LONGITUDINAL CANCER EVOLUTION FROM SINGLE CELLS

Daniele Ramazzotti[1,†], **Fabrizio Angaroni**[2,†], **Davide Maspero**[2,3,4,†], **Gianluca Ascolani**[2], **Isabella Castiglioni**[5,4], **Rocco Piazza**[1], **Marco Antoniotti**[2], **Alex Graudenzi**[4,*]

[1] School of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy
[2] Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy
[3] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[4] Inst. of Molecular Bioimaging and Physiology,
Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
[5] Department of Physics "Giuseppe Occhialini", Univ. of Milan-Bicocca, Milan, Italy
† Equal contributors
* Corresponding author: alex.graudenzi@ibfm.cnr.it

ABSTRACT

The rise of longitudinal single-cell sequencing experiments on patient-derived cell cultures, xenografts and organoids is opening new opportunities to track cancer evolution in single tumors and to investigate intra-tumor heterogeneity. This is particularly relevant when assessing the efficacy of therapies over time on the clonal composition of a tumor and in the identification of resistant subclones.

We here introduce LACE (Longitudinal Analysis of Cancer Evolution), the first algorithmic framework that processes single-cell somatic mutation profiles from cancer samples collected at different time points and in distinct experimental settings, to produce longitudinal models of cancer evolution. Our approach solves a Boolean matrix factorization problem with phylogenetic constraints, by maximizing a weighted likelihood function computed on multiple time points, and we show with simulations that it outperforms state-of-the-art methods for both bulk and single-cell sequencing data.

Remarkably, as the results are robust with respect to high levels of data-specific errors, LACE can be employed to process single-cell mutational profiles as generated by calling variants from the increasingly available scRNA-seq data, thus obviating the need of relying on rarer and more expensive genome sequencing experiments. This also allows to investigate the relation between genomic clonal evolution and phenotype at the single-cell level.

To illustrate the capabilities of LACE, we show its application to a longitudinal scRNA-seq dataset of patient-derived xenografts of $BRAF^{V600E/K}$ mutant melanomas, in which we characterize the impact of concurrent BRAF/MEK-inhibition on clonal evolution, also by showing that distinct genetic clones reveal different sensitivity to the therapy.

Introduction

The advent of single-cell omics measurements has fueled an exceptional growth of high-resolution quantitative studies on complex biological phenomena [1, 2]. This is extremely relevant in the analysis of cancer evolution and in the characterization of intra-tumor heterogeneity (ITH), which is a major cause of drug resistance and relapse [3, 4, 5, 6].

In recent years, a large array of targeted cancer therapies such as kinase inhibitors, monoclonal antibodies and, more recently, immunomodulatory agents and clinical grade CAR-T has been developed [7]. However, the availability of such personalized therapies requires comparably advanced diagnostic and monitoring tools to study the response of cancer cells under the selective pressure generated by the treatment. In this respect, a highly-awaited major experimental advancement is provided by *longitudinal* single-cell sequencing experiments on samples taken at different time points from the same tumor, or from patient-derived cell cultures, xenografts or *organoids* [8, 9]. In most cases, single-cell transcriptomes are sequenced, e.g., via scRNA-seq experiments [10], yet recently some methods for genotyping single cells with good quality and reasonable costs have been proposed [11, 12].

Longitudinal single-cell sequencing data might allow to track cancer evolution at unprecedented resolution, as they can be employed – in principle – to call somatic variants in each single cell of a tumor sample, at any given time point. Accordingly, this may allow to draw a high-resolution picture of evolutionary history of that tumor, as well as to measure the effect of any possible external intervention, such as a therapeutic strategy. Yet, no technique can explicitly process longitudinal single-cell mutational profiles.

On the one hand, in fact, the existing list of approaches that process single-cell data and extend phylogenetic methods by handling data-specific errors [13, 14, 15, 16], are not suitable to handle multiple temporally ordered samples derived from the same tumor, and cannot be used to investigate the clonal prevalence variation in time. On the other hand, even though methods for longitudinal *bulk* sequencing data are starting to produce noteworthy results [17, 18, 19], they usually require complex computational strategies to deconvolve the signal coming from intermixed cell subpopulations. Furthermore, there is an ongoing debate whether multi-sample trees from bulk samples are indeed phylogenies or, conversely, if they might lead to erroneous evolutionary inferences [20].

We here propose LACE (Longitudinal Analysis of Cancer Evolution), a new computational method for the reconstruction of longitudinal clonal trees of tumor evolution from longitudinal single-cell somatic mutation profiles of tumor samples. In the output tree, each vertex corresponds to a clone – which we here consider as a subpopulation of single cells sharing the same *drivers* –, whereas edges represent the parental relations among them. LACE estimates the prevalence of each clone at each time point, hence allowing to identify clones that emerge, expand, shrink or disappear during the history of the tumor, e.g., as a consequence of a therapy or a selection sweep.

Our method manages noise in single-cell sequencing data by estimating false positive and false negative rates – which might be different in distinct time points –, and takes full advantage of the temporal information present in the data, by returning the longitudinal clonal tree that maximizes a weighted likelihood function computed on all data points. In this way, our method is able to handle possible differences in quality, sample size and error rates of the experiments performed at distinct time points. It is well known, in fact, that extremely different error rates are observed in single-cell experiments performed via distinct experimental platforms, and that even experiments made with the same platform might display highly heterogeneous noise levels [21].

The search is performed by solving a Boolean matrix factorization problem [22], either via exhaustive search in case of very small models or via a Markov Chain Monte Carlo (MCMC), which ensures high scalability and convergence (with infinite samplings). LACE can also return the posterior probability of the output model with respect to each data point, therefore allowing to quantitatively assess the statistical confidence of the inference.

Importantly, the robustness of our approach allows its application to the highly-available scRNA-seq data – which are usually employed to characterize the gene expression patterns of single-cells in a variety of experimental settings [23] –, by calling somatic variants in transcribed regions with standard pipelines [24] and by selecting a set of putative drivers. This allows to overcome the limitation of relying on longitudinal single-cell whole genome/exome sequencing experiments, which are currently rarer and significantly more expensive. In addition, by applying standard data analysis pipelines for the analysis of transcriptomes [23], our method allows to investigate the relation between somatic evolution and gene expression profiles in cancer clones and in single cells, especially in relation with possible external interventions, such as therapies.

There are several advantages in employing a formulation of the problem based on clonal trees, instead of standard phylogenetic trees in which single cells are displayed as leaves. First, currently available single-cell data cannot guarantee the identifiability of a unique and reliable phylogenetic tree, mostly due to noise, to insufficient information and to the huge number of features of the output model [13]. Second, from the biological perspective, the resolution at the clone level is an effective choice to explain and predict cancer evolution and generate hypotheses with translational relevance [25], whereas phylogenetic trees including hundreds or even thousands of cells might be extremely difficult to query and interpret, especially with respect to the possible effect of therapies.

In order to assess the accuracy and robustness of the results produced by LACE, we performed extensive simulations, and compared with CALDER [19], a recent method for the reconstruction of longitudinal phylogenetic trees from bulk samples, SCITE [13] and TRaIT [16], two state-of-the-art tools for the inference of mutational trees from single-cell sequencing data.

We finally applied LACE to a longitudinal scRNA-seq dataset of patient-derived xenografts (PDXs) of $\text{BRAF}^{\text{V600E/K}}$ mutant melanomas [26], by first performing mutational profiling via the widely-used GATK pipeline [27] and by selecting a panel of somatic marker variants. We here show that LACE produces robust results on longitudinal cancer evolution, even with noisy and incomplete data, and in particular, that it can characterize the efficacy of BRAF/MEK-inhibitor therapy on the clonal dynamics, also allowing to portray the phenotypic properties of the distinct (sub)clones.

