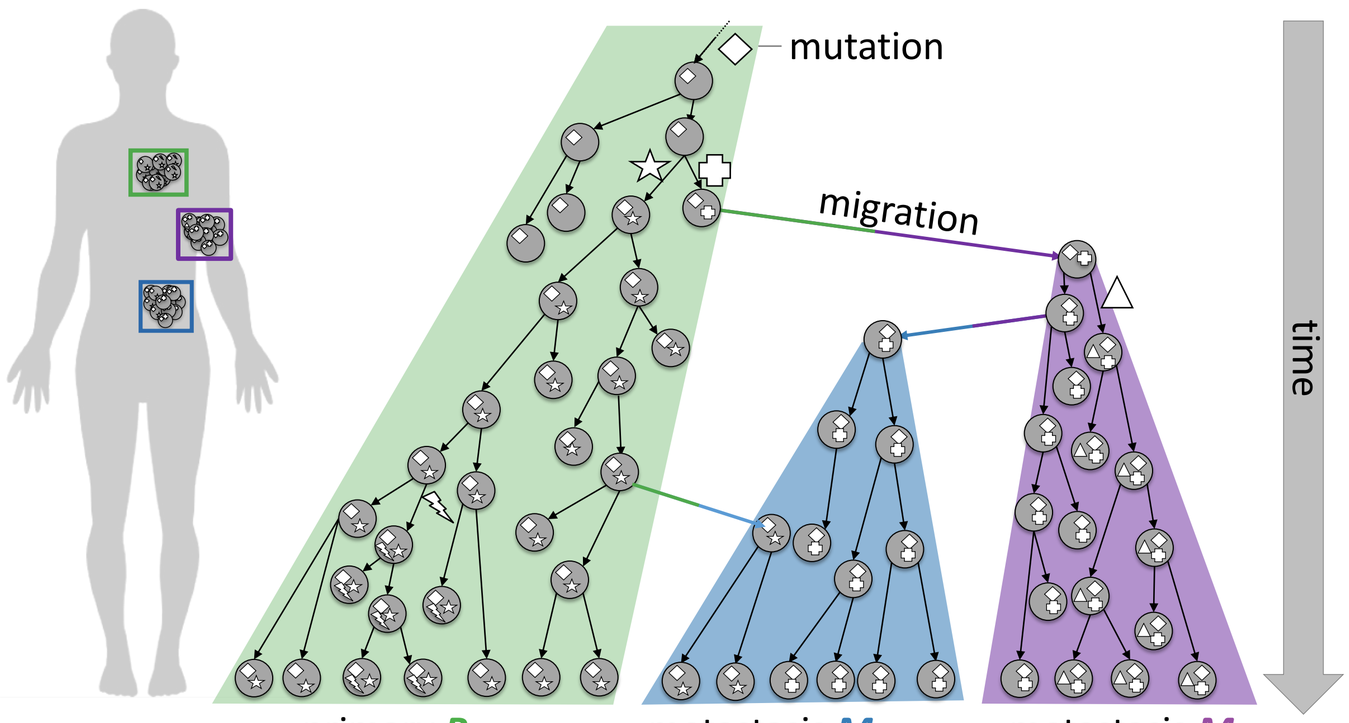
Combinatorial Algorithms for Tumor Phylogenetics

