacing each parameter with a probability distributions that encodes prior expectations. This prior is then combined with the model assumptions (*likelihood*) using Bayes theorem to yield the posterior distribution of the parameters $\boldsymbol{\theta}$, given an observed dataset D:

$$p(\boldsymbol{\theta}|D) = \frac{L(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)} \tag{1.34}$$

$p(D)$ is the probability of observing the data after integrating over all possible model parameters $\boldsymbol{\theta} \in \Theta$. Intuitively, this means that our model includes an expectation over the distribution of θ that is "updated" by observing D. Therefore $\boldsymbol{\theta}$ retains information about the entire space of possible solutions including their uncertainty instead of a point estimate. To use this idea practically, a common approach would be to compare the posterior $\boldsymbol{\pi}(M|D)$ two models M_0, M_1 with different parameter specifications accounting for the prior $\boldsymbol{\pi}(M)$:

$$B = \frac{\boldsymbol{\pi}(M_1|D)}{\boldsymbol{\pi}(M_0|D)} \Big/ \frac{\boldsymbol{\pi}(M_1)}{\boldsymbol{\pi}(M_0)} \tag{1.35}$$

$$= \frac{\boldsymbol{\pi}(\theta_1|D)}{\boldsymbol{\pi}(\theta_0|D)} \Big/ \frac{\boldsymbol{\pi}(\theta_1)}{\boldsymbol{\pi}(\theta_0)} \tag{1.36}$$

$$= \frac{L(D|\theta_1)}{L(D|\theta_0)} \tag{1.37}$$

B is called the *Bayes Factor* and is equal to the likelihood ratio of the two models. Of note, this approach directly accounts for and penalizes model complexity, which is difficult with frequentist methods. The disadvantage is that calculation of the posterior is often analytically intractable and has to be replaced by computationally intensive sampling methods [Argelaguet et al., 2018].

1.4.6 Accounting for structure using gLMMs

The final extension to linear models relevant to this thesis comes in the form of *mixed* models. These are generally used when there is additional information about the structure of samples that extends beyond simple covariates. The idea is that we can encode this information as a covariance matrix, so given fixed effect covariates $\boldsymbol{X}$ and associated parameters $\boldsymbol{\beta_X}$, our covariate of interest $\boldsymbol{z}$ with parameter β we can write

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta_X} + \boldsymbol{z}\beta, \sigma_u^2\boldsymbol{K} + \sigma^2\boldsymbol{I_N}) \tag{1.38}$$

where $\boldsymbol{K}$ is a covariances matrix encoding a *random effect* and σ_u^2 controls its strength. While this examples illustrates normally distributed data, this approach can be linked to counting noise models as in (24), transforming the linear mixed model into a *generalized linear mixed model* (gLMM).

This approach has two main applications: First of all, $\boldsymbol{K}$ can represent known confounding structure in the data that one wants to account for. One well known example are LMMs for genetic association tests in the context of quantitative trait locus mapping [Yu et al., 2006], where $\boldsymbol{K}$ controls for relatedness between subjects. A second example are gLMMs used for the analysis of single-cell data. When comparing multiple replicates, cells within the same sample will share some correlation structure and will not be independent. When not accounted for, this leads to a pseudo-replication problem and inflated test statistics. Using a random effect encoding sample identity, this can be accounted for [Zimmerman et al., 2021]. There is currently an active debate over whether in practice, this approach provides an advantage over simple summation across similar cells ("pseudo-bulk" analysis) for differential expression analysis between multi-condition single cell datasets. It should be noted that sample identity in this case can also be given as a fixed effect.

The second broad use case of mixed models are variance component tests to associate a given structure in the data with the dependent variable, rather than regressing it out. Mathematically, this is realized using the hypothesis test $H_1 : \sigma_u^2 \neq 0$ and can be done using the parameter tests outlined in the previous section. Of note this approach allows for a relative quantification of the variance explained, since for a model with $M - 1$ random effect components

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{X}}, \sigma_{u,1}^2\boldsymbol{K_1} + \sigma_{u,2}^2\boldsymbol{K_2} + ... + \sigma_N^2\boldsymbol{I_N}) \tag{1.39}$$

we can compute the fraction of explained variance for K_i as $\rho_i = \frac{\sigma_i^2}{\sum_1^{M-1}\sigma_i^2+\sigma^2}$ and in particular the total non-technical variance as $\rho_{tot} = \frac{\sum_1^{M-1}\sigma_i^2}{\sum_1^{M-1}\sigma_i^2+\sigma^2}$.

One recently published model called *scDALI* provides such a gLMM for the analysis of allele-specific single-cell sequencing data and to test for cell type-dependency thereof [Heinen et al., 2021]. Given counts (k_i, n_i) and a covariate matrix $\boldsymbol{X}$ the scDALI model can be cast as

$$\boldsymbol{k} \sim BetaBin(\boldsymbol{n}, \boldsymbol{p}, \theta) \tag{1.40}$$

$$\phi(\boldsymbol{p}) = u \tag{1.41}$$

$$u \sim \mathcal{N}(\boldsymbol{u_0} + \boldsymbol{X}\boldsymbol{a}, \sigma_{hom}^2\boldsymbol{1}\boldsymbol{1}^T + \sigma_{het}^2\boldsymbol{K}) \tag{1.42}$$

Here, two variance components assess the evidence for differences in allelic imbalance relative to a null imbalance $\boldsymbol{u_0}$ (usually $\boldsymbol{u_0} = \boldsymbol{0} \implies \boldsymbol{p} = \boldsymbol{0.5}$) after accounting for the covariates. The homogeneous (persistent) imbalance that is shared across cells is quantified by σ_{hom}^2, while a heterogeneous imbalance is captured by σ_{het}^2 and depends on a cell type structure given by $\boldsymbol{K}$. scDALI uses score tests to assess the significance of the two components or their combination. The pseudo-replication issue is solved by introducing a sample-level covariate into $\boldsymbol{X}$. Accounting for overdispersion can be done either globally or on the gene level. In practice, the authors estimate a shared overdispersion parameter θ across all features which is used for all feature-level tests.

1.4.7 Gaussian process regression

This use of LMMs to encode structural information in the data is closely linked to gaussian process regression (GPR). GPR assumes the data is distributed as a multivariate Gaussian with a structured covariance matrix:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{x}), \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})) \tag{1.43}$$

where $\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{x})$ is called a mean function and $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})$ is called the *kernel function* which encodes our prior belief over the relationship between individual datapoints. A popular choice is the radial basis function (RBF) kernel, given by

$$K(x_i, x_i) = exp(-\frac{(x_i - x_j)^2}{\sigma^2}) \tag{1.44}$$

which is high when points are close in $\boldsymbol{x}$ and therefore enforces smoothness in the fit, depending on the length scale hyperparameter σ. However, kernels can also explicitely encode parametric functions such as constant, linear, polynomial or periodic [Rasmussen, 2004].

Formally, the GP specifies a prior on the functions that we want to fit to the data. For this reason, the normal distribution in (1.43) is actually infinitely-dimensional, as it spans the whole input value space $\boldsymbol{x}$. To use this prior to make predictions on experimental data of finite dimensionality, one uses the joint distribution over the training and test points $\boldsymbol{f}, \boldsymbol{f_x}$:

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f_x} \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu_x} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma_x} \\ \boldsymbol{\Sigma}_x^T & \boldsymbol{\Sigma_{x,x}} \end{bmatrix}) \tag{1.45}$$

This is using element-wise vector notation for the training and test mean-functions $\boldsymbol{\mu}, \boldsymbol{\mu_x}$ and denote the covariances with and between training and test data as $\boldsymbol{\Sigma}, \boldsymbol{\Sigma_{x,x}}$ as well as $\boldsymbol{\Sigma_x}$. To predict the unknown distribution of test data based on training data, we write the conditional distribution $\boldsymbol{f}|\boldsymbol{f_x}$:

$$\boldsymbol{f}|\boldsymbol{f_x} \sim \mathcal{N}(\boldsymbol{\mu_x} + \boldsymbol{\Sigma_x^T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{f} - \boldsymbol{\mu}), \boldsymbol{\Sigma_{x,x}} - \boldsymbol{\Sigma_x^T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma_x}) \tag{1.46}$$

GPR therefore represents the usage of a prior on the functions that $\boldsymbol{f_x}$ defines. The output is a posterior distribution with credible function values defined by the input data. If one assumes gaussian noise on the data measurement, as done in linear regression, this amounts to adding additional variance to the GP covariance:

$$\begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon_x} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} \sigma^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0}^T & \sigma^2\boldsymbol{I} \end{bmatrix}) \tag{1.47}$$

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f_x} \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu_x} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} + \sigma^2\boldsymbol{I} & \boldsymbol{\Sigma_x} \\ \boldsymbol{\Sigma}_x^T & \boldsymbol{\Sigma_{x,x}} + \sigma^2\boldsymbol{I} \end{bmatrix}) \tag{1.48}$$

GPR models can be fit by optimizing the log-likelihood, which involves optimization of hyper-parameters included in the kernel functions. Their main benefits are their flexibility in describing non-linear patterns in data, while being robust to overfitting, their main disadvantage is computational complexity since computation of the covariance scales quadratically with the number of input data points. In particular, GPs have been useful to describe spatiotemporal data, where the covariance function encodes similarity of close data points. Indeed, one of the main applications for GPs has been geostatistical data analysis, to which spatial transcriptomics has obvious parallels [Stegle et al., 2010, Svensson et al., 2018, Velten et al., 2022]. Similarly to the extension of linear models using gLMs, GPs can be applied to non-gaussian data by coupling the function modelled by a GP to a noise model [BinTayyash et al., 2021].

1.5 Aims and outline of this thesis

Overarchingly, this thesis explores how single-cell resolved measurements of allele-specific expression (and chromatin accessibility) can be used to inform the investigation of gene

regulation. I make use of interspecific mouse hybrids, where two inbred strains with a high density of single-nucleotide variants (around 1 / 100bp) are crossed and sequencing can assign reads to paternal haplotypes based on the resulting heterozygous variants. Using single-cell readouts, I can then probe ASE at cellular resolution, link it to the total expression state of the cell, and both characterize its cell type-specificity and use cell-to-cell variation to infer regulatory mechanisms. In **Chapter 2**, I first probe how genetically driven ASE varies during cellular differentiation *in vivo*, for which I use spermatogenesis as a model system. I find that interactions between cell type and ASE are remarkably pervasive, and demonstrate that between mouse sub-species, cell type-specific transcriptional divergence is mainly driven by *cis*-genetic effects. **Chapter 3** moves to explore allele-specific expression driven by X-chromosome inactivation. I develop an approach to classify X-inactivation haplotypes which allows us to quantify escapee expression at the single-cell level, revealing unappreciated variability in escape among immune cells. I use this approach to test the hypothesis that ageing leads to loss of integrity of XCI and increased escape, and observe that this is only the case in clonally expanding lymphocytes. In **Chapter 4**, I finally attempt to provide a potential mechanism of variable escape, by showing that the *Xist* long-noncoding RNA can silence escapees chromosome-wide. I also use this system to quantify the resilience of escapees to silencing. These results showcase the power of coupling allele-specific expression analysis to single-cell readouts. I will finally discuss future directions, such as perturbation experiments with allelic readout, as well as challenges in transferring these approaches to human samples.

1.6 Publications

The results in this thesis are part of the following publication:

Jasper Panten, Tobias Heinen, Nils Eling, Christina Ernst, Rebecca E. Wagner, Maja Satorius, John Marioni, Oliver Stegle*, Duncan T. Odom*; **The dynamic genetic determinants of increased transcriptional divergence in spermatids.** *Nature Communications (accepted, equal contribution)*

Manuscripts on the results in **Chaper 2 / 3** are currently in preparation.

1.6.1 Other contributions

Contributions to published papers or preprints that are not covered in this thesis:

Ivana Winkler, Alexander Tolkachov, Fritjof Lammers, Perrine Lacour, Nina Schneider, Marie-Luise Koch, Jasper Panten, Florian Grünschläger, Klaudija Daugelaite, Tanja Poth, Simon Haas, Duncan T. Odom, Angela Goncalves; **The function and decline of the female reproductive tract at single-cell resolution.** *2022, biorxive*

Jasper Panten*, Stefania Del Prete*, James P. Cleland*, Lauren M. Saunders, Job van Riet, Anja Schneider, Paul Ginno, Nina Schneider, Marie-Luise Koch, Moritz Gerstung, Oliver Stegle, James M. A. Turner, Edith Heard*, Duncan T. Odom*. **Four-Core Genotypes mice harbour a 3.2MB X-Y translocation that perturbs Tlr7 dosage.** *2023, biorxive*.

Chapter 2

The dynamic genetic determinants of increased transcriptional divergence in spermatids

Overview: *Cis*-genetic effects are known to drive gene expression changes in tissues and discrete cell types. However, it is still poorly understood how *cis*- and *trans*-effects act across continuous trajectories of cellular differentiation *in vivo*. Here, I use an F1 hybrid mouse system to quantify allele-specific expression during spermatogenesis at the single-cell level, which allows me to comprehensively quantify *cis*- and *trans*-genetic effects, and in particular their dynamic changes across cellular differentiation. In total, I show that almost half of the genes subject to genetic regulation show evidence for dynamic *cis*-effects that vary during the differentiation process. My approach also allows me to robustly identify dynamic changes in *trans*-effects, which are substantially less common than *cis*-driven ones. I demonstrate that genetic effects shows the strongest effect sizes in round spermatids, which contributes to their increased transcriptional divergence that we detect across multiple species. My work quantifies genetic effects in an example of cellular differentiation *in vivo*, and demonstrates a generalizable strategy to dissect the impact of regulatory variants on gene regulation in dynamic systems.

Contributions: This study is joint work between the Stegle and Odom labs. The project was conceived by Oliver Stegle, Duncan T. Odom and me. Tobias Heinen developed the scDALI model and software. Christina Ernst and Nils Eling generated and performed preliminary analysis of the cross-species dataset in Figure 5, under supervision by John C. Marioni. Rebecca E. Wagner contributed to functional genomics experiments and Maja Satorious assisted with the generation of scRNA-Seq libraries. I generated all other data, performed computational analysis and designed all figures. Duncan T. Odom and Oliver Stegle supervised the study and wrote the manuscript with me. The paper has been accepted for publication as:
Jasper Panten, Tobias Heinen, Nils Eling, Christina Ernst, Rebecca E. Wagner, Maja Satorius, John C. Marioni, Oliver Stegle, Duncan T. Odom. **The dynamic genetic determinants of increased transcriptional divergence in spermatids.** *Nature Communications*

2.1 Introduction

The probably most common cause of allelic imbalance in gene expression is the presence of regulatory variants in the genome that differentially affect the two copies of a gene [Cleary and Seoighe, 2021]. In recent years, it has become increasingly clear that the interactions of these variants with the cell type-specific regulatory machinery are critical to understand when and how they act, including for trait- and disease-associated variants [Kim-Hellmuth et al., 2020]. However, we still lack a complete understanding about how these interactions function and how pervasive they are. Recently, single-cell based approaches have expanded the ability to fine-map these effects to specific cell types with unprecedented resolution, but have also been shown to capture more subtle variation between sub-celltypes and cell states. These studies mapping the context-specificity of common genetic variants show that dynamic effects are the rule rather than the exception. A specific interest has been the detection of differentiation-dependent genetic effects during developmental processes, which might be critical for phenotypic variation [Zhernakova et al., 2017, Jerber et al., 2021, Cuomo et al., 2020, Strober et al., 2019, Sarkar et al., 2019, van der Wijst et al., 2020, Elorbany et al., 2022].

While working with human data has obvious practical interest to understand the genetic effects underlying disease and general phenotypic variability, samples are generally difficult to obtain, limited to accessible specimens like blood or post-mortem tissues. Furthermore environmental exposure effects and the complex genetics of outbred populations introduce variability, necessitating large sample sizes to perform powered mapping of genetic effects. Therefore, studies in model organisms have purpose in particular to understand the general architecture underlying sequence-to-phenotype links or to nominate molecular mechanisms that turned out to be conserved in humans [Albert and Kruglyak, 2015]. An advantage is that experiments using model organisms can be designed to have defined genetics. For example, the collaborative cross generated offspring by crossing eight inbred, fully homozygous mouse strains that can be used to map the effects of parental genotypes that are randomized through meiotic recombination [Threadgill et al., 2011]. The known set of variants combined with the consistent environmental conditions in these animals make studies using the mice more powerful at the same sample number than for example human studies. This suggests that these approaches should be beneficial to detect *trans*-effects and context-dependent effects, that require larger power. To further reduce the complexity, one can perform an outcross between two species only. In the F2 generation meiotic shuffling will then allow to perform QTL mapping of all differences between the founder strains. These studies have revealed development and cell type-specific action of *cis*- and *trans*-effects, in particular in nematodes [Francesconi and Lehner, 2014, Ben-David et al., 2021].

An even more simplified system is provided by generating an F1 cross between two homozygous founders. In this setting, *cis*-effects can be read out directly by assessing allelic imbalance in the hybrid animal. Additionally, one can compare the allelic usage in the hybrid to the expression differences in the two founder strains. Since, as described earlier, allelic imbalance isolates the *cis*-driven portion expression divergence since the two alleles are measured in the same *trans*-regulatory environment, the difference between allelic ratios between F1 and F0 has to be driven in *trans*. This approach has been powerful to demonstrate that the vast majority of gene expression divergence between closely related species is driven by *cis*-effects, that is, by the direct action of sequence changes [Wittkopp et al., 2004, Wittkopp and Kalay, 2011, Goncalves et al., 2012, Tirosh et al., 2009].

However, it is not known if context-specificity in genetic effects is primarily driven by interactions of *cis*-effects with the cellular environment or context-specific *trans*-effects.

Of note, this analysis directly interrogates the expression divergence that drives phenotypes between sub-species that can still be hybridized. This setup is therefore a powerful approach to understand how small genetic changes drive changes in expression regulation and thereby novel phenotypes. It has been suggested that *trans*-effects are more likely to be cell type-specific than *cis*-effects, as the presence of trans-actors will depend on the cell type, but this has not been comprehensively addressed [Yazar et al., 2022].

Here, I sought to develop an approach that comprehensively maps the variability of genetic effects during differentiation *in vivo*. I achieve this by combining an F1 hybrid approach with single-cell RNA-Sequencing, which allows me to dissect the *cis*- and *trans*-contributions to transcriptional divergence at high resolution.

2.2 Separation of *cis*- and *trans*-effects in F1 hybrid designs

The F1 hybrid trio model has been originally introduced in *Wittkopp, 2004* and is based on the following assumptions: When allele-specific expression of two alleles is measured from the same cell, one can assume that the sum of *trans*-factors acting on expression of each allele is the same **(Fig. 2.1)** [Wittkopp et al., 2004]. Therefore, any allelic difference detected will be caused by *cis*-acting effects, which impact the allele that is on the same chromosome as the variant. In this way, measuring ASE in F1 hybrids shows the sum of all *cis*-acting effects on a gene, and measuring ASE between F0s gives the sum of aggregate *cis* and *trans*-effects. By combining this setup with single-cell resolved measurements of expression, I can now measure *cis*-effects at single-cell resolution and *trans*-effects by comparing cells in similar transcriptional states across the F0s and the F1. Consequently, this approach quantifies genetic effects across (continually changing) cell states, and therefore characterizes the dynamics of genetic regulation.

I chose spermatogenesis as a model differentiation process for multiple reasons. Most importantly, it represents a simplified unidirectional and highly continuous process that has been shown to be captured well in single-cell transcriptomics and involves minimal tissue treatment during the experimental protocol [Ernst et al., 2019]. Secondly, spermatogenesis is known to show rapid transcriptional divergence between closely related and distant species [Murat et al., 2022, Shami et al., 2020, Soumillon et al., 2013, Brawand et al., 2011, Kopania et al., 2022].

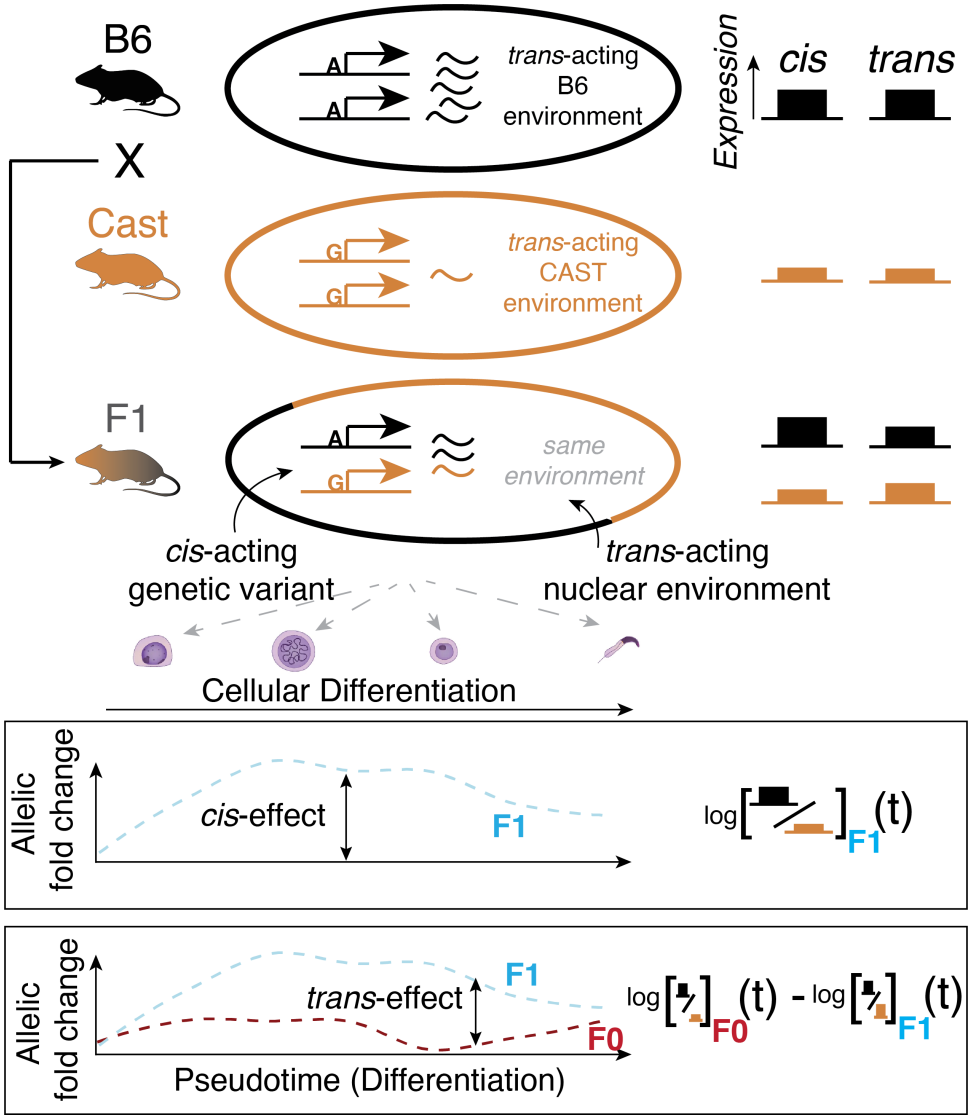

Figure 2.1: Conceptual outline of the project. I am using a previously established system to dissect *cis*- and *trans*-effects. By coupling it to single-cell expression readouts, I can then quantify variability in these effects across cell types. Additionally, the cells can be ordered according to their position in the differentiation timeline using the whole transcriptome information.

2.2.1 Single-cell RNA-Sequencing of testis cells from an inter-specific F1 mouse cross and the parental strains

To analyse the context-specificity of genetic effects in primary mouse tissues, I generated single-cell RNA-Sequencing data from 6 mice each of the standard lab strain C57BL6 (specifically C57BL6-Ly5.1, abbreviated B6 from here), the wild-derived CAST/EiJ (CAST) strain and their immediate F1 cross (2x B6xCAST, 4x CASTxB6 where the first strain denotes the female in the cross) **(Fig. 2.2a)**. C57BL6 most closely represents the *Mus musculus domesticus* sub-species, while CAST is *Mus musculus castaneus*, with approximately 500 thousand years seperating the two sub-species [Wong et al., 2015]. Previous whole genome sequencing efforts identified 17361111 million SNVs between the two strains, corresponding of a SNP density of 6.94 SNPs / kb [Keane et al., 2011]. The experiments were performed in 3 independent runs, where each run consisted of 2 samples per strain.

I first processed droplet-based scRNA-Seq data from testis in an analogous way as previously published [Ernst et al., 2019]. To allow for an integrated analysis of gene expression levels across strains, I used the standard mm10 genome (derived from the B6 strain) as a reference where the CAST-derived SNPs were N-masked to reduce mapping bias. In order to facilitate cross-comparisons between the different species, I used mutual-nearest-neighbour (MNN) integration methods across all individual libraries to define a joint gene expression space [Haghverdi et al., 2018]. Upon visualization using Uniform Manifold Approximation and Projection (UMAP), samples were found to readily integrate within and across species **(Fig. 2.2b)**. I then used clustering and marker gene expression to classify cell types in the dataset. In the UMAP-projection, germ cells formed a stereotypical one-dimensional differentiation trajectory, allowing for the annotation of different spermatogenic sub-stages, including meiotic stages (Leptotene, Diplotene, Meiosis I and II) and different stages of Spermiogenesis (Round Spermatids 1-6 and Elongating Spermatids 1-6) **(Fig. 2.2c)**.

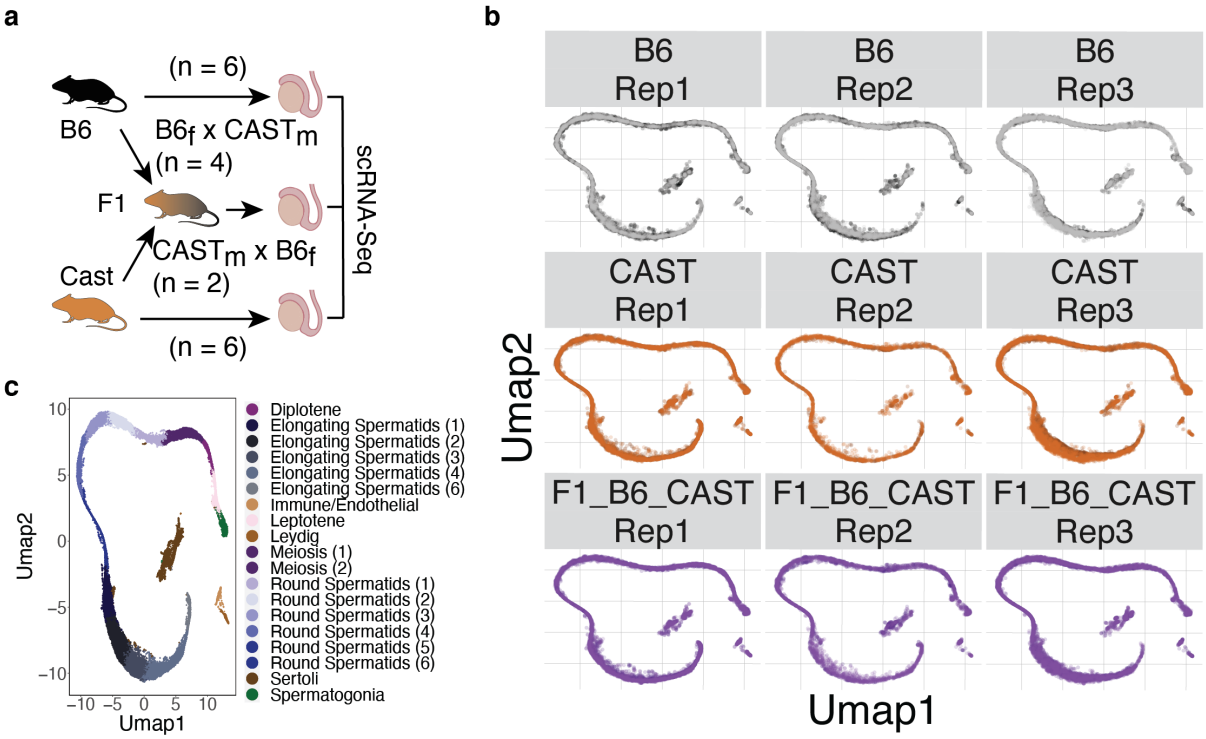

Figure 2.2: Integration of F0 and F1 testis datasets. (a) Experimental outline of the data generation. B6 and CAST mice are crossed into a hybrid, and 6 individual animals per experimental group are processed using 10x-based scRNA-Seq. **(b)** UMAP of MNN-integrated samples, split by experimental replicates (two biological replicates each). **(c)** Joint UMAP across all samples, with annotated cell types.

Across libraries, the dataset contained 54,863 cells split into 1,202-4,361 cell per library **(Fig. 2.3a)**. Quality control metrics such as genes detected per cell and cell type marker gene expression were highly reproducible within and across species **(Fig. 2.3b, d)**. In particular, I validated marker gene expression for Sertoli (marker gene *Cst12, Cldn11*), Leydig (*Insl3, Fabp3*) and endothelial / immune cells (*Acta2*)) as well as germ cells (spermatogonia (*Dazl, Stra8, Dmrt1*), spermatocytes (*Hormad1, Piwil1, Pou5f2, Aurka*), round spermatids (*Tex21*) and elongating spermatids (*Prm1, Atg7, Akap4, Tnp1*)). Finally, cell type compositions were consistent across samples, with spermatids (RS, ES) and spermatocytes (SC) making up the majority of sequenced cells **(Fig. 2.3c)**. These results are in line with previous single-cell

surveys and textbook knowledge of testicular cell types, indicating a high quality dataset with strong technical reproducibility ([Ernst et al., 2019]). Furthermore, they demonstrate the functional conservation of core gene expression patterns across closely related mammals. For all subsequent analysis, I removed the somatic cell types and focussed on the differentiation of spermatocytes and spermatids.

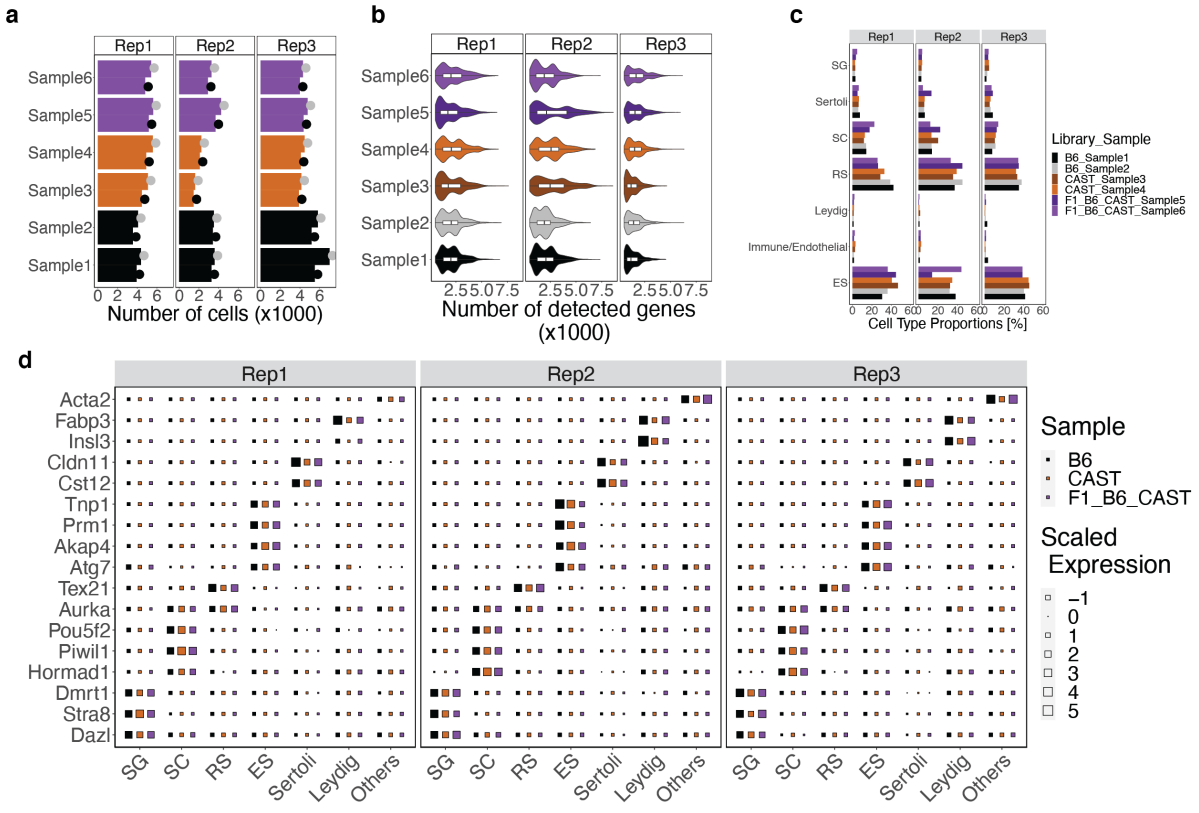

Figure 2.3: Dataset metrics across species and experiments. (**a**) Barplot showing the number of cells per experimental replicate before (grey dot) and after (black dot) filtering out technical outliers. (**b**) Violin plot showing the distribution of detected genes across experimental replicates. (**c**) Barplots showing the cell type distributions per experimental replicate. (**d**) Dotplot showing gene expression of marker genes across replicates. For each gene, the normalized gene expression is aggregated across cells per cell type and library.

2.2.2 Mapping of allele-specific gene expression from droplet-based scRNA-Seq data

In the second step, I aimed to verify that allele-specific gene expression could be quantified in a robust and reproducible manner from droplet-based scRNA-Seq data. Expression profiles generated using the 10x genomics platform are 1) sparse (median 4602 UMI / cell and detected 1791 genes / cell in the dataset) and 2) shows a strong 3' bias, as the final sequencing library only contains fragments that include the cell and UMI barcodes at the end of the transcript, both of which limit the detectability of allele-specific expression. I therefore sought to assess the resolution of allele-specific signals in the data. Using SNV sets between B6 and CAST I determined which reads were derived from the maternal and the paternal haplotypes and performed the same quantifications in the parental strains as a control. Reassuringly, virtually all (>99.7%) allele-resolved reads in the parental strains mapped to the expected parental

genotype, while in the F1 mice, approximately half of the reads were derived from each strain **(Fig. 2.3a-d)**. Furthermore, I found that maternally inherited chromosomes (X and mitochondrial) were only expressed from the maternal haplotype in F1 mice **(Fig. 2.3f)**. In total, 25.82% of reads were assignable to a parental haplotype, which corresponded to robust allelic quantifications of 6,495 genes, (54.98% of detected genes at more than 50 reads per sample). In total, these results indicate that 10x Chromium-based single-cell RNA-Seq is a valuable tool for the detection of allele-specific signals in complex tissues, although there is a limitation in the detection of lowly expressed genes.

During differentiation, germ cells undergo ploidy changes from 2n (spermatogonia) to 4n (at the beginning of Meiosis I) and again to 2n (after the first meiotic reduction) and 1n (in round spermatids, after both meiotic reductions) [Ernst et al., 2019]. From the DNA content, it would therefore be expected that spermatids should show gene expression only from a single parental haplotype. However, I observed chromosome-wide bi-allelic gene expression in individual round and elongating spermatids, similar to di- and tetraploid cells **(Fig. 2.3e)**. This is due to the fact that spermatids do not undergo full cytokinesis in the second meiotic division, but remain connected through ”cytoplasmic bridges” that allow transfer of proteins and mRNA between sister cells. From an evolutionary perspective, this phenomenon is thought to provide essential X-linked gene products to cells with a Y-chromosome, thus avoiding reduced fitness of sperm with different sex chromosomes [Bhutani et al., 2021]. In this context, I can therefore treat spermatids as functionally diploid when assessing allele-specific expression as I cannot distinguish whether transcripts are derived from a neighbouring cell.

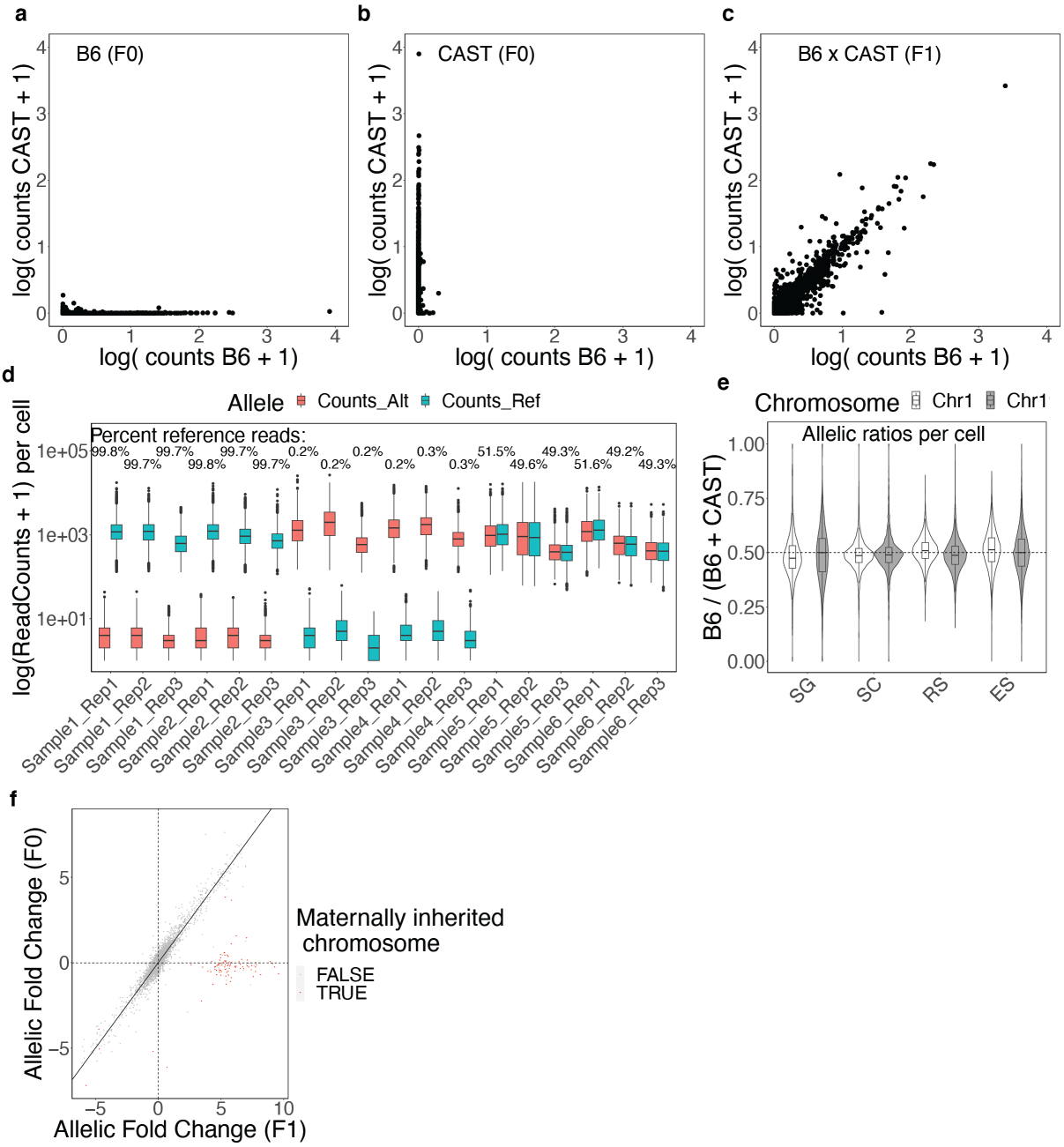

Figure 2.4: Quantifying allele-specific expression in mouse germ cells. (**a-c**) Scatterplot showing allele-specific read counts in B6 F0 (**a**), CAST F0 (**b**) and hybrid mice (**c**). Allele-specific counts are aggregated across all samples. (**d**) Boxplots showing allelic ratios per cell. Alternative (Alt) counts are CAST-dervied, reference (Ref) from B6. Percentages above the plots indicate the proportion of reference counts. **e** Violin plots showing allelic ratios per cell for chromosomes 1 and 10 in spermatogonia (SG), spermatocytes (SC), round spermatids (RS) and elongating spermatids (ES). Note that haploid cells do show diploid expression, similarly to di- and tetraploid cells. **f** Scatterplot show log allelic fold changes $\frac{B6}{CAST}$ in F1 against F0 for samples from the forward cross (female B6 x male CAST). Genes on maternally (B6) inherited chromosomes show clear bias in the F1, but not in the F0.

2.2.3 *Cis*- and *trans*-regulatory contributions to strain-specific expression in testis

Having established the high quality of cell state, gene expression and allelic signals in the dataset I next moved to analyse expression differences between F0 animals of the B6 and CAST strains. Using DESeq2, I tested for differential expression across pooled whole germ cell pseudobulk, as well as separated pseudobulk libraries of spermatocytes, round and elongating spermatids [Love et al., 2014b]. I furthermore characterize the strength of differential expression using the log fold change between allelic read counts x_{B6} and x_{CAST} (allelic fold change, [Mohammadi et al., 2017]):

$$aFC = log2\frac{x_{B6}}{x_{CAST}} \tag{2.1}$$

Across the different comparisons, this analysis demonstrates widespread transcriptional changes between both strains, detecting between 27.13 - 46.3% of genes as differentially expressed (Benjamini-Hochberg adjusted p-value < 0.1) and 12.23 - 20.01% with an absolute allelic fold change of at least 1 **(Fig. 2.5e)**.

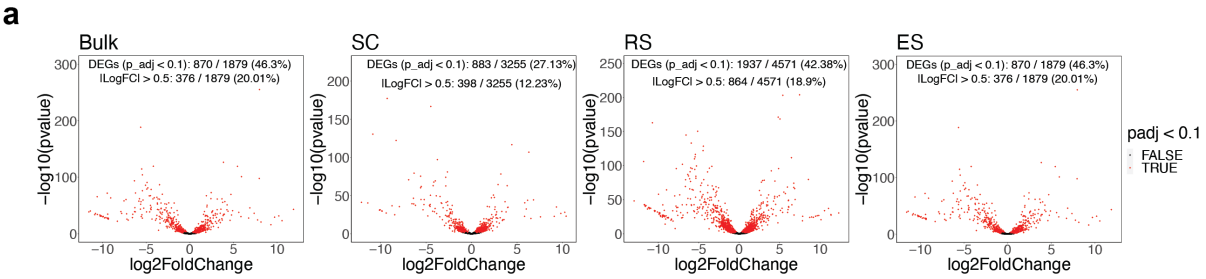

Figure 2.5: Allelic differential expression analysis on main germ cell types. (a) Volcano plots showing the result of a differential expression analysis between the two alleles. Genes are considered differentially expressed if they are detected at an FDR < 10% and stratified by whether their absolute fold change exceeds 0.5.

Differential expression between two strains can be the result of genetic variation affecting the expression of nearby genes in *cis* or by altering function or expression level of diffusible factors which in turn affect gene expression levels in *trans* [Wittkopp et al., 2004, Goncalves et al., 2012]. These two scenarios can be distinguished by measuring gene expression of the two strain-specific alleles in the same *trans*-regulatory environment, for example in the nucleus of an F1 animal, and comparing it to the allelic expression levels in the parental strains, where both *cis*- and *trans*-acting factor influence transcript levels. Mathematically, I therefore define F0, F1, *cis*- and *trans*-allelic fold changes as follows:

$$aFC_{F0} = log2\frac{x^{F0}_{B6}}{x^{F0}_{CAST}}$$
$$aFC_{F1} = log2\frac{x^{F1}_{B6}}{x^{F1}_{CAST}}$$
$$aFC_{cis} = aFC_{F1}$$
$$aFC_{trans} = aFC_{F1} - aFC_{F0}$$

where x_a^S represents the number of reads mapping to an allele a in strain S. I next sought to classify genome-wide, whether the differential expression of genes were driven by *cis*- or *trans*-effects. To this end, I re-implemented a statistical model first presented in [Goncalves et al., 2012]. In this framework, it is assumed that the presence of *cis*- and *trans*-effects on allelic-specific gene expresion will lead to allelic fold changes as follows:

conserved: No difference in expression levels between alleles in F0 or F1,
$aFC_{F0} = aFC_{F1} = 0$

cis: Different expression levels between alleles, with the same difference in F0 and F1,
$aFC_{F0} = aFC_{F1} \neq 0$

trans: No difference in expression levels between alleles in F1 but a difference in F0,
$aFC_{F0} \neq aFC_{F1} = 0$

cis + trans: No difference in expression levels between alleles in F0 or F1,
$aFC_{F0} \neq aFC_{F1} \neq 0$

Next, the model places distributional assumptions on the generative process the data is derived from. For a given gene, let x_i and y_j be the observed reference and alternative F0 counts that are adjusted for sequencing depth by downsampling in a given replicate. Let k_l, n_l be the reference and total counts in in a given F1 replicate. I assume that x_i, y_i are generated from Poisson-Gamma distributions with mean parameters p_x, p_y and over-dispersion r and that k_l, n_l are generated from a beta-binomial distribution with mean parameter π and over-dispersion θ:

$$
\begin{aligned}
x_i &\sim Poi(\lambda_x), \lambda_x \sim Ga(r, p_x/(1-p_x)) \\
y_j &\sim Poi(\lambda_y), \lambda_y \sim Ga(r, p_y/(1-p_y)) \\
k_l &\sim Bin(n_l, p_l), p_l \sim Beta(\pi/\theta, (1-\pi)/\theta)
\end{aligned}
$$

For the Poisson-Gamma distributions, I used DESeq2 to estimate gene-specific over-dispersion parameters r.
As $\frac{p_x}{(1-p_x)}, \frac{p_x}{(1-p_x)}$ and π correspond to the average gene expression levels of both strains and within the F1, I can translate the regulatory categories defined above into restrictions on the mean parameters:

***conserved*:** $p_x = p_y$ and $\pi = 0.5$

***cis*:** $p_x \neq p_y$ and $\pi = p_x/(1-p_x)/(p_x/(1-p_x) + p_y/(1-p_y))$

***trans*:** $p_x \neq p_y$ and $\pi = 0.5$

cis + *trans*: $p_x = p_y$ and $\pi \neq 0.5$

For statistical inference, I now fit the observed data to each statistical model using maximum likelihood with each parameter restriction and identify the most likely model using the Bayesian Information Criterion (BIC). Specifically, I computed the BIC for each model and assign a gene to the conserved category, if for no alternative model the difference in BIC $\Delta BIC = BIC_{conserved} - BIC_{alternative} > 4$. If multiple alterantive categories showed $\Delta BIC > 4$, I chose the category with the highest difference.

I first assessed under which mode of regulation genes were across all germ cells, mirroring an analysis of the tissue in bulk (I removed somatic cells, but these are a small fraction), as performed in [Goncalves et al., 2012]. I found that the vast majority of genes showed either no genetic effects (conserved) or *cis*-effects, while a much smaller fraction showed *trans*-effects or *cis* and *trans*-effect **(Fig. 2.6a)**. These results show that as in liver, the majority of gene expression changes are driven by *cis*-regulatory effects **(Fig. 2.6e)**. I observe substantially fewer *trans*-regulatory effects coinciding with *cis*-regulatory effects which might reflect differences in regulatory effects between tissues, or differences in the technology utilized (bulk vs scRNA-Seq). Indeed, I found significant but modest overlap between the regulated genes in testis and liver. Besides being more common, *cis*-effects also showed stronger effect sizes **(Fig. 2.6c, d)**.

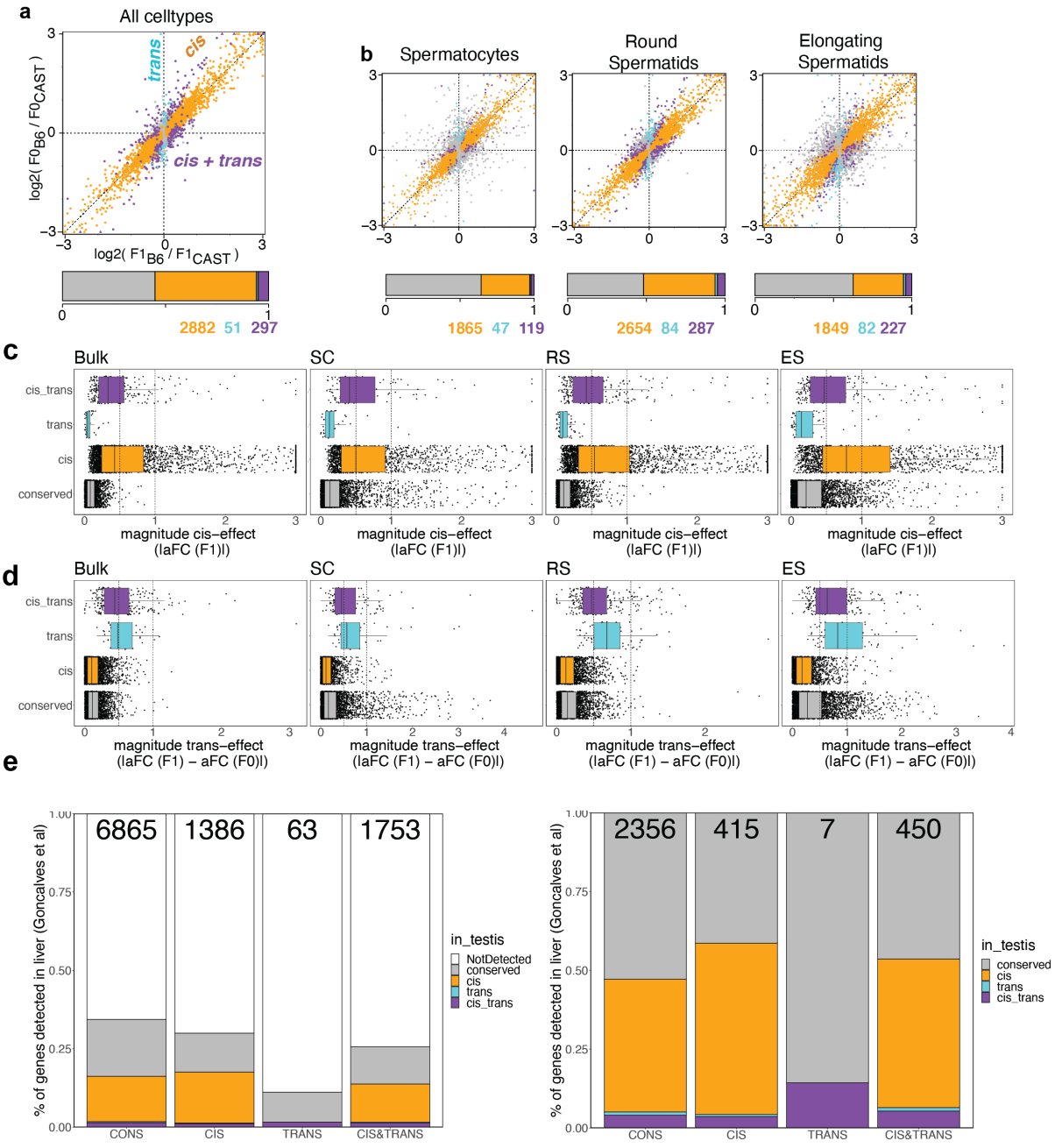

Figure 2.6: (Caption on the next page.)

Figure 2.6: Classification of genes into categories of genetic effects. (**a**) Scatterplot showing allelic fold changes in F1 and F0 across all germ cells. Genes are classified using the framework described above and colored by their category. The barplot underneath shows the fraction adn number of genes annotated into the different categories. (**b**) As (**a**), but a separate analysis for spermatocytes (SC in the following), round spermatids (RS) and elongating spermatids (ES). (**c**) Boxplot showing the magnitude of *cis*-effects across categories. (**d**) Boxplot showing the magnitude of *trans*-effects across categories. (**e**) Barplot showing the overlap between detected *cis*-, *trans*, and *cis*+*trans*-categorized genes. All genes that are undetected in the testis dataset are excluded, which is shown more directly on the right. Annotations of genetic effects in liver are used directly from the original publication [Goncalves et al., 2012].

2.2.4 Identification of cell type-specific *cis*- and *trans*-contributions

Having confirmed that most strain-specific expression is driven in *cis*, I next asked whether I could identify differences in regulatory variation between cell types. To this end, I repeated the assignment to regulatory categories for each of the major cell types (spermatocytes, round spermatids and elongating spermatids). In each cell type I found that the majority of genes was either conserved or regulated by *cis*-effects, with *trans*-effects playing a smaller role **(Fig. 2.6a, b)**. Interestingly, I found more genes with genetic effects when aggregating all discoveries across 3 cell types than by the whole-tissue analysis, suggesting that effects can be masked by bulk level analysis **(Fig. 2.7a)**. I next asked how often genes were categorized to show the same effect category across cell types. This analysis showed that, first, a substantial number of genes only shows genetic effects in specific cell types and, second, that a small proportion changes regulatory categories **(Fig. 2.7b)**. Finally, I aimed to test directly whether *cis*-effects differed substantially between cell types. To this end, I fit a binomial linear model to the sample- and cell type-level allelic counts k_{cs}, n_{cs} and used a likelihood ratio test to identify genes with significant and strong differences:

$$k_{cs} \sim Bin(n_{cs}, p_c) \tag{2.2}$$

$$logit(p_c) = \beta_0 + \beta_c x_s \tag{2.3}$$

where p_c represents the allelic ratio in cell type c and β_c represents the cell type-specific difference in logit-transformed allelic ratios. This analysis showed 411 genes with differential *cis*-effects between spermatocytes and round spermatids and 411 genes between spermatocytes and elongating spermatids **(Fig. 2.7c, d)**. Given that there seemed to be cell type-specificity in allelic ratios, I asked whether these alone were sufficient to reconstruct a latent space that would allow to separate cell type in the same way as total expression would. To this end, I quantified the evidence for allelic imbalance ratios from 0.5 as the binomial deviance score in every cell:

$$D = log(\frac{Bin(n, k, \frac{n}{k})}{Bin(n, k, p_0)}) \tag{2.4}$$

$$= k \cdot log(\frac{k}{n \cdot p_0}) + (n - k) \cdot log(\frac{n - k}{n \cdot (1 - p_0)}) \tag{2.5}$$

Where $p_0 = \frac{k_{total}}{n_{total}}$ summed acros all cells. In then subjected the resulting score matrix to principal component analysis and performed UMAP. The resulting UMAP structure groups cells of the same cell type close to each other, suggesting that there is cell type-specific information in

allelic ratio profiles (**Fig. 2.7e**. This analysis should be caveated on whether this score is fully independent on total gene expression levels.
Collectively, these results demonstrate that using single-cell readouts of allelic usage, I can identify genes under genetic control in a transcriptome-wide manner and that genetic effects vary substantially across cell types.

