Inferring cancer type-specific patterns of metastatic spread

Divya Koyyalagunta[1, 2], **Karuna Ganesh**[3, 4], **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.
[3]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[4]Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, 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 is more accurate and is up to 50x faster than current state-of-the-art. Given a cohort of patients, Metient can calibrate its parsimony criteria, thereby identifying shared patterns of metastatic dissemination in the cohort. 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 many patients, Metient identifies more plausible migration histories than experts, and further finds that polyclonal seeding of metastases is more common than previously reported. By removing the need for hard constraints on what patterns of metastatic spread are most likely, Metient introduces a way to further our understanding of cancer type-specific metastatic spread.

migration history inference | metastasis | mixed-variable combinatorial optimzation

Correspondence: *morrisq@mskcc.org*

Introduction

Metastasis is associated with 90% of cancer deaths, yet its causes and physiology remain poorly understood[1]. It remains unclear how often multiple clones seed metastases, how often metastases are capable of seeding other metastases, and if there is a relationship between seeding clones and organ-specific metastases[2–10]. It is also not known whether metastatic potential is rare, and thus gained once in the same cancer, or common, and thus gained multiple times[11–14]. The answers to all these questions would improve the understanding and clinical management of metastasis, but doing so requires reconstructing migration histories of metastatic clones from clinical sequencing data which, until recently, was very challenging[2–4].

Recent algorithms have tackled this challenge using maximum parsimony principles. These algorithms identify parsimonious migration histories that explain the clonal compositions of primary tumors and one or more matched metastatic tumors[5,15–17]. However, different definitions of parsimony can disagree on the best solution, and current algorithms resolves these conflicts using ad hoc rules[15–17]. For example, a common rule is to only allow metastases to be seeded from the primary[14], whereas determining whether metastases can seed other metastases is, itself, an important question. Indeed, one prevailing model in oncology, the "sequential progression model" – which posits that lymph node metastases give rise to distant metastases – is the rationale for surgical removal of lymph nodes[18]. However, a recent phylogenetic analysis found that the sequential model only applied to a third of patients in a colorectal cohort[19]. By pre-biasing their reconstructions with ad hoc rules, current algorithms undermine a key goal in making these reconstructions: determining which patterns of metastatic spread are prevalent in different cancer types.

To address this dilemma and overcome the limitations of previous tools (Supplementary Table 1), we introduce Metient (**met**astasis + grad**ient**). Metient is a principled statistical algorithm that proposes multiple potential hypotheses of metastatic spread in a patient and resolves parsimony conflicts using other, readily-available data. Metient achieves this through two key innovations. First, it adapts recent stochastic optimization algorithms for discrete variables to the problem of combinatorial optimization, thereby enabling efficient sampling of multiple parsimonious solutions. Second, it introduces new biological criteria, termed metastasis priors, to calibrate its parsimony criteria and select among equally parsimonious solutions. These calibrated criteria can also be used to uncover cancer type-specific trends in metastatic spread.

On realistic simulated data, Metient outperforms parsimony-only models in accurately recovering the true migration history. When applied to patient cohorts with metastatic breast[20], skin[3], ovarian[4], neuroblastoma[9], and lung cancer[14], Metient automatically identifies all plausible expert-assigned migration histories. In notable cases, it also uncovers more plausible reconstructions, often when prior expert analyses pre-selected a favored seeding pattern.

Through its unbiased automated approach, Metient reveals that metastases are often seeded polyclonally and that most metastatic seeding follows a single, shared evolutionary trajectory. The cancer type-specific models learned by Metient reflect known differences in metastasis biology, suggesting that Metient can offer insights into metastatic

dissemination for new cancer cohorts.

Metient is free, open-source software that includes easy-to-use visualization tools to compare multiple hypotheses on metastatic dissemination. Metient is accessible at https://github.com/morrislab/metient/.

Results

The Metient algorithm

Migration history inference algorithms take DNA sequencing data from primary and metastatic tumor samples as input, along with an unlabeled clone tree that encodes the genetic ancestry of cancer clones (Figure 1a). These inputs are used to estimate the proportions of clonal populations in anatomical sites (referred to as "witness nodes" in Figure 1b). The internal nodes of the clone tree are then labeled with anatomical sites, defining the historical migrations: a clone that migrates to a new site receives a different label than its parent clone (Figure 1b) and the tree edge that connects them is deemed a “migration edge”. The final output is referred to as a “migration history”[17] (Figure 1b).

MACHINA[17] 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[4,15–17]; **comigrations**—the number of migration events in which one or more clones travel from one site to another[17]; and **seeding sites**—the number of anatomical sites that seed another site[17]. MACHINA searches for the most parsimonious history by minimizing these three metrics.

This search involves solving a mixed-variable combinatorial optimization problem, consisting of continuous variables (the clone porportions matrix $\mathbf{U}$ in Figure 1b), and discrete variables (the labeled clone tree matrix $\mathbf{V}$ in Figure 1b). MACHINA, and other prior approaches, formulate this problem as a mixed integer linear programming (MILP) problem that they solve using commercial solvers[21]. However, using an MILP imposes strong limitations on the types of scoring functions that can be applied to migration histories, as MILPs require hard constraints and a linear objective. Moreover, MILP solvers identify only a single optimal solution, whereas there are often multiple solutions which are either equally parsimonious, or that trade-off one parsimony metric for the another (e.g., reducing the number of seeding sites by increasing the number of migration events). Returning a single solution obscures these possibilities, and the ad hoc rules used to distinguish among multiple solutions often introduce implicit bias into the reconstructions.

To address these issues, Metient takes a more systematic approach by first defining a “Pareto front”[22] for each patient (Figure 1c). To do so, Metient searches for migration histories under a wide range of parsimony models (Supplementary Table 2). A parsimony model is represented by a set of parsimony weights – w_m, w_c, and w_s – assigned, respectively, to the number of migrations (indicated by m), comigrations (c), and seeding sites (s). A migration history’s parsimony score, p, is the model-weighted average of these three parsimony metrics, i.e., $p = w_m m + w_c c + w_s s$. Different parsimony models favor different histories on the Pareto front. Efficiently recovering this Pareto front required replacing the current state-of-the-art MILP with newly developed stochastic gradient descent methods that employ a low-variance gradient estimator for the discrete categorical distribution over migration histories parameterized by the parsimony model[23,24] ($\mathbf{V}$ in Figure 1b; Methods, Supplementary Information). Metient’s gradient descent approach converges to a solution many times faster than the MILP, and

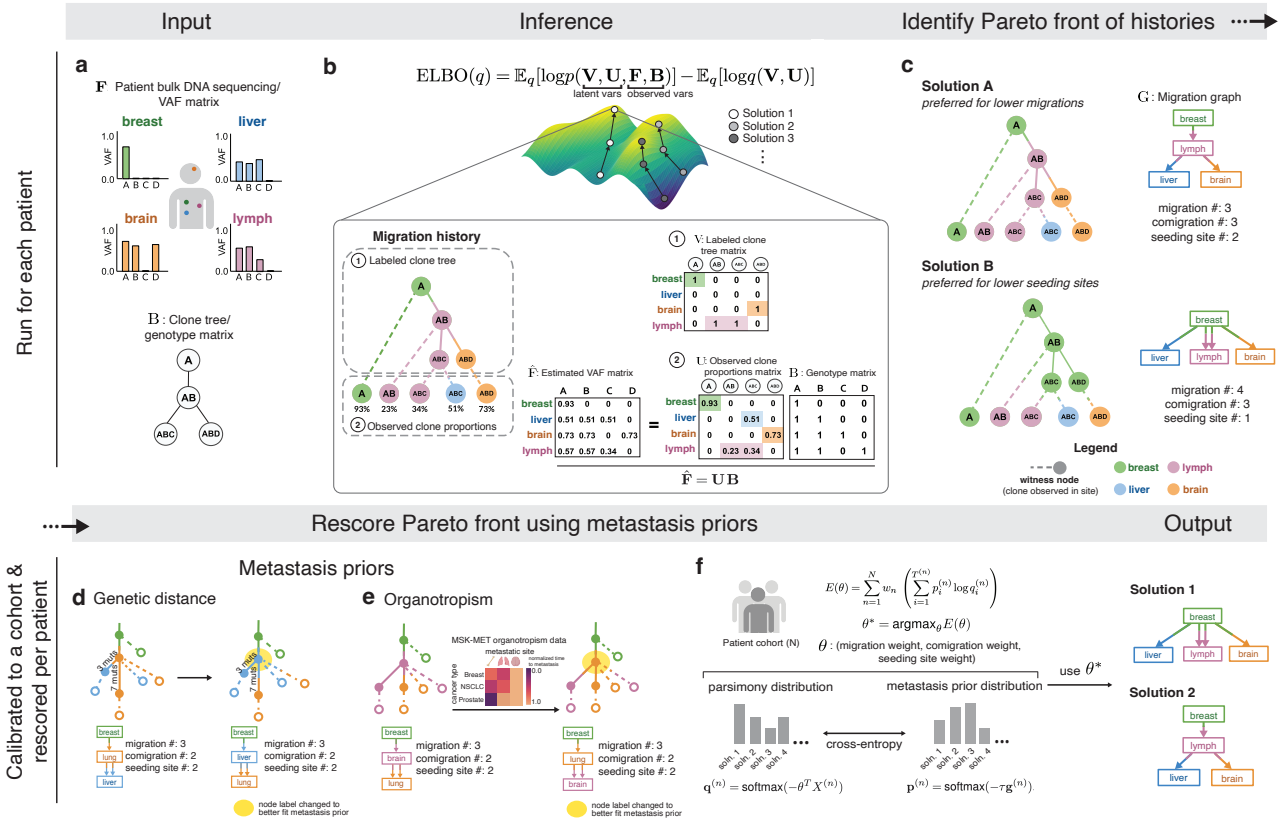

Figure 1. Overview of the Metient method. (a) Input: (top) bulk DNA sequencing sampled from multiple tumors in a single patient, and (bottom) a clone tree which represents the evolutionary relationship of mutations. AB refers to a clone with mutations or mutation clusters A and B. **(b) Inference**: 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 proportion 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 a 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 Pareto front of migration histories produced for each patient in that cohort.

it also helps to define the Pareto front by identifying multiple local maxima of the migration history score for each parsimony model (Methods, Supplementary Information). In addition, this approach reduces a large combinatorial search space of possible migration histories to only the most plausible explanations of metastatic spread for a given patient.

Metient-calibrate fits cancer type-specific parsimony models

To illustrate the importance of defining a Pareto front of multiple possible patterns of metastatic spread, we defined four different cancer type-specific patient cohorts consisting of genomic sequencing of matched primary and multiple metastases: melanoma[3], high-grade serous ovarian cancer (HGSOC)[4], high-risk neuroblastoma (HR-NB)[9], 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) in total from 167 patients (melanoma: n=7, HGSOC: n=7, HR-NB: n=27, NSCLC: n=126). Applying Metient to these patients, we discovered that 45% (75/167) had multiple Pareto-optimal migration histories, and that the complexity of the Pareto front increased with the number of metastases: 79% (27/34) of patient cases with three or more metastases had multiple Pareto-optimal histories. Often the choice among these different Pareto-optimal histories substantially impacted the interpretation of metastatic spread. For example, Figure 1c shows a patient with metastatic breast cancer with two Pareto-optimal 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, forcing 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[4,15,17], resolve parsimony conflicts by minimizing migrations first, and then comigrations, thus implementing a parsimony model where $w_m >> w_c >> w_s$. However, no single parsimony model is appropriate for all cancer types. For example, in ovarian cancer, clusters of metastatic cells are thought to "passively" disseminate to the peritoneum or omentum through peritoneal fluid[25–27]. As such, metastatic events are more likely to be polyclonal, i.e., multiple clones seed metastases, so we might expect many more migrations than comigrations. In many solid cancers, metastatic cells make a "pit stop" at regional lymph nodes before disseminating to other distant sites[28], and for the estimated 23.4% of patients with lymph node metastases across cancer types[29], multiple seeding sites may be common. Different cancer type-specific patterns of metastatic spread are reflected in differences in trends in the relative numbers of migrations, comigrations, and seeding sites, and prespecifying a cancer type-independent parsimony model can prevent the recovery of these patterns. Furthermore, in our cohorts, we found that there were often multiple, equally parsimonious migration histories. MACHINA selects among these randomly, or via predefined constraints on the allowable patterns of metastatic spread.

