

# Supplementary Note

## Inferring Parsimonious Migration Histories for Metastatic Cancers

Mohammed El-Kebir<sup>1,3</sup>, Gryte Satas<sup>1,2</sup>, and Benjamin J. Raphael<sup>1,\*</sup>

<sup>1</sup>Department of Computer Science, Princeton University, Princeton, NJ 08540.

<sup>2</sup>Department of Computer Science, Brown University, Providence, RI 02912.

<sup>3</sup>Present address: Department of Computer Science, University of Illinois at Urbana-Champaign,  
Urbana, IL 61801

\*Correspondence: braphael@princeton.edu

### Contents

<b>A Results</b>	<b>4</b>
A.1 Simulations . . . . .	4
A.1.1 Simulation Setup . . . . .	4
A.1.2 Methods . . . . .	7
A.1.3 Results . . . . .	8
A.2 Metastatic Colorectal Cancer . . . . .	12
A.3 Metastatic Ovarian Cancer . . . . .	13
A.4 Metastatic Prostate Cancer . . . . .	14
A.5 Metastatic Melanoma . . . . .	16
A.6 Metastatic Breast Cancer . . . . .	16
<b>B Methods</b>	<b>18</b>
B.1 Preliminaries . . . . .	18
B.1.1 Migration Number $\mu(T, \ell)$ . . . . .	20
B.1.2 Comigration Number $\gamma(T, \ell)$ . . . . .	21
B.1.3 Seeding Site Number $\sigma(T, \ell)$ . . . . .	23
B.2 Problem Statements . . . . .	23
B.2.1 Parsimonious Migration History . . . . .	23

B.2.2	Parsimonious Migration History with Tree Refinement . . . . .	24
B.2.3	Parsimonious Migration History with Tree Inference . . . . .	25
B.3	MACHINA . . . . .	27
B.3.1	Unconstrained Parsimonious Migration History Problem . . . . .	27
B.3.2	Constrained Parsimonious Migration History Problem . . . . .	28
B.3.3	Parsimonious Migration History with Tree Refinement Problem . . . . .	31
B.3.4	Parsimonious Migration History with Tree Inference Problem . . . . .	32

**C References** **37**

**List of Figures**

1	The assumption of sample homogeneity in colorectal patient 2 leads to an unlikely clone tree . . . . .	43
2	MACHINA infers parsimonious monoclonal single-source seeding histories for breast cancer patient A7	44
3	Mutation clusters do not necessarily correspond to extant clones . . . . .	45
4	MACHINA accurately infers clone trees and migration histories for the $\Sigma_{\max} = 5$ simulated instances	46
5	MACHINA accurately infers clone trees and migration histories for the $\Sigma_{\max} = 8$ simulated instances	47
6	Polytomy refinement leads to more parsimonious solutions for ovarian cancer 1 . . . . .	48
7	Different vertex labelings exist for ovarian cancer 3 with fewer migrations . . . . .	49
8	Resolving polytomies in ovarian cancer patient 7 suggests a more likely primary tumor . . . . .	49
9	MACHINA infers parsimonious migration histories without met-to-met spread for four prostate cancers	50
10	Prostate cancers A10, A31 and A32 may have undergone parallel seeding . . . . .	51
11	MACHINA recapitulates the cases of polyclonal seeding and non-serial progression in melanoma . . .	52
12	The method by McPherson et al. identifies complex mM seeding patterns for breast cancer A7 . . . .	52
13	MACHINA infers a polyclonal parallel single-source seeding migration history for breast cancer pa- tient A1 . . . . .	53
14	Treomics and breast cancer patient A1 . . . . .	54
15	An agent-based model for simulating metastatic tumors . . . . .	55
16	Comparison of different clustering algorithms on the $\Sigma_{\max} = 5$ simulation instances . . . . .	56
17	MACHINA performance with different mutation clustering algorithms . . . . .	57
18	Performance of MACHINA for varying purity, sequence coverage, and number of samples . . . . .	58
19	Performance of MACHINA for varying number of SNVs . . . . .	59
20	There exists a tradeoff between the migration number $\mu$ and the comigration number $\gamma$ . . . . .	60
21	The comigration number $\gamma(T, \ell)$ may not equal the number of multi-edges in a migration graph $G$ with directed cycles . . . . .	60
22	The migration history helps resolve polytomies in clone trees . . . . .	61
23	Bulk samples of a tumor are mixtures of the leaves of an unknown phylogenetic tree . . . . .	62

24	Binarizations are spanning trees of a directed graph . . . . .	63
----	--	----

## List of Tables

1	Reported clustering [19] of mutations 47 and 81 in patient A7 is likely incorrect . . . . .	64
2	Simulated instances and MACHINA results with $\Sigma_{\max} = 5$ anatomical sites . . . . .	65
3	Simulated instances and MACHINA results with $\Sigma_{\max} = 8$ anatomical sites . . . . .	66
4	The clone trees inferred by Kim et al. [23] have extensive homoplasy . . . . .	67
5	Results of the Sankoff enumeration algorithm for seven ovarian cancer patients . . . . .	68
6	Results for seven ovarian cancer patients . . . . .	69
7	Results for five prostate cancer patients . . . . .	70
8	Results for five metastatic melanoma patients . . . . .	70
9	Taxonomy of migration patterns . . . . .	71

# A Results

## A.1 Simulations

We describe the simulation setup in Section A.1.1. In Section A.1.2, we provide additional details on the methods considered in the benchmarking experiments. Finally, we present the results of these experiments in Section A.1.3.

### A.1.1 Simulation Setup

We simulate the life history of a metastatic tumor by generating a cell tree using an agent-based model. This model is an extension of the model used to simulate tumor growth in [37], with the equal multiplicative fitness landscape and the logistic growth function described in [2]. We extend the model to include migrations and passenger mutations. Moreover, we constrain mutations to adhere to the infinite sites assumption and to be copy neutral. A discrete time-step, or *generation*, is composed of two phases. The first phase consists of cell cycle events—replication and death. The second phase consists of cell migration events. In the following, we will elaborate on the two phases and describe the overall simulation algorithm.

**Cell Cycle Events: Replication and Death** A cell  $c$  has two quantities associated to it: (1) the anatomical site  $a(c)$  in which the cell resides, (2) the set  $\sigma(c)$  of mutations the cell harbors. Each mutation  $i$  has a fitness effect  $s_i$ . We model mutations as either drivers or passengers, with a driver mutation  $i$  having positive fitness effect  $s_i = 0.01$  and passenger mutation  $i$  having fitness effect  $s_i = 0$ . The subset of mutations of cell  $c$  that are drivers are denoted by  $\hat{\sigma}(c)$ . As described in [37], we use a phenotype-based density-limited model, where the phenotype of a cell  $c$  is defined by the set  $\hat{\sigma}(c)$  of driver mutations the cell contains. Let  $N(\hat{\sigma}(c))$  be the number of cells with phenotype  $\hat{\sigma}(c)$  in anatomical site  $a(c)$ , and let  $K(\hat{\sigma}(c))$  be the carrying capacity for phenotype  $\hat{\sigma}(c)$  in anatomical site  $a(c)$ . Then, we define the birth probability as

$$b(c) = \frac{1}{2} \prod_{i \in \sigma(c)} \left[ (1 + s_i) \cdot \left( 1 - \frac{N(\hat{\sigma}(c))}{K(\hat{\sigma}(c))} \right) \right]$$

where the carrying capacity for phenotype  $\hat{\sigma}(c)$  depends on the number  $|\hat{\sigma}(c)|$  of driver mutations in cell  $c$ , i.e.

$$K(\hat{\sigma}(c)) = 50000 \cdot |\hat{\sigma}(c)|.$$

In each generation, cell  $c$  replicates with probability  $b(c)$  and dies with probability  $d(c) = 1 - b(c)$ . If cell  $c$  replicates, we generate two new daughter cells  $c_1$  and  $c_2$ . Daughter cell  $c_1$  is identical to  $c$ , i.e.

$$\begin{aligned} a(c_1) &= a(c), \\ \sigma(c_1) &= \sigma(c). \end{aligned}$$

With a probability 0.1, daughter cell  $c_2$  acquires one new mutation such that

$$\begin{aligned} a(c_2) &= a(c), \\ \sigma(c_2) &= \sigma(c) \cup \{n + 1\}. \end{aligned}$$

where  $n$  is the total number of distinct mutations across all cells. This mutation rate is consistent with observed per-nucleotide somatic mutation rates [29] and whole-exome sequencing. With probability  $\hat{p} = 2 \cdot 10^{-7}$ , mutation  $n + 1$  is a driver with  $s_{n+1} = 0.01$ , and else, with probability  $1 - \hat{p}$ , mutation  $n + 1$  is a passenger with  $s_{n+1} = 0$ .

**Migration** We restrict migrations so that the resulting migration graph  $G$  adheres to a pre-specified migration pattern defined as follows.

- Monoclonal single-source seeding (mS), where every generation at most one single cell per anatomical site  $s$  migrates to a new anatomical site  $t$ .
- Polyclonal single-source seeding (pS), where every generation at most one group of cells migrates from an anatomical site  $s$  to a new anatomical site  $t$ , or to an existing anatomical site  $t'$  that was previously seeded from anatomical site  $s$ .
- Polyclonal multi-source seeding (pM), where every generation at most one group of cells migrates from an anatomical site  $s$  to a new anatomical site  $t$ , or to an anatomical site  $t'$  that has not seeded from an ancestor of  $s$ .
- Polyclonal reseeding (pR), where every generation at most one group of cells migrates from an anatomical site  $s$  to a new anatomical site  $t$ , or to an existing anatomical site  $t' \neq s$ .

Migration occurs after the cell cycle events (i.e. replication or death). Let  $N(s, y) = \{c \mid a(c) = s, \hat{\sigma}(c) = y\}$  be the number of cells  $c$  in anatomical site  $s$  with phenotype (driver mutations)  $\hat{\sigma}(c) = y$ . Let  $Y(s)$  be the set of phenotypes in anatomical site  $s$ . For each anatomical site  $s$ , we decide to initiate a migration event with a probability that is proportional to the number of cells in  $s$  and the number of drivers they contain, i.e.

$$10^{-6} \prod_{y \in Y(s)} (N(s, y) \cdot |y|).$$

The exact migration event depends on the specified migration pattern as follows.

- In the case of mS, we select a single cell  $c$  uniformly at random from the cells residing in anatomical site  $s$ . Subsequently, we set  $a(c) = t$  where  $t$  is a new anatomical site.
- In the case of pS, we draw the number  $k$  of migrating cells from  $\text{Poisson}(1)$ . We then pick  $k$  cells  $c_1, \dots, c_k$  without replacement from anatomical site  $s$ . Next, with probability 0.5, we decide to migrate to a new anatomical site  $t$  and set  $a(c_i) = t$  for all  $i \in [k]$ . Or, with probability 0.5, we migrate to an existing anatomical site  $t'$  that has been previously seeded from  $s$  and thus set  $a(c_i) = t'$  for all  $i \in [k]$ .
- In the case of pM, we draw the number  $k$  of migrating cells from  $\text{Poisson}(1)$ . We then pick  $k$  cells  $c_1, \dots, c_k$  without replacement from anatomical site  $s$ . Next, with probability 0.5, we decide to migrate to a new anatomical site  $t$  and set  $a(c_i) = t$  for all  $i \in [k]$ . Or, with probability 0.5, we migrate to an existing anatomical site  $t'$  that has not been seeded from an ancestor of  $s$  and set  $a(c_i) = t'$  for all  $i \in [k]$ .

- In the case of pR, we draw the number  $k$  of migrating cells from  $\text{Poisson}(1)$ . We then pick  $k$  cells  $c_1, \dots, c_k$  without replacement from anatomical site  $s$ . Next, with probability 0.5, we decide to migrate to a new anatomical site  $t$  and set  $a(c_i) = t$  for all  $i \in [k]$ . Or, with probability 0.5, we migrate to an existing anatomical site  $t' \neq s$ .

**Initialization and Termination** We initialize the simulation with a single cell  $c$  in anatomical site  $a(c) = P$  (where  $P$  is the primary tumor) with no passenger mutations and a single driver mutation, i.e.  $\sigma(c) = \hat{\sigma}(c) = \{1\}$ . An anatomical site  $s$  is *detectable* if it has at least 5000 cells. We have two termination conditions:

1. No more cells are alive.
2. The number of detectable anatomical sites is greater than a pre-specified parameter  $\Sigma_{\max}$ .

**From Cell Tree to Clone Tree** We start by identifying the set  $X$  of mutations that occur at 5% frequency in at least one anatomical site. For each cell  $c$ , we set  $\bar{\sigma}(c) = \sigma(c) \cap X$ . This yields a partition of cells into clones, i.e. each clone  $\bar{c}_s$  is a set of cells  $c$  with identical mutations  $\bar{\sigma}(c) = \sigma(\bar{c}_s)$  in the same anatomical site  $\sigma(c) = s$ . Recall that mutations in the cell tree adhere to the infinite sites assumption. As such, we obtain the clone tree from clones  $\bar{c}_s$  using the perfect phylogeny theorem [16]. More specifically, we construct a binary matrix  $B = [b_{\bar{c}_s, i}]$ , whose rows correspond to clones and columns to mutations  $X$ ; an entry  $b_{\bar{c}_s, i}$  is 1 if  $i \in \sigma(\bar{c}_s)$  and 0 otherwise. By construction, matrix  $B$  is conflict free and thus the corresponding clone tree  $T$  can be obtained in linear time [16]. Since  $B$  may contain repeated columns, the edges of  $T$  may be labeled by a set of co-occurring mutations. Each such set forms a mutation cluster. Each leaf  $u$  of  $T$  corresponds to a clone  $\bar{c}_s$  and is thus labeled by  $\hat{\ell}(u) = s$ . In addition, we label each leaf corresponding to clone  $\bar{c}_s$  by the proportion  $\alpha(\bar{c}_s)$  of cells of anatomical site  $s$  that comprise clone  $\bar{c}_s$ .

**Simulating Read Counts** Let  $T$  be the simulated clone tree and let  $m = |\Sigma|$  be the number of anatomical sites. We generate bulk sequencing samples for each anatomical site  $s$ . Each sample  $p$  is a mixture of the clones present in the corresponding anatomical site  $s$ . Let  $k$  be the number of clones in anatomical site  $s$ . We model mixture proportions  $u_{p,1}, \dots, u_{p,k}$  as a draw from  $\text{Dir}(5 \cdot \alpha(\bar{c}_{s,1}), \dots, 5 \cdot \alpha(\bar{c}_{s,k}))$ , where  $\alpha(\bar{c}_{s,j})$  is the proportion of cells in anatomical site  $s$  of clone  $\bar{c}_{s,j}$ . For each mutation  $i \in X$ , we define the frequency  $f_{p,i}$  as follows:

$$f_{p,i} = \frac{1}{2} \sum_{j \in [k]: i \in \sigma(\bar{c}_{s,j})} u_{p,j} \quad (1)$$

In other words,  $f_{p,i}$  is the true proportion of reads in sample  $p$  that contain SNV  $i$  under the assumption that the corresponding locus is part of a copy-neutral region of an autosomal chromosome. We then simulate the total number  $d_{p,i}$  of reads as a draw from  $\text{Poisson}(200)$ , yielding a number of reads that is typical in a whole-exome sequencing experiment. Next, we draw  $v_{p,i}$  variant reads from  $\text{Binomial}(d_{p,i}, f_{p,i})$ . The number  $r_{p,i}$  of reference reads is  $d_{p,i} - v_{p,i}$ . Note that we assume the tumor samples to be pure and do not model normal admixture.

**Simulated Instances** We varied the maximum number  $\Sigma_{\max} \in \{5, 8\}$  of detectable anatomical sites and considered different migration patterns: mS, pS, pM and pR. For each combination of  $\Sigma_{\max}$  and migration pattern, we simulated

ten clone trees and migration histories. Subsequently, we obtained two bulk samples of the primary tumor and a single bulk sample for each metastasis. Thus, we have 80 simulated instances in total (Tables 2 and 3).

### A.1.2 Methods

**MACHINA** As described in the main text, MACHINA’s PMH-TI mode consists of three steps:

1. Clustering of mutations, yielding frequency matrices  $(F^-, F^+)$  whose entries  $f_{p,C}^-$  and  $f_{p,C}^+$  are, respectively, the lower bound and upper bound of the confidence interval of the frequency of cluster  $C$  in sample  $p$ .
2. Inferring the set  $\bar{\mathcal{T}}$  of mutation trees given  $(F^-, F^+)$ .
3. Solving the PMH-TI for each mutation tree  $T \in \bar{\mathcal{T}}$  given  $(F^-, F^+)$ .

First, to obtain mutation clusters and their frequencies, we use the clustering procedure of AncesTree [8], which we describe briefly in the following. Let  $\hat{X}_{p,i}$  be a random variable describing the variant allele frequency (VAF) for a sample  $p$  and mutation  $i$ . For mutation  $i$  in sample  $p$ , we model the observed number  $\tilde{v}_{p,i}$  of variant reads as  $\tilde{v}_{p,i} \sim \text{Binomial}(\tilde{d}_{p,i}, \hat{X}_{p,i})$  where  $\tilde{d}_{p,i}$  is the total number of reads at the mutation locus. Assuming a uniform prior on the binomial proportion, the posterior distribution over VAF  $\hat{X}_{p,i}$  given  $\tilde{v}_{p,i}$  and  $\tilde{d}_{p,i}$  is  $\text{Beta}(1 + \tilde{v}_{p,i}, 1 + (\tilde{d}_{p,i} - \tilde{v}_{p,i}))$ . To infer mutation clusters, we compute 99.9% confidence intervals  $[\hat{f}_{p,i}^-, \hat{f}_{p,i}^+]$  on the VAF posterior distribution of  $\hat{X}_{p,i}$  for each mutation  $i$  in sample  $p$ . Using these intervals, we construct an undirected graph  $\bar{G}$ , where each vertex  $v_i \in V(\bar{G})$  corresponds to a unique mutation  $i$  and there is an edge  $(v_i, v_j) \in E(\bar{G})$  if and only if the confidence intervals  $[\hat{f}_{p,i}^-, \hat{f}_{p,i}^+]$  and  $[\hat{f}_{p,j}^-, \hat{f}_{p,j}^+]$  overlap for each sample  $p$ . Each connected component  $C$  of  $\bar{G}$  corresponds to a mutation cluster. To infer a confidence interval  $[f_{p,C}^-, f_{p,C}^+]$  on the frequency of mutation cluster  $C$  in each sample  $p$ , we combine the read counts for all mutations in the same mutation cluster  $C$ , yielding a combined variant read count  $v_{p,C} = \sum_{i \in C} v_{p,i}$  and combined total read count  $d_{p,C} = \sum_{i \in C} d_{p,i}$ . We assume that each cluster  $C$  has a fixed VAF  $X_{p,C}$ , such that for all mutations  $i \in C$ ,  $\hat{X}_{p,i} = X_{p,C}$ . Then, with a uniform prior, the posterior distribution of  $X_{p,C}$  given  $v_{p,C}$  and  $d_{p,C}$  is  $\text{Beta}(1 + v_{p,C}, 1 + (d_{p,C} - v_{p,C}))$ . For each mutation cluster  $C$  and sample  $p$ , we infer 95% confidence intervals  $[\hat{f}_{p,C}^-, \hat{f}_{p,C}^+]$  on the VAF distribution  $X_{p,C}$ . Since each mutation is copy-neutral, we obtain frequencies  $[f_{p,C}^-, f_{p,C}^+]$  by multiplying the VAFs  $[\hat{f}_{p,C}^-, \hat{f}_{p,C}^+]$  by 2. Frequency matrices  $(F^-, F^+)$  contain all frequency intervals  $[f_{p,C}^-, f_{p,C}^+]$  for each sample  $p$  and mutation cluster  $C$ .

Second, given  $(F^-, F^+)$ , we enumerate all mutation trees  $\bar{\mathcal{T}}$  using the SPRUCE algorithm [9]. Third, given  $(F^-, F^+)$ , we solve the PMH-TI for each mutation tree  $T \in \bar{\mathcal{T}}$  under various restrictions on the topology of the resulting migration graph  $G$ . That is, we restrict  $G$  to (1) an S pattern, (2) either an S or an M pattern or (3) leave  $G$  unrestricted. All instances were solved to optimality by MACHINA.

**Treomics** We run Treomics [38] twice on each instance, with and without subclone detection. For both runs, we use default arguments. All simulated instances were solved to optimality by Treomics.

**Neighbor Joining** We run the neighbor joining algorithm [42] implemented in the phangorn R package [46]. To do so, we first obtain a binary mutation matrix  $B = [b_{p,i}]$  using a variant allele frequency threshold of 0.01, i.e.  $b_{p,i} = 1$  if and only if the variant allele frequency of mutation  $i$  in sample  $p$  is at least 0.01. Next, following the default settings, we measure the distance between a pair of samples using the Hamming distance.

**PhyloWGS** We ran PhyloWGS [7] with default arguments. We define a clone  $i$  to be present in a sample  $p$  only if the inferred mixing proportion  $u_{p,i}$  is at least 0.05.

**AncesTree** We run AncesTree [8] with default arguments ( $\beta = 0.8$  and  $\gamma = 0.01$ ) and provide it with the same mutation clusters used by MACHINA. We define a clone  $i$  to be present in a sample  $p$  only if the inferred mixing proportion  $u_{p,i}$  is at least 0.05.

### A.1.3 Results

**Robinson-Foulds Distance** We use the Robinson-Foulds distance [39] to assess the accuracy of each inferred phylogenetic tree  $T$  by comparing the topology of the anatomical site labels in  $T$  to the simulated tree  $T^*$ . Each edge  $(u, v)$  of  $T$  induces a bipartition, or *split*, of the leaf set  $L(T)$  into  $\{U, V\}$  such that upon removal of  $(u, v)$  the set  $U$  contains those leaves present in the connected component that contains  $u$  and the set  $V$  contains the leaves that are present in the connected component that contains  $v$ . Rather than considering splits  $\{U, V\}$  directly, we consider  $\ell$ -splits composed of the two sets  $\{\ell(U), \ell(V)\}$  of anatomical site labels of  $\{U, V\}$ , i.e.

$$\{\ell(U), \ell(V)\} = \{\{\ell(u) \mid u \in U\}, \{\ell(v) \mid v \in V\}\}. \quad (2)$$

We note that an  $\ell$ -split  $\{\ell(U), \ell(V)\}$  is a multi-set. Let  $\mathcal{L}(T)$  be the set of all  $\ell$ -splits of a clone tree  $T$ . The Robinson-Foulds distance  $d(T^*, T)$  between the simulated tree  $T^*$  and the inferred tree  $T$  is the size of the symmetric difference between  $\mathcal{L}(T)$  and  $\mathcal{L}(T^*)$ , i.e.

$$d(T^*, T) = |(\mathcal{L}(T^*) \setminus \mathcal{L}(T)) \cup (\mathcal{L}(T) \setminus \mathcal{L}(T^*))|. \quad (3)$$

Note that [38] used a similar measure but only considered the set  $\ell(V)$  for each split  $\{U, V\}$ .

**Clone Tree Inference** Supplementary Fig. 4A and Supplementary Fig. 5A show that the clone trees inferred by MACHINA are more accurate than those inferred by Treeomics and neighbor joining. In addition, Treeomics with and without subclone detection outperforms neighbor joining. Enabling subclone detection in Treeomics-sub improves the results considerably, which is not surprising given that our simulations result in anatomical sites each composed of multiple clones. However, Treeomics-sub cannot match the performance of MACHINA and AncesTree, likely because these methods use variant allele frequencies to deconvolve mixed samples and thus are better able to detect subclones. PhyloWGS achieves similar performance to Treeomics-sub but performs worse than AncesTree and MACHINA. Importantly, MACHINA outperforms both PhyloWGS and AncesTree, thus showing that MACHINA’s performance is not only due to deconvolution but also due to the integrative inference of parsimonious clone trees and migration histories.



**Migration History Inference** Next, we assess the performance of MACHINA by considering the inferred migration graphs  $G$  and vertex labelings  $\ell$  identified for each simulated instance. We only consider the results obtained by MACHINA run without any topological constraints on  $G$ . We do not show results for Treeomics and neighbor joining as these methods do not yield a vertex labeling and migration graph. Supplementary Fig. 4B and Supplementary Fig. 5B show that MACHINA successfully determines the migration patterns of the mS instances. However, simulated instances with more complex migration patterns (e.g. pM and pR) can often be explained by simpler migration patterns. To further investigate this, we compare the inferred migration graph  $G$  to the simulated migration graph  $G^*$  by computing multi-edge recall and precision as follows:

$$\text{recall}(G, G^*) = \frac{|E(G) \cap E(G^*)|}{|E(G^*)|} \quad \text{and} \quad \text{precision}(G, G^*) = \frac{|E(G) \cap E(G^*)|}{|E(G)|}.$$

As  $E(G)$  and  $E(G^*)$  are multi-sets, we take the multiplicity of each multi-edge into account when computing recall and precision; e.g. a recall of 100% means that each multi-edge of  $G^*$  composed of  $c$  edges corresponds to a multi-edge in  $G$  composed of at least  $c$  edges between the same anatomical sites. We use the  $F_1$  score as a summary statistic, which is the harmonic mean between recall and precision, i.e.

$$F_1(G, G^*) = 2 \cdot \frac{\text{precision}(G, G^*) \cdot \text{recall}(G, G^*)}{\text{precision}(G, G^*) + \text{recall}(G, G^*)}.$$

We find that the precision and recall of the migration graph identified by MACHINA decreases with increasing complexity of the simulated migration pattern (Supplementary Fig. 4C and Supplementary Fig. 5C).

To identify drivers of metastasis, the set of migrating clones must be determined accurately. We assess this by comparing the set  $U(T, \ell)$  of migrating clones in the inferred clone tree  $T$  and vertex labeling  $\ell$  to the set  $U(T, \ell^*)$  of migrating clones in the simulated clone tree  $T^*$  and the simulated vertex labeling  $\ell^*$ . More specifically, the set  $U(T, \ell)$  is composed of the mutations present in the vertices that are incident to migration edges, i.e.  $U(T, \ell) = \{\bar{\sigma}(u) \mid (u, v) \in E(T), \ell(u) \neq \ell(v)\}$  where  $\bar{\sigma}(u)$  denotes the set of mutations present in vertex  $u$ . We compute recall and precision of the migrating clones as:

$$\text{recall}(U(T, \ell), U(T^*, \ell^*)) = \frac{|U(T, \ell) \cap U(T^*, \ell^*)|}{|U(T^*, \ell^*)|} \quad \text{and} \quad \text{precision}(U(T, \ell), U(T^*, \ell^*)) = \frac{|U(T, \ell) \cap U(T^*, \ell^*)|}{|U(T, \ell)|}.$$

Again, we consider the  $F_1$  score defined as

$$F_1(U(T, \ell), U(T^*, \ell^*)) = 2 \cdot \frac{\text{precision}(U(T, \ell), U(T^*, \ell^*)) \cdot \text{recall}(U(T, \ell), U(T^*, \ell^*))}{\text{precision}(U(T, \ell), U(T^*, \ell^*)) + \text{recall}(U(T, \ell), U(T^*, \ell^*))}. \quad (4)$$

Supplementary Fig. 4D and Supplementary Fig. 5D show that MACHINA identifies the clones that migrate and seed metastases with high precision and recall across all simulated migration patterns.

**Minimum-Migration Vertex Labelings** McPherson et al. [28] use the Sankoff algorithm [45] to find a vertex labeling with minimum migration number  $\mu$  given a clone tree. We enumerate all minimum-migration vertex labelings given the simulated clone tree  $T^*$  for each instance. Supplementary Fig. 4E and Supplementary Fig. 5E show that for each simulated clone tree  $T^*$  many such vertex labelings exist and that the migration number  $\mu$  does not distinguish

between different migration patterns of varying complexity. For instance, the simulated clone tree with seed 2 and simulated pattern pS (Supplementary Fig. 5E) has 800 minimum-migration vertex labelings, the majority of which (60%) correspond to more complex migration patterns. These findings illustrate the importance of using more sophisticated score functions that distinguish between migration patterns of varying complexity.

**Mutation Clustering** In the above results, we used the clustering algorithm described in Section A.1.2. In the following, we explore the performance of MACHINA when the output of different mutation clustering algorithms – including PhyloWGS [7], PyClone [41], Clomial [53] and SciClone [30] – are input to MACHINA. We note that PhyloWGS simultaneously clusters mutations and arranges the resulting clusters into a tree; in the following comparisons we discard the tree and only use the inferred mutation clusters; we refer the reader to Figs. 4 and 5 for PhyloWGS’s clone tree inference performance results. We run PhyloWGS with default arguments. We use the binomial model in PyClone and specify a default sequencing error rate of 0.001, a purity of 1 and use 10,000 MCMC iterations with a burnin of 1,000. We use the default maximum number of 10 clusters in Clomial. For SciClone, we use the default beta mixture model and the default number of at most 10 mutation clusters. Using larger numbers of mutation clusters in both Clomial and SciClone leads to prohibitive running times.

We assess the performance of each clustering method on the  $\Sigma_{\max} = 5$  instances. Supplementary Fig. 16A shows that the width of the 95% confidence intervals of the mutation cluster frequencies is largely unaffected by the clustering algorithm. Supplementary Fig. 16B shows that PyClone, Clomial and SciClone infer far fewer clusters than MACHINA and PhyloWGS. We now assess whether this affects the accuracy of the inferred clusterings. A clustering is a partition of the set of  $n$  mutations. Let  $\mathcal{C}$  denote the inferred clustering and let  $\mathcal{C}^*$  denote the simulated clustering. To assess the similarity of  $\mathcal{C}$  to  $\mathcal{C}^*$  typically all  $\binom{n}{2}$  pairs of mutations are considered. For each pair  $(a, b)$  of distinct mutations the following four cases are distinguished.

1. True positive (TP): if  $(a, b)$  co-occur in a cluster in  $\mathcal{C}$  and  $\mathcal{C}^*$ .
2. False positive (FP): if  $(a, b)$  co-occur in a cluster in  $\mathcal{C}$  but do not co-occur in  $\mathcal{C}^*$ .
3. True negative (TN): if  $(a, b)$  do not co-occur in  $\mathcal{C}$  nor in  $\mathcal{C}^*$ .
4. False negative (FN): if  $(a, b)$  do not co-occur in  $\mathcal{C}$  but do co-occur in  $\mathcal{C}^*$ .

