



Mutational Signatures

February 23, 2021

CS 598 MEB

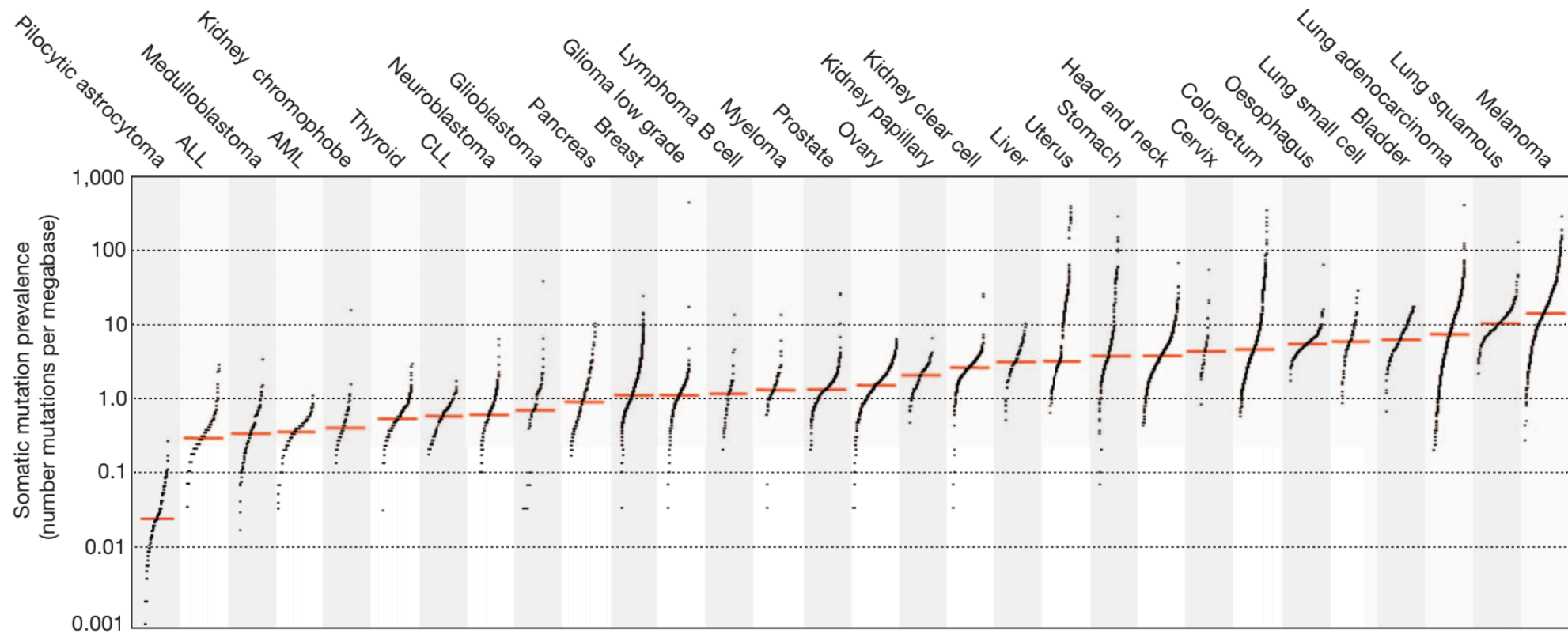
Today's Outline

1. **What** are mutational signatures?
2. **How** are mutational signatures estimated?
3. **Why** are mutational signatures useful?

Today's Outline

1. **What** are mutational signatures?
2. **How** are mutational signatures estimated?
3. **Why** are mutational signatures useful?

Prevalence of mutations in cancer

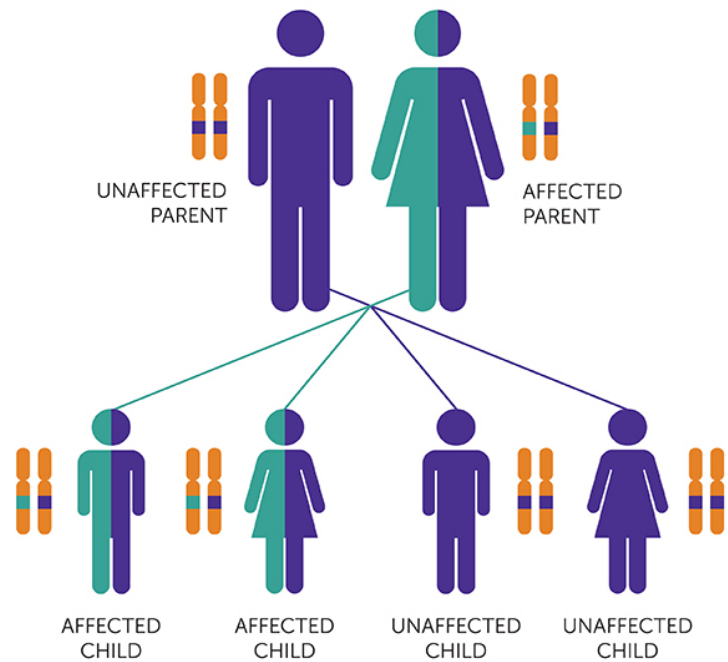


On average, 1,000-10,000 mutations/genome and 10-100 mutations/exome

Where do mutations come from?

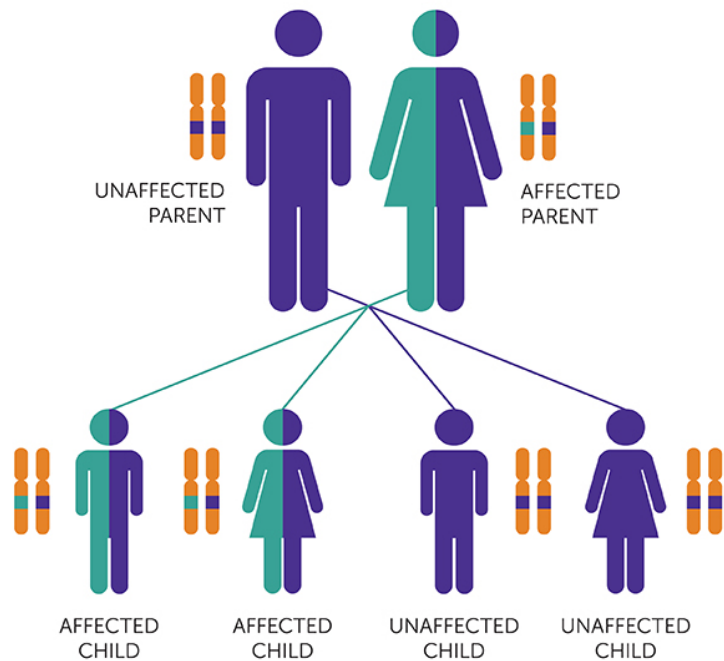
Germline

Somatic



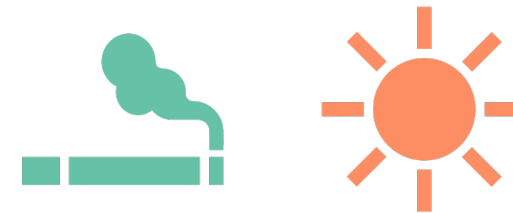
Where do mutations come from?

Germline



Somatic

Environmental Interference



Replication Errors



Somatic mutations accumulate over time

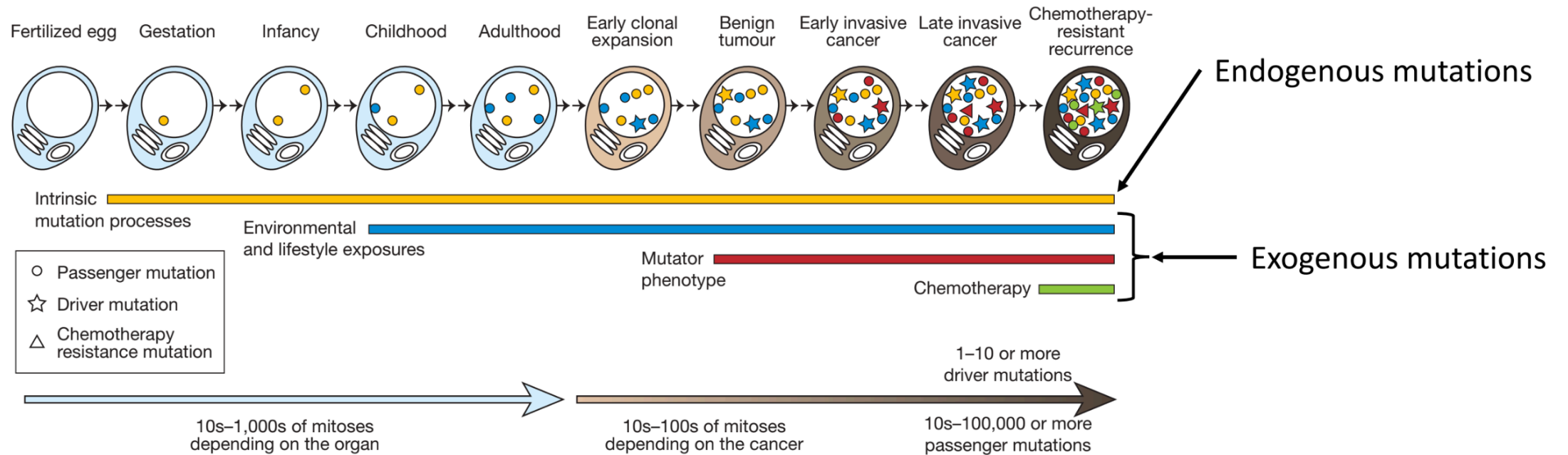
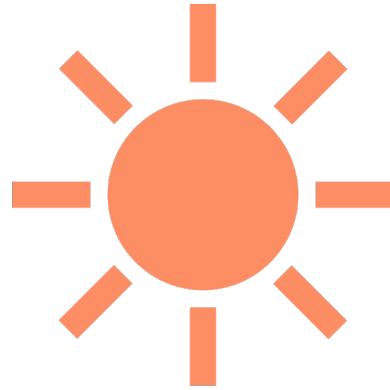


Figure 1: Stratton et al. (Nature 2009)

Somatic mutations accumulate over time

GERMLINE: cagaccagatagggcatgagatacatccgcgagtgattggattacatctggggagtgattgatctaaactcttctcaagg

CANCER: aagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg



Problem! No labels

GERMLINE: cagaccagatagggcatgagatacatccgcgagtgattggattacatctggggagtgattgatctaaactcttctcaagg

CANCER: aagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

Mutational Signatures

GOAL 1: Identify distinct mutational patterns associated with mutational processes (i.e., estimate *signatures*).

GOAL 2: Identify exposure to each pattern for each patient (i.e., estimate *exposures*).