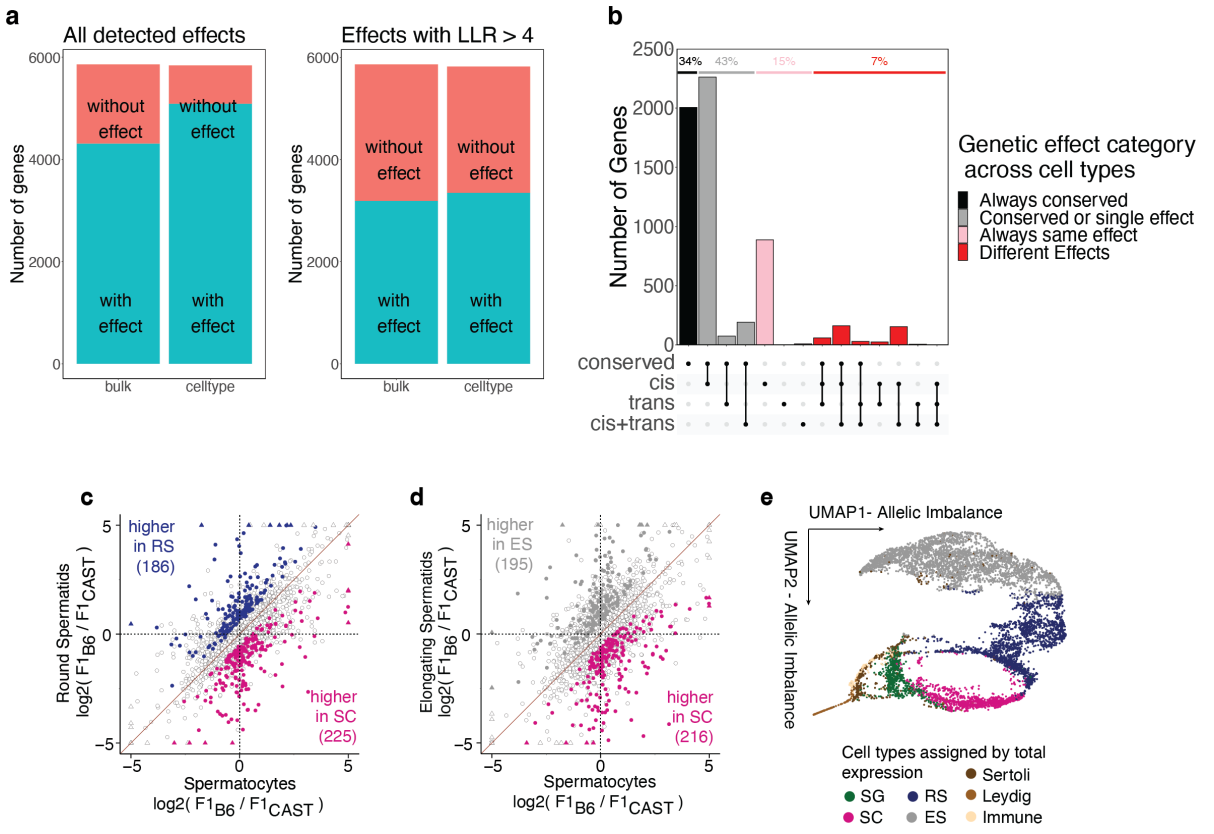

Figure 2.7: Variation in genetic effects across cell types. (**a**) Barplot showing how many genes are assigned to have a genetic effect (*cis, trans, cis+trans*) when assigning them to counts derived from aggregating the entire sample ("bulk") as opposed to assigning them to major cell types seperately and counting the number of genes with an effect in at least one cell type. The right panel shows the same analysis, when subsetting on effects with a log-likelihood ratio of > 4. (**b**) Upset plot categorizing genes into categories based on their genetic effects. (**c**) Scatterplot showing allelic fold changes between spermatocytes and round spermatids, colored based on whether differential allelic imbalance had been detected (binomial linear model, FDR < 10% by likelihood ratio test). (**d**) as (**c**), comparing spermatocytes and elongating spermatids. (**e**) UMAP embedding of allelic ratios, overlaid with cell types annotated based on total expression values.

2.3 Dynamic models of *cis*- and *trans*-contributions to species-specific expression

In the previous section, I analysed genetic effects in testes by treating the 3 major cell types as distinct entities. However, this is clearly a simplification of the process by which spermatogonia differentiate into spermatids, which features multiple continuous transitions through differentiation stages. One exciting prospect of single-cell based methods is that it can reconstruct such processes as, given sufficient sampling, one will be likely to capture cells in all states and sub-states. Indeed, the dimensionality reduction in **(Fig. 2.2c)** shows that the cluster definitions during germ cell differentiation are somewhat artificial as cell states transition into each other.

2.3.1 Allelic variability during cellular differentiation

I therefore aimed to switch from a cell type-centric to a differentiation-centric view and attempt to model genetic effects as a continuous function f: cell state → strength of a genetic effect. To this end, I used principal curve analysis to fit a 1-dimensional trajectory to the PC-compressed gene expression data. This analysis assigns a single coordinate in differentiation to each cell in both F1 and parental datasets and clearly shows a progression from the stem cell to the most differentiated states **(Fig. 2.2a, b)**. I can consider this trajectory as a pseudo-time of differentiation, as it captures its temporal progression through repeated sampling from multiple asynchronous differentiation processes [Trapnell et al., 2014].

I next asked to which extent *cis*-effects were modulated by differentiation stage. From a statistical point of view, this analysis corresponds to the question of whether allelic imbalance in F1 animals varies significantly across pseudotime. For this analysis I employed *scDALI*, a binomial generalized linear mixed model to test genes for a) *persistent cis*-effects, that is, unequal expression of the two alleles that is constant across pseudotime and b) *dynamic cis*-effects, where the allelic imbalance varies [Heinen et al., 2022]. Visually, this corresponds to a constant function centered at an allelic imbalance of 0.5 for genes with no *cis*-effect, while a constant shift represents a persistent and any non-constant function represents a dynamic effect **(Fig. 2.8c, d)**. The statistical tests are implemented as scDALI-hom and scDALI-het and correspond to testing for significant variance components σ_p^2 and σ_d^2 respectively in the follow model:

$$
\begin{aligned}
k &\sim Bin(n, p) \\
p &\sim Beta(\mu, \theta) \\
logit(\mu) &= u \\
u &\sim \mathcal{N}(\alpha, \sigma_p^2 I + \sigma_d^2 K)
\end{aligned}
$$

where k and n are the cell-specific observed reference and total allelic counts and p is the underlying allelic rate. In the scDALI-model, p is drawn from a Beta-distribution with over-dispersion parameter θ to account for unmodelled technical or biological variability. The vector of (logit-transforemd) mean allelic rates is then drawn from a multivariate normal distribution with mean of a null allelic rate α (usually and in this case, $\alpha = logit(0.5) = 0$) with a covariance matrices that allow for constant shifts ($\sigma_p^2 I$) and temporal variation ($\sigma_d^2 K$).

Similar to gaussian process regression, here K represents a kernel matrix that determines the shape of the underlying function fit to the allelic rate. For simplicity, here I use an order-3 polynomial kernel $\mathbf{P} = [t, t^2, t^3], K = \mathbf{P}^T\mathbf{P}$.

2.3.2 Pervasive differentiation-dependence of *cis*-effects

I tested 4,039 genes with at least 1000 allele-specific counts per sample for persistent and dynamic allelic imbalance to identify persistent and dynamic *cis*-effects. As an example, I observed strong cell type-dependent allelic imbalance for the gene *Ddt*, which is highly expressed in spermatocytes and whose expression drops on both alleles throughout spermiogenesis. However, as the B6-allele drops faster in expression than the CAST-allele in round spermatids I observed a strong stage-dependent effect on the allelic ratio (FDR < 10%) **(Fig. 2.8c, d)**. Notably, in the genome-wide analysis, I found persistent effects in 2275 (56%) of genes and dynamic effects in 1152 (29%) of genes. This corresponds to 44% of genes for which the cell type-specific regulatory environment modulates the magnitude of the genetic effect. Notably, 334 of dynamic effects did not show a persistent effect, meaning that they would be missed if analyzing the tissue in bulk **(Fig. 2.8d)**. Specifically, I found examples of these genes that a) showed allelic imbalance in an extremely narrow window of differentiation or b) showed opposite directions in allelic imbalance, both of which lead to a aggregate imbalance close to 0.5 **(Fig. 2.8e)**. I also showed that when encoding the cell state not as a continuous, but a discrete variable (block covariance encoding cell types) I found fewer allelic effects, demonstrating that in systems not well described by discrete cell types, power can be gained by dynamic modelling **(Fig. 2.8f-g)**.

Next, I asked when during cellular differentiation dynamic *cis*-effects occured. To this end, I first derived estimated allelic trajectories using the Gaussian process model in scDALI. Here, the model is simplified to a Gaussian data likelihood and I used Gaussian process regression with a RBF-Kernel to derive a non-parametric fit to the allelic trajectory:

$$\frac{k}{n} \sim \mathcal{N}(\alpha, \sigma^2 K)$$

I then performed hierarchical clustering on the z-scored inferred trajectories and identified seven clusters of genes with similar changes in allelic usage **(Fig. 2.9a)**. This analysis first showed that groups of genes with changes in *cis*-effects exist across the entire differentiation process. However, cell type transitions showed a concentration of changes and the two largest clusters featured strongest changes during the transition from round to elongating spermatids. This process features global chromatin remodelling and progressive silencing of transcription caused by the exchange of histones to protamines, which might exacerbate differential allelic effects on transcription [Rathke et al., 2014] . I next asked where during a gene expression trajectory allelic balance would be strongest. To this end, I also used hierarchical clustering to group genes with similar changes in total gene expression (the sum of both alleles) during differentiation. I found that dynamic genetic effects occured similarly in genes that peaked in expression early, intermediate or late and that there was little association between when a gene was most strongly expressed and when it showed strongest allelic imbalance at a global scale **(Fig. 2.8b)**. Indeed, there was little association between allelic imbalance and expression clustering across pseudotime **(Fig. 2.8c)**.

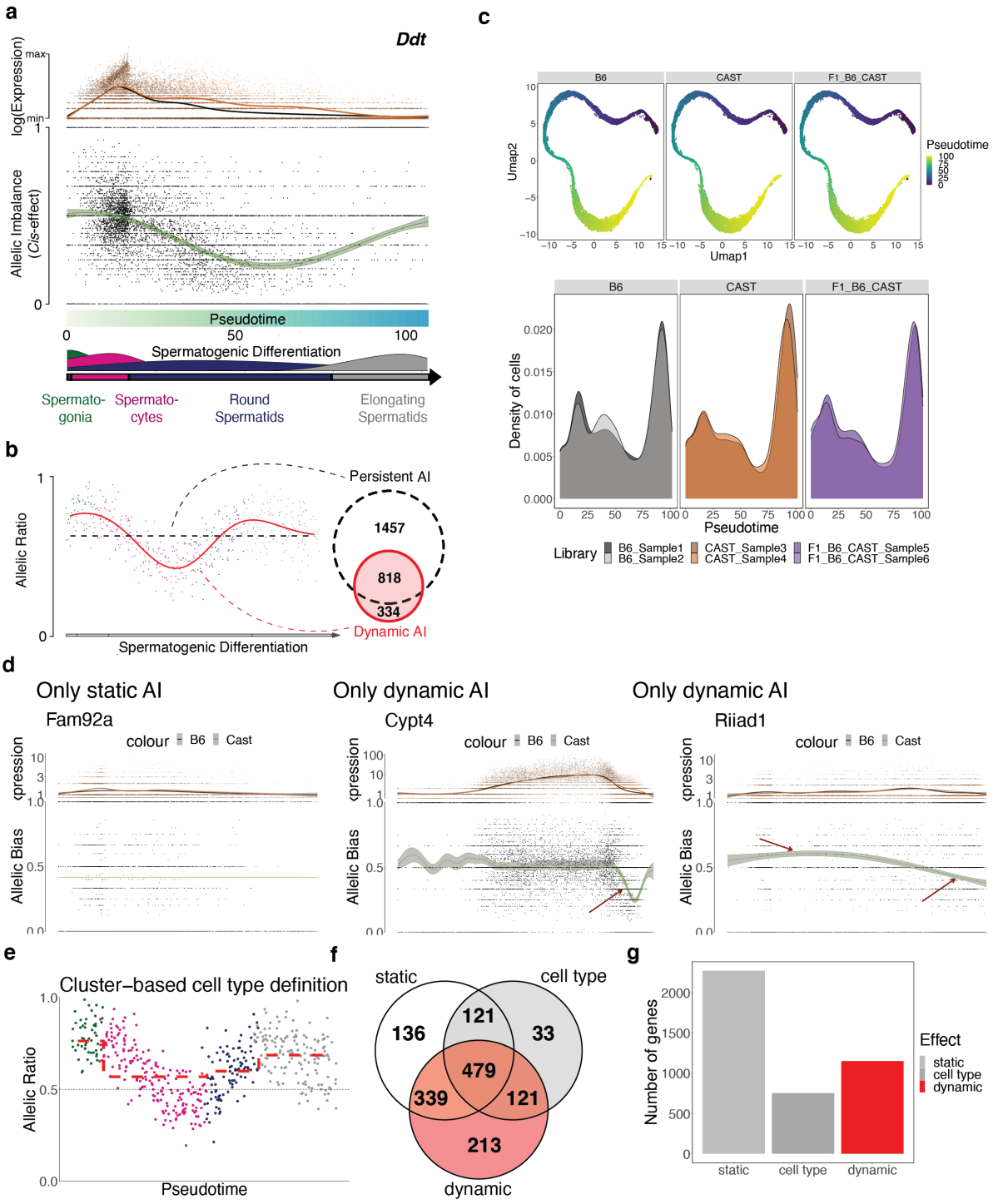

Figure 2.8: Dynamic modelling of allelic imbalance during spermatogenic differentiation. **a** UMAP showing the progression of pseudotime estimates across the differentiation process, separated by strains. **b** pseudotime scores are shown as a density, showing an equal distribution of cells across the assigned pseudotime. **c** Top panel: Scatterplots showing expression of the B6 (black) and CAST alleles across pseudotemporal ordering for the *Ddt* gene. The lines represent a LOESS interpolation. Middle panel: The same data, visualized as allelic ratios. The green line and shading represent mean and variance of the interpolated ratio of the scDALI fit. Lower panel: Schematic and cell type distributions of germ cells across spermatogenic differentiation. **d** Schamtic showing the analysis approach. Genes with persistent and dynamic allelic imbalance are detected using scDALI. On the right a Venn diagram shows the number and overlap of detected genes. **e** Examples of scDALI results, showcasing different classes of genetic effects. **f** Schematic showing an alternative analysis approach where cell types are defined as discrete cell types instead of pseudotemporal ordering. **f** Venn diagram showing the overlap between a static, dynamic and the cell type-specific modelling approach. **f** Bar plot showing the number of genes with genetic effects between the three approaches.

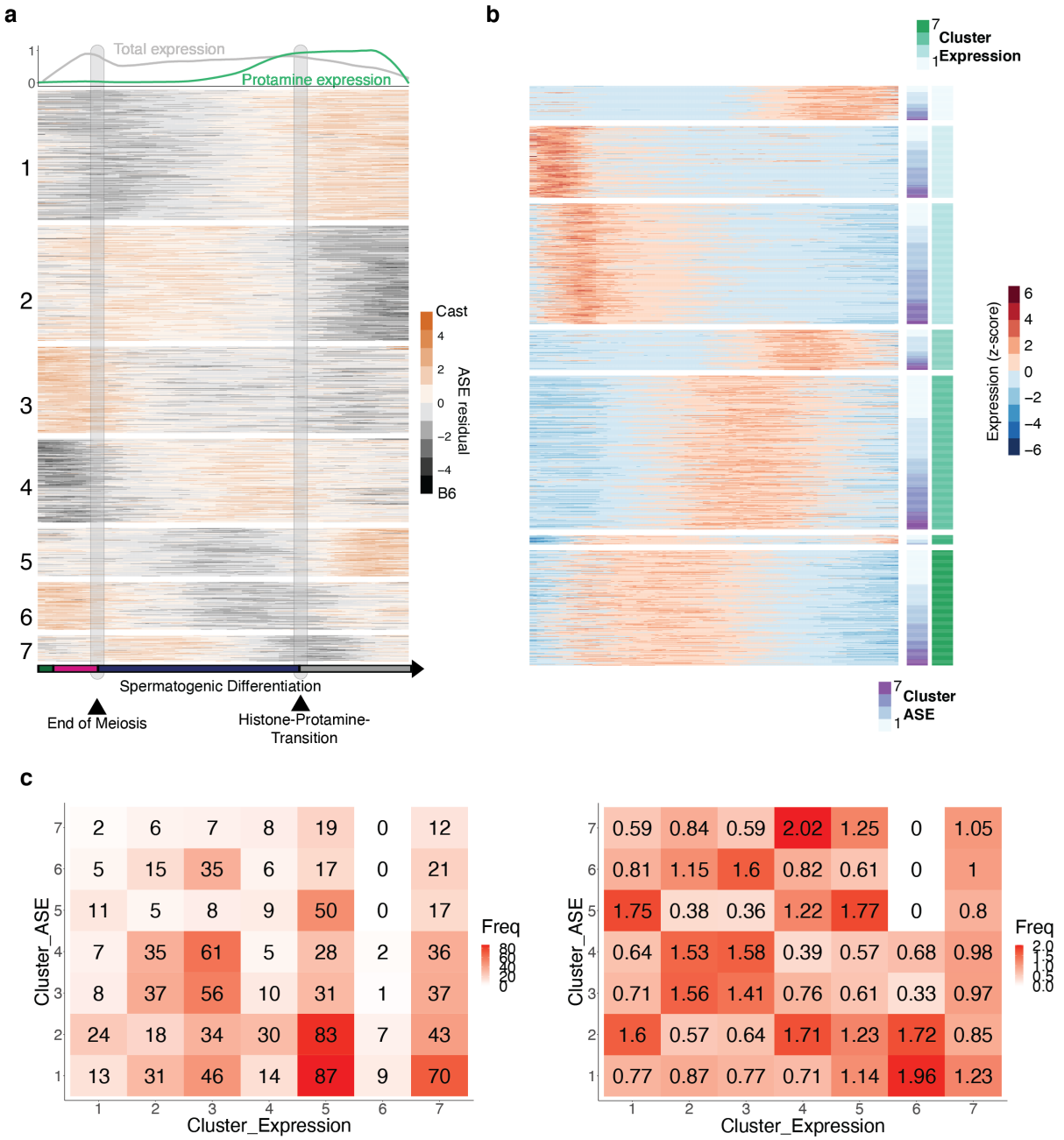

Figure 2.9: Clustering of ASE trajectories across spermatogenesis. **a** Heatmap showing z-scores of allelic rate interpolations calculated by scDALI, where black represents a relative bias towards the B6 and brown towards the CAST allele. The genes are ordered into seven groups using hierarchical clustering. The top panel shows total expression levels of all genes (grey) and histone protamine transition genes (*Prm1, Prm2, Tnp1, Tnp2*). The bottom arrow indicates the stages of spermatogenic differentiation across pseudotime. (**b**) Similarly, this heatmap shows z-scored total expression for the same genes. Right annotations show hierarchical clustering results based on allelic trajectories (as in **a**, purple) and total expression trajectories (green). (**c**) Heatmap showing the overlap between the two clustering analyses in number of genes per cluster combination (left). On the right, ratios of observed against expected number of genes if all clusters were equally distributed.

2.3.3 Dynamic *cis*-effects are associated with transcriptional regulation and chromatin accessibility

To investigate further which regulatory mechanisms might drive dynamic allelic im balance, I specifically asked whether for individual genes, allelic imbalance would be strongest during gene up-regulation, down-regulation or its highest point of expression. To this end I identified 726 genes that showed such a behaviour (from low to high and again to low expression) and quantified allelic balance during the start and the end of this trajectory **(Fig. 2.8a)**. I calculated the strength of an allelic effect as the absolute difference to allelic balance $|AI - 0.5|$ and quantified it across 20% quantiles of the pseudotime where the gene in question is expressed. I found both genes that were strongest early and late in differentiation, with allelic imbalance more frequently escalating late **(Fig. 2.8b)**. Late effects were also generally stronger **(Fig. 2.8c)** and were as common as genes peaking in AI throughout the trajectory **(Fig. 2.8d)**. This analysis suggests that allelic imbalance is associated with phases in gene expression where a gene is actively regulated (up- or down). Also, the enrichment of genes with a strong allelic imbalance during down-regulation suggests a possible mechanistic explanation for differential allelic imbalance: When a gene is up-regulated, its transcript levels are mainly driven by its transcription rate, while down-regulation will be driven both by changes in transcription rate and changes in RNA stability, for example through differential activity of RNA-binding proteins.

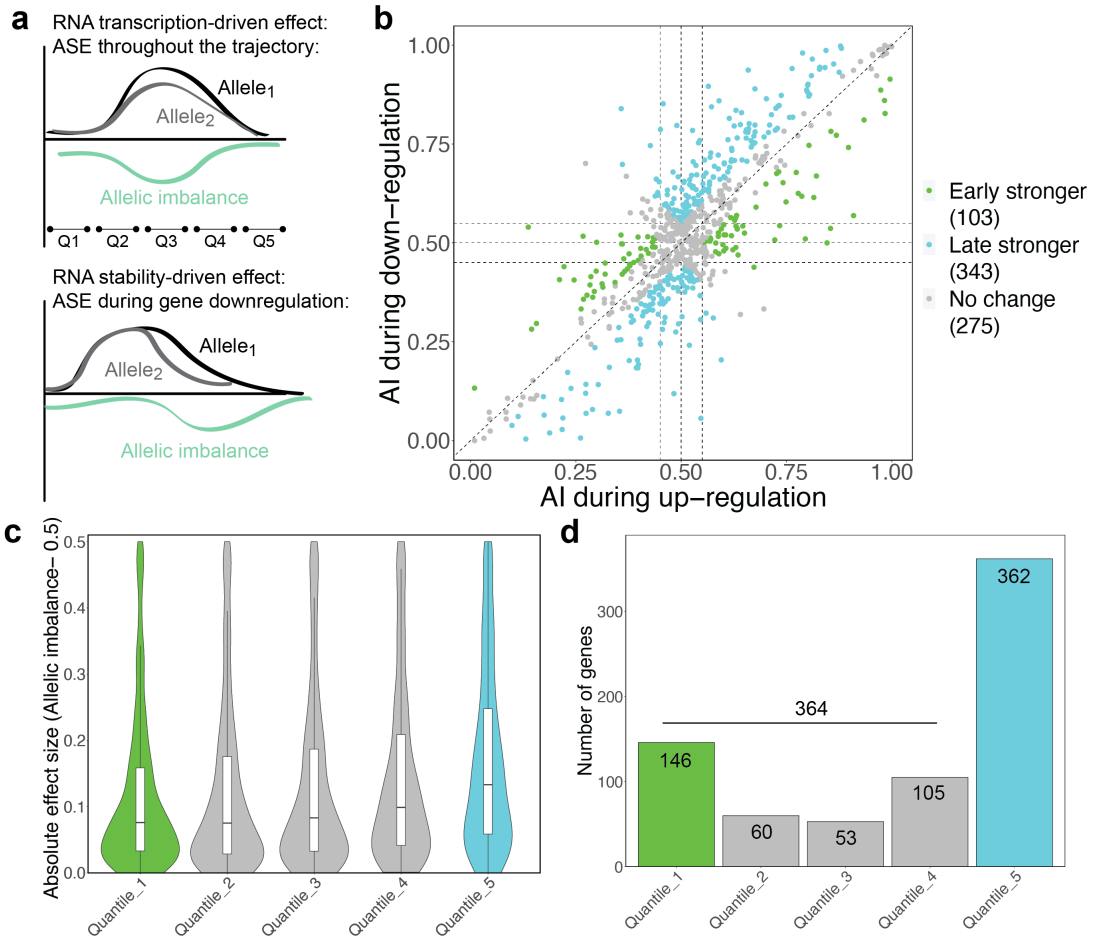

Figure 2.10: (Caption on the next page.)

Figure 2.10: Strength of allelic imbalance across gene expression trajectories. (**a**) Schematic showing the analysis approach. A gene expression-driven effect will be visible as a multiplicative change across the gene expression trajectory. A stability-driven effect will lead to an imbalance only during down-regulation. (**b**) Scatterplot showing allelic imbalance in the beginning (first 20% quantile) and the end (80% quantile) for all considered genes. Genes are highlighted as "early stronger" (green) if their allelic ratio is at least 0.1 stronger than late and "late stronger" (blue) if the reverse is true, or without change if neither is true. (**c**) Violin plots showing the distribution of the absolute allelic effect across 20% quantiles through the expression. (**d**) Barplots showing the number of genes which show the strongest allelic effect in the respective quantile of pseudotime.

To further investigate the impact of transcriptional regulation on allelic imbalance, I next asked if allele-specific chromatin accessibility was associated with allele-specific expression. To this end, I performed ATAC-Seq in spermatocytes by using FACS-sorting to isolate tetraploid cells. After data processing and identification of regions with open chromatin, I confirmed that across two replicates, both total read count and allelic imbalance were well reproducible and that I found many sites with allelic bias in chromatin accessibility **(Fig. 2.11a,b)**. When analyzing the genomic location of identified peaks, I observed that the majority overlapped promoters or gene bodies, and only around 15% were located in intergenic regions (at least 3 kb away from any gene) **(Fig. 2.11c)**. This is unlike chromatin accessibility profiles in most somatic cells [Buenrostro et al., 2015], and shows that meiotic cells at this stage are depleted for accessibility at distal regulatory sites.

To assess whether allele-specific chromatin accessibility (asCA) was associated with allele-specific gene expression (asGE), I performed an enrichment analysis as follows: I identified 6,844 sites with an asCA > 0.1 (calculated as the absolute difference of the allelic ratio $\frac{B6}{B6+CAST}$ to 0.5). I then calculated the fraction of genes asGE that showed at least one peak with asCA nearby. To obtain a null expectation, I shuffled the association between peaks and genes randomly throughout the genome. I found that genes with asGE were more likely to show asCA than expected randomly, and that this differences was stronger for genes with higher asGE, suggesting that asCA might underly allelic differences in expression **(Fig. 2.11d)**. To ask whether asCA was also able to predict cell type-specific asGE, I repeated this analysis considering genes with asGE in specific cell types (SC, RS and ES). Spermatocytes showed a higher coordination between asGE and asCA, which is expected as the ATAC-Seq profiles were measured in SCs **(Fig. 2.11e)**. These results show that there is a cell type-specific association between asGE and asCA. asGE was also associated with spermatocytes-asCA in other cell types, likely because the majority of genes show asGE across cell types (even if it is variable). To finally assess coordination between dynamic asGE and chromatin accessibility, I ranked all genes with dynamic AI by how much stronger AI was in spermatocytes. As expected, genes with higher AI in SCs were more likely to show asCA nearby, reflecting an association between cell type-specific chromatin accessibility and cell type-specific asGE **(Fig. 2.11f,g)**.

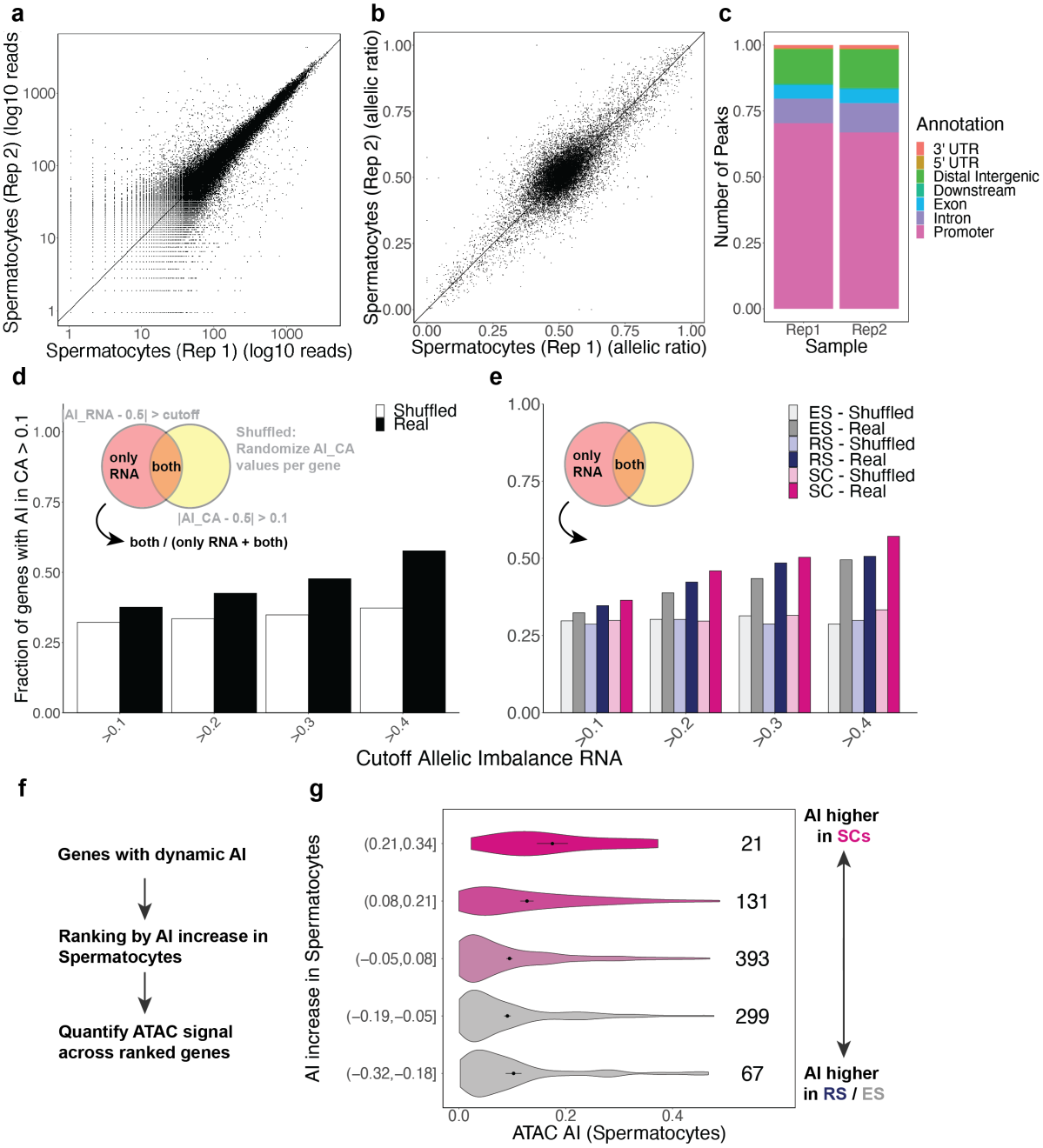

Figure 2.11: Allele-specific chromatin accessibility in spermatocytes. (**a**) Scatterplot showing the number of reads per peak across the two analyzed replicates. (**b**) Scatterplot showing allelic ratios between the two analyzed replicates. (**c**) Distribution of genomic location of the accessible sites. Exons, Introns and untranslated regions are defined based on gene models, promoters as 3kb regions up- and downstream the transcription start site. All other peaks are considered "distal intergenic". (**d**) Barplots showing association analysis between allelic imbalance measurements in RNA and accessibility. (**e**) As (**d**), but seperated by major cell types. (**f, g**) Analysis approach to associate dynamic allelic imbalance to chromatin accessibility. Violin plots show asCA of peaks assigned to the nearest gene which are ranked by the difference of allelic imbalance in spermatocytes and spermatids.

2.3.4 Sequence conservation, but not other genomic features are predictive of dynamic *cis*-effects

Finally, I aimed to identify other genomic features associated with dynamic allelic imbalance. To this end, I adopted a machine learning procedure by asking if I could train a classification model to predict whether genes showed *cis*-effects based on genomic properties. I included information 1) from my scRNA-Seq dataset such as expression level and intra- and inter-celltype expression variability, 2) genic features such as gene length, promoter structure and the size of untranslated regions, 3) sequence conservation and specifically SNV-densities between the B6 and CAST strains and 4) the presence of promoters and enhancers derived from previous publications [Roller et al., 2021].

I transformed the features if necessary and then trained two random forest classification models to first predict whether a gene had any *cis*-effect (persistent or dynamic) and secondly, among these genes, whether they had a dynamic as opposed to only a persistent *cis*-effect. To assess model performance, I performed 5-fold cross-validation and then ranked features by their mean decrease in accuracy. I obtained an average cross-validation AUC of 0.633 for predicting persistent and of 0.605 for dynamic effects, showing that genomic features were generally not strongly predictive **(Fig. 2.12a)**. However, I found that genes with genetic effects were under generally lower sequence conservation as measured both by PhastCons scores and the SNV-density between B6 and CAST. Indeed, genes with dynamic *cis*-effects showed even lower sequence constraint than genes with persistent effects, showing that variation in gene expression is associated with sequence variability. I also found gene expression level to be weakly predictive which likely reflects that the power to detect genetic effects is generally lower for lowly expressed genes **(Fig. 2.12b)**.

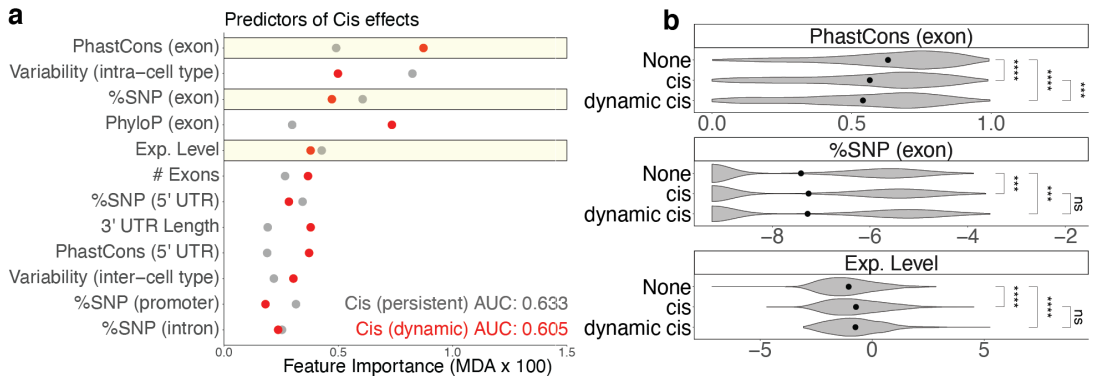

Figure 2.12: Prediction of genes with *cis*-effects from sequence features. ((**a**)) Summary of prediction results. Feature importance is measured by mean decrease in accuracy when the respective feature is remoed. Grey values represent the results when predicting genes with *cis*-effects from all tested genes. Red values represent the prediction of genes with dynamic *cis*-effects from genes with persistent and dynamic effects. **(b)** Violin plots showing the distribution of predictive features across effect categories. Significance is estimated by Wilcoxons rank sum test, *< 0.1, **< 0.05, ***< 0.01, ****< 0.001.

In this section, I use dynamic modelling to provide a comprehensive characterization of genetic effects acting in *cis* that vary across differentiation. Surprisingly, the strength of genetic effects varies for more than 40% of *cis*-effects, highlighting their frequent context-dependency. These dynamics can be partly explained to allele-specific chromatin accessibility, but to a lesser extent

by genomic features. Finally, I show that increased allelic imbalance occurs during up- and down-regulation of genes, providing a possible relation to differential RNA stability.

2.3.5 A novel approach to identify genes with dynamic *trans*-effects

Besides dynamic *cis*-effects, my experimental design uniquely provides the opportunity to identify dynamic *trans*-effects that occur when a specific allele of a *trans*-acting factor influences gene expression. Classical genetic association tests are usually focussed on *cis*-effects (that are close to the gene of interest). Meanwhile, *trans*-association testing effectively requires regression every genetic variant against the expression of all individual genes, which requires extremely large sample sizes for a powered analysis, and this is especially true when context-specificity is considered. Meanwhile, an F1 design provides *trans*-components of genetic effects by comparing allelic usage differences between F1 alleles (which are in the same trans-regulatory environment and therefore isolate the *cis*-effect) and F0 alleles (which are affected by both *cis* and *trans*). This controlled analysis is possible with few replicates and can be easily extended to test for differentiation dependency. As an example, I visualized the allelic expression trajectories *Dnajc2* gene for which I observe a deviation between allelic trajectories specifically in spermatids **(Fig. 2.13a)**.

I therefore sought to extend the dynamic modelling approach outlined by scDALI to test for differences in allelic usage between two trajectories, rather than just testing for variability in one. To this end, I first needed to quantify allele ratios in F0 over differentiation. Since both alleles are not measured in the same cell in the parental strains, I estimated a joint pseudotime for all samples (strains) together and then grouped cells with similar pseudotime values into 100 evenly spaced intervals. I then quantify a bin-wise allelic ratio between F0s and with the F1. This results in two allelic trajectories for which deviations show the presence of (dynamic) *trans*-effects. Statistically, I describe the two trajectories through co-localized gaussian process regression **(Fig. 2.13b)**.

Specifically, I quantified F1 and F0 read counts k_c, n_c and x_c, y_c, per interval $t_0 < t_1 < ... < t_{100}$ across pseudo-time and compute average allelic rates in an interval $[t_i, t_{i+1}]$

$$r_{F0}^{t_i} = \frac{\sum x_c}{\sum x_c + \sum y_c}, r_{F1}^{t_i} = \frac{\sum k_c}{\sum n_c}, \tag{2.6}$$

where the index c considers all cells in the interval t_i. I then assumed for $\mathbf{r}_{F0} = (r_{F0}^1, r_{F0}^2, ..., r_{F0}^n)$ and $\mathbf{r}_{F1} = (r_{F1}^1, r_{F1}^2, ..., r_{F1}^n)$ and with $\phi(x) = log(\frac{x}{1-x})$ applied element-wise:

$$\begin{pmatrix} \mathbf{u}_{F_0} \\ \mathbf{u}_{F_1} \end{pmatrix} = \begin{pmatrix} \phi(\mathbf{r}_{F_0}) \\ \phi(\mathbf{r}_{F_1}) \end{pmatrix} \tag{2.7}$$

$$\begin{pmatrix} \mathbf{u}_{F_0} \\ \mathbf{u}_{F_1} \end{pmatrix} \sim \mathcal{N}(\begin{pmatrix} \boldsymbol{\mu}_{F_0} \\ \boldsymbol{\mu}_{F_1} \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{F_0,F_0} & \mathbf{K}_{F_1,F_0} \\ \mathbf{K}_{F_0,F_1} & \mathbf{K}_{F_1,F_1} \end{pmatrix} + \begin{pmatrix} \sigma_1\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \sigma_2\mathbf{1} \end{pmatrix}) \tag{2.8}$$

,

$\boldsymbol{\mu}_{F_0}$ and $\boldsymbol{\mu}_{F_1}$ represent constant mean functions that model shifts of the allelic imbalance trajectory from $\phi(0.5) = 0$. $\mathbf{K}_{F_0,F_0}$ and $\mathbf{K}_{F_1,F_1}$ represent the covariance functions of the F0 and F1 trajectories respectively and $\mathbf{K}_{F_1,F_0}$, $\mathbf{K}_{F_0,F_1}$ are cross-covariance functions modelling correlated or uncorrelated behaviour of the F0 and F1 functions. The normal approximation to the

Beta-binomial likelihood in (2) and (3) is justified for high total counts $\sum x_c + \sum y_c$ and $\sum n_c$. I therefore only considered bins with at least 100 aggregated reads across F0 and F1. I assumed a single kernel function $\mathbf{K}$ with joint hyper-parameters across F0 and F1 (but potentially different realizations), allowing me to write

$$\mathcal{N}(\begin{pmatrix}\boldsymbol{\mu}_{F_0}\\ \boldsymbol{\mu}_{F_1}\end{pmatrix}, \begin{pmatrix}\mathbf{K}_{F_0,F_0} & \mathbf{K}_{F_1,F_0}\\ \mathbf{K}_{F_0,F_1} & \mathbf{K}_{F_1,F_1}\end{pmatrix} + \begin{pmatrix}\sigma_1\mathbf{1} & \mathbf{0}\\ \mathbf{0} & \sigma_2\mathbf{1}\end{pmatrix}) \tag{2.9}$$

$$= \mathcal{N}(\begin{pmatrix}\boldsymbol{\mu}_{F_0}\\ \boldsymbol{\mu}_{F_1}\end{pmatrix}, \begin{pmatrix}\sigma_{F_0,F_0}\mathbf{K} & \sigma_{F_1,F_0}\mathbf{K}\\ \sigma_{F_0,F_1}\mathbf{K} & \sigma_{F_1,F_1}\mathbf{K}\end{pmatrix} + \begin{pmatrix}\sigma_1\mathbf{1} & \mathbf{0}\\ \mathbf{0} & \sigma_2\mathbf{1}\end{pmatrix}) \tag{2.10}$$

$$= \mathcal{N}(\begin{pmatrix}\boldsymbol{\mu}_{F_0}\\ \boldsymbol{\mu}_{F_1}\end{pmatrix}, \begin{pmatrix}\sigma_{F_0,F_0} & \sigma_{F_1,F_0}\\ \sigma_{F_0,F_1} & \sigma_{F_1,F_1}\end{pmatrix} \otimes \mathbf{K} + \begin{pmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{pmatrix} \otimes \mathbf{I}) \tag{2.11}$$

$$= \mathcal{N}(\begin{pmatrix}\boldsymbol{\mu}_{F_0}\\ \boldsymbol{\mu}_{F_1}\end{pmatrix}, \mathbf{C} \otimes \mathbf{K} + \begin{pmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{pmatrix} \otimes \mathbf{I}) \tag{2.12}$$

where $\otimes$ is the kronecker product and σ_1 and σ_2 encode residual variation not explained by the cell state. Based on this general GP model, I can now derive specific models for static and dynamic *trans* effects:

no or only *cis* effects (1): No difference in allelic trajectory between F1 and F0:

$$\boldsymbol{\mu}_{F_0} = \boldsymbol{\mu}_{F_1}, \mathbf{C} = \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$$

only static *trans*-effect (2): Only constant shift between F0 and F1:

$$\boldsymbol{\mu}_{F_0} \neq \boldsymbol{\mu}_{F_1}, \mathbf{C} = \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$$

static + dynamic *trans*-effect (3): Variable allelic trajectories:

$$\boldsymbol{\mu}_{F_0} \neq \boldsymbol{\mu}_{F_1}, \mathbf{C} = \begin{pmatrix}\sigma_{F_0,F_0} & \sigma_{F_1,F_0}\\ \sigma_{F_0,F_1} & \sigma_{F_1,F_1}\end{pmatrix}$$

I fit all models using variational Gaussian processes (VGP model in the package gpflow) and used the ELBO (evidence lower bound) as an approximation to the marginal likelihood of the data under each model. Based on this, I defined Bayes Factors as the ratio of these marginal likelihoods between two models. I defined evidence for static *trans*-effects as the Bayes factor for model **(2)** against model **(1)** and dynamic *trans*-effects as evidence for model **(3)** against model **(1)**. Moreover, this framework allows to detect dynamic *cis*-effects by substituting $\mathbf{K}$ with a constant kernel **(Fig. 2.13c)**.

I next tested genome-wide by considering genes with significant evidence for *trans*-effects when they showed a difference in log posterior probability > 10. This analysis shows 313 genes with only persistent and 117 genes with dynamic *trans*-effects. To validate this analysis, I considered the effect sizes for *trans*-effects. Let $d_i = |r_i^{F0} - r_i^{F1}|$ be the *trans*-residual for a given gene in the pseudotime interval i, I can define a persistent *trans*-effect size $d_{pers} = mean(d_i)$ and a variable effect size as $d_{dyn} = quantile(d_i, 0.9) - d_i, 0.9)$ **(Fig. 2.14c)**. Genes with higher evidence for *trans*-effect also generally show stronger effect sizes, showing that my analysis identifies relevant genes. The observed *trans*-effects occur across different differentiation stages and bias both alleles, which I specifically analyze for the genes *Rnaseh2a, Tcp1, Guk1* and *4930503B20Rik* and genome-wide **(Fig. 2.14d)**. In the genome-wide analysis, I found that

trans-effects more commonly peak late in differentiation **(Fig. 2.14e)**. Finally, I showed that similar to genes with *cis*-effects, *trans*-effects are associated with lower sequence conservation, as evidenced by PhastCons scores. However, I do not find a stronger reduction of conservation in dynamic genes **(Fig. 2.13d)**.

I finally sought to analyze the co-occurence of *cis*- and *trans*-effects on specific genes. For this, I needed to repeat the classification of dynamic *cis*-effects in my Gaussian process framework, in order to derive Bayes Factors comparable between both analyses. To this end, I identified dynamic *cis*-genes using the co-localized GP model, comparing between the static and dynamic *cis*-models. I observed comparable numbers of genes with dynamic *cis*-effects at a log BF of 10 and show that this method and scDALI-het show strong overlap **(Fig. 2.14a, b)**. Based on this classification, I identified that in general, much fewer genes show *trans*-effects, and that these are not more and potentially less often dynamic **(Fig. 2.13e)**. I also observed that both static and dynamic *cis* / *trans*-effects commonly co-occured on the same genes, with a stronger enrichment for dynamic effects **(Fig. 2.13f, g)**. This is in line with previous observations of a large number of *cis*+*trans* effects in previous studies and suggests possible interactions of different genetic effects on the same genes [Goncalves et al., 2012].

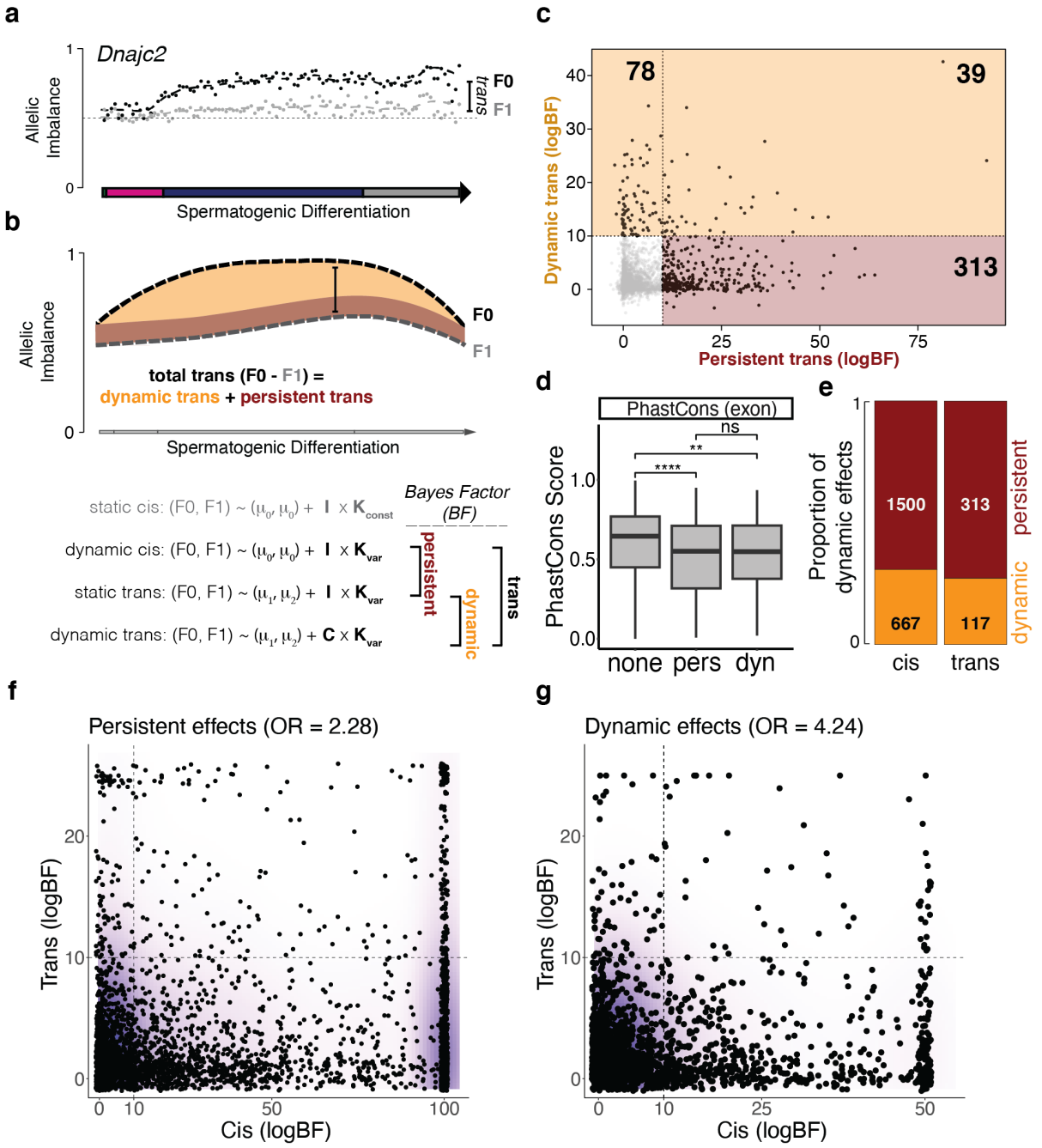

Figure 2.13: Gaussian process models to detect dynamic *trans*-effects. (**a**) Allelic trajectories measured in F1 and between F0 mice for the gene *Dnajc2*. Individual points represent aggregates in 100 evenly spaced intervals of pseudotime. The underneath arrow represents the most common cell type at the respective pseudotime interval. (**b**) Schematic showing the analysis approach. Differences between F0 and F1 trajectories, are decomposed into an additive shift and additional variation, which captures persistent and dynamic *trans*-effects. (**c**) Scatterplot showing log Bayes Factors for persistent and dynamic *trans*-effects. (**d**) Boxplot showing average exonic PhastCons scores for genes without, persistent or dynamic *trans*-effects. (**e**) Barplots showing the absolute numbers and fractions of effects that are dynamic, for *cis* and *trans* respectively. (**f, g**) Scatterplots comparing the evidence for *cis*- and *trans*-effects that are persistent or dynamic. OR indicates the odds ratio of both effects co-occurring.

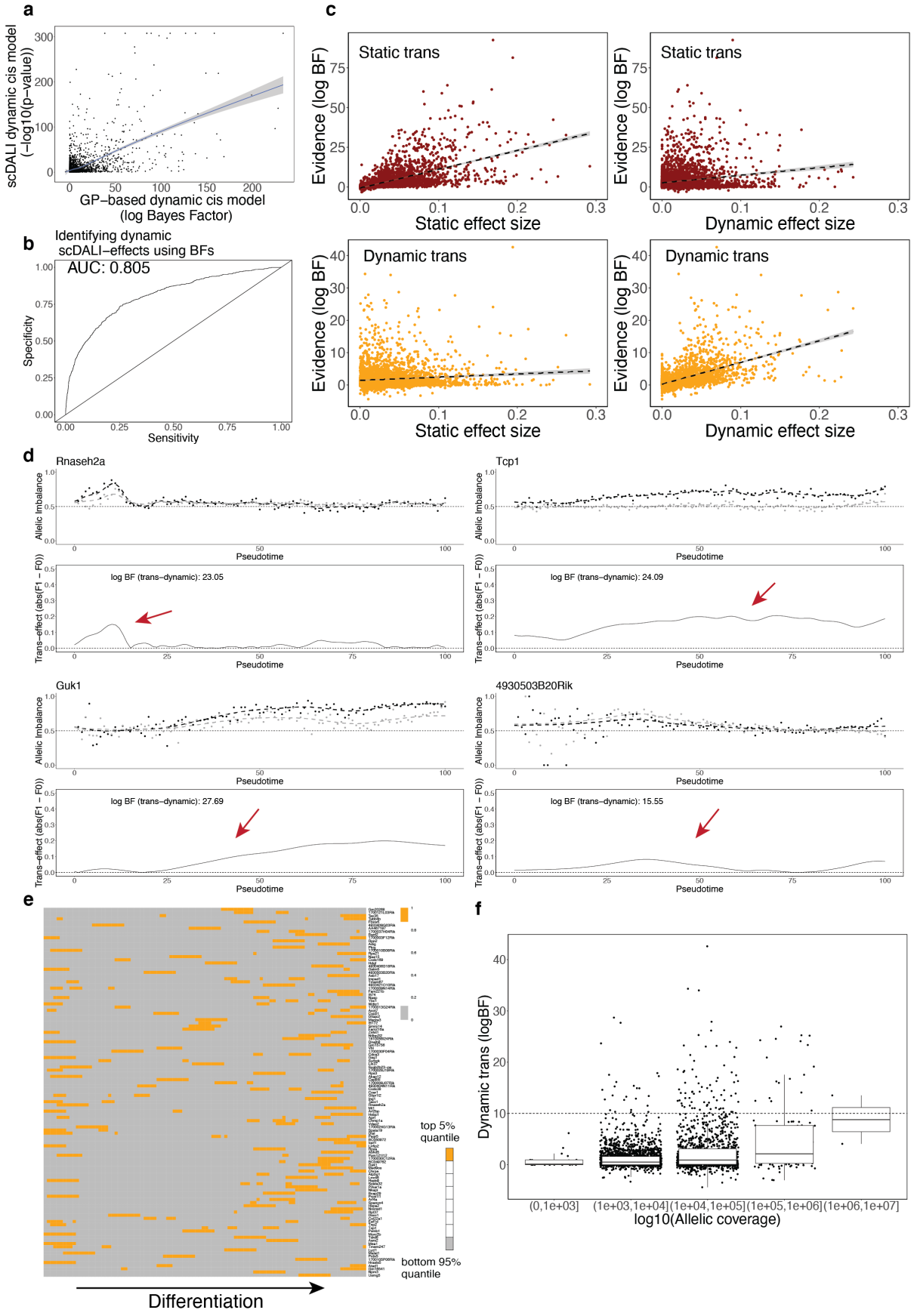

Figure 2.14: (Caption on the next page.)

