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

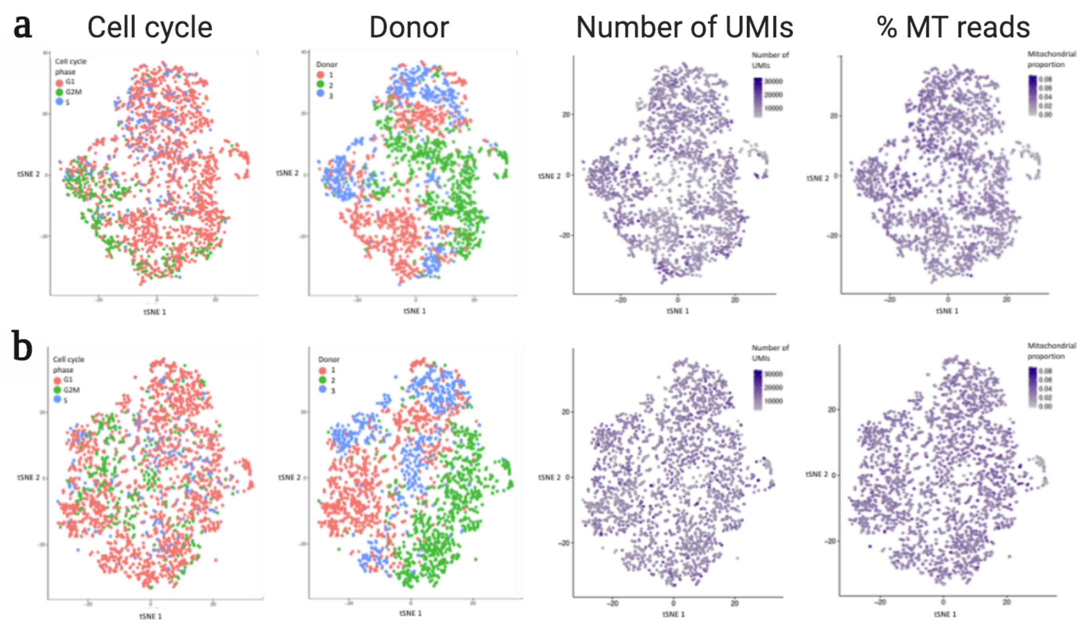

Fig. 3.1 An overview of a pilot droplet scRNA-seq dataset. a) tSNE visualisations coloured by cell cycle phase, donor, number of UMIs, and mitochondrial read proportion. b) Repeat of the tSNE visualisations after regression of cell cycle, number of UMIs and mitochondrial proportion.

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.

