ARTICLES
https://doi.org/10.1038/s41592-020-0766-3

nature methods

Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes

Davis J. McCarthy[1,2,3,22], Raghd Rostom[1,4,22], Yuanhua Huang[1,5,22], Daniel J. Kunz[4,6,7], Petr Danecek[4], Marc Jan Bonder[1], Tzachi Hagai[1,4,8], Ruqian Lyu[2,3], HipSci Consortium*, Wenyi Wang[9], Daniel J. Gaffney[4], Benjamin D. Simons[6,7,10], Oliver Stegle[1,4,11,12] ✉ and Sarah A. Teichmann[1,4,6] ✉

Bulk and single-cell DNA sequencing has enabled reconstructing clonal substructures of somatic tissues from frequency and cooccurrence patterns of somatic variants. However, approaches to characterize phenotypic variations between clones are not established. Here we present cardelino (https://github.com/single-cell-genetics/cardelino), a computational method for inferring the clonal tree configuration and the clone of origin of individual cells assayed using single-cell RNA-seq (scRNA-seq). Cardelino flexibly integrates information from imperfect clonal trees inferred based on bulk exome-seq data, and sparse variant alleles expressed in scRNA-seq data. We apply cardelino to a published cancer dataset and to newly generated matched scRNA-seq and exome-seq 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 role for cell division genes in somatic evolution in healthy skin.

Aging, environment and genetic factors can affect mutational processes, thereby shaping the acquisition of somatic mutations across the life span[1–5]. The maintenance and evolution of somatic mutations can result in clonal structure. Whole-genome and whole-exome DNA sequencing of bulk cell populations has been used to reconstruct the mutational processes that underlie somatic mutagenesis[6–10] as well as clonal trees[11–13]. These strategies have been complemented by single-cell DNA sequencing (scDNA-seq)[14–16], which together with new computational approaches have helped to improve the reconstruction of clonal trees[17–23]. However, the functional differences between clones and their molecular phenotypes remain largely unknown.

An important step toward 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 of clone-to-clone gene expression differences in disease[24–27]. Targeted genotyping strategies linking known mutations of interest to single-cell transcriptomes have proved useful in particular settings, but remain limited by technical challenges and the requirement for strong previous information on informative variants[28–30]. Generally applicable methods for inferring the clone of origin of single cells to study genotype-transcriptome relationships have not yet been 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 assess its performance in comparison to existing methods. Finally, we demonstrate the use of cardelino by applying it to align single-cell transcriptome profiles to clonal substructure in five patients with melanoma cancer and 32 dermal fibroblast lines derived from healthy donors.

Results

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

Initially, we assess the accuracy of cardelino using simulated data that mimic clonal structures and properties of scRNA-seq as observed in real data (Methods and Supplementary Fig. 2). In addition to simulated variant profiles, we supply a guide clonal tree with a 10% error rate compared to the true simulated tree.

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [2]St Vincent's Institute of Medical Research, Fitzroy, Victoria, Australia. [3]Melbourne Integrative Genomics, School of Mathematics and Statistics/School of Biosciences, University of Melbourne, Parkville, Victoria, Australia. [4]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [5]Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. [6]Department of Physics, Cavendish Laboratory, Cambridge, UK. [7]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. [8]School of Molecular Cell Biology and Biotechnology, George S Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [9]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [10]The Wellcome Trust/Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge, UK. [11]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [12]Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. [22]These authors contributed equally: Davis J. McCarthy, Raghd Rostom, Yuanhua Huang. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: o.stegle@dkfz.de; st9@sanger.ac.uk

