Inferring the origin of metastases from cancer phylogenies

Woo Suk Hong¹, Max Shpak²,³,⁴,*, Jeffrey P. Townsend⁵,⁶,⁷**

¹. Yale School of Medicine, New Haven, CT 06510
². NeuroTexas Institute, St. David’s Medical Center, Austin TX 78705
³. Center for Systems and Synthetic Biology, University of Texas, Austin TX 78712
⁴. Fresh Pond Research Institute, Cambridge MA 02140
⁵. Department of Biostatistics, Yale School of Public Health, New Haven CT 06510
⁶. Department of Ecology and Evolutionary Biology, Yale University, New Haven CT 06520
⁷. Program in Computational Biology and Bioinformatics, Yale University, New Haven CT 06510

* corresponding author
**order of authorship is alphabetical, indicating equal contributions

Running Title: Phylogeny of metastases

Keywords: Metastasis, Phylogeny, Somatic Evolution, Tumor Heterogeneity

Financial Support: WSH was supported by Yale School of Medicine Medical Student Research Fellowship. MS was supported by the St. David’s Foundation impact fund. JPT was supported by the Notsew Orm Sands Foundation and by NIH NCI 1U01CA176067-01A1.

Corresponding Author: Max Shpak, 1015 E. 32nd St, Suite 404, St. David’s Medical Center, Austin TX 78705, 512-544-8077, 512-544-5547 (fax), shpak.max@gmail.com

Conflicts of Interest: None

Notes: 3024 words (excluding references and figure captions). 2 Figures, 0 Tables.
Abstract
Determining the evolutionary history of metastases is a key problem in cancer biology. Several recent studies have presented inferences regarding the origin of metastases based on phylogenies of cancer lineages. Many of these studies have concluded that the observed monophyly of metastatic subclones favored metastasis-to-metastasis spread (“a metastatic cascade” rather than parallel metastases from the primary tumor). In this article, we argue that identifying a monophyletic clade of metastatic subclones does not provide sufficient evidence to unequivocally establish a history of metastatic cascades. In the absence of a complete phylogeny of the subclones within the primary tumor, a scenario of parallel metastatic events from the primary tumor is an equally plausible interpretation. Future phylogenetic studies on the origin of metastases should obtain a complete phylogeny of subclones within the primary tumor. This complete phylogeny may be obtainable by ultra-deep sequencing and phasing of large sections or by targeted sequencing of many small, spatially heterogeneous sections, followed by phylogenetic reconstruction using well-established molecular evolutionary models. In addition to resolving the evolutionary history of metastases, a complete phylogeny of subclones within the primary tumor facilitates the identification of driver mutations by application of phylogeny-based tests of natural selection.
Introduction

Cancer progression, from the formation of the first neoplastic cell to metastasis, occurs through the origin and proliferation of subclonal lineages characterized by unique somatic mutations. Provided sufficient numbers of mutations occur during the branching process of cancer lineage evolution [1-3], this process generates a hierarchical genealogy that can be analyzed using the phylogenetic tools of evolutionary biology. Numerous studies have leveraged the reconstructed phylogenies of tumor subclones to infer the genetic history of cancer progression, including addressing the origin of metastases. For example, metastasis-to-metastasis spread (a “metastatic cascade”), as opposed to parallel metastatic events, has been inferred from the tree-like topology of tumor subclone phylogenies [4-7]. Central to this conclusion was the assumption that parallel metastatic events from a single primary tumor would necessarily result in a star topology with a polyphyletic group of metastases, while a metastatic cascade would result in a tree topology with a monophyletic clade of metastases. This assumption—and others like it—require careful scrutiny to ensure that the inferred phylogenies truly support proposed models of cancer progression and evolution. In this commentary, we investigate the roles such assumptions play in these studies. For instance, star phylogenies are not the only possible outcomes of parallel metastases. We argue that the chronology of metastatic events cannot be established without complete information on the phylogeny of subclonal lineages within the primary tumor. Below, we will show why this is the case and suggest approaches that are more likely to provide insights into the chronology of metastatic events.

