Inferring cancer type-specific patterns of metastatic spread

Divya Koyyalagunta[1, 2] **and Quaid Morris**[1, 2]

[1]Tri-Institutional Graduate Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, NY 10065, USA
[2]Computational and Systems Biology Program, Sloan Kettering Institute, New York, NY 10065, USA.

The metastatic spread of a cancer can be reconstructed from DNA sequencing of primary and metastatic tumours, but doing so requires solving a challenging combinatorial optimization problem. This problem often has multiple solutions that cannot be distinguished based on current maximum parsimony principles alone. Current algorithms use ad hoc criteria to select among these solutions, and decide, a priori, what patterns of metastatic spread are more likely, which is itself a key question posed by studies of metastasis seeking to use these tools. Here we introduce Metient, a freely available open-source tool which proposes multiple possible hypotheses of metastatic spread in a cohort of patients and rescores these hypotheses using independent data on genetic distance of metastasizing clones and organotropism. Metient adapts Gumbel-softmax gradient estimators, to quickly map out a Pareto front of migration histories that cover the range of histories that are parsimonious under some criteria. Given a cohort of patients, Metient can calibrate its parsimony criteria, thereby identifying shared patterns of metastatic dissemination in the cohort. Compared with the current state-of-the-art, Metient recovers more migration histories, is more accurate, and is more than 40x faster. Reanalyzing metastasis in 169 patients based on 490 tumors, Metient automatically identifies cancer type-specific trends of metastatic dissemination in melanoma, high-risk neuroblastoma and non-small cell lung cancer. Metient's reconstructions usually agree with semi-manual expert analysis, however, in 24 patients, Metient identifies more plausible migration histories than experts, and thus finds that polyclonal seeding of metastases is more common than previously reported.

migration history inference | metastasis | mixed-variable combinatorial optimzation

Correspondence: *morrisq@mskcc.org*

Main

Metastasis is associated with 90% of cancer deaths, yet its causes and physiology remain poorly understood[1]. It remains unclear whether multiple clones within the same primary cancer can seed metastases, and if so, what the relationship between the clone and the seeding site is[2–8]. In some cancers, metastases can seed other metastases, though perhaps not all[2,9,10]. It is not known whether metastatic potential is rare, and thus gained once, or common, and thus gained multiple times, in the same cancer[11–14]. The answers to all these questions impact the understanding and clinical management of metastasis.

To address these questions, one can use molecular data from primary and, ideally multiple, matched metastatic tumors to reconstruct the migratory histories of the metastatic clones[3,15]. Specifically, one can reconstruct the clonal composition of each sample from bulk DNA sequencing data[16], and use maximum parsimony algorithms applied to these clones to recover migration histories[15,17,18]. However, current parsimony algorithms rely on ad hoc assumptions to resolve differences among various possible parsimonious solutions[15,17,18], and these assumptions bias the output of these tools because they pre-select what patterns of metastasis are allowed. Whereas, a main goal of reconstructing migration histories is to determine what the likely patterns of metastatic spread in each cancer type actually are. For example, a prevailing model in oncology, the "sequential progression model", states that lymph node metastases give rise to distant metastases, and is the rationale for surgical removal of lymph nodes[19]. However, a phylogenetic analysis of colorectal tumors revealed that distant metastases are more often seeded from both the lymph node metastases and the primary, compared to just the lymph nodes alone[20]. Pre-defined constraints on metastatic seeding patterns, especially simplistic ones that assume primary-only seeding, prevent analyses from recovering the metastatic patterns that are the best fit to real data.

To resolve this dilemma and overcome weaknesses shared by prior migration history reconstruction algorithms (Supplementary Table 1), we present Metient (**met**astasis + grad**ient**), a principled statistical algorithm that identifies multiple possible hypotheses of metastatic spread in a patient, and scores each in an unbiased way by making reference to other relevant data. To accomplish this, Metient makes two key innovations. First, Metient adapts recently proposed combinatorial optimization approaches to efficiently sample multiple parsimonious solutions. Then, Metient introduces new biological criteria, which we refer to as metastasis priors, to learn overall trends of a particular cancer type and resolve conflicting parsimonious solutions.

On realistic simulated data, Metient is more accurate at recovering the ground truth migration history than a parsimony-only model, and when applied to patient cohorts with metastatic skin[9], ovarian[10], neuroblastoma[7], breast[21] and lung cancer[14], Metient automatically recovers all plausible expert-assigned migration histories, and in some notable cases also identifies more plausible reconstructions, particularly when, in order to use previous methods, expert analyses pre-select a particular seeding pattern of metastasis. In these cases, patients in the cohort almost certainly exhibit alternate patterns of spread.

Using its automated, unbiased approach, Metient finds that metastases are often seeded polyclonally, and that most

metastatic seeding evolves along a single, shared evolutionary trajectory. The cancer type-specific model learned by Metient for each cohort reflects known differences in metastatic biology, thus Metient offers a way to infer new biology about metastatic dissemination in cancer types where patterns are unknown. Metient is freely-available, open source software and we provide it along with easy visualizations to help generate and compare multiple hypotheses on metastatic dissemination at https://github.com/morrislab/metient/.

Results

Identifying multiple possible hypotheses of metastatic spread in a patient

Migration history inference algorithms take as input DNA sequencing data from primary and metastatic tumor samples and an unlabeled clone tree which encodes the genetic ancestry of cancer clones (Figure 1a). These inputs are used to estimate the proportions of clonal populations in anatomical sites ("witness nodes" in Figure 1b,c). Then, the internal nodes of the clone tree are labeled with anatomical sites, thereby defining the historical migrations: clones that migrated to a new site have a different label than their parent clone (Figure 1b,c). We refer to the final output as a "migration history"[15] (Figure 1c).

MACHINA[15] is the most widely used, and most advanced, migration history reconstruction algorithm. It scores migration histories using three parsimony metrics: **migrations**, the number of times a clone migrates to a different site[10,15,17,18]; **comigrations**, the number of migration events in which one or more clones travel from one site to another[15]; and **seeding sites**, the number of anatomical sites that seed another site[15]. MACHINA searches for the most parsimonious history by minimizing these three metrics. However, parsimony conflicts can frequently arise, in cases where multiple histories can be equally parsimonious, or it is possible to trade-off one metric in favor of another, e.g., the number of seeding sites can be reduced by increasing the number of migration events.

To offer a more systematic approach, we designed Metient to define a "Pareto front"[22] for each patient, which captures the relative trade offs between the three parsimony metrics (Figure 1b,c). Defining a Pareto front allows Metient to reduce a large combinatorial search space to only the plausible explanations of metastatic spread for a given patient. Recovering a Pareto front required an alternative approach to solving the combinatorial mixed-variable optimization underlying migration history inference. This inference involves solving for a continuous variable (observed clone percentages ($\mathbf{U}$; in Figure 1b)), and a discrete variable (labeled clone tree ($\mathbf{V}$ in Figure 1b)). Previous approaches have formulated this as a mixed integer linear programming (MILP) problem and rely on commercial solvers[23]. However, MILP solvers only identify a single optimal solution, and furthermore require hard constraints and a linear objective, severely limiting the types of scoring functions that can be applied to migration histories. In contrast, Metient uses a flexible statistical approach in which the migration history score is proportional to its posterior probability, and the Pareto front corresponds to different modes of this distribution (Methods). Metient uses state-of-the-art gradient descent methods to optimize this objective, relying on a low variance gradient estimator for the discrete categorical distribution over migration histories[24,25] ($\mathbf{V}$ in Figure 1b; Methods, Supplementary Information).

A maximum parsimony-only model relies on ad hoc assumptions

To test the utility of recovering multiple possible hypotheses of metastatic spread (the Pareto front) on patient data, we identified cohorts with publicly available genomic sequencing of matched primary and multiple metastases from four cancer types: melanoma[9], high-grade serous ovarian cancer (HGSOC)[10], high-risk neuroblastoma (HR-NB)[7],

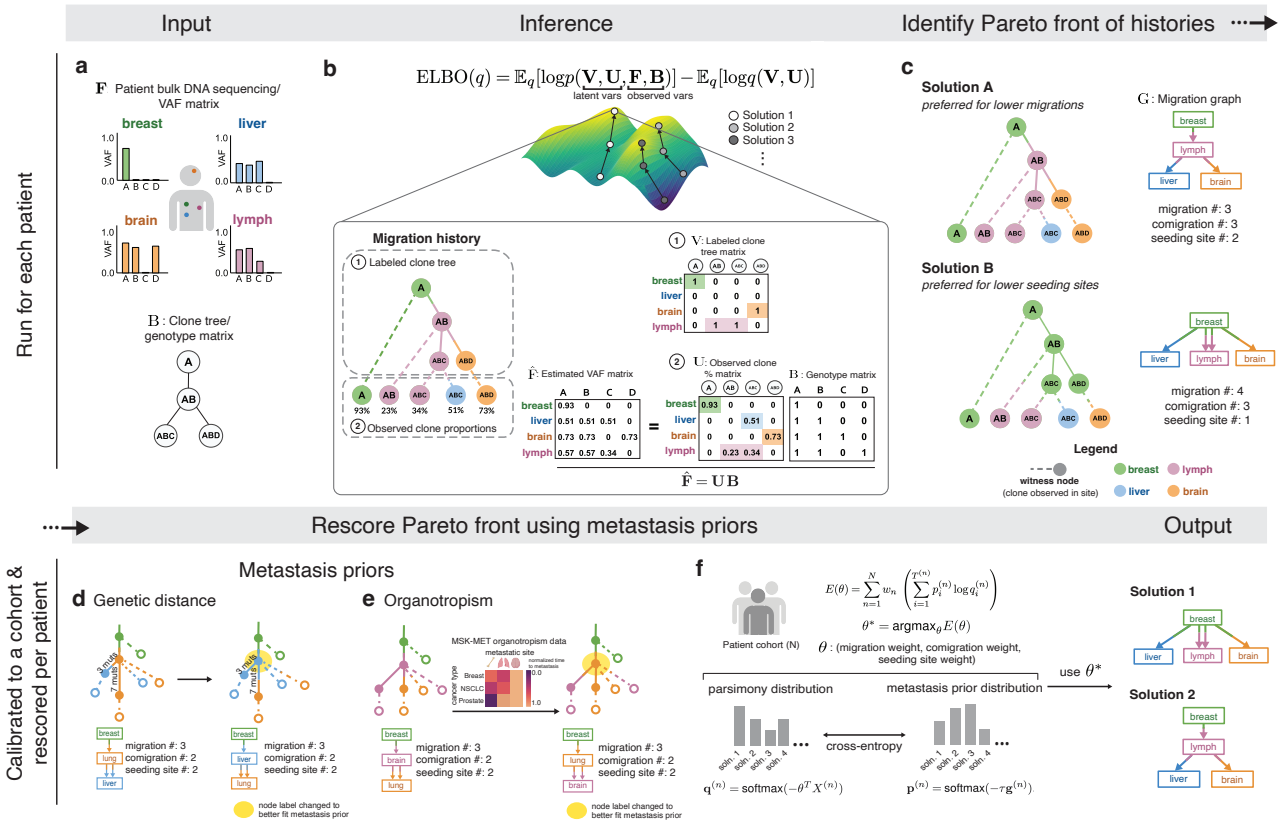

Figure 1. Overview of the Metient method. (a) Input: (top) bulk DNA sequencing sampled from multiple tumors in a single patient, and (bottom) a clone tree which represents the evolutionary relationship of mutations. AB refers to a clone with mutations or mutation clusters A and B. **(b) Inference**: Using the inputs as observed variables, we infer the latent variables (1) $\mathbf{V}$ (representing the labeled clone tree) and (2) $\mathbf{U}$ (representing the percent of each clone in each anatomical site). $\hat{\mathbf{F}}$ is the estimated VAF matrix produced by $\mathbf{UB}$, where $\mathbf{B}_{ij} = 1$ if clone i contains mutation j. Each migration history solution can be represented by a migration history, which is a clone tree with (1) an anatomical site labeling of its internal nodes, and (2) leaf nodes representing the observed clone proportions in anatomical sites. **(c) Identify Pareto front of histories**: We infer the Pareto front of migration histories as defined by the three parsimony metrics (migration, comigration and seeding site number). A migration graph $\mathbf{G}$ summarizes the migration edges of the migration history. **(d) Genetic distance**: An example of how using genetic distance can promote migration histories with migrations on longer edges with more mutations. The anatomical site label of the yellow shaded node is changed. **(e) Organotropism**: An example of how using organotropism can promote migration histories that do not contain unlikely metastatic patterns, such as subsequent metastasis from the brain. The anatomical site label of the yellow shaded node is changed. **(f) Metient-calibrate**: Weights on the parsimony metrics (θ) are fit by minimizing the cross entropy loss between each patient's migration histories' probability distribution as scored by the metastasis priors (target distribution) and the probability distribution as scored by the parsimony metrics (source distribution). These weights are fit across a cohort of patients, and then used to rescore the trees produced for each patient in that cohort.

and non-small cell lung cancer (NSCLC)[14]. After applying quality control (Supplementary Information), we arrived at a dataset of 479 tumors (143 with multi-region sampling) from 167 patients (melanoma: n=7, HGSOC: n=7, HR-NB: n=27, NSCLC: n=126). When applying Metient to these patients, we found that ranking the migration histories on the Pareto front frequently involved non-trivial decisions (74/167 patients have multiple Pareto optimal migration histories), and that these choices have a substantial impact on the interpretation of metastatic spread. Figure 1c shows an example patient with metastatic breast cancer with two equally parsimonious reconstructions: one in which a lymph node metastasis gives rise to all other metastatic tumors, and another where most metastases are seeded directly from the primary tumor. Here, an arbitrary choice between the two reconstructions determines whether one concludes that the lymph node acted as a staging site for metastatic spread.

MACHINA and all previous methods[10,15,17] resolve conflicts posed by two equally plausible migration histories by minimizing migrations first. This decision is problematic because no single set of preselected constraints will be appropriate for all cancer types. For example, in many solid cancers, metastatic cells make a "pit stop" at nearby lymph nodes before disseminating to other distant sites[26], and for the estimated 23.4% of patients with lymph node metastases across cancer types[27], multiple seeding sites may be common.

In ovarian cancer, clusters of metastatic cells "passively" disseminate to the peritoneum or omentum through peritoneal fluid[28–30]. Here, metastatic events are more likely to be polyclonal, i.e., multiple clones seed metastases, so we might expect more migrations than comigrations. Thus, different modes of metastatic spread may be reflected in differences in the relative numbers of migrations, comigrations, and seeding sites, and prespecifying the relative importance of parsimony metrics would interfere with methods' ability to detect these patterns.

Rather than pre-deciding on which metastatic processes are more likely, Metient assigns a weighted parsimony score, p, to a history with m migrations, c comigrations, and s seeding sites, with weights w_m, w_c, and w_s assigned to each respective parsimony metric. Different migration histories on the Pareto front are favored under different settings of $[w_m\ w_c\ w_s]$. To fit these weights, Metient makes use of biologically relevant data, which we call the "metastasis priors" to identify an optimal setting of $[w_m\ w_c\ w_s]$ across a cohort of patients with the same cancer type (Figure 1d-f; Methods). If parsimony weights are already available, or users want to run Metient on individual patients, they can do so in the Metient-evaluate mode (Methods). We also provide weights learned across the combined four cohorts as the default parameters for Metient-evaluate. Since the aforementioned calibration is Metient's default mode, whenever unspecified, Metient refers to Metient-calibrate.

The first metastasis prior we use is genetic distance, which measures the number of mutations between a clone and its parent clone in the clone tree. The genetic distance between clones can serve as a proxy for time[31–33]. Indeed, metastases across many cancer types have moderately or significantly higher tumor mutation burden (TMB) than matched primaries[27,34,35]. As such, the genetic distance prior scores migration histories based on the averaged genetic distances between migrating clones and their parent. For example, the two hypothetical migration histories of a breast cancer patient in Figure 1d are equally likely under a maximum parsimony-only model, but genetic distance helps choose between the two by promoting the history with a migration on a longer tree edge.

