Detecting Evolutionary Patterns of Cancers using Consensus Trees

Sarah Christensen[1], Juho Kim[2], Nicholas Chia[3,4], Oluwasanmi Koyejo[1], and Mohammed El-Kebir[1]

[1] Dept. of CS, University of Illinois at Urbana-Champaign

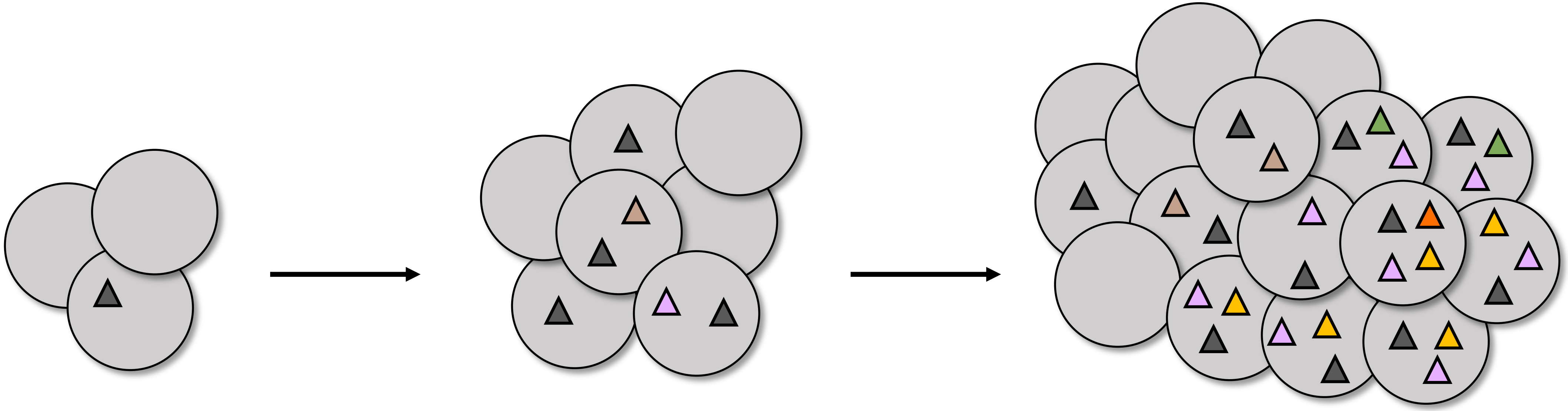
[2] Dept. of ECE, University of Illinois at Urbana-Champaign

