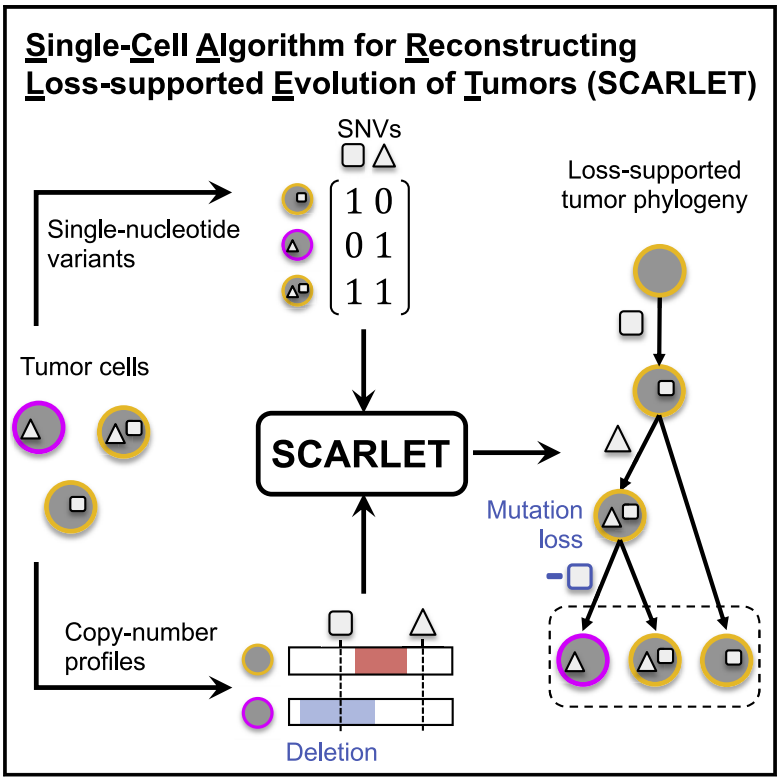
Cell Systems

SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses

Graphical Abstract



Highlights

- Single-nucleotide variants (SNVs) and CNAs are markers of cancer evolution
- Copy-number aberrations (CNAs) may overlap SNVs and result in SNV loss
- Loss-supported model constrains SNV losses to loci with a decrease in copy number
- SCARLET integrates SNVs and CNAs yielding more accurate single-cell phylogenies

Authors

Gryte Satas, Simone Zaccaria, Geoffrey Mon, Benjamin J. Raphael

Correspondence

braphael@princeton.edu

In Brief

Both single-nucleotide variants (SNVs) and copy-number aberrations (CNAs) accumulate during cancer evolution, and these mutations may overlap on the genome. We introduce SCARLET (single-cell algorithm for reconstructing loss-supported evolution of tumors), an algorithm to construct phylogenies from single-cell DNA sequencing data using both SNVs and CNAs.

Satas et al., 2020, Cell Systems *10*, 323–332
April 22, 2020
https://doi.org/10.1016/j.cels.2020.04.001

CellPress

Cell Systems
Focus on RECOMB

CellPress
OPEN ACCESS

Methods

SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses

Gryte Satas,[1,2] Simone Zaccaria,[2] Geoffrey Mon,[2] and Benjamin J. Raphael[2,3,*]
[1]Department of Computer Science, Brown University, Providence, RI 02912, USA
[2]Department of Computer Science, Princeton University, Princeton, NJ 08540, USA
[3]Lead Contact
*Correspondence: braphael@princeton.edu
https://doi.org/10.1016/j.cels.2020.04.001

SUMMARY

A small number of somatic mutations drive the development of cancer, but all somatic mutations are markers of the evolutionary history of a tumor. Prominent methods to construct phylogenies from single-cell-sequencing data use single-nucleotide variants (SNVs) as markers but fail to adequately account for copy-number aberrations (CNAs), which can overlap SNVs and result in SNV losses. Here, we introduce SCARLET, an algorithm that infers tumor phylogenies from single-cell DNA sequencing data while accounting for both CNA-driven loss of SNVs and sequencing errors. SCARLET outperforms existing methods on simulated data, with more accurate inference of the order in which mutations were acquired and the mutations present in individual cells. Using a single-cell dataset from a patient with colorectal cancer, SCARLET constructs a tumor phylogeny that is consistent with the observed CNAs and suggests an alternate origin for the patient's metastases. SCARLET is available at: github.com/raphael-group/scarlet.

INTRODUCTION

Cancer arises from an evolutionary process during which somatic mutations accumulate in a population of cells. Different cells within a tumor acquire distinct complements of somatic mutations, resulting in a heterogeneous tumor. Quantifying this intra-tumor heterogeneity and reconstructing the evolutionary history of a tumor is crucial for diagnosis and treatment of cancer (Burrell et al., 2013; Tabassum and Polyak, 2015). The evolution of a tumor is typically described by a phylogenetic tree or phylogeny, whose leaves represent the cells observed at the present time and whose internal nodes represent ancestral cells (see Box 1). Tumor phylogenies are challenging to reconstruct using DNA sequencing data from bulk tumor samples, since these data contain mixtures of mutations from thousands to millions of heterogeneous cells in the sample (Jiao et al., 2014; El-Kebir et al., 2015, 2016 Malikic et al., 2015; Popic et al., 2015; Deshwar et al., 2015; Jiang et al., 2016; Alves et al., 2017; Satas and Raphael, 2017; Pradhan and El-Kebir, 2018; Zaccaria et al., 2018; Miura et al., 2019; Myers et al., 2019). Recently, single-cell DNA sequencing (scDNA-seq) of tumors has become more common, and new technologies, such as those from 10 X Genomics (10X Genomics, 2018), Mission Bio (Mission Bio, 2019), and others (Gawad et al., 2016; Zahn et al., 2017; Navin, 2015) are improving the efficiency and lowering the costs of isolating, labeling, and sequencing individual cells. While scDNA-seq overcomes the difficulties of phylogeny reconstruction from bulk samples, it introduces a new challenge of higher rates of missing data and errors due to DNA amplification errors, undersampling, and sequencing errors (Gawad et al., 2016).

Early work in phylogeny inference from scDNA-seq data uses single-nucleotide variants (SNVs) as phylogenetic markers. A particular challenge for SNV-based analysis is high rates (up to 30% for high-depth scDNA-seq; Gawad et al., 2016) of allele dropout errors, where only one of two alleles is observed at a heterozygous site. Methods address this challenge by using an evolutionary model to infer a phylogeny while simultaneously imputing missing data and correcting errors in the observed SNVs. Algorithms such as SCITE (Jahn et al., 2016), OncoNEM (Ross and Markowetz, 2016), SciΦ (Singer et al., 2018), and B-SCITE (Malikic et al., 2019a, 2019b) use the simplest phylogenetic model for SNVs—the infinite-sites model. In this model, a locus in a cell has one of two states: an SNV (or mutation) is either present at the locus (state 1) or absent (state 0). Transitions between states are constrained in the phylogeny such that each mutation is gained ($0 \rightarrow 1$) at most once during evolution, and never subsequently lost ($1 \rightarrow 0$). A phylogeny that respects the infinite-sites model is known as a perfect phylogeny, and the state of mutations in the leaves of the phylogeny is summarized by a mutation matrix, whose binary entries indicate the presence (state 1) or absence (state 0) of every mutation in each observed cell (Figure 1A). On error-free data, the perfect phylogeny is unique (Gusfield, 1991). However, on typical scDNA-seq data, errors in the mutation matrix must be corrected to yield a perfect phylogeny model. Because many such corrections are possible, multiple phylogenies are typically equally consistent with the data (Figure 1B). An additional challenge in inferring

Box 1. Primer

Cancer is an evolutionary process where cells in a tumor accumulate somatic mutations over time. While only a small number of these somatic mutations drive the development of cancer, all somatic mutations are a marker of the evolutionary history of a tumor. Recent scDNA-seq technologies enable the measurement of somatic mutations in individual cells from a tumor, providing data to construct a phylogenetic tree, or phylogeny, that represents the past evolution of the tumor.

Constructing a phylogenetic tree that describes the ancestral relationships between cells in a tumor relies on a choice of markers, or characters, that distinguish the individual cells as well as an evolutionary model describing how these markers change over time. For scDNa-seq of tumors, a popular choice of markers is single-nucleotide differences between cancer cells, known as SNVs. However, current scDNA-seq technologies measure SNVs with high rates of missing data and errors due to technical limitations such as DNA amplification artifacts, undersampling, and sequencing errors. Standard phylogenetic methods do not handle such high rates of missing data and errors. Thus, specialized algorithms have been developed to construct phylogenetic trees from single-cell measurements of SNVs. Early works used the simplest evolutionary model for SNVs, the infinite-sites model, where a position in the genome is mutated at most once.

However, SNVs are not the only type of somatic mutation that occur in cancer. In particular, most solid tumors have many CNAs, mutations that duplicate or delete segments of the genome that range in scale from hundreds of nucleotides through whole chromosomes. CNAs often overlap SNVs; for example, a deletion may remove SNVs. The infinite-sites model does not allow loss of SNVs, and thus methods that use this model do not accurately reconstruct the phylogenetic trees of tumors with many CNAs. More general evolutionary models that allow loss of SNVs, or mutation losses, have recently been used in single-cell phylogenetic analysis, such as the Dollo and finite-sites models. However, these models do not examine the underlying DNA sequencing data for evidence of CNAs. Thus, such models are generally too permissive, admitting many different phylogenies even when these contradict the observed CNA data.

In this paper, we introduce a loss-supported evolutionary model that allows SNV losses only when accompanied by evidence in the DNA sequencing data of a deletion at the same locus. We use this loss-supported model as the basis for an algorithm, SCARLET (single-cell algorithm for reconstructing loss-supported evolution of tumors), that infer tumor phylogenies from scDNA-seq data, accounting for both mutation loss and sequencing errors. We show that SCARLET infers single-cell tumor phylogenies more accurately than existing methods.

phylogenies from cancer sequencing data is that somatic mutations in tumors occur across all genomic scales from SNVs to copy-number aberrations (CNAs), which amplify or delete larger genomic regions. CNAs may overlap SNVs and affect the state of SNVs in cells; e.g., a deletion that overlaps an SNV may result in a mutation loss ($1 \rightarrow 0$). The infinite-sites model does not allow mutation losses and therefore may yield incorrect phylogenies when applied to SNVs in regions containing CNAs. One solution is to exclude regions containing CNAs and build phylogenies from SNVs in diploid or copy-neutral regions. However, $\approx$90% of solid tumors are highly aneuploid (Taylor et al., 2018) containing extensive CNAs, and $\approx$30% of solid tumors have whole-genome duplications (Bielski et al., 2018). Identifying collections of SNVs with no possibility of overlapping CNAs during evolution of such tumors may be challenging. Recently, several methods (El-Kebir, 2018; Ciccolella et al., 2018; McPherson et al., 2016; Zafar et al., 2017, 2019; Malikic et al., 2019a, 2019b) have been introduced for single-cell phylogeny inference that allows loss of mutations. SPhyR (El-Kebir, 2018), SASC (Ciccolella et al., 2018), and PyDollo (McPherson et al., 2016) use the Dollo model (Dollo, 1893), which relaxes the infinite-sites model. In the Dollo model, a mutation may be gained ($0 \rightarrow 1$) at most once but may be lost ($1 \rightarrow 0$) multiple times. SiFit (Zafar et al., 2017), SiCloneFit (Zafar et al., 2019), and PhiSCS (Malikic et al., 2019a, 2019b) use the finite-sites model, a further relaxation that allows a mutation to be gained more than once. A challenge in using these less-stringent evolutionary models is that they increase the ambiguity in phylogenetic reconstruction (Figure 1C). Even in simple cases with no error, multiple phylogenies are consistent with the data, and the number of phylogenies further increase when there are errors and uncertainty in the mutation matrix. Both the errors in scDNA-seq data and the mutation losses in the phylogeny conspire to yield considerable challenges and ambiguity in the single-cell phylogeny inference problem. This ambiguity is further amplified because both sequencing errors and losses result in the same signal in the observed data: an observed "0" in the mutation matrix instead of a "1." Thus, it is particularly difficult to distinguish between errors in the data and potential mutation losses.

A major limitation in using the Dollo or finite-sites models to allow mutation losses is that neither of these models consider evidence from CNAs that support or refute a mutation loss at a locus. While more general multi-state models of tumor evolution have been used to infer phylogenies from bulk tumor sequencing data (Deshwar et al., 2015; El-Kebir et al., 2016; Jiang et al., 2016), these approaches neither model the errors in scDNA-seq data nor scale to hundreds to thousands of observed cells. Since mutation losses are the major complication in SNV evolution and responsible for most of the violations of the infinite-sites model in scDNA-seq data (Kuipers et al., 2017; McPherson et al., 2016), the full generality of a multi-state model may not be necessary to obtain accurate phylogenies from scDNA-seq data. Rather, we describe an approach that constrains mutation losses by using copy-number data from the same cells.