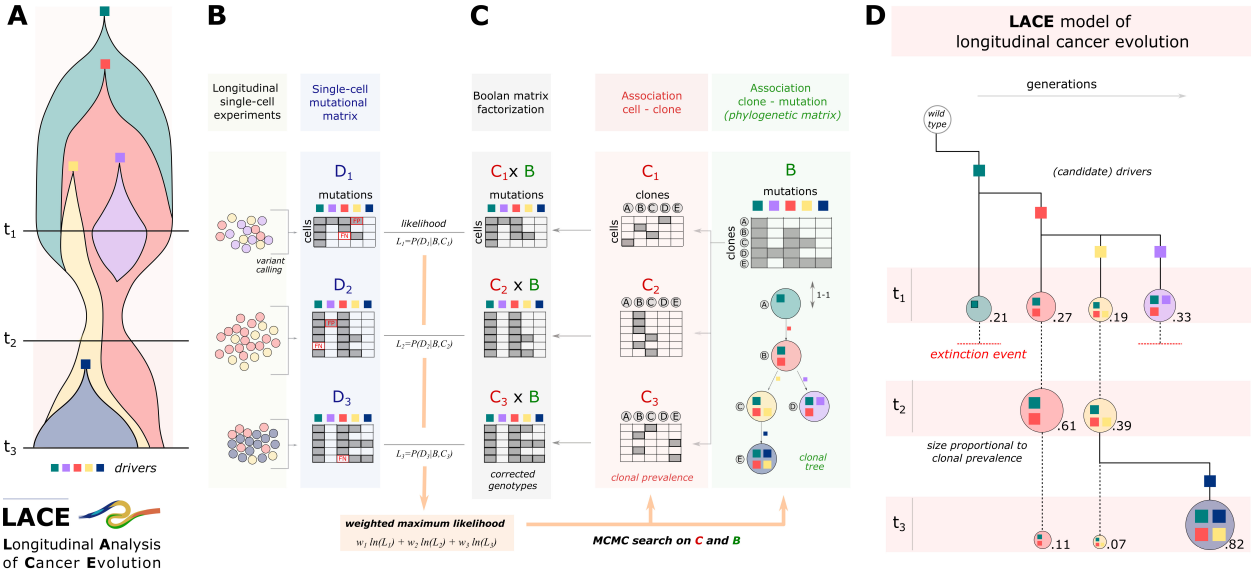


Figure 1: **The LACE framework**. **(A)** The branching evolution of a single tumor is described via a fishplot, and is characterized by the accumulation of 5 driver mutations in distinct clones. **(B)** Single cells are sampled and sequenced at three subsequent time points: t_1, t_2 and t_3, e.g., via scRNA-seq experiments. By applying pipelines for variant calling, somatic mutation profiles for each time point are generated (matrices $\mathbf{D}_i$), which might include false positives and false negatives. **(C)** LACE solves a Boolean matrix factorization problem $\mathbf{C}_i \cdot \mathbf{B} = \mathbf{D}_i$, in which $\mathbf{B}$ is the phylogenetic matrix and $\mathbf{C}_i$ is the cell attachment matrix, by maximizing a weighted likelihood function computed on all time points, via a MCMC search scheme. **(D)** As a result, a unique longitudinal clonal tree is returned, which may include both observed and unobserved clones and the ancestral relations between them, as well as the clonal prevalence variation, as measured at distinct time points. Solid lines represent parental relations between clones, each one characterized by a unique candidate driver, whereas dotted lines are displayed for graphical purposes to connect clones through time points or to extinction events.

Results

The LACE framework. LACE is a computational framework that processes multiple temporally ordered mutational profiles of single cells, collected from cancer samples or patient-derived cell cultures, xenografts or organoids, even in distinct experimental settings (e.g., pre- and post-treatment). Such profiles can be derived from whole-genome/exome or targeted sequencing experiments, but also by calling variants from single-cell RNA-seq data.

LACE takes as input a binary matrix for each time point/experiment, in which an entry is 1 if a somatic variant (e.g., single-nucleotide variants – SNVs, structural variants, etc.) is present in a given cell, 0 if not present and NA if the data point is missing. In order to identify putative cancer (sub)clones, LACE allows the selection of a set of candidate driver mutations, which can leverage on standard practices of driver selection, on biological knowledge and from the application of clustering techniques to mutation co-occurrence patterns, as proposed for instance in [25]. Furthermore, our method can be inputted with false positive and false negative rates for each time point, in the case when such information can be derived from the technical features of the experiments. Otherwise, LACE includes a noise estimation procedure, performed via a parameter grid search.

LACE then solves a Boolean matrix factorization problem with standard phylogenetic constraints, by maximizing a weighted likelihood function on all time points. The rationale is that experiments collected at distinct time points may include even extremely different sample sizes and technical errors: LACE allows to balance such differences, by setting proper weights on the likelihood function. As default, the weights are set to be inversely proportional to the sample size of each dataset, in order to have comparable likelihood values through distinct experiments. LACE employs a MCMC search scheme on the phylogenetic matrix, which is defined by the association between clones and sets of mutations, and which identifies a unique clonal tree. The weighted likelihood is maximized by exhaustively scanning the attachment of single cells to the clones of the tree.

LACE returns: (1) the maximum likelihood clonal tree describing the longitudinal evolution of a tumor, in which nodes are putative clones and edges represent the parental relations among them, as proposed for instance in [28]; (2) the attachment of single cells to clones, which can be used to estimate the clonal prevalence at any considered time

point, as well as to identify macro evolutionary events, such as extinction or the emergence of new clones; (3) the posterior probability of the output model; (4) (optional) the error rates as estimated from data. Furthermore, as the cell attachment induces a partitioning of the single cells, this might be used in turn for mapping clones on gene expression clusters, obtained via standard transcriptome analyses, if data are available.

Performance evaluation with synthetic simulations. In order to assess the performance of LACE and compare it with competing approaches, we performed extensive tests on synthetic datasets.

We generated a total of 450 independent datasets for distinct experimental scenarios (see below) and compared LACE with SCITE [13] and TRaIT [16], two high-performing tools for the inference of mutational trees from single-cell sequencing data on single time points, and with CALDER [19], the benchmark tool for longitudinal phylogenetic tree inference from bulk sequencing data. As synthetic datasets need to be adapted to be processed by such tools, with respect to CALDER, we computed the cancer cell fraction of driver mutations from the observed single-cell genotypes, and by assuming a uniform sampling of single cells and a read depth of 200X, which is typical for whole-exome sequencing experiments. Input data for SCITE and TRaIT were generated by concatenating the longitudinal datasets in a unique mutational profile matrix.

We compared the performance in terms of precision: $\frac{\text{true positives}}{\text{true positives}+\text{false positives}}$ and recall: $\frac{\text{true positives}}{\text{true positives}+\text{false negatives}}$. The complete parameter settings of the simulations are provided as Supplementary Information (SI).

