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**).

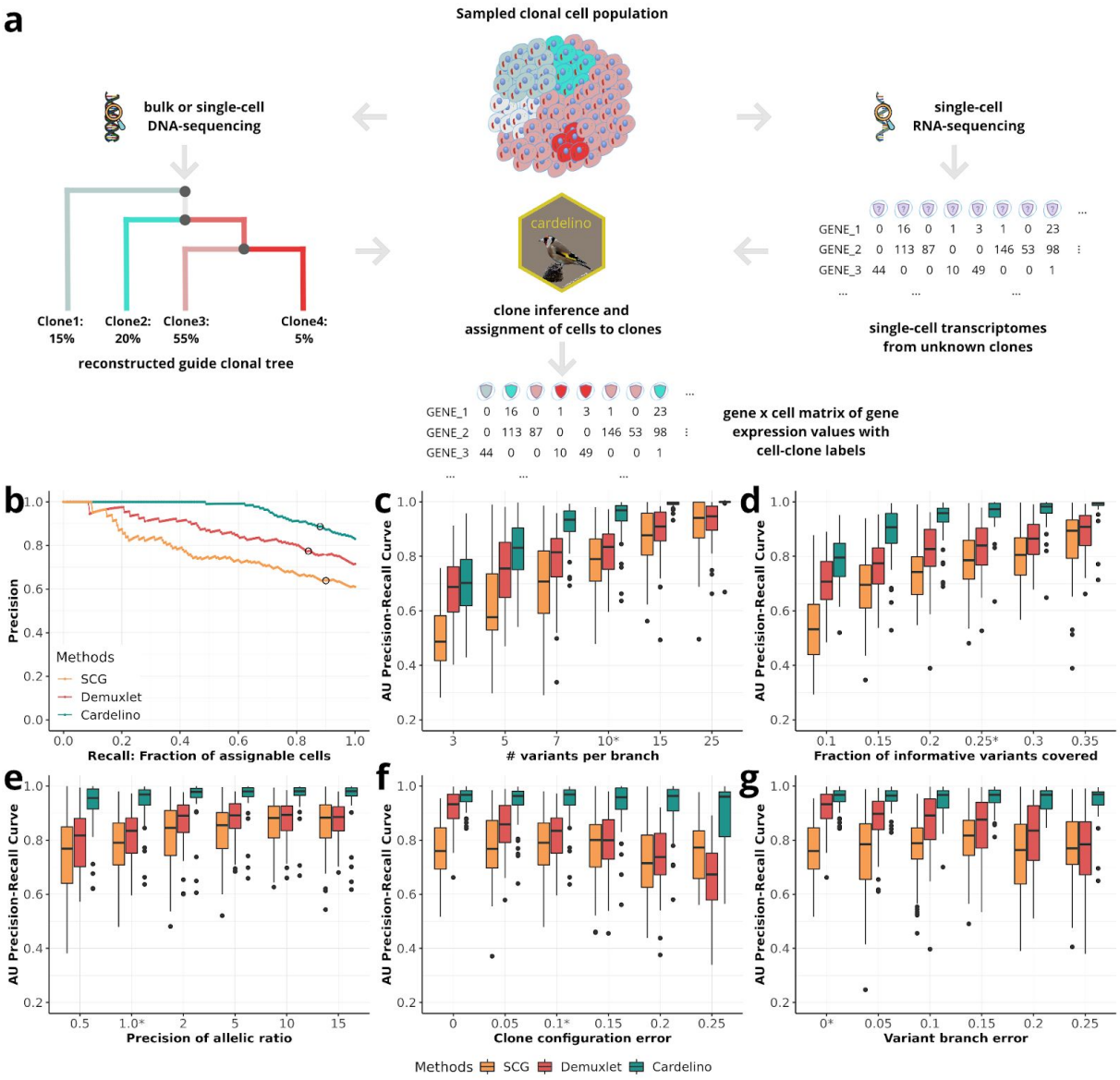


Figure 1 | Overview and validation of the cardelino model. (a) Overview and approach. A clonal tree is reconstructed using DNA-sequencing (e.g. deep exome sequencing) data to derive a guide clone configuration. Cardelino then performs probabilistic clustering of single-cell transcriptomes based on variants 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**).

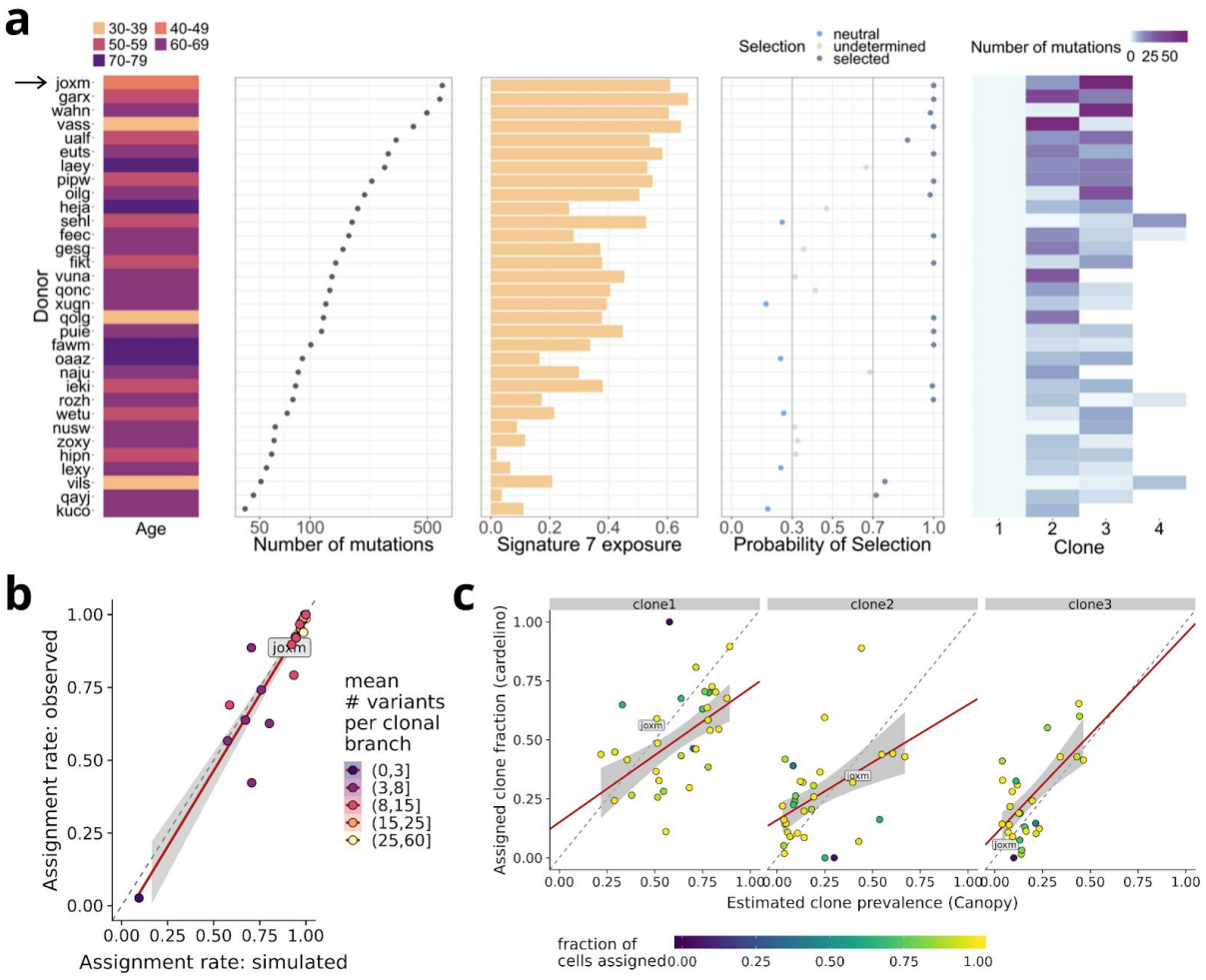


Figure 2 | Parallel deep exome sequencing and scRNA-seq profiling of 32 human dermal fibroblast lines. (a) Overview and somatic mutation profiles across lines, from left to right: donor age; number of somatic SNVs; estimated exposure of COSMIC mutational signature 7; probability of selection estimated 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.

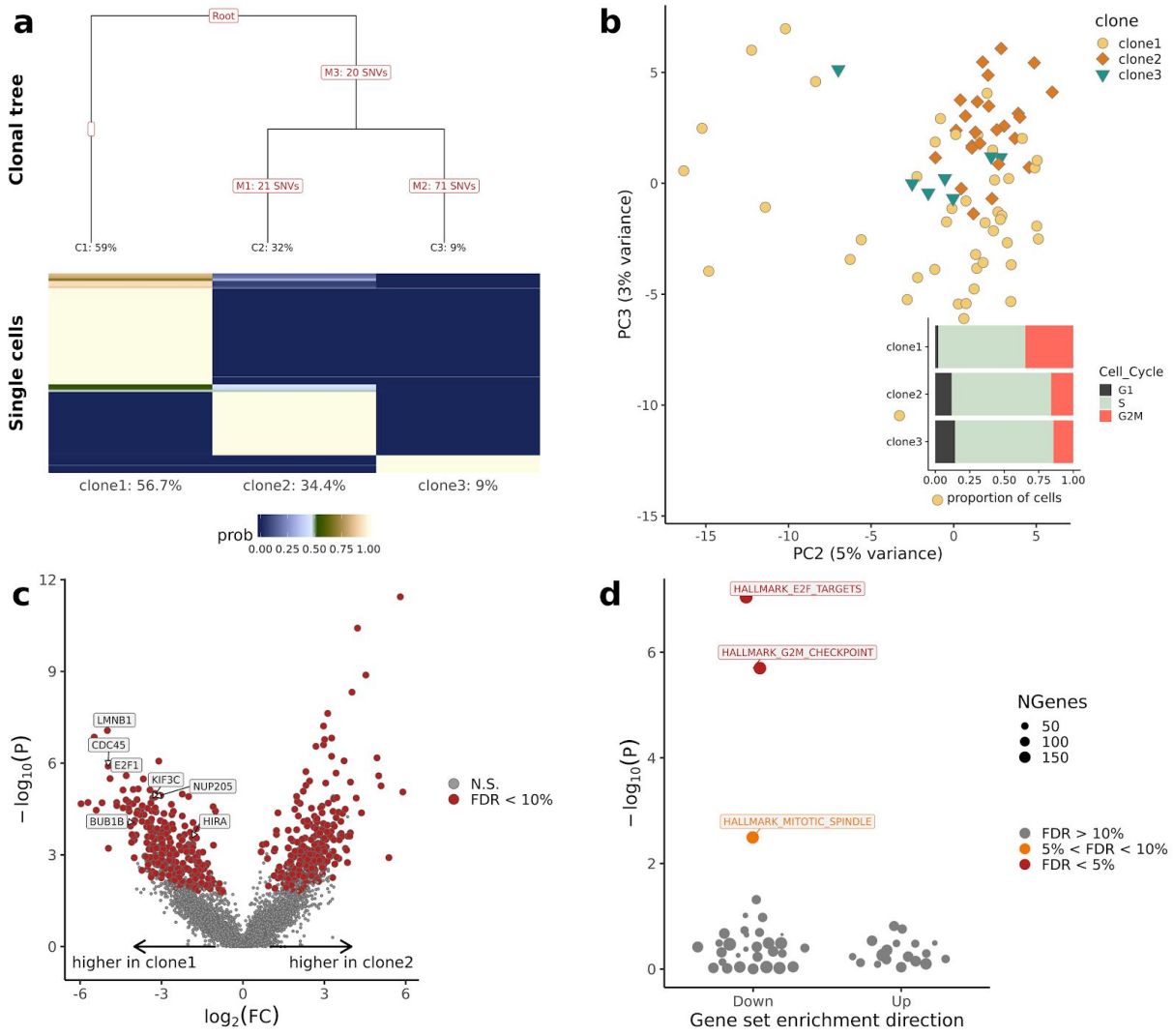


Figure 3 | Clone-specific transcriptome profiles reveal gene expression differences for *joxm,* one example line. (a) Top: Clonal tree inferred using Canopy (Jiang et al., 2016). The number of variants tagging each branch and the expected prevalence (fraction) of each clone is shown. Bottom: cardelino 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.

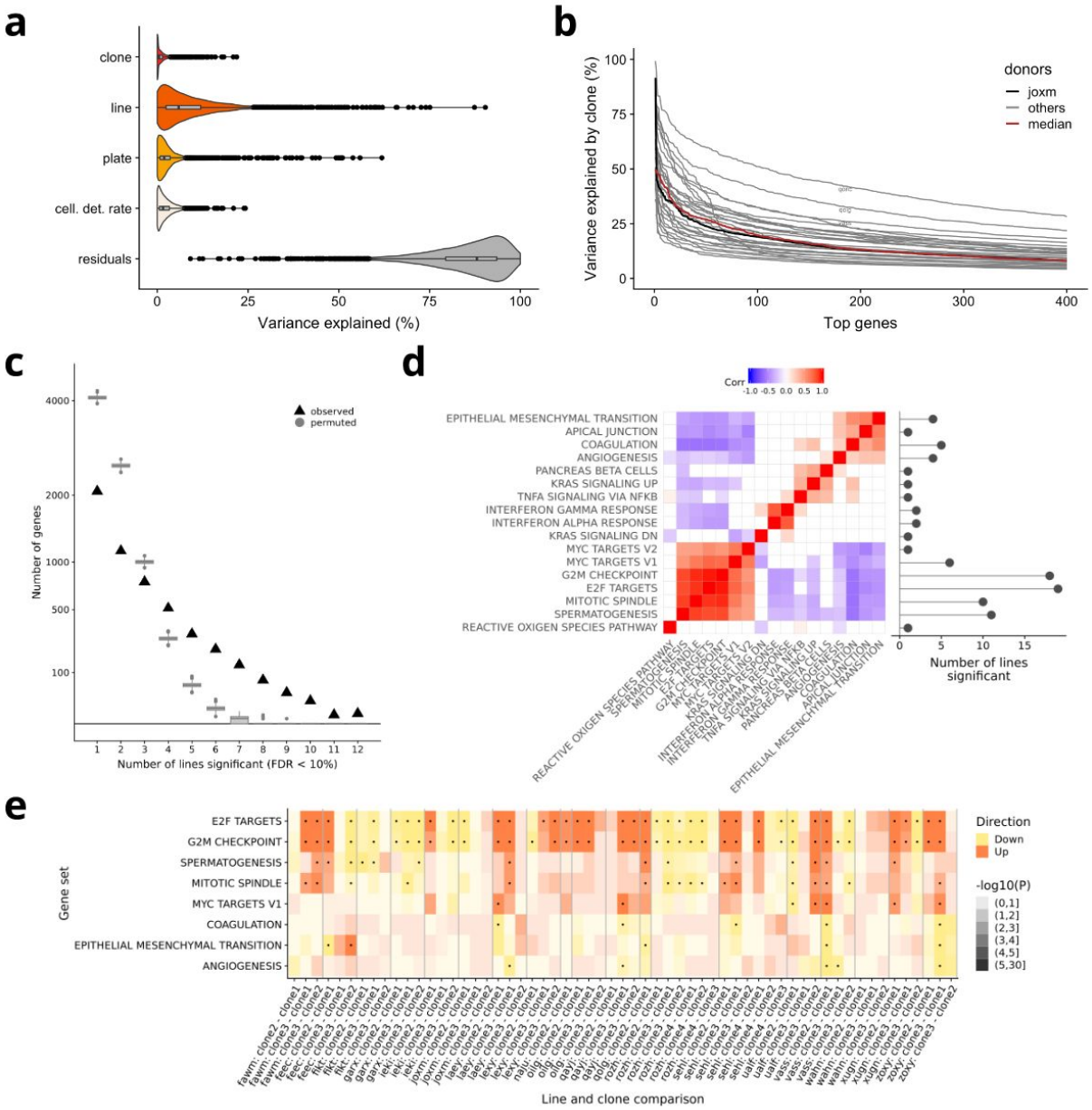


Figure 4 | Signatures of transcriptomic clone-to-clone variation across 31 lines. (a) Violin and box plots show the percentage of variance explained by clone, line, experimental plate and cellular detection rate for 4,998 highly variable genes, estimated using a linear mixed model (**Methods**). **(b)**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

