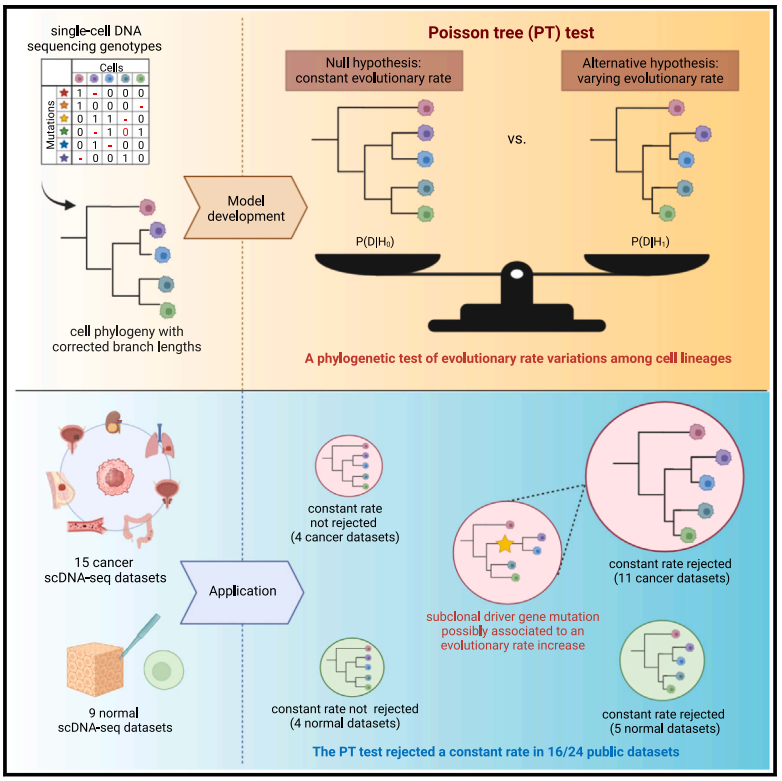
Cell Genomics

Article

Single-cell phylogenies reveal changes in the evolutionary rate within cancer and healthy tissues

Graphical abstract



Highlights

- Phylogenetic test for varying evolutionary rates in scDNA-seq data
- Constant rate rejection in most cancer and half of the healthy datasets analyzed
- Mutation in driver genes in cancer datasets could explain the rate acceleration

Authors

Nico Borgsmüller, Monica Valecha, Jack Kuipers, Niko Beerenwinkel, David Posada

Correspondence

niko.beerenwinkel@bsse.ethz.ch (N.B.), dposada@uvigo.es (D.P.)

In brief

Borgsmüller et al. develop a phylogenetic test of evolutionary rate variation among cell lineages using scDNA-seq data. In several normal and most cancer scDNA-seq datasets analyzed, they reject a constant evolutionary rate and identify mutations in driver genes that could explain the rate acceleration in tumor tissues.

Borgsmüller et al., 2023, Cell Genomics *3*, 100380
September 13, 2023
https://doi.org/10.1016/j.xgen.2023.100380

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Single-cell phylogenies reveal changes in the evolutionary rate within cancer and healthy tissues

Nico Borgsmüller,[1,2,6] Monica Valecha,[3,4,6] Jack Kuipers,[1,2] Niko Beerenwinkel,[1,2,*] and David Posada[3,4,5,7,*]
[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland
[2]SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland
[3]CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[4]Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain
[5]Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, 36310 Vigo, Spain
[6]These authors contributed equally
[7]Lead contact
*Correspondence: niko.beerenwinkel@bsse.ethz.ch (N.B.), dposada@uvigo.es (D.P.)
https://doi.org/10.1016/j.xgen.2023.100380

SUMMARY

Cell lineages accumulate somatic mutations during organismal development, potentially leading to pathological states. The rate of somatic evolution within a cell population can vary due to multiple factors, including selection, a change in the mutation rate, or differences in the microenvironment. Here, we developed a statistical test called the Poisson Tree (PT) test to detect varying evolutionary rates among cell lineages, leveraging the phylogenetic signal of single-cell DNA sequencing (scDNA-seq) data. We applied the PT test to 24 healthy and cancer samples, rejecting a constant evolutionary rate in 11 out of 15 cancer and five out of nine healthy scDNA-seq datasets. In six cancer datasets, we identified subclonal mutations in known driver genes that could explain the rate accelerations of particular cancer lineages. Our findings demonstrate the efficacy of scDNA-seq for studying somatic evolution and suggest that cell lineages often evolve at different rates within cancer and healthy tissues.

INTRODUCTION

Somatic evolution is the process by which cell populations accumulate genetic and epigenetic mutations during the lifetime of a multicellular organism. Recent technological advances have enabled the study of individual cell lineages, providing a more detailed picture of how this process unfolds, particularly in humans.[1–3] Understanding somatic evolution has clear implications regarding development, aging, and disease.[4–6]

Cancer is one of the most prominent examples of somatic evolution. It has long been recognized as a Darwinian process in which different cell lineages compete for space and resources, and only the fittest lineages survive, eventually leading to clonal cell expansions and tumor progression.[7–9] In recent years, analyses of genomic data from tumor cohorts have demonstrated that cancer progression is driven primarily by mutations in particular genes (often referred to as ''driver genes'') that provide specific cell lineages with a selective advantage.[10–16] However, the predominant role of selection after malignant transformation is under debate.[17,18] Several authors have proposed that, once established, some tumors might evolve neutrally, accumulating mutations that do not alter cell fitness.[19–21] Williams et al.[22,23] proposed a test of neutral evolution for bulk tumor sequencing data based on the observed distribution of the variant allele frequencies (VAFs) (see also Tung and Durrett[24]). When they applied their test to genomic data from 14 tumor types, it failed to reject neutral evolution in one-third of the 904 datasets analyzed.[22,25] Several studies have questioned these findings,[26–29] suggesting that the proportion of tumors that evolve neutrally is smaller and that selection is the main driver of tumor progression.

The somatic evolutionary rate is the number of mutations per time unit acquired by a cell lineage and results from the mutation rate per cell division times the cell division rate per time unit. If the evolution of a cell population is neutral, different cell lineages will evolve at the same rate, accumulating mutations at a constant pace, as a ''molecular clock.''[30,31] On the other hand, if selection is acting on a cell population, then the evolutionary rates among cell lineages will differ.[32,33] Once the fittest lineage has outcompeted the other lineages, the evolutionary rate within the cell population will be effectively neutral again until the next selective event.

In organismal evolutionary biology, deviations from the molecular clock are often interpreted as signals of selection,[34–36] but under the implicit assumption that the (germline) mutation rate is constant. However, the mutation rate in cancer cells can increase during tumor progression, for example, due to genetic alterations in DNA repair pathways.[37–42] A change in the somatic

mutation rate of particular cell lineages will lead to varying evolutionary rates in the cell population, even under neutral evolution, possibly leading to false rejections of neutrality by the VAF-based tests.[22,26] Still, the interplay between somatic mutation rates and selection is complex, and changes in the mutation rate might, in some cases, result from selection favoring higher mutability in some genes.[37,43]

Identifying evolutionary rate changes in cell populations is therefore critical to understand somatic evolution in cancer and healthy tissues. However, assessing evolutionary rate variation from bulk tissue samples is challenging, as millions of cells are sequenced simultaneously. Cell lineages are mixed in this case, and their deconvolution is complex and error-prone.[44] In contrast, single-cell DNA sequencing (scDNA-seq) provides immediate information on the genotypes of individual cells. From scDNA-seq data, it is possible to infer cell phylogenies in which branch lengths represent the evolutionary rates of the cell lineages.[45–48] However, scDNA-seq data suffer from technical errors like amplification errors and allele dropout that can result in spurious mutation calls (false negatives and false positives),[49] potentially biasing the estimation of the branch lengths in the cell phylogeny.

Here, we introduce a test of evolutionary rate variation among cell lineages based on scDNA-seq data. The Poisson Tree (PT) test uses single-nucleotide variant (SNV) calls to implement a phylogenetic likelihood ratio test that accounts for the technical errors in scDNA-seq data. The null hypothesis of the PT test is that the evolutionary rate of the sampled cell lineages is constant. Therefore, rejecting the null hypothesis might point to ongoing selection or changes in the mutation rate within the sampled cell population.

Using simulated data, we show that the PT test can identify rate variation among cell lineages while being robust to scDNA-seq noise. We applied the PT test to 24 scDNA-seq datasets from 15 cancer and nine healthy tissue samples, identifying rate variation among cell lineages in 11 cancer and five healthy datasets. In six cancer datasets with significant evolutionary rate variation, we identified mutations in known driver genes on internal branches of the phylogenetic trees, suggesting that selection could explain the acceleration of the evolutionary rate of particular cell lineages.

RESULTS

A PT test of evolutionary rate variation

To carry out the PT test, it is necessary to specify as input a matrix of SNVs, a phylogeny of contemporaneously sampled cells, and scDNA-seq false positive and false negative genotype error rates (Figures 1A–1C). We first map all SNVs to the branches of the cell phylogeny and weigh the branches with the probability of missing true SNVs due to scDNA-seq errors (Figure 1D). Assuming that the number of SNVs per branch follows a Poisson distribution, we estimate the likelihood of two competing models, one with a constant evolutionary rate for all branches (null hypothesis; Figure 1E) and the other with varying evolutionary rates among branches (alternative hypothesis; Figure 1F). Under the constant-rate model, branch lengths are constrained, while they are independent under the varying-rate model. Finally, we compare the two models with a likelihood ratio test (LRT) (Figure 1G). See the STAR Methods for details regarding the PT test.

The PT test detects evolutionary rate variation reliably

To evaluate the performance of the PT test, we used CellCoal[50] to simulate scDNA-seq data with constant or varying evolutionary rates, with and without scDNA-seq errors. In the simulations with scDNA-seq errors, we varied the false negative rate from 2.5% to 30% while maintaining a fixed false positive rate of 1% or varied the false positive rate from 0.1% to 2% (see Huang et al.[51]) while maintaining a fixed false negative rate of 10%. All simulation runs generated samples of 30 cells and were repeated 1,000 times. See the STAR Methods for details regarding the simulation conditions.

In the absence of rate variation among cell lineages and without scDNA-seq errors, the distribution of pvalues of the PT test was uniform under the null hypothesis, as expected for an unbiased test (Figure 2A, first panel). Without evolutionary rate variation, but with scDNA-seq errors, p values were strongly shifted toward 1, making the PT test conservative (Figure 2A, second to fourth panels). With a false negative rate of 30%, low p values became more common, indicating that the test may not distinguish between high scDNA-seq error rates and changes in the evolutionary rate (Figure S1). On the other hand, even a high false positive rate of 2% did not bias the PT test toward low p values (Figure S2A). In all cases, the difference in the p value distributions between using the inferred cell phylogeny and scDNA-seq error rates by CellPhy[47] (blue) or the true phylogeny and scDNA-seq error rates (red) was marginal. Likewise, p value distributions did not change when using infSCITE[52] instead of CellPhy to infer the cell tree and the scDNA-seq error rates (Figure S1).

