Algorithms for Phylogenetic Tree Correction in Species and Cancer Evolution

Dissertation Defense

Sarah Christensen

November 6, 2020

Dissertation Idea

Develop biologically *meaningful optimization* problems
with corresponding *efficient algorithms*
that leverage *auxiliary data* to address challenges
in species and tumor *phylogeny estimation*.

Completed Work at Prelim

Species phylogenies

Chapter 1. Christensen S., Molloy E.K., Vachaspati P. & Warnow T. (2018). OCTAL: Optimal Completion of Gene Trees in Polynomial Time. *Algorithms for Molecular Biology*.

Chapter 2. Christensen S., Molloy E.K., Vachaspati P., Yammanuru A. & Warnow T. (2020). Non-parametric correction of estimated gene trees using TRACTION. *Algorithms for Molecular Biology*.

Tumor phylogenies

Chapter 3. Christensen S., Leiserson M.D.M., & El-Kebir M. (2020). PhySigs: Phylogenetic Inference of Mutational Signature Dynamics. *Pacific Symposium on Biocomputing*.

Chapter 4. Christensen S., Kim J., Koyejo S., Chia N. & El-Kebir M. (2020). Detecting Evolutionary Patterns of Cancers using Consensus Trees**.** [Submitted to ECCB 2020].

New Work Since Prelim

Tumor phylogenies

Chapter 3. Developed R package and visualization tool.

Github: *https://github.com/elkebir-group/PhySigs_R*

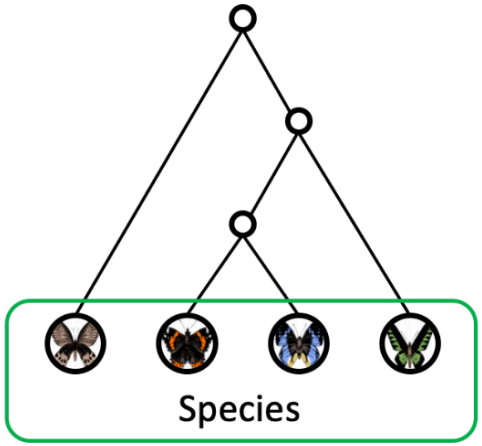
Visualization App: *https://physigs-tree-browser.herokuapp.com/*

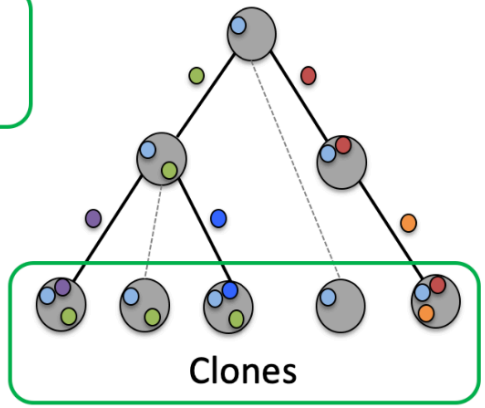
Chapter 4. Presented at ECCB and published in *Bioinformatics*.

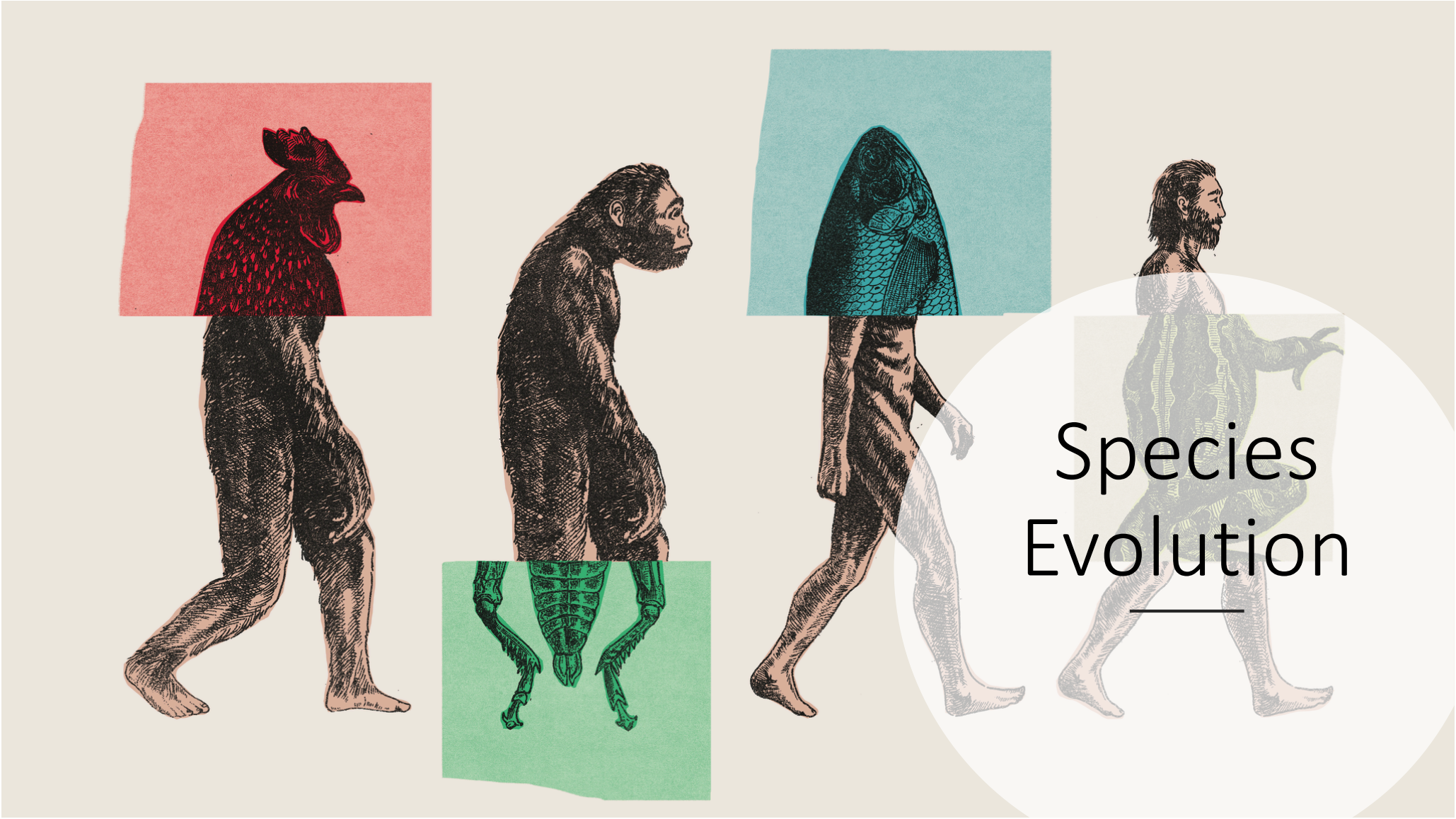
Christensen S. & El-Kebir M. (2020). Expanding Detection of Evolutionary Patterns of Cancers to Broader Biologically Realistic Conditions. *Bioinformatics*.

Overview of Talk

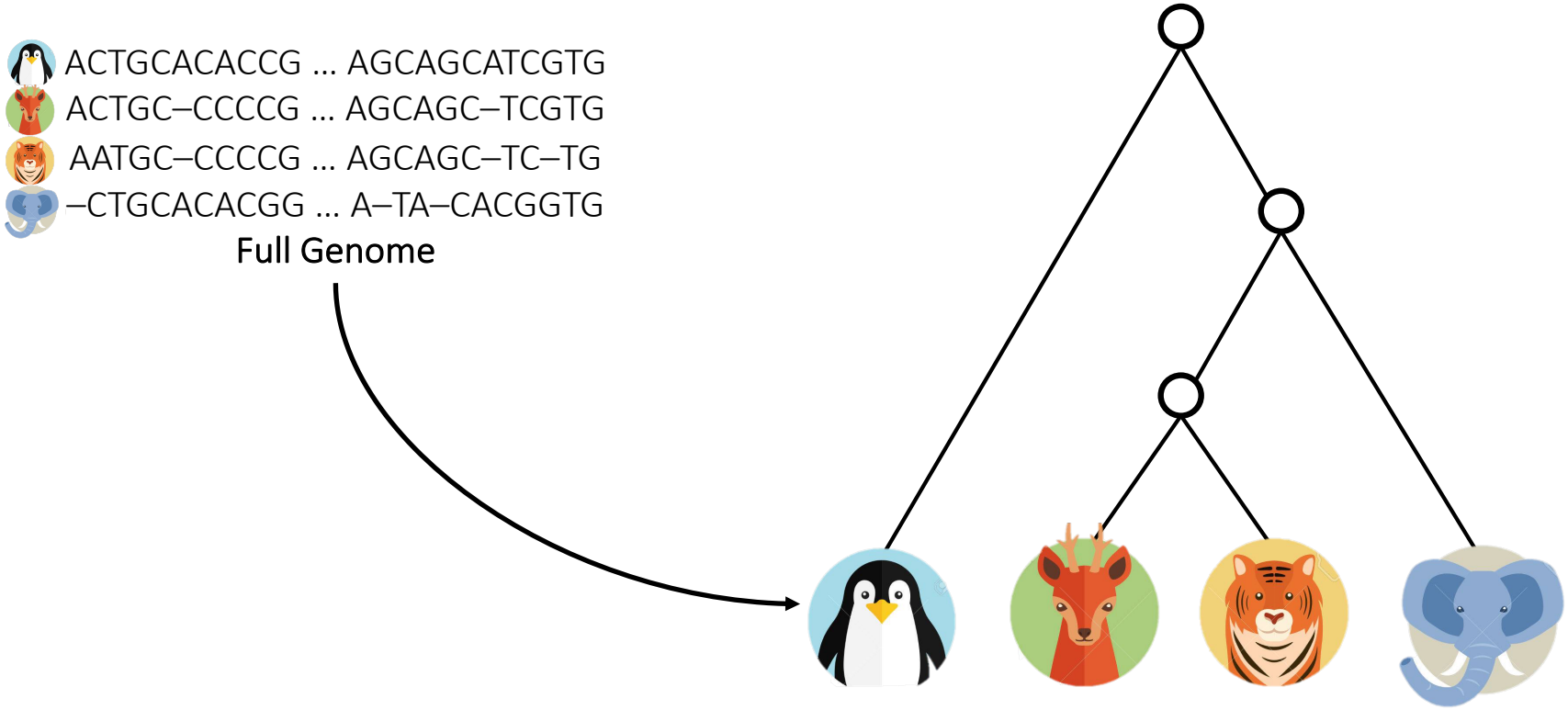
- Species Evolution
  - Background
  - Contributions from Chapter 2
- Tumor Evolution
  - Background
  - Contributions from Chapter 4
- Conclusions