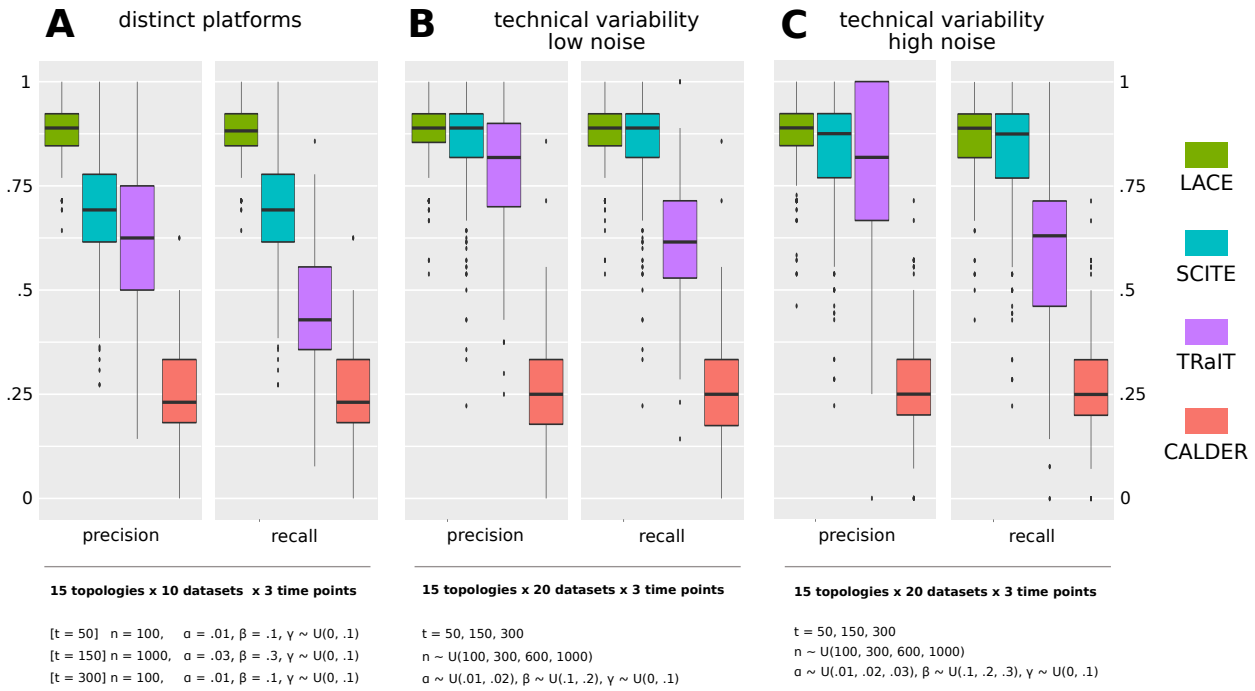


Figure 2: **Comparison on simulated data.** The population dynamics of cancer subpopulations was simulated with the tool from [29] – see the SI for the parameter setting. 15 scenarios were selected in which a number of drivers between 7 and 15 was observed during the simulation, resulting in branching evolution topologies. For each topology, a number of independent single-cell mutational profile datasets were sampled at 3 distinct time points of the dynamics. LACE was compared with SCITE [13], TRaIT [16] and CALDER [19], on precision and recall with respect to the ground-truth topology. **(A)** In order to simulate single-cell data from distinct experimental platforms, we generated 10 independent datasets for each topology, including $n = 100$, $\alpha = 0.01$, $\beta = 0.1$ for time points $t = 50$ and $t = 300$, and $n = 1000$, $\alpha = 0.03$, $\beta = 0.3$ for time point $t = 150$. Each generated dataset was subsequently inflated with 10% of missing data, with a 0.5 probability. **(B)** In order to model technical variability, 20 independent datasets were sampled for each topology, in which at each time point ($t = 50, 150, 300$), each dataset includes with uniform and independent probability: a number of cells in the set $\{100, 300, 600, 1000\}$, α in the set $\{0.01, 0.02\}$ and β in the set $\{0.1, 0.2\}$. Also in this case, approximately half of the datasets include 10% of missing data. **(C)** To account for higher error rates, 20 independent datasets for each topology were generated as in (B), with values of α in the set $\{0.01, 0.02, 0.03\}$ and β in the set $\{0.1, 0.2, 0.3\}$.

Synthetic longitudinal datasets generated from distinct experimental platforms. We designed a first in-silico scenario to account for sequencing experiments performed via distinct platforms at different time points, a likely setting for studies involving patient-derived cell cultures or organoids. In particular, we simulated the case in which the first and the third time points are characterized by a smaller number of cells ($n_{t_1} = n_{t_3} = 100$) and a lower noise rate ($\alpha_{t_{1,3}} = 0.01, \beta_{t_{1,3}} = 0.1$), thus modeling a plausible setting resembling a Smart-seq protocol and the second time point in which a much larger number of cells is sequenced ($n_{t_2} = 1000$), yet with a significantly higher noise rate ($\alpha_{t_2} = 0.03, \beta_{t_2} = 0.3$), therefore modeling the features of a typical droplet-based experiment such as 10X Genomics protocol [21]. We also assume that standard pipelines for mutation profiling from scRNA-seq data are employed, and we set a 0.5 probability for each dataset to have $\gamma = 0.1$ of missing entries.

In Figure 2A one can see the distribution of precision and recall in this scenario, with respect to the four approaches. By managing different error rates in distinct time points and by weighting the likelihood function with respect to the sample size, LACE is able to achieve the highest values in both precision and recall, outperforming all competing methods. This proves that LACE produces robust results when dealing with experiments from distinct protocols and with high differences in sample size and noise rates, as it might be common in real-world settings.

Synthetic longitudinal datasets with technical variability. To assess the consequences of noise and of technical variability in a larger experimental setting, we generated 300 independent datasets with a number of cells chosen at each time point with uniform probability in the set $n = \{100, 300, 600, 1000\}$. We modeled a first setting with low noise (i.e., α and β randomly chosen in the set $\{0.01, 0.02\}$ and $\{0.1, 0.2\}$, respectively) and a second setting with higher values of noise (i.e., α and β in the range $\{0.01, 0.02, 0.03\}$ and $\{0.1, 0.2, 0.3\}$, respectively).

In Figure 2B-C one can see that also in this case LACE performs better than all competing methods in both precision and recall and proves to be robust with increasing noise levels. All in all, these results show the applicability of LACE to experimental settings in which a high variability in sample size and error rates is observed across different temporally ordered single-cell sequencing experiments.

We also recall that LACE's output is more expressive than those of methods designed to process single-time point datasets, such as SCITE and TRaIT, as it allows to quantify the clonal prevalence variation in time, as well as to estimate the temporal positioning of phenomena such as clone emergence or extinction, for instance as a consequence of a therapy.

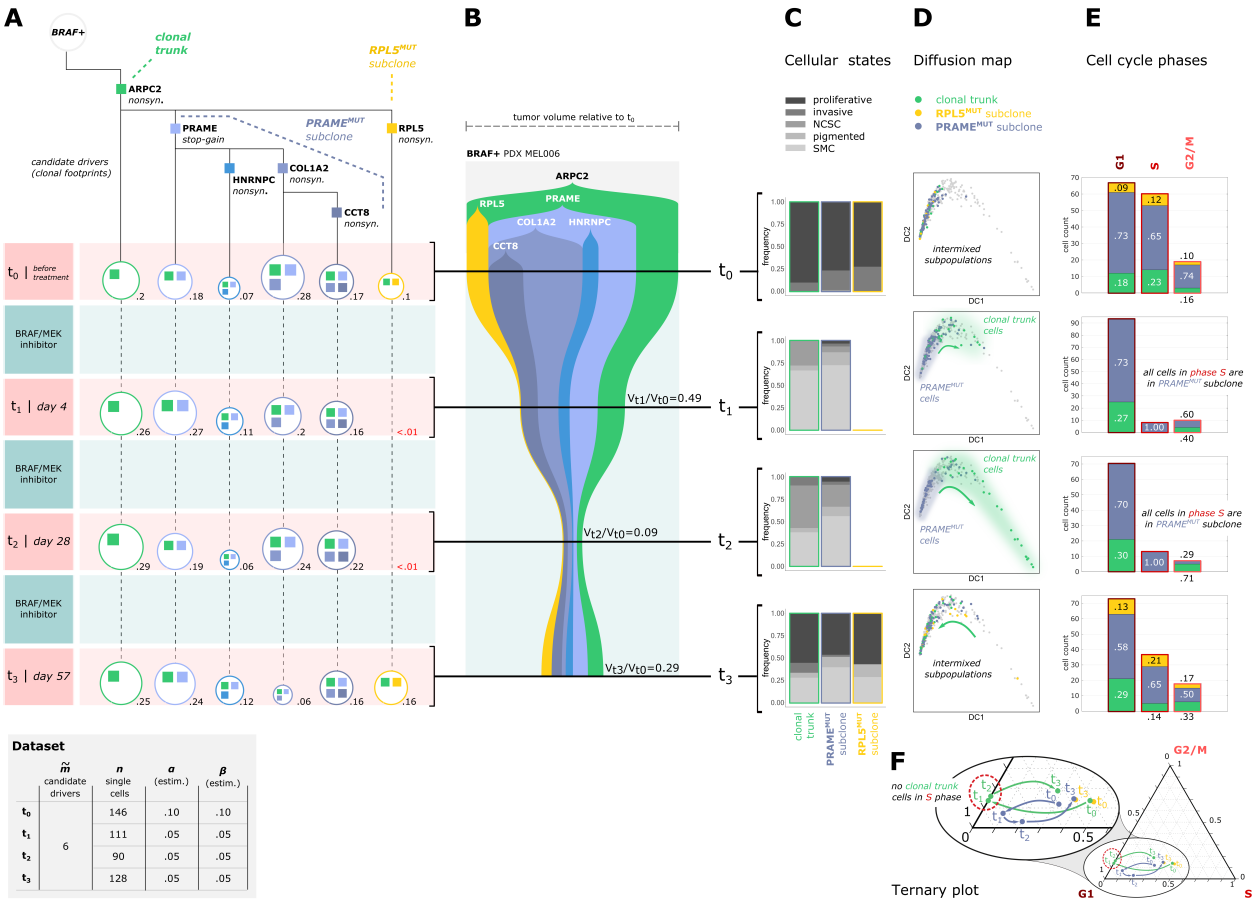
Application of LACE to longitudinal datasets from PDX of BRAF mutant melanomas. We applied LACE to a longitudinal dataset from [26]. In the study, the authors analyze multiple omics data generated from both bulk and single-cell experiments, to investigate minimal residual disease (MRD) in patient-derived xenografts from BRAF-mutant melanomas. In particular, they expose PDXs to $\text{BRAF}^{\text{V600E/K}}$ inhibitor (i.e., *dabrafenib*), either alone or in combination with a MEK inhibitor (i.e., *trametinib*), and they perform multiple sequencing experiments at different time points.