The commonly-used Rand index [36] is defined as the accuracy, i.e.  $(TP + TN) / (TP + FP + TN + FN)$ . We find for each clustering method and migration pattern that the number of false negatives is around 0 (data not shown). In other words, the recall of each clustering method is around 1. Therefore, instead of using the Rand index, we compute the precision, defined as  $TP / (TP + FP)$ . Supplementary Fig. 16C shows that MACHINA and PhyloWGS achieve higher clustering precision than PyClone, Clomial and SciClone, which is likely due to the larger number of inferred clusters by the former methods.

Next, we assess the impact of using different clustering algorithms as input to MACHINA. Supplementary Fig. 17A shows that the clone trees inferred by MACHINA are more similar to the simulated clone trees than those inferred

by neighbor joining [42], Treeomics [38], PhyloWGS [7] and AncestryTree [8]. We note, however, that MACHINA’s performance decreases with the precision of the clustering algorithm used to generate input clusters for MACHINA. Indeed, Supplementary Fig. 17B shows that the recall and precision of the migration clones also decreases when MACHINA is used together with clustering algorithms that achieved lower precision on these simulated instances. Supplementary Fig. 17C shows that despite the differences in the inference of the migrating clones, the inferred migration graphs are robust to the choice of clustering algorithm, with very minor variation in  $F_1$  scores.

**Downsampling Sequencing Coverage, Sample Purity and Number of Sequenced Samples** To evaluate the robustness of MACHINA and the used mutation clustering algorithm, we perform downsampling experiments, where we reduce the number of sequenced samples, the sample purity and the sequencing coverage. To do so, we simulate 10 metastatic tumors with  $\Sigma_{\max} = 5$  anatomical sites for each of the four migration patterns (mS, pS, pM and pR). For each simulated tumor, we generate a baseline instance containing three bulk samples per anatomical site with a mean sequencing coverage of 10,000x, a nucleotide sequencing error rate of 0.001 and a sample purity of 1. We then retain for each anatomical site  $\{1, 2, 3\}$  bulk samples. Next, for each retained sample, we downsample the sequencing depth of each SNV to  $\{200x, 500x, 1000x, 10000x\}$  using different purities  $\{0.5, 0.8, 1.0\}$ . Thus, for each simulated tumor we have  $3 \cdot 4 \cdot 3 = 36$  sets of sequencing reads, corresponding to all combinations of number of samples, coverage and purity.

We run MACHINA’s clustering algorithm on each instance and compute 95% confidence intervals (as described previously). Supplementary Fig. 18A shows that the width of the 95% confidence intervals of the mutation clusters frequencies decreases with increasing coverage and number of samples. Concordantly, the clustering precision increases with increasing coverage and number of samples (Supplementary Fig. 18B). By comparing each downsampled instance to the baseline, we find that MACHINA benefits from higher resolution sequencing data, with the number of samples having the largest impact, followed by the depth of sequencing and the sample purity (Supplementary Fig. 18C-E).

**Downsampling Number of SNVs** We now evaluate the effect of the number of sequenced SNVs on the performance of MACHINA. Recall that previously we used a mutation rate of 0.1, corresponding to whole exome sequencing data (Section A.1.1). We increase this rate to 10 mutations every cell division, corresponding to whole genome sequencing data [29]. For each migration pattern (mS, pS, pM and pR), we simulate 10 metastatic tumors with  $\Sigma_{\max} = 8$  anatomical sites, amounting to a total of  $4 \cdot 10 = 40$  instances. We generate two bulk samples of the primary tumor and a single bulk sample for each metastasis; each with a purity of 0.8. We sequence each sample at an average depth of 200x and an error rate of 0.001. Subsequently, we downsample the number of SNVs to  $\{100\%, 50\%, 10\%, 5\%, 1\%, 0.5\%, 0.1\%\}$  of all SNVs. We use the MACHINA clustering algorithm to group SNVs with similar variant allele frequencies across all samples, and we run MACHINA in PMH-TI mode to obtain clone trees and migration histories.

We find that the performance of MACHINA is affected by the number of sequenced SNVs. Importantly, MACHINA achieves good performance in the regime of whole exome data (with only 1 to 5% of all SNVs). More specifically,

we find that the number of inferred mutation clusters increases with the number of SNVs (Supplementary Fig. 19B). Similarly, we find that with more SNVs the uncertainty in mutation clusters decreases, as shown by the mean width of the 95% confidence intervals of frequencies of mutation clusters (Supplementary Fig. 19C). The clustering precision and recall, as defined above, are affected by the number of SNVs, with the precision decreasing and recall increasing with increasing numbers of SNVs (Supplementary Fig. 19D-E). On the other hand, the clone tree distance  $d(T^*, T)$  between the simulated tree  $T^*$  (containing all SNVs) and each inferred tree  $T$  (containing a subset of the SNVs) decreases with increasing number of SNVs (Supplementary Fig. 19F). Finally, the precision and recall (summarized by the  $F_1$  score) of the clones that migrate to different anatomical sites (Supplementary Fig. 19G), and of the migration graph (Supplementary Fig. 19H) increases with increasing numbers of SNVs. Note that in the computation of the  $F_1$  score of the migrating clones, we define a clone by the subset of SNVs that were present in the downsampled instance. That is, in Equation (4) we define  $U(T^*, \ell^*) = \{\bar{\sigma}(u) \cap X \mid (u, v) \in E(T^*), \ell^*(u) \neq \ell^*(v)\}$  where  $\bar{\sigma}(u)$  is the set of mutations present in vertex  $u$  of  $T^*$  and  $X$  is the set of mutations present in the downsampled instance.

## A.2 Metastatic Colorectal Cancer

Most of the published studies [3–5, 12–14, 20, 23–25, 27, 47, 49–51, 54, 55] derive clone trees using standard phylogenetic techniques based on neighbor-joining [42], maximum parsimony [11] or maximum likelihood [10]. The implicit assumption that these studies make is that the sequenced anatomical sites are homogeneous, i.e. composed of a single clone. Here, we show that the sample homogeneity assumption can lead to clone trees and consequently migration histories that are likely incorrect. In particular, we show that the resulting clone trees have multiple cases of *homoplasy*, where identical single-nucleotide mutations occur independently on different branches of the tree. While homoplasy cannot definitively be ruled out in cancer evolution, it is highly implausible to see multiple (even dozens) of such events. Indeed, no homoplasy, or the infinite sites assumption, is the standard assumption for single-nucleotide mutations in cancer [17, 26, 32, 48].

A recent commentary [1] discusses the issues of the sample homogeneity assumption on heterogeneous tumor sequencing data. We illustrate how the issue of extensive homoplasy arises in a study of five patients with metastatic colorectal cancers [23]. In this study, the authors sequenced for each patient 2–6 metastases and 2–5 regions from the primary tumor, and used maximum parsimony to infer phylogenetic trees, assuming that each region consists of a single clone. We find that the published trees exhibit extensive homoplasy, with many violations of the infinite sites assumption (Supplementary Table 4). The most extreme example is patient CRC2, which has 412 identified SNVs, 41 of which violate the infinite sites assumption (Supplementary Fig. 1A). An evolutionary history where  $\approx 10\%$  of SNVs occur independently multiple times is highly unlikely, particularly because the majority of these mutations are not known cancer-causing mutations that may have undergone positive selection. Moreover, the violations cluster on the tree: 24 of the violations occur on the branch leading to region P3 from the primary tumor, with 18 of these mutations also occurring on a single branch leading to region P1 from the primary tumor and metastases M1 and M2. When comparing the distribution of the variant allele frequencies (VAFs) of these 18 mutations to the VAF distribution of the other mutations, we find a clear indication that P3 is not homogeneous and contains a subclone composed of

the 18 homoplasmy mutations (Supplementary Fig. 1B). This example clearly demonstrates the differences between the clone tree and the migration graph and shows that ignoring intra-tumor heterogeneity results in a tree that is neither representative of the evolutionary relationships between clones/cells nor representative of the migrations of cells between anatomical sites.

### A.3 Metastatic Ovarian Cancer

We run the Sankoff enumeration algorithm (Supplementary Methods) on the reported clone trees of all metastatic ovarian cancers in [28]. In all cases, the reported vertex labeling is among the enumerated vertex labelings (Supplementary Table 5). For patients 1, 3 and 7, we find multiple vertex labelings with the same minimum migration number  $\mu^*$  but different comigration number  $\gamma$ . We also consider the effect of resolving polytomies (Supplementary Table 6) and find for most patients that this results in fewer migrations than reported. We describe these results in more detail in the following.

None of the enumerated vertex labelings of patient 1 have comigration number  $\gamma_{\min} = m - 1 = 6$  and consequently do not correspond to a single-source seeding (S) pattern. When enforcing an S pattern using the ILP (Supplementary Methods), we find that with LOv as the primary,  $\mu = 15$  migrations are required, and that with an ROv primary,  $\mu = 14$  are required. This suggests that ROv is a more likely primary under a single-source seeding constraint. However, by resolving the polytomies in this patient’s clone tree under an S pattern, we find that in both cases only  $\mu = 12$  migrations are needed (Supplementary Fig. 6C and D). When allowing for reseeded (R) in the polytomy resolution, we find that the most parsimonious solution is obtained with LOv as the primary with migration number  $\mu = 11$  but with comigration number  $\gamma = 7$  and reseeded between LOv and ROv (Supplementary Fig. 6E).

For patient 3 with  $m = 8$  anatomical sites, the minimum migration number  $\mu^* = 27$  is achieved with both LOv and ROv as the primary, with 4 and 20 vertex labelings, respectively (Supplementary Table 5). The reported vertex labeling has comigration number  $\gamma = 9$  with LOv as the primary [28] (Supplementary Fig. 7B). However, there are two vertex labelings where the primary is either LOv or ROv, that both achieve the minimum comigration number  $\gamma_{\min} = m - 1 = 7$  (Supplementary Fig. 7C and D). These labelings correspond to a polyclonal parallel single-source seeding (pPS) pattern, where each metastasis is seeded only once from the primary by multiple comigrating clones. Since McPherson et al. [28] do not reconstruct the migration pattern nor analyze the number of comigrations, they overlook these more parsimonious explanations. When further exploring binarizations, we arrive at fewer migrations  $\mu = 25$  with both LOv and ROv as the primary (Supplementary Fig. 7E and F).

For patient 7, the authors report a vertex labeling with migration number  $\mu = 11$ . This labeling, however, has a non-ovarian anatomical site (RUt, right uterosacral ligament) as the primary, which might be unlikely for ovarian cancer (Supplementary Fig. 8B). We find that the reported labeling is the only minimum migration labeling with RUt as the primary (Supplementary Table 5). When using one of the ovaries (LOv or ROv) as the primary, we find for LOv four labelings with migration number  $\mu = 12$  and for ROv 32 labelings with migration number  $\mu = 13$ . Among these, there are multiple labelings that achieve the minimum comigration number  $\gamma_{\min} = 6$  (e.g., Supplementary Fig. 8C and D). Thus with either of the ovaries as the primary more migrations are required than with an RUt primary.

The clone tree, however, has four polytomies, whose resolution leads to a migration history with migration number  $\mu = 11$  and comigration number  $\gamma_{\min} = 6$  for either LOv and ROv as the primary Supplementary Fig. 8E and F). Moreover, solving the PHM-TR problem with RUt as the primary does not result in fewer than  $\mu = 11$  migrations (Supplementary Fig. 8G). These findings suggest that analysis of the migration history provides no evidence for the authors’ statement that the primary tumor is located in the right uterosacral ligament (RUt) as opposed to either of the ovaries (LOv or ROv).

For the remaining patients, we find that polytomy resolution allows for the identification of more parsimonious migration patterns than reported. For instance, for patient 2 there exists a binarization and subsequent labeling with a monoclonal single-source seeding (mS) pattern. Moreover, for patients 4 and 9 we can no longer distinguish between LOv and ROv as the primary when resolving polytomies.

Interestingly, in our analyses of the ovarian cancer dataset, where metastasis typically proceeds via the intraperitoneal cavity lacking physical barriers, we find large numbers of comigrating clones, while our analyses of the breast, prostate and melanoma datasets (in the next sections), where metastasis follows the more common mechanisms of lymphogenous or hematogenous spread, have few numbers of comigrating clones. This suggests that the extent of comigration may differ depending on the mode of metastatic spread. However, more comprehensive analyses with larger sample numbers are required to support this hypothesis.

#### A.4 Metastatic Prostate Cancer

In a seminal paper, Gundem et al. [15] studied the evolutionary history of ten metastatic prostate cancers. The authors sequenced matched primary and metastasis samples using whole genome sequencing (WGS) technology. Among the ten patients, the authors concluded that five patients exhibited polyclonal seeding and eight patients exhibited metastasis-to-metastasis spread. In our nomenclature, a parallel single-source seeding (PS) pattern indicates the absence of metastasis-to-metastasis spread (Supplementary Table 9). As described in the main text, the presence of a primary tumor sample is essential for reconstructing the complete migration history. However, only five patients (A10, A22, A29, A31 and A32) included a sequencing sample from the primary prostate anatomical site (Supplementary Table 7). We analyze the migration history of these five patients with MACHINA.

Patient A10 has 9472 SNVs distributed over  $m = 4$  anatomical sites. Gundem et al. [15] identified a clone tree with nine mutation clusters and upon manual analysis hypothesized that each metastasis of this patient was seeded by a single clone and that metastasis-to-metastasis spread took place. MACHINA (in PHM-TR mode) resolves a single polytomy in the reported clone tree and identifies a vertex labeling with a monoclonal parallel single-source seeding pattern (mPS), migration number  $\mu = 3$ , comigration number  $\gamma = 3$  and seeding site number  $\sigma = 1$  (Supplementary Fig. 9A and Supplementary Fig. 10A). We thus recapitulate the authors’ finding of monoclonal seeding. However, in the migration history inferred by MACHINA each metastasis is seeded directly from the primary tumor, which is an alternative explanation that should be evaluated alongside the metastasis-to-metastasis spread hypothesis reported by the authors. Careful analysis of the vertex labeling inferred by MACHINA reveals that the ‘gold’ mutation cluster (containing 1218 mutations) is key in distinguishing the two alternative scenarios (Supplementary Fig. 10A). For

a parallel migration pattern this cluster must have originated in the primary prostate tumor, whereas in the case of metastasis-to-metastasis spread this cluster must have originated in either the periportal or the perigastric lymph node. Gundem et al. [15] reported a cancer cell fraction (CCF) of 0.4% in the prostate sample of the ‘gold’ mutation cluster. If this value is accurate then the clone composed of the mutations in the ‘grey’ and ‘gold’ clusters was present in only a small proportion in the prostate at the time of sequencing and must have seeded both the periportal and the perigastric lymph node, ruling out metastasis-to-metastasis spread. In case the observed CCF is due to sequencing artifacts and in reality is 0, the scenario of parallel seeding of the two lymph nodes would imply that the clone containing both the ‘gold’ and ‘grey’ mutation clusters must have become extinct in the prostate, whereas the existence of such an extinct clone is not required in the case of metastasis-to-metastasis spread. These considerations must be taken into account before drawing conclusions regarding the mode of seeding. Thus, Gundem et al.’s finding of metastasis-to-metastasis spread for this patient is not conclusive.

Patient A29 has 8275 SNVs in  $m = 2$  anatomical sites. MACHINA finds a vertex labeling with migration number  $\mu = 1$ , comigration number  $\gamma = 1$  and seeding site number  $\sigma = 1$  (Supplementary Fig. 9B), thus confirming the authors’ finding of monoclonal parallel single-source seeding (mPS).

Patient A31 has 4852 SNVs in  $m = 5$  anatomical sites. Gundem et al. [15] identified a clone tree with ten mutation clusters and hypothesized that polyclonal seeding and metastasis-to-metastasis spread have taken place for this patient. MACHINA (in PHM-TR mode) resolves three polytomies in the reported clone tree and finds that the most parsimonious vertex labeling has migration number  $\mu = 10$ , comigration number  $\gamma = 4$  and seeding site number  $\sigma = 2$ , corresponding to a polyclonal single-source seeding (pS) pattern (Supplementary Fig. 9C and Supplementary Fig. 10B). Although a pPS pattern requires two more migrations ( $\mu = 12$ ) for this patient (Supplementary Fig. 9D), the presence of the ‘dark green’ and ‘light blue’ mutation clusters in the prostate—with CCFs of 6.4% and 1.2%, respectively—indicate that parallel seeding of all metastases from the primary is a likelier explanation than metastasis-to-metastasis spread.

Patient A32 has 9388 SNVs in  $m = 6$  anatomical sites. The authors identified a clone tree with 12 mutation clusters, and upon manual analysis hypothesize that several metastases of this patient were seeded polyclonally via a metastatic cascade. Indeed, we find that the most parsimonious vertex labeling has migration number  $\mu = 7$ , comigration number  $\gamma = 5$  and seeding site number  $\sigma = 2$  and corresponds to a polyclonal multi-source seeding (pM) migration pattern with metastasis-to-metastasis spread (Supplementary Fig. 9E). Under a parallel single-source seeding (PS) constraint (defined in Section B.2) we find a polyclonal parallel single-source seeding (pPS) migration history with a single additional migration ( $\mu = 8$ ), comigration number  $\gamma = 4$  and seeding site number  $\sigma = 1$  (Supplementary Fig. 9F and Supplementary Fig. 10C). This alternative migration history with only one additional migration must be taken into account when deciding whether metastasis-to-metastasis spread took place, especially given that minimum-migration labeling follows a complex polyclonal multi-source seeding pattern. Similarly to patient A10, there is a single mutation cluster whose presence in the primary prostate tumor would allow one to definitively rule out metastasis-to-metastasis spread (Supplementary Fig. 10C). This is the ‘light blue’ cluster; Gundem et al. [15] report a CCF of 0.3% in the prostate for this cluster. Thus, targeted sequencing of the mutations in the ‘light blue’ cluster

would be highly informative. Note that even if targeted sequencing fails to establish the presence of this clone in the primary tumor, parallel seeding might still be an explanation for this patient. In this case, the reason for not observing the clone could be either extinction of the clone or insufficient sequencing resolution.

Finally, patient A22 has 10262 SNVs in  $m = 10$  anatomical sites. Using the clone tree reported by the authors, MACHINA finds that the most parsimonious vertex labeling for this patient has migration number  $\mu = 26$ , comigration number  $\gamma = 12$  and seeding site number  $\sigma = 5$ , corresponding to a polyclonal reseeding (pR) pattern (Supplementary Table 7). Enforcing a parallel single-source seeding (PS) migration pattern results in 10 more migrations. Contingent on the accuracy of the reported clone tree, this finding indicates that metastasis-to-metastasis may have likely taken place for this patient.

In summary, while MACHINA’s results show complete agreement with Gundem et al. [15] in concluding polyclonal seeding, we find that for three patients (A10, A31 and A32) a scenario of parallel seeding of all metastases from the primary tumor is also consistent with the data.

## A.5 Metastatic Melanoma

To study whether migration in melanoma follows a *serial* progression from primary tumor to regional metastases to distant metastases, Sanborn et al. [44] performed whole exome sequencing on matched primary and metastases from eight patients with metastatic melanoma. More specifically, the authors aimed to find evidence for the seeding of multiple distinct anatomical sites from the primary tumor, likely ruling out a serial progression. In addition, the authors aimed to detect polyclonal seeding of anatomical sites by identifying mutations that are subclonal in distinct metastatic sites. Two patients were found to have polyclonal-seeded metastases and non-serial progression was concluded for six patients. We use MACHINA to analyze the migration history of the six patients reported to have a non-serial progression (Supplementary Table 8).

We only consider SNVs that occur in copy-neutral regions and cluster these using the AncesTree clustering algorithm [8], described in Section A.1.2. We exclude patient H from our analysis due to the absence of copy-neutral SNVs. By solving the PMH-TI problem, MACHINA finds that only patients C and E have undergone polyclonal seeding, in agreement with [44]. MACHINA infers a parallel single-source seeding (PS) pattern for patients A, C, D and F, and infers a migration history with  $\sigma = 2$  seeding sites and migration number  $\mu = 5$  for patient E. These findings are in line with the conclusion of the likely absence of serial progression by Sanborn et al. [44].

## A.6 Metastatic Breast Cancer

Hoadley et al. [19] restricted their analyses to single-nucleotide variants (SNVs) that occur in copy-neutral regions. The authors used SciClone [30] to cluster mutations with the same variant allele frequency (VAF) across all anatomical sites. We correct the reported cluster assignments of mutations 47 and 81, as described in Supplementary Table 1. We use the reported mutation clusters to infer confidence intervals on the frequency of cells containing the mutations present in a cluster. We do this by viewing variant reads as draws from a binomial distribution similarly to SciClone.



More specifically, let  $C$  be a mutation cluster and let  $v_{s,i}$  and  $r_{s,i}$  denote, respectively, the number of variant and reference reads of SNV  $i \in C$  in anatomical site  $s$ . Assuming a uniform prior on the true frequency of cluster  $C$  leads to a beta posterior distribution given the observed variant and reference read counts of the mutations present in cluster  $C$ , i.e.  $\text{Beta}\left(1 + \sum_{i \in C} v_{s,i}, 1 + \sum_{i \in C} r_{s,i}\right)$ . For each mutation cluster  $C$  and anatomical site  $p$ , we use the Bonferroni correction to infer confidence intervals  $[\hat{f}_{p,C}^-, \hat{f}_{p,C}^+]$  on the posterior distribution such that the family-wise type-I error rate is  $1 - \alpha = 5\%$ . That is, given  $m = 6$  samples,  $n = 10$  mutations clusters and  $\alpha = 0.95$ , we use a confidence of

$$1 - \frac{1 - \alpha}{mn} \approx 99.92\%.$$

Multiplying each interval  $[\hat{f}_{p,C}^-, \hat{f}_{p,C}^+]$  by 2 yields frequency matrices  $(F^-, F^+)$ .

**Patient A7** Patient A7 has  $m = 6$  anatomical sites, including the breast primary and five metastases from the brain, kidney, liver, lung and rib. The authors identified 478 SNVs that occur in copy-neutral genomic regions. These SNVs were clustered into 10 clusters using SciClone [30] (Supplementary Fig. 2A). SPRUCE finds two different mutation trees that differ in their ordering of mutation clusters 3 and 5 (Supplementary Fig. 2B). One of the mutation trees, denoted by  $\bar{T}$ , corresponds to the mutation tree reported by the authors. The authors conflated mutation clusters and clones, and erroneously inferred a clone tree from  $\bar{T}$  by assigning an anatomical site to each mutation cluster if the corresponding VAF was greater than 0. This yielded a clone tree with 22 extant clones (Supplementary Fig. 2C). By manually assigning anatomical sites to internal vertices and resolving polytomies, Hoadley et al. [19] inferred two distinct migration histories: (i) a polyclonal multi-source seeding (pM) history with migration number  $\mu = 12$  and comigration number  $\gamma = 6$  (Main Text), and (ii) a polyclonal single-source seeding (pS) with  $\mu = 15$  and  $\gamma = 5$  (Supplementary Fig. 2D). By running MACHINA to solve the PMH-TI problem given  $(F^-, F^+)$  and either of the mutation trees, we arrive at parsimonious monoclonal single-source seeding (mS) histories with  $\mu_{\min} = 5$  migrations and  $\gamma_{\min} = 5$  comigrations (Supplementary Fig. 2E). Note that for both mutation trees there is ambiguity in the order of migrations between between kidney and liver, relative to the lung metastasis (2F).

We ran Treeomics, with and without subclone detection enabled, given the variant allele frequencies of patient A7. In both cases, Treeomics infers a phylogenetic tree that contains only a single clone for each anatomical site (Supplementary Fig. 2G). This apparent homogeneity of the metastases might be an artifact of the discretization of the variant allele frequencies in the Treeomics algorithm. As the migration history of this patient most likely follows an mS pattern, the clones in each metastatic site will correspond to a single clade. Therefore, the discretization step will not result in violations of the infinite sites assumption and subsequently Treeomics will not invoke the subclone detection heuristic (which, in fact, is a variation of the split row operation described in [18]). We observed the same phenomenon in all simulated mS instances, where no subclones were detected in the metastases by Treeomics.

We provide the method by McPherson et al. [28] the clone tree that MACHINA inferred from the reported mutation tree (leftmost tree in Supplementary Fig. 2B). This clone tree has two polytomies (Supplementary Fig. 2E). Since the McPherson et al. method [28] only considers the migration number  $\mu$  and does not resolve polytomies, it is unable to infer the mS migration history with migration number  $\mu = 5$  identified by MACHINA (Supplementary

Fig. 2F). Instead, it infers monoclonal multi-source seeding (mM) migration patterns with one additional migration (Supplementary Fig. 12).

**Patient A1** This patient has  $m = 5$  anatomical sites, including the primary breast tumor and four metastases from the lung, spinal, adrenal glands and liver. The authors report 329 copy-neutral SNVs, distributed over 9 mutation clusters (Supplementary Fig. 13A). SPRUCE [9] finds four different mutation trees that differ in the ordering of mutation clusters 4, 6 and 7 (Supplementary Fig. 13B). Two of the four mutation trees correspond to mutation trees inferred by the authors using ClonEvol [6]. Hoadley et al. [19] infer a polyclonal parallel single-source seeding (pPS) migration history with migration number  $\mu = 13$  and comigration number  $\gamma_{\min} = 4$  (Supplementary Fig. 13C). By jointly analyzing the cell division, mutation and migration history we infer, for each of the four mutation trees, clone trees that admit migration histories with migration number  $\mu = 6$  and comigration number  $\gamma_{\min} = 4$  (Supplementary Fig. 13D). These results provide a more parsimonious explanation of the history of this metastatic cancer than previously reported in [19].

Treomics identifies a clone tree that misses two subclones in the breast and the adrenal samples (Supplementary Fig. 14A). This is a likely consequence of the VAF discretization step in Treomics. In the breast sample, the discretization incorrectly removes a cluster of 48 mutations; the presence of this cluster is supported by the number of variant read of the comprising mutations (Supplementary Fig. 14B). On the other hand, in the adrenal sample, there are two mutation clusters that have distinct VAFs, one of which with the aforementioned 48 mutations; this information is lost upon discretization (Supplementary Fig. 14B). Consequently, the minimum migration history of the clone tree inferred by Treomics misses the polyclonal seeding of the adrenal metastasis (Supplementary Fig. 14C). The method employed by McPherson et al. [28] identifies the same migration graph when given the clone tree inferred by MACHINA.

## B Methods

In Section B.1 we define the various mathematical concepts that we use for analyzing the migration history of a metastatic cancer. Next, we introduce in Section B.2 the three problem statements for (1) inferring a migration history given a clone tree (PMH), (2) refining polytomies in a given clone tree using the migration history (PHM-TR), and (3) inferring the clone tree and migration history jointly given mutation frequencies from bulk sequencing data (PMH-TI). Finally, in Section B.3 we describe MACHINA (Metastatic And Clonal History INtegrative Analysis), an algorithm that solves the three problems and provides a framework for comprehensively analyzing the cell division, mutation and migration history of metastatic cancers.

### B.1 Preliminaries

We start by introducing notation that we use throughout the manuscript. Let  $T$  be a rooted tree with vertex set  $V(T)$  and edge set  $E(T)$ . We further split the vertex set  $V(T)$  into a leaf set  $L(T)$  and internal vertex set  $I(T)$ . Vertex  $r(T)$

denotes the root of  $T$ . We denote the children of a vertex  $u$  by  $\delta_T(u)$ . We write  $u \preceq_T v$  if and only if vertex  $v$  is reachable from vertex  $u$ , and we write  $u \prec_T v$  if and only if  $u \neq v$  and  $v$  is reachable from  $u$ . We denote by  $T_v$  the subtree of  $T$  rooted at vertex  $v$ .

We consider a metastatic cancer composed of  $n$  mutations that are present in clones from  $m$  distinct anatomical sites denoted by  $\Sigma$ . The *primary tumor*, denoted by  $P \in \Sigma$ , is the origin of the tumor. The mutation tree  $\bar{T}$  describes the mutation history of the metastatic cancer and is formally defined as follows.

**Definition 1.** A mutation tree  $\bar{T}$  is an edge-labeled rooted tree whose  $n$  edges are labeled uniquely by mutations  $[n]$ .

Requiring each mutation to label at exactly one edge of the mutation tree is known in population genetics as the *infinite sites assumption*, or in phylogenetics as the *perfect phylogeny condition* [16]. This assumption states that a mutation only occurs once throughout the entire history of a tumor and is never lost. Since the root vertex  $r(\bar{T})$  corresponds to a normal cell, we have that  $r(\bar{T})$  has state 0 for every character/mutation  $i \in [n]$ . According to the clonal theory of cancer [34], we have that all tumor cells are descendant from the same tumor cell, i.e. we require  $|\delta(r(\bar{T}))| = 1$ . Each vertex  $v_i \in V(\bar{T}) \setminus \{r(\bar{T})\}$  corresponds to a unique mutation  $i \in [n]$  that labels the incoming edge of  $v_i$ . As such, each vertex  $v_i \in V(\bar{T})$  defines a clone that is composed of the mutations that label the edges of the unique path from  $v_i$  to the root  $r(\bar{T})$ . A mutation tree  $\bar{T}$  does not assign clones to anatomical sites. To that end, we define a clone tree  $T$  as follows.

**Definition 2.** A clone tree  $T$  is a rooted tree whose leaves are labeled by anatomical sites  $\Sigma$  via the function  $\ell : L(T) \rightarrow \Sigma$  such that for each anatomical site  $s$  there exists a leaf  $u$  labeled by  $\ell(u) = s$ . The edges of  $T$  are labeled via the function  $\lambda : E(T) \rightarrow [n] \cup \{\perp\}$  such that for each mutation  $i \in [n]$  there is exactly one edge  $(u, v)$  where  $\lambda(u, v) = i$ . Edges labeled by  $\perp$  do not introduce a new mutation.