We introduce SCARLET (single-cell algorithm for reconstructing loss-supported evolution of tumors), an algorithm that infers phylogenies from scDNA-seq data by integrating SNVs and copy-number data. SCARLET is based on the loss-supported

CellPress
OPEN ACCESS

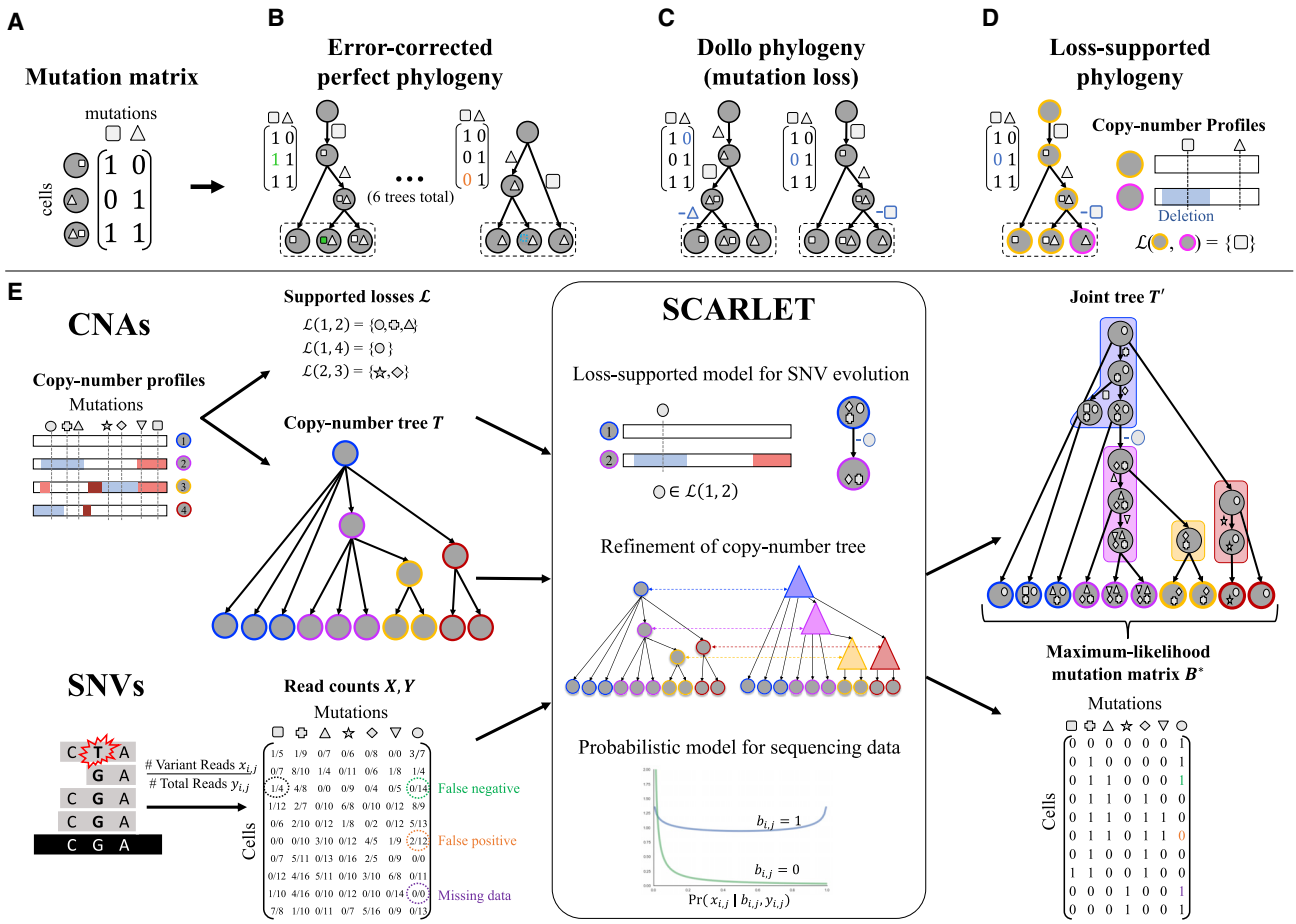


Figure 1. Loss-Supported Phylogeny Model and SCARLET Algorithm for the Maximum-Likelihood Loss-Supported Refinement Problem

(A) A mutation matrix with two mutations in three cells does not admit a perfect phylogeny. This may be due to either errors or mutation losses.

(B) Under the infinite -sites model, existing methods correct errors in the observed matrix to yield a perfect phylogeny.

(C) Under the Dollo model, existing methods identify mutation losses to explain violations of the infinite -sites model. Both the infinite-sites and Dollo models yield multiple equally plausible phylogenies.

(D) The loss-supported model overcomes this ambiguity by using copy-number data to constrain mutation losses.

(E) SCARLET algorithm for the maximum-likelihood loss-supported refinement problem. SCARLET integrates SNVs and CNAs for tumor phylogeny inference. For CNAs, observed copy-number profiles indicate amplified (red) or deleted (blue) genomic regions along the entire genome and are used to obtain two inputs for SCARLET. First, supported loss sets $\mathcal{L}(c, c')$ for pairs of copy-number profiles (empty sets are not shown) indicate mutations that are affected by deletions. Second, a copy-number tree T, which describes the ancestral relationships between observed cells (leaves) as determined by copy-number profiles. For SNVs, variant $\boldsymbol{X}$ and total $\boldsymbol{Y}$ read counts are provided to SCARLET for every cell and every mutation. SCARLET computes a joint tree T' on the observed cells and a maximum-likelihood mutation matrix B^* by constraining mutation losses to the supported loss sets $\mathcal{L}$, computing a refinement T' of T, and selecting the maximum-likelihood $\boldsymbol{B}^*$ using a probabilistic model for the presence ($b_{i,j} = 1$) or absence ($b_{i,j} = 0$) of each SNV in each cell.

phylogeny model that constrains mutation losses to loci, where the copy-number data have evidence of a deletion (Figure 1D). The loss-supported phylogeny generalizes the infinite-sites and Dollo models. SCARLET also relies on a probabilistic model of the read counts for each SNV to address errors and missing data that are common in scDNA-seq. On simulated data, we show that SCARLET infers more accurate phylogenies compared with existing methods. We then use SCARLET to analyze scDNA-seq data from a metastatic colorectal cancer patient (Leung et al., 2017). We find that the published phylogeny—constructed from SNVs under the infinite-sites model—has the implausible conclusion that genome-wide copy-number profiles evolved twice independently during the evolution of this tumor. In contrast, SCARLET infers a loss-supported phylogeny that has three mutation losses, with each loss supported by a copy-number change at the locus. Moreover, the SCARLET phylogeny supports the hypothesis of a single migration between the colon primary tumor and liver metastasis (monoclonal seeding). In contrast, previous published phylogenies (Leung et al., 2017; Zafar et al., 2019) reported a more complex origin of the metastasis with multiple migrations (polyclonal seeding). By integrating information from both SNVs and CNAs, SCARLET obtains more accurate reconstructions of tumor evolution at single-cell resolution.

RESULTS

SCARLET Algorithm for Loss-Supported Phylogeny Model

We developed an algorithm, SCARLET, to infer phylogenetic trees from scDNA-seq data by integrating data from both SNVs and CNAs. SCARLET has three important features (Figure 1E): (1) the loss-supported phylogeny model, which constrains mutation losses to loci where there is a corresponding decrease in copy number; (2) an algorithm to compute a loss-supported phylogeny by refinement of a coarse phylogenetic tree derived from copy-number data alone; and (3) maximum-likelihood inference of SNVs using a probabilistic model of observed read counts in scDNA-seq data. We describe each of these key features below.

The loss-supported model is a model of SNV evolution where mutation gains $(0 \rightarrow 1)$ occur at most once, but mutation losses $(1 \rightarrow 0)$ are constrained by sets $\mathcal{L}$ of supported losses that are defined by CNAs in the same cells (Figure 1D). Specifically, we assume that for each cell we measure both a mutation profile **b** of SNVs and a copy-number profile c. For each pair (c, c') of copy-number profiles, we define the supported loss set $\mathcal{L}(c, c')$ as the set of SNVs at loci where there is a decrease in copy number (e.g., due to a deletion or loss-of-heterozygosity [LOH] event) between profiles c and c'. In the loss-supported phylogeny, a mutation loss at an SNV loci a is allowed between cells v and w only if a is in $\mathcal{L}(c_v, c_w)$. The loss-supported model can thus be viewed as a generalization of other models for SNV evolution: the perfect phylogeny model is the special case where $\mathcal{L} = \varnothing$, while the Dollo model and finite-sites model corresponds to $\mathcal{L}$ being the complete set of all mutations. In contrast to these extremes, the loss-supported model allows for intermediate values of $\mathcal{L}$ derived from copy-number data. The loss-supported model depends on the copy-number profiles of both the observed and ancestral cells. However, we do not directly measure the copy-number profiles of the ancestral cells. To overcome this limitation, SCARLET takes as input a copy-number tree T, which is derived from the copy-number profiles of the observed cells using copy-number phylogenetic reconstruction algorithms (such as described in Schwarz et al., 2014; Chowdhury et al., 2015; El-Kebir et al., 2017; Zaccaria et al., 2018) (Figure 1E). SCARLET computes the supported loss sets $\mathcal{L}$ from the copy-number profiles of the observed cells (leaves of T) and the copy-number profiles of the ancestral cells (internal vertices of T). Typically, scDNA-seq data of SNVs (e.g., from targeted sequencing) measures copy-number profiles with low resolution, and thus tumor cells share a limited number of distinct copy-number profiles. Consequently, the copy-number tree T has many multifurcations or unresolved ancestral vertices with more than two children. SCARLET finds a joint tree T' that is a loss-supported phylogeny and a refinement (Wang et al., 2014) of T by resolving multifurcations in T using the mutation profiles of the observed cells (Figure 1E). Data from scDNA-seq typically have high error rates in identifying SNVs, and particularly high rates of false negatives and missing data due to amplification bias and allele dropout (Gawad et al., 2016). SCARLET models these errors using a beta-binominal distribution (Singer et al., 2018) of the observed read counts. As such, SCARLET computes the loss-supported refinement T' that maximizes the likelihood of the observed sequencing data under this probabilistic model (Figure 1E).

Simulated Data

We compared SCARLET to four existing algorithms that build phylogenies from single-cell sequencing data, SCITE (Jahn et al., 2016), SciΦ (Singer et al., 2018), SPhyR (El-Kebir, 2018), and SiFit (Zafar et al., 2017), on simulated data. We simulated 50 trees, each with 20 mutations, 4 copy-number profiles, and 1–8 mutation losses per tree. From these trees, we simulated 100 observed cells with each cell having equal probability of being a child of any vertex in the simulated tree, and simulated sequencing data with an expected sequencing depth of 100× and allelic dropout rate of 0.15. Additional details of simulated data and parameters of each method are included in STAR Methods. We evaluated the phylogenies output by the methods by two measures, the *mutation matrix error* and the *pairwise ancestral relationship error*, that have been previously used in tumor evolution studies (Ciccolella et al., 2018; Myers et al., 2019; Govek et al., 2018; Singer et al., 2018; El-Kebir, 2018; Satas and Raphael, 2017). The mutation matrix error $M(\mathbf{B}, \widehat{\mathbf{B}}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} |b_{ij} - \widehat{b}_{ij}|$ is the normalized Hamming distance between the inferred binary mutation matrix $\widehat{\mathbf{B}}$ and the true binary mutation matrix **B** and assesses the accuracy of the error-corrected mutation profiles for each observed cell. The pairwise ancestral relationship error $E(T, \widehat{T})$ is the proportion of pairwise ancestral relationships between mutations in the inferred tree $\widehat{T}$ that differ from the ancestral relationships in the true tree T. Specifically, every pair a, a' of mutations has one of four possible ancestral relationships in $\widehat{T}$ and in T: (1) a and a' occur on the same branch; (2) a is ancestral to a'; (3) a' is ancestral to a; and (4) a and a' are incomparable. Note that only mutation gains are considered in the calculation of this error, so that all methods are evaluated on the same set of mutations. We do not calculate the pairwise ancestral relationship error for SiFit because it uses a finite-sites model, which allows mutations to recur and, consequently, pairs of mutations may not have a unique relationship.

SCARLET outperforms all other methods on both mutation matrix error and ancestral relationship error (Figures 2A and 2B). The high errors of SCITE and SciΦ were expected since these methods use an infinite-sites model while the simulations include mutation losses, which violates the model assumptions. However, the methods that do allow mutation losses, SPhyR (based on the k-Dollo model) and SiFit (based on the finite-sites model), do not exhibit improvement over the other methods and perform worse than SCARLET. These results confirm that models that include unconstrained mutation losses have significant ambiguity as it is difficult to distinguish between true mutation losses and false positives or negatives in the data (Figure 1). By using copy-number information to constrain mutation losses, SCARLET overcomes the ambiguity in phylogeny reconstruction and obtains lower error in the inferred mutation matrix and phylogeny.

