Mutational Signatures

February 20, 2020

CS 598 MEB

Today’s Outline

1. **What** are mutational signatures?

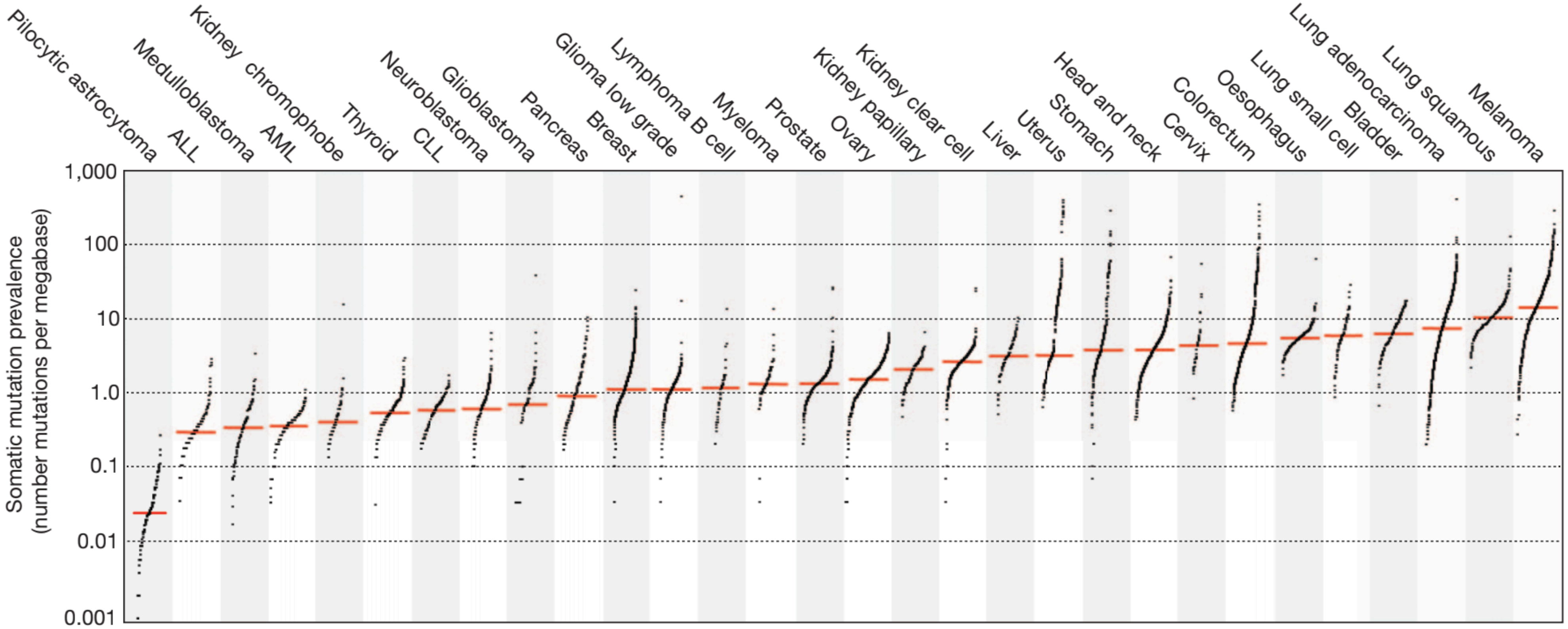
2. **How** are mutational signatures estimated?

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

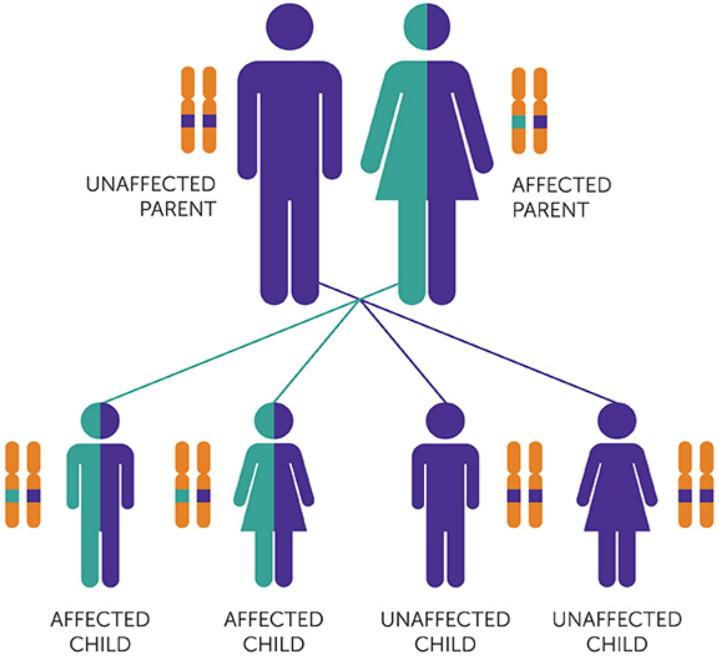


On average, 1,000-10,000 mutations/genome and 10-100 mutations/exome

Alexandrov, et al. (Nature 2013)

Where do mutations come from?

Germline

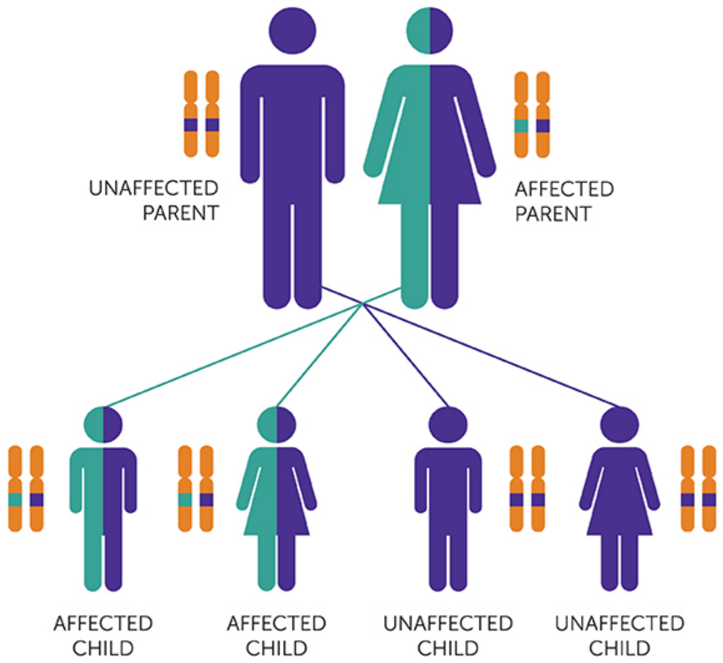


Somatic

Image credits: https://impactgenetics.com/testing-services/bap1-tumor-predisposition-syndrome-bap1-tpds/bap1-tpds-info-patients-clinicians/

Where do mutations come from?

Germline



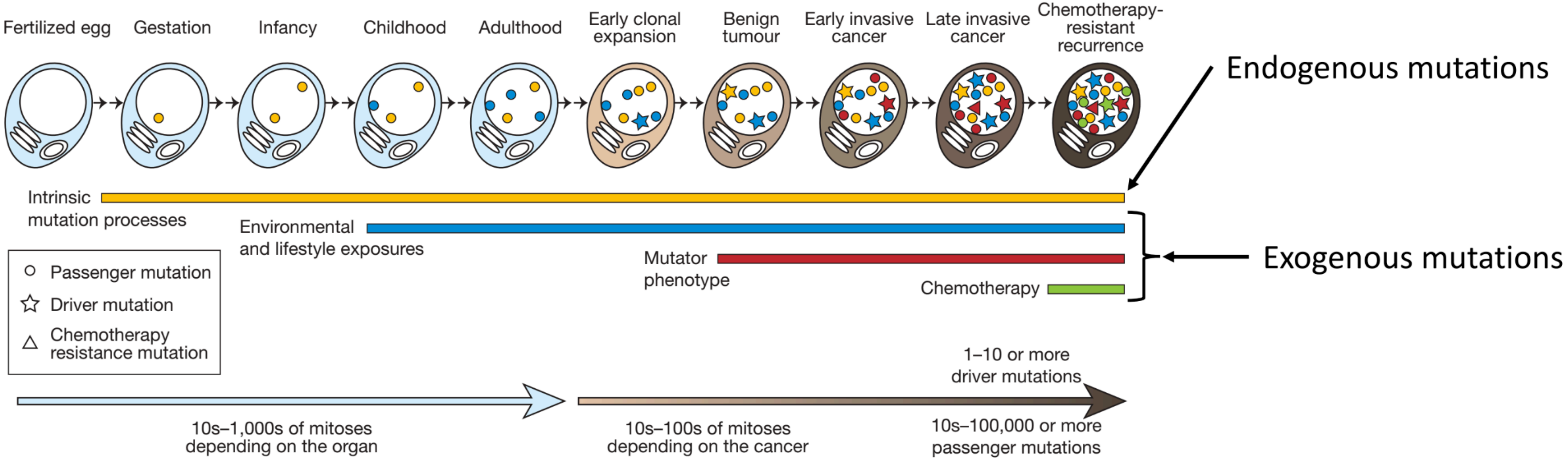
Somatic

Environmental Interference

Replication Errors

Image credits: https://impactgenetics.com/testing-services/bap1-tumor-predisposition-syndrome-bap1-tpds/bap1-tpds-info-patients-clinicians/

Somatic mutations accumulate over time


Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
1–10 or more driver mutations
10s–100,000 or more passenger mutations

Endogenous mutations

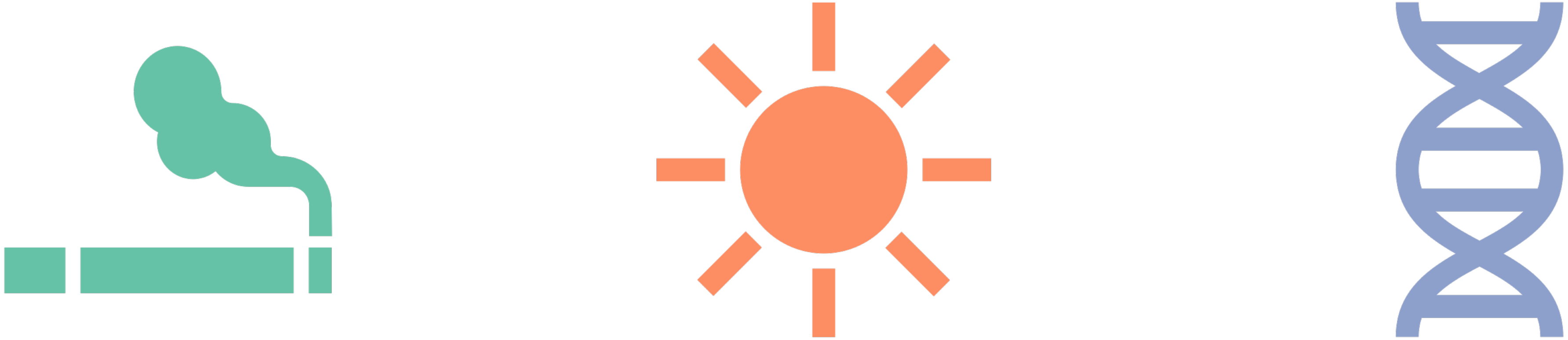
Exogenous mutations

Figure 1: Stratton et al. (Nature 2009)

Somatic mutations accumulate over time

GERMLINE: cagacccagatagggcatgagatacatccgcgagtgattggattacatctggggagtgattgatctaaactcttctcaagg

CANCER: aagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg



Problem! No labels

GERMLINE: cagacccagatagggcatgagatacatccgcgagtgattggattacatctggggagtgattgatctaaactcttctcaagg

CANCER: aagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

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: aagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 2: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 3: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

.