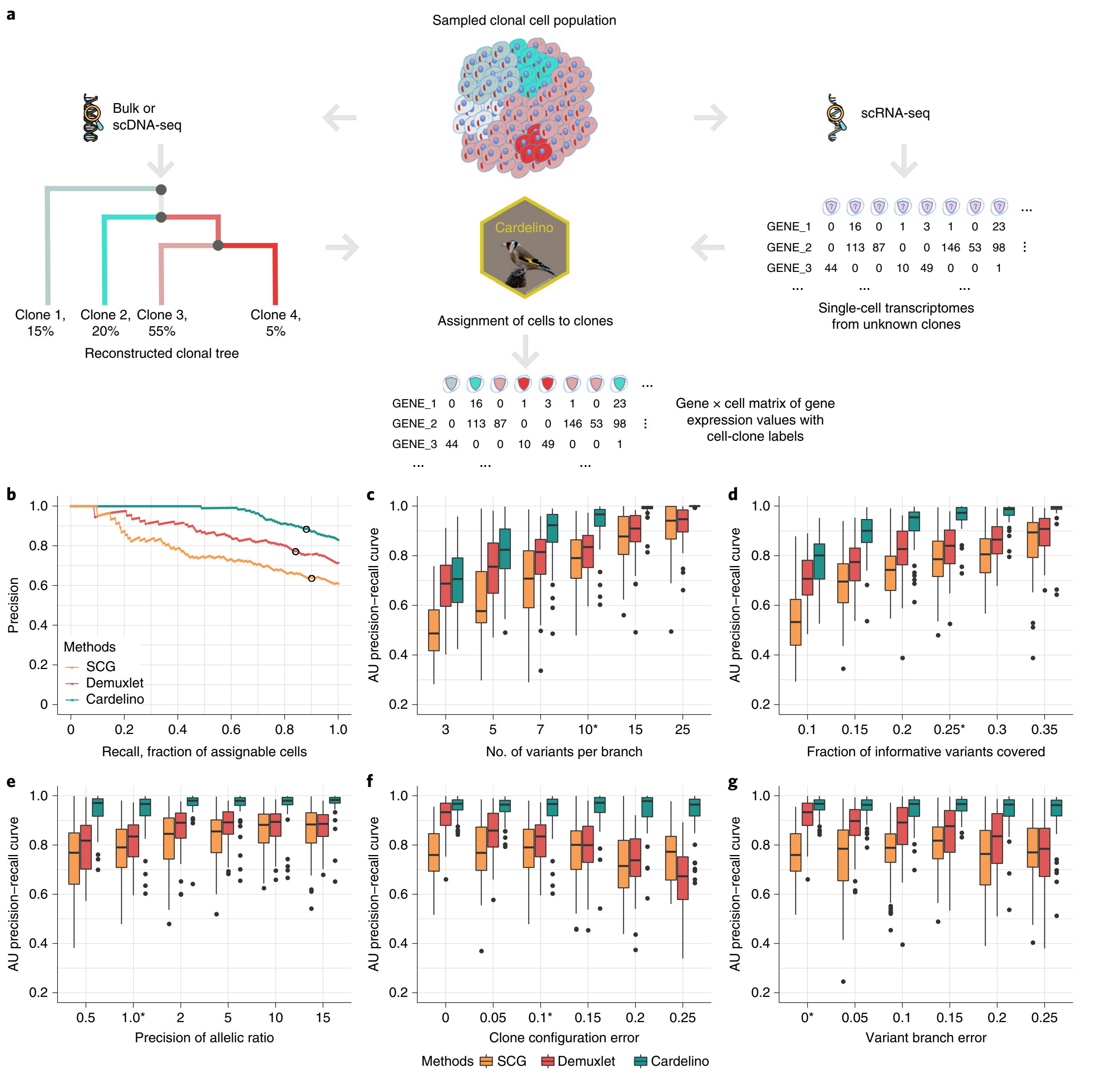

Fig. 1 | Overview and validation of the cardelino model. a, A guide clonal tree can be reconstructed using DNA sequencing (for example, WES) data. Cardelino then performs probabilistic clustering of single-cell transcriptomes based on somatic variants detected in scRNA-seq reads, assigning cells to clones in the clonal tree. **b**–**g**, Benchmarking of cell assignment using simulated data. Default parameter values used in the simulations are denoted with an asterisk. **b**, Overall assignment performance for a dataset comprising 200 cells, simulated assuming a four-clone structure with ten variants per branch and nonzero read coverage for 20% of the variants and an error rate of 10% on the mutation states between the guide clonal tree 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**, Box plot of the area under precision-recall curve (that is, area under (AU) curves such as shown in **b**) across $n = 50$ repeated simulations, when varying the numbers of variants per clonal branch (**c**), the fraction of informative variants covered (that is, nonzero scRNA-seq read coverage) (**d**), the precision (that is, 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**). Standard box plots here show median, box covering the 25 to 75% quartiles, whiskers extending up to 1.5 times the interquartile range above and below the box and outliers shown as points. For further details see Methods and Supplementary Note.

Alongside cardelino, we consider two alternative methods: Single Cell Genotyper (SCG[21]) and an implementation of Demuxlet[31], a method designed for sample demultiplexing rather than assigning single-cell transcriptomic profiles to clones (‘clone assignment’ of cells, see Methods). Cardelino achieves high overall performance (area under precision recall curve, 0.965; Fig. 1b), outperforming both SCG and Demuxlet.

We explore the effect of key dataset characteristics on clone assignment, including the number of variants per clonal branch (Fig. 1c) and the expected number of variants with nonzero scRNA-seq coverage per cell (Fig. 1d). As expected, the number of variants per clonal branch and their scRNA-seq read coverage are positively associated with the performance of all methods, with cardelino consistently outperforming alternatives, particularly in settings with low coverage. We further explore the effect of allelic imbalance on cell assignment (Fig. 1e), finding that cardelino is more robust than SCG and Demuxlet in regimes where a larger fraction of variants has high allelic imbalance: a source of error that is explicitly accounted for in cardelino. We also vary the error rate in the guide clonal tree, either by introducing uniform errors in the tree's configuration matrix (Fig. 1f) or by swapping variants between branches in the tree (Fig. 1g). In both settings, cardelino is markedly more robust than Demuxlet, which assumes that the defined guide clonal tree is error-free. In contrast, cardelino retains high performance (area under precision recall curve, >0.96) for error rates as high as 25% (Fig. 1f,g).

We also consider two simplified variants of cardelino, one that performs de novo tree reconstruction without considering a guide clonal tree (cardelino-free), and a second that treats the guide tree as error-free (cardelino-fixed). These comparisons, further investigating the parameters assessed in Fig. 1, confirm the benefits of the data-driven modeling of the guide clonal tree as prievious information that is refined while assigning scRNA-seq profiles to clones (Supplementary Figs. 3 and 4). We also explore the effects of the number of clones and the tree topology, again finding that cardelino is robust to these parameters (Supplementary Fig. 4).

Cardelino assigns single-cell transcriptomes to clones in human dermal fibroblasts. Next, we apply cardelino to a dataset from 32 human dermal fibroblast lines, allowing investigation of the clonal structure of skin samples from healthy donors. The lines were derived from nominally healthy donors that are part of the UK human induced pluripotent stem cell initiative (HipSci[32]; Supplementary Table 1). For each line, we generated deep whole-exome sequencing data (WES) (median read coverage, 254), and 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 quality control (QC); median coverage, 484,000 reads and median genes observed, 11,108; see Supplementary Table 2).

Initially, we consider high-confidence somatic single-nucleotide variants (SNVs) identified from WES data (Methods), which 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, two-sided Fisher exact test false discovery rate (FDR) ≤ 0.1; see Methods). Most SNVs can be attributed to the well-documented ultraviolet signature (COSMIC Signature 7; primarily C to T mutations[8]), agreeing with expected mutational patterns from ultraviolet exposure of skin tissues (Fig. 2a, Supplementary Fig. 5 and Methods). To explore whether the somatic SNVs confer a selective advantage, we used SubConalSelection at a per-line level[33]. Alternative methods such as dN/dS[34] and methods using the SNV frequency distribution[35,36] are not conclusive on this dataset, likely due to lack of statistical power resulting from the low number of mutations detected in our samples. The SubClonalSelection analysis identifies at least 12 lines with a clear fit to their selection model, suggesting positive selection of clonal subpopulations (Fig. 2a, Supplementary Fig. 6 and Methods). These patterns motivate the investigation of clone-specific variations in terms of transcriptome phenotypes.

Next, we reconstruct the clonal trees in each line using WES-derived estimates of the variant allele frequency (VAF) of somatic variants that are also present in scRNA-seq reads (Methods). Canopy[13] identifies two to four clones per line (Fig. 2a), which correspond to subpopulations of cells that share (and are identified by) a common somatic mutation profile and includes a 'base clone' without somatic mutations (Methods). Following Canopy's inference of clones, we use cardelino to map scRNA-seq profiles from 1,732 cells (posterior probability >0.9 for clone assignment, from 2,044 cells in total) to the underlying clones using the Canopy inference as guide clonal structure (Methods; for Canopy and cardelino output for all 32 lines see Supplementary File 1). 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 (Supplementary Fig. 7). 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 use of the adaptive phylogeny error model employed by cardelino. We also run the four alternative methods on these 32 lines for comparison (Supplementary Fig. 8).

To assess the confidence of these cell assignments further, we consider simulated cells drawn from a clonal structure that matches the corresponding line, again finding that cardelino outperforms alternative methods (Supplementary Fig. 8). 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 harbor fewer distinguishing variants are associated with lower assignment rates, at consistently high cell assignment accuracy (median 0.965, mean 0.939; see Supplementary Fig. 9), 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 (Supplementary Note). Consistent with these results, the clone prevalences estimated from Canopy and the fractions of cells assigned to the corresponding clones are qualitatively 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 and Supplementary Fig. 10).

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 reveals global transcriptome substructure aligned with the somatic clones identified in this cell population (Fig. 3b). Additionally, we observe differences in the fraction of cells in different cell-cycle stages, where clone 1 has the fewest cells in G1, and the largest fraction in S and G2/M (Fig. 3b, inset plot and Supplementary Note), suggesting that clone 1 is proliferating most rapidly. Next, we consider differential expression (DE) analysis of individual genes between the two largest clones (clone 1, 46 cells versus clone 2, 25 cells), which identifies 901 DE genes (two-sided edgeR QL *F*-test; FDR < 0.1; 549 at FDR < 0.05; Fig. 3c). Consistent with the cell-cycle stage assignment, we observe that genes that are overexpressed in clone 1 are enriched for processes associated with the cell cycle and cell proliferation[37–40] (gene sets E2F targets, G2/M checkpoint, mitotic spindle; two-sided camera test; FDR < 0.1; see Fig. 3d).