We evaluated the effect of the input copy-number tree on SCARLET's accuracy by running SCARLET in two modes:

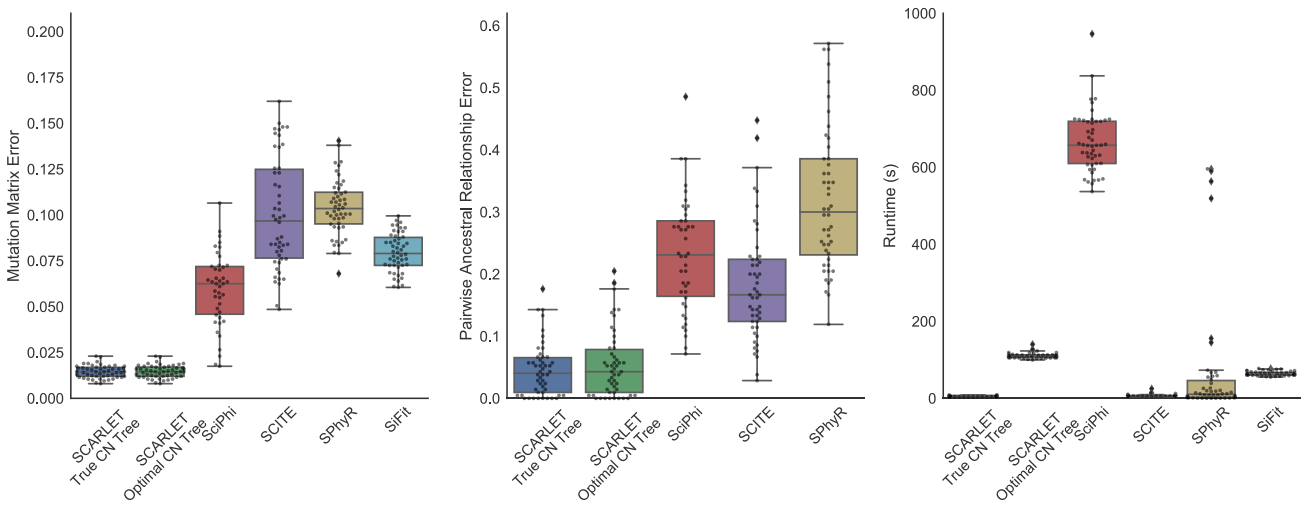


Figure 2. SCARLET Outperforms Existing Methods for Phylogeny Inference on Simulated Single-Cell Data

(Left) Mutation matrix error, (center) Pairwise ancestral relationship error, and (right) runtime for each method. SCARLET was run either knowing (true CN tree) or not knowing (optimal CN tree) the true copy-number tree.

when the true copy-number tree is either known ("SCARLET true CN tree") or unknown ("SCARLET optimal CN tree"). In the latter case, we enumerated all copy-number trees, ran SCARLET once for each copy-number tree, and output the solution with the highest likelihood. In both cases, we provided SCARLET with the true copy-number profiles of each cell and the true set $\mathcal{L}$ of supported losses. SCARLET exhibited comparable performance when running with or without knowledge of the copy-number tree (Figures 2A and 2B). Notably, in 46/50 simulated instances, the maximum-likelihood solution obtained when running SCARLET with unknown copy-number tree was identical to the solution found when providing the true copy-number tree. Clearly, running SCARLET with all possible copy-number trees (16 copy-number trees in this simulation) increases the runtime (Figure 2C), but the runtime remains reasonable when the number of copy-number profiles is small, which is the case for many real datasets (see below).

We further tested SCARLET to evaluate scalability to datasets for larger numbers of mutations (up to $m = 100$) and larger number of copy-number profiles (up to $k = 10$) (Figures 3A and 3B). SCARLET has no loss of accuracy for larger datasets and in some cases has better accuracy on larger datasets. The runtime of SCARLET increases with m, and increases moderately with k, but remains reasonable (with all simulated instances taking < 3 min to run). In addition, we tested how errors in inferring the correct number of copy-number profiles affected the accuracy of SCARLET (Figure 3C). In particular, we tested two types of errors. In "merge" errors, two sets of cells with distinct copy-number profiles are merged together into one. In "split" errors, one set of cells is split and inferred to have two distinct copy-number profiles. With either type of error, SCARLET outperforms other algorithms, with "split" errors leading to a larger reduction in performance than "merge" errors.

Single-Cell Phylogeny of Metastatic Colorectal Cancer

We used SCARLET to analyze scDNA-seq of a metastatic colorectal cancer patient CRC2 from Leung et al. (2017). This dataset included targeted sequencing of 1,000 genes in 141 cells from a primary colon tumor and 45 cells from a matched liver metastasis (Figure 4A). The authors identified 36 SNVs and used SCITE (Jahn et al., 2016) to derive a perfect phylogeny from these SNVs (Figure 4B). This perfect phylogeny tree shows two distinct branches of metastatic cells and (Leung et al., 2017) concluded that this was evidence of polyclonal seeding of the liver metastasis; i.e., two distinct cells (or groups of cells) with different complements of mutations migrated from the primary colon tumor to the liver metastasis. Examining the copy-number data, one finds a curious discrepancy between the SCITE tree and the single-cell copy-number profiles. Whole-genome sequencing of 42 single cells from the same patient reveals that all metastatic cells share losses of chromosomes 2, 3p, 4, 7, 9, 16, and 22 relative to the cells in the primary tumor (Figure 4C). According to the SCITE tree, all of these large CNAs would had to have occurred twice independently in the two distinct branches of metastatic cells. Although CNAs can exhibit homoplasy, this high rate of occurrence of the exact same events seems highly unlikely. Thus, we observe an inconsistency between the copy-number data and the SCITE tree constructed using only SNV data. Notably, this same dataset was recently analyzed by SiCloneFit (Zafar et al., 2019) using a finite-sites model. The SiCloneFit tree also showed two branches of metastatic cells and concluded that there was polyclonal seeding of the metastases. Thus, the SiCloneFit phylogeny also has the same inconsistency between the SNV phylogeny and copy-number data.

We analyzed this dataset using SCARLET to see whether joint analysis of SNVs and CNAs data could help resolve the inconsistency between the tree derived from SNVs and the observed copy-number profiles. We first derived four distinct copy-number profiles by hierarchical clustering of ploidy-corrected read-depth ratios from the targeted single-cell sequencing data. These copy-number profiles included an aneuploid profile for all primary cells (P), two different aneuploid profiles for metastatic cells (M1 and M2), and the profile of diploid cells (D); (Leung

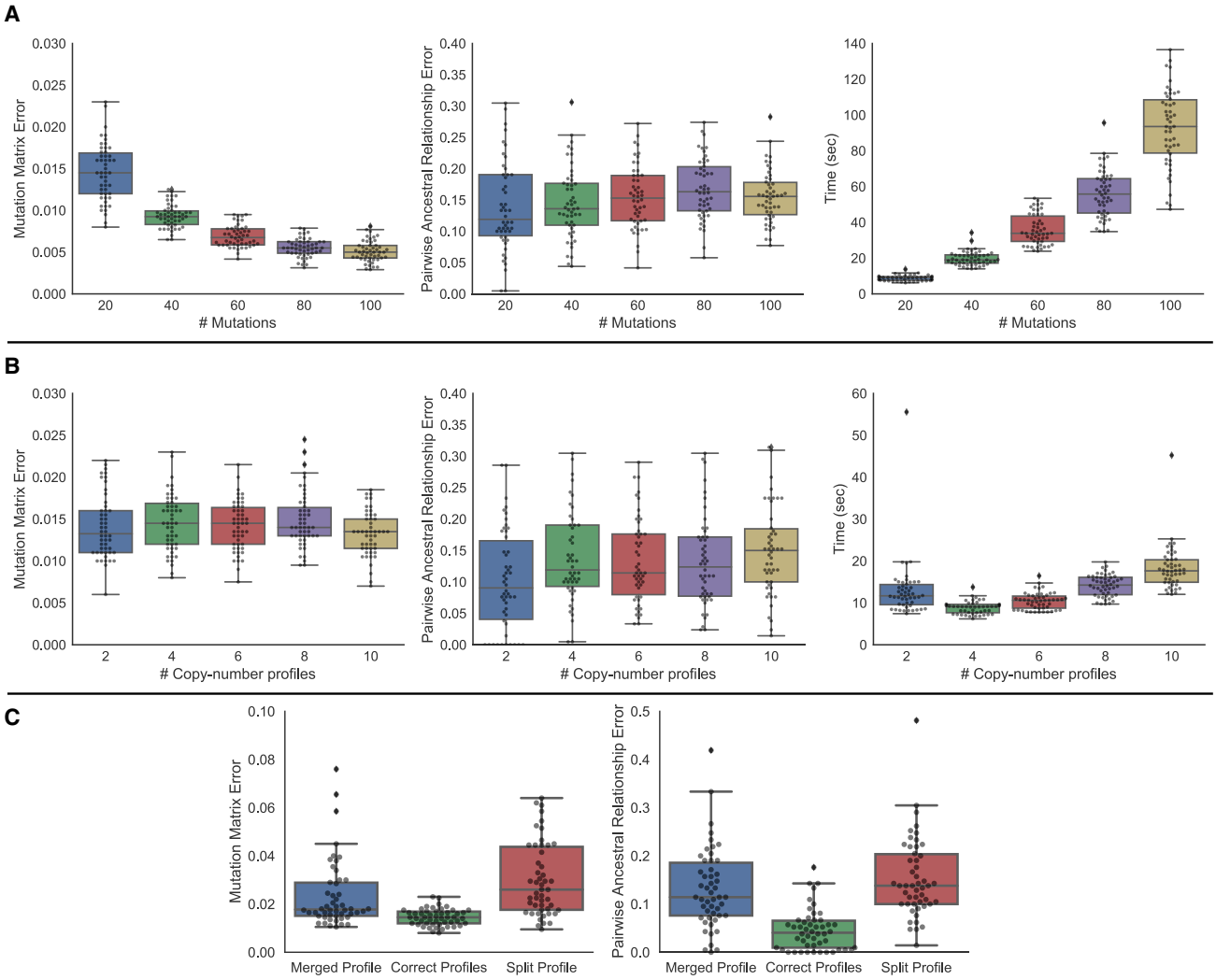


Figure 3. SCARLET Scales to Larger Datasets and Tolerates Errors in Copy-Number Profiles

(A) SCARLET results on simulated data with varying number of mutations, $n = 100$ cells and $k = 4$ copy-number profiles. (Left) Mutation matrix error; (Center) pairwise ancestral relationship error; (Right) runtime.

(B) SCARLET results on simulated data with varying number of copy-number profiles, $n = 100$ cells and $m = 20$ mutations. (Left) Mutation matrix error; (center) pairwise ancestral relationship error; (right) runtime.

(C) SCARLET results on simulated data with incorrect number of copy-number profiles. We introduce errors in the number of copy-number profiles by either merging two distinct copy-number profiles together (merged profile), or splitting one copy-number profile into two (split profile) and compare performance against the correct number of copy-number profiles.

et al., 2017) and similarly derived four copy-number profiles from whole-genome sequencing of a different set of 42 cells from the same patient. Since four copy-number profiles is a small number to infer a tree using a copy-number evolution model, we instead ran SCARLET in the "optimal CN tree" setting selecting the copy-number tree that produced the highest likelihood. Specifically, we ran SCARLET on all nine possible rooted copy-number trees with the root having the diploid profile (D) and internal vertices labeled by one of the three aneuploid copy profiles (P, M1, and M2). For each copy-number tree, we derived the set $\mathcal{L}$ of supported losses as the mutation loci that exhibited significant decreases in read depth (i.e., number of aligned sequencing reads). Additional details are included in STAR Methods.

SCARLET constructed a tree (Figure 4D and 4E) with a single clade containing all metastatic cells. This is consistent with the copy-number data, since the shared chromosomal losses could have occurred once in a common ancestor of all metastatic cells. Moreover, this tree suggests that the liver metastasis was the product of monoclonal seeding; i.e., a single cell (or small group of cells) with the same somatic mutations migrated from the primary colon tumor to the metastasis and all metastatic cells descended from the founder cells present in this single migration. This result contradicts previous results (Leung et al., 2017; Zafar et al., 2019) of a more complicated polyclonal seeding of the metastasis. The SCARLET tree contains three mutation losses: in genes FHIT, LRP1B, and LINGO2. Each of these losses is supported by a significant decrease in read depth (Figure 4D),

CellPress
OPEN ACCESS

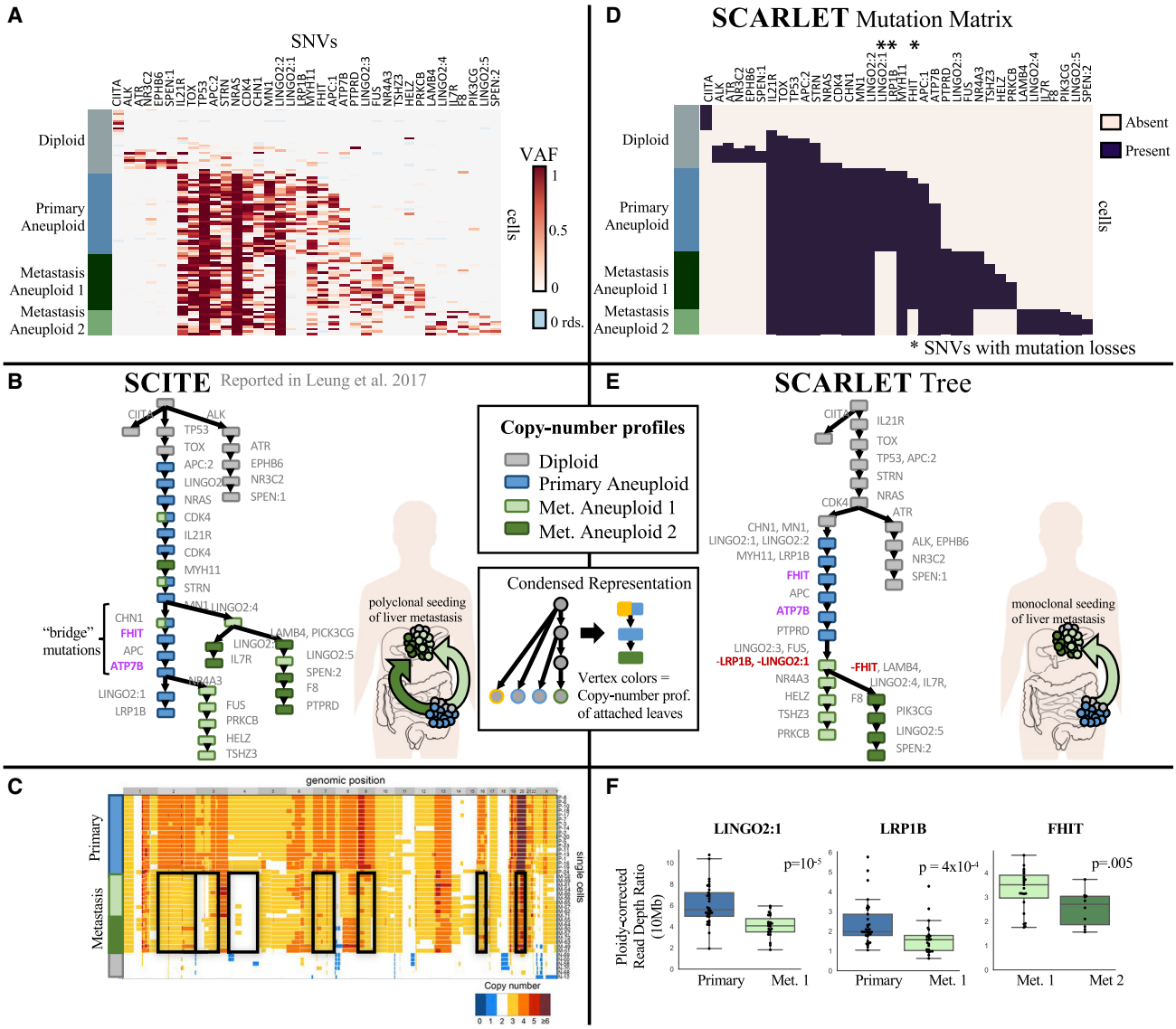


Figure 4. SCARLET Infers a Loss-Supported Phylogeny Consistent with Copy-Number Profiles from a Metastatic Colorectal Cancer Patient