In contrast, Metient uses metastasis priors to both define a cancer type-specific parsimony model and to rank equally parsimonious histories. These priors incorporate additional biological constraints relevant to migration histories. We provide a tool, Metient-calibrate, that fits a patient cohort-specific parsimony model using the metastasis priors (Figure 1d-f; Methods). This calibrated model is used to rank Pareto-optimal histories that differ in their metrics. Metient also provides a pan-cancer parsimony model, calibrated to all four cohorts combined, for use when an appropriate patient cohort is not available.

Metient provides two metastasis priors. One, genetic distance, can be applied to any cohort. The other, organotropism, can be used when appropriate tissue-type information are available for the sequenced tumor samples. The genetic distance prior considers the average genetic distance of migration edges in the labeled clone tree; where the genetic distance on an edge is the number of mutations gained in the child clone and not present in the parent clone. In general, we expect genetic distance to tend to be higher on migration edges than other clone tree edges for a number of reasons. First, the colonizing clones of a metastasis have undergone a clonal expansion in their metastatic site, which makes their private mutations more easily detectable by finite depth sequencing. In

contrast, the vast majority of private mutations in the source tumor will not be at high enough cellular frequency to be detectable, and subclones detected in the source tumor need not have undergone a clonal expansion[30]. In addition to increased mutation detectability, colonizing cells likely have more mutations than randomly selected cells in the source population due to the strong selection pressures they faced in metastasizing, as strong selection pressures select, perhaps indirectly, for higher mutation rates in asexually reproducing populations[31–33]. Finally, metastases exhibit greater genomic instability[29,34,35], possibly as a consequence of these selection pressures, which is associated with heightened mutation rates[36]. Indeed, metastases across many cancer types have moderately or significantly higher tumor mutation burden (TMB) than matched primaries[29,35,37]. Metient's genetic distance prior deems more probable those migration histories with higher averaged genetic distances on migration edges (Methods, Supplementary Information). Figure 1d illustrates an example of using the genetic distance prior to select between two equally parsimonious migration histories.

The second metastassis prior, organotropism, is derived from data from 25,775 Memorial Sloan Kettering metastatic cancer patients[29] on the preference that some cancer types have to colonize other organs[38]. We used these data to construct a matrix for 27 common cancer types, where each entry is the frequency of metastasis to a particular anatomical site that is observed in patients with that cancer type (Figure 1e). Note that there are no direct data for frequencies of migrations from one metastatic site to another metastatic site, so Metient only uses this matrix to score migrations coming from the primary site (Methods). For example, breast cancer metastasizes to lung more often than brain, so Metient's organotropism prior favors a solution with migrations to the brain from a breast-seeded lung metastasis over one with migrations from a breast-seeded brain metastasis to the lung (Figure 1e). Indeed, brain to lung metastasis is rare[39]. As we illustrate in later sections, our metastasis priors lead to better performance on simulated benchmarks, and more plausible migration history reconstructions than using maximum-parsimony rules and cancer type-independent rules. Nonetheless, Metient reports all Pareto-optimal solutions; in this example, both solutions in Figure 1e are visualized in a simple summary report, so that these multiple hypotheses can be easily evaluated by the user.

Importantly, Metient uses its metastasis priors to complement but not replace its parsimony model. In our benchmarking analyses on simulated data, we find that using genetic distance alone to score migration histories performs poorly and can result in the inference of highly non-parsimonious migration histories (Supplementary Tables 4, 3, see also PathFinder[40]). Instead, the metastasis priors are only used once the Pareto front is defined, to calibrate parsimony models and to rank equally parsimonious solutions.

Simulated data validates the genetic distance prior and shows that Metient is state-of-the-art

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[17]) on simulated data, originally used to validate MACHINA, for 80 patients with 5-11 tumor sites and various patterns of metastatic spread.

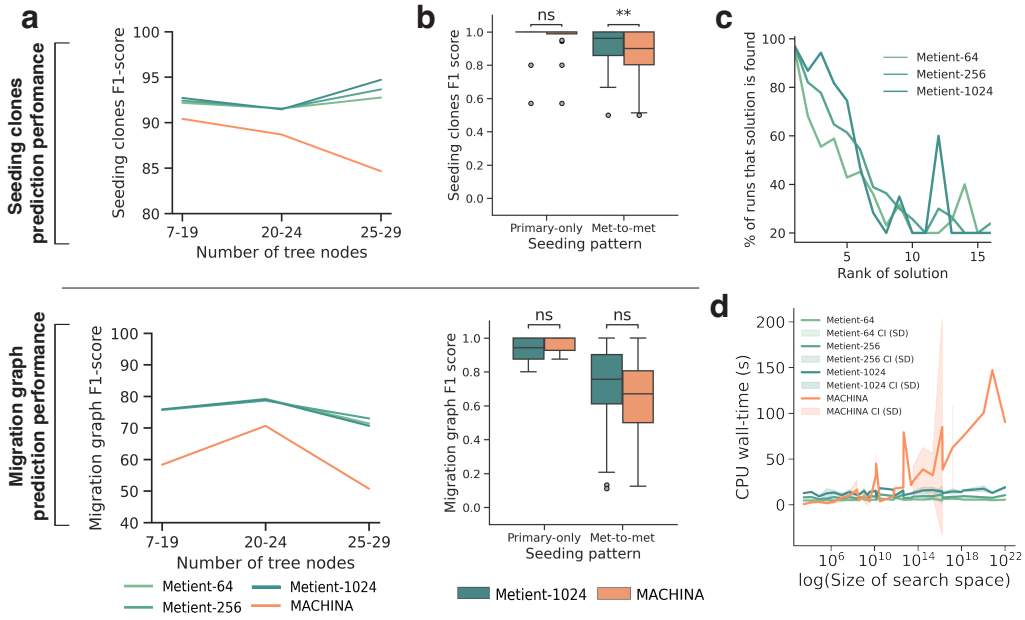

Figure 2. Metient achieves state-of-the-art performance on simulated data. All results shown for Metient are in calibrate mode using genetic distance as the metastasis prior. Metient-1024 refers to a model configuration where 1024 solutions are sampled. For a given simulated input, for MACHINA (which outputs one solution) the top solution is used, and for Metient we evaluate all top (lowest loss) solutions. **(a)** The averaged F1-score for predicting seeding clones (top) and migration graph (bottom), within three buckets of input tree sizes. **(b)** The distribution of F1-scores for predicting seeding clones (top) and migration graph (bottom) on different broad seeding patterns. Statistical significance assessed by a Wilcoxon signed rank test; ns: not significant, **: p=0.0021. **(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.

First, to assess the added value of the genetic distance prior, we used Metient-calibrate to fit a calibrated parsimony model, and compared calibrated Metient with a version of Metient that used the parsimony model implied by MACHINA. We fit two calibrated models, one on a cohort with primary-only seeding and another on a cohort with metastasis-to-metastasis seeding. Metient-calibrate improved recovery of the ground truth migration graph (Figure 1c) over fixed parsimony model (Calibrate vs. Evaluate (MP) in Supplementary Table 3), showcasing the ability of the metastasis priors to learn metastatic patterns specific to a cohort and improve overall accuracy. In addition, Metient-calibrate predicts ground truth seeding clones and migrations graphs at least as accurately as MACHINA, with overall improvements as tree sizes get larger (Figure 2a,b) and significant improvements in inferring the seeding clones for patients with more complex metastasis-to-metastasis seeding (Figure 2b top; p=0.0021).

Notably, although the Metient framework is non-deterministic, it identifies the same top solution 97% of the time across multiple runs (Figure 2c). Furthermore, in addition to its improved accuracy, Metient runs up to 55x faster (3.95s with Metient-64 vs. 221.19s 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).

Validation of organotropism prior

To validate the organotropism prior, we ran Metient, using the pan-cancer parsimony model, on samples available from two patients with metastatic breast cancer[20] where site labels could be mapped to those used in our

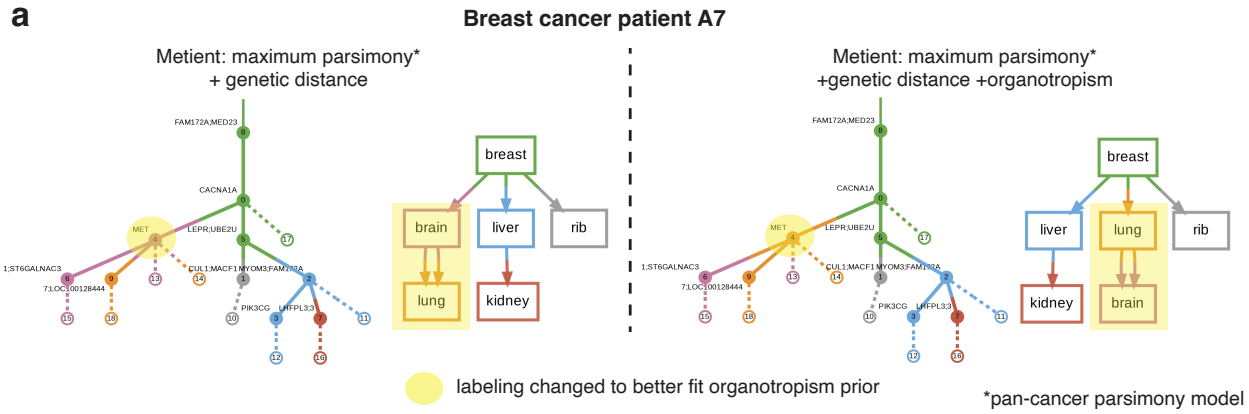

Figure 3. Organotropism prior corrects unlikely patterns of seeding. (a) The inferred migration history for breast cancer patient A7[20] without (left) and with (right) the inclusion of the organotropism prior. The addition of an organotropism prior changes the vertex labeling of clone 4 from originating in the brain to originating in the lung. Solid edges are edges in the clone tree, and dashed edges indicate the presence of the clone in the corresponding colored anatomical site (i.e., witness nodes).

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[39] (Figure 3a).

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.

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 4a). The other two criteria are clonality, i.e., the number of distinct clones seeding metastases (Figure 4b), and phyleticity, i.e., whether metastatic potential is gained in one or multiple evolutionary trajectories of the clone tree (Figure 4c; Methods). We distinguish between genetic polyclonality, in which more than one clone seeds metastases in a patient, and site polyclonality, in which more than one clone seeds an individual site (Figure 4b; Methods). We introduce this distinction to highlight cases where each metastasis is seeded by a single clone, but all sites are not seeded by the same clone (i.e., the cancer is genetically polyclonal but site monoclonal), because these may be cases where different site-specific mutations are needed for metastasis. We also update the previous definitions of metastasis-initiating clones (commonly called seeding clones). We define a seeding or colonizing clone as a node in a migration history whose parent has a different label than itself (Methods), because this clone is the only one guaranteed to have the mutations necessary to establish the metastasis. Previous work often refers to the parent of the colonizing clone as the seeding clone[14,17], although this clone may not have all of mutations required for the observed metastasis.