Despite finding de novo mutations in known oncogenes (e.g., MEK1 and NRAS) in resistant cells, the analyses of the copy number alteration profiles, performed via massively parallel sequencing of single-cell genomes, was not sufficient to effectively characterize the clonal architecture and evolution of the tumor, whose composition appear to be similar prior to and after the treatment. Conversely, by analyzing transcriptomic data from both bulk and single-cell RNA-seq experiments, the authors were able to identify four distinct cell subpopulations, characterized by specific transcriptional states (i.e., *neural crest stem cell* – NCSC, *invasive*, *pigmented* and *starved-like melanoma cell* – SMC), which are unsensitive to treatment and eventually lead to relapse, whereas the remaining cell subpopulations get quickly extinct. Based on these findings, the authors hypothesize that that the co-emergence of drug-tolerant states within MRD is predominantly due to the phenotipic plasticity of melanoma cells, which results in transcriptional reprogramming.

We here aim at refining the analysis of the clonal evolution of the tumor, by employing single-cell mutational profiles, as generated by calling variants from scRNA-seq data. In particular, we employed the GATK Best Practices [30] to identify good-quality variants and we selected 6 putative drivers/clonal footprints, by applying a number of filters based on statistical and biological significance (see Methods).

The LACE model shown in Figure 3 reveals the presence of a clonal trunk including a nonsynonimous somatic mutation on ARPC2 – a known melanoma marker [34] –, and of two distinct subclones, with somatic mutations on PRAME and RPL5 as initiating events. In particular, PRAME is a melanoma-associated antigen and known prognostic and diagnostic marker, which was recently targeted for immunotherapy [35]. RPL5 is a candidate tumor suppressor gene for many tumor types and displays inactivating mutations or focal deletions in around 28% of melanomas, which usually result in somatic ribosome defects [36].

The $\text{PRAME}^{\text{MUT}}$ subclone is characterized by the accumulation of further nonsynonimous mutations in HNRNPC, COL1A2 and CCT8 and displays an overall prevalence around $\sim 70\%$ before treatment (t_0). In particular, HNRNPC is



	$\widetilde{m}$ candidate drivers	n single cells	α (estim.)	β (estim.)
t_0	6	146	.10	.10
t_1		111	.05	.05
t_2		90	.05	.05
t_3		128	.05	.05

Figure 3: **LACE model – PDX MEL006.** **(A)** The longitudinal evolution of the PDX MEL006 derived from a BRAF mutant melanoma [26] returned by LACE is displayed. Single-cells were isolated and sequenced via scRNA-seq at four subsequent time points: (t_0) before treatment ($n = 146$ single-cells); (t_1) after 4 days of concurrent treatment with BRAF inhibitor (i.e, dabrafenib) and MEK inhibitor (i.e, trametinib) ($n = 111$), (t_2) after 28 days of treatment ($n = 90$), (t_3) after 57 days of treatment ($n = 128$). Single-cell mutational profiles from scRNA-seq datasets were generated by applying the GATK Best Practices [30], and $\widetilde{m} = 6$ candidate drivers/clonal footprints were selected to be used as input in LACE, with the procedure described in the main text. Each node in the output model represents a candidate (sub)clone, characterized by a set of clonal footprints (colored squares). Solid edges represent parental relations. The clonal prevalence, as measured by normalizing the single-cell count, is displayed near the nodes and marked in red if lower than 1%. In the gray box, the number of candidate drivers $\widetilde{m}$, the number of single cells n and the estimated false positive and false negative rates, α and β, are returned for each time point. **(B)** The representation via a standard fishplot, generated via TimeScape [31], is displayed. The volume size at different time points is taken from [26]. **(C)** The composition of the clonal subpopulation (green) and of both the RPL5$^{\text{MUT}}$ (yellow) and PRAME$^{\text{MUT}}$ (blue) subclones with respect to the cellular states identified via single-cell transcriptomics analysis in [26] is shown for each time point (cells labeled by the authors as "other" were excluded from the analysis). **(D)** The diffusion maps [32] computed on 58 differentially expressed genes identified via ANOVA test (FDR $p < 0.10$) is shown; plots are generated via SCANPY [33]. A distinct diffusion map is shown for each time point, in which only the cells sampled at each time point are colored according to the clonal identity (i.e., clonal trunk, RPL5$^{\text{MUT}}$ subclone or PRAME$^{\text{MUT}}$ subclone). **(E)** The proportion of cells in G1, S and G2/M phases with respect to the distinct (sub)clones is shown with barplots for each time point. Cell phases are estimated on 97 cell cycle genes via SCANPY. At time point t_1 and t_2 all cells in phase S belong to the PRAME$^{\text{MUT}}$ subclone. **(F)** A ternary plot showing the trajectories of the (sub)clonal subpopulations in the cell cycle space is shown. In the barycentric plot the three variables represent the ratio of cells belonging to phase G1, S and G2/M, respectively, and sum up to 1.

a heterogeneous ribonucleoprotein involved in mRNA processing and stability, and its regulation appears to be involved

in the PLK1-mediated P53 expression pathway, as it was shown via quantitative proteomic analysis on $BRAF^{V600/E}$ mutant melanoma cells treated with PLK1-specific inhibitor [37]. COL1A2 is an extracellular matrix protein that is supposed to maintain tissue integrity and homeostasis, and was identified as a melanoma marker [38], as well as a candidate prognostic factors in several cancer types [39]. CCT8 encodes the theta subunit of the CCT chaperonin, and was found as up-regulated or mutated in several cancer types [40].

After 4 days of therapy (t_1), the tumor volume halves. In particular, the RPL5^{MUT} subclone seems to disappear, whereas both the clonal subpopulation and the $\text{PRAME}^{\text{MUT}}$ subclone maintain a stable prevalence. A further reduction of tumor volume is observed after 28 days of treatment ($\sim 9\%$ of the volume at t_0), whereas at day 57 (t_3) a significant growth is observed, in which the tumor reaches $\sim 29\%$ of the initial volume, hinting at a possible relapse likely due to cells developing resistance. RPL5^{MUT} subclone reappears at time t_3 with a prevalence around $\sim 16\%$, suggesting that its absence at time points t_1 and t_2 may be due to sampling limitations, which however do not affect the capability of LACE of inferring a correct evolution model. At time t_3, the prevalence of all (sub)clonal subpopulations is similar to that of time t_0, hinting at the absence of significant clonal selection.

We then analyzed the composition of (sub)clones with respect to the cellular states identified in [26] via single-cell transcriptomics analysis (Figure 3C). Consistently with the findings in the article, the majority of cells in all (sub)clones are in a proliferative state before treatment, whereas at time point t_1 and t_2 no proliferative cells are left and all (sub)clones undergo transcriptional reprogramming, by displaying heterogeneous cell states (mostly starving-like melanoma cells, SMC), which result in acquired resistance at time t_3, when cells restart proliferating in all (sub)clones.

As only minor differences are observed among (sub)clones with respect to cellular states, we refined the transcriptomic analyisis, by first focusing on differentially expressed genes. The differential expression analysis, performed via standard ANOVA among all (sub)clones on all time points (data normalized by library size), allowed to identify 21 significant genes with FDR $p < 0.2$, among which only PRAME displays a FDR $p < 0.05$ and is significantly up-regulated in $\text{PRAME}^{\text{MUT}}$ cells, with $\log_2$-FC = 0.71 (see the Supplementary File 1 for the list of differentially expressed genes and the SI for further details on the analysis).

In order to analyze in depth the transition leading cells to resistance, we performed the same analysis with respect to the distinct time points. Interestingly, no genes are found as differentially expressed among (sub)clones at time t_0 and t_3. At time t_1 only 2 genes display a FDR $p < 0.2$, whereas at time t_2, 156 and 58 genes display a FDR $p < 0.2$ and < 0.1, respectively (see the Supplementary File 1). Within the latter group, 5 genes are significantly up-regulated in $\text{PRAME}^{\text{MUT}}$ cells and display a $\log_2$-FC larger than 3, namely NGLY1 ($\log_2$-FC $= 4.28$), CDCA7 ($\log_2$-FC $= 3.45$), HK1 ($\log_2$-FC $= 3.27$), DNAJB4 ($\log_2$-FC $= 3.27$), ISOC2 ($\log_2$-FC $= 3.11$; the distribution of gene expression values at time t_2 in $\text{PRAME}^{\text{MUT}}$ and PRAME^{WT} cell subpopulations is shown in Supplementary Figure 2). The results of this analysis suggest that, in addition to shifting their cellular states, distinct (sub)clones may differently respond to the therapy, and this would result in a transient increase of phenotypic heterogeneity at time t_1 and, especially, at time t_2.

