Early and multiple origins of metastatic lineages within primary tumors

Zi-Ming Zhao[a,1,2], Bixiao Zhao[b,1], Yalai Bai[c], Atila Iamarino[a], Stephen G. Gaffney[a], Joseph Schlessinger[d,2], Richard P. Lifton[b], David L. Rimm[c], and Jeffrey P. Townsend[a,e,f,2]

[a]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510; [b]Department of Genetics, Yale School of Medicine, New Haven, CT 06520; [c]Department of Pathology, Yale School of Medicine, New Haven, CT 06520; [d]Department of Pharmacology, Yale School of Medicine, New Haven, CT 06520; [e]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520; and [f]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511

Contributed by Joseph Schlessinger, January 5, 2016 (sent for review July 29, 2015; reviewed by Andrew G. Clark and Allen Rodrigo)

Many aspects of the evolutionary process of tumorigenesis that are fundamental to cancer biology and targeted treatment have been challenging to reveal, such as the divergence times and genetic clonality of metastatic lineages. To address these challenges, we performed tumor phylogenetics using molecular evolutionary models, reconstructed ancestral states of somatic mutations, and inferred cancer chronograms to yield three conclusions. First, in contrast to a linear model of cancer progression, metastases can originate from divergent lineages within primary tumors. Evolved genetic changes in cancer lineages likely affect only the proclivity toward metastasis. Single genetic changes are unlikely to be necessary or sufficient for metastasis. Second, metastatic lineages can arise early in tumor development, sometimes long before diagnosis. The early genetic divergence of some metastatic lineages directs attention toward research on driver genes that are mutated early in cancer evolution. Last, the temporal order of occurrence of driver mutations can be inferred from phylogenetic analysis of cancer chronograms, guiding development of targeted therapeutics effective against primary tumors and metastases.

tumor phylogenetics | ancestral reconstruction | cancer | chronograms | oncogenes

It has long been understood that tumorigenesis is an evolutionary process (1) associated with the accumulation of somatic mutations (2). However, many aspects of that process that are fundamental to cancer biology and targeted treatment have been challenging to reveal, such as the divergence times and genetic clonality of metastatic lineages (3, 4). Somatic mutations have revealed tumor type-specific drivers by comparison of primary tumor and normal tissues (5, 6), and studies examining the evolutionary process of cancer across multiple sites have used a handful of subjects to identify ubiquitous, shared, and private mutations (1) and to reconstruct a number of tumor phylogenies using parsimony or unweighted pair group methods with arithmetic mean (1, 7) but have lacked the power to generalize about the tumorigenic or metastatic process across cancer types (1).

Tumor phylogenetics, using a larger sample with explicit evolutionary models, can be applied using molecular evolutionary models to reconstruct ancestral states of somatic mutations and infer cancer chronograms, revealing novel information about the timing of gene mutations and their contributions to tumorigenesis and metastasis and addressing three fundamental aspects of cancer biology. First, the topology of divergence of primary and metastatic lineages can differentiate between a linear model of cancer progression, in which all metastatic tumors are descended from a single original primary cell such that all metastases are more closely related to each other than they are to any tissue in the primary tumor, and a branched model, in which metastases can originate from divergent lineages within primary tumors. Second, molecular evolutionary trees and chronograms can quantify how early metastatic lineages arise in tumor development, clarifying the role of mutations in facilitating metastasis. Last, integration of temporal inferences across patients can convey the order of occurrence of driver mutations, guiding development of targeted therapeutics effective against primary tumors and metastases.

Here, we perform tumor phylogenetics to address these questions. Although ascertaining variable degrees of tumor heterogeneity (1) by computational analyses of subclonality within primary tumors has proven challenging (8), another approach to revealing heterogeneity is analysis of the sequence divergence of major clones extracted from distant sites. We replayed the "tape of cancer" and mapped genetic mutations on the tree of cancer evolution extending from normal tissue to primary tumor and metastases. Analyzing new exome sequence data from primary and metastatic sites, we applied maximum likelihood and Bayesian approaches to reveal phylogenetic relationships and tumor evolution chronology. We identified genetic mutations associated with tumorigenesis that commonly precede the first genetic divergence of all cancer lineages, also examining those that precede all metastases. Furthermore, we quantified the temporal distributions of the first genetic divergence of metastases from the primary tumor and evaluated the temporal order of gene mutations in cancer.

Results and Discussion

We generated a unique dataset of sequenced normal, primary, and matched metastatic tumor tissues from 40 subjects with 13 types of cancer, including 13 subjects with lung cancer and 7 subjects with pancreatic cancer. We sequenced exomes of normal, primary, and matched metastatic tumor tissues, including

Significance

The knowledge that cancer is an evolutionary process is old, but only recently can sequencing technology provide data for clinically relevant evolutionary analyses of cancer. Approaches developed for evolutionary biology can reveal the relationship among clonal lineages, the ancestral states of gene sequences, and the timing of evolutionary events. We performed whole exome sequencing of cancer tissues from multiple sites of dozens of subjects, demonstrating nonlinear patterns of tumor progression and early origins of metastatic lineages and quantifying the times of occurrence of driver mutations. These findings direct research attention away from the search for genes that induce metastasis toward genes that are mutated early in tumorigenesis, providing therapeutic targets effective against both primary tumors and metastases.

Author contributions: J.S., R.P.L., D.L.R., and J.P.T. designed research; Z.-M.Z., B.Z., Y.B., A.I., S.G.G., and J.P.T. performed research; Z.-M.Z., A.I., S.G.G., D.L.R., and J.P.T. contributed new reagents/analytic tools; Z.-M.Z., B.Z., A.I., S.G.G., and J.P.T. analyzed data; and Z.-M.Z., B.Z., Y.B., A.I., S.G.G., J.S., R.P.L., D.L.R., and J.P.T. wrote the paper.

Reviewers: A.G.C., Cornell University; and A.R., Australia National University.

The authors declare no conflict of interest.