A clone tree  $T$  describes the cell division and mutation history of the clones currently present in anatomical sites  $\Sigma$ . One infers a clone tree  $T$  using specialized phylogenetic inference techniques from tumor sequencing data [7–9, 21, 22, 26, 33, 35, 40, 48, 52]. Note that multiple regions from a single anatomical site may be sequenced, and doing so may improve the accuracy of the inferred clone tree. The leaves  $L(T)$  of the clone tree  $T$  are the extant clones of the tumor and each leaf  $u \in L(T)$  is labeled by the anatomical site  $\ell(u)$  in which it occurs. The internal vertices  $V(T) \setminus L(T)$  are the ancestral clones. Each directed edge  $(u, v) \in E(T)$  is labeled by the somatic mutation  $\lambda(u, v)$  that distinguish clone  $v$  from clone  $u$ . Note that each clone tree  $T$  corresponds to a unique mutation tree  $\bar{T}$  obtained by condensing all unlabeled edges of  $T$ . We will elaborate on this in Section B.2.2. Unless explicitly mentioned, we omit the edge labeling of  $T$  in the following.

While the clone tree describes the lineage of the clones according to cell divisions and mutations, it does not directly describe the migration history of the clones of a tumor. To describe the migration history, we need to assign each internal vertex  $v$  an anatomical site  $\ell(v) \in \Sigma$ .

**Definition 3.** Let  $T$  be a clone tree on anatomical sites  $\Sigma$  with leaf labeling  $\hat{\ell}$ . The vertex labeling  $\ell : V(T) \rightarrow \Sigma$  extends  $\hat{\ell}$  to all vertices of  $T$  such that  $\ell(r(T)) = P$  and  $\ell(u) = \hat{\ell}(u)$  for all leaves  $u$ .

We call a clone-tree edge  $(u, v)$  a *migration edge* provided  $\ell(u) \neq \ell(v)$ . We model the migration patterns, i.e. the origin and destination anatomical site of each migration, by a directed multigraph called the *migration graph*, which is formally defined as follows.

**Definition 4.** A migration graph  $G$  on anatomical sites  $\Sigma$  is a connected, vertex-labeled directed multigraph whose  $m = |\Sigma|$  vertices are in 1-1 correspondence with  $\Sigma$ .

We denote the edge multiset of  $G$  by  $E(G)$ . As such,  $E(G)$  may contain multiple edges  $(s, t)$ —all such occurrences combined form a multi-edge.

Migration patterns, as described by the migration graph, can be distinguished in two different ways. First, by the number of clones that migrate between two anatomical sites, i.e. only a single clone migrates in the case of *monoclonal seeding* (m), whereas multiple clones migrate in the case of *polyclonal seeding* (p). Second, by the topology of migrations: with *parallel single-source seeding* (PS) all migrating clones originate from the primary  $P$ , with *single-source seeding* (S) all the clones that migrate to an anatomical site originate from the same source anatomical site, whereas they have multiple origins with *multi-source seeding* (M), and with *reseeding* (R) clones migrate back and forth between anatomical sites. Supplementary Table 9 shows the different migration patterns.

Each vertex labeling  $\ell$  of a clone tree  $T$  determines the migration graph  $G$ . That is, each migration edge  $(u, v)$  in  $T$ , with  $\ell(u) \neq \ell(v)$ , corresponds to an edge  $(\ell(u), \ell(v))$  in  $G$ . There exist many different vertex labelings  $\ell$  of a clone tree  $T$ , yielding different migration graphs (Supplementary Fig. 20), which we distinguish by the migration number  $\mu(T, \ell)$ , the comigration number  $\gamma(T, \ell)$  and the seeding site number  $\sigma(T, \ell)$ , defined in the following sections.

### B.1.1 Migration Number $\mu(T, \ell)$

Given a clone tree  $T$  and vertex labeling  $\ell$ , we define the *migration number*  $\mu(T, \ell)$  as the number of migration edges, i.e.

$$\mu(T, \ell) = |\{(u, v) \in E(T) \mid \ell(u) \neq \ell(v)\}|. \quad (5)$$

Equivalently, the migration number  $\mu(T, \ell)$  is the number  $|E(G)|$  of edges of  $G$ . Since each of the  $m - 1$  metastases must be seeded, we make the following observation.

**Observation 1.** The migration number  $\mu(T, \ell)$  is at least  $\mu_{\min} = m - 1$ .

We note that the case of  $\mu_{\min} = m - 1$  migrations corresponds to *monoclonal single-source seeding* (mS), where every metastasis is seeded by a single migrating clone. Not every clone tree admits a vertex labeling with migration number  $\mu_{\min}$  (Supplementary Fig. 20). The following proposition presents a necessary and sufficient condition for the existence of a labeling with migration number  $\mu_{\min}$ .

**Proposition 1.** Let  $\nu(s)$  be the lowest common ancestor of all leaves labeled by anatomical site  $s \neq P$ , and let  $\nu(P)$  be  $r(T)$ . Then, a clone tree  $T$  with leaf labeling  $\bar{\ell}$  has a vertex labeling  $\ell^*$  with migration number  $\mu(T, \ell) = \mu_{\min}$  if and only if for all distinct anatomical sites  $s, t$  and clone-tree leaves  $u, v$  where  $\bar{\ell}(u) = s$  and  $\bar{\ell}(v) = t$  the paths from  $\nu(s)$  to  $u$  and from  $\nu(t)$  to  $v$  are vertex disjoint.

*Proof.* ( $\Rightarrow$ ) Let  $\ell^*$  be a vertex labeling of  $T$  with migration number  $\mu(T, \ell^*) = \mu_{\min}$ . Consider two distinct clone-tree leaves  $u, v$  such that  $\ell^*(u) = s$ ,  $\ell^*(v) = t$  and  $s \neq t$ . Assume for a contradiction that the paths from  $\nu(s)$  to  $u$  and from  $\nu(t)$  to  $v$  overlap at clone-tree vertex  $w$ . We claim that  $\ell^*(\nu(s)) = s$  and  $\ell^*(\nu(t)) = t$ , as otherwise  $\mu(T, \ell^*) > \mu_{\min}$ . Thus, it holds that  $\ell^*(w) = s$  and  $\ell^*(w) = t$ . However,  $s \neq t$ , which is a contradiction. Hence, the paths from  $\nu(s)$  to  $u$  and from  $\nu(t)$  to  $v$  are vertex disjoint.

( $\Leftarrow$ ) We show constructively how to obtain a valid clone-tree vertex labeling  $\ell^*$  from  $\nu$  given that for all distinct anatomical sites  $s, t \in \Sigma$  and clone-tree leaves  $u, v \in L(T)$  such that  $\bar{\ell}(u) = s$  and  $\bar{\ell}_T(v) = t$  the paths from  $\nu(s)$  to  $u$  and from  $\nu(t)$  to  $v$  are vertex disjoint. For each clone-tree vertex  $u \in V(T)$ , set  $\ell^*(u) = s$  if  $\nu(s)$  is the first clone-tree vertex encountered on the unique path from  $u$  to  $r(T)$ . We claim that  $\ell^*$  is a vertex labeling of  $T$  subject to the two conditions of Definition 3.

1. By definition of  $\nu$ , we have that  $\nu(P) = r(T)$  and thus  $\ell^*(r(T)) = P$ .
2. Suppose for a contradiction that there exists a clone-tree leaf  $u$  such that  $\ell^*(u) \neq \bar{\ell}(u)$ . Let  $\bar{\ell}(u) = s$ . Now, by definition of  $\nu$ , it holds that  $\nu(s) \preceq u$ . Let  $\ell^*(u) = t$ . Note that by the premise that  $t \neq s$ . By construction, it holds that  $\nu(t) \preceq u$ . Since  $T$  is a tree, it must hold that either  $\nu(s) \prec \nu(t) \preceq u$  or  $\nu(t) \prec \nu(s) \preceq u$ . Either case would be a contradiction with the fact that the paths from  $\nu(s)$  to  $u$  and from  $\nu(t)$  to  $v$  are vertex disjoint. Hence,  $\ell^*(u) = \bar{\ell}(u)$  for all clone-tree leaves  $u$ .

By construction of  $\ell^*$ , we have that  $T[\{u \in V(T) \mid \ell^*(u) = s\}]$  is connected for each anatomical site  $s$ . Thus,  $\ell^*$  has migration number  $\mu_{\min} = m - 1$ .  $\square$

Note that the above condition can be checked in polynomial time. In the main text we described that a clone tree may admit many vertex labelings with the same minimum number of migrations but with very distinct migration graphs. To distinguish such vertex labelings, we consider additional objective functions.

### B.1.2 Comigration Number $\gamma(T, \ell)$

Given a clone tree  $T$  and vertex labeling  $\ell$ , a *comigration* is a subset of migration edges between the same anatomical sites that occur on distinct branches in the clone tree.

**Definition 5.** Given a clone tree  $T$  and vertex labeling  $\ell$ , a comigration is a subset  $X \subseteq E(T)$  of edges that (1) occur on distinct branches of  $T$ , and (2) there exist distinct anatomical sites  $s \neq t$  such that  $\ell(u) = s$  and  $\ell(v) = t$  for each  $(u, v) \in X$ .

Note that a comigration may consist of just a single migration edge. The comigration number  $\gamma(T, \ell)$  is the size of the smallest partition of migration edges into comigrations. Equivalently, one can determine  $\gamma(T, \ell)$  by counting the number  $\gamma(s, t)$  of comigrations between distinct anatomical sites  $s \neq t$ :

$$\gamma(T, \ell) = \sum_{s \neq t} \gamma(s, t). \quad (6)$$

The number  $\gamma(s, t)$  of comigrations between distinct anatomical sites  $s \neq t$  is the maximum number of migration edges  $(u, v)$  with  $\ell(u) = s$  and  $\ell(v) = t$  that are on the same path of  $T$  starting from the root  $r(T)$ .

If  $G$  does not contain a directed cycle then every path in  $T$  starting at  $r(T)$  contains at most one migration edge  $(u, v)$  labeled by the same sites  $\ell(u) = s$  and  $\ell(v) = t$ . The following observation thus follows.

**Observation 2.** *If  $G$  does not contain a directed cycle then the comigration number  $\gamma(T, \ell)$  equals the number of multi-edges in  $G$ , i.e.  $\gamma(T, \ell) = |\{(s, t) \in E(G)\}|$ .*

Supplementary Fig. 21 shows an example where  $G$  has a directed cycle and the comigration number  $\gamma(T, \ell)$  does not equal the number of multi-edges of  $G$ .

Since every comigration is composed of at least one migration, we make the following observation.

**Observation 3.** *The comigration number  $\gamma(T, \ell)$  is at most the number  $\mu(T, \ell)$  of migrations.*

Since there are  $m - 1$  metastases that each must be seeded, we note the following.

**Observation 4.** *The comigration number  $\gamma(T, \ell)$  is at least  $\gamma_{\min} = m - 1$ .*

The case where each metastasis is seeded by a single comigration results in the migration graph  $G$  being a multi-tree, as noted in the following observation.

**Observation 5.** *Let  $G$  be the migration graph of a vertex labeling  $\ell$  and clone tree  $T$ . Then,  $\gamma(T, \ell) = \gamma_{\min} = m - 1$  if and only if  $G$  is a multi-tree.*

Vertex labelings with  $\gamma_{\min}$  comigrations correspond to single-source seeding (S) migration patterns. There always exist such vertex labelings with  $\gamma_{\min}$  comigrations, as noted in the following.

**Observation 6.** *Let  $\ell$  be a vertex labeling of a clone tree  $T$  such that  $\ell(u) = P$  for every internal vertex  $u$ . Then,  $\gamma(T, \ell) = \gamma_{\min}$ .*

From Proposition 1 and Observation 3, the following result follows for the case of clone trees  $T$  with *homogeneous* anatomical sites, i.e. where each anatomical site is composed of a single clone.

**Proposition 2.** *A homogeneous clone tree has a vertex labeling with migration number  $\mu_{\min}$  and comigration number  $\gamma_{\min}$ .*

Recall that in Section A.2 we described that many analyses of metastatic cancers in the literature are based on the implicit assumption that samples are homogeneous, and derive a sample tree using standard phylogenetic techniques. If in addition only one sample were taken from each anatomical site then the above result implies that there exist a vertex labeling of the sample tree with a monoclonal single-source seeding (mS) pattern, migration number  $\mu_{\min}$  and comigration number  $\gamma_{\min}$ .

### B.1.3 Seeding Site Number $\sigma(T, \ell)$

We now introduce a third objective, the seeding site number  $\sigma(T, \ell)$ . A *seeding site* is an anatomical site  $s$  that is a source of migrations, that is, there exists a migration edge  $(u, v)$  where  $\ell(u) = s$ . Thus, the number  $\sigma(T, \ell)$  of seeding sites is defined as

$$\sigma(T, \ell) = |\{s \in \Sigma \mid \exists (u, v) \in E(T) : \ell(u) = s, \ell(v) \neq s\}|. \quad (7)$$

With  $m > 1$  anatomical sites, the number of seeding sites must be at least 1.

**Observation 7.** *The seeding site number  $\sigma(T, \ell)$  is at least  $\sigma_{\min} = 1$ .*

Since the clone-tree root  $r(T)$  must be labeled by the primary  $P$  (Definition 3), we make the following observation.

**Observation 8.** *Let  $G$  be the migration graph of a vertex labeling  $\ell$  and clone tree  $T$ . Then,  $\sigma(T, \ell) = \sigma_{\min} = 1$  if and only if  $G$  is a multi-tree where only the vertex corresponding to anatomical site  $P$  has out-degree greater than 0.*

Combining the above result with Observation 5, we obtain the following proposition.

**Proposition 3.** *If  $\sigma(T, \ell) = \sigma_{\min} = 1$  then  $\gamma(T, \ell) = \gamma_{\min} = m - 1$ .*

We extend Observation 4 to include the number of seeding sites.

**Observation 9.** *Let  $\ell$  be a vertex labeling of a clone tree  $T$  such that  $\ell(u) = P$  for every internal vertex  $u$ . Then,  $\gamma(T, \ell) = \gamma_{\min}$  and  $\sigma(T, \ell) = \sigma_{\min}$ .*

Supplementary Table 9 shows that the different scores distinguish different migration patterns.

## B.2 Problem Statements

### B.2.1 Parsimonious Migration History

The first problem we tackle is that of inferring a migration history (vertex labeling) given a clone tree. As mentioned, there are many different objective functions one could use to score a vertex labeling. From Observation 9 it follows that labeling every internal vertex  $u$  of  $T$  by  $P$  results in the minimum comigration number  $\gamma_{\min} = m - 1$  and the minimum seeding site number  $\sigma_{\min} = 1$ . Thus, as an objective it makes sense to consider the migration number  $\gamma(T, \ell)$  first. Based on Proposition 3, ties must be broken by considering the comigration number  $\gamma(T, \ell)$  followed by the seeding site number  $\sigma(T, \ell)$ . The Parsimonious Migration History (PMH) problem adheres to this prioritization of the different objectives.

**Problem 1** (Parsimonious Migration History (PMH)). *Given a clone tree  $T$ , find a vertex labeling  $\ell$  with the minimum migration number  $\mu(T, \ell) = \mu^*(T)$  and subsequently the smallest comigration number  $\gamma(\ell, T) = \hat{\gamma}(T)$  and smallest seeding site number  $\sigma(\ell, T) = \hat{\sigma}(T)$ , where*

$$\hat{\gamma}(T) = \min_{\ell: \mu(T, \ell) = \mu^*(T)} \gamma(T, \ell) \quad \text{and} \quad \hat{\sigma}(T) = \min_{\ell: \mu(T, \ell) = \mu^*(T), \gamma(T, \ell) = \hat{\gamma}(T)} \sigma(T, \ell). \quad (8)$$

*The resulting migration graph  $G$  is (1) restricted to an PS pattern, or (2) restricted to either an PS or S pattern, or (3) restricted to an PS, S or M pattern, or (4) unrestricted.*

## B.2.2 Parsimonious Migration History with Tree Refinement

A clone tree is a coarse-grained representation of a *cell tree*, whose vertices are cells and directed edges relate parental cells to their daughters. As the division of a cell results in two daughter cells, the cell tree is a rooted full binary tree, i.e. every internal vertex has out-degree 2 (Supplementary Fig. 22). Direct observation of the cell tree is nearly impossible as longitudinal high-fidelity measurements of somatic mutations in large numbers of single tumor cells is a huge technical challenge [31]. Conceptually, given a set  $X$  of mutations, collapsing all vertices of the cell tree with the same subset  $X' \subset X$  of mutations yields the clone tree. In practice, we infer a clone tree given measurements of only the extant clones. As such, the resulting clone tree will typically be non-binary and contain many *polytomies*, i.e. vertices with out-degree greater than 2 that reflect the uncertainty in the ancestral relationships of their children. We may refine polytomies by analyzing a non-binary clone tree in light of the migration history.

We start by defining two operations that can be performed on any clone tree  $T$ :  $\text{SPLIT}(u)$  and  $\text{CONTRACT}(u, v)$ . The operation  $\text{SPLIT}(u)$  takes as input a vertex  $u$  of  $T$  such that  $|\delta(u)| > 2$  and transform  $T$  into  $T'$  by introducing a new vertex  $u'$ , an edge  $(u, u')$  labeled by  $\lambda(u, u') = \perp$  and between 2 and  $|\delta(u)| - 1$  children of  $u$  are moved to  $u'$ . On the other hand, the operation  $\text{CONTRACT}(u, v)$  takes as input an edge  $(u, v)$  of  $T$  such that  $v \notin L(T)$  and  $\lambda(u, v) = \perp$ . This operation results in a new clone tree  $T'$  through the reassignment of the children  $\delta(v)$  of  $v$  to  $u$ , and the removal of edge  $(u, v)$  and vertex  $v$ . Since in a canonical clone tree  $T_{(\bar{T}, U)}$  the only edges labeled by  $\perp$  are incoming to leaves, the  $\text{CONTRACT}$  operation cannot be applied to  $T_{(\bar{T}, U)}$ . We write  $T \sim T'$  if  $T'$  can be obtained from  $T$  through a sequence of zero, one or multiple  $\text{SPLIT}$  and  $\text{CONTRACT}$  operations (Supplementary Fig. ??). A *refinement*  $T'$  is a clone tree obtained from  $T$  through a sequence of (zero or more)  $\text{SPLIT}$  operations. We have the following problem.

**Problem 2** (Parsimonious Migration History with Tree Refinement (PHM-TR)). *Given a clone tree  $T$ , find a refinement  $T'$  of  $T$  and vertex labeling  $\ell$  of  $T'$  with the minimum migration number  $\mu(T', \ell) = \mu^*(T')$  and subsequently the smallest comigration number  $\gamma(T', \ell) = \hat{\gamma}(T')$  and smallest seeding site number  $\sigma(T', \ell) = \hat{\sigma}(T')$ . The resulting migration graph  $G$  is (1) restricted to an PS pattern, or (2) restricted to either an PS or S pattern, or (3) restricted to an PS, S or M pattern, or (4) unrestricted.*

Since in a refinement  $T'$  of  $T$  all internal vertices with out-degree 1 are retained, we have the following proposition, which states that tree refinement does not lead to worse solutions for the migration history.

**Proposition 4.** *Let  $T'$  be a refinement of  $T$  and let  $\ell$  be a vertex labeling of  $T$ . Then, there exists a vertex labeling  $\ell'$  such that  $(\mu(T, \ell), \gamma(T, \ell), \sigma(T, \ell)) = (\mu(T', \ell'), \gamma(T', \ell'), \sigma(T', \ell'))$ .*

*Proof.* We define

$$\ell'(v') = \ell(\beta(v'))$$

and claim that

$$(\mu(T, \ell), \gamma(T, \ell), \sigma(T, \ell)) = (\mu(T', \ell'), \gamma(T', \ell'), \sigma(T', \ell')).$$



We show that  $\mu(T, \ell) = \mu(T', \ell')$  and  $\sigma(T, \ell) = \sigma(T', \ell')$  by showing that there exists a bijection between the migration edges of  $T$  and the migration edges of  $T'$ . More specifically, let  $Y \subseteq E(T)$  be the subset of migration edges in  $T$  and let  $Y' \subseteq E(T')$  be the subset of migration edges in  $T'$ . We define the function  $\xi : Y' \rightarrow Y$  such that

$$\xi(u', v') = (\beta(u'), \beta(v')).$$

We claim that  $\xi$  is a bijection. To see this, let  $(u, v) \in Y$  be a migration edge of  $T$ . Consider the subtree  $T'_v$  of  $T'$  as defined previously, and let  $v'$  be the root of this subtree and  $u'$  its parent. Now,  $u'$  must be in the tree  $T'_u$ . Thus,  $\beta(u') = u$  and  $\beta(v') = v$ . Hence,  $\xi(u', v') = (\beta(u'), \beta(v')) = (u, v)$ . Moreover,  $\ell'(u') = \ell(\beta(u')) = \ell(u)$  and  $\ell'(v') = \ell(\beta(v')) = \ell(v)$ . Therefore  $(u', v')$  is a migration edge in  $T'$ .

It remains to show that  $\gamma(T, \ell) = \gamma(T', \ell')$ . By construction, we have that  $u' \preceq_{T'} v'$  if and only if  $\beta(u') \preceq_T \beta(v')$ . Thus, the order of the migration edges in  $T'$  is respected in  $T$  by  $\xi$ . Hence,  $\gamma(T, \ell) = \gamma(T', \ell')$  and  $\sigma(T, \ell) = \sigma(T', \ell')$ .  $\square$

### B.2.3 Parsimonious Migration History with Tree Inference

In this section we introduce the problem of jointly inferring parsimonious clone trees and migration histories from bulk measurements of a metastatic cancer. We start by reviewing our previous results on clone tree inference from bulk samples [8], where we made the infinite sites assumption. This assumption assumes the absence of homoplasy and requires that a mutation only occurs once and is never lost. This is a reasonable assumption considering the length of the human genome and underlies many published methods for tumor phylogeny inference [7–9, 21, 22, 26, 33, 35, 40, 48, 52]. As a consequence of the infinite sites assumption, each edge of a mutation tree  $\bar{T}$  is labeled uniquely by a mutation from  $[n] = \{1, \dots, n\}$ . We represent a mutation tree  $\bar{T}$  by an  $n \times n$  binary matrix  $B$  called the mutation matrix, which is defined as follows.

**Definition 6** ([8]). *A matrix  $B \in \{0, 1\}^{n \times n}$  is a mutation matrix provided:*

1. *There exists exactly one  $r \in [n]$ , corresponding to the founder mutation, such that  $\sum_{i=1}^n b_{r,i} = 1$ .*
2. *For each  $j \in [n] \setminus \{r\}$  there exists exactly one  $i \in [n]$  such that  $\mathbf{b}_i \subseteq \mathbf{b}_j$  and  $\sum_{l=1}^n (b_{j,l} - b_{i,l}) = 1$ .*
3.  *$b_{i,i} = 1$  for all  $i \in [n]$ .*

We note that there is a 1-1 correspondence between the set  $\mathcal{B}_n$  of mutation matrices and the set  $\bar{\mathcal{T}}_n$  mutation trees on  $n$  mutations, and refer to [8] for additional details.

**Lemma 1** ([8]). *There is a 1-1 correspondence between  $\bar{\mathcal{T}}_n$  and  $\mathcal{B}_n$ .*

Most cancer sequencing studies consider one or more bulk samples from  $m$  anatomical sites. Let  $k$  be the total number of sequenced samples. We denote with  $\sigma(s)$  the set of samples taken from anatomical site  $s$ . Since each bulk sample  $p$  is a mixture of thousands to millions of cells, we do not directly observe a mutation tree  $\bar{T}$  (or mutation matrix  $B$ ). Instead, we measure various quantities that allow us to infer an  $k \times n$  frequency matrix  $F = [f_{p,i}]$  whose entries  $f_{p,i}$  describe the frequency of cells in sample  $p$  that harbor mutation  $i$  [9].

**Definition 7** ([8]). An  $k \times n$  matrix  $F = [f_{p,i}]$  is a frequency matrix provided  $0 \leq f_{p,i} \leq 1$  for each sample  $p$  and mutation  $i$ . Moreover, for each mutation  $i$  there exists a sample  $p$  such that  $f_{p,i} > 0$ .

Given a mutation tree  $\bar{T}$ , we describe the clonal composition of all samples by an  $k \times n$  mixture matrix  $U$ , defined as follows.

**Definition 8** ([8]). An  $k \times n$  matrix  $U = [u_{p,j}]$  is a mixture matrix provided  $u_{p,j} \geq 0$  and  $\sum_j u_{p,j} \leq 1$  for all samples  $p$ .

Frequency matrix  $F$  follows directly from mixture matrix  $U$  and mutation matrix  $B$ , as stated in the following equation.

$$F = UB. \quad (9)$$

We have the inverse problem.

**Problem 3** (Tree Inference Problem). Given frequency matrix  $F$  infer mixture matrix  $U$  and mutation matrix  $B$  such that  $F = UB$ .

The infinite sites assumption, used in the definitions of mutation matrix  $B$  and mutation tree  $T$ , leads to the following two important results.

**Lemma 2** ([8]). Given frequency matrix  $F$  and mutation matrix  $B$ , matrix  $U = [u_{p,j}]$  with entries

$$u_{p,j} = f_{p,j} - \sum_{v_l \in \delta(v_j)} f_{p,l} \quad (10)$$

is the unique matrix  $U$  such that  $F = UB$ .

Not every mutation matrix  $B$  yields a mixture matrix  $U$  with nonnegative entries given  $F$ . In previous work, we proved that nonnegativity is a necessary and sufficient condition for solutions to the Frequency Matrix Factorization problem [8].

**Lemma 3** ([8]). Given frequency matrix  $F$ , a mutation matrix  $B$  admits a mixture matrix  $U$  if and only if

$$f_{p,i} \geq \sum_{v_j \in \delta(v_i)} f_{p,j} \quad (\text{SC})$$

for each mutation  $i$  and sample  $p$ .

Thus, the set of extant clones and their anatomical sites is fully determined given a frequency matrix  $F$  and a mutation tree  $\bar{T}$  (mutation matrix  $B$ ). We say that a mutation matrix  $B$  (or equivalently mutation tree  $\bar{T}$ ) generates frequency matrix  $F$  if and only if there exists a mixture matrix  $U$  such that  $F = UB$ . While there is a unique  $U$  for a given mutation matrix  $B$  (mutation tree  $\bar{T}$ ), the problem of finding a mutation matrix given  $F$  is underdetermined, i.e. multiple mutation matrices  $B$  may explain the observed frequencies  $F$  [9]. On the other hand, the problem of deciding whether there exists a mutation matrix  $B$  (or mutation tree  $\bar{T}$ ) given  $F$  is NP-complete [8]. Supplementary Fig. 23 summarizes the results of this section.

**Clustering Mutations** There is extensive uncertainty in the measurements used to obtain  $F$ . We model the uncertainty of the frequencies  $F$  by considering confidence intervals denoted by  $F^- = [f_{s,i}^-]$  and  $F^+ = [f_{s,i}^+]$ . Given frequency intervals  $F^- = [f_{s,i}^-]$ ,  $F^+ = [f_{s,i}^+]$  and a mutation matrix  $B = [b_{j,i}]$ , there may be many mixture matrices  $U = [u_{s,j}]$  such that

$$\sum_j u_{s,j} \cdot b_{j,i} \in [f_{s,i}^-, f_{s,i}^+].$$

Thus, a given mutation matrix  $B$  may correspond to different mixture matrices  $U$  that assign different complements of extant clones to anatomical sites. We propose to deal with this ambiguity by jointly inferring parsimonious clone trees and migration histories.

**Problem 4** (Parsimonious Migration History with Tree Inference (PMH-TI)). *Given a mutation tree  $\bar{T}$ , whose vertices  $v_i$  are labeled by intervals  $[f_{s,i}^-, f_{s,i}^+]$  for each anatomical site  $s$ , find an assignment  $\hat{F} = [\hat{f}_{s,i}]$  of frequencies such that  $\hat{f}_{s,i} \in [f_{s,i}^-, f_{s,i}^+]$  and*

$$\hat{f}_{s,i} \geq \sum_{v_j \in \delta(v_i)} \hat{f}_{s,j} \tag{SC}$$

for each vertex  $v_i$  and anatomical site  $s$ , and the resulting clone tree  $T$  admits a refinement  $T'$  and vertex labeling  $\ell$  with the minimum migration number  $\mu(T', \ell) = \mu^*(T')$  and subsequently the smallest comigration number  $\gamma(T', \ell) = \hat{\gamma}(T')$  and smallest seeding site number  $\sigma(T', \ell) = \hat{\sigma}(T')$ . The resulting migration graph  $G$  is (1) restricted to an PS pattern, or (2) restricted to either an PS or S pattern, or (3) restricted to an PS, S or M pattern, or (4) unrestricted.

Note that multiple mutation matrices may be compatible with the observations  $(F^-, F^+)$ . Thus, we must solve the above problem for each mutation matrix  $B$  that explains the observations  $(F^-, F^+)$ . These mutation matrices may be inferred using our enumeration algorithm SPRUCE [9].

Since we only sequence a tiny subset of all the reads that are present in a bulk sample, we typically do not have the resolution to infer the ancestral relationships of every pair  $i, j$  of mutations. Specialized clustering algorithms have been proposed to groups of mutations with similar frequencies into *mutation clusters* [30, 41, 43, 53]. Importantly, a mutation cluster does not correspond to a clone but corresponds to an edge label of a mutation tree (Supplementary Fig. 3). Application of these clustering algorithms yields frequency interval matrices  $(F^-, F^+)$  whose columns correspond to mutation clusters rather than individual mutations. The problems that we introduced in the last two sections extend trivially to mutation clusters.