Figure 2.14: Features and examples of dynamic *trans*-effects during spermatogenesis. (**a**) Scatterplot showing the correlation between log Bayes factor and -log10(p-value) to test for dynamic *cis*-effects. Smoothing line is derived by linear regression. (**b**) The same information, visualized as a ROC-curve, to show correspondence between the two models. (**c**) Scatterplots showing the association between persistent and dynamic *trans* effects sizes and evidence from the two models. (**d**) Examples of dynamic *trans*-effects. Derived estimates for F0 and F1 allelic trajectories are shown in every case, with the *trans*-effect size, of which the most extreme points are highlighted by arrows. (**e**) Heatmap showing the intervals with strongest *trans*-effect sizes across pseudotime. (**f**) Evidence for dynamic *trans*-effects as a function of allele-specific coverage, showing that expression level is not the main determining factor in detecting genetic effects.

In summary, I describe a novel approach to detect dynamic *trans*-effects in a mouse model system. These are relatively uncommon and are not more dynamic than *cis*-effects. I also show that similar to *cis*-effects, *trans*-effects are associated with reduced sequence constraint. Finally, I show that there is a moderate co-occurence of dynamic *cis*- and *trans*-effects on the same genes, that is stronger than the co-localization of persistent effects.

2.4 Species-specific gene expression dynamics are mainly driven by *cis*-effects

Through the previous sections, I have demonstrated that a single-cell snapshot of gene expression in a complex mouse tissue allows the analysis of both differentiation-dependent genetic effects. I found that *cis*-effects are both more common and stronger than *trans*-effects, which is in line with previous studies highlighting the predominance of *cis*-effects in shaping strain- and species-specific gene expression [Wittkopp et al., 2004, Goncalves et al., 2012]. My data now allows us to determine the impact of *cis*- and *trans*-effects on cell type-specific expression divergence: One could theorize that a) it would be mainly a function of cell type-specific *trans*-actors or b) it could be mainly a function of how the *trans*-regulatory environment interacts with *cis*-acting effects in different contexts.

2.4.1 Dynamics in species-specific gene expression is mainly caused by *cis*-effects

To address this question, I asked how similar the allelic trajectories were between F0 parental (*cis*+*trans*) and F1 alleles (*cis* only). To this end, I first identified genes where strain-specific expression varied across differentiation using a dynamic modelling approach equivalent to the on used in the previous section. This analysis identified 924 genes, a similar amount as genes with dynamic *cis*-effects. I compared these dynamics to trajectories of genetic effects by using joint hierarchical clustering into seven clusters. This analysis shows which genes show comparable behaviour over pseudotime, and whether divergence is more similar to *cis*- or *trans*-effects. I found that differential divergence and *cis*-components showed highly similar behaviours over pseudotime, whereas *trans*-effects differed markedly **(Fig. 2.15a-c)**. Indeed, 63.6% of genes with dynamic expression divergence showed dynamic *cis*-effects, as opposed to only 12.0% showing dynamic *trans*-effects **(Fig. 2.15d)**. Also, clusters largely contained the same genes between divergence and *cis*-effect patterns **(Fig. 2.15e)**. These results demonstrate that species-specific expression dynamics are very similar to the activity of dynamic *cis*-effects, demonstrating that as persistent *cis*-effects drive tissue-level differential expression, dynamic *cis*-effects drive species-specific context-specific divergence [Goncalves et al., 2012].

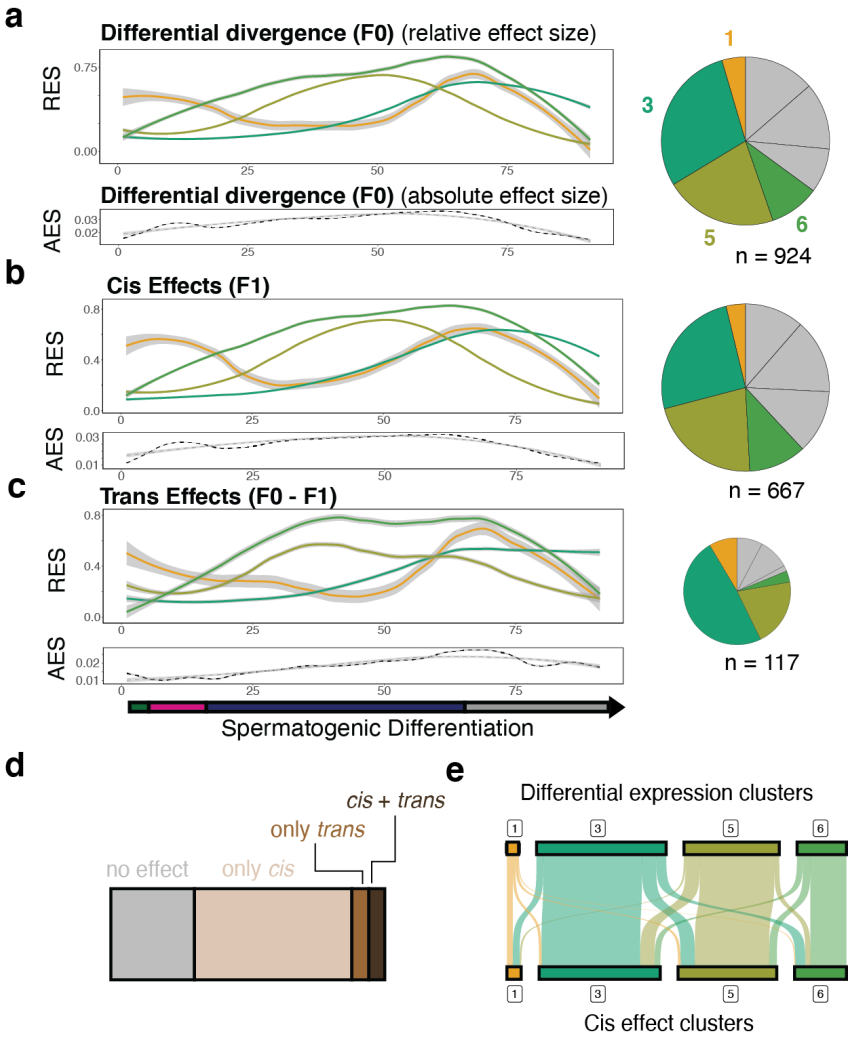

Figure 2.15: Trajectories of genetic effects compared to expression divergence. (**a**) Average F0 allelic ratio effect sizes (absolute deviation from 0.5) trajectories for different clusters. These are scaled to their maximum (relative effect size, RES). The lower panel shows the average absolute effect size (AES) interpolated during pseudotime. The pie-chart shows how genes are distributed across clusters. (**b**) As (**a**) but for allelic ratios measured in the F1. (**c**) As (**a**), but showing the absolute difference between F0 and F1 trajectories.

2.4.2 Round spermatids show higher transcriptional divergence across sub-species

I finally addressed if genetic effects are, on average, stronger in specific cell types. To this end, I quantified the average strength of expression divergence, *cis*- and *trans*-effects across pseudotime intervals **(Fig. 2.15a-c, lower panels)**. This analysis revealed a strongly increased concentration of expression divergence during the late round spermatid stage in all three quantifications. Previous work has demonstrated that spermatids show stronger gene expression divergence than other testicular cell types and that this drives accelerated evolution of transcription levels in testes [Shami et al., 2020, Murat et al., 2022, Soumillon et al., 2013]. These results were shown to hold across primate species and broad vertebrate evolution. One specific study analysed expression divergence between FACS-sorted mouse spermatocyte and spermatids, also suggesting that transcription levels changed faster in spermatids [Kopania et al., 2022]. My results confirm and extend these findings to sub-species of mice. I also show that expression divergence escalates during the course of differentiation. Furthermore, previous studies mainly implicated pervasive transcription of testis-specific genes in this phenomenon. I demonstrate that this increased divergence is partly driven by stronger genetic effects on widely expressed genes, which suggests that transriptional regulation is under lower constraint in this cell type.

2.4.3 Stronger transcriptional divergence extends to closely related species

I finally sought to validate and extend these findings by using a third strain of mice, *Mus caroli*, that represents a full sub-species which can not produce offspring with *Mus musculus domesticus* or *Mus musculus castaneus*. I used a previously generated single-cell RNA-Seq dataset from testis of C57BL6, CAST/EiJ and CAROLI/EiJ in biological duplicates. I processed the scRNA-Seq data in an analogous way to the F1 dataset and confirmed high reproducibility in detected genes and celltypes between samples and strains **(Fig. 2.16a-c)**. As in the previous dataset, the individual samples readily integrated providing a cross-species map of spermatogenesis with all expected cell types **(Fig. 2.16d,e)**. Similarly, marker gene expression and cell type proportions were highly consistent **(Fig. 2.16f,g)**.

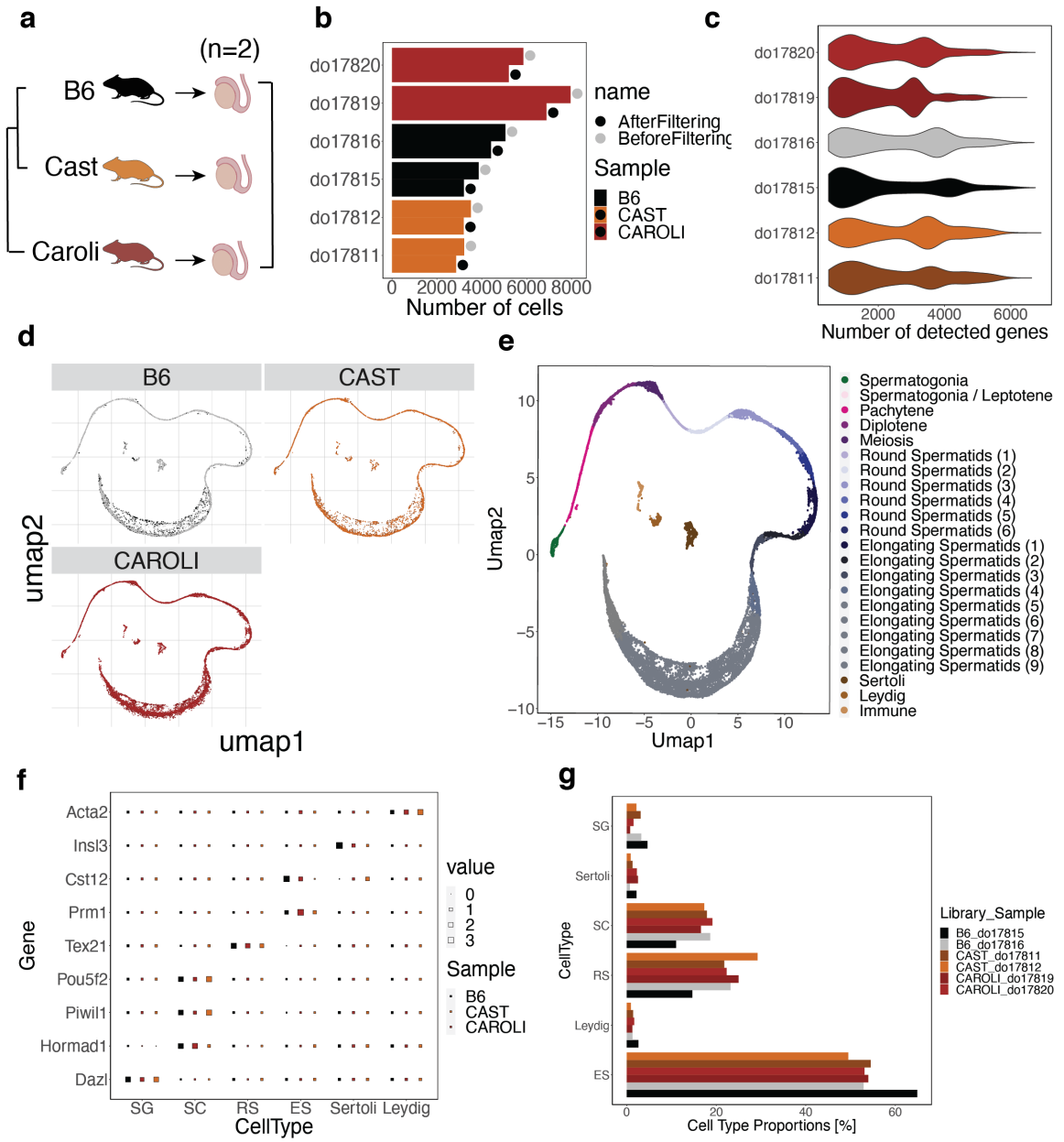

Figure 2.16: Quality control of the three-species dataset. (**a**) Overview of the sample generation including their evolutionary relationship and the number of biological replicates (n = 2). (**b**) Barplot showing the number of cells before and after filtering for individual samples. (**c**) Distribution of UMI counts per cell across individual libraries. (**d**) UMAP showing the joint embedding of individual libraries. (**e**) Cell type annotations across three mouse species. (**f**) Gene expression of marker genesacross cell types and species, for comparison see **(Fig. 2.3d)**. (**g**) Cell type distributions as fractions of total for each sample, showing high consistency.

I asked whether similar to B6 and CAST, CAROLI mice showed increased transcriptional divergence late in differentiation, using the integrated representation of spermatogenesis across the three species **(Fig. 2.17a-c)**. To this end, I performed two sets of analysis: First, similar to the last section, I asked which genes showed cell type-specific transcriptional divergence across differentiation. To make the analysis comparable to the previous dataset, I used the differentiation pseudotime defined in the F1 dataset as a reference. To this end, I identified the 5 nearest neighbours in the F1 dataset based on pearson correlation coefficients for every sample in the 3-species dataset and assigned the mean pseudotime value to the cell.

Based on this definition of differentiation, I next used Gaussian Process Regression to identify genes that varied in species-specific expression measured as the absolute difference of the allelic ratio to 0.5 over pseudotime **(Fig. 2.17d,e)**. I identified 1521 genes in the B6/CAST and 4135 genes in the B6/CAROLI comparison. Using hierarchical clustering, I found that the majority of genes peaked in differential expression late in differentiation, similarly as in the F1 dataset. Indeed, the average absolute divergence was strongest in round spermatids for both comparisons **(Fig. 2.17d,e, lower panels)**. As a second analysis strategy, I used a correlation-based approach. To this end, I partitioned cells into their respective cell types and then computed pairwise pearson correlation coefficients between 1) all pairs of B6 cells within that cell types and 2) between pairs of B6 and CAST cells. I then repeated the same analysis with B6-CAROLI pairs. I then compared the average correlation coefficients per cell type and found lower correlations between species in round spermatid stages than in other cell types, closely mirroring the results of the previous analysis **(Fig. 2.17f)**. Collectively, these analyses demonstrate increased transcriptional divergence in round spermatids across mouse sub-species, mirroring previous results in divergent species and primates.

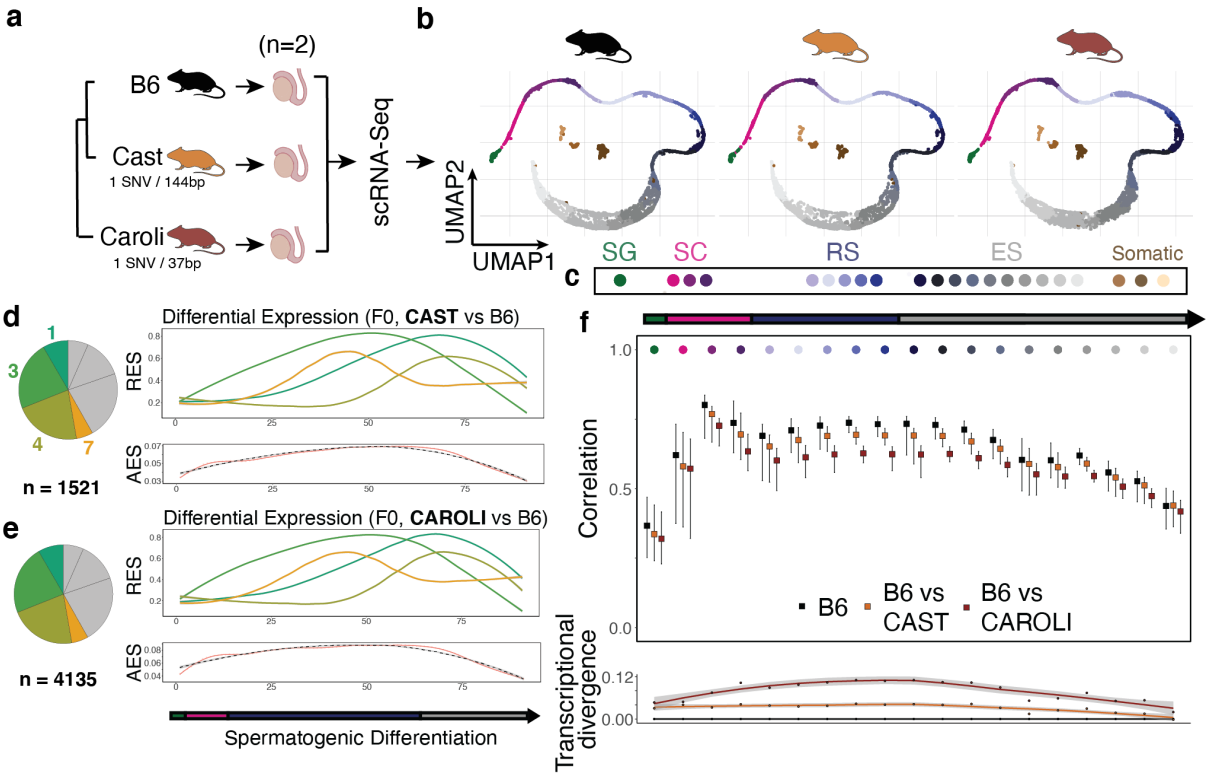

Figure 2.17: Expression divergence analysis across three mouse species. (**a**) Overview over the experimental design, including sequence divergence in SNV frequency compared to the reference genome (B6). (**b**) UMAP plots giving showing the cross-species embedding of cells. (**c**) Legend of colors to cell type groups. (**d, e**) Trajectories of expression divergence between B6 and CAST (**d**) or CAROLI (**e**) across clusters. Lower panels show average absolute divergence across genes and pie charts show distribution of clusters. (**f**) Results of correlation-based divergence quantification. Boxes show median cell-cell correlation values within B6 cells (black) and B6-CAST (light brown) and B6-CAROLI cell pairs (dark brown). Whiskers show 0.1 and 0.9 quantiles. Lower panel shows the difference in median correlation values to B6, with an interpolated line.

2.5 Discussion

Recently, a flurry of studies with increasingly high power has used single-cell RNA-Sequencing to discover context-specific genetic effects using variations of QTL mapping in human samples. Their primary conclusion is that to understand genetic effects, it is essential to measure them in the right context, which primarily means cell type, but extends to for example disease, sex and age. However, these studies are still largely underpowered, as measurements from human samples are intrinsically noisy due to unobserved covariates and genetic association studies carry a large multiple testing burden. This is particularly severe for both context-specific tests and in particular *trans*-effects, that require $\sim 10^6$ variants to be tested against $\sim 10^4$ expressed genes. Based on these studies, it is therefore difficult to assess, in a global manner, what the extent of context-dependency is for either class of genetic effects.

Our approach allows for the discovery and analysis of context-specific variation in primary mouse tissues. While genetic effects between specific mouse strains are generally of lower interest than the genetic impact of variants on human traits, my approach provides a number of unique advantages. I only require a small sample size, as variation between inbred strains with defined genetics and under equal conditions is reproducible. Furthermore, up to the detection limit of single-cell RNA-Seq, my approach comprehensively assays all genes. I use this strength to show that during spermatogenesis, context-dependency is remarkably common, affecting close to half of genetic effects. Furthermore, I observed that *trans*-acting variants are less common and do not show higher cell type-specificity. I also show some evidence that both chromatin accessibility and RNA stability might contribute to allelic effects. In the future, a key contribution of genetics in model organisms could be to quantify the effect of different regulatory mechanisms to context-dependency of genetic effects by employing single-cell multi-modal readouts. This approach is starting to be used in QTL analysis settings.

One additional advantage of model organisms is that the analysis of genetic effects can be easily coupled to treatment regimens. For example, allele-specific expression in human T-cells has been analyzed to map interactions with immune responses [Gutierrez-Arcelus et al., 2020]. In mice, much more complex experimental designs can be used and analyzed *in vivo*.

A limitation of the F1 hybrid trio approach is that it does not directly link specific variants to the gene of interest, but instead integrates over all genetic variation affecting it. On the one hand, direct links will increase my understanding of the genetic architecture of expression variability. On the other, summing over multiple variants means that I likely underestimate their total action because they might cancel each other out. However, it might be reasonable to assume that for the majority of genes, single or few variants explain most of the observed divergence. Direct mapping of variant effects between two strains can be achieved through meiotic shuffling in the F2 generation followed by QTL mapping across a large number of individuals. Recent advances in single-cell technologies now allow for large sample sizes through multiplexing [Srivatsan et al., 2020], which makes these experiments feasible, when coupled to an opimized design in terms of individual and cell numbers [Cuomo et al., 2021].

In this chapter, I also present a robust approach to detect persistent and dynamic *trans*-effects by joint modelling of allelic imbalance trajectories using Gaussian Process regression. This extends the scDALI model to compare differences in allelic interpolations between conditions. However, I used a simplified model that assumes Gaussian data noise instead of using a bino-

mial likelihood, and this requires me to aggregate measurements across intervals of pseudotime and to restrict the analysis to regions of high gene expression. A full model of *trans*-effects could include the noise assumptions from [Goncalves et al., 2012] and be cast as:

$$\mathbf{y}_{B6} \sim NegBinom(\boldsymbol{\mu}_{B6}, \boldsymbol{\theta_1}) \tag{2.13}$$

$$\mathbf{y}_{CAST} \sim NegBinom(\boldsymbol{\mu}_{CAST}, \boldsymbol{\theta_1}) \tag{2.14}$$

$$\mathbf{k}_{B6} \sim BetaBinom(\mathbf{k}_{B6} + \mathbf{k}_{CAST}, \mathbf{r}_{F_1}, \boldsymbol{\theta_2}) \tag{2.15}$$

$$\mathbf{r}_{F_0} = \frac{\boldsymbol{\mu}_{B6}}{\boldsymbol{\mu}_{B6} + \boldsymbol{\mu}_{CAST}} \tag{2.16}$$

$$\mathbf{u}_{F_0} = logit(\frac{\mathbf{r}_{F_0}}{1 - \mathbf{r}_{F_0}}) \tag{2.17}$$

$$\mathbf{u}_{F_1} = logit(\frac{\mathbf{r}_{F_1}}{1 - \mathbf{r}_{F_1}}) \tag{2.18}$$

$$\begin{pmatrix}\mathbf{u}_{F_0}\\ \mathbf{u}_{F_1}\end{pmatrix} \sim \mathcal{N}(\begin{pmatrix}\boldsymbol{\mu}_{F_0}\\ \boldsymbol{\mu}_{F_1}\end{pmatrix}, \begin{pmatrix}\mathbf{K}_{F_0,F_0} & \mathbf{K}_{F_1,F_0}\\ \mathbf{K}_{F_0,F_1} & \mathbf{K}_{F_1,F_1}\end{pmatrix} + \begin{pmatrix}\sigma_1\mathbf{1} & \mathbf{0}\\ \mathbf{0} & \sigma_2\mathbf{1}\end{pmatrix}) \tag{2.19}$$

This would allow to better account for variation in total expression levels and give reliable results at low read counts.

Besides serving as a model for the structure of human genetics, closely related mouse species can serve as a model for evolutionary dynamics. The F1 trio approach has previously shown that pure *trans*-effects are rare, and shown that expression and transcription factor binding changes are usually directly resulting from sequence variation acting in *cis* [Wittkopp et al., 2004, Goncalves et al., 2012, Wong et al., 2015, Stefflova et al., 2013]. I have now shown that cell type-specific action of genetic effects can be similarly predicted from *cis*-action alone.

In particular, my analysis confirms results from divergent species which states that the testis and in particular spermatids show a faster accumulation of transcriptional changes compared to other tissues and cell types [Murat et al., 2022, Shami et al., 2020]. Indeed, I show that this increased divergence is already observable in closely related species. I did not observe the same increase in divergence for elongating spermatids, however, this is likely due to lack of active transcription and reduced mRNA amounts in these cells. An important contribution is that previous studies largely attributed this increase to pervasive and spermatid-specific expression of quickly evolving genes. I show directly, that *cis*- and *trans*-driven genetic effects show increased activity in spermatid differentiation stages. This suggests that round spermatids not only tolerate pervasive gene expression, but also show less constraint against changes in gene regulation through genetic variants.

Chapter 3

The landscape of escape from X-inactivation in immune cells

Overview: In female cells, one X-chromosome is inactivated, leading to mono-allelic expression of most X-linked genes. However, a number of genes escapes X-inactivation (XCI), retain full or partial bi-allelic expression from both active and inactive X (Xa/Xi) and might increase gene dosage in a sex-specific manner. It has been suggested that these dosage changes contribute to sex-specific phenotypes, in particular autoimmune disease. However, it is largely unknown which genes escape in adult immune cells, how their expression level varies across cell types and how it evolves during ageing. I therefore developed a single-cell based approach to deconvolve random XCI across cells and to quantify escapee expression. I show that X-inactivation is largely complete in adult immune cells, but that cell type-specific escape exists. Strikingly, T-cell types that expand strongly during ageing show a global increase in escape, while other cell types retain a stable Xi. This global rise in escape affects leads to an increased expression dosage of potentially functional genes, including *Med14* or *Cxcr3*. Furthermore, integration of escape measurements with continuous cell state annotations links instability of the Xi to exhaustion phenotypes in T-cells. I finally confirm these results at the level of chromatin accessibility. Cell type-specific escape is associated with an active chromatin state and ageing leads to a globally more accessible Xi in CD8+ memory T-cells. This work shows an unanticipated level of variation in epigenetically driven allelic imbalance across cell types and ages, while also providing a first map of X-inactivation in complex tissues.

Contributions: This study represents joint work between the Stegle, Odom and Heard labs. The project was conceived by Stefania del Prete and Agnese Loda. Stefania del Prete performed all experiments related to this project and parts of the single-cell RNA-Sequencing data preprocessing. I performed all other computational analysis and generated all figures used in this chapter. Agnese Loda and Edith Heard gave input on the interpretation of results. Duncan T. Odom and Oliver Stegle supervised the project. A paper including most findings from this chapter is currently in preparation.

3.1 Introduction

Sex bias in disease is widespread, particularly for disorders with an immunological component. This is true for cancer, infections and especially autoimmune diseases which show sex biased incidences up to 90% [Billi et al., 2019]. Sex bias is multifactorial and can be driven by differential exposures to environmental effects, lifestyle choices or the microbiome, but also intrinsic biological factors including sex hormone levels and chromosome complement [Dhakal et al., 2022, Xing et al., 2022, Takahashi and Iwasaki, 2021]. NK cell numbers are higher in males, although they are less functional, and CD4+/CD8+ T-cell ratios are higher in females [Cheng et al., 2023]. Innate immune responses including Toll-like receptor activity and antigen-presenting cell efficiency are likewise higher in females. These cellular and molecular features are thought to cause enhanced immune responses associated with better clearance of infections, but also an increased propensity to autoimmune disease [Takahashi and Iwasaki, 2021].

An emerging area of research has linked escape from X-inactivation to changes in immune function through specific X-linked genes known to participate in immunity, including *Kdm6a, Tlr7, Cd40lg* and *Cxcr3* [Souyris et al., 2018, Youness et al., 2021, Oghumu et al., 2019]. Lymphocytes have further been suggested to show alternative mechanisms to maintain silencing of the Xi which might predispose them for increased escape [Yu et al., 2021, Wang et al., 2016]. However, a direct link between chromosome complement and immune phenotypes has not been established. Also, reports of immune cell specific escape remain anecdotal and the heterogeneity across the many cell types of the immune system is not well described.

The immune system changes during ageing, which presents a risk factor for autoimmunity [Goronzy and Weyand, 2012] and drives organismal changes across non-immune tissues ("inflammaging") [Yousefzadeh et al., 2021, Chung et al., 2019, Franceschi et al., 2018]. For example, there is an increase in inflammatory cytokines, a shift from naive to memory lymphocytes and a reduction of the antibody repertoire [Santoro et al., 2021]. These processes have been associated with loss of epigenetic integrity [Keenan and Allan, 2019] and increased transcriptional variability across cells [Martinez-Jimenez et al., 2017]. Ageing has also been hypothesized to destabilize the Xi and lead to increased escape [Grigoryan et al., 2021, Peeters et al., 2014]. In particular, the *Otc* has been shown to increase in escape during ageing, which might be due to telomere shortening resulting in a loss of silencing integrity on the Xi [Brown and Rastan, 1988, Schoeftner et al., 2009]. However, this hypothesis has not been tested directly at genome and tissue scale.

In general, it remains unclear to which extent escape from XCI is context-specific. In mice, X-linked genes with Y-linked gametologs (*Kdm6a, Kdm5c, Ddx3x, Eif2s3x*) and genes in the pseudo-autosomal region in humans are known to escape across contexts, likely because this retains dosage balance between XX and XY [Tukiainen et al., 2017]. Independently, genes have been found to re-activate after XCI or to escape in a context-specific manner, where context can refer to different tissues, cell lines or developmental stages ([**?**, Richart et al., 2022]). However, escape might also vary between cell types and even individual cells in either dynamic or clonally heritable patterns [Peeters et al., 2014, Carrel and Willard, 2005].

To close these gaps in knowledge, I used profiled escape from XCI in a dataset encompassing

mouse immune cell types and ages *in vivo*. I develop a novel approach to quantify escape using single-cell sequencing that does not rely on transgenic mice or clonal populations and directly provides cell type-specific measurements. I show that silencing of the Xi is largely stable, but specific genes show cell type-specific escape. By integrating single-cell transcriptomics, chromatin accessibility and allelic expression, I show that within cell types, the level of escape is associated with specific transcriptional programs. During ageing, the Xi is stable in most cell types, but cells with high proliferative capacity like CD8+ memory T-cells show increased escape and a more open chromatin landscape. These effects may confer clonal advantages and affect the stability and function of the T-lymphocyte pool in a sex-specific manner, which might impact disease-specific immune responses in people with multiple X-chromosomes.

3.2 Profiling sex- and age-specific gene expression in the mouse spleen at single-cell resolution

I sought to set up an experimental and computational system that would allow for the quantification of escape from XCI in different cell types and tissues using RNA-sequencing technologies. Escape has previously been inferred from sex-biased gene expression, or from allelic bias in the presence of skewed XCI using bulk transcriptomics [Balaton et al., 2015, Oliva et al., 2020, Qu et al., 2015, Berletch et al., 2015, Carrel and Willard, 2005, Tukiainen et al., 2017, Andergassen et al., 2017]. However, to definitively measure escape, it is required to 1) identify the active and inactive haplotype in the sample of interest and 2) to measure expression with full allelic resolution to resolve how much is derived from the inactive one. This is complicated by the fact that XCI is random and therefore healthy tissues contain a mixture of cells with maternal or paternal haplotypes inactive.

So far, most studies assessing escape have circumvented this problem by creating an experimental system with clonal, fully biased XCI whose status can be determined *a priori* **Fig. 3.1a)**. This can be achieved by using cell lines with clonal origin, as the XCI status of expanded cell will be passed through mitosis. While this approach can not be used to directly measure escape *in vivo*, transgenic mouse models have been employed to create clonal XCI populations *in vivo*. By monoallelic deletion of the *Xist* gene, all cells in the blastocyst will inactivate the other allele, where *Xist* is normally expressed. In these clonal populations escape can be directly measured from the allele-specific expression levels of the inactive X chromosome. Using the mouse as a model provides the further advantage that interspecific crosses with high SNP densities allow for high-resolution mapping of allele-specific expression. While powerful, this model requires transgenics that might affect the X-inactivation process itself. I reasoned that single-cell sequencing approaches should be able to assess expression from the inactive X directly, because the Xi of each individual cell can be directly identified from the data, while also providing information on specific cell types and states. This approach has been used sporadically in human cells, but never to globally assess the landscape of XCI in complex tissues [Tukiainen et al., 2017, Wainer Katsir and Linial, 2019, Garieri et al., 2018].
To allow for high resolution allele-specific measurements, I utilized data generated from an F1 cross between the C57BL6 and CAST/EiJ strains, equivalently to the one used in **Chapter 2**. Native spleen cells were isolated from whole tissues by dissociation, MACS-based dead cell exclusion and were immediately used for parallel single-cell RNA- and ATAC-sequencing. No further FACS-based enrichment of specific cell types was performed, but instead all cells

present in the tissue were captured. Using the Chromium 10x based platform, 4-6 samples per experimental condition were generated **Fig. 3.1b)**. This included both males and females to compare escape and sex-biased gene expression, young (2 months) and aged (24 months) animals to assess the impact of ageing on escape and both cross directions (that is, swapping the maternal and paternal genotypes) to rule out imprinting effects. Finally, 21 mice from a secondary cross (SPRET/EiJ x B6) were included, to validate the results in an additional genetic background that has an even higher SNP-density. In total, the dataset I am using contains 79 animals processed for scRNA-Seq **Fig. 3.1c)**.

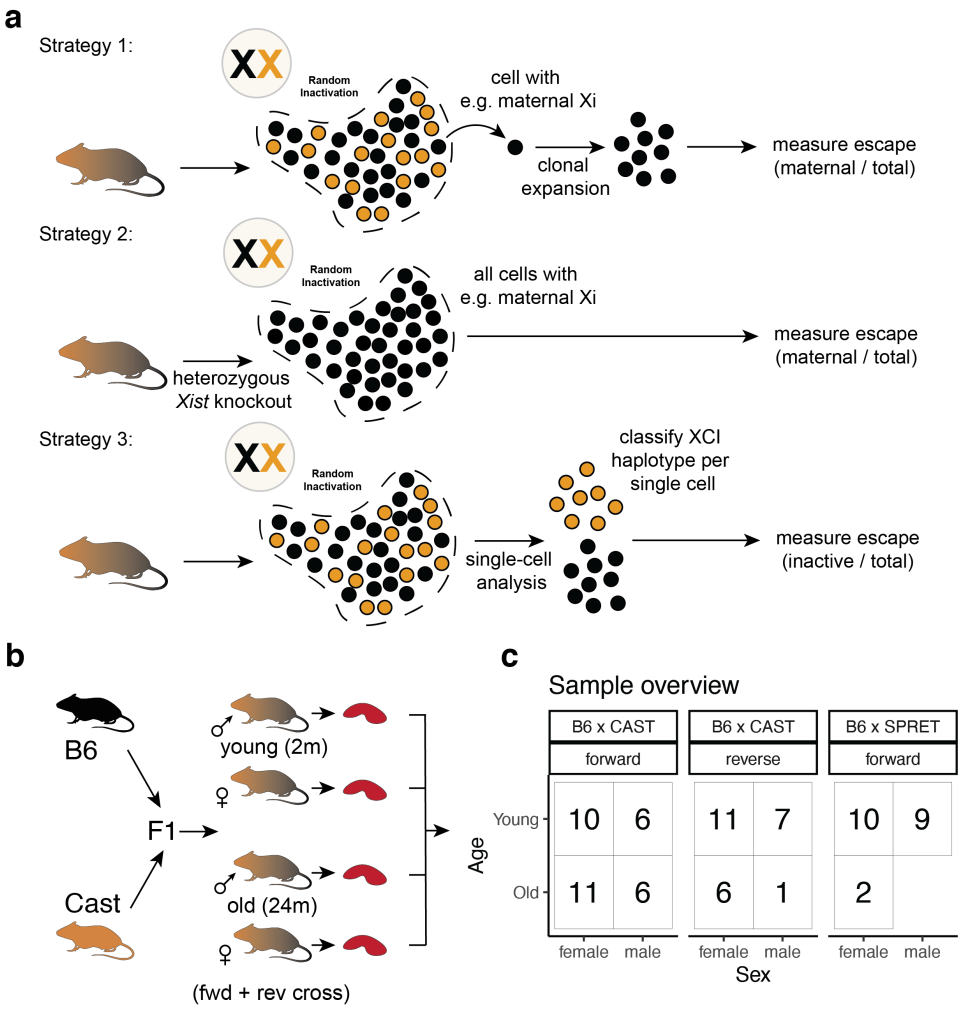

Figure 3.1: Strategies to measure escape using allele-specific RNA-Sequencing. (**a**) To quantify expression from the Xi by RNA-Seq, one needs both allelic resolution and the haplotype of the Xi has to be known in each cell. This can be achieved by clonal expansion of a single cell which propagates the X-inactivation status into the population (Strategy 1). Strategy 2 involves fully biasing XCI towards one allele by deleting *Xist* on the Xi. In this project I am using allele-specific single-cell measurements to identify the Xi and expression thereof directly within each cell (Strategy 3). (**b**) Experimental overview. First generation F1 hybrids are generated by crossing female B6 and male CAST mice (forward) or male CAST and female B6 mice (reverse). For both young (2 months) and old (24 months) mice, spleens are isolated and processed for single-cell RNA-sequencing. (c) Table depicting the number of analyzed mice. These include 21 mice from a B6 / SPRET cross which are not considered for any subsequent analysis.

3.2.1 Computational analysis of single-cell RNA-Sequencing data

I first assessed the data quality, reproducibility and representation of cell types in the two datasets. Low-level analysis of single-cell RNA-Sequencing data largely follows the procedure outlined in **Chapter 2**. I used CellRanger to align and quantify per cell-expression using an N-masked mm10 reference genome. I therefore only assess loci with comparable genomic structure in the B6 and CAST strains. I noted that inclusion of intronic reads substantially increased the number of UMIs per cell (not shown), likely because resting lymphocytes of which the spleen is primarily composed are expected to show comparatively little cytoplasmic RNA, and therefore included these in the quantification. This also likely provides a subtantial improvement on allele-specific quantifications, as a larger genomic space is used for the quantification (see **Section 3.3**).

I identified high quality cells by removing barcodes outside of four median absolute deviatons (MADs) from the sample average in both UMIs and detected genes per cell, as well as cells with >5% of mitochondrial transcripts. Across samples, I retained a median of 5693 cells per sample and 1283 UMIs per cell, which was comparable across biological conditions and experimental days. I excluded a small number of samples with low cell numbers and low average UMIs per cell (4 samples) **(Fig. 3.2a,b)**. Among the remaining libraries, the different covariates distributed similarly across dimensionality reductions generated by tSNE **(Fig. 3.2c-e)**.

To classify cell types in the final dataset, I used a semi-supervised strategy. After correcting for sequencing depth and sample-effects using mutual nearest neigbours-based batch correction, louvain clustering identified 35 clusters in the data **(3.2f, g)**. I then used a per-cell reference-based annotation to assign most similar cell types from the ImmGen database and annotated each cluster as B-cells, T-cells or other based on a majority vote within each cluster. I then re-clustered the data and used per-cluster marker genes, automatic annotations and dimensionality reductions to assign a second level cell types to every secondary cluster **(3.2h-j)**. I further identified likely doublet clusters with high doublet detection scores and removed clusters that were difficult to assign as outliers as well as two clusters likely composed of erythrocytes. I validated these cell type assignments by assessing marker gene expression **(3.2h-j, bottom panels)**.

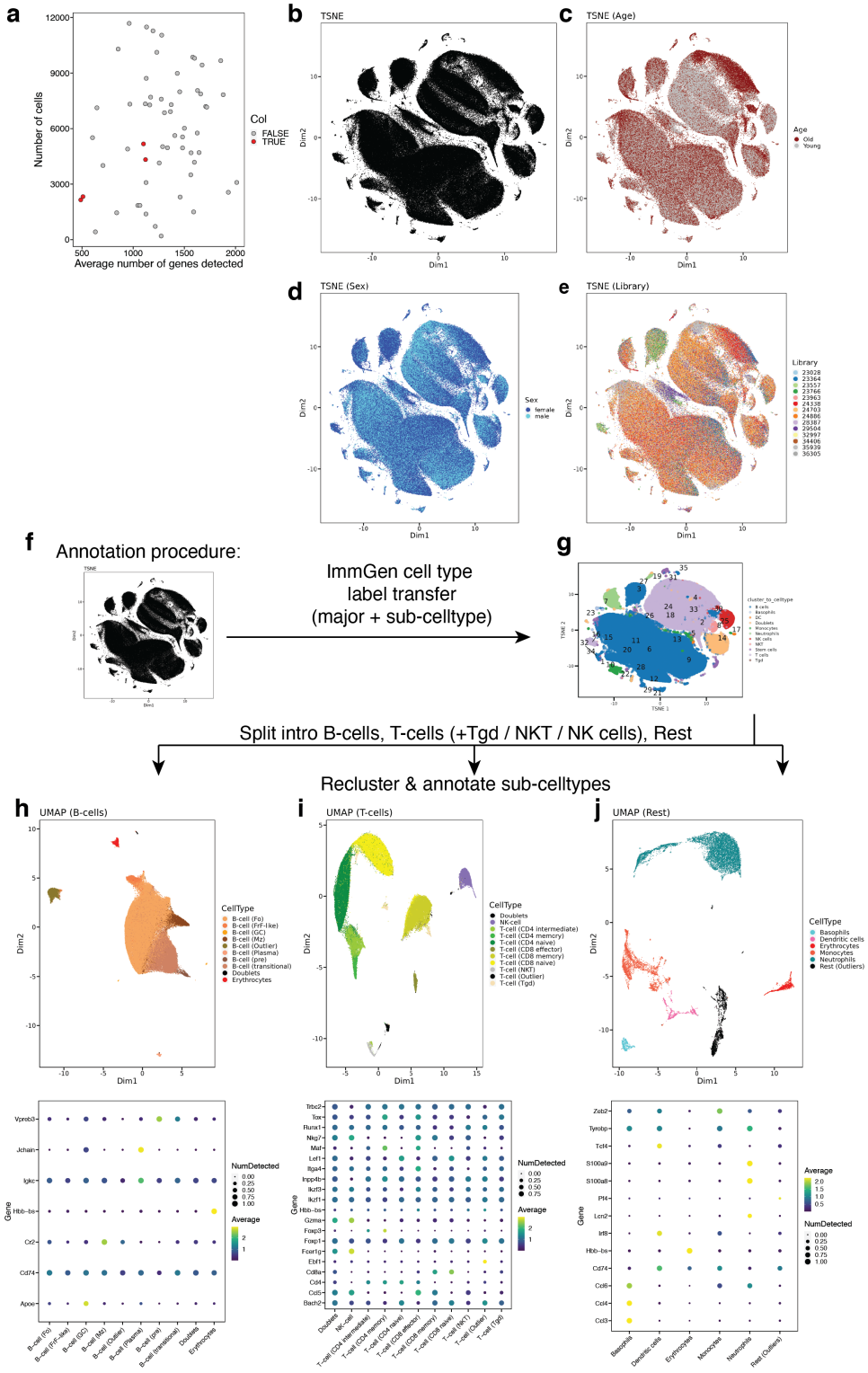

Figure 3.2: QC metrics, dimensionality reductions and cell type annotation of the immune XCI dataset. (**a**) Scatterplot showing the median number of genes detected per sample against the amount of cells in that sample. Color showing four excluded samples. (**b**) TSNE plot depicting all cells in the scRNA-Seq dataset and resolved by age (**c**), sex (**d**) and batch (sequencing library) (**e**). (**f, h**) Overview over the annotation procedure. Label-transferred cell types are collapsed at the cluster level (indicated by numbers in (**f**)). B-cells, T-cells and others are then isolated, re-clustered and annotated at the sub-celltype level, represented as UMAP plots (**h-j**). Dotplots in the bottom row indicate marker gene expression level for sub-celltypes.

3.2.2 Age-effects on cell type-distribution and gene expression

After removing doublets, technical artefacts and erythrocytes, all cells in the dataset could be assigned to expected splenic immune cell populations. The dataset is primarily composed of B- and T-lymphocytes, small myeloid populations of basophil and neutrophil granulocytes as well as antigen-presenting monocytes and dendritic cells. Among B-cells, I found pre-B-cells which are the first stage migrating into secondary lymphoid organs, differentiate into transitional and ultimately follicular B-cells which make up the majority of lymphoid follicles [Lewis et al., 2019]. I furthermore observe distinct marginal zone (Mz) and germinal center (GC) B-celltypes which correspond to distinct anatomic locations in the spleen, as well as rare plasma B-cells. Among T-cells, I could distinguish regulatory CD4+ and cytotoxic CD8+ subtypes, and both unstimulated naive and stimulated effector as well as memory populations. Within CD4+ cells I observed a population with naive and memory features which I dubbed ”intermediate”, and I observed small numbers natural-killer like T-cells and gamma-delta T-cells, as well as distinct natural killer cells. I are using these main classes of cell types for all subsequent analysis **(Fig. 3.3a)**.

To more directly assess the consistency of expression measurements, I next assessed sample- and cell type-level similarities between individual libraries. Samples showed largely reproducible cell type fractions, with follicular B-cells being the most common cell type, followed by naive and effector T-cell types, NK-cells, marginal zone and transitional B-cells and finally small populations of myeloid cells **(Fig. 3.3b)**. To assess reproducibility at the gene expression level, I performed principal component analysis on sample-level pseudo-bulk aggregates **(Fig. 3.3c)**. I found that clustering of a subset of samples was strongly driven by experimental batch (first experiments performed), which I excluded to minimize technical artefacts. After removing these samples, clustering in principal component space was less strong, and largely driven by differences between young and old samples **(Fig. 3.3d)**. I next performed principal component analysis on pseudo-bulk aggregates of individual cell types for each cell type which represents more 100 cells in at least five samples. Differences between sexes and ages were apparent in some PCs, in particular for CD8+ memory T-cells, where PC1 separated samples by age **(Fig. 3.3e)**. I note that experimental batch and condition is confounded in this dataset (not shown), and detected differences should be validated at the level of batches.

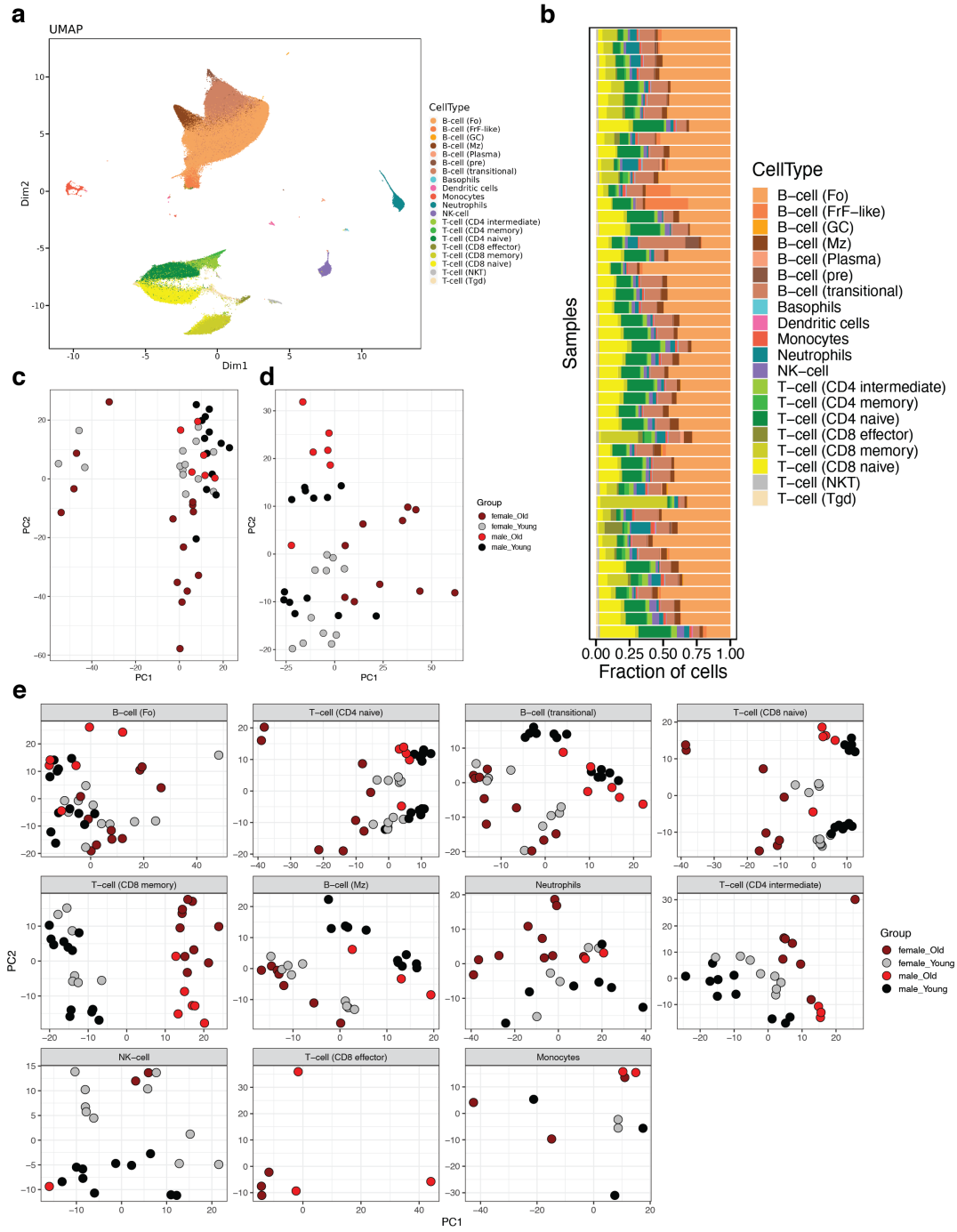

Figure 3.3: Exploratory analysis and outlier detection. (**a**) Scatterplot showing a UMAP-projection of the filtered dataset, with doublets and outlier cell types removed. (**b**) Barplot showing the fraction of cell types across individual samples. (**c**) Scatterplot showing the first two principal components from a PCA on pseudo-bulk aggregates of all samples. Colors indicate sex- and age of the shown libraries. (**d**) as (**c**), but with early batches excluded. (**e**) as (**d**), but PCA is run on pseudo-bulk aggregates per cell type seperately. Only samples with at least 100 cells of the indicated cell type are shown, and cell types are shown if at least 5 samples show at least 100 cells.

Previous single-cell studies in mice have revealed profound changes in gene expression and cell type distributions in aged compared to young spleens [Mogilenko et al., 2022, Kimmel et al., 2019, Lewis et al., 2019] I therefore quantitatively assessed variation in cell type distributions by performing variance component analysis on a binomial generalized linear model of each cell type fit to cell type fractions. This model can be written as

$$k_{c,s} \sim Bin(n_s, p_c) \tag{3.1}$$

$$log(\frac{p_c}{1-p_c}) = \beta_0 + \beta_{Age} + \beta_{Sex} + \beta_{Sex,Age} \tag{3.2}$$

where $k_{c,s}$ is the number of cells in a sample for the tested celltype, n_s is the total number of cells in that sample p_c is the latent fraction of cells. p_c is then decomposed into age, sex and interaction effects β_{Age}, β_{Sex} and $\beta_{Sex,Age}$. To obtain a global measure of each covariates effect, I computed model divergences when removing individual predictors for each cell type. The fraction compared to the intercept-only model stands in as a variance-explained-like measurement as in ordinary linear regression.

I found that sample age explained the most variance, whereas sex, cross direction and experimental batch only explained little, suggesting that biological variation induced by age exceeds technical biases **(Fig. 3.4a)**. Directly testing for the significance of β_{Age} and β_{Sex} showed no sex-biased cell types, but confirmed strong age-related changes (likelihood ratio test, FDR < 10%). Comparing across cell types, all classes of naive, effector and memory T-cells, as well as Tgd and NK-cells were strongly affected by ageing with differentiated cells increasing at the expense of naive cells **(Fig. 3.4b, c)**. These effects, including age-related loss of NK cells has previously been shown, and suggested to affect immune responses [Hazeldine and Lord, 2013, Mittelbrunn and Kroemer, 2021]. Additionally, I observed an increase in Plasma B-cells, consistent with previous results [Frasca and Blomberg, 2009]. Finally, follicular B-cells, neutrophils and monocytes increase slightly in proportion, which is potentially an artefact resulting from a relative increase of these cell types due to the massive loss of naive T-cells, a drawback of the non-absolute quantification through single-cell RNA-sequencing. Finally, I tested for interaction effects between sex and age in cell type distributions, that is, for significance of the $\beta_{Sex,Age}$. I only found one cell type with a significant interaction effect after multiple testing correction, namely for neutrophils **(Fig. 3.4d)**. It should be noted that their low transcriptional complexity makes it challenging to identify these in scRNA-Seq data [Grieshaber-Bouyer et al., 2021]. These results suggest that splenic ageing is largely sex-independent at the cell type level, at least for phenotypes that can be quantified reliably by single-cell RNA-sequencing.

