Title: Continuous lineage recording reveals rapid, multidirectional metastasis in a lung cancer xenograft model in mouse

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

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.

*The authors contributed equally to this work.

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

One Sentence Summary: Detailed single-cell phylogenies capture the frequency, tissue routes, and seeding patterns of metastasis *in vivo*.

Abstract:

Consequential events in cancer progression are typically rare and occur in the unobserved past. Detailed cell phylogenies can capture the history and chronology of such transient events – including metastasis. Here, we applied our Cas9-based lineage tracer to study metastatic progression in a lung cancer xenograft mouse model, revealing the underlying rates, routes, and patterns 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 highly aggressive populations, and these differences are associated with characteristic changes in transcriptional state, including differential expression of metastasis-related genes like *IFI27* and *ID3*. We further show that metastases transit via tissue routes that are diverse, complex, and multidirectional, and identify examples of reseeding, seeding cascades, and parallel seeding topologies. More broadly, we demonstrate the power of next-generation lineage tracers to record cancer evolution at high resolution and vast 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 the 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 important step in cancer progression to study because it is chiefly responsible for disease relapse and mortality (*7*). Yet because metastatic events are intrinsically rare, transient, and stochastic (*8*, *9*), they are challenging 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 critical 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, these 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*); some of these technical limitations can be mitigated using single-cell resolution measurements or whole-genome sequencing.

The recent development of Cas9-enabled lineage tracing techniques with single-cell RNA readouts (*21–24*) provides the potential to explore cancer progression at vastly larger scale and finer resolution than was previously possible. These methods most commonly rely on similar technical principles (reviewed in (*25*, *26*)). Briefly, Cas9 targets and cuts a defined genomic locus (i.e., "scratch pad" or "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 each individual cell by sequencing and paired with single-cell expression profiles of the cell state (*21*, *22*). Then, as in retrospective tracing approaches, various computational approaches (*27–32*) can reconstruct a phylogenetic tree that best models subclonal cellular relationships (e.g., by maximum-parsimony) from the observed 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 (*22*, *33*) and neural development in zebrafish (*21*). Additionally, resources now exist for studying other phylogenetic processes in mouse (*22*, *33*), and analytical tools are available for computationally reconstructing and benchmarking trees from large lineage tracing datasets (*32*, *34*).

Here we apply lineage tracing to explore metastatic dynamics in an orthotopic xenograft model of lung cancer in mice (*35*). Specifically, we have modified our previously described "molecular recorder" for lineage tracing (*22*), 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 (*32*), 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 rate, transcriptional signatures, and routes of metastatic dissemination 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 a mouse lung orthotopic xenograft model because this system is characterized by aggressive metastases (*35*). Orthotopic xenografting experiments such as this are useful for modeling cancer progression *in vivo* (Francia et al. 2011). We engineered A549 cells with a refined version of our lineage tracing technology (*22*) (**Fig. S2**; Methods). Specifically, the engineered cells contain: (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 are expressed and can be captured by single-cell RNA-sequencing, and finally (iv) triple-sgRNAs to direct Cas9 to the three cut-sites in the Target Sites, thereby initiating lineage recording (**Fig. 1A; Fig. S2A–C**). To enable tracing over the timescale of months, we carefully designed the sgRNAs with nucleotide mismatches to the Target Sites, thus decreasing their affinity (*36*, *37*) and slowing the lineage recording rate (*22*, *38*). Approximately 5,000 engineered cells ("A549-LT") were then embedded in matrigel and surgically implanted into the left lung of an immunocompromised (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 tumorous tissues were identified by GFP-positive foci. Anatomically, small tumors studded all five lung lobes, tumor cells predominated the mediastinal lymph tissue, and a small tumor nodule (likely lymphatic) was found on the liver (**Fig. 1D**). This tissue distribution is consistent with previous studies involving A549 xenografts (*35*). From these tissues, we collected six samples: one from the left lung (including the primary site), two from lobes of the right lung, two from the mediastinum, and one from the liver (**Fig. 1E, left**). The tumor samples were dissociated, fluorescence-sorted to exclude normal mouse cells, and finally processed for emulsion-based 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 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 from two cohort experiments (called "M10k", "M100k", and "M30k"), each injected with varying numbers of A549-LT cells that were engineered with different versions of the lineage tracing technology (**Figs. S17–S18;** Methods). We focus our discussion of the results on mouse M5k mouse because it yielded the richest lineage tracing dataset (i.e., the most cells and distinct lineages). However, as discussed, the key results described below are reproduced in the other collected mice (**Figs. S19–S20**).

Distinguishing clonal cancer populations

Our lineage recorder "Target Site" (*22*) carries two orthogonal units of lineage information: (i) a static 14bp-randomer barcode ("intBC") that is unique and distinguishes between the 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 strong 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 (*32*). 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 (i.e., 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. Based on sequencing a sample of A549-LT cells pre-implantation, we estimate that the implanted pool of 5,000 cells initially contained 2,150 distinguishable clones (**Fig. S2D**). We assigned the vast majority of cancer cells collected post-sacrifice (97.7%) to the largest 100 clonal populations based on their intBC sets (**Figs. S4A-B**), ranging in size from >11,000 (Clone #1, “CP001”) to ~30 cells (CP100) (**Fig. S4C**). Though there were some smaller clonal populations, we focused on the largest ones because lineage tracing in few cells is less informative. Notably, the observation of only ~100 successful clonal populations in mouse tumors (and the absence of the vast majority of initial clones) suggests that a minority of cells may be competent for engraftment and survival *in vivo* (as low as 5%; **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* (*39*, *40*).

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. S5A-C**). Though most clonal populations exhibited adequate parameters 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; **Fig. S5D**); these clones were excluded from tree reconstruction and downstream analysis (Methods).