[3] Microbiome Program, Center for Individualized Medicine, Mayo Clinic

[4] Division of Surgical Research, Department of Surgery, Mayo Clinic

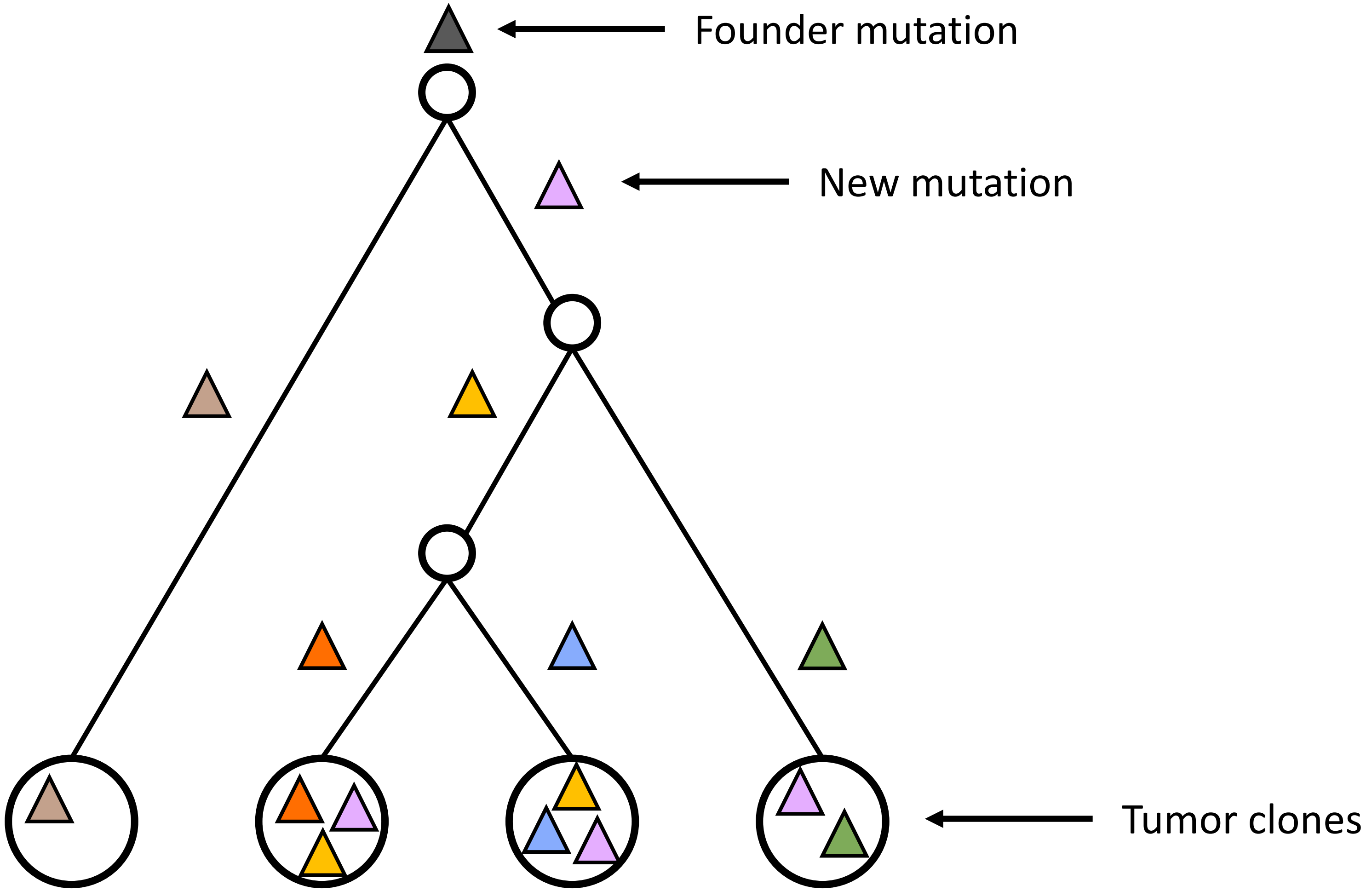
ECCB 2020

Evolution in Cancer



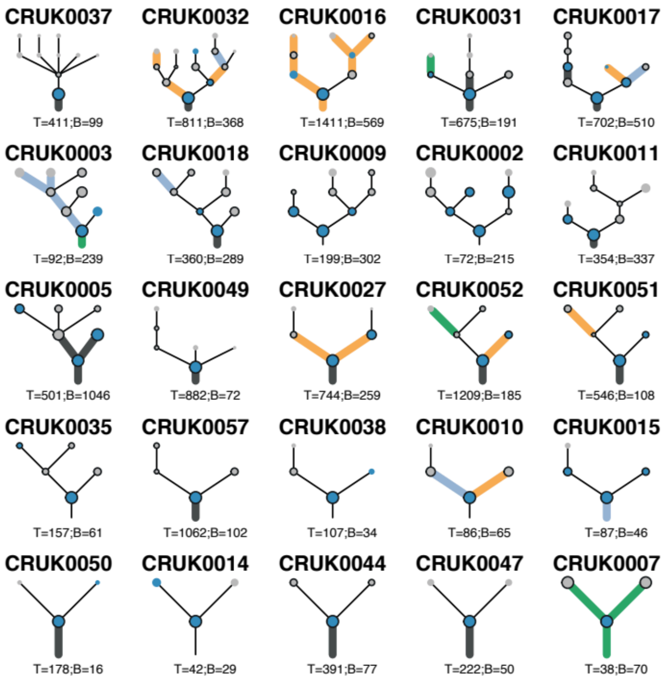
Clonal Evolution Theory of Cancer
[Nowell, 1976]

Phylogenetic Trees in Cancer



Phylogenies have potential to improve stratification of cancer patients into subtypes

Goal: Find repeated patterns defined by ordering of recurrent driver mutations

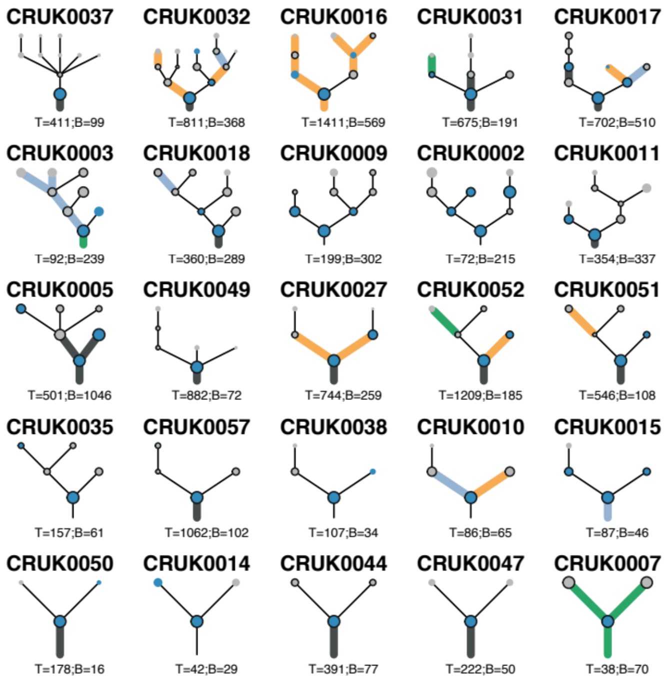


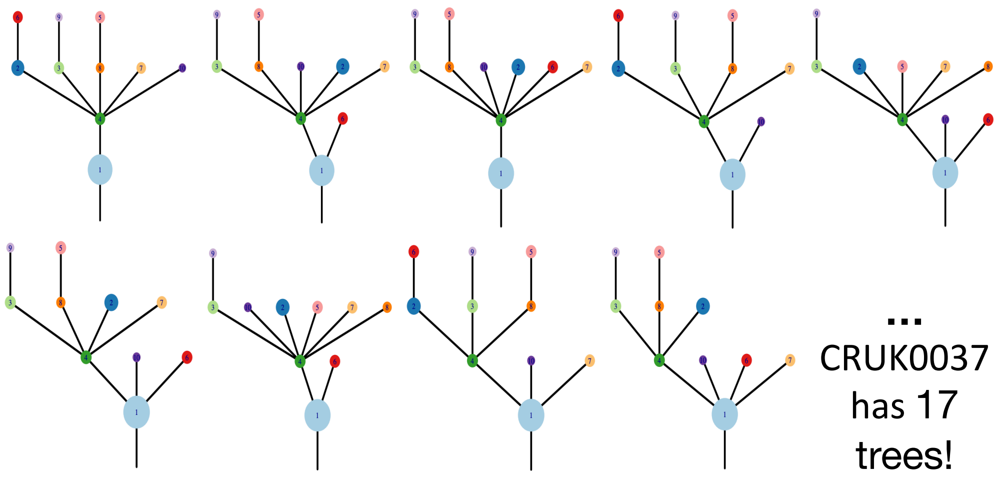
Images from [Jamal-Hanjani et al., *NEJM* 2017]

Phylogenies have potential to improve stratification of cancer patients into subtypes

Goal: Find repeated patterns defined by ordering of recurrent driver mutations

Challenge: Obfuscated by alternative phylogenies at the individual patient level





...
CRUK0037
has 17
trees!

