Title: Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts

One Sentence Summary: Single-cell lineage tracing and RNA-seq capture diverse metastatic behaviors and drivers in lung cancer xenografts in mice.

Authors: Jeffrey J. Quinn*[1,2], Matthew G. Jones*[1,2,3,4,10], Ross A. Okimoto[5,6], Shigeki Nanjo[5,6], Michelle M. Chan[1,2], Nir Yosef†[7,8,9,10], Trever G. Bivona†[1,5,6], & Jonathan S. Weissman†[1,2,11,12]

Affiliations:

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA, USA.

[2]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA, USA.

[3]Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, CA, USA.

[4]Integrative Program in Quantitative Biology, University of California, San Francisco, San Francisco, CA, USA.

[5]UCSF Department of Medicine, University of California, San Francisco, San Francisco, CA, USA.

[6]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, USA.

[7]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA.

[8]Chan Zuckerberg Biohub Investigator, San Francisco, CA, USA.

[9]Ragon Institute of Massachusetts General Hospital, MIT and Harvard University, Cambridge, MA, USA.

[10]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA.

[11]Whitehead Institute, Cambridge, MA, USA.

[12]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.

*The authors contributed equally to this work.

†Corresponding authors. Email: weissman@wi.mit.edu (J.S.W.); Trever.Bivona@ucsf.edu (T.G.B.); niryosef@berkeley.edu (N.Y.).

Abstract:

Cancer progression is characterized by rare, transient events which are nonetheless highly consequential to disease etiology and mortality. Detailed cell phylogenies can recount the history and chronology of these critical events – including metastatic seeding. Here, we applied our Cas9-based lineage tracer to study the subclonal dynamics of metastasis in a lung cancer xenograft mouse model, revealing the underlying rates, routes, and drivers of metastasis. We report deeply resolved phylogenies for tens of thousands of metastatically disseminated cancer cells. We observe surprisingly diverse metastatic phenotypes, ranging from metastasis-incompetent to aggressive populations. These phenotypic distinctions result from pre-existing, heritable, and characteristic differences in gene expression, and we demonstrate that these differentially expressed genes can drive invasiveness. Furthermore, metastases transit via diverse, multidirectional tissue routes and seeding topologies. Our work demonstrates the power of tracing cancer progression at unprecedented resolution and scale.

Main Text:

Cancer progression is governed by evolutionary principles (reviewed in (*1*)), which leave clear phylogenetic signatures upon every step of this process (*2*, *3*), from early acquisition of oncogenic mutations (i.e., the relationships between normal and malignantly transformed cells (*4*)), to metastatic colonization of distant tissues (i.e., the relationship between a primary tumor and metastases (*5*)), and finally adaptation to therapeutic challenges (i.e., the relationship between sensitive and resistant clones (*6*)). Metastasis is a particularly critical step in cancer progression to study because it is chiefly responsible for cancer-related mortality (*7*). Yet because metastatic events are intrinsically rare, transient, and stochastic (*8*, *9*), they are typically impossible to monitor in real time. Analogous to the cell fate maps that have played an essential role in deepening our understanding of organismal development and cell type differentiation (*10*, *11*), accurately reconstructed phylogenetic trees of tumors and metastases can reveal key features of this process, such as the clonality, timing, frequency, origins, and destinations of metastatic seeding (*12*).

Lineage tracing techniques allow one to map the genealogy of related cells, providing a crucial tool for exploring the phylogenetic principles of biological processes like cancer progression and metastasis. Classical lineage tracing strategies can infer tumor ancestry from the pattern of shared sequence variations across tumor subpopulations (e.g., naturally occurring mutations, like single-nucleotide polymorphisms or copy-number variations) (*13*, *14*). These "retrospective" tracing approaches are particularly valuable for studying the subclonal dynamics of cancer in patient-derived samples, such as elucidating which mutations contribute to metastasis and when they occur (*15–18*). However, the resolution of these approaches is typically low because of the limited number of distinguishing natural mutations, and the conclusions can be confounded by incomplete or impure bulk tumor sampling (*19*), sequencing artifacts (*20*), varying levels of intratumor heterogeneity, and non-neutral mutations (*1*, *5*). Alternatively, so-called "prospective" lineage tracing approaches – wherein cells are marked with a static label (e.g. genetic barcode or fluorescent tag) – can measure gross population dynamics at *clonal* resolution (*21*), but cannot resolve important and fine *subclonal* features of cancer biology, like evolution and the rate, timing, and directionality of metastatic events.

The recent development of Cas9-enabled lineage tracing techniques with single-cell RNA-sequencing readouts (*22–26*) provides the potential to explore cancer progression at vastly larger scales and finer resolution than has been previously possible with classical prospective or retrospective tracing approaches. These new methods most commonly rely on similar technical principles (reviewed in (*27*, *28*)). Briefly, Cas9 targets and cuts a defined genomic locus (hereafter "Target Site"), resulting in a stable insertion/deletion (indel) "allele" that is inherited over subsequent generations; as the cells divide, they accrue more Cas9-induced indels at additional sites that further distinguish successive clades of cells (**Fig. 1A**, **Fig. S1**). At the end of the lineage tracing experiment, the indel alleles are collected from individual cells by sequencing and paired with single-cell expression profiles of the cell state (*22*, *23*). Then, as in retrospective tracing approaches, various computational approaches (*29–34*) can reconstruct a phylogenetic tree that best models subclonal cellular relationships (e.g., by maximum-parsimony) from the observed shared or distinguishing alleles. Thus far, Cas9-enabled tracing has been successfully applied to study important aspects of metazoan biology, like the cellular progenitor landscape in early mammalian embryogenesis (*23*, *35*), hematopoiesis (*36*), and neural development in zebrafish (*22*). Additionally, resources now exist for studying other phylogenetic processes in mouse (*23*, *35*), and analytical tools are available for computationally reconstructing and benchmarking trees from large lineage tracing datasets (*34*, *37*).

Here we apply lineage tracing to explore the subclonal dynamics of metastatic dissemination in an orthotopic xenograft model of lung cancer in mice (*38*). Specifically, we have modified our previously described "molecular recorder" for lineage tracing (*23*), now enabling the capture of highly detailed, single-cell-resolution phylogenies across tens of thousands of cells with continuous tracing *in vivo* over several months. Additionally, we have expanded on our analytical toolkit, Cassiopeia (*34*), with algorithms for inference of unobserved events from a phylogeny, which we applied here to resolve metastatic transitions between tissues. These and other advances allowed us to study the rates, transcriptional drivers, and tissue routes of metastasis at unprecedented scale and resolution.

Tracing metastasis in a mouse xenograft model

We chose to study metastasis using a human *KRAS*-mutant lung adenocarcinoma line (A549 cells) in an orthotopic xenograft model in mice because this system is characterized by aggressive metastases (*38*) and orthotopic xenografting experiments such as this are useful for modeling cancer progression *in vivo (39)*. We engineered A549 cells with a refined version of our molecular recorder technology (*23*) (**Fig. S2**; Methods). Specifically, the engineered cells contained: (i) luciferase for live imaging; (ii) Cas9 for generating heritable indels; (iii) ~10 uniquely barcoded copies of the Target Site for recording lineage information, which can be captured as expressed transcripts by single-cell RNA-sequencing; and finally (iv) triple-sgRNAs to direct Cas9 to the Target Sites, thereby initiating lineage recording (**Fig. 1A; Fig. S2A–C**). To enable tracing over months-long timescales, we carefully designed the sgRNAs with nucleotide mismatches to the Target Sites, thereby decreasing their affinity (*40*, *41*) and tuning the lineage recording rate (*23*, *42*). Approximately 5,000 engineered cells ("A549-LT") were then embedded in matrigel and surgically implanted into the left lung of an immunodeficient (C.B-17 *SCID*) mouse (**Fig. 1B**). We followed bulk tumor progression by live luciferase-based imaging (**Fig. 1C**): early bioluminescent signal was modest and restricted to the primary site (left lung), consistent with engraftment; with time, the signal progressively increased and spread throughout the thoracic cavity, indicating tumor growth and metastasis. After 54 days, the mouse was sacrificed and tumors were identified in the five lung lobes, throughout the mediastinal lymph tissue, and on the liver (**Fig. 1D**), in a pattern that is consistent with previous studies in this model (*38*). From these tumorous tissues, we collected six samples, including one from the left lung (i.e., including the primary site; **Fig. 1E, left**). The tumor samples were dissociated, fluorescence-sorted to exclude normal mouse cells, and finally processed for single-cell RNA-sequencing. To simultaneously measure the transcriptional states and phylogenetic relationships of the cells, we prepared separate RNA expression and Target Site amplicon libraries, respectively, resulting in 41,487 paired single-cell profiles from six tissue samples (**Fig. 1E, right**; **Fig. S3**; Methods).

In addition to the mouse described above (hereafter "M5k"), we also performed lineage tracing in three other mice (called "M10k", "M100k", and "M30k"), using A549-LT cells engineered with slightly different versions of the lineage tracing technology (**Fig. S4;** Methods). Unless otherwise noted, we focus our primary discussion of the results on mouse M5k because it yielded the richest lineage tracing dataset with the most cells and distinct lineages.

Distinguishing clonal cancer populations

Our lineage recorder "Target Site" (*23*) carries two orthogonal units of lineage information: (i) a static 14bp-randomer barcode ("intBC") that is unique and distinguishes between multiple integrated Target Site copies within each cell; and (ii) three independently evolving Cas9 cut-sites per Target Site that record heritable indel alleles and are used for subclonal tree reconstruction (**Fig. 1A**). Each Target Site is expressed from a constitutive promoter allowing it to be captured by single-cell RNA-sequencing. After amplifying and sequencing the Target Site mRNAs, the reads were analyzed using the Cassiopeia processing pipeline (*34*). Briefly, this pipeline leverages unique molecular identifier (UMI) information and redundancy in sequencing reads to confidently call intBCs and indel alleles from the lineage data, which inform subsequent phylogenetic reconstruction (**Fig. S1**; Methods).

We first determined the number of clonal populations (that is, groups of related cells that descended from a single clonogen at the beginning of the xenograft experiment), which are each associated with a set of intBCs. Importantly, the A549-LT cells were prepared at high diversity such that clones carry distinct intBC sets. By sampling the A549-LT cells before implantation, we estimate that the implanted pool of 5,000 cells initially contained 2,150 distinguishable clones (**Fig. S2D**). Based on their intBC sets, we assigned the vast majority of cancer cells collected from the mouse (97.7%) to 100 clonal populations (**Figs. S5A-B**), ranging in size from >11,000 (Clone #1, “CP001”) to ~30 cells (CP100) (**Fig. S5C**). Though there were some smaller clonal populations, we focused on these largest 100 because lineage tracing in few cells is less informative. Furthermore, despite initially implanting ~2,150 distinct clones, only ~100 clones successfully engrafted and proliferated; this indicates that only a small minority of cells may be competent for engraftment and survival *in vivo* (**Fig. S2D**). Moreover, we find no correlation between initial (pre-implantation) and final (post-sacrifice) clonal population size (Spearman’s ρ=-0.026; **Fig. S2E**), suggesting that clone-intrinsic characteristics that confer greater fitness *in vitro* do not necessarily confer greater fitness *in vivo* (*43*, *44*).

Features that influence the lineage recording capacity and tree reconstructability differed between clonal populations, such as the copy-number of Target Sites, the percentage of recording sites bearing indel alleles, and allele diversity (**Fig. S6A-C; Fig. S7**). Though most clonal populations exceeded parametric standards for confident phylogenetic reconstruction, some had slow recording kinetics or low allele diversity and failed to pass quality-control filters (17 clones, 7.3% of total cells in mouse M5k, **Fig. S6D; Fig. S7B**); these clones were excluded from tree reconstruction and downstream analyses (Methods).

We observed that the clonal populations exhibited distinct distributions across the six tissues (**Fig. S8A-C**), ranging from being present exclusively in the primary site (e.g., CP029, CP046), to overrepresented in a tissue (CP003, CP020), or distributed broadly over all sampled tissues (CP002, CP013). The level of tissue dispersal is a consequence of metastatic dissemination and thus can inform on the frequency of past metastatic events as follows: clonal populations that reside exclusively in the primary site likely never metastasized; those that did not broadly colonize tissues likely metastasized rarely; and those with more broad dispersal across all tissues likely metastasized more frequently. To quantify the relationship between tissue distribution and metastatic phenotype, we defined a statistical measure of the observed-versus-expected tissue distributions of cells (termed “Tissue Dispersion Score”; Methods) to operate as a coarse, tissue-resolved approximation of the metastatic rate. Across the 100 clonal populations in this mouse, we observed a wide range of Tissue Dispersion Scores (**Fig. S8D**), suggesting broad metastatic heterogeneity across the tumor populations. We next explored this suggested metastatic heterogeneity more directly and at far greater resolution using the evolving lineage information.

Single-cell-resolved cancer phylogenies

The key advantage of our lineage tracer is not in following *clonal* lineage dynamics (i.e. from cells’ static intBCs, as described in the section above) but rather in reconstructing *subclonal* lineage dynamics (i.e. from cells’ continuously evolving indel alleles). As such, we next reconstructed high-resolution phylogenetic trees using the Cassiopeia suite of phylogenetic inference algorithms (*34*) with modified parameters tailored to this dataset’s unprecedented complexity and scale (Methods). Each of the resulting trees comprehensively describes the phylogenetic relationships between all cells within the clonal population (**Fig. 2A**), thus summarizing their life histories from initial, clonogenic founding in the mouse to final dissemination across tissues and tumors. The trees are intricately complex (mean tree depth of 7.25; **Fig. S6E**) and highly resolved (consisting of 37,888 cells with 33,266 (87.8%) unique lineage states; **Fig. S6C**).

To illustrate the intricate complexity of the trees in this dataset, we present the reconstructed phylogram and lineage alleles for a representative clonal population of 5,616 cells (CP003; **Fig. 2B**) with 99.0% (5,560) unique cell lineage states, mean tree depth of 10.0, and maximum tree depth of 20. Intuitively, cells that are more closely related to one another ought to share more lineage alleles, which is evident from the patterns of shared alleles within clades and distinguishing alleles between clades (**Fig. 2B, zoomed inlays**). Indeed, we find systematic agreement between phylogenetic distance (i.e., the distance between two cells in the tree) and allelic distance (the difference between two cells' lineage alleles) for this example (**Fig. 2C**) and across all other trees (**Fig. S10**), thus supporting their accuracy. The high diversity of distinguishable Cas9-induced indels (9,936 unique alleles across all M5k cells; evident in the array of unique allele colors in **Fig. 2B**) also reduces the probability of homoplasy, an issue which complicates tree reconstruction and impairs tree accuracy (*34*, *45*). Altogether, these features indicate that the reconstructed trees accurately model the true phylogenetic relationships between cells.

Inferring and quantifying past metastatic events from phylogenies

A striking feature revealed by the reconstructed phylogenies is the varying extent to which closely related cells reside in different tissues (**Fig. 2A**), patterns which directly result from ancestor cells having physically transited from one tissue to another in the past (i.e., metastatic seeding). Varying rates of metastasis produce different patterns of concordance between phylogeny and tissue (**Fig. 3A**). For example, non-metastatic populations result in all clades remaining within a single tissue (**Fig. 3A-B, left**); conversely, highly metastatic populations result in closely related cells residing in different tissues (**Fig. 3A-B, right**). Finally, intermediate levels of metastasis can similarly lead to a dispersed tissue distribution as in the highly metastatic regime, though with fewer metastatic transitions, thus supporting the need to reconstruct trees in order to distinguish such cases (**Fig. 3A-B, middle**).

To quantitatively study the relationship between metastatic phenotype and phylogenetic topology, we used the Fitch-Hartigan maximum parsimony algorithm (*46*, *47*). Our implementation of this algorithm provides the minimal number of ancestral (i.e., not directly observed) metastatic transitions that are needed to explain the final (i.e., observed) tissue location of each cell in a given tree. We defined a score of the metastatic potential (termed "TreeMetRate") by dividing the inferred minimal number of metastatic transitions by the total number of possible transitions (i.e., edges in the tree). Empirically, we observe a distribution of clonal populations that spans the full spectrum of metastatic phenotypes between low (non-metastatic) and high (very metastatic) TreeMetRates (**Fig. 3B,C**). The TreeMetRate is stable across bootstrapping experiments in simulated trees (**Fig. S9E-F**) and when using an alternative phylogenetic reconstruction method (Neighbor-Joining (*29*)) on empirical data (**Fig. S11A**; Pearson's ρ=0.94), indicating that the TreeMetRate is a robust measurement of metastatic behavior – though, notably, Cassiopeia trees are more parsimonious than those reconstructed by Neighbor-Joining (**Fig. S11B**). Empirically, the Tissue Dispersal Score agrees with the TreeMetRate at low metastatic rates (**Fig. S12A,C**), however, the TreeMetRate more accurately captures the underlying metastatic rate over a broad range of simulated metastatic rates because it can distinguish between moderate and high metastatic rates (**Fig. S9D**), which both result in broad dispersion across tissues (**Fig. 3A**), whereas the Tissue Dispersal Score saturates at intermediate metastatic rates (**Fig. S9B**). Furthermore, the TreeMetRate also agrees with the probability that a cell's closest relative (by lineage allele similarity) resides in a different tissue for each clonal population (termed "AlleleMetRate"; **Fig. S12B, D**); importantly, the AlleleMetRate is an alternative metric of metastatic potential that exploits the evolving nature of our lineage tracer but is independent of tree reconstruction. Again, however, simulations indicate that the TreeMetRate is the superior measurement of the underlying metastatic rate (**Fig. S9A-D**), underscoring the value of the reconstructed phylogenies in helping identify aspects of metastatic behavior that would otherwise be invisible.

We further extended our parsimony-based approach to quantify the metastatic phenotype at the resolution of individual cells (termed the "scMetRate") by averaging the TreeMetRate for all subclades containing a given cell (Methods). This measurement is sensitive to *subclonal* differences in metastatic behavior (**Fig. 3C**), and highlighted intriguing bimodal metastatic behavior for clone CP007 (**Fig. 5G;** discussed below).

Additionally, we find that the scMetRate is uncorrelated to clonal population size, proliferation signatures (*48*, *49*), or cell cycle stage (*50*) (**Fig. S13**), indicating that it can measure metastatic potential uncoupled from proliferative capacity. Overall, these results indicate that cancer cells in this dataset exhibit diverse metastatic phenotypes both between and within clonal populations, which can be meaningfully distinguished and quantified by virtue of the lineage tracer, but would have otherwise been hidden from classical barcoding approaches.

Transcriptional drivers of differences in metastatic phenotype

A central question in cancer biology is the extent to which cellular properties (e.g., transcriptional state) underlie cancer phenomena (*51*), like metastatic capacity. By comparing the paired transcriptional and lineage datasets, we found that different metastatic behaviors corresponded to differential expression of genes, many with known roles in metastasis. First, after filtering and normalizing the scRNA-sequencing data, we applied *Vision* (*52*), a tool for assessing the extent to which the variation in cell-level quantitative phenotypes can be explained by transcriptome-wide variation in gene expression. While we found little transcriptional effect attributable to clonal population assignment, we found a modest association between a cell's transcriptional profile and its tissue sample or metastatic rate (**Fig. S14**). We next performed pairwise differential expression analyses comparing cells from completely non-metastatic clonal populations (i.e., four clones that never metastasized from the primary tissue in the left lung, like CP029) to metastatic clones in the same tissue (**Fig. S15**). This clone-resolution analysis identified several genes with significant expression changes which were also consistent across each non-metastatic clone (log2 fold-change > 1.5, FDR < 0.01), such as IFI6. These initial results suggested that differences in metastatic phenotype may manifest in characteristic differences in gene expression, and motivated deeper analysis.

Next, we sought to comprehensively identify genes that are associated with metastatic behavior by regressing single-cell gene expression against the scMetRates (over all observed cells, clonal populations, and tissues; **Fig. 4A**; Methods), thereby leveraging both the scRNA-seq dataset and the single-cell phylogenies.

Many of the identified positive metastasis-associated candidates (i.e., genes with significantly higher expression in highly metastatic cells) have known roles in potentiating tumorigenicity (**Fig. 4B, top**). For example, IFI27 is an interferon-induced factor that is anti-apoptotic and promotes epithelial-mesenchymal transition (EMT), cell migration, and cancer stemness in various carcinomas (*53*, *54*); REG4 enhances cell migration and invasion in colorectal carcinoma (*55*) and *KRAS*-driven lung adenocarcinoma (*56*); and TNNT1 has elevated expression in many cancers and may promote EMT and invasiveness (*57*). Similarly, many negative metastasis-associated candidates (i.e., genes with significantly lower expression in highly metastatic cells) have known roles in attenuating metastatic potential (**Fig. 4B, bottom**). For example, NFKBIA (IκBα) is a pan-cancer tumor suppressor via inhibition of pro-tumoral NFκB signaling (*58*); lower expression of ID3 enhances tumor cell migration and invasion *in vitro* and in lung adenocarcinoma xenograft models (*59*); and downregulation of ASS1 supports tumor metabolism and proliferation (*60*). (Interestingly, our most significant negative candidate was KRT17, which has previously been implicated in *promoting* invasiveness in lung adenocarcinoma (*61*) and its overexpression has been associated with poor prognosis in some cancers (*62*); we follow-up on this unexpected finding below.) Additionally, many of the identified genes were significantly reproduced across every mouse in this study (**Fig. 4C-D**; **Fig. S17**). And more generally, the gene-level expression trends are broadly supported by significant correlation between the TreeMetRate and several gene expression signatures (*63*) (**Fig. S16**), including interferon signaling programs (*64*), RAS pathways (*65*) (A549 cells are *KRAS*-mutant), cancer invasiveness (*66*), and EMT (*67*) (consistent with increased NFκB signaling (*68*, *69*)).

While we identified many interesting and reproducible gene candidates in our regression analysis above, it was unclear whether they were directly driving the metastatic phenotype or were merely associated with it. To address this important point, we next explored the functional impact on metastatic behavior of modulating the expression of five high-scoring gene candidates (IFI6, IFI27, KRT17, ID3, and ASS1). First, we engineered A549 cells to enable CRISPR-inhibition or -activation perturbations (CRISPRi/a; activity validated in **Fig. S18C,D)**, then increased or decreased expression, respectively, of the five gene targets using two independent sgRNAs per gene. Finally, we measured the perturbed cells' invasion phenotype *in vitro* using a transwell invasion assay (**Fig. 4E,F**; Methods). As hypothesized, CRISPRi knock-down resulted in *decreased* invasiveness for positive metastasis-associated genes (IFI6 and IFI27; p=0.001, 0.005, respectively) and *increased* invasiveness for negative metastasis-associated genes (KRT17, ID3, and ASS1; p=0.054, 0.003, and 0.062, respectively; **Fig. 4E**). Conversely, we found that elevating candidate gene expression by CRISPRa produced the exact opposite results (**Fig. 4F**), indicating that the invasion phenotype can be quantitatively altered by both increased or decreased expression for each of the five candidate genes tested, including notably KRT17. We confirmed that the modulation of expression of each of these genes strongly and significantly modulated invasiveness (p<0.01) in a separate human lung cancer cell line (H1299 cells, which are *KRAS* wild-type, *TP53*-mutant, and harbor endogenous *NRAS*^{Q61K}; **Fig. S18A,B**); though, for two of the genes (IFI27, IFI6), CRISPRa had a significant effect (p<0.01) while CRISPRi did not. Taken together, these results indicate that (i) the lineage tracer can meaningfully identify metastasis-associated genes *in vivo*, (ii) some of these gene candidates are sufficient to drive differences in metastatic phenotype, and (iii) these genes' roles in mediating invasiveness extend beyond the one A549 cancer model and across different oncogenic backgrounds.

Heterogeneity and heritability of metastatic behavior in pre-implantation cells

We next used the positive and negative metastasis-associated genes identified above (**Fig. 4A**) to define a *de novo* transcriptional signature (hereafter, "Metastasis Signature"; **Fig. 5A**; **Fig. S19A**). We found that even prior to implantation into the mice, the cells already exhibited meaningful heterogeneity in the Metastatic Signature (**Fig. 5B**), and metastasis-associated genes like ID3 and TNNT1 were similarly heterogeneously expressed pre-implantation (**Fig. 5C**). Next, we used the lineage barcodes to map cells from the *in vitro* pre-implantation pool to the clonal populations that engrafted *in vivo* (**Fig. S19B**). We then segregated these mapped cells into the top and bottom halves by their corresponding TreeMetRate, and queried their pre-implantation Metastatic Signatures. We found that cells from more metastatic clones in the mouse had modestly yet significantly higher metastatic signatures prior to implantation, and vice versa (**Fig. S19C**). This indicates that the pre-implantation transcriptional signature is mildly predictive of *in vivo* metastatic phenotype (**Fig. S19D**), though the distinction becomes more amplified *in vivo* (**Fig. S19C, D**). This result suggests that even before cells were xenografted into the mouse, they were primed for greater or lesser metastatic capacity *in vitro*.

While the pre-existing transcriptional heterogeneity in the pre-implantation cells was noteworthy, it remained unclear whether these differences were stochastic or intrinsic properties of the cells that could be robustly propagated *in vitro* and *in vivo*. One way to address this question is by implanting two cells from the same clone into two distinct mice and querying how well their metastatic phenotype is reproduced. Using the cells' intBCs, which statically mark clones, we identified two such instances where cells from the same clonal population seeded tumors in two different mice (**Fig. 5D; Fig. S20**). Strikingly, for each of the two pairs of clonal populations, the TreeMetRates were nearly identical (**Fig. 5E**). In fact, one of these pairs had the most similar TreeMetRates across all pairs of clones in the two mouse experiments (Δ(TreeMetRate)=0.0005, *p*=0.0049; **Fig. 5F**). Taken together, these results indicate that (i) the diverse metastatic phenotypes *in vivo* are determined before implantation (also **Fig. 5B, C**), (ii) the metastatic phenotype is reproducible over generations and is thus heritable (**Fig. 5E,F**; also **Fig. S19C,D**), and (iii) our analytical approaches for quantifying the metastatic rate, including reconstruction of the phylogenies, are experimentally robust (**Fig. 5E**).

Evolution of metastatic phenotype

Though we have thus far discussed how metastatic phenotype is clone-intrinsic and stably inherited, we identified a single example within the dataset that was the exception to this general rule. Specifically, Clone #7 (CP007) exhibited exceptionally distinct subclonal metastatic behaviors, wherein one clade metastasized frequently to other tissues while another clade remained predominantly in the right lung (**Fig. 5G**). This distinction is reflected in a bimodal distribution of scMetRates (**Fig. 3C**; **Fig. 5H**). To explore the relationship between subclonal structure and gene expression, we applied *Hotspot* (*70*) and identified two modules of correlated genes that exhibit heritable expression programs (**Fig. S21A**). Strikingly, the cumulative expression of genes in Module 1 is correlated with lower metastatic rates, while the opposite holds for Module 2 (**Fig. 5I; S21B,C**). Consistently, the two modules broadly correspond to the two clades with diverging metastatic phenotypes (**Fig. 5J**). This result is reproduced even in a control analysis of CP007 cells from the right lung only (**Fig. S21D-G**), indicating that these differences in gene expression indeed reflect differences in metastatic phenotype rather than tissue-specific effects. This example serves to illustrate that though the metastatic rate is stably inherited, it can also evolve – albeit rarely – within a clonal population, alongside concordant changes in transcriptional signature. Importantly, this finding could only be appreciated by virtue of the subclonal resolution of the lineage tracer.

Tissue routes and topologies of metastasis

The phylogenetic reconstructions also made it possible to describe detailed histories about the tissue routes and the directionality of metastatic seeding. For example, the phylogenetic tree for CP095 reveals five distinct metastatic events from the left lung to different tissues, in a paradigmatic example of simple primary seeding (**Fig. 6A–B**). Other phylogenies revealed more complicated trajectories, such as CP019, wherein early primary seeding to the mediastinum was likely followed by intra-mediastinal transitions and later seeding from the mediastinum to the liver and right lung (**Fig. 6D–E**). To more systematically characterize the tissue transition routes revealed by the phylogenetic trees, we extended the Fitch-Hartigan algorithm (*46*, *47*) to infer the directionality of each tissue transition (i.e., the origin and destination of each metastatic event) along a clonal population's ancestry. Our algorithm, called *FitchCount*, builds on other ancestral inference algorithms like MACHINA (*71*) by scaling to large inputs and providing tissue transitions frequencies that are aggregated across all ancestries that satisfy the maximum parsimony criterion (Methods; Supplemental Text). Through simulation we show that *FitchCount* can accurately recover underlying transition probabilities better than a naive application of the Fitch-Hartigan algorithm (**Fig. S9G–H**; Methods), likely because the naive approach summarizes only a single optimal assignment solution, whereas *FitchCount* summarizes all optimal solutions. The resulting conditional probabilities of metastasis to and from each tissue are summarized in a tissue transition probability matrix (**Fig. 6C, F**). Notably, we found that these transition matrices are varied and distinct to each clone (**Fig. G; Fig. S22**). We next used principal component analysis (PCA) to stratify clones by their transition matrices (**Fig. 6H**) and identified descriptive features that capture differences in the metastatic tissue routes traversed by each clone (**Fig. 6I; Fig. S23**). These descriptive features include primary seeding from the left lung (as in CP095, **Fig. 6A–C**), metastasis from and within the mediastinum (CP098, **Fig. 6G, left**), or metastasis between lung lobes (CP070, **Fig. 6G**), and may reflect intrinsic differences in tissue tropism. From this feature analysis we also note that many clones primarily metastasized via the mediastinal lymph tissue (**Fig. 6H–I**), suggesting that the mediastinum may act as a nexus for metastatic dissemination in this mouse model. This observation is consistent with previous experiments in this model (*38*), bulk live imaging during tumor progression in this experiment wherein tumors appear to quickly colonize the mediastinum (**Fig. 1C**), and the terminal disease state wherein the mediastinum harbors the majority of the tumor burden (**Fig. S8**).

Many models of metastatic seeding topology (i.e, the sequence and directionality of metastatic transitions) have been described in cancer (*1*), including reseeding, seeding cascades, parallel seeding, and others; and each is characterized by a distinct phylogenetic signature (**Fig. 6J**). These different metastatic topologies can critically influence the progression, relapse, and treatment of cancers (*9*, *72–74*); for example, reseeding of metastatic cells returning to the primary tumor site can contribute genetic diversity, resistance to treatment, and metastatic potential to tumors (*75*, *76*). Within this single dataset, we find numerous examples of all of these topologies (**Fig. 6K**); in fact, we most often observe examples of all topologies within every clone (**Fig. S24**), as well as more complex topologies that defy simple classifications (e.g., **Fig. 6D** and **G, right**), further underscoring the aggressive metastatic nature of A549 cells in this xenograft model.

Discussion

We applied our next-generation, Cas9-based lineage tracer to study metastasis in a lung cancer xenograft model in mice. We tracked metastatic spread at unprecedented scale and resolution that would have been unattainable by static barcoding experiments (i.e., "prospective" tracing), bulk tissue sampling, or single-cell RNA-sequencing methods alone, and were only made possible by experimental and algorithmic advances that we made on our molecular recorder platform (*23*). Experimentally, we increased lineage recorder information capacity and tuned the tracer dynamics for longer experimental timescales, thus allowing us to uniquely mark tens of thousands of cells descended from dozens of clonogens over several months. Analytically, enhancements to Cassiopeia, including *FitchCount*, allowed us to reconstruct accurate, informative, and deeply resolved phylogenetic trees, and interpret them to identify rare, transient events in cells' ancestry (here, metastasis). Beyond the utility of this experimental and analytical framework for exploring many facets of cancer biology, we believe this tracing approach is broadly applicable to study the phylogenetic foundations of many biological processes that transpire over multiple cell generations.

When we applied this tracing strategy to a lung cancer xenograft model, several key insights emerged: (1) Single-cell lineage tracing revealed meaningful and reproducible differences in the frequency and directionality of metastatic dissemination that would not have been discernible from bulk experiments (e.g., gross distribution of clones across tumorous tissues). For example, even among clonal populations that are broadly disseminated across tissues, the lineage tracer reveals substantive differences in the underlying metastatic rates. (2) Even within a single cancer line and xenograft model, we found surprisingly diverse metastatic phenotypes, ranging from metastasis-incompetent to highly metastatic, which we then used to generate and test hypotheses regarding metastatic biology. We show that these distinctions in metastatic behavior were driven by specific differentially expressed genes, like IFI6, ID3, and KRT17, that can directly modulate invasion phenotype in multiple cancer contexts. Furthermore, we empirically demonstrate that these distinct metastatic phenotypes are pre-programmed prior to implantation, stably inherited over generations, and predictive of future metastatic behavior. (3) Metastatic dissemination was rapid, frequent, and complex in this model, transiting via different complicated seeding topologies, such as seeding cascades, parallel seeding, and more. Additionally, we illustrated that it is possible to capture subtle differences in tissue tropism and, using this strategy, we identify the mediastinum as a hub for metastatic seeding, perhaps because the mediastinal lymph is a favorable niche with extensive tissue connections (*77*). Extending beyond this xenograft model, these findings suggest that metastatic seeding patterns can be highly complex and possibly patient-specific.

As a first report, this work by necessity focuses on a single model of metastasis. Looking forward, it will be valuable to explore these findings in other experimental contexts, including those described below, wherein the lineage tracer may also be deployed. First, because A549s exhibit highly aggressive metastatic spread, the rapid and frequent metastatic events we observe may be most relevant to advanced stages of cancer progression. Future work could apply this lineage tracing approach to models of any stage in cancer progression, such as (i) other cell lines that represent earlier cancer stages, (ii) genetic models of inducible tumor initiation (*78*), or (iii) patient-derived xenograft (PDX) models (*79*, *80*). Second, as is typical in xenograft experiments, this model requires immunodeficient host mice, and therefore does not reflect the pervasive influence wielded by the immune system on natural cancer progression (*81*–*83*); lineage tracing in syngeneic lines or autochthonous models of cancer could chart how an intact immune system influences cancer progression. Third, future work could build on our study by investigating the roles that other metastasis-associated gene candidates identified here possibly play in metastasis, or elucidate the molecular mechanism by which KRT17 diminishes metastatic phenotype *in vitro* and *in vivo*. Fourth, this work describes metastasis at the spatial resolution of tissues (i.e., not individual tumors) because we bulk tissues that potentially contain multiple tumors (including extensive micrometastases). An important direction would be to merge lineage data with high-resolution spatial information using the rapidly advancing techniques for spatial single-cell approaches (*24*, *84*–*86*); this would clarify, among other features, the clonality of micrometastases, monophyletic versus polyphyletic dissemination (*12*), and the spatial constraints of tumor growth and metastasis.

More broadly, Cas9-enabled lineage tracing technologies should be readily deployable to explore aspects of cancer progression and evolution, especially the history and chronology of rare and transient events like metastasis. This will empower future work to more comprehensively describe cancer – as well as other biological phenomena – at unprecedented resolution and scale.