Clonal tissue distributions and metastatic potential

We observed that the clonal populations exhibited distinct distributions across the six tumorous tissues (**Fig. 1F**; **Fig. S6**), 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 direct consequence of metastatic spread 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. Some populations’ tissue distributions are more difficult to interpret, such as CP022 which resides entirely within the right lung and has no relatives in the primary site. This tissue distribution may have resulted from an early metastasis from the primary to the right lung, followed by the extinction of cells remaining at the primary site; alternatively, it is possible that we did not sample cells from this clonal population in the left lung, presumably because of their rarity. Separately, we cannot exclude the possibility that tissue samples carry hematogenous cancer cells, though these are likely rare (*41*) and therefore would not contribute significantly to differences in tissue distribution.

In silico modeling of the metastatic process indicates that tissue dispersal reports on the underlying metastatic rate, albeit imperfectly (e.g., it saturates at intermediate metastatic rates; **Fig. S7B**). 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 phenotype. Across the 100 clonal populations in this mouse, we observed a wide range of Tissue Dispersion Scores (**Fig. 1G**), suggesting wide metastatic heterogeneity. We next explored this metastatic heterogeneity more directly and at far greater resolution using the evolving lineage information.

Reconstructed cancer cell phylogenies

The key advantage of our lineage tracer is not in following *clonal* lineage dynamics (i.e. from cells' static intBCs, as described above) but rather in reconstructing *subclonal* lineage dynamics (i.e. from cells' continuously evolving indel alleles, as in retrospective approaches). As such, we next reconstructed high-resolution phylogenetic trees using the Cassiopeia suite of phylogenetic inference algorithms (*32*) with modified parameters tailored to this dataset's unprecedented complexity and scale (Methods). The resulting trees comprehensively describe the phylogenetic relationships between the cells within each clonal population (**Fig. 2A**). The trees are intricately complex (mean tree depth 7.25; **Fig. S5E**) and highly resolved (consisting of 37,888 cells with 33,266 (87.8%) unique lineage allele states; **Fig. S5C**).

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, inlays**). Indeed, we find systematic agreement between phylogenetic distance (the distance between two cells in the tree) and allelic distance (the difference between two cells' alleles) for this example (**Fig. 2C**) and across all other trees (**Fig. S8**), thus supporting their accuracy. The high diversity of indel alleles here (9,936 unique indels across all cells in mouse M5k; represented by the array of unique colors in the "character matrix", **Fig. 2B**) also reduces the probability of homoplasy, an issue that can complicate tree reconstruction and diminish tree accuracy (*32*, *42*). Altogether, these features indicate that the reconstructed trees accurately model the true phylogenetic relationships between cells.

Quantification of 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 ancestral cells physically transitioning from one tissue to another (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 inhabiting 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 (*43*, *44*). This algorithm provides the minimal number of ancestral (unobserved) metastatic transitions that are needed to explain the observed tissue assignment of cells in a given tree. We defined a score of the metastatic potential (termed "Tree MetRate") by dividing the inferred minimal number of metastatic transitions by the 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) Tree MetRates (**Fig. 3B**). The Tree MetRate is stable across bootstrapping experiments in simulated trees (**Fig. S7E-F**) and when using an alternative phylogenetic reconstruction method (i.e., Neighbor-Joining (*27*); **Fig. S9A**; Pearson's ρ=0.94), indicating that the Tree MetRate is a robust measurement of metastatic behavior – though, notably, Cassiopeia trees are more parsimonious than those reconstructed by Neighbor-Joining (**Fig. S9B**). Though the Tissue Dispersal Score Tree agrees with Tree MetRate at low metastatic rates (**Fig. 3C**), the Tree MetRate more accurately captures the underlying metastatic rate over a broad range of simulated metastatic rates (**Fig. S7D**) because it can distinguish between moderate and high metastatic rates, which both result in broad dispersion across tissues (**Fig. S7B;** e.g., **Fig 3B**). Furthermore, the Tree MetRate 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 "Allele MetRate"; **Fig. 3D**); importantly, the Allele MetRate 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 Tree MetRate is a superior measurement of the underlying metastatic rate (**Fig. S7A-D**), underscoring the value of the reconstructed phylogenies in helping identify aspects of metastatic behavior that would otherwise be invisible.

Though metastatic phenotype appeared to be generally consistent within a clonal population, there may have been subclonal variations. To this end, we extended our parsimony-based approach to quantify the metastatic phenotype at the resolution of single-cells (termed the "single-cell Tree MetRate") by averaging the Tree MetRate for all subclades containing a given cell (Methods). Importantly, this measurement is sensitive to subclonal differences in metastatic behavior (**Fig. 3E**), such as the bimodal metastatic rates observed in CP007 cells (discussed below; **Fig. 4E**). Additionally, we find that the single-cell Tree MetRate is uncorrelated to clonal population size, proliferation signatures (*45*, *46*), or cell cycle stage (*47*) (**Fig. S10**), 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.

Transcriptional signatures of distinct metastatic phenotypes

By overlaying the single-cell transcriptional information and the cell phylogenies, we found that different metastatic behaviors corresponded to differential expression of genes with known roles in metastasis. First, after filtering and normalizing the scRNA-sequencing data, we applied *Vision* (*48*), 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 state and its tissue or single-cell Tree MetRate (**Fig. S11**). To more sensitively identify the transcriptional features of metastatic cells within a single tissue, we next performed pairwise differential expression analyses comparing cells from completely non-metastatic clonal populations (i.e., four that never metastasized from the primary tissue in the left lung, like CP029) to all other cells observed in the left lung (**Fig. 4A**). Many genes, such as IFI6, exhibited significant expression changes that were consistent across each non-metastatic clone (log2 fold-change > 1.5, FDR < 0.01). This suggests that differences in metastatic phenotype are manifested in characteristic differences in transcriptional state. Importantly, this differential expression analysis is limited to cells from a single tissue (i.e., the left lung), so gene expression differences we identified are unlikely due to tissue-specific differences.