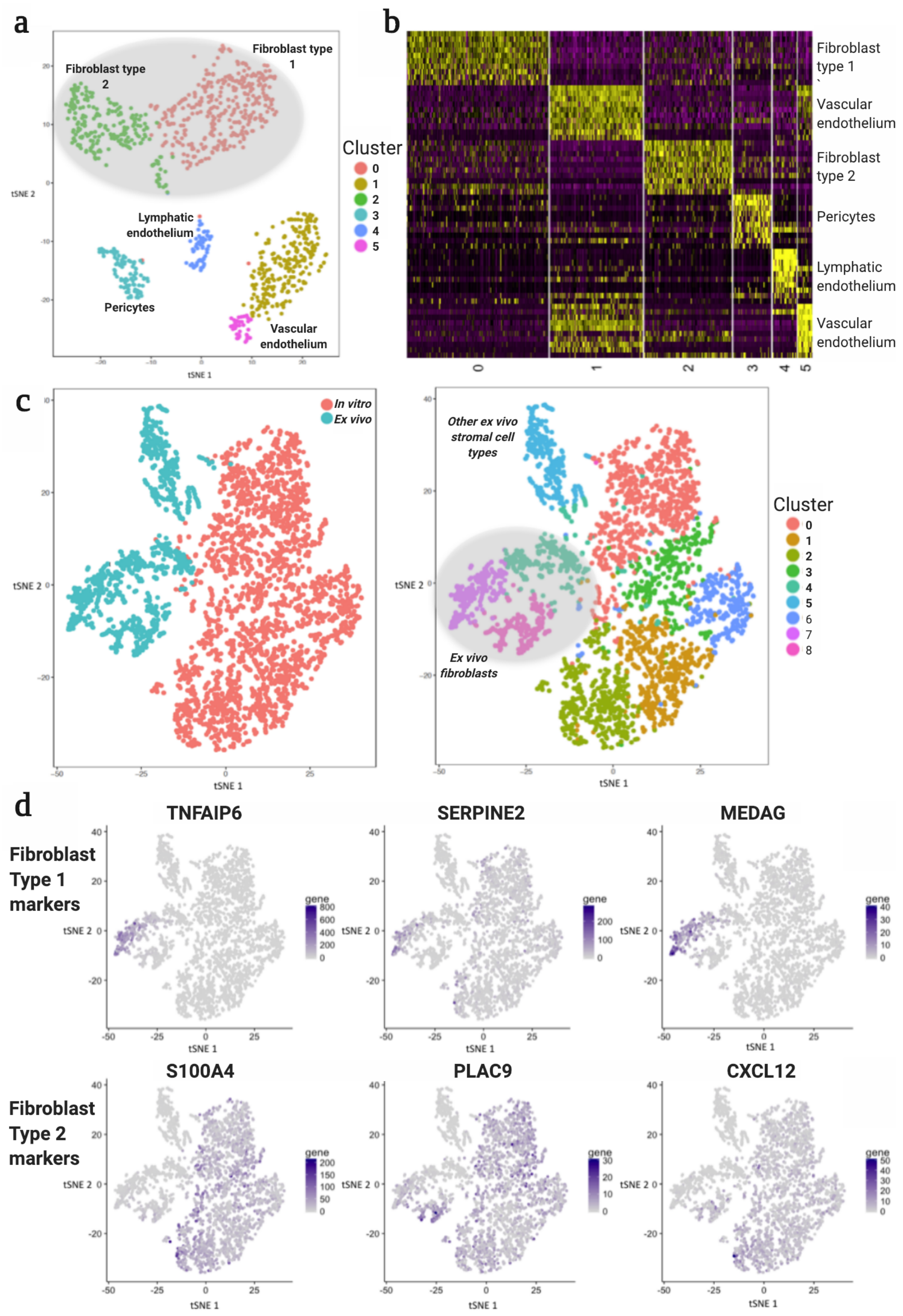

Fig. 3.2 Comparison of *in vitro* and *ex vivo* fibroblasts. a) tSNE visualisation and clustering of *ex vivo* skin cells; fibroblasts are shaded in grey. b) Top 10 differentially expressed markers for each cluster; full list with gene names in Table B.1. c) tSNE of merged *ex vivo* and *in vitro* datasets. 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.