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

Prior work on inferring phylogenies and finding evolutionary patterns using patient cohorts

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

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

- Current methods do not account for cancer subtypes.

- Current methods do not scale to large patient trees.

- Current methods have trouble dealing with varying mutation sets as well as mutation clusters.

Our approach

We pose an optimization problem MCCT (*Multiple Choice Consensus Tree*) and algorithm RECAP (*Revealing Evolutionary Consensus Across Patients*).

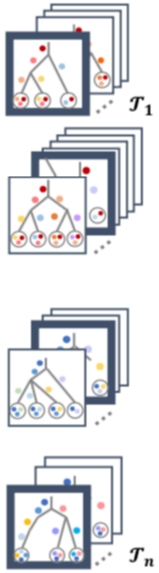
Our approach leverages common patterns of evolution

found in subtypes of patients

to resolve ambiguities in patient data.

Multiple Choice Consensus Tree (MCCT) Problem

Inputs



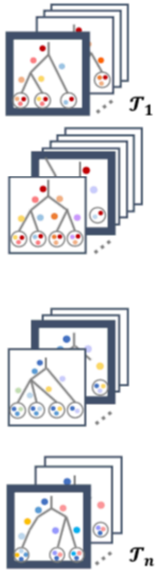
A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Multiple Choice Consensus Tree (MCCT) Problem

Inputs

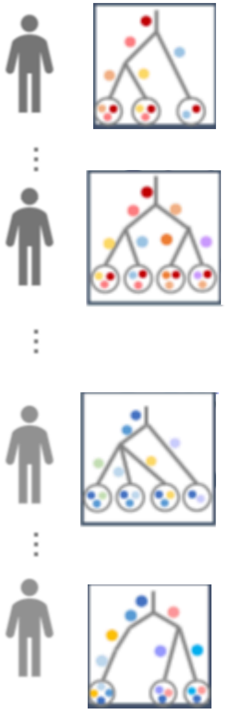


A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

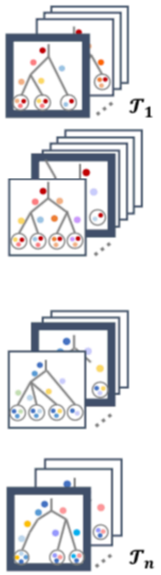
Output



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

Multiple Choice Consensus Tree (MCCT) Problem

Inputs

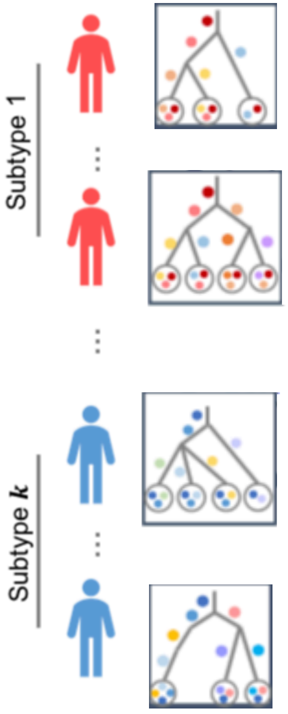


A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Output

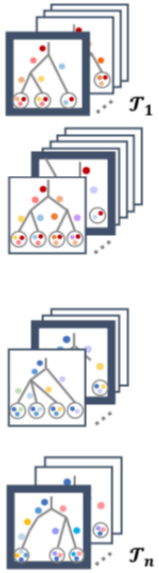


Select a tree $S_i \in \mathcal{T}_i$ for each patient i,

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

Multiple Choice Consensus Tree (MCCT) Problem

Inputs

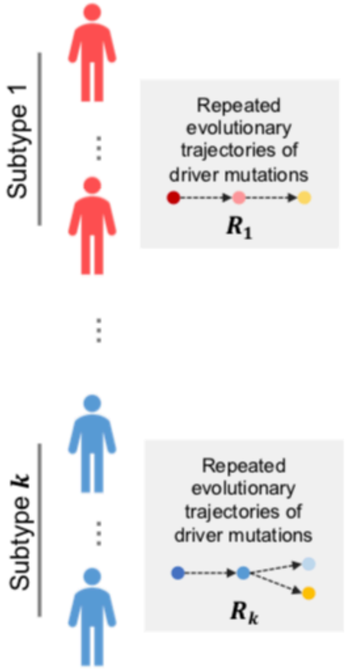


A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Output



Subtype 1
Repeated evolutionary trajectories of driver mutations
R_1
Subtype k
Repeated evolutionary trajectories of driver mutations
R_k

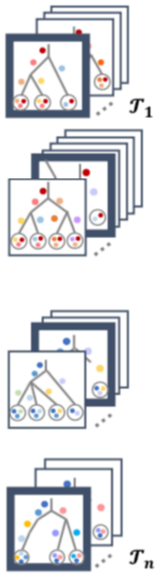
Select a tree $S_i \in \mathcal{T}_i$ for each patient i,