[1]Z.-M.Z. and B.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: Jeffrey.Townsend@Yale.edu, Joseph.Schlessinger@Yale.edu, or Zi-ming.Zhao@Yale.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1525677113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1525677113

32 primary tumors and 139 sites of metastases, ranging from two to seven metastases per subject (Fig. 1*A*, Fig. S1*A*, Dataset S1, and *SI Materials and Methods*). Twenty-four of these primary tumors were identified clinically without ambiguity whereas, in eight lung cancer subjects, the primary tumor was identified with less certainty. Exomes were captured from normal tissues and tumors of sufficient purity and sequenced to an average of nearly 200× coverage per targeted base, 95% of which were covered by more than 20 independent reads (Fig. S1*B* and Dataset S2). After alignment to the reference human genome, comparison of normal and all matched tumor sequences identified 20–5,370 somatic mutations in each subject. Variant calls were tested for a subset of somatic mutations by Sanger sequencing and were 100% validated (Dataset S3).

Tumor Phylogenetics of Multiregion Tumors Revealed the Origins of Metastatic Lineages from Divergent Lineages Within Primary Tumors. We constructed 40 multiple sequence alignments, each alignment including the somatic variants from all tumors within each subject and their matched normal tissue sequence (Dataset S4). To determine the genetic relationship of these tumors, we applied parsimony-based (9), maximum likelihood (10), and Bayesian inference (11) to the multiple sequence alignment, estimating phylogenies of the tumor samples within subjects (Fig. 1*B* and Fig. S2). We then calibrated these phylogenies with tumor type-specific data on tumor cell division times (12) and with the clinical timings of diagnosis, biopsy, resection, and autopsy for each patient (Datasets S5 and S6) to evaluate the evolutionary pattern and tempo of the genetic divergence of primary and metastatic tumor lineages (Fig. 1*C* and Fig. S3).

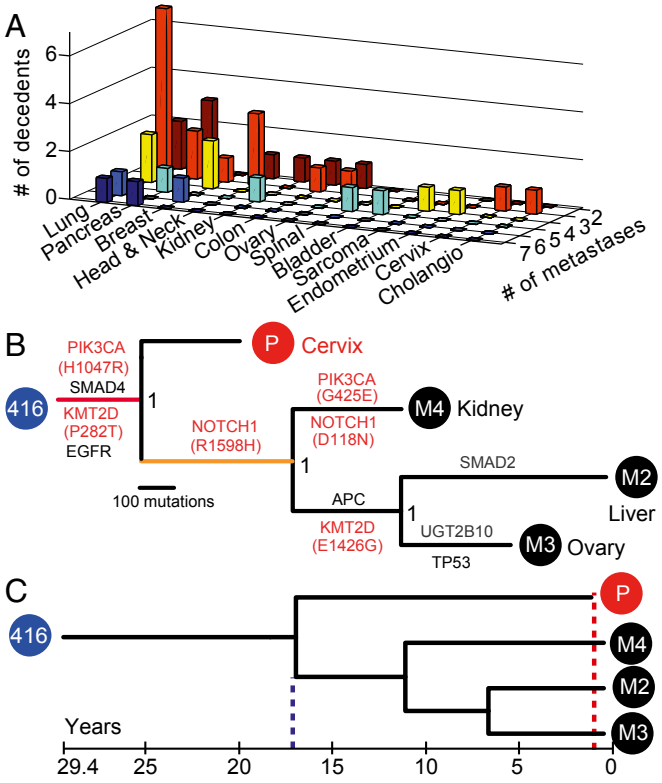

Fig. 1. Tumor samples and methodology. (*A*) Cancer type and number of metastases analyzed for 40 subjects in our study. (*B*) Maximum likelihood tree of normal tissue (blue circle, 416), primary tumor of the cervix (red circle, P), and metastatic tumors (black circles, M2, M3, and M4) from subject 416, with a horizontal scale proportional to the number of mutations, and with all known driver mutations mapped to branches. The red internode represents the lineage ancestral to the primary tumor and all metastases. The orange internode represents the lineage ancestral to all metastases but not to the primary tumor. Genes in red have more than one mutation occurring on multiple branches; mutation locations are indicated in parentheses. Numbers at each internode indicate the statistical support for the corresponding branch (1 means 100% support). (*C*) Inferred cancer chronogram for subject 416, scaled in years, encompassing the first genetic divergence from Normal sequence (29.4 y), the first genetic divergence of metastases (17 y, blue dashes), and the diagnosis time (8 mo, red dashes). The phylogeny for subject 416 exhibited a diversity of cancer driver mutations occurring across multiple branches.