Cell cycle and proliferation pathways frequently vary between clones. To quantify the overall effect of somatic substructure on gene expression variation across the full dataset, we fit a linear mixed model to individual genes (Methods), partitioning gene expression variation into a line component, a clone component, technical batch (that is, processing plate), cellular detection rate (proportion of genes with nonzero expression per cell) and residual noise. While globally, the line component explains a larger fraction

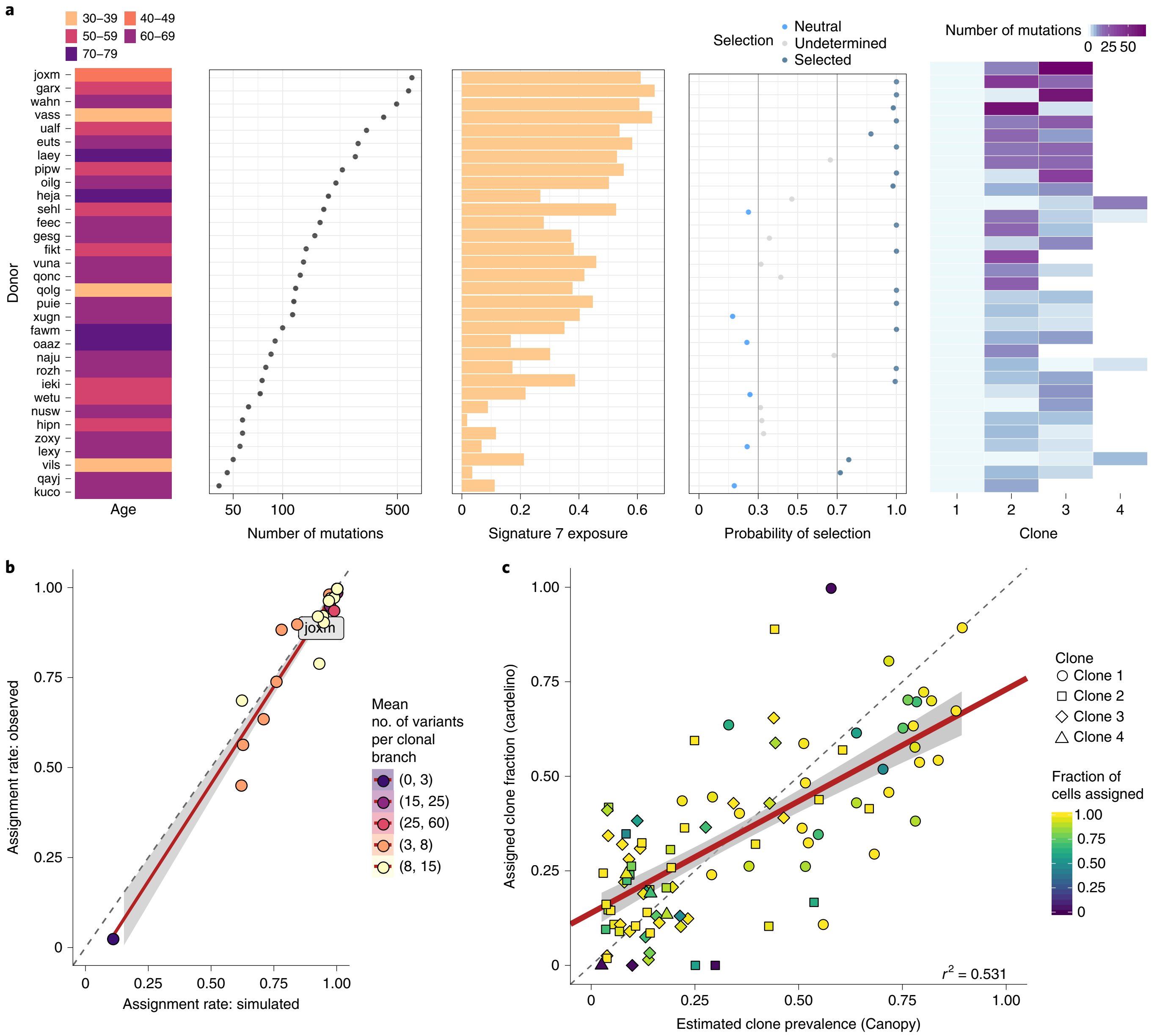

Fig. 2 | Parallel deep-exome sequencing and scRNA-seq profiling of 32 human dermal fibroblast lines. a, Overview and somatic mutation profiles across lines, from left to right: donor age; number of somatic SNVs; estimated exposure of COSMIC mutational signature 7; probability of selection estimated by SubClonalSelection[33]. Color denotes the selection status based on probability cut-offs (gray lines); the gray background indicates results with high uncertainty due to the low number of mutations detected; number of clones inferred using Canopy, with color indicating the number of informative somatic SNVs for cell assignment to each clone (nonzero read coverage in scRNA-seq data). **b**, Assignment rate (fraction of cells assigned) using matched simulated single-cell transcriptomes (*x* axis; see Methods) versus the empirical assignment rate (*y* axis) for each of $n = 32$ lines (at assignment threshold posterior $P > 0.5$). Color 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 a 95% confidence interval shown in gray. **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 ($n = 96$ dots) are the fractions of cells assigned to clones 1 to 4 as in **a**, considering the most likely assignment for assignable cells (posterior probability $P > 0.5$) with each point representing a cell line. Shape denotes the clone assigned (1, 2, 3 or 4) and color 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 gray.

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

Next, we assess transcriptomic differences between any pair of clones for each line by testing for DE genes (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). Most, 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 lines. Comparison to data with permuted gene labels demonstrates an excess of recurrently differentially expressed genes compared to chance expectation

