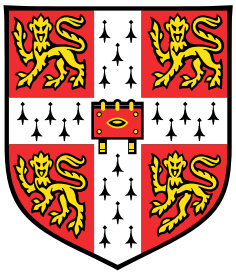
Human cellular genetics of innate immunity



Raghd Rostom

Wellcome Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Christ's College

August 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding tables, footnotes, bibliography and appendices.

Raghd Rostom

August 2019

Human Cellular Genetics of Innate Immunity

Raghd Rostom

The type I interferon response is a key part of the innate immune system, responding to infection and inducing an antiviral intracellular state. While there is known to be variability in this signalling pathway between individuals, alongside cell-to-cell heterogeneity in a genetically identical cell population, the basis of this variation is not fully understood.

In this PhD, I established large-scale single-cell RNA sequencing experiments to study cellular variation in the innate immune response in fibroblasts of 70 healthy human individuals from the HipSci initiative. Chapter 2 describes optimisation of stimulation conditions to induce an antiviral response, and the experimental work carried out on the panel of donors.

In Chapter 3, I analyse heterogeneity in resting (unstimulated) fibroblasts. By comparing to *ex vivo* skin data containing multiple cell types, I confirm the relative homogeneity of the *in vitro* cultured fibroblasts used, mapping to one sub-population of *ex vivo* skin fibroblasts. Using matched whole exome sequencing data, somatic mutations in sub-populations of cells within each donor were detected, and clonal populations identified. A novel computational method, cardelino, was developed for inference of the clonal tree configuration and the clone of origin of individual cells that have been assayed using scRNA-seq. Applying cardelino to 32 fibroblast lines identifies hundreds of differentially expressed genes between cells from different somatic clones, with cell cycle and proliferation pathways frequently enriched.

Returning to innate immunity, Chapters 4 and 5 centre on variability in the type I interferon response. I first describe work linking variability in the innate immune response and evolutionary divergence across mammalian species. Focusing on human variability, the large dataset described above is used to characterise the innate immune response at single cell resolution, elucidating the dynamics of the response across donors in Chapter 4. Chapter 5 describes the application of quantitative trait loci approaches to innate immune phenotypes. This work characterises both inter- and intra-individual heterogeneity in innate immunity.

Acknowledgements

I am deeply grateful to Sarah Teichmann and Oliver Stegle for giving me the opportunity to undertake this research, and for the mentorship throughout. Their immense knowledge, motivation and drive inspired me to take on this PhD, and continues to inspire me going forward. I would also like to thank the members of both the Teichmann and Stegle groups for their support over the years, both scientifically and socially, without whom day-to-day research would have been an entirely different experience. In particular, I would like to express my gratitude to Tzachi Hagai and Davis McCarthy, whose advice and assistance along the way had a huge impact upon my PhD, along with Mike Stubbington and Kerstin Meyer for their invaluable research guidance.

I gratefully acknowledge the funding received towards my PhD from the Biomedical and Biological Sciences Research Council (BBSRC) and the Wellcome Sanger Institute (WSI), along with my second research home at the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI).

My deepest appreciation goes to my friends, both nearby and far away, for keeping me sane and reminding me about the world outside of research. They have truly shaped the non-scientific part of my PhD experience, both in Cambridge and beyond.

Finally, I could not have made it to this point without my loving family, who always believed in me and encouraged me to follow my dreams. Their inspiration and support throughout has motivated me to work hard and kept me going through challenging times.

Contributions

Chapter 1

The section on single cell RNA sequencing analysis was adapted from a review written with the input of Valentine Svensson, published in FEBS journal.

Chapter 2

Bulk RNA sequencing data for protocol optimisation was generated by Tzachi Hagai. During the expansion and stimulation of HipSci lines, invaluable support was provided by the Cellular Genotyping and Phenotyping facility. Data processing was conducted with the help of Davis McCarthy and the Cellular Genetics Informatics team, WSI.

Chapter 3

Primary skin data was generated by the lab of Muzlifah Haniffa.
The study of clonal structure in fibroblasts was carried out as part of a close collaboration with Davis McCarthy and Yuanhua Huang, who developed the computational method - cardelino - underpinning this analysis, and final figures for the paper. The full manuscript is included in Appendix B.

Chapter 4

The cross-mammalian dataset presented in Section 4.1 was produced by Tzachi Hagai. This work was published in Nature, 2018, and the full paper is included in Appendix C.

Chapter 5

QTL analysis was conducted using a pipeline developed by Marc Jan Bonder, and run with the support of Ni Huang.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AMD Age-related macular degeneration

BASiCS Bayesian analysis of single-cell sequencing

BBKNN Batch balanged k nearest neighbours

CGI CpG island

CyTOF Cytometry by time of flight

DC Diffusion component

DM Distance to median

DPT Diffusion pseudotime

eQTL Expression quantitative trait loci

FACS Fluorescence-activated cell sorting

FISH Fluorescence in situ hybridisation

FPKM Fragments per kilobase per million

GLM Generalised linear model

GPLVM Gaussian process latent variable model

GWAS Genome wide association study

HipSci Human Induced Pluripotent Stem Cell Initiative

IFNs Interferons

IIG Innate immune gene

IVT *In vitro* transcription

LF Lipofectamine

LMM Linear mixed model

LPS Lipopolysaccharide

LRT Likelihood ratio test

MDS Multidimensional Scaling

MNN Mutual nearest neighbour

MST Minimum spanning tree

NLRs NOD-like receptors

LD Linkage disequilibrium

PAMPs Pathogen associated molecular patterns

PCA Principal Component Analysis

pDCs Plasmacytoid dendritic cells

Poly(I:C) Polyinosinic:polycytidylic acid

PRRs Pattern recognition receptors

RLRs RIG-I-like receptors

ROS Reactive oxygen species

RT Reverse transcription

scDNA-seq Single cell DNA sequencing

SCG Single Cell Genotyper

scLVM Single cell latent variable model

scMT-seq Single-cell methylome and transcriptome sequencing

scRNA-seq Single-cell RNA-sequencing

scRNA-seq Single cell RNA sequencing

scRRBS Single cell reduced representation bisulfite sequencing

SNN Shared nearest neighbour

SNP Single nucleotide polymorphism

SNV Single nucleotide variants

TLRs Toll-like receptors

TPM Transcripts per million

tSNE t-Distributed Stochastic Neighbour Embedding

UMAP Uniform manifold approximation and projection

UMIs Unique Molecular Identifiers

WGCNA Weighted gene co-expression network analysis

ZIFA Zero-inflated factor analysis

Chapter 1

Introduction

1.1 Human genetic variation

Over the last century, there has been an increasing effort to understand and map genetic variation between humans, and the consequent functional effect of this. This has been accelerated in recent decades with the development of microarray and sequencing technologies, and in particular of high-throughput genetic sequencing. Initiatives such as the Human Genome Project [1], 1000 Genomes Project [2], HapMap project [3], and most recently the 100,000 Genomes Project highlight efforts within the field to chart common variation and the increasing scale at which this is being achieved.

1.1.1 The basis of genetic variation

Genetic variation stems from alteration of DNA sequences, referred to as 'variants' or 'mutations'. These events can occur as a result of endogenous processes, such as errors in DNA replication, chromosome segregation and recombination, or as a result of damage from endogenous or exogenous chemicals (Fig 1.1, [4]). While processes involving chromosome and DNA function are highly regulated, they - like any cellular function

- are not 100% efficient. In the case of DNA replication, around $6x10^9$ nucleotides must be copied in each cell division. Although the major DNA polymerases involved in this DNA synthesis have intrinsic proofreading and exonuclease capacity, allowing the ability to detect and remove incorrectly inserted bases, this does not occur in a very small proportion of cases and errors are maintained. This is often the case at regions in the genome with repeat sequences. In these areas, where there are repeats of particular nucleotide or oligonucleotide sequences, replication slippage may occur, leading to insertion or deletion of nucleotides. On a larger scale, errors in chromosome segregation and recombination may lead to variation in copy number of substantial regions of DNA or entire chromosomes. In the germline, these events often lead to embryonic lethality or developmental disorders, however they can also occur in somatic cells - a typical occurrence in cancer development.

Alongside faults in the processes described above, chemical damage to DNA can cause mutations, deriving from both endogenous and exogenous sources (Figure 1.1). Given the aqueous environment within cells, hydrolytic damage is common. This can lead to the cleavage of covalent N-glycosylic bonds between a base and its sugar, producing an abasic site, or to the deamination of some bases to leave a carbonyl group. Further elements of the cellular milieu produced by normal metabolic reactions can lead to oxidative damage, in particular reactive oxygen species (ROS). The sugar-phosphate backbone can be damaged as a result, or DNA bases can be attacked leading to the production of derivatives, many of which are mutagenic. An alternate source of endogenous damage is the erroneous methylation of adenosine. This causes distortion of the double helix and disrupts DNA-protein interactions. While the majority of chemical damage derives from these intrinsic mechanisms, exogenous agents can also play a role. Examples are the production of ROS within cells due to ionizing radiation from external sources, leading to oxidative damage as described above. Non-ionizing

ultraviolet radiation can also cause damage, resulting from the covalent bonding between adjacent pyrimidines. Finally, environmental chemicals can covalently bond to and distort the DNA helix, such as the large aromatic hydrocarbons found in the smoke of cigarettes and vehicles.

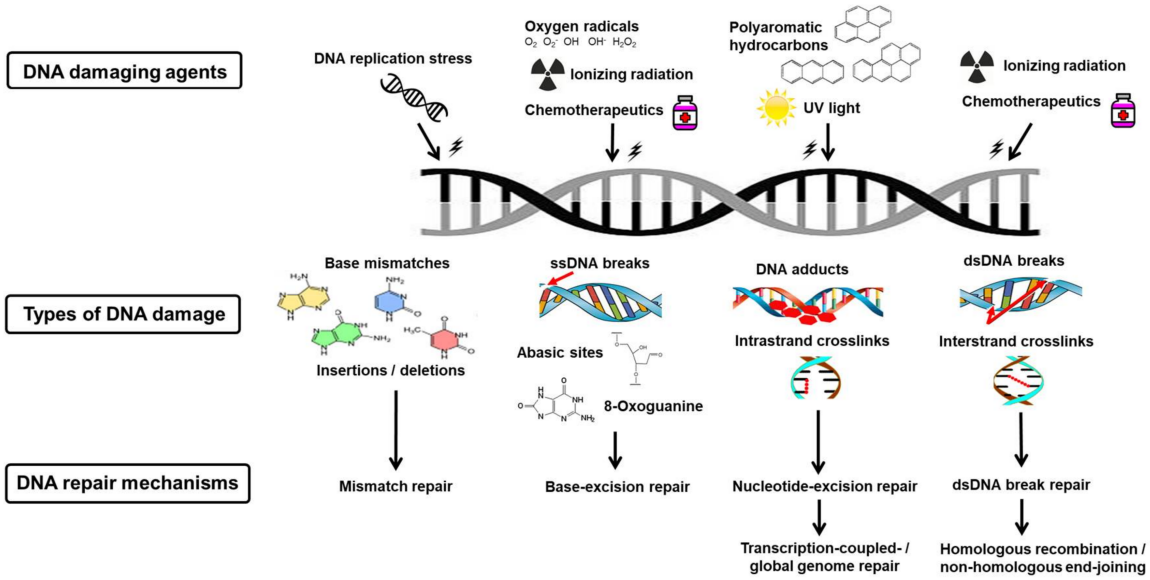


Fig. 1.1 Mechanisms of DNA damage and repair, from Helena *et al.* [4]

If not repaired, these changes in DNA sequence may have a wide range of consequences, or no discernible effect at all. A large proportion of variants do not have a functional effect for several reasons: firstly, much of the genome is non-coding and has no known function. Secondly, there is a high level of genetic redundancy, with substitutions at the third base in a codon sequence often producing the same amino acid (synonymous mutations), and also redundancy in the sequence as a whole - for example, there are hundreds of almost identical ribosomal RNA genes. Finally, even in situations where a variant in a coding gene results in a different amino acid produced (nonsynonymous), this may be functionally unimportant within the protein and therefore tolerated. Despite this, variants with phenotypic effects do arise, and while they may occasionally have a beneficial effect, and may even become positively

selected for in the population, in many cases the mutations are harmful. Variants which have become fixed in the population have been catalogued in dbSNP [5]. Many methods have been used over the years to study the effects of genetic variants, and a brief history and description of modern methods follows.

1.1.2 Approaches for studying human genetic variation

In the early decades of genetic research, familial history was used in genetic linkage studies. The basis of this approach is the increased frequency of co-inheritance of genetic markers close in genomic location than would be expected by chance. Huntington's disease was the first for which the locus - on chromosome 4 - was identified purely by linkage [6]. Following this, developments were made in mapping cystic fibrosis to chromosome 7 [7–10], While this method provided a lot of novel insight in disorders arising from a single gene and with high penetrance - the percentage of individuals with a given genotype who exhibit the associated phenotype - these approaches were more difficult for complex diseases arising from the combination of many low penetrance variants. With the evolution of technologies to assay genome sequences, however, it has become increasingly possible to understand the role of common genetic variants both in disease and healthy phenotypes.

Accelerated by these next-generation sequencing technologies, it has been possible to deeply characterise genetic variation in the population as a whole. This has been highlighted by large-scale international consortia such as the HapMap project [3] and 1000 Genomes Project [2]. The scale of these studies will continue to grow, exemplified by the 100,000 Genomes Project currently underway. This extensive work to map common genetic variation opened the door to genome wide association studies (GWAS).

The GWAS approach is to ask whether a particular variant appears more often in individuals with a phenotype of interest than expected by chance (Figure 1.2). It

is common to use a case-control set up, where two groups are compared: those with the disease/phenotype of interest, and controls without. An odds ratio is calculated, reflecting the odds of the variant in the two groups, with an OR > 1 signifying higher prevalence in the case group. The power to detect significant effects depends on the sample size, distribution of effect sizes of causal genetic variants, and the frequency of these in the population, and the linkage disequilibrium (LD) between the observed genotyped DNA variants and the unknown causal variants. GWAS approaches have also been applied to quantitative phenotypes, such as height or concentration of given biomarkers.

The first GWAS, published in 2005, focused on age-related macular degeneration (AMD). In a comparison of 96 cases and 50 controls, Klein *et al.* identified a role of the CFH gene in AMD [11]. A major breakthrough followed in 2007 with the publication of the Wellcome Trust Case Control Consortium [12], in which 3000 shared controls were compared with around 2000 patients for each of seven common disease phenotypes. Not only was this the largest study of its kind at the time, but it also set the precedent for future GWAS studies in a number of ways. Population stratification was carefully considered, HapMap data was used for genotype imputation in a novel manner [13], and significant attention was given to genotype calling.

Since then, the number of GWAS has increased year-on-year, and vast progress has been made in identifying and understanding genetic variation in the human population. However, the nature of the studies above means that the findings often do not reveal the mechanistic basis or causative role of genetic variants, as the causal variant is usually not directly genotyped but rather in linkage disequilibrium with the genotyped SNPs. This necessitates methods to move closer to an understanding of the biology underlying a process or phenotype of interest caused by observed genetic differences (Figure 1.2b).

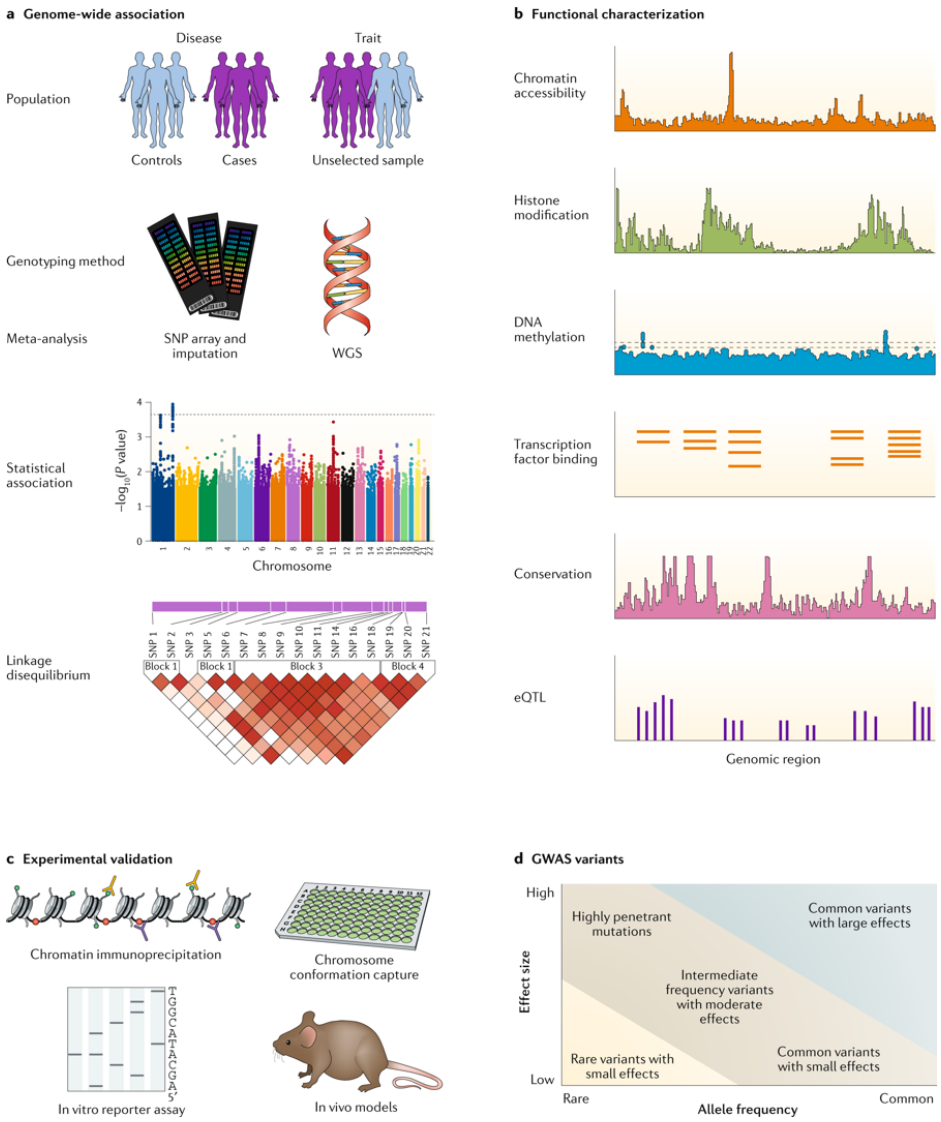


Fig. 1.2 Genome-wide association study design, from Tam *et al.* [14]: a) The aim of a genome-wide association study (GWAS) is to detect associations between allele or genotype frequency and trait status. The first step is to identify the disease or trait to be studied and select an appropriate study population. Genotyping can be performed using SNP arrays combined with imputation or whole-genome sequencing. Association tests are used to identify regions of the genome associated with the phenotype of interest at genome-wide significance, and meta-analysis is a common step to increase the statistical power to detect associations. b) Functional characterization of genetic variants is often required to move from statistical association to causal variants and genes, especially in the non-coding genome. Computational methods are used to predict the regulatory effect of non-coding variants on the basis of functional annotations. c) Target genes can be identified or confirmed using chromatin immunoprecipitation and chromosome conformation capture methods, and experimentally validated using cell-based systems and model organisms. d) Genetic variants exist along a spectrum of allele frequencies and effect sizes. Most risk variants identified by GWAS lie within the two diagonal lines. Rare variants with small effect sizes are difficult to identify using GWAS, and common variants with large effects are unusual for common complex diseases.

A key example is the expression quantitative trait loci (eQTL) approach, in which SNPs driving differences in expression levels of particular genes are identified. These eQTLs can be described as acting in *cis*, typically considered with a 1Mb window, or in *trans* from a more distant genomic location, typically 5Mb or further, or on a different chromosome entirely (Figure 1.3). By studying the transcription of genes, captured in RNA sequencing experiments, a more direct output of genetic variation can be captured. This intermediate phenotype can explain cellular events at a level closer to mechanism, uncovering novel biological insight into the disease or process of interest.

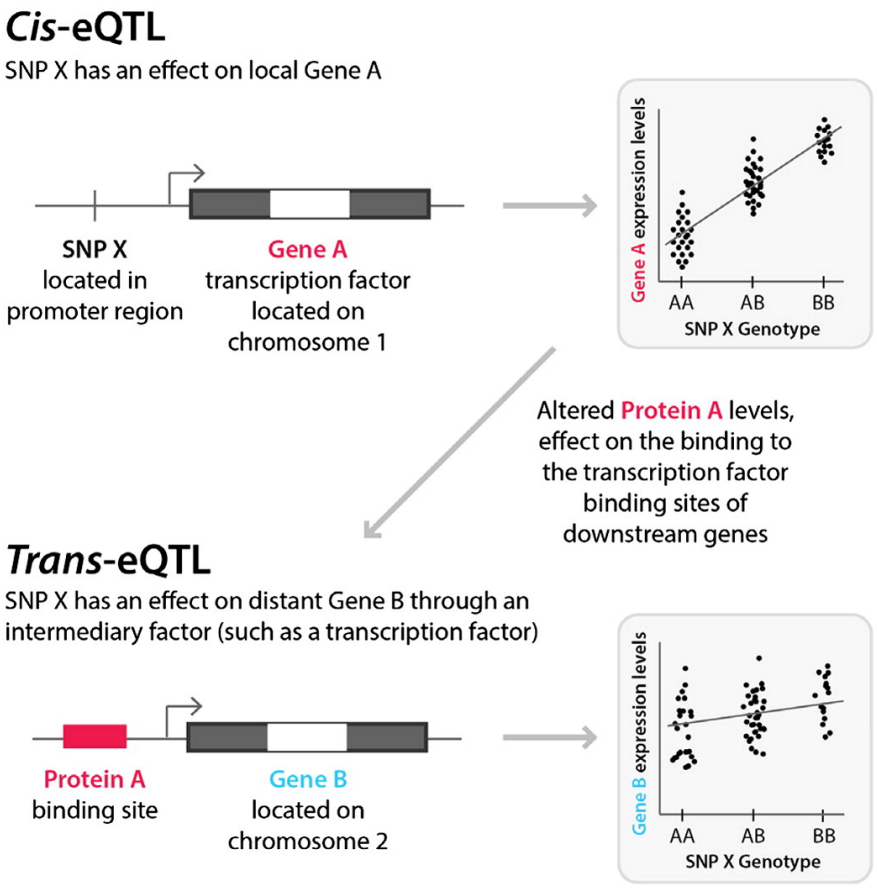


Fig. 1.3 The expression quantitative trait loci (eQTL) approach, from Westra & Franke [15]: eQTLs can be either local effects (cis-eQTLs), or distant, indirect effects (trans-eQTLs).

1.2 Single cell RNA sequencing

1.2.1 Evolution of single cell RNA sequencing technologies

While transcriptomic studies have, for many years, provided insight into mRNA expression and regulation, technological advances have allowed the quantification of transcripts at an unprecedented resolution. By sequencing the mRNA component of individual single cells, it has now become possible to study gene expression at an entirely new level, opening the door to novel biological questions which could not be addressed using population-level RNA sequencing. For example, the variability in splicing [16–20] and allelic expression [18, 21–23], between cells has been shown, along with analysis of the stochastic gene expression and transcriptional kinetics [24, 25]. Furthermore, single-cell RNA-sequencing (scRNA-seq) data have allowed fine-grained analysis of developmental trajectories [26–28] and identification of rare cell types [29, 30].

In order to obtain scRNA-seq data, cells must first be isolated individually in an accurate and rapid manner. Initially, microscopic manipulation provided a reliable method to isolate single cells through physical separation using a capillary pipette, and may still play an important role in systems where few cells are available. However, the high labour and low-throughput nature of this technique has resulted in it being surpassed by higher throughput methods. Fluorescence-activated cell sorting (FACS) provides an efficient way to isolate a large number of cells in a rapid manner, and also allows the selection of cells based on fluorescent labelling. Size or marker selection is commonly used, and through 'index sorting', the data for each cell can be recorded and used in downstream analysis. Despite the prevalence of this method, the high number of starting cells required, along with the potential damage caused by the staining and physical stress of the process, means it may be a problematic approach. More recently, microfluidic techniques have emerged as a key method for capturing

single cells, allowing isolation in small volumes within a closed system, often followed directly by amplification and downstream reactions. The small volume in which these reactions occur increases the capture efficiency and lowers the reagent cost. Finally, techniques involving the isolation of single cells in microdroplets, such as DropSeq [31] and InDrop [32], have rapidly expanded the high-throughput nature of scRNA-seq–allowing processing of tens of thousands of cells in a short space of time. The small volume of reactions, once again, decreases the cost per cell. Over time, these methods will continue to increase in speed, efficiency and reliability, further improving throughput of single-cell isolation.

Many protocols exist for the subsequent reverse transcription (RT), amplification, and library preparation prior to sequencing. Poly(T) priming is used to select polyadenylated mRNA for reverse transcription, however, only an estimated 10–20 percent of transcripts are sampled, particularly affecting lowly expressed genes [33]. Methods then differ in their approach to second-strand synthesis, either using poly(A) tailing, leading to a 3' bias, or template-switching to produce full-transcript coverage. Amplification can be achieved through two methods: linear *in vitro* transcription (IVT) or exponential PCR, each with its own advantages and drawbacks. Ziegenhain *et al.*[34] and Svensson *et al.*[35] provide a comprehensive experimental and computational comparison of most of the protocols commonly used. Following cDNA amplification, library preparation is most commonly carried out using the commercially available Nextera kit and sequencing on the Illumina platform, although other methods are available.

As a relatively new field, it is key to understand the structure and complexities of scRNA-seq data, ensuring that appropriate analytical and statistical methods are applied [36]. Particularly challenging is the high level of noise [37, 38], which derives primarily from the nature of single-cell experiments (called 'technical variation' and is mainly due to factors such as mRNA capture efficiency and cDNA amplification bias),

along with the biological heterogeneity of cells ('biological variation'). Furthermore, unlike with conventional RNA-sequencing where experimental biases are well studied [39, 40], there are biases which are still not fully understood in single-cell experiments, such as 'dropouts' due to the low amounts of starting material, leading to false negative expression.

Single-cell RNA-sequencing is a lossy technique, and it is not completely understood what causes the different failure modes for samples. Practically, this means the first step after acquiring reads from a scRNA-seq experiment is to perform quality control. Reads are processed in a similar manner to bulk RNA-seq, allowing expression quantification. There are several methods to do this, broadly split into those that use a genome reference for alignment, such as STAR [41], TopHat/TopHat2 [42, 43] and HISAT/HISAT2 [44, 45], and those that perform 'pseudoalignment', a quicker alternative, such as Kallisto [46] and Salmon [47].

It is important to check the quality of both the raw data (which can be performed using tools developed for bulk RNA-seq, such as FastQC [48] or Kraken [49]), along with the aligned output. Imperative in scRNA-seq is the cell-by-cell quality control [50], ensuring that cells of poor quality are removed from subsequent analysis. Many metrics can be used to measure cell quality, such as the number of reads or genes detected, the proportion of reads mapping to mitochondrial genes (which may signify leaking of cytoplasmic RNA or cells undergoing apoptosis), or the proportion of reads mapping to externally spiked-in RNA molecules if used in the experiment [51].

Depending on the analysis task, appropriate normalization of the data is needed. Several normalization methods have been developed, many of which adjust for differences in sequencing depth and/or make use of spike-in molecules and/or unique molecular identifiers (UMIs) when available (reviewed in detail in [52]). Once cleaned data are obtained, there are many routes of analysis depending on the biological

question under investigation (Figure 1.4). In the next section, I will consider these analysis from two viewpoints: cell-level approaches, such as the grouping of cells and trajectory ordering, along with gene-level investigations, such as gene variability and noise, co-expression, and identification of differentially expressed genes.

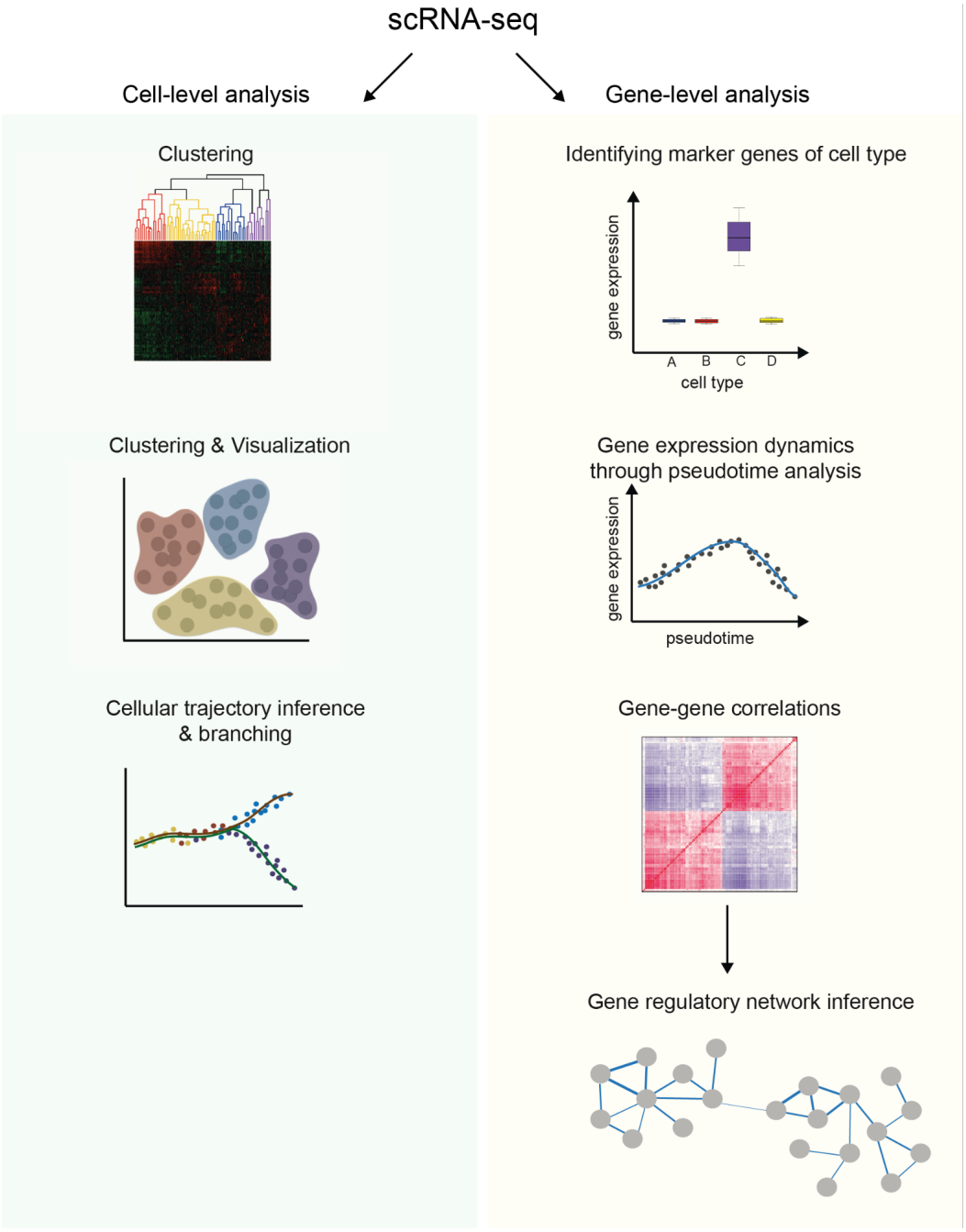


Fig. 1.4 Overview of analysis methods for the interpretation of scRNA-seq data.

1.2.2 Analysis of scRNA-sequencing data

Cell level analyses

Visualising and clustering cells

The cataloging and classification of cells is a long-standing biological challenge. Traditionally, cell types were determined morphologically or based on molecular cell surface markers. However, with the availability of genome-wide expression data, the possibility of transcriptome-based analysis of cell similarity provides an alternative indicator of cell type.

The first step in understanding the distribution of cells is often to apply dimensionality reduction techniques: this represents the thousands of dimensions (genes) found in scRNA-sequencing data with a much smaller number, attempting to maintain a representation of some variation of interest. Furthermore, by considering only a two or three dimensional space, visualisation provides a mean to qualitatively explore the data. There are hundreds of dimensionality reduction methods available which the researcher can elect to apply either to all observed genes or a selected subset of genes of interest. The most widespread is Principal Component Analysis (PCA) [53], where weighted sums of dimensions represent the data. The dimensions for each sample are known as principal components. These dimensions explain decreasing amounts of variation in the original data, with the first principal component capturing as much of the variance as possible. Another commonly applied method is t-SNE (t-Distributed Stochastic Neighbour Embedding) [54], a non-linear visualization technique which considers local distances between data points (cells) by combining dimensionality reduction with random walks on the nearest-neighbour network with the goal of separating far-apart clusters, while also ensuring all data points can be seen by eye to allow for comparisons of cluster size. This is a variation of Multidimensional Scaling (MDS), where PCA is applied on pairwise Euclidean distances to preserve pairwise distances

in a low-dimensional space. A recent, and increasingly adopted method, is uniform manifold approximation and projection (UMAP) [55], which has been shown to preserve more of the global structure within datasets, with an improvement in run time and reproducibility [56].

While powerful, and popular, these techniques can be heavily affected by the problematic abundance of zeroes in single cell data; an issue which several methods account for. ZIFA (zero-inflated factor analysis) [57] extends the linear factor analysis framework, (based on correlations in the data rather than covariances), accounting for dropout characteristics in the data. The R-package Destiny provides an alternative, non-linear method using diffusion maps [58]: distance between cells reflects the transition probability based on several paths of random walks between the cells. This assumes a smooth nature of the data, and also includes imputation of drop-outs.

Unsupervised clustering techniques provide a mechanism to group cells by similarity. While this unbiased approach has benefits, the small number of samples and absence of a way to validate if groupings are "real" poses a problem, along with prior information on the number or type of groups. The features of single cell data discussed above, such as dropouts, biases and noise, also add to the difficulty of accurate clustering. Despite these problems, several tools have been developed for use with scRNA-seq, along with traditional methods such as hierarchical clustering [59]. SNN-Cliq [60] achieves clustering by considering similarity calculated using a graph-based approach in which a shared nearest neighbour (SNN) network is constructed using rankings of similarities based on expression levels; dense clusters of nodes (cells) are then found. RaceID [29], while also using similarity in expression between cells (based on Pearson correlation), utilises a different approach: k-means clustering. In k-means clustering each sample is associated with a one of k prototypes, so that the total squared distance (inverse of similarity) from samples to prototypes is minimal. After the initial step, RaceID uses

an outlier detection algorithm and identifies cells which do not fit the model accounting for technical and biological noise. This has been used in the detection of rare cell populations. Another k-means-based tool, Single Cell Consensus Clustering (SC3) [61], uses consensus clustering [62], an ensemble strategy, to average over parameter choices in an attempt to make cluster assignments more robust. Another method, SIMLR [63], uses multiple-kernel learning to infer similarity in a gene expression matrix with a given number of cell populations. As multiple kernels are used, it is possible to learn a distance measurement between cells that is specific to the statistical properties of the scRNA-seq set under investigation. Two widely adopted strategies using a community detection approach are Louvain [64] and Leiden [65] clustering, In the first method, clusters are identified by moving nodes individually between groups until the quality of clusters can no longer be improved. The network is then aggregated, with each cluster becoming a node, and the steps of node movement and aggregation repeated. While this leads to an efficient approach, clusters may be badly connected - a problem which the Leiden method tackles by improving upon the aggregation step, allowing clusters to be split.

Cellular trajectory inference and branching analysis

Trajectory analysis is a simpler version of dimensionality reduction, where the assumption is that a 1-dimensional "time" can describe the high-dimensional expression values. The theory is that during a biological process, changes will happen gradually, so biological observations can be ordered compared to each other in terms of pairwise similarity. While clustering techniques have been used to define discrete population and states for a long time, trajectory inference is younger in the field of scRNA-seq. However, many methods have been developed in recent years, and Saelens *et al.* recently conducted a comprehensive benchmarking of 45 of these methods [66]. Here, just a subset of methods are described.

One of the initial methods for so called pseudotime analysis of single cells was Monocle [67], which used a minimum spanning tree (MST) strategy to order cells by the distance to a start cell, based on a technique for putting microarray samples on a trajectory [68]. In the updated versions of Monocle, the MST strategy has been replaced by a more sophisticated tree embedding strategy [69, 70].

Diffusion pseudotime (dpt) [27] offers an alternative, in which geodesic pairwise distances between samples on the data manifold are approximated using a diffusion map representation. Trajectory is then defined as the distance from a start cell along these distances. A different strategy for trajectory inference is to consider a generative model for the data, treating "time-points" as hidden (or latent). This leads to the probabilistic interpretation of PCA, which in turn leads to factor analysis and ZIFA. Here the expression of each gene can be described as a linear function of an unknown "time".

Non-linearity in the data, as described in [67] precludes PCA from being an effective technique for this task. The Gaussian Process Latent Variable Model (GPLVM) allows gene expression to follow any smooth (non-linear) function over time [71]. While more computationally demanding than linear versions, this allows cells to be put in the most likely ordering [71, 72]. This means that the most number of genes exhibit smooth expression curves with as little noise as possible. Being a probabilistic model, the benefit is that uninteresting structure in the data can be accounted for directly, such as batch effects or technical factors. It is also possible to incorporate more information about experimental design through priors [28].

The Ouija method [73] takes a different approach to pseudotime in a couple of ways. Firstly, it defines a generative model for gene expression in scRNA-seq data based on ZIFA, to deal with the most common types of measurement noise. Secondly, it is based on the assumption that a small number of switch-like markers for a biological process

of interest are known. The cells are then ordered according to the most likely ordering to confer with the switching genes.

A unique problem in single cell developmental data is that a set of progenitor cells can develop into multiple distinct cell types. This means the cells will not follow a single trajectory in the high-dimensional space. A couple of heuristics have been published: in Wishbone [74], cells are clustered by the pairwise detour distance relative to a reference cell, using geodesic distance. This method is reported to be correctly recovering the known stages and bifurcation point of T-cell development in mouse. Another method, that has been introduced by Haghverdi *et al.* [27], measures transition between cells using a random-walk-based distance.

More principled model based approaches have been presented with SCUBA, which considers transition of cell clusters over time [75], as well as with GPfates / OMGP [28], where multiple smooth trajectories are explicitly modeled. After inference, each cell gets assigned a posterior probability of having been sampled from a particular trajectory. This method has been shown to be efficient in reconstructing the developmental trajectories of Th1 and Tfh cell populations during Plasmodium infection in mice.

An interesting recently developed method, partition-based graph abstraction (PAGA), generates a graph-like map, estimating connectivity of partitions in the data [76]. This approach provides a way to bridge the clustering type of analysis, as discussed above, with the continuous nature of many biological processes, as modelled with conventional pseudotime approaches.

Gene level analyses

Unwanted factor removal

Uninteresting, largely technical variation can be observed in both bulk RNA-seq and scRNA-seq experiments. This variation is usually correlated with some common experimental factor, such as room temperature or stock of reagents. This form of variation are known as batch effects. It is possible to handle batch effects by having a careful balanced experimental design, such as uniformly distributing replicate conditions across batches. For statistical analysis and inference, if the samples are spread over multiple batches, this information can directly be accounted for [77]. Additionally, several statistical methods have been developed to adjust for batch effects [78, 79]. One example is ComBat, which removes known batch effects using a linear model of expression from batches where variance is based on an empirical Bayesian framework [78].

Technical variation in scRNA-seq experiments could be due to mRNA capture efficiency, cDNA amplification bias and the rate cDNAs in a library are sequenced. To estimate technical variation, several methods use spike-in molecules, which are added with each cell in the same quantity. Risso *et al.* developed a sleuth of strategies called RUVSeq that either performs factor analysis on a set of control genes such as ERCC spike-ins or samples within replicate libraries to identify technical factors which can be adjusted for [80]. Similar strategies have also been made by others [81–83].

Substantial amount of variation also results from differences in cell size or cell cycle stage of each cell. To adjust for cell cycle effects, Buettner *et al.* have developed single-cell latent variable model (scLVM), which is a two-step approach that reconstructs cell cycle state before using this information to obtain adjusted gene expression levels by linear regression [84]. They have also shown that removing cell cycle effects in T cells reveals sub-populations associated with T-cell differentiation [84]. This highlights the

importance of dissecting biological variation into interesting and uninteresting parts in correctly characterizing sub-populations.

In recent years, many further methods have been developed for the integration of discrete experimental batches. One example is canonical correlation analysis (implemented in Seurat [85], which identifies a shared gene correlation structure across datasets, using this structure to align the datasets. Haghverdi *et al.* developed a mutual nearest neighbours (MNN) approach [86] to correct expression between batches. This method uses 'landmark' cells, which are representative of cell types or clusters across all datasets to be integrated. Park *et al.* provide an alternative approach, using a batch balanced k nearest neighbour graph (BBKNN) [87] to combine batches. While these examples highlight just some of many methods available, there will undoubtedly be further work in this area, particularly given the increasing scale of scRNA-seq data generated and desire to integrate across experiments.

Identification of highly variable genes

Several methods have been developed to identify genes that show high biological variability. Brennecke *et al.* have first estimated technical noise using spike-in molecules and modeled mean-variance relationship to identify highly variable genes [37]. Kim *et al.* have presented a statistical framework to decompose the total variance into the technical and biological variance based on a generative model, which would help in identifying variable genes [22]. Another method, BASiCS, uses a Bayesian model which jointly models spike-ins and endogenous genes and provides posterior distributions for the extent of biological variability [88].

Identification of differentially expressed genes and marker genes

Identification of differentially expressed genes and marker genes of subpopulations is a simple yet important analysis in scRNA-seq studies. Although originally developed for bulk RNA-seq experiments, methods such as DESeq2 [89] and EdgeR [90] are also

widely used in scRNA-seq experiments. DESeq2 identifies differentially expressed genes by fitting a generalised linear model (GLM) for each gene, uses shrinkage estimation to stabilize variance and fold changes, and applies a Wald or likelihood ratio (LR) test for significance testing [89]. EdgeR fits a GLM with negative binomial (NB) noise for each gene, estimates dispersions by conditional maximum likelihood, and identifies differential expression using an exact test adapted for overdispersed data [90]. Monocle also fits a GLM, but dispersion is estimated directly from the data for each gene, since most single cell studies have enough samples to allow this [67]. For relative abundance data, dropouts are handled by using a tobit noise model, while using a NB noise model with imputed dropouts for count data.

One method developed for scRNA-seq experiments, called MAST, uses two-part generalized linear model that is adjusted for cellular detection rate (dropouts) [91]. Another method, M3Drop, applies Michaelis-Menten modelling of dropouts in scRNA-seq, that is used to identify genes differentially dropped out [92]. SCDE is a Bayesian method to compare two groups of single cells, taking into account variability in scRNAseq data due to drop-out and amplification biases and uses a two-component mixture for testing for differences in expression between conditions [93]. Another method, SINCERA identifies differentially expressed genes based on simple statistical tests such as Wilcoxon rank sum and t-tests [94]. In comparison to these methods, scDD identifies genes where the overall distribution of values have changed between conditions. This answers a different question which might be of interest in scRNA-seq experiments [95]. Using a Bayesian modeling framework, scDD classifies each gene into one of the four types of changes across two biological conditions: shifts in unimodal distribution, differences in the number of modes, differences in the proportion of cells within modes, or both differences in the number of modes and shifts in unimodal distribution [95].

Gene-centric expression dynamics through pseudotime analysis

Using an inferred trajectory as described above, samples can be analysed using a continuous time covariate instead of a few discrete time points. This enables the use of more sophisticated time-series based analysis techniques for modeling gene expression dynamics, and allows us to ask more complex questions from the data.

The popular scRNA-seq package Monocle provides a wrapper for the vector generalised additive model (VGAM) package to investigate how expression changes over the trajectory. Splines are used to model expression dependence on pseudotime to allow non-linear trends. The VGAM package allows for more than just expression levels to be modelled by the splines: with appropriate link functions, allelic expression balance or isoform usage can be modelled [18]. Splines require several parameters to be chosen however, and the choices greatly affect the results. A non-parametric non-linear alternative to spline regression is Gaussian Process regression, which can be used in a likelihood ratio based fashion to identify genes which are dependent on pseudotime [71, 96].

Often, we want to ask particular questions from the data, in which case parametric models are useful. In the SwitchDE method, genes which sequentially switch on or off can be identified, along with a parameter letting you learn when the switch happens [97]. Similarly, an assumption can be that genes are described as a transient pulse over the pseudotime. The package ImpulseDE identifies such genes, while providing parameters for when in pseudotime the pulse occurs [98].

Correlation analysis and network inference

One important application of scRNA-seq studies is the identification of co-regulated modules of genes and gene-regulatory networks constructed using gene-to-gene expression correlations. Here, genes with highly correlated expression levels across cells are assumed to be co-regulated. Using single-cell transcriptomic data of Th2 cells, Mahata

et al. demonstrated how gene-gene correlations can be used to reveal novel mechanistic insights; they have applied correlation analysis between steroidogenic enzyme Cyp11a1 and cell surface genes and identified Ly6c1/2 as a marker of the steroid-producing cell population in mouse [99].

One method to elucidate regulatory interactions in bulk RNA-seq studies is called the weighted gene co-expression network analysis (WGCNA) [100]. In such a network, nodes represent genes and edges represent co-expression as defined by correlation and relative interconnectedness. The method has also been applied in a scRNA-seq study where the authors have identified a number of functional modules of co-expressed genes that can describe each embryonic developmental stage in mouse [101].

Although these methods are useful, the inferred networks are undirected; that is, they do not provide direct regulatory relationships among genes. One method, SCENIC, aims to address this by constructing gene regulatory networks from scRNA-seq data [102]. SCENIC defines co-regulated modules, or 'regulons', using GENIE3 [103] to identify the targets of transcription factors, and cis-regulatory motif analysis.

1.3 The human innate immune system

The innate immune system is the body's first line of defence against damage, rapidly sensing and responding to infectious or harmful agents. Due to the diversity of potential threats, a range of mechanisms are utilised to act against infections. These are prompted by detection of pathogen-associated molecular patterns (PAMPs) - conserved structures which are predominantly expressed by large groups of pathogens, rather than the host. One major group of PAMPs is nucleic acids (Gurtler Bowie, 2013). Although there may be complexity in detecting various RNA and DNA structures due to the similarity with host nucleic acids, their essential role for pathogen survival means they are a useful signal, particularly for viruses in which there may be a limited amount of alternative distinguishing molecular features.

1.3.1 The type I interferon response

To detect these pathogenic signals, there are several classes of pattern recognition receptors (PRRs) in various cytoplasmic or membrane-bound locations. These include Toll-like receptors (TLRs), RIG-I-like receptors (RLRs) and NOD-like receptors (NLRs), among others. These receptors may function in signalling - activating downstream pathways to instigate a response - or have a direct effector function, blocking pathogenic replication and propagation [104]. In the case of viral infections, distinct sensors play a role in the recognition of viral RNA and DNA. In the case of RNA, RIG-I and MDA5 sense cytosolic non-self RNA, with specificity towards differing lengths of dsRNA [105]. In contrast, the presence of viral DNA is sensed through cGAS, leading to the production of cGAMP and consequent activation of STING [106].

Despite specific recognition pathways, activation of viral sensors converges in downstream signalling, leading to activation of NF-κB, TBK1 and IRF3 to induce

production of type I interferons (IFNs). In this thesis, the response to RNA viruses is studied, using synthetic dsRNA to mimic the presence of viral nucleic acids in host cells. The induction of the type I interferon response through dsRNA sensing is shown in Figure 1.5.

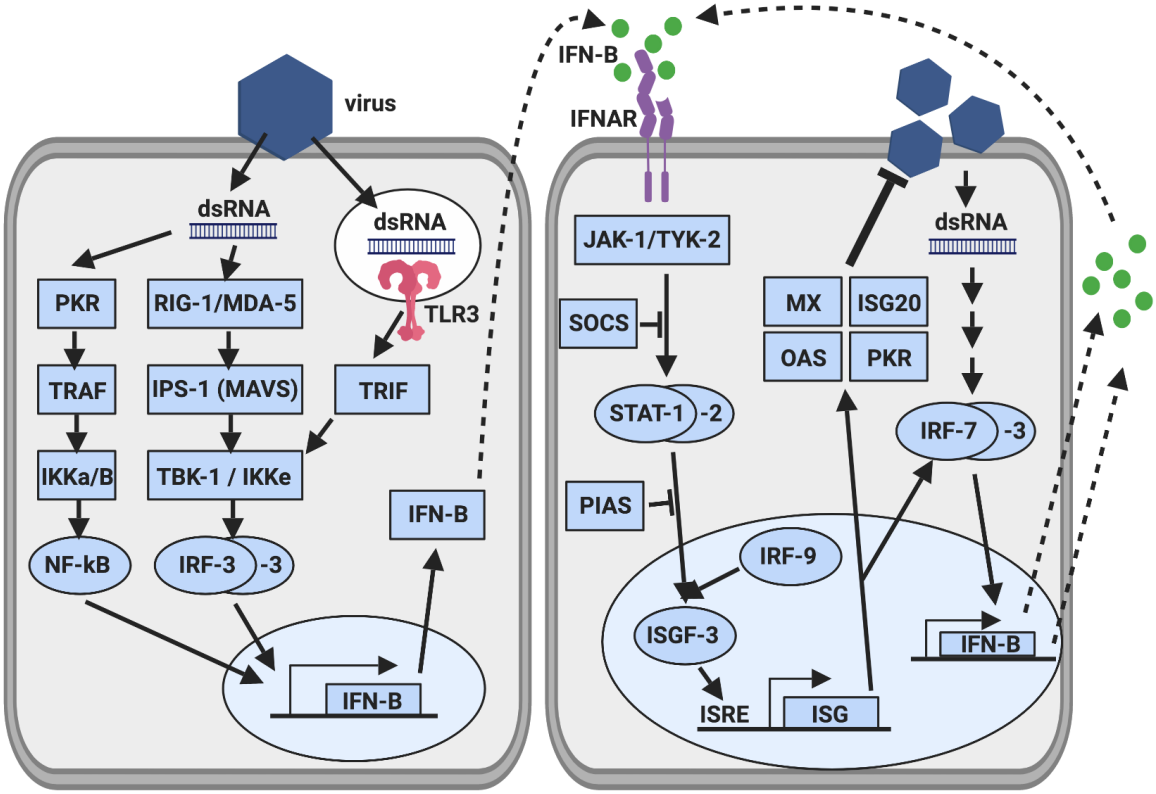


Fig. 1.5 Induction of the Type I Interferon response.

Interferons are a subset of the class of immune signalling cytokines, and can be subdivided into type I (IFN-α, IFN-β, IFN-ϵ, IFN-κ, IFN-ω), type II (IFN-γ) and type III (IFN-λ), based upon receptor specificity [107]. Of the type I IFNs, which bind the heterodimeric IFNAR1-IFNAR2 receptor, IFN-α and IFN-β are the most studied. There are 14 IFN-α genes and only one IFN-β gene in humans. When bound to type I IFNs, the IFNAR1-IFNAR2 heterodimer activates JAK1 and TYK2 [108], leading to phosphorylation of STAT1-STAT2 heterodimers [109]. Consequent migration into the

nucleus, association with IFN regulation factor 9 (IRF9) and binding to IFN-stimulated response elements instigates transcription of IFN-inducible genes.

Type I interferons play an important role in the response to viral infections (reviewed in [110]), and are able to be produced at low levels by most cell types. Certain cells have been shown to function in producing high levels of these proteins, invoking a systemic response. Plasmacytoid dendritic cells (pDCs), for example, were identified as 'natural interferon producing cells' [111]. However, fibroblasts mainly produce IFN-β, considered the central cytokine responsible for stimulating cells locally. This leads to altered gene expression, chemokine production, antigen presentation and the induction of an adaptive immune response (Figure. 1.6).

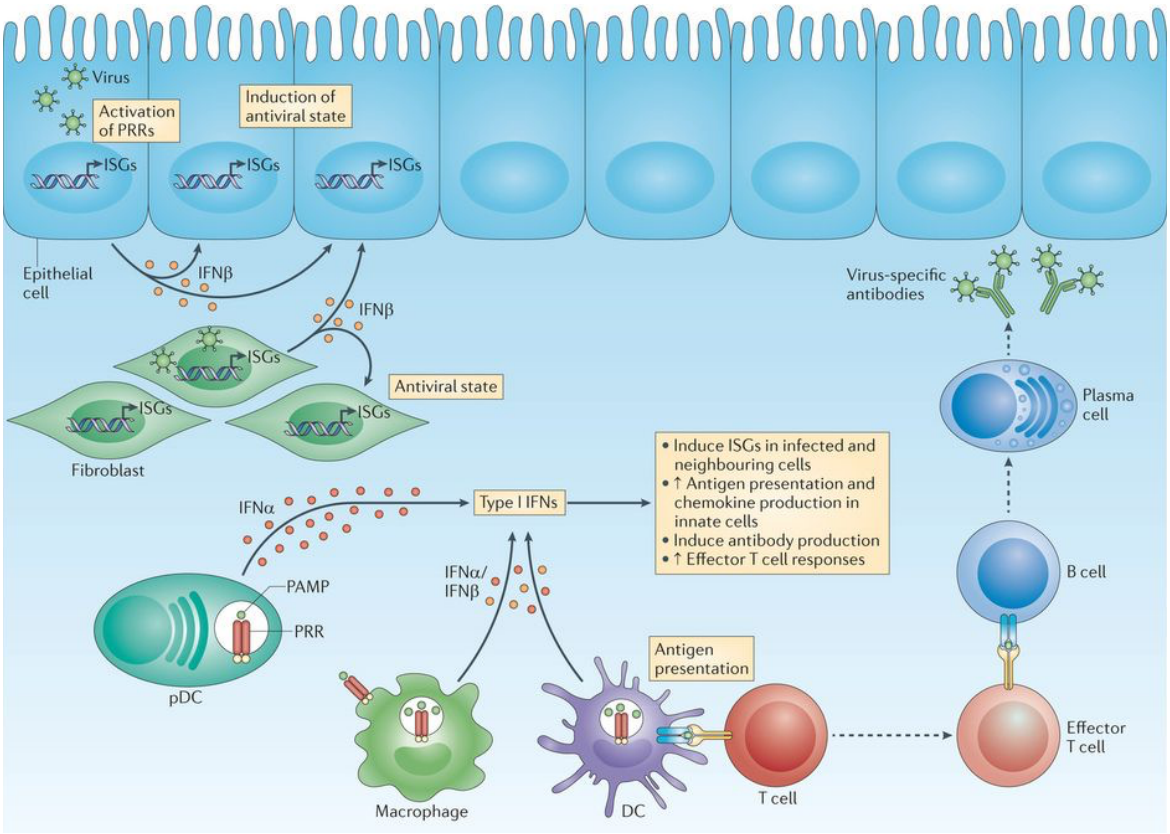


Fig. 1.6 The role of interferons and inter-cellular communication in the immune response. (Ivashkiv Donlin, 2014 [112])

1.3.2 Cell-to-cell heterogeneity in innate immunity

As with many biological processes, the innate immune response is considerably heterogenous between cells in an individual, despite cells being genetically identical. This derives both from the stochastic nature of biochemical reactions within cells (intrinsic) along with communication between cells and other environmental factors (extrinsic). Fluctuation in gene expression is one of the largest causes of variation in clonal populations, reviewed in [113]. This is due to the low molecular abundance of some key elements (such as transcription factors) along with the number of chemical reactions required to turn genetic sequence to functional product. Although modelled for many systems, a severely reduced view is often taken, with only transcription and translation included.

Within immunology, stochastic expression of many interleukin genes has been observed, such as IL-2 [114] and IL-10 [115]. In fibroblasts, large variation in the induction of IFN-β in response to viral infections has been shown [116]. Furthermore, heterogeneity in response of coarse-grained cell populations has been studied, such as macrophages [117, 118] and monocytes [**?**]. The considerable advance in single-cell technologies in recent years, however, will allow further illumination of inter-cellular variability in innate immunity. For example, scRNA-seq was recently used to reveal novel dendritic cell and monocyte sub-populations [119]. However, scRNA-seq holds exciting possibility not only for cell classification, but also in understanding the innate immune response. One example is the discovery of bimodal transcript splicing in bone-marrow derived dendritic cells in response to lipopolysaccharide (LPS) treatment [16].

1.3.3 Genetic variability in the innate immune response

The hereditary nature of some susceptibility to infectious diseases has long been known, alongside the variation between individuals in the response to particular pathogens. Early investigations involved twin studies, which showed a higher concordance in identical to non-identical twins for some infections, particular those which are chronic and have low infectivity. Examples of these findings for viral infections include poliomyelitis and hepatitis B [120]. More recently, the development of high throughput technologies, as described above, has enabled the identification of single nucleotide polymorphisms (SNPs) associated with particular infections. Just two examples of many available are Hepatitis C clearance, for which a SNP in IL28B has been identified [121], and reduced influenza virus clearance (a SNP in IFITM3; [122]).

Unlike adaptive immunity, in which receptor sequences undergo rearrangement in somatic cells, the innate immune system's pattern recognition receptors are germline encoded, along with signalling components and effector mechanisms. Therefore all aspects of innate immunity, from recognition to action, are likely to be subject to genetic variation. Genetic analysis of patients with susceptibility to particular infections has pinpointed elements of the innate immune, and more specifically type I interferon, response as playing a key role. For example, Zhang *et al.* [123] described two children with herpes simplex encephalitis, both with a heterozygous mutation in TLR3. In dermal fibroblasts from these individuals, treatment with a synthetic dsRNA (polyinosinic:polycytidylic acid; also known as poly(I:C)) did not induce expression of IFN-β, IFN-γ or IL-6. More recently, Ciancanelli *et al.* [124] characterised a patient with compound heterozygous null mutations in interferon response factor 7, who suffered a life threatening primary influenza infection. In this case, dermal fibroblasts and iPSC-derived epithelial cells from the patient produced reduced amounts of type I IFN and showed increased viral replication. There have been further studies showing

deficiency in innate immune signalling pathways in the fibroblasts of affected individuals, such as those with an IRAK1 [125] or DOCK2 [126] mutation.

Moving beyond individuals with a deficient innate response or specific susceptibility, expression quantitative trait loci (eQTL) approaches been used to characterise genetic variability within healthy populations. Some studies have identified SNPs in particular mechanisms, such as the TLR4 pathway [127]. In recent years, however, investigations have expanded from studying one pathway or pathogen to eQTL mapping in broader innate immune stimulation. Two studies in which this has been conducted are Fairfax *et al.*, 2014 [128], where primary CD14+ monocytes were treated with IFN-γ or LPS, and Lee *et al.*, 2014 [129], in which dendritic cells were stimulated with influenza virus, LPS, or IFN-β. These studies identified treatment-specific eQTLs, highlighting the importance of considering genetic variation within the biological context of interest. However, in these studies changes in expression were measured only at a cell population level and at distinct time points. Further insight is needed into the genetic effect on variability of innate immune components, gained from single-cell expression studies, along with the dynamics and regulation of the response.

1.4 Using single-cell RNA sequencing data to study genetic variation in the innate immune response

While there have been significant advances in understanding the genetic basis of variation in the innate immune response in recent years, there is still a way to go in defining the intra- and inter-individual components of this variability, particularly in healthy individuals. In order to do this, a dataset spanning a large number of donors, profiled at single cell resolution, is required. To this end, this thesis outlines the establishment of an experimental system using relatively homogenous dermal fibroblast populations of 70 human individuals obtained from the Human Induced Pluripotent Stem Cell Initiative (HipSci). Assaying these cells using two stimulation conditions - a synthetic dsRNA, and Interferon-β - over time allows us to study key questions:

(1) How does the interferon response vary between human individuals and can this variation be attributed to common genetic variants?

(2) How do different cells from the same donor respond to a danger signal that should elicit interferon, and how do they respond to a direct interferon stimulus?

Alongside this, the heterogeneity in unstimulated human fibroblasts is characterised, to understand the variation and clonality seen in genetically identical populations of fibroblasts. The single-cell resolution provides unprecedented insight into not only the human genetics of the innate immune response, but also the role of cell-to-cell variation in this response.

Chapter 2

Establishing a system for studying innate immune responses in human fibroblasts

> ***Declaration***
>
> *Bulk RNA sequencing data for protocol optimisation was generated and processed by Tzachi Hagai.*
>
> *During the expansion and stimulation of HipSci lines, invaluable assistance was provided by the Cellular Genotyping and Phenotyping facility at the Wellcome Sanger Institute (WSI).*
>
> *Data processing was conducted with the help of the Cellular Genetics Informatics group, WSI, and Davis McCarthy, EMBL-EBI.*

2.1 Defining optimal stimulation conditions

It was first shown many years ago that synthetic dsRNA - polyinosinic:polycytidylic acid, also known as poly(I:C) - could induce an antiviral response and interferon production in treated cells [130, 131]. However, there are many factors in the stimulation procedure which may affect the response, such as concentration of poly(I:C) used, time after stimulation, and reagents used. Furthermore, the effect of these variables may differ between cell types[132]. It is also possible to induce interferon signalling more directly through administration of interferons, capturing the second cascade of signalling and removing the effect of upstream PAMP sensing.

In order to determine the effects of differing stimulation conditions in human fibroblasts, and enable optimisation of large-scale experiments, bulk RNA-sequencing data generated by Tzachi Hagai was analysed. In these experiments, primary human fibroblasts (HipSci resource) were either stimulated directly with poly(I:C) or interferons. Fibroblasts were seeded approximately 18 hours prior to stimulation, at a density of 100,000 cells per well (6 well plate) or equivalent numbers on smaller plates (12 well and 24 well plate). Cells were cultured either in specialised fibroblast medium (ATCC-PCS-201-041), or in alpha-MEM supplemented with 10% FBS, non-essential amino acids, vitamin C and L-glutamine. In order to achieve sufficient intracellular levels of poly(I:C), addition with lipofectamine 2000 (LF) in a transfection medium (opti-MEM) was used, at a ratio of 1 μg poly(I:C) : 2 μg LF : 100 μg opti-MEM. Interferon was added directly at a concentration of 1000 U/ml. At stated time points (1, 2, 3, 4, 6, 8, 12, 18 or 24 hours) post-stimulation, cells were lysed with RLT buffer containing 1% β-mercaptoethanol, and collected. Library preparation was performed according to Illumina Truseq/KAPA protocols and samples sequences using Illumina Hi-Seq (125bp paired-end sequencing).

The raw RNA-seq data was processed using two pipelines, either mapping with TopHat2 and using cuffLinks to quantify reads in order to identify differentially expressed genes with the cuffDiff tool, or mapping with Kallisto for analysis with Sleuth.

TopHat2 and CuffLinks

Reads were first mapped to the human reference genome (hg37) using TopHat2 [43], before calculating a normalized count for each gene (fragments per kilobase per million, FPKM) using cuffLinks [133]. Differentially expressed genes were identified using the cuffDiff command. These stages were performed by Tzachi Hagai, using default parameters. For analysis of patterns of expression, FPKM was first converted to TPM using the relationship derived by Lior Pachter [20, 134]. A threshold for TPM expression of 2 was chosen, and any transcripts which had expression below this across samples were discarded. The basis for this threshold was the finding by Wagner et al., [135], that RNA-seq data can be modelled as a mixture of two distributions: an exponential distribution for transcripts from inactive genes and a negative binomial distribution for actively transcribed genes. It is shown that the probability of TPM 2 for the exponential distribution (inactive genes) is $< 10^{-8}$, while the probability of a gene with TPM 2 belonging to the class of non-expressed genes is $< 1\%$, for all datasets considered. In order to compare the behaviour of genes across samples, in response to stimulation/time, it is also important to normalise within each gene. The normalised value for a gene in a given condition (z) was calculated using z-score normalisation: $z = (x - \mu)/\sigma$ where: x = raw TPM in the sample considered, μ = average TPM across all conditions, σ = standard deviation of the TPM values across all samples for the gene.

Kallisto and Sleuth

In order to model the time-course response to stimulants with the Sleuth program, reads were first pseudoaligned using kallisto [46]. This software is able to accurately quantify transcript abundances without the need for alignment. Using estimated counts from the kallisto output, a time-based model was fitted in sleuth, using natural splines with five degrees of freedom. Any transcripts for which a likelihood ratio test against a null model had a q-value < 0.01 were considered significant and included in further analysis.

2.1.1 Stimulation with Poly(I:C)

Effects of the transfection procedure

While the primary focus was the investigation of varying poly(I:C) concentration and duration of treatment, the effect of lipofecatmine transfection was first checked, as this is thought to be able to induce up-regulation of gene expression. To identify the effect of LF in fibroblasts, pairwise comparisons between poly(I:C) + LF samples and the media used and LF alone with media were conducted, with differentially expressed genes identified using the cuffLinks software.

Figure 2.1a shows the total number of differentially expressed genes for each sample. It is clear to see that lipofectamine alone does not cause up-regulation of many genes at any time point, while poly(I:C) causes an increase in the expression of hundreds/thousands of genes. Interestingly, at 12 hours there are more differentially expressed genes when the transfection medium is not added, however when only genes involved in innate immunity are considered (Figure 2.1b) this difference is reduced. This trend may be a result of less efficient transfection in the absence of the transfection medium, leading to a delayed but similar induction of the immune response.

To verify induction of the type 1 interferon response, expression of IFN-β across conditions was considered, Figure 2.1c. This confirms up-regulation of IFN-β gene expression when treated with poly(I:C) but not lipofectamine. Furthermore, there is a sustained and slightly higher expression when transfection enhancing medium is added.

Concentration of poly(I:C)

As poly(I:C) may have harmful effects at high concentrations, the effect of reducing the concentration on induction of the type I interferon response was examined. To see whether there was increased sensitivity in the detection of response genes at higher poly(I:C) concentration, the number of differentially expressed genes in the standard concentration (1 μg/ml) over reduced concentrations (either 0.5 μg/ml or 0.1 μg/ml), samples were compared directly using cuffLinks. While there are limited differences between different concentrations of poly(I:C) at 4 and 12 hours, there is a clear increase in differentially expressed genes at 8 hours (Figure 2.1d), particularly compared to 0.1 μg/ml.

The dynamic response to poly(I:C) in two individuals

In order to investigate the response over time after poly(I:C) stimulation, a model using null splines to capture dynamics over the time-course was fitted as described above. Any transcripts for which this model explained behaviour significantly better than a null model (likelihood ratio test, q-value <0.01) were selected. This process was carried out for experimental data from two individuals separately. Many of the most significant genes (smallest q-value) are known to be involved in the antiviral response, such as the IFIT genes, IRF1, CCL2 and OAS1. Plotting the expression profiles of several of these genes (Figure 2.2a) highlights different types of expression patterns and variability between individuals. For example, some transcripts show fairly rapid up-regulation, with a peak at around 8 hours, before a decrease in expression level (IRF1) while

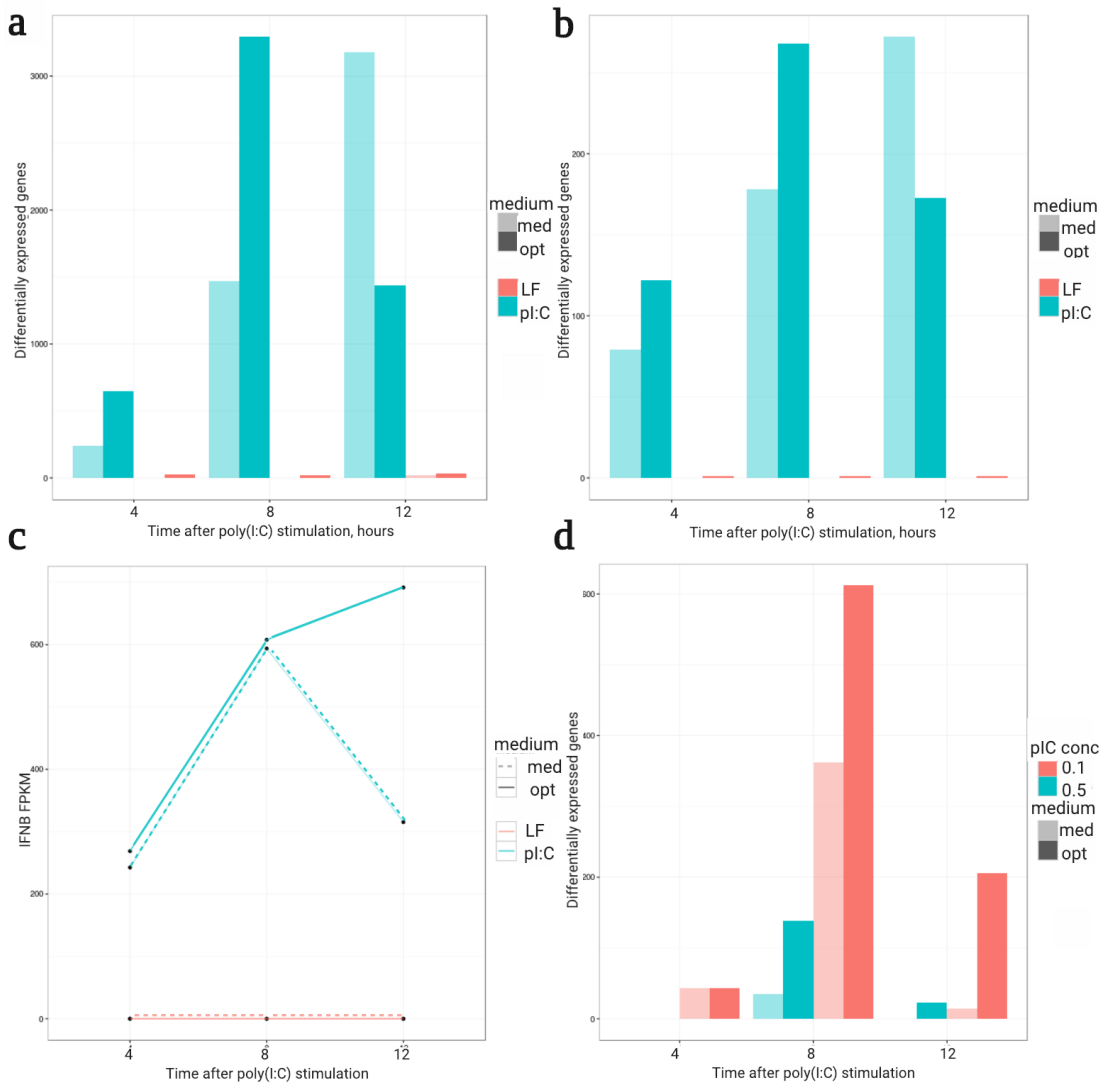


Fig. 2.1 Effects of the poly(I:C) transfection procedure. a-b) Number of genes up-regulated under different stimulant conditions. Colour indicates stimulant: red = lipofectamine alone (LF), blue = p(I:C) + LF (pI:C), while shade denotes transfection medium: light = medium alone ('med'), dark = medium + opti-MEM ('opt') a) Total differentially expressed genes identified. b) Number of differentially expressed innate immune genes. c) Expression of IFN-β (FPKM) across poly(I:C) and transfection control conditions. d) Comparison of response induced by different concentrations of poly(I:C): number of differentially expressed genes identified between 1 μg poly(I:C) and 0.5 or 0.1 μg.

others show up-regulation followed by sustained expression (IFIT3). In several cases, increase in expression is slower in individual 2 (IRF1, CCL2, OAS1) and there are large differences in transcript levels across time points between the individuals (IRF1, CCL2, OAS1). Figure 2.2b shows the behaviour of the entire set of significant transcripts (z-score normalised across conditions for each transcript) in the two individuals. Three groups of genes with distinct expression patterns can be seen, and similar groups are present in both individuals. The first cluster (orange) appears to be 'slow-response' genes, in which expression increases after 8-12 hours and reaches a maximum at 24 hours. In contrast, the cluster highlighted in purple are 'quick-response' genes, peaking at 8 hours. Finally, there are a group of transcripts (blue) which are expressed in the control and earliest time point, indicating genes which are down-regulated in response to poly(I:C). As cells show morphological changes signifying higher levels of apoptosis in later time points, the slow-response genes may be involved in this process. The enrichment (hypergeometric test) of genes marked as 'apoptotic' or 'interferon response' (GO term annotation) was investigated for each cluster, shown in Table 2.1. While all clusters showed enrichment of apoptotic genes, which may suggest that this is a more general feature of the transcripts selected as significant, only clusters 1 and 2 show enrichment of interferon response genes, suggesting that these specific patterns of expression reflect dynamic antiviral responses.

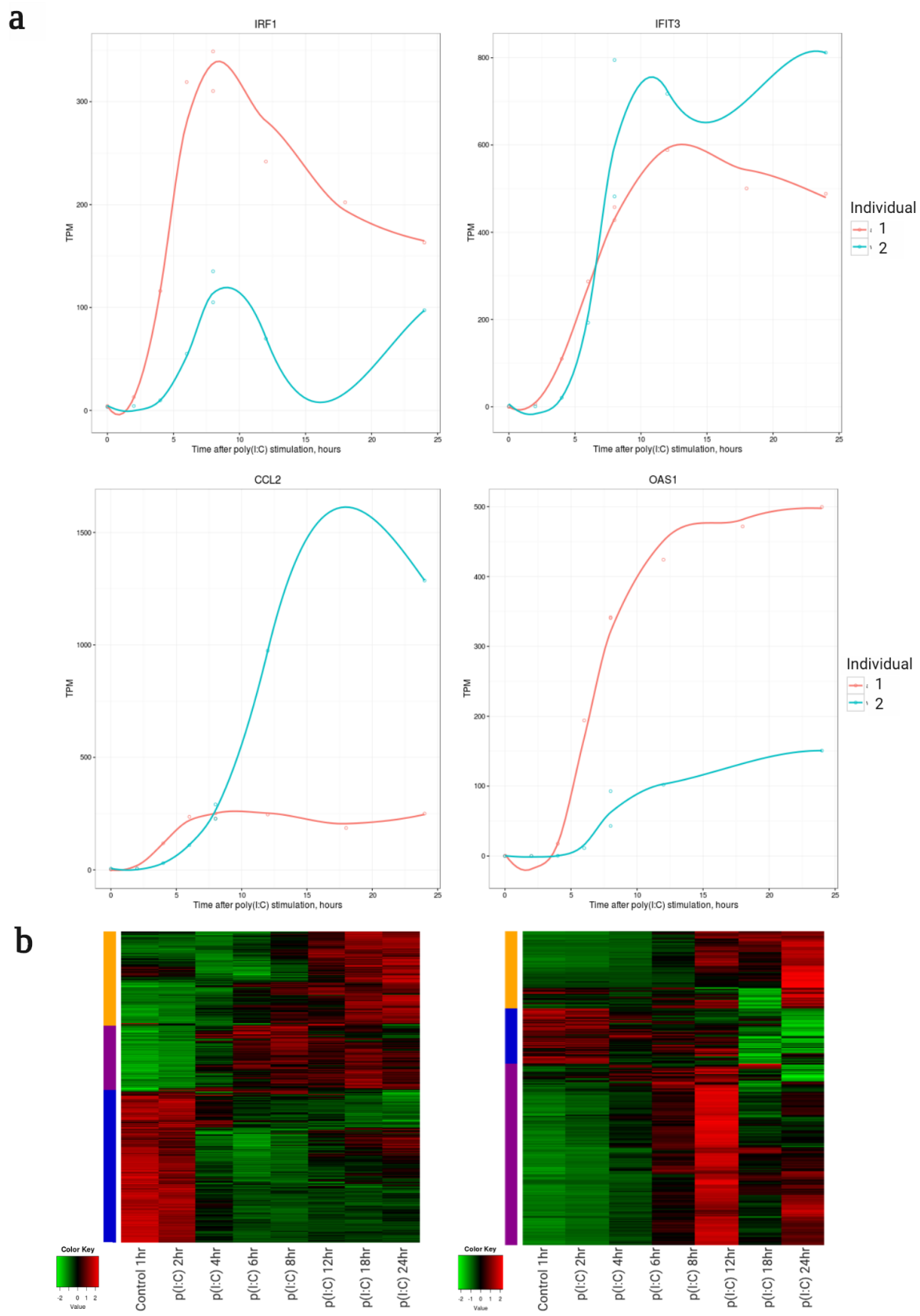


Fig. 2.2 Response to poly(I:C) stimulation over time in two individuals. a) TPM profile of IRF1, IFIT3, CCL2 amd OAS1, respectively, at timepoints of 0-24 hours after poly(I:C) treatment. b) Z-score normalised TPM of all transcripts for which the spline-based model was significant (likelihood ratio test, q value < 0.01) in two individuals.

Table 2.1 Enrichment of apoptotic v.s. IFN response genes in response to poly(I:C).

Group	Total	Apoptosis genes	Enrichment p-value	IFN response genes	Enrichment p-value
Background	*37978*	*1988*	-	*122*	-
Slow-response (orange)	2035	302	0	51	0
Quick-response (purple)	1904	285	0	36	0
Down-regulated (blue)	3024	387	0	6	0.87

2.1.2 Stimulation with interferons

While the results above show that poly(I:C) is capable of inducing an antiviral state in transfected cells, it is also possible to induce interferon signaling in a direct fashion through administration of interferons. As cell types respond differently to distinct interferons, an initial investigation into the response in fibroblasts was conducted, before looking at a more comprehensive time course of interferon-induced changes.

Type I vs Type II interferons

The response to IFN-α and IFN-β (type I) and IFN-γ (type II) at 1 and 4 hours, along with combined IFN-β and IFN-γ stimulation for 4 hours, was studied in two individuals. The heatmaps in Figure 2.3a show two distinct sets of genes in both individuals: one group which responds to interferons at 4 hours (blue), while another which shows higher expression in the controls and at 1 hour after stimulation (orange). As expected, the former group is very strongly enriched for interferon response genes (18/230, background proportion = 122/37978, p = 0), while the latter is not (1/279, p = 0.23). From these heatmaps, it appears that IFN-α and IFN-β elicit similar changes, while IFN-γ shows a distinct response (although up-regulation of type 1 response genes is seen in the sample stimulated with both IFN-β and IFN-γ).

To consider directly the similarity between response to the different interferons, Pearson correlation of gene expression (TPM) between samples was calculated, shown in Figure 2.3b. As would be expected, there is a high level of correlation between control and 1 hour time points, and between 4 hour time points of IFN-α and IFN-β stimulation. As seen above, IFN-γ treatment alone yields a more distinct response.

Interferon β time course dynamics

In order to elucidate the dynamics of response to interferon β, the same modelling approach as discussed for poly(I:C) above was utilised. Similarly, any transcripts for which the spline-based model significantly explained the data (compared to a null model, q-value <0.01) were selected. Again, many of these are known to function in the innate immune response, and the expression of example transcripts after IFN-β stimulation is shown in Figure 2.4a. The difference in dynamics is highlighted in these plots, in which some transcripts are more quickly up-regulated before decreasing (TAP1) or plateauing (DTX3L), while others steadily increase over time (ISG15, STAT1). Interestingly, in several of the transcripts, increase in expression in the second individual begins at later time points compared to the first individual. This may signify a broader delay in response to type I interferons. Although there is a lack of replicates for each individual, suggesting that further investigation may be needed in order to conclude differences between the individuals, the presence of several close time points in each time course deriving from different experimental wells adds reliability to the findings. Considering all significant transcripts (Figure 2.4b), there appear to be four main patterns of expression displayed in both individuals. In order of the timing of peak expression, there is first a group in which expression is highest in the control and at 1 and 2 hours of expression, but down-regulated after this (blue). Another group has similar expression but with a later peak (3-4 hours, brown). Neither of these groups are significantly enriched for genes involved in the interferon response (Table 2.2). The

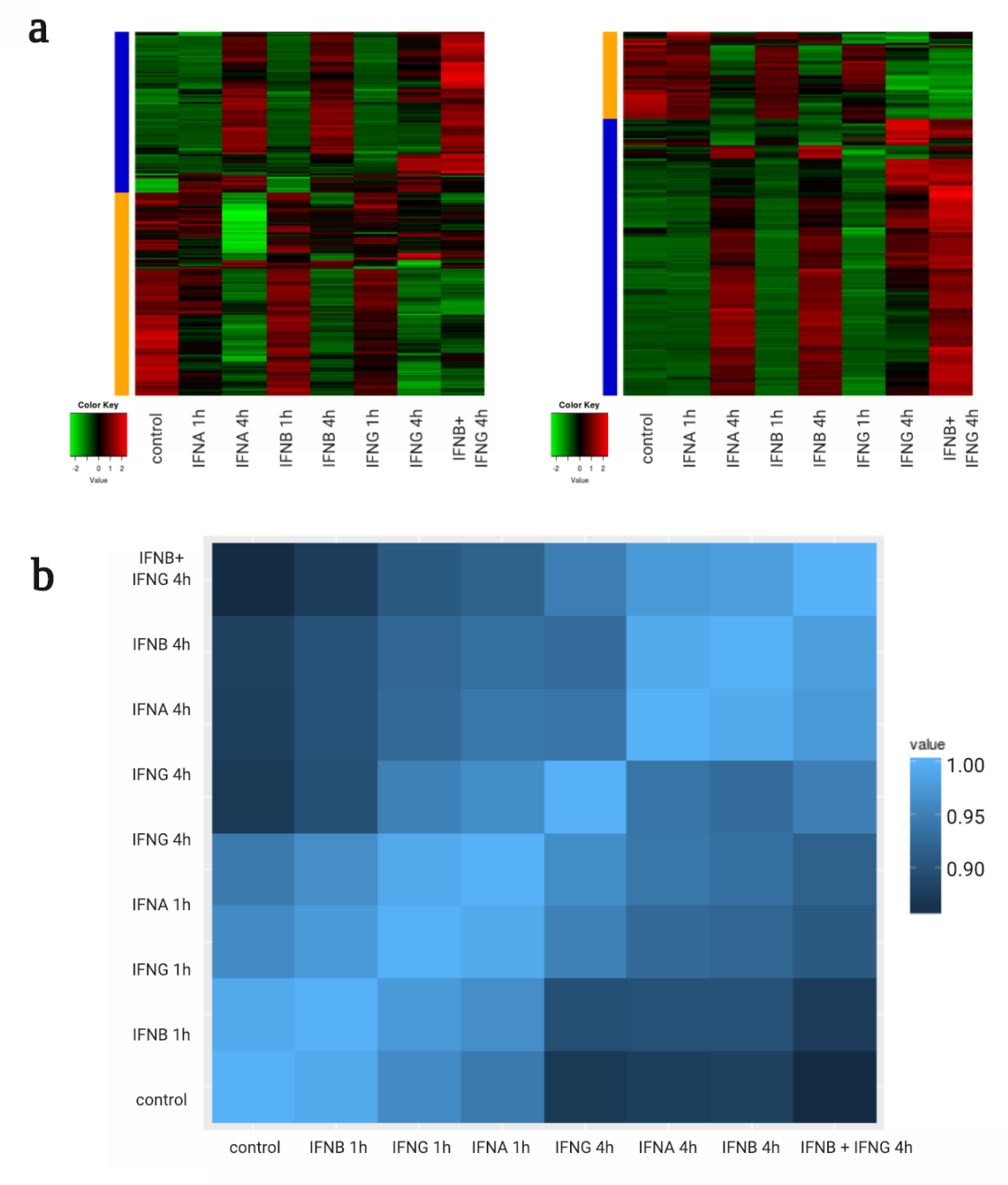


Fig. 2.3 Response to different interferon stimulations. a) Z-score normalised TPM across different interferon treatments (IFN-α, IFN-β, and IFN-γ) at 1 and 4 hours in two individuals; b) Correlation (Pearson coefficient) of TPM between different IFN treatments.

remaining two groups both show up-regulation in response to IFN-β, although in the group highlighted purple this is focused at 6-12 hours post-stimulation, while in the final group (orange) expression is highest at the latest time points. In both cases, there is high enrichment of genes known to play a role in the interferon response.

Table 2.2 Enrichment of apoptotic v.s. IFN response genes in response to IFN-β.

Group	Total	Apoptosis genes	Enrichment p-value	IFN response genes	Enrichment p-value
Background	*37978*	*1988*	-	*122*	-
Early (brown)	288	44	3.1 x $10^{-}13$	2	0.066
Down-regulated (blue)	448	68	8.9 x $10^{-}16$	2	0.17
Slow-response (orange)	480	81	0	26	0
Intermediate-response (purple)	366	68	0	27	0

2.1.3 Innate immunity vs. apoptotic genes across conditions

Thus far, the response to poly(I:C) and IFN-β stimulation has been considered separately, and presence of genes known to be involved in the innate immune response only seen through enrichment values. To further investigate this, alongside the presence of genes known to be involved in apoptosis (a factor in deciding optimal experimental conditions), the expression across control, poly(I:C) and IFN-β treated cells at many time points were considered for the set of innate immune and apoptotic genes. As there is overlap in these two sets of genes, only those which are annotated with one but not the other term were considered. Figure 2.5a) shows the normalised TPM across samples for innate immune and apoptotic genes respectively. While there are similar expression patterns, for example genes expressed most highly in poly(I:C) treatment (highlighted in orange), genes expressed across control and IFN-β samples (blue) and

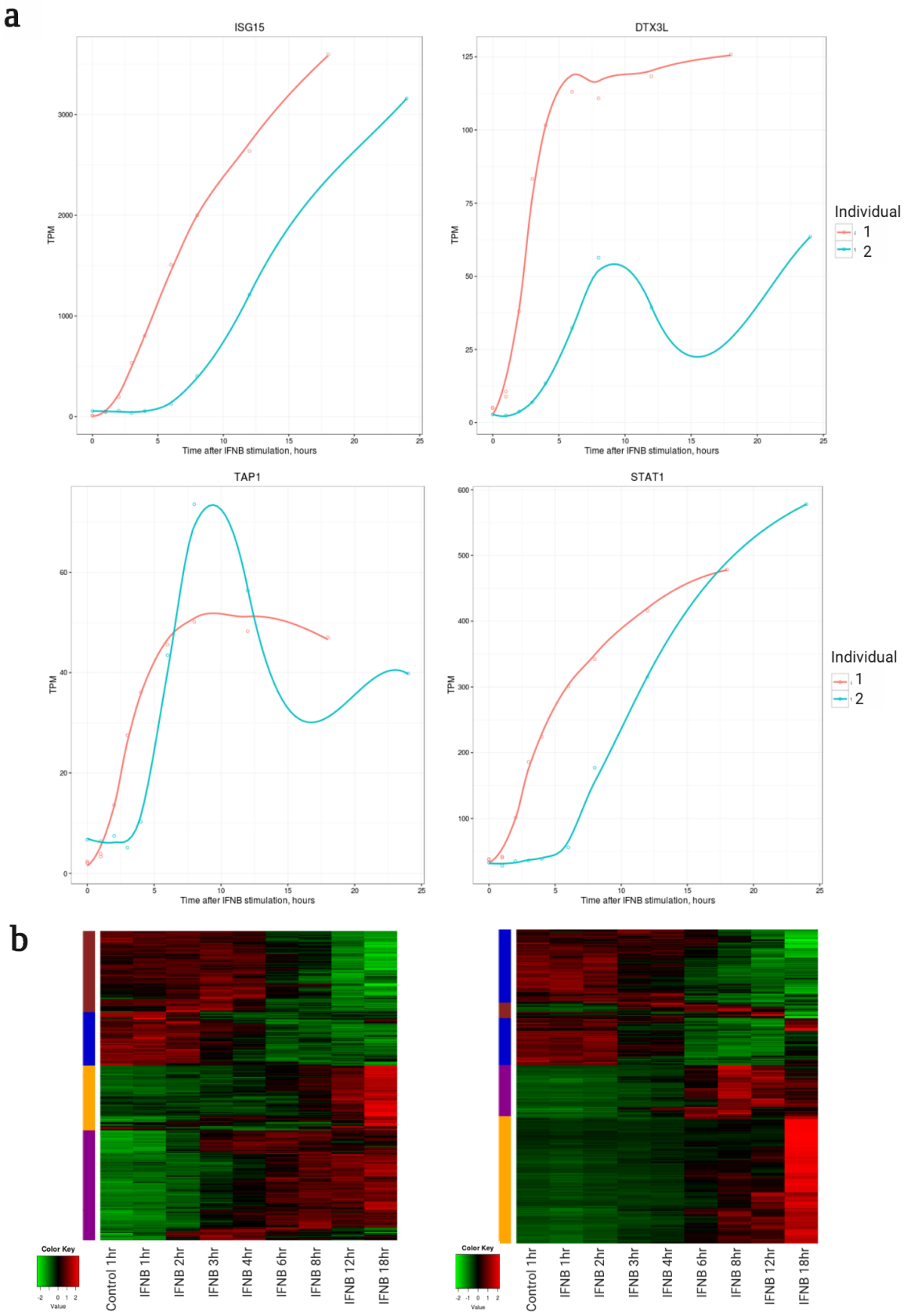


Fig. 2.4 Response to IFN-β stimulation over time in two individuals. a) TPM profile of ISG15, DTX3L, TAP1 amd STAT1, respectively, at timepoints of 0-24 hours after IFN-β treatment. e) Z-score normalised TPM of all transcripts for which the spline-based model was significant (likelihood ratio test, q value < 0.01) in two individuals.

those expressed mostly in control samples and at the latest time points after poly(I:C) stimulation (purple), there are some key differences between the innate immunity and apoptosis set of genes. In the heat map on the left (innate immunity), there are more distinct waves of expression through the poly(I:C) time-course, and there is a set of genes in the lower part of the top cluster which are expressed in both IFN-β and poly(I:C) stimulated cells – a feature missing from the right-hand heatmap.