## B.3 MACHINA

### B.3.1 Unconstrained Parsimonious Migration History Problem

As noted by McPherson et al. [28], finding a vertex labeling  $\ell$  of a clone tree  $T$  with the minimum migration number  $\mu(T, \ell) = \mu^*(T)$  is an instance of the small phylogeny problem under a maximum parsimony objective with a single character, and can be solved in polynomial time with the Sankoff algorithm [45]. In particular, the recurrence  $M[u, s]$  yields the minimum number of state transitions, or the minimum number of migrations, when clone-tree vertex  $u$  is

assigned anatomical site  $s$ , and is defined as

$$M[u, s] = \begin{cases} \infty, & \text{if } u \in L(T) \text{ and } \ell(u) \neq s, \\ 0, & \text{if } u \in L(T) \text{ and } \ell(u) = s, \\ \sum_{v \in \delta(u)} \min_{t \in \Sigma} \{c_{s,t} + M[v, t]\}, & \text{if } u \notin L(T) \end{cases} \quad (11)$$

where  $c_{s,t} = 0$  if  $s = t$  and  $c_{s,t} = 1$  otherwise. Since the root vertex  $r(T)$  must be labeled by  $P$ , the minimum migration number  $\mu^*(T)$  is given by  $M[r(T), P]$ . We note that the entries of  $M$  can be computed using dynamic programming bottom-up from the leaves of  $T$  (Algorithm 2). To compute all minimum-migration labelings, we store for each vertex-anatomical site pair  $(v, s)$  the set  $\Delta(v, s)$  composed of all pairs  $(w, t)$  where  $w \in \delta(v)$  and  $t = \arg \min_{t \in \Sigma} \{c_{s,t} + M[v, t]\}$ .

We now describe how to adapt the backtrace step of the Sankoff algorithm to enumerate all vertex labelings that have  $\mu^*(T)$  migrations. The idea is to maintain a frontier  $H$ , which is a queue, composed of vertex-anatomical site pairs  $(v, s)$  that lead to a minimum-migration labeling. Initially, we set  $H = \{(r(T), P)\}$ . Then, at every iteration we remove the first pair  $(v, s)$  from the queue  $H$  and set  $\ell(v) = s$ . Next, we make a copy of  $H$  called  $H'$  from which we remove all pairs  $(w, t)$  where  $t = s$ . In addition, we add all pairs in  $\Delta(v, s)$  to the front of  $H'$ . We record whether we actually removed entries from  $H'$  and then recurse. To avoid enumerating duplicate labelings, we empty the queue if no entries were removed from  $H'$ . We report a labeling whenever  $H$  is empty (Algorithms 1 and 2, 3).

---

**Algorithm 1:** SANKOFF( $T, \Sigma, \ell$ )

---

**Input:** Clone tree  $T$  with anatomical sites  $\Sigma$  and leaf labeling  $\ell$ .

**Output:** All vertex labelings  $\ell$  of  $T$  with the minimum migration number  $\mu(T, \ell) = \mu^*(T)$ .

- 1  $M \leftarrow \emptyset$
  - 2 SOLVE( $T, \Sigma, \ell, M, r(T)$ )
  - 3  $H \leftarrow \{(r(T), P)\}$
  - 4 BACKTRACE( $T, M, \Delta, H, \ell$ )
- 

### B.3.2 Constrained Parsimonious Migration History Problem

The Sankoff algorithm, presented in the previous section, solves the unconstrained PMH problem, i.e. among the enumerated vertex labelings with migration number  $\mu^*(T)$  return those labelings with the comigration number  $\hat{\gamma}(T)$  and seeding site number  $\hat{\sigma}(T)$ . As such, the resulting migration graph  $G$  may contain directed cycles and correspond to a reseeded (R) pattern. Here, we present an integer linear program (ILP) for solving the PMH problem subject to one of the following restrictions where the migration graph  $G$  has (1) a PS pattern, (2) either a PS or S pattern, (3) a PS, S or an M pattern.

We order the anatomical sites  $\Sigma = \{1, \dots, m\}$  such that anatomical site 1 is the primary  $P$ . The clone-tree vertices  $V(T)$  are ordered such that  $v_1$  is the clone-tree root  $r(T)$ . We introduce variables  $\mathbf{x} \in \{0, 1\}^{|V(T)| \times m}$  such that  $x_{i,s}$

indicates whether clone-tree vertex  $v_i$  is labeled by anatomical site  $s$ , i.e.  $x_{i,s} = 1$  if and only if  $\ell(v_i) = s$ .

$$\sum_{s=1}^m x_{i,s} = 1 \quad \forall v_i \in V(T) \quad (12)$$

$$x_{1,1} = 1 \quad (13)$$

$$x_{i,s} = 1 \quad \forall v_i \in L(T), s = \ell(v_i) \quad (14)$$

$$x_{i,s} = 0 \quad \forall v_i \in L(T), s \neq \ell(v_i) \quad (15)$$

Constraints (12) model that each clone-tree vertex is labeled by only one anatomical site. Constraint (13) ensures that the clone-tree root is labeled by the primary tumor  $P$ , whereas Constraints (14) and (15) fix the labels of the clone-tree leaves.

To model whether an edge is a migration edge, we introduce variables  $\mathbf{z} \in \{0, 1\}^{|E(T)| \times m}$  such that  $z_{i,j,s} = 1$  if and only if the edge  $(v_i, v_j)$  is not a migration edge, i.e.  $\ell(v_i) = \ell(v_j)$ . Equivalently, variables  $z_{i,j,s}$  correspond to the product  $x_{i,s} \cdot x_{j,s}$ , which we model using the following constraints.

$$z_{i,j,s} \leq x_{i,s} \quad \forall (v_i, v_j) \in E(T), 1 \leq s \leq m \quad (16)$$

$$z_{i,j,s} \leq x_{j,s} \quad \forall (v_i, v_j) \in E(T), 1 \leq s \leq m \quad (17)$$

Note that we do not need to enforce that  $z_{i,j,s} \geq x_{i,s} + x_{j,s} - 1$  as this constraint is implied by the objective function where we minimize the number of migrations.

Next, we introduce variables  $\mathbf{y} \in \{0, 1\}^{|E(T)|}$ , which indicate whether the incident vertices of clone-tree edge  $(v_i, v_j)$  are labeled by different anatomical sites, i.e.  $y_{i,j} = 1$  if and only if  $\ell(v_i) \neq \ell(v_j)$ . Now, a clone-tree edge  $(v_i, v_j)$  is *not* a migration edge if and only if  $v_i$  and  $v_j$  are not labeled by the same anatomical site  $s$ . This is captured by the following constraints.

$$\sum_{s=1}^m z_{i,j,s} = 1 - y_{i,j} \quad \forall (v_i, v_j) \in E(T) \quad (18)$$

We now introduce variables  $\mathbf{c} \in \{0, 1\}^{m \times m}$  such that each variable  $c_{s,t}$  indicates whether there exists a migration edge  $(v_i, v_j)$  where  $v_i$  is labeled by anatomical site  $s$  and  $v_j$  is labeled by anatomical site  $t$ .

$$c_{s,t} \geq x_{i,s} + x_{j,t} - 1 \quad \forall (v_i, v_j) \in E(T), 1 \leq s, t \leq m \quad (19)$$

$$c_{s,s} = 0 \quad \forall 1 \leq s \leq m \quad (20)$$

Constraints (19) force  $c_{s,t}$  to be 1 if there exists an edge  $(v_i, v_j)$  such that  $x_{i,s} = 1$  and  $x_{j,t} = 1$ . Note that the objective function ensures that  $c_{s,t} = 0$  if no such edge exists. Constraints (20) set  $c_{s,s} = 0$  in accordance with the definitions of a migration and a comigration.

Finally, we introduce variables  $\mathbf{d} \in \{0, 1\}^m$  such that  $d_s = 1$  if and only if there exists a migration edge  $(v_i, v_j)$  where  $\ell(v_i) = s$ , as captured by the Constraints (21). Constraint (22) encodes that the primary  $P$  by definition is a

seeding site.

$$d_s \geq c_{s,t} \quad \forall 1 \leq s, t \leq m \quad (21)$$

$$d_1 = 1 \quad (22)$$

We have now introduced all the variables and constraints needed to formulate the PMH problem as an ILP. In the objective function we consider the migration number. We assume that the comigration number equals the number of multi-edges, which is the case for labelings that result in migration graphs that do not contain directed cycles (Observation 2). We multiply this number by  $1/|V(T)|$  to break ties in favor of labelings with smaller comigration number. In the case of further ties, i.e. labelings with migration number  $\mu^*$  and comigration number  $\hat{\gamma}$ , we consider the seeding site number, which we multiply by  $1/(m \cdot |V(T)|)$ . We thus have the following ILP.

$$\min_{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{c}, \mathbf{d}} \sum_{(v_i, v_j) \in E(T)} y_{i,j} + \frac{1}{|V(T)|} \sum_{s=1}^m \sum_{t=1}^m c_{s,t} + \frac{1}{m \cdot |V(T)|} \sum_{s=1}^m d_s$$

s.t. (12), (13), (14), (15), (16), (17), (18), (19), (20), (21) and (22)

$$x_{i,s} \in \{0, 1\} \quad \forall v_i \in V(T), 1 \leq s \leq m \quad (23)$$

$$y_{i,j} \in \{0, 1\} \quad \forall (v_i, v_j) \in E(T) \quad (24)$$

$$z_{i,j,s} \in \{0, 1\} \quad \forall (v_i, v_j) \in E(T), 1 \leq s \leq m \quad (25)$$

$$c_{s,t} \in \{0, 1\} \quad \forall 1 \leq s, t \leq m \quad (26)$$

$$d_s \in \{0, 1\} \quad \forall 1 \leq s \leq m \quad (27)$$

We now describe how to impose constraints on the topology of the migration graph, as described by variables  $\mathbf{c} \in \{0, 1\}^{m \times m}$ . Recall that for parallel single-source seeding (PS) it holds that  $s = P$  and  $t \neq P$  for all edges  $(s, t)$  of the migration graph. We encode this using the following constraints.

$$\sum_{s=1}^m c_{s,1} = 0 \quad (28)$$

$$c_{s,t} = 0 \quad \forall 2 \leq s, t \leq m \quad (29)$$

$$c_{1,t} = 1 \quad \forall 2 \leq t \leq m \quad (30)$$

Constraints (28) prevent reseeding into the primary  $P$ . Constraints (29) prevent migration edges between different metastases and Constraints (30) enforce that each metastasis is seeded from  $P$ .

For single-source seeding (S), we require that variables  $\mathbf{c} \in \{0, 1\}^{m \times m}$  are constrained to be a tree. In addition to Constraints (28), we introduce the following constraints that ensure that each metastatic site is seeded from only one other site.

$$\sum_{s=1}^m c_{s,t} = 1 \quad \forall 2 \leq t \leq m \quad (31)$$

For multi-source seeding (M), we require the migration graph  $G$  to be a multi-DAG, i.e.  $G$  must not contain any directed cycles. To model this, we denote the set of all directed cycles by  $\mathcal{C}$  and introduce the following cycle

inequalities.

$$\sum_{(s,t) \in C} c_{s,t} \leq |C| - 1 \quad \forall 1 \leq t \leq m, C \in \mathcal{C} \quad (32)$$

To summarize, we model various seeding patterns by extending the ILP with different constraints: we model PS using (28), (29) and (30), S by (28) and (31), and M by (32).

We note that the above ILP is not guaranteed to an yield optimal solution when one does not impose a topological constraint (PS, S or M). However, should, for a specific unconstrained instance, the resulting migration graph be a multi-DAG, then optimality and correctness are guaranteed.

### B.3.3 Parsimonious Migration History with Tree Refinement Problem

We consider the PHM-TR problem, where given a clone tree  $T$  we seek a binarization  $T'$  of  $T$  and vertex labeling  $\ell$  of  $T'$  such that  $(\mu(T', \ell), \gamma(T', \ell), \sigma(T', \ell))$  is minimum (Problem 2). We refer to Section B.2.2 for the formal definition of a binarization. Here, we describe an ILP for solving this problem.

We define a directed simple graph  $S$ , called the *search graph*, that will contain all binarizations  $T'$  of a given clone tree  $T$  as constrained spanning trees. Similarly to the definition of a binarization, we do this recursively using the function EXPAND. We initialize  $S$  to contain a single vertex  $r(S)$  and invoke EXPAND( $r(T)$ ,  $r(S)$ ). For each call EXPAND( $u$ ,  $u'_1$ ), let  $\delta(u) = \{v_1, \dots, v_k\}$  be the children of  $u$ . If  $k = 1$ , we add the vertex  $v'_1$  to  $S$  as well as the edge  $(u'_1, v'_1)$ . If  $k > 1$ , we add the vertices  $\{u'_2, \dots, u'_{k-1}, v'_1, \dots, v'_k\}$ . In addition, we introduce edges  $(u'_i, v'_j)$  for each  $i \in \{1, \dots, k-1\}$  and  $j \in \{1, \dots, k\}$ , and the edges  $(u'_i, u'_j)$  for each  $i \in \{1, \dots, k-1\}$  and  $j \in \{i+1, \dots, k-1\}$ . We recurse on each child  $v_i$  of  $u$  and its counterpart  $v'_i$ . Algorithm 4 shows the pseudocode.

We now proceed with presenting an ILP for solving the PHM-TR problem. We introduce variables  $\mathbf{w} \in \{0, 1\}^{|E(S)|}$  such that each variable  $w_{i,j}$  indicates whether edge  $(v_i, v_j) \in E(S)$  is in the solution binary clone tree. We require that each vertex  $v_i \neq r(S)$  has in-degree 1 and that each vertex  $v_i \neq L(S)$  has either out-degree 1 or 2.

$$\sum_{(v_i, v_j) \in E(S)} w_{i,j} = 1 \quad \forall v_j \in V(S) \setminus \{r(S)\} \quad (33)$$

$$\sum_{v_j \in \delta(v_i)} w_{i,j} = 1 \quad \forall v_j \in I(S), |\delta(v_i)| = 1 \quad (34)$$

$$\sum_{v_j \in \delta(v_i)} w_{i,j} = 2 \quad \forall v_j \in I(S), |\delta(v_i)| > 1 \quad (35)$$

Moreover, we change the domains of variables  $\mathbf{x} \in \{0, 1\}^{|V(T)| \times m}$ ,  $\mathbf{y} \in \{0, 1\}^{|E(T)|}$  and  $\mathbf{z} \in \{0, 1\}^{|E(T)| \times m}$  and their corresponding constraints to  $\mathbf{x} \in \{0, 1\}^{|V(S)| \times m}$ ,  $\mathbf{y} \in \{0, 1\}^{|E(S)|}$  and  $\mathbf{z} \in \{0, 1\}^{|E(S)| \times m}$ .

We force  $y_{i,j}$  to be 0 if the edge  $(v_i, v_j)$  of  $S$  is not picked using the following constraint.

$$y_{i,j} \leq w_{i,j} \quad \forall (v_i, v_j) \in E(S) \quad (36)$$

We replace (18) by the following constraint, where we only consider edges  $(v_i, v_j)$  that are in the solution.

$$\sum_{s=1}^m z_{i,j,s} = w_{i,j} - y_{i,j} \quad \forall (v_i, v_j) \in E(S) \quad (37)$$

We replace (19) by the following constraint, where in addition to  $x_{i,s} = 1$  and  $x_{j,s} = 1$  we require  $w_{i,j} = 1$  for  $c_{s,t}$  to be 1.

$$c_{s,t} \geq x_{i,s} + x_{j,t} + w_{i,j} - 2 \quad \forall (v_i, v_j) \in E(S), 1 \leq s, t \leq m \quad (38)$$

Thus, we have the following ILP.

$$\min_{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{c}, \mathbf{d}, \mathbf{w}} \sum_{(v_i, v_j) \in E(S)} y_{i,j} + \frac{1}{|V(S)|} \sum_{s=1}^m \sum_{t=1}^m c_{s,t} + \frac{1}{m \cdot |V(S)|} \sum_{s=1}^m d_s$$

s.t. (12), (13), (14), (15), (16), (17), (20), (21), (22) and (33) – (38)

$$x_{i,s} \in \{0, 1\} \quad \forall v_i \in V(S), 1 \leq s \leq m \quad (39)$$

$$y_{i,j} \in \{0, 1\} \quad \forall (v_i, v_j) \in E(S) \quad (40)$$

$$z_{i,j,s} \in \{0, 1\} \quad \forall (v_i, v_j) \in E(S), 1 \leq s \leq m \quad (41)$$

$$c_{s,t} \in \{0, 1\} \quad \forall 1 \leq s, t \leq m \quad (42)$$

$$d_s \in \{0, 1\} \quad \forall 1 \leq s \leq m \quad (43)$$

$$w_{i,j} \in \{0, 1\} \quad \forall (v_i, v_j) \in E(S) \quad (44)$$

Similarly to the PMH problem, we consider restrictions of the allowed seeding patterns by introducing additional constraints: we model PS using (28), (29) and (30), S by (28) and (31), and M by (32).

### B.3.4 Parsimonious Migration History with Tree Inference Problem

We consider the PMH-TI problem. In this problem we are given a mutation tree  $\bar{T}$  (or mutation matrix  $B$ ) whose vertices  $v_i$  are labeled by frequency intervals  $[f_{s,i}^-, f_{s,i}^+]$  for each anatomical site  $s$ . Each frequency assignment  $\hat{F} = [\hat{f}_{s,i}]$  yields a clone tree  $T$  potentially containing polytomies. The task is to find an assignment  $\hat{F} = [\hat{f}_{s,i}]$  of frequencies such that for each vertex  $v_i$  and anatomical site  $s$  the following three conditions hold: (1)  $\hat{f}_{s,i} \in [f_{s,i}^-, f_{s,i}^+]$ , (2)  $\hat{f}_{s,i} \geq \sum_{v_j \in \delta(v_i)} \hat{f}_{s,j}$ , and (3) the resulting clone tree  $T$  obtained from  $\hat{F}$  and  $B$  admits a binarization  $T'$  with the minimum migration number  $\mu^*(T')$  and subsequently the smallest comigration number  $\hat{\gamma}(T')$  and smallest seeding site number  $\hat{\sigma}(T')$ .

We start by defining the extended search graph  $R$ , a directed graph that will contain all binarized clone trees as subtrees. To that end, we need to identify all potential extant clones given a mutation tree  $\bar{T}$  and frequency matrices  $(F^-, F^+)$ . In [9] we defined a lower bound  $\hat{F}^- = [\hat{f}_{s,i}^-]$  as follows.

$$\hat{f}_{s,i}^- = \begin{cases} f_{s,i}^-, & \text{if } v_i \in L(\bar{T}). \\ \max \left\{ f_{s,i}^-, \sum_{v_j \in \delta(v_i)} \hat{f}_{s,j}^- \right\}, & \text{if } v_i \notin L(\bar{T}). \end{cases} \quad (45)$$

Observe that  $\hat{f}_{s,i}^-$  can be computed bottom-up from the leaves of  $\bar{T}$ . Intuitively,  $\hat{f}_{s,i}^-$  is the minimum frequency that a mutation-tree vertex  $v_i$  can attain for anatomical site  $s$  while satisfying the sum condition and respecting the provided



lower bounds  $F^-$ . Similarly, we define

$$\hat{f}_{s,i}^+ = \begin{cases} f_{s,i}^+, & \text{if } v_i = r(\bar{T}), \\ \min \left\{ f_{s,i}^+, \hat{f}_{s,\pi(v_i)}^+ - \sum_{v_j \in \delta(\pi(v_i)) \setminus \{v_i\}} \hat{f}_{s,j}^- \right\}, & \text{if } v_i \neq r(\bar{T}). \end{cases} \quad (46)$$

Intuitively,  $\hat{f}_{s,i}^+$  is the maximum frequency that a mutation-tree vertex  $v_i$  can attain for anatomical site  $s$  while satisfying the sum condition and respecting the provided frequency bounds ( $F^-, F^+$ ). This quantity can be computed top-down from the root  $r(\bar{T})$ . Now, we can define the maximum mixture proportion  $u_{s,i}^+$  as follows.

$$u_{s,i}^+ = \hat{f}_{s,i}^+ - \sum_{v_j \in \delta(v_i)} \hat{f}_{s,i}^-. \quad (47)$$

Clone  $v_i \in V(\bar{T})$  is potentially present in anatomical site  $s$  if  $u_{s,i}^+ > 0$ . If this is the case, we introduce a new vertex  $v_{i,s}$  labeled by  $\ell(v_{i,s}) = s$  and directed edge  $(v_i, v_{i,s})$ . Let  $T$  denote the tree obtained by following this procedure. We obtain the extended search graph  $R$  by invoking  $\text{EXPAND}(r(T), r(R))$ , where initially  $R$  only contains a single vertex  $r(R)$ .

To represent that a clone  $v_i \in V(\bar{T})$  is absent in anatomical site  $s$ , or equivalently that vertex  $v_{i,s}$  is absent, we introduce a dummy anatomical site, which we represent by index 0. We change the domains of variables  $\mathbf{x} \in \{0, 1\}^{|V(T)| \times m}$ ,  $\mathbf{y} \in \{0, 1\}^{|E(T)|}$  and  $\mathbf{z} \in \{0, 1\}^{|E(T)| \times m}$  and their corresponding constraints to  $\mathbf{x} \in \{0, 1\}^{|V(R)| \times (m+1)}$ ,  $\mathbf{y} \in \{0, 1\}^{|E(R)|}$  and  $\mathbf{z} \in \{0, 1\}^{|E(R)| \times (m+1)}$ . Importantly, the constraints involving variables  $\mathbf{z} \in \{0, 1\}^{|E(R)| \times (m+1)}$  will not include anatomical site  $s$  unless they are reintroduced below.

Recall that  $L(R)$  corresponds to the set of vertices with out-degree 0, which in our case are vertices  $v_{i,s}$  indicating that clone  $v_i$  occurs in anatomical site  $s > 0$ . We introduce the following constraints.

$$\sum_{s=0}^m x_{i,s} = 1 \quad \forall v_i \in V(R) \quad (48)$$

$$x_{i,s} + x_{i,0} = 1 \quad \forall v_{i,s} \in L(R) \quad (49)$$

$$x_{i,t} = 0 \quad \forall v_{i,s} \in L(T), t \neq s \quad (50)$$

Constraints (48) impose that each vertex is assigned exactly one anatomical site. Constraints 49 state that the leaf vertices  $v_{i,s} \in L(R)$  are either assigned anatomical site  $s$  or anatomical site 0. Constraints (50) state that anatomical site 0 cannot be assigned to the internal vertices of  $R$ .

We introduce fractional variables  $\mathbf{u}, \mathbf{f} \in [0, 1]^{m \times |I(R)|}$ , where  $u_{s,i}$  denotes the mixture proportion of clone  $v_i$  in anatomical site  $s$  and  $f_{s,i}$  denotes the frequency of mutation  $i$  in anatomical site  $s$ . We have the following constraints.

$$f_{s,i} \geq \hat{f}_{s,i}^- \quad \forall 1 \leq s \leq m, v_i \in I(R) \quad (51)$$

$$f_{s,i} \leq \hat{f}_{s,i}^+ \quad \forall 1 \leq s \leq m, v_i \in I(R) \quad (52)$$

$$f_{s,i} \geq \sum_{v_j \in \delta(v_i) \setminus L(R)} f_{s,j} \quad \forall 1 \leq s \leq m, v_i \in I(R) \quad (53)$$

$$u_{s,i} = f_{s,i} - \sum_{v_j \in \delta(v_i) \setminus L(R)} f_{s,j} \quad \forall 1 \leq s \leq m, v_i \in I(R) \quad (54)$$

Constraints (51) and (52) restrict the frequencies to be within the supplied confidence intervals. Constraints (53) model the sum condition and constraints (54) model the mixture proportion.

We relate mixture proportions to anatomical site labelings using the following big M constraint.

$$x_{i,s} \geq u_{s,i} \qquad \forall v_{i,s} \in L(R) \qquad (55)$$

$$z_{i,(i,s),0} \leq 1 - u_{s,i} \qquad \forall (v_i, v_{i,s}) \in E(R), v_{i,s} \in L(R) \qquad (56)$$

$$z_{i,j,0} = 0 \qquad \forall (v_i, v_j) \in E(R), v_{i,j} \notin L(R) \qquad (57)$$

Constraints (55) force variables  $x_{i,s}$  to be 1 if clone  $v_i$  is present in anatomical site  $s > 0$ . Constraints (56) prevent  $z_{i,(i,s),0}$  from being 1 if clone  $v_i$  is present in anatomical site  $s > 0$ . Anatomical site 0 can only be used for the leaf vertices of  $R$  as encoded by Constraints (57). We replace (37) by the following constraint, where we only consider edges  $(v_i, v_j)$  that are in the solution.

$$\sum_{s=0}^m z_{i,j,s} = w_{i,j} - y_{i,j} \qquad \forall (v_i, v_j) \in E(R) \qquad (58)$$

Thus, we have the following mixed integer linear program.

$$\min_{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{c}, \mathbf{d}, \mathbf{w}, \mathbf{f}, \mathbf{u}} \sum_{(v_i, v_j) \in E(R)} y_{i,j} + \frac{1}{|V(R)|} \sum_{s=1}^m \sum_{t=1}^m c_{s,t} + \frac{1}{m \cdot |V(R)|} \sum_{s=1}^m d_s$$

s.t. (13), (16), (17), (20), (21), (22) and (33) – (36), (38), (48) – (58)

$$x_{i,s} \in \{0, 1\} \qquad \forall v_i \in V(R), 0 \leq s \leq m \qquad (59)$$

$$y_{i,j} \in \{0, 1\} \qquad \forall (v_i, v_j) \in E(R) \qquad (60)$$

$$z_{i,j,s} \in \{0, 1\} \qquad \forall (v_i, v_j) \in E(R), 0 \leq s \leq m \qquad (61)$$

$$c_{s,t} \in \{0, 1\} \qquad \forall 1 \leq s, t \leq m \qquad (62)$$

$$d_s \in \{0, 1\} \qquad \forall 1 \leq s \leq m \qquad (63)$$

$$w_{i,j} \in \{0, 1\} \qquad \forall (v_i, v_j) \in E(R) \qquad (64)$$

$$0 \leq f_{s,i} \leq 1 \qquad \forall 1 \leq s \leq m, v_i \in V(\bar{T}) \qquad (65)$$

$$0 \leq u_{s,i} \leq 1 \qquad \forall 1 \leq s \leq m, v_i \in V(\bar{T}) \qquad (66)$$

Similarly to the PMH and PHM-TR problem, we consider restrictions of the allowed seeding patterns by introducing additional constraints: we model PS using (28), (29) and (30), S by (28) and (31), and M by (32).

---

**Algorithm 2:** SOLVE( $T, \Sigma, \ell, M, u$ )

---

**Input:** Clone tree  $T$  with anatomical sites  $\Sigma$  and leaf labeling  $\ell$ , table  $M$  and clone-tree vertex  $u$ .

**Output:** Computes the entries  $M[u, s]$  for each anatomical site  $s \in \Sigma$ .

```
1 if  $u \in L(T)$  then
2   foreach  $s \in \Sigma$  do
3     if  $\ell(u) = s$  then
4        $M[u, s] \leftarrow 0$ 
5        $\Delta[u, s] \leftarrow \emptyset$ 
6     else
7        $M[u, s] \leftarrow \infty$ 
8        $\Delta[u, s] \leftarrow \emptyset$ 
9   else
10    foreach  $v \in \delta(u)$  do
11      SOLVE( $T, \Sigma, \ell, M, v$ )
12    foreach  $s \in \Sigma$  do
13       $M[u, s] \leftarrow 0$ 
14      foreach  $v \in \delta(u)$  do
15         $c \leftarrow \infty$ 
16        foreach  $t \in \Sigma$  do
17          if  $\text{COST}(M, u, s, v, t) < c$  then
18             $c \leftarrow \text{COST}(M, u, s, v, t)$ 
19           $M[u, s] \leftarrow M[u, s] + c$ 
20           $\Delta[u, s] \leftarrow \emptyset$ 
21          foreach  $t \in \Sigma$  do
22            if  $\text{COST}(M, u, s, v, t) = c$  then
23               $\Delta[u, s] \leftarrow \Delta[u, s] \cup \{(v, t)\}$ 
```

---

---

**Algorithm 3:** BACKTRACE( $T, \ell, M, \Delta, H$ )

---

**Input:** Clone tree  $T$  with anatomical sites  $\Sigma$  and leaf labeling  $\ell$ , table  $M$ , back pointers  $\Delta$ , and  $H$  is the frontier.

**Output:** Enumerates all vertex labelings  $\ell$  of  $T$  with minimum migration number  $\mu(T, \ell) = \mu^*(T)$ .

```
1 if  $H = \emptyset$  then
2   Report  $\ell$ 
3 else
4   done  $\leftarrow$  False
5   while  $H \neq \emptyset$  and not done do
6      $(v, s) \leftarrow \text{POP}(F)$ 
7      $\ell(v) = s$ 
8      $H' \leftarrow \emptyset$ 
9     foreach  $(w, t) \in H$  do
10      if  $s \neq t$  then  $H' \leftarrow \{(w, t)\} \cup H$ 
11      if  $H' = H$  then done  $\leftarrow$  True
12      foreach  $(w, t) \in \Delta(v, s)$  do
13         $H \leftarrow \{(w, t)\} \cup H$ 
14      BACKTRACE( $T, \ell, M, \Delta, H'$ )
```

---

---

**Algorithm 4:** EXPAND( $u, u'_1$ )

---

**Input:** Vertex  $u \in V(T)$  and corresponding vertex  $u'_1 \in V(S)$ .

**Output:** Constructs a subgraph rooted at  $u'_1$  that contains all binarizations of the subtree of  $T$  rooted at  $u$  as spanning trees.

```
1 Let  $\delta(u) = \{v_1, \dots, v_k\}$ 
2 if  $k = 1$  then
3    $V(S) \leftarrow V(S) \cup \{v'_1\}$ 
4    $E(S) \leftarrow E(S) \cup \{(u'_1, v'_1)\}$ 
5 else
6    $V(S) \leftarrow V(S) \cup \{u'_2, \dots, u'_{k-1}\}$ 
7    $V(S) \leftarrow V(S) \cup \{v'_1, \dots, v'_k\}$ 
8   for  $i \leftarrow 1$  to  $k - 1$  do
9     for  $j \leftarrow i + 1$  to  $k - 1$  do
10       $E(S) \leftarrow E(S) \cup \{(u'_i, u'_j)\}$ 
11     for  $j \leftarrow 1$  to  $k$  do
12       $E(S) \leftarrow E(S) \cup \{(u'_i, v'_j)\}$ 
13 for  $i \leftarrow 1$  to  $k$  do
14   EXPAND( $v_i, v'_i$ )
```

---

## C References

- [1] Joao M Alves, Tamara Prieto, and David Posada. Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*, July 2017.
- [2] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550, 2010.
- [3] David Brown, Dominiek Smeets, Borbála Székely, Denis Larsimont, A Marcell Szász, Pierre-Yves Adnet, Françoise Rothé, Ghizlane Rouas, Zsófia I Nagy, Zsófia Faragó, Anna-Mária Tókécs, Magdolna Dank, Gyöngyvér Szentmártoni, Nóra Udvarhelyi, Gabriele Zoppoli, Lajos Pusztai, Martine Piccart, Janina Kulka, Diether Lambrechts, Christos Sotiriou, and Christine Desmedt. Erratum: Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nature communications*, 8:15759, June 2017.
- [4] Peter J Campbell, Shinichi Yachida, Laura J Mudie, Philip J Stephens, Erin D Pleasance, Lucy A Stebbings, Laura A Morsberger, Calli Latimer, Stuart McLaren, Meng-Lay Lin, David J McBride, Ignacio Varela, Serena A Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P Butler, Jon W Teague, Constance A Griffin, John Burton, Harold Swerdlow, Michael A Quail, Michael R Stratton, Christine Iacobuzio-Donahue, and P Andrew Futreal. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113, October 2010.
- [5] Youn Jin Choi, Je-Keun Rhee, Soo Young Hur, Min Sung Kim, Sung Hak Lee, Yeun-Jun Chung, Tae-Min Kim, and Sug-Hyung Lee. Intra-individual genomic heterogeneity of high-grade serous carcinoma of the ovary and clinical utility of ascitic cancer cells for mutation profiling. *The Journal of Pathology*, 241(1):57–66, January 2017.
- [6] H X Dang, B S White, S M Foltz, C A Miller, J Luo, R C Fields, and C A Maher. ClonEvol: clonal ordering and visualization in cancer sequencing. *Annals of Oncology*, September 2017.
- [7] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun H Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, February 2015.
- [8] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, June 2015.
- [9] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J Raphael. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, 3(1):43–53, July 2016.