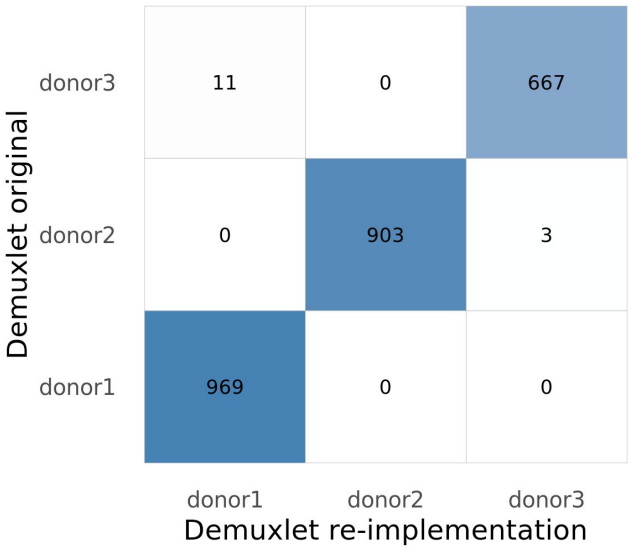


Figure S1: Comparison of donor assignment results from the original Demuxlet software and our implementation. The confusion matrix of cells assigned to three donors by two methods, which are highly concordant. Note, those unmatched cells are all identified as doublets by Demuxlet. The data is generated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

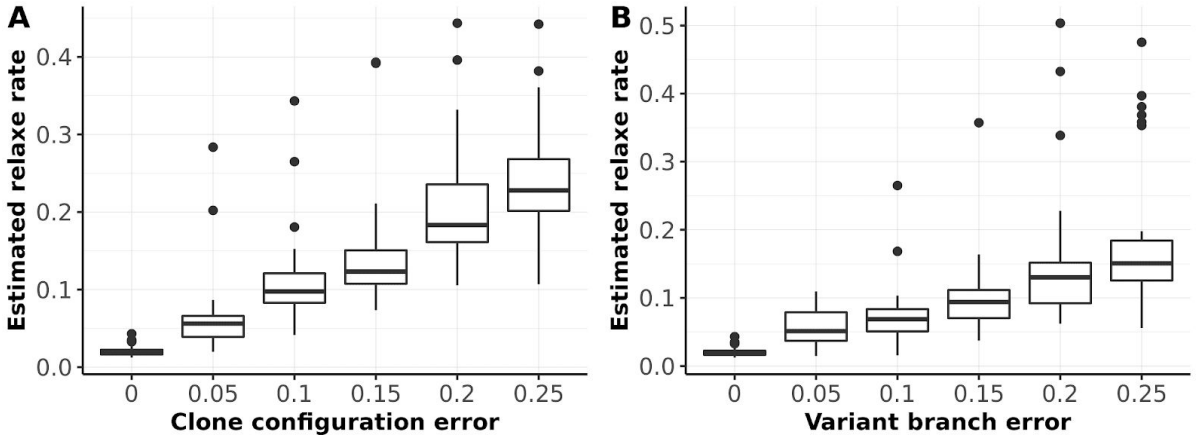


Figure S2: Evaluation of the inferred relax (error) rate using simulations. (A) The estimated relax rate as a function of the simulated error rates. Errors are simulated by uniformly swapping the mutation states in the guide clonal configuration matrix, except the base clone which has no mutations. (B)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

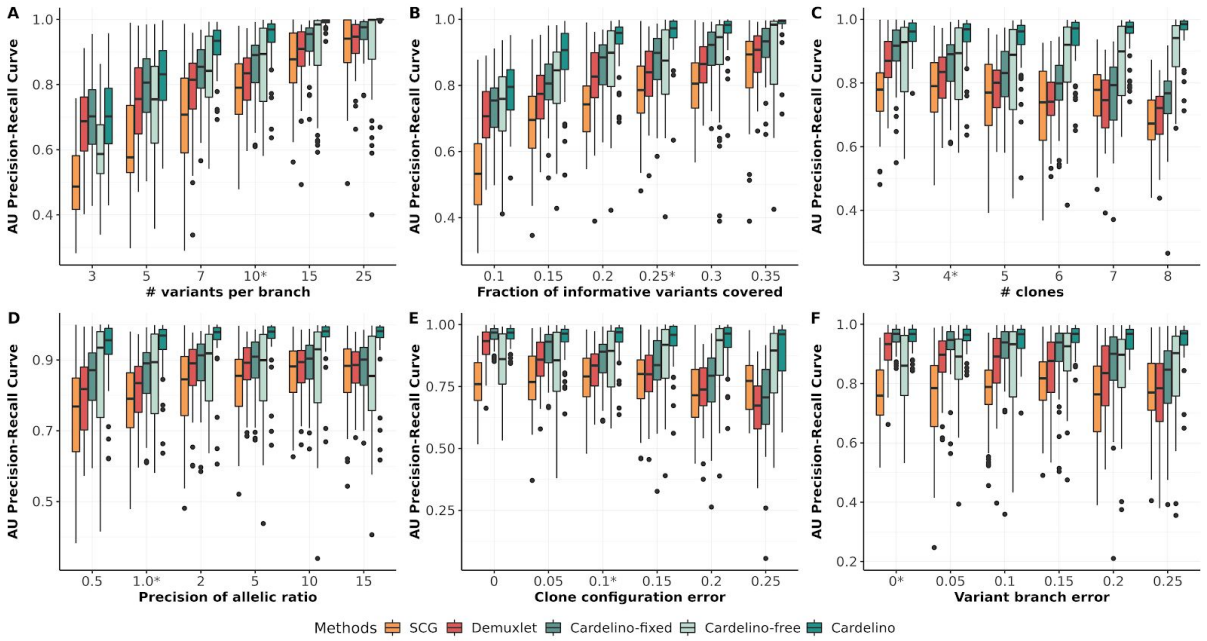


Figure S3. Assessment of cell assignment to clones across a variety of simulation settings, considering SingleCellGenotyper (SCG), Demuxlet (our implementation to avoid the requirement of .bam format), cardelino and its two versions: cardelino-free without any informative clone configuration prior and 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.

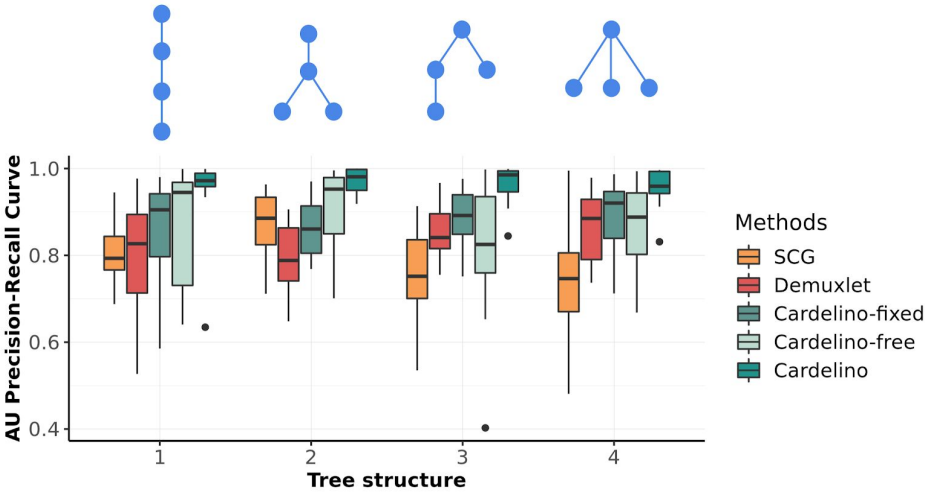
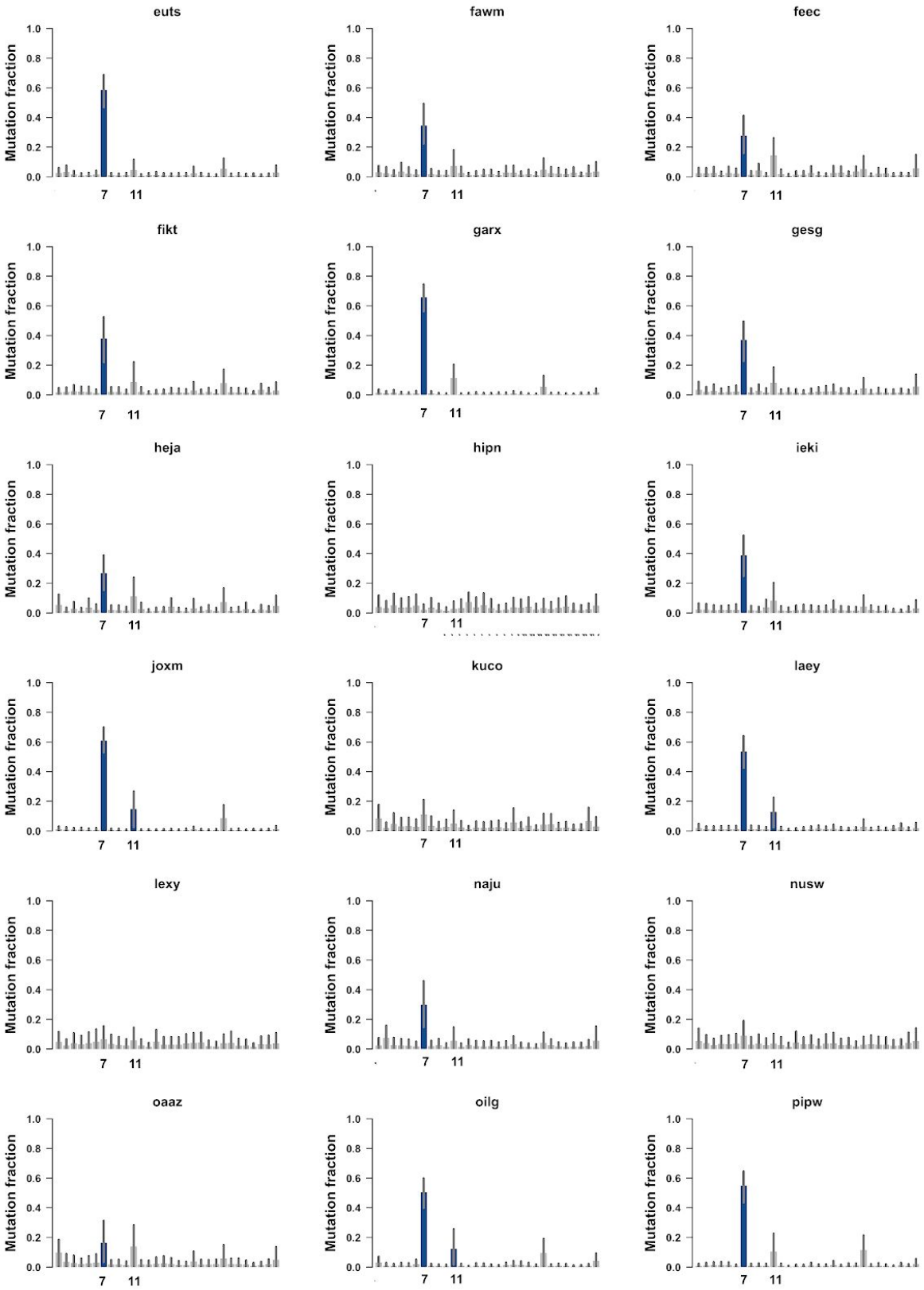


Figure S4. The effects of the tree topology on the cell assignment accuracy. In the simulations in Fig. 1 and Supp Fig. S2, there are 20 repeats for each parameter, where one of the tree topology candidates are randomly selected in each repeat. For the four-clone configuration, there are four different 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

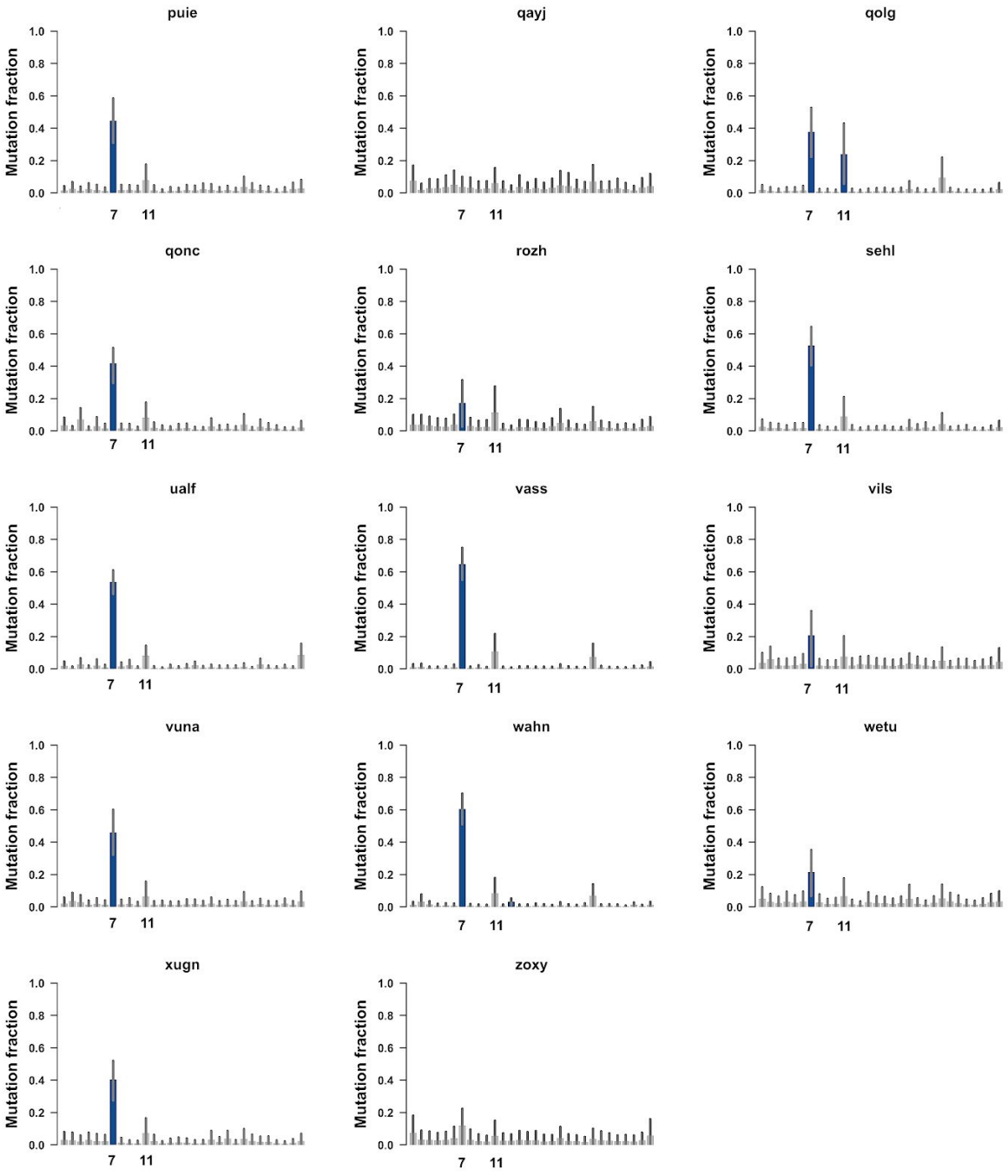


Figure S5. Estimated mutational signature exposures based upon the tri-nucleotide context of somatic SNVs called from whole-exome sequencing (WES) data for 32 HipSci human fibroblast lines. The x-axis shows 30 COSMIC mutational signatures, in order, and the y-axis shows estimated exposures (mutation 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.