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

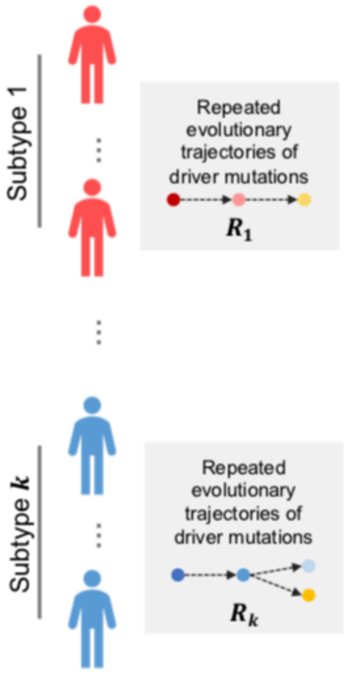


A set of possible trees for each patient

+

Parameter *k* for desired number of clusters

Output


Subtype 1
Repeated evolutionary trajectories of driver mutations
R_1
Subtype k
Repeated evolutionary trajectories of driver mutations
R_k

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

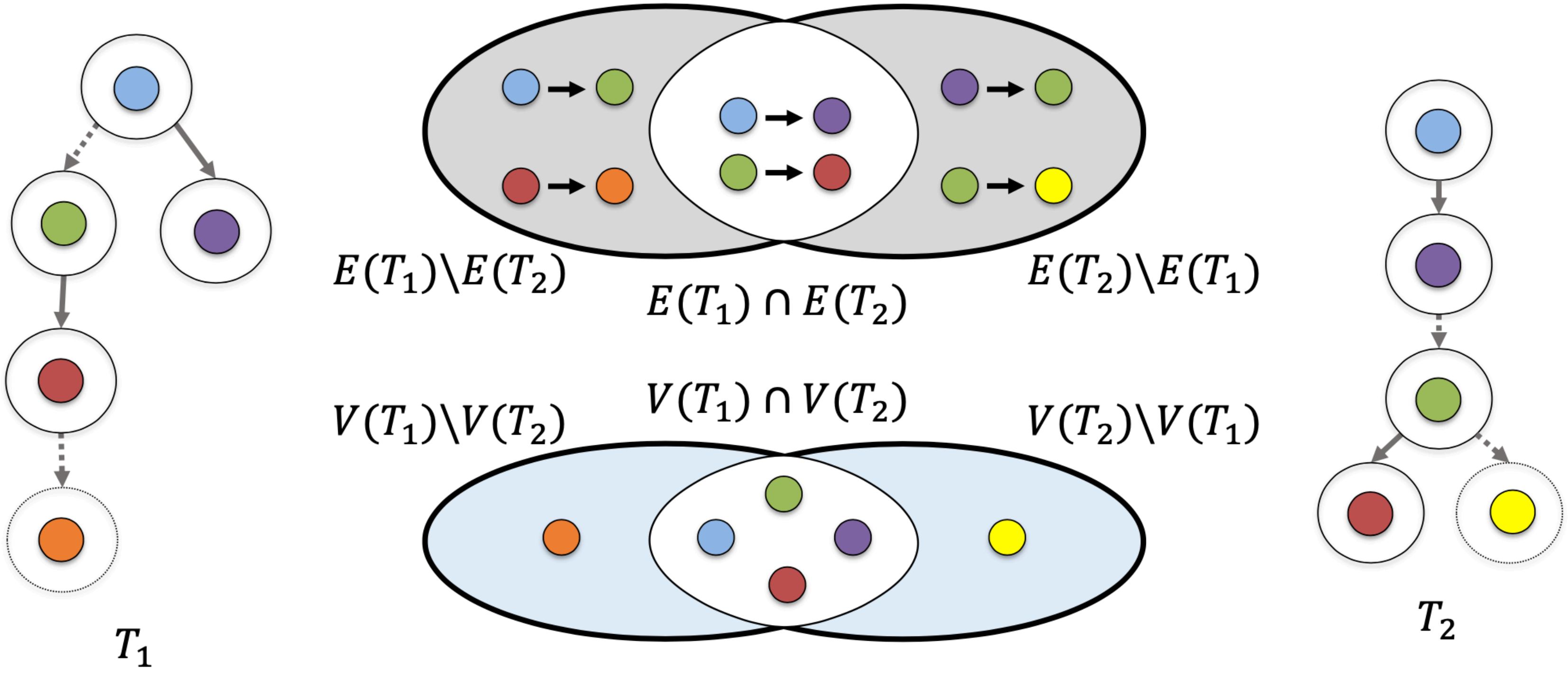
Assign each patient i to a cluster $\sigma(i) \in [k]$,

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.

Repeated evolutionary trajectories of driver mutations
R_1

Distance function accounts for varying mutation sets and tree sizes



$$d_N(T_1,T_2) = \frac{|E(T_1)\,\Delta\,E(T_2)| + |V(T_1)\,\Delta\,V(T_2)|}{2|\Sigma|} = \frac{4+2}{12} = 0.5$$

RECAP: Summary of results

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

Addresses prior limitations: RECAP allows for different patient subtypes, different mutation sets, scales to larger sets of mutations, and includes a DP subroutine to handle mutation clusters.

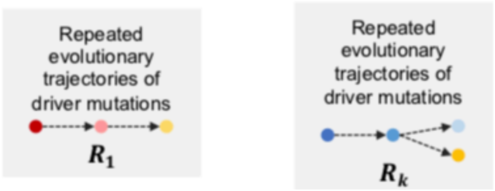
Empirical performance: RECAP outperforms existing methods on *simulated* data where there are different underlying subtypes and resolves ambiguity for patient phylogenies on *biological* data.