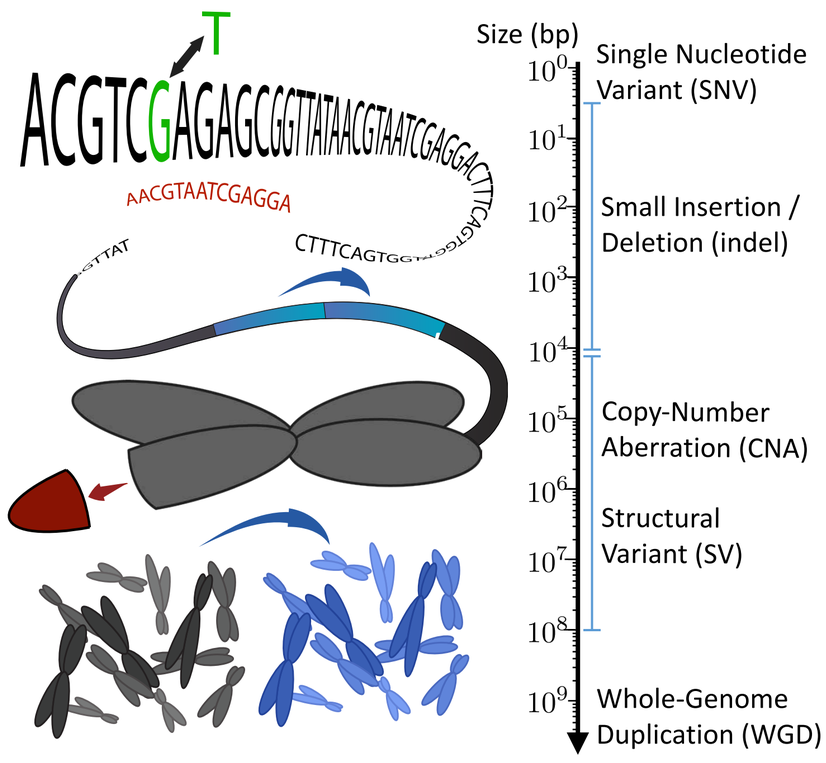
Mohammed El-Kebir[1]

[1]Department of Computer Science, UIUC

(A) Cellular history of a tumor: cell division, mutation and migration

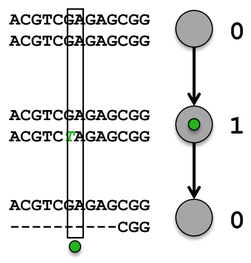


(B) Somatic mutations occur at distinct genomic scales



(C) Tumor phylogeny estimation from single-cell DNA-seq data with SPhyR (Single-cell Phylogeny Reconstruction)

SNVs can be **lost** due to CNAs

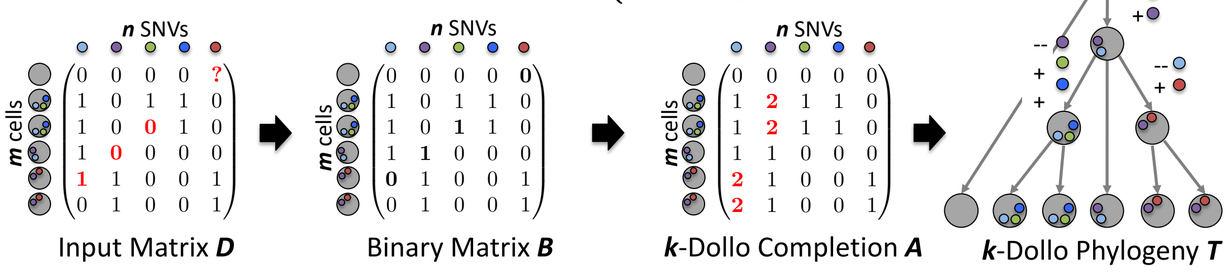


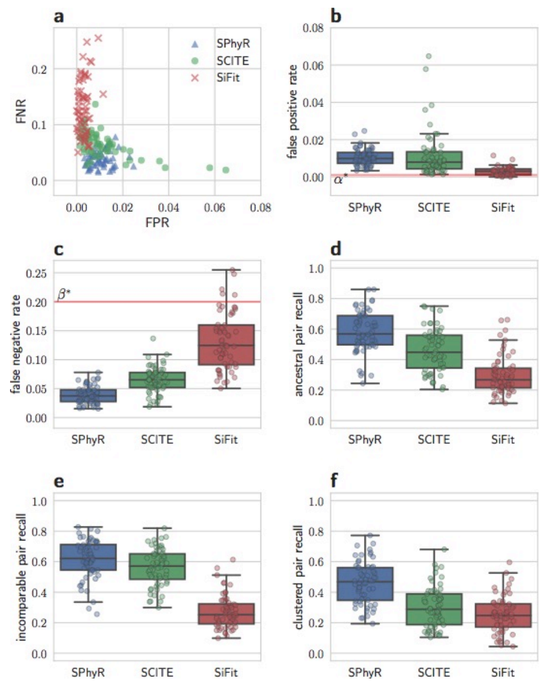
Infinite sites assumption:

- No parallel evolution of SNVs
- No loss of SNVs
- SCITE [Jahn et al. 2016]
- OncoNEM [Ross and Markowetz, 2016]

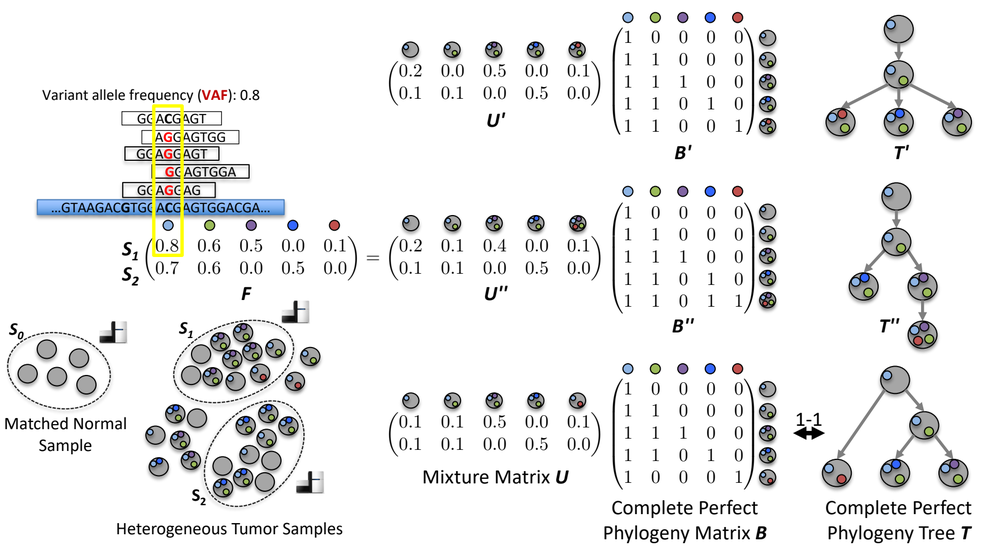
***k*-Dollo Phylogeny Flip and Cluster (*k*-DPFC) problem.** Given matrix $D \in \{0,1,?\}^{m \times n}$, error rates $\alpha, \beta \in [0,1]$, integers $k, s, t \in \mathbb{N}$, find matrix $B \in \{0,1\}^{m \times n}$ and tree T such that: (1) B has at most s unique rows and at most t unique columns; (2) $\Pr(D \mid B, \alpha, \beta)$ is maximum; and (3) T is a k-Dollo phylogeny for B.