Today's Outline

1. **What** are mutational signatures?

2. **How** are mutational signatures estimated?

3. **Why** are mutational signatures useful?

Idea! Look across patients

PATIENT 1: aagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 2: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 3: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

•
•
•

PATIENT n: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

Idea! Look across patients

PATIENT 1: aagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 2: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 3: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

•
•
•

PATIENT n: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

What patterns should we look at?

PATIENT 1: aagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

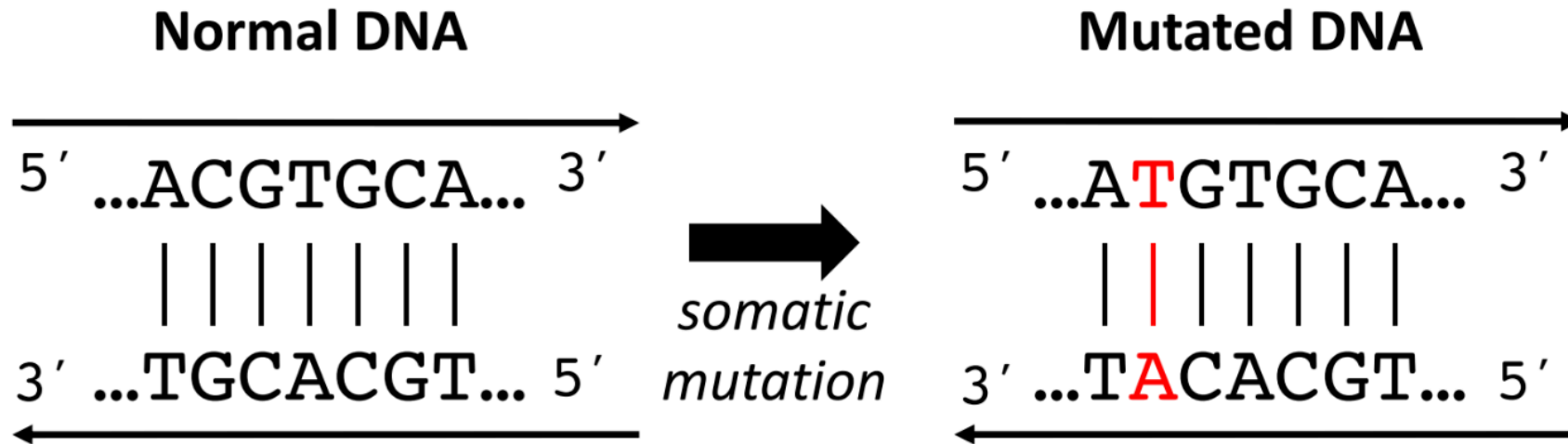
PATIENT 2: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 3: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

•
•
•

PATIENT n: agaagacctagttagggcatgagaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

Mutational Category: Single base substitutions



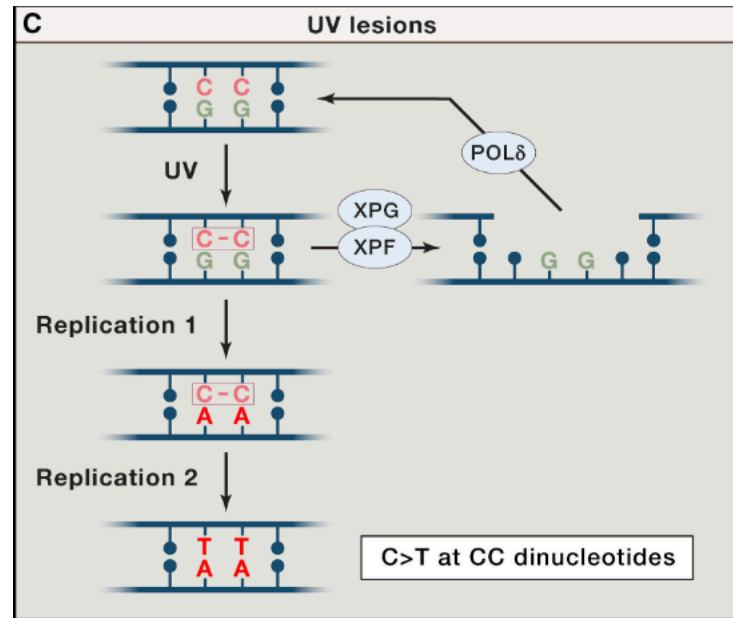
Did a C>T mutation occur? Or a G>A?

Mutational Category: Single base substitutions

	A	T	C	G
T	T > A		T > C	T > G
C	C > A	C > T		C > G

Six substitutions patterns (pyrimidine first)

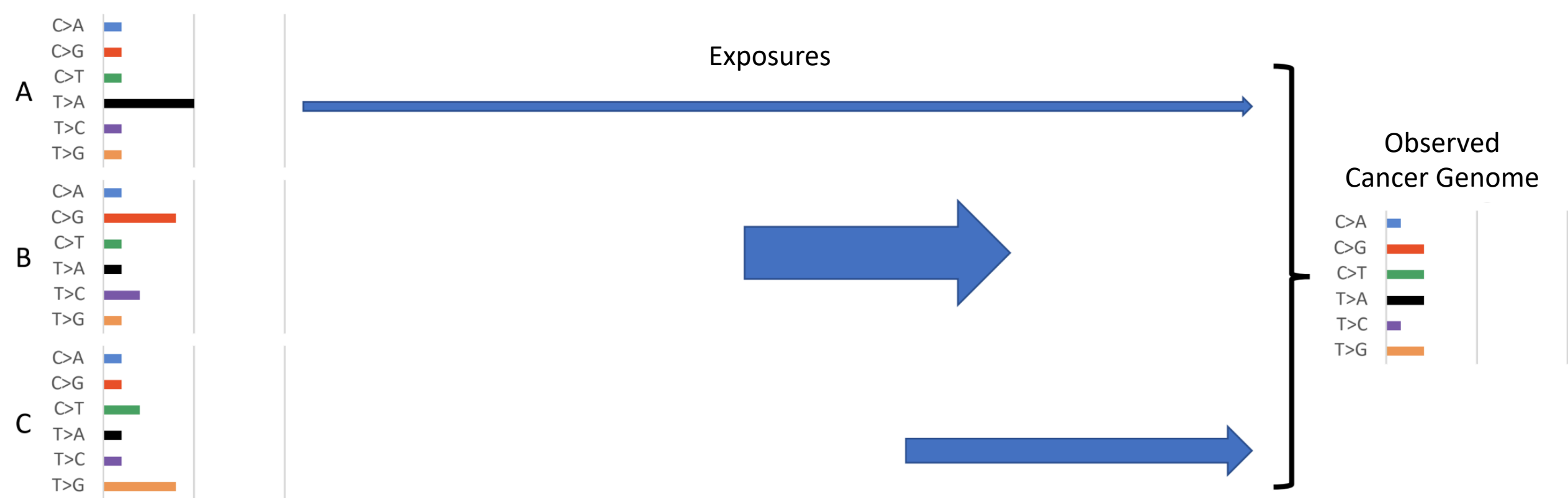
Mutational Category: Single base substitutions



Environmental exposures and repair errors can lead to consistent substitution patterns.

Integrating mutational categories into problem

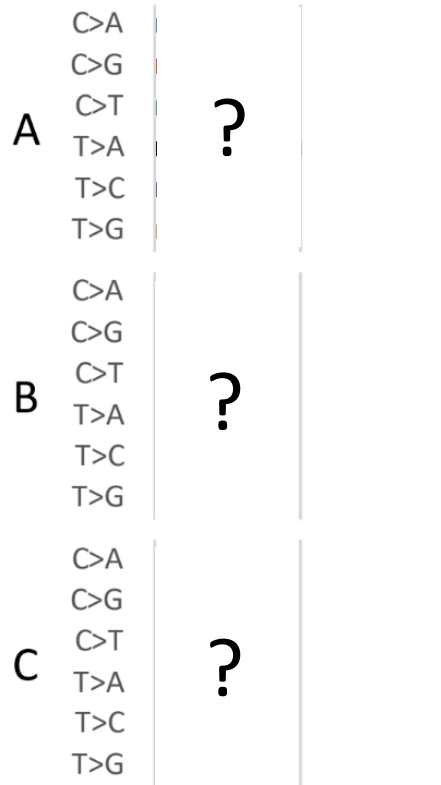
Mutational Processes



Adapted from Ludmil Alexandrov

Integrating mutational categories into problem

Mutational Processes



Exposures

?

?

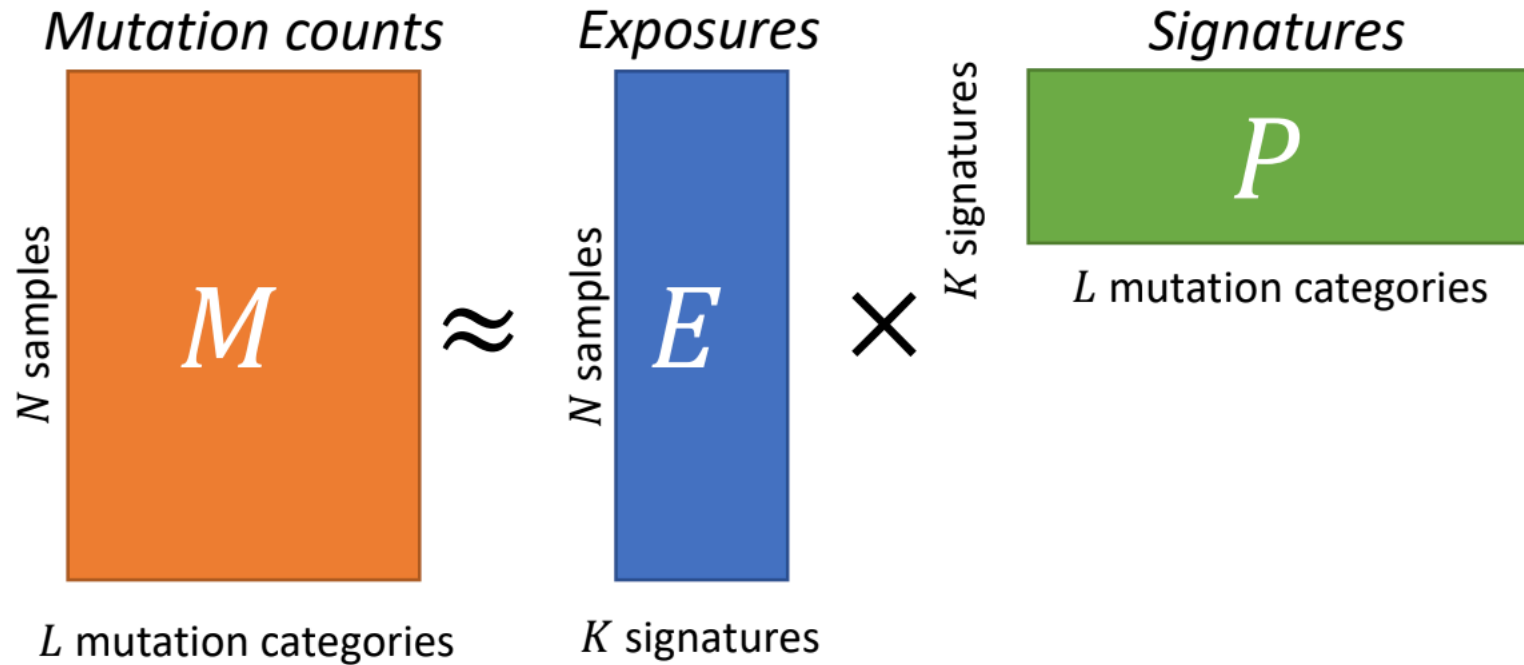
?