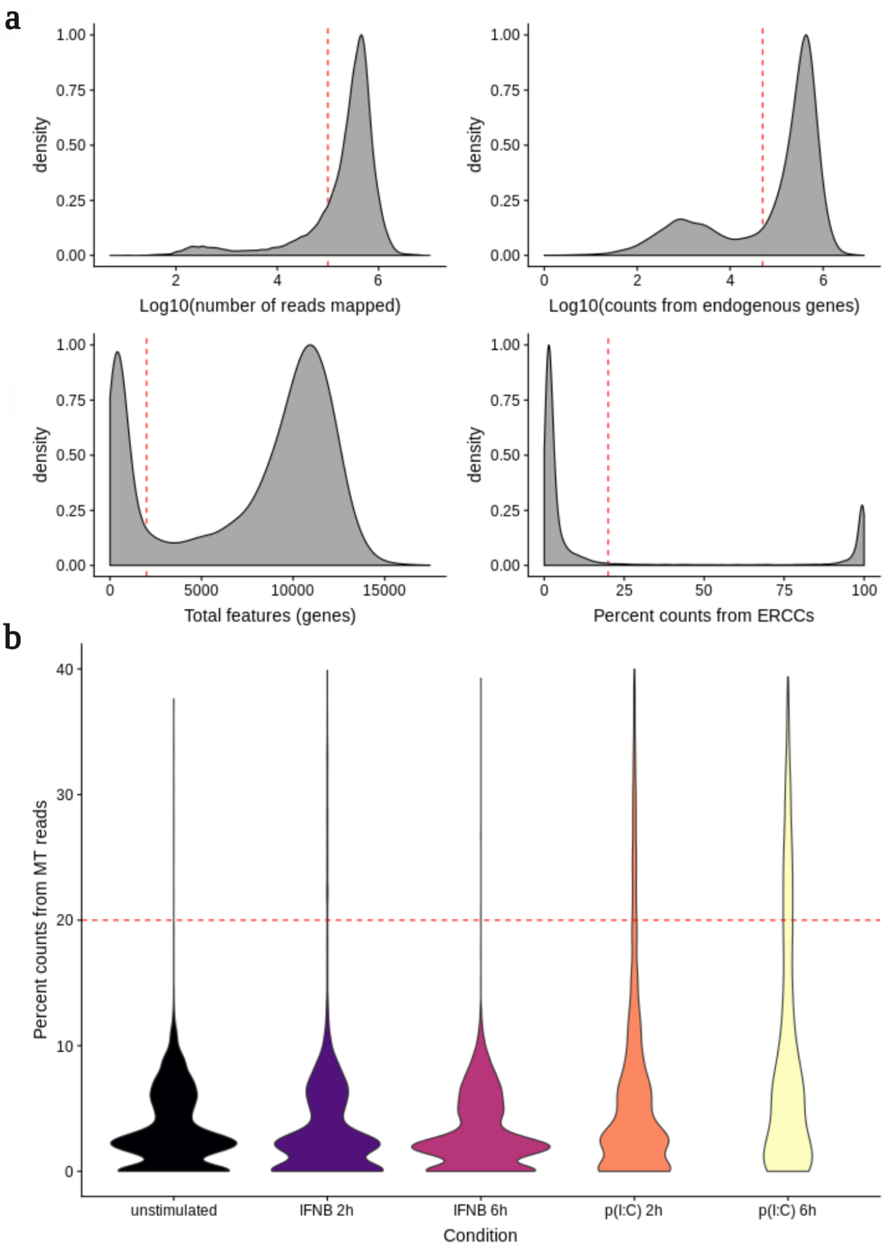

Fig. 3.3 Quality control of scRNA-seq data. a) Distribution of technical factors across cells: number of mapped reads, counts from endogenous reads, total features, ERCC percentage. Thresholds used for filtering cells shown in red: greater than 100,000 reads mapped, greater than 50,000 counts from endogenous 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

