PhyDOSE: Design of Follow-up Single-cell Sequencing Experiments of Tumors

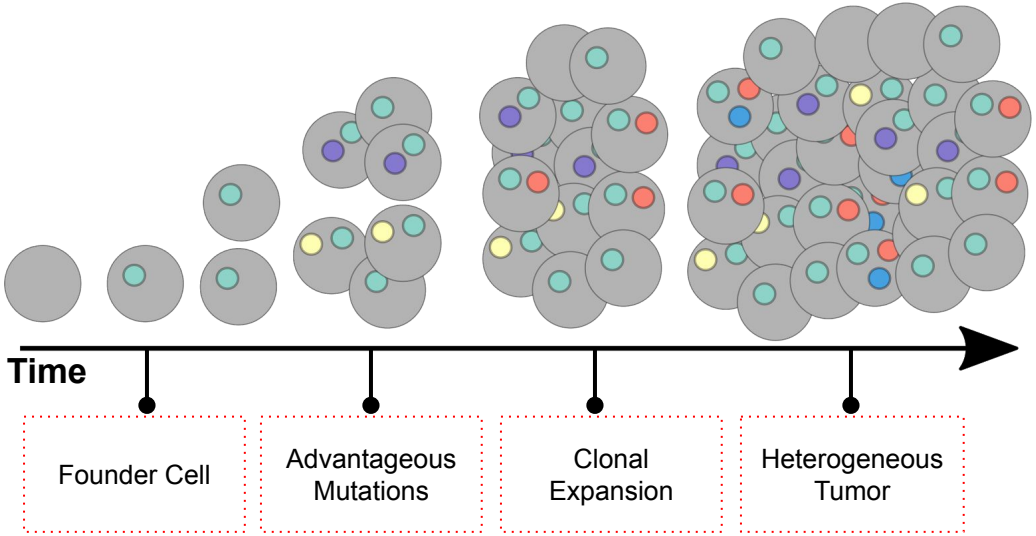
Leah Weber[1]*, Nuraini Aguse[1]*, Nicholas Chia[2,3] and Mohammed El-Kebir[1]

1 University of Illinois at Urbana-Champaign, Department of Computer Science
2 Microbiome Program, Center for Individualized Medicine, Mayo Clinic
3 Division of Surgical Research, Department of Surgery, Mayo Clinic

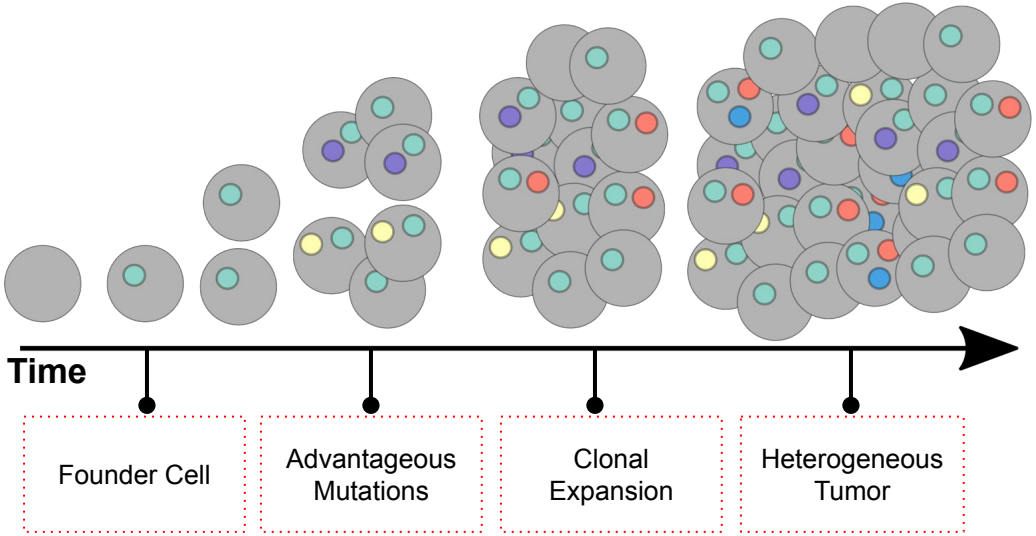
RECOMB-CCB 2020
June 18, 2020

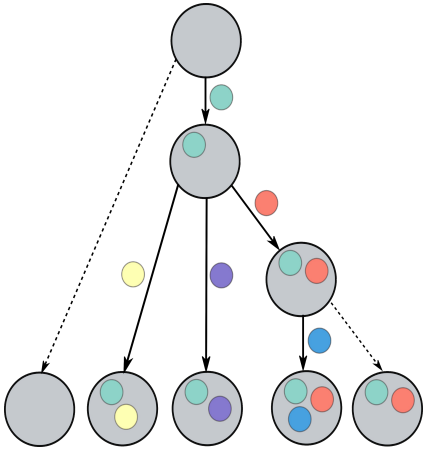
*These authors contributed equally to this work

Cancer is an evolutionary process



Cancer is an evolutionary process





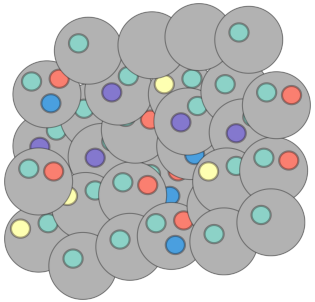
Phylogenetic Tree

Identify treatment targets	Understand metastatic development	Compare evolutionary patterns across patients

DNA sequencing of tumors

Bulk DNA Sequencing ($)

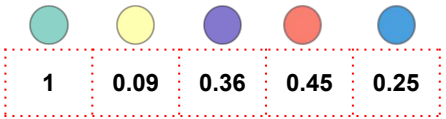
Single-cell DNA Sequencing ($$$)

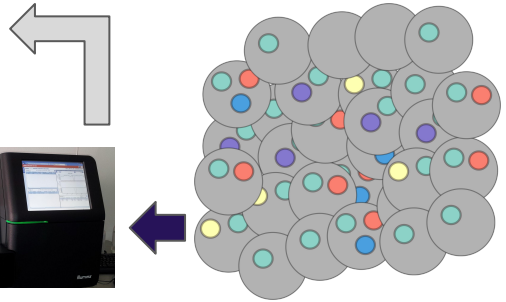


DNA sequencing of tumors

Bulk DNA Sequencing ($)

Cancer Cell Fractions



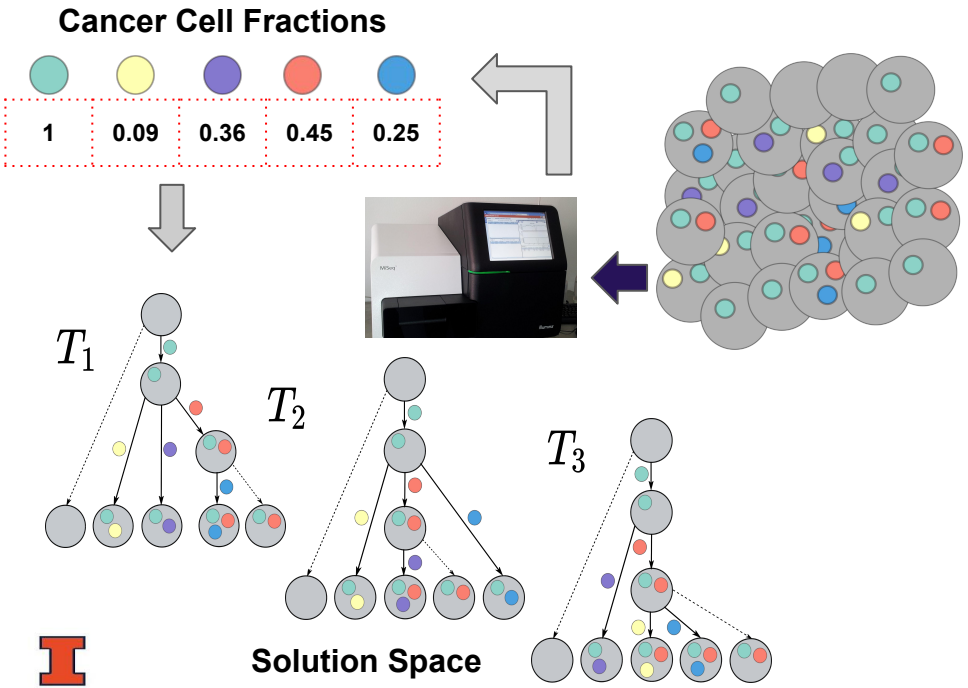


Single-cell DNA Sequencing ($$$)

DNA sequencing of tumors

Bulk DNA Sequencing ($)

