Inferring parsimonious migration histories for metastatic cancers

Mohammed El-Kebir\textsuperscript{1,3}, Gryte Satas\textsuperscript{1,2} and Benjamin J. Raphael\textsuperscript{1,4}\textsuperscript{*}

Metastasis is the migration of cancerous cells from a primary tumor to other anatomical sites. Although metastasis was long thought to result from monoclonal seeding, or single cellular migrations, recent phylogenetic analyses of metastatic cancers have reported complex patterns of cellular migrations between sites, including polyclonal migrations and reseeding. However, accurate determination of migration patterns from somatic mutation data is complicated by intratumor heterogeneity and discordance between clonal lineage and cellular migration. We introduce MACHINA, a multi-objective optimization algorithm that jointly infers clonal lineages and parsimonious migration histories of metastatic cancers from DNA sequencing data. MACHINA analysis of data from multiple cancers shows that migration patterns are often not uniquely determined from sequencing data alone and that complicated migration patterns among primary tumors and metastases may be less prevalent than previously reported. MACHINA’s rigorous analysis of migration histories will aid in studies of the drivers of metastasis.

Cancer is an evolutionary process in which somatic mutations accumulate in a population of cells, yielding a heterogeneous primary tumor that is composed of multiple cellular subpopulations with different complements of mutations. During cancer progression, cancerous cells may migrate to other anatomical sites, seeding new metastases at these sites. Because metastasis causes up to 90% of deaths from solid tumors\textsuperscript{1}, understanding this process is of critical importance in improving the diagnosis and treatment of cancer\textsuperscript{2–5}.

Until recently, the dominant model of metastasis was the ‘monoclonal theory’, which posits that each metastasis is founded by a single founder cell\textsuperscript{6}. Recent analyses of high-throughput DNA sequencing data have suggested more complex migration patterns between primary tumors and metastases\textsuperscript{6–8}. In particular, several studies\textsuperscript{9–13} have reported ‘polyclonal seeding’ (in which cancer cells from one or more anatomical sites seed a metastasis) and ‘reseeding’ (in which cancer cells migrate from a metastasis back to the primary tumor or back to other metastases). Such complex migration patterns can result in highly heterogeneous metastases with aggressive phenotypes\textsuperscript{7,14}. Polyclonal seeding may be explained either by the simultaneous migration, or conigration, of multiple cells from distinct clones or by multiple waves of migrating cells, with each wave being composed of cells from the same anatomical site\textsuperscript{6}. Of note, evidence from mouse models suggests that cancer cells migrate together in clusters and that these cell clusters may be more efficient at forming metastases than single cells\textsuperscript{14–20}.

Many recent analyses infer a cancer cell migration history for an individual patient from phylogenetic trees constructed from the somatic mutations measured in multiple anatomical sites. In most cases, this inference relies on a combination of two assumptions that do not generally hold in cancer sequencing data. The first assumption, sample homogeneity, assumes that each sequenced sample is a homogeneous population of cells with identical somatic mutations. Many published analyses of bulk sequencing data from matched primary tumors and metastases\textsuperscript{21–25} implicitly rely on this assumption by using standard phylogenetic techniques such as neighbor joining, maximum parsimony or maximum likelihood to the

mutations measured at each sequenced region (Fig. 1a). However, it is well established that tumors exhibit extensive intratumor heterogeneity\textsuperscript{6,39–40} and thus it is unlikely that a bulk tumor sample is homogeneous. Multi-region sequencing\textsuperscript{1} reduces, but does not eliminate, this heterogeneity, as each region remains a mixture of cells. The assumption of sample homogeneity can result in the construction of phylogenetic trees that have unexpected implications for tumor evolution, such as suspiciously high rates of homoplasy, or convergent evolution\textsuperscript{11} (Supplementary Fig. 1). To avoid the assumption of sample homogeneity, one can identify subpopulations of cells with the same somatic mutations (referred to as clones) by clustering mutations according to their variant allele frequencies (VAFs)\textsuperscript{41,42}. One then uses specialized phylogenetic algorithms to construct ‘clone trees’ from mixed samples\textsuperscript{39–40} (Fig. 1b). However, these specialized techniques have been used only sporadically in the analysis of metastasis\textsuperscript{6}.

The second assumption, mutation–migration concordance, states that a tree constructed from the mutations present in clones at multiple anatomical sites determines the history of cellular migrations. In other words, the migration history follows directly from the topology and branch lengths of a phylogenetic tree constructed from mutations. The mutation–migration concordance assumption underlies many recent analyses of metastatic cancers\textsuperscript{21–25,27–29,34–36,38} as well as a recent method, Treeomics\textsuperscript{41}, for reconstructing clone trees from sequencing data of metastatic tumors\textsuperscript{32,33}. However, there are two problems with the mutation–migration concordance assumption. First, the migration history does not uniquely follow from the structure of a phylogenetic tree because the phylogenetic tree does not encode the anatomical sites of ancestral clones, as has been noted previously\textsuperscript{32,34}. Second, although somatic mutations can be used as a marker for cellular lineage, mutations do not directly model the history of cellular migrations between anatomical sites. In particular, although cellular lineage is appropriately modeled as a tree—because a cell divides into two daughter cells—migrations do not necessarily follow a tree topology (Fig. 1, center). Indeed, complex migration patterns, such as polyclonal seeding or reseeding, cannot be modeled by a tree.

\textsuperscript{1}Department of Computer Science, Princeton University, Princeton, NJ, USA. \textsuperscript{2}Department of Computer Science, Brown University, Providence, RI, USA. \textsuperscript{3}Present address: Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA. \textsuperscript{*}e-mail: braphael@princeton.edu
An essential missing ingredient for evaluating different hypotheses about the migration pattern that occurred in metastasis is an explicit model that assesses how well each hypothesis fits the observed sequencing data. Here we introduce a rigorous computational model that represents migration patterns using a migration graph, a directed multigraph that describes the migration of cells between anatomical sites (Fig. 1b). We introduce a taxonomy of migration patterns, streamlining the ambiguous language in the literature. Notably, we show that minimizing the number of migrations is insufficient to distinguish different migration patterns and that additional biologically motivated criteria are necessary to distinguish parsimonious migration histories.