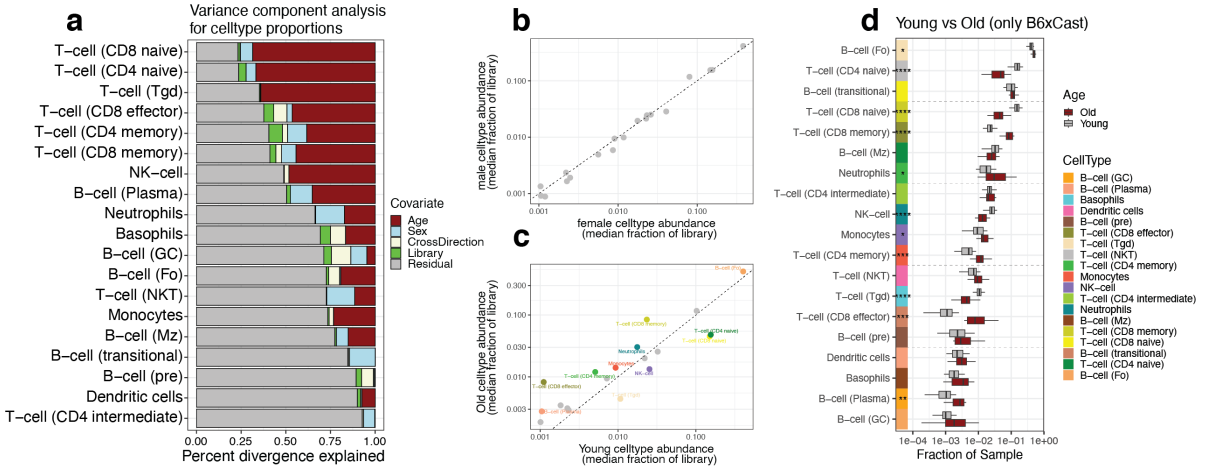

Figure 3.4: Analysis of cell type proportion differences between sexes and ages. (**a**) Barplot showing the results of a variance component analysis of cell type fractions against multiple covariates in the dataset. On the x-axis, I show the relative change in model divergence when including the respective covariate. Residual divergence is measured against a saturated model that exactly predicts the data. (**b**) Scatterplot showing the average cell type fractions across female samples against the fractions of male samples. No significant differences (through testing in the model defined by equation 3.1) are observed. (**c**) As (**b**), but comparing young and old conditions. Cell types with significant differences are shown by color and name (likelihood ratio test, FDR < 10%). (**d**) Boxplots showing the distribution of cell type fractions across sexes and ages for each cell types. Only neutrophils show significant interaction coefficients $\beta_{Sex,Age}$ (likelihood ratio test, FDR < 10%).

3.3 Mapping of escapees at single-cell resolution

3.3.1 Elimination of mapping bias in allele-specific mapping of X-linked genes

Having established that the dataset is of high quality, recapitulates expected differences between sexes and ages, and that these conditions are well comparable, I next sought to explore escape from XCI by quantifying allele-specific expression from the X-chromosome. I considered that it would be critical to ensure reliable allele-specific mapping between the B6 and CAST haplotypes to obtain unbiased measurements. I quantified allele-specific expression (ASE) by mapping sequencing reads to an N-masked genome and assigning reads to the maternal or paternal haplotype based on the presence of known heterozygous single-nucleotide variants. This quantification can be affected by problems in mapping (specific alleles might cause reads to align differently than the alternative allele), genotyping (SNVs might differ between the reference database and the analyzed animal), sequencing or PCR errors (read might contain a wrong base) or ambient RNA (dying cells are known to release RNA, which might then be measured in a different cell). I therefore analyzed allelic bias at the gene level in male mice, where, depending on the maternal genotype, only one X-chromosome is expected. The majority of genes in male mice showed only genes derived from the maternal haplotype, although a small fraction of reads ($< 1\%$) were derived from the opposite haplotype, likely representing sequencing errors. However, a smaller subset of genes showed biallelic or opposite allelic polarization, which is easiest explained by genotyping errors or strong mapping bias **(Fig. 3.5a,b)**.

As some of these genes were highly expressed, in a genome-wide analysis, I found that this lead to not fully polarized measurements from the X **Fig. 3.5c)**. I therefore removed all SNPs with more than 1% of reads mapping to the wrong haplotype in males, which eliminated the off-target X-linked expression. I observed that by excluding reads that overlap SNPs biased in male samples removed a number of genes that were outliers compared to the proportional relationship between the two alleles in female samples as well **(Fig. 3.5e,f)**. Meanwhile, autosomes showed the expected allelic ratio of 0.5 in both sexes, whereas mitochondrial haplotypes were only expressed from the maternal haplotype. Females show X-expression ratios of around 0.3, suggesting unequal X-inactivation between the haplotypes **(Fig. 3.5c, d)**.

In total, the corrected allelic ratios follow the expected distributions across sexes and chromosomes. The presence of genotyping artefacts in this widely used model with published genomes and variant sets highlights the need to quality control the genotype assignments used for allelic analysis. The here presented strategy of including male hybrid mice should generally be effective to reduce false-positive signals during the analysis of escape.

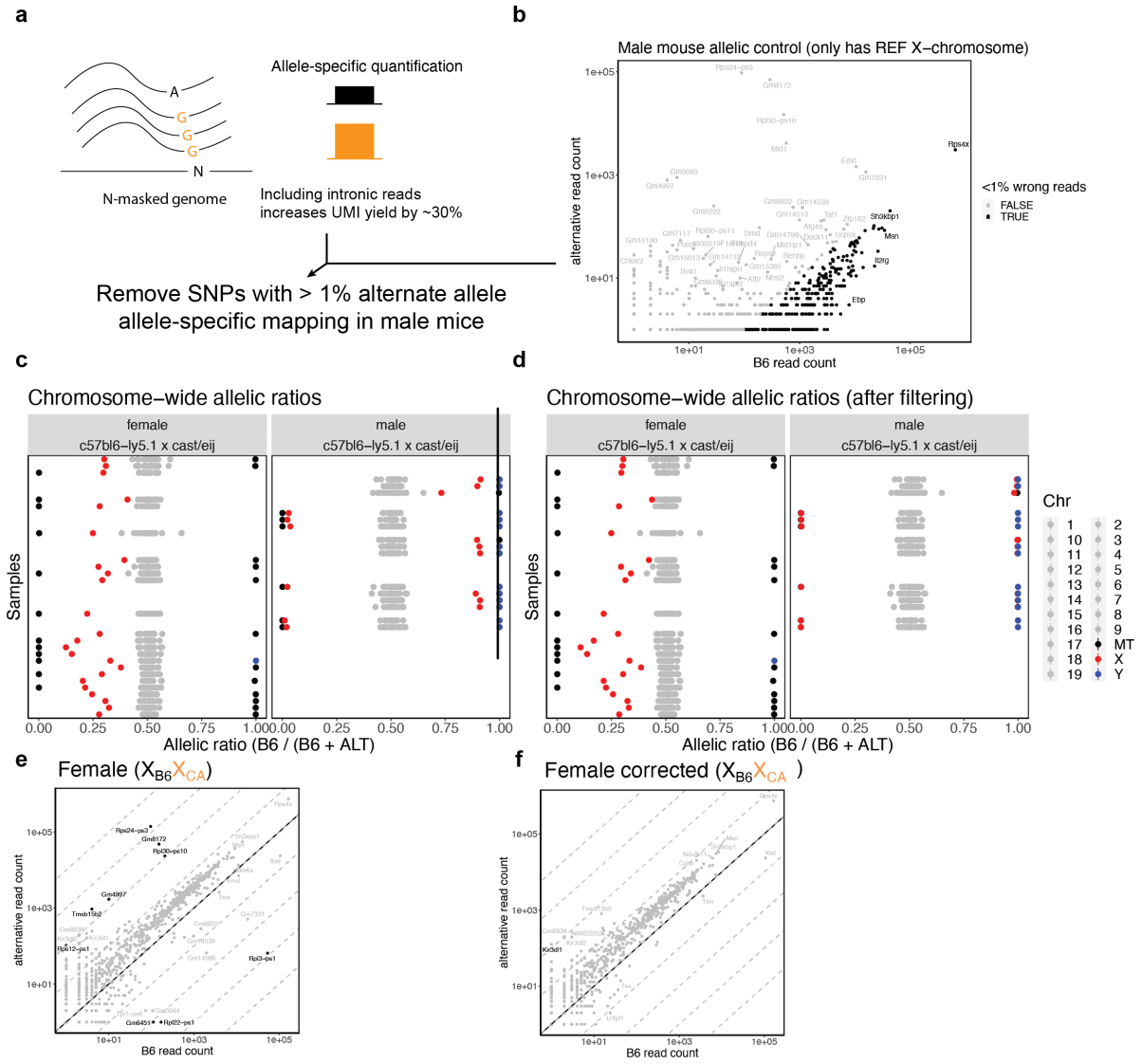

Figure 3.5: Addressing quantification biases in X-linked allelic counts. (**a**) Overview over the mapping strategy. (**b**) Scatterplot showing per-gene allelic read counts of all X-linked genes in a single male mouse with a B6 X-chromosome. Genes with more than 1% of UMIs assigned to the wrong genotype are coloured grey. (**c, d**) Chromosome-wide allelic ratios obtained by summing reads across genes, separating female and male individuals. Grey points represent autosomes, black the mitochondrial genome, red the X-chromosome and blue the Y-chromosome, which shows unreliable quantifications due to the low number of allelically assignable reads (**c**). On the right, after excluding SNPs identified in (**b**) from the quantification. (**e, f**) as **b**, but showing a single female sample (**e**), and the same data after excluding SNPs with off-target mapping identified in male mice (**f**). Note the reduced number of outliers from the proportional relationship between B6 and CAST counts.

3.3.2 Identifying the inactive X in single cells

In general, two approaches can be considered to assess expression from the Xi. The direct approach would be to identify genes with bi-allelic signal in individual cells. However, this requires high read counts at the gene-level, which my dataset did not contain, and might not be achievable genome-wide with an scRNA-sequencing technology in cells with a low RNA content like resting lymphocytes. Indeed, I obtained a median of 1200 allele-specific UMIs per cell which corresponds to an assignment rate of 40%, and around 20 UMIs mapping to the X-chromosome in an allele-specific manner **(Fig. 3.6a-c)**. However, given ASE data with phased genome information, it is possible to assign an X-inactivation status to individual cells to quantify haplotype-specific expression by aggregating across all X-linked genes, and to then quantify expression from the Xi even with only few reads or no supporting it for a given gene. Indeed, individual cells showed a clear bimodal distribution in chromosome-wide ASE on the X, while the autosomal ratio was close to the expected 0.5 **(Fig. 3.6d)**. To ensure accurate assignments and to minimize impact of wrongly assigned cells, especially those with few X-linked reads, I designed a beta-binomial mixture model to assign cells to inactivation haplotypes:

$$k_i \sim Bin(n_i, p) \tag{3.3}$$
$$p \sim Beta(\mu, \theta) \tag{3.4}$$
$$\mu = (1-\epsilon) * a + \epsilon * (1-a) \tag{3.5}$$
$$a \sim Ber(f) \tag{3.6}$$

Here, (k_i, n_i) are the B6 and total read counts for cell i and p encodes, via μ and ϵ, the average fraction of reads from the non-inactive haplotype while accounting for overdispersion of these counts, and f encodes the probability that a cell has a B6 inactive haplotype (via the indicator variable a). Evaluating the posterior of this model given (k_i, n_i) allows for the quantification of evidence for either class as opposed to the other. I quantified haplotype-specific read counts while excluding *Xist*, fit the above model using variational inference and called cells with at least a posterior ratio >7 as B6 inactive, <7 as CAST inactive and unassigned otherwise. This procedure allowed for the assignment of more than 90% of female cells **(Fig. 3.6e)**.

Without *Xist*, this model estimates the fraction of expression from the Xi $\epsilon = 1.95\%$, confirming that XCI is nearly complete in adult cells. I note that I can not exclude the presence of individual cells that largely lack inactivation, which this approach excludes as technical artefacts. I compared this strategy to basic thresholding-based approaches. Assigning cells only based on the allelic ratio, while calling cells with less than 25% bias as unassigned, lead to a number of cells not being excluded as unassigned, whereas thresholding out cells with fewer than 10 allelic reads over-called unassigned cells that the model-based approach could classify (**Fig. 3.6f**). Finally, I confirmed that *Xist* was expressed only from the Xi, which was the case as *Xist* allelic ratios were close to zero for B6-active and close to 1 for CAST-active cells, consistently across samples (**Fig. 3.6g**). Unassigned cells were showing intermediate *Xist*-ratios, and therefore likely represent doublets.

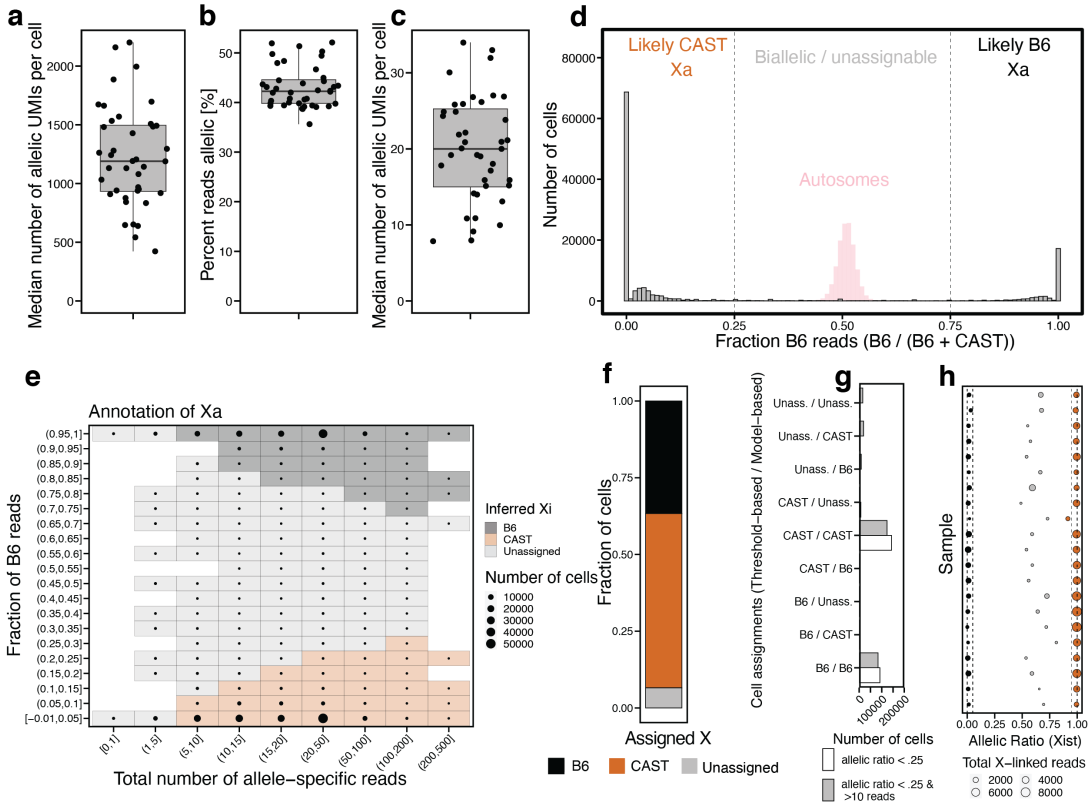

Figure 3.6: Annotating the Xi haplotype in single cells. (**a-c**) Boxplots showing for each B6 x CAST F1 sample the number of UMIs with allelic assignment (**a**), the fraction of allically assignable reads among total UMIs per cell (**b**) and the number of assignable and X-linked UMIs of the total (**c**). In each case, the shown value is the median within each sample. (**d**) Histogram showing the fraction of reads assigned to the B6 haplotype on the X-chromosome per cell. Pink indicates the same allelic ratio, but for autosomes. (**e**) Heatmap showing Xa annotation results. Cells were binned by the number of X-linked allele-specific reads and the resulting allelic ratio. The background of each tile shows whether the majority of cells in that bin are assigned as an CAST Xa (brown), B6 Xa (black) or unassigned (grey). The size of the points indicates the number of cells. (**f**) Barplot showing the fraction of cells assigned to the respective haplotypes, or unassigned. (**f**) Barplots showing comparisons between thresholding-based annotation strategies. (**f**) Dotplot showing the average allelic ratio of *Xist* across samples and stratified by inactivation haplotype. The size of the points indicates the number of cells assigned to the respective haplotype in the given sample.

To my knowledge, this is the first demonstration of an X-inactivation status map of a mammalian tissue using single-cell methods. In particular, this approach allows to assign skewed inactivation in different cell types, which, in humans, is linked to ageing and genetic disease [Shvetsova et al., 2019]. Across samples, the CASTxB6 F1 crosses showed skewed X-inactivation of a median of 75% (as opposed to unskewed XCI at 50%) **(3.7a)**. This result confirms previous reports that link different alleles in the X-chromosome control region to a higher probability of the CAST X to stay active during XCI, likely due to strain-specific *Xist* mRNA levels [Calaway et al., 2013]. I also confirm a slightly increased skew reported in the reciprocal cross (CAST female x B6 male, reverse). I did not observe significant differences in skew in aged animals. I also noted that individual mice showed variable skew in XCI, consistent with a bottleneck during X inactivation that leads to overdispersed fractions **(3.7b)**.

Second, my XCI classification assignments gives the opportunity to measure skewed XCI in

different cell types. Since all splenocytes differentiate from the same hematopoietic stem cell pool, it would be expected for them to have the same XCI skew across populations. However, I observed substantial variation in skew values across cell types. Using a binomial linear model that accounts for the baseline skew per sample, I found that all T-cell types show decreased skewing compared to the sample average, that is, more cells within the T-cell population have an active B6 X-chromosome than expected from the sample average. This holds when comparing T-cells to myeloid cells (Macrophages, DCs and Neutrophils) which are more distantly related to T and B-cells in the hematopoietic differentiation hierarchy, but also when comparing them to closely related B-cells, suggesting a T-cell specific effect (**3.7c**).

Given the base assumption that the XCI status is fixed during early development, there are two likely explanations for this effect: First, the higher or lower expression of an X-linked factor could improve T-cell proliferation or survival, giving cells with an active B6 chromosome a long term advantage. I reasoned that in this case, I should be able to identify a corresponding transcriptional signature in the B6 X-active cells, however, differential expression analysis between the two inactivation genotypes did not reveal any such signature in any T-cell subtypes (not shown). Secondly, the X-linked factor could bias differentiation in B6 X-active T-cells towards the T-cell lineage, in which case the effect would not be visible in mature cells. While this data can not provide a definitive mechanism, differential expression analysis identified X-linked genes differentially expressed between B6 X-active and CAST X-active cells, suggesting potential candidates that might drive this effect (not shown).

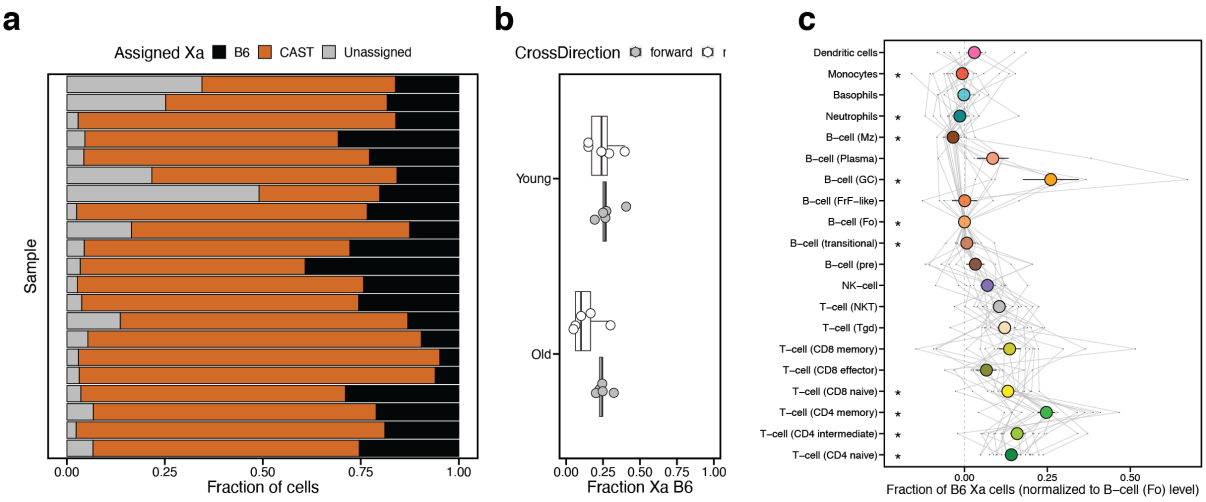

Figure 3.7: Skewed X-inactivation across cell types. (**a**) Barplot showing the fraction of cells with CAST Xa (brown), B6 Xa (black) and unassigned. (**b**) Boxplots showing the fraction of B6 Xa cells across samples, stratified by cross direction and age. Forward indicates the B6 (female) x CAST (male) cross direction, reverse the opposite. (**c**) Fraction of B6 Xa across cell types, which are ordered by relatedness in the hematopoietic hierarchy. Only instances of at least ten cells per cell type and sample are shown in small points, and grey lines connect measurements in the same sample (mouse). Large coloured dots represent mean and with standard deviations across measurements. Significance is evaluated using a beta-binomial GLM (FDR 10%), where the fractions are tested in the respective against all cell types and significant tests are shown with by a star.

3.3.3 The set of escapees in mouse immune cells

During the previous sections, I have shown that this dataset efficiently resolves both cell types, X-inactivation status and allele-specific expression on the X-chromosome. I therefore next sought to identify the set of genes that were expressed from the Xi in immune cells. As this analysis relies on the presence of variants segregating between the parental genotypes, genes without these might be missed. I therefore assessed how many genes with significant expression levels in spleen did not show sufficient variants to assess allelic expression, and found that the majority of expressed genes (>100 total reads) showed at least 50 allele-specific reads (87%, and 85% when considering genes >1000 total and >500 allele-specific reads). I note that the high observed coverage is likely facilitated by reads originating from secondary priming during reverse transcription, which increases the coverage of variants across the transcript compared to only 3'-based profiling [La Manno et al., 2018].

Allelic bulk RNA-Seq of spleen in artificially skewed models has been published previously, giving me the opportunity to directly compare the sets of identified escapees between the two technologies [Berletch et al., 2015]. To identify genes with significant expression from the Xi, I summed all reads mapping to the Xi across all cells based on the assigned Xi haplotype and compared them to the number of reads mapping to the Xa (allelic ratio $\frac{Xi}{Xi+Xa}$) **(3.7a)**. Under around a 100 allele-specific reads, measurements were noisy and likely preclude accurate quantification of escape. As expected, the majority of genes with sufficient signal (351 genes) showed low allelic ratios, with 321 genes under 5% of Xi expression and the only gene with full Xi bias is *Xist*. 30 genes possibly shows evidence for escape (allelic ratio greater 5%) **(3.7b)**. I observed that the majority of genes show a small fraction of reads mapping to the Xi, which I attribute to likely background due to sequencing errors and ambient RNA **(3.7c)**. I therefore defined escapees as genes with significant evidence for an allelic ratio greater than 5% and consistent observation across replicates. I used a beta-binomial model and compared the banded null hypothesis of an allelic rate $H_0 : p <= 0.05$ against $H_1 : p > 0.05$.

Aggregating full samples, I found that 12 genes (other than *Xist* which as the causative factor in XCI I do not consider as an escapee) showed escape from XCI (*Kdm5c, Kdm6a, Ddx3x, Eif2s3x, Jpx, Ftx, Utp14a, Pbdc1, Cxcr3, Bcorl1, 1810030O07Rik, Gm15228*) **(3.7d)**. Most of these genes are known to escape XCI in spleen and many other tissues and are either X-Y orthologs (*Kdm5c, Kdm6a, Ddx3x, Eif2s3x*), associated with the X-inactivation center (*Ftx, Jpx*) or known escapees with other functions (*Pbdc1, Utp14a*) [Loda et al., 2022, Berletch et al., 2015]. *Cxcr3*, for which I estimate 5.64% expression from the Xi, has been previously shown to escape in a small population of T-cells [Oghumu et al., 2019]. I also tested seperately if the same genes were detected from both inactive haplotypes, finding broadly comparable results for both, although the detection power on the CAST haplotype, which is three times as likely to be active than the B6 haplotype, is substantially lower **(3.7e)**. *Berletch et al.* additionally showed escape of 6 genes *5430427O19Rik, 5530601H04Rik, 5730416F02Rik, Cfp, Vsig4, Firre*, all of which are lowly expressed or inconsistent across replicates in the dataset (not shown), and confirmed *Xist, Kdm6a, Eif2s3x, Pbdc1* and *Kdm5c* with similar allelic ratios. In summary, I find that single-cell RNA-Seq reliably identifies escapees, and that the set of these in adult tissues is small.

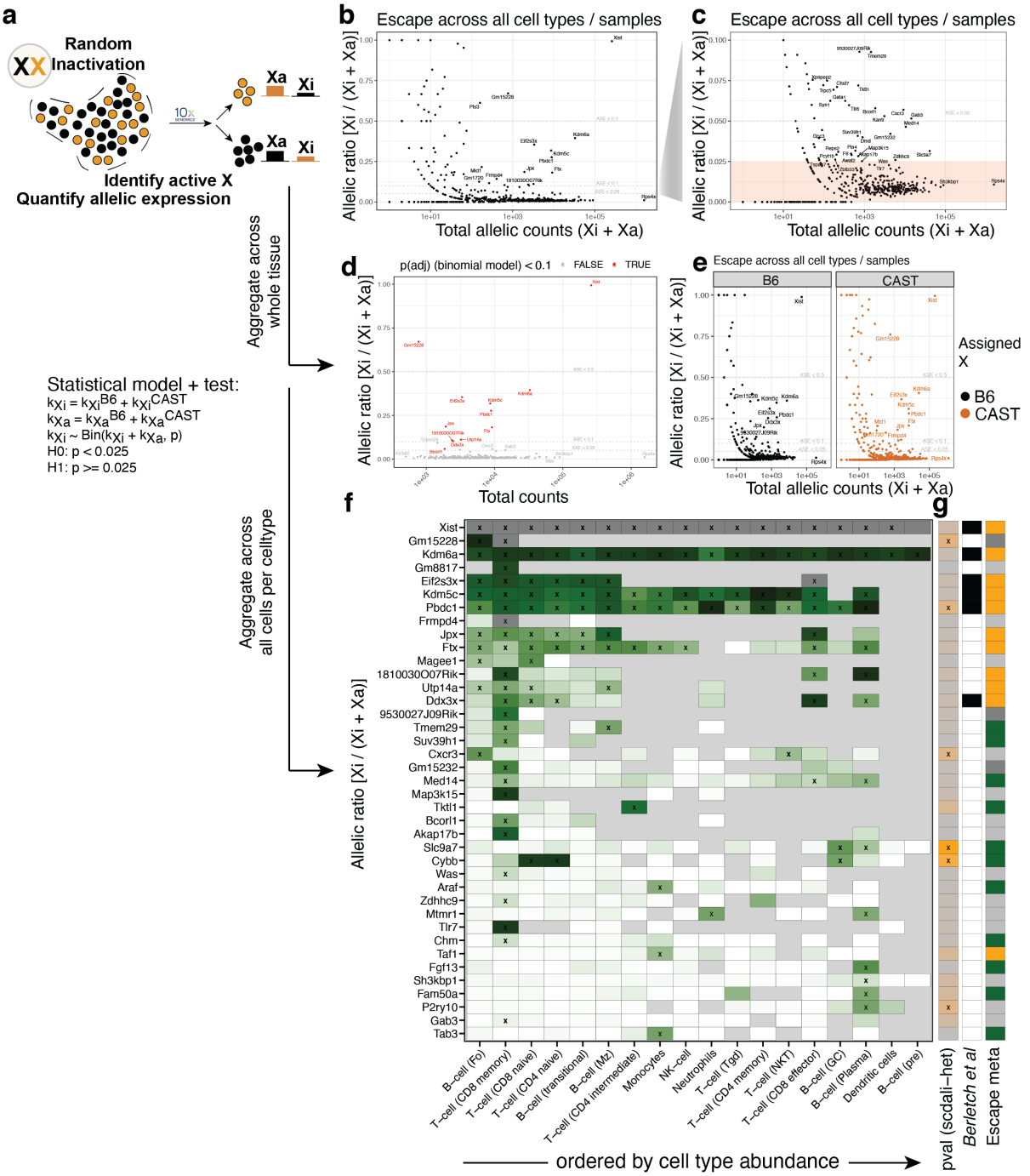

Figure 3.8: Reliable identification of escapees from single-cell data. (**a**) Overview over the analysis procedure. Cells are classified by their Xi status which allows for the direct quantification of Xi-specific expression as an allelic ratio Xi / (Xi + Xa). (**a**) Scatterplot showing total allelic signal (x) and allelic ratios (y) summarized across all female cells in the dataset. Note that allelic ratios in lowly expressed genes (left of the plot) are unreliable. (**c**) Zoom-in into (**b**), highlighting baseline allelic ratios. (**d**) as (**a**), but removing lowly expressed genes and genes with significant escape are coloured (binomial GLM, banded Wald test FDR 10%). (**e**) as (**a**), but seperating cells by assigned Xi haplotype. (**f**) Heatmap showing allelic ratios across cell types. The colour scale indicates the allelic ratio Xi / (Xi + Xa), and is capped at 0.7 (grey cells). Cells are only shown if the gene is detected in at least 3 samples in the respective cell type. An "x" in a cell indicates significant escape as in (**d**). The colour bars on the side indicate: (left) significantly variable escape across cell types (scDALI FDR of 10%), (middle): whether the gene escapes in *Berletch et al* (black box), (right) if the gene is classified as constitutive, facultative or non-escapee in the **Chapter 4** meta-analysis.

Next, I tested for escape in every cell type separately. This analysis identified 39 genes, of which 18 showed escape in at least 2 and 21 in only one cell type **(3.7f)**. These genes include *Cxcr3*, which I identify as uniquely escaping in NK-T cells. I also find that *Med14*, whose escape has recently been linked to increased proliferation in cancer cells specifically escapes in T- and most in CD8+ memory T-cells [Richart et al., 2022]. I also found that for genes with escape across multiple cell types, the escape level, that is the fraction of expression from the Xi, varied between these cell types. I therefore used scDALI to identify statistically significant variation in allelic ratios and found 6 genes *Gm15228, Pbdc1, Cxcr3, Slc9a7, Cybb, P2ry10* at an FDR of 10% . The majority of genes I found in specific cell types were not identified by *Berletch et al.*, likely because low expression from the Xi in subpopulations is difficult to detect. I also compared this set of escapees to a meta-analysis I conducted that classified genes as constitutive escapees if they were detected in the majority of a set of studies and facultative if genes were variably expressed or silenced (see also **Chapter 4**). I found that cell type-unspecific escapees largely corresponded to the set of constitutive escapees, while escapees only found in few cell types were more likely to be facultative, demonstrating that some previously observed variability in escape extends into inter-celltype variation. Finally, I note that all escapees show allelic ratios well below 0.5, which corresponds to full biallelic expression, suggesting that either the regulatory mechanisms on the Xi do not permit full expression, or that escape only occurs in a fraction of cells.

These results provide the first map of escape in a complex tissue at cell type resolution. Single-cell measurements both resolve more escapees than bulk assays, where escape in small sub-populations is masked, and identify genes with inter-cell type variation in escape. My results show previously unappreciated variability, including critical genes that have been undetected (*Med14, Bcorl1*). My approach also provides a framework to measure escape even with very limited signal per cell, and still covers the X-linked transcriptome broadly (> 80% of expressed genes), even though I acknowledge that the signal in small populations is sparse and limits the detection power.

While I identified genes with significant expression from the Xi, it is unclear whether this leads to dosage differences in gene expression. The increase in transcription due to the presence of two alleles might be offset by secondary sex-specific effects (for example due to hormone-induced regulation) and regulatory feedback or X-upregulation can modify Xa expression between sexes. In particular, it is unclear whether cell type-specific escape leads to sex-specific dosage in the affected cell types, especially since the allelic ratios and their differences are often low. I therefore compared the measured escape to the observed bias between sexes. Assuming equal expression from the Xa in both sexes and no secondary effects, the predicted sex-bias $log2(\frac{y_f}{y_m})$ due to escape would be with $d := \frac{y_{Xi}}{y_{Xi}+y_{Xa}}$:

$$log_2(\frac{y_f}{y_m}) = log_2(\frac{y_{Xa} + y_{Xi}}{y_{Xa}}) \tag{3.7}$$

$$= log_2(1/\frac{y_{Xa}}{y_{Xa} + y_{Xi}}) \tag{3.8}$$

$$= log_2(1/(1 - \frac{y_{Xi}}{y_{Xa} + y_{Xi}})) \tag{3.9}$$

$$= log_2(\frac{1}{1-d}) \tag{3.10}$$

I compared expected to the observed sex bias and found a that they did not necessarily corre-

spond **Fig. (3.9a)**. In particular, I observed genes that showed female expression bias without broad escape for *Cybb*, although this was rare, and a number of genes with male bias, even though they showed escape. I performed the same analysis across cell type, and found that genes generally showed similar patterns across the cell types they were expressed in **(Fig. 3.9b)**. However, I do confirm that some genes, including *Med14 and Suv39h1*, showed female expression bias in parallel to their escape. I conclude that escape is somewhat predictive of female expression bias, but has to be separately evaluated for each individual gene. Also, this analysis only addresses RNA, not protein levels, which might vary further.

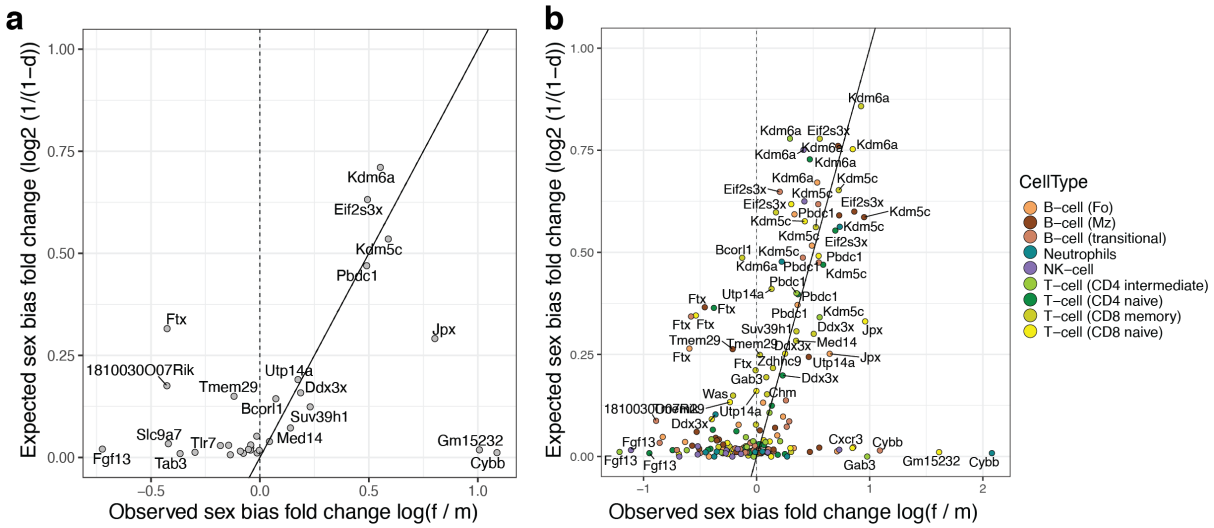

Figure 3.9: Escape partly explains expression sex bias. **a** Scatterplot showing observed (x) sex bias derived from a differential expression analysis between female and male spleens against expected (y) sex bias from measured escape. (**b**) The same analysis, showing sex-bias and escape measurements in individually different cell types.

3.4 The total activity of the X-chromosome varies across cell types and ages

Focussing on cell types rather than genes, I found the most escapees in CD8+ memory T-cells. Little is known on whether specific cell types show globally higher or lower escape **Fig. (3.8b)**. However, it has been hypothesized that in aged cells, escape might increase due to a loss of heterochromatic structure on the Xi, but this has not been experimentally demonstrated genome-wide [Schoeftner et al., 2009] **(Fig. 3.10a)**. In general, changes in total Xi expression could result from either 1) coordinated activation of many cell type-specific escapees, or 2) random activity of genes in individual cells. To capture both of these effects, I defined the global level of escape as the fraction of all Xi-linked reads compared to all X-linked reads. I excluded *Xist*, as its high expression level will strongly impact the escape measurement and its regulation is likely independent of global Xi activity, and *Rps4x* which is uniquely highly expressed, most likely globally silenced in this dataset, and therefore prone to introduce false positive background signal. I first inspected total escape ratios derived from summing all reads across cells and genes per individual mouse. As above, I found that approximately 2% of gene expression is derived from the Xi in young mice **Fig. (3.10b)**. I observed a very modest increase in total escape during ageing, suggesting that silencing of the Xi is largely stable. I considered that this small increase could be due to a larger gain in specific cell types. I therefore aimed to quantify total escape in single cells. As the total number of reads used to derive global escape estimates per cell is still low (median 20 reads), I used scDALI to interpolate the total rate of escape r between cells of similar cell types, represented by a 10-dimensional principal component-space:

$$\begin{aligned} logit(r) &= u \\ u &\sim \mathcal{N}(\alpha, \sigma_d^2 K) \\ K &= \mathbf{X}_{PC_{10}} \end{aligned}$$

Comparing these estimated escape rates between ages demonstrated that the majority of cells showed similar global escape in aged cells, but a subpopulation of cells showed increased rates **Fig. (3.10c)**. Visualizing rates on a UMAP showed an increase in global escape in specific cell types, most obviously in aged CD8+ memory T-cells **Fig. (3.10d)**. I validated these results by computing aggregated allelic ratios across cell types and individuals, focussing on cell types that were detected with at least 50 cells in 3 samples in both young and aged mice. This analysis showed that significant differences between ages were present only in NK-T cells, CD8 memory T-cells and marginal zone B-cells **Fig. 3.8e, f)**. These results are in line with the higher number of detected escapees in these cell types **Fig. (3.8f)**, and show that neutrophils show the lowest escape, followed by NK cells, follicular and transitional B-cells, Monocytes, and naive T-cells, but higher escape in all differentiated T-cells. Of note, differentiated CD4+ and CD8+ T-cells showed higher levels of global escape than their naive counterparts, suggesting that activation and differentiation of T-cells might generally be accompanied by increases in escape. Strikingly, I observed the highest levels of global escape in aged CD8+ memory T-cells, and aged cells did generally show higher levels of escape in the T-cell compartment. However, this effect is highly cell type specific, as transitional and follicular B-cells did not show this increase. This suggests that ageing might indeed affect the global stability of the Xi, but that this occurs in a cell type-specific manner.

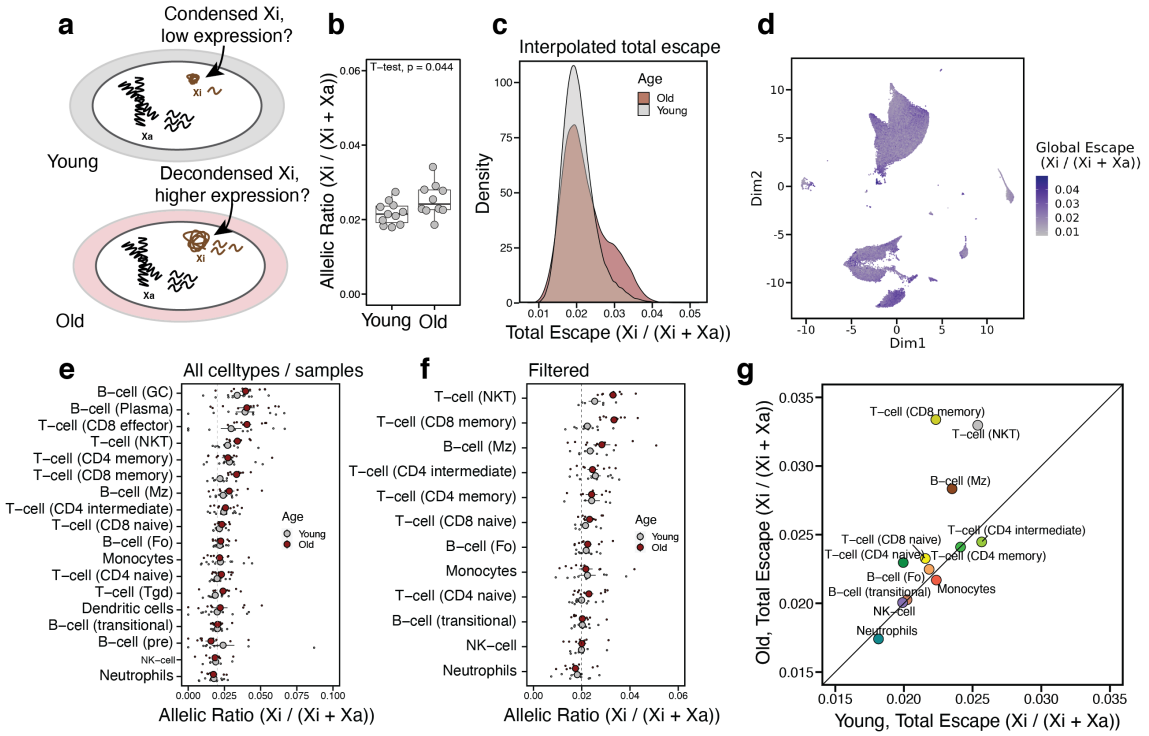

Figure 3.10: Global escape increases during ageing in specific cell types. (**a**) Diagram outlining the hypothesis of increased escape in aged cells due to loss of epigenetic integrity on the Xi. (**b**) Quantification of global escape in young and old mice. Each data point represents one profiled individual. The y-axis shows the fraction of reads from the Xi across all genes from the X-chromosome (exluding *Xist* and *Rps4x*), summed across all cells in the respective individual. Significance is assessed using a t-test. (**c**) Density plots showing expression from the Xi across individual cells. Per-cell allelic ratios are smoothed between similar cells using scDALI. (**d**) UMAP based on gene expression profiles showing the same estimated allelic ratios in different cell types. (**e**) Boxplots comparing allelic ratios between young and old samples, as in (**b**), but aggregating per individual and cell type. Significance is assessed using a t-test. (**f**) Scatterplots showing average total escape estimates per cell type across samples.

Having identified an increase of escape with age in T-cells, I next investigated which genes were affected. T-lymphocytes arise from progenitors that migrate into the thymus, where they are counter-selected to avoid auto-reactivity, and differentiate into regulatory CD4+ and cytotoxic CD8+ cells [Miller, 2011]. These naive T-cells migrate to the secondary lymphoid tissues, including the spleen and lymph nodes, where they await stimulation by clone-specific antigens **Fig. (3.11a)**. Activated cells start proliferating and differentiate into effector or memory cells and establish long-term memory of immune responses. I found that naive T-cells showed comparable escape profiles to B-cells and other immune populations, showing mainly canonical escapees and others (*Cybb*), with little changes in aged mice. Meanwhile CD8+ memory T-cells show an abundance of genes with escape in aged mice **Fig. (3.11b, c)**. However, I observe a depletion of naive and an increase in differentiated T-cells upon ageing, which makes it difficult to compare within both populations between ages. I therefore compared mixtures of naive and memory CD8+ T-cells between ages, and found that constitutive escapees showed moderate increases in escape **Fig. (3.11d)**. Additionally I found a number of age-specific escapees in CD8+ cells, in particular *1810030O07Rik, Med14, Bcorl1, Gm15232, Gab3, Zdhhc9* and *Chm*. I note that many of these *de novo* escapees are close to other escapees, in particular *1810030O07Rik* which neighbours *Med14* or *Bcorl1* and *Zdhhc9* which are close to the constitutive escapee *Utp14a*. I next asked whether escape of these genes in individual mice was a

deterministic event (for example triggered by age-related changes in the splenic environment) or stochastic event (for example caused by escape in specifically expanded clones). I found that *Med14* escape in ageing occured in all profiled individuals, *1810030O07Rik* showed escape in multiple mice, whereas *Akap17b, Tlr7* and *Marp3k15* only escaped in a single individual **Fig. (3.11e, f)**. This suggests that escape can be both stochastic and deterministic.

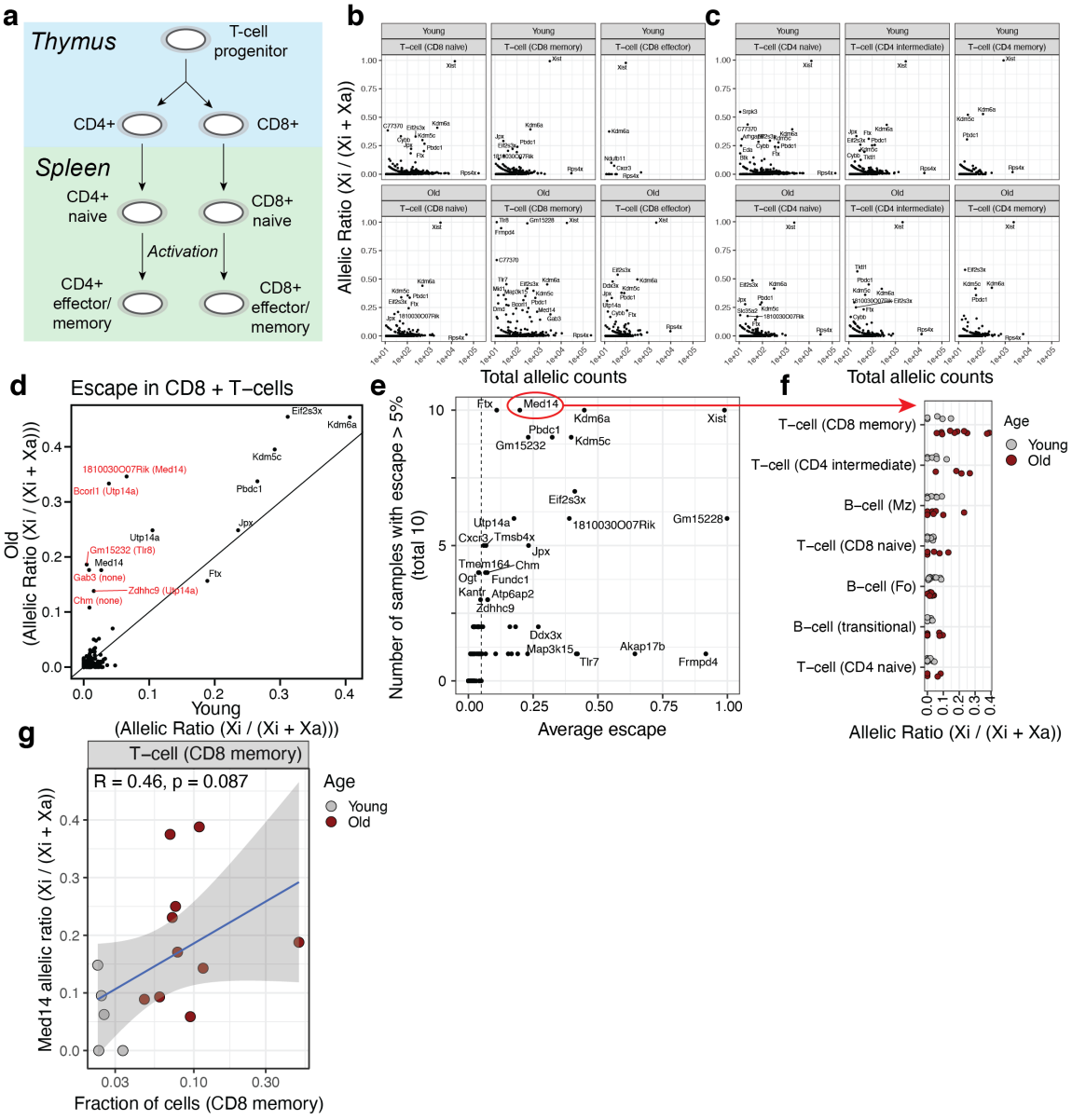

Figure 3.11: Variation in escape during T-cell ageing. (**a**) A schematic overview of T-cell development in thymus and periphery, showing the relationships between CD4 / CD8 naive / effector cells. (**b, c**) Scatterplots showing escape in young (top) and aged (bottom) mice for multiple T-cell types. (**d**) Scatterplot showing the changes in escape within CD8+ T-cells, aggregating over naive and differentiated cell types. (**e**) Plot showing consistency in escape changes across individuals. Some genes show age-related escape in all, some only in single or few individuals. (**f**) Strip plot showing changes in escape during ageing for the *Med14* gene in multiple cell types. (**g**) Scatterplot showing the relationship between fraction of CD8+ memory T-cells and Med14 escape.

I therefore propose to distinguish facultative escape that appears as ”structured” variability, for example through age-related tissue remodelling or ”random” variability, for example through clone-specific effects. *Med14* has been described as a facultative escapee previously and has been linked to proliferation and progression in breast cancer. My results present further evidence that escape of this gene might be selected for. Indeed, I observed a moderate correlation between the abundance of CD8+ memory T-cells in aged mice and the extent of Med14 escape **Fig. (3.11g)**.

3.4.1 Total escape correlates with an exhaustion phenotype in T-cells

In the previous section, I have demonstrated the analysis of escape from XCI in populations of cells using cell type and XCI haplotype annotations, to obtain cell type-level pseudobulk measurements. However, single-cell resolved data allows for more fine-grained analysis of cellular phenotypes in relation to escape. This can be achieved by correlating cellular phenotypes of interest, measured by total RNA analysis to the escape phenotypes at the single-cell level. Specifically, I considered whether in CD8+ memory T-cells, specific gene expression programs could be associated with the global level of escape from XCI. To this end, I used scDALI to perform logistic regression to predict global escape from individual highly variable genes **Fig. (3.12a)**.

I found that many genes significantly associated with global escape, some even after regressing out sample identity, were markers of T-cell exhaustion (*S100a4, S100a6, Ctla4, Inpp4b*, and in particular the exhaustion-associated transcription factor *Maf* [Verdeil, 2016] **Fig. (3.12b-e)**. Meanwhile, the escape *Kdm6a* was strongly associated, as expected since total escape is partly computed from it and is expected to increase expression levels, but *Xist* dosage was not associated, suggesting that steady-state fluctuations in *Xist* do not impact escape substantially in immune cells **Fig. (3.12c)**.

Collectively, these results demonstrate a remarkable variability in the global extent of escape, both across individual cell types and ages. Then, I have shown that T-cell exhaustion is associated with the escape state. This regression-based analysis can also be applied to all other cell types, or any other dataset, to discover associated transcriptional signatures. It can also be directly applied to individual genes, to for example identify links between upstream regulators and escapees. Of note, this correlation-based approach does not distinguish between cause and consequence. Consequently, it might yield downstream targets of overexpressed genes, or upstream regulators. It might also be a mere association, as for example exhaustion and loss of escape might be the result of clonal hyper-proliferation. Integrating functional genomics data might help with this distinction by assessing escapee-specific gene regulation.

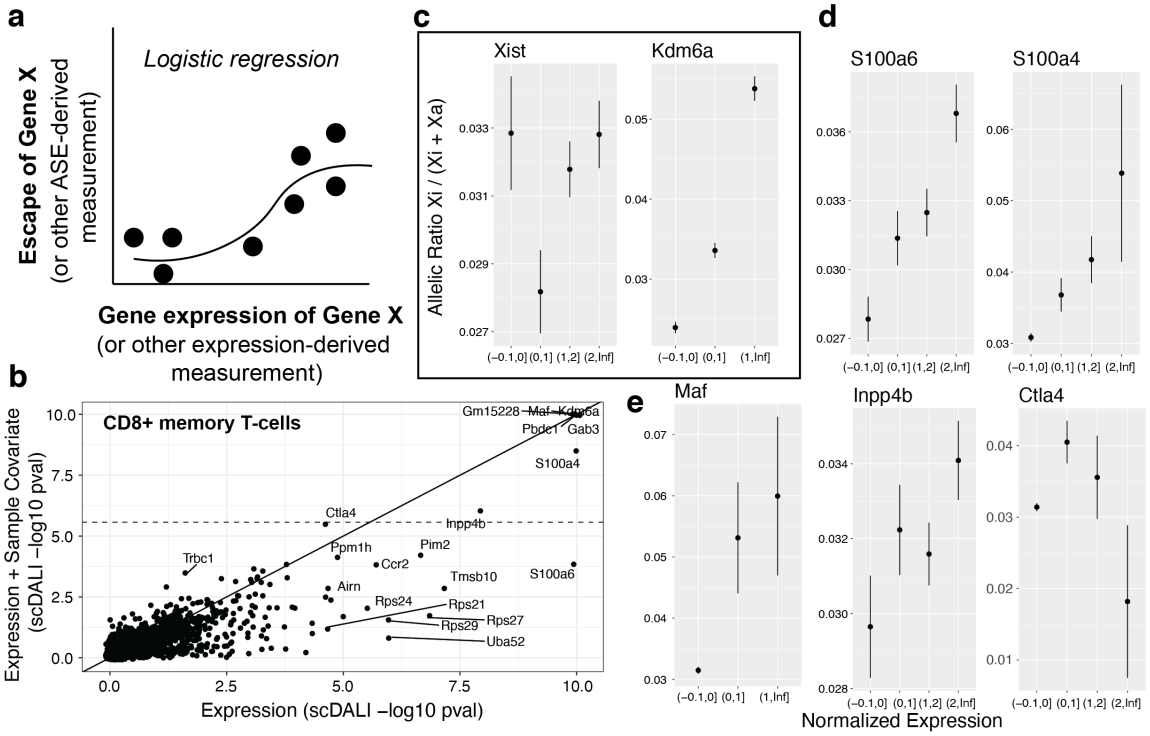

Figure 3.12: Global escape is associated with an exhausted T-cell phenotype. (**a**) Cartoon of the analysis approach. Logistic regression is performed to identify correlations between total expression-derived phenotypes and escape-derived measurements. (**b**) Scatterplot showing p-values derived from using scDALI to predict total escape from individual gene expression measurements in CD8+ T-cells. X-axis shows the results from a test without covariates, the y-axis from a test where sample identity is blocked, to remove the effects of sample-specific bias or ageing. Shown is are -log10(p-values), and the dashed line shows significance at 10% FDR. (**c-d**) Aggregated allelic ratios across binned expression levels of different genes, specifically *Xist* and *Kdm6a* (**c**), *S100a4* and *S100a6* (**d**) and *Maf*, *Ctla4* and *Inpp4b* (**e**).