.

.

PATIENT n: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

Idea! Look across patients

PATIENT 1: aagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 2: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 3: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

.

.

.

PATIENT n: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

What patterns should we look at?

PATIENT 1: aagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 2: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

PATIENT 3: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

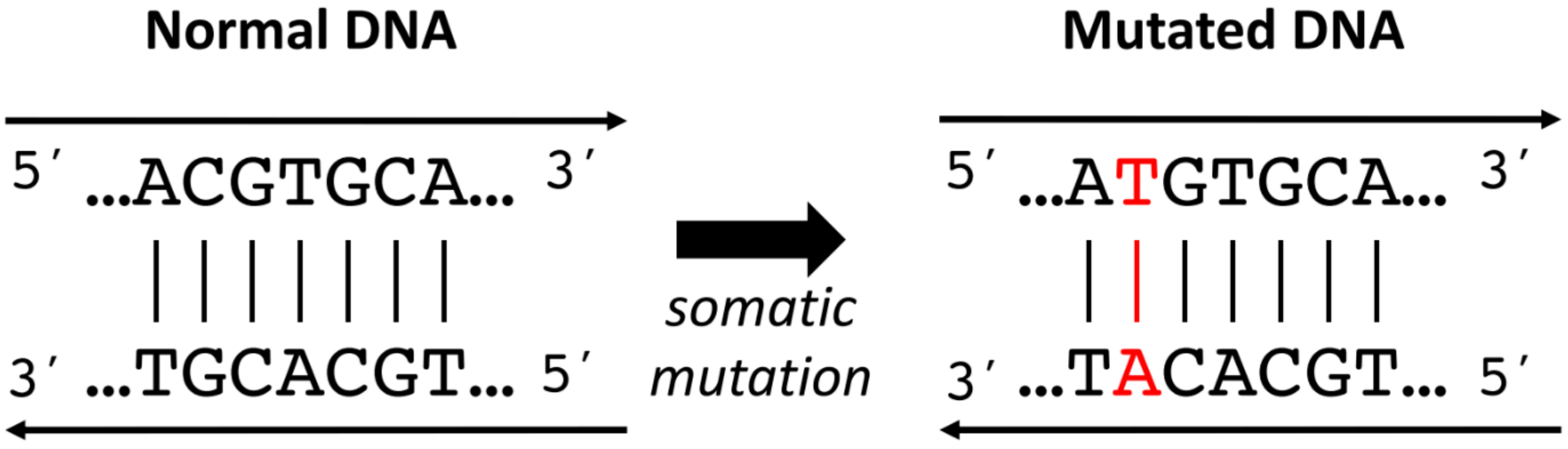
.

.

.

PATIENT n: agaagacctagttagggcatgagaaacatgcgccagtgatcggaattcatgtgggtggtgattcatctaaagtcttcgcatgg

Mutational Category: Single base substitutions



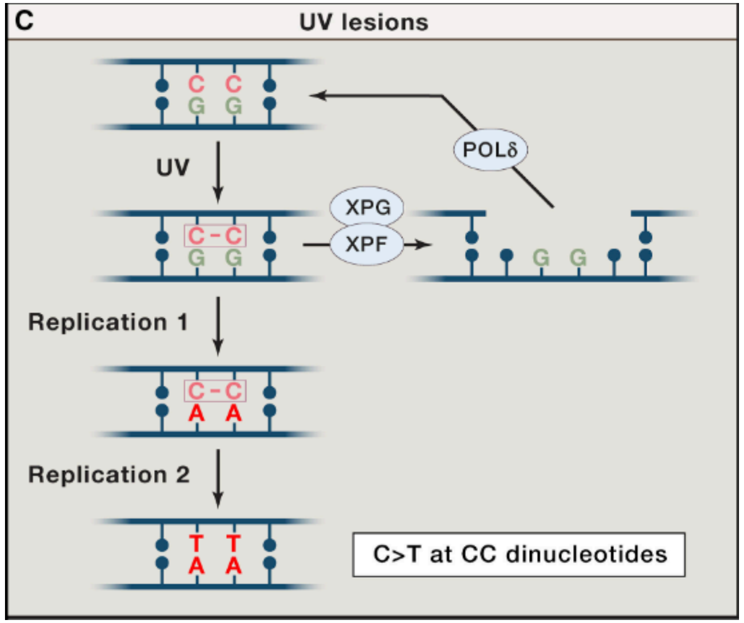
Did a C>T mutation occur? Or a G>A?

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)

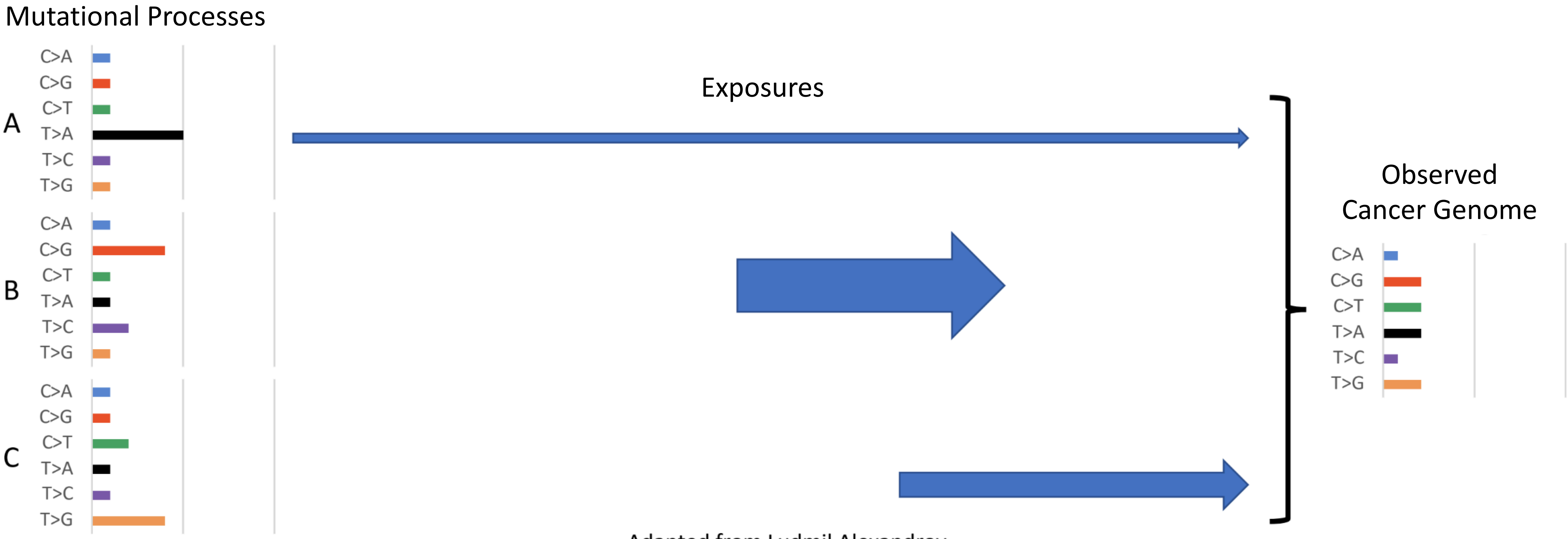
Mutational Category: Single base substitutions



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

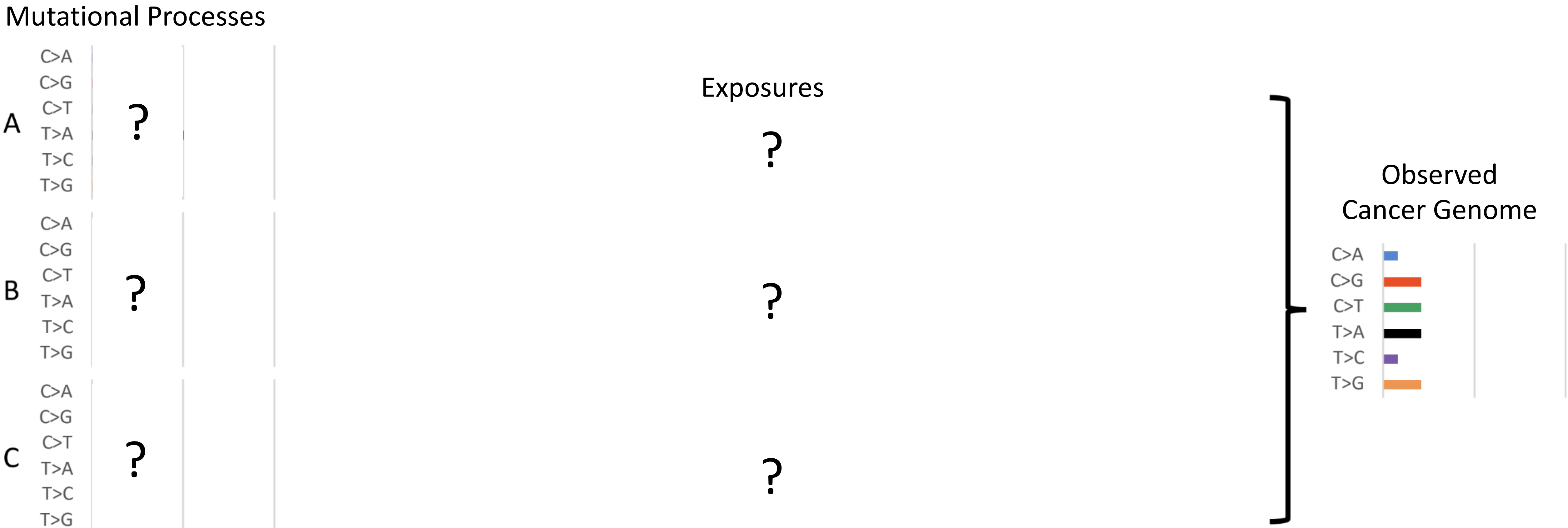
Figure 2: Tubbs & Nussenzweig. (Cell, 2017)