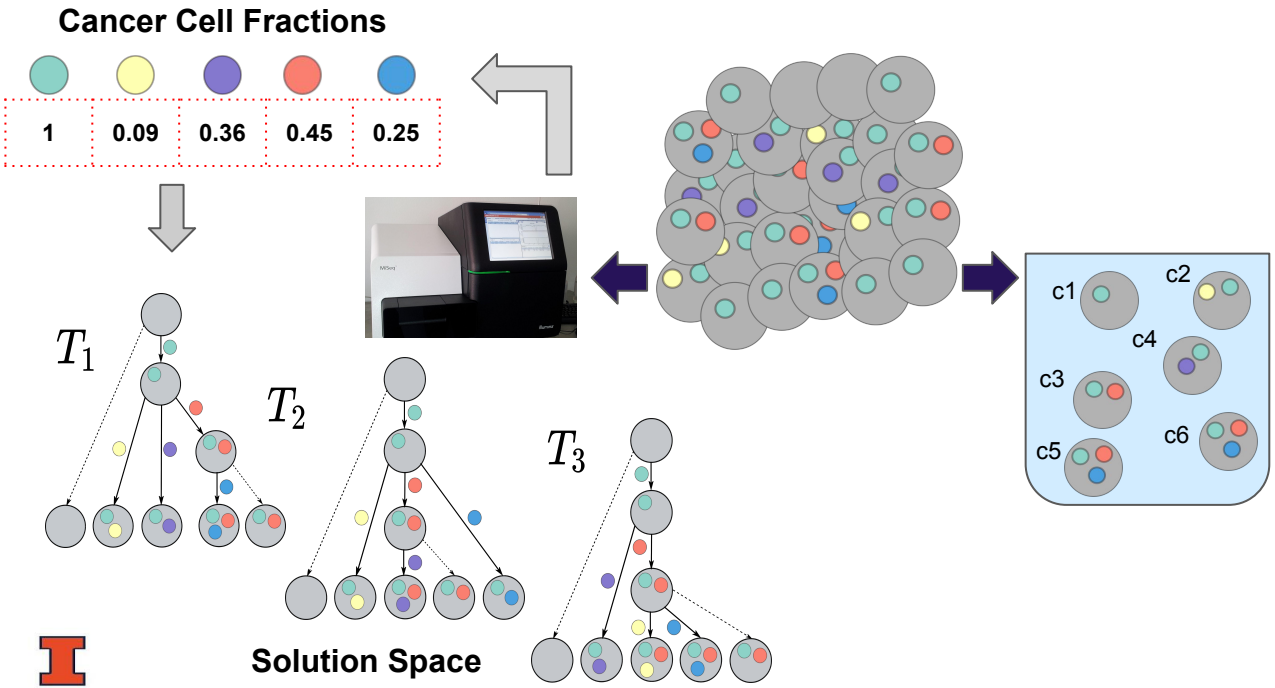
Single-cell DNA Sequencing ($$$)



DNA sequencing of tumors

Bulk DNA Sequencing ($)

Single-cell DNA Sequencing ($$$)



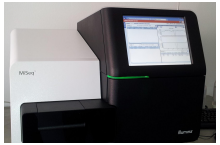
DNA sequencing of tumors

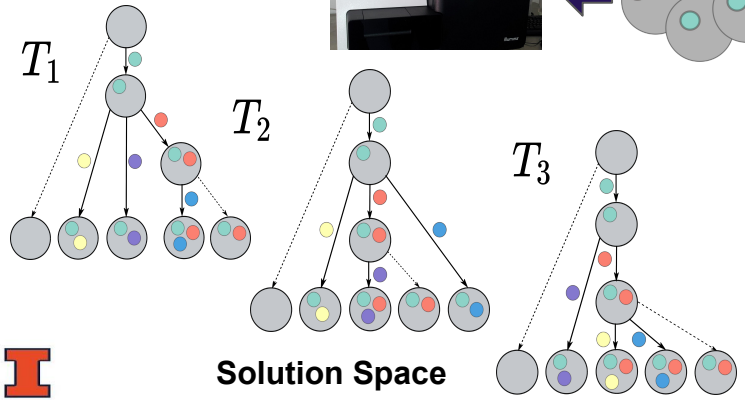
Bulk DNA Sequencing ($)

Single-cell DNA Sequencing ($$$)

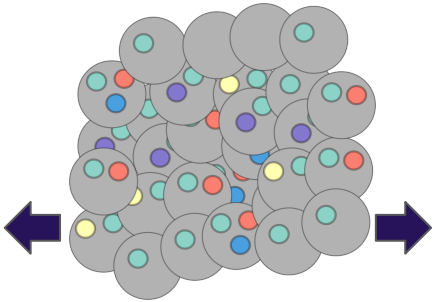
Cancer Cell Fractions

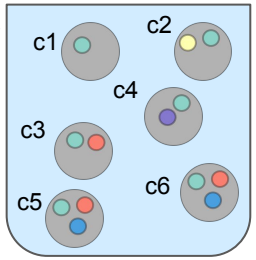
1	0.09	0.36	0.45	0.25





Solution Space





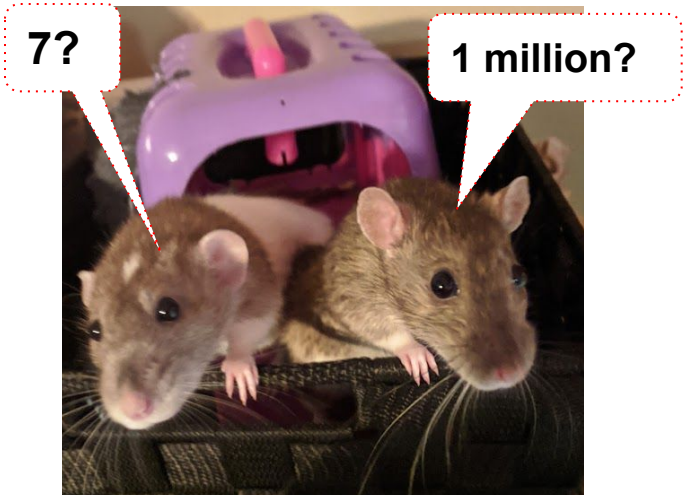
c1	1	0	0	0	0
c2	1	1	0	0	0
c3	0	0	0	1	0
c4	1	0	0	1	0
c5	1	?	0	1	1
c6	1	0	0	1	0

0	False Negative

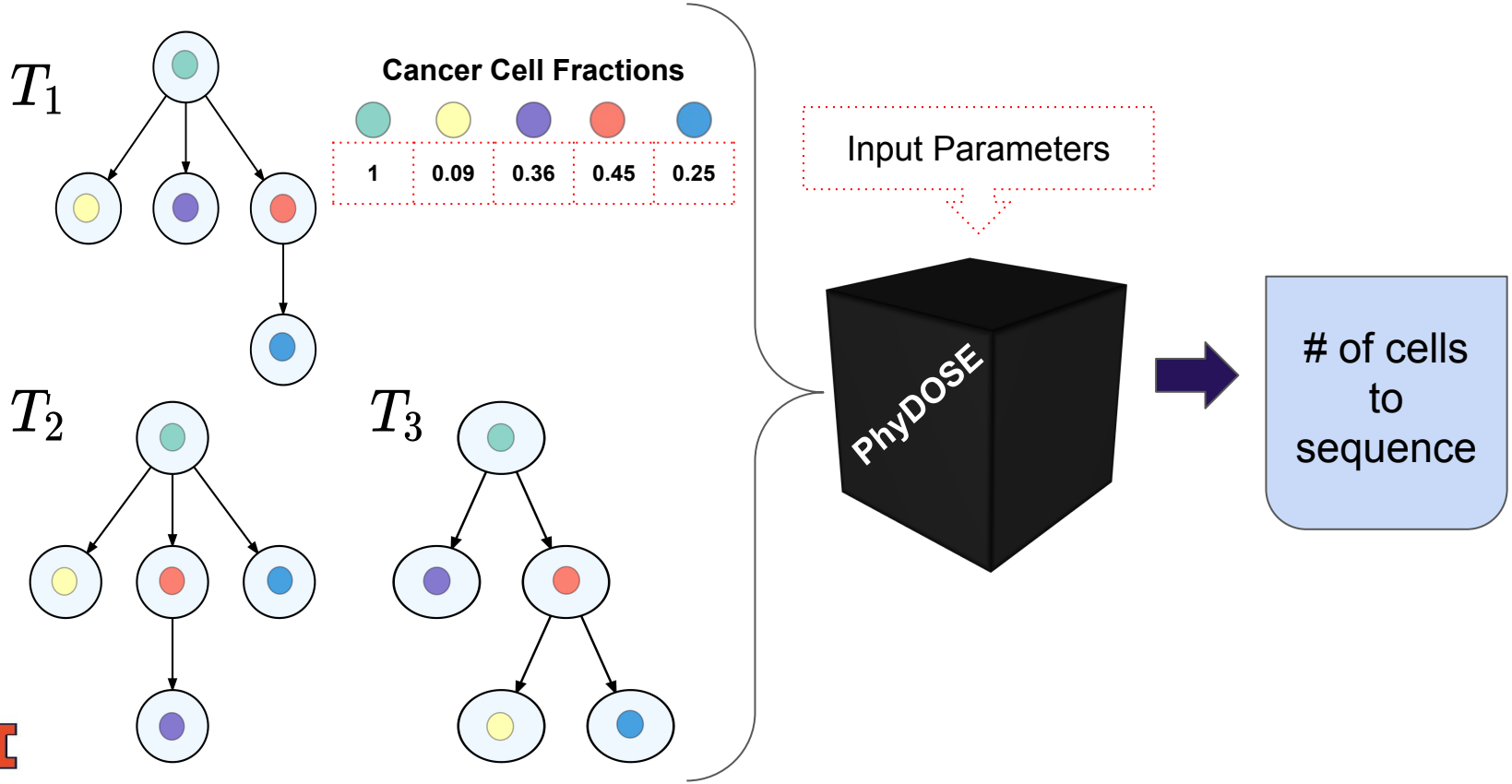
Phylogeny inference from DNA sequencing

Method	Bulk Sequencing Data	Single-cell Data
SCITE [Jahn et al., 2016]		X
OncoNEM [Ross & Markowetz, 2017]		X
SPhyR [El-Kebir, 2018]		X
SiCloneFit [Zafar et al., 2019]		X
PhiSCS [Malikic et al., 2019a]	X	X
B-SCITE [Malikic et al. 2019b]	X	X

How many single-cells should you sequence to minimize costs?