3.5 The cell type-specific chromatin landscape of the Xi

Previous work has assessed the landscape of accessible chromatin on the Xi, but only rarely in adult cell types and never in specific cell types [Berletch et al., 2015, Giorgetti et al., 2016]. I therefore used scATAC-sequencing data to characterise domains of accessible chromatin on the Xi and its inter-celltype variation. To process scATAC-sequencing data, I followed a conceptually similar approach as for the scRNA-sequencing dataset using the *ArchR* package [Granja et al., 2021]. After CellRanger-based alignment to the same N-masked reference genome, I defined cells with >5000 unique read pairs and >12 transcription start site enrichment. To assign cell types, I used the annotated single-cell RNA-Seq dataset as a reference using the *SingleR* workflow, and assigned equivalent cell types manually **(Fig. 3.13a)**. The resulting structure of the UMAP-space and cell type proportions were broadly in agreement with the scRNA-sequencing data and largely reproducible across samples **(Fig. 3.13b)**. Cell filtering left the dataset with a median of 6760 cells per library with average 26502.5 read pairs and 20.6 transcription start-site enrichment, evenly distributed across experimental conditions **(Fig. 3.13c-e)**. As expected, accessible chromatin was concentrated at promoters, but was also present in intergenic regions, likely representing *cis*-regulatory elements **(Fig. 3.13f, g)**.

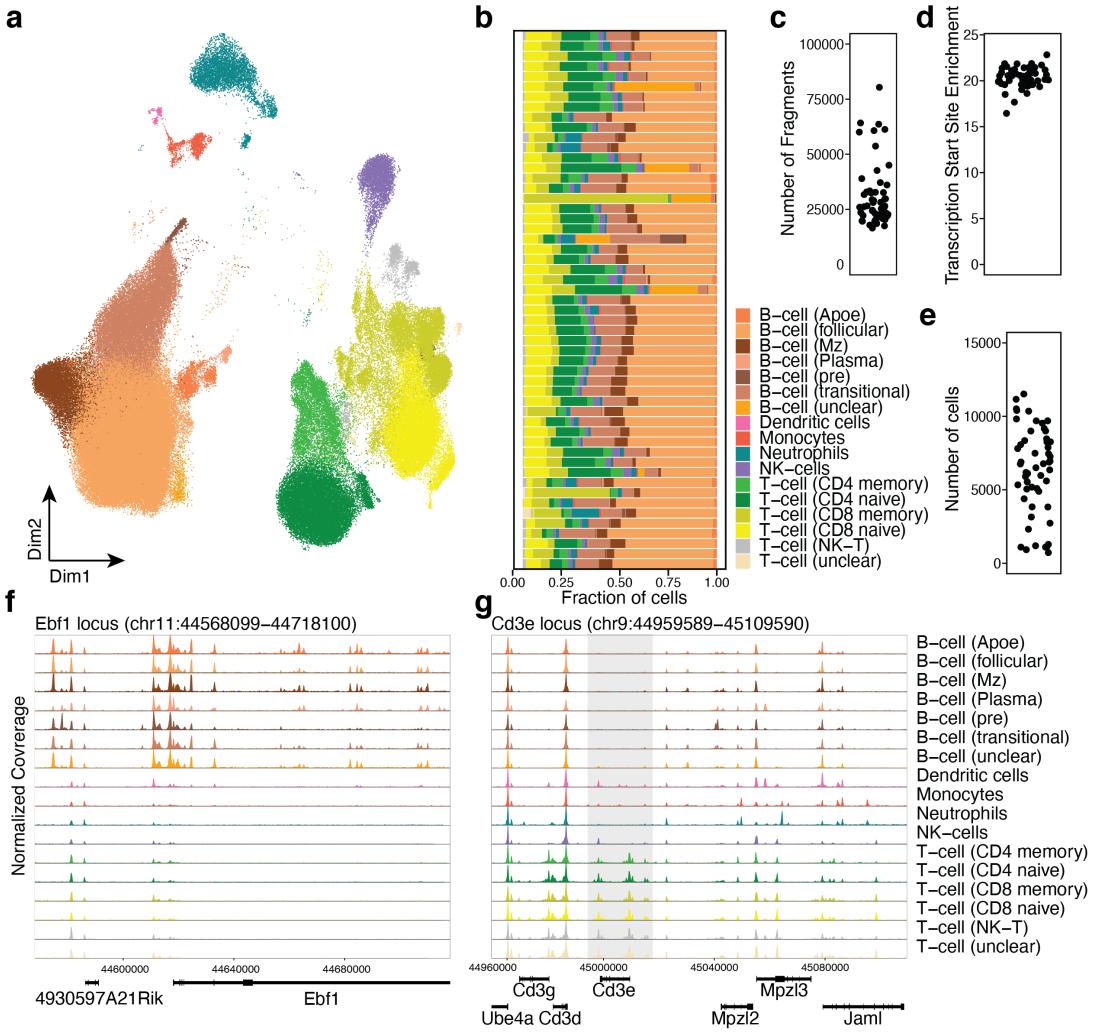

Figure 3.13: (Caption on the next page.)

Figure 3.13: Processing and cell type annotation of the scATAC-Seq dataset. (**a**) UMAP representation of all integrated samples, showing the variety of cell types. For the cell type legend, see **b**. (**b**) Cell type distribution across individual samples, showing the fractions relative to the number of cells per sample. With some exceptions, these cell type representations are highly consistent. (**c-e**) Per sample statistics, demonstrating reproducibility of quality control metrics. (**c**) Number of individual read pairs per cell. Each point shows the median value per cell for the respective sample. (**d**) Transcription start site enrichment per cell. (**e**) Number of sequenced cells per sample. (**f**) Cell type-specific coverage tracks at the *Ebf1* locus, showing B-cell specific accessibility. (**g**) Cell type-specific coverage tracks around the *Cd3e* locus (highlighted in grey), showing T-cell specific accessibility and further variation across cell types.

There was substantial variability in accessibility, reflecting he activity of marker genes, for example at the B-cell-specific *Ebf1* locus. Similarly, the T-cell marker *Cd3e* indeed showed accessibility only in T-cells. Furthermore, other regulatory elements in the region showed consistent or variable accessibility.

3.5.1 Chromatin accessibility corresponds to escapee expression on the Xi

Similar to expression, chromatin accessibility is expected to be greatly reduced on the inactive X [Giorgetti et al., 2016, Berletch et al., 2015]. I therefore used the same approach to classify cells in the scATAC-sequencing dataset as the transcriptomics-based analysis. Haplotype-specific allelic ratios in chromatin accessibility were clearly bimodal, as they were for expression **(Fig. 3.14a)**.

Using the beta-binomial mixture model presented previously I was able to classify 86.3% into B6 and CAST Xa haplotypes and finding an average Xi activity of 10.4%, showing that as a global measure, more signal is derived from the Xi in chromatin accessibility than in expression. However, this is at the cost of a reduced number of assignable cells and therefore potentially more background signal. On a UMAP-projection, different cell types show mixing of Xa haplotypes. This analysis also visually confirms the higher probability of T-cells to show an active X-chromosome than other cell types which I observed in the scRNA-sequencing data.

Across libraries I demonstrate consistent Xa haplotype ratios, and the same skewed XCI derived from the transcriptomics datasets and previously observed **(Fig. 3.14c)**. I next assessed the tissue-wide open chromatin profiles generated across all cells in the dataset. As described previously, the Xi is largely devoid of accessible chromatin with punctuate exceptions **(Fig. 3.14d)**. Furthermore, both active and inactive accessibility profiles were highly concordant between the two strains, when analyzed at this resolution.

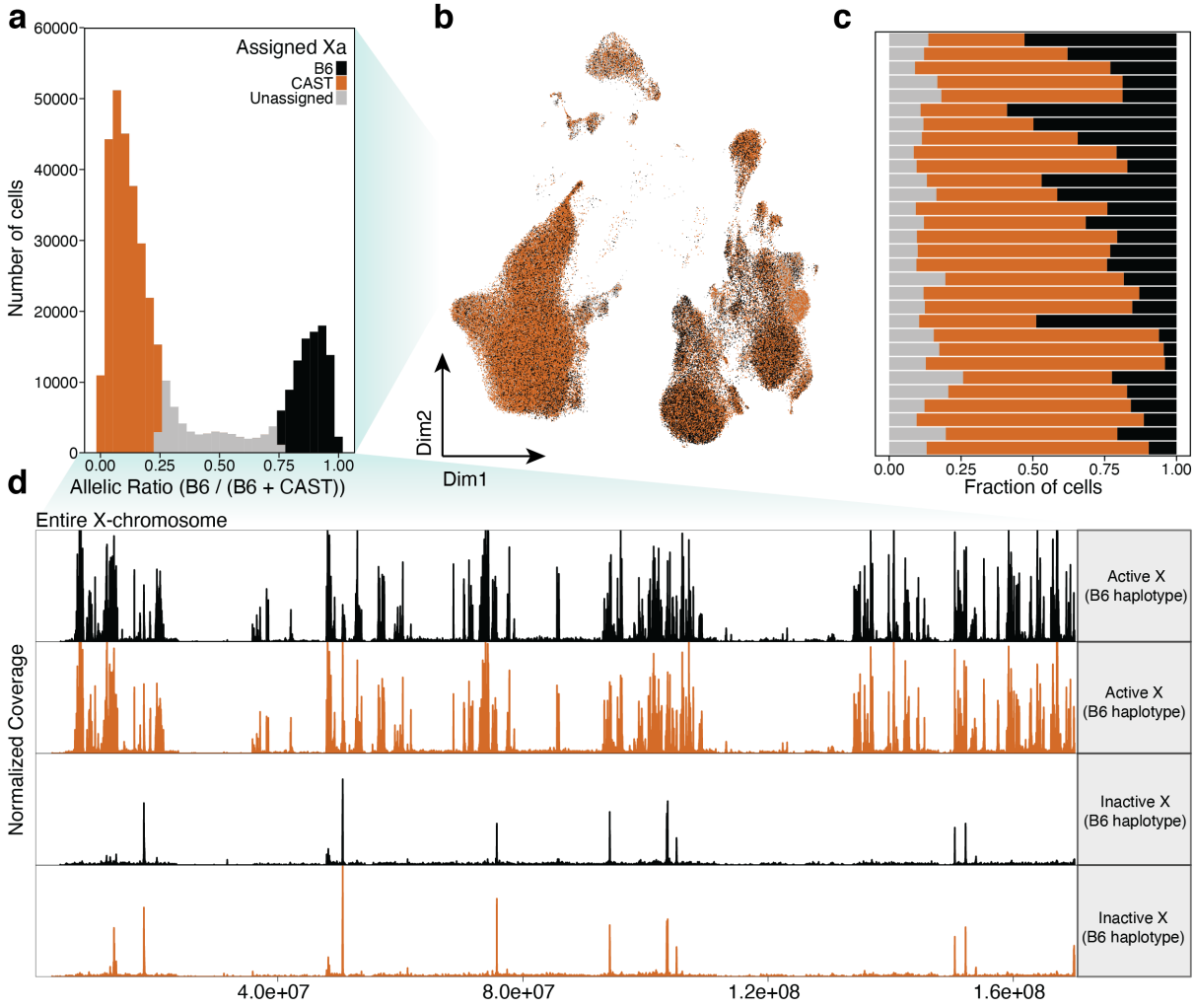

Figure 3.14: Chromatin accessibility of the Xi. (**a**) Histogram showing the allelic ratios when computing the ratio across all haplotype-specific reads. Only female cells are shown in this plot. (**b**) UMAP projection showing annotated Xa haplotypes based on mixture modelling. (**c**) Barplots showing the fraction of Xa haplotypes across individual samples. (**d**) Coverage plot showing accessibility profiles on the Xa and Xi for both haplotypes, normalized by total signal across transcription start sites and number of cells [Granja et al., 2021].

I next identified the sites of open chromatin on the Xi in a peak-based method. I used the *addReproduciblePeakSet* function from the *ArchR*-package to identify a consensus peak set based on total accessibility of the X-chromosome. I then calculated allelic read count ratios $X_i/(X_i+X_a)$ per peak across all cells. It has been suggesteed that chromatin regions accessible on the Xi are depleted of distal enhancers [Giorgetti et al., 2016]. I therefore also annotated peaks depending on whether they overlapped exons, introns, promoters (defined as regions 1kb upstream of the transcription start site of a gene) or none of those, in which case I refer to them as ”distal” **(Fig. 3.15a)**. I observed peaks of all categories to show allelic ratios greater than 0.05, indicating accessibility on the Xi. Interestingly, promoters were generally depleted from accessibility on the Xi, whereas intronic peaks made up the majority of peaks with Xi-biased accessibility (>0.5). Distal peaks were more common among low allelic ratio sites **(Fig. 3.15b, c)**. I also performed the same statistical test as in **(Fig. 3.8d)**. This analysis identified that 536 out of 3244 peaks showed significant escape when aggregating counts across all immune cell types. When focussing on specific genes carrying Xi-accessible chromatin, I found that

many known constitutive escapees showed high allelic ratios, evenly spread across accessibility levels as measured by the total number of allelic reads. In particular, *Kdm6a, Kdm5c, Utp14a, Tmem29, Ftx, Eif2s3x and Pbdc1* showed accessibility in their promoter regions **(Fig. 3.15d, e)**. In contrast to the gene expression measurements, *Xist* shows accessibility on both Xa and Xi. Finally, I compared allelic ratios derived from RNA-sequencing to ATAC-sequencing based measurements. This analysis revealed high and moderate escape in gene expression, defined as genes with an RNA-derived allelic ratio >0.2 as well as > 0.1 (and < 0.2) respectively, was associated with high or moderate ATAC-derived allelic ratios **(Fig. 3.15f)**. Indeed, the set of escapees derived from both methods overlapped strongly and allelic ratios showed a reasonable correlation **(Fig. 3.15g)**.

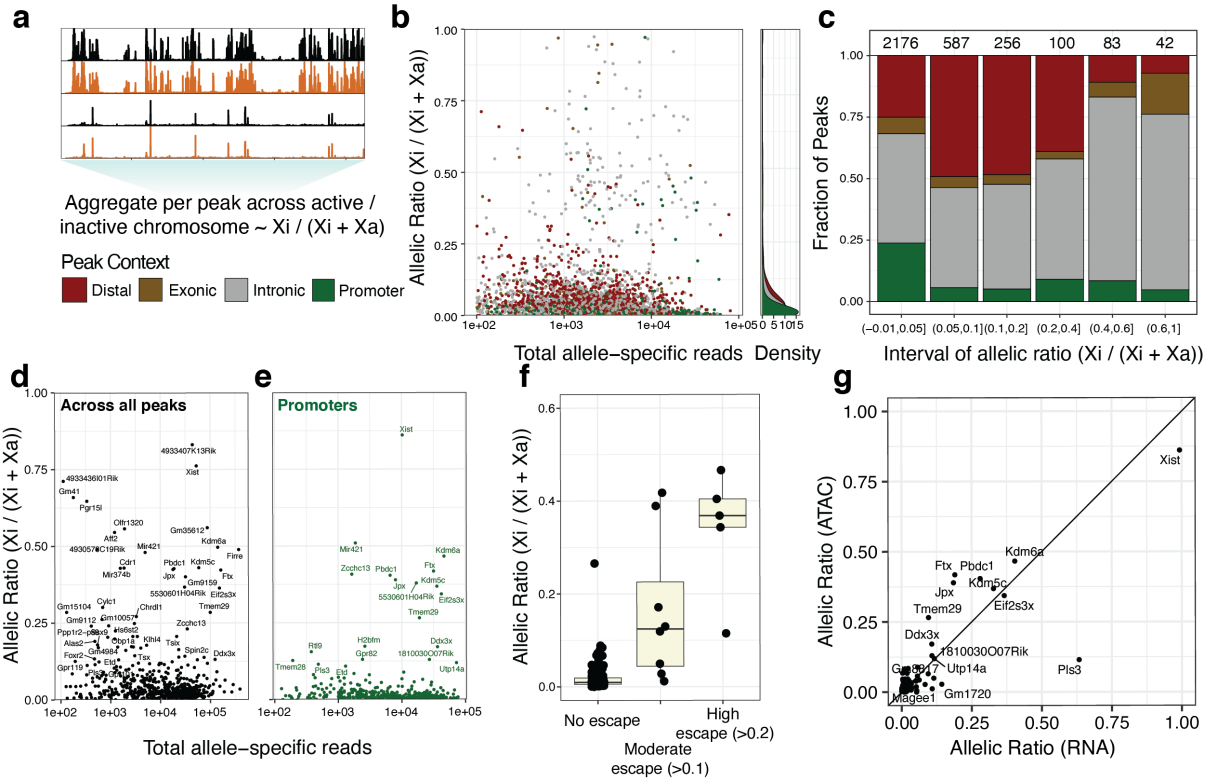

Figure 3.15: Chromatin accessibility of the Xi. (**a**) Overview of the analysis strategy. (**b**) Scatterplot showing the total accessibility, as derived from allele-specific counts, against the fractions of Xi-derived reads. (**c**) Barplots stratifying peaks by allelic ratio, and by annotated region. (**d**) Allelic ratios of accessibility measurements across genes. Each peak is annotated to the nearest gene, and allelic ratios are aggregated across them. (**e**) As **d**, showing promoter allelic ratios. (**f**) Boxplots comparing promoter allelic ratios measured from RNA, see **Secion 3.8**. (**g**) Scatterplot showing the same data as in (**f**), demonstrating a correlation between the two omics datasets.

Next, I asked whether chromatin accessibility varied across cell types. To this end, I employed scDALI with a cell type-based kernel and detected cell type-specific accessible peaks at an FDR <10%. This analysis revealed that intronic peaks were the most likely to be variable across cell types. This was followed by promoter-overlapping, distal, and last, exonic peaks **(Fig. 3.16a)**. Overall, this analysis demonstrates remarkable variability in the regulatory landscapes of adult immune cell types. As specific examples, I examined chromatin accessibility around the *Med14* and *Cxcr3* loci, which showed variable escape in expression. I observed that CD8+ memory T-cells, among other cell types, showed Xi-specific accessibility on the *Med14* promoter and the end of the gene body, whereas follicular B-cells and other cell types did not. Similarly,

Cxcr3 showed accessibility only in NK-T- and CD8+ memory T-cells.

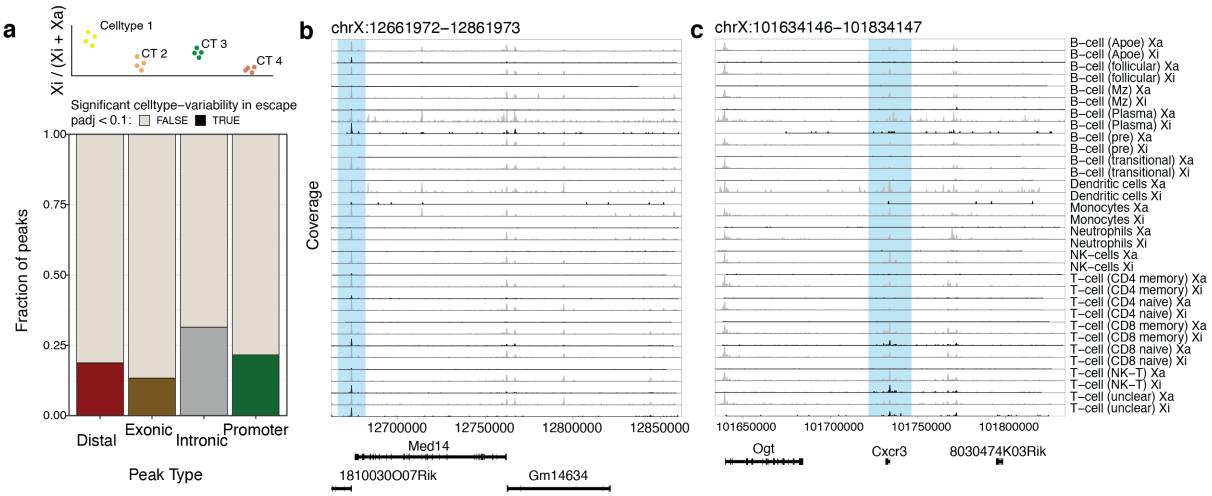

Figure 3.16: Cell type-specific chromatin accessibility of the Xi. (**a**) Outline of the scDALI-based analysis of cell type-specific accessibility ratios (top). On the bottom, a barplot shows the fraction of peaks with significant variability, stratified by genomic context. (**b**) Coverage plot across the Xa (grey) and Xi (black) across cell types, around the *Med14* locus. (**b**) Coverage plot across the Xa (grey) and Xi (black) across cell types, around the *Cxcr3* locus.

Finally, I sought to replicate the increase in global escape derived from scRNA-sequencing data, considering that a general loss of silencing integrity should lead to increased chromatin accessibility across the X. To this end, I calculated allelic read ratios $\frac{X_i}{X_i+X_a}$ across the Xi for each cell, and for each population of cells within a sample **(Fig. 3.17a)**. I observed that for the vast majority of cell types, average allelic ratios stayed the same of increased during ageing. Notably, CD8+ memory T-cells showed a stark increase, paralleling the results derived from scRNA-sequencing data **(Fig. 3.17b)**. This was largely consistent across analyzed samples. Of note, the scATAC-sequencing data also showed an increase in CD8+ naive T-cells, suggesting that the effect is indeed ageing-related and not only differentiation-dependent. Furthermore, I observed strikingly high allelic ratios in Plasma B-cells **(Fig. 3.17c)**.

I next asked whether this increase in accessibility corresponded to specific genomic features, and stratified the analysis by promoter, intronic, exonic and distal peaks. Globally, exonic and promoter peaks shpwed the lowest accessibility, relative to the Xa, and less variation across cell types **(Fig. 3.17d)**. However, CD8+ memory T-cells showed and increase also at the promoter level, which is most closely associated with gene expression. Of note, distal and intronic peaks showed striking variability in allelic ratios across cell types and the strongest increase in CD8+ memory T-cells. These results indicate that ageing-related increases in accessibility affect the entire genome, and include regulatory elements. I confirmed this by analyzing the allelic ratios between young and old CD8+ T-cells, pooling across naive and memory cells whose populations contract and expand during ageing respectively. This analysis revealed a set of peaks with increased allelic ratios in aged cells, including sites close to the *Tmsb4x* and *Elf4* genes **(Fig. 3.17e)**. The former has been shown to escape in different contexts, while the latter is a transcription factor relevant to T-cell biology [Mousavi et al., 2020, Yamada et al., 2009].

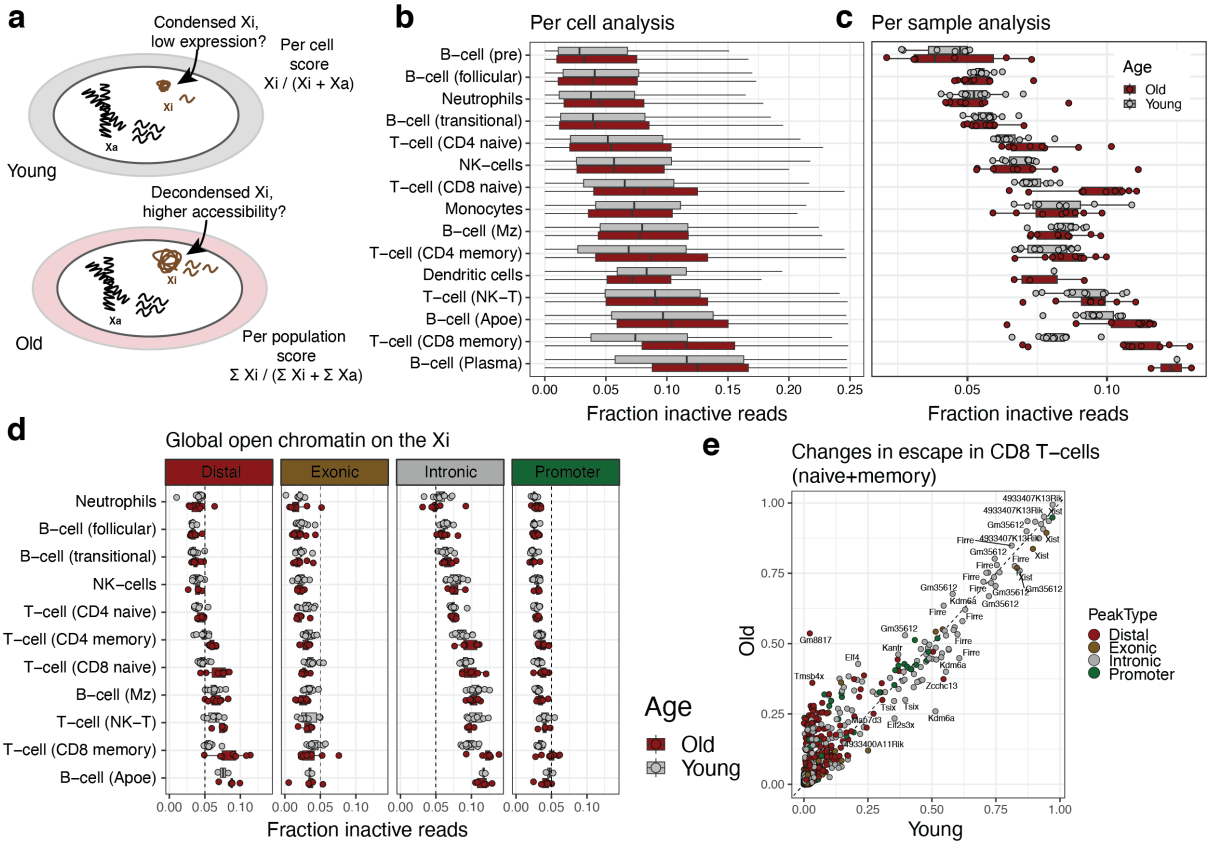

Figure 3.17: Global increase of accessibility on the Xi during ageing. (**a**) Scheme of the analysis approach to quantify global accessibility on the Xi. I both consider X-chromosome-wide ratios by summing reads across genes, and sample-cell type population bulks. (**b**) Boxplots showing distributions of Xi ratios across cell types and ages. Boxplots show medians, 25 and 75% and 1.5 times inter-quartile ranges (whiskers). (**c**) Boxplots showing aggregated counts per cell type and sample. (**d**) The same plot as (**c**), but stratified by peak category. (**e**) Scatterplot showing allelic ratios in young and old CD8+ T-cells (naive and memory).

3.6 Discussion

My work in this chapter dissects the cell type-specificity of escape from X-inactivation in an adult tissue. By employing single-cell methods, I show that it is possible to assign the X-inactivation status of individual cells. Thereby, one obtains a multi-modal dataset containing information on expression levels, including cell types and states, allele-specific expression, specifically escape from XCI and skewed XCI together. Using this data, I show that XCI is largely complete in all immune cell types, but that both quantitative variation in escape and cell type-specific escapees exist. Then, I move to organismal ageing, and test a long-standing hypothesis that the inactive X-chromosome loses its stability in aged cells. I show that global escape does indeed increase, but that this is a cell type-specific, not a universal feature of ageing. Specifically, this increase is most prominent in CD8+ T-cells and might be due to their clonal proliferation history. Finally, I show that chromatin accessibility can be analyzed using scATAC-Seq data in an equivalent way. These results show that expression and chromatin accessibility are largely correlated at escape loci, and confirm that ageing of CD8+ T-cells destabilizes the Xi.

How high the total fraction of genes escaping XCI is has been under some debate. The prevailing view has been that in mice and humans, around 3-7% and 20% of genes escape, largely due to the larger pseudo-autosomal regions in humans which retain expression on the Xi [Loda et al., 2022]. However, some studies suggested that in individual cell types escape might be widespread, affecting more than 50% of genes [Sierra et al., 2023]. My results show that escape is largely stable across different cell types, and applying the approach presented here to other tissues has largely confirmed this (preliminary results from James Cleland). Reassuringly, some of the genes I observe to escape have been suggested to do so before and are likely to impact cell function [Oghumu et al., 2019, Richart et al., 2022]. This approach can be used to generate tissue-level escape maps, and fully define the extent of expression from the Xi.

The most important contribution of this chapter is the observation that global escape increases in some cell types, but that the Xi is largely stable during ageing. It is known that *Xist* is dispensable for maintenance of XCI, which is instead maintained by repressive regulatory machinery and epigenetic erosion is thought to be a hallmark of ageing [López-Otín et al., 2013]. This has been shown for different epigenetic marks as well as chromatin acessibility [Bozukova et al., 2022, Tauc et al., 2021, Benayoun et al., 2019]. It will therefore be interesting to investigate whether repressive marks are affected in parallel to increased escape. In particular, it is noteworthy that CD8+ memory T-cells show the strongest clonal expansion [Sun et al., 2022]. It is therefore tempting to speculate that excessive proliferation drives loss of escape, possibly through telomere shortening [Schoeftner et al., 2009, Goronzy and Weyand, 2019]. This parallels the erosion of XCI in induced pluripotent stem cells *in vitro* [Cloutier et al., 2022].

Furthermore, plasticity of escape during clonal expansion might explain recurrent increase of *Med14*, which has been shown to affect proliferation [Richart et al., 2022]. Increases in expression of these genes might be selected for by proliferative pressure, and indeed provide female-specific phenotypic variability. These results need to be confirmed in human samples, where large single-cell datasets of immune cells across ages are available [Yazar et al., 2022]. Furthermore, these results suggest specific functional experiments: If increased escape is due to hyper-proliferation, it should be possible to induce this change by long-term culture of T-cells

in vitro. Furthermore, it should be validated that *Med14* has an impact on T-cell proliferation or function, and that aged female cells show proliferative advantages over male cells. A largely unaddressed topic is what regulates escapees, and whether cell-type specific gene regulation on the Xi is different to the Xa. My single-cell approach can use correlative analysis to address this question in two ways: On the one hand, the integrative total expression measurements with escape quantifications can nominate transcription factors which correlate with mRNA levels from the Xi. This type of analysis has been used to derive gene regulatory networks [Aibar et al., 2017]. However, it can be challenging to distinguish genuine causal links from joint correlation with an upstream cell state variable, which is correlated to both.
Secondly, chromatin accessibility can be used in a similar manner to identify links between *cis*-regulatory elements and target genes [Pliner et al., 2018]. The dataset I preliminarily analysed will be useful to identify CREs associated with escapee activity, since the Xi is known to show a strikingly different accessibility landscape than the Xa [Giorgetti et al., 2016]. In the future, multi-modal measurements of RNA, ATAC and others will be useful to refine these links.

A recent study has used a similar approach as presented here to assess escape in human immune cells [Tomofuji et al., 2023]. This preprint proposes an approach to partially phase the X-chromosome based on expression data alone, which is particularly useful when no genomic phasing information is present. However, the number of genes and loci of accessible chromatin assessed in that study is limited, partly because of the lack of phased genome information and the rarity of heterozygous variants. Furthermore, that study did not use statistical models of allele-specific read counts to account for sparsity.

Although single-cell data in principle provides the ability to quantify escape at the single-cell level, neither I nor [Tomofuji et al., 2023] use the data to quantify inter-cell variability beyond comparing cell types. This is due to the inherent sparsity of allele-specific data. It is also an open question whether other methods such as Smart-Seq2 which suffer from PCR amplification bias due to the lack of UMIs will improve these quantifications, as escapees are generally not highly expressed. Given enough signal, it will be interesting to model overdispersion of escape directly, and to assess whether the incomplete escape I observed is due to lower expression in all cells or lack of escape in some cells. A specific model for a given gene could be similar to the BASICS model for total expression variability:

$$k_i \sim BetaBin(n_i, p, log(\theta_t) + log(\theta_0) + log(\theta_x)) \tag{3.11}$$

where θ_t captures technically induced overdispersion which can be estimated using spike-ins, θ_0 is the baseline biological overdispersion and θ_x captures condition or cell type-specific overdispersion. When read counts are low, mean and variance are indistinguishable, as the read counts are close to Bernoulli-distributed. However, by sharing information on θ across genes or increasing read counts, this model could quantify the extent to which escape is technically driven or varies across cells and conditions.

Chapter 4

Xist modulates the expression of escapees

Overview: Little is known about how escape from X-inactivation is regulated. In humans and mice, the long non-coding RNA *Xist* is the causative factor driving X-inactivation during embryogenesis, but appears to be largely dispensable for long-term silencing. However, deletion of *Xist* leads to moderate de-repression of the Xi, and its expression levels and localization have been suggested to impact escape. Here, I systematically test whether *Xist* can silence escapees in a post-silencing context. In neural progenitor cells, *Xist* overexpression leads to a time-dependent loss of escape of almost all genes. Genes show variable resistance to silencing, partly dependent on their chromosomal position. As during X-inactivation, silencing is dependent on the *Xist* co-factor *SPEN*. When *Xist*-levels are reduced back to normal, genes show partial reversibility, again depending on how long they had been silenced. Finally, increased *Xist*-levels reduce escape during imprinted XCI. These results demonstrate a potential mechanism by which physiological or pathological variation in *Xist* activity might impact escape from XCI.

Contributions: This study represents joint work between the Heard, Stegle and Odom labs. The project was conceived by Edith Heard, Antonia Hauth and Agnese Loda. Antonia Hauth and Agnese Loda performed all experiments. The embryo data was generated by Agnese Loda and Emma Kneuss. Sequencing data was pre-processed by Antonia Hauth and me, using a workflow designed by Yuvia Perez. The escape meta-analysis was conducted by Yuvia Perez and extended by me. I performed all other computational analysis. The project was supervised by Edith Heard, Oliver Stegle and Duncan Odom. A paper including all of the results presented in this chapter is currently in preparation.

4.1 Introduction

In the previous chapter, I have performed a comprehensive mapping of escape in adult cell types, but that study did not address the gene regulatory mechanisms acting on escapees. I have shown, as others previously, that expression from the Xi is associated with a retention of accessible chromatin at their gene bodies and promoters [Berletch et al., 2015]. Furthermore, it has been demonstrated that they show marks of an active chromatin state, and a depletion of DNA methylation [Balaton et al., 2015]. In contrast, distal regulatory elements that

are critical for context-specific gene expression on the Xa, are rarely accessible on the Xi [Giorgetti et al., 2016]. Additionally, there is a clear connection to chromosome conformation, with intact TAD structures at escapees [Giorgetti et al., 2016]. This is driven by binding of CTCF, suggesting that local compartmentalization is critical to resist full condensation of the Xi [Berletch et al., 2015, Fang et al., 2023]. These results suggest that a partially active chromatin environment facilitates Xi-specific expression. Apart from CTCF, there is little evidence of *trans*-acting factors driving escape, although escapee promoters are enriched for motifs of the transcription factor YY1 [Peeters et al., 2023].

As the term ”escape” from XCI implies, these genes resist silencing during the inactivation process, as opposed to them being re-activated later during development or adult life. It has been suggested that by the mechanics of *Xist* spreading across the Xi, some genes are protected from silencing [Engreitz et al., 2013]. Furthermore, repetitive DNA, especially LINE1 elements are associated with *Xist* progression and are depleted from escapees [Lyon, 1961]]. Therefore, sequence features contribute to the intrinsic ability of a subset of genes to resist silencing, which has in particular been shown for constitutive escapees which escape in all tissue contexts [Barros de Andrade E Sousa et al., 2019, Pacini et al., 2021]. However, it remains open whether these parameters play a role in their propensity to become re-activated as well. In particular, facultative escape which is only observed in specific tissues or other contexts, might result from secondary loss of silencing later in life.

Whether *Xist* RNA can impact escape remains an open question. The prevailing model is that during XCI, it needs to be expressed for a sufficient period of time to make XCI irreversible [Wutz et al., 2002]. Afterwards, silencing is driven through secondary actors, and is thought to be *Xist*-independent. In accordance with this, its experimental deletion does not lead to full re-activation of the Xi. It does however lead to mild increase in expression, and has hematological phenotypes [Yildirim et al., 2013, Yang et al., 2020, Yu et al., 2021, Yang et al., 2022b]. This suggests that the continued presence of *Xist* on the Xi contributes to silencing, even though it is not strictly required. A similar phenomenon was observed in breast cancer, where loss of Xist-mediated silencing has been shown to contribute to aggressiveness [Richart et al., 2022].

Here, I use expression data generated in a reduced model system of escape to directly and quantitatively test the impact of *Xist* on escape from XCI. Using an inducible *Xist*-transgene in neural progenitor cell lines, I show that increased *Xist* mRNA levels silence escapees. Silencing progresses as long as *Xist* is overexpressed and leads to almost full, chromosome-wide loss of escape. However, dynamic modelling shows that constitutive escapees show the strongest resistance to silencing. I further show that as during XCI, *Xist* action depends on the presence of its co-factor *SPEN*. The partial silencing induced by short-term *Xist*-overexpression is largely reversible when *Xist*-levels are reduced again. Meanwhile, long-term overexpression leads to irreversible inactivation of many genes, especially those strongly susceptible to *Xist*-mediated silencing. Finally, I show that increased *Xist*-levels reduce escape in pre-implantation embryos, validating these findings *in vivo*. Collectively, these results show that *Xist* has the capacity to re-initiate X-inactivation in adult cells and especially facultative escapees can be inactivated fully. Furthermore, I use this process to quantify a genes intrinsic resistance to inactivation in adult cells, which might predict how likely a gene is to re-activate.

4.2 *Xist* overexpression silences genes that escape X-inactivation

4.2.1 Measuring escape from XCI using clonal NPC lines

To assess the role of *Xist* in the silencing of escapees, I used data generated in a cell line system that allows to model escape. A commonly used model are neural progenitor cells generated from mouse embryonic stem cells (mESCs) derived from an F1 hybrid cross. In naive mESCs, both X-chromosomes are active. By changing medium conditions mESCs are grown in to remove self-renewal cues, they undergo spontaneous differentiation, mainly committing to the neural lineage [Ying et al., 2008]. During this process, X-inactivation proceeds, which has been used as an *in vitro* model for XCI in previous studies [Wutz et al., 2002]. As in the embryo, XCI is random, but by clonal expansion of individual NPCs, cell lines with fully biased XCI can be obtained **(Fig. 4.1a)**. The NPC lines described in this chapter were generated from a C57B6 x CAST/EiJ ESC line that carries a dox-inducible transgene on the B6 X-chromosome (Clone E6) (see next section) [Dossin et al., 2020]. I also make use of transgenic cell lines that carry an inducible degron for SPEN (Clones CL30/CL31, and subcloned cell lines CL31.16/CL30.7) (see **Section 4.3**). In this study the RNA-Sequencing data is has full transcript length, which facilitates the quantification of allele-specific expression. I first validated that for the vast majority of expressed genes, total read counts were strongly correlated with allelic read counts, suggesting that only few genes do not permit quantification of escape **(Fig. 4.1b)**. In total, I can assess escape in 442 genes out of 573 expressed X-linked genes (>10 reads per sample, 77.1%). Next, I confirmed that autosomes show expected bi-allelic expression, and that the majority of X-linked genes show mono-allelic expression from the Xa, which is the CAST haplotype in all cell lines used here. To this end, I calculated the ratio between allelic read counts

$$\frac{X_a}{X_a + X_i} = \frac{X_{CAST}}{X_{CAST} + X_{B6}} \tag{4.1}$$

which revealed strongly Xa (CAST)-biased expression for most X-linked genes in all cell lines **(Fig. 4.1c)**.

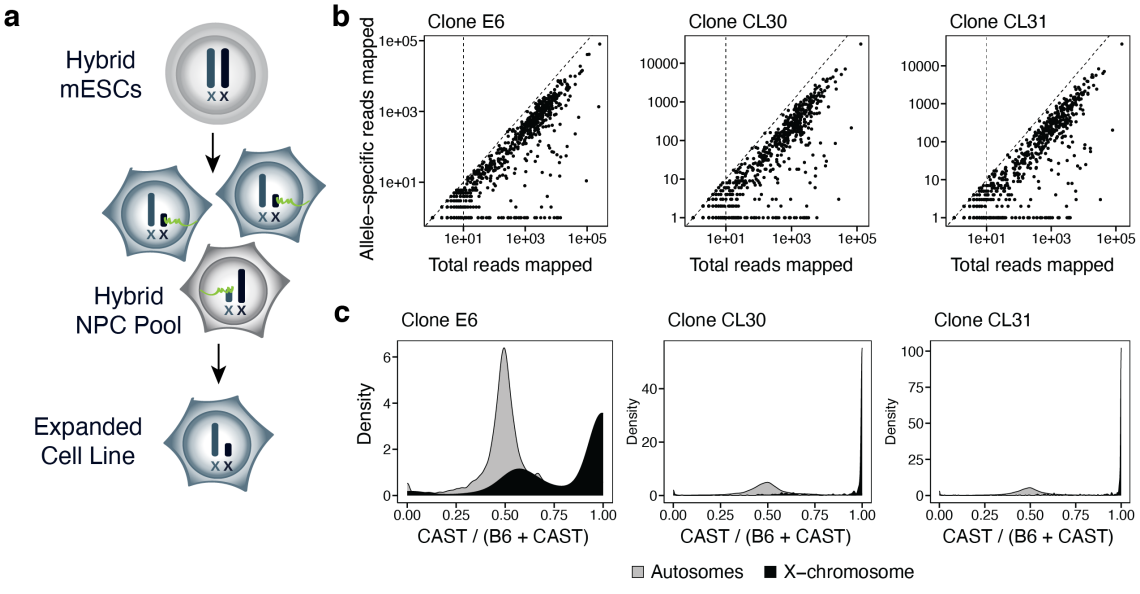

Figure 4.1: (Caption on the next page.)

Figure 4.1: Allele-specific quantifications of escape in NPC lines (**a**) Scheme demonstrating the derivation of neural progenitor cell lines used in this study, modified from graphics generated by Agnese Loda. (**b**) Scatterplot showing total (x-axis) against allele-specific (y-axis) read counts of X-linked genes in the three analyzed cell lines, showing one replicate each. Only a small number of genes shows no allele-specific discrimination and can therefore not be analyzed in this study. (**c**) Density plot showing distributions of allelic ratio for autosmal (grey) and X-linked genes in the indicated cell lines. Note that in each cell line, the majority of genes is only expressed from the active (CAST) haplotype.

4.2.2 Widespread escape from XCI in neural progenitor cells

I found that bi-allelic expression of X-linked genes, indicating escape from XCI, was common in the E6 cell line, but scarcer in CL30 and CL31 **(Fig. 4.2a)**. Escapees were highly similar across experimental replicates and most genes that I found in the low escape clones CL30/CL31 were also found in E6. As expected, *Xist* was fully biased to the Xi **(Fig. 4.2a)**. As *Xist* is the only gene expected to be only expressed from the Xi, I excluded a small number of genes with an allelic ratio > 0.8, as these likely represent mapping artefacts. Furthermore, I observed that escapees tended to appear in clusters along the length of the X-chromosome, in accordance with previous work that suggests regions of the Xi to be conducive for expression [Giorgetti et al., 2016].

Previous work has identified many escapees in different mouse tissues, with a subset of them being found in most datasets analysed [Fang et al., 2023]. These ”constitutive” escapees might have different molecular features than sporadic ”facultative” escapees. I therefore extended a previously conducted survey of studies profiling escapees to classify genes as constitutive and facultative, or usually silenced. Across 19 studies, the escape status of 863 genes was assessed in at least one dataset **(Fig. 4.2b)**. I then classified genes that were observed to escape in more than three and at least half the studies as constitutive and genes that never escaped as silenced. All other genes I termed as facultative. Previously, escape has been defined as the fraction of Xi-derived reads being greater than 10% [Balaton et al., 2021]. In the NPC dataset, 13-14 out of the 14 annotated constitutive escapees escaped, with 12 of them being observed in all cell lines, showing that also in NPCs, constitutive escapees are generally expressed from the Xi **(Fig. 4.2c, d)**. In contrast, 30 to 95 facultative escapees were escaping across samples, with 23 being shared among cell lines, and an excess of genes in the E6 cell line. Of note, 17 - 46 silenced genes also escaped (10 shared), which might represent NPC-specific escape, or might have been missed due to lack of power in previous studies. I therefore consider these genes ”NPC-specific” escapees for the rest of this chapter.

In total, the presented analysis demonstrates widespread escape that varies across NPC lines, and is consistent with previous surveys of escape. Also, the dataset shows that the E6 cell line shows extensive escape and the majority of expressed genes has sufficient allele-specific coverage, making it an ideal model system to study regulation of escape genome-wide.

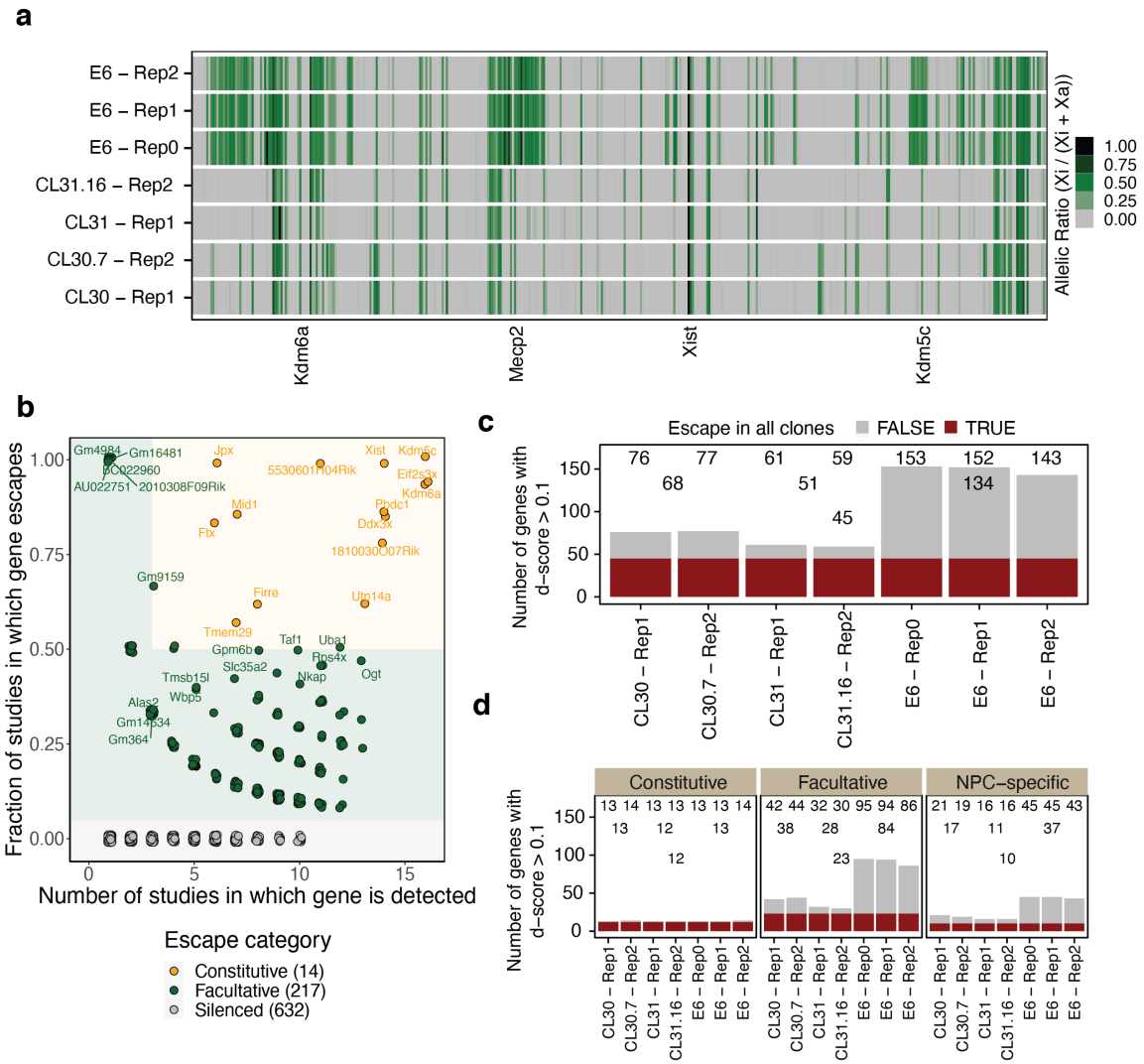

Figure 4.2: Measuring escape in neural progenitor cell lines. (**a**) Heatmap showing allelic ratios for X-linked genes, ordered by their position on the X-chromosome. Individual replicates are shown for the E6 cell line, and parental and subcloned cell lines for C30/C31. (**b**) Scatterplot showing the results of a meta-analysis of escape across RNA-Sequencing studies. For each X-linked gene, the number of datasets in which its escape was assessed is plotted against the fraction in which it was reported an escapee. Genes which are never considered escapees are termed "silenced" (grey). Genes which escape more than half the time and at least in three studies are termed "constitutive" escapees (orange) all other genes are termed "facultative". (**c**) Number of genes escaping in cell lines, and common escapees in different combinations. (**d**) As (**c**), but stratified by escape category.

4.2.3 Time-dependent silencing of escape by *Xist*-overexpression

I next asked whether increased *Xist*-levels would lead to reduction of escape across the X-chromosome. All used cell lines contain an inducible Tet-On *Xist*-transgene that allows for controlled over-expression by addition of doxycycline (Dox) to the culture medium **(Fig. 4.3a)**. The transgene is located on the inactive B6 X-chromosome and *Xist* is known to act in *cis*, the overexpression is therfore specific to the Xi, and is not expected to affect the Xa or autosomes. Across three replicate experiments performed in the E6 cell line, I found that Dox addition over 3, 7, 14 and 21 days lead to a robust and stable average 7-fold increase of *Xist*-mRNA levels **(Fig. 4.3b)**. I next assessed reduction of escape as measured by the allelic ratio of 134

escapees in E6, excluding *Xist*. Addition of Dox induced a clear and time-dependent reduction of escape across the X-chromosome **(Fig. 4.3c)**. Stratifying between escape categories shows that the vast majority of genes showed a reduction in allelic ratios **(Fig. 4.3d)**. In particular, NPC-specific and facultative escapees are almost fully silenced, while constitutive escapees are strongly reduced, but retain more expression from the Xi. I next asked to which extent genes were silenced at the different time points. To this end I quantified the extent of silencing as the fraction of retained escape after treatment, and also assessed the fraction of genes that retained more than 10% of expression from the Xi. I also used a binomial linear model to assess significant reductions in allelic ratios derived from active and inactive read counts k_{Xa}, k_{Xi} by modelling

$$k_{Xi} \sim Bin(k_{Xa} + k_{Xi}, p)$$
$$logit(p) = u$$
$$u = \beta_0 + \beta_{Dox}$$

and estimating the significance of β_{Dox}. This analysis shows progressively increasing numbers of genes that decrease significantly in escape, and genes that fall below the fraction threshold of 10%. Importantly, most constitutive escapees saw a reduction in escape, even though half of the genes still retained > 10% expression from the Xi. A majority of NPC-specific and facultative escapees were already largely silenced after seven days of Dox overexpression **(Fig. 4.3e)**.

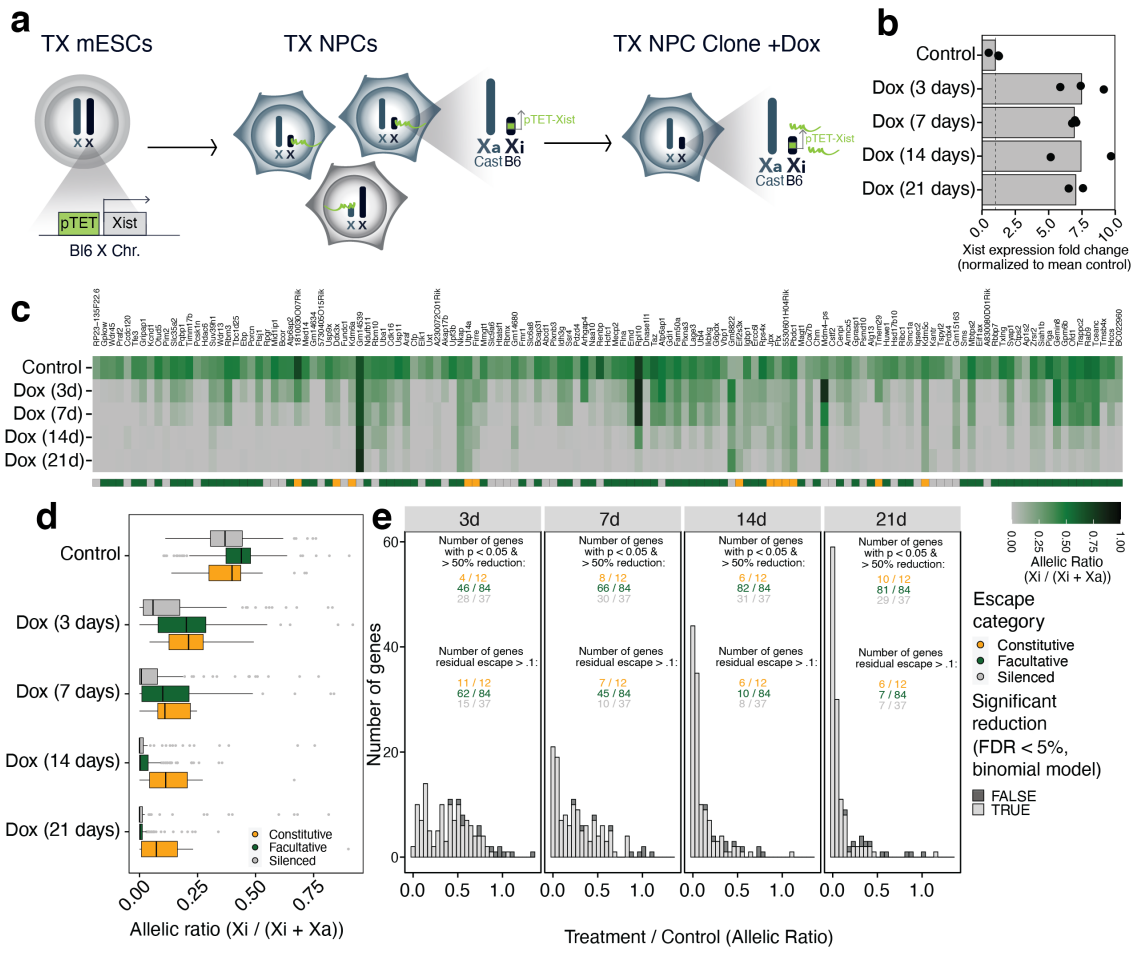

Figure 4.3: (Caption on the next page.)