Consistent with expert annotations[3,4,9,14,17], Metient finds that single-source seeding from the primary tumor is the

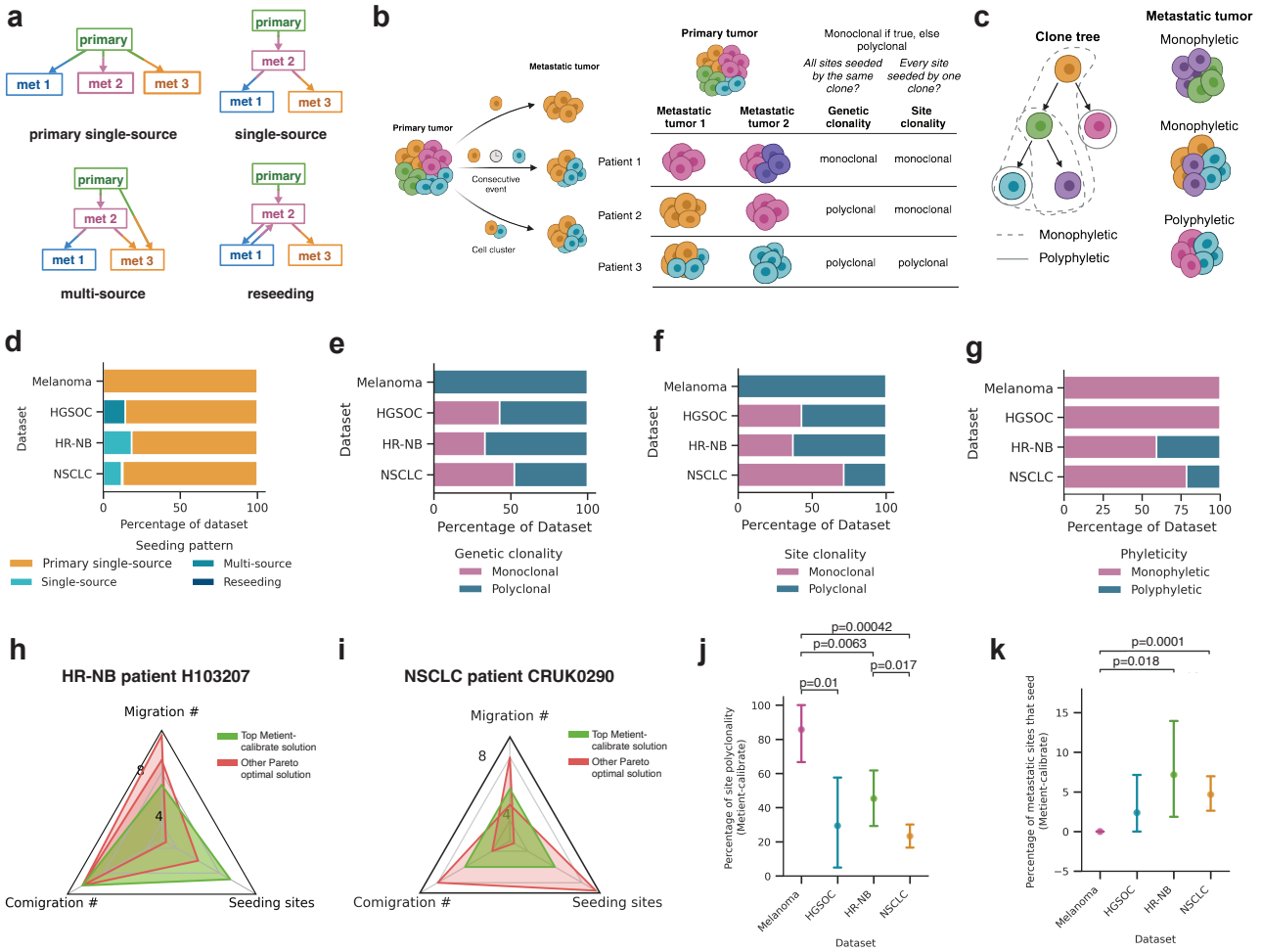

Figure 4. Clonal, phyletic and seeding patterns of four cancer types. (a) Schematic describing the four metastatic seeding patterns. met: metastasis. **(b)** Schematic depicting how metastases can get seeded by either one or multiple clones and the definitions of genetic clonality and site clonality. 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 monophyletic and polyphletic seeding. Monophyletic indicates that the colonizing clone closest to the root can reach every other colonizing clone on the clone tree. Colors represent genetically distinct cancer cell populations. Distribution of **(d)** seeding patterns, **(e)** genetic clonality, **(f)** site clonality and **(g)** phyleticity for each dataset, as inferred by Metient's top migration history. **(h)** Radar plot showing the unique Pareto-optimal metrics for migration histories inferred by Metient for HR-NB patient H103207. **(i)** Radar plot showing the unique Pareto-optimal metrics for migration histories inferred by Metient for NSCLC patient CRUK290. **(j)** Comparing across datasets the percent of migrations that are polyclonal for the top Metient solution. Statistical significance assessed by a Welch's t-test. Error bars are the standard error for each dataset. **(k)** Comparing across datasets the percent of metastatic sites that seed for the top Metient solution. Statistical significance assessed by a Welch's t-test. Error bars are the standard error for each dataset.

most common pattern in every cohort (Figure 4d). However, Metient identifies a larger fraction of polyclonal migration patterns than previous reports[8,14]: 53.3% of patients have sites that are seeded by different clones, i.e., genetically polyclonal (Figure 4e), and 38.3% of patients have at least one site seeded by multiple clones, i.e. site polyclonal (Figure 4f). Overall, Metient estimates that 34.1% of sites (107/314) are seeded by multiple clones; nearly double prior estimates of site polyclonality (19.2%) based on an analysis of breast, colorectal and lung cancer patients[8].

Notably, parsimony model choice influences the polyclonality of migration histories, because reducing the number of seeding sites tends to increase the number of polyclonal migrations (Supplementary Figure S1a). However, the higher polyclonality in Metient's reconstructions does not result from an assumption of primary-only seeding, as done in prior work, which would result in even more polyclonal migrations (Supplementary Figure S1a, Supplementary Information).

Metient's phyleticity estimates mirror previous reports: 77.2% of patients (129/167) have a monophyletic tree where metastatic potential is gained once and maintained (Figure 4g). For some patients, this is due to the root clone being observed in one or more metastatic sites (Supplementary Figure S1b), and for other patients, all colonizing 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.

Cancer type-specific metastasis trends

We next examined cancer type-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 (Methods), and found that 98.4% of patients ranked the same top solution across bootstrap samples, indicating that Metient can learn a reproducible cancer type-specific model for the melanoma and HGSOC cohorts which have only 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. colonizing clones) for HR-NB patient H103207 (Figure 4h), but the solution with the median value of each metric for NSCLC patient CRUK0290 (Figure 4i). To systematically assess the impact of cohort-specific rankings we computed the percentage of polyclonality and number of seeding sites in the top ranked solution for patients with each cancer type. Overall, we found a significantly higher fraction of polyclonal migrations in melanoma than HGSOC, HR-NB and NSCLC patients (Figure 4j). 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 traveling together through hematogeneous or lymphatic routes to seed new localized tumors[41]. The HR-NB and NSCLC cohorts had significantly higher percentages of metastasis-to-metastasis seeding than melanoma (Figure 4k). As described below, in the HR-NB

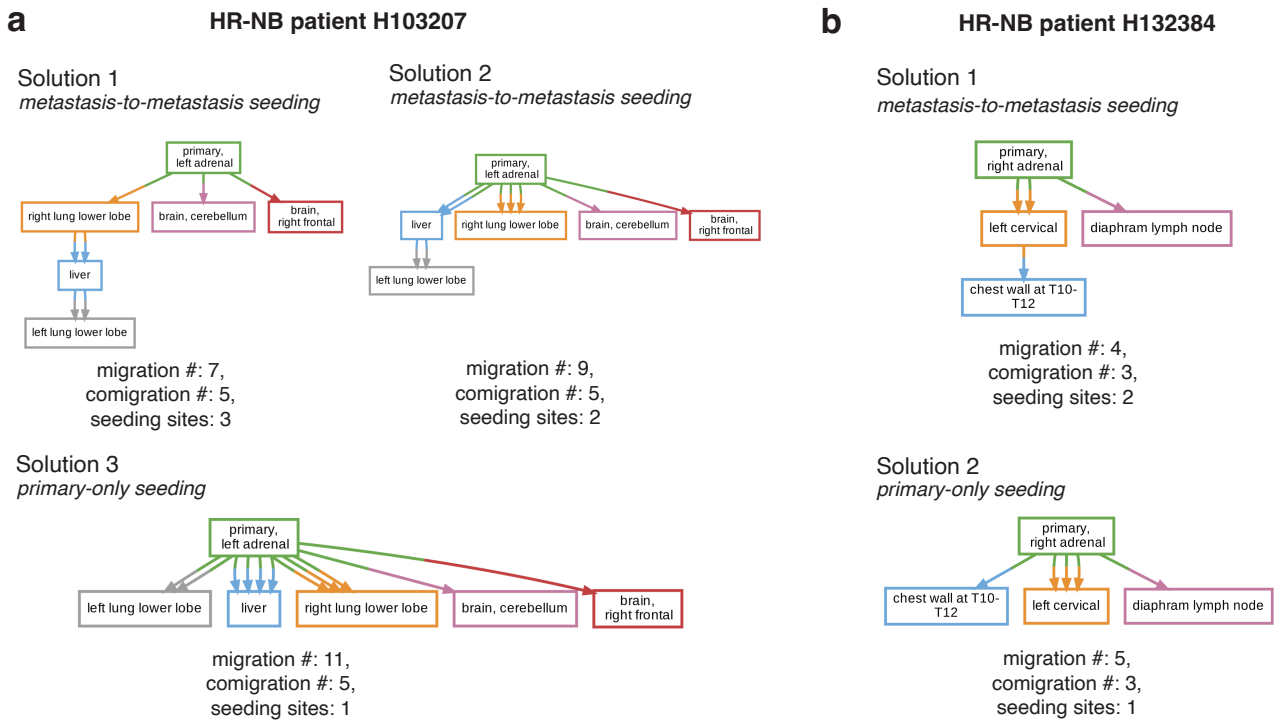

Figure 5. Metient finds biologically relevant trees. (a) All ranked Pareto-optimal migration graphs inferred by Metient-calibrate for HR-NB patient H103207. **(b)** All ranked Pareto-optimal migration graphs inferred by Metient-calibrate for HR-NB patient H132384.

cohort, multiple patients exhibit metastasis-to-metastasis seeding within an organ or between commonly metastatic sites. In the NSCLC cohort, 76.2% of patients have lymph node metastases, from which it is known that further metastases are commonly seeded[42]. Indeed, Metient predicted that 75% (12/16) of NSCLC patients who had metastasis-to-metastasis seeding had seeding from a lymph node to other metastases.

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.

Of the 167 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 identifies a more parsimonious history or recovers the expert annotation on the Pareto front but the metastasis priors prefer a different history than the expert. 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 reported to have initially seeded directly from the primary[9]. HR-NB patient H103207 shows evidence

of two possible metastasis-to-metastasis seeding scenarios. One, which is ranked the highest by the calibrated parsimony metrics posits a serial progression of metastatic seeding from the primary to the right lung, then to the liver, and finally to the left lung. The other, which has the second highest rank, posits seeding from the primary to the liver and then the left lung (Figure 5a). While the exact prevalence of metastasis-to-metastasis seeding between the liver and lung in HR-NB is unknown, both are common sites of metastases across cancer types due to cancer cells' ability to take advantage of rich blood supply, vascular organization and physiology[38]. Colonization of the lung by clones from a primary liver tumor is common[38,43,44] and, similarly, the liver is a common site of metastasis for primary lung cancer patients[38,45], suggesting that transitions from a liver-competent cancer clone to a lung-competent one and vice versa could also be common. For this patient, multiple colonizing clones emerge on distinct branches of the clone tree, providing another line of evidence that the suggested metastasis-to-metastasis seeding probably occurred (Supplementary Figure S2a). Specifically, the CNS-colonizing clones appear on a shared branch, and the lung- and liver-colonizing clones appear on a separate, shared branch after further primary tumor evolution occurred (Supplementary Figure S2a). This suggests that evolution within the primary tumor gave rise to multiple clones with organ-specific metastatic competence, and is concordant with the clonal analysis reported by Gundem et al.[9] for this patient. Patient H132384 also shows evidence of metastasis-to-metastasis seeding, but from bone-to-bone, first to the left cervical and secondarily to the chest wall (Figure 5b). Metastasizing cells exhibit organ-specific genetic and phenotypic changes to survive in a new microenvironment[38], suggesting that seeding an additional tumor within the same organ microenvironment is more likely than a secondary migration from the primary adrenal tumor in this case. In addition, prior experimental evidence shows that bone metastases prime and reprogram cells to form further secondary metastases[46,47]. These posited metastasis-to-metastasis seedings are thus upported by site proximity or organotropism, or both, and these Metient reconstructions were made without providing such information.

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 enforces 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 (12.8%; 16/126) patients (Figure 6a). CRUK0484 is one such patient where Metient proposes that an initial metastasizing clone to the rib leads to secondary metastasis formation in the scapula (Figure 6b), which we propose is a more plausible solution based on the same line of reasoning described for the bone-to-bone metastasis predicted in HR-NB patient H132384 above.