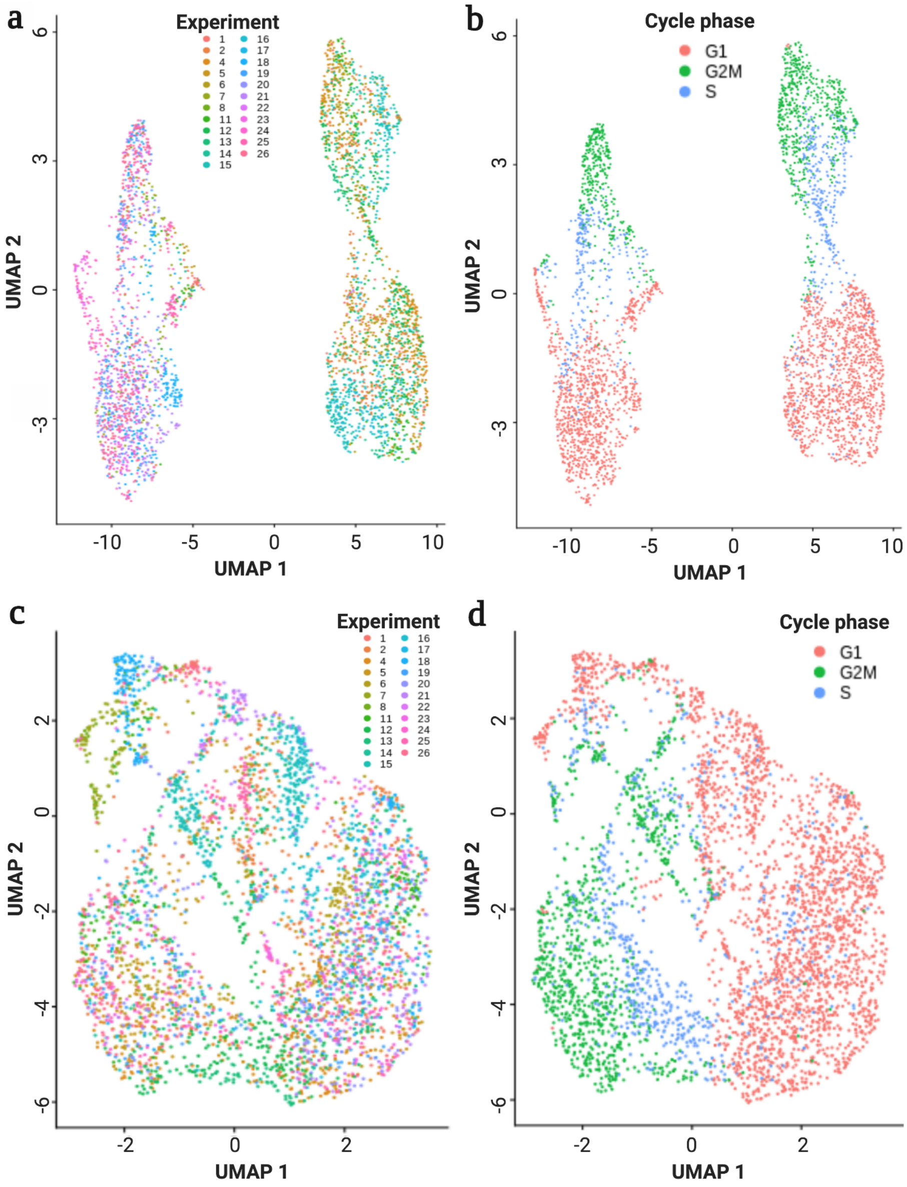

Fig. 3.4 Integration of scRNA-seq batches with Seurat. a) Dimensionality reduction using UMAP on uncorrected data: left, coloured by experimental batch, right, coloured by cell cycle phase. The first two UMAP dimensions are shown. b) UMAP plots after using Seurat v3's 'integrate' method: left, by batch,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