In Figure 2.5b), the correlation between all samples is considered. While the similarity between samples of the same treatment type is to be expected, a clear difference in the similarity between IFN-β samples and controls can be seen – the correlation is much lower in innate immune genes than apoptotic genes.

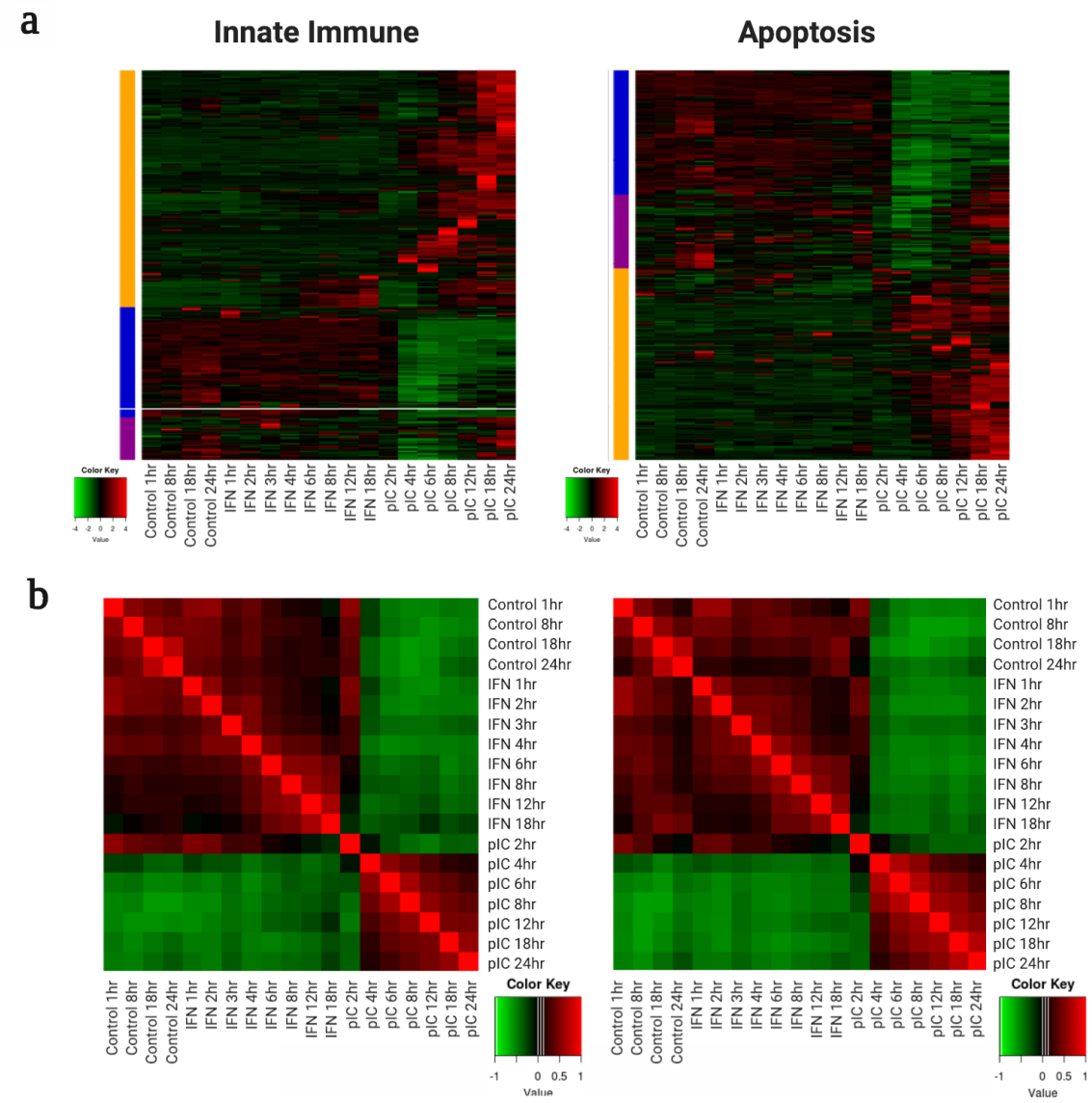


Fig. 2.5 Comparison of control, poly(I:C) and IFN-β treated cells with time. a) Z-score normalised TPM, and b) Correlation between samples, for genes with 'innate immune' (left) and 'apoptosis' (right) functions.

2.2 Large-scale stimulation experiments

In order to study the effect of genetic variation on the innate immune response, a large number of individuals was required. To this end, primary fibroblast cells from the Human Induced Pluripotent Stem Cell Initiative (HipSci; http://www.hipsci.org/) were used. These samples were collected initially for reprogramming into induced Pluripotent Stem Cells (iPSCs), however they provided a ideal resource for stimulation experiments, especially given the genetic profiling carried out through the intiative. The cells derived from healthy individuals spanning a range of ages and both genders (Appendix A).

2.2.1 Expansion of lines

As the initial sample from each line was one vial of 1 million cells, expansion of cells was required to ensure there were enough for stimulation experiments and further studies. Cells were cultured in supplemented DMEM (high glucose, pyruvate, GlutaMAX - Life Technologies), with 10% FBS and 1% penicillin-streptomycin added, until they had expanded at least three-fold. The passage numbers of fibroblasts ranged, as did the apparent quality of the cells, leading to the introduction of a 'grading' system. Not all lines were graded, however this qualitative score - based upon morphology under the microscope - was recorded for the majority of cultures, and cell viability scores were recorded for all lines. Only lines with a grade 3 or above were used in further experimental work; grades, where available, are shown in Appendix A.

2.2.2 Stimulation experiments

The aim of the stimulation experiments is to mimic a viral infection, inducing an effective type I interferon response while minimizing apoptosis. On the basis of the data

presented in section 2.1, it was determined that the following experimental conditions would be used:

- 0.5 μg/ml p(I:C), at 2 and 6 hours
- 1000 U/ml IFN-β, at 2 and 6 hours
- Unstimulated, medium-only (control) cells.

The concentration of p(I:C) was chosen based upon the observation that 1 μg/ml p(I:C) induced a similar response, and 0.1 μg/ml was significantly less effective (Figure 2.1d). The time points were chosen in order to capture the early induction of response at 2 hours, followed by later response at 6 hours, while minimising observation of the apoptotic effect seen at later times (Figure 2.5). To capture the secondary wave of type I interferon signalling, IFN-β was applied directly to cells. Both IFN-α and IFN-β induced a type I interferon response (Figure 2.3), however IFN-β was chosen due to its physiological relevance in fibroblast cells.

A schematic of the experimental setup is shown in Figure 2.6. As can be seen, this was carried out for fibroblasts from many individuals, with three donors being profiled in each experiment. Using the same experimental protocol as above, fibroblasts were stimulated directly with either rhodamine-conjugated poly(I:C) or human recombinant IFN-β. Poly(I:C) was mixed with 1 μl lipofectamine 2000 in 50 μl optiMEM, per well (6 well plate), for 5 minutes prior to transfection. IFN-β was diluted in the media immediately prior to addition. After the relevant period of time, cells were trypsinised and mixed (for example, 'unstimulated' cells from the three donors would be pooled together). The primary aim of this mixing step is to reduce downstream experimental variability between donors, while simultaneously streamlining the collection stage. However, this consequently necessitates the *in silico* deduction of the donor of origin for each cell, as described below.

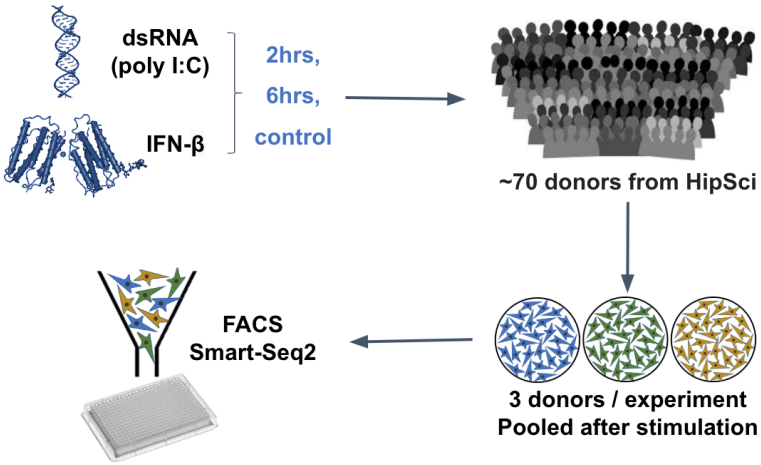


Fig. 2.6 An overview of stimulation experiments on HipSci fibroblast lines. Cells were stimulated with either 0.5 μg/ml p(I:C) or 1000 U/ml IFN-β for 2 or 6 hours, or left unstimulated as a control. Three donors per experiment were stimulated, and pooled together prior to FACS and consequent processing.

In a pilot study of three lines, cells were captured using both a droplet capture method (10X Genomics) and a flow cytometry plate based method. In the droplet capture protocol, control and stimulated cells were (separately) washed with PBS, trypsinised, and resuspended in PBS + 4% BSA. Cells were captured in droplet suspension on 10x Genomics' Chromium machine, and processed to sequencing libraries following the supplier's protocol. Multiplexed libraries were sequenced on an Illumina MiSeq instrument.

For the plate-based method used throughout, cells were washed with PBS, trypsinised, and resuspended in PBS + 0.1% DAPI. Cells were sorted on a Becton Dickinson IN-FLUX into plates containing 2 μl/well lysis buffer. Single cells were sorted individually (using FSC-W vs FSC-H), and apoptotic cells were excluded using DAPI. Rhodamine-positive cells were selected in the poly(I:C) treatments. Reverse transcription and cDNA amplification was performed according to the SmartSeq2 protocol (Picelli et al., 2014), and library preparation was performed using an Illumina Nextera kit. Samples were sequenced using paired-end 75bp reads on an Illumina HiSeq 2500 machine.

2.3 Data processing

For both bulk and single cell data the reads were submitted to fastqQC (version 0.11.7), mapped using Salmon (version 0.9.1) on genome build GRCh37, quantified by featureCounts from the Subread package (version 1.6.2). These programs were run using the Nextflow pipeline [1]. All programs, excepting the Subread package, were installed in the conda environment specified by the file 'env-rnaseq1.6.yml' in the same repository. The Subread package was installed separately. The index for salmon was built on reference v29lift37.

[1] https://github.com/cellgeni/rnaseq

2.4 Additional datasets

2.4.1 Primary skin data

To study the similarity of *in vitro* cultured fibroblasts to *ex vivo* skin cells, a primary skin tissue sample was used. This derived from a collaboration with Professor Muzlifah Haniffa (Newcastle University), and Roser Vento, Felipe Vieira Braga and Gozde Kar.

A skin sample taken from a human female was digested overnight in RPMI, 10% FCS, 100 U/ml penicillin, 100 μg/ml streptomycin, 1% L-Glutamine and 1.6 mg/ml collagenase. Dead cells were removed using beads from Miltenyi Biotec, followed by use of CD45+ beads (Miltenyi Biotec) to remove immune cells according to standard manufacturer protocol. To profile non-immune cells, the CD45- fraction was processed in a 10X Chromium machine (10X Genomics). Libraries were prepared according to the manufacturer's protocol. The resulting libraries were sequenced on two lanes of Illumina Hiseq 2,500 (rapid run mode). Droplet-based sequencing data was aligned, filtered and quantified using the Cell Ranger Single-Cell Software Suite, against the GRCh38

human reference genome provided by Cell Ranger. The output of this procedure (filtered matrix files) was used with the Seurat package. Low-quality cells (cells with less than 500 expressed genes and above 10% mitochondrial reads) were removed prior to further analysis.

2.4.2 Cross-mammalian data

These data were generated by Tzachi Hagai, and involved stimulation of primary dermal fibroblasts from sexually-mature females of four different species (human (European ancestry), rhesus macaque, C57BL/6 (black 6) mouse and brown Norway rat). All skin samples were taken from shoulders. Human cells were obtained from the Hipsci project, as described above. Rhesus macaque cells were extracted from skin tissues that were incubated for 2 h with 0.5% collagenase B after mechanical processing, and then filtered through 100 µm strainers before being plated and passaged before cryo-banking. Rodent cells were obtained from PeloBiotech where they were extracted using a similar protocol.

Prior to stimulation, cells were thawed and grown for several days in ATCC fibroblast growth medium with Fibroblast Growth Kit-Low serum (supplemented with Primocin and penicillin/streptomycin) - a controlled medium that has proven to provide good growing conditions for fibroblasts from all species, with slightly less than 24 h doubling times. About 18 h before stimulation, cells were trypsinized, counted and seeded into 6-well plates (100,000 cells per well). Cells were stimulated as follows: (1) stimulated with 1 μg/ml high-molecular mass poly(I:C) transfected with 2 μg/ml Lipofectamin 2,000; (2) mock transfected with Lipofectamin 2,000; (3) stimulated with 1,000 IU of IFNB for 8 h (human IFN-β for human and macaque cells, rat IFN-β for rat cells, mouse IFN-β for mouse cells; all IFNs were obtained from PBL, and had activity units based on similar virological assays); or (4) left untreated.

Chapter 3

Heterogeneity in primary human fibroblasts

> ***Declaration***
>
> *Primary skin data was generated and processed by the lab of Muzlifah Haniffa. The study of clonal structure in fibroblasts was carried out as part of a close collaboration with Davis McCarthy and Yuanhua Huang (Stegle Group, EMBL-EBI), who developed and benchmarked the computational method - cardelino - underpinning this analysis, and final figures for the paper. Daniel Kunz (Teichmann Group, WSI) conducted the selection analysis. The full manuscript, under review at the time of writing, is included along with supplementary figures and methods as Appendix C.*

3.1 Introduction

Prior to characterising differences in the innate immune response within and between individuals, it is important to understand heterogeneity in resting fibroblasts. Fibroblasts are a diverse cell type, characterised by synthesis of structural proteins and role in the extracellular matrix. It is known that there are a variety of subtypes across tissues, however the breadth and molecular functions in humans are incompletely characterised. Within the skin, there are several fibroblast classes, such as papillary, reticular, and hair follicle fibroblasts. Fibroblast sub-types in the skin are reviewed in depth in Lynch Watt, 2018 [136]. In this chapter, I investigate heterogeneity in cultured dermal fibroblasts by comparing to scRNA-seq data from primary skin samples.

Even within cells classified as the same type, there can be considerable transcriptional heterogeneity. This is reviewed in depth in [38], where the distinction is made between the stochasticity in biochemical processes (termed 'noise') and variability in the observable molecular phenotypes. In brief, this phenotypic variability, which can be assayed with single cell technologies, is a combination of stochastic noise along with deterministic regulatory mechanisms. While the role of variability across biological contexts has yet to be fully elucidated, it is particularly important in immune-stimulation contexts to first understand sources of transcriptional heterogeneity within the resting state prior to activation. The second part of this chapter is focused on characterising heterogeneity in unstimulated cultured fibroblasts.

Thus far, heterogeneity has been considered solely at a transcriptional level. However, elements such as ageing, environment and genetic factors can impact mutational processes, thereby shaping the acquisition of somatic mutations across the life span [137–141]. The maintenance and evolution of somatic mutations in different sub-populations of cells can result in clonal structure, both within healthy and disease tissues.

Targeted, whole-genome and whole-exome DNA sequencing of bulk cell populations has been utilized to reconstruct the mutational processes that underlie somatic mutagenesis [142–146] as well as clonal trees [147–149]. Availability of single-cell DNA sequencing methods (scDNA-seq; [150–152] combined with new computational approaches have helped to improve the reconstruction of clonal populations [153–159]. However, the functional differences between clones and their molecular phenotypes remain largely unknown. Systematic characterisation of the phenotypic properties of clones could reveal mechanisms underpinning healthy tissue growth and the transition from normal to malignant behaviour.

An important step towards such functional insights would be access to genome-wide expression profiles of individual clones, yielding genotype-phenotype connections for clonal architectures in tissues. Recent studies have explored mapping scRNA-seq profiles to clones with distinct copy number states in cancer, thus providing a first glimpse at clone-to-clone gene expression differences in disease [160–163]. Targeted genotyping strategies linking known mutations of interest to single-cell transcriptomes have proven useful in particular settings, but remain limited by technical challenges and the requirement for strong prior information [164–166]. Generally-applicable methods for inferring the clone of origin of single cells to study genotype-transcriptome relationships are not yet established. In the final part of this chapter, I present a method developed by Davis McCarthy and Yuanhua Huang to infer clones from scRNA-seq data. Using cultured fibroblasts from the HipSci resource, I investigate mutational and transcriptional heterogeneity across clones.

3.2 A comparison of *in vitro* and *ex vivo* fibroblasts

A pilot experiment was used to investigate heterogeneity in the HipSci fibroblast samples used. In this study, fibroblasts from three individuals were pooled together before droplet capture (10X Genomics) and further processing, in order to minimise confounding batch effects. Using a novel method - cardelino, described further below in Section 3.2 - the donor of origin for each cell was deduced, using the scRNA-seq data and genotype information available for these lines as part of the HipSci project.

Dimensionality reduction techniques were used to map the high dimensional transcriptomic data onto a more easily interpreted low dimension space. Figure 3.1a shows the effect of various cellular factors, both technical and biological, using t-Stochastic Neighbourhood Embedding (tSNE) - a non-linear dimensionality reduction method. Cell cycle, assigned using the Seurat package on the basis of cycle phase marker expression, and donor of origin are major factors that differentiate the cells (leftmost panels). Number of unique molecular identifiers (UMIs), an indicator of transcript capture and sequencing depth, along with mitochondrial percentage, an indicator of cell quality, appear to have a less distinct distribution (rightmost panels), however this analysis only contains cells which passed the quality control (greater than 500 detected genes and less than 10% mitochondrial reads). Three variables were regressed out - cell cycle phase, number of UMIs and mitochondrial percentage - to allow analysis of biological differences of interest. This reduces the contribution of these factors (Figure 3.1b), while retaining donor differences.

As the fibroblasts described within this thesis have been in culture and passaged several times prior to use, a primary skin dataset produced by the lab of Muzlifah Haniffa was used for comparison (Chapter 2.4). These data contain several cell types in addition to fibroblast sub-populations (Figure 3.2a). Cluster-specific markers were

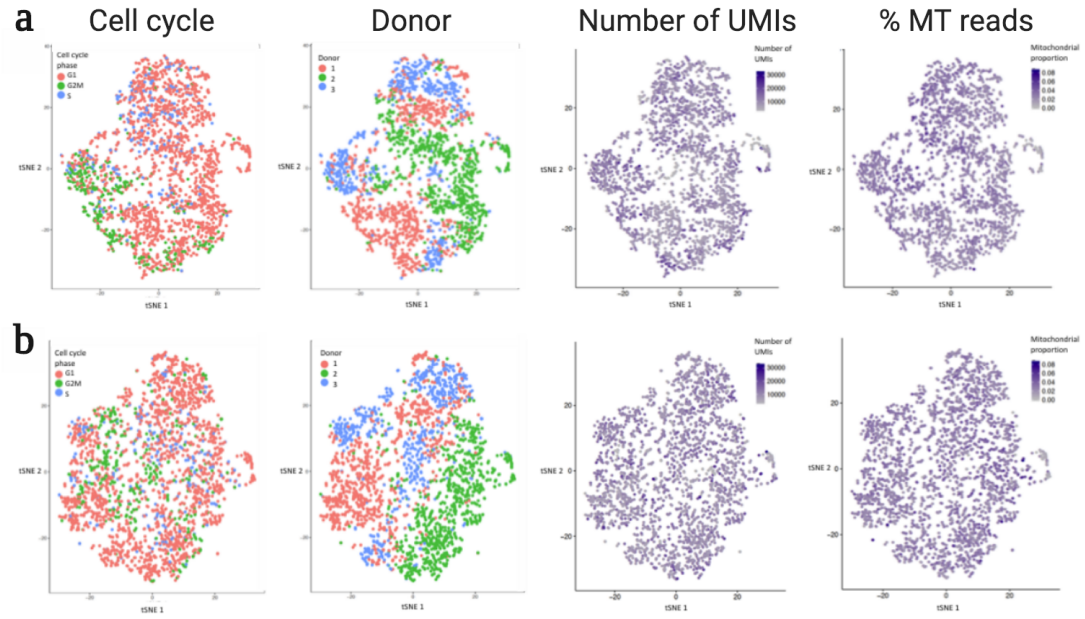


Fig. 3.1 An overview of a pilot droplet scRNA-seq dataset. a) tSNE visualisations coloured by cell cycle phase, donor, number of UMIs, and mitochondrial read proportion. b) Repeat of the tSNE visualisations after regression of cell cycle, number of UMIs and mitochondrial proportion.

identified using the Seurat v1 package [85], and are more uniquely expressed between clusters (Figure 3.2b; list of marker genes in Table B.1; Appendix B). To compare directly between these cells and the *in vitro* cultured fibroblasts mentioned above, the datasets were combined and clustering performed again (Figure 3.2c). The two datasets cluster separately in the combined analysis, however this is likely due to the large experimental and technical differences driving distribution in the tSNE plot.

The expression of markers indicative of *ex vivo* fibroblasts (Figure 3.2a-b, clusters 0 and 2 - referred to as fibroblast type 1 and 2 respectively) were plotted on the combined dataset (Figure 3.2d). From these plots, it appears that the *in vitro* cells are most similar to a subset of primary fibroblasts (type 2), and that expression of these marker genes is widespread and relatively homogenous across the *in vitro* cells. This not only confirms the isolation of the *in vitro* fibroblasts to a particular subset, but also the exclusion of other skin cell types from the population after extraction.

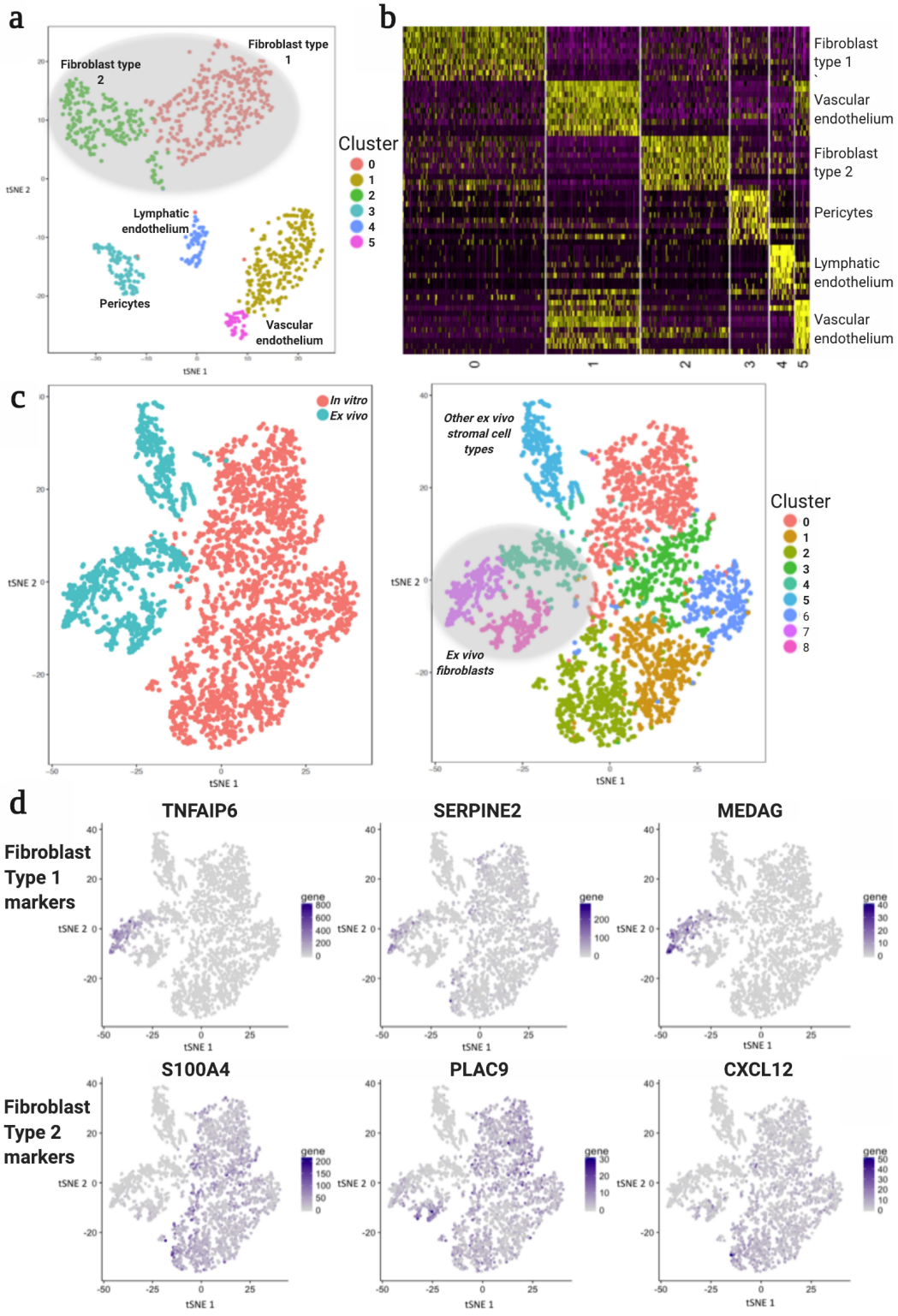


Fig. 3.2 Comparison of *in vitro* and *ex vivo* fibroblasts. a) tSNE visualisation and clustering of *ex vivo* skin cells; fibroblasts are shaded in grey. b) Top 10 differentially expressed markers for each cluster; full list with gene names in Table B.1. c) tSNE of merged *ex vivo* and *in vitro* datasets. d) Clustering of merged datasets, with *ex vivo* fibroblast populations once again shaded in grey. d) Expression of selected *ex vivo* fibroblast cluster markers in the merged dataset.

3.3 Transcriptional heterogeneity in unstimulated fibroblasts

While the fibroblasts studied appear to derive from one type, there may be other sources of heterogeneity within the cell populations. To investigate this further, unstimulated cells from the large stimulation experiment described in Chapter 2.2 were studied.

3.3.1 An overview of the scRNA-seq dataset

The quantified scRNA-seq data was first examined to gain an overview of the entire dataset. Prior to applying any filtering steps, there were 32367 cells. Looking at technical features of this dataset, it is clear that there is a large amount of variability in the quality and coverage of cells, highlighted by considering the number of reads mapped per cell, and the number of exogenous spike-in RNAs (ERCCs); Figure 3.3a.

Given the nature of scRNA-seq data, it is critical to perform stringent quality control prior to downstream analysis. In the biological context presented, this is both particularly relevant and challenging given the high levels of apoptosis induced alongside the antiviral response, as seen in Chapter 2.3. While early timepoints were selected to minimise apoptosis, there is a significant amount of cell death in samples treated with poly(I:C) for six hours. This is apparent transcriptionally when considering the number of mitochrondrial transcripts in each cell, which can be used as a transcriptional indicator of cell death, and is highest in the final stimulation condition (Figure 3.3).

Considering these technical factors, the following thresholds for retaining cells were applied: greater than 100,000 reads mapped, greater than 40% reads mapped, greater than 50,000 counts from endogenous genes, greater than 2,000 features (genes), fewer than 20% of counts from ERCCs and fewer than 20% of counts from mitochondrial reads. This resulted in 16929 cells being retained.

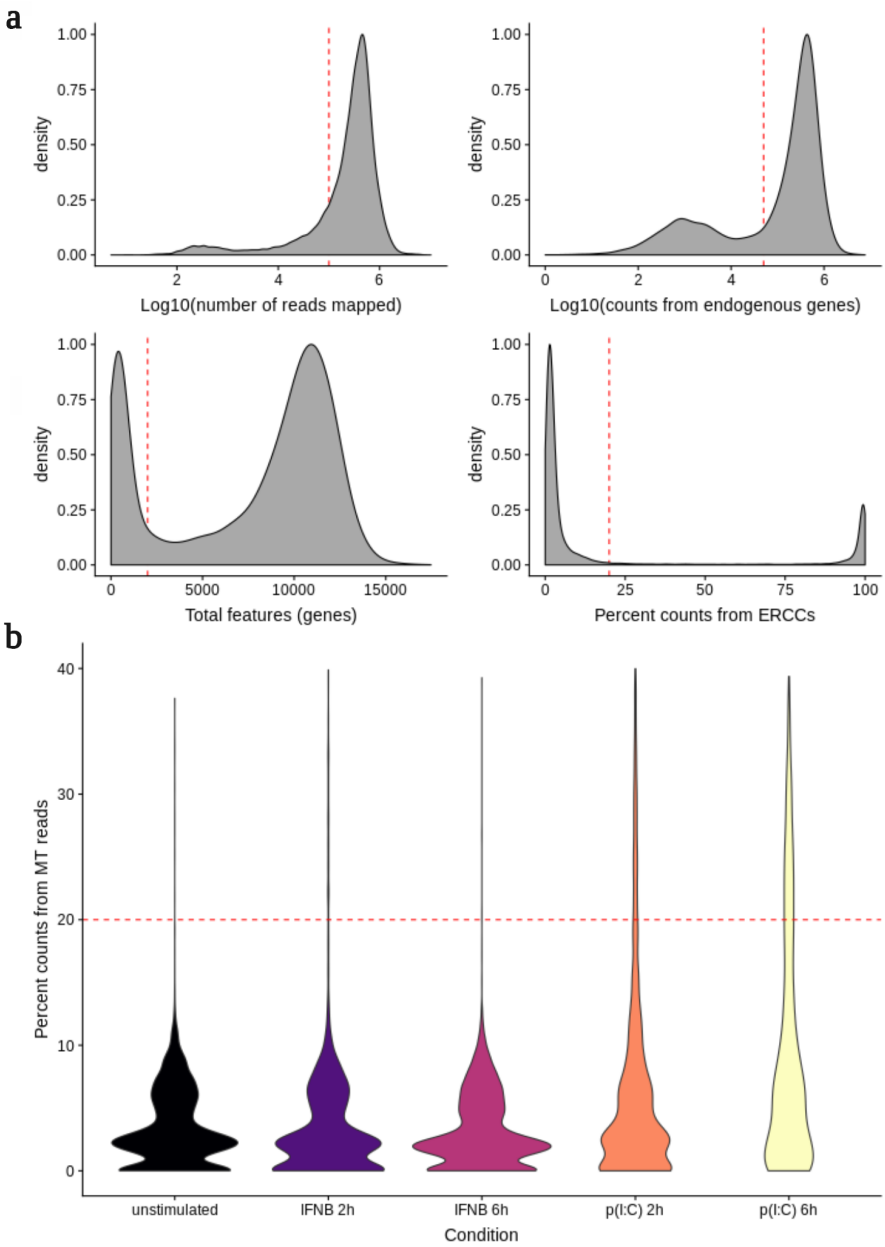


Fig. 3.3 Quality control of scRNA-seq data. a) Distribution of technical factors across cells: number of mapped reads, counts from endogenous reads, total features, ERCC percentage. Thresholds used for filtering cells shown in red: greater than 100,000 reads mapped, greater than 50,000 counts from endogenous genes, greater than 2,000 features (genes), fewer than 20% of counts from ERCCs and fewer than 20% of counts from mitochondrial reads.. b) Number of reads from mitochondrial (MT) genes across stimulation conditions.

3.3.2 Clustering analysis of unstimulated fibroblasts

Following the quality control step, there were 3979 unstimulated cells across 61 individuals. Using UMAP (Uniform Manifold Approximation and Projection), it is clear to see that a major driver of variation is experimental batch effect, although cells also cluster by cell cycle phase (Figure 3.4a). The batch divide arises from experimental date - it seems that samples from the first 16 experiments form one batch, while the remainder of samples form a discrete second batch. Although every effort was made to ensure reagents and protocols remained constant across all experiments, it appears that there was some variation arising from the processing of single-cell samples (this batch effect is not present in bulk RNA samples obtained in parallel). In order to characterise the dataset as a whole, it is important to correct the expression data to ensure it is comparable across experiments. In order to do this, the 'integrate' function from the Seurat v3 package was applied. This resulted in good mixing of the two batches in UMAP space, with cell cycle phase now being the major driver of variation in the dataset (Figure 3.4b).

To further investigate heterogeneity within unstimulated fibroblasts, the cells were clustered using the Seurat v3 package [167]. This uses a graph-based approach, first constructing a K-nearest neighbours (KNN) graph, using 'FindNeighbours' function. This uses the first 10 principal components to build the graph, refining weights between cells considering the shared overlap in their local neighbourhood. The 'FindClusters' function, which determines 'communities' of cells using a modularity optimisation approach, was then applied with a resolution of 0.2. This resulted in identification of five clusters (Figure 3.5a).

To characterise these clusters further, the top 10 marker genes per cluster were identified using a Wilcoxon rank sum test implemented in the 'FindMarkers' function. The expression of these genes across clusters is shown in Figure 3.5b. Enrichment

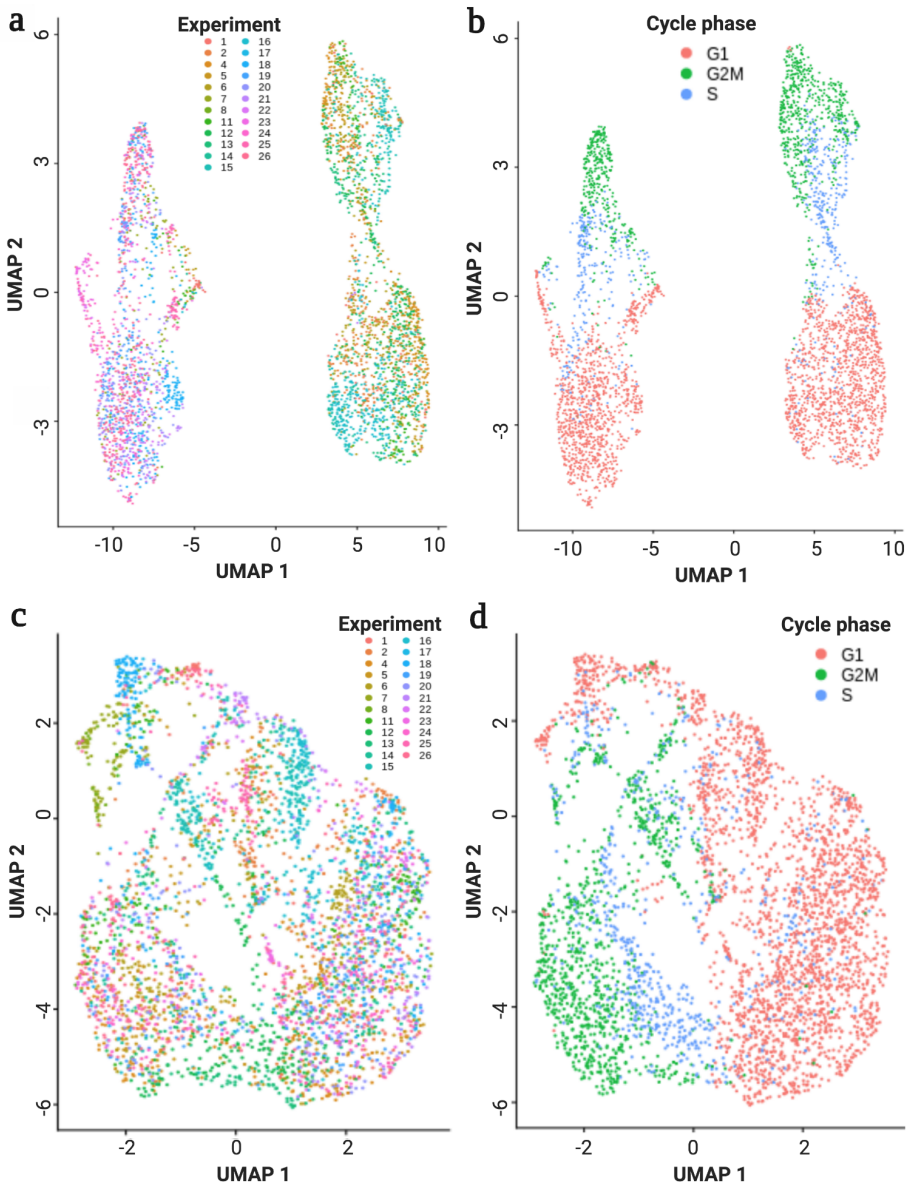


Fig. 3.4 Integration of scRNA-seq batches with Seurat. a) Dimensionality reduction using UMAP on uncorrected data: left, coloured by experimental batch, right, coloured by cell cycle phase. The first two UMAP dimensions are shown. b) UMAP plots after using Seurat v3's 'integrate' method: left, by batch, right, by cell cycle phase.

of gene ontology (GO) terms was examined to identify biological processes that may define these clusters; the significant GO terms are shown in Table B.2.

From this analysis, it appears that there are two major cycling clusters, both enriched for GO terms such as "cell cycle" and "cell division". The distinction may lie in the modules of cell cycle genes most highly expressed. Cycling cluster 1, for example, appears to have a predominance of spindle-related genes, such as ASPM and the centromeric proteins CENPF and CENPE.

Conversely, there are two clusters which represent non-cycling cells. Both these clusters have marker genes involved in cell-to-cell interaction and the extracellular matrix, such as FN1, COL3A1 and POSTN in non-cycling cluster 1, and B4GALT1, EMP3 in cluster 2. Cluster 1 also has enriched GO terms reflecting these processes. Again, although there are shared biological functions, cells in the two clusters may differ in expression level of subsets of these genes.

The final cluster, composed of a small number of cells, has GO terms related to diverse processes. However, many of the genes appear to relate to 'regulation of proliferation' (UBC, S100A4, S100A6,LGALS1, TMSB4X) or myofibril assembly (ACTC1, ACTG1, TMSB4X). This cluster comprises a mixed distribution of cell cycle phases, and could represent proliferative cells which are at a transition between cell cycle phases.

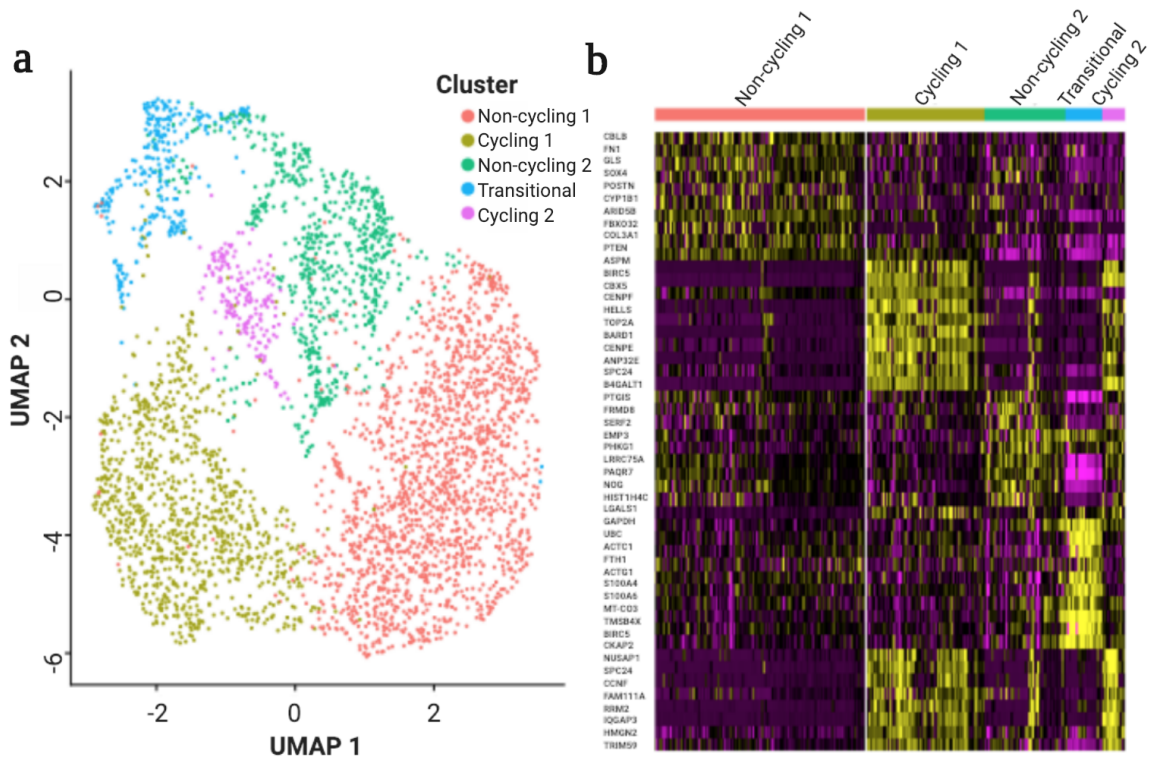


Fig. 3.5 Clustering analysis of unstimulated fibroblasts a) Clusters identified using Seurat v3's graph-based clustering approach, applying the 'FindNeighbours' and 'FindClusters' functions, with a resolution of 0.2. Five clusters are identified. b) The top 10 marker genes for each cluster are shown, with genes and cells ordered by cluster.

3.4 Identifying common variants and somatic mutations in scRNA-seq data

In collaboration with Davis McCarthy and Yuanhua Huang, we undertook a study to define clones within fibroblast populations. The project aimed to harness the ability to identify somatic mutations in transcriptomes of individual cells, mapping the cells to a clonal tree defined on the basis of shared clonal mutations, followed by investigation of the phenotypic differences between these clones. We used the scRNA-seq data of HipSci fibroblast lines, described in Chapter 2, focusing on 32 lines for which matching deep whole exome-sequencing data was available through the HipSci consortium. The full manuscript, including Supplementary Material, is included in Appendix C.

3.4.1 Cardelino: a method for assigning cells to clones using scRNA-seq data

Cardelino is a Bayesian method for integrating somatic clonal substructure and transcriptional heterogeneity within a population of cells. Briefly, cardelino models the expressed variant alleles in single cells as a clustering model, with clusters corresponding to somatic clones with (unknown) mutation states (Figure. 3.6a). Critically, cardelino leverages imperfect but informative clonal tree configurations obtained from complementary technologies, such as bulk or single-cell DNA sequencing data, as prior information, thereby mitigating the sparsity of scRNA-seq variant coverage. Cardelino employs a variant specific beta-binomial error model that accounts for stochastic dropout events as well as systematic allelic imbalance due to mono-allelic expression or genetic factors.

Initially, we assessed the accuracy of cardelino using simulated data that mimic typical clonal structures and properties of scRNA-seq as observed in real data (4 clones, 10 variants per branch, 25% of variants with read coverage, 200 cells, 50 repeat

experiments). By default, we consider an input clone configuration with a 10% error rate compared to the true simulated tree (namely, 10% of the values in the clone configuration matrix are incorrect). Alongside cardelino, we considered two alternative approaches: Single Cell Genotyper (SCG; [157]) and an implementation of Demuxlet, which was designed for sample demultiplexing rather than clone assignment ([168]; see Methods and Supp. Fig. S1). In the default setting, cardelino achieves high overall performance (Precision-Recall AUC=0.965; Figure. 3.6b), outperforming both SCG and Demuxlet. For example, at a cell assignment confidence threshold (posterior probability of cell assignment) of P=0.5, cardelino assigns 88% of all cells with an overall accuracy of 88.6%.

We explored the effect of key dataset characteristics on cell assignment, including the number of variants per clonal branch (Figure. 3.6c) and the expected number of variants with non-zero scRNA-seq coverage per cell (Figure. 3.6d). As expected, the number of variants per clonal branch and their read coverage in scRNA-seq are positively associated with the performance of all methods, with cardelino consistently outperforming alternatives, in particular in settings with low coverage. We further explored the effects of allelic imbalance on cell assignment (Figure. 3.6e), and found that cardelino is more robust than SCG and Demuxlet when there is a larger fraction of variants with high allelic imbalance. We attribute cardelino's robustness to its approach of modelling the allelic imbalance per variant, whereas SCG and Demuxlet both use a global parameter and hence cannot account for variability of allelic imbalance across sites. We also varied the error rate in the guide clone configuration, either introducing uniform errors in the configuration matrix by swapping the mutation states of any variants in any clone (Figure. 3.6f) or by swapping variants between branches (Figure. 3.6g). In both settings, cardelino is markedly more robust than Demuxlet, which assumes that the defined reference clonal structure is error free.

Notably, cardelino retains excellent performance (AUPRC>0.96) at error rates up to 25% (Figure. 3.6f-g), by modelling deviations between the observed and the true latent tree (Appendix C; Supplementary Figure S2).

We also considered two simplified variants of cardelino, one of which does not consider the guide clone tree and performs *de novo* tree reconstruction (cardelino-free), and a second model that treats the guide tree as fixed without modelling any errors (cardelino-fixed). These comparisons, further investigating the parameters assessed in Figure. 3.6, confirm the benefits of the data-driven modelling of the guide clone configuration as a prior that is adapted jointly while assigning scRNA-seq profiles to clones (Appendix C; Supplementary Figure S3). We also explored the effects of the number of clones (Appendix C; Supplementary Figure S3c), and the tree topology (Appendix C; Supplementary Figure S4), again finding that cardelino is robust to these parameters.

Taken together, these results demonstrate that cardelino is broadly applicable to robustly assign individual single-cell transcriptomes to clones, thereby reconstructing clone-specific transcriptome profiles.

3.4.2 Mutational analysis of *in vitro* fibroblasts

Between 30 and 107 unstimulated cells were assayed per line (median 61 cells after QC; median coverage: 484k reads; median genes observed: 11,108; Appendix C Supplementary Table S2). Initially, we considered high-confidence somatic single nucleotide variants (SNVs) identified based on whole exome sequencing (WES) data (Appendix C; Methods) to explore the mutational landscape across lines. This reveals considerable variation in the total number of somatic SNVs, with 41–612 variants per line (Figure. 3.7a; coverage of 20 reads, 3 observations of alternative allele, Fisher's exact test FDR0.1).

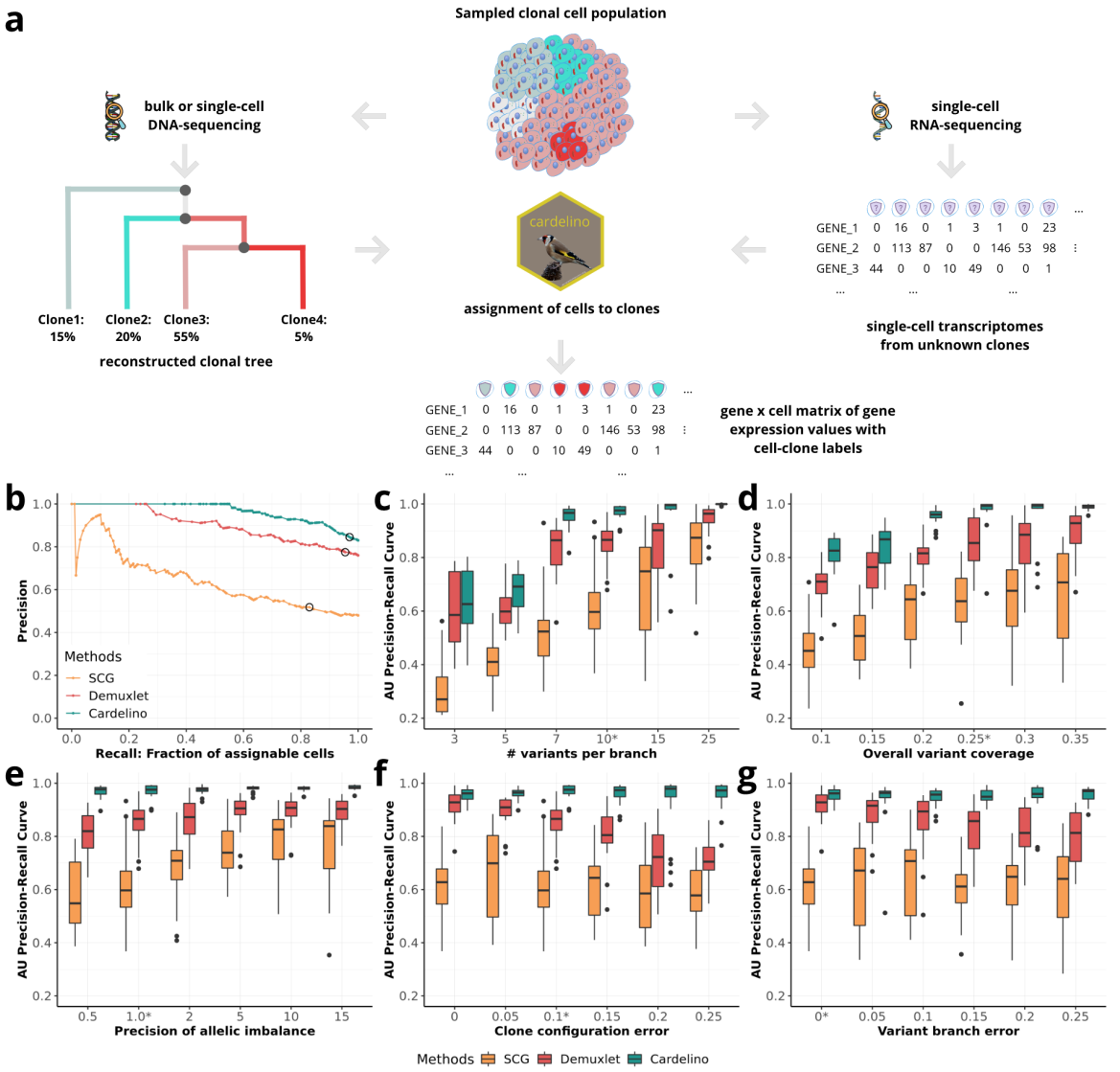


Fig. 3.6 Overview and validation of the cardelino model. (a) Overview and approach. A clonal tree is reconstructed using DNA-sequencing data to derive a guide clone configuration. Cardelino then performs probabilistic clustering of single-cell transcriptomes based on variants detected in scRNA-seq reads, assigning cells to clones in the mutation tree. (b-g) Benchmarking of the cell assignment using simulated data by changing one variable each time. The default values are highlighted with a star. (b) Overall assignment performance for a dataset consisting of 200 cells, simulated assuming a 4-clone structure with 10 variants per branch and non-zero read coverage for 20% of the variants. An error rate of 10% on the mutation states between the guide clone configuration and the true clonal tree was used. Shown is the fraction of true positive cell assignments (precision) as a function of the fraction of assigned cells (recall), when varying the threshold of the cell assignment probability. The black circle corresponds to the posterior cell assignment threshold of P=0.5. (c-g) Area Under (AU) precision-recall curve (i.e. area under curves such as shown in b), when varying the numbers of variants per clonal branch (c), the fraction of informative variants covered (i.e., non-zero scRNA-seq read coverage) (d), the precision (i.e., inverse variance) of allelic ratio across genes; lower precision means more genes with high allelic imbalance (e), the error rate of the mutation states in clone configuration matrix (f), and the fraction of variants that are wrongly assigned to branches (g).

Mutational signature exposures were estimated using the sigfit package [169], providing the COSMIC 30 signatures as reference [144], and with a highest posterior density (HPD) threshold of 0.9. Signatures were determined to be significant when the HPD did not overlap zero. Two signatures (7 and 11) were significant in two or more donors (Appendix C; Supplementary Figure S5). The majority of SNVs can be attributed to the well-documented UV signature, COSMIC Signature 7 (primarily C to T mutations; [144], agreeing with expected mutational patterns from UV exposure of skin tissues (Figure. 3.7a).

To understand whether the somatic SNVs confer any selective advantage in skin fibroblasts, we used the SubClonalSelection package to identify neutral and selective dynamics at a per-line level [170]. Other established methods such as dN/dS [171] and alternative methods using the SNV frequency distribution [172, 173] are not conclusive in the context of this dataset, likely due to lack of statistical power resulting from the low number of mutations detected in each sample. The SubClonalSelection analysis identifies at least 10 lines with a clear fit to their selection model, suggesting positive selection of clonal sub-populations (Figure. 3.7a). In other words, a third of the samples from this cohort of healthy donors contain clones evolving adaptively, which we can investigate in more detail in terms of transcriptome phenotype.

Next, we reconstructed the clonal trees in each line using WES-derived estimates of the variant allele frequency of somatic variants that are also covered by scRNA-seq reads (Appendix C; Methods). Canopy [149] identifies two to four clones per line (Figure. 3.7a). Briefly, Canopy models the phylogeny of cell growth in a tissue by depicting a bifurcating tree arising from a diploid germline cell whose daughter cells are subject to progressive waves of somatic mutations. When a sample of a tissue is taken, the tree is sliced horizontally, cutting the branches to form “leaves” or “clones”. Thus each clone represents a subpopulation of cells that share (and are

identified by) the somatic mutations in their most recent common ancestral cell. To handle the presence of a subpopulation of cells without somatic mutations, "clone1" is defined to represent a non-bifurcating, somatic mutation-free branch of the clonal tree. Thus, with any somatic variants present at sub-clonal frequencies (the case for all cell lines here), Canopy will infer the presence of at least two clones. Following Canopy's inference of clones, we used cardelino to confidently map scRNA-seq profiles from 1,732 cells (out of a total of 2,044 cells) to clones from the corresponding lines. Cardelino estimates an error rate in the guide clone configuration of less than 25% in most lines (median 18.6%), and assigns a large fraction of cells confidently (>90% for 23 lines; at posterior probability P>0.5). The model identifies four lines with an error rate between 35-46% and an outlier (vils, a line with few somatic variants), which demonstrates the utility of the adaptive phylogeny error model employed by cardelino. We also ran the other four alternative methods on these 32 lines (Appendix C; Supplementary Figure S12), and found that the *de novo* methods appear to suffer from higher uncertainty in recontrustructing clonal trees from scRNA-seq data only (Appendix C; Supplementary Figure S12C), while using the fixed-guide clonal tree from bulk exome-seq data may be over-simplified and leads to reduced stability when considering alternative high-confidence trees (Appendix C; Supplementary Figure S12D-E).

To further assess the confidence of these cell assignments, we considered, for each line, simulated cells drawn from a clonal structure that matches the corresponding line, finding that cardelino gives high accuracy (AUPRC>0.9) in 29 lines, again clearly outperforming competing methods (Appendix C; Supplementary Figure S13). Additionally, we observed high concordance (R2 = 0.94) between the empirical cell-assignment rates and the expected values based on the corresponding simulation for the same line (Figure. 3.7b). Lines with clones that harbour fewer distinguishing variants

are associated with lower assignment rates (Appendix C; Supplementary Figure S14), at consistently high cell assignment accuracy (median 0.965, mean 0.939 - Appendix C; Supplementary Figure SS15), indicating that the posterior probability of assignment is calibrated across different settings. We also considered the impact of technical features of scRNA-seq data on cell assignment, finding no evidence of biased cell assignments (Appendix C; Supplementary Figure S16-20). Finally, clone prevalences estimated from Canopy and the fractions of cells assigned to the corresponding clones are reasonably concordant (adjusted R2 = 0.53), providing additional confidence in the cardelino cell assignments, while highlighting the value of cardelino's ability to update input clone structures using single-cell variant information (Figure. 3.7c).

3.4.3 Transcriptional analysis of *in vitro* fibroblasts

Initially, we focused on the fibroblast line with the largest number of somatic SNVs (joxm; white female aged 45-49; Figure. 3.8a), with 612 somatic SNVs (112 detected both in WES and scRNA-seq) and 79 QC-passing cells, 99% of which could be assigned to one of three clones (Figure. 3.8a). Principal component analysis of the scRNA-seq profiles of these cells reveals global transcriptome substructure that reflects to a degree the somatic clonal structure in this population of cells (Figure. 3.8b). Additionally, we observed differences in the fraction of cells in different cell cycle stages, where clone1 has the fewest cells in G1, and the largest fraction in S and G2/M (Figure. 3.8b, inset plot; global structure and cell cycle plots for all lines in Appendix C; Supplementary Figures S24-33). This suggests that clone 1 is proliferating most rapidly. Next, we considered differential expression analysis of individual genes between the two largest clones (clone 1: 46 cells versus clone 2: 25 cells), which identifies 901 DE genes (edgeR QL F-test; FDR<0.1; 549 at FDR<0.05; Figure. 3.8c). These genes are approximately evenly split into up- and down-regulated sets. However, the down-regulated genes are

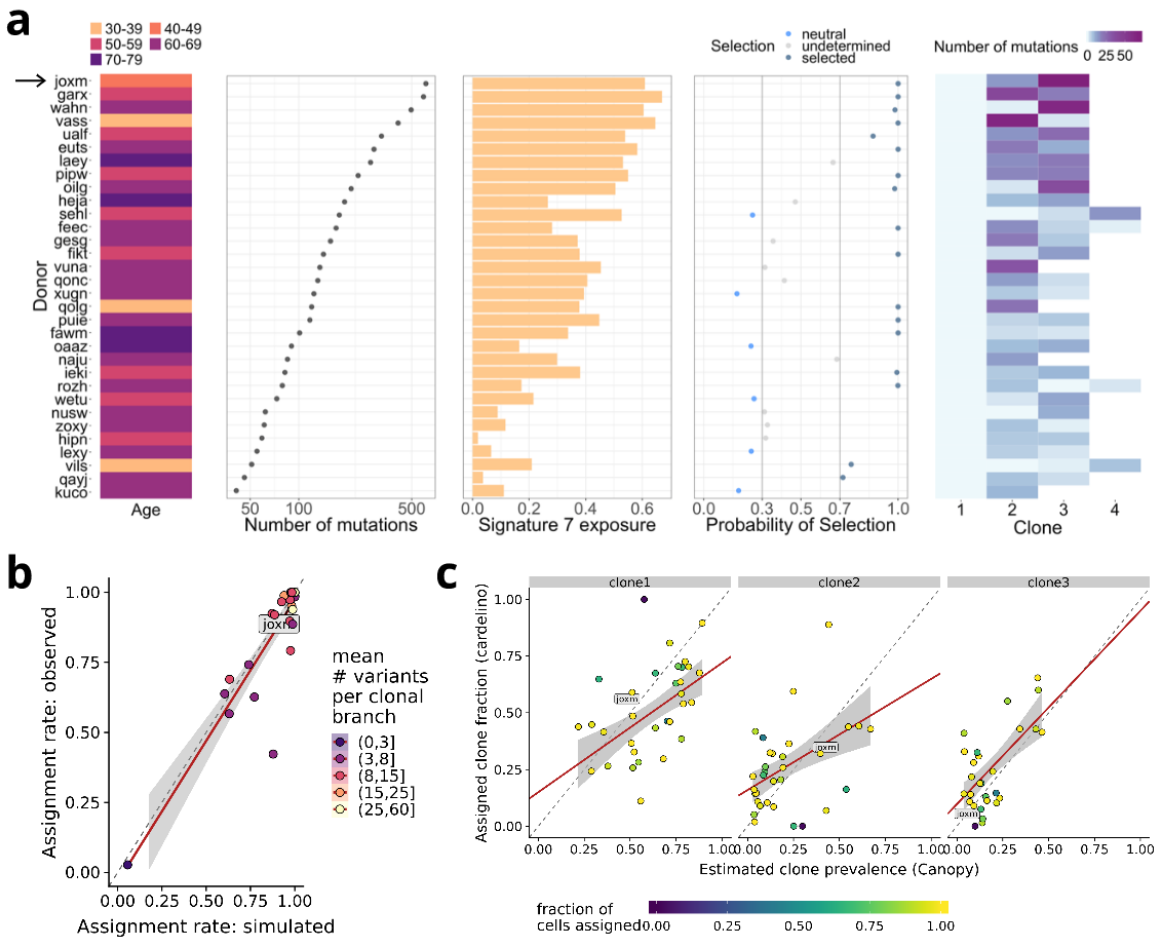


Fig. 3.7 Characterisation of mutational and clonal structure in 32 fibroblast lines. a) Overview and somatic mutation profiles across lines (donors), from left to right: donor age; number of somatic SNVs; estimated exposure of COSMIC mutational signature 7; probability of selection estimated by SubClonalSelection [170], colour denotes the selection status based on probability cut-offs (grey lines), the grey background indicates results with high uncertainty due to the low number of mutations detected; number of clones inferred using Canopy [149], with colour indicating the number of informative somatic SNVs for cell assignment to each clone (non-zero read coverage in scRNA-seq data). (b) Assignment rate (fraction of cells assigned) using simulated single-cell transcriptomes (x-axis) versus the empirical assignment rate (y-axis) for each line (at assignment threshold posterior P>0.5). Colour denotes the average number of informative variants across clonal branches per line. The line-of-best fit from a linear model is shown in red, with 95% confidence interval shown in grey. (c) Estimated clone prevalence from WES data (x-axis; using Canopy) versus the fraction of single-cell transcriptomes assigned to the corresponding clone (y-axis; using cardelino). Shown are the fractions of cells assigned to clones one to three as in a, considering the most likely assignment for assignable cells (posterior probability P>0.5) with each point representing a cell line. Colour denotes the total fraction of assignable cells per line (P>0.5). A line-of-best fit from a weighted regression model is shown in red with 95% confidence interval shown in grey.

enriched for processes involved in the cell cycle and cell proliferation. Specifically, the three significantly enriched gene sets are all up-regulated in clone 1 (camera; FDR<0.1; Figure. 3.8d). All three gene sets (E2F targets, G2/M checkpoint and mitotic spindle) are associated with the cell cycle, so these results are consistent with the cell-cycle stage assignments suggesting increased proliferation of clone 1. Taken together, the results suggest that somatic substructure in this cell population results in clones that exhibit measurably different expression phenotypes across the transcriptome, with significant differential expression in cell cycle and growth pathways.

To quantify the overall effect of somatic substructure on gene expression variation across the entire dataset, we fitted a linear mixed model to individual genes (Appendix C; Methods), partitioning gene expression variation into a line (likely donor) component, a clone component, technical batch (i.e. processing plate), cellular detection rate (proportion of genes with non-zero expression per cell) and residual noise. As expected, the line component typically explains a substantially larger fraction of the expression variance than clone (median 5.5% for line, 0.5% for clone), but there are 194 genes with a substantial clone component (>5% variance explained by clone; Figure. 3.9a). Even larger clone effects are observed when estimating the clone component in each line separately, which identifies between 331 and 2,162 genes with a substantial clone component (>5% variance explained by clone; median 825 genes; Figure. 3.9b). This indicates that there are line-specific differences in the set of genes that vary with clonal structure.

Next, we carried out a systematic differential expression (DE) analysis to assess transcriptomic differences between any pair of clones for each line (considering 31 lines with at least 15 cells for DE testing - Appendix C; Methods). This approach identifies up to 1,199 DE genes per line (FDR<0.1, edgeR QL F test). A majority, 61%, of the total set of 5,289 unique DE genes, are detected in two or more lines, and 39% are

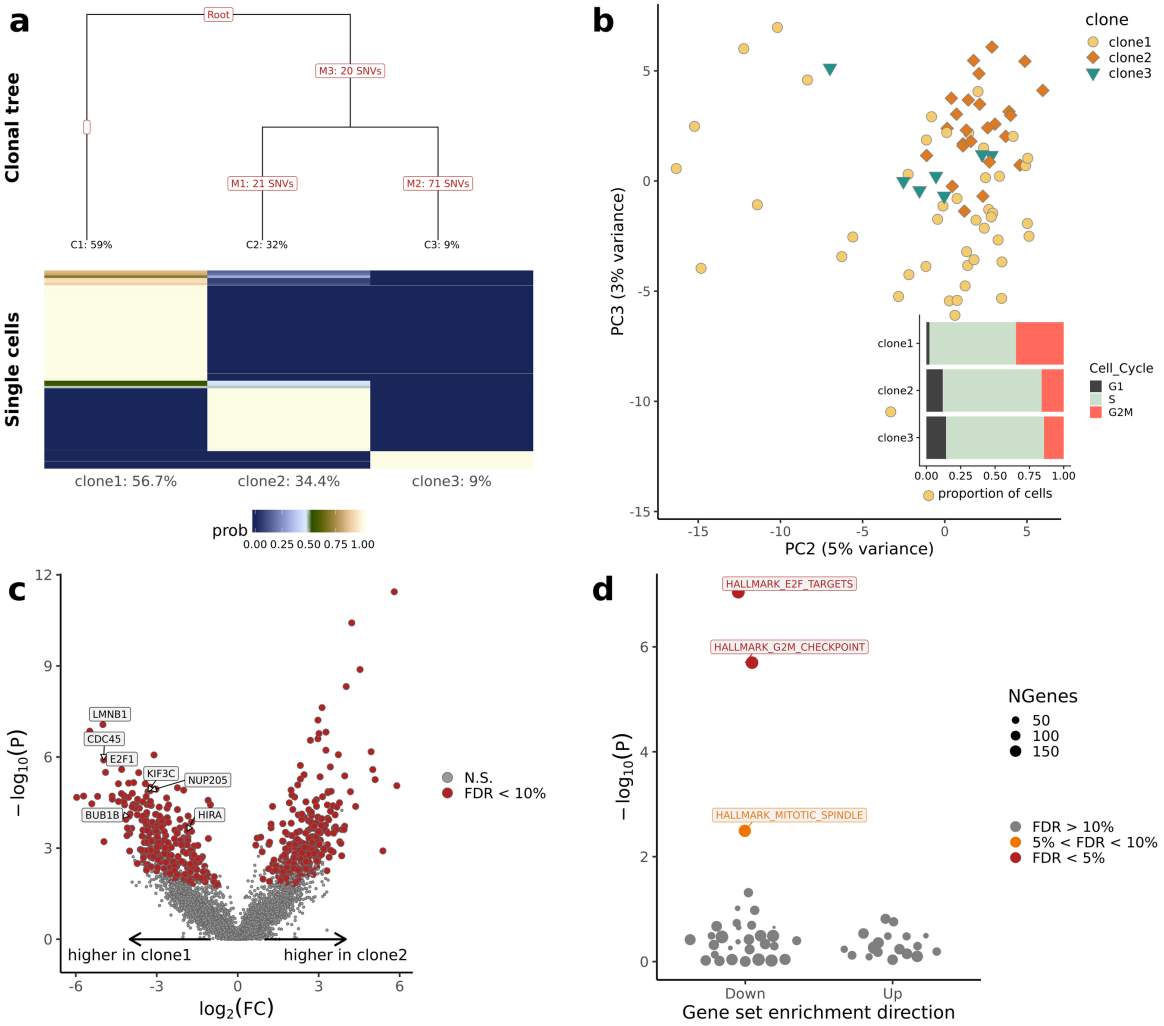


Fig. 3.8 Clone-specific transcriptome profiles reveal gene expression differences for joxm, one example line. (a) Top: Clonal tree inferred using Canopy [149]. The number of variants tagging each branch and the expected prevalence (fraction) of each clone is shown. Bottom: cardelino cell assignment matrix, showing the assignment probability of individual cells to three clones. Shown below each clone is the fraction of cells assigned to each clone. (b) Principal component analysis of scRNA-seq profiles with colour indicating the most likely clone assignment. Inset plot: Cell-cycle phase fractions for cells assigned to each clone (using cyclone [174]). (c) Volcano plot showing negative log10 P values versus log fold changes (FC) for differential expression between cells assigned to clone 2 and clone 1. Significant differentially expressed genes (FDR<0.1) are highlighted in red. (d) Enrichment of MSigDB Hallmark gene sets using camera [175] based on log2 FC values between clone 2 and clone 1 as in c. Shown are negative log10 P values of gene set enrichments, considering whether gene sets are up-regulated in clone 1 or clone 2, with significant (FDR < 0.05) gene sets highlighted and labelled. All results are based on 78 out of 79 cells that could be confidently assigned to one clone (posterior P>0.5).

detected in at least three of the 31 lines. Comparison to data with permuted gene labels demonstrates an excess of recurrently differentially expressed genes compared to chance expectation (Figure. 3.9c, P<0.001; 1,000 permutations - Appendix C; Methods). We also identify a small number of genes that contain somatic variants in a subset of clones, resulting in differential expression between wild-type and mutated clones (Appendix C; Supplementary Figure S34).

To investigate the transcriptomic changes between cells in more detail, we used gene set enrichment analysis in each line. This approach reveals whether there is functional convergence at a pathway level (using MSigDB Hallmark gene sets; Methods; [176]). Of 31 lines tested, 19 have at least one significant MSigDB Hallmark gene set (FDR<0.05, camera; Methods), with key gene sets related to cell cycle and growth being significantly enriched in all of those 19 lines. Directional gene expression changes of gene sets for the E2F targets, G2M checkpoint, mitotic spindle and MYC target pathways are highly coordinated (Figure. 3.9d), despite limited overlap of individual genes between the gene sets (Appendix C; Supplementary Figure S35).

Similarly, directional expression changes for pathways of epithelial-mesenchymal transition (EMT) and apical junction are correlated with each other. Interestingly, these are anti-correlated with expression changes in cell cycle and proliferation pathways (Figure. 3.9d). Within individual lines, the enrichment of pathways often differs between pairs of clones, highlighting the variability in effects of somatic variants on the phenotypic behaviour of cells (Figure. 3.9e).

These consistent pathway enrichments across a larger set of donors point to somatic variants commonly affecting the cell cycle and cell growth in fibroblast cell populations. These results indicate both deleterious and adaptive effects of somatic variants on proliferation, suggesting that a significant fraction of these variants are non-neutral in the majority of donors in our study.

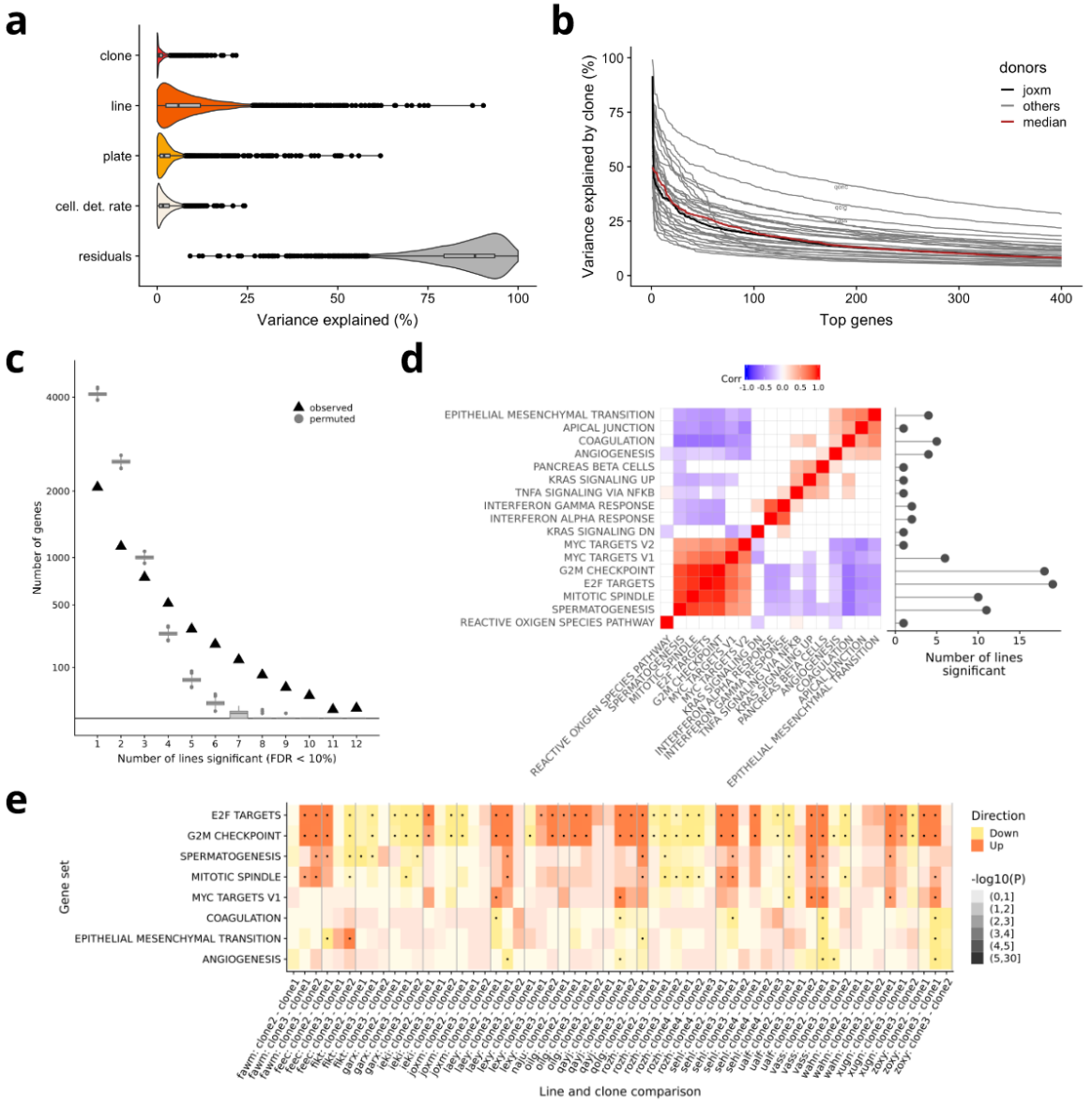


Fig. 3.9 Signatures of transcriptomic clone-to-clone variation across 31 lines. (a) Violin and box plots show the percentage of variance explained by clone, line, experimental plate and cellular detection rate for 4,998 highly variable genes, estimated using a linear mixed model (Methods; Appendix C). (b) Percentage of gene expression variance explained by clone when fitting a linear mixed model for each individual line for the 400 genes with the most variance explained by clone per line (Methods; Appendix C). Individual lines correspond to cell lines (donors), with joxm highlighted in black and the median across all lines in red. (c) The number of recurrently differentially expressed (DE) genes between any pair of clones (FDR<0.1; edgeR QL F test), detected in at least one to 12 lines, with box plots showing results expected by chance (using 1,000 permutations). (d) Left panel: Heatmap showing pairwise correlation coefficients (Spearman R, only nominal significant correlations shown (P<0.05)) between signed P-values of gene set enrichment across lines, based on differentially expressed genes between clones. Shown are the 17 most frequently enriched MSigDB Hallmark gene sets. Right panel: number of lines in which each gene set is found to be significantly enriched (FDR<0.05). (e) Heatmap depicting signed P-values of gene set enrichments for eight Hallmark gene sets in 19 lines. Dots denote significant enrichments (FDR<0.05).

3.5 Discussion

Within the fibroblast categorisation, several types of fibroblast have been defined within the skin [136]. Studies into expression of collagen and proteoglycans with immunohistochemistry have revealed differences between papillary and reticular layers [177, 178] although differences in fibroblast type are confounded by other differences in these layers. However, taking an explant culture allows isolation of papillary and reticular fibroblasts. Applying this approach to human dermis has identified several differences between these fibroblast types, such as rate of cell division [179, 180] and expression of collagens and proteoglycans [181].

In this chapter, I have shown isolation of the fibroblasts used in my work to one subtype of fibroblast. However, given the additional complexity within the skin, it is important to consider that studying one fibroblast type alone will not illuminate the full *in vivo* role of fibroblast innate immune response in the skin. Homogeneity in the resting state provides the benefit of a standardised experimental system, particularly key when conducting experiments across many donors. However, it is important to place any findings within the full dermal context, taking into account both fibroblast heterogeneity and the interaction between fibroblasts and the remaining cell types within the local environment.

Within the *in vitro* fibroblasts assayed, the largest source of variation in the scRNA-seq data derived from experimental batch. However, after integrating experimental batches, I showed that the largest source of biological heterogeneity in the dataset arises from cell cycle effect. Partitioning of cells highlighted clusters of cycling and non-cycling cells. In the latter, clusters showed enrichment for GO terms relating to cell-to-cell communication and involvement in the extracellular matrix, reflecting the role of fibroblasts within the wider tissue environment.

Considering intra-individual genetic variability within the fibroblast populations profiled, we identified clonal structure in 32 of the fibroblast lines for which WES data was available. Harnessing transcriptomic information for cells assigned to clones, we identified substantial and convergent gene expression differences between clones across lines. Analysis of clonal evolutionary dynamics using somatic variant allele frequency distributions revealed evidence for positive selection of clones in ten of 32 lines. These results support previous observations of clonal populations undergoing positive selection in normal human eyelid epidermis assayed by targeted DNA sequencing [138, 172, 182].

We shed light on the phenotypic effects of this adaptive evolution, identifying differential expression of gene sets implicated in proliferation and cancer such as the E2F and MYC pathways. This surprising result in healthy tissue suggests pervasive inter-clonal phenotypic variation with important functional consequences, although clonal dynamics *in vivo* in primary tissue may differ from what we observe in the fibroblast cell lines. It is intriguing to speculate about potential mechanisms driving these inter-clonal phenotypic differences, which might stem solely from observed somatic variants, could involve unobserved variants, or could arise through indirect mechanisms involving (post-)transcriptional regulation or epigenetic differences. Further work is needed to identify drivers of molecular differences between clones.

Chapter 4

Cell-to-cell variability in the innate immune response

> ***Declaration***
>
> *The cross-mammalian dataset presented in Section 4.1 was produced by Tzachi Hagai. This work was published in Nature, 2018, and the full paper is included in Appendix D. Analysis of bulk RNA-seq data and calculation of response divergence were conducted by Tzachi, along with production of final figures for the manuscript.*

4.1 Introduction

The innate immune system acts as a first line of defence across cell types and species, inhibiting pathogen replication and signalling pathogen presence to other cells. A key feature of this response is the rapid evolution that many of the genes have undergone along the vertebrate lineage, often attributed to pathogen-driven selection. As described in Chapter 1.3, another characteristic of the response is the high level of heterogeneity among responding cells, however the functional importance of this variability is unclear.

These two characteristics - rapid evolutionary divergence and high cell-to-cell variability — seem to be at odds with the strong regulatory constraints imposed on the host immune response: the need to execute a well-coordinated and carefully balanced programme to avoid tissue damage and pathological immune conditions. How this tight regulation is maintained despite rapid evolutionary divergence and high cell-to-cell variability remains unclear, but it is central to our understanding of the innate immune response and its evolution.

In this chapter, I present two angles of this question. Firstly, in a study led by Tzachi Hagai, we studied the evolution of the innate immune programme using two cells types — fibroblasts and mononuclear phagocytes — in different mammalian clades challenged with several immune stimuli. The results presented here focus on the fibroblast results; the experimental methods are described in Chapter 2.4. I then go on to use a larger human scRNA-seq dataset, described in Chapter 2.2, to define the dynamics of the response at a single cell resolution, characterising response gene modules.

4.2 Innate immune variability: a cross-mammalian study

4.2.1 Transcriptional divergence in immune response

First, we studied the transcriptional response of fibroblasts to stimulation with dsRNA (poly(I:C)) across the four species (human, macaque, rat and mouse). Bulk RNA-sequencing (RNA-seq) data was generated for each species after 4 h of stimulation, along with respective controls (Figure 4.1a).

In all species, dsRNA treatment induced rapid upregulation of genes that encode expected antiviral and inflammatory products, including IFN-β, TNF, IL1A and CCL5 (Figure 4.1b). A similar transcriptional response between species was observed when considering one-to-one orthologues (Spearman correlation, $P < 10^{-}10$ in all comparisons), as reported in other immune contexts [183–185]. Furthermore, as seen in other expression programmes [186–188], the response tended to be more strongly correlated between closely related species than between more distantly related species (Appendix D; Extended Data Figure 1).

Using these cross-species bulk transcriptomics data, we characterized the differences in response to dsRNA between species for each gene. While some genes, such as those encoding the NF-B subunits RELB and NFKB2, respond similarly across species, other genes respond differently in the primate and rodent clades (Figure 4.1c). For example, Ifi27 (which encodes a restriction factor against numerous viruses) is strongly upregulated in primates but not in rodents, whereas Daxx (which encodes an antiviral transcriptional repressor) exhibits the opposite behaviour.

To quantify transcriptional divergence in immune responses between species, we focused on genes that were differentially expressed during the stimulation (see Appendix

D; Methods) referred to as 'responsive genes' (Figure 4.1d). In this analysis, we study the subset of these genes with one-to-one orthologues across the studied species, of which there are 955 such responsive genes in dsRNA-stimulated human fibroblasts. We define a measure of response divergence by calculating the differences between the fold-change estimates while taking the phylogenetic relationship into account (Appendix D; Methods).

For subsequent analyses, we split the 955 responsive genes into three groups on the basis of their level of response divergence: (1) high-divergence dsRNA-responsive genes (the top 25% of genes with the highest divergence values in response to dsRNA across the four studied species); (2) low-divergence dsRNA-responsive genes (the bottom 25%); and (3) genes with medium divergence across species (the middle 50%; Figure 4.1d).

4.2.2 Cell-to-cell variability in immune response

As described in Chapter 1.3, previous studies have shown that the innate immune response displays high variability across responding cells. However, the relationship between cell-to-cell transcriptional variability and response divergence between species is not well understood. To study heterogeneity across individual cells, single cell RNA-seq was performed in all species in a stimulation time course (Figure 4.1a).

Cell-to-cell variability was quantitatively measured using an established measure for variability: distance to median (DM) [25]. We found a clear trend in which genes that were highly divergent in response between species were also more variable in expression across individual cells within a species (Figure 4.2); observed across the stimulation time points and in different species.

Next, we examined the relationship between the presence of promoter elements (CpG islands - CGIs - and TATA-boxes) and a gene's cell-to-cell variability. Genes that are predicted to have a TATA-box in their promoter had higher transcriptional

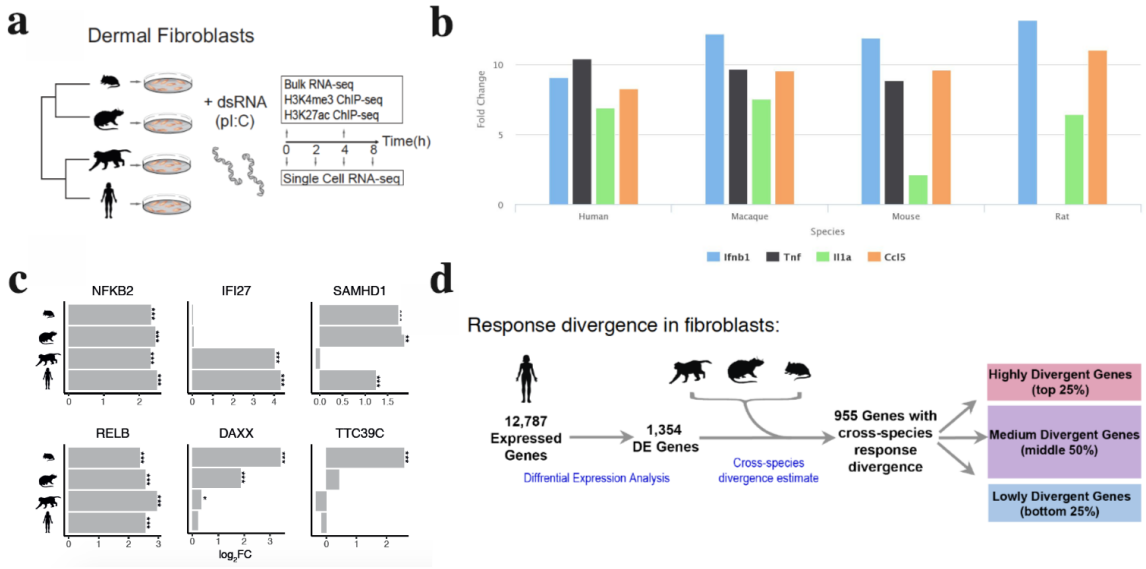


Fig. 4.1 Response divergence across species in innate immune response. a) Study design. Primary dermal fibroblasts from mouse, rat, human and macaque stimulated with dsRNA or controls. Samples were collected for bulk and single cell RNA-seq and ChIP–seq. b) Fold change of example genes (IFNB1, TNF, IL1A and CCL5) across the four species after 4h dsRNA stimulation. c) Fold-change (FC) after 4h dsRNA stimulation in fibroblasts for sample genes across species (edgeR exact test, based on n = 6, 5, 3 and 3 individuals from human, macaque, rat and mouse, respectively). False discovery rate (FDR)-corrected P values are shown (***P<0.001, **P<0.01, *P<0.05). d) Estimating each gene's level of cross-species divergence in transcriptional response to dsRNA stimulation. Using differential expression analysis, fold-change in dsRNA response was assessed for each gene in each species. We identified 1,358 human genes as differentially expressed (DE) (FDR-corrected q<0.01), of which 955 had one-to-one orthologues across the four studied species. For each gene with one-to-one orthologues across all species, a response divergence measure was estimated using: response divergence = $log[1/4 \text{ x } \Sigma_i, j(log[FCprimate_i] - log[FCrodent_j])^2]$. Genes were grouped into low, medium and high divergence according to their response divergence values for subsequent analysis.