Simulation procedure allows patient subtypes

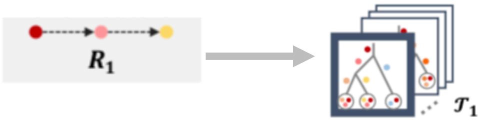
Randomly draw patient clustering

Subtype 1
Subtype *k*
...
...
...

Construct cluster consensus tree by using Prüfer sequence on random subset of mutations


Repeated evolutionary trajectories of driver mutations
R_1
Repeated evolutionary trajectories of driver mutations
R_k

Generate patient trees by simulating bulk sequencing experiment seeded by consensus tree



600 different simulated instances parameterized by four variables:

# of mutations across cohort: 5 or 12

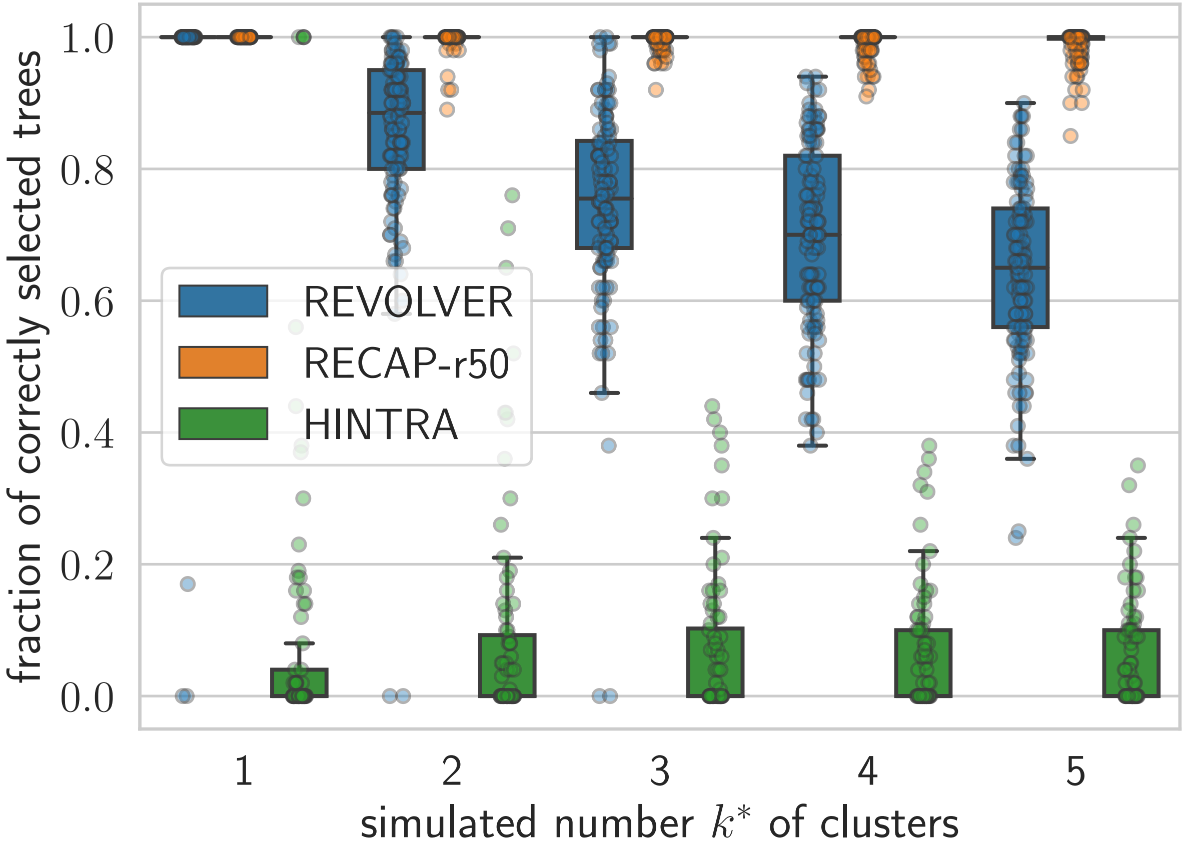
# of mutations in patient trees: 5, 7, or 12