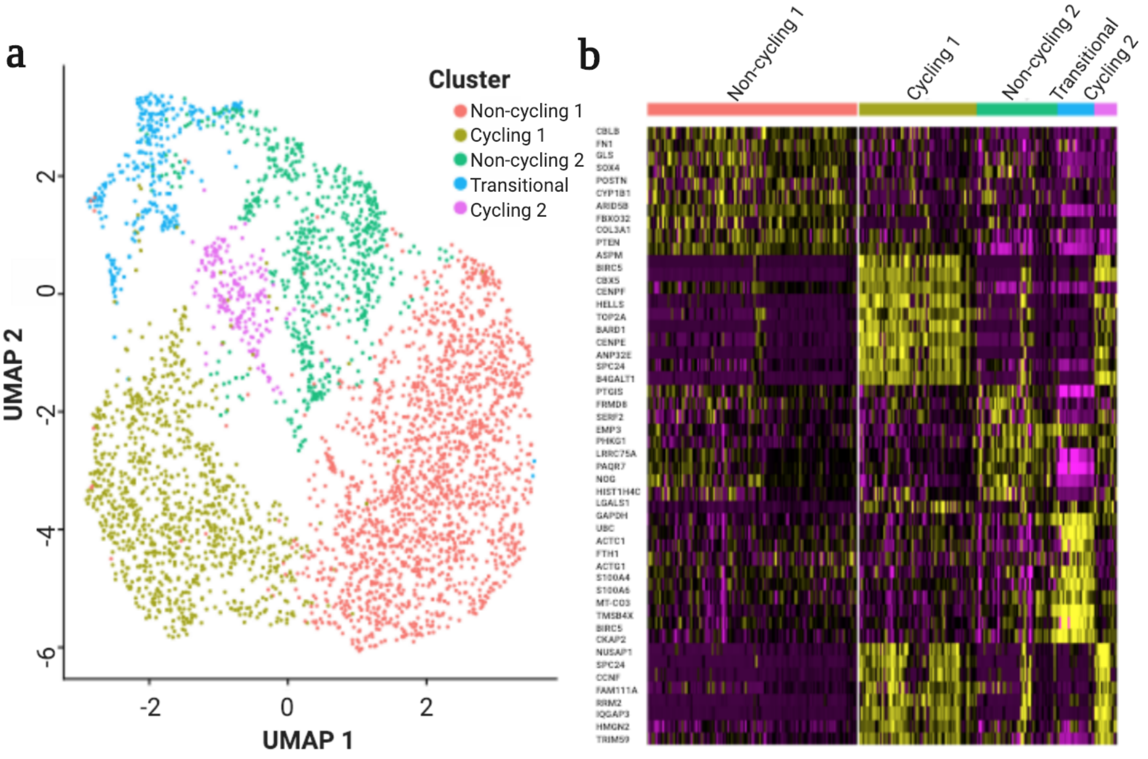

Fig. 3.5 Clustering analysis of unstimulated fibroblasts a) Clusters identified using Seurat v3's graph-based clustering approach, applying the 'FindNeighbours' and 'FindClusters' functions, with a resolution of 0.2. Five clusters are identified. b) The top 10 marker genes for each cluster are shown,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

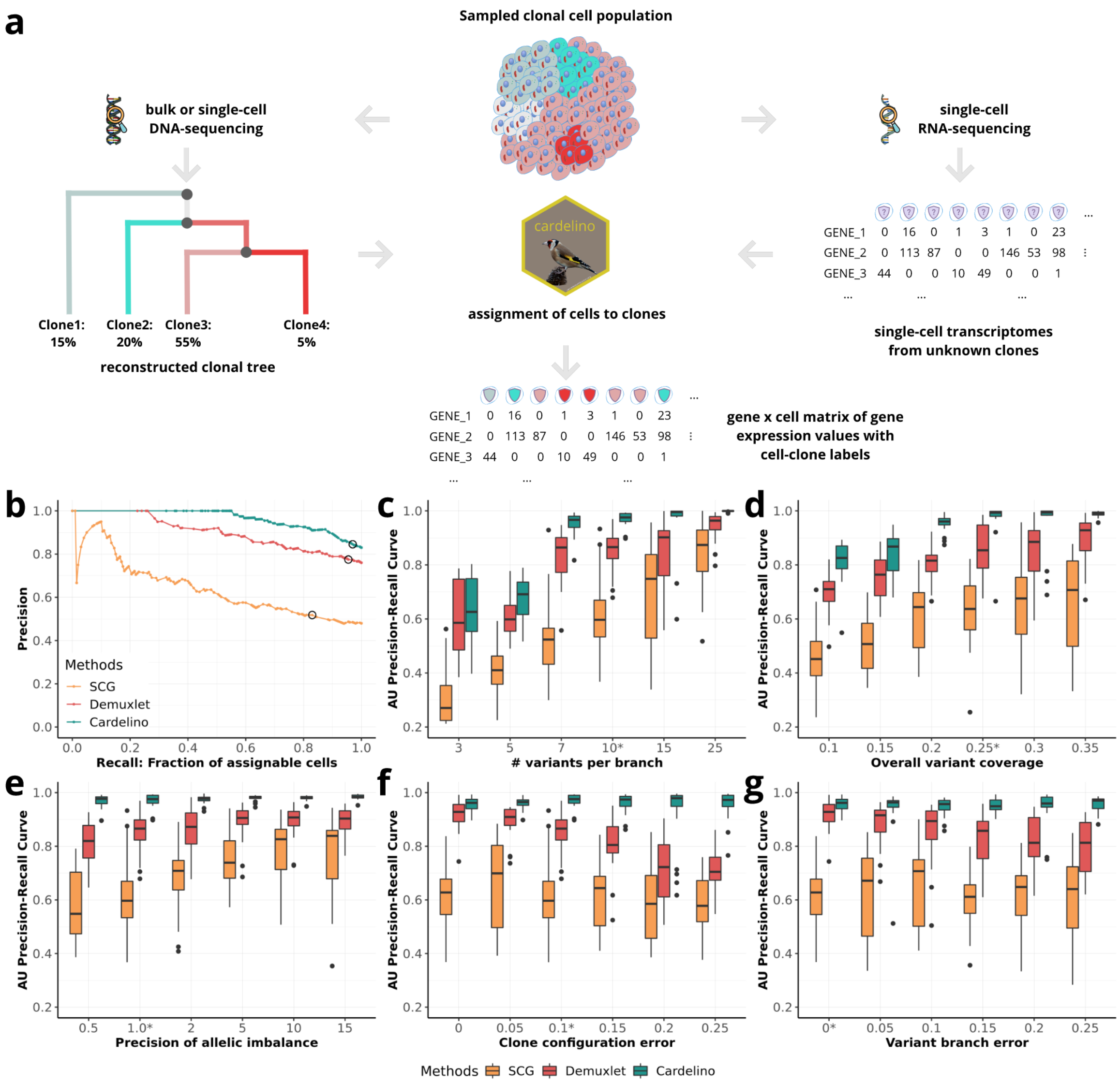

Fig. 3.6 Overview and validation of the cardelino model. (a) Overview and approach. A clonal tree is reconstructed using DNA-sequencing data to derive a guide clone configuration. Cardelino then performs probabilistic clustering of single-cell transcriptomes based on variants detected in scRNA-seq reads, 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