SCARLET was applied to targeted scDNA-seq of 141 single cells from the primary colon tumor (blue) and 45 single cells from the liver metastasis (green) of patient CRC2.

(A) Variant allele frequencies of 36 somatic SNVs in 96 cells as inferred by SCITE.

(B) Perfect phylogeny tree inferred by SCITE in Leung et al. (2017) of patient CRC2. Two distinct branches of metastatic cells—suggesting polyclonal seeding of the liver metastasis—are separated by the four indicated "bridge mutations" occurring in cells of the primary tumor.

(C) Published copy-number profiles from DOP-PCR (degenerate oligonucleotide-primed polymerase chain reaction) whole-genome sequencing of 42 single cells from both the primary tumor and metastasis of CRC2; figure adapted from Leung et al. (2017). All metastatic cells share deletions of six chromosomes (black boxes) but are separated into two groups (light and dark green) by a small number of additional CNAs.

(D) Mutation matrix derived from the loss-supported phylogeny inferred by SCARLET on the same data.

(E) The loss-supported phylogeny inferred by SCARLET has a single branch containing all metastatic cells —suggesting monoclonal seeding of the liver metastasis and consistent with the similar copy-number profiles of all metastatic cells. SCARLET identifies mutation losses (red) in LINGO2, LRP1B, and FHIT.

(F) Significant decreases in read depths are observed at the loci of the three mutation losses identified by SCARLET (p values derived from Wilcoxon Rank-Sum test).

providing evidence that the loci containing these mutations were likely affected by deletions. Notably FHIT and LRP1B are located in fragile sites in the genome (Smith et al., 2006), which are known regions of genomic instability. In addition, the loss of the mutation LINGO2:1 in LINGO2 is further supported by a shift in the variant allele frequency of another mutation, LINGO2:2, in the same gene. Specifically, the variant allele frequency of LINGO2:2 is ≈1 in the metastatic cells (Figure 4A), suggesting that this mutant allele is homozygous, consistent with a deletion or LOH event where the LINGO2:1 mutation was lost.

We examined further the evidence for polyclonal seeding in the initial study of this patient. Leung et al. (2017) included a statistical analysis of the variant read counts of the four "bridge mutations," ATP7B, FHIT, APC, and CHN1 that occurred between the first and second metastatic branches in the SCITE tree. This analysis showed that mutations in ATP7B and FHIT were present in a subset of primary tumor cells and in the second metastatic branch (detected in 10/13 and 13/13 cells, respectively) while being absent in the second metastatic branch (detected in 1/15 and 1/15 cells, respectively). Under the infinite-sites model used by SCITE, mutation loss is not allowed and thus polyclonal seeding is necessary to explain the absence of these mutations. The same analysis found high uncertainty regarding the placement of mutations in APC and CHN1 and thus these were not cited as evidence for polyclonal seeding.

The loss-supported model used by SCARLET provides an alternate explanation for the absence of FHIT and ATP7B. SCARLET identifies a supported mutation loss to explain the presence of the mutation in FHIT only in a subset of metastatic cells (M1). This loss is supported by a shift in read depth ($p = 0.005$) in the 10-Mb region containing the locus (Figure 4F). SCARLET does not identify a supported mutation loss to similarly explain ATP7B as we did not observe a significant decrease in read depth for the corresponding locus ($p = 0.34$). However, this lack of a significant decrease in read depth at the ATP7B locus does not necessarily imply that there was no mutation loss. In particular, because targeted sequencing was performed for only 1,000 genes, the copy-number data are fairly low resolution, and we calculated read depth in 10-Mb bins. Thus, we may lack the statistical power to identify a shorter deletion, especially a deletion present in only the 10 metastatic cells with copy-number profile M2. In summary, we argue that the sequencing data provide stronger evidence for the phylogeny constructed by SCARLET, which is consistent with both SNV and copy-number data and supports a more parsimonious explanation of monoclonal seeding of the liver metastasis.

DISCUSSION

Somatic mutations in tumors range across all genomic scales, from SNVs through large CNAs. To date, most methods for constructing phylogenies from scDNA-seq data (Jahn et al., 2016; Singer et al., 2018; Ross and Markowetz, 2016; Zafar et al., 2017, 2019; El-Kebir, 2018; Ciccolella et al., 2018; McPherson et al., 2016; Malikic et al., 2019a, 2019b) used only SNVs, ignoring CNAs and thus throwing out important information for phylogenetic inference. Here, we introduced SCARLET, which uses measurements of both SNVs and CNAs to reconstruct tumor phylogenies from scDNA-seq data. SCARLET is based on a loss-supported evolutionary model, which constrains mutation losses to loci containing evidence of a CNA. By using the information about CNAs that is readily available in scDNA-seq data, the loss-supported model has less ambiguity in the phylogeny inference than the Dollo and finite-sites models that allow mutation losses to occur anywhere on the tree. In scDNA-seq data, where there is often considerable uncertainty in the mutations present in each cell, this reduction in ambiguity enables more accurate phylogeny inference. On simulated scDNA-seq data, we find that SCARLET outperforms existing methods that do not utilize copy-number data. On targeted scDNA-seq data from a metastatic colorectal cancer patient, we showed that SCARLET found a phylogeny containing three mutation losses. Notably, SCARLET's tree was both more consistent with the copy-number data and provided a simpler explanation of monoclonal seeding of the liver metastasis compared with the more complex phylogenies reported previously (Leung et al., 2017; Zafar et al., 2019). Thus, accurate modeling of mutations losses results in different conclusions regarding the migration patterns of metastasis.

There are a number of directions for future improvement. First, the current implementation of SCARLET either requires the copy-number tree in input or enumerates all possible copy-number trees and selects the maximum-likelihood result. This approach is applicable when the number of distinct copy-number profiles is small; e.g., in the case of targeted scDNA-seq data (Leung et al., 2016; Xu et al., 2012; Mission Bio, 2019), where copy-number data typically are lower resolution. However, with higher-quality copy-number data, extensions to larger numbers of copy-number profiles are needed. One approach is to use copy-number evolution models (Chowdhury et al., 2015; Schwarz et al., 2014; El-Kebir et al., 2017; Zaccaria et al., 2018) to identify a modest number of copy-number trees that summarize the uncertainty in the copy-number evolutionary history. Second, one could extend the loss-supported model into a unified evolutionary model for SNVs and CNAs. Indeed, the loss-supported model provides a natural framework to integrate SNVs directly with evolutionary models of CNAs. As single-cell sequencing technologies continue to improve, higher quality measurements of both SNVs and CNAs from the same sets of cells will become available. We anticipate that SCARLET and the loss-supported model will play a crucial role in the analysis of these data.

STAR★METHODS

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

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Loss-Supported Phylogeny Model
  - Loss-Supported Refinement Problem
  - Solving the Loss-Supported Refinement problem
  - Base Step
- RECURSIVE STEP
  - Maximum Likelihood Loss-supported Refinement Problem
  - SCARLET Algorithm for Maximum-Likelihood Loss-Supported Refinement Problem
  - Finding Maximum Likelihood Subtree Roots
  - Finding Refinement Subtrees
  - Proofs
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Simulation Details
  - Copy-Number Analysis of Colorectal Cancer Patient
- DATA AND CODE AVAILABILITY

CellPress
OPEN ACCESS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.04.001.

ACKNOWLEDGMENTS

This work is supported by US National Institutes of Health (NIH) grants R01HG007069 and U24CA211000, US National Science Foundation (NSF) CAREER award (CCF-1053753), and Chan Zuckerberg Initiative DAF grant 2018-182608 to B.J.R.

AUTHOR CONTRIBUTIONS

Conceptualization, G.S. and B.J.R.; Methodology, G.S., S.Z., and B.J.R.; Software, G.S.; Formal Analysis, G.S., S.Z., and G.M.; Investigation, G.S., S.Z., G.M., and B.J.R.; Data Curation, G.S.; Writing – Original Draft, G.S., S.Z., and B.J.R.; Writing – Review & Editing, G.S., S.Z., and B.J.R.; Supervision, B.J.R.; Funding Acquisition, B.J.R.

DECLARATION OF INTERESTS

B.J.R. is a founder of Medley Genomics and a member of its board of directors.

Received: February 10, 2020
Revised: March 3, 2020
Accepted: March 25, 2020
Published: April 22, 2020; corrected online May 22, 2020

REFERENCES

Alves, J.M., Prieto, T., and Posada, D. (2017). Multiregional tumor trees are not phylogenies. Trends Cancer *3*, 546–550.

Bielski, C.M., Zehir, A., Penson, A.V., Donoghue, M.T.A., Chatila, W., Armenia, J., Chang, M.T., Schram, A.M., Jonsson, P., Bandlamudi, C., et al. (2018). Genome doubling shapes the evolution and prognosis of advanced cancers. Nat. Genet. *50*, 1189–1195.

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

Chen, D., Eulenstein, O., Fernandez-Baca, D., and Sanderson, M. (2006). Minimum-flip supertrees: complexity and algorithms. EEE/ACM Trans. Comput. Biol. Bioinform. *3*, 165–173.

Chowdhury, S.A., Gertz, E.M., Wangsa, D., Heselmeyer-Haddad, K., Ried, T., Schäffer, A.A., and Schwartz, R. (2015). Inferring models of multiscale copy number evolution for single-tumor phylogenetics. Bioinformatics *31*, i258–i267.

Ciccolella, S., Gomez, M.S., Patterson, M., Della Vedova, G., Hajirasouliha, I., and Bonizzoni, P. (2018). Inferring cancer progression from single cell sequencing while allowing loss of mutations. bioRxiv. https://doi.org/10.1101/268243v2.

Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., and Morris, Q. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol *16*, 35.

Dollo, L. (1893). The laws of evolution. Bull. Soc. Bel. Geol. Palaeontol. *7*, 164–166.

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

El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B.J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics *31*, i62–i70.

El-Kebir, M., Raphael, B.J., Shamir, R., Sharan, R., Zaccaria, S., Zehavi, M., and Zeira, R. (2017). Complexity and algorithms for copy-number evolution problems. Algor. Mol. Biol. *12*, 13.

El-Kebir, M., Satas, G., Oesper, L., and Raphael, B.J. (2016). Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Syst *3*, 43–53.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. *17*, 175–188.

Govek, K., Sikes, C., and Oesper, L. (2018). A consensus approach to infer tumor evolutionary histories. In Proceedings of the 2018 Acm international conference on bioinformatics, computational biology, and health informatics, pp. 63–72.

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

Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. Genome Biol *17*, 86.

Jiang, Y., Qiu, Y., Minn, A.J., and Zhang, N.R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc. Natl. Acad. Sci. USA *113*, E5528–E5537.

Jiao, W., Vembu, S., Deshwar, A.G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics *15*, 35.

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.

Leung, M.L., Davis, A., Gao, Ruli, Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S., and Navin, N.E. (2017). Single-cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res *27*, 1287–1299.

Leung, M.L., Wang, Y., Kim, C., Gao, Ruli, Jiang, J., Sei, E., and Navin, N.E. (2016). Highly multiplexed targeted dna sequencing from single nuclei. Nat. Protoc. *11*, 214–235.

Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S.C., and Beerenwinkel, N. (2019a). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. Nat. Commun. *10*, 2750.

Malikic, S., McPherson, A.W., Donmez, N., and Sahinalp, C.S. (2015). Clonality inference in multiple tumor samples using phylogeny. Bioinformatics *31*, 1349–1356.

Malikic, S., Mehrabadi, F.Rashidi, Ciccolella, S., Rahman, M.K., Ricketts, C., Haghshenas, E., Seidman, D., Hach, F., Hajirasouliha, I., and Sahinalp, S.C. (2019b). PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. Genome Res *29*, 1860–1877.

McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A.W., Ha, G., Biele, J., Yap, D., Wan, A., et al. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nat. Genet. *48*, 758–767.

Miura, S., Vu, T., Deng, J., Buturla, T., Choi, J., and Kumar, S. (2019). Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. bioRxiv. https://doi.org/10.1101/697318v1.

Myers, M.A., Satas, G., and Raphael, B.J. (2019). CALDER: inferring phylogenetic trees from longitudinal tumor samples. Cell Syst *8*, 514–522.e5.

Navin, N.E. (2015). The first five years of single-cell cancer genomics and beyond. Genome Res *25*, 1499–1507.

Pe'er, I., Pupko, T., Shamir, R., and Sharan, R. (2004). Incomplete directed perfect phylogeny. SIAM J. Comput. *33*, 590–607.

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R.B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. Genome Biol *16*, 91.

Pradhan, D., and El-Kebir, M. (2018). On the non-uniqueness of solutions to the perfect phylogeny mixture problem. In Comparative Genomics. RECOMB-CG 2018. Lecture Notes in Computer Science, *vol 11183*, M. Blanchette and A. Ouangraoua, eds. (Springer), pp. 277–293.

Ross, E.M., and Markowetz, F. (2016). OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol *17*, 69.

Satas, G., and Raphael, B.J. (2017). Tumor phylogeny inference using tree-constrained importance sampling. Bioinformatics *33*, i152–i160.

Schwarz, R.F., Trinh, A., Sipos, B., Brenton, J.D., Goldman, N., and Markowetz, F. (2014). Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput. Biol. *10*, e1003535.

Singer, J., Kuipers, J., Jahn, K., and Beerenwinkel, N. (2018). Single-cell mutation identification via phylogenetic inference. Nat. Commun. *9*, 5144.

Smith, D.I., Zhu, Y., McAvoy, S., and Kuhn, R. (2006). Common fragile sites, extremely large genes, neural development and cancer. Cancer Lett *232*, 48–57.

Tabassum, D.P., and Polyak, K. (2015). Tumorigenesis: it takes a village. Nat. Rev. Cancer *15*, 473–483.

Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell *33*, 676–689.e3.

Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155–160.

Wu, T., Moulton, V., and Steel, M. (2009). Refining phylogenetic trees given additional data: an algorithm based on parsimony. IEEE/ACM Trans Comput Biol Bioinform (TCBB) *6*, 118–125.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, Aifa, Song, Luting, 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.

Zaccaria, S., El-Kebir, M., Klau, G.W., and Raphael, B.J. (2018). Phylogenetic copy-number factorization of multiple tumor samples. J. Comput. Biol. *25*, 689–708.

Zafar, H., Navin, N., Chen, K., and Nakhleh, L. (2019). SiCloneFit: bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. Genome Res *29*, 1847–1859.

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.

Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S.P., Aparicio, S., and Hansen, C.L. (2017). Scalable whole-genome single-cell library preparation without Preamplification. Nat. Methods *14*, 167–173.

10 X Genomics. (2018) Assessing tumor heterogeneity with single cell cnv. https://www.10xgenomics.com/solutions/single-cell-cnv.

Mission Bio. (2019). Copy number variants and single nucleotide variants simultaneously detected in single cells. https://missionbio.com/cnv_application_note.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Simulated Data	This paper	https://github.com/raphael-group/scarlet
Colorectal cancer single-cell DNA sequencing	NCBI Sequence Read Archive	https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP074289
Software and Algorithms		
SCARLET	This paper	https://github.com/raphael-group/scarlet
SCITE	Jahn et al., 2016	https://github.com/cbg-ethz/SCITE/
SiFit	Zafar et al., 2017	https://bitbucket.org/hamimzafar/sifit/
SciPhi	Singer et al., 2018	https://github.com/cbg-ethz/SCIPhI/
SPhyR	El-Kebir, 2018	https://github.com/elkebir-group/SPhyR

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Ben Raphael (braphael@princeton.edu). This study did not generate any reagents.

METHOD DETAILS

Loss-Supported Phylogeny Model

We model the evolutionary history of a tumor as a rooted, directed phylogenetic tree $T = (V(T), E(T))$, whose vertex set $V(T) = L(T) \cup I(T)$ consists of a set $L(T)$ of n leaves corresponding to *observed cells* and a set $I(T)$ of inner vertices corresponding to *ancestral cells*. A directed edge $(v, w) \in E(T)$ indicates that cell v is an ancestor of cell w. We do not directly observe T but rather we measure a set of phylogenetic markers for every observed cell $v \in L(T)$. In the case where the markers are somatic single-nucleotide variants (SNV), the measurements correspond to a binary *mutation profile* $\mathbf{b}_v \in \{0,1\}^m$ for each observed cell v, where $b_{v,a} = 1$ indicates that cell v has a somatic mutation at locus a and $b_{v,a} = 0$ indicates that cell v does not have a somatic mutation at locus a. We assume that the mutation profile $\mathbf{b}_r$ of the root r is $\mathbf{b}_r = \overrightarrow{0}$ since the root represents the normal cell that preceded the tumor. We define the *mutation matrix* $\mathbf{B} = [\mathbf{b}_v]_{v \in L(T)}$ to be the matrix whose rows are the mutation profiles of leaves $v \in L(T)$.

The problem of phylogenetic tree inference is to find a tree T and an *augmented mutation matrix* $\mathbf{B}' = [\mathbf{b}'_v]_{v \in V(T)}$ whose rows correspond to binary mutation profiles of the vertices of T and where the submatrix $[\mathbf{b}'_v]_{v \in L(T)}$ is equal to $\mathbf{B}$. Since there are many possible trees that relate the observed cells, methods for phylogeny inference find T and $\mathbf{B}'$ that best fit a specific evolutionary model. The simplest evolutionary model for SNVs is the *infinite sites*, or *perfect phylogeny* model. In this model, each mutation is gained ($0 \rightarrow 1$) at most once, and is never subsequently lost. A more general model the Dollo model allows mutations to be gained ($0 \rightarrow 1$) at most once, but lost ($1 \rightarrow 0$) multiple times. Formally, the Dollo model is defined as follows.

Definition 1 *A phylogenetic tree T is a Dollo phylogeny with respect to augmented mutation matrix* $\mathbf{B}'$ *provided that for every locus a, there is at most one edge* $(v, w) \in E(T)$ *such that* $b'_{v,a} = 0$ *and* $b'_{w,a} = 0$.

In contrast to the perfect phylogeny model, under the Dollo model there are often multiple phylogenies that are consistent with input data (Figure 1).

DNA sequencing data often contains contains additional information about the genomic locations where mutation losses are possible. Specifically, we assume that for each cell v, we also observe a copy-number profile $\mathbf{p}_v = [p_{v,1}, \ldots, p_{v,N}]$ where $p_{v,i}$ indicates the number of copies of genomic segment i in cell v. For simplicity, we label the unique copy-number profiles observed for all the cells by integers $\{1, \ldots, k\}$, such that the vector $\mathbf{c} = [c_v]$ represents the copy-number profile assignment $c_v \in \{1, \ldots, k\}$ of every cell v. The copy-number profiles of cells provide constraints on mutation losses. In particular, we allow mutation losses only at loci where an overlapping deletion or loss-of-heterozygosity (LOH) distinguishes the copy-number profiles. We record the information about the loci where losses are allowed in a collection $\mathcal{L}$ of *supported loss sets*. For each pair c, c' of distinct copy-number profiles we define the set $\mathcal{L}(c, c') \subseteq \{1, \ldots, m\}$ of *supported losses* to be the set of all the mutation loci located in genomic regions with a decrease in copy

number (indicating possible deletion or LOH) between c and c'. We define $\mathcal{L}(c,c)=\varnothing$ for all c. We denote the collection of supported losses as $\mathcal{L}=\{\mathcal{L}(c,c'):(c,c')\in\{1,\ldots,k\}\times\{1,\ldots,k\}\}$. We define a *loss-supported phylogeny* as a Dollo phylogeny where all mutation losses are supported.

Definition 2 Given copy number profiles $\boldsymbol{c}'=[c_v]_{v\in V(T)}$ and supported losses $\mathcal{L}$, a phylogenetic tree T is a loss-supported phylogeny with respect to augmented mutation matrix $\boldsymbol{B}'$ provided that: (1) T is a Dollo phylogeny; (2) If $b'_{v,a}=1$ and $b'_{w,a}=0$ for edge (v,w) then $a\in\mathcal{L}(c'_v,c'_w)$.

The loss-supported phylogeny inference problem is to infer a loss-supported phylogeny T given a mutation matrix $\mathbf{B}$ and copy-number profile vector $\mathbf{c}$ that label the leaves of T, as well as a set $\mathcal{L}$ of supported losses. However, this general problem has a major complication: the copy-number profiles of the ancestral cells are unknown. Without knowledge of ancestral copy-number profiles, the loss sets $\mathcal{L}$ cannot be used to constrain mutation losses. Ideally, one might infer copy-number profiles of ancestral cells (e.g., using a copy-number evolution model (Schwarz et al., 2014; Chowdhury et al., 2015; El-Kebir et al., 2017; Zaccaria et al., 2018)) while *simultaneously* inferring a loss-supported phylogeny on the SNVs. The derivation of a score/likelihood for such joint model is not straightforward, and is left for future work. Instead, in the next section, we describe an algorithm that infers a loss-supported phylogeny by *refining* a *copy-number tree* given in input.

Loss-Supported Refinement Problem

In this section, we introduce the *Loss-Supported Refinement (LSR) problem*, a special case of the loss-supported phylogeny inference problem, where we have additional information about the evolutionary relationships between copy-number profiles. In particular, we assume that we are given a copy-number tree $b'_{v,a}=1$ and a copy-number profile vector $b'_{w,a}=0$ for all vertices in T. A copy-number tree is a phylogenetic tree constructed using CNAs as evolutionary markers. Leaves of T correspond to observed cells, inner vertices of T to ancestral cells with distinct copy-number profiles, and edges to ancestral relationships. As single-cell DNA sequencing data of SNVs typically measures copy-number profiles with low-resolution, this copy-number tree typically has many multifurcations (i.e., unresolved ancestral vertices with more than two children). We use the mutation matrix (v,w) for all $a\in\mathcal{L}(c'_v,c'_w)$ to refine vertices in T, which results in a *joint tree* T' that reflects the evolutionary history of both the SNVs and CNAs. This sequential approach is inspired by an asymmetry between SNVs and CNAs in the loss-supported model: CNAs affect the observed state transitions of SNVs as deletions result in SNV loss, but SNVs do not result in changes in copy-number state. The joint tree T' is a *refinement* (Wu et al., 2009) of T; i.e., $L(T')=L(T)$ and T may be obtained by contracting edges in T'.

A refinement is formalized as a mapping $\gamma:V(T)\to 2^{V(T')}$, where for all $v\in V(T)$, $\gamma(v)$ is a rooted subtree $T'[\gamma(v)]$ in T'. Given T' one can obtain T by contracting each subtree $T'[\gamma(v)]$ into a single vertex $v\in V(T)$. We refer to the set of subtrees defined by γ as the *refinement subtrees*.

We define the LSR problem as the problem of finding a refinement T' of a copy-number tree T such that T' is a loss-supported phylogeny.

Problem 1 ***Loss-Supported Refinement (LSR) problem*** *Given a copy-number tree T, a copy-number profile vector $\boldsymbol{c}=[c_v]_{v\in V(T)}$, a mutation matrix $\boldsymbol{B}=[\boldsymbol{b}_v]_{v\in L(T)}$, and supported losses $\mathcal{L}$, find a refinement T' of T, a copy-number profile vector $\boldsymbol{c}'=[c_{v'}]_{v'\in V(T')}$, and an augmented mutation matrix $\boldsymbol{B}'=[\boldsymbol{b}'_{v'}]_{v'\in V(T')}$ with $\boldsymbol{b}_{v'}=\boldsymbol{b}'_{v'}$ for all $v'\in L(T')$, such that*

(1) $c'_{v'}=c_v$ for all $v\in V(T)$ and $v'\in\gamma(v)$, and
(2) T' is a loss-supported phylogeny with respect to $\boldsymbol{B}'$, $\boldsymbol{c}'$, and $\mathcal{L}$.

We provide four necessary and sufficient conditions for a solution $T',\mathbf{c}',\mathbf{B}'$ to the LSR problem. These conditions constrain the set of refinement subtrees defined by γ. The four conditions state that (1) each mutation occurs at most once, (2) mutations are not lost *within* refinement subtrees, (3) all mutation losses *between* refinement subtrees are supported, and (4) refinement subtree copy-number profiles are preserved. We formally define these four conditions as follows, using $r(v)$ to denote the root of subtree $T'[\gamma(v)]$ and $p(r(v))$ to denote the parent of $r(v)$.

Theorem 1 *Given copy-number tree T, copy-number profile vector $\boldsymbol{c}$, mutation matrix $\boldsymbol{B}$, and supported losses $\mathcal{L}$, a refinement T' of T, copy-number profile vector $\boldsymbol{c}'$, and augmented mutation matrix $\boldsymbol{B}'$ are a solution to the LSR problem if and only if*

(1) For all loci a, there exists exactly one edge $(v',w')\in E(T')$ with $b'_{v',a}=0$ and $b'_{w',a}=1$;

And for all $v\in V(T)$:

(2) There does not exist any edge $(v',w')\in E(T'[\gamma(v)])$ with $b'_{v',a}=1$ and $b'_{w',a}=0$;
(3) If $b'_{p(r(v)),a}=1$ and $b'_{r(v),a}=0$, then $a\in\mathcal{L}(c'_{p(r(v))},c'_{r(v)})$;
(4) $c'_{v'}=c_v$ for all $v'\in\gamma(v)$

Note that, taken together, conditions (1) and (2) imply that each of these subtrees $T'[\gamma(v)]$ is a perfect phylogeny with respect to submatrix $\mathbf{B}'[\gamma(v)]$. We use this structure to solve the LSR problem in the next section.