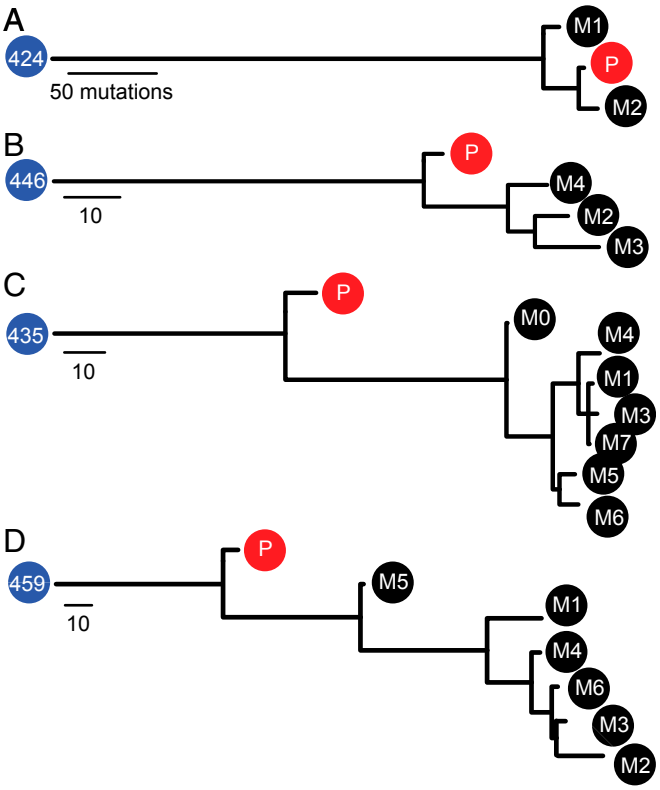

Fig. 2. Four maximum likelihood cancer molecular evolutionary trees, with a horizontal scale proportional to the number of mutations. (*A*) Subject 424 had a colon primary tumor and metastases to the duodenum (M1) and liver (M2). The primary tumor was an ingroup to all metastases in 80.4% of the Bayesian posterior of trees for subject 424. (*B*) Subject 446 had a pancreatic adenocarcinoma primary tumor and metastases to the kidney (M2), bowel (M3), and liver (M4). The primary tumor was an outgroup to all metastases in 99.8% of the Bayesian posterior of trees for subject 446. (*C*) Subject 435 had a poorly differentiated lung adenocarcinoma primary tumor and metastases to the lung (M0), liver (M1), pancreas (M3), hilar lymph node (M4), paraprostatic soft tissue (M5), perirenal soft tissue (M6), and mediastinum (M7). The primary tumor was an outgroup to all metastases in 100% of the Bayesian posterior of trees for subject 435. (*D*) Subject 459 had a lung adenocarcinoma primary tumor and metastases to the lung (M1), liver (M2), spleen (M3), kidney (M4), adrenal (M5) and paratracheal lymph node (M6). The primary tumor was an outgroup to all metastases in 100% of the Bayesian posterior of trees for subject 459.

Rare lineage-specific events, such as new somatic mutations or epigenetic marks, might trigger metastasis, a hypothesis that is a component of linear models of progression (13). If such rare lineage-specific events induce metastasis from the primary tumor, then primary tumor lineages would be expected to produce a single monophyletic clade of metastatic lineages with the primary tumor and normal tissue as outgroups (i.e., all metastatic lineages would share a common ancestor that is more recent than their most recent common ancestor with the primary tumor; e.g., Fig. 1*B*). In contrast, if production of metastatic lineages is rare but independent of causative genetic and epigenetic alterations (i.e., stochastic production of metastatic lineages from all extant lineages), then primary lineages would not be an outgroup to all metastatic lineages in all patients. The stochastic expectation for the primary lineage being the outgroup in a patient with n tumor samples is equal to the number of possible phylogenetic trees with

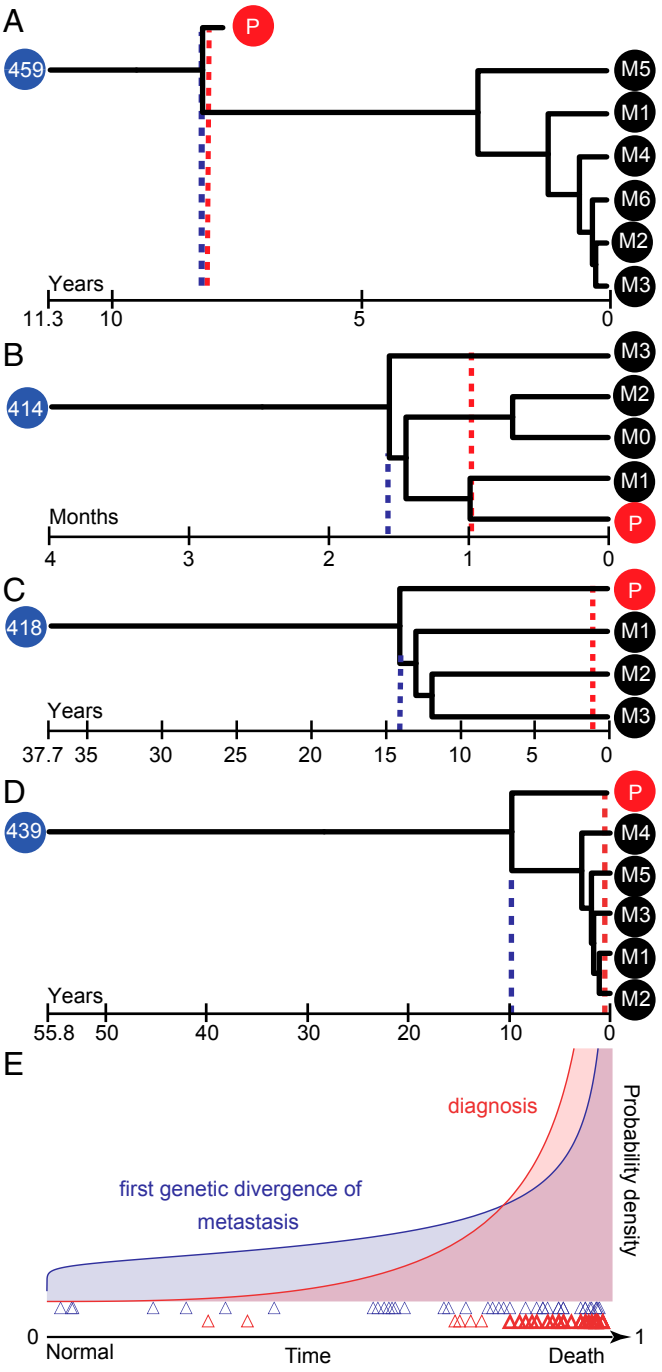

Fig. 3. Timings of the first genetic divergence from normal tissue sequence (blue circle), of the first genetic divergence of metastases (blue dashes) and of diagnosis (red dashes) during tumor progression. (*A*) Subject 459 (lung adenocarcinoma, aged 54 y at death) provided an example of early diagnosis of the primary tumor without diagnosis of metastases, but also early divergence of the metastases. (*B–D*) Subjects 414 (lung, large cell, aged 25 y), 418 (ovarian, aged 47 y), and 439 (renal clear cell carcinoma, aged 58 y) provided examples of late metastasis in which diagnosis of the primary tumor and metastases occurred after the first genetic divergence of metastasis. (*E*) Probability density for the occurrence of the first genetic divergence of metastases and for the time of diagnosis. The *x* axis is scaled from 0 (the first genetic divergence of primary tumor tissue from normal tissue) to 1 (death). In our set of 40 lethal cancers, the first genetic divergences of metastatic lineages (blue triangles) are distributed so as to often occur earlier than diagnosis time (red triangles).

the primary tumor constrained to be an outgroup, $(2n - 5)!!$, divided by the number of possible phylogenetic trees, $(2n-3)!!$, which simplifies to $1/(2n - 3)$.