# of clusters in ground truth: 1 to 5

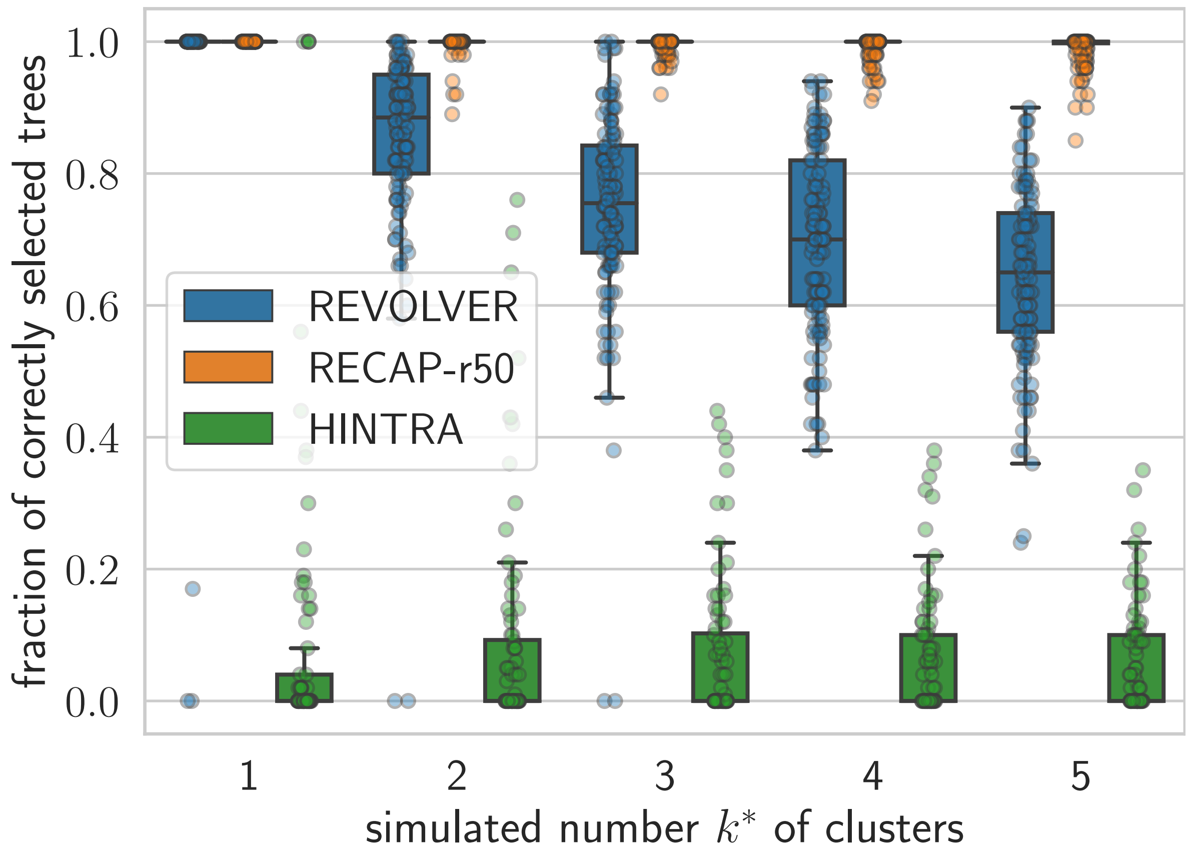
# of patients in cohort: 50 or 100

RECAP improves performance, especially with many patient subtypes, on simulated data

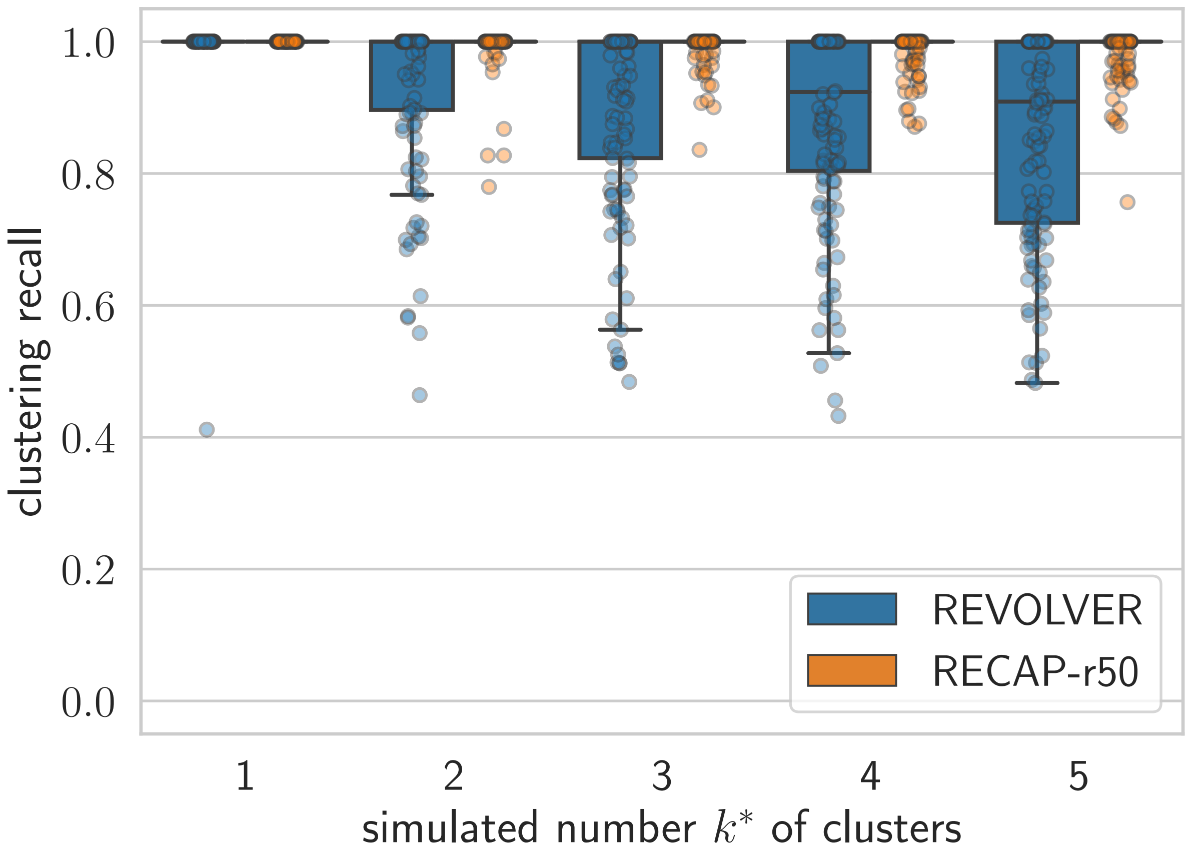
Is the correct patient tree selected?