This aspect is particularly evident by looking at the projection of single cells in the space of the 58 most significant differentially expressed genes (FDR $p < 0.1$), represented via diffusion maps [32] in Figure 3D. Before treatment (t_0), almost all cells are positioned in the left region of the map and appear to be highly intermixed, proving the existence of a homogeneous phenotypic behaviour. At time t_1, the RPL5^{MUT} subclone (yellow) disappears, whereas the clonal subpopulation (green) undertakes an apparent shift toward the right region of the map, which is characterized by transcriptional patterns that progressively diverge from those observed prior to the therapy, and which may possibly indicate high levels of cellular stress. This effect is notably amplified at time t_2, where an explicit split of the clonal and the $\text{PRAME}^{\text{MUT}}$ subpopulations can be observed, also in correspondence of the maximum dispersion of the cells on the map.

This outcome would further prove that distinct genetic clones may indeed suffer the effects of BRAF/MEK inhibition in different ways, during the resistance development phase, and this would result in different transcriptional patterns. At time point t_3, when cells have achieved resistance and restart proliferating, RPL5^{MUT} subclone expands, and all cells appear to be intermixed on the left portion of the diffusion map once again.

We analyzed the cell cycle phase of the single cells at different time points, as estimated on 97 cell cycle genes via SCANPY [33]. In Figure 3E one can see that cells are distributed across phases G1, S and G2/M in expected proportions in all subclones before therapy (t_0). Strikingly, at time point t_1 and t_2 all cells in phase S belong to subclone $\text{PRAME}^{\text{MUT}}$, whereas all cells of the clonal subpopulation are found in phase G1 or G2/M. At time t_3 the scenario resembles that of time t_0 and all (sub)clones include cells in all cell cycle phases. By looking at the ternary plot representing the proportion of cells in different cell cycle phases (Figure 3E), it is possible to notice that the clonal and the $\text{PRAME}^{\text{MUT}}$ subpopulations indeed undertake distinct trajectories in presence of therapy, before returning to a a state similar to the initial one.

This major result proves that the concurrent BRAF/MEK inhibition indeed affects in distinct ways different genetic clones during the resistance development stage. Apparently, cells lacking the PRAME mutation would be prevented from proceeding into S phase, whereas this effect would be highly mitigated in $\text{PRAME}^{\text{MUT}}$ cells. All in all, these results cast a new light on the relation between clonal evolution and phenotype at the single-cell level, and suggest that distinct genetic clones may respond to therapy in significantly different ways.

Discussion

Cancer is an evolutionary process in which cells progressively accumulate somatic variants and undergo selection, while competing in a complex microenvironment [41]. Such variants can be used to track the evolution of a single tumor, to characterize intra-tumor heterogeneity and to identify the genomic makeup of subclones responsible for therapy resistance or phenotypic switches [42].

The advent of single-cell sequencing has recently paved the way for high-resolution analyses of cancer evolution. We are currently entering a new phase in which longitudinal studies on cancer samples and on patient-derived cell cultures, xenografts and organoids are becoming available [43, 26]. This experimental shift is expected to improve our knowledge of cancer evolution, by allowing to explicitly consider the temporal scale at the single-cell level.

There are at least two major advantages in employing longitudinal instead of cross-sectional single-cell data. The first is that longitudinal data allows to dramatically improve the statistical robustness of the inference, and this is extremely relevant with highly noisy and often incomplete data, such as single-cell mutational profiles. The second is that only with an explicit temporal dimension it is possible to evaluate the clonal dynamics, the emergence or the extinction of subclones, especially with respect to the efficacy of therapeutic strategies.

LACE is the first algorithmic framework that can process single-cell datasets collected at different time points to produce longitudinal clonal trees of tumor evolution. Remarkably, our approach can explicitly model different error rates and sample size in distinct experiments, which is typical in longitudinal studies. Accordingly, it can leverage the information extracted from possibly biased or non-exhaustive samplings of the tumor's cells, which, instead, might lead to erroneous evolutionary inference when using single-time point datasets.

Moreover, as the results are noise-tolerant, LACE can deliver reliable results even with extremely imperfect mutational profiles, as those derived by calling variants from transcriptome. This allows to exploit transcriptomic data, commonly available for most single-cell studies, to assess the clonal composition and the history of a tumor, and to directly investigate for the first time the relation between genomic clonal evolution and phenotype at the single-cell level, for instance in response to a certain treatment.

Even though the application of longitudinal single-cell sequencing in clinical settings is still in its infancy, it is reasonable to expect a rapid diffusion of these techniques, for instance in the context of hematological clonal disorders, where cancer cells are readily accessible over time. The availability of tools dedicated to the analysis of longitudinal single-cell data will allow the study of the detailed molecular mechanisms triggered by the therapies directly in primary cancer cells. Notably, in many hematological clonal disorders, such as chronic myeloid leukemia, the leukemic stem cells and in particular the quiescent subset proved to be resistant even to targeted therapies such as Imatinib or second generation BCR-ABL1 inhibitors [44, 45]. Unfortunately, the nature of this resistance is still elusive, owing to the technical challenges involved in studying rare cell populations during treatment by using conventional approaches. In this scenario, LACE may be employed to characterize the resistance mechanisms selectively occurring in small subsets of the cancer cells pool.

As shown in the case study, the innovative features of LACE allow to deliver experimental hypotheses with translational relevance. On the one hand, our model produced a high-resolution picture of the evolutionary history of the tumor, as shaped by events such as the administration of a therapy. Furthermore, LACE allowed to identify (sub)clones that show different sensitivity to the therapy. In particular, we detected an unexpected behavior of the cell cycle machinery in different (sub)clones upon treatment, with only $\text{PRAME}^{\text{MUT}}$ (sub)clone maintaining a fraction of cells in S phase. These findings suggest that longitudinal single-cell analyses are effective in dissecting the mechanisms by which cancer cells react upon treatment, at least for clonal disorders where tumor cells are readily accessible.

All in all, by explicitly allowing a mapping between the clonal evolution and the phenotypic properties of single cells, LACE proved to be a powerful and expressive tool to decipher intra-tumor heterogeneity on multiple scales.

Our framework might be easily extended to account for violations of the Infinite Sites Assumption, as proposed, e.g., in [15, 46, 47], and by possibly leveraging the information on copy-number alterations. Furthermore, as both bulk and single-cell sequencing data may be increasingly available in longitudinal studies, the integration of both data types within our framework might allow to improve the clone identification and the inference quality, as proposed in [48, 25].

Finally, once a significant number of longitudinal single-cell datasets would be available on specific cancer types, techniques based on transfer learning might be applied to our models to identify possible patterns of recurrent evolution across tumors, as proposed by some of the authors in [49].

Methods

Input data preprocessing. LACE requires a distinct input data matrix including single-cell mutational profiles for any time point or experiment. As we are interested in somatic evolution of cancer subpopulations, mutational profiles may include, for instance, single-nucleotide variants (SNVs) or structural variants.

Variant calling can be performed from DNA sequencing data – either whole-genome/exome or targeted sequencing – but also from transcriptomic data, such as scRNA-seq (see the SI for further details). The latter represents a cost-effective and highly-available alternative, despite known technological issues, such as the presence of reads encompassing intronic regions and the impossibility of calling variants from non-transcribed regions. In fact, LACE is robust with respect to high levels of noise, also thanks to the phylogenetic constraints implied by the process of accumulation of somatic mutations.

Notice that, as any statistical inference method, the reliability of the results is higher when the sampling bias is limited, as single-cell samplings should be ideally significant of the tumor's composition at any time point. However, by exploiting the information extracted from multiple datasets sampled from the same evolutionary history, LACE can deliver statistically significant output models even with possibly biased, incomplete or imperfect samplings, as opposed to standard methods for single-time point data.

In our framework, a (sub)clone is defined as a subset of cells sharing the same set of driver alterations, i.e., cells with different genotypes may indeed belong to the same (sub)clone. As it is difficult to know a priori which alterations are drivers for a given tumor, at least by relying on data of single patients, LACE requires to select a set of candidate driver mutations prior to the inference.

To this end, the presence of known variants involving oncogenes and tumor suppressor genes should be first verified. Previously uncharacterized mutations might be also selected, if significantly present in the samples. In both cases, filters on statistical significance (e.g., recurrence thresholds) and on clinical/functional features should be employed, in order to reduce the impact of noise of single-cell measurements and to exclude non functional variants (e.g., rare polymorphisms). Furthermore, it might be also sound to cluster the mutations selected after filtering, on the basis of co-occurrence patterns across single-cells, even in distinct experiments, as proposed with a different scope in [25].