Of the 24 cancer phylogenies that featured a clinically unambiguous primary tumor and two or more metastases, 16 featured a well-supported topological position of the primary tumor also consistent with patterns of loss of heterozygosity (Fig. S2). Of these 16, we found that 6 (38%) exhibited a most likely topology in which metastatic tumor lineages were not monophyletic and the primary tumor was not the outgroup to all metastases (e.g., Fig. 2*A*) [Bayesian posterior probability (BPP) of the primary tumor as the outgroup ranged from 13–30%, median 21%] (Dataset S7). Monophyly and paraphyly of metastases were not correlated with any particular tumor type. This remarkable frequency of metastatic tumor paraphyly rejects a general model stating that a rare, heritable, and lineage-specific event is necessary to induce metastasis. Instead, it favors a model where a metastatic event can occur stochastically from a heterogeneous primary tumor. Our phylogenies clearly demonstrate a branched evolution model of tumorigenesis and metastasis, as advocated by Burrell et al. (14). Furthermore, they demonstrate that not only do metastatic lineages evolve in parallel with the lineage sequenced from the primary tumor (3), but also they originate from divergent lineages within primary tumors.

In contrast, 10 phylogenies exhibited a topology in which the unambiguous primary tumor was the outgroup to all metastatic tumors (BPP 67–100%; median 100%; e.g., Fig. 2 *B–D* and Dataset S7). This proportion [62.5%; 95% credible interval (CI) 35–85%] is significantly higher than the 19% random expectation for this 16-subject subset (Dataset S7). To incorporate the cancer phylogenies of the 8 additional subjects whose inferred topology with regard to the primary tumor was indicative but moderately to highly uncertain, we integrated over uncertainty of all trees using the Bayesian posterior distributions for all 24 subjects with clinically unambiguous primary tumors. The result was consistent with our previous analysis, yielding a posterior average of 14.3 out of 24 subjects (60%; CI 46–75%), with their primary tumor as the outgroup (Dataset S7).

We then included an additional eight cancer phylogenies with two or three metastases for which the clinical identification of the resected tumors as primary was deemed to be of moderate confidence, yielding a consistent 55% (CI 44–69%) of phylogenies with the primary tumor as the outgroup (Dataset S7). The results are significantly higher than the random expectation (21% for the 32 subjects). This higher value demonstrates that heritable genetic, epigenetic, or other lineage-specific events can contribute a proclivity within lineages toward metastasis of the primary tumor. However, this result also demonstrates that this lineage-specific effect is not so strong as to universally lead to monophyletic metastases (Dataset S7; $P < 10^{-11}$ that at least one of the subjects had a primary tumor that was an ingroup to metastases). Thus, either heritable genetic or epigenetic events at best induce a proclivity toward increased metastasis, or they typically occur early in the evolution of the primary tumor and thus are regularly present in all or nearly all primary tissue long before detection.

The simple linear model specified—that all metastatic tumors are descended from a single original primary cell, such that all metastases are more closely related to each other than they are to any tissue in the primary tumor—requires no explicit modeling of the processes of tumorigenesis and metastasis to imply a specific phylogenetic pattern of metastases and primary tumors. There are many additional questions about the process of metastasis that cannot be addressed without more complex models of the processes underlying tumorigenesis and metastasis and comprehensive intratumor sampling (15). Nevertheless, our phylogenetic analysis demonstrates that (*i*) paraphyly happens, violating the simplest of linear null hypotheses about how primary and metastatic tumors are related (the hypotheses that all metastases are more closely related to each other than they are to any tissue in the primary tumor) and that (*ii*) paraphyly happens less than would

be expected completely by chance, indicating that there is a proclivity within some lineages toward metastasis.

Inference of Cancer Chronograms Revealed the Early Genetic Divergence of Primary and Metastatic Tumor Lineages. Examination of the inferred molecular evolutionary trees illustrates that metastatic lineages can diverge genetically from the primary tumor early in the cancer history. In the cancer molecular evolutionary trees for 11 out of 40 subjects (Fig. S2), the shared ancestral lineage of all tumors was shorter than the subsequent branch lengths leading to a metastatic tissue sampled at autopsy. To examine this pattern of early metastasis further, we analyzed the implications of the inferred tree topology combined with constraints from the timing of clinical sample acquisition. For this purpose, we transformed the molecular evolutionary trees into chronograms (Fig. 1*C*), applying a molecular clock calibrated with the timings of diagnosis, biopsy, surgical resection, and autopsy, and parameterized by cell division times of primary tumor cells (Datasets S5 and S6). For all subjects, we mapped the time of the first genetic divergence of tumor lineages to a 0–1 timescale, where 0 represented the time of genetic origin of tumorigenesis from normal tissue, and 1 represented death of the patient. In the cancer chronograms of 7 subjects, the most recent common ancestor of the primary tumor and metastases occurred in the first half of both the chronological timescale and the tree with branches scaled to mutations (e.g., Fig. 3 *A* vs. *B–D*). Thus, the divergence of metastatic lineages from the primary tumor can occur closer to the first genetic divergence of the tumor from normal tissue, than to death.