Our computational model forms the basis for ‘metastatic and clonal history integrative analysis’ (MACHINA), an algorithm that does not assume sample homogeneity and mutation–migration concordance. MACHINA operates in three distinct modes. In the first mode, MACHINA infers parsimonious migration histories (PMH) for a given clone tree. In the second mode, MACHINA infers parsimonious migration histories and simultaneously resolves uncertainties in a given clone tree (PMH-TR). In the third mode, MACHINA jointly infers a parsimonious migration history and a clone tree that best fit measured mutation frequencies (PMH-TI). Using simulated metastatic cancers, we show that MACHINA more accurately recovers clone trees and migration histories than existing approaches that assume sample homogeneity and/or that use limited models of migration. Using DNA sequencing data from metastatic ovarian, prostate, breast and skin cancers, we demonstrate that MACHINA provides a rigorous approach to evaluate alternative migration histories. We show that some previous reports of metastasis-to-metastasis migrations, polyclonal migrations, reseeding or multisource seeding are not well supported by the data. By improving the analysis of migration histories, MACHINA will enable further studies of the drivers and mechanisms of metastasis.

**Results**

A computational model for migration histories. Suppose we measure all clones that are present in m anatomical sites of a metastatic cancer. These clones are distinguished by their somatic mutations and their anatomical locations. We can then define a clone tree T that describes the cell division history, or cell lineage, of the clones and the mutations that accumulated as a result of these cell divisions (Fig. 1b). However, a clone tree does not describe the migration history, the process by which cells or clones moved between anatomical sites. This is because, although we know the anatomical sites and the mutations that occurred in the clones, we do not know the internal vertices of a metastatic cancer (Fig. 1b). MACHINA jointly infers a parsimonious migration history and a clone tree that best fit measured mutation frequencies (PMH-TI). Using simulated metastatic cancers, we show that MACHINA more accurately recovers clone trees and migration histories than existing approaches that assume sample homogeneity and/or that use limited models of migration. Using DNA sequencing data from metastatic ovarian, prostate, breast and skin cancers, we demonstrate that MACHINA provides a rigorous approach to evaluate alternative migration histories. We show that some previous reports of metastasis-to-metastasis migrations, polyclonal migrations, reseeding or multisource seeding are not well supported by the data. By improving the analysis of migration histories, MACHINA will enable further studies of the drivers and mechanisms of metastasis.
The ability to distinguish different labelings $\ell$ of a clone tree $T$ requires a biologically motivated scoring function. In the simplest model, we assume that each migration between anatomical sites is monoclonal and comprises cells from a single clone. The migration number $\mu(T, \ell)$ counts the number of such monoclonal migrations. We assume that migrations are rare events in the evolutionary history of the cancer and that migrations between all pairs of anatomical sites are equally likely. Thus, we appeal to the maximum-parsimony principle and aim to find a labeling with the minimum migration number $\mu^*(T) = \min \mu(T, \ell)$. Finding the minimum migration labeling is an instance of the small phylogeny problem and can be solved by using the Sankoff algorithm, as noted by Slatkin and Maddison, and later by McPherson et al.

Notably, the maximum-parsimony labeling is not unique and there are typically many vertex labelings of $T$ with the same minimum migration number $\mu(T)$ but markedly different structures of migration between anatomical sites (Fig. 2). We introduce the migration graph $G$ as a mathematical representation of this structure. The vertices of $G$ are anatomical sites, and directed edges indicate migrations between anatomical sites. From a vertex labeling $\ell$ of a clone tree $T$, we obtain the migration graph $G$ by collapsing all of the vertices labeled by the same anatomical site into a single vertex and by removing any self-loops that connect the same vertex. Formally, the migration graph is a multigraph, as there may exist multiple directed edges between the same pair of anatomical sites.

We can classify a migration graph $G$ according to its migration pattern, which is defined by two criteria (Fig. 3a). The first criterion is the presence of a single edge between a pair of anatomical sites, which indicates monoclonal migration, versus the presence of multiple edges between a pair of anatomical sites, which indicates polyclonal migration. We say that a migration graph is polyclonal (p) if the graph contains at least one multi-edge; otherwise, the graph is monoclonal (m). The second criterion classifies the topology of migration graph $G$ into three types: (i) single-source seeding (S), in which for each anatomical site all of the migrations into the site originate from a single anatomical site, and thus $G$ is a tree; (ii) multisource seeding (M), in which at least one anatomical site has clones that originate from different anatomical sites but no migrations return to the originating sites, and thus $G$ is a directed acyclic graph; and (iii) reseeding (R), in which at least one migration returns to an originating anatomical site, and thus $G$ has a directed cycle. We denote a migration pattern by combining the two criteria, for example, ‘mS’ denotes monoclonal (m) single-source seeding (S).

Recent experimental evidence suggests that tumor cells can simultaneously travel in groups through the bloodstream or the lymphatic system and settle at other anatomical sites, which suggests that polyclonal migrations may not be unusual. Thus, we introduce a second model, the comigration model, which counts simultaneous, or polyclonal, migration of multiple clones between the same anatomical sites as a single event. We define the comigration number $\gamma(T, \ell)$ as the smallest number of monoclonal and polyclonal migrations incurred by vertex labeling $\ell$ of a clone tree $T$. Although the migration number $\mu(T, \ell)$ equals the number of edges in the migration graph $G$, the comigration number $\gamma(T, \ell)$ equals the number of multi-edges in the case $G$ is acyclic (see the Supplementary Note for the precise definition of $\gamma(T, \ell)$).

Each migration pattern constrains the migration number $\mu$ and comigration number $\gamma$ in a few ways (Fig. 3b). First, because each metastasis is seeded by at least one migrating clone, any vertex labeling of a clone tree $T$ with $m$ anatomical sites must have a migration number of at least $\mu_{\min} = m - 1$ and a comigration number of at least $\gamma_{\min} = m - 1$. Second, a vertex labeling $\ell$ of $T$ corresponds to an $S$ pattern if and only if $\gamma(T, \ell) = \gamma_{\min} = m - 1$. In general, vertex labelings with an $S$ pattern always exist. If such a labeling also has the minimum possible migration number $\mu_{\min}$, then the labeling corresponds to an $mS$ pattern; otherwise, the labeling corresponds to a $pS$ pattern. Finally, vertex labelings $\ell$ correspond to $M$ and $R$ patterns if and only if the comigration number $\gamma(T, \ell) > \gamma_{\min}$.