Species
Evolution

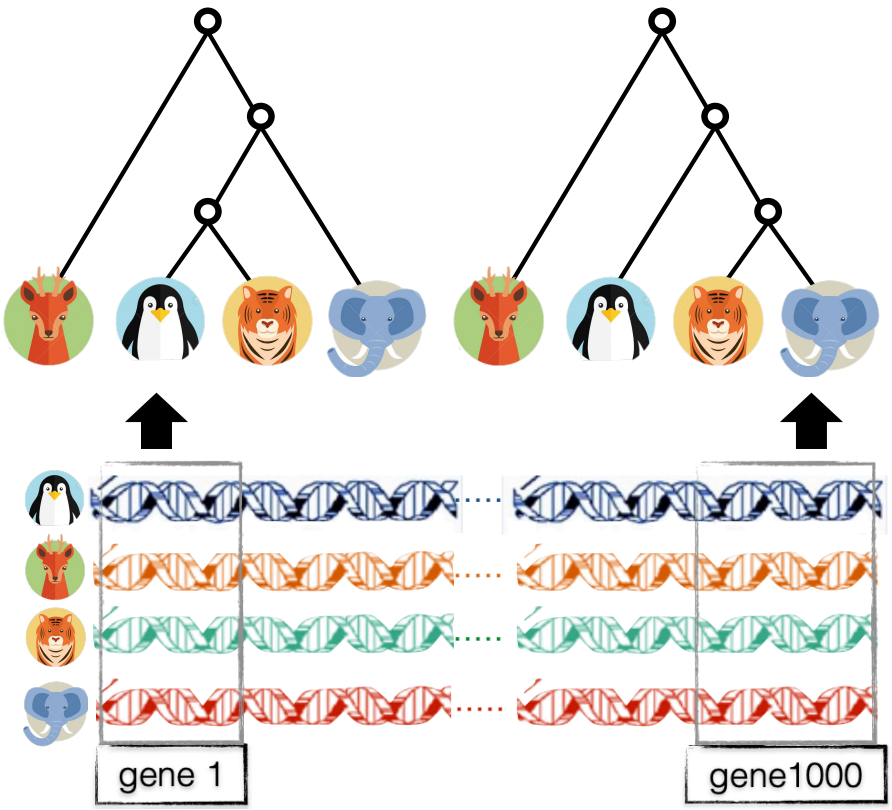
Species Tree



Gene Trees

- Gene trees may differ from each other as well as from the species tree

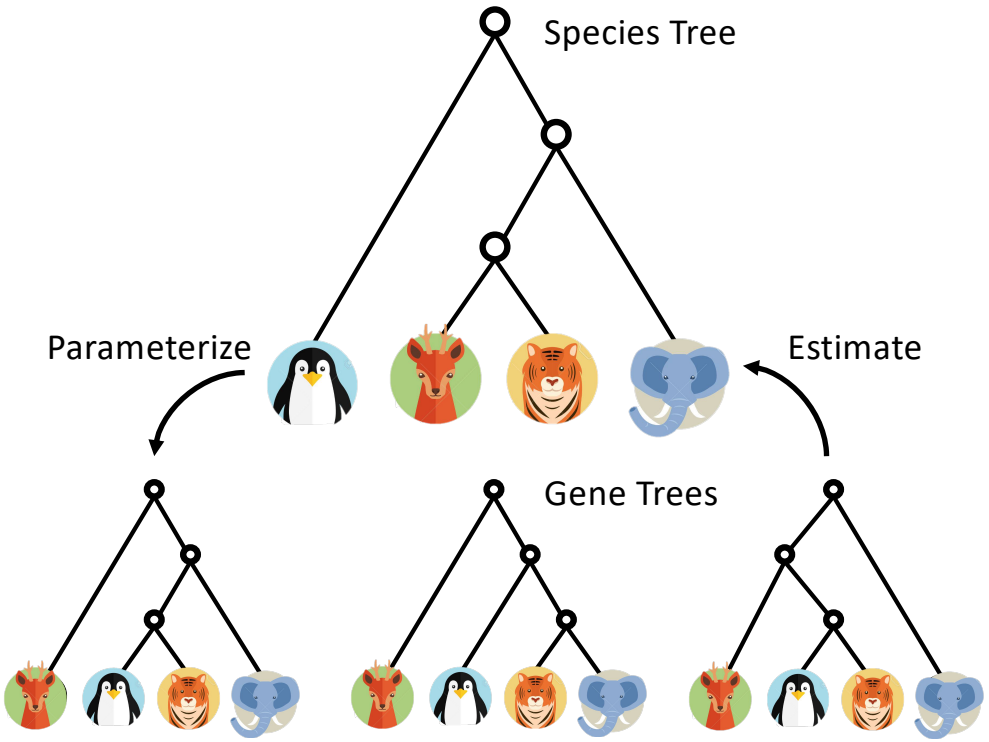
- Causes of tree heterogeneity
  - Incomplete lineage sorting (ILS)
  - Gene duplication and loss (GDL)
  - Horizontal gene transfer (HGT)



Gene and Species Trees Related

- Species tree parameterizes distribution over gene trees under models of gene evolution.

- Gene trees may likewise be used to recover a species tree.



[Pamilo and Nei, 1988; Rannala and Yang, 2003]

Challenges for Gene Tree Estimation

- Estimated gene trees can have missing species as well as low-confidence branches.
  - Avian Phylogenomic Project average branch support below 30% [Jarvis et al., 2014]
- These challenges may impact downstream analysis.
- Idea: Can we improve gene tree estimation by using species trees?

Leading Gene Tree Correction Methods All Assume GDL

- **ecceTERA** [Jacox et al, 2016]
- **NOTUNG** [Durand, 2006]
- **ProfileNJ** [Noutahi et at, 2016]
- **TreeFix** [Bansal et at, 2015]

- Gene tree correction methods just use the topology of a species tree to improve the gene tree.
- Integrative methods also incorporate sequencing data.

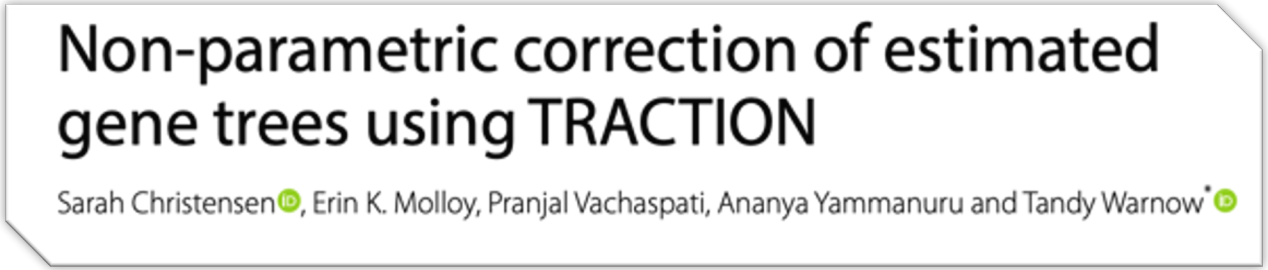
Our Non-Parametric Approach

Chapter 1: We add in missing species using a reference tree with OCTAL.

[Christensen et al., *Algorithms For Molecular Biology* 2018]

Chapter 2: We add in missing species and correct low-support branches using a reference tree with TRACTION.

[Christensen et al., *Algorithms For Molecular Biology* 2020]


Non-parametric correction of estimated gene trees using TRACTION

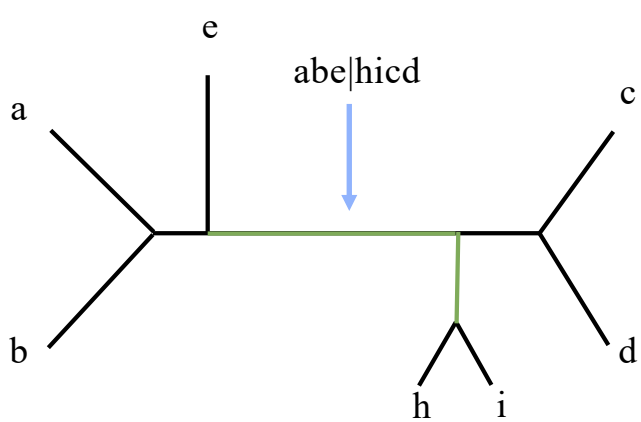
Sarah Christensen, Erin K. Molloy, Pranjal Vachaspati, Ananya Yammanuru and Tandy Warnow*

Chapter 2: TRACTION

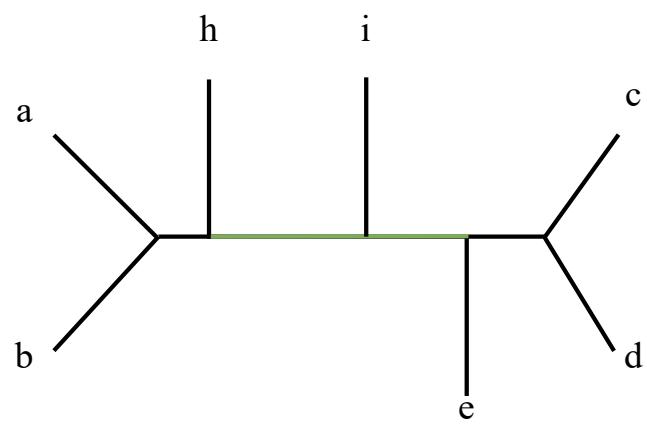
Christensen S., Molloy E.K., Vachaspati P., Yammanuru A. & Warnow T. (2020). Non- parametric correction of estimated gene trees using TRACTION. *Algorithms for Molecular Biology*.

Robinson-Foulds (RF) Distance

Given two trees on the same leaf set, the *RF distance* is the total number of unique bipartitions in each tree.



Here,
RF distance is 4



RF-OTRC Optimization Problem

The Optimal Tree Refinement and Completion Problem

Input: An unrooted, singly-labeled, binary tree T on leaf set S and an unrooted, singly-labeled tree t on $R \subseteq S$.

Output: An unrooted, singly-labeled, binary tree T' on S with two key properties:

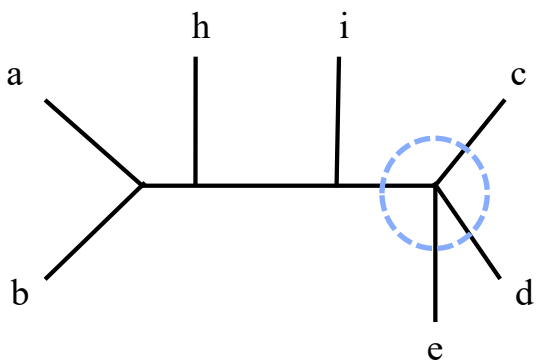
1. T' contains all the leaves of S and is compatible with t (i.e., $T'|_R$ is a refinement of t) and
2. T' minimizes the RF distance to T among all binary trees satisfying condition (1).