The distribution of the times of the earliest divergences of the metastatic and primary tumor lineages is not uniform (Fig. 3*E*; $P < 0.001$). The mean time of the most recent common ancestors of all metastatic and primary tumor lineages was 0.72 on our 0–1 timescale whereas the mean cancer diagnosis time was 0.90, late in the unique genetic history of the cancer (Fig. 3*E*; $P < 0.001$). Saliently, in 35 of our subjects (87.5%), genetic divergence of the first metastatic lineage had already occurred by the time of diagnosis of the primary tumor.

In six cases (15%) in which no metastatic tumors were clinically identified at diagnosis of the primary tumor, we found that metastases had in four cases already genetically diverged. In another five subjects (12.5%), genetic divergence of the first metastatic lineage occurred after diagnosis. In three of these five, however, metastases were diagnosed together with the primary tumor at clinical presentation. Thus, either the metastases sampled were not those present at diagnosis or the true genetic divergences of metastases in these subjects were earlier than inferred by our methods. Timing of divergence of the metastatic and primary tumor lineages was uncorrelated with tumor type. Based on these results, lineages that proceed to metastasis can genetically differentiate from the primary tumor lineage early in the evolutionary and temporal history of cancer, and genetic divergence of metastases often predates diagnosis even when they are not evident at diagnosis. When metastases genetically diverge early, there is less time for the accumulation of any mutations conferring the ability to metastasize subsequent to tumorigenesis. Therefore, early genetic divergence of metastases favors a parallel model (16) in which frequently disseminated cancer cells rarely establish themselves, rather than a linear model (13) of cancer evolution requiring somatic mutations to produce metastases. The observed early genetic divergence of metastases has implications for the retrospective examination of medical care, including whether surgery is seeding distant metastases (17). For example, claims that metastases necessarily arose subsequent to a delay in diagnosis and resection could be refuted.

Our results indicate that metastases from primary tumors can be produced early and stochastically. They suggest that clinical treatments addressing genetically heterogeneous metastases could be warranted even when the sole diagnosis is of a primary tumor (18). In subjects with early or stochastic metastases, the likelihood that crucial metastasis-inducing mutations will be identified as suitable targets for pharmacological intervention is low. Nevertheless, if present, mutations conveying an increased chance of metastasis would be expected to occur before the divergence of any metastatic tumor lineages from the primary tumor. Thus, their identification by examination of the shared ancestry of metastatic lineages separate from the evolution of the sampled primary tumor lineage would be of particular importance (19).

Ancestral State Reconstruction Revealed the Temporal Occurrence of Mutations in Driver Genes Along Tumor Progression. To identify mutations that are divergent from the primary tumor lineage but shared by evolved metastatic lineages, we applied molecular evolutionary models to estimate ancestral sequences in the phylogeny, inferred the gene sequence at every branch point, and mapped all mutations to internodes. Internodes diverging from a primary tumor lineage that constitute a shared ancestral lineage of all metastatic lineages might be particularly likely to be associated with mutations that facilitate metastasis (Fig. 1*B*). We evaluated such mutations with MutSigCV (5) and found genes exhibiting an abnormal burden of mutations.

The early mutations disposing a cell lineage toward cancer are often thought to play key functional roles because they are shared by all tumors (primary and metastatic) and are far more likely to be drivers of tumorigenesis that are key to the origin and persistence of cancer than those that are acquired subsequently. Therefore, we examined whether any genes exhibited an abnormal burden of nonsilent mutations mapping specifically to the origin of the primary tumor (Fig. 1*B*). MutSigCV identified the well-known tumor suppressor *TP53* and the oncogene *KRAS* (false discovery rate of ≤ 0.1) as genes repeatedly mutated early in the evolution of these diverse tumors. Their frequent presence in the root of cancer lineages implies that they play key formative roles in the origin of cancer and that they deserve redoubled attention for their roles in tumorigenesis. The potential to therapeutically target these longstanding drivers of cancer continues to evolve, including targeting TP53 via the TP53/MDM2 interaction (20), targeting KRAS via KRAS G12C (21), targeting post-translational modifications of KRAS (22), or targeting upstream or downstream effectors (23).

To investigate whether mutations of a larger set of known cancer driver genes (6, 19; Dataset S8) tend to occur early in cancer evolution, we compared the occurrence of somatic mutations in driver versus nondriver genes on the early branch shared by all primary and metastatic tumors (Fig. 1*B*) with their occurrence in subsequent branches. Across all subjects, we found that nonsilent tumor type-specific driver mutations (Datasets S9 and S10) are enriched in the early branch (35 early drivers, 13 late drivers, 2,080 early nondrivers, 3,056 late nondrivers; $P = 0.0001$, Fisher's exact test) after removing hypermutated subject 430 (24). Statistical significance held even when retaining subject 430 ($P = 0.001$; Dataset S11) or when tallying all mutations in cancer drivers ($P = 0.001$; Dataset S11).

Furthermore, in the early branch, the ratio of nonsilent versus silent mutations (dN/dS) was significantly higher in tumor type-specific drivers than in nondrivers (41 dN driver, 1 dS driver; 4,418 dN nondriver, 1,908 dS nondriver; $P = 0.0001$). This significance still held when we tallied mutations in all cancer genes from Dataset S8 for all subjects ($P = 0.0027$; Dataset S12). However, dN/dS was not significantly higher in tumor-specific drivers in the late branches than in nondrivers ($P = 0.11$; Dataset S12). The significance of results for both early and late branches remained unchanged after removing hypermutated tumors from subject 430 from the analysis (Dataset S12).

