# Inferring Tumor Phylogenies from Multi-region Sequencing

Zheng Hu[1,2] and Christina Curtis[1,2,*]
[1]Departments of Medicine and Genetics
[2]Stanford Cancer Institute
Stanford University School of Medicine, Stanford, CA 94305, USA
*Correspondence: cncurtis@stanford.edu
http://dx.doi.org/10.1016/j.cels.2016.07.007

A new computational method illuminates the heterogeneity and evolutionary histories of cells within a tumor.

As cancer cells divide, they accrue somatic alterations that, in principle, offer tantalizing opportunities to infer evolutionary relationships among the genetically distinct cell populations (clones) that comprise a tumor. In practice, however, reconstructing a phylogenetic tree representing these ancestries is complicated because samples from bulk tumors contain a mixture of cells, where both the number of clones and their relative proportions are unknown. In this issue of *Cell Systems*, El-Kebir and colleagues (El-Kebir et al., 2016) describe an elegant mathematical formulation for the problem of "deconvolving" genome sequencing data obtained from such a mixture of clones, and they present a combinatorial algorithm to reconstruct individual tumor phylogenies from multi-region sequencing data. This important methodological advance more fully exploits the spectrum of somatic alterations that arise during tumor evolution toward a better understanding of intra-tumor heterogeneity.

An accurate and quantitative picture of intra-tumor heterogeneity is a major focus of current research. This genetic heterogeneity encodes the evolutionary history of the tumor and has important clinical implications, as sampling bias can obscure the interpretation of genomic profiles, and elevated heterogeneity may be associated with poor treatment response. In an effort to characterize intra-tumor heterogeneity and infer tumor ancestry, several studies have profiled multiple regions of the same tumor (Gerlinger et al., 2012, Sottoriva et al., 2013, McPherson et al., 2016) and even single cancer glands (Sottoriva et al., 2015), revealing important complementary molecular information. Phylogenetic trees are the canonical structure for representing tumor ancestries, where clones signify nodes. But in such multi-region sequencing strategies, it is essential to account for cellular admixture and the fact that somatic single-nucleotide variants occur in the presence of copy number alterations; otherwise, tumor phylogenies inferred from these data may be inaccurate.

Early efforts more than two decades ago used somatic microsatellite markers as a molecular clock toward reconstruction of tumor phylogenies (Shibata et al., 1996), borrowing from approaches to compare natural populations (Takezaki and Nei, 1996). Since then, high-throughput genomic profiling has enabled characterization of intra-tumor heterogeneity at increasing resolution, and it is now appreciated that this is a common feature of diverse solid tumors (Marusyk et al., 2012, Gerlinger et al., 2012, Sottoriva et al., 2013, 2015, McPherson et al., 2016). Numerous computational approaches aimed at inferring tumor phylogenies from single or multi-region bulk sequencing data have recently been proposed. Most of these methods utilize the variant allele fraction or cancer cell fraction for somatic single-nucleotide variants restricted to diploid regions to infer a two-state perfect phylogeny, assuming an infinite-site model such that each site can mutate only once and persists. In practice, convergent evolution could result in the acquisition of the same mutation more than once, thereby violating this assumption. Similarly, mutations could be lost due to loss of heterozygosity.

Indeed, both single-nucleotide variants and copy number alterations arise during tumor evolution, and both the variant allele fraction and cancer cell fraction depend on the copy number state whose inference reciprocally relies on the relative ordering of these alterations such that joint analysis can help resolve their ancestral relationship (Figure 1). To tackle this outstanding problem, El-Kebir et al. (2016) formulated the multi-state perfect phylogeny mixture deconvolution problem to infer clonal genotypes, clonal fractions, and phylogenies by simultaneously modeling single-nucleotide variants and copy number alterations from multi-region sequencing of individual tumors. Based on this framework, they present SPRUCE (Somatic Phylogeny Reconstruction Using Combinatorial Enumeration), an algorithm designed for this task. This new approach uses the concept of a "character" to represent the status of a variant in the genome. Commonly, binary characters have been used to represent single-nucleotide variants—that is, the variant is present or absent. In contrast, El-Kebir et al. use multi-state characters to represent copy number alterations, which may be present in zero, one, two, or more copies in the genome.