Observed
Cancer Genome



Adapted from Ludmil Alexandrov

Alexandrov Problem Statement



Objective function:







$$\min_{E, P} \|M - EP\|_2^2 \text{ such that } E, P \geq 0.$$





Alexandrov Algorithmic Technique

Non-negative Matrix Factorization (NMF)

- Unique solution not guaranteed
- Popular search heuristics use alternating optimization
- Used in many applications

	M1	M2	M3	M4	M5
 Comedy	3	1	1	3	1
 Action	1	2	4	1	3

	 Comedy	 Action
 A	✓	✗
 B	✗	✓
 C	✓	✗
 D	✓	✓

	M1	M2	M3	M4	M5
	3	1	1	3	1
	1	2	4	1	3
	3	1	1	3	1
	4	3	5	4	4

Alexandrov Model Selection

The number of mutational processes is not known.

Algorithm is run over a range of values.

Pick number of processes based on:

- Low reconstruction error for number of processes
- Process reproducibility with random initializations

Alexandrov Mutational Category: SBS with flanking

A C > T G

$$4 \times 6 \times 4 = 96$$

Alexandrov Data (v2 - March 2015)

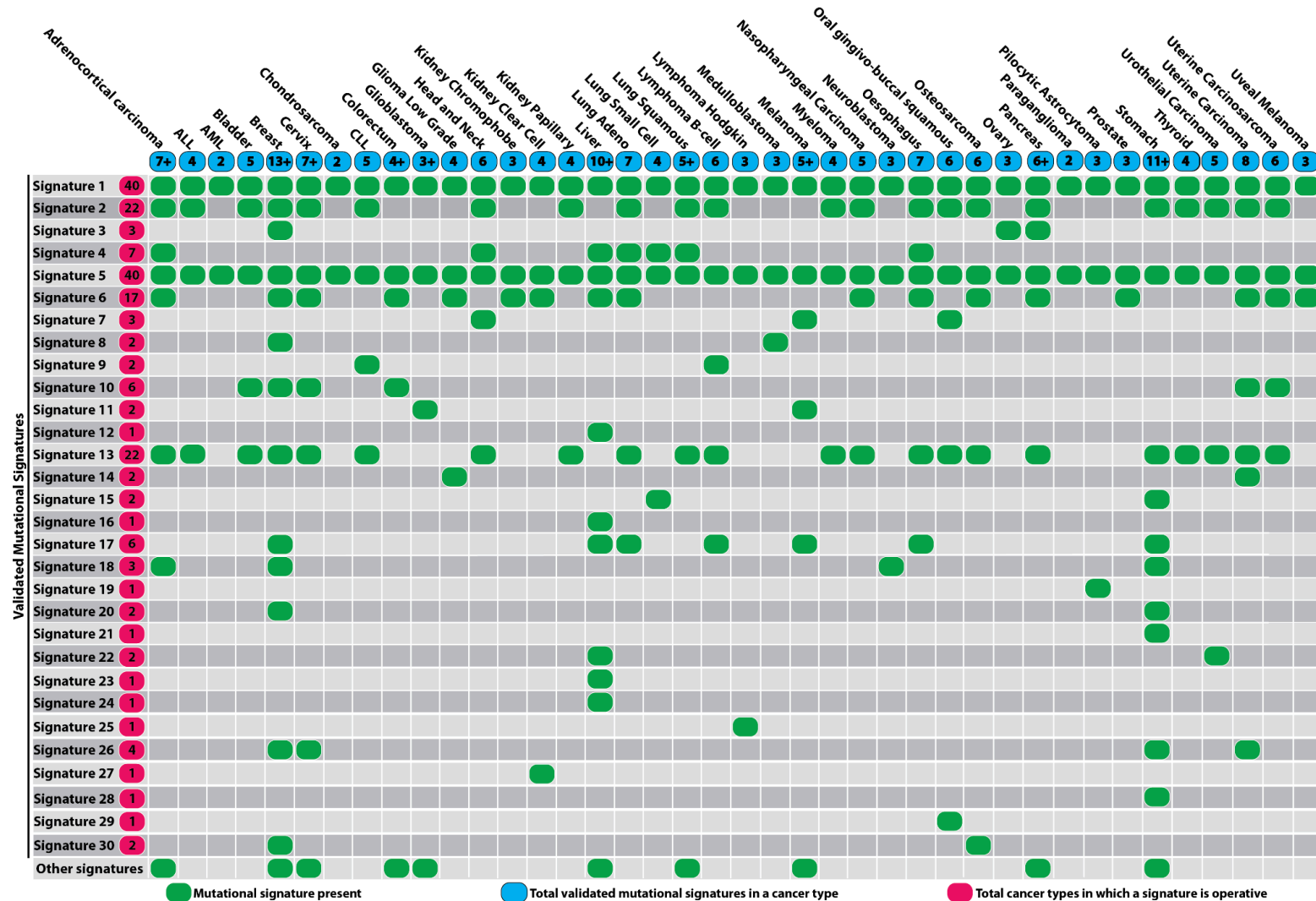
10,952 exomes and 1,048 whole-genomes

40 distinct types of human cancer

Estimated per cancer type and per class (WGS, WXS)

Maintained at: <https://cancer.sanger.ac.uk/cosmic/>

Alexandrov Results (v2 - March 2015)



Alexandrov Results (v2 - March 2015)



Future Work

Improve the estimation of signatures

Look at different types of mutational categories

Study NMF solutions of comparable reconstruction error

Other ideas?

Today's Outline

1. **What** are mutational signatures?
2. **How** are mutational signatures estimated?
3. **Why** are mutational signatures useful?

Application Directions

Prevention

- Establish correlations between environmental factor and mutations

Treatment

- Identify subtypes of patients based on signatures

Orthogonal signal for prioritizing solutions

PhySigs: Phylogenetic Inference of Mutational Signature Dynamics

Sarah Christensen¹, Mark D.M. Leiserson², and Mohammed El-Kebir¹

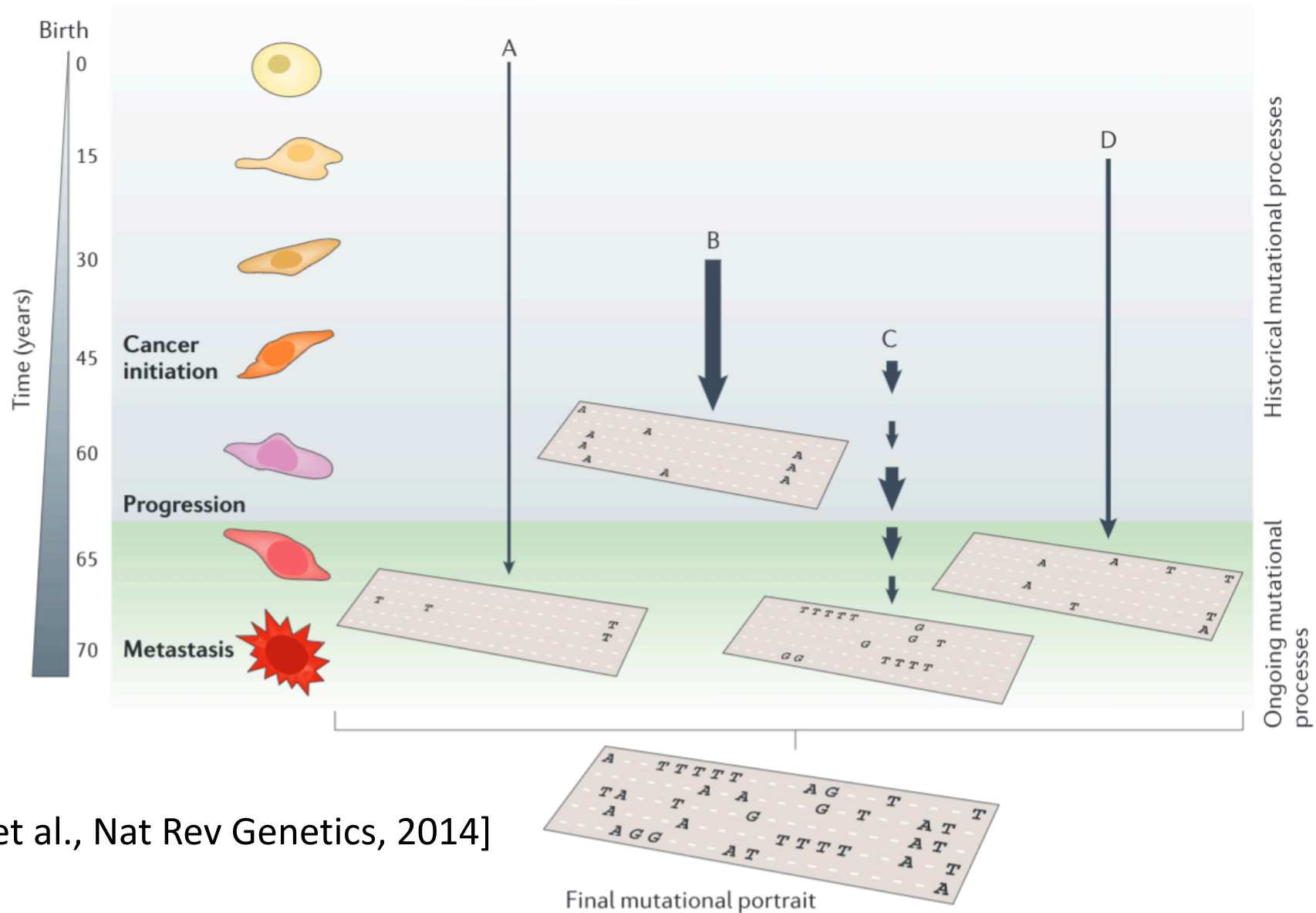
¹Dept. of CS, University of Illinois at Urbana-Champaign