Intra-Tumor Phylogeny and Metastasis.

To illustrate how parallel metastatic events can generate a monophyletic subclade of genotypes from different metastatic tumors, consider the tumor progression histories and phylogenies shown in Figures 1a–b. In both examples, subclones A and E represent the most common and the only recognized subclones in the primary tumor, while subclones B, C, and D represent metastases. Figure 1a illustrates a single metastatic event from the primary tumor followed by a metastatic cascade, while Figure 1b shows three parallel metastases from the primary tumor. In spite of two fundamentally different histories of metastatic events, both scenarios generate the same phylogenetic tree, with a monophyletic clade of metastases. Figure 1c shows a third scenario where parallel metastatic events can be unambiguously inferred from the polyphyly of metastatic genotypes.

In the absence of a complete phylogeny of all cell lineages within the primary tumor, the most that can be inferred from a monophyletic clade of metastases is that all metastases descended from a single subclone within the primary tumor, rather than descending from different sampled subclones within the primary tumor (Figure 1c). The monophyly of metastatic genotypes in itself provides no means of distinguishing between a single metastatic event and parallel metastatic events.
The relative plausibility of each of these three scenarios (1a, 1b, or 1c) depends on the biology of metastasis. If most cells in the primary tumor have a high potential for generating metastases, we should expect parallel metastatic events from the primary tumor that will result in unambiguous polyphyly of metastatic genotypes (i.e. some metastases will be more closely related to primary tumor subclones than to other metastases). In contrast, if we assume a linear model in which there are a limited number of mutations that permit cells to metastasize, then we should expect all metastatic cells to arise from a single subclonal lineage. Nevertheless, it is crucial to note that the set of shared mutations present among all members of a metastatic subclade are not equivalent to a metastatic phenotype; mutations instigating metastasis might have occurred later in evolution and might be specific to subsets of lineages. This distinction between shared mutations that define a lineage and the unique phenotypes of its extant descendants is analogous to a problem in the molecular phylogenies of higher organisms: it cannot be assumed that the phenotypic traits defining extant “crown taxa” of extant related taxa were present at the times when the stem lineages diverged from their common ancestor [8,9]. For example, it is unlikely that defining morphological traits such as upright posture/locomotion appeared in the most recent common ancestor of humans and australopithecines at the moment this monophyletic lineage split from the shared common ancestor with chimpanzees. Similarly, we cannot assume that a monophyletic lineage of metastatic subclones necessarily had a metastatic cell as a most recent common ancestor.

Consequently, only a phylogeny replete with a complete representation of subclones within the primary tumor will make it possible to distinguish parallel metastases of primary tumor origin from a metastatic cascade. Unless we can rule out the possibility of clonal diversification within the primary tumor followed by parallel metastases, we cannot claim clonal diversification within metastases. The challenge is made even worse if we admit a possibility of subclones within the primary tumor going extinct as a consequence of natural immune surveillance or of early therapeutic intervention, in which case no purely topological inference will differentiate between a scenario of parallel metastases and a scenario of a metastatic cascade.

Examples from the recent literature
Two recent publications have wrestled with the problem of inferring the presence or absence of metastatic cascades without gathering complete representations of subclonal heterogeneity within the primary tumor. Schwarz et al [5] inferred primarily metastasis-to-metastasis spread (consistent with Figure 1a) based on the monophyly of metastatic genotypes in 8 out of 9 subjects with high-grade serous ovarian cancer. However, their samples analyzed include only one major clone from the primary tumor of each subject, and thus did not differentiate between the metastatic cascade scenario and the alternative hypothesis of
parallel metastatic events (as in Figure 1b), because of the absence of necessary information on spatially distinct subclonal heterogeneity within the primary tumor.