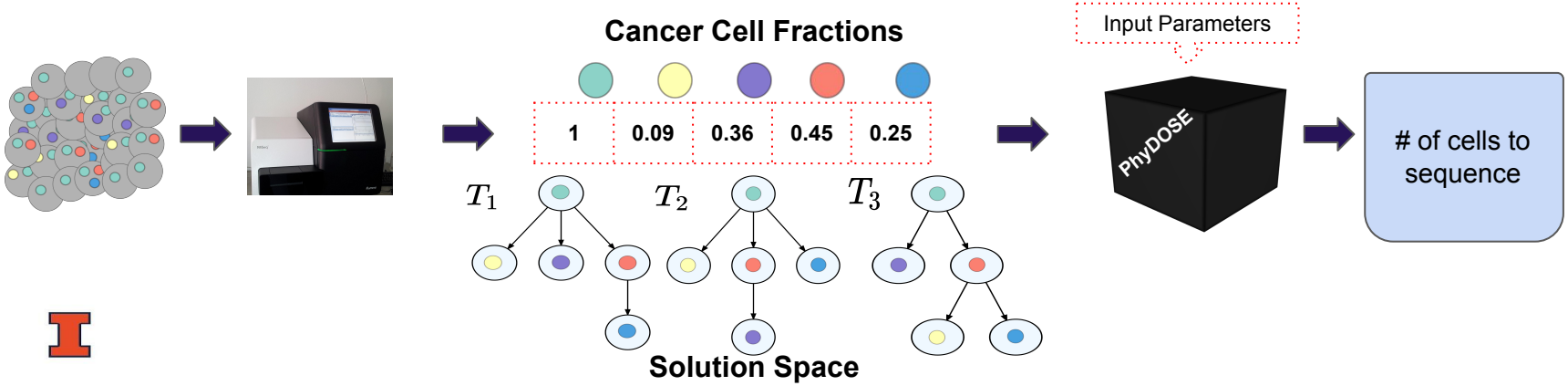


Key idea: Design a cost-effective single-cell sequencing experiment using bulk DNA data

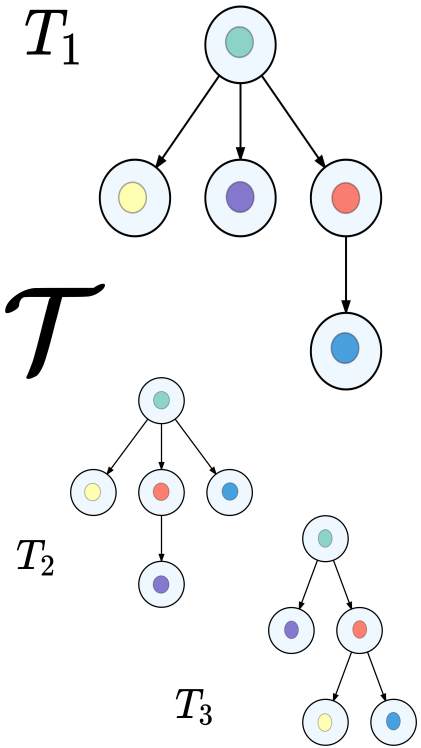


Outline

- Problem statement
- Methods
- Complexity
- Simulation study
- Application to real data
- Conclusions and future work



Key idea: Bulk data guides cost effective single-cell experiment design



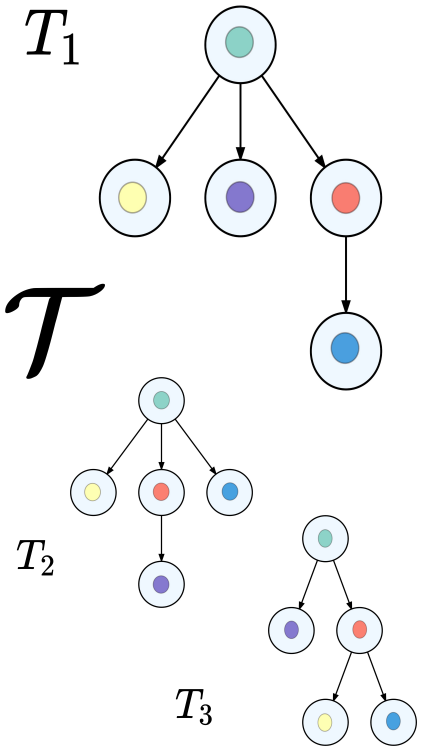
SINGLE-CELL SEQUENCING POWER CALCULATION (SCS-PC)

Given a set $\mathcal{T}$ of candidate phylogenies, frequencies **f**

Cancer Cell Fractions f

1	0.09	0.36	0.45	0.25

Key idea: Bulk data guides cost effective single-cell experiment design



SINGLE-CELL SEQUENCING POWER CALCULATION (SCS-PC)

Given a set $\mathcal{T}$ of candidate phylogenies, frequencies $\mathbf{f}$ and confidence level γ,

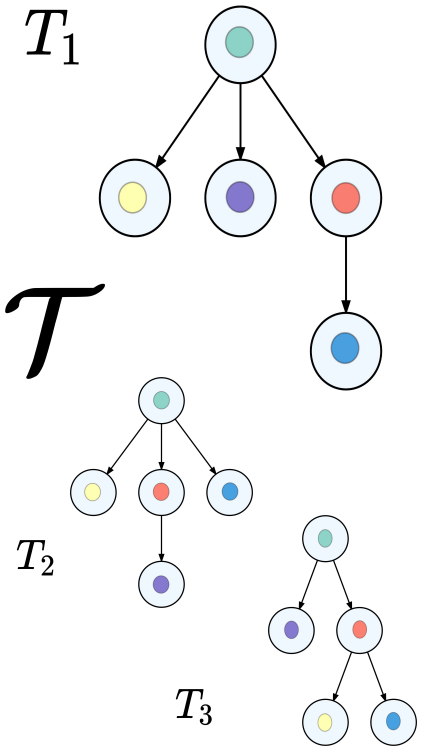
Cancer Cell Fractions $\mathbf{f}$

1	0.09	0.36	0.45	0.25

Confidence Level

$$\gamma = 0.95$$

Key idea: Bulk data guides cost effective single-cell experiment design



Single-cell Sequencing Power Calculation (SCS-PC)

Given a set $\mathcal{T}$ of candidate phylogenies, frequencies $\mathbf{f}$ and confidence level γ, find the minimum number k^* of single cells needed to determine the true phylogeny T among $\mathcal{T}$ with probability at least γ.