Perhaps a high somatic selection intensity for these driver mutations within clonal cancer lineages means that they either are critical initiating events or are likely to outreplicate other mutations within a small clonal neoplasm. Accordingly, previous assessments of the important genes in cancer (6, 19; Dataset S8) correlate with our findings of their timing early in tumorigenesis. For example, six cancer driver gene mutations known to play key roles in cancer (*KRAS*, *TP53*, *PIK3CA* encoding a phosphoinositide 3-kinase, *KMT2C* and *KMT2D* encoding histone methytransferases, and *ALK* encoding a receptor tyrosine kinase)

occurred in multiple cancer patients across tumor types. Based on the inferred locations of these mutations in the cancer chronograms—and consistent with our early driver finding—*KRAS* and *TP53*, involved in cancer-related pathways such as MAPK signaling and telomere maintenance, are likely to be key early mutations in tumorigenesis. Mutations in *KMT2D* and *PIK3CA* frequently occurred midway along the cancer history ($P < 0.001$), and *ALK* and *KMT2C* frequently occurred late ($P < 0.001$) in the evolution of cancer (Fig. 4). Note that *TP53* would be expected to more frequently receive its first mutation than a gene with a smaller mutational target like *KRAS*, which is oncogenic only with mutations at codon sites 12, 13, 18, 61, and 117, because the mutational target size of genes is highly relevant to when they are likely to receive their first disabling mutation. The temporal interdispersion of mutations of a tumor suppressor with high mutational target size (*TP53*) within oncogenes with low mutational target sizes (*KRAS*, *PIK3CA*, *KMT2D*, *ALK*, and *KMT2C*) implies that the temporal order inferred is partially driven by positive selection and possibly epistatic interactions rather than solely being driven by the waiting time to the first driver mutation.

Conclusion

We have demonstrated that genetic lineages of metastases can arise early in primary tumors, sometimes long before primary tumor diagnosis. This result directs research efforts away from prevention of metastasis-facilitating mutations and toward a better understanding of fundamental drivers of tumorigenesis. Second, we demonstrate conclusively that, in contrast with the longstanding model of linear progression of cancer, metastases can originate from divergent lineages within primary tumors. This result argues that, although evolved genetic changes in cancer lineages seem to affect the proclivity of tumor cells to metastasize, it is unlikely that there are single genetic changes that are necessary or sufficient for metastasis. Lastly, we demonstrate the temporal order of occurrence of relevant driver mutations, indicating their relative roles in tumorigenesis. Although many studies have identified therapeutically targetable alterations associated with late-stage tumors (25), evolutionary analyses of the timing of mutations in driver genes in cancerous tissues will be important not only because pharmaceutical treatment could have a large therapeutic effect on tumors exhibiting these early-appearing mutations, but also because any such therapeutic effect is likely to be exerted across otherwise heterogeneous primary and metastatic tumor lineages. These alleles represent targets for pharmaceutical intervention that exist in nearly all subsequent lineages, and their targeting would provide the greatest potential for long-term success of therapy.

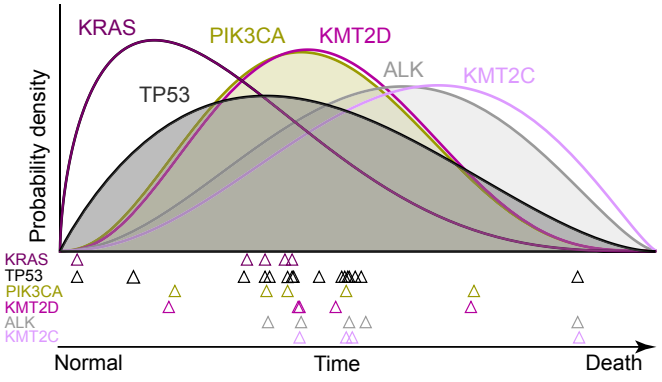

Fig. 4. Inferred distributions of the temporal occurrence of mutations (Δ) in cancer driver genes *KRAS* (dark purple line, no shading), *TP53* (black line, gray shading), *PIK3CA* (olive line, olive shading), *KMT2D* (medium purple line, no shading), *ALK* (light gray line, light gray shading), and *KMT2C* (light purple line, no shading) across diverse cancer types. Probability densities for the appearance of alleles with nonsilent mutations across these cases indicate that mutations of *KRAS* and *TP53* tend to occur earlier than mutations of *PIK3CA* or *KMT2D* ($P < 0.001$), which, in turn, tend to occur earlier than mutations of *ALK* or *KMT2C* ($P < 0.001$). Mutations are depicted at the midpoint of the interval during which the gene was inferred to be mutated in the particular subject.

Materials and Methods

Ethics Statement and Specimen Sampling. The tissues assessed in this study were obtained from the Yale Pathology Archives based on Yale Human Investigation Committee at Yale University, Protocol no. 0304025173 to D.L.R., which allows retrieval of tissue from archives that was consented or has been approved for use with waiver of consent. Formalin-fixed paraffin-embedded (FFPE) primary tumor tissues, metastases, and normal tissues were obtained from 40 subjects (Dataset S1). Clinical characteristics, tumor type, numbers of metastatic samples, and locations are summarized in Fig. 1.

DNA Extraction, Exome Sequencing, Variant Calling, and Variant Validation. Genomic DNA was extracted from FFPE core biopsies using the BiOstic FFPE Tissue DNA isolation kit (MO BIO) following standard procedures that we have used successfully in the past to recover single nucleotide variants (26). Genomic DNA underwent targeted exome capture using the NimbleGen SeqCap EZ Human Exome Library v2.0 (Roche), followed by sequencing on the Illumina HiSeq platform to generate 74-base paired end reads, as in Choi et al. (27). To determine variant sites, sequences from all tumor tissues were aligned to the human reference genome (hg19) using Eland (Illumina), and single nucleotide variants were called using SAMtools (28). Tumor purity was estimated using the mean of minor allele frequencies of SNPs within regions of loss of heterozygosity (LOH). Somatic mutations were identified by comparing read counts from tumor samples (primary and/or metastasis) with those from corresponding normal tissues, as described in Choi et al. (27).

To accurately characterize the nucleotide state at the variant sites across tumor tissues within a patient, somatic mutations were further analyzed to account for tumor impurity, loss of heterozygosity, and other factors that confound variant calling (Fig. S4). To eliminate false positive calls present in common SNP databases, variants were removed if they were present in the National Heart, Lung, and Blood Institute Exome Sequencing Project (release ESP6500SI-V2), 1000 Genomes (release no. 84, March 2012), or the Yale Human Exome Database. To validate somatic mutations identified (Dataset S4), we performed direct Sanger sequencing of known cancer genes after targeted PCR (Dataset S3) for patients depicted in Figs. 1 and 4.