- [10] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, November 1981.
- [11] Walter M Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.
- [12] Marco Gerlinger et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*, 366(10):883–92, Mar 2012.
- [13] Marco Gerlinger et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multi-region sequencing. *Nat Genet*, 46(3):225–33, Mar 2014.
- [14] William J Gibson, Erling A Hoivik, Mari K Halle, Amaro Taylor-Weiner, Andrew D Cherniack, Anna Berg, Frederik Holst, Travis I Zack, Henrica M J Werner, Kjersti M Staby, Mara Rosenberg, Ingunn M Stefansson, Kanthida Kusunmano, Aaron Chevalier, Karen K Mauland, Jone Trovik, Camilla Krakstad, Marios Giannakis, Eran Hodis, Kathrine Woie, Line Bjorge, Olav K Vintermyr, Jeremiah A Wala, Michael S Lawrence, Gad Getz, Scott L Carter, Rameen Beroukhim, and Helga B Salvesen. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nature Genetics*, 48(8):848–855, August 2016.
- [15] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose M C Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini M L Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O’Meara, Kevin J Dawson, William Isaacs, Michael R Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C Whitaker, ICGC Prostate UK Group, David E Neal, Colin S Cooper, Rosalind A Eeles, Tapio Visakorpi, Peter J Campbell, Ultan McDermott, David C Wedge, and G Steven Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, April 2015.
- [16] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.
- [17] Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–86, June 2014.
- [18] Iman Hajirasouliha and Benjamin J Raphael. Reconstructing Mutational History in Multiply Sampled Tumors Using Perfect Phylogeny Mixtures. In *Algorithms in Bioinformatics*, pages 354–367. Springer, Berlin, Heidelberg, Berlin, Heidelberg, September 2014.
- [19] Katherine A Hoadley, Marni B Siegel, Krishna L Kanchi, Christopher a Miller, Li Ding, Wei Zhao, Xiaping He, Joel S Parker, Michael C Wendl, Robert S Fulton, Ryan T Demeter, Richard K Wilson, Lisa A Carey, Charles M Perou, and Elaine R Mardis. Tumor Evolution in Two Patients with Basal-like Breast Cancer: A Retrospective Genomics Study of Multiple Metastases. *PLOS Med*, 13(12):e1002174, December 2016.
- [20] Hedayatollah Hosseini, Milan M S Obradović, Martin Hoffmann, Kathryn L Harper, Maria Soledad Sosa, Melanie Werner-Klein, Lahiri Kanth Nanduri, Christian Werno, Carolin Ehrl, Matthias Maneck, Nina Patwary,

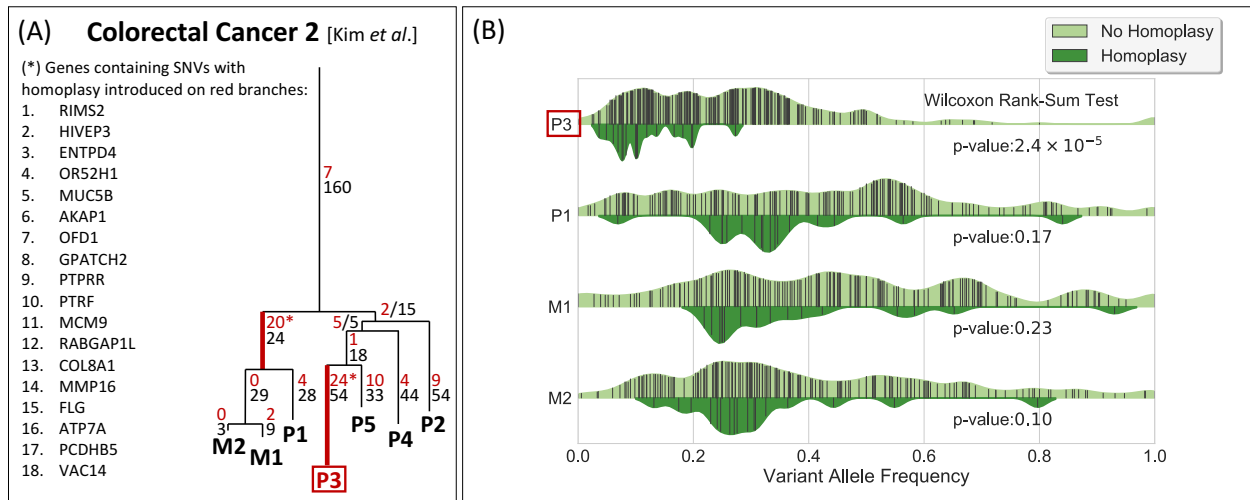
- Gundula Haunschild, Miodrag Gužvić, Christian Reimelt, Michael Grauvogl, Norbert Eichner, Florian Weber, Andreas D Hartkopf, Florin-Andrei Taran, Sara Y Brucker, Tanja Fehm, Brigitte Rack, Stefan Buchholz, Rainer Spang, Gunter Meister, Julio A Aguirre-Ghiso, and Christoph A Klein. Early dissemination seeds metastasis in breast cancer. *Nature*, 540(7634):552–558, December 2016.
- [21] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, May 2016.
- [22] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35, 2014.
- [23] Tae-Min Kim, Seung-Hyun Jung, Chang Hyeok An, Sung Hak Lee, In-Pyo Baek, Min Sung Kim, Sung-Won Park, Je-Keun Rhee, Sug-Hyung Lee, and Yeun-Jun Chung. Subclonal Genomic Architectures of Primary and Metastatic Colorectal Cancer Based on Intratumoral Genetic Heterogeneity. *Clinical Cancer Research*, 21(19):4461–4472, October 2015.
- [24] Bin Liu, Yoshitsugu Mitani, Xiayu Rao, Mark Zafereo, Jianjun Zhang, Jianhua Zhang, P Andrew Futreal, Guillermina Lozano, and Adel K El-Naggar. Spatio-Temporal Genomic Heterogeneity, Phylogeny, and Metastatic Evolution in Salivary Adenoid Cystic Carcinoma. *JNCI: Journal of the National Cancer Institute*, 109(10), October 2017.
- [25] H Lote, I Spiteri, L Ermini, A Vatsiou, A Roy, A McDonald, N Maka, M Balsitis, N Bose, M Simbolo, A Mafficini, A Lampis, J C Hahne, F Trevisani, Z Eltahir, G Mentrasti, C Findlay, E A J Kalkman, M Punta, B Werner, S Lise, A Aktipis, C Maley, M Greaves, C Braconi, J White, M Fassan, A Scarpa, A Sottoriva, and N Valeri. Carbon dating cancer: defining the chronology of metastatic progression in colorectal cancer. *Annals of Oncology*, 28(6):1243–1249, June 2017.
- [26] Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, May 2015.
- [27] Melissa Q McCreery, Kyle D Halliwill, Douglas Chin, Reyno Delrosario, Gillian Hirst, Peter Vuong, Kuang-Yu Jen, James Hewinson, David J Adams, and Allan Balmain. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nature Medicine*, 21(12):1514–1520, November 2015.
- [28] Andrew McPherson, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, Leah M Prentice, Jaswinder Khattri, Maia A Smith, Cydney B Nielsen, Sarah C Mullaly, Steve Kalloger, Anthony Karnezis, Karey Shumansky, Celia Siu, Jamie Rosner, Hector Li Chan, Julie Ho, Nataliya Melnyk, Janine Senz, Winnie Yang, Richard Moore, Andrew J Mungall, Marco a Marra, Alexandre Bouchard-Côté, C Blake Gilks, David G Huntsman, Jessica N McAlpine, Samuel Aparicio, and Sohrab P Shah. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*, May 2016.

- [29] Brandon Milholland, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, 8, 2017.
- [30] Christopher A Miller et al. Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665, Aug 2014.
- [31] Nicholas E Navin. Cancer genomics: one cell at a time. *Genome Biol*, 15(8):452, 2014.
- [32] Daniel E Newburger, Dorna Kashef-Haghighi, Ziming Weng, Raheleh Salari, Robert T Sweeney, Alayne L Brunner, Shirley X Zhu, Xiangqian Guo, Sushama Varma, Megan L Troxell, Robert B West, Serafim Batzoglou, and Arend Sidow. Genome evolution during progression to breast cancer. *Genome Research*, 23(7):1097–1108, July 2013.
- [33] Serena Nik-Zainal et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, May 2012.
- [34] P C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8, Oct 1976.
- [35] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):91, May 2015.
- [36] William M Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, April 2012.
- [37] Johannes G Reiter, Ivana Bozic, Krishnendu Chatterjee, and Martin A Nowak. Ttp: tool for tumor progression. In *International Conference on Computer Aided Verification*, pages 101–106. Springer, 2013.
- [38] Johannes G Reiter, Alvin P Makohon-Moore, Jeffrey M Gerold, Ivana Bozic, Krishnendu Chatterjee, Christine A Iacobuzio-Donahue, Bert Vogelstein, and Martin A Nowak. Reconstructing metastatic seeding patterns of human cancers. *Nature communications*, 8:14114, January 2017.
- [39] D F Robinson and L R Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, February 1981.
- [40] Edith M Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):69, 2016.
- [41] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, April 2014.
- [42] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, July 1987.

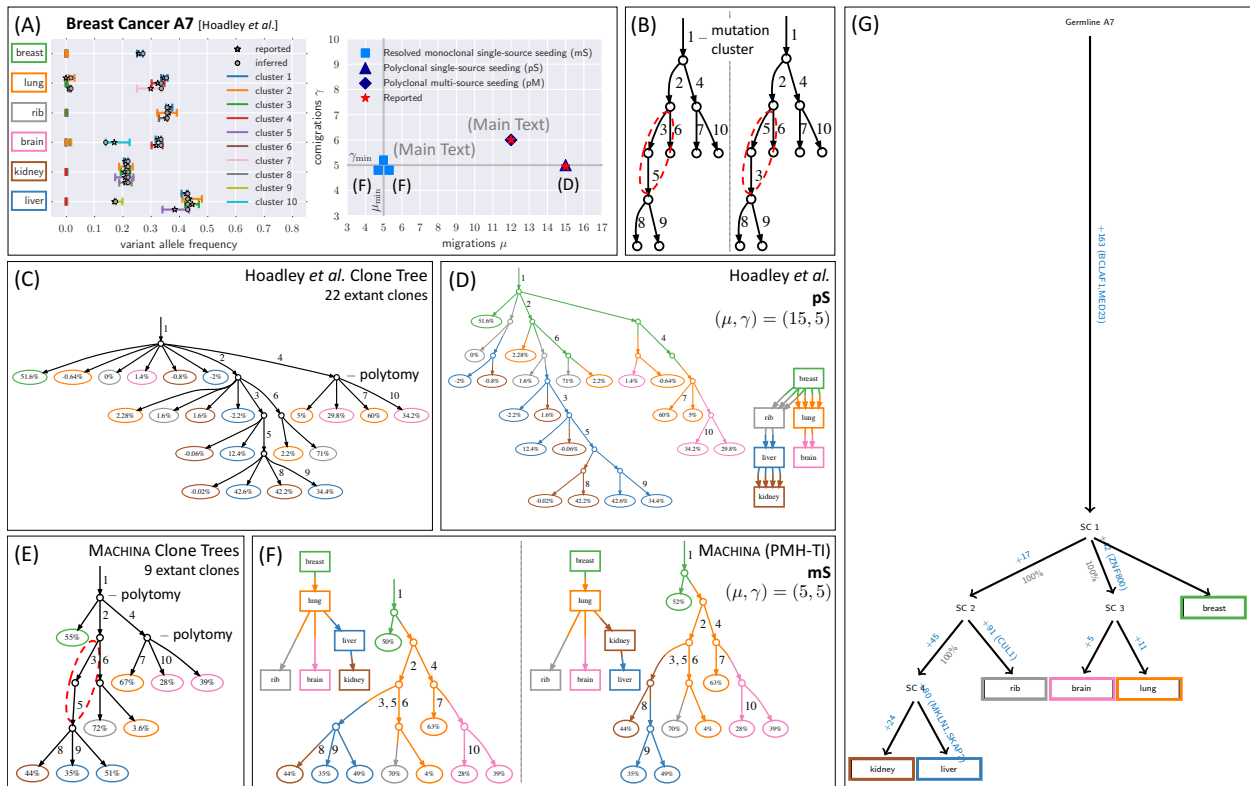


- [43] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. dd-Clone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome biology*, 18(1):44, March 2017.
- [44] J Zachary Sanborn, Jongsuk Chung, Elizabeth Purdom, Nicholas J Wang, Hojabr Kakavand, James S Wilmott, Timothy Butler, John F Thompson, Graham J Mann, Lauren E Haydu, Robyn P M Saw, Klaus J Busam, Roger S Lo, Eric a Collisson, Joe S Hur, Paul T Spellman, James E Cleaver, Joe W Gray, Nam Huh, Rajmohan Murali, Richard A Scolyer, Boris C Bastian, and Raymond J Cho. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, 112(35):10995–11000, September 2015.
- [45] David Sankoff. Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, January 1975.
- [46] K.P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.
- [47] Roland F Schwarz, Charlotte K Y Ng, Susanna L Cooke, Scott Newman, Jillian Temple, Anna M Piskorz, Davina Gale, Karen Sayal, Muhammed Murtaza, Peter J Baldwin, Nitzan Rosenfeld, Helena M Earl, Evis Sala, Mercedes Jimenez-Linan, Christine A Parkinson, Florian Markowitz, and James D Brenton. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLOS Medicine*, 12(2):e1001789, February 2015.
- [48] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165, 2013.
- [49] Qiang Tan, Jian Cui, Jia Huang, Zhengping Ding, Hao Lin, Xiaomin Niu, Zhiming Li, Guan Wang, Qingquan Luo, and Shun Lu. Genomic Alteration During Metastasis of Lung Adenocarcinoma. *Cellular Physiology and Biochemistry*, 38(2):469–486, 2016.
- [50] Mathilde B H Thomsen, Iver Nordentoft, Philippe Lamy, Søren Høyer, Søren Vang, Jakob Hedegaard, Michael Borre, Jørgen B Jensen, Torben F Ørntoft, and Lars Dyrskjøt. Spatial and temporal clonal evolution during development of metastatic urothelial carcinoma. *Molecular Oncology*, 10(9):1450–1460, November 2016.
- [51] Ruidong Xue, Ruoyan Li, Hua Guo, Lin Guo, Zhe Su, Xiaohui Ni, Lisha Qi, Ti Zhang, Qiang Li, Zemin Zhang, Xiaoliang Sunney Xie, Fan Bai, and Ning Zhang. Variable Intra-Tumor Genomic Heterogeneity of Multiple Lesions in Patients With Hepatocellular Carcinoma. *Gastroenterology*, 150(4):998–1008, April 2016.
- [52] Ke Yuan, Thomas Sakoparnig, Florian Markowitz, and Niko Beerenwinkel. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):1, 2015.
- [53] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C Anthony Blau, and William Stafford Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol*, 10(7):e1003703, Jul 2014.

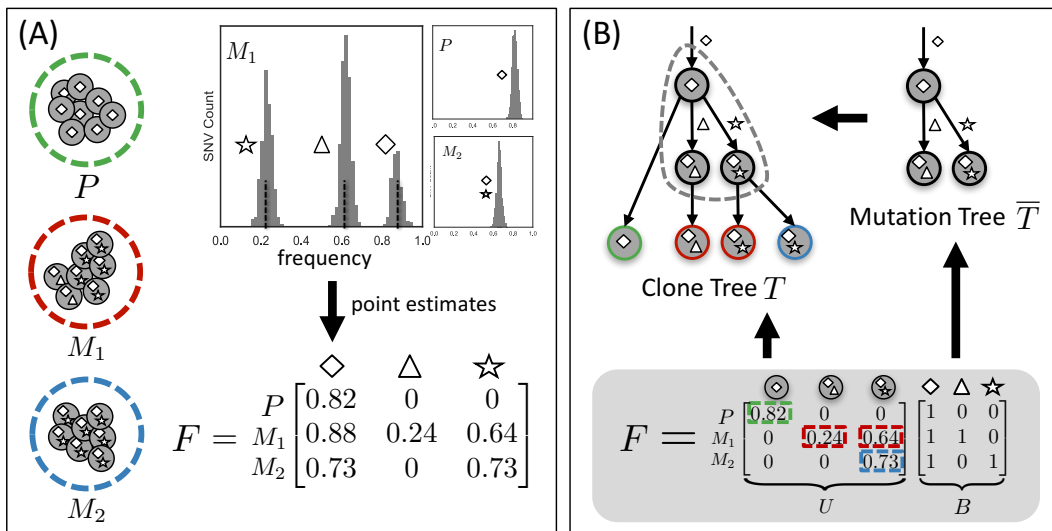
- [54] Weiwei Zhai, Tony Kiat-Hon Lim, Tong Zhang, Su-Ting Phang, Zenia Tiang, Peiyong Guan, Ming-Hwee Ng, Jia Qi Lim, Fei Yao, Zheng Li, Poh Yong Ng, Jie Yan, Brian K Goh, Alexander Yaw-Fui Chung, Su-Pin Choo, Chiea Chuen Khor, Wendy Wei-Jia Soon, Ken Wing-Kin Sung, Roger Sik-Yin Foo, and Pierce Kah-Hoe Chow. The spatial organization of intra-tumour heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma. *Nature communications*, 8:4565, February 2017.
- [55] Zi-Ming Zhao, Bixiao Zhao, Yalai Bai, Atila Iamarino, Stephen G Gaffney, Joseph Schlessinger, Richard P Lifton, David L Rimm, and Jeffrey P Townsend. Early and multiple origins of metastatic lineages within primary tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 113(8):2140–2145, February 2016.



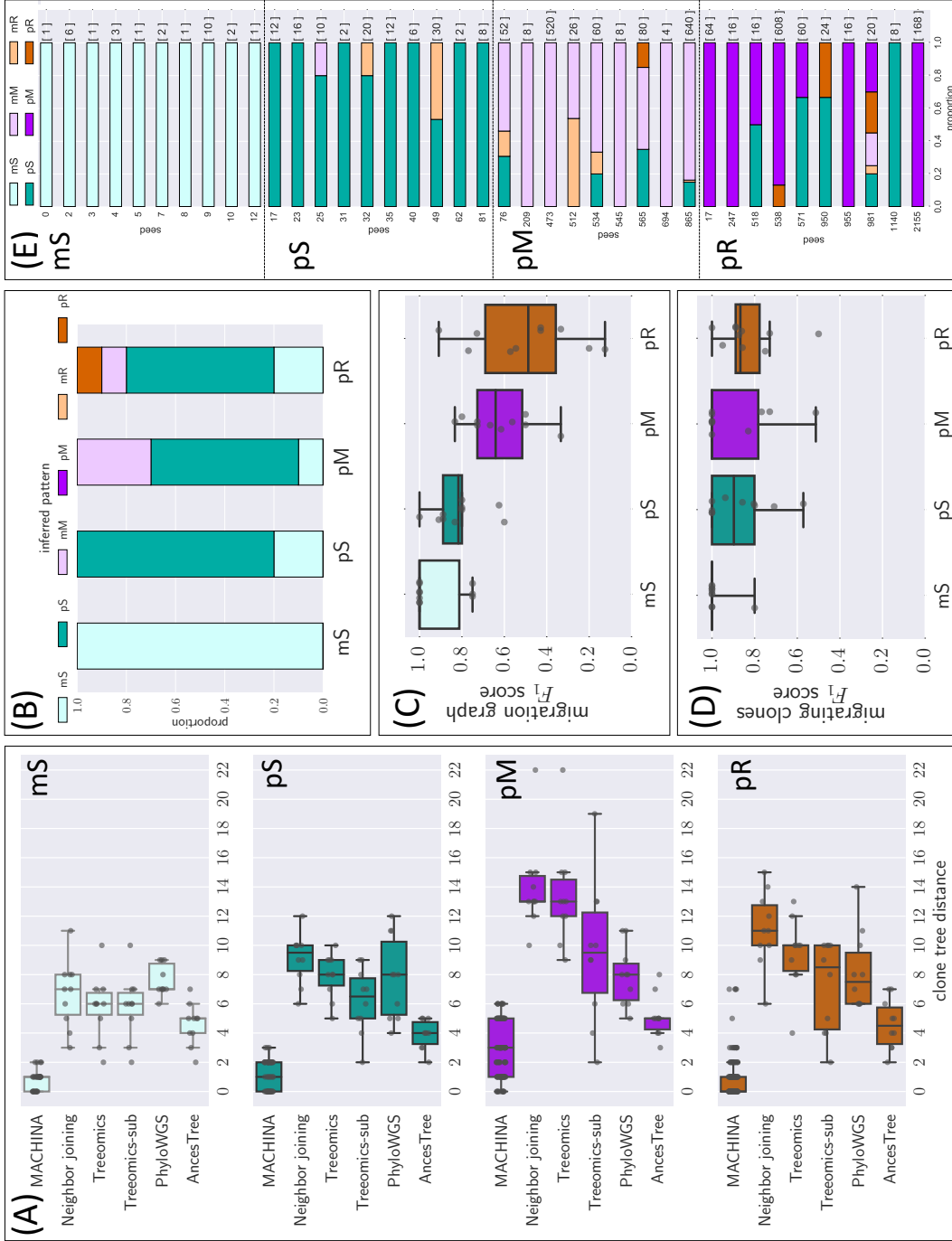
Supplementary Figure 1: **Not accounting for intra-tumor heterogeneity within regions leads to an unlikely clone tree with extensive homoplasy in colorectal patient 2 [23].** (A) The published phylogenetic tree of colorectal cancer patient CRC2 was inferred by maximum parsimony [23]. Of the 412 sequenced single-nucleotide variants (SNVs), 41 exhibit homoplasy, occurring independently on multiple branches of the tree, as indicated by the red numbers on the branches (number of introduced mutations in black). A subset of 18 SNVs have undergone homoplasy and occur in P1, M1, M2 and P3 (indicated by '\*'). (B) For each region P1, M1, M2 and P3, we show the variant allele frequency (VAF) distribution of the 18 SNVs with homoplasy (light green), and the VAF distribution of the remaining SNVs that are present in the region (dark green). In region P3, the homoplasy SNVs have significantly lower VAFs than the other SNVs in the region (Wilcoxon rank-sum p-value:  $2.4 \cdot 10^{-5}$ ), indicating that P3 is not homogeneous and contains a subclone composed of the homoplasy SNVs. On the other hand, in regions P1, M1 and M2, the homoplasy SNVs do not have significantly different VAFs than the other SNVs in each region, possibly indicating that they are clonal in these regions.



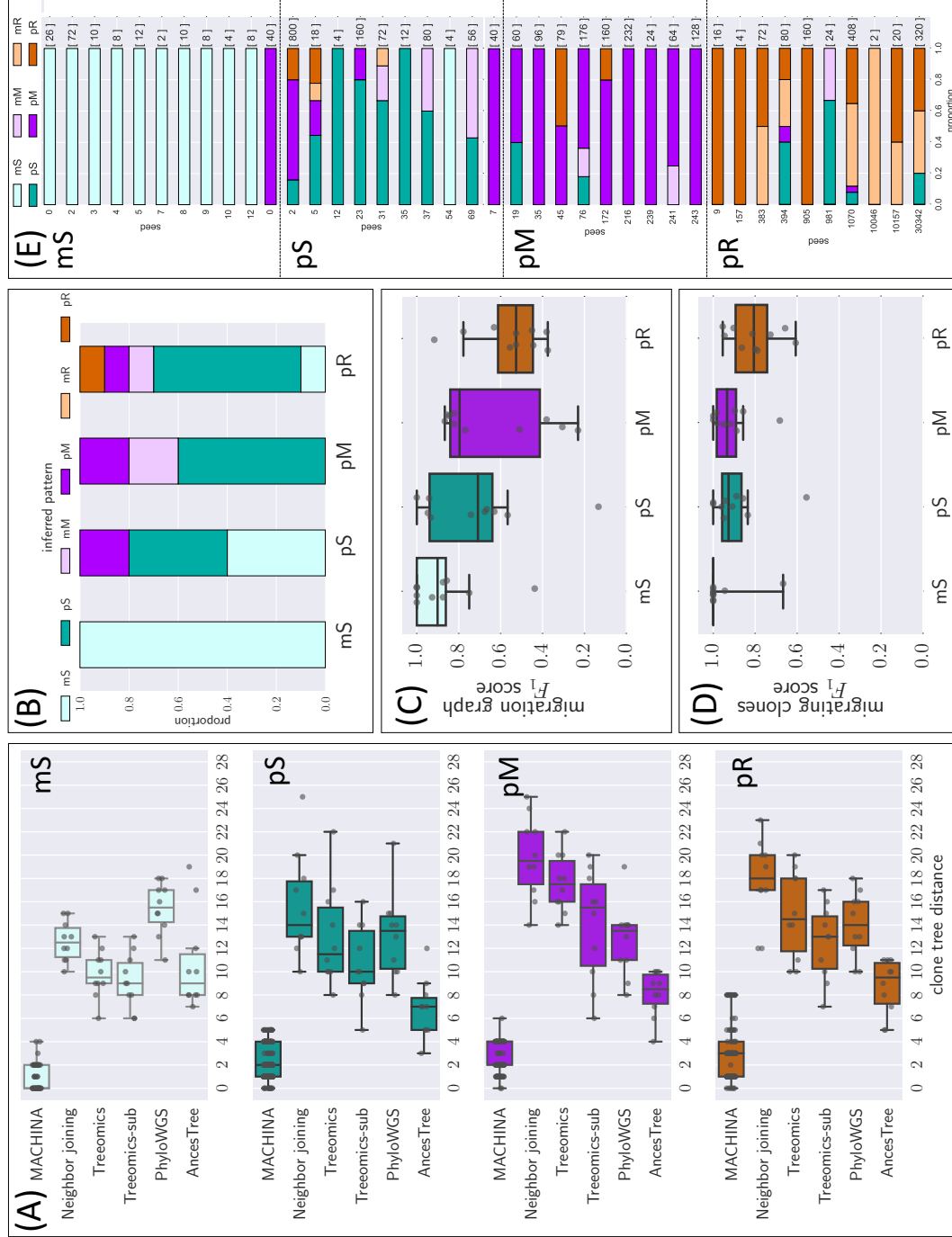
Supplementary Figure 2: **MACHINA** infers a parsimonious monoclonal single-source seeding history for breast cancer patient A7 [19]. (A) Patient A7 is composed of  $m = 6$  anatomical sites. We show 99.92% confidence intervals on the SciClone posterior distribution. (B) In the mutation tree reported by the authors cluster 3 precedes cluster 5. Using SPRUCE [9], we find an alternative mutation tree where cluster 5 precedes cluster 3. (C) Hoadley et al. [19] report a clone tree with 22 extant clones based on the leftmost mutation tree shown in (B). Viewing these clones in light of the reported VAFs (indicated by ‘ $\star$ ’ in (A)) shows that many are unsupported and have negative proportions in their respective anatomical sites. (D) Subsequent analysis by Hoadley et al. [19] yielded a polyclonal single-source seeding (pS) migration pattern with migration number  $\mu = 15$  and comigration number  $\gamma_{\min} = 5$ . (E) There exists a frequency matrix  $\hat{F}$  (indicated by ‘ $\circ$ ’ in (A)) that results in two clone trees with only 9 extant clones. We show only one clone tree; in the other clone tree, the order of the edges labeled by mutation clusters 3 and 5 is swapped. (F) We find more parsimonious migration patterns with  $\mu_{\min} = 5$  and  $\gamma_{\min} = 5$ . In addition to ambiguity in the order of mutation clusters 3 and 5, we find that there is ambiguity in the seeding order between kidney and liver, relative to the lung metastasis. (G) Treeomics does not identify the two subclones that MACHINA detected in the liver and brain.