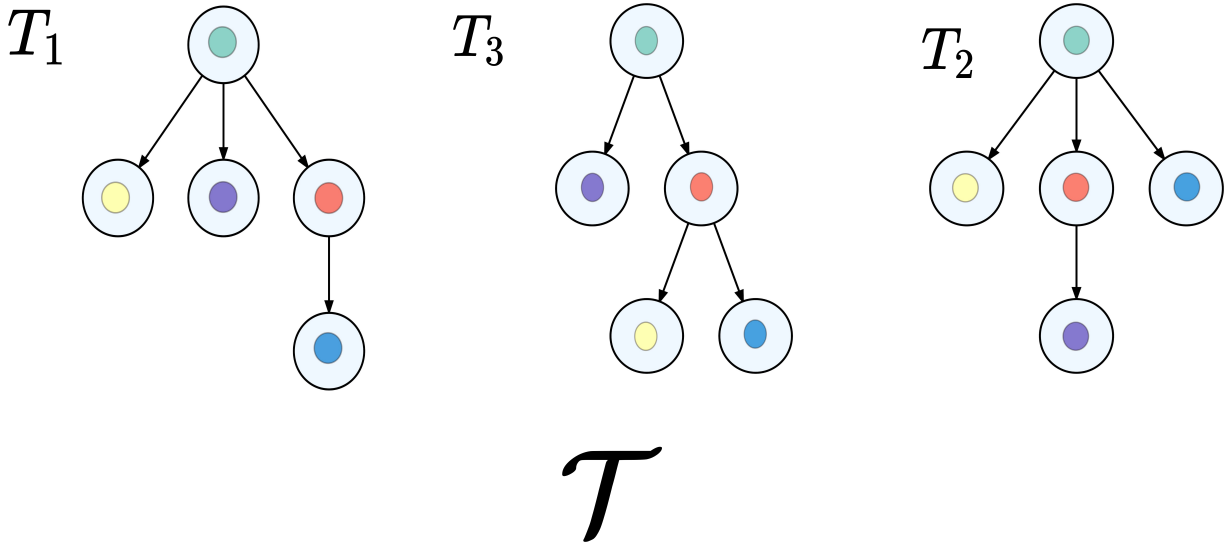
Cancer Cell Fractions $\mathbf{f}$

1	0.09	0.36	0.45	0.25

Confidence Level

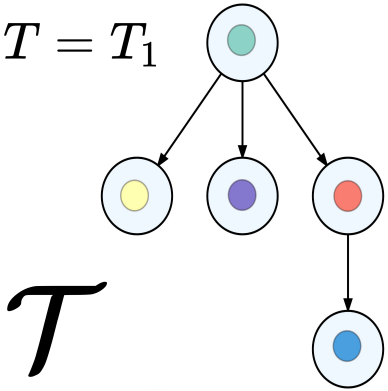
$\gamma = 0.95$ → k^*

Solving the SCS-PC



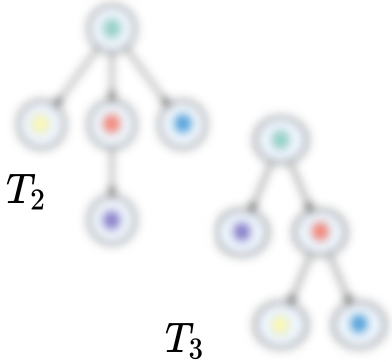
True phylogeny unknown

Key idea: condition on each tree being the true tree and solve SCS-PC



SCS Power Calculation for Phylogeny T (T-SCS-PC)

Given a set $\mathcal{T}$ of candidate phylogenies and a phylogeny $T \in \mathcal{T}$, frequencies $\mathbf{f}$ and confidence level γ,



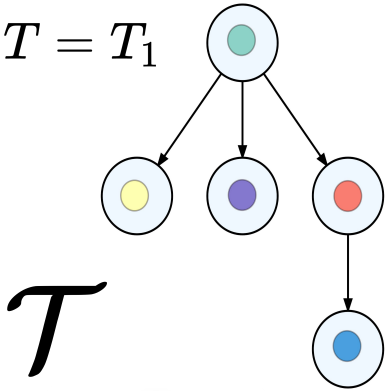
Cancer Cell Fractions $\mathbf{f}$

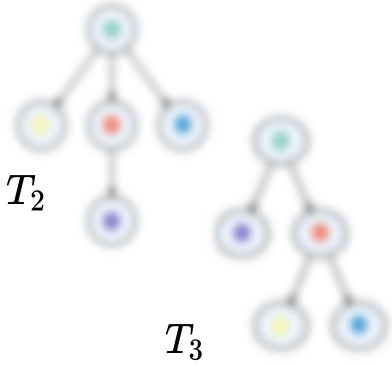
1	0.09	0.36	0.45	0.25

Confidence Level

$$\gamma = 0.95$$

Key idea: condition on each tree being the true tree and solve SCS-PC





SCS Power Calculation for Phylogeny T (T-SCS-PC)

Given a set $\mathcal{T}$ of candidate phylogenies and a phylogeny $T \in \mathcal{T}$, frequencies $\mathbf{f}$ and confidence level γ, find the minimum number k^* of single cells needed such that the probability of a successful SCS experiment is greater than or equal to γ.

Cancer Cell Fractions $\mathbf{f}$

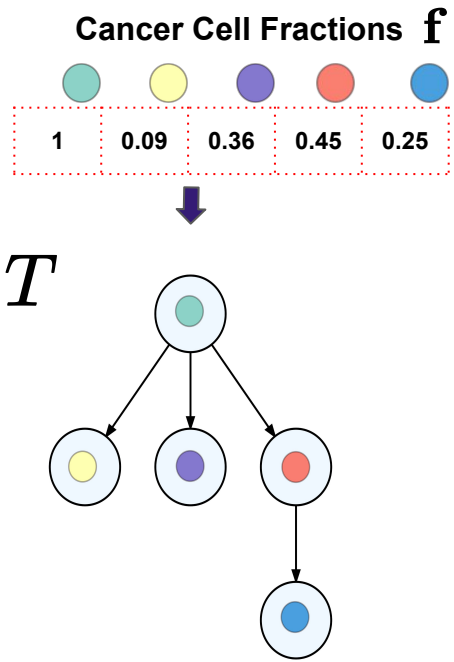
1	0.09	0.36	0.45	0.25

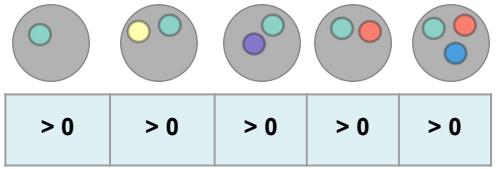
Confidence Level

$\gamma = 0.95$ → k^*

$$k^* = \arg\min_{k} P(Success \mid T, \mathcal{T}, k, \mathbf{f}) \geq \gamma$$

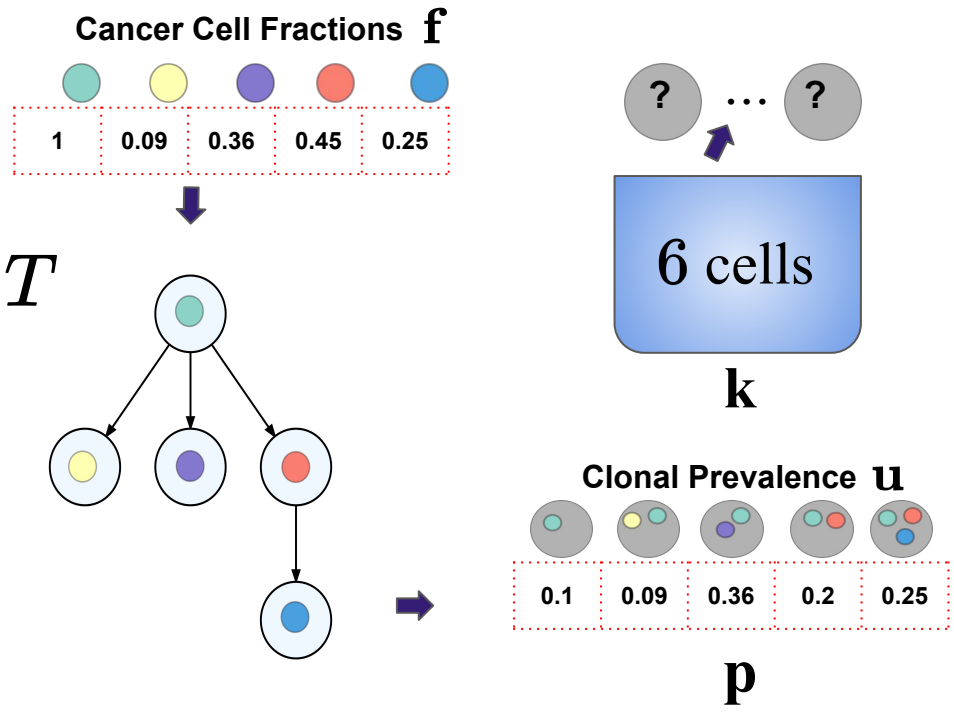
What is a successful experiment given T?