References:

1. S. Turajlic, C. Swanton, Metastasis as an evolutionary process. *Science*. **352**, 169–175 (2016).

2. N. E. Navin, J. Hicks, Tracing the tumor lineage. *Mol. Oncol.* **4**, 267–283 (2010).

3. P. C. Nowell, The clonal evolution of tumor cell populations. *Science*. **194**, 23–28 (1976).

4. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature*. **458**, 719–724 (2009).

5. A. R. Brannon, E. Vakiani, B. E. Sylvester, S. N. Scott, G. McDermott, R. H. Shah, K. Kania, A. Viale, D. M. Oschwald, V. Vacic, A.-K. Emde, A. Cercek, R. Yaeger, N. E. Kemeny, L. B. Saltz, J. Shia, M. I. D'Angelica, M. R. Weiser, D. B. Solit, M. F. Berger, Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol.* **15**, 454 (2014).

6. H.-E. C. Bhang, D. A. Ruddy, V. Krishnamurthy Radhakrishna, J. X. Caushi, R. Zhao, M. M. Hims, A. P. Singh, I. Kao, D. Rakiec, P. Shaw, M. Balak, A. Raza, E. Ackley, N. Keen, M. R. Schlabach, M. Palmer, R. J. Leary, D. Y. Chiang, W. R. Sellers, F. Michor, V. G. Cooke, J. M. Korn, F. Stegmeier, Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat. Med.* **21**, 440–448 (2015).

7. C. L. Chaffer, R. A. Weinberg, A perspective on cancer cell metastasis. *Science*. **331**, 1559–1564 (2011).

8. A. W. Lambert, D. R. Pattabiraman, R. A. Weinberg, Emerging Biological Principles of Metastasis. *Cell*. **168**, 670–691 (2017).

9. J. Massagué, A. C. Obenauf, Metastatic colonization by circulating tumour cells. *Nature*. **529**, 298–306 (2016).

10. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).

11. X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang, Y. Xu, W. Huang, M. Jiang, X. Jiang, J. Mao, Y. Chen, C. Lu, J. Xie, Q. Fang, Y. Wang, R. Yue, T. Li, H. Huang, S. H. Orkin, G.-C. Yuan, M. Chen, G. Guo, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. **173**, 1307 (2018).

12. N. J. Birkbak, N. McGranahan, Cancer Genome Evolutionary Trajectories in Metastasis. *Cancer Cell*. **37**, 8–19 (2020).

13. S.-H. S. Wu, J.-H. Lee, B.-K. Koo, Lineage Tracing: Computational Reconstruction Goes Beyond the Limit of Imaging. *Mol. Cells*. **42**, 104–112 (2019).

14. R. Schwartz, A. A. Schäffer, The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**, 213–229 (2017).

15. M. Jamal-Hanjani, A. Hackshaw, Y. Ngai, J. Shaw, C. Dive, S. Quezada, G. Middleton, E. de Bruin, J. Le Quesne, S. Shafi, M. Falzon, S. Horswell, F. Blackhall, I. Khan, S. Janes, M. Nicolson, D. Lawrence, M. Forster, D. Fennell, S.-M. Lee, J. Lester, K. Kerr, S. Muller, N. Iles, S. Smith, N. Murugaesu, R. Mitter, M. Salm, A. Stuart, N. Matthews, H. Adams, T. Ahmad, R. Attanoos, J. Bennett, N. J. Birkbak, R. Booton, G.

Brady, K. Buchan, A. Capitano, M. Chetty, M. Cobbold, P. Crosbie, H. Davies, A. Denison, M. Djearman, J. Goldman, T. Haswell, L. Joseph, M. Kornaszewska, M. Krebs, G. Langman, M. MacKenzie, J. Millar, B. Morgan, B. Naidu, D. Nonaka, K. Peggs, C. Pritchard, H. Remmen, A. Rowan, R. Shah, E. Smith, Y. Summers, M. Taylor, S. Veeriah, D. Waller, B. Wilcox, M. Wilcox, I. Woolhouse, N. McGranahan, C. Swanton, Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol.* **12**, e1001906 (2014).

16. Z. Hu, J. Ding, Z. Ma, R. Sun, J. A. Seoane, J. Scott Shaffer, C. J. Suarez, A. S. Berghoff, C. Cremolini, A. Falcone, F. Loupakis, P. Birner, M. Preusser, H.-J. Lenz, C. Curtis, Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* **51**, 1113–1122 (2019).

17. M. Gerlinger, A. J. Rowan, S. Horswell, M. Math, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, C. Swanton, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

18. D. J. H. Shih, N. Nayyar, I. Bihun, I. Dagogo-Jack, C. M. Gill, E. Aquilanti, M. Bertalan, A. Kaplan, M. R. D’Andrea, U. Chukwueke, F. M. Ippen, C. Alvarez-Breckenridge, N. D. Camarda, M. Lastrapes, D. McCabe, B. Kuter, B. Kaufman, M. R. Strickland, J. C. Martinez-Gutierrez, D. Nagabhushan, M. De Sauvage, M. D. White, B. A. Castro, K. Hoang, A. Kaneb, E. D. Batchelor, S. H. Paek, S. H. Park, M. Martinez-Lage, A. S. Berghoff, P. Merrill, E. R. Gerstner, T. T. Batchelor, M. P. Frosch, R. P. Frazier, D. R. Borger, A. J. Iafrate, B. E. Johnson, S. Santagata, M. Preusser, D. P. Cahill, S. L. Carter, P. K. Brastianos, Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. *Nat. Genet.* (2020), doi:10.1038/s41588-020-0592-7.

19. W. S. Hong, M. Shpak, J. P. Townsend, Inferring the Origin of Metastases from Cancer Phylogenies. *Cancer Res.* **75**, 4021–4025 (2015).

20. J. G. Reiter, A. P. Makohon-Moore, J. M. Gerold, I. Bozic, K. Chatterjee, C. A. Iacobuzio-Donahue, B. Vogelstein, M. A. Nowak, Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8**, 14114 (2017).

21. M. B. Woodworth, K. M. Girskis, C. A. Walsh, Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).

22. B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, A. F. Schier, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

23. M. M. Chan, Z. D. Smith, S. Grosswendt, H. Kretzmer, T. M. Norman, B. Adamson, M. Jost, J. J. Quinn, D. Yang, M. G. Jones, A. Khodaverdian, N. Yosef, A. Meissner, J. S. Weissman, Molecular recording of mammalian embryogenesis. *Nature*. **570**, 77–82 (2019).

24. K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, L. Cai, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

25. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. van Oudenaarden, Whole-organism clone

tracing using single-cell sequencing. *Nature*. **556**, 108–112 (2018).

26. B. Spanjaard, B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov, J. P. Junker, Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).

27. C. S. Baron, A. van Oudenaarden, Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.* **20**, 753–765 (2019).

28. D. E. Wagner, A. M. Klein, Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* (2020), doi:10.1038/s41576-020-0223-2.

29. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

30. J. H. Camin, R. R. Sokal, A Method for Deducing Branching Sequences in Phylogeny. *Evolution*. **19**, 311 (1965).

31. H. Zafar, C. Lin, Z. Bar-Joseph, Single-cell Lineage Tracing by Integrating CRISPR-Cas9 Mutations with Transcriptomic Data. *bioRxiv* (2019), p. 16.

32. K. Sugino, T. Lee, Robust Reconstruction of CRISPR and Tumor Lineage Using Depth Metrics. *bioRxiv* (2019).

33. J. Feng, W. S. DeWitt III, A. McKenna, N. Simon, A. Willis, F. A. Matsen IV, Estimation of cell lineage trees by maximum-likelihood phylogenetics. *bioRxiv* (2019), p. 14.

34. M. G. Jones, A. Khodaverdian, J. J. Quinn, M. M. Chan, J. A. Hussmann, R. Wang, C. Xu, J. S. Weissman, N. Yosef, Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biol.* **21**, 64 (2020).

35. R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, G. M. Church, Developmental barcoding of whole mouse via homing CRISPR. *Science*. **361** (2018), doi:10.1126/science.aat9804.

36. S. Bowling, D. Sritharan, F. G. Osorio, M. Nguyen, P. Cheung, A. Rodriguez-Fraticelli, S. Patel, W.-C. Yuan, Y. Fujiwara, B. E. Li, S. H. Orkin, S. Hormoz, F. D. Camargo, An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell*. **181**, 1693–1694 (2020).

37. A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, J. Shendure, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353**, aaf7907 (2016).

38. R. A. Okimoto, F. Breitenbuecher, V. R. Olivas, W. Wu, B. Gini, M. Hofree, S. Asthana, G. Hrustanovic, J. Flanagan, A. Tulpule, C. M. Blakely, H. J. Haringsma, A. D. Simmons, K. Gowen, J. Suh, V. A. Miller, S. Ali, M. Schuler, T. G. Bivona, Inactivation of Capicua drives cancer metastasis. *Nat. Genet.* **49**, 87–96 (2017).

39. G. Francia, W. Cruz-Munoz, S. Man, P. Xu, R. S. Kerbel, Mouse models of advanced spontaneous metastasis for experimental therapeutics. *Nat. Rev. Cancer*. **11**, 135–141 (2011).

40. E. A. Boyle, J. O. L. Andreasson, L. M. Chircus, S. H. Sternberg, M. J. Wu, C. K. Guegler, J. A. Doudna,

W. J. Greenleaf, High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5461–5466 (2017).

41. S. K. Jones Jr, J. A. Hawkins, N. V. Johnson, C. Jung, K. Hu, J. R. Rybarski, J. S. Chen, J. A. Doudna, W. H. Press, I. J. Finkelstein, Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *bioRxiv* (2019), p. 214.

42. M. Jost, D. A. Santos, R. A. Saunders, M. A. Horlbeck, J. S. Hawkins, S. M. Scaria, T. M. Norman, J. A. Hussmann, C. R. Liem, C. A. Gross, J. S. Weissman, Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nat. Biotechnol.* (2020), doi:10.1038/s41587-019-0387-5.

43. N. McGranahan, C. Swanton, Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. **168**, 613–628 (2017).

44. A. N. Hata, M. J. Niederst, H. L. Archibald, M. Gomez-Caraballo, F. M. Siddiqui, H. E. Mulvey, Y. E. Maruvka, F. Ji, H.-E. C. Bhang, V. Krishnamurthy Radhakrishna, G. Siravegna, H. Hu, S. Raoof, E. Lockerman, A. Kalsy, D. Lee, C. L. Keating, D. A. Ruddy, L. J. Damon, A. S. Crystal, C. Costa, Z. Piotrowska, A. Bardelli, A. J. Iafrate, R. I. Sadreyev, F. Stegmeier, G. Getz, L. V. Sequist, A. C. Faber, J. A. Engelman, Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.* **22**, 262–269 (2016).

45. I. Salvador-Martínez, M. Grillo, M. Averof, M. J. Telford, Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *Elife*. **8** (2019), doi:10.7554/eLife.40292.

46. W. M. Fitch, Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool.* **20**, 406 (1971).

47. J. A. Hartigan, Minimum Mutation Fits to a Given Tree. *Biometrics*. **29** (1973), p. 53.

48. I. Ben-Porath, M. W. Thomson, V. J. Carey, R. Ge, G. W. Bell, A. Regev, R. A. Weinberg, An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).

49. C. Rosty, M. Sheffer, D. Tsafrir, N. Stransky, I. Tsafrir, M. Peter, P. de Crémoux, A. de La Rochefordière, R. Salmon, T. Dorval, J. P. Thiery, J. Couturier, F. Radvanyi, E. Domany, X. Sastre-Garau, Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma. *Oncogene*. **24**, 7094–7104 (2005).

50. M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, D. Botstein, Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*. **13**, 1977–2000 (2002).

51. T. Celià-Terrassa, Y. Kang, Distinctive properties of metastasis-initiating cells. *Genes Dev.* **30**, 892–908 (2016).

52. D. DeTomaso, M. G. Jones, M. Subramaniam, T. Ashuach, C. J. Ye, N. Yosef, Functional interpretation of single cell similarity maps. *Nat. Commun.* **10**, 4376 (2019).

53. H. Wang, X. Qiu, S. Lin, X. Chen, T. Wang, T. Liao, Knockdown of IFI27 inhibits cell proliferation and invasion in oral squamous cell carcinoma. *World J. Surg. Oncol.* **16**, 64 (2018).

54. S. Li, Y. Xie, W. Zhang, J. Gao, M. Wang, G. Zheng, X. Yin, H. Xia, X. Tao, Interferon alpha-inducible protein 27 promotes epithelial-mesenchymal transition and induces ovarian tumorigenicity and stemness. *J. Surg. Res.* **193**, 255–264 (2015).

55. Y. Guo, J. Xu, N. Li, F. Gao, P. Huang, RegIV potentiates colorectal carcinoma cell migration and invasion via its CRD domain. *Cancer Genet. Cytogenet.* **199**, 38–44 (2010).

56. S. Sun, Z. Hu, S. Huang, X. Ye, J. Wang, J. Chang, X. Wu, Q. Wang, L. Zhang, X. Hu, H. Yu, REG4 is an indicator for KRAS mutant lung adenocarcinoma with TTF-1 low expression. *J. Cancer Res. Clin. Oncol.* **145**, 2273–2283 (2019).

57. Y.-H. Hao, S.-Y. Yu, R.-S. Tu, Y.-Q. Cai, TNNT1, a prognostic indicator in colon adenocarcinoma, regulates cell behaviors and mediates EMT process. *Biosci. Biotechnol. Biochem.* **84**, 111–117 (2020).

58. M. Bredel, D. M. Scholtens, A. K. Yadav, A. A. Alvarez, J. J. Renfrow, J. P. Chandler, I. L. Y. Yu, M. S. Carro, F. Dai, M. J. Tagge, R. Ferrarese, C. Bredel, H. S. Phillips, P. J. Lukac, P. A. Robe, A. Weyerbrock, H. Vogel, S. Dubner, B. Mobley, X. He, A. C. Scheck, B. I. Sikic, K. D. Aldape, A. Chakravarti, G. R. Harsh 4th, NFKBIA deletion in glioblastomas. *N. Engl. J. Med.* **364**, 627–637 (2011).

59. F.-F. Chen, Y. Liu, F. Wang, X.-J. Pang, C.-D. Zhu, M. Xu, W. Yu, X.-J. Li, Effects of upregulation of Id3 in human lung adenocarcinoma cells on proliferation, apoptosis, mobility and tumorigenicity. *Cancer Gene Ther.* **22**, 431–437 (2015).

60. S. Rabinovich, L. Adler, K. Yizhak, A. Sarver, A. Silberman, S. Agron, N. Stettner, Q. Sun, A. Brandis, D. Helbling, S. Korman, S. Itzkovitz, D. Dimmock, I. Ulitsky, S. C. Nagamani, E. Ruppin, A. Erez, Diversion of aspartate in ASS1-deficient tumours fosters de novo pyrimidine synthesis. *Nature*. **527**, 379–383 (2015).

61. J. Liu, L. Liu, L. Cao, Q. Wen, Keratin 17 Promotes Lung Adenocarcinoma Progression by Enhancing Cell Proliferation and Invasion. *Med. Sci. Monit.* **24**, 4782–4790 (2018).

62. R. P. Hobbs, A. S. Batazzi, M. C. Han, P. A. Coulombe, Loss of Keratin 17 induces tissue-specific cytokine polarization and cellular differentiation in HPV16-driven cervical tumorigenesis in vivo. *Oncogene*. **35**, 5653–5662 (2016).

63. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

64. U. Einav, Y. Tabach, G. Getz, A. Yitzhaky, U. Ozbek, N. Amariglio, S. Izraeli, G. Rechavi, E. Domany, Gene expression analysis reveals a strong signature of an interferon-induced pathway in childhood lymphoblastic leukemia as well as in breast and ovarian cancer. *Oncogene*. **24**, 6367–6375 (2005).

65. C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, K. H. Buetow, PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–9 (2009).

66. D. Anastassiou, V. Rumjantseva, W. Cheng, J. Huang, P. D. Canoll, D. J. Yamashiro, J. J. Kandel, Human cancer cells express Slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. *BMC Cancer*. **11**, 529 (2011).

67. M. Jechlinger, S. Grunert, I. H. Tamir, E. Janda, S. Lüdemann, T. Waerner, P. Seither, A. Weith, H. Beug,

N. Kraut, Expression profiling of epithelial plasticity in tumor progression. *Oncogene*. **22**, 7155–7169 (2003).

68. M. A. Huber, N. Azoitei, B. Baumann, S. Grünert, A. Sommer, H. Pehamberger, N. Kraut, H. Beug, T. Wirth, NF-kappaB is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression. *J. Clin. Invest.* **114**, 569–581 (2004).

69. Y. Zhang, R. A. Weinberg, Epithelial-to-mesenchymal transition in cancer: complexity and opportunities. *Front. Med.* **12**, 361–373 (2018).

70. D. DeTomaso, N. Yosef, Identifying Informative Gene Modules Across Modalities of Single Cell Genomics. *bioRxiv* (2020), p. 54.

71. M. El-Kebir, G. Satas, B. J. Raphael, Inferring parsimonious migration histories for metastatic cancers. *Nat. Genet.* **50**, 718–726 (2018).

72. R. R. Langley, I. J. Fidler, The seed and soil hypothesis revisited--the role of tumor-stroma interactions in metastasis to different organs. *Int. J. Cancer*. **128**, 2527–2535 (2011).

73. I. J. Fidler, M. L. Kripke, The challenge of targeting metastasis. *Cancer Metastasis Rev.* **34**, 635–641 (2015).

74. T. Oskarsson, E. Batlle, J. Massagué, Metastatic stem cells: sources, niches, and vital pathways. *Cell Stem Cell*. **14**, 306–321 (2014).

75. A. Heyde, J. G. Reiter, K. Naxerova, M. A. Nowak, Consecutive seeding and transfer of genetic diversity in metastasis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14129–14137 (2019).

76. E. Comen, L. Norton, Self-seeding in cancer. *Recent Results Cancer Res.* **195**, 13–23 (2012).

77. E. R. Pereira, D. Kedrin, G. Seano, O. Gautier, E. F. J. Meijer, D. Jones, S.-M. Chin, S. Kitahara, E. M. Bouta, J. Chang, E. Beech, H.-S. Jeong, M. C. Carroll, A. G. Taghian, T. P. Padera, Lymph node metastases can invade local blood vessels, exit the node, and colonize distant organs in mice. *Science*. **359**, 1403–1407 (2018).

78. M. DuPage, T. Jacks, Genetically engineered mouse models of cancer reveal new insights about the antitumor immune response. *Curr. Opin. Immunol.* **25**, 192–199 (2013).

79. X. Zhang, S. Claerhout, A. Prat, L. E. Dobrolecki, I. Petrovic, Q. Lai, M. D. Landis, L. Wiechmann, R. Schiff, M. Giuliano, H. Wong, S. W. Fuqua, A. Contreras, C. Gutierrez, J. Huang, S. Mao, A. C. Pavlick, A. M. Froehlich, M.-F. Wu, A. Tsimelzon, S. G. Hilsenbeck, E. S. Chen, P. Zuloaga, C. A. Shaw, M. F. Rimawi, C. M. Perou, G. B. Mills, J. C. Chang, M. T. Lewis, A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.* **73**, 4885–4897 (2013).

80. M. Hidalgo, F. Amant, A. V. Biankin, E. Budinská, A. T. Byrne, C. Caldas, R. B. Clarke, S. de Jong, J. Jonkers, G. M. Mælandsmo, S. Roman-Roman, J. Seoane, L. Trusolino, A. Villanueva, Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.* **4**, 998–1013 (2014).

81. H. Gonzalez, C. Hagerling, Z. Werb, Roles of the immune system in cancer: from tumor initiation to

metastatic progression. *Genes Dev.* **32**, 1267–1284 (2018).

82. M. Angelova, B. Mlecnik, A. Vasaturo, G. Bindea, T. Fredriksen, L. Lafontaine, B. Buttard, E. Morgand, D. Bruni, A. Jouret-Mourin, C. Hubert, A. Kartheuser, Y. Humblet, M. Ceccarelli, N. Syed, F. M. Marincola, D. Bedognetti, M. Van den Eynde, J. Galon, Evolution of Metastases in Space and Time under Immune Selection. *Cell*. **175**, 751–765.e16 (2018).

83. M. Binnewies, E. W. Roberts, K. Kersten, V. Chan, D. F. Fearon, M. Merad, L. M. Coussens, D. I. Gabrilovich, S. Ostrand-Rosenberg, C. C. Hedrick, R. H. Vonderheide, M. J. Pittet, R. K. Jain, W. Zou, T. K. Howcroft, E. C. Woodhouse, R. A. Weinberg, M. F. Krummel, Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).

84. S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, E. Z. Macosko, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. **363**, 1463–1467 (2019).

85. P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, J. Frisén, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. **353**, 78–82 (2016).

86. A. Askary, L. Sanchez-Guardado, J. M. Linton, D. M. Chadly, M. W. Budde, L. Cai, C. Lois, M. B. Elowitz, In situ readout of DNA barcodes and single base edits facilitated by in vitro transcription. *Nat. Biotechnol.* **38**, 66–75 (2020).

87. M. Wang, Y. Zhao, B. Zhang, Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* **5**, 16923 (2015).

88. M. Jost, Y. Chen, L. A. Gilbert, M. A. Horlbeck, L. Krenning, G. Menchon, A. Rai, M. Y. Cho, J. J. Stern, A. E. Prota, M. Kampmann, A. Akhmanova, M. O. Steinmetz, M. E. Tanenbaum, J. S. Weissman, Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent. *Mol. Cell*. **68**, 210–223.e6 (2017).

89. L. A. Gilbert, M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C. Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, L. S. Qi, M. Kampmann, J. S. Weissman, Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*. **159**, 647–661 (2014).

90. M. A. Horlbeck, L. A. Gilbert, J. E. Villalta, B. Adamson, R. A. Pak, Y. Chen, A. P. Fields, C. Y. Park, J. E. Corn, M. Kampmann, J. S. Weissman, Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*. **5** (2016), doi:10.7554/eLife.19760.

91. B. Adamson, T. M. Norman, M. Jost, M. Y. Cho, J. K. Nuñez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev, J. S. Weissman, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*. **167**, 1867–1882.e21 (2016).

92. Y. E. Tak, B. P. Kleinstiver, J. K. Nuñez, J. Y. Hsu, J. E. Horng, J. Gong, J. S. Weissman, J. K. Joung, Inducible and multiplex gene regulation using CRISPR-Cpf1-based transcription factors. *Nat. Methods*. **14**, 1163–1166 (2017).

93. W. Bergsma, A bias-correction for Cramér's and Tschuprow's. *Journal of the Korean Statistical Society*. **42** (2013), pp. 323–328.

94. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods*. **15**, 1053–1058 (2018).

95. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. **27**, 1739–1740 (2011).

96. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Acknowledgements: We thank D. Yang, A. Khodaverdian, and R. Zhang for discussions; M. Jost and J.K. Nuñez for plasmids; and UCSF Center for Advanced Technology (E. Chow and S. Elmes) and UCSF Preclinical Therapeutics Core (B. Hann). **Funding:** This work was supported by NIH-NIGMS F32GM125247 (J.J.Q.); UCSF Discovery Fellowship (M.G.J.); NIH K08CA222625 (R.A.O.); NIH R01DA036858 and 1RM1HG009490 (J.S.W.); NIH R01CA231300, U54CA224081, R01CA204302, R01CA211052, and R01CA169338, and the Pew and Stewart Foundations (T.G.B.); NIH-NIAID U19AI090023 (N.Y.), and Chan-Zuckerberg Initiative 2018-184034. J.S.W. is a Howard Hughes Medical Institute Investigator. M.M.C. is a Gordon and Betty Moore fellow of the Life Sciences Research Foundation. **Author contributions:** All authors contributed to the design of experiments and analysis. J.J.Q. engineered cell lines, processed tissues, and prepared sequencing libraries. R.A.O. performed mouse surgeries and imaging. M.G.J. and J.J.Q. processed lineage tracing sequencing data. S.N. and J.J.Q. performed invasion assays. M.G.J. performed phylogenetic reconstruction and analyzed the trees and single-cell RNA-sequencing data. M.G.J. and N.Y. conceived and implemented the *FitchCount* algorithm. All authors aided in the interpretation of the analyses. J.J.Q., M.G.J., and J.S.W. wrote the manuscript, and all authors read and approved the final manuscript. **Competing interests:** J.S.W. is an advisor and/or has equity in KSQ Therapeutics, Maze Therapeutics, Amgen, Tenaya, and 5 AM Ventures. T.G.B. is an advisor to Novartis, Astrazeneca, Revolution Medicines, Array, Springworks, Strategia, Relay, Jazz, Rain and receives research funding from Novartis and Revolution Medicines. **Data and materials availability:** The A549-LT cell line will be available via material transfer agreement. Raw sequencing reads and processed lineage tracing data files will be made available via GEO accession upon acceptance. Lineage tracer processing pipeline, phylogenetic reconstruction algorithm, and *FitchCount* are publicly available via the GitHub repository for Cassiopeia (github.com/YosefLab/Cassiopeia); all other analysis scripts and notebooks will be made publicly available upon acceptance on Github (github.com/mattjones315/MetastasisTracing).

List of Supplementary Materials:

Materials and Methods

Figs. S1-S24

Supplementary Text

sgRNA sequences for CRISPRi/a

Gene Lists

Figures and Figure Legends: (embedded below)

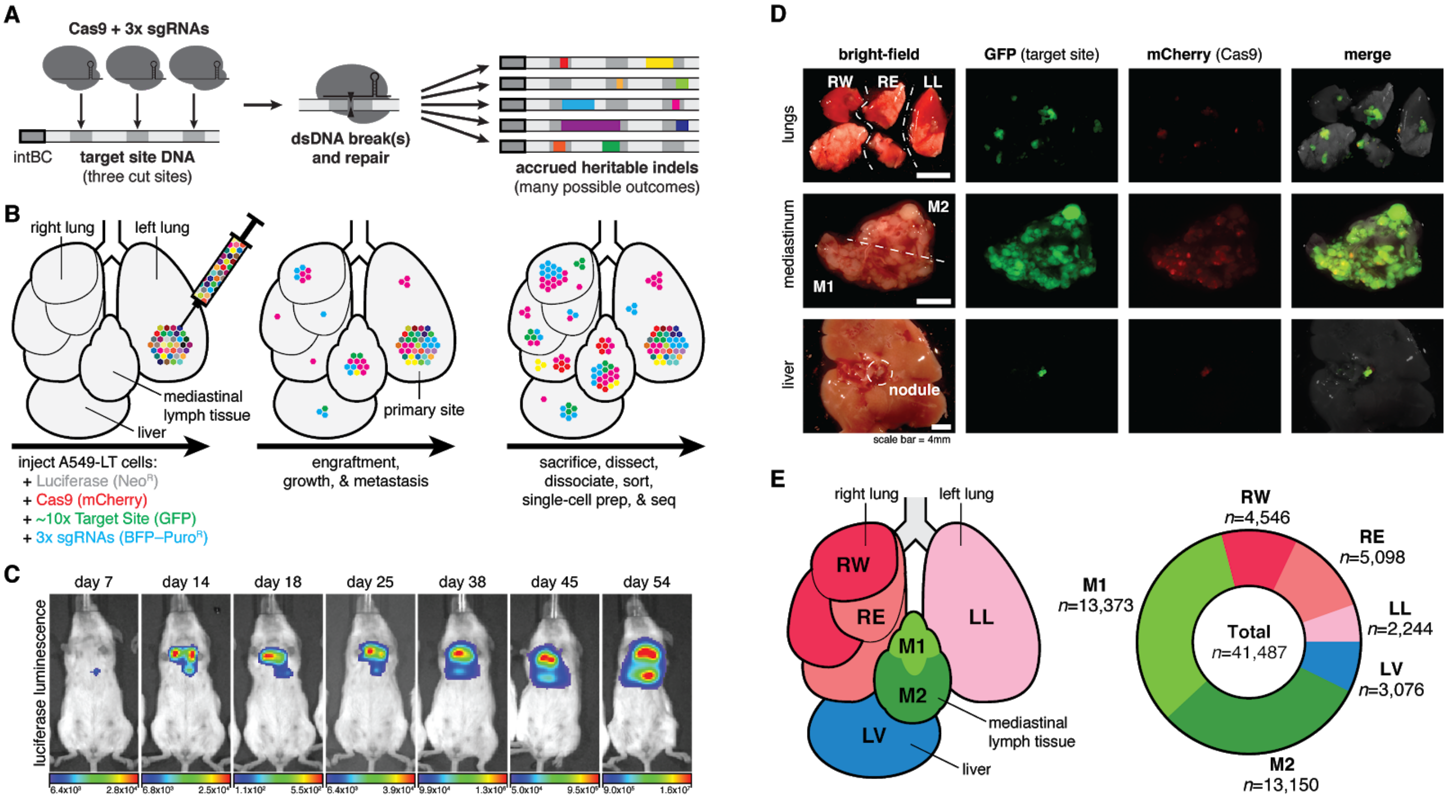


Fig. 1. Lineage tracing in a lung cancer xenograft model in mice. (**A**) Our Cas9-enabled lineage tracing technology. Cas9 and three sgRNAs bind and cut cognate sequences on genomically integrated Target Sites, resulting in diverse indel outcomes (multicolored rectangles), which act as heritable markers of lineage. (**B**) Xenograft model of lung cancer metastasis. Approximately 5,000 A549-LT cells were surgically implanted into the left lung of immunodeficient mice. The cells engrafted at the primary site, proliferated, and metastasized within the five lung lobes, mediastinal lymph, and liver. (**C**) *In vivo* bioluminescence imaging of tumor progression over 54 days of lineage recording, from early engraftment to widespread growth and metastasis. (**D**) Fluorescent imaging of collected tumorous tissues. (**E**) Anatomical representation of the six tumorous tissue samples (**left**), and the number of cells collected with paired single-cell transcriptional and lineage datasets (**right**).

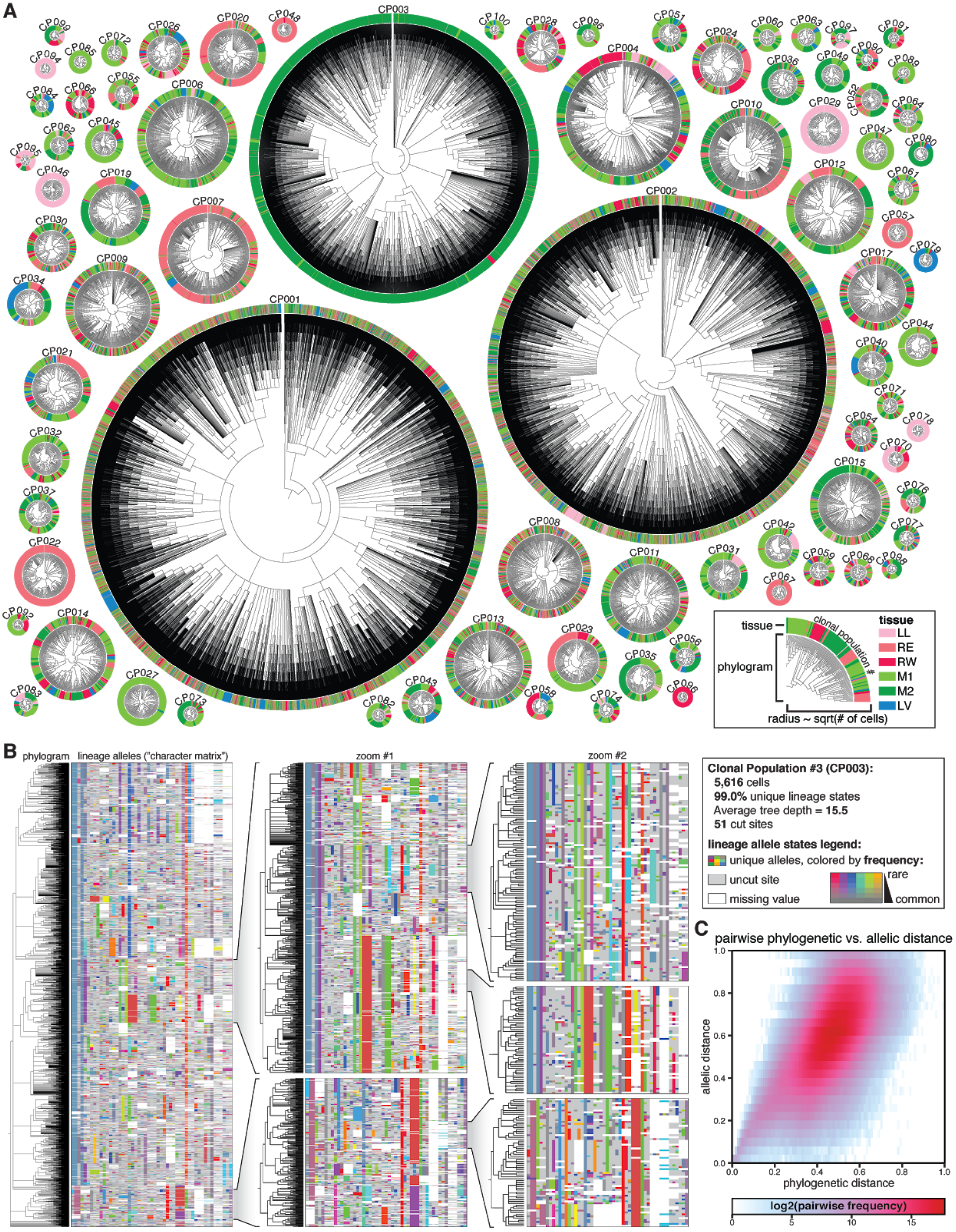

A
CP099
CP094
CP085
CP072
CP026
CP020
CP048
CP003
CP100
CP028
CP096
CP051
CP024
CP060
CP063
CP097
CP091
CP084
CP066
CP055
CP006
CP004
CP036
CP049
CP090
CP089
CP062
CP045
CP010
CP029
CP052
CP064
CP095
CP046
CP019
CP047
CP080
CP030
CP007
CP012
CP061
CP002
CP057
CP009
CP017
CP079
CP034
CP001
CP044
CP040
CP021
CP071
CP054
CP078
CP032
CP070
CP015
CP037
CP076
CP008
CP042
CP077
CP022
CP011
CP031
CP059
CP068
CP098
CP067
CP092
CP014
CP013
CP023
CP056
CP035
CP027
CP043
CP058
CP086
CP083
CP073
CP082
CP074
tissue
phylogram
clonal population #
radius ~ sqrt(# of cells)
tissue
LL
RE
RW
M1
M2
LV
B
phylogram
lineage alleles ("character matrix")
zoom #1
zoom #2
Clonal Population #3 (CP003):
5,616 cells
99.0% unique lineage states
Average tree depth = 15.5
51 cut sites
lineage allele states legend:
unique alleles, colored by frequency:
uncut site
missing value
rare
common
C
pairwise phylogenetic vs. allelic distance
allelic distance
phylogenetic distance
0.0
0.2
0.4
0.6
0.8
1.0
log2(pairwise frequency)
0
5
10
15

Fig. 2. High-resolution phylogenetic trees capture the histories of clonal cancer populations. (**A**) Highly detailed phylogenetic reconstructions for each clonal population, represented as radial phylograms. Each cell is represented along the circumference and colored by tissue, as in **Fig. 1E** and legend. Trees differ in size, tissue distribution, and frequency of tissue transitions. Each tree is scaled by the square-root of the number of cells. (**B**) Phylogenetic tree and lineage alleles of one clonal population (CP003; *N*=5,616 cells). The phylogram (**left**) represents cell–cell relationships and the matrix (**right**) represents the lineage alleles for each cell. Alleles are uniquely colored, where saturation indicates allele rarity (legend). (**B, inlays**) Nested zooms of individual clades show the patterns of shared and distinguishing indel alleles, and highlight indel diversity, tree depth, and tree complexity. (**C**) Correspondence between phylogenetic distance (the normalized pairwise tree distance between two cells) and allelic distance (the normalized pairwise difference in alleles between two cells) for CP003, indicating that the tree accurately models phylogenetic relationships.