As another prior we also use organotropism to score certain migration edges as more likely than others. Organotropism refers to the preference that some cancer types have to colonize other organs[36]. To make use of this preference in scoring migration histories, we used clinical annotations from more than 25,000 Memorial Sloan Kettering metastatic cancer patients[27] to construct a matrix for 27 common cancer types, where each entry is the frequency that metastasis to a particular anatomical site is observed in that cancer type (Figure 1e). Note that there is no direct data for frequencies of migrations from metastatic sites to other metastatic sites, so we use the organotropism matrix to score migrations coming from the primary site (Methods). In general, we expect frequently observed metastatic sites for a cancer to be migrations directly from the primary site, thereby implicitly assuming that less frequently observed migrations may represent migrations from a metastatic site. For example, according to the organotropism matrix, breast cancer metastasizes more often to lung than brain. Our organotropism prior thus favors a solution with migrations from the lung to the brain, over a solution with migrations from the brain to the lung, which is clinically rare[37] (Figure 1e).

In our benchmarking analyses on simulated data, we find that genetic distance and organotropism alone can result in the inference of highly non-parsimonious migration histories. This is the motivation behind first choosing a likely set of trees (the Pareto front) using parsimony metrics, and then calibrating the parsimony metric weights to the metastasis priors to choose between multiple plausible trees. A previous method, PathFinder[38], which uses Bayesian inference with a genetic distance objective, found that applying parsimony criteria did not improve overall performance. However, in our analyses, we find that genetic distance alone does not perform as well as maximum parsimony with genetic distance (Supplementary Tables 2, 3).

Metient achieves state-of-the-art performance

To assess Metient's new objective and gradient-based optimization on data with a provided ground-truth, we ran benchmarking analyses along with the state-of-the-art migration history inference method (MACHINA[15]) on a simulated dataset of 80 patients with 5-11 tumor sites and various patterns of metastatic spread originally introduced by El-Kebir et al[15]. Metient predicts ground truth in the simulations at least as accurately as MACHINA (Figure 2a,b), with larger improvements in higher input tree sizes and more complex seeding patterns such as polyclonal multi-source and reseeding (Figure 2a,b).

To further evaluate if our adaptive model (Metient-calibrate) is better at choosing the best migration history from the Pareto front than a fixed model (Metient-evaluate with $w_m >> w_c >> w_s$, i.e., the same weighting scheme as MACHINA), we categorized the simulated patients into cohorts based on their ground-truth seeding pattern and compared performance of these two models. Metient-calibrate improves performance over the fixed model in most seeding pattern categories (Supplementary Tables 2, 3), showcasing the ability of the metastasis priors to learn metastatic patterns specific to a cohort and improve overall accuracy.

Notably, although the Metient framework is non-determenistic, it identifies the same top solution 97% of the time across multiple runs (Metient-calibrate with a sample size of 1024; Figure 2c). Furthermore, in addition to its

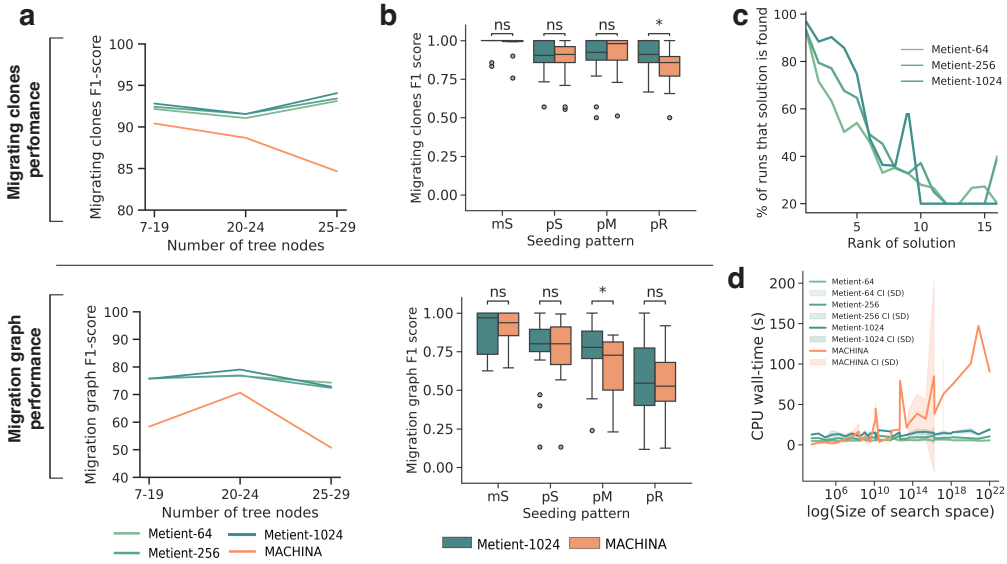

Figure 2. Metient achieves state-of-the-art performance on simulated data while producing multiple possible explanations of metastatic spread. All results shown for Metient are in calibrate mode using genetic distance as the metastasis prior. Metient-1024 refers to a model configuration where 1024 solutions are sampled. For a given simulated input, for MACHINA (which outputs one solution) the top solution is used, and for Metient we evaluate all top solutions. **(a)** The averaged F1-score for predicting migrating clones (top) and migration graph (bottom), within three buckets of input tree sizes. **(b)** The distribution of F1-scores for predicting migrating clones (top) and migration graph (bottom) on different seeding patterns. m: monoclonal, p: polyclonal, S: single-source seeding (includes primary single-source seeding), M: multi-source seeding, R: reseeding. Statistical significance assessed by a paired t-test: *: $p < 0.05$. **(c)** After running Metient five times, the percentage of runs that a certain solution is found as a function of its averaged rank across runs. **(d)** CPU wall-time needed to run Metient vs. MACHINA as a function of the search space size. CI: confidence interval, SD: standard deviation.

improved accuracy, Metient runs up to 44x faster (5.02s with Metient-64 vs. 221.2s with MACHINA for a cancer tree with 18 clones and 9 tumors), showcasing our framework's scalability even as tree sizes get very large (Figure 2d).

Multi-cancer analysis of clonality, phyleticity, and dissemination patterns

Having established that Metient can accurately recover ground-truth and learn cohort-specific metastatic patterns on simulated data, we next sought to apply the method to real patient data from the melanoma, HGSOC, HR-NB and NSCLC cohorts to investigate shared and unique patterns of metastatic dissemination. Due to missing or inadequate anatomical site labels for many patients in these cohorts, we were unable to use Metient's organotropism matrix on these cohorts, and we only calibrated to genetic distance. We validated the organotropism prior on a breast cohort described later.

Using Metient, we examined three aspects of metastatic dissemination across the four cohorts. The first aspect is seeding pattern, which can be sub-categorized as single-source from the primary or from another site, multi-source or reseeding (Figure 3a). The other two criteria are clonality, i.e., the number of distinct clones seeding metastases (Figure 3b,c), and phyleticity, i.e., whether metastatic potential is gained in one or multiple evolutionary trajectories of the clone tree (Figure 3d; Methods). We distinguish between genetic polyclonality, in which more than one clone seeds metastases, and site polyclonality, in which more than one clone migrates to an individual site (Figure 3c;

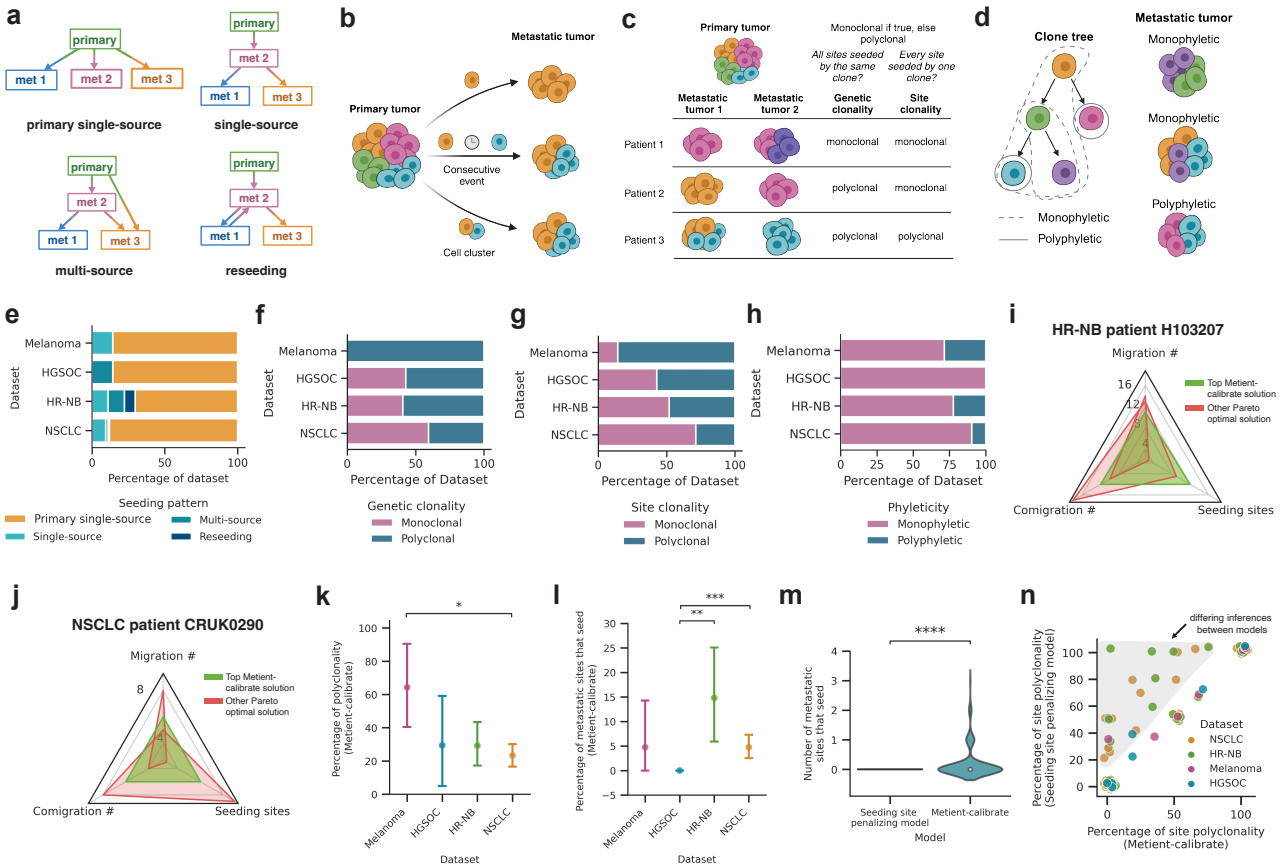

Figure 3. Clonal, phyletic and seeding patterns of four cancer types. (a) Schematic describing the four metastatic seeding patterns. met: metastasis. **(b)** Schematic depicting how metastases can get seeded by either one or multiple clones. When a site is seeded by multiple clones, this can be a result of multiple clones traveling in a cluster to the same anatomical site, or because of two clones traveling one after the other to the same site. Colors represent genetically distinct cancer cell populations. **(c)** Schematic depicting the definitions of genetic clonality and site clonality. Colors represent genetically distinct cancer cell populations. **(d)** Schematic depicting the definitions of monophyletic and polyphletic seeding. Monophyletic indicates that the seeding clone closest to the root can reach every other seeding clone on the clone tree. Colors represent genetically distinct cancer cell populations. Distribution of **(e)** seeding patterns, **(f)** genetic clonality, **(g)** site clonality and **(h)** phyleticity for each dataset, as inferred by Metient's top migration history. **(i)** Radar plot showing the unique Pareto optimal metrics for migration histories inferred by Metient for HR-NB patient H103207. **(j)** Radar plot showing the unique Pareto optimal metrics for migration histories inferred by Metient for NSCLC patient CRUK290. **(k)** Comparing across datasets the percent of migrations that are polyclonal for the top Metient solution. Statistical significance assessed by a Welch's t-test; *: 1e-02 < p <= 5e-02. Error bars are the standard error for each dataset, **(l)** Comparing across datasets the percent of metastatic sites that seed for the top Metient solution. Statistical significance assessed by a Welch's t-test; **: 1e-03 < p <= 1e-02, ***: 1e-04 < p <= 1e-03. Error bars are the standard error for each dataset. **(m)** A comparison of the number of metastatic sites that seed other sites between migration histories chosen by a model which chooses the migration history with the lowest number of seeding sites and Metient. Statistical significance assessed by a paired t-test, p=1.948e-06. **(n)** A comparison of the percent of site polyclonal migrations for each patient's migration history when using the best migration history chosen by Metient (x-axis) vs. a model that chooses the migration history with the lowest number of seeding sites (y-axis).

Methods).

Consistent with expert annotations[7,9,10,14,15], Metient finds that single-source seeding from the primary tumor is the most common pattern in every cohort (Figure 3e). However, Metient identifies a larger fraction of polyclonal migration patterns than previous reports[6,14]: 46.7% of patients have sites that are seeded by different clones, i.e., genetically polyclonal (Figure 3f), and 35.3% of patients have at least one site seeded by multiple clones, i.e. site polyclonal (Figure 3g). In an analysis of breast, colorectal and lung cancer patients, Hu et al.[6] estimate that 19.2% of sites across cancer types, compared to 31.3% of sites in our Metient analysis (100/320), are seeded by multiple clones. This heightened sensitivity for Metient to detect polyclonal migration can be attributed to its clone-tree-guided reconstruction of the seeding clones, which previous analyses neglect.

Metient's phyleticity estimates are largely consistent with previous reports: 88% of patients (147/167) have a monophyletic tree where metastatic potential is gained once and maintained (Figure 3h). For some patients, this is due to the root clone (i.e. clonal population) being observed in one or more metastatic sites (Supplementary Figure S1a), and for other patients, all seeding clones belong to a single path of the clone tree. Either scenario suggests that metastatic potential is less likely to be gained via multiple, independent evolutionary trajectories across cancers.

Validation of organotropism prior

To validate the organotropism prior, we ran Metient-evaluate on samples available from two patients with metastatic breast cancer[21] where site labels could be mapped to those used in our organotropism matrix. When faced with multiple parsimonious migration histories, Metient chooses a more plausible tree, wherein lung to brain seeding is preferred over brain to lung seeding, which is clinically rare[37] (Figure 4a).

Cancer-specific metastasis trends

We next examined cancer-specific differences in metastatic trends, first using a bootstrapping approach to ensure that the parsimony metric weights were reproducible and reflective of population level patterns for a particular cancer type. We fit parsimony metric weights to 100 bootstrapped samples of patients within the cohort, and found that 98.4% of patients ranked the same top solution across bootstrap samples, indicating that Metient is learning a reproducible cancer type-specific model even with lower sample sizes (e.g. melanoma and HGSOC cohorts with seven patients each).

These cancer type-specific parsimony metric weights lead to cohort-specific choices on how Metient ranks a patient's Pareto front of migration histories. For example, Metient chooses the solution on the Pareto front with lowest migration number (i.e. migrating clones) for HR-NB patient H103207 (Figure 3i), but the solution with the median value of each metric for NSCLC patient CRUK0290 (Figure 3j). To systematically assess the impact of cohort-specific rankings we computed the percentage of polyclonality and number of seeding sites for each cancer type. Overall, we found a significantly higher fraction of polyclonal migrations in melanoma than NSCLC patients (Figure 3k). One explanation for this heightened polyclonality in melanoma patients is that all patients in the cohort had locoregional

skin metastases, a common “in-transit” metastatic site around the primary melanoma or between the primary melanoma and regional lymph nodes. These locoregional sites could have multiple cancer cells travel through hematogeneous or lymphatic routes to seed new localized tumors[39]. HR-NB and NSCLC had significantly higher percentages of metastasis-to-metastasis seeding than HGSOC (Figure 3l). We show that multiple HR-NB patients exhibit metastasis-to-metastasis seeding within an organ (below). Also, 76.2% of NSCLC patients have lymph node metastases, from which it is known that further metastases are commonly seeded. Indeed, 66% of NSCLC patients who had metastasis-to-metastasis seeding (10/15) had seeding from a lymph node to other metastases.