In any case, however, the resulting list of candidate drivers may include false positives, i.e., alterations unrelated to tumor progression, such as passengers. For this reason, in real-world analyses one should prudently refer to the alterations in the list as *clonal footprints*. In fact, the inference of the clonal architecture and evolution of the tumor is preserved when clonal footprints include non-drivers, as LACE simply relies on the existence of a consistent mutation accumulation process. Yet, in the interpretation of the output model one should remind that some of the candidate (sub)clones may include non-drivers; in this case, it may be also sound to examine subgraphs in the output tree, which might represent a suitable resolution to describe the clonal composition of the tumor. Clearly, the trade-off between expected false positives (i.e., non-drivers included in the list) and false negatives (i.e., drivers non included in the list) should guide the user toward either stricter or looser criteria for candidate driver selection.

Single-cell data factorization problem (single time point). Let us consider k different clones, m observed *putative drivers* (i.e., *clonal footprints*) and n single cells $c_1, \ldots, c_n$ sampled in a given experiment. We can then define the following matrices:

1. The *single-cell data matrix* $\mathbf{D}$: a binary $n \times m$ matrix where each row represents a single cell and each column a mutation; an element of $\mathbf{D}$, $d_{i,j} = 1$ if we observe mutation j in cell i, otherwise $d_{i,j} = 0$.
2. The *phylogenetic matrix* $\mathbf{B}$: a binary $k \times m$ matrix where each row represents a clone and each column a mutation; $b_{ij} = 1$ if we observe mutation j in clone i, otherwise $b_{i,j} = 0$. Each $\mathbf{B}$ can uniquely be represented by a tree and vice versa [50, 28]. Moreover, if we assume that the phylogenetic process is *perfect* and that the Infinite Site Assumption (ISA) holds [51], i.e., mutations are never lost and there is only one root. $\mathbf{B}$ has the following properties: (i) $\mathbf{B}$ is a square matrix ($k = m$), i.e., the number of clones is equal to the number of mutations; (ii) the rank of $\mathbf{B}$ is k; (iii) the Hamming distance between any pair of rows of $\mathbf{B}$ is ≥ 1 and there is a column of $\mathbf{B}$ where all entries are 1; (iv) $\mathbf{B}$ is a lower triangular matrix and all the elements of its diagonal are equal to 1 [28].
3. The *cell attachment matrix* $\mathbf{C}$: a binary $n \times k$ matrix, where each row represents a single cell and each column a clone; $c_{i,j} = 1$ if cell i is associate to clone j, otherwise $c_{i,j} = 0$. We notice that each cell is attached exactly

to one clone, i.e., the sum of any row of $\mathbf{C}$ is equal to 1. Note that this allows to easily compute the clonal prevalence.

On these premises and by assuming that the number of sampled cells is larger than the number of clonal footprints, i.e., $n > m$, we can state the following proposition.

Proposition 1 *For every single-cell matrix data* $\mathbf{D}$ *the following factorization holds:*

$$\mathbf{D} = \mathbf{C} \cdot \mathbf{B}. \tag{1}$$

Its existence is guaranteed by construction, whereas the proof of uniqueness is straightforward.

Likelihood function definition (single time point). The factorization defined in Eq. (1) may not hold if $\mathbf{D}$ includes false positives, false negatives and missing values, which is typical in real-world scenarios. For this reason standard phylogenetic methods needs to be extended by modeling error rates, i.e., α as false positive rate and β as false negative rate, as proposed for instance in [13, 14].

LACE aims at maximizing the posterior probability of the output model (i.e., $\mathbf{C}, \mathbf{B}$), given a noisy dataset $\mathbf{D}$. By applying Bayes rule, the posterior probability can be written as:

$$P(\mathbf{B}, \mathbf{C}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{C}, \mathbf{B})P(\mathbf{B}, \mathbf{C})}{P(\mathbf{D})}. \tag{2}$$

As typically done, we assume that $P(\mathbf{D})$ is identical for all models, thus:

$$P(\mathbf{B}, \mathbf{C}|\mathbf{D}) \propto P(\mathbf{D}|\mathbf{C}, \mathbf{B})P(\mathbf{B}, \mathbf{C}). \tag{3}$$

In the current implementation of LACE, we assume a uniform and uninformative prior on the phylogenetic matrix $\mathbf{B}$ and the cell attachment $\mathbf{C}$. Whether knowledge on the underlying biological phenomenon would be available, our framework could be directly extended to include priors for $P(\mathbf{B}, \mathbf{C})$.

Let us define the estimated genotype matrix $\mathbf{G} = \mathbf{C} \cdot \mathbf{B}$, which is subsumed by the LACE model. Therefore, the problem is reduced to the maximization of the following likelihood function:

$$P(\mathbf{D}|\mathbf{B}, \mathbf{C}) = \prod_{i=1}^{n} \prod_{j=1}^{m} P(d_{i,j}|\mathbf{G}_{i,j}), \tag{4}$$

where $d_{i,j}$ is a entry of $\mathbf{D}$ and:

$$P(d_{i,j}|\mathbf{G}_{i,j}) = \begin{cases} \alpha, & if \quad d_{i,j} = 1 \quad and \quad \mathbf{G}_{i,j} = 0, \\ 1 - \alpha, & if \quad d_{i,j} = 1 \quad and \quad \mathbf{G}_{i,j} = 1, \\ \beta, & if \quad d_{i,j} = 0 \quad and \quad \mathbf{G}_{i,j} = 1, \\ 1 - \beta, & if \quad d_{i,j} = 0 \quad and \quad \mathbf{G}_{i,j} = 0, \\ 1, & if \quad d_{i,j} = NA, \end{cases} \tag{5}$$

with α being the false positive rate, β the false negative rate, and NA labeling a missing entry. Clearly, in this case the maximum a posteriori solution is equivalent to the maximum likelihood solution. Notice that α and β can be provided as input to LACE, or they can be estimated (see below). We refer to the SI for details and discussions on how to account for high levels of noise in the input data by marginalizing the cell attachment.

Longitudinal single-cell data factorization problem. We now generalize the problem to the case of y experiments taken at different time points $t_1 \leq t_2 \leq \cdots \leq t_y$. For every experiment comprising $n_1, \ldots, n_y$ different single cells and $m_1, \ldots, m_y$ mutations, respectively, we can construct y independent data matrices $\mathbf{D}_1, \ldots, \mathbf{D}_y$ as defined in (1).

As we assume that the longitudinal experiments are sampled from a unique generative phylogenetic matrix $\mathbf{B}$, we need to expand the input matrices $\mathbf{D}_1, \ldots, \mathbf{D}_y$ and $\mathbf{C}_1, \ldots, \mathbf{C}_y$ in order to include the union $\widetilde{m}$ of all the mutations detected in at least one cell over all the experiments.

Let $\widetilde{\mathbf{D}}_s$ and $\widetilde{\mathbf{C}}_s$ be the expanded matrices for s-th time point, we define the longitudinal single-cell data factorization problem as follow:

$$\widetilde{\mathbf{D}}_s = \widetilde{\mathbf{C}}_s \cdot \mathbf{B} \qquad s = 1, 2, \ldots, y. \tag{6}$$

Notice that the formulation of an analogous problem for longitudinal bulk sequencing data was introduced in [19].

Weighted likelihood function for longitudinal single-cell data. Since we are interested in learning a unique clonal tree $\mathbf{B}$ from different longitudinal datasets, we here define a weighted likelihood function as follows:

$$P(\widetilde{\mathbf{D}}_\mathbf{1}, \ldots, \widetilde{\mathbf{D}}_y | \mathbf{B}, \widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y) = \prod_{s=1}^{y} P(\widetilde{\mathbf{D}}_s | \mathbf{B}, \widetilde{\mathbf{C}}_s)^{w_s}, \qquad s = 1, 2, \ldots, y, \tag{7}$$

where $P(\widetilde{\mathbf{D}}_s | \mathbf{B}, \widetilde{\mathbf{C}}_s)$ can be computed as for Eq. (4) and w_s are *weights* aimed at modelling possible idiosyncrasies of multiple longitudinal experiments, e.g., due to possible differences in quality and/or in the number of sampled cells. The definition of a weighted likelihood function allows us to explicitly account for experimental and technological differences among experiments collected at distinct time points and represents one of the major novelties of our approach.

The choice of appropriate weights is problem- and data-specific and can benefit from a broad literature in statistical inference (see, e.g., [52]). In general, uniform weights would bias the solution toward datasets with larger sample sizes. For this reason, if no prior is available on the quality of the single experiments, we suggest as default weight for the s-th dataset composed by n_s cells: $w_s = (1 - \frac{n_s}{n_T})/(y-1)$, where n_T is the total number of cells of all the experiments, and $y \geq 2$ is the number of experiments.