Next, we more comprehensively identified genes that are associated with metastatic behavior – either positively or negatively – by regressing single-cell gene expression against the single-cell Tree MetRates (over all observed cells, clonal populations, and tissues; **Fig. 4B**; Methods). Many positive hits (i.e., genes with significantly higher expression in highly metastatic cells) have known roles in potentiating tumorigenicity (**Fig. 4C, 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 (*49*, *50*), REG4 enhances cell migration and invasion in colorectal carcinoma (*51*) and KRAS-driven lung adenocarcinoma (*52*), and TNNT1 has elevated expression in many cancers and may promote cell invasion via EMT (*53*). Similarly, many negative hits (genes with significantly lower expression in highly metastatic cells) have known roles in attenuating metastatic potential (**Fig. 4C, bottom**); for example, NFKBIA (IκBα) is a pan-cancer tumor suppressor via inhibition of pro-tumoral NFκB signaling (*54*), overexpression of ID3 inhibits tumor cell migration and invasion *in vitro* and in similar xenograft models of lung adenocarcinoma xenograft (*55*), and downregulation of ASS1 supports tumor metabolism and proliferation (*56*). Paradoxically, our most significant negative hit is KRT17, which has previously been implicated in promoting invasiveness in lung adenocarcinoma (*57*) and its overexpression is associated with poor prognosis in many cancers (*58*). This suggests that KRT17 may play a context-specific role in metastatic progression in this model, which could be further explored experimentally. Indeed, all of the identified genes here are prime candidates for deeper study to elucidate their possible molecular roles in this xenograft model of metastasis. Overall, the gene-level expression trends are broadly supported by significant correlation between the Tree MetRate and several gene expression signatures (*59*) (**Fig. S12**), including interferon signaling programs (*60*), RAS pathways (*61*) (A549 cells are *KRAS*-mutant), cancer invasiveness (*62*), and EMT (*63*) (consistent with increased NFκB signaling (*64*, *65*)). Additionally, the identified gene hits are significantly reproduced across all mice in this study (**Fig. S20**).

Clone #7 (CP007) exhibits exceptionally distinct subclonal metastatic behaviors, wherein one large clade metastasized frequently to other tissues and another large clade remained predominantly in the right lung (**Fig. 4D**). This distinction is reflected in a bimodal distribution of single-cell Tree MetRates (**Fig. 3E**; **Fig. 4E**). To explore the relationship between subclonal structure and gene expression, we applied *Hotspot* (*66*) and identified two modules of correlated genes that exhibit heritable expression programs (**Fig. S13A**). Strikingly, the cumulative expression of genes in one module is correlated with lower metastatic rates, while the opposite holds for the other module (**Fig. 4F; S13B-C**). Consistently, the two modules correspond to the two phylogenetic subclades with diverging metastatic phenotypes (**Fig. 4G**). This result is reproduced even in a control analysis of CP007 cells from the right lung only (**Fig. S13D-G**), indicating that these differences in gene expression indeed reflect differences in metastatic phenotype rather than tissue-specific effects. Thus, this example illustrates that the metastatic phenotype is not an intrinsically immutable characteristic of each clonal population, and that metastatic rate, alongside concordant changes in transcriptional state, can change substantially during tumor development.

Tissue routes and topologies of metastasis

The phylogenetic reconstructions also make it possible to describe detailed histories about the tissue routes and 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. 5A–B**). Other phylogenies reveal 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. 5D–E**). To more systematically characterize the tissue transition routes revealed by the phylogenetic trees, we extended the Fitch-Hartigan algorithm (*43*, *44*) to infer each tissue transition event along a clonal population's ancestry. Our algorithm, called *FitchCount*, builds on other ancestral inference algorithms like MACHINA (*67*) 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. S7G–H**; Methods), likely because the naive approach summarizes only a single optimal solution (i.e., assigning tissues to ancestral nodes to minimize the number of transitions), whereas *FitchCount* summarizes all optimal solutions (the number of which scales exponentially with the tree size). The resulting conditional probabilities of metastasis to and from each tissue are summarized in transition matrices, suggesting the most probable tissue transition routes in a clone's past. Notably, we found that the transition matrices are varied and distinct to each clone (**Fig. 5C, F, G; Fig. S14**). We next used principal component analysis (PCA) to classify clones by their transition matrices (**Fig. 5H**) and identified descriptive features that capture differences in the metastatic routes traversed by each clone (**Fig. 5I; Fig. S15**). These features include primary seeding from the left lung (as in CP095, **Fig. 5A–C**), metastasis from and within the mediastinum (CP098, **Fig. 5G, left**), or metastasis between lung lobes (CP070, **Fig. 5G**), and may reflect intrinsic or stochastic differences in tissue tropism. From this feature analysis we note that many clones primarily metastasized via the mediastinal lymph tissue (**Fig. 5H–I**), suggesting that the mediastinum may act as a nexus for metastatic dissemination in this model. This observation is consistent with past experiments in this model (*35*), 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. 1E**).