When comparing the TRACERx classifications of clonality and phyleticity for each patient to those implied by Metient's highest-scoring solution, we find 84.1% agreement (106/126) in clonality (Figure 6c) and 78% agreement (96/123) in phyleticity (Figure 6d) (three patients classified as "mixed" phyleticity by TRACERx were excluded). The discrepancies between these classifications stem from the way in which metastatis initiating clones are defined. TRACERx identifies shared clones between a primary tumor and its metastases, defining the seeding clone as the most recent shared clone between the primary tumor and the metastasis. In contrast, Metient uses the entire

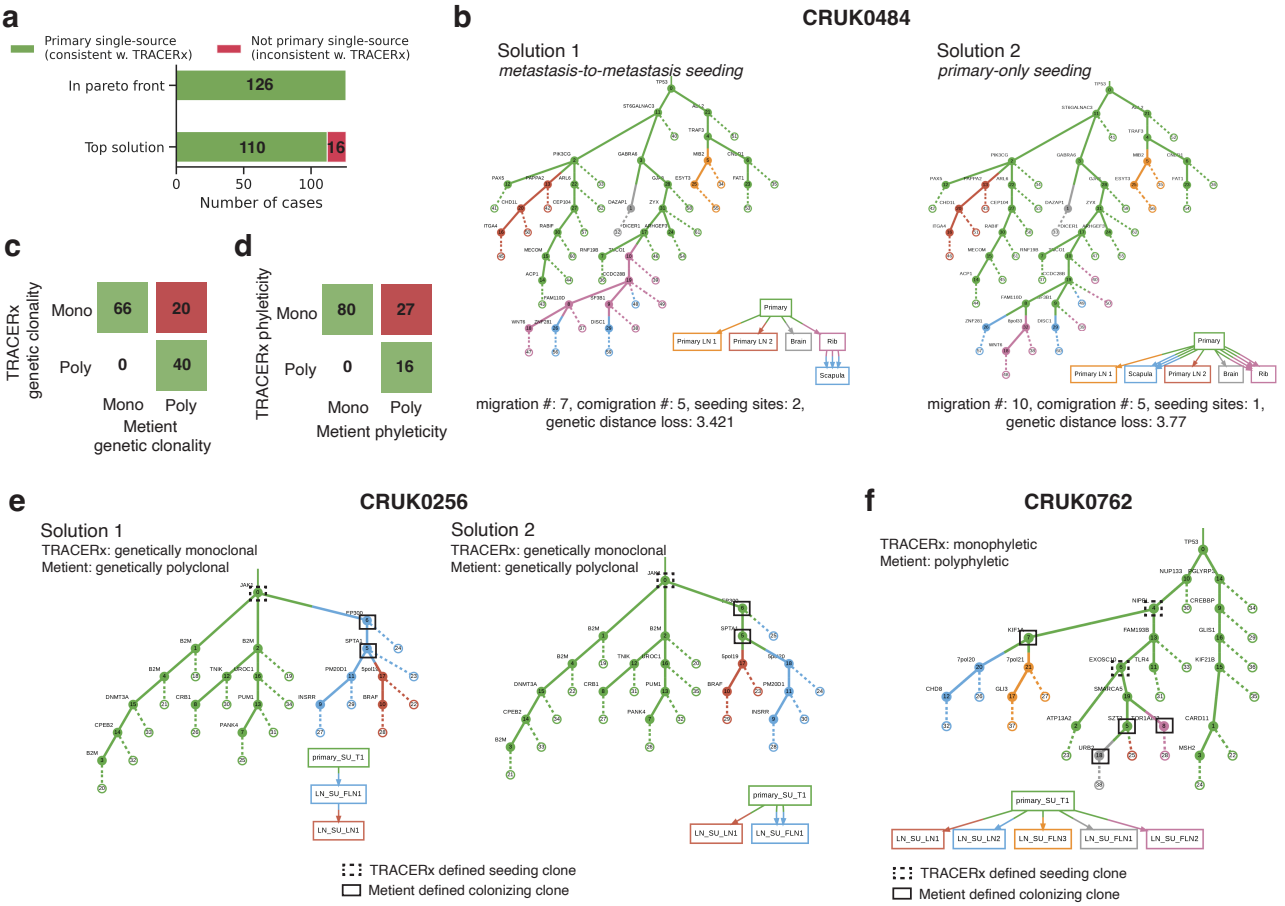

Figure 6. TRACERx NSCLC cohort. **(a)** The number of solutions that are classified as primary single-source (the assumed seeding pattern of TRACERx) vs. other when looking only at the Pareto-optimal solutions vs. the top solution. **(b)** The top two Pareto-optimal solutions for NSCLC patient CRUK0484 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 CRUK0762 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.

migration history to define seeding clones (Methods) and accounts for metastasis-to-metastasis seeding, rather than assuming that seeding occurs only from the primary tumor. As a result, Metient has significantly higher sensitivity in detecting colonizing populations within metastases and, subsequently, increases the detection of polyclonal and polyphyletic events.

In 20 NSCLC patients, Metient inferred that multiple colonizing clones are needed to explain the full migration history, whereas no history is consistent with the TRACERx identified colonizing clones. For example, for patient CRUK0256 (Figure 6e), 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 6e solution 1), or two clones originally from the primary (no longer detectable in the metastatic samples due to either ongoing evolution or undersampling) that seeded the metastases (Figure 6e 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.

Inference of phyleticity is also impacted by the use of the clone tree to determine colonizing clones, as the path connecting colonizing clones is used to determine if metastatic competence arises once or multiple times during evolution. Because the number of colonizing clones is underestimated in the TRACERx analysis, monoclonal seeding is inferred more often, automatically classifying these histories as monophyletic. Furthermore, we find 27 cases where TRACERx classifies a patient as monophyletic and Metient classifies the same patient 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 6f). Therefore, while we agree that monophyleticity is the majority pattern in NSCLC (63%), we suggest that polyphyleticity might be underestimated due to less sensitivity in previous methods' ability to detect colonizing clones.

Discussion

We have presented and validated Metient, a new framework for reconstructing the migration histories of metastases. In contrast to prior work, Metient defines a Pareto front of possible migration histories, and then uses metastasis priors to resolve parsimony conflicts in a data-dependent manner. Another key innovation is that it adapts Gumbel straight-through stochastic gradient estimation to optimize the combinatorial problem required for history reconstruction. Collectively, these advances improve performance on simulated data, improve biological interpretation on real data, and define a Pareto front in a fraction of the time that MACHINA, the current state-of-the-art, takes to output a single solution. Notably, Metient uses open source software packages, whereas other methods rely on commercial MILP solvers. Metient, due to its much improved speed, could easily be adapted to much larger migration history reconstruction problems, such as those posed by single-cell data.

Here we show that by selecting among Pareto-optimal solutions using a pre-specified parsimony model and ad hoc rules, previous algorithms biased the conclusions of studies of metastatic spread. In one study[14], primary-only seeding was assumed when analyzing migration histories, thus plausible histories with metastasis-to-metastasis seeding were ignored, even when they were identified by MACHINA. Metient thus provides an unbiased means of identifying cancer-type specific trends in metastasis biology, thus addressing a critical problem in metastasis research.

Metient's increased precision in identifying colonizing clones allowed it to detect almost twice as much polyclonality as previously reported, suggesting that it is common for multiple clones to contribute to metastatic progression. Despite this, Metient still inferred that metastatic potential rarely emerges independently in separate evolutionary paths.

Currently, Metient uses genetic distance and organotropism as its metastasis priors, however, the Metient framework is designed to be easily extensible. Adding a new prior simply requires writing a scoring function because Metient incorporates auto-differentiation to compute its gradient updates. For instance, the framework could be easily extended to incorporate mutational signatures as a prior, since metastases exhibit shifts in mutational signature composition[48,49].

Metient has some limitations. It scales well in compute time for larger clone trees or more samples but, because the loss landscape complexity increases substantially, in some cases (less than 1%), Metient became stuck in local minima. This problem was resolved when we ran Metient multiple times and with larger sample sizes, and we recommend this practice with larger reconstruction problems. One criteria to ensure convergence is when the Pareto front remains unchanged. Other migration history algorithms are also highly sensitive to the complexity of the loss landscape, and convergence issues that they face are not necessarily resolved by rerunning the algorithm. Also, Metient is not designed to consider subclonal copy number alternations (CNAs) when correcting its estimated variant allele frequencies for CNAs. Using the descendant cell fraction (DCF)[50] or phylogenetic cancer cell fraction (phyloCCF)[51] as inputs to Metient could solve this. Alternatively, one could input which clones are in which samples

directly into Metient instead of the allele frequencies. Finally, we note that choice of clustering and tree inference algorithm used when inputting data into Metient can impact both the clonality and phyleticity classifications. In an attempt to most accurately compare our migration histories to previously reported results, where possible, we use the same clustering and trees inferred for the original datasets.

In conclusion, we show that Metient offers a fast and adaptable, fully automated framework that leverages 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[52], PairTree[53], SPRUCE[54], CITUP[55], or EXACT[56]. Building on a previous approach as described by Wintersinger et al.[53], 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:

- A_{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 mutation cellular frequency to variant 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 $p(x_{kj}|\,\mathbf{F}_{kj}) = \mathrm{Binom}(A_{kj}|A_{kj} + R_{kj}, \omega_{kj}\,\mathbf{F}_{kj})$. Where $\mathbf{F} = \mathbf{U}\,\mathbf{B}$ is the mutation cellular frequency matrix, $\mathbf{B} \in \{0,1\}^{C\times M}$ 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 $\mathbf{U}_{kc}$ is the fraction of anatomical site k made up by 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. An alternative way to find a point estimate of $\mathbf{U}$ is using a previously described projection algorithm for this problem[52,53,56,57]. A point estimate $\mathbf{U}$ can be found by optimizing the following quadratic approximation to the binomial likelihood of $\mathbf{U}$ given $\mathbf{B}$ and $\mathbf{F}$:

$$LP(\mathbf{U}|\,\mathbf{B},\mathbf{F},\mathbf{W}) = min_{\hat{\mathbf{F}},\mathbf{U}}||\mathbf{W}\odot(\mathbf{F}-\hat{\mathbf{F}})||^2\ s.t.\ \mathbf{U}\,\mathbf{1} \leq 1,\ \mathbf{U} \geq 0,\ \hat{\mathbf{F}} = \mathbf{U}\,\mathbf{B} \quad \textbf{(1)}$$

where $||\cdot||$ is the Frobenius norm, $\mathbf{1}$ is a vector of 1s, $\mathbf{F}$ are the observed mutation frequencies, $\mathbf{W}$ is a $K \times M$ matrix of inverse-variances for each mutation in each sample derived from $\mathbf{F}$, and $\odot$ is the Hadamard, i.e., element-wise product. The definition for $\mathbf{W}$ is as described in previous work[53,56].

We use $\mathbf{U}$ (estimated in either of the previously described ways) to determine if a clone c is present in an anatomical site k. If c is present, we attach a witness node with label k (leaf nodes connected by dashed lines in Figure 1b, c) to clone c in clone tree $\mathbf{T}$. We deem c to be present in k if $\mathbf{U}_{kc} > 5\%$ 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.

Here the term "anatomical site" is used to describe a distinct tumor mass. If multiple samples are taken from the

same tumor mass, we combine them as described in "Bulk DNA sequencing pre-processing: Non-small Cell Lung Cancer Dataset".

Note that read count data are only used to determine which clones are present in which sites, if a matrix indicating the presence or absence of each clone in each anatomical site is available, it can be used as an input to replace the read count data. These clone-to-site assignment matrices can be derived, e.g., from single-cell data.

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, i.e., nodes with more than two children. Polytomy resolution is discussed in the section "Resolving polytomies".

Because we are interested in identifying multiple hypotheses of metastatic spread, Metient seeks 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 is first detected 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:

$$\text{ELBO}(q) = \mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})] + \mathbb{H}(\mathbf{V}) \quad \textbf{(2)}$$

Where $\mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})]$ evaluates a labeling based on 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 optimize using 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 $\mathbf{V}$ via the straight-through estimator of the Gumbel-Softmax distribution[23,24]. Starting with a matrix $\psi \in \{0,1\}^{K \times C}$, of randomly initialized values, where K is the number of anatomical sites and C is the number of clones, and each column represents the unnormalized log probabilities of clone c being labeled in site k:

1. At every iteration, for each clone c, we sample $g_{1c} ... g_{kc}$, k i.i.d. samples from Gumbel(0,1) and compute $y_{ic} = \psi_{ic} + g_{ic}$.