Finally, we define the log-likelihood objective function as follows:

$$\ln(P(\widetilde{\mathbf{D}}_\mathbf{1}, \ldots, \widetilde{\mathbf{D}}_y | \mathbf{B}, \widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y)) = \sum_{s=1}^{y} w_s \ln(P(\widetilde{\mathbf{D}}_s | \mathbf{B}, \widetilde{\mathbf{C}}_s)), \qquad s = 1, 2, \ldots, y. \tag{8}$$

We notice that the values of α and β might be even extremely different among datasets – e.g., due to technological features of the experiments. LACE can explicitly model different error rates, which will be indicated as α_s and β_s for the s-th experiment.

Error rates estimation. When error rates α_s and β_s are unknown, LACE includes a noise estimation procedure. By assuming that error rates are independent both from the phylogenetic matrix $\mathbf{B}$ and the cell attachments $\widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y$, we can extend Eq. (7):

$$\begin{aligned} &P(\mathbf{B}, \widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y, \alpha_1, \ldots, \alpha_y, \beta_1, \ldots, \beta_y | \widetilde{\mathbf{D}}_\mathbf{1}, \ldots, \widetilde{\mathbf{D}}_y) \propto \\ &P(\widetilde{\mathbf{D}}_\mathbf{1}, \ldots, \widetilde{\mathbf{D}}_y | \mathbf{B}, \widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y) P(\alpha_1, \ldots, \alpha_y, \beta_1, \ldots, \beta_y) P(\mathbf{B}, \widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y). \end{aligned} \tag{9}$$

Finally, by assuming $P(\mathbf{B}, \widetilde{\mathbf{C}}_\mathbf{1}, \ldots, \widetilde{\mathbf{C}}_y)$ and $P(\alpha_1, \ldots, \alpha_y, \beta_1, \ldots, \beta_y)$ to be uniform, our problem can be reduce to solving Eq. (8), given fixed values of α and β. Therefore, the optimal values for $\alpha_1, \ldots, \alpha_y$ and $\beta_1, \ldots, \beta_y$ can be directly estimated by performing a parameter scan (see below).

Search Scheme. LACE's output model is composed by three components: the clonal tree $\mathbf{B}$, the cell attachment matrices $\widetilde{\mathbf{C}}_s$, and (optionally) the estimated error rates α_s and β_s, with $s = 1, \ldots, y$, where y is the number of time points.

The search space of the possible solutions is huge, in fact given $\widetilde{m}$ mutations (i.e., the union the mutations occurring at least once in all experiments) it includes continuous terms for α_s and β_s, a discrete term of dimension $\frac{(2\widetilde{m}-3)!}{(\widetilde{m}-2)!2^{\widetilde{m}-2}}$ for $\mathbf{B}$, and another discrete term of dimension $\prod_{s=1}^{y} n_s^{\widetilde{m}}$ for $\widetilde{\mathbf{C}}_s$.

As an exhaustive search can be achieved only for very small models, LACE employs a Markov Chain Monte Carlo (MCMC) scheme to sample from the joint posterior distribution given the input data. In particular, the MCMC includes two ergodic moves on $\mathbf{B}$: (i) node relabeling, (ii) prune and reattach of a single node and its descendants. For each

proposed configuration $\mathbf{B}'$, we find the maximum likelihood $\widetilde{\mathbf{C}}'$ via an exhaustive search (the default probability for selecting move (i) or move (ii) is equal to 0.5).

Thus, given $\widetilde{\mathbf{C}}'$ and by assuming that all the $\mathbf{B}$ are equally probable, the acceptance ratio ρ is given by:

$$\rho_{\mathbf{B}',\widetilde{\mathbf{C}}'_{\mathbf{1}},\ldots,\widetilde{\mathbf{C}}'_{y}} = \min\left\{\left(\frac{P(\widetilde{\mathbf{D}}_{\mathbf{1}},\ldots,\widetilde{\mathbf{D}}_{y}|\mathbf{B}',\widetilde{\mathbf{C}}'_{\mathbf{1}},\ldots,\widetilde{\mathbf{C}}'_{y})}{P(\widetilde{\mathbf{D}}_{\mathbf{1}},\ldots,\widetilde{\mathbf{D}}_{y}|\mathbf{B},\widetilde{\mathbf{C}}_{\mathbf{1}},\ldots,\widetilde{\mathbf{C}}_{y})}\right)^{1/T},1\right\}, \tag{10}$$

where T is a learning rate parameter, which could be used to speed up convergence as proposed in [13] (default value $T = 1$). The move is accepted if it results in a higher weighted likelihood, otherwise it is rejected. With proper move probabilities, acceptance ratio and an infinite number of moves, the MCMC is ensured to converge to the posterior [13].

In the current implementation of LACE we decided to exclude α_s and β_s from the MCMC, as they are continous variables and this would dramatically increase the computational burden of the search. Therefore, LACE allows to perform a grid search on such parameters, by running multiple parallel MCMC searches with fixed error rates for each time point. The pseudocode of the algorithm can be found in the SI.

Synthetic data generation. To generate synthetic datasets we employed the tool from [29], which simulates a branching process modeling the population dynamics of cancer subpopulations, characterized by the accumulation of random mutations, which can either be drivers – i.e., inducing a certain proliferative advantage –, or passengers – i.e., with no effect. The simulator eventually returns the list of existing genotypes (including both passengers and drivers) and the relative prevalence at any time point, where a single time step corresponds to a generation (i.e., a replication event – see the SI for further details and the parameter settings).

We selected 15 simulation scenarios in which a number of drivers between 7 and 15 was observed, from which we sampled a large number of independent longitudinal single-cell mutational profile datasets including the drivers only, on three distinct time points, i.e., $t = 50, 150$ and 300. Single cells were randomly sampled with a probability proportional to cellular prevalence and we finally inflated the resulting binary mutational profiles with various rates of: false positives, α, false negatives, β and missing entries γ.

Mutational profiling from scRNA-seq data. LACE requires longitudinal single-cell mutational profiles, as computed on a panel of candidate drivers (also named clonal footprints, see above). Such profiles might be optimally derived by whole-genome and whole-exome single-cell sequencing experiments, which allow to call both known and uncharacterized somatic mutations, or by genotyping single cells, when target gene panels are available. In both cases, however, genome sequencing at the single-cell level is a still expensive and error-prone option, mostly related to technical issues in cell isolation and genome amplification [53].

Conversely, single-cell RNA-seq data are commonly used for an increasingly larger number of research goals and can be effectively used to call variants on exonic regions, even at the single-cell level [54]. It is known that scRNA-seq data are affected by different sources of noise [55] and cannot be used to call variants in non transcribed regions. However, we have shown that the results of LACE are noise-tolerant, also due to the phylogenetic constraints implied by the process of accumulation of mutations (see Fig. 2) and, therefore, it can be efficiently applied to possibly incomplete or noisy mutational profiles.

In particular, we here applied LACE to a longitudinal scRNA-seq dataset originally analysed in [26]. In the study, a number of PDXs were derived from $\text{BRAF}^{\text{V600E/K}}$ mutant melanoma patients and were treated with concurrent BRAF/MEK-inhibition. In our analysis, we selected PDX MEL006, for which four temporally-ordered scRNA-seq datasets are available: (i) pre-treatment, (ii) after 4 days of treatment, (iii) after 28 days of treatment, (iv) after 57 days of treatment. In the study, whole transcriptome amplification was made with a modified SMART-seq2 protocol and libraries preparation were performed using the Nextera XT Illumina kit. Samples were sequenced on the Illumina NextSeq 500 platform, by using 75bp single-end reads. Low-quality cells were filtered-out based on library size, number of genes expressed per cell, ERCCs, house-keeping gene expression and mitochondrial DNA reads, and a total of 674 cells was finally included in the dataset [26].

We applied further filters to remove cells displaying a fraction of counts on mitochondrial genes larger than 20% and cells displaying outlier values with respect to library size. As a result, we selected 475 single cells for downstream analysis. In order to call SNVs and indels from such dataset, we employed the GATK Best Practices [30], which are proven to be effective even with single-cell data [54] (see the SI for the parameter settings). A VCF file including 272674 unique variants was generated and subsequently annotated with Annovar [56].

A first filtering step was applied to discard low-quality or non functional variants. First, synonymous and unknown mutations were removed (196320 unique mutations left); second, variants observed in less than two reads in each single

cell were filtered-out (195931 unique mutations left); third, we employed a threshold of 1% on minor allele frequency, to remove possible germline mutations, as no normal tissue was included in the study (191599 unique mutations left).