Solving the Loss-Supported Refinement problem

In this section, we derive an efficient algorithm to solve the LSR problem. This algorithm decomposes the LSR problem into $k=|I(T)|$ instances of the Incomplete Directed Perfect Phylogeny (IDP) problem (Pe'er et al., 2004) – one instance for each copy-number profile

– using the characterization given in Theorem 1. Specifically, Theorem 1 characterizes LSR solutions by giving conditions on the set of refinement subtrees of T. We design an algorithm to find a set $\mathcal{T} = \{T'_v : v \in V(T)\}$ of subtrees, an augmented mutation matrix $\mathbf{B}'$, and copy-number profiles $\mathbf{c}'$ that satisfy Theorem 1. Using $\mathcal{T}$ and $\mathbf{B}'$, we then construct a refinement T' such that $T'[\gamma(v)] = T'_v$ and $c'_{v'} = c_v$ for all vertices $v' \in V(T'_v)$.

We present a recursive algorithm that refines T from the leaves to the root. The algorithm relies on three additional constraints on the solution T', $\mathbf{c}'$ and $\mathbf{B}'$ that do not effect the existence of a solution, described in the following lemma.

Lemma 1 *If there exists a solution to the LSR problem for a given T, $\boldsymbol{c}$, $\boldsymbol{B}$, $\mathcal{L}$, then there exists a solution T', $\boldsymbol{c}'$, $\boldsymbol{B}'$ that satisfies the following conditions*.

(1) For all $(v,w) \in E(T)$, $p(r(w))$ is a leaf of subtree $T'[\gamma(v)]$.
(2) For all $v \in V(T) \setminus \{r\}$, if $b'_{v',a} = 1$ for all $v' \in L(T'[\gamma(v)])$ then $b'_{r(v),a} = 1$.
(3) For all $v \in V(T)$ and all loci a, $b'_{p(r(v)),a} \geq b'_{r(v),a}$.

Our recursive algorithm is composed of a base and recursive step.

Base Step

The base step determines T'_v and $\mathbf{B}'[V(T'_v)]$ for leaf vertices $a \in \mathcal{L}(c'_v, c'_w)$. For any leaf in a refinement T', $\gamma(v) = \{v\}$. Thus the subtree $T'_v \in \mathcal{T}$ is composed of a single vertex v, with mutation profile $\mathbf{b}'_v = \mathbf{b}_v$ and copy-number profile $c'_v = c_v$.

RECURSIVE STEP

The recursive step aims to find T'_v and $\mathbf{B}'[V(T'_v)]$ for internal vertices $v \in I(T)$. We find T'_v in two steps. First, we identify the set of constraints on the leaves $L(T'_v)$ of T'_v given by Theorem 1. Second, given these constraints, we find mutation profiles $\mathbf{B}'[L(T'_v)]$ for the leaves that respect a perfect phylogeny, as required by condition (1) and (2) of Theorem 1. These mutation profiles uniquely determine the structure of T'_v as T'_v is a perfect phylogeny (Gusfield, 1991). We describe these steps in detail below.

By condition (i) of Lemma 1, T'_v has a leaf for every $(v,w) \in E(T)$; thus $L(T'_v) = \{p(r(w)) : (v,w) \in E(T)\}$. We first recursively solve for T'_w and $\mathbf{B}'[V(T'_w)]$ for every vertex w such that $(v,w) \in E(T)$. Thus, we know the mutation profile $\mathbf{b}'_{r(w)}$ of the root of each child subtree. We do not directly observe $\mathbf{B}'[L(T_v)]$, but the mutation profile of a vertex is constrained by condition (3) of Theorem 1 and constraint (iii) of Lemma 1 given the mutation profile of a child. Specifically, the parent has the same mutation profile as the child, except if there is a mutation loss. The mutation profiles are further constrained by condition (1) of Theorem 1 as each mutation occurs at most once across all subtrees. We respect this condition by minimizing the number of mutation gains per locus, by only having a mutation gain at locus a in a subtree if a strict subset of the leaves have the mutation.

We summarize these leaf constraints on $\mathbf{B}'[L(T_v)]$ as a ternary matrix $\overline{\mathbf{B}}'_v = [\overline{\mathbf{b}}'_v]_{v \in L(T_v)}$ where $\overline{\mathbf{b}}_v = [\overline{b}'_{v,a}] \in \{0,1,?\}^m$. The first constraint fixes the values for some entries of $\mathbf{B}'[L(T_v)]$, such that $\overline{b}'_{p(r(w)),a} = 1$ when $b'_{r(w),a} = 1$, and $\overline{b}'_{p(r(w)),a} = 0$ when $b'_{r(w),a} = 0$ and $a \notin \mathcal{L}(c_v, c_w)$. The second constraint further sets some of the previously non-fixed entries in $\mathbf{B}'[L(T_v)]$ to minimize the total number of mutation gains in T'_v. If there exist leaves $v', w' \in L(T'_v)$ where $b_{v',a} = 0$ and $b_{w',a} = 1$), then mutation a must be gained in subtree T'_v. To achieve the minimum number of mutation gains, we thus maximize the number of all-zero and all-one columns of $\overline{\mathbf{B}}'_v$: we set to 0 any previously undetermined entries $\overline{b}_{v',a}$ for columns of $\overline{\mathbf{B}}'$ that only have '0' ('1', resp.) entries (setting of $\overline{b}_{v',a} = 0$, $\overline{b}_{v',a} = 1$ resp.). At last, we set any remaining undetermined entry of $\overline{\mathbf{B}}'_v$ to be '?'.

Finally, we aim to find $\mathbf{B}'[L(T_v)]$ by filling the '?' entries of $\overline{\mathbf{B}}'_v$. More specifically, given $\overline{\mathbf{B}}'_v$, we seek $\mathbf{B}'[L(T_v)]$ such that if $\overline{b}_{v',a} \in \{0,1\}$ then $b'_{v',a} = \overline{b}'_{v',a}$ for all mutations a and $\mathbf{B}'[L(T_v)]$ is a perfect phylogeny matrix. This problem is known as the Incomplete Directed Perfect Phylogeny (IDP) problem and has been shown to be solvable in $O(n^2 m)$ time (Pe'er et al., 2004). In our case $n = |L(T_v)| = d_v$ where d_v is the out-degree of vertex v in T. Solving an instance of the IDP problem yields a perfect phylogeny mutation matrix $\mathbf{B}'[L(T_v)]$, which in turn determines the perfect phylogeny tree T'_v and mutation matrix $\mathbf{B}'[L(T'_v)]$.

Maximum Likelihood Loss-supported Refinement Problem

The LSR problem assumes that the mutation matrix $\mathbf{B}$ is error-free. In practice, we do not observe this mutation matrix $\mathbf{B}$, but instead we observe read counts from a sequencing experiment. Specifically, we measure a variant read count matrix $\mathbf{X} = [\mathbf{x}_v]_{v \in L(T)}$ and a total read count matrix $\mathbf{Y} = [\mathbf{y}_v]_{v \in L(T)}$, where $x_{v,a} \in \mathbb{N}$ is the number of variant reads at locus a in cell v and $y_{v,a} \in \mathbb{N}$ is the total number of reads. Whole-genome amplification (Gawad et al., 2016), which typically precedes single-cell DNA sequencing, introduces a considerable amount of error into these read count matrices. Specifically, single-cell sequencing SNV data has high rates of false negative errors (i.e., $x_{v,a} = 0$ when $b_{v,a} = 1$) and missing data (i.e., $y_{v,a} = 0$). In addition, sequencing and whole-genome amplification introduce false positive errors (i.e., $x_{v,a} > 0$ when $b_{v,a} = 0$) as well. Most existing methods (Jahn et al., 2016; Malikic et al., 2019a, 2019b; Zafar et al., 2017, 2019; El-Kebir, 2018; Ross and Markowetz, 2016) for single-cell phylogeny inference discretize read counts into an

observed mutation matrix $\tilde{B}$, using either two or three genotypes in addition to missing data (i.e $\tilde{b}_{v,a} \in \{0, 1, ?\}$ or $\tilde{b}_{v,a} \in \{00, 01, 11, ?\}$). However, discretizing the mutation data loses information about the likelihood of errors. For example, a locus with a single variant read is far more likely to be a false positive error than a locus with hundreds of variant reads, but a discretized mutation matrix does not distinguish between these cases.

We use a maximum-likelihood approach to model the observed variant and total read counts. Specifically, we aim to find the mutation matrix $\mathbf{B}^* = \text{argmax}\ \Pr(\mathbf{XY}, \mathbf{B})$ that admits a solution $T', \mathbf{B}', \mathbf{c}'$ to the LSR problem and maximizes the likelihood of the observed variant read counts $\mathbf{X}$ given the total read counts $\mathbf{Y}$. Our approach to compute $\mathbf{B}^*$ is not specific to a particular likelihood model for read counts but does assume that the likelihood has the form $\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}) = \prod_{v=1}^{n} \prod_{a=1}^{m} \Pr(x_{v,a}|y_{v,a}, b_{v,a})$; i.e. the variant read counts $\mathbf{X}$ are independent of each other across cells and loci given $\mathbf{Y}$ and $\mathbf{B}$. In this work, we used a beta-binomial model similar to the one previously used by SciΦ (Singer et al., 2018). If mutation a is absent in cell v (i.e., $b_{v,a} = 0$), then the probability of observing a variant read corresponds to the per-nucleotide rate of sequencing error ε. For Illumina sequencing reads, we use $\varepsilon = 0.001$. If mutation a is present in cell v (i.e., $b_{v,a} = 1$), then we model the variant counts at a locus using a beta-binomial distribution. We estimate parameters α and β empirically from the distribution of heterozygous germline single-nucleotide polymorphisms (SNPs) in the data. We thus define the data likelihood for observing $x_{v,a}$ variant reads at locus a in cell v as follows:

$$\Pr(x_{v,a}|y_{v,a},\ b_{v,\ a}) = \begin{cases} \text{Beta} - \text{Binomial}(x_{v,a}|n = y_{v,a},\ \alpha,\ \beta) & \text{if } b_{v,a} = 1, \\ \text{Binomial}(x_{v,a}|n = y_{v,a},\ p = \varepsilon) & |\text{if } b_{v,a} = 0. \end{cases}$$

Let $\mathcal{B}_{T,\mathbf{c},\mathcal{L}}$ be the set of mutation matrices $\mathbf{B}$ such that there exists a solution $T', \mathbf{c}', \mathbf{B}'$ to the LSR problem given T, $\mathbf{c}$, $\mathcal{L}$, and B. We formulate the problem as follows.

Problem 2 ***Maximum Likelihood Loss-Supported Refinement (ML-LSR) problem*** *Given variant read counts* $\boldsymbol{X} = [\boldsymbol{x}_v]_{v \in L(T)}$*, total read counts* $\boldsymbol{Y} = [\boldsymbol{y}_v]_{v \in L(T)}$*, copy-number tree* T*, copy-number profile vector* $\boldsymbol{c} = [c_v]_{v \in V(T)}$*, and supported losses* $\mathcal{L}$*, find* $\mathbf{B}^* = \text{argmax}_{\mathcal{B}_{T,\mathbf{c},\mathcal{L}}}\ \Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B})$.

We show the ML-LSR is NP-hard by reduction from the Minimum Flip Problem (Chen et al., 2006) in "Proofs". Since current datasets have mutation matrices with hundreds–thousands of cells, we derive an algorithm in the next section that finds an approximate solution to the ML-LSR problem by subdividing the ML-LSR problem into k instances of the maximum likelihood Incomplete Directed Perfect Phylogeny problem.

SCARLET Algorithm for Maximum-Likelihood Loss-Supported Refinement Problem

We introduce SCARLET (Single-Cell Algorithm for Reconstructing Loss-supported Evolution of Tumors), an algorithm to find a loss-supported phylogeny from single-cell DNA sequencing data. SCARLET aims to solve the ML-LSR problem, defined above in Problem 2, by finding the maximum likelihood mutation matrix $\mathbf{B}^*$. Since a solution $\mathbf{B}^*$ of the ML-LSR is in $\mathcal{B}_{T,\mathbf{c},\mathcal{L}}$, there exists at least one tree T', a copy-number profile vector $\mathbf{c}'$, and an augmented mutation matrix $\mathbf{B}'$ of $\mathbf{B}^*$ such that $(T', \mathbf{B}', \mathbf{c}')$ is a solution to the LSR problem. Given solution $(T', \mathbf{B}', \mathbf{c}')$, $\mathbf{B}^*$ is uniquely determined as $\mathbf{B}^* = [\mathbf{b}'_v]_{v \in L(T)}$. We thus proceed here by finding a solution $(T', \mathbf{B}', \mathbf{c}')$ to the LSR that yields a maximum-likelihood $\mathbf{B}^*$. To solve ML-LSR problem, we extend the algorithm we previously presented to solve the LSR problem. The LSR problem decomposes into a set of IDP instances if we know the mutation profiles $\mathbf{R} = [\mathbf{b}'_{r(v)}]_{v \in I(T)}$ of the roots of subtrees $\mathcal{T}$. In the LSR, we computed $\mathbf{R}$ recursively, starting with the leaves $L(T)$ whose mutation profiles are given by $\mathbf{B}$. In the ML-LSR, however, we are not given $\mathbf{B}$, and thus do not know $\mathbf{R}$. Therefore, SCARLET uses two-step procedure where we first compute the maximum-likelihood mutation profiles $\mathbf{R}^*$ of the roots and then independently infer each maximum-likelihood refinement subtree given $\mathbf{R}^*$. Note that this two-step procedure is not guaranteed to find the overall maximum likelihood solution $\mathbf{B}^*$, as there may be cases where $\mathbf{B}^*$ does not admit a solution with the maximum-likelihood roots $\mathbf{R}*$. However, we show in Results that SCARLET is both accurate and fast in practice.