SPRUCE outperforms existing methods on simulated data, yielding higher recall rates under a variety of scenarios. Moreover, it is more robust to noise in variant allele frequency estimates, which is a significant feature of tumor genome sequencing data. Importantly, El-Kebir and colleagues demonstrate that there is often an ensemble of phylogenetic trees consistent with the underlying data. This uncertainty calls for caution in deriving definitive conclusions about the evolutionary process from a single solution. Along these lines, the two-state perfect

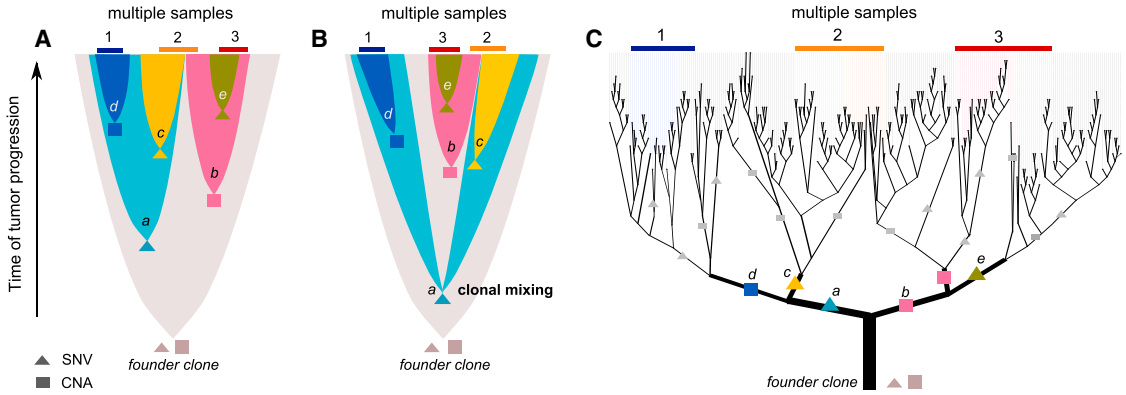


**Figure 1. Tumor Phylogeny Reconstruction from Multi-region Bulk Sequencing Data**

(A) Schematic representation of tumor evolution where cells are related by a genealogical tree and growth occurs in the presence of spatial constraints. Each clone is defined by a distinct constellation of single-nucleotide variants (SNVs) and/or copy number alteratiosn (CNAs) and localizes to a particular region(s) of the tumor. For illustration, three bulk samples obtained at the time of primary tumor diagnosis are shown, each consisting of one or more clone, as indicated by the colored segments. Each SNV/CNA creates a new clone according to the infinite-allele model assumed in El-Kebir et al. (2016). Sample 1 is composed of clone *d*, sample 2 by clones *b* and *c*, and sample 3 by *b* and *e*. Although the mutation defining clone *a* can be detected in both sample 1 and sample 2, clone *a* is not detected in these two samples due to replacement by successive clones, *d* and *c*.

(B) While an equivalent phylogeny can be reconstructed as in (A), information on the topography of samples within the tumor, coupled with inference of the mutational timeline, can reveal patterns of clone mixing and can aid delineation of the underlying growth dynamics (Sottoriva et al., 2015). For example, clonal mixing in an early tumor (turquoise SNV clone) could give rise to patterns of genetic heterogeneity where the same somatic alteration(s) is detected in distant regions (samples 1 and 2) of the tumor.

(C) Schematic illustration of a tumor phylogenetic tree whose leaves correspond to mixtures of cells (clones) harboring somatic alterations in varied proportions and the edges describe their ancestral relationship. Phylogeny deconvolution aims to reconstruct the tree (including the relative timing of SNVs/CNAs) and mixing proportions from the somatic alterations that underlie the evolutionary process given *m* mixtures of the leaves of the tree. Light gray triangles and squares denote undetectable SNVs and CNAs, respectively.

phylogeny inferred from single-nucleotide variants may represent only one possible scenario that explains the copy number alteration data. Importantly, the simulation studies also formalize the intuition that inclusion of additional samples provides greater benefit than sequencing a small number of samples to greater depth, highlighting opportunities for improved study design.

Thus, El-Kebir et al. provide theoretical and conceptual advances to this challenging problem while pointing to several opportunities for further improvement. For example, although the infinite-allele model is more suitable for copy number alterations, allowing every copy number alteration to generate a new allelic type (i.e., copy number alterations may change state more than once but can change to the same state at most once), it is likely to be violated in tumors with defects in DNA damage repair pathways and high genomic instability. Such violations may occur in a significant proportion of tumors in practice, hindering accurate tree inference. As the authors note, more generalized phylogenetic models such as maximum parsimony may better capture somatic evolution but will require additional methodological developments for the specialized case of phylogenetic mixtures. However, in the absence of ground-truth human tumor phylogenies, comparison of the accuracy of different models is limited to simulation studies, and the biological interpretation of these processes remains challenging.

Like other methods, SPRUCE cannot accommodate the full set of somatic variants (characters) derived from whole-genome or exome-sequencing studies, thereby necessitating prior filtering and clustering of single-nucleotide variants. Although tree inference can be performed using relatively few variants, the identification of unique phylogenies is challenging. Moreover, since the number of possible tree shapes grows super-exponentially with the number of taxa (clones) (St. John, 2016), interpretation of the resultant solution space quickly becomes intractable. This could perhaps be aided by employing heuristics such as estimating similarities among tree shapes or statistical fit to the underlying data. As the authors propose, a promising future direction could leverage the combinatorial multi-state ancestry graph enumerated by SPRUCE to generate informative priors for use within a probabilistic framework such as Markov Chain Monte Carlo sampling, thereby drawing on the complementary strengths of these approaches.