RF-OTRC Optimization Problem

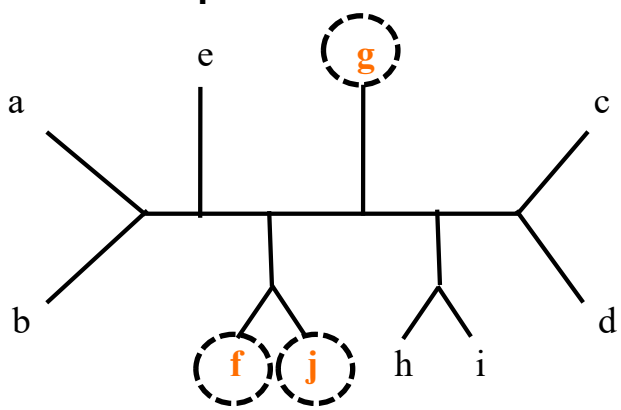
Inputs

Output

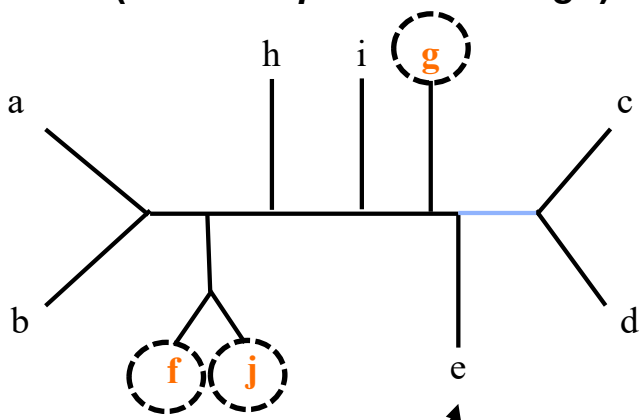
Gene tree *t* on *R*⊆*S*



Species tree *T* on *S*



***T'* (an *S-completion* refining *t*)**



RF distance between T and T' is 8.
This is optimal.

TRACTION: Main Contributions

Theorem: TRACTION solves the RF-OTRC Problem in $O(n^{1.5} \log n)$ time where n is the number of leaves in the reference tree.

Generalization to multi-label trees: We show a naïve generalization is possible, but can produce degenerate results.

Empirical results: Simulation studies show some advantages over leading methods. We will show this here.

Questions for Simulation Study

- Can we improve estimated gene trees with estimated species trees?
- Which correction methods perform best and under what conditions?
- How do model conditions impact absolute and relative performance?

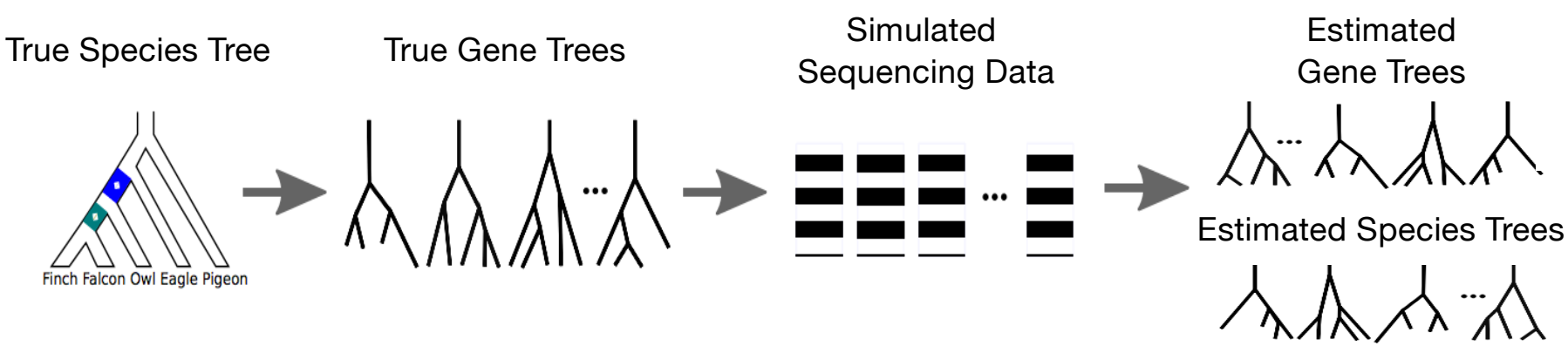
TRACTION: Tree Correction Simulation Design

ILS only datasets

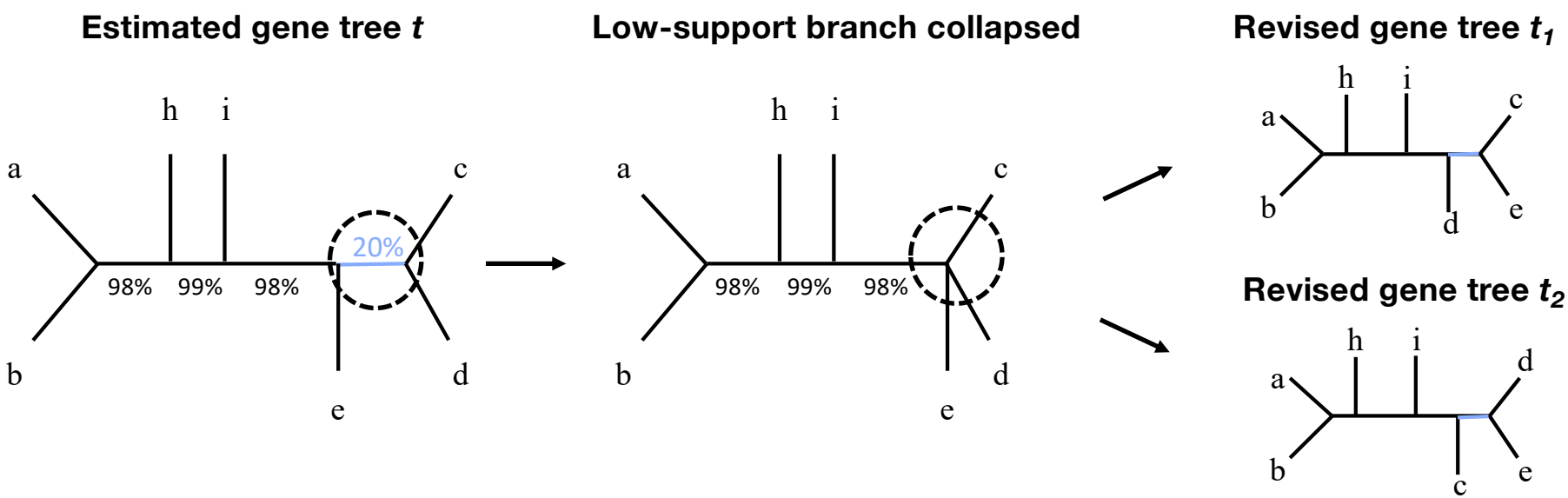
- 26 species per true gene tree
- 8,000 gene trees in total
- 2 levels of ILS; varying sequence lengths

HGT+ILS datasets

- 51 species per true gene tree
- 60,000 gene trees in total
- 2 levels of HGT; 3 different sequence lengths

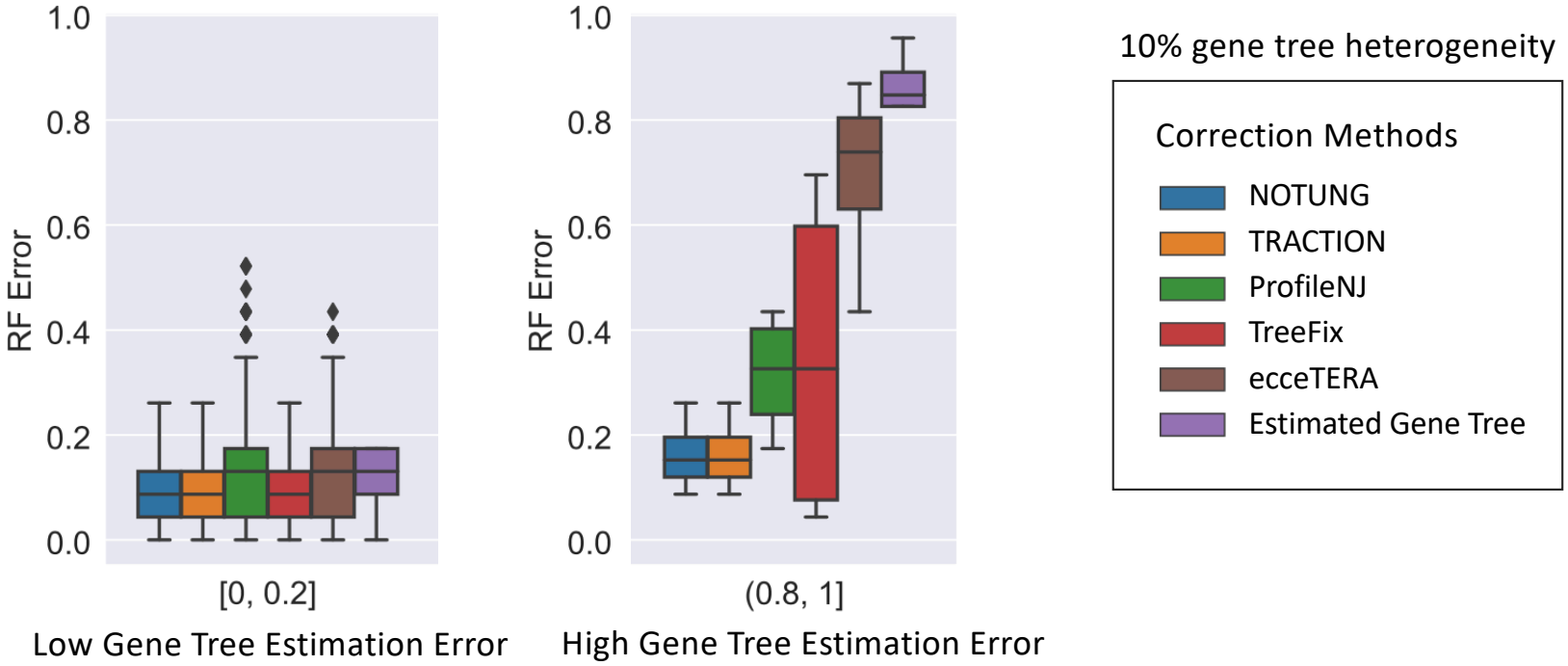


Correct Low-Support Branches with Species Tree



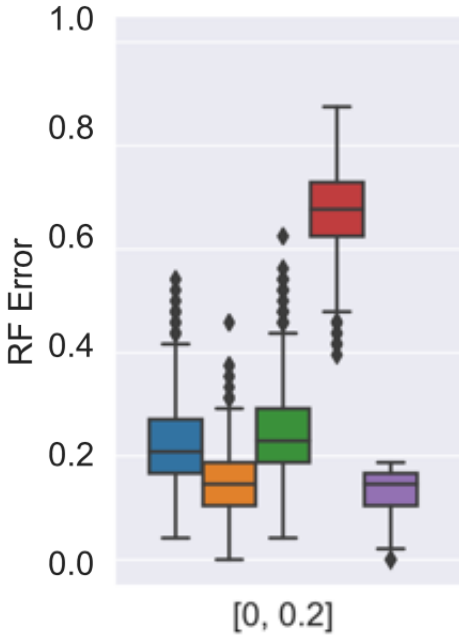
ILS-Only Results

All methods improve estimated gene trees, but NOTUNG and TRACTION improve the most.