²Dept. of CS, University of Maryland College Park

PSB 2020



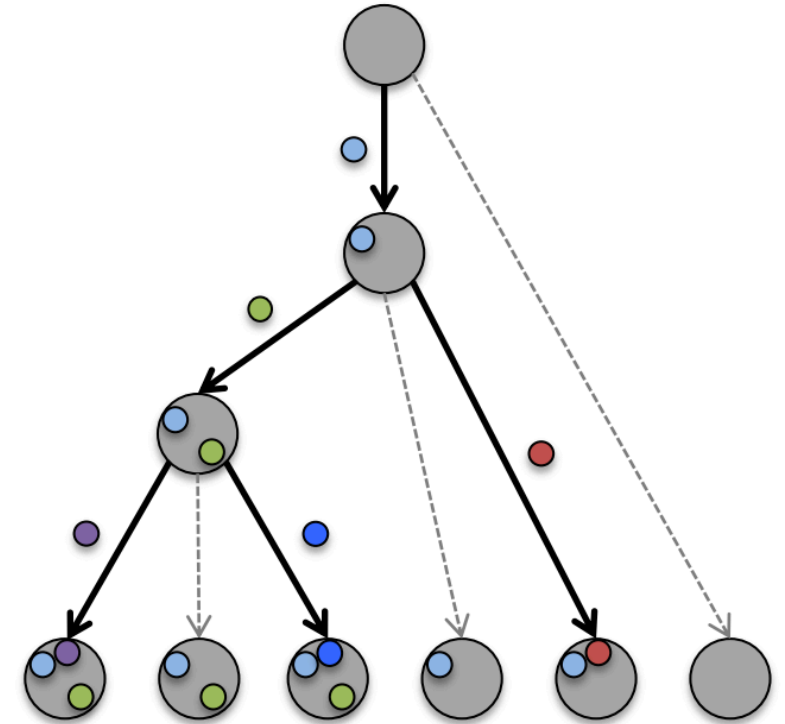
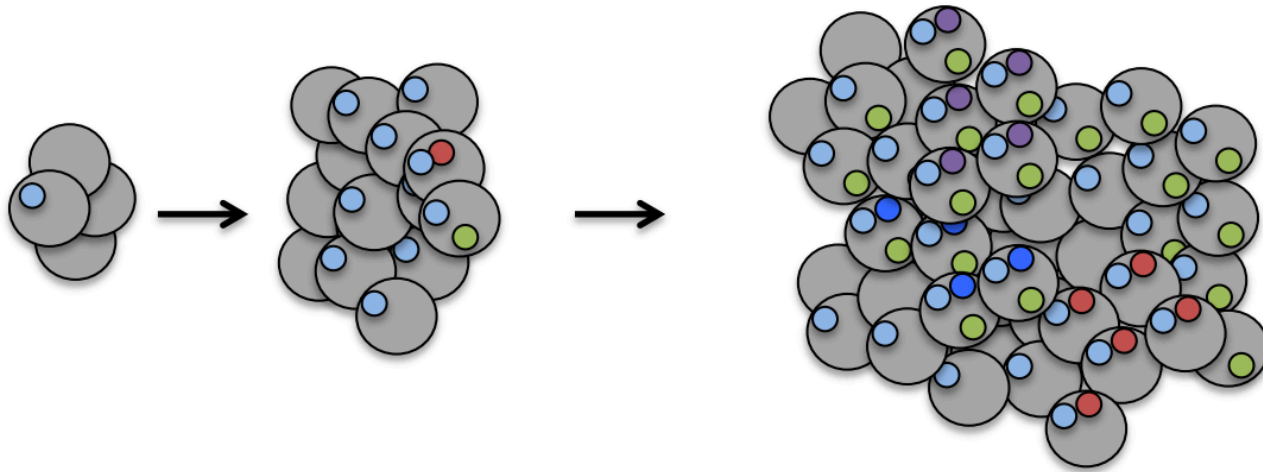
Signature Exposures at the Tumor Level



[Helleday et al., Nat Rev Genetics, 2014]

Intra-tumor Heterogeneity

Clonal Evolution Theory of Cancer [Nowell, 1976]

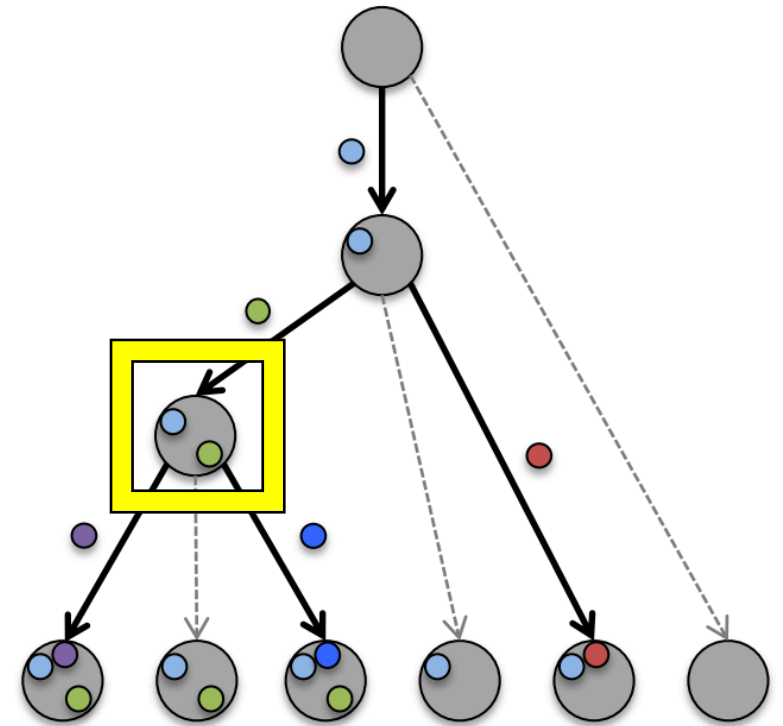


[Schwartz and Schäffer, Nat Rev Genetics, 2017]

Heterogeneity in Signatures in Tumor Clones?

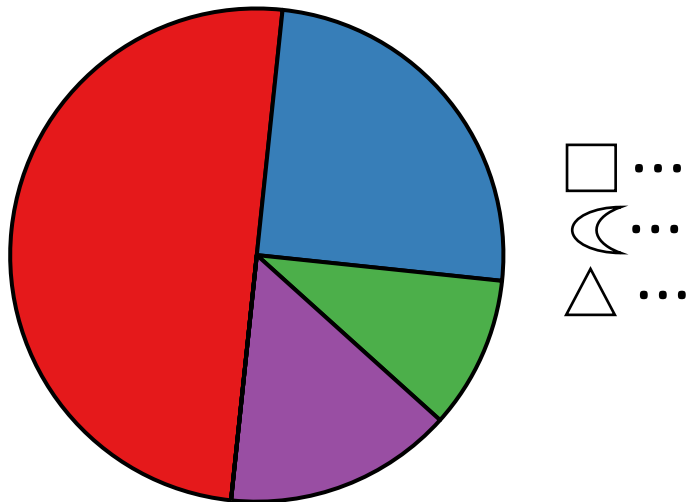
A clone can be distinguished from its parent by the set of newly introduced mutations.

Idea is to look at differences in exposures for these *newly introduced* mutations across the tree.

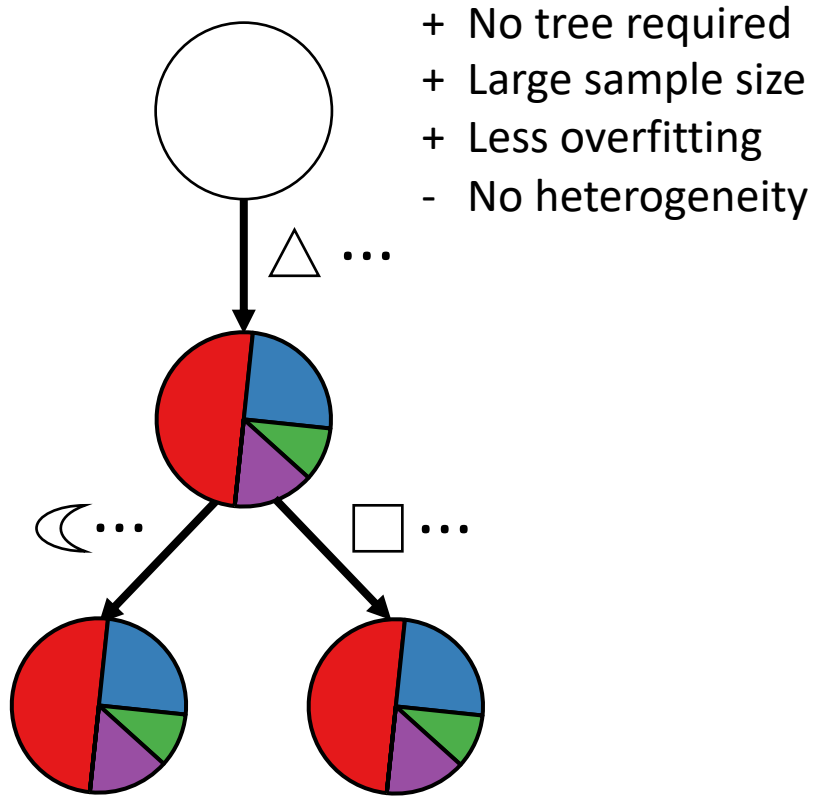


Previous: Single Exposure Inference

■ sig. 1 ■ sig. 2 ... ■ sig. 30



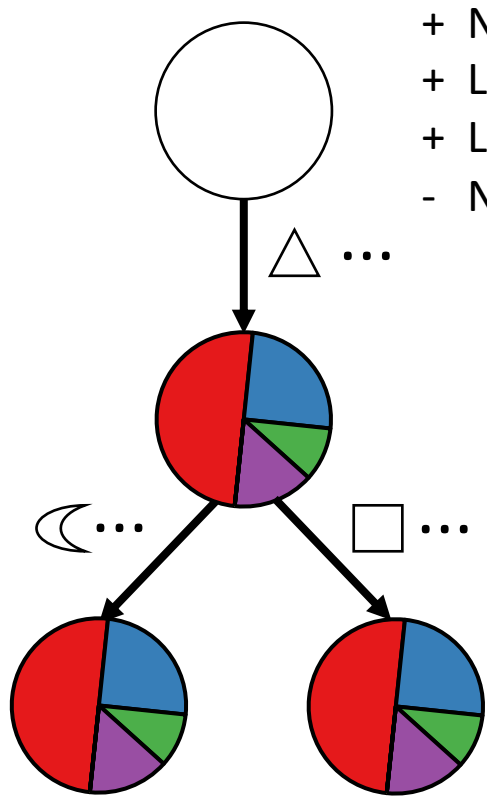
Previous: Single Exposure Inference



- + No tree required
- + Large sample size
- + Less overfitting
- No heterogeneity