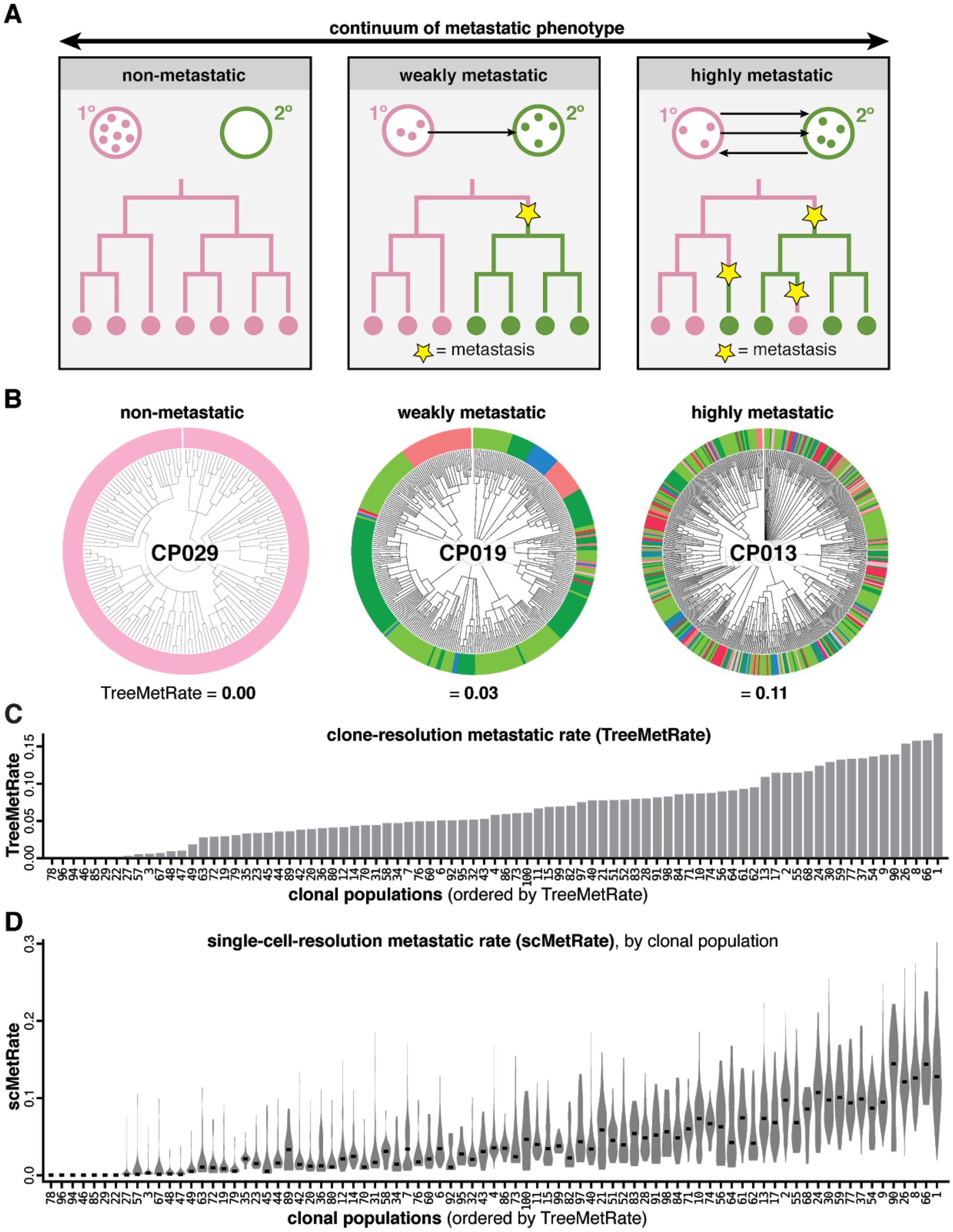


Fig. 3. Quantifying the diverse metastatic phenotypes of clonal populations directly from cell lineages. (**A**) Theoretical continuum of metastatic phenotypes, spanning non-metastatic (never exiting the primary site) to highly metastatic (frequently transitioning between tumors; arrows). Ancestral metastatic events between tissues leave clear phylogenetic signatures (yellow stars). (**B**) Example clonal populations that illustrate the wide range of metastatic phenotypes observed: a non-metastatic population that never exits the primary site (CP029); a moderately metastatic population that infrequently transitions between different tissues (CP019); and a

frequently metastasizing population with closely related cells residing in different tissues (CP013). Cells colored by tissue as in **Fig. 1E**; metastatic phenotypes scored by the TreeMetRate. (**C**) The distribution of TreeMetRates for each clonal population. (**D**) The distributions of single-cell-resolution metastatic phenotypes (scMetRates) for each clonal population, rank-ordered by TreeMetRate; median scMetRate indicated in black.

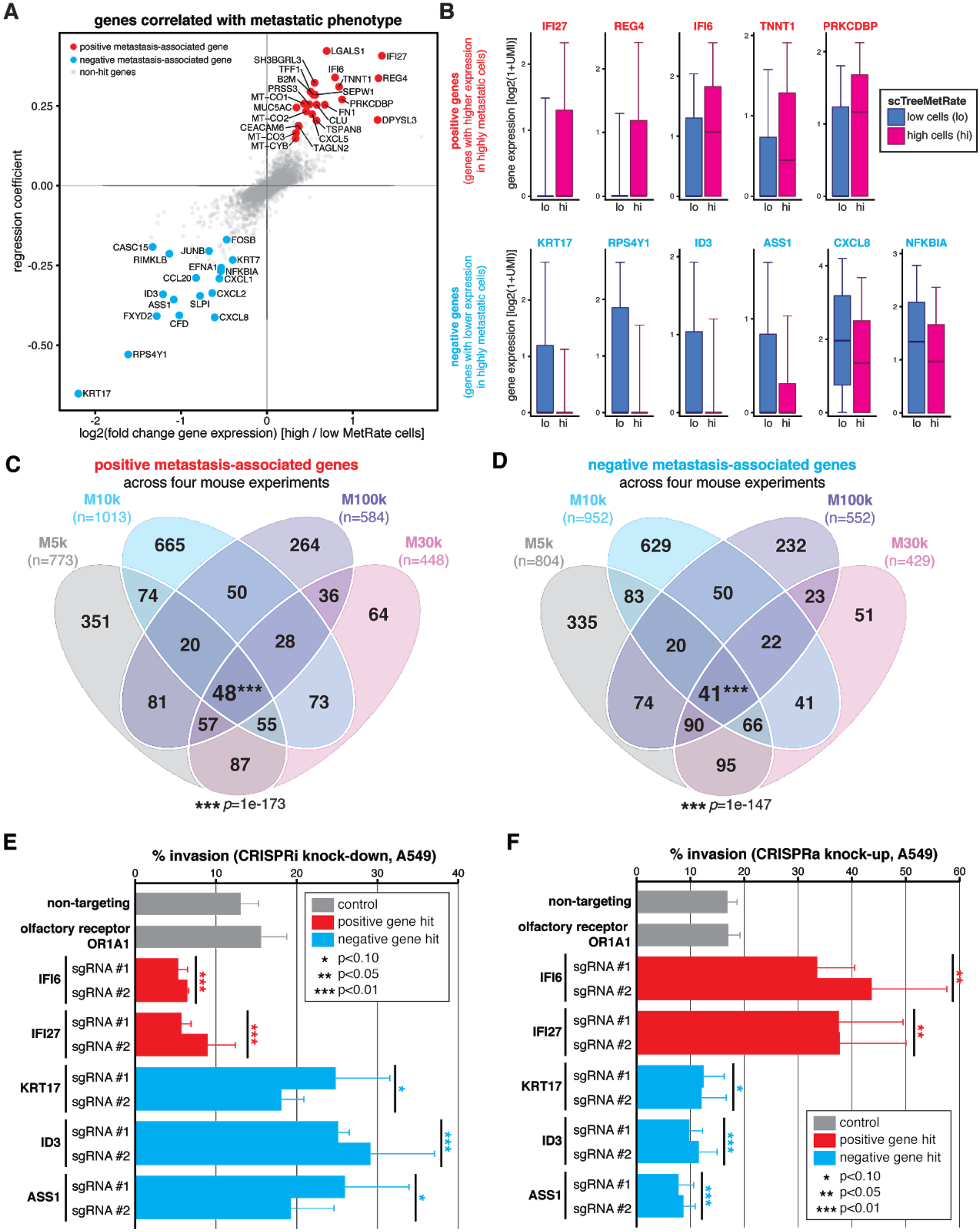

A
genes correlated with metastatic phenotype
positive metastasis-associated gene
negative metastasis-associated gene
non-hit genes
regression coefficient
0.25
0.00
-0.25
-0.50
log2(fold change gene expression) [high / low MetRate cells]
-2
-1
0
1
LGALS1
IFI27
SH3BGRL3
TFF1
IFI6
B2M
TNNT1
REG4
PRSS3
SEPW1
MT-CO1
PRKCDBP
MUC5AC
FN1
MT-CO2
CLU
DPYSL3
CEACAM6
TSPAN8
MT-CO3
CXCL5
MT-CYB
TAGLN2
FOSB
CASC15
JUNB
RIMKLB
KRT7
EFNA1
NFKBIA
CCL20
CXCL1
ID3
CXCL2
ASS1
SLPI
FXYD2
CFD
CXCL8
RPS4Y1
KRT17
B
positive genes
(genes with higher expression in highly metastatic cells)
gene expression [log2(1+UMI)]
IFI27
REG4
IFI6
TNNT1
PRKCDBP
scTreeMetRate
low cells (lo)
high cells (hi)
negative genes
(genes with lower expression in highly metastatic cells)
KRT17
RPS4Y1
ID3
ASS1
CXCL8
NFKBIA
lo
hi
C
positive metastasis-associated genes
across four mouse experiments
M10k (n=1013)
M100k (n=584)
M5k (n=773)
M30k (n=448)
665
264
74
50
36
351
64
20
28
48***
81
73
57
55
87
*** p=1e-173
D
negative metastasis-associated genes
across four mouse experiments
M10k (n=952)
M100k (n=552)
M5k (n=804)
M30k (n=429)
629
232
83
50
23
335
51
20
22
41***
74
41
90
66
95
*** p=1e-147
E
% invasion (CRISPRi knock-down, A549)
0
10
20
30
40
non-targeting
olfactory receptor OR1A1
IFI6
IFI27
KRT17
ID3
ASS1
sgRNA #1
sgRNA #2
control
positive gene hit
negative gene hit
* p<0.10
** p<0.05
*** p<0.01
F
% invasion (CRISPRa knock-up, A549)
0
10
20
30
40
50
60
non-targeting
olfactory receptor OR1A1
IFI6
IFI27
KRT17
ID3
ASS1
sgRNA #1
sgRNA #2
control
positive gene hit
negative gene hit
* p<0.10
** p<0.05
*** p<0.01

Fig. 4. Divergent metastatic phenotypes are driven by differences in gene expression. (**A**) Poisson regression analysis of single-cell gene expression and scMetRate for all cells and all tissues; fold-change and coefficient of regression shown. The strongest and most significant positive and negative genes are annotated (red and blue, respectively; Methods). (**B**) Expression level of several positive and negative metastasis-associated gene candidates (top and bottom rows, respectively) in cells with low or high scMetRate (blue and magenta box-plots, respectively). Boxes: first, second, and third quartiles; whiskers: 9th and 91st percentiles of expression distribution. (**C** and **D**) Overlap of identified positive and negative metastasis-associated genes, respectively, from the four mouse experiments; number of genes indicated. Four-way intersections between gene sets are significant by *SuperExactTest* (*87*) multi-set intersection test. (**E** and **F**) *In vitro* transwell invasion assays following CRISPRi or CRISPRa gene perturbation, respectively, in A549 cells. Perturbations were performed using two independent sgRNAs per gene. Differences in invasion phenotype relative to two negative control guides (non-targeting and olfactory receptor) were significant by two-tailed *t*-test; error bars show standard deviation across triplicates.

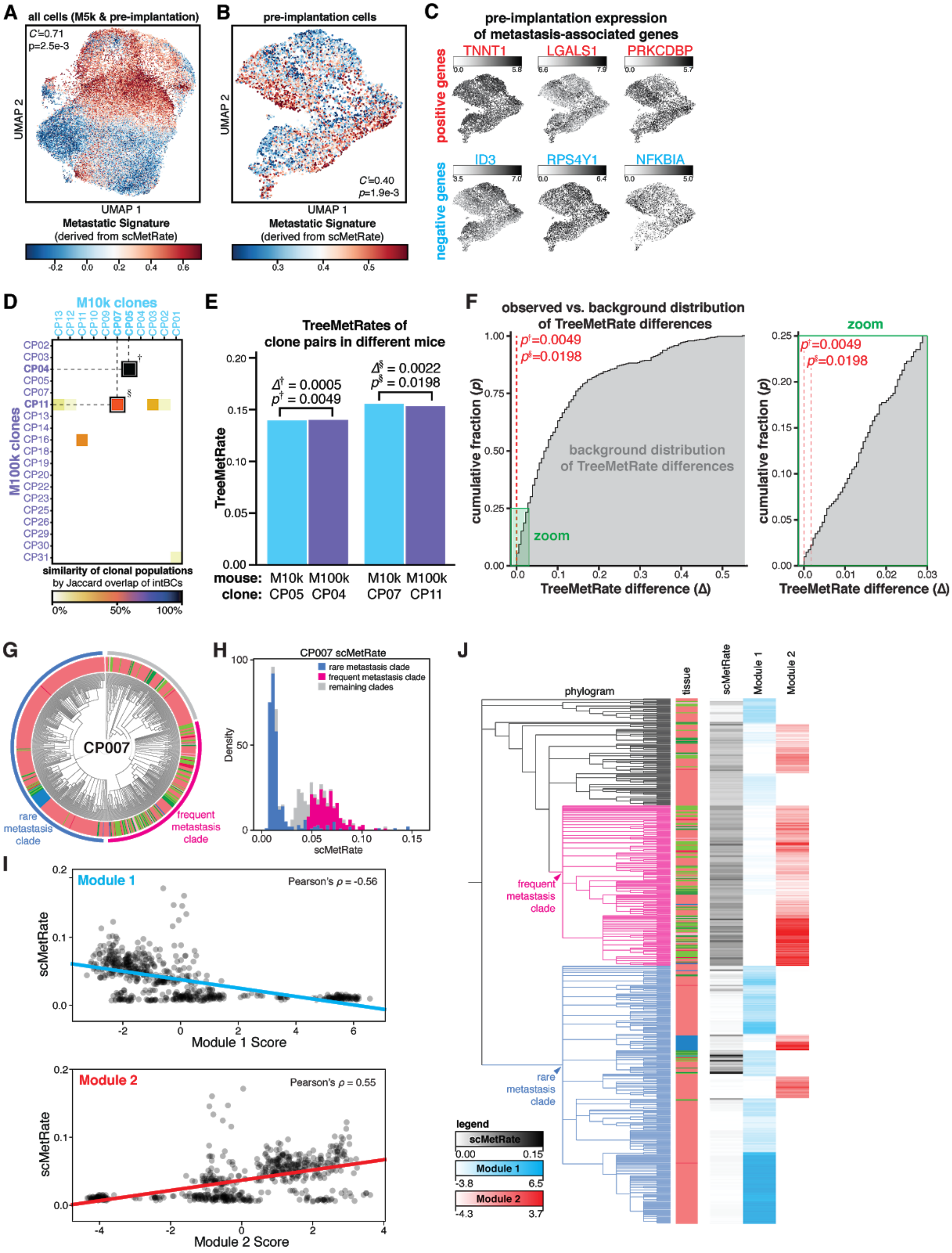

A
all cells (M5k & pre-implantation)
C′=0.71
p=2.5e-3
UMAP 2
UMAP 1
Metastatic Signature
(derived from scMetRate)
-0.2 0.0 0.2 0.4 0.6
B
pre-implantation cells
C′=0.40
p=1.9e-3
UMAP 2
UMAP 1
Metastatic Signature
(derived from scMetRate)
0.3 0.4 0.5
C
pre-implantation expression
of metastasis-associated genes
positive genes
TNNT1 0.0 5.8
LGALS1 6.6 7.9
PRKCDBP 0.0 5.7
negative genes
ID3 3.5 7.0
RPS4Y1 0.0 6.4
NFKBIA 0.0 5.0
D
M10k clones
CP13 CP12 CP11 CP10 CP09 CP07 CP05 CP04 CP03 CP02 CP01
M100k clones
CP02 CP03 CP04 CP05 CP07 CP11 CP13 CP14 CP16 CP18 CP19 CP20 CP22 CP23 CP25 CP26 CP29 CP30 CP31
†
§
similarity of clonal populations
by Jaccard overlap of intBCs
0% 50% 100%
E
TreeMetRates of
clone pairs in different mice
Δ† = 0.0005
p† = 0.0049
Δ§ = 0.0022
p§ = 0.0198
TreeMetRate
0.00 0.05 0.10 0.15 0.20
mouse: M10k M100k M10k M100k
clone: CP05 CP04 CP07 CP11
F
observed vs. background distribution
of TreeMetRate differences
p†=0.0049
p§=0.0198
cumulative fraction (p)
0.00 0.25 0.50 0.75 1.00
background distribution
of TreeMetRate differences
zoom
0.0 0.1 0.2 0.3 0.4 0.5
TreeMetRate difference (Δ)
zoom
p†=0.0049
p§=0.0198
cumulative fraction (p)
0.00 0.05 0.10 0.15 0.20 0.25
0.0 0.01 0.02 0.03
TreeMetRate difference (Δ)
G
CP007
rare
metastasis
clade
frequent
metastasis
clade
H
CP007 scMetRate
rare metastasis clade
frequent metastasis clade
remaining clades
Density
0 50 100
0.00 0.05 0.10 0.15
scMetRate
I
Module 1
Pearson's ρ = -0.56
scMetRate
0.0 0.1 0.2
-2 0 2 4 6
Module 1 Score
Module 2
Pearson's ρ = 0.55
scMetRate
0.0 0.1 0.2
-4 -2 0 2 4
Module 2 Score
J
phylogram
tissue
scMetRate
Module 1
Module 2
frequent
metastasis
clade
rare
metastasis
clade
legend
scMetRate 0.00 0.15
Module 1 -3.8 6.5
Module 2 -4.3 3.7

Fig. 5. Metastatic phenotype is predetermined, heritable, and reproducible. (**A** and **B**) Projections of transcriptional states of M5k cancer cells and pre-implantation cells (**A**) or pre-implantation cells alone (**B**), colored by Metastatic Signature. Association between transcriptional state and Metastatic Signature is measured by inverted Geary's *C'* and significance by false discovery rate (*p*). (**C**) Pre-implantation cells exhibit heterogeneity in expression of metastasis-associated genes. (**D**) Jaccard overlap of intBC sets between clonal populations in M10k and M100k mice. Two pairs of clonal populations (indicated by † and §) were related between the two mouse experiments (Jaccard overlap>50%). (**E**) Comparison of TreeMetRates from related clones implanted in M10k and M100k, showing minimal difference in metastatic rate (Δ) between clone pairs. (**F**) Cumulative distribution plot of the background distribution of all possible pairwise TreeMetRate differences between M10k and M100k clones (gray), with zoom to show low-Δ regime. Both of the observed differences are statistically smaller than expected ($p^{\dagger}$=0.0049 and $p^{\S}$=0.0198; red dashes). (**G**) Divergent subclonal metastatic behavior exhibited in the phylogenetic tree of clonal population #7, with annotated subclades; cells colored by tissue as in **Fig. 1E**. (**H**) The bimodal distribution of scMetRates for cells in CP007, with cells from the divergent subclades indicated. (**I**) Comparison of single-cell metastatic phenotype and *Hotspot* transcriptional module scores. (**J**) Overlay illustrating concordance between CP007 phylogeny, scMetRates, and *Hotspot* Module scores.

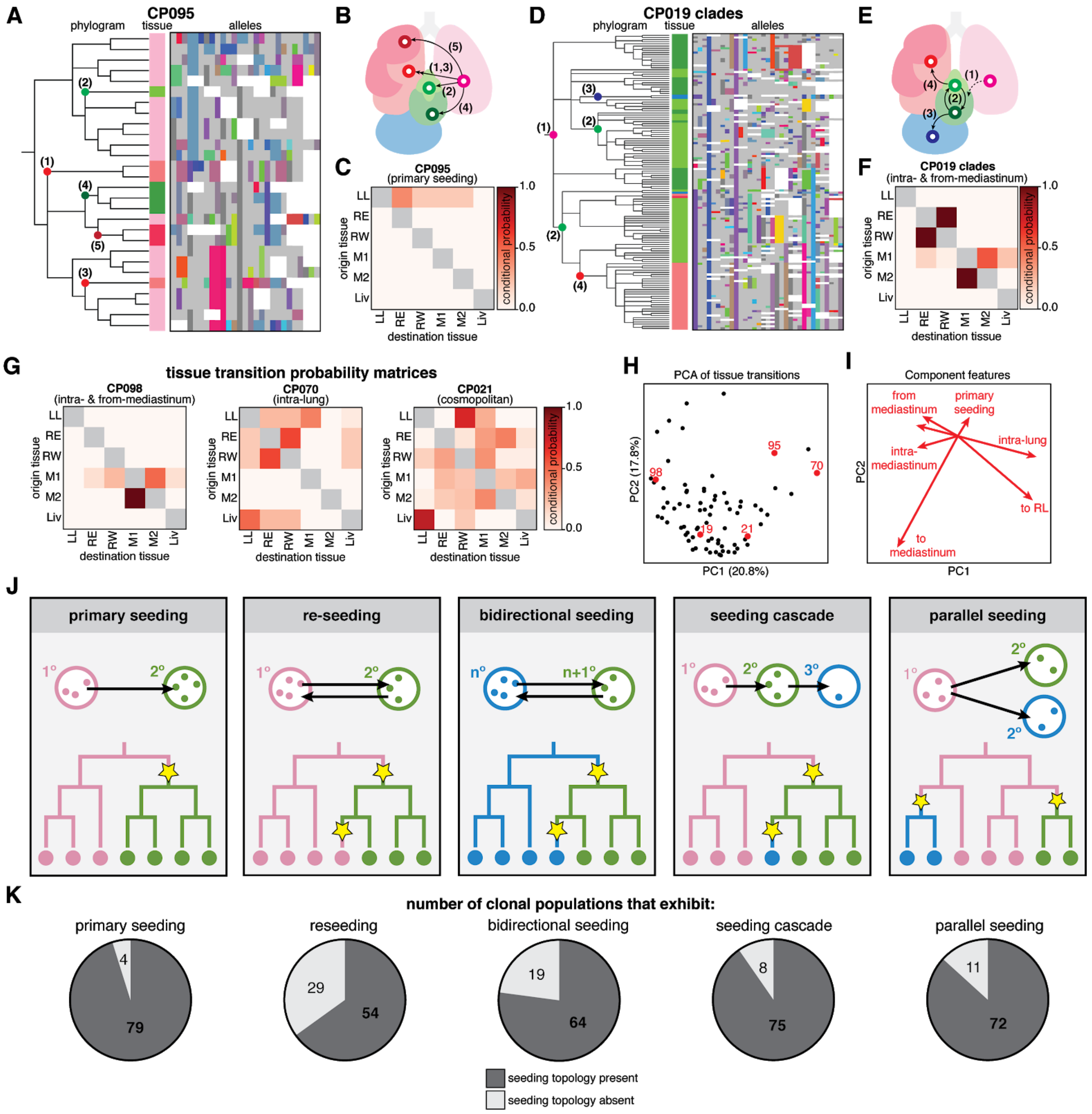


Fig. 6. Metastases were seeded via complex tissue routes and multidirectional topologies. (**A** and **D**) Phylogenetic trees and lineage alleles for clonal population #95 and #19 clades, respectively. Notable metastatic events are annotated in the phylogram and represented graphically as arrows (**B** and **E**); cells colored by tissue as in **Fig. 1E**; lineage alleles colored as in **Fig. 2B**; dashed arrow indicates an assumed transition. (**C** and **F**)

Tissue transition matrices representing the conditional probability of metastasizing from and to tissues, defining the tissue routes of metastasis for each clonal population. CP095 solely exhibits primary seeding from the left lung, whereas CP019 shows more complex seeding routes. (**G**) Tissue transition matrices illustrating the diversity of tissue routes, including metastasis from and within the mediastinum (**left**), between the lung lobes (**middle**), or amply to and from all tissues (**right**). (**H**) Principal component analysis (PCA) of tissue transition probabilities for each clonal population. Displayed clones are annotated in red; percentage of variance explained by components indicated on axes. (**I**) Component vectors of PCA with descriptive features. (**J**) Possible topologies of metastatic seeding, represented graphically and phylogenetically as in **Fig. 3A**. (**K**) Number of clonal populations that exhibit each metastatic seeding topology.

Supplemental Materials for "Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts" (Quinn et al. 2020)

List of Supplemental Materials:
Materials and Methods
Supplemental Figures and Legends
Supplementary Text

Materials and Methods:

Cell culture. A549 cells (human lung adenocarcinoma line, American Type Culture Collection CCL-185) and H1299 cells (ATCC CRL-5803) were maintained in Dulbecco's modified eagle medium (DMEM, Gibco) supplemented with 10% (v/v) fetal bovine serum (FBS; VWR Life Science Seradigm), 2 mM glutamine, 100 units/mL penicillin, and 100 µg/mL streptomycin (hereafter "complete DMEM"). Cells were cultured at 37ºC in a humidified 5% (v/v) CO_2 atmosphere. Cells were split into fresh culture medium every two to three days by trypsinization with TrypLE reagent (Gibco) quenched with complete DMEM, and maintained at cell density as recommended by ATCC.

Plasmid design and cloning. The triple-sgRNA-BFP-Puromycin^R lentivector, PCT62 (to be made available on Addgene), was constructed using four-way Gibson assembly (NEB) as previously described (*88*), and expresses three sgRNA cassettes driven by distinct U6 promoters, with constitutive BFP and puromycin-resistance markers for selection. The three sgRNAs are complementary to the three cut-sites in the Target Site (PCT48), except for precise single base-pair mismatches that decrease their avidity for the cognate cut-sites and subsequently slow lineage recording kinetics (*89*). Guide RNAs for CRISPRi and CRISPRa experiments were chosen from the human CRISPRi/a v.2 libraries (Addgene #83969 and #83978, respectively) and were cloned into the pLG1 lentiviral backbone (Addgene #84832) using the annealing and ligation as previously described (*90*).

Lentivirus preparation and infection method. Lentivirus was produced by transfecting HEK293T cells with standard 4th generation packaging vectors delivered by TransIT-LTI transfection reagent (Mirus) as described in (*34*) and (*91*). Target Site (PCT48) lentiviral supernatant was concentrated 10-fold using Lenti-X Concentrator (Takara Bio) according to manufacturer's instructions. Viral preparations were filtered and frozen prior to infection. Triple-sgRNA lentiviral preparation (PCT62) was titered and diluted to a concentration to yield approximately 50% infection rate. All infections were performed in 6-well plates with 2 mL media and $2x10^5$ adhered cells per well; 1 mL of titered lentivirus was added to the culture medium with 8 µg/mL of polybrene and incubated at 37ºC overnight, after which media was replaced and cells were expanded to larger culture volumes as needed.

Cell line engineering and selection strategies. Two separate lineage tracing-competent A549 cell lines (hereafter, "A549-LT1" and "A549-LT2") were engineered separately and by different methods, primarily distinguished by (i) the method of high copy-number transduction of the Target Site (piggyBac transposition and high-titer lentiviral infection, respectively) and (ii) the version of the "molecular recorder" technology used (the original components described in (*23*) and the improved components described in (*34*), respectively). For both A549-LT1 and -LT2, the cells were first infected with Luciferase-Neomycin^R lentivirus; two days following infection, *neomycin(+)* cells were selected by treatment with 800 µg/mL G418 Geneticin (Thermo Fisher) every second day for 10 days, expanding the cells as necessary. Next, the A549+Luciferase cells were infected with either Cas9-P2A-BFP or Cas9-P2A-mCherry lentivirus, respectively; three days following infection, *BFP(+)* or *mCherry(+)* cells were collected by fluorescence-activated cell sorting (FACS) using the

BD FACS Aria II at the UCSF Center for Advanced Technology core. Next, the A549+Luciferase+Cas9 cells were serially transduced with different versions of the GFP-TargetSite vector (transposon-based PCT17 or lentivector-based PCT48, respectively). For the A549-LT1 line, the transposon transduction was performed by electroporation using an Amaxa Nucleofector (Lonza) with 100 ng of PiggyBac transposase plasmid (SBI System Biosciences) and 500 ng of PCT17 transposon plasmid. For the A549-LT2 line, the high-titer PCT48 lentiviral infections were performed as above, but in triplicate and pooled to further maximize diversity. Following GFP-TargetSite transduction, *GFP(+)* cells were collected by FACS. These steps were repeated two additional times for a total of three serial transductions of the Target Site, fluorescence-sorting cells with progressively higher GFP fluorescence after each transduction (**Fig. S2**). Finally, four days prior to implantation, A549+Luciferase+Cas9+TargetSite cells were transduced by titered triple-sgRNA (with BFP-Puromycin^{R} markers; PCT61) lentivirus; 36 hours following infection, *BFP(+)* cells were collected by FACS and returned to *in vitro* culture until final preparation for implantation, thus producing lineage tracing-competent A549-LT1 and -LT2 cell lines.

Mouse care. Mouse experiments were performed as in (*38*). Six- to eight-week-old female SCID C.B-17 mice (C.B-*Igh-1*^{b}/IcrTac-*Prkdc*^{scid}; Taconic) were maintained in specific-pathogen-free conditions in facilities approved by the American Association for Accreditation of Laboratory Animal Care. Surgical procedures were reviewed and approved by the UCSF Institutional Animal Care and Use Committee (IACUC), Protocol #AN107889-03C.

Orthotopic lung xenografts. To prepare A549-LT cell suspensions for implantation, cells were collected from culture by trypsinization and quenched with complete DMEM. Cells were then washed in cold PBS and resuspended with cold Matrigel matrix (BD Bioscience) at the appropriate final concentration (500, 1000, 3000, and 10,000 cells/µL for mice M5k, M10k, M30k, and M100k, respectively). The Matrigel cell suspensions were gently mixed and transferred into a 1-mL syringe and remained on ice until implantation. Orthotopic implantations were performed as in (*38*): mice were placed in the right lateral decubitus position and anesthetized with 2.5% inhaled isoflurane. A 1-cm surgical incision was made along the posterior medial line of the left thorax. Fascia and adipose tissue layers were dissected and retracted to expose the lateral ribs, the intercostal space, and the left lung parenchyma. Upon recognition of left lung respiratory variation, a 30-gauge hypodermic needle was used to advance through the intercostal space approximately 3 mm into the lung tissue. Care was taken to inject 10 µL (5,000, 10,000, 30,000, or 100,000 cells, respectively) of cell suspension directly into the left lung. Mice were observed post-procedure for 1–2 hours, and their body weights and wound healing were monitored weekly. Experiments were performed across two mouse cohorts: A549-LT1 cells were implanted in the first mouse cohort (including mice M10k and M100k); A549-LT2 cells were implanted in the second cohort (including mice M5k and M30k).

Bioluminescence imaging. Mice were imaged with the Xenogen IVIS 100 bioluminescent imaging system at the UCSF Preclinical Therapeutics Core. Bioluminescence monitoring of the tumor engraftment and metastatic progression was performed biweekly. Before imaging, mice were anesthetized with inhaled isoflurane and injected intraperitoneally with 200 µL of *D*-Luciferin at a dose of 150 mg/kg body weight. For each mouse, bioluminescent signal was measured from the thoracic cavity in the supine position and calculated automatically using Living Image Software; radiance units are photons/s/cm^{2}/steradian. Mice were sacrificed after bioluminescent signal was anatomically extensive throughout the thorax but before the mice exhibited labored breathing (53 to 80 days post-implantation; varied by mouse). Mice were injected with *D*-Luciferin as before and subsequently killed; the heart and lungs were resected *en bloc*, the heart was removed, and the right and left lung lobes were separated. Tumorous tissues were imaged *ex vivo* by either bioluminescence (performed as before) or fluorescence using a stereo fluorescent microscope (Nikon).

CRISPRi/a perturbations and invasion assays. CRISPRi/a cell lines were first generated by infecting parental cell lines with lentivirus carrying either the CRISPRi (dCas9-BFP-KRAB; Addgene #46911) or CRISPRa (dCas9-XTEN-VPR-GFP (*92*)) genetic component, respectively; stably transduced cells were selected to purity by FACS. CRISPRi and CRISPRa activity was confirmed by perturbing the gene expression of the cell-surface marker genes CD81 and CD151 (for CRISPRi) and CXCR4 (for CRISPRa); perturbations were performed by infecting with lentivirus carrying sgRNA against the target marker genes (see sgRNA spacer sequences in the Supplemental Materials) and transduced cells were selected for 4 days with puromycin as described above. Seven days after transduction, the perturbed cells were collected and stained with APC-labelled antibodies against CD81, CD151, and CXCR4 (BioLegend), and the fluorescence of >10^3 individual cells was measured by flow cytometry (Attune NxT, Thermo Fisher Scientific). For invasion assays, CRISPRi and CRISPRa cell lines were treated with lentivirus carrying sgRNAs against the targeted metastasis-associated gene candidates (see sgRNA spacer sequences in the Supplemental Materials). We selected five candidate genes of interest to assay based on the following criteria: the strongest positive (IFI27) or negative (KRT17) genes identified from mouse M5k; in the same gene family as IFI27 (interferon-induced; IFI6); or previous evidence of modulating invasion phenotype in similar models (ASS1, ID3) (*59*, *60*). Cells transduced with sgRNAs against these genes were selected and cultured for one week, as above. For the invasion assays, DMEM supplemented with 10% FBS was added to the bottom chamber of a 24-well trans-well plate. The perturbed cells were counted, and 1.5×10^4 cells were resuspended in serum-free media and added to the top chamber of 8-μm pore matrigel-coated (invasion) or non-coated (migration) trans-well inserts (Corning BioCoat). After 16 hours, non-invading cells on the apical side of inserts were scraped off and the trans-well membrane was fixed in methanol for 15 minutes and stained with Crystal Violet for 30 minutes. The basolateral surface of the membrane was visualized with a Zeiss Axioplan II immunofluorescence microscope at 10×. Each trans-well insert was counted manually at 10× and performed in triplicate. Invasion phenotype was calculated as the number of invading cells through the matrigel membrane divided by the mean number of cells migrating through control insert. Differential invasiveness was assessed using a two-tailed *t*-test comparing the invasion rate of the perturbation to both negative controls.