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. 5J**). These different metastatic topologies can critically influence the progression, relapse, and treatment of cancers (*9*, *68–70*); for example, reseeding of metastatic cells returning to the primary tumor site can contribute genetic diversity, resistance to treatment, and metastatic potential to tumors (*71*, *72*). Within this single dataset, we find numerous examples of all of these topologies (**Fig. 5K**); in fact, we most often observe examples of all topologies within every clone (**Fig. S16**), as well as more complex topologies that defy simple classifications (e.g., **Fig. 5D** 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, tracking metastatic spread at unprecedented resolution. These observations were made possible by experimental and algorithmic advances that we made on our "molecular recorder" platform (*22*). 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. 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 reveals the frequency and directionality of metastatic dissemination that would not be 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 find surprisingly diverse metastatic phenotypes, ranging from metastasis-incompetent to highly metastatic, which are heritable and correspond to distinct, reproducible transcriptional states involving important hallmarks of metastasis; these transcriptional differences nominate these genes as candidates for further study to determine their possible molecular roles in metastasis. (3) Metastatic dissemination is rapid, frequent, and complex in this model, transiting via different complicated seeding topologies, such as seeding cascades, parallel seeding, and more. Furthermore, we illustrate 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 (*73*). 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 important 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 (*74*), or (iii) patient-derived xenograft (PDX) models (*75*, *76*). Second, this xenograft model requires immunodeficient host mice, and therefore does not reflect the pervasive influence wielded by the immune system on natural cancer progression (*77*–*79*); lineage tracing in syngeneic lines or spontaneous models of cancer could chart how an intact immune system affects cancer progression. Third, it would be valuable to test how transcriptional differences relates to (or determine) metastatic capacity, for example by genetically perturbing the candidate metastasis-related gene targets identified here via Cas9 knock-out or CRISPRi knock-down (*80*, *81*). These perturbations could be executed alongside lineage tracing, allowing for the simultaneous readout of transcriptional, phylogenetic, and phenotypic changes *in vivo*. Indeed, it should be possible to perform a pooled screen of a subset of gene candidates in a single perturbation experiment or in a single mouse, using the lineage tracer to demultiplex perturbation conditions (*82*). Fourth, this work describes metastasis at the spatial resolution of tumorous tissues (i.e., not individual tumors) because we bulk together tissues that 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 (*23*, *83–85*); 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. 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).

22. 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).

23. 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).

24. 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).

25. 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).

26. 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.

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

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

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

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

31. 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.

32. 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).

33. 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.

34. 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).

35. 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).

36. 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).

37. 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.

38. 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.

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

40. 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).

41. K. J. Luzzi, I. C. MacDonald, E. E. Schmidt, N. Kerkvliet, V. L. Morris, A. F. Chambers, A. C. Groom, Multistep nature of metastatic inefficiency: dormancy of solitary cells after successful extravasation and limited survival of early micrometastases. *Am. J. Pathol.* **153**, 865–873 (1998).

42. 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.

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

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

45. 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).

46. 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).

47. 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).

48. 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).

49. 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).

50. 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).

51. 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).

52. 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).

53. 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).

54. 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).

55. 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).

56. 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).

57. 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).

58. 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).

59. 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).

60. 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).

61. 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).

62. 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).

63. 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).

64. 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).

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

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

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

68. 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).

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

(2015).

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

71. 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).

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

73. 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).

74. 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).

75. 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).

76. 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).

77. 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).

78. 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).

79. 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).

80. 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).

81. 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).

82. I. P. Winters, C. W. Murray, M. M. Winslow, Towards quantitative and multiplexed in vivo functional cancer genomics. *Nat. Rev. Genet.* **19**, 741–755 (2018).

83. 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).

84. 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).

85. 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).

86. 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).

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

88. 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).

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

90. 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, and UCSF Center for Advanced Technology, 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. **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. 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-S20

Supplementary Text

Supplementary Text

Figures and Figure Legends: (embedded below)