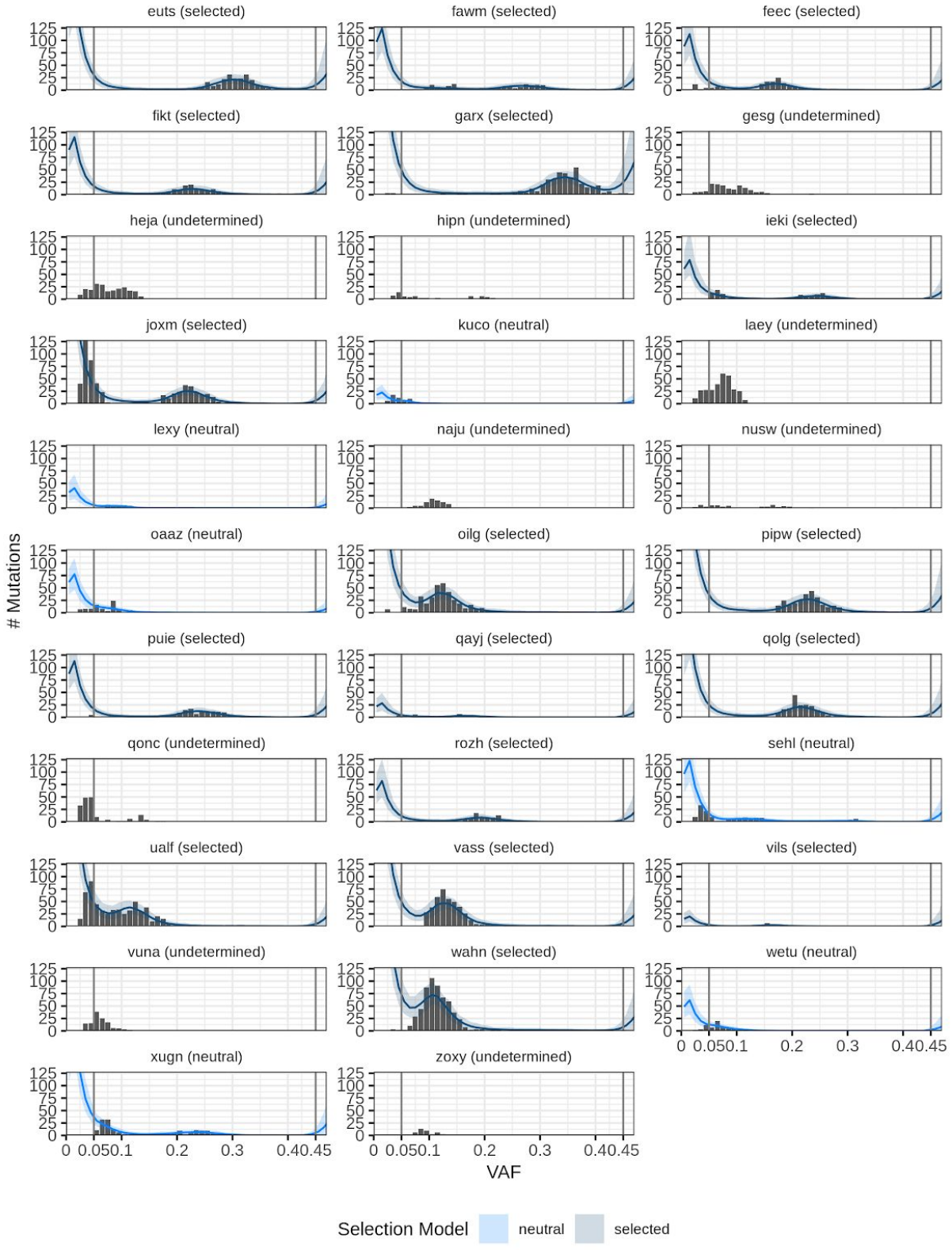


Figure S6. Variant allele frequency (VAF) distributions for somatic variants called from whole exome sequencing data for the 32 fibroblast lines. The grey lines indicate the cut-offs on the allele frequency distribution (**Methods**). The blue lines indicate the model (neutral/selected) inferred by SubClonalSelection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

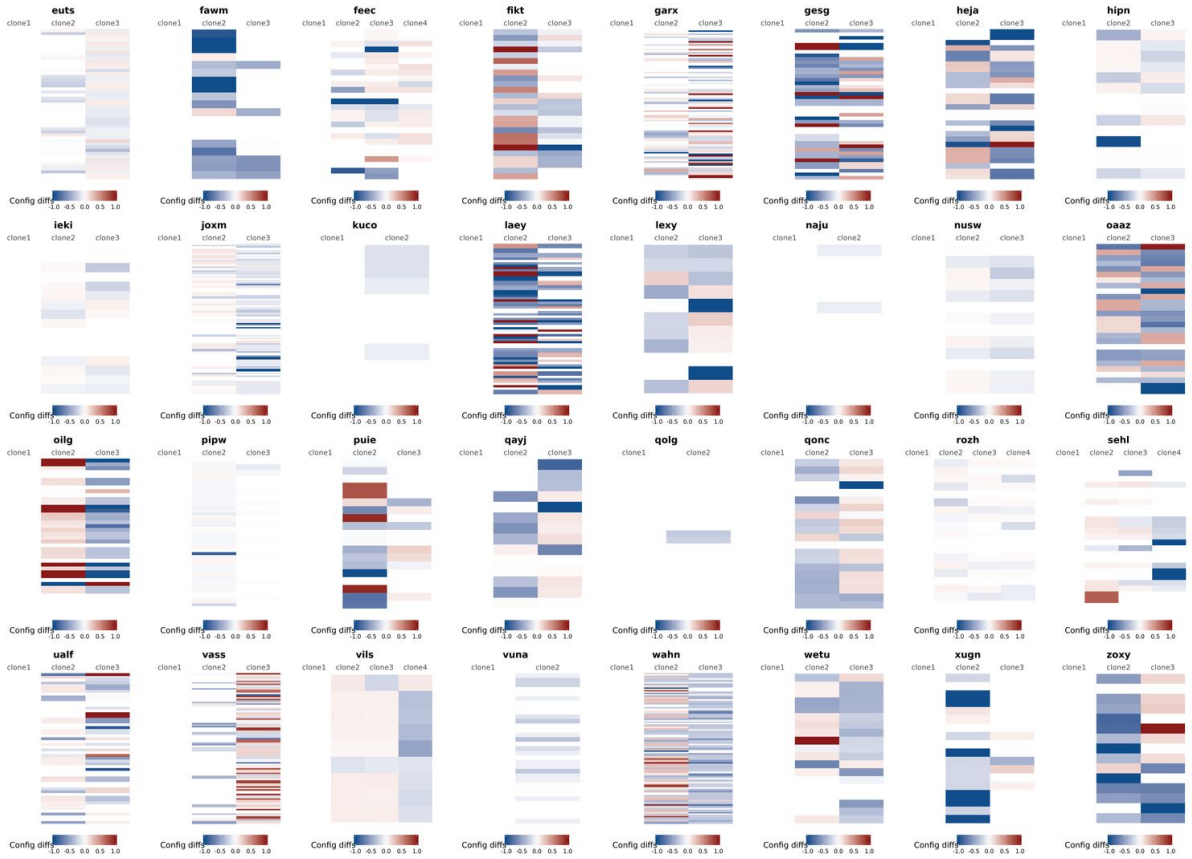


Figure S10. Differences in configuration matrices (rows represent single-nucleotide variants and columns represent clones) between Canopy trees and updated trees from cardelino (average configuration matrix over 4,750 posterior samples from the cardelino model minus the configuration matrix for the tree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

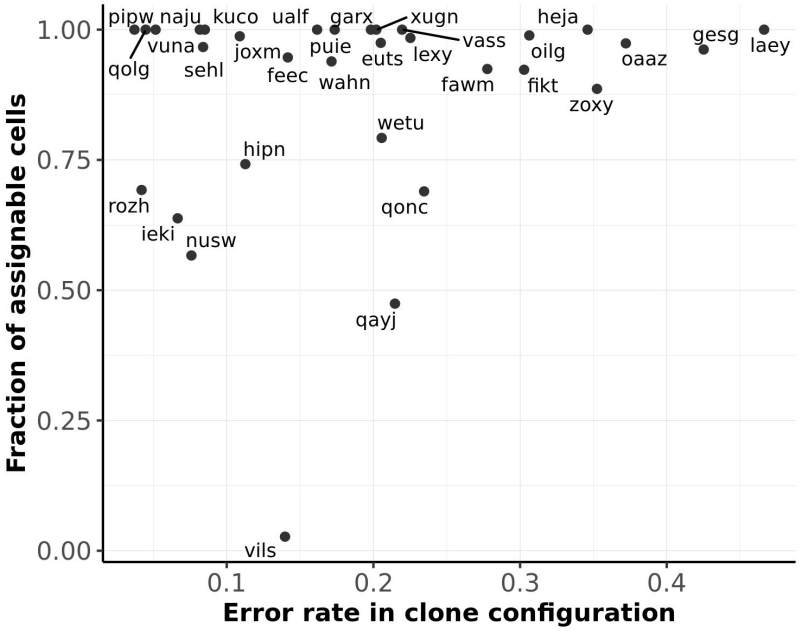


Figure S11. Estimated error rate in the clonal tree configuration derived from bulk exome-seq data (based on cardelino) for each of 32 lines versus fraction of confidently assigned cells. Even though some lines have high error rate in the input clonal tree configuration, cardelino can still assign a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

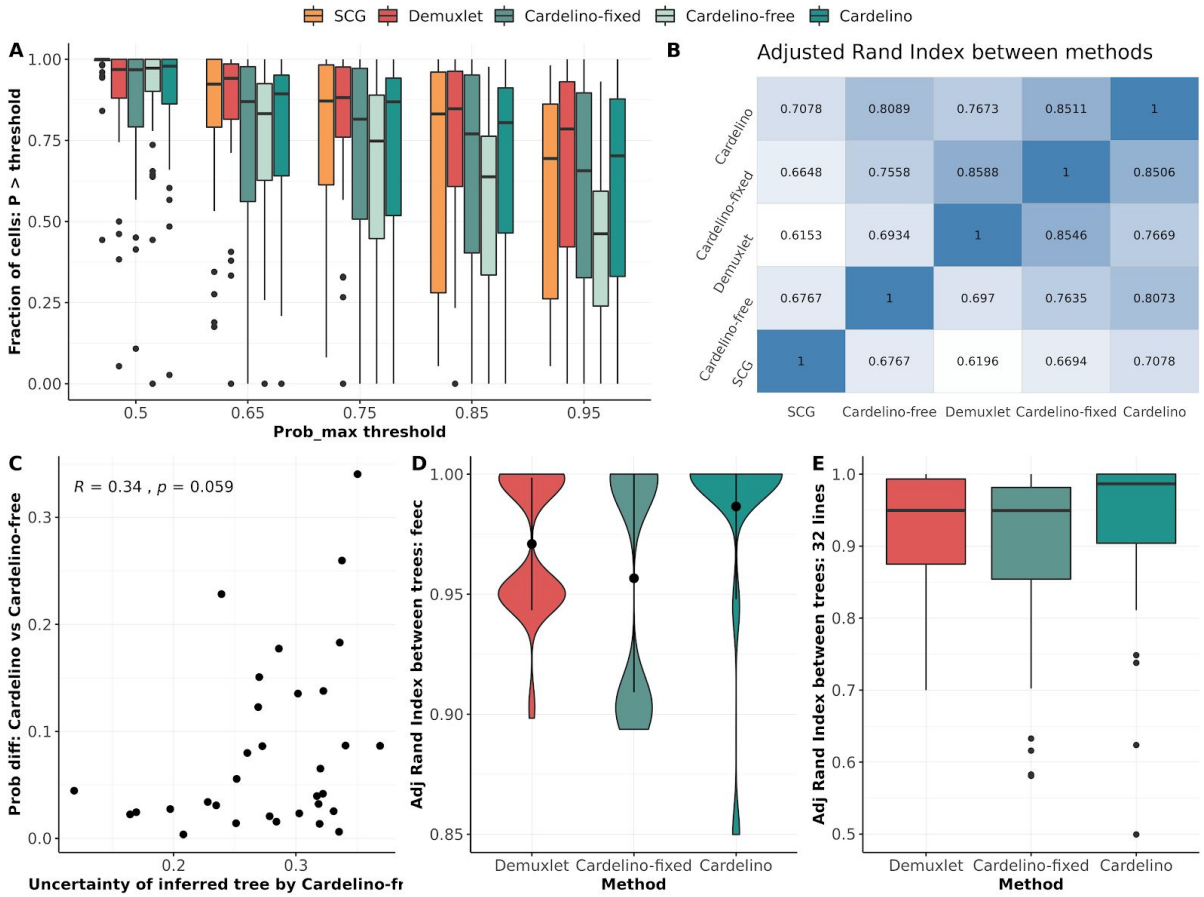


Figure S12. Comparison of cell assignment between five methods across 32 lines. (**A**) The fraction of assignable cells (i.e., highest P > thresholds) when varying the thresholds from 0.5 to 0.95. Shown are box plots depicting median and the first and third quantiles of the 32 lines. (**B**) The adjusted 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.

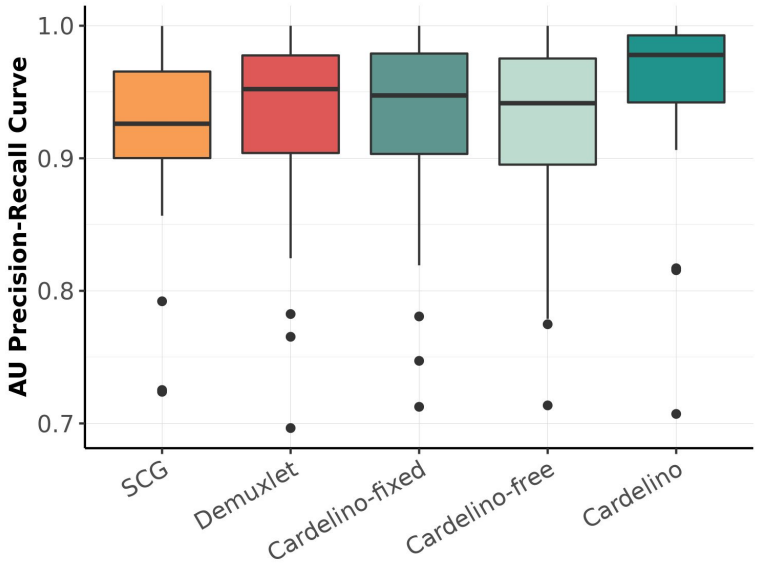


Figure S13. Assessment of cell assignment to clones across a variety of simulation settings, considering SingleCellGenotyper (SCG), Demuxlet (our implementation to avoid the requirement of .bam format), cardelino and its two versions: cardelino-free without any informative clone configuration prior and 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.

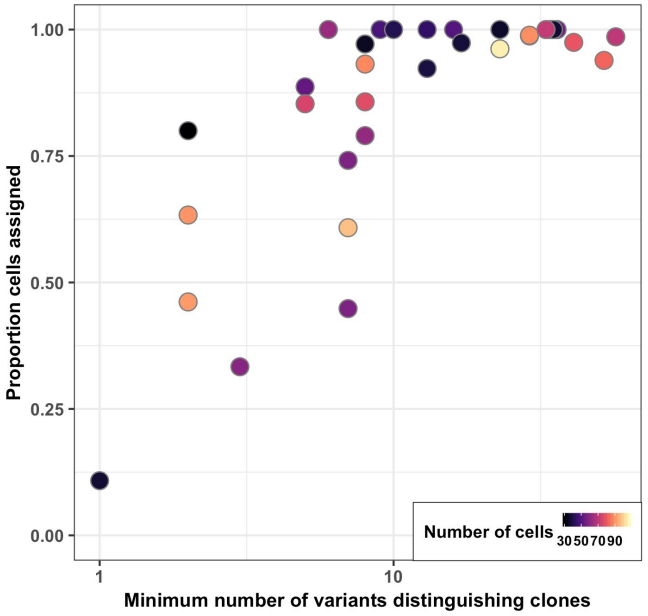


Figure S14. Scatter plot of the fraction of cells assigned in each cell line using cardelino (at posterior probability > 0.5) as a function of the minimum number of clone-specific variants for the corresponding line (minimum Hamming distance between clones for a given donor), for 32 fibroblast lines.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

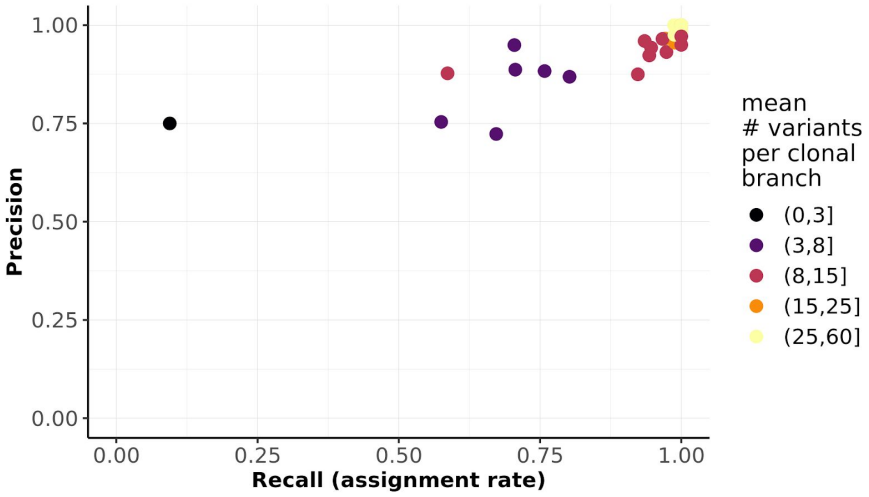


Figure S15. Scatter plot of recall (assignment rate) versus precision (assignment accuracy) when assigning cells using cardelino (at posterior probability > 0.5). Shown are data from for 32 simulated lines, using parameters that match the observed data characteristics in the set of 32 real fibroblast 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.