Tissue homogenization and single-cell preparation. Collected tumorous tissues were minced on ice and placed in separate gentleMACS C Tubes (Miltenyi Biotec) with 5 mL each of tissue dissociation buffer (30 mL DMEM/F-12 culture (Gibco) supplemented with 16 mg collagenase (Worthington Biochemical), 5 mg Liberase TL (Roche), and 100 U DNase I (Worthington Biochemical)). Tissues were homogenized on a gentleMACS Octo Dissociator (Miltenyi Biotec) for one minute, then incubated in a 37ºC water bath for five minutes followed by repeated trituration using a 1000 μL pipette tip; incubation and trituration were repeated until tissues were completely dissociated, three to six times in total. Cells were pelleted by centrifugation at 500 rcf for 5 minutes at 4ºC. To remove red blood cells, the cell pellets were resuspended in 1 mL ACK Lysing Buffer (Gibco) and incubated at room temperature for 3 minutes, followed by centrifugation. To achieve a single-cell resuspension, the cell pellet was resuspended in 500 μL TrypLE reagent and incubated at 37ºC for 5 minutes followed by trituration; digestion was quenched with 1 mL complete DMEM medium, centrifuged, and resuspended in 1 mL of cold complete DMEM. Cells were strained through a 40 μm nylon filter-cap tube (Corning). Cancer cells were fluorescence-sorted (identified by high GFP and BFP fluorescence) at 4ºC to remove mouse cells and debris, pelleted, resuspended in 1 mL cold phosphate-buffered saline with 0.04% bovine serum albumin, and filtered through a 40 μm FlowMi filter-tip (Bel-Art). The concentration of cells was determined on an Accuri flow cytometer (BD Biosciences), then pelleted. Finally, the cells were thoroughly resuspended in 20 μL phosphate-buffered saline with 0.04% bovine serum albumin, mixed with the Chromium Single Cell 3′ V2 Kit reagents (10X Genomics) according to the User Guide, and then loaded into droplet emulsions using the Chromium Controller, thus facilitating single-cell transcriptional measurements. Separate Chromium lanes were used for each tissue sample.

Sequencing library preparations. Gene expression libraries were prepared according to the Chromium Single Cell 3′ V2 Kit User Guide. To prepare the Target Site libraries, the amplified cDNA libraries were further amplified with Target Site-specific primers containing Illumina-compatible adapters and sample indices (forward: 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAATCCAGCTAGCTGTGCAGC; reverse: 5′-AATGATACGGCGACCACCGAGATCTACACNNNNNNNNTCTTTCCCTACACGACGCTCTTCCGATCT; "N" denotes sample indices) using Kapa HiFi HotStart ReadyMix (Roche), as described in (*34*). Approximately 24 fmol of template cDNA was used per sample, divided between four identical reactions to avoid possible PCR-induced library biases. PCR products were purified and size-selected using SPRI magnetic beads (Beckman) and quantified by BioAnalyzer (Agilent).

Library sequencing. Sequencing libraries from each tissue sample were pooled to yield approximately equal coverage per cell per sample; gene expression libraries and Target Site amplicon libraries were pooled in an approximately 10:1 molar ratio to yield more RNA reads than Target Site reads. Libraries were further pooled with approximately 5% PhiX genomic DNA library added for quality-control. The libraries were sequenced using a custom sequencing strategy on the NovaSeq S2 platform (Illumina) in order to read the full-length Target Site amplicons. Sample identities were read as dual indices (I1 and I2: 8 cycles each); the 10X cell barcode and unique molecular identifier (UMI) sequences were read first (R1: 26 cycles) and the Target Site sequence was read second (R2: >250 cycles). Over 4.70 billion sequencing clusters passed Illumina QC filters and were processed as described below. All raw and processed data will be made available on GEO (accession pending). Only the first 98 bases per read were used for analysis in the RNA expression libraries to mask the longer reads required to sequence the Target Sites.

TargetSite sequencing data processing. Raw Target Site sequencing data was processed using the Cassiopeia processing pipeline as defined (*34*). Briefly, reads with identical cellBC and UMI were collapsed into a single, error-corrected consensus sequence representing a single expressed transcript. Consensus sequences with poor quality or sequencing coverage were removed, according to pipeline-defined thresholds. Each consensus sequence was aligned to the wild-type reference Target Site sequence, and the intBC and indel alleles were called from the alignment. These data are summarized in a molecule table which records the cellBC, UMI, intBC, indel allele, read depth, and other relevant information.

Calling clonal populations. Collected cells were grouped into "clonal populations", defined as populations of cells which descended from a single engineered clone and which therefore shared the same set of intBCs. This was accomplished using an iterative strategy of defining *de novo* sets of intBCs and assigning cells to clonal populations based on the sets, and repeating this process to further refine the assignments, similar to the strategy described in (*34*), as follows: (1) The most frequently observed intBC ("top intBC") was identified from the molecule table. (2) A "clustered intBC set" was defined as the intBCs present in >20% of cells containing the top intBC. (3) Cells with >25% of their Target Site UMIs pertaining to the clustered intBC set were then collected into a "cluster" and removed from the molecule table. (4) This process of identifying the top intBC, clustered intBC set, and cell clusters was iterated until at least one of two stopping criteria was met: (i) the returned cluster of cells was empty, indicating that the clustering strategy had exhausted well defined clusters, or (ii) the total fraction of clustered cells exceeded 98% of all cells in the molecule table. (5) Next, a "clonal population intBC set" was defined for each cell cluster, defined as intBCs that were present in >20% of cells in the cluster. Some clusters were manually redefined based on overlapping clonal population intBC sets likely resulting from the serial transduction strategy of cell line engineering; i.e., some clusters were expected to share intBCs because they were clonally related during serial integration of the Target Site (see **Fig. S20B** for an example). (6) Finally, cells from the initial molecule table were assigned to a clonal population if >60% of its

UMIs pertained to a single clonal population intBC set. Clonal populations with fewer than 25 cells were not considered, as lineage tracing over small cell numbers is less informative.

In the pre-implantation sample, we assigned cells to clonal populations observed in M5k by evaluating the proportion of TargetSite UMIs that corresponded to a clonal population intBC set. As above, we summed together the UMI proportions of each clonal population intBC set for each cell, thus yielding a similarity score that we could use to assign cells to clones. Here, we assigned a cell to a clone if at least 75% of its UMIs pertained to that clonal population intBC set.

Filtering cells and assembly of allele tables. Cells with poor Target Site capture (defined as cellBCs with <10 Target Site UMIs) were removed due to poor representation. Cells with ambiguous allele states were removed if >10% of the UMIs had conflicting allele states (i.e., intBC-distinguished Target Sites with more than one allele state), likely resulting from PCR errors or cell doublets (wherein both cells are in the same clonal population; so-called "intra-doublets"). Cells with ambiguous clonal population assignment were removed if <60% of its UMIs pertained to a single clonal population's set of intBCs; this likely results from cell-free Target Site transcripts, PCR errors, or cell doublets (wherein cells are from different clonal populations; so-called "inter-doublets"). Finally, the filtered molecule table with clonal population assignments is collapsed into an allele table, which summarizes the cellBC, intBC, indel allele, number of UMIs, assigned clonal population, and other relevant information.

Calculation of Tissue Dispersal Score. The Tissue Dispersal Score is the inverted Cramér's V (1 - V), a statistical measure of the association between two variables derived from the chi-squared test, ranging from 0 (no deviation from the background) to 1 (complete deviation from the background). For a given clonal population, we first perform a chi-squared test by forming a contingency table X over summarizing the number of cells found in each tissue for the clonal population, and the number of cells found in each tissue aggregated across all other clonal populations (referred to as the "background"). Importantly, the number of cells found in each tissue in the background are scaled such that the sums for both columns in X are equal. After performing a chi-squared test on the $r \times k$ contingency table, X, with a total of N counts across both columns, we derive the bias-corrected Cramér's V test (*93*) statistic:

$$V = \sqrt{\frac{\phi'}{min\,(\hat{k}-1, \hat{r}-1)}}$$

where $\phi' = max(0, \frac{\phi}{N} - \frac{(k-1)(r-1)}{N-1})$, $\hat{k} = \frac{(k-1)^2}{N-1}$, and $\hat{r} = \frac{(r-1)^2}{N-1}$. Finally, this statistic is inverted to obtain the Tissue Dispersal Score.

Filtering of clonal populations for tree reconstruction. We removed a minority of clonal populations that exhibited suboptimal lineage tracing parameters, as defined by <15% of cut-sites bearing indels (i.e., an estimate of the lineage recording kinetics) and <66.7% of unique cell allele states (i.e., an estimate of the lineage diversity and information content). Phylogenetic reconstruction in such poor parameter regimes resulted in low information trees (indicated by the shallow tree depth of filtered trees in **Figs. S6E** and **S7C**) that would not be interpretable for sensitive downstream analysis, such as calculating the inferred rate of metastasis.

Assembly of character matrices for phylogenetic tree reconstruction. To reconstruct lineages, we created "character matrices" from the allele tables of each clonal population using the Cassiopeia software. Specifically, we summarized the indels observed in each of the N cells in a clonal population across the M cut-sites to form an NxM matrix where each entry is an integer-representation of the indel observed in that cell at that cut-site. These entries are referred to as "character-states" or "states"; the columns of this matrix are referred to as "characters"; and the rows are referred to as "cells" or "samples", interchangeably. Missing data was specified as the string '-', and uncut sites were specified as '0'.

Cassiopeia-build pipeline overview. We reconstructed each clonal population's phylogeny using the Cassiopeia-Hybrid module, as described in (*34*). Briefly, Cassiopeia-Hybrid takes as input the *N*x*M* character matrix and first uses the Cassiopeia-Greedy heuristic to split cells into small groups based on the presence, or absence, of states that occurred early in the phylogeny (referred to as "character-splits"); then, each subproblem is solved precisely with the Steiner-Tree-based Cassiopeia-ILP module; finally, the completed subproblems are merged together to form the final tree. All clonal populations were reconstructed with a maximum neighborhood size of 10,000 (--max_neighborhood_size 10000) and a maximum time to convergence of 3.5 hr (--time_limit 12600).

While missing data is handled in Cassiopeia-ILP by considering all possible assignments, Cassiopeia-Greedy handled missing data with two heuristics: first, cells with missing data in a character-split were classified based on their 10 closest neighbors by a modified Hamming distance normalized by the number of overlapping characters two cells shared (using the "knn" greedy missing data mode in Cassiopeia: --greedy_missing_data_mode knn --num_neighbors 10); second, characters with greater than 30% missing data were not selected as character-splits (i.e., --greedy_max_missing_rep 0.3).

We determined prior-probabilities for each indel by using the proportion of times it appeared, independently, on an intBC in a clonal population (accounting for all clonal populations in our data). These priors were used in two ways: first, they were used to select indels as character-splits that likely occurred early in the phylogeny (i.e. they appeared in several cells, but had a relatively low prior) for Cassiopeia-Greedy; second, we used the priors to weight edges in the Steiner-Tree optimization within Cassiopeia-ILP (invoked with the --weighted_ilp argument to Cassiopeia's command line interface). Priors were provided to Cassiopeia's reconstruction algorithm using the --mutation_map argument.

To transition from Cassiopeia-Greedy to -ILP, we introduced a modified criteria based on the maximum distance from a cell population's Latest Common Ancestor (LCA) to any given cell in the population. Specifically, a new Cassiopeia-ILP subproblem is spawned if the distance to an LCA of a group of cells is below a user-defined threshold (--cutoff <lca_cutoff>). An appropriate LCA distance threshold was determined, iteratively, such that no Cassiopeia-ILP subproblem exceeded the user-defined maximum neighborhood size. The LCA transitioning criteria was invoked with --hybrid_lca_mode.

Neighbor-Joining reconstructions. For neighbor-joining reconstructions, we used the --neighbor_joining_weighted reconstruction option in the Cassiopeia package, which uses a scikit-bio implementation of neighbor-joining (version 0.5.5). To define distances between cells as input, we utilized a version of the modified Hamming distance function that incorporated prior-probabilities (prior-probabilities are passed in via the --mutation_map argument). Specifically, we defined $d'(a, b)$ as the sum of all the pairwise $h'_p(a_i, b_i)$ values across the cut-sites in a clonal population:

$$h'_p(a_i, b_i) = \begin{cases} -log(p(a_i)) - log(p(b_i)) & \text{if } a_i \neq b_i \text{ and } a_i, b_i \text{ are mutated} \\ -log(p(a_i)) & \text{if } a_i \text{ mutated, } b_i \text{unmutated} \\ -log(p(b_i)) & \text{if } b_i \text{ mutated, } a_i \text{unmutated} \\ log(p(a_i)) + log(p(b_i)) & \text{if } a_i, b_i \text{ mutated and } a_i == b_i \\ 0 & o.w. \end{cases}$$

Finally, distances $d'(a, b)$ were normalized by the number of cut-sites overlapping between cells a and b.

Allelic vs. phylogenetic distance calculations. The concordance between allelic and phylogenetic distances were used to assess the agreement between a tree's phylogenetic structure and the observed mutations. To quantify allelic distances, we defined a modified Hamming distance between cells *a* and *b*,

$$d(a,b) = \sum_{i \in M} h'(a_i, b_i)$$

Where M is the number of cut-sites in the clonal population that cells a and b come from, a_i and b_i are the states observed at the i^{th} character, and h' is defined as follows:

$$h'(a_i, b_i) = \begin{cases} 2 & \text{if } a_i \neq b_i \text{ and } a_i \neq 0 \text{ and } b_i \neq 0 \\ 1 & \text{if } (a_i == 0 \text{ or } b_i == 0) \text{ and } a_i \neq b_i \\ 0 & \text{o.w.} \end{cases}$$

Importantly, these values were normalized by $2 * M$, the maximum distance for a pair of cells.

The phylogenetic distance was calculated for all pairs of cells as the number of non-zero length edges that separated the two cells in the tree. The phylogenetic distances were normalized by the "diameter" of the phylogeny, i.e., the maximum distance between any pair of cells.

Derivation of the AlleleMetRate. The AlleleMetRate is provided as a "tree-agnostic" measurement of a clonal population's intrinsic metastatic potential. Intuitively, the rate measures the proportion of cells that do not reside in the same tissue as their closest relative (as determined by the modified Hamming distance between two cells' character-states). Importantly, if a cell has more than one closest relative, each of their votes are normalized by the number of relatives this cell has. More formally, the contribution a cell has to the AlleleMetRate is

$$m(c) = \frac{1}{K} \sum_{i \in Neighbors(c)} I(tissue(i) \neq tissue(c))$$

Where K is the number of closest relatives a cell has, and *I(*)* is an indicator function that equals 1 if the tissue of cell *i* is different from the tissue of cell *c*. The AlleleMetRate reported is

$$M_a = \frac{1}{|L|} \sum_{l \in L} m(l)$$

Where L is the set of all leaves in the tree.

Derivation of the TreeMetRate. The TreeMetRate is derived from the Fitch-Hartigan maximum parsimony algorithm (Fitch, 1971; Hartigan, 1973). Briefly, the Fitch-Hartigan algorithm takes in a rooted tree with labelings at the leaves (in our case labels indicate which tissue the cell was obtained from) and in our case reports two items: (a) a distribution of labels at the internal nodes that minimize the number of transitions between tissues (i.e., achieves maximum parsimony); and (b) the minimum number of transitions between tissues (referred to as the "parsimony score"). The Fitch-Hartigan algorithm is an efficient (scaling with $n \cdot k$ where n is the number of cells and k is the number of tissues) and common algorithm for solving the "Small Parsimony Problem", as opposed to the "Large Parsimony Problem" which attempts to find the *tree* of maximum parsimony.

The Fitch-Hartigan algorithm operates in two phases: first, by propagating up from the leaves the set of possible labelings at each internal node; second, by traversing in depth-first order from the root of the phylogeny and selecting one state for each internal node from its set of optimal labels. Conveniently, the parsimony score can be obtained from this second phase. To finally report the TreeMetRate, we simply normalize this parsimony score by the number of edges in the tree.

Derivation of the single-cell MetRate. The scMetRate for a given cell is defined as the average of all TreeMetRates for the clades that contain that cell. Specifically, we employ the following algorithm:

```
compute_scMetRate(tree, x):
     rate = 0
     num_ancestors = 0
     for n in depth_first_search(tree, x):
          rate += compute_tree_metrate(n)
          num_ancestors += 1

     rate /= num_ancestors
     return rate
```

Where *compute_tree_metrate(n)* is a function that computes the TreeMetRate for the subtree rooted at an internal node *n* (as described above).

scRNA-seq preprocessing and normalization. RNA-seq libraries were quantified using CellRanger version 2.1.1 with the GRCh38 genome build for M5k and M30k; and CellRanger version 2.0.0 with the hg19 genome build for M10k and M100k. Cells not found in the Target Site Library were filtered out. All RNA-seq datasets were normalized identically – we first normalized the UMI counts in every cell to the median number of UMIs found in each library (referred to hereafter as "UMI-normalized" counts) and then each matrix of UMI-normalized counts was log-transformed (after adding a 1-pseudocount to preserve 0's in the data; hereafter referred to as "log-transformed" counts). In the analysis presented in Figure 5A we applied scVI (version 0.6.5) to the raw counts of the M5k and the pre-implantation cells to obtain a shared, batch-corrected latent space (learning rate of 1e-3, 400 epochs, 10 latent dimensions). This space was used for downstream projections and *Vision* analyses.

Vision analysis on M5k. *Vision* (version 2.1) was applied to the UMI-normalized counts for the 35,006 filtered and clonal population-assigned cells that overlapped between M5k's scRNA-seq and TargetSite libraries. Genes were filtered out using the Fano-filter procedure in *Vision*, leaving 4,005 genes. To provide a latent space to *Vision,* we computed a reduced dimension embedding using scVI (*94*) on the filtered gene matrix of raw counts, using a learning rate of 1e-3 and 40 epochs to produce 10 latent components. No batch correction was used. Signatures were obtained from MSigDB (*95*).

Differential expression of left lung samples. We used the UMI-normalized counts from M5k to perform differential expression between cells in the left lung, stratified by whether or not they were from a clonal population that metastasized from the left lung at any point in the experiment. All cells found in the left lung that were from clonal populations that metastasized are labeled as "Metastatic". We performed 5 differential expression tests for the cells in each group: CP029 vs. "Metastatic"; CP036 vs. "Metastatic"; CP078 vs. "Metastatic"; CP094 vs. "Metastatic"; and "Metastatic" vs. all cells in CP029, CP036, CP078, and CP094. Tests were performed using the Wilcoxon rank sums test as implemented in Scanpy (*96*).

Differential expression of single-cell MetRate by Poisson regression. Genes differentially expressed between highly and lowly metastatic cells were determined by using a Poisson regression scheme (as implemented in the

GLM package in *Julia*, version 1.3.7). Cells were first segmented into "Low" and "High" groups (referred to as *m*) based on the scMetRate. Then, the following model was employed on the UMI-normalized counts:

$$H_o : log(1 + expr) \sim 1 + SizeFactor$$
$$H_a : log(1 + expr) \sim 1 + SizeFactor + m$$

Where *SizeFactor* was defined as the number of genes detected in that cell and *expr* refers to the UMI-normalized expression count of a particular gene. A Likelihood Ratio Test (LRT) was used to determine the significance of the alternative hypothesis (in particular, the significance of the model fit improvement due to the variable *m*). Log2 fold-changes (Log2FC's) were calculated by comparing the mean expression of a gene in the "High" group to that of the "Low" group of the UMI-normalized counts with a pseudocount of 0.01 added to account for zero counts. P-values were adjusted using the Benjamini-Hochberg false discovery rate procedure. Genes were considered significant if their FDR-adjusted p-value was less than 0.01.

Cells with greater than 20% counts attributed to the mitochondrial genome were filtered out and genes observed in fewer than 10% of cells were filtered. We proceeded with 35,005 cells and 4,993 genes for M5k; 1,492 cells and 11,171 genes for M10k; 17,175 cells and 7,620 genes for M30k; and 2,105 cells and 10,371 genes for M100k. Analysis was performed similarly for the M10k, M30k, and M100k mice, with considerations of their underlying scMetRate distributions taken into account. To ensure an accurate reflection of High and Low metastatic groups, cells were stratified according to the median in M10k, the 25th percentile in M100k; and in M30k because the scMetRate was not bimodal, we stratified cells into groups below the 25th and above the 75th percentiles.

Positive and negative gene "hits" in the M5k analysis were determined using a "discriminant" score defined as the absolute value of log2(fold-change) times the negative log10(FDR) of the gene. Genes with a discriminant score greater than 600 were annotated as "hits". For the analysis in Fig. 4C,D we identified significant gene sets as those with an FDR-corrected p-value < 0.01 and log2FC > 0 or log2FC < 0 for positive and negative sets, respectively. Significance of gene set overlap was assessed with the R package *SuperExactTest (87)*, version 1.0.7) Finally, genes were considered "reproducible" in the M10k, M30k, and M100k analyses if they were found to be significant (FDR < 0.01) and their effect was in the same direction as in M5k.

A consensus "Metastatic Signature" was formed from the top genes in the positive and negative direction from each mouse. Specifically, we first ranked each list by the discriminant score described above. Then, for each direction in each mouse as measured by log2FC, we selected the top 30 genes. We then took the unique genes in each direction across all mice and formed a directional signature from these two lists. Scores for this signature were obtained for each cell with *Vision*.

Identifying gene signatures correlated with TreeMetRates. Gene signature scores for single cells were calculated with *Vision*, and pseudo-bulked by clonal population by taking the mean signature score for each signature across the cells in a clonal population. We then calculated the Spearman correlation using the scipy Python package (version 1.2.2) between the pseudo-bulked gene signature score and the TreeMetRate for a given clonal population.

Hotspot *analysis for CP007.* *Hotspot* (version 0.9.0) was acquired from Github (https://github.com/YosefLab/Hotspot) and run on the UMI-normalized counts for the 603 cells in CP007 and all genes expressed in at least 10% of cells.We used the UMI-adjusted negative binomial model ("danb") as the background gene expression model and distances between cells were computed from the CP007 tree reconstructed with Cassiopeia. We used 40 neighbors for the *Hotspot* analysis. Modules of at least 120 genes were identified from all genes with an FDR < 0.1, and the *Hotspot* module scores were used to annotate the tree.

The right lung (E) only ("RE-only") control was performed identically, except for first removing cells that were not from the RE tissue sample in CP007 (437 cells).

***Inference of Tissue Transition Matrices with* FitchCount.** To infer the relative propensities of a clonal population's cells to transition between any two tissues, we developed an efficient algorithm for aggregating together optimal solutions proposed by the Fitch-Hartigan algorithm (see **Supplementary Text** for algorithm and proof). Briefly, *FitchCount* is a dynamic programming algorithm that operates on a rooted phylogeny. It begins by performing the "bottom-up" phase of the Fitch-Hartigan algorithm to obtain a distribution of optimal labels over each internal node and then employs a computationally efficient algorithm for computing the number of times a given transition occurs in all optimal solutions proposed by the Fitch-Hartigan algorithm. Finally, this algorithm returns a square matrix, *M*, where each value $m_{i,j}$ indicates the number of times a transition from tissue *i* to tissue *j* was observed in the tree over all optimal solutions.

In **Fig. 6** and **Fig. S22**, we report transition matrices for the probability of a cell metastasizing from one tissue to another, given that the cell metastasizes; i.e., $P(m_{i,j} \mid i \mathrel{!}= j)$. To obtain these conditional probability tables, *P*, we first set *diag(M)* to 0 (indicating that the probability of self-transition is 0) and re-normalize each row to sum to 1.

Feature selection and Principal Component Analysis (PCA) of tissue transition matrices. To identify the trends in the tissue transition matrices presented in **Fig. 6**, we performed dimensionality reduction using Principal Component Analysis (PCA) on the flattened tissue transition matrices. Beyond all the conditional probability of transitioning between tissue samples summarized in the tissue transition matrix *M*, we included additional features which we hypothesized would aggregate important signals. In particular, the following were added by deriving statistics from *P*, the conditional probability matrix, and *M*, the unnormalized tissue transition matrix (recall that we observed 6 tissue samples in M5k: left lung [LL], right lung W [RW], right lung E [RE], mediastinum 1 & 2 [M1 & M2] and liver [Liv]):

- The Left Lung reseeding rate (defined as the sum of the probabilities in the LL-column of *P)*
- The Mediastinum tropism rate (defined as the sum of the probabilities in the M1 or M2 columns of *P*)
- The Liver tropism rate (defined as the sum of the probabilities in the Liver column of *P*)
- The Right Lung tropism rate (defined as the sum of the probabilities in the Right Lung column of *P*)
- The Left lung to Right Lung seeding rate (defined as the sum of the probabilities $p_{LL,\,RW}$ and $p_{LL,\,RE}$)
- The Right Lung to Left Lung reseeding rate (defined as the sum of the probabilities $p_{RW,\,LL}$ and $p_{RW,\,LL}$)
- The "intra lung" metastatic rate (defined as the sum of the probabilities leading from the LL to either RL sample and vice versa in *P*)
- The "intra mediastinum" metastatic rate (defined as the sum of probabilities going between *M1* and *M2* samples in *P*)
- The Primary Seeding density (defined as the density of transitions observed in the LL row of *T*).
- The M1 seeding density (defined as the density of transitions in the M1 row of *T*).
- The M2 seeding density (defined as the density of transitions in the M2 row of *T*).
- The Mediastinum seeding density (defined as the sum of densities of transitions in the M1 and M2 rows of *T*).

This resulted in a set of 48 features used for dimensionality reduction. To perform PCA on this matrix, we concatenated the 48 features for each clonal population, standardized the features, and used PCA as implemented in the scikit-learn Python package (version 0.21.3).

Classification of Seeding Topologies. To detect seeding topologies, as presented in **Fig. 5J** and **K**, we devised simple algorithms from the tree structure and the conditional probability tissue transition matrices, *P*.

- To identify clonal populations that exhibited primary seeding, we evaluated if there existed some $p_{LL,x} > 0$ for some $x \in \{RW, RE, M1, M2, Liv\}$.
- To identify clonal populations that experienced reseeding to the left lung, we evaluated if there existed some $p_{x,LL} > 0$ for some $x \in \{RW, RE, M1, M2, Liv\}$.
- To identify clonal populations that exhibited bidirectional seeding (i.e. seeding to any tissue that previously served as a source for a metastatic event), we evaluate whether or not there exist two metastatic events, the second of which returns to the source of the first metastatic event. To do so, we sampled 100 solutions from the Fitch-Hartigan top-down procedure (which assigns labels to internal nodes) and evaluated whether or not we observe a path from the root to any leaf that follows a bidirectional seeding pattern (not necessarily on consecutive edges).
- To identify clonal populations that exhibited a seeding cascade, we evaluated whether or not there existed any tissue transition in a clonal population from $s_i \rightarrow s_j$ such that $s_i \neq s_j \neq LL.$ Even if we do not observe any cells in the left lung of a given tumor, we know that the pattern previously described is evidence of a seeding cascade because all tumors began in the left lung of the mouse.
- To identify clonal population that exhibited parallel seeding, we evaluated whether or not there existed two conditional probabilities in P, p_{x,y_1} and p_{x,y_2}, that were non-zero from any tissue x to any pair of tissues y_1 and y_2 such that $y_1 \neq y_2$.

Model framework, parameters, and assumptions of metastasis simulator. To simulate metastatic events, we built on the simulation framework present in Cassiopeia as previously described in (*34*). The framework simulates a series of D binary splits over cells with M characters & S states per character; mutations are introduced every generation for each cell according to a per-character mutation rate, m, and dropout is simulated at the end according to the dropout rate, d.

To simulate metastatic processes, we introduced three new parameters: a cell-division rate (α), a metastatic rate (μ), and a probability map of transitioning between tissues should a metastatic event occur (*P)* that follows the same structure as the conditional tissue transition matrices discussed above. We assume that every generation, a cell first has the opportunity to divide (by evaluating whether $Unif(0, 1) < \alpha$) and then if so, each of its offspring has the opportunity to metastasize (by evaluating $Unif(1) < \mu$). If the daughter cell metastasizes, a tissue is chosen randomly according to the probabilities specified in P. To note, in this new simulation framework, we choose to control the size of a clonal population randomly with α, instead of subsampling at a predetermined rate as described previously.

Importantly, we can also apply this simulation framework on top of a tree by simply traversing the edges of the tree and simulating metastasis without cell divisions. For the results presented in this study, we used the M5k mouse to parameterize the simulations - especially in regards to the number of tissues that cells can metastasize to (6) and the distribution of metastatic rates (between 0 and 0.3).

Assessing accuracy of the TreeMetRate. We used the simulation framework to evaluate how well the Tissue Dispersion Score, AlleleMetRate, and TreeMetRate measurements were able to capture the underlying metastatic rate μ of a simulation. To do, we simulated trees of variable depth D and parameterized with some doubling rate, $\alpha \sim Unif(0, 0.7)$, and metastatic rate, $\mu \sim Unif(0, 0.3)$. We simulated 1000 such trees, with $D \in \{10, 14\}$ (90% of trees were simulated with $D = 10$, and 10% of trees were simulated with $D = 14$, to roughly reflect the distribution of trees we see in our data). Target Sites were simulated using empirically relevant parameters: 40 characters, 40 states, a dropout rate of 18%, and a mutation rate of 2.5%. For the purposes of estimating the TreeMetRate, P contained uniform probabilities of transitioning to any other tissue from a given tissue.

For each tree, the AlleleMetRate and TreeMetRate were calculated directly from the cells or tree; the Tissue Dispersion Score was calculated with a background derived from the tissue distribution across the 1,000 trees. Performance was assessed by the agreement between a specific score and the underlying rate of metastasis for that tree, as in **Fig. S9A-D**.

Bootstrapping analysis of TreeMetRate stability. Bootstrapping analysis was performed on a set of 10 simulated trees from (*34*). For each simulated tree, we sampled with replacement the cut-sites of the character matrices 100 times, resulting in 100 "bootstrapped character matrices" for each tree. We reconstructed each of these bootstrapped character matrices with Cassiopeia-ILP (maximum_neighborhood_size = 10,000, time_limit = 12,600). This left us with 1,000 reconstructions where each reconstruction corresponded to one of 10 simulated trees.

Then, for each simulated tree, we overlaid 50 metastatic processes on to the tree (here, because the tree was already defined, it was not necessary to use α). Each time, we transferred the labels to each of the 100 reconstructions for that simulated tree and evaluated the mean TreeMetRate, as well as the standard error, defined as:

$$\hat{mean}(TMR) = \frac{1}{B}\sum compute\_tree\_metrate(tree)$$

$$\hat{se}(TMR) = \frac{1}{B}\sum(compute\_tree\_metrate(tree) - \hat{mean}(TMR))$$

where B is the number of bootstrapped samples for a given metastatic process (here, $B = 100$). The coefficient variation for each metastatic process was defined as the ratio $\hat{se}/\hat{mean}$.

Assessing accuracy of the tissue transition matrices. We evaluated the accuracy of conditional tissue transition matrix inference by utilizing a similar approach above: we simulated trees ($\alpha \sim Unif(0, 0.7)$, $\mu \sim Unif(0, 0.3)$) with $D \in \{10, 14\}$ that allowed cells to move between 6 tissues (these probabilities were specified in the conditional probability matrix *P*). For each simulation, we evaluated how well a particular statistic was able to capture the underlying *conditional* tissue transition matrix *P'* (i.e. the non-diagonal probabilities, corresponding to the probability that a cell will move to some tissue, given it is found in a particular tissue and is metastasizing) calculated from the ground-truth internal labels.

Conditional tissue transition matrices were simulated according to two models: one where the rates of metastasizing to any other tissue were uniform (hereafter referred to as the "Uniform" model) and one where rates could be biased to some tissues, i.e., experience various tissue tropisms (hereafter referred to as the "Biased" model). To model both scenarios, we sampled each tissue's conditional transition probabilities from a Dirichlet distribution, parameterized with a length 5 array $\bar{\alpha}$ (recall that we are simulating conditional tissue transitions, and are not concerned with the probability that a cell remains in place) that essentially provides the "density" towards any outcome. To simulate the "Uniform" model, $\bar{\alpha}$ was parameterized as [50, 50, 50, 50, 50] - yielding roughly uniform probabilities across the 5 tissues. To simulate the "Biased" model, we used a "flat Dirichlet distribution", corresponding to $\bar{\alpha}$ parameterized as [1, 1, 1, 1, 1], which yielded potentially very biased probability distributions. We simulated 500 trees per model.

We used three algorithms for inferring the conditional tissue transition matrix *P* for a given tree:

1. *FitchCount* (as described above, and in the **Supplementary Materials**).
2. Naive-Fitch (in which the internal nodes were assigned from a single solution drawn from the Fitch-Hartigan top-down algorithm and used to infer *P'*).
3. Majority-Vote (in which the internal nodes were assigned the label that appeared at the greatest frequency at the leaves below the node and these assignments were used to infer *P'*).

In both the Naive-Fitch and Majority-Vote case, after internal nodes were assigned a label we performed a depth-first traversal on the tree and counted the number of times a parent transitioned to a child. Ground-truth conditional transition matrices were found similarly, using the simulated ground-truth labels for internal nodes.

To evaluate accuracy, we computed the Spearman correlation (using Python's scipy library, version 1.2.2) between the flattened matrix of the ground-truth conditional transition matrix and the transition matrices inferred by one of the three algorithms described above (*FitchCount,* Naive-Fitch, and Majority-Vote).