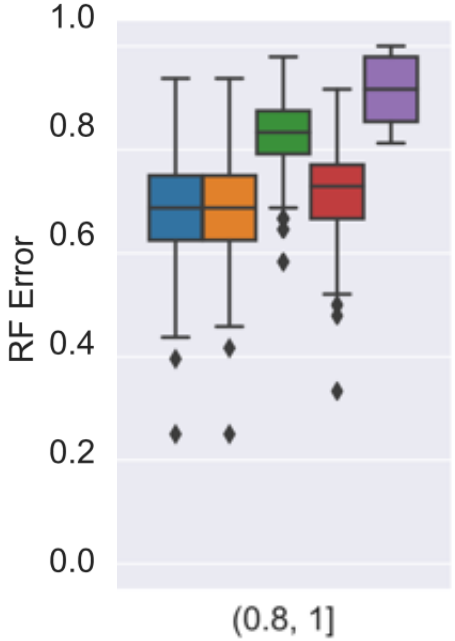


ILS+HGT Results

In many cases, correction methods reduce accuracy.
Only TRACTION consistently maintains or improves accuracy.



Low Gene Tree Estimation Error



High Gene Tree Estimation Error

68% gene tree heterogeneity

Correction Methods

- NOTUNG
- TRACTION
- ProfileNJ
- TreeFix
- ecceTERA
- Estimated Gene Tree

ecceTERA is not shown because it did not complete on this dataset

Answers from Simulation Study

- Can we improve estimated gene trees with estimated species trees?
  - Yes, in many cases.
- Which correction methods perform best and under what conditions?
  - NOTUNG and TRACTION consistently perform well
  - Slight advantage to TRACTION under HGT+ILS
- How do model conditions impact absolute and relative performance?
  - All methods perform well on ILS-only condition where ILS is low to moderate
  - NOTUNG and TRACTION performed best relative to other methods
  - TRACTION consistently maintains or improves accuracy on HGT+ILS

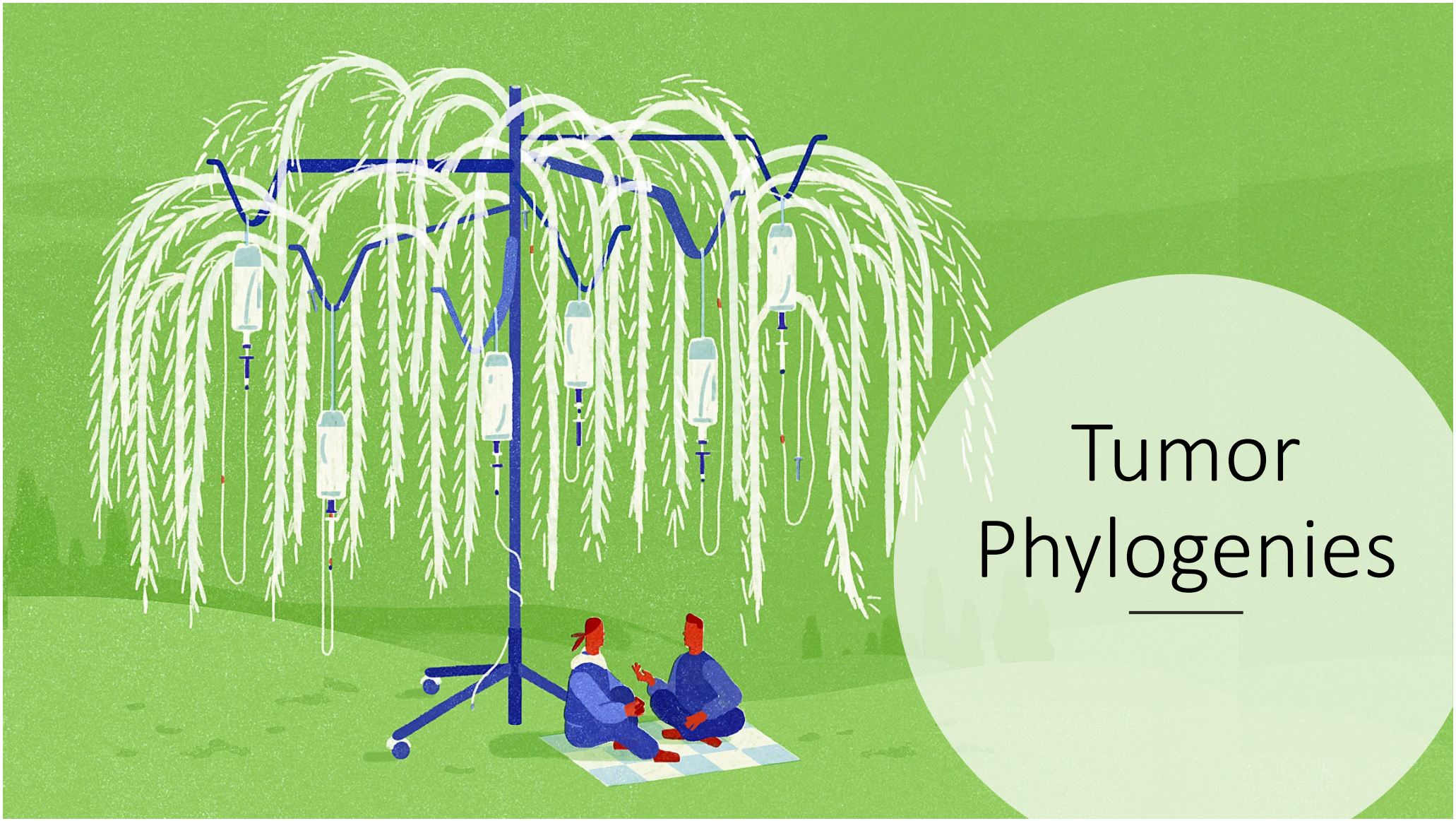
TRACTION: Future Directions

Explore other reference trees: We used species trees, but other types of reference trees should be tried.

Explore other distance measures: We used RF distance, but other distances could be tried.

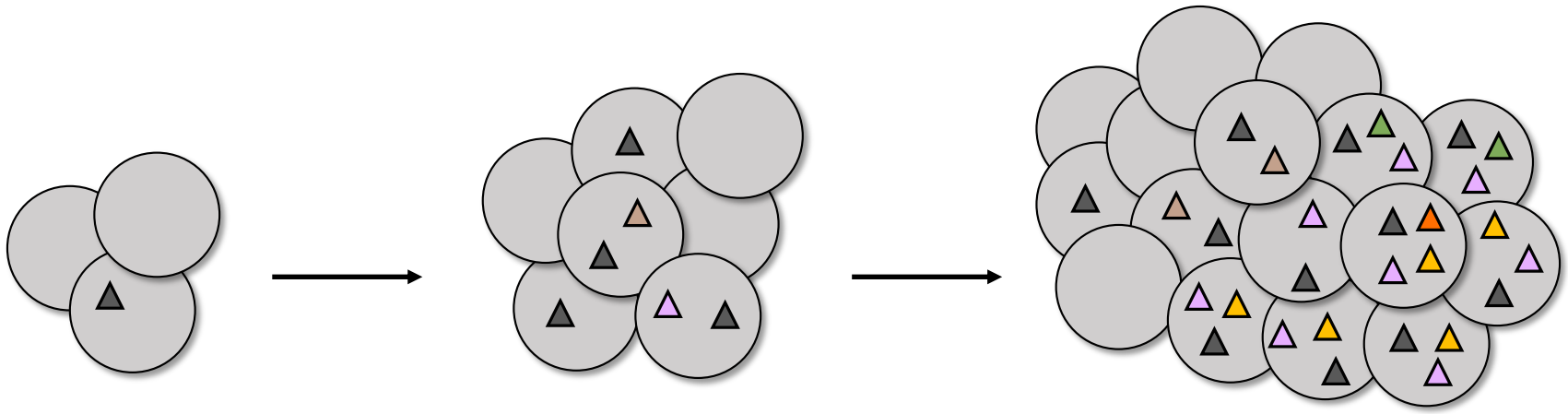
Measure impact on downstream analysis: The effect of gene tree correction on downstream tasks should be evaluated.

Continue to pursue multi-label trees: Other extensions of RF distance to multi-label tree have been proposed.