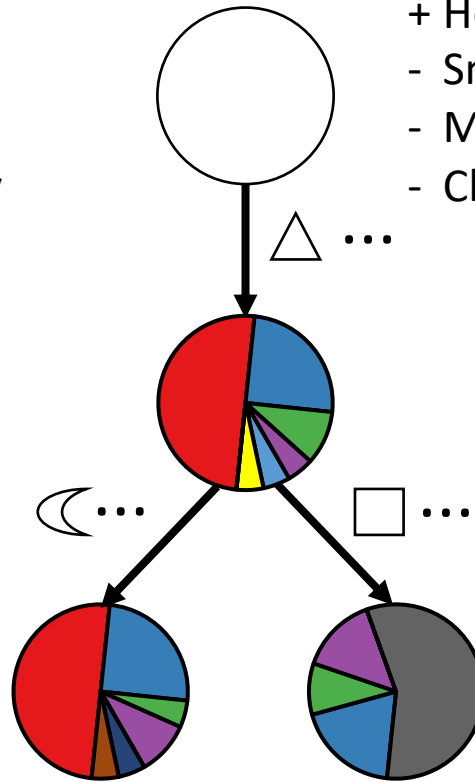
Same exposures for every clone

Previous: Independent Exposure Inference



- + No tree required
- + Large sample size
- + Less overfitting
- No heterogeneity

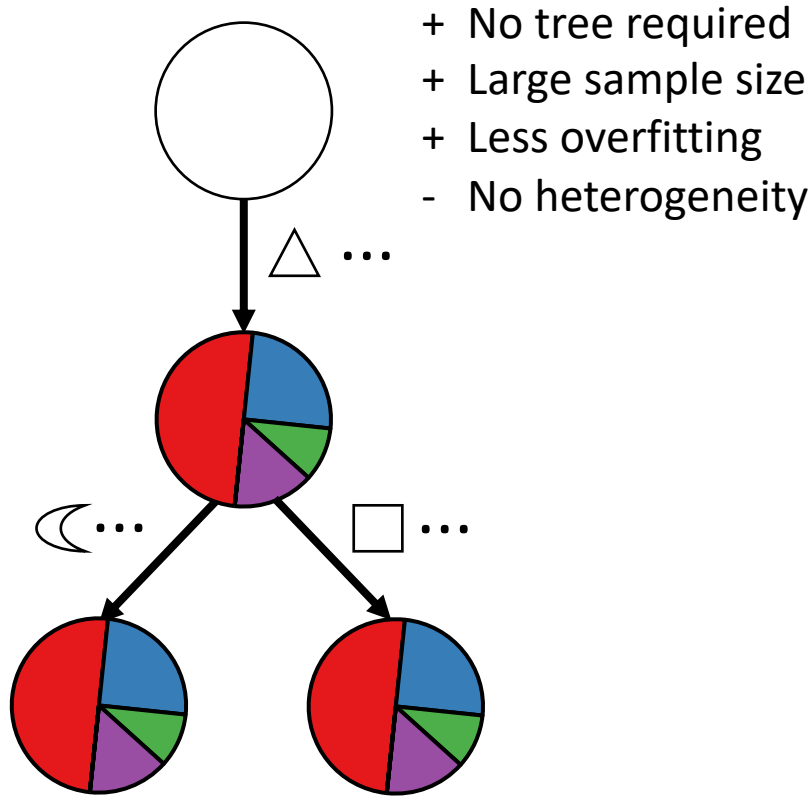
Same exposures for every clone



- + Heterogeneity
- Small sample size
- May overfit
- Clones required

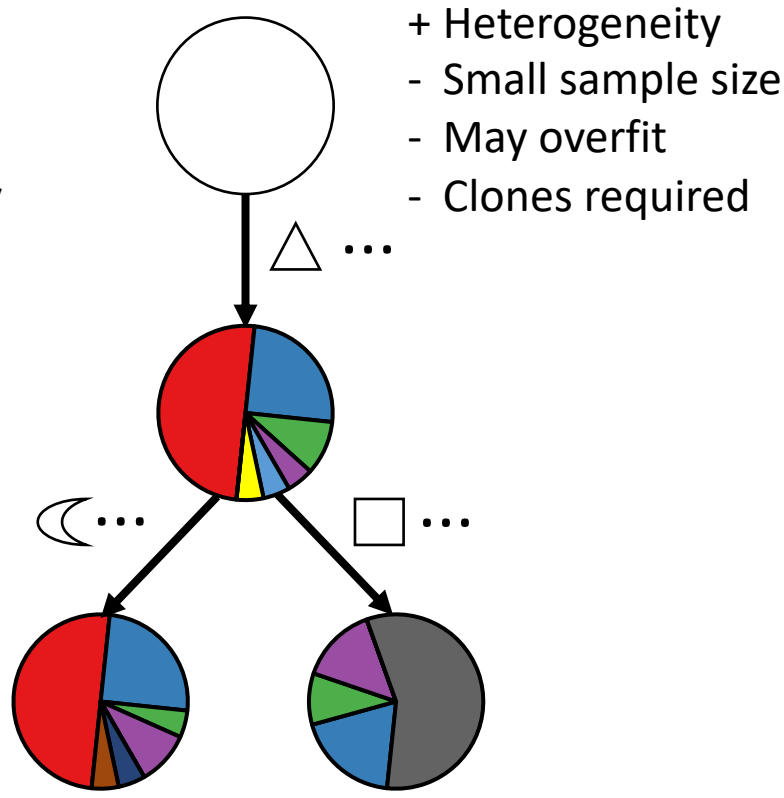
Different exposures for every clone

This Work: Tree Constrained Exposure Inference



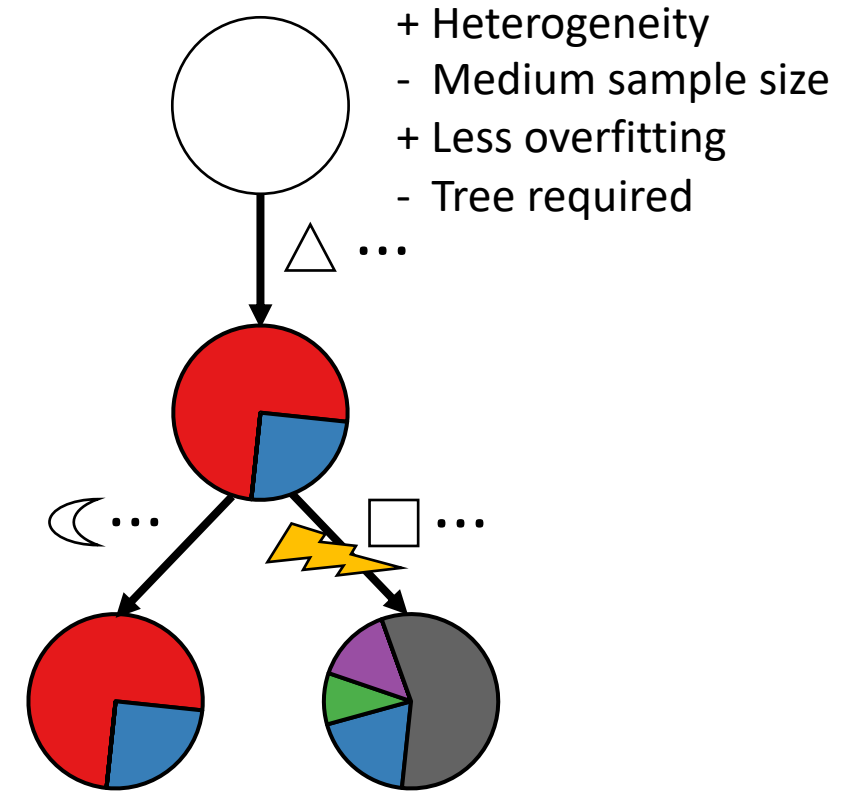
Same exposures for every clone

Alexandrov et al., 2013; Rosenthal et al., 2016
Huang et al., 2017; Blokzijl et al., 2018



Different exposures for every clone

McPherson et al., 2016;
Jamal-Hanjani et al., 2017



Clone exposures separated by **shifts**

PhySigs

Problem Statement and Methodology

Tree-constrained Exposure Problem

Problem 3 (Tree-constrained Exposure (TE)). *Given feature matrix P , corresponding count matrix C , signature matrix S , phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $\|P - SDC\|_F$ is minimum and D is composed of k sets of identical columns, each corresponding to a connected subtree of T .*

P : Mutation Count Matrix

$$\begin{matrix} & \textcircled{\triangle} & \textcircled{\smile} & \textcircled{\square} \\ \text{AC>AA TC>GG} & \begin{pmatrix} 0 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \end{matrix}$$

$$m \times n$$

Tree-constrained Exposure Problem

Problem 3 (Tree-constrained Exposure (TE)). Given feature matrix P , corresponding count matrix C , signature matrix S , phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $\|P - SDC\|_F$ is minimum and D is composed of k sets of identical columns, each corresponding to a connected subtree of T .

P : Mutation Count Matrix S : Signature Matrix

$$\begin{array}{c}
 \begin{array}{ccc}
 \textcircled{\triangle\triangle} & \textcircled{\curvearrowright} & \textcircled{\square\square} \\
 \text{AC>AA TC>GG} & & \\
 \begin{pmatrix} 0 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} & \approx & \begin{array}{cc} \text{Sig. 1} & \text{Sig. 2} \\ \begin{pmatrix} 0 & .5 \\ 1 & .5 \end{pmatrix} \end{array} \\
 m \times n & & m \times r
 \end{array}
 \end{array}$$

Tree-constrained Exposure Problem

Problem 3 (Tree-constrained Exposure (TE)). Given feature matrix P , corresponding count matrix C , signature matrix S , phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $\|P - SDC\|_F$ is minimum and D is composed of k sets of identical columns, each corresponding to a connected subtree of T .