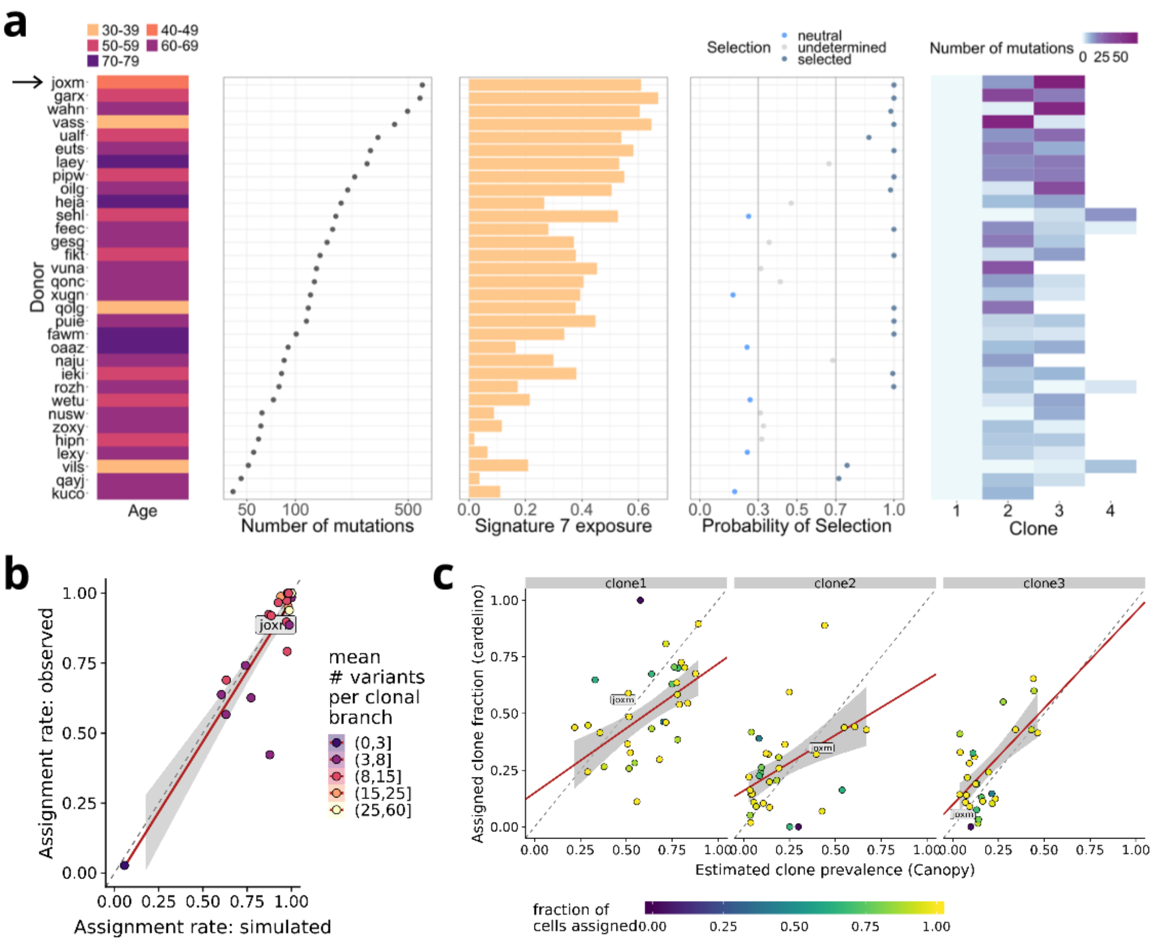

Fig. 3.7 Characterisation of mutational and clonal structure in 32 fibroblast lines. a) Overview and somatic mutation profiles across lines (donors), from left to right: donor age; number of somatic SNVs; estimated exposure of COSMIC mutational signature 7; probability of selection estimated by SubClonalSelection [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