In another recent example, Hong et al [6] infer from the monophyly of metastatic genotypes in a subject with prostate cancer a cancer history including a complex metastatic cascade, with cross-seedings across metastatic sites and re-seedings into the surgical bed (site of primary tumor resection). In the phylogeny of the first subject in Hong et al. (Fig. 2a), four subclones (A, B, C, and D) were identified in the primary tumor through extensive spatial sampling from seven different locations. Subclone E was unique to the single clinically identifiable metastasis and was clearly derived from a single metastatic event from the primary. In contrast, the primary tumor was sequenced in only two locations in the second subject, yielding three subclones (Figure 2b). Perhaps it is no coincidence that it is in this subject that a complex metastatic cascade is inferred, based on the monophyly of metastatic genotypes with primary tumor subclones as an outgroup. Because the sampling of the primary tumor in the second subject was limited in comparison to the sampling of the primary tumor in their first subject, there is no way to formally exclude the possibility of parallel metastatic events from unsampled subclones in the primary tumor (Figure 1b), as opposed to a metastatic cascade (Figure 1a). In the third subject (Figure 2c), a single primary tumor subclone was inferred from a single sample. In this case, because the metastatic subclones are polyphyletic (i.e. metastatic subclone B shares a more recent common ancestor with primary tumor subclone A than with the other metastatic subclones), the inference of parallel metastatic events is unambiguous (e.g. Figure 1c).

Suggested Approaches and Applications

A number of recent studies have more or less explicitly acknowledged the need for a complete characterization of subclonal heterogeneity within the primary tumor in order to unambiguously estimate the location of the genetic divergence of metastases. In these studies, the most frequent approach taken has been to perform deep next generation sequencing on as large a portion of the tumor as possible, followed by the execution of phasing algorithms for genotyping [11-13]. This thorough approach was applied by Gundem et al [7], in which evidence was marshaled for the presence of metastatic cascades in subjects with lethal prostate cancer, and in Yates et al[10], which investigated subclonal diversification within primary breast cancer tissue. Gundem et al [7] point out that “we cannot formally exclude an alternative explanation for the observed patterns, that each of these metastases has seeded from an undetected subclone in the primary tumor. However, targeted resequencing of a subset of mutations failed to detect any such subclones, despite a median sequencing depth of 471×” (p. 354). Care must be taken, however, to differentiate between depth of sequencing and extent of spatial sampling. Sequencing depth is irrelevant to the inference of genetic variation within the primary tumor unless it is accompanied by broad
spatial sampling of the tumor, since genetic diversity in tumors is spatially partitioned and increasing sequence depth for a small section of the tumor will not necessarily capture any additional subclones. For the question of discerning parallel metastases from metastatic cascades, sequencing a sample of limited spatial extent at great depth is the proverbial equivalent of rereading the same paragraph of a newspaper article to validate its truth.

Two approaches can be applied to address these issues and better resolve the question of parallel metastases or metastatic cascades. First, one can sample a large number of small, spatially separated sections of the tumor, under the assumption that genetic diversity is spatially partitioned and that each small portion is relatively homogeneous. If the assumptions are valid, this approach eliminates the need for especially deep sequencing and avoids the challenging task of haplotype inference from multiple genotypes. Alternatively, one can perform highly parallelized single-cell sequencing from entire tumors or from spatially heterogeneous samples of a single tumor to provide the same information. Even with single-cell sequencing, it remains a requirement that single cells from sufficiently many tumor sections are sampled to capture the full diversity of the primary tumor. The number of cells needed to be sequenced would depend—in a way that is not yet worked out—on the extent of subclonal heterogeneity and on the quality of each single-cell exome or genome sequence.

To be clear, next generation sequencing of tumor sections [14-16] and single cell sequencing [17-21] have successfully reconstructed plausibly complete subclone phylogenies within the primary tumor. Importantly, results of these studies only provided evidence for the polyphyly of metastatic genotypes. There were many instances in these studies where metastases shared unique somatic mutations with subclones in the primary tumor that were not shared with other metastases. Moreover, the metastases and their sister lineages in the primary tumor were typically descendants of single, comparatively recent subclones in the tumor, as depicted in Figure 1b. In the absence of what appears to have been a complete phylogeny within the primary tumor, each of these studies could easily have erroneously identified metastatic tumors as monophyletic and arising as a consequence of a metastatic cascade.