P : Mutation Count Matrix

S : Signature Matrix

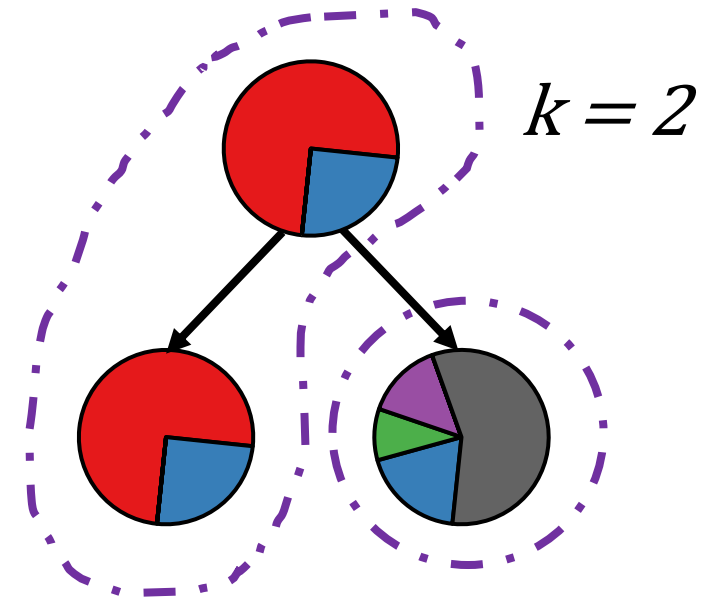
D : Relative Exposure Matrix

C : Weight Matrix

$$\begin{array}{c}
 \begin{array}{ccc}
 \begin{array}{c} \triangle \triangle \\ \circlearrowleft \end{array} & \begin{array}{c} \circlearrowleft \end{array} & \begin{array}{c} \square \square \\ \circlearrowleft \end{array} \\
 \begin{array}{c} AC>AA \\ TC>GG \end{array} \begin{pmatrix} 0 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \approx \begin{array}{cc} \text{Sig. 1} & \text{Sig. 2} \\ \begin{array}{c} AC>AA \\ TC>GG \end{array} \begin{pmatrix} 0 & .5 \\ 1 & .5 \end{pmatrix} & \begin{array}{cc} \begin{array}{c} \triangle \triangle \\ \circlearrowleft \end{array} & \begin{array}{c} \circlearrowleft \end{array} \\ \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} & \begin{array}{c} \begin{array}{c} \square \square \\ \circlearrowleft \end{array} \\ 0 & 1 \end{array} & \begin{array}{ccc} \begin{array}{c} \triangle \triangle \\ \circlearrowleft \end{array} & \begin{array}{c} \circlearrowleft \end{array} & \begin{array}{c} \square \square \\ \circlearrowleft \end{array} \\ \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} & \begin{array}{c} \triangle \triangle \\ \circlearrowleft \end{array} & \begin{array}{c} \circlearrowleft \end{array} & \begin{array}{c} \square \square \\ \circlearrowleft \end{array} \\
 m \times n & m \times r & r \times n & n \times n \\
 \underbrace{\hspace{15em}} & & & \\
 E: \text{Exposure Matrix} & & &
 \end{array}
 \end{array}$$

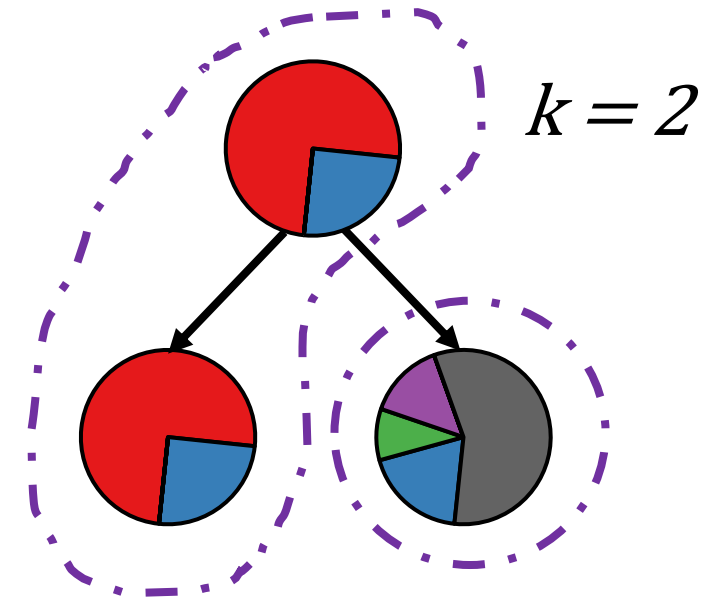
PhySigs Algorithm

Step 1: Solve TE Problem for each possible number k of clusters

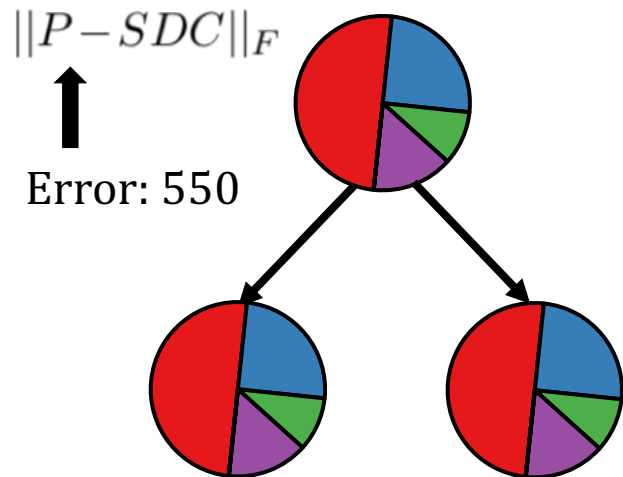


PhySigs Algorithm

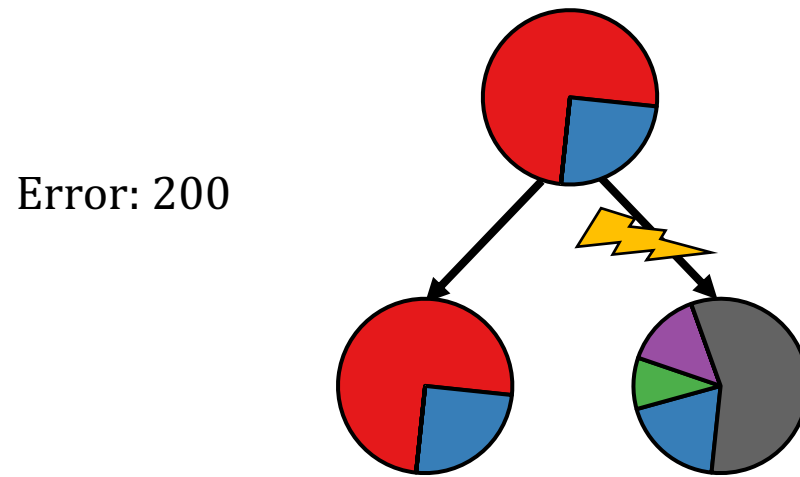
Step 1: Solve TE Problem for each possible number k of clusters



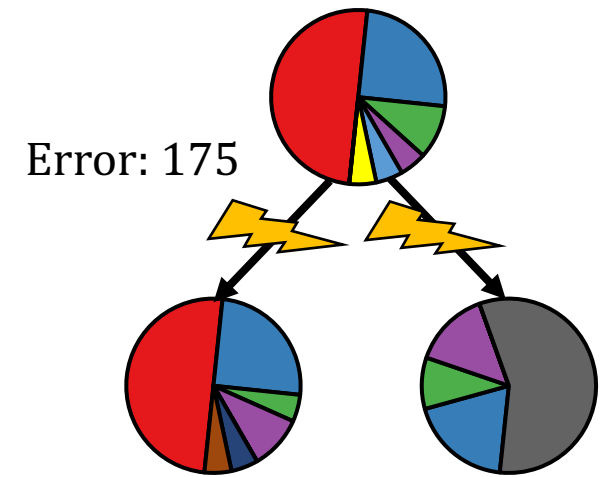
No Shift ($k=1$)



One Shift ($k=2$)



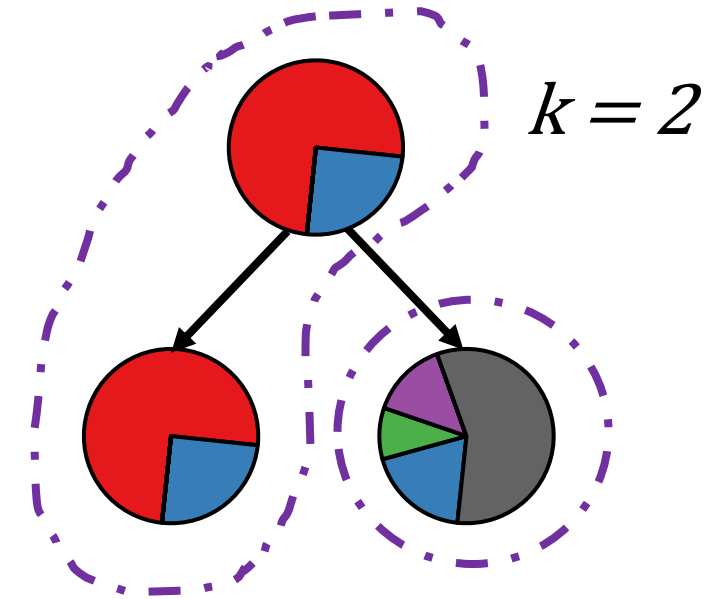
Two Shifts ($k=3$)



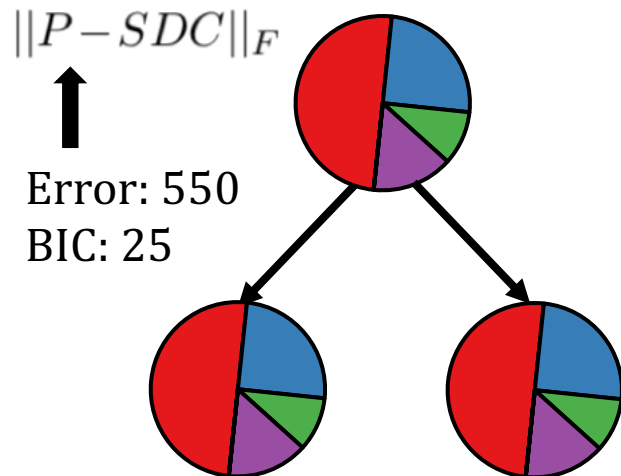
PhySigs Algorithm

Step 1: Solve TE Problem for each possible number k of clusters

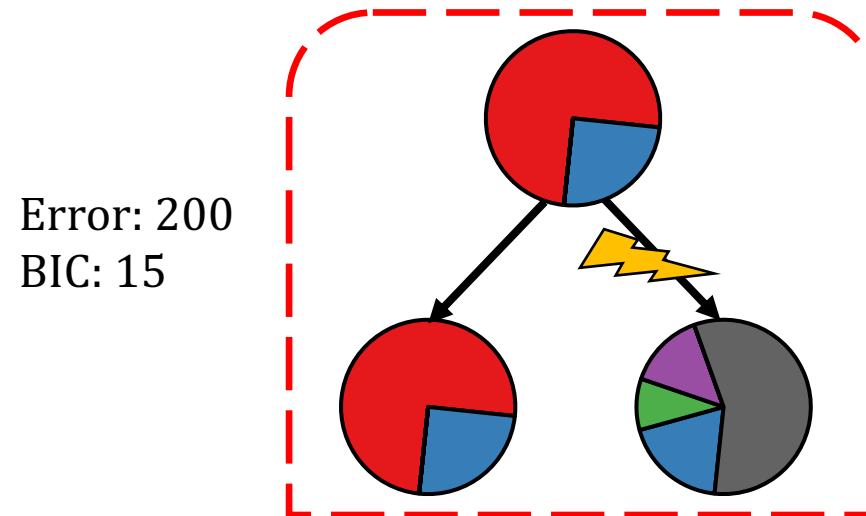
Step 2: Choose best number k of clusters using Bayesian Information Criterion (BIC)