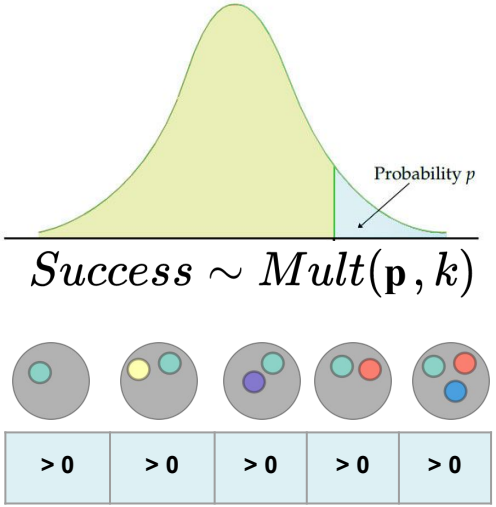




SCOPIT
[Davis et al. 2019]

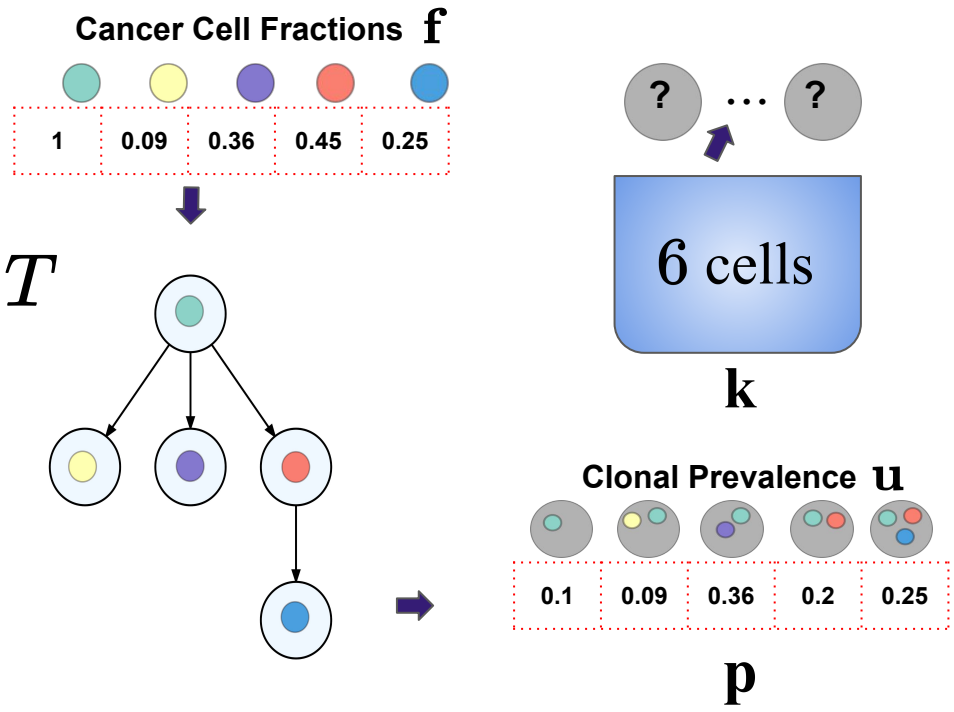
What is a successful experiment given T?

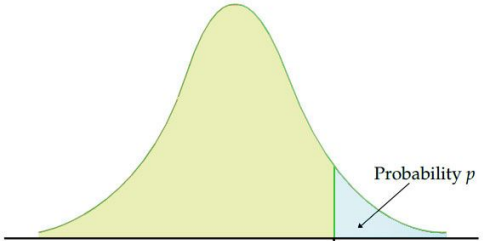




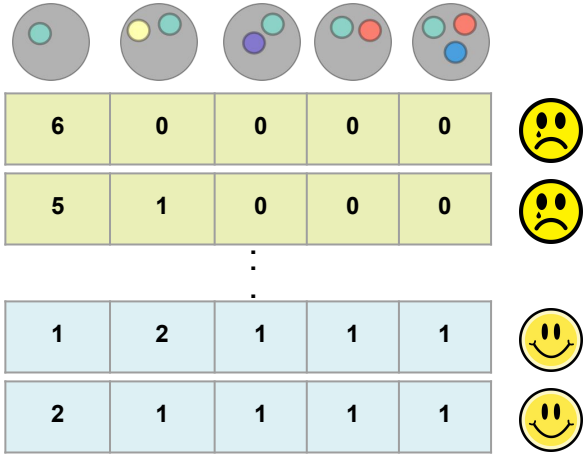
SCOPIT
[Davis et al. 2019]

What is a successful experiment given T?



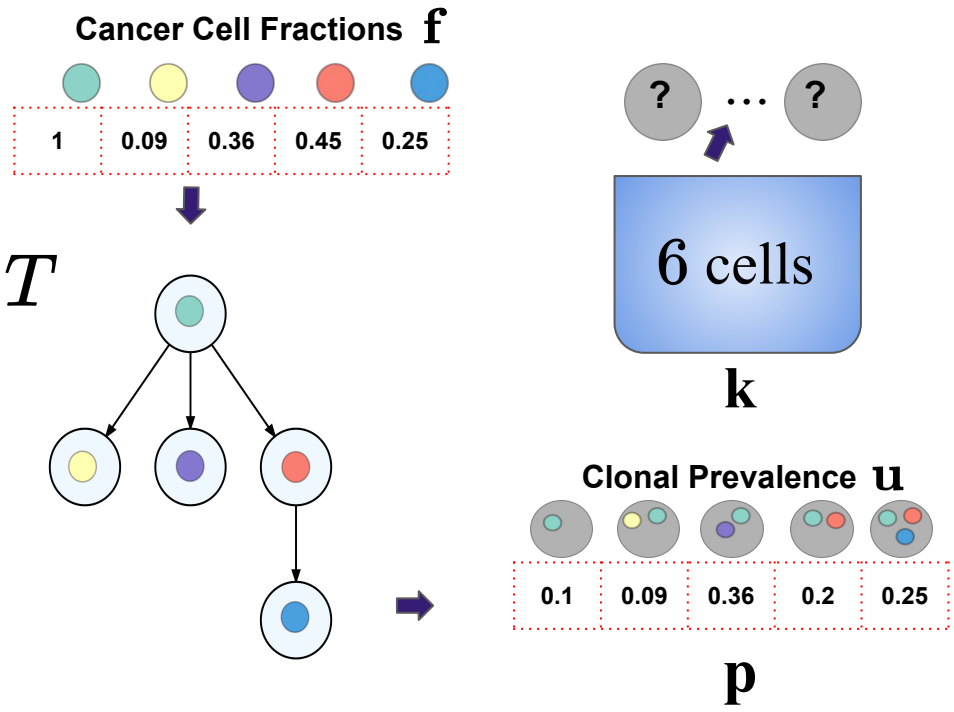


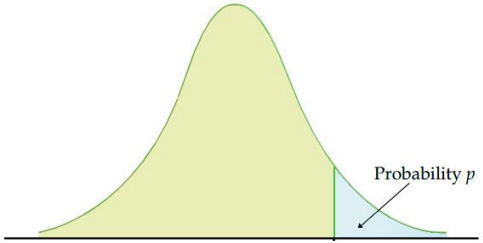
$$Success \sim Mult(\mathbf{p}, k)$$



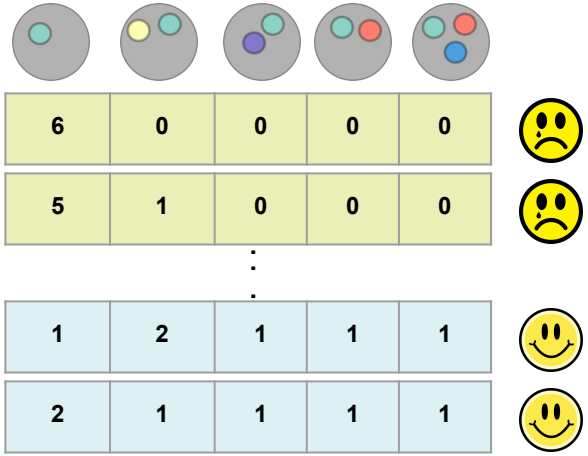
SCOPIT
[Davis et al. 2019]

What is a successful experiment given T?





$$Success \sim Mult(\mathbf{p}, k)$$

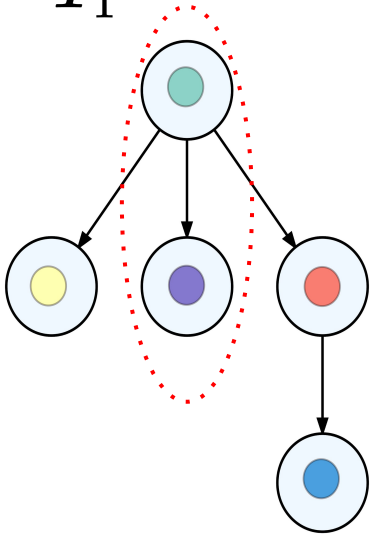


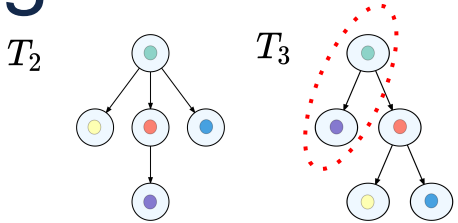
But we don't always need to observe all clones for a successful experiment!