Model choice impacts downstream analyses

As we were analyzing different aspects of metastatic dissemination, we asked how these answers might change if a seeding model is enforced when reconstructing a patient’s migration history. To highlight how the choice of seeding model can impact the analysis and interpretation of metastatic dissemination, we compared the migration histories produced by three models: (1) assumption of primary, single-source seeding, (2) the MACHINA assumptions, which first minimize migrations, and then break ties based on comigration count followed by seeding site number, and finally (3) the adaptive Metient model fit to each cohort. As expected, a primary, single-source seeding model chooses a primary, single-source dissemination pattern for 100% of patients (Supplementary Figure S1b). The migration penalizing model chooses a primary single-source seeding explanation in 82% of patients, and Metient falls in between the two, choosing a primary single-source seeding explanation in 85% of patients (Figure S1c). Importantly, since Metient can recover and evaluate the relative trade-offs of the parsimony metrics, when choosing a primary single-source solution, our model has either not found a plausible metastasis-to-metastasis explanation for a patient’s data on the Pareto front, or has used the metastasis priors to deem such an explanation less likely. In contrast, previous models do not automatically recover multiple possible hypotheses, therefore reducing confidence in these algorithms’ choice of best history.

In addition to having an impact on the inferred seeding patterns, a model that assumes primary single-source seeding also changes other interpretations of metastatic seeding. We asked two questions about the best migration histories produced by the two extremes of models, i.e. the seeding-site penalizing model and Metient: (1) the frequency in which a new seeding site is added, and (2) the frequency of polyclonal migrations between two sites. As expected, a model which penalizes seeding sites promotes migration histories with only one seeding site (Figure 3m). In turn, such a model infers a higher fraction of polyclonal migrations (Figure 3n) compared to the histories prioritized by Metient. The trade-off between polyclonality and seeding sites occurs because additional seeding sites reduce the number of migration edges that must be placed between the primary and all other metastases. Balancing this trade-off correctly is important as it impacts the interpretation of seeding clonality as well as which clones perform seeding. Specifically, 10.2% (17/167) of patients have differing seeding clones between the two models, significantly changing the inference of which clones, and therefore which mutations, have metastatic competence.

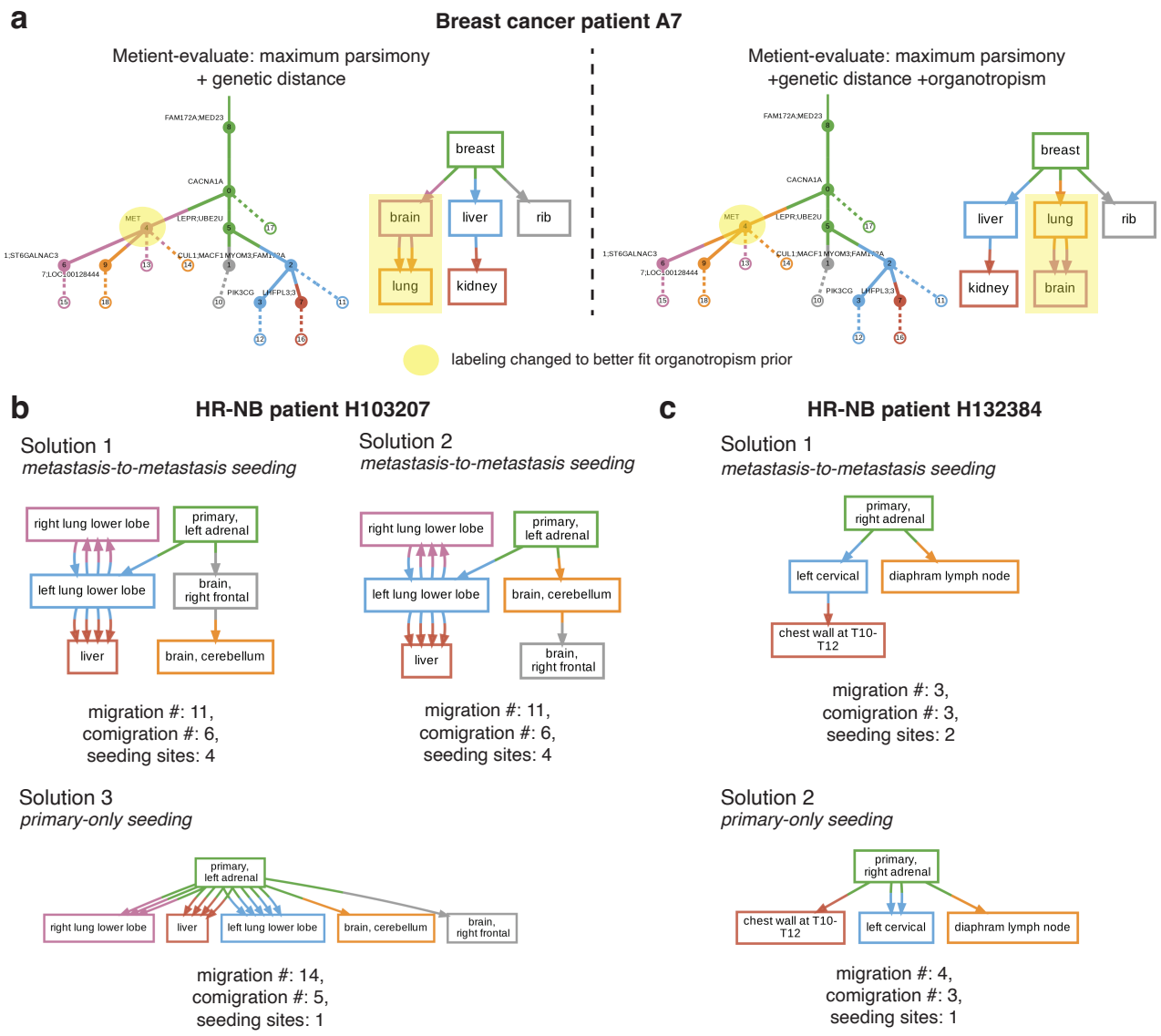

Figure 4. Metient finds biologically relevant trees. (a) Migration history for breast cancer patient A7[21], with Metient-evaluate using genetic distance (left) and Metient-evaluate using genetic distance and organotropism (right). Solid edges are a part of the clone tree, and dashed edges indicate the presence of the clone in the corresponding colored anatomical site. For example, clone 7 is present in the kidney. The addition of an organotropism prior changes the vertex labeling of the clone 4 from originating in the brain to originating in the lung. **(b)** The top three Pareto optimal migration graphs inferred by Metient-calibrate for HR-NB patient H103207. Note that this patient has additional trees on the Pareto front that are not displayed. **(c)** All ranked Pareto optimal migration graphs inferred by Metient-calibrate for HR-NB patient H132384.

Metastasis priors identify biologically relevant migration histories and alternative explanations of spread

A core advance of Metient is its ability to identify and rank the Pareto optimal histories of a patient's cancer. To assess how well our top ranked solution aligns with the most biologically plausible explanation, we compared our inferred migration histories to previously reported, expert-annotated seeding patterns. This decision is not trivial, as almost half of all patients have multiple migration histories on the Pareto front, and 40% of those patients have migration histories with different parsimony metrics.

Of the 169 patients analyzed, 152 patients had an expert or model-derived annotation available. Because the HR-NB annotations only indicate the presence of a migration between two sites and not the directionality, for an overall comparison of these 152 patients we compared our site-to-site migrations to those that were previously reported (i.e., a binarized representation of migration graph $\mathbf{G}$ (Figure 1c)). In 84% of patients (128/152), Metient-calibrate's highest ranked solution aligns with the previously reported migration history. For the remaining 24 patients, Metient either promotes a simpler explanation, infers the expert annotation on the Pareto front but prefers another migration history, or chooses a migration history better supported by prior understanding of common organ-specific metastatic dissemination patterns. We provide a detailed case-by-case comparison in the Supplementary Information and Supplementary Figures S2, S3, S4, S5, and highlight some of the interesting cases below.

Metient predicted metastasis-to-metastasis seeding for two HR-NB cases (H103207, H132384), which were previously thought to have seeded directly from the primary[7]. Even though information about site proximity or organotropism was not provided to Metient, for these two patients it predicted metastasis-to-metastasis seeding to occur within an organ. For example, patient H103207 shows evidence of seeding within the brain, first to the right frontal lobe and then the cerebellum, as well as extensive seeding between the right and left lobes of the lung (Figure 4b). Patient H132384 shows evidence of seeding from bone-to-bone, first to the left cervical and secondarily to the chest wall (Figure 4c). Metastasizing cells exhibit organ-specific genetic and phenotypic changes to survive in a new microenvironment[36], suggesting that colonizing an additional tumor within the same organ microenvironment is more likely than a secondary migration from the primary adrenal tumor in these cases.

Next we compared the inferred migration histories from the NSCLC samples we analyzed, to an in-depth analysis of the same samples by the TRACERx consortium[14]. The TRACERx analysis assumes a primary single-source dissemination model, i.e., that metastases are only seeded from the lung, for its analysis of clonality and phyleticity. While Metient generally agrees with this dissemination model, Metient predicts metastasis-to-metastasis seeding for several patients (Figure 5a). CRUK0484 is one such patient where Metient proposes that an initial metastasizing clone to the rib leads to secondary metastasis formation in the scapula, which is a more plausible solution based on prior evidence that bone metastases prime and reprogram cells to form further secondary metastases[40,41].

When comparing the TRACERx classifications of clonality and phyleticity for each patient to those implied by Metient's highest-scoring solution, we find 91.3% agreement (115/126) in clonality (Figure 5c) and 92.7% agreement (114/123) in phyleticity (Figure 5d); three patients classified as "mixed" phyleticity by TRACERx were excluded). The discrepancies between these classifications stem from the way in which seeding clones are defined. While

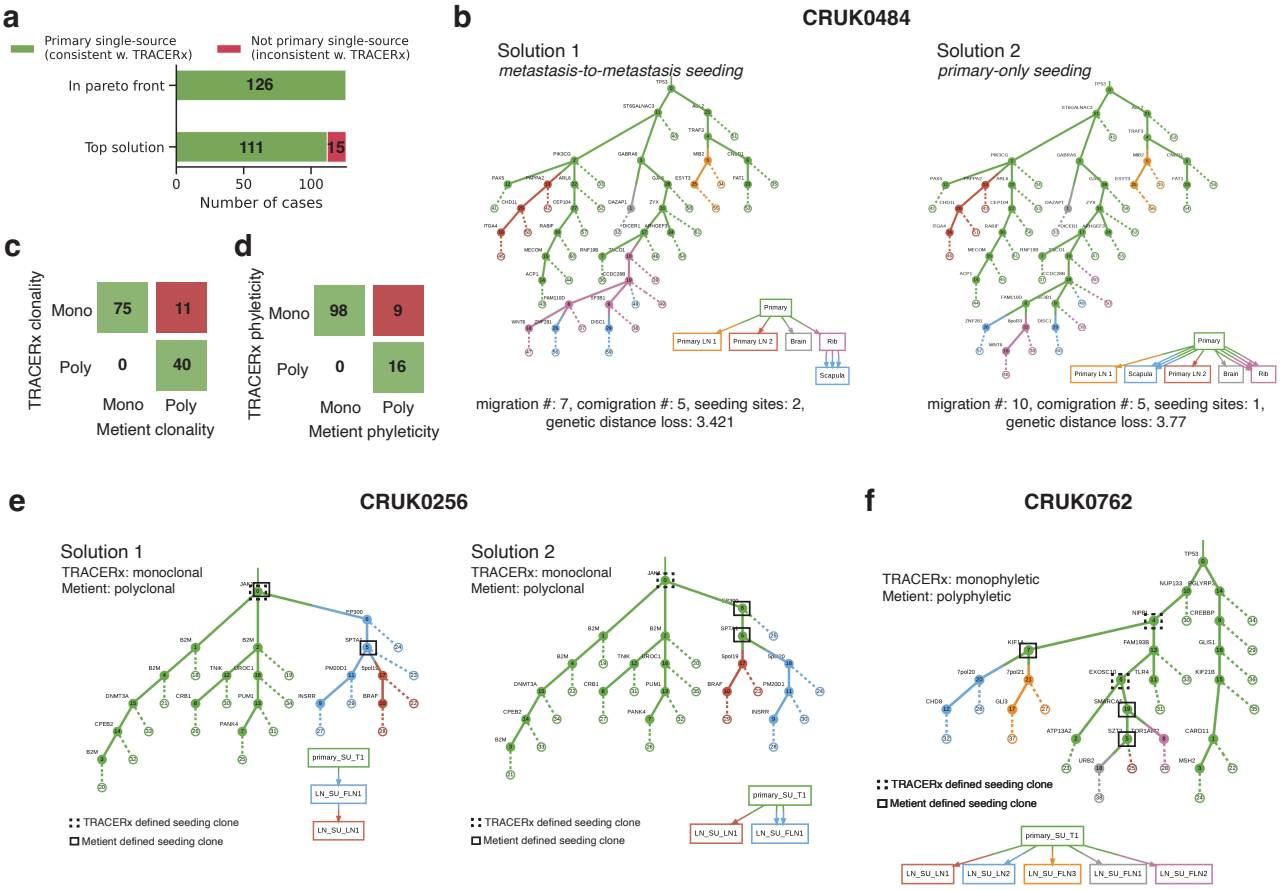

Figure 5. TRACERx NSCLC cohort. **(a)** The number of solutions that are classified as primary single-source (the assumed seeding pattern of TRACERx) vs. other when looking only at the Pareto optimal solutions vs. the top solution. **(b)** The top two Pareto optimal solutions for NSCLC patient CRUK0484 as ranked by Metient-calibrate. Comparison of Metient's inference to TRACERx's **(c)** clonality and **(d)** phyleticity classification. Numbers in boxes indicate the number of patients in agreement or disagreement. **(e)** All Pareto optimal solutions for NSCLC patient CRUK0256 as ranked by Metient-calibrate. **(f)** Patient CRUK0762 where seeding pattern and clonality are in agreement between Metient-calibrate but phyleticity differs due to which clones are classified as seeding.

TRACERx identifies shared clones between the primary and each metastasis, Metient uses the full migration history to define seeding clones. Therefore, in 11 cases we identify multiple seeding clones needed to explain the full migration history, which cannot be resolved from the TRACERx identified seeding clones alone. For example, for patient CRUK0256 (Figure 5e), only the root clone is shared between primary and metastases, making it the only seeding clone by TRACERx's definition. However, according to the clone tree and the observed presence of clone 6 in LN_SU_FLN1 and clone 5 in both LN_SU_FLN1 and LN_SU_LN1, we conclude that there must have been either a metastasis-to-metastasis seeding event (Figure 5e solution 1), or two clones originally from the primary (no longer detectable in those samples due to either ongoing evolution or undersampling) that seeded the metastases (Figure 5e solution 2). In either migration history, multiple clones had to participate in seeding in order to explain the clone tree and observed clones inferred from the sequencing data.

Not using the clone tree to determine seeding clones also impacts the inferred phyleticity, as the path connecting seeding clones is used to determine if metastatic competence arises once or multiple times. Because the number of seeding clones is underestimated in the TRACERx analysis, monoclonal seeding is inferred more often, automatically classifying these histories as monophyletic. However, we find nine cases where TRACERx classifies a patient as monophyletic and Metient classifies as polyphyletic; in such cases the multiple clones needed to explain seeding occur on separate paths of the clone tree (e.g. patient CRUK0762, Figure 5f).

Discussion

In this work, we proposed and extensively validated a novel framework for metastasis migration history inference, which previously required substantial expert input due to poorly specified parsimony models. Metient introduces two major advances: defining a Pareto front of possible alternative solutions using an innovative way to sample and optimize solutions of mixed-variable combinatorial optimization problems, as well as a step to calibrate these multiple solutions to independent evidence of correct migration histories, namely genetic distance and organotropism. These advances improve performance on simulated data compared to the state-of-the-art method, improve biological interpretation on real data, and sample multiple probable solutions in a fraction of the time. This scalability is essential for use cases such as application to single-cell sequencing data, which is becoming increasingly available and which can be inputted into Metient as is (as long as the observed clone proportions are provided).