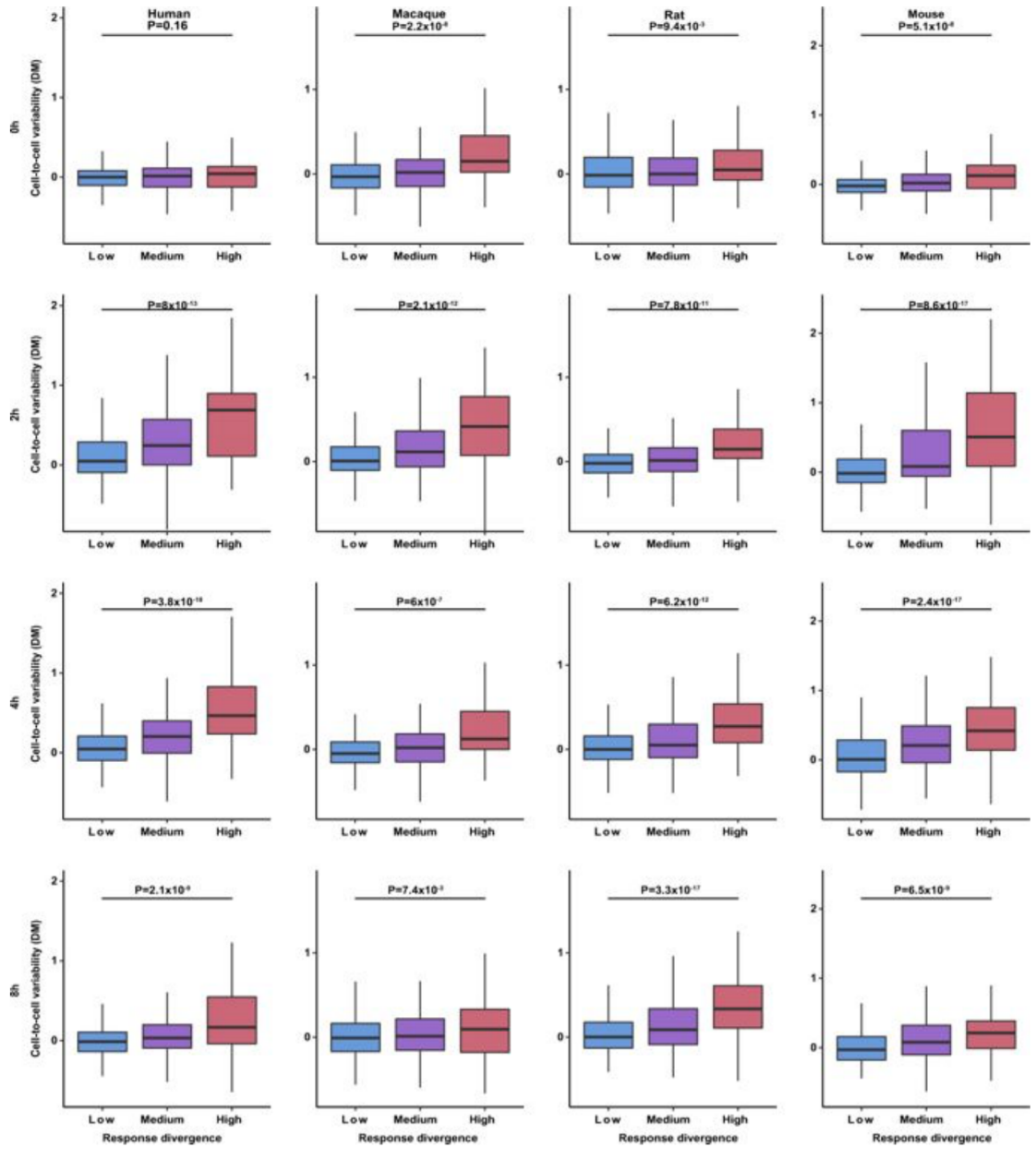


Fig. 4.2 Cell-to-cell variability versus response divergence across species and conditions. Cell-to-cell variability values, as measured with DM across individual cells, compared with response divergence between species (grouped into low, medium and high divergence). Variability values are based on n =29, 56, 55, 35 human cells, n = 20, 32, 29, 13 rhesus cells, n = 33, 70, 65, 40 rat cells, and n = 53, 81, 59, 30 mouse cells, stimulated with dsRNA for 0, 2, 4 and 8 h, respectively. Rows represent different time points (0, 2, 4 and 8 h), and columns represent different species. High-divergence genes were compared with low-divergence genes using a one-sided Mann–Whitney test. Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the IQR.

variability, whereas CGI-containing genes tended to have lower variability (Figure 4.3a), in agreement with previous findings [189]. This finding also applied to transcriptional divergence between species (Figure 4.3b), showing that both these characteristics are associated with the presence of specific promoter elements.

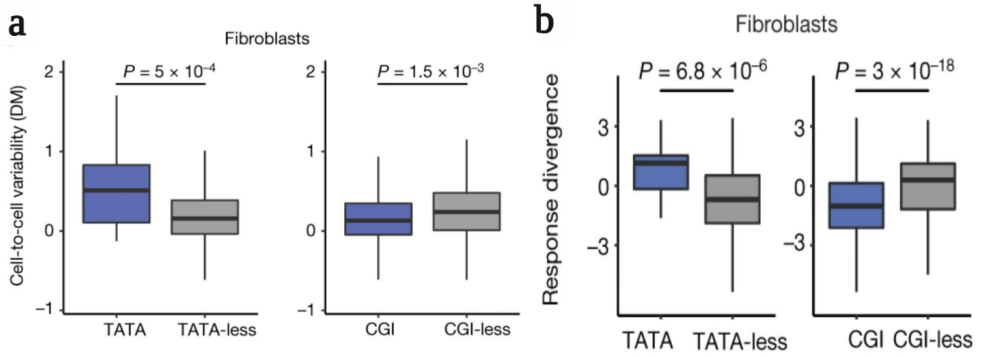


Fig. 4.3 Promoter architecture versus transcriptional divergence and variability. a) Comparison of cell-to-cell variability of genes with and without a TATA-box and a CGI (one-sided Mann–Whitney test). Cell-to-cell variability values are from DM estimations of human fibroblasts stimulated with dsRNA for 4 h (n = 55 cells). b) Comparison of divergence in response of genes with and without a TATA-box and a CGI in fibroblast dsRNA stimulation.

4.2.3 Transcriptional divergence and variability of cytokines

We next investigated whether different functional classes among responsive genes are characterized by varying levels of transcriptional divergence. To this end, we divided responsive genes into categories according to function (such as cytokines, transcriptional factors and kinases) or the processes in which they are known to be involved (such as apoptosis or inflammation). Genes related to cellular defence and inflammation—most notably cytokines, chemokines and their receptors (hereafter 'cytokines')—tended to diverge in response significantly faster than genes involved in apoptosis or immune regulation (chromatin modulators, transcription factors, kinases and ligases) (Figure 4.4).

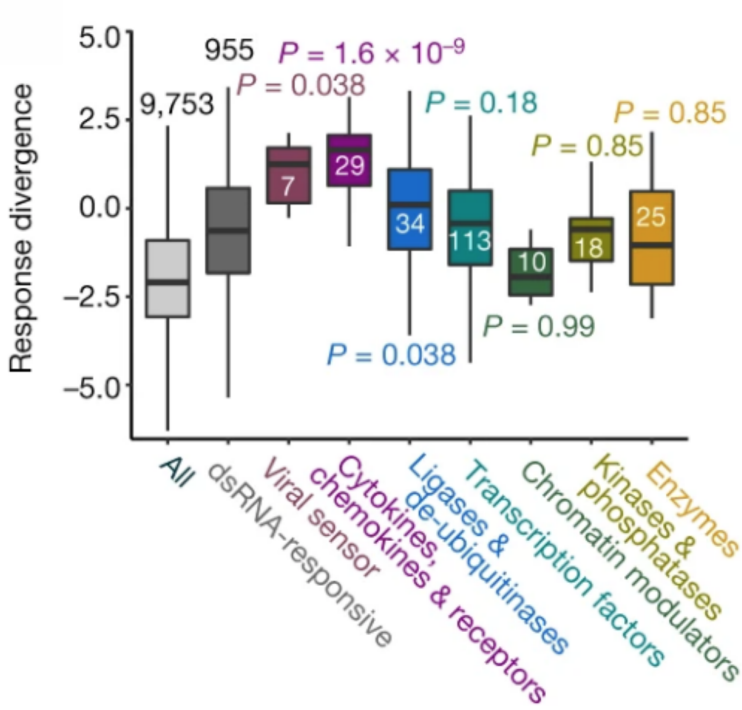


Fig. 4.4 Transcriptional divergence in genes of different functional categories. Distributions of divergence values of 9,753 expressed genes in fibroblasts, 955 dsRNA-responsive genes and different functional subsets of the dsRNA-responsive genes (each subset is compared with the set of 955 genes using a one-sided Mann–Whitney test and FDR-corrected P values are shown).

We subsequently compared the response divergence across species with the transcriptional cell-to-cell variability of three groups of responsive genes with different functions: cytokines, transcription factors, and kinases and phosphatases (referred to as 'kinases'). In contrast to kinases and transcription factors, many cytokines display relatively high levels of cell-to-cell variability across time points (Figure 4.5a). Furthermore, these are expressed only in a small subset of responding cells (Figure 4.5b).This has previously been reported for several cytokines, as described in Chapter 1.3. Here, we find that cells show high levels of variability in expression of cytokines from several families (for example, IFN-β, CXCL10 and CCL2).

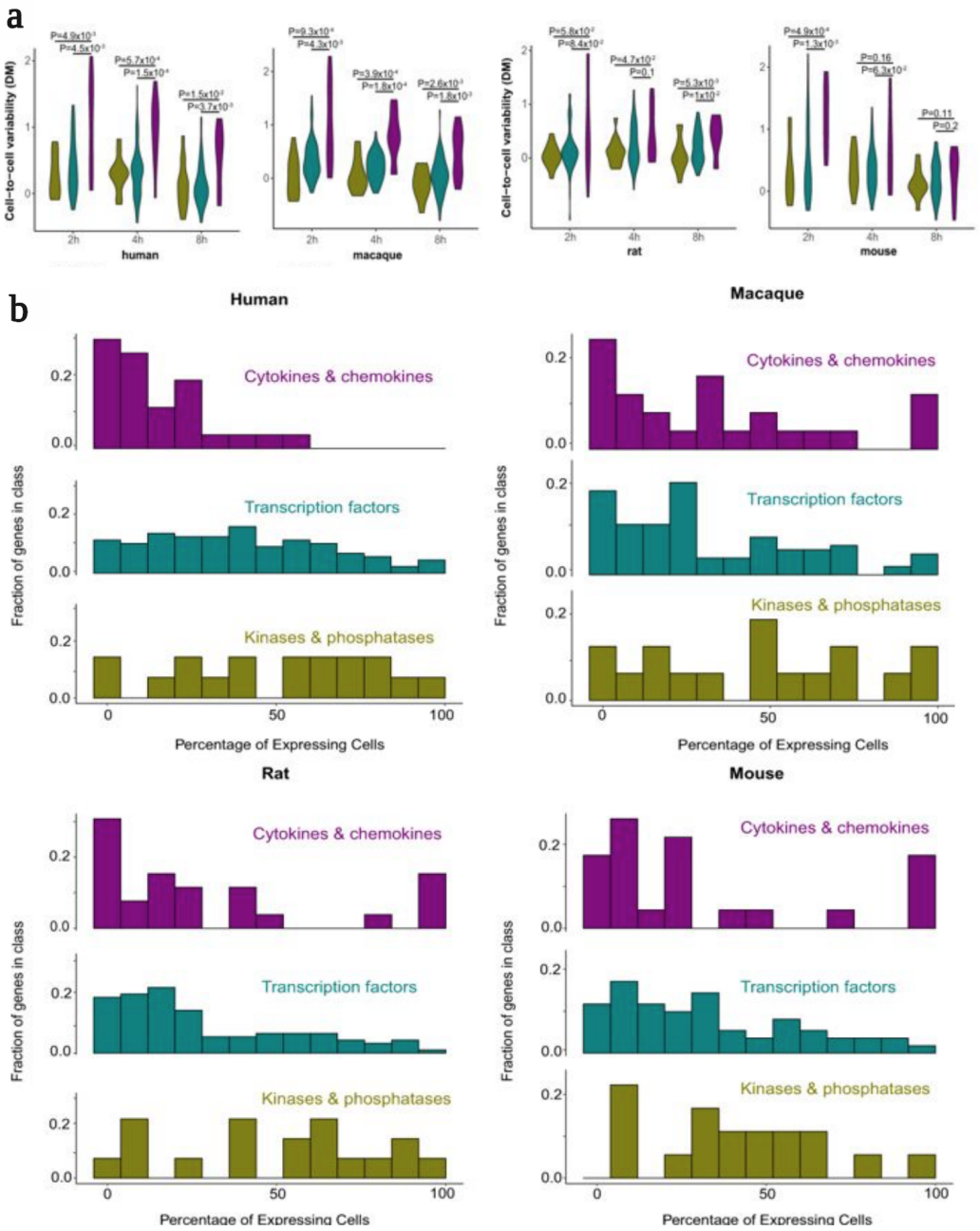


Fig. 4.5 Cell-to-cell variability levels in cytokines, transcription factors and kinases. a) Violin plots showing the distribution of cell-to-cell variability values (DM) of cytokines, transcription factors and kinases during a dsRNA stimulation time course in fibroblasts. Number of cells used in each species (at 2, 4, 8 h dsRNA, respectively): human, 56, 55, 35; macaque, 32, 29, 13; rat, 70, 65, 40; mouse, 81, 59, 30. Purple, cytokines; green, transcription factors; beige, kinases. Comparisons between groups of genes were performed using one-sided Mann–Whitney tests. Violin plots show the kernel probability density of the data. b) Histograms showing the percentage of fibroblasts expressing cytokines (top), transcription factors (middle) and kinases (bottom) following 4 h dsRNA stimulation, in human, macaque, rat and mouse cells. The percentage of expressing cells is divided into 13 bins (x-axis). The y-axis represents the fraction of genes from this gene class (for example, cytokines) that are expressed in each bin.

4.3 Characterising the Type I interferon response in human fibroblasts

4.3.1 Single-cell RNA-sequencing data

Having characterised variability in the innate immune response from an evolutionary perspective, the question of heterogeneity within the human population remains. In order to address this, a comprehensive dataset comprising both bulk and single cell RNA-sequencing data at two timepoints and with two stimulation conditions, along with a control, was generated - as described in Chapter 2.

The single cell dataset was filtered as described in Chapter 3.2, and UMAP dimensionality reduction was used to gain an oversight of the full dataset (Figure 4.6). It is clear to see that, as before, a major driver of variation is experimental batch effect, although cells also cluster by experimental condition. Once again, the 'integrate' function from the Seurat v3 package [167] was applied. This resulted in good mixing of the two batches in UMAP space, with experimental condition now being the major driver of variation in the dataset (Figure 4.6). The separation in unstimulated and interferon-treated cells seen in the 'condition' plot arises from cell cycle state, with cycling cells forming the cluster of mixed conditions on the left side of the plot.

4.3.2 The temporal dynamics of the response

Harnessing the resolution available within the single-cell data generated, it is possible to comprehensively study the innate immune response over time. Although both poly(I:C) and IFN-β induce antiviral signalling within treated cells, the two elicit different responses, as can be seen in Figure 4.6.

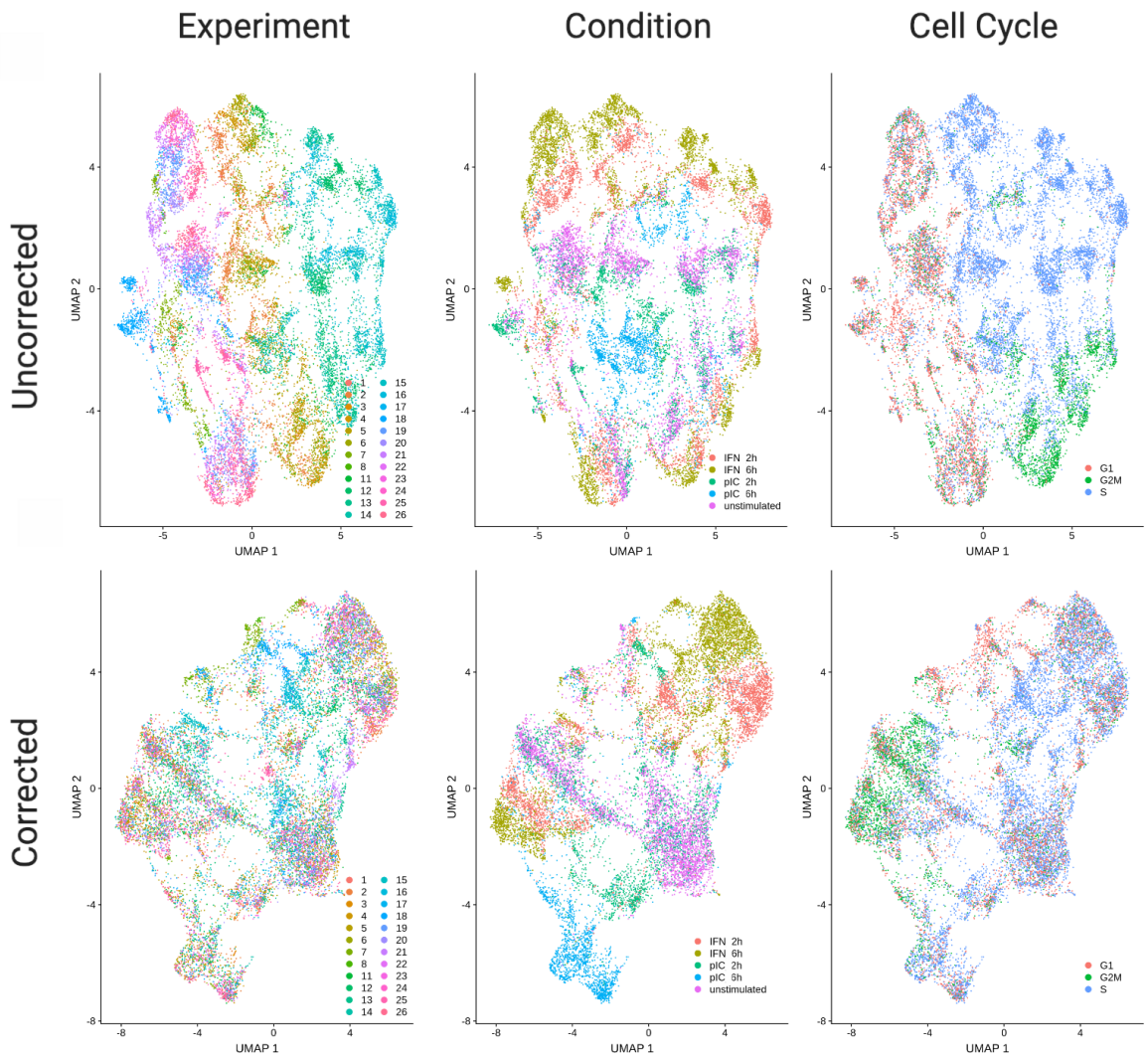


Fig. 4.6 Integration of scRNA-seq batches with Seurat. Dimensionality reduction using UMAP on uncorrected data (upper panel) and corrected data using Seurat v3's 'integrate' method (lower panel). Colours indicate, in order, experimental batch, stimulation condition, and cell cycle phase. The first two UMAP dimensions are shown.

In order to appropriately characterise the two response pathways, cells treated with poly(I:C) were separated from those treated with IFN-β. The two time points for each condition were considered together, along with control unstimulated cells. The benefit of combining all treated cells is particularly apparent after poly(I:C) stimulation, in which many cells after two hours of stimulation are transcriptionally similar to unstimulated cells, highlighting heterogeneity in this response.

To create a pseudotime, the destiny package was used, which employs a diffusion map approach [190]. This was applied to the 5000 most highly variable genes, calculated with Seurat's 'findVariableGenes' function, to the IFN-β and poly(I:C) pathways separately. Figure 4.7a shows Diffusion Components (DCs) 1 and 2 for each of these responses. This demonstrates that the largest source of variability, segregating along DC1, is stimulation condition. DC2 shows separation, particularly of unstimulated cells, representing cell cycle effects. This is confirmed by GO term enrichment analysis of the genes most highly correlated with DC2, along with visual inspection of the cell cycle phase distribution versus DC2 - shown in the inset plots in Figure 4.7a.

Given the correlation of DC1 with stimulation response in both treatment conditions, this is used as a 'response pseudotime'. In the case of poly(I:C) stimulation, the reverse of DC1 is taken as unstimulated cells lie on the right hand side. The distribution of cells in each stimulation condition across this response pseudotime highlights the heterogeneity in response (Figure 4.7b). This is particularly true in the response to poly(I:C) treatment, where the earlier timepoint shows a bimodality in the cells. Many of the cells show high similarity to the unstimulated state, while a subset are shifted to the right in response pseudotime, overlapping with the peak of the poly(I:C) 6 hour distribution (which itself has a broad distribution). In the IFN-β response pseudotime, the peak around the middle of DC1 corresponds to cell cycle state, however a breadth in distribution within responding cells can be seen.

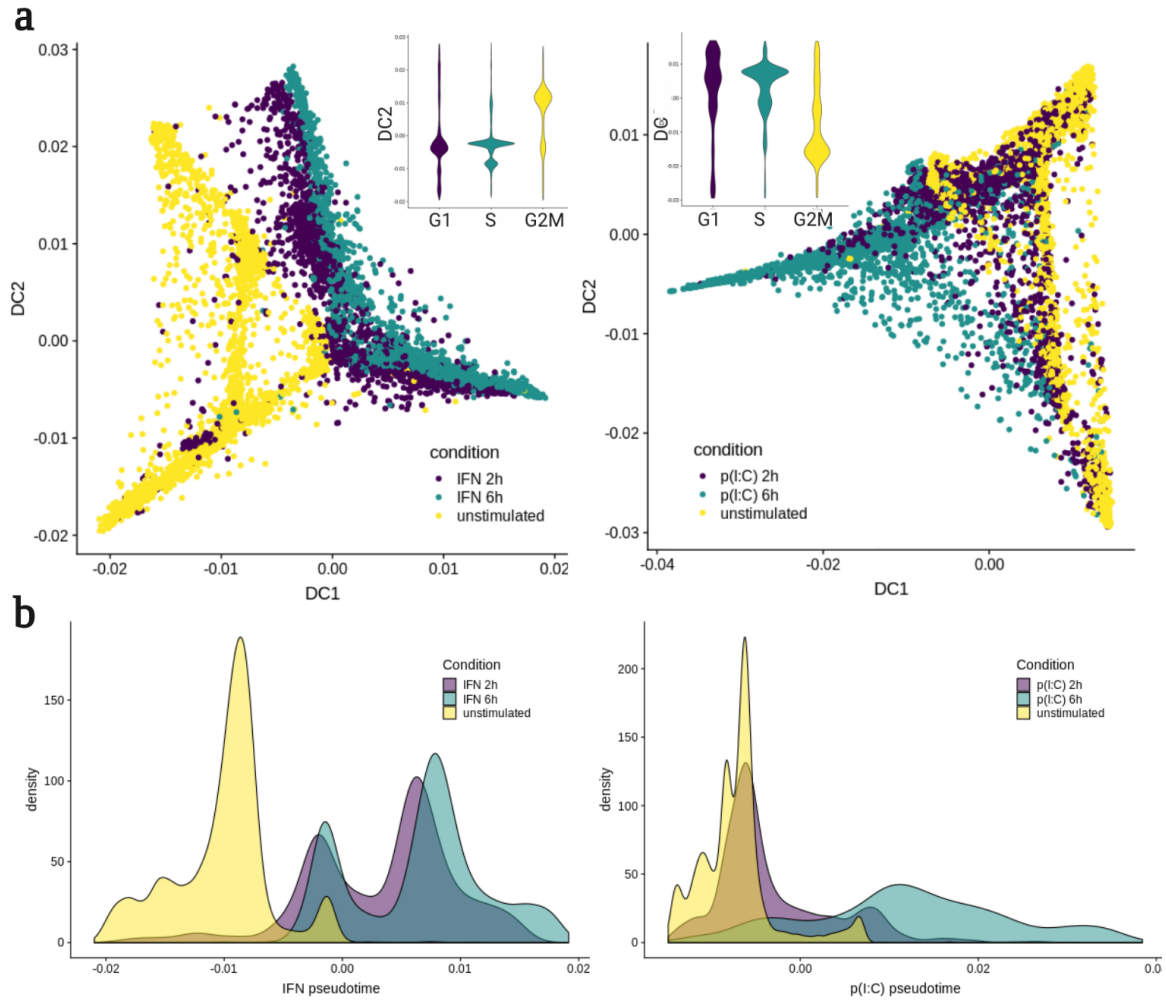


Fig. 4.7 A pseudotime of poly(I:C) and interferon response pathways. a) Cells plotted after dimensionality reduction with 'destiny' [190]; diffusion components (DCs) 1 and 2 shown for the 'IFN pathway', left, and 'poly(I:C) pathway', right, coloured by stimulation condition. Inset plots show the number of cells per cell cycle phase, assigned using 'cyclone' [174] against DC2. b) Density of cells from each stimulation condition across response pseudotimes, for the 'IFN pathway', left, and 'poly(I:C) pathway', right.

To confirm that the calculated pseudotimes capture the innate immune response, one can look at expression of known response genes, such as ISG15 and IFN-β (Figure 4.9a). It is worth noting that IFN-β treatment is not expected to induce IFN-β expression itself, and that in response to poly(I:C) treatment only a subset of cells produce IFN-β (rightmost panel), as discussed previously. Beyond example genes, it is possible to verify the expression of an entire innate immune response gene set. Deschamps *et al.* curated a set of 1553 innate immune genes (IIGs) from GO term annotation, InnateDB and manual addition [191]. These genes are classified into different functions, and examples of the genes and their annotated functions are shown in Figure 4.8. Looking at expression across the pseudotimes defined above, IIGs increase in the response to both IFN-β and poly(I:C) (Figure 4.9b), whereas the opposite is true for the remainder of genes (referred to as 'non-IIGs').

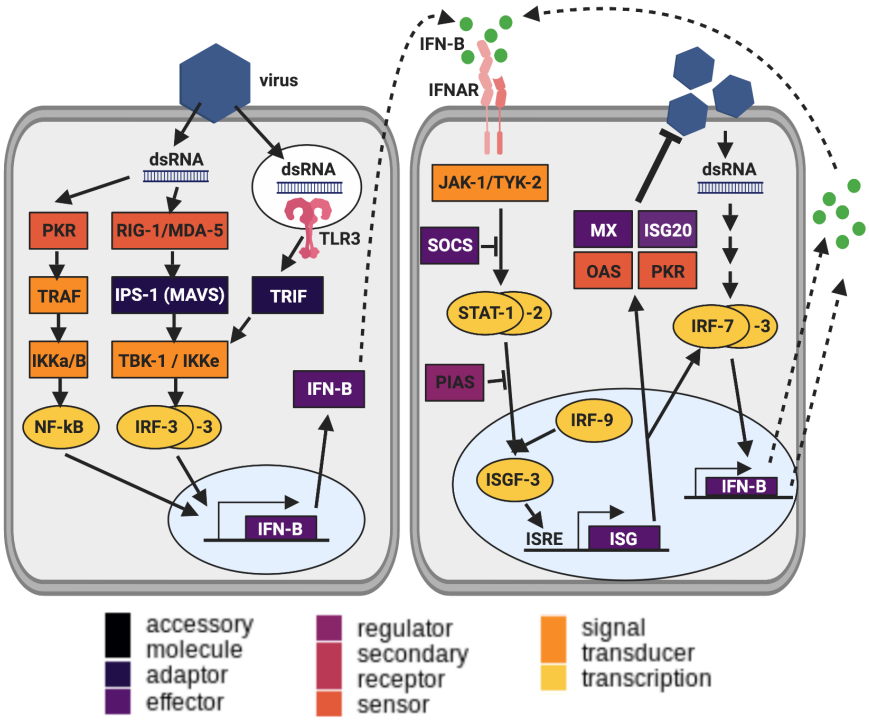


Fig. 4.8 Functional classification of innate immune genes. A curated list of innate immune genes (IIGs) was obtained from Deschamps *et al.* [191]. Examples involved in Type I interferon signalling, and their functional classification, are shown here.

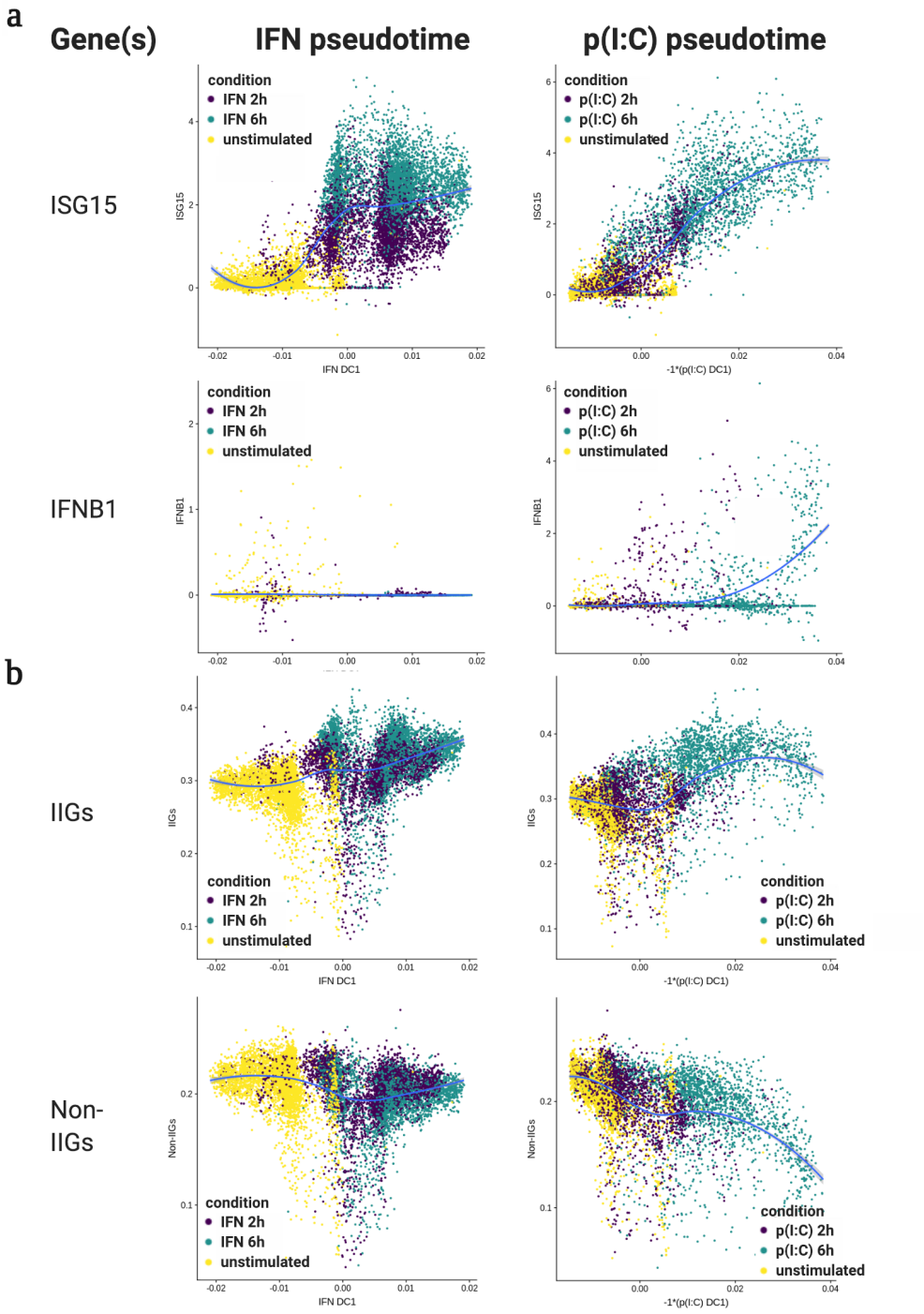


Fig. 4.9 Expression of innate immune genes across response pseudotime. a) The expression of ISG15 and IFNB1 against IFN pseudotime, left, and poly(I:C) pseudotime, right. b) Average expression of the set of innate immune genes (IIGs) [191] and non-IIGs over IFN pseudotime, left, and poly(I:C) pseudotime, right.

4.3.3 Defining gene modules in the innate immune response

To define the dynamics of gene expression for each gene independently, it is possible to fit a model to the expression of each gene across response pseudotime. Using the SwitchDE package (Appendix E; Figure E.1a) [97], parameters for the activation time (t0), expression level (μ) and slope of activation (k) were inferred for the IFN-β and poly(I:C) responses. Using these inferred models, it is possible to look at the pattern of gene expression across time. Genes were categorised as 'on' or 'off' for each response pathway based upon whether they had a positive or negative k value, respectively. The top 500 most significant genes (taking the q-value from the SwitchDE model) in each direction were considered, and hierarchical clustering was used to define clusters of genes with shared temporal expression patterns. The number of clusters was determined visually based upon the dendrograms of gene similarity, for the IFN-β and poly(I:C) pseudotimes respectively. These clusters show good concordance with modules defined using an alternative approach: WGCNA (Weighted gene correlation network analysis)[100]; Appendix E, Figure E.1.

Each module was tested for enrichment of the IIGs described above, and results are shown in Tables 4.1 and 4.2 for the response to IFN-β and poly(I:C) respectively. The distribution of functional categories for these IIGs was considered, and is shown in the right hand plots of Figures 4.10 and 4.11. Furthermore, GO term enrichment analysis was conducted for each cluster, and the list of significantly enriched terms (pvalue $<$ 0.05) is shown in Tables E.1 and E.2 (Appendix E).

In both the IFN-β and poly(I:C) response, there is one major cluster which represents the canonical Type I interferon response. In the case of IFN-β treatment, this is cluster coloured in black (Figure 4.10). This module of genes shows low expression in unstimulated cells (visible in the heatmap), a high enrichment of IIGs (hypergeometric test; pvalue = 1.98e-30), and inclusion of typical genes (such as DDX58, MYD88,

OAS3, ISG15, ISG20, IRF7, IFIT2, TRIM25, SAMHD1, IFI6, IFI35 and STAT1). The GO terms for this cluster reflect this signalling pathway, with the two most significant terms being "defense response to virus" and "type I interferon signaling pathway". This cluster has a particularly high representation of IIGs in the classes 'effector', 'regulator', and 'sensor'. highlighting functions across the pathway (Figure 4.10a; right panel). The same trend can be seen in the 'black' cluster in poly(I:C), which shows highest expression in later stages of the poly(I:C) response pseudotime. The most significant GO terms include "innate immune response", "defense response to virus" and "cytokine-mediated signaling pathway", and all of the example genes listed above fall within the cluster (with the exception of IRF7, however IRF9 is included). Again, IIGs are highly enriched (hypergeometric test; pvalue = 3.58e-42), and show functions across the pathway (Figure 4.11a; right panel).

Beyond these two major clusters, modules of genes with discrete innate immune response functions can be identified. For example, in response to IFN-β, there is a co-expressed set of genes (pink) which show involvement in signal transduction and regulation. This cluster includes genes such as DHX58, JAK2, STAT3 and TRADD. The third cluster, on the other hand, shows a higher level of effector function, with enrichment of GO terms relating to cytokine production, and genes such as CCl2, CXCL11 and CXCL16.

These alternative modules are less clear in the poly(I:C) response. Here, the second group of genes is a small cluster dominated by mitochondrial genes, which is reflected in the enriched GO terms. The two annotated IIGs within this cluster are IFITM2 and IFITM3, both of which are classified as 'effector' proteins. The third cluster, while enriched in IIGs, shows less ubiquitous expression in responding cells than cluster 1. There are no significant GO terms for this cluster, however members include genes

known to be involved in the type I interferon response, such as NFKBIA + NFKBID, SOCS3, CCL5, DDX3X and JUN, and particularly in signal transduction.

Along with categorising up-regulated gene sets, it is interesting to consider the set of genes down-regulated in response to the mock-viral stimulations. In response to IFN-β treatment, two major clusters of genes are down-regulated. One cluster, expressed in unstimulated cells but switched off in the response (Figure 4.10b; left panel) reflects processes around chromatin organisation and nucleic acid processing. Example genes in this cluster are HDAC2, SMARCA2, and ZNF287. The other cluster represents the cell cycle, with strongly enriched GO terms such as 'cell cycle' and 'DNA metabolic process'. Genes include CDK1, CCNA2, CDCA2, CCDC18, and several members of the CENP family.

In response to poly(I:C) stimulation, there are two major functions of down-regulated genes. The largest cluster of genes, which show decreased expression in responding cells, are involved in biological processes such as 'organelle organisation' and 'establishment of localisation in cell'. The second and third cluster are less clearly defined, however one cluster (pink) shows enrichment of GO terms highlighting metabolic processes, while the other centres on protein localisation and processing. Furthermore, these two modules show different temporal dynamics across the response pseudotime (Figure 4.11b; left panel).

The definition of these modules across response pseudotimes highlights a tightly regulated type I interferon response, with coordinated modules of genes showing discrete innate immune functions.

Table 4.1 Enrichment of IIGs in modules of co-expressed genes in the IFN-β response.

Gene module	Total group size	Number of IIGs	Enrichment p-value
Canonical Type I IFN	81	44	1.98e-30
Regulator/signal transduction	197	53	4.66e-18
Effector	222	40	2.65e-08
Cell cycle	244	19	0.33
Chromatin organisation	256	16	0.69

Table 4.2 Enrichment of IIGs in modules of co-expressed genes in the poly(I:C) response.

Group	Total group size	Number of IIGs	Enrichment p-value
Canonical Type I IFN	311	103	3.58e-42
Mitochondrial	25	2	0.27
Signal transduction	164	24	0.00032
Organelle localisation	298	26	0.15
Metabolic processes	127	21	0.00011
Protein regulation	75	10	0.018

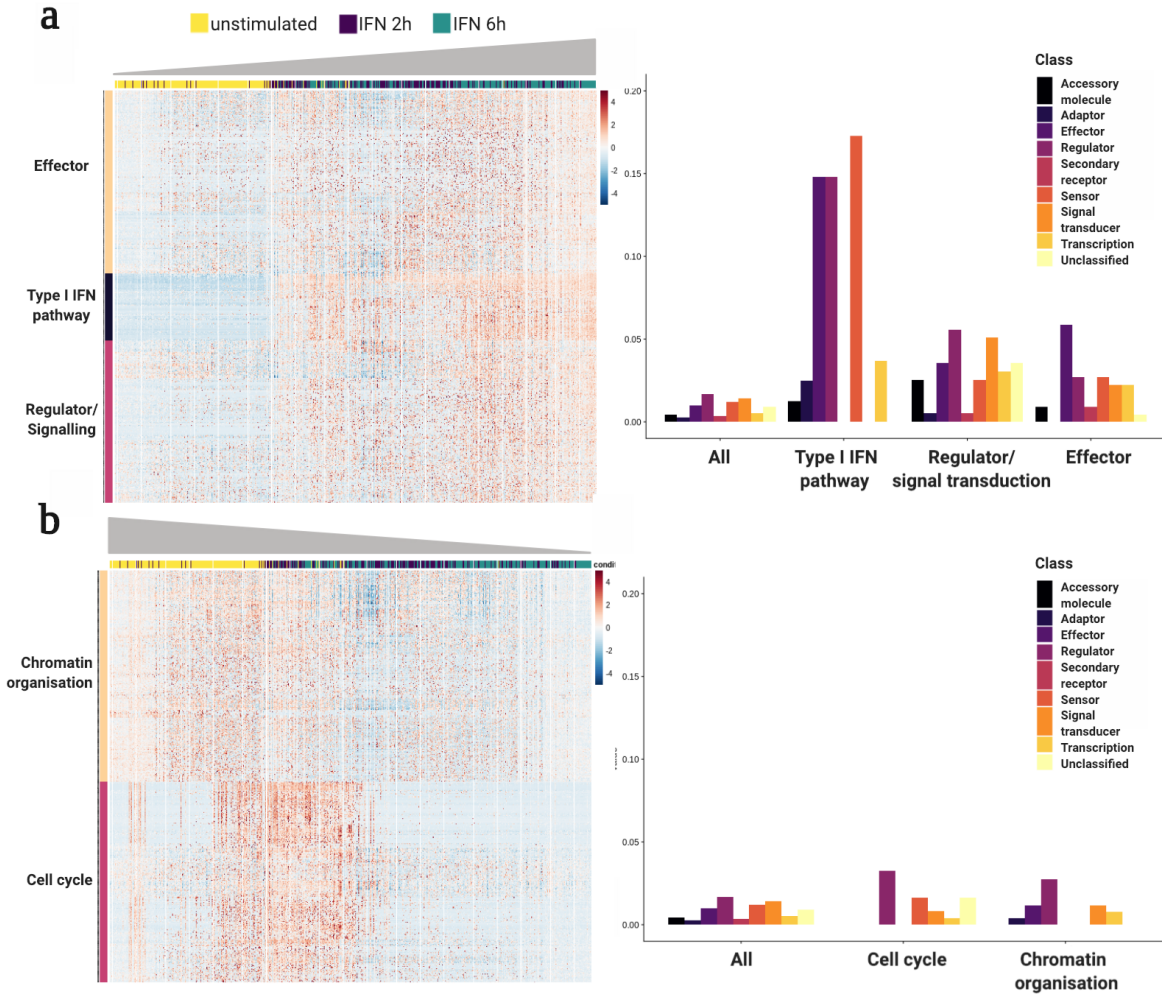


Fig. 4.10 Modules of co-expressed genes in the response to IFN-β. The SwitchDE package [97] was used to infer a dynamic model of expression for each gene. Genes with a positive 'k' value were termed 'on' genes and those with a negative 'k' as 'off' genes, shown in panels a and b respectively. The 500 'on' and 'off' genes with the most significant qvalue were selected. Their z-score normalised expression across the pseudotime defined in Figure 4.7 is shown on the left; genes were clustered using hierarchical clustering with the ward method. Right: proportion of genes from each IIG functional category within the total cluster; the background set shows representation in the entire set of 15363 genes tested in SwitchDE.

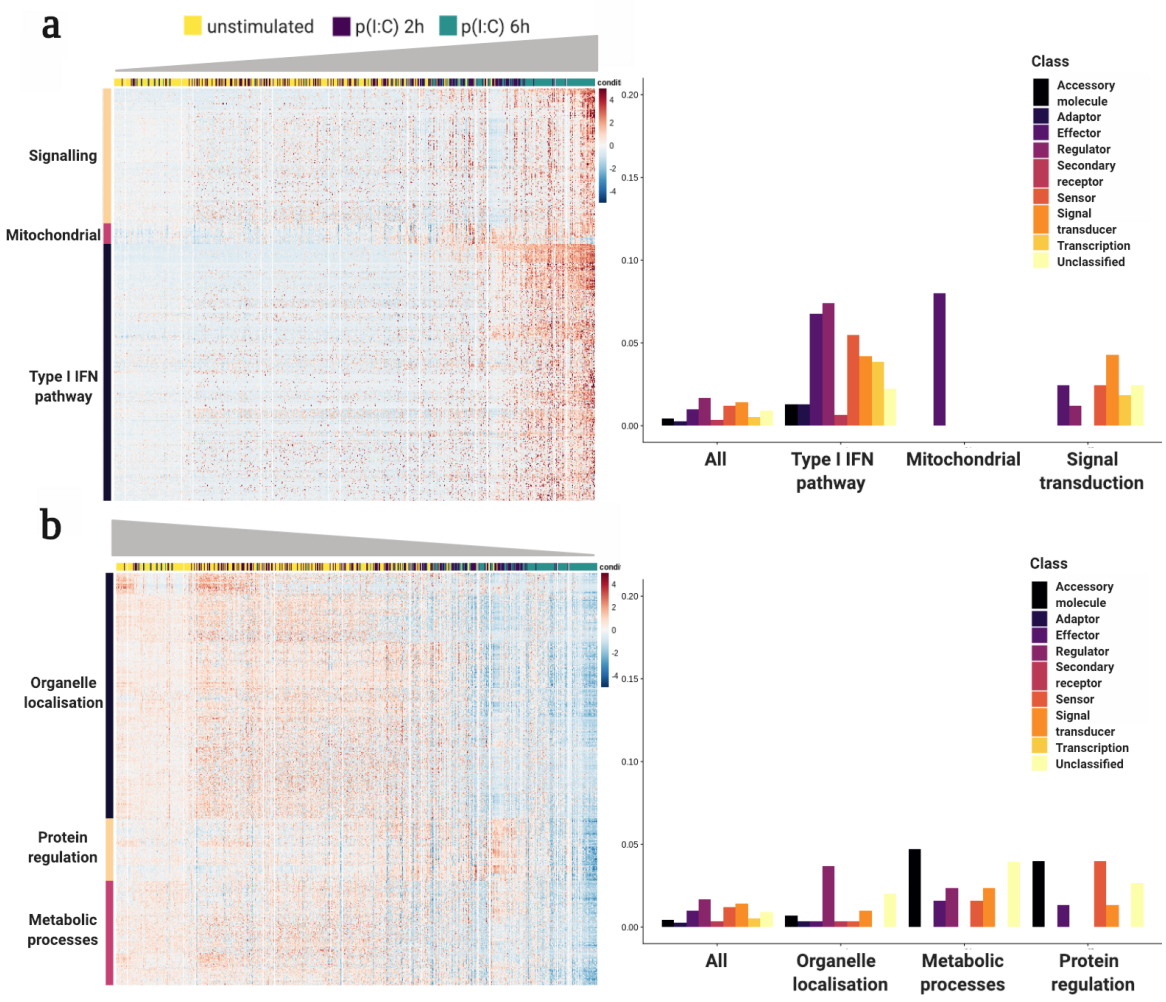


Fig. 4.11 Modules of co-expressed genes in the response to poly(I:C). The SwitchDE package [97] was used to infer a dynamic model of expression for each gene. Genes with a positive 'k' value were termed 'on' genes and those with a negative 'k' as 'off' genes, shown in panels a and b respectively. The 500 'on' and 'off' genes with the most significant qvalue were selected. Their z-score normalised expression across the pseudotime defined in Figure 4.7 is shown on the left; genes were clustered using hierarchical clustering with the ward method. Right: proportion of genes from each IIG functional category within the total cluster; the background set shows representation in the entire set of 15363 genes tested in SwitchDE.

4.4 Discussion

In this chapter, I have described work charting the evolutionary architecture of the innate immune response. We showed that genes that diverge rapidly between species show higher levels of variability in their expression across individual cells than genes that diverge more slowly. Both of these characteristics are associated with a similar promoter architecture, enriched in TATA-boxes and depleted of CGIs. Notably, such promoter architecture is also associated with the high transcriptional range of genes during the immune response. Thus, transcriptional changes between conditions (stimulated versus unstimulated), species (transcriptional divergence), and individual cells (cell-to-cell variability) may all be mechanistically related to the same promoter characteristics. In yeast, TATA-boxes are enriched in promoters of stress-related genes, displaying rapid transcriptional divergence between species and high variability in expression [192, 193]. This finding suggests intriguing analogies between the mammalian immune and yeast stress responses—two systems that have been exposed to continuous changes in external stimuli during evolution.

We have also shown that genes involved in regulation of the immune response—such as transcription factors and kinases—are relatively conserved in their transcriptional responses. These genes might be under stronger functional and regulatory constraints, owing to their roles in multiple contexts and pathways, which would limit their ability to evolve. This limitation could represent an Achilles' heel that is used by pathogens to subvert the immune system. Cytokines, on the other hand, diverge rapidly between species, owing to their promoter architecture and because they have fewer constraints imposed by intracellular interactions or additional non-immune functions. Cytokines may therefore represent a successful host strategy to counteract rapidly evolving pathogens as part of the host–pathogen evolutionary arms race.

Cytokines also display high cell-to-cell variability and tend to be co-expressed with other cytokines and cytokine regulators in a small subset of cells, and this pattern is conserved across species. As prolonged or increased cytokine expression can result in tissue damage [194–196], restriction of cytokine production to only a few cells may enable a rapid, but controlled, response across the tissue to avoid long-lasting and potentially damaging effects. This cellular variability in response is also observed in the larger human scRNA-seq dataset, where cells in each stimulation condition show a wide distribution of positions across the IFN-β and poly(I:C) response pseudotimes. This further strengthens the notion that the response is heterogenous but highly regulated.

One mechanism to achieve a strongly coordinated response is the regulation of gene modules with discrete functions. By characterising genes whose expression changes across response stimulation, I showed that it is possible identify distinct modules. In both stimulation timepoints, a gene module representing the canonical type I interferon pathway was observed. Further discrete gene clusters were seen, such as those involved in signalling or effector functions. These modules showed differences in temporality and variability of expression. For example, the 'effector' module showed less ubiquitous expression across cells in response to IFN-β treatment compared to the type I interferon module, mirroring the cytokine heterogeneity seen in the cross-mammalian work. These features further suggest tight regulation of expression within each gene set, and across the response as a whole.

Chapter 5

Inter-individual variability in the innate immune response

> ***Declaration***
>
> *QTL analysis was conducted using a pipeline developed by Marc Jan Bonder (Stegle Group, EMBL-EBI). The pipeline was run with the support of Ni Huang (Teichmann Group, WSI).*

5.1 Introduction

Alongside characterising the innate immune response across the scRNA-seq dataset as a whole, as seen in the previous chapter, it is possible to consider the variability between donors, illuminating differences in response to infections within the healthy human population. One method to do so is by fitting a linear mixed model to each gene, partitioning variation in gene expression into components. This can give insight into the source of variation within a dataset globally, but doesn't pinpoint any effects that may derive from specific genetic differences.

To elucidate the effect of genetic variation, an alternative approach that is commonly used is the eQTL approach, as described in Chapter 1.1. In the case of scRNA-sequencing data, it is possible to generate a mean expression value in two ways: either averaging across all cells to create a single 'pseudobulk' expression level per donor, or treating each cell from a donor as an independent replicate. While the latter approach increases sample size, it comes at the price of increased noise and computational cost. For this reason, the 'pseudobulk' expression value was used in this work.

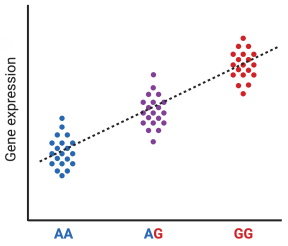
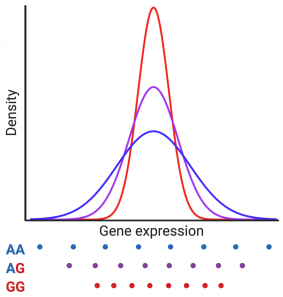
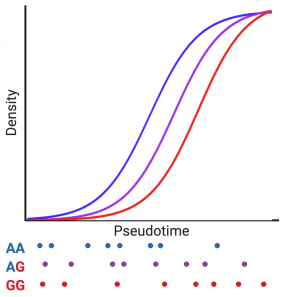
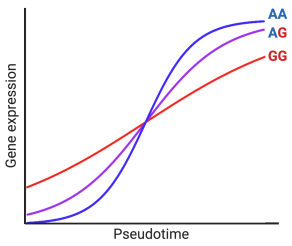
Alongside mean expression level, scRNA-seq also opens up the possibility of identifying variation in the heterogeneity of expression within each donor. For the current study, the simple metric of variance of each gene per donor per response pathway was calculated. As discussed in Chapter 1.2, however, there are alternate methods to reflect variability in expression, such as BASiCS [88] and DM [25].

Beyond metrics around the average and variance of expression between cells, it is possible to define alternative phenotypes capturing the difference in response between individuals (Table 5.1). One example is the proportion of cells expressing a gene, particularly of importance for cytokines and other signalling molecules that show stochastic expression across cells (Chapter 1.3).

Another element of variability is not just the level of expression, but the temporality of this. It may be the case that certain donors respond 'earlier' than others - a phenotype that is not captured by considering expression alone. However, this only provides one phenotype per donor, rather than a per-gene value allowing testing against the specific gene in question. The dynamics of expression may be inferred on a individual gene basis, for example using the SwitchDE package [97]. This infers parameters for the activation time (t0), expression level (μ) and slope of activation (k) (Table 5.1).

In this chapter, I describe the application of these approaches to the IFN-β and poly(I:C) stimulation dataset previously described, using both bulk and single cell RNA sequencing. By considering variability in gene expression within these data, I aim to identify a genetic basis for differences in innate immune response between individuals.

Table 5.1 Phenotypes derived from scRNA-seq data.

Phenotype	Description	Schematic representation
Mean	The mean expression level per gene per sample. Cells from each donor and condition are averaged to produce a 'pseudobulk'.	
Variance	The variance of expression per gene per sample. As above, cells from each donor and condition are combined together.	
Cell proportion	The proportion of cells expressing a gene, per donor and condition	
Average pseudotime	For each donor, the average of all cells for the IFN and poly(I:C) pseudotimes	
SwitchDE parameters	Applying 'switchDE' [97] to each donor for the IFN and poly(I:C) pathway, inferring parameters t0, μ and k per gene	

5.2 | Variance partitioning of gene expression

To investigate variability within the scRNA-seq data, a variance partitioning approach, as described above, was taken. This was applied to the 5000 most highly variable genes, using the variancePartition package [197]]. The components of 'donor', 'condition' and 'log10(counts)' were included in the model, along with a residual noise component.

Figure 5.1a highlights the large proportion of variance explained by residual noise in this single cell dataset. Despite this, there are 674 genes for which the donor component explains more than 5% of variance in gene expression, and 362 genes for which condition explains more than 5% of variance. The latter threshold may be used to define a set of 'response' genes - those that vary most with stimulation conditions.

While many genes show large variance explained by donor or condition independently (Figure 5.1b), there are 139 genes for which both components explain more than 5% of gene expression variation. This shows promise for the ability to identify genes that vary between individuals in a stimulation-specific manner.

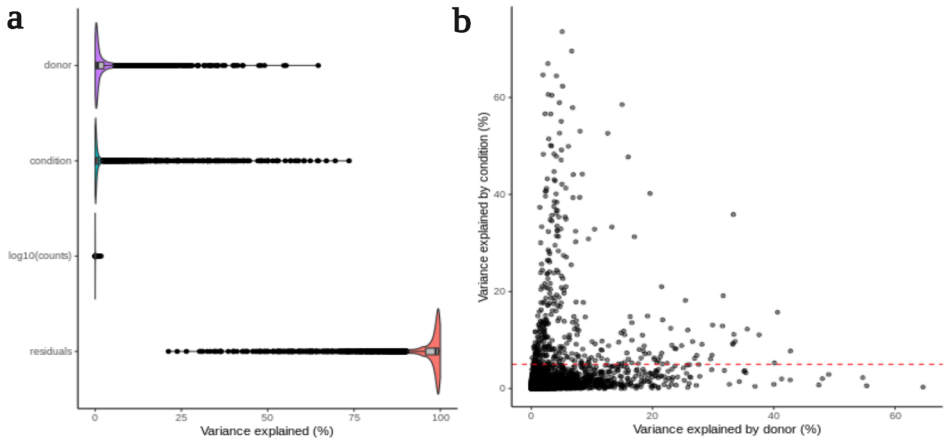


Fig. 5.1 Variance partitioning of gene expression across scRNA-seq data. a) A linear mixed model was fitted for each gene, and variance in expression was partitioned into the following components: 'donor', 'condition' (stimulation and time point), 'log10(total counts)', and a residual noise component. b) For each gene, the variance explained by 'donor' (x-axis) and 'condition' (y-axis) is shown. The red line indicates a 5% threshold for variance explained by condition, used to define a set of 'response' genes.

5.3 | eQTL analysis on bulk RNA-seq data

In order to identify the effect of common variants on the innate immune response, differences in the expression of response genes were examined using an eQTL approach, described further below. Gene sets were defined for the IFN-β response and poly(I:C) response independently.

To identify genes with a change in expression in response to the two stimuli, the generalised linear model quasi-likelihood F test (glmQLF) in the edgeR package [90] was applied to the bulk RNA-seq data generated in parallel to the scRNA-seq described above. The test was conducted in a pairwise manner between each stimulation condition (for example the IFN-β 2 hour time point) and the unstimulated sample, and genes were labelled as 'response genes' using an FDR threshold of 0.05. The union for the two time points per stimulus were taken, yielding an overall set of 'IFN-β response' and 'poly(I:C) response' genes. These were supplemented with genes determined as having a high condition-dependent variance in the single cell data: the 362 genes for which the 'condition' component explained more than 5% of variance in expression.

A consistent eQTL mapping strategy was applied to bulk RNA-seq expression and expression traits derived from scRNA-seq. We considered common variants (minor allele frequency $> 5\%$) within a cis-region spanning 100kb up- and downstream of the gene body for cis QTL analysis. Association tests were performed using a linear mixed model (LMM), accounting for population structure and sample repeat structure as random effects (using a kinship matrix estimated using PLINK [198]). All models were fitted using LIMIX [199]. The significance was tested using a likelihood ratio test (LRT). To adjust for global differences in expression across samples, we included the first 10 principal components, calculated on the 500 mostly highly variable genes,

as covariates. To control for multiple testing, we then applied Benjamini-Hochberg correction [200].

The results of eQTL testing on the bulk RNA-seq data for each condition are shown in Figure 5.2: panel a shows testing of the set of IFN-β response genes, while panel b shows the equivalent for poly(I:C) response genes. The values refer to the number of significant genes using a multiple testing-corrected p-value threshold of 0.1, with the lower part of each panel showing the condition (or overlap of conditions) in which this set was significant. The total number of significant hits in each condition is shown in the bottom left.

For both the IFN-β and poly(I:C) response, the eQTL effects identified are largely context specific, with low overlap seen between the different conditions. Unfortunately the poly(I:C) 6 hour timepoint had a lower number of samples, reducing the power to detect QTL genes. However, the remaining conditions show detection of many response genes. Within these sets, several identified genes are within the list of known innate immune genes (IIGs) described in the previous chapter. These are listed in Table 5.2.

The identification of IIGs with a genetically-determined variation in the unstimulated state raises the intriguing possibility of differences between individuals in their ability to respond based upon expression prior to infection. This is highlighted by QTLs in DDX1 and UNC93B1, both of which are involved in sensing the presence of viral dsRNA. In the case of DDX1, this is in a complex with DDX21, DDx36 and TRIF [201], while UNC93B1 is involved in direct interaction with TLRs [202]. The expression differences across genotypes and conditions is shown for DDX1 as an example (Figure 5.3a). The presence of eQTLs in DDX1 has been found elsewhere; significant results can be seen in multiple tissues of the GTEx resource [203] (Figure 5.3b), including the SNP shown in panel a. Interestingly, this eQTL is most significant in transformed fibroblast cells, although significant effects can be seen in other tissues.

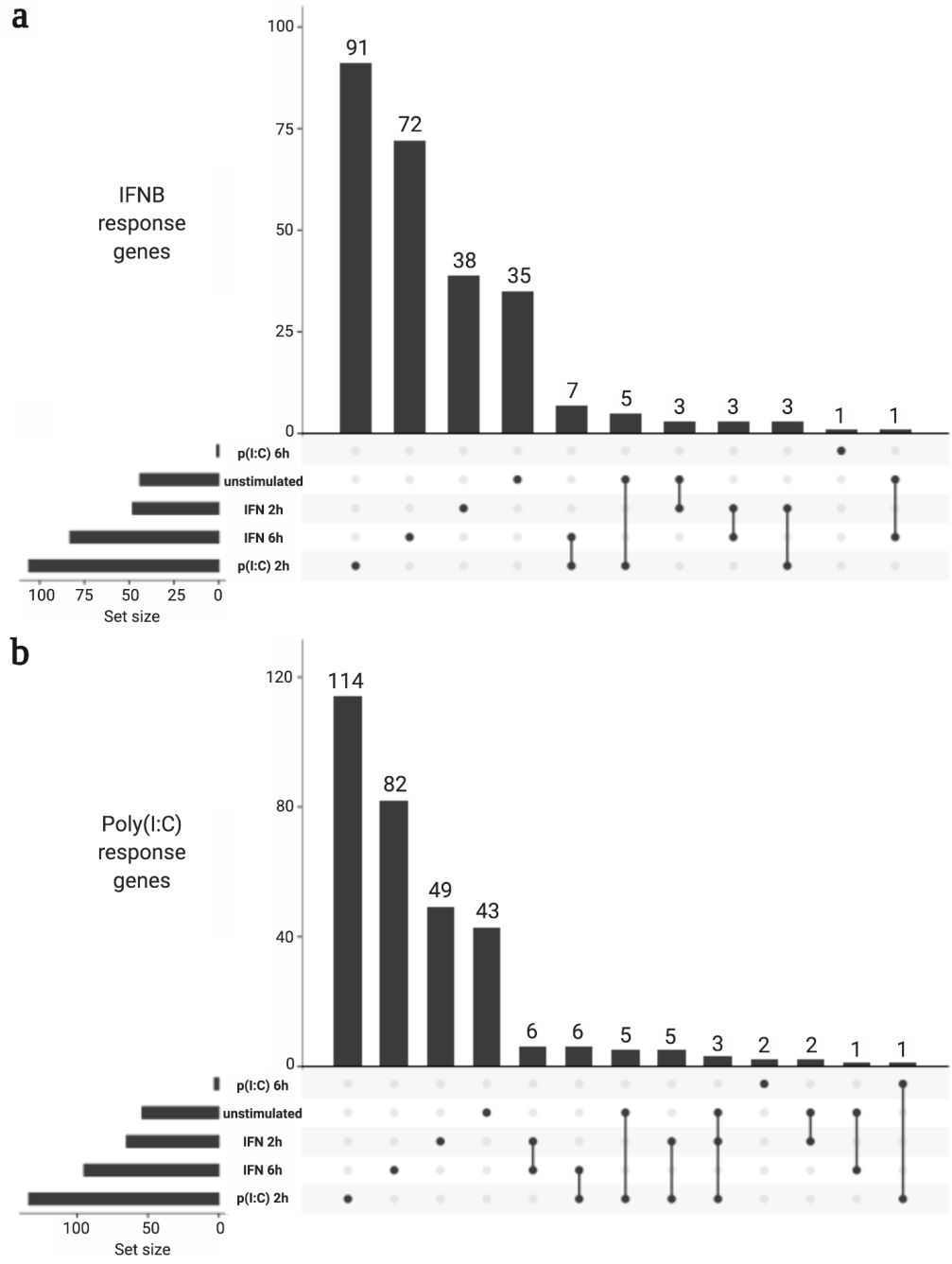


Fig. 5.2 Overlap of eQTLs across stimulation conditions in bulk RNA-seq data, for a) IFN-β response genes, and b) poly(I:C) response genes. Values refer to the number of significant genes (multiple testing-corrected p-value < 0.1). The total number of significant hits in each condition is shown in the bottom left.

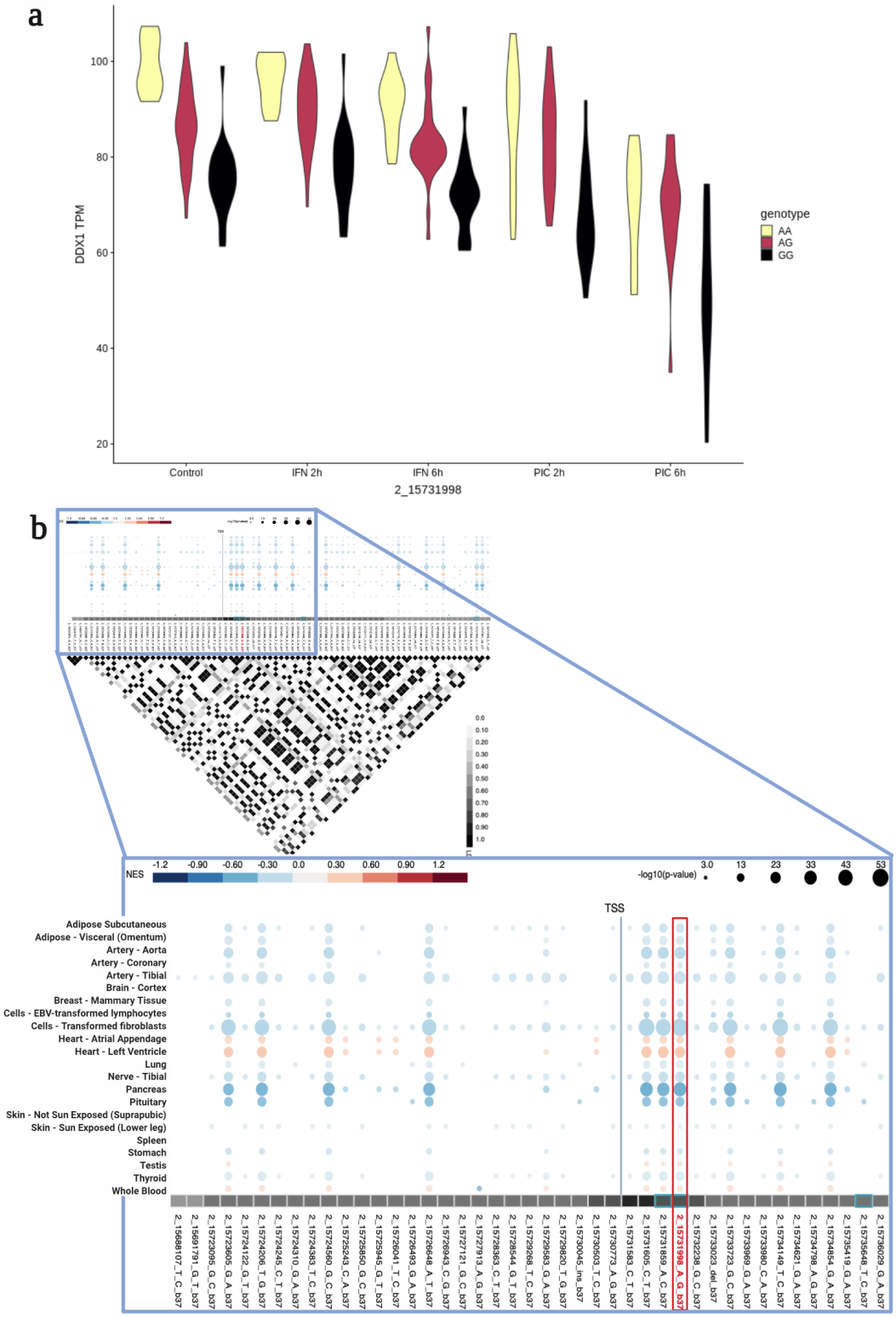


Fig. 5.3 Expression of DDX1 varies with genotype and conditions. a) Expression level (TPM) in bulk RNA-seq data of the DDX1 gene, grouped by stimulation condition and coloured by genotype. b) Presence of eQTLs in the DDX1 gene in the GTEx project. Upper panel shows a zoomed out view of the LD structure, while the lower panel shows the region around the TSS. The SNP identified in panel a is highlighted in red.

Looking at the remaining list of IIGs, several genes known to play a genetically-causative role in disease are identified. For example, mutations in TREX1 have been shown to play a role in both systemic lupus erythematosus and Aicardi–Goutieres syndrome [204–206]. As described in Chapter 1, heterozygous null mutations in IRF7 have been shown to lead to life-threatening influenza infection, and alter type I interferon signalling capacity of dermal fibroblasts [124]. The observation of variability in expression of these genes in healthy individuals could underpin differences in the response to infections within the phenotypically normal human population.

Table 5.2 Significant eQTL hits from bulk RNA-seq - known IIGs.

Condition	Innate Immune Genes
IFN-β response genes	
Unstimulated	*AMACR DDX1 UNC93B1*
IFN-β 2h	*DNAJA3 TRIM69 TREX1 PRKAR2A UBA7*
IFN-β 6h	*TREX1 BTN3A2 AMACR IRF7 TRIM69 TRIM4 APOBEC3F CALCOCO2 DUSP7 CCL2*
Poly(I:C) 2h	*AMACR PLEC LGALS9 IFIT5 BTN3A2 FES CTSS PRDX1 DDX1 IRAK1BP1 OAS3 CASP7 DUSP7*
Poly(I:C) 6h	-
Poly(I:C) response genes	
Unstimulated	*AMACR DDX1 UNC93B1*
IFN-β 2h	*TRIM69 TREX1 PRKAR2A UBA7 PRKAR2A*
IFN-β 6h	*TREX1 BTN3A2 AMACR IRF7 TRIM69 TRIM4 CASP12 APOBEC3F CALCOCO2 DUSP7 CCL2*
Poly(I:C) 2h	*ABCF1 AMACR PLEC LGALS9 IFIT5 BTN3A2 ULBP3 CTSS PRDX1 DDX1 IRAK1BP1 OAS3 CASP7 ACE*
Poly(I:C) 6h	-

5.4 QTL analysis on single cell phenotypes

5.4.1 Mean expression

The results of eQTL testing on 'pseudobulk' expression values for each condition are shown in Figure 5.4, with panel a showing IFN-β response genes, and poly(I:C) response genes in panel b, as before. While the overall number of genes identified is lower than bulk RNA-seq data, likely due to a slightly smaller sample size and increased noise within the dataset, it is still possible to detect significant QTLs at a multiple-testing corrected p-value threshold of 0.1. Once again, these effects are highly context specific.

Considering the innate immune genes within these sets (Table 5.3), several of the previously identified genes from bulk eQTL analysis appear, such as DDX1, IFIT5, OAS3 and BTN3A2. However, novel genes are identified through this analysis, such as ZC3HAV1, TRIM23 and TRIM25, highlighting the potential of scRNA-seq as an orthogonal data type in eQTL discovery. An example is shown for TRIM25 in Figure 5.5, in which expression in single cell (panel a) versus bulk (panel b) data is shown. The expression level in bulk RNA-seq is low (values between 1-3 TPM), which may be the cause of lack of ability to detect a significant effect in this dataset.

Table 5.3 Significant eQTL hits from scRNA-seq 'pseudobulk' values - known IIGs. The genes detected are all classified as both IFN-β and Poly(I:C) response genes.

Condition	Innate Immune Genes
IFN-β and poly(I:C) response genes	
Unstimulated	*TRIM5 ZC3HAV1 DDX1 TRIM23 IFIT5*
IFN-β 2h	*TRIM5 ZC3HAV1*
IFN-β 6h	*BTN3A2 OAS3 ZC3HAV1 TRIM25*
Poly(I:C) 2h	*TRIM5*
Poly(I:C) 6h	*TRIM5 ZC3HAV1*

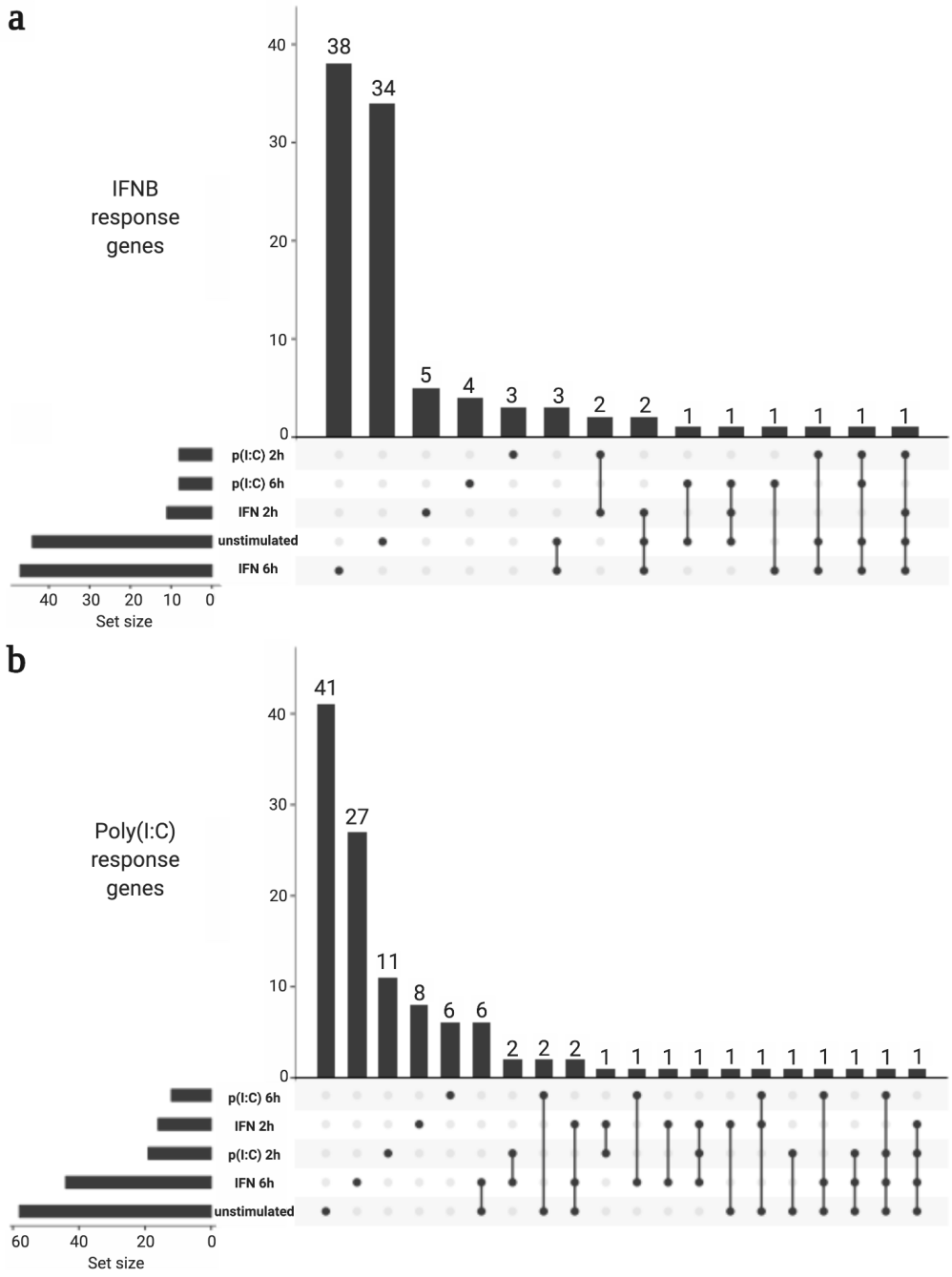


Fig. 5.4 Overlap of eQTLs across stimulation conditions in scRNA-seq derived 'pseudobulk' data, for a) IFN-β response genes, and b) poly(I:C) response genes. Values refer to the number of significant genes (multiple testing-corrected p-value < 0.1). The total number of significant hits in each condition is shown in the bottom left.

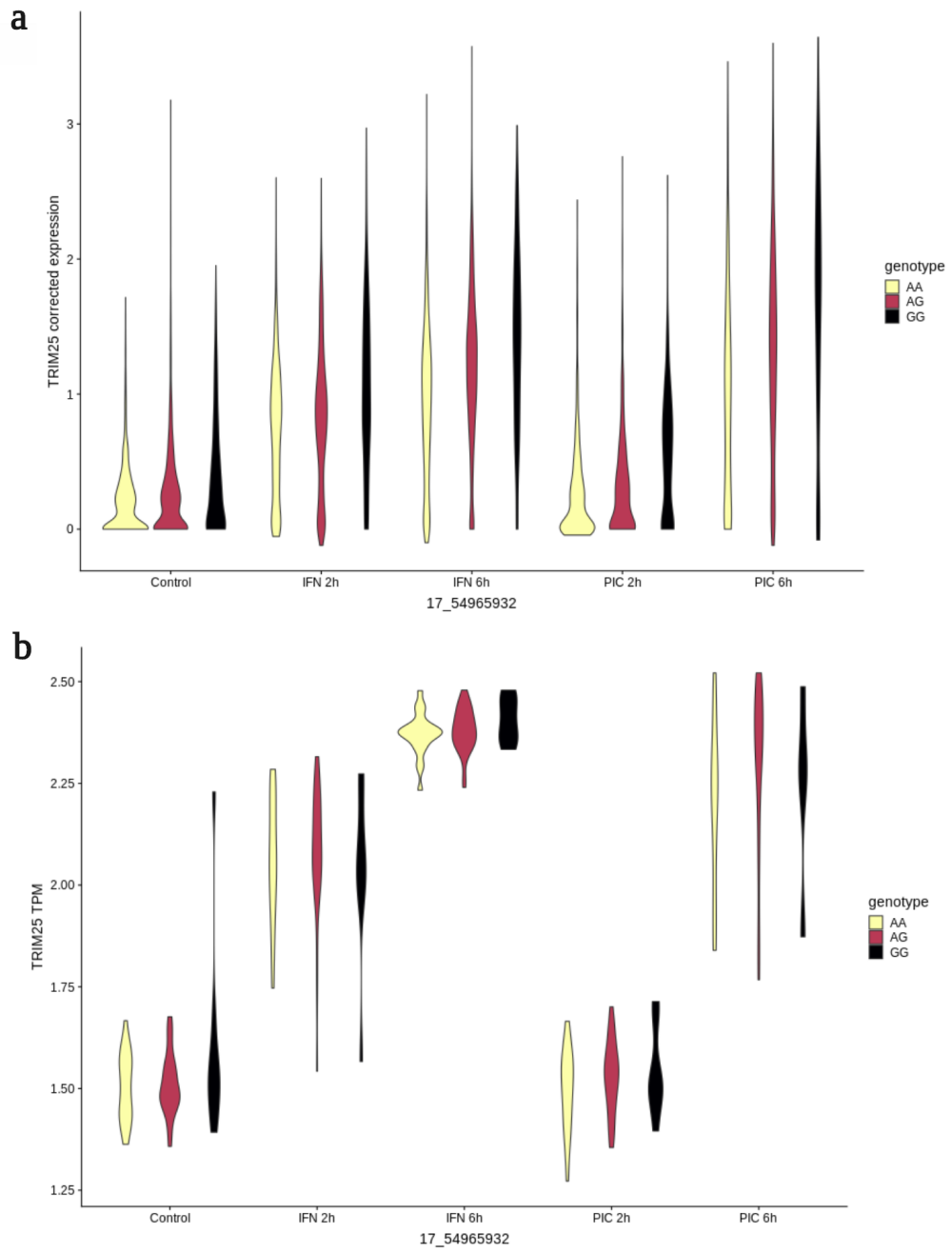


Fig. 5.5 Detection of a TRIM25 eQTL in scRNA-seq data. a) Expression level (Seurat-corrected value) in scRNA-seq data of the TRIM25 gene, grouped by stimulation condition and coloured by genotype. b) Expression level (TPM) in bulk RNA-seq data of the TRIM25 gene, grouped by stimulation condition and coloured by genotype.

5.4.2 Other response phenotypes

Using the IFN-β and poly(I:C) response pseudotimes described in Chapter 4, the average position of cells for each donor was calculated, for each pseudotime independently. As the pseudotime was inferred based upon all donors, the pseudotime average for each donor should reflect the speed of response relative to other donors. Furthermore, the parameters reflecting dynamics of gene expression (t0, mu and k) were inferred for each donor across the two pseudotimes using the SwitchDE package [97]. The variance of gene expression, along with the 'cell proportion' (i.e. the number of cells expression each gene), across the two responses was also calculated.

In preliminary attempts at using these scRNA-seq derived phenotypes, only the variance of genes and proportion of expressing cells across the IFN-β and poly(I:C) response pseudotimes showed significant QTL genes. While this did not result in many innate immune genes being identified, two novel hits were TECPR1, which had a significant cell proportion QTL in both the IFN-β and poly(I:C) response, and SMARCE1, which has a significant cell proportion QTL in the poly(I:C) response. Furthermore, genes identified above - ZC3HAV1 and TRIM5 - were also detected as cell proportion QTLs. These results show the potential of scRNA-seq to identify variation in the proportion of cells expressing innate immune genes between individuals. ZC3HAV1 is shown as an example in Figure 5.6. For this gene, no bulk eQTL was detected (Figure 5.6a), however there appears to be a shift in the distribution of cells expressing the gene (and also a shift in expression level) between the genotypes.

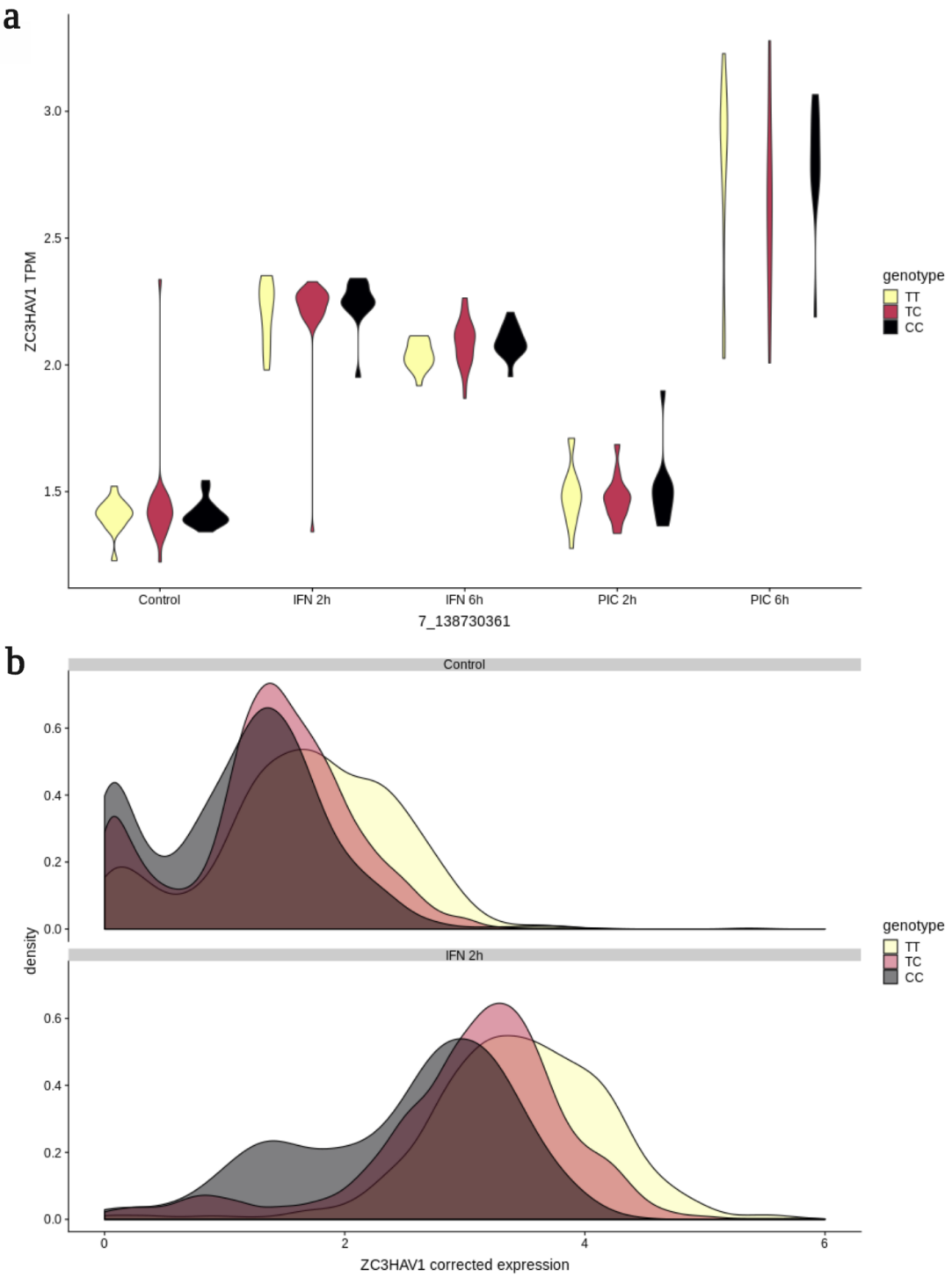


Fig. 5.6 Detection of a ZC3HAV1 cell proportion QTL in scRNA-seq data. a) Expression level (TPM) in bulk RNA-seq data of the ZC3HAV1 gene, grouped by stimulation condition and coloured by genotype. b) Expression level (Seurat-corrected value) in SCRNA-seq data of the ZC3HAV1 gene coloured by genotype. Distribution of expression is shown for the unstimulated cells (upper panel) and cells after 2 hours of IFN-β treatment (lower panel).

5.5 Characterisation of QTL innate immune genes

Taking a combination of genes identified across QTL approaches yields a total set of 391 genes. While the majority were identified through bulk eQTL analysis (Figure 5.7a), use of the single cell data identified 89 additional genes.

In order to characterise these further, enrichment for particular functional categories was investigated for genes within the known IIG set (Figure 5.7b). Each functional class was compared against the background number in the entire scRNA-seq dataset. Sensors were found to be significantly enriched (hypergeometric test, pvalue = 0.021), which could suggest an interesting source of variability in response to infection through differences in detection of pathogens.

Having identified many response genes with a genetic basis for variation between individuals, it is interesting to consider whether these genes show co-regulated expression at a single cell level and across pseudotime. To this end, the expression of all genes with significant QTLs identified was plotted against the IFN-β and poly(I:C) response pseudotimes defined in Chapter 4 (Figures 5.7c-d). From this analysis, it appears that there are modules of co-expressed genes, particularly in response to poly(I:C) treatment, however further work will be needed to elucidate whether there is a genetic mechanism underpinning co-regulation of these genes.

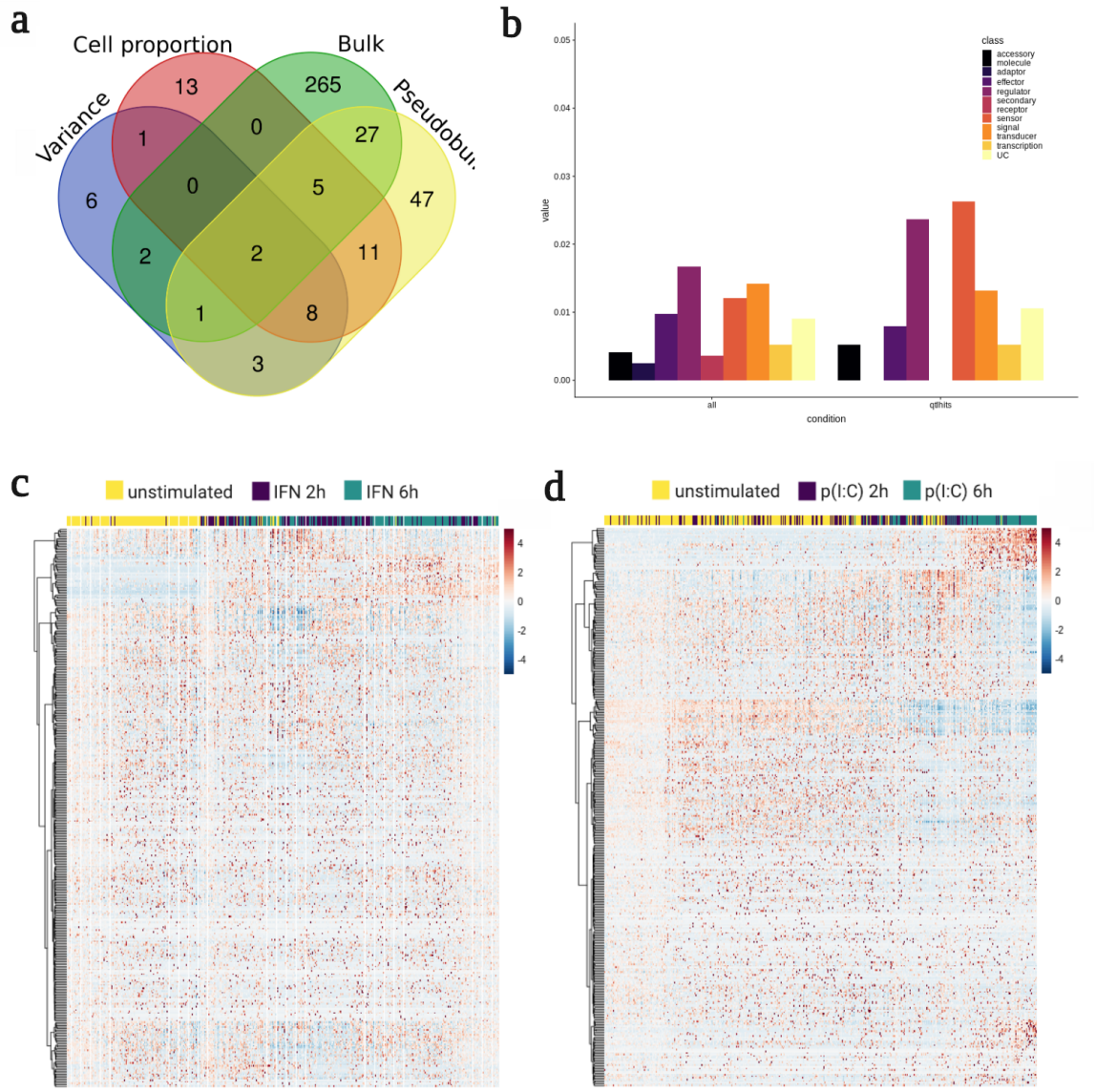


Fig. 5.7 Investigation of significant genes across QTL approaches. a) Overlap in genes identified from bulk RNA-seq data and scRNA-seq (pseudobulk, variance and cell proportion QTLs). b) Distribution of IIGs across functional categories - background on the left, vs significant QTL genes on the right. c-d) Expression of QTL genes across the IFN-β and poly(I:C) response pseudotimes respectively.