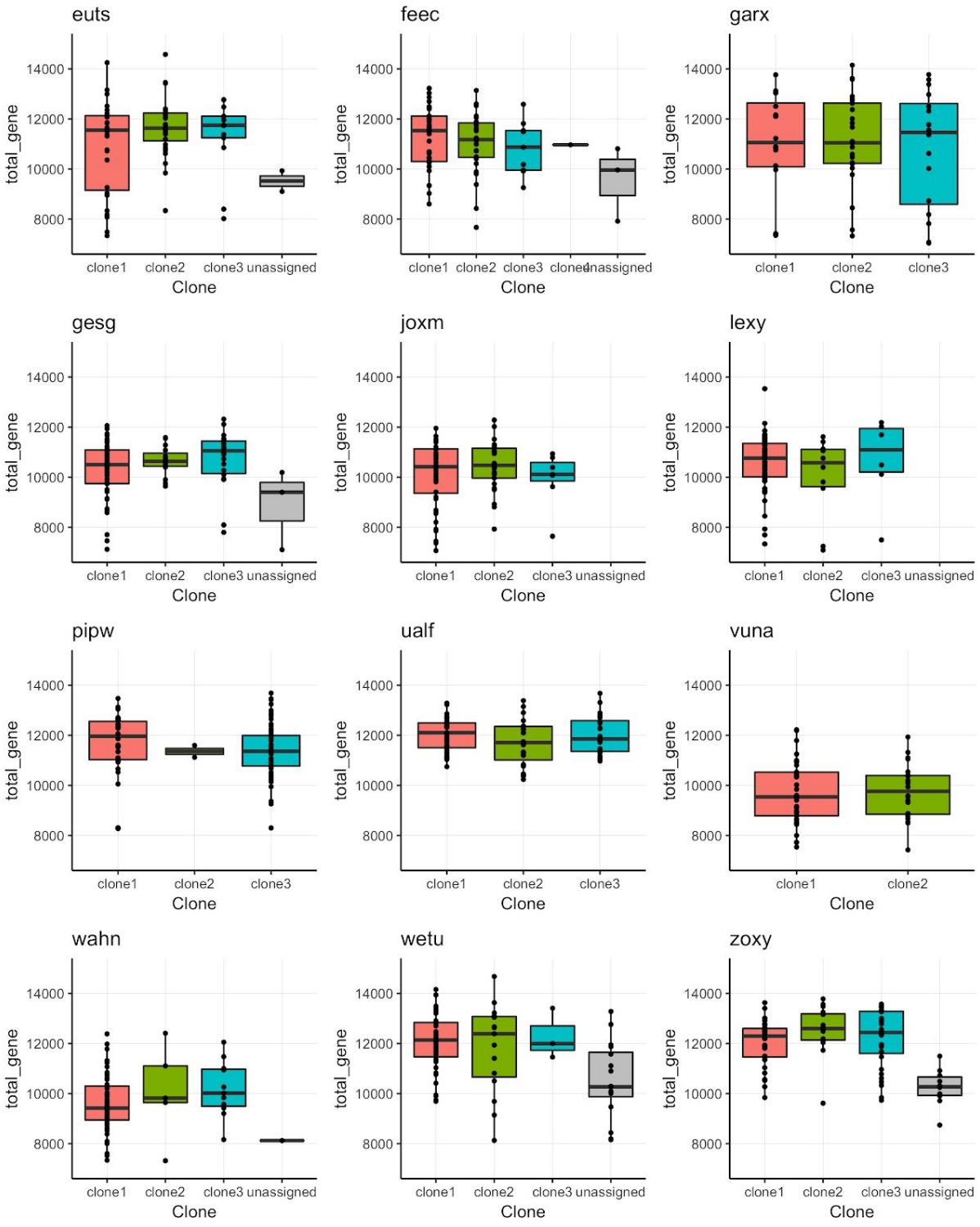


Figure S16. Boxplots of the total number of expressed genes in each cell, grouped by the clone assigned by cardelino. Twelve lines with more than 60 assignable cells are presented. Globally, clone assignment is not linked to the total number of expressed genes in a given cell.

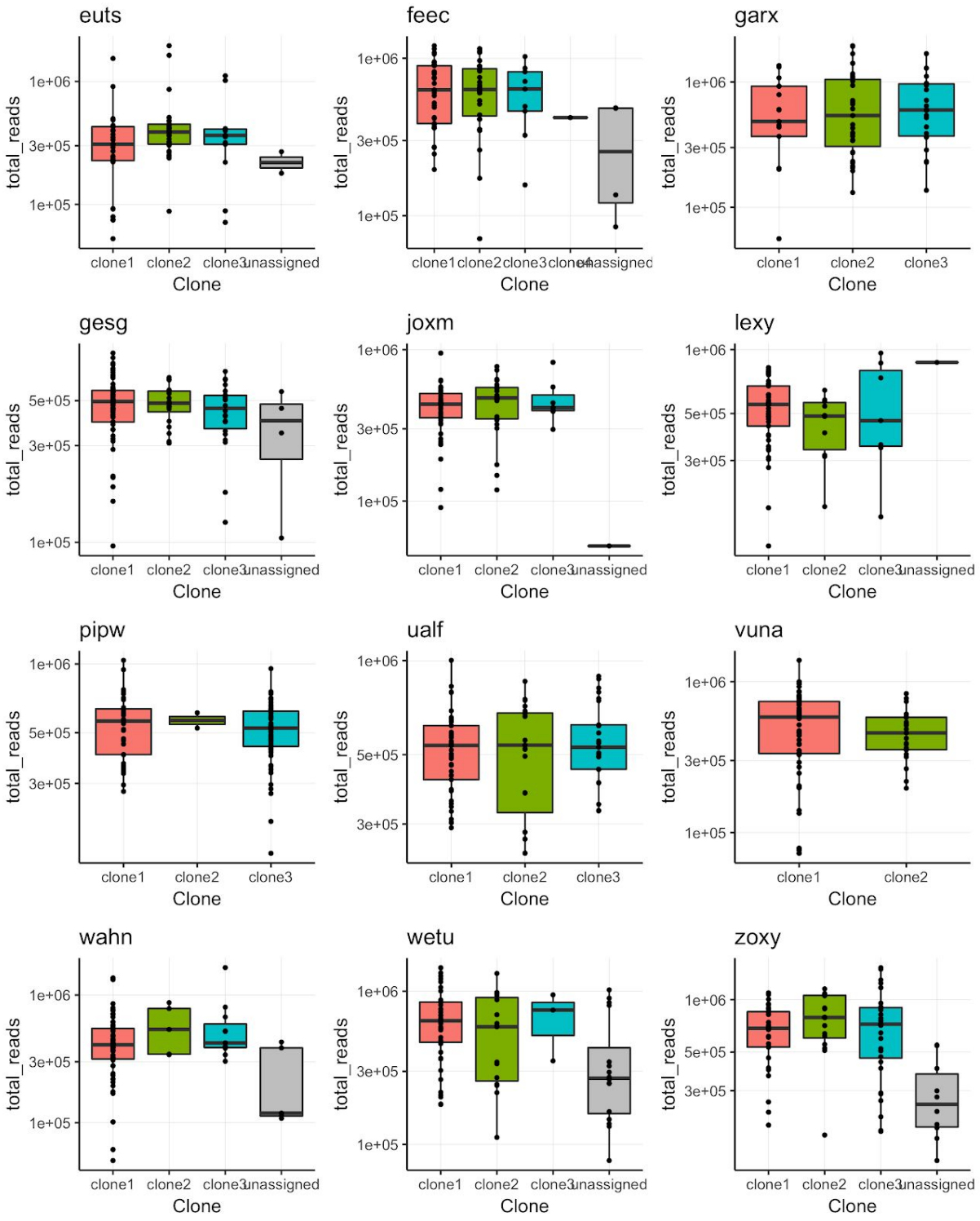


Figure S17. Boxplots of the total number of sequenced read counts from endogenous genes in each cell, grouped by the clone assigned by cardelino. Twelve lines with more than 60 assignable cells are presented. Globally, clone assignment is not linked to the total number of read counts in a given cell.

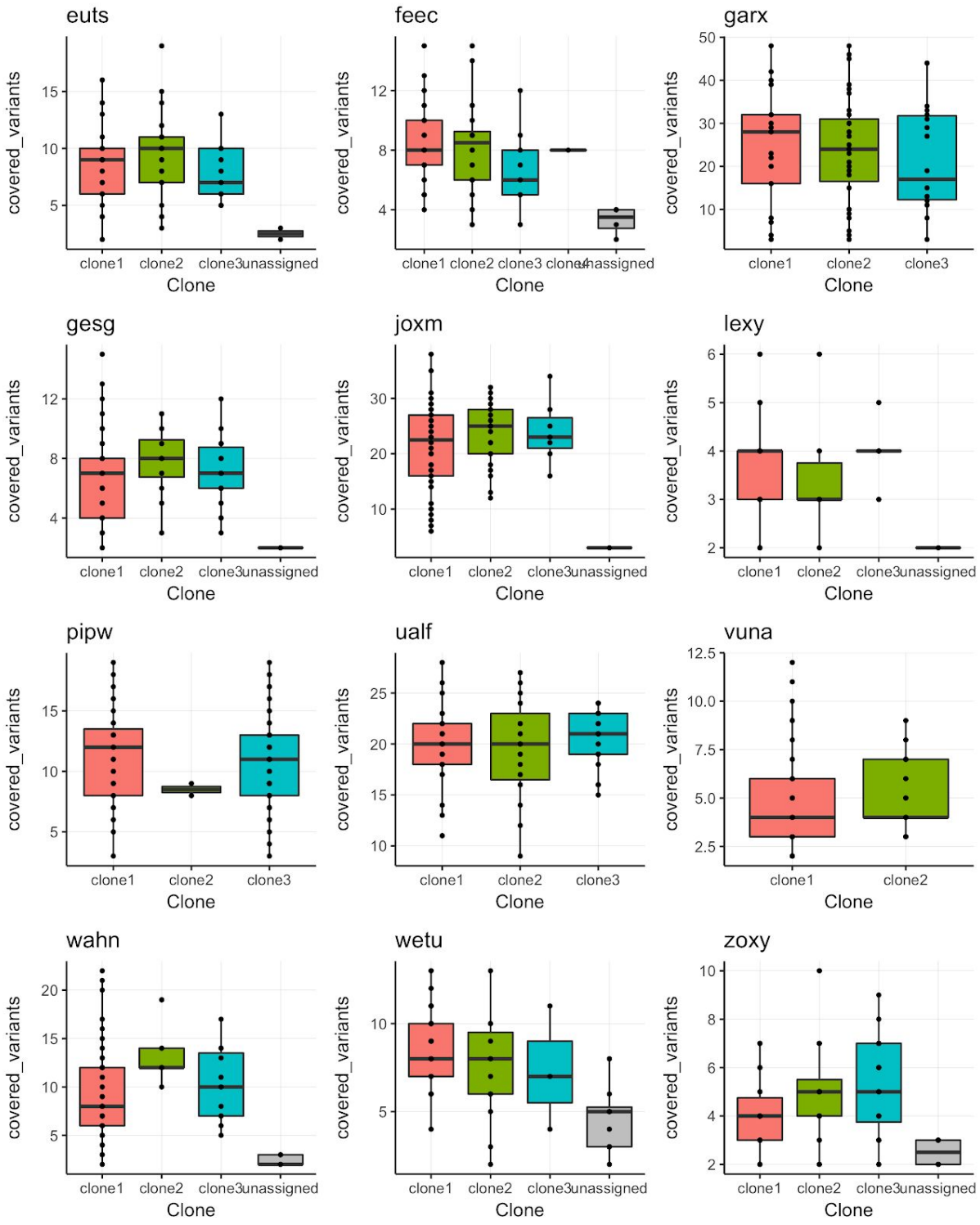


Figure S18. Boxplots of the number of variants for clone identification with read coverage in each cell, grouped by the clone assigned by cardelino. Twelve lines with more than 60 assignable cells are presented. Globally, clone assignment is not linked to the number expressed variant loci in a given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

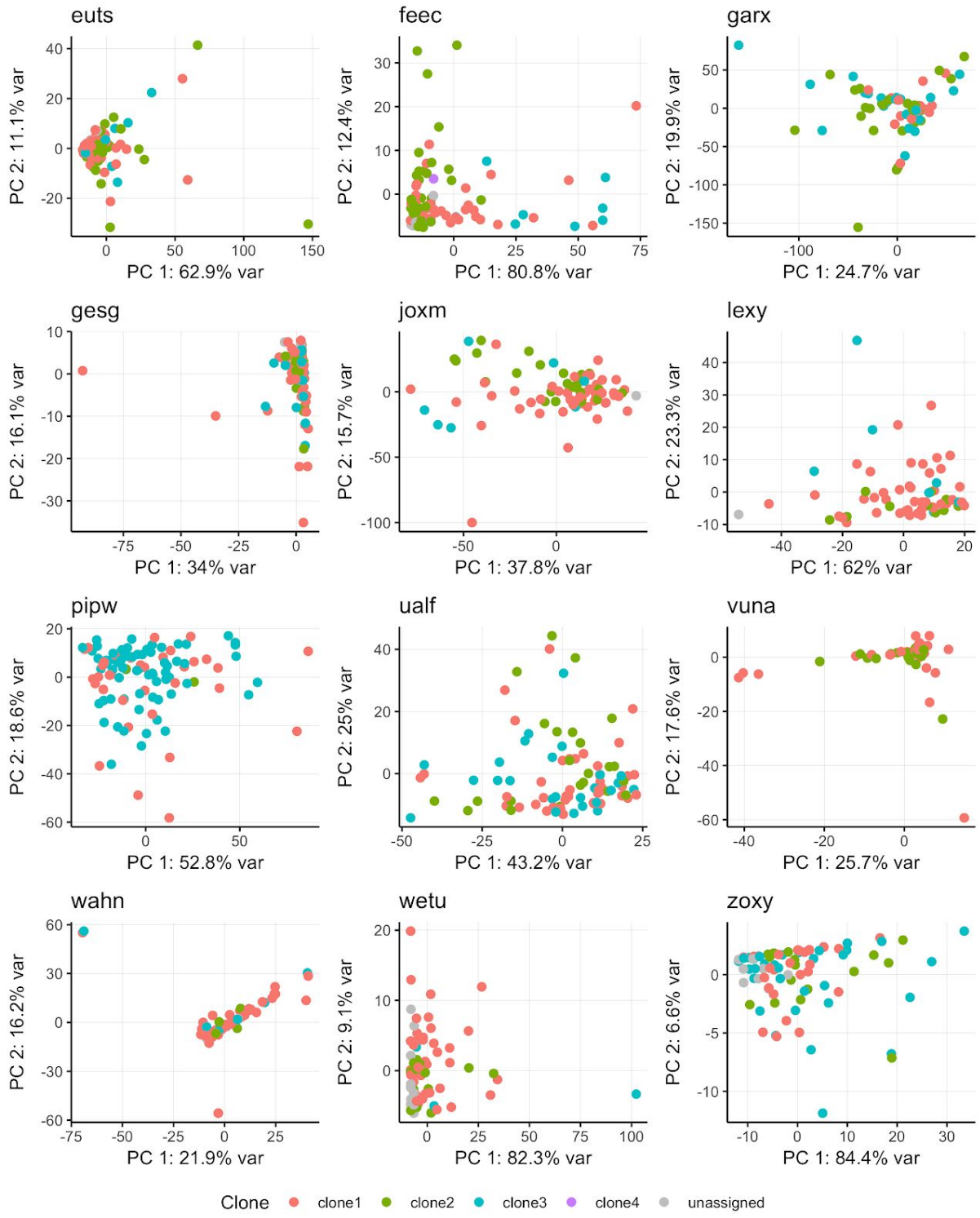


Figure S19. Scatter plot of the first two principal components calculated on the read coverage of the set of somatic variant sites used for clone assignment. Shown are data from twelve lines with at least 60 assignable cells. The first two PCs do not segregate cells from different clones, suggesting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