Finding Maximum Likelihood Subtree Roots

SCARLET aims to find the maximum-likelihood subtree roots $\mathbf{R}^* = [\mathbf{r}_v]_{v \in V(T)}$ such that there exists a loss-supported refinement T', $\mathbf{c}'$, $\mathbf{B}'$ with subtree roots $\mathbf{R}^*$. The existence of a solution T', $\mathbf{c}'$, $\mathbf{B}'$ constrains the possible mutation profiles of the roots. Specifically, by Definition [def:lsp] of a loss-supported phylogeny, a mutation at locus a is gained at most once in T'. Matrix $\mathbf{R}$ is a valid mutation state assignment for roots provided for each locus a, it is possible that a mutation at locus a occurred exactly once and was only lost when the loss was supported. Specifically, (1) there exists a subtree T_a of T such that for all $v \in V(T)$, $r_{v,a} = 1$ if $v \in V(T_a)$ and v is not the root of T_a and $r_{v,a} = 0$ otherwise; and (2) for any edge $(v, w) \in T$ such that $v \in T_a$ and $w \notin T_a$, $a \in \mathcal{L}(c_v, c_w)$. Any valid $\mathbf{R}$ uniquely defines a subtree T_a for each locus a. Roots $\mathbf{R}$ admit a mutation profile $\mathbf{b}_a$ for locus a provided that $\mathbf{b}_a$ satisfies the following.

1. If $v \notin T_a$ then mutation a is absent in all cells v' such that $c_{v'} = c_v$.
2. If $v \in T_a$ and v is not the root of T_a then mutation a is present in all cells v' such that $c_{v'} = c_v$.
3. If $v \in T_a$ and v is the root of T_a then mutation a is either present or absent, as mutation a occurred in T'_v

The likelihood given roots $\mathbf{R}$ is computed by marginalizing over all admitted mutation profiles. Let $\beta_{\mathbf{R}} = \{\mathbf{b}_a : \mathbf{R} \text{ admits } \mathbf{b}_a\}$ be the set of mutation profiles for mutation a admitted by roots $\mathbf{R}$. Then:

$$\Pr(\mathbf{X}|\mathbf{Y},\mathbf{R}) = \prod_{a=1}^{m} \prod_{\mathbf{b}_a \in \beta_{\mathbf{R}}} \Pr(\mathbf{x}_a|\mathbf{y}_a,\mathbf{b}_a) = \prod_{a=1}^{m} \prod_{v=1}^{n} Pr(x_{v',a}|y_{v',a},\mathbf{R})$$

such that

$$Pr(x_{v',a}|y_{v',a},\mathbf{R}) = \begin{cases} \Pr(x_{v',a}|y_{v',a},b_{v',a}=0) & \text{if } c_{v'} = c_v \text{ and } v \notin S_a \\ \Pr(x_{v',a}|y_{v',a},b_{v',a}=1) & \text{if } c_{v'} = c_v, v \in S_a \text{ and } v \text{ is not root of } S_a \\ \frac{1}{2}\Pr(x_{v',a}|y_{v',a},b_{v',a}=1) + \frac{1}{2}\Pr(x_{v',a}|y_{v',a},b_{v',a}=0) & \\ & \text{otherwise.} \end{cases}$$

We thus find $\mathbf{R}^*$ by enumerating valid mutation state assignments for roots for each mutation locus a, then computing the maximum likelihood as above.

Finding Refinement Subtrees

As input for the ML-IDP, we define a ternary matrix $\overline{\mathbf{B}}'_v = [\overline{\mathbf{b}'}_w]_{w \in L(T'_v)}$ for each vertex $v \in V(T)$ as before. For $v \in I(T)$, we define $\overline{\mathbf{b}'}_{p(r(v))}$ as previously given the mutation profile $\mathbf{b}'_{r(v)}$ of the root $r(v)$ of T'_v. For $a \in \mathcal{L}(c'_v, c'_w)$, we have that $\overline{\mathbf{b}'}_{p(r(v))} = \mathbf{b}_v$ but unlike in the LSR problem, we are not given the mutation profile $\mathbf{b}_v$ in the ML-LSR problem. Instead, we compute the likelihood of $\mathbf{b}_v$ as in Equation [eq:likelihood]. As such, finding $\mathbf{B}^*$ is equivalent to find the maximum likelihood submatrices $\{\mathbf{B}^*[\{v : c_w = c_v\}] : v \in I(T)\}$ such that $\overline{\mathbf{B}}'_v$ admits an incomplete directed perfect phylogeny.

We describe an integer-linear programming (ILP) formulation to compute these maximum likelihood submatrices. Given ternary matrix $\overline{\mathbf{B}}'_v$ and read count matrices $\mathbf{X}_v = [\mathbf{x}_w]_{w \in L(T_v)}$, $\mathbf{Y}_v = [\mathbf{y}_w]_{w \in L(T_v)}$, we aim to find matrix $\mathbf{B}'$ where $\Pr(\mathbf{X}_v|\mathbf{Y}_v, \mathbf{B}'_v)$ is maximized subject to two constraints: (1) $\mathbf{B}'_v$ is perfect phylogeny matrix, and (2) $b'_{v,w,a} = \overline{b}'_{v,w,a}$ if $\overline{b}'_{v,w,a} \neq ?$. Let $\mathbf{B}'_v \in \mathcal{P}_{\overline{\mathbf{B}}'_v}$ indicate that constraints (1) and (2) are met. We thus aim to find $\mathbf{B}'^*_v = \text{argmax}_{\mathbf{B}'_v \in \mathcal{P}_{\overline{\mathbf{B}}'_v}} \Pr(\mathbf{X}_v|\mathbf{Y}_v, \mathbf{B}'_v)$ and we design an integer linear program (ILP) to find $\mathbf{B}'^*_v$. For simplicity in the remainder of this section, we do not include subscripts for v – e.g., $\mathbf{B}' = \mathbf{B}'_v$, $\mathbf{X} = \mathbf{X}_v$, $\mathbf{Y} = \mathbf{Y}_v$. Below, we derive a linear objective for the ILP.

$$\begin{aligned} \Pr(\mathbf{X}|\mathbf{Y},\mathbf{B}') &= \text{argmax}_{\mathbf{B}'_v \in \mathcal{P}_{\overline{\mathbf{B}}'_v}} \sum_w \sum_a \log\Pr\left(x_{w,a}|y_{w,a},b'_{w,a}\right) \\ &= \text{argmax}_{\mathbf{B}'_v \in \mathcal{P}_{\overline{\mathbf{B}}'_v}} \sum_w \sum_a \left[\log\Pr\left(x_{w,a}|y_{w,a},b'_{w,a}\right) - \log\Pr\left(x_{w,a}|y_{w,a},b'_{w,a}=0\right)\right] \\ &= \text{argmax}_{\mathbf{B}'_v \in \mathcal{P}_{\overline{\mathbf{B}}'_v}} \sum_w \sum_a b'_{w,a}\left[\log\Pr\left(x_{w,a}|y_{w,a},b'_{w,a}=1\right) - \log\Pr\left(x_{w,a}|y_{w,a},b'_{w,a}=0\right)\right] \\ &= \text{argmax}_{\mathbf{B}'_v \in \mathcal{P}_{\overline{\mathbf{B}}'_v}} \sum_w \sum_a b'_{w,a} \cdot C_{w,a} \end{aligned}$$

where for observed cells w,

$$C_{w,a} = \log\Pr(x_{w,a}|y_{w,a},b_{w,a}=1) - \log\Pr(x_{w,a}|y_{w,a},b_{w,a}=0).$$

For unobserved cells, we constrain that $b'_{w,a} = \overline{b}'_{w,a}$ if $\overline{b}'_{w,a} \neq ?$ by setting $C_{w,a}$ as follows:

$$C_{w,a} = \begin{cases} M & \overline{b}'_{w,a} = 1 \\ -M & \overline{b}'_{w,a} = 0 \\ 0 & \overline{b}'_{w,a} = ?, \end{cases}$$

where M is a large constant. We use an ILP to maximize $\sum_w \sum_a b'_{w,a} \cdot C_{w,a}$ subject to $\mathbf{B}'$ being a perfect phylogeny matrix. We introduce a set of auxiliary variables F, G, H to enforce the three gametes condition, where $F_{a,b}$, $G_{a,b}$ and $H_{a,b}$ indicate that a pair of columns a, b show the $(1,1)$, $(0,1)$ and $(1,0)$ gametes respectively, and $F_{w,a,b}$, $G_{w,a,b}$ and $H_{w,a,b}$ indicate that $(b'_{w,a}, b'_{w,b})$ show the $(1,1)$, $(0,1)$ and $(1,0)$ gametes respectively. All auxiliary variables are constrained to be binary. This yields the following ILP.

$$
\begin{aligned}
\text{maximize}\quad & \sum_{w,a} b'_{w,a} \cdot C_{w,a} \\
& \text{subject to} \\
& F_{a,b} + G_{a,b} + H_{a,b} \leq 2 && \text{for all } a,b \\
& b'_{w,a} + b'_{w,b} - 1 \leq F_{w,a,b} \leq \min\left(b'_{w,a}, b'_{w,b}\right) && \text{for all } a,b,w \\
& \max_w F_{w,a,b} \leq F_{a,b} \leq \sum_w F_{w,a,b} && \text{for all } a,b \\
& -b'_{w,a} + b'_{w,b} \leq G_{w,a,b} \leq \min\left(1 - b'_{w,a}, b'_{w,b}\right) && \text{for all } a,b,w \\
& \max_w G_{w,a,b} \leq G_{a,b} \leq \sum_w G_{w,a,b} && \text{for all } a,b \\
& b'_{w,a} - b'_{w,b} \leq H_{w,a,b} \leq \min\left(b'_{w,a}, 1 - b'_{w,b}\right) && \text{for all } a,b,w \\
& \max_w H_{w,a,b} \leq H_{a,b} \leq \sum_w H_{w,a,b} && \text{for all } a,b
\end{aligned}
$$

Proofs

Proof of Theorem 1

Proof.

We first show that any solution that meets these constraints is a solution to the LSR problem. Constraint (1) of the LSR problem is explicitly enforced by condition (4) of Theorem 1. Constraint (2) is that T' is a loss-supported phylogeny, i.e., every mutation occurs at most once (enforced by condition (1) of Theorem 1), and every mutation loss is supported. By condition (2), there are no mutation losses between cells that have the same copy-number state, and by (3) mutation losses that are not supported are not allowed between cells with different copy-number states. Thus, if $(T', \mathbf{B}', \mathbf{c}')$ meet the conditions of Theorem 1, then $(T', \mathbf{B}', \mathbf{c}')$ is a solution to the LSR.

We next show that any solution $(T', \mathbf{B}', \mathbf{c}')$ to the LSR meets the four conditions stated in Theorem 1. We will do this by showing that any (T', $\mathbf{B}'$, $\mathbf{c}'$) that violates any one of these constraints cannot be a solution to the LSR.

1. This condition is directly required by the definition of a loss-supported phylogeny
2. If this condition does not hold, there exists a mutation that is lost in subtree $T'[\gamma(v)]$. As every vertex in $T'[\gamma(v)]$ has the same copy-number state (by condition (4), this mutation loss is not supported and thus T' is not a solution to the LSR problem.
3. If this condition is violated, then there is a mutation loss that is not supported and thus T' is not a solution to the LSR problem.
4. This condition is directly required by the LSR problem statement.

Proof of Lemma 1

Proof.

We will show by construction that for any solution $(T', \mathbf{B}', \mathbf{c}')$ that violates these constraints, there exists another solution $(T'', \mathbf{c}'', \mathbf{B}'')$ that meets these constraints.

(1) For all $(\boldsymbol{v},\boldsymbol{w}) \in \boldsymbol{E}(\boldsymbol{T})$, $\boldsymbol{p}(\boldsymbol{r}(\boldsymbol{w}))$ is a leaf of subtree $\boldsymbol{T}'[\boldsymbol{\gamma}(\boldsymbol{v})]$.

Consider an edge $(p(r(w)), r(w)) \in E(T')$ such that $p(r(w))$ is not a leaf of $T'[\gamma(v)]$. That is, $p(r(w))$ has another child in $T'[\gamma(v)]$. We construct T'' by splitting $p(r(w))$ into two vertices u and u' such that there is an edge $(u,u') \in E(T'')$, $\mathbf{b}''_u = \mathbf{b}''_{u'}$ and $c''_u = c''_{u'}$, and the only outgoing edge from u' is $(u', r(w))$. Thus, u' is now the parent of $r(w)$ and u' is a leaf. This split preserves the rest of the tree and does not introduce violations of any of the conditions in Theorem 1 or any of the other assumptions in this Lemma. Thus T'', $\mathbf{B}''$, $\mathbf{c}''$ is a solution to the LSR problem.