For comparison, we also applied the molecular clock LRT implemented in PAUP*[53,54] and the Poisson dispersion test.[55] The former is typically used in organismal phylogenetics to test for a constant evolutionary rate among lineages and assumes error-free data. The latter tests if the number of SNVs per cell is sampled from a Poisson distribution, ignoring the underlying cell phylogeny. In our simulations, the PAUP* LRT was biased toward low p values, even when using the true cell phylogeny and without scDNA-seq errors (Figure 2B, orange). The p values of the Poisson dispersion test were biased toward 1 in the absence of scDNA-seq errors (Figure 2B, green) but became biased toward 0 otherwise, resulting in a high number of false rejections of the null hypothesis. We concluded that both tests are unsuited for detecting changes in the evolutionary rate among cell lineages from scDNA-seq data.

To assess the power of the PT test, we simulated evolutionary rate variation by introducing changes in the evolutionary rate of a given cell lineage. We chose an internal branch with probability proportional to its length and increased its length and that of all descendant branches by 2×, 5×, or 10×. To explore the effect of the sample size, we simulated 100 cells and subsampled 10, 30, 50, 70, and 90 cells, excluding replicates without cells affected by the rate change. As expected, the power of the PT test increased with more drastic evolutionary rate changes and larger sample sizes (Figures 2C, and S2B). Without scDNA-seq errors,

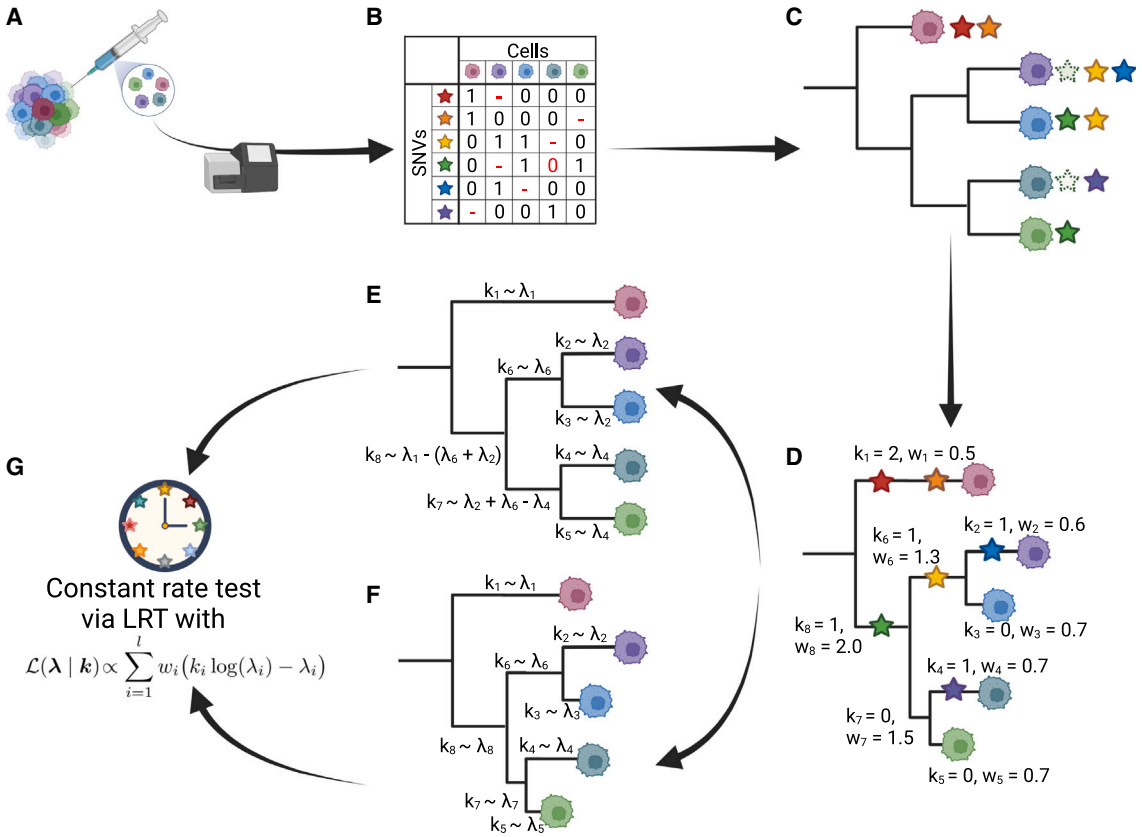


Figure 1. Overview of the PT test of evolutionary rate variation

(A) Single cells are isolated from a tissue, and their genome is amplified and sequenced.

(B and C) Based on the sequencing reads, single-nucleotide variants (SNVs) are called (B) and used to infer a cell phylogeny (C).

(D) SNVs are mapped onto the branches of the cell phylogeny, specifying their length k, and branch weights w are determined based on how likely an SNV on each branch might be missed due to single-cell DNA sequencing errors. Branch lengths k are modeled by a Poisson distribution with rate parameter λ, where λ represents the total (genome-wide) number of SNVs that occur in a branch.

(E) Under the null model, the evolutionary rate is constant, implying that the cumulative branch length from the root to any cell is expected to be similar, and the rate parameters λ are constrained accordingly.

(F) Under the alternative model, branch lengths are independent and, therefore, can be variable.

(G) The likelihood of the data under the null and the alternative model is computed and compared with a likelihood-ratio test (LRT).

the power of the PT test was 92%–100% already for 2× rate changes. With scDNA-seq errors, the power of the PT was above 90% for 5× and 10× rate changes and for samples with more than 10 cells. For the 2× rate change, the power dropped below 50%, especially for small sample sizes and high error rates. Overall, we conclude that the PT test can reliably detect changes in the evolutionary rates among cell lineages using scDNA-seq data.

VAF-based selection tests detect evolutionary rate changes poorly

As selection is one of the main causes for changes in the evolutionary rate among cell lineages, we also explored customized bulk approaches for comparing neutral and adaptive evolution, specifically the 1/*f* test[22] and Mobster.[23] Both approaches try to identify lineages with a growth advantage based on the VAF distribution.

We simulated bulk data at 100× sequencing depth without scDNA-seq errors using the same evolutionary rates as for the single-cell data (250 repetitions each). Under a constant evolutionary rate, the p value distribution of the 1/*f* test was biased toward 0, resulting in wrong rejections of neutrality in 20% of the simulations. With a 2×–10× increase in the evolutionary rate, the 1/*f* test rejected neutrality in nearly 35% of the cases (Figures 2C, first panel, yellow, and S3A). Mobster wrongly inferred subclonal selection in 73% of the simulations with a constant evolutionary rate and reported subclonal selection in up to 83% of the cases with a rate increase (Figure S3B). A possible explanation for this finding could be that the extent of rate variation we simulated, or the number of affected cells, was not large enough for the VAF-based tests to reject neutrality.

The PT test infers evolutionary rate variation in scDNA-seq data

We applied the PT test to 24 scDNA-seq datasets (12 whole genome and 12 whole exome) from 16 patients containing between 7 and 71 cells (Table 1). Fifteen datasets were derived

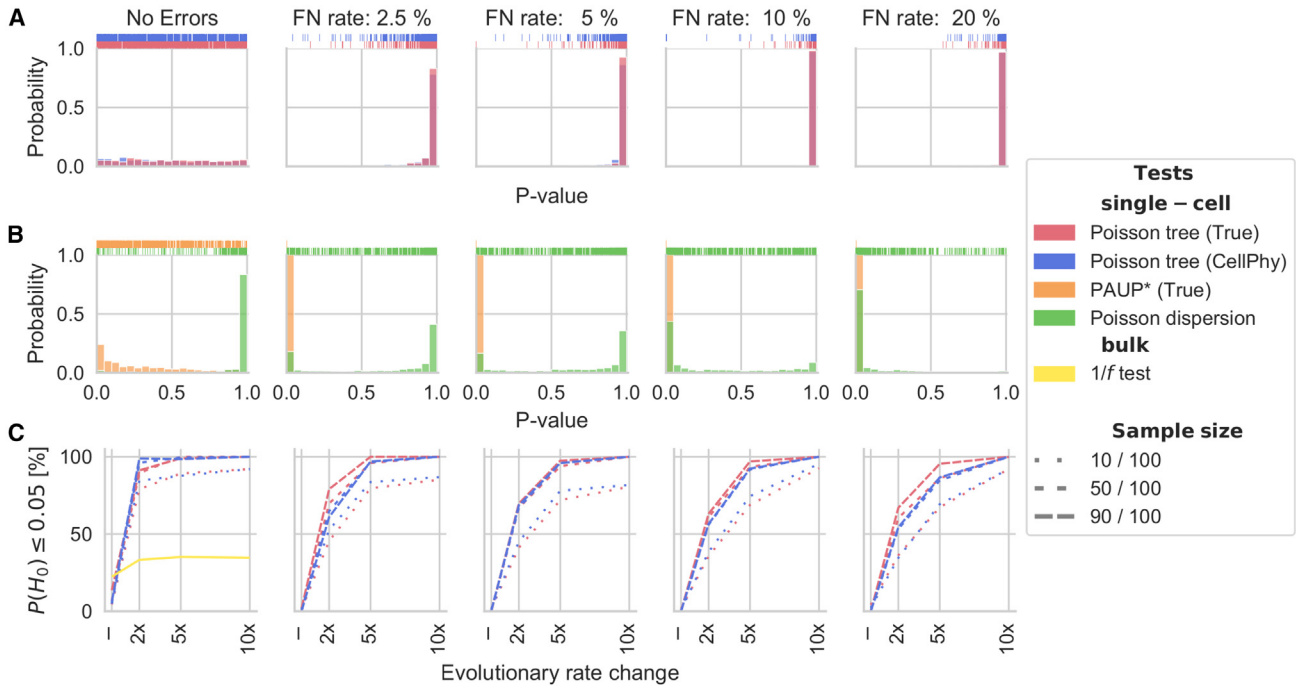


Figure 2. The PT test detects evolutionary rate variation reliably

(A) p value distribution of the PT test under a constant evolutionary rate (null hypothesis) for different scDNA-seq false negative (FN) rates, using the true (red) or inferred (blue) cell phylogeny and the scDNA-seq error rates. The rug plots above each panel display the p values for each replicate (n = 1,000).

(B) p value distribution of PAUP*'s LRT (orange) and the Poisson dispersion test (green) under a constant evolutionary rate for different scDNA-seq FN rates. The rug plots above each panel display the p values for each replicate (n = 1,000).

(C) Statistical power of the PT test for detecting variable evolutionary rates. Rate changes are introduced by increasing the rate for a given lineage by 2×, 5×, or 10×. Distinct line styles represent different sample sizes (total: 100 cells). In the left panel, the yellow line represents the proportion of bulk datasets in which the 1/*f* test rejected neutrality.

from cancer tissues (blood, bladder, lung, prostate, breast, colorectal [CRC], and renal cancers), and nine were from normal, healthy tissues. Additionally, all datasets contained a bulk normal sample, and all but four cancer datasets contained a bulk tumor sample. Tables S1 and S3 describe these datasets in detail. We ran the PT test using the cell phylogenies and scDNA-seq error rates inferred by CellPhy. Then, we mapped the SNVs to specific branches (see the STAR Methods for further details) and identified cancer-specific driver SNVs using IntOGen.[15]

Out of the 24 scDNA-seq datasets, we rejected a constant evolutionary rate in 11 out of 15 cancer and five out of nine normal datasets (Figure 3). We found no relationship between the PT test results and the number of SNVs, cells, or inferred scDNA-seq false negative rates (Figure S4). If changes in the evolutionary rate are driven by selection, we might be able to locate an SNV in a driver gene in one of the internal branches of the cell phylogenies. Early driver events, i.e., those mapped to the trunk of the phylogenetic tree (the branch representing the ancestral lineage to all sampled cancer cells), will likely be involved in tumor initiation or complete selective sweeps and therefore do not result in rate changes. Later driver events, mapped to internal branches, will result in subclonal selection and, therefore, in heterogeneous evolutionary rates. As known cancer driver events also occur in healthy tissues,[56–58] we explored their existence in our normal datasets as a possible explanation for changes in the evolutionary rate.