The interplay between the migration pattern, the migration number and the comigration number implies that the analysis of migration histories is a constrained multi-objective optimization problem. However, parsimonious vertex labelings found by the Sankoff algorithm optimize only a single objective, the migration number, and do not consider tradeoffs between the migration pattern, the migration number and the comigration number. We developed an algorithm, MACHINA, that is able to solve three constrained multi-objective optimization problems: the PMH problem, the PMH-TR problem and the PHM-TI problem (Fig. 4). We validated MACHINA using simulated metastatic tumors (Fig. 5; see Methods for further details).

Comigrations in ovarian cancer. We used MACHINA to analyze the cancer cell migration history of seven patients with metastatic ovarian cancer from McPherson et al. These data included whole-genome and targeted sequencing on a total of 68 samples from different anatomical sites, including the left (LOv) and right (ROv) ovary, as well as various metastases. McPherson et al. constructed a clone tree $T$ for each patient by clustering mutations that had similar cell frequencies across different anatomical sites, and then they determined the evolutionary relationships between the clusters using a Dollo parsimony model. Next, they found a minimum migration labeling, i.e., a labeling of the internal vertices of $T$ by anatomical sites with the minimum migration number $\mu(T)$. Because the anatomical site of the primary tumor was unknown, McPherson et al. selected the primary tumor to be at the anatomical site that incurred the fewest number of migrations. For six of the seven patients, the primary site was inferred to be either the left or the right ovary; however, for one patient the primary site was inferred to be the right uterosacral ligament.

We used MACHINA (in PMH mode) to find a parsimonious migration history for the reported clone tree $T$ for each patient. For all of the patients, we found that the reported vertex labeling...
Fig. 3 | Migration history analysis requires evaluation of tradeoffs between migration pattern, migration number and comigration number. a. The taxonomy of migration patterns is defined using two different criteria. First, the migration graph \( G \) is polyclonal (\( p \)) if it contains multi-edges; otherwise, the graph is monoclonal (\( m \)). Second, the topology of the migration graph defines the migration pattern. In single-source seeding (\( S \)), each anatomical site is seeded by clones originating from at most one anatomical site and the migration graph \( G \) is a tree. In multisource seeding (\( M \)), at least one anatomical site is seeded by clones originating from more than one site; however, the migration graph \( G \) is acyclic. In reseeding (\( R \)), clones migrate back and forth between anatomical sites and the migration graph has a directed cycle. b. The 2D plot shows that each migration pattern constrains the migration number \( \mu \) and the comigration number \( \gamma \). Note that \( \mu \geq \gamma \), as the comigration model allows for simultaneous migrations of clones. Moreover, with \( m \) anatomical sites, the minimum possible comigration number \( \gamma_{\min} = m - 1 \), and it is achieved by only a single-source seeding (\( S \)) pattern. Additionally, the monoclonal single-source seeding (\( mS \)) pattern also has the minimum possible migration number \( \mu_{\min} = m - 1 \). In contrast, the \( M \) and \( R \) patterns have comigration number \( \gamma \geq m \), with the polyclonal \( pM \) and \( pR \) patterns having migration numbers \( \mu \geq m + 1 \). Labels of individual points on the graph indicate the scores of the corresponding clone tree labelings in Fig. 2 (as indicated in the plot).

Fig. 4 | The MACHINA algorithm for joint clone tree inference and migration history analysis. a. One mode of MACHINA solves the PMH problem. Here one is given a set \( P \) of allowed migration patterns and a clone tree \( T \), whose leaves correspond to extant clones and in which each leaf is labeled by the anatomical site where the clone is present. The task is to infer a vertex labeling \( \ell \) of \( T \) that minimizes the migration number \( \mu \) and comigration number \( \gamma \). b. Another mode of MACHINA solves the PMH-TI problem. Here one does not directly observe the clone tree but only mutation frequencies, whose uncertainty is recorded as confidence intervals with corresponding frequency matrices \( (F, F') \). In addition, one is given a set \( P \) of allowed migration patterns. The task is to infer a frequency matrix \( F' \), a clone tree \( T \) and vertex labeling \( \ell \) of \( T \) that minimize the migration number \( \mu \) and comigration number \( \gamma \).
was among the output of vertex labelings derived by MACHINA (Supplementary Table 5). We found that three of the seven patients (patients 1, 3 and 7) admitted multiple vertex labelings that achieved the minimum migration number \( \mu(T) \) but that differed considerably in the comigration number and the structure of migrations. For example, McPherson et al.12 reported vertex labeling \( \ell_C \) for patient 1 with migration number \( \mu(T, \ell_C) = \mu = 13 \), designating ROv as the primary tumor (Fig. 6a). This labeling had comigration number \( \gamma(T, \ell_C) = 10 \) (Fig. 6b). MACHINA found two additional vertex labelings \( \ell_D \) and \( \ell_E \) with the same minimum migration number \( \mu(T) = 13 \) but with a smaller comigration number \( \gamma(T, \ell_D) = 7 \) (Fig. 6d,e).

In addition to different designations of the primary tumor, these different labelings produced migration graphs with different migration patterns (Fig. 6b–e). For example, the authors reported that the left ovary (LOv) is polyphyletic, i.e., composed of clones from distinct phylogenetic branches, which they reported to be indicative of polyclonal migration. Indeed, in our nomenclature, the migration graph corresponding to \( \ell_C \) had a polyclonal multisource seeding (pM) pattern, with multisource seeding of the LOv from multiple clones from the ROv and a single clone from the small bowel (SBwl). Moreover, the SBwl was both a destination of clones from the ROv primary site and a source of clones for various anatomical sites, including the LOv and several metastases (Fig. 6c).

Metastasis-to-metastasis migrations in prostate cancer. We applied MACHINA (in PHM-TR mode) to clone trees of ten metastatic prostate cancers that were reported in Gundem et al.11, MACHINA’s analysis supported the reported polyclonal migrations in these patients (Supplementary Note). However, the evidence for metastasis-to-metastasis migrations was not conclusive. For three of the eight patients for which Gundem et al.11 reported metastasis-to-metastasis migrations (patients A10, A31 and A32), MACHINA found alternative migration histories with parallel seeding of all metastases from the primary tumor that were also consistent with the data (Supplementary Note).
from Hoadley et al.10, who reported that metastases in both patients resulted from “multiclonal seeding instead of a single cell of origin”. For patient A7, Hoadley et al.10 sequenced m = 6 anatomical sites and identified ten clusters of somatic mutations, using SciClone42 (Fig. 7a). There is considerable uncertainty in the VAF for each mutation (Fig. 7b), and this uncertainty propagates through the construction of the clone tree and the clonal composition of each sequenced anatomical site. Ignoring this uncertainty, Hoadley et al.10 reported a clone tree with 22 extant clones (Fig. 7c). Using manual analyses, the authors described two different migration histories, both of which were recovered by MACHINA (Fig. 7d). The migration history with the smallest migration number (\( \mu = 6 \)) corresponded to a polyclonal multisource seeding (pM) pattern (\( \gamma = 6 \), in which the lung was seeded by clones from the rib and breast, and polyclonal migrations occurred from the lung to the brain and from the liver to the kidney (Fig. 7e).