Figure 4.3: Measuring escape in neural progenitor cell lines. (**a**) Scheme demonstrating the derivation of neural progenitor cell lines and *Xist* overexpression, modified from graphics generated by Agnese Loda. (**b**) Barplot showing the increase of *Xist* RNA in the different timepoints. Each points represents on replicate experiment (n = 3). (**c**) Heatmap showing loss of escape during Xist overexpression over time. The heatmap only shows detected escapee genes, and genes are ordered by their position on the X-chromosome. (**d**) Boxplots shpwing the distribution of allelic ratios in (**c**) across escape categories. (**e**) Histogram shpwing the distribution of allelic ratios after Xist-overexpression, normalized to the untreated control. White fill shows the number of genes with significant reduction in a generalized linear model (FDR < 5%).

These results demonstrate clearly that an increase in *Xist* levels leads to almost complete silencing of the Xi in the E6 cell line. I next assessed whether this phenotype was reproducible in independent cell lines. In an equivalent experiment, seven days of Dox exposure lead to *Xist* levels being increased by an average of 11-fold in CL30 and CL31 (two experimental replicates). As in E6, there was a clear reduction of allelic ratios after 3 and 7 days, with NPC-specific genes being most affected **(Fig. 4.4a, b)**. These results demonstrate that *Xist* mRNA levels reduce escape in a chromosome-wide manner. Although almost all genes are efficiently silenced, constitutive escapees seem to be most resistant to inactivation.

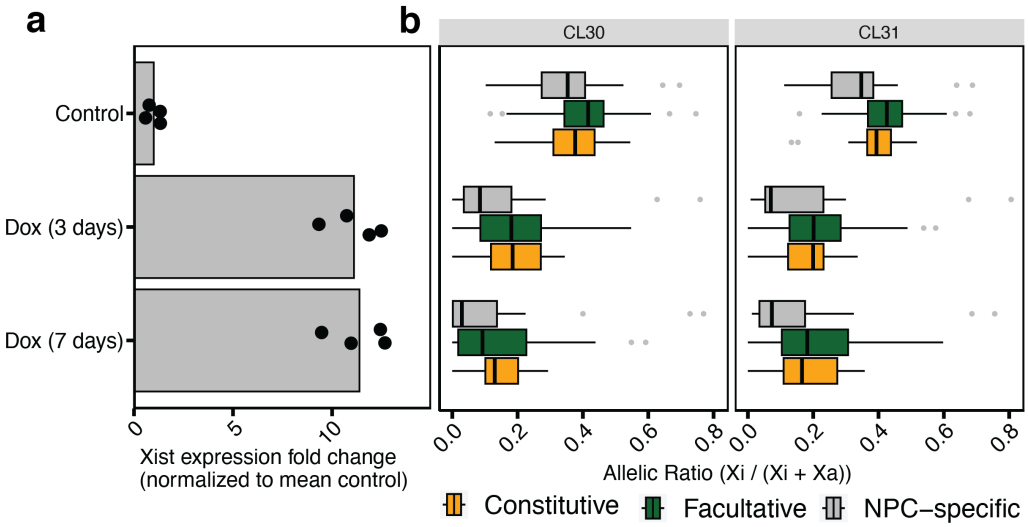

Figure 4.4: Validating the silencing phenotype in independent cell lines. (**a**) Barplots showing Xist overexpression in CL30 and CL31. Expression levels are normalized to the median of all control samples. (**b**) Boxplots showing loss of escape in CL30 and CL31. Genes are stratified by their classification in the escape meta-analysis.

4.2.4 Modelling the kinetics of escapee silencing

The results presented in the previous section suggest that while all X-linked genes can be silenced by *Xist*, there is variation in how quickly expression is lost and whether some residual escape remains, particularly for constitutive escapees. To understand this variation quantitatively, I devised exponential decay models to describe the observed silencing data. As the read counts derived from bulk RNA-Seq are reasonably high, I approximated the allelic rates as following a normal distribution, instead of modelling the read counts directly using a binomial distribution. For each gene, I compute allelic ratios $r = \frac{k_{Xi}}{k_{Xi}+k_{Xa}}$ and assume that r is distributed

as

$$r \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu = e^{-kt} A_0 + b_0$$

where t is the timepoint of measurement in days. This model assumes that the allelic ratios start at A_0 to account for variable basal escape levels and then reduce with a decay constant k, which captures the speed of silencing. The half-time of reduction, until $r = \frac{1}{2}A_0$ can be computed by $t_{1/2} = \frac{k}{ln(2)}$. Finally, b_0 accounts for the possibility that escape is not reduced to zero, but that the exponential decay approaches a non-zero value over time. I refer to this value as residual escape **(Fig. 4.5a)**. By comparing the full model to a reduced model without b_0, I can assess whether genes get fully silenced or not.

I estimated model parameters using non-linear regression (*nls* function, *stats* R package) and assessed the validity of function fits by the R^2 value, which showed that 136 genes were fit well by regression including or excluding b_0. At the level of specific genes, there were fully and quickly silenced genes such as *Mecp2*, slowly but fully affected genes such as *Gpm6b* and varying levels of residual escape, for example at *Kdm5c* and *Pbdc1* **(Fig. 4.5b)**. As the dataset encompasses only a limited set of timepoints, it is not necessarily clear whether genes are fully silenced or silencing is very slow. I therefore assessed whether there was statistical evidence for residual escape. To this end, I compared two model fits with and without b_0 and computed the Bayesian Information Criterion to account for the additional parameter and found that 33 out of 136 genes had a higher BIC in the offset model **(Fig. 4.5c)**. Across genes, I found that silencing speeds, measured by escape half-life, varied broadly across genes, and that constitutive escapees showed slower silencing than facultative ones which were themselves slower than NPC-specific ones **(Fig. 4.5d)**. Similarly, residual escape was strongest in constitutive escapees than the other classes **(Fig. 4.5e)**. However, after accounting for escape category, there was still large variation in all categories, suggesting that other features impact the silencing behaviours of genes.

Indeed, visual inspection of **(Fig. 4.3d)** suggests that genes in close proximity to each other were silenced at similar rates. To quantify this effect I defined clusters of escapees in close proximity. Traversing the X-chromosome, I split genes into groups that were within 100kb of each other, and split groups if they were further apart. Having defined 11 clusters of escapees with greater than two genes, I then computed the average absolute difference in silencing half times between genes within the same cluster, and random gene pairs **(Fig. 4.5f)**. Genes in close proximity showed significantly more similar silencing dynamics than random background, suggesting that the local chromosome environment affects escape.

These results demonstrate differential resistance of genes to silencing. This variability is partly explained by gene-intrinsic factors, associated with constitutive and facultative categories, but also the local context on the X-chromosome. This is in line with previous results demonstrating that escape is organized in topologically associating domains [Giorgetti et al., 2016].

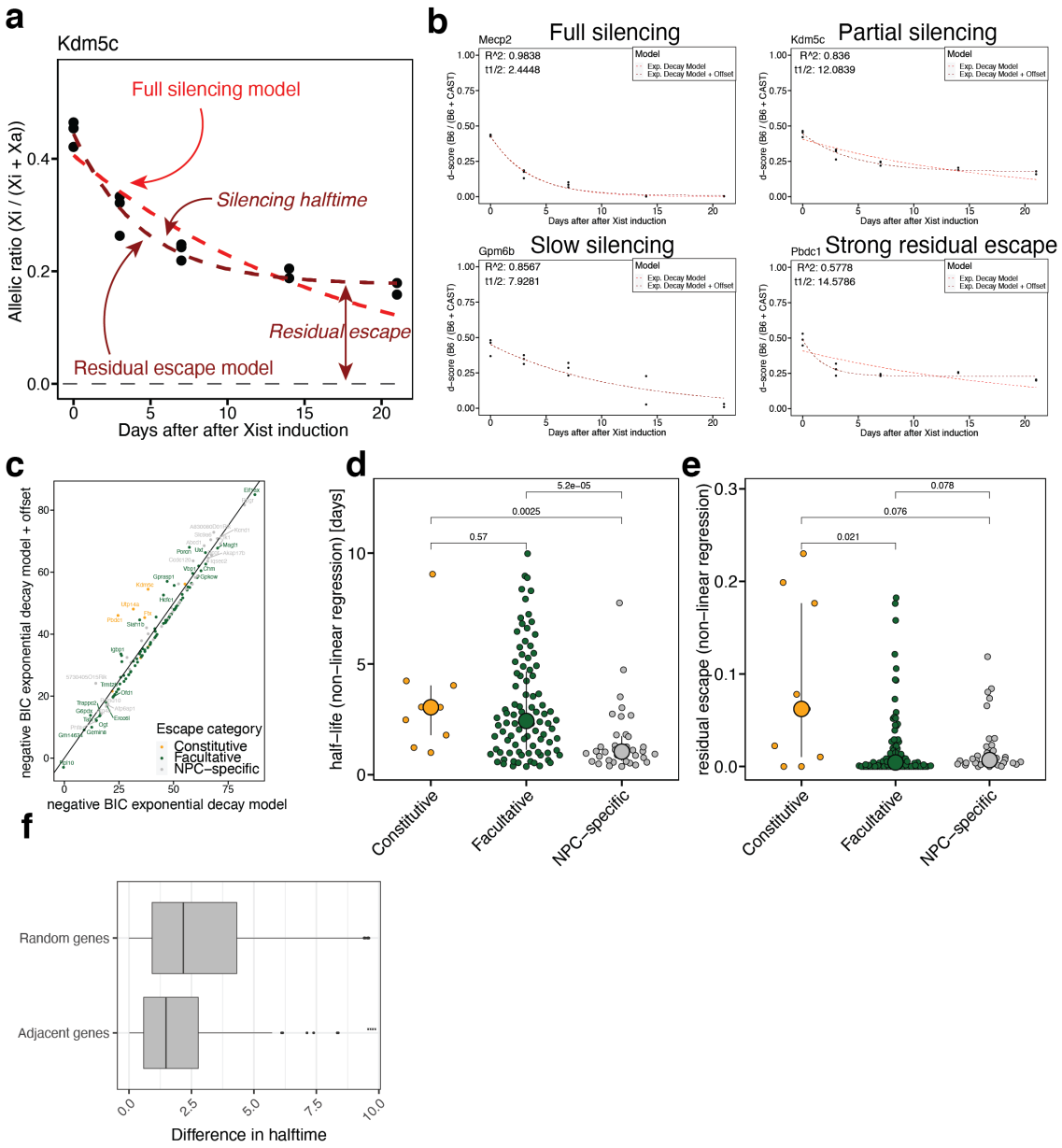

Figure 4.5: Modelling the silencing dynamics of escapees. (**a**) Scheme demonstrating the modelling approach. Exponential decay functions with and without a residual parameter are fit to the escape measurements of each gnee (**b**) Examples of decay curves for genes with varying dynamics, showing a fully silenced gene (*Mecp2*), a partially silenced gene (*Kdm5c*), a slowly silenced gene (*Gpm6c*) and one with high residual escape (*Pbdc1*). (**c**) Significance analysis of residual escape models. A scatterplot shows the Bayesian Information Criterion for the normal exponential decay model and one with residual escape. (**d**) Boxplot showing the distribution of estimated half-lifes and residual escape (**e**), stratified by escape categories. (**f**) Boxplots showing distributions of half-time similarities between adjacent and random gene pairs.

4.2.5 Effects of escapee-silencing on total expression

While the previous results show that escape is strongly reduced by *Xist*-overexpression, it does not directly show whether combined gene expression levels from both X-chromosomes is affected. I therefore next used DESeq2 to assess differential expression separately for the Xi and Xa. As expected, the Xi showed significantly reduced expression of almost all X-linked

genes after 7 and 21 days (**Fig. 4.6a** and **Fig. 4.6b** respectively). In contrast, almost no genes on the Xa were affected, showing that X-linked expression is not changed in *trans*, and that there is no compensation from the Xa. Across both X-chromosomes, a subset of the genes showed a significant reduction in expression levels, demonstrating that increasing *Xist* abundance can reduce X-linked expression levels.

Next, I assessed treatment effects on autosomal chromosomes. There have been mixed reports of whether *Xist* will expand from the Xi and silence genes in *trans* [Markaki et al., 2021, Jachowicz et al., 2022]. Furthermore, reducing the expression levels of escapees might have secondary effects on autosomes. Using DESeq2, I found only few differentially expressed genes, suggesting that neither of these effects are strong **(Fig. 4.6c)**. However, it is necessary to acknowledge that these might be weak effects that will only be visible using targeted analysis. As expression from the Xi is either less or equal to the Xa, even a full loss of escape will only decrease gene expression levels by at most 50%, and secondary effects from this difference might be difficult to detect. These results collectively show that *Xist*-mediated silencing is Xi-specific and reduces escapee expression levels, but has only a weak effect on gene expression globally. This validates the NPC system as suitable to study loss of escape, as secondary effects from silencing are limited.

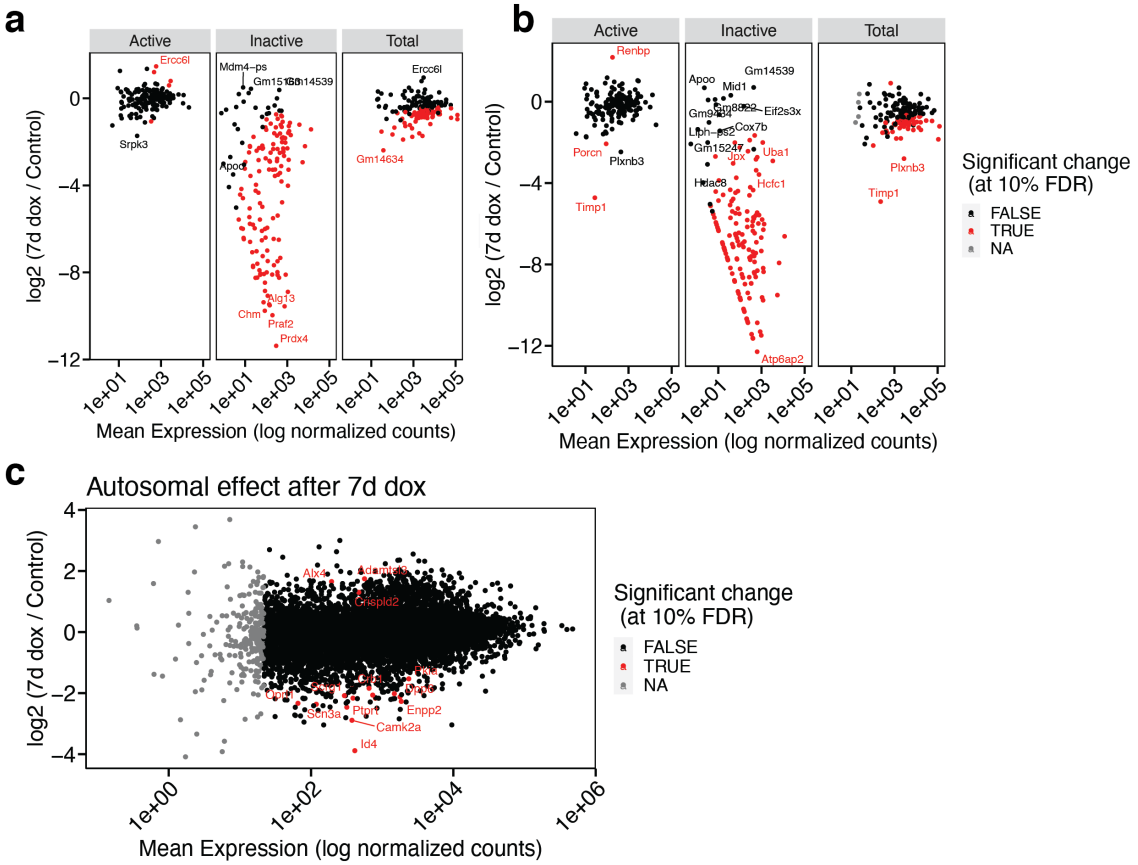

Figure 4.6: Effect of escapee silencing on gene expression. (**a**) DESeq2 analysis for differential expression of X-linked genes after seven days of dox treatment. The left plot shows counts from the active X, the middle from the inactive X and the right combining both together. Genes with a significant effect at 10% FDR are highlighted in red. (**b**) as (**a**), but showing tests after 21 days of dox treatment. (**c**) DESeq2 test of total read counts on autosomal genes.

4.3 Escapee silencing depends on the silencing co-factor *SPEN*

The mechanisms downstream of *Xist* during random XCI in the post-implantation embryo are relatively well understood. After establishment of *Xist*-expression on one chromosome, the lncRNA spreads across the X and recruits various chromatin silencing factors. These include the two polycomb repressive complexes (PRC1, through hnRNPK, and PRC2), methyltransferase proteins, including *Dnmt3b*, and histone de-acetylase complexes. A key factor that binds to *Xist* is the protein *SPEN*, which is thought to mediate silencing through interaction with Hdacs and the NuRD complex [Jachowicz et al., 2022, Loda et al., 2022].

I next addressed whether the inactivation mechanism during X-inactivation was the same as during silencing of escapees in differentiated cells. To this end, I made use of data generated from the CL30.7 and CL31.16 cell lines. Besides the Dox-inducible *Xist*-transgene, these contain SPEN alleles with an auxine-inducible degradation tag [Dossin et al., 2020]. Adding the molecule auxin to the culture medium will lead to recruitment of SPEN to the proteasome, and thereby to rapid degradation of the protein [Yesbolatova et al., 2020] **(Fig. 4.7a)**. For the combined treatment, auxin was added two days before Dox-induction to ensure full SPEN-depletion ahead of silencing. In both 3 and 7 day timepoints, I found that there was no reduction of allelic ratios in the absence of SPEN, compared to the untreated control, while allelic ratios were strongly reduced in the treatment with Dox only **Fig. (4.7b)**. This result was consistent between the two assayed cell lines **Fig. (4.7c)**. I note that this analysis only contains the sub-cloned cell lines presented in **Section 4.2.2**, as the parental lines were not genotypically pure (not shown).

I next visualized the distribution of allelic ratios across the different conditions, again demonstrating that SPEN-depletion abolished *Xist*-mediated silencing almost fully and consistently across cell lines. This was furthermore consistent for all escape categories, demonstrating that also in differentiated cells *Xist* requires *SPEN* as a co-factor **(Fig. 4.7d, e)**. These results show that escapee-silencing through *Xist* requires at least some of the same co-factors and suggests strongly that it follows the same mechanism as gene-silencing during random XCI.

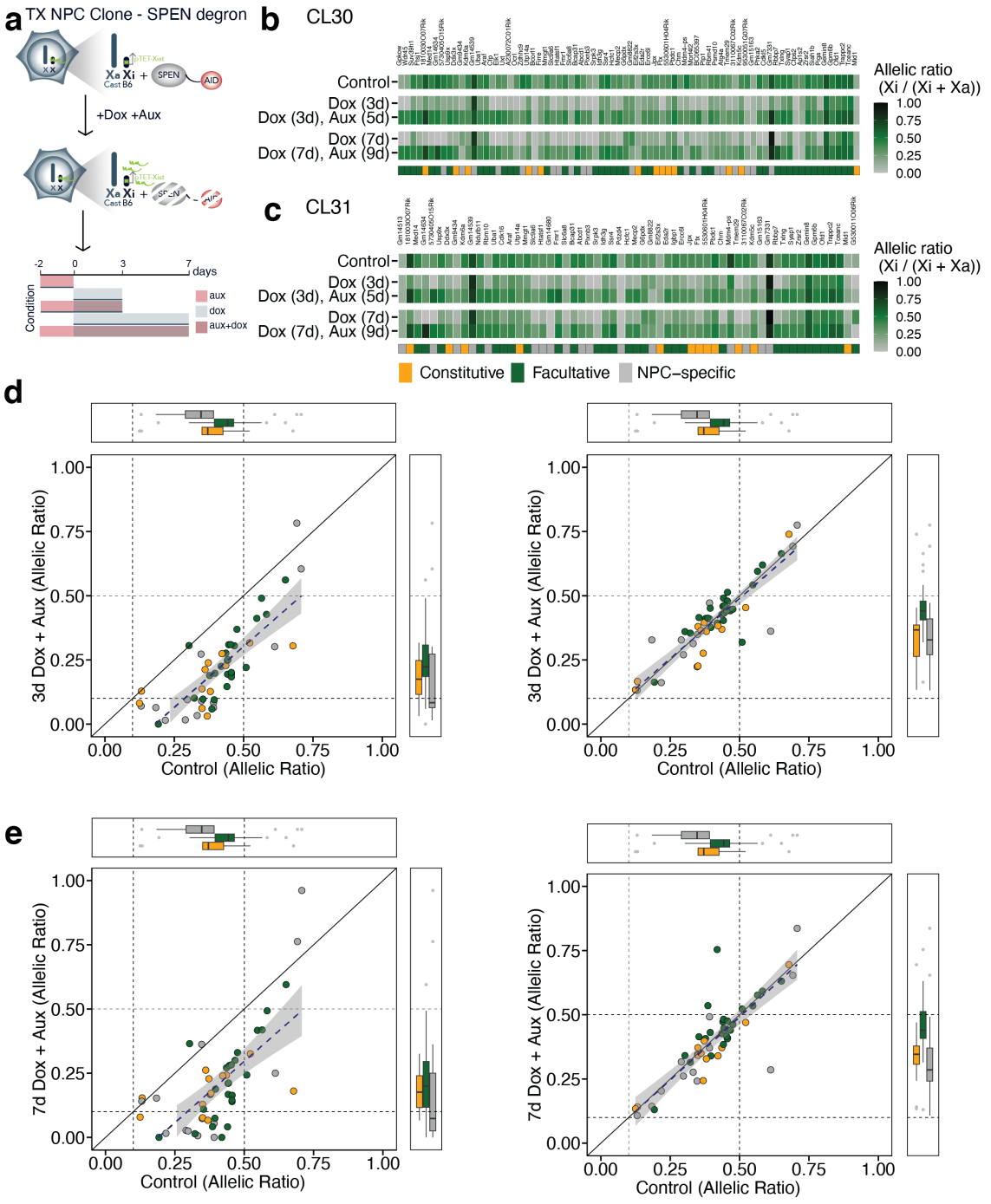

Figure 4.7: Escapee silencing depends on the Xist co-factor SPEN. (**a**) Scheme demonstrating *Xist* overexpression and SPEN depletion, modified from graphics generated by Agnese Loda. (**b**) Heatmap showing allelic ratios of escapees in CL30. Shown are untreated control, Xist overexpression (Dox) and additional depletion of *SPEN* (Dox, Aux). Genes are shown in order by chromosomal location. The lower annotation bar classifies genes as NPC-specific, facultative or constitutive escapees. (**c**) as (**b**), for CL31. (**d**) Scatterplot of allelic ratios in control condition against 3 day dox condition (left) and 3 day dox with auxin (right). The diagonal represents equal values between the two conditions, that is, no treatment effect. The marginal distributions are shown as boxplots. Values are averaged between CL30 and CL31. (**e**) As (**d**), but showing 7 day treatment effects.

4.4 Silencing of escapees is partially reversible

During XCI, gene silencing becomes mostly *Xist*-independent after completion of XCI. Inactivation proceeds from an initial reversible state into irreversibility within a few days, both *in vivo* and in a cell line model of the process [Wutz et al., 2002]. As this study uses a Dox-inducible transgene, irreversibility can be directly tested by removing Dox from the culture medium **(Fig. 4.8a)**.

I first confirmed that after 4 or 7 days of washout, irrespective of previous treatment time, *Xist* abundances is reduced to control levels **(Fig. 4.8b)**. I then inspected escape after removal of Dox for seven days, to assure that re-establishment of expression can proceed. I found that after seven days, many genes returned to the same allelic ratio as in control conditions, while others stayed reduced. After 21 days, this effect was substantially more pronounced, with a majority staying silenced and a subset returning to full escape **(Fig. 4.8c)**. Quantifying the distributions of these effects, as silencing became close to complete over time, the amount of recovery was also reduced **(Fig. 4.8d)**. I next quantified reversibility as the fraction of the allelic ratio after washout, compared to control conditions. This analysis showed that constitutive escapees returned to around 80% of escape after 7 day treatment and washout, which did not change much for longer silencing time points. In contrast, facultative escapees showed similar reversibility as constitutive ones after 7 days, but much more stayed silenced after 21 days with washout. NPC-specific genes were largely irreversible already after 7 days, and almost fully irreversible after 21 days **(Fig. 4.8e)**. These results could be confirmed after discretizing irreversibility into reversible, partially reversible and irreversible categories, based on a recovery of 80% or 10% respectively. The vast majority of genes were reversible after seven, but partially or irreversible after 14 and 21 days **(Fig. 4.8f)**.

These results show that sustained *Xist*-overexpression is necessary to maintain the almost fully silenced state of the Xi, but that it can irreversibly silence a subset of genes. I considered that the observed variation in reversibility could correspond to the variability in resistance to silencing I observed in **Section 4.2.3**. I therefore correlated the silencing half-time to the irreversibility fraction. Indeed, there was a small but significant correlation (Pearsons correlation coefficient $r = 0.4$), showing that slow-to-silence genes recover escape more readily, and recapitulating the differences between categories **(Fig. 4.8g)**. Similarly, reversibility of a gene could be associated with its chromosomal context. To this end, I defined locally connected groups of genes. Specifically, I calculated the distance of every escapee to the second nearest one, and traversed the X-chromosome, assigning genes to a group as long as they were within 100kb of each other, and splitting groups whenever the distance was greater. This partitioned genes into 11 clusters with at least three genes, and the rest being singlets or doublets. Region 11 (purple) encompassed four genes, all of which were silenced slowly, and recovered escape largely. Similarly, genes in region 7 (pink) had intermediate half-times and recovery, while the adjecent region 6 (green) was quickly silenced and did not recover. In contrast, region 1 genes (light green) were silenced quickly and were irreversible **(Fig. 4.8h)**.

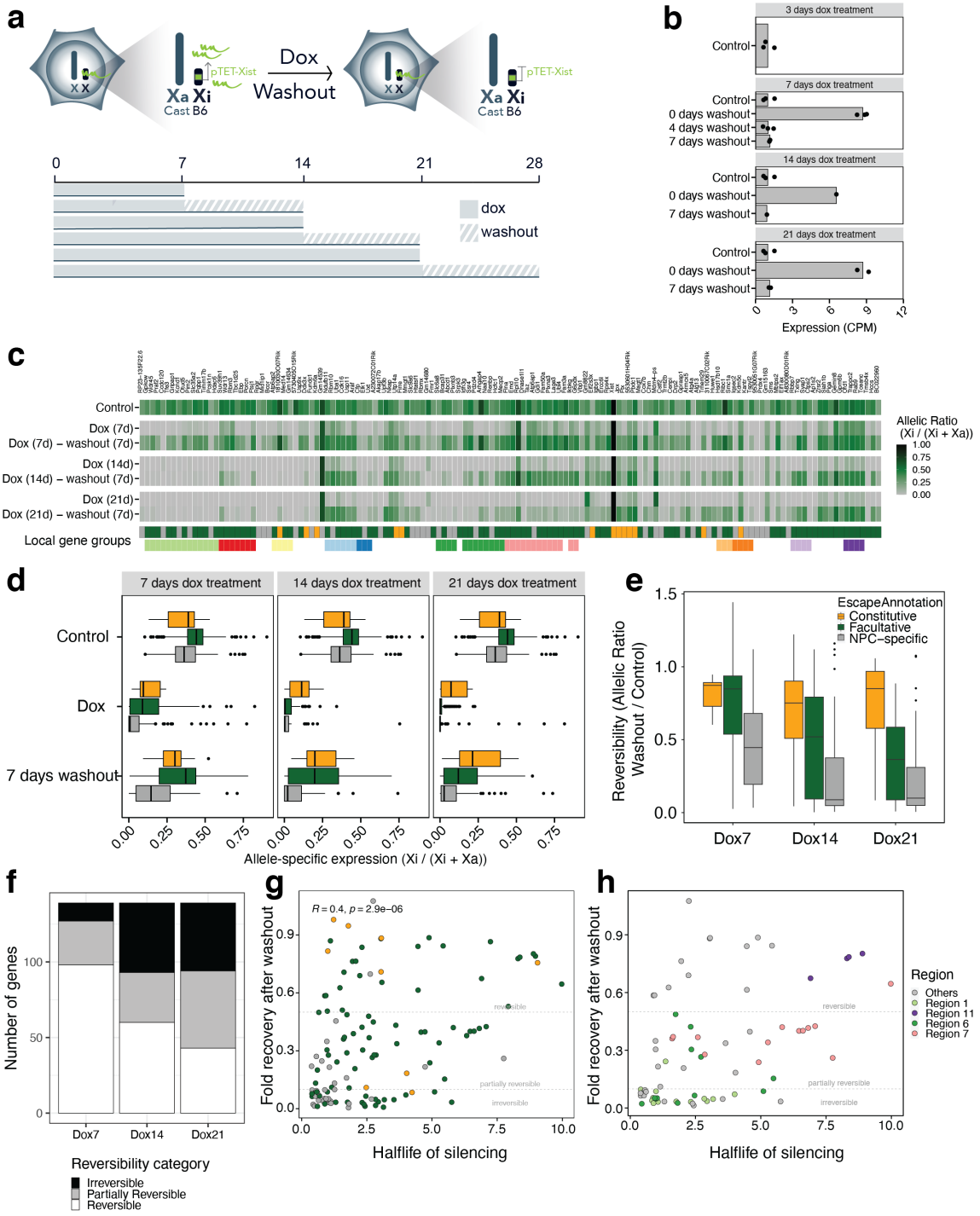

Figure 4.8: Escapee silencing reversibility is gene dependent. (**a**) Scheme demonstrating *Xist* overexpression and washout experiments, modified from graphics generated by Agnese Loda. (**b**) Barplot showing the increase of *Xist* overexpression and loss thereof after washout. Each point represents one replicate experiment (n = 2). (**c**) Heatmap showing allelic ratios of X-linked genes in dox treatment conditions and washout. Genes are ordered by chromosomal position. (**d**) Boxplots showing the distriution of allelic ratios, averaged across replicates. Genes are stratified by escape categories. The colors below indicate clusters of locally close genes as described above (**e**) Boxplots showing "reversibility" of allelic ratios calculated as the ratio recovered after releasing dox. Results are stratified by escape category and silencing duration. (**f**) Discretization of the data in (**e**), by termining genes with less than 10% of escape recovered "irreversible", genes with at least 80% recovered as reversible, and all others as partially reversible. (**g**) Scatterplots comparing reversibility (y-axis) against speed of silencing as calculated by exponential decay fits. Genes are colored by escape category. (**h**) Same results as in (**g**), but colored by clusters of silencing as determined in **(4.5)**

However, I also observed many genes with slow silencing and no recovery, or the other way around, suggesting that other factors affect these behaviours. For example, a number of constitutive singlet escapees showed slow silencing. Collectively, these results show that gene silencing by *Xist* becomes progressively irreversible, as it does during XCI. I also show that the intrinsic resistance to silencing correlates with a genes ability to reactivate. Indeed, it is known that constitutive escapees resist silencing during XCI *in vivo* [Barros de Andrade E Sousa et al., 2019].

4.4.1 Quantifying escape from imprinted XCI

I finally sought to validate these results *in vivo*. At genome-scale, escape from XCI is difficult to measure due to random X-inactivation, unless single-cell methods are employed (as in **Chapter 3**). In mice, an alternative setting in which escape can be assessed is presented by imprinted X-inactivation in the pre-implantation embryo [van den Berg et al., 2011]. The samples used here are female embryos derived from mating a male C57B6 *TetOn-Xist / TetOn-Xist; wt / Rtta* with a female wild type JF1. The TetOn-Xist allele is the same as in the previous figures and allows for the overexpression of *Xist*. Dox-induction in a TetOn system requires a tetracycline-induced trans-activator (Rtta) which is introduced through an autosome **(Fig. 4.9a)**. Since the Rtta is a heterozygous allele, half the resulting embryos will carry the trans-activator and overexpress *Xist* after Dox-treatment, and half will not. Since female embryos inherit an X from the C57B6 and JF1 strains each, the resulting embryos are also hybrids which allow for allelic quantifications, similar to the C57B6 x CAST cross in the previous sections **(Fig. 4.9b)**. As imprinted X-inactivation silences the paternally inherited, in this case the B6 X-chromosome, full allelic bias towards the JF1 allele is expected.

The individual embryos showed equal expression from the two haplotypes across autosomes, and expected skewed expression on the Xi **(Fig. 4.9c)**. While most genes showed inactivation, there was a subset of genes with bi-allelic expression, consistent with escape. I first used principal component analysis (PCA) to group the individual embryos by their autosomal gene expression patterns. The samples seperated across PC1 by their morphology-based annotation into morulas (earliest embryonic stage profiled), early and late blastocysts **(Fig. 4.9d)**. I therefore used PC1 to order embryos through their developmental progression. I next computed average allelic ratios of Xi expression across the X for every embryo to capture the extent of escape. This measure was anti-correlated with PC1, consistent with the progression of imprinted XCI. I next mapped sequencing reads to the Rtta transgene and detected reads in around half the sequenced samples. Rtta+ (Rtta positive) embryos tended to have lower expression from the Xi than Rtta- ones, suggesting that overexpression of *Xist* affects escape from imprinted XCI. Importantly, Rtta+ embryos were distributed similarly across pseudotime in late blastocyst stage samples. This suggests that *Xist*-overexpression does not affect their developmental progression and that the embryos tolerate this perturbation well, making it less likely for the effects on escape to be secondary.

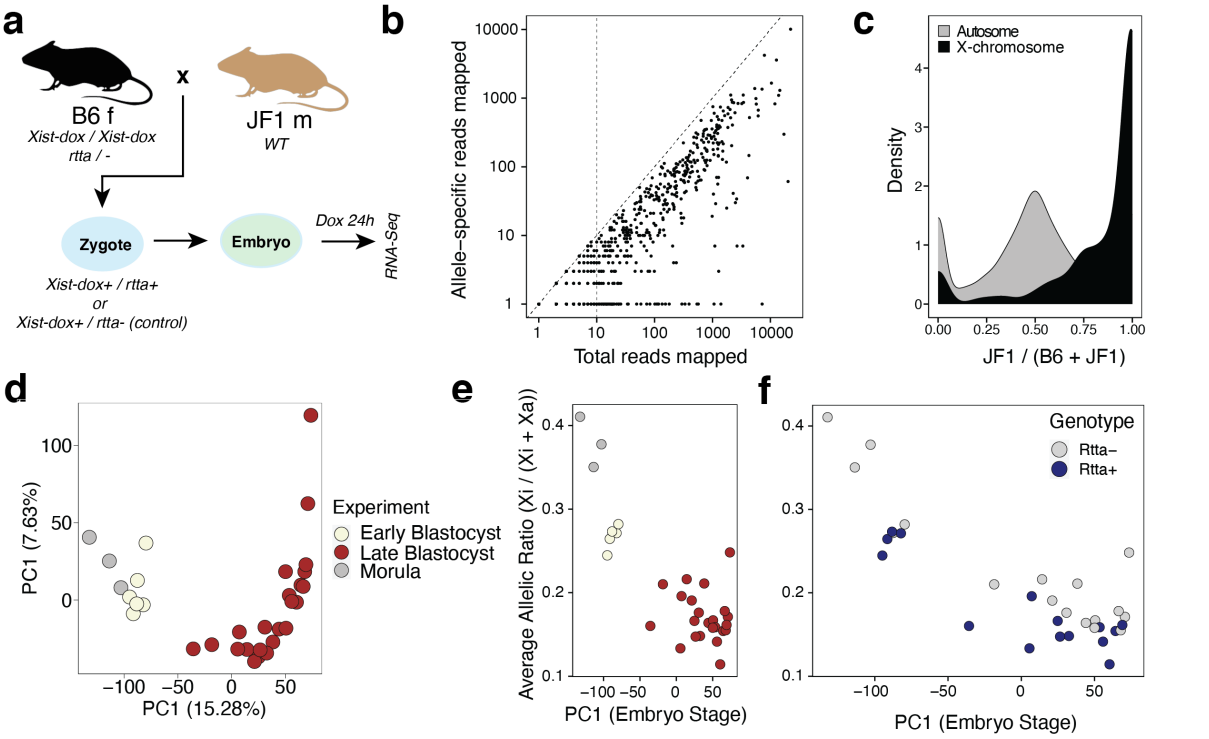

Figure 4.9: Quantifying escape from imprinted XCI. (**a**) Experimental outline of the embryo experiments. Male B6 *TetOn-Xist / TetOn-Xist; wt / Rtta* were crossed with wild type JF1 mice. All female offspring will carry the *TetOn-Xist* allele, and half will or will not carry the *Rtta* transactivator allele. The *Rtta*+ embryos will overexpress *Xist* in response to Doxycycline exposure, while the *Rtta*- will not and serve as a control. (**b**) Scatterplots showing total reads per gene against allele-resolved reads for one sample, demonstrating high resolution of allelic quantifications. (**c**) Density plot showing allelic ratios on autosomes and the X-chromosome on the same sample. (**d**) PCA of total expression data of the embryo dataset. Samples are annotated by morphological assignment of developmental stages. The first principal component can be used as a pseudo-temporal ordering. (**e**) Scatterplot showing PC1 against the average allelic ratio of X-linked genes. (**f**) as (**e**), but annotating by the presence of reads mapping to the RTTA-transgene.

4.5 Xist overexpression silences escapees during imprinted XCI

I next assessed silencing of escapees *in vivo* at the level of individual genes. Since the dataset only contained sufficient embryos in late blastocyst stage in both conditions, I ignored the other two groups in this analysis. I first computed average allelic ratios in Rtta- and Rtta+ embryos and visualized these across their chromosomal position. This analysis showed mild reductions of allelic ratios across a small number of genes **(Fig. 4.10a)**. Quantitative aggregation over all genes show that facultative and NPC-specific categories showed a clear reduction in allelic ratios **(Fig. 4.10b)**. At the level of single genes, a substantial number of expressed genes showed a significant reduction in escape, a smaller amount did not change, but no genes increased in escape. Among the most affected genes were *Atrx, Usp9x* and *Hcfc1*. These results collectively demonstrate that increasing *Xist*-levels affect escapees also during embryogenesis. As expected, the reduction is substantially more subtle than in the cell line, as *Xist*-exposure is both shorter (24 hours) and likely increases its transcript levels to a lower extent.

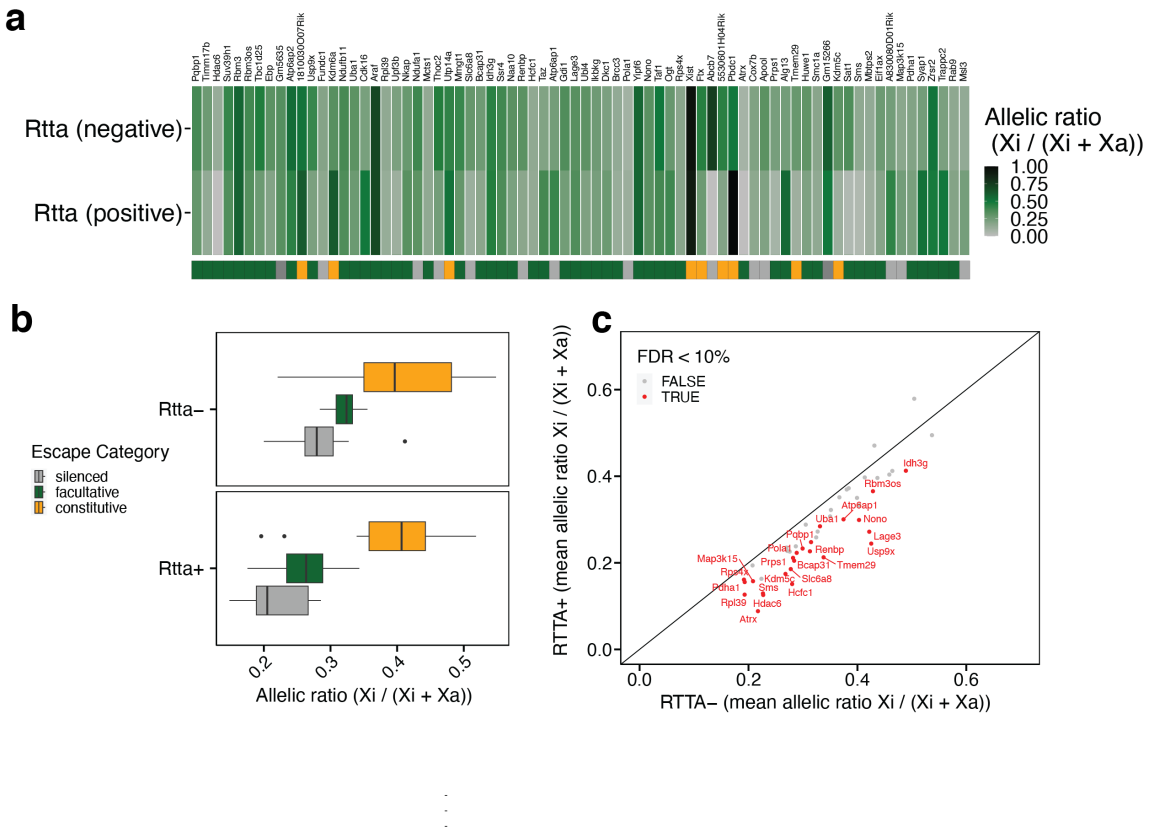

Figure 4.10: Xist levels control escape from imprinted XCI. (**a**) Heatmap showing average allelic ratios of X-linked genes in pre-implantation embryos, separated by controls (Rtta-) and Dox-treated samples (Rtta+). Annotations below the heatmap show escapee categories, for the legend, refer to **b**. (**b**) Boxplots summarizing the results in (**a**). (**c**) Scatterplot showing allelic ratios in Rtta- and Rtta+ conditions where values represent averages across all embryos. Significance was assessed using Students t-test.

4.5.1 Discussion

In this chapter, I used data from RNA-Seq experiments where *Xist* levels are titrated to demonstrate that the lncRNA has the capacity to exert its silencing function even after XCI is completed and in particular, that it can elimeinate escapee expression almost fully. This silencing is dependent on the co-factor SPEN, suggesting that mechanistically, *Xist* acts through the same pathways in pluripotent as well as differentiated cells and relies on the XCI initiation machinery *Dossin*. Furthermore, I use modelling of allelic trajectories to show that genes differ in their resistance to silencing in two parameters, namely residual escape after long-term *Xist*-exposure and their speed in loss of expression. Although constitutive escapees show the highest resistance in both of these characteristics, other genes show similar resistance, which might be linked to their genomic location or other gene-dependent features. Interestingly, inactivation is mostly reversible after short overexpression, and partly irreversible after long-term exposure, with the extent of re-activation being dependent on their initial resistance. Finally, modulation of *Xist*-levels impacts escape during imprinted XCI *in vivo*.

These results provide a possible pathway through which escape might be controlled in different contexts. A critical question is to which extent *Xist*-levels vary across physiological or pathological contexts. Recently, it has been suggested using single-cell analysis, that

its variation during the cell cycle controls escapee expression levels [Garieri et al., 2018]. While I see that *Xist* is robustly expressed across cell types and ages in **Chapter 3**, it might vary specifically cancer [Ma et al., 2023, Richart et al., 2022]. Furthermore, differential function might be driven by localization differences in the absence of expression differences [Jacobson et al., 2022, Wang et al., 2016]. Finally, changes in *Xist* expression might exert effects on gene expression only as a long-term consequence.

The fact that resistance to silencing seems to be locus-specific adds to the knowledge that *Xist* acts in chromatin domains which are proximal in 2d or 3d space [Engreitz et al., 2013]. Indeed I and others observe that escape usually occurs in clusters of genes, and that these show similar responses during inactivation [Barros de Andrade E Sousa et al., 2019]. It will be critical to quantitatively assess to which extent escape is driven by gene-specific sequence elements, the general epigenomic state of the locus, and its local chromatin environment. So far, different kinds of escapees have mainly been classified empirically based on the number of studies in which they are detected. My results might help to instead describe constitutive and facultative escape in terms of genomic and epigenomic features, and to test whether there is a genuine difference between these two categories.

Finally, this study presents a single perturbation of a *trans*-acting factor on escape, but recent technical advances allow to perform such assays at scale. By performing pooled perturbations and reading out the responses using single-cell sequencing, one can assay the genome-wide effects of these perturbations at scale [Morris et al., 2023]. This can be done using classical CRISPR-knockout experiments, or to tune expression levels using CRISPR-interference or activation. By combining pooled screens with allele-specific single-cell assays as I outline in **Chapter 3**, one could assess the effects of many perturbations on escape in a parallel manner. In particular, this would allow to compare the effects across genes and identify mechanisms that drive resistance to silencing.

Chapter 5

Discussion and future perspectives

In this thesis, I have demonstrated how single-cell RNA- and ATAC-sequencing focussed on allele-specific quantifications is a powerful approach to map cell type-specific allelic usage caused by genetic and non-genetic effects, and to use this data to infer gene regulatory principles. The presented work relies on mouse models that allow for genome-wide allelic discrimination, high-throughput single-cell sequencing assays, and statistical models of fractional read count data using binomial (mixed) models. In **Chapter 2**, I have generated a first of its kind dataset to map *cis*- and *trans*-genetic effects at cellular resolution using a classic F1 hybrid approach. I reveal that cell type-specific modulation of genetic effects is pervasive during cellular differentiation *in vivo*, and that cell type-specific genetic effects, at least between mouse species, are usually *cis*-driven. By comparison to an additional mouse strain, I then show that cell type-specific evolution rates are driven by dynamic genetic effects. Next, I move to the analysis of allele-specific expression and chromatin accessibility caused by X-chromosome inactivation (XC) in **Chapter 3** (data generated by Stefania del Prete). I develop an approach that ”phases” the inactive X-chromosomal haplotypes at the single-cell level, and thereby quantifies both skew in X-inactivation and escape from XCI, even though the allele-specific data is sparse at the single-cell level. This approach reveals unappreciated heterogeneity in escape from XCI in immune cells, which might contribute to sex-biased autoimmune disease. I next test the hypothesis that ageing leads to a loss of robust X-inactivation, and find that this only holds in differentiated T-cell populations. I also generate cell type-specific maps of accessible chromatin on the inactive X. Following up on these results, in **Chapter 4**, I use data generated by Antonia Hauth and Agnese Loda to test whether the *Xist* long non-coding RNA can silence escapees in cells after X-inactivation is completed. Using a simplified neural progenitor cell system, I show that escape from XCI is almost fully abolished after prolonged increase in *Xist*-levels. Furthermore, I use time series modelling to show how silencing behaviours are strongly gene-dependent, and that stronger resistance is associated with reversibility, suggesting intrinsic variation in a genes propensity to escape XCI. Finally, I show that *Xist* can silence genes *in vivo*. These results demonstrate that allelic imbalance can be strongly cell type-specific, motivating further analysis in different contexts. These results demonstrate the power of combining cell type and state measurements obtained from total single-cell data with the orthogonal information from allelic readouts, which is conceptually equivalent to vertical integration approaches using multi-modal single-cell measurements [Argelaguet et al., 2021]. Additionally, the last project describes a potential mechanism that drives context-specific allelic imbalance, which so far has been largely unexplored.

However, these results strongly rely on model organisms and the advances are in some respects rather conceptual. I will discuss future methodological challenges, especially when transferring these approaches to non-model organisms. Then I will outline open biological questions regarding allelic imbalance and its implications for gene regulation, and how single-cell approaches might help to address them.

5.1 Technical challenges when measuring allele-specific expression

5.1.1 Application to human samples

All sequencing-based measurements of allelic imbalance rely on the presence of heterozygous variants. The use of interspecific F1 crosses provides SNPs at a frequency of, depending on the cross, at least 1 / 130bps, which means that every gene will contain multiple variants. In this way, even RNA-Sequencing approaches that do not measure the entire length of a transcript will contain sufficient information to assay a majority of expressed genes, as demonstrated in this work. However, in a typical human genome, heterozygous variant frequencies of around 1 / 1300bps are expected, a ten-fold decrease [Lander et al., 2001]. Furthermore, genes or sequences with regulatory function are further depleted of variants through selection. Therefore, 3'-biased droplet-based assays might not be suitable outside of using hybrids between inbred strains. Previous work has used full-length RNA-Seq such as the Smart-Seq2 protocol, which increases the probability of detecting variants, even if the majority of reads are not allelically resolvable, and this approach has been extended to droplet-based assays [Picelli et al., 2014, Hagemann-Jensen et al., 2020, Salmen et al., 2022]. However, these approaches do not use single-molecule counting and resulting counts can be difficult to interpret due to PCR amplification bias.

Recently, long-read sequencing based on PacBio and Oxford Nanopore technologies has been continuously improved and is now being applied to single-cell RNA-Sequencing. This approach is powerful and generalizable, as any sequencing protocol generating full-length cDNA libraries can be used for this approach, and drawbacks of long-read methods including low throughput and in the case of Oxford Nanopore, high error rates are continually being improved. A recent study has demonstrated the possibility of using long-read sequencing in conjunction with allele-specific analysis to map *cis*-effects in human tissues [Glinos et al., 2022]. While improving allelic resolution, this approach also directly quantifies transcript usage and links it to allelic balance, which is difficult with short read sequencing [Leigh-Brown et al., 2015]. Indeed, differential transcript presents a complication to the analysis of genetic variants based on *cis*-effects if isoforms cover different SNVs. Long-read single-cell expression profiling with a focus on allelic imbalance will likely resolve the cell type-specificity of genetic variants on expression and transcript usage in the near future [Philpott et al., 2021, Shi et al., 2023]. An additional possibility is opened up by targeted sequencing approaches [Schraivogel et al., 2020]. By PCR-amplification of specific transcripts or regions of these, sequencing assays can be focussed on measuring only variant-containing parts of the library. For hybrid mice used in this study, only 40% of reads contain allelic information, and the benefit will likely be much larger in human samples. A similar focus on the discernable fraction of the transcriptome can be achieved using adaptive-sampling in Nanopore sequencing [Weilguny et al., 2023].

5.1.2 Alternative technologies to measure single-cell allelic imbalance

Besides improvements of sequencing assays, other ways of measuring allelic imbalance have to be considered. In parallel to single-cell genomics, fluorescence in situ hybridization (FISH) has been the second main approach to quantify transcript abundances in single cells. So called single-molecule FISH (smFISH) technologies have a number of advantages over sequencing-based methods. While transcript capture in scRNA-Seq is inherently lossy (it is estimated that 1-10% of transcripts are recovered), smFISH can in principle achieve full sensitivity [Chen et al., 2015, Borm et al., 2022]. Furthermore, FISH assays cells without dissociation in native tissues and therefore informs the spatial context of each cell, and even sub-cellular localization of transcripts. However, smFISH is an imaging-based readout, and therefore suffers from lower throughput, with the number of genes depending on multiplexing schemes that are limited to 100s of individual measurements. For highly expressed genes, single transcripts can not be distinguished in the resulting microscopy images and the sample-level throughput is low. FISH has been used to detect allelic variants of the same gene by designing probes that overlap sequence variants and are not cross-reactive [Ginart et al., 2016, Herzing et al., 2002]. In principle, this approach can be scaled up in parallel to recent smFISH developments, although each gene would require multiple multiplex-slots to be detected at allelic resolution.

A fundamental limitation to all so far discussed approaches is that they require genetic variation between the alleles of interest, which do not necessarily exist. In the absence of sequence variation, the only natural discerning feature between two homologous chromosomes is their position in the nucleus. As FISH-based methods provide subcellular spatial resolution, it is in principle possible to measure active transcription at two sites in a single nucleus. Indeed, this is a common approach to characterize active and inactive X-chromosomes, where the Xi can be identified by co-localization with *Xist* RNA, and active transcription can be measured using intronic probes [Yue et al., 2014]. Further technical improvements are necessary to increase the number of assayed features and to make the measurement truly quantitative, but this provides an approach to directly measure allelic usage without genetic variation. Finally, alleles can be distinguished by genome engineering. Classical papers studying single genes used dual fluorescent reporters to measure allelic usage [Oghumu et al., 2019, Wu et al., 2014]. It is conceivable that these approaches can be scaled up to many genes, potentially through targeted insertion of heterozygous variants which are predicted to have no effect on expression or gene regulation, which can then be used in the same way as natural variation.

5.2 Expanding the scope to novel omics approaches

5.2.1 Multimodal allele-specific profiling

This thesis largely focus on allelic imbalance in gene expression, but it is clear that allelic measurements can and should be extended to other molecular layers. I have addressed chromatin accessibility in **Chapters 2** & **3**, which has been the most common target of allelic analysis previously [Yang et al., 2022a]. My results show coordination between chromatin accessibility and expression at escape loci and due to genetic effects. Previous studies have investigated allele-specific CHIP-Seq, chromosome conformation, histone marks, chromatin accessibility, RNA stability and translation through ribosome-bound RNA [Wong et al., 2015, Giorgetti et al., 2016, Yang et al., 2019, Sun et al., 2018b, Ozadam et al., 2023]. Some of these have recently been demonstrated to be feasible in single

cells [Collombet et al., 2020, Ozadam et al., 2023, Heinen et al., 2022]. An exciting direction is the possibility of integrating allelic signals into multi-modal measurements, which become increasingly feasible. A recent study has examined the partial correlations between allelic usage in chromatin accessibility, histone marks and gene expression, demonstrating different modes interactions particular, a specific coordination as compared to total expression levels [Floc'hlay et al., 2021]. Performing these assays at the single-cell level will for example allow to connect allelic signal between any other pair of molecular layers. In particular, correlation analysis has proven powerful to link the activity of regulatory elements to their target gene expression in multimodal RNA and ATAC data [Pliner et al., 2018]. However, this correlation often spans diverse cell types, which is prone to introduce false positive interactions based on cell type-specific regulatory element activity. A more direct approach would correlate allelic CRE usage to allele-specific expression in homogeneous populations which should yield stronger candidates for functional relevance.

A specific application of multi-modal allelic data will be the full decoding of gene expression dynamics. Recent work [Gorin and Pachter, 2022] has suggested that mechanistic models of gene expression, including transcriptional bursting and mRNA processing are necessary to assess differential expression or its variability between conditions. This description should also include allelic resolution and inferences of transcriptional parameters will benefit from it.