A further filtering step was then employed to identify the list of putative drivers/clonal footprints to be used in LACE. We first selected the mutations showing a frequency greater than 5% in at least one time point (595 unique mutations left). We then marked as missing entries (i.e., NA) the variants in a position with coverage lower than 3, as they might be miscalled due to gene expression down-regulation. We kept the mutations displaying less than 40% of missing data on all time points (151 unique mutations left), a median coverage larger than 10 and a median alternative read count larger than 4 (82 unique mutations left). As a final step, we manually curated the list of 82 remaining variants, to verify the possible presence of errors due to amplification (i.e., strand slippages) or alignment artifacts.

As a result, we selected the following 6 clonal footprints to be provided as input to LACE: ARPC2 (chr2:218249894, C>T, nonsynonymous substitution), CCT8 (chr21:29063389, G>A, nonsyn.), COL1A2 (chr7:94422978, C>A, nonsyn.), HNRNPC (chr14:21211843, C>T, nonsyn.), PRAME (chr22:22551005, T>A, stop-gain), RPL5 (chr1:92837514, C>G, nonsyn.). The oncoprint including the mutational profiles of the single cells is displayed in Fig. 1 of the SI.

Data availability

LACE is available as an open source R tool at: `https://github.com/BIMIB-DISCo/LACE`. The scRNA-seq datasets used in our analyses were downloaded from GEO: `https://www.ncbi.nlm.nih.gov/geo/`, accession code: `GSE116237`. The source code used to replicate all our analyses, including synthetic and real datasets, is available at this link: `https://github.com/BIMIB-DISCo/LACE-UTILITIES`.

Acknowledgements

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Università e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures by the AIRC-IG grant 22082. Support was also provided by the CRUK/AIRC Accelerator Award #22790, "Single-cell Cancer Evolution in the Clinic". We thank Giulio Caravagna, Chiara Damiani, Francesco Craighero and Lucrezia Patruno for helpful discussions.

Competing Interests

The authors declare that they have no competing financial interests.

Contributions

D.R., F.A., D.M. and A.G. designed the approach, defined the method and implemented it. D.R. and A.G. performed the simulations. D.R., D.M., G.A., R.P. and A.G. executed the experimental data analysis pipeline. D.R., F.A., D.M., I.C., R.P., M.A. and A.G. analyzed the data and interpreted the results. A.G. and D.R. supervised the study. All authors discussed, drafted and approved the manuscript.

References

[1] Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).

[2] Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14**, 618 (2013).

[3] Gillies, R. J., Verduzco, D. & Gatenby, R. A. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nature Reviews Cancer* **12**, 487–493 (2012).

[4] Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).

[5] Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

[6] Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences* **110**, 4009–4014 (2013).

[7] Jackson, H. J., Rafiq, S. & Brentjens, R. J. Driving car t-cells forward. *Nature reviews Clinical oncology* **13**, 370 (2016).

[8] Goodspeed, A., Heiser, L. M., Gray, J. W. & Costello, J. C. Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research* **14**, 3–13 (2016).

[9] Clevers, H. Modeling development and disease with organoids. *Cell* **165**, 1586–1597 (2016).

[10] Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

[11] Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods* **12**, 519 (2015).

[12] Nam, A. S. *et al.* Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature* **571**, 355–360 (2019).

[13] Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome biology* **17**, 1 (2016).

[14] Ross, E. M. & Markowetz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome biology* **17**, 1 (2016).

[15] Zafar, H., Tzen, A., Navin, N., Chen, K. & Nakhleh, L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology* **18**, 178 (2017).

[16] Ramazzotti, D., Graudenzi, A., De Sano, L., Antoniotti, M. & Caravagna, G. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC bioinformatics* **20**, 210 (2019).

[17] Siravegna, G. *et al.* Radiologic and genomic evolution of individual metastases during HER2 blockade in colorectal cancer. *Cancer Cell* **34**, 148–162 (2018).

[18] Khan, K. H. *et al.* Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial. *Cancer discovery* **8**, 1270–1285 (2018).

[19] Myers, M. A., Satas, G. & Raphael, B. J. Calder: Inferring phylogenetic trees from longitudinal tumor samples. *Cell Systems* **8**, 514 – 522.e5 (2019).

[20] Alves, J. M., Prieto, T. & Posada, D. Multiregional tumor trees are not phylogenies. *Trends in cancer* **3**, 546–550 (2017).

[21] Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2017).

[22] Schachtner, R., Pöppel, G., Tomé, A. & Lang, E. From binary NMF to variational bayes NMF: A probabilistic approach. In *Non-negative Matrix Factorization Techniques*, 1–48 (Springer, 2016).

[23] Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology* **15** (2019).

[24] Liu, F. *et al.* Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome biology* **20**, 1–15 (2019).

[25] Salehi, S. *et al.* ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome biology* **18**, 44 (2017).

[26] Rambow, F. *et al.* Toward minimal residual disease-directed therapy in melanoma. *Cell* **174**, 843–855 (2018).

[27] McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–1303 (2010).

[28] El-Kebir, M., Oesper, L., Acheson-Field, H. & Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* **31**, i62–70 (2015).

[29] El-Kebir, M., Satas, G. & Raphael, B. J. Inferring parsimonious migration histories for metastatic cancers. *Nature genetics* **50**, 718 (2018).

[30] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491 (2011).

[31] Smith, M. A. *et al.* E-scape: interactive visualization of single-cell phylogenetics and cancer evolution. *Nature methods* **14**, 549 (2017).

[32] Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

[33] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

[34] Kashani-Sabet, M. *et al.* A multi-marker assay to distinguish malignant melanomas from benign nevi. *Proceedings of the National Academy of Sciences* **106**, 6268–6272 (2009).

[35] Orlando, D. *et al.* Adoptive immunotherapy using prame-specific t cells in medulloblastoma. *Cancer research* **78**, 3337–3349 (2018).

[36] Pelletier, J., Thomas, G. & Volarević, S. Ribosome biogenesis in cancer: new players and therapeutic avenues. *Nature Reviews Cancer* **18**, 51 (2018).

[37] Cholewa, B. D., Pellitteri-Hahn, M. C., Scarlett, C. O. & Ahmad, N. Large-scale label-free comparative proteomics analysis of polo-like kinase 1 inhibition via the small-molecule inhibitor bi 6727 (volasertib) in brafv600e mutant melanoma cells. *Journal of proteome research* **13**, 5041–5050 (2014).

[38] Koga, Y. *et al.* Genome-wide screen of promoter methylation identifies novel markers in melanoma. *Genome research* **19**, 1462–1470 (2009).

[39] Li, J., Ding, Y. & Li, A. Identification of col1a1 and col1a2 as candidate prognostic factors in gastric cancer. *World journal of surgical oncology* **14**, 297 (2016).

[40] Huang, X. *et al.* Chaperonin containing tcp 1, subunit 8 (cct 8) is upregulated in hepatocellular carcinoma and promotes hcc proliferation. *Apmis* **122**, 1070–1079 (2014).

[41] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).

[42] Caravagna, G. *et al.* Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences* **113**, E4025–E4034 (2016).

[43] Sharma, A. *et al.* Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. *Nature communications* **9**, 4931 (2018).

[44] Perl, A. & Carroll, M. Bcr-abl kinase is dead; long live the cml stem cell. *The Journal of clinical investigation* **121**, 22–25 (2011).

[45] Kinstrie, R. *et al.* Cd93 is expressed on chronic myeloid leukemia stem cells and identifies a quiescent population which persists after tyrosine kinase inhibitor therapy. *Leukemia* 1–13 (2020).

[46] El-Kebir, M. Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics* **34**, i671–i679 (2018).

[47] Bonizzoni, P., Ciccolella, S., Della Vedova, G. & Gomez, M. S. Does relaxing the infinite sites assumption give better tumor phylogenies? an ILP-based comparative approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2018).

[48] Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C. & Beerenwinkel, N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications* **10**, 2750 (2019).

[49] Caravagna, G. *et al.* Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods* **15**, 707 (2018).

[50] Gusfield, D. Efficient algorithms for inferring evolutionary trees. *Networks* **21**, 19–28 (1991).

[51] Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893 (1969).

[52] Hu, F. & Zidek, J. V. The relevance weighted likelihood with applications. In *Empirical Bayes and Likelihood Inference*, 211–235 (Springer, 2001).

[53] Navin, N. E. & Chen, K. Genotyping tumor clones from single-cell data. *Nature Methods* **13**, 555–556 (2016).

[54] Schnepp, P. M., Chen, M., Keller, E. T. & Zhou, X. SNV identification from single-cell RNA sequencing data. *Human Molecular Genetics* (2019).

[55] Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature communications* **6**, 8687 (2015).

[56] Wang, K., Li, M. & Hakonarson, H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164–e164 (2010).