Supplementary Figures and Legends: (embedded below)

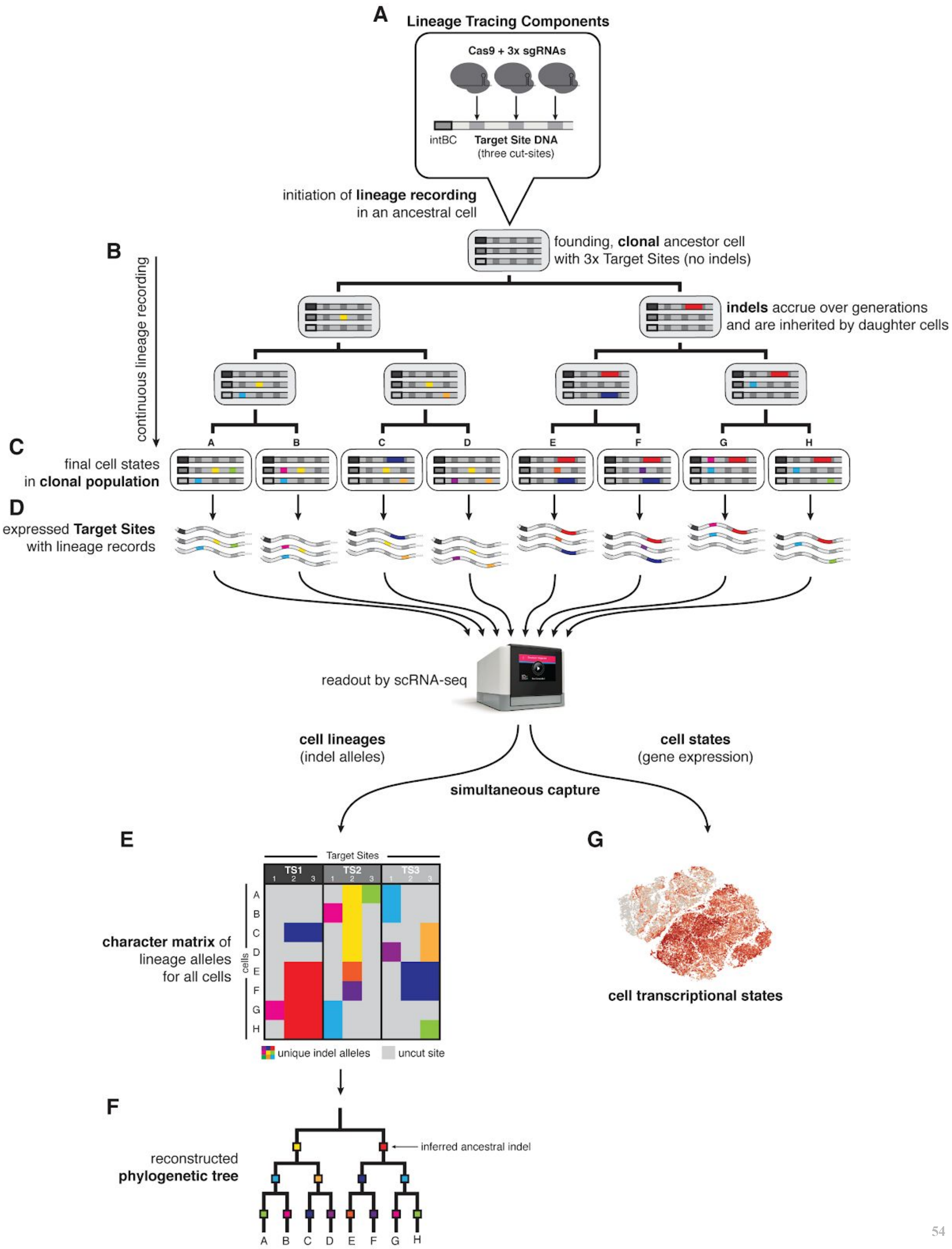

A
Lineage Tracing Components
Cas9 + 3x sgRNAs
intBC
Target Site DNA
(three cut-sites)
initiation of lineage recording
in an ancestral cell
B
continuous lineage recording
founding, clonal ancestor cell
with 3x Target Sites (no indels)
indels accrue over generations
and are inherited by daughter cells
C
final cell states
in clonal population
A
B
C
D
E
F
G
H
D
expressed Target Sites
with lineage records
readout by scRNA-seq
cell lineages
(indel alleles)
cell states
(gene expression)
simultaneous capture
E
Target Sites
TS1
TS2
TS3
1
2
3
1
2
3
1
2
3
character matrix of
lineage alleles
for all cells
cells
A
B
C
D
E
F
G
H
unique indel alleles
uncut site
F
reconstructed
phylogenetic tree
inferred ancestral indel
A
B
C
D
E
F
G
H
G
cell transcriptional states

Fig. S1. Detailed schematic of lineage tracing methodology. (**A**) Cells are genetically engineered with the lineage tracing components: (i) Cas9, (ii) multiple copies of the Target Site, and finally (iii) three sgRNAs that are complementary to three cut-sites on each Target Site. Multiple copies of the Target Site per cell are distinguished by unique integration barcodes (intBCs). Cas9-induced double-stranded breaks at cut-sites are repaired with high-diversity insertions or deletions (indels), which act as stable, heritable markers of cell lineage. (**B**) Upon initiation of lineage recording in a founding cell (i.e., one "clone"), indels continuously accrue on the Target Sites (colored boxes), which are inherited by descendents over subsequent generations. (**C**) At the end of the lineage recording experiment, the final population of descendent cells (i.e., one "clonal population") are collected and (**D**) their expressed Target Site mRNAs are captured by single-cell RNA-sequencing (e.g., by the 10X Genomics Chromium platform) alongside transcriptome-wide expressed genes. (**E**) The indel allele information is read from the Target Site sequences and summarized in a "character matrix" of indel allele states (values) for each cut-site (columns) in each cell (rows). (**F**) From the pattern of cells' shared and distinguishing indel alleles, a tree reconstruction algorithm builds a phylogenetic model that best captures cell–cell relationships (e.g., by maximizing parsimony), thus producing a detailed map of cell lineage. (**G**) Simultaneously, single-cell RNA-sequencing captures the transcriptional profiles of the observed cells, allowing for direct comparisons between cell lineage and cell state.

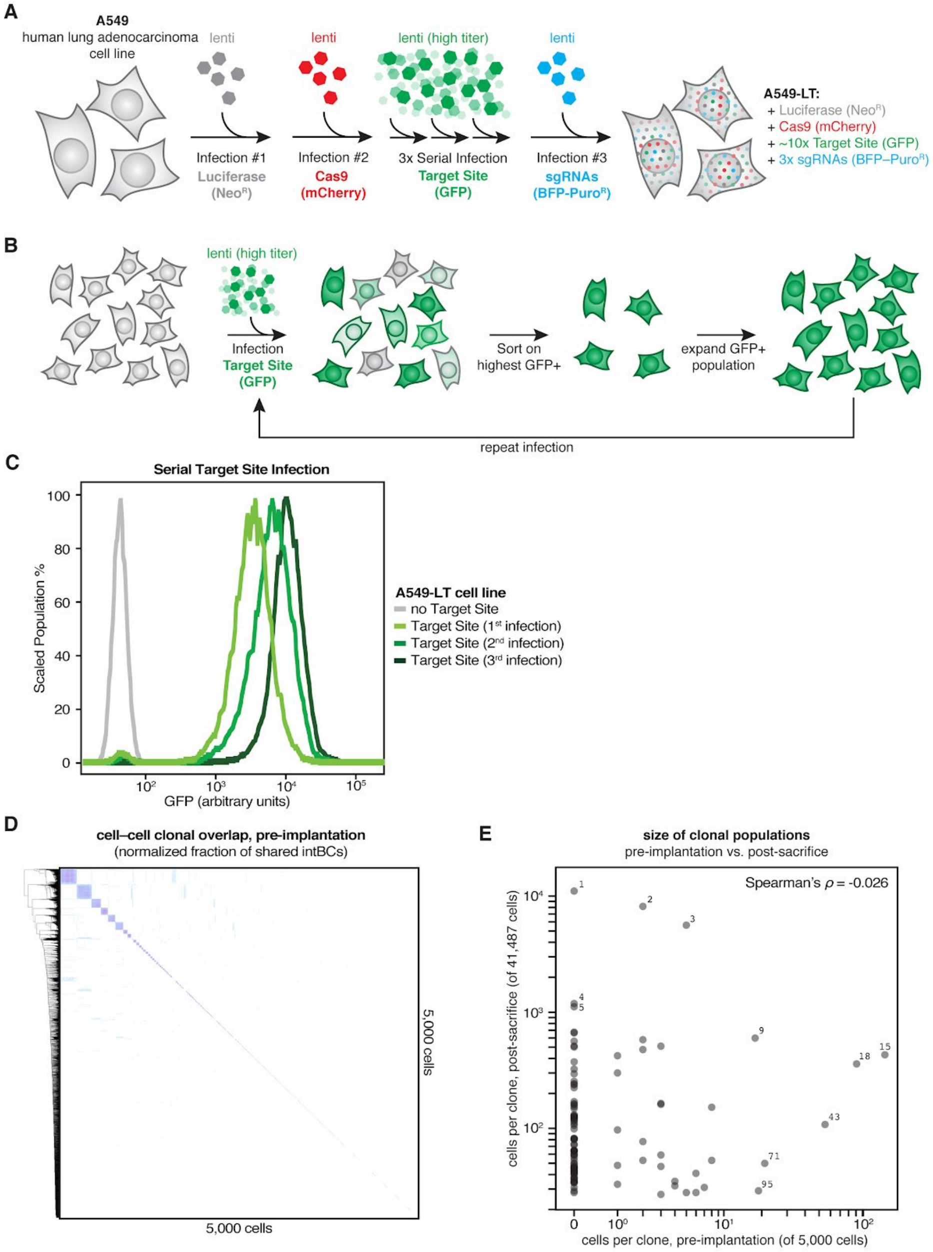
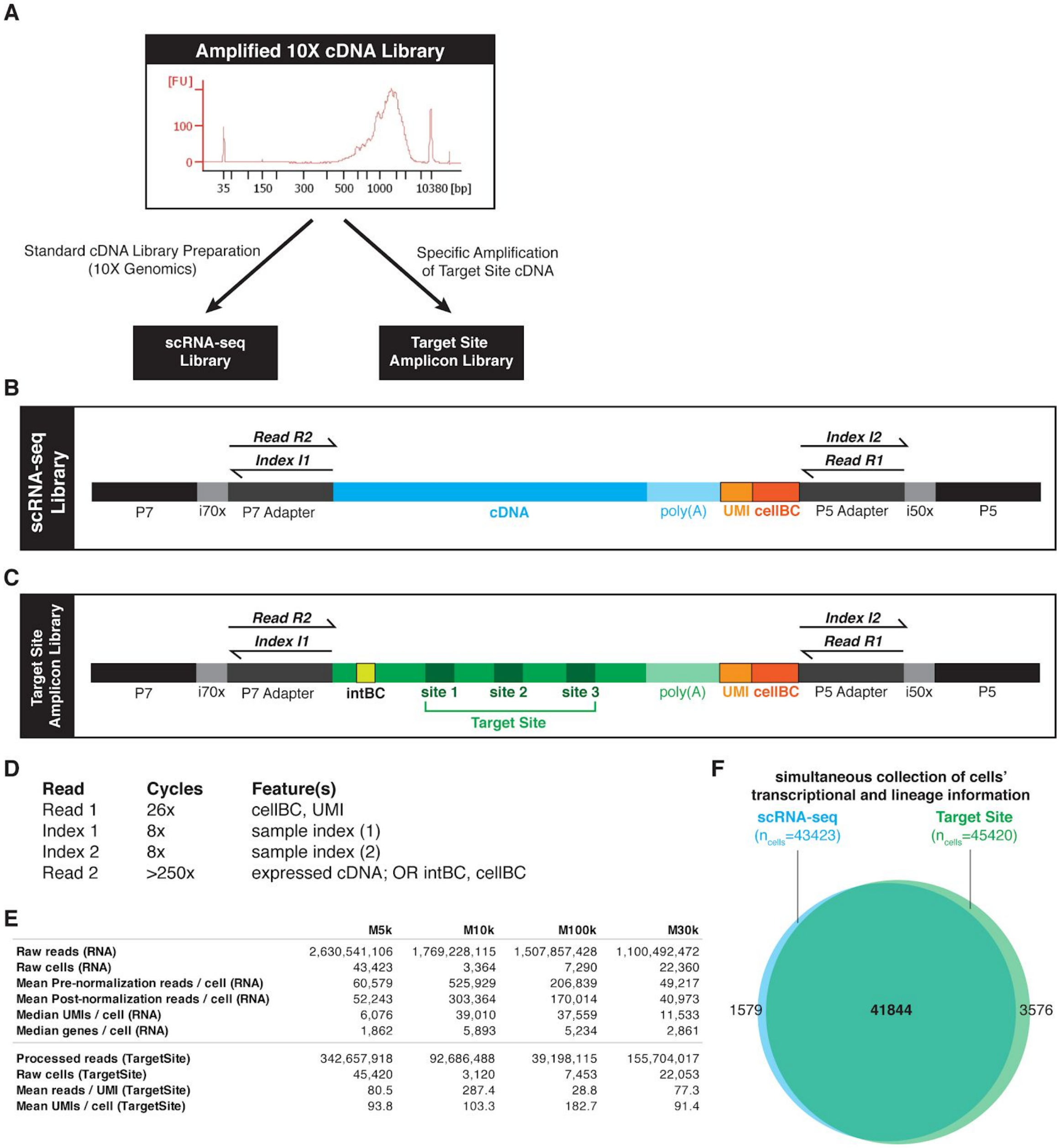


Fig. S2. Cell line engineering strategy and estimation of clonal diversity. (**A**) Human lung adenocarcinoma (A549) cells were genetically engineered with the lineage tracing components by lentiviral transduction with (1) Luciferase-Neomycin^R and antibiotic-selected, (2) Cas9-mCherry and fluorescence-sorted, (3) serial, high-titer TargetSite-GFP and fluorescence-sorted, and finally (4) triple-sgRNA BFP-Puromycin^R and fluorescence-sorted, thus producing lineage tracing-competent A549-LT cells. (**B**) Serial, high-titer TargetSite-GFP lentiviral transduction strategy to achieve high copy-number. (**C**) Cells with high-shifted GFP fluorescence after successive TargetSite-GFP infections, indicating increasing copy-number of the Target Site. (**D**) Sample of 5,000 A549-LT cells prior to injection to estimate initial clonal diversity. Shown is a heat-map of the fraction of shared intBCs in all cell–cell comparisons. On average, approximately 22,000 unique, high-quality intBCs were identified per random sample of 5,000 cells; thus, we estimate approximately 2,150 distinct clones per 5,000 cells (assuming 10.3 intBCs per clonal population; **Fig. S6A**) at the beginning of the experiment. (**E**) Comparison of the size of each clonal population observed in mouse M5k in a pre-implantation sample of 5,000 cells (*in vitro*) and post-sacrifice (*in vivo*). There is no correlation between the *in vitro* and *in vivo* population sizes (Spearman's ρ=-0.026).



D

Read	Cycles	Feature(s)
Read 1	26x	cellBC, UMI
Index 1	8x	sample index (1)
Index 2	8x	sample index (2)
Read 2	>250x	expressed cDNA; OR intBC, cellBC

E

	M5k	M10k	M100k	M30k
Raw reads (RNA)	2,630,541,106	1,769,228,115	1,507,857,428	1,100,492,472
Raw cells (RNA)	43,423	3,364	7,290	22,360
Mean Pre-normalization reads / cell (RNA)	60,579	525,929	206,839	49,217
Mean Post-normalization reads / cell (RNA)	52,243	303,364	170,014	40,973
Median UMIs / cell (RNA)	6,076	39,010	37,559	11,533
Median genes / cell (RNA)	1,862	5,893	5,234	2,861
Processed reads (TargetSite)	342,657,918	92,686,488	39,198,115	155,704,017
Raw cells (TargetSite)	45,420	3,120	7,453	22,053
Mean reads / UMI (TargetSite)	80.5	287.4	28.8	77.3
Mean UMIs / cell (TargetSite)	93.8	103.3	182.7	91.4

Fig. S3. Sequencing library construction and metrics. (**A**) The amplified cDNA (shown here as a BioAnalyzer trace) from the Chromium 3' Single Cell V2 kit (10X Genomics) serves as the template for both (**B**) the single-cell gene expression ("RNA") library and (**C**) the single-cell Target Site amplicon library (Methods). (**D**) The cellBC and unique molecular identifier (UMI) are sequenced from Read R1, the sample identities are sequenced from Indices I1 and I2, and the expressed cDNA or the Target Site amplicon (including the intBC and cut-sites 1, 2, and 3) are sequenced from Read R2. (**E**) Library sequencing metrics for TargetSite and RNA libraries for each mouse. (**F**) There is vast overlap in the cells identified from the gene expression and lineage sequencing datasets, as shown for mouse M5k.

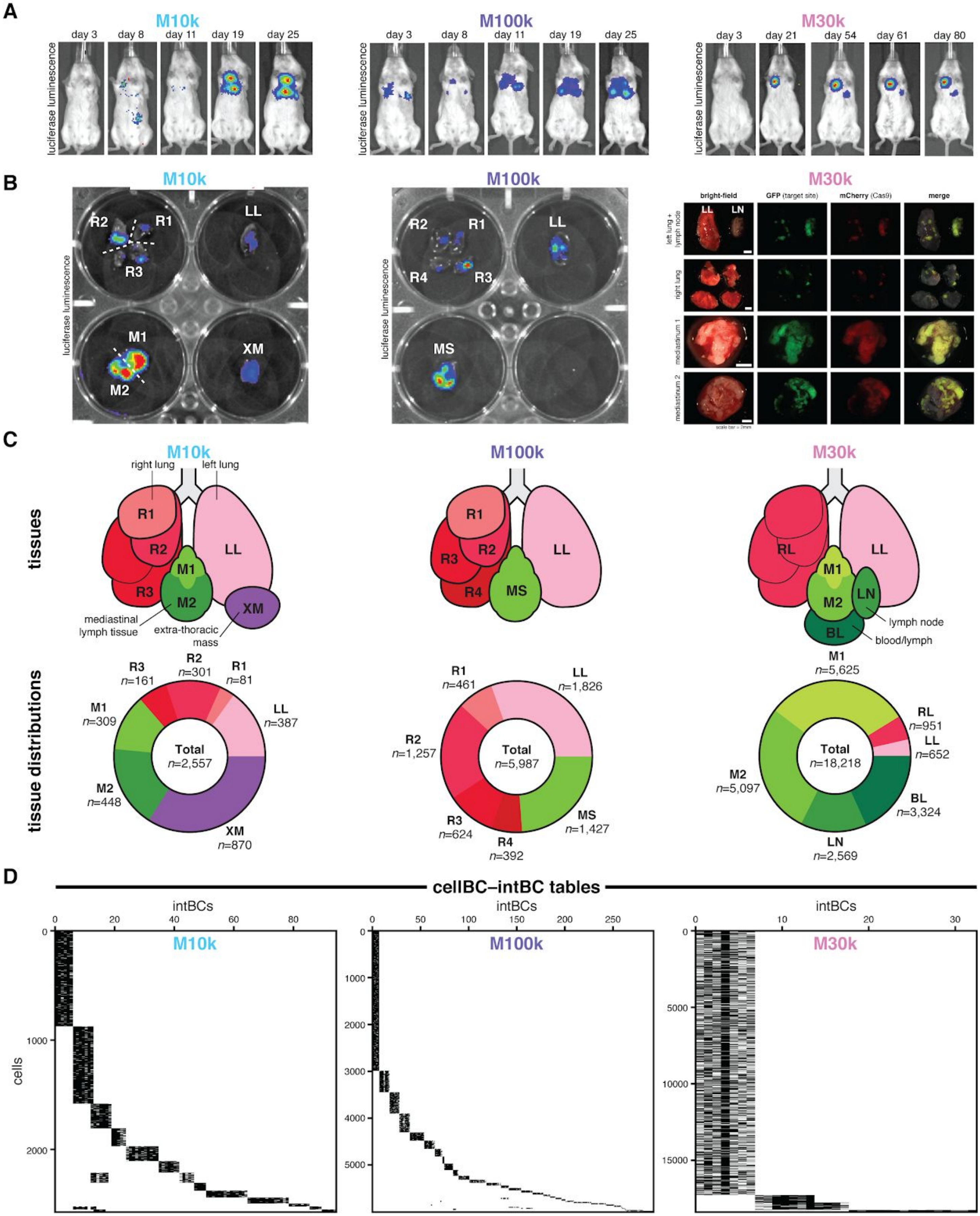

A
M10k
day 3
day 8
day 11
day 19
day 25
luciferase luminescence
M100k
day 3
day 8
day 11
day 19
day 25
luciferase luminescence
M30k
day 3
day 21
day 54
day 61
day 80
luciferase luminescence
B
M10k
luciferase luminescence
R2
R1
R3
LL
M1
M2
XM
M100k
luciferase luminescence
R2
R1
R4
R3
LL
MS
M30k
bright-field
GFP (target site)
mCherry (Cas9)
merge
left lung + lymph node
LL
LN
right lung
mediastinum 1
mediastinum 2
scale bar = 2mm
C
tissues
M10k
right lung
left lung
R1
R2
R3
LL
M1
M2
XM
mediastinal lymph tissue
extra-thoracic mass
M100k
R1
R2
R3
R4
LL
MS
M30k
RL
LL
M1
M2
LN
BL
lymph node
blood/lymph
tissue distributions
R3 n=161
R2 n=301
R1 n=81
LL n=387
M1 n=309
M2 n=448
XM n=870
Total n=2,557
R1 n=461
LL n=1,826
R2 n=1,257
R3 n=624
R4 n=392
MS n=1,427
Total n=5,987
M1 n=5,625
RL n=951
LL n=652
BL n=3,324
LN n=2,569
M2 n=5,097
Total n=18,218
D
cellBC–intBC tables
intBCs
0
20
40
60
80
M10k
cells
0
1000
2000
intBCs
0
50
100
150
200
250
M100k
0
1000
2000
3000
4000
5000
intBCs
0
10
20
30
M30k
0
5000
10000
15000

Fig. S4. Tracing the cell lineages of metastatic progression in three additional mice. In addition to mouse M5k discussed in throughout the main text, we also traced the lineages of metastatic dissemination in three additional mice orthotopically xenografted with 10,000 (M10k, left), 100,000 (M100k, middle), and 30,000 (M30k, right) A549-LT cells in two cohort experiments (first cohort, A549-LT1: M10k and M100k; second cohort, A549-LT2: M5k and M30k). (**A**) *In vivo* bioluminescence imaging of cancer cell engraftment and metastatic spread at indicated times post-implantation. The mice were sacrificed 53, 67, and 80 days post-implantation, respectively. (**B**) *Ex vivo* imaging of tumorous tissues by bioluminescence (left and middle) or fluorescence (right) imaging with the tissue samples indicated. In addition to extensive tumors in the lungs and mediastinum, M10k had one large solid tumor located ventral to the left lung and embedded in the ribcage (called here an "extra-thoracic mass"; XM). (**C**) Anatomical representation of collected tumorous tissues (top) and the number of cells collected for each tissue and each mouse (bottom). Notably, M30k had a large lymph node (LN) on the left lung, as well as diffuse bloody lymph (BL) in the thoracic cavity upon sacrificing. (**D**) CellBC-intBC tables showing clonal populations of cancer cells in each mouse, as in **Fig. S5A,B**. Some intBCs are shared between some clones (most notably in M10k) which likely resulted from cells that were clonally related at the stage of serial Target Site transduction during cell line engineering (Methods).

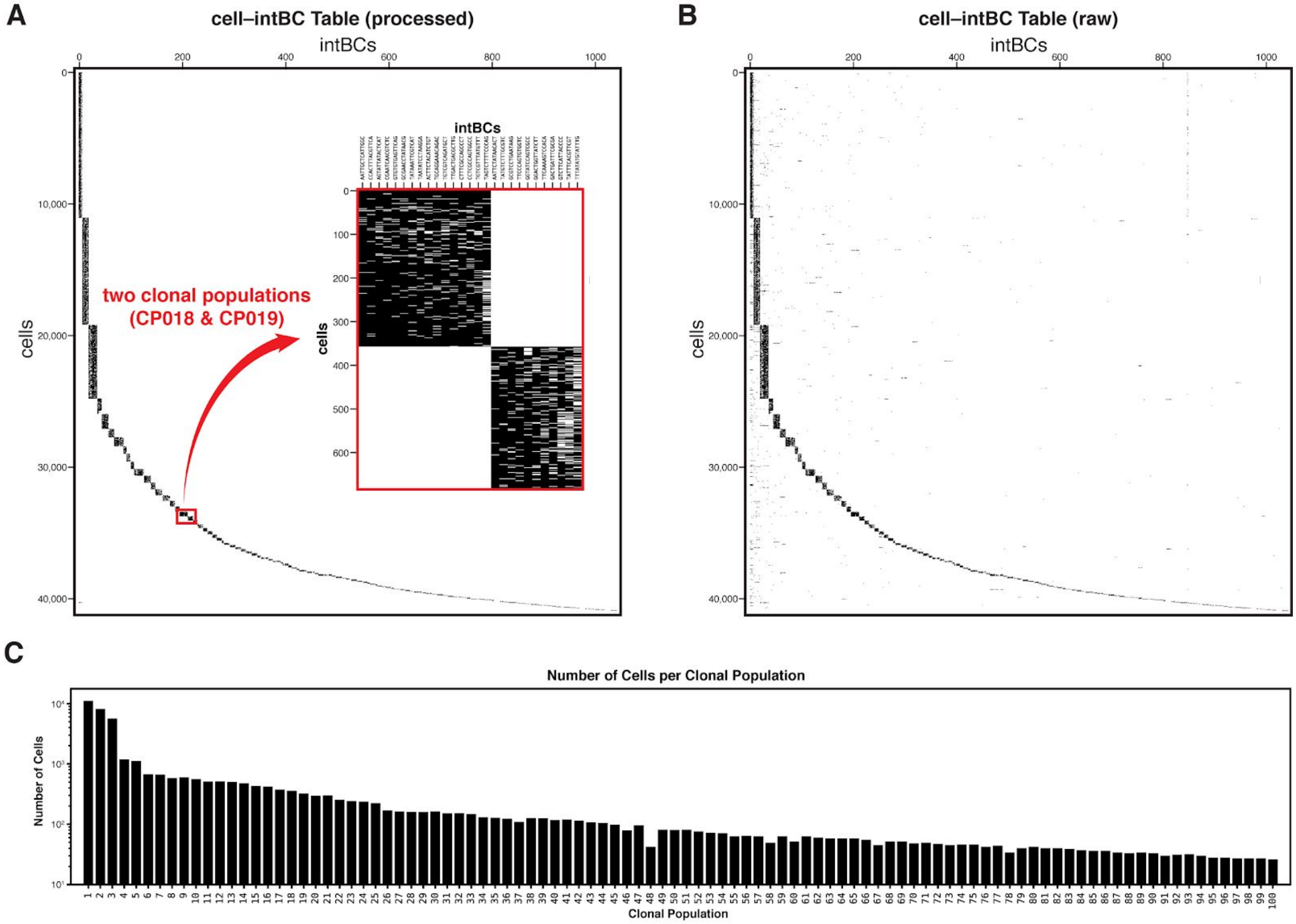


Fig. S5. Identifying clonal populations by shared integration barcodes (intBCs). (**A** and **B**) Tables representing the >1,000 unique integration barcodes (intBCs; columns) and >40,000 cells (rows) observed in mouse M5k, after processing (**A**) and before processing (**B**). Because intBCs are clonally inherited, cells that share identical intBCs are grouped into clonal populations (here, black blocks). (**A; inset**) The set of intBCs is generally exclusive to a single clonal population, such as those observed in clonal populations #18 and #19 (CP018 and CP109). (**B**) In the raw, unprocessed cell–intBC table, cells may be associated with intBCs that are not in their defined clonal set (indicated here by density outside of the black blocks). Through the processing pipeline, these conflicting intBCs (and/or the cells to which they pertain) are identified as cell doublets, cell-free transcripts in emulsion droplets, or sequencing artifacts and removed (Methods). (**C**) The number of cells per clonal population, generally numbered by descending population size.

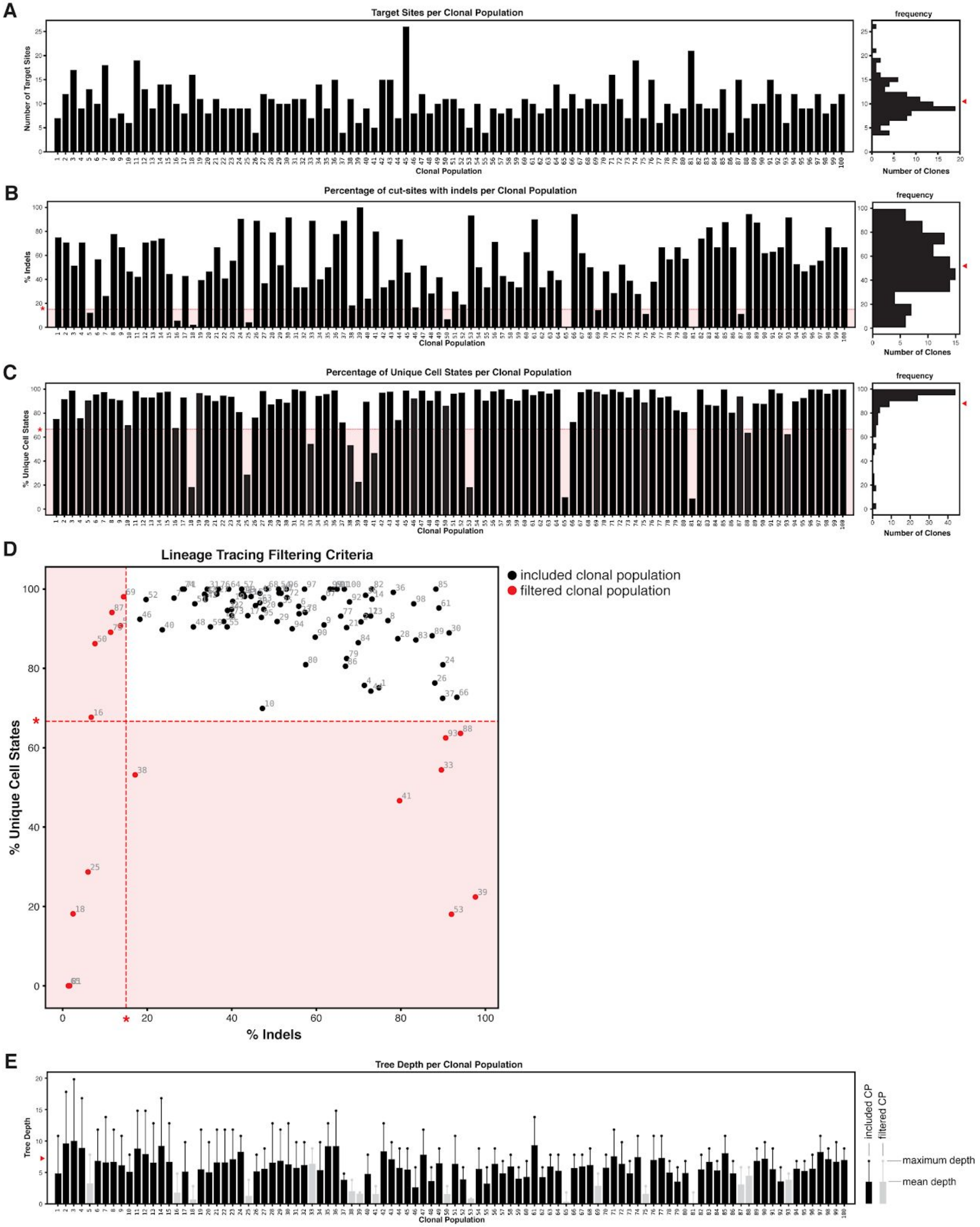

A
Target Sites per Clonal Population
Number of Target Sites
Clonal Population
frequency
Number of Clones
B
Percentage of cut-sites with indels per Clonal Population
% Indels
Clonal Population
frequency
Number of Clones
C
Percentage of Unique Cell States per Clonal Population
% Unique Cell States
Clonal Population
frequency
Number of Clones
D
Lineage Tracing Filtering Criteria
● included clonal population
● filtered clonal population
% Unique Cell States
% Indels
E
Tree Depth per Clonal Population
Tree Depth
Clonal Population
included CP
filtered CP
maximum depth
mean depth

Fig. S6. Characteristics of the lineage tracer and quality-control of clonal populations. (**A**) The copy-number of integrated Target Sites per clonal population, as determined by the number of unique intBCs. (**A, right**) The distribution of Target Site copy-number per clonal population; mean copy-number of Target Sites indicated (red arrowhead). (**B**) The percentage of cut-sites bearing lineage indel alleles per clonal population. Clonal populations with <15% indels are excluded (red asterisk, red underlay). (**B, right**) The distribution of indel-bearing cut-sites per clonal population; mean percentage indicated (red arrowhead). (**C**) The percentage of unique cell lineage states per clonal population (i.e., lineage diversity). Clonal populations with <66.7% diversity were excluded (red asterisk, red underlay). (**C, right**) The distribution of lineage diversity per clonal population; mean percentage indicated (red arrowhead). (**D**) Comparison of lineage tracing characteristics (% indels and % unique cell states) to define quality-control filtering criteria. Clonal populations removed by the filter (red closed circles) and filtering thresholds (red asterisks, red underlay) are indicated. (**E**) The depth of the reconstructed phylogenetic trees for each clonal population. Mean tree depths are shown as closed bars; maximum tree depths are shown as whiskers. Clonal populations that were excluded due to suboptimal lineage tracing characteristics (gray) and the average tree depth across all clonal populations (red arrowhead) are indicated.