Multimodal assays also address the missing connection between gene expression and protein levels. Whether allelic imbalance actually leads to differential contribution to protein levels remains unclear in the studies presented here. There has been an attempt to use mass spectrometry to directly detect proteins derived from specific alleles, but this requires exceedingly rare coding variants. The most proximal readout is ribosome-bound RNA profiled through single-cell ribosome profiling [Ozadam et al., 2023].

5.2.2 Allele-specific perturbation experiments

After identifying context-specific allelic imbalance, the next step is to find causes of it. This usually requires functional experiments in which candidate factors are perturbed, to determine whether they are causally linked to allelic bias. Recently, scalable designs for CRISPR screening assays allow to test multiple perturbations in a single experiment. This is achieved by introducing randomly selected guide RNAs into individual cells, which can be read out in parallel to single-cell genomics data. The result is a dataset in which small groups of cells each carry one of many perturbations, and the transcriptome-wide effecst can be assessed. In the most basic setting, CRISPR-based deletion of target genes is coupled to scRNA-Seq (Perturb-Seq), but recent work has extended this to different measurement modalities (ATAC-Seq and isoform usage [Rubin et al., 2019, Kowalski et al., 2023]) and various methods of gene perturbation (CRISPR-inhibition and activation [Morris et al., 2023]). These approaches can be readily extended to assaying cellular contributions to allelic imbalance (for possible applications, see **Section 5.7**).

High-throughput perturbation experiments can be similarly used to determine the effect of allelic imbalance on downstream phenotypes. Guide RNAs that overlap variants in critical positions can be used to generated allele-specific perturbations [Li et al., 2020], and gene expression can be tuned using CRISPR-inference to match observed variation between alleles [Noviello et al., 2023].

5.3 Directions in computational modelling

This thesis heavily relies on statistical modelling to describe allele-specific read count data. In particular, I make use of the scDALI model which models cell state-driven variation in allelic imbalance which is inferred from total gene expression profiles. The idea of quantifying allelic variation based on transcriptomic state has been used in several different models, that further explore its application to context-specific genetics in multi-sample datasets [Qi et al., 2023] and in spatial variation of ASE [Zou et al., 2021]. Further extensions of this class of mixed models to multiple random effects will allow to, for example, detect changes in allelic usage driven by different experimental conditions, in parallel to the increased complexity of mixed models for QTL analysis [Cuomo et al., 2022].

While the beta-binomial likelihood is a natural choice for allelic read counts, there has been no systematic comparison against alternative models. For example, beta- and binomial distributions are difficult to estimate when the mean parameter is close to the boundaries. In this case, it might be beneficial to switch to allelic fold change models, where the allelic information is encoded by an additional parameter in a standard poisson- or negative binomial linear model, and which are well established as suitable for sequencing count data [Mohammadi et al., 2017].

Another opportunity for further developments are models that integrate total expression information and allelic signals in an unbiased manner. I have previously suggested that the joint analysis of allelic and total expression levels resembles the integration of multi-modal datasets, where multiple data modalities are observed in the same cell. For multi-omics datasets, dimensionality reduction approaches based on factor analysis or non-linear embeddings based on variational autoencoders have proven powerful [Argelaguet et al., 2018, Lopez et al., 2018]. These methods may be a blueprint for an unbiased method to relate allelic to non-allelic signals.

Finally, much work has been done to reliably quantify variability in gene expression [Vallejos et al., 2015]. For allelic data, this is represented by the overdispersion parameter of a beta-binomial distribution. In the presented data, most of the allelic counts so low at the single cell level that they are best described by Bernoulli-distributions, which do not allow for a disambiguation of mean and overdispersion without further assumptions. When the technology allows for more dense allelic data, these approaches will be critical to distinguish technical from biological variability within homogeneous populations.

5.4 What are biological implications?

A critical question is for which genes it is necessary to treat the two (or more, in the case of polyploidy) copies separately in order to understand their biology. With the described technological advantages, it will become possible to comprehensively identify all allelic variation, which mechanisms cause allelic bias and its impact on human variation or disease. Using single-cell technologies, it will be especially important to quantify and distinguish the different sources of allelic imbalance. On the one hand, much of the variation caused by genetic effects, imprinting or XCI will be identifiable across cells and cell types, although possibly with modulation by cellular context. Additional variation will stem from transcriptional dynamics, and it is somewhat open which of these contributes the most. A relatively unexplored third contribution is be allelic imbalance that is mitotically heritable, but not due to genetic effects, imprinting or XCI. In clonal cell lines, such allelic bias has been

observed and termed ”random mono-allelic expression” (RAME), or analogously RAMA for chromatin accessibility [Xu et al., 2017]. While subject genes seem to be consistent between studies and are associated with changes in chromatin state and DNA methylation, these effects remain largely unexplored mechanistically. Importantly, the existence of these genes has not been confirmed *in vivo*, largely because clonal tracing in animals is challenging. Synthetic cell type labels, for example through CRISPR-scarring methods now allow to trace clones organism-wide, but their resolution is still limited [Goyal et al., 2023]. The same is true for the use of natural barcode such as mitochondrial mutations in primary samples [Ludwig et al., 2019]. In principle, integration of such lineage markers might answer whether somatic mutations or epigenetic mechanisms confer allelic bias sub-clonally and how prevalent that might be. A recent study has taken a related approach in testing for clonally heritable gene expression states, without alellic resolution, and found clone-dependent differences in expression levels for a subset of genes in T-lymphocytes. Whether this principle generalizes to other tissues, especially those with fast turnover, remains open. A complete organismal single-cell map, analogous to the studies of the Human Cell Atlas and across-tissue mouse atlases should include allele-specific analysis using appropriate sequencing technologies [Regev et al., 2017, Cusanovich et al., 2018]. Extending these surveys to developmental stages will help to elucidate the extent to which allelic bias is set during embryogenesis, or is a result of somatic evolution in the adult.

My results identifying age-related increase of escape in memory T-cells in **Chapter 2** is likely provide an example of clonally inherited changes in allelic balance. This could be directly tested by using clonal tracing through T-cell receptor sequences, a naturally generated barcode. Furthermore, **Chapter 3** demonstrates a potential mechanism that changes escape across cell types, or of specific cells within a population. Notably, even though I do not observe changes in *Xist* levels across cell types and ages, *Xist* activity might differ through changes in cofactors or differential localization. These results suggest that escape is likely more plastic than anticipated. This has been confirmed by further recent studies that address variation in escape across tissues [Tukiainen et al., 2017, Berletch et al., 2015]. Finally, I have directly addressed the question whether ageing drives changes in escape. I demonstrate that this is not a universal property of the Xi, but is likely due to strong proliferation bottlenecks during T-cell activation and expansion. Lymphocytes represent a particular case, as they are both long-lived and proliferating, which might be required for a loss of silencing to become apparent. Highly proliferative cell types are usually found in organs with high turnover (for example, epithelial tissues), which where cells are shed before they can accumulate greater changes in escape. On the other hand, long-lived cells such as neurons do not proliferate further, which might maintain a stable inactive X-chromosome. While my results show some suggestion of selection for escape, this will have to be shown directly through overexpression assays.

Furthermore, there are stark differences in escape when comparing mice and humans. Although the same set of escapees in mice tends to be conserved in humans, escape is thought to be much more frequent in humans. To address the impact of escape on human sex-biased phenotypes such as autoimmune disease, it will be necessary to directly assess escape across cell types and ageing. A recent study has attempted such an analysis in 10x Genomics single-cell data of PBMCs in japanese individuals [Tomofuji et al., 2023]. This study provides a powerful demonstration that a phasing-based approach similar to the one I developed can be extended to human samples. In particular, it showcases how Xi assignments can be made without fully phased genomic data. However, it also suffers from low signal due to the use

of 3'-biased RNA-sequencing in human samples, profiling only a small fraction of X-linked genes. Furthermore, it did not carefully assess the validity of their Xi assignments or allelic quantifications, which I have shown to introduce false positive signal even in highly defined genomes.

My results on cell type-specific allelic imbalance caused by genetic effects supplements many recent studies on cell type- and differentiation-dependent QTLs in human samples [Cuomo et al., 2021, Jerber et al., 2021, Findley et al., 2021, Ward et al., 2021] and other model organisms [Francesconi and Lehner, 2014]. It is becoming increasingly apparent that *cis*-effects are strongly dependent on the cell type they are measured in. The prevailing model of these dynamic QTLs is that they affect regulatory elements, usually through transcription factor binding changes. Using perturbation assays, it will be possible to comprehensively test the effect of transcription factors on genetic effects. A recent proof-of-concept study deleted all predicted DNA binding proteins and tested their effects on gene expression [Joung et al., 2023], an approach that could be used equivalently to determine allelic effects of TF deletions in different cell types. This will reveal how much variation can indeed be explained by TF abundances, nominate candidate mechanisms for genetic traits, and provide an estimate for the extent to which TFs determine cell types. Furthermore, this approach will suggest for which QTLs transcription factor abundance does not explain variability, which might be driven through chromosome structure or RNA-RNA interactions.

A larger question is in which way accumulated sequence changes lead to gene expression evolution between species. The F1 hybrid setup is a powerful demonstration that most transcriptional changes between sub-species are derived directly from sequence variation rather than secondary changes in *trans*-acting factors. Notably, this approach can be extended to species that can not generate viable offspring (including humans) by cell fusion experiments, where tetraploid hybrid cells are generated artifically *in vitro*. This approach has been of particular interest in models of brain development, where humans are thought to present specific evolutionary innovations [Gokhman et al., 2021]. These results suggest that *cis*-driven changes are the stronger contributor to the divergence of expression levels, but also highlights the importance of specific *trans*-actors.

Chapter 6

Appendix

6.1 Materials and Methods

This section contains experimental methods and additional details on computational analysis where it is not fully described in the results section.

6.1.1 The dynamic genetic determinants of increased transcriptional divergence in spermatids

All experiments in this chapter were performed by me unless indicated and the methods descriptions here are published in similar form in the aforementioned paper.

Mouse strains. The mouse strains used in this work were purchased from Jackson laboratories and maintained in house. Specifically, C57BL6-Ly5.1, CAST/EiJ and CAROLI /EiJ strains were used (Strains #002014 and #000928, #000926). F1 hybrid mice were generated by mating C57BL6-Ly5.1 (female) and CAST/EiJ (male) mice and in the opposite direction (C57BL6-Ly5.1 (male) and CAST/EiJ (female)). Mice were used after two months of age, when spermatogenesis is fully established. Mice were housed in the DKFZ mouse facility or the Biological Ressources Unit (Cancer Research UK center) under specific pathogen-free conditions, fixed 12h day-night cycles, ad libitum access to food and water, a humidity of 80% and 24 degrees celsius. Sacrificing was performed through cervical dislocation, according to ethics guidelines for experimental animals (approved by the Regierungspräsidium Karlsruhe or the Animal Welfare and Ethics Review Board, following the Cambridge Institute guidelines).

Single-cell RNA-Sequencing of mouse testicular tissue. The protocol for 10x Genomics-based single-cell RNA-Sequencing of mouse testis cells follows [Ernst et al., 2019]. For the cross-evolution dataset (**Figure 2.16**) Testes were surgically dissected and the tunica albuginea was removed using forceps. The extracted seminiferous tubules were digested for 30 minutes at 37 degrees celsius using 25mg/ml Collagenase A (Sigma, 10103578001), 25mg/ml Dispase II (Sigma, D4693) and 2.5mg/ml DNAse I (Sigma, 10104159001). During digestion, the tissue was gently triturated using a 1ml pipette. The resulting single-cell solution was passed through a 40 μm strainer, counted and used immediately for single-cell RNA-Sequencing. 10000 cells were loaded into one channel of the Chromium™ Single Cell A Chip (10X Genomics®, 1000009) and scRNA-Seq libraries were generated using the Chromium™ Single Cell 3' Library & Gel Bead Kit v2 (10X Genomics®, 120237) according

to the manufacturer's instructions. Resulting libraries were sequenced using an Illumina HiSeq2500 machine with read lengths of 26bp for read 1 and 98bp for read 2. The sequencing data for the two B6 samples in this dataset was previously published [Ernst et al., 2019] and is available under the accession number E-MTAB-6946. *These experiments were performed by Christina Ernst.*
The F1 dataset was generated using the same procedure, with modifications. The dataset was generated was three independent sequencing experiments each comprising two individuals of C57BL6-Ly5.1, CAST/EiJ and F1 mice. The enzyme concentrations were reduced 5-fold, which yielded similar digestion progress after the same time as the original protocol. Libraries were generated using the Chromium™ Single Cell B Chip (10X Genomics ® 1000073) and Single Cell 3' Library & Gel Bead Kit v3 (10X Genomics ®, 1000075) kits and sequencing was performed on a NovaSeq 6k with 28 bp read 1 and 94 bp read 2.

ATAC-Seq of F1 spermatocytes. Spermatocytes were isolated based on nuclear DNA content by fluorescence-activated cell sorting as described in [Ernst et al., 2019]. Cells were incubated for 45 minutes at 37°C with 5μg/μl Hoechst 33342 (R37165, ThermoFisher). Next, the cells were resuspended in phosphate buffered saline (PBS, Sigma) with 1% Fetal Calf Serum (FCS, Gibco, 16140071) with propidium iodide (P4170) at a final concentration of 1 μg/ml to detect live cells, and 50.000 cells were sorted on a BD FACSAria Fusion machine (Hoechst: Excitation 405nm, 450/50 filter, PI: Excitation 488nM, filter 616/23). I performed ATAC-Seq on bulk populations as described with modifications [Corces et al., 2017]. Cells were washed with 500μl ice-cold PBS and then incubated on ice for 3 minutes in 50μl cell lysis buffer (10mM Tris-HCl pH 7.5 (AM9850G, LIFE Technologies), 10mM NaCl (AM9760G, ThermoFisher), 3mM MgCl (AM9530G, ThermoFisher), 0.1% NP-40 (85124, LIFE Technologies), 0.1% Tween-20 (P1379, Sigma), 0.01% Digitonin (BN2006, LIFE Technologies)). 1ml wash buffer (10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM MgCl, 0.1% Tween-20) was used to rinse the lysis buffer, and permeabilized cells were centrifuged for 10 minutes at 500g and 4°C. Transposition of open chromatin was performed by adding 50mul transposition mix (25mul 2x TD buffer (Illumina, 20034197), 16.5mul PBS, 0.5mul 10% Tween-20, 0.5mul 1% Digitonin, 2.5mul tagment DNA enzyme (Illumina, 20034197)) to the permeabilized cell pellet which was then incubated at 37°C for 30 minutes. Tagmented DNA was isolated using MinElute PCR Purification Kit (28006, Qiagen) and libraries were amplified using barcoded primers and the NEBNext® High-Fidelity 2X PCR Master Mix (M0541S, NEB). Libraries were subjected to quality control and sequenced on an Illumina NextSeq2000 sequencer with paired end 100bp read lengths.

Expression quantification of 10x scRNA-Seq data for different mouse strains. For alignment of 10x scRNA-Seq data, genomic references for the strains C57BL/6 (GRCm38), CAST/EiJ and CAROLI/EiJ were created using CellRanger mkref (v3.1) from the ensembl release 94. The resulting filtered count matrices after cell identification and per gene expression quantification were generated using CellRanger count (v3.1) with default settings. For the F1 dataset, a C57BL/6 - CAST/EiJ cross-strain reference was generated based on the GRCm38 sequence with SNP positions between mm10 and CAST/EiJ replaced by the ambiguous nucleotide N. These variants were derived from ftp://ftp-mouse.sanger.ac.uk/current_snps/mgp.v5.merged.snps_all.dbSNP142.vcf.gz) [Keane et al., 2011]. Against this reference, total expression was quantified by CellRanger count (v3.1) for C57BL/6, CAST/EiJ and hybrid mice.

Low-level analysis of scRNA-Seq data and cell type annotation. This is an expanded version of the preprocessing which is described in the results section of **Chapter 2** and is similar to the procedure described in [Ernst et al., 2019]. It largely relies on functions from the scran (v1.20.1) and scater (v1.20.1) R packages [McCarthy et al., 2017, Lun et al., 2016]. Cells were excluded if the showed less than 500 UMIs and 500 detected genes. The resulting counts were normalised with the computeSumFactors function and log-transformed. For cell type annotation, mutual nearest neighbour-based batch correction was performed using the function MNNcorrect with the individual sample as batch variable to exclude both species-specific and technical variation across samples [Haghverdi et al., 2018]. For dimensionality reduction, the resulting corrected count matrix was used for dimensionality reduction through principal component analysis (prcomp, stats, v4.4.0) followed by tSNE (Rtsne, Rtsne, v0.15) and UMAP (umap, umap, v0.2.7.0). To identify clusters corresponding to different cell types, graph-based community detection using the Louvain algorithm was used which is implemented by the functions buildSNNGraph and cluster_louvain in the package igraph (v1.2.10). Cell type labels were then defined by comparison to the [Ernst et al., 2019] samples: For the cross-species dataset the cell type labels could be used to annotate cluster identities. For the F1 dataset, clusters were annotated using marker genes for somatic cells (Sertoli, Leydig and Immune / other structural cells, which were discarded for most further analysis) and the established order of cell types across the latent structure of the expression space during spermatogenic differentiation. Pseudo-temporal ordering of germ cells was generated using principal curve analysis implemented in the package princurve (v2.1.6). This was based on the first two principal components computed across cells. For the cross-strain dataset, a pseudotemporal ordering for cells was derived by computing the median pseudotime value for the 50 nearest neighbours in F1 dataset. *Preliminary sripts for preprocessing were written by Nils Eling and modified by me.*

Quantifications of allele-specific read counts. To quantify allele-specific read counts, first, the output bam file from cellranger (possorted.bam) was annotated with the B6/CAST heterozygous SNPs using a script from the WASP-pipeline with modificaitons (find_intersecting_snps_10x.py) [van de Geijn et al., 2015]. Then, individual reads that contained one or more alleles from the maternal or paternal haplotype were counted, while discarding UMI duplicates, reads overlapping indels and reads with conflicting SNP identities which constituted less than 0.1% of all reads as likely sequencing errors. To compare allelic-specific expression between F1 mice and F0 parents, all libraries were quantified against the same N-masked reference. This approach validated that >98% of reads from F0 animals that were assignable indeed mapped to the correct reference allele. To obtain similar sequencing depth per allele for the F0 and F1 mice which only contain half the UMIs per genotype, F0 libraries were downsampled to 50% of reads. Furthermore, 71 mitochondrial and X-chromosomal genes which only showed reads mapping to the reference (maternal) allele were discarded and 7 genes with a strong paternal bias in the F1 but no bias in the F0 which likely represent mapping errors. Around 25.82% of reads were assignable to either the maternal or the paternal haplotype across samples. Depending on the specific analysis, we allele-specific expression was quantified either as the allelic ratio B6 / (B6 + CAST) or as the (log2) allelic fold-change, log2 (B6 / CAST).

Allele-specific analysis of ATAC-sequencing data. ATAC-Seq reads were trimmed using trimmomatic (v0.38) and mapped to the mm10 genome build using bowtie2 (v2.3.5.1) [Bolger et al., 2014, Langmead and Salzberg, 2012]. Peak calling was performed using macs2 (v2.1.2.1) with the –nomodel –extsize 200 –shift –100 –call-summits parame-

ters [Zhang et al., 2008]. Finally the find_intersecting_snps.py script from the WASP package (v0.3.4) was used to annotate reads with SNPs between the B6 and CAST genomes. Then, a modified version of the count_allelic.py script from the scDALI software (https://github.com/tohein/scai_utils) was used to count reads with allele-specific mapping within each peak which were then used for further analysis. A consensus peak set between the two replicates was defined using the mergeByOverlaps function (GenomicRanges, v1.44.0, with the argument minoverlap = 0.9). Then, the genomic distribution was quantified using the annotatePeak function from the ChIPseeker package (v1.28.3) [Wang et al., 2022]. Each peak was then annotated with its closest gene if there was one within 20kb distance and only peaks with at least an average of 50 allele-specific reads were retained. Allelic imbalance (AI) in chromatin accessibility was then quantified as the average read count ratio B6 / (B6 + CAST) across both replicates. Then, an over-representation analysis was performed to relate allelic imbalance in chromatin accessibility to allelic imbalance in gene expression. To this end, the effect size of allelic imbalance for a gene or peak was defined as $d = |AI - 0.5|$ and the fraction of genes with an associated ATAC-Seq peak with $d > 0.1$ was computed. Then, a random distribution was obtained by shuffling the observed AI estimates randomly across peaks. Finally, to investigate the association between dynamic AI and chromatin accessibility, the set of genes with dynamic AI was considered and ranked by the differential in AI between spermatocytes and spermatids and the strength of AI in associated ATAC-Seq peaks was quantified.

Joint clustering of dynamic *cis*, *trans* and differential expression effects. To derive joint patterns of total expression, genetic effects and differential expression trajectories, first, bin-wise estimates of *cis*-effects ($|AI - 0.5|$ in F1), differential expression ($|allelicratio - 0.5|$) and *trans*-effects ($|AI\_F1 - AI\_F0|$) were computed. This was done for all genes with detected dynamic effects (log BF >10). Then, these trajectories were smoothed by averaging across 5 bins. Also, the average across all genes was computed as a measure of the total dynamic effect estimate. Each individual genes was then scaled to the 0 - 1 range and subjected jointly to hierarchical clustering (potentially including *cis*, *trans* or differential expression-effects for the same gene). Then average cluster trajectories per effects group (*cis*, *trans*, differential expression) and cluster were computed.

6.1.2 The landscape of escape from X-inactivation in immune cells

All experiments to generate scRNA-Seq and scATAC-Seq data were performed by Stefania Del Prete.

Mouse strains. The same strains were used in this project as in **Chapter 2**. Additionally, SPRET/EiJ mice (Strain number #001146) were used to generate F1 hybrids which are included in the dataset but were not used for any analysis shown in this thesis. Young and old male and female mice were used at around 2 months and 24 months of age.

Single-cell RNA-Sequencing of mouse splenocytes. Dissected spleen fragments were passed through a 40-μm cell strainer (Greiner 542040). The cell suspension was strained twice using DMEM (Gibco 41966029). Strainers were then washed with additional DMEM supplemented with 5% FBS and cells were pelleted by centrifugation at 300g for 4 minutes at 4 degrees celsius. Cell pellets were then resuspende for 1 minute in ice in 300 μl of ACK lysis buffer (Lonza BP10-548E). Cells were then washed with 1mL of DMEM with 5% FBS and pelleted.

Cells were counted using 0.4% trypan blue stain (Invitrogen) and a Countess II automated cell counter (Invitrogen). 1×10^7 cells were subjected to dead cell depletion using Dead Cell Removal Kit (Miltenyi Biotec 130-090-101) by the manufacturer's instructions. Dead cells were depleted using MS MACS Columns (Miltenyi Biotec 130-042-201) on a MiniMACS Separator (Miltenyi Biotec 130-042-102). The flowthrough containing live cells was resuspended in 1X PBS with 0.04% bovine serum albumin (BSA, Miltenyi Biotec 130-091-376) at 1,000 cells per μl. Single-cell RNA was performed with 20.000 cells per sample analogously to the previous chapter. Libraries were paired-end sequenced on an Illumina NextSeq2000 with read lengths of 30bp (read 1) and 199bp (read2). The read lengths were increased to improve allelic coverage.

Single-cell ATAC-Sequencing of mouse splenocytes. The single-cell solution generated for single-cell RNA-Sequencing was processed for single-cell ATAC-Sequencing in parallel. For nuclei isolation, cells were incubated on ice for 3 minutes in 100μl cell lysis buffer (10mM Tris-HCl pH 7.4 (AM9850G, LIFE Technologies), 10mM NaCl (AM9760G, ThermoFisher), 3mM MgCl (AM9530G, ThermoFisher), 0.1% NP-40 (85124, LIFE Technologies), 0.1% Tween-20 (P1379, Sigma), 0.01% Digitonin (BN2006, LIFE Technologies)). 1ml wash buffer (10mM Tris-HCl pH 7.5, 10mM NaCl, 3mM MgCl, 0.1% Tween-20) was used to rinse the nuclei which were then centrifugated at 500g for 5 minutes at 4 degress celsius. scATAC libraries were generated from 20000 cells using the Chromium™ Next GEM Single Cell ATAC Reagent Kits v1.1 (10X Genomics® 1000175) and the Chromium™ Next GEM Chip H Single Cell Kit (10X Genomics® 1000161).

Expression quantifications of 10x scRNA-Seq and 10x scATAC-Seq data for different mouse strains. Total and allele-specific expression quantifications for scRNA-Seq data were generated analogously to **Chapter 2**, with the main difference that the most recent CellRanger version (v7.0) was used, which includes intronic read counts for quantification. For scATAC-Seq data, a similar approach was used. The Cellranger-atac software was used to create N-masked references and to align samples (v2.1.0). For the allele-specific quantifications, read pairs were annotated using heterozygous variants using the WASP software as for the scRNA-Seq data. The resulting annotated alignment file was split by assigned haplotypes and processed into fragment files using the sinto package (sinto fragments, v0.9.0, https://github.com/timoast/sinto).

Low-level analysis of scRNA- and scATAC-Seq data. The scRNA-Seq data was processed similarly to the testis dataset, as described in the methods. The scATAC-Seq data was analyzed using the ArchR-package [Granja et al., 2021]. The cellranger-atac output was converted into arrow files and aggregated into an ArchR-project. Cells were defined as barcodes with more than 5000 and less than 1000000 unique fragments and a TSS enrichment of at least 13. Next, counts were tiled across the genome in 500bp bins and reads per bin were counted (*addTileMatrix*). The resulting matrix was subjected to latent semantic indexing (*addIterativeLSI*) and Harmony was used to correct for sample-specific effects (*addHarmony*) [Korsunsky et al., 2019]. Samples were then subjected to Louvain clustering on the Harmony output (*addClusters*) and doublets were assigned (*addDoubletScores*), leading to exclusion of four likely doublet clusters. To annotate cell types, gene-level scores were computed (*addGeneScoreMatrix*) and used to assign cell types from the annotated scRNA-Seq data using the SingleR package. Similarly to the scRNA-Seq datasets, the cells were split into B-cells, T-cells and others and fine-mapped cell types were assigned to match the scATAC-Seq and scRNA-Seq annotaions.

Allele-specific analysis of scATAC-Seq data. The haplotype-specific fragment files were aggregated into an ArchR-project by treating both haplotypes as individual samples and without excluding possible cells. The resulting ArchR-project was subset on the samples identified using the total analysis and active and inactive X-chromosomes were annotated as described in the results.

6.1.3 *Xist* modulates the expression of escapees

All experiments were performed by Antonia Hauth and Dr. Agnese Loda. The nextflow-pipeline used to quantify gene expression in RNA-Seq samples was written by Dr. Yuvia Perez-Rico and for some of the samples, preprocessing was performed by Antonia Hauth.

Cell culture and treatments All experiments were performed in neural precursor cell lines grown in N2B27 medium (1:1 Neurobasal / DMEM-F12, N2 + B27 supplements) with FGF and EGF supplemented at 10ng/ml on gelatin-coated flasks. *Xist* overexpression was induced by addition of Doxycyclin at 1μg/ml. SPEN degradation in was induced by addition of auxin at 500μM. All cell lines were checked for the presence of two X-chromosomes in each experiment using DNA-FISH.

Processing of RNA-Sequencing datasets. Allele-specific processing of RNA-Seq data was performed using a nextflow-pipline (https://github.com/yuviaapr/allele-specific_RNA-seq). The implemented steps largely mirror the analysis of allele-specific RNA-Sequencing data in **Chapers 2 / 3**. Raw reads were trimmed using trim_galore (v0.6.6, https://github.com/FelixKrueger/TrimGalore). Trimmed reads were aligned to an N-masked genome using STAR (v2.5.3a). Reads were assigned to parental haplotypes using SNP-split (v0.5.0) and counted using featureCounts (v2.0.1). For all cell line data, cells were of a B6xCAST background. The embryos were of a B6xJF1 background, and SNP-sets were chosen accordingly.

Bibliography

[Adrianse et al., 2018] Adrianse, R. L., Smith, K., Gatbonton-Schwager, T., Sripathy, S. P., Lao, U., Foss, E. J., Boers, R. G., Boers, J. B., Gribnau, J., and Bedalov, A. (2018). Perturbed maintenance of transcriptional repression on the inactive x-chromosome in the mouse brain after xist deletion. *Epigenetics Chromatin*, 11(1):50.

[Aibar et al., 2017] Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., and Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, 14(11):1083–1086.

[Albert and Kruglyak, 2015] Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, 16(4):197–212.

[Andergassen et al., 2017] Andergassen, D., Dotter, C. P., Wenzel, D., Sigl, V., Bammer, P. C., Muckenhuber, M., Mayer, D., Kulinski, T. M., Theussl, H.-C., Penninger, J. M., Bock, C., Barlow, D. P., Pauler, F. M., and Hudson, Q. J. (2017). Mapping the mouse allelome reveals tissue-specific regulation of allelic expression. *Elife*, 6.

[Arendt et al., 2016] Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., and Wagner, G. P. (2016). The origin and evolution of cell types. *Nat. Rev. Genet.*, 17(12):744–757.

[Argelaguet et al., 2019] Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., Dean, W., Nichols, J., Stegle, O., Marioni, J. C., and Reik, W. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491.

[Argelaguet et al., 2021] Argelaguet, R., Cuomo, A. S. E., Stegle, O., and Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.*, 39(10):1202–1215.

[Argelaguet et al., 2018] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.*, 14(6):e8124.

[Ay and Arnosti, 2011] Ay, A. and Arnosti, D. N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit. Rev. Biochem. Mol. Biol.*, 46(2):137–151.

[Badia-i Mompel et al., 2023] Badia-i Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., and Saez-Rodriguez, J. (2023). Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.*, pages 1–16.

[Balaton et al., 2015] Balaton, B. P., Cotton, A. M., and Brown, C. J. (2015). Derivation of consensus inactivation status for x-linked genes from genome-wide studies. *Biol. Sex Differ.*, 6:35.

[Balaton et al., 2021] Balaton, B. P., Fornes, O., Wasserman, W. W., and Brown, C. J. (2021). Cross-species examination of x-chromosome inactivation highlights domains of escape from silencing. *Epigenetics Chromatin*, 14(1):12.

[Banerji et al., 1983] Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3):729–740.

[Baran et al., 2015] Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E. K., Rivas, M. A., Pirinen, M., Gutierrez-Arcelus, M., Smith, K. S., Kukurba, K. R., Zhang, R., Eng, C., Torgerson, D. G., Urbanek, C., the GTEx Consortium, Li, J. B., Rodriguez-Santana, J. R., Burchard, E. G., Seibold, M. A., MacArthur, D. G., Montgomery, S. B., Zaitlen, N. A., and Lappalainen, T. (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.*, 25(7):927–936.

[Barros de Andrade E Sousa et al., 2019] Barros de Andrade E Sousa, L., Jonkers, I., Syx, L., Dunkel, I., Chaumeil, J., Picard, C., Foret, B., Chen, C.-J., Lis, J. T., Heard, E., Schulz, E. G., and Marsico, A. (2019). Kinetics of xist-induced gene silencing can be predicted from combinations of epigenetic and genomic features. *Genome Res.*, 29(7):1087–1099.

[Bartel, 2004] Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297.

[Bartman et al., 2016] Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A., and Blobel, G. A. (2016). Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol. Cell*, 62(2):237–247.

[Bartosovic et al., 2021] Bartosovic, M., Kabbe, M., and Castelo-Branco, G. (2021). Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.*, 39(7):825–835.

[Baysoy et al., 2023] Baysoy, A., Bai, Z., Satija, R., and Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.*, 24(10):695–713.

[Behm et al., 2023] Behm, M., Centurión, P. B., Penso-Dolfin, L., Bataller, F. J. B., Hirschmüller, N., Delaunay, S., Koch, M.-L., Del Prete, S., Sohn, D., Reifenberg, C., Schopp, M., Lammers, F., Solé-Boldo, L., Dutton, J., Begall, S., Khaled, W. T., St. J. Smith, E., Odom, D. T., Frye, M., and Goncalves, A. (2023). An interactive cellular ecosystem blocks epithelial transformation in naked mole-rat.

[Bell et al., 2011] Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, 12(8):554–564.

[Ben-David et al., 2021] Ben-David, E., Boocock, J., Guo, L., Zdraljevic, S., Bloom, J. S., and Kruglyak, L. (2021). Whole-organism eQTL mapping at cellular resolution with single-cell sequencing. *Elife*, 10.

[Benayoun et al., 2019] Benayoun, B. A., Pollina, E. A., Singh, P. P., Mahmoudi, S., Harel, I., Casey, K. M., Dulken, B. W., Kundaje, A., and Brunet, A. (2019). Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Res.*, 29(4):697–709.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300.

[Benko et al., 2009] Benko, S., Fantes, J. A., Amiel, J., Kleinjan, D.-J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C. T., McBride, D., Golzio, C., Fisher, M., Perry, P., Abadie, V., Ayuso, C., Holder-Espinasse, M., Kilpatrick, N., Lees, M. M., Picard, A., Temple, I. K., Thomas, P., Vazquez, M.-P., Vekemans, M., Crollius, H. R., Hastie, N. D., Munnich, A., Etchevers, H. C., Pelet, A., Farlie, P. G., FitzPatrick, D. R., and Lyonnet, S. (2009). Highly conserved non-coding elements on either side of SOX9 associated with pierre robin sequence. *Nat. Genet.*, 41(3):359–364.

[Berger, 2002] Berger, S. L. (2002). Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.*, 12(2):142–148.

[Berletch et al., 2015] Berletch, J. B., Ma, W., Yang, F., Shendure, J., Noble, W. S., Disteche, C. M., and Deng, X. (2015). Escape from X inactivation varies in mouse tissues. *PLoS Genet.*, 11(3):e1005079.

[Berletch et al., 2010] Berletch, J. B., Yang, F., and Disteche, C. M. (2010). Escape from X inactivation in mice and humans. *Genome Biol.*, 11(6):213.

[Bhutani et al., 2021] Bhutani, K., Stansifer, K., Ticau, S., Bojic, L., Villani, A.-C., Slisz, J., Cremers, C. M., Roy, C., Donovan, J., Fiske, B., and Friedman, R. C. (2021). Widespread haploid-biased gene expression enables sperm-level natural selection. *Science*, 371(6533).

[Bienko, 2023] Bienko, M. (2023). How Hi-C ignited the era of 3D genome biology. *Nat. Rev. Genet.*, 24(7):418–418.

[Billi et al., 2019] Billi, A. C., Kahlenberg, J. M., and Gudjonsson, J. E. (2019). Sex bias in autoimmunity. *Curr. Opin. Rheumatol.*, 31(1):53–61.

[BinTayyash et al., 2021] BinTayyash, N., Georgaka, S., John, S. T., Ahmed, S., Boukouvalas, A., Hensman, J., and Rattray, M. (2021). Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. *Bioinformatics*, 37(21):3788–3795.

[Blackledge and Klose, 2021] Blackledge, N. P. and Klose, R. J. (2021). The molecular principles of gene regulation by polycomb repressive complexes. *Nat. Rev. Mol. Cell Biol.*, 22(12):815–833.

[Bolger et al., 2014] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

[Bonder et al., 2021] Bonder, M. J., Smail, C., Gloudemans, M. J., Frésard, L., Jakubosky, D., D'Antonio, M., Li, X., Ferraro, N. M., Carcamo-Orive, I., Mirauta, B., Seaton, D. D., Cai, N., Vakili, D., Horta, D., Zhao, C., Zastrow, D. B., Bonner, D. E., Wheeler, M. T., Kilpinen, H., Knowles, J. W., Smith, E. N., Frazer, K. A., Montgomery, S. B., and Stegle, O. (2021). Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nat. Genet.*, 53(3):313–321.

[Borensztein et al., 2017] Borensztein, M., Syx, L., Ancelin, K., Diabangouaya, P., Picard, C., Liu, T., Liang, J.-B., Vassilev, I., Galupa, R., Servant, N., Barillot, E., Surani, A., Chen, C.-J., and Heard, E. (2017). Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. *Nat. Struct. Mol. Biol.*, 24(3):226–233.

[Borm et al., 2022] Borm, L. E., Mossi Albiach, A., Mannens, C. C. A., Janusauskas, J., Özgün, C., Fernández-García, D., Hodge, R., Castillo, F., Hedin, C. R. H., Villablanca, E. J., Uhlén, P., Lein, E. S., Codeluppi, S., and Linnarsson, S. (2022). Scalable in situ single-cell profiling by electrophoretic capture of mRNA using EEL FISH. *Nat. Biotechnol.*, 41(2):222–231.

[Bozukova et al., 2022] Bozukova, M., Nikopoulou, C., Kleinenkuhnen, N., Grbavac, D., Goetsch, K., and Tessarz, P. (2022). Aging is associated with increased chromatin accessibility and reduced polymerase pausing in liver. *Mol. Syst. Biol.*, 18(9):e11002.

[Brawand et al., 2011] Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

[Brown and Rastan, 1988] Brown, S. and Rastan, S. (1988). Age-related reactivation of an x-linked gene close to the inactivation centre in the mouse. *Genet. Res.*, 52(2):151–154.

[Browning and Browning, 2011] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, 12(10):703–714.

[Buenrostro et al., 2015] Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

[Cabezas-Wallscheid et al., 2014] Cabezas-Wallscheid, N., Klimmeck, D., Hansson, J., Lipka, D. B., Reyes, A., Wang, Q., Weichenhan, D., Lier, A., von Paleske, L., Renders, S., Wünsche, P., Zeisberger, P., Brocks, D., Gu, L., Herrmann, C., Haas, S., Essers, M. A. G., Brors, B., Eils, R., Huber, W., Milsom, M. D., Plass, C., Krijgsveld, J., and Trumpp, A. (2014). Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell*, 15(4):507–522.

[Calaway et al., 2013] Calaway, J. D., Lenarcic, A. B., Didion, J. P., Wang, J. R., Searle, J. B., McMillan, L., Valdar, W., and Pardo-Manuel de Villena, F. (2013). Genetic architecture of skewed X inactivation in the laboratory mouse. *PLoS Genet.*, 9(10):e1003853.

[Calo and Wysocka, 2013] Calo, E. and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, 49(5):825–837.

[Cao et al., 2017] Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667.

[Carpenter et al., 2014] Carpenter, S., Ricci, E. P., Mercier, B. C., Moore, M. J., and Fitzgerald, K. A. (2014). Post-transcriptional regulation of gene expression in innate immunity. *Nat. Rev. Immunol.*, 14(6):361–376.

[Carrel and Brown, 2017] Carrel, L. and Brown, C. J. (2017). When the lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 372(1733).

[Carrel and Willard, 2005] Carrel, L. and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in x-linked gene expression in females. *Nature*, 434(7031):400–404.

[Castel et al., 2015] Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biol.*, 16:195.

[Chatterjee et al., 2016] Chatterjee, S., Kapoor, A., Akiyama, J. A., Auer, D. R., Lee, D., Gabriel, S., Berrios, C., Pennacchio, L. A., and Chakravarti, A. (2016). Enhancer variants synergistically drive dysfunction of a gene regulatory network in hirschsprung disease. *Cell*, 167(2):355–368.e10.

[Chen et al., 2015] Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090.

[Chen et al., 2022] Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., and Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608(7924):733–740.

[Chen et al., 2016] Chen, X., Shen, Y., Draper, W., Buenrostro, J. D., Litzenburger, U., Cho, S. W., Satpathy, A. T., Carter, A. C., Ghosh, R. P., East-Seletsky, A., Doudna, J. A., Greenleaf, W. J., Liphardt, J. T., and Chang, H. Y. (2016). ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat. Methods*, 13(12):1013–1020.

[Cheng et al., 2023] Cheng, M. I., Li, J. H., Riggan, L., Chen, B., Tafti, R. Y., Chin, S., Ma, F., Pellegrini, M., Hrncir, H., Arnold, A. P., O'Sullivan, T. E., and Su, M. A. (2023). The x-linked epigenetic regulator UTX controls NK cell-intrinsic sex differences. *Nat. Immunol.*, 24(5):780–791.

[Chess, 2005] Chess, A. (2005). Monoallelic expression of protocadherin genes. *Nat. Genet.*, 37(2):120–121.

[Choi et al., 2019] Choi, K., Raghupathy, N., and Churchill, G. A. (2019). A bayesian mixture model for the analysis of allelic expression in single cells. *Nat. Commun.*, 10(1):1–11.

[Chubb et al., 2006] Chubb, J. R., Trcek, T., Shenoy, S. M., and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Curr. Biol.*, 16(10):1018–1025.

[Chung et al., 2019] Chung, H. Y., Kim, D. H., Lee, E. K., Chung, K. W., Chung, S., Lee, B., Seo, A. Y., Chung, J. H., Jung, Y. S., Im, E., Lee, J., Kim, N. D., Choi, Y. J., Im, D. S., and Yu, B. P. (2019). Redefining chronic inflammation in aging and Age-Related diseases: Proposal of the senoinflammation concept. *Aging Dis.*, 10(2):367–382.

[Cleary and Seoighe, 2021] Cleary, S. and Seoighe, C. (2021). Perspectives on Allele-Specific expression. *Annu Rev Biomed Data Sci*, 4:101–122.

[Cloutier et al., 2022] Cloutier, M., Kumar, S., Buttigieg, E., Keller, L., Lee, B., Williams, A., Mojica-Perez, S., Erliandri, I., Rocha, A. M. D., Cadigan, K., Smith, G. D., and Kalantry, S. (2022). Preventing erosion of x-chromosome inactivation in human embryonic stem cells. *Nat. Commun.*, 13(1):1–18.

[Collombet et al., 2020] Collombet, S., Ranisavljevic, N., Nagano, T., Varnai, C., Shisode, T., Leung, W., Piolot, T., Galupa, R., Borensztein, M., Servant, N., Fraser, P., Ancelin, K., and Heard, E. (2020). Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature*, 580(7801):142–146.

[Corces et al., 2017] Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., Montine, T. J., Greenleaf, W. J., and Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*, 14(10):959–962.

[Crane et al., 2017] Crane, G. M., Jeffery, E., and Morrison, S. J. (2017). Adult haematopoietic stem cell niches. *Nat. Rev. Immunol.*, 17(9):573–590.

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

[Csankovszki et al., 2001] Csankovszki, G., Nagy, A., and Jaenisch, R. (2001). Synergism of xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J. Cell Biol.*, 153(4):773–784.

[Cuomo et al., 2021] Cuomo, A. S. E., Alvari, G., Azodi, C. B., single-cell eQTLGen consortium, McCarthy, D. J., and Bonder, M. J. (2021). Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.*, 22(1):188.

[Cuomo et al., 2022] Cuomo, A. S. E., Heinen, T., Vagiaki, D., Horta, D., Marioni, J. C., and Stegle, O. (2022). CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Mol. Syst. Biol.*, 18(8):e10663.

[Cuomo et al., 2020] Cuomo, A. S. E., Seaton, D. D., McCarthy, D. J., Martinez, I., Bonder, M. J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., Knights, A., Natarajan, K. N., HipSci Consortium, Vallier, L., Marioni, J. C., Chhatriwala, M., and Stegle, O. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.*, 11(1):810.

[Cusanovich et al., 2018] Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee,

C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., and Shendure, J. (2018). A Single-Cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324.e18.

[Das et al., 2021] Das, S., Vera, M., Gandin, V., Singer, R. H., and Tutucci, E. (2021). Intracellular mRNA transport and localized translation. *Nat. Rev. Mol. Cell Biol.*, 22(7):483–504.

[Deng et al., 2014a] Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014a). Single-Cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196.

[Deng et al., 2014b] Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014b). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196.

[Deng et al., 2022] Deng, Y., Bartosovic, M., Ma, S., Zhang, D., Kukanja, P., Xiao, Y., Su, G., Liu, Y., Qin, X., Rosoklija, G. B., Dwork, A. J., Mann, J. J., Xu, M. L., Halene, S., Craft, J. E., Leong, K. W., Boldrini, M., Castelo-Branco, G., and Fan, R. (2022). Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature*, 609(7926):375–383.

[Deshmukh et al., 2018] Deshmukh, S., Ponnaluri, V. C., Dai, N., Pradhan, S., and Deobagkar, D. (2018). Levels of DNA cytosine methylation in the drosophila genome. *PeerJ*, 6:e5119.

[Dhakal et al., 2022] Dhakal, S., Chaulagain, S., and Klein, S. L. (2022). Sex biases in infectious diseases research. *J. Exp. Med.*, 219(6).

[Doane and Elemento, 2017] Doane, A. S. and Elemento, O. (2017). Regulatory elements in molecular networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 9(3).

[Donovan et al., 2019] Donovan, B. T., Huynh, A., Ball, D. A., Patel, H. P., Poirier, M. G., Larson, D. R., Ferguson, M. L., and Lenstra, T. L. (2019). Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *EMBO J.*, 38(12).

[Dossin et al., 2020] Dossin, F., Pinheiro, I., Żylicz, J. J., Roensch, J., Collombet, S., Le Saux, A., Chelmicki, T., Attia, M., Kapoor, V., Zhan, Y., Dingli, F., Loew, D., Mercher, T., Dekker, J., and Heard, E. (2020). SPEN integrates transcriptional and epigenetic control of x-inactivation. *Nature*, 578(7795):455–460.

[Duboule, 2007] Duboule, D. (2007). The rise and fall of hox gene clusters. *Development*, 134(14):2549–2560.

[Elorbany et al., 2022] Elorbany, R., Popp, J. M., Rhodes, K., Strober, B. J., Barr, K., Qi, G., Gilad, Y., and Battle, A. (2022). Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. *PLoS Genet.*, 18(1):e1009666.

[ENCODE Project Consortium, 2012] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

[Engreitz et al., 2013] Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K., and Guttman, M. (2013). The xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, 341(6147):1237973.

[Ernst et al., 2019] Ernst, C., Eling, N., Martinez-Jimenez, C. P., Marioni, J. C., and Odom, D. T. (2019). Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat. Commun.*, 10(1):1251.

[Fan et al., 2021] Fan, J., Wang, X., Xiao, R., and Li, M. (2021). Detecting cell-type-specific allelic expression imbalance by integrative analysis of bulk and single-cell RNA sequencing data. *PLoS Genet.*, 17(3):e1009080.

[Fan et al., 2005] Fan, T., Hagan, J. P., Kozlov, S. V., Stewart, C. L., and Muegge, K. (2005). Lsh controls silencing of the imprinted cdkn1c gene. *Development*, 132(4):635–644.

[Fang et al., 2019] Fang, H., Disteche, C. M., and Berletch, J. B. (2019). X inactivation and escape: Epigenetic and structural features. *Front Cell Dev Biol*, 7:219.

[Fang et al., 2023] Fang, H., Tronco, A. R., Bonora, G., Nguyen, T., Thakur, J., Berletch, J. B., Filippova, G. N., Henikoff, S., Shendure, J., Noble, W. S., Disteche, C. M., and Deng, X. (2023). CTCF-mediated insulation and chromatin environment modulate car5b escape from X inactivation. *bioRxiv*.

[Ferguson-Smith, 2011] Ferguson-Smith, A. C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.*, 12(8):565–575.

[Ferguson-Smith and Bourc'his, 2018] Ferguson-Smith, A. C. and Bourc'his, D. (2018). The discovery and importance of genomic imprinting. *Elife*, 7.

[Ferrón et al., 2011] Ferrón, S. R., Charalambous, M., Radford, E., McEwen, K., Wildner, H., Hind, E., Morante-Redolat, J. M., Laborda, J., Guillemot, F., Bauer, S. R., Fariñas, I., and Ferguson-Smith, A. C. (2011). Postnatal loss of dlk1 imprinting in stem cells and niche astrocytes regulates neurogenesis. *Nature*, 475(7356):381–385.

[Findley et al., 2021] Findley, A. S., Monziani, A., Richards, A. L., Rhodes, K., Ward, M. C., Kalita, C. A., Alazizi, A., Pazokitoroudi, A., Sankararaman, S., Wen, X., Lanfear, D. E., Pique-Regi, R., Gilad, Y., and Luca, F. (2021). Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. *Elife*, 10.

[Fleck et al., 2023] Fleck, J. S., Camp, J. G., and Treutlein, B. (2023). What is a cell type? *Science*, 381(6659):733–734.

[Floc'hlay et al., 2021] Floc'hlay, S., Wong, E. S., Zhao, B., Viales, R. R., Thomas-Chollier, M., Thieffry, D., Garfield, D. A., and Furlong, E. E. M. (2021). Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. *Genome Res.*, 31(2):211–224.

[Franceschi et al., 2018] Franceschi, C., Garagnani, P., Parini, P., Giuliani, C., and Santoro, A. (2018). Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nat. Rev. Endocrinol.*, 14(10):576–590.

[Francesconi and Lehner, 2014] Francesconi, M. and Lehner, B. (2014). The effects of genetic variation on gene expression dynamics during development. *Nature*, 505(7482):208–211.

[Frasca and Blomberg, 2009] Frasca, D. and Blomberg, B. B. (2009). Effects of aging on B cell function. *Curr. Opin. Immunol.*, 21(4):425–430.

[Frieda et al., 2017] Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B., and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111.

[Fukaya et al., 2016] Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer control of transcriptional bursting. *Cell*, 166(2):358–368.

[Furlan and Galupa, 2022] Furlan, G. and Galupa, R. (2022). Mechanisms of choice in X-Chromosome inactivation. *Cells*, 11(3).

[Garieri et al., 2018] Garieri, M., Stamoulis, G., Blanc, X., Falconnet, E., Ribaux, P., Borel, C., Santoni, F., and Antonarakis, S. E. (2018). Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. *Proceedings of the National Academy of Sciences*, 115(51):13015–13020.

[Gebauer and Hentze, 2004] Gebauer, F. and Hentze, M. W. (2004). Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.*, 5(10):827–835.

[Gehart and Clevers, 2019] Gehart, H. and Clevers, H. (2019). Tales from the crypt: new insights into intestinal stem cells. *Nat. Rev. Gastroenterol. Hepatol.*, 16(1):19–34.

[Gendrel et al., 2014] Gendrel, A.-V., Attia, M., Chen, C.-J., Diabangouaya, P., Servant, N., Barillot, E., and Heard, E. (2014). Developmental dynamics and disease potential of random monoallelic gene expression. *Dev. Cell*, 28(4):366–380.

[Gendrel et al., 2016] Gendrel, A.-V., Marion-Poll, L., Katoh, K., and Heard, E. (2016). Random monoallelic expression of genes on autosomes: Parallels with x-chromosome inactivation. *Semin. Cell Dev. Biol.*, 56:100–110.

[Gimelbrant et al., 2007] Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. *Science*, 318(5853):1136–1140.

[Ginart et al., 2016] Ginart, P., Kalish, J. M., Jiang, C. L., Yu, A. C., Bartolomei, M. S., and Raj, A. (2016). Visualizing allele-specific expression in single cells reveals epigenetic mosaicism in an H19 loss-of-imprinting mutant. *Genes Dev.*, 30(5):567–578.

[Giorgetti et al., 2016] Giorgetti, L., Lajoie, B. R., Carter, A. C., Attia, M., Zhan, Y., Xu, J., Chen, C. J., Kaplan, N., Chang, H. Y., Heard, E., and Dekker, J. (2016). Structural organization of the inactive X chromosome in the mouse. *Nature*, 535(7613):575–579.

[Glinos et al., 2022] Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K. L., Garimella, K., Bowers, T., Costello, M., Ardlie, K., Jian, R., Tucker, N. R., Ellinor, P. T., Harrington, E. D., Tang, H., Snyder, M., Juul, S., Mohammadi, P., MacArthur, D. G., Lappalainen, T., and Cummings, B. B. (2022). Transcriptome variation in human tissues revealed by long-read sequencing. *Nature*, 608(7922):353–359.

[Gokhman et al., 2021] Gokhman, D., Agoglia, R. M., Kinnebrew, M., Gordon, W., Sun, D., Bajpai, V. K., Naqvi, S., Chen, C., Chan, A., Chen, C., Petrov, D. A., Ahituv, N., Zhang, H., Mishina, Y., Wysocka, J., Rohatgi, R., and Fraser, H. B. (2021). Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat. Genet.*, 53(4):467–476.

[Goncalves et al., 2012] Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D. T., and Marioni, J. C. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.*, 22(12):2376–2384.

[Gorin and Pachter, 2022] Gorin, G. and Pachter, L. (2022). Modeling bursty transcription and splicing with the chemical master equation. *Biophys. J.*, 121(6):1056–1069.

[Goronzy and Weyand, 2012] Goronzy, J. J. and Weyand, C. M. (2012). Immune aging and autoimmunity. *Cell. Mol. Life Sci.*, 69(10):1615–1623.

[Goronzy and Weyand, 2019] Goronzy, J. J. and Weyand, C. M. (2019). Mechanisms underlying T cell ageing. *Nat. Rev. Immunol.*, 19(9):573–583.

[Goyal et al., 2023] Goyal, Y., Busch, G. T., Pillai, M., Li, J., Boe, R. H., Grody, E. I., Chelvanambi, M., Dardani, I. P., Emert, B., Bodkin, N., Braun, J., Fingerman, D., Kaur, A., Jain, N., Ravindran, P. T., Mellis, I. A., Kiani, K., Alicea, G. M., Fane, M. E., Ahmed, S. S., Li, H., Chen, Y., Chai, C., Kaster, J., Witt, R. G., Lazcano, R., Ingram, D. R., Johnson, S. B., Wani, K., Dunagin, M. C., Lazar, A. J., Weeraratna, A. T., Wargo, J. A., Herlyn, M., and Raj, A. (2023). Diverse clonal fates emerge upon drug treatment of homogeneous cancer cells. *Nature*, 620(7974):651–659.

[Granja et al., 2021] Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., and Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, 53(3):403–411.

[Grieshaber-Bouyer et al., 2021] Grieshaber-Bouyer, R., Radtke, F. A., Cunin, P., Stifano, G., Levescot, A., Vijaykumar, B., Nelson-Maney, N., Blaustein, R. B., Monach, P. A., and Nigrovic, P. A. (2021). The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments. *Nat. Commun.*, 12(1):1–21.