Integrating mutational categories into problem



Adapted from Ludmil Alexandrov

Integrating mutational categories into problem

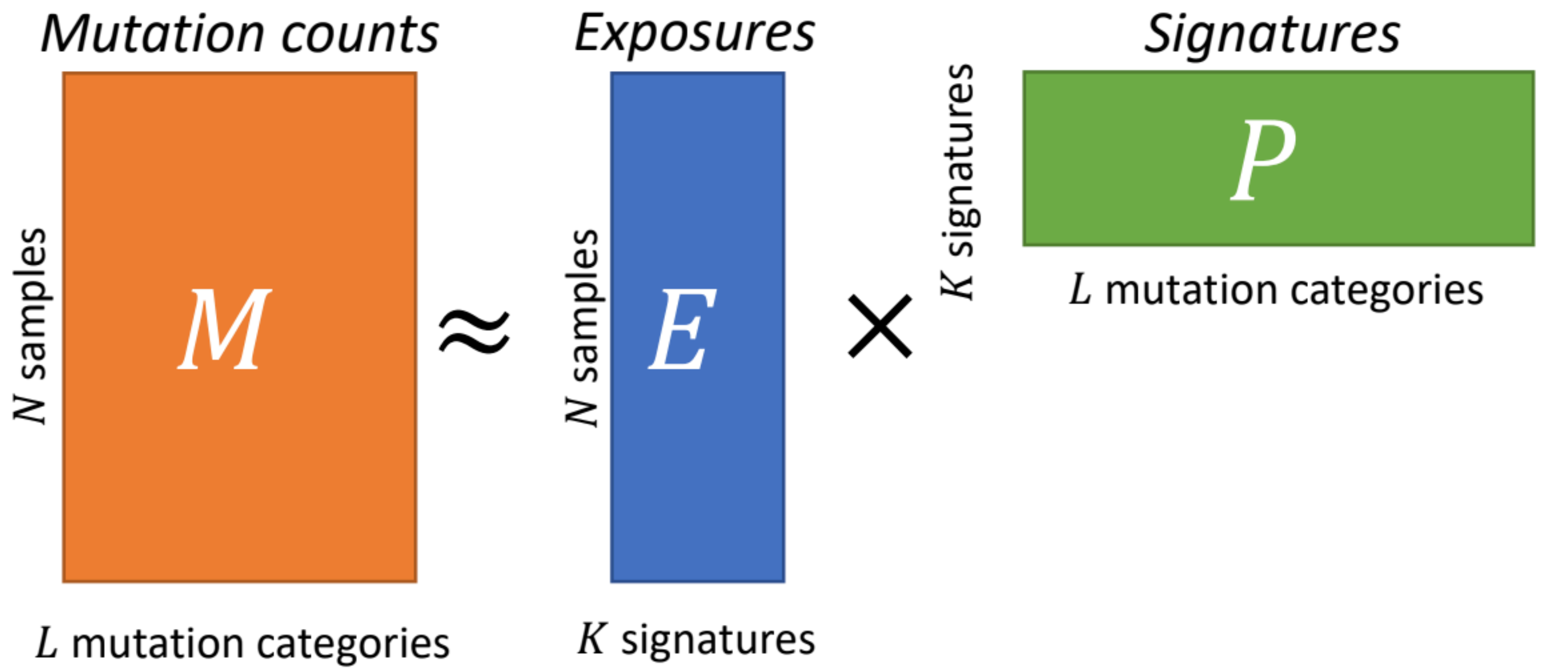

Mutational Processes

Exposures

Observed Cancer Genome

Adapted from Ludmil Alexandrov

Alexandrov Problem Statement



Objective function:

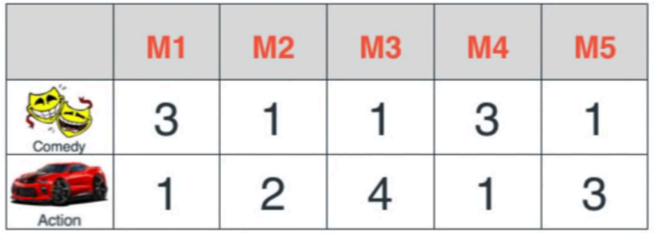
$$\min_{E,P} ||M - EP||_2^2 \text{ such that } E, P \geq 0.$$

Alexandrov, et al. (2013)

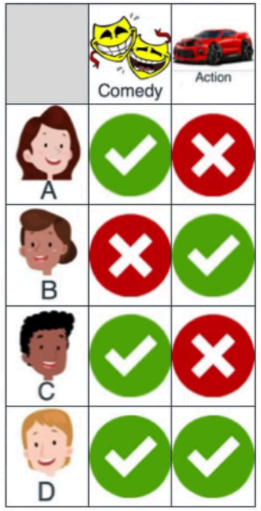
Alexandrov Algorithmic Technique

Non-negative Matrix Factorization (NMF)

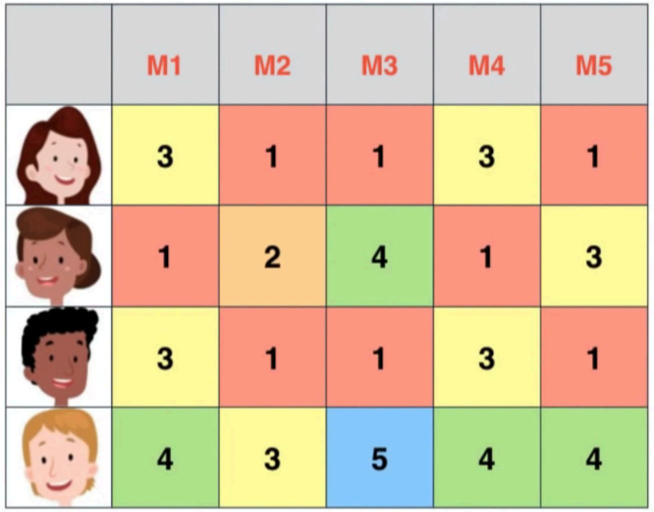
- Unique solution not guaranteed
- Popular search heuristics use alternating optimization
- Used in many applications



	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

https://www.youtube.com/watch?v=ZspR5PZemcs

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

https://www.youtube.com/watch?v=ZspR5PZemcs

Alexandrov Mutational Category: SBS with flanking

$$\underline{\quad \text{A} \quad} \qquad \underline{\quad \text{C} > \text{T} \quad} \qquad \underline{\quad \text{G} \quad}$$

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

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)

Validated Mutational Signatures

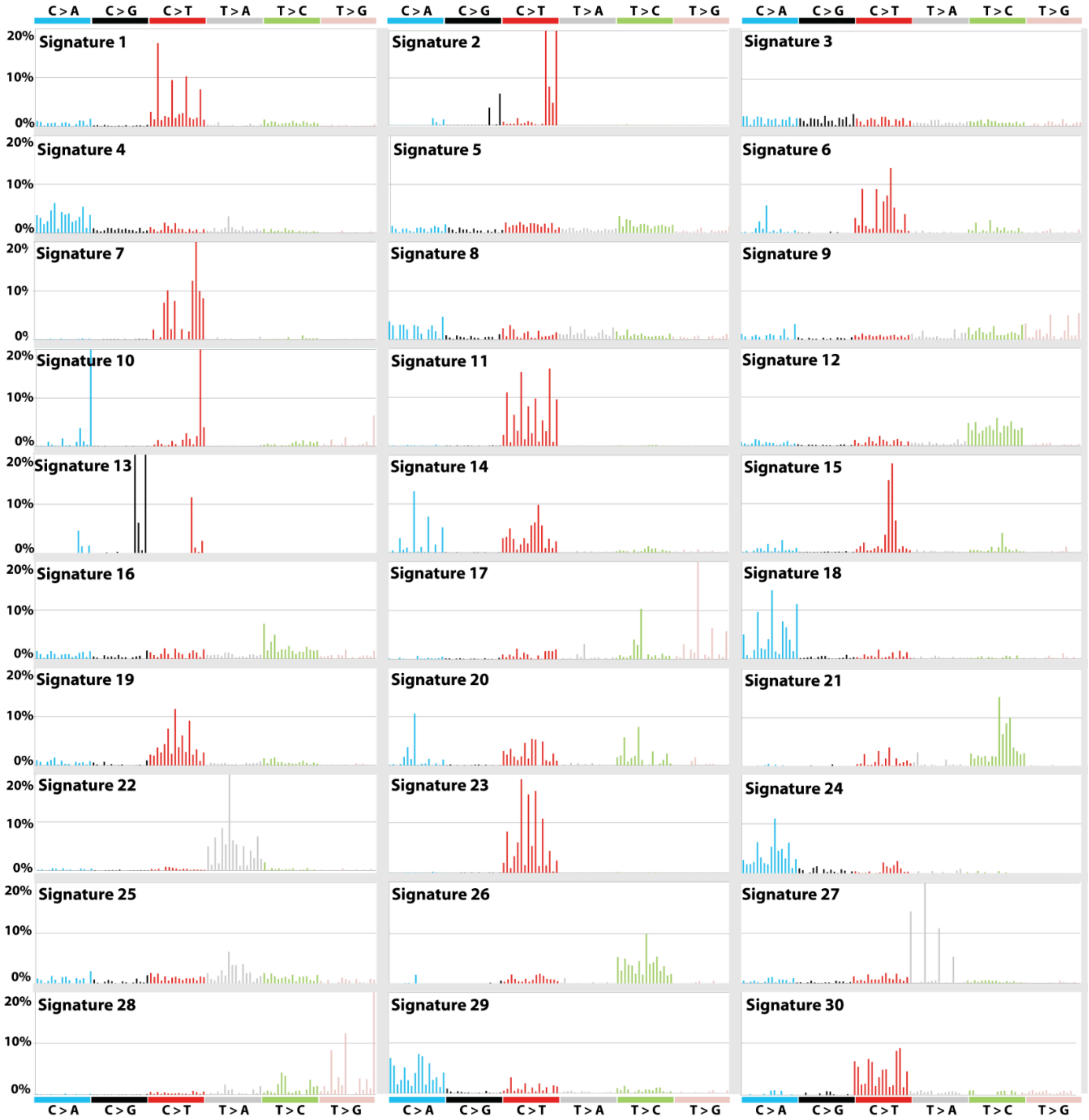
		Adrenocortical carcinoma	ALL	AML	Bladder	Breast	Cervix	Chondrosarcoma	CLL	Colorectum	Glioblastoma	Glioma Low Grade	Head and Neck	Kidney Chromophobe	Kidney Clear Cell	Kidney Papillary	Liver	Lung Adeno	Lung Small Cell	Lung Squamous	Lymphoma B-cell	Lymphoma Hodgkin	Medulloblastoma	Melanoma	Myeloma	Nasopharyngeal Carcinoma	Neuroblastoma	Oesophagus	Oral gingivo-buccal squamous	Osteosarcoma	Ovary	Pancreas	Paraganglioma	Pilocytic Astrocytoma	Prostate	Stomach	Thyroid	Urothelial Carcinoma	Uterine Carcinoma	Uterine Carcinosarcoma	Uveal Melanoma
		7+	4	2	5	13+	7+	2	5	4+	3+	4	6	3	4	4	10+	7	4	5+	6	3	3	5+	4	5	3	7	6	6	3	6+	2	3	3	11+	4	5	8	6	3
Signature 1	40	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Signature 2	22	🟩	🟩		🟩	🟩	🟩		🟩				🟩			🟩		🟩		🟩	🟩				🟩	🟩		🟩	🟩	🟩		🟩				🟩	🟩	🟩	🟩	🟩	
Signature 3	3					🟩																									🟩	🟩									
Signature 4	7	🟩											🟩				🟩	🟩	🟩	🟩								🟩													
Signature 5	40	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Signature 6	17	🟩				🟩	🟩			🟩		🟩		🟩	🟩		🟩	🟩								🟩		🟩		🟩		🟩			🟩				🟩	🟩	🟩
Signature 7	3												🟩											🟩					🟩												
Signature 8	2					🟩																	🟩																		
Signature 9	2								🟩												🟩																				
Signature 10	6				🟩	🟩	🟩			🟩																													🟩	🟩	
Signature 11	2										🟩													🟩																	
Signature 12	1																🟩																								
Signature 13	22	🟩	🟩		🟩	🟩	🟩		🟩				🟩			🟩		🟩		🟩	🟩				🟩	🟩		🟩	🟩	🟩		🟩				🟩	🟩	🟩	🟩	🟩	
Signature 14	2											🟩																											🟩		
Signature 15	2																		🟩																	🟩					
Signature 16	1																🟩																								
Signature 17	6					🟩											🟩	🟩			🟩			🟩				🟩								🟩					
Signature 18	3	🟩				🟩																					🟩									🟩					
Signature 19	1																																	🟩							
Signature 20	2					🟩																														🟩					
Signature 21	1																																			🟩					
Signature 22	2																🟩																					🟩			
Signature 23	1																🟩																								
Signature 24	1																🟩																								
Signature 25	1																					🟩																			
Signature 26	4					🟩	🟩																													🟩			🟩		
Signature 27	1														🟩																										
Signature 28	1																																			🟩					
Signature 29	1																												🟩												
Signature 30	2					🟩																								🟩											
Other signatures		🟩				🟩	🟩			🟩	🟩						🟩			🟩				🟩								🟩				🟩					