We used MACHINA (in PMH-TI mode) to jointly infer the clone tree and migration history from confidence intervals on the VAFs that were derived from the SciClone clustering (Fig. 7f). We obtained a clone tree with only nine extant clones and a monoclonal single-source seeding (mS) migration pattern with migration number \( \mu_{\text{mS}} = 5 \) and comigration number \( \gamma_{\text{mS}} = 5 \) (Fig. 7f). This finding contradicts the reports of polyclonal migrations in patient A7 in Hoadley et al.10.

For comparison, we also ran Treeomics51 on the sequencing data of patient A7. Treeomics was unable to identify the two subclones that MACHINA detected in the liver and brain (Supplementary Fig. 2g). To demonstrate the advantages of MACHINA’s ability to resolve polytomies, we ran the minimum migration labeling method on the unresolved clone tree inferred by MACHINA. This method inferred a more complex monoclonal multisource seeding (mM) history with migration number \( \mu = 6 \) and comigration number \( \gamma = 6 \) (owing to two polytomies in the clone tree), which were resolved by MACHINA (Fig. 7f) but not by the minimum migration labeling method (Fig. 7g).

For patient A1 from Hoadley et al., MACHINA identified more parsimonious clone trees and migration histories than previously reported10 (Supplementary Note). Our results on these two patients show that ambiguities in the sequencing data and inaccuracies in the clone tree may lead to the inference of unnecessarily complex migration patterns. By accounting for uncertainty in bulk sequencing data and jointly inferring parsimonious clone trees and migration histories,
MACHINA is able to find simpler migration histories that explain the observed mutation data.

**Metastatic progression in melanoma.** We applied MACHINA (in PMH-TI mode) to eight patients with metastatic melanoma from Sanborn et al. MACHINA recapitulated the findings reported by Sanborn et al. and identified parsimonious migration histories in which multiple anatomical sites were seeded directly from the primary tumor. These results provide additional support for Sanborn et al.’s rejection of the commonly accepted serial progression model in melanoma, in which migration proceeds from primary tumor to regional metastases to distant metastases (Supplementary Note).

**Discussion**

The increasing availability of DNA and RNA sequencing data from matched samples of primary tumors and metastases provides opportunities to improve our understanding of the drivers of metastasis. Recent phylogenetic analyses have shown that the process of metastasis may be more complicated than monoclonal migration of individual cells between anatomical sites—with polyclonal migrations of cells, multisource seeding and reseeding between the primary tumor and metastases. Here we showed that deriving such conclusions about migration patterns without a precise quantitative model is a risky endeavor and can lead to statements about migration patterns that are not adequately supported by the data. In particular, the simplest migration model, monoclonal single-source seeding, should be definitively ruled out before invoking more complicated migration patterns to explain the data.

We introduced MACHINA, a multi-objective optimization algorithm that jointly infers the cell division, mutation and migration history from DNA sequencing data while simultaneously resolving uncertainty in bulk samples. MACHINA is based on a mathematical model that distinguishes the process of cell division from the process of cell migration. This model evaluates migration histories according to three criteria: the migration pattern, the migration number and the comigration number, the latter of which is motivated by experimental evidence describing clusters of tumor cells that simultaneously migrate and seed metastases. We used MACHINA to analyze sequencing data from metastatic ovarian, prostate, breast and skin cancers. Notably, in each case we found that multiple migration histories were consistent with the sequencing data and that, in many cases, these migration histories were simpler than those reported. These alternative migration histories contradict reports in some patients regarding metastasis-to-metastasis migrations, the anatomical site of a primary tumor or the occurrence of polyclonal migrations.

MACHINA fills a critical need in studies of metastasis, enabling researchers to rigorously assess the validity of different migration patterns in individual patients and to evaluate the prevalence of these patterns across large cohorts of patients and tumor types. However, we note several limitations of our analyses. First, although MACHINA relaxes the assumption of sample homogeneity, the subpopulations of cells (or clones) inferred by MACHINA are not homogeneous: complete homogeneity is obtained only at the level of individual cells. At the same time, we emphasize that single-cell

---

**Fig. 7 | Joint analysis of mutations and migrations shows a monoclonal single-source migration history for a patient with metastatic breast cancer.**

- **a**. VAFs for ten mutation clusters across six anatomical sites reported in breast cancer A7 from Hoadley et al.
- **b**. 95% confidence intervals for the VAFs of each cluster in the kidney of the patient in a; overlapping intervals indicate the presence of a single clone in kidney.
- **c**. In the reported mutation tree of breast cancer A7, each edge is labeled by a mutation cluster and each vertex corresponds to a clone comprising the mutation clusters on the unique path to the root.
- **d**. Migration number, comigration number and migration pattern for migration histories reported in Hoadley et al. for patient A7 (e), reported by MACHINA (f and Supplementary Note) and identified by the minimum migration labeling (g). The reported migration history has migration number \( \mu = 12 \), comigration number \( \gamma = 6 \) and a polyclonal multisource seeding (pM) pattern. Each leaf label is the proportion of the extant clone in the corresponding anatomical site; small or negative proportions are due to analysis of mutation clusters and not clones in the published work. Note that the clone tree has many polytomies. MACHINA inferred a monoclonal single-source seeding (mS) migration history, the simplest possible migration history, implying that the sequencing data did not strongly support complicated polyclonal migration patterns in this patient. Without MACHINA’s polytree resolution, a minimum migration labeling of the unresolved clone tree selected one of the possible labelings that minimized the migration number \( \mu \), leading to more complicated migration history with multisource seeding of the lung.
sequencing alone does not resolve migration histories—even with perfect cell trees, the anatomical site of ancestral cells remains unknown. The PMH and PHM-TR modes of MACHINA are directly applicable to cell trees derived from single-cell sequencing data. Second, although MACHINA allows for uncertainty in the frequencies of the mutation clusters, MACHINA does not account for uncertainty in the composition of the clusters themselves. Inferred migration patterns may be affected by such uncertainty. We evaluated how the number, purity and sequencing depth of samples affected the inference of migration patterns (Supplementary Note). Third, copy-number aberrations are an additional source of uncertainty, as they lead to a divergence between the fraction of cancer cells containing a mutation (often called the cancer cell fraction (CCF)) and the VAF of the mutation\(^{4,19}\). Although CCFs cannot be uniquely determined from sequencing data\(^{4}\), one might be able to jointly infer CCFs, mutation clusters, clone trees and migration histories by extending the parsimony objectives we have introduced here. Fourth, there is evidence that primary tumors from different tissues are biased in the anatomical sites of metastasis\(^{7}\). One can encode such differences as different weights in the current MACHINA algorithm (Supplementary Note).