These points highlight the intrinsic challenges associated with clonal deconvolution from bulk sequencing data, which is aggravated by extensive genetic diversity, sampling bias, and uncertainty in single-nucleotide variant and copy number alteration estimates. Indeed, the extent of intra-tumor heterogeneity is vastly underestimated, as has been demonstrated in colorectal tumors, which often exhibit Big Bang dynamics, wherein after transformation of the founding clone, the tumor grows in the absence of stringent selection, compatible with effectively neutral evolution (Sottoriva et al., 2015). Hence, for tumors that follow a Big Bang model, timing is the primary determinant of mutational frequency, such that late-arising mutations will be largely undetectable (Figure 1C). Theoretical approaches also suggest that nearly every tumor cell may be genetically distinct (Ling et al., 2015). This abundant genetic diversity provides an opportunity to infer tumor ancestry but also poses challenges for the conceptual definition of a clone if all cells are unique.

Single-cell sequencing mitigates many of the above issues by enabling direct analysis of the unit of interest but poses additional methodological problems due to noisy and incomplete measurements. Currently, accurate, scalable, and complete single-cell tumor genomes are limited by practical and technical considerations. Nonetheless, certain questions may be informed by a combination of bulk and single-cell sequencing (McPherson et al., 2016). Additionally, new technologies enabling simultaneous mutational and transcriptional profiling of the same cell (Macaulay et al., 2015) will facilitate mapping of clonal tumor genotypes to phenotypic changes. Another paper in this issue (Li et al., 2016) describes an approach for simultaneously calling structural variants and allele-specific copy number alterations within a cancer genome. Accurate identification of somatic variants is a critical input to methods such as those of El Kebir et al.

It is important to appreciate that the inferred tree alone does not provide a complete view of a tumor's evolutionary history and dynamics. For instance, early clonal mixing may contribute to spatial heterogeneity in certain cancers, as we have observed in colon cancer (Sottoriva et al., 2015) and as illustrated in Figure 1B. However, without topographical information regarding the origin of a sample within a tumor, the reconstructed phylogeny cannot be distinguished from that arising in the absence of clone mixing (Figure 1A). Since solid tumors exhibit hierarchical tissue structure and are not well-mixed populations, such spatial organization is likely to be generally important.

Crucially, the resultant patterns of intratumor heterogeneity reflect the evolutionary forces that gave rise to them and may be exploited to infer both phylogenies and patient-specific tumor dynamics within a spatial computational model of tumor growth, as we have previously shown (Sottoriva et al., 2015). Such approaches can also help to illuminate the range of evolutionary trajectories that occur at different stages of tumor progression in distinct cancer types, including gradual (linear) progression versus punctuated evolution. These strategies can also be applied to delineate outstanding questions concerning mechanisms of metastatic progression and the extent of clonal dynamics under treatment selective pressure. Ultimately, the interpretation of such data within population genetic models will provide quantitative insights into human tumor evolution toward the development of predictive models and patient-tailored treatment strategies.

## ACKNOWLEDGMENTS

The authors thank Ruping Sun and Chris Probert for useful discussions.

## REFERENCES

El-Kebir, M., Satas, G., Oesper, L., and Raphael, B.J. (2016). Cell Syst. *3*, this issue, 43–53.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). N. Engl. J. Med. *366*, 883–892.

Li, Y., Zhou, S., Schwartz, D.C., and Ma, J. (2016). Cell Syst. *3*, this issue, 21–34.

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al. (2015). Proc. Natl. Acad. Sci. USA *112*, E6496–E6505.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). Nat. Methods *12*, 519–522.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Nat. Rev. Cancer *12*, 323–334.

McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A.W., Ha, G., Biele, J., Yap, D., Wan, A., et al. (2016). Nat. Genet. *48*, 758–767.

Shibata, D., Navidi, W., Salovaara, R., Li, Z.H., and Aaltonen, L.A. (1996). Nat. Med. *2*, 676–681.

Sottoriva, A., Spiteri, I., Piccirillo, S.G., Touloumis, A., Collins, V.P., Marioni, J.C., Curtis, C., Watts, C., and Tavaré, S. (2013). Proc. Natl. Acad. Sci. USA *110*, 4009–4014.

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., and Curtis, C. (2015). Nat. Genet. *47*, 209–216.

St. John, K. (2016). Syst. Biol., Published online June 22, 2016. http://dx.doi.org/10.1093/sysbio/syw025.

Takezaki, N., and Nei, M. (1996). Genetics *144*, 389–399.