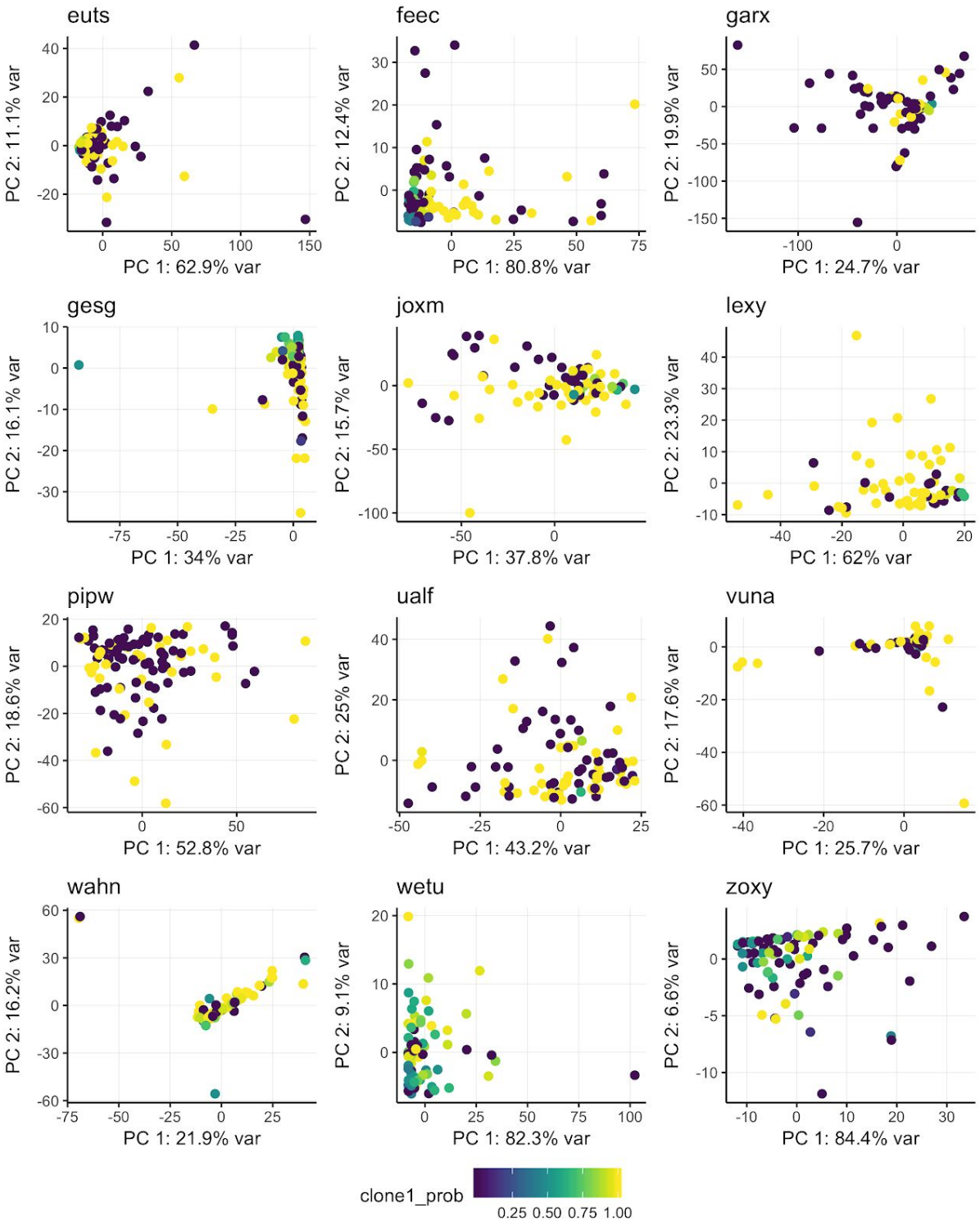


Figure S20. Scatter plot of the first two principal components calculated on the read coverage of the set of somatic variant sites used for clone assignment. Cells are colored by the assignment probability of clone 1 (*i.e.* the "base clone" which by definition contains no unique somatic variants). Shown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

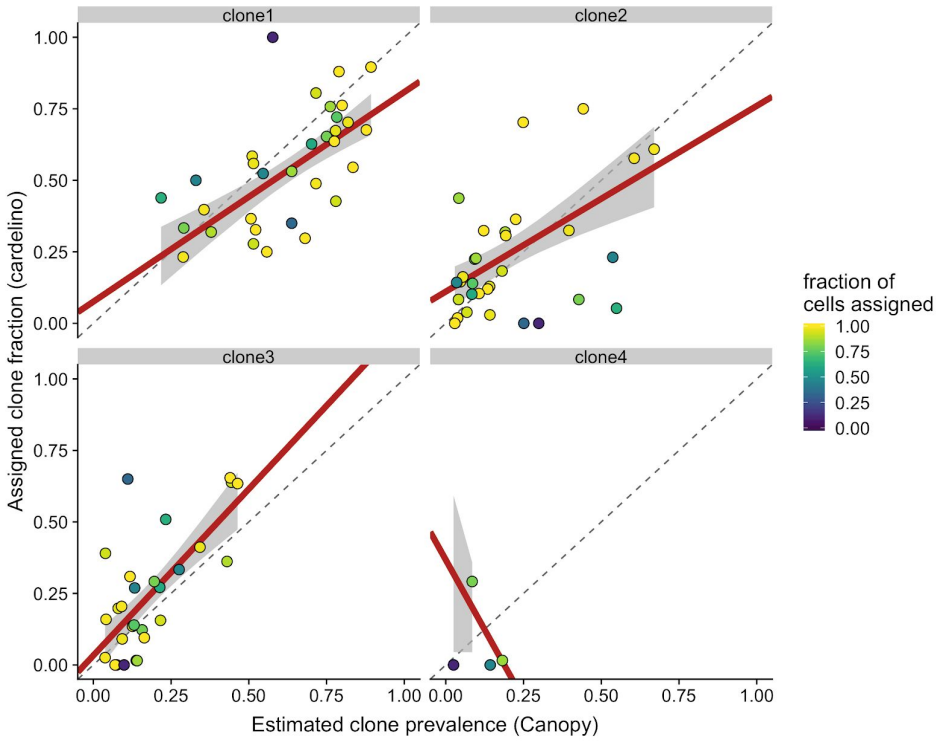


Figure S21. Clone prevalence estimates from WES data (x-axis; using Canopy) *versus* the fraction of single-cell transcriptomes assigned to the clone (y-axis; using *cardelino*), for each clone across lines. Points are coloured by the overall fraction of single-cell transcriptomes assigned for a given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

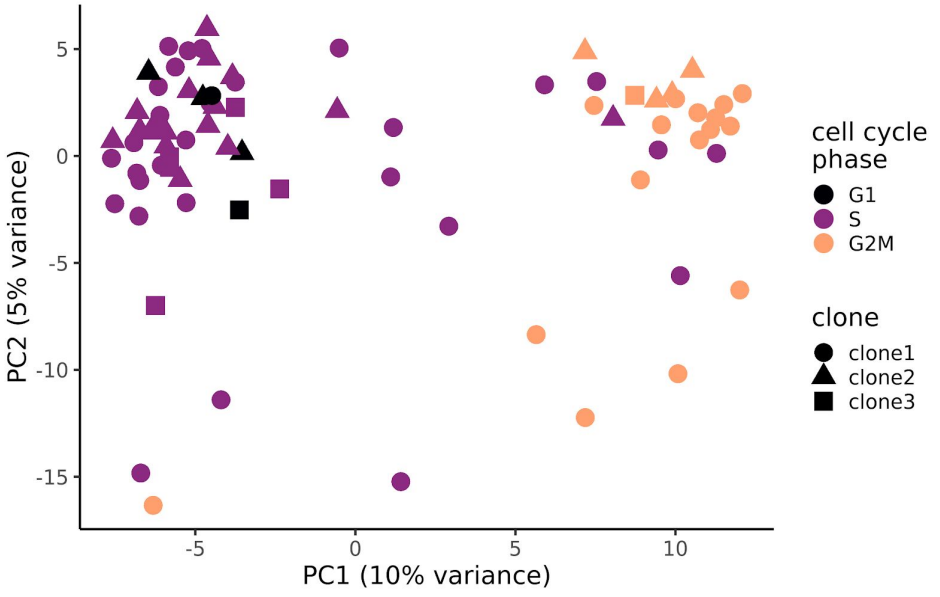


Figure S22. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) for clone-assigned cells for the example line *joxm*. Cells are coloured by the cell cycle phase inferred by the *cyclone* method implemented in the scran package, and shape denotes the assigned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

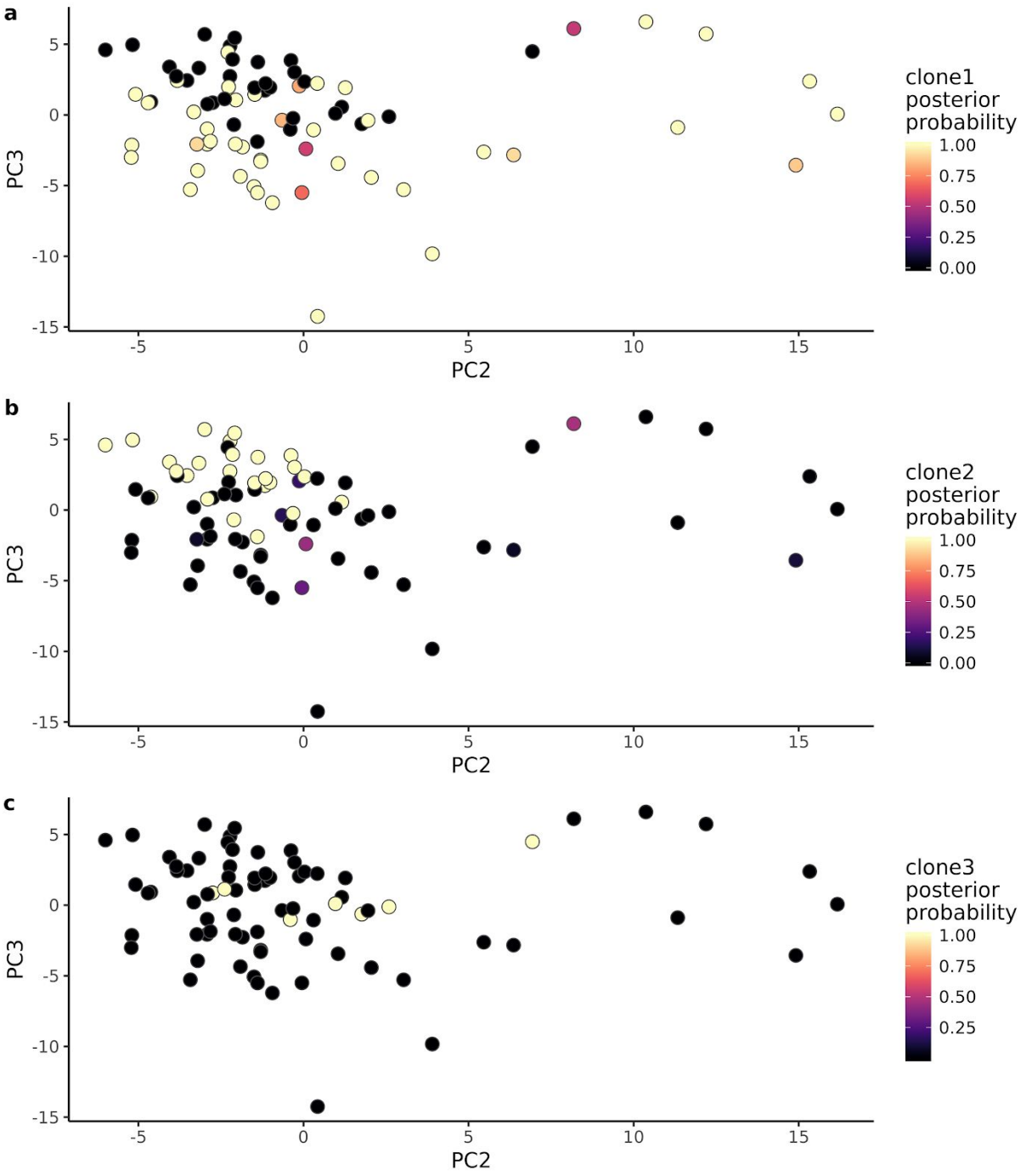


Figure S23. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) for clone-assigned cells for the donor *joxm*, plotting principal component 3 against principal component 2. Cells are coloured by the posterior probability from cardelino that the cell belongs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

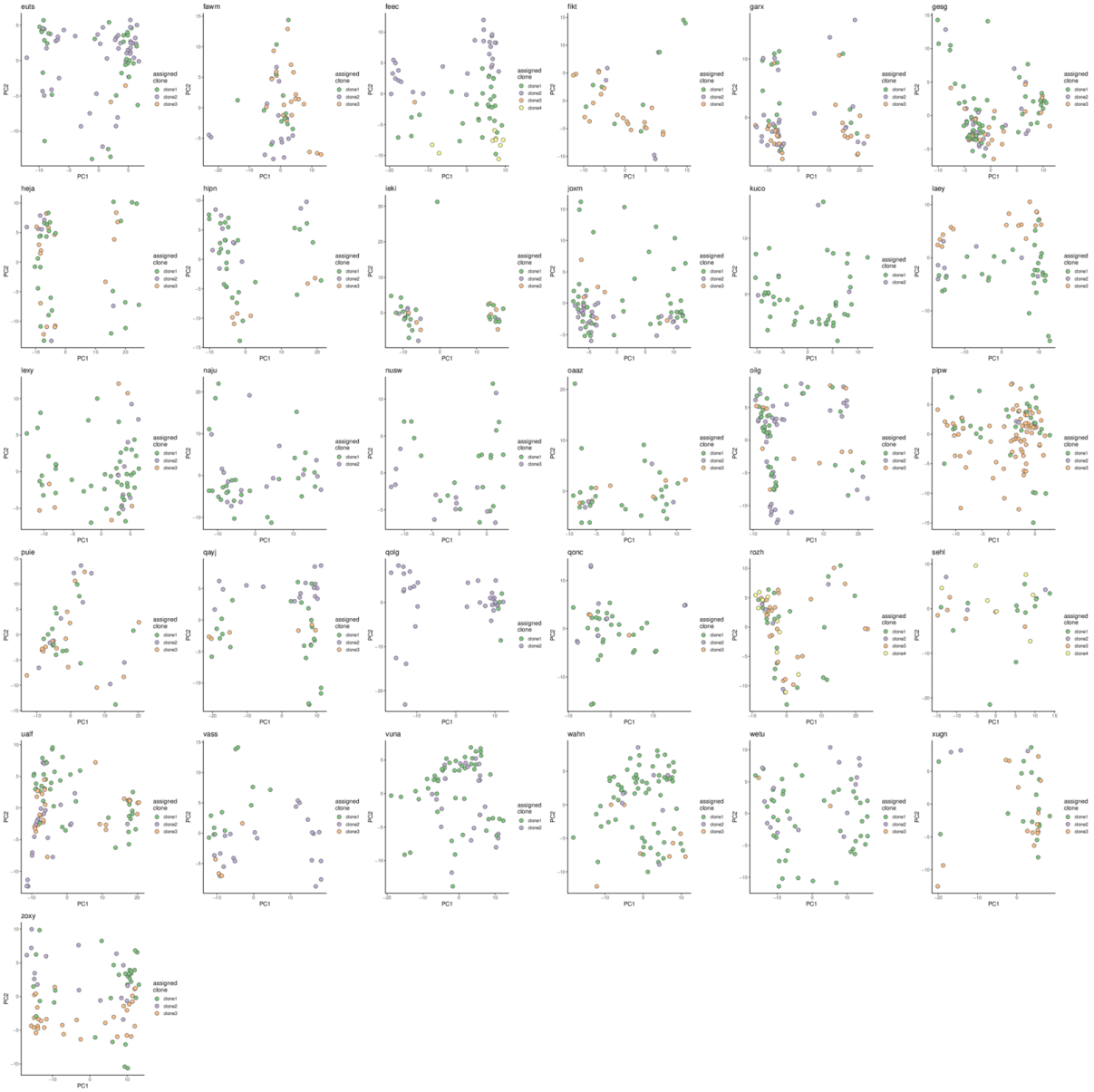


Figure S24. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the assigned clone from cardelino.