In four normal datasets (Lodato-P2-N, Wang-ER+-N, Wu-CRC0827-N, and Wu-CRC0907-P), the PT test did not reject a constant rate (Figure 3A). In Wang-ER+-N, we did not detect driver mutations, while in Lodato-P2-N, all known driver mutations were placed on the trunk. In the polyp dataset Wu-CRC0907-P, we detected an activating SNV in the oncogene *BRAF*, which was not reported in the original study. *BRAF* activation is a known early event in CRC tumor initiation,[59] indicating that the polyp might have been malignant already despite being classified as normal based on histopathology. In Wu-CRC0827-N, we inferred a *PARP4* SNV on an internal branch. *PARP4*'s mode of action is labeled as "ambiguous" in IntOGen, and it is not listed as a driver gene in the Cancer Gene Census.[60]

The PT test rejected a constant rate in the remaining five normal datasets (Figure 3B). In Kang-N, we found a known cancer driver SNV on an internal branch (present in 9/14 cells), namely an activation of *NCOR2*. For Lodato-P1-N, Lodato-P3-N, and Li-N, we inferred fully ladder-like trees with limited bootstrap support. Therefore, these results should be interpreted with caution. In Wang-TNBC-N, only 68 SNVs were called, out of which 37 were mapped to the trunk. However, despite the low number of SNVs at internal branches, the PT test did reject a constant evolutionary rate for this dataset. In the four cancer

Table 1. Poisson Tree (PT) test results and called SNVs in cancer-specific driver genes for scDNA-seq datasets

Dataset	Tissue	Subset	N cell	N SNVs	FN rate (%)	PT test (pvalue)	Driver SNVs	
							Trunk	Internal br.
Li-C[56]	bladder	cancer	54	885	11	0.005[a]	*SF3B1*	*ATM*
Li-N	bladder	healthy	8	644	17	$<1 \times 10^{-6}$[a]	–	–
Hou-C[57]	blood	cancer	71	1,387	9	0.082	*ATM*, *PRKD2*	–
Wang-ER+-C[58]	breast	cancer	46	355	4	0.999	*PIK3CA*, *MAP3K1*	–
Wang-ER+-N	breast	healthy	12	300	12	0.231	–	–
Wang-TNBC-C[58]	breast	cancer	15	1,472	10	$<1 \times 10^{-6}$[a]	*SPEN*, *NOTCH2*, *NTRK1*, *ZFHX3*	*ARID1B*, *SMAD4*, *ERBB4*, *GNAS*
Wang-TNBC-N	breast	healthy	15	68	3	$<1 \times 10^{-6}$[a]	–	–
Alves-L-C[59]	colon	cancer	22	5,089	2	$<1 \times 10^{-6}$[a]	*APC*	*SOX9*
Alves-LR-C[59]	colon	cancer	30	6,850	4	$<1 \times 10^{-6}$[a]	*APC*, *SOX9*, *MYH9*	–
Kang-C[49]	colon	cancer	30	2,645	6	0.131	*NCOR2*, *CARD11*	–
Kang-N	colon	healthy	9	769	7	$<4 \times 10^{-5}$[a]	–	*NCOR2*
Kozlov-C[48]	colon	cancer	23	3,503	3	$<1 \times 10^{-6}$[a]	*NRAS*	–
Wu-CRC0827-C[60]	colon	cancer	50	652	9	$<1 \times 10^{-6}$[a]	–	*PARP4*, *NBEA*, *TP53*, *FAT4*, *TBX3*
Wu-CRC0827-P	colon	cancer	19	379	10	$<2 \times 10^{-4}$[a]	–	*PARP4*
Wu-CRC0827-N	colon	healthy	15	298	10	0.491	–	*PARP4*
Wu-CRC0907-C[60]	colon	cancer	49	574	10	$<1 \times 10^{-6}$[a]	–	*SMARCA4*, *APC*, *GNAS*, *ARID1A*
Wu-CRC0907-P	colon	healthy	25	181	4	0.336	*SMARCA4*	*BRAF*
Xu-C[61]	kidney	cancer	20	747	4	0.158	–	–
Ni-C[62]	lung	cancer	8	340	20	$<1 \times 10^{-6}$[a]	*PIK3CA*, *RB1*, *TP53*	*SETD2*
Lodato-P1-N[63]	neurons	healthy	10	935	5	$<4 \times 10^{-4}$[a]	–	–
Lodato-P2-N[63]	neurons	healthy	15	747	4	0.128	*ZNRF3*	–
Lodato-P3-N[63]	neurons	healthy	8	928	9	$<1 \times 10^{-4}$[a]	*TET2*	–
Su-P1-C[64]	prostate	cancer	7	23,130	14	$<1 \times 10^{-6}$[a]	–	–
Su-P2-C[64]	prostate	cancer	8	15,394	4	$<1 \times 10^{-6}$[a]	–	–

N, number of; SNVs, single-nucleotide variants; FN, false negative; br, branch.
[a]p values are below a significance level of 0.05 for the PT test.

datasets Hou-C, Wang-ER+-C, Kang-C, and Xu-C, a constant rate was not rejected (Figure 3C). We did not identify driver SNVs on internal branches in any of them. In three datasets (Hou-C, Wang-ER+-C, and Kang-C), however, we identified drivers on the trunk, possibly being involved in tumor initiation or past selective sweeps. For the remaining 11 cancer datasets, the PT test rejected a constant rate (Figure 3D). In seven of these, we identified at least one known driver SNV on an internal branch of the tree. In the remaining four, we identified either no driver (Su-P1-C and Su-P2-C) or only drivers on the trunk (Alves-LR-C and Kozlov-C).

Overall, the PT test rejected a constant rate in the majority of normal datasets, although only a few of them harbored known cancer driver mutations. In contrast, we identified SNVs in known driver genes in all but three cancer datasets. In cancer datasets where the PT test did not reject a constant rate, drivers were absent or placed on the trunk, whereas in seven out of 11 datasets where the PT test reported variable rates, a driver SNV was mapped to an internal branch, possibly explaining the change in evolutionary rates.

Additionally, we merged all the SNVs detected for each cell into a pseudo-bulk dataset and calculated the global non-synonymous over synonymous substitution rate ratios (dN/dS) for 369 cancer driver genes and genome-wide dN/dS ratios for all genes[11] (Table S1). In 11 datasets, out of which seven were derived from normal tissue, the dN/dS ratio for known drivers could not be computed, as no or just one SNV was located in a cancer driver gene. Where computable, the confidence intervals of the dN/dS ratios included 1 or smaller values (Hou-C and Su-P1-C). Genome-wide dN/dS values could be calculated for all datasets except Lodato-P3-N. Their confidence intervals included 1 or only values smaller than 1 (Li-N, Hou-C, Wang-ER+-C, Wang-ER+-N, Kozlov-C, Xu-C, and Su-P1-C) in all datasets, showing no evidence for positive selection.

Bulk selection tests produce ambiguous results on scDNA-seq data

To compare bulk and single-cell approaches on biological data, we applied the 1/*f* test and Mobster to 16 matched tumor bulk samples (Table S2). Here, we also calculated the global dN/dS

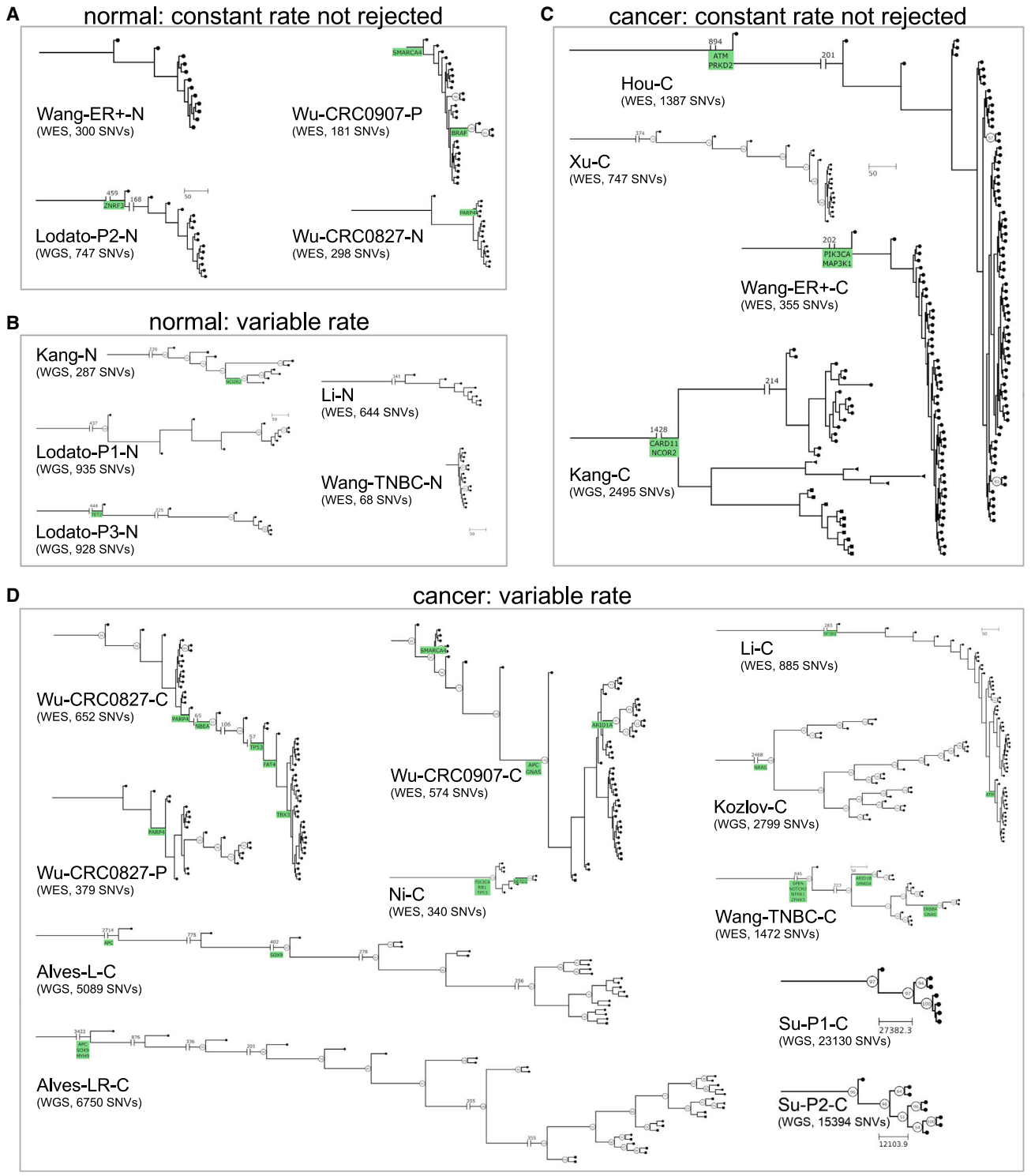


Figure 3. The PT test infers evolutionary rate variations in scDNA-seq data

(A) In four healthy tissue datasets, we detected no deviation from a constant evolutionary rate. SNVs in known driver genes were absent, located on the trunk, or present in four cells at most.