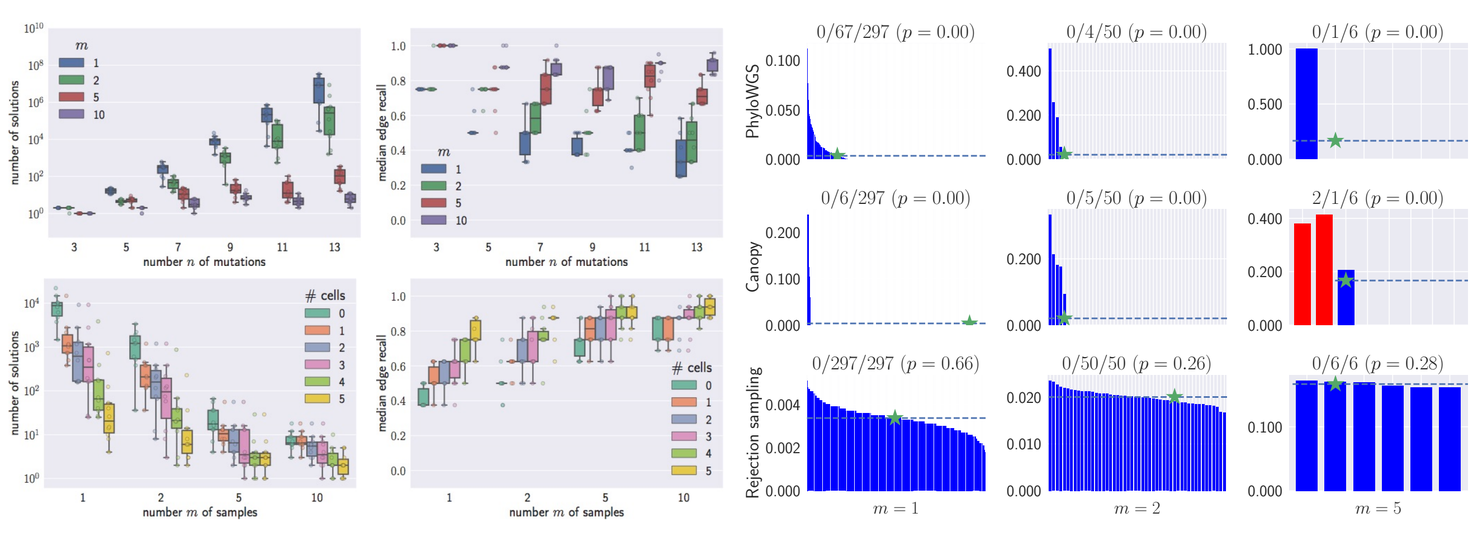
$$\Pr(D \mid B, \alpha, \beta) = \prod_{p=1}^{m} \prod_{c=1}^{n} \begin{cases} \alpha, & d_{p,c} = 1 \text{ and } b_{p,c} = 0 \\ 1-\alpha, & d_{p,c} = 1 \text{ and } b_{p,c} = 1, \\ \beta, & d_{p,c} = 0 \text{ and } b_{p,c} = 1, \\ 1-\beta, & d_{p,c} = 0 \text{ and } b_{p,c} = 0, \\ 1, & d_{p,c} = ? \end{cases}$$





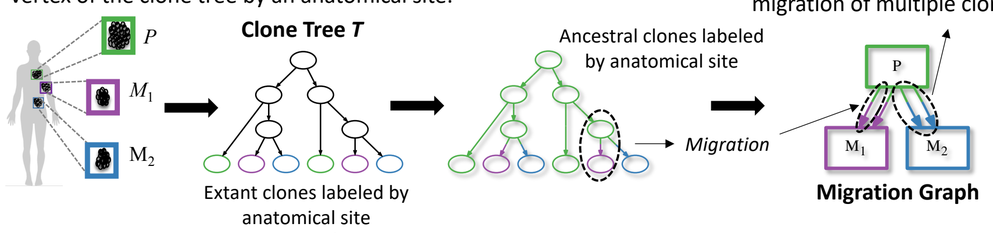
(D) Non-uniqueness in tumor phylogeny estimation from bulk DNA-seq data





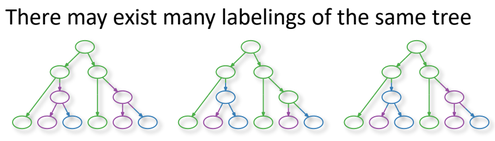
(E) Migrations and comigrations distinguish migration patterns

The migration history is determined by a labeling of each vertex of the clone tree by an anatomical site.

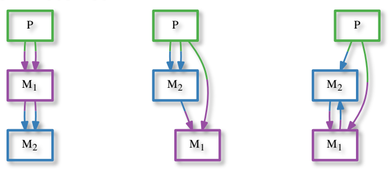


Comigration: simultaneous migration of multiple clones

There may exist many labelings of the same tree



and thus many possible migration histories.