2. We then sample from the categorical distribution represented by the column vector $\psi_{:c}$ by setting $i^* = \text{argmax}_i \, y_{ic}$ and represent that sample with a one-hot encoding in $\mathbf{V}$, i.e., $\mathbf{V}_{ic} = 1$ if $i = i^*$, 0 otherwise.

3. Then we evaluate the ELBO(ν) where

$$\nu_{ic} = \frac{\text{exp}(y_{ic}/\tau)}{\sum_{j=1}^{k} \text{exp}(y_{jc}/\tau)} \qquad \text{for } i = 1, ..., k,$$

using a stochastic approximation based on $\mathbf{V}$, and take the gradient of this ELBO in the backward pass, thus implementing the straight-through estimator.

4. 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, ν resembles a one-hot vector.

At the end of training, as τ approaches 0, then the gradient becomes unbiased and ν approaches $\mathbf{V}$. In order to capture multiple modes of the posterior distribution, each representing different hypotheses about the migration history, 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[58] (with a different random initialization for each parallel process), 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 fit a patient cohort-specific parsimony model using the metastasis priors. 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.

To optimize the parsimony metric weights, Metient identifies a Pareto front of labeled trees for each patient and scores these trees based on (1) the weighted parsimony metrics and (2) the metastasis priors: genetic distance and, if appropriate anatomical labels are available, organotropism. These form the parsimony distribution and metastasis prior distribution, respectively. We initialize with equal weights and use gradient descent to minimize the cross entropy loss between the parsimony distribution and metastasis prior distribution for all patients in the cohort. Once the optimization converges, Metient rescores the trees on the Pareto front using the fitted weights, to identify the maximum calibrated parsimony solution, and genetic distance and organotropism are used to break ties between equally parsimonious migration histories. See Supplementary Information for a more detailed 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 a pan-cancer parsimony model calibrated to the four cohorts (melanoma, HGSOC, HR-NB, NSCLC) discussed in this work.

Defining the organotropism matrix

Data from the MSK-MET study[29] for 25,775 patients with annotations of distant metastases locations was downloaded from the publicly available cbioportal[59]. 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 the negative log frequencies as a “relative time to metastasis”, and only score 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 for users to instead input their own organotropism vector for each patient.

Evaluations on simulated data

We use the simulated data for 80 patients provided by MACHINA[17] to benchmark our method's performance. To prepare inputs to Metient, we use the same clustering algorithm and clone tree inference algorithm used in MACHINA (MACHINA[17] and SPRUCE[54], respectively) in order to accurately compare only our migration history inference algorithm (including polytomy resolution) against MACHINA's. 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 it does not improve performance on simulated data. (Supplementary Tables 4, 3). However, this performance is not necessarily indicative of polytomy resolution working poorly, because it actually finds more parsimonious solutions than the ground truth solution in 75% of simulated data (Supplementary Figure S6).

Evaluation metrics. We use the same migration graph and seeding clones F1-scores as MACHINA. Given a reconstructed migration graph **G**, its recall and precision with respect to the ground truth migration graph $\mathbf{G}^*$ 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 **G**, and multiple edges between the same two sites are included in $E(\mathbf{G})$. When there are multiple edges from site i to site j, $|E(\mathbf{G}) \cap E(\mathbf{G}^*)| = \min(a, b)$, where a and b are the number of edges from site i to site j in **G** and $\mathbf{G}^*$, respectively.

Recall and precision of the seeding 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, i.e., the subclone, associated with the clone nodes that have an outgoing migration edge. For example, $C(\mathbf{U}, \mathbf{V}) = \mathrm{A}, \mathrm{B}, \mathrm{C}$ in solution A of Figure 1c. The definition for seeding clones used in these evaluations is distinct from how we define seeding clones in the rest of the paper ("Defining colonizing clones, clonality, and phyleticity" in Methods). Specifically, if there is an edge between two nodes (u, v), where the labeling of u and v are not equal, we define the seeding clone as v. However in order to consistently compare to MACHINA in these evaluations, we use their definition and define the seeding clone as u. We note that identifying the mutations of v is generally a harder problem.

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 with a sample size of 1024, 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 colonizing clones, clonality, and phyleticity

A colonizing clone is defined as a node in a migration history whose parent is a different color than itself. There are two exceptions to this rule: when node a has a parent with a different color than itself, but the node is a witness node (Figure 1c) or a polytomy resolver node (e.g. A_POL in Supplementary Figure S7a). In these cases, these nodes do not represent any new mutations, but rather contain the same mutations as its parent. For these two cases, the colonizing clone is defined to be a's parent node.

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

Genetic clonality and site clonality are depicted schematically in Figure 4b.

To define phyleticity, we first extract all colonizing clones from a migration history. We then identify the colonizing clone closest to the root, s, i.e., the colonizing clone with the shortest path to the root. If all other colonizing clones are descendants of the tree rooted at s, the migration history is monophyletic, otherwise, it is polyphyletic. Under this definition, if a tree is monophyletic, then there are no independent evolutionary trajectories that give rise to colonizing clones. This is depicted schematically in Figure 4c.

In order to accurately compare our phyleticity measurements to TRACERx, we use their definition in Figure 6c and the TRACERx comparison analysis. To apply their definition to our migration histories, we extract colonizing clones as described above, and then determine if there is a Hamiltonian path in the clone tree that connects the colonizing clones. I.e., we determine if there is a path in the clone tree that visits each colonizing clone exactly once. If such a Hamiltonian path exists, we call this migration history monophyletic under the TRACERx definition, and polyphyletic otherwise.

Bootstrap sampling for fitting parsimony metric weights

Running Metient-calibrate on the 167 patients from the melanoma, HGSOC, HR-NB and NSCLC datasets infers a Pareto front of migration histories for each patient. For each dataset, we subset patients that have a Pareto front with size greater than one, and take 100 bootstrap samples of patients from this subset. Patients with a single solution on the Pareto front do not have an impact on the cross-entropy loss used to fit the parsimony metric weights. For

each bootstrap sample of patients, their Pareto front migration histories are used to fit the parsimony metric weights ("Calibrate alignment" in Supplementary Information). For each of the parsimony metric weights fit to a bootstrap sample, we evaluated how these weights would order the Pareto front, and evaluated how consistently the same top solution was chosen. We average the percent of times the same solution is ranked as the top solution across the four datasets.

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[3], breast[20], HGSOC[4], NSCLC[14], MSK-MET[29].

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. 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.
4. 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.
5. Nicolai J Birkbak and Nicholas McGranahan. Cancer genome evolutionary trajectories in metastasis. *Cancer cell*, 37(1):8–19, 2020.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
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. 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.

16. 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.
17. Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature genetics*, 50(5):718–726, 2018.
18. 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.
19. 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.
20. 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.
21. Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.
22. Joseph E Stiglitz. Pareto optimality and competition. *The Journal of Finance*, 36(2):235–251, 1981.
23. Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
24. 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.
25. Ernst Lengyel. Ovarian cancer development and metastasis. *The American journal of pathology*, 177(3):1053–1064, 2010.
26. Anirban K Mitra. *Ovarian cancer metastasis: a unique mechanism of dissemination*. IntechOpen, 2016.
27. Philippe Gui and Trever G Bivona. Evolution of metastasis: New tools and insights. *Trends in Cancer*, 8(2):98–109, 2022.
28. 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.
29. 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.
30. Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature genetics*, 48(3):238–244, 2016.
31. François Taddei, Miroslav Radman, John Maynard-Smith, Bruno Toupance, Pierre-Henri Gouyon, and Bernard Godelle. Role of mutator alleles in adaptive evolution. *Nature*, 387(6634):700–702, 1997.
32. Emily F Mao, Laura Lane, Jean Lee, and Jeffrey H Miller. Proliferation of mutators in a cell population. *Journal of bacteriology*, 179(2):417–422, 1997.
33. Christopher F Gentile, Szi-Chieh Yu, Sebastian Akle Serrano, Philip J Gerrish, and Paul D Sniegowski. Competition between high-and higher-mutating strains of escherichia coli. *Biology letters*, 7(3):422–424, 2011.
34. Samra Turajlic and Charles Swanton. Metastasis as an evolutionary process. *Science*, 352(6282):169–175, 2016.
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. Laurent Sansregret and Charles Swanton. The role of aneuploidy in cancer evolution. *Cold Spring Harbor perspectives in medicine*, 7(1):a028373, 2017.

37. 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.
38. 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.
39. Ellsworth C Alvord. Why do gliomas not metastasize? *Archives of Neurology*, 33(2):73–75, 1976.
40. 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.
41. 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.
42. Jonathan Sleeman, Anja Schmid, and Wilko Thiele. Tumor lymphatics. In *Seminars in cancer biology*, volume 19, pages 285–297. Elsevier, 2009.
43. Yeu-Tsu Margaret Lee and Deborah A Geer. Primary liver cancer: pattern of metastasis. *Journal of surgical oncology*, 36(1):26–31, 1987.
44. Wenrui Wu, Xingkang He, Dewi Andayani, Liya Yang, Jianzhong Ye, Yating Li, Yanfei Chen, and Lanjuan Li. Pattern of distant extrahepatic metastases in primary liver cancer: a seer based study. *Journal of Cancer*, 8(12):2312, 2017.
45. Matias Riihimäki, A Hemminki, Mahdi Fallah, Hauke Thomsen, Kristina Sundquist, Jan Sundquist, and Kari Hemminki. Metastatic sites and survival in lung cancer. *Lung cancer*, 86(1):78–84, 2014.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. Ethan Kulman, Rui Kuang, and Quaid Morris. Orchard: building large cancer phylogenies using stochastic combinatorial search. *arXiv preprint arXiv:2311.12917*, 2023.
53. 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.
54. 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.

55. 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.
56. 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.
57. Bei Jia, Surjyendu Ray, Sam Safavi, and José Bento. Efficient projection onto the perfect phylogeny model. *Advances in Neural Information Processing Systems*, 31, 2018.
58. 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.
59. 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.
60. David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.
61. 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.
62. 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.
63. Sierra Gillis and Andrew Roth. Pyclone-vi: scalable inference of clonal population structures using whole genome data. *BMC bioinformatics*, 21:1–16, 2020.
64. 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 Julia Simundza for her valuable feedback on this manuscript, and Deeksha Madala for coming up with the method name Metient. K.G. is supported by NIH grants R37CA266185, U2CCA233284 and U54CA274492. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 227260-01 (D.K.) and NIH/NCI Cancer Center Support Grant P30 CA008748 (Vickers).

Supplementary Figures and Tables

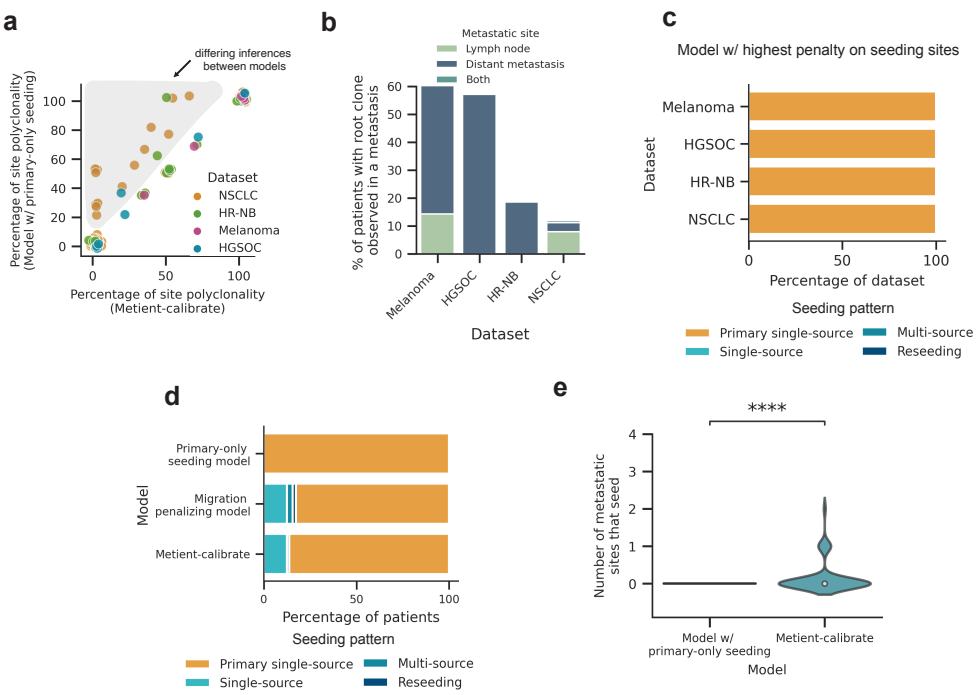

Figure S1. (a) A comparison of the percent of site polyclonal migrations for each patient's migration history when using the best migration history chosen by Metient (x-axis) vs. a model that assumes primary-only seeding (y-axis). **(b)** Percent of patients in each dataset with the root cancerous clone observed in a metastatic site. **(c)** 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. **(d)** 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 (primary-only seeding model), lowest number of migrations (migration penalizing model), or the top Metient-calibrate solution. **(e)** 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 a model that assumes primary-only seeding vs. Metient. Statistical significance assessed by a paired t-test, p=2.233e-06.