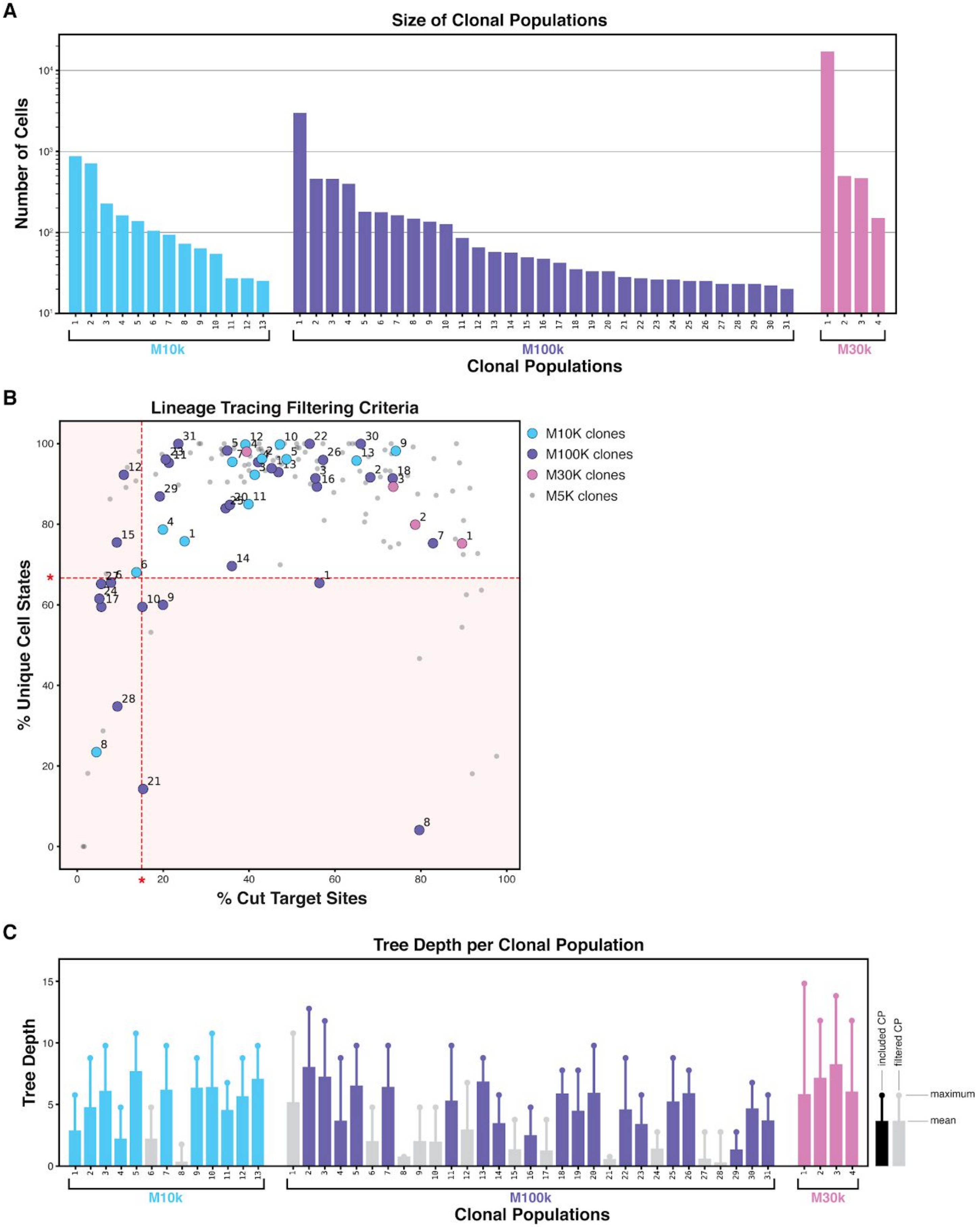

A
Size of Clonal Populations
Number of Cells
10¹
10²
10³
10⁴
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
1 2 3 4
M10k
M100k
M30k
Clonal Populations
B
Lineage Tracing Filtering Criteria
% Unique Cell States
0
20
40
60
80
100
*
% Cut Target Sites
0
*
20
40
60
80
100
M10K clones
M100K clones
M30K clones
M5K clones
C
Tree Depth per Clonal Population
Tree Depth
0
5
10
15
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
1 2 3 4
M10k
M100k
M30k
Clonal Populations
included CP
filtered CP
maximum
mean

Fig. S7. Lineage tracing characteristics of clonal populations in additional mice. (**A**) Number of cells in each clonal population in M10k, M100k, and M30k mice. (**B**) Scatter plot of the percentage of cut-sites bearing indels and the percentage of unique cell states per clonal population, which are characteristics of the lineage tracer that influence tree reconstructability. Some clonal populations exhibited suboptimal parameters (red asterisks and red field) and were excluded from reconstruction and downstream analyses. (**C**) The mean and maximum depths of the reconstructed phylogenetic trees for each clonal population in additional mice, as in **Fig. S6E**.

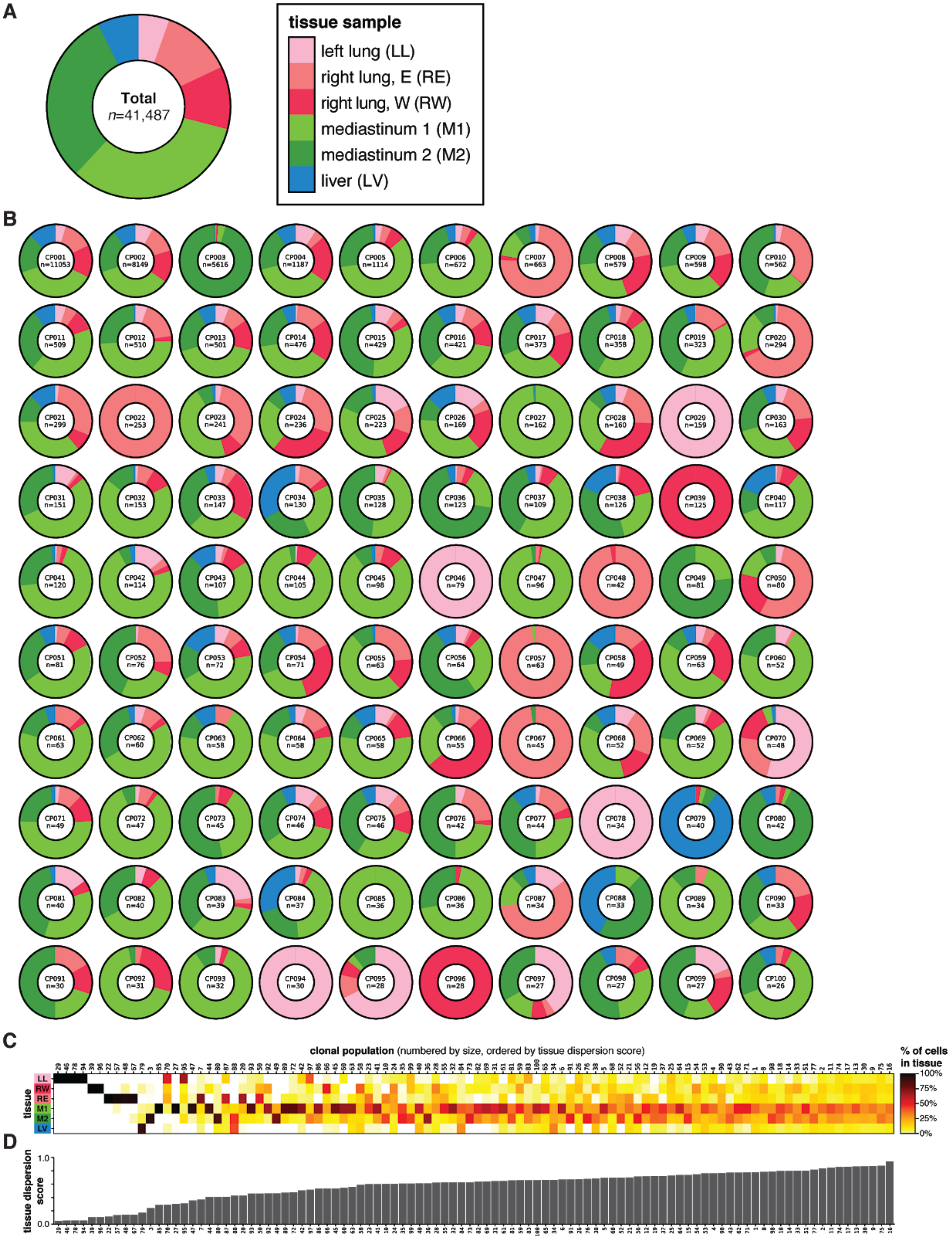

A
Total
n=41,487
tissue sample
left lung (LL)
right lung, E (RE)
right lung, W (RW)
mediastinum 1 (M1)
mediastinum 2 (M2)
liver (LV)
B
CP001 n=11053
CP002 n=8149
CP003 n=5616
CP004 n=1187
CP005 n=1114
CP006 n=672
CP007 n=663
CP008 n=579
CP009 n=598
CP010 n=562
CP011 n=509
CP012 n=510
CP013 n=501
CP014 n=476
CP015 n=429
CP016 n=421
CP017 n=373
CP018 n=358
CP019 n=323
CP020 n=294
CP021 n=299
CP022 n=253
CP023 n=241
CP024 n=236
CP025 n=223
CP026 n=169
CP027 n=162
CP028 n=160
CP029 n=159
CP030 n=163
CP031 n=151
CP032 n=153
CP033 n=147
CP034 n=130
CP035 n=128
CP036 n=123
CP037 n=109
CP038 n=126
CP039 n=125
CP040 n=117
CP041 n=120
CP042 n=114
CP043 n=107
CP044 n=105
CP045 n=98
CP046 n=79
CP047 n=96
CP048 n=42
CP049 n=81
CP050 n=80
CP051 n=81
CP052 n=76
CP053 n=72
CP054 n=71
CP055 n=63
CP056 n=64
CP057 n=63
CP058 n=49
CP059 n=63
CP060 n=52
CP061 n=63
CP062 n=60
CP063 n=58
CP064 n=58
CP065 n=58
CP066 n=55
CP067 n=45
CP068 n=52
CP069 n=52
CP070 n=48
CP071 n=49
CP072 n=47
CP073 n=45
CP074 n=46
CP075 n=46
CP076 n=42
CP077 n=44
CP078 n=34
CP079 n=40
CP080 n=42
CP081 n=40
CP082 n=40
CP083 n=39
CP084 n=37
CP085 n=36
CP086 n=36
CP087 n=34
CP088 n=33
CP089 n=34
CP090 n=33
CP091 n=30
CP092 n=31
CP093 n=32
CP094 n=30
CP095 n=28
CP096 n=28
CP097 n=27
CP098 n=27
CP099 n=27
CP100 n=26
C
clonal population (numbered by size, ordered by tissue dispersion score)
29 46 78 94 39 96 22 57 48 67 79 3 85 70 27 95 47 7 44 80 87 88 20 93 50 92 49 89 72 42 97 86 66 45 60 63 58 23 41 10 24 35 99 40 36 28 55 32 84 73 82 69 31 61 81 59 83 100 65 34 6 91 26 76 38 5 68 52 21 56 12 19 37 25 64 15 54 53 4 90 43 62 71 1 8 98 18 14 33 51 77 2 11 74 17 13 30 9 75 16
tissue
LL
RW
RE
M1
M2
LV
% of cells in tissue
100%
75%
50%
25%
0%
D
tissue dispersion score
1.0
0.0
29 46 78 94 39 96 22 57 48 67 79 3 85 70 27 95 47 7 44 80 87 88 20 93 50 92 49 89 72 42 97 86 66 45 60 63 58 23 41 10 24 35 99 40 36 28 55 32 84 73 82 69 31 61 81 59 83 100 65 34 6 91 26 76 38 5 68 52 21 56 12 19 37 25 64 15 54 53 4 90 43 62 71 1 8 98 18 14 33 51 77 2 11 74 17 13 30 9 75 16

Fig. S8. Clonal populations exhibit distinct tissue distributions. (**A**) The bulk distribution of all collected cells across the six tissue samples. (**B**) The distributions of cells from each clonal population across the six tissue samples. Some clonal populations were exclusive to the primary tissue (e.g., clonal population CP046), whereas some clones exhibited biased tissue distributions (e.g., CP003) and many others were observed broadly distributed across all tissues (e.g., CP011). (**C**) Tissue distributions of the largest 100 clonal populations. (**D**) The Tissue Dispersion Score is a statistical measurement of the distribution across tissues for each clonal population; *x*-axis shared with (**C**).

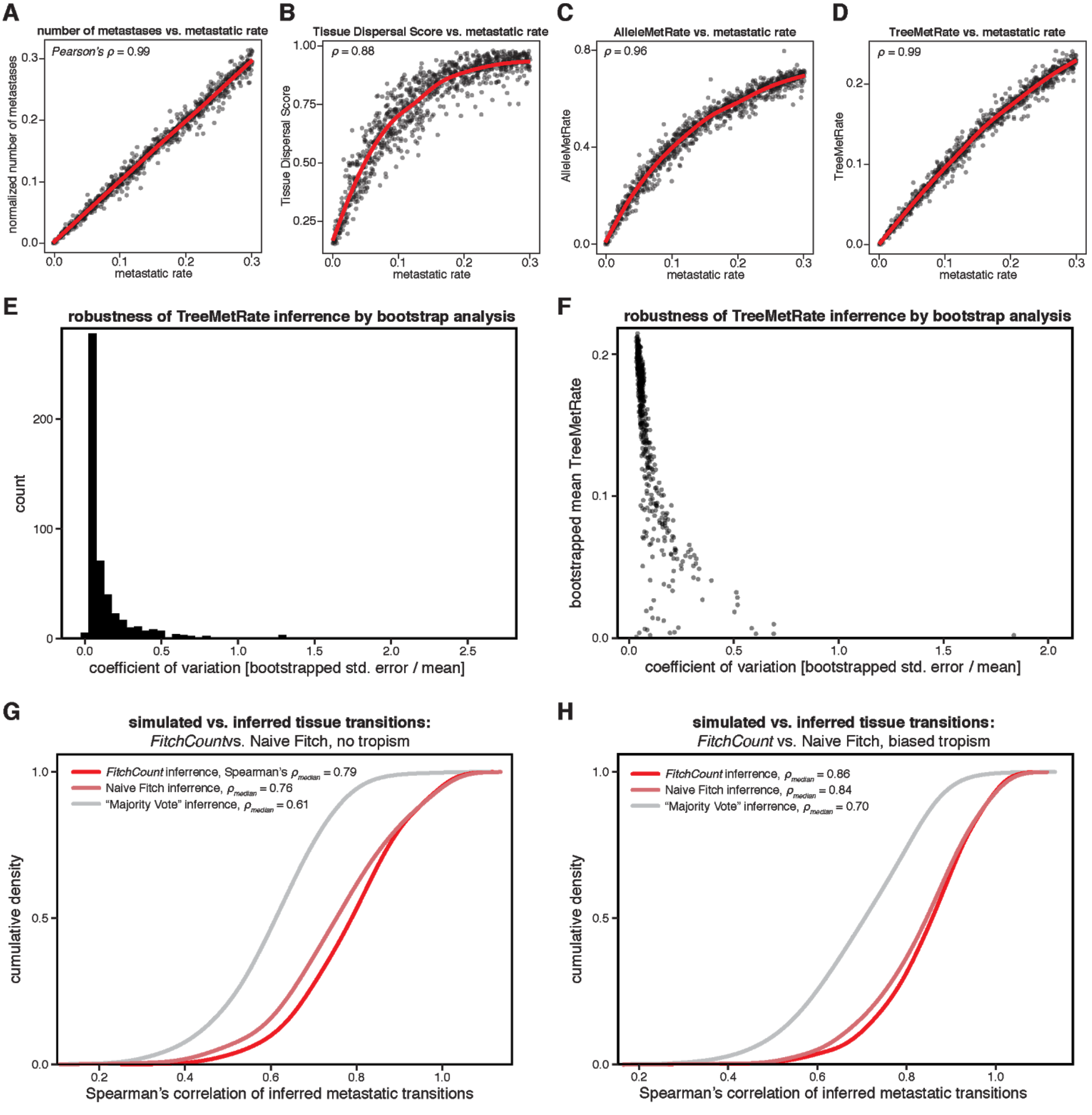


Fig. S9. Assessing the accuracy of different measurements of metastatic rate and inference of tissue transitions using simulated lineages. (**A–D**) Comparison between the simulated metastatic rate and various lineage tracer-derived statistics, with the correlation (Pearson's ρ) indicated; red lines represent moving average. (**A**) The normalized count of simulated metastatic transitions is very well correlated with the simulated metastatic rates, and serves as a ground-truth benchmark. (**B**) The Tissue Dispersal Score, which is a statistical measure of how closely a clone's tissue distribution matches the background tissue distributions, is correlated with the metastatic rate, but saturates at intermediate metastatic regimes. (**C**) The AlleleMetRate, or the proportion of cells whose closest relative by allele similarity is in a different tissue, is better correlated with

metastatic rate. (**D**) The TreeMetRate, or the proportion of inferred metastases in a reconstructed phylogeny, is the best lineage-derived measurement of metastatic phenotype by correlation. (**E** and **F**) Robustness of the TreeMetRate inference by bootstrap analysis. (**E**) The distribution of the coefficients of variation of the TreeMetRate across 50,000 simulated, bootstrapped phylogenies (Methods). The small coefficient of variation indicates that the TreeMetRate is a robust measurement. (**F**) The TreeMetRate coefficient of variation is only relatively large when the mean TreeMetRate is small. (**G** and **H**) Cumulative density plots assessing the accuracy of the *FitchCount* strategy for inferring ancestral tissue transition from simulated phylogenies with or without simulated biased tissue transitions (thus approximating tropism; Methods: "*Assessing accuracy of the tissue transition matrices*"). *FitchCount* outperforms other inference approaches, as measured by the Spearman correlation between the inferred and the ground-truth conditional tissue transition probability matrices. *FitchCount* was benchmarked against two other inference methods: (i) a "naive" single-solution implementation of the Fitch-Hartigan maximum parsimony algorithm or (ii) "Majority Vote", which infers ancestral tissue location as the tissue in which the majority of cells below each clade-level reside. Conditional probabilities for each algorithm were obtained by row-normalizing the count matrix with respect to the non-diagonal counts in each row.

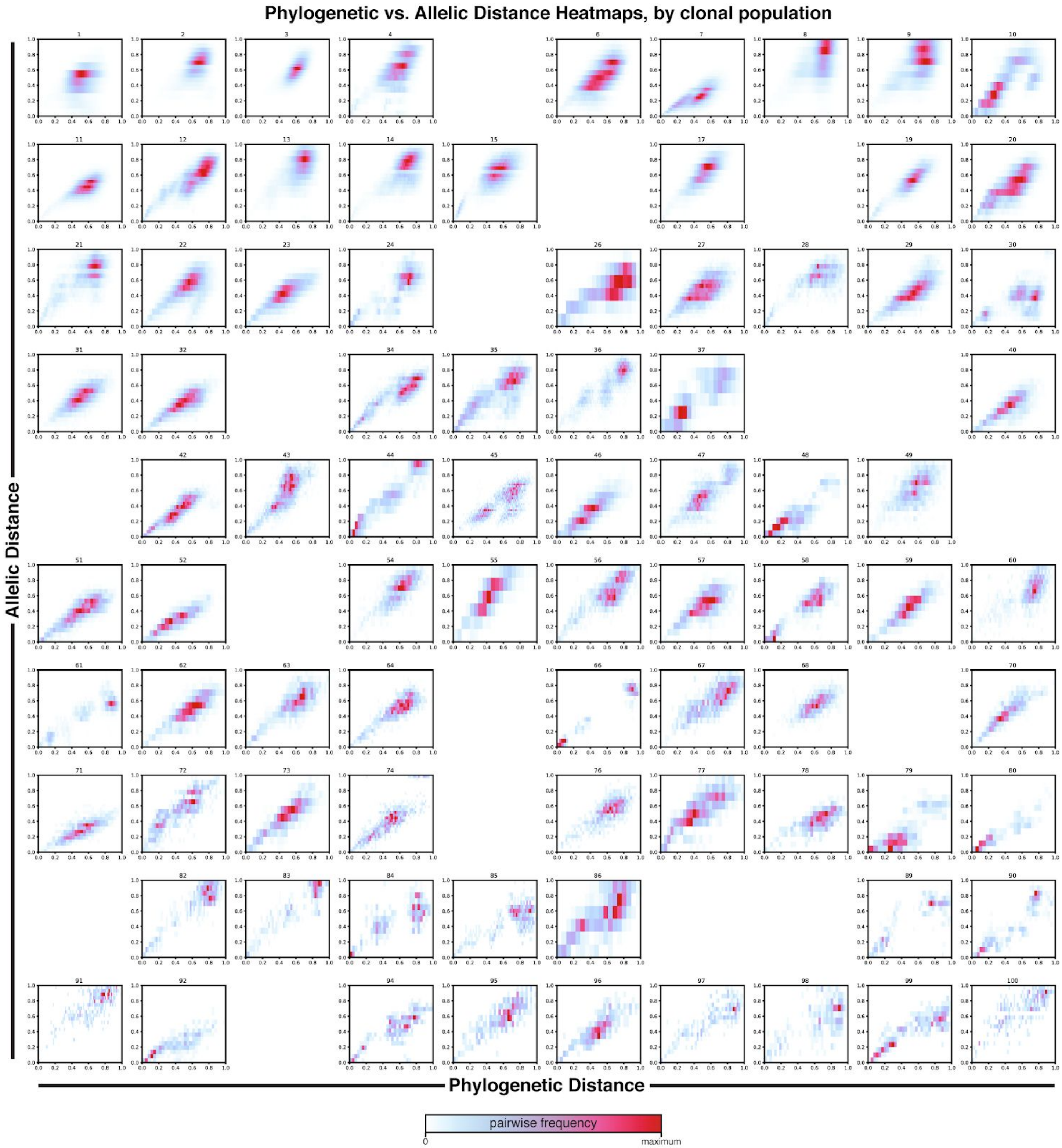


Fig. S10. Relationship between phylogenetic distance and allelic distance for each clonal population. Density heat-maps comparing phylogenetic distance (i.e., the normalized tree branch distance between two cells) and allelic distance (i.e., the normalized difference in lineage allele state between two cells) for all pairwise cell–cell relationships and for each clonal population, as in **Fig. 2C**. As expected, phylogenetic and allelic distances are correlated, suggesting that the reconstructed trees are a good phylogenetic model of cell–cell relationships. Excluded clonal populations are not shown.

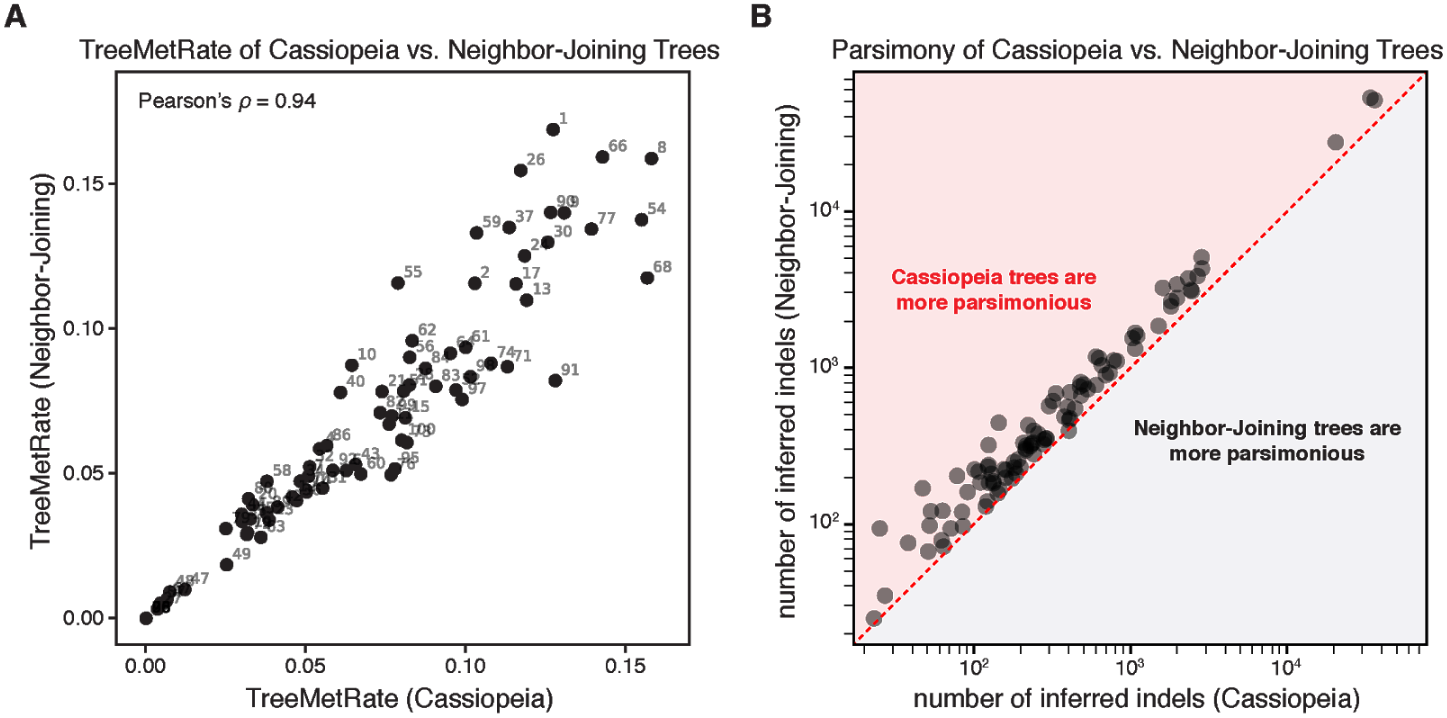


Fig. S11. The TreeMetRate is stable across tree reconstruction algorithms (Cassiopeia versus Neighbor-Joining). (**A**) Comparison of the TreeMetRates for each clonal population from Cassiopeia trees and Neighbor-Joining trees. The TreeMetRates are correlated for both the Cassiopeia and Neighbor-Joining trees (Pearson's ρ=0.94). (**B**) Comparison of the parsimony of Cassiopeia trees and Neighbor-Joining trees, defined as the number of inferred indels in each tree. Notably, the Cassiopeia trees are more parsimonious than the Neighbor-Joining trees (i.e., they have fewer inferred indels; red overlay).

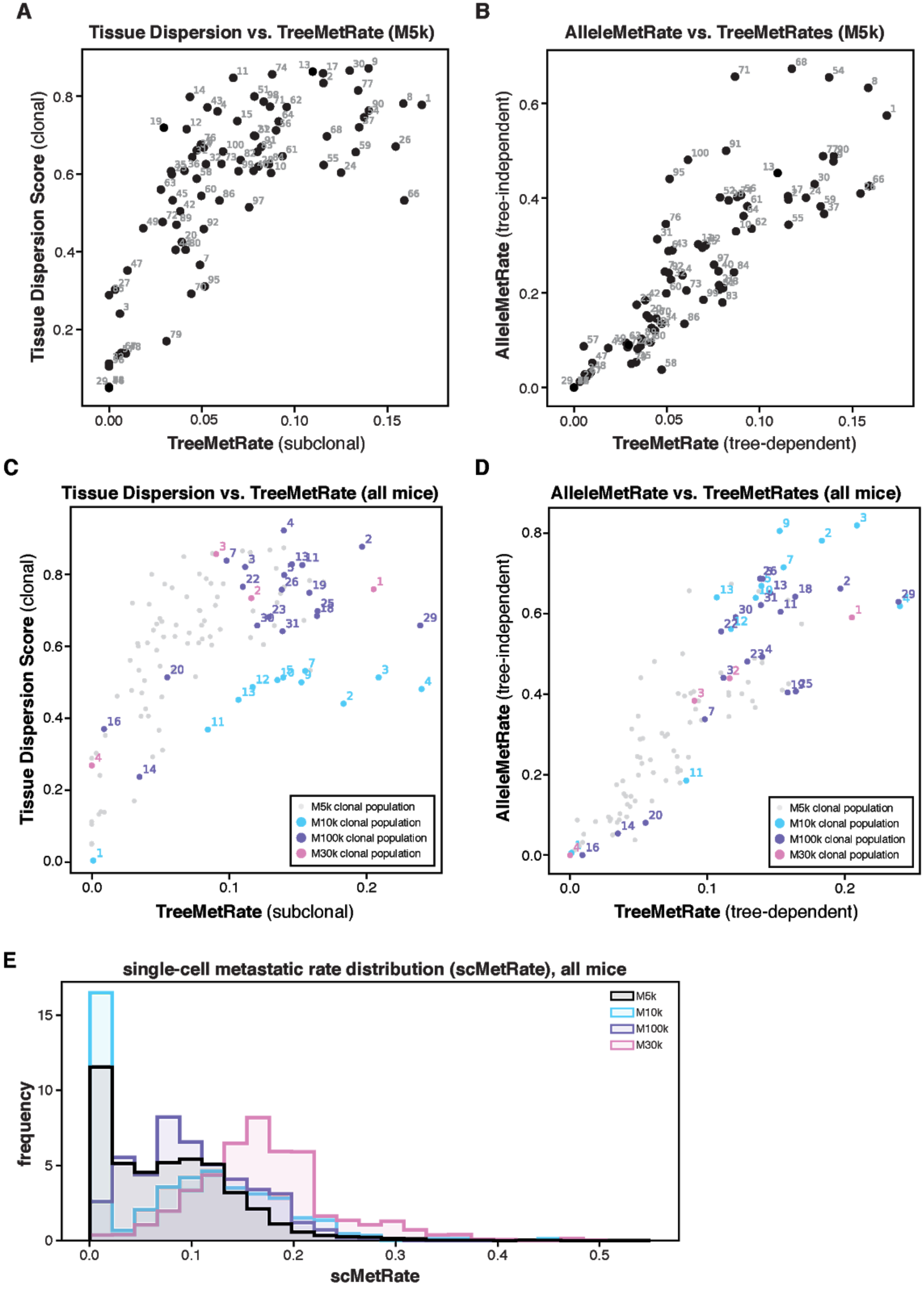

A
Tissue Dispersion vs. TreeMetRate (M5k)
Tissue Dispersion Score (clonal)
TreeMetRate (subclonal)
0.2
0.4
0.6
0.8
0.00
0.05
0.10
0.15
B
AlleleMetRate vs. TreeMetRates (M5k)
AlleleMetRate (tree-independent)
TreeMetRate (tree-dependent)
0.0
0.2
0.4
0.6
0.00
0.05
0.10
0.15
C
Tissue Dispersion vs. TreeMetRate (all mice)
Tissue Dispersion Score (clonal)
TreeMetRate (subclonal)
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
M5k clonal population
M10k clonal population
M100k clonal population
M30k clonal population
D
AlleleMetRate vs. TreeMetRates (all mice)
AlleleMetRate (tree-independent)
TreeMetRate (tree-dependent)
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
M5k clonal population
M10k clonal population
M100k clonal population
M30k clonal population
E
single-cell metastatic rate distribution (scMetRate), all mice
frequency
scMetRate
0
5
10
15
0.0
0.1
0.2
0.3
0.4
0.5
M5k
M10k
M100k
M30k

Fig. S12. Clonal populations exhibit broad metastatic phenotypes, measured by Tissue Dispersal Score, AlleleMetRate, and TreeMetRate. (**A–D**) We evaluated the metastatic phenotype of each clonal population from mouse M5k (**A, B**) and the additional mice M10k, M100k, and M30k (**C, D**) using our three lineage-derived measurements: Tissue Dispersal Score (as in **Fig. S8D**), AlleleMetRate, and TreeMetRate (as in **Fig. 3C**). These three measurements follow similar relative trends across all four mouse experiments. First, the clonal populations exhibit a broad range of metastatic phenotypes. Second, all three measurements are correlated with one another, though simulations indicate that the TreeMetRate is the most accurate for estimating the underlying metastatic rate (**Fig. S9**). Third, Tissue Dispersal Score saturates at intermediate metastatic regimes (**A** and **C**). (**E**) The distribution of single-cell-resolution metastatic rates (scMetRates) across all cells for each mouse (as in **Fig. 3D**). Though all mice have broad distributions of metastatic phenotypes, mouse M5k (black) is particularly well represented by cells in low-to-intermediate metastatic regimes whereas mouse M30k (pink) has very few cells in the low metastatic regime.

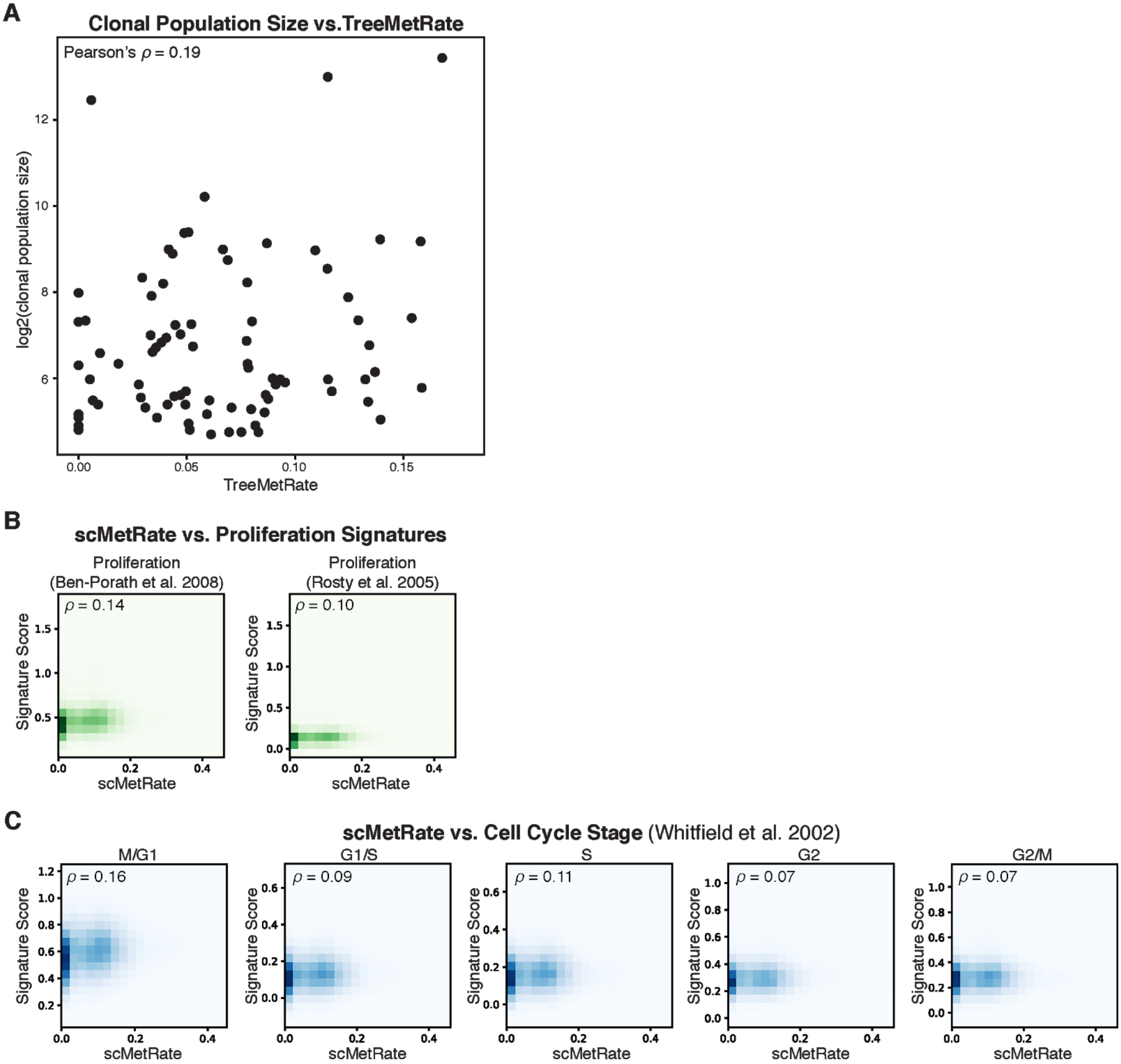


Fig. S13. The scMetRate measures metastatic potential decoupled from proliferative capacity. (**A**) There is poor correlation between the scMetRate and the log2-transformed clonal population size, a proxy for clonal fitness. (**B** and **C**) Density heat-maps comparing the scMetRate and various transcriptional signatures. Notably, the scMetRate is poorly correlated with transcriptional signatures of proliferation (**B**) nor stages of the cell cycle (**C**). Pearson's correlations (ρ) are indicated for each subplot.