There are additional future directions. Other types of data can be used to infer migration histories, including DNA methylation, circulating tumor DNA and circulating tumor cells. In addition, the computational model of migrations introduced here could be used to study spatial heterogeneity and cellular migrations during the growth of a tumor in a single anatomical site. Another possibility is to apply MACHINA to non-cancer data, for example, to analyze migrations of individuals and pathogens between geographically isolated populations, for which previous work has constrained migrations to particular topologies\(^{8,9}\). Finally, on the theoretical side, the computational complexity of the PMH, PHM-TR and PMH-TI problems is unknown.

High-throughput DNA sequencing has revolutionized studies of cancer evolution. The complexity, sublety and unique features of these data necessitate the use of robust and reproducible analysis approaches based on quantitative models. In particular, such models allow researchers to evaluate the evidence for simple explanations for the data before proposing complex evolutionary scenarios. Just as it is necessary to rigorously examine the evidence for neutral evolution in a tumor before one can reliably conclude that selection has occurred\(^{10}\), it is also necessary to rigorously evaluate the evidence for simple migration patterns in a metastatic cancer before concluding that complex migration patterns have occurred. In the coming years, the marriage of high-throughput genomics and epigenomics data with appropriate quantitative analysis will further elucidate the mysteries of metastasis.

URLs. MACHINA code repository, http://github.com/raphael-group/machina; Gurobi Optimizer, http://www.gurobi.com/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0106-z.

Received: 20 June 2017; Accepted: 9 March 2018; Published online: 26 April 2018

References

1. Gupta, G. P. & Massagué, J. Cancer metastasis: building a framework. Cell 127, 679–695 (2006).
2. Comen, E., Norton, L. & Massagué, J. Clinical implications of cancer self-seeding. Nat. Rev. Clin. Oncol. 8, 369–377 (2011).
3. Faries, M. B., Steen, S., Ye, X., Sim, M. & Morton, D. L. Late recurrence in melanoma: clinical implications of lost dormancy. J. Am. Coll. Surg. 217, 27–34 (2013).
4. Sanborn, J. Z. et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. Proc. Natl. Acad. Sci. USA 112, 10995–11000 (2015).
5. Tabasam, D. P. & PolyaK, K. Tumorigenesis: it takes a village. Nat. Rev. Cancer 15, 473–483 (2015).
6. Macintyre, G. et al. How subclinical modeling is changing the metastatic paradigm. Clin. Cancer Res. 23, 630–635 (2017).
7. Naxerova, K. & Jain, R. K. Using tumor phylogenetics to identify the roots of metastasis in humans. Nat. Rev. Clin. Oncol. 12, 258–272 (2015).
8. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. Science 352, 169–175 (2016).
9. Choi, Y. J. et al. Intra-individual genomic heterogeneity of high-grade serous carcinoma of the ovary and clinical utility of ascitic cancer cells for mutation profiling. J. Pathol. 241, 57–66 (2017).
10. Houssay, K. A. et al. Tumor evolution in two patients with basal-like breast cancer: a retrospective genomics study of multiple metastases. PLoS Med. 13, e1002174 (2016).
11. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. Nature 520, 353–357 (2015).
12. McPherson, A. et al. Divergent modes of clonal spread and intraepithelial mixing in high-grade serous ovarian cancer. Nat. Genet. 48, 758–767 (2016).
13. Brown, D. et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy-number aberrations. Nat. Commun. 8, 14944 (2017).
14. Aceto, N. et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell 158, 1110–1122 (2014).
15. Marrinucci, D. et al. Fluid biopsy in patients with metastatic prostate, pancreatic and breast cancers. Phys. Biol. 9, 016003 (2012).
16. Maddipati, R. & Stanger, B. Z. Pancreatic cancer metastases harbor evidence of polyclonality. Cancer Discov. 5, 1086–1097 (2015).
17. Yu, M. et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. Science 339, 580–584 (2013).
18. Cheung, K. J. et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin-14-expressing tumor cell clusters. Proc. Natl. Acad. Sci. USA 113, E854–E863 (2016).
19. Dadiani, M. et al. Real-time imaging of lymphogenic metastasis in orthotopic human breast cancer. Cancer Res. 66, 8037–8041 (2006).
20. Cheung, K. J. & Ewald, A. J. A collective route to metastasis: seeding by tumor cell clusters. Science 352, 167–169 (2016).
21. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multi-region sequencing. N. Engl. J. Med. 366, 883–892 (2012).
22. Gerlinger, M. et al. Genomic architecture and evolution of clear-cell renal cell carcinomas defined by multi-region sequencing. Nat. Genet. 46, 225–233 (2014).
23. Kim, T.-M. et al. Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. Clin. Cancer Res. 21, 4461–4472 (2015).
24. Zhao, Z.-M. et al. Early and multiple origins of metastatic lineages within primary tumors. Proc. Natl. Acad. Sci. USA 113, 2140–2145 (2016).
25. McCrery, M. Q. et al. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. Nat. Med. 21, 1514–1520 (2015).
26. Liu, B. et al. Spatiotemporal genomic heterogeneity, phylogeny and metastatic evolution in salivary adenoid cystic carcinoma. J. Natl. Cancer Inst. 109, 109 (2017).
27. Zhai, W. et al. The spatial organization of intratumor heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma. Nat. Commun. 8, 4565 (2017).
28. Gibson, W. J. et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. Nat. Genet. 48, 848–855 (2016).
29. Lote, H. et al. Carbon dating cancer: defining the chronology of metastatic progression in colorectal cancer. Ann. Oncol. 28, 1243–1249 (2017).
30. Thomsen, M. B. H. et al. Spatial and temporal clonal evolution during development of metastatic urothelial carcinoma. Mol. Oncol. 10, 1450–1460 (2016).
31. Tan, Q. et al. Genomic alteration during metastasis of lung adenocarcinoma. Cell. Physiol. Biochem. 38, 469–486 (2016).
32. Hosseini, H. et al. Early dissemination seeds metastasis in breast cancer. Nat. 540, 552–558 (2016).
33. Xue, R. et al. Variable intratumor genomic heterogeneity of multiple lesions in patients with hepatocellular carcinoma. Gastroenterology 150, 998–1008 (2016).
34. Campbell, P. J. et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467, 1109–1113 (2010).
35. Schwarz, R. F. et al. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. PLoS Med. 12, e1001789 (2015).
36. Beltran, H. et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat. Med.* **22**, 298–305 (2016).