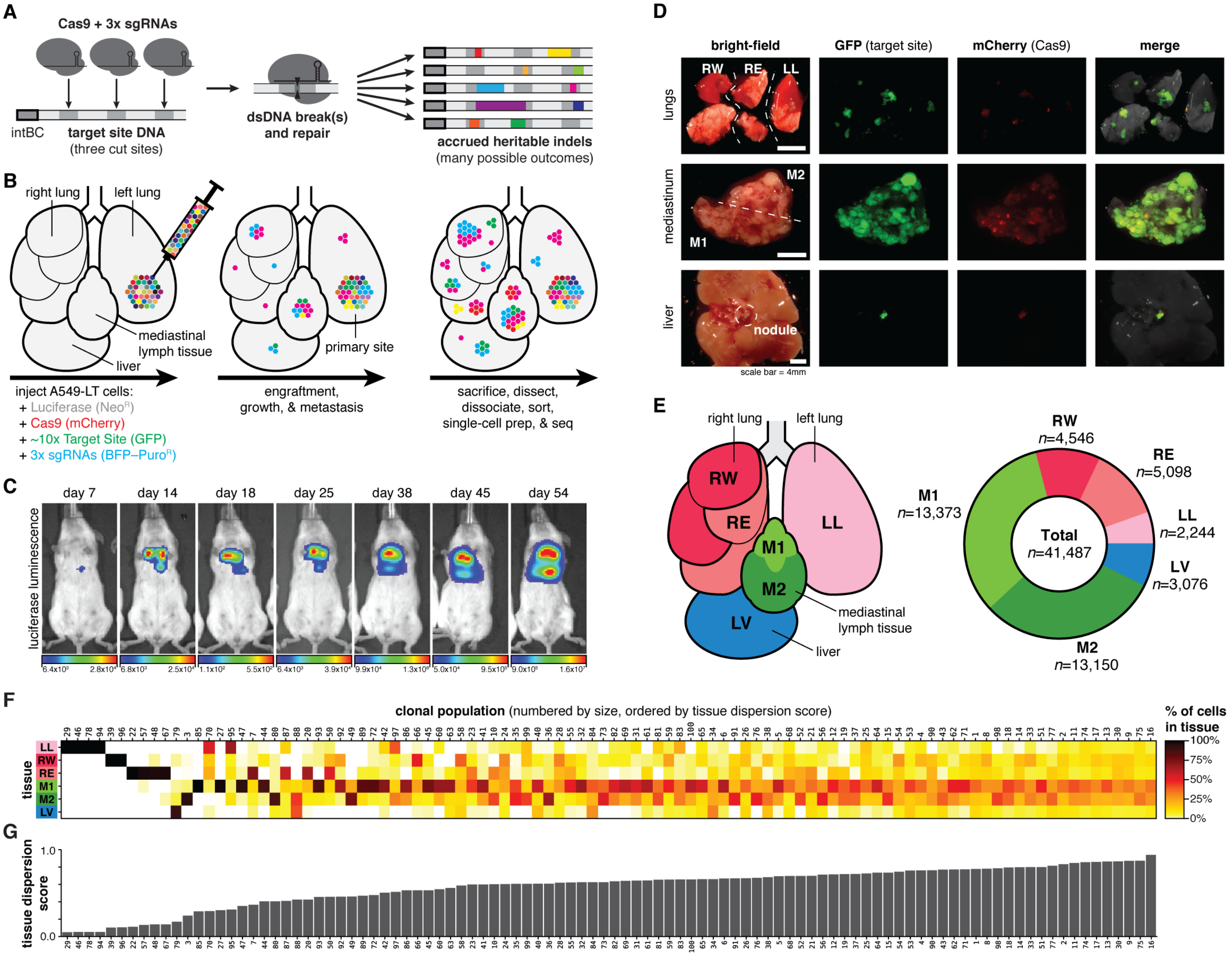

Fig. 1. Lineage tracing in a lung cancer xenograft model in mouse. (**A**) Our Cas9-enabled lineage tracing technology. Cas9 and three sgRNAs bind and cut cognate sequences on genomically integrated Targets Sites, resulting in diverse indel outcomes (multicolored rectangles), which act as heritable markers 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**) Tumorous tissues collected, featuring tumors widespread throughout the lungs and mediastinum and a preponderance of tumor cells in the mediastinum. (**E**) Anatomical representation of the six tumorous tissue samples (**left**), and the number of cells collected with paired transcriptional and lineage datasets (**right**). (**F**) Tissue

distributions of the largest 100 clonal populations. (**G**) The Tissue Dispersion Score is a statistical measurement of the distribution across tissues for each clonal population.

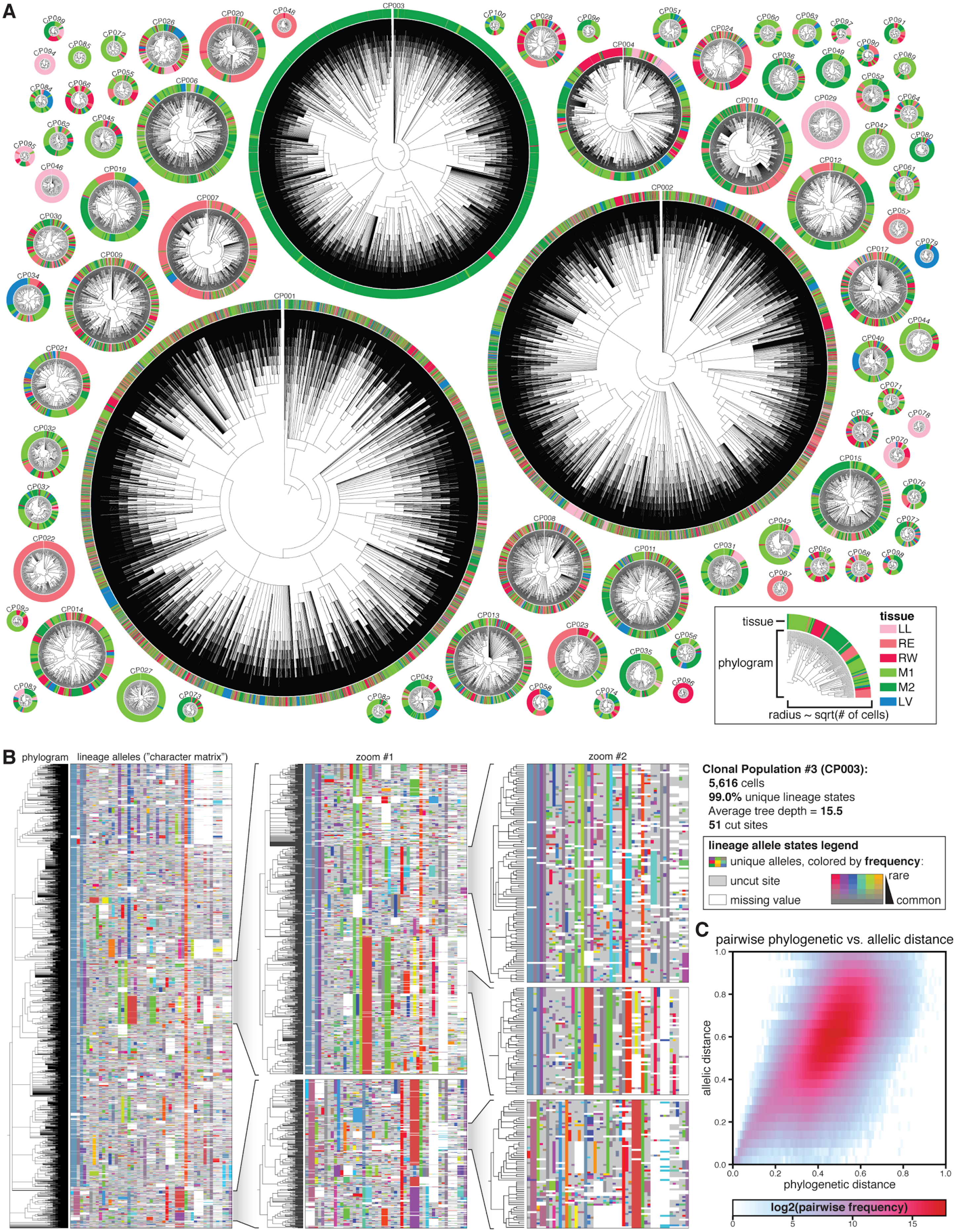
A
CP099
CP020
CP048
CP003
CP100
CP028
CP096
CP051
CP060
CP063
CP097
CP091
CP026
CP072
CP085
CP094
CP024
CP004
CP090
CP036
CP049
CP089
CP084
CP066
CP055
CP006
CP052
CP029
CP064
CP062
CP045
CP010
CP047
CP080
CP095
CP046
CP019
CP012
CP061
CP030
CP007
CP002
CP057
CP079
CP009
CP017
CP034
CP001
CP044
CP040
CP021
CP071
CP054
CP078
CP032
CP070
CP015
CP037
CP076
CP008
CP077
CP022
CP042
CP011
CP031
CP059
CP068
CP098
CP067
CP092
CP014
CP013
CP023
CP056
CP035
CP027
CP083
CP043
CP058
CP096
CP073
CP082
CP074
tissue
phylogram
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**) Phylogenetic reconstructions for each clonal population represented as radial phylograms, with each cell along the circumference colored by tissue as in **Fig. 1E**. Trees are scaled by the square-root of the number of cells in the clonal population. (**B**) Phylogenetic tree and lineage alleles of one clonal population (CP003; *N*=5,616 cells). The phylogram (**left**) represents cell relationships and the color matrix (**right**) represents the lineage alleles for each cell. Alleles are uniquely colored and color saturation represents allele rarity (legend). (**B, inlays**) Nested zooms of individual clades highlight allele state diversity, tree depth, and tree complexity. (**C**) Correlation 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.