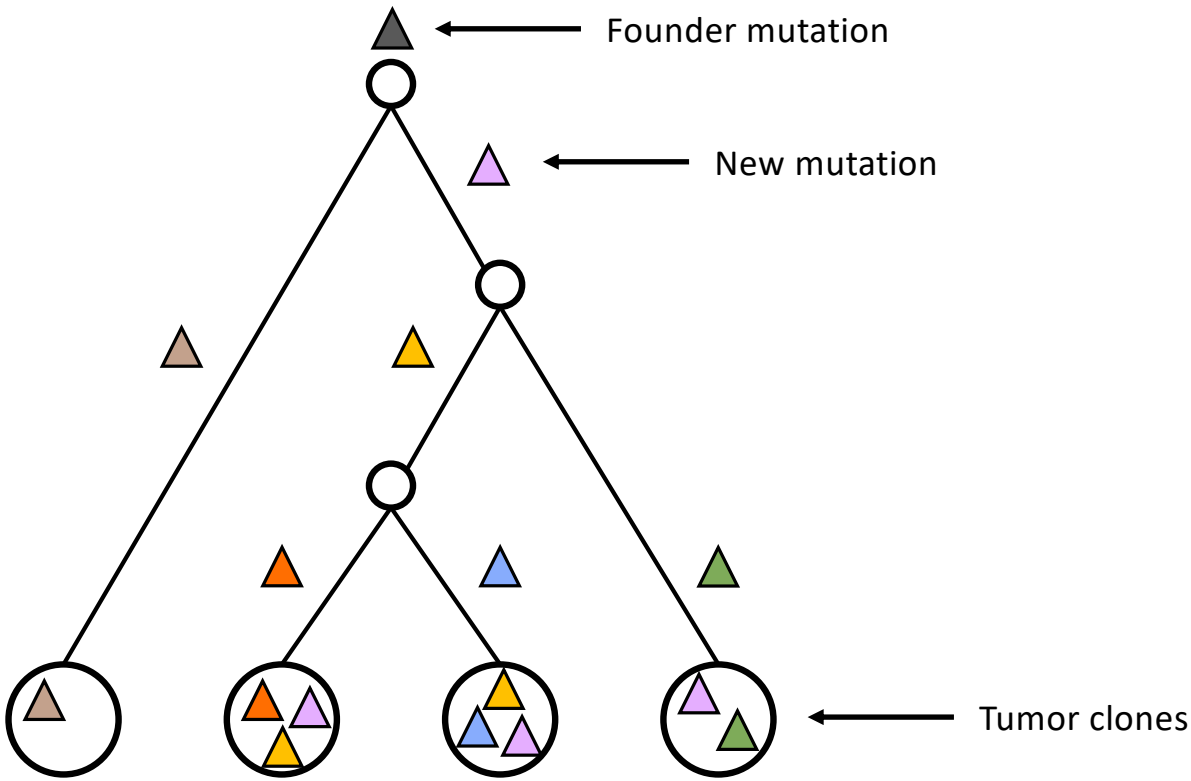
Tumor Phylogenies

Evolution in Cancer



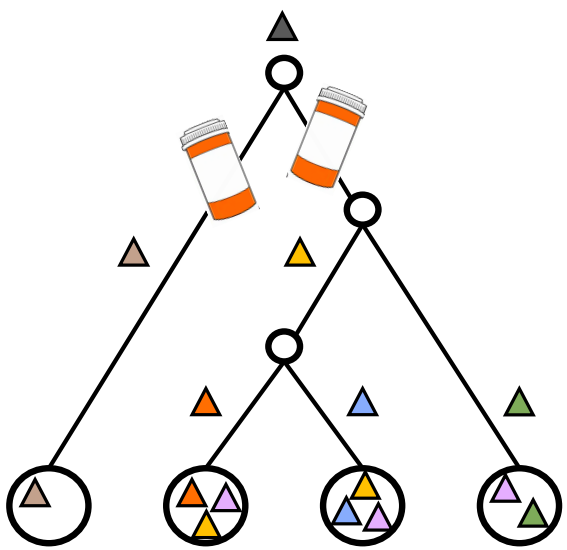
Clonal Evolution Theory of Cancer
[Nowell, 1976]

Phylogenetic Trees in Cancer



Downstream Analysis Requires Accurate Tumor Phylogeny Inference

Identify treatment targets



Understand metastasis

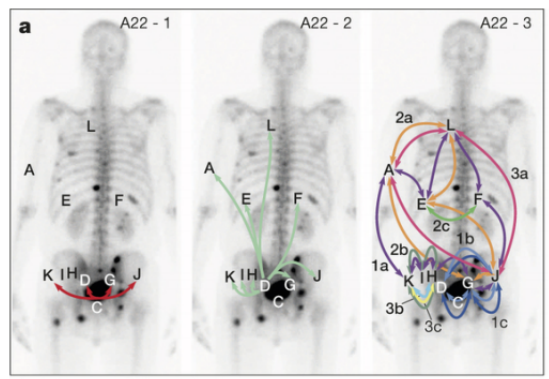


Image from [Gundem et al., 2015]

Find common patterns

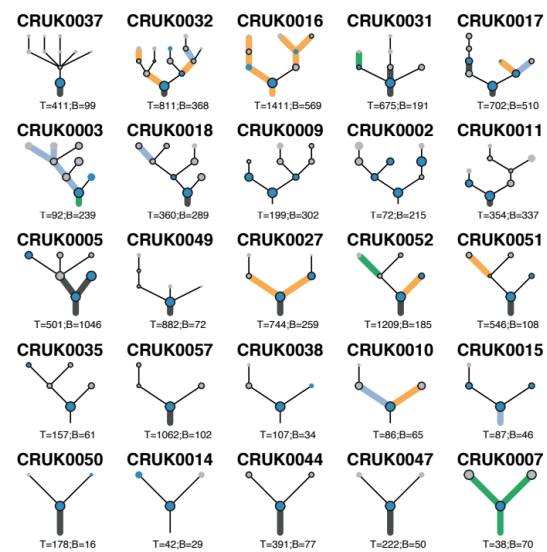
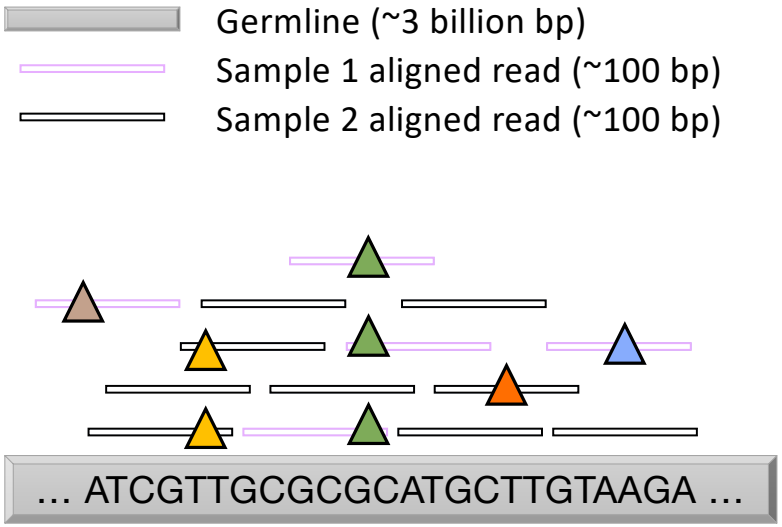
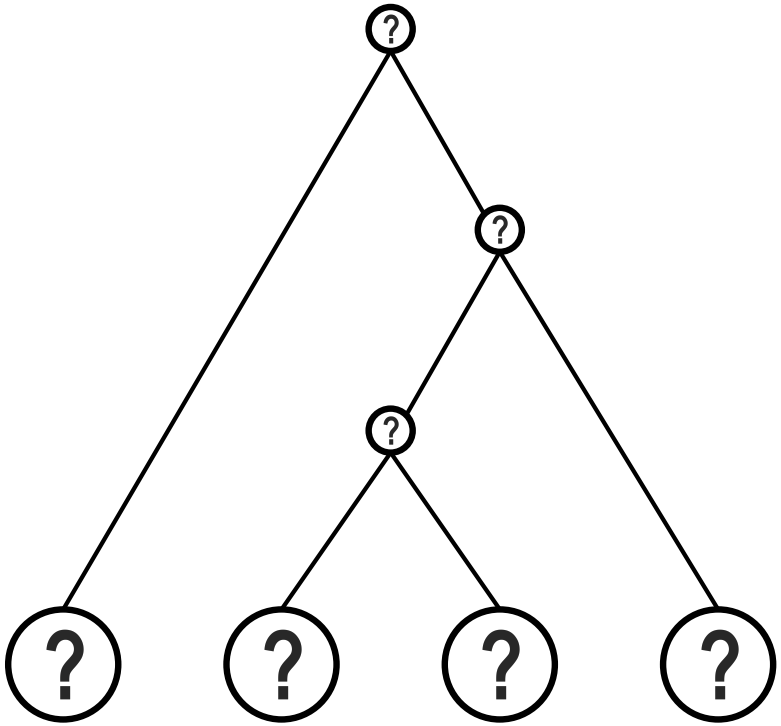


Image from [Jamal-Hanjani et al., 2017]

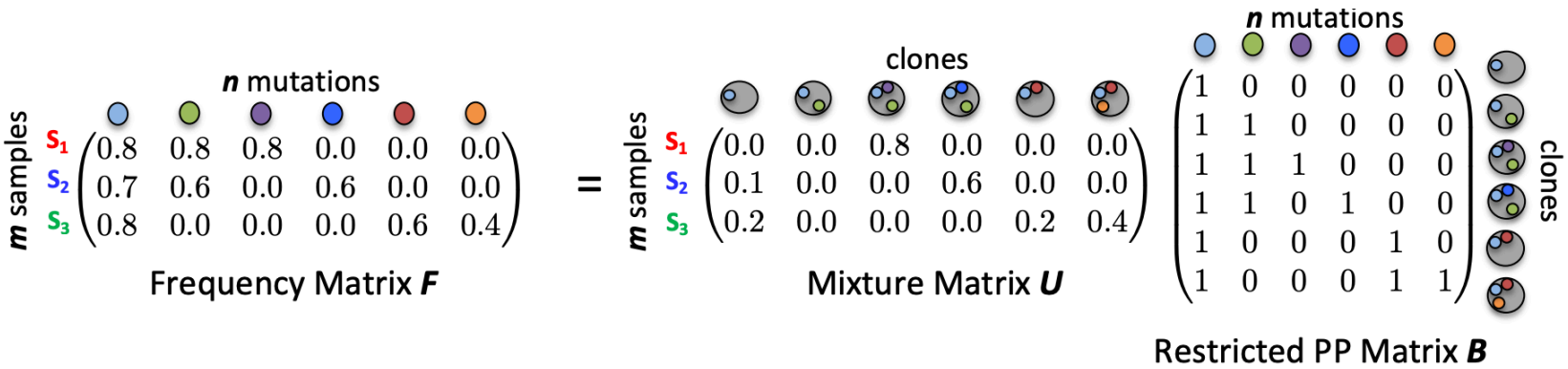
Bulk Sequencing Adds New Challenge



Only observe *mutations* and *frequencies* across *samples*, not co-occurrence in cells.