🟩 Mutational signature present

🟦 Total validated mutational signatures in a cancer type

🟥 Total cancer types in which a signature is operative

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

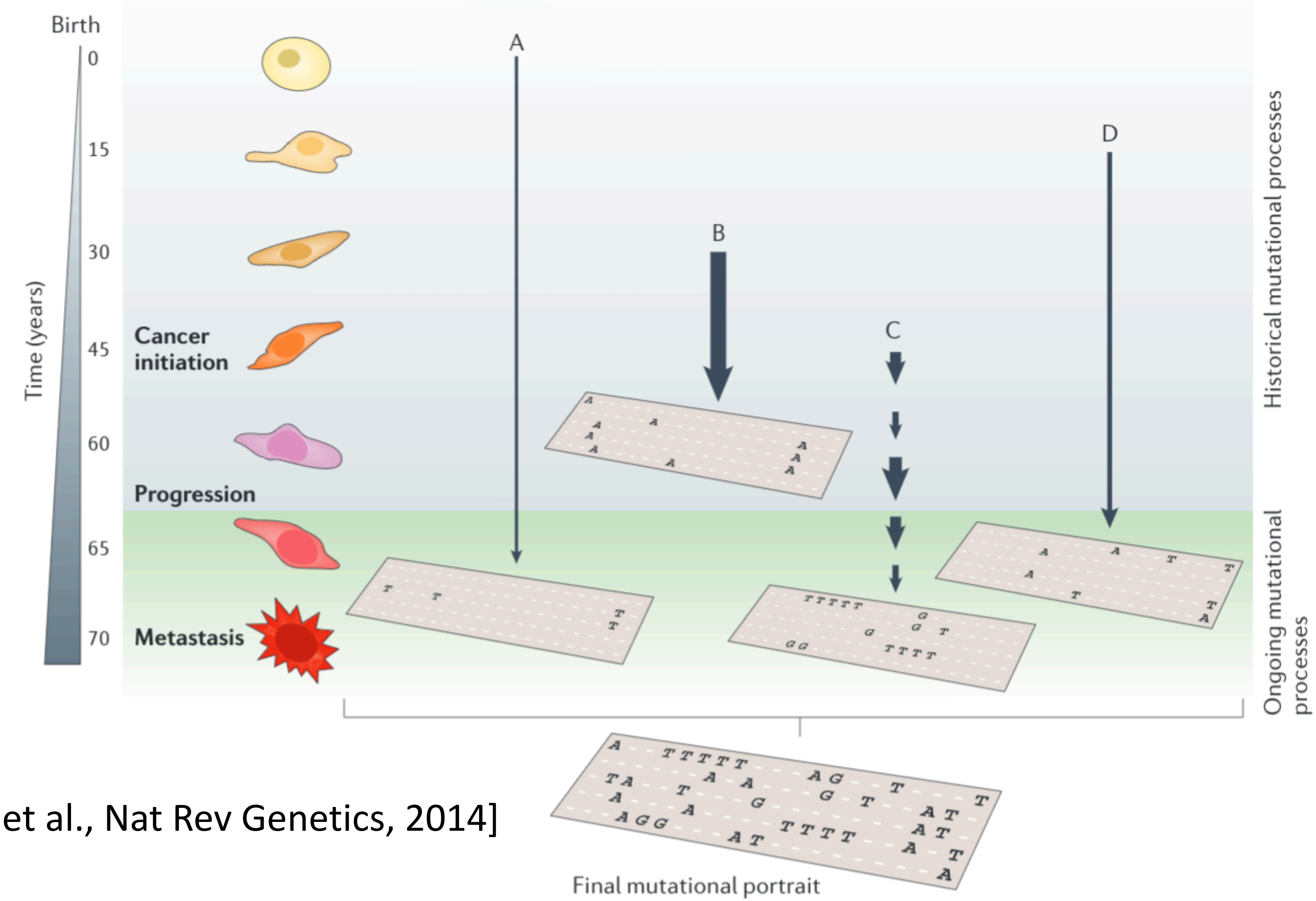
PhySigs: Phylogenetic Inference of Mutational Signature Dynamics

Sarah Christensen[1], Mark D.M. Leiserson[2], and Mohammed El-Kebir[1]

[1]Dept. of CS, University of Illinois at Urbana-Champaign
[2]Dept. of CS, University of Maryland College Park

PSB 2020

Signature Exposures at the Tumor Level

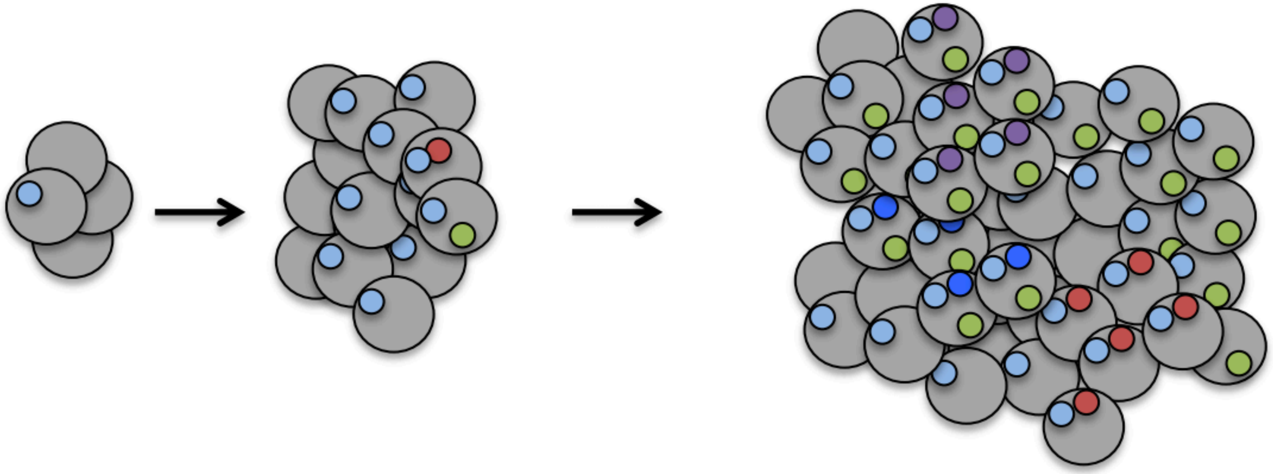

Birth
0
15
30
45
60
65
70
Time (years)
Cancer initiation
Progression
Metastasis
A
B
C
D
Historical mutational processes
Ongoing mutational processes
Final mutational portrait

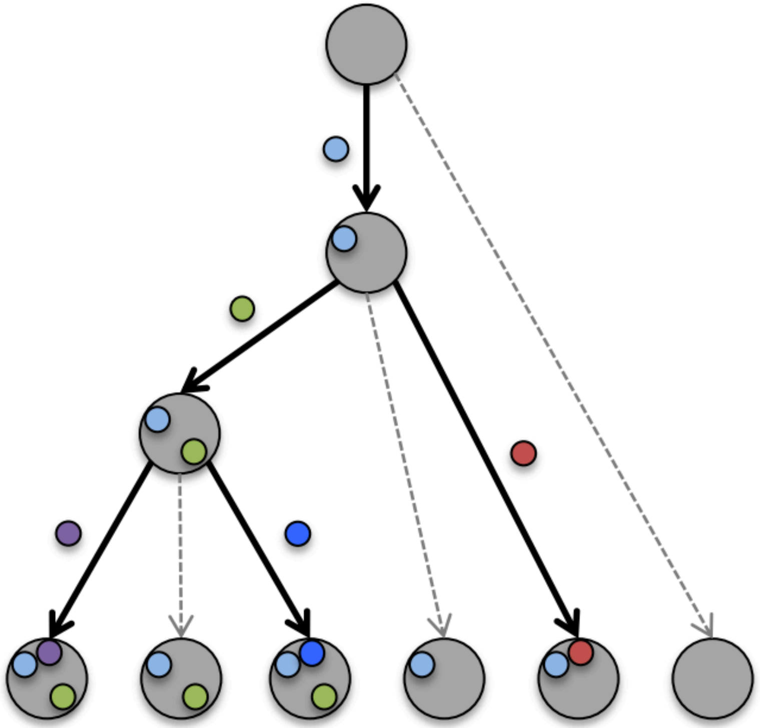
[Helleday et al., Nat Rev Genetics, 2014]

Intra-tumor Heterogeneity

Clonal Evolution Theory of Cancer

[Nowell, 1976]



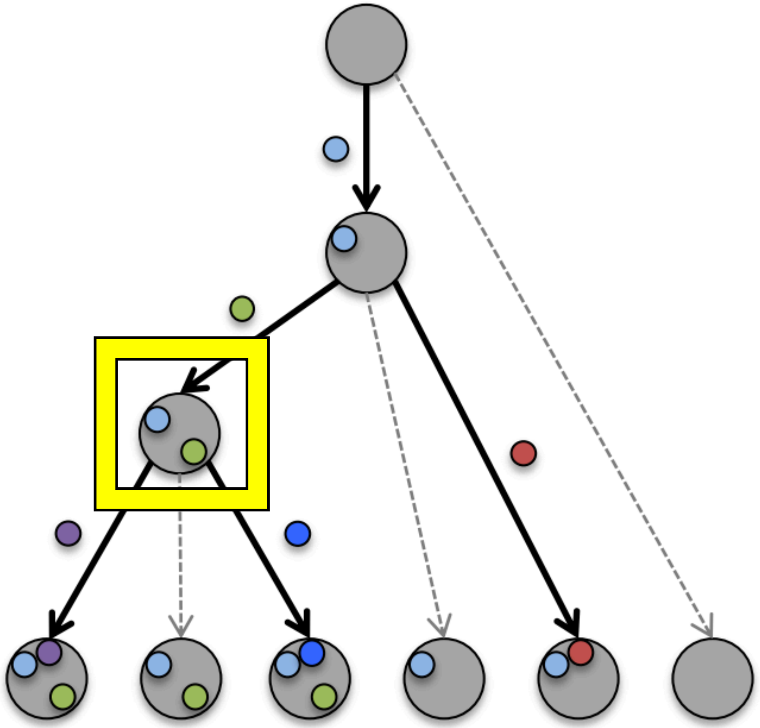


[Schwartz and Schäffer, Nat Rev Genetics, 2017]

Heterogeneity in Signatures in Tumor Clones?