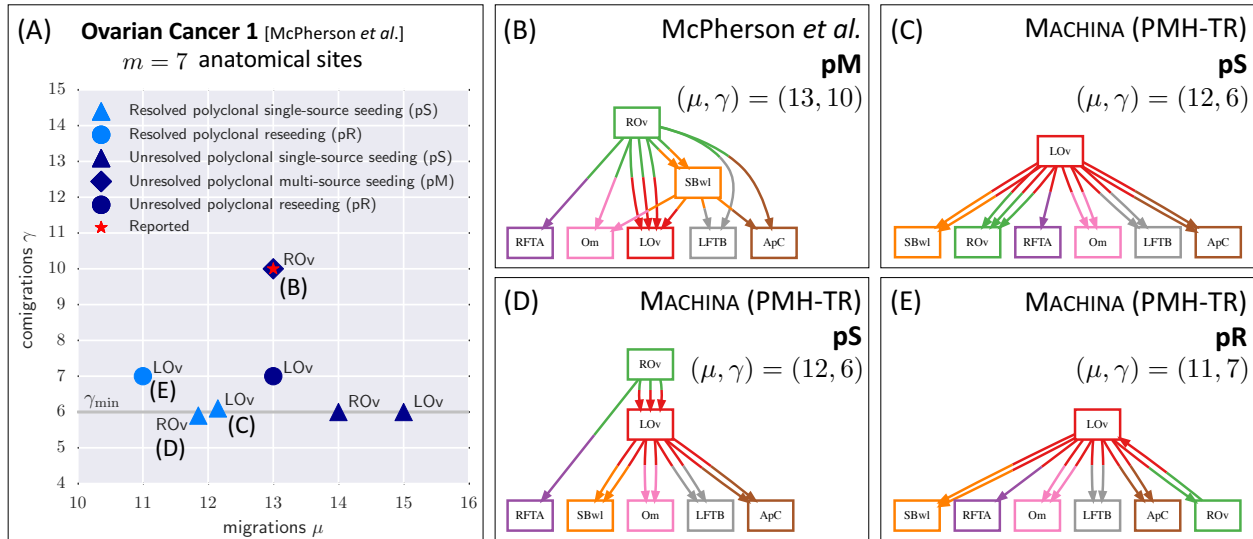
Supplementary Figure 3: **Mutation clusters do not necessarily correspond to extant clones.** (A) A first step of analysis in many studies is clustering mutations that appear at similar frequency across all samples. Here, sequencing sample  $M_1$  shows three apparent mutation clusters. (B) The samples are composed of two extant clones, one containing the diamond set and triangle set of mutations, and one containing the diamond and star set of mutations. Although the diamond cluster is present in high frequency in all samples, there is no clone in a sample containing just the diamond set of mutations. This example illustrates that mutation clusters and clones are distinct concepts.



Supplementary Figure 4: **MACHINA accurately infers clone trees and migration histories for the  $\Sigma_{\max} = 5$  simulated instances.** (A) We compare MACHINA against neighbor joining [42], Treomics [38], PhyloWGS [7] and Ancestree [8] by computing the distance  $d(T^*, T)$  between the simulated tree  $T^*$  and each inferred tree  $T$ . We show for each simulated pattern the distribution of the distances of the clone trees inferred by each method (mS in light green, pS in dark green, pM in purple and pR in orange). (B) We show for each simulated instance the patterns of the migration graphs  $G$  inferred by MACHINA run without any topological restrictions on  $G$ . These results show that the simulated instances with more complex migration patterns (pM and pR) can often be explained by simpler migration patterns. (C) Indeed, the precision and recall of the migration graph identified by MACHINA decreases with increasing complexity of the migration pattern, as summarized by the harmonic mean  $F_1$  between these two quantities. (D) MACHINA identifies the clones that migrate to different anatomical sites with high precision and recall across all migration patterns. (E) McPherson et al. [28] use the Sankoff algorithm [45] to find a vertex labeling with minimum migration number  $\mu$ . We enumerate all minimum-migration vertex labelings given the simulated clone tree  $T^*$  and show the number of such labelings in square brackets. These results indicate that the migration number  $\mu$  does not determine the migration pattern and that more sophisticated scoring functions are required for this.

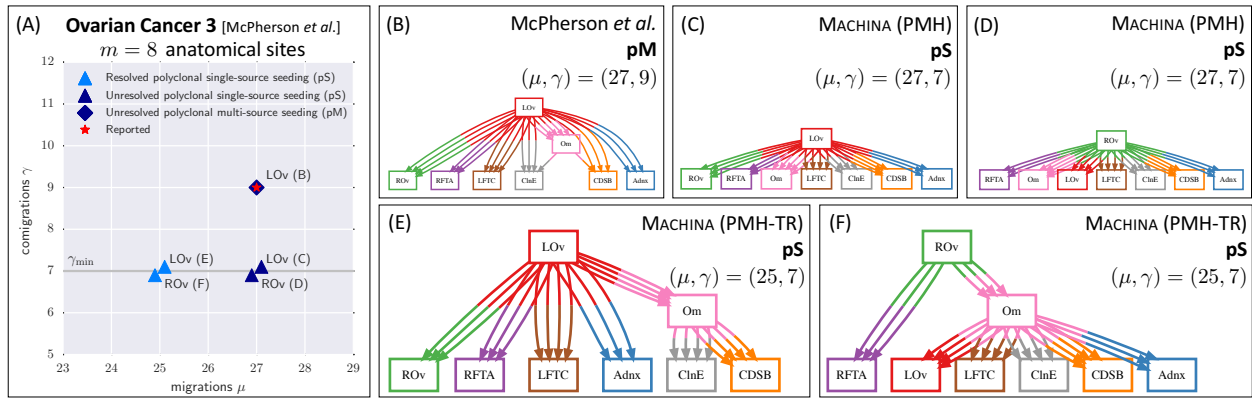


Supplementary Figure 5: **MACHINA accurately infers clone trees and migration histories for the  $\Sigma_{\max} = 8$  simulated instances.** (A) We compare MACHINA against neighbor joining [42], Treemomics [38], PhyloWGS [7] and Ancestree [8] by computing the distance  $d(T^*, T)$  between the simulated tree  $T^*$  and each inferred tree  $T$ . We show for each simulated pattern the distribution of the distances of the clone trees inferred by each method (mS in light green, pS in dark green, pM in purple and pR in orange). (B) We show for each simulated instance the patterns of the migration graphs  $G$  inferred by MACHINA run without any topological restrictions on  $G$ . These results show that the simulated instances with more complex migration patterns (pM and pR) can often be explained by simpler migration patterns. (C) Indeed, the precision and recall of the migration graph identified by MACHINA decreases with increasing complexity of the migration pattern, as summarized by the harmonic mean  $F_1$  between these two quantities. (D) MACHINA identifies the clones that migrate to different anatomical sites with high precision and recall across all migration patterns. (E) McPherson et al. [28] use the Sankoff algorithm [45] to find a vertex labeling with minimum migration number  $\mu$ . We enumerate all minimum-migration vertex labelings given the simulated clone tree  $T^*$  and show the number of such labelings in square brackets. These results indicate that the migration number  $\mu$  does not determine the migration pattern and that more sophisticated scoring functions are required for this.

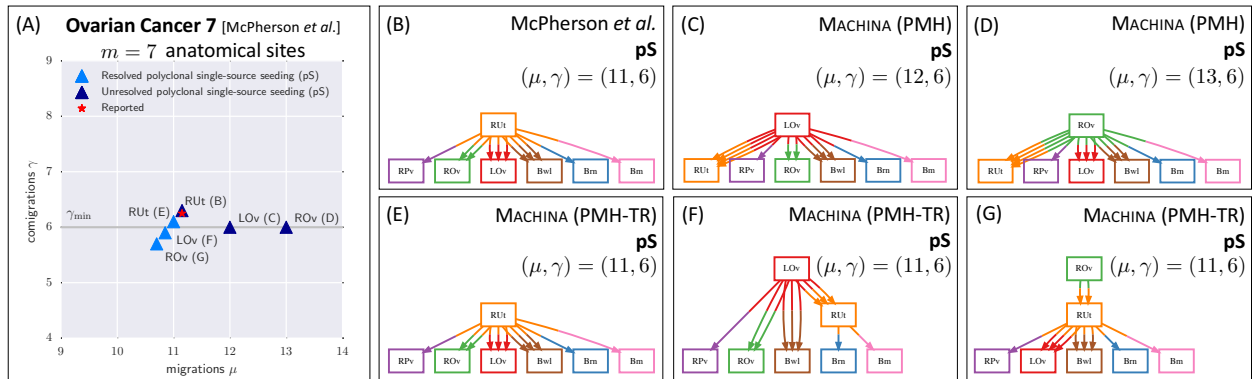


Supplementary Figure 6: **Resolving polytomies in the clone tree of ovarian cancer patient 1 leads to more parsimonious migration histories than reported in [28].** (A) Clone tree refinements under different topological constraints lead to varying migration number  $\mu$  and comigration number  $\gamma$ . Shapes indicate different migration patterns. (B) The migration graph obtained from the reported vertex labeling has comigration number  $\gamma = 10$  and corresponds to a complex migration pattern where the small bowel (SBwl) metastasis is both a destination for clones from the ROv (right ovary) primary and a source of clones for multiple anatomical sites, including LOv (left ovary) and several other metastases. (C-D) Solving the PHM-TR problem under an S constraint results in migration number  $\mu = 12$  and comigration number  $\gamma_{\min} = 6$  for both LOv and ROv as the primary. (E) With LOv as the primary, solving the unconstrained PHM-TR problem results in migration number  $\mu = 11$  and comigration number  $\gamma = 7$ , corresponding to a pR pattern.

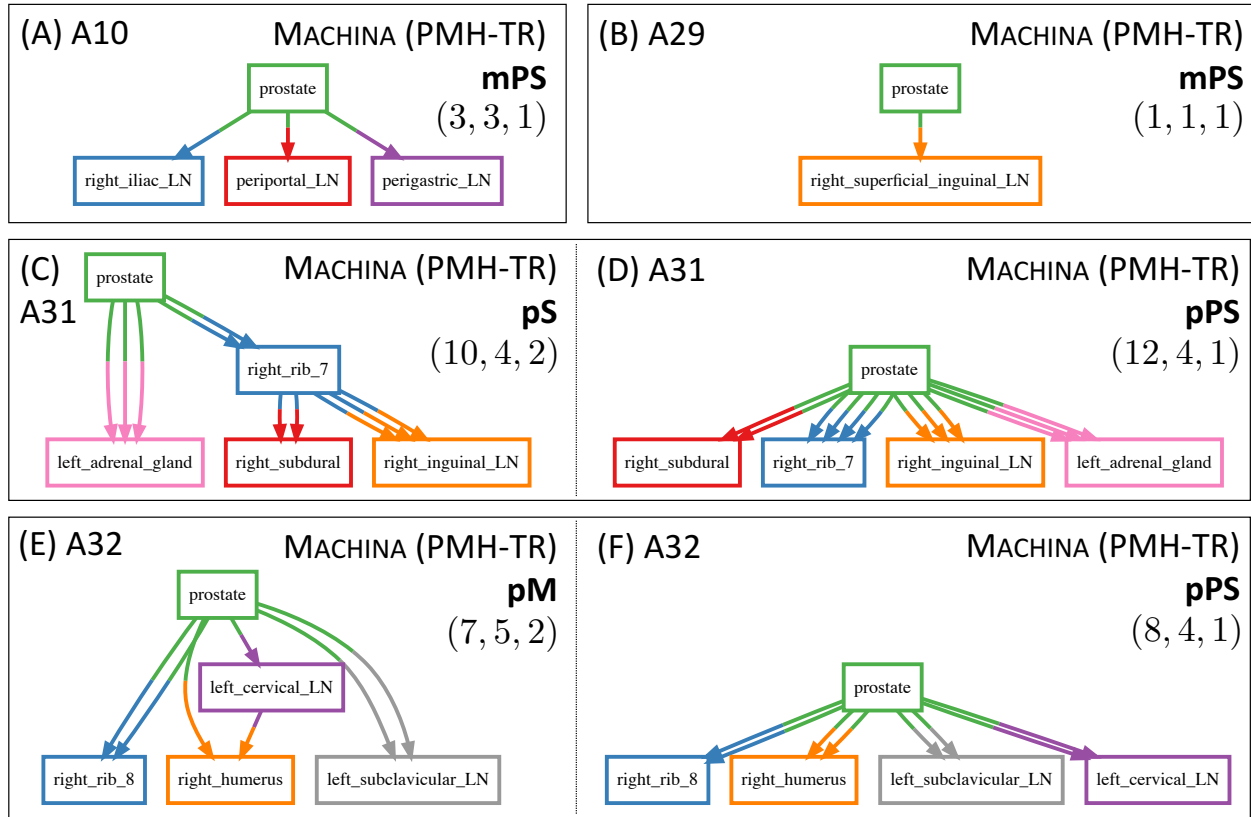




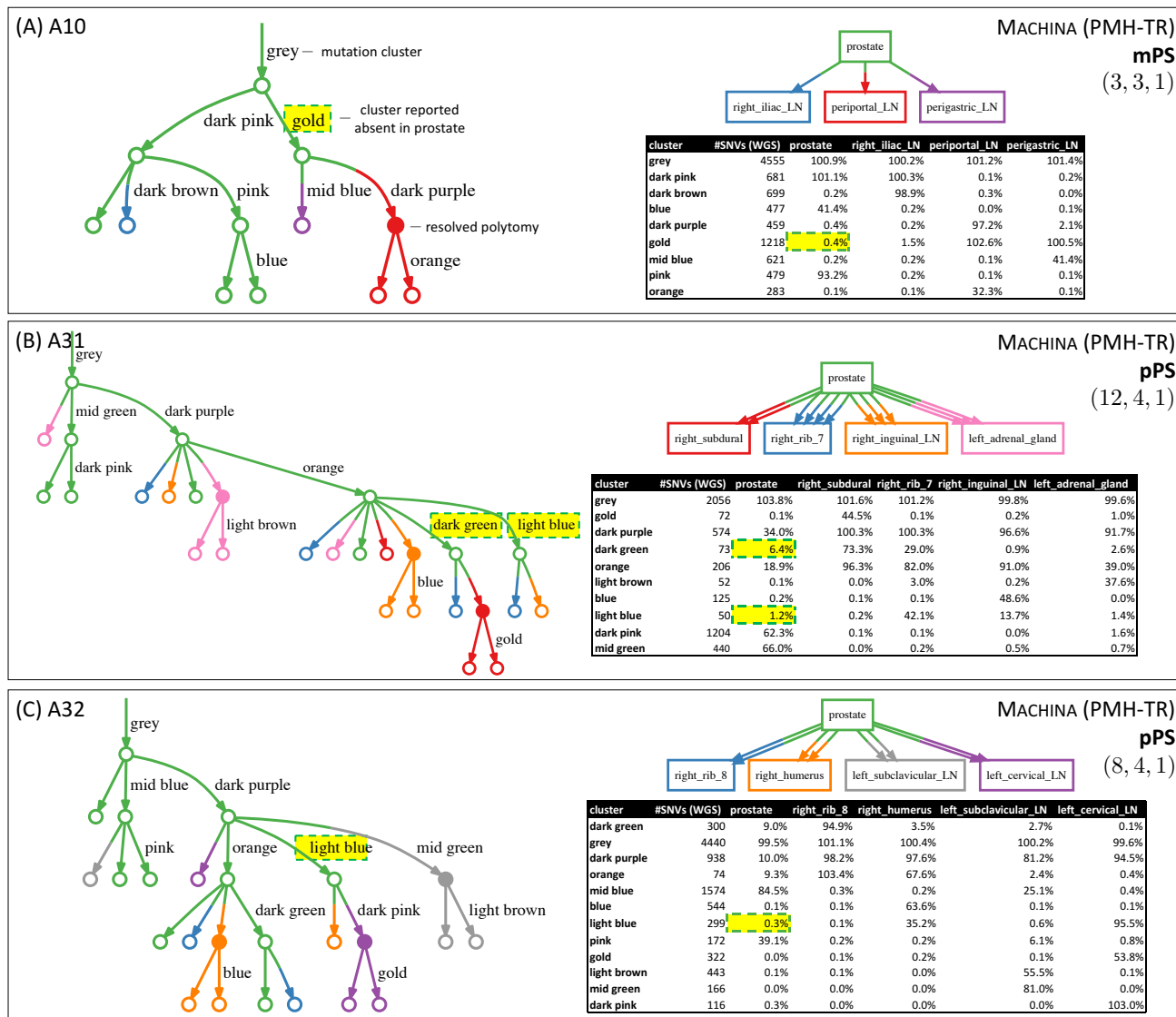
Supplementary Figure 7: **Different vertex labelings exist for ovarian cancer patient 3 with fewer migrations than reported in [28].** (A) Patient 3 has multiple vertex labelings with the minimum migration number  $\mu^* = 27$  but with different comigration number  $\gamma$ . Shapes indicate different migration patterns. (B) The reported vertex labeling has comigration number  $\gamma = 9$ , and corresponds to a polyclonal multi-source seeding (pM) pattern where both the sigmoid colon deposit (ClnE) and the cul de sac (CDSB) are seeded by clones from the left ovary and the omentum (Om). (C-D) There exist two vertex labelings with the same minimum migration number but with comigration number  $\gamma_{\min} = 7$ , where the metastases are only seeded from the left ovary (LOv) or the right ovary (ROv). (E-F) Resolving polytomies in the clone tree results in fewer migrations  $\mu = 25$  for the same comigration number  $\gamma_{\min} = 7$  with either LOv or ROv as the primary.



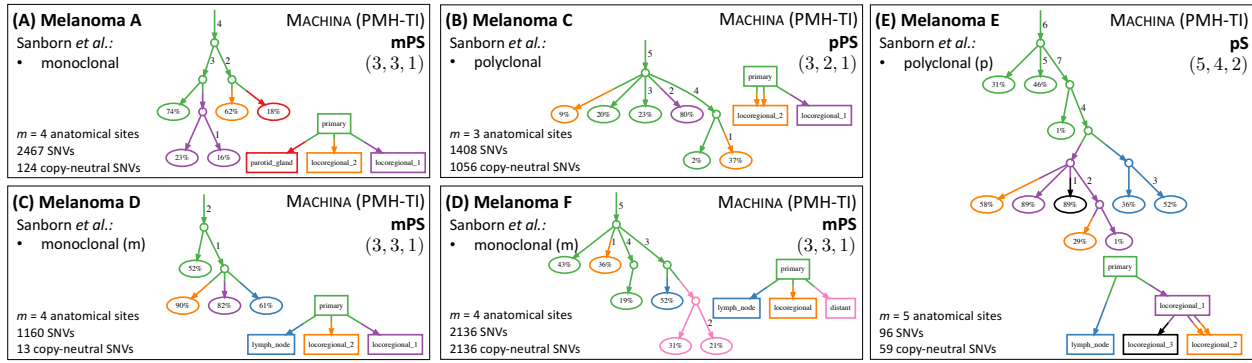
Supplementary Figure 8: **Resolving polytomies in ovarian cancer patient 7 suggests an LOv or ROv primary as opposed to the RUt primary reported in [28].** (A) Not resolving polytomies leads to varying migration numbers  $\mu$  using different primary tumors. Shapes indicate different migration patterns. (B-D) Not resolving polytomies leads to the same comigration number  $\gamma_{\min} = 6$  but different migration numbers depending on the primary tumor:  $\mu = 11$  with the right uterosacral ligament (RUt),  $\mu = 12$  with the left ovary (LOv), and  $\mu = 13$  with the right ovary (ROv). (E-G) Resolving the polytomies leads to migration number  $\mu = 11$  and comigration number  $\gamma_{\min} = 6$  in all cases.



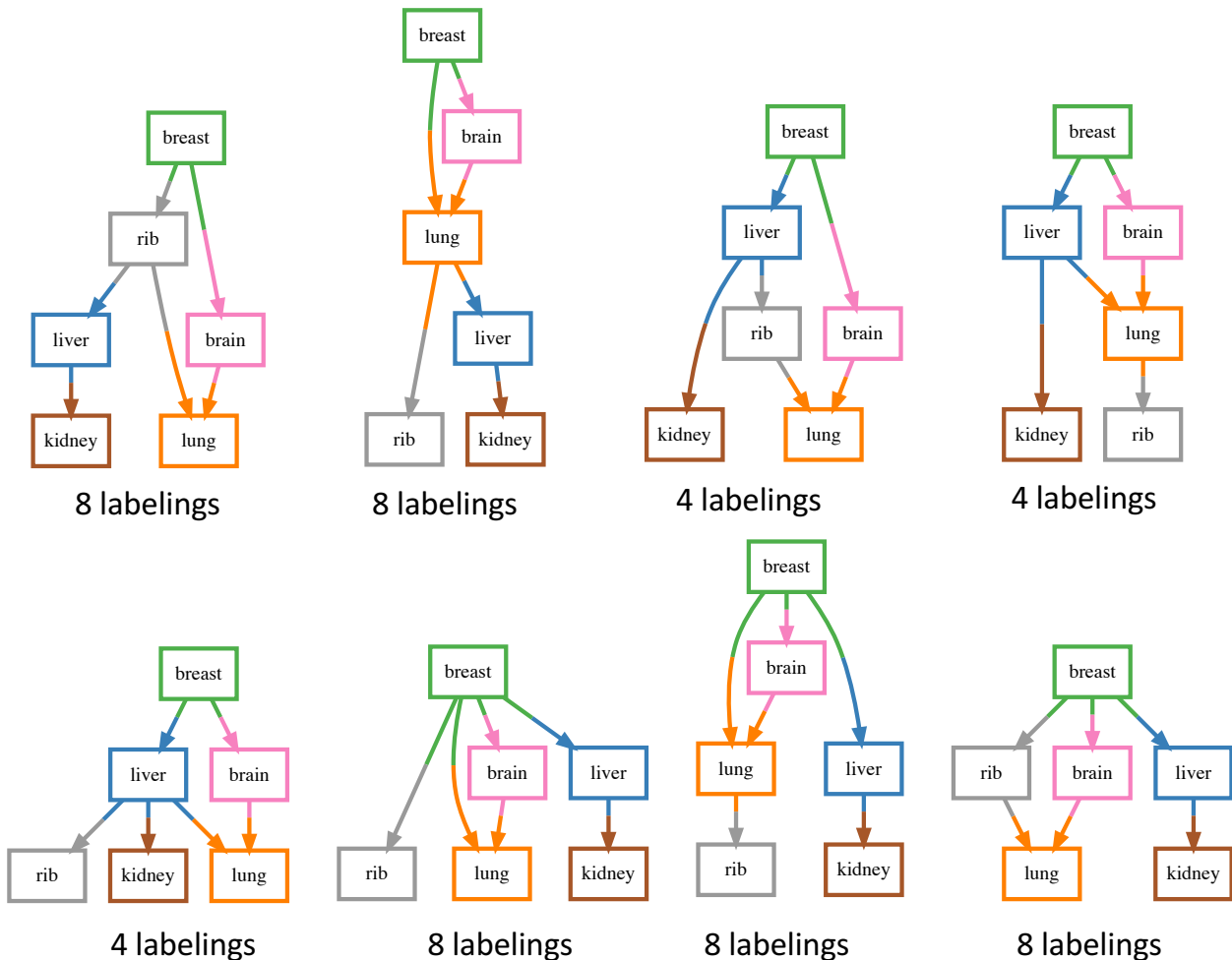
Supplementary Figure 9: **MACHINA infers parsimonious migration histories without metastasis-to-metastasis spread for four prostate cancer patients [15].** We solve the PHM-TR problem with MACHINA for each of the clone trees reported in [15]. (A) MACHINA finds a monoclonal parallel single-source seeding (mPS) migration history for patient A10, which does not support the authors’ claim of metastasis-to-metastasis spread. (B) Similarly to Gundem et al. [15], we find a migration history with an mPS migration pattern. (C-D) MACHINA finds that two more migrations are required to obtain a polyclonal parallel single-source seeding (pPS) migration history for patient A31, indicating that metastasis-to-metastasis spread may have taken place. (E-F) On the other hand, only a single additional migration is required to obtain a polyclonal parallel single-source (pPS) migration history for patient A32 instead of a complex polyclonal multi-source seeding (pM) pattern.



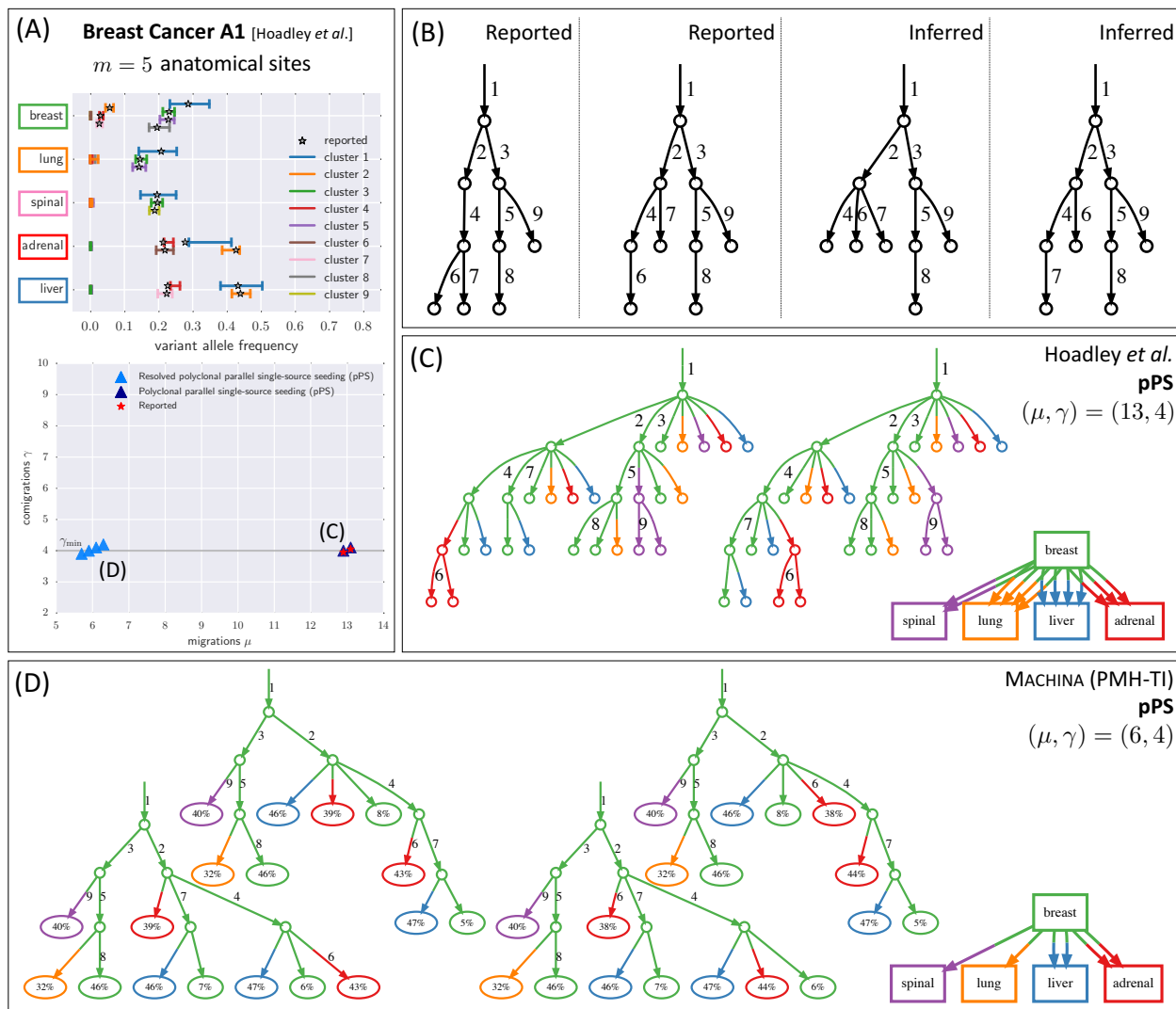
Supplementary Figure 10: **The clone trees of prostate cancers A10, A31 and A32 reported in [15] support parallel single-source seeding (PS) of all metastases from the primary tumor.** Panels show vertex labelings and resolved clone trees inferred by MACHINA (in PHM-TR) mode. The edges are labeled by mutation clusters (obtained from [15]) and filled vertices correspond to polytomies resolved by MACHINA. Tables show cancer cell fractions (CCFs) of each mutation cluster in each anatomical site. (A) The ‘gold’ mutation cluster in patient A10 has a CCF of 0.4% in the prostate. If this value is accurate then the ‘gold’ clone must have seeded both the periportal and the perigastric lymph node, ruling out metastasis-to-metastasis spread. (B) The ‘dark green’ and ‘light blue’ mutation clusters in patient A31 have a CCF in the prostate of 6.4% and 1.2%, respectively. This indicates that metastasis-to-metastasis spread is not a likely explanation. (C) Similarly to patient A10, there is a single mutation cluster whose presence in the primary prostate tumor would allow one to conclude parallel seeding (Supplementary Fig. 10C). The ‘light blue’ cluster (CCF of 0.3% in the prostate) is key to ruling out metastasis-to-metastasis spread. Note that a parallel seeding explanation for these three patients does not require the presence of the indicated mutation clusters in the primary tumor.



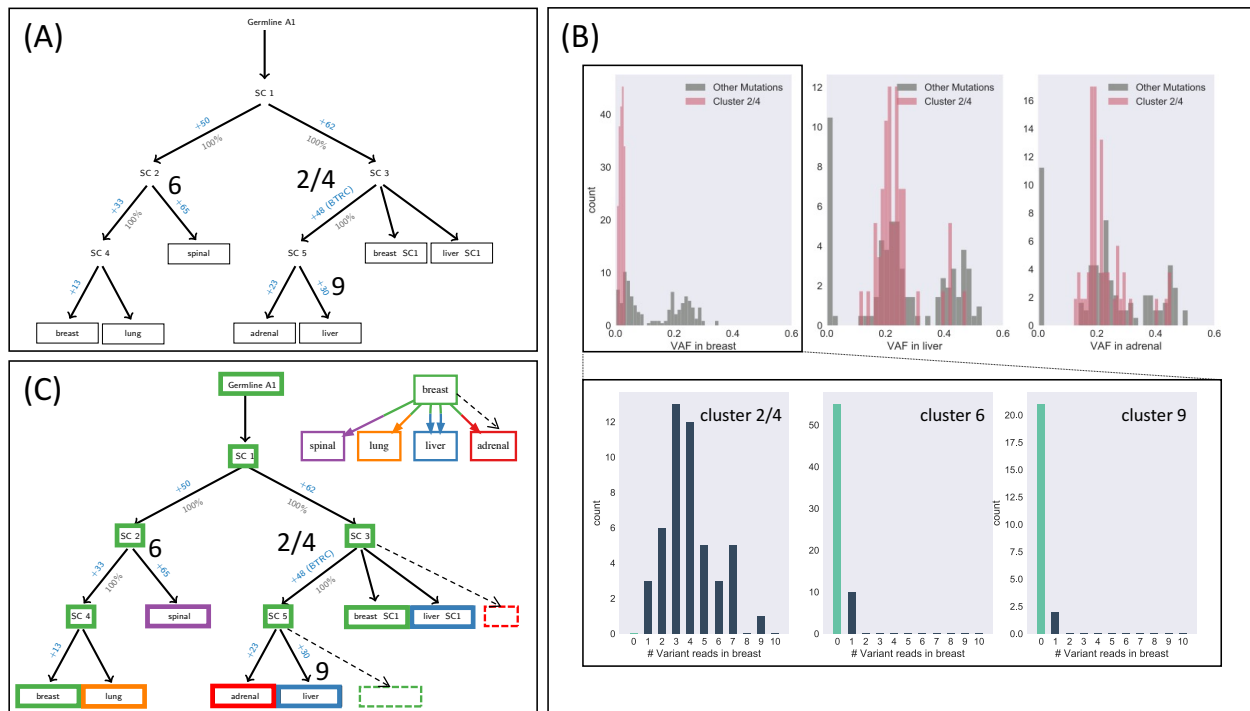
Supplementary Figure 11: **MACHINA recapitulates the cases of polyclonal seeding and non-serial progression reported in [44].** (A-E) We show the clone tree and migration graph inferred by MACHINA.