[Grigoryan et al., 2021] Grigoryan, A., Pospiech, J., Krämer, S., Lipka, D., Liehr, T., Geiger, H., Kimura, H., Mulaw, M. A., and Florian, M. C. (2021). Attrition of X chromosome inactivation in aged hematopoietic stem cells. *Stem Cell Reports*, 16(4):708–716.

[Grima and Esmenjaud, 2023] Grima, R. and Esmenjaud, P.-M. (2023). Quantifying and correcting bias in transcriptional parameter inference from single-cell data.

[GTEx Consortium et al., 2017] GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Battle, A., Brown, C. D., Engelhardt, B. E., and Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

[Guo et al., 2005] Guo, L., Hu-Li, J., and Paul, W. E. (2005). Probabilistic regulation in TH2 cells accounts for monoallelic expression of IL-4 and IL-13. *Immunity*, 23(1):89–99.

[Gutierrez-Arcelus et al., 2020] Gutierrez-Arcelus, M., Baglaenko, Y., Arora, J., Hannes, S., Luo, Y., Amariuta, T., Teslovich, N., Rao, D. A., Ermann, J., Jonsson, A. H., Navarrete, C., Rich, S. S., Taylor, K. D., Rotter, J. I., Gregersen, P. K., Esko, T., Brenner, M. B., and Raychaudhuri, S. (2020). Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat. Genet.*, 52(3):247–253.

[Haberle and Stark, 2018] Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.*, 19(10):621–637.

[Hafner et al., 2021] Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., and Zavolan, M. (2021). CLIP and complementary methods. *Nature Reviews Methods Primers*, 1(1):1–23.

[Hagemann-Jensen et al., 2020] Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J. M., Faridani, O. R., and Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using smart-seq3. *Nat. Biotechnol.*, 38(6):708–714.

[Haghverdi et al., 2018] Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36(5):421–427.

[Halow et al., 2021] Halow, J. M., Byron, R., Hogan, M. S., Ordoñez, R., Groudine, M., Bender, M. A., Stamatoyannopoulos, J. A., and Maurano, M. T. (2021). Tissue context determines the penetrance of regulatory DNA variation. *Nat. Commun.*, 12(1):2850.

[Hanna, 2020] Hanna, C. W. (2020). Placental imprinting: Emerging mechanisms and functions. *PLoS Genet.*, 16(4):e1008709.

[Hazeldine and Lord, 2013] Hazeldine, J. and Lord, J. M. (2013). The impact of ageing on natural killer cell function and potential consequences for health in older adults. *Ageing Res. Rev.*, 12(4):1069–1078.

[Heinen et al., 2021] Heinen, T., Secchia, S., Reddington, J., Zhao, B., Furlong, E. E. M., and Stegle, O. (2021). scDALI: Modelling allelic heterogeneity of DNA accessibility in single-cells reveals context-specific genetic regulation.

[Heinen et al., 2022] Heinen, T., Secchia, S., Reddington, J. P., Zhao, B., Furlong, E. E. M., and Stegle, O. (2022). scDALI: modeling allelic heterogeneity in single cells reveals context-specific genetic regulation. *Genome Biol.*, 23(1):8.

[Heinz et al., 2015] Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.*, 16(3):144–154.

[Henning et al., 2018] Henning, A. N., Roychoudhuri, R., and Restifo, N. P. (2018). Epigenetic control of CD8+ T cell differentiation. *Nat. Rev. Immunol.*, 18(5):340–356.

[Hershey et al., 2012] Hershey, J. W. B., Sonenberg, N., and Mathews, M. B. (2012). Principles of translational control: an overview. *Cold Spring Harb. Perspect. Biol.*, 4(12).

[Herzing et al., 2002] Herzing, L. B. K., Cook, Jr, E. H., and Ledbetter, D. H. (2002). Allele-specific expression analysis by RNA-FISH demonstrates preferential maternal expression of UBE3A and imprint maintenance within 15q11- q13 duplications. *Hum. Mol. Genet.*, 11(15):1707–1718.

[Hwang et al., 2018] Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, 50(8):1–14.

[Ianevski et al., 2022] Ianevski, A., Giri, A. K., and Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.*, 13(1):1–10.

[Jachowicz et al., 2022] Jachowicz, J. W., Strehle, M., Banerjee, A. K., Blanco, M. R., Thai, J., and Guttman, M. (2022). Xist spatially amplifies SHARP/SPEN recruitment to balance chromosome-wide silencing and specificity to the X chromosome. *Nat. Struct. Mol. Biol.*, 29(3):239–249.

[Jacobson et al., 2022] Jacobson, E. C., Pandya-Jones, A., and Plath, K. (2022). A lifelong duty: how xist maintains the inactive X chromosome. *Curr. Opin. Genet. Dev.*, 75:101927.

[Janiszewski et al., 2019] Janiszewski, A., Talon, I., Chappell, J., Collombet, S., Song, J., De Geest, N., To, S. K., Bervoets, G., Marin-Bejar, O., Provenzano, C., Vanheer, L., Marine, J.-C., Rambow, F., and Pasque, V. (2019). Dynamic reversal of random X-Chromosome inactivation during iPSC reprogramming. *Genome Res.*, 29(10):1659–1672.

[Jerber et al., 2021] Jerber, J., Seaton, D. D., Cuomo, A. S. E., Kumasaka, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M. J., Mountjoy, E., Ghoussaini, M., Lancaster, M. A., HipSci Consortium, Marioni, J. C., Merkle, F. T., Gaffney, D. J., and Stegle, O. (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.*, 53(3):304–312.

[Jerkovic and Cavalli, 2021] Jerkovic, I. and Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.*, 22(8):511–528.

[Joung et al., 2023] Joung, J., Ma, S., Tay, T., Geiger-Schuller, K. R., Kirchgatterer, P. C., Verdine, V. K., Guo, B., Arias-Garcia, M. A., Allen, W. E., Singh, A., Kuksenko, O., Abudayyeh, O. O., Gootenberg, J. S., Fu, Z., Macrae, R. K., Buenrostro, J. D., Regev, A., and Zhang, F. (2023). A transcription factor atlas of directed differentiation. *Cell*, 186(1):209–229.e26.

[Keane et al., 2011] Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunção, J. A., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.

[Keenan and Allan, 2019] Keenan, C. R. and Allan, R. S. (2019). Epigenomic drivers of immune dysfunction in aging. *Aging Cell*, 18(1):e12878.

[Khodosevich et al., 2002] Khodosevich, K., Lebedev, Y., and Sverdlov, E. (2002). Endogenous retroviruses and human evolution. *Comp. Funct. Genomics*, 3(6):494–498.

[Kim and Marioni, 2013] Kim, J. K. and Marioni, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.*, 14(1):R7.

[Kim-Hellmuth et al., 2020] Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S. E., Hamel, A. R., Viñuela, A., Roberts, A. L., Mangul, S., Wen, X., Wang, G., Barbeira, A. N., Garrido-Martín, D., Nadel, B. B., Zou, Y., Bonazzola, R., Quan, J., Brown, A., Martinez-Perez, A., Soria, J. M., GTEx Consortium, Getz, G., Dermitzakis, E. T., Small, K. S., Stephens, M., Xi, H. S., Im, H. K., Guigó, R., Segrè, A. V., Stranger, B. E., Ardlie, K. G., and Lappalainen, T. (2020). Cell type-specific genetic regulation of gene expression across human tissues. *Science*, 369(6509).

[Kimmel et al., 2019] Kimmel, J. C., Penland, L., Rubinstein, N. D., Hendrickson, D. G., Kelley, D. R., and Rosenthal, A. Z. (2019). Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res.*, 29(12):2088–2103.

[Kinjo et al., 2020] Kinjo, K., Yoshida, T., Kobori, Y., Okada, H., Suzuki, E., Ogata, T., Miyado, M., and Fukami, M. (2020). Random X chromosome inactivation in patients with klinefelter syndrome. *Mol Cell Pediatr*, 7(1):1.

[Kinoshita and Smith, 2018] Kinoshita, M. and Smith, A. (2018). Pluripotency deconstructed. *Dev. Growth Differ.*, 60(1):44–52.

[Kopania et al., 2022] Kopania, E. E. K., Larson, E. L., Callahan, C., Keeble, S., and Good, J. M. (2022). Molecular evolution across mouse spermatogenesis. *Mol. Biol. Evol.*, 39(2).

[Korsunsky et al., 2019] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12):1289–1296.

[Kowalski et al., 2023] Kowalski, M. H., Wessels, H.-H., Linder, J., Choudhary, S., Hartman, A., Hao, Y., Mascio, I., Dalgarno, C., Kundaje, A., and Satija, R. (2023). CPA-Perturb-seq: Multiplexed single-cell characterization of alternative polyadenylation regulators. *bioRxivorg*.

[Kravitz et al., 2023] Kravitz, S. N., Ferris, E., Love, M. I., Thomas, A., Quinlan, A. R., and Gregg, C. (2023). Random allelic expression in the adult human body. *Cell Rep.*, 42(1):111945.

[Kumar et al., 2018] Kumar, B. V., Connors, T. J., and Farber, D. L. (2018). Human T cell development, localization, and function throughout life. *Immunity*, 48(2):202–213.

[La Manno et al., 2018] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, 560(7719):494–498.

[Lambert et al., 2018] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359.

[Lappalainen and MacArthur, 2021] Lappalainen, T. and MacArthur, D. G. (2021). From variant to function in human disease genetics. *Science*, 373(6562):1464–1468.

[Larsson et al., 2019] Larsson, A. J. M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O. R., Reinius, B., Segerstolpe, Å., Rivera, C. M., Ren, B., and Sandberg, R.

(2019). Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254.

[Latham, 1995] Latham, K. E. (1995). Stage-specific and cell type-specific aspects of genomic imprinting effects in mammals. *Differentiation*, 59(5):269–282.

[Laukoter et al., 2020] Laukoter, S., Pauler, F. M., Beattie, R., Amberg, N., Hansen, A. H., Streicher, C., Penz, T., Bock, C., and Hippenmeyer, S. (2020). Cell-Type specificity of genomic imprinting in cerebral cortex. *Neuron*, 107(6):1160–1179.e9.

[Lavin et al., 2014] Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., and Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell*, 159(6):1312–1326.

[Lawrence et al., 2016] Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral thinking: How histone modifications regulate gene expression. *Trends Genet.*, 32(1):42–56.

[Leigh-Brown et al., 2015] Leigh-Brown, S., Goncalves, A., Thybert, D., Stefflova, K., Watt, S., Flicek, P., Brazma, A., Marioni, J. C., and Odom, D. T. (2015). Regulatory divergence of transcript isoforms in a mammalian model system. *PLoS One*, 10(9):e0137367.

[Leitch et al., 2013] Leitch, H. G., McEwen, K. R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J. G., Smith, A., Surani, M. A., and Hajkova, P. (2013). Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.*, 20(3):311–316.

[Lentini et al., 2022] Lentini, A., Cheng, H., Noble, J. C., Papanicolaou, N., Coucoravas, C., Andrews, N., Deng, Q., Enge, M., and Reinius, B. (2022). Elastic dosage compensation by x-chromosome upregulation. *Nat. Commun.*, 13(1):1854.

[Lewis et al., 2019] Lewis, S. M., Williams, A., and Eisenbarth, S. C. (2019). Structure and function of the immune system in the spleen. *Sci Immunol*, 4(33).

[Li et al., 2012] Li, S. M., Valo, Z., Wang, J., Gao, H., Bowers, C. W., and Singer-Sam, J. (2012). Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. *PLoS One*, 7(2):e31751.

[Li and Sasaki, 2011] Li, Y. and Sasaki, H. (2011). Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. *Cell Res.*, 21(3):466–473.

[Li et al., 2020] Li, Y., Weng, Y., Bai, D., Jia, Y., Liu, Y., Zhang, Y., Kou, X., Zhao, Y., Ruan, J., Chen, J., Yin, J., Wang, H., Teng, X., Wang, Z., Liu, W., and Gao, S. (2020). Precise allele-specific genome editing by spatiotemporal control of CRISPR-Cas9 via pronuclear transplantation. *Nat. Commun.*, 11(1):1–12.

[Liu et al., 2016] Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3):535–550.

[Loda et al., 2022] Loda, A., Collombet, S., and Heard, E. (2022). Gene regulation in time and space during x-chromosome inactivation. *Nat. Rev. Mol. Cell Biol.*, 23(4):231–249.

[Lopez et al., 2018] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058.

[López-Otín et al., 2013] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell*, 153(6):1194–1217.

[Love et al., 2014a] Love, M. I., Huber, W., and Anders, S. (2014a). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550.

[Love et al., 2014b] Love, M. I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550.

[Ludwig et al., 2019] Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A., and Sankaran, V. G. (2019). Lineage tracing in humans enabled by mitochondrial mutations and Single-Cell genomics. *Cell*, 176(6):1325–1339.e22.

[Luecken and Theis, 2019] Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, 15(6):e8746.

[Lun et al., 2016] Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Res.*, 5:2122.

[Lyko et al., 2000] Lyko, F., Ramsahoye, B. H., and Jaenisch, R. (2000). DNA methylation in drosophila melanogaster. *Nature*, 408(6812):538–540.

[Lyon, 1961] Lyon, M. F. (1961). Gene action in the x-chromosome of the mouse (mus musculus l.). *Nature*, 190:372–373.

[Ma et al., 2023] Ma, Y., Zhu, Y., Shang, L., Qiu, Y., Shen, N., Wang, J., Adam, T., Wei, W., Song, Q., Li, J., Wicha, M. S., and Luo, M. (2023). LncRNA XIST regulates breast cancer stem cells by activating proinflammatory IL-6/STAT3 signaling. *Oncogene*, 42(18):1419–1437.

[Macneil and Walhout, 2011] Macneil, L. T. and Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.*, 21(5):645–657.

[Magklara and Lomvardas, 2013] Magklara, A. and Lomvardas, S. (2013). Stochastic gene expression in mammals: lessons from olfaction. *Trends Cell Biol.*, 23(9):449–456.

[Markaki et al., 2021] Markaki, Y., Gan Chong, J., Wang, Y., Jacobson, E. C., Luong, C., Tan, S. Y. X., Jachowicz, J. W., Strehle, M., Maestrini, D., Banerjee, A. K., Mistry, B. A., Dror, I., Dossin, F., Schöneberg, J., Heard, E., Guttman, M., Chou, T., and Plath, K. (2021). Xist nucleates local protein gradients to propagate silencing across the X chromosome. *Cell*, 184(25):6174–6192.e32.

[Marks et al., 2015] Marks, H., Kerstens, H. H. D., Barakat, T. S., Splinter, E., Dirks, R. A. M., van Mierlo, G., Joshi, O., Wang, S.-Y., Babak, T., Albers, C. A., Kalkan, T., Smith, A., Jouneau, A., de Laat, W., Gribnau, J., and Stunnenberg, H. G. (2015). Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome Biol.*, 16(1):149.

[Martinez-Jimenez et al., 2017] Martinez-Jimenez, C. P., Eling, N., Chen, H.-C., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., Stojic, L., Rayner, T. F., Stubbington, M. J. T., Teichmann, S. A., de la Roche, M., Marioni, J. C., and Odom, D. T. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, 355(6332):1433–1436.

[McCarthy et al., 2017] McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186.

[Miller, 2011] Miller, J. F. A. P. (2011). The golden anniversary of the thymus. *Nat. Rev. Immunol.*, 11(7):489–495.

[Mittelbrunn and Kroemer, 2021] Mittelbrunn, M. and Kroemer, G. (2021). Hallmarks of T cell aging. *Nat. Immunol.*, 22(6):687–698.

[Miura et al., 2019] Miura, H., Takahashi, S., Poonperm, R., Tanigawa, A., Takebayashi, S.-I., and Hiratani, I. (2019). Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nat. Genet.*, 51(9):1356–1368.

[Mogilenko et al., 2022] Mogilenko, D. A., Shchukina, I., and Artyomov, M. N. (2022). Immune ageing at single-cell resolution. *Nat. Rev. Immunol.*, 22(8):484–498.

[Mohammadi et al., 2017] Mohammadi, P., Castel, S. E., Brown, A. A., and Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.*, 27(11):1872–1884.

[Monahan et al., 2019] Monahan, K., Horta, A., and Lomvardas, S. (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature*, 565(7740):448–453.

[Monk et al., 2019] Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R., and Riccio, A. (2019). Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat. Rev. Genet.*, 20(4):235–248.

[Moore et al., 2020] Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E. L., Freese, P., Gorkin, D. U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B. A., Mortazavi, A., Keller, C. A., Zhang, X.-O., Elhajjajy, S. I., Huey, J., Dickel, D. E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J. C., Rozowsky, J., Zhang, J., Chhetri, S. B., Zhang, J., Victorsen, A., White, K. P., Visel, A., Yeo, G. W., Burge, C. B., Lécuyer, E., Gilbert, D. M., Dekker, J., Rinn, J., Mendenhall, E. M., Ecker, J. R., Kellis, M., Klein, R. J., Noble, W. S., Kundaje, A., Guigó, R., Farnham, P. J., Cherry, J. M., Myers, R. M., Ren, B., Graveley, B. R., Gerstein, M. B., Pennacchio, L. A., Snyder, M. P., Bernstein, B. E., Wold, B., Hardison, R. C., Gingeras, T. R., Stamatoyannopoulos, J. A., and Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

[Morris et al., 2023] Morris, J. A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D. A., Hao, S., Mimitou, E. P., Smibert, P., Roeder, K., Katsevich, E., Lappalainen, T., and Sanjana, N. E. (2023). Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science*, 380(6646):eadh7699.

[Moulton, 2018] Moulton, V. R. (2018). Sex hormones in acquired immunity and autoimmune disease. *Front. Immunol.*, 9:2279.

[Mousavi et al., 2020] Mousavi, M. J., Mahmoudi, M., and Ghotloo, S. (2020). Escape from X chromosome inactivation and female bias of autoimmune diseases. *Mol. Med.*, 26(1):127.

[Murat et al., 2022] Murat, F., Mbengue, N., Winge, S. B., Trefzer, T., Leushkin, E., Sepp, M., Cardoso-Moreira, M., Schmidt, J., Schneider, C., Mößinger, K., Brüning, T., Lamanna, F., Belles, M. R., Conrad, C., Kondova, I., Bontrop, R., Behr, R., Khaitovich, P., Pääbo, S., Marques-Bonet, T., Grützner, F., Almstrup, K., Schierup, M. H., and Kaessmann, H. (2022). The molecular evolution of spermatogenesis across mammals. *Nature*.

[Neikes et al., 2023] Neikes, H. K., Kliza, K. W., Gräwe, C., Wester, R. A., Jansen, P. W. T. C., Lamers, L. A., Baltissen, M. P., van Heeringen, S. J., Logie, C., Teichmann, S. A., Lindeboom, R. G. H., and Vermeulen, M. (2023). Quantification of absolute transcription factor binding affinities in the native chromatin context using BANC-seq. *Nat. Biotechnol.*, 41(12):1801–1809.

[Noviello et al., 2023] Noviello, G., Gjaltema, R. A. F., and Schulz, E. G. (2023). CasTuner is a degron and CRISPR/Cas-based toolkit for analog tuning of endogenous gene expression. *Nat. Commun.*, 14(1):1–17.

[Nutt et al., 1999] Nutt, S. L., Heavey, B., Rolink, A. G., and Busslinger, M. (1999). Commitment to the b-lymphoid lineage depends on the transcription factor pax5. *Nature*, 401(6753):556–562.

[Oghumu et al., 2019] Oghumu, S., Varikuti, S., Stock, J. C., Volpedo, G., Saljoughian, N., Terrazas, C. A., and Satoskar, A. R. (2019). Cutting edge: CXCR3 escapes X chromosome inactivation in T cells during infection: Potential implications for sex differences in immune responses. *J. Immunol.*, 203(4):789–794.

[Oliva et al., 2020] Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A. D. H., Cotter, D. J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., Castel, S. E., Mohammadi, P., Aguet, F., Zou, Y., Khramtsova, E. A., Skol, A. D., Garrido-Martín, D., Reverter, F., Brown, A., Evans, P., Gamazon, E. R., Payne, A., Bonazzola, R., Barbeira, A. N., Hamel, A. R., Martinez-Perez, A., Soria, J. M., GTEx Consortium, Pierce, B. L., Stephens, M., Eskin, E., Dermitzakis, E. T., Segrè, A. V., Im, H. K., Engelhardt, B. E., Ardlie, K. G., Montgomery, S. B., Battle, A. J., Lappalainen, T., Guigó, R., and Stranger, B. E. (2020). The impact of sex on gene expression across human tissues. *Science*, 369(6509).

[Orozco et al., 2012] Orozco, L. D., Bennett, B. J., Farber, C. R., Ghazalpour, A., Pan, C., Che, N., Wen, P., Qi, H. X., Mutukulu, A., Siemers, N., Neuhaus, I., Yordanova, R., Gargalovic, P., Pellegrini, M., Kirchgessner, T., and Lusis, A. J. (2012). Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. *Cell*, 151(3):658–670.

[Osumi-Sutherland et al., 2021] Osumi-Sutherland, D., Xu, C., Keays, M., Levine, A. P., Kharchenko, P. V., Regev, A., Lein, E., and Teichmann, S. A. (2021). Cell type ontologies of the human cell atlas. *Nat. Cell Biol.*, 23(11):1129–1135.

[Ozadam et al., 2023] Ozadam, H., Tonn, T., Han, C. M., Segura, A., Hoskins, I., Rao, S., Ghatpande, V., Tran, D., Catoe, D., Salit, M., and Cenik, C. (2023). Single-cell quantification of ribosome occupancy in early mouse development. *Nature*, 618(7967):1057–1064.

[Pacini et al., 2021] Pacini, G., Dunkel, I., Mages, N., Mutzel, V., Timmermann, B., Marsico, A., and Schulz, E. G. (2021). Integrated analysis of xist upregulation and x-chromosome inactivation with single-cell and single-allele resolution. *Nat. Commun.*, 12(1):1–17.

[Pai et al., 2015] Pai, A. A., Pritchard, J. K., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.*, 11(1):e1004857.

[Papalexi and Satija, 2017] Papalexi, E. and Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.*, 18(1):35–45.

[Pappalardo and Barra, 2021] Pappalardo, X. G. and Barra, V. (2021). Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics Chromatin*, 14(1):25.

[Park, 2009] Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10(10):669–680.

[Patel et al., 2023] Patel, H. P., Coppola, S., Pomp, W., Aiello, U., Brouwer, I., Libri, D., and Lenstra, T. L. (2023). DNA supercoiling restricts the transcriptional bursting of neighboring eukaryotic genes. *Mol. Cell*, 83(10):1573–1587.e8.

[Peeters et al., 2023] Peeters, S., Leung, T., Fornes, O., Farkas, R. A., Wasserman, W. W., and Brown, C. J. (2023). Refining the genomic determinants underlying escape from x-chromosome inactivation. *NAR Genom Bioinform*, 5(2):lqad052.

[Peeters et al., 2014] Peeters, S. B., Cotton, A. M., and Brown, C. J. (2014). Variable escape from x-chromosome inactivation: identifying factors that tip the scales towards expression. *Bioessays*, 36(8):746–756.

[Petropoulos et al., 2016] Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026.

[Philpott et al., 2021] Philpott, M., Watson, J., Thakurta, A., Brown, T., Oppermann, U., and Cribbs, A. P. (2021). Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat. Biotechnol.*, 39(12):1517–1520.

[Picelli et al., 2014] Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171–181.

[Pierce et al., 2021] Pierce, S. E., Granja, J. M., and Greenleaf, W. J. (2021). High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.*, 12(1):2969.

[Piunti and Shilatifard, 2021] Piunti, A. and Shilatifard, A. (2021). The roles of polycomb repressive complexes in mammalian development and cancer. *Nat. Rev. Mol. Cell Biol.*, 22(5):326–345.

[Plath et al., 2002] Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A., and Panning, B. (2002). Xist RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.*, 36:233–278.

[Pliner et al., 2018] Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., and Trapnell, C. (2018). Cicero predicts cis-regulatory DNA interactions from Single-Cell chromatin accessibility data. *Mol. Cell*, 71(5):858–871.e8.

[Porubsky et al., 2020] Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., Ebler, J., Munson, K. M., Sorensen, M., Sulovari, A., Haukness, M., Ghareghani, M., Lansdorp, P. M., Paten, B., Devine, S. E., Sanders, A. D., Lee, C., Chaisson, M. J. P., Korbel, J. O., Eichler, E. E., and Marschall, T. (2020). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.*, 39(3):302–308.

[Potter, 2018] Potter, S. S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.*, 14(8):479–492.

[Qi et al., 2023] Qi, G., Strober, B. J., Popp, J. M., Keener, R., Ji, H., and Battle, A. (2023). Single-cell allele-specific expression analysis reveals dynamic and cell-type-specific regulatory effects. *Nat. Commun.*, 14(1):6317.

[Qu et al., 2015] Qu, K., Zaba, L. C., Giresi, P. G., Li, R., Longmire, M., Kim, Y. H., Greenleaf, W. J., and Chang, H. Y. (2015). Individuality and variation of personal regulomes in primary human T cells. *Cell Syst*, 1(1):51–61.

[Quinones-Valdez et al., 2022] Quinones-Valdez, G., Fu, T., Chan, T. W., and Xiao, X. (2022). scallele: A versatile tool for the detection and analysis of variants in scRNA-seq. *Science Advances*, 8(35):eabn6398.

[Ragavan and Patel, 2022] Ragavan, M. and Patel, M. I. (2022). The evolving landscape of sex-based differences in lung cancer: a distinct disease in women. *Eur. Respir. Rev.*, 31(163).

[Raj and van Oudenaarden, 2008] Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226.

[Rasmussen, 2004] Rasmussen, C. E. (2004). Gaussian processes in machine learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Rathke et al., 2014] Rathke, C., Baarends, W. M., Awe, S., and Renkawitz-Pohl, R. (2014). Chromatin dynamics during spermiogenesis. *Biochim. Biophys. Acta*, 1839(3):155–168.

[Reddy et al., 2019] Reddy, P. C., Gungi, A., and Unni, M. (2019). Cellular and molecular mechanisms of hydra regeneration. *Results Probl. Cell Differ.*, 68:259–290.

[Regev et al., 2017] Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen,

N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., and Human Cell Atlas Meeting Participants (2017). The human cell atlas. *Elife*, 6.

[Reik and Walter, 2001] Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, 2(1):21–32.

[Reinius et al., 2016] Reinius, B., Mold, J. E., Ramsköld, D., Deng, Q., Johnsson, P., Michaëlsson, J., Frisén, J., and Sandberg, R. (2016). Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.*, 48(11):1430–1435.

[Reinius and Sandberg, 2015] Reinius, B. and Sandberg, R. (2015). Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.*, 16(11):653–664.

[Richart et al., 2022] Richart, L., Picod-Chedotel, M.-L., Wassef, M., Macario, M., Aflaki, S., Salvador, M. A., Héry, T., Dauphin, A., Wicinski, J., Chevrier, V., Pastor, S., Guittard, G., Le Cam, S., Kamhawi, H., Castellano, R., Guasch, G., Charafe-Jauffret, E., Heard, E., Margueron, R., and Ginestier, C. (2022). XIST loss impairs mammary stem cell differentiation and increases tumorigenicity through mediator hyperactivation. *Cell*, 185(12):2164–2183.e25.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Robles-Espinoza et al., 2021] Robles-Espinoza, C. D., Mohammadi, P., Bonilla, X., and Gutierrez-Arcelus, M. (2021). Allele-specific expression: applications in cancer and technical considerations. *Curr. Opin. Genet. Dev.*, 66:10–19.

[Rodriguez and Larson, 2020] Rodriguez, J. and Larson, D. R. (2020). Transcription in living cells: Molecular mechanisms of bursting. *Annu. Rev. Biochem.*, 89:189–212.

[Rodriguez et al., 2019] Rodriguez, J., Ren, G., Day, C. R., Zhao, K., Chow, C. C., and Larson, D. R. (2019). Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, 176(1-2):213–226.e18.

[Roller et al., 2021] Roller, M., Stamper, E., Villar, D., Izuogu, O., Martin, F., Redmond, A. M., Ramachanderan, R., Harewood, L., Odom, D. T., and Flicek, P. (2021). LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.*, 22(1):62.

[Rubin et al., 2019] Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., Mumbach, M. R., Ji, A. L., Kim, D. S., Cho, S. W., Zarnegar, B. J., Greenleaf, W. J., Chang, H. Y., and Khavari, P. A. (2019). Coupled Single-Cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, 176(1):361–376.e17.

[Salmen et al., 2022] Salmen, F., De Jonghe, J., Kaminski, T. S., Alemany, A., Parada, G. E., Verity-Legg, J., Yanagida, A., Kohler, T. N., Battich, N., van den Brekel, F., Ellermann, A. L., Arias, A. M., Nichols, J., Hemberg, M., Hollfelder, F., and van Oudenaarden, A. (2022). High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol.*, 40(12):1780–1793.

[Santoni et al., 2017] Santoni, F. A., Stamoulis, G., Garieri, M., Falconnet, E., Ribaux, P., Borel, C., and Antonarakis, S. E. (2017). Detection of imprinted genes by Single-Cell Allele-Specific gene expression. *Am. J. Hum. Genet.*, 100(3):444–453.

[Santoro et al., 2021] Santoro, A., Bientinesi, E., and Monti, D. (2021). Immunosenescence and inflammaging in the aging process: age-related diseases or longevity? *Ageing Res. Rev.*, 71:101422.

[Sarkar et al., 2019] Sarkar, A. K., Tung, P.-Y., Blischak, J. D., Burnett, J. E., Li, Y. I., Stephens, M., and Gilad, Y. (2019). Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.*, 15(4):e1008045.

[Sarropoulos et al., 2021] Sarropoulos, I., Sepp, M., Frömel, R., Leiss, K., Trost, N., Leushkin, E., Okonechnikov, K., Joshi, P., Giere, P., Kutscher, L. M., Cardoso-Moreira, M., Pfister, S. M., and Kaessmann, H. (2021). Developmental and evolutionary dynamics of cis-regulatory elements in mouse cerebellar cells. *Science*, 373(6558).

[Schadt et al., 2003] Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302.

[Schoeftner et al., 2009] Schoeftner, S., Blanco, R., Lopez de Silanes, I., Muñoz, P., Gómez-López, G., Flores, J. M., and Blasco, M. A. (2009). Telomere shortening relaxes X chromosome inactivation and forces global transcriptome alterations. *Proc. Natl. Acad. Sci. U. S. A.*, 106(46):19393–19398.

[Schraivogel et al., 2020] Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., and Steinmetz, L. M. (2020). Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods*, 17(6):629–635.

[Segert et al., 2021] Segert, J. A., Gisselbrecht, S. S., and Bulyk, M. L. (2021). Transcriptional silencers: Driving gene expression with the brakes on. *Trends Genet.*, 37(6):514–527.

[Shami et al., 2020] Shami, A. N., Zheng, X., Munyoki, S. K., Ma, Q., Manske, G. L., Green, C. D., Sukhwani, M., Orwig, K. E., Li, J. Z., and Hammoud, S. S. (2020). Single-Cell RNA sequencing of human, macaque, and mouse testes uncovers conserved and divergent features of mammalian spermatogenesis. *Dev. Cell*, 54(4):529–547.e12.

[Shen et al., 2014] Shen, S. Q., Turro, E., and Corbo, J. C. (2014). Hybrid mice reveal parent-of-origin and cis- and trans-regulatory effects in the retina. *PLoS One*, 9(10):e109382.

[Shi et al., 2016] Shi, Y., Inoue, H., Wu, J. C., and Yamanaka, S. (2016). Induced pluripotent stem cell technology: a decade of progress. *Nat. Rev. Drug Discov.*, 16(2):115–130.

[Shi et al., 2023] Shi, Z.-X., Chen, Z.-C., Zhong, J.-Y., Hu, K.-H., Zheng, Y.-F., Chen, Y., Xie, S.-Q., Bo, X.-C., Luo, F., Tang, C., Xiao, C.-L., and Liu, Y.-Z. (2023). High-throughput and high-accuracy single-cell RNA isoform analysis using PacBio circular consensus sequencing. *Nat. Commun.*, 14(1):1–13.

[Shvetsova et al., 2019] Shvetsova, E., Sofronova, A., Monajemi, R., Gagalova, K., Draisma, H. H. M., White, S. J., Santen, G. W. E., Chuva de Sousa Lopes, S. M., Heijmans, B. T., van Meurs, J., Jansen, R., Franke, L., Kiełbasa, S. M., den Dunnen, J. T., 't Hoen, P. A. C., BIOS consortium, and GoNL consortium (2019). Skewed x-inactivation is common in the general female population. *Eur. J. Hum. Genet.*, 27(3):455–465.

[Sierra et al., 2023] Sierra, I., Pyfrom, S., Weiner, A., Zhao, G., Driscoll, A., Yu, X., Gregory, B. D., Vaughan, A. E., and Anguera, M. C. (2023). Unusual X chromosome inactivation maintenance in female alveolar type 2 cells is correlated with increased numbers of x-linked escape genes and sex-biased gene expression. *Stem Cell Reports*, 18(2):489–502.

[Signor and Nuzhdin, 2018] Signor, S. A. and Nuzhdin, S. V. (2018). The evolution of gene expression in cis and trans. *Trends Genet.*, 34(7):532–544.

[Smallwood et al., 2014] Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*, 11(8):817–820.

[Smith et al., 2020] Smith, J., Sen, S., Weeks, R. J., Eccles, M. R., and Chatterjee, A. (2020). Promoter DNA hypermethylation and paradoxical gene activation. *Trends Cancer Res.*, 6(5):392–406.

[Soumillon et al., 2013] Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., Dym, M., de Massy, B., Mikkelsen, T. S., and Kaessmann, H. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.*, 3(6):2179–2190.

[Souyris et al., 2018] Souyris, M., Cenac, C., Azar, P., Daviaud, D., Canivet, A., Grunenwald, S., Pienkowski, C., Chaumeil, J., Mejía, J. E., and Guéry, J.-C. (2018). TLR7 escapes X chromosome inactivation in immune cells. *Sci Immunol*, 3(19).

[Spitz and Furlong, 2012] Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626.

[Splinter et al., 2011] Splinter, E., de Wit, E., Nora, E. P., Klous, P., van de Werken, H. J. G., Zhu, Y., Kaaij, L. J. T., van IJcken, W., Gribnau, J., Heard, E., and de Laat, W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on xist RNA. *Genes Dev.*, 25(13):1371–1383.

[Srivatsan et al., 2020] Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., Zhang, F., Steemers, F., Shendure, J., and Trapnell, C. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51.

[Srivatsan et al., 2021] Srivatsan, S. R., Regier, M. C., Barkan, E., Franks, J. M., Packer, J. S., Grosjean, P., Duran, M., Saxton, S., Ladd, J. J., Spielmann, M., Lois, C., Lampe, P. D., Shendure, J., Stevens, K. R., and Trapnell, C. (2021). Embryo-scale, single-cell spatial transcriptomics. *Science*, 373(6550):111–117.

[Stark et al., 2019] Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.*, 20(11):631–656.

[Stefflova et al., 2013] Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D. J., Talianidis, I., Marioni, J. C., Flicek, P., and Odom, D. T. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540.

[Stegle et al., 2010] Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol.*, 17(3):355–367.

[Stoeckius et al., 2017] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, 14(9):865–868.

[Strober et al., 2019] Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., and Gilad, Y. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290.

[Stuart and Satija, 2019] Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.*, 20(5):257–272.

[Sudlow et al., 2015] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779.

[Sun et al., 2018a] Sun, J. H., Zhou, L., Emerson, D. J., Phyo, S. A., Titus, K. R., Gong, W., Gilgenast, T. G., Beagan, J. A., Davidson, B. L., Tassone, F., and Phillips-Cremins, J. E. (2018a). Disease-Associated short tandem repeats co-localize with chromatin domain boundaries. *Cell*, 175(1):224–238.e15.

[Sun et al., 2018b] Sun, W., Gao, Q., Schaefke, B., Hu, Y., and Chen, W. (2018b). Pervasive allele-specific regulation on RNA decay in hybrid mice. *Life Sci Alliance*, 1(2):e201800052.

[Sun et al., 2022] Sun, X., Nguyen, T., Achour, A., Ko, A., Cifello, J., Ling, C., Sharma, J., Hiroi, T., Zhang, Y., Chia, C. W., Wood, 3rd, W., Wu, W. W., Zukley, L., Phue, J.-N., Becker, K. G., Shen, R.-F., Ferrucci, L., and Weng, N.-P. (2022). Longitudinal analysis reveals age-related changes in the T cell receptor repertoire of human T cell subsets. *J. Clin. Invest.*, 132(17).

[Svensson, 2020] Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, 38(2):147–150.

[Svensson et al., 2018] Svensson, V., Teichmann, S. A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nat. Methods*, 15(5):343–346.

[Takahashi and Yamanaka, 2006] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676.

[Takahashi and Iwasaki, 2021] Takahashi, T. and Iwasaki, A. (2021). Sex differences in immune responses. *Science*, 371(6527):347–348.

[Tam and Loebel, 2007] Tam, P. P. L. and Loebel, D. A. F. (2007). Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.*, 8(5):368–381.

[Tang et al., 2009] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382.

[Tauc et al., 2021] Tauc, H. M., Rodriguez-Fernandez, I. A., Hackney, J. A., Pawlak, M., Ronnen Oron, T., Korzelius, J., Moussa, H. F., Chaudhuri, S., Modrusan, Z., Edgar, B. A., and Jasper, H. (2021). Age-related changes in polycomb gene regulation disrupt lineage fidelity in intestinal stem cells. *Elife*, 10:e62250.

[Teixeira and Lehmann, 2019] Teixeira, F. K. and Lehmann, R. (2019). Translational control during developmental transitions. *Cold Spring Harb. Perspect. Biol.*, 11(6).

[Threadgill et al., 2011] Threadgill, D. W., Miller, D. R., Churchill, G. A., and de Villena, F. P.-M. (2011). The collaborative cross: a recombinant inbred mouse population for the systems genetic era. *ILAR J.*, 52(1):24–31.

[Tirosh et al., 2009] Tirosh, I., Reikhav, S., Levy, A. A., and Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, 324(5927):659–662.

[Tomofuji et al., 2023] Tomofuji, Y., Edahiro, R., Shirai, Y., Kock, K. H., Sonehara, K., Wang, Q. S., Namba, S., Moody, J., Ando, Y., Suzuki, A., Yata, T., Ogawa, K., Namkoong, H., Lin, Q. X. X., Buyamin, E. V., Le Min, T., Sonthalia, R., Han, K. Y., Tanaka, H., Lee, H., Asian Immune Diversity Atlas Network, Japan COVID-19 Task Force, The BioBank Japan Project, Okuno, T., Liu, B., Matsuda, K., Fukunaga, K., Mochizuki, H., Park, W.-Y., Yamamoto, K., Hon, C.-C., Shin, J. W., Prabhakar, S., Kumanogoh, A., and Okada, Y. (2023). Quantification of the escape from X chromosome inactivation with the million cell-scale human single-cell omics datasets reveals heterogeneity of escape across cell types and tissues.

[Trapnell, 2015] Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.*, 25(10):1491–1498.

[Trapnell et al., 2014] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386.

[Travers et al., 2013] Travers, M. E., Mackay, D. J. G., Dekker Nitert, M., Morris, A. P., Lindgren, C. M., Berry, A., Johnson, P. R., Hanley, N., Groop, L. C., McCarthy, M. I., and Gloyn, A. L. (2013). Insights into the molecular mechanism for type 2 diabetes susceptibility at the KCNQ1 locus from temporal changes in imprinting status in human islets. *Diabetes*, 62(3):987–992.

[Tukiainen et al., 2017] Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., Lappalainen, T., Regev, A., Ardlie, K. G., Hacohen, N., and MacArthur, D. G. (2017). Landscape of X chromosome inactivation across human tissues. *Nature*, 550(7675):244–248.

[Tunnacliffe and Chubb, 2020] Tunnacliffe, E. and Chubb, J. R. (2020). What is a transcriptional burst? *Trends Genet.*, 36(4):288–297.

[Uffelmann et al., 2021] Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., and Lappalainen, T. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21.

[Vallejos et al., 2015] Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian analysis of Single-Cell sequencing data. *PLoS Comput. Biol.*, 11(6):e1004333.

[van de Geijn et al., 2015] van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11):1061–1063.

[van den Berg et al., 2011] van den Berg, I. M., Galjaard, R. J., Laven, J. S. E., and van Doorninck, J. H. (2011). XCI in preimplantation mouse and human embryos: first there is remodelling. . . . *Hum. Genet.*, 130(2):203–215.

[van der Wijst et al., 2020] van der Wijst, M., de Vries, D. H., Groot, H. E., Trynka, G., Hon, C. C., Bonder, M. J., Stegle, O., Nawijn, M. C., Idaghdour, Y., van der Harst, P., Ye, C. J., Powell, J., Theis, F. J., Mahfouz, A., Heinig, M., and Franke, L. (2020). The single-cell eQTLGen consortium. *Elife*, 9.

[VanInsberghe et al., 2021] VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., Clevers, H., and van Oudenaarden, A. (2021). Single-cell ribo-seq reveals cell cycle-dependent translational pausing. *Nature*, 597(7877):561–565.

[Velten et al., 2022] Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., and Stegle, O. (2022). Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods*, 19(2):179–186.

[Velten et al., 2017] Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., Hirche, C., Lutz, C., Buss, E. C., Nowak, D., Boch, T., Hofmann, W.-K., Ho, A. D., Huber, W., Trumpp, A., Essers, M. A. G., and Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.*, 19(4):271–281.

[Verdeil, 2016] Verdeil, G. (2016). MAF drives CD8+ t-cell exhaustion. *Oncoimmunology*, 5(2):e1082707.

[Vervoort et al., 2022] Vervoort, S. J., Devlin, J. R., Kwiatkowski, N., Teng, M., Gray, N. S., and Johnstone, R. W. (2022). Targeting transcription cycles in cancer. *Nat. Rev. Cancer*, 22(1):5–24.

[Võsa et al., 2021] Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M.-J., Agbessi, M.,

Christiansen, M. W., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez Flitman, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg, J., Kettunen, J., Lee, B., Zhang, F., Qi, T., Hernandez, J. A., Arindrarto, W., Beutner, F., BIOS Consortium, i2QTL Consortium, Dmitrieva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Hewitt, A. W., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M. G., Penninx, B. W. J. H., Pritchard, J. K., Raitakari, O. T., Rotzschke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., 't Hoen, P. A. C., Thiery, J., Tönjes, A., van Dongen, J., van Iterson, M., Veldink, J. H., Völker, U., Warmerdam, R., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., van Meurs, J., Prokisch, H., Ahsan, H., Pierce, B. L., Lehtimäki, T., Boomsma, D. I., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W. H., Downes, K., Stegle, O., Battle, A., Visscher, P. M., Yang, J., Scholz, M., Powell, J., Gibson, G., Esko, T., and Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.*, 53(9):1300–1310.

[Wainer Katsir and Linial, 2019] Wainer Katsir, K. and Linial, M. (2019). Human genes escaping x-inactivation revealed by single cell expression data. *BMC Genomics*, 20(1):201.

[Wang et al., 2016] Wang, J., Syrett, C. M., Kramer, M. C., Basu, A., Atchison, M. L., and Anguera, M. C. (2016). Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. *Proc. Natl. Acad. Sci. U. S. A.*, 113(14):E2029–38.

[Wang et al., 2022] Wang, Q., Li, M., Wu, T., Zhan, L., Li, L., Chen, M., Xie, W., Xie, Z., Hu, E., Xu, S., and Yu, G. (2022). Exploring epigenomic datasets by ChIPseeker. *Curr Protoc*, 2(10):e585.

[Ward et al., 2021] Ward, M. C., Banovich, N. E., Sarkar, A., Stephens, M., and Gilad, Y. (2021). Dynamic effects of genetic variation on gene expression revealed following hypoxic stress in cardiomyocytes. *Elife*, 10:e57345.

[Weedon et al., 2013] Weedon, M. N., Cebola, I., Patch, A.-M., Flanagan, S. E., De Franco, E., Caswell, R., Rodríguez-Seguí, S. A., Shaw-Smith, C., Cho, C. H.-H., Allen, H. L., Houghton, J. A. L., Roth, C. L., Chen, R., Hussain, K., Marsh, P., Vallier, L., Murray, A., Ellard, S., Ferrer, J., and Hattersley, A. T. (2013). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.*, 46(1):61–64.

[Weilguny et al., 2023] Weilguny, L., De Maio, N., Munro, R., Manser, C., Birney, E., Loose, M., and Goldman, N. (2023). Dynamic, adaptive sampling during nanopore sequencing using bayesian experimental design. *Nat. Biotechnol.*, 41(7):1018–1025.

[Wilkinson et al., 2007] Wilkinson, L. S., Davies, W., and Isles, A. R. (2007). Genomic imprinting effects on brain development and function. *Nat. Rev. Neurosci.*, 8(11):832–843.

[Wilkinson et al., 2020] Wilkinson, M. E., Charenton, C., and Nagai, K. (2020). RNA splicing by the spliceosome. *Annu. Rev. Biochem.*, 89:359–388.

[Willard, 1996] Willard, H. F. (1996). X chromosome inactivation, XIST, and pursuit of the X-Inactivation center. *Cell*, 86(1):5–7.

[Wittkopp et al., 2004] Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88.

[Wittkopp and Kalay, 2011] Wittkopp, P. J. and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.*, 13(1):59–69.

[Wong et al., 2015] Wong, E. S., Thybert, D., Schmitt, B. M., Stefflova, K., Odom, D. T., and Flicek, P. (2015). Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.*, 25(2):167–178.

[Wood et al., 2010] Wood, M. D., Hiura, H., Tunster, S. J., Arima, T., Shin, J.-Y., Higgins, M. J., and John, R. M. (2010). Autonomous silencing of the imprinted cdkn1c gene in stem cells. *Epigenetics*, 5(3):214–221.

[Wu et al., 2014] Wu, H., Luo, J., Yu, H., Rattner, A., Mo, A., Wang, Y., Smallwood, P. M., Erlanger, B., Wheelan, S. J., and Nathans, J. (2014). Cellular resolution maps of X chromosome inactivation: implications for neural development, function, and disease. *Neuron*, 81(1):103–119.

[Wutz et al., 2002] Wutz, A., Rasmussen, T. P., and Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of xist RNA. *Nat. Genet.*, 30(2):167–174.

[Xing et al., 2022] Xing, E., Billi, A. C., and Gudjonsson, J. E. (2022). Sex bias and autoimmune diseases. *J. Invest. Dermatol.*, 142(3 Pt B):857–866.

[Xu et al., 2017] Xu, J., Carter, A. C., Gendrel, A.-V., Attia, M., Loftus, J., Greenleaf, W. J., Tibshirani, R., Heard, E., and Chang, H. Y. (2017). Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells. *Nat. Genet.*, 49(3):377–386.

[Yamada et al., 2009] Yamada, T., Park, C. S., Mamonkin, M., and Lacorazza, H. D. (2009). Transcription factor ELF4 controls the proliferation and homing of CD8+ T cells via the krüppel-like factors KLF4 and KLF2. *Nat. Immunol.*, 10(6):618–626.

[Yamasaki et al., 2003] Yamasaki, K., Joh, K., Ohta, T., Masuzaki, H., Ishimaru, T., Mukai, T., Niikawa, N., Ogawa, M., and Kishino, T. (2003). Neurons but not glial cells show reciprocal imprinting of sense and antisense transcripts of ube3a. *Hum. Mol. Genet.*, 12(8):837–847.

[Yang et al., 2019] Yang, E.-W., Bahn, J. H., Hsiao, E. Y.-H., Tan, B. X., Sun, Y., Fu, T., Zhou, B., Van Nostrand, E. L., Pratt, G. A., Freese, P., Wei, X., Quinones-Valdez, G., Urban, A. E., Graveley, B. R., Burge, C. B., Yeo, G. W., and Xiao, X. (2019). Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nat. Commun.*, 10(1):1–15.

[Yang et al., 2020] Yang, L., Yildirim, E., Kirby, J. E., Press, W., and Lee, J. T. (2020). Widespread organ tolerance to xist loss and X reactivation except under chronic stress in the gut. *Proceedings of the National Academy of Sciences*, 117(8):4262–4272.

[Yang et al., 2022a] Yang, M. G., Ling, E., Cowley, C. J., Greenberg, M. E., and Vierbuchen, T. (2022a). Characterization of sequence determinants of enhancer function using natural genetic variation. *Elife*, 11.

[Yang et al., 2022b] Yang, T., Ou, J., and Yildirim, E. (2022b). Xist exerts gene-specific silencing during XCI maintenance and impacts lineage-specific cell differentiation and proliferation during hematopoiesis. *Nat. Commun.*, 13(1):1–19.

[Yazar et al., 2022] Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R. P., Clarke, L., Maccora, K., Chen, C., Cook, A. L., Ye, C. J., Fairfax, K. A., Hewitt, A. W., and Powell, J. E. (2022). Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041.

[Ye et al., 2020] Ye, Y., Zhang, Z., Liu, Y., Diao, L., and Han, L. (2020). A Multi-Omics perspective of quantitative trait loci in precision medicine. *Trends Genet.*, 36(5):318–336.

[Yesbolatova et al., 2020] Yesbolatova, A., Saito, Y., Kitamoto, N., Makino-Itou, H., Ajima, R., Nakano, R., Nakaoka, H., Fukui, K., Gamo, K., Tominari, Y., Takeuchi, H., Saga, Y., Hayashi, K.-I., and Kanemaki, M. T. (2020). The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice. *Nat. Commun.*, 11(1):1–13.

[Yildirim et al., 2013] Yildirim, E., Kirby, J. E., Brown, D. E., Mercier, F. E., Sadreyev, R. I., Scadden, D. T., and Lee, J. T. (2013). Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell*, 152(4):727–742.

[Ying et al., 2008] Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. *Nature*, 453(7194):519–523.

[Youness et al., 2021] Youness, A., Miquel, C.-H., and Guéry, J.-C. (2021). Escape from X chromosome inactivation and the female predominance in autoimmune diseases. *Int. J. Mol. Sci.*, 22(3).

[Yousefzadeh et al., 2021] Yousefzadeh, M. J., Flores, R. R., Zhu, Y., Schmiechen, Z. C., Brooks, R. W., Trussoni, C. E., Cui, Y., Angelini, L., Lee, K.-A., McGowan, S. J., Burrack, A. L., Wang, D., Dong, Q., Lu, A., Sano, T., O'Kelly, R. D., McGuckian, C. A., Kato, J. I., Bank, M. P., Wade, E. A., Pillai, S. P. S., Klug, J., Ladiges, W. C., Burd, C. E., Lewis, S. E., LaRusso, N. F., Vo, N. V., Wang, Y., Kelley, E. E., Huard, J., Stromnes, I. M., Robbins, P. D., and Niedernhofer, L. J. (2021). An aged immune system drives senescence and ageing of solid organs. *Nature*, 594(7861):100–105.

[Yu et al., 2021] Yu, B., Qi, Y., Li, R., Shi, Q., Satpathy, A. T., and Chang, H. Y. (2021). B cell-specific XIST complex enforces x-inactivation and restrains atypical B cells. *Cell*, 184(7):1790–1803.e17.

[Yu et al., 2006] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, 38(2):203–208.

[Yue et al., 2014] Yue, M., Charles Richard, J. L., Yamada, N., Ogawa, A., and Ogawa, Y. (2014). Quick fluorescent in situ hybridization protocol for xist RNA combined with immunofluorescence of histone modification in x-chromosome inactivation. *J. Vis. Exp.*, (93):e52053.

[Zaitlen and Kraft, 2012] Zaitlen, N. and Kraft, P. (2012). Heritability in the genome-wide association era. *Hum. Genet.*, 131(10):1655–1664.

[Zeng, 2022] Zeng, H. (2022). What is a cell type and how to define it? *Cell*, 185(15):2739–2755.

[Zhang et al., 2008] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137.

[Zheng et al., 2017] Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8(1):1–12.

[Zheng et al., 2011] Zheng, W., Gianoulis, T. A., Karczewski, K. J., Zhao, H., and Snyder, M. (2011). Regulatory variation within and between species. *Annu. Rev. Genomics Hum. Genet.*, 12:327–346.

[Zhernakova et al., 2017] Zhernakova, D. V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., Bonder, M. J., van Rooij, J., Verkerk, M., Jhamai, P. M., Moed, M., Kielbasa, S. M., Bot, J., Nooren, I., Pool, R., van Dongen, J., Hottenga, J. J., Stehouwer, C. D. A., van der Kallen, C. J. H., Schalkwijk, C. G., Zhernakova, A., Li, Y., Tigchelaar, E. F., de Klein, N., Beekman, M., Deelen, J., van Heemst, D., van den Berg, L. H., Hofman, A., Uitterlinden, A. G., van Greevenbroek, M. M. J., Veldink, J. H., Boomsma, D. I., van Duijn, C. M., Wijmenga, C., Slagboom, P. E., Swertz, M. A., Isaacs, A., van Meurs, J. B. J., Jansen, R., Heijmans, B. T., 't Hoen, P. A. C., and Franke, L. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.*, 49(1):139–145.

[Zhou et al., 2023] Zhou, Y., Zhan, X., Jin, J., Zhou, L., Bergman, J., Li, X., Rousselle, M. M. C., Belles, M. R., Zhao, L., Fang, M., Chen, J., Fang, Q., Kuderna, L., Marques-Bonet, T., Kitayama, H., Hayakawa, T., Yao, Y.-G., Yang, H., Cooper, D. N., Qi, X., Wu, D.-D., Schierup, M. H., and Zhang, G. (2023). Eighty million years of rapid evolution of the primate Y chromosome. *Nature Ecology & Evolution*, 7(7):1114–1130.

[Ziegenhain et al., 2017] Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4.

[Zimmerman et al., 2021] Zimmerman, K. D., Espeland, M. A., and Langefeld, C. D. (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.*, 12(1):1–9.

[Zou et al., 2021] Zou, L. S., Zhao, T., Cable, D. M., Murray, E., Aryee, M. J., Chen, F., and Irizarry, R. A. (2021). Detection of allele-specific expression in spatial transcriptomics with spASE.