5.6 Discussion

Variability in the response to viruses has long been studied, predominantly through investigation of individuals with susceptibility to particular infections, or with a genetic disorder in innate immune genes. In this work, I showed variability in this response across healthy individuals, first using a variance partitioning approach to highlight genes whose expression was explained by condition or donor.

Applying QTL approaches to bulk RNA-seq data revealed hundreds of genes with a genetic basis for inter-individual variation. Several of these have been previously implicated in disease, such as TREX1 [204–206] and IRF7 [124], validating the detection of biologically interesting hits. However, the identification of other innate immune genes in donors with a normal phenotype highlights the potential to understand variability in the response within the population as a whole. Furthermore, characterisation of QTL genes which are not annotated as known innate immune genes may yield novel insight into the type I interferon response.

Using scRNA-seq data, it was possible to expand the set of QTL genes, primarily through calculation of a 'pseudobulk' expression metric. In the case of phenotypes reflecting differences in temporal dynamics, such as average position in response pseudotime, or SwitchDE parameters, it is likely that the number of cells and donors in the current study is not large enough given the amount of noise within the data, and hence the difficulty in robustly inferring dynamic parameters. For such insights, it will be necessary to further develop the phenotypes used in such approaches. Furthermore, an increase in sample size will be required to improve power in single cell-based QTL studies. In a recent computational analysis, Sarker *et al.* estimated the sample size that would be required to detect dispersion QTLs in scRNA seq data (QTLs that affect the variability but not mean expression level) [207]. They showed that 4,015

individuals would be the lower bound to achieve 80% power to detect the strongest dQTLs in iPSCs. While this number will be lower for phenotypes that also affect mean expression level, an increase in number of individuals profiled will broaden the range of molecular phenotypes interrogatable with QTL approaches.

Moving forward, further analysis will shed light on the nature of the QTL hits identified. Through combination with ChIP-seq and ATAC-seq data, it will be possible to overlap identified genomic loci with regulatory regions. This will shed light on the mechanism of regulation, for example through transcription factor binding sites or enhancer regions. This will also allow detection of regions that may be 'primed' in an unstimulated state, as shown previously in human macrophages [208].

Chapter 6

Concluding remarks

In this PhD, I optimised and conducted large-scale single-cell RNA sequencing experiments to study cellular variation in the type I interferon response in fibroblasts of 70 healthy human individuals. Using this dataset, I first studied heterogeneity in the unstimulated state, comparing to *ex vivo* skin data to confirm the relative homogeneity of the *in vitro* cultured fibroblasts used. Using matched whole exome sequencing data, somatic mutations in sub-populations of cells within each donor were detected, and clonal populations identified. Applying cardelino to 32 of the HipSci fibroblast lines identified hundreds of differentially expressed genes between cells from different somatic clones, with cell cycle and proliferation pathways frequently enriched.

Returning to innate immunity, I performed analyses into the variability in the innate immune response across mammalian species, showing a link with evolutionary divergence. Within the human dataset I generated, I characterised the innate immune response at single cell resolution, elucidating the dynamics of the response across donors and defining discrete gene modules. Harnessing the scRNA-seq data, I defined several phenotypes to capture variability in this response. Applying quantitative trait loci approaches to study the genetic basis of this heterogeneity in innate immunity, I

identified 391 response genes with a QTL from either bulk, pseudobulk, or single-cell expression traits.

Moving into the future, experimental work will be required for functional validation of genetic variants altering the innate immune response. One approach is the use of knock down or knock out experiments (for example, through transfection with siRNAs) prior to innate immune stimulation with poly(I:C)/IFN-β. This would allow elucidation of the role of individual genes in the type I interferon response, and could be applied to genes identified through temporal analysis (Chapter 4) or genetic analysis (Chapter 5) in order to investigate regulation within the system. For the validation of specific genetic variants, such as those described in Chapter 5, a CRISPR approach could be used. With this, cell lines could be engineered to contain the alternative genotype at the specific site of interest, prior to monitoring the effect on response. Where variants are suspected to affect binding of transcription factors, this could be confirmed by ChIP-seq experiments. Furthermore, stimulation experiments may be extended to understand how genetic variants relate to susceptibility phenotypes in particular individuals. *In vitro* infection with specific viruses, rather than poly(I:C) and IFN-β, could allow a more focused look at the role of variation in infectious diseases.

As described in Chapter 3, fibroblasts form just one element of the skin millieu. To place this component of the innate immune response within the tissue environment, transcription can be measured spatially. Single-molecule fluorescence in situ hybridisation (smFISH) [209] provides a method to detect individual mRNA molecules of tissue sections, but is limited in the number of transcripts assayable. Recent developments in multiplexing, for example multiplex error-robust FISH (MERFISH) [210] and sequential FISH (seqFISH) [211] have addressed this bottleneck. Applying these methods to innate immune stimulation in the skin would deepen our understanding of the spatial

nature of type I interferon signalling, which may shed light on the heterogeneity in this response.

While this work has focused on variability in transcription, technologies to profile single cells at different molecular levels have vastly evolved over recent years. For example, there are several techniques to capture epigenetic regulation, such as single cell reduced representation bisulfite sequencing (scRRBS) [212, 213] and single-cell methylome and transcriptome sequencing (scMT-seq) [214], which profile DNA methylation. Single cell proteomic assays are developing rapidly, however they are still limited in the number of proteins that can be studied within an experiment. Proteomic methods, such as fluorescence-activated cell sorting (FACS) and cytometry by time of flight (CyTOF) do allow higher throughput of cells than sequencing-based technologies. To gain a more complete picture of the heterogeneity in innate immune response, it will be necessary to utilise and integrate these assays.

This work has highlighted the role of scRNA-sequencing technology in understanding variability within healthy donors, both in the unstimulated (Chapter 3) and activated (Chapters 4 and 5) states. We are currently at the boundary of throughput for the use of single cell sequencing in population genetics. However, the increasing scale of these technologies will soon allow this approach to become more commonplace, allowing application to many biological processes.

Looking further into the future, it is intriguing to speculate on the ability to use our understanding of variability in the innate immune response in a translational context. We are at a point of technological advance in two directions: an increasing characterisation of genetic variability, with initiative such as the 100,000 Genomes Project, and a rapid increase in the resolution and methodologies with which we can profile individual cells. This will need to be accompanied by development of sophisticated computational methods to handle such large -omics data. However, with

increased data availability, we may be able to link molecular phenotypes to physiological responses, incorporating information such as infection history. This will pave the way for translating an understanding of the molecular basis and impact of variability in innate immune response to personalised therapies.

References

[1] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Mei Lee Hong, Joann Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe,

Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa De La Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G.R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F.A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw Pyng Yang, Ru Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[2] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[3] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[4] Jolene Michelle Helena, Anna Margaretha Joubert, Simone Grobbelaar, Elsie Magdalena Nolte, Marcel Nel, Michael Sean Pepper, Magdalena Coetzee, and Anne Elisabeth Mercier. Deoxyribonucleic acid damage and repair: Capitalizing on our understanding of the mechanisms of maintaining genomic integrity for therapeutic purposes. *International Journal of Molecular Sciences*, 19(4):E1148, 2018.

[5] S T Sherry, M H Ward, M Kholodov, J Baker, L Phan, E M Smigielski, and K Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[6] James F. Gusella, Nancy S. Wexler, P. Michael Conneally, Susan L. Naylor, Mary Anne Anderson, Rudolph E. Tanzi, Paul C. Watkins, Kathleen Ottina, Margaret R. Wallace, Alan Y. Sakaguchi, Anne B. Young, Ira Shoulson, Ernesto

Bonilla, and Joseph B. Martin. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940):234–238, 1983.

[7] L C Tsui, D W Cox, P J McAlpine, and M Buchwald. Cystic fibrosis: analysis of linkage of the disease locus to red cell and plasma protein markers. *Cytogenetics and cell genetics*, 39(3):238–9, 1985.

[8] Robert G. Knowlton, Odile Cohen-Haguenauer, Nguyen Van Cong, Jean Frézal, Valerie A. Brown, David Barker, Jeffrey C. Braman, James W. Schumm, Lap Chee Tsui, Manuel Buchwald, and Helen Donis-Keller. A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature*, 318(6044):380–382, 1985.

[9] Brandon J. Wainwright, Peter J. Scambler, Jorg Schmidtke, Eila A. Watson, Hai Yang Law, Martin Farrall, Howard J. Cooke, Hans Eiberg, and Robert Williamson. Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature*, 318(6044):384–385, 1985.

[10] R White, S Woodward, M Leppert, P O'Connell, M Hoff, J Herbst, J M Lalouel, M Dean, and G Vande Woude. A closely linked genetic marker for cystic fibrosis. *Nature*, 318(6044):382–4, 1985.

[11] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, Michael B Bracken, Frederick L Ferris, Jurg Ott, Colin Barnstable, and Josephine Hoh. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720):385–389, 2005.

[12] WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[13] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, 2007.

[14] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[15] Harm Jan Westra and Lude Franke. From genome to function by studying eQTLs. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1842(10):1896–1902, 2014.

[16] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell

transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–40, 2013.

[17] Georgi K. Marinov, Brian A. Williams, Ken McCue, Gary P. Schroth, Jason Gertz, Richard M. Myers, and Barbara J. Wold. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, 2014.

[18] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi An Ma, and Cole Trapnell. Single-cell mRNA quantification and differential analysis with Census. *Nature Methods*, 14(3):309–315, 2017.

[19] Joshua D. Welch, Yin Hu, and Jan F. Prins. Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, 44(8):e73, 2016.

[20] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, 2013.

[21] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[22] Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Tomislav Illicic, Sarah A. Teichmann, and John C. Marioni. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, 6:8687, 2015.

[23] Björn Reinius, Jeff E. Mold, Daniel Ramsköld, Qiaolin Deng, Per Johnsson, Jakob Michaëlsson, Jonas Frisén, and Rickard Sandberg. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nature Genetics*, 48(11):1430–1435, 2016.

[24] Jong Kyoung Kim and John C. Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14(1):R7, 2013.

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

[26] Sean C. Bendall, Kara L. Davis, El Ad David Amir, Michelle D. Tadmor, Erin F. Simonds, Tiffany J. Chen, Daniel K. Shenfeld, Garry P. Nolan, and Dana Pe'Er. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.

[27] Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016.

[28] Tapio Lönnberg, Valentine Svensson, Kylie R. James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan S.F. Soon, Lily G. Fogg, Arya Sheela Nair, Urijah N. Liligeto, Michael J.T. Stubbington, Lam Ha Ly, Frederik Otzen Bagger, Max Zwiessele, Neil D. Lawrence, Fernando Souza-Fonseca-Guimaraes, Patrick T. Bunn, Christian R. Engwerda, William R. Heath, Oliver Billker, Oliver Stegle, Ashraful Haque, and Sarah A. Teichmann. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Science Immunology*, 3(21):eaat1469, 2017.

[29] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander Van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

[30] Amit Zeisel, Ana B. Moz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–42, 2015.

[31] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, 2015.

[32] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[33] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, 2014.

[34] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell*, 65(4):631–643, 2017.

[35] Valentine Svensson, Kedar Nath Natarajan, Lam Ha Ly, Ricardo J. Miragaia, Charlotte Labalette, Iain C. Macaulay, Ana Cvejic, and Sarah A. Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, 14(4):381–387, 2017.

[36] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

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

[38] Nils Eling, Michael D. Morgan, and John C. Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, 2019.

[39] Michael I Love, John B Hogenesch, and Rafael A Irizarry. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology*, 34(12):1287–1291, 2016.

[40] D C Jones, K T Kuppusamy, N J Palpant, X Peng, E Charles, H Ruohola-baker, and W L Ruzzo. Isolator: accurate and stable analysis of isoform-level expression in RNA-Seq experiments. *bioRxiv*, DOI:10.1101/088765, 2016.

[41] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[42] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[43] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.

[44] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 2015.

[45] Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, 2019.

[46] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[47] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017.

[48] Simon Andrews and Babraham Bioinformatics. FastQC: A quality control tool for high throughput sequence data. Online:http://www.bioinformatics.babraham.ac.uk/projects/, 2010.

[49] Matthew P.A. Davis, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J. Enright. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49, 2013.

[50] Davis J. McCarthy, Kieran R. Campbell, Aaron T.L. Lun, and Quin F. Wills. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186, 2017.

[51] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17:29, 2016.

[52] Dominic Grün and Alexander Van Oudenaarden. Design and Analysis of Single-Cell Sequencing Experiments. *Cell*, 163(4):799–810, 2015.

[53] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572, 1901.

[54] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[55] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, 2018.

[56] Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37:38–44, 2019.

[57] Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16:241, 2015.

[58] Stéphane Lafon and Ann B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–403, 2006.

[59] Joe H. Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[60] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015.

[61] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods*, 14(5):483–486, 2017.

[62] Alexander Strehl and Joydeep Ghosh. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[63] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–416, 2017.

[64] Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, 2008.

[65] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

[66] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019.

[67] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

[68] Paul M. Magwene, Paul Lizardi, and Junhyong Kim. Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, 19(7):842–850, 2003.

[69] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, 2017.

[70] Qi Mao, Li Wang, Steve Goodison, and Yijun Sun. Dimensionality Reduction Via Graph Structure Learning. In *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 765–774, 2015.

[71] Iain C. Macaulay, Valentine Svensson, Charlotte Labalette, Lauren Ferreira, Fiona Hamey, Thierry Voet, Sarah A. Teichmann, and Ana Cvejic. Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell Reports*, 14(4):966–977, 2016.

[72] Kieran R. Campbell and Christopher Yau. Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. *PLoS Computational Biology*, 12(11):e1005212, 2016.

[73] Kieran Campbell and Christopher Yau. Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. *bioRxiv*, DOI:10.1101/060442, 2016.

[74] Manu Setty, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology*, 34(6):637–645, 2016.

[75] Eugenio Marco, Robert L. Karp, Guoji Guo, Paul Robson, Adam H. Hart, Lorenzo Trippa, and Guo-Cheng Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 111(52):E5643–50, 2014.

[76] F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, 2019.

[77] E. H. Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[78] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[79] Monica Benito, Joel Parker, Quan Du, Junyuan Wu, Dong Xiang, Charles M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114, 2004.

[80] Davide Risso, John Ngai, Terence P. Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902, 2014.

[81] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[82] Jeffrey T. Leek. Svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21):e161, 2014.

[83] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

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

[85] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–410, 2018.

[86] Laleh Haghverdi, Aaron T.L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[87] Jong-Eun Park, Krzysztof Polański, Kerstin Meyer, and Sarah A. Teichmann. Fast Batch Alignment of Single Cell Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape. *bioRxiv*, DOI:10.1101/397042, 2018.

[88] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology*, 11(6):e1004333, 2015.

[89] M. I. Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data - the DESeq2 package. *DESeq2 manual*, 2016.

[90] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[91] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16:278, 2015.

[92] Tallulah S. Andrews and Martin Hemberg. Modelling dropouts allows for unbiased identification of marker genes in scRNASeq experiments. *bioRxiv*, DOI:10.1101/065094, 2016.

[93] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[94] Minzhe Guo, Hui Wang, S. Steven Potter, Jeffrey A. Whitsett, and Yan Xu. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Computational Biology*, 11(11):e1004575, 2015.

[95] Keegan D. Korthauer, Li Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, 2016.

[96] Alfredo A. Kalaitzis and Neil D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12:180, 2011.

[97] Kieran R. Campbell and Christopher Yau. Switchde: Inference of switch-like differential expression along single-cell trajectories. *Bioinformatics*, 33(8):1241–1242, 2017.

[98] Jil Sander, Joachim L. Schultze, and Nir Yosef. ImpulseDE: Detection of differentially expressed genes in time series data using impulse models. *Bioinformatics*, 33(5):757–759, 2017.

[99] Bidesh Mahata, Xiuwei Zhang, Aleksandra A. Kolodziejczyk, Valentina Proserpio, Liora Haim-Vilmovsky, Angela E. Taylor, Daniel Hebenstreit, Felix A. Dingler, Victoria Moignard, Berthold Göttgens, Wiebke Arlt, Andrew N.J. McKenzie, and Sarah A. Teichmann. Single-cell RNA sequencing reveals T helper cells synthesizing steroids De Novo to contribute to immune homeostasis. *Cell Reports*, 7(4):1130–42, 2014.

[100] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[101] Zhigang Xue, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun Yan Jiang, Yun Feng, Zhenshan Liu, Qiao Zeng, Liming Cheng, Yi E. Sun, Jia Yin Liu, Steve Horvath, and Guoping Fan. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597, 2013.

[102] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean Christophe Marine, Pierre Geurts, Jan Aerts, Joost Van Den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, 2017.

[103] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, 2010.

[104] Surya Pandey, Taro Kawai, and Shizuo Akira. Microbial sensing by toll-like receptors and intracellular nucleic acid sensors. *Cold Spring Harbor Perspectives in Biology*, 7(1):a016246, 2015.

[105] Hiroki Kato, Osamu Takeuchi, Eriko Mikamo-Satoh, Reiko Hirai, Tomoji Kawai, Kazufumi Matsushita, Akane Hiiragi, Terence S. Dermody, Takashi Fujita, and Shizuo Akira. Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid-inducible gene-I and melanoma differentiation-associated gene 5. *Journal of Experimental Medicine*, 205(7):1601–10, 2008.

[106] Lijun Sun, Jiaxi Wu, Fenghe Du, Xiang Chen, and Zhijian J. Chen. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science*, 339(6121):786–791, 2013.

[107] M.O. Diaz, S. Bohlander, and G. Allen. Nomenclature of the Human Interferon Genes a. *Journal of Interferon & Cytokine Research*, 16(2):179–180, 1996.

[108] M. Cristina Gauzzi, Laura Velazquez, Roslyn McKendry, Knud E. Mogensen, Marc Fellous, and Sandra Pellegrini. Interferon-α-dependent activation of Tyk2 requires phosphorylation of positive regulatory tyrosines by another kinase. *Journal of Biological Chemistry*, 271(34):20494–500, 1996.

[109] Xiaoxia Li, Stewart Leung, Sajjad Qureshi, James E. Darnell, and George R. Stark. Formation of STAT1-STAT2 heterodimers and their role in the activation of IRF-1 gene transcription by interferon-α. *Journal of Biological Chemistry*, 271(10):5790–4, 1996.

[110] Hans Heinrich Hoffmann, William M. Schneider, and Charles M. Rice. Interferons and viruses: An evolutionary arms race of molecular interactions. *Trends in Immunology*, 36(3):124–138, 2015.

[111] Marco Colonna, Giorgio Trinchieri, and Yong-Jun Liu. Plasmacytoid dendritic cells in immunity. *Nature immunology*, 5(12):1219–1226, 2004.

[112] Lionel B. Ivashkiv and Laura T. Donlin. Regulation of type I interferon responses. *Nature Reviews Immunology*, 14(1):36–49, 2014.

[113] Benjamin B. Kaufmann and Alexander van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Current Opinion in Genetics and Development*, 17(2):107–112, 2007.

[114] Georg A. Holländer. On the stochastic regulation of interleukin-2 transcription. *Seminars in Immunology*, 11(5):357–367, 1999.

[115] D. P. Calado, T. Paixao, D. Holmberg, and M. Haury. Stochastic Monoallelic Expression of IL-10 in T Cells. *The Journal of Immunology*, 177(8):5358–5364, 2006.

[116] Mingwei Zhao, Jiangwen Zhang, Hemali Phatnani, Stefanie Scheu, and Tom Maniatis. Stochastic expression of the interferon-β gene. *PLoS Biology*, 10(1):e1001249, 2012.

[117] T. Ravasi, C. Wells, A. Forest, D. M. Underhill, B. J. Wainwright, A. Aderem, S. Grimmond, and D. A. Hume. Generation of Diversity in the Innate Immune System: Macrophage Heterogeneity Arises from Gene-Autonomous Transcriptional Probability of Individual Inducible Genes. *The Journal of Immunology*, 168(1):44–50, 2002.

[118] S. Ramsey, A. Ozinsky, A. Clark, K. D. Smith, P. De Atauri, V. Thorsson, D. Orrell, and H. Bolouri. Transcriptional noise and cellular heterogeneity in mammalian macrophages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1467):495–506, 2006.

[119] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, Laura Jardine, David Dixon, Emily Stephenson, Emil Nilsson, Ida Grundberg, David McDonald, Andrew Filby, Weibo Li, Philip L. De Jager, Orit Rozenblatt-Rosen, Andrew A. Lane, Muzlifah Haniffa, Aviv Regev, and Nir Hacohen. Single-cell RNA-Seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Transplantation*, 356(6335):eaah4573, 2017.

[120] T. M. Lin, C. J. Chen, M. M. Wu, C. S. Yang, J. S. Chen, C. C. Lin, T. Y. Kwang, S. T. Hsu, S. Y. Lin, and L. C. Hsu. Hepatitis B virus marker in Chinese twins. *Anticancer Research*, 9(3):737–741, 1989.

[121] Andri Rauch, Zoltán Kutalik, Patrick Descombes, Tao Cai, Julia Di Iulio, Tobias Mueller, Murielle Bochud, Manuel Battegay, Enos Bernasconi, Jan Borovicka, Sara Colombo, Andreas Cerny, Jean François Dufour, Hansjakob Furrer, Huldrych F. Günthard, Markus Heim, Bernard Hirschel, Raffaele Malinverni, Darius Moradpour, Beat Müllhaupt, Andrea Witteck, Jacques S. Beckmann, Thomas Berg, Sven Bergmann, Francesco Negro, Amalio Telenti, and Pierre Yves Bochud. Genetic Variation in IL28B Is Associated With Chronic Hepatitis C and Treatment Failure: A Genome-Wide Association Study. *Gastroenterology*, 138(4):1338–1345, 2010.

[122] Aaron R. Everitt, Simon Clare, Thomas Pertel, Sinu P. John, Rachael S. Wash, Sarah E. Smith, Christopher R. Chin, Eric M. Feeley, Jennifer S. Sims, David J. Adams, Helen M. Wise, Leanne Kane, David Goulding, Paul Digard, Verneri Anttila, J. Kenneth Baillie, Tim S. Walsh, David A. Hume, Aarno Palotie, Yali

Xue, Vincenza Colonna, Chris Tyler-Smith, Jake Dunning, Stephen B. Gordon, Rosalind L. Smyth, Peter J. Openshaw, Gordon Dougan, Abraham L. Brass, and Paul Kellam. IFITM3 restricts the morbidity and mortality associated with influenza. *Nature*, 484(7395):519–523, 2012.

[123] Shen Ying Zhang, Emmanuelle Jouanguy, Sophie Ugolini, Asma Smahi, Gaëlle Elain, Pedro Romero, David Segal, Vanessa Sancho-Shimizu, Lazaro Lorenzo, Anne Puel, Capucine Picard, Ariane Chapgier, Sabine Plancoulaine, Matthias Titeux, Céline Cognet, Horst Von Bernuth, Cheng Lung Ku, Armanda Casrouge, Xin Xin Zhang, Luis Barreiro, Joshua Leonard, Claire Hamilton, Pierre Lebon, Bénédicte Héron, Louis Vallée, Lluis Quintana-Murci, Alain Hovnanian, Flore Rozenberg, Eric Vivier, Frédéric Geissmann, Marc Tardieu, Laurent Abel, and Jean Laurent Casanova. TLR3 deficiency in patients with herpes simplex encephalitis. *Science*, 317(5844):1522–7, 2007.

[124] Michael J. Ciancanelli, Sarah X.L. Huang, Priya Luthra, Hannah Garner, Yuval Itan, Stefano Volpi, Fabien G. Lafaille, Céline Trouillet, Mirco Schmolke, Randy A. Albrecht, Elisabeth Israelsson, Hye Kyung Lim, Melina Casadio, Tamar Hermesh, Lazaro Lorenzo, Lawrence W. Leung, Vincent Pedergnana, Bertrand Boisson, Satoshi Okada, Capucine Picard, Benedicte Ringuier, Françoise Troussier, Damien Chaussabel, Laurent Abel, Isabelle Pellier, Luigi D. Notarangelo, Adolfo García-Sastre, Christopher F. Basler, Frédéric Geissmann, Shen Ying Zhang, Hans Willem Snoeck, and Jean Laurent Casanova. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. *Science*, 348(6233):448–453, 2015.

[125] Erika Della Mina, Alessandro Borghesi, Hao Zhou, Salim Bougarn, Sabri Boughorbel, Laura Israel, Ilaria Meloni, Maya Chrabieh, Yun Ling, Yuval Itan, Alessandra Renieri, Iolanda Mazzucchelli, Sabrina Basso, Piero Pavone, Raffaele Falsaperla, Roberto Ciccone, Rosa Maria Cerbo, Mauro Stronati, Capucine Picard, Orsetta Zuffardi, Laurent Abel, Damien Chaussabel, Nico Marr, Xiaoxia Li, Jean-Laurent Casanova, and Anne Puel. Inherited human IRAK-1 deficiency selectively impairs TLR signaling in fibroblasts. *Proceedings of the National Academy of Sciences*, 114(4):E514–E523, 2017.

[126] Kerry Dobbs, Cecilia Domínguez Conde, Shen-Ying Zhang, Silvia Parolini, Magali Audry, Janet Chou, Emma Haapaniemi, Sevgi Keles, Ivan Bilic, Satoshi Okada, Michel J. Massaad, Samuli Rounioja, Adel M. Alwahadneh, Nina K. Serwas, Kelly Capuder, Ergin Çiftçi, Kerstin Felgentreff, Toshiro K. Ohsumi, Vincent Pedergnana, Bertrand Boisson, Şule Haskoloğlu, Arzu Ensari, Michael Schuster, Alessandro Moretta, Yuval Itan, Ornella Patrizi, Flore Rozenberg, Pierre Lebon, Janna Saarela, Mikael Knip, Slavé Petrovski, David B. Goldstein, Roberta E. Parrott, Berna Savas, Axel Schambach, Giovanna Tabellini, Christoph Bock, Talal A. Chatila, Anne Marie Comeau, Raif S. Geha, Laurent Abel, Rebecca H. Buckley, Aydan İkincioğulları, Waleed Al-Herz, Merja Helminen, Figen Doğu,

Jean-Laurent Casanova, Kaan Boztuğ, and Luigi D. Notarangelo. Inherited DOCK2 Deficiency in Patients with Early-Onset Invasive Infections. *New England Journal of Medicine*, 372(25):2409–2422, 2015.

[127] Sarah Kim, Jessica Becker, Matthias Bechheim, Vera Kaiser, Mahdad Noursadeghi, Nadine Fricker, Esther Beier, Sven Klaschik, Peter Boor, Timo Hess, Andrea Hofmann, Stefan Holdenrieder, Jens R. Wendland, Holger Fröhlich, Gunther Hartmann, Markus M. Nöthen, Bertram Müller-Myhsok, Benno Pütz, Veit Hornung, and Johannes Schumacher. Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. *Nature Communications*, 5:5236, 2014.

[128] Benjamin P Fairfax, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, and Julian C Knight. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science*, 343(6175):1246949, 2014.

[129] Mark N. Lee, Chun Ye, Alexandra Chloé Villani, Towfique Raj, Weibo Li, Thomas M. Eisenhaure, Selina H. Imboywa, Portia I. Chipendo, F. Ann Ran, Kamil Slowikowski, Lucas D. Ward, Khadir Raddassi, Cristin McCabe, Michelle H. Lee, Irene Y. Frohlich, David A. Hafler, Manolis Kellis, Soumya Raychaudhuri, Feng Zhang, Barbara E. Stranger, Christophe O. Benoist, Philip L. De Jager, Aviv Regev, and Nir Hacohen. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175):1246980, 2014.

[130] A. K. Field, A. A. Tytell, G. P. Lampson, and M. R. Hilleman. Inducers of interferon and host resistance. II. Multistranded synthetic polynucleotide complexes. *Proceedings of the National Academy of Sciences*, 58(3):1004–1010, 1967.

[131] A. Billiau, C. E. Buckler, F. Dianzani, C. Uhlendorf, and S. Baron. Induction of the Interferon Mechanism by Single-Stranded RNA: Potentiation by Polybasic Substances. *Experimental Biology and Medicine*, 132(2):790–796, 1969.

[132] M. Firoz Mian, Amna N. Ahmed, Mehrnaz Rad, Artem Babaian, Dawn Bowdish, and Ali A. Ashkar. Length of dsRNA (poly I:C) drives distinct innate immune responses, depending on the cell type. *Journal of Leukocyte Biology*, 94(5):1025–36, 2013.

[133] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-seq. *Bioinformatics*, 27(17):2325–9, 2011.

[134] Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, 2012.

[135] Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences*, 132(3):159–164, 2013.

[136] Magnus D. Lynch and Fiona M. Watt. Fibroblast heterogeneity: implications for human disease. *Journal of Clinical Investigation*, 128(1):26–35, 2018.

[137] F. M. Burnet. Intrinsic mutagenesis: A genetic basis of ageing. *Pathology*, 6(1):1–11, 1974.

[138] Iñigo Martincorena and Peter J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–9, 2015.

[139] Nicolas Stransky, Ann Marie Egloff, Aaron D. Tward, Aleksandar D. Kostic, Kristian Cibulskis, Andrey Sivachenko, Gregory V. Kryukov, Michael S. Lawrence, Carrie Sougnez, Aaron McKenna, Erica Shefler, Alex H. Ramos, Petar Stojanov, Scott L. Carter, Douglas Voet, Maria L. Cortés, Daniel Auclair, Michael F. Berger, Gordon Saksena, Candace Guiducci, Robert C. Onofrio, Melissa Parkin, Marjorie Romkes, Joel L. Weissfeld, Raja R. Seethala, Lin Wang, Claudia Rangel-Escareño, Juan Carlos Fernandez-Lopez, Alfredo Hidalgo-Miranda, Jorge Melendez-Zajgla, Wendy Winckler, Kristin Ardlie, Stacey B. Gabriel, Matthew Meyerson, Eric S. Lander, Gad Getz, Todd R. Golub, Levi A. Garraway, and Jennifer R. Grandis. The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–60, 2011.

[140] Eran Hodis, Ian R. Watson, Gregory V. Kryukov, Stefan T. Arold, Marcin Imielinski, Jean Philippe Theurillat, Elizabeth Nickerson, Daniel Auclair, Liren Li, Chelsea Place, Daniel Dicara, Alex H. Ramos, Michael S. Lawrence, Kristian Cibulskis, Andrey Sivachenko, Douglas Voet, Gordon Saksena, Nicolas Stransky, Robert C. Onofrio, Wendy Winckler, Kristin Ardlie, Nikhil Wagle, Jennifer Wargo, Kelly Chong, Donald L. Morton, Katherine Stemke-Hale, Guo Chen, Michael Noble, Matthew Meyerson, John E. Ladbury, Michael A. Davies, Jeffrey E. Gershenwald, Stephan N. Wagner, Dave S.B. Hoon, Dirk Schadendorf, Eric S. Lander, Stacey B. Gabriel, Gad Getz, Levi A. Garraway, and Lynda Chin. A landscape of driver mutations in melanoma. *Cell*, 150(2):251–263, 2012.

[141] Kuan-Lin Huang, R Jay Mashl, Yige Wu, Deborah I Ritter, Jiayin Wang, Clara Oh, Marta Paczkowska, Sheila Reynolds, Matthew A Wyczalkowski, Ninad Oak, Adam D Scott, Michal Krassowski, Andrew D Cherniack, Kathleen E Houlahan, Reyka Jayasinghe, Liang-Bo Wang, Daniel Cui Zhou, Di Liu, Song Cao, Young Won Kim, Amanda Koire, Joshua F McMichael, Vishwanathan Hucthagowder, Tae-Beom Kim, Abigail Hahn, Chen Wang, Michael D McLellan, Fahd Al-Mulla, Kimberly J Johnson, Cancer Genome Atlas Research Network, Olivier Lichtarge, Paul C Boutros, Benjamin Raphael, Alexander J Lazar, Wei Zhang, Michael C Wendl, Ramaswamy Govindan, Sanjay Jain, David Wheeler, Shashikant Kulkarni, John F Dipersio, Jüri Reimand, Funda Meric-Bernstam,

Ken Chen, Ilya Shmulevich, Sharon E Plon, Feng Chen, and Li Ding. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, 173(2):355–370, 2018.

[142] Serena Nik-Zainal, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A. Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, Richard Rance, King Wai Lau, Laura J. Mudie, Ignacio Varela, David J. McBride, Graham R. Bignell, Susanna L. Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, Patrick S. Tarpey, Helen R. Davies, Elli Papaemmanuil, Philip J. Stephens, Stuart McLaren, Adam P. Butler, Jon W. Teague, Göran Jönsson, Judy E. Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerod, Andrew Tutt, John W.M. Martens, Samuel A.J.R. Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne Lise Borresen-Dale, Andrea L. Richardson, Michael S. Neuberger, P. Andrew Futreal, Peter J. Campbell, and Michael R. Stratton. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 2012.

[143] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Marcin Imielinsk, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[144] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai YinKok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2017.

[145] Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C. Wendl, Jaegil Kim, Brendan Reardon, Patrick Kwok Shing Ng, Kang Jin Jeong, Song Cao, Zhining Wang, Jianjiong Gao, Qingsong Gao, Fang Wang, Eric Minwei Liu, Loris Mularoni, Carlota Rubio-Perez, Niranjan Nagarajan, Isidro Cortés-Ciriano, Daniel Cui Zhou, Wen Wei Liang, Julian M. Hess, Venkata D. Yellapantula, David Tamborero, Abel Gonzalez-Perez, Chayaporn Suphavilai, Jia Yu Ko, Ekta Khurana, Peter J. Park, Eliezer M. Van Allen, Han Liang, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok Shing Ng, Arvind Rao, Michael Ryan, Jioajiao Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, Harsha Vardhan

Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric

Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra

Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl Friedrich Becker, Klaus Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M.Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Ken Chen, and Rachel Karchin. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 174(4):1034–1035, 2018.

[146] Li Ding, Matthew H Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L Gibbs, Amila Weerasinghe, Kuan-Lin Huang, Collin Tokheim, Isidro Cortés-Ciriano, Reyka Jayasinghe, Feng Chen, Lihua Yu, Sam Sun, Catharina Olsen, Jaegil Kim, Alison M Taylor, Andrew D Cherniack, Rehan Akbani, Chayaporn Suphavilai, Niranjan Nagarajan, Joshua M Stuart, Gordon B Mills, Matthew A Wyczalkowski, Benjamin G Vincent, Carolyn M Hutter, Jean Claude Zenklusen, Katherine A Hoadley, Michael C Wendl, Llya

Shmulevich, Alexander J Lazar, David A Wheeler, Gad Getz, and Cancer Genome Atlas Research Network. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, 173(3):305–320, 2018.

[147] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.

[148] Amit G. Deshwar, Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16:35, 2015.

[149] Yuchao Jiang, Yu Qiu, Andy J. Minn, and Nancy R. Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–37, 2016.

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

[151] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014.

[152] Nicholas E. Navin. The first five years of single-cell cancer genomics and beyond. *Genome Research*, 25(10):1499–1507, 2015.

[153] Kyung I. Kim and Richard Simon. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, 15:27, 2014.

[154] Nicholas E Navin and Ken Chen. Genotyping tumor clones from single-cell data. *Nature Methods*, 13(7):555–556, 2016.

[155] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17:86, 2016.

[156] Jack Kuipers, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, 27(11):1885–1894, 2017.

[157] Andrew Roth, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A. Smith, Cydney B. Nielsen, Jessica N. McAlpine, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature Methods*, 13(7):573–576, 2016.

[158] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. ddClone: Joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biology*, 18(1):44, 2017.

[159] Salem Malikic, Katharina Jahn, Jack Kuipers, Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *bioRxiv*, DOI:10.1101/234914, 2017.

[160] Sören Müller, Siyuan John Liu, Elizabeth Di Lullo, Martina Malatesta, Alex A Pollen, Tomasz J Nowakowski, Gary Kohanbash, Manish Aghi, Arnold R Kriegstein, Daniel A Lim, and Aaron Diaz. Single-cell sequencing maps gene expression to mutational phylogenies in PDGF- and EGF-driven gliomas. *Molecular Systems Biology*, 12(11):889, 2016.

[161] Itay Tirosh, Andrew S. Venteicher, Christine Hebert, Leah E. Escalante, Anoop P. Patel, Keren Yizhak, Jonathan M. Fisher, Christopher Rodman, Christopher Mount, Mariella G. Filbin, Cyril Neftel, Niyati Desai, Jackson Nyman, Benjamin Izar, Christina C. Luo, Joshua M. Francis, Aanand A. Patel, Maristela L. Onozato, Nicolo Riggi, Kenneth J. Livak, Dave Gennert, Rahul Satija, Brian V. Nahed, William T. Curry, Robert L. Martuza, Ravindra Mylvaganam, A. John Iafrate, Matthew P. Frosch, Todd R. Golub, Miguel N. Rivera, Gad Getz, Orit Rozenblatt-Rosen, Daniel P. Cahill, Michelle Monje, Bradley E. Bernstein, David N. Louis, Aviv Regev, and Mario L. Suvà. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309–313, 2016.

[162] Jean Fan, Hae Ock Lee, Soohyun Lee, Daeun E. Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolaos Barkas, Peter J. Park, Woong Yang Park, and Peter V. Kharchenko. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Research*, 28(8):1217–1227, 2018.

[163] Kieran R. Campbell, Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Hossein Farahani, Farhia Kabeer, Ciara O'Flanagan, Justina Biele, Jazmine Brimhall, Beixi Wang, Pascale Walters, Imaxt Consortium, Alexandre Bouchard-Côté, Samuel Aparicio, and Sohrab P. Shah. Clonealign: Statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biology*, 20(1):54, 2019.

[164] Alice Giustacchini, Supat Thongjuea, Nikolaos Barkas, Petter S. Woll, Benjamin J. Povinelli, Christopher A.G. Booth, Paul Sopp, Ruggiero Norfo, Alba Rodriguez-Meira, Neil Ashley, Lauren Jamieson, Paresh Vyas, Kristina Anderson, Åsa Segerstolpe, Hong Qian, Ulla Olsson-Strömberg, Satu Mustjoki, Rickard Sandberg, Sten Eirik W. Jacobsen, and Adam J. Mead. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nature Medicine*, 23(6):692–702, 2017.

[165] Lih Feng Cheow, Elise T. Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S.W. Tan, Paul Robson, Yuin Han Loh, Stephen R. Quake, and William F. Burkholder. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13(10):833–836, 2016.

[166] Mridusmita Saikia, Philip Burnham, Sara H Keshavjee, Michael F Z Wang, Michael Heyang, Pablo Moral-Lopez, Meleana M Hinchman, Charles G Danko, John S L Parker, and Iwijn De Vlaminck. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nature methods*, 16(1):59–62, 2019.

[167] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 2019.

[168] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018.

[169] Kevin Gori and Adrian Baez-Ortega. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv*, DOI:10.1101/372896, 2018.

[170] Marc J. Williams, Benjamin Werner, Timon Heide, Christina Curtis, Chris P. Barnes, Andrea Sottoriva, and Trevor A. Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50(6):895–903, 2018.

[171] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041, 2017.

[172] Benjamin D. Simons. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proceedings of the National Academy of Sciences*, 113(1):128–133, 2016.

[173] Marc J. Williams, Benjamin Werner, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238–244, 2016.

[174] Antonio Scialdone, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

[175] Di Wu and Gordon K. Smyth. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, 2012.

[176] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–40, 2011.

[177] Wilhelm N. Meigel, Steffen Gay, and Lutz Weber. Dermal architecture and collagen type distribution. *Archives for Dermatological Research*, 259(1):1–10, 1977.

[178] J. Michael Sorrell, David A. Carrino, Marilyn A. Baber, Daniel Assclineau, and Arnold I. Caplan. A monoclonal antibody which recognizes a glycosaminoglycan epitope in both dermatan sulfate and chondroitin sulfate proteoglycans of human skin. *Histochemical Journal*, 31(8):549–558, 1999.

[179] Robert A. Harper and Gary Grove. Human skin fibroblasts derived from papillary and reticular dermis: Differences in growth potential in vitro. *Science*, 204(4392):526–527, 1979.

[180] Irwin A. Schafer, Maureen Pandy, Roderick Ferguson, and Bryan R. Davis. Comparative observation of fibroblasts derived from the papillary and reticular dermis of infants and adults: Growth kinetics, packing density at confluence and surface morphology. *Mechanisms of Ageing and Development*, 31(3):275–293, 1985.

[181] Atsushi Akagi, Shingo Tajima, Akira Ishibashi, Noriko Yamaguchi, and Yutaka Nagai. Expression of type XVI collagen in human skin fibroblasts: Enhanced expression in fibrotic skin diseases. *Journal of Investigative Dermatology*, 113(2):246–250, 1999.

[182] Iñigo Martincorena, Philip H. Jones, and Peter J. Campbell. Constrained positive selection on cancer mutations in normal skin. *Proceedings of the National Academy of Sciences*, 113(9):E1128–9, 2016.

[183] Luis B. Barreiro, John C. Marioni, Ran Blekhman, Matthew Stephens, and Yoav Gilad. Functional comparison of innate immune signaling pathways in primates. *PLoS Genetics*, 6(12):e1001249, 2010.

[184] K. Schroder, K. M. Irvine, M. S. Taylor, N. J. Bokil, K.-A. Le Cao, K.-A. Masterman, L. I. Labzin, C. A. Semple, R. Kapetanovic, L. Fairbairn, A. Akalin, G. J. Faulkner, J. K. Baillie, M. Gongora, C. O. Daub, H. Kawaji, G. J. McLachlan, N. Goldman, S. M. Grimmond, P. Carninci, H. Suzuki, Y. Hayashizaki, B. Lenhard, D. A. Hume, and M. J. Sweet. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proceedings of the National Academy of Sciences*, 109(16):E944–53, 2012.

[185] T. Shay, V. Jojic, O. Zuk, K. Rothamel, D. Puyraimond-Zemmour, T. Feng, E. Wakamatsu, C. Benoist, D. Koller, and A. Regev. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proceedings of the National Academy of Sciences*, 110(8):2946–51, 2013.

[186] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.

[187] Alex T. Kalinka, Karolina M. Varga, Dave T. Gerrard, Stephan Preibisch, David L. Corcoran, Julia Jarrells, Uwe Ohler, Casey M. Bergman, and Pavel Tomancak. Gene expression divergence recapitulates the developmental hourglass model. *Nature*, 468(7325):811–814, 2010.

[188] Philipp Khaitovich, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. Evolution of primate gene expression. *Nature Reviews Genetics*, 7(9):693–702, 2006.

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

[190] Philipp Angerer, Laleh Haghverdi, Maren Büttner, Fabian J. Theis, Carsten Marr, and Florian Buettner. Destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32(8):1241–1243, 2016.

[191] Matthieu Deschamps, Guillaume Laval, Maud Fagny, Yuval Itan, Laurent Abel, Jean Laurent Casanova, Etienne Patin, and Lluis Quintana-Murci. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *American Journal of Human Genetics*, 98(1):5–21, 2016.

[192] John R.S. Newman, Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. DeRisi, and Jonathan S. Weissman. Single-cell

proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

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

[194] Yanick J. Crow and Nicolas Manel. Aicardi-Goutières syndrome and the type I interferonopathies. *Nature Reviews Immunology*, 15(7):429–440, 2015.

[195] John C. Hall and Antony Rosen. Type I interferons: Crucial participants in disease amplification in autoimmunity. *Nature Reviews Rheumatology*, 6(1):40–49, 2010.

[196] J. R. Tisoncik, M. J. Korth, C. P. Simmons, J. Farrar, T. R. Martin, and M. G. Katze. Into the Eye of the Cytokine Storm. *Microbiology and Molecular Biology Reviews*, 76(1):16–32, 2012.

[197] Gabriel E. Hoffman and Eric E. Schadt. variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, 17(1):483, 2016.

[198] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[199] C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle. LIMIX: genetic analysis of multiple traits. *Nature Methods*, 12(8):755–758, 2015.

[200] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[201] Zhiqiang Zhang, Taeil Kim, Musheng Bao, Valeria Facchinetti, Sung Yun Jung, Amir Ali Ghaffari, Jun Qin, Genhong Cheng, and Yong Jun Liu. DDX1, DDX21, and DHX36 Helicases Form a Complex with the Adaptor Molecule TRIF to Sense dsRNA in Dendritic Cells. *Immunity*, 34(6):866–78, 2011.

[202] Hiroki Itoh, Megumi Tatematsu, Ayako Watanabe, Katsunori Iwano, Kenji Funami, Tsukasa Seya, and Misako Matsumoto. UNC93B1 physically associates with human TLR8 and regulates TLR8-mediated signaling. *PLoS ONE*, 6(12):e28500, 2011.

[203] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[204] Yanick J. Crow, Bruce E. Hayward, Rekha Parmar, Peter Robins, Andrea Leitch, Manir Ali, Deborah N. Black, Hans Van Bokhoven, Han G. Brunner, Ben C. Hamel, Peter C. Corry, Frances M. Cowan, Suzanne G. Frints, Joerg Klepper, John H. Livingston, Sally Ann Lynch, Roger F. Massey, Jean François Meritet, Jacques L. Michaud, Gerard Ponsot, Thomas Voit, Pierre Lebon, David T. Bonthron, Andrew P. Jackson, Deborah E. Barnes, and Tomas Lindahl. Mutations in the gene encoding the 3-5 DNA exonuclease TREX1 cause Aicardi-Goutières syndrome at the AGS1 locus. *Nature Genetics*, 38(8):917–920, 2006.

[205] Gillian Rice, William G. Newman, John Dean, Teresa Patrick, Rekha Parmar, Kim Flintoff, Peter Robins, Scott Harvey, Thomas Hollis, Ann O'Hara, Ariane L. Herrick, Andrew P. Bowden, Fred W. Perrino, Tomas Lindahl, Deborah E. Barnes, and Yanick J. Crow. Heterozygous Mutations in TREX1 Cause Familial Chilblain Lupus and Dominant Aicardi-Goutières Syndrome. *The American Journal of Human Genetics*, 80(4):811–815, 2007.

[206] Jason M. Fye, Clinton D. Orebaugh, Stephanie R. Coffin, Thomas Hollis, and Fred W. Perrino. Dominant mutations of the TREX1 exonuclease gene in lupus and aicardi-goutières syndrome. *Journal of Biological Chemistry*, 286(37):32373–82, 2011.

[207] Abhishek K. Sarkar, Po Yuan Tung, John D. Blischak, Jonathan E. Burnett, Yang I. Li, Matthew Stephens, and Yoav Gilad. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS genetics*, 15(4):e1008045, 2019.

[208] Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J. Knights, Alice L. Mann, Kousik Kundu, Christine Hale, Gordon Dougan, and Daniel J. Gaffney. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, 50(3):424–431, 2018.

[209] Anna Lyubimova, Shalev Itzkovitz, Jan Philipp Junker, Zi Peng Fan, Xuebing Wu, and Alexander Van Oudenaarden. Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols*, 8(9):1743–1758, 2013.

[210] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[211] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron*, 92(2):342–357, 2016.

[212] Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-Cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135, 2013.

[213] Hongshan Guo, Ping Zhu, Fan Guo, Xianlong Li, Xinglong Wu, Xiaoying Fan, Lu Wen, and Fuchou Tang. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nature Protocols*, 10(5):645–659, 2015.

[214] Christof Angermueller, Stephen J. Clark, Heather J. Lee, Iain C. Macaulay, Mabel J. Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A. Smallwood, Chris P. Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle, and Wolf Reik. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–232, 2016.

Appendix A

Overview of HipSci fibroblast lines

Table A.1 Overview of HipSci lines used in stimulation experiments.

Donor	Grade	Passage	Gender	Age	Ethnicity
bima	3/4	12	male	40-44	White - White British
bubh	3/4	7	female	35-39	White - White British
ceik	NA	8	male	50-54	White - White British
ciwj	4	6	female	35-39	White - White British
cuhk	NA	7	female	45-49	White - White British
deyz	4	7	female	55-59	White - White British
diku	NA	7	female	60-64	White - White British
dons	4	7	female	55-59	White - White British
eika	4	6	male	45-49	White - White British
eipl	3/4	6	female	40-44	White - White British
eiwy	NA	13	female	65-69	White - White British
eofe	4	7	female	45-49	White - White British
euts	NA	7	male	60-64	White - White British
fawm	4	8	female	70-74	White - White British
feec	3		male	60-64	White - White British
fiaj	NA	6	male	55-59	White - White British
fikt	NA	7	male	50-54	White - White British

Donor	Grade	Passage	Gender	Age	Ethnicity
garx	4	6	female	50-54	White - White British
gesg	4	7	male	60-64	White - White British
gifk	4	8	male	55-59	White - White British
hehd	3/4	7	female	60-64	White - White British
heja	NA	7	male	70-74	White - White British
hiaf	NA	6	male	65-69	White - White British
hipn	3	8	male	55-59	White - White British
jogf	4/5	6	male	30-34	White - White British
joxm	4	7	female	45-49	White - Other
kuco	4	6	female	65-69	White - White British
laey	4/5	7	female	70-74	White - White British
lexy	NA	7	female	60-64	White - White British
lise	3	5	female	45-49	White - White British
melw	4	7	male	60-64	White - White British
miaj	4/5	7	male	50-54	White - White British
naju	4/5	6	male	60-64	White - White British
nusw	NA	7	male	65-69	White - White British
oaaz	NA	5	male	70-74	White - White British
oaqd	4	7	male	55-59	White - White British
oicx	4/5	7	female	65-69	White - White British
oilg	3/4	8	male	65-69	White - White British
ouvb	4		female	50-54	White - White British
pahc	4	7	female	55-59	White - White British
pamv	NA	6	male	65-69	White - White British
pelm	NA	12	female	40-44	White - White British
pipw	4	6	male	50-54	White - White British
puie	5	6	male	60-64	White - White British
qaqx	NA	6	female	60-64	White - White British
qolg	4	7	male	35-39	White - White British
qonc	NA	6	female	65-69	White - White British

Donor	Grade	Passage	Gender	Age	Ethnicity
quls	4	7	male	55-59	White - White British
rozh	3/4	8	female	65-69	White - White British
rutc	4/5	6	female	60-64	White - White British
sebz	NA	6	female	55-59	White - White British
sehl	4	8	female	55-59	White - White British
sohd	4	7	female	70-74	White - White British
tixi	4	7	female	70-74	White - White British
tolg	4	6	male	70-74	White - White British
toss	NA	6	male	65-69	White - White British
tuju	3/4	7	female	50-54	White - White British
ualf	3/4	9	female	55-59	White - White British
vabj	4	6	female	50-54	White - White British
vass	4	6	female	30-34	White - White British
vils	4	7	female	35-39	White - White British
vuna	4/5	11	female	65-69	White - White British
wahn	4/5	6	female	65-69	White - White British
wetu	4	10	female	55-59	White - White British
wigw	3/4	7	male	65-69	White - White British
wopl	3/4	6	male	55-59	White - White British
wuye	4/5	7	female	30-34	White - White British
xojn	4	7	female	50-54	White - White British
xugn	4	8	male	65-69	White - White British
xuja	4	6	female	45-49	White - White British
zihe	3/4	7	female	75-79	White - White British
zoxy	4	6	female	60-64	White - White British

Appendix B

Heterogeneity in primary human fibroblasts

Table B.1 Marker genes of *ex vivo* skin clusters

Gene name	P-value	Cluster
Fibroblast Type 1		
TNFAIP6	2.99E-269	0
SERPINE2	1.58E-177	0
MEDAG	1.15E-169	0
CTSL	9.59E-162	0
THBS2	9.21E-158	0
PTGES	2.56E-135	0
PDPN	3.43E-132	0
AKR1C1	9.68E-122	0
BNIP3	5.09E-110	0
NAMPT	1.55E-107	0
GLUL	2.18E-100	0
IL6	1.85E-87	0
FST	6.85E-85	0
PTX3	2.59E-74	0
MGST1	3.74E-72	0
MT1X	1.21E-68	0
ACKR3	2.83E-52	0
CXCL1	4.75E-52	0
G0S2	5.90E-21	0
COMP	2.20E-16	0
Vascular Endothelium		
TM4SF1	1.21E-207	1
DSTN	2.73E-205	1
ACTB	1.44E-165	1
HLA-DRB1	1.94E-160	1
ACTG1	6.35E-158	1
UPP1	3.84E-143	1
NCOA7	1.15E-133	1
HLA-DRA	6.80E-131	1
TSC22D1	4.41E-123	1
PLS3	5.67E-121	1
GBP2	9.47E-116	1
PRSS23	2.72E-115	1
HLA-DQB1	6.10E-112	1
SAT1	1.46E-110	1

Gene name	P-value	Cluster
SERPINE1	9.65E-110	1
YWHAH	4.29E-109	1
NEDD9	1.46E-102	1
CYR61	4.72E-101	1
SOX17	1.04E-99	1
EDN1	1.10E-60	1
Fibroblast Type 2		
PLAC9	4.67E-104	2
SFRP2	2.72E-103	2
CXCL12	4.64E-87	2
PPIC	4.18E-79	2
S100A4	6.55E-78	2
SEPP1	1.86E-76	2
TPPP3	5.44E-76	2
OLFML3	6.50E-76	2
TSC22D3	7.26E-75	2
ARL6IP5	1.28E-74	2
CRIP1	3.10E-70	2
CTSK	6.79E-70	2
ADH1B	1.00E-68	2
PTGDS	6.25E-65	2
CRABP2	6.33E-63	2
COL1A1	8.38E-43	2
COL3A1	9.42E-42	2
SOSTDC1	2.89E-35	2
GADD45B	5.53E-34	2
APOE	1.51E-21	2
Pericytes		
RGS5	2.38E-93	3
NDUFA4L2	1.40E-75	3
CALD1	1.57E-64	3
C11orf96	6.90E-53	3
LURAP1L	6.87E-52	3
MTHFD2	2.06E-49	3
CPM	4.33E-48	3
CHN1	1.19E-45	3
ID4	6.04E-43	3
EDNRB	6.38E-43	3

Gene name	P-value	Cluster
TFPI	1.40E-42	3
RRAD	2.79E-42	3
KCNE4	9.81E-37	3
VEGFA	1.86E-36	3
TPM2	1.89E-34	3
CPE	5.59E-33	3
HES4	1.32E-32	3
SRGN	6.27E-30	3
ACTA2	4.86E-29	3
MT1A	1.26E-27	3
Lymphatic Endothelium		
CCL21	6.47E-114	4
TFF3	3.94E-87	4
MMRN1	4.83E-84	4
CLDN5	2.03E-51	4
FABP5	9.07E-50	4
PPFIBP1	7.79E-46	4
LYVE1	4.55E-45	4
GNAS	1.78E-37	4
LAPTM5	2.06E-37	4
GNG11	1.58E-26	4
SDPR	1.90E-25	4
RAMP2	2.95E-25	4
EGLN3	9.43E-24	4
HYAL2	2.94E-22	4
SNCG	4.56E-20	4
IRF8	1.00E-18	4
ANGPT2	3.56E-18	4
FABP4	3.63E-16	4
CXCL8	2.38E-10	4
CXCL2	2.55E-08	4
Vascular Endothelium		
CCL14	4.38E-45	5
GNG11	1.94E-41	5
ACKR1	1.62E-39	5
CD74	2.00E-39	5
HLA-DRA	5.77E-36	5
CYTL1	2.13E-34	5

Gene name	P-value	Cluster
AQP1	1.11E-32	5
ITM2A	1.09E-30	5
HLA-DPA1	3.09E-30	5
CD34	3.43E-30	5
TXNIP	5.15E-30	5
PECAM1	9.57E-30	5
TSPAN7	6.29E-29	5
CTGF	5.58E-28	5
RND1	2.09E-25	5
SELE	1.52E-22	5
FOS	9.76E-19	5
DNAJB1	9.94E-16	5
HSPA1B	1.53E-15	5
HSPA1A	3.74E-14	5

Table B.2 GO term enrichment in unstimulated fibroblast clusters

GO term ID	GO term name	Enrichment p-value
	Non-cycling 1	
GO:0071705	nitrogen compound transport	0.0363
GO:0007044	cell-substrate junction assembly	0.0194
GO:0050765	negative regulation of phagocytosis	0.0405
GO:2000808	negative regulation of synaptic vesicle clustering	0.0132
GO:2001202	negative regulation of transforming growth factor-beta secretion	0.0132
GO:0030198	extracellular matrix organization	0.00337
GO:0018149	peptide cross-linking	0.00442
GO:2000761	positive regulation of N-terminal peptidyl-lysine acetylation	0.0132
GO:0006543	glutamine catabolic process	0.0186
GO:0001666	response to hypoxia	0.0477
GO:0002902	regulation of B cell apoptotic process	0.0436
GO:0061002	negative regulation of dendritic spine morphogenesis	0.0301
GO:0060179	male mating behavior	0.0268
GO:0007166	cell surface receptor signaling pathway	0.00639
GO:0071257	cellular response to electrical stimulus	0.0381
GO:0048681	negative regulation of axon regeneration	0.0375
GO:0009612	response to mechanical stimulus	0.0375
GO:0098698	postsynaptic specialization assembly	0.0301
GO:0052047	interaction with other organism via secreted substance involved in symbiotic interaction	0.0186
GO:0060070	canonical Wnt signaling pathway	0.0424
GO:1990138	neuron projection extension	0.0301
GO:0007160	cell-matrix adhesion	0.00306
GO:2000134	negative regulation of G1/S transition of mitotic cell cycle	0.0203
GO:0071310	cellular response to organic substance	0.00306
GO:0001667	ameboidal-type cell migration	0.00337
GO:0048468	cell development	0.00852
GO:0071603	endothelial cell-cell adhesion	0.0268

GO term ID	GO term name	Enrichment p-value
GO:1904209	positive regulation of chemokine (C-C motif) ligand 2 secretion	0.0186
GO:0032286	central nervous system myelin maintenance	0.0186
GO:0051674	localization of cell	0.00337
GO:1903984	positive regulation of TRAIL-activated apoptotic signaling pathway	0.0241
GO:0090071	negative regulation of ribosome biogenesis	0.0268
GO:0046466	membrane lipid catabolic process	0.0489
GO:0033622	integrin activation	0.0436
GO:0051090	regulation of DNA-binding transcription factor activity	0.0186
GO:0097105	presynaptic membrane assembly	0.0351
GO:1990523	bone regeneration	0.0186
GO:0002328	pro-B cell differentiation	0.0371
GO:0010663	positive regulation of striated muscle cell apoptotic process	0.0371
GO:0003174	mitral valve development	0.0371
GO:1903690	negative regulation of wound healing, spreading of epidermal cells	0.0268
GO:0046855	inositol phosphate dephosphorylation	0.0371
GO:0060024	rhythmic synaptic transmission	0.0268
GO:0051705	multi-organism behavior	0.0124
GO:0003215	cardiac right ventricle morphogenesis	0.0405
GO:0060134	prepulse inhibition	0.0363
GO:0097267	omega-hydroxylase P450 pathway	0.0335
GO:0009888	tissue development	0.00306
GO:1904706	negative regulation of vascular smooth muscle cell proliferation	0.047
GO:0048514	blood vessel morphogenesis	0.00306
GO:0009404	toxin metabolic process	0.0449
GO:0071702	organic substance transport	0.0427
GO:0060414	aorta smooth muscle tissue morphogenesis	0.0241
GO:0097107	postsynaptic density assembly	0.0301
GO:1990314	cellular response to insulin-like growth factor stimulus	0.0363

GO term ID	GO term name	Enrichment p-value
GO:0035025	positive regulation of Rho protein signal transduction	0.0489
GO:0071307	cellular response to vitamin K	0.0186
GO:0045475	locomotor rhythm	0.0375
GO:0032228	regulation of synaptic transmission, GABAergic	0.0489
GO:0048589	developmental growth	0.00852
GO:0007270	neuron-neuron synaptic transmission	0.0351
GO:0048870	cell motility	0.00337
GO:1904668	positive regulation of ubiquitin protein ligase activity	0.0405
GO:0008361	regulation of cell size	0.0335
GO:1901564	organonitrogen compound metabolic process	0.0407
GO:0006469	negative regulation of protein kinase activity	0.0415
GO:0042574	retinal metabolic process	0.0375
GO:0008285	negative regulation of cell proliferation	0.0363
GO:2000272	negative regulation of signaling receptor activity	0.00639
GO:0006537	glutamate biosynthetic process	0.0268
GO:0003284	septum primum development	0.0268
GO:0060044	negative regulation of cardiac muscle cell proliferation	0.0392
GO:0070372	regulation of ERK1 and ERK2 cascade	0.0392
GO:0060613	fat pad development	0.0335
GO:0060736	prostate gland growth	0.0375
GO:0019373	epoxygenase P450 pathway	0.0392
GO:0014067	negative regulation of phosphatidylinositol 3-kinase signaling	0.0381
GO:0051895	negative regulation of focal adhesion assembly	0.0424
GO:0030334	regulation of cell migration	0.00337

Cycling 1

GO:0046826	negative regulation of protein export from nucleus	0.0282
GO:0007049	cell cycle	3.01e-06
GO:0051301	cell division	2.81e-06
GO:0051382	kinetochore assembly	0.00174
GO:0031441	negative regulation of mRNA 3'-end processing	0.0344
GO:0002052	positive regulation of neuroblast proliferation	0.0466

GO term ID	GO term name	Enrichment p-value
GO:0051253	negative regulation of RNA metabolic process	0.00887
GO:0000280	nuclear division	0.000116
GO:0045870	positive regulation of single stranded viral RNA replication via double stranded DNA intermediate	0.0106
GO:0045769	negative regulation of asymmetric cell division	0.0106
GO:0000712	resolution of meiotic recombination intermediates	0.0449
GO:0085020	protein K6-linked ubiquitination	0.0327
GO:0045786	negative regulation of cell cycle	0.00651
GO:0030263	apoptotic chromosome condensation	0.0263
GO:0031508	pericentric heterochromatin assembly	0.0152
GO:0099606	microtubule plus-end directed mitotic chromosome migration	0.0106
GO:0021873	forebrain neuroblast division	0.0263
GO:0042981	regulation of apoptotic process	0.0114
GO:0034508	centromere complex assembly	0.000175
GO:0007292	female gamete generation	0.0206
GO:0032501	multicellular organismal process	0.0344
GO:0045892	negative regulation of transcription, DNA-templated	0.0241
GO:0006259	DNA metabolic process	0.0199
GO:0007051	spindle organization	0.0039
GO:0051304	chromosome separation	0.000643
GO:0061469	regulation of type B pancreatic cell proliferation	0.0366
GO:1990001	inhibition of cysteine-type endopeptidase activity involved in apoptotic process	0.0327

Non-cycling 2

	No significant gene sets	

Transitional

GO:0051651	maintenance of location in cell	0.0479
GO:0048145	regulation of fibroblast proliferation	0.0479
GO:0034118	regulation of erythrocyte aggregation	0.0371
GO:0002317	plasma cell differentiation	0.0479
GO:0030239	myofibril assembly	0.0371

GO term ID	GO term name	Enrichment p-value
GO:1905273	positive regulation of proton-transporting ATP synthase activity, rotational mechanism	0.0371
GO:0035606	peptidyl-cysteine S-trans-nitrosylation	0.0371
GO:0007249	I-kappaB kinase/NF-kappaB signaling	0.0371
	Cycling 2	
GO:0000226	microtubule cytoskeleton organization	0.0163
GO:0051494	negative regulation of cytoskeleton organization	0.0495
GO:0007049	cell cycle	0.00457
GO:0000281	mitotic cytokinesis	0.0315
GO:0051301	cell division	0.00457
GO:1903901	negative regulation of viral life cycle	0.0358

Appendix C

Manuscript: *Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants*

Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants

Davis J. McCarthy[1,4,10*], Raghd Rostom[1,2,*], Yuanhua Huang[1,11*], Daniel J. Kunz[2,5,6], Petr Danecek[2], Marc Jan Bonder[1], Tzachi Hagai[1,2], HipSci Consortium, Wenyi Wang[8], Daniel J. Gaffney[2], Benjamin D. Simons[5,6,7], Oliver Stegle[1,3,9,#], Sarah A. Teichmann[1,2,5,#]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK; [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK; [3]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany; [4]St Vincent's Institute of Medical Research, Fitzroy, Victoria 3065, Australia. [5]Cavendish Laboratory, Department of Physics, JJ Thomson Avenue, Cambridge, CB3 0HE, UK. [6]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, CB2 1QN, UK. [7]The Wellcome Trust/Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge, UK. [8]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [9]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany; [10]Melbourne Integrative Genomics, School of Mathematics and Statistics/School of Biosciences, University of Melbourne, Parkville, 3010, Australia. [11]Department of Clinical Neurosciences, University of Cambridge, CB2 0QQ, Cambridge, UK

** These authors contributed equally to this work.*

 # Corresponding authors.

Key findings

- **A novel approach for integrating DNA-seq and single-cell RNA-seq data to reconstruct clonal substructure for single-cell transcriptomes.**
- **Evidence for non-neutral evolution of clonal populations in human fibroblasts.**
- **Proliferation and cell cycle pathways are commonly distorted in mutated clonal populations.**

Abstract

Decoding the clonal substructures of somatic tissues sheds light on cell growth, development and differentiation in health, ageing and disease. DNA-sequencing, either using bulk or using single-cell assays, has enabled the reconstruction of clonal trees from frequency and co-occurrence patterns of somatic variants. However, approaches to systematically characterize phenotypic and functional variations between individual clones are not established. Here we present cardelino (https://github.com/PMBio/cardelino), a computational method for inferring the clonal tree configuration and the clone of origin of individual cells that have been assayed using single-cell RNA-seq (scRNA-seq). Cardelino allows effective integration of information from imperfect clonal tree inferences based on bulk exome-seq data, and sparse variant alleles expressed in scRNA-seq data. After validating our model using simulations, we apply cardelino to matched scRNA-seq and exome sequencing data from 32 human dermal fibroblast lines, identifying hundreds of differentially expressed genes between cells from different somatic clones. These genes are frequently enriched for cell cycle and proliferation pathways, indicating a key role for cell division genes in non-neutral somatic evolution.

Keywords: single cell, somatic mutations, clonality

Introduction

Ageing, environment and genetic factors can impact mutational processes, thereby shaping the acquisition of somatic mutations across the life span (Burnet 1974; Martincorena and Campbell 2015; Stransky et al. 2011; Hodis et al. 2012; Huang et al. 2018). The maintenance and evolution of somatic mutations in different sub-populations of cells can result in clonal structure, both within healthy and disease tissues. Targeted, whole-genome and whole-exome DNA sequencing of bulk cell populations has been utilized to reconstruct the mutational processes that underlie somatic mutagenesis (Nik-Zainal et al. 2012; Alexandrov et al. 2013; Forbes et al. 2017; Bailey et al. 2018; Ding et al. 2018) as well as clonal trees (Roth et al. 2014; Deshwar et al. 2015; Jiang et al. 2016).

Availability of single-cell DNA sequencing methods (scDNA-seq; (N. Navin et al. 2011; Wang et al. 2014; N. E. Navin 2015) combined with new computational approaches have helped to improve the reconstruction of clonal populations (K. I. Kim and Simon 2014; N. E. Navin and Chen 2016; Jahn, Kuipers, and Beerenwinkel 2016; Kuipers et al. 2017; Roth et al. 2016; Salehi et al. 2017; Malikic et al. 2017). However, the functional differences between clones and their molecular phenotypes remain largely unknown. Systematic characterisation of the phenotypic properties of clones could reveal mechanisms underpinning healthy tissue growth and the transition from normal to malignant behaviour.

An important step towards such functional insights would be access to genome-wide expression profiles of individual clones, yielding genotype-phenotype connections for clonal architectures in tissues. Recent studies have explored mapping scRNA-seq profiles to clones with distinct copy number states in cancer, thus providing a first glimpse at clone-to-clone gene expression differences in disease (Müller et al. 2016; Tirosh et al. 2016; Fan et al. 2018; Campbell et al. 2019). Targeted genotyping strategies linking known mutations of interest to single-cell transcriptomes have proven useful in particular settings, but remain limited by technical challenges and the requirement for strong prior information (Giustacchini et al. 2017; Cheow et al. 2016; Saikia et al. 2019). Generally-applicable methods for inferring the clone of origin of single cells to study genotype-transcriptome relationships are not yet established.

To address this, we have developed cardelino: a computational method that exploits variant information in scRNA-seq reads to map cells to their clone of origin. We validate our model using simulations and compare its performance to two alternative versions of the cardelino model, Single-Cell Genotyper (Roth et al. 2016), designed for clonal inference from scDNA-seq data, and Demuxlet (Kang et al. 2018), designed to infer sample identity for cells using scRNA-seq and reference genotype data. We demonstrate that cardelino allows for accurate assignment of full-length single-cell transcriptomes to the clonal substructure in 32 normal dermal fibroblast lines. With linked

somatic variants, clone and gene expression information, we investigate gene expression differences between clones at the level of individual genes and in pathways, which provides new insights into the dynamics of clones. These findings also extend recent studies using bulk DNA-seq data, predominantly in epithelial cells, that have revealed oncogenic mutations and evidence of selective clonal dynamics in normal tissue samples (Behjati et al. 2014; Martincorena et al. 2015; Simons 2016b; Martincorena, Jones, and Campbell 2016; Simons 2016a). Our approach can be applied to a broad range of somatic substructure analyses in population or disease settings to reveal previously inaccessible differences in molecular phenotypes between cells from the same individual.

Results

Mapping single-cell transcriptomes to somatic clones with cardelino

We present cardelino, a Bayesian method for integrating somatic clonal substructure and transcriptional heterogeneity within a population of cells. Briefly, cardelino models the expressed variant alleles in single cells as a clustering model, with clusters corresponding to somatic clones with (unknown) mutation states (**Fig. 1a**). Critically, cardelino leverages imperfect but informative clonal tree configurations obtained from complementary technologies, such as bulk or single-cell DNA sequencing data, as prior information, thereby mitigating the sparsity of scRNA-seq variant coverage. Cardelino employs a variant specific beta-binomial error model that accounts for stochastic dropout events as well as systematic allelic imbalance due to mono-allelic expression or genetic factors.

Initially, we assess the accuracy of cardelino using simulated data that mimic typical clonal structures and properties of scRNA-seq as observed in real data (4 clones, 10 variants per branch, 25% of variants with read coverage, 200 cells, 50 repeat experiments; **Methods**). By default, we consider an input clone configuration with a 10% error rate compared to the true simulated tree (namely, 10% of the values in the clone configuration matrix are incorrect). Alongside cardelino, we consider two alternative approaches: Single Cell Genotyper (SCG; Roth et al. 2016) and an implementation of Demuxlet, which was designed for sample demultiplexing rather than clone assignment (Kang et al. 2018; see **Methods** and **Supp. Fig. S1)**. In the default setting, cardelino achieves high overall performance (Precision-Recall AUC=0.965; **Fig. 1b**), outperforming both SCG and Demuxlet. For example, at a cell assignment confidence threshold (posterior probability of cell assignment) of P=0.5, cardelino assigns 88% of all cells with an overall accuracy of 88.6%.

We explore the effect of key dataset characteristics on cell assignment, including the number of variants per clonal branch (**Fig. 1c**) and the expected number of variants with non-zero scRNA-seq coverage per cell (**Fig. 1d**). As expected, the number of variants per clonal branch and their read coverage in scRNA-seq are positively associated with the performance of all methods, with cardelino consistently outperforming alternatives, in particular in settings with low coverage. We further explore

the effects of allelic imbalance on cell assignment (**Fig. 1e**), and find that cardelino is more robust than SCG and Demuxlet when there is a larger fraction of variants with high allelic imbalance. We attribute cardelino's robustness to its approach of modelling the allelic imbalance per variant, whereas SCG and Demuxlet both use a global parameter and hence cannot account for variability of allelic imbalance across sites. We also vary the error rate in the guide clone configuration, either introducing uniform errors in the configuration matrix by swapping the mutation states of any variants in any clone (**Fig. 1f**) or by swapping variants between branches (**Fig. 1g**). In both settings, cardelino is markedly more robust than Demuxlet, which assumes that the defined reference clonal structure is error free. Notably, cardelino retains excellent performance (AUPRC>0.96) at error rates up to 25% (**Fig. 1f-g)**, by modelling deviations between the observed and the true latent tree (**Supp. Fig. S2**).

We also consider two simplified variants of cardelino, one of which does not consider the guide clone tree and performs *de novo* tree reconstruction (cardelino-free), and a second model that treats the guide tree as fixed without modelling any errors (cardelino-fixed). These comparisons, further investigating the parameters assessed in **Fig. 1**, confirm the benefits of the data-driven modelling of the guide clone configuration as a prior that is adapted jointly while assigning scRNA-seq profiles to clones (**Supp. Fig. S3**). We also explore the effects of the number of clones (**Supp. Fig. S3c**), and the tree topology (**Supp. Fig. S4**), again finding that cardelino is robust to these parameters.

Taken together, these results demonstrate that cardelino is broadly applicable to robustly assign individual single-cell transcriptomes to clones, thereby reconstructing clone-specific transcriptome profiles.

Cardelino assigns single cell transcriptomes to clones in human dermal fibroblasts

Next, we apply cardelino to 32 human dermal fibroblast lines derived from healthy donors that are part of the UK human induced pluripotent stem cell initiative (HipSci; Kilpinen *et al.*, 2017; **Supp. Table S1**). For each line, we generated deep whole exome sequencing data (WES; median read coverage: 254), and matched Smart-seq2 scRNA-seq profiles using pools of three lines in each processing batch (**Methods**). We assayed between 30 and 107 cells per line (median 61 cells after QC; median coverage: 484k reads; median genes observed: 11,108; **Supp. Table S2**).

Initially, we consider high-confidence somatic single nucleotide variants (SNVs) identified based on WES data (**Methods**) to explore the mutational landscape across lines. This reveals considerable variation in the total number of somatic SNVs, with 41–612 variants per line (**Fig. 2a;** coverage of ≥20 reads, ≥3 observations of alternative allele, Fisher's exact test FDR≤0.1; see **Methods**). The majority of SNVs can be attributed to the well-documented UV signature, COSMIC Signature 7 (primarily C to T mutations; (Forbes et al. 2017), agreeing with expected mutational patterns from UV exposure of skin tissues (**Fig. 2a; Supp. Fig. S5; Methods**).

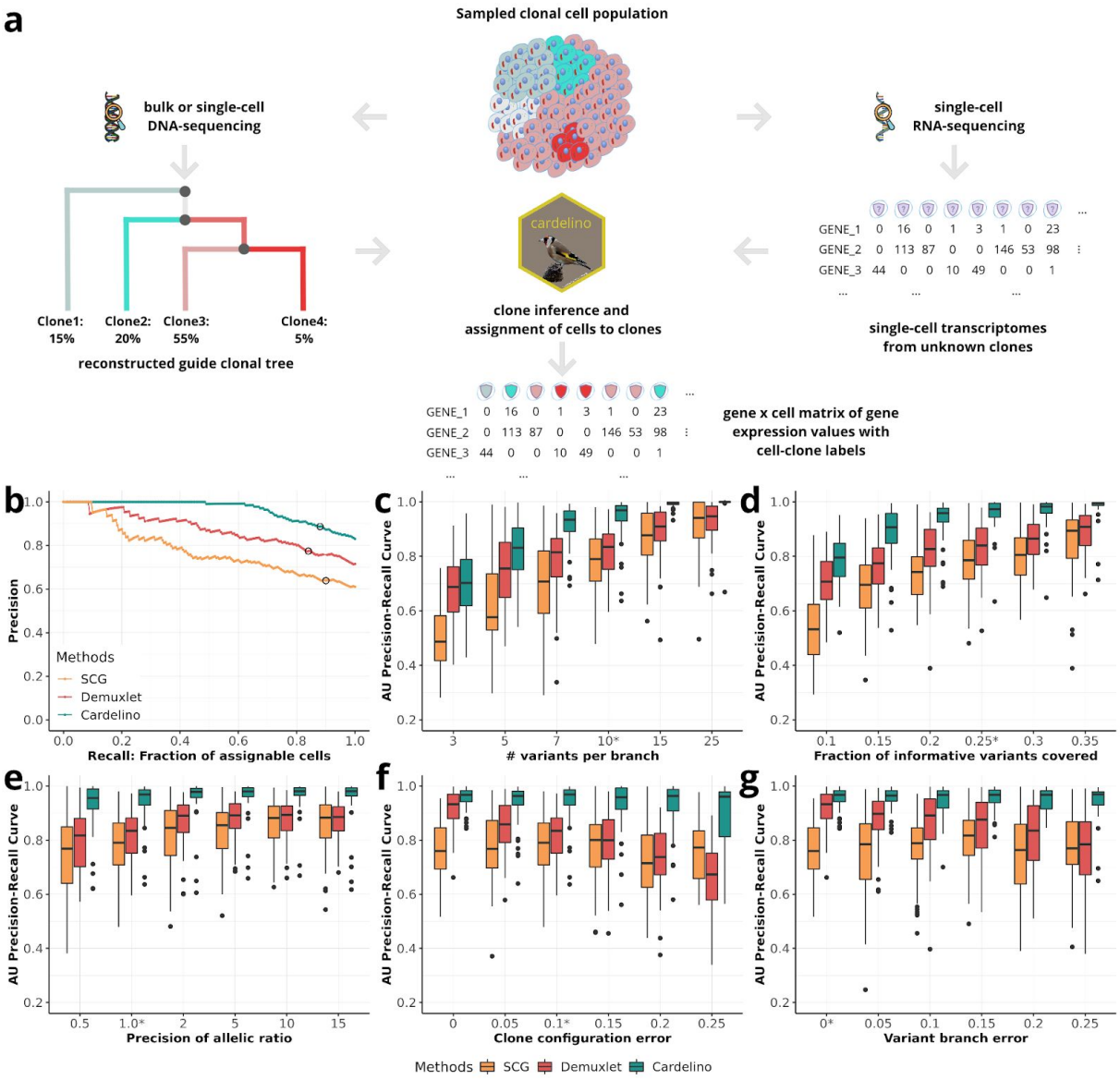


Figure 1 | Overview and validation of the cardelino model. (a) Overview and approach. A clonal tree is reconstructed using DNA-sequencing (e.g. deep exome sequencing) data to derive a guide clone configuration. Cardelino then performs probabilistic clustering of single-cell transcriptomes based on variants detected in scRNA-seq reads, assigning cells to clones in the mutation tree. **(b-g)** Benchmarking of the cell assignment using simulated data by changing one variable each time. The default values are highlighted with a star. **(b)** Overall assignment performance for a dataset consisting of 200 cells, simulated assuming a 4-clone structure with 10 variants per branch and non-zero read coverage for 20% of the variants and simulating an error rate of 10% on the mutation states between the guide clone configuration and the true clonal tree (**Methods**). Shown is the fraction of true positive cell assignments (precision) as a function of the fraction of assigned cells (recall), when varying the threshold of the cell assignment probability. The black circle corresponds to the posterior cell assignment threshold of P=0.5. **(c-g)** Area Under (AU) precision-recall curve (*i.e.* area under curves such as shown in **b**), when varying the numbers of variants per clonal branch (**c**), the fraction of informative variants covered (i.e., non-zero scRNA-seq read coverage) (**d**), the precision (i.e., inverse variance) of allelic ratio across genes; lower precision means more genes with high allelic imbalance (e), the error rate of the mutation states in clone configuration matrix (f), and the fraction of variants that are wrongly assigned to branches (g). For details and default parameter settings see **Methods**.

To understand whether the somatic SNVs confer any selective advantage in skin fibroblasts, we used SubConalSelection to identify neutral and selective dynamics at a per-line level (Williams et al. 2018). Other established methods such as dN/dS (Martincorena et al. 2018) and alternative methods using the SNV frequency distribution (Simons 2016a; Williams et al. 2016) are not conclusive in the context of this dataset, likely due to lack of statistical power resulting from the low number of mutations detected in each sample. The SubClonalSelection analysis identifies at least 10 lines with a clear fit to their selection model, suggesting positive selection of clonal sub-populations (**Fig. 2a**; **Supp. Fig. S6**; **Methods**). In other words, a third of the samples from this cohort of healthy donors contain clones evolving adaptively, which we can investigate in more detail in terms of transcriptome phenotype.

Next, we reconstruct the clonal trees in each line using WES-derived estimates of the variant allele frequency of somatic variants that are also covered by scRNA-seq reads (**Methods**). Canopy (Jiang et al. 2016) identifies two to four clones per line (**Fig. 2a**). Briefly, Canopy models the phylogeny of cell growth in a tissue by depicting a bifurcating tree arising from a diploid germline cell whose daughter cells are subject to progressive waves of somatic mutations. When a sample of a tissue is taken, the tree is sliced horizontally, cutting the branches to form “leaves” or “clones”. Thus each clone represents a subpopulation of cells that share (and are identified by) the somatic mutations in their most recent common ancestral cell. To handle the presence of a subpopulation of cells without somatic mutations, “clone1” is defined to represent a non-bifurcating, somatic mutation-free branch of the clonal tree. Thus, with any somatic variants present at sub-clonal frequencies (the case for all cell lines here), Canopy will infer the presence of at least two clones. Following Canopy’s inference of clones, we use cardelino to confidently map scRNA-seq profiles from 1,732 cells (out of a total of 2,044 cells) to clones from the corresponding lines (**Methods;** for Canopy input trees and output from cardelino for all lines see **Supp. Fig. S7-10**). Cardelino estimates an error rate in the guide clone configuration of less than 25% in most lines (median 18.6%), and assigns a large fraction of cells confidently (>90% for 23 lines; at posterior probability P>0.5; **Supp. Fig. S11**). The model identifies four lines with an error rate between 35-46% and an outlier (*vils*, a line with few somatic variants), which demonstrates the utility of the adaptive phylogeny error model employed by cardelino. We also run the other four alternative methods on these 32 lines (**Supp. Fig. S12**), and find that the *de novo* methods appear to suffer from higher uncertainty in recontrustructing clonal trees from scRNA-seq data only (**Supp. Fig. S12C**), while using the fixed-guide clonal tree from bulk exome-seq data may be over-simplified and leads to reduced stability when considering alternative high-confidence trees (**Supp. Fig. S12D-E**).

To further assess the confidence of these cell assignments, we consider, for each line, simulated cells drawn from a clonal structure that matches the corresponding line, finding that cardelino gives high accuracy (AUPRC>0.9) in 29 lines, again clearly outperforming competing methods (**Supp. Fig. S13**). Additionally, we observe high concordance (R^2 = 0.94) between the empirical cell-assignment rates

and the expected values based on the corresponding simulation for the same line (**Fig. 2b**). Lines with clones that harbour fewer distinguishing variants are associated with lower assignment rates (**Supp. Fig. S14**), at consistently high cell assignment accuracy (median 0.965, mean 0.939; **Supp. Fig. S15**), indicating that the posterior probability of assignment is calibrated across different settings. We also consider the impact of technical features of scRNA-seq data on cell assignment, finding no evidence of biased cell assignments (**Supp. Fig. S16-20**). Finally, clone prevalences estimated from Canopy and the fractions of cells assigned to the corresponding clones are reasonably concordant (adjusted R^2 = 0.53), providing additional confidence in the cardelino cell assignments, while highlighting the value of cardelino's ability to update input clone structures using single-cell variant information (**Fig. 2c**).

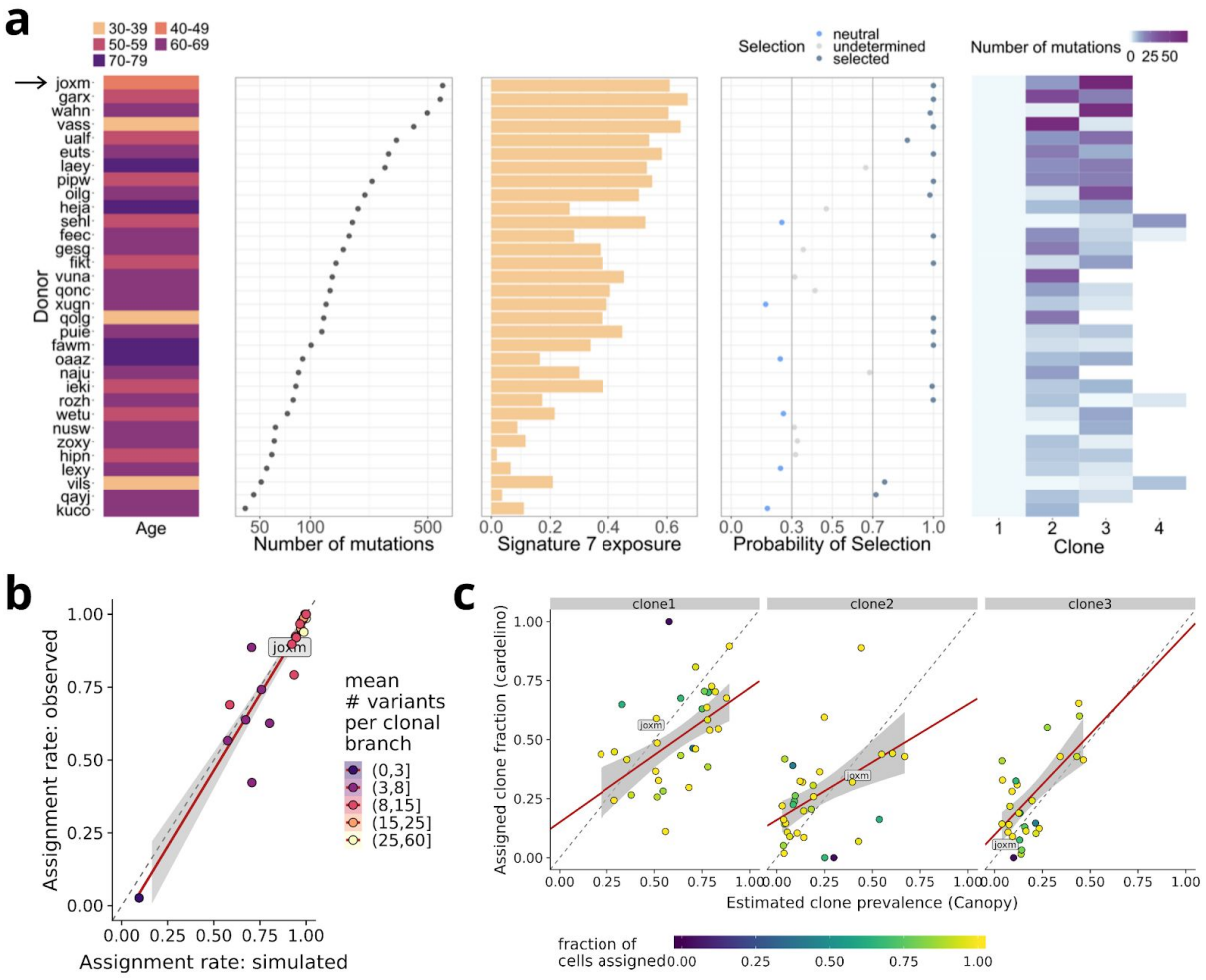


Figure 2 | Parallel deep exome sequencing and scRNA-seq profiling of 32 human dermal fibroblast lines. (a) Overview and somatic mutation profiles across lines, from left to right: donor age; number of somatic SNVs; estimated exposure of COSMIC mutational signature 7; probability of selection estimated by SubClonalSelection (Williams et al. 2018), colour denotes the selection status based on probability cut-offs (grey lines), the grey background indicates results with high uncertainty due to the low number of mutations detected; number of clones inferred using Canopy (Jiang et al., 2016), with colour indicating the number of informative somatic SNVs for cell assignment to each clone (non-zero read coverage in scRNA-seq data). **(b)** Assignment rate (fraction of cells assigned) using matched simulated single-cell transcriptomes (x-axis; **Methods**) *versus* the empirical assignment rate (y-axis) for each line (at assignment threshold posterior P>0.5). Colour denotes the average number of informative variants across clonal branches per line. The line-of-best fit from a linear model is shown in red, with 95%

confidence interval shown in grey. **(c)** Estimated clone prevalence from WES data (x-axis; using Canopy) versus the fraction of single-cell transcriptomes assigned to the corresponding clone (y-axis; using cardelino). Shown are the fractions of cells assigned to clones one to three as in **a**, considering the most likely assignment for assignable cells (posterior probability P>0.5) with each point representing a cell line; see **Supp. Fig. S21** for results from four donors with >3 clones). Colour denotes the total fraction of assignable cells per line (P>0.5). A line-of-best fit from a weighted regression model is shown in red with 95% confidence interval shown in grey.