(B) In five healthy tissue datasets, we identified variable rates. Three of these showed a ladder-like pattern, meaning that each internal node is an ancestor to at least one leaf node (Lodato-P1, Lodato P3, and Li-N), and in the Wang-TNBC-N dataset, we called only 68 SNVs and mapped most to the trunk.

(legend continued on next page)

ratios for cancer driver genes and the genome-wide dN/dS ratios for all genes. The 1/*f* test rejected neutrality in six cases, including three datasets that the PT test did not (Kang-C, Wu-CRC0907-P, and Xu-C). On the other hand, the 1/*f* test did not reject neutrality in four datasets where the PT test found evidence for rate variation (Wu-CRC0827-C, Wang-TNBC-C, Ni-C, and Kozlov-C). These findings may be limited, as a sequencing depth above 100× and cellularity above 0.5, required for the 1/*f* test to be robust,[23] were only achieved in the Li-BC sample. Mobster only produced results for three datasets. In the bladder cancer bulk sample corresponding to Li-C, a clone with selective disadvantage ($s = -1.1$) was inferred, and no clones were inferred in the bulk samples corresponding to Wang-TNBC-C (PT test p value $< 1 \times 10^{-6}$) and Xu-C (PT test $p = 0.16$). The global dN/dS ratio confidence intervals for SNVs in the 369 driver genes included 1.0 in all cancer samples. When calculated for all genes, only the Li dataset showed evidence of positive selection (dN/dS: 1.1–2.7). The dN/dS confidence interval for all other datasets included 1.0 or only smaller values (Kang-BC, Hou-BC1, and Hou-BC2).

DISCUSSION

Somatic evolution plays an important role in multiple aspects of biology and medicine, including development, aging, and disease. Recent technological advances have enabled studying somatic evolution at the single-cell level, where different cell lineages might evolve at distinct rates due to various factors, including genetic and epigenetic mutations or changes in the microenvironment. In this work, we have introduced a phylogenetic test to detect evolutionary rate variation among cell lineages using scDNA-seq data. Our simulations suggest that the PT test is conservative, even in the presence of scDNA-seq noise, and that it is more powerful than existing tests assuming error-free data or using bulk data. When we applied the PT test to scDNA-seq data, we rejected a constant evolutionary rate in both healthy and cancer cell populations. In the latter, we identified potential driver mutations that might be involved in the rate acceleration of particular tumor lineages.

The observation of distinct tumor cell lineages evolving at different rates is expected under the prevalent cancer progression model, in which selection drives the expansion of the fittest clones.[7–9] Househam et al.[33] recently leveraged single-colorectal-gland phylogenies to distinguish between neutral evolution (constant evolutionary rate) and subclonal selection (varying evolutionary rates), finding putative driver mutations associated with the latter. In the six scDNA-seq datasets analyzed here, we also identified potential driver mutations that could explain the acceleration of particular cell lineages. Some of the mutated genes, like *BRAF* or *GNAS*, regulate cellular proliferation,[61,62] while others, e.g., *ATM* and *SETD2*, are related to changes in the mutation rate through their involvement in DNA damage repair pathways.[63,64] We also detected varying rates in half of the healthy tissue samples tested. Recent studies have shown that clonal expansions occur in many healthy tissues[56–58] and that some of these expansions might result from selective pressures[65,66] Therefore, finding evolutionary rate variation in some healthy populations can be expected, although the underlying mechanisms responsible for this variation are still largely unknown. At the same time, we could not reject a constant evolutionary rate in four healthy and four cancer datasets. There are different possible explanations for this result. If the evolutionary rate is truly constant, it might imply that these populations are evolving neutrally. Otherwise, failure of the test to detect true variation in the evolutionary rate might be due to minor differences in the evolutionary rates among the cell lineages or due to a limited sample size in terms of cells or mutations.

In recent years, multiple models have been proposed to explain the various evolutionary trajectories inferred from genomic data in most cancer types (e.g., Davis et al.,[17] Vendramin et al.,[18] and Williams et al.[67]). For example, a two-phased process has been proposed where rapid changes induced by genome-level alterations alternate with phases of gradual evolution with changes occurring at the gene level.[68,69] Clearly, large structural events at the chromosomal and genome levels will often affect the fitness and, therefore, modify the effective growth rate of the affected cell lineage. Importantly, changes in the growth rate, regardless of their origin, will also alter the rate of accumulation of SNVs.

We want to stress that while the PT test assesses differences in the evolutionary rates among cell lineages using single-cell SNVs, different causes, despite SNVs themselves, can change the rate of SNV accumulations. Any somatic events, like large or small structural variations, epigenetic modifications, or changes in the transcriptome,[70–72] can also alter the cell division or SNV mutation rate and, therefore, the evolutionary rate of particular cell lineages. Likewise, changes in the evolutionary rate can also result from environmental effects and do not necessarily need to have a genetic origin.[73] For example, spatial constraints,[74] cell dormancy,[75] or variations in the tumor microenvironment[76] might result in heterogeneous evolutionary rates.

In summary, we introduced a test for the homogeneity of evolutionary rates among cell lineages. Applying this test, we found that cell lineages often evolve at different rates within cancer and healthy populations and that mutations in driver genes could explain some of these differences in the case of cancer. Apart from helping pinpoint cell lineages of particular interest, testing rate homogeneity could validate tumor age estimates[77] or developmental inferences based on a molecular clock.[78–80] New methods combining concepts from evolutionary biology with advances in single-cell technologies, as showcased in this work, offer great potential to study the evolution of somatic tissues and the underlying mechanisms.

(C) In four cancer datasets, we did not reject a constant rate. We either identified no driver SNV (Xu-C) or mapped all known drivers to the trunk (Hou-C, Kang-C, and Wang-ER+-C).

(D) In 11 cancer datasets, we found evidence for variable evolutionary rates. In seven of these, we identified at least one known driver SNV on an internal branch (Wu-CRC0827-C, Wu-CRC0827-P, Alves-L-C, Wu-CRC0907-C, Ni-C, Li-C, Wang-TNBC-C). The leaf node shapes correspond with different spatial sampling locations. Bootstrap values above 50 are indicated.

Table 2. CellCoal simulation parameters

	Constant rate			Variable rates	
Scenario	No SNV errors	Varying FN errors	Varying FP errors	No SNV errors	Varying FN errors
Number of cells	30	30	30	100	100
Genome length	10,000	10,000	10,000	10,000	10,000
Mutation rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	8×10^{-7}, 5×10^{-7}, 3×10^{-7}	8×10^{-7}, 5×10^{-7}, 3×10^{-7}
Rate change factor	–	–	–	2×, 5×, 10×	2×, 5×, 10×
Number of cells subsampled	–	–	–	10, 30, 50, 70, 90	10, 30, 50, 70, 90
Sequencing depth mean (×)	20	20	20	20	20
Sequencing depth dispersion (×)	0	5	5	0	5
Sequencing error (%)	0	1	1	0	1
ADO per cell (mean) (%)	0	5, 10, 20, 40, 60	20	0	5, 10, 20, 40, 60
ADO per cell (variance) (%)	0	10	10	0	10
Amplification error (%)	0	1	0.1, 1, 2	0	1
Number of replicates	1,000	1,000	1,000	3,000	3,000

SNV, single-nucleotide variant; FN, false negative; FP, false positive; ADO, allelic dropout.

Limitations of the study

scDNA-seq is laborious and expensive, and the number of cells sequenced tends to be limited accordingly. However, if the sample size is small, cell lineages with different evolutionary rates might be missed. While bulk approaches sequence more cells than single-cell strategies, the analysis of bulk data relies on summary statistics, potentially precluding the detection of subtle changes in the evolutionary rate.[44] The VAF distribution, for example, depends highly on the number, type, and spatial location of biopsies taken.[81] Multiregional sampling should facilitate a more detailed exploration of the spatial rate heterogeneity, thus increasing the statistical power of the PT test. Nevertheless, despite the relatively small sample size of the scDNA-seq datasets analyzed here, we detected changes in the evolutionary rate in most of them, suggesting that changes in evolutionary rates within tissues might be common ground. In addition, although the PT test detects rate changes independent of their origin, drawing reliable conclusions about their cause would probably imply obtaining multiomic, spatial, and microenvironmental data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Poisson tree test model and input data
  - Null model: Constant evolutionary rate
  - Alternative model: Constraint-free evolutionary rates
  - Likelihood ratio test
  - Computation of the observed branch lengths
  - Branch weights
  - Implementation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Simulation of scDNA-seq data with constant and variable evolutionary rates
  - Inference of cell phylogenies
  - VAF-based tests of neutrality and subclonal selection for bulk data
  - dN/dS ratio estimation
  - Biological data processing

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100380.

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie CONTRA grant agreement no. 766030, the European Research Council (ERC) agreement nos. 617457 to D.P. and 609883 to N. Beerenwinkel, and the Spanish Ministry of Science and Innovation - MICINN (PID2019- 106247GB-I00 to D.P.). D.P. also receives support from Xunta de Galicia.

AUTHOR CONTRIBUTIONS

Conceptualization, N. Beerenwinkel, N. Borgsmüller, and D.P; methodology, N. Borgsmüller, J.K., and D.P.; software, N. Borgsmüller and M.V.; validation, N. Borgsmüller and M.V.; formal analysis, N. Borgsmüller; investigation, N. Borgsmüller and M.V.; resources, N. Beerenwinkel and D.P.; data curation, N. Borgsmüller and M.V.; writing – original draft, N. Borgsmüller and D.P.; writing – review & editing, N. Beerenwinkel, N. Borgsmüller, J.K., D.P., and M.V.; visualization, N. Borgsmüller; supervision, N. Beerenwinkel and D.P.; funding acquisition, N. Beerenwinkel and D.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research.