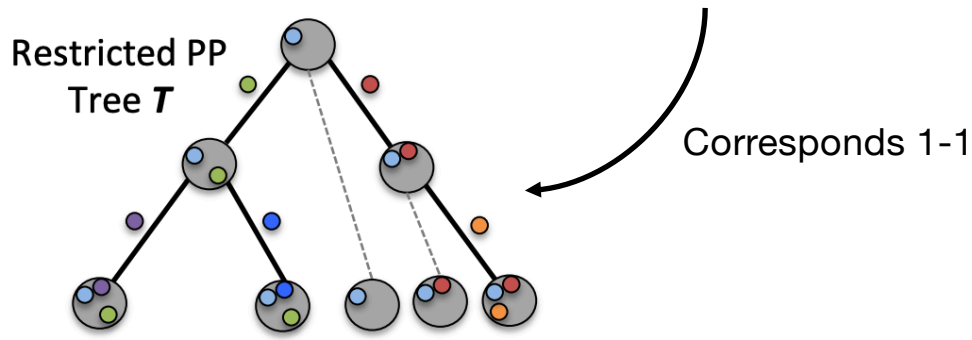


Perfect Phylogeny (PP) Mixture Problem

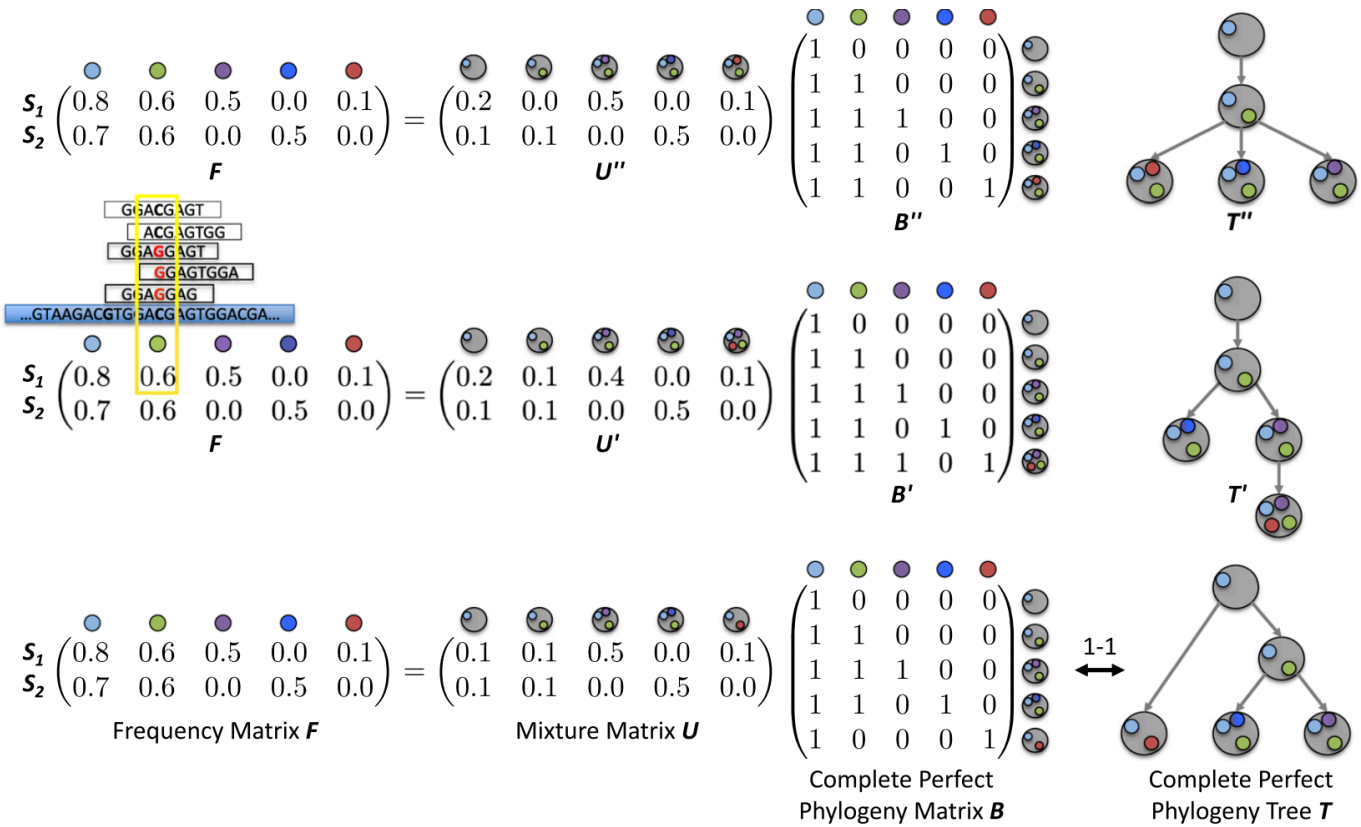


PPM Variants

TrAp [Strino et al., 2013]
PhyloSub [Jiao et al., 2014]
AncesTree [El-Kebir et al., 2015]
CITUP [Malikic et al., 2015]
BitPhylogeny [Yuan et al., 2015]
LICHeE [Popic et al., 2015]
…

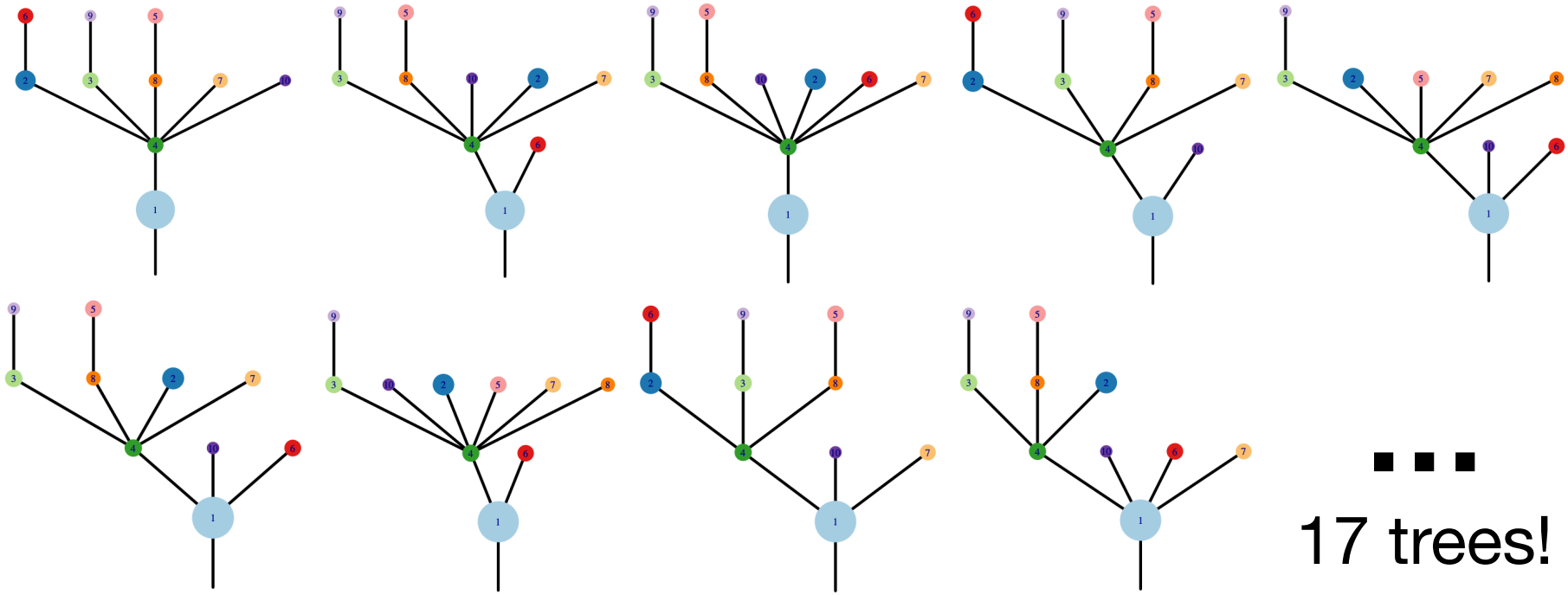


Challenge: Many Optimal Solutions



From [Qi et al., 2019]

Lung Cancer Patient: CRUK0037



...

17 trees!

From [Jamal-Hanjani et al., *NEJM* 2017]

Current Approaches for Reducing Optimal Solution Space

Long-read sequencing (e.g., [Deshwar et al., 2015])

Able to obtain reads with millions of basepairs.

Mutations on the same read originate from a single cell.

Single-cell sequencing (e.g., [Jahn et al., 2015; Zafar et al., 2017; El-Kebir 2018; Malikic et al., 2019])

Must account for sequencing errors.

Mutations comprising single cell form a connected path.

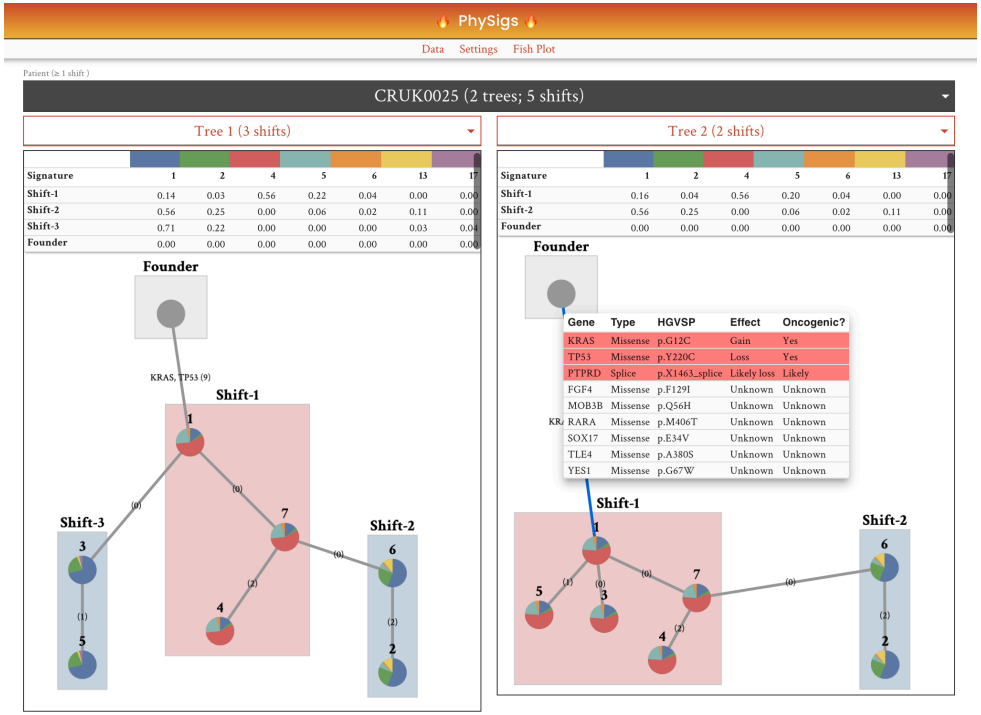
Our Approach

Chapter 3: We reduce the solution space using mutational signatures with PhySigs. **[Christensen et al., PSB 2020]**

Our Approach

Chapter 3: We reduce the solution space using mutational signatures with PhySigs.