SCOPIT
[Davis et al. 2019]

Key idea: distinguishing feature

$T = T_1$

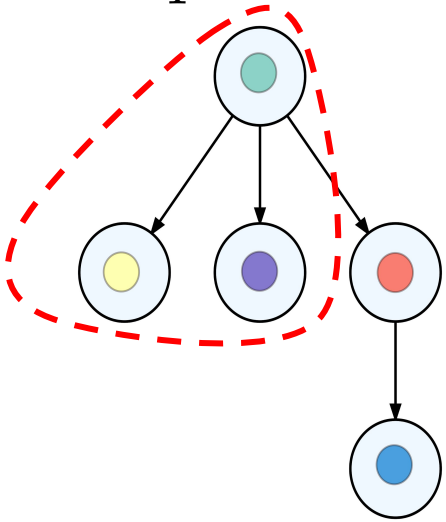




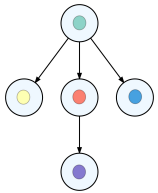
Featurette /clone in T	T_2	T_3
	+	+
	+	+
	-	+
	+	-
	-	+

Key idea: distinguishing feature

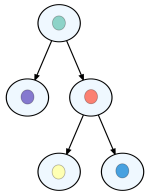
$T = T_1$



T_2



T_3



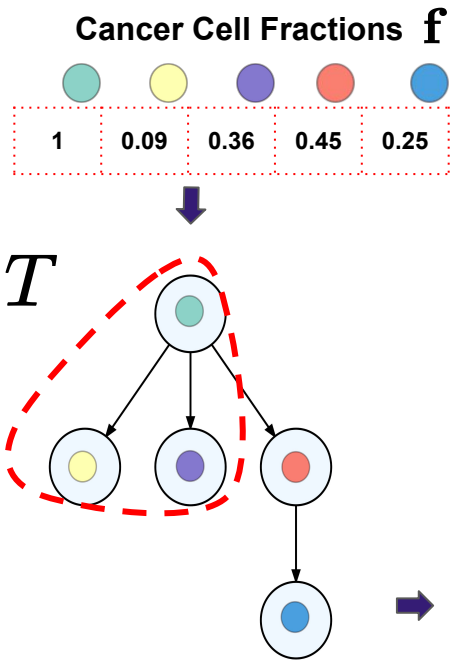
Featurette /clone in T

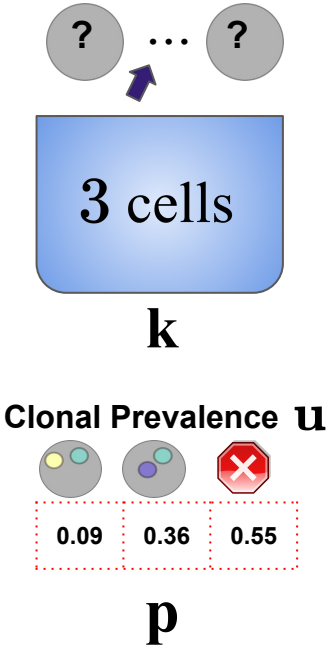
	+	+
	+	+
	-	+
	+	-
	-	+

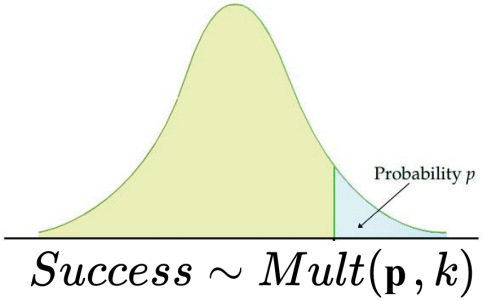
Distinguishing feature

Success is defined as observing a distinguishing feature.

Probabilistic model







0	0	3
0	1	2
⋮		
1	1	1
1	2	0
2	1	0

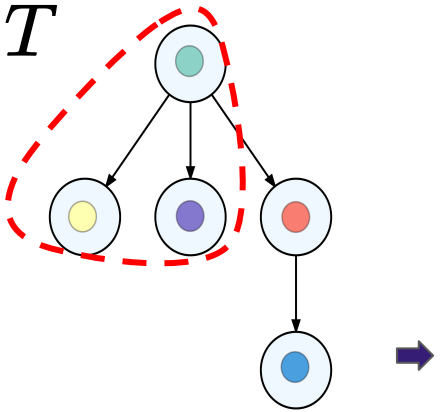
Success is defined as observing a distinguishing feature.

Probabilistic model

Cancer Cell Fractions $\mathbf{f}$

1	0.09	0.36	0.45	0.25

T



? … ?

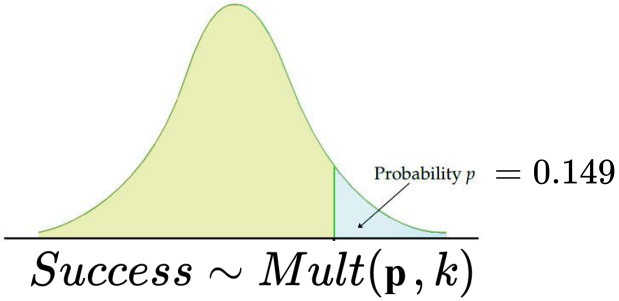
3 cells

$\mathbf{k}$

Clonal Prevalence $\mathbf{u}$

0.09	0.36	0.55

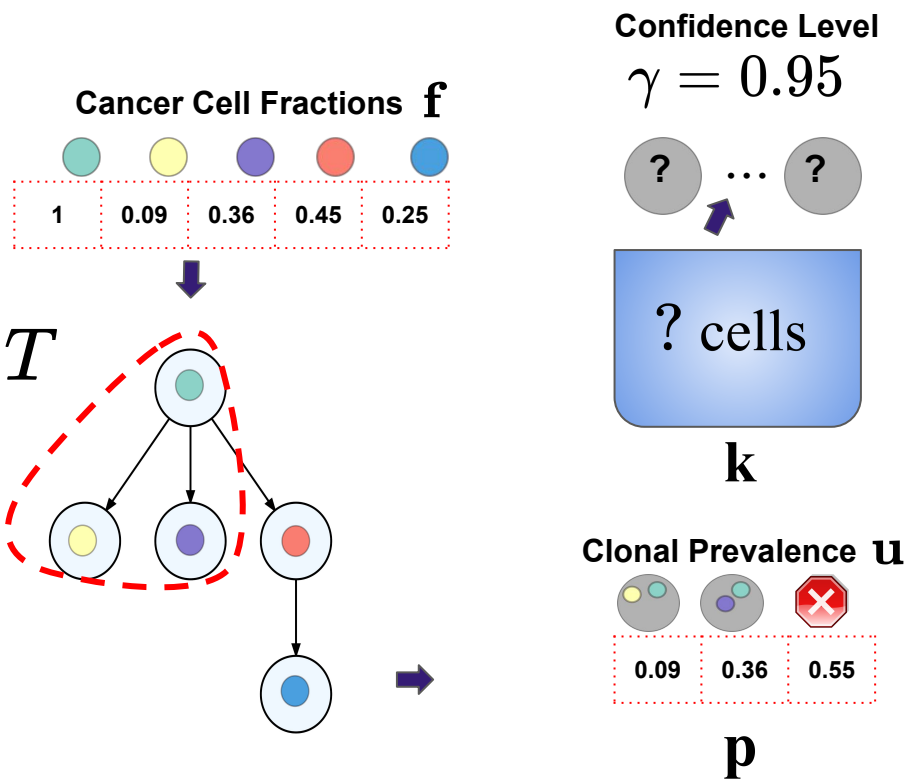
$\mathbf{p}$



			prob.
0	0	3	0.0
0	1	2	0.0
⋮			
1	1	1	0.11
1	2	0	0.03
2	1	0	0.009

Success is defined as observing a distinguishing feature.

Power calculation for fixed tree T

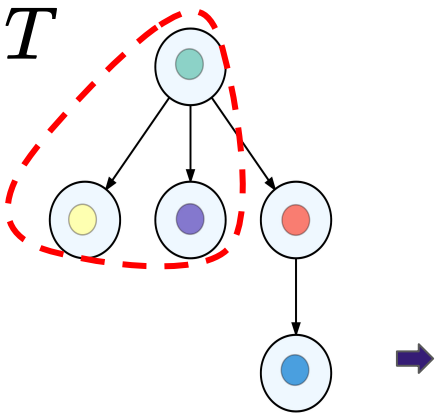


$$k^* = \arg\min_k P(Success \mid T, \mathcal{T}, k, \mathbf{f}) \geq \gamma$$

Power calculation for fixed tree T

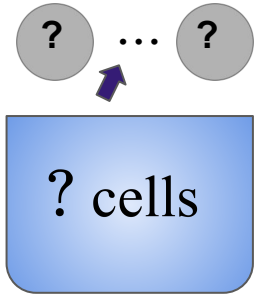
Cancer Cell Fractions $\mathbf{f}$

1	0.09	0.36	0.45	0.25



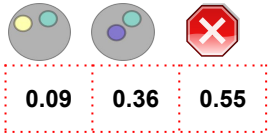
Confidence Level

$$\gamma = 0.95$$

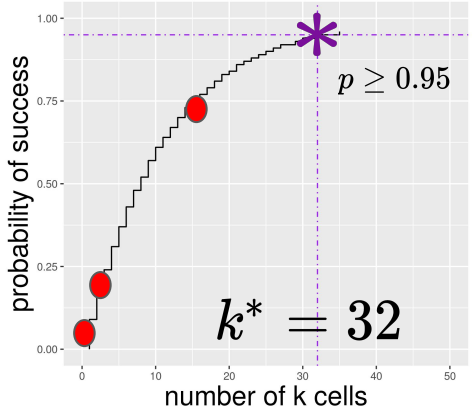


$\mathbf{k}$

Clonal Prevalence $\mathbf{u}$



$\mathbf{p}$



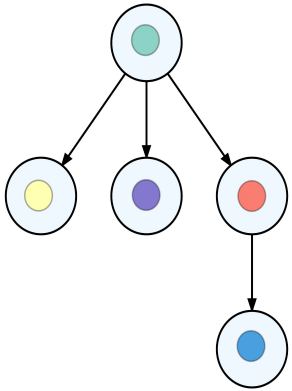
$$k^* = \arg\min_k P(Success \mid T, \mathcal{T}, k, \mathbf{f}) \geq \gamma$$

k	prob.
3	0.15
4	0.25
⋮	
15	0.75
⋮	
32	0.95

k*= 32 is the solution to the T-SCS-PC problem.