Phylogenetic Analyses and Ancestral State Reconstruction. Somatic variant sites across each group of corresponding tumor and normal samples were concatenated into multiple sequence alignments and analyzed by phylogenetic and evolutionary methods (Dataset S4). Trees that maximized parsimony were conclusively identified with an exhaustive tree bisection–reconnection search by PAUP 4.0 (9). Maximum likelihood trees were estimated using GARLI v2.0 (10), which applies a modified genetic algorithm to infer the best topology, branch lengths, and substitution model parameters simultaneously, accessed by the convergence of log Likelihood (lnL) scores in two runs. Bayesian inference of tree topology was performed using MrBayes (11).

To better understand tumor progression, we inferred sequences for the ancestor of all tumors and all other ancestral/internal nodes in the phylogeny. Ancestral state reconstruction was performed using the package baseml from PAML 4.7 (29), implemented with the tree topology from the maximum likelihood tree based on the alignment using the K80 model. The inferred ancestral state for each internal node was compared with the sequence state of the adjacent node to infer mutations that occurred on each branch of the phylogeny (Fig. 1*B*). All of the mutated genes were then mapped on each branch of the phylogeny for each of the 40 subjects for further temporal analyses.

Inference of Cancer Chronograms. Because mutations in self-renewing tissues are cumulative over time and directly correlated with age (30), we used a molecular clock phylogenetic approach to infer the order of genetic divergence of tumors, calibrating it with the timings of diagnosis, biopsy, and/or surgical resection of the primary tumor. We inferred chronograms for all cancer phylogenies (Fig. S3), under a relaxed uncorrelated-clock model specifying uniform branching priors using mcmctree implemented in PAML 4.7 (29), which accommodates rate heterogeneity, including increases in the rate of mutations (31) among tumor tissues over time and across the phylogeny. We specified a gamma prior distribution for the rate of somatic cell substitutions at 8.4×10^{-8} site^{-1}·y^{-1} (32), scaled to the observed potential doubling time (T_{pot}) and SD for each specific tumor (12).

Temporal Inference of Genetic Divergence of Tumors and Diagnosis Time. To know the distribution of the times of the first genetic divergences of tumor lineages, we mapped them to the 0–1 timescale inferred for each chronogram, where zero represented the genetic origin of tumorigenesis from normal tissue and one represented death of the patient (Fig. 1*C*). For each subject, we then drew 1,000 values from the posterior distribution for the genetic origin of tumorigenesis from the normal tissue, and 1,000 values from the posterior distribution for the time of first diagnosis. We calculated the mean as a central estimate for each of these distributions for each subject and also calculated the maximum likelihood beta distribution for each of these sets of 40 time points using Mathematica 9. For comparison, we plotted the beta distributions superimposed on our relative timescale (Fig. 3). To evaluate whether inferred beta distributions were statistically significantly different from uniform, the inferred maximum likelihood model was compared with the maximum likelihood model of a uniform beta distribution ($\alpha = 1$, $\beta = 1$) via a likelihood ratio test evaluated against the χ^2 distribution with two degrees of freedom. To establish the significance of difference between distributions of tumor ancestor times and diagnosis times, we compared the maximum likelihood null model of a single beta distribution for all inferred times to the maximum likelihood of two distinct beta distributions for tumor ancestor times and diagnosis times via a nested likelihood ratio test evaluated against the χ^2 distribution with two degrees of freedom.

Temporal Inference of Gene Occurrence Along Tumor Progression. We performed a Fisher's exact test to check whether tumor driver genes were occurring significantly early during tumor progression. To detect the temporal order of mutated alleles, we assembled the time intervals when genes were inferred to be mutated across all subjects on a 0–1 timescale (Fig. 1*C*). Each mutation was associated with a branch in the cancer phylogeny by our ancestral state inference. We integrated over uncertainty in the timing of occurrence of the mutation by sampling uniformly from the associated branch on 1,000 of our 0–1-scaled chronograms in our Bayesian posterior and averaging the resulting sample set. Using Mathematica 9, we fitted these averages (Fig. 4) to the maximum likelihood beta distribution for when the mutations commonly arise on the 0–1 timescale.

ACKNOWLEDGMENTS. We thank Douglas Brash, Gilbert Omenn, Andrew Clark, Tandy Warnow, Daniel Hartl, Allen Rodrigo, and Angelika Hofmann for providing feedback on draft versions of the manuscript. We also thank the Yale High Performance Computing Center for providing computational resources and Robert Bjornson for assistance regarding disposition of clusters and servers to the project. We thank Irina Tikhonova and Christopher Castaldi in the Yale Center for Genome Analysis for sample preparation and sequencing. This project was supported by Gilead Sciences, Inc. A.I. was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo scholarships 12/04818-5 and 13/15144-8.