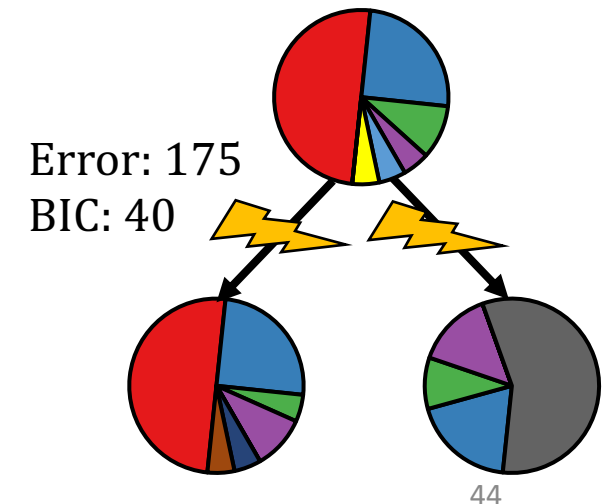
No Shift ($k=1$)



One Shift ($k=2$)



Two Shifts ($k=3$)

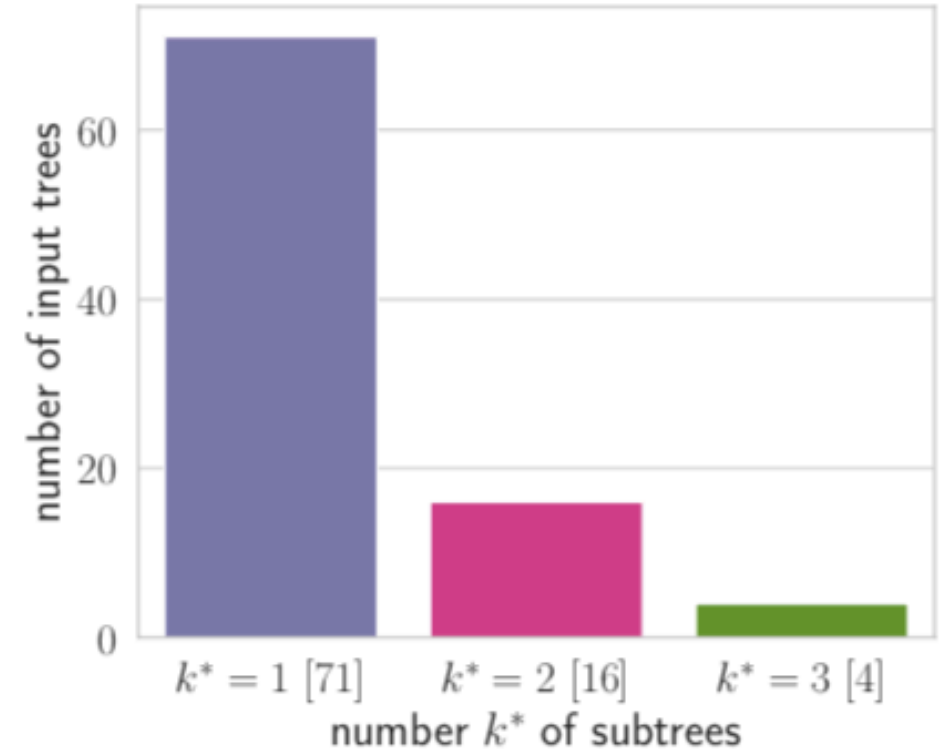


Results

On Simulated and Biological Data

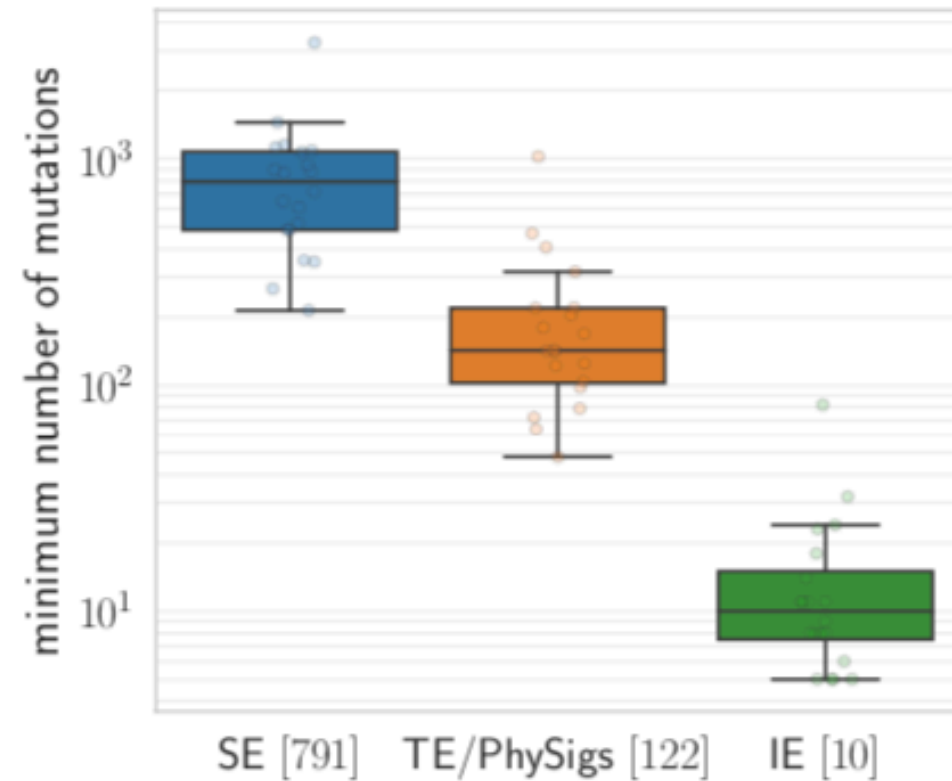
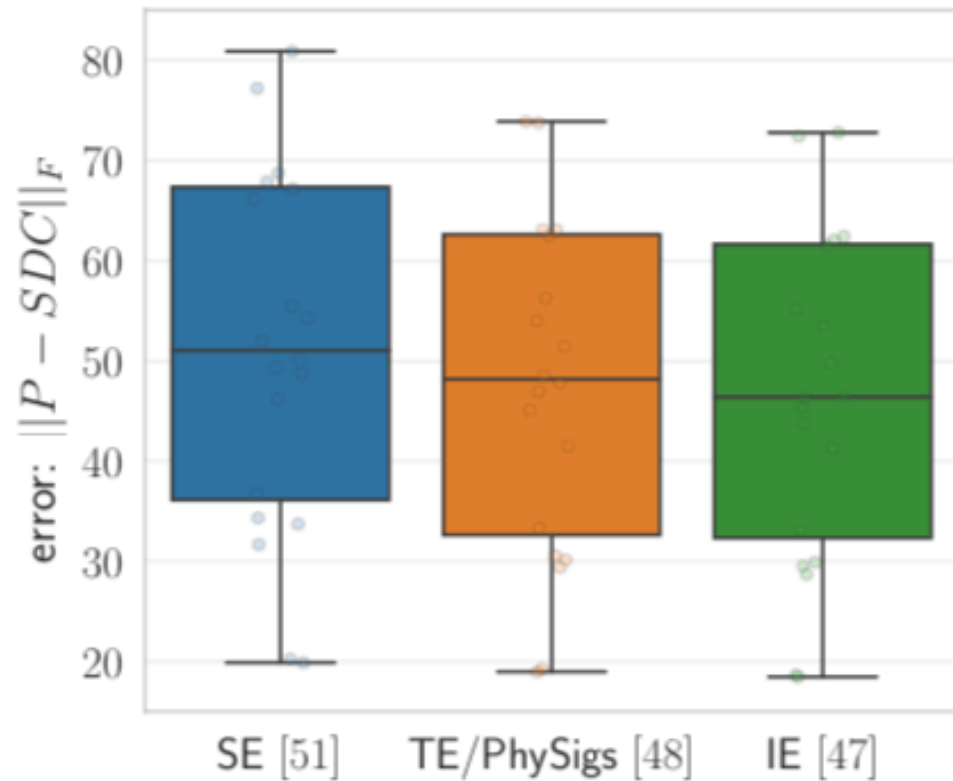
PhySigs on Lung Cancer Cohort

- 91 patient tumors
- Number of clones per patient ranges from 2 to 15 (median of 5).
- Number of equally likely trees per patient ranges from 1 to 17 (median of 1).