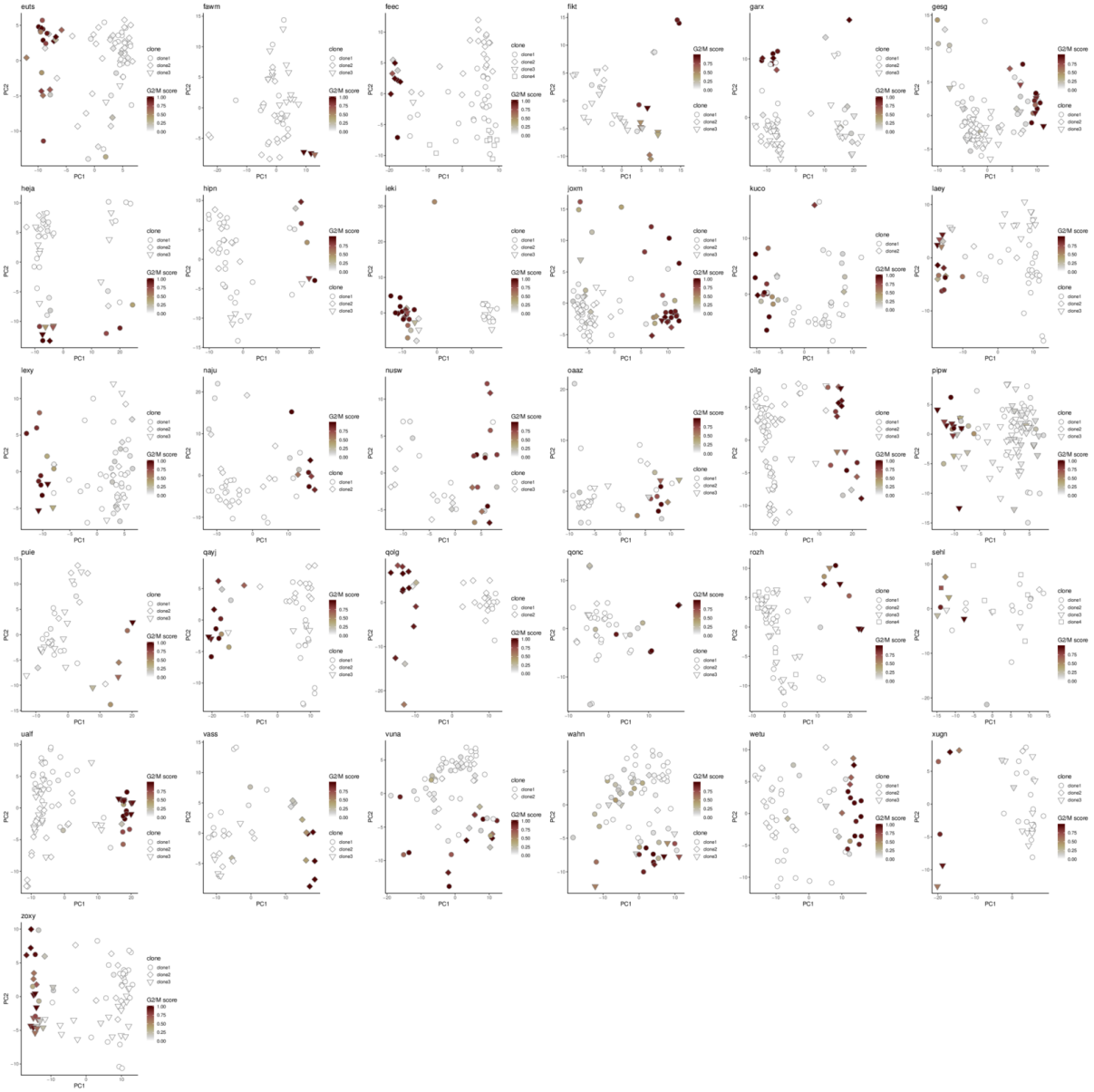


Figure S25. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the G2M cell cycle phase score calculated with the cyclone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

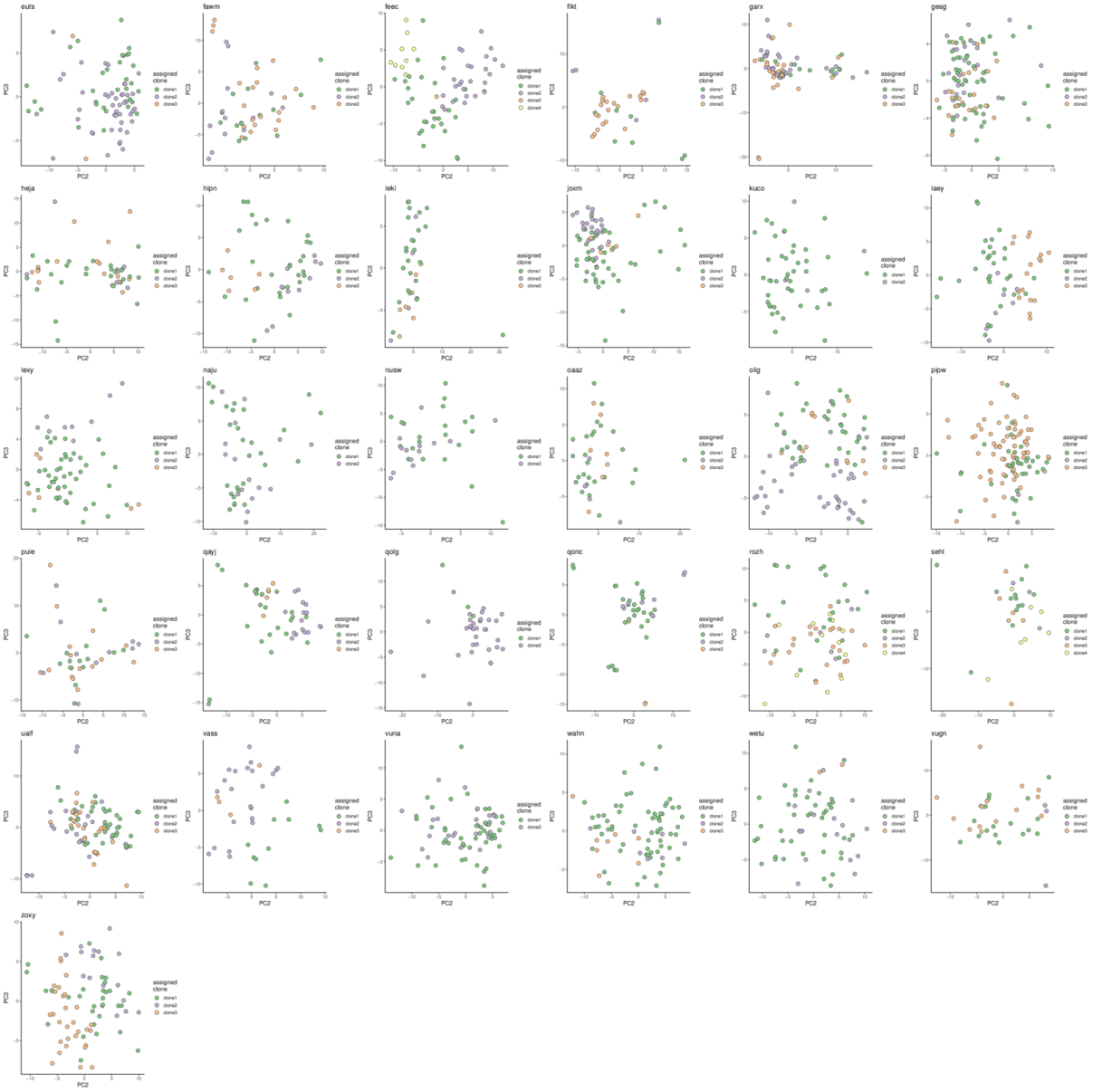


Figure S26. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC3 plotted against PC2 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the assigned clone from cardelino.

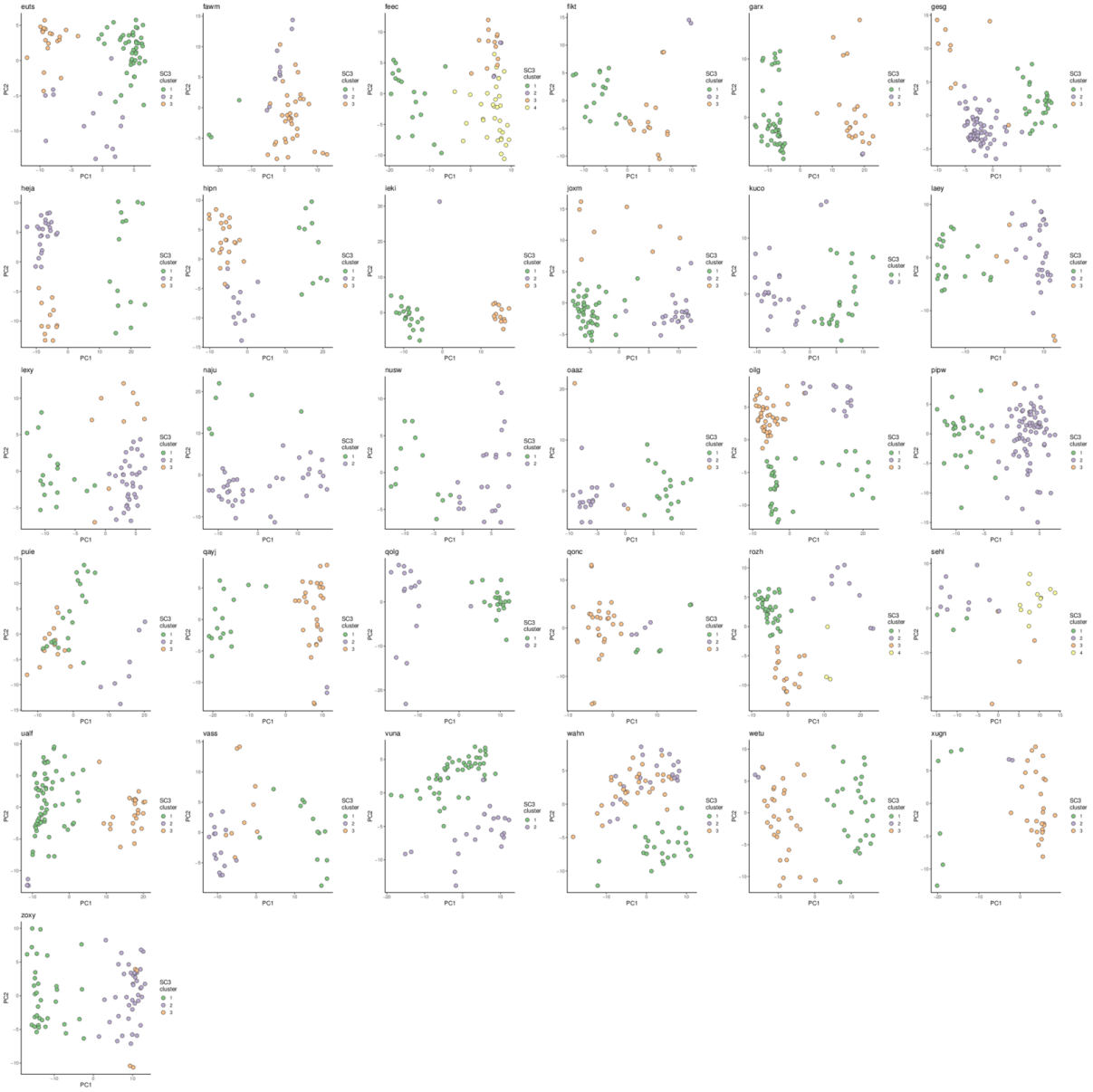


Figure S27. Principal component analysis from single-cell gene expression data (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail in the manuscript. Cells are coloured by the clusters identified by SC3 (Kiselev et al, *Nature Methods*,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

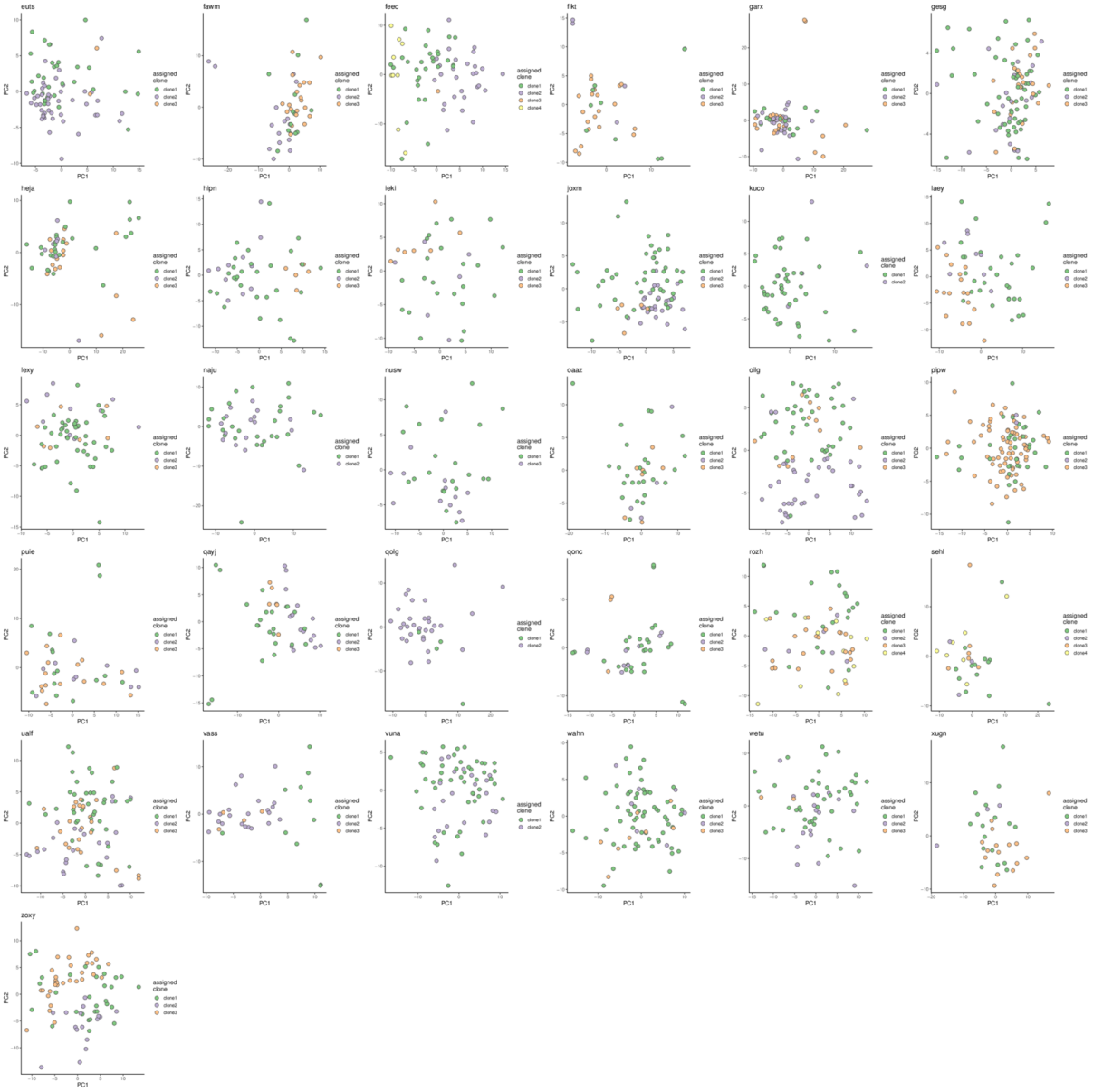


Figure S28. Principal component analysis from single-cell gene expression data after regressing out *cyclone* G1, G2M and S cell cycle scores from the normalised expression values (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail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