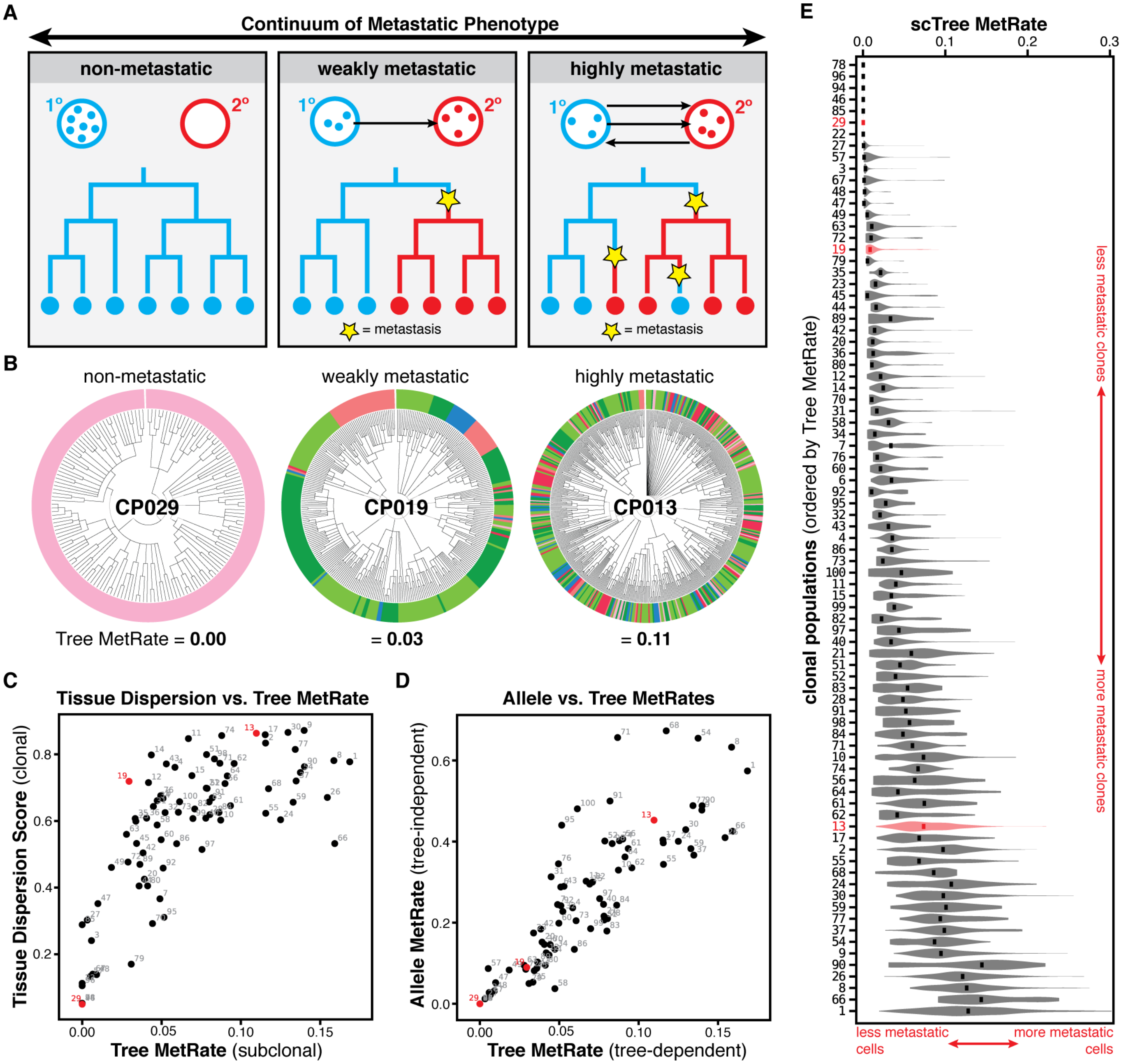

Fig. 3. Phylogenetic trees reveal that clonal populations exhibit diverse metastatic phenotypes. (**A**) Theoretical continuum of metastatic phenotypes, spanning non-metastatic (never exiting the primary site) to highly metastatic (frequently transitioning between tumors; arrows). Ancestral metastatic events between tissues leave clear phylogenetic signatures (yellow stars). (**B**) Three clonal populations 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 Tree MetRate. (**C** and **D**) Comparison of three lineage tracer-derived measurements of metastatic phenotype: "Tissue Dispersion Score" (as in **Fig. 1F**) is a statistical measure of the clone's distribution across tissues; "Allele MetRate" measures the probability that a cell's closest relative by allele state is in a different tissue; and "Tree MetRate" measures the inferred frequency of metastatic transitions from the reconstructed phylogeny. Examples from (B) shown in red. (**E**) The distributions of single-cell-resolution metastatic phenotypes (single-cell Tree MetRates) for each clonal population, ordered by Tree MetRate; median rate indicated in black; examples from (B) shown in red.