While Metient scales well in compute time to large trees, for inputs with very large clone trees or many tumor samples, we recommend users run Metient multiple times to ensure the best results. Specifically, the use of the Gumbel-Softmax reparamaterization, which is a low variance but high bias estimate of the gradient, can cause the inference procedure to get stuck in local minima, although in practice, we've found that this impacts <1% of the real data evaluated in this study. Previous optimization methods also face convergence and runtime issues with large inputs, and Metient provides a means to address these problems via rerunning or increasing the number of samples used. Finally, our current method performs variant allele frequency (VAF) correction for SNVs based on copy number alterations (CNAs) that are clonal in a sample. Subclonal CNAs, such as subclonal deletions, are therefore not taken into account, but one could potentially handle this by using the descendant cell fraction (DCF)[42] or phylogenetic cancer cell fraction (phyloCCF)[43] to estimate observed clone proportions and input this into Metient.

The metastasis priors are a step towards identifying the correct seeding pattern for a cancer type when this is not known a priori, a critical problem in metastasis research. We show multiple examples of real patients where choosing the best migration history using inputted weights of maximum parsimony metrics can lead to ad hoc decisions between what biological processes in metastasis are more likely. Instead, we allow metastasis priors (genetic distance and organotropism) to make this decision for us. In this way, Metient finds solutions not anticipated by experts by removing the hard constraints on metastatic patterns used by previous studies. We show that the metastasis priors are informative and necessary in some cases for appropriate biological interpretation. While we have only explored the use of genetic distance and organotropism in this study, our framework is easily extensible to additional priors in the future. For example, our framework could be extended to use mutational signatures as a prior, since metastases exhibit shifts in mutational signature composition[44,45].

Applying Metient to four cancer cohorts with sampling of primary and metastatic tumors reveals predominant monophyleticity, suggesting that it is rare for metastatic potential to be gained in independent evolutionary trajectories, or that metastatic potential is gained early on in a cancer's evolution and maintained. In addition, a reanalysis of the data reveals that polyclonality is more common than anticipated due to a lack of sensitivity in

previous methods to detect all seeding clones. In conclusion, we show that Metient offers a fast and adaptable framework to leverage bulk DNA sequencing data to probe enduring questions in metastasis research.

Methods

Estimating observed clone proportions

The first step of Metient is to estimate the binary presence or absence of clone tree ($\mathbf{T}$) nodes in each site. The clone tree $\mathbf{T}$ can either be provided as input, or inferred from the DNA sequencing data using, e.g., Orchard[46], PairTree[47], SPRUCE[48], CITUP[49], or EXACT[50]. Building on a previous approach as described by Wintersinger et al.[47], Metient estimates the proportion of clones in each site using the input clone tree $\mathbf{T}$ and read count data from bulk DNA sequencing. For a genomic locus j in anatomical site k, the probability of observing read count data x_{kj} is defined using the following:

- V_{kj} is the number of reads that map to genomic locus j in anatomical site k with the variant allele
- R_{kj} is the number of reads that map to genomic locus j in anatomical site k with the reference allele
- ω_{kj} is a conversion factor from subclonal frequency to varaiant allele frequency (VAF) for genomic locus j in anatomical site k

Using a binomial model, we then estimate the proportion of anatomical site k containing clone c using $p(x_{kj}|(\mathbf{U}\,\mathbf{B})_{kj}) = \mathrm{Binom}(V_{kj}|V_{kj} + R_{kj}, \omega_{kj}(\mathbf{U}\,\mathbf{B})_{kj})$. $\mathbf{B} \in \{0,1\}^{C\times M}$ is 1:1 with a clone tree, where C is the number of clones and M is the number of mutations or mutation clusters, and $\mathbf{B}_{cm} = 1$ if clone c contains mutation m (Figure 1b). $\mathbf{U} \in [0,1]^{K\times C}$, where K is the number of anatomical sites, and $\mathbf{U}_{sc}$ is the fraction of anatomical site k that contains clone c (Figure 1b). An L1 regularization is used to promote sparsity, since we expect most values in $\mathbf{U}$ to be zero. For details on how to set ω_{kj}, see "Variant read probability calculation (ω)" in Supplementary Information.

We attach a witness node with label k (leaf nodes connected by dashed lines in Figure 1b, c) to clone c in clone tree $\mathbf{T}$, if $\mathbf{U}_{sc} > 5\%$ for a given anatomical site k and clone c. If a clone c does not make up 5% of any of the K anatomical sites, and c is a leaf node of the clone tree $\mathbf{T}$, we remove this node since it is not well estimated by the data. Alternatively, the binary presence or absence of clone tree nodes in anatomical sites can be provided as input, which we use to attach witness nodes instead of estimating $\mathbf{U}$.

Note that here the term "anatomical site" is used to describe a distinct tumor mass. If multiple samples are taken from the same tumor, we combine them as described in "Bulk DNA sequencing pre-processing: Non-small Cell Lung Cancer Dataset".

Labeling the clone tree

The next step in inferring a migration history is to jointly infer a labeling of the clone tree and resolve polytomies (nodes with more than two children). Polytomy resolution is discussed in the section "Resolving polytomies".

Since we are interested in identifying multiple hypotheses of metastatic spread, we aim to find multiple possible labelings of a clone tree $\mathbf{T}$. Each possible labeling is represented by a matrix $\mathbf{V} \in \{0,1\}^{K\times C}$, where K is the

number of anatomical sites and C is the number of clones, and $\mathbf{V}_{kc} = 1$ if clone c originated in anatomical site k. Each column of $\mathbf{V}$ is a one-hot vector. We solve for an individual $\mathbf{V}$ by optimizing the evidence lower bound, or ELBO, as defined by:

$$\text{ELBO}(q) = \mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})] + \mathbb{H}(\mathbf{V}) \tag{1}$$

Where $\mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})]$ evaluates a labeling based on maximum parsimony, genetic distance, and organotropism, and the second term is the entropy term. $\mathbf{U}$ has been optimized as described in the previous section "Estimating observed clone proportions", or taken as input from the user. See Supplementary Information for a full derivation of this objective. Because $\mathbf{V}$ is a matrix of discrete categorical variables, we do not optimize $\mathbf{V}$ directly, but rather the underlying probabilites of each category that we feed through a Gumbel-softmax estimator (see "Gumbel-softmax optimization").

Gumbel-softmax optimization

In the previous section, we described how to score the matrix representation of the labeled clone tree, $\mathbf{V}$. Here, we describe how to optimize $\mathbf{V}$. Starting with a matrix $\psi \in \{0,1\}^{K \times C}$, of randomly initialized values, where each column represents the class probabilities of clone c being labeled in site k:

1. At every iteration, take a sample from the Gumbel-Softmax distribution[24,25] independently for each clone:

$$v_i = \frac{\exp((\log(\psi_{ic}) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\psi_{jc}) + g_j)/\tau)} \qquad \text{for } i = 1, ..., k.$$

   where $g_1 ... g_k$ are i.i.d. samples drawn from Gumbel(0,1), and the k-dimensional sample vector y is a column of ψ.

2. Evaluate the ELBO by setting $\mathbf{V}$ to a discretized ψ (taking argmax along the columns, i.e. the straight-through estimator[24]), but using the continuous approximation of the gradients in the backward pass.

3. During training, start with a high τ to permit exploration, then gradually anneal τ to a small but non-zero value so that the Gumbel-Softmax distribution resembles a one-hot vector.

At the end of training, as τ approaches 0, ψ is a matrix where each column is a one-hot vector (the labeling of each clone). Therefore, we set $\mathbf{V} = \psi$ at the end of optimization. In order to capture multiple modes of the posterior distribution, we optimize multiple $\mathbf{V}$s in parallel. To do this, we set up steps 1-3 such that x ψs are solved for in parallel[51], where x is equal to the sample size and is calculated according to the size of the inputs ($\propto K^C$). See Supplementary Information for further explanation.

Resolving polytomies

An overview of the algorithm to resolve polytomies is given in Supplementary Figure S7a and b.

1. If a node i in $\mathbf{T}$ has more than 2 children, we create a new "resolver" node for every site where either i or i's children are observed in. Specifically, for every node i in $\mathbf{T}$, we look at the set of nodes P, which contains node i and node i's children. We then tally the anatomical sites of all witness nodes for nodes in P. If any anatomical site is counted at least twice, a resolver node with that anatomical site label is added as a new child of i. The genetic distance between the parent node i and its new resolver node is set to 0 since there are no observed mutations between the two nodes.

2. We allow the children of i to stay as a child of i, or become a child of one of the resolver nodes of i.

3. Any resolver nodes that are unused (i.e. have no children) or which do not improve the migration history (i.e. the parsimony metrics without the resolver node are the same or worse) are removed.

Fixing optimal subtrees

To improve convergence, we perform two rounds of optimization when solving for a labeled clone tree and resolving polytomies:

1. Solve for labeled trees and resolve polytomies jointly (as described in previous sections).

2. For each pair of labeled tree and polytomy resovled tree, find optimal subtrees. I.e., find the largest subtrees, as defined by the most number of nodes, where all labels for all nodes are equal. This means that there is no other possible optimal labeling for this subtree (there are 0 migrations, 0 comigrations, 0 seeding sites), and we can keep it fixed. Fix these nodes' labelings and adjacency matrix connections (if using polytomy resolution).

3. Repeat step 1 for any nodes that have not been fixed in step 2.

Metient-calibrate

In Metient-calibrate, we aim to find a weighting for the maximum parsimony metrics that best matches other biological data relevant to metastasis. We take the Pareto front of trees for each patient and score these trees based on (1) the maximum parsimony metrics and (2) the metastasis priors (genetic distance and organotropism). These form the parsimony distribution and metastasis prior distribution, respectively.

To score a migration history using genetic distance, we use the following equation: $\sum_{ij} -log(\mathbf{D_{ij}})\mathbf{K}_{ij}$, where $\mathbf{D}$ contains the normalized number of mutations between clones, and $\mathbf{K} = 1$ if clone i is the parent of clone j and clone i and clone j have different anatomical site labels.

To score a migration history using organotropism, we use the following equation: $\sum_{i=1}^{K} -log(\mathbf{o_i})\mathbf{g_i}$, where vector $\mathbf{o}$ contains the frequency at which the primary seeds other anatomical sites, and vector $\mathbf{g}$ contains the number of migrations from the primary site to all other anatomical sites for a particular migration history.

We start with equal weighting of the parsimony metrics, and using gradient descent, minimize the cross entropy loss between the parsimony distribution and metastasis prior distribution for all patients in the cohort. These optimized parsimony weights are used to rank the solutions on the Pareto front, and genetic distance and organotropism are used to break ties between equally parsimonious migration histories. See Supplementary Information for further derivation.

Metient-evaluate

In Metient-evaluate, weights for each maximum parsimony metric (migrations, comigrations, seeding sites) and optionally, genetic distance and organotropism, are taken as input. These weights are used to rank the solutions on the Pareto front. If no weights are inputted, we provide pre-calibrated weights from the four cancer types/datasets discussed in this work.

Evaluations on simulated data

We use the simulated data for 80 patients provided by MACHINA[15] to benchmark our method's performance. All performance scores are reported using MACHINA's PMH-TI mode and Metient-calibrate with a sample size of 1024, both with default configurations. We do not use polytomy resolution for Metient-calibrate in these results, since we do not find that it improves performance on simulated data (Supplementary Tables **??**).

Evaluation metrics. We calculate migration graph and migrating clones F1-scores the same way as MACHINA. Using an inferred migration graph $\mathbf{G}$ and comparing it to the ground truth migration graph $\mathbf{G}^*$, recall and precision are calculated as follows:

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

where $E(\mathbf{G})$ are the edges of $\mathbf{G}$, and multiple edges between the same two sites are included in $E(\mathbf{G})$. Recall and precision of the migrating clones in the inferred migration history (which includes inference of both the clone tree labeling and observed clone proportions) is calculated as follows:

$$\text{recall} = \frac{|C(\mathbf{U}, \mathbf{V}) \cap C(\mathbf{U}^*, \mathbf{V}^*)|}{|C(\mathbf{U}^*, \mathbf{V}^*)|} \quad \text{precision} = \frac{|C(\mathbf{U}, \mathbf{V}) \cap C(\mathbf{U}^*, \mathbf{V}^*)|}{|C(\mathbf{U}, \mathbf{V})|}$$

where $C(\mathbf{U}, \mathbf{V})$ is the set of mutations that have an outgoing migration edge. For example, $C(\mathbf{U}, \mathbf{V}) = \mathrm{A}, \mathrm{B}, \mathrm{C}$ in solution 1 of Figure 1c.

Timing benchmarks. All timing benchmarks (Figure 2e) were run on 8 Intel(R) Xeon(R) CPU E5-2697 v4 @ 2.30GHz CPU cores with 8 gigabytes of RAM per core. Runtime of each method is the time needed to run inference and save dot files of the inferred migration histories (and for Metient, an additional serialized file with the results of the top k migration histories). We compare MACHINA's PMH-TI mode to Metient-calibrate, both with default configurations. These are the same modes used to report comparisons in F1-scores. Each value in Figure 2e is the time needed to

run one patient's tree. Because Metient-calibrate has an additional inference step where parsimony metric weights are fit to a cohort, we take the time needed for this additional step and divide it by the number of patient trees in the cohort, and add this time to each patient's migration history runtime.

Defining clonality and phyleticity

In order to rectify different meanings of the terms "monoclonal" and "polyclonal" used in previous work, we define two terms:

- genetic clonality: if all sites are seeded by the same seeding clone, this patient is genetically monoclonal, otherwise, genetically polyclonal.
- site clonality: if each site is seeded by one seeding clone, but not necessarily the same seeding clone, this patient is site monoclonal, otherwise, site polyclonal.

A seeding clone is a node in a migration history whose child is a different color than itself. These definitions are depicted in Figure 3c. We emphasize that while monoclonal and polyclonal have previously been used interchangeabley for either definition, they depict fundamentally distinct biological phenomena. For example, site clonality does not capture the fact that there might be site-specific mutations that are needed for colonization in that organ.

To define phyleticity, we first extract all seeding clones from a migration history. We then identify the seeding clone closest to the root, s, i.e. the first seeding clone when doing a breadth first search of the clone tree. From s, if all other seeding clones can be reached, i.e., they are descendants of the tree rooted at s, the migration history is monophyletic, otherwise, it is polyphyletic. This captures the fact that when a tree is monophyletic, there are no independent evolutionary trajectories that give rise to seeding clones.

In order to accurately compare our phyleticity measurements to TRACERx, we use their definition in Figure 5c and the TRACERx comparison analysis. To apply their definition to our migration histories, we extract seeding clones as described above, and then determine if there is a Hamiltonian path in the clone tree that connects the seeding clones. If such a Hamiltonian path exists, we call this migration history monophyletic under the TRACERx definition, and polyphyletic otherwise.

Extracting organotropism data from MSK-MET

Data from the MSK-MET study[27] for 25,775 patients with annotations of distant metastases locations was downloaded from the publicly available cbioportal[52]. Each patient had annotations of one of 27 primary cancer types and the presence or absence of a metastasis in one of 21 distant anatomical sites. The original authors extracted this data from electronic health records and mapped it to a reference set of anatomical sites. We sum over all patients to build a 27 x 21, cancer type by metastatic site occurrence matrix. We then normalize the rows to turn these into frequencies. We interpret these frequencies as a "normalized time to metastasis", and only weigh

migrations from the primary site to other sites, because there is no data to indicate frequencies of seeding from metastatic sites to other metastatic sites, or back to the primary. We make this data available for users, with the option to input your own organotropism vector for each patient as well.