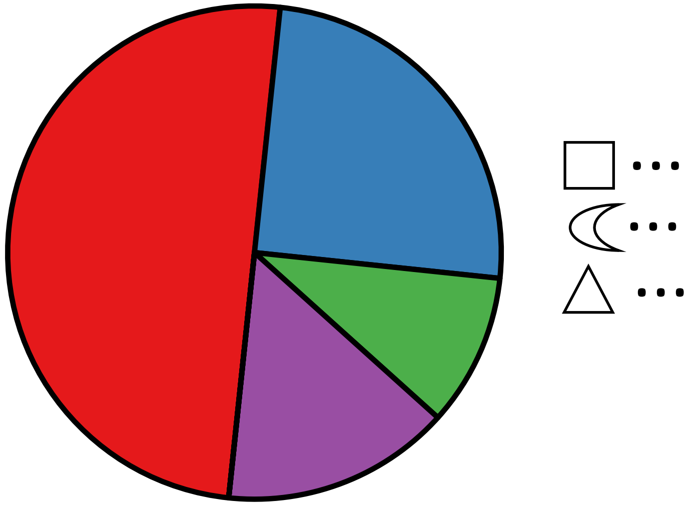
A clone can be distinguished from its parent by the set of newly introduced mutations.

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



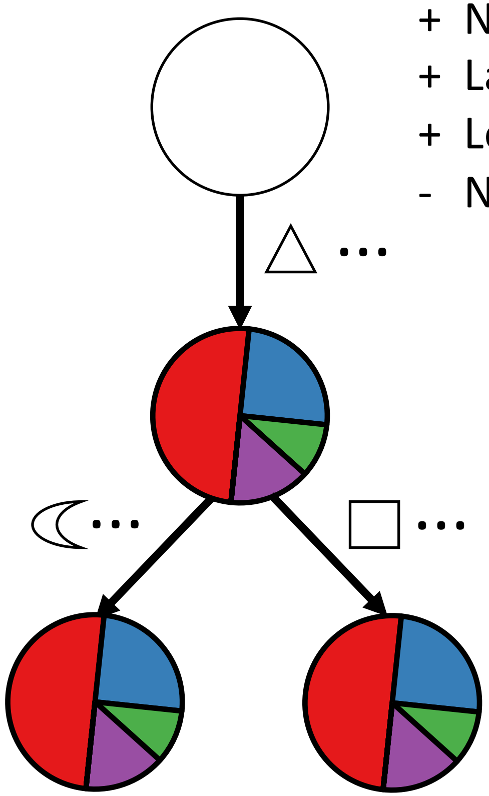
Previous: Single Exposure Inference

🟥 sig. 1 🟦 sig. 2 ··· 🟪 sig. 30



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

Previous: Single Exposure Inference

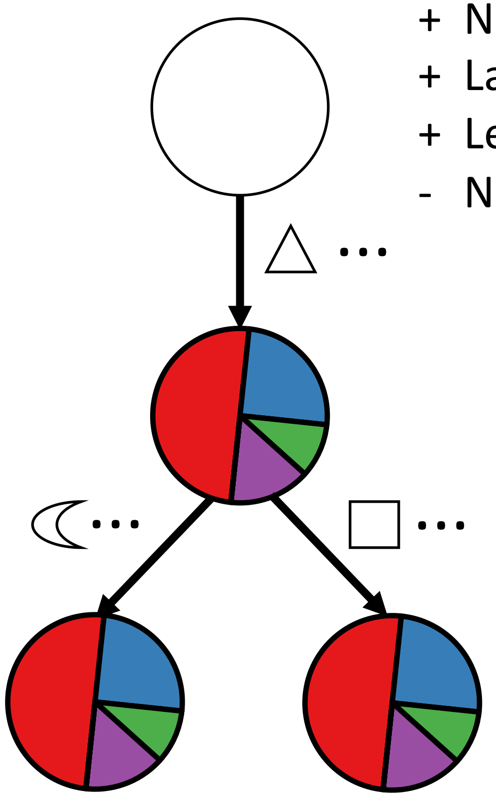


+ No tree required
+ Large sample size
+ Less overfitting
- No heterogeneity

Same exposures for every clone

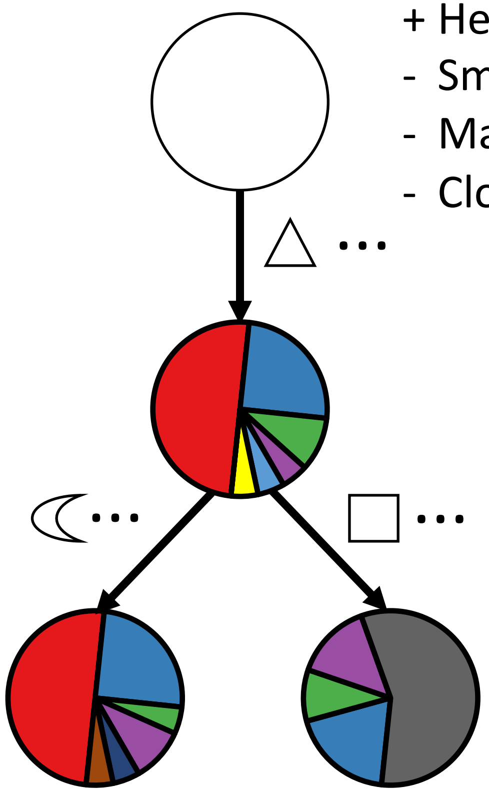
Alexandrov et al., 2013; Rosenthal et al., 2016
Huang et al., 2017;Blokzijl et al,. 2018

Previous: Independent Exposure Inference



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

Same exposures for every clone



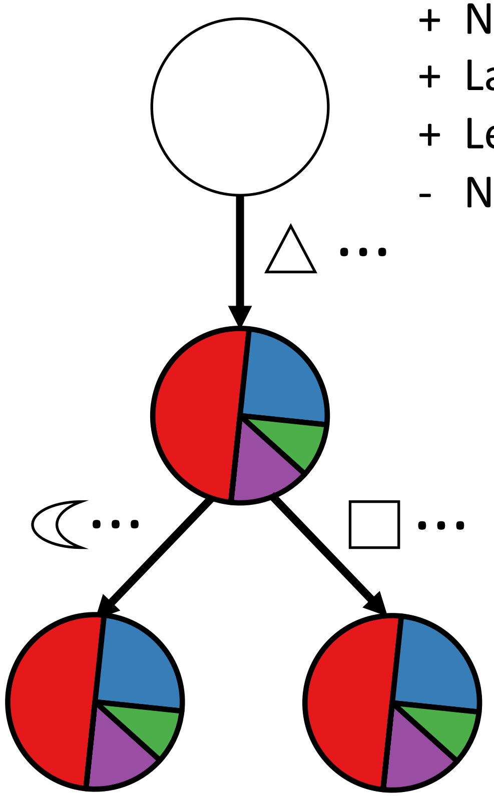
+ Heterogeneity
- Small sample size
- May overfit
- Clones required

Different exposures for every clone

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

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

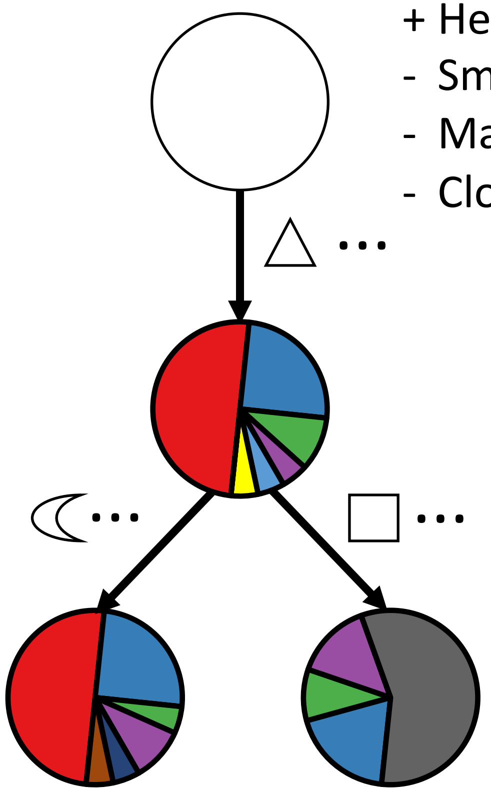
This Work: Tree Constrained Exposure Inference



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

<span style="color:red">Same</span> exposures for every clone

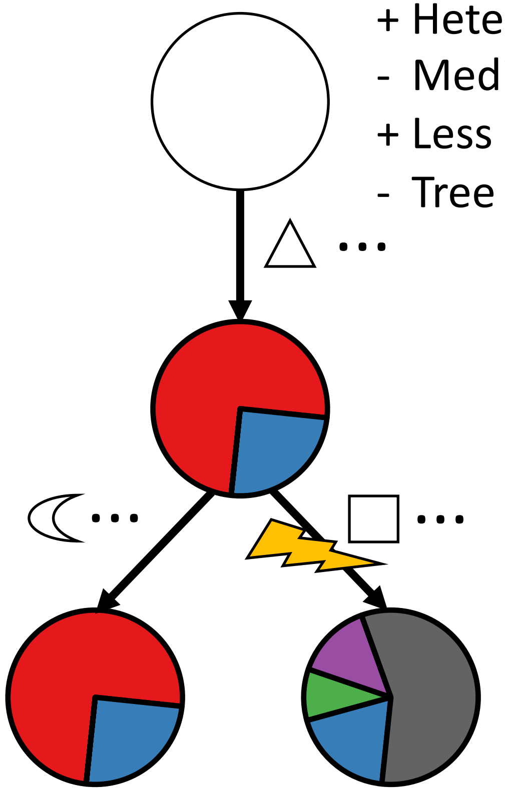
Alexandrov et al., 2013; Rosenthal et al., 2016
Huang et al., 2017;Blokzijl et al,. 2018