Received: October 8, 2022
Revised: May 3, 2023
Accepted: July 18, 2023
Published: August 17, 2023

REFERENCES

1. Zhang, L., Dong, X., Lee, M., Maslov, A.Y., Wang, T., and Vijg, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. Proc. Natl. Acad. Sci. USA *116*, 9014–9019. https://doi.org/10.1073/pnas.1902510116.
2. Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y., et al. (2021). Somatic mutation landscapes at single-molecule resolution. Nature *593*, 405–410. https://doi.org/10.1038/s41586-021-03477-4.
3. Yang, D., Jones, M.G., Naranjo, S., Rideout, W.M., 3rd, Min, K.H.J., Ho, R., Wu, W., Replogle, J.M., Page, J.L., Quinn, J.J., et al. (2022). Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. Cell *185*, 1905–1923.e25. https://doi.org/10.1016/j.cell.2022.04.015.
4. Marioni, J.C., and Arendt, D. (2017). How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. Annu. Rev. Cell Dev. Biol. *33*, 537–553. https://doi.org/10.1146/annurev-cellbio-100616-060818.
5. Wiedmeier, J.E., Noel, P., Lin, W., Von Hoff, D.D., and Han, H. (2019). Single-Cell Sequencing in Precision Medicine. In Precision Medicine in Cancer Therapy (Springer International Publishing), pp. 237–252.
6. Stadler, T., Pybus, O.G., and Stumpf, M.P.H. (2021). Phylodynamics for cell biologists. Science *371*, eaah6266. https://doi.org/10.1126/science.aah6266.
7. Nowell, P.C. (1976). The Clonal Evolution of Tumor Cell Populations. Science *194*, 23–28. https://doi.org/10.1126/science.959840.
8. Merlo, L.M.F., Pepper, J.W., Reid, B.J., and Maley, C.C. (2006). Cancer as an evolutionary and ecological process. Nat. Rev. Cancer *6*, 924–935. https://doi.org/10.1038/nrc2013.
9. Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313. https://doi.org/10.1038/nature10762.
10. Ostrow, S.L., Barshir, R., DeGregori, J., Yeger-Lotem, E., and Hershberg, R. (2014). Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes. PLoS Genet. *10*, 10042399–e1004311. https://doi.org/10.1371/journal.pgen.1004239.
11. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.
12. Bakhoum, S.F., and Landau, D.A. (2017). Chromosomal Instability as a Driver of Tumor Heterogeneity and Evolution. Cold Spring Harb. Perspect. Med. *7*, a029611. https://doi.org/10.1101/cshperspect.a029611.
13. Cannataro, V.L., Gaffney, S.G., Townsend, J.P., and Townsend, J.P. (2018). Effect Sizes of Somatic Mutations in Cancer. J. Natl. Cancer Inst. *110*, 1171–1177. https://doi.org/10.1093/jnci/djy168.
14. Frankell, A.M., Jammula, S., Li, X., Contino, G., Killcoyne, S., Abbas, S., Perner, J., Bower, L., Devonshire, G., Ococks, E., et al. (2019). The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. Nat. Genet. *51*, 506–516. https://doi.org/10.1038/s41588-018-0331-5.
15. Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. Nat. Rev. Cancer *20*, 555–572. https://doi.org/10.1038/s41568-020-0290-x.
16. Boström, M., and Larsson, E. (2022). Somatic mutation distribution across tumour cohorts provides a signal for positive selection in cancer. Nat. Commun. *13*, 7023. https://doi.org/10.1038/s41467-022- 34746-z.
17. Davis, A., Gao, R., and Navin, N. (2017). Tumor evolution: Linear, branching, neutral or punctuated? Biochimica et biophysica acta. Reviews on cancer *1867*, 151–161. https://doi.org/10.1016/j.bbcan. 2017.01.003.
18. Vendramin, R., Litchfield, K., and Swanton, C. (2021). Cancer evolution: Darwin and beyond. EMBO J. *40*, e108389. https://doi.org/10.15252/embj.2021108389.
19. Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., and Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. Nat. Genet. *47*, 209–216. https://doi.org/10.1038/ng.3214.
20. Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non Darwinian cell evolution. Proc. Natl. Acad. Sci. USA *112*, E6496–E6505. https://doi.org/10.1073/pnas.1519556112.
21. Sun, R., Hu, Z., Sottoriva, A., Graham, T.A., Harpak, A., Ma, Z., Fischer, J.M., Shibata, D., and Curtis, C. (2017). Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nat. Genet. *49*, 1015–1024. https://doi.org/10.1038/ng.3891.
22. Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. Nat. Genet. *48*, 238–244. https://doi.org/10.1038/ng.3489.
23. Williams, M.J., Werner, B., Heide, T., Curtis, C., Barnes, C.P., Sottoriva, A., and Graham, T.A. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. Nat. Genet. *50*, 895–903. https://doi.org/10.1038/s41588-018-0128-6.
24. Tung, H.-R., and Durrett, R. (2021). Signatures of neutral evolution in exponentially growing tumors: A theoretical perspective. PLoS Comput. Biol. *17*, e1008701–e1008712. https://doi.org/10.1371/journal.pcbi.1008701.
25. Heide, T., Zapata, L., Williams, M.J., Werner, B., Caravagna, G., Barnes, C.P., Graham, T.A., and Sottoriva, A. (2018). Reply to 'Neutral tumor evolution? Nat. Genet. *50*, 1633–1637. https://doi.org/10.1038/s41588-018-0256-z.
26. Tarabichi, M., Martincorena, I., Gerstung, M., Leroi, A.M., Markowetz, F., PCAWG Evolution and Heterogeneity Working Group; Spellman, P.T., Morris, Q.D., Lingjærde, O.C., Wedge, D.C., and Van Loo, P. (2018). Neutral tumor evolution? Nat. Genet. *50*, 1630–1633. https://doi.org/10.1038/s41588-018-0258-x.
27. McDonald, T.O., Chakrabarti, S., and Michor, F. (2018). Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. Nat. Genet. *50*, 1620–1623. https://doi.org/10.1038/s41588-018-0217-6.
28. Balaparya, A., and De, S. (2018). Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. Nat. Genet. *50*, 1626–1628. https://doi.org/10.1038/s41588-018-0219- 4.
29. Bozic, I., Paterson, C., and Waclaw, B. (2019). On measuring selection in cancer from subclonal mutation frequencies. PLoS Comput. Biol. *15*, 10073688–e1007415. https://doi.org/10.1371/journal.pcbi.1007368.
30. Zuckerkandl, E., and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. In Evolving Genes and Proteins. Ed. by Bryson,

Vernon and Vogel (Henry J. Academic Press), pp. 97–166. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.

31. Kimura, M. (1968). Evolutionary Rate at the Molecular Level. Nature *217*, 624–626. https://doi.org/10.1038/217624a0.
32. Niida, A., Iwasaki, W.M., and Innan, H. (2018). Neutral Theory in Cancer Cell Population Genetics. Mol. Biol. Evol. *35*, 1316–1321. https://doi.org/10.1093/molbev/msy091.
33. Househam, J., Heide, T., Cresswell, G.D., Spiteri, I., Kimberley, C., Zapata, L., Lynn, C., James, C., Mossner, M., Fernandez-Mateos, J., et al. (2022). Phenotypic plasticity and genetic control in colorectal cancer evolution. Nature *611*, 744–753. https://doi.org/10.1038/s41586-022-05311-x.
34. Edwards, S.V. (2009). Natural selection and phylogenetic analysis. Proc. Natl. Acad. Sci. USA *106*, 8799–8800. https://doi.org/10.1073/pnas.0904103106.
35. Pertea, M., Pertea, G.M., and Salzberg, S.L. (2011). Detection of lineage-specific evolutionary changes among primate species. BMC Bioinf. *12*, 274. https://doi.org/10.1186/1471-2105-12-274.
36. Hedge, J., and Wilson, D.J. (2016). Practical Approaches for Detecting Selection in Microbial Genomes. PLoS Comput. Biol. *12*, 10047399–e1004812. https://doi.org/10.1371/journal.pcbi.1004739.
37. S Datta, R., Gutteridge, A., Swanton, C., Maley, C.C., and Graham, T.A. (2013). Modelling the evolution of genetic instability during tumour progression. Evol. Appl. *6*, 20–33. https://doi.org/10.1111/eva.12024.
38. Huang, S. (2013). Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. Cancer Metastasis Rev. *32*, 423–448. https://doi.org/10.1007/s10555-013-9435-7.
39. Asatryan, A.D., and Komarova, N.L. (2016). Evolution of genetic instability in heterogeneous tumors. J. Theor. Biol. *396*, 1–12. https://doi.org/10.1016/j.jtbi.2015.11.028.
40. Aguadé-Gorgorió, G., and Solé, R. (2018). Adaptive dynamics of unstable cancer populations:The canonical equation. Evol. Appl. *11*, 1283–1292. https://doi.org/10.1111/eva.12625.
41. Sun, S., Klebaner, F., Zhang, X., and Tian, T. (2018). Instantaneous mutation rate in cancer initiation and progression. BMC Syst. Biol. *12*, 110. https://doi.org/10.1186/s12918-018-0629-z.
42. Fisk, J.N., Mahal, A.R., Dornburg, A., Gaffney, S.G., Aneja, S., Contessa, J.N., Rimm, D., Yu, J.B., and Townsend, J.P. (2022). Premetastatic shifts of endogenous and exogenous mutational processes support consolidative therapy in EGFR-driven lung adenocarcinoma. Cancer Lett. *526*, 346–351. https://doi.org/10.1016/j.canlet.2021.11.011.
43. Russo, M., Crisafulli, G., Sogari, A., Reilly, N.M., Arena, S., Lamba, S., Bartolini, A., Amodio, V., Magrì, A., Novara, L., et al. (2019). Adaptive mutability of colorectal cancers in response to targeted therapies. Science *366*, 1473–1480. https://doi.org/10.1126/science.aav4474.
44. Shi, W., Ng, C.K.Y., Lim, R.S., Jiang, T., Kumar, S., Li, X., Wali, V.B., Piscuoglio, S., Gerstein, M.B., Chagpar, A.B., et al. (2018). Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. Cell Rep. *25*, 1446–1457. https://doi.org/10.1016/j.celrep.2018.10.046.
45. Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. Genome Biol. *17*, 86. https://doi.org/10.1186/s13059-016-0936-x.
46. Zafar, H., Tzen, A., Navin, N., Chen, K., and Nakhleh, L. (2017). SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome Biol. *18*, 178. https://doi.org/10.1186/s13059-017-1311-2.
47. Kozlov, A., Alves, J.M., Stamatakis, A., and Posada, D. (2022). CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. Genome Biol. *23*, 37. https://doi.org/10.1186/s13059-021-02583-w.
48. Kang, S., Borgsmüller, N., Valecha, M., Kuipers, J., Alves, J.M., Prado-López, S., Débora, Beerenwinkel, N., Posada, D., and Szczurek, E. (2022). SIEVE: joint inference of single nucleotide variants and cell phylogeny from single-cell DNA sequencing data. Genome Biol. *23*, 248. https://doi.org/10.1186/s13059-022-02813-9.
49. Navin, N.E. (2014). Cancer genomics: one cell at a time. Genome Biol. *15*, 452. https://doi.org/10.1186/s13059-014-0452-9.
50. Posada, D. (2020). CellCoal: coalescent simulation of single-cell sequencing samples. Mol. Biol. Evol. *37*, 1535–1542. https://doi.org/10.1093/molbev/msaa025.
51. Huang, L., Ma, F., Chapman, A., Lu, S., and Xie, X.S. (2015). Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annu. Rev. Genom. Hum. Genet. *16*, 79–102. https://doi.org/10.1146/annurev-genom-090413-025352.
52. Kuipers, J., Jahn, K., Raphael, B.J., and Beerenwinkel, N. (2017). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome Res. *27*, 1885–1894. https://doi.org/10.1101/gr.220707.117.
53. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. *17*, 368–376. https://doi.org/10.1007/BF01734359.
54. Swofford, D.L. (2003). PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) (Sinauer Associates). *Version 4*.
55. Ota, T., and Kimura, M. (1971). On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. *1*, 18–25. https://doi.org/10.1007/BF01659391.
56. Martincorena, I. (2019). Somatic mutation and clonal expansions in human tissues. Genome Med. *11*, 35. https://doi.org/10.1186/s13073-019-0648-4.
57. Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P., Dentro, S.C., Dawson, K.J., Butler, T., Rahbari, R., Mitchell, T.J., et al. (2020). The mutational landscape of normal human endometrial epithelium. Nature *580*, 640–646. https://doi.org/10.1038/s41586-020-2214-z.
58. Kakiuchi, N., and Ogawa, S. (2021). Clonal expansion in non-cancer tissues. Nat. Rev. Cancer *21*, 239–256. https://doi.org/10.1038/s41568-021-00335-3.
59. Fanelli, G.N., Dal Pozzo, C.A., Depetris, I., Schirripa, M., Brignola, S., Biason, P., Balistreri, M., Dal Santo, L., Lonardi, S., Munari, G., et al. (2020). The heterogeneous clinical and pathological landscapes of metastatic Braf-mutated colorectal cancer. Cancer Cell Int. *20*, 30. https://doi.org/10.1186/s12935-020-1117-2.
60. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.
61. Preto, A., Figueiredo, J., Velho, S., Ribeiro, A.S., Soares, P., Oliveira, C., and Seruca, R. (2008). BRAF provides proliferation and survival signals in MSI colorectal carcinoma cells displaying BRAFV600E but not KRAS mutations. J. Pathol. *214*, 320–327. https://doi.org/10.1002/path.2295.
62. Jin, X., Zhu, L., Cui, Z., Tang, J., Xie, M., and Ren, G. (2019). Elevated expression of GNAS promotes breast cancer cell proliferation and migration via the PI3K/AKT/Snail1/E-cadherin axis. Clin. Transl. Oncol. *21*, 1207–1219. https://doi.org/10.1007/s12094-019-02042-w.
63. Li, J., Duns, G., Westers, H., Sijmons, R., van den Berg, A., and Kok, K. (2016). SETD2: an epigenetic modifier with tumor suppressor functionality. Oncotarget *7*, 50719–50734. https://doi.org/10.18632/oncotarget.9368.
64. Vidotto, T., Nersesian, S., Graham, C., Siemens, D.R., and Koti, M. (2019). DNA damage repair gene mutations and their association with tumor immune regulatory gene expression in muscle invasive bladder cancer subtypes. Journal for ImmunoTherapy of Cancer *7*. https://doi.org/10.1186/s40425- 019-0619-8.
65. Poon, G.Y.P., Watson, C.J., Fisher, D.S., and Blundell, J.R. (2021). Synonymous mutations reveal genome-wide levels of positive selection in