RECAP improves performance, especially with many patient subtypes, on simulated data



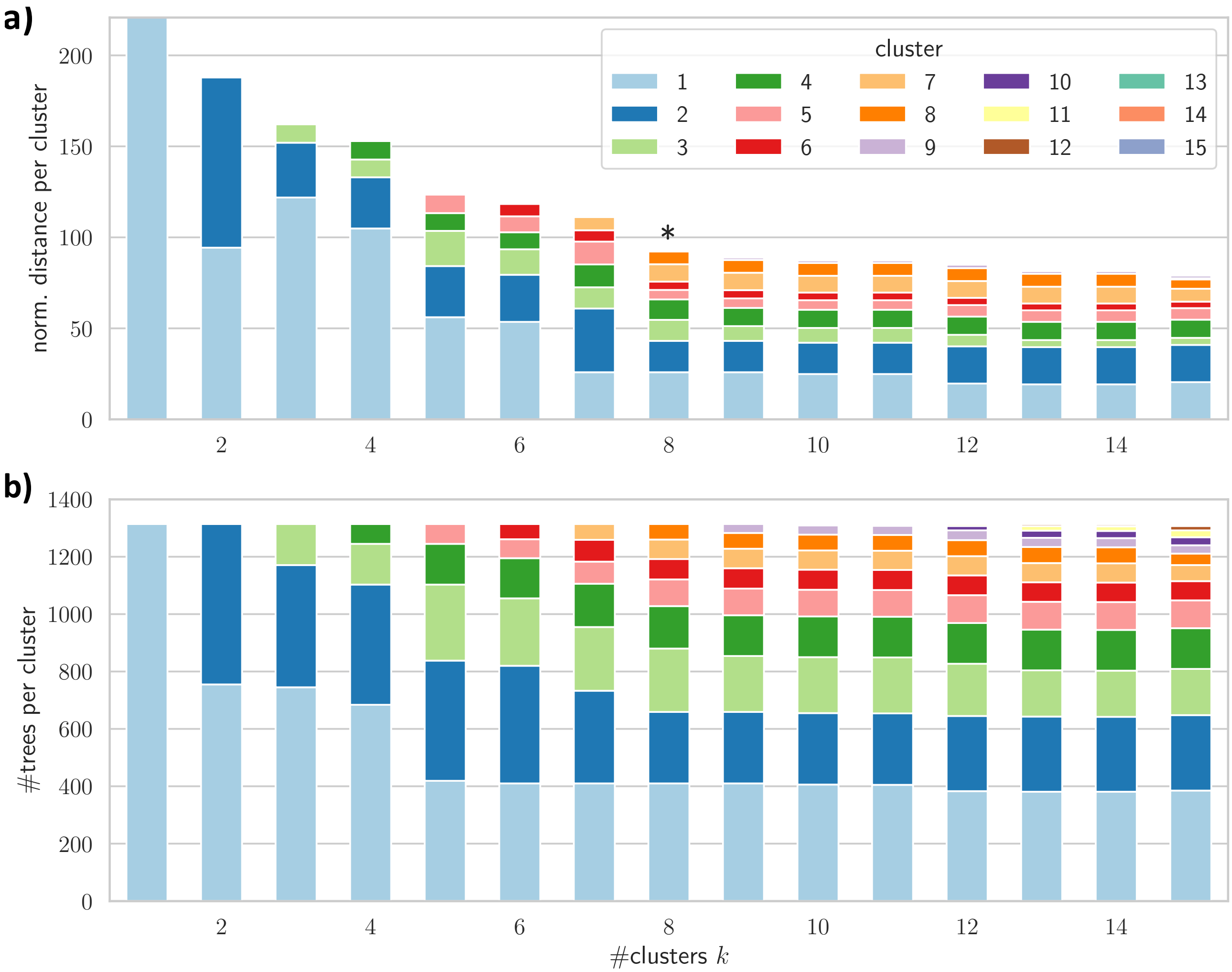
Is the correct patient tree selected?



Is the patient clustering correct?

RECAP finds clusters in breast cancer cohort

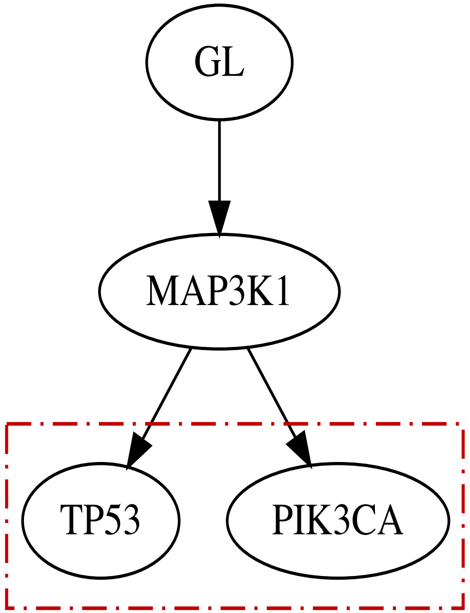
- 1,315 patients with SNVs in copy neutral regions
- 1 to 6,332 trees per patient calculated using SPRUCE
- Restricted to 8 mutations, occurring in >100 patients
- Identified 8 clusters with 55 to 400 patients in each