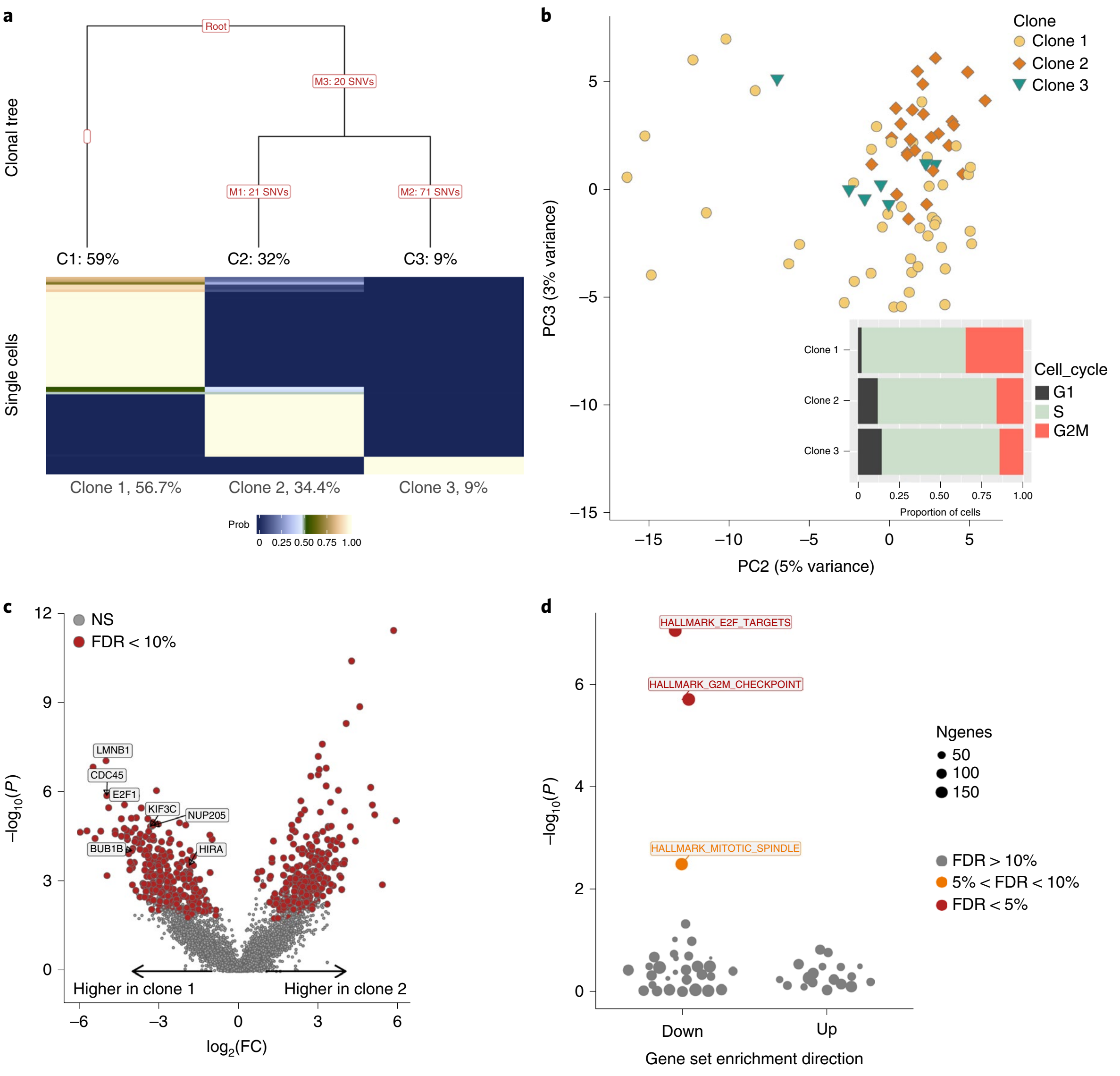

Fig. 3 | Clone-specific transcriptome profiles reveal gene expression differences for joxm, one example line. a, Top, clonal tree inferred using Canopy (Jiang et al.[13]). The number of variants tagging each branch and the expected prevalence (fraction) of each clone is shown. Bottom, cardelino cell-to-clone 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 $n = 78$ scRNA-seq profiles with color indicating the most likely clone assignment. Inset plot, cell-cycle phase fractions for cells assigned to each clone (using cyclone, see Methods). **c**, Volcano plot showing negative $\log_{10} P$ values (quasi-likelihood F-test from edgeR, two-sided) versus $\log_2$-fold changes (FC) for DE between cells assigned to clone 2 and clone 1 ($n = 44$ versus 27 cells). Significant differentially expressed genes (FDR < 0.1) are highlighted in red. **d**, Enrichment of MSigDB Hallmark gene sets using the two-sided camera test (Methods) based on $\log_2$ fold change values between clone 2 and clone 1 as in **c**. Shown are negative $\log_{10} P$ values of gene set enrichments, considering whether gene sets are up-regulated in clone 1 or clone 2, with significant (FDR < 0.05) gene sets highlighted and labeled. All results are based on 78 out of 79 cells that could be confidently assigned to one clone (posterior $P > 0.5$, Methods).

(Fig. 4c, $P < 0.001$, 1,000 permutations; see Methods). We also identify a small number of genes that contain somatic variants in a subset of clones, resulting in DE between the base clone (clone 1, no somatic variants) and mutated clones (Supplementary Fig. 11).

To align the transcriptomic changes across lines, we assessed shared patterns of enriched pathways (Methods). Of 31 lines, 19 show significant enrichment for at least one MSigDB Hallmark gene set (FDR < 0.05, camera; see Methods), with key gene sets related to cell cycle and growth enriched recurrently (Fig. 4d). Furthermore, the directionality of the expression changes of the 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 (Supplementary Fig. 12). A second cluster of pathways that vary in a coordinated fashion, related to epithelial-mesenchymal transition and apical junction, was anticorrelated with expression changes in cell cycle and proliferation pathways (Fig. 4d). There is substantial heterogeneity in gene expression programs across clones within individual lines, where we observe that the enrichment of pathways often differs between different pairs of clones (Fig. 4e, all lines shown in Supplementary Fig. 12). Overall, a picture of substantial complexity in clonal gene expression emerges, highlighting the variability in effects of somatic variants on the phenotypic behavior of cells.

Cardelino assigns single-cell transcriptomes to clones in human melanoma samples. Finally, we consider a second dataset and apply