Currently, whether one interprets the monophyly of metastatic subclones—when observed—as the result of a metastatic cascade or as the result of parallel metastases depends to a large degree on one’s assumptions about whether the capacity for metastasis is innate in most primary tumor cells or a derived trait found in a uniquely derived subpopulation of cells. In a detailed, spatially explicit, and spatially thorough sequencing study of metastatic pancreatic cancer [15], it was argued that the capacity for metastasis is acquired later in the course of tumor progression because none of the metastatic tumor genotypes were derived from the ancestral tumor genotype. However, Yachida et al. [16] found no evidence for metastatic cascades: the two distinct liver metastases and the lung metastases are derived from different (albeit recent) subclones, while the subclones associated with local invasion (metastasis to
the peritoneum) occurred comparatively early in the subclone genealogy. The latter result is consistent
with studies that indicate that the capacity for invasive migration is innate to many cancer types [22-24].
It implies that early seeding of metastatic lineages is likely in many cases, and argues that with sufficient
sampling of subclones, we should generally expect to find phylogenies resembling Fig. 1c rather than 1a.

Completely characterizing subclonal variation within tumors and correctly inferring their
phylogeny provides a the frame for addressing many key aspects of cancer evolutionary genomics, such
as distinguishing somatic natural selection from neutral evolution. The implications of the subclonal
phylogeny are wide-ranging. For instance, in Schwarz et al. [4], a “clonal expansion (C.E.) index” is
proposed that compares the distribution of subclones within and among tumors to a uniform distribution
as a model for distinguishing selection from the neutral effects of mutation and genetic drift. However,
the uniform null for tumor genotypes is not justified by neutral models of genome evolution in tumor
cells. It is not the null expectation for any population evolving along a phylogenetic tree, nor is it the null
expectation for the distributions of genotypes of organisms in a population, all for the same reasons:
genealogical relatedness, finite time, and spatial structure. This realization is the basis for comparative
method [25] statistics in evolutionary biology. Attempts to validate causal relations between traits that
map to subclonal genotypes must make us of similar approaches applied to tumor subclones. Specifically,
reconstructing the genealogy of tumor subclones provides a foundation for deriving null distributions of
subclone frequencies, potentially by use of a coalescent model of neutral evolution. Under a neutral null
model, it is assumed that all changes in the frequency distribution of tumor subclones result from
mutations and stochastic genetic drift, in contrast to any deterministic process of natural selection
favoring some subclones over others. A neutral distribution would be the correct null on which to base
tests of natural selection on driver genes, as is used in tests based on the distribution of pairwise genetic
distances among genotypes that are now standard in population genetics research [26,27]. Finally, a
complete phylogeny would also facilitate tests of natural selection—such as comparisons of silent vs.
nonsynonymous nucleotide substitutions, e.g. [28]—that are performed on the basis of genealogical
information rather than under an arbitrary assumption of null uniformity.

Future Directions:
We have illustrated the challenges inherent to inferring the mode of metastasis spread from cancer
phylogenies in the absence of a complete characterization of subclone diversity within the primary tumor.
Specifically, we have demonstrated how parallel metastatic events can result in the monophyly of
metastases, refuting the idea that monophyly of metastases necessarily implies metastasis-to-metastasis
spread, without a complete accounting of subclonal heterogeneity within the primary tumor. A complete
phylogeny of subclones within the primary tumor requires extensive sampling of the subclone diversity; it
also requires the application of algorithms that can reliably infer the evolutionary history of subclone-defining mutations. We remark that many of the studies of tumor metastases referenced above either use distance-based clustering methods or heuristic reconstructions of genealogy done by inspection, and argue that these approaches need to be refined by the application of well-established character-based phylogeny reconstruction methods based on molecular evolutionary models.