[Christensen et al., PSB 2020]



Package installation

PhySigs is an R package that can be conveniently installed from GitHub.

```
install.packages("devtools")
devtools::install_github("elkebir-group/PhySigs_R")
```

Our Approach

Chapter 3: We reduce the solution space using mutational signatures with PhySigs. **[Christensen et al., PSB 2020]**

Chapter 4: We reduce the solution space using other patients' bulk data with RECAP. **[Christensen et al., ECCB 2020]**

Chapter 4: RECAP

Christensen S., Kim J., Koyejo S., Chia N. & El-Kebir M. (2020). Detecting Evolutionary Patterns of Cancers using Consensus Trees**.** [Presented at ECCB 2020].

Common Patterns in Patient Cohorts

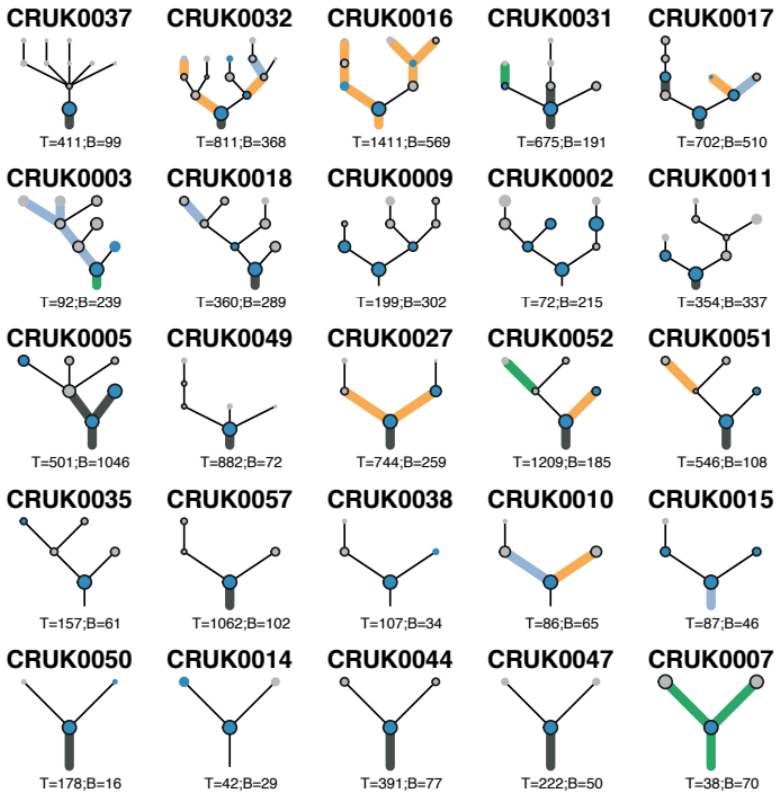


Image from [Jamal-Hanjani et al., 2017]

Prior Work using Other Patient Data

REVOLVER [Caravagna et al., *Nat. Methods* 2018]

Hintra [Khakabimamaghani et al., *Bioinformatics/ISMB* 2019]

- Current methods do not account for patient subtypes, a phenomenon that has been documented in other contexts.
- Current methods do not scale to large patient trees.
- Current methods have trouble dealing with varying mutation sets as well as mutation clusters.

RECAP Idea

To resolve ambiguities in patient data,
we could leverage common patterns of evolution
found in subtypes of patients.

Multiple Choice Consensus Tree (MCCT) Problem

Inputs

A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Multiple Choice Consensus Tree (MCCT) Problem

Inputs	Output

A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Select a tree $S_i \in \mathcal{T}_i$ for each patient i,

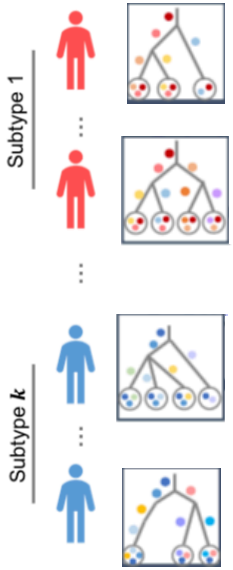
Multiple Choice Consensus Tree (MCCT) Problem

Inputs	Output

A set of possible trees for each patient

+

Parameter *k* for desired number of clusters



Select a tree $S_i \in \mathcal{T}_i$ for each patient i,

Assign each patient i to a cluster $\sigma(i) \in [k]$,

Multiple Choice Consensus Tree (MCCT) Problem

Inputs	Output

A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Select a tree $S_i \in \mathcal{T}_i$ for each patient i,

Assign each patient i to a cluster $\sigma(i) \in [k]$,

Construct consensus tree R_j for each cluster j,

Multiple Choice Consensus Tree (MCCT) Problem

Inputs	Output

A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

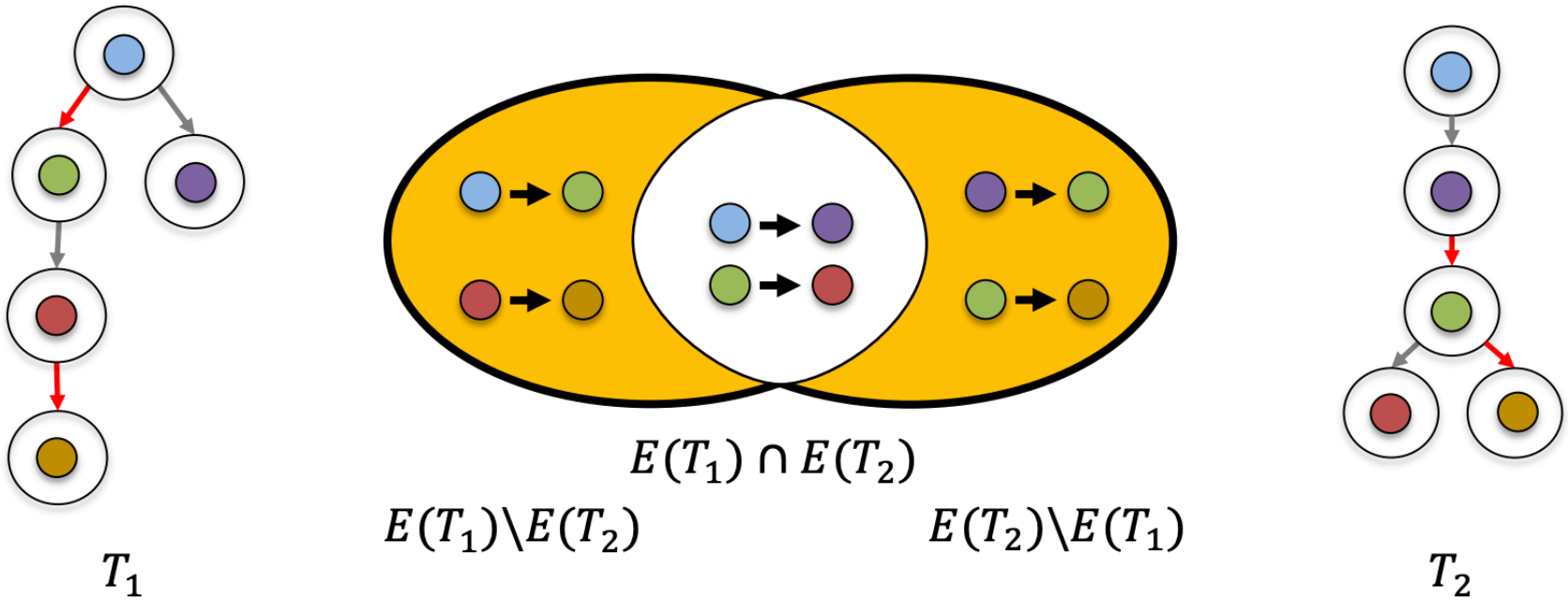
Select a tree $S_i \in \mathcal{T}_i$ for each patient i,

Assign each patient i to a cluster $\sigma(i) \in [k]$,

Construct consensus tree R_j for each cluster j,

Such that the sum of distances from each selected tree to the corresponding consensus tree is minimized.

Parent-Child (PC) Distance Function



Symmetric difference is 4.

RECAP: Main Contributions

Hardness: Proved MCCT NP-Hard via a reduction from 3-SAT and proposed gradient descent heuristic RECAP to use in practice.

Addresses prior limitations: RECAP allows for different patient subtypes and scales to larger sets of mutations.

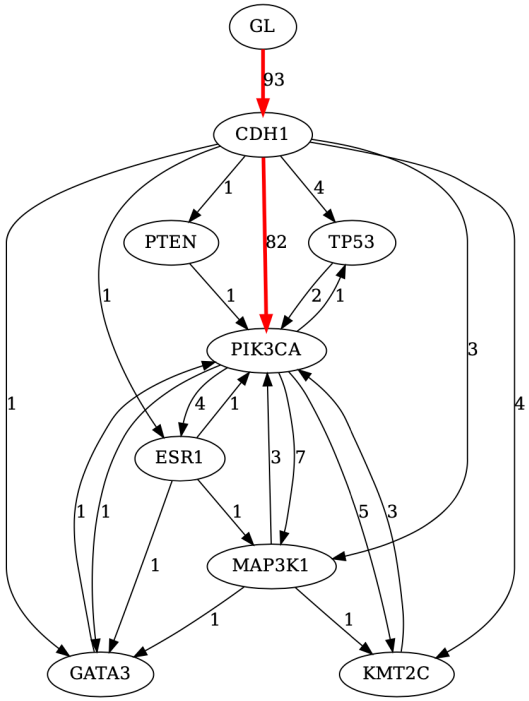
Simulation performance: Encouraging results on simulated data where there are different underlying subtypes.

Real Data performance: Uncover well-supported evolutionary trajectories in non-small cell lung cancer and breast cancer cohorts.

RECAP recovers known cancer subtype based on evolutionary trajectories

- Khakabimamaghani et al. (2019) previously used HINTRA to analyze breast cancer dataset
  - Manually split patients into four subtypes based on receptor status
  - In the HR+/HER2- subtype, found CDH1 commonly precedes PIK3CA.