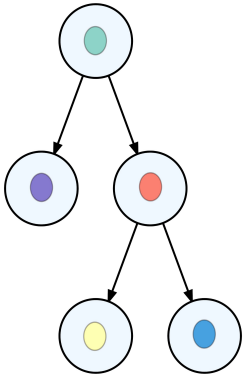
Solving the SCS-PC

T_1



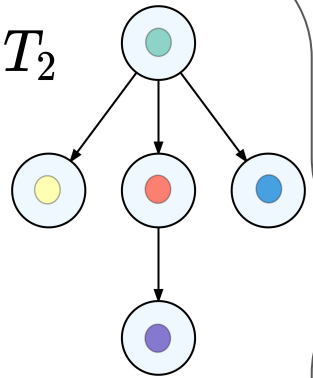
$k^* = 32$

T_3



$k^* = 32$

T_2



$k^* = 4$

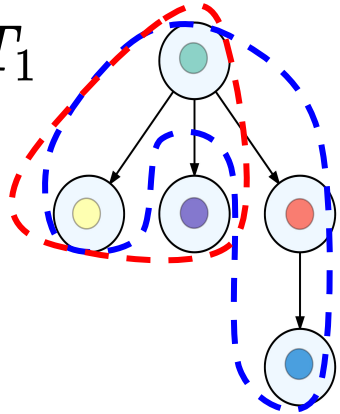
Taking the maximum yields and upper bound

$k^* = 32$

k*= 32 is the solution to the SCS-PC problem.

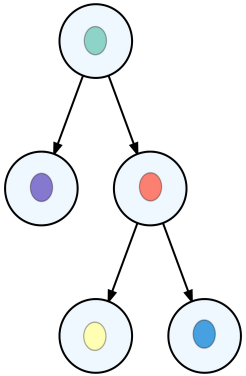
Solving the SCS-PC

T_1



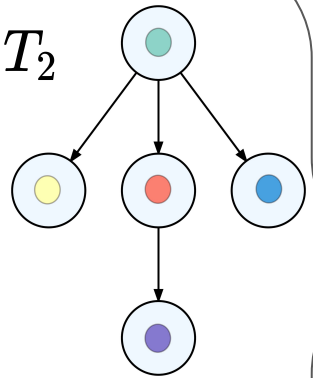
$k^* = 32$

T_3



$k^* = 32$

T_2



$k^* = 4$

Taking the maximum yields and upper bound

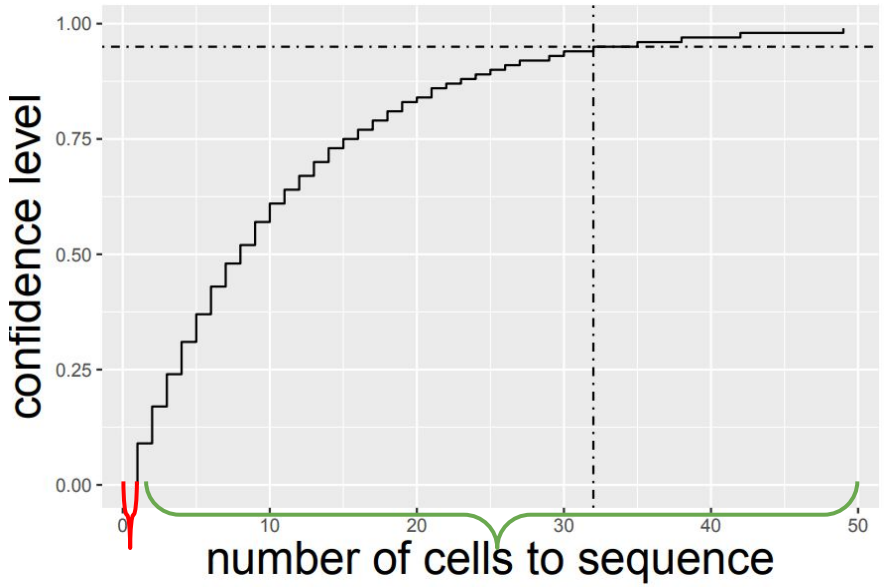
$k^* = 32$

Account for multiple distinguishing features

Adjust for false negatives

k*= 32 is the solution to the SCS-PC problem.

T-SCS-PC is NP-hard by reduction from Set Cover



Probability of success is 0

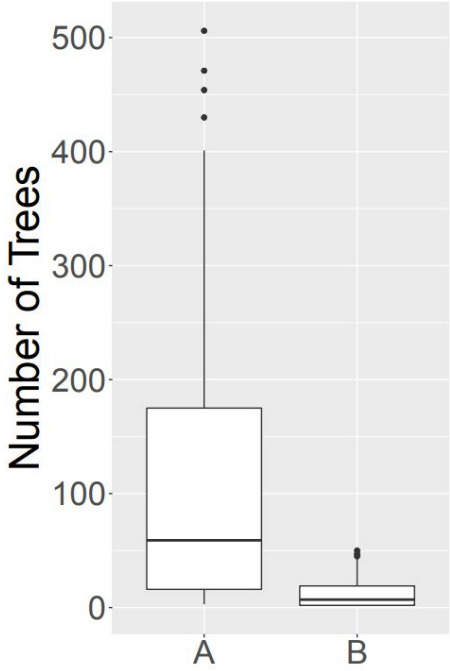
Probability of success is > 0

Set Cover ⇔ T-SCS-PC

Lemma: Let $(\mathcal{T}, T_0, \mathbf{f}, \gamma = \epsilon)$ be the *T*-SCS-PC instance corresponding to Set Cover instance $(U, \mathcal{F})$. A minimum cover has size k^* if and only if k^* is the smallest integer such that

$\cdot \Pr(Y_{k^*} \mid u(T_0, \mathbf{f})) \geq \gamma$

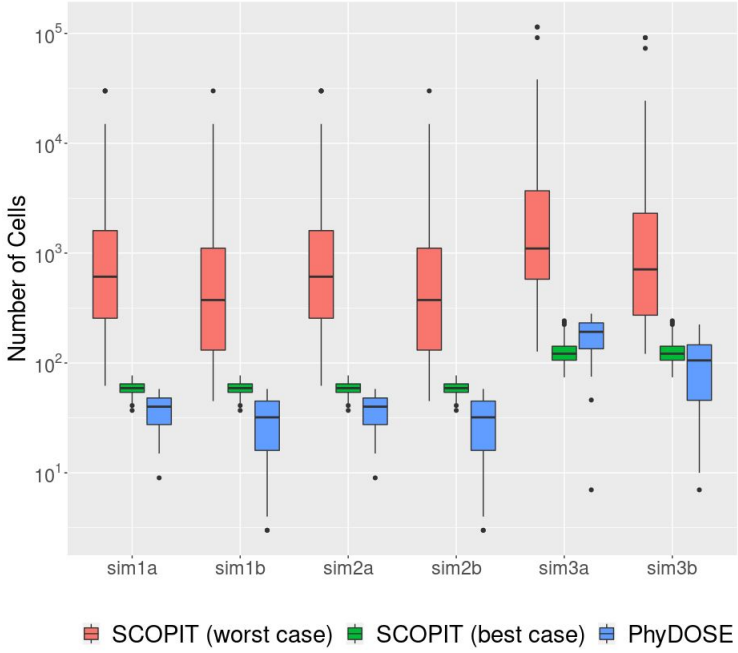
Simulation design



ID	% of Trees	Clones	Mutations	Prevalence Noise	FNR β	Doublet δ
sim1a	100%	7	7	0%	0	0
sim1b	10%	7	7	0%	0	0
sim2a	100%	7	7	5%	0	0
sim2b	10%	7	7	5%	0	0
sim2c	100%	7	7	20%	0	0
sim3a	100%	7	7	5%	0.2	0.1
sim3b	10%	7	7	5%	0.2	0.1
sim4a	100%	10	100	5%	0.2	0.1

- 100 replications
- SCOPIT comparison
- SPhyR phylogeny inference
- $\gamma = 0.95$