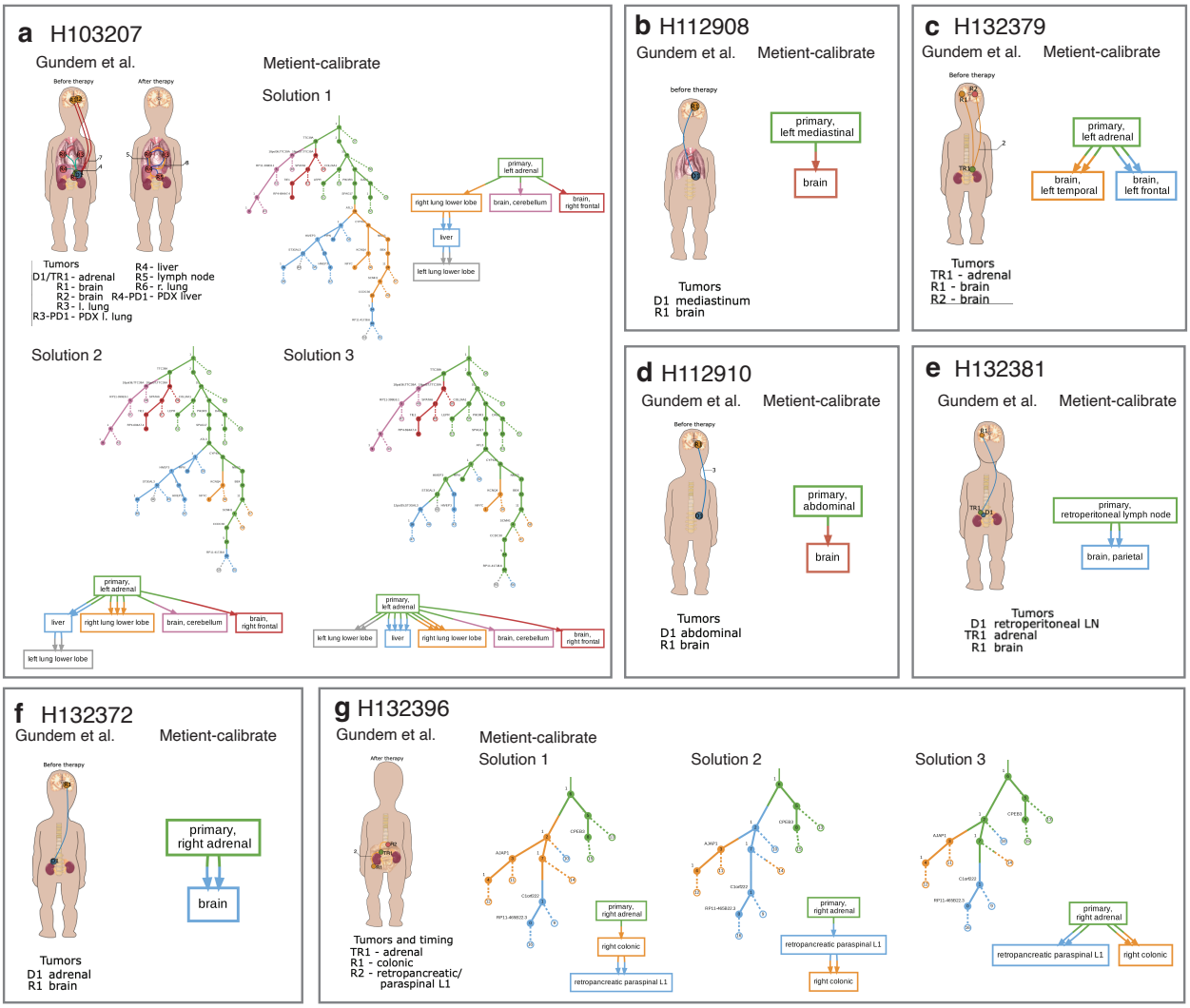

Figure S2. Comparison of Gundem et al.[9] reported body maps (left of each square) and Metient-calibrate inferred histories. The Metient-calibrate solutions with unique migration graphs on the Pareto front are shown. For example, in cases where there are multiple Pareto optimal migration histories with the same migration graph, only the migration history with the lowest loss is shown.

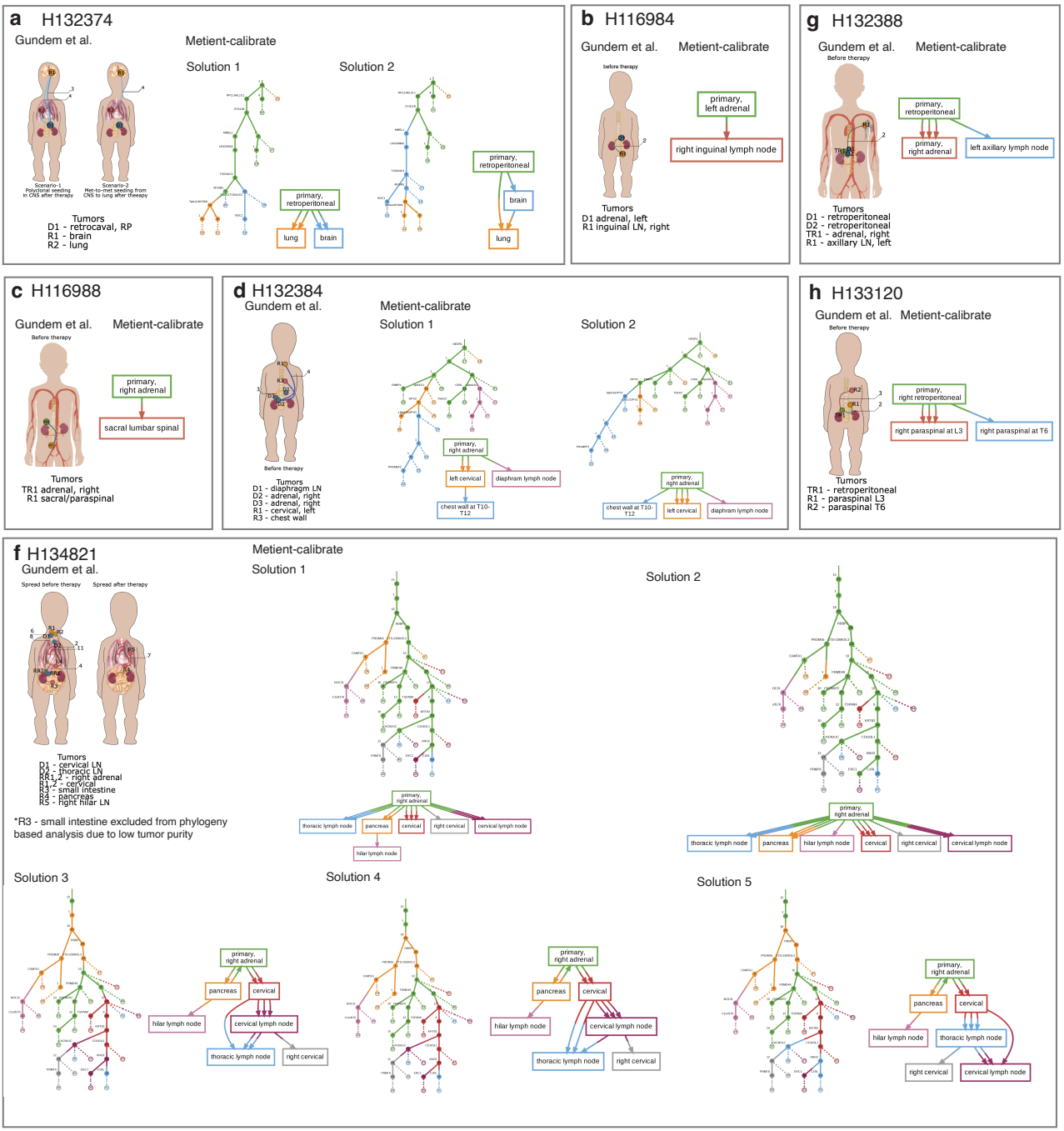

Figure S3. Comparison of Gundem et al.[9] reported body maps (left of each square) and Metient-calibrate inferred histories. The Metient-calibrate solutions with unique migration graphs on the Pareto front are shown. For example, in cases where there are multiple Pareto optimal migration histories with the same migration graph, only the migration history with the lowest loss is shown.

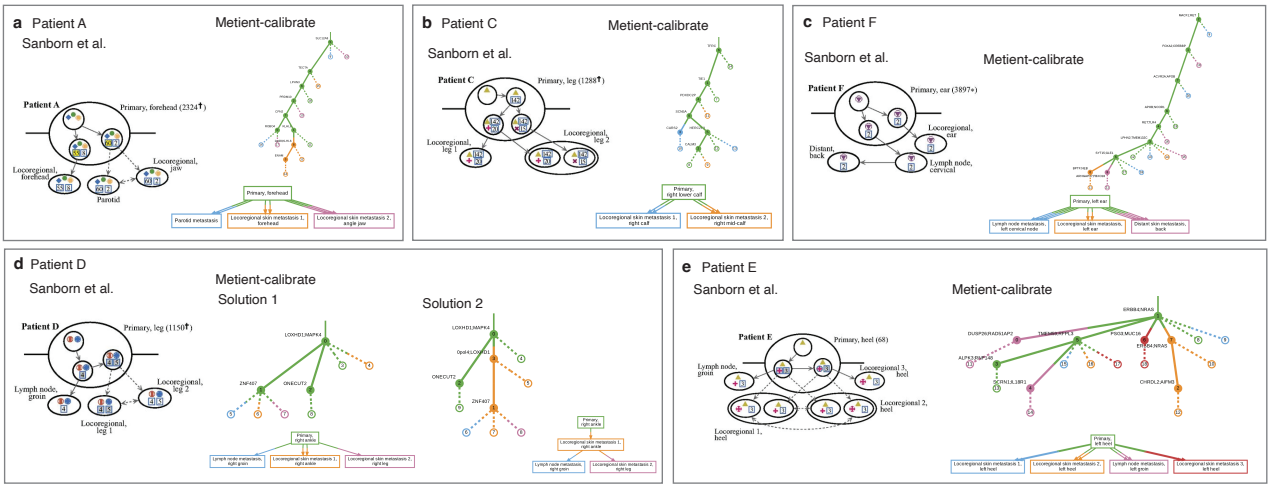

Figure S4. Comparison of Sanborn et al.[3] reported histories and Metient-calibrate inferred histories. In the Sanborn et al.[3] reported histories, solid lines denote probable dissemination patterns and dashed lines denote multiple possible paths. The Metient-calibrate solutions with unique migration graphs on the Pareto front are shown. For example, in cases where there are multiple Pareto optimal migration histories with the same migration graph, only the migration history with the lowest loss is shown.