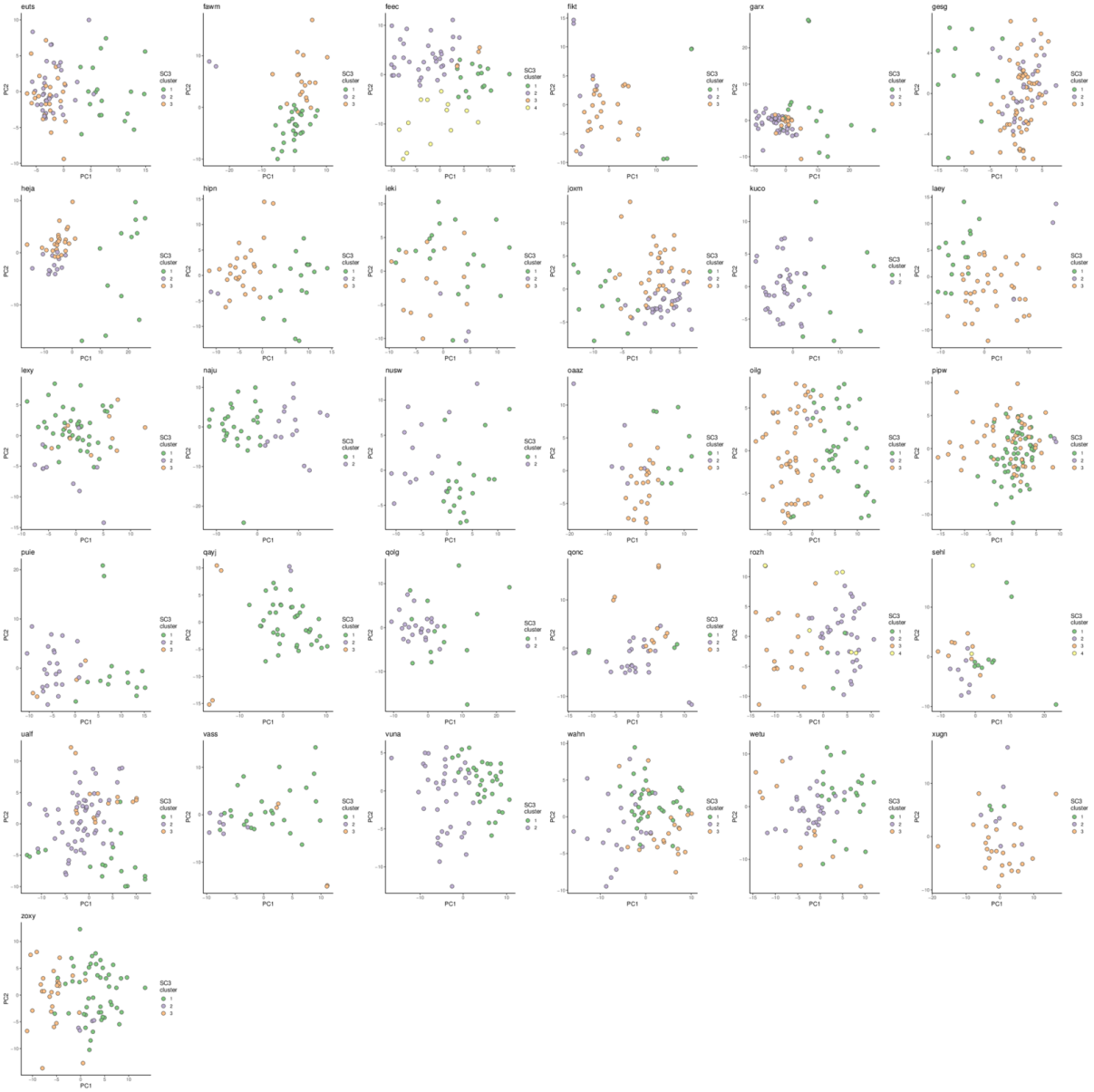


Figure S29. Principal component analysis from single-cell gene expression data after regressing out *cyclone* G1, G2M and S cell cycle scores from the normalised expression values (top 500 most-variable genes) showing PC2 plotted against PC1 for clone-assigned cells for the 31 lines analysed in detail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

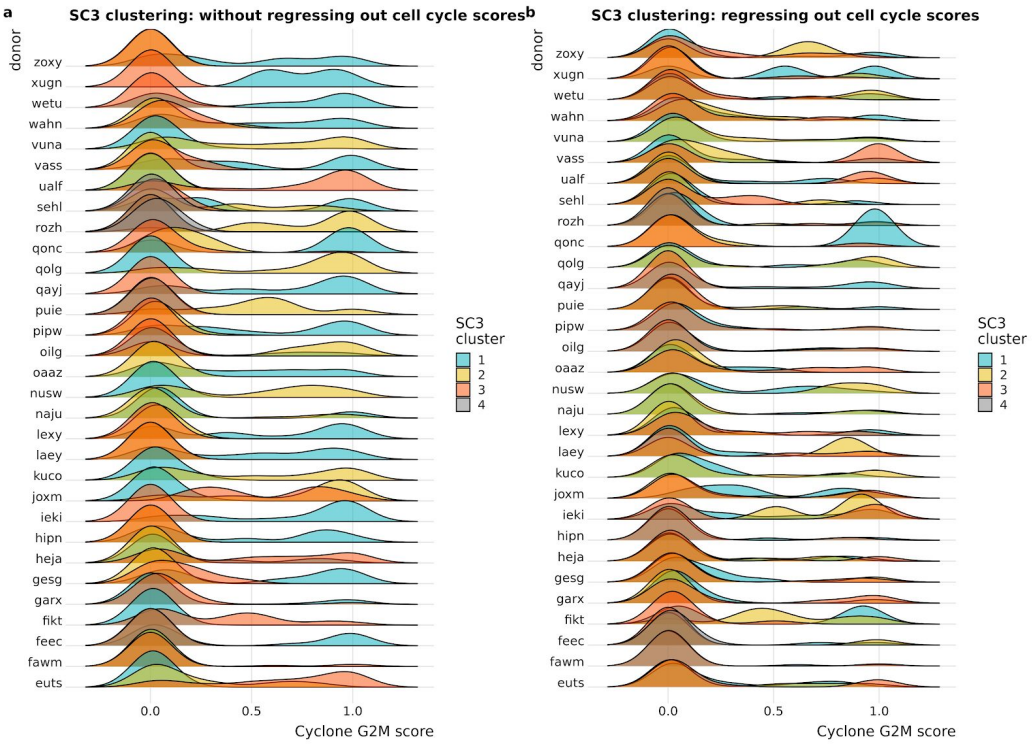


Figure S30. Distributions of *cyclone* G2M scores for each cell line (donor) stratified (coloured) by the clusters identified by SC3 when **(a)** applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and **(b)** applying SC3 to gene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

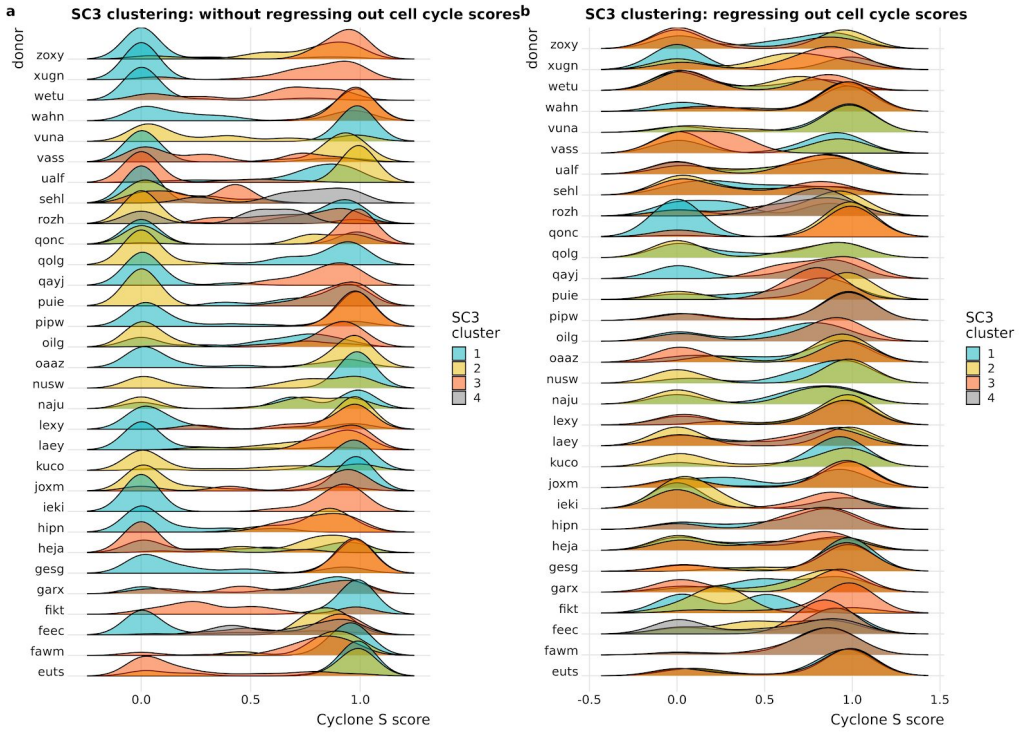


Figure S31. Distributions of *cyclone* S scores for each cell line (donor) stratified (coloured) by the clusters identified by SC3 when **(a)** applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and **(b)** applying SC3 to gene expression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

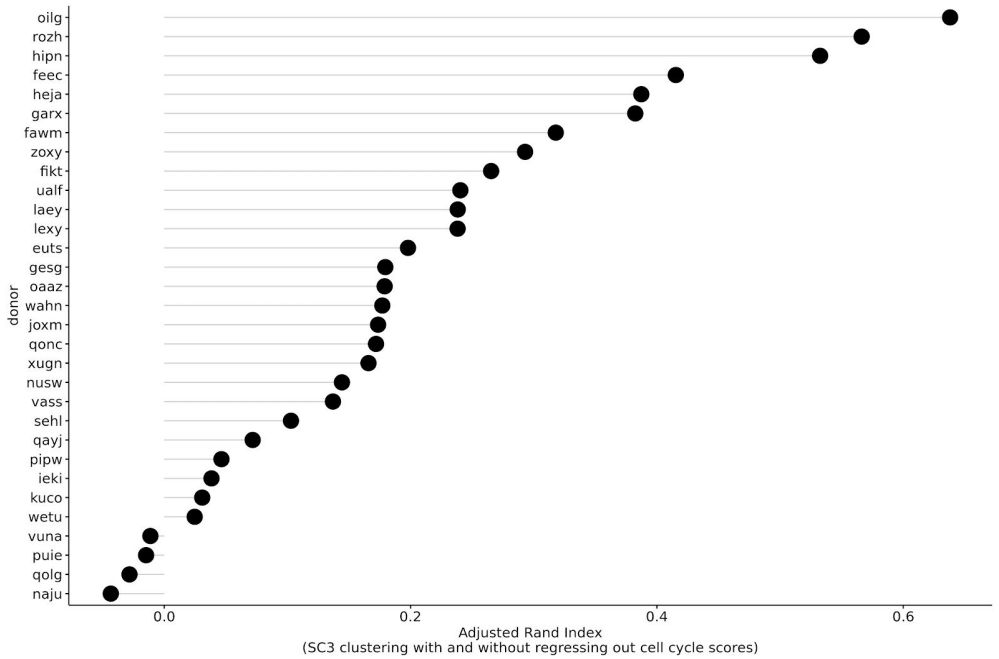


Figure S32. Adjusted Rand Index values comparing the clusters identified by SC3 when applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and when applying SC3 to gene expression values after regressing out *cyclone* cell cycle scores.

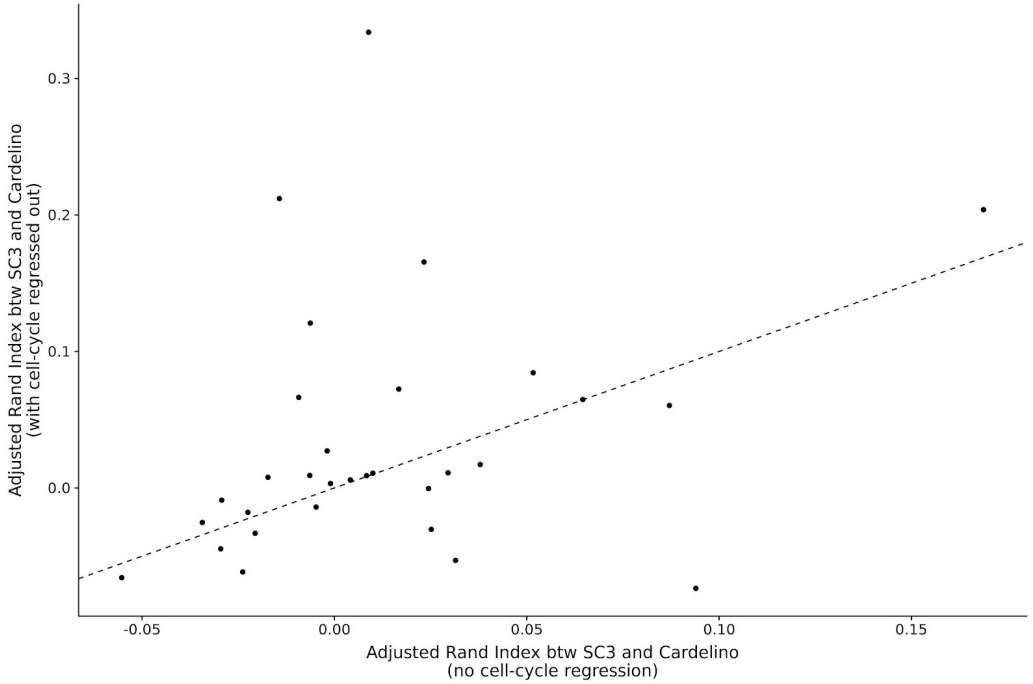


Figure S33. Adjusted Rand Index values comparing the clusters identified by SC3 and the clone assignments from cardelino when applying SC3 to normalised gene expression values, without regressing out *cyclone* G1, G2M and S cell-cycle phase scores, and when applying SC3 to gene expression values after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

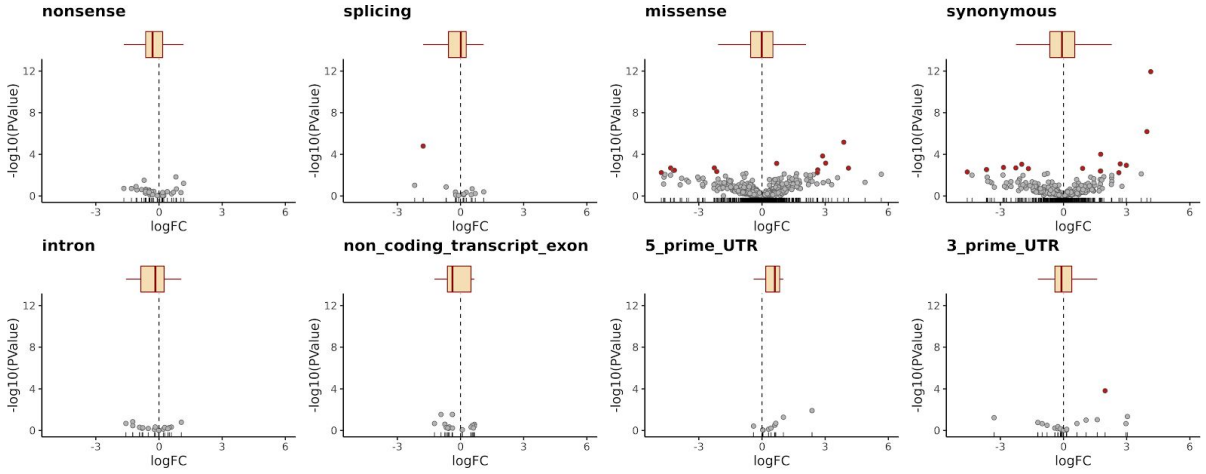


Figure S34. Direct effects of somatic variants on genes overlapping the variant. Volcano plot showing negative log P values versus $\log_2$-fold change from testing differential expression for genes with a somatic mutation between cells with the mutation and cells without the mutation, faceted by VEP 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.