37. De Mattos-Arruda, L. et al. Establishing the origin of metastatic deposits in the setting of multiple primary malignancies: the role of massively parallel sequencing. *Mol. Oncol.* **8**, 150–158 (2014).

38. Naxerova, K. et al. Origins of lymphatic and distant metastases in human colorectal cancer. *Science* **357**, 55–60 (2017).

39. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

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

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

42. Miller, C. A. et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10**, e1003665 (2014).

43. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).

44. Deshwar, A. G. et al. PhylotWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).

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

46. Popic, V. et al. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* **16**, 91 (2015).

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

48. Yuan, K., Sakoparnig, T., Markowitz, F. & Beerenwinkel, N. BitPhylogeny: a probabilistic framework for reconstructing intratumor phylogenies. *Genome Biol.* **16**, 36 (2015).

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

50. Dang, H. X. et al. ClonEvol: clonal ordering and visualization in cancer sequencing. *Annu. Oncol.* **28**, 3076–3082 (2017).

51. Reiter, J. G. et al. Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8**, 14114 (2017).

52. Vakiani, E. et al. Local recurrences at the anastomotic area are clonally related to the primary tumor in sporadic colorectal carcinoma. *Oncotarget* **8**, 42487–42494 (2017).

53. Makohon-Moore, A. P. et al. Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat. Genet.* **49**, 358–366 (2017).

54. Hong, W. S., Shpak, M. & Townsend, J. P. Inferring the origin of metastases from cancer phylogenies. *Cancer Res.* **75**, 4021–4025 (2015).

55. Sankoff, D. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **28**, 35–42 (1975).

56. Slatkin, M. & Maddison, W. P. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613 (1989).

57. Qiu, M., Hu, J., Yang, D., Cosgrove, D. P. & Xu, R. Pattern of distant metastases in colorectal cancer: a SEER-based study. *Oncotarget* **6**, 38658–38666 (2015).

58. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

59. Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A. & Holmes, E. C. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog.* **3**, 1220–1228 (2007).

60. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).

61. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
Methods
We developed an algorithm, 'metastatic and clonal history integrative analysis' (MACHINA), to solve three versions of the migration analysis problem. Each version is a constrained multi-objective optimization problem. In the following sections, we describe these three versions and provide validation and benchmarking results.

Parsimonious migration history (PMH). The first version of the migration analysis problem is the PMH problem. In this problem, we are given as input a clone tree T, which has been derived from some other data, such as bulk sequencing data or single-cell sequencing data. We are also given a set P of allowed migration patterns. Our goal is to find a labeling of the vertices of T by anatomical sites that first minimizes the migration number μ(T) and then minimizes the smallest comigration number γ(T).

Parsimonious migration history (PMH). Given a clone tree T and a set P of allowed migration patterns, find a vertex labeling ρ with the minimum migration number μ(T) and subsequently the smallest comigration number γ(T).

We consider three different sets P of allowed migration patterns: (i) P = {S}, which requires the migration graph G to be a single-source (S) migration pattern; (ii) P = {S,M}, which requires the migration graph to be either an S or an M pattern; (iii) P = {S,M,R}, which means that G is unrestricted. Note that because these cases are nested, the migration number decreases monotonically from case (i) to case (ii) to case (iii). In contrast, the most restrictive case P = {S} leads to the minimum comigration number γmin. Thus, by using different sets of allowed migration patterns, one can explore the tradeoff between the migration pattern, migration number and comigration number. For example, one could assess the evidence for reseeding (migration pattern R) by examining the difference in number of migrations reported when P = {S,M,R} versus when P = {S,M}, which restricts the migration graph to either an S or an M pattern.

Parsimonious migration history with tree resolution (PMH-TR). The second version of the migration problem, the PHM-TR problem, aims to infer a migration history while simultaneously resolving polyomies of a given clone tree T, where a polyomy is an internal vertex of T with more than two children. Because cells divide into exactly two daughter cells, such polyomies reflect uncertainty in the ancestral relationships of clones (Fig. i). The PHM-TR problem is useful for analyzing DNA sequencing data from both bulk tumors and single cells, as polyomies are common in both datasets; for example, most currently published clone trees derived from single-cell data are not fully resolved and contain polyomies.

Parsimonious migration history with tree resolution (PMH-TR). Given a clone tree T and a set P of allowed migration patterns, find a refinement T′ of T and vertex labeling ℓ of T′ with the minimum migration number μ(T′), and subsequently the smallest comigration number γ(T′).

The resolution of polyomies has been previously studied in species phylogenetics, and Maddison provides an exponential time algorithm for constructing a tree with the minimum number of polyomies under some constraints (Supplementary Fig. 3). A clone contains not only the mutations introduced on its incoming edge, but also all of the mutations of its ancestral clones. From the output of a clustering algorithm, we obtain confidence intervals (F, F') for the frequencies of each cluster per anatomical site. These mutation-frequency confidence intervals are the input to the PHM-TR problem.

Parsimonious migration history with tree inference (PMH-TI). The third version of the migration problem, the PHM-TI problem, jointly infers a clone tree and a migration history directly from bulk sequencing data. In bulk sequencing data, there is often substantial uncertainty in the clone tree due to the fact that the sequenced samples are generally not homogeneous but instead are mixtures of populations of cells, or clones, with different complements of somatic mutations. Analysis of these mixed samples with standard phylogenetic techniques, such as neighbor-joining, maximum parsimony or maximum likelihood, yields phylogenetic trees with high rates of homoplasy. Thus, many specialized deconvolution algorithms that model bulk samples as mixtures have been proposed for inferring clone trees. The computational problem that these approaches solve can be viewed as a constrained matrix factorization problem.