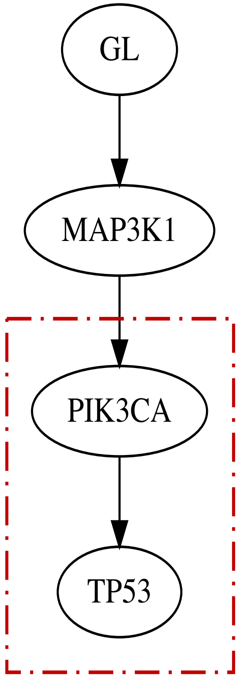
Raw data from [Razavi et al., 2018]

RECAP resolves ambiguity for patient P-0004859

P-0004859 Input Trees

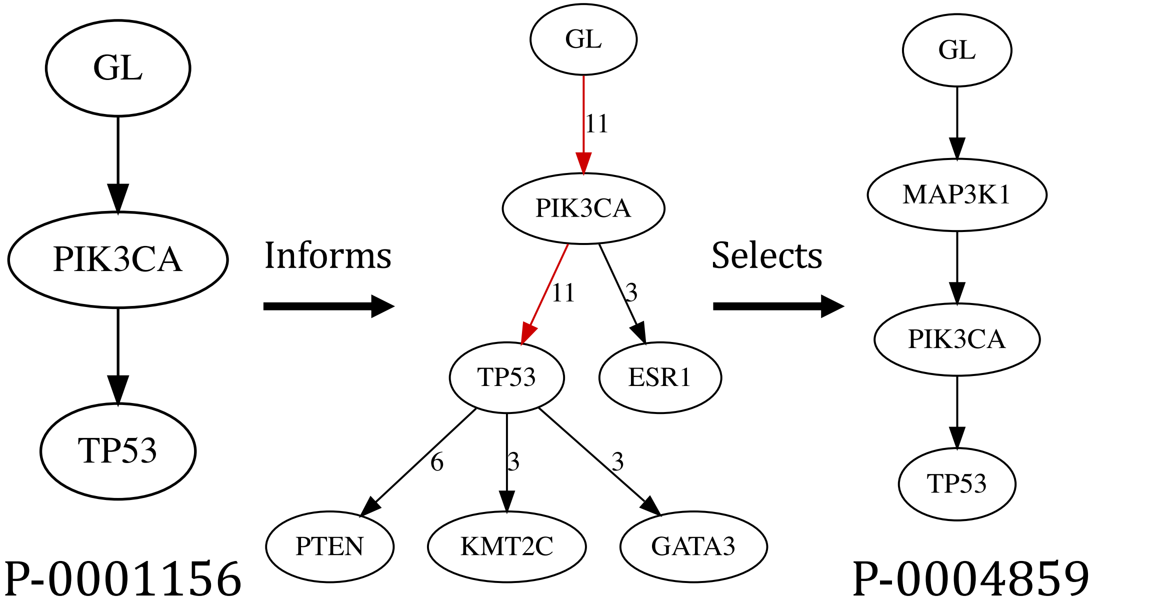


Tree 1



Tree 2

RECAP Outputs Selection



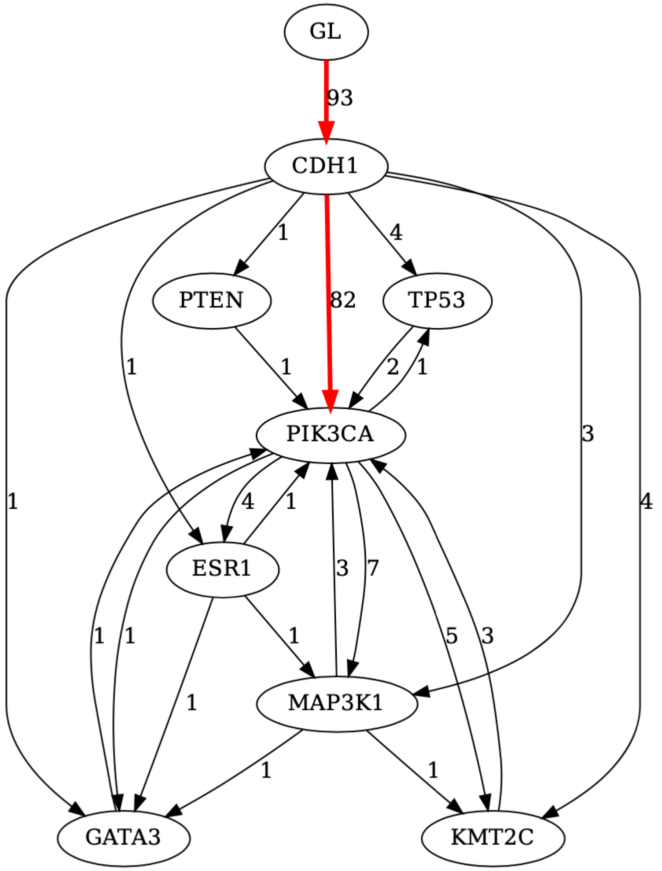
P-0001156
Input Tree

Consensus Graph 1

P-0004859
Output Tree 2

RECAP recovers known cancer subtype based on evolutionary trajectories

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

Conclusion and discussion

RECAP leverages common patterns of evolution to simultaneously resolve ambiguities in sequencing data and identify cancer subtypes.

RECAP expands on previous work by testing for different subtypes and running on varying mutation sets along with mutation clusters.

MCCT is an adaptable framework leading to avenues of future work (e.g., changing distance metric, consensus graph, evolutionary model).

Availability: https://github.com/elkebir-group/RECAP

Acknowledgements

Co-authors (pictured above)

Layla Oesper for helpful discussions

El-Kebir lab for thoughtful feedback

NSF (CCF 18-50502 to MEK)