Data availability

The HR-NB dataset was accessed from the NCI's Cancer Research Data Commons (https://datacommons.cancer.gov) under the study phs03111.v1.p1. The anatomical site labels for TRACERx patients used data generated by The TRAcking Non-small Cell Lung Cancer Evolution Through Therapy (Rx) (TRACERx) Consortium and provided by the UCL Cancer Institute and The Francis Crick Institute. The TRACERx study is sponsored by University College London, funded by Cancer Research UK and coordinated through the Cancer Research UK and UCL Cancer Trials Centre. The organotropism matrix derived from MSK-MET is available at https://github.com/morrislab/metient/blob/main/metient/data/msk_met/msk_met_freq_by_cancer_type.csv. The following publicly available datasets were used: melanoma[9], breast[21], HGSOC[10], NSCLC[14], MSK-MET[27].

Code availability

Metient is available as a software package installable with pip at https://github.com/morrislab/metient/. Tutorials for usage can be found at https://github.com/morrislab/metient/tree/main/tutorial. Code to reproduce figures from this manuscript can be found at https://github.com/morrislab/metient/tree/main/metient/jupyter_notebooks.

Bibliography

1. Karuna Ganesh and Joan Massagué. Targeting metastatic cancer. *Nature medicine*, 27(1):34–44, 2021.
2. Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose MC Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini ML Kallio, Gunilla Högnäs, Matti Annala, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015.
3. Nicolai J Birkbak and Nicholas McGranahan. Cancer genome evolutionary trajectories in metastasis. *Cancer cell*, 37(1):8–19, 2020.
4. Q Wei, Z Ye, X Zhong, L Li, C Wang, RE Myers, JP Palazzo, D Fortuna, A Yan, SA Waldman, et al. Multiregion whole-exome sequencing of matched primary and metastatic tumors revealed genomic heterogeneity and suggested polyclonal seeding in colorectal cancer metastasis. *Annals of oncology*, 28(9):2135–2141, 2017.
5. Zheng Hu, Jie Ding, Zhicheng Ma, Ruping Sun, Jose A Seoane, J Scott Shaffer, Carlos J Suarez, Anna S Berghoff, Chiara Cremolini, Alfredo Falcone, et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nature genetics*, 51(7):1113–1122, 2019.
6. Zheng Hu, Zan Li, Zhicheng Ma, and Christina Curtis. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nature genetics*, 52(7):701–708, 2020.
7. Gunes Gundem, Max F Levine, Stephen S Roberts, Irene Y Cheung, Juan S Medina-Martínez, Yi Feng, Juan E Arango-Ossa, Loic Chadoutaud, Mathieu Rita, Georgios Asimomitis, et al. Clonal evolution during metastatic spread in high-risk neuroblastoma. *Nature Genetics*, pages 1–12, 2023.
8. David Brown, Dominiek Smeets, Borbála Székely, Denis Larsimont, A Marcell Szász, Pierre-Yves Adnet, Françoise Rothé, Ghizlane Rouas, Zsófia I Nagy, Zsófia Faragó, et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nature communications*, 8(1):14944, 2017.
9. 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, et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proceedings of the National Academy of Sciences*, 112(35):10995–11000, 2015.
10. Andrew McPherson, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature genetics*, 48(7):758–767, 2016.
11. Priscilla K Brastianos, Scott L Carter, Sandro Santagata, Daniel P Cahill, Amaro Taylor-Weiner, Robert T Jones, Eliezer M Van Allen, Michael S Lawrence, Peleg M Horowitz, Kristian Cibulskis, et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer discovery*, 5(11):1164–1177, 2015.
12. Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Tim Chambers, Jose I Lopez, David Nicol, Tim O'Brien, James Larkin, Stuart Horswell, et al. Tracking cancer evolution reveals constrained routes to metastases: Tracerx renal. *Cell*, 173(3):581–594, 2018.
13. Ayesha Noorani, Xiaodun Li, Martin Goddard, Jason Crawte, Ludmil B Alexandrov, Maria Secrier, Matthew D Eldridge, Lawrence Bower, Jamie Weaver, Pierre Lao-Sirieix, et al. Genomic evidence supports a clonal diaspora model for metastases of esophageal adenocarcinoma. *Nature genetics*, 52(1):74–83, 2020.
14. Maise Al Bakir, Ariana Huebner, Carlos Martínez-Ruiz, Kristiana Grigoriadis, Thomas B. K. Watkins, Oriol Pich, David A. Moore, Selvaraju Veeriah, Sophia Ward, Joanne Laycock, and et al. The evolution of non-small cell lung cancer metastases in tracerx. *Nature*, Apr 2023.
15. Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature genetics*, 50(5):718–726, 2018.

16. Maxime Tarabichi, Adriana Salcedo, Amit G Deshwar, Máire Ni Leathlobhair, Jeff Wintersinger, David C Wedge, Peter Van Loo, Quaid D Morris, and Paul C Boutros. A practical guide to cancer subclonal reconstruction from dna sequencing. *Nature methods*, 18(2):144–155, 2021.
17. HX Dang, BS White, SM Foltz, CA Miller, Jingqin Luo, RC Fields, and CA Maher. Clonevol: clonal ordering and visualization in cancer sequencing. *Annals of oncology*, 28(12):3076–3082, 2017.
18. 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(1):14114, 2017.
19. Chong Zhang, Lin Zhang, Tianlei Xu, Ruidong Xue, Liang Yu, Yuelu Zhu, Yunlong Wu, Qingqing Zhang, Dongdong Li, Shuohao Shen, et al. Mapping the spreading routes of lymphatic metastases in human colorectal cancer. *Nature communications*, 11(1):1993, 2020.
20. Kamila Naxerova, Johannes G Reiter, Elena Brachtel, Jochen K Lennerz, Marc Van De Wetering, Andrew Rowan, Tianxi Cai, Hans Clevers, Charles Swanton, Martin A Nowak, et al. Origins of lymphatic and distant metastases in human colorectal cancer. *Science*, 357(6346):55–60, 2017.
21. 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, et al. Tumor evolution in two patients with basal-like breast cancer: a retrospective genomics study of multiple metastases. *PLoS medicine*, 13(12):e1002174, 2016.
22. Joseph E Stiglitz. Pareto optimality and competition. *The Journal of Finance*, 36(2):235–251, 1981.
23. Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.
24. Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
25. Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
26. Nathan E Reticker-Flynn, Weiruo Zhang, Julia A Belk, Pamela A Basto, Nichole K Escalante, Genay OW Pilarowski, Alborz Bejnood, Maria M Martins, Justin A Kenkel, Ian L Linde, et al. Lymph node colonization induces tumor-immune tolerance to promote distant metastasis. *Cell*, 185(11):1924–1942, 2022.
27. Bastien Nguyen, Christopher Fong, Anisha Luthra, Shaleigh A Smith, Renzo G DiNatale, Subhiksha Nandakumar, Henry Walch, Walid K Chatila, Ramyasree Madupuri, Ritika Kundra, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell*, 185(3):563–575, 2022.
28. Ernst Lengyel. Ovarian cancer development and metastasis. *The American journal of pathology*, 177(3):1053–1064, 2010.
29. Anirban K Mitra. *Ovarian cancer metastasis: a unique mechanism of dissemination*. IntechOpen, 2016.
30. Philippe Gui and Trever G Bivona. Evolution of metastasis: New tools and insights. *Trends in Cancer*, 8(2):98–109, 2022.
31. Zachary R Chalmers, Caitlin F Connelly, David Fabrizio, Laurie Gay, Siraj M Ali, Riley Ennis, Alexa Schrock, Brittany Campbell, Adam Shlien, Juliann Chmielecki, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome medicine*, 9:1–14, 2017.
32. Thomas J Mitchell, Samra Turajlic, Andrew Rowan, David Nicol, James HR Farmery, Tim O'Brien, Inigo Martincorena, Patrick Tarpey, Nicos Angelopoulos, Lucy R Yates, et al. Timing the landmark events in the evolution of clear cell renal cell cancer: Tracerx renal. *Cell*, 173(3):611–623, 2018.
33. Clemency Jolly and Peter Van Loo. Timing somatic events in the evolution of cancer. *Genome biology*, 19:1–9, 2018.
34. Ditte S Christensen, Johanne Ahrenfeldt, Mateo Sokač, Judit Kisistók, Martin K Thomsen, Lasse Maretty, Nicholas McGranahan, and Nicolai J Birkbak. Treatment represents a key driver of metastatic cancer evolution. *Cancer Research*, 82(16):2918–2927, 2022.

35. Francisco Martínez-Jiménez, Ali Movasati, Sascha Remy Brunner, Luan Nguyen, Peter Priestley, Edwin Cuppen, and Arne Van Hoeck. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature*, pages 1–9, 2023.
36. Yang Gao, Igor Bado, Hai Wang, Weijie Zhang, Jeffrey M Rosen, and Xiang H-F Zhang. Metastasis organotropism: redefining the congenial soil. *Developmental cell*, 49(3):375–391, 2019.
37. Ellsworth C Alvord. Why do gliomas not metastasize? *Archives of Neurology*, 33(2):73–75, 1976.
38. Sudhir Kumar, Antonia Chroni, Koichiro Tamura, Maxwell Sanderford, Olumide Oladeinde, Vivian Aly, Tracy Vu, and Sayaka Miura. Pathfinder: Bayesian inference of clone migration histories in cancer. *Bioinformatics*, 36(Supplement_2):i675–i683, 2020.
39. Ingrid H Wolf, Erika Richtig, Daisy Kopera, and Helmut Kerl. Locoregional cutaneous metastases of malignant melanoma and their management. *Dermatologic surgery*, 30:244–247, 2004.
40. Igor L Bado, Weijie Zhang, Jingyuan Hu, Zhan Xu, Hai Wang, Poonam Sarkar, Lucian Li, Ying-Wooi Wan, Jun Liu, William Wu, et al. The bone microenvironment increases phenotypic plasticity of er+ breast cancer cells. *Developmental cell*, 56(8):1100–1117, 2021.
41. Weijie Zhang, Igor L Bado, Jingyuan Hu, Ying-Wooi Wan, Ling Wu, Hai Wang, Yang Gao, Hyun-Hwan Jeong, Zhan Xu, Xiaoxin Hao, et al. The bone microenvironment invigorates metastatic seeds for further dissemination. *Cell*, 184(9):2471–2486, 2021.
42. Gryte Satas, Simone Zaccaria, Mohammed El-Kebir, and Benjamin J Raphael. Decifering the elusive cancer cell fraction in tumor heterogeneity and evolution. *Cell systems*, 12(10):1004–1018, 2021.
43. Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.
44. Charles W Ashley, Arnaud Da Cruz Paula, Rahul Kumar, Diana Mandelker, Xin Pei, Nadeem Riaz, Jorge S Reis-Filho, and Britta Weigelt. Analysis of mutational signatures in primary and metastatic endometrial cancer reveals distinct patterns of dna repair defects and shifts during tumor progression. *Gynecologic oncology*, 152(1):11–19, 2019.
45. Lindsay Angus, Marcel Smid, Saskia M Wilting, Job van Riet, Arne Van Hoeck, Luan Nguyen, Serena Nik-Zainal, Tessa G Steenbruggen, Vivianne CG Tjan-Heijnen, Mariette Labots, et al. The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nature genetics*, 51(10):1450–1458, 2019.
46. Ethan Kulman, Rui Kuang, and Quaid Morris. Orchard: building large cancer phylogenies using stochastic combinatorial search. *arXiv preprint arXiv:2311.12917*, 2023.
47. Jeff A Wintersinger, Stephanie M Dobson, Ethan Kulman, Lincoln D Stein, John E Dick, and Quaid Morris. Reconstructing complex cancer evolutionary histories from multiple bulk dna samples using pairtreereconstructing cancer evolutionary histories using pairtree. *Blood Cancer Discovery*, pages OF1–OF12, 2022.
48. 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, 2016.
49. Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.
50. Surjyendu Ray, Bei Jia, Sam Safavi, Tim van Opijnen, Ralph Isberg, Jason Rosch, and José Bento. Exact inference under the perfect phylogeny model. *arXiv preprint arXiv:1908.08623*, 2019.
51. Yaoxin Li, Jing Liu, Guozheng Lin, Yueyuan Hou, Muyun Mou, and Jiang Zhang. Gumbel-softmax-based optimization: a simple general framework for optimization problems on graphs. *Computational Social Networks*, 8(1):1–16, 2021.
52. Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the

cbioportal. *Science signaling*, 6(269):pl1–pl1, 2013.

53. David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.
54. 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, 2014.
55. Serena Nik-Zainal, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

Acknowledgments

We thank Karuna Ganesh and Julia Simundza for their valuable feedback on this manuscript, and Deeksha Madala for coming up with the method name Metient. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 227260-01 (D.K.). Q. M. holds a Canada CIFAR AI chair.

Supplementary Figures

Previous Methods for Migration History Inference

Method	Ancestral migration history	Subclonal presence in sites	Models Complex Seeding	Multiple solutions	Organotropism	Genetic Distance	Polytomy Resolution
ClonEvol[17]	Y	Y	N	Y	N	N	N
Treeomics[18]	N	Y	N	N	N	N	N
MACHINA[15]	Y	Y	Y	N	N	N	Y
PathFinder[38]	Y	N	N	Y	N	Y	Y
Metient	Y	Y	Y	Y	Y	Y	Y

Table 1. Summary of previous methods which perform at least one inference task of migration history. Ancestral migration history refers to the inference of internal vertices in the clone tree (e.g. labeling subclone AB as originating in lung in the top dashed box under output in Figure 1c). Subclonal presence in sites refers to the inference of the leaf nodes (e.g. identifying that subclone ABC is present in both lung and pancreas in the bottom dashed box under output in Figure 1c). Multiple solutions indicates whether a method only outputs one solution when multiple solutions are equally likely.

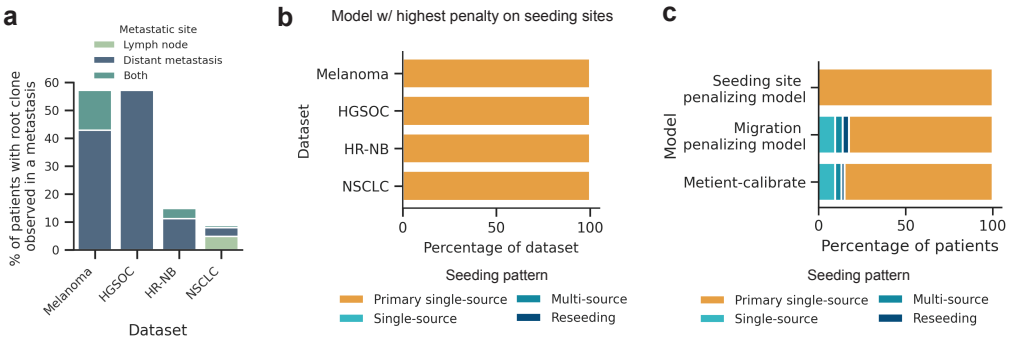

Figure S1. (a) Percent of patients in each dataset with the root cancerous clone observed in a metastatic site. **(b)** The distribution of seeding patterns in each dataset when taking the migration history on the approximate Pareto front with the lowest number of seeding sites, run with Metient-calibrate. **(c)** The distribution of seeding patterns across all patients if we choose the migration history on the Pareto front with the lowest number of seeding sites (Seeding site penalizing model), lowest number of migrations (Migration penalizing model), or the top Metient-calibrate solution.