Differences in gene expression between clones suggest phenotypic impact of somatic variants

Initially, we focus on the fibroblast line with the largest number of somatic SNVs (*joxm;* white female aged 45-49; **Fig. 2a**), with 612 somatic SNVs (112 detected both in WES and scRNA-seq) and 79 QC-passing cells, 99% of which could be assigned to one of three clones (**Fig. 3a**). Principal component analysis of the scRNA-seq profiles of these cells reveals global transcriptome substructure that is aligned with the somatic clonal structure in this population of cells (**Fig. 3b**). Additionally, we observe differences in the fraction of cells in different cell cycle stages, where clone1 has the fewest cells in G1, and the largest fraction in S and G2/M (**Fig. 3b, inset plot;** PC1 in **Supp. Fig. S22-23;** global structure and cell cycle plots for all lines in **Supp. Figs. S24-33**). This suggests that clone1 is proliferating most rapidly. Next, we consider differential expression analysis of individual genes between the two largest clones (clone1: 46 cells *versus* clone2: 25 cells), which identifies 901 DE genes (edgeR QL F-test; FDR<0.1; 549 at FDR<0.05; **Fig. 3c**). These genes are approximately evenly split into up- and down-regulated sets. However, the down-regulated genes are enriched for processes involved in the cell cycle and cell proliferation. Specifically, the three significantly enriched gene sets are all up-regulated in clone1 (camera; FDR<0.1; **Fig. 3d**). All three gene sets (E2F targets, G2/M checkpoint and mitotic spindle) are associated with the cell cycle, so these results are consistent with the cell-cycle stage assignments suggesting increased proliferation of clone1.

Taken together, the results suggest that somatic substructure in this cell population results in clones that exhibit measurably different expression phenotypes across the transcriptome, with significant differential expression in cell cycle and growth pathways.

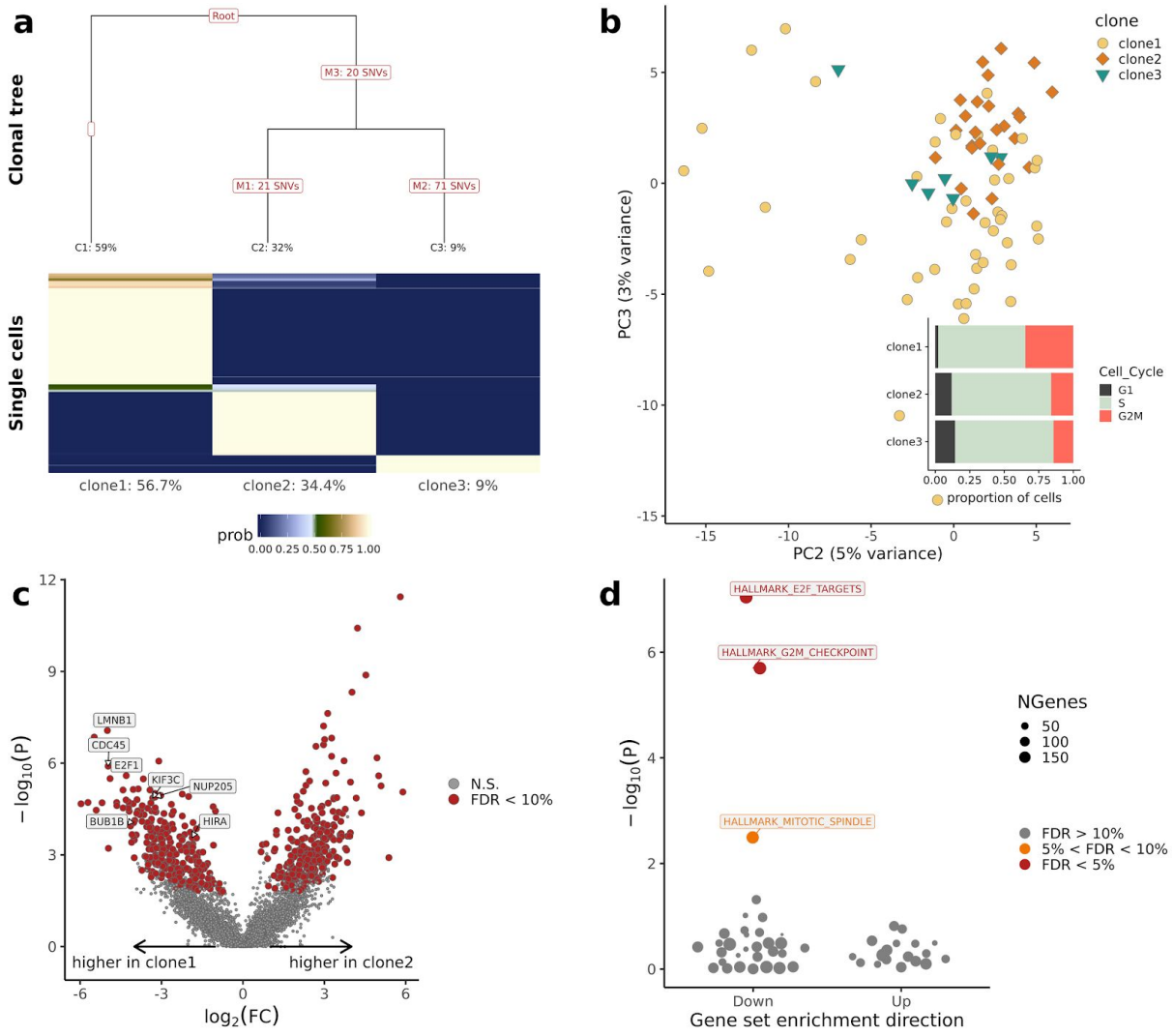


Figure 3 | Clone-specific transcriptome profiles reveal gene expression differences for *joxm,* one example line. (a) Top: Clonal tree inferred using Canopy (Jiang et al., 2016). The number of variants tagging each branch and the expected prevalence (fraction) of each clone is shown. Bottom: cardelino cell assignment matrix, showing the assignment probability of individual cells to three clones. Shown below each clone is the fraction of cells assigned to each clone. **(b)** Principal component analysis of scRNA-seq profiles with colour indicating the most likely clone assignment. Inset plot: Cell-cycle phase fractions for cells assigned to each clone (using cyclone; Scialdone et al., 2015). **(c)** Volcano plot showing negative $\log_{10}$ P values versus log fold changes (FC) for differential expression between cells assigned to clone2 and clone1. Significant differentially expressed genes (FDR<0.1) are highlighted in red. **(d)** Enrichment of MSigDB Hallmark gene sets using camera (Wu and Smyth, 2012) based on $\log_2$ FC values between clone2 and clone1 as in **c**. Shown are negative $\log_{10}$ P values of gene set enrichments, considering whether gene sets are up-regulated in clone1 or clone2, with significant (FDR < 0.05) gene sets highlighted and labelled. All results are based on 78 out of 79 cells that could be confidently assigned to one clone (posterior P>0.5; **Methods**).

Cell cycle and proliferation pathways frequently vary between clones

To quantify the overall effect of somatic substructure on gene expression variation across the entire dataset, we fit a linear mixed model to individual genes (**Methods**), partitioning gene expression variation into a line (likely donor) component, a clone component, technical batch (*i.e.* processing plate), cellular detection rate (proportion of genes with non-zero expression per cell) and residual noise. As expected, the line component typically explains a substantially larger fraction of the

expression variance than clone (median 5.5% for line, 0.5% for clone), but there are 194 genes with a substantial clone component (>5% variance explained by clone; **Fig. 4a**). Even larger clone effects are observed when estimating the clone component in each line separately, which identifies between 331 and 2,162 genes with a substantial clone component (>5% variance explained by clone; median 825 genes; **Fig. 4b)**. This indicates that there are line-specific differences in the set of genes that vary with clonal structure.

Next, we carry out a systematic differential expression (DE) analysis to assess transcriptomic differences between any pair of clones for each line (considering 31 lines with at least 15 cells for DE testing; **Methods**). This approach identifies up to 1,199 DE genes per line (FDR<0.1, edgeR QL F test). A majority, 61%, of the total set of 5,289 unique DE genes, are detected in two or more lines, and 39% are detected in at least three of the 31 lines. Comparison to data with permuted gene labels demonstrates an excess of recurrently differentially expressed genes compared to chance expectation (**Fig. 4c**, P<0.001; 1,000 permutations; **Methods**). We also identify a small number of genes that contain somatic variants in a subset of clones, resulting in differential expression between wild-type and mutated clones (**Supp. Fig. S34**).

To investigate the transcriptomic changes between cells in more detail, we use gene set enrichment analysis in each line. This approach reveals whether there is functional convergence at a pathway level (using MSigDB Hallmark gene sets; **Methods**; (Liberzon et al. 2011)). Of 31 lines tested, 19 have at least one significant MSigDB Hallmark gene set (FDR<0.05, camera; **Methods**), with key gene sets related to cell cycle and growth being significantly enriched in all of those 19 lines. Directional gene expression changes of gene sets for the *E2F* targets, G2M checkpoint, mitotic spindle and MYC target pathways are highly coordinated (**Fig. 4d**), despite limited overlap of individual genes between the gene sets (**Supp. Fig. S35**).

Similarly, directional expression changes for pathways of epithelial-mesenchymal transition (EMT) and apical junction are correlated with each other. Interestingly, these are anti-correlated with expression changes in cell cycle and proliferation pathways (**Fig. 4d**). Within individual lines, the enrichment of pathways often differs between pairs of clones, highlighting the variability in effects of somatic variants on the phenotypic behaviour of cells (**Fig. 4e;** all lines shown in **Supp. Fig. S36**).

These consistent pathway enrichments across a larger set of donors point to somatic variants commonly affecting the cell cycle and cell growth in fibroblast cell populations. These results indicate both deleterious and adaptive effects of somatic variants on proliferation, suggesting that a significant fraction of these variants are non-neutral in the majority of donors in our study.

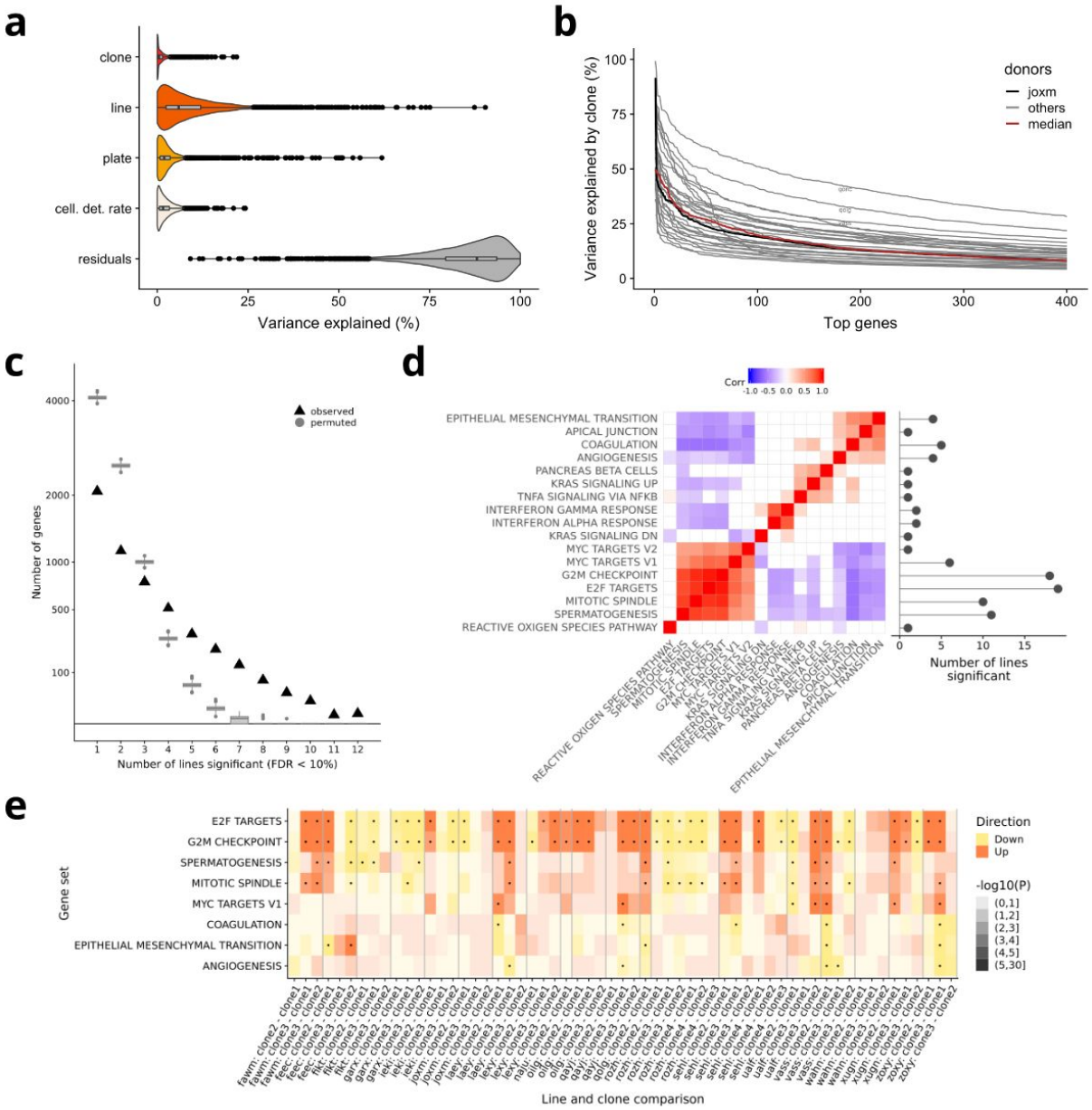


Figure 4 | Signatures of transcriptomic clone-to-clone variation across 31 lines. (a) Violin and box plots show the percentage of variance explained by clone, line, experimental plate and cellular detection rate for 4,998 highly variable genes, estimated using a linear mixed model (**Methods**). **(b)** Percentage of gene expression variance explained by clone when fitting a linear mixed model for each individual line for the 400 genes with the most variance explained by clone per line (**Methods**). Individual lines correspond to cell lines (donors), with *joxm* highlighted in black and the median across all lines in red. **(c)** The number of recurrently differentially expressed (DE) genes between any pair of clones (FDR<0.1; edgeR QL F test), detected in at least one to 12 lines, with box plots showing results expected by chance (using 1,000 permutations). **(d)** Left panel: Heatmap showing pairwise correlation coefficients (Spearman R, only nominal significant correlations shown (P<0.05)) between signed P-values of gene set enrichment across lines, based on differentially expressed genes between clones. Shown are the 17 most frequently enriched MSigDB Hallmark gene sets. Right panel: number of lines in which each gene set is found to be significantly enriched (FDR<0.05). **(e)** Heatmap depicting signed P-values of gene set enrichments for eight Hallmark gene sets in 19 lines. Dots denote significant enrichments (FDR<0.05).

Discussion

Here, we develop and apply a computational approach for integrating somatic clonal structure with single-cell RNA-seq data. This allows us to identify molecular signatures that differ between clonal cell populations. Our approach is based on first inferring clonal structure in a population of cells using WES data, followed by the assignment of individual single-cell transcriptomes to clones using a computational approach called cardelino. Our method enables the efficient reconstruction of clone-specific transcriptome profiles from high-throughput assays. Our integrative analysis of bulk WES and scRNA-seq data from 32 human fibroblast cell lines reveals substantial phenotypic effects of somatic variation, including in healthy tissue.

Central to our approach is cardelino, a robust model for clone inference and the probabilistic assignment of cells to clones based on variants contained in scRNA-seq reads. Our approach is conceptually related to de-multiplexing methods for single-cell transcriptomes from multiple genetically distinct individuals (Kang et al. 2018). However, cardelino addresses a substantially more challenging problem: to distinguish cells from the same individual based on the typically small number of somatic variants (*e.g.* dozens) that segregate between clones in a population of cells. Cardelino simultaneously infers the clonal tree configuration and the clone of origin of individual cells based on sparse variant alleles observed in scRNA-seq data, while leveraging imperfect clonal trees derived from complementary assays such as bulk exome-seq data.

Inferring clonal trees from any type of data remains a hard problem and all clonal inference methods produce clonal trees with substantial uncertainty, so cardelino's flexible approach to integrating variant information from scRNA-seq and other data sources is a key strength of the method. Our results show that cardelino outperforms methods that use an input clonal tree as fixed and error-free (Demuxlet, cardelino-fixed) and methods that do not use any guide tree at all (SCG, cardelino-free), confirming the utility of flexible, data-driven incorporation of multiple sources of information on clonal structure. Surprisingly, cardelino-free also performs strongly, better than SCG and almost as well as cardelino in some settings, demonstrating that our underlying modeling of allele counts in scRNA-seq data works well enough to yield excellent clone inference and cell-clone assignment results even when no external information about clonal structure is available.

Harnessing transcriptomic phenotypic information for cells assigned to clones in fibroblast lines, we identify substantial and convergent gene expression differences between clones across lines, which are enriched for pathways related to proliferation and the cell cycle. Analysis of clonal evolutionary dynamics using somatic variant allele frequency distributions from WES data reveals evidence for positive selection of clones in ten of 32 lines. These results support previous observations of clonal populations undergoing positive selection in normal human eyelid epidermis assayed by targeted DNA sequencing (Martincorena et al. 2015; Simons 2016b; Martincorena, Jones, and Campbell 2016;

Simons 2016a). We shed light on the phenotypic effects of this adaptive evolution, consistently identifying differential expression of gene sets implicated in proliferation and cancer such as the E2F and MYC pathways. This surprising result in healthy tissue suggests pervasive inter-clonal phenotypic variation with important functional consequences, although we do note that clonal dynamics *in vivo* in primary fibroblast tissue may differ somewhat from what we observe in the fibroblast cell lines. It is intriguing to speculate about potential mechanisms driving these inter-clonal phenotypic differences, which might stem solely from observed somatic variants, could involve unobserved variants, or could arise through indirect mechanisms involving (post-)transcriptional regulation or epigenetic differences. Further work will be needed to identify drivers of molecular differences between clones across biological systems.

The clones studied here each represent a subpopulation of cells that share and are identified by the somatic variants in their most recent common ancestral cell. Individual cells in each clone would be undergoing further mutation that could lead to genetic and molecular differences between cells grouped into the same clone, and so cells assigned to a given clone will not be completely genetically or transcriptomically homogenous. Thus, within-clone heterogeneity could limit the ability of downstream analyses to identify differences in expression or molecular phenotypes between clones. Clonal inference depends heavily on the set of somatic variants supplied, so careful calling of somatic SNVs is a vital step before clonal inference with Canopy, cardelino and other tools. We found clonal inference methods to perform better with strictly filtered somatic SNVs, so here we preferred a conservative somatic variant calling approach that emphasised specificity over sensitivity. Future studies would therefore benefit from higher-depth sequencing of DNA, either with bulk or single-cell approaches, to better identify somatic variants and thus enable confident inference of more complex clonal structures. Increasing both the number of genetically distinct individuals and the numbers of cells assayed per individual would further improve power to find molecular differences between clones.

While we use clonal trees from bulk WES data as input to cardelino in this study, our method is general and can exploit prior information on clonal substructure inferred from either bulk or single-cell DNA-seq data. Our cardelino-free method also works when no external information on clonal structure is available. The methods presented here can be applied to any system in which somatic variants tag clonal populations of cells and can be accessed with scRNA-seq assays. Though not explored here, we also expect the cardelino model to be effective for other single-cell 'omics assays that capture somatic variant information, such as those profiling chromatin accessibility (Buenrostro et al. 2015) or methylation (Guo et al. 2013; Smallwood et al. 2014). Assignment of cells to clones relies on coverage of somatic variants in scRNA-seq reads, so cell populations with relatively fewer somatic variants may require full-length transcriptome sequencing at higher coverage per cell to enable confident assignments. Our inference methods in cardelino are computationally efficient, so will comfortably scale to multi-site samples and many thousands of cells. Thus, cardelino will be applicable to

high-resolution studies of clonal gene expression in both healthy and malignant cell populations as well as *in vitro* models.

Taken together, our results highlight the utility of cardelino to study gene expression variability in clonal cell populations and suggest that even in nominally healthy human fibroblast cell lines there are clonal populations with growth advantages, opening new avenues to study cell behaviour in clonal populations.

Methods

The cardelino model

The cardelino model jointly infers the clonal tree configuration and assigns single cells to one of the clones by modelling the expressed alleles with a probabilistic clustering model (see graphical model in **Supp. Fig. S37**). The unobserved clonal tree configuration *C* is an N-by-K binary matrix for N variants and K clones encoding the mutation profile for each clone. We let $c_{i,k}$=1 if somatic variant i is present in clone k and $c_{i,k}$=0 otherwise. Cardelino allows for incorporating a guide clone configuration Ω (an analogous binary matrix) as prior, for which an appropriate relaxation (or error) rate ξ is inferred. The probability of the entries in the latent clonal configuration matrix C are modelled as

$$P(c_{i,k} = 1|\Omega, \xi) = \xi^{(1-\Omega_{i,k})}(1-\xi)^{\Omega_{i,k}} \quad (1)$$

The prior clone configuration Ω is assumed to be informative but imperfect. In this study, we used the clone configuration derived from bulk exome-seq data by Canopy to define the prior Ω and to estimate the number of clones.

Based on scRNA-seq data, we extract for each cell and variant that segregates between clones the number of sequencing reads that support the reference allele (reference read count) or the alternative allele (alternate read count) respectively. We denote the variant-by-cell matrix of alternate read counts by A with element $a_{i,j}$ denoting the number of reads supporting the alternative allele for variant *i* in cell *j* and similarly the variant-by-cell matrix of total read counts (sum of reference and alternate read counts) by D. Entries in A and D matrices are non-negative integer values, with missing entries in the matrix D indicating zero read coverage for a given cell and variant.

Fundamentally, we model the alternate read count using a binomial model, using a variant-specific beta distribution on the binomial rate, thereby modelling overdispersion as well as systematic errors. For a given site in a given cell, there are two possibilities: the variant is “absent” in the clone the cell is assigned to or the variant is “present”, as encoded in the configuration matrix C. Thus, the “success probability” $\boldsymbol{\theta}$ for the binomial model for each variant, where success is defined as observing an alternate read in the scRNA-seq reads, is modelled using two (sets of) parameters: θ_0 for homozygous reference alleles (variant absent), and $\boldsymbol{\theta}_1$ for heterozygous variants (variant present). The likelihood for cell j given an assignment to clone k follows then as a product of binomial distributions,

$$P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = k, C, \boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_i)^{c_{i,k}} \times \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_0)^{1-c_{i,k}} \right\} \quad (3)$$

where I_j is the identity of the specific clone cell j is assigned to, and $\boldsymbol{a_j}$ and $\boldsymbol{d_j}$ are the observed alternate and total read count vectors, respectively, for variants 1 to N in cell j. The parameter vector $\boldsymbol{\theta}$ is a set of the unknown binomial success parameters of binomial distributions for modelling the allelic

read counts as described above. Specifically, θ_0 denotes the binomial success rate for the alternative allele when $c_{i,k}=0$ (variant absent), thereby accounting for sequencing errors or errors in the clonal tree configuration, and $\boldsymbol{\theta}_1=\{\theta_1,\theta_2...,\theta_N\}$ denotes a vector of binomial parameters, one for each variant, for $c_{i,k}=1$. The latter binomial rates model the effect of allelic imbalance, which means the probability of observing alternate reads at frequencies that differ from 0.5 for true heterozygous sites (see **Supp. Methods** for details).

To capture the uncertainty in the binomial success probabilities, we introduce beta prior distributions on θ_0 and $\boldsymbol{\theta}_1$. To ensure sensible prior distributions, we estimate the beta parameters from the scRNA-seq at known germline heterozygous variants for highly expressed genes (**Supp. Fig. S38**). For example, in the fibroblast dataset considered here, this approach yielded prior parameters of beta (0.2, 99.8) for θ_0 and beta (0.45, 0.55) for θ_i, $i>0$. The prior probability that cell j belongs to clone k is modelled using an uniform prior such that $P(I_j= k|\boldsymbol{\pi}) = \pi_k = 1/K$ for all k.

The joint posterior probability of clonal tree configuration *C*, cell assignment *I* and the parameters $\boldsymbol{\theta}$ and ξ can be described as follows.

$$P(C, I, \xi, \boldsymbol{\theta}|A, D) \propto P(A, D|I, C, \boldsymbol{\theta})P(C|\Omega, \xi)P(\xi|\kappa)P(I|\boldsymbol{\pi})P(\boldsymbol{\theta}|\nu_1, \nu_0) \quad (4)$$

We use a Gibbs sampler to infer this posterior distribution, and the details of the algorithm can be found in **Supp. Methods**, where we also present two alternative versions of cardelino: cardelino-free without any informative clone configuration prior and cardelino-fixed assuming that the clone configuration prior is fixed and error-free (see **Supp. Methods** and **Supp. Fig. S3**). Despite the full Bayesian approach, cardelino is computationally efficient, enabling the assignment of hundreds of cells within minutes using a single compute node. These methods will comfortably scale to datasets with many thousands of cells.

Alternative methods

Different from cardelino, two alternative methods with distinct strategies are compared: Demuxlet which assumes the guide clonal tree is perfect (Kang et al. 2018), and SingleCellGenotyper (SCG) which does not take any guide clonal tree (Roth et al. 2016).

Demuxlet requires a BAM file as input to obtain an empirical sequencing error rate from the sequencing quality score, which is not compatible with our simulated allelic read count matrices. Therefore, we re-implemented the core model of Demuxlet by following the third equation in the online method and the Supplementary Table S3 in its original paper (Kang et al. 2018). We set the sequencing error rate to 0.003 for all reads, by matching our simulation settings. We also compared our implementation and the original implementation on demultiplexing pooled scRNA-seq data, and found they are perfectly concordant (**Supp. Fig. S1**).

For SCG, the input is a matrix of categorical values denoting the measured genotype states for each variant in each cell. Here, our raw observation is the alternative and reference allelic read counts, hence we need to transform the observed raw counts into genotype states. As the false positive rate is mostly very low (i.e., observing an alternative allelic read from homozygous reference genotype), we simply take the genotype g_{ij} for variant i in cell j as 1 (i..e, heterozygous) if there is any alternative allelic read (i.e., a_{ij}>0), otherwise we take g_{ij}=0 (i.e., homozygous reference). In case there is no expression, we give a missing value g_{ij}=3. For running SCG, we used the run_singlet_model mode and configured the hyper-parameters as follows: kappa_prior=1, gamma_prior=[[30, 0.3], [4, 4]], and state_prior=[1, 1], which match our simulation settings. Note, we ran SCG from a Python wrap function in order to fix the first clone as a base clone, i.e., with no mutations.

Additionally, we included two variants of cardelino with similar strategies to SCG and Demuxlet: cardelino-fixed, similar to Demuxlet, assumes the guide clonal tree is perfect and cardelino-free, similar to SCG, does not use any guide clonal tree. The implementation of these two cardelino variants are described in the Supp. Methods. These five methods are compared with simulations in different settings (**Supp. Fig. S3** and **S13**).

The inferred clone labels may not be best aligned to the simulated clones, especially for SCG and cardelino-free that do not use any guide clone configuration, hence before evaluation we aligned the inferred clones to the simulated truth (or the input guide clones) by re-ordering the inferred clones to reach the lowest number of conflicting mutation states between two configuration matrices.

Cell culture

Dermal fibroblasts, derived from skin-punch samples from the shoulder of 32 donors (White British, age range 30-75), were obtained from the HipSci project (http://hipsci.org). Following thawing, fibroblasts were cultured in supplemented DMEM (high glucose, pyruvate, GlutaMAX (Life Technologies / 10569-010), with 10% FBS (Lab Tech / FB-1001) and 1% penicillin-streptomycin (Life Technologies / 15140122) added. 18 hours prior to collection, cells were trypsinised (Life Technologies / 25300054), counted, and seeded at a density of 100,000 cells per well (6 well plate).

Cell pooling, capture and full-length transcript single-cell RNA sequencing

Cells were washed with PBS, trypsinised, and resuspended in PBS (Gibco / 14190-144) + 0.1% DAPI (AppliChem / A1001). Cells from three lines were pooled and consequently sorted on a Becton Dickinson INFLUX machine into plates containing 2uL/well lysis buffer. Single cells were sorted individually (using FSC-W vs FSC-H), and apoptotic cells were excluded using DAPI. Cells from each three-plex cell pool were sorted across four 96-well plates. Reverse transcription and cDNA amplification was performed according to the Smart-seq2 protocol (Picelli et al. 2014), and library

preparation was performed using an Illumina Nextera kit. Samples were sequenced using paired-end 75bp reads on an Illumina HiSeq 2500 machine.

Bulk whole-exome sequencing data and somatic variant calling

We obtained bulk whole-exome sequencing data from HipSci fibroblast (median read coverage: 254) and derived iPS cell lines (median read coverage: 79) released by the HipSci project (Streeter et al. 2016; Kilpinen et al. 2017). Sequenced reads were aligned to the GRCh37 build of the human reference genome (Church et al. 2011) using *bwa* (Li 2013). To identify single-nucleotide somatic variant sites in the fibroblast lines, we compared variant allele frequencies for putative somatic variants in the fibroblast and matching iPS samples, using the iPS line as the reference "normal" sample in the absence of true germline samples for these lines. As the iPS lines were derived from their matching fibroblast lines, this comparison flips the usual tumour-normal comparison exploited in standard somatic mutation calling pipelines. As such, somatic variants present in a fibroblast sample are also expected to be present in the matching iPS sample, violating key assumptions of established somatic variant callers such as MuTect2 (Cibulskis et al. 2013) and Strelka2 (S. Kim et al. 2018). Thus, we apply a variant calling approach specific to our experimental setting here.

For each exome sample, we searched for sites with a non-reference base in the read pileup using *bcftools/mpileup* (Li et al. 2009). In the initial pre-filtering we retained sites with a per-sample coverage of at least 20 reads, at least three alternate reads in either fibroblast or iPS samples and an allele frequency less than 5% in the ExAC browser (Karczewski et al. 2017) and 1000 Genomes data (The 1000 Genomes Project Consortium 2015). A Fisher exact test (Fisher 1922) implemented in *bcftools/ad-bias* was then used to identify sites with significantly different variant allele frequency (VAF) in the exome data between fibroblast and iPS samples for a given line (Benjamini-Hochberg FDR < 10%). Sites were removed if any of the following conditions held: VAF < 1% or VAF > 45% in high-coverage fibroblast exome data; fewer than two reads supporting the alternative allele in the fibroblast sample; VAF > 80% in iPS data (to filter out potential homozygous alternative mutations); neither the iPS VAF or fibroblast VAF was below 45% (to filter out variants with a "significant" difference in VAF but are more likely to be germline than somatic variants). We further filtered sites to require uniqueness of sites across donors as it is highly unlikely to observe the same point mutation in more than one individual, so such sites almost certainly represent technical artefacts. Overall, this somatic variant calling approach aims to achieve higher specificity at the cost of lower sensitivity, so is conservative and should limit the inclusion of false-positive somatic variants in our callset.

We used bcftools/cnv (Danecek et al., 2016) to call copy number aberrations in fibroblasts. Calls were filtered to exclude CNAs with quality score <2, deletions with <10 markers and duplications with <10 heterozygous markers. We also excluded any calls that were smaller than 200Kb.

Identification of mutational signatures

Signature exposures were estimated using the *sigfit* package (Gori and Baez-Ortega 2018), providing the COSMIC 30 signatures as reference (Forbes et al. 2017), and with a highest posterior density (HPD) threshold of 0.9. Signatures were determined to be significant when the HPD did not overlap zero. Two signatures (7 and 11) were significant in two or more donors.

Identification of selection dynamics

Several methods have been developed to detect deviations from neutral growth in cell populations (Simons 2016a; Williams et al. 2016, 2018; Martincorena et al. 2018). Methods such as dN/dS or models assessing the fit of neutral models to the data need a high number of mutations to determine selection/neutrality. Given the relatively low number of mutations found in the donors in this study, these models are not applicable. We used the package *SubClonalSelection* (https://github.com/marcjwilliams1/SubClonalSelection.jl) in *Julia 0.6.2* which works with a low number of mutations (> 100 mutations; Williams et al. 2018). The package simulates the fit of a neutral and a selection model to the allele frequency distribution, and returns a probability for the selection model to fit the data best.

At small allele frequencies the resolution of the allele frequency distribution is limited by the sequencing depth. We chose a conservative lower resolution limit of $f_{min} = 0.05$ (Shin et al. 2017). At the upper end of the allele frequency distribution we chose a cut-off at $f_{max} = 0.45$ to account for ploidy ($= 2$). For the classification of the donors, we introduced cut-offs on the resulting selection probability of the algorithm. Donors with a selection probability below 0.3 are classified as 'neutral', above 0.7 as 'selected'. Donors which are neither 'selected' nor 'neutral' remain 'undetermined'. See **Fig. 2a** and **Supp. Fig. S6** for the results of the classification and fit of the models to the data. *subClonalSelection* assumes that the total population of cells is expanding exponentially and unfortunately does not allow to check for alternative growth hypotheses. However, we expect the growth dynamics not to have a big impact on the VAF distributions (in the extreme case of a constant population the VAF decay dynamics change to 1/f from 1/f^2 but still show peaks for selected clones; compare Figure 1 in Williams et al. 2018). Hence, the comparison of the selection model versus the neutral model should lead to meaningful results.

Single-cell gene expression quantification and quality control

Raw scRNA-seq data in CRAM format was converted to FASTQ format with *samtools* (v1.5), before reads were adapter- and quality-trimmed with *TrimGalore!* (github.com/FelixKrueger/TrimGalore) (Martin 2011). We quantified transcript-level expression using Ensembl v75 transcripts (Flicek et al. 2014) by supplying trimmed reads to *Salmon* v0.8.2 and using the "--seqBias", "--gcBias" and "VBOpt" options (Patro et al. 2017). Transcript-level expression values were summarised at gene level (estimated counts) and quality control of scRNA-seq data was done with the *scater* package

(McCarthy et al. 2017) and normalisation with the *scran* package (Lun, Bach, and Marioni 2016; Lun, McCarthy, and Marioni 2016). Cells were retained for downstream analyses if they had at least 50,000 counts from endogenous genes, at least 5,000 genes with non-zero expression, less than 90% of counts from the 100 most-expressed genes in the cell, less than 20% of counts from ERCC spike-in sequences and a *Salmon* mapping rate of at least 40% (**Supp. Table S2**). This filtering approach retains 63.7% of assayed cells.

Deconvolution of donors from pools

To increase experimental throughput in processing cells from multiple distinct donor individuals (*i.e.* lines), and to ensure an experimental design robust to batch effects, we pooled cells from three lines in each processing batch, as described above. As such, we do not know the donor identity of each cell at the time of sequencing and cell-donor identity must be inferred computationally. Thus, for both donor and, later, clone identity inference it is necessary to obtain the count of reads supporting the reference and alternative allele at informative germline and somatic variant sites. Trimmed FASTQ reads (described above) were aligned to the GRCh37 p13 genome with ERCC spike-in sequences with STAR in basic two-pass mode (Dobin et al. 2012) using the GENCODE v19 annotation with ERCC spike-in sequences (Searle et al. 2010). We further use *picard* (Broad Institute 2015) and *GATK* version 3.8 (McKenna et al. 2010) to mark duplicate reads (*MarkDuplicates*), split cigar reads (*SplitNCigarReads*), realign indels (*IndelRealigner*), and recalibrate base scores (*BaseRecalibrator*).

For cell-donor assignment we used the *GATK HaplotypeCaller* to call variants from the processed single-cell BAM files at 304,405 biallelic SNP sites from dbSNP (Sherry et al. 2001) build 138 that are genotyped on the Illumina HumanCoreExome-12 chip, have MAF > 0.01, Hardy-Weinberg equilibrium P < 1e-03 and overlap protein-coding regions of the 1,000 most highly expressed genes in HipSci iPS cells (as determined from HipSci bulk RNA-seq data). We merged the per-cell VCF output from *GATK HaplotypeCaller* across all cells using *bcftools* version 1.7 (Danecek et al. 2011, 2016) and filtered variants to retain those with MAF > 0.01, quality score > 20 and read coverage in at least 3% of cells. We further filtered the variants to retain only those that featured in the set of variants in the high-quality, imputed, phased HipSci genotypes and filtered the HipSci donor genotype file to include the same set of variants.

We used the *donor_id* function in the cardelino package to assign cells to donors. This function assigns cells to donors by modelling alternative allele read counts with given genotypes of input donors. For a single germline variant, the three base genotypes (as minor allele counts) can be 0, 1 and 2. For doublet genotype profiles generated by combining pairs of donor genotypes, two additional combinatory genotypes, 0.5 and 1.5 are allowed. We assume that each genotype has a unique binomial distribution whose parameters are estimated by an EM algorithm in a framework similar to clone assignment (described above; see **Supp. Methods**). When we enable doublet detection, the

posterior probabilities that a cell comes from any of the donors provided, including doublet donors, are calculated for donor assignment. There are 490 available HipSci donors, so we run cardelino in two passes on each plate of scRNA-seq data separately. In the first pass, the model outputs the posterior probability that each cell belongs to one of the 490 HipSci donors, ignoring the possibility of doublets. In the second pass, only those donors with a posterior probability greater than 0.95 in at least one cell are considered by the model as possible donors and doublet detection is enabled. After the second pass, if the highest posterior probability is greater than 0.95, more than 25 variants have read coverage, and the doublet probability is less than 5% then we provisionally assign the cell to the donor with the highest posterior probability. If the provisionally assigned donor is one of the three donors known to have been pooled together for the specific plate, then we deem the cell to be confidently assigned to that donor, otherwise we deem the cell to have "unassigned" donor. With this approach, 97.4% of cells passing QC (see above) are confidently assigned to a donor (**Supp. Fig. S39**). Of the cells that are not confidently assigned to a donor, 2.1% are identified as doublets by cardelino and 0.5% remain "unassigned" due to low variant coverage or low posterior probability. Thus, we have 2,338 QC-passing, donor-assigned cells for clonal analysis.

Clonal inference

We inferred the clonal structure of the fibroblast cell population for each of the 32 lines (donors) using Canopy (Jiang et al. 2016). We used read counts for the variant allele and total read counts at filtered somatic variant sites from high-coverage whole-exome sequencing data from the fibroblast samples as input to Canopy. In addition to the variant filtering described above, input sites were further filtered for tree inference to those that had non-zero read coverage in at least one cell assigned to the corresponding line. We used the BIC model selection method in Canopy to choose the optimal number of clones per line. Here, for each of the 32 lines, we considered the highest-likelihood clonal tree produced by Canopy, along with the estimated prevalence of each clone and the set of somatic variants tagging each clone as the given clonal tree for cell-clone assignment.

Cell-clone assignment

For cell-clone assignment we required the read counts supporting reference and alternative alleles at somatic variant sites. We used the *bcftools* version 1.7 *mpileup* and *call* methods to call variants at somatic variant sites derived from bulk whole-exome data, as described above, for all confidently assigned cells for each given line. Variant sites were filtered to retain variants with more than three reads observed across all cells for the line and quality greater than 20. We retained cells with at least two somatic variants with non-zero read coverage (2,044 cells across 32 lines). From the filtered VCF output of *bcftools* we obtained the number of reads supporting the alternative allele and the total read coverage for each somatic variant site with more than three reads covering the site, in total, across all the line's cells. In general, read coverage of somatic variant sites in scRNA-seq data is sparse, with over 80% of sites for a given cell having no overlapping reads. We used the scRNA-seq read counts at

the line's somatic variant sites to assign QC-passing cells from the line to clones using the *clone_id* function in the cardelino R package.

Simulations to benchmark cell to clone assignment

We simulated data to test the performance of cardelino as follows. First, given a clonal tree configuration C (N-by-K binary matrix), a given number of cells are generated (e.g. 200, see below), whose genotypes are sampled from K clones following a multinomial distribution parameterised by clonal fractions F. Second, given a matrix D (N-by-M matrix) of sequencing coverage for N sites in M cells, we uniformly sample the coverage profiles from these M cells into a given number of cells for simulation. Third, after having the genotype $h_{ij}=c_{i,lj}$ and the sequencing depth d_{ij} for variant i in cell j from the previous two steps, we can generate the read count a_{ij} for the alternative allele by sampling from a binomial distribution with success parameter θ_0 if h_{ij}=0 or with an allele-specific expression parameter θ_i if h_{ij}=1. Note, both θ_0 and θ_i are randomly generated from beta prior distributions, whose parameters are estimated from experimental data.

Based on the above simulation workflow, two simulation experiments are performed to evaluate the accuracy and robustness of cardelino. One simulation was performed with synthesizing the same number of cells as seen for each of the 32 lines, where input parameters are from the observed matrices C and D, clonal fraction F, and cardelino-learned $\boldsymbol{\theta}$ from each line. To match the error rate in the guide clone configuration as observed in experimental data, we swapped the same fraction of mutation states for non-base clones in the guide configuration matrix *C* when running cardelino. We repeat the simulation 50 times on each line, permuting the position of the errors in the tree configuration. This simulation tries to mimic all settings in each line, which not only evaluates the accuracy of the model, but also reflects the quality of the data in each line for clonal assignment.

Additionally, in a second set of simulations, we change one of these parameters each time to systematically assess cardelino. The clonal configuration is defined by the number of clones, K, a perfect phylogenetic matrix $((K\text{-}1)\text{-by-}K)$ including a base clone, and the number of unique variants per clonal branch n, which returns a configuration matrix C with a shape of $n(K\text{-}1)\text{-by-}K$. With setting K clones, one out of all possible clonal tree structures is randomly selected to generate the clonal configuration matrix. Then the sequencing depth matrix D for these $n(K\text{-}1)$ variants are sampled from a line with 439 variants across 151 cells (see distribution in **Supp. Fig. S40**). In order to increase or decrease the missingness rate of D, zero coverages are respectively added or removed linearly according to the expression level of the gene corresponding to the variant. The allelic expression balance can be adjusted by changing the parameters of its beta prior distribution. We set uniform clonal prevalence in the second simulation. With each parameter setting, 200 cells are randomly synthesized and this procedure is repeated 50 times to vary the random selection of errors in the tree configuration, the branch position of each variant, and the tree structure. When one setting parameter

varies, others are used at the default values: number of variants per clonal branch = 10, variant coverage = 0.25, clone number = 4, precision of allelic ratio = 1 (i.e. shape1+shape2 of beta prior, lower precision means more variants with high allelic imbalance), error rate of the mutation states in the input clone configuration = 0.1, and fraction of wrongly clustered variants = 0 (though this is coupled with the error rate). These default values are representative of the 32 experimental lines (**Supp. Fig. S11, S38, S41**).

Variance component, differential expression and pathway analysis

Expression analyses between clones required further filtering of cells for each line. Analyses were conducted using cells that passed the following filtering procedure for each line: (1) clones identified in the line were retained if at least three cells were confidently assigned to the clone; (2) cells were retained if they were confidently assigned to a retained clone. Lines were retained for DE testing if they had at least 15 cells assigned to retained clones, allowing us to conduct expression analyses for 31 out of the 32 lines (all except *vils*).

Expression variance across cells is decomposed into multiple components in a linear mixed model, including cellular detection rate (proportion of genes with non-zero expression per cell) as a fixed effect and plate (i.e. experimental batch), donor (i.e. line; only when combining cells across all donors) and clone (nested within donor for combined-donor analysis) as random effects. We fit the linear mixed model on a per-gene basis using the *variancePartition* R package (Hoffman and Schadt 2016).

Differential gene expression (DE) testing was conducted using the quasi-likelihood F-test method (Lund et al. 2012) in the *edgeR* package (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012) as recommended by Soneson and Robinson (Soneson and Robinson 2018). To test for differences in expression between cells assigned to different clones in a line, we fit a linear model for single-cell gene expression with cellular detection rate (proportion of QC-filtered genes expressed in a cell; numeric value), plate on which the cell was processed (a factor) and assigned clone (a factor) as predictor variables. The quasi-likelihood F test was used to identify genes with: (1) any difference in average expression level between clones (analogous to analysis of variance), and (2) differences in average expression between all pairs of clones (“pairwise contrasts”). We considered 10,876 genes that were sufficiently expressed (an average count >1 across cells in all lines) to test for differential expression.

To test for significance of overlap of DE genes across donors, we sampled sets of genes without replacement the same size as the number of DE genes (FDR < 10%) for each line. For each permutation set, we then computed the number of sampled genes shared between between donors. We repeated this procedure 1,000 times to obtain distributions for the number of DE genes shared by multiple donors if shared genes were obtained purely by chance.

Gene set enrichment (pathway) analyses were conducted using the *camera* (Wu and Smyth 2012) method in the *limma* package (Smyth 2004; Ritchie et al. 2015). Using $\log_2$-fold-change test statistics for 10,876 genes for pairwise contrasts between clones from the *edgeR* models above as input, we applied *camera* to test for enrichment for the 50 Hallmark gene sets from MSigDB, the Molecular Signatures Database (Liberzon et al. 2011). For all differential expression and pathway analyses we adjusted for multiple testing by estimating the false discovery rate (FDR) using independent hypothesis weighting (Ignatiadis et al. 2016), as implemented in the *IHW* package, with average gene expression supplied as the independent covariate.

Code availability

The cardelino methods are implemented in an open-source, publicly available R package (github.com/PMBio/cardelino). The code used to process and analyse the data is available (github.com/davismcc/fibroblast-clonality), with a reproducible workflow implemented in Snakemake (Köster and Rahmann 2012). Descriptions of how to reproduce the data processing and analysis workflows, with html output showing code and figures presented in this paper, are available at davismcc.github.io/fibroblast-clonality. Docker images providing the computing environment and software used for data processing (hub.docker.com/r/davismcc/fibroblast-clonality/) and data analyses in R (hub.docker.com/r/davismcc/r-singlecell-img/) are publicly available.

Data availability

Single-cell RNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7167. Whole-exome sequencing data is available through the HipSci portal (www.hipsci.org). Metadata, processed data and large results files are available under the DOI 10.5281/zenodo.1403510 (doi.org/10.5281/zenodo.1403510).

Author contributions

R.R., T.H. and S.A.T. conceived and planned the experiments. R.R. and T.H. carried out the experiments. Y.H., D.J.M. and O.S. developed the computational methods. Y.H. developed the statistical model and the implementation. Y.H. and D.J.M. wrote the software. Y.H. carried out all simulation experiments and benchmarked alternative methods. The HipSci Consortium provided the cell lines and exome sequencing data. P.D. conducted somatic variant calling from exome sequencing data. D.J.G. advised on somatic variant calling approaches and the mutational signatures analysis carried out by R.R. D.J.M. and M.J.B. developed data processing workflows and D.J.M. processed the single-cell RNA-sequencing data. D.J.K. conducted the selection analyses, supervised by B.D.S. D.J.M. and Y.H. carried out clonal inference and cell assignment analyses. D.J.M. conducted differential gene and pathway expression analyses and integrated the computational analyses into a reproducible workflow. D.J.M. and R.R. took the lead in writing the manuscript. D.J.M., R.R. and Y.H.

drafted the manuscript and designed the figures. W.W. suggested improvements to somatic variant calling and differential expression analyses. S.A.T. and O.S. conceived of the study, planned and supervised the work. All authors contributed to the interpretation of results and commented on and approved the final manuscript. The HipSci Consortium generated and provided early access to the fibroblast lines used in this work (see **Supp. Material** for a full list of consortium members).

Acknowledgements

We would like to thank David Jörg for highly constructive discussions. We would like to acknowledge the Wellcome Sanger Institute Cellular Genetics and Phenotyping teams (in particular, Alex Alderton, Celine Gomez, Rachel Boyd, Sharad Patel and Sam Barnett) and DNA pipelines for their invaluable assistance in generating the data for this study. We would like to thank Gerda Kildisiute for assisting in CNV analysis of the lines.

References

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* 500 (7463): 415–21.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. “Comprehensive Characterization of Cancer Driver Genes and Mutations.” *Cell* 173 (2): 371–85.e18.

Behjati, Sam, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C. Wedge, Asif U. Tamuri, Iñigo Martincorena, et al. 2014. “Genome Sequencing of Normal Cells Reveals Developmental Lineages and Mutational Processes.” *Nature* 513 (7518): 422–25.

Broad Institute. 2015. “Picard Tools.” Picard Tools - By Broad Institute. 2015. http://broadinstitute.github.io/picard/.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” *Nature* 523 (7561): 486–90.

Burnet, F. M. 1974. “Intrinsic Mutagenesis: A Genetic Basis of Ageing.” *Pathology* 6 (1): 1–11.

Campbell, Kieran R., Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Hossein Farahani, et al. 2019. “Clonealign: Statistical Integration of Independent Single-Cell RNA and DNA Sequencing Data from Human Cancers.” *Genome Biology* 20 (1): 54.

Cheow, Lih Feng, Elise T. Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S. W. Tan, et al. 2016. “Single-Cell Multimodal Profiling Reveals Cellular Epigenetic Heterogeneity.” *Nature Methods* 13 (10): 833–36.

Church, Deanna M., Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, et al. 2011. “Modernizing Reference Genome Assemblies.” *PLoS Biology* 9 (7): e1001091.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. 2013. “Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples.” *Nature Biotechnology* 31 (3): 213–19.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58.

Danecek, Petr, Shane A. McCarthy, HipSci Consortium, and Richard Durbin. 2016. “A Method for Checking Genomic Integrity in Cultured Cell Lines from SNP Genotyping Data.” *PloS One* 11 (5): e0155014.

Deshwar, Amit G., Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. 2015. “PhyloWGS: Reconstructing Subclonal Composition and Evolution from Whole-Genome Sequencing of Tumors.” *Genome Biology* 16 (1): 35.

Ding, Li, Matthew H. Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L. Gibbs, et al. 2018. “Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics.” *Cell* 173 (2): 305–20.e10.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2012. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* , October. https://doi.org/10.1093/bioinformatics/bts635.

Fan, Jean, Hae-Ock Lee, Soohyun Lee, Da-Eun Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, et al. 2018. “Linking Transcriptional and Genetic Tumor Heterogeneity through Allele Analysis of Single-Cell RNA-Seq Data.” *Genome Research*, June. https://doi.org/10.1101/gr.228080.117.

Fisher, Ronald A. 1922. “On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P.” *Journal of the Royal Statistical Society* 85 (1): 87–94.

Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, et al. 2014. “Ensembl 2014.” *Nucleic Acids Research* 42 (Database issue): D749–55.

Forbes, Simon A., David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, et al. 2017. “COSMIC: Somatic Cancer Genetics at High-Resolution.” *Nucleic Acids Research* 45 (D1): D777–83.

Giustacchini, Alice, Supat Thongjuea, Nikolaos Barkas, Petter S. Woll, Benjamin J. Povinelli, Christopher A. G. Booth, Paul Sopp, et al. 2017. “Single-Cell Transcriptomics Uncovers Distinct Molecular Signatures of Stem Cells in Chronic Myeloid Leukemia.” *Nature Medicine* 23 (6): 692–702.

Gori, Kevin, and Adrian Baez-Ortega. 2018. “Sigfit: Flexible Bayesian Inference of Mutational Signatures.” *bioRxiv*. https://doi.org/10.1101/372896.

Guo, Hongshan, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. 2013. “Single-Cell Methylome Landscapes of Mouse Embryonic Stem Cells and Early Embryos Analyzed Using Reduced Representation Bisulfite Sequencing.” *Genome Research* 23 (12): 2126–35.

Hodis, Eran, Ian R. Watson, Gregory V. Kryukov, Stefan T. Arold, Marcin Imielinski, Jean-Philippe Theurillat, Elizabeth Nickerson, et al. 2012. “A Landscape of Driver Mutations in Melanoma.” *Cell* 150 (2): 251–63.

Hoffman, Gabriel E., and Eric E. Schadt. 2016. “variancePartition: Interpreting Drivers of Variation in Complex Gene Expression Studies.” *BMC Bioinformatics* 17 (1): 483.

Huang, Kuan-Lin, R. Jay Mashl, Yige Wu, Deborah I. Ritter, Jiayin Wang, Clara Oh, Marta Paczkowska, et al. 2018. “Pathogenic Germline Variants in 10,389 Adult Cancers.” *Cell* 173 (2): 355–70.e14.

Ignatiadis, Nikolaos, Bernd Klaus, Judith B. Zaugg, and Wolfgang Huber. 2016. “Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing.” *Nature Methods* 13 (7): 577–80.

Jahn, Katharina, Jack Kuipers, and Niko Beerenwinkel. 2016. “Tree Inference for Single-Cell Data.” *Genome Biology* 17 (1): 86.

Jiang, Yuchao, Yu Qiu, Andy J. Minn, and Nancy R. Zhang. 2016. “Assessing Intratumor Heterogeneity and Tracking Longitudinal and Spatial Clonal Evolutionary History by next-Generation Sequencing.” *Proceedings of the National Academy of Sciences* 113 (37): E5528–37.

Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, et al. 2018. “Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation.” *Nature Biotechnology* 36 (1): 89–94.

Karczewski, Konrad J., Ben Weisburd, Brett Thomas, Matthew Solomonson, Douglas M. Ruderfer, David Kavanagh, Tymor Hamamsy, et al. 2017. “The ExAC Browser: Displaying Reference Data Information from over 60 000 Exomes.” *Nucleic Acids Research* 45 (D1): D840–45.

Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, et al. 2017. “Common Genetic Variation Drives Molecular Heterogeneity in Human iPSCs.” *Nature* 546 (7658): 370–75.

Kim, Kyung In, and Richard Simon. 2014. “Using Single Cell Sequencing Data to Model the Evolutionary History of a Tumor.” *BMC Bioinformatics* 15 (January): 27.

Kim, Sangtae, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, et al. 2018. “Strelka2: Fast and Accurate Calling of Germline and Somatic Variants.” *Nature Methods*, July, 1.

Köster, Johannes, and Sven Rahmann. 2012. “Snakemake--a Scalable Bioinformatics Workflow Engine.” *Bioinformatics* 28 (19): 2520–22.

Kuipers, Jack, Katharina Jahn, Benjamin J. Raphael, and Niko Beerenwinkel. 2017. “Single-Cell Sequencing Data Reveal Widespread Recurrence and Loss of Mutational Hits in the Life Histories of Tumors.” *Genome Research* 27 (11): 1885–94.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. “Molecular Signatures Database (MSigDB) 3.0.” *Bioinformatics* 27 (12): 1739–40.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Lun, Aaron T. L., Karsten Bach, and John C. Marioni. 2016. “Pooling across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts.” *Genome Biology* 17 (1): 75.

Lun, Aaron T. L., Davis J. McCarthy, and John C. Marioni. 2016. “A Step-by-Step Workflow for Low-Level Analysis of Single-Cell RNA-Seq Data.” *F1000Research* 5 (August). https://doi.org/10.12688/f1000research.9501.1.

Lund, Steven P., Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth. 2012. “Detecting Differential Expression in RNA-Sequence Data Using Quasi-Likelihood with Shrunken Dispersion Estimates.” *Statistical Applications in Genetics and Molecular Biology* 11 (5). https://doi.org/10.1515/1544-6115.1826.

Malikic, Salem, Katharina Jahn, Jack Kuipers, Cenk Sahinalp, and Niko Beerenwinkel. 2017. “Integrative Inference of Subclonal Tumour Evolution from Single-Cell and Bulk Sequencing Data.” *bioRxiv*. https://doi.org/10.1101/234914.

Martincorena, Iñigo, and Peter J. Campbell. 2015. “Somatic Mutation in Cancer and Normal Cells.” *Science* 349 (6255): 1483–89.

Martincorena, Iñigo, Philip H. Jones, and Peter J. Campbell. 2016. “Constrained Positive Selection on Cancer Mutations in Normal Skin.” *Proceedings of the National Academy of Sciences* 113 (9): E1128–29.

Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. 2018. “Universal Patterns of Selection in Cancer and Somatic Tissues.” *Cell* 173 (7): 1823.

Martincorena, Iñigo, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C. Wedge, et al. 2015. “High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin.” *Science* 348 (6237): 880.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.

McCarthy, Davis J., Kieran R. Campbell, Aaron T. L. Lun, and Quin F. Wills. 2017. “Scater: Pre-Processing, Quality Control, Normalization and Visualization of Single-Cell RNA-Seq Data in R.” *Bioinformatics* 33 (8): 1179–86.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. “Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation.” *Nucleic Acids Research* 40 (10): 4288–97.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303.

Müller, Sören, Siyuan John Liu, Elizabeth Di Lullo, Martina Malatesta, Alex A. Pollen, Tomasz J. Nowakowski, Gary Kohanbash, et al. 2016. “Single-cell Sequencing Maps Gene Expression to Mutational Phylogenies in PDGF- and EGF-driven Gliomas.” *Molecular Systems Biology* 12 (11): 889.

Navin, Nicholas E. 2015. “The First Five Years of Single-Cell Cancer Genomics and beyond.” *Genome Research* 25 (10): 1499–1507.

Navin, Nicholas E., and Ken Chen. 2016. “Genotyping Tumor Clones from Single-Cell Data.” *Nature Methods* 13 (7): 555–56.

Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, et al. 2011. “Tumour Evolution Inferred by Single-Cell Sequencing.” *Nature* 472 (7341): 90–94.

Nik-Zainal, Serena, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, et al. 2012. “Mutational Processes Molding the Genomes of 21 Breast Cancers.” *Cell* 149 (5): 979–93.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Picelli, Simone, Omid R. Faridani, Asa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. “Full-Length RNA-Seq from Single Cells Using Smart-seq2.” *Nature Protocols* 9 (1): 171–81.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47–e47.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.

Roth, Andrew, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. 2014. “PyClone: Statistical Inference of Clonal Population Structure in Cancer.” *Nature Methods* 11 (March): 396.

Roth, Andrew, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A. Smith, et al. 2016. “Clonal Genotype and Population Structure Inference from Single-Cell Tumor Sequencing.” *Nature Methods* 13 (7): 573–76.

Saikia, Mridusmita, Philip Burnham, Sara H. Keshavjee, Michael F. Z. Wang, Michael Heyang, Pablo Moral-Lopez, Meleana M. Hinchman, Charles G. Danko, John S. L. Parker, and Iwijn De Vlaminck. 2019. “Simultaneous Multiplexed Amplicon Sequencing and Transcriptome Profiling in Single Cells.” *Nature Methods* 16 (1): 59–62.

Salehi, Sohrab, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. 2017. “ddClone: Joint Statistical Inference of Clonal Populations from Single Cell and Bulk Tumour Sequencing Data.” *Genome Biology* 18 (1): 44.

Searle, S., A. Frankish, A. Bignell, B. Aken, T. Derrien, M. Diekhans, R. Harte, et al. 2010. “The GENCODE Human Gene Set.” *Genome Biology* 11. https://doi.org/10.1186/gb-2010-11-s1-p36.

Sherry, S., M. Ward, M. Kholodov, J. Baker, L. Phan, E. Smigielski, and K. Sirotkin. 2001. “dbSNP: The NCBI Database of Genetic Variation.” *Nucleic Acids Research* 29 (1): 308–11.

Shin, Hyun-Tae, Yoon-La Choi, Jae Won Yun, Nayoung K. D. Kim, Sook-Young Kim, Hyo Jeong Jeon, Jae-Yong Nam, et al. 2017. “Prevalence and Detection of Low-Allele-Fraction Variants in Clinical Cancer Samples.” *Nature Communications* 8 (1): 1377.

Simons, Benjamin D. 2016a. “Deep Sequencing as a Probe of Normal Stem Cell Fate and Preneoplasia in Human Epidermis.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (1): 128–33.

———. 2016b. “Reply to Martincorena et Al.: Evidence for Constrained Positive Selection of Cancer Mutations in Normal Skin Is Lacking.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (9): E1130–31.

Smallwood, Sébastien A., Heather J. Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R. Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. 2014. “Single-Cell Genome-Wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity.” *Nature Methods* 11 (8): 817–20.

Smyth, Gordon K. 2004. “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.” *Statistical Applications in Genetics and Molecular Biology* 3 (February): Article3.

Soneson, Charlotte, and Mark D. Robinson. 2018. “Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis.” *Nature Methods*, February. https://doi.org/10.1038/nmeth.4612.

Stransky, Nicolas, Ann Marie Egloff, Aaron D. Tward, Aleksandar D. Kostic, Kristian Cibulskis, Andrey Sivachenko, Gregory V. Kryukov, et al. 2011. “The Mutational Landscape of Head and Neck Squamous Cell Carcinoma.” *Science* 333 (6046): 1157–60.

Streeter, Ian, Peter W. Harrison, Adam Faulconbridge, The HipSci Consortium, Paul Flicek, Helen Parkinson, and Laura Clarke. 2016. “The Human-Induced Pluripotent Stem Cell Initiative—data Resources for Cellular Genetics.” *Nucleic Acids Research*, October. https://doi.org/10.1093/nar/gkw928.

The 1000 Genomes Project Consortium. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Tirosh, Itay, Andrew S. Venteicher, Christine Hebert, Leah E. Escalante, Anoop P. Patel, Keren Yizhak, Jonathan M. Fisher, et al. 2016. “Single-Cell RNA-Seq Supports a Developmental Hierarchy in Human Oligodendroglioma.” *Nature* 539 (7628): 309–13.

Wang, Yong, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, et al. 2014. “Clonal Evolution in Breast Cancer Revealed by Single Nucleus Genome Sequencing.” *Nature* 512 (7513): 155–60.

Williams, Marc J., Benjamin Werner, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva. 2016. “Identification of Neutral Tumor Evolution across Cancer Types.” *Nature*

Genetics 48 (January): 238.
Williams, Marc J., Benjamin Werner, Timon Heide, Christina Curtis, Chris P. Barnes, Andrea Sottoriva, and Trevor A. Graham. 2018. “Quantification of Subclonal Selection in Cancer from Bulk Sequencing Data.” *Nature Genetics* 50 (6): 895–903.
Wu, Di, and Gordon K. Smyth. 2012. “Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation.” *Nucleic Acids Research* 40 (17): e133.

Supplementary Material

Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants

Davis J. McCarthy[1,4,10*], Raghd Rostom[1,2,*], Yuanhua Huang[1,*], Daniel J. Kunz[2,5,6], Petr Danecek[2], Marc Jan Bonder[1], Tzachi Hagai[1,2], HipSci Consortium, Wenyi Wang[8], Daniel J. Gaffney[2], Benjamin D. Simons[5,6,7], Oliver Stegle[1,3,9,#], Sarah A. Teichmann[1,2,5,#]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Hinxton, Cambridge, UK; [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK; [3]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany; [4]St Vincent's Institute of Medical Research, Fitzroy, Victoria 3065, Australia. [5]Cavendish Laboratory, Department of Physics, JJ Thomson Avenue, Cambridge, CB3 0HE, UK. [6]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, CB2 1QN, UK. [7]The Wellcome Trust/Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge, UK. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [9]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany; [10]Melbourne Integrative Genomics, School of Mathematics and Statistics/School of Biosciences, University of Melbourne, Parkville, 3010, Australia.

** These authors contributed equally to this work.*

 # Corresponding authors.

ORCIDs:

- DJM: 0000-0002-2218-6833
- RR: 0000-0002-4453-3357
- YH: 0000-0003-3124-9186
- DJK: 0000-0003-3597-6591
- PD:
- MJB: 0000-0002-8431-3180
- TH:
- WW: 0000-0003-0617-9438
- DJG: 0000-0002-1529-1862
- BDS: 0000-0002-3875-7071
- OS: 0000-0002-8818-7193
- ST: 0000-0002-6294-6366

HipSci consortium members

Helena Kilpinen[2,8], Angela Goncalves[2], Andreas Leha[2,10], Vackar Afzal[3], Kaur Alasoo[2], Sofie Ashford[4], Sendu Bala[2], Dalila Bensaddek[3], Marc Jan Bonder[1], Francesco Paolo Casale[1], Oliver J Culley[5], Anna Cuomo[1], Petr Danecek[2], Adam Faulconbridge[1], Peter W Harrison[1], Annie Kathuria[5] , Davis J McCarthy[1,9], Shane A McCarthy[2], Ruta Meleckyte[5], Yasin Memari[2], Bogdan Mirauta[1], Nathalie Moens[5], Filipa Soares[6], Alice Mann[2], Daniel Seaton[1], Ian Streeter[1], Chukwuma A Agu[2], Alex Alderton[2], Rachel Nelson[2], Sarah Harper[2], Minal Patel[2], Alistair White[2], Sharad R Patel[2], Laura Clarke[1], Reena Halai[2], Christopher M Kirton[2], Anja KolbKokocinski[2], Philip Beales[8], Ewan Birney[1], Davide Danovi[5], Angus I Lamond[3], Willem H Ouwehand[2,4,7], Ludovic Vallier[2,6], Fiona M Watt[5], Richard Durbin[2,11], Oliver Stegle[1,12,13] , Daniel J Gaffney[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.
[2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom.
[3]Centre for Gene Regulation & Expression, School of Life Sciences, University of Dundee, DD1 5EH, United Kingdom.
[4]Department of Haematology, University of Cambridge, Cambridge, United Kingdom.
[5]Centre for Stem Cells & Regenerative Medicine, King's College London, Tower Wing, Guy's Hospital, Great Maze Pond, London SE1 9RT, United Kingdom.
[6]Wellcome Trust and MRC Cambridge Stem Cell Institute and Biomedical Research Centre, Anne McLaren Laboratory, University of Cambridge, CB2 0SZ, United Kingdom.
[7]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom.
[8]UCL Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, United Kingdom.
[9]St Vincent's Institute of Medical Research, Fitzroy Victoria 3065, Australia.
[10]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
[11]Department of Genetics, University of Cambridge, Cambridge, United Kingdom.
[12]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany.
[13]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120, Heidelberg, Germany.

Line name	Gender	Age	Number of Variants	Signature 7 Mean Exposure	Number Clones With Cells	Minimum Hamming Distance	Total cells	Assigned Cells	Proportion Assigned Cells
euts	male	60-64	292	0.585	3	29	79	78	0.987
fawm	female	70-74	101	0.337	3	5	53	47	0.887
feec	male	60-64	170	0.281	4	5	75	64	0.853
fikt	male	50-54	142	0.378	3	13	39	36	0.923
garx	female	50-54	592	0.670	3	57	70	69	0.986
gesg	male	60-64	157	0.372	3	23	105	101	0.962
heja	male	70-74	192	0.266	3	16	50	50	1.000
hipn	male	55-59	59	0.019	3	8	62	49	0.790
ieki	female	55-59	82	0.381	3	7	58	26	0.448
joxm	female	45-49	612	0.609	3	41	79	77	0.975
kuco	female	65-69	41	0.112	2	9	48	48	1.000
laey	female	70-74	278	0.532	3	36	55	55	1.000
lexy	female	60-64	55	0.069	3	6	63	63	1.000
naju	male	60-64	85	0.296	2	13	44	44	1.000
nusw	male	65-69	62	0.091	3	3	60	20	0.333
oaaz	male	70-74	90	0.172	3	17	38	37	0.974
oilg	male	65-69	211	0.505	3	2	90	57	0.633
pipw	male	50-54	233	0.551	3	34	107	107	1.000
puie	male	60-64	117	0.448	3	10	41	41	1.000
qayj	female	60-64	46	0.035	3	7	97	59	0.608
qolg	male	35-39	120	0.381	2	23	36	36	1.000
qonc	female	65-69	131	0.406	3	7	58	43	0.741
rozh	female	65-69	79	0.173	4	2	91	42	0.462
sehl	female	55-59	178	0.527	4	2	30	24	0.800
ualf	female	55-59	325	0.540	3	29	89	88	0.989
vass	female	30-34	412	0.647	3	35	37	37	1.000
vils	female	35-39	51	0.206	4	1	37	4	0.108
vuna	female	65-69	135	0.456	2	33	71	71	1.000
wahn	female	65-69	496	0.605	3	52	82	77	0.939
wetu	female	55-59	73	0.212	3	8	77	66	0.857
xugn	male	65-69	124	0.398	3	8	35	34	0.971
zoxy	female	60-64	61	0.117	3	8	88	82	0.932

Table S1: Biological and technical metadata for each of the 32 HipSci human fibroblast lines used. Number of variants refers to somatic variants identified from whole-exome sequencing data (**Methods**); Signature 7 exposure refers to Signature 7 (UV) from the COSMIC set of mutational signatures; Minimum Hamming distance denotes the minimum number of variants distinguishing between two clones in the inferred clonal tree for the line (**Methods**).

	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	% passing filter
Before QC filtering	Total counts from endog. genes	178	123,489	383,929	442,353	621,738	5,833,292	-
	Total genes expressed	174	6,772	10,446	8,801	11,790	16,243	-
	% counts from ERCCs	0	0.97	1.81	14.47	3.34	99.90	-
	% counts top 100 expressed genes	29.4	40.8	55.6	57.8	62.8	100.0	-
	% reads mapped	7.69	68.71	75.59	74.80	81.67	100.0	-
After QC filtering	Total counts from endog. genes	50,464	316,033	484,887	559,742	710,028	2,659,889	80.6
	Total genes expressed	5,083	9,960	11,108	10,846	12,100	14,804	79.3
	% counts from ERCCs	0.001	0.96	1.63	1.86	2.39	18.1	85.3
	% counts top 100 expressed genes	29.4	38.6	52.4	49.2	58.2	89.0	86.1
	% reads mapped	44.1	70.3	76.0	74.8	79.1	92.7	99.3

Table S2: Summaries of QC metrics for single-cell RNA-seq data before and after QC filtering. Cells were required to have more than 50,000 counts from endogenous genes, more than 5,000 genes expressed (*i.e*. with non-zero expression), less than 20% of counts from ERCC transcripts, less than 90% of counts from the 100 most-expressed genes in the cell and at least 40% of reads mapped using *Salmon*. Metrics were computed using the *scater* package (**Methods**).

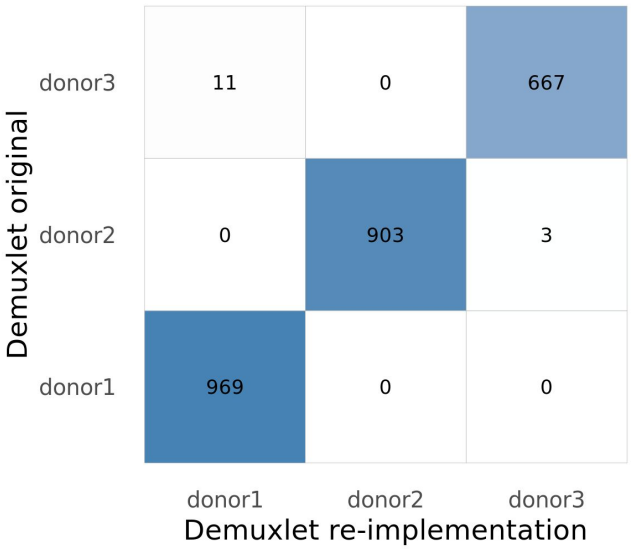


Figure S1: Comparison of donor assignment results from the original Demuxlet software and our implementation. The confusion matrix of cells assigned to three donors by two methods, which are highly concordant. Note, those unmatched cells are all identified as doublets by Demuxlet. The data is generated by 10x genomics platform by pooling three HipSci lines.

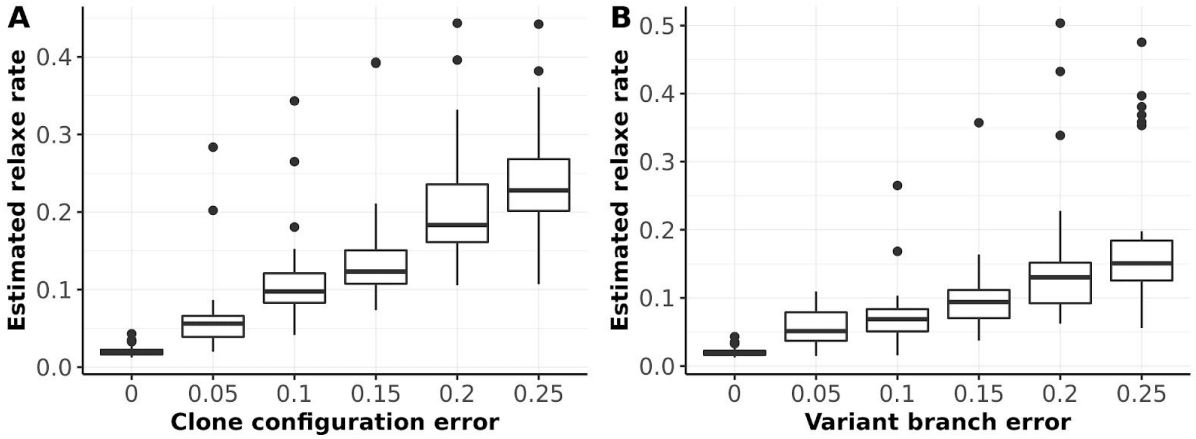


Figure S2: Evaluation of the inferred relax (error) rate using simulations. (A) The estimated relax rate as a function of the simulated error rates. Errors are simulated by uniformly swapping the mutation states in the guide clonal configuration matrix, except the base clone which has no mutations. (B) The estimated relax rate across different fractions of variants that have wrong branch configuration. Errors are added by swapping branches for variants.

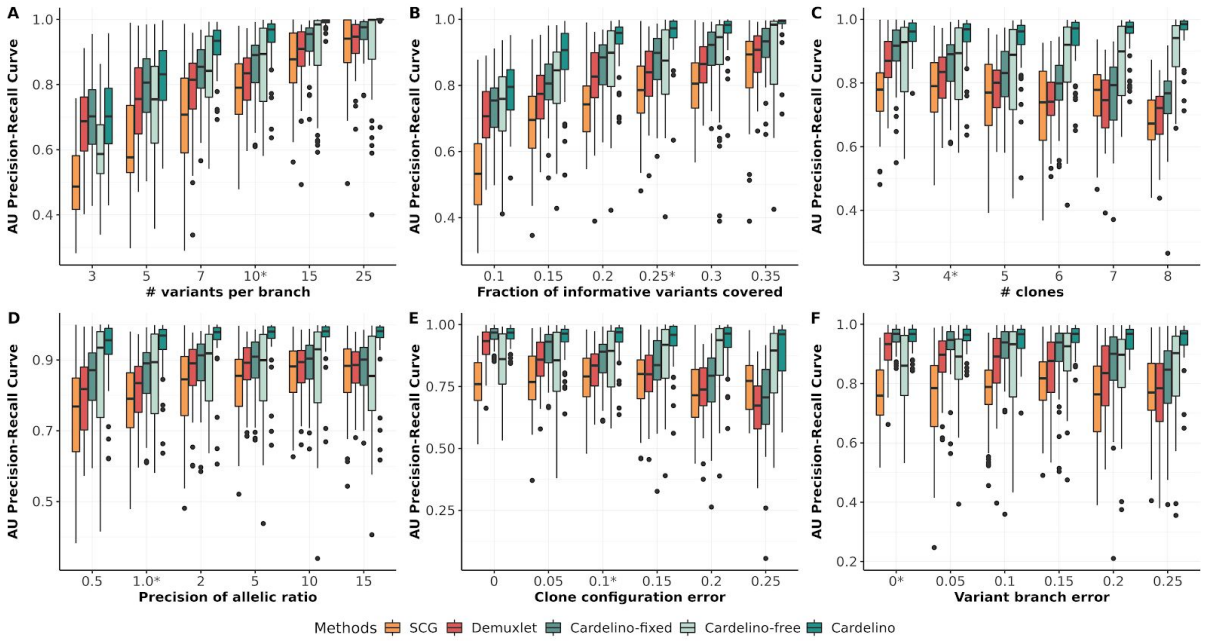


Figure S3. Assessment of cell assignment to clones across a variety of simulation settings, considering SingleCellGenotyper (SCG), Demuxlet (our implementation to avoid the requirement of .bam format), cardelino and its two versions: cardelino-free without any informative clone configuration prior and cardelino-fixed assuming that the clone configuration prior is correct (**Methods** and **Supp Methods**). All methods were applied to simulated data with known ground truth, varying (**A**) the number of informative variants per clonal branch, (**B**) the fraction of informative variants covered (i.e., non-zero scRNA-seq read coverage), (**C**) the total number of clones, (**D**) the precision (i.e., inverse variance) of allelic ratio across genes; lower precision means more genes with high allelic imbalance, (**E**) the rate of general errors of mutation states in the clone configuration matrix, (**F**) the fraction of wrongly clustered variants in the input clonal tree branch. Default parameter values are marked with an asterisk and are retained when varything other parameters.

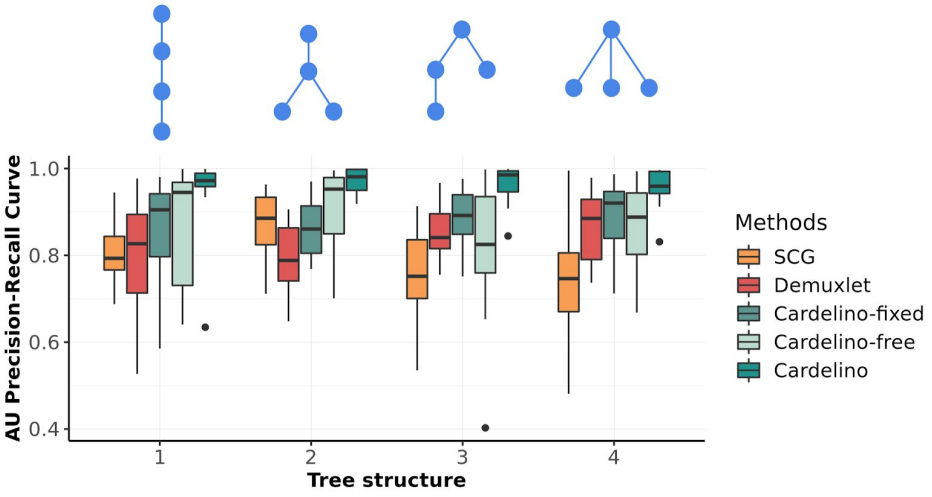
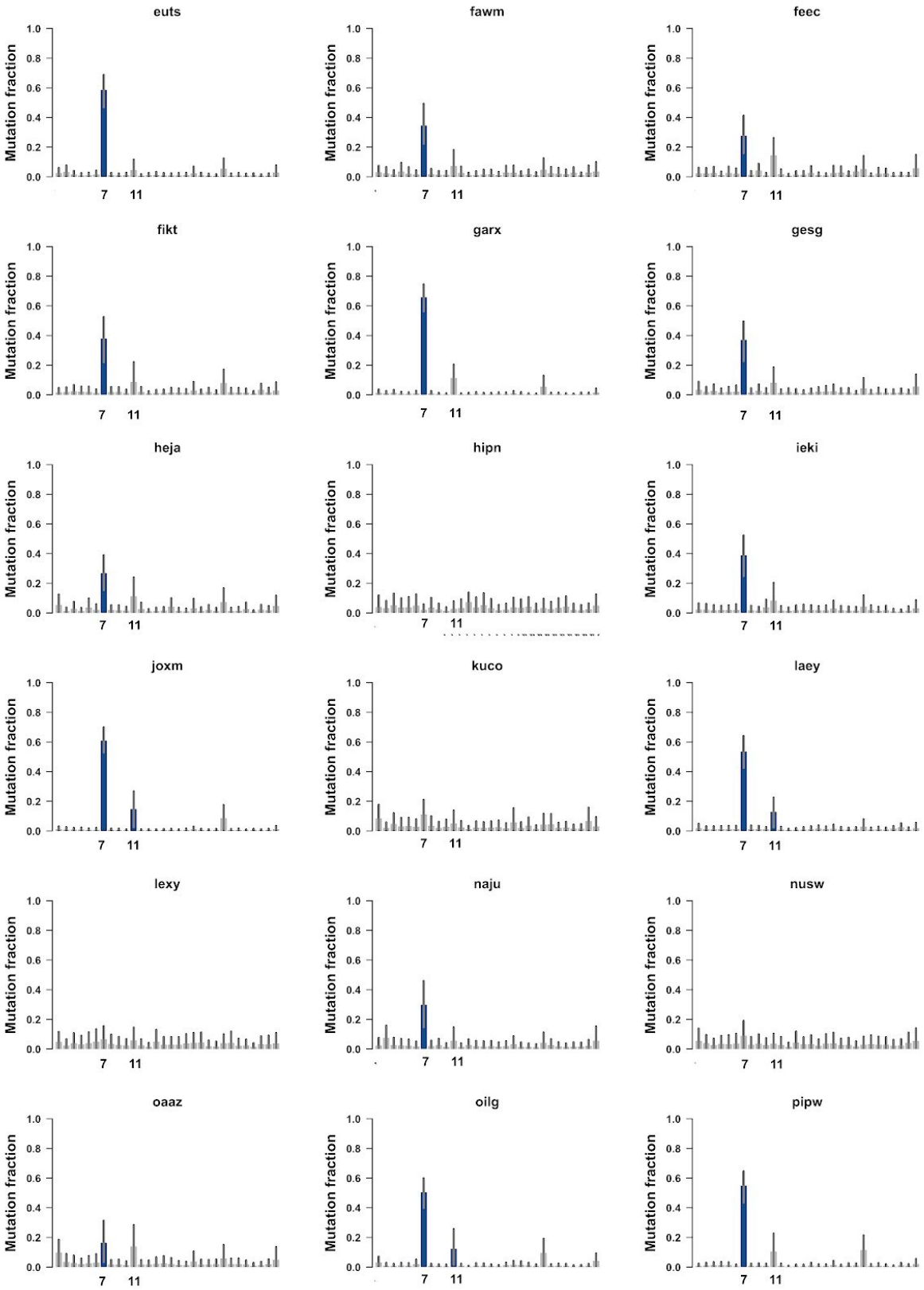


Figure S4. The effects of the tree topology on the cell assignment accuracy. In the simulations in Fig. 1 and Supp Fig. S2, there are 20 repeats for each parameter, where one of the tree topology candidates are randomly selected in each repeat. For the four-clone configuration, there are four different tree topologies (upper panel), and their performance (area under the precision-recall curve) for the five different methods are splitted (bottom panel).


euts
fawm
feec
fikt
garx
gesg
heja
hipn
ieki
joxm
kuco
laey
lexy
naju
nusw
oaaz
oilg
pipw
Mutation fraction
1.0
0.8
0.6
0.4
0.2
0.0
7
11

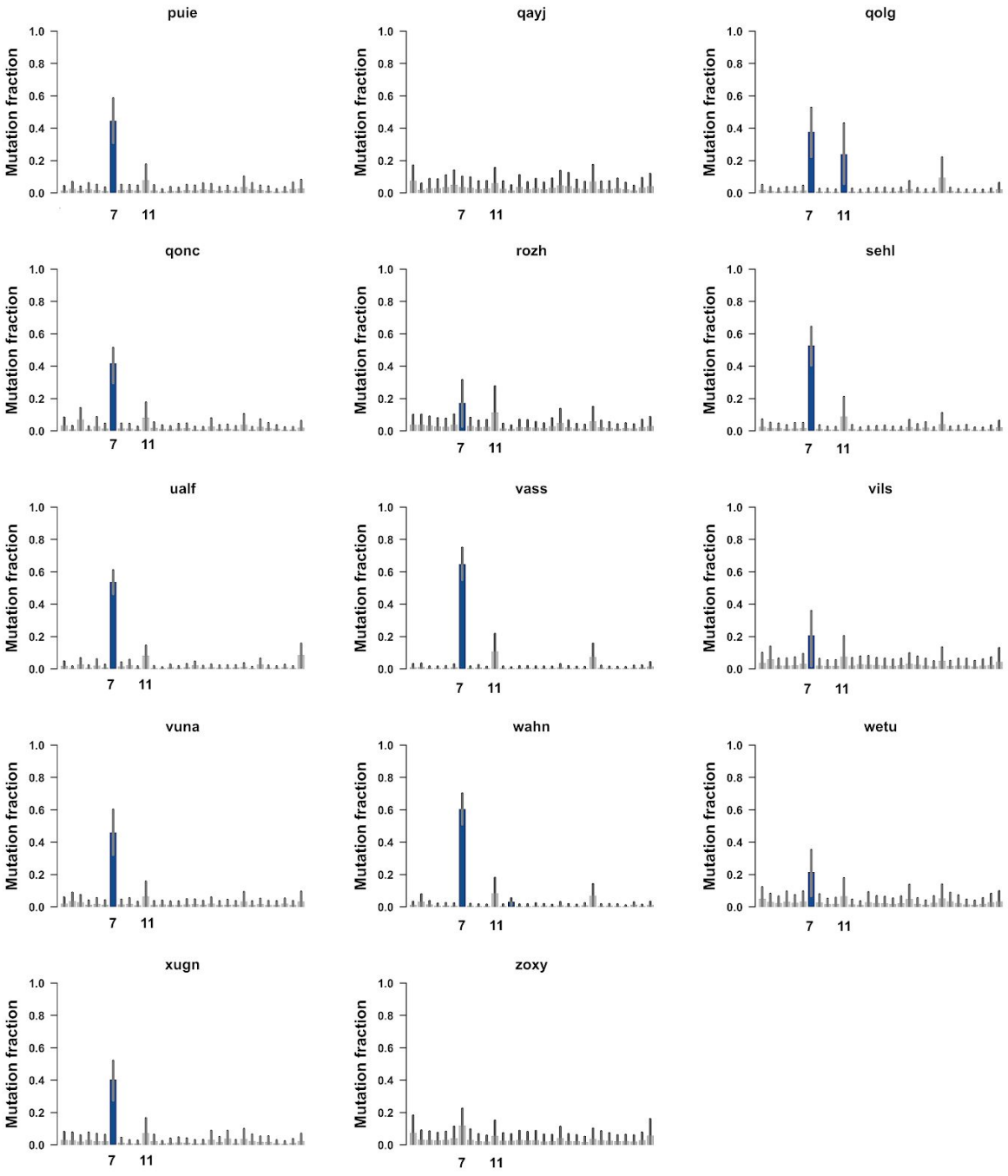


Figure S5. Estimated mutational signature exposures based upon the tri-nucleotide context of somatic SNVs called from whole-exome sequencing (WES) data for 32 HipSci human fibroblast lines. The x-axis shows 30 COSMIC mutational signatures, in order, and the y-axis shows estimated exposures (mutation fraction) using the *sigfit* package (**Methods**), with significant signatures highlighted in blue. Across lines, the only significant signatures are Signature 7 (UV mutagenic process) and Signature 11.