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

Different exposures for every clone

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



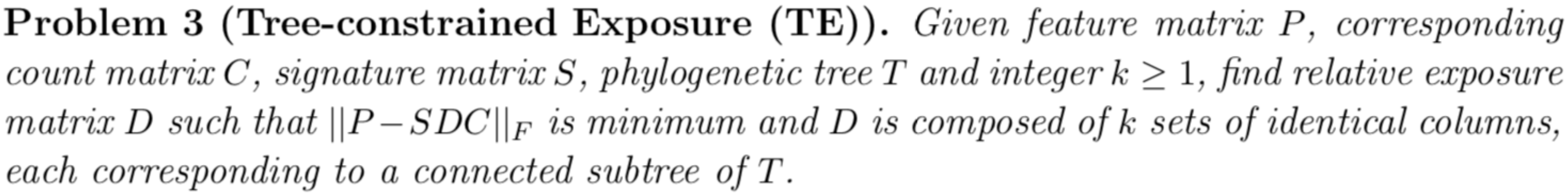
- + Heterogeneity
- - Medium sample size
- + Less overfitting
- - Tree required

Clone exposures separated by shifts

PhySigs

Problem Statement and Methodology

Tree-constrained Exposure Problem



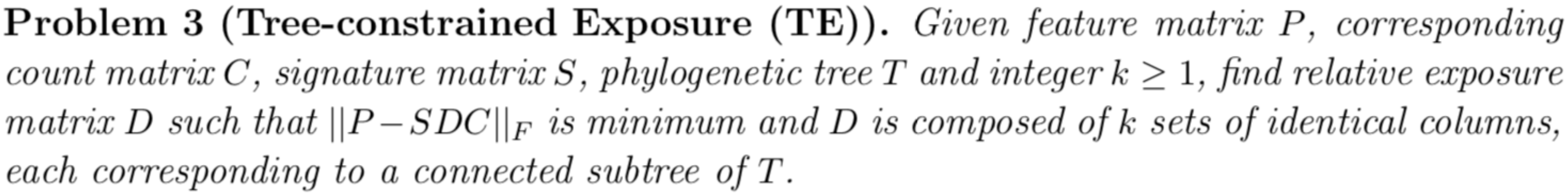
Problem 3 (Tree-constrained Exposure (TE)). *Given feature matrix P, corresponding count matrix C, signature matrix S, phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $||P - SDC||_F$ is minimum and D is composed of k sets of identical columns, each corresponding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

P: Mutation Count Matrix

$$\begin{array}{c|ccc} & △ & ☾ & □ \\ \hline \text{TC>GG} & 0 & 0 & 1 \\ \text{AC>AA} & 2 & 1 & 1 \end{array}$$

$$m \times n$$

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

$$\begin{array}{cccc} & \triangle\triangle & \text{☾} & \square\square \\ \text{TC>GG} & 0 & 0 & 1 \\ \text{AC>AA} & 2 & 1 & 1 \end{array} \approx \begin{array}{ccc} & \text{Sig. 1} & \text{Sig. 2} \\ \text{TC>GG} & 0 & .5 \\ \text{AC>AA} & 1 & .5 \end{array}$$

$$m \times n \qquad\qquad m \times r$$

Tree-constrained Exposure Problem

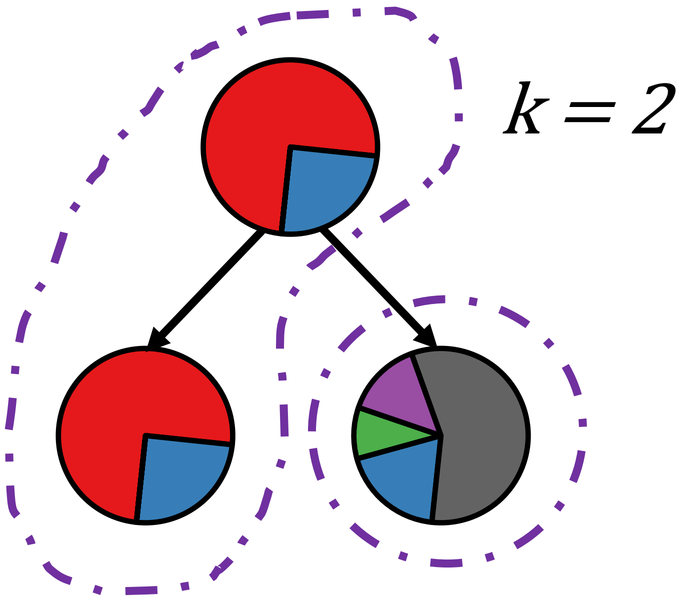
Problem 3 (Tree-constrained Exposure (TE)). *Given feature matrix P, corresponding count matrix C, signature matrix S, phylogenetic tree T and integer $k \geq 1$, find relative exposure matrix D such that $||P - SDC||_F$ is minimum and D is composed of k sets of identical columns, each corresponding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

P: Mutation Count Matrix *S*: Signature Matrix *D*: Relative Exposure Matrix *C*: Weight Matrix

$$\underset{m \times n}{\begin{array}{c|ccc} & △ & ☾ & ⧉ \\ \hline \text{TC>GG} & 0 & 0 & 1 \\ \text{AC>AA} & 2 & 1 & 1 \end{array}} \approx \underset{m \times r}{\begin{array}{c|cc} & \text{Sig. 1} & \text{Sig. 2} \\ \hline \text{TC>GG} & 0 & .5 \\ \text{AC>AA} & 1 & .5 \end{array}} \underbrace{\underset{r \times n}{\begin{array}{ccc} △ & ☾ & ⧉ \\ \hline \boxed{\begin{matrix}1 & 1 \\ 0 & 0\end{matrix}} & & \boxed{\begin{matrix}0 \\ 1\end{matrix}} \end{array}} \underset{n \times n}{\begin{array}{ccc|c} △ & ☾ & ⧉ & \\ \hline 2 & 0 & 0 & △ \\ 0 & 1 & 0 & ☾ \\ 0 & 0 & 2 & ⧉ \end{array}}}_{E\text{: Exposure Matrix}}$$

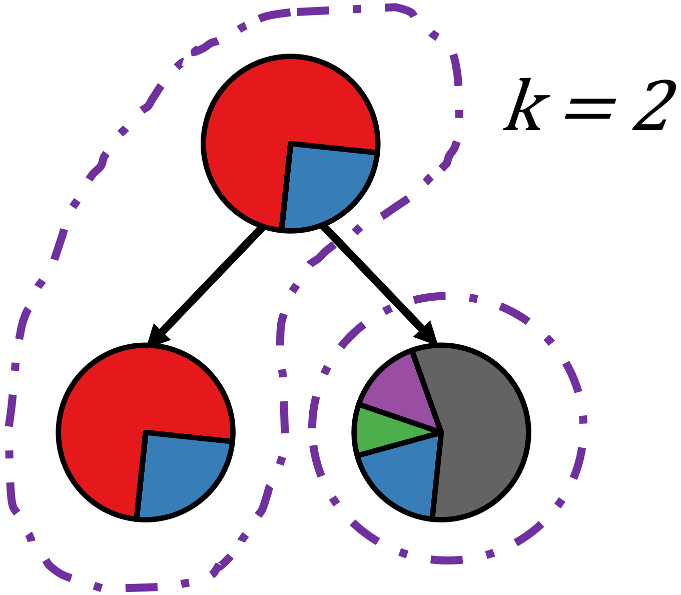
PhySigs Algorithm

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



PhySigs Algorithm

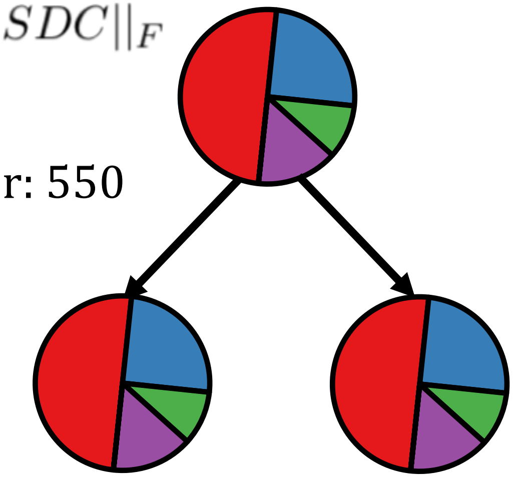
Step 1: Solve TE Problem for each possible number *k* of clusters



No Shift (k=1)

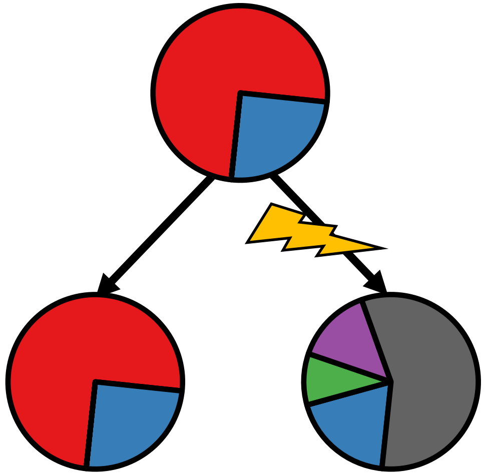
$||P - SDC||_F$

Error: 550



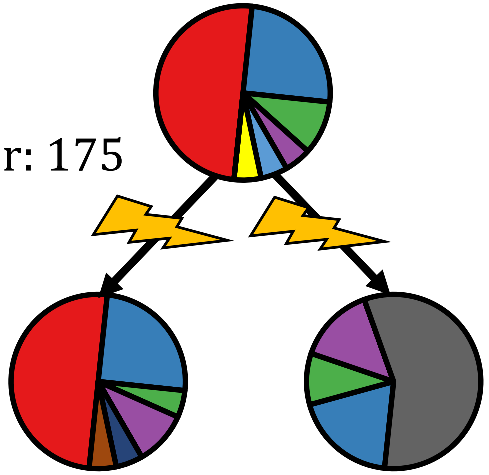
One Shift (k=2)

Error: 200



Two Shifts (k=3)

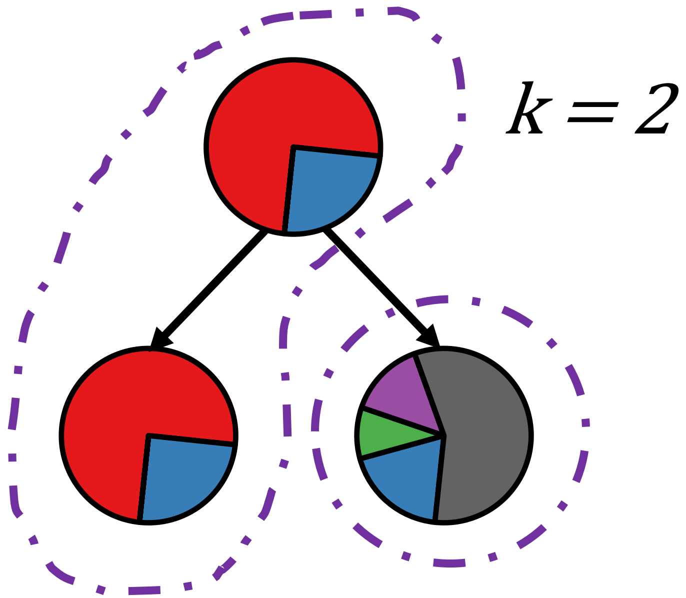
Error: 175



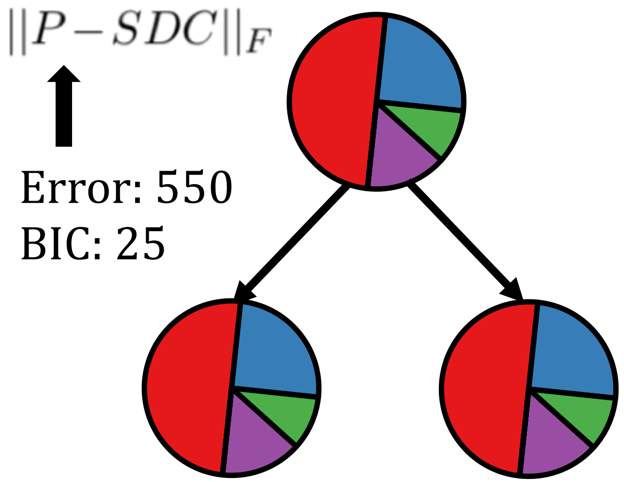
PhySigs Algorithm

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)