1. Gerlinger M, et al. (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 46(3):225–233.
2. Weinberg RA (1991) Tumor suppressor genes. *Science* 254(5035):1138–1146.
3. Harbst K, et al. (2014) Molecular and genetic diversity in the metastatic process of melanoma. *J Pathol* 233(1):39–50.
4. Nguyen DX, Bos PD, Massagué J (2009) Metastasis: From dissemination to organ-specific colonization. *Nat Rev Cancer* 9(4):274–284.
5. Lawrence MS, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214–218.
6. Lawrence MS, et al. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505(7484):495–501.
7. Naxerova K, Jain RK (2015) Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat Rev Clin Oncol* 12(5):258–272.
8. Miller CA, et al. (2014) SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Comput Biol* 10(8):e1003665.
9. Swofford DL (2003) PAUP*, Phylogenetic Analysis Using Parsimony (* and Other Methods), Version 4.
10. Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis (The University of Texas, Austin).
11. Ronquist F, et al. (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61(3):539–542.
12. Rew DA, Wilson GD (2000) Cell production rates in human tissues and tumours and their significance. Part II: Clinical data. *Eur J Surg Oncol* 26(4):405–417.
13. Hynes RO (2003) Metastatic potential: Generic predisposition of the primary tumor or rare, metastatic variants-or both? *Cell* 113(7):821–823.
14. Burrell RA, McGranahan N, Bartek J, Swanton C (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501(7467):338–345.
15. Hong WS, Shpak M, Townsend JP (2015) Inferring the origin of metastases from cancer phylogenies. *Cancer Res* 75(19):4021–4025.
16. Klein CA (2009) Parallel progression of primary tumours and metastases. *Nat Rev Cancer* 9(4):302–312.
17. Baum M, Demicheli R, Hrushesky W, Retsky M (2005) Does surgery unfavourably perturb the "natural history" of early breast cancer by accelerating the appearance of distant metastases? *Eur J Cancer* 41(4):508–515.
18. Gray JW (2003) Evidence emerges for early metastasis and parallel evolution of primary and metastatic tumors. *Cancer Cell* 4(1):4–6.
19. Vogelstein B, et al. (2013) Cancer genome landscapes. *Science* 339(6127):1546–1558.
20. Wang S, et al. (2014) SAR405838: An optimized inhibitor of MDM2-p53 interaction that induces complete and durable tumor regression. *Cancer Res* 74(20):5855–5865.
21. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM (2013) K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* 503(7477):548–551.
22. Ahearn IM, Haigis K, Bar-Sagi D, Philips MR (2012) Regulating the regulator: Post-translational modification of RAS. *Nat Rev Mol Cell Biol* 13(1):39–51.
23. Downward J (2003) Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer* 3(1):11–22.
24. Harfe BD, Jinks-Robertson S (2000) DNA mismatch repair and genetic instability. *Annu Rev Genet* 34:359–399.
25. Vasan N, et al. (2014) A targeted next-generation sequencing assay detects a high frequency of therapeutically targetable alterations in primary and metastatic breast cancers: Implications for clinical practice. *Oncologist* 19(5):453–458.
26. Goh G, et al. (2014) Recurrent activating mutation in PRKACA in cortisol-producing adrenal tumors. *Nat Genet* 46(6):613–617.
27. Choi M, et al. (2011) K+ channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. *Science* 331(6018):768–772.
28. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
29. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
30. Tomasetti C, Vogelstein B, Parmigiani G (2013) Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci USA* 110(6):1999–2004.
31. Palles C, et al.; CORGI Consortium; WGS500 Consortium (2013) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 45(2):136–144.
32. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* 107(3):961–968.
33. Tokuda Y, et al. (1990) Fundamental study on the mechanism of DNA degradation in tissues fixed in formaldehyde. *J Clin Pathol* 43(9):748–751.
34. Yu YP, Michalopoulos A, Ding Y, Tseng G, Luo J-H (2014) High fidelity copy number analysis of formalin-fixed and paraffin-embedded tissues using Affymetrix Cytoscan HD chip. *PLoS One* 9(4):e92820.
35. Townsend JP, Su Z, Tekle YI (2012) Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. *Syst Biol* 61(5):835–849.
36. Sukumaran J, Holder M (2008) SumTrees, Summarization of Split Support on Phylogenetic Trees (part of the DendroPy Phylogenetic Computation Library), Version 2(3).
37. Yang Z (2006) *Computational Molecular Evolution* (Oxford Univ Press, Oxford).
38. Rambaut A, Drummond A, Suchard M (2013) Tracer, Version 1. 6.
39. Erba E, et al. (1994) Cell kinetics of human ovarian cancer with in vivo administration of bromodeoxyuridine. *Ann Oncol* 5(7):627–634.
40. Choi J-H, Wong AST, Huang H-F, Leung PCK (2007) Gonadotropins and ovarian cancer. *Endocr Rev* 28(4):440–461.
41. Rew DA, Thomas DJ, Coptcoat M, Wilson GD (1991) Measurement of in vivo urological tumour cell kinetics using multiparameter flow cytometry: Preliminary study. *Br J Urol* 68(1):44–48.
42. Oda T, et al. (2001) Growth rates of primary and metastatic lesions of renal cell carcinoma. *Int J Urol* 8(9):473–477.
43. Zhang J, Kang SK, Wang L, Touijer A, Hricak H (2009) Distribution of renal tumor growth rates determined by using serial volumetric CT measurements. *Radiology* 250(1):137–144.
44. Schofield MJ, Hsieh P (2003) DNA mismatch repair: Molecular mechanisms and biological function. *Annu Rev Microbiol* 57:579–608.
45. Kariola R, et al. (2004) MSH6 missense mutations are often associated with no or low cancer susceptibility. *Br J Cancer* 91(7):1287–1292.
46. Edelmann W, et al. (1997) Mutation in the mismatch repair gene Msh6 causes cancer susceptibility. *Cell* 91(4):467–477.
47. Rowland MM, Schonhoft JD, McKibbin PL, David SS, Stivers JT (2014) Microscopic mechanism of DNA damage searching by hOGG1. *Nucleic Acids Res* 42(14):9295–9303.
48. Wood RD, Mitchell M, Lindahl T (2005) Human DNA repair genes, 2005. *Mutat Res* 577(1-2):275–283.
49. Stanton PD, Cooke TG, Forster G, Smith D, Going JJ (1996) Cell kinetics in vivo of human breast cancer. *Br J Surg* 83(1):98–102.
50. Nik-Zainal S, et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149(5):979–993.
51. Ostrow SL, Barshir R, DeGregori J, Yeger-Lotem E, Hershberg R (2014) Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS Genet* 10(3):e1004239.
52. Sottoriva A, et al. (2015) A Big Bang model of human colorectal tumor growth. *Nat Genet* 47(3):209–216.