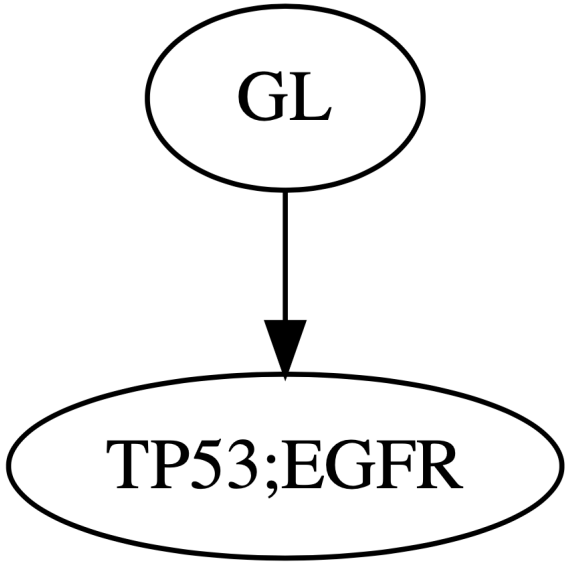
- RECAP finds subtype de novo in Cluster 7.
  - Consensus tree has CDH1 as parent of PIK3CA
  - 87 out of 93 patients (93.5%) in Cluster 7 belong to the HR+/HER2- subtype.



RECAP Cluster 7 Consensus Graph

Handling Mutation Clusters

- Mutations with similar frequencies in each sample from same tumor are typically clustered
- Lack of signal to resolve ordering represents a kind of ambiguity
- Resolving the ordering of driver mutations important for understanding common evolutionary trajectories



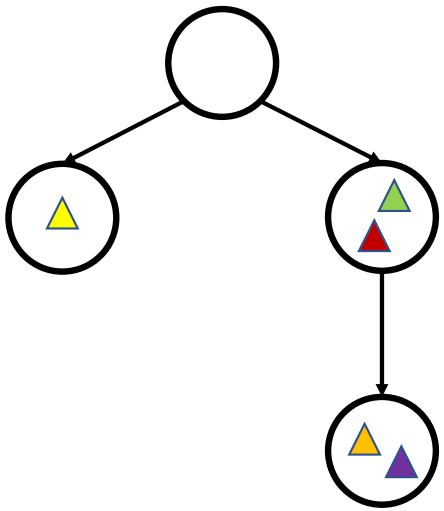
Lung cancer patient CRUK0004

Tree from [Jamal-Hanjani et al., 2017]

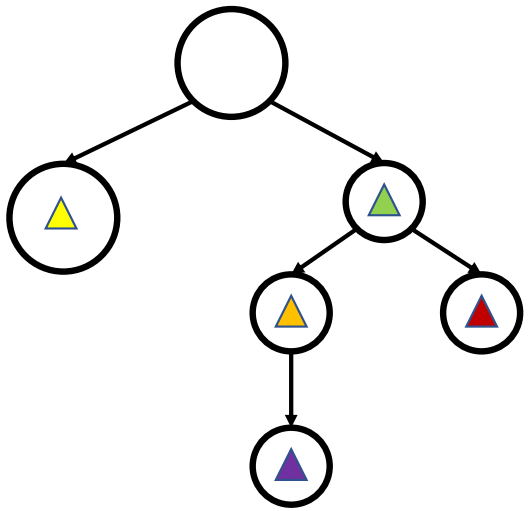
PC Optimal Cluster Expansion Problem

Inputs

Tree *T* with mutation clusters



Reference tree *R*



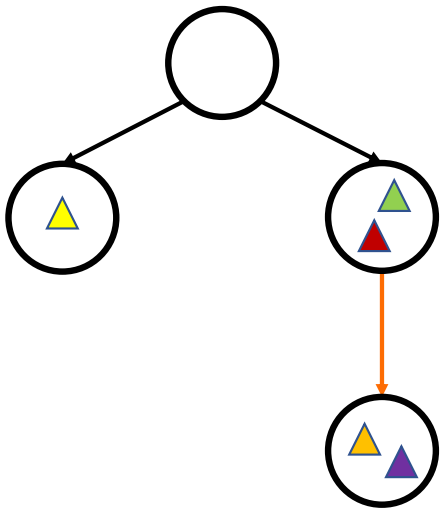
Expand clusters in tree *T* minimizing PC distance to *R*.

PC Optimal Cluster Expansion Problem

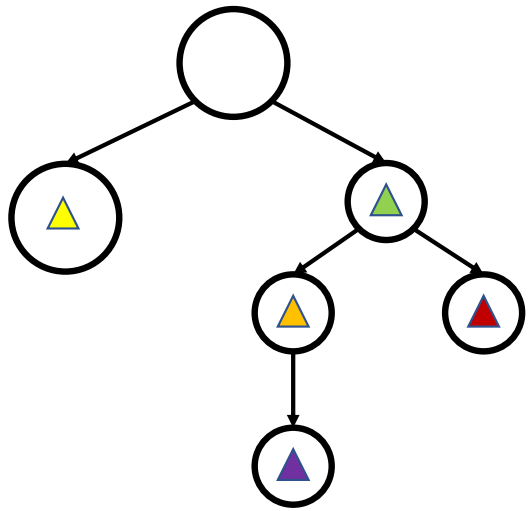
Inputs

Output

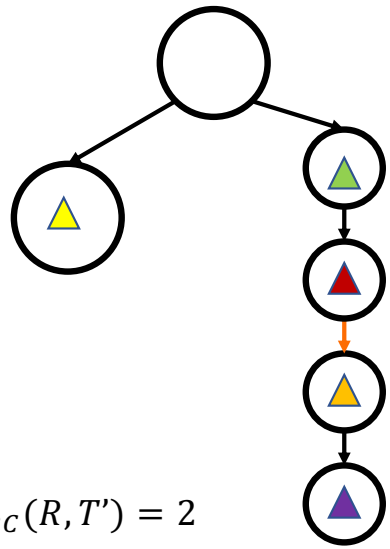
Tree *T* with mutation clusters



Reference tree *R*



T′* expanding *T



$D_{PC}(R, T') = 2$

PC Optimal Cluster Expansion Problem

Inputs

Output

Tree *T* with mutation clusters

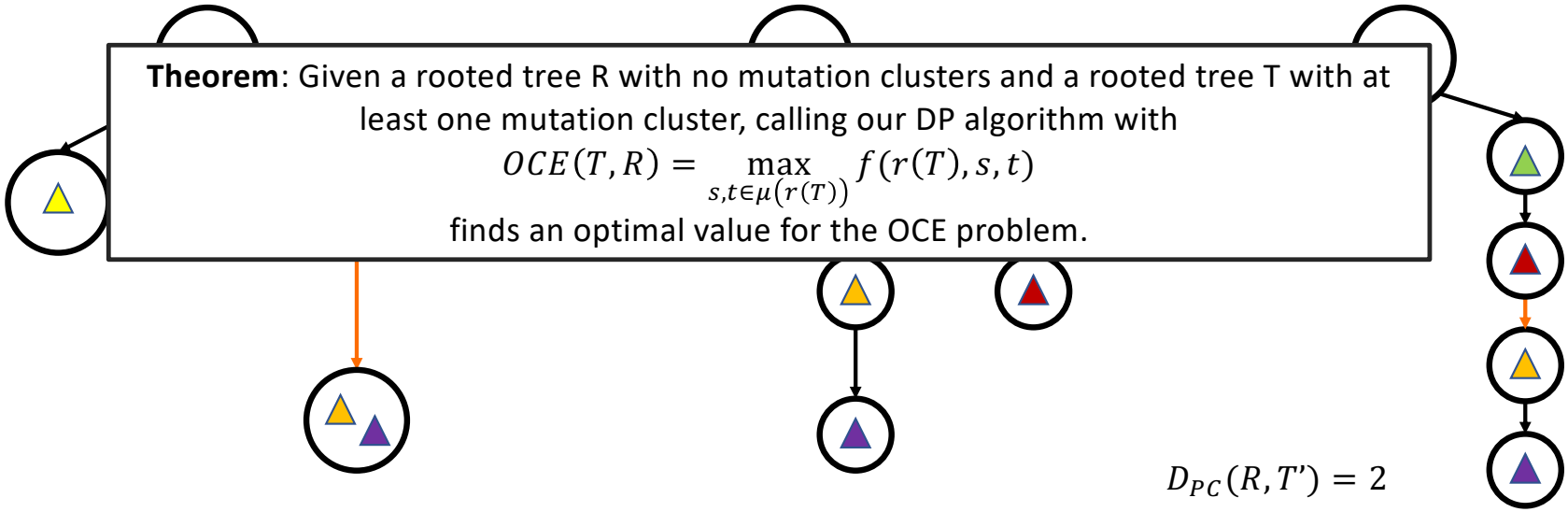
Reference tree *R*

T'* expanding *T

Theorem: Given a rooted tree R with no mutation clusters and a rooted tree T with at least one mutation cluster, calling our DP algorithm with

$$OCE(T,R) = \max_{s,t \in \mu(r(T))} f(r(T),s,t)$$

finds an optimal value for the OCE problem.



$D_{PC}(R,T') = 2$

RECAP: Future Directions

Try other distance measures: We use parent-child distance but other measures, such as ancestor-descendent, can be explored.

Move beyond infinite sites assumption: We can explore measures of similarity that do not assume a mutation is gained once and not lost.

Consider other consensus graphs: This is useful for incorporating mutual exclusivity of driver mutations that occur in the same pathway.

Incorporate into visualization: We could also support visualizing common evolutionary trajectories in our tool.

Conclusions

Returning to Dissertation Idea

Develop biologically *meaningful optimization* problems

RF-OTC, **RF-OTRC**, TCE, and **MCCT**

Returning to Dissertation Idea

Develop biologically *meaningful optimization* problems

RF-OTC, **RF-OTRC**, TCE, and **MCCT**

with corresponding *efficient algorithms*

OCTAL, **TRACTION**, PhySigs, and **RECAP**

Returning to Dissertation Idea

Develop biologically *meaningful optimization* problems

RF-OTC, **RF-OTRC**, TCE, and **MCCT**

with corresponding *efficient algorithms*

OCTAL, **TRACTION**, PhySigs, and **RECAP**

that leverage *auxiliary data* to address challenges

Species tree, mutational signatures, and **other patients**

in species and tumor *phylogeny estimation*.

Looking Forward

- Introduced four methods for improving phylogeny estimation
  - Posed a biologically meaningful optimization problem
  - Established computational complexity and conceived of approach
  - Implemented and benchmarked empirical performance

- Several directions for future research
  - Expanding to more realistic models of evolution
  - Explore use of other graph theoretic objects
  - Assess implications for downstream analysis

Thank you to my sources of funding.

Chirag Fellowship

State Farm Doctoral Award

C.L. and Jane Liu Award

Ira & Debra Cohen Graduate Fellowship

National Science Foundation (Grant No. CCF-1535977 & IIS 15-13629 to Tandy Warnow).

Thank you to my many mentors and collaborators.