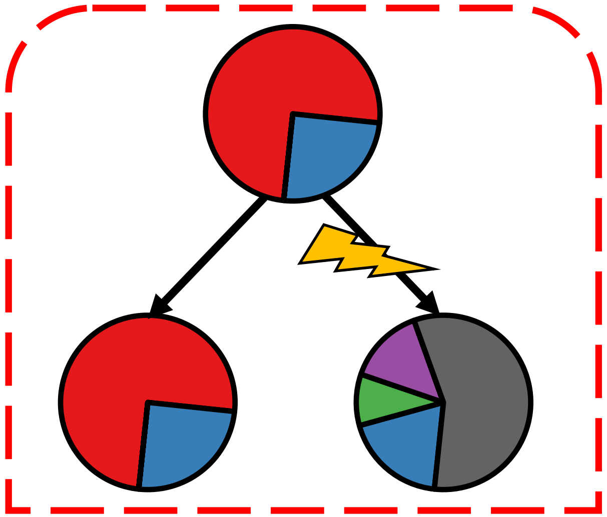
No Shift (k=1)



Error: 550
BIC: 25

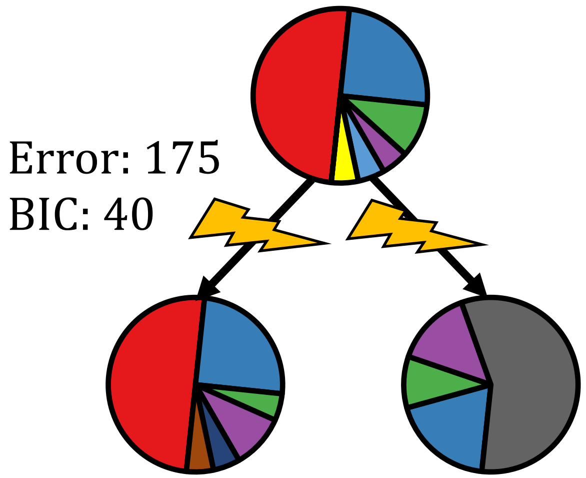
One Shift (k=2)

Error: 200
BIC: 15



Lowest BIC Selected

Two Shifts (k=3)



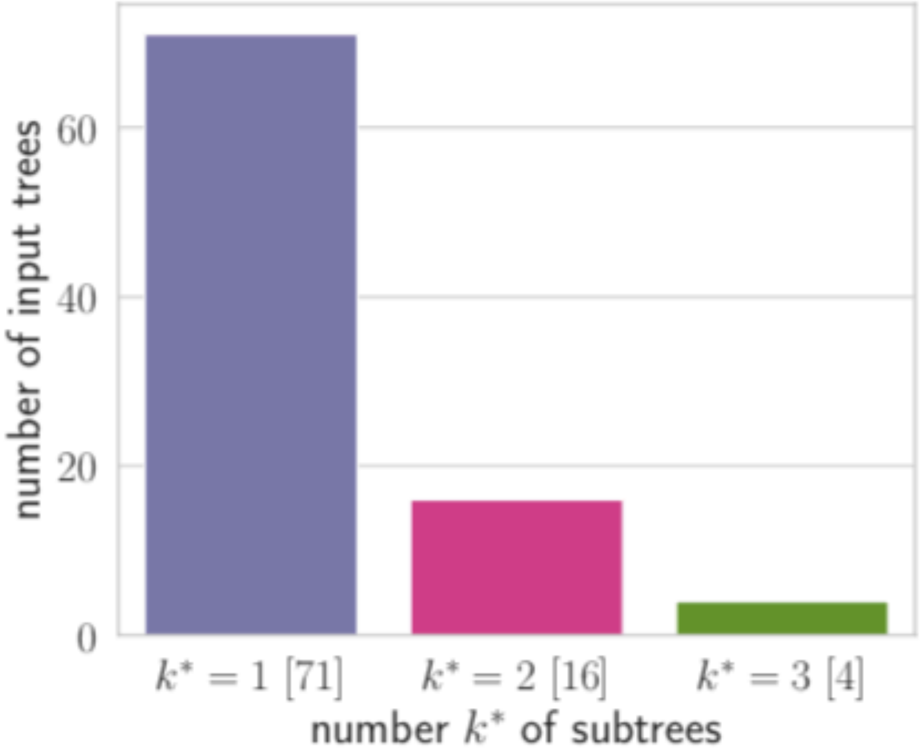
Error: 175
BIC: 40

Results

On Simulated and Biological Data

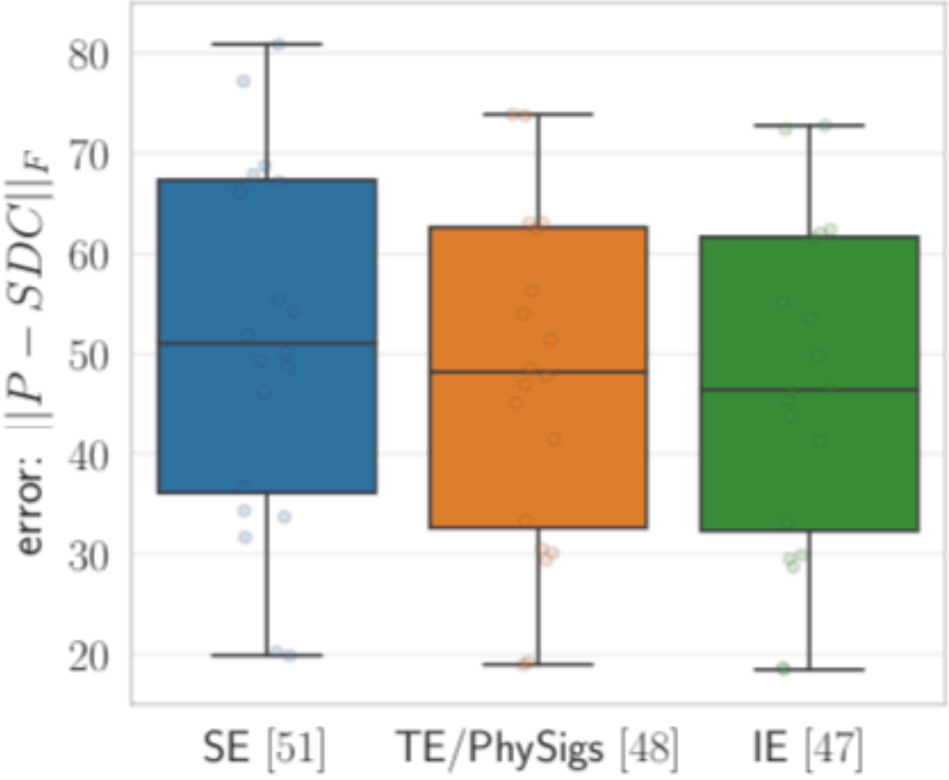
PhySigs on Lung Cancer Cohort

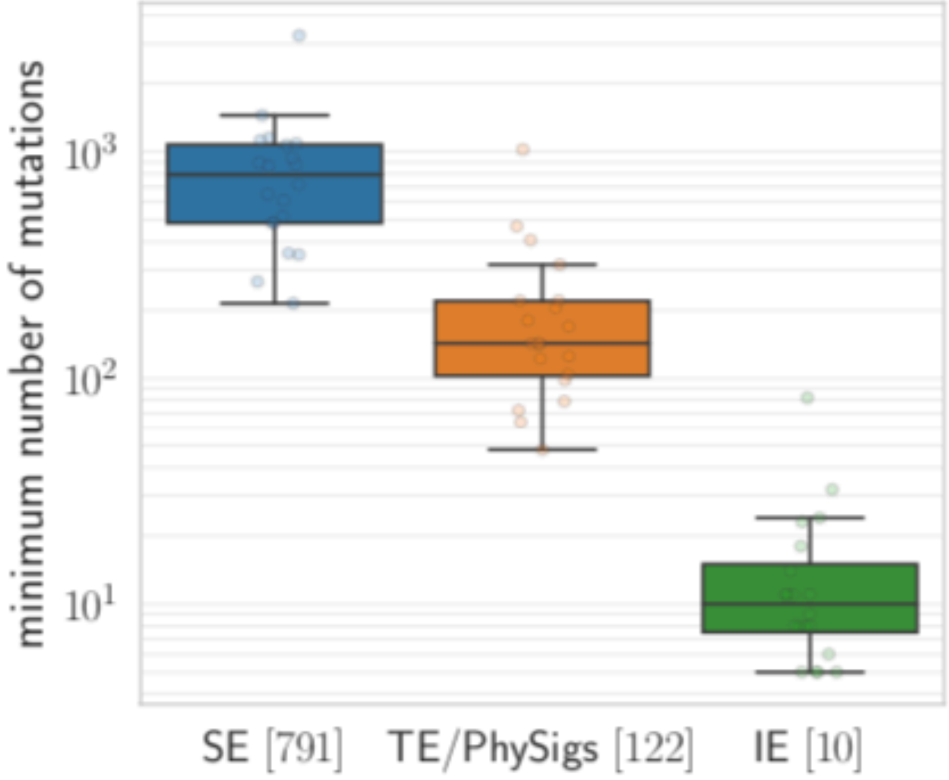
- 91 patient tumors
- Number of clones per patient ranges from 2 to 15 (median of 5).
- Number of equally likely trees per patient ranges from 1 to 17 (median of 1).