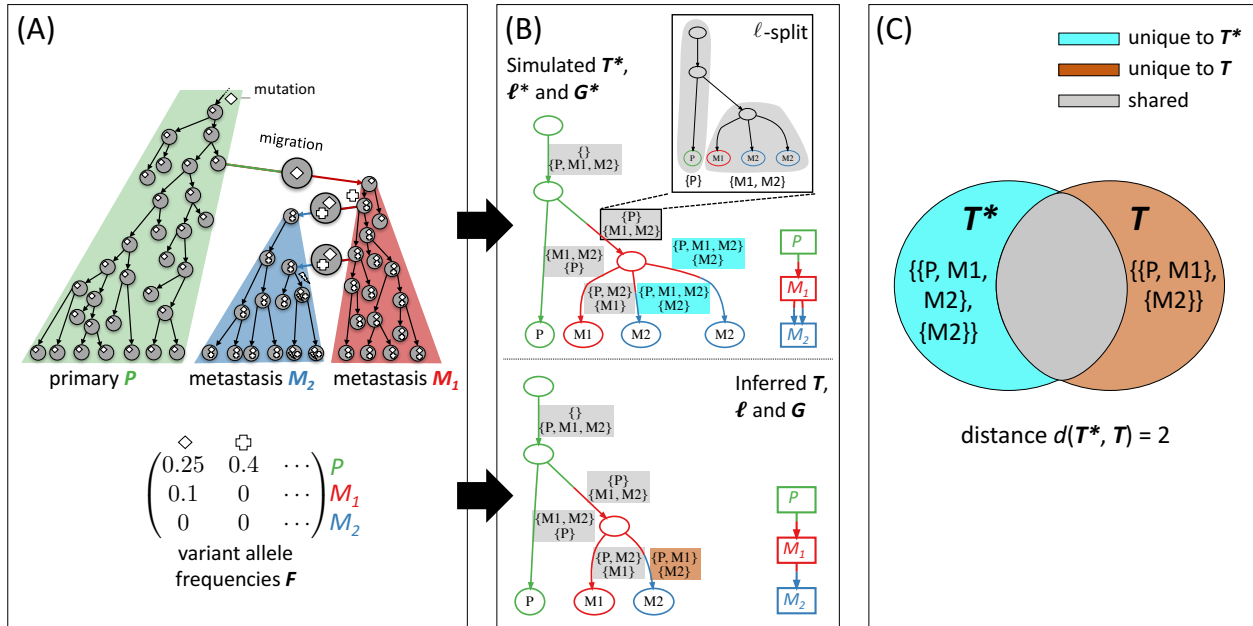
Supplementary Figure 12: **The method by McPherson et al. [28] identifies complex mM patterns for breast cancer A7.** In total, the number of minimum migration labelings in the clone tree inferred by MACHINA is 52, each having a monoclonal multi-source seeding (mM) pattern.



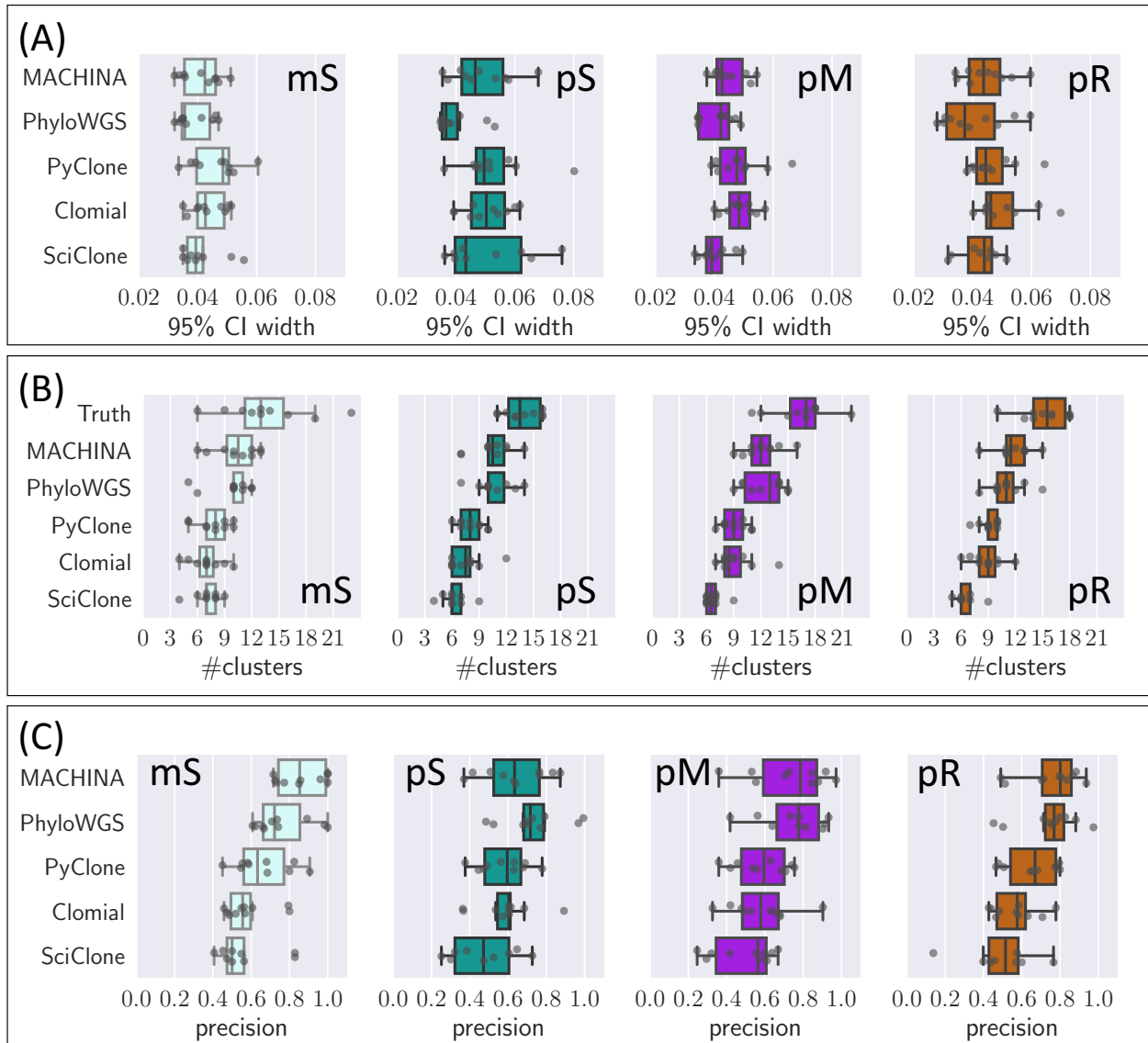
Supplementary Figure 13: **Joint analysis of mutation, cell division and migration history reveals a polyclonal parallel single-source seeding (pPS) migration history for breast cancer patient A1 [19].** (A) Patient A1 is composed of  $m = 5$  anatomical sites. We show 99.99% confidence intervals on the SciClone posterior distribution. (B) SPRUCE [9] infers four mutation trees, two of which were previously identified by the authors using ClonEvol [6]. (C) The authors infer a polyclonal parallel single-source seeding (pPS) history with migration number  $\mu = 13$  and comigration number  $\gamma_{\min} = 4$ . (D) By analyzing the four mutation trees jointly with the migration history, MACHINA finds that the liver and adrenal metastases are seeded by two clones each. The resulting migration history has migration number  $\mu = 6$  and comigration number  $\gamma_{\min} = 4$ , corresponding to a polyclonal parallel single-source seeding (pPS) history.



Supplementary Figure 14: **Treomics does not detect the additional subclone in adrenal that likely resulted from polyclonal seeding in breast cancer patient A1 [19].** (A) Treomics detects subclones in the breast and liver for patient A1 [19]. (B) The 48 mutations on the edge incoming to ‘SC 5’ correspond to mutations from clusters 2 and 4. Inspection of the variant allele frequencies of these 48 mutations reveals that they are present in breast, liver and adrenal. Although these mutations have small variant allele frequencies in the breast, all 48 mutations have at least one variant read, indicating the clusters’ likely presence in the breast, especially when compared to the variant read counts of the mutations of clusters that are absent in the breast (e.g., clusters 6 and 9 are shown here). Due to the variant allele frequency discretization step in Treomics, this cluster was not detected in the breast in the resulting tree. Moreover, the VAF distribution in adrenal indicates the presence of two distinct clusters of mutations, one containing the 48 mutations and one not. Again, the discretization step discards this important information. (C) As such, the minimum migration history of the clone tree inferred by Treomics misses the polyclonal seeding of adrenal (dashed), which MACHINA recovers.

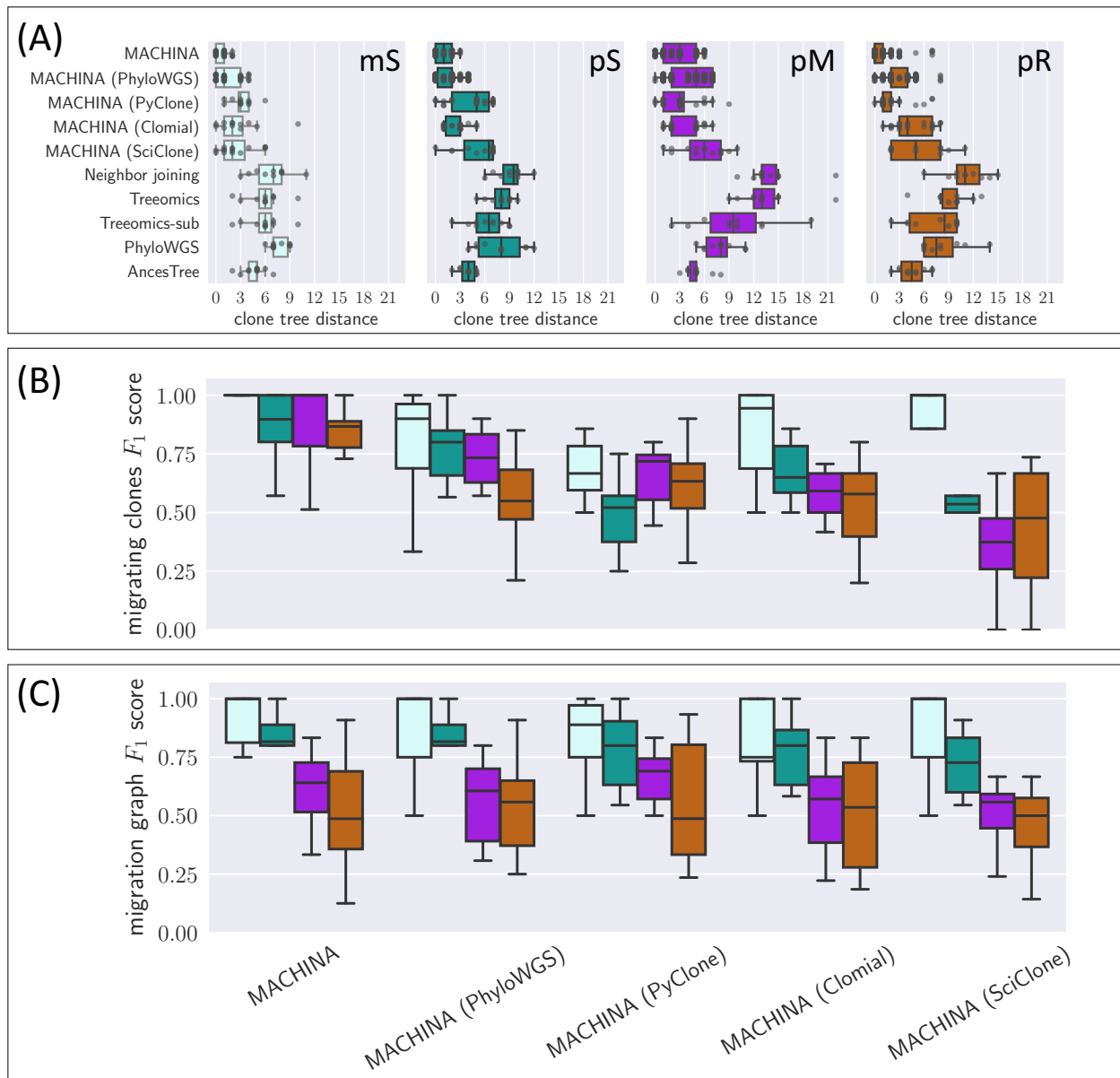


Supplementary Figure 15: **We simulate a metastatic tumor by extending an existing agent-based model [37].** (A) In this model, tumor cells accumulate mutations and migrate following different migration patterns: monoclonal and polyclonal single-source seeding (mS and pS, respectively), polyclonal multi-source seeding (pM) and polyclonal reseeding (pR). Subsequent *in silico* sequencing of the resulting tumor cells results in a simulated clone tree  $T^*$ , vertex labeling  $\ell^*$ , migration graph  $G^*$  and variant allele frequencies  $F$ . (B) To assess the similarity of a clone tree  $T$  inferred from  $F$  to  $T^*$ , we adapt the Robinson-Foulds distance. For both  $T^*$  and  $T$ , we first determine the set of  $\ell$ -splits. (C) The distance  $d(T^*, T)$  is the number of  $\ell$ -splits unique to either  $T^*$  (cyan) or  $T$  (orange).

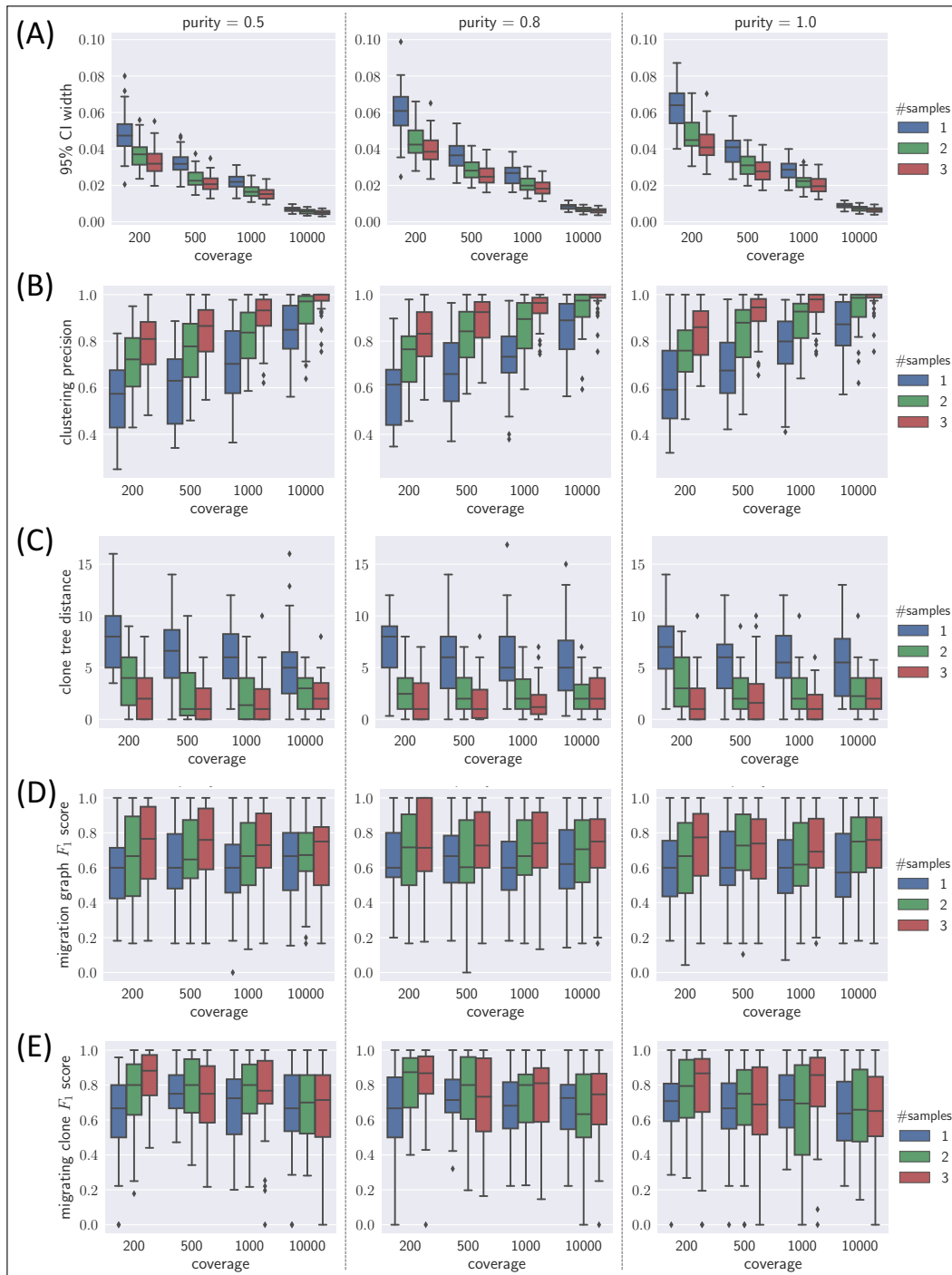


Supplementary Figure 16: **Comparison of different clustering algorithms on the  $\Sigma_{\max} = 5$  simulation instances.** We compare the mutation clusters produced by MACHINA to mutation clusters produced by PhyloWGS [7], PyClone [41], Clomial [53] and SciClone [30]. (A) We find that across different migration patterns (mS in light green, pS in dark green, pM in purple and pR in orange) the mean 95% confidence interval widths of the cluster mutation frequencies show little variation. (B) MACHINA and PhyloWGS infer far more clusters than the other methods. The simulated number of mutation clusters is indicated by ‘truth’. (C) As a result, MACHINA and PhyloWGS achieve higher clustering precision than PyClone, Clomial and SciClone.

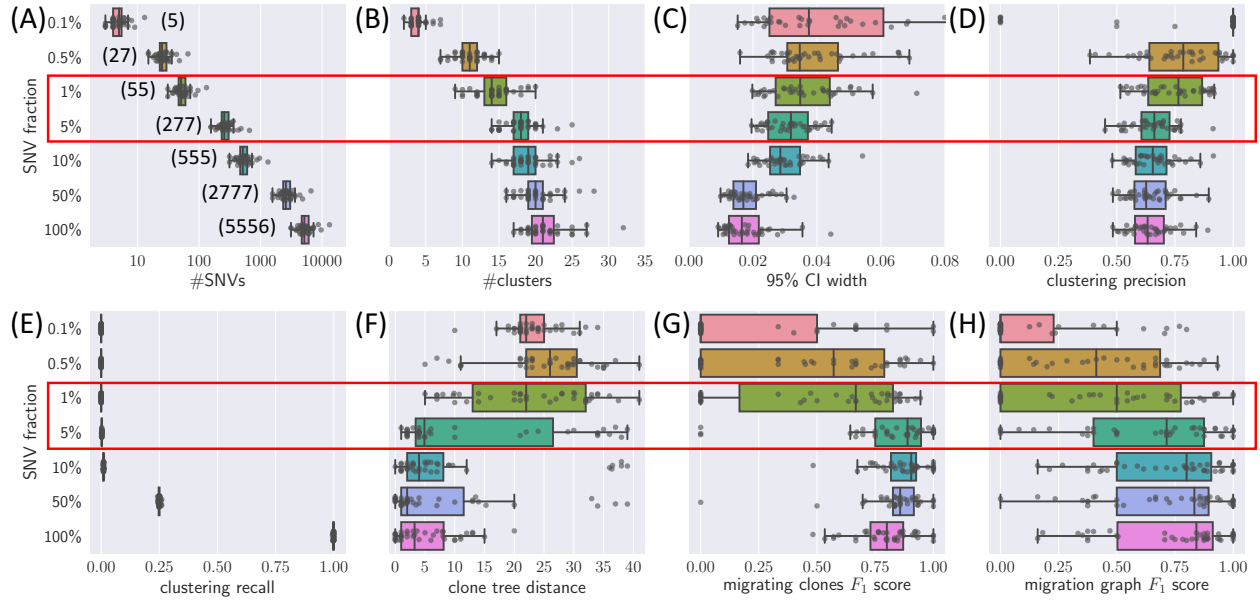




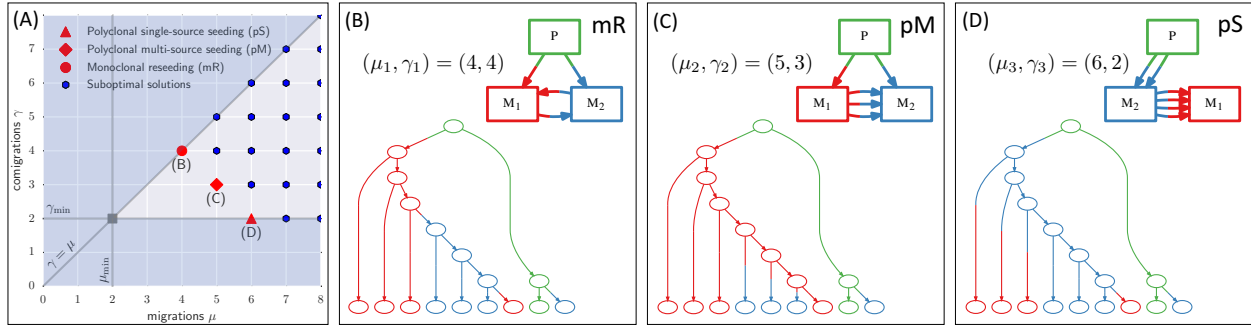
Supplementary Figure 17: **MACHINA performance with different mutation clustering algorithms.** We run MACHINA (in PMH-TI mode) on the  $\Sigma_{\max} = 5$  simulation instances given mutation clusters obtained by PhyloWGS [7], PyClone [41], Clomial [53] and SciClone [30]. (A) We show the distribution of clone tree distances  $d(T^*, T)$  between the simulated tree  $T^*$  and each inferred tree  $T$  for each simulated migration pattern (mS in light green, pS in dark green, pM in purple and pR in orange). MACHINA’s performance decreases when used in conjunction with clustering methods with lower clustering precision. However, regardless of the used clustering algorithm, MACHINA outperforms existing clone tree inference algorithms (neighbor joining [42], Treeomics [38], PhyloWGS [7] and AncesTree [8]). (B) The precision and recall of clones that migrate to different anatomical sites, as summarized by the  $F_1$  score, is affected by the used clustering algorithm and is correlated with the clustering precision of each method (Supplementary Fig. 16A). (C) Although the precision and recall of the inferred migrating clones varies across clustering methods, the inferred migration graphs are robust to these differences.



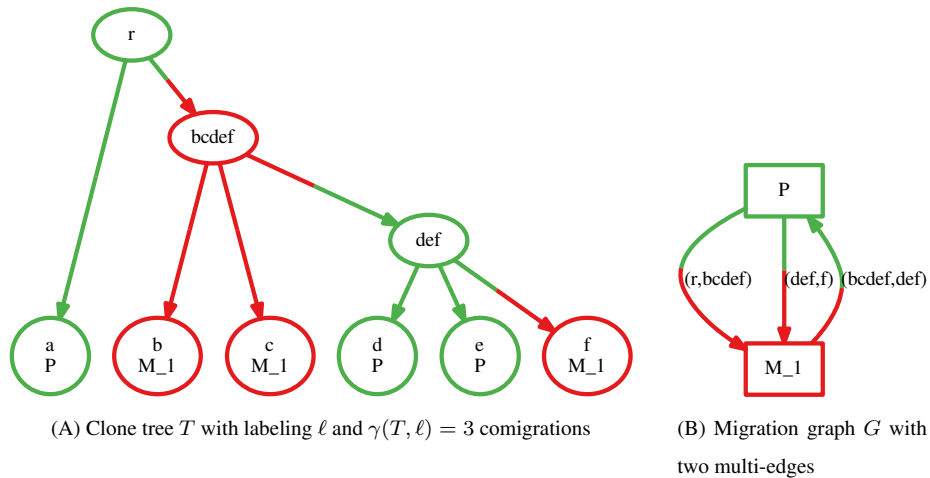
Supplementary Figure 18: **Performance of MACHINA for varying purity, sequence coverage, and number of samples.** Ten metastatic tumors for each of the four migration patterns (mS, pS, pM and pR) were simulated. From each anatomical site, simulated DNA sequencing data from three bulk samples with a purity of 1, a nucleotide sequencing error rate of 0.001 and a target coverage of 10,000x was generated. Columns correspond to purity values, colors correspond to number of samples and the  $x$ -axis corresponds to the sequencing coverage. (A) The mean 95% confidence interval widths of the cluster mutation frequencies inferred by MACHINA's clustering algorithm. (B) The clustering precision. (C) The distribution of clone tree distances  $d(T^*, T)$  between the simulated tree  $T^*$  and each inferred tree  $T$ . (D-E) The precision and recall of the migration graph (D) and clones that migrate to different anatomical sites (E), as summarized by the  $F_1$  score.



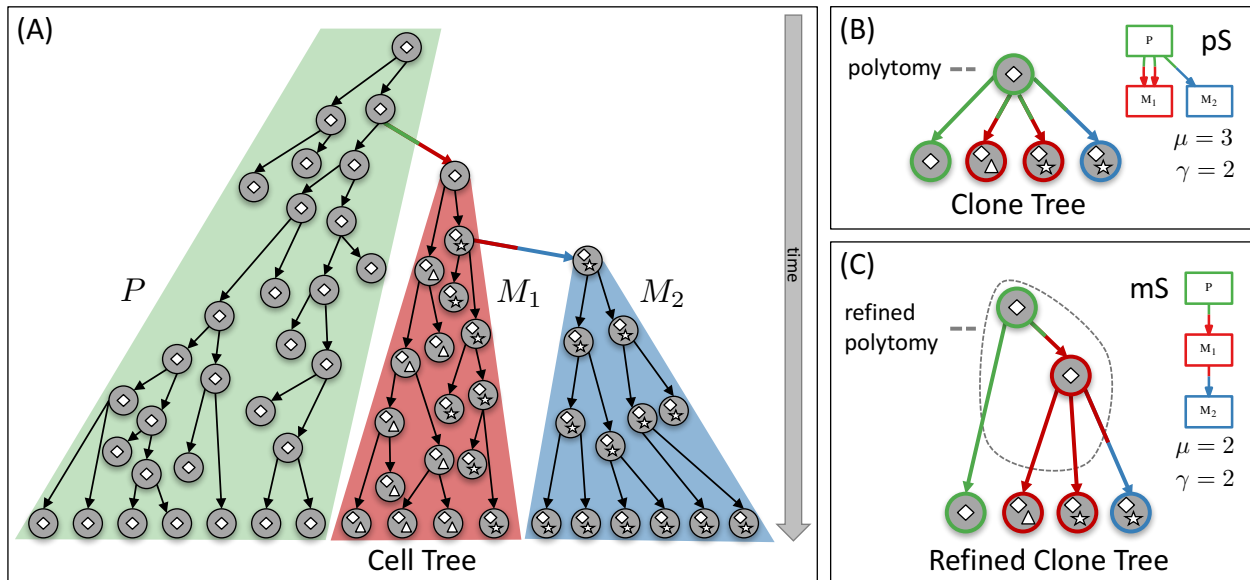
Supplementary Figure 19: **MACHINA’s performance varies with the number of mutations, achieving good performance in the regime of whole exome sequencing data (indicated in red).** For each migration pattern (mS, pS, pM and pR), we simulate 10 metastatic tumors with  $\Sigma_{\max} = 8$  anatomical sites and a rate of 10 mutations every cell division, corresponding to whole genome sequencing data. We downsample the initial number of simulated SNVs from 100% to 0.1%. (A) The plot shows the resulting number of SNVs using a log scale; the mean numbers of SNVs are shown in parentheses. Fractions of 1-5% correspond to the number of mutations observed in whole exome sequencing datasets. (B) We use the MACHINA clustering algorithm to group SNVs with similar variant allele frequencies. The number of inferred mutation clusters increases with the number of SNVs. (C) Similarly, with more SNVs the uncertainty in mutation clusters decreases, as shown by the mean width of the 95% confidence intervals of frequencies of mutation clusters. (D-E) The clustering precision and recall are affected by the number of SNVs, with the precision decreasing and recall increasing with increasing numbers of SNVs. (F) The clone tree distance  $d(T^*, T)$  between the simulated tree  $T^*$  (containing all SNVs) and each inferred tree  $T$  (containing a subset of the SNVs) decreases with increasing number of SNVs. (G-H) Similarly, the precision and recall (summarized by the  $F_1$  score) of the clones that migrate to different anatomical sites (G), and of the migration graph (H) increases with increasing numbers of SNVs. Note that in (H) we define a clone by the subset of SNVs present in the downsampled instance.