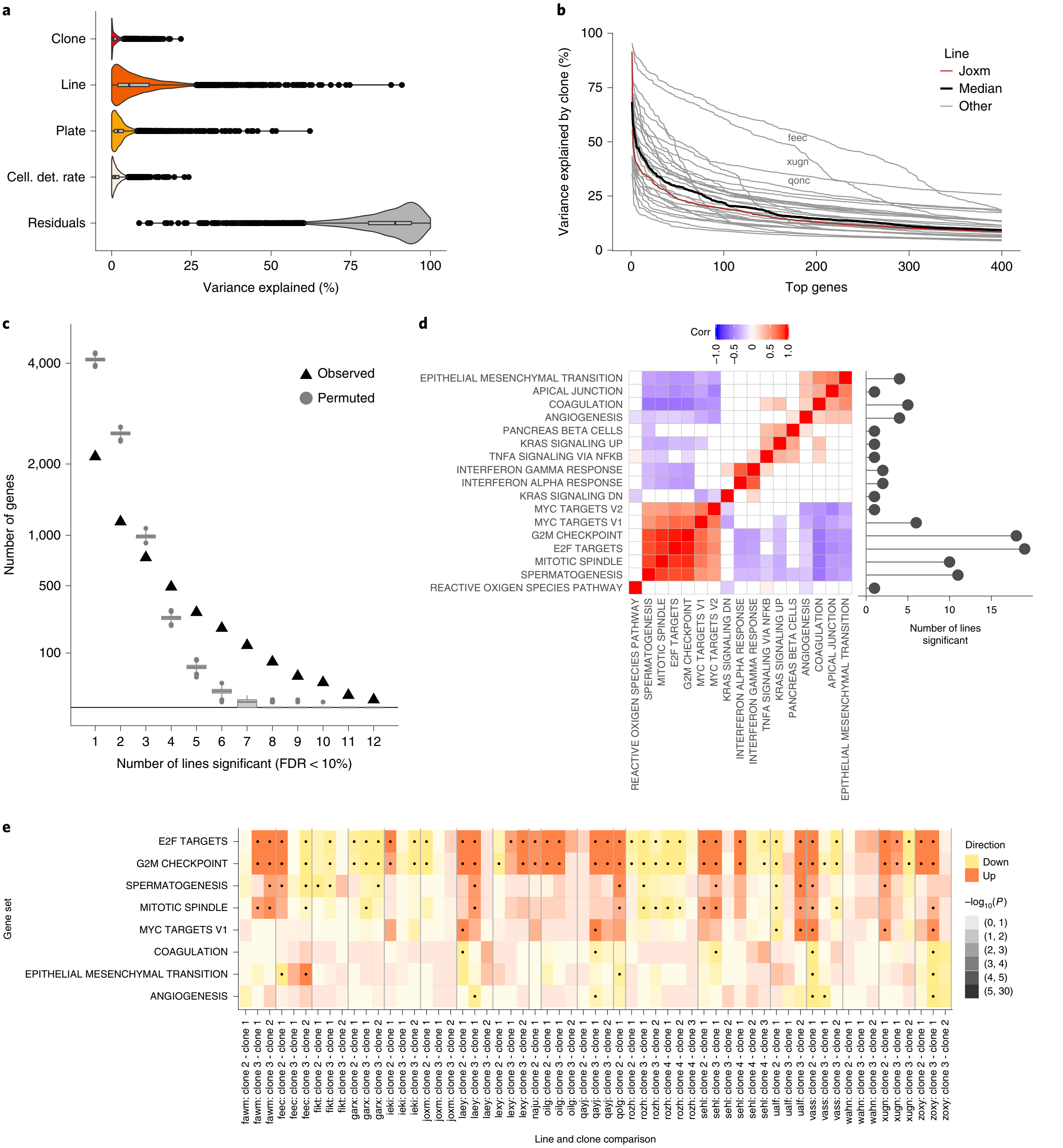

Fig. 4 | Signatures of transcriptomic clone-to-clone variation across 31 lines. a, Violin and box plots show the percentage of variance explained by clone, line, experimental plate and cellular detection (Cell. det.) rate for $n = 4{,}998$ highly variable genes, estimated using a linear mixed model (Methods). **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 genes between any pair of clones (FDR < 0.1; edgeR QL *F*-test, *n* ranges from 29 to 107 cells across lines), detected in at least 1–12 lines, with box plots showing results expected by chance (using $n = 1{,}000$ permutations). **d**, Left panel, heatmap showing pairwise correlation coefficients (Spearman's *R*, only nominal significant correlations shown, $P < 0.05$) between signed *P* values of gene set enrichment (two-sided camera test) across $n = 32$ lines, based on differentially expressed genes (*n* ranges from 2 to 1,113 DE genes across lines) 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 (same test as for the left panel, FDR < 0.05). **e**, Heatmap depicting signed *P* values of gene set enrichments (two-sided camera test) for eight Hallmark gene sets in $n = 19$ lines. Dots denote significant enrichments (FDR < 0.05). Violin and box plots in **a** and **c** are standard: violin plots show data density with default bandwidth calculation and box plots show median, box covering 25 to 75% quartiles, whiskers extending up to 1.5 times the interquartile range above and below the box and outliers shown as points.

cardelino to 1,290 single cells from five patients with metastatic melanoma for which both somatic variant calls from WES data and scRNA-seq data are available[41]. As for the fibroblast data, we infer clonal trees using Canopy and assign cells to clones using cardelino (Methods and Supplementary Note). Cardelino is able to assign most of the cells with high-confidence (Supplementary Note). As expected, across confidently assigned immune cells (defined using marker genes from the original study), we find 96–100% assigned to the base clone (Supplementary Fig. 13a; see Supplementary Note for more details and results). Between 0 and 21% of melanoma cells (median 7%) are assigned to the base clone, with the rest assigned to other clones harboring somatic variants. We find hundreds of differentially expressed genes across all pairwise comparisons between clones (edgeR two-sided QL F-test, FDR < 0.05; Supplementary Note). Consistent with the results observed on the fibroblast data, we observe patterns of DE related to cell proliferation. In three of the five patients with melanoma we observe at least five pathways to be significantly enriched between clones (two-sided camera test, FDR < 0.05; see Methods and Supplementary Note), with E2F targets, MTORC1 signaling, mitotic spindle and G2M checkpoint among the Hallmark gene sets significantly enriched in at least two patients (Supplementary Fig. 13b). These gene sets are also recurrently significantly enriched when restricting the analysis only to melanoma cells (Supplementary Fig. 13c). Collectively, these results show that cardelino is able to distinguish between tumor and nontumor cells with very high accuracy, and also to differentiate between tumor clones on the basis of distinct SNVs expressed in scRNA-seq reads.

Discussion

Here we develop and apply a computational approach for integrating somatic clonal structure with scRNA-seq data. Cardelino leverages a clonal tree inferred from WES or scDNA-seq as prior information to then assign single-cell transcriptomes to individual clones. This assignment step is conceptually related to existing demultiplexing methods for single-cell transcriptomes from multiple genetically distinct individuals[31]. However, cardelino addresses a substantially more challenging problem: to distinguish cells from the same individual based on the typically small number of somatic variants (for example, 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 taking advantage of imperfect clonal trees derived from complementary assays such as bulk exome-seq data.

Inferring clonal trees from any type of data remains a challenging task and existing clonal inference methods produce clonal trees with substantial uncertainty. Consequently, the ability of cardelino to adapt these trees in the light of variant information from scRNA-seq is a key strength of the method. Our results show that cardelino outperforms methods that consider an input clonal tree as fixed and error-free (Demuxlet, cardelino-fixed) and those that do not use any guide tree at all (SCG, cardelino-free). Comparing cardelino and cardelino-free highlights complementary strengths and use cases. Where external data for inference of the clonal tree exist (such as bulk exome or scDNA-seq), cardelino can be more accurate than cardelino-free. In settings where such data are not available, cardelino-free offers de novo clonal inference and assignment with competitive accuracy. Currently, cardelino only considers SNV data; future development could incorporate information on copy number and structural variation into the model to improve clonal tree inference and cell assignment. Cardelino also currently requires somatic variants to have been called using bulk or scDNA-seq data, which limits its use to settings where both DNA-seq and scRNA-seq data are available. Future development of an accurate somatic variant caller for scRNA-seq data would relax this need for complementary DNA-seq data and allow cardelino to be applied in many more settings.

Harnessing transcriptomic phenotypic information for cells assigned to clones in 32 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 VAF distributions from WES data reveals evidence for positive selection of clones in 12 of 32 lines, consistent with the results from our gene expression analysis. This surprising result in healthy tissue suggests inter-clonal phenotypic variation with important functional consequences, a result that will, however, require independent experimental validation. Additionally, the clonal dynamics in fibroblast cell lines may deviate from the in vivo situation in primary fibroblast tissue. Nevertheless, 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 cardelino method is general and can exploit available information on clonal substructure inferred from bulk or scDNA-seq data or other sources. If no external information on clonal structure is available, the cardelino-free method simultaneously infers clonal structure and assigns cells to clones with high accuracy. Cardelino's inference procedure is computationally efficient, so will 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. Although not explored here, the cardelino model may also be effective for other single-cell 'omics assays that capture somatic variant information, such as those profiling chromatin accessibility[42] or methylation[43,44].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0766-3.

Received: 26 November 2018; Accepted: 31 January 2020;
Published online: 16 March 2020

References

1. Burnet, F. M. Intrinsic mutagenesis: a genetic basis of ageing. *Pathology* **6**, 1–11 (1974).
2. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
3. Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
4. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
5. Huang, K.-L. et al. Pathogenic germline variants in 10,389 adult cancers. *Cell* **173**, 355–370.e14 (2018).
6. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
7. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
8. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
9. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
10. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320.e10 (2018).
11. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396 (2014).

12. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
13. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **113**, E5528–E5537 (2016).
14. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
15. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
16. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25**, 1499–1507 (2015).
17. Kim, K. I. & Simon, R. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinf.* **15**, 27 (2014).
18. Navin, N. E. & Chen, K. Genotyping tumor clones from single-cell data. *Nat. Methods* **13**, 555–556 (2016).
19. Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome Biol.* **17**, 86 (2016).
20. Kuipers, J., Jahn, K., Raphael, B. J. & Beerenwinkel, N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27**, 1885–1894 (2017).
21. Roth, A. et al. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nat. Methods* **13**, 573–576 (2016).
22. Salehi, S. et al. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.* **18**, 44 (2017).
23. Malikic, S. et al. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commun.* **10**, 2750 (2019).
24. Müller, S. et al. Single-cell sequencing maps gene expression to mutational phylogenies in PDGF- and EGF-driven gliomas. *Mol. Syst. Biol.* **12**, 889 (2016).
25. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
26. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).
27. Campbell, K. R. et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.* **20**, 54 (2019).
28. Giustacchini, A. et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
29. Cheow, L. F. et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods* **13**, 833–836 (2016).
30. Saikia, M. et al. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nat. Methods* **16**, 59–62 (2019).
31. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
32. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
33. Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
34. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **173**, 1823 (2018).
35. Simons, B. D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl Acad. Sci. USA* **113**, 128–133 (2016).
36. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238 (2016).
37. Ramaker, R. C. et al. RNA sequencing-based cell proliferation analysis across 19 cancers identifies a subset of proliferation-informative cancers with a common survival signature. *Oncotarget.* **8**, 38668–38681 (2017).
38. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
39. Tsang, J. C. H. et al. Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. *Genome Biol.* **16**, 178 (2015).
40. Kolodziejczyk, A. A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
41. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
42. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
43. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
44. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

HipSci Consortium

Helena Kilpinen[13,14], Angela Goncalves[13], Andreas Leha[13,15], Vackar Afzal[16], Kaur Alasoo[13], Sofie Ashford[17], Sendu Bala[13], Dalila Bensaddek[16], Marc Jan Bonder[1], Francesco Paolo Casale[1], Oliver J. Culley[18], Anna Cuomo[1], Petr Danecek[13], Adam Faulconbridge[1], Peter W. Harrison[1], Annie Kathuria[18], Davis J. McCarthy[1,2], Shane A. McCarthy[13], Ruta Meleckyte[18], Yasin Memari[13], Bogdan Mirauta[1], Nathalie Moens[18], Filipa Soares[19], Alice Mann[13], Daniel Seaton[1], Ian Streeter[1], Chukwuma A. Agu[13], Alex Alderton[13], Rachel Nelson[13], Sarah Harper[13], Minal Patel[13], Alistair White[13], Sharad R. Patel[13], Laura Clarke[1], Reena Halai[13], Christopher M. Kirton[13], Anja KolbKokocinski[13], Philip Beales[14], Ewan Birney[1], Davide Danovi[18], Angus I. Lamond[16], Willem H. Ouwehand[13,17,20], Ludovic Vallier[13,19], Fiona M. Watt[18], Richard Durbin[13,21], Oliver Stegle[1,11,12] and Daniel J. Gaffney[13]

[13]Wellcome Genome Campus, Wellcome Trust Sanger Institute, Hinxton, UK. [14]UCL Great Ormond Street Institute of Child Health, University College London, London, UK. [15]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee, Germany. [16]Centre for Gene Regulation & Expression, School of Life Sciences, University of Dundee, Dundee, UK. [17]Department of Haematology, University of Cambridge, Cambridge, UK. [18]Centre for Stem Cells & Regenerative Medicine, King's College London, Tower Wing, Guy's Hospital, Great Maze Pond, London, UK. [19]Wellcome Trust and MRC Cambridge Stem Cell Institute and Biomedical, Research Centre, Anne McLaren Laboratory, University of Cambridge, Cambridge, UK. [20]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK. [21]Department of Genetics, University of Cambridge, Cambridge, UK.

Methods

The cardelino model. Given a list of N common variants and an estimated number of clones, K, the cardelino model jointly infers the clonal tree configuration and assigns single cells to one of the K clones by modeling the expressed alleles with a probabilistic clustering model (see graphical model in Supplementary Fig. 1).

The clonal tree configuration is represented as *an* $N \times K$ binary matrix C, with entries $c_{i,k} = 1$ if somatic variant i is present in clone k and $c_{i,k} = 0$ otherwise. Entries in matrix C are latent variables (unobserved) and can be inferred by incorporating a prior (an observed clone configuration matrix Ω with an appropriate relaxation (or error) rate $\xi \in [0, 1)$). The guide matrix Ω (with entries $\Omega_{i,k} \in \{0,1\}$) can be seen as an observed analogous matrix of C, which is considered informative but imperfect with error ξ. Specifically, we model entries in C as

$$P\left(c_{i,k} = 1 | \Omega_{i,k}, \xi\right) = \xi^{\left(1-\Omega_{i,k}\right)}(1-\xi)^{\Omega_{i,k}} \qquad (1)$$

This indicates that if $\Omega_{i,k} = 1$, then $c_{i,k} = 1$ with probability $1 - \xi$ and $c_{i,k} = 1$ with probability ξ. The special case $\xi = 0$, corresponds to Ω being a perfect representation of the clonal tree configuration; that is, $C = \Omega$. In this study, Ω is obtained from bulk exome sequencing data source using Canopy[13], which also yields the estimate of K. For the prior of ξ, we introduce a conjugate prior beta distribution, with hyper-parameter vector $\boldsymbol{\kappa}$: $P(\xi|\boldsymbol{\kappa}) = Beta(\xi; \kappa_0, \kappa_1)$. By default, we set $(\kappa_0, \kappa_1) = (1,9)$. The *beta* distribution is in format Beta(α, β) here and below.

Let $I_j \in \{1, \ldots, K\}$ denote cell identity, that is the clone to which cell j is assigned. Then I_j is a categorical variable with a (prior) probability distribution $F = (f_1, \ldots, f_K)$, with elements $f_k \in [0, 1]$ satisfying $\sum_{k=1}^{K} f_k = 1$. One possible specification of F could be the clonal fraction resulting from Canopy[13] with each f_k representing the relative prevalence of clone k. However, to avoid biasing cell assignment toward highly prevalent clones for cells with little read information, we assume an uninformative uniform prior:

$$P\left(I_j = k|F\right) = 1/K \text{ for all } k. \qquad (2)$$

Let M be the number of cells. For each cell and variant that segregates between clones, we extract the number of sequencing reads that support the reference allele (reference read count) and the alternative allele (alternative read count). The core part of the cardelino model is to model the alternative read count using a binomial model.

Specifically, let $a_{i,j}$ and $d_{i,j}$ denote, respectively, the alternative and total read count for variant i in cell j, where total read count is the sum of reference and alternative read counts. By definition, both $a_{i,j}$ and $d_{i,j}$ are nonnegative integers, collected in two $N \times M$ matrices A and D, respectively. We assume that $a_{i,j}$ follows a binomial distribution Binom(n,p), where the 'number of trials' n represents the total number of observed reads (that is, $n = d_{i,j}$) and 'success' is defined as observing an alternative-allele read.