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

PhySigs Explains Data without Overfitting

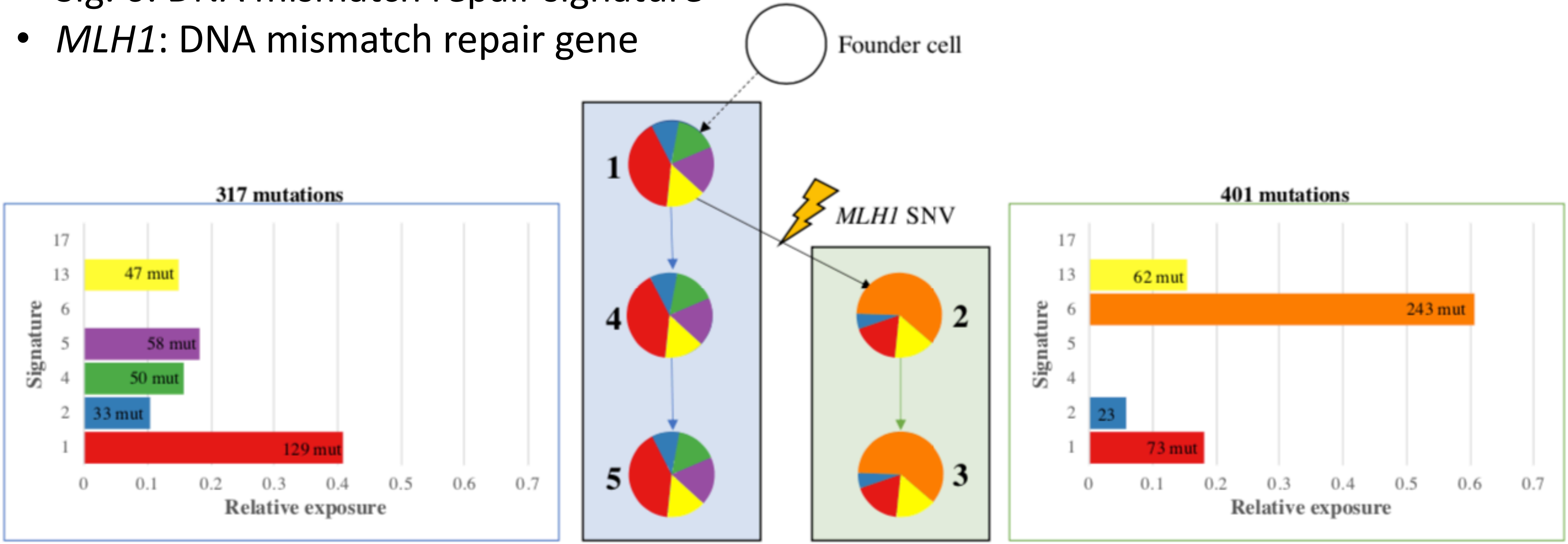




Data from [Jamal-Hanjani et al., 2017]

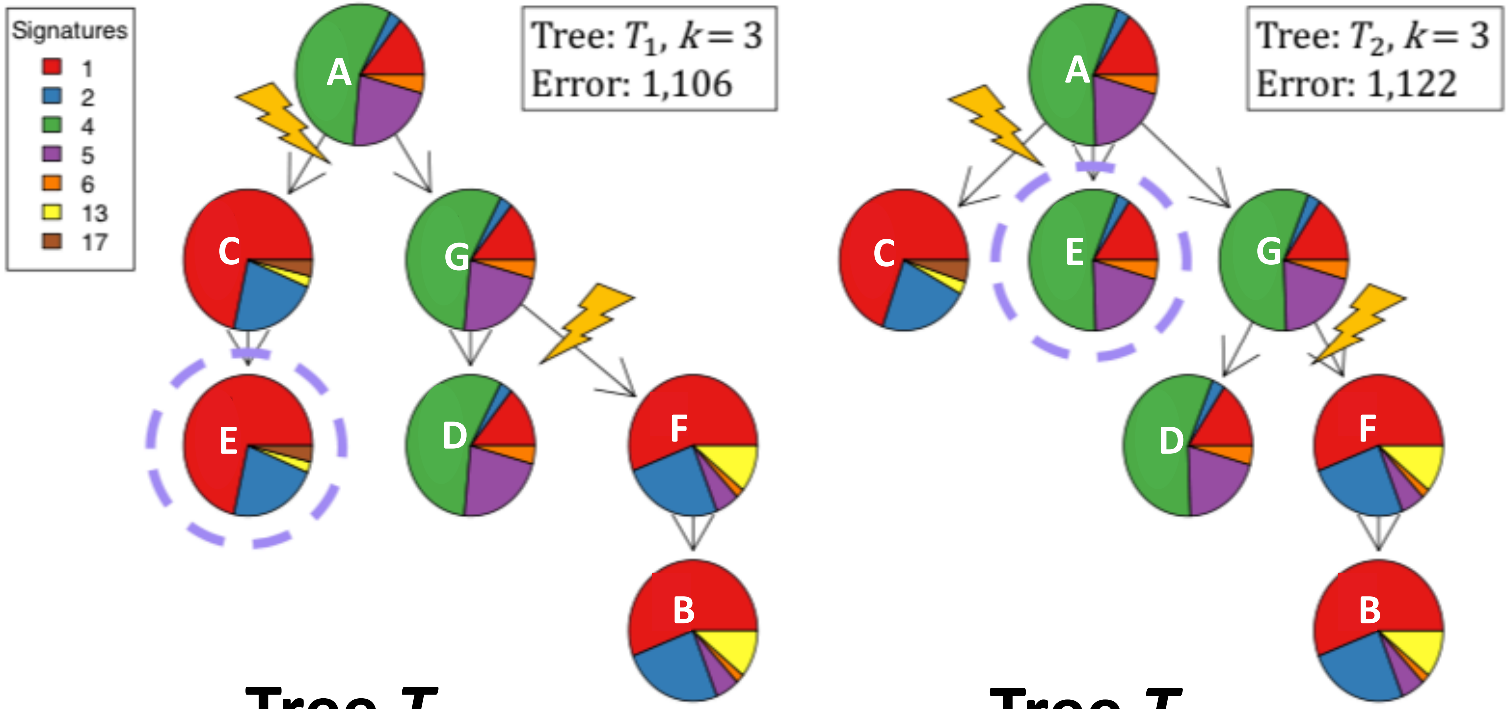
PhySigs Finds Explainable Shift for Patient CRUK0064

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



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

PhySigs Constrains Solutions for Patient CRUK0025



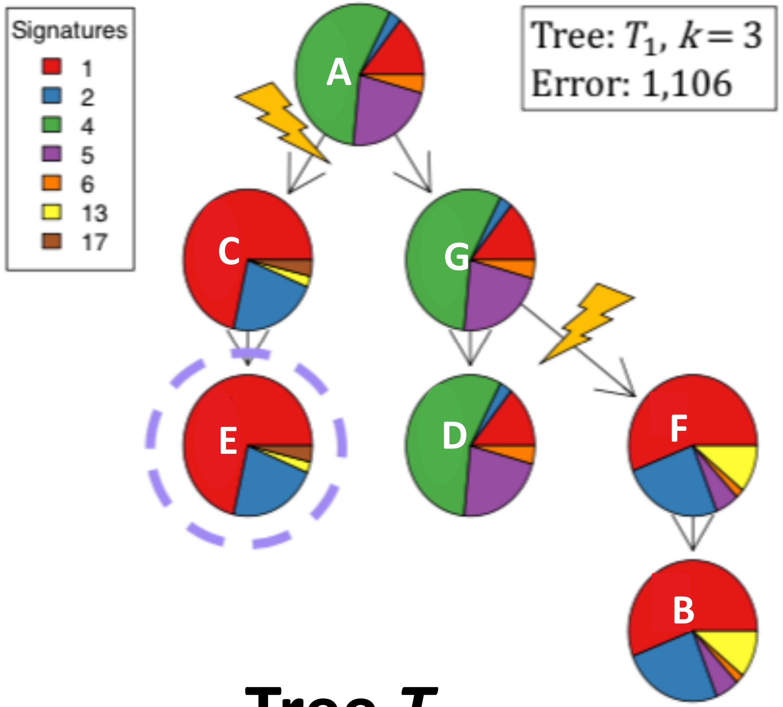
Tree T_1

- 2 Shifts
- Small Error

Tree T_2

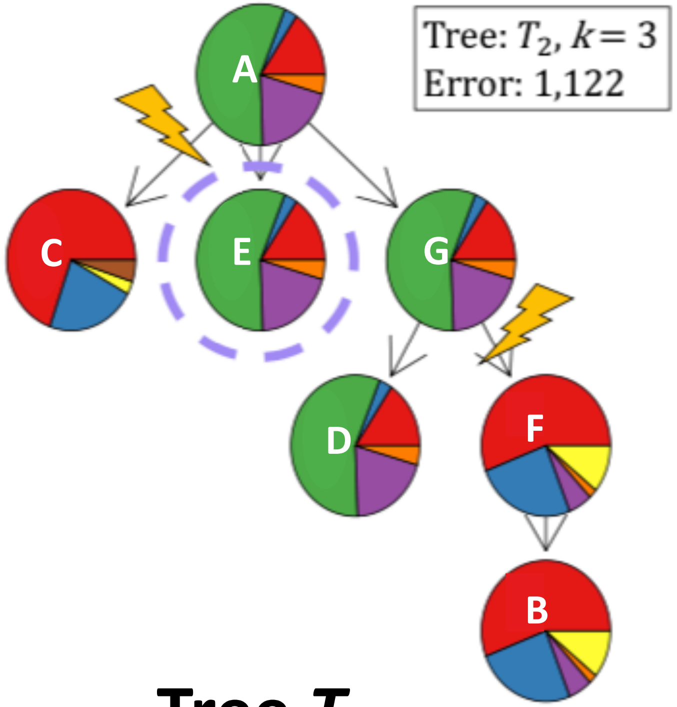
- 2 Shifts
- Big Error

PhySigs Constrains Solutions for Patient CRUK0025



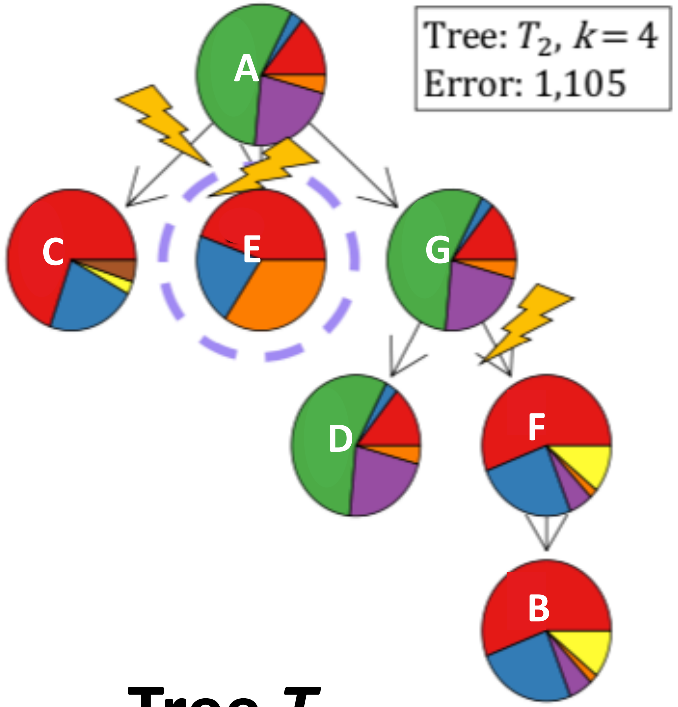
Tree T_1

- 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)