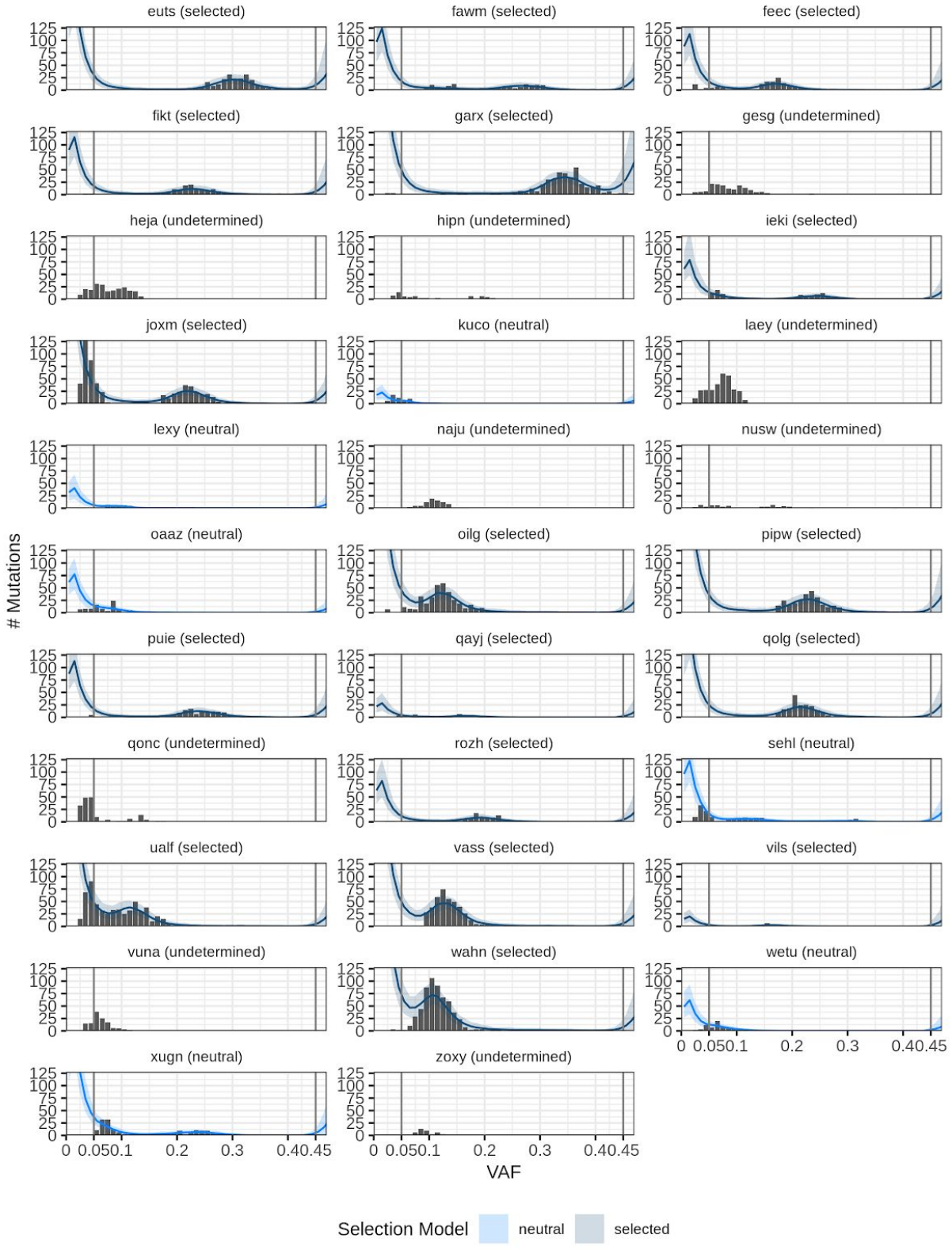


Figure S6. Variant allele frequency (VAF) distributions for somatic variants called from whole exome sequencing data for the 32 fibroblast lines. The grey lines indicate the cut-offs on the allele frequency distribution (**Methods**). The blue lines indicate the model (neutral/selected) inferred by SubClonalSelection (shading 95% confidence interval).

Figure S7. Clonal tree inferred by Canopy and then updated by cardelino (shown is output from cardelino) and posterior probability of assignment of each cell to each clone from cardelino for the 32 lines analysed in detail in the manuscript.

Figure S8. Clonal tree inferred by Canopy (unaltered tree output from Canopy is shown) and posterior probability of assignment of each cell to each clone from cardelino for the 32 lines analysed in detail in the manuscript.

Figure S9. Comparison of the clonal tree inferred by Canopy and the updated tree after running cardeilno for the 32 lines analysed in detail in the manuscript.

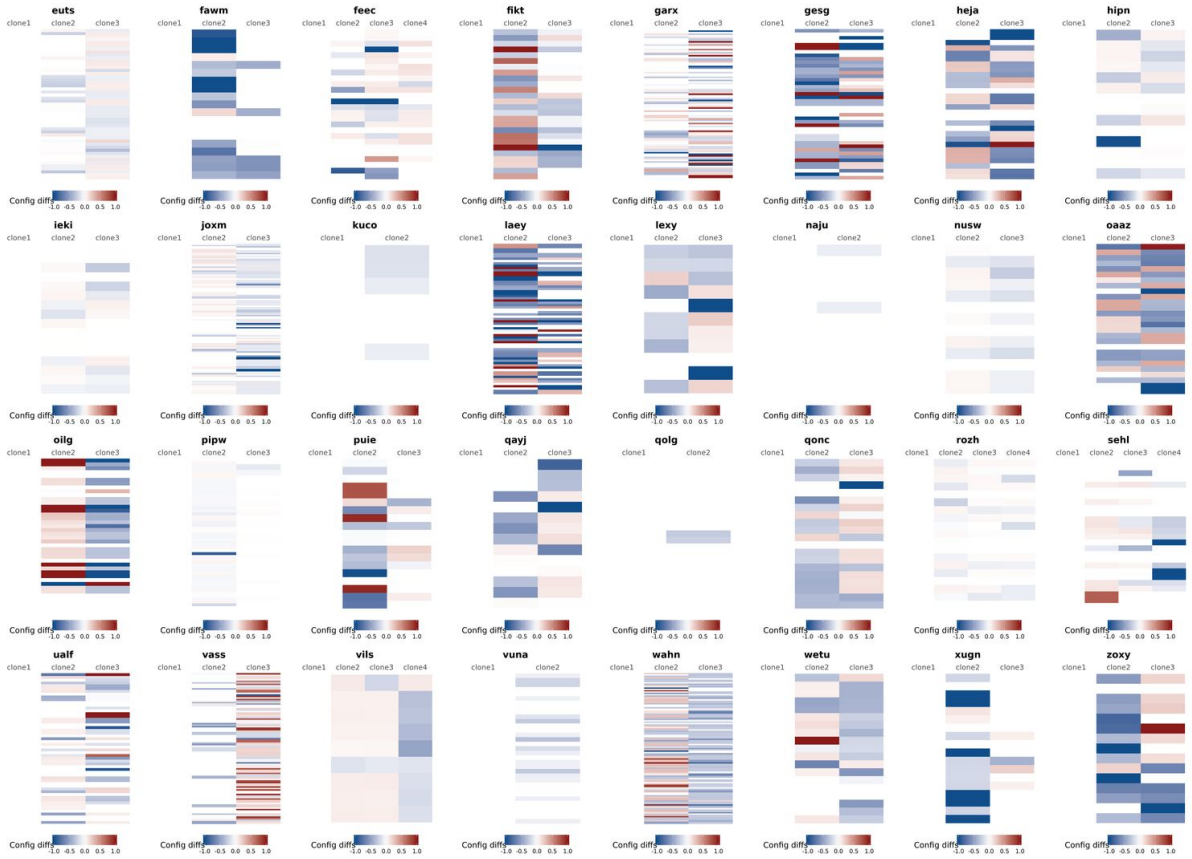


Figure S10. Differences in configuration matrices (rows represent single-nucleotide variants and columns represent clones) between Canopy trees and updated trees from cardelino (average configuration matrix over 4,750 posterior samples from the cardelino model minus the configuration matrix for the tree inferred by Canopy).

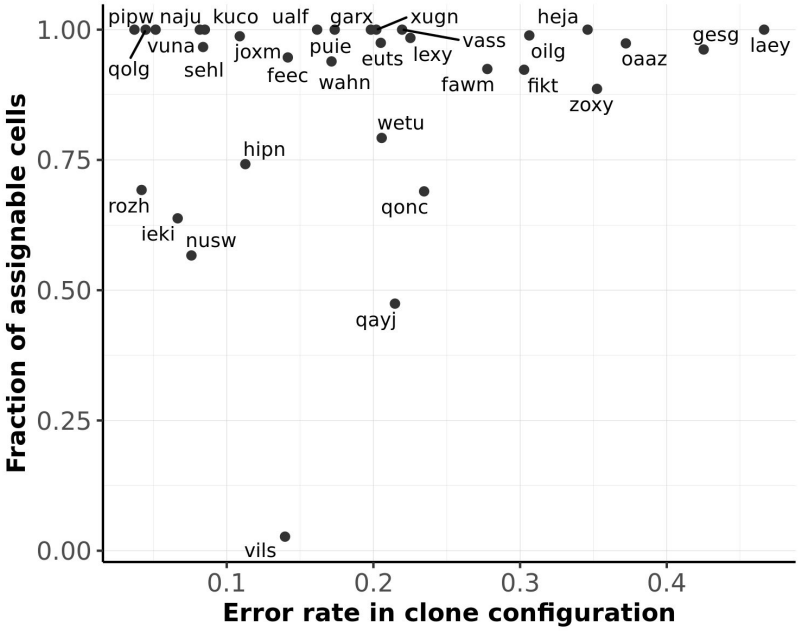


Figure S11. Estimated error rate in the clonal tree configuration derived from bulk exome-seq data (based on cardelino) for each of 32 lines versus fraction of confidently assigned cells. Even though some lines have high error rate in the input clonal tree configuration, cardelino can still assign a high fraction of cells to clones confidently.

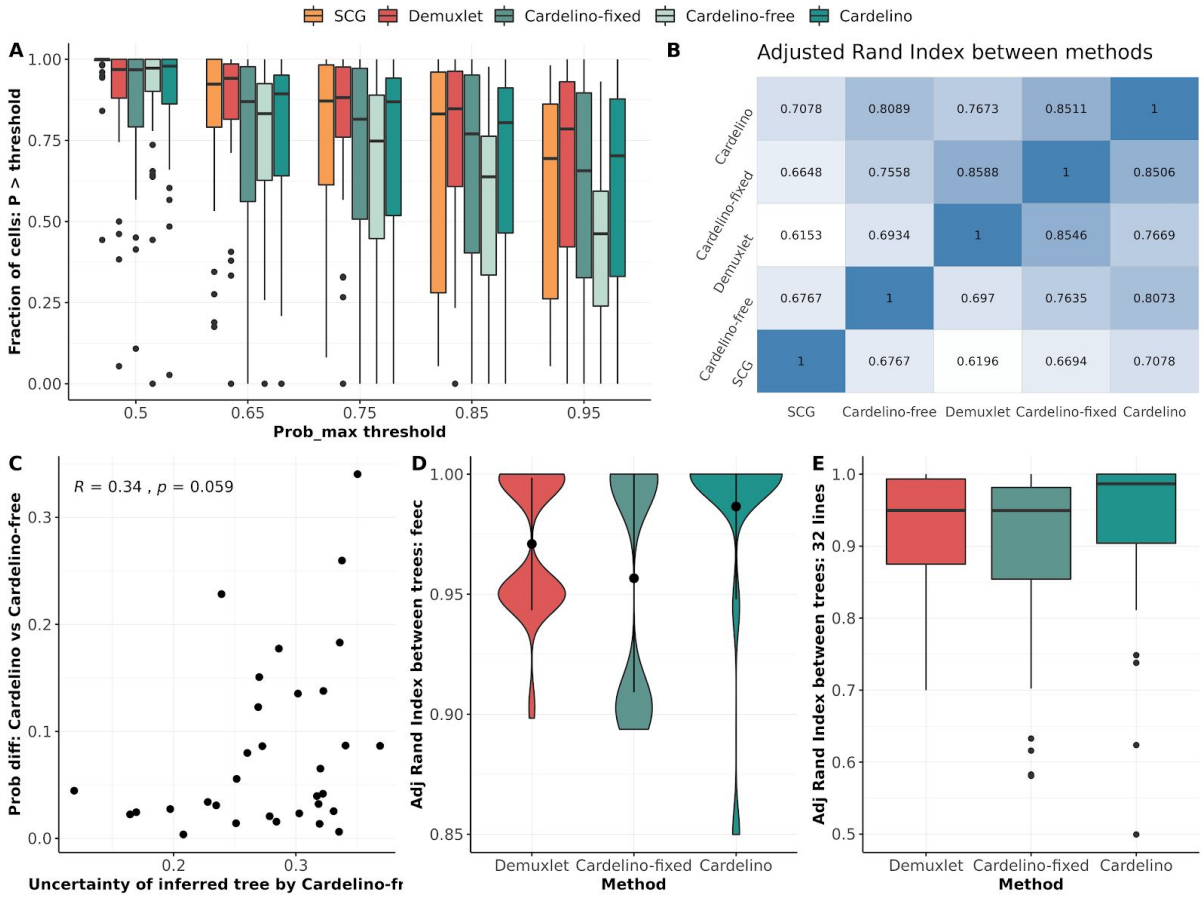


Figure S12. Comparison of cell assignment between five methods across 32 lines. (**A**) The fraction of assignable cells (i.e., highest P > thresholds) when varying the thresholds from 0.5 to 0.95. Shown are box plots depicting median and the first and third quantiles of the 32 lines. (**B**) The adjusted Rand index of cell assignment to clones between the five considered methods. The values is averaged across 32 lines. (**C**) Scatter plot between the uncertainty of the inferred tree from cardelino-free (x-axis) and the mean absolute difference of the assignment probability between cardelino-free and cardelino (y-axis). The output posterior clonal configuration matrix from cardelino-free consists of the probability of each variant being present in each clone. A completely uninformative clonal tree would have all entries equal to 0.5. Thus, we measure the uncertainty of the output tree from cardelino-free by taking 0.5 minus the mean absolute difference of the posterior probability configuration matrix and the uninformative configuration probability matrix of all of entries equal to 0.5. WIth this measure, a value of 0.5 indicates a posterior configuration indistinguishable from the uninformative configuration and a value of 0 indicates very high confidence from the model in the posterior configuration. (**D**) Pairwise comparison of clone assignments by adjusted Rand Index for high-probability Canopy tree solutions on one representative line: feec. Shown are pairwise comparisons for the thirty most probable trees derived from bulk exome-seq data with Canopy, leading to 435 tree pairs for each line. (**E**) The adjusted Rand index of cell assignment between two different guide clonal trees across all 32 lines. Each dot in the boxplot denotes a line, which is the average of these 435 pairwise comparisons.

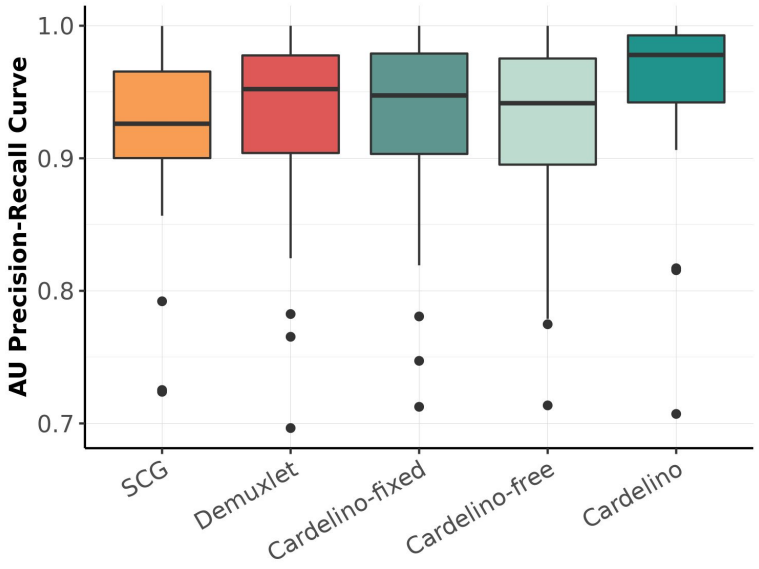


Figure S13. Assessment of cell assignment to clones across a variety of simulation settings, considering SingleCellGenotyper (SCG), Demuxlet (our implementation to avoid the requirement of .bam format), cardelino and its two versions: cardelino-free without any informative clone configuration prior and cardelino-fixed assuming that the clone configuration prior is all correct (Methods and Supp Methods). Considered were simulated data based on empirical characteristics observed in 32 fibroblast lines. For each line, the sequence coverage, clone configuration (i.e., number of clones, variants on each branch), and allelic imbalance parameters were obtained to derive simulation parameters. 200 cells are synthesised per line and a clone configuration with 10% errors are used as a guide. The main Fig. 2b and Supp. Fig. S13 are both based on this simulation.

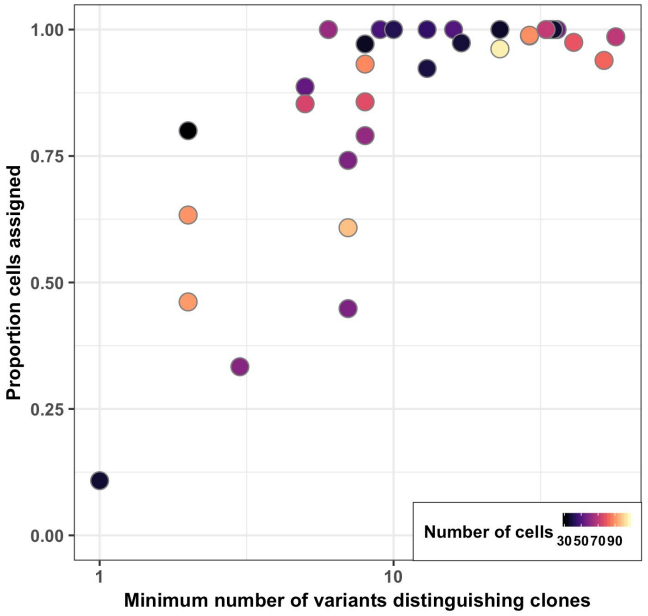


Figure S14. Scatter plot of the fraction of cells assigned in each cell line using cardelino (at posterior probability > 0.5) as a function of the minimum number of clone-specific variants for the corresponding line (minimum Hamming distance between clones for a given donor), for 32 fibroblast lines. Total number of cells that were considered for this analysis (QC passed) per line indicated by colour.

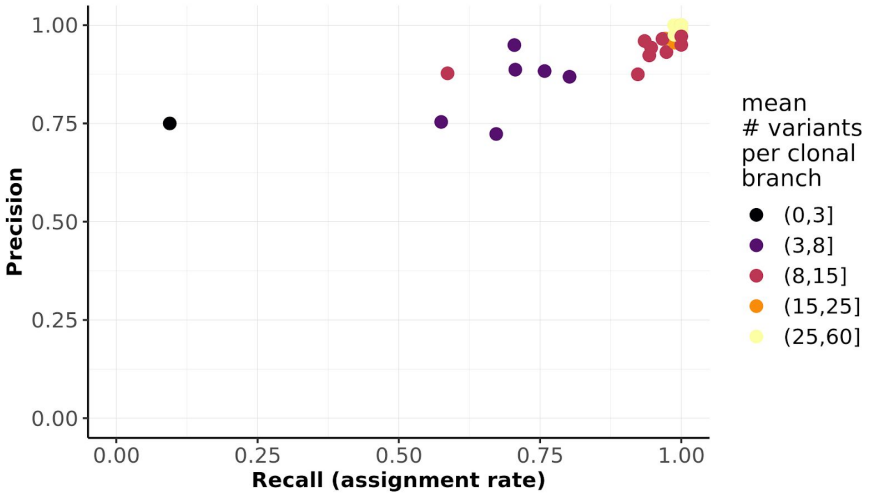


Figure S15. Scatter plot of recall (assignment rate) versus precision (assignment accuracy) when assigning cells using cardelino (at posterior probability > 0.5). Shown are data from for 32 simulated lines, using parameters that match the observed data characteristics in the set of 32 real fibroblast lines (**Methods**). The average number of variants per clonal branch (*i.e.*, #variant / (#clone - 1)) is shown by point colour (slightly different from Supp. Fig. S4 which uses the minimum number of variants distinguishing between pairs of clones, as shown in Fig. 3a). Lines with fewer informative variants per branch tend to have lower assignment rates, but the precision remains high.

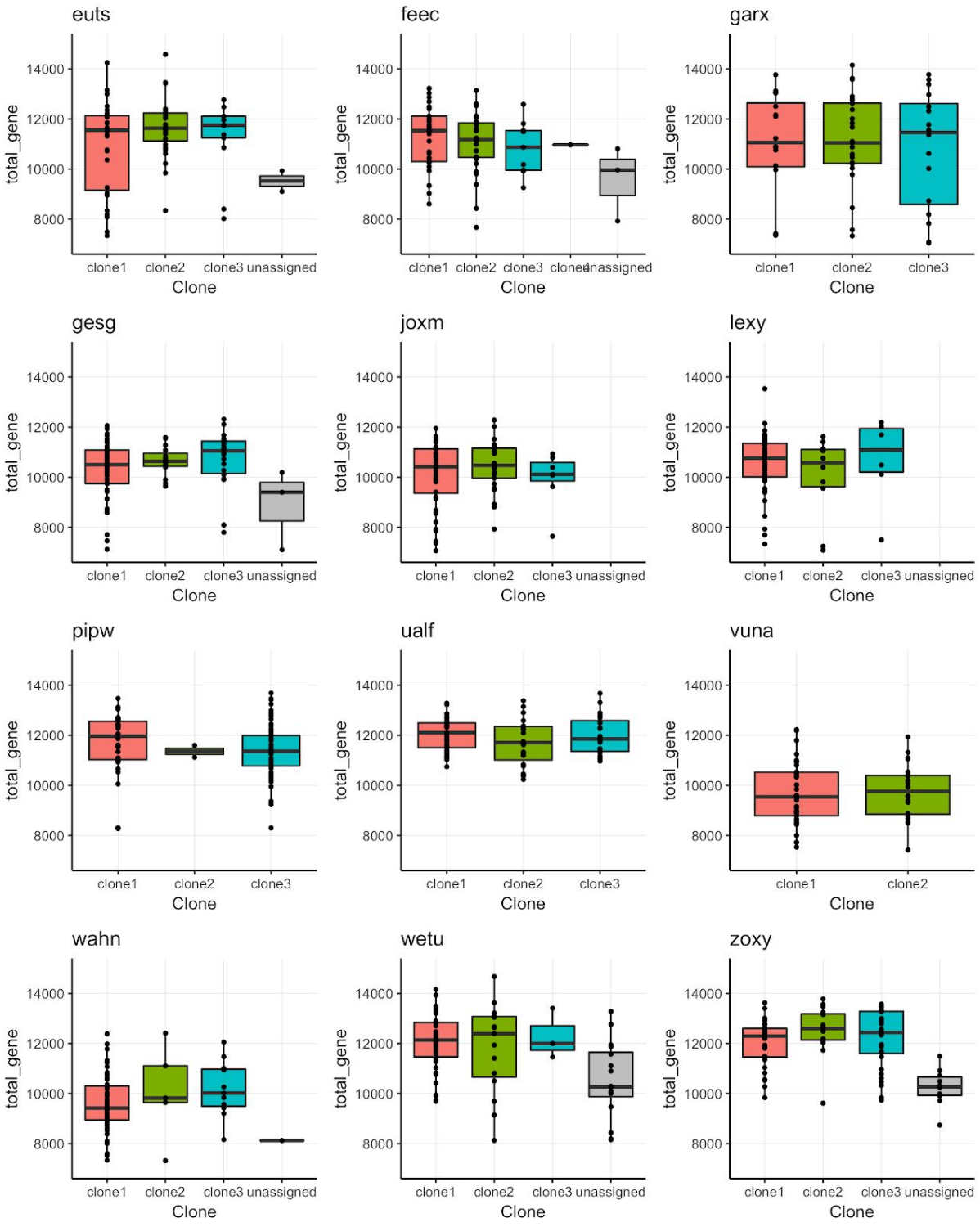


Figure S16. Boxplots of the total number of expressed genes in each cell, grouped by the clone assigned by cardelino. Twelve lines with more than 60 assignable cells are presented. Globally, clone assignment is not linked to the total number of expressed genes in a given cell.

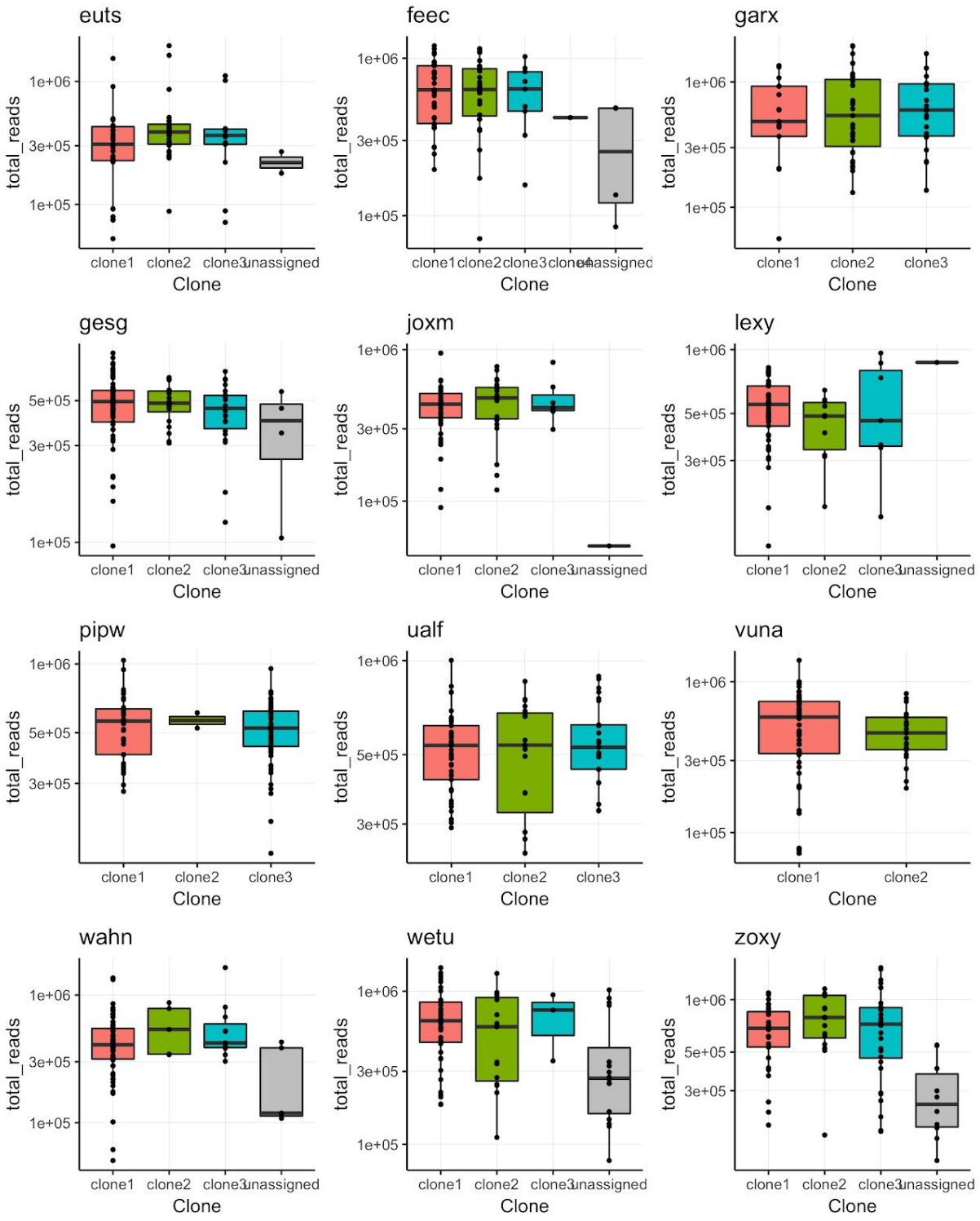


Figure S17. Boxplots of the total number of sequenced read counts from endogenous genes in each cell, grouped by the clone assigned by cardelino. Twelve lines with more than 60 assignable cells are presented. Globally, clone assignment is not linked to the total number of read counts in a given cell.

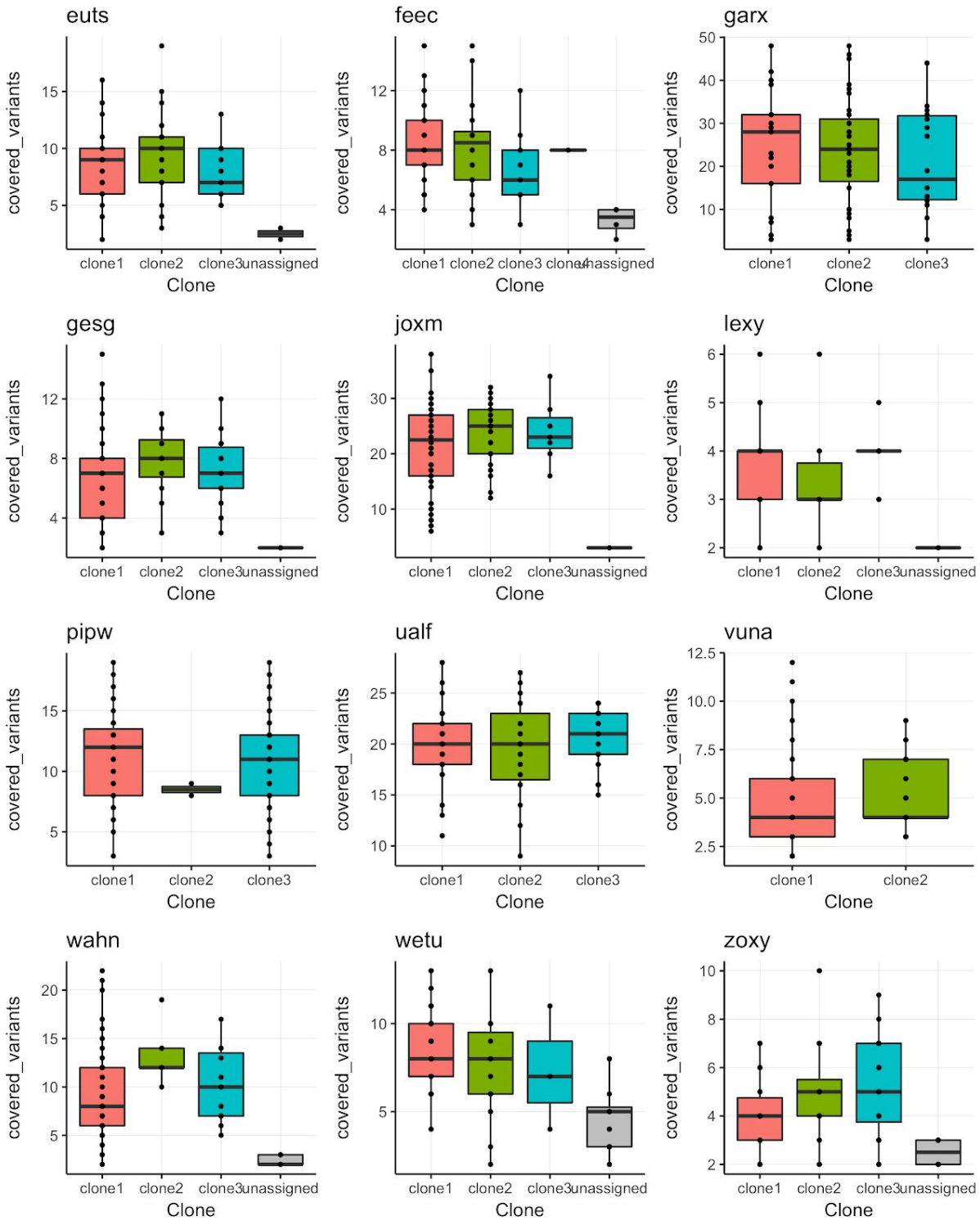


Figure S18. Boxplots of the number of variants for clone identification with read coverage in each cell, grouped by the clone assigned by cardelino. Twelve lines with more than 60 assignable cells are presented. Globally, clone assignment is not linked to the number expressed variant loci in a given cell, with the exception of the “unassigned” category which is enriched for cells with low coverage.

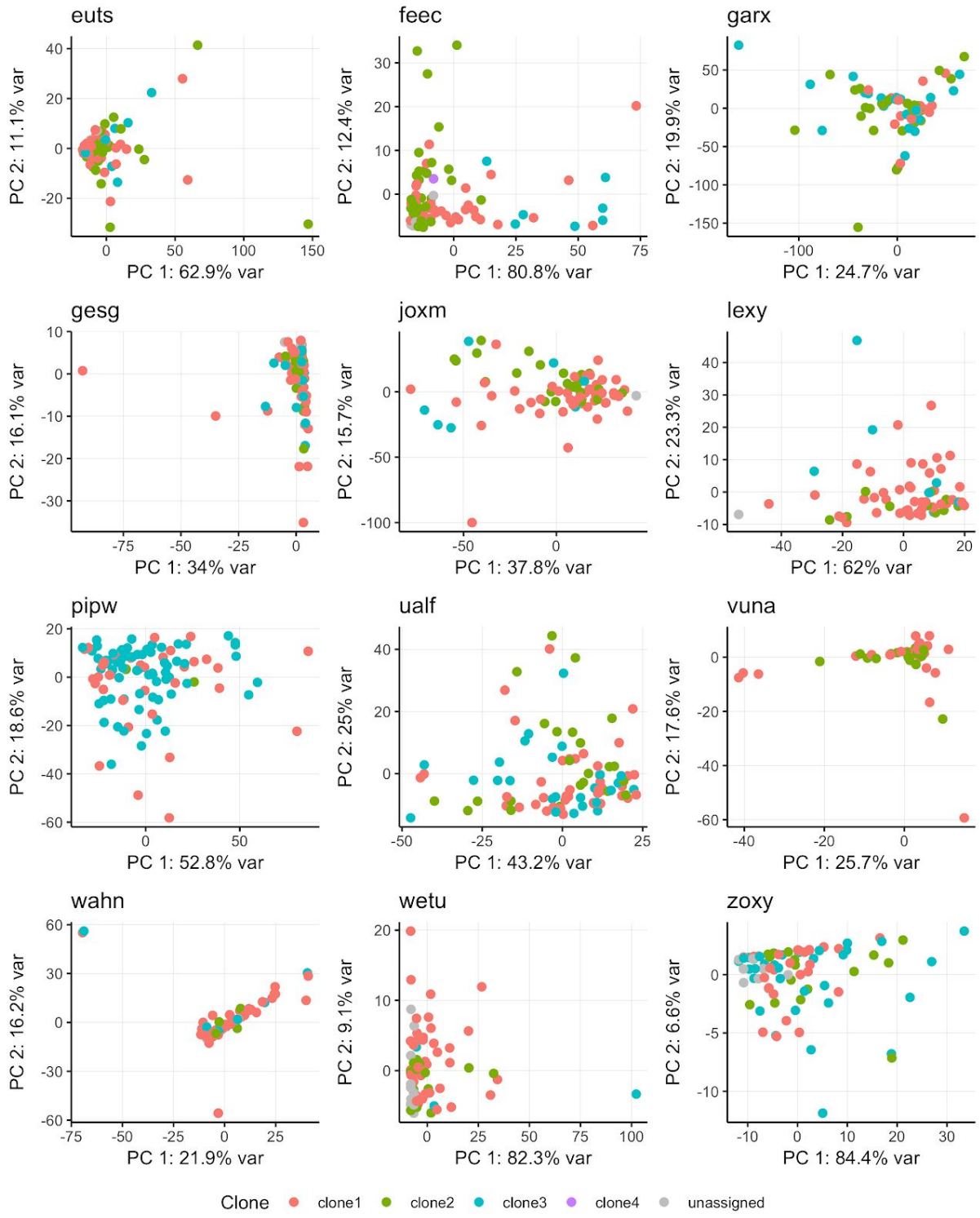


Figure S19. Scatter plot of the first two principal components calculated on the read coverage of the set of somatic variant sites used for clone assignment. Shown are data from twelve lines with at least 60 assignable cells. The first two PCs do not segregate cells from different clones, suggesting that read coverage of somatic variants does not associate with or bias clone assignment.

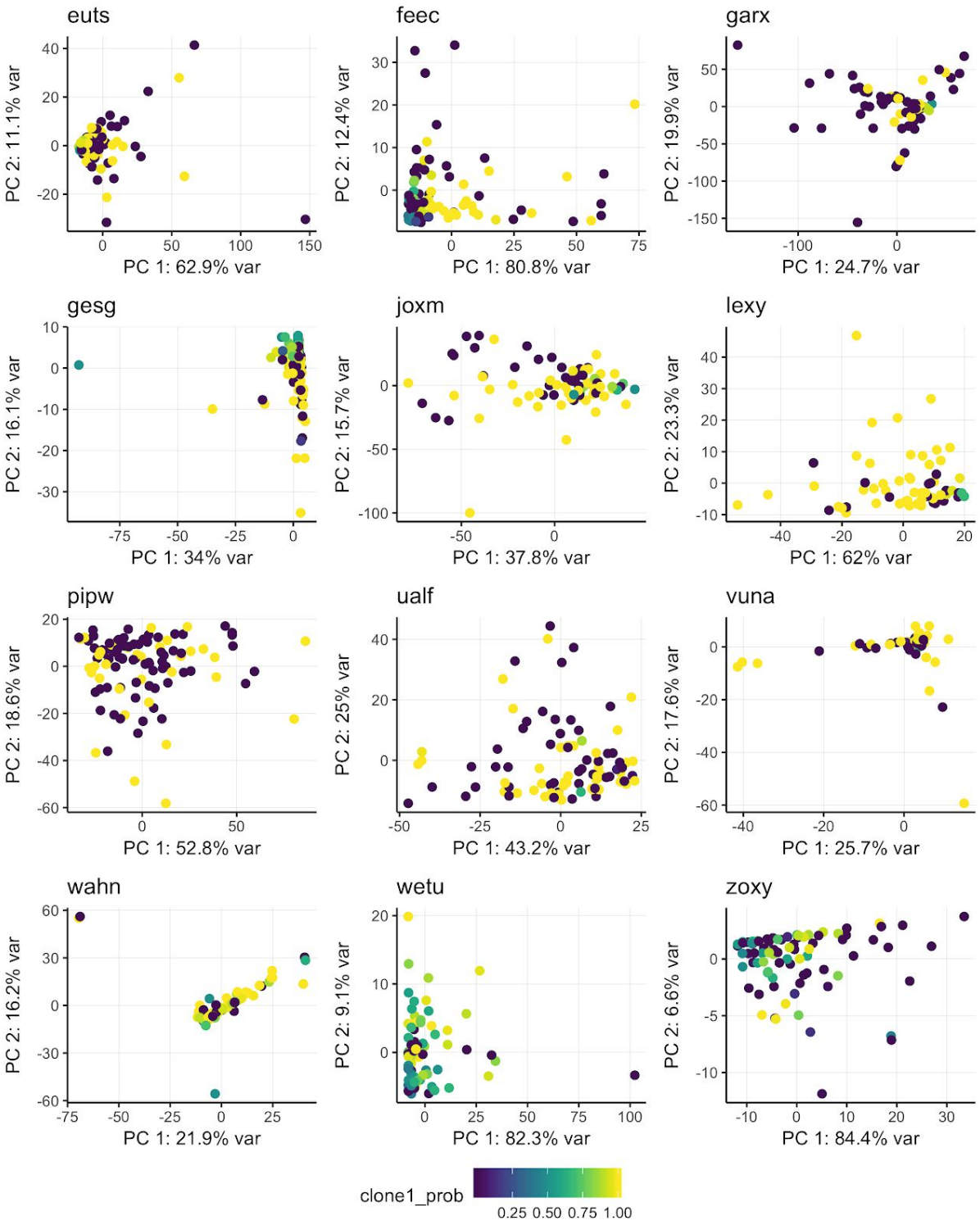


Figure S20. Scatter plot of the first two principal components calculated on the read coverage of the set of somatic variant sites used for clone assignment. Cells are colored by the assignment probability of clone 1 (*i.e.* the "base clone" which by definition contains no unique somatic variants). Shown are data from twelve lines with at least 60 assignable cells.

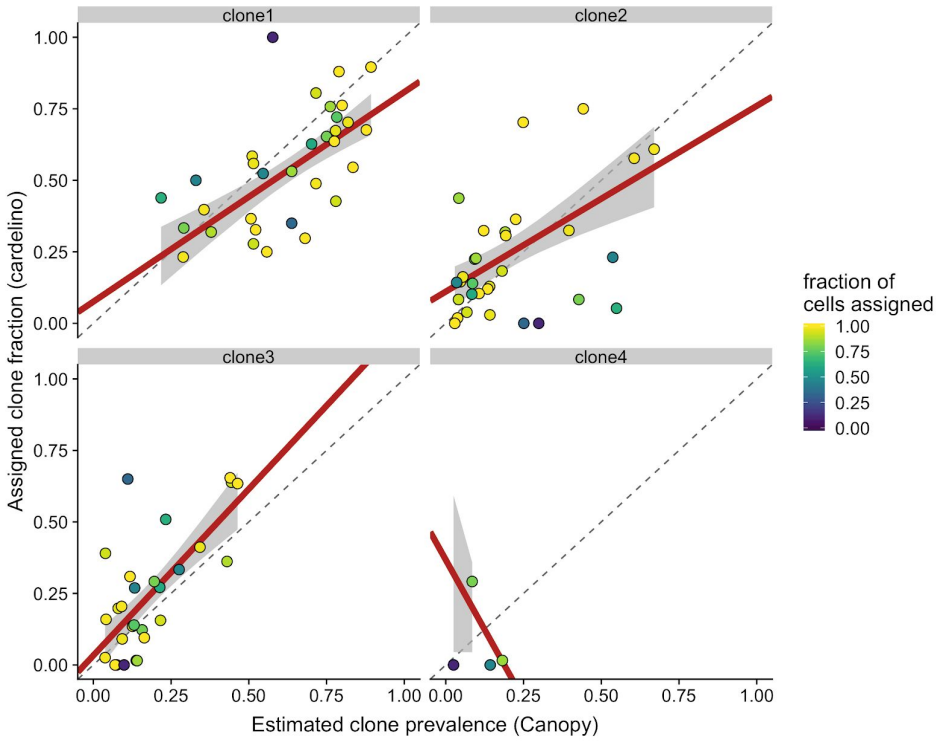


Figure S21. Clone prevalence estimates from WES data (x-axis; using Canopy) *versus* the fraction of single-cell transcriptomes assigned to the clone (y-axis; using *cardelino*), for each clone across lines. Points are coloured by the overall fraction of single-cell transcriptomes assigned for a given line (*i.e.* cells with posterior P>0.5 for assignment).

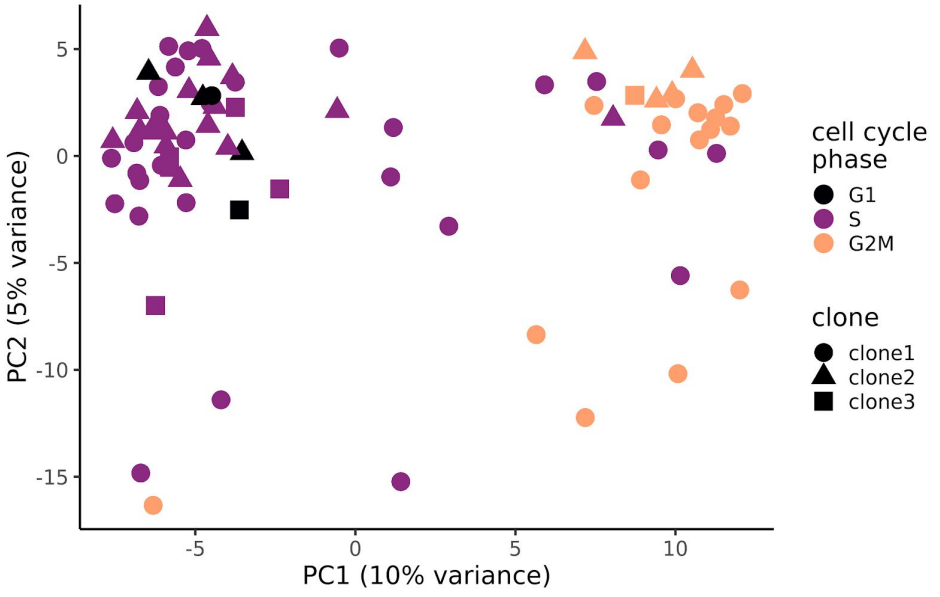


Figure S22. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) for clone-assigned cells for the example line *joxm*. Cells are coloured by the cell cycle phase inferred by the *cyclone* method implemented in the scran package, and shape denotes the assigned clone from cardelino.

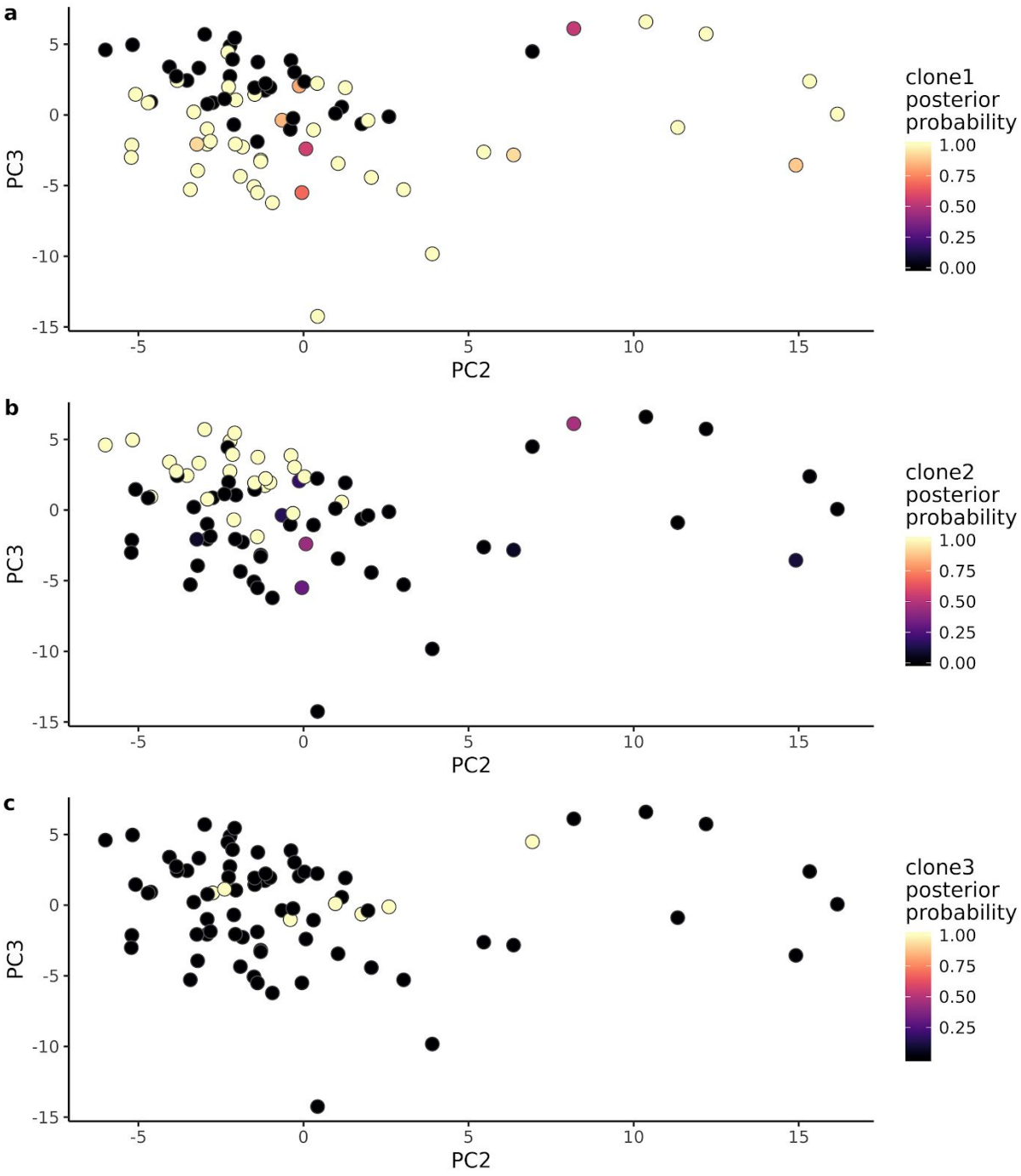


Figure S23. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) for clone-assigned cells for the donor *joxm*, plotting principal component 3 against principal component 2. Cells are coloured by the posterior probability from cardelino that the cell belongs to clone1 **(a)**, clone2 **(b)** or clone3 **(c)**.

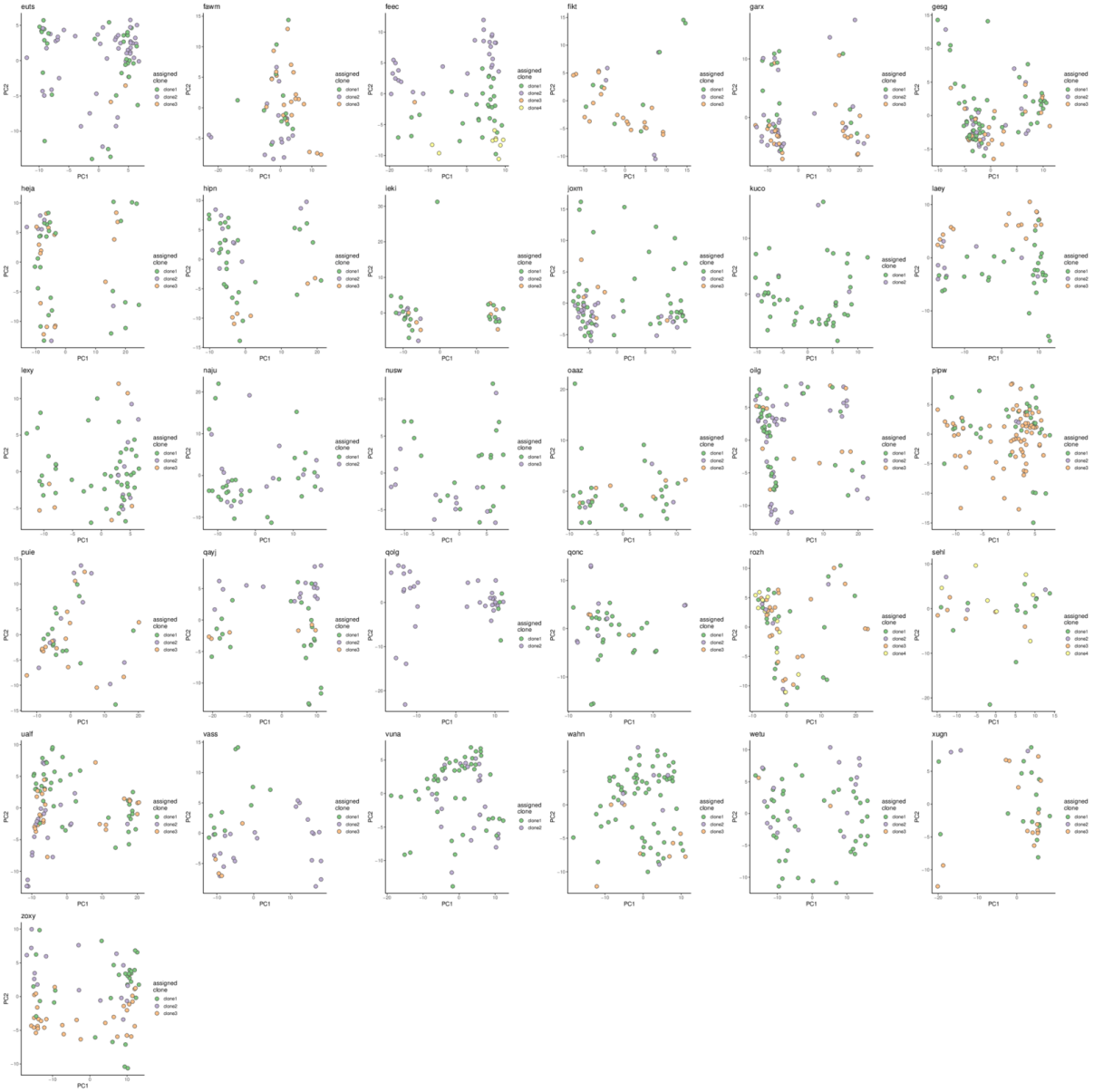


Figure S24. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the assigned clone from cardelino.

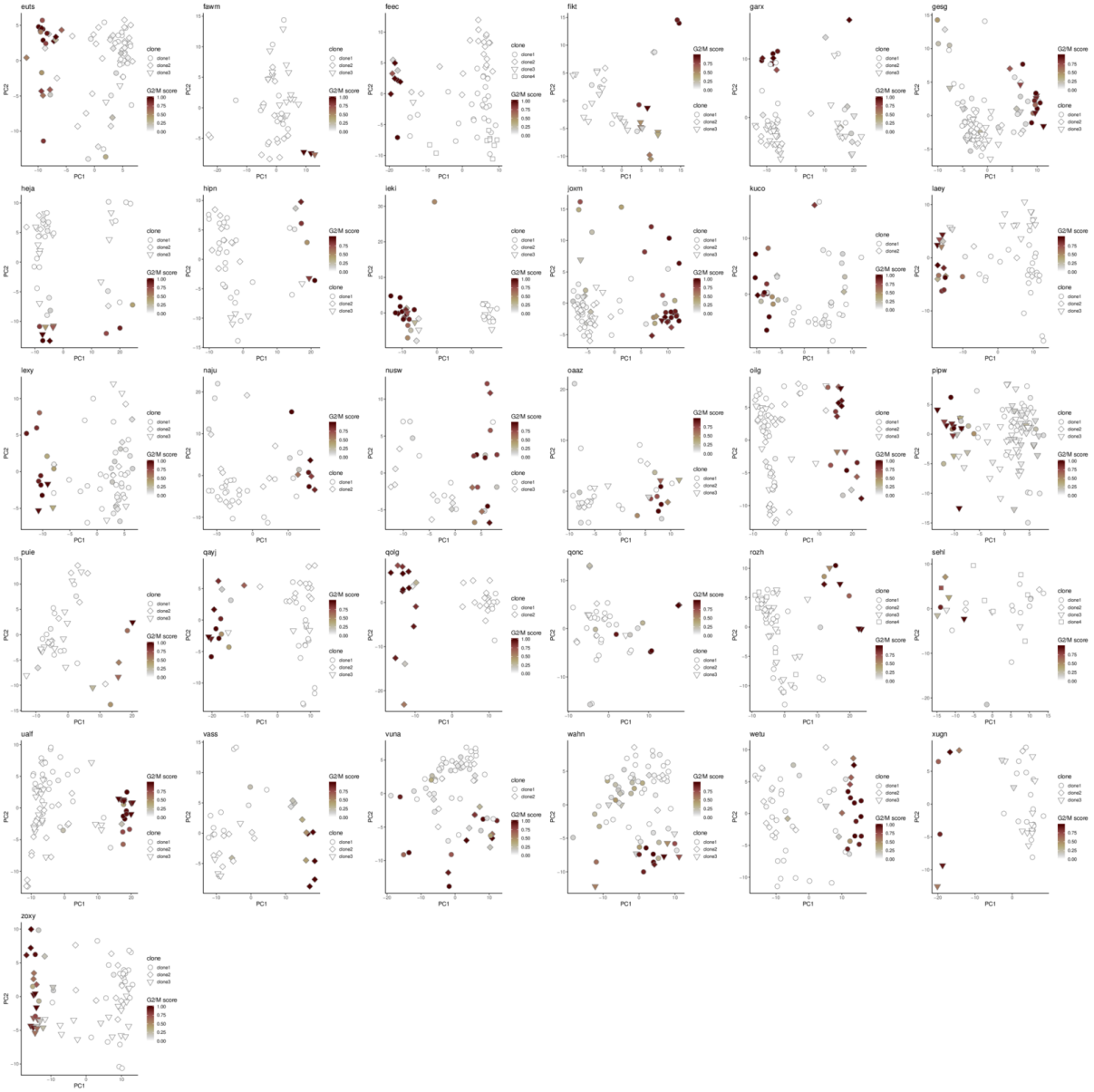


Figure S25. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the G2M cell cycle phase score calculated with the cyclone method implemented in the scran package, and shape denotes the assigned clone from cardelino.

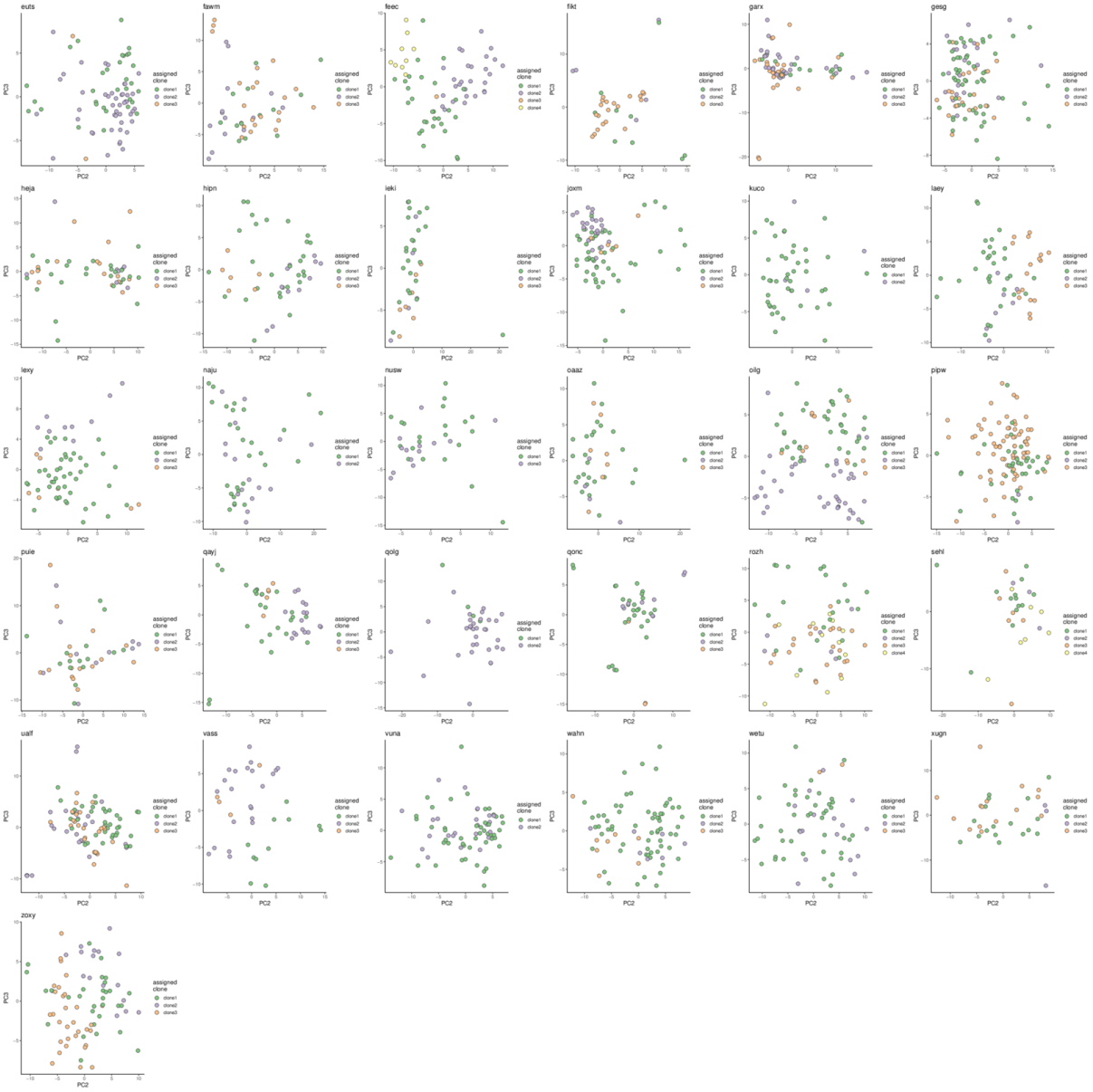


Figure S26. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC3 plotted against PC2 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the assigned clone from cardelino.

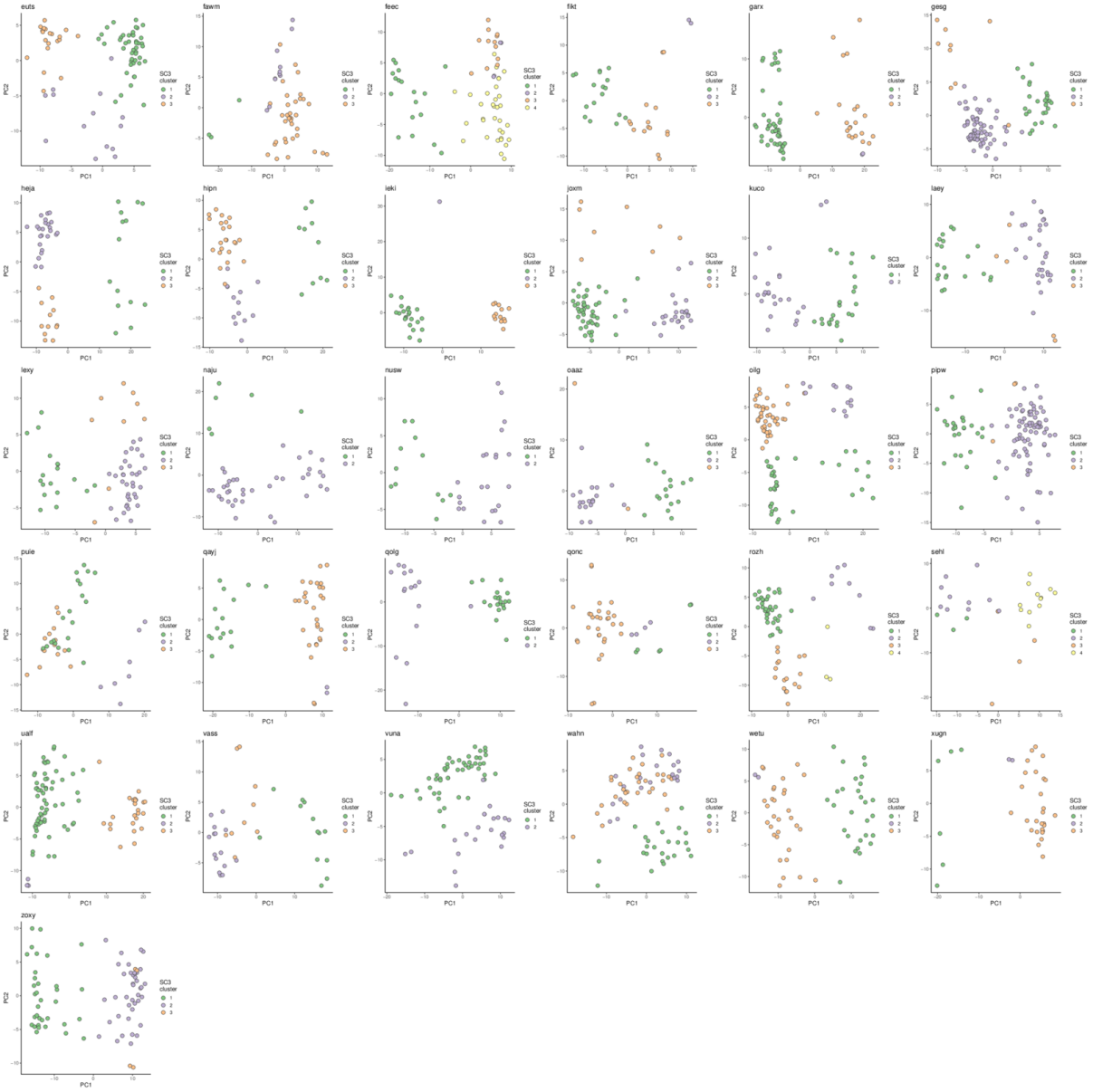


Figure S27. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the clusters identified by SC3 (Kiselev et al, *Nature Methods*, 2017).

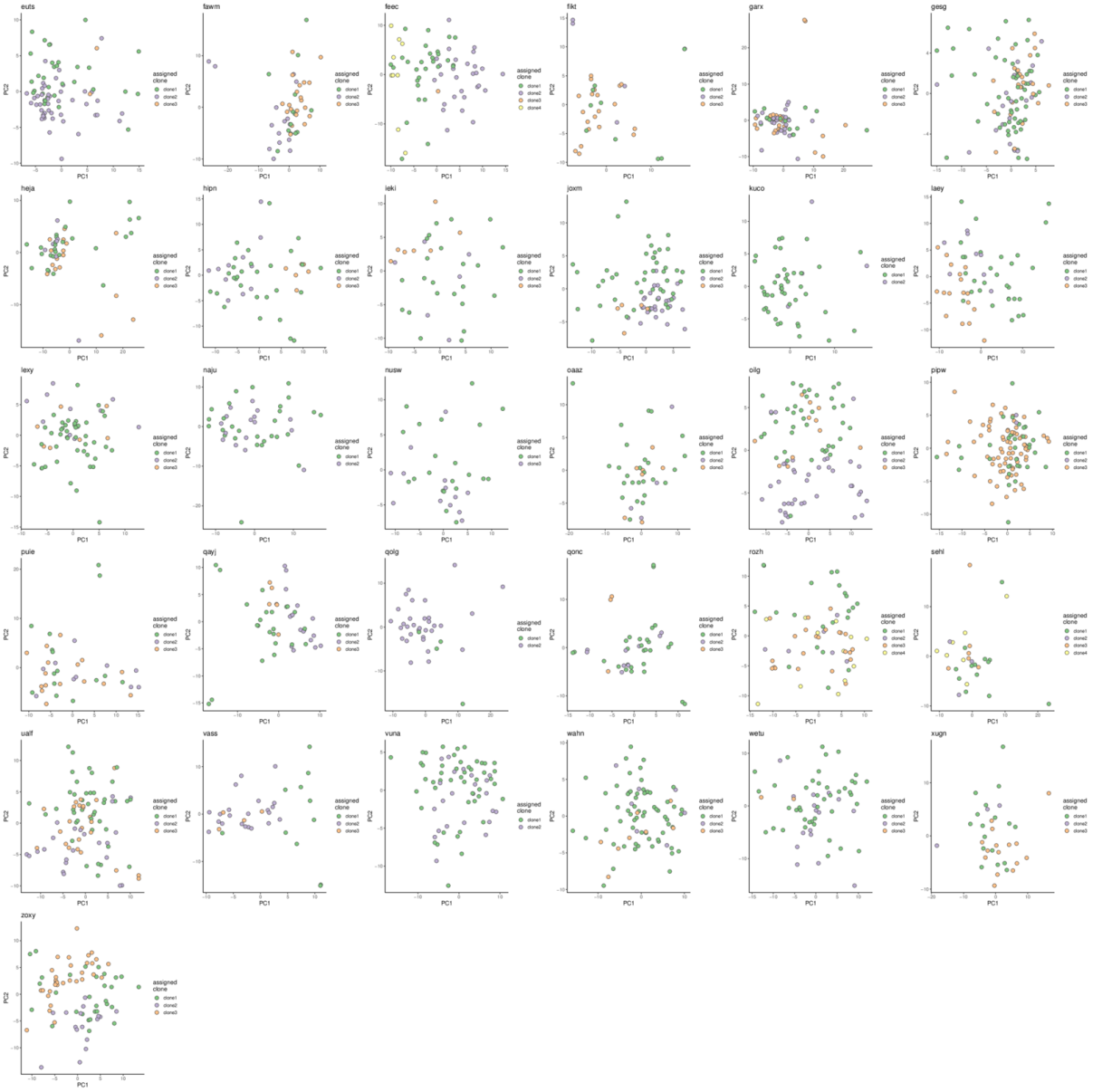


Figure S28. Principal component analysis from single-cell gene expression data after regressing out *cyclone* G1, G2M and S cell cycle scores from the normalised expression values (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the assigned clone from cardelino.

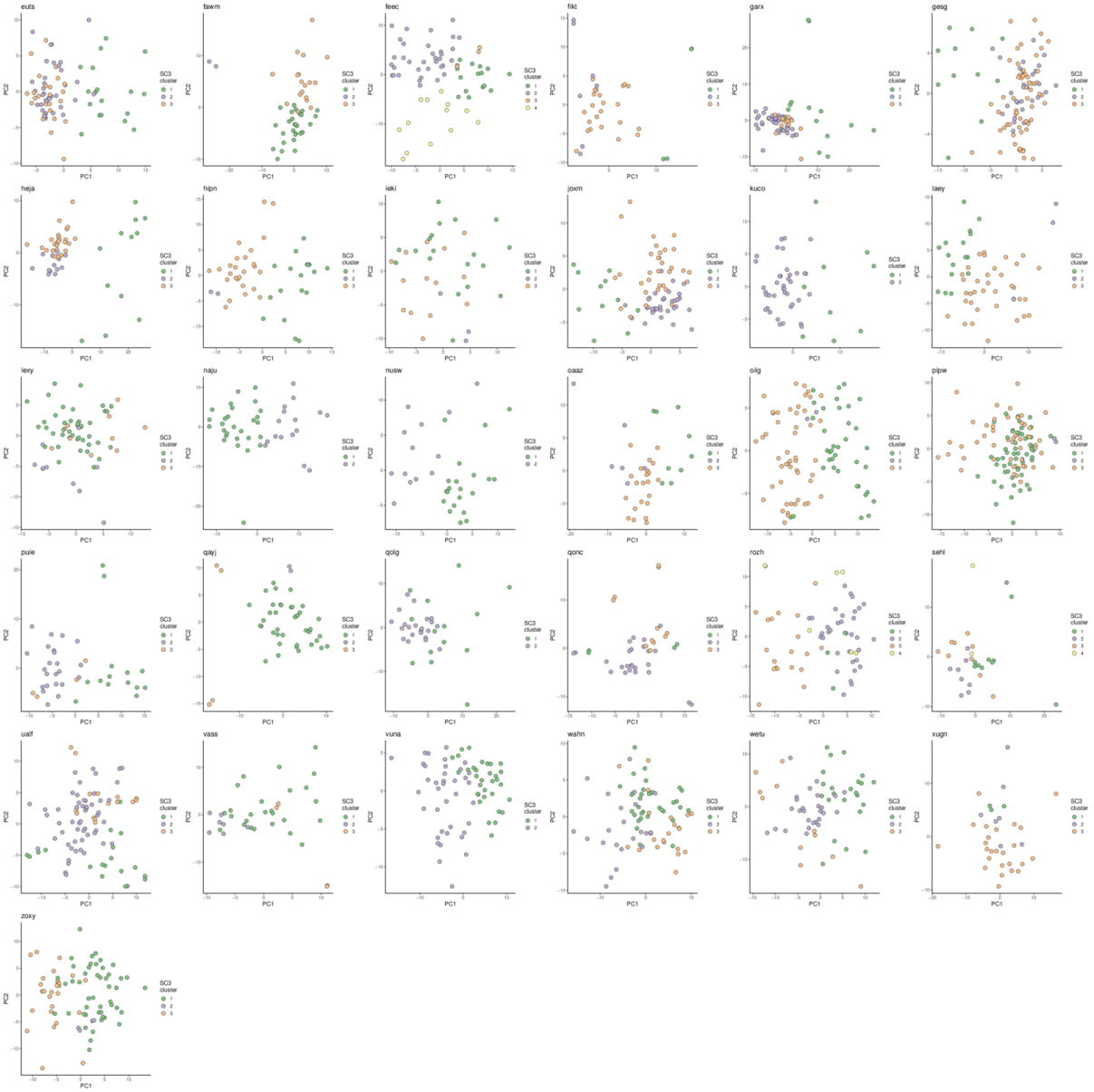


Figure S29. Principal component analysis from single-cell gene expression data after regressing out *cyclone* G1, G2M and S cell cycle scores from the normalised expression values (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the clusters identified by SC3.

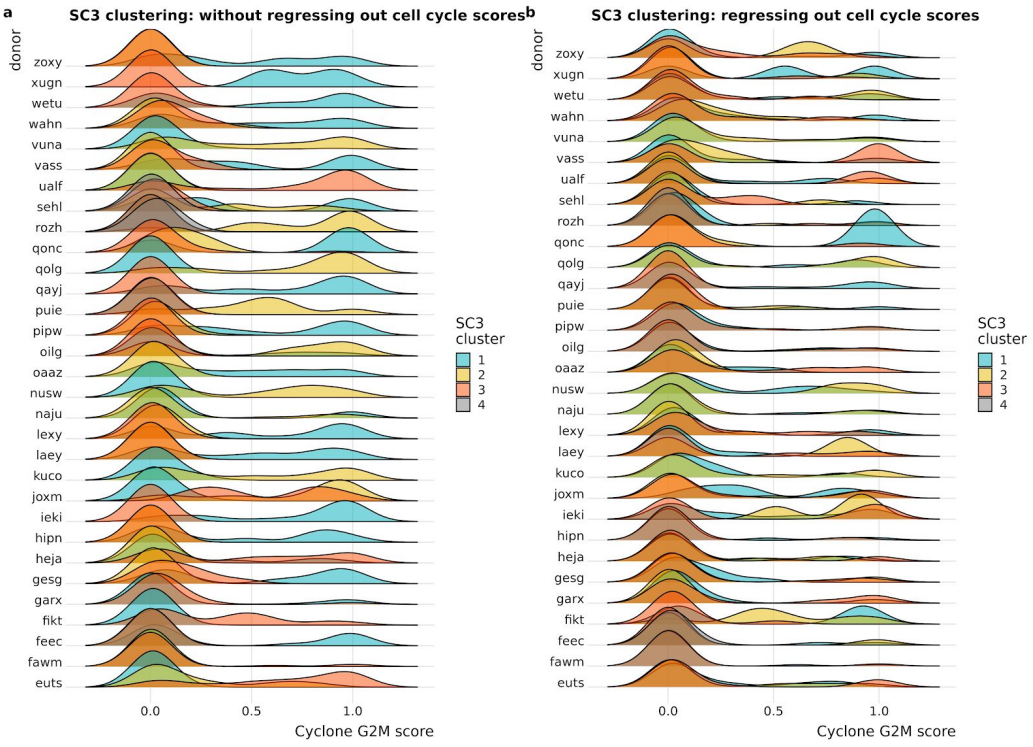


Figure S30. Distributions of *cyclone* G2M scores for each cell line (donor) stratified (coloured) by the clusters identified by SC3 when **(a)** applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and **(b)** applying SC3 to gene expression values after regressing out *cyclone* cell cycle scores.

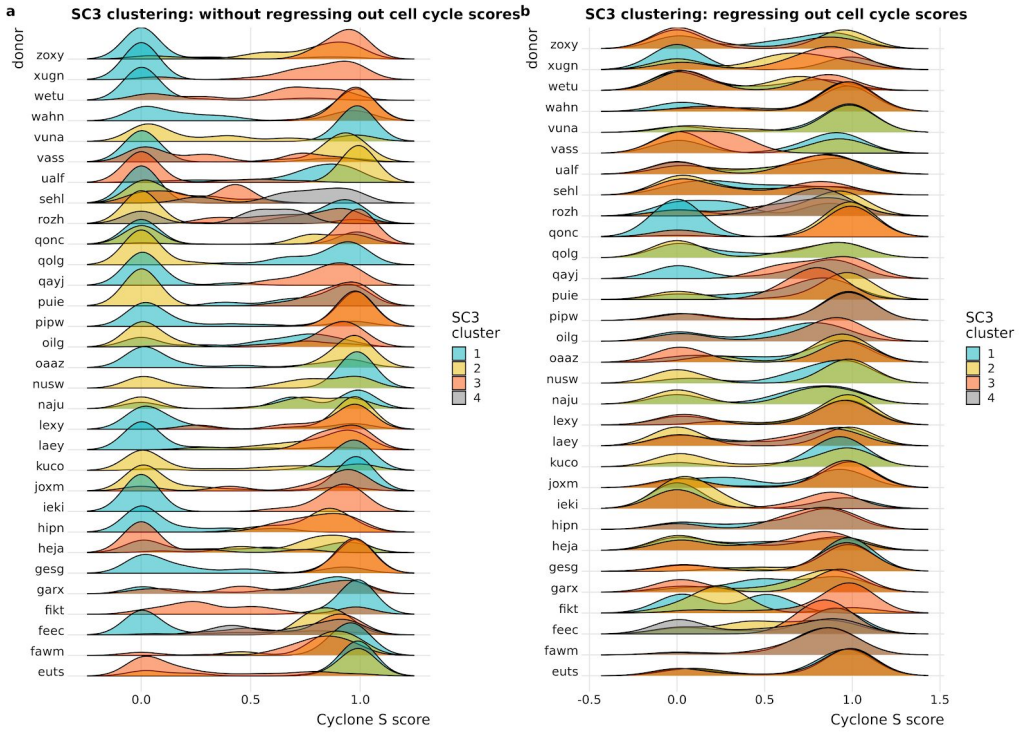


Figure S31. Distributions of *cyclone* S scores for each cell line (donor) stratified (coloured) by the clusters identified by SC3 when **(a)** applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and **(b)** applying SC3 to gene expression values after regressing out *cyclone* cell cycle scores.

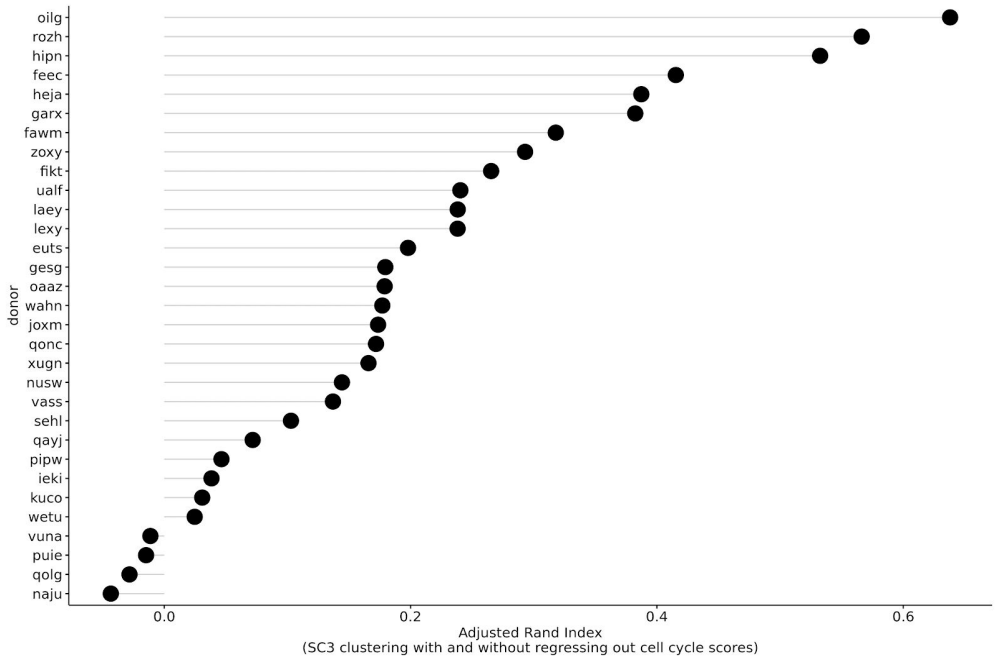


Figure S32. Adjusted Rand Index values comparing the clusters identified by SC3 when applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and when applying SC3 to gene expression values after regressing out *cyclone* cell cycle scores.

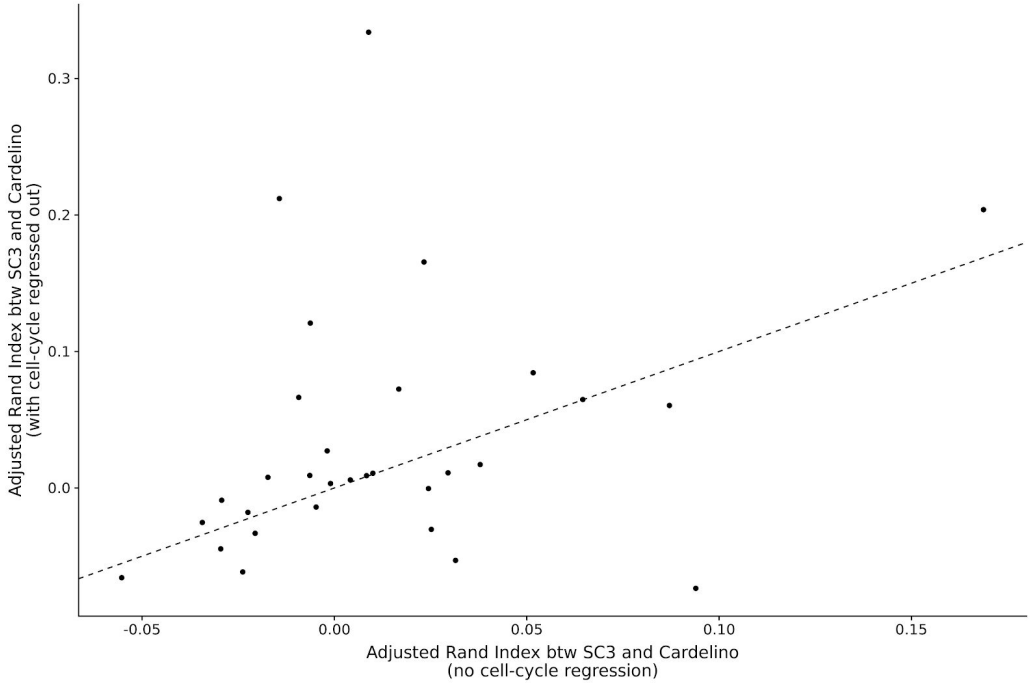


Figure S33. Adjusted Rand Index values comparing the clusters identified by SC3 and the clone assignments from cardelino when applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and when applying SC3 to gene expression values after regressing out *cyclone* cell cycle scores.

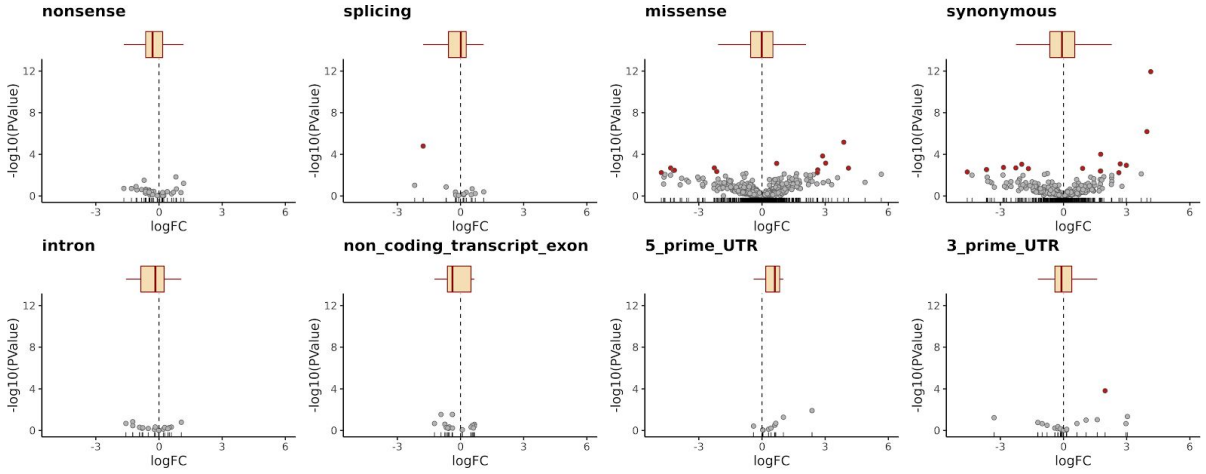


Figure S34. Direct effects of somatic variants on genes overlapping the variant. Volcano plot showing negative log P values versus $\log_2$-fold change from testing differential expression for genes with a somatic mutation between cells with the mutation and cells without the mutation, faceted by VEP annotation category (**Methods**). Each point represents a gene, and boxplots show the overall $\log_2$-fold change distribution for each annotation category. DE tests are conducted within each line (donor) separately, and results shown here are aggregated across 32 lines. Genes are categorised by simplified functional annotations from VEP of the somatic mutation, and genes significantly DE at an FDR threshold of 20% are shown in red.

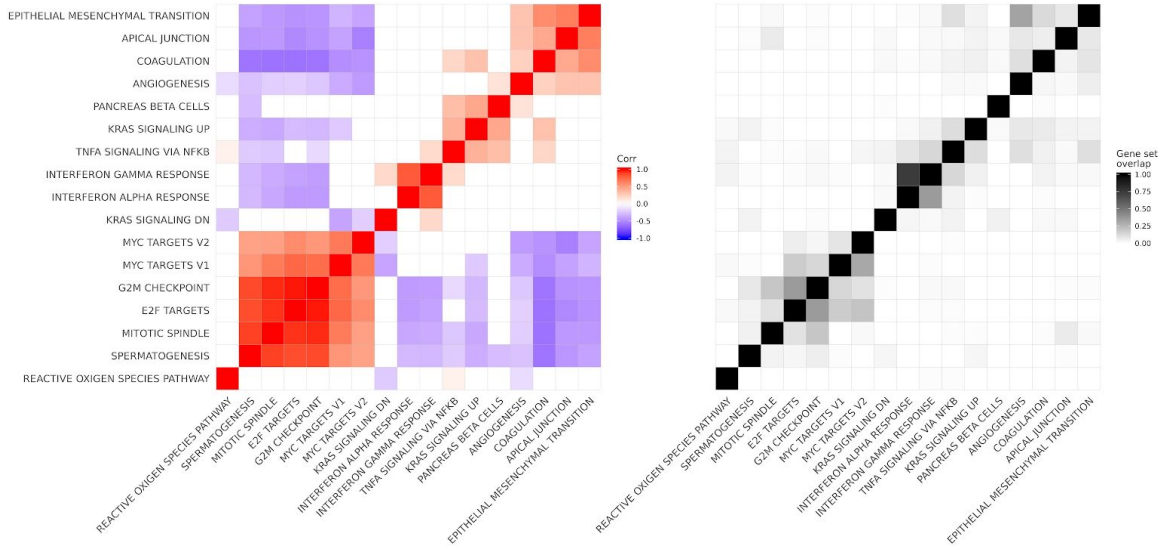


Figure S35. (left) Heatmap showing Spearman correlation between gene set enrichment results for the 16 most frequently enriched MSigDB Hallmark gene sets across 31 lines. Colour indicates the correlation between pairs of gene sets and is only shown if the correlation is significant ($P < 0.05$). (**right**) Heatmap showing proportion of overlap in genes between pairs of gene sets (matching those in left panel).