(2) For all $\boldsymbol{v} \in \boldsymbol{V}(\boldsymbol{T})$ such that $\boldsymbol{v}$ is not the root of $\boldsymbol{T}$, $\boldsymbol{b}_{\boldsymbol{r}(\boldsymbol{v})} = 1$ if $\boldsymbol{b}_{\boldsymbol{v}'} = 1$ for all $\boldsymbol{v}' \in \boldsymbol{L}(\boldsymbol{T}'[\boldsymbol{\gamma}(\boldsymbol{v})])$.

Assume that $T', \mathbf{c}', \mathbf{B}'$ meet condition 1. If this constraint is violated, this means that there is some mutation a that is gained in a subtree $T'[\gamma(v)]$ but there are no leaves of $T'[\gamma(v)]$ that do not contain a. Let $T'' = T'$, $c'' = c'$. Let $b''_{v',a} = 1$ if $v \in T'[\gamma(v)]$. This change does not violate any of the conditions in Theorem 1. Specifically, this change does not introduce new mutation gains, and as this only alters the mutation profiles of internal vertices of $T'[\gamma(v)]$ so this cannot introduce new mutation losses. As T'' and $\mathbf{c}''$ are preserved, refinement and copy-number consistency conditions are automatically met. This change may introduce violations to Assumption 3 in this Lemma that can subsequently be corrected as below.

(3) **For all $\boldsymbol{v} \in \boldsymbol{V}(\boldsymbol{T})$ and all loci $\boldsymbol{a}$, $\boldsymbol{b}_{\boldsymbol{p}(\boldsymbol{r}(\boldsymbol{v})),\boldsymbol{a}} \geq \boldsymbol{b}_{\boldsymbol{r}(\boldsymbol{v}),\boldsymbol{a}}$.**

This constraint states that there are no mutation gains on edges between subtrees. We construct T'' by performing a similar split as we did for constraint (1). Suppose there's an edge $(p(r(v)), r(v)) \in E(T')$ such that $b_{p(r(v)),a} = 0$ and $b_{r(v),a} = 1$. Split $p(r(v))$ into vertices u, u' such that $b_{u,a} = 0$ and $b_{u',a} = 1$, and for all $a' \neq a$, $b_{u,a'} = b_{u',a'}$ and the only outgoing edge from u' is $(u', r(w))$. This split preserves the

rest of the tree and does not introduce violations of any of the conditions in Theorem 1 or any of the other assumptions in this Lemma. Thus T'', $\mathbf{B}''$, $\mathbf{c}''$ is a solution to the LSR problem.

Hardness of ML-LSR

Lemma 2 *The ML-LSR is NP-hard*.

Proof.

We show this by reduction from the Flip problem (Chen et al., 2006) which is known to be NP-Complete.

Given a binary matrix $\mathbf{B} \in \{0,1\}^{m \times n}$ and integer $\kappa \in \mathbb{N}$, decide whether there exists a directed perfect phylogeny matrix $\mathbf{B}' \in \{0,1\}^{m \times n}$ such that no more than κ entries in $\mathbf{B}'$ differ from $\mathbf{B}$.

Let $(\mathbf{B}, \kappa)$ be an instance of the Flip problem. For the corresponding instance of the ML-LSR problem, we let $k = 1$ and define the inputs as follows:

1. T is the star phylogeny, where all leaves $v \in V(T)$ are attached to a single internal vertex;
2. $\mathbf{c} = \overrightarrow{1}$;
3. No mutation losses are supported in $\mathcal{L}$;
4. $\mathbf{X} = \mathbf{B}$, and $\mathbf{Y} = [1]^{m \times n}$

Define a likelihood function $\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}^*)$ that is symmetric when $x_{v,a} \in \{0,1\}$ and $y_{v,a} = 1$

$$\log\left(\Pr\left(x_{v,a}|y_{v,a}, b_{v,a}\right)\right) = \begin{cases} \alpha & \text{if } x_{v,a} = b_{v,a} \\ \beta & \text{if } x_{v,a} \neq b_{v,a} \end{cases}$$

such that $\beta < \alpha$. Thus the log-likelihood of a matrix $\mathbf{B}^*$ is

$$\begin{aligned} \log\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}^*) &= \sum_{v=1}^{n} \sum_{a=1}^{m} \log\left(\Pr\left(x_{v,a}|y_{v,a}, b_{v,a}\right)\right) \\ &= \lambda \cdot \beta + (mn - \lambda)\alpha \end{aligned}$$

We claim that there exists a perfect phylogeny matrix $\mathbf{B}'$ with at most κ changes if and only if there exists a solution $\mathbf{B}^*$ to the ML-LSR

$$\log\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}^*) \geq \kappa \cdot \beta + (mn - \kappa) \cdot \alpha$$

We first show the forward direction. If there exists a perfect phylogeny matrix $\mathbf{B}'$ with at most κ changes from $\mathbf{B}' = \mathbf{X}$, then the log-likelihood $\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}') \geq \kappa \cdot \beta + (mn - \kappa) \cdot \alpha$. Thus for the maximum likelihood solution, $\mathbf{B}^*$, $\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}^*) \geq \Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}')$.

We next show the reverse direction. If $\log\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{B}^*) \geq \kappa \cdot \beta + (mn - \kappa)$ then $\mathbf{B}^*$ has at most κ changes from $\mathbf{X} = \mathbf{B}$. Thus, there exists a $\mathbf{B}' = \mathbf{B}^*$.

QUANTIFICATION AND STATISTICAL ANALYSIS

Simulation Details

We simulated 50 single-cell DNA sequencing datasets, each data set containing n=100 observed cells that were related by a phylogenetic tree containing m=20 mutations, and k=4 copy-number profiles. We simulated each data set in four steps. First, we simulated the topology of a tree. $m + k + 1 = 25$ vertices were randomly assigned to be in the trunk of the tree or in one of the k copy-number profiles. Vertices in the trunk were joined into a linear path, and vertices not in the trunk were assigned attachments uniformly at random, such that vertices in the same copy-number profile form connected subtrees. We assign the m mutations onto the 20 edges without copy-number profile changes. For each edge (v, w) in the simulated tree with a change in the copy-number profiles c and c', we also simulated the set $\mathcal{L}(c, c')$ of supported losses by selecting a random subset of the m loci such that $\left|\mathcal{L}(c, c')\right| \sim \text{Poisson}(0.2 * m)$. Third, we introduced with probability 0.5 a mutation loss in every genomic locus $a \in \mathcal{L}(c, c')$ if the mutation is contained in the parent p. To respect the k-Dollo model with $k = 1$, we enforce that the same mutation is lost at most once in the simulated tree. We thus obtained simulated trees with 1–8 mutation losses. Last, we add 100 leaves, corresponding to the observed cells, and we append those to a random vertex of the simulated tree.

We simulated read counts from all the cells of each simulated tree with errors specific of single-cell DNA sequencing data. Specifically, we generated a total read count $y_{v,a}$ and a variant read count $x_{v,a}$ for each locus a in cell v with an allelic dropout rate of $d = 0.15$, according to previous analyses (Gawad et al., 2016). First, we generated $y_{v,a}$ according to a Poisson distribution and assuming an expected sequencing coverage of 100× such that $y_{v,a} \sim \text{Poisson}(100)$. Note that when both the alleles drop out, $y_{v,a} = 0$. Second, we generated $x_{v,a}$ according to either the absence or presence of a mutation in locus a. If the variant is absent, $x_{v,a} \sim \text{Binomial}(t_{v,a}, \varepsilon)$ where $\varepsilon = 0.001$ models the sequencing error rate. If the variant is present, we model the overdispersion in the variant read count $x_{v,a}$ resulting from whole-genome amplification using a Beta-Binomial model as in previous studies (Singer et al., 2018) such that $x_{v,a} \sim \text{Binomial}(t_{v,a}, f_{v,a})$ and $f_{v,a} \sim \max\{\text{Beta}(\alpha, \alpha), \varepsilon\}$ (with $\alpha = 0.25$ in order to obtain an allele dropout rate of $d \approx 0.15$).

Copy-Number Analysis of Colorectal Cancer Patient

We describe the analysis of copy number aberrations in colorectal cancer patient CRC2 from Leung et al. (2017), which provides part of the input data for SCARLET. Leung et al. (2017) performed single-cell DNA sequencing of a 1000 cancer gene panel from 186 cells from a primary tumor and metastasis. We computed copy-number profiles **c** and supported losses $\mathcal{L}$ from read-depth ratios as follows.

First, we computed read depth ratios in 10-Mb genomic bins by calculating the read depth $r_{v,i}$ in every bin i of every cell v, as the number of sequencing reads that align to the bin. To account for context-specific variation in read depth, we normalized $r_{v,i}$ using the corresponding read depth n_i in a matched normal sample. Moreover, to account for shifts in read depth due to differences in the ploidy of each cell, we further corrected $r_{v,i}$ by using the ploidy φ_v of cells v measured by DAPI staining (Leung et al., 2017) ($\varphi_v =$ 3.3 for primary aneuploid cells, $\varphi_v = 3.0$ for metastatic aneuploid cells, $\varphi_v = 2.0$ for diploid cells). Therefore, we obtain the resulting corrected read-depth ratio $\hat{r}_{v,i} = \frac{r_{v,i}}{n_i}\frac{\varphi_v}{2}$ for every bin i in cell v. We performed hierarchical clustering on read depth ratios $\hat{r}_v$ for all cells v to infer copy-number profiles c. In particular, we fixed the number of clusters to 4 according to the the number of copy-number clones previously identified (Leung et al., 2017).

We identified sets of supported losses in the same 186 cells by identifying significant shifts in the read depths of the bins that contain the 36 somatic single-nucleotide variants previously identified by Leung et al. (2017). To test whether there was a loss of variant a in bin i between copy-number profiles j and k, we performed a signed Wilcoxon rank-sum test. The two groups of observations correspond to cells with copy-number profiles j and k, such that $G_j = \{\hat{r}_{v,a}; c_v = j\}$ and $G_k = \{\hat{r}_{v,a}; c_v = k\}$. The Wilcoxon Rank-sum Test tests whether observations in G_j and G_k are drawn from the same distribution. A mutation loss was supported if the test yielded a p-value $p<.01$.

DATA AND CODE AVAILABILITY

SCARLET software, simulated data, and processed CRC2 data are available at github.com/raphael-group/scarlet. Original CRC2 data was downloaded from NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP074289.

Cell Systems, Volume *10*

Supplemental Information

SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses

Gryte Satas, Simone Zaccaria, Geoffrey Mon, and Benjamin J. Raphael

Supplementary Figures

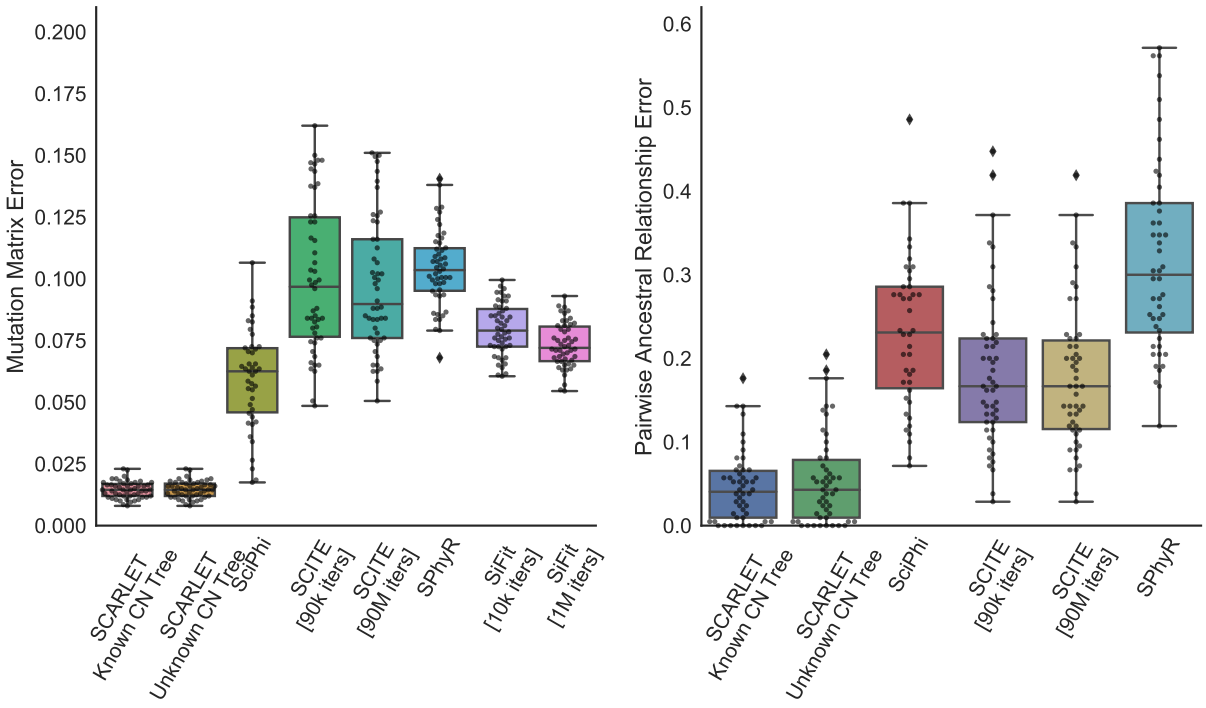


Figure 1: SCARLET Results on Simulated Data, with extended runtime of SCITE and SiFit. Related to Figure 2.