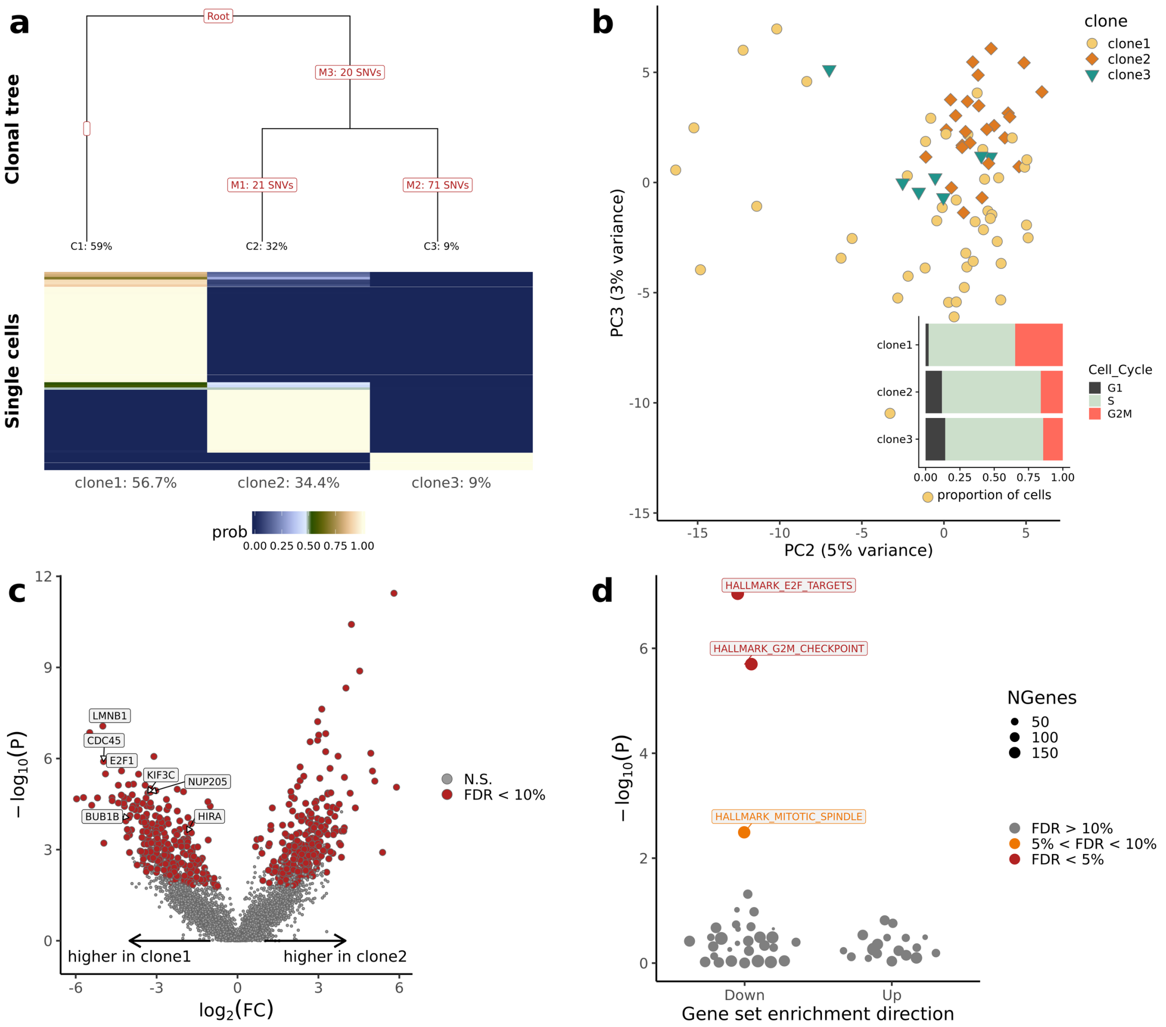

Fig. 3.8 Clone-specific transcriptome profiles reveal gene expression differences for joxm, one example line. (a) Top: Clonal tree inferred using Canopy [149]. The number of variants tagging each branch and the expected prevalence (fraction) of each clone is shown. Bottom: cardelino cell assignment matrix, 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.