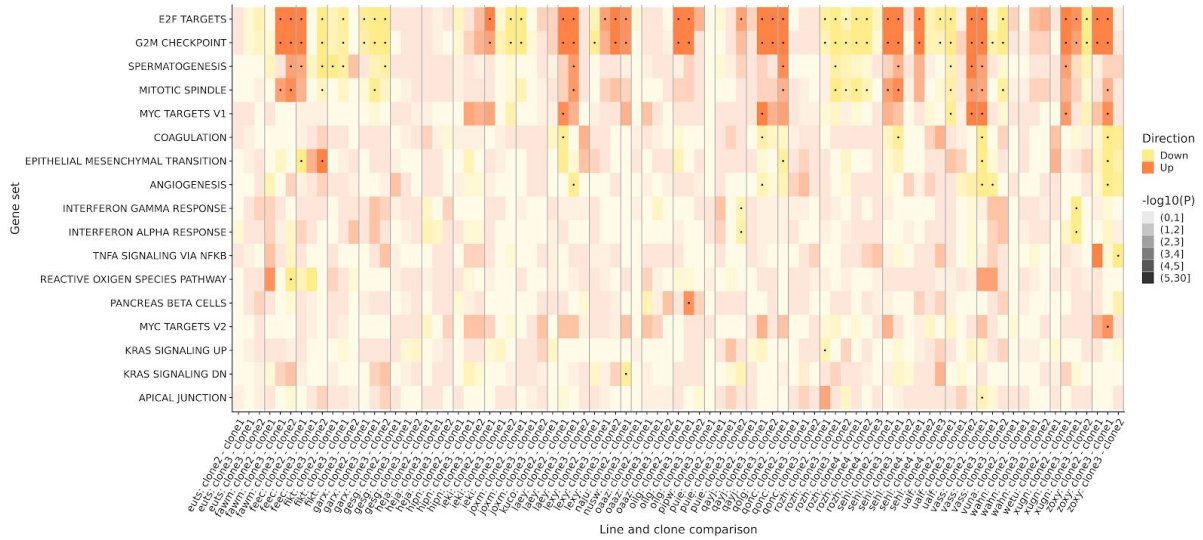


Figure S36. Heatmap showing the direction (first listed clone relative to second listed clone; in colour) and strength of enrichment (-log10(P) as degree of shading) for Hallmark gene sets tested with camera (Methods) for all pairwise comparisons between clones across 31 lines. Gene sets that are significantly enriched at an FDR threshold of 5% are indicated with dots. Gene sets are shown if significant in at least one line, and are ordered by number of lines in which they are significant.

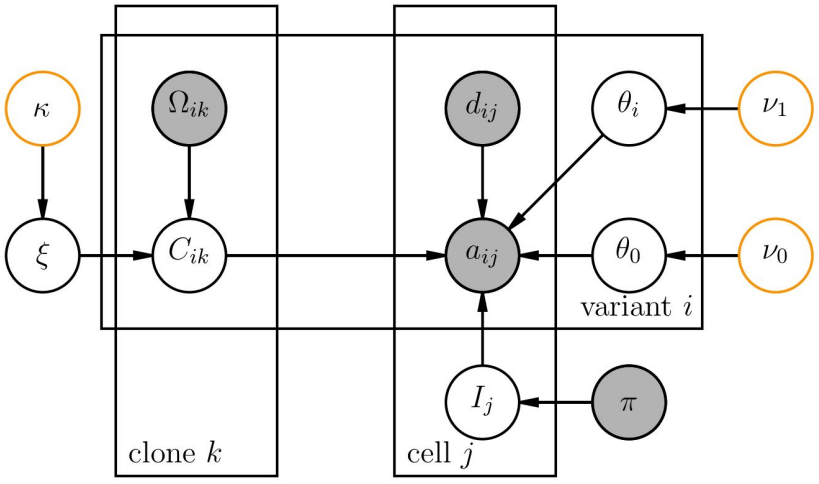


Figure S37. Graphical representation of the cardelino model. The clonal tree configuration matrix C is a random variable and follows a Bernoulli distribution encoded by an input tree configuration Ω that is provided to the model (e.g. estimated from bulk or single-cell DNA-seq data using existing methods such as Canopy) as well as an error rate ξ, which follows a beta prior distribution with hyper parameters κ. The indicator matrix I defines the assignment of cells to clones, which is another unknown variable, and assumed to follow a multinomial prior with fixed parameter π for each cell. The clone configuration C and cell identity I together encode the genotype $c_{i,Ij}$ of each variant i in each cell j. If $c_{i,Ij}$ is 1, the alternative allelic read count will follow a binomial distribution with gene specific parameter θ_i, otherwise with error related parameter θ_0. Both θ_i and θ_0 have a beta prior distribution, but with different parameters. Shaded nodes represent observed variables; unshaded nodes represent unknown variables; yellow circled nodes represent fixed hyper parameters.

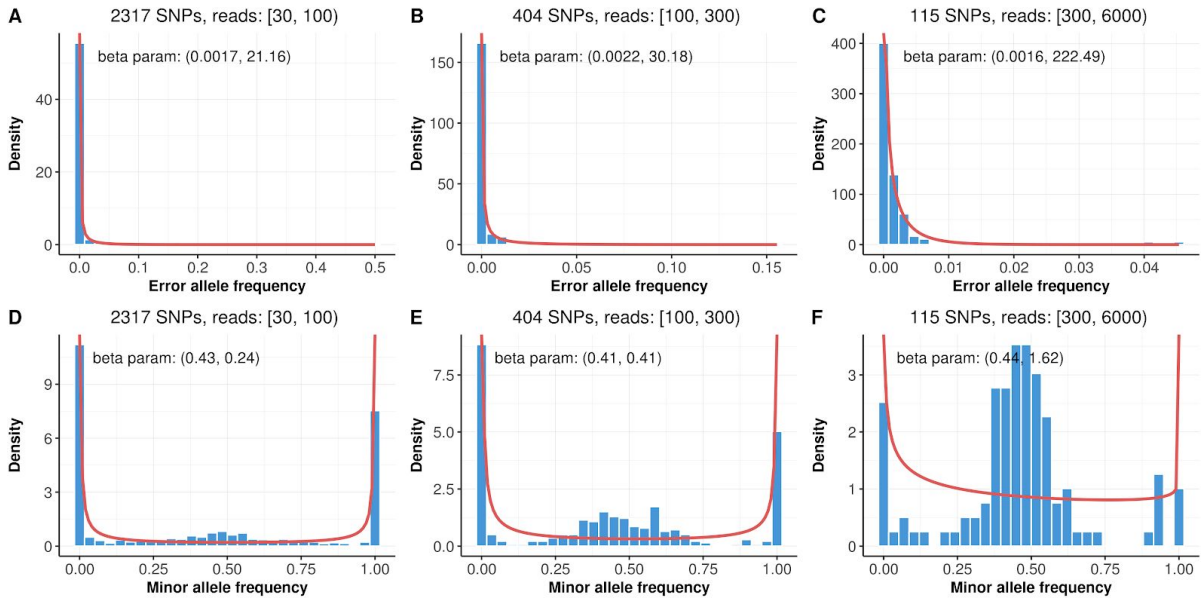


Figure S38. Estimated beta-binomial distribution of the "sequencing error rate" (theta0; **A-C**) and the alternative allele count rate given a variant is present (theta1; **D-F**) in single cells from germline heterozygous variants across three expression levels in donor vass. For each germline heterozygous variant, we select the cell with the highest expression to represent its minor allele frequency and the sequencing error rate, namely the fraction of reads from other alleles instead of either reference or alternative alleles. The parameters of beta-binomial distribution is obtained by a maximum likelihood estimate with VGAM R package. The Format of beta distribution parameters: (mean, shape1 + shape2).

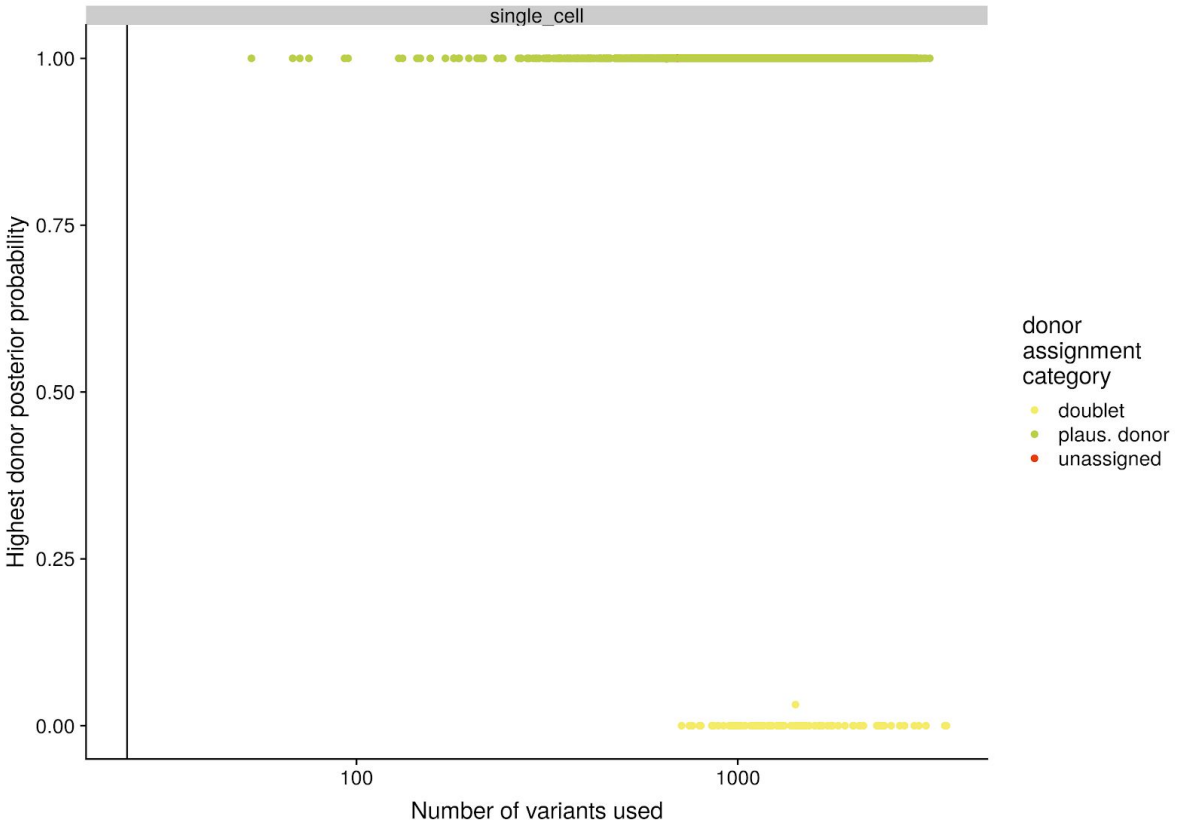


Figure S39. Donor identification results from cardelino for QC-passing cells for 32 fibroblast lines (*i.e.* donors) used to demultiplex cells from plates on which cells from three lines were pooled. The y-axis shows the highest posterior probability for donor assignment from cardelino (**Methods**) for a little over 2,000 cells passing QC using expression-based metrics (real Smart-seq2 data from our study; not simulated data). The donor ID results are emphatic, with posterior probabilities either very close to 1 or very close to zero, meaning that the model is very confident about assigning each cell either to a specific donor (*i.e.* line) or that the "cell" is actually doublet, or that it matches none of the plausible donors. The x-axis shows the number of germline variants with read coverage in the cells that were informative for donor assignment of the cell. Cells are coloured by donor assignment category: either "plausible donor" *(i.e.* a donor/line that was known to have been used on the processing plate), "doublet" (nominal single cells that have been inferred to be doublets) or "unassigned" (too few variants for assignment or posterior probability of assignment less than 0.95). NB: 21 unassigned cells are not visible due to overplotting by doublet cells.

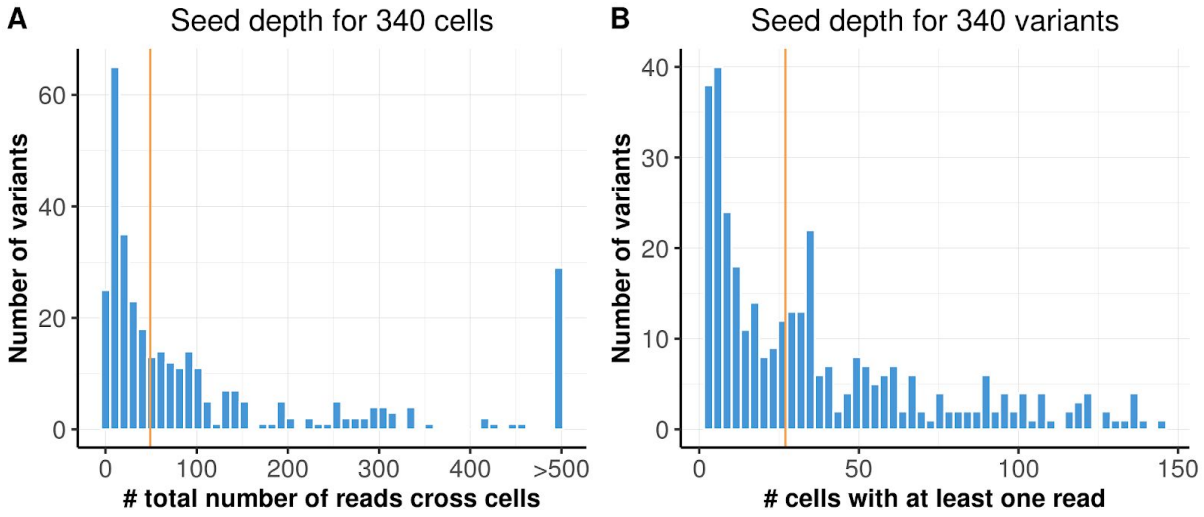


Figure S40. Summary of sequencing depths of 340 variants across a pool of 151 cells. **(A)** Histogram of total read counts on each variant across 151 cells, median number is shown in yellow; **(B)** Histogram of the number of cells with non-zero read coverage for each variant; median number is shown in yellow. This matrix is used as a seed to generate sequencing depths for simulations in Fig. 1(b-g) and Supp. Fig. S2.

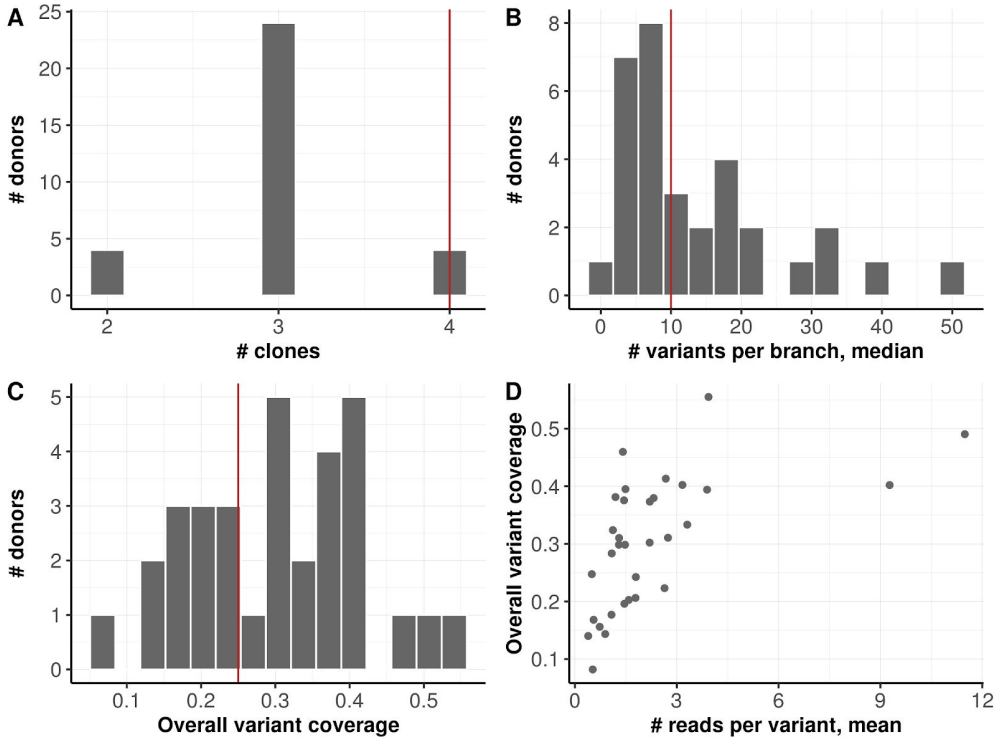


Figure S41. Distribution of key parameters in single cells assignment to clones across 32 donors: (**A**) number of clones inferred from bulk exome-seq data. (**B**) the median number of variants per clonal branch; (**C**) the overall coverage of variants, namely the fraction of variants with at least one read; (**D**) the scatter plot between the mean number of reads per variant per cell and the overall coverage of variants in the same donor. The default simulation parameters are highlighted with the red line.

Supplementary methods for "Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants"

Davis J. McCarthy†, Raghd Rostom†, Yuanhua Huang†, Daniel J. Kunz, Petr Danecek, Marc Jan Bonder, Tzachi Hagai, HipSci Consortium, Wenyi Wang, Daniel J. Gaffney, Benjamin D. Simons, Oliver Stegle, Sarah A. Teichmann

1 The Cardelino model

As input for Cardelino, we assume that an informative clonal structure configuration is first inferred from another data source such as deep, bulk exome-sequencing using a tool such as Canopy [1]. This inference yields an estimate of the number of clones present, K, clonal fractions $F = (f_1, .., f_K)$, where f_k denotes the relative prevalence of a given clone k ($\sum_{k=1}^{K} f_k = 1$), and a clonal tree configuration matrix C (an N-by-K binary matrix) for N variants and K clones, where $c_{i,k} = 1$ if somatic variant i is present in clone k and $c_{i,k} = 0$ otherwise. Given C and F, Cardelino aims to assign individual cells to one of K clones based on their expressed alleles using a probabilistic clustering model (see graphical representation in **Supp. Fig. S21**). From scRNA-seq data we extract, for each cell and variant that segregates between clones, the number of sequencing reads supporting the reference allele (reference read count) and the number of reads supporting the alternative allele (alternate read count). We denote the variant-by-cell matrix of alternate read counts by A and the variant-by-cell matrix of total read counts (sum of reference and alternate read counts) by D. Entries in A and D are therefore non-negative integers, with missing entries in the matrix D indicating zero read coverage for a given cell and variant.

The prior probability that cell j belongs to clone k could be taken as the clonal fraction f_k, but to avoid biasing cell assignment towards highly prevalent clones for cells with little read information (where the prior is more influential) we use a uniform prior F such that $P(I_j = k|F) = 1/K$ for all k. Note, the variable F is used to denote a uniform prior for convenience here, which can be different from the output of Canopy or another clonal inference method. Given this prior distribution, the posterior probability of cell j belonging to clone k can be expressed as:

$$P(I_j = k|\boldsymbol{a_j}, \boldsymbol{d_j}, C, F, \boldsymbol{\theta}) = \frac{P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = k, C, \boldsymbol{\theta})P(I_j = k|F)}{\sum_{t=1}^{K} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = t, C, \boldsymbol{\theta})P(I_j = t|F)}, \quad (1)$$

where I_j is the identity of the specific clone cell j is assigned to, and $\boldsymbol{a_j}$ and $\boldsymbol{d_j}$ are the observed alternate read count and total read count vectors, respectively, for variants 1 to N in cell j. The parameter vector $\boldsymbol{\theta}$ is a set of unknown parameters to model the allelic counts, which will be discussed in next section.

It is typically challenging to obtain a perfect clonal configuration from bulk exome-seq data only. Hence errors are likely to exist in the input configuration C. To account for errors in clonal configurations, we can use the input configuration as an informative prior (we use Ω

for this prior configuration) rather than as fixed and true. We can then learn the posterior configuration (we use C for consistency with other sections) and its corresponding error rate ξ. Therefore, we aim to have the full posterior distribution as follows,

$$P(\boldsymbol{\theta}, C, \xi | A, D, \Omega, F). \tag{2}$$

2 Modelling allelic expression

The core part of the Cardelino model is to model the alternate read count using a binomial model. For a given site in a given cell, there are two possibilities: the variant is "absent" in the clone a cell is assigned to (i.e. the cell is homozygous reference at that position) or the variant is "present" in the clone the cell is assigned to (i.e. the cell is heterozygous at that position), as encoded in the configuration matrix C. When considering the "success probability" $\boldsymbol{\theta}$ for the binomial model, where here a success is defined as observing an alternate read, we consider two alternative (sets of) parameters for each of these settings: θ_0 for homozygous reference alleles (variant absent), and $\boldsymbol{\theta_1} = \{\theta_1, ..., \theta_N\}$ for the case with heterozygous variants (variant present). Note, here we use a common parameter θ_0 for homozygous reference alleles in all variants, but $\theta_i, i \geq 1$ for each variant i to account for the gene specific level of allelic imbalance that causes the probability of observing alternate reads to differ from 0.5. Therefore, the allelic counts base model for the two genotypes can be written in the following binomial distributions,

$$p(a_{i,j}|d_{i,j}, h_{i,j}, \boldsymbol{\theta}) = \begin{cases} \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_0), & \text{if } h_{i,j} = 0. \\ \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_i), & \text{if } h_{i,j} = 1. \end{cases} \tag{3}$$

where $h_{i,j} = c_{i,I_j} \in \{0, 1\}$ is the genotype of variant i in cell j, which is encoded by clonal configuration C and cell identity I_j. Furthermore, the likelihood of cell j from clone k can be formalised as follows,

$$\begin{aligned} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = k, C, \boldsymbol{\theta}) &= \prod_{i=1}^{N} p(a_{i,j}|d_{i,j}, h_{i,j}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} \left\{ \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_i)^{c_{i,k}} \times \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_0)^{1-c_{i,k}} \right\} \end{aligned} \tag{4}$$

Then, we could have the likelihood of parameters $\boldsymbol{\theta} = \{\theta_0, \theta_1, ..., \theta_N\}$ to observe a full data set across M cells by marginalizing the mixture of cell assignments, as follows

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{j=1}^{M} \sum_{I_j=1}^{K} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j, C, \boldsymbol{\theta}) P(I_j|F). \tag{5}$$

Furthermore, we could view the the clonal assignment in a Bayesian way, and introduce informative prior distributions for unknown parameters $\boldsymbol{\theta}$. By multiplying the prior probability by the likelihood, we could have the posterior probability as follows,

$$\begin{aligned} P(\boldsymbol{\theta}|A, D, C, F, \boldsymbol{\nu}) &\propto P(\boldsymbol{\theta}|\boldsymbol{\nu}) \times \prod_{j=1}^{M} \sum_{I_j=1}^{K} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j, C, \boldsymbol{\theta}) P(I_j|F) \\ &= \texttt{Beta}(\theta_0|\alpha_0, \beta_0) \prod_{i=1}^{N} \texttt{Beta}(\theta_i|\alpha_1, \beta_1) \times \prod_{j=1}^{M} \sum_{I_j=1}^{K} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j, C, \boldsymbol{\theta}) P(I_j|F), \end{aligned} \tag{6}$$

where we use a beta prior distribution, a conjugate distribution to the binomial distribution, for each θ, and the hyperparameters $\boldsymbol{\nu} = \{\alpha_0, \beta_0, \alpha_1, \beta_1\}$ of the prior are learned from germline heterozygous variants.

Accounting for the uncertainty of $\boldsymbol{\theta}$, this unknown parameter can be marginalised in the posterior probability of clonal assignment, as follows,

$$P(I_j = k|\boldsymbol{a_j}, \boldsymbol{d_j}, C, F) = \int_{\boldsymbol{\theta}} P(I_j = k|\boldsymbol{a_j}, \boldsymbol{d_j}, C, F, \boldsymbol{\theta}) P(\boldsymbol{\theta}|A, D, C, F, \boldsymbol{\nu}) d\boldsymbol{\theta}. \tag{7}$$

3 Inference for the Cardelino model

In the above section, we defined the posterior probability of clonal assignment I and binomial parameters $\boldsymbol{\theta}$, the configuration matrix C and its error rate ξ. With conjugate prior distributions, a Gibbs sampler can be used to generate a set of samples following the posterior distribution.

In this Gibbs sampling algorithm, we sample cell assignment I, parameters $\boldsymbol{\theta}$, the configuration matrix C and its error rate ξ alternately. Given that three of these four unknown variables are fixed, the elements of the other parameter are conditionally independent. Therefore, given $\boldsymbol{\theta}$ and C, we could sample the clonal identity I_j via a categorical distribution, taking Eq(4,3), as follows

$$\begin{aligned} P(I_j = k|I_{-j}, A, D, C, F, \boldsymbol{\theta}) &= P(I_j = k|\boldsymbol{a_j}, \boldsymbol{d_j}, C, F, \boldsymbol{\theta}) \\ &\propto P(I_j = k|F) P(\boldsymbol{a_j}|I_j = k, \boldsymbol{d_j}, C, \boldsymbol{\theta}). \end{aligned} \tag{8}$$

Similarly, given the clonal identity I and configuration C in a previous step, $\theta_i, 0 \leq i \leq N$ are independent from each other, and the posterior probability in Eq(6) can be rewritten by inserting the base model in Eq(3) as follows,

$$\begin{aligned} P(\boldsymbol{\theta}|A, D, C, I, \boldsymbol{\nu}) &\propto \texttt{Beta}(\theta_0|\alpha_0, \beta_0) \prod_{i=1}^{N} \texttt{Beta}(\theta_i|\alpha_1, \beta_1) \\ &\times \prod_{j=1}^{M} \prod_{i=1}^{N} \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_0)^{1-c_{i,I_j}} \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_i)^{c_{i,I_j}} \\ &= \texttt{Beta}(\theta_0|\alpha_0, \beta_0) \prod_{j=1}^{M} \prod_{i=1}^{N} \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_0)^{1-c_{i,I_j}} \\ &\times \prod_{i=1}^{N} \left\{ \texttt{Beta}(\theta_i|\alpha_1, \beta_1) \prod_{j=1}^{M} \texttt{Binom}(a_{i,j}|d_{i,j}, \theta_i)^{c_{i,I_j}} \right\}. \end{aligned} \tag{9}$$

Therefore, we could sample individual θ values via a beta distribution as follows,

$$\theta_0|I \sim \texttt{beta}(\alpha_0 + u_0, \beta_0 + v_0); \qquad \theta_i|I \sim \texttt{beta}(\alpha_1 + u_i, \beta_1 + v_i), i > 0 \tag{10}$$

where

$$\begin{aligned} u_0 &= \sum_{i=1}^{N} \sum_{j=1}^{M} a_{i,j}(1 - c_{i,I_j}), & v_0 &= \sum_{i=1}^{N} \sum_{j=1}^{M} (d_{i,j} - a_{i,j})(1 - c_{i,I_j}), \\ u_i &= \sum_{j=1}^{M} a_{i,j} c_{i,I_j}, \ \ i > 0, & v_i &= \sum_{j=1}^{M} (d_{i,j} - a_{i,j}) c_{i,I_j}, \ \ i > 0. \end{aligned} \tag{11}$$

Furthermore, given the cell assignment I and the binomial parameters $\boldsymbol{\theta}$ and the error rate ξ, we can obtain the distribution of the configuration C as follows,

$$P(C_{i,k} = 1|C_{-i,k}, A, D, I, F, \boldsymbol{\theta}, \xi) = \frac{|\Omega_{i,k} - \xi| \prod_{j=1}^{M} \mathbb{I}(I_j = k)\text{binom}(a_{i,j}|d_{i,j}, \theta_i)}{|\Omega_{i,k} - \xi| \prod_{j=1}^{M} \mathbb{I}(I_j = k)\text{binom}(a_{i,j}|d_{i,j}, \theta_i) + |\Omega_{i,k} - \xi - 1| \prod_{j=1}^{M} \mathbb{I}(I_j = k)\text{binom}(a_{i,j}|d_{i,j}, \theta_0)} \tag{12}$$

Given the configuration C, we can also have the distribution of the error rate ξ. Here, we introduce a conjugate prior beta distribution with hyper-parameter κ_0, κ_1, hence we can write the posterior of ξ as follows,

$$P(\xi|C, \Omega, \kappa_0, \kappa_1) = \text{beta}(\kappa_0 + \sum_{i,k} \mathbb{I}(\Omega_{i,k} \neq C_{i,k}), \kappa_1 + \sum_{i,k} \mathbb{I}(\Omega_{i,k} = C_{i,k})) \tag{13}$$

Now, based on Eq (8-13), we could sample the full joint distribution of I, θ, C and ξ with Gibbs sampling in the following Algorithm 1.

Algorithm 1: Gibbs sampling for Cardelino model

```
Initialize θ = {θ_0, θ_1, ..., θ_N}
for t = 1 to H do
    for j = 1 to M do
        Sample: I_j = k|I_{-j}, A, D, C, F, θ with Eq(8)
    for i = 0 to N do
        Sample: θ_i|I, A, D, C, θ_{-i} with Eq (10)
    for i = 0 to N do
        for k = 1 to K do
            Sample: C_{i,k} = 1|C_{-i,k}, A, D, I, F, θ, ξ with Eq (12)
    Sample: ξ|C, Ω, κ_0, κ_1 with Eq (13)
```

In practice, we could sample 3,000 iterations and check the convergence with Geweke's convergence diagnostic (Z score) by using the first 10% and the last 50% iterations of the sampled chain. If $|Z| > 2$, then 100 more iterations will be added until the criterion is passed. Usually, this algorithm converges very quickly, even with as few as 100 iterations in some cases.

4 Inference with the EM algorithm to assign cells to donors

With a couple of tweaks the Cardelino model described above is also useful for assigning cells to the donor from which they originate in experimental settings where cells from multipled donors are pooled together before they are assayed ("multiplexed"). For the task of assigning cells to donors of origin rather than clone, we assume that the clonal tree configuration is fixed (here we interpret the "clonal tree configuration" as the reference genotypes of the donors, which we have access to), and all sites have a common parameter when variant is "present", i.e., $\theta_1 = \theta_2 = ... = \theta_N$. For simplicity, we use θ_1 to denote this shared parameter and ignore the conflict with the symbol in the Cardelino model. Therefore, the alternative model only has two parameters θ_0 and θ_1, for the "success probability" for variant absent and present, respectively.

In this donor-assignment setting, an attractive alternative possibility for inference in the Cardelino model is to use the Expectation-Maximisation (EM) algorithm. The EM algorithm has the advantage of being much more computationally efficient than the Gibbs sampler described above. However, EM inference yields only point estimates of parameter values and will

lose the uncertainty in the parameters for clonal assignment. Consequently, it can suffer from over-fitting if there are very few sequencing reads, especially in lowly expressed genes. Therefore, for EM inference it is important to use a single parameter for all variants and turn off the gene specific parameters in original Eq(3) to retain sufficient reads for a robust point estimate.

This setting proves very useful in assigning cells to donors given genotypes in multiplexed experiments, where the statistical framework is fundamentally the same but the error in the genotypes is much lower than from a clonal tree, and the large number of variants benefits from the high computational efficiency of the EM algorithm. Here, we introduce the algorithm with all $\theta_i, 1 \leq i \leq N$ turned into a single shared parameter θ_1; all equations in above sections still hold. In the real data analysis in the main text, we use this EM inference method to assign cells to donors from our three-donor multiplexed experimental design.

In order to maximise the likelihood in Eq(5) (or log likelihood for convenience), let us first rewrite the likelihood of assigning a single cell j to a certain clone k by extending the binomial probability as follows,

$$\begin{aligned} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = k, C, \boldsymbol{\theta}) &= \prod_{i=1}^{N} P(a_{i,j}|d_{i,j}, \theta, c_{i,k}) = \prod_{i=1}^{N} \mathcal{B}(a_{i,j}; d_{i,j}, \theta_{c_{i,k}}) \\ &= w_j \times \theta_0^{S^1_{j,k}} \times (1-\theta_0)^{S^2_{j,k}} \times \theta_1^{S^3_{j,k}} \times (1-\theta_1)^{S^4_{j,k}}, \end{aligned} \tag{14}$$

where $w_j = \prod_{i=1}^{N} \binom{d_{i,j}}{a_{i,j}}$ is a product of binomial coefficients. $S^1_{j,k}, S^2_{j,k}, S^3_{j,k}, S^4_{j,k}$ are the summarized read counts of alternative and reference alleles in genotypes without or with variant, respectively, as follows,

$$\begin{aligned} S^1_{j,k} &= \sum_{i=1}^{N} a_{i,j}\mathbb{I}(c_{i,k} = 0), \qquad S^2_{j,k} = \sum_{i=1}^{N} (d_{i,j} - a_{i,j})\mathbb{I}(c_{i,k} = 0), \\ S^3_{j,k} &= \sum_{i=1}^{N} a_{i,j}\mathbb{I}(c_{i,k} = 1), \qquad S^4_{j,k} = \sum_{i=1}^{N} (d_{i,j} - a_{i,j})\mathbb{I}(c_{i,k} = 1). \end{aligned} \tag{15}$$

These values can be equivalently taken from dot products of matrices $S^1 = A^\top(1-C)$, $S^2 = (D-A)^\top(1-C)$, $S^3 = A^\top C$, and $S^4 = (D-A)^\top C$.

Now, we can estimate the clonal assignment I_j and the parameters $\boldsymbol{\theta} = \{\theta_0, \theta_1\}$ with an EM algorithm. In the initialization, we set the parameter θ randomly. Then we iterate the E step and M step in the EM algorithm. In the E-step, given the parameter in the previous step, we calculate the posterior of the cell assignment

$$\gamma_{j,k} = P(I_j = k|\boldsymbol{a_j}, \boldsymbol{d_j}, C, F, \theta) = \frac{P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = k, C, \boldsymbol{\theta})P(I_j = k|F)}{\sum_{t=1}^{K} P(\boldsymbol{a_j}|\boldsymbol{d_j}, I_j = t, C, \boldsymbol{\theta})P(I_j = t|F)}, \tag{16}$$

which is often called component responsibility in the EM algorithm. In the M-step, given the posterior of cell assignment, we optimize the parameter to maximize the likelihood. By setting the derivation of the log likelihood Eq (5) (taking Eq (14)) to 0, we could have the following condition to satisfy,

$$\frac{\log \mathcal{L}(\theta)}{\theta_0} = \sum_{j=1}^{M}\sum_{k=1}^{K} \gamma_{j,k}\left[\frac{S^1_{j,k}}{\theta_0} - \frac{S^2_{j,k}}{1-\theta_0}\right] = 0. \tag{17}$$

Therefore, we can have a closed form solution for θ_0 (and θ_1 similarly) as follows,

$$\theta_0 = \frac{\sum_{j=1}^{M}\sum_{k=1}^{K} \gamma_{j,k} S^1_{j,k}}{\sum_{j=1}^{M}\sum_{k=1}^{K} \gamma_{j,k}(S^1_{j,k} + S^2_{j,k})} \qquad \theta_1 = \frac{\sum_{j=1}^{M}\sum_{k=1}^{K} \gamma_{j,k} S^3_{j,k}}{\sum_{j=1}^{M}\sum_{k=3}^{K} \gamma_{j,k}(S^3_{j,k} + S^4_{j,k})}. \tag{18}$$

Here, we summarize the EM algorithm for the cell assignment and parameter estimate in the following Algorithm 2. To end the algorithm, we could check if the improvement of the log likelihood is lower than a threshold or set a fixed number of iterations (e.g. 100 iterations are sufficient in many cases).

Algorithm 2: EM algorithm for cell assignments to clones

Initialize $\theta = \{\theta_0, \theta_1\}$ and evaluate $\log \mathcal{L}(\theta)$
while *not converged* **do**
  E step: Calculate $\gamma_{j,k}$ with current parameters
  $\gamma_{j,k} = \frac{P(A_j|I_j=k,D_j,C,F,\theta)P(I_j=k)}{\sum_{t=1}^{K} P(A_j|I_j=t,D_j,C,F,\theta)P(I_j=t)}$
  M step: Maximizing likelihood on parameters with current responsibilities
  $\theta_0^{\texttt{new}} = \frac{\sum_{j=1}^{M}\sum_{k=1}^{K}\gamma_{j,k}S_{j,k}^{1}}{\sum_{j=1}^{M}\sum_{k=1}^{K}\gamma_{j,k}(S_{j,k}^{1}+S_{j,k}^{2})}$; $\theta_1^{\texttt{new}} = \frac{\sum_{j=1}^{M}\sum_{k=1}^{K}\gamma_{j,k}S_{j,k}^{3}}{\sum_{j=1}^{M}\sum_{k=3}^{K}\gamma_{j,k}(S_{j,k}^{3}+S_{j,k}^{4})}$
  Update $\log \mathcal{L}(\theta)$ and check convergence
return $\theta, \boldsymbol{\gamma}, \log \mathcal{L}(\theta)$

In addition, the binomial distribution can be switched into simpler Bernoulli model by setting a threshold s (e.g. 1) as $\hat{a}_{i,j} = \mathbb{I}(a_{i,j} \geq s)$ and $\hat{d}_{i,j} = \mathbb{I}(d_{i,j} \geq s)$, and all above equations and inference methods remain applicable. The Bernoulli base model can be useful when the sequencing coverage is highly even, e.g., in scDNA-seq [2] or when the variance of allelic expression is extremely high.

References

[1] Yuchao Jiang, Yu Qiu, Andy J Minn, and Nancy R Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537, 2016.

[2] Andrew Roth, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A Smith, Cydney B Nielsen, Jessica N McAlpine, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature Methods*, 13:573, may 2016.

Appendix D

Manuscript: *Gene expression variability across cells and species shapes innate immunity*

ARTICLE

https://doi.org/10.1038/s41586-018-0657-2

Gene expression variability across cells and species shapes innate immunity

Tzachi Hagai[1,2]*, Xi Chen[1], Ricardo J. Miragaia[1,3], Raghd Rostom[1,2], Tomás Gomes[1], Natalia Kunowska[1], Johan Henriksson[1], Jong-Eun Park[1], Valentina Proserpio[4,5], Giacomo Donati[4,6], Lara Bossini-Castillo[1], Felipe A. Vieira Braga[1,7], Guy Naamati[2], James Fletcher[8], Emily Stephenson[8], Peter Vegh[8], Gosia Trynka[1], Ivanela Kondova[9], Mike Dennis[10], Muzlifah Haniffa[8,11], Armita Nourmohammad[12,13], Michael Lässig[14] & Sarah A. Teichmann[1,2,15]*

As the first line of defence against pathogens, cells mount an innate immune response, which varies widely from cell to cell. The response must be potent but carefully controlled to avoid self-damage. How these constraints have shaped the evolution of innate immunity remains poorly understood. Here we characterize the innate immune response's transcriptional divergence between species and variability in expression among cells. Using bulk and single-cell transcriptomics in fibroblasts and mononuclear phagocytes from different species, challenged with immune stimuli, we map the architecture of the innate immune response. Transcriptionally diverging genes, including those that encode cytokines and chemokines, vary across cells and have distinct promoter structures. Conversely, genes that are involved in the regulation of this response, such as those that encode transcription factors and kinases, are conserved between species and display low cell-to-cell variability in expression. We suggest that this expression pattern, which is observed across species and conditions, has evolved as a mechanism for fine-tuned regulation to achieve an effective but balanced response.

The innate immune response is a cell-intrinsic defence program that is rapidly upregulated upon infection in most cell types. It acts to inhibit pathogen replication while signalling the pathogen's presence to other cells. This programme involves the modulation of several cellular pathways, including production of antiviral and inflammatory cytokines, upregulation of genes that restrict pathogens, and induction of cell death[1,2].

An important characteristic of the innate immune response is the rapid evolution that many of its genes have undergone along the vertebrate lineage[3,4]. This rapid evolution is often attributed to pathogen-driven selection[5–7].

Another hallmark of this response is its high level of heterogeneity among responding cells: there is extensive cell-to-cell variability in response to pathogen infection[8,9] or to pathogen-associated molecular patterns (PAMPs)[10,11]. The functional importance of this variability is unclear.

These two characteristics—rapid divergence in the course of evolution and high cell-to-cell variability—seem to be at odds with the strong regulatory constraints imposed on the host immune response: the need to execute a well-coordinated and carefully balanced programme to avoid tissue damage and pathological immune conditions[12–15]. How this tight regulation is maintained despite rapid evolutionary divergence and high cell-to-cell variability remains unclear, but it is central to our understanding of the innate immune response and its evolution.

Here, we study the evolution of this programme using two cells types—fibroblasts and mononuclear phagocytes—in different mammalian clades challenged with several immune stimuli (Fig. 1a).

Our main experimental system uses primary dermal fibroblasts, which are commonly used in immunological studies[8,13]. We compare the response of fibroblasts from primates (human and macaque) and rodents (mouse and rat) to polyinosinic:polycytidylic acid (poly(I:C)), a synthetic double-stranded RNA (dsRNA; Fig. 1a, left). Poly(I:C) is frequently used to mimic viral infection as it rapidly elicits an antiviral response[16].

We comprehensively characterize the transcriptional changes between species and among individual cells in their innate immune response. We use population (bulk) transcriptomics to investigate transcriptional divergence between species, and single-cell transcriptomics to estimate cell-to-cell variability in gene expression. Using promoter sequence analyses along with chromatin immunoprecipitation with sequencing (ChIP–seq), we study how changes in the expression of each gene between species and across cells relate to the architecture of its promoter. Furthermore, we examine the relationship between cross-species divergence in gene coding sequence and expression and constraints imposed by host–pathogen interactions.

Additionally, we use a second system—bone marrow-derived mononuclear phagocytes from mouse, rat, rabbit and pig challenged with lipopolysaccharide (LPS), a commonly used PAMP of bacterial origin (Fig. 1a, right).

Together, these two systems provide insights into the architecture of the immune response across species, cell types and immune challenges.

Transcriptional divergence in immune response

First, we studied the transcriptional response of fibroblasts to stimulation with dsRNA (poly(I:C)) across the four species (human, macaque,

[1]Wellcome Sanger Institute, Cambridge, UK. [2]EMBL- European Bioinformatics Institute, Cambridge, UK. [3]Centre of Biological Engineering, University of Minho, Braga, Portugal. [4]Department of Life Sciences and Systems Biology, University of Turin, Torino, Italy. [5]Italian Institute for Genomic Medicine (IIGM), Torino, Italy. [6]Molecular Biotechnology Center, University of Turin, Torino, Italy. [7]Open Targets, Wellcome Sanger Institute, Cambridge, UK. [8]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. [9]Division of Pathology and Microbiology, Animal Science Department, Biomedical Primate Research Centre, Rijswijk, The Netherlands. [10]Research Department, Public Health England, National Infection Service, Porton Down, UK. [11]Department of Dermatology and NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [12]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany. [13]Department of Physics, University of Washington, Seattle, WA, USA. [14]Institute for Biological Physics, University of Cologne, Cologne, Germany. [15]Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Cambridge, UK. *e-mail: tzachi@ebi.ac.uk; st9@sanger.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

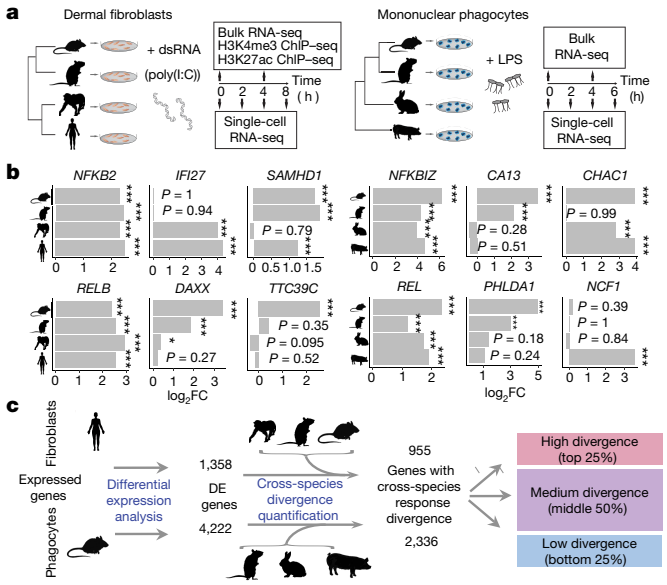


Fig. 1 | Response divergence across species in innate immune response. **a**, Study design. Left, primary dermal fibroblasts from mouse, rat, human and macaque stimulated with dsRNA or controls. Samples were collected for bulk and single-cell RNA-seq and ChIP–seq. Right, primary bone marrow-derived mononuclear phagocytes from mouse, rat, rabbit and pig stimulated with LPS or controls. Samples were collected for bulk and single-cell RNA-seq. **b**, Left, fold-change (FC) in dsRNA stimulation in fibroblasts for sample genes across species (edgeR exact test, based on $n = 6$, 5, 3 and 3 individuals from human, macaque, rat and mouse, respectively). Right, fold-change in LPS stimulation in phagocytes for sample genes across species (Wald test implemented in DESeq2, based on $n = 3$ individuals from each species). False discovery rate (FDR)-corrected P values are shown (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$). **c**, Top, estimating each gene's level of cross-species divergence in transcriptional response to dsRNA stimulation in fibroblasts. Using differential expression analysis, fold-change in dsRNA response was assessed for each gene in each species. We identified 1,358 human genes as differentially expressed (DE) (FDR-corrected $q < 0.01$), of which 955 had one-to-one orthologues across the four studied species. For each gene with one-to-one orthologues across all species, a response divergence measure was estimated using: response divergence $= \log[1/4 \times \sum_{i,j}(\log[\text{FC primate}_i] - [\log\text{FC rodent}_j])^2]$. Genes were grouped into low, medium and high divergence according to their response divergence values for subsequent analysis. Bottom, estimating each gene's level of cross-species divergence in LPS response in mononuclear phagocytes. A response divergence measure was estimated using: response divergence $= \log[1/3 \times \sum_{j}(\log[\text{FC pig}] - \log[\text{FC glire}_j])^2]$ (where glires are mouse, rat and rabbit).

rat and mouse). We generated bulk RNA-sequencing (RNA-seq) data for each species after 4 h of stimulation, along with respective controls (see Fig. 1a and Methods).

In all species, dsRNA treatment induced rapid upregulation of genes that encode expected antiviral and inflammatory products, including *IFNB*, *TNF*, *IL1A* and *CCL5* (see also Supplementary Table 3). Focusing on one-to-one orthologues, we performed correlation analysis between species and observed a similar transcriptional response (Spearman correlation, $P < 10^{-10}$ in all comparisons; Extended Data Fig. 1), as reported in other immune contexts[17–19]. Furthermore, the response tended to be more strongly correlated between closely related species than between more distantly related species, as in other expression programmes[20–24].

We characterized the differences in response to dsRNA between species for each gene, using these cross-species bulk transcriptomics data. While some genes, such as those encoding the NF-κB subunits RELB and NFKB2, respond similarly across species, other genes respond differently in the primate and rodent clades (Fig. 1b, left). For example, *Ifi27* (which encodes a restriction factor against numerous viruses) is strongly upregulated in primates but not in rodents, whereas *Daxx* (which encodes an antiviral transcriptional repressor) exhibits the opposite behaviour.

Similarly, in our second experimental system, which consists of lipopolysaccharide (LPS)-stimulated mononuclear phagocytes from mouse, rat, rabbit, and pig (Fig. 1b, right), some genes responded similarly across species (for example, *Nfkb2*), whereas others were highly upregulated only in specific clades (for example, *Phlda1*).

To quantify transcriptional divergence in immune responses between species, we focused on genes that were differentially expressed during the stimulation (see Methods). For simplicity, we refer to these genes as 'responsive genes' (Fig. 1c). In this analysis, we study the subset of these genes with one-to-one orthologues across the studied species. There are 955 such responsive genes in dsRNA-stimulated human fibroblasts and 2,336 in LPS-stimulated mouse phagocytes. We define a measure of response divergence by calculating the differences between the fold-change estimates while taking the phylogenetic relationship into account (Methods, Supplementary Figs. 1–7 and Supplementary Table 4).

For subsequent analyses, we split the 955 genes that were responsive in fibroblasts into three groups on the basis of their level of response divergence: (1) high-divergence dsRNA-responsive genes (the top 25% of genes with the highest divergence values in response to dsRNA across the four studied species); (2) low-divergence dsRNA-responsive genes (the bottom 25%); and (3) genes with medium divergence across species (the middle 50%; Fig. 1c). We performed an analogous procedure for the 2,336 LPS-responsive genes in phagocytes.

Promoter architecture of diverging genes

Next, we tested whether divergence in transcriptional responses is reflected in the conservation of promoter function and sequence. Using ChIP–seq, we profiled active histone marks in the fibroblasts of all species. The presence of trimethylation of lysine 4 on histone H3 (H3K4me3) in promoter regions of high-divergence genes was significantly less conserved between humans and rodents than was the presence of H3K4me3 in promoters of low-divergence genes (Extended Data Fig. 2).

We then used the human H3K4me3 ChIP–seq peaks to define active promoter regions of the responsive genes in human fibroblasts. The density of transcription factor binding motifs (TFBMs) was significantly higher in the active promoter regions of high-divergence genes than in low-divergence genes (Fig. 2a). Notably, when comparing the conservation of the core promoter regions in high- versus low-divergence dsRNA-responsive genes, we found that genes that diverge highly in response to dsRNA show higher sequence conservation in this region (Fig. 2b).

This unexpected discordance may be related to the fact that promoters of high- and low-divergence genes have distinctive architectures, associated with different constraints on promoter sequence evolution[18,25,26]. Notably, promoters containing TATA-box elements tend to have most of their regulatory elements in regions immediately upstream of the transcription start site (TSS). These promoters are thus expected to be more conserved. The opposite is true for CpG island (CGI)[26,27] promoters. Indeed, we found that TATA-boxes are associated with higher transcriptional divergence, while genes with CGIs diverge more slowly, both in fibroblasts and phagocytes (Fig. 2c; Extended Data Fig. 3). Thus, a promoter architecture enriched in TATA-boxes and depleted of CGIs is associated with higher transcriptional divergence, while entailing higher sequence conservation upstream of these genes[18,26,27].

Transcriptional divergence of cytokines

We next investigated whether different functional classes among responsive genes are characterized by varying levels of transcriptional divergence. To this end, we divided responsive genes into categories according to function (such as cytokines, transcriptional factors and kinases) or the processes in which they are known to be involved (such as apoptosis or inflammation).

© 2018 Springer Nature Limited. All rights reserved.

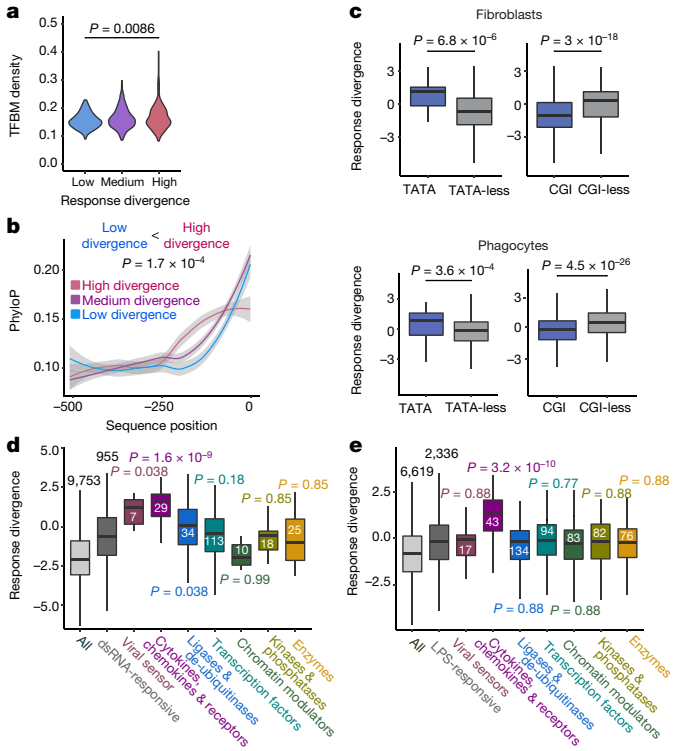


Fig. 2 | Transcriptionally divergent genes have unique functions and promoter architectures. a, TFBM density in active promoters and response divergence. For each gene studied in fibroblast dsRNA stimulation, the total number of TFBM matches in its H3K4me3 histone mark was divided by the length of the mark (human marks were used; $n = 879$ differentially expressed genes with ChIP–seq data). High-divergence genes have higher TFBM density than low-divergence genes (one-sided Mann–Whitney test). **b**, Promoter sequence conservation and response divergence in fibroblast dsRNA stimulation. Sequence conservation values are estimated with phyloP7 for 500 base pairs upstream of the transcription start site (TSS) of the human gene. Mean conservation values of each of the 500 base pairs upstream of the TSS are shown for high-, medium- and low-divergence genes ($n = 840$ genes). Genes that are highly divergent have higher sequence conservation (one-sided Kolmogorov–Smirnov test). The 95% confidence interval for predictions from a linear model computed by geom_loess function is shown in grey. **c**, Comparison of divergence in response of genes with and without a TATA-box and a CGI in fibroblast dsRNA stimulation and phagocyte LPS stimulation. TATA-box matches and CGI overlaps were computed with respect to the TSS of human genes in fibroblasts ($n = 955$ genes), and to the TSS of mouse genes in phagocytes ($n = 2{,}336$). **d**, Distributions of divergence values of 9,753 expressed genes in fibroblasts, 955 dsRNA-responsive genes and different functional subsets of the dsRNA-responsive genes (each subset is compared with the set of 955 genes using a one-sided Mann–Whitney test and FDR-corrected *P* values are shown). **e**, Distributions of divergence values of 6,619 expressed genes in phagocytes, 2,336 LPS-responsive genes and different functional subsets of the LPS-responsive genes (each subset is compared with the set of 2,336 genes using a one-sided Mann–Whitney test and FDR-corrected *P* values are shown). Violin plots show the kernel probability density of the data. Boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the interquartile range (IQR).

Genes related to cellular defence and inflammation—most notably cytokines, chemokines and their receptors (hereafter 'cytokines')—tended to diverge in response significantly faster than genes involved in apoptosis or immune regulation (chromatin modulators, transcription factors, kinases and ligases) (Fig. 2d, e, Extended Data Fig. 4, Supplementary Fig. 1).

Cytokines also had a higher transcriptional range in response to immune challenge (a higher fold-change). Regressing the fold-change from the divergence estimates resulted in reduction of the relative divergence of cytokines versus other responsive genes, but the difference still remained (Supplementary Fig. 2). Cytokine promoters are enriched in TATA-boxes (17% versus 2.5%, $P = 1.1 \times 10^{-3}$, Fisher's exact test) and depleted of CGIs (14% versus 69%, $P = 1.6 \times 10^{-9}$), suggesting that this promoter architecture is associated both with greater differences between species (response divergence) and larger changes between conditions (transcriptional range).

Cell-to-cell variability in immune response

Previous studies have shown that the innate immune response displays high variability across responding cells[28,29]. However, the relationship between cell-to-cell transcriptional variability and response divergence between species is not well understood.

To study heterogeneity in gene expression across individual cells, we performed single-cell RNA-seq in all species in a time course following immune stimulation. We estimated cell-to-cell variability quantitatively using an established measure for variability: distance to median (DM)[30].

We found a clear trend in which genes that were highly divergent in response between species were also more variable in expression across individual cells within a species (Fig. 3a). The relationship between rapid divergence and high cell-to-cell variability held true in both the 955 dsRNA-responsive genes in fibroblasts and the 2,336 LPS-responsive genes in phagocytes. This can be observed across the stimulation time points and in different species (Extended Data Figs. 5, 6). We analysed in depth the relationship between transcriptional divergence and cell-to-cell variability by using additional immune stimulation protocols (Supplementary Figs. 8, 9), and different experimental and computational approaches (Extended Data Fig. 7, Supplementary Figs. 10–13). Notably, the trends we observed are not a result of technical biases due to low expression levels in either the bulk or the single-cell RNA-seq data (Supplementary Figs. 14, 15).

Next, we examined the relationship between the presence of promoter elements (CGIs and TATA-boxes) and a gene's cell-to-cell variability. Genes that are predicted to have a TATA-box in their promoter had higher transcriptional variability, whereas CGI-containing genes tended to have lower variability (Fig. 3b), in agreement with previous findings[31]. Thus, both transcriptional variability between cells (Fig. 3b) and transcriptional divergence between species (Fig. 2c) are associated with the presence of specific promoter elements.

Transcriptional variability of cytokines

We subsequently compared the response divergence across species with the transcriptional cell-to-cell variability of three groups of responsive genes with different functions: cytokines, transcription factors, and kinases and phosphatases (hereafter 'kinases'; Fig. 3c, Extended Data Fig. 8). In contrast to kinases and transcription factors, many cytokines display relatively high levels of cell-to-cell variability (Extended Data Fig. 9), being expressed only in a small subset of responding cells (Extended Data Fig. 10). This has previously been reported for several cytokines[29]. For example, IFNB is expressed in only a small fraction of cells infected with viruses or challenged with various stimuli[8,11,32]. Here, we find that cells show high levels of variability in expression of cytokines from several families (for example, IFNB, CXCL10 and CCL2).

Cell-to-cell variability of cytokines remains relatively high in comparison to kinases and transcription factors during a time course of 2, 4 and 8 h after dsRNA stimulation of fibroblasts (Extended Data Fig. 9). This pattern is similar across species, and can also be observed in LPS-stimulated phagocytes (Extended Data Fig. 9). Thus, the high variability of cytokines and their expression in a small fraction of stimulated cells across all time points is evolutionarily conserved.

Cytokines tended to be co-expressed in the same cells, raising the possibility that their expression is coordinated (see Supplementary Information and Supplementary Fig. 16). We also identified genes whose expression was correlated with cytokines in human fibroblasts and showed that their orthologues tend to be co-expressed with cytokines in other species. This set is enriched with genes known to be involved in cytokine regulation (Supplementary Table 5).

© 2018 Springer Nature Limited. All rights reserved.

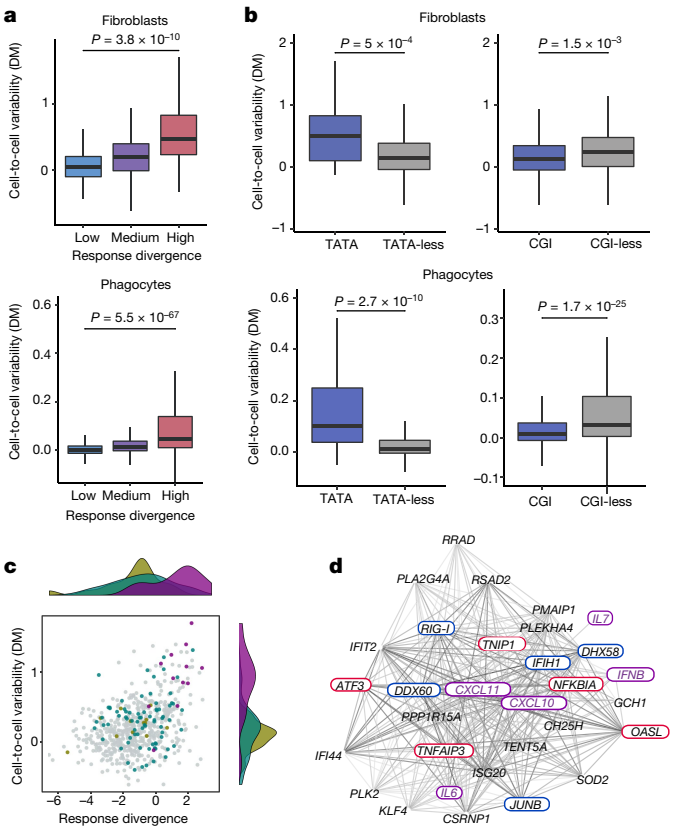


Fig. 3 | Cell-to-cell variability in immune response corresponds to response divergence. a, Comparison of divergence in response across species with transcriptional variability between individual cells. Top, fibroblast dsRNA stimulation (variability measured in $n = 55$ human cells, following 4 h dsRNA stimulation). Bottom, phagocyte LPS stimulation (variability measured in $n = 3{,}293$ mouse cells, following 4 h LPS stimulation). Genes classified as high-, medium- or low-divergence according to level of response divergence. Cell-to-cell variability values of high-divergence genes were compared with those of low-divergence genes (one-sided Mann–Whitney test). **b**, Comparison of cell-to-cell variability of genes with and without a TATA-box and a CGI, in fibroblast dsRNA stimulation and phagocyte LPS stimulation (one-sided Mann–Whitney test). Cell-to-cell variability values are from DM estimations of human fibroblasts stimulated with dsRNA for 4 h ($n = 55$ cells) and from mouse phagocytes stimulated with LPS for 4 h ($n = 3{,}293$ cells). **c**, Scatter plot showing divergence in response to dsRNA in fibroblasts across species and transcriptional cell-to-cell variability in human cells following 4 h of dsRNA stimulation ($n = 684$ dsRNA-responsive genes). Purple, cytokines; green, transcription factors; beige, kinases. The distributions of divergence and variability values of these groups are shown above and to the right of the scatter plot, respectively. **d**, A network showing genes that correlate positively in expression with the chemokine gene *CXCL10* across cells (Spearman correlation, $\rho > 0.3$), in at least two species (one of which is human), following dsRNA treatment in fibroblasts (based on $n = 146$, 74, 175 and 170 human, macaque, rat and mouse cells, respectively). Purple, cytokines; red, positive regulators of cytokine expression; blue, negative regulators. Colours of lines, from light to dark grey, reflect the number of species in which this pair of genes was correlated. Boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 × IQR.

As an example, we focused on the genes whose expression is positively correlated with the chemokine CXCL10 in at least two species (Fig. 3d). This set includes four cytokines co-expressed with CXCL10 (in purple), as well as known positive regulators of the innate immune response and cytokine production (in blue), such as the viral sensors IFIH1 (also known as MDA5) and RIG-I (also known as DDX58) This is in agreement with previous evidence that IFNB expression is limited to cells in which important upstream regulators are expressed at sufficiently high levels[8,11,32]. Here, we show that this phenomenon of co-expression with upstream regulators applies to a wider set of cytokines and is conserved across species. Notably, cytokines were co-expressed not only with their positive regulators but also with genes that are known to act as negative regulators of cytokine expression or cytokine signalling (in red), suggesting that cytokine expression and function is tightly controlled at the level of individual cells.

The evolutionary landscape of innate immunity

Many immune genes, including several cytokines and their receptors, have been shown to evolve rapidly in coding sequence[3,33]. However, it is not known how divergence in coding sequence relates to transcriptional divergence in innate immune genes. Using the set of 955 dsRNA-responsive genes in fibroblasts, we assessed coding sequence evolution in the three subsets of low-, medium- and high-divergence genes (as defined in Fig. 1c).

We compared the rate at which genes evolved in their coding sequences with their response divergence by considering the ratio of non-synonymous (dN) to synonymous (dS) nucleotide substitutions. Genes that evolved rapidly in transcriptional response had higher coding sequence divergence (higher dN/dS values) than dsRNA-responsive genes with low response divergence (Fig. 4a).

Rapid gene duplication and gene loss have been observed in several important immune genes[34–39] and are thought to be a result of pathogen-driven pressure[40,41]. We therefore tested the relationship between a gene's divergence in response and the rate at which the gene's family has expanded and contracted in the course of vertebrate evolution. We found that transcriptionally divergent dsRNA-responsive genes have higher rates of gene gain and loss (Fig. 4b) and consequently are also evolutionarily younger (Fig. 4c, Supplementary Fig. 17).

Previous reports have suggested that proteins encoded by younger genes tend to have fewer protein–protein interactions (PPIs) within cells[42]. Indeed, we found that rapidly diverging genes tend to have fewer PPIs (Fig. 4d). Together, these results suggest that transcriptionally divergent dsRNA-responsive genes evolve rapidly through various mechanisms, including fast coding sequence evolution and higher rates of gene loss and duplication events, and that their products have fewer interactions with other cellular proteins than those of less divergent genes.

The interaction between pathogens and the host immune system is thought to be an important driving force in the evolution of both sides. We therefore investigated the relationship between transcriptional divergence and interactions with viral proteins by compiling a data set of known host–virus interactions in humans[6,43,44]. Notably, genes whose products had no known viral interactions showed higher response divergence than genes encoding proteins with viral interactions (Fig. 4e). Furthermore, the transcriptional divergence of genes targeted by viral immunomodulators[45]—viral proteins that subvert the host immune system—was lower still (Fig. 4e). These observations suggest that viruses have evolved to modulate the immune system by interacting with immune proteins that are relatively conserved in their response. Presumably, these genes cannot evolve away from viral interactions, unlike host genes that are less constrained[46].

The summary of our results in Fig. 4f highlights the differences in both regulatory and evolutionary characteristics between cytokines and other representative dsRNA-responsive genes. Cytokines evolve rapidly through various evolutionary mechanisms and have higher transcriptional variability across cells. By contrast, genes that are involved in immune response regulation, such as transcription factors and kinases, are more conserved and less heterogeneous across cells. These genes encode proteins that have more interactions with other cellular proteins, suggesting that higher constraints are imposed on their evolution. This group of conserved genes is more often targeted by viruses, revealing a relationship between host–pathogen dynamics and the evolutionary landscape of the innate immune response.

Discussion

Here, we have charted the evolutionary architecture of the innate immune response. We show that genes that diverge rapidly between

© 2018 Springer Nature Limited. All rights reserved.

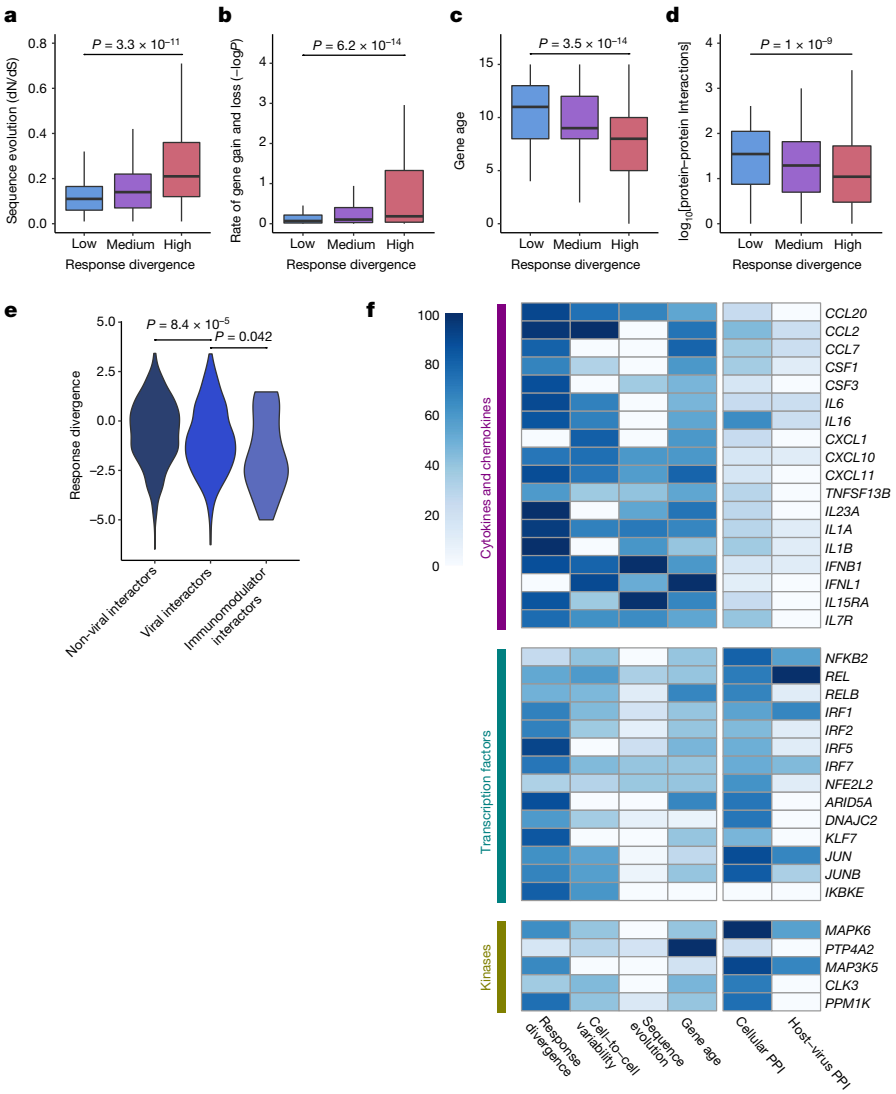


Fig. 4 | Relationship of response divergence and other evolutionary modes. a–**d**, dsRNA-responsive genes in fibroblasts are divided by level of response divergence into three groups, as in Fig. 1c. **a**, Coding sequence divergence, as measured using dN/dS values across 29 mammals. Higher dN/dS values denote faster coding sequence evolution ($n = 567$ genes). **b**, Rate at which genes were gained and lost within the gene family across the vertebrate clade (plotted as –log$P$). Higher values denote faster gene gain and loss rate ($n = 955$ genes). **c**, Evolutionary age (estimated with Panther7 phylogeny and Wagner reconstruction algorithm). Values denote the branch number with respect to human (distance from human in the phylogenetic tree); higher values indicate greater age ($n = 931$ genes). **d**, Number of known physical interactions with other cellular proteins ($n = 955$ genes). **e**, Distribution of transcriptional response divergence values among dsRNA-responsive genes whose protein products do not interact with viral proteins, interact with at least one viral protein, or interact with viral immunomodulators ($n = 648$, 307 and 25 genes, respectively). **a**–**e**, One-sided Mann–Whitney tests. **f**, A scaled heat map showing values of response divergence (as in Fig. 1c), cell-to-cell variability (as in Fig. 3a), coding sequence divergence (dN/dS values, as in **a**), gene age (as in **c**; younger genes have darker colours), number of cellular PPIs (as in **d**) and number of host–virus interactions (as in **e**), for example genes from three functional groups: cytokines, transcription factors, and kinases. Values are shown in a normalized scale between 0 and 100, with the gene with the highest value assigned a score of 100. Missing values are shown in white. Boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 × IQR. Violin plots show the kernel probability density of the data.

species show higher levels of variability in their expression across individual cells than genes that diverge more slowly. Both of these characteristics are associated with a similar promoter architecture, enriched in TATA-boxes and depleted of CGIs. Notably, such promoter architecture is also associated with the high transcriptional range of genes during the immune response. Thus, transcriptional changes between conditions (stimulated versus unstimulated), species (transcriptional divergence), and individual cells (cell-to-cell variability) may all be mechanistically related to the same promoter characteristics. In yeast, TATA-boxes are enriched in promoters of stress-related genes, displaying rapid transcriptional divergence between species and high variability in expression[30,47]. This finding suggests intriguing analogies between the mammalian immune and yeast stress responses—two systems that have been exposed to continuous changes in external stimuli during evolution.

We have also shown that genes involved in regulation of the immune response—such as transcription factors and kinases—are relatively conserved in their transcriptional responses. These genes might be under stronger functional and regulatory constraints, owing to their roles in multiple contexts and pathways, which would limit their ability to evolve. This limitation could represent an Achilles' heel that is used by pathogens to subvert the immune system. Indeed, we found that viruses interact preferentially with conserved proteins of the innate immune response. Cytokines, on the other hand, diverge rapidly between species, owing to their promoter architecture and because they have fewer constraints imposed by intracellular interactions or additional

© 2018 Springer Nature Limited. All rights reserved.

non-immune functions. We therefore suggest that cytokines represent a successful host strategy to counteract rapidly evolving pathogens as part of the host–pathogen evolutionary arms race.

Cytokines also display high cell-to-cell variability and tend to be co-expressed with other cytokines and cytokine regulators in a small subset of cells, and this pattern is conserved across species. As prolonged or increased cytokine expression can result in tissue damage[48–50], restriction of cytokine production to only a few cells may enable a rapid, but controlled, response across the tissue to avoid long-lasting and potentially damaging effects.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0657-2.

Received: 24 August 2017; Accepted: 17 August 2018;
Published online 24 October 2018.

1. Borden, E. C. et al. Interferons at age 50: past, current and future impact on biomedicine. *Nat. Rev. Drug Discov.* **6**, 975–990 (2007).
2. Iwasaki, A. A virological view of innate immune recognition. *Annu. Rev. Microbiol.* **66**, 177–196 (2012).
3. Nielsen, R. et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
4. Haygood, R., Babbitt, C. C., Fedrigo, O. & Wray, G. A. Contrasts between adaptive coding and noncoding changes during human evolution. *Proc. Natl Acad. Sci. USA* **107**, 7853–7857 (2010).
5. Fumagalli, M. et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* **7**, e1002355 (2011).
6. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
7. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).
8. Zhao, M., Zhang, J., Phatnani, H., Scheu, S. & Maniatis, T. Stochastic expression of the interferon-β gene. *PLoS Biol.* **10**, e1001249 (2012).
9. Avraham, R. et al. Pathogen cell-to-cell variability drives heterogeneity in host immune responses. *Cell* **162**, 1309–1321 (2015).
10. Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
11. Hwang, S. Y. et al. Biphasic RLR-IFN-β response controls the balance between antiviral immunity and cell damage. *J. Immunol.* **190**, 1192–1200 (2013).
12. Porritt, R. A. & Hertzog, P. J. Dynamic control of type I IFN signalling by an integrated network of negative regulators. *Trends Immunol.* **36**, 150–160 (2015).
13. Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nat. Rev. Immunol.* **14**, 36–49 (2014).
14. Brinkworth, J. F. & Barreiro, L. B. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Curr. Opin. Immunol.* **31**, 66–78 (2014).
15. Kobayashi, K. S. & Flavell, R. A. Shielding the double-edged sword: negative regulation of the innate immune system. *J. Leukoc. Biol.* **75**, 428–433 (2004).
16. Kumar, H., Kawai, T. & Akira, S. Pathogen recognition by the innate immune system. *Int. Rev. Immunol.* **30**, 16–34 (2011).
17. Barreiro, L. B., Marioni, J. C., Blekhman, R., Stephens, M. & Gilad, Y. Functional comparison of innate immune signaling pathways in primates. *PLoS Genet.* **6**, e1001249 (2010).
18. Schroder, K. et al. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proc. Natl Acad. Sci. USA* **109**, E944–E953 (2012).
19. Shay, T. et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl Acad. Sci. USA* **110**, 2946–2951 (2013).
20. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
21. Kalinka, A. T. et al. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814 (2010).
22. Khaitovich, P., Enard, W., Lachmann, M. & Pääbo, S. Evolution of primate gene expression. *Nat. Rev. Genet.* **7**, 693–702 (2006).
23. Levin, M. et al. The mid-developmental transition and the evolution of animal body plans. *Nature* **531**, 637–641 (2016).
24. Reilly, S. K. & Noonan, J. P. Evolution of gene regulation in humans. *Annu. Rev. Genomics Hum. Genet.* **17**, 45–67 (2016).
25. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* **38**, 830–834 (2006).
26. Haberle, V. & Lenhard, B. Promoter architectures and developmental gene regulation. *Semin. Cell Dev. Biol.* **57**, 11–23 (2016).
27. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).
28. Franz, K. M. & Kagan, J. C. Innate immune receptors as competitive determinants of cell fate. *Mol. Cell* **66**, 750–760 (2017).
29. Satija, R. & Shalek, A. K. Heterogeneity in immune responses: from populations to single cells. *Trends Immunol.* **35**, 219–229 (2014).
30. Newman, J. R. et al. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
31. Faure, A. J., Schmiedel, J. M. & Lehner, B. Systematic analysis of the determinants of gene expression noise in embryonic stem cells. *Cell Syst.* **5**, 471–484.e474 (2017).
32. Rand, U. et al. Multi-layered stochasticity and paracrine signal propagation shape the type-I interferon response. *Mol. Syst. Biol.* **8**, 584 (2012).
33. Fumagalli, M. & Sironi, M. Human genome variability, natural selection and infectious diseases. *Curr. Opin. Immunol.* **30**, 9–16 (2014).
34. Johnson, W. E. & Sawyer, S. L. Molecular evolution of the antiretroviral TRIM5 gene. *Immunogenetics* **61**, 163–176 (2009).
35. Choo, S. W. et al. Pangolin genomes and the evolution of mammalian scales and immunity. *Genome Res.* **26**, 1312–1322 (2016).
36. Braun, B. A., Marcovitz, A., Camp, J. G., Jia, R. & Bejerano, G. Mx1 and Mx2 key antiviral proteins are surprisingly lost in toothed whales. *Proc. Natl Acad. Sci. USA* **112**, 8036–8040 (2015).
37. Xu, L. et al. Loss of RIG-I leads to a functional replacement with MDA5 in the Chinese tree shrew. *Proc. Natl Acad. Sci. USA* **113**, 10950–10955 (2016).
38. Sackton, T. B., Lazzaro, B. P. & Clark, A. G. Rapid expansion of immune-related gene families in the house fly, *Musca domestica*. *Mol. Biol. Evol.* **34**, 857–872 (2017).
39. Brunette, R. L. et al. Extensive evolutionary and functional diversity among mammalian AIM2-like receptors. *J. Exp. Med.* **209**, 1969–1983 (2012).
40. Malfavon-Borja, R., Wu, L. I., Emerman, M. & Malik, H. S. Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes. *Proc. Natl Acad. Sci. USA* **110**, E583–E592 (2013).
41. Barber, M. F., Lee, E. M., Griffin, H. & Elde, N. C. Rapid evolution of primate type 2 immune response factors linked to asthma susceptibility. *Genome Biol. Evol.* **9**, 1757–1765 (2017).
42. Saeed, R. & Deane, C. M. Protein–protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics* **7**, 128 (2006).
43. Calderone, A., Licata, L. & Cesareni, G. VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res.* **43**, D588–D592 (2015).
44. Halehalli, R. R. & Nagarajaram, H. A. Molecular principles of human virus protein-protein interactions. *Bioinformatics* **31**, 1025–1033 (2015).
45. Pichlmair, A. et al. Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* **487**, 486–490 (2012).
46. Dyer, M. D., Murali, T. M. & Sobral, B. W. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.* **4**, e32 (2008).
47. Tirosh, I. & Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* **18**, 1084–1091 (2008).
48. Crow, Y. J. & Manel, N. Aicardi-Goutières syndrome and the type I interferonopathies. *Nat. Rev. Immunol.* **15**, 429–440 (2015).
49. Hall, J. C. & Rosen, A. Type I interferons: crucial participants in disease amplification in autoimmunity. *Nat. Rev. Rheumatol.* **6**, 40–49 (2010).
50. Tisoncik, J. R. et al. Into the eye of the cytokine storm. *Microbiol. Mol. Biol. Rev.* **76**, 16–32 (2012).

Acknowledgements We thank N. Eling, M. Fumagalli, Y. Gilad, O. Laufman, A. Marcovitz, J. Marioni, K. Meyer, M. Muffato, D. Odom, O. Stegle, A. Stern, M. Stubbington, V. Svensson and M. Ward for discussions; G. Emerton, A. Jinat, L. Mamanova, K. Polanski, A. Fullgrabe, N. George, S. Barnett, R. Boyd, S. Patel and C. Gomez for technical assistance; the Hipsci consortium for human fibroblast lines; and members of the Teichmann laboratory for support at various stages. This project was supported by ERC grants (ThDEFINE, ThSWITCH) and an EU FET-OPEN grant (MRG-GRAMMAR No 664918) and Wellcome Sanger core funding (Grant No WT206194). T.H. was supported by an HFSP Long-Term Fellowship and by EMBO Long-Term and Advanced fellowships. V.P. is funded by Fondazione Umberto Veronesi.

Reviewer information *Nature* thanks L. Barreiro, I. Yanai and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.H. and S.A.T. designed the project; T.H., X.C., R.J.M., R.R., N.K. and J.-E.P. performed experiments with help from V.P., G.D. and F.A.V.B.; T.H., X.C., R.J.M., R.R., T.G. and J.H. analysed the data with help from G.N., L.B.-C., G.T, A.N. and M.L.; J.F., E.S., P.V., I.K., M.D. and M.H. provided samples; S.A.T. supervised the project; T.H., R.R., N.K. and S.A.T. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0657-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0657-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.H. or S.A.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Ethical compliance. This project was approved by the Wellcome Sanger Institute Animal Welfare and Ethical Review Body, and complied with all relevant ethical regulations regarding animal research and human studies. Human cells were obtained from the Hipsci project[51], where they were collected from volunteers recruited from the NIHR Cambridge BioResource (written consent was given). Human skin profiling was performed in accordance with protocols approved by the Newcastle Research Ethics Committee (REC approval 08/H0906/95+5). Macaque skin samples were obtained from animals assigned to unrelated non-infectious studies, provided by Public Health England's National Infection Service in accordance with Home Office (UK) guidelines and approved by the Public Health England Ethical Review Committee under an appropriate UK Home Office project license.

Cross-species dermal fibroblast stimulation with dsRNA and IFNB. *Tissue culture.* We cultured primary dermal fibroblasts from low passage cells (below 10) that originated from females from four different species (human (European ancestry), rhesus macaque, C57BL/6 (black 6) mouse and brown Norway rat). All skin samples were taken from shoulders. Stimulation experiments and library preparations were done in identical conditions across all species and for all genomics techniques. Details on the numbers of individuals used in each technique are listed in each technique's section and in Supplementary Table 1.

Human cells were obtained from the Hipsci project[51] (http://www.hipsci.org/). Rhesus macaque cells were extracted from skin tissues that were incubated for 2 h with 0.5% collagenase B (Roche; 11088815001) after mechanical processing, and then filtered through 100-μm strainers before being plated and passaged before cryo-banking. Rodent cells were obtained from PeloBiotech where they were extracted using a similar protocol. In vitro cultured fibroblasts from all four species resemble a particular in vivo cluster of dermal fibroblasts (see Supplementary Information). Cells were not tested for mycoplasma contamination.

Prior to stimulation, cells were thawed and grown for several days in ATCC fibroblast growth medium (Fibroblast Basal Medium (ATCC, ATCC-PCS-201-030) with Fibroblast Growth Kit-Low serum (ATCC, PCS-201-041) (supplemented with Primocin (Invivogen, ant-pm-1) and penicillin/streptomycin (Life Technologies, 15140122)) - a controlled medium that has proven to provide good growing conditions for fibroblasts from all species, with slightly less than 24 h doubling times. About 18 h before stimulation, cells were trypsinized, counted and seeded into 6-well plates (100,000 cells per well). Cells were stimulated as follows: (1) stimulated with 1 μg/ml high-molecular mass poly(I:C) (Invivogen,tlrl-pic) transfected with 2 μg/ml Lipofectamin 2,000 (ThermoFisher, 11668027); (2) mock transfected with Lipofectamin 2,000; (3) stimulated with 1,000 IU of IFNB for 8 h (human IFNB: 11410-2 (for human and macaque cells); rat IFNB: 13400-1; mouse IFNB: 12401-1; all IFNs were obtained from PBL, and had activity units based on similar virological assays); or (4) left untreated. Interferon stimulation was used as a control, to study how genes that were upregulated in the secondary wave of the innate immune response diverge between species.

Additional human and mouse samples were stimulated with 1,000 IU of cross-mammalian IFN (CMI, or Universal Type I IFN Alpha, PBL, 11200-1). The latter stimulation was done to assess the effects of species-specific and batch-specific IFNB.

In all of the above-mentioned stimulations, we used a longer time course for single-cell RNA-seq than for bulk RNA-seq, for two main reasons: (1) in the bulk, we chose to focus on one main stimulation time point for simplicity and to obtain an intuitive fold-change between stimulated and unstimulated conditions; (2) in single cells, when studying cell-to-cell variability, we chose to profile, in addition to the main stimulation time point, cells in earlier and later stages of the response. This is important for studying how the dynamics and magnitude of the response affect gene expression variability between responding cells.

The poly(I:C) we used tested negative for the presence of bacterial beta-endotoxin using a coagulation test (PYROGENT Plus, 0.06 EU/ml sensitivity, N283-06).

Bulk RNA-seq: library preparation and sequencing. For bulk transcriptomics analysis, cells from individuals from different species were grown in parallel and stimulated with dsRNA, IFNB (and cross-mammalian IFN) and their respective controls. In total, samples from 6 humans, 6 macaques, 3 mice and 3 rats were used. Total RNA was extracted using the RNeasy Plus Mini kit (Qiagen, 74136), using QIAcube (Qiagen). RNA was then measured using a Bioanalyzer 2100 (Agilent Technologies), and samples with RIN < 9 were excluded from further analysis (one macaque sample stimulated with poly(I:C) and its control).

Libraries were produced using the Kapa Stranded mRNA-seq Kit (Kapa Biosystems, KK8421). The Kapa library construction protocol was modified for automated library preparation by Bravo (Agilent Technologies). cDNA was amplified in 13 PCR cycles, and purified using Ampure XP beads (Beckman Coulter, A63882) (1.8× volume) using Zephyr (Perkin Elmer). Pooled samples were sequenced on an Illumina HiSeq 2500 instrument, using paired-end 125-bp reads.

ChIP–seq: library preparation and sequencing. Samples from three individuals from each of the four species were grown and stimulated (with poly(I:C) for 4 h or left untreated, as described above) in parallel to samples collected for bulk RNA-seq. Following stimulation, samples were crosslinked in 1% HCHO (prepared in 1× DPBS) at room temperature for 10 min, and HCHO was quenched by the addition of glycine at a final concentration of 0.125 M. Cells were pelleted at 4 °C at 2,000*g*, washed with ice-cold 1× DPBS twice, and snap-frozen in liquid nitrogen. Cell pellets were stored at −80 °C until further stages were performed. ChIPmentation was performed according to version 1.0 of the published protocol[52] with a few modifications (see additional details in Supplementary Methods).

Library preparation reactions contained the following reagents: 10 μl purified DNA (from the above procedure), 2.5 μl PCR Primer Cocktails (Nextera kit, Illumina, FC-121-1030), 2.5 μl N5xx (Nextera index kit, Illumina FC-121-1012), 2.5 μl N7xx (Nextera index kit, Illumina, FC-121-1012), 7.5 μl NPM PCR Master Mix (Nextera kit, Illumina, FC-121-1030). PCR cycles were as follows: 72 °C, 5 min; 98 °C, 2 min; [98 °C, 10 s, 63 °C, 30 s, 72 °C, 20 s] × 12; 10 °C hold.