Thus, the 'success probability' p corresponds to the probability that an observed read is from alternative allele, in theory equivalent to the fraction of alternative allele among the two alleles in expression. To this end, we parameterize p for two different settings: $p = \boldsymbol{\theta}_0$ for homozygous reference alleles (variant absence, $c_{i,k} = 0$ and $p = \boldsymbol{\theta}_i$ for heterozygous variants i (variant present, $c_{i,k} = 1$). Note that when $c_{i,k} = 0$, that is, a given variant is not present in clone k, we model the observed number of (false positive) reads that carry the variant using a common parameter value $\boldsymbol{\theta}_0$, which encodes the expected frequency of sequencing errors or errors in the clonal tree configuration. In contrast for $c_{i,k} = 1$, that is, if the variant is present in the clone, the binomial rate denotes the expected fraction of reads that carry the mutation. In the absence of assay noise and assuming heterozygous sequence variants this rate would be 0.5. To account for the gene-specific differences in allelic imbalance, which causes the probability of observing alternative reads to vary, we fit a gene-specific success probability. Denote the collection of all binomial success probability parameters as $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_{1,\ldots,}\boldsymbol{\theta}_N)$. Therefore, the binomial model for the alternative read count can be written as

$$P\left(a_{i,j}|d_{i,j}, I_j = k, c_{i,k}, \boldsymbol{\theta}\right) = \begin{cases} \text{Binom}\left(a_{i,j}|d_{i,j}, \boldsymbol{\theta}_0\right) & \text{if } c_{i,k} = 0 \\ \text{Binom}\left(a_{i,j}|d_{i,j}, \boldsymbol{\theta}_i\right) & \text{if } c_{i,k} = 1 \end{cases}. \qquad (3)$$

The prior distribution for parameters $\boldsymbol{\theta}$ is taken as a beta distribution with hyper-parameters v (see equations (10) and (11) in the Supplementary Note for details of $P(\boldsymbol{\theta}|\nu)$). To ensure sensible prior distributions, we estimate ν from scRNA-seq data at known germline heterozygous variants for highly expressed genes (Supplementary Note). For example, in the fibroblast dataset considered here, this approach yield prior parameters of *Beta*(0.2,99.8) for $\boldsymbol{\theta}_0$ and *Beta*(0.45,0.55) for $\theta_i, i \in \{1, \ldots, N\}$.

The posterior distribution of the clonal tree configuration C, cell assignment $I = (I_1, \ldots, I_M)$, parameters $\boldsymbol{\theta}$ and ξ then follows as

$$P(C, I, \xi, \theta|\mathbf{A}, \mathbf{D}, \Omega, \mathbf{F}) \propto P(A|D, I, C, \theta)P(C|\Omega, \xi)P(I|F)P(\xi|\kappa)P(\boldsymbol{\theta}|\nu), \qquad (4)$$

where $P(A|D, I, C, \boldsymbol{\theta}) = \prod_i \prod_j P\left(a_{i,j}|d_{i,j}, I_j = k, c_{i,k}, \boldsymbol{\theta}\right)$ is the likelihood across all cells and variants (as defined in equation (3)).

When conditioning on the state of the latent variables C and $\boldsymbol{\theta}$, the posterior probability of cell j coming from clone k follows as

$$P\left(I_j = k|a_j, d_j, C, F, \boldsymbol{\theta}\right) = \frac{P\left(I_j = k|F\right) \prod_i P\left(a_{i,j}|d_{i,j}, I_j = k, c_{i,k}, \boldsymbol{\theta}\right)}{\sum_{t=1}^{K} P\left(I_j = t|F\right) \prod_i P\left(a_{i,j}|d_{i,j}, I_j = t, c_{i,k}, \boldsymbol{\theta}\right)}. \qquad (5)$$

Following a Bayesian treatment of the model, we account for the uncertainty of these two variables C and $\boldsymbol{\theta}$ by marginalizing them out from the full posterior (equation (4)) using Gibbs sampling. See the Supplementary Note for details. In addition to the full model, we also provide implementations for two (simplified) variants of cardelino: cardelino-free without any informative clone configuration Ω and cardelino-fixed assuming that the clone prior is fixed and error-free ($\xi = 0$). Despite the fully 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 up to tens of thousands of cells.

Alternative methods. Aside from cardelino, two alternative methods with distinct strategies are considered: Demuxlet that assumes the guide clonal tree is perfect[31], and SCG that does not take any guide clonal tree[21].

The software implementation of 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 in its original paper[31]. We set the sequencing error rate to 0.003 for all reads, by matching our simulation settings. We also compared our implementation to the original implementation by assessing the ability to demultiplex pooled scRNA-seq data, finding perfectly concordance (Supplementary Note).

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 (that is, observing an alternative allelic read from homozygous reference genotype), we simply take the genotype g_{ij} for variant i in cell j as 1 (that is, heterozygous) if there is any alternative allelic read (that is, $a_{ij} > 0$), otherwise we take $g_{ij} = 0$ (that is, 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 to fix the first clone as a base clone; that is, with no mutations.

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 reordering 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. Then, 18 h before collection, cells were trypsinized (Life Technologies/25300054), counted and seeded at a density of 100,000 cells per well (six-well plate).

Cell pooling, capture and full-length transcript scRNA sequencing. Cells were washed with PBS, trypsinized 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 2 µl per well lysis buffer. Single cells were sorted individually (using FSC-W versus 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[45], and library preparation was performed using an Illumina Nextera kit. Samples were sequenced using paired-end 75 base pair reads on an Illumina HiSeq 2500 machine.

Bulk WES data and somatic variant calling. We obtained bulk WES data from HipSci fibroblast (median read coverage, 254) and derived iPS cell lines (median read coverage, 79) released by the HipSci project[32,46]. Sequenced reads were aligned to the GRCh37 build of the human reference genome[47] using bwa-mem[48]. To identify single-nucleotide somatic variant sites in the fibroblast lines, we compared VAFs 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 tumor-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. Thus, we apply a variant calling approach specific to our experimental setting here.

For each exome sample, we searched for sites with a nonreference base in the read pileup using bcftools/mpileup[49]. In the initial prefiltering 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[50] and 1,000 Genomes data[51]. A two-sided Fisher exact test[52] implemented in bcftools/ad-bias was then used to identify sites with significantly different VAFs 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 nor 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 artifacts. 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 to call copy number aberrations in fibroblasts. Calls were filtered to exclude copy number aberrations with a quality score less than two, deletions with fewer than ten markers and duplications with fewer than ten heterozygous markers. We also excluded any calls that were smaller than 200 kilobases.

Estimation of mutational signatures. Signature exposures were estimated using the sigfit package[53], providing the COSMIC 30 signatures as a reference[8], and with a highest posterior density threshold of 0.9. Signatures were determined to be significant when the highest posterior density 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[33–36]. 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 v.0.6.2, which works with a low number of mutations (>100 mutations[33]). 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 $f_{\mathrm{min}} = 0.05$ (ref. [54]). At the upper end of the allele frequency distribution we chose a cut-off at $f_{\mathrm{max}} = 0.45$ 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 that are neither 'selected' nor 'neutral' remain 'undetermined'. See Fig. 3a and Supplementary Fig. 8 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 shows peaks for selected clones; compare Fig. 1 in ref. [33]). 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 (v.1.5), before reads were adapter- and quality-trimmed with TrimGalore! (github.com/FelixKrueger/TrimGalore)[55]. We quantified transcript-level expression using Ensembl v.75 transcripts[56] by supplying trimmed reads to Salmon v.0.8.2 and using the '–seqBias', '–gcBias' and 'VBOpt' options[57]. Transcript-level expression values were summarized at gene level (estimated counts) and quality control of scRNA-seq data was done with the scater package[58] and normalization with the scran package[59,60]. Cells were retained for downstream analyses if they had at least 50,000 counts from endogenous genes, at least 5,000 genes with nonzero 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% (Supplementary Note). This filtering approach retains 63.7% of assayed cells. We used a pooled experimental design, whereby cells from three individual cell lines were pooled before scRNA-sequencing. Deconvolution of cells to donor cell lines was subsequently conducted using common genetic variants as natural cell barcodes (Supplementary Note), resulting in 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[13]. We used read counts for the variant allele and total read counts at filtered somatic variant sites from high-coverage WES 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 nonzero 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 v.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 nonzero 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.