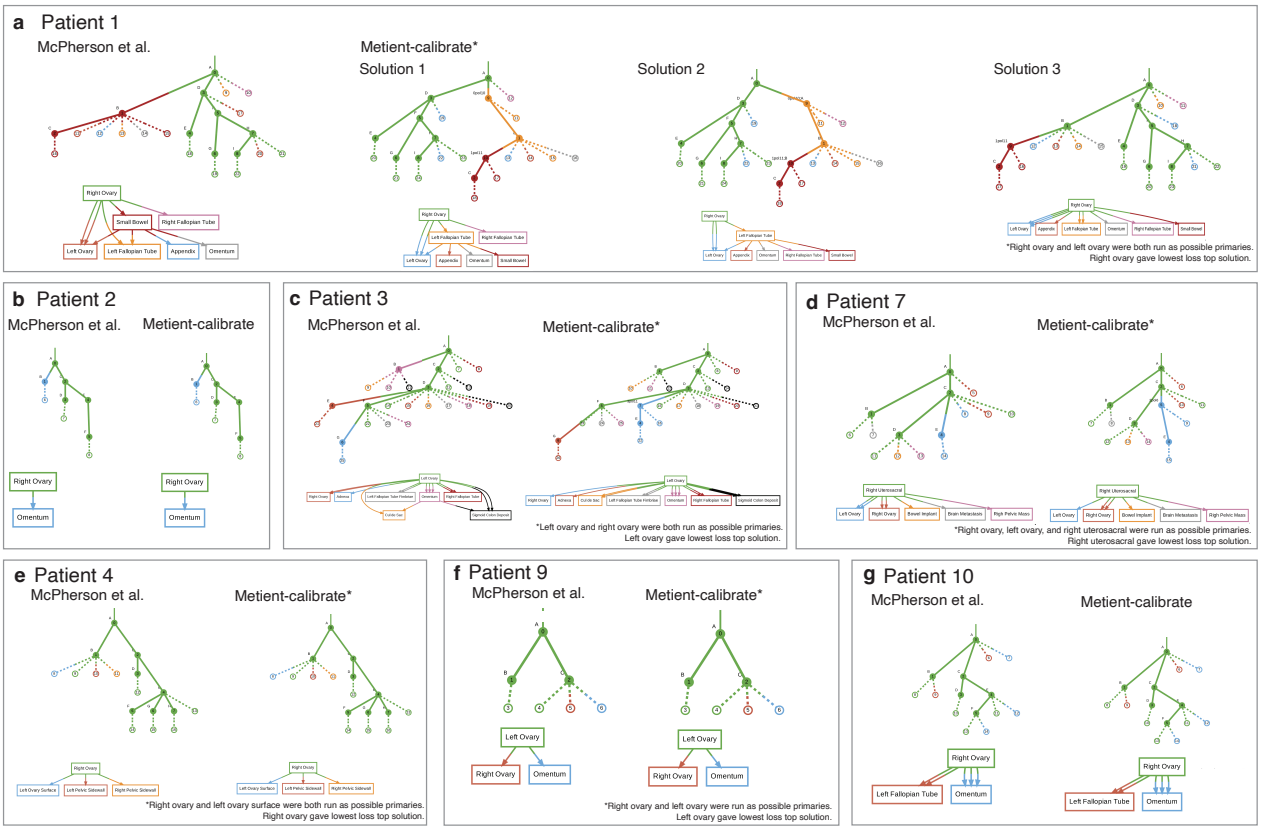

Figure S5. Comparison of McPherson et al.[4] reported histories and Metient-calibrate inferred histories. The Metient-calibrate solutions with unique migration graphs on the Pareto front are shown. For example, in cases where there are multiple Pareto optimal migration histories with the same migration graph, only the migration history with the lowest loss is shown. When multiple possible primaries were available, Metient-calibrate was run once with each possible primary, and the primary with the lowest loss solution is shown.

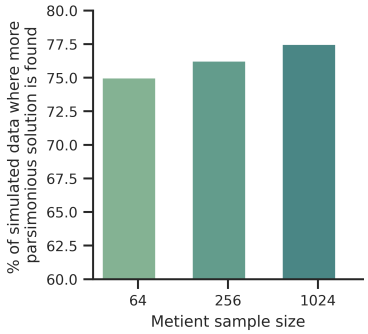

Figure S6. The percent of simulated data where a more parsimonious solution than ground truth is found when running Metient-1024 in calibrate mode with polytomy resolution. More parsimonious is defined as at least one of the parsimony metrics (migration, comigration and seeding site number) being less than the ground truth and all other metrics being equal.

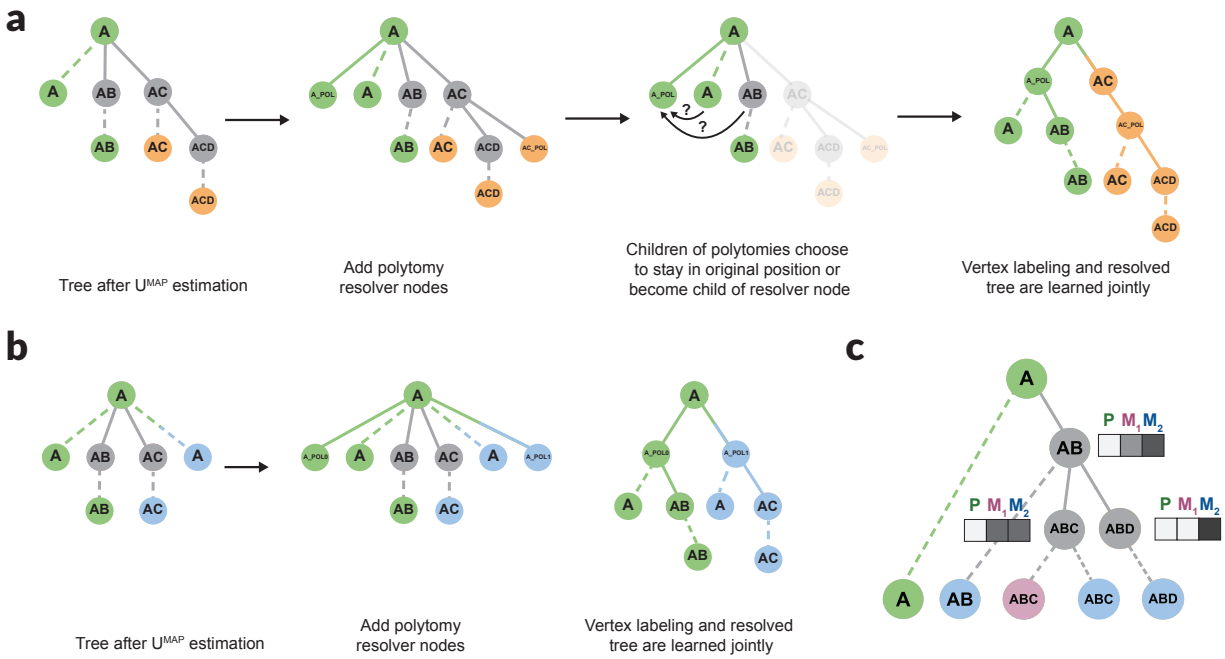

Figure S7. (a) Polytomy resolution algorithm with two nodes (A and AC) that have polytomies that can be resolved. **(b)** Polytomy resolution algorithm for a single node with four children and thus two resolver nodes. **(c)** Weight initialization is done such that nodes start with higher probabilities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

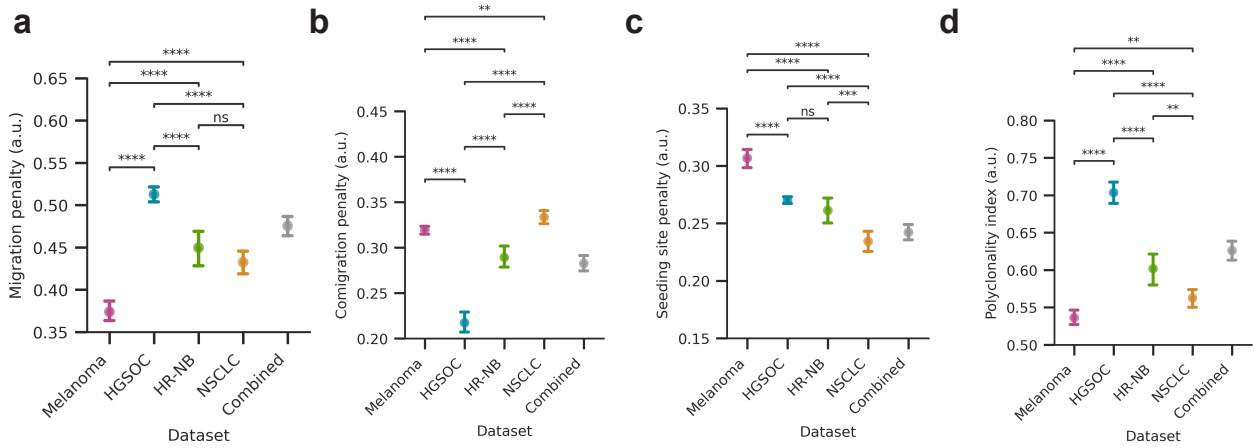

Figure S8. The **(a)** migration penalty/weight, **(b)** comigration penalty/weight, and **(c)** seeding site penalty/weight for each cohort, when taking 100 bootstrap samples of each cohort and fitting the weights to the bootstrapped sample. **(d)** The polyclonality index, which is $1 - (w_c/(w_m + w_c))$, 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.

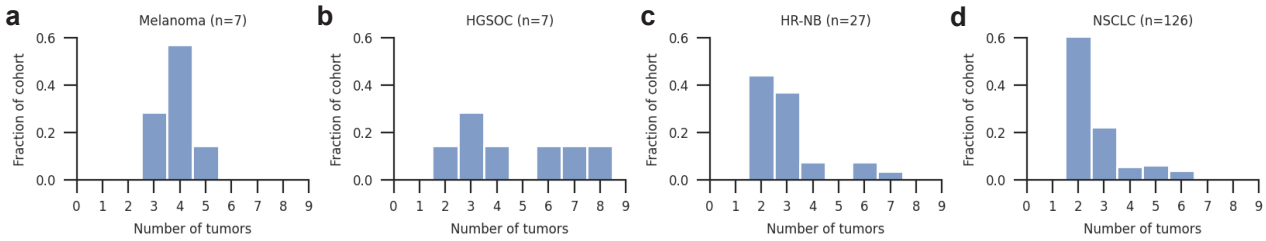

Figure S9. The distribution of tumors (number of distinct anatomical sites) for each cohort: **(a)** melanoma, **(b)** high-grade serous ovarian cancer (HGSOC), **(a)** high-risk neuroblastoma (HR-NB) and **(a)** non-small cell lung cancer (NSCLC).

Previous Methods for Migration History Inference

Method	Labels clone tree	Estimates clone proportions in sites	Models Complex Seeding	Multiple solutions	Organo-tropism	Genetic Distance	Polytomy Resolution
ClonEvol[15]	Y	Y	N	Y	N	N	N
Treeomics[16]	N	Y	N	N	N	N	N
MACHINA[17]	Y	Y	Y	N	N	N	Y
PathFinder[40]	Y	N	N	Y	N	Y	Y
Metient	Y	Y	Y	Y	Y	Y	Y

Table 1. Summary of previous methods which perform some aspect of migration history inference. Y = yes, N = no. Labels clone tree refers to whether the method infers the labels of the internal vertices of a clone tree (e.g. labeling clone AB as originating in lymph in Figure 1c, solution A). Estimates clone proportions in sites refers to whether the method infers the leaf nodes (witness nodes) (e.g. identifying that clone ABC is present in both lymph and liver in Figure 1c, solution A). Multiple solutions indicates whether a method outputs multiple possible migration histories.

Parsimony model	Migration number weight (w_m)	Comigration number weight (w_c)	Seeding site number weight (w_s)
$w_m >> w_c >> w_s$ (MACHINA model)	1000	100	1
$w_c >> w_m >> w_s$	100	1000	1
$w_s >> w_m >> w_c$	100	1	1000
$w_s >> w_c >> w_m$	1	100	1000
$w_c >> w_s >> w_m$	1	1000	100
$w_m >> w_s >> w_c$	1000	1	100

Table 2. The multiple parsimony models that Metient uses to build a Pareto front of solutions for a patient's data. Each parsimony model has a different relative weighting on each parsimony metric.