Amplified libraries were purified by double AmpureXP bead purification: first with 0.5× bead ratio, keep supernatant, second with 1.4× bead ratio, keep bound DNA. Elution was done in 20 μl Buffer EB (QIAGEN).

One microlitre of library was run on a Bioanalyzer (Agilent Technologies) to verify normal size distribution. Pooled samples were sequenced on an Illumina HiSeq 2000 instrument, using paired-end 75-bp reads.

Flow cytometry for single-cell RNA-seq. For scRNA-seq, we performed two biological replicates, with each replicate having one individual from each of the four studied species. A time course of dsRNA stimulation of 0, 4, and 8 h was used in one replicate (divided into two technical replicates), while the second replicate included a time course of 0, 2, 4, and 8 h. Poly(I:C) transfection was done as described above. In the case of sorting with IFNLUX, we used rhodamine-labelled poly(I:C).

Cells were sorted with either Beckman Coulter MoFlo XDP (first replicate) or Becton Dickinson INFLUX (second replicate) into wells containing 2 μl lysis buffer (1:20 solution of RNase Inhibitor (Clontech, 2313A) in 0.2% v/v Triton X-100 (Sigma-Aldrich, T9284)), spun down and immediately frozen at −80 °C.

When sorting with MoFlo, a pressure of 15 psi was used with a 150-μm nozzle, using the 'Single' sort purity mode. Dead or late-apoptosis cells were excluded using propidium iodide at 1 μg/ml (Sigma, Cat Number P4170) and single cells were selected using FSC W versus FSC H. When sorting with INFLUX, a pressure of 3 psi was used with a 200-μm nozzle, with the 'single' sort mode. Dead or late-apoptosis cells were excluded using 100 ng/ml DAPI (4′,6-diamidino-2-phenylindole) (Sigma, D9542). DAPI was detected using the 355-nm laser (50 mW), using a 460/50 nm bandpass filter. Rhodamine was detected using the 561-nm laser (50mW), using a 585/29 nm bandpass filter. Single cells were collected using FSC W versus FSC H.

Library preparation from full-length RNA from single cells and sequencing. Sorted plates were processed according to the Smart-seq2 protocol[53]. Oligo-dT primer (IDT), dNTPs (ThermoFisher, 10319879) and ERCC RNA Spike-In Mix (1:25,000,000 final dilution, Ambion, 4456740) were added to each well, and reverse transcription (using 50 U SmartScribe, Clontech, 639538) and PCR were performed following the original protocol with 25 PCR cycles. cDNA libraries were prepared using Nextera XT DNA Sample Preparation Kit (Illumina, FC-131-1096), according to the protocol supplied by Fluidigm (PN 100-5950 B1). Quality Checks on cDNA were done using a Bioanalyser 2100 (Agilent Technologies). Libraries were quantified using the LightCycler 480 (Roche), pooled and purified using AMPure XP beads (Beckman Coulter) with Hamilton 384 head robot (Hamilton Robotics). Pooled samples were sequenced on an Illumina HiSeq 2500 instrument, using paired-end 125-bp reads.

Read mapping to annotated transcriptome. For bulk RNA-seq samples, adaptor sequences and low-quality score bases were first trimmed using Trim Galore (version 0.4.1) (with the parameters '–paired–quality 20–length 20 -e 0.1–adapter AGATCGGAAGAGC'). Trimmed reads were mapped and gene expression was quantified using Salmon (version 0.6.0)[54] with the following command: 'salmon quant -i [index_file_directory] -l ISR -p 8–biasCorrect–sensitive–extraSensitive -o [output_directory] -1 -g [ENSEMBL_transcript_to_gene_file]–useFSPD–numBootstraps 100'. Each sample was mapped to its respective species' annotated transcriptome (downloaded from ENSEMBL, version 84: GRCh38 for human, MMUL_1 for macaque, GRCm38 for mouse, Rnor_6.0 for rat). We included only the set of coding genes (*.cdna.all.fa files). We removed annotated secondary haplotypes of human genes by removing genes with 'CHR_HSCHR'.

Quantifying differential gene expression in response to dsRNA. To quantify differential gene expression between treatment and control for each species and for each treatment separately, we used edgeR (version 3.12.1)[55] using the rounded estimated counts from Salmon. This was done only for genes that had a significant level of expression in at least one of the four species (TPM >3 in at least $N - 1$ libraries, where N is the number of different individuals we have for this species with libraries that passed quality control, and TPM is transcripts per million). Differential

© 2018 Springer Nature Limited. All rights reserved.

expression analysis was performed using the edgeR exact test, and *P* values were adjusted for multiple testing by estimating the false discovery rate (FDR).

Conservation and divergence in immune response: fold-change-based analysis. We compared the overall change in response to treatment (dsRNA or IFNB) between pairs of species, by computing the Spearman correlation of the fold-change in response to treatment across all one-to-one orthologues that were expressed in at least one species (Extended Data Fig. 1a–h). Fold-change was calculated with edgeR, as described above. Spearman correlations of all expressed genes appear in grey. Correlations of the subset of differentially expressed genes (genes with FDR-corrected $P < 0.01$ in at least one of the compared species) appear in black.

In Extended Data Fig. 1a–c, we show comparisons in response to dsRNA. In Extended Data Fig. 1d–f, we show comparisons in response to IFNB, which we use here to study the similarity of the secondary immune response between species.

We constructed a tree based on a gene's change in expression in response to dsRNA and to IFNB, using expressed genes that had one-to-one orthologues across all four species and were expressed in at least one species in at least one condition (Extended Data Fig. 1i). We used hierarchical clustering, with the hclust command from the stats R package, with the distance between samples computed as $1 - \rho$, where ρ is the pairwise Spearman correlation between each pair of species mentioned above (a greater similarity, reflected in a higher correlation, results in a smaller distance) and 'average' as the clustering method.

The above-mentioned analyses focus on one-to-one orthologues between the compared species. In Supplementary Table 6, we quantify the similarity in response between species (based on Spearman correlations) when adding genes with one-to-many orthologues.

Quantifying gene expression divergence in response to immune challenge. To quantify transcriptional divergence in immune response between species, we focus on genes that have annotated one-to-one orthologues across the studied species (human, macaque, mouse and rat). 9,753 of the expressed genes have annotated one-to-one orthologues in all four species, out of which 955 genes are differentially expressed in human in response to dsRNA treatment (genes with an FDR-corrected $P < 0.01$).

We define a measure of response divergence (based on a previous study[56]) by calculating the differences between the fold-change estimates across the orthologues: response divergence $= \log[1/4 \times \sum_{ij}(\log[\text{FC primate}_i] - \log[\text{FC rodent}_j])^2]$. This measure takes into account the structure of the phylogeny, and gives a relative measure of divergence in response across all genes with one-to-one orthologues.

To consider differences between species, we focus on between-clade differences (primates versus rodents), rather than on within-clade differences. In this way, we map the most significant macro-evolutionary differences along the longest branches of our four-species phylogeny. In addition, averaging within clades acts as a reduction of noise[56].

We compared this divergence measure to two other measures that use models (and incorporate both between- and within-clade divergence) and found a strong correlation between the divergence estimates across the three approaches (Supplementary Figs. 3, 4).

In most of the subsequent analyses, we focus on the 955 dsRNA-responsive genes: genes that were differentially expressed in response to dsRNA (genes that have an FDR-corrected $P < 0.01$ in human, and have annotated one-to-one orthologues in the other three species). For some of the analyses, we split these 955 genes based on quartiles, into genes with high, medium and low divergence (Fig. 1c).

We also studied how imprecisions in the fold-change estimates affected the response divergence estimates and subsequent analyses (Supplementary Figs. 5, 6).

Comparison of response divergence between different functional groups. To compare the divergence rates between sets of dsRNA-responsive genes that have different functions in the innate immune response, we split these 955 genes into the following functional groups (all groups are mutually exclusive, and any gene that belongs to two groups was excluded from the latter group; human gene annotations were used).

We first grouped genes by annotated molecular functions: viral sensors (genes that belong to one of the GO categories: GO:0003725 (dsRNA binding), GO:0009597 (detection of virus), and GO:0038187 (pattern recognition receptor activity)); cytokines, chemokines and their receptors (GO:0005125 (cytokine activity), GO:0008009 (chemokine activity), GO:0004896 (cytokine receptor activity), and GO:0004950 (chemokine receptor activity)); transcription factors (taken from the Animal Transcription Factor DataBase (version 2.0)[57]); chromatin modulators (GO:0016568 (chromatin modification), GO:0006338 (chromatin remodelling), GO:0003682 (chromatin binding), and GO:0042393 (histone binding)); kinases and phosphatases (GO:0004672 (protein kinase activity) and GO:0004721 (phosphoprotein phosphatase activity)); ligases and deubiquitinases (GO:0016579 (protein deubiquitination), GO:0004842 (ubiquitin-protein transferase activity) and GO:0016874 (ligase activity)); and other enzymes (mostly involved in metabolism rather than regulation: GO:0003824 (catalytic activity)). The divergence response values of these functional subsets were compared to the entire group of 955 dsRNA-responsive genes (Fig. 2d, e).

Next, we grouped genes by biological processes that are known to be important in the innate immune response: antiviral defence (GO:0051607 (defence response to virus)); inflammation (GO:0006954 (inflammatory response)); apoptosis (GO:0006915 (apoptotic process)); and regulation (GO annotations related to regulation of innate immune response pathways include only few genes. We thus used as the group of genes related to regulation, the merged group of genes that are annotated as transcription factors, chromatin modulators, kinases and phosphates or ligases and deubiquitinases, since all these groups include many genes that are known to regulate the innate immune response.)

Gene lists belonging to the mentioned GO annotations were downloaded using QuickGo[58]. The distribution of response divergence values for each of the functional groups was compared with the distribution of response divergence of the entire set of dsRNA-responsive genes. Cytokines, chemokines and their receptors are merged in Fig. 2d, e, 3c. Analogous comparisons of functional groups in IFNB response (with 841 IFNB-responsive genes) are shown in Supplementary Fig. 1. See additional analyses in Supplementary Information.

Alignment and peak calling of ChIP–seq reads. ChIP–seq reads were trimmed using trim_galore (version 0.4.1) with '–paired–trim1–nextera' flags. The trimmed reads were aligned to the corresponding reference genome (hg38 for human, rheMac2 for macaque, mm10 for mouse, rn6 for rat; all these genomes correspond to the transcriptomes used for RNA-seq mapping) from the UCSC Genome Browser[59] using bowtie2 (version 2.2.3) with default settings[60]. In all four species, we removed the Y chromosome. In the case of human, we also removed all alternative haplotype chromosomes. Following alignment, low-confident mapped and improperly paired reads were removed by samtools[61] with '-q 30 –f 2' flags.

Enriched regions (peaks) were called using MACS2 (v.2.1.1)[62] with a corrected *P* value cutoff of 0.01 with '-f BAMPE -q 0.01 -B–SPMR' flags, using input DNA as control. The genome sizes (the argument for '-g' flag) used were 'hs' for human, 'mm' for mouse, 3.0×10^9 for macaque and 2.5×10^9 for rat. Peaks were considered reproducible when they were identified in at least two of the three biological replicates and overlapped by at least 50% of their length (non-reproducible peaks were excluded from subsequent analyses). Reproducible peaks were then merged to create consensus peaks from overlapping regions of peaks from the three replicates by using mergeBed from the bedtools suite[63].

Gene assignment and conservation of active promoters and enhancers. We subsequently linked human peaks with the genes they might be regulating as follows: H3K4me3 consensus peak was considered the promoter region of a given gene if its centre was between 2 kb upstream and 500 bp downstream of the annotated TSS of the most abundantly expressed transcript of that gene.

Similarly, H3K27ac was considered the enhancer region of a given gene if its centre was in a distance above 1 kb and below 1 Mb, and there was no overlap (of 1 bp or more) with any H3K4me3 peak.

In each case where, based on the distance criteria, more than a single peak was linked to a gene (or more than a single gene was linked to a peak), we took only the closest peak–gene pair (ensuring that each peak will have up to one gene and vice versa).

To compare active promoters and enhancers between species, we excluded any human peak that could not be uniquely mapped to the respective region in the other species. This was done by looking for syntenic regions of human peaks in the other three species by using liftOver[64], and removing peaks that had either unmapped regions or more than one mapped region in the compared species. We considered syntenic regions with at least 70% sequence similarity between the species (minMatch = 0.7, and 0.8 in the case of human-macaque comparison), with a minimal length (minSizeQ and minSizeT) corresponding to the length of the shortest peak (128 bp in H3K4 and 142 bp in H3K27).

We defined an active human promoter or enhancer as conserved if a peak was identified in the corresponding region of the other species (we repeated this analysis by comparing human with each of the other three species separately). We compared the occurrence of conserved promoters and enhancers in genes that are highly divergent in response to dsRNA with low-divergence genes, and used Fisher's exact test to determine the statistical significance of the observed differences between high- and low-divergence genes (Extended Data Fig. 2).

Promoter sequence analysis. To calculate the total number of transcription factor binding motifs in a gene's active promoter region, we downloaded the non-redundant JASPAR core motif matrix (pfm_vertebrates.txt) from the JASPAR 2016 server[65] and searched for significant matches for these motifs using FIMO[66] in human H3K4me3 peaks. The TFBM density of peaks was calculated by dividing the total number of motif matches in a peak by the peak's length. TBFM density values in human H3K4me3 peaks linked with high- and low-divergence genes were compared (Fig. 2a).

PhyloP7 values were used to assess promoter sequence conservation[67]. Sequence conservation quantification was performed by taking the estimated nucleotide substitution rate for each nucleotide along the promoter sequence (500 bp upstream of the TSS of the relevant human gene). When several annotated transcripts existed,

© 2018 Springer Nature Limited. All rights reserved.

the TSS of the most abundantly expressed transcript was used (based on bulk RNA data). The substitution rate values from all genes were aligned, based on their TSS position, and a mean for each of the 500 positions was calculated separately for the group of genes with high, medium and low response divergence. The two-sample Kolmogorov–Smirnov test was used to compare the paired distribution of rates between the means of the high-divergence and low-divergence sets of genes. To plot the mean values of the three sets of divergent genes, the geom_smooth function from the ggplot2 R package was used with default parameters (with loess as the smoothing method) (Fig. 2b).

Human CGI annotations were downloaded from the UCSC genome table browser (hg38), and CGI genes were defined as those with a CGI overlapping their core promoter (300 bp upstream of the TSS reference position, and 100 bp downstream of it, as suggested previously[18]). Genes were defined as having a TATA box if they had a significant match to the Jaspar TATA box matrix (MA0108.1) in the 100 bp upstream of their TSS by FIMO[66] with default settings (we used a 100 bp window owing to possible inaccuracies in TSS annotations). We note that only 28 out of 955 dsRNA-responsive genes had a matching TATA-box motif in this region. For both TATA and CGI analyses, the promoter sequences of the human orthologues were used.

Read mapping and quality control of scRNA-seq (full-length RNA). Gene expression was quantified in a manner similar to the quantification for bulk transcriptomics libraries described above. Low-quality cells were filtered using quality control criteria (cells with at least 100,000 mapped reads, with at least 2,000 expressed genes with TPM >3, with ERCC $<10\%$ and MT $<40\%$, where ERCC and MT refer to reads mapped to synthetic RNA Spike-In genes and mitochondrial genes). This quality control filtering resulted in 240 cells from a first biological replicate, including two technical replicates (with a time course of 0, 4, 8 h). In a second larger biological replicate (with a dsRNA stimulation time course of 0, 2, 4, 8 h), 728 cells passed quality control. Results throughout the manuscript relate to the second cross-species biological replicate in which a higher proportion of cells passed QC, and the lower-quality first replicate data were not considered further.

Cell-to-cell variability analysis. To quantify the biological cell-to-cell variability of genes, we applied the DM (Distance to Median) approach—an established method, which calculates the cell-to-cell variability in gene expression while accounting for confounding factors such as gene expression level[30]. This is done by first filtering out genes that are expressed at low levels: for Smart-seq2 data we included only genes that had an average expression of at least 10 size-factor normalized reads (except for Extended Data Fig. 9a, in which we reduced the threshold to 5, to allow a larger number of genes to be included in the comparisons). This procedure was done to filter genes that displayed higher levels of technical variability between samples owing to low expression. Second, to account for gene expression level, the observed cell-to-cell variability of each gene was compared with its expected variability, based on its mean expression across all samples and in comparison with a group of genes with similar levels of mean expression. This DM value is also corrected by gene length (in the case of Smart-seq2 data), yielding a value of variability that can be compared across genes regardless of their length and mean expression values[68]. As a second approach, we used BASiCS[69,70] (see Supplementary Information).

We note that the relationship observed in Fig. 3a between response divergence and cell-to-cell variability is not an artefact, stemming from differences in expression levels: (A) With respect to cell-to-cell variability, a gene's expression level is controlled for by DM calculations, where expression level is regressed by using a running median (Supplementary Fig. 14). (B) Similarly, we can regress the expression level measured in bulk RNA-seq from the quantified response divergence by subtracting the running median of expression from the divergence estimates. When repeating the analysis comparing cell-to-cell variability versus regressed response divergence, the relationship between the two is maintained (Supplementary Fig. 15).

Cytokine co-expression analysis. For the chemokine gene *CXCL10*, we built a network (using CytoScape[71]) of genes that correlate with *CXCL10* in dsRNA-stimulated human fibroblasts and in at least one more species, using genes with a Spearman correlation value above 0.3 (see Fig. 3d and Supplementary Information).

Coding sequence evolution analysis. The ratio dN/dS (non-synonymous to synonymous codon substitutions) of human genes across the mammalian clade was obtained from a previous study that used orthologous genes from 29 mammals[72]. Distributions of dN/dS values were computed for each of the three groups of genes with low, medium and high divergence in response to dsRNA, and are plotted in Fig. 4a.

Rate of gene gain and loss analysis. The significance at which a gene's family has experienced a higher rate of gene gain and loss in the course of vertebrate evolution, in comparison with other gene families, was retrieved from ENSEMBL[73]. The statistics provided by ENSEMBL are calculated using the CAFE method[74], which estimates the global birth and death rate of gene families and identifies gene families that have accelerated rates of gain and loss. Distributions of the *P* values from this statistic were computed for each of the three groups of genes with low, medium and high divergence in response to dsRNA and are plotted as the negative logarithm values in Fig. 4b.

Gene age analysis. Gene age estimations were obtained from ProteinHistorian[75]. To ensure that the results were not biased by a particular method of ancestral protein family reconstruction or by specific gene family assignments, we used eleven different estimates for mammalian genes (combining five different databases of protein families with two different reconstruction algorithms for age estimation, as well as an estimate from the phylostratigraphic approach). For each gene, age was defined with respect to the species tree, where a gene's age corresponds to the branch in which its family is estimated to have appeared (thus, larger numbers indicate evolutionarily older genes).

Data for gene age in comparison with divergence in response to dsRNA are shown in Fig. 4c (using Panther7 phylogeny and Wagner reconstruction algorithm) and in Supplementary Fig. 17a (for all 11 combinations of gene family assignments and ancestral family reconstructions). See additional analyses in Supplementary Information.

Cellular protein–protein interaction analysis. Data on the number of experimentally validated PPIs for human genes were obtained from STRING (version 10)[76]. Distributions of PPIs for genes with low, medium and high divergence in response to dsRNA are plotted in Fig. 4d.

Host–virus interaction analysis. Data on host–virus protein–protein interactions were downloaded from the VirusMentha database[43], and combined with two additional studies that have annotated host–virus protein–protein interactions[6,44]. We split the 955 dsRNA-responsive genes into genes with known viral interactions (genes whose protein products were reported to interact with at least one viral protein), and genes with no known viral interactions: 'viral interactors' and 'no viral interactions', respectively, in Fig. 4e. In addition, we define a subset of genes within the viral interactors set: those known to interact with viral proteins that are immunomodulators (proteins known to target the host immune system and modulate its response[45]).

We note that the results presented in Fig. 4e are in agreement with previous analyses that are based on all human genes and on coding sequence evolution[46]. However, the overlap in the sets of genes between the previous analyses and the one presented here is small (for example, in one published study[46] there were 535 human genes with known interactions with pathogens, 57 of which overlap with the 955 genes that are the basis of the current analysis).

Additional experiments with human fibroblasts and human skin tissue. Additional experiments were performed with human dermal fibroblasts and with cells extracted from human skin tissues to study in greater detail the relationship between response divergence across species and cell-to-cell variability. See Supplementary Methods and Supplementary Discussion for details.

Cross-species bone marrow-derived phagocyte stimulation with LPS and dsRNA. *Tissue culture.* Primary bone marrow-derived mononuclear phagocytes originating from females of four different species (black 6 mouse, brown Norway rat, rabbit and pig) and cultured with GM-CSF, were obtained from PeloBiotech. Twenty-four hours before the start of the stimulation time course, cells were thawed and split into 12-well plates (500,000 cells per well). Cells were stimulated with: (1) 100 ng/ml LPS (Invivogen, tlrl-smlps), or with (2) 1 μg/ml high-molecular mass poly(I:C) (Invivogen, tlrl-pic) transfected with 2 μl/ml Lipofectamin 2,000 (ThermoFisher, 11668027). LPS stimulation time courses of 0, 2, 4, 6 h were performed for all species. Poly(I:C) stimulations were performed for rodents for 0, 2, 4, 6 h. We also processed cells for bulk RNA-seq for 0 and 4 h stimulation time points. Details on the individuals used in each technique are listed in Supplementary Table 2.

Library preparation for single cells using microfluidic droplet cell capture. Following stimulation, cells were collected using Cell Dissociation Solution Non-enzymatic (Sigma-Aldrich, C5914), washed and resuspended in $1\times$ PBS with 0.5% (w/v) BSA. Cells were then counted and loaded on the 10x Chromium machine aiming for a targeted cell recovery of 5,000 cells according to the manual. Libraries were prepared following the Chromium Single Cell 3′ v2 Reagent Kit Manual[77]. Libraries were sequenced on an Illumina HiSeq 4000 instrument with 26 bp for read 1 and 98 bp for read 2.

Library preparation and sequencing for bulk RNA-seq. Total RNA was extracted and libraries were prepared as described in the fibroblasts section. Pooled samples were sequenced on an Illumina HiSeq 4000 instrument, using paired-end 75-bp reads.

Quantifying gene expression in bulk RNA-seq data. Adaptor sequences and low-quality score bases were trimmed using Trim Galore (version 0.4.1). Trimmed reads were mapped and gene expression was quantified using Salmon: (version 0.9.1)[54] with the following command: 'salmon quant -i [index_file_directory] / -l ISR -p 8–seqBias–gcBias–posBias -q -o [output_directory] -1 -g [ENSEMBL_transcript_to_gene_file]–useVBOpt–numBootstraps 100'. Mouse samples were mapped to mouse transcriptome (ENSEMBL, version 84). We note that we used

© 2018 Springer Nature Limited. All rights reserved.

the bulk data only for TSS analysis. For differential expression analysis, we used an in silico bulk from the single-cell data (see below).

Quantifying gene expression in microfluidic droplet cell capture data. Microfluidic droplet cell capture data was first quantified using 10x Genomics' Cell Ranger Single-Cell Software Suite (version 2.0, 10× Genomics Inc.)[77] against the relevant genome (ENSEMBL, version 84). We removed cells with fewer than 200 genes or more than 10% mitochondrial reads. To remove potential doublets, we excluded the top 10% of cells expressing the highest numbers of genes. Genes expressed in less than 0.5% of the cells were excluded from the calculations. We then filtered cells that expressed fewer than 10% of the total number of filtered genes.

Since bone marrow-derived phagocytes may include secondary cell populations, we focused our analysis on the major cell population. We identified clusters within each data set, using the Seurat[78] functions RunPCA, followed by FindClusters (using 20 dimensions from the PCA, default perplexity and a resolution of 0.1) and have taken the cells belonging to the largest cluster for further analysis, resulting in a less heterogeneous population of cells. A lower resolution of 0.03 was used for rabbit-LPS4, rabbit-LPS2, mouse-PIC2, mouse-PIC4; and 0.01 for rabbit-LPS6.

Quantifying gene expression divergence in response to immune challenge. We created an in silico bulk table by summing up the UMIs of the post-QC single cells belonging to the largest cluster of cells, in each of the samples. We then used the three replicates in unstimulated conditions and in 4 h LPS stimulation to perform a differential expression analysis using DESeq2[79] Wald test, and *P* values were adjusted for multiple testing by estimating the FDR. A similar procedure was performed with mouse and rat dsRNA stimulation (with 4 h dsRNA stimulation versus unstimulated conditions).

To quantify transcriptional divergence in immune response between species, we focused on genes that have annotated one-to-one orthologues across the studied species.

We define a measure of response divergence by calculating the differences between the fold-change estimates across the orthologues: response divergence $= \log[1/3 \times \sum_j (\log[\text{FC pig}] - \log[\text{FC glire}_j])^2]$. For each gene, the fold-change in the outer group (pig), is subtracted from the fold-change in the orthologues of the three glires (mouse, rat and rabbit), and the average of the square values of these subtractions is taken as the response divergence measure. In most of the analyses, we focus on the 2,336 LPS-responsive genes—genes that are differentially expressed in response to LPS (genes that have an FDR-corrected $P < 0.01$ in mouse, and have annotated one-to-one orthologues in the other three species).

Promoter elements, gene function and cell-to-cell variability analyses. Promoter elements (TATA and CGIs), gene function and cell-to-cell variability analyses were performed as described in the fibroblasts section. Mouse genes were used as the reference for gene function and TSS annotations. For variability analysis, we used one representative replicate out of three.

Statistical analysis and reproducibility. Statistical analyses were done with R version 3.3.2 for Fisher's exact test, two-sample Kolmogorov–Smirnov test and Mann–Whitney test. Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the interquartile range (as implemented by the R function geom_boxplot). Violin plots show the kernel probability density of the data (as implemented by the R function geom_violin).

All cross-species bulk RNA-seq replicates were successful, except for one macaque individual in which the treated sample had a low RNA quality and was removed from the analysis (along with the matching control). All cross-species ChIP–seq replicates were successful. Cross-species scRNA-seq of fibroblasts was performed in two biological replicates. Results throughout the manuscript relate to the second cross-species biological replicate, for which a higher proportion of cells passed technical quality control. Three out of three replicates for each species and condition were successful when preparing single-cell libraries for mononuclear phagocytes, except for two libraries that failed at the emulsion preparation stage. Two out of two replicates of single-cell in situ RNA hybridization assay were performed and both are shown.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

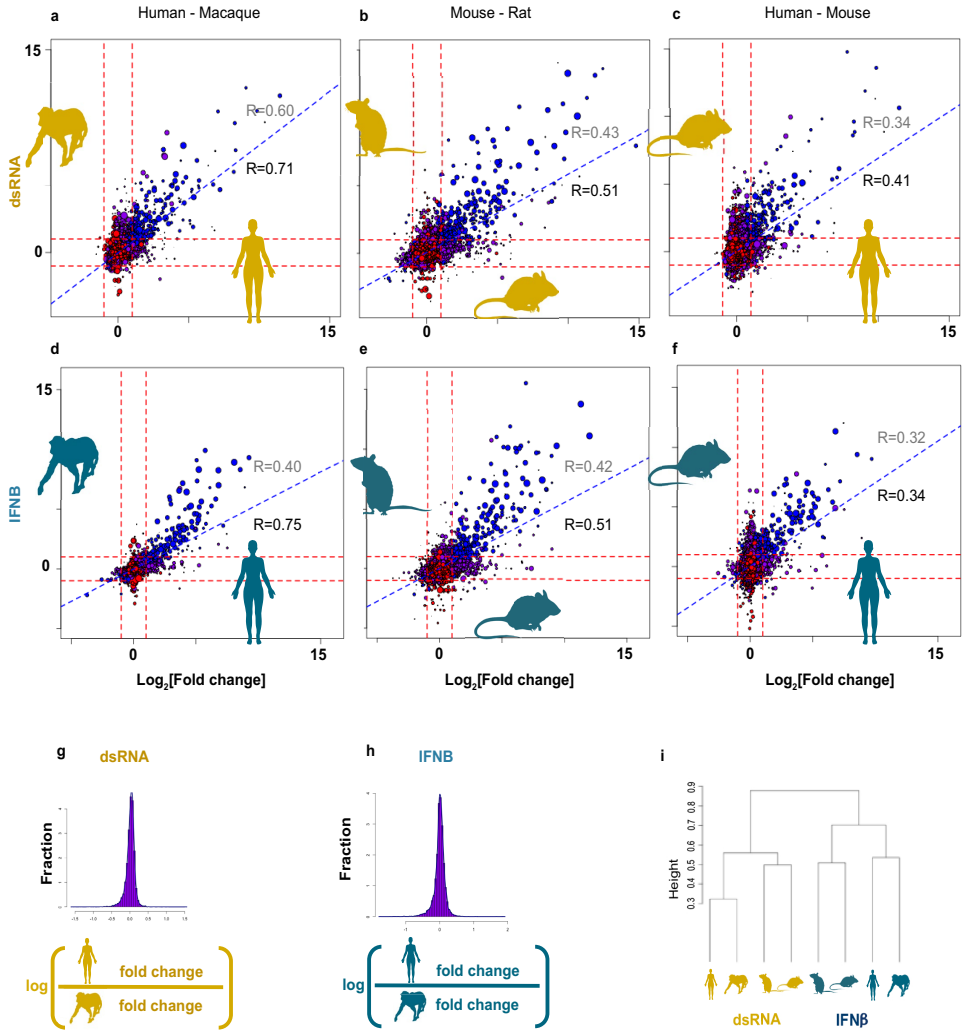
Code availability. Scripts for major analyses are available at https://github.com/Teichlab/innate_evo.

Data availability

Sequencing data have been deposited in ArrayExpress with the following accessions: E-MTAB-5918, E-MTAB-5919, E-MTAB-5920, E-MTAB-6754, E-MTAB-6773, E-MTAB-5988, E-MTAB-5989, E-MTAB-6831, E-MTAB-6066, E-MTAB-7032, E-MTAB-7037, E-MTAB-7051 and E-MTAB-7052.

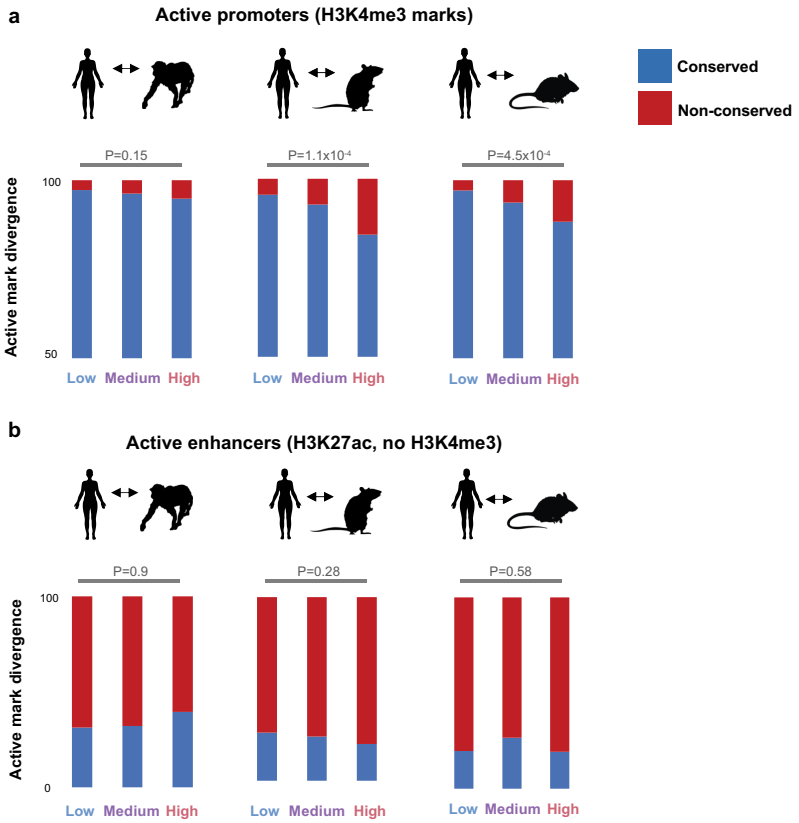
51. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
52. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods* **12**, 963–965 (2015).
53. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
54. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
55. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
56. Nourmohammad, A. et al. Adaptive evolution of gene expression in *Drosophila*. *Cell Reports* **20**, 1385–1395 (2017).
57. Zhang, H. M. et al. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.* **40**, D144–D149 (2012).
58. Binns, D. et al. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
59. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
61. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
64. Kuhn, R. M. et al. The UCSC genome browser database: update 2007. *Nucleic Acids Res.* **35**, D668–D673 (2007).
65. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
66. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
67. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
68. Kolodziejczyk, A. A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
69. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLOS Comput. Biol.* **11**, e1004333 (2015).
70. Martinez-Jimenez, C. P. et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
71. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
72. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
73. Herrero, J. et al. Ensembl comparative genomics resources. *Database* **2016**, bav096 (2016).
74. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
75. Capra, J. A., Williams, A. G. & Pollard, K. S. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLOS Comput. Biol.* **8**, e1002567 (2012).
76. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
77. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
78. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
79. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold-change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

© 2018 Springer Nature Limited. All rights reserved.



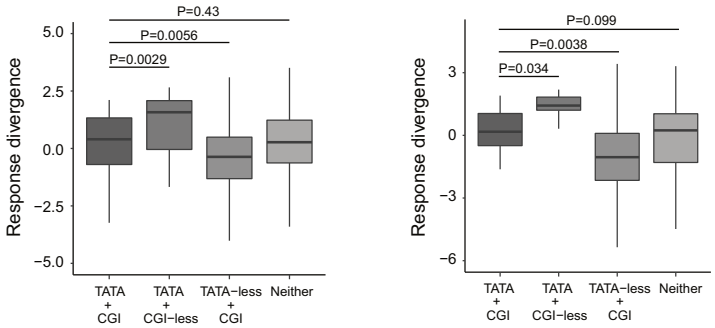
Extended Data Fig. 1 | Fibroblast response to dsRNA and IFNB across species. To study the similarity in response to treatment across species, we plotted the fold-change values of all expressed genes (with one-to-one orthologues) between pairs of species (human–macaque, mouse–rat and human–mouse) in response to dsRNA (poly(I:C)) (**a**–**c**). As a control, we performed the same procedure with IFNB stimulations (**d**–**f**). Fold-changes were inferred from differential expression analyses, determined by the exact test in the edgeR package[6] and based on $n = 6$, 5, 3 and 3 individuals from human, macaque, rat and mouse, respectively. Spearman correlations between all expressed one-to-one orthologues are shown in grey, Spearman correlations between the subset of differentially expressed genes (FDR-corrected $P < 0.01$ in at least one species) appear in black. Number of genes shown is $n = 11{,}035$, 11,005, 11,137, 10,851, 10,826 and 10,957 in **a**–**f**, respectively. Genes are coloured blue if they were differentially expressed (FDR-corrected $P < 0.01$) in both species, purple if they were differentially expressed in only one species, or red if they were not differentially expressed. **g**, **h**, Density plots of ratio of fold-change in response to dsRNA or to IFNB. **g**, Comparison between human and macaque orthologues in dsRNA response. **h**, Comparison between human and mouse orthologues in IFNB response. **i**, Dendrogram based on the fold-change in response to dsRNA or to IFNB across 9,835 one-to-one orthologues in human, macaque, rat and mouse.

© 2018 Springer Nature Limited. All rights reserved.



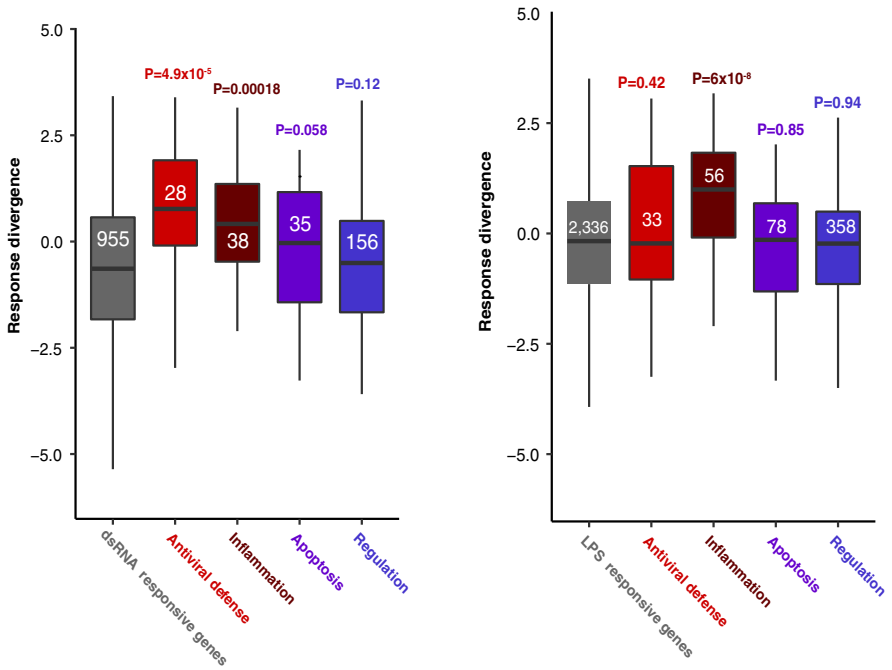
Extended Data Fig. 2 | Correspondence of transcriptional divergence and divergence of active promoters and enhancers. Comparison of divergence in transcriptional response to dsRNA with divergence of active chromatin marks in active promoters (**a**, profiled using H3K4me3 in proximity to gene's TSS) and enhancers (**b**, H3K27ac without overlapping H3K4me3). Chromatin marks were linked to genes on the basis of their proximity to the gene's TSS. Chromatin marks were obtained from $n = 3$ individuals in each of the four species, from fibroblasts stimulated with dsRNA or left untreated. The statistics are based on $n = 855$, 818 and 813 human genes that have a linked H3K4me3 mark with a syntenic region in macaque, rat and mouse, respectively (**a**); and on $n = 326$, 241 and 242 human genes that have a linked H3K27ac mark with a syntenic region in macaque, rat and mouse, respectively (**b**). Each panel shows the fraction of conserved marks between human and macaque, rat or mouse, in genes that have high, medium and low divergence in their transcriptional response. In each column, the histone mark's signal was compared between human and the syntenic region in one of the three other species. If an active mark was found in the corresponding syntenic region, the linked gene was considered to have a conserved active mark (promoter or enhancer). The fractions of genes with conserved promoters (or enhancers) in each pair of species were compared between high- and low-divergence genes using a one-sided Fisher's exact test. When comparing active promoter regions of high- versus low-divergence genes, we observe that low-divergence genes have a significantly higher fraction of conserved marks in rodents. This suggests an agreement between divergence at the transcriptional and chromatin levels in active promoter regions. In active enhancer regions, we do not observe these patterns, suggesting that the major contribution to divergence comes from promoters.

© 2018 Springer Nature Limited. All rights reserved.



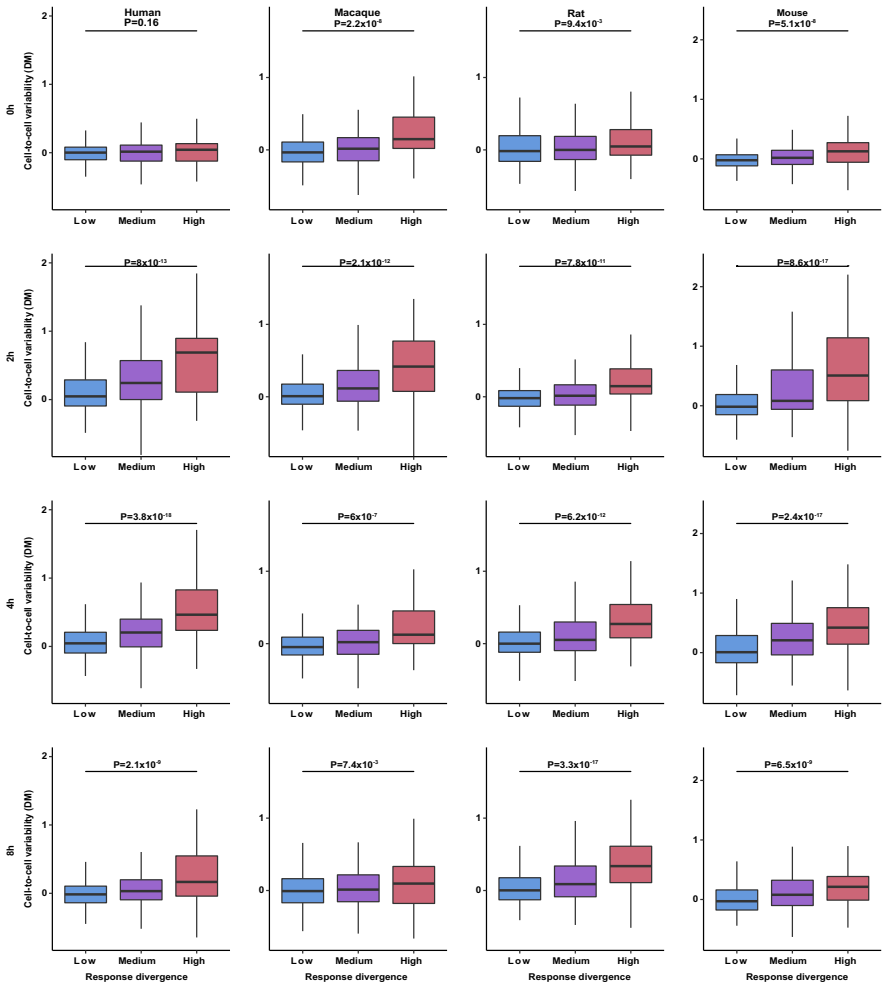
Extended Data Fig. 3 | Comparison of response divergence of genes containing various promoter elements. Comparison of response divergence between genes with and without a TATA-box and a CGI. Left, fibroblasts ($n = 14$, 14, 633 and 294 differentially expressed genes with only TATA-box element, with both CGI and TATA-box elements, with only CGI, and with neither element in their promoters, respectively); right, phagocytes ($n = 13$, 29, 1,718 and 576 differentially expressed genes with only a TATA-box element, with both CGI and TATA-box elements, with only a CGI, and with neither element in their promoters, respectively). Genes with a TATA-box without a CGI have higher response divergence than genes with both elements. Genes with a CGI but without a TATA-box diverge more slowly than genes with both elements. Genes with both elements do not differ significantly in their divergence from genes lacking both elements (one-sided Mann–Whitney test). Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the IQR.

© 2018 Springer Nature Limited. All rights reserved.



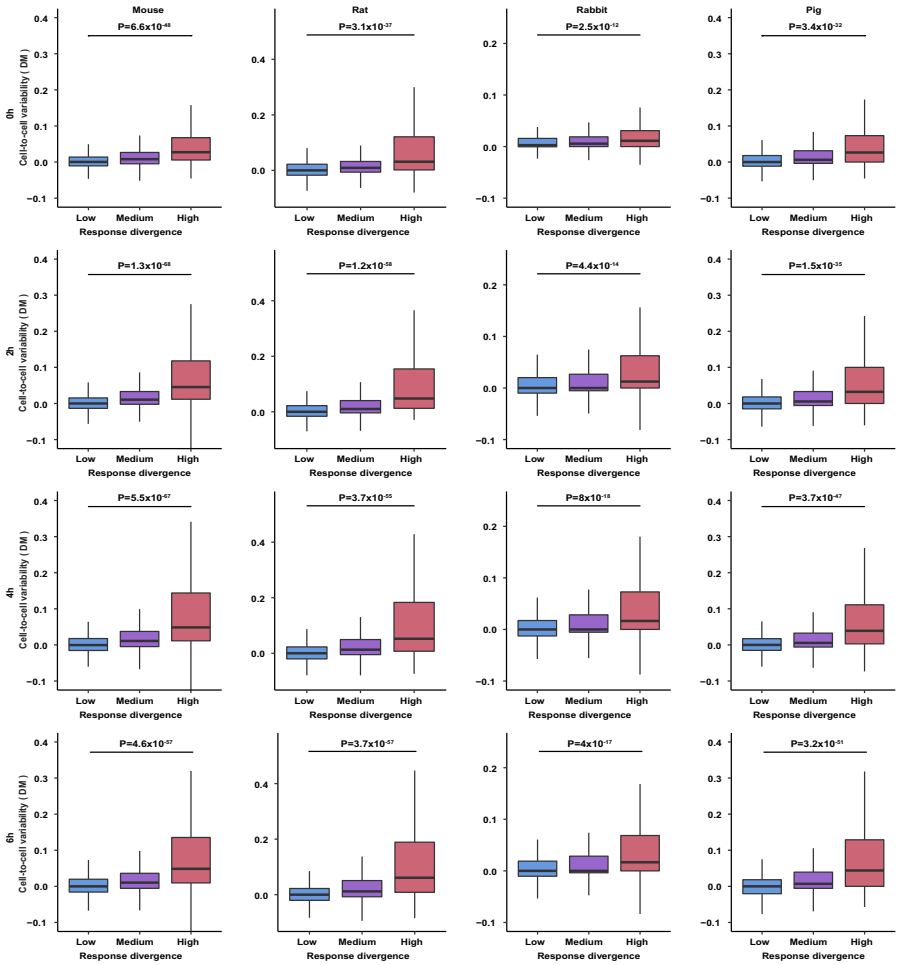
Extended Data Fig. 4 | Response divergence of molecular processes upregulated in immune response. Left, distributions of divergence values of $n = 955$ dsRNA-responsive genes in fibroblasts and subsets of this group belonging to different biological processes. For each functional subset, the distribution of divergence values is compared with the set of 955 dsRNA-responsive genes using a one-sided Mann–Whitney test. FDR-corrected *P* values are shown above each group and group size is shown inside each box. Right, distributions of divergence values of $n = 2{,}336$ LPS-responsive genes in mononuclear phagocytes and subsets of this group belonging to different biological processes. For each functional subset, the distribution of divergence values is compared with the set of 2,336 LPS-responsive genes. FDR-corrected *P* values (one-sided Mann–Whitney test) are shown above each group and group size is shown inside each box. Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the IQR.

© 2018 Springer Nature Limited. All rights reserved.



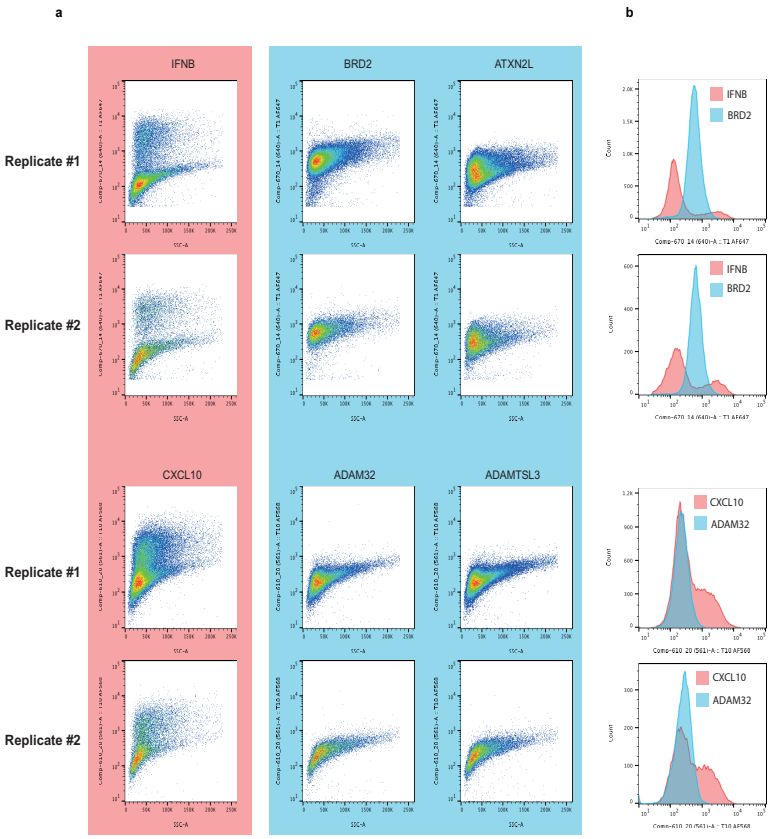
Extended Data Fig. 5 | Cell-to-cell variability versus response divergence across species and conditions in fibroblasts after dsRNA stimulation. Cell-to-cell variability values, as measured with DM across individual cells, compared with response divergence between species (grouped into low, medium and high divergence). Variability values are based on $n = 29, 56, 55, 35$ human cells, $n = 20, 32, 29, 13$ rhesus cells, $n = 33, 70, 65, 40$ rat cells, and $n = 53, 81, 59, 30$ mouse cells, stimulated with dsRNA for 0, 2, 4 and 8 h, respectively. Rows represent different dsRNA stimulation time points (0, 2, 4 and 8 h), and columns represent different species as shown. High-divergence genes were compared with low-divergence genes using a one-sided Mann–Whitney test. Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the IQR.

© 2018 Springer Nature Limited. All rights reserved.



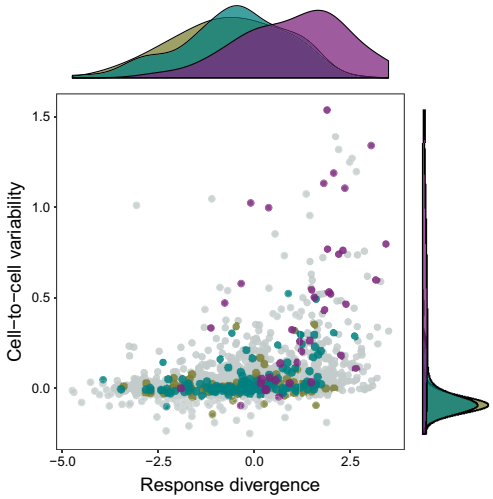
Extended Data Fig. 6 | Cell-to-cell variability versus response divergence across species and conditions in mononuclear phagocytes after LPS stimulation. Cell-to-cell variability values, as measured with DM across cells, compared with response divergence between species (grouped into low, medium and high divergence). Variability values are based on $n = 3{,}519$, 4,321, 3,293, 2,126 mouse cells, $n = 2{,}266$, 2,839, 1,963, 1,607 rat cells, $n = 3{,}275$, 1,820, 1,522, 1,660 rabbit cells, and $n = 1{,}748$, 1,614, 1,899, 1,381 pig cells, stimulated with LPS for 0, 2, 4 and 6 h, respectively. Rows represent different LPS stimulation time points (0, 2, 4 and 6 h), and columns represent different species as shown. High-divergence genes were compared with low-divergence genes using a one-sided Mann–Whitney test. Data in boxplots represent the median, first quartile and third quartile with lines extending to the furthest value within 1.5 of the IQR.

© 2018 Springer Nature Limited. All rights reserved.



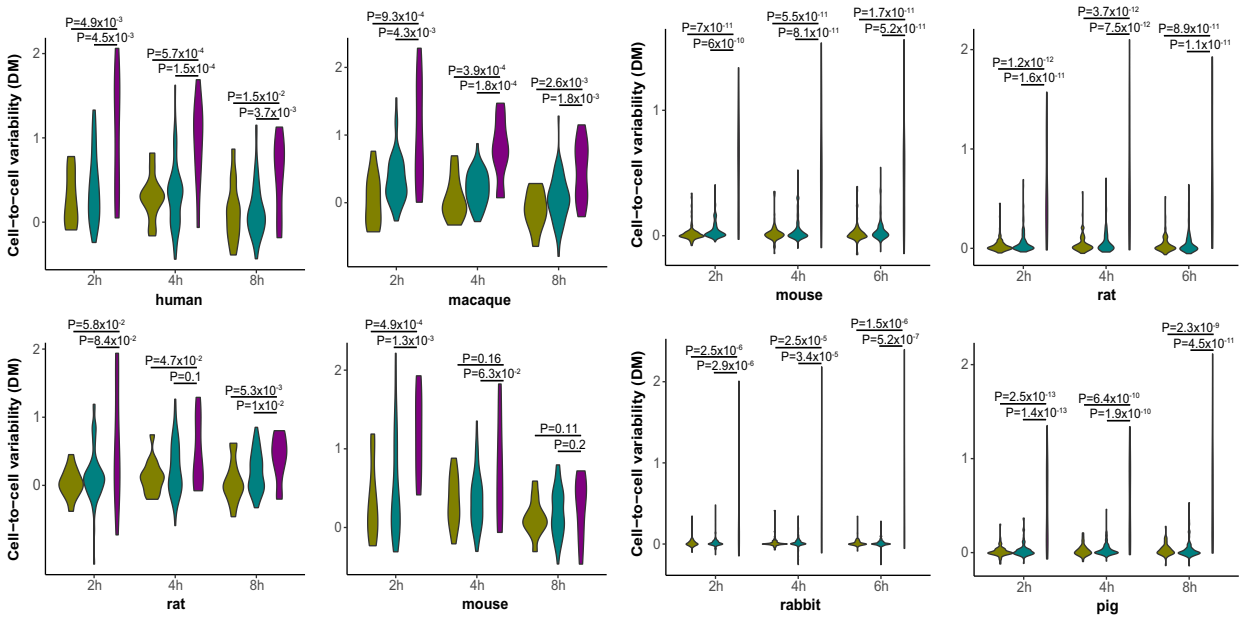
Extended Data Fig. 7 | Cell-to-cell variability of cytokine expression in single cell in situ RNA hybridization assay combined with flow cytometry (PrimeFlow). PrimeFlow measurement of two cytokine genes (*IFNB* and *CXCL10*) that show high cell-to-cell variability in scRNA-seq. As controls, two genes matched on expression levels (*ATXN2L* and *ADAM32*) but that show low cell-to-cell variability in scRNA-seq data are shown. As the expression of cytokines is at the low end of the distribution, we also chose two genes with middle-range expression values (*ADAMTSL3* and *BRD2*) as additional controls. The experiment was performed in $n = 2$ independent replicates, originating from the same individual. Both replicates are shown. **a**, Pseudocolour contour plot for RNA target expression in dsRNA-stimulated human fibroblasts. The *x*-axis shows area of side scatter (SSC-A) and the *y*-axis shows fluorescent signal for target RNA probes. RNA targets detected by the same fluorescent channel are displayed together. Top, *IFNB* and control genes *BRD2* and *ATXN2L*, type 1 probe, Alexa FluorTM 647. Bottom, *CXCL10* and control genes *ADAMTSL3* and *ADAM32*, type 10 probe, Alexa FluorTM 568. The cytokine genes display a broader range of fluorescence signal than the controls. **b**, Histograms comparing fluorescence of cytokine and control pairs (*IFNB*–*BRD2* for type 1 probe and *CXCL10*–*ADAM32* for type 10 probe). The histograms show a bimodal distribution of expression signal for the two cytokine genes (*IFNB* and *CXCL10*, red), but not for controls (blue). This agrees with scRNA-seq data in which *CXCL10* and *IFNB* display high levels of cell-to-cell variability.

© 2018 Springer Nature Limited. All rights reserved.



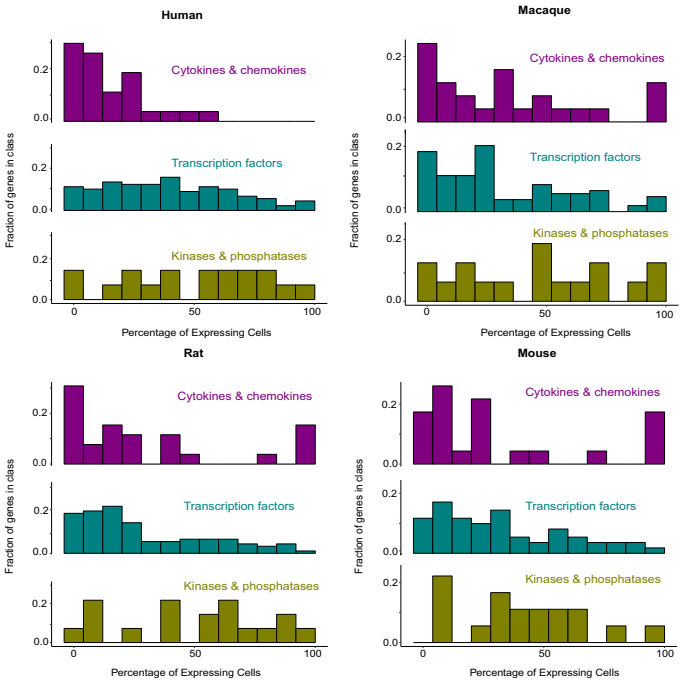
Extended Data Fig. 8 | Cell-to-cell variability levels and response divergence of cytokines, transcription factors and kinases in response to LPS stimulation of phagocytes. A scatter plot showing divergence in response to LPS across species and transcriptional cell-to-cell variability in mouse mononuclear phagocytes following 4 h of LPS treatment, in $n = 2{,}262$ LPS-responsive genes. Purple, cytokines; green, transcription factors; beige, kinases. The distributions of divergence values and cell-to-cell variability values of each of the three functional groups are shown above and to the right of the scatter plot, respectively.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Cell-to-cell variability levels in cytokines, transcription factors and kinases across species and stimulation time points. Violin plots showing the distribution of cell-to-cell variability values (DM) of cytokines, transcription factors and kinases during immune stimulation. Left, fibroblast dsRNA stimulation time course. Number of cells used in each species (at 2, 4, 8 h dsRNA, respectively): human, 56, 55, 35; macaque, 32, 29, 13; rat, 70, 65, 40; mouse, 81, 59, 30. Right, phagocyte LPS stimulation time course. Number of cells used in each species (at 2, 4, 6 h LPS, respectively): mouse, 4,321, 3,293, 2,126; rat, 2,839, 1,963, 1,607; rabbit, 1,820, 1,522, 1,660; pig, 1,614, 1,899, 1,381. For both panels, colours as in Fig. 3c. Comparisons between groups of genes were performed using one-sided Mann–Whitney tests. Violin plots show the kernel probability density of the data.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Percentage of cells expressing cytokines, transcription factors and kinases. Histograms showing the percentage of fibroblasts expressing cytokines (top), transcription factors (middle) and kinases (bottom) following 4 h dsRNA stimulation, in human, macaque, rat and mouse cells (based on $n = 55$, 29, 65 and 59 cells, respectively). The percentage of expressing cells is divided into 13 bins (x-axis). The y-axis represents the fraction of genes from this gene class (for example, cytokines) that are expressed in each bin (for example, in human, nearly 30% of the cytokine genes (y-axis) are expressed in the first bin, corresponding to expression in fewer than 8% of cells).

© 2018 Springer Nature Limited. All rights reserved.

Appendix E

Innate immune response modules

Table E.1 GO term enrichment in IFN-β response gene modules

GO term ID	GO term name	Enrichment p-value
	Canonical Type I IFN	
GO:0051607	defense response to virus	2.36e-44
GO:0060337	type I interferon signaling pathway	8.34e-38
GO:0032479	regulation of type I interferon production	3.18e-13
GO:0070647	protein modification by small protein conjugation or removal	2.4e-05
GO:0044248	cellular catabolic process	0.000405
GO:0006508	proteolysis	0.00116
GO:0071360	cellular response to exogenous dsRNA	0.0016
GO:0032020	ISG15-protein conjugation	0.00591
GO:0006471	protein ADP-ribosylation	0.00613
GO:0003373	dynamin family protein polymerization involved in membrane fission	0.00769
GO:0090503	RNA phosphodiester bond hydrolysis, exonucleolytic	0.0103
GO:2001034	positive regulation of double-strand break repair via nonhomologous end joining	0.0114
GO:1904469	positive regulation of tumor necrosis factor secretion	0.0114
GO:0052548	regulation of endopeptidase activity	0.0114

GO term ID	GO term name	Enrichment p-value
GO:0000266	mitochondrial fission	0.0115
GO:0034058	endosomal vesicle fusion	0.0136
GO:0019985	translesion synthesis	0.0138
GO:0070206	protein trimerization	0.0185
GO:0051770	positive regulation of nitric-oxide synthase biosynthetic process	0.0213
GO:0035563	positive regulation of chromatin binding	0.0213
GO:0034356	NAD biosynthesis via nicotinamide riboside salvage pathway	0.0268
GO:0009200	deoxyribonucleoside triphosphate metabolic process	0.0299
GO:0051248	negative regulation of protein metabolic process	0.0299
GO:0034162	toll-like receptor 9 signaling pathway	0.033
GO:0061025	membrane fusion	0.0403
GO:0002737	negative regulation of plasmacytoid dendritic cell cytokine production	0.0411
GO:0072308	negative regulation of metanephric nephron tubule epithelial cell differentiation	0.0411
GO:1905795	cellular response to puromycin	0.0411
GO:0090616	mitochondrial mRNA 3'-end processing	0.0411
GO:0031324	negative regulation of cellular metabolic process	0.0448
GO:0051100	negative regulation of binding	0.0498
GO:0048661	positive regulation of smooth muscle cell proliferation	0.0498
	Regulator/signal transduction	
GO:0006952	defense response	8.86e-17
GO:0060333	interferon-gamma-mediated signaling pathway	1.09e-08
GO:0008219	cell death	5.81e-06
GO:0010952	positive regulation of peptidase activity	1.79e-05
GO:0001817	regulation of cytokine production	1.84e-05
GO:0032940	secretion by cell	0.000203
GO:0002274	myeloid leukocyte activation	0.00033
GO:0045055	regulated exocytosis	0.00102
GO:0043687	post-translational protein modification	0.00147

GO term ID	GO term name	Enrichment p-value
GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling	0.0016
GO:0016485	protein processing	0.00245
GO:0061180	mammary gland epithelium development	0.00426
GO:1902728	positive regulation of growth factor dependent skeletal muscle satellite cell proliferation	0.00482
GO:0048872	homeostasis of number of cells	0.00933
GO:0006775	fat-soluble vitamin metabolic process	0.0117
GO:0010957	negative regulation of vitamin D biosynthetic process	0.0186
GO:1903903	regulation of establishment of T cell polarity	0.0186
GO:0071360	cellular response to exogenous dsRNA	0.0199
GO:1901224	positive regulation of NIK/NF-kappaB signaling	0.0219
GO:0042325	regulation of phosphorylation	0.0223
GO:0046902	regulation of mitochondrial membrane permeability	0.0286
GO:1903599	positive regulation of autophagy of mitochondrion	0.0299
GO:1902895	positive regulation of pri-miRNA transcription by RNA polymerase II	0.0338
GO:0032020	ISG15-protein conjugation	0.0368
GO:0048050	post-embryonic eye morphogenesis	0.0368
GO:0070647	protein modification by small protein conjugation or removal	0.0374
GO:0001503	ossification	0.0405
GO:0001885	endothelial cell development	0.0463
GO:0002291	T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell	0.0468
GO:0003382	epithelial cell morphogenesis	0.0497
	Effector	
GO:0051707	response to other organism	0.00596
GO:0034097	response to cytokine	0.00596

GO term ID	GO term name	Enrichment p-value
GO:0006955	immune response	0.00596
GO:0019079	viral genome replication	0.00596
GO:0045069	regulation of viral genome replication	0.0112
GO:0010529	negative regulation of transposition	0.0143
GO:0006952	defense response	0.0143
GO:0010528	regulation of transposition	0.0143
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	0.0188
GO:0051172	negative regulation of nitrogen compound metabolic process	0.0315
GO:0045944	positive regulation of transcription by RNA polymerase II	0.0317
GO:0051254	positive regulation of RNA metabolic process	0.0317
GO:0010557	positive regulation of macromolecule biosynthetic process	0.0423
GO:0048705	skeletal system morphogenesis	0.0453
GO:0010463	mesenchymal cell proliferation	0.0467
GO:0043374	CD8-positive, alpha-beta T cell differentiation	0.0467
GO:0065007	biological regulation	0.0467
GO:0001816	cytokine production	0.0468
GO:0060324	face development	0.0471
GO:0097152	mesenchymal cell apoptotic process	0.0483
GO:0010628	positive regulation of gene expression	0.0483
	Cell cycle	
GO:0007049	cell cycle	1.42e-58
GO:0006259	DNA metabolic process	6.06e-26
GO:0006974	cellular response to DNA damage stimulus	7.19e-13
GO:0051310	metaphase plate congression	3.75e-08
GO:0031145	anaphase-promoting complex-dependent catabolic process	6.72e-08
GO:0051169	nuclear transport	3.11e-06
GO:0034502	protein localization to chromosome	8.81e-06
GO:0000723	telomere maintenance	6e-05
GO:0007019	microtubule depolymerization	0.000169
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.000367

GO term ID	GO term name	Enrichment p-value
GO:0009263	deoxyribonucleotide biosynthetic process	0.000591
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	0.000611
GO:0006403	RNA localization	0.000798
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	0.00111
GO:0006606	protein import into nucleus	0.00166
GO:0016572	histone phosphorylation	0.00234
GO:0085020	protein K6-linked ubiquitination	0.0025
GO:0009411	response to UV	0.00308
GO:0006405	RNA export from nucleus	0.00342
GO:0002200	somatic diversification of immune receptors	0.004
GO:0001556	oocyte maturation	0.00416
GO:0045814	negative regulation of gene expression, epigenetic	0.00416
GO:0051347	positive regulation of transferase activity	0.00583
GO:0031100	animal organ regeneration	0.00639
GO:0031291	Ran protein signal transduction	0.00754
GO:1905448	positive regulation of mitochondrial ATP synthesis coupled electron transport	0.0139
GO:0006235	dTTP biosynthetic process	0.0139
GO:0046075	dTTP metabolic process	0.0139
GO:0009123	nucleoside monophosphate metabolic process	0.0207
GO:0035519	protein K29-linked ubiquitination	0.0215
GO:0044314	protein K27-linked ubiquitination	0.0215
GO:0006189	'de novo' IMP biosynthetic process	0.0294
GO:0031503	protein-containing complex localization	0.0297
GO:0009314	response to radiation	0.0322
GO:0072383	plus-end-directed vesicle transport along microtubule	0.0381
GO:0000056	ribosomal small subunit export from nucleus	0.0381
GO:1904666	regulation of ubiquitin protein ligase activity	0.0423
GO:0009157	deoxyribonucleoside monophosphate biosynthetic process	0.0483
GO:0055015	ventricular cardiac muscle cell development	0.0483
GO:0000055	ribosomal large subunit export from nucleus	0.0483

GO term ID	GO term name	Enrichment p-value
GO:0075713	establishment of integrated proviral latency	0.0483
	Chromatin organisation	
GO:0051171	regulation of nitrogen compound metabolic process	1.42e-06
GO:0016070	RNA metabolic process	2.46e-06
GO:0006325	chromatin organization	5.81e-06
GO:0018394	peptidyl-lysine acetylation	2.75e-05
GO:0010605	negative regulation of macromolecule metabolic process	3.4e-05
GO:0006403	RNA localization	0.000411
GO:0050657	nucleic acid transport	0.000854
GO:0048511	rhythmic process	0.00139
GO:1902400	intracellular signal transduction involved in G1 DNA damage checkpoint	0.00189
GO:0072431	signal transduction involved in mitotic G1 DNA damage checkpoint	0.00189
GO:0010604	positive regulation of macromolecule metabolic process	0.00217
GO:0071426	ribonucleoprotein complex export from nucleus	0.00218
GO:0030330	DNA damage response, signal transduction by p53 class mediator	0.00241
GO:0016575	histone deacetylation	0.00308
GO:0006337	nucleosome disassembly	0.00361
GO:2000773	negative regulation of cellular senescence	0.00361
GO:0000289	nuclear-transcribed mRNA poly(A) tail shortening	0.00412
GO:0080182	histone H3-K4 trimethylation	0.0044
GO:0016447	somatic recombination of immunoglobulin gene segments	0.00476
GO:0032233	positive regulation of actin filament bundle assembly	0.00793
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.00842
GO:0050872	white fat cell differentiation	0.01

GO term ID	GO term name	Enrichment p-value
GO:0045023	G0 to G1 transition	0.0167
GO:0060338	regulation of type I interferon-mediated signaling pathway	0.0182
GO:0090500	endocardial cushion to mesenchymal transition	0.0194
GO:0051315	attachment of mitotic spindle microtubules to kinetochore	0.0212
GO:0010948	negative regulation of cell cycle process	0.0346
GO:0072655	establishment of protein localization to mitochondrion	0.0361
GO:0031647	regulation of protein stability	0.0396
GO:1902177	positive regulation of oxidative stress-induced intrinsic apoptotic signaling pathway	0.0469
GO:0070345	negative regulation of fat cell proliferation	0.0469

Table E.2 GO term enrichment poly(I:C) response gene modules

GO term ID	GO term name	Enrichment p-value
	Canonical Type I IFN	
GO:0045087	innate immune response	4.73e-41
GO:0051607	defense response to virus	5.82e-33
GO:0019221	cytokine-mediated signaling pathway	8.24e-27
GO:0032479	regulation of type I interferon production	6.38e-17
GO:0012501	programmed cell death	1.1e-12
GO:0051092	positive regulation of NF-kappaB transcription factor activity	8.8e-11
GO:0042127	regulation of cell proliferation	6.1e-05
GO:1903463	regulation of mitotic cell cycle DNA replication	0.000226
GO:0042270	protection from natural killer cell mediated cytotoxicity	0.000226
GO:0070383	DNA cytosine deamination	0.00155
GO:0045869	negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	0.00274
GO:0031087	deadenylation-independent decapping of nuclear-transcribed mRNA	0.00274

GO term ID	GO term name	Enrichment p-value
GO:0001568	blood vessel development	0.0035
GO:0061180	mammary gland epithelium development	0.00398
GO:0001525	angiogenesis	0.00479
GO:0016553	base conversion or substitution editing	0.00888
GO:0070423	nucleotide-binding oligomerization domain containing signaling pathway	0.00947
GO:0007569	cell aging	0.00965
GO:0030218	erythrocyte differentiation	0.0101
GO:0097343	ripoptosome assembly	0.0144
GO:2000045	regulation of G1/S transition of mitotic cell cycle	0.0179
GO:0007159	leukocyte cell-cell adhesion	0.0201
GO:0010594	regulation of endothelial cell migration	0.0201
GO:0046208	spermine catabolic process	0.0243
GO:1904798	positive regulation of core promoter binding	0.0243
GO:0035282	segmentation	0.0253
GO:0010528	regulation of transposition	0.0258
GO:0032088	negative regulation of NF-kappaB transcription factor activity	0.0294
GO:0034356	NAD biosynthesis via nicotinamide riboside salvage pathway	0.03
GO:0032495	response to muramyl dipeptide	0.0349
GO:0014732	skeletal muscle atrophy	0.0349
GO:0042359	vitamin D metabolic process	0.0349
GO:0140052	cellular response to oxidised low-density lipoprotein particle stimulus	0.0349
GO:0002730	regulation of dendritic cell cytokine production	0.0349
GO:0048289	isotype switching to IgE isotypes	0.0349
GO:0010667	negative regulation of cardiac muscle cell apoptotic process	0.0397
GO:0046135	pyrimidine nucleoside catabolic process	0.0397
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.0417
GO:0022612	gland morphogenesis	0.0466
GO:0010966	regulation of phosphate transport	0.0467

GO term ID	GO term name	Enrichment p-value
GO:1990668	vesicle fusion with endoplasmic reticulum-Golgi intermediate compartment (ERGIC) membrane	0.0467
GO:0032436	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	0.0482
	Mitochondrial	
GO:0006119	oxidative phosphorylation	3.38e-15
GO:0022904	respiratory electron transport chain	2.84e-14
GO:0032981	mitochondrial respiratory chain complex I assembly	1.06e-08
GO:1902600	proton transmembrane transport	5.77e-07
GO:0046597	negative regulation of viral entry into host cell	0.00232
GO:0006979	response to oxidative stress	0.00263
GO:0055093	response to hyperoxia	0.0029
GO:0000028	ribosomal small subunit assembly	0.00348
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	0.00348
GO:0035455	response to interferon-alpha	0.00348
GO:0035456	response to interferon-beta	0.00494
GO:0046688	response to copper ion	0.0054
GO:0046677	response to antibiotic	0.00578
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	0.00578
GO:0042493	response to drug	0.00694
GO:0042407	cristae formation	0.00727
GO:0035094	response to nicotine	0.00761
GO:0045071	negative regulation of viral genome replication	0.0134
GO:0010729	positive regulation of hydrogen peroxide biosynthetic process	0.0168
GO:0006413	translational initiation	0.0181
GO:0071357	cellular response to type I interferon	0.0261
GO:0010035	response to inorganic substance	0.0261
GO:0072513	positive regulation of secondary heart field cardioblast proliferation	0.0283
GO:0021549	cerebellum development	0.0306
GO:0007568	aging	0.0316

GO term ID	GO term name	Enrichment p-value
GO:0002181	cytoplasmic translation	0.0377
GO:0042538	hyperosmotic salinity response	0.0388
GO:0010940	positive regulation of necrotic cell death	0.0388
GO:0048318	axial mesoderm development	0.0439
GO:0046689	response to mercury ion	0.0439
	Signal transduction	
	No significant gene sets	
	Organelle localisation	
GO:0006996	organelle organization	7.97e-06
GO:0051649	establishment of localization in cell	3.28e-05
GO:0090066	regulation of anatomical structure size	0.00111
GO:0007165	signal transduction	0.00111
GO:0006464	cellular protein modification process	0.00111
GO:0042060	wound healing	0.00111
GO:0036211	protein modification process	0.00111
GO:0016032	viral process	0.00123
GO:0016477	cell migration	0.00123
GO:0009057	macromolecule catabolic process	0.00615
GO:0070936	protein K48-linked ubiquitination	0.00664
GO:0006793	phosphorus metabolic process	0.0076
GO:0070534	protein K63-linked ubiquitination	0.0145
GO:1901660	calcium ion export	0.0146
GO:0044265	cellular macromolecule catabolic process	0.0146
GO:0007167	enzyme linked receptor protein signaling pathway	0.0147
GO:0007049	cell cycle	0.0175
GO:1902309	negative regulation of peptidyl-serine dephosphorylation	0.0185
GO:0090435	protein localization to nuclear envelope	0.0186
GO:1990314	cellular response to insulin-like growth factor stimulus	0.0186
GO:0031329	regulation of cellular catabolic process	0.0238
GO:0010769	regulation of cell morphogenesis involved in differentiation	0.026
GO:0061037	negative regulation of cartilage development	0.0262
GO:0030900	forebrain development	0.0262

GO term ID	GO term name	Enrichment p-value
GO:0007599	hemostasis	0.0291
GO:1903984	positive regulation of TRAIL-activated apoptotic signaling pathway	0.0363
GO:0060978	angiogenesis involved in coronary vascular morphogenesis	0.0363
GO:0097350	neutrophil clearance	0.0363
GO:0001655	urogenital system development	0.0369
GO:0051301	cell division	0.0374
GO:0043010	camera-type eye development	0.0381
GO:0044794	positive regulation by host of viral process	0.0423
GO:0048609	multicellular organismal reproductive process	0.0443
GO:0021543	pallium development	0.046
GO:0050765	negative regulation of phagocytosis	0.0488
	Metabolic processes	
GO:0006810	transport	9.37e-13
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3.09e-08
GO:0006518	peptide metabolic process	2.12e-06
GO:0006735	NADH regeneration	2.87e-05
GO:0061621	canonical glycolysis	2.87e-05
GO:0021762	substantia nigra development	3e-05
GO:1901844	regulation of cell communication by electrical coupling involved in cardiac conduction	0.000644
GO:0034976	response to endoplasmic reticulum stress	0.00124
GO:0036500	ATF6-mediated unfolded protein response	0.00124
GO:0075206	positive regulation by host of symbiont cAMP-mediated signal transduction	0.00321
GO:0010524	positive regulation of calcium ion transport into cytosol	0.00383
GO:0006094	gluconeogenesis	0.00473
GO:0048013	ephrin receptor signaling pathway	0.00473
GO:0006936	muscle contraction	0.00532
GO:0051343	positive regulation of cyclic-nucleotide phosphodiesterase activity	0.00562
GO:0051186	cofactor metabolic process	0.00666

GO term ID	GO term name	Enrichment p-value
GO:0097066	response to thyroid hormone	0.00852
GO:0060314	regulation of ryanodine-sensitive calcium-release channel activity	0.00955
GO:0001568	blood vessel development	0.00965
GO:0044092	negative regulation of molecular function	0.0148
GO:0050790	regulation of catalytic activity	0.017
GO:0001893	maternal placenta development	0.018
GO:0097064	ncRNA export from nucleus	0.0183
GO:2001235	positive regulation of apoptotic signaling pathway	0.0194
GO:0045792	negative regulation of cell size	0.0223
GO:0001765	membrane raft assembly	0.0223
GO:0038093	Fc receptor signaling pathway	0.0223
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	0.0225
GO:0002478	antigen processing and presentation of exogenous peptide antigen	0.0228
GO:0033631	cell-cell adhesion mediated by integrin	0.0262
GO:0045214	sarcomere organization	0.0269
GO:0050999	regulation of nitric-oxide synthase activity	0.0269
GO:0018279	protein N-linked glycosylation via asparagine	0.0269
GO:0031623	receptor internalization	0.0274
GO:0042060	wound healing	0.0288
GO:0035304	regulation of protein dephosphorylation	0.0301
GO:0000910	cytokinesis	0.0306
GO:0022898	regulation of transmembrane transporter activity	0.0334
GO:0001667	ameboidal-type cell migration	0.036
GO:0031952	regulation of protein autophosphorylation	0.0383
GO:0019511	peptidyl-proline hydroxylation	0.0393
GO:0042744	hydrogen peroxide catabolic process	0.0433
GO:0022604	regulation of cell morphogenesis	0.0446
GO:0007596	blood coagulation	0.0463
GO:0034381	plasma lipoprotein particle clearance	0.0469
GO:0031639	plasminogen activation	0.0469

GO term ID	GO term name	Enrichment p-value
GO:0032516	positive regulation of phosphoprotein phosphatase activity	0.0469
GO:0007166	cell surface receptor signaling pathway	0.047
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	0.0486
GO:0043687	post-translational protein modification	0.0489
GO:0044147	negative regulation of development of symbiont involved in interaction with host	0.0491
GO:1905581	positive regulation of low-density lipoprotein particle clearance	0.0491
GO:1903673	mitotic cleavage furrow formation	0.0491
GO:1904313	response to methamphetamine hydrochloride	0.0491
GO:1903609	negative regulation of inward rectifier potassium channel activity	0.0491
GO:0002842	positive regulation of T cell mediated immune response to tumor cell	0.0491
GO:1905152	positive regulation of voltage-gated sodium channel activity	0.0491
GO:1904695	positive regulation of vascular smooth muscle contraction	0.0491
GO:0071528	tRNA re-export from nucleus	0.0491
GO:1904401	cellular response to Thyroid stimulating hormone	0.0491
GO:0050832	defense response to fungus	0.0491
GO:2000811	negative regulation of anoikis	0.0491
GO:0035606	peptidyl-cysteine S-trans-nitrosylation	0.0491
GO:0003081	regulation of systemic arterial blood pressure by renin-angiotensin	0.0491
GO:0051621	regulation of norepinephrine uptake	0.0491
GO:0002368	B cell cytokine production	0.0491
GO:0060051	negative regulation of protein glycosylation	0.0491
GO:0010801	negative regulation of peptidyl-threonine phosphorylation	0.0491
GO:0034238	macrophage fusion	0.0491

GO term ID	GO term name	Enrichment p-value
GO:2000147	positive regulation of cell motility	0.0491
GO:1905597	positive regulation of low-density lipoprotein particle receptor binding	0.0491
GO:0001998	angiotensin-mediated vasoconstriction involved in regulation of systemic arterial blood pressure	0.0491
GO:1900085	negative regulation of peptidyl-tyrosine autophosphorylation	0.0491
	Protein regulation	
GO:0070972	protein localization to endoplasmic reticulum	5.21e-13
GO:0006413	translational initiation	4.39e-08
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	8.04e-08
GO:0044403	symbiont process	5.71e-05
GO:0070887	cellular response to chemical stimulus	0.000588
GO:0043687	post-translational protein modification	0.00139
GO:0097435	supramolecular fiber organization	0.00208
GO:0050821	protein stabilization	0.0028
GO:0002376	immune system process	0.00356
GO:0001666	response to hypoxia	0.00753
GO:0051897	positive regulation of protein kinase B signaling	0.00803
GO:0034309	primary alcohol biosynthetic process	0.0114
GO:0048251	elastic fiber assembly	0.017
GO:0043312	neutrophil degranulation	0.0193
GO:0017015	regulation of transforming growth factor beta receptor signaling pathway	0.0196
GO:2000121	regulation of removal of superoxide radicals	0.0198
GO:0006089	lactate metabolic process	0.0303
GO:0048678	response to axon injury	0.0336
GO:0044409	entry into host	0.0346
GO:0018208	peptidyl-proline modification	0.0346
GO:0010608	posttranscriptional regulation of gene expression	0.0363
GO:0031333	negative regulation of protein complex assembly	0.0363
GO:1903206	negative regulation of hydrogen peroxide-induced cell death	0.0477

GO term ID	GO term name	Enrichment p-value
GO:0009651	response to salt stress	0.0477
GO:1901388	regulation of transforming growth factor beta activation	0.0479
GO:1903189	glyoxal metabolic process	0.0479
GO:0048693	regulation of collateral sprouting of injured axon	0.0479
GO:1903200	positive regulation of L-dopa decarboxylase activity	0.0479
GO:1903072	regulation of death-inducing signaling complex assembly	0.0479
GO:0034120	positive regulation of erythrocyte aggregation	0.0479
GO:1901194	negative regulation of formation of translation preinitiation complex	0.0479
GO:0045454	cell redox homeostasis	0.0479
GO:0036531	glutathione deglycation	0.0479
GO:1990478	response to ultrasound	0.0479
GO:1903195	regulation of L-dopa biosynthetic process	0.0479
GO:1902546	positive regulation of DNA N-glycosylase activity	0.0479
GO:0036529	protein deglycation, glyoxal removal	0.0479
GO:0140041	cellular detoxification of methylglyoxal	0.0479
GO:1903197	positive regulation of L-dopa biosynthetic process	0.0479
GO:2001272	positive regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis	0.0479
GO:0048689	formation of growth cone in injured axon	0.0479
GO:0018323	enzyme active site formation via L-cysteine sulfinic acid	0.0479
GO:0018032	protein amidation	0.0479
GO:1903168	positive regulation of pyrroline-5-carboxylate reductase activity	0.0479
GO:2000277	positive regulation of oxidative phosphorylation uncoupler activity	0.0479
GO:0106046	guanine deglycation, glyoxal removal	0.0479

GO term ID	GO term name	Enrichment p-value
GO:0015872	dopamine transport	0.0479
GO:2001023	regulation of response to drug	0.0479
GO:1990262	anti-Mullerian hormone signaling pathway	0.0479
GO:1905578	regulation of ERBB3 signaling pathway	0.0479
GO:1903122	negative regulation of TRAIL-activated apoptotic signaling pathway	0.0479
GO:0032535	regulation of cellular component size	0.0479
GO:1903659	regulation of complement-dependent cytotoxicity	0.0479
GO:1903176	regulation of tyrosine 3-monooxygenase activity	0.0479
GO:1905572	ganglioside GM1 transport to membrane	0.0479

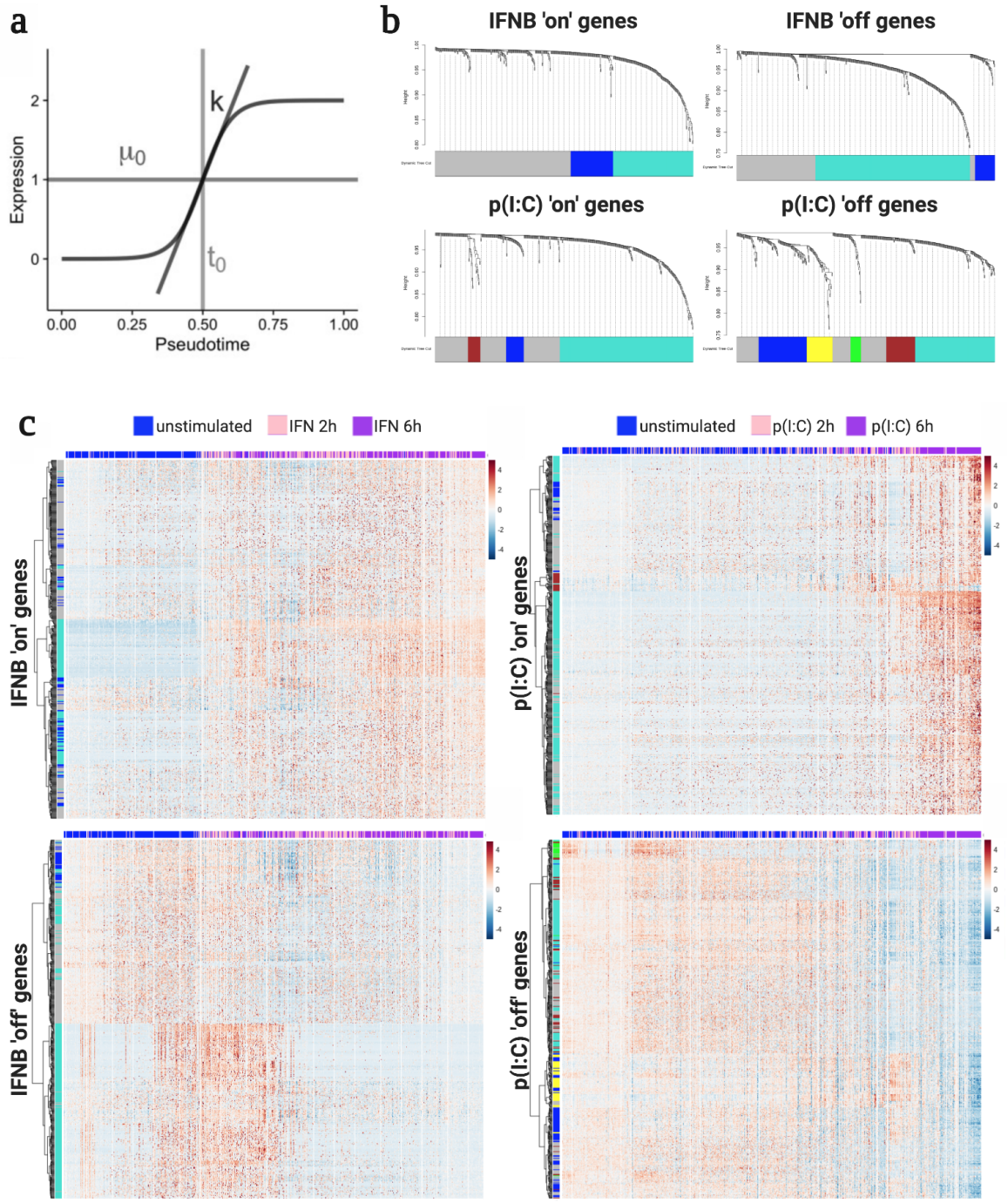


Fig. E.1 Modules of co-expressed innate immune response genes using WGCNA. a) The SwitchDE package [97] was used to infer a dynamic model of expression for each gene. b) WGCNA was applied to detect modules of co-expressed genes. The 500 most significant up- and down- regulated genes in each response pathway were used; dendrogram and inferred clusters are shown. Dynamic tree cutting approach was used to determine the optimum number of gene clusters. c) The expression of these genes over 'IFN pathway', left, and 'poly(I:C) pathway', right, are shown. Expression values are scaled within each row, and WGCNA cluster assignment is shown in the left-side colour bar.