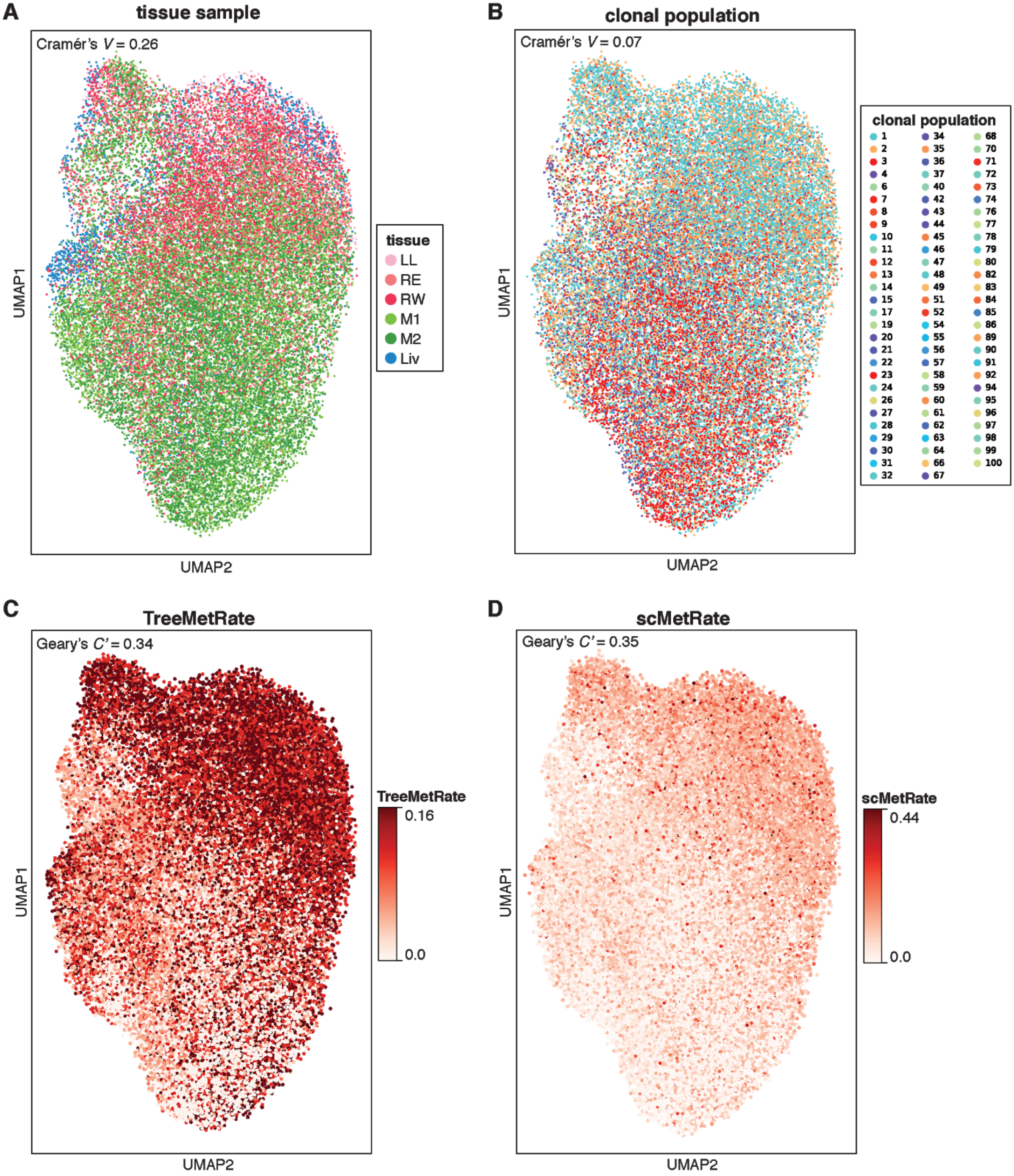


Fig. S14. Effects of tissue sample, clonal population, and metastatic rate on transcriptional state. To identify global trends in the gene expression data, we used *Vision* (*52*) to statistically assess the transcriptional effect of four features: (**A**) tissue sample, (**B**) clonal population identity, (**C**) TreeMetRate, and (**D**) scMetRate. The transcriptional states are represented here as a two-dimensional projection using Uniform Manifold

Approximation and Projection (UMAP; B). Distinctions in transcriptional state are not predominantly explained by the clonal population (**B**; by Cramér's *V*), though there is modest association between transcriptional state and both tissue sample and metastatic phenotype (by Cramér's *V* and inverted Geary's *C'*, respectively).

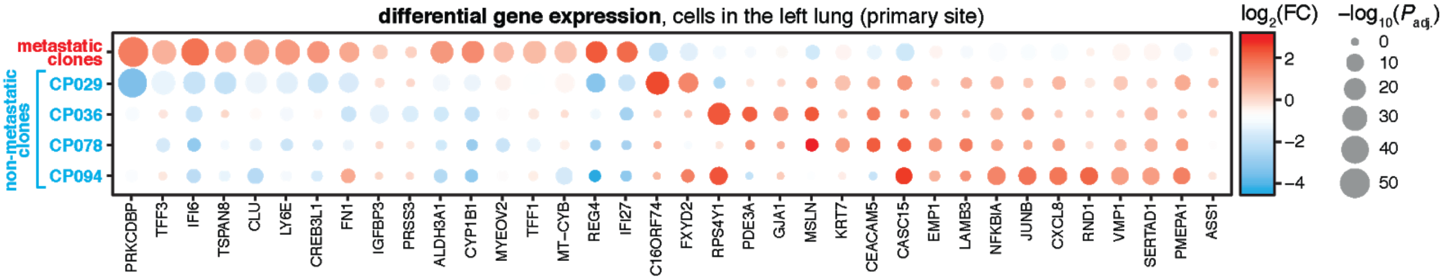


Fig. S15. Differential expression between non-metastatic and metastatic clonal populations in the primary tissue. Differential gene expression analysis comparing four non-metastatic clonal populations (CP029, 36, 78, and 94) and all metastatic clonal populations in the primary tumor tissue (i.e., all other cells in the left lung). Significantly differentially expressed genes are colored by the log2-transformed fold-change in gene expression and scaled by the adjusted Wilcoxon rank-sum test *P*-value.

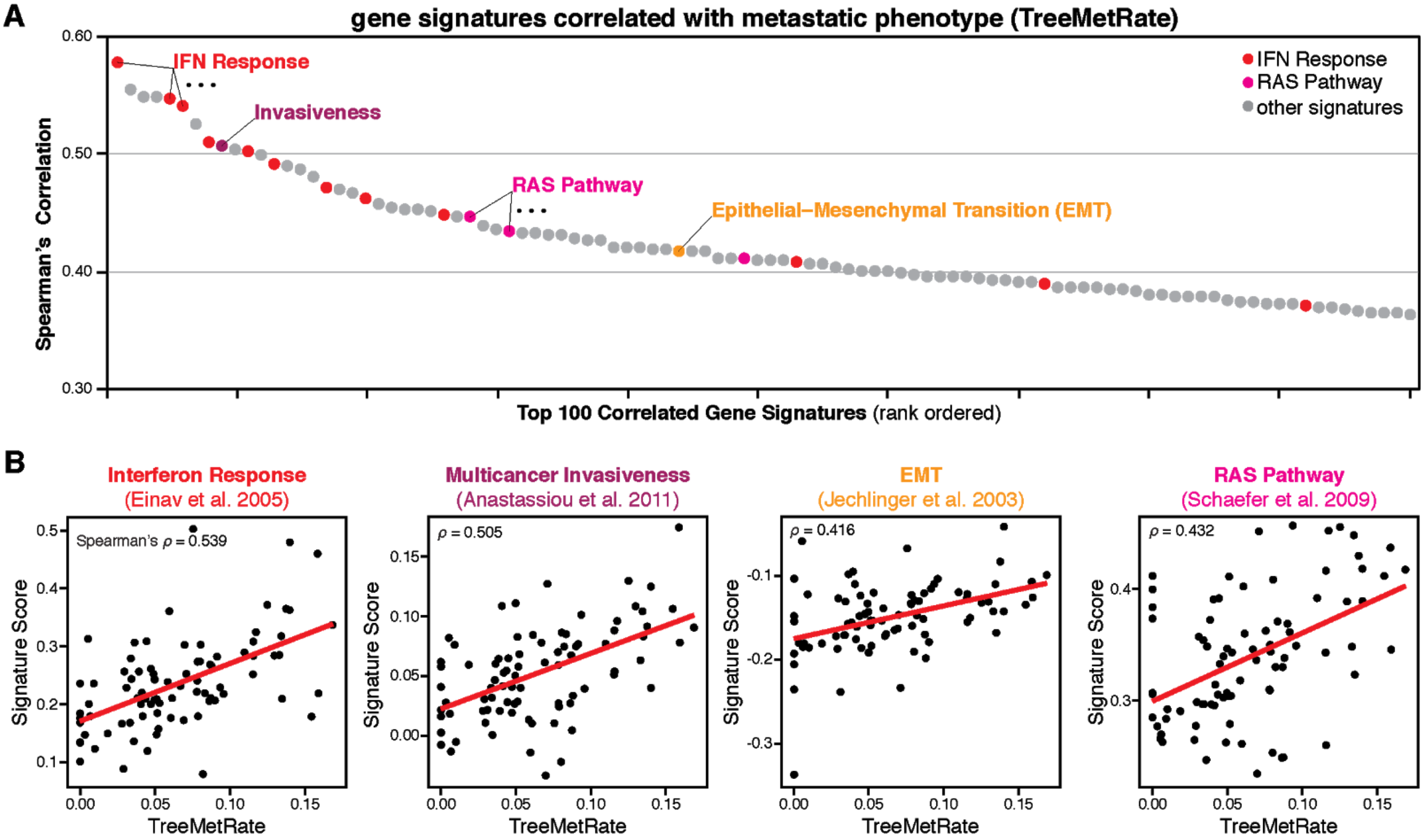


Fig. S16. Metastasis-related gene signatures are correlated with metastatic potential. (**A**) Rank-ordered gene signatures that are the most positively correlated with TreeMetRate, including many related to interferon response (red) and RAS pathways (magenta), as well as other metastasis-related signatures. (**B**) Scatter plots showing the correlation between TreeMetRate and noted gene signature scores per clonal population; Spearman's correlation (ρ) indicated.

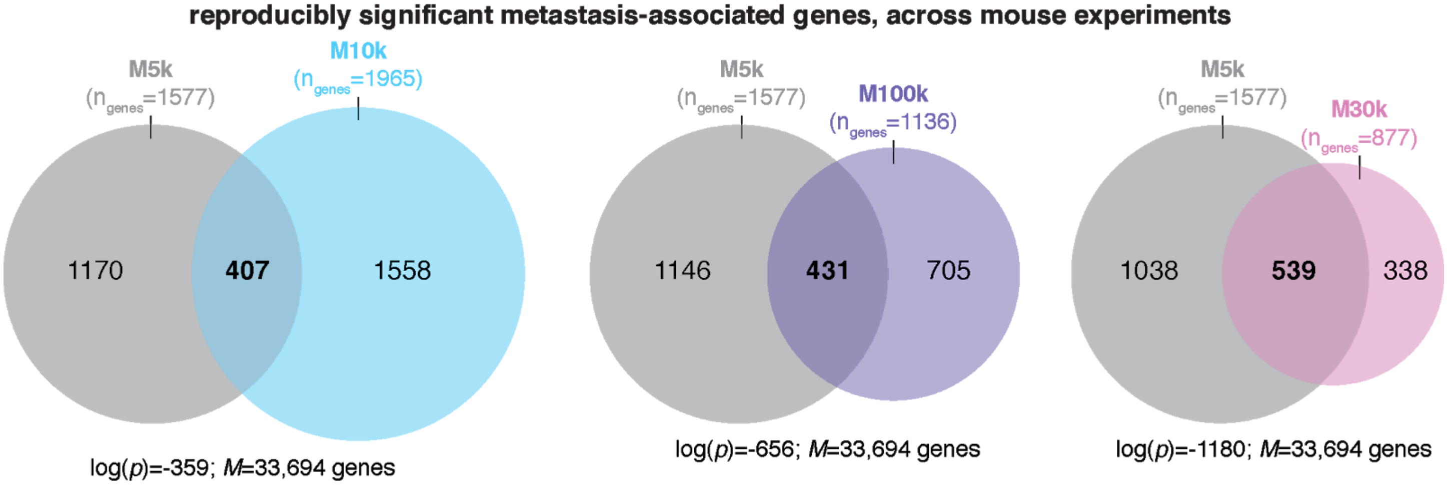


Fig. S17. Many of the same genes are associated with metastatic phenotype across all mice. Using the same regression strategy as in the analysis of mouse M5k, we found many genes with expression that is significantly associated with high or low scMetRates. The number of significant genes for each mouse (FDR < 0.01) and their overlap in the same direction with mouse M5k (gray) are shown (n_{genes}). In all cases, the overlap between mouse M5k and each additional mouse is significant by hypergeometric test (*p*-value and *M* indicated).

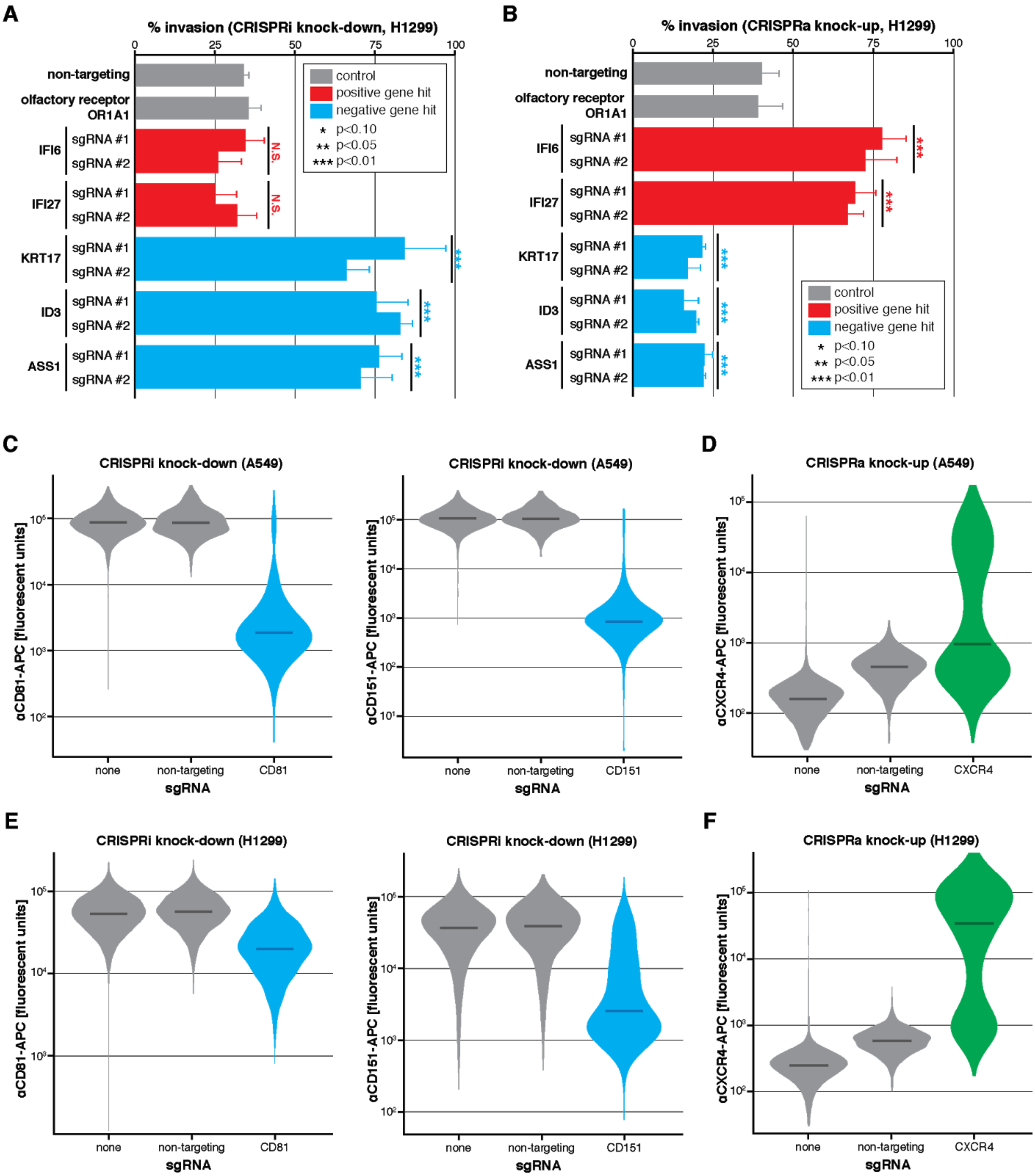


Fig. S18. Functional validation of five gene candidates in a different cell line (H1299s) and validation of CRISPRi and CRISPRa activity. (**A** and **B**) *In vitro* transwell invasion assays following CRISPRi or CRISPRa gene perturbation, respectively, in H1299 cells; as in **Fig.4E, F**. Perturbation of positive and negative metastasis-associated gene candidates were performed in triplicate using two independent sgRNAs per gene.

Differences in invasion phenotype relative to two negative control guides (non-targeting and olfactory receptor) were significant by two-tailed *t*-test. N.S., not significant; error bars show standard deviation. (**C, E**) A549-CRISPRi cells or H1299-CRISPRi cells, respectively, were treated with no sgRNA, non-targeting sgRNA, or sgRNAs against either CD81 or CD151, two highly expressed cell-surface markers. One week following treatment, the cells were collected, stained with APC-labelled anti-CD81 or anti-CD151 antibodies and their fluorescence was measured by flow cytometry. Shown here is substantial knock-down of CD81 or CD151 gene expression relative to no sgRNA or non-targeting sgRNA controls. (**D, F**) The same validation experiment as conducted in **C** and **E**, but for A549-CRISPRa cells or H1299-CRISPRa cells, respectively, treated with an sgRNA against CXCR4, a lowly expressed cell-surface marker. Shown here is substantially increased CXCR4 gene expression relative to no sgRNA or non-targeting sgRNA controls. Violin plots show distribution of fluorescent signal for each cell identified by flow cytometry; median marked by dark bar.

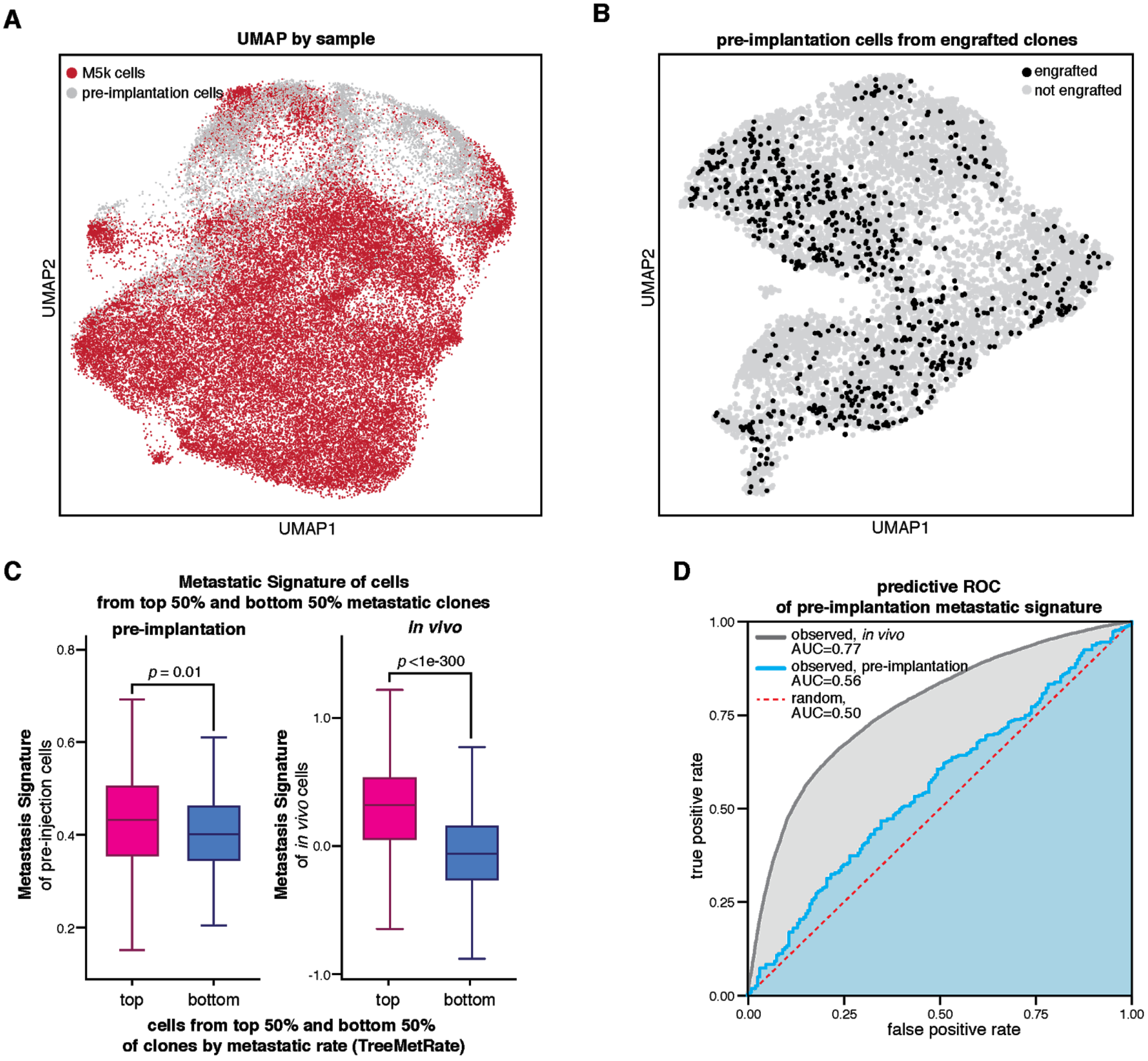


Fig. S19. The cells in the pre-implantation pool heterogeneously express metastasis-associated genes, which are modestly predictive of their *in vivo* metastatic phenotype. (**A**) Projection of transcriptional states of M5k and pre-implantation cells, colored by sample, as in **Fig. 5A**. (**B**) Some of the cells from the pre-implantation pool could be assigned to the ~100 clonal populations that engrafted and proliferated in mouse M5k based on their clonal barcodes (i.e., intBCs). Shown are the pre-implantation cells that could be assigned to an engrafted clone (black) on a projection of pre-implantation transcriptional states, as in **Fig. 5B,C**. (**C, left**) Pre-implantation cells from the top 50% (most) metastatic clones *in vivo* have higher Metastatic Signature scores than pre-implantation cells from the bottom 50% (least) metastatic clones *in vivo* (Mann-Whitney U p-value=0.01). (**C, right**) The Metastatic Signature of the most and least metastatic clones is more pronounced *in vivo* than in the pre-implantation cells (p-value<1e-300). (**D**) For the pre-implantation cells, the difference in Metastatic Signature scores between the most and least metastatic clones is modest, yet significant by ROC (receiver operator characteristic) analysis of false positives vs. true positives (area under the curve, AUC=0.56), indicating that the Metastatic Signature score pre-implantation is a modest predictor of *in vivo* metastatic phenotype. The predictive power for the *in vivo* population of cells is greater (AUC=0.77).

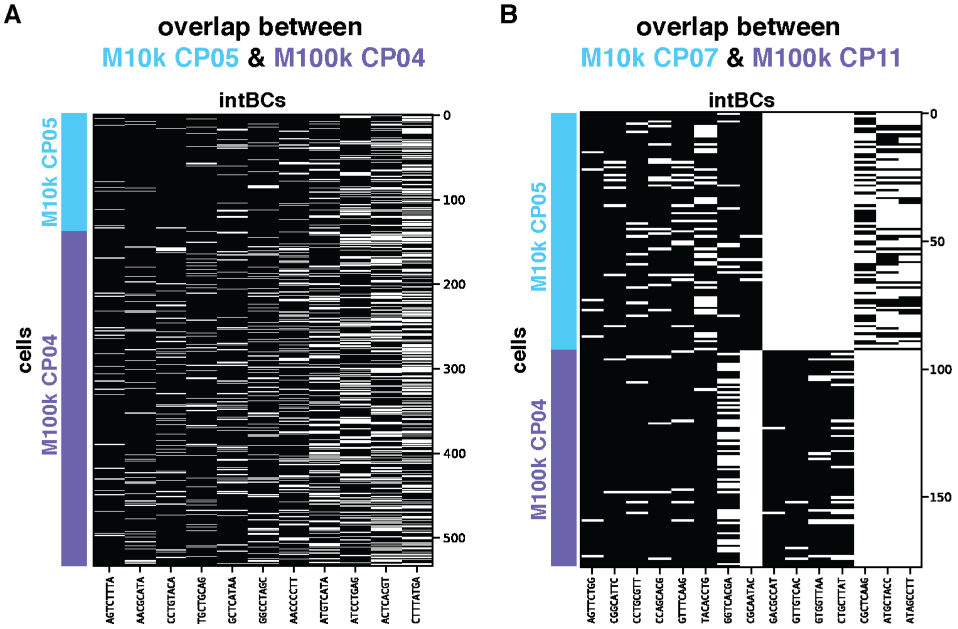


Fig. S20. Two pairs of clonal populations from mice M10k and M100k are related, enabling an experiment to determine the robustness and reproducibility of metastatic phenotype across independent mouse experiments. Each intBC (columns) observed for each cell (rows) from (**A**) M10k CP05 and M100k CP04 and (**B**) M10k CP07 and M100k CP11. Cells from M10k are shown in light blue; M100k in purple. The clonal populations in each of these pairs are related to one another based on their shared sets of intBCs, as in **Fig. 5D**. The paired clonal populations here are the most closely related between the two mouse experiments.

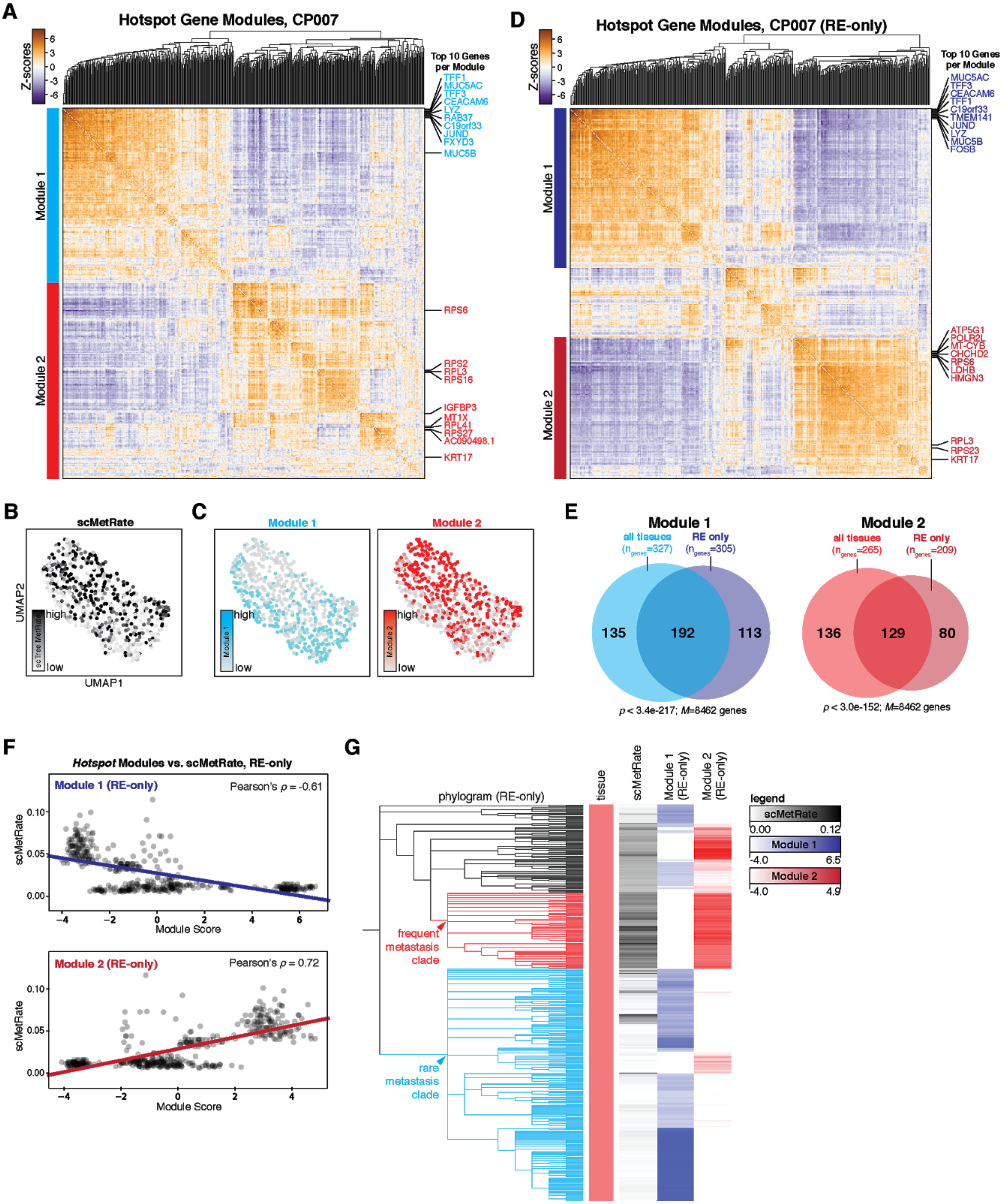


Fig. S21. Distinct transcriptional modules underlie distinct clade-specific metastatic behaviors in Clone #7. (**A**) *Hotspot* analysis identifies two gene modules that have heritable expression patterns in CP007 (Modules 1 and 2; indicated in cyan and red, respectively). Pairwise local correlations of genes (with FDR < 0.1)

calculated with *Hostpot* are shown by heat-map; the top 10 most significant genes each for Modules 1 and 2 are annotated in cyan and red, respectively. (**B** and **C**) The transcriptional states for all cells in CP007 represented in a two-dimensional projection (UMAP) and colored by scMetRate (**B**) or *Hotspot* Module scores (**C**). (**D**) When restricting *Hotspot* analysis to only cells from the "RE" tissue sample in CP007, two heritable gene modules are identified (dark blue and dark red). As before, the top 10 genes from each module are annotated. (**E**) The gene modules identified from all cells or from only RE-only cells in CP007 overlap significantly by hypergeometric test. Number of genes in each set (n_{genes}), number of expressed genes in at least 10% of cells in CP007 (*M*), and *p*-value of the hypergeometric test are indicated. (**F**) RE-only Modules 1 and 2 are negatively and positively associated with the scMetRate, respectively, as in the analysis for all cells from CP007 (**Fig. 5I**). (**G**) Overlay of the phylogram for RE-only cells from CP007, scMetRate, and *Hotspot* module scores (RE-only), showing concordance between the frequently metastasizing clade (red) and Module 2 and the rarely metastasizing clade (cyan) and Module 1, as in **Fig. 5J.**

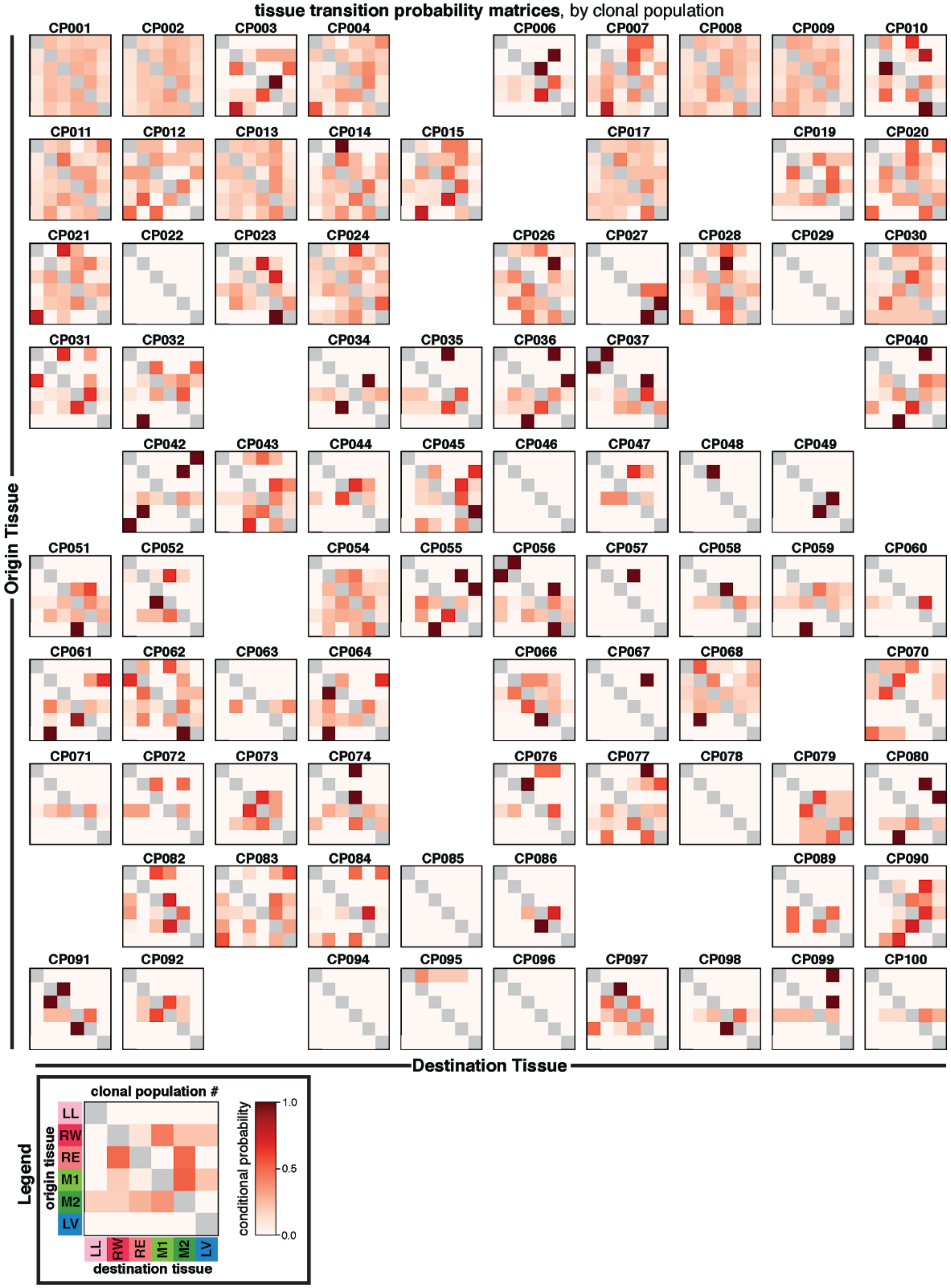

tissue transition probability matrices, by clonal population
CP001 CP002 CP003 CP004 CP006 CP007 CP008 CP009 CP010
CP011 CP012 CP013 CP014 CP015 CP017 CP019 CP020
CP021 CP022 CP023 CP024 CP026 CP027 CP028 CP029 CP030
CP031 CP032 CP034 CP035 CP036 CP037 CP040
CP042 CP043 CP044 CP045 CP046 CP047 CP048 CP049
CP051 CP052 CP054 CP055 CP056 CP057 CP058 CP059 CP060
CP061 CP062 CP063 CP064 CP066 CP067 CP068 CP070
CP071 CP072 CP073 CP074 CP076 CP077 CP078 CP079 CP080
CP082 CP083 CP084 CP085 CP086 CP089 CP090
CP091 CP092 CP094 CP095 CP096 CP097 CP098 CP099 CP100
Origin Tissue
Destination Tissue
Legend
clonal population #
origin tissue
LL
RW
RE
M1
M2
LV
destination tissue
conditional probability
1.0
0.5
0.0

Fig. S22. Tissue transition probability matrices for each clonal population. The conditional probability of transition from and to each tissue inferred from the phylogenetic trees (calculated with *FitchCount)* of each clonal population, thus summarizing the most probable tissue routes of metastasis. Legend (lower left) indicates the color bar showing conditional probability and the tissue labels, as in **Fig. 1E**. Notably, the transition matrices are varied and distinct to each clonal population.