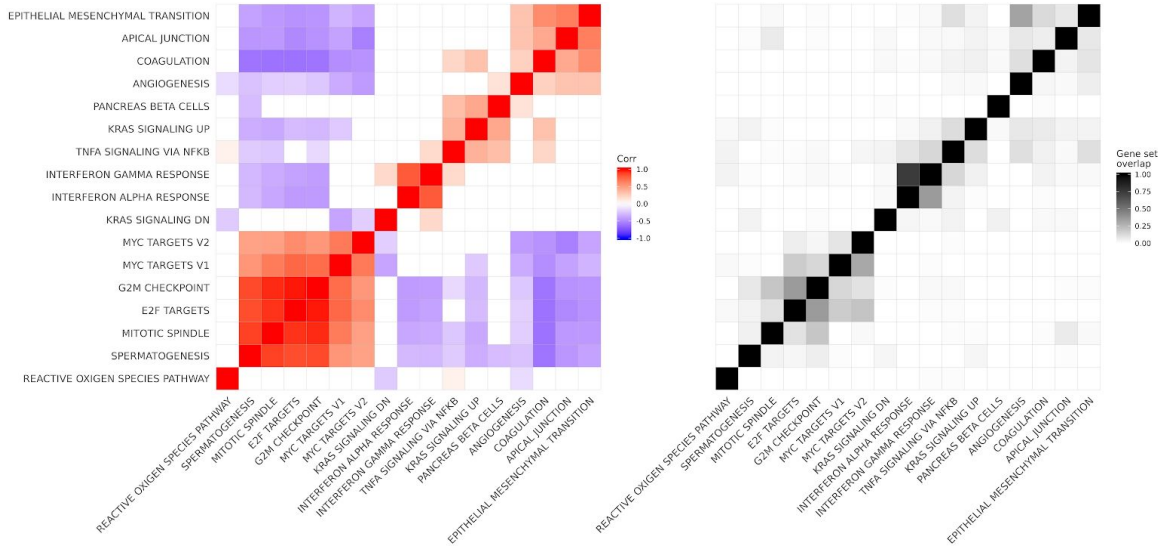


Figure S35. (left) Heatmap showing Spearman correlation between gene set enrichment results for the 16 most frequently enriched MSigDB Hallmark gene sets across 31 lines. Colour indicates the correlation between pairs of gene sets and is only shown if the correlation is significant ($P < 0.05$). (**right**)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

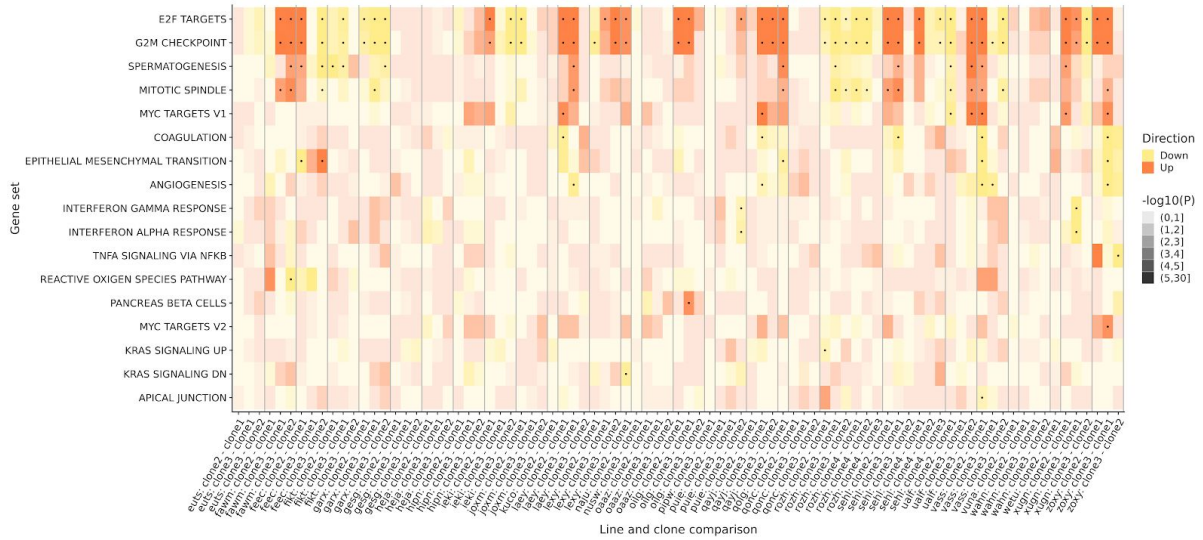


Figure S36. Heatmap showing the direction (first listed clone relative to second listed clone; in colour) and strength of enrichment (-log10(P) as degree of shading) for Hallmark gene sets tested with camera (Methods) for all pairwise comparisons between clones across 31 lines. Gene sets that are significantly 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.

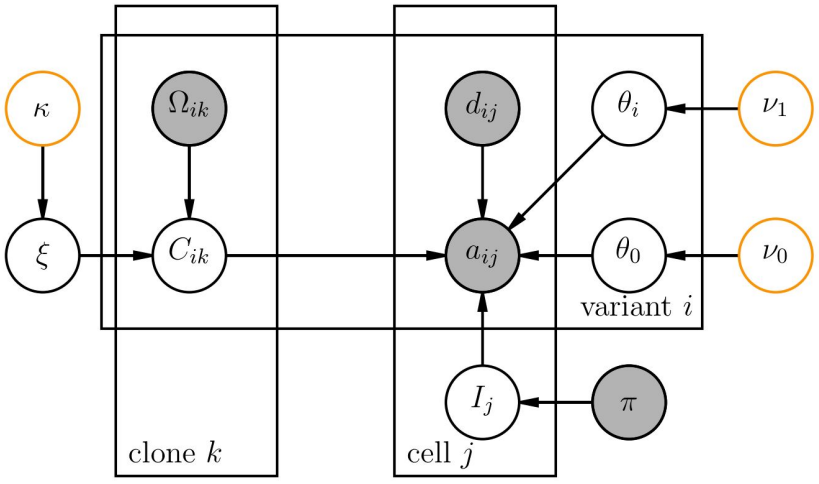


Figure S37. Graphical representation of the cardelino model. The clonal tree configuration matrix C is a random variable and follows a Bernoulli distribution encoded by an input tree configuration Ω that is provided to the model (e.g. estimated from bulk or single-cell DNA-seq data using existing methods 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.

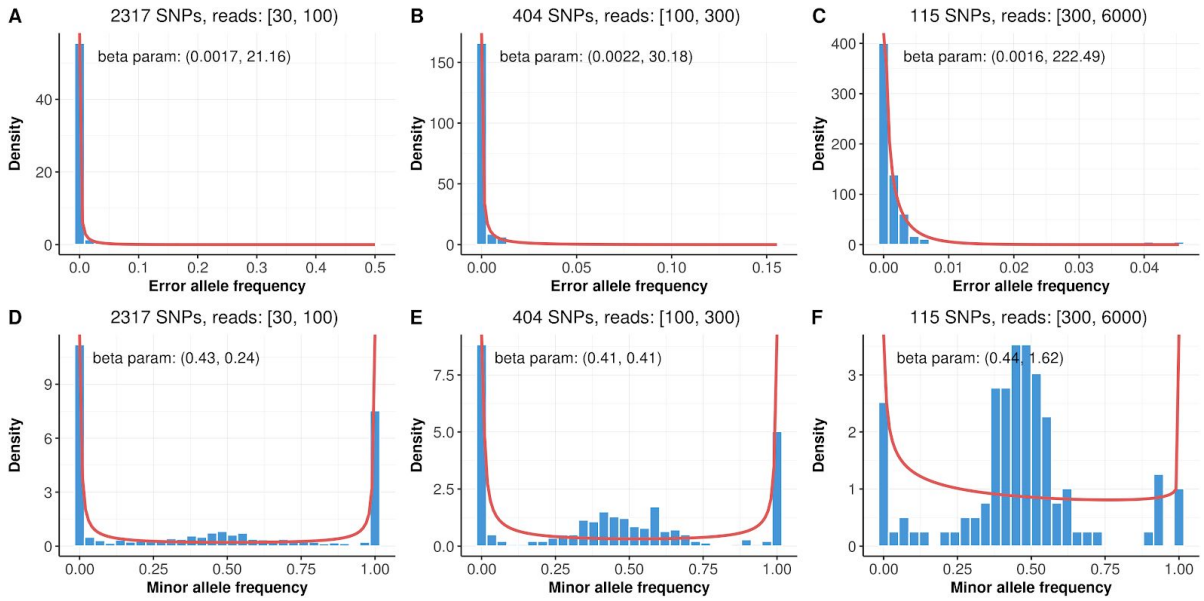


Figure S38. Estimated beta-binomial distribution of the "sequencing error rate" (theta0; **A-C**) and the alternative allele count rate given a variant is present (theta1; **D-F**) in single cells from germline heterozygous variants across three expression levels in donor vass. For each germline heterozygous 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).

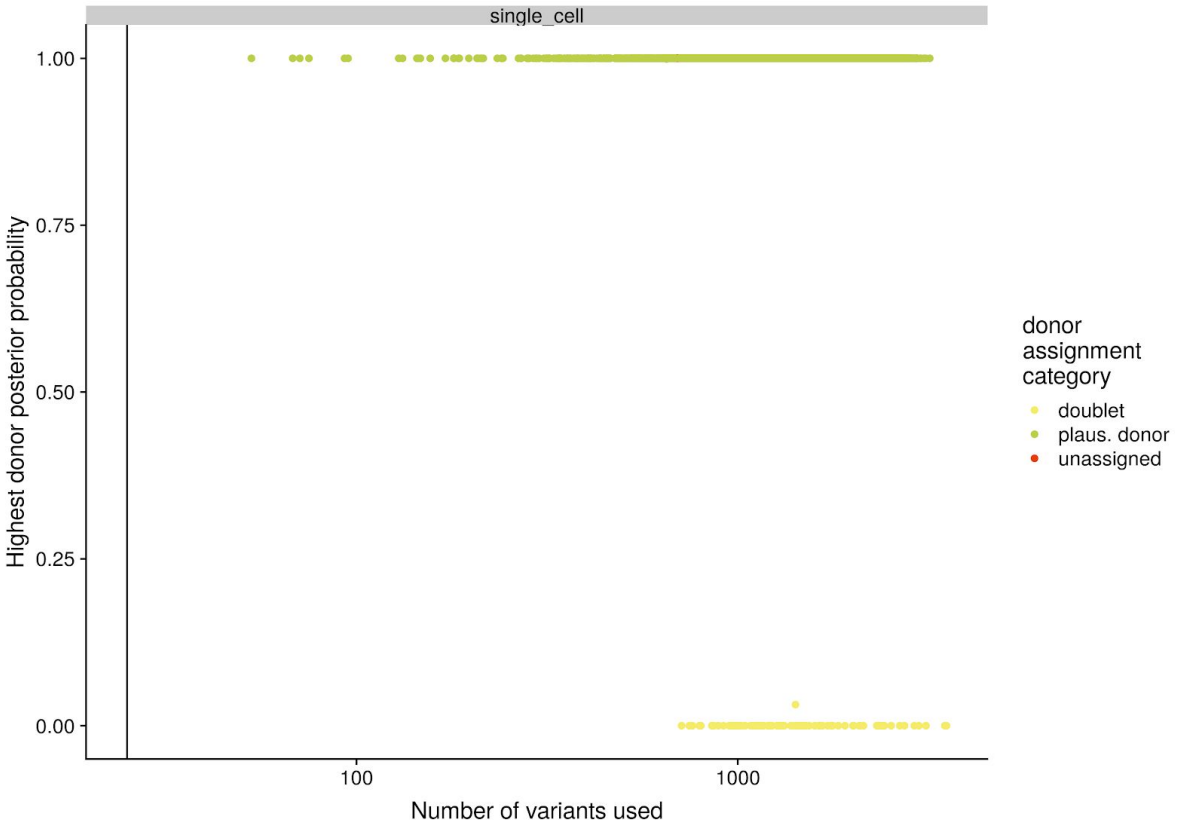


Figure S39. Donor identification results from cardelino for QC-passing cells for 32 fibroblast lines (*i.e.* donors) used to demultiplex cells from plates on which cells from three lines were pooled. The y-axis shows the highest posterior probability for donor assignment from cardelino (**Methods**) 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.

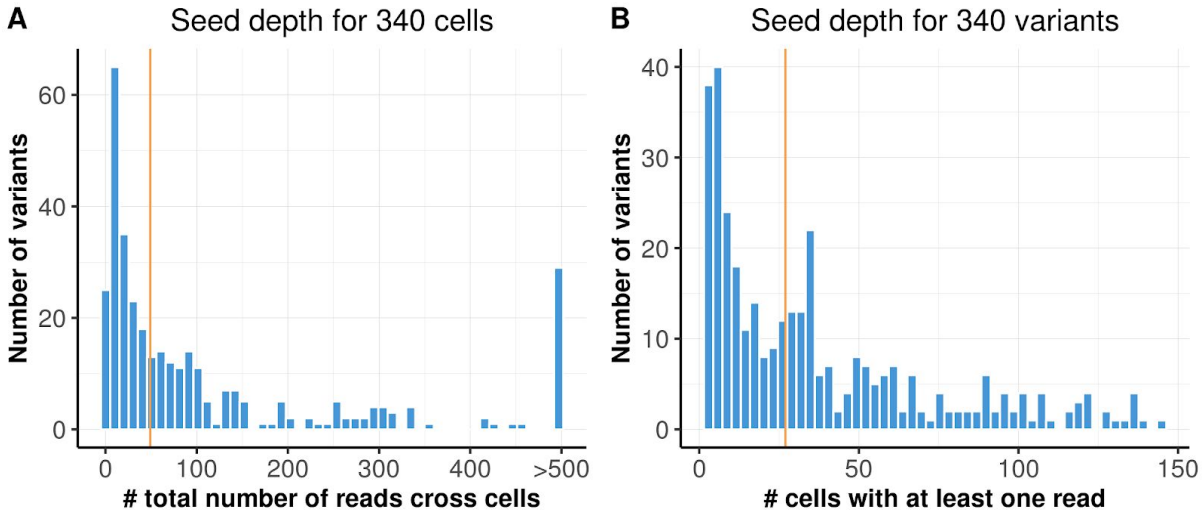


Figure S40. Summary of sequencing depths of 340 variants across a pool of 151 cells. **(A)** Histogram of total read counts on each variant across 151 cells, median number is shown in yellow; **(B)** Histogram of the number of cells with non-zero read coverage for each variant; median number is shown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

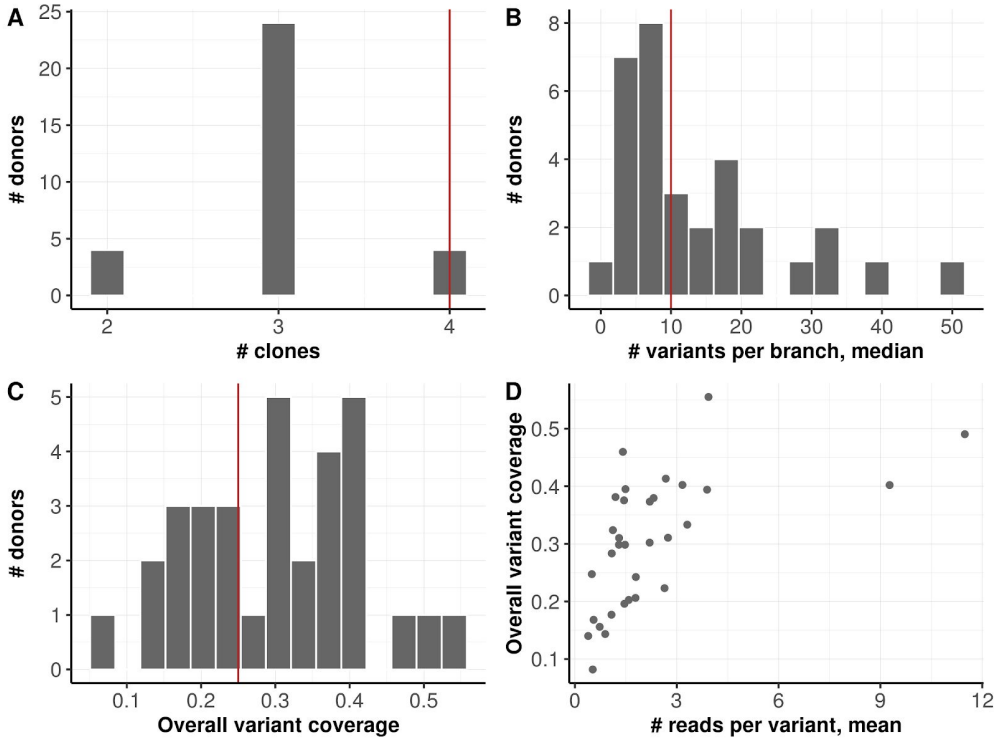


Figure S41. Distribution of key parameters in single cells assignment to clones across 32 donors: (**A**) number of clones inferred from bulk exome-seq data. (**B**) the median number of variants per clonal branch; (**C**) the overall coverage of variants, namely the fraction of variants with at least 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.