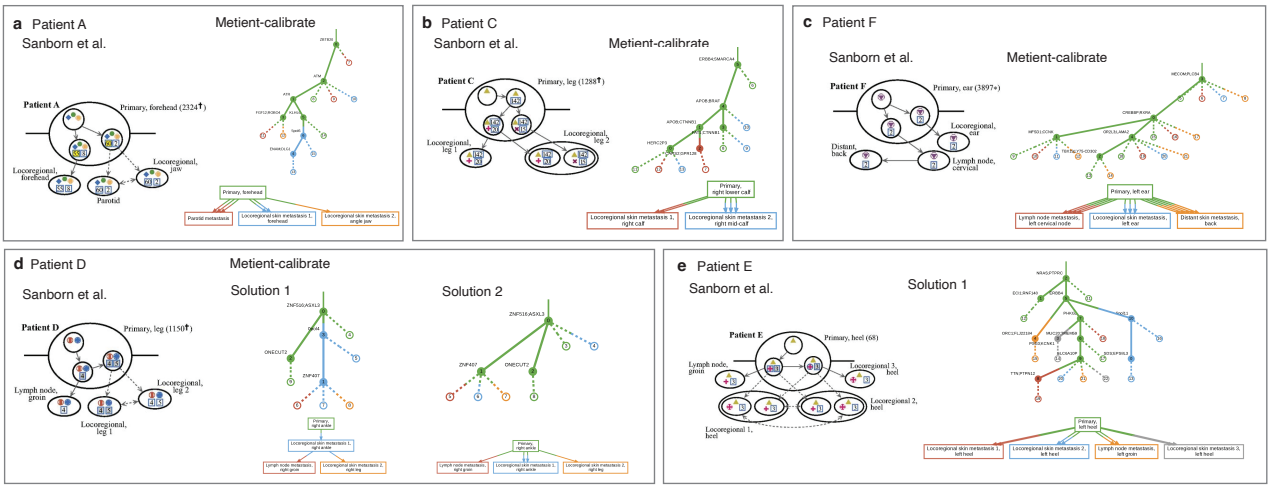

Figure S2. Comparison of Sanborn et al.[9] reported histories and Metient-calibrate inferred histories.

Average migrating clone F1-scores

Method	mS	pS	pM	pR	Macro-F1	Micro-F1
Evaluate (MP)	0.770	0.785	0.825	0.765	0.786	0.788
Evaluate (MP) + polyres	0.813	0.817	0.838	0.822	0.823	0.824
Evaluate (GD)	0.954	0.936	0.923	0.905	0.929	**0.926**
Evaluate (GD) + polyres	0.960	0.948	0.926	0.887	**0.930**	0.925
Evaluate (MP+GD)	0.973	**0.918**	0.918	0.892	0.925	0.919
Evaluate (MP+GD) + polyres	0.944	0.898	0.889	0.904	0.909	0.904
Calibrate	**0.983**	0.917	**0.928**	0.893	**0.930**	0.924
Calibrate + polyres	0.959	0.901	0.869	**0.895**	0.906	0.899
MACHINA	0.965	0.874	0.907	0.824	0.893	0.883

Table 2. Average F1-scores of migrating clones for each seeding pattern on simulated data. All Metient models were run with a sample size of 1024. When multiple solutions are found for a given input, all lowest loss solutions were taken. Evaluate (MP): Metient in evaluate mode with maximum parsimony only. Evaluate (GD): Metient in evaluate mode with genetic distance only. Evaluate (MP+GD): Metient in evaluate mode with maximum parsimony and genetic distance. Calibrate: Metient in calibrate mode, using genetic distance as the metastasis prior. polyres: polytomy resolution is used. mS: monoclonal single-source seeding. pS: polyclonal single-source seeding. pM: polyclonal multi-source seeding. pR: polyclonal reseeding.

Average migration graph F1-scores

Method	mS	pS	pM	pR	Macro-F1	Micro-F1
Evaluate (MP)	0.818	0.796	0.656	0.650	0.730	0.716
Evaluate (MP) + polyres	0.869	0.730	0.695	0.579	0.718	0.697
Evaluate (GD)	0.819	0.818	0.643	0.673	0.738	0.725
Evaluate (GD) + polyres	0.799	0.777	0.568	0.624	0.692	0.675
Evaluate (MP+GD)	0.849	**0.821**	0.689	**0.662**	0.755	0.740
Evaluate (MP+GD) + polyres	0.897	0.744	0.692	0.599	0.733	0.710
Calibrate	0.852	0.817	**0.719**	0.657	**0.761**	**0.747**
Calibrate + polyres	0.893	0.761	0.657	0.556	0.717	0.691
MACHINA	**0.903**	0.757	0.627	0.542	0.707	0.679

Table 3. Average F1-scores of migration graph for each seeding pattern on simulated data. AAll Metient models were run with a sample size of 1024. When multiple solutions are found for a given input, all lowest loss solutions were taken. Evaluate (MP): Metient in evaluate mode with maximum parsimony only. Evaluate (GD): Metient in evaluate mode with genetic distance only. Evaluate (MP+GD): Metient in evaluate mode with maximum parsimony and genetic distance. Calibrate: Metient in calibrate mode, using genetic distance as the metastasis prior. polyres: polytomy resolution is used. mS: monoclonal single-source seeding. pS: polyclonal single-source seeding. pM: polyclonal multi-source seeding. pR: polyclonal reseeding.

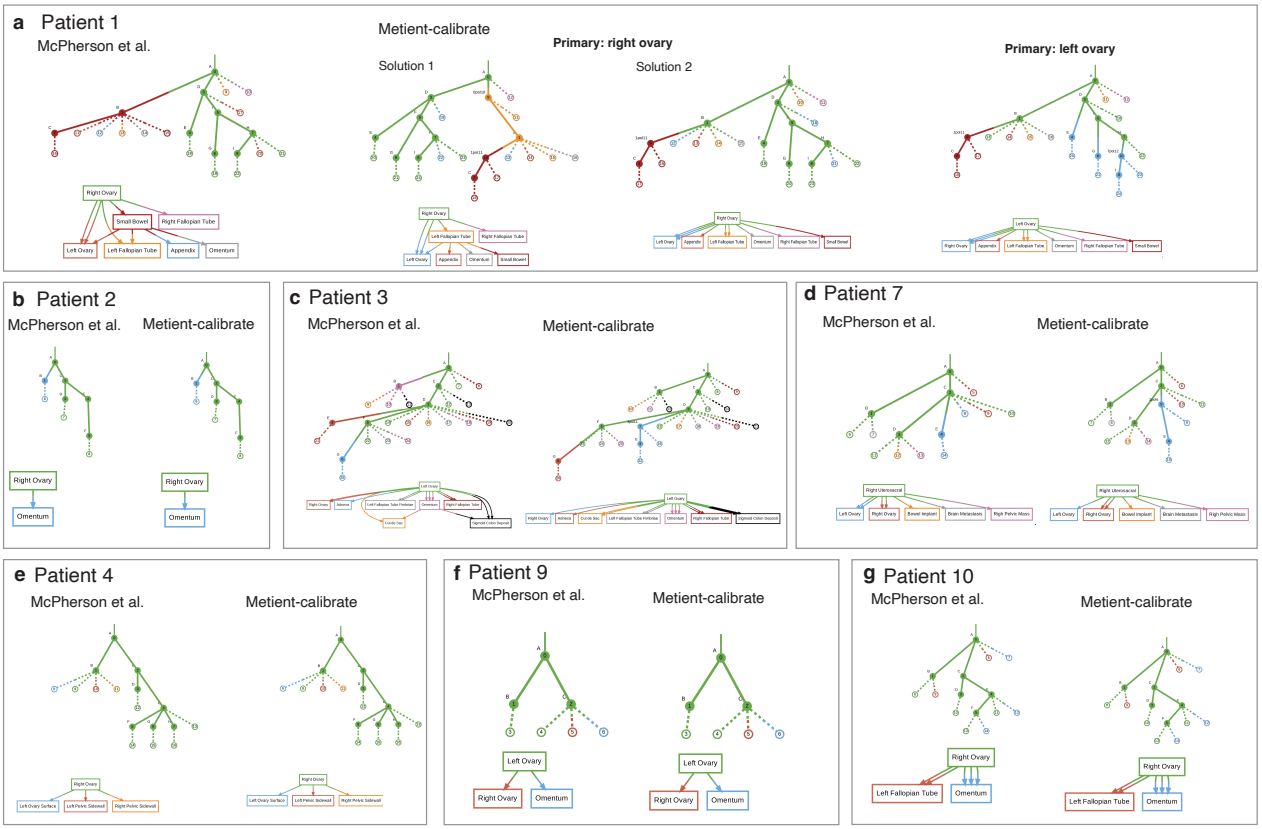

Figure S3. Comparison of McPherson et al.[10] reported histories and Metient-calibrate inferred histories. When multiple possible primaries were available, Metient-calibrate was run once with each possible primary. The most parsimonious possible solution of these is shown first.

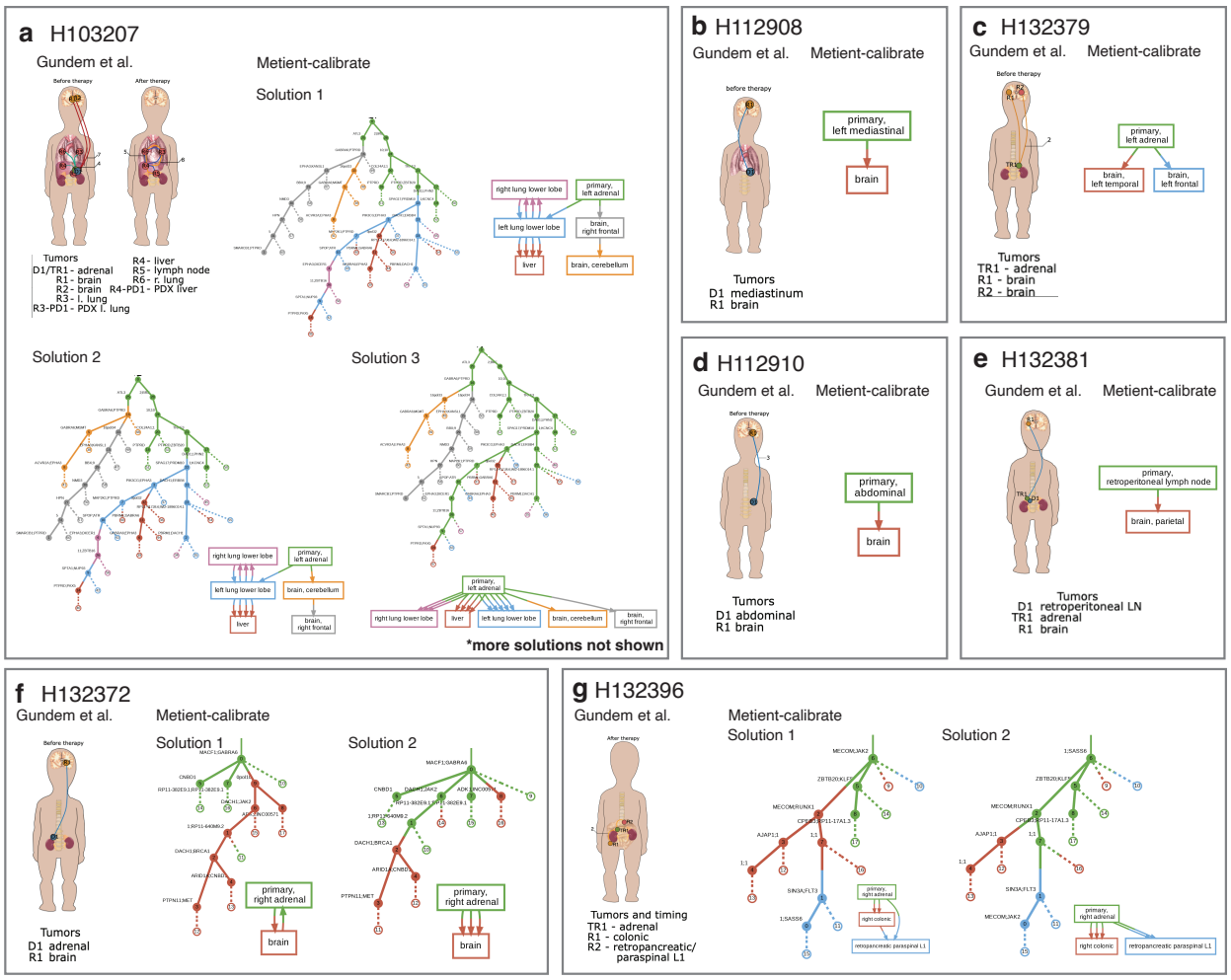

Figure S4. Comparison of Gundem et al.[7] reported body maps (left of each square) and Metient-calibrate inferred histories.

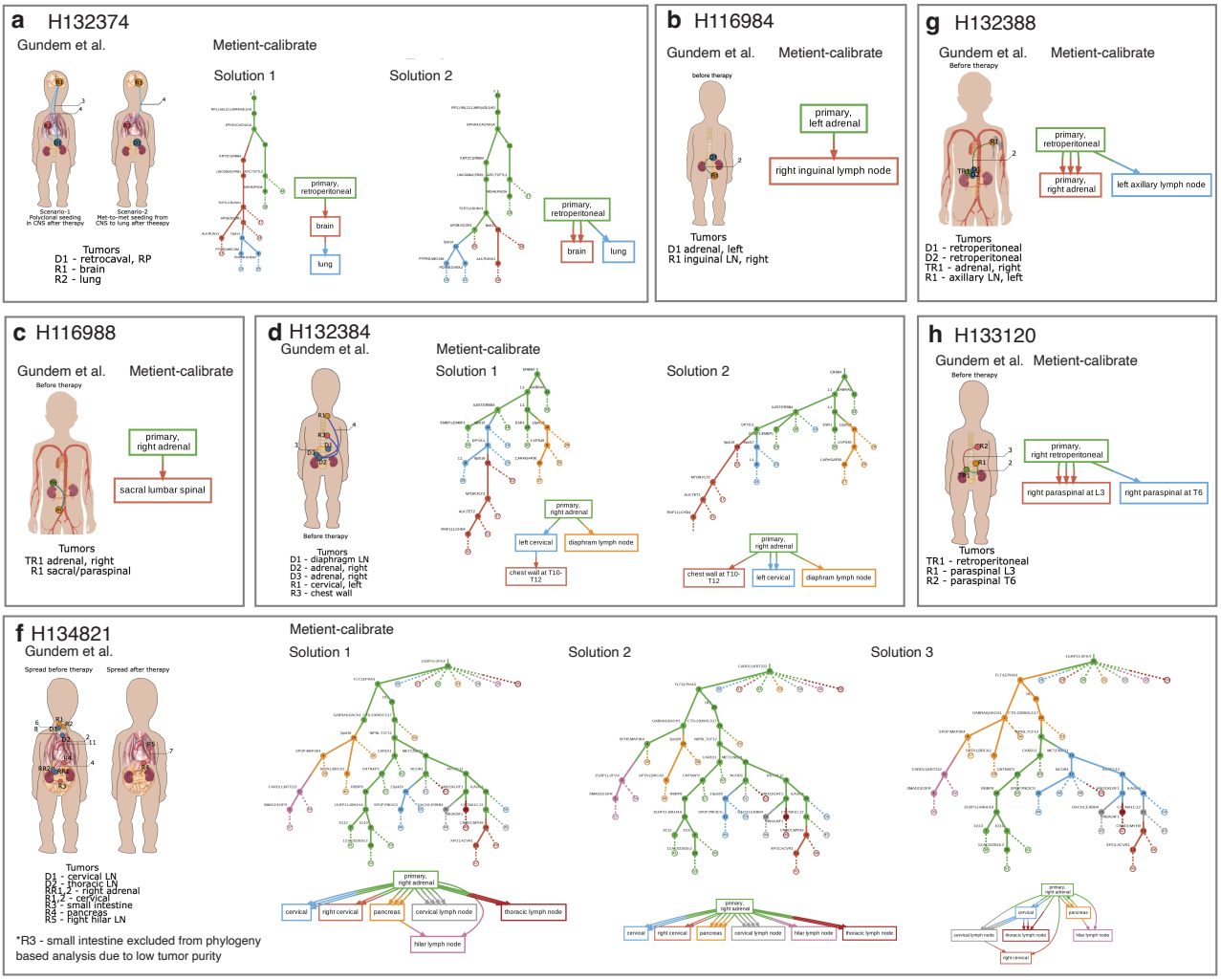

Figure S5. Comparison of Gundem et al.[7] reported body maps (left of each square) and Metient-calibrate inferred histories.