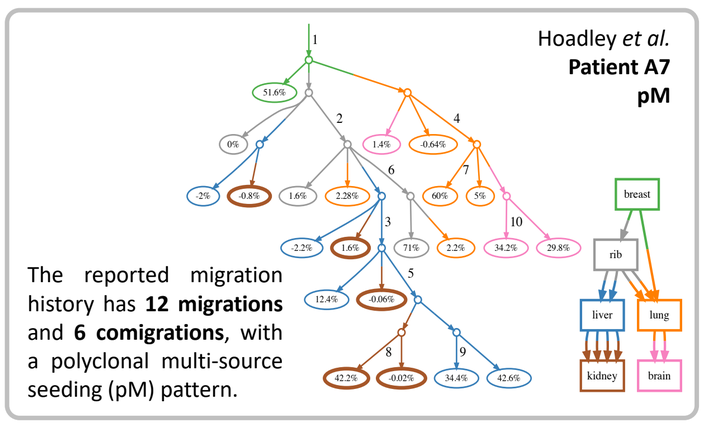
We describe the complexity of a migration history according to:

1. The number of migrations
2. The number of comigrations
3. The migration pattern

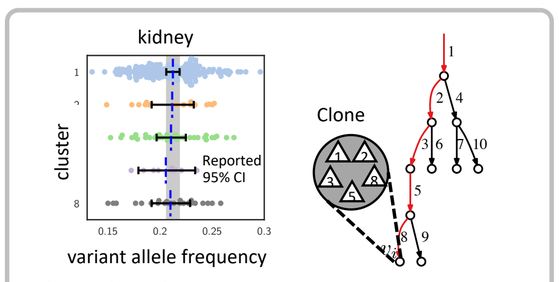
Taxonomy of Migration Patterns

	single-source seeding (S)	multi-source seeding (M)	reseeding (R)
monoclonal (m)	tree	directed acyclic graph	directed graph
polyclonal (p)	multi-tree	directed acyclic multi-graph	directed multi-graph

(F) Simultaneous analysis of mutation and migration history reveals monoclonal single-source seeding in breast cancer

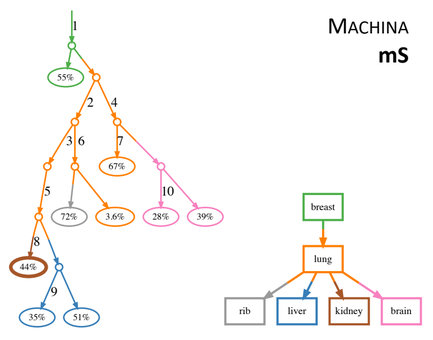


The reported migration history has **12 migrations** and **6 comigrations**, with a polyclonal multi-source seeding (pM) pattern.



Jointly inferring the present clones and the migration history suggests the presence of fewer clones, including a single clone in kidney.

In patient A7, MACHINA infers a **monoclonal single-source seeding (mS) migration history**, the simplest possible migration history, with **5 migrations and 5 comigrations** implying that the sequencing data does not strongly support complicated polyclonal migration patterns in this patient.



El-Kebir, M., Oesper, L., Acheson-Field, H., & Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, *31*(12), i62–i70.
El-Kebir, M., Satas, G., Oesper, L., & Raphael, B. J. (2016). Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, *3*(1), 43–53.
El-Kebir, M., Satas, G., & Raphael, B. J. (2018). Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, *50*(5), 718–726.
El-Kebir, M. (2018). SPhyR: Tumor Phylogeny Estimation from Single-Cell Sequencing Data under Loss and Error. *Bioinformatics/ECCB 2018*. To appear.
Pradhan, D., & El-Kebir, M. (2018). On the non-uniqueness of solutions to the perfect phylogeny mixture problem. *In submission.*