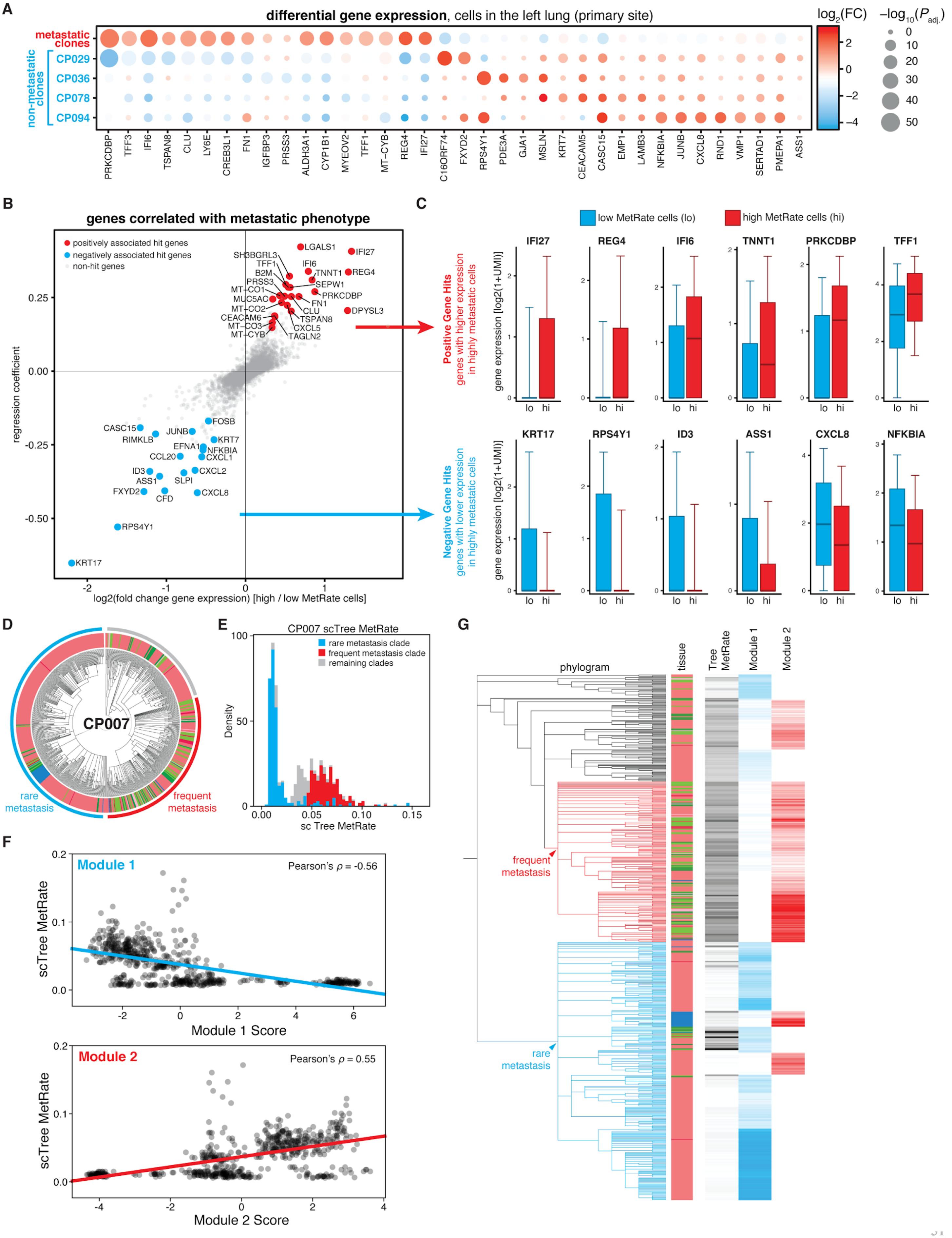
A
differential gene expression, cells in the left lung (primary site)
metastatic clones
non-metastatic clones
CP029
CP036
CP078
CP094
PRKCDBP
TFF3
IFI6
TSPAN8
CLU
LY6E
CREB3L1
FN1
IGFBP3
PRSS3
ALDH3A1
CYP1B1
MYEOV2
TFF1
MT-CYB
REG4
IFI27
C16ORF74
FXYD2
RPS4Y1
PDE3A
GJA1
MSLN
KRT7
CEACAM5
CASC15
EMP1
LAMB3
NFKBIA
JUNB
CXCL8
RND1
VMP1
SERTAD1
PMEPA1
ASS1
$\log_2$(FC)
2
0
−2
−4
$-\log_{10}(P_{adj.})$
0
10
20
30
40
50
B
genes correlated with metastatic phenotype
positively associated hit genes
negatively associated hit genes
non-hit genes
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
CXCL8
CFD
RPS4Y1
KRT17
regression coefficient
0.25
0.00
-0.25
-0.50
-2
-1
0
1
log2(fold change gene expression) [high / low MetRate cells]
C
low MetRate cells (lo)
high MetRate cells (hi)
Positive Gene Hits
genes with higher expression in highly metastatic cells
gene expression [log2(1+UMI)]
IFI27
REG4
IFI6
TNNT1
PRKCDBP
TFF1
Negative Gene Hits
genes with lower expression in highly metastatic cells
KRT17
RPS4Y1
ID3
ASS1
CXCL8
NFKBIA
lo
hi
D
CP007
rare metastasis
frequent metastasis
E
CP007 scTree MetRate
rare metastasis clade
frequent metastasis clade
remaining clades
Density
100
50
0
0.00
0.05
0.10
0.15
sc Tree MetRate
F
Module 1
Pearson's ρ = -0.56
scTree MetRate
0.2
0.1
0.0
-2
0
2
4
6
Module 1 Score
Module 2
Pearson's ρ = 0.55
-4
-2
0
2
4
Module 2 Score
G
phylogram
tissue
Tree MetRate
Module 1
Module 2
frequent metastasis
rare metastasis