The input is a mutation frequency matrix F = (fij), where fij is the proportion of cells in anatomical site s that have mutation i. Given F, the goal is to find a non-negative matrix U = [uij] and a binary mutation matrix B = [bij] such that F = UB. Here, bij = 1, if and only if mutation i is present in clone j, otherwise bij = 0. Entry uij is the proportion of clone j in anatomical site s; clone j is present in an anatomical site s, if and only if uij > 0.

Most deconvolution methods assume the absence of homoplasy, i.e., they require mutations to only occur once in the clone tree and never to be lost. This no-homoplasy assumption (also known as the infinite sites assumption) has two important implications for the computational problem. First, given F and B, there is only one matrix U such that F = UB. Second, under this assumption there is a one-to-one correspondence between mutation matrices B and mutation trees T, which describes the relationships between mutations. In contrast to a clone tree, the leaves of a mutation tree are not labeled by anatomical sites. From T and U, we obtain a clone tree T by attaching a leaf to a vertex v in T and setting the label ℓ(v) = s for each entry ℓ(v) > 0, i.e., clone j is present in anatomical site s. Thus, to determine the presence of clones in anatomical sites, one must know both the mutation tree T (or equivalently matrix B) and U.

Deciding whether there exists a mutation tree T that respects the no-homoplasy assumption explaining F (i.e., whether there exists a mutation matrix B corresponding to T and a mixture matrix B with non-negative entries satisfying F = U×B) is nondeterministic polynomial time complete (NP-complete) . Moreover, when such a tree exists, the problem is typically underdetermined, i.e., there may exist many trees that explain F. As described previously, mutations trees that explain the observed mutation frequencies F are constrained spanning trees of a directed acyclic graph obtained from F that satisfy the condition (S), defined as

\[ f_{i,j} \geq \sum_{r \in \text{children}(v)} f_{r,j} \]

for each vertex v and anatomical site s. In practice, there is extensive uncertainty in the mutation frequencies in F, as these frequencies must be estimated from the proportion of DNA sequence reads that contain a mutation at a locus. One way to model this uncertainty is to define confidence intervals [f_{i,j}, f_{i,j}'] for each mutation i in sample s. Given confidence intervals F = (f_{i,j}, f_{i,j}') for the frequency of each mutation in each sample s, and a mutation matrix B = [b_{ij}], there may be many mixture matrices U = [u_{ij}] such that \[ \sum_{j} b_{ij} \leq \sum_{k} f_{i,k} \]. Thus, in addition to many mutation trees T explaining the observed data, each mutation tree T may correspond to multiple mixture matrices U. As such, the presence of clones in anatomical sites is no longer fully determined given a mutation tree T (or mutation matrix B). Moreover, many different mutation trees may explain the observed (F, F'). This leads to the following problem (Fig. 4b).

Parsimonious migration history with tree inference (PMH-TI). Given a set P of allowed migration patterns and mutation frequency confidence intervals (F = (f_{i,j}, f_{i,j}'), F = (f_{i,j}', f_{i,j}'')) of each clone tree T, and a vertex labeling ℓ of T such that: (i) \[ \ell_j \in [f_{i,j}', f_{i,j}''], \] (ii) F satisfies the sum condition for T; and (iii) vertex labeling ℓ of T has minimum migration number μ(T) and subsequently smallest comigration number γ(T).

Validation of MACHINA on simulated data. We assessed the performance of MACHINA on simulated metastatic tumors, which we generated by extending an existing agent-based simulation of tumor growth to include cell migration. In this simulation, tumor cells could migrate to an existing anatomical site or seed a new anatomical site following a user-defined migration pattern (mS, pS, PM and pR) and a number m of anatomical sites. For simplicity, we simulated only single-nucleotide mutations and excluded copy-number aberrations. Each simulation resulted in a clone tree T, a vertex labeling ℓ and migration graph G, which together described the clonal and migration history of the simulated tumor. We generated 40 simulated tumors with either m = 5 or m = 8 anatomical sites and varying migration patterns. For each simulated tumor, we generated DNA sequencing data from a single bulk sequencing sample from each metastasis and from single-cell sequencing samples from the primary tumor. Each sample had a purity of 1 and a target DNA sequence coverage of 200× corresponding to a typical whole-exome sequencing experiment.

We first benchmarked MACHINA (run in PMH-TI mode) against four other methods that construct phylogenetic trees from metastatic tumors: (i) neighbor-joining, a standard approach for computing phylogenetic trees that has been used frequently for analysis of recent sequencing datasets from samples of metastases (for example, refs 30,31,32,33); (ii) Treemomics, a recent phylogenetic reconstruction algorithm designed for analyzing samples of metastases;
Thus, choosing one of the minimum migration labelings will often not result in the migration patterns that were more complex than pS, such as pM and pR (Fig. 5b).

Vertex labelings with the same minimum number of migrations but varying correct migration pattern. For simulated polyclonal single-source seeding (pS) inferred by MACHINA closely resembles the simulated migration graph and to identify the clones that migrated and seeded anatomical sites. Figure 5c shows migration graph and thus differ considerably from MACHINA. Moreover, both neighbor-joining and Treeomics assume sample homogeneity, as they record each mutation as present (1) or absent (0) in each sample, by thresholding of VAFs. Specifically, Treeomics finds the most likely ‘error corrections’ required to transform the 0- to 1-valued mutation matrix into a perfect phylogeny matrix, which describes an evolutionary tree with no homoplasy (i.e., infinite sites assumption). Additionally, Treeomics provides an optional heuristic, denoted as Treeomics-sub, that detects subclones by resolving violations of the infinite sites assumption by using a variation of the ‘split row’ operation that was previously described.69,70 In contrast, PhyloWGS and AncesTree analyze VAFs and infer subclones within a sample. Further details of the simulations and the mutation-clustering algorithm in MACHINA are in the Supplementary Note.