healthy tissues. Nat. Genet. *53*, 1597–1605. https://doi.org/10.1038/s41588-021-00957-1.

66. Wijewardhane, N., Dressler, L., and Ciccarelli, F.D. (2021). Normal Somatic Mutations in Cancer Transformation. Cancer Cell *39*, 125–129. https://doi.org/10.1016/j.ccell.2020.11.002.
67. Williams, M.J., Sottoriva, A., and Graham, T.A. (2019). Measuring Clonal Evolution in Cancer with Genomics. Annu. Rev. Genom. Hum. Genet. *20*, 309–329. https://doi.org/10.1146/annurev genom-083117-021712.
68. Notta, F., Chan-Seng-Yue, M., Lemire, M., Li, Y., Wilson, G.W., Connor, A.A., Denroche, R.E., Liang, S.B., Brown, A.M.K., Kim, J.C., et al. (2016). A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. Nature *538*, 378–382. https://doi.org/10.1038/nature19823.
69. Heng, J., and Heng, H.H. (2022). Genome chaos: Creating new genomic information essential for cancer macroevolution. Semin. Cancer Biol. *81*, 160–175. https://doi.org/10.1016/j.semcancer.2020.11.003.
70. Watkins, T.B.K., Lim, E.L., Petkovic, M., Elizalde, S., Birkbak, N.J., Wilson, G.A., Moore, D.A., Grönroos, E., Rowan, A., Dewhurst, S.M., et al. (2020). Pervasive chromosomal instability and karyotype order in tumour evolution. Nature *587*, 126–132. https://doi.org/10.1038/s41586-020-2698-6.
71. Perdigoto, C.N. (2019). Epigenetic cancer evolution, one cell at a time. Nat. Rev. Genet. *20*, 434–435. https://doi.org/10.1038/s41576-019-0143-1.
72. Kester, L., van Oudenaarden, A., and Alexander. (2018). Single-Cell Transcriptomics Meets Lineage Tracing. Cell Stem Cell *23*, 166–179. https://doi.org/10.1016/j.stem.2018.04.014.
73. Somarelli, J.A., Gardner, H., Cannataro, V.L., Gunady, E.F., Boddy, A.M., Johnson, N.A., Fisk, J.N., Gaffney, S.G., Chuang, J.H., Li, S., et al. (2020). Molecular Biology and Evolution of Cancer: From Discovery to Action. Mol. Biol. Evol. *37*, 320–326. https://doi.org/10.1093/molbev/msz242.
74. Noble, R., Burri, D., Le Sueur, C., Cécile, L., Viossat, Y., Kather, J.N., Beerenwinkel, N., and Beerenwinkel, N. (2022). Spatial structure governs the mode of tumour evolution. Nat. Ecol. Evol. *6*, 207–217. https://doi.org/10.1038/s41559-021-01615-9.
75. Phan, T.G., and Croucher, P.I. (2020). The dormant cancer cell life cycle. Nat. Rev. Cancer *20*, 398–411. https://doi.org/10.1038/s41568-020-0263-0.
76. Maley, C.C., Aktipis, A., Graham, T.A., Sottoriva, A., Boddy, A.M., Janiszewska, M., Silva, A.S., Gerlinger, M., Yuan, Y., Pienta, K.J., et al. (2017). Classifying the evolutionary and ecological features of neoplasms. Nat. Rev. Cancer *17*, 605–619. https://doi.org/10.1038/nrc.2017.69.
77. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122–128. https://doi.org/10.1038/s41586-019-1907-7.
78. Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H., Hasaart, K., de la Fonteijne, L., Varela, I., Camargo, F.D., and van Boxtel, R. (2018). Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. *25*, 2308–2316.e4. https://doi.org/10.1016/j.celrep.2018.11.014.
79. Coorens, T.H.H., Moore, L., Robinson, P.S., Sanghvi, R., Christopher, J., Hewinson, J., Przybilla, M.J., Lawson, A.R.J., Spencer Chapman, M., Cagan, A., et al. (2021). Extensive phylogenies of human development inferred from somatic mutations. Nature *597*, 387–392. https://doi.org/10.1038/s41586-021-03790-y.
80. Mitchell, E., Spencer Chapman, M., Williams, N., Dawson, K.J., Mende, N., Calderbank, E.F., Jung, H., Mitchell, T., Coorens, T.H.H., Spencer, D.H., et al. (2022). Clonal dynamics of haematopoiesis across the human lifespan. Nature *606*, 343–350. https://doi.org/10.1038/s41586-022-04786-y.
81. Chkhaidze, K., Heide, T., Werner, B., Williams, M.J., Huang, W., Caravagna, G., Graham, T.A., and Sottoriva, A. (2019). Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLoS Comput. Biol. *15*, 1–26. https://doi.org/10.1371/journal.pcbi.1007243.
82. Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., Li, F., Im, K.M., Wu, K., Wu, H., Ye, X., et al. (2012). Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. GigaScience *1*. https://doi.org/10.1186/2047-217X-1-12.
83. Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., et al. (2012). Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. Cell *148*, 873–885. https://doi.org/10.1016/j.cell.2012.02.028.
84. Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155–160. https://doi.org/10.1038/nature13600.
85. Alves, J.M., Prado-López, S., Tomás, L., Valecha, M., Estévez-Gómez, N., Alvariño, P., Geisel, D., Modest, D.P., Sauer, I.M., Pratschke, J., Raschzok, N., et al. (2022). Clonality and timing of relapsing colorectal cancer metastasis revealed through whole-genome single-cell sequencing. Cancer Lett. *543*, 215767. https://doi.org/10.1016/j.canlet.2022. 215767.
86. Wu, H., Zhang, X.Y., Hu, Z., Hou, Q., Zhang, H., Li, Y., Li, S., Yue, J., Jiang, Z., Weissman, S.M., et al. (2017). Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. Oncogene *36*, 2857–2867. https://doi.org/10.1038/onc.2016.438.
87. Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. Cell *148*, 886–895. https://doi.org/10.1016/j.cell.2012.02.025.
88. Ni, X., Zhuo, M., Su, Z., Duan, J., Gao, Y., Wang, Z., Zong, C., Bai, H., Chapman, A.R., Zhao, J., et al. (2013). Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc. Natl. Acad. Sci. USA *110*, 21083–21088. https://doi.org/10.1073/pnas.1320659110.
89. Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science (New York, N.Y.) *350*, 94–98. https://doi.org/10.1126/science.aab1785.
90. Su, F., Zhang, W., Zhang, D., Zhang, Y., Pang, C., Huang, Y., Wang, M., Cui, L., He, L., Zhang, J., et al. (2018). Spatial Intratumor Genomic Heterogeneity within Localized Prostate Cancer Revealed by Single-nucleus Sequencing. Eur. Urol. *74*, 551–559. https://doi.org/10.1016/j.eururo.2018.06.005.
91. Caravagna, G., Heide, T., Williams, M.J., Zapata, L., Nichol, D., Chkhaidze, K., Cross, W., Cresswell, G.D., Werner, B., Acar, A., et al. (2020). Subclonal reconstruction of tumors by using machine learning and population genetics. Nat. Genet. *52*, 898–907. https://doi.org/10.1038/s41588-020-0675-5.
92. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.
93. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.
94. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
95. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy

number and mutation profiles from tumor sequencing data. Ann. Oncol. *26*, 64–70. https://doi.org/10.1093/annonc/mdu479.

96. Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A.Y., Wang, T., and Vijg, J. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat. Methods *14*, 491–493. https://doi.org/10.1038/nmeth.4227.
97. Zafar, H., Wang, Y., Nakhleh, L., Navin, N., and Chen, K. (2016). Monovar: single-nucleotide variant detection in single cells. Nat. Methods *13*, 505–507. https://doi.org/10.1038/nmeth.3835.
98. Lalee, M., Nocedal, J., and Plantenga, T. (1998). On the Implementation of an Algorithm for Large Scale Equality Constrained Optimization. SIAM J. Optim. *8*, 682–706. https://doi.org/10.1137/S1052623493262993.
99. Self, S.G., and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. J. Am. Stat. Assoc. *82*, 605–610. https://doi.org/10.2307/2289471.
100. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Res. *50*, D20–D26. https://doi.org/10.1093/nar/gkab1112.
101. Broad Institute. Picard Tools. http://broadinstitute.github.io/picard/.
102. Van der, A., Geraldine, A., Connor, O.', and Brian, D. (1920). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (O'Reilly Media).
103. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
1000G Reference Genome hs37d5, GRCh37	Genome Reference Consortium	https://www.internationalgenome.org/category/reference/
GATK Resource Bundle b37	Broad Institute	https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/
scDNA WES data Li	Li et al.[82]	SRA: SRA051489
scDNA WES data Hou	Hou et al.[83]	SRA: SRA050202
scDNA WES data Wang	Wang et al.[84]	SRA: SRA053195
scDNA WGS data Kang	Kang et al.[48]	SRA: SRP067815
scDNA WES data Wu	Wu et al.[85]	SRA: SRP067815
scDNA WES data Xu	Xu et al.[86]	SRA: SRA050201
scDNA WES data Ni	Ni et al.[87]	SRA: SRP029757
scDNA WGS data Lodato	Lodato et al.[87]	SRA: SRP041470, SRP061939
scDNA WGS data Su	Su et al.[88]	SRA: SRP127755
scDNA WGS data Alves	Alves et al.[89]	Mendeley Data: https://doi.org/10.17632/pbbx6gckck.1
scDNA WGS data Kozlov	Kozlov et al.[47]	BioProject: PRJNA789841
Software and algorithms		
PT test	This paper	https://doi.org/10.5281/zenodo.7998185
Python v3.7.6 with packages numpy v1.18.5, pandas v1.0.5, scipy v1.5.1, and ete3 v3.1.2	Python Software Foundation	https://www.python.org
CellCoal v1.3.1	Posada[50]	https://github.com/dapogon/cellcoal
Cellphy v0.9.2	Kozlov et al.[47]	https://github.com/amkozlov/cellphy
infSCITE	Kuipers et al.[52]	https://github.com/cbg-ethz/infSCITE
neutralitytestr v0.0.3	Williams et al.[22]	https://CRAN.R-project.org/package=neutralitytestr
Mobster v1.0.0	Caravagna et al.[90]	https://caravagnalab.github.io/mobster
dndscv v0.1.0	Martincorena et al.[11]	https://github.com/im3sanger/dndscv
cutadapt v1.18	Martin et al.[91]	https://github.com/marcelm/cutadapt
bwa v0.7.17	Li et al.[92]	https://github.com/lh3/bwa
Picard SortSam v2.18.14	Broad Institute	https://broadinstitute.github.io/picard/
Picard MarkDuplicates v2.18.14	Broad Institute	https://broadinstitute.github.io/picard/
GATK IndelRealignement v3.7.0	Broad Institute	https://gatk.broadinstitute.org/hc/en-us
GATK BaseRecalibrator v4.0.10	Broad Institute	https://gatk.broadinstitute.org/hc/en-us
samtools v1.9	Danecek et al.[93]	https://www.htslib.org/
Sequenza, v3.0.0	Favero et al.[94]	https://github.com/oicr-gsi/sequenza
SCCcaller v2.0.0	Dong et al.[95]	https://github.com/NBMueller/SCcaller
Monovar	Zafar et al.[96]	https://github.com/NBMueller/MonoVar
Ensembl Variant Effect Predictor, release/100.0	McLaren et al.[97]	https://github.com/Ensembl/ensembl-vep

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, David Posada (dposada@uvigo.es).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Poisson tree test model and input data

The Poisson tree (PT) test requires as input a matrix of SNV genotypes $\boldsymbol{X}$, false positive and false negative genotype error rates (α and β, respectively), and a rooted tree topology τ, representing the phylogeny of a sample of contemporaneous cells. The genotype matrix $\boldsymbol{X}$ has dimensions $m \times n$, where m is the number of SNV loci and n is the number of cells. Each element of $\boldsymbol{X} \in \{0, 1, -\}^{m \times n}$ indicates if an SNV is present (1), absent (0), or missing ($-$) in a given cell.

The false positive rate α is the probability that an unmutated genotype is wrongly identified as mutated (i.e., true 0's called as 1's), the false negative rate β is the probability of a mutated genotype (1) to be wrongly identified as unmutated (0) (i.e., true 1's called as 0's). In scDNA-seq data, false positives arise mainly from DNA lesions occurring during cell isolation and manipulation, single-cell whole-genome amplification (scWGA) errors, and sequencing errors. False negatives also arise from these errors, but most of them result from allele dropout (ADO) during scWGA. Missing genotypes arise from ADO and insufficient sequencing coverage. The fraction of missing genotypes in each cell j is $\gamma_j = \frac{1}{m}\sum_{p=1}^{m}[x_{p,j} = -]$, where $[\cdot]$ is the indicator function.

The cell tree topology τ consists of $2n-1$ nodes and $2n-2$ branches I. The n leaf nodes L correspond to sampled cells and the $n-1$ internal nodes I represent unobserved ancestor cells. τ can be represented by a binary matrix $\boldsymbol{A} \in \{0,1\}^{I \times n}$, where rows represent branches and columns represent cells, and $a_{i,j} = 1$ if branch i belongs to the path from the root to cell j, and 0 otherwise. The length of a branch (λ) represents the expected number of SNVs accumulated along that branch and is the product of the expected number of SNVs taking place per time unit (i.e., the evolutionary rate θ) and the time elapsed along the branch (t):

$$\lambda = \theta \cdot t \qquad \text{(Equation 1)}$$

The evolutionary rate is the product of the mutation rate per cell division times the number of cell divisions along the branch, but these two quantities are indistinguishable.

To assess the constancy of θ, we must first infer the expected branch lengths λ. To do so, we model the observed branch lengths k (see STAR Methods section 'Computation of the observed branch lengths' for further details) as Poisson variables with mean $\lambda = [\lambda_1, \ldots, \lambda_I] \in \mathbb{R}_{>0}$. The likelihood of the observed branch lengths k given the expected branch lengths λ is

$$L(\boldsymbol{k}|\boldsymbol{\lambda}) = \prod_{i=1}^{I} \text{Poisson}(k_i|\lambda_i)^{w_i} = \prod_{i=1}^{I} \left(\frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!}\right)^{w_i} \qquad \text{(Equation 2)}$$

where w_i weighs the impact of branch length k_i on the likelihood according to its probability of being affected by scDNA-seq errors (see STAR Methods section 'Branch weights' for further details). The log likelihood is then:

$$\mathcal{L}(\boldsymbol{k}|\boldsymbol{\lambda}) = \log(L(\boldsymbol{k}|\boldsymbol{\lambda})) \propto \sum_{i=1}^{I} w_i(k_i \log(\lambda_i) - \lambda_i) \qquad \text{(Equation 3)}$$

where we have omitted the constant $\sum_{i=1}^{I}\log(k_i!)$, which cancels out in the likelihood ratio below and, thus, does not affect the maximum likelihood solution. The maximum likelihood estimate (MLE) of λ can be obtained by solving the following equation:

$$\min_{\lambda \in \mathbb{R}_{>0}^{I}} -\sum_{i=1}^{I} w_i(k_i \log(\lambda_i) - \lambda_i) \qquad \text{(Equation 4)}$$

Null model: Constant evolutionary rate

The null model assumes a constant evolutionary rate θ along the tree (i.e., a molecular clock). Consequently, the expected cumulative branch length from any internal node to any cell should be similar. This imposes $n-1$ constraints on the expected branch lengths $\boldsymbol{\lambda}$, which can be written as a system of linear equations defined by a constraint matrix $\boldsymbol{C} \in \{-1, 0, 1\}^{(n-1) \times I}$. Each row in $\boldsymbol{C}$ corresponds to an internal node and each column to a branch:

$$c_{ij} = \begin{cases} 1 & \text{if } E_j \in P(I_i, L_{j\rightarrow}) \\ -1 & \text{if } E_j \in P(I_i, L_{j\leftarrow}) \\ 0 & \textit{else} \end{cases} \qquad \text{(Equation 5)}$$

where $P(l_x, L_y)$ is the path between the internal node x and cell y, i.e., the set of all branches connecting the two nodes, and $L_{i\rightarrow}$ and $L_{i\leftarrow}$ are arbitrary cells from the left or right subtree succeeding node l_i, respectively. There are several equivalent parametrizations of $\boldsymbol{C}$, as both the left and right subtrees and the cells are chosen arbitrarily. A scheme of the constraint matrix is displayed in Figure S5A. Given that the sum of Poisson variables is also a Poisson variable, we can write the constraints imposed by the null model as

$$\boldsymbol{C} \cdot \boldsymbol{\lambda} = 0 \qquad \text{(Equation 6)}$$

To find the MLE (Equation 4) subject to the rate constraints (Equation 6) and the boundary constraints $\lambda > 0$, we used the Byrd-Omojokun Trust-Region Sequential Quadratic Programming algorithm,[98] a gradient-based numerical optimizer.

Alternative model: Constraint-free evolutionary rates

The alternative model allows the evolutionary rate θ to change along the tree. In this scenario, there are no constraints on $\boldsymbol{\lambda}$, i.e., $\boldsymbol{C} = 0$, and Equation 4 can be solved analytically. The likelihood of this model is maximal when setting the expected branch lengths equal to the observed branch lengths (Data S1.1). For zero branch lengths (with no SNVs), we use the limit $\lim_{k \to 0^+} k_i \log(k_i) = 0$.

Likelihood ratio test

To test the constancy of the evolutionary rate, we compare the two competing models using a likelihood ratio test (LRT). The LRT test statistic Λ is twice the negative log likelihood ratio:

$$\Lambda = -2[\mathcal{L}(\mathbf{k}|\boldsymbol{\lambda}_0) - \mathcal{L}(\mathbf{k}|\boldsymbol{\lambda}_1)]$$

$$= -2\left[\sum_{i=1}^{I} w_i(k_i(\log(\lambda_{0,i}) - \log(\lambda_{1,i})) - \lambda_{0,i} + \lambda_{1,i})\right] \qquad \text{(Equation 7)}$$

As the null model is nested in the alternative model, Λ follows a X^2 distribution with $n-1$ degrees of freedom, except for the case when optimized parameters are on the constraint boundary, i.e., zero. In that case, Λ is distributed as a mixture of X^2 distributions with $n, n-1, \ldots n-k$ degrees of freedom, weighted by normalized binomial coefficients.[99]

Computation of the observed branch lengths

We compute the observed branch lengths by mapping the SNVs to specific branches of the input cell tree. We define the matrix $\boldsymbol{M}^{m\times 1} \in [0,1]$, where $m_{p,i}$ is the probability that SNV p is assigned to branch i. We assume SNVs are i.i.d. and follow an infinite sites model (i.e., mutations can only occur once at a given site). Therefore, if an SNV is mapped to a branch, we expect to see it in all the descendant cells unless there are false negative SNV calls or missing data. We can calculate the probability of assigning an SNV to a given branch[45] by considering the set of cells that harbor the SNV and the set of cells we would expect to harbor it if mapped to that branch:

$$m_{p,i} = \frac{\prod_{j=1}^{n} P\left(x_{p,j} \middle| a_{i,j}\right)}{\prod_{j=1}^{n} P\left(x_{p,j} \middle| 0\right) + \sum_{i'=1}^{I} \prod_{j=1}^{n} P\left(x_{p,j} \middle| a_{i',j}\right)} \qquad \text{(Equation 8)}$$

The first product of the denominator is the probability of a false SNV call if the SNV is not present in any cell. Given a false positive rate α and a false negative rate β, we can calculate the probability of the observed SNV x given the expected genotype y:

$$\mathrm{P}(x|y) = \begin{cases} 1-\alpha & \text{if } x=0 \wedge y=0 \\ \beta & \text{if } x=0 \wedge y=1 \\ \alpha & \text{if } x=1 \wedge y=0 \\ 1-\beta & \text{if } x=1 \wedge y=1 \\ 1 & \text{if } x=- \end{cases} \qquad \text{(Equation 9)}$$

The number of SNVs mapped to a branch j, required for solving Equation 4, is the column sum over $\boldsymbol{M}$:

$$k_j = \sum_{p=1}^{m} m_{p,j} \qquad \text{(Equation 10)}$$

The SNV mapping procedure is depicted in Figure S5B.