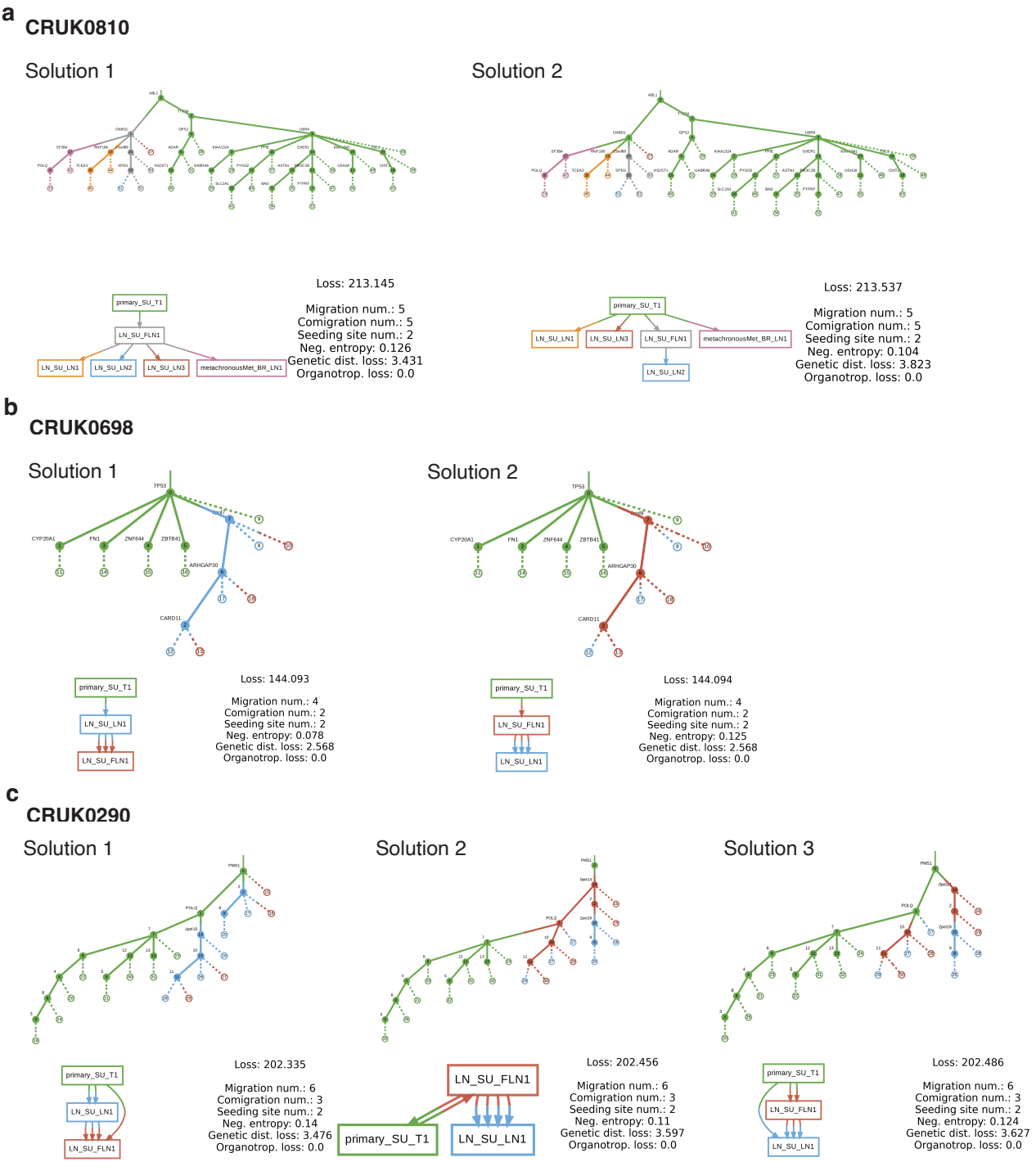

Figure S6. Examples of NSCLC patients with multiple migration histories on the Pareto front with the same parsimony metrics, but different migration graphs.

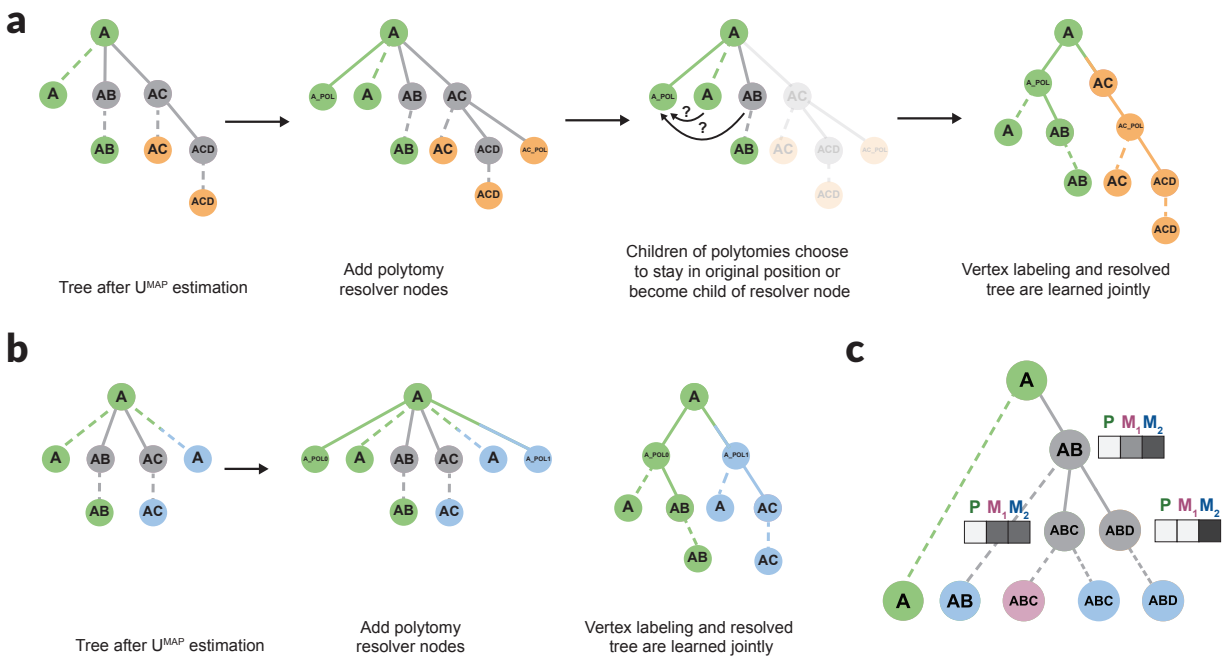

Figure S7. **(a)** Polytomy resolution algorithm with two nodes (A and AC) that have polytomies that can be resolved. **(b)** Polytomy resolution algorithm for a single node with four children and thus two resolver nodes. **(c)** Weight initialization is done such that nodes start with higher probabilities of being in the same site as the site that they or their children are detected in (after $\mathbf{U}^{\mathrm{MAP}}$ estimation).

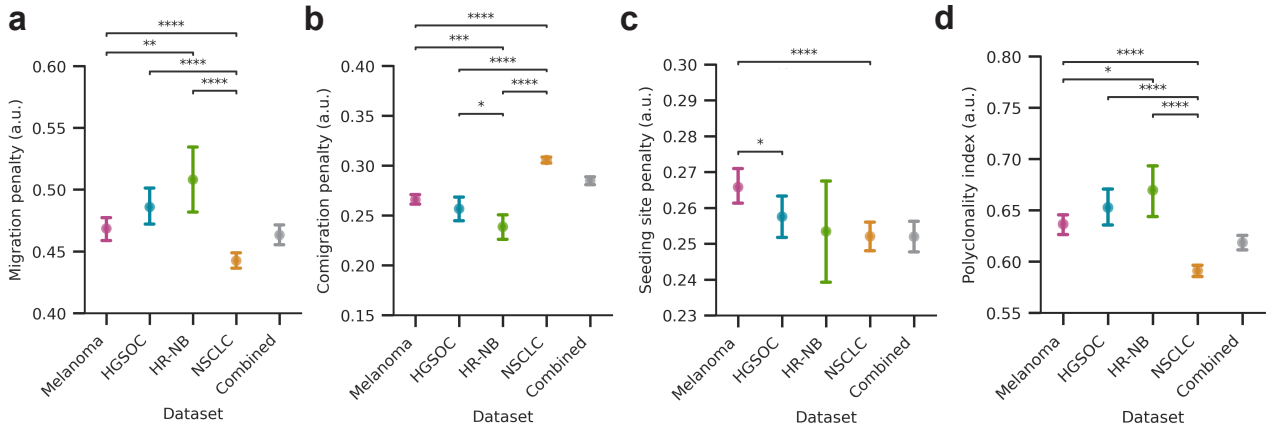

Figure S8. The **(a)** migration penalty/weight, **(b)** comigration penalty/weight, and **(c)** seeding site penalty/weight for each cohort, when taking 100 bootstrap samples of each cohort and fitting the weights to the bootstrapped sample. **(d)** The polyclonality index, which is $1 - (w_c/(w_m + w_c))$, where w_m is the migration penalty/weight and w_c is the comgiration penalty/weight. Statistical significance tested through a Welch's t-test; ns: 5e-02 < p <= 1, *: 1e-02 < p <= 5e-02, **: 1e-03 < p <= 1e-02, ***: 1e-04 < p <= 1e-03, ****: p <= 1e-04.

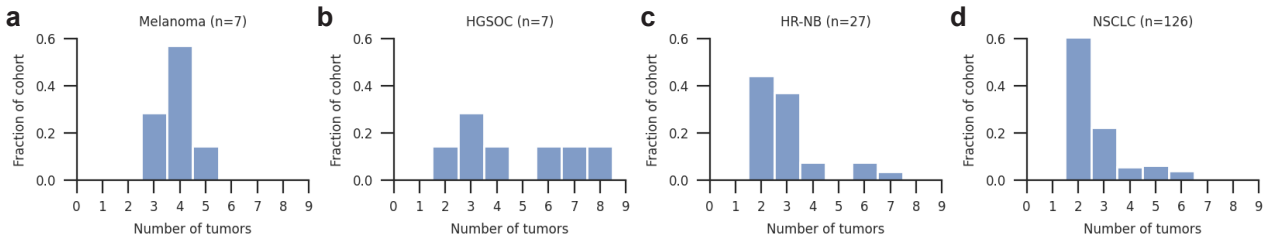

Figure S9. The distribution of tumors (number of distinct anatomical sites) for each cohort: **(a)** melanoma, **(b)** high-grade serous ovarian cancer (HGSOC), **(a)** high-risk neuroblastoma (HR-NB) and **(a)** non-small cell lung cancer (NSCLC).

Supplementary Information

A. Evaluating migration histories

We present our technique for optimizing migration histories in the context of variational inference. Our goal is to approximate the conditional density of latent variable $\mathbf{V}$ given observed variables $\mathbf{U}$ and $\mathbf{T}$: $p(\mathbf{V} \mid \mathbf{U}, \mathbf{T})$. $\mathbf{U}$ has been optimized as described in the section "Estimating observed clone proportions" in Methods. $p(\mathbf{V} \mid \mathbf{U}, \mathbf{T})$ can be written as:

$$p(\mathbf{V} \mid \mathbf{U}, \mathbf{T}) = \frac{p(\mathbf{U}, \mathbf{T} \mid \mathbf{V}) p(\mathbf{V})}{p(\mathbf{U}, \mathbf{T})} \tag{S1}$$

We cannot calculate the denominator, or the evidence, as its derivation is intractable (there are many possible values of $\mathbf{V}$):

$$p(\mathbf{U}, \mathbf{T}) = \sum_{\mathbf{V}} p(\mathbf{U}, \mathbf{T}, \mathbf{V}) \tag{S2}$$

We approximate the posterior distribution $p(\mathbf{V} \mid \mathbf{U}, \mathbf{T})$ with a simpler distribution $q(\mathbf{V})$, and we aim to minimize the Kullback-Leibler (KL) divergence between $q(\mathbf{V})$ and the true posterior $p(\mathbf{V} \mid \mathbf{U}, \mathbf{T})$. The Evidence Lower Bound (ELBO) is given by:

$$\text{ELBO}(q) = \mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})] + \mathbb{H}(\mathbf{V}) \tag{S3}$$

Where the second term is the entropy term.

To handle the categorical nature of $\mathbf{V}$, we use the Gumbel-Softmax reparameterization trick. For each categorical variable v_i in $\mathbf{V}$ (i.e., columns of $\mathbf{V}$), parameterized by ψ_i, we have:

$$v_i \sim \text{Categorical}(\psi_i) \tag{S4}$$

The Gumbel-Softmax trick provides a continuous approximation for differentiable sampling:

1. Sample Gumbel noise g_j for each category j:

$$g_j = -\log(-\log(u_j)) \quad \text{where} \quad u_j \sim \text{Uniform}(0, 1) \tag{S5}$$

2. Compute the Gumbel-Softmax sample:

$$\tilde{v}_{ij} = \frac{\exp((\log(\psi_{ij}) + g_j)/\tau)}{\sum_k \exp((\log(\psi_{ik}) + g_k)/\tau)} \tag{S6}$$

Here, τ is the temperature parameter controlling the smoothness of the approximation. As $\tau \to 0$, $\tilde{v}_{ij}$ approaches a one-hot encoded vector.

Using the Gumbel-Softmax reparameterization, we approximate the expectation in the ELBO with a sample of $\mathbf{V}$,

which we denote $\tilde{\mathbf{V}}$:

$$\mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})] \approx \log p(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) \tag{S7}$$

$$\mathbb{H}(\mathbf{V}) \approx -\sum_{j=1}^{C}\sum_{k=1}^{K} q(\tilde{\mathbf{V}}_{jk}) \log q(\tilde{\mathbf{V}}_{jk}) \tag{S8}$$

In order to capture multiple modes of the posterior distribution, we optimize a set of variational distributions $q(\mathbf{V})$ in parallel, as described in Methods, "Gumbel-softmax optimization".

In the following sections, we describe how we calculate $p(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T})$, which is broken down into (1) $p_m(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T})$, i.e., the scoring of $\tilde{\mathbf{V}}$ using maximum parsimony, (2) $p_g(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T})$, i.e., the scoring of $\tilde{\mathbf{V}}$ using genetic distance, and (3) $p_o(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T})$, i.e., the scoring of $\tilde{\mathbf{V}}$ using organotropism.

A.1. Evaluating maximum parsimony. As previously described by MACHINA[15], the maximum parsimony metrics are defined as:

- **migration number** m: Given clone tree $\mathbf{T}$ and vertex labeling $\mathbf{V}$, the migration number is the number of edges in $\mathbf{T}$ where the outgoing node and incoming node have a different label. It is the number of edges in migration graph $\mathbf{G}$.
- **comigration number** c: Given clone tree $\mathbf{T}$ and vertex labeling $\mathbf{V}$, the comigration number is a subset of the migration edges between two anatomical sites, such that the migration edges occur on distinct branches of the clone tree. It is the number of multi-edges in migration graph $\mathbf{G}$ if $\mathbf{G}$ does not contain cycles.
- **seeding site number** s: Given a clone tree $\mathbf{T}$ and vertex labeling $\mathbf{V}$, the seeding site number is the number of unique anatomical sites with an outgoing edge. It is the number of edges in migration graph $\mathbf{G}$ with an outgoing edge.

Maximum parsimony scoring calculates the number of migrations m , comigrations c, and seeding sites.

$$\begin{aligned} p_m(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) &= w_m \cdot m + w_c \cdot c + w_s \cdot s \\ m &= \sum_{ij} \mathbf{G} - Tr(\mathbf{G}) \\ s &= \sum_{j=1}^{n} \left(\left(\sum_{i=1}^{m} \left(\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K)\right)_i \right)^* \right)_j \\ c &= \sum_{ij} \mathbf{G}^*_{ij} - Tr(\mathbf{G}^*) + \sum_{ij} \left(\sum_{l=1}^{m} \left(\mathbf{P} \odot (\mathbf{W} \odot \mathbf{X})\right)_l \right)_{ij} \end{aligned} \tag{S9}$$

where $\mathbf{G} = \mathbf{V}\,\mathbf{T}\,\mathbf{V}^T, \mathbf{P} = \left(\mathbf{T} \vee \mathbf{I}_N\right)^{N-1}, \mathbf{X} = \mathbf{V}^T\,\mathbf{V}, \mathbf{Y} = \sum_{i=1}^{m} \left((\mathbf{V}\,\mathbf{T}\,\mathbf{V}^T \odot (\mathbf{J}_{CK} - \mathbf{V}^T)\right), \mathbf{Z}^* = \mathbf{sign}(\mathbf{Z})$. $\vee$ represents boolean matrix multiplication, $\mathbf{I_n}$ is a $n \times n$ identity matrix, and $\mathbf{J}_{mn}$ is a matrix of ones with dimensions $m \times n$.