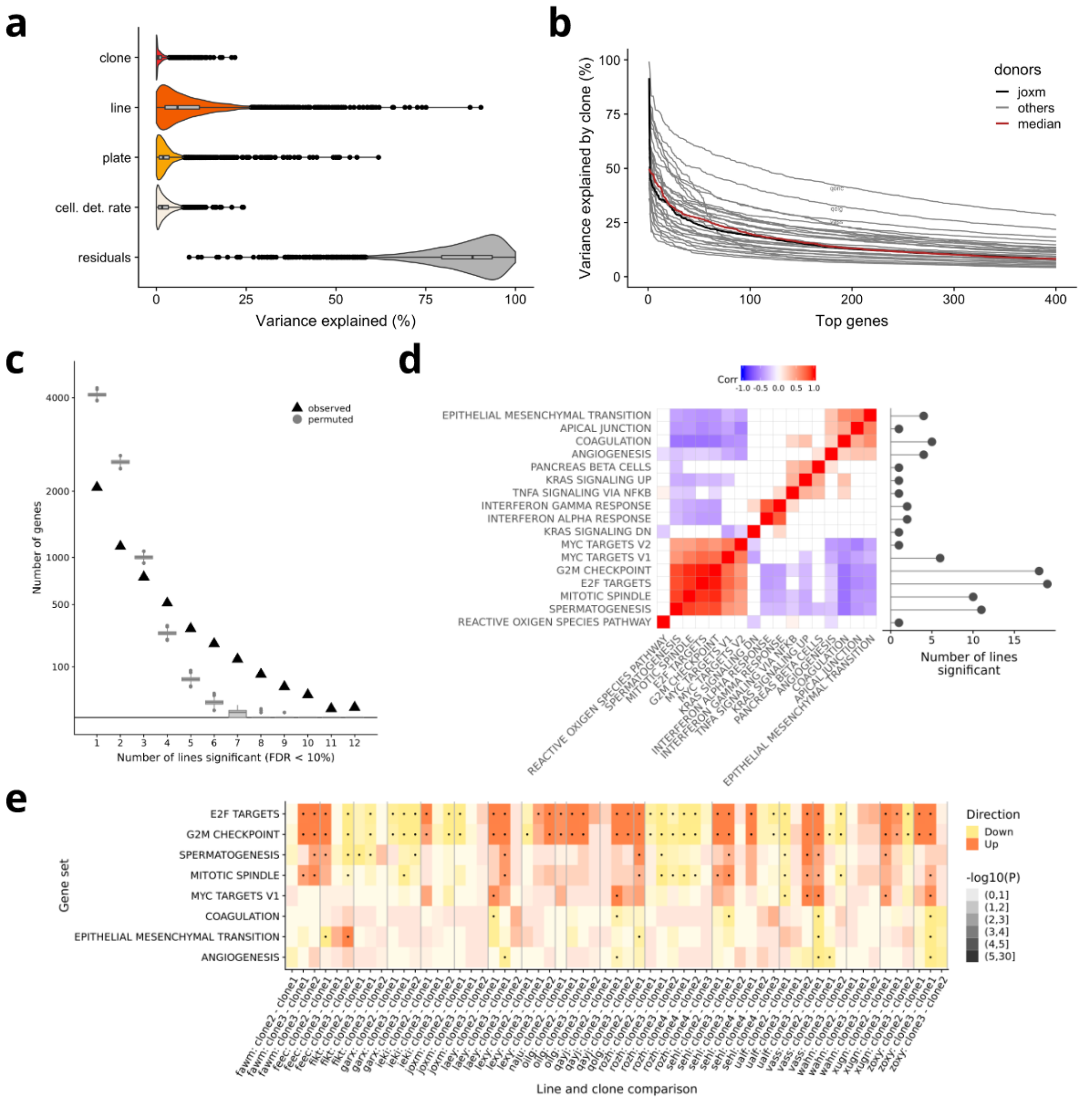

Fig. 3.9 Signatures of transcriptomic clone-to-clone variation across 31 lines. (a) Violin and box plots show the percentage of variance explained by clone, line, experimental plate and cellular detection rate for 4,998 highly variable genes, estimated using a linear mixed model (Methods; Appendix C). (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.