Fig. 4. Divergent metastatic phenotypes correspond to differences in gene expression. (**A**) Differential gene expression analysis comparing four non-metastatic clonal populations and all metastatic clonal populations in the primary tumor tissue. 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. (**B**) Poisson regression analysis of gene expression and single-cell Tree MetRate for all cells and all tissues; fold-change and coefficient of regression shown. The strongest and most significant positive and negative gene hits are annotated (red and blue, respectively). (**C**) Expression level of positive and negative gene hits (top and bottom, respectively) in cells with high or low single-cell Tree MetRate (red and blue, respectively). Boxes and whiskers represent first, second, and third quartiles, and 9th and 91st percentiles of expression distribution. (**D**) Divergent subclonal metastatic behavior exhibited in the phylogenetic tree of clonal population #7, with annotated subclades; cells colored by tissue as in **Fig. 1E**. (**E**) The bimodal distribution of single-cell Tree MetRates for cells in CP007, with cells from the divergent subclades indicated. (**F**) Comparison of single-cell metastatic phenotype and *Hotspot* transcriptional module scores. (**G**) Overlay illustrating concordance between CP007 phylogeny, single-cell Tree MetRates, and *Hotspot* Module scores.

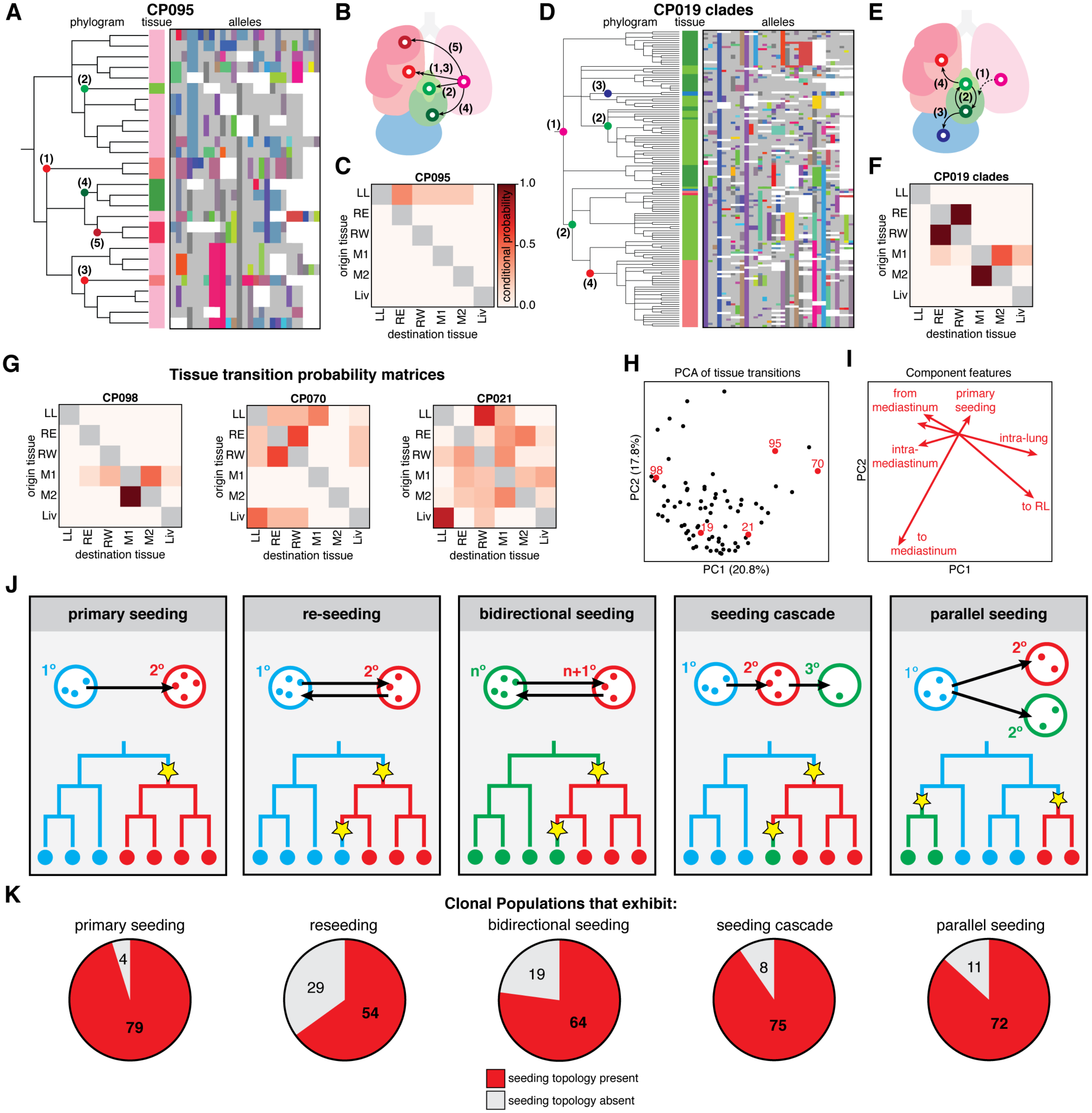

Fig. 5. Metastases follow complex, multidirectional tissue routes and seeding topologies. (**A** and **D**) Phylogenetic trees and lineage alleles for clonal population #95 and #19 clades, respectively. Notable metastatic events are annotated in the phylogram and represented graphically as arrows (**B** 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.