A.2. Evaluating genetic distance. Genetic distance is a measure of the number of mutations between clones. Given a distance matrix $\mathbf{D}$ which has normalized genetic distances between every clone:

$$p_g(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) = \frac{w_g}{m} \sum_{ij} -log(\mathbf{D}) \odot \mathbf{T} \odot (\mathbf{J}_C - \mathbf{X}) \tag{S10}$$

where $\mathbf{T} \odot \mathbf{J}_C - \mathbf{X}$ tells us if two nodes have an edge between them and they are in different sites. Taking the hadamard product of this with the negative log of $\mathbf{D}$ gives lower scores to edges with higher genetic distances. We normalize by the migration number m so we don't penalize migration histories with more migrations through this scoring.

A.3. Evaluating organotropism. Organotropism refers to the observation that certain cancers metastasize to specific organs. We penalize migration edges between organs that are less likely to occur based on clinical data. Given a vector $\mathbf{o}$ which contains the frequency that a primary tumor seeds other anatomical sites:

$$p_o(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) = \frac{w_o}{m_p} \sum_{i=1}^{K} -log(\mathbf{o}) \odot (\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K))_{p,i} \tag{S11}$$

where $(\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K))$ contains the number of migrations between different sites, and taking the hadamard product of this with the negative log of $\mathbf{o}$ gives lower scores to migration edges with higher organotropism frequencies. The subscript p, i represents taking the row of $(\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K))$ which represents the primary site index and summing over the columns at every other anatomical site i. We normalize by m_p, the number of migrations originating from the primary site, so we don't penalize migration histories with more migrations through this scoring.

B. Calibrate alignment

To fit the parameters of the the maximum parsimony part of the objective $\theta = [w_m \ w_s \ w_d]$ (Equation S9), which enforce a seeding model, we look at how well the maximum parsimony distribution (under many possible θs) aligns with the genetic distance distribution of each patient's migration history trees.

Take a cohort of N patients, where each patient, n, is associated with a set,

$$S^{(n)} = \left\{ t_i^{(n)} \right\}_{i=1}^{T^{(n)}},$$

of $T^{(n)}$ trees. Each tree t is associated with a genetic distance g_t (or, alternatively, an organotropism score), and a vector of parsimony metrics $\mathbf{x}_t = [m_t \ c_t \ s_t]$ (i.e., the counts of migrations, comigrations, and seeding sites, respectively). The goal is to set the parameters, $\theta = [w_m \ w_c \ w_s]$ of the parsimony prior $q(t) \propto \exp\left(-\mathbf{x}_t^T \theta\right)$ so that it matches, as best as possible, a target distribution, $p(t)$, over the trees t implied by the g_t, where $p(t) \propto \exp(-\tau g_t)$ and τ is a user-defined "temperature" hyper-parameter.

To fit these parameters, we define patient-specific categorical distributions $p^{(n)}(t)$ and $q^{(n)}(t)$ as follows. Let $\mathbf{g}^{(n)}$ be the vector of length $T^{(n)}$ of genetic distances of the trees for patient n, where $g_i^{(n)}$ is the genetic distance for the i-th tree. And let the column vector $\mathbf{x}_i^{(n)}$ be the parsimony metrics for the i-th tree associated with patient n. We will append the $T^{(n)}$ vectors $\mathbf{x}_i^{(n)}$ to make a $3 \times T^{(n)}$ design matrix $X^{(n)}$. Also we define the vector-valued softmax function in the typical way, i.e.,

$$\text{softmax}(\mathbf{v})_i = \frac{\exp(v_i)}{\sum_{j=1}^{|\mathbf{v}|} \exp(v_j)}$$

where $\text{softmax}(\mathbf{v})_i$ is the i-th element of the vector output by $\text{softmax}(\mathbf{v})$. Then the "parsimony" probability distribution over the trees for patient n is represented by the vector $\mathbf{q}^{(n)}$

$$\mathbf{q}^{(n)} = \text{softmax}(-\theta^T X^{(n)})$$

and the target distribution by the vector $\mathbf{p}^{(n)}$

$$\mathbf{p}^{(n)} = \text{softmax}(-\tau \mathbf{g}^{(n)}).$$

Then we define the cohort calibration objective $E(\theta)$ as an average cross-entropy over the patient cohort, i.e.,

$$E(\theta) = \sum_{n=1}^{N} \left(\sum_{i=1}^{T^{(n)}} p_i^{(n)} \log q_i^{(n)} \right)$$

and the MLE estimate of the parameters is $\theta^* = \text{argmax}_\theta E(\theta)$.

B.1. Specifying the target distribution by setting the temperature parameter. The use of $E(\theta)$ to set θ requires that for a patient n that, generally speaking, the genetic distance $g_i^{(n)}$ for a potential migration history, represented by a tree i, is lower for more probable histories. However, because $E(\theta)$ is minimized when $\tau \mathbf{g}^{(n)} = \theta X^{(n)} + c\mathbf{1}$ for some constant c, this could be a very strong assumption, one that we might not always be comfortable making.

Fortunately, we can set τ to increase the correctness of this assumption. Notice that in the limit of large τ that

$$\lim_{\tau \to \infty} E(\theta) = \sum_{n=1}^{N} w_n \ \log q_{i_n^*}^{(n)}$$

where $i_n^* = \text{argmin}_i \ g_i^{(n)}$, assuming that the minimum is unique. If the minimum is not unique then the above is true if we replace $\log q_{i_n^*}^{(n)}$ with the average of $\log q_t^{(n)}$ of all the trees t that have the minimum genetic distance for patient n.

So, in other words, if we set τ to be very large, then $E(\theta)$ is just the (weighted) sum of the log probabilities of the minimum genetic distance trees in each patient, and optimizing $E(\theta)$ corresponds to maximizing the parsimony probabilities of the best scoring trees per patient under the genetic distance score.

$$\prod_i \frac{exp(X^{(i)^\tau}\theta)}{\sum_{j|rank(j)\geq rank(i)} exp(X^{(j)^\tau}\theta)}$$

So, we set τ to be large, such that τ is multiple times the maximum genetic distance (assuming that the genetic distance is always positive). We do the same for the organotropism prior.

C. Case-by-case differences to expert annotations

C.1. Comparisons to Melanoma patients from Sanborn et al.. Migration histories generated for the metastatic melanoma cohort using Metient-calibrate agree with the expert analysis that most melanoma patients exhibit primary single-source seeding (6/7 patients; Supplementary Figure S2). For patient F (Supplementary Figure S2c), our reconstruction of the clone tree and observed clones does not suggest that a lymph node to distant metastasis seeding event is likely, but that this patient also likely exhibits a primary-only seeding pattern. For patient D, we predict that a locoregional skin metastasis from the right ankle could have given rise to subsequent metastases, supporting one of the possible paths (in dotted lines) that the original authors propose (Supplementary Figure S2d). We also predict a primary single-source solution on the Pareto front which is another possible path proposed by the authors (Supplementary Figure S2d).

C.2. Comparisons to HGSOC patients from McPherson et al. In the seven HGSOC patients, predicted migration histories by McPherson et al.[10] were made available using an algorithm that only minimizes migrations (Sankoff algorithm[53]). We find that five out of seven patients are in complete agreement (Supplemental Figure S3). For patient 1, by resolving polytomies, we offer an explanation with less migrations and comigrations, and predict that the left fallopian tube rather than the small bowel served as a possible intermediate site before further metastatic dissemination (Supplemental Figure S3a). For patient 3, we offer an explanation with less migrations, comigrations and seeding sites, suggesting that all metastases were seeded from the primary (Supplemental Figure S3c). Finally for patient 7, solving for polytomies allows us to reduce the migration number by 1 from the right uterosacral to left ovary, although the overall seeding pattern is in agreement (Supplemental Figure S3d).

C.3. Comparisons to HR-NB patients from Gundem et al. Because the HR-NB annotations only indicate the presence of a migration between two sites and not the directionality, we compared our site-to-site migrations (i.e., a binarized representation of migration graph $\mathbf{G}$ (Figure 1c)) to those that were previously reported. We looked at the 14 HR-NB patients for which there were manual expert annotations from Gundem et al.[7], and found that we predict the same overall site-to-site migrations for 9 out of 14 cases. For patient H103207, we predict their before therapy pattern on the Pareto front (Solution 3 in Figure S4a), but we prioritize two solutions with lobe-to-lobe brain metastasis, and extensive seeding between the lobes of the lung as well as the liver. This seeding between the liver and two lobes of the lung is suggested in their after therapy hypothesis of spread (Figure S4a). For patients H132372 and H132396, Metient prioritizes migration histories with fewer migrations (Figure S4f, g), but presents the expert annotations on the Pareto front. For patient H132384, Metient proposes bone-to-bone secondary metastasis formation, but

again presents the expert annotations on the Pareto front (Figure S5d). For patient H134821, we offer a simpler explanation of seeding for patient, where four of the metastasis-to-metastasis transitions predicted by Gundem et al. are not needed to explain the observed clones and clone tree (Supplementary Figure S5f). We do however support the metastasis-to-metastasis seeding from pancreas to the hilar lymph node as proposed by the authors.

D. Bulk DNA sequencing pre-processing

D.1. Variant read probability calculation (ω). In order to account for non-diploid copy number and tumor purities, we require a variant read probability ω to be input for every genomics locus at each anatomical site. The variant read probability is the probability of observing a read with the variant allele at that locus in a cell with the mutation, and is calculated as

$$\omega = \rho * (c_{maj})/(\rho * (c_{maj} + c_{min}) + (1 - \rho) * 2) \quad \textbf{(S12)}$$

Where ρ is tumor purity, c_{maj} is major copy number and c_{min} is minor copy number, assuming that the major copy number maps to the variant allele.

If clustering is used, we have to properly combine multiple SNV loci with different potential variant read probabilites. To do this, we rescale the reference and variant allele read counts for each locus and then set its variant read probability to 0.5 before combining variants within a cluster (where we add the reference and variant allele read counts for all variants within a cluster). This rescaling allows us to effectively treat the variant as coming from a diploid locus. To achieve this, we use the following rescaling formulas, which has been previously described in Wintersinger et al.[47]:

$$\begin{aligned}
T_{js} &= V_{js} + R_{js} \\
\hat{T}_{js} &= 2\omega_{js} T_{js} \\
\hat{V}_{js} &= \min(V_{js}, \hat{T}_{js}) \\
\hat{R}_{js} &= \hat{T}_{js} - \hat{V}_{js} \\
\hat{\omega}_{js} &= \frac{1}{2}
\end{aligned}$$

Where T_{js} is the input count of total reads, V_{js} is the input count of variant reads, R_{js} is the input count of reference reads, and ω_{js} is the variant read probability at a genomic locus j in anatomical site s. The rescaled total, reference, and variant allele read counts and variant read probability are $\hat{T}_{js}, \hat{V}_{js}, \hat{R}_{js}$ and $\hat{\omega}_{js}$, respectively.

D.2. Breast Cancer Dataset. The single nucleotide variant calls from two breast cancer patients with whole genome sequencing data were taken from Hoadley et al.[21]. The variant calls were in copy number neutral variant positions and tumor purity was not reported, so reference and variant counts along with defaults for tumor purity, major

copy number and minor copy number (defaults are 1.0, 1, 1, respectively) were inputted into PyClone-0.13.1 clonal analysis[54]. PyClone's MCMC chain was run for 100,000 iterations, discarding the first 50,000 as burnin. Orchard was run using the PyClone clusters as input with -p flag to force trees to be monoprimary (come from a singular root cancer clone) and all variant read probabilities set to the default of 0.5, since SNVs from regions with CNAs were excluded, and tumor purity was not reported and thus assumed to be 1. We ran Metient-evaluate on this data using all default configurations (dynamically calculated sample size based on size of input clone tree and number of anatomical sites).

D.3. High-grade Serous Ovarian Cancer Dataset. To better compare to McPherson et al.'s own migration history analysis, we used the mutation clusters, clone trees and cellular prevalences of each clone that they estimate and report[10]. Metient was run with the $\mathbf{U}$ matrix inputted, and we solve for $\mathbf{V}$ for each patient. We ran Metient-calibrate on this data using all default configurations (dynamically calculated sample size based on size of input clone tree and number of anatomical sites) and with polytomy resolution.

D.4. Melanoma Dataset. The single nucleotide variant and copy number calls from eight melanoma patients with whole exome sequencing data were taken from Sanborn et al.[9], along with estimated tumor purity. Only SNVs in copy number neutral regions were considered. Patient H was excluded due to a lack of copy number neutral SNVs. Reference and variant read counts along with major and minor copy number and tumor purity were inputted into PyClone-0.13.1 for clonal analysis[54]. PyClone's MCMC chain was run for 10,000 iterations, discarding the first 5,000 as burnin. The maximum number of clusters to use was set to 30 and clusters with less than 5 mutations were discarded. Orchard was run using the PyClone clusters as input with -p flag to force trees to be monoprimary (come from a singular root cancer clone). Variant read probabilities for Orchard and Metient were calculated using major copy number, minor copy number and tumor purity according to Equation S12. We ran Metient-calibrate on this data using all default configurations (dynamically calculated sample size based on size of input clone tree and number of anatomical sites) and with polytomy resolution.

D.5. Neuroblastoma Dataset. Access to multi-WGS data for 45 neuroblastoma patients was provided through dbGaP accession phs03111[7]. Of these 45 patients, 27 patients had at least one primary and one metastatic tumor sample with a tumor purity of >10%, and all analysis was conducted on this patient subset. Single nucleotide variant, copy number calls and tumor purities were collected from this dataset, and clusters produced from the original paper using DPClust[55] were used. Multiple samples for the same anatomical site and sample time (i.e., diagnosis, therapy-naive re-resection, therapy resection during induction chemotherapy, relapse or further relapse) were combined by pooling reference and variant allele counts. Orchard was run using the DPClust clusters as input with -p flag to force trees to be monoprimary (come from a singular root cancer clone). Variant read probabilities for Orchard and Metient were calculated using major copy number, minor copy number and tumor purity according to Equation S12. We ran Metient-calibrate on this data using all default configurations (dynamically calculated sample size based on size of input clone tree and number of anatomical sites) and with polytomy resolution.

For three patients (H103207, H132388, H134822), multiple primary tumor samples were collected at different time points (diagnosis and resection during therapy). For these patients, we treated the therapy resection and diagnosis tumor as multiple samples from the same anatomical site if the anatomical site was labeled the same, and as two different primaries if the anatomical sites were different. The therapy resections were usually taken a few months after diagnosis tumor samples.

D.6. Non-small Cell Lung Cancer Dataset. We used the clustered SNVs, clone trees and observed clone proportions made available by the TRACERx consortium for 126 non-small cell lung cancer (NSCLC) patients (downloaded from https://zenodo.org/record/7649257). When samples for multiple regions of a tumor were available, the reference and variant allele counts were summed together to generate reference and variant allele counts for the entire tumor. Since we model variant allele counts as binomially distributed with n total reads (variant + reference) and p probability of generating a variant read, this summing assumes that each sampled region of a tumor has the same probability p. Metient was run with the $\mathbf{U}$ matrix inputted, and we solve for $\mathbf{V}$ for each patient. We ran Metient-calibrate on this data using all default configurations (dynamically calculated sample size based on size of input clone tree and number of anatomical sites) and with polytomy resolution.