Variance component, DE and pathway analysis. Gene expression analyses were carried out in exactly the same way for the fibroblast and melanoma datasets, as follows. Expression analyses between clones required further filtering of cells for each line (or patient in the melanoma data; below we will use 'line' to denote either line or patient as appropriate). 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; and (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 fibroblast lines (all except vils). All patients with melanoma were retained for expression analyses.

Expression variance across cells is decomposed into multiple components in a linear mixed model, including cellular detection rate (proportion of genes with nonzero expression per cell) as a fixed effect and plate (that is, experimental batch), donor (that is, 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[61].

Differential gene expression testing was conducted using the quasi-likelihood *F*-test method[62] in the edgeR package[63,64] as recommended by Soneson and Robinson[60]. 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 (15,649 genes in the melanoma data) that were sufficiently expressed (an average count >1 across cells in all lines) to test for DE.

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 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[61] method in the limma package[65,66]. Using $\log_2$-fold-change test statistics for 10,876 genes (15,649 genes in the melanoma data) 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[62]. For all DE and pathway analyses we adjusted for multiple testing by estimating the FDR using independent hypothesis weighting[63], as implemented in the IHW package, with average gene expression supplied as the independent covariate.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

scRNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7167. WES data is available through the HipSci portal (www.hipsci.org). The lines used in this study have the identifiers: euts, fawm, feec, fikt, garx, gesg, heja, hipn, ieki, joxm, kuco, laey, lexy, naju, nusw, oaaz, oilg, pipw, puie, qayj, qolg, qonc, rozh, sehl, ualf, vass, vils, vuna, wahn, wetu, xugn, zoxy. Metadata, processed data and large results files are available at https://doi.org/10.5281/zenodo.1403510

Code availability

The cardelino methods are implemented in an open-source, publicly available R package (github.com/single-cell-genetics/cardelino). The code used to process and analyse the data is available (github.com/davismcc/fibroblast-clonality), with a reproducible workflow implemented in Snakemake[64]. 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.

References

45. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
46. Streeter, I. et al. The human-induced pluripotent stem cell initiative—data resources for cellular genetics. *Nucleic Acids Res.* **45**, 691–697 (2016).
47. Church, D. M. et al. Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).
48. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv [q-bio.GN]* (2013).
49. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. Karczewski, K. J. et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).
51. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
52. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85**, 87–94 (1922).
53. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. Preprint at *bioRxiv* https://doi.org/10.1101/372896 (2018).
54. Flicek, P. et al. Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
55. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
56. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
57. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
58. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinf.* **17**, 483 (2016).
59. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* **11**, https://doi.org/10.1515/1544-6115.1826 (2012).
60. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
61. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
62. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
63. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).
64. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
65. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 1–25 (2004).
66. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

Acknowledgements

We thank D. Jörg for highly constructive discussions and P. Qiao for valuable comments on the manuscript. We acknowledge the Wellcome Sanger Institute Cellular Genetics and Phenotyping teams (in particular, A. Alderton, C. Gomez, R. Boyd, S. Patel and S. Barnett) and DNA pipelines for their invaluable assistance in generating the data for this study. We thank G. Kildisiute for assisting in CNV analysis of the fibroblast lines. This project was supported by Wellcome Sanger core funding (WT206194) and the Human Induced Pluripotent Stem Cell Initiative. Research in the Stegle laboratory is supported by the BMBF, the Volkswagen Foundation, the Chan Zuckerberg Initiative and the EU (ERC project DECODE, grant agreement 732546). D.J.M. is supported by the National Health and Medical Research Council of Australia (grants APP1112681 and APP1162829), seed funding from the Baker Foundation and the Holyoake Research Fellowship at St Vincent's Institute of Medical Research and the University of Melbourne. R.R. is supported the BBSRC Doctoral Training Programme. Y.H. is supported by the University of Cambridge and EMBL-EBI through an EBPOD postdoctoral fellowship. D.J.K. is supported by the Wellcome Trust under grants 203828/Z/16/A and 203828/Z/16/Z. T.H. is supported by a Human Frontier Science Program Fellowship, an EMBO Long-term Fellowship and an EMBO Advanced Fellowship.

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 fibroblast scRNA-sequencing data. R.L. and D.J.M. processed the melanoma scRNA-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 DE 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 Supplementary Note for a full list of consortium members).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0766-3.

Correspondence and requests for materials should be addressed to O.S. or S.A.T.

Peer review information Nicole Rusk and Lin Tang were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Sarah A. Teichmann

Last updated by author(s): 2019-12-23

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

samtools(v1.5,v1.8) TrimGalore(v 0.4.5) Salmon (v0.8.2) STAR(v2.6.0b) picard (v2.18.4) GATK(3.8) fastp(v0.20.0)

Data analysis

R (v3.6.0)
Snakemake (v5.5.0)
edgeR (v3.26.8)
limma (v3.40.6)
Cardelino (v0.6.4)
Demuxlet (no version information)
Single-Cell Genotyper(SCG) (v0.3.1)
SubClonalSelection (v0.0.0)
Canopy (1.3.0)
bcftools (v1.7,v1.8)
Scater (v1.12.2)
Scran (v1.12.1)
sctransform (v0.2.0)
variancePartition (v1.14.1)
IHW (v1.12.0)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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; identifiers for the lines used are: euts, fawm, feec, fikt, garx, gesg, heja, hipn, ieki, joxm, kuco, laey, lexy, naju, nusw, oaaz, oilg, pipw, puie, qayj, qolg, qonc, rozh, sehl, ualf, vass, vils, vuna, wahn, wetu, xugn, zoxy). Metadata, processed data anc large results files are available under the DOI 10.5281/zenodo.1403510 (doi.org/10.5281/zenodo.1403510).

Figures 3, 4, & 5 have associated raw data. Figure 1 uses simulated data. Figure 2 uses published data from Tirosh et al (2016).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Sample size was limited by the number of available HipSci lines for which whole-exome sequence data (at least 10 somatic mutations identified) and single-cell RNA-seq data (for at least 15 QC-passing cells) was available (Methods), and all available samples were used.
Data exclusions	No data were excluded from the study beyond the quality control and data processing steps described in the Methods.
Replication	Due to limited material available from sampled cell populations no experimental replication was undertaken.
Randomization	This is not relevant as there were no predefined experimental groups in this study.
Blinding	Blinding was not possible for this study and not relevant since we did not attempt to find associations between sample groups (there were none) and sample phenotypes or other characteristics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Fibroblasts were derived from healthy adults of European ancestry from the Hipsci consortium. See http://www.hipsci.org/ for more details. Specifically, the following HipSci lines were used: euts, fawm, feec, fikt, garx, gesg, heja, hipn, ieki, joxm, kuco, laey, lexy, naju, nusw, oaaz, oilg, pipw, puie, qayj, qolg, qonc, rozh, sehl, ualf, vass, vils, vuna, wahn, wetu, xugn, zoxy.
Authentication	No authentication was made except for visual morphology scan.
Mycoplasma contamination	Cell were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None.