We found that the clone trees identified by MACHINA better resembled the simulated clone trees than those inferred by neighbor-joining, Treeomics, PhyloWGS and AncesTree across all migration patterns. Figure 5a showed that the distance between the inferred clone tree T and the simulated clone tree $T'$ on simulations with $m = 8$ anatomical sites was substantially smaller for MACHINA. (Supplementary Fig. 4 shows similar results with $m = 5$ anatomical sites.) Here we used a modified Robinson–Foulds distance75 to compute the distance between trees. This demonstrated the deficiency of the sample-homogeneity assumption on heterogeneous data. Treeomics-sub did not assume sample homogeneity and performed better than Treeomics. However, Treeomics-sub could not match the performance of MACHINA and AncesTree, likely because the latter two methods used VAFs to deconvolve mixed samples and thus were better able to detect subclones. PhyloWGS achieved similar performance to Treeomics-sub but performed worse than AncesTree and MACHINA. Notably, MACHINA outperformed both PhyloWGS and AncesTree, thus showing that MACHINA’s advantage was not only a result of its analysis of subclones but also because of its simultaneous inference of clone trees and migration histories.

Next we examined whether the goal of minimizing migrations, as done in McPherson et al., using the Sankoff algorithm, was sufficient to determine the correct migration pattern. For simulated polyclonal single-source seeding (pS) migration patterns, we found that simulated clone trees typically had multiple vertex labelings with the same minimum number of migrations but varying migration patterns that were more complex than pS, such as pM and pR (Fig. 5b). The fraction of minimum migration labelings of the correct clone tree that were pS ranged from 0 to 1, with a median of 0.52 across the ten simulated instances. Thus, choosing one of the minimum migration labelings will often not result in the correct migration pattern. Moreover, one of the trials, indicated by the asterisk (*) in Fig. 5b, did not admit a minimum-migration vertex labeling with a pS pattern. However, by solving the PMH problem with a single-source seeding constraint, MACHINA determined that a vertex labeling with such a pS pattern required only one additional migration. These findings demonstrated the importance of more sophisticated scoring functions that accounted for the complexity of the resulting migration pattern. See Supplementary Note for corresponding results for the mS, pM and pR patterns.

Finally, we assessed MACHINAs ability to infer the correct migration graph and to identify the clones that migrated and seeded anatomical sites. Figure 5c shows migration graph $G$ inferred by MACHINA closely resembles the simulated migration graph $G'$ for the mS and pS simulations. The more complicated pM and pR patterns were more difficult to infer correctly, with MACHINA sometimes inferring simpler migration patterns than those obtained by the simulations (Supplementary Fig. 3). This may have been due to the resolution of the data, which did not allow us to detect the minor subclones involved in more complex seeding events. Finally, Fig. 5d shows that MACHINA was able to identify the mutations present in migrating clones with high precision and recall for all simulated migration patterns. Accurate identification of such mutations is an important prerequisite for further experimental validation of the mutations that drive metastatic progression. In summary, our simulations showed that MACHINA accurately infers clone trees and migration histories, outperforming existing methods.

In the Supplementary Note, we show that MACHINA continues to outperform existing clone-tree inference methods when given mutation clusters from different clustering algorithms. Moreover, we show that MACHINA benefits from having more samples and higher coverage sequencing, with the number of samples having the largest impact, followed by the depth of sequencing and the sample purity. Finally, we performed subsampling experiments by running MACHINA on subsets of mutations from simulated WGS data. We found that MACHINA performed well, given only a small subset (<5%) of mutations, demonstrating that MACHINA could accurately infer migration patterns given only whole-exome sequencing (WES) data.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. MACHINA is open source and available on GitHub (see URLs).

Data availability. The datasets analyzed during the current study are available from the database of Genotypes and Phenotypes (dbGaP) database under accessions phs000076 and phs000941.v1.p1 and from the European Genome-phenome Archive (EGA) under accessions EGAS00001000547, EGAS00001000262, EGAS00001000730 and EGAS00001000756.

References
62. Ross, E. M. & Markowitz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. 17, 69 (2016).
63. Jahn, K., Kuipers, J. & Beerwinkel, N. Tree inference for single-cell data. Genome Biol. 17, 86 (2016).
64. Zafar, H., Tien, A., Navin, N., Chen, K. & Nakleh, L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome Biol. 18, 178 (2017).
65. Leung, M. L. et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res. 27, 1287–1299 (2017).
66. Maddison, W. Reconstructing character evolution on polytomous cladograms. Cladistics 5, 365–377 (1989).
67. Pitch, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Biol. 20, 406–416 (1971).
68. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum-likelihood approach. J. Mol. Evol. 17, 368–376 (1981).
69. Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. Nucleic Acids Res. 41, e165 (2013).
70. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single-nucleotide somatic mutations. BMC Bioinformatics 15, 35 (2014).
71. Satas, G. & Raphael, B. J. Tumor phylogeny inference using tree-constrained importance sampling. Bioinformatics 33, i152–i160 (2017).
72. Zare, H. et al. Inferring clonal composition from multiple sections of a breast cancer. PLoS Comput. Biol. 10, e1003703 (2014).
73. Salehi, S. et al. ddClone: joint statistical inference of clonal populations from single-cell and bulk-tumor sequencing data. Genome Biol. 18, 44 (2017).
74. Reiter, J.G., Bozic, I., Chatterjee, K. & Nowak, M.A. in Computer-Aided Verification (eds. Sharygina, N. and Veith, H.) 101–106 (Springer, Berlin and Heidelberg, Germany, 2013).
75. Hajirasouliha, I. & Raphael, B.J. in Algorithms in Bioinformatics (eds. Brown, D. and Morgenstern, B.) 354–367 (Springer-Verlag, Berlin and Heidelberg, Germany, 2014).
76. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147 (1981).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size
   Describe how sample size was determined.
   No experimental data was generated for this study. Instead, we analyzed previously published datasets of metastatic ovarian cancer (McPherson et al., Nature Genetics, 2016), metastatic breast cancer (Hoadley et al., PLOS Medicine, 2016), metastatic prostate cancer (Gundem et al., Nature 2015) and metastatic melanoma (Sanborn et al., PNAS 2015).

2. Data exclusions
   Describe any data exclusions.
   No experimental data was generated for this study. Moreover, no data points were excluded from the previously datasets generated that we analyzed in this paper.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   No experimental data was generated for this study.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   No experimental data was generated for this study.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   No experimental data was generated for this study.

6. Statistical parameters
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   - [ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - [ ] A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - [ ] A statement indicating how many times each experiment was replicated
   - [ ] The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   - [ ] A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - [ ] The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
   - [ ] A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - [ ] Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study. We developed and used MACHINA to analyze the migration histories of the studied datasets. MACHINA is publicly available at: https://github.com/raphael-group/machina

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company. No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used. No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used. No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination. No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study. This study did not involve research animals and/or animal-derived materials.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants. This study did not involve human research participants.