Applying phylogenetic analysis to the study of cancer with an understanding of its power and limitations will be a great boon to cancer analytic methodology. Inferring the mode of metastasis spread is just one area to which applications of longstanding research on evolutionary biology can be applied to the evolution of cancer; for example, the timing of key mutations underlying tumorigenesis can also be inferred using ancestral state inference and mutation mapping. It has long been recognized that genetic heterogeneity within and among tumors in a patient affects clinical outcomes such as response to therapy. In such cases, regardless of the originating basis of that heterogeneity, the degree of genealogical affinity among tumor lineages should be predictive of the degree of response at dispersed sites. A reliable inference of the evolutionary history of genetic variants will inform clinical decision-making as well as provide insights into the basic biology of tumors and metastases. Ultimately, phylogenetic information has significant potential to be useful in the clinic, though the relation of cancer phylogenies to clinical trajectories has yet to be demonstrated. Phylogenies will likely exhibit predictive value for a subject’s prognosis and the cancer’s responses to therapy—they are the natural conceptual framework for conveying an understanding of tumor heterogeneity, which is the putative source of emergent tumor resistance to current pharmaceutical therapeutics. As sampling multiple sites becomes increasingly feasible with decreasing sequencing costs, and as clinicians become more aware of the relevance and interpretation of phylogenetic analysis, we expect to see studies that map clinical trajectories to phylogenies, infer ancestral states and the timing of mutations, and even describe the full temporal history of cancer within subjects. Such studies will provide a complex but manageable scheme of how cancer evolves, and will lead to testable hypotheses regarding the optimal means of providing personalized care informed by phylogenetic and evolutionary analysis.

Acknowledgements

WSH was supported by Yale School of Medicine Medical Student Research Fellowship. MS was supported by the St. David’s Foundation impact fund. JPT was supported by the Notsew Orm Sands Foundation and by NIH NCI 1U01CA176067-01A1.

References


**Figure Legends**

**Figure 1a.** Assume that the subclones A and E are the most common in the primary tumor, while F and G are rarer and are not sampled, and that there is a unique primary tumor metastasis from which later metastatic subclones (B, C, and D) are derived.

**Figure 1b.** This case makes similar assumptions about sampling as in 1a, except that parallel metastases are independently derived from the lineage containing the rarer (unsampled) subclones F and G.

**Figure 1c.** Metastatic tumor subclades B–D are derived from polyphyletic subclonal lineages in the primary tumor.

**Figure 2.** Phylogenies of the first three subjects from Hong et al.[6] (a: Subject 299, b: Subject 498, c: Subject 177). These phylogenetic trees are not the result of a new analysis. They are a preferred graphical representation to those presented in Hong et al.[6] that is topologically equivalent. In this depiction, observed subclones are represented as terminal taxa (leaves) rather than as direct ancestors (internal
nodes). The topologies of the phylogenetic relationships among subclones shown here are entirely equivalent to the topologies presented in Hong et al.[6].
Figure 1.

a. A diagram showing the relationship between primary and metastatic cancer samples, with labels A, B, C, D, E, F, and G. The diagram indicates that samples F and G were not sampled.

b. A similar diagram to a, but with a different structure.

c. A third diagram, also showing the relationship between primary and metastatic cancer samples, with labels A, B, C, D, E, F, and G. The diagram indicates that samples F and G were not sampled.
Figure 2.

a. [Diagram of a tree with nodes labeled A, B, C, D, E, where A and B are primary, C, D, and E are metastatic.]

b. [Diagram of a tree with nodes labeled A, B, C, D, E, F, G, H, where A and B are primary, C, D, E, F, G, and H are metastatic.]

c. [Diagram of a tree with nodes labeled A, B, C, D, where A and B are primary, C and D are metastatic.]
Inferring the origin of metastases from cancer phylogenies

Woo Suk Hong, Max Shpak and Jeffrey P. Townsend

Cancer Res Published OnlineFirst August 10, 2015.

Updated version
Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-15-1889

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.