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

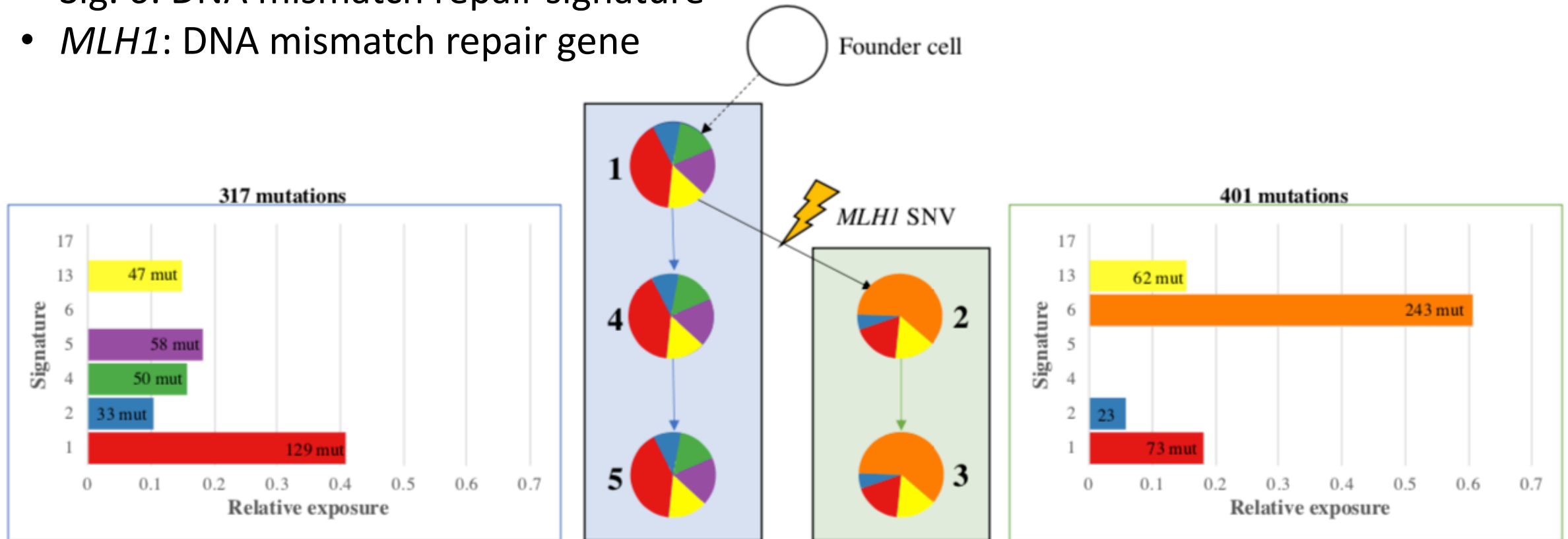
PhySigs Explains Data without Overfitting



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

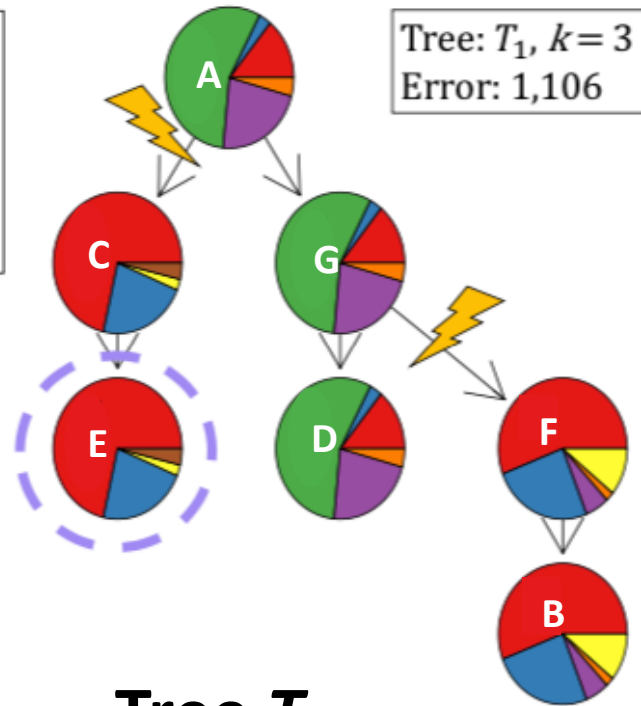
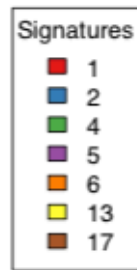
PhySigs Finds Explainable Shift for Patient CRUK0064

- Sig. 6: DNA mismatch repair signature
- *MLH1*: DNA mismatch repair gene



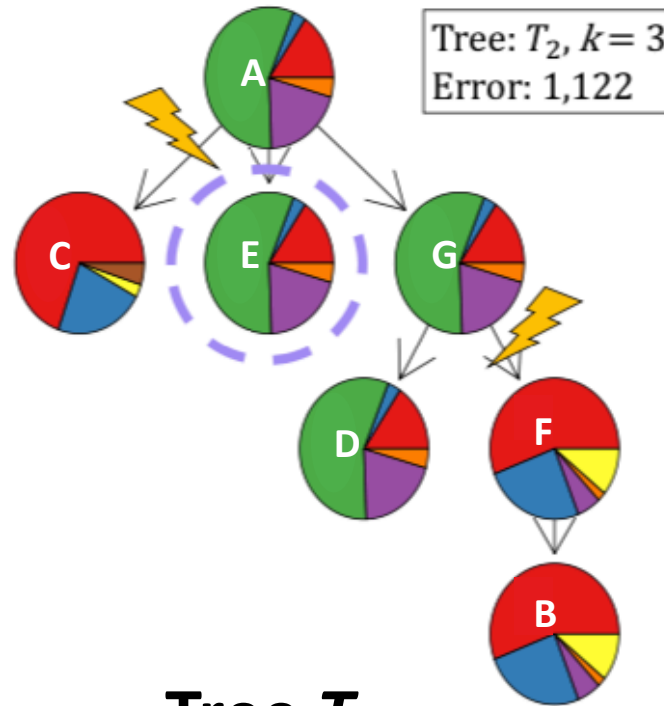
Data from [Jamal-Hanjani et al., 2017]

PhySigs Constrains Solutions for Patient CRUK0025



Tree T_1

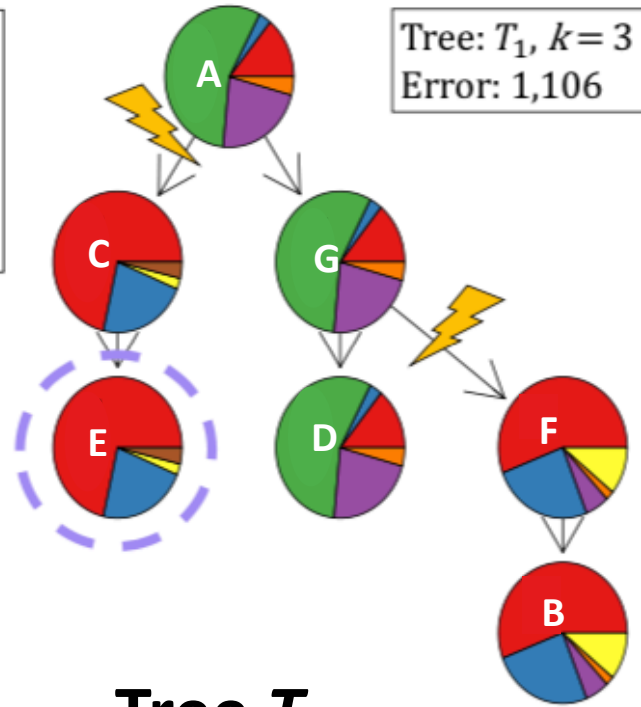
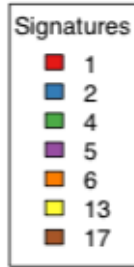
- 2 Shifts
- Small Error



Tree T_2

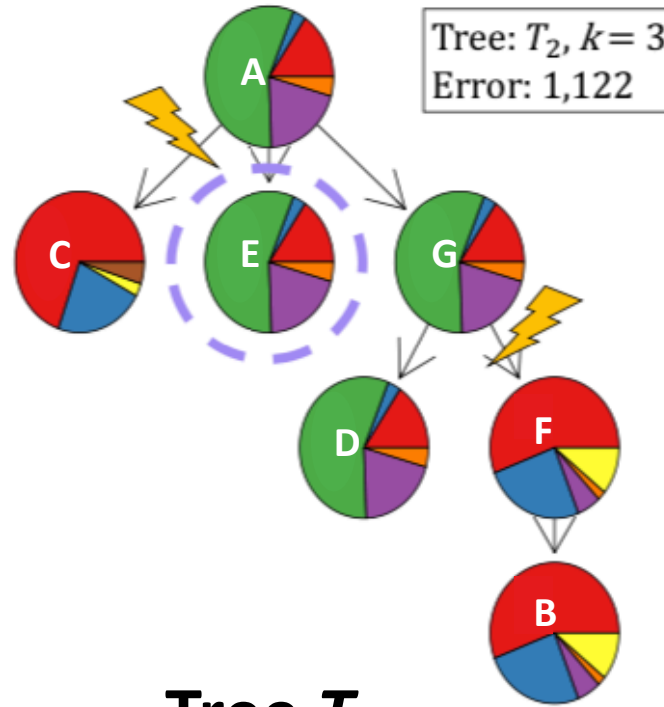
- 2 Shifts
- Big Error

PhySigs Constrains Solutions for Patient CRUK0025



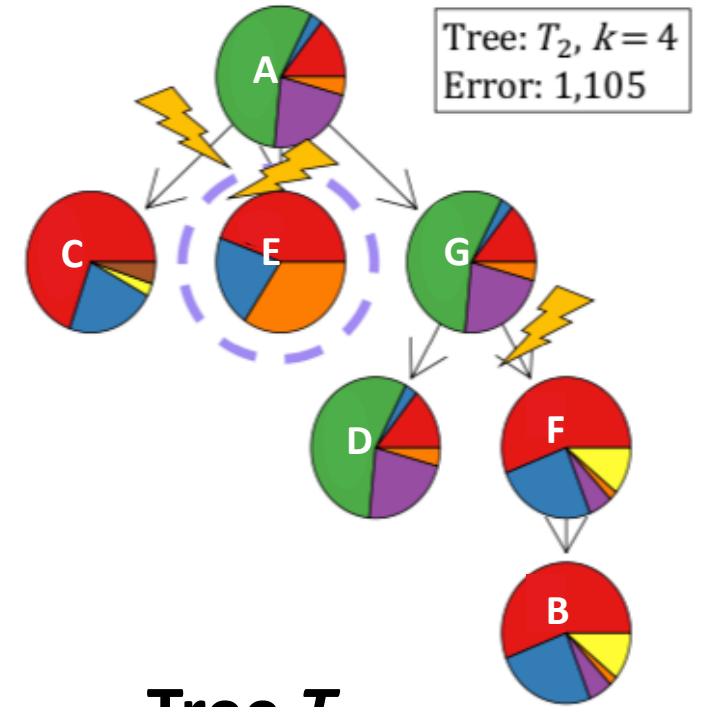
Tree T_1

- 2 Shifts
- Small Error



Tree T_2

- 2 Shifts
- Big Error



Tree T_2

- 3 Shifts
- Small Error

Conclusions and Discussion

Key concept: mutational signature exposure may not be constant across clones due to intra-tumor heterogeneity

PhySigs identifies shifts along edges of a tumor's evolutionary tree

- May want to consider additional patterns for shifts

PhySigs works by reducing to single exposure problem and then can be solved with existing algorithms

- Hardness of tree constrained exposure for a fixed k remains open

Availability: <https://github.com/elkebir-group/PhySigs>

Acknowledgements

El-Kebir lab:

- Chuanyi Zhang
- Jiaqi Wu
- Juho Kim
- Leah Weber
- Nuraini Aguse
- Sarah Christensen
- Yerong Li
- Yuanyuan Qi



National Science Foundation (IIS 15-13629 to SC, CCF 18-50502 to MEK)