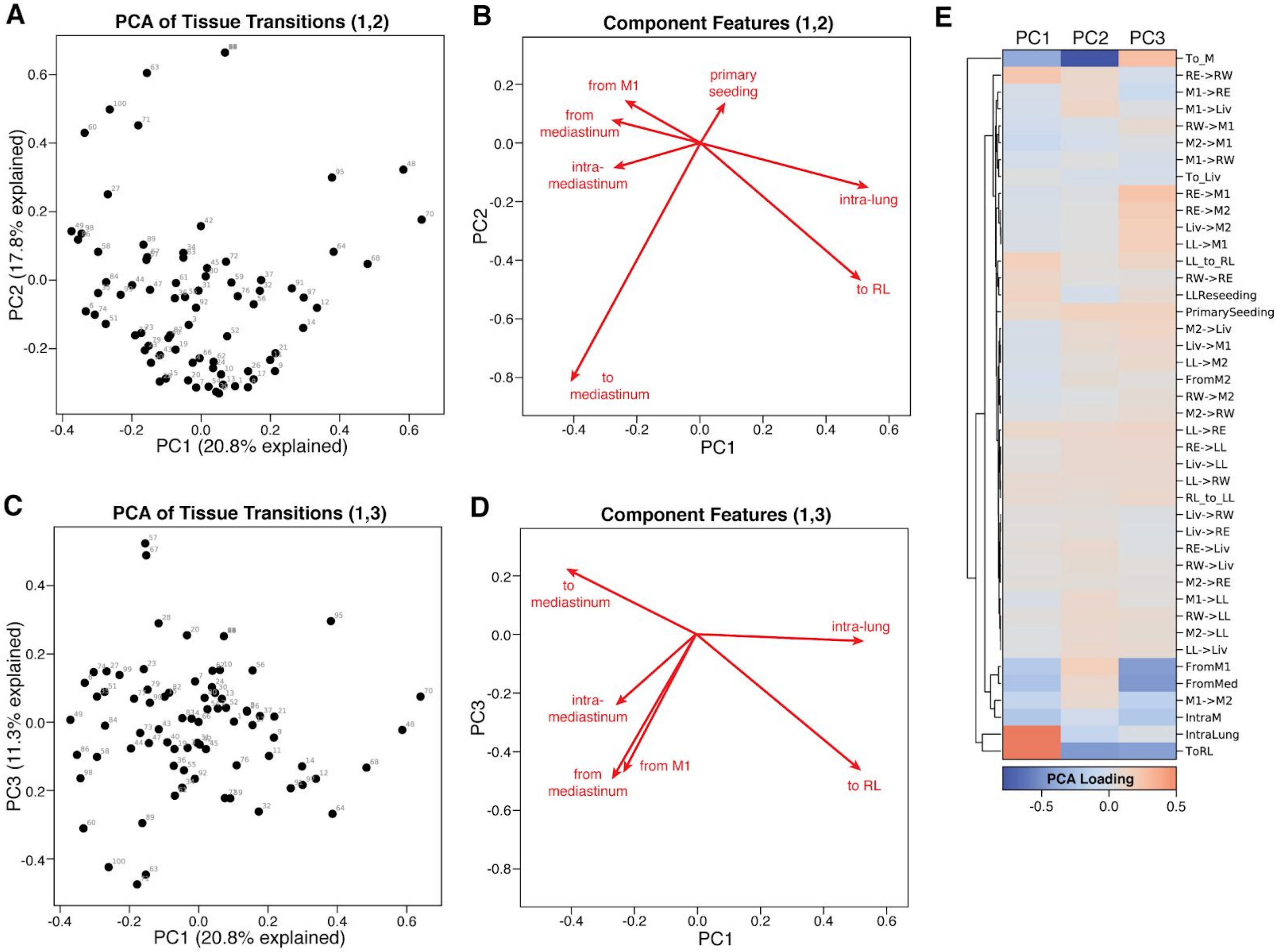


Fig. S23. Describing the principal features of metastatic seeding routes. (**A**, **C**) PCA projections of the metastatic tissue transitions for each clonal population (annotated). The percentage of the variance explained by each component is indicated on the axes for the first and second (**A**) or first and third (**C**) components. (**B**, **D**) Biplot vectors representing the most explanatory features of the first, second, and third principal components, annotated by descriptive features of metastatic transitions. The length and angle of the vector describe the scale and direction, respectively, of each descriptive feature. (**F**) The PCA loadings of the metastatic transition features for each principal component.

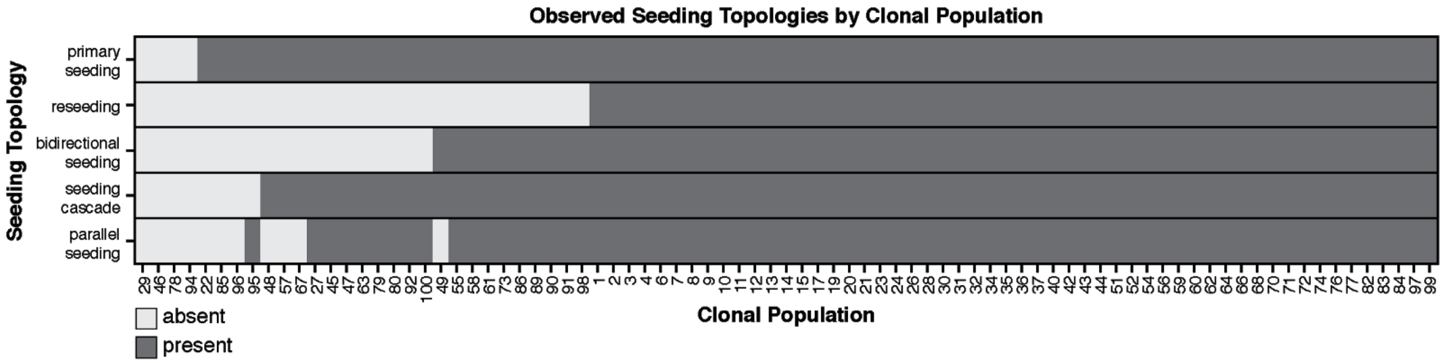


Fig. S24. Seeding topologies observed in each clonal population. A table describing classified seeding topologies (rows) that are present or absent (dark or light gray, respectively) in each clonal population (columns). The majority of clonal populations exhibit examples of all seeding topologies.

Supplementary Text: (embedded below)

FitchCount: an Efficient Algorithm for Inferring Transition Matrices on Phylogenetic Trees

April 16, 2020

Contents

1 Deriving transition matrices from phylogenetic trees

In this document, we describe our approach for solving the following problem: Consider a phylogenetic tree $\mathcal{T}$ rooted at vertex r over V vertices and E edges which we denote as $\mathcal{T}^r$. In this tree, each leaf l is assigned a state $state(l)$, where states are drawn from a state space Σ (i.e. $\forall v, state(v) \in \Sigma$). Here, leaves are single cells derived from a single-cell lineage tracing experiment (Chan et al. 2019) studying cancer metastasis, the tree describes the cells' evolutionary history (as inferred by Cassiopeia (Jones et al. 2020)), and the states represent the tissues from which the cells were obtained. Our task is to derive summary statistics obtained by assigning states to the internal nodes of the tree(i.e., ancestral cells that were not observed in the study). Specifically, the summary statistics we are interested at are: (1) the overall number of metastatic events (i.e., transition between tissues) that occurred in a clone's history and (2) the frequency (i.e. number) of transitions between each pair of tissues s_i and s_j.

This class of problems typically requires what is known as "ancestral state reconstruction" (Joy et al. 2016; Slatkin and Maddison 1989; McPherson et al. 2016), which in essence attempts to assign an ancestral states to every node in a given tree that minimizes some function. Here, we use parsimony as our objective function. Our work relies on classical algorithms (Fitch 1971; Hartigan 1973) and is also inspired by recent algorithms using these principles to infer metastatic histories like MACHINA (El-Kebir, Satas, and Raphael 2018).

As noted by El-Kebir et al (El-Kebir, Satas, and Raphael 2018), while there exist several algorithms for effectively inferring clone trees from DNA samples of metastatic cancers (Reiter et al. 2017; Deshwar et al. 2015; El-Kebir et al. 2015), none of them except for MACHINA propose an ancestral node labeling and subsequent classification of metastatic topologies. This is because a metastatic history does not follow uniquely from a tree structure, and metastasis itself is not necessarily unidirectional (i.e. there exist polyclonal & reseeding events that may introduce cyclic topologies). Though MACHINA represents a significant advance, we build on it by reporting summary statistics over *all* optimal solutions rather than one, and additionally circumvent the computationally-intensive Integer Linear Programming (ILP) optimization routine in favor for a dynamic programming approach.

Below, we describe our algorithmic strategy (and prove it) for inferring both summary statistics. We begin by describing how the overall number of transitions can be derived from the Fitch-Hartigan algorithm (Fitch 1971; Hartigan 1973). Next, we introduce *FitchCount*, an algorithm for counting the number of transitions between any two tissues in a given phylogeny over all optimal solutions to the Fitch-Hartigan algorithm.

2 Algorithmic strategy

The first summary statistic that we are interested ind is the minimal possible number of state transitions in the tree that is sufficient to explain the state assignment to the leaves (which is given as an input). In other words, out of all possible assignment of tissue labels to the ancestral cells (which can be exponentially many), consider the assignments that entail the minimum number of cases in which a parent node and a child node come from a different tissue (i.e., state transition). Our first goal is to retrieve the number of state transitions in these optimal assignments, but not the assignments themselves (note that the number of transition is the same [i.e., minimum possible] in all optimal assignments). In our second goal, we are interested in the number of specific transitions across all optimal assignments, which means that we would have to also look at the optimal assignments themselves.

While our first goal can be readily addressed by the classical algorithm of Fitch (Fitch 1971) and Hartigan (Hartigan 1973) or using another algorithm by Sankoff (Sankoff 1975), the second goal requires an additional procedure. The reason for this is that the existing algorithms are able to retrieve only one specific optimal assignment in each run through the tree. However, we would ideally like to base our summary statistics on the space of all possible optimal assignments. Since there can be exponentially many optimal assignments, we needed to find an efficient way to extract the summary statistic without actually enumerating all algorithms.

Notably, the Fitch (Fitch 1971) algorithm was originally designed for binary trees. An important property of this algorithm is that the optimal assignments that it produces guarantee optimality even if we consider every sub-tree in isolation. This is different from a scenario where we allow state assignments that may not be optimal when we are considering only a certain sub-tree but become optimal due to compensation elsewhere in the tree. The Hartigan algorithm extends it to non-binary trees and can also be modified such that its optimal assignments remain optimal in every sub-tree. We refer to this modification as the Fitch-Hartigan algorithm. This algorithm operates in a time linear in the input size (i.e., it scales proportionally to $n \cdot k$ where n is the number of cells and k is the number of possible states [tissues, in our case]). The Sankoff algorithm (Sankoff 1975) uses a more involved formulation with a slower run time (it scales proportionally to $n \cdot k^2$) that can account for different penalties to different state transitions. It also returns all possible optimal solutions, including solutions that may not be optimal for every sub-tree, if it is considered in isolation. Here, we employ the Fitch-Hartigan approach due to its simplicity and speed and since we reasoned that it is desirable that our solutions remain optimal, not just for the entire clone, but also for every sub-clone individually.

3 Finding the minimal number of transitions

The Fitch-Hartigan algorithm begins with a "bottom-up" procedure in which labels at the leaves are propagated up to internal nodes in the tree. This "bottom-up" phase assigns a set of labels (tissues) $opt[v]$ to each node v in the tree that satisfy the optimality demands (namely, maximum parsimony). Specifically, for every $s \in opt[v]$ there exists at least one state assignment to the nodes in T^v (the tree rooted by v) where the state of v is s and that is optimal for T^v and for every sub-tree of T^v. Furthermore, the set $opt[v]$ includes all such states. For completeness, we provide a proof for these claims in the appendix (Claim 3).

In addition to generating the sets *opt* the algorithm can also count for every node v the minimum possible number of state transitions $n(v)$ required in the sub-tree rooted by v (Observe that $n(r)$ for the tree rooted at r corresponds to the number of transitions across the entire tree). The "bottom-up" procedure *opt* is applied in **post-order traversal** (evoked by applying it on the root node) with the following pseudocode:

```
function OPT($node = v$)
    if is_leaf(v) then
        return $\{state(v)\}$    ▷ for leaves return their assignment, which was provided as input
    $\forall s \in \Sigma, c(s) = \#\{u \in child(v)$ s.t. $s \in opt(u)\}$
    $k = max_{s \in \Sigma} c(s)$
    $n(v) =\mid child(v) \mid -k + sum_{u \in child(v)} n(u)$
    return $\{s \in \Sigma$ s.t. $c(s) = k\}$
```

Note that in the original formulation by Hartigan, an additional complication is added to increase the number of optimal assignments that can be retrieved by the algorithm. This is done by accumulation of an additional set of states $opt_2(v)$ for every node v that can be used to derive solutions that are globally optimal, but are not optimal in the sub-tree rooted by v. We therefore excluded this part of the algorithm (see appendix for proof [Claim 3]).

4 Inferring the frequency of different state transition events

The second part of the Fitch-Hartigan algorithm is a top-down procedure for finding one (out of potentially many) optimal solution, i.e., a labeling $state : V \rightarrow \Sigma$ of each node $v \in V$ in the tree with a state $s \in \Sigma$. It starts by randomly selecting a state for the root node r out of the set $opt(r)$ and then continues to select legal (see Definition 1) states for child nodes, based on the value assigned to their parent.

```
function STATE-ASSIGNMENT($node = v$)
    if is_root(v) then
        $state(v)$ = random selection out of $opt(v)$
    else
        if $state(parent(v)) \in opt(v)$ then $state(v) = state(parent(v))$
        else$state(v)$ = random selection out of $opt(v)$
```

The above procedure can face many ties during its execution, and can thus potentially return all optimal solutions (provided that they remain optimal for every sub-tree) if it is applied many times. When we compare Fitch-Hartigan to FitchCount in the main text, we use one top-down round (i.e., consider one optimal solution) for the former.

The *FitchCount* procedure was designed to provide a comprehensive evaluation of state transition frequencies, by basing its estimation on the space of all possible optimal state assignments, instead of only a single or few optimal assignments. Compared to the naive approach to enumerate all possible optimal state assignments given by the Fitch-Hartigan algorithm (which may require exponential number of executions), *FitchCount* performs in $\mathcal{O}(n \cdot k^3)$ time for a tree with n leaf nodes and k possible states (assuming each internal node has at least two child nodes, which is the case in our work).

4.1 The algorithm

We define several arrays for storing necessary information:

1. $opt[v]$: The set of optimal assignments for a node v given by the Fitch-Hartigan bottom up approach procedure (defined in the algorithm opt).
2. $N[v, s]$: The number of optimal solutions below the node v given that it takes on the state s.
3. $C[v, s, s_i, s_j]$: The number of transitions from state s_i to state s_j in all optimal solutions of the tree rooted at v, given that v takes on the state s.
4. $M[i, j]$: The number of transitions between s_i and s_j observed across all optimal solutions to the Fitch-Hartigan algorithm.

Our overall objective is to fill in the dynamic programming matrix M, which will subsequently require knowledge of the other dynamic programming arrays. Note that in the following we refer to the arrays using either rounded parenthesis or rectangular parenthesis. The former denotes a function call and the latter denotes a retrieval of an already-computed entry (which we assume to get populated automatically after the respective function call and available globally to the algorithms). Our Main function is:

```
function MAIN($tree = T$, $states = \Sigma$).
    $r$ = root of $T$
    Call opt(r)                ▷ Note that opt(r) fills out the array opt in post-order from the root
    for all $s \in opt[r]$ do
        Call $N(r, s)$
    for all $s \in opt[r]$ do
        for all $\{s_i, s_j\} \in \Sigma^2$ do
            Call $C(r, s, s_i, s_j)$
    for all $\{s_i, s_j\} \in S^2$ do
        $M[s_i, s_j] = sum(C[r, :, s_i, s_j])$
    return $M$
```

The following algorithms for filling in specific entries to $N[v, s]$, and $C[v, s, s_i, s_j]$:

```
function N($node = v, state = s$)
    if is_leaf(v) then
        return 1
    $A = []$ ▷ an array storing the number of solutions in each subtree below $v$ given its state $s$
    for all $u \in child(v)$ do
        $LS = \emptyset$                    ▷ The set of legal states for node $u$ given $state(v) = s$
        if $s \in opt[u]$ then
            $LS = \{s\}$
        else
            $LS = opt[u]$
      $A[u] = \sum_{s' \in LS} N(u, s')$
    return $\prod_{u \in child(v)} A[u]$

function C($node = v, state = s, from = s_i, to = s_j$)
    if is_leaf(v) then
        return 0
    $K = []$        ▷ A temporary array to store the number of transitions observed for each child
    for all $u \in child(v)$ do
        $LS[u] = \emptyset$                 ▷ The set of legal states for node $u$ given $state(v) = s$
        if $s \in opt[u]$ then
            $LS[u] = \{s\}$
        else
            $LS[u] = opt[u]$
        $K[u] = \sum_{s' \in LS[u]} C(u, s', s_i, s_j)$
        if $(s_i == s) \wedge (s_j \in LS[u])$ then
            $K[u] += N[u, s_j]$
    return $\sum_{u \in child(v)} \left( K[u] \prod_{u' \in child(v) \setminus \{u\}} \left( \sum_{s' \in LS[u']} N[u', s'] \right) \right)$
```

4.2 Proof

We prove correctness of the dynamic programming matrices N, and C.

Claim 1. *For any node v and state $s \in opt[v]$, $N[v, s]$, is precisely the number of optimal solutions in $T^{(v)}$ where $state(v) = s$.*

Proof. To prove the correctness of the dynamic programming array N, we'll proceed by induction over the height of the tree $height(T) = h$. For convenience, we'll also make use of a temporary dynamic programming array A which stores the number of solutions for each child $c \in child(v)$, aggregated across each possible state of that child given the parent's state s.

Base Case #1, $h = 0$. $N[l, state(l)] = 1$. This relation is trivially true for the case where $h = 0$ and there exists a single leaf l_1 in which case the only solution consists of $state(l) = s$.

Base Case #2, $h = 1$. Consider a tree $T^{(v)}$ rooted at node v, where $child(v) = \{l_1, ..., l_m\}$. We know that each leaf l_i has a single assignment $state(l_i)$ from the definition of the small-parsimony problem. Thus, the number of possible solutions for this tree with $state(v) = s$ is always one, namely with each leaf taking on their only state. Specifically, in this base case, we observe that $A[u] = 1 \forall u \in \{l_1, .., l_m\}$. To show this, we consider two cases:

- If $s \notin opt(l_i)$: $A[l_i] = \sum_{s' \in opt[l_i]} N[l_i, s'] = N[l_i, state(l_i)] = 1$ since $height(T^{(l_i)}) = 0$.
- If $s \in opt[l_i]$: $s == state(l_i)$ and $A[l_i] = N[l_i, s] = 1$ since $height(T^{(l_i)}) = 0$.

and the relation

$$N[v, s] = \prod_{u \in child(v)} A[u] = A[l_1] * ... * A[l_m] = 1$$

thus is correct.

Inductive Hypothesis. For a tree $T^{(v)}$ of height h, and some $s \in opt[v]$, $N[v, s]$ exactly stores the number of optimal solutions in the tree rooted at v.

Inductive Step. Consider a tree $T^{(v)}$ of height $h + 1$ and some state $s \in opt[v]$. We will show that both the array A correctly stores the number of solutions for the child u given $state(v) = s$ and that the relation $N[v, s] = \prod_{u \in child(v)} A[u]$ is correct.

First, we note that for the tree to be globally optimal, for each $u \in child(v)$, $state(u)$ must be s if $s \in opt[u]$; else, any state from $opt[u]$ can be assigned to u as each incurs a cost of 1 to the overall parsimony of the tree (see Claim 3). These choices for "optimal" states are stored in the array $LS[u]$.

Second, we know from our inductive hypothesis that $N[u, s']$ is correct for any child $u \in child(v)$ and any state $s' \in opt[u]$ as the tree $T^{(u)}$ has a height h. Thus, it is clear that $A[u] = \sum_{s' \in LS[u]} N[u, s']$ correctly returns the number of solutions in the subtree rooted at u over all possible legal states that u can take on.

Finally, we observe that given $state(v) = s$, each child can be treated independently as we consider global solutions that in the tree $T^{(v)}$ with $state(v) = s$. Because of this, the number of such solutions is the size of the permutation of all optimal sub-trees rooted at each $u \in child(v)$ - i.e. the product of all $A[u]$. To show this, consider v has m children. Let the set of optimal internal labellings to $T^{(u_j)}$ be denoted as $\tau_j = \{t_i^{(j)}\}_{i=1}^{A[u_j]}$ where $t_i^{(j)}$ is the i^{th} solution for the tree rooted at u_j given $state(v) = s$. Then, the possible set of solutions is the Cartesian Product between $\tau_1, \tau_2, ..., \tau_m$:

$$\tau_1 \times ... \times \tau_k = \left\{ \{t_1^{(1)}, t_1^{(2)}\}, \{t_2^{(1)}, t_1^{(2)}\}, ..., \{t_{A[u_1]}^{(1)}, t_1^{(2)}\}, ..., \{t_1^{(m-1)}, t_2^{(m)}\}, ..., \{t_{A[u_{m-1}]}^{(m-1)}, t_{[A[u_m]}^{(m)}\} \right\}$$

This Cartesian product has a size $|\tau_1| \times |\tau_2| ... \times |\tau_m| = A[u_i] \times ... \times A[u_k] = \prod_{u \in child(v)} A[u]$. Thus, this relation holds $T^{(v)}$ where $height(T^{(v)}) = h + 1$. □

Claim 2. *For any node v and state $s \in opt[v]$ assigned to v and $\{s_i, s_j\} \in 2^{\Sigma}$, the array $C[v, s, s_i, s_j]$ correctly stores the number of transitions from $s_i \to s_j$ in $T^{(v)}$.*

Proof. We will prove by induction over the height of the tree, h, that for a node v, a state $state(v) = s$, and some $(s_i, s_j) \in 2^{\Sigma}$ both $K[u] \forall u \in child(v)$ and

$$C[v, s, s_i, s_j] = \sum_{u \in child(v)} \left(K[u] \prod_{u' \in child(v) \setminus \{u\}} \left(\sum_{s' \in LS[u']} N[u', s'] \right) \right)$$

are correct. Here $K[u]$ is the number of transitions from s_i to s_j that exist by considering child u of node v given $state(v) = s$ and $LS[u]$ is a function that finds the set of legal (Definition 1) assignments to u given the parent's state is s.

Base Case #1, $h = 0$. The relation trivially holds for a tree of height 0 as there cannot exist any transitions for a tree without edges. As calculated, $C[v, s, s_i, s_j] = 0$ for all leaves and thus the relation holds.

Base Case #2, $h = 1$. Consider a tree of height 1, $T^{(v)}$, where $child(v) = \{l_1, ..., l_m\}$ and that $N[l_i, state(l_i)] = 1$. We can count the number of transitions by considering for every edge the following:

- $s \neq s_i$ then $C[v, s, s_i, s_j]$ is necessarily 0.
- $s == s_i$, then $C[v, s, s_i, s_j]$ is the number of leaves that have state s_j.

By construction, for some child l_i, $C[v, s, s_i, s_j]$ must be

$$K[l_i] = 1[s == s_i \wedge s_j \in LS[u]] == 1[s == s_i \wedge s_j == state(l_i)]$$

Then, $C[v, s, s_i, s_j] = \sum_{l \in child(v)} K[l_i]$. We'll prove that this is equal to the relation described above:

$$\begin{aligned} C[v, s, s_i, s_j] &= \sum_{l \in child(v)} \left(K[l] \prod_{l' \in child(v) \setminus \{l\}} \left(\sum_{s' \in LS[u']} N[u', s'] \right) \right) \\ &= \sum_{l \in child(v)} \left(K[l] \prod_{l' \in child(v) \setminus \{l\}} 1 \right) \\ &= \sum_{l \in child(v)} K[l] \end{aligned}$$

Thus, our relation holds for $h = 1$.

Inductive Hypothesis. Assume for a tree $T^{(v)}$ of height h where $state(v) = s$, $C[v, s, s_i, s_j]$ correctly computes the number of transitions from $s_i, s_j \in 2^{\Sigma}$ for the tree.

Induction Step. Now consider a tree of height $h + 1$ rooted at v where $state(v) = s$. We'll show that both $K[u]$ is correct for all $u \in child(v)$ and that the relation for calculating $C[v, s, s_i, s_j]$ holds.

We'll first show that $K[u]$ is correct $\forall u \in child(v)$. As defined above, $K[u]$ is the number of $s_i \rightarrow s_j$ transitions that are due to the node u given $state(v) = s$. We know that given our inductive hypothesis, for the subtree rooted at u, $T^{(u)}$, $C[u, s', s_i, s_j]$ for any state $s' \in opt[u]$ is correct. Then, the number of $s_i \rightarrow s_j$ transitions under u, given $state(v) = s$, is equal to the sum of all $C[u, s', s_i, s_j]$ for those $s' \in LS[u]$ as those are the only solutions that would be considered by the Fitch-Hartigan algorithm (note that we are guaranteed to have optimal state assignments to chose from for $LS[u]$ as we prove in Claim 3), plus the transition (if it exists) from v to u.

Now, we'll show that the relation for $C[v, s, s_i, s_j]$ holds. For the tree $T^{(v)}$ let's assume that v has m children: $child(v) = \{u_i\}_{i=1}^{m}$. As above, we'll maintain the notation that the set of legal assignments given that $state(v) = s$ for u_i to be $LS[u_i]$. Furthermore, let the set of $s_i \rightarrow s_j$ transitions underneath u_j be $\rho_j = \{r_i^{(j)}\}_{i=1}^{K[u_j]}$ and the set of trees that are legal, optimal assignments under u_j be $\tau_j = \{t_i^{(j)}\}_{i=1}^{\Lambda(u_j)}$ where $\Lambda(u_j) = \sum_{s' \in LS[u_j]} N[u_j, s']$, assuming $state(v) = s$. We can see then that the total number of transitions from $s_i \to s_j$ is

$$\begin{aligned}
&\Big\{\{r_1^{(1)}, t_1^{(2)}\}, \{r_2^{(1)}, t_1^{(2)}\}, ..., \{r_{K[u_1]}^{(1)}, t_1^{(2)}\},\\
&\{r_1^{(1)}, t_2^{(2)}\}, ..., \{r_{K[u_1]}^{(1)}, t_{\Lambda(u_2)}^{(2)}\}, ..., \{r_{K[u_1]}^{(1)}, t_{\Lambda(u_m)}^{(m)}\},\\
&\{r_1^{(2)}, t_1^{(1)}\}, ..., \{r_{K[u_2]}^{(2)}, t_{\Lambda(u_m)}^{(m)}\},\\
&\{r_1^{(m)}, t_1^{(1)}\}, ..., \{r_{A[u_m]}^{(1)}, t_1^{(2)}\}, \{r_{A[u_m]}^{(m)}, t_{\Lambda(u_{m-1})}^{(m-1)}\}\Big\}
\end{aligned}$$

Which is equal to the sum of the following Cartesian Products:

$$\rho_1 \times \{(\tau_2, ..., \tau_m)\} + \rho_2 \times \{(\tau_1, \tau_3, ..., \tau_m)\} + ... + \rho_m \times \{(\tau_1, ..., \tau_{m-1})\}$$

where the cardinality of this set is

$$K[u_1] \prod_{i \in 2..m} \Lambda(u_i) + K[u_2] \prod_{i \in 1,3,..m} \Lambda(u_i) + ... + K[u_m] \prod_{i \in 1,...,m-1} \Lambda(u_i)$$

which can be further simplified to

$$\sum_{u \in child(v)} K[u] \prod_{u' \in child(v) \setminus \{u\}} \sum_{s' \in LS[u']} N(u', s')$$

Thus the relation is correct and $C[v, s, s_i, s_j]$ is correct by induction.

□

5 Appendix

Definition 1. *(Legal Assignment). An assignment $state(v) = s$ is **legal** for a node v and given $state(parent(v)) = s'$ if either $s == s'$ or $s' \notin opt(v)$. Observe that only legal assignments are explored in the Fitch-Hartigan algorithm, and are guaranteed to be optimal in the sub-tree rooted at v.*

Claim 3. *Consider any node v and let $T(v)$ denote the sub-tree rooted at v. The bottom up procedure above returns a set $opt(v)$ such that for every $s \in opt(v)$: (1) there exists a solution (state assignment) that is optimal for $T(v)$ and every sub-tree of $T(v)$, in which the state of v is s; and (2) there does not exist a solution that is optimal for $T(v)$ and for every sub-tree of $T(v)$, in which the state of v is some $s' \notin opt(v)$*

Proof. Proof by induction on tree height h (max length from root to any leaf).

Base Case #1, $h = 0$. In this case the tree consists of a single leaf node, for which state assignment is already fixed and the claim follows trivially.

Base Case #2, $h = 1$. In this case, we have one internal node v with n leaves as its immediate descendants. Here, we define $opt(v)$ as the set of all states that are found in k out of n descendants, where k is maximal. Clearly, the value of the optimal solution in this case has $n - k$ state transitions, which can be obtained by assignment of any state in $opt(v)$ to v. Furthermore, the solution is trivially optimal for every sub-tree (i.e., singleton), thus proving the first part of the claim. Furthermore, assignment of any $s' \notin opt(v)$ to v will necessarily entail strictly more than

$n - k$ state transitions, thus proving the second part of the claim.

Inductive step. Assume an internal node v whose corresponding sub-tree is of height h. Let C denote the set of its child nodes. From the induction, we assume that the claim holds for every node $u \in C$. $opt(v)$ is defined as the set of all states that are found in k out of the $m =\mid C \mid$ child nodes, where k is maximal. First let us denote by $optval(v)$ the value (number of state transitions) of the optimal solution for $T(v)$. From the assumption of the induction, it is easy to see that $optval(v) = sum_{u \in C} optval(u) + m - k$. This value can be reached following the top-down procedure of the Fitch- Hartigan algorithm: (i) assign v with some state $s \in opt(v)$; (ii) assign s to all child nodes u where $s \in opt(u)$ (iii) assign each remaining child node u' with some other state from $opt(u')$ (iv) consider some optimal solution for each of the sub-trees that are rooted by the child nodes. Note that these optimal solutions must exist due to the assumption of our induction. Clearly, the solution that we built satisfies the first part of our claim. For the second part of our claim, assume by contradiction that there exists a state assignment in which the state of v is $s' \notin opt(v)$ that achieves optimality for $T(v)$ and all of its sub-trees. However, from the assumption of the induction we must choose an assignment for every child u out of its set $opt(u)$. It therefore follows that the value of any such solution must be at least $sum_{u \in C} optval(u) + m - k + 1$. We note that in the original paper by Hartigan, a solution is possible where for one or more child nodes u we select an assignment that is identical to the state of the parent v but is not from $opt(u)$. However, while this solution reaches optimality for $T(v)$, it will not be optimal for $T(u)$.

□

References

Chan, Michelle M, Zachary D Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas M Norman, Britt Adamson, Marco Jost, Jeffrey J Quinn, Dian Yang, Matthew G Jones, et al. 2019. "Molecular recording of mammalian embryogenesis." *Nature* 570 (7759): 77–82.

Deshwar, Amit G, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. 2015. "PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors." *Genome biology* 16 (1): 35.

Fitch, Walter M. 1971. "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology." *Systematic Biology* 20, no. 4 (December): 406–416. ISSN: 1063-5157. doi:10.1093/sysbio/20.4.406.

Hartigan, J. A. 1973. "Minimum Mutation Fits to a Given Tree." *Biometrics* 29 (1): 53–65. ISSN: 0006341X, 15410420. http://www.jstor.org/stable/2529676.

Jones, Matthew G., Alex Khodaverdian, Jeffrey J. Quinn, Michelle M. Chan, Jeffrey A. Hussmann, Robert Wang, Chenling Xu, Jonathan S. Weissman, and Nir Yosef. 2020. "Inference of single-cell phylogenies from lineage tracing data using Cassiopeia." *Genome Biology* 21 (1): 92. ISSN: 1474-760X. doi:10.1186/s13059-020-02000-8. https://doi.org/10.1186/s13059-020-02000-8.

Joy, Jeffrey B., Richard H. Liang, Rosemary M. McCloskey, T. Nguyen, and Art F. Y. Poon. 2016. "Ancestral Reconstruction." *PLOS Computational Biology* 12, no. 7 (July): 1–20. doi:10.1371/journal.pcbi.1004763. https://doi.org/10.1371/journal.pcbi.1004763.

El-Kebir, Mohammed, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. 2015. "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data." *Bioinformatics* 31 (12): i62–i70.

El-Kebir, Mohammed, Gryte Satas, and Benjamin J. Raphael. 2018. "Inferring parsimonious migration histories for metastatic cancers." *Nature Genetics* 50 (5): 718–726. ISSN: 1546-1718. doi:10.1038/s41588-018-0106-z.

McPherson, Andrew, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, et al. 2016. "Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer." *Nature genetics* 48 (7): 758.

Reiter, Johannes G., Alvin P. Makohon-Moore, Jeffrey M. Gerold, Ivana Bozic, Krishnendu Chatterjee, Christine A. Iacobuzio-Donahue, Bert Vogelstein, and Martin A. Nowak. 2017. "Reconstructing metastatic seeding patterns of human cancers." *Nature Communications* 8 (1): 14114. ISSN: 2041-1723. doi:10.1038/ncomms14114.

Sankoff, David. 1975. "Minimal Mutation Trees of Sequences." *SIAM Journal on Applied Mathematics* 28 (1): 35–42. doi:10.1137/0128004.

Slatkin, Montgomery, and Wayne P Maddison. 1989. "A cladistic measure of gene flow inferred from the phylogenies of alleles." *Genetics* 123 (3): 603–613.