Branch weights

The estimated branch lengths $\boldsymbol{k}$ are subject to uncertainty due to scDNA-seq noise and the stochasticity of the mutation process. The soft assignment in Equation 8 accounts for uncertainty in the SNV placement, including false positive and false negative calls, if an SNV is called in at least one cell. However, if a true SNV is unobserved due to false negative or missing data in all the cells harboring it, this SNV is not reported. We call this event an SNV loss. The probability of an SNV loss differs for each branch and is proportional to

the number of descendant cells. The more descendant cells harbor the SNV, the less likely it will be lost. Accordingly, branches with many descendant cells are less prone to SNV losses than those with few, making the observed number of SNVs on these branches more reliable. The probability for an SNV loss on a branch is:

$$P_{loss}(E_i|\boldsymbol{A},\alpha,\beta,\boldsymbol{\gamma}) = \prod_{j=1}^{n}\left[(1-\gamma_j)\beta+\gamma_j\right]^{a_{ij}}\cdot\left[(1-\gamma_j)(1-\alpha)+\gamma_j\right]^{1-a_{ij}} \quad \text{(Equation 11)}$$

The first term in Equation 11 describes the probability of a false negative or missing genotype (e.g., due to a lack of sequencing coverage) in all cells containing the SNV; the second term is the probability of no false positive or missing genotypes in any other cell. Given the probability for an SNV loss and by assuming that it is Binomial-distributed, i.e., an SNV is either lost or not, we can weigh the branches by their inverse-variance of their SNV loss probability by defining:

$$\tilde{w}_i = min\left\{\left(P_{loss}(E_i|\boldsymbol{A},\alpha,\beta,\boldsymbol{\gamma})(1-P_{loss}(E_i|\boldsymbol{A},\alpha,\beta,\boldsymbol{\gamma}))\right)^{-1}, w_{max}\right\} \quad \text{(Equation 12)}$$

To ensure that the test statistic of the LRT can be approximated with a X^2 distribution, we rescaled the weights to retain the original degrees of freedom

$$w_i = I\cdot\frac{\tilde{w}_i}{\sum_{i'=1}^{I}\tilde{w}_{i'}} \quad \text{(Equation 13)}$$

such that $\sum_{i=1}^{I} w_i = I$. The branch weighting is depicted in Figure S5C.

The inverse variance of a Bernoulli distribution is unbounded. Therefore, we need to define an upper limit w_{max}. Without this limit, the weight for a single or few branches with a very low probability for SNV losses would be several magnitudes higher than for other branches. A w_{max} value of 1000, for instance, caps the probability of an SNV loss at 0.001. As the weights are rescaled, w_{max} also regulates their dispersion: low w_{max} values lead to weights closer to 1, and larger w_{max} values lead to weights dispersed more widely.

We evaluated the impact of w_{max} on the accuracy of the PT test using simulations. We found that $w_{max} = 1000$ ensured a false rejection rate of the null hypothesis close to zero (Data S1.2, Figure S1). Therefore, we used this value for all the calculations in this study.

Implementation

The PT test is implemented in Python and requires called mutations in VCF format, a phylogenetic tree in Newick format, and estimated false negative and false positive rates of the called mutations as input.

QUANTIFICATION AND STATISTICAL ANALYSIS

Simulation of scDNA-seq data with constant and variable evolutionary rates

We used CellCoal[50] to evaluate the accuracy of the PT test. CellCoal simulates the phylogeny of a sample of cells together with SNV genotypes subject to scDNA-seq errors. By default, CellCoal simulates phylogenies in which all lineages share a constant evolutionary rate, resulting in ultrametric or clock-like cell trees in which the evolutionary distance from the root to any cell is equivalent.

To simulate scenarios with evolutionary rate variation, we implemented in CellCoal the possibility of introducing one or more changes in the evolutionary rate along the tree. For each rate change, we chose a branch with probability proportional to its length and multiplied its length, and that of its descendant branches, by a factor. In our simulations, we introduced single rate changes with 2x, 5x, and 10x factors. We only included samples where the fraction of cells affected by the rate change was between 10% and 90%, as rate variability would be hardly distinguishable in the excluded cases.[22]

We simulated samples of 30 and 100 cells for the constant and varying rate scenarios, respectively. From the latter scenario, we obtained subsamples of 10, 30, 50, 70, and 90 cells to assess the effect of the sample size on the statistical power, i.e., on the ability of the PT test to detect variable evolutionary rates. We filtered subsamples without cells affected by the rate change as the evolutionary rate is constant in these. In all cases, we simulated a genome length of 10,000 sites and a mutation rate of 10^{-6}, except for the scenarios with variable rates, in which we scaled the mutation rate to obtain a similar number of SNVs across scenarios (Table 2). We kept the expected sequencing depth constant across scenarios (20x) and explored multiple levels of scDNA-seq bias by considering different combinations of ADO, sequencing depth overdispersion, and sequencing and amplification error rates (Table 2). For each set of conditions, we simulated 1,000 (constant rate scenarios) or 3,000 (variable rates scenarios) replicates.

The SNV genotypes resulting from the CellCoal simulations have attached a particular sequencing depth (depending on the expected sequencing depth, its overdispersion, and the ADO rate) and the relative likelihood of observing this genotype. In all cases, we filtered out SNV sites with sequencing depth below 5x or conditional genotype quality (GQ) below one, where GQ is the Phred-scaled likelihood difference between the most likely and the second most likely genotype.

The pipelines for simulating data and all subsequent analyses were implemented in Snakemake.

Inference of cell phylogenies

Using the simulated data, we inferred the cell phylogenies with CellPhy[47] (-o healthycell -l), which operates a constraint-free model for the branch lengths, and with infSCITE[52] (-r 1 -n 5e5 -d 0.01 -ad 0.2 -e 0.1 -z -a -transpose), which infers the tree topology but not the branch lengths). Cellphy infers the amplification/sequencing error rate (ERR), corresponding to the probability of observing a wrong allele and the ADO rate. To translate CellPhy's estimates of ADO and ERR into false positive and false negative rates, we used half of the estimated ADO rate plus one-third of the estimated ERR rate as false negative rate. Half of the ADO rate represents the chance that the mutated allele is affected by an ADO, and one-third of the ERR rate represents the chance that the mutated allele appears as the reference allele due to an error. As the false positive rate for the PT test, we used CellPhy's estimate of the ERR rate, representing the chance that an unmutated allele appears as a mutated allele due to an error. To root the phylogenetic trees inferred by CellPhy, we added a synthetic cell lacking SNVs and used it for outgroup rooting.

VAF-based tests of neutrality and subclonal selection for bulk data

For the benchmark of the 1/*f* test[22] and Mobster,[91] we simulated samples of 100 cells with a genome length of 10,000 sites, with 1x sequencing depth and without scDNA-seq errors, imitating bulk data. To simulate evolutionary rate variation, we used the same rate change factors as for the single-cells, 2x, 5x, and 10x, and the corresponding mutation rates. As before, we included only datasets where the fraction of cells affected by the rate change was between 20% and 70%, as the 1/*f* test detects deviations from neutrality only in that VAF range.[23]

We then calculated the VAF for each SNV by summing up the number of reads supporting the mutated allele in all cells and dividing it by the total number of reads in all cells at this SNV site. In total, we simulated 250 and 750 pseudo-bulk samples with a constant or variable evolutionary rate, respectively (for the variable cases, 250 replicates for each rate change factor: 2x, 5x, and 10x).

We used the VAFs calculated from the pseudo-bulk samples as input for the 1/*f* test and Mobster. For the 1/*f* test, we set the sequencing depth to 100x, the ploidy to 2, and the tumor purity to 1. For Mobster, we set the number of possible "subclones" to 1. For the datasets with a constant evolutionary rate, we expect zero subclones; for the datasets with rate variation, we expect one subclone differing in its evolutionary rate.

dN/dS ratio estimation

We calculated dN/dS ratios with the R package dndscv[11] and default parameters, once including only mutations in the 369 cancer driver genes identified by Martincorena et al.[11] and once for all genes. For the scDNA-seq data, we used pseudo-bulk data, as the number of coding SNVs per cell was not enough to calculate dN/dS ratios for individual cells.

Biological data processing

We downloaded 24 scDNA-seq datasets in FASTQ format from the NCBI's Sequence Read Archive (SRA) database.[100] We trimmed library adapters and amplification protocol-specific adapters with cutadapt,[92] mapped the reads to the 1000G Reference Genome hs37d5 with bwa,[93] and sorted them with Picard[101] SortSam. We marked read duplicates with Picard MarkDuplicates and realigned around indels with GATK[102] IndelRealignement using the 1000G Phase 1 and the Mills and 1000G gold standard databases. GATK BaseRecalibrator was used to recalibrate base scores considering dbSNP (build 138) and indels from the 1000G Phase 1. We calculated sequencing depth and breadth with samtools.[94] The pipelines for processing scDNA-seq data were implemented as a Snakemake workflow.

For each cell, we computed the ADO rates as described in Lodato et al.[89] Cells with an extremely high ADO rate (above Q3 + 1.5 IQR per dataset), as well as cells with <40% coverage breadth, were excluded from the analyses (Table S3). To maximize statistical power, we used all available SNVs. We applied stringent filters to all SNV calls, ensuring the same quality for on- and off-target sites. Similar pre-processing was done for the normal and tumor bulk data, followed by the estimation of copy numbers using Sequenza.[95] We called SNVs in single cells using a modified version of SCCcaller[96] with default parameters (https://github.com/NBMueller/SCcaller - modifications listed) and a modified version of Monovar[97] with default parameters except for the consensus filtering step (https://github.com/NBMueller/MonoVar - modifications listed). As input for Monovar, we generated read pileups with a minimum mapping quality of 40 using samtools mpileup. When tumor bulk samples were available, we called SNVs with GATK Mutect2, following the GATK best practice workflow for "Somatic short variant discovery (SNVs + Indels)". Finally, we generated a set of high-confidence SNVs for each dataset by 1) excluding SNVs with a quality score below ten or a read depth below ten, and 2) excluding SNVs that were called in only one cell and were not supported by both single-cell callers or by the bulk tumor sample. Additionally, we excluded SNVs with missing data in more than 50% of the cells. To identify potential phenotypic effects, we annotated the SNV calls with the Ensembl Variant Effect Predictor[103]

For the scDNA-seq data, we inferred cell phylogenies with CellPhy with the same settings as for the simulated data. For the bulk data, we ran the 1/*f* test with ploidy and cellularity values inferred by Sequenza and Mobster with default parameters.