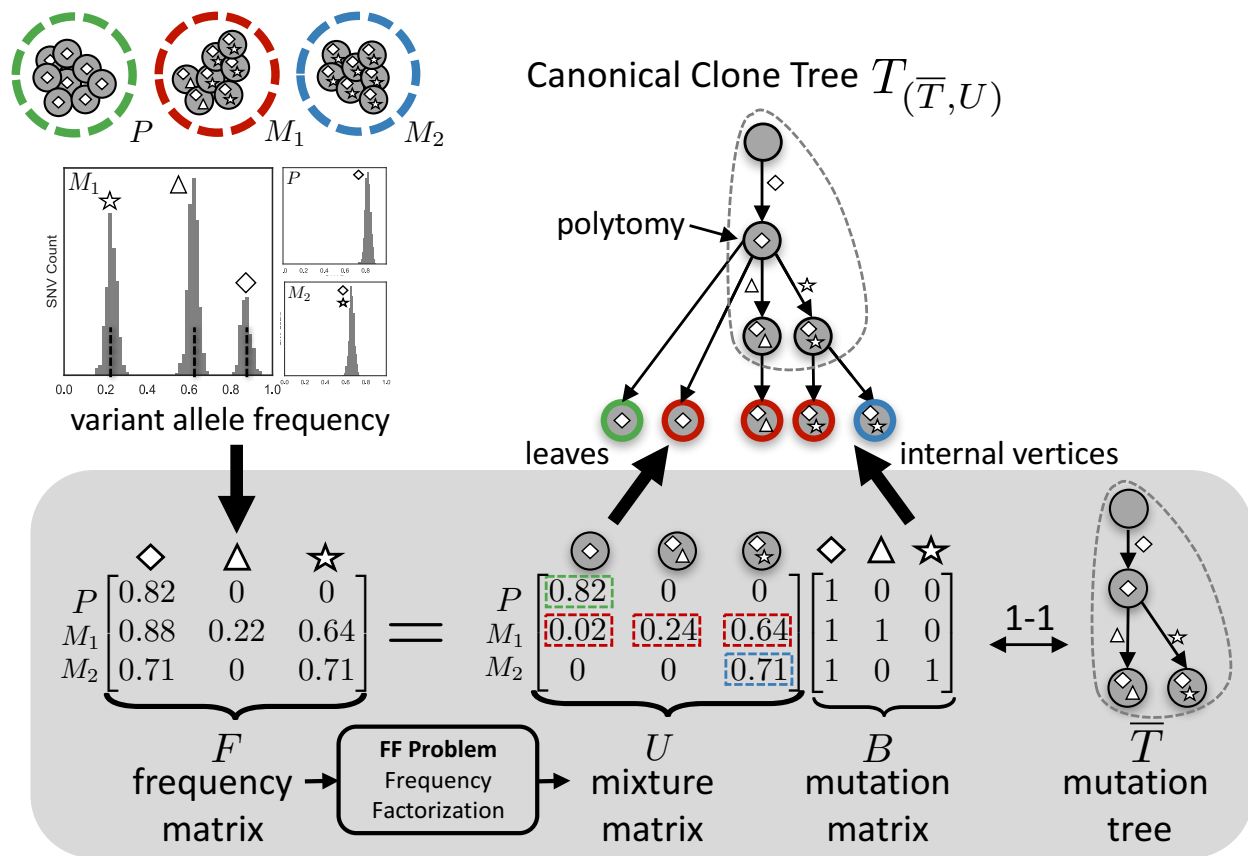
Supplementary Figure 20: **There exists a tradeoff between the migration number  $\mu$  and the comigration number  $\gamma$ .** (A) The given clone tree  $T$  with  $m = 3$  anatomical sites has three labelings on the Pareto front: (B)  $(\mu_1, \gamma_1) = (4, 4)$ , (C)  $(\mu_2, \gamma_2) = (5, 3)$  and (D)  $(\mu_3, \gamma_3) = (6, 2)$ . Thus, the minimum migration number  $\mu^*(T)$  is 4 and the minimum comigration number  $\gamma_{\min}$  is  $m - 1 = 2$ . There exists no vertex labeling with both migration number  $\mu^*(T)$  and comigration number  $\gamma_{\min}$ , thus showing that there is a tradeoff between both scores. Moreover, there exists no vertex labeling with migration number  $\mu_{\min} = m - 1$  and comigration number  $\gamma_{\min} = m - 1$  (gray box in (A)), indicating that  $T$  does not admit a mS migration history. The shaded area contains infeasible points with migration number  $\mu < \mu_{\min}$  (Observation 1),  $\mu < \gamma$  migrations (Observation 3), and comigration number  $\gamma < \gamma_{\min}$  (Observation 4).



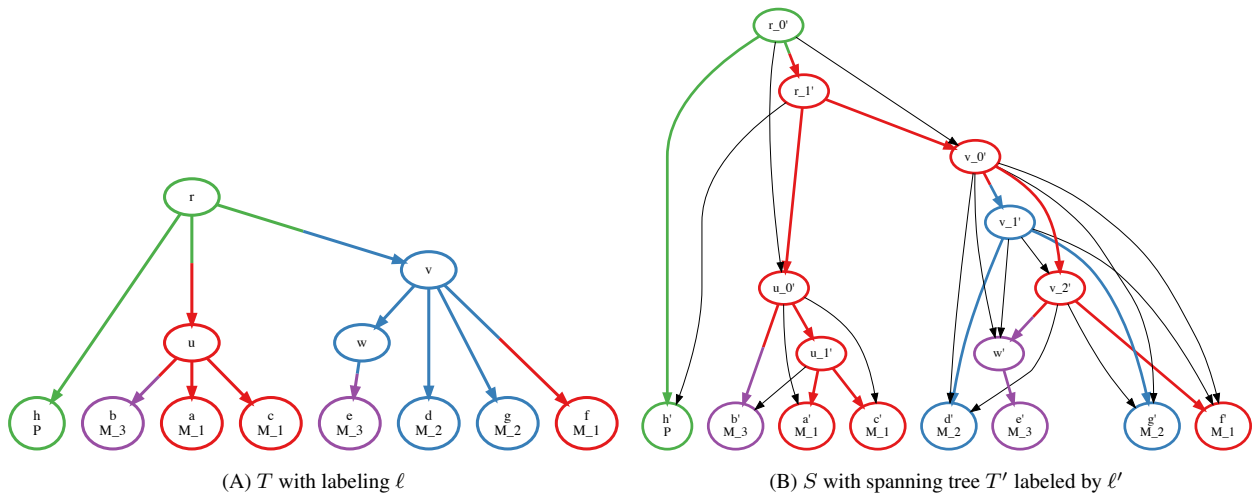
Supplementary Figure 21: **The comigration number  $\gamma(T, \ell)$  may not equal the number of multi-edges in a migration graph  $G$  with directed cycles.** (A) Vertex labeling  $\ell$  of clone tree  $T$  results in three migration edges:  $(r, bcdef)$ ,  $(bcdef, def)$  and  $(def, f)$ . (B) These migration edges result in a directed cycle in the migration graph  $G$ . Multi-graph  $G$  has two multi-edges:  $(P, M_1)$  and  $(M_1, P)$ , where multi-edge  $(P, M_1)$  has multiplicity two and corresponds to migration edges  $(r, bcdef)$  and  $(def, f)$ . Inspecting  $T$  reveals that  $(r, bcdef)$  and  $(def, f)$  occur on the same branch of  $T$  and consequently clone  $def$  must have occurred after the migration of clone  $r$  to anatomical site  $P$ . Thus, there are two separate comigrations between  $(P, M_1)$  leading to comigration number  $\gamma(T, \ell) = 3$ , which is different from the number of two multi-edges in  $G$ .



Supplementary Figure 22: **Constraints on the migration history help resolve polytomies in clone trees.** (A) A cell tree is a full binary tree that represents the cell division, mutation and migration history of a tumor. Due to lack of data, the cell tree is unknown. The leaves of the cell tree are extant cells and are labeled by the anatomical site in which they occur. (B) Instead, using specialized phylogeny inference techniques, one infers a clone tree from DNA sequencing data, which in general contains polytomies and is non-binary. Solving the PMH problem using this non-binary clone tree results in an incorrect migration history and migration pattern, where two clones migrate from  $P$  and seed  $M_1$  and a single clone migrates from  $P$  and seeds  $M_2$ , corresponding to a polyclonal single-source seeding (pS) pattern. (C) By solving the PHM-TR problem, we find a refinement of the original clone tree with a single migration from  $P$  to  $M_1$  followed by one migration from  $M_1$  to  $M_2$ , corresponding to a parsimonious monoclonal single-source seeding (mS) pattern.



Supplementary Figure 23: **With tumor bulk sequencing we do not directly observe the mutation tree, instead we observe variant allele frequencies which are mixed measurements of the leaves of an unknown clone tree.** Combining these with copy-number calls yields a frequency matrix  $F$ . By subsequently solving the frequency factorization problem, we obtain potentially many mutation matrices  $B$ , each associated with a unique mixture matrix  $U$  such that  $F = UB$ . From  $U$  and  $B$  (or equivalently  $\bar{T}$ ) we obtain the canonical clone tree  $T_{(\bar{T}, U)}$ . The shown canonical clone tree contains a polytomy, a vertex with more than two children.



Supplementary Figure 24: **All binarizations of clone tree  $T$  can be modeled as constrained spanning trees of graph  $S$ .** (A) Clone tree  $T$  admits a labeling  $\ell$  with cost  $(\mu(T, \ell), \gamma(T, \ell), \sigma(T, \ell)) = (5, 5, 3)$ . (B) The graph  $S$  contains all binarizations as spanning trees, among which  $T'$  with labeling  $\ell'$  and cost  $(\mu(T', \ell'), \gamma(T', \ell'), \sigma(T', \ell')) = (4, 3, 2)$ .

	breast	brain	kidney	liver	lung	rib
Mutation 47	0.01	0.01	0.18	0.48	0.00	0.36
Mutation 81	0.00	0.03	0.24	0.50	0.00	0.29
Cluster 2	0.00	0.01	0.21	0.44	0.02	0.36
Cluster 3	0.00	0.01	0.21	0.45	0.00	0.00

Supplementary Table 1: **Reported clustering [19] of mutations 47 and 81 in patient A7 is likely incorrect.** The table shows variant allele frequencies (VAFs) for mutations 47 (chromosome 7, position 12163423) and 81 (chromosome 7, position 57562948), as well as cluster mean VAFs for clusters 2 and 3. Mutations 47 and 81 were reported by Hoadley et al. [19] as belonging to cluster 3. The two mutations are more consistent with cluster 2 due to their high VAFs in the rib sample (red box).



seed	#anatomical sites $m$	#mut. trees $ \overline{\mathcal{T}} $	simulated		MACHINA (S)		MACHINA (S, M)		MACHINA (S, M, R)	
			pattern	$(\mu, \delta)$	pattern	$(\mu, \delta)$	pattern	$(\mu, \delta)$	pattern	$(\mu, \delta)$
0	5	2	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
2	5	4	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
3	6	4	mS	(5, 5)	mS	(5, 5)	mS	(5, 5)	mS	(5, 5)
4	5	3	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
5	5	1	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
7	5	2	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
8	5	12	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
9	5	1	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
10	5	2	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
12	5	2	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
17	5	1	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
23	5	2	pS	(7, 4)	pS	(7, 4)	pS	(7, 4)	pS	(7, 4)
25	5	4	pS	(6, 4)	pS	(6, 4)	pS	(6, 4)	pS	(6, 4)
31	5	36	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
32	5	4	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
35	5	1	pS	(5, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
40	6	12	pS	(6, 5)	pS	(6, 5)	pS	(6, 5)	pS	(6, 5)
49	5	4	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
62	5	4	pS	(5, 4)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
81	5	15	pS	(6, 4)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
76	5	4	pM	(6, 5)	pS	(6, 4)	pS	(6, 4)	pS	(6, 4)
209	5	12	pM	(7, 5)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
473	7	80	pM	(10, 8)	pS	(10, 6)	pS	(10, 6)	pS	(10, 6)
512	5	12	pM	(6, 5)	pS	(6, 4)	mM	(5, 5)	mM	(5, 5)
534	5	24	pM	(6, 5)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
545	6	8	pM	(7, 6)	mS	(5, 5)	mS	(5, 5)	mS	(5, 5)
565	5	4	pM	(6, 5)	pS	(6, 4)	pS	(6, 4)	pS	(6, 4)
694	5	30	pM	(6, 5)	pS	(6, 4)	pS	(6, 4)	pS	(6, 4)
865	5	8	pM	(7, 5)	pS	(7, 4)	mM	(5, 5)	mM	(5, 5)
907	6	2	pM	(7, 6)	pS	(8, 5)	mM	(6, 6)	mM	(6, 6)
17	6	48	pR	(9, 6)	pS	(9, 5)	pS	(9, 5)	pS	(9, 5)
247	5	78	pR	(7, 5)	pS	(7, 4)	pS	(7, 4)	pS	(7, 4)
518	5	2	pR	(6, 5)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
538	5	4	pR	(8, 5)	pS	(7, 4)	mM	(6, 6)	mM	(6, 6)
571	6	2	pR	(7, 6)	pS	(6, 5)	pS	(6, 5)	pS	(6, 5)
950	7	8	pR	(8, 7)	mS	(6, 6)	mS	(6, 6)	mS	(6, 6)
955	5	8	pR	(6, 5)	pS	(6, 4)	pS	(6, 4)	pS	(6, 4)
981	5	4	pR	(6, 5)	mS	(4, 4)	mS	(4, 4)	mS	(4, 4)
1140	5	12	pR	(6, 5)	pS	(5, 4)	pS	(5, 4)	pS	(5, 4)
2155	7	120	pR	(9, 8)	pS	(9, 6)	pS	(9, 6)	pR	(8, 7)

Supplementary Table 2: **Simulated instances and MACHINA results with  $\Sigma_{\max} = 5$  anatomical sites.** For each simulated instance, we show the used random number generator seed, the simulated migration pattern, the number  $m$  of resulting anatomical sites, the simulated migration and comigration numbers  $(\mu, \gamma)$  and the number  $|\overline{\mathcal{T}}|$  of mutation trees enumerated by SPRUCE [9]. Next, we show the migration pattern and the migration and comigration numbers  $(\mu, \gamma)$  inferred by MACHINA under various topological constraints.

seed	#anatomical sites $m$	#mut. trees $ \overline{\mathcal{T}} $	simulated		MACHINA (S)		MACHINA (S, M)		MACHINA (S, M, R)	
			pattern	$(\mu, \delta)$	pattern	$(\mu, \delta)$	pattern	$(\mu, \delta)$	pattern	$(\mu, \delta)$
0	9	2	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)
2	9	2	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)
3	9	8	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)
4	8	8	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
5	8	1	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
7	8	2	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
8	9	2	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)
9	9	1	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)
10	8	4	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
12	9	108	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)	mS	(8, 8)
0	10	192	pS	(12, 9)	pS	(12, 9)	pM	(11, 10)	pM	(11, 10)
2	8	144	pS	(11, 7)	pS	(11, 7)	pM	(10, 8)	pM	(10, 8)
5	8	6	pS	(9, 7)	pS	(8, 7)	pS	(8, 7)	pS	(8, 7)
12	8	6	pS	(8, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
23	8	4	pS	(10, 7)	pS	(9, 7)	pS	(9, 7)	pS	(9, 7)
31	8	60	pS	(8, 7)	pS	(8, 7)	pS	(8, 7)	pS	(8, 7)
35	8	24	pS	(9, 7)	pS	(9, 7)	pS	(9, 7)	pS	(9, 7)
37	8	4	pS	(8, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
54	8	1	pS	(8, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
69	8	2	pS	(8, 7)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
7	8	6	pM	(10, 9)	pS	(9, 7)	mM	(8, 8)	mM	(8, 8)
19	8	48	pM	(9, 8)	pS	(9, 7)	pS	(9, 7)	pS	(9, 7)
35	9	48	pM	(11, 9)	pS	(11, 8)	pM	(10, 9)	pM	(10, 9)
45	8	2	pM	(11, 9)	pS	(10, 7)	pS	(10, 7)	pS	(10, 7)
76	9	4	pM	(10, 9)	pS	(10, 8)	pS	(10, 8)	pS	(10, 8)
172	10	180	pM	(14, 11)	pS	(14, 9)	pM	(12, 10)	pM	(12, 10)
216	8	6	pM	(9, 8)	pS	(8, 7)	pS	(8, 7)	pS	(8, 7)
239	8	2	pM	(9, 8)	pS	(8, 7)	pS	(8, 7)	pS	(8, 7)
241	8	6	pM	(9, 8)	pS	(9, 7)	mM	(8, 8)	mM	(8, 8)
243	9	12	pM	(13, 9)	pS	(13, 8)	pS	(13, 8)	pS	(13, 8)
9	8	16	pR	(9, 8)	pS	(10, 7)	pS	(10, 7)	pR	(9, 8)
157	11	48	pR	(12, 11)	pS	(12, 10)	pS	(12, 10)	pS	(12, 10)
383	8	16	pR	(10, 9)	pS	(10, 7)	pM	(9, 8)	pM	(9, 8)
394	9	12	pR	(10, 9)	pS	(9, 8)	pS	(9, 8)	pS	(9, 8)
905	8	10	pR	(9, 8)	pS	(9, 7)	pS	(9, 7)	pS	(9, 7)
981	9	8	pR	(12, 10)	pS	(9, 8)	pS	(9, 8)	pS	(9, 8)
1070	8	20	pR	(9, 8)	mS	(7, 7)	mS	(7, 7)	mS	(7, 7)
10046	8	2	pR	(11, 8)	pS	(9, 7)	mM	(8, 8)	mM	(8, 8)
10157	8	32	pR	(9, 8)	pS	(9, 7)	pS	(9, 7)	pS	(9, 7)
30342	8	2	pR	(10, 8)	pS	(8, 7)	pS	(8, 7)	pS	(8, 7)

Supplementary Table 3: **Simulated instances and MACHINA results with  $\Sigma_{\max} = 8$  anatomical sites.** For each simulated instance, we show the used random number generator seed, the simulated migration pattern, the number  $m$  of resulting anatomical sites, the simulated migration and comigration numbers  $(\mu, \gamma)$  and the number  $|\overline{\mathcal{T}}|$  of mutation trees enumerated by SPRUCE [9]. Next, we show the migration pattern and the migration and comigration numbers  $(\mu, \gamma)$  inferred by MACHINA under various topological constraints.

<b>patient</b>	<i>m</i>	<b>#primary regions</b>	<b>#metastases</b>	<b>#SNVs</b>	<b>#SNVs w/ homoplasy</b>
CRC1	7	4	3	209	12
CRC2	7	5	2	412	41
CRC3	11	5	6	253	15
CRC4	6	4	2	191	5
CRC5	4	2	2	144	1

Supplementary Table 4: **The clone trees inferred by Kim et al. [23] have extensive homoplasy.** The last column shows the number of characters that violate the infinite sites assumption after the removal of the specified region.

patient	#anatomical sites $m$	primary	#migrations $\mu$	#comigrations $\gamma$	pattern	#solutions
<b>1</b>	<b>7</b>	<b>LOv</b>	<b>13</b>	<b>7</b>	<b>pR</b>	<b>2</b>
1	7	LOv	13	11	pR	2
1	7	ROv	13	10	pM	1
<b>2</b>	<b>2</b>	<b>ROv</b>	<b>2</b>	<b>1</b>	<b>pS</b>	<b>1</b>
<b>3</b>	<b>8</b>	<b>LOv</b>	<b>27</b>	<b>7</b>	<b>pS</b>	<b>1</b>
3	8	LOv	27	9	pM	3
<b>3</b>	<b>8</b>	<b>ROv</b>	<b>27</b>	<b>7</b>	<b>pS</b>	<b>1</b>
3	8	ROv	27	9	pM	3
3	8	ROv	27	10	pM	4
3	8	ROv	27	12	pM	12
4	4	LOv	7	3	pS	1
<b>4</b>	<b>4</b>	<b>ROv</b>	<b>6</b>	<b>3</b>	<b>pS</b>	<b>1</b>
<b>7</b>	<b>7</b>	<b>LOv</b>	<b>12</b>	<b>6</b>	<b>pS</b>	<b>2</b>
7	7	LOv	12	7	pM	2
7	7	ROv	13	6	pS	2
7	7	ROv	13	7	pM	2
7	7	ROv	13	8	pM	6
7	7	ROv	13	9	pM	8
7	7	ROv	13	9	pR	6
7	7	ROv	13	10	pM	2
7	7	ROv	13	10	pR	6
<b>7</b>	<b>7</b>	<b>RUt</b>	<b>11</b>	<b>6</b>	<b>pS</b>	<b>1</b>
<b>9</b>	<b>3</b>	<b>LOv</b>	<b>4</b>	<b>2</b>	<b>pS</b>	<b>1</b>
9	3	ROv	5	2	pS	1
10	3	ROv	6	2	pS	1

Supplementary Table 5: **Results of the Sankoff enumeration algorithm for seven ovarian cancer patients [28]**. We enumerate all vertex labelings that achieve the minimum migration number  $\mu^*$  with either LOv (left ovary) or ROv (right ovary) as the primary. Patients 1, 3 and 7 admit multiple vertex labelings with the same minimum migration number  $\mu^*$ . These labelings yield different migration patterns with different comigration numbers  $\gamma$ ; the migration patterns are polyclonal single-source seeding (pS), polyclonal multi-source seeding (pM) and polyclonal reseeded (pR). For patient 1, the comigration number varies from 7 to 11, for patient 3 it varies from 7 to 12, and for patient 7 it varies from 6 to 10. Note that for patient 1, none of the vertex labelings have a single-source seeding migration pattern, i.e. they do not achieve the minimum comigration number  $\gamma_{\min} = m - 1 = 6$ . The minimum migration number  $\mu^* = 11$  for patient 7 is achieved when using a non-ovary anatomical site (RUt, right uterosacral ligament) as the primary.

patient	#sites $m$	#mutation clusters	#polytomies	Reported [28]		$P = \text{LOv}$			$P = \text{ROv}$				
				$P$	$(\mu^*, \gamma)$	(PS, S, M, R)	#sols	(PS, S, M)	(PS, S)	(PS, S, M, R)	#sols	(PS, S, M)	(PS, S)
1	7	9	5	ROv	(13, 10)	(13, 7)	2/4	(15, 6)	(15, 6)	(13, 10)	1/1	(13, 10)	(14, 6)
				refined		(11, 7)	n/a	(12, 6)	(12, 6)	(12, 6)	(12, 6)	n/a	(12, 6)
2	2	6	2	ROv	(2, 1)	n/a	n/a	n/a	n/a	(2, 1)	1/1	(2, 1)	(2, 1)
				refined		n/a	n/a	n/a	n/a	n/a	(1, 1)	n/a	(1, 1)
3	8	7	5	LOv	(27, 9)	(27, 7)	1/4	(27, 7)	(27, 7)	(27, 7)	1/20	(27, 7)	(27, 7)
				refined		(25, 7)	n/a	(25, 7)	(25, 7)	(25, 7)	(25, 7)	n/a	(25, 7)
4	4	8	4	ROv	(6, 3)	(7, 3)	1/1	(7, 3)	(7, 3)	(6, 3)	1/1	(6, 3)	(6, 3)
				refined		(6, 3)	n/a	(6, 3)	(6, 3)	(6, 3)	(6, 3)	n/a	(6, 3)
7	7	5	4	RUt*	(11*, 6)	(12, 6)	2/4	(12, 6)	(12, 6)	(13, 6)	2/32	(13, 6)	(13, 6)
				refined		(11, 6)	n/a	(11, 6)	(11, 6)	(11, 6)	(11, 6)	n/a	(11, 6)
9	3	3	2	LOv	(4, 2)	(4, 2)	1/1	(4, 2)	(4, 2)	(5, 2)	1/1	(5, 2)	(5, 2)
				refined		(4, 2)	n/a	(4, 2)	(4, 2)	(4, 2)	(4, 2)	n/a	(4, 2)
10	3	6	3	ROv	(6, 2)	n/a	n/a	n/a	n/a	(6, 2)	1/1	(6, 2)	(6, 2)
				refined		n/a	n/a	n/a	n/a	n/a	(6, 2)	n/a	(6, 2)

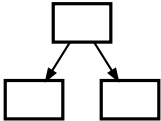
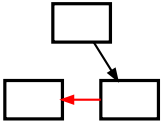
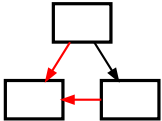
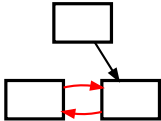
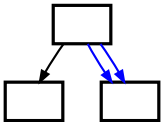
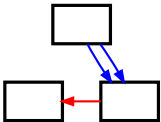
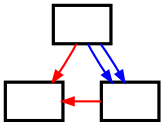
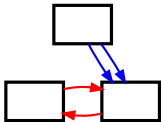
Supplementary Table 6: **Results for seven ovarian cancer patients [28]**. For each patient, we show the number  $m$  of anatomical sites, the number of mutation clusters, the number of polytomies, the anatomical site  $P$  denoted as the primary tumor, the migration and comigration number  $(\mu, \gamma)$  in the reported labeling in [28]. We enumerate all labelings with minimum migration number  $\mu^*$  and the primary tumor  $P$  set to either LOv (left ovary) or ROv (right ovary) using the Sankoff algorithm. We report the minimum-migration solution in the R column, as well as the number of labelings that correspond to this solution and the total number of enumerated solutions. In addition, we use the ILP to solve the PMH problem under various topological constraints, including reseeding (R), multi-source seeding (M) and single-source seeding (S). The second row of each instance reports the solution to the PHM-TR problem. (\*) In patient 7, the reported vertex labeling in [28] has one fewer migration due to setting the primary to right uterosacral ligament (RUt), suggesting a different non-ovarian tissue of origin than the remaining ovarian cancer patients in the study.

patient	$m$	#SNVs	Gudem et al. [15]			MACHINA (PS)		MACHINA (PS, S)		MACHINA (PS, S, M, R)	
			#clusters	monoclonal seeding	met-to-met spread	pattern	$(\mu, \gamma, \sigma)$	pattern	$(\mu, \gamma, \sigma)$	pattern	$(\mu, \gamma, \sigma)$
A10	4	9472	9	yes (m)	yes	mPS	(3,3,1)	mPS	(3,3,1)	mPS	(3,3,1)
A29	2	8275	5	yes (m)	no (PS)	mPS	(1,1,1)	mPS	(1,1,1)	mPS	(1,1,1)
A31	5	4852	10	no (p)	yes	pPS	(12,4,1)	pS	(10,4,2)	pS	(10,4,2)
A32	6	9388	12	no (p)	yes	pPS	(8,4,1)	pPS	(8,4,1)	pM	(7,5,2)
A22	10	10262	16	no (p)	yes	pPS	(36,9,1)	pS	(32,9,3)	pR	(26,12,5)

Supplementary Table 7: **Results for five prostate cancer patients [15]**. From left to right, we show the patient identifier, the number  $m$  of anatomical sites, the reported presence or absence of monoclonal seeding and a metastatic cascade. Next, we show the inferred migration pattern and score using different constraints for the PHM-TR problem (parallel single-source (PS), single-source (S), multi-source (M) and reseeded (R)).

patient	$m$	#SNVs	Sanborn et al. [44]		MACHINA		
			#clusters	polyclonal	#clusters	pattern	$(\mu, \gamma, \sigma)$
A	4	124/2467	7	no	4	mPS	(3,3,1)
C	3	1056/1408	5	yes	5	pPS	(3,2,1)
D	4	13/1160	3	no	2	mPS	(3,3,1)
E	5	59/96	6	yes	7	pS	(5,4,2)
F	4	2136/2136	7	no	5	mPS	(3,3,1)

Supplementary Table 8: **Results for five metastatic melanoma patients**. From left to right, we show the patient, the number  $m$  of anatomical sites, the number of copy-neutral and total number of SNVs, the number of reported clusters and whether polyclonal seeding was reported in [44]. Next, we show the number of clusters inferred by the AncestryTree clustering algorithm and finally the migration pattern and parsimony score inferred by MACHINA.

	parallel single-source seeding (PS)	single-source seeding (S)	multi-source seeding (M)	reseeding (R)
monoclonal (m)	 <p>tree</p> $\mu = \mu_{\min} = m - 1$ $\gamma = \gamma_{\min} = m - 1$ $\sigma = \sigma_{\min} = 1$	 <p>tree</p> $\mu = \mu_{\min}$ $\gamma = \gamma_{\min}$ $\sigma > \sigma_{\min}$	 <p>directed acyclic graph</p> $\mu > \mu_{\min}$ $\gamma > \gamma_{\min}$ $\sigma > \sigma_{\min}$	 <p>directed graph</p> $\mu > \mu_{\min}$ $\gamma > \gamma_{\min}$ $\sigma > \sigma_{\min}$
polyclonal (p)	 <p>multi-tree</p> $\mu > \mu_{\min}$ $\gamma = \gamma_{\min}$ $\sigma = \sigma_{\min}$	 <p>multi-tree</p> $\mu > \mu_{\min}$ $\gamma = \gamma_{\min}$ $\sigma > \sigma_{\min}$	 <p>directed acyclic multi-graph</p> $\mu > \mu_{\min} + 1$ $\gamma > \gamma_{\min}$ $\sigma > \sigma_{\min}$	 <p>directed multi-graph</p> $\mu > \mu_{\min} + 1$ $\gamma > \gamma_{\min}$ $\sigma > \sigma_{\min}$

Supplementary Table 9: **Taxonomy of migration patterns between anatomical sites.** Migration patterns can be distinguished in two different ways. First, by the number of clones that migrate between two anatomical sites: each metastasis is seeded by a single clone in the case of monoclonal (m) seeding, whereas with polyclonal (p) seeding multiple clones migrate from one anatomical site to another. Second, by the migration topology: each metastasis is seeded only from the primary tumor with parallel single-source seeding (PS), is seeded from a single anatomical site with single-source seeding (S), and has multiple anatomical seeding sites with multi-source seeding (M) and clones migrate back and forth between anatomical sites in the case of reseeding (R). With  $m$  anatomical sites  $\mu_{\min} = m - 1$  and  $\gamma_{\min} = m - 1$  are lower bounds on the migration and comigration number, respectively. The migration pattern affects the migration number  $\mu$  and the comigration number  $\gamma$ .