Average migration graph F1-scores

Method	Primary-only	Met-to-met	Macro-F1	Micro-F1
Evaluate (MP)	0.930	0.688	0.809	0.736
Evaluate (MP) + polyres	**0.983**	0.648	0.816	0.715
Evaluate (GD)	0.857	0.691	0.774	0.724
Evaluate (GD) + polyres	0.829	0.649	0.739	0.685
Calibrate	0.930	**0.716**	**0.823**	**0.759**
Calibrate + polyres	**0.983**	0.662	**0.823**	0.726
MACHINA	0.968	0.643	0.806	0.708

Table 3. Average F1-scores of migration graph for each broad seeding pattern (primary-only seeding or metastasis-to-metastasis seeding) 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. 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 migrating clone F1-scores

Method	Primary-only	Met-to-met	Macro-F1	Micro-F1
Evaluate (MP)	0.795	0.781	0.788	0.784
Evaluate (MP) + polyres	0.873	0.791	0.832	0.808
Evaluate (GD)	0.954	0.876	0.915	0.892
Evaluate (GD) + polyres	**0.979**	**0.928**	**0.954**	**0.939**
Calibrate	0.961	0.916	0.938	0.925
Calibrate + polyres	0.961	0.890	0.926	0.905
MACHINA	0.954	0.876	0.915	0.892

Table 4. Average F1-scores of migrating clones for each broad seeding pattern (primary-only seeding or metastasis-to-metastasis seeding) 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. Calibrate: Metient in calibrate mode, using genetic distance as the metastasis prior. polyres: polytomy resolution is used.

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 to optimize $\mathbf{V}$. Starting with a matrix $\psi \in \{0, 1\}^{K \times C}$, of randomly initialized values, where K is the number of anatomical sites and C is the number of clones, and each column represents the unnormalized log probabilities of clone c being labeled in site k:

1. At every iteration, for each clone c, we sample $g_{1c} ... g_{kc}$, k i.i.d. samples from Gumbel(0,1) and compute $y_{ic} = \psi_{ic} + g_{ic}$. Where a sample g from the Gumbel is computed as:

$$g = -\log(-\log(u)) \quad \text{where} \quad u \sim \text{Uniform}(0, 1) \tag{S4}$$

2. We then sample from the categorical distribution represented by the column vector $\psi_{:c}$ by setting $i^* = \text{argmax}_i \, y_{ic}$ and represent that sample with a one-hot encoding in $\mathbf{V}$, i.e., $\mathbf{V}_{ic} = 1$ if $i = i^*$, 0 otherwise.

3. Then we evaluate the ELBO(ν) where

$$\nu_{ic} = \frac{\exp(y_{ic}/\tau)}{\sum_{j=1}^{k} \exp(y_{jc}/\tau)} \qquad \text{for } i = 1, ..., k,$$

using a stochastic approximation based on $\mathbf{V}$, and take the gradient of this ELBO in the backward pass, thus implementing the straight-through estimator.

4. 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, ν resembles a one-hot vector.

At the end of training, as τ approaches 0, then the gradient becomes unbiased and ν approaches $\mathbf{V}$. In order to capture multiple modes of the posterior distribution, each representing different hypotheses about the migration history, 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[58] (with a different random initialization for each parallel process), where x is equal to the sample size and is calculated according to the size of the inputs ($\propto K^C$).

Using the Gumbel-Softmax reparameterization as described above, we approximate the expectation in the ELBO with a sample of $\mathbf{V}$, which we denote $\tilde{\mathbf{V}}$:

$$\mathbb{E}_{q(\mathbf{V})}[\log p(\mathbf{U}, \mathbf{T}, \mathbf{V})] \approx \log p(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) \tag{S5}$$

$$\mathbb{H}(\mathbf{V}) \approx -\sum_{j=1}^{C} \sum_{k=1}^{K} q(\tilde{\mathbf{V}}_{jk}) \log q(\tilde{\mathbf{V}}_{jk}) \tag{S6}$$

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[17], the maximum parsimony metrics are defined as:

- **migration number** m: Given clone tree $\mathbf{T}$ and clone tree 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 clone tree 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 clone tree 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 s.

$$p_m(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) = w_m \cdot m + w_c \cdot c + w_s \cdot s \tag{S7}$$

$$m = \sum_{ij} \mathbf{G} - Trace(\mathbf{G})$$

$$s = \sum_{j=1}^{n} \left(\left(\sum_{i=1}^{m} \left(\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K) \right)_i \right)^* \right)_j$$

$$c = \sum_{ij} \mathbf{G}^*_{ij} - Tr(\mathbf{G}^*) + \sum_{ij} \left(\sum_{l=1}^{m} \left(\mathbf{P} \odot (\mathbf{W} \odot \mathbf{X}) \right)_l \right)_{ij}$$

where $\mathbf{G} = \tilde{\mathbf{V}}\mathbf{T}\tilde{\mathbf{V}}^T, \mathbf{P} = \left(\mathbf{T} \vee \mathbf{I}_N\right)^{N-1}, \mathbf{X} = \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}, \mathbf{Y} = \sum_{i=1}^{m} \left((\tilde{\mathbf{V}}\mathbf{T}\tilde{\mathbf{V}}^T \odot (\mathbf{J}_{CK} - \mathbf{V}^T) \right), \mathbf{Z}^* = \mathbf{sign}(\mathbf{Z})$. $\vee$ represents boolean matrix multiplication, $\mathbf{I_n}$ is a $n \times n$ identity matrix, $\odot$ is the Hadamard, i.e., element-wise product, 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:

$$p_g(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) = \frac{w_g}{m} \sum_{ij} -log(\mathbf{D}) \odot \mathbf{T} \odot (\mathbf{J}_C - \mathbf{X}) \tag{S8}$$

where $\mathbf{J}_C$ is a square matrix of ones, $\odot$ is the Hadamard, i.e., element-wise product, and $\mathbf{X} = \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}$. The product $\mathbf{T} \odot \mathbf{J}_C - \mathbf{X}$ 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 implicitly 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:

$$p_o(\tilde{\mathbf{V}}, \mathbf{U}, \mathbf{T}) = \frac{w_o}{m_p} \sum_{i=1}^{K} -log(\mathbf{o}) \odot (\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K))_{p,i} \tag{S9}$$

where $\mathbf{G} = \tilde{\mathbf{V}}\mathbf{T}\tilde{\mathbf{V}}^T$, $\odot$ is the Hadamard, i.e., element-wise product, $\mathbf{J}_K$ is a square matrix of ones, and $\mathbf{I_K}$ is the identity matrix. The product $(\mathbf{G} \odot (\mathbf{J}_K - \mathbf{I}_K))$ 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 implicitly penalize migration histories with more migrations through this scoring.

B. Calibrate alignment

A parsimony model is represented by a set of parsimony weights – w_m, w_c, and w_s – assigned, respectively, to the number of migrations (indicated by m), comigrations (c), seeding sites (s). A migration history's parsimony score, p, is the model-weighted average of these three parsimony metrics, i.e., $p = w_m m + w_c c + w_s s$ (Equation S7). Different parsimony models favor different histories on the Pareto front. To fit a parsimony model to a cancer type-specific cohort, we look at how well the maximum parsimony distribution 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 $T^{(n)}$ migration histories. Each migration history 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 migration histories 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 migration histories 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 migration history 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.,

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

and the MLE estimate of the parameters is $\theta^* = \text{argmax}_\theta E(\theta)$. w_n is set to $log(E/(r \cdot b))$, where E is the number of internal edges of a patient's clone tree, r is the number of possible primaries for the patient, and b is the number of possible clone trees for a given patient (so as not to bias towards patients with multiple possible primaries or multiple possible clone trees). Since the number of edges is equal to the maximum number of migrations possible in a tree, it is also equal to the number of possible genetic distance observations that that tree can provide in the genetic distance scoring of that tree. Therefore, w_n is representative of the information content that a patient can provide when fitting θ.

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 (7/7 patients; Supplementary Figure S4). For patient F (Supplementary Figure S4c), 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. In the second best solution predicted for patient D, Metient predicts 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 S4d). We also predict a primary single-source solution on the Pareto front which is another possible path proposed by the authors (Supplementary Figure S4d).

C.2. Comparisons to HGSOC patients from McPherson et al.. In the seven HGSOC patients, predicted migration histories by McPherson et al.[4] were made available using an algorithm that only minimizes migrations (Sankoff algorithm[60]). We find that four out of seven patients are in complete agreement (Supplemental Figure S5). 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 S5a). 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 S5c). 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 S5d).

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.[9], and found that we predict the same overall site-to-site migrations for 10 out of 14 cases. For patient H103207, we predict their before therapy pattern on the Pareto front (Solution 3 in Figure S2a), but we prioritize two solutions with metastasis-to-metastasis seeding between the lung and the liver. A subset of this seeding between the liver and two lobes of the lung is suggested in their after therapy hypothesis of spread (Figure S2a). While Gundem et al. suggest seeding between the two lobes of the lung as well as from each lobe of the lung to the liver, we infer a simpler, serial progression, where the right lung lower lobe seeds the liver, which subsequently seed the left lung lower lobe (Solution 1 in Figure S2a). For patient H132396, Metient prioritizes migration histories with fewer migrations (Solutions 1 and 2 in Figure S2g), but presents the expert annotation on the Pareto front (Solution 3 in Figure S2g). For patient H132384, Metient proposes bone-to-bone secondary metastasis formation (Solution 1 in Figure S3d), but again presents the expert annotation on the Pareto front (Solution 2 in Figure S3d). For patient H134821, we infer the same pancreas to hilar lymph node seeding proposed by the authors as spread after therapy, but suggest that all other metastases were seeded directly by the primary (Solution 1 in Figure S3f). However, we report the same metastasis-to-metastasis seeding between the cervical and thoracic lymph nodes and cervical metastases as the authors in alternative solutions on the Pareto front (Solutions 3-5 in Figure S3f).

D. 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 number 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 S1c). The migration penalizing model chooses a primary single-source seeding explanation in 82.6% of patients, and Metient falls in between the two, choosing a primary single-source seeding explanation in 86.2% of patients (Supplementary Figure S1d). 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 assumption of primary, single-source seeding 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 assumes primary, single-source seeding promotes migration histories with only one seeding site (Supplementary Figure S1e). In turn, such a model infers a higher fraction of polyclonal migrations (Supplementary Figure S1a) 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, 9% (15/167) of patients have differing colonizing clones between the two models, changing the inference of which clones, and therefore which mutations, have metastatic competence.

E. Bulk DNA sequencing pre-processing

E.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 genomic locus in each sample. For a given sample s and variant allele j, the variant read probability ω_{js} is the probability of observing a read with the variant allele at that locus in a cell with the mutation, and is calculated as:

$$\omega_{js} = M_{js}/N_{js} \tag{S10}$$

where M_{js} is the number of copies of the mutant allele j in sample s in the cells that contain the mutant allele, and N_{js} is the average number of copies at the genomic locus of the mutation j in all cells in s.

To account for the fact that cancer cells frequently have different numbers of copies at genomic loci compared to normal cells, N_{js} is calculated as:

$$N_{js} = \rho_s N_{js}^{(c)} + (1 - \rho_s) N_{js}^{(h)} \tag{S11}$$

where:

- $N_{js}^{(c)}$ is the population average copy number of the locus which contains mutant allele j in the cancer cell population

- $N_{js}^{(h)}$ is the copy number at the genomic locus of mutation j in the normal cell population. In diploid cells this is 2, and in haploid cells this is 1.

- ρ_s is the tumor purity of sample s

ρ_s and $N_{js}^{(c)}$ (and sometimes N_{js}) are normal outputs from a copy number calling pipeline. We suggest setting $M_{js} = 1$ unless there is strong evidence that the j allele has been amplified. In this case, allele-specific copy number callers provide the major allele copy number A_{js} and minor allele copy number B_{js}, where $N_{js}^{(c)} = A_{js} + B_{js}$, and $M_{js} = A_{js}$. When a locus is impacted by many different CNAs, accurately estimating M_{js} is challenging since there are likely subclonal changes in the multiplicity of the j allele, in which case we recommend excluding these mutations. For additional information on how to estimate M_{js} and N_{js} please refer to Tarabichi et al.[61].

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.[53]:

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

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.

E.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.[20]. 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[62]. 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).

E.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[4]. 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.

E.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.[3], 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-VI 0.1.3 for clonal analysis[63]. PyClone-VI's fit command was run with all default parameters. 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 were calculated using major copy number, minor copy number and tumor purity according to Equation S10. We ran Metient-calibrate with the clonal proportions estimated

by running Orchard (i.e., η in Orchard's output) using all default configurations and with polytomy resolution.

E.5. Neuroblastoma Dataset. Access to multi-WGS data for 45 neuroblastoma patients was provided through dbGaP accession phs03111[9]. 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[64] 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 S10. We ran Metient-calibrate with the clonal proportions estimated by running Orchard (i.e., η in Orchard's output) using all default configurations 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.

E.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.