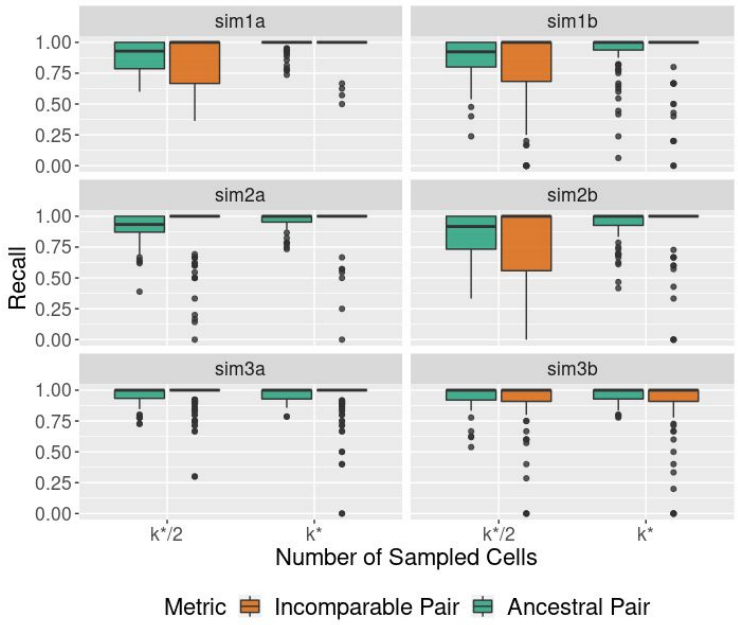
SCOPIT comparison



ID	% of Trees	Clones	Mutations	Prevalence Noise	FNR β	Doublet δ
sim1a	100%	7	7	0%	0	0
sim1b	10%	7	7	0%	0	0
sim2a	100%	7	7	5%	0	0
sim2b	10%	7	7	5%	0	0
sim2c	100%	7	7	20%	0	0
sim3a	100%	7	7	5%	0.2	0.1
sim3b	10%	7	7	5%	0.2	0.1
sim4a	100%	10	100	5%	0.2	0.1

- 100 replications
- SCOPIT comparison
- SPhyR phylogeny inference
- $\gamma = 0.95$

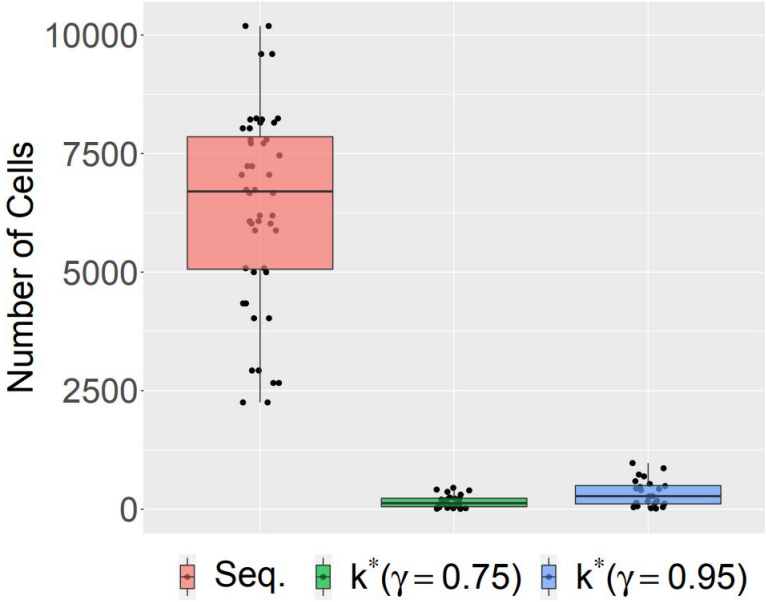
Phylogeny inference with SPhyR



ID	% of Trees	Clones	Mutations	Prevalence Noise	FNR β	Doublet δ
sim1a	100%	7	7	0%	0	0
sim1b	10%	7	7	0%	0	0
sim2a	100%	7	7	5%	0	0
sim2b	10%	7	7	5%	0	0
sim2c	100%	7	7	20%	0	0
sim3a	100%	7	7	5%	0.2	0.1
sim3b	10%	7	7	5%	0.2	0.1
sim4a	100%	10	100	5%	0.2	0.1

- 100 replications
- SCOPIT comparison
- SPhyR phylogeny inference
- $\gamma = 0.95$

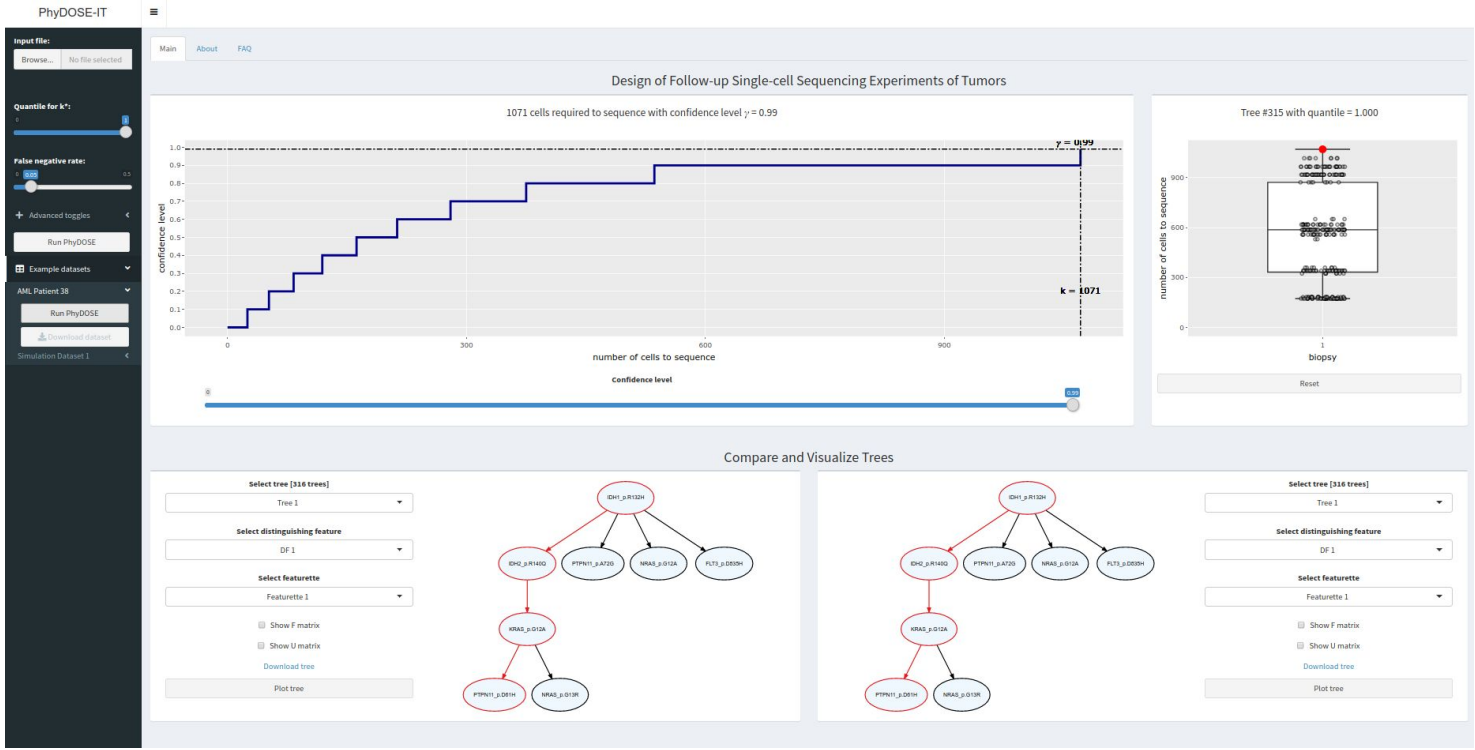
Acute myeloid leukemia (AML) cohort



PhyDOSE k compared with the original number of cells sequenced*

- Morita et al. (2020) performed high throughput targeted microfluidic single cell DNA sequencing on a cohort of 77 patients with AML.
- Based on the published variant allele frequencies, we enumerated between 2 and 316 candidate trees for 24 patients and used PhyDOSE to estimate k*.

PhyDOSE-IT and phydoser R package



https://phydose.shinyapps.io/PhyDOSE-IT/ https://github.com/elkebir-group/phydoser

Conclusions and future work

PhyDOSE Conclusions

- Proposes cost-efficient single-cell experiment design to yield high-fidelity phylogenies
- Agnostic to the type of single-cell sequencing technology used
- Available as both a web-application and an R package

Future Work

- Optimally determine the number of cells to sequence across multiple biopsies
- Explore evolutionary models beyond the infinite sites model
- Formulate and solve the RE-SCS-PC problem
  - Find out next time what it means to me...

Acknowledgements

El-Kebir group

- **Mohammed El-Kebir**
- **Nuraini Aguse**
- Yuanyuan Qi
- Jiaqi Wu
- Sarah Christensen
- Palash Sashittal
- Juho Kim
- Jackie Oh
- Chuanyi Zhang

UIUC Center for Computational Biotechnology and Genomic Medicine (grant: CSN 1624790)

National Science Foundation (CCF-1850502)

CENTER FOR COMPUTATIONAL BIOTECHNOLOGY & GENOMIC MEDICINE