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

Leah Weber\textsuperscript{1*}, Nuraini Aguse\textsuperscript{1*}, Nicholas Chia\textsuperscript{2,3} and Mohammed El-Kebir\textsuperscript{1}

\textsuperscript{1} University of Illinois at Urbana-Champaign, Department of Computer Science
\textsuperscript{2} Microbiome Program, Center for Individualized Medicine, Mayo Clinic
\textsuperscript{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

- Founder Cell
- Advantageous Mutations
- Clonal Expansion
- Heterogeneous Tumor

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 ($)

Single-cell DNA Sequencing ($$$)

Cancer Cell Fractions

| Fraction | Value |
|----------|-------|
| 1        | 0.09  |
| 2        | 0.36  |
| 3        | 0.45  |
| 4        | 0.25  |
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
DNA sequencing of tumors

Bulk DNA Sequencing ($)

Cancer Cell Fractions

Single-cell DNA Sequencing ($$$)

Solution Space
DNA sequencing of tumors

Bulk DNA Sequencing ($)  

DNA sequencing of tumors

Single-cell DNA Sequencing ($$$)

Cancer Cell Fractions

| Cancer Cell Fractions | T1 | T2 | T3 |
|-----------------------|----|----|----|
|                       | 1  | 0.09 | 0.36 | 0.45 | 0.25 |

Solution Space

| Cancer Cell Fractions | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | ? | 0 | 1 | 1 | 0 |
|-----------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c1                    | 1 | 0 | 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 |

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 |
| SPPhyR [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?

7? 1 million?
**Key idea:** Design a cost-effective single-cell sequencing experiment using bulk DNA data

Cancer Cell Fractions:
- 1
- 0.09
- 0.36
- 0.45
- 0.25

Input Parameters:
- # of cells to sequence

PhyDOSE

# of cells to sequence
Outline

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

11

Input Parameters

# of cells to sequence

Cancer Cell Fractions

Solution Space

PhyDOSE

T1  T2  T3

1  0.09  0.36  0.45  0.25
Key idea: Bulk data guides cost effective single-cell experiment design

\[
T_1
\]

\[
\text{SINGLE-CELL SEQUENCING POWER CALCULATION (SCS-PC)}
\]

Given a set \( T \) of candidate phylogenies, frequencies \( f \)

\[
T
\]

\[
T_2
\]

\[
T_3
\]

Cancer Cell Fractions \( f \)

\[
\begin{array}{c|c|c|c|c}
\text{Index} & 1 & 0.09 & 0.36 & 0.45 & 0.25 \\
\end{array}
\]
**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 $\gamma$,

| Cancer Cell Fractions $\mathbf{f}$ | Confidence Level $\gamma = 0.95$ |
|-----------------------------------|----------------------------------|
| 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 $\gamma$, find the minimum number $k^*$ of single cells needed to determine the true phylogeny $\mathcal{T}$ among $\mathcal{T}$ with probability at least $\gamma$.

| Cancer Cell Fractions $\mathbf{f}$ | Confidence Level |
|-----------------------------------|------------------|
| $1$ | $0.09$ | $0.36$ | $0.45$ | $0.25$ | $\gamma = 0.95$ |

$\mathbf{k^*}$
Solving the SCS-PC

True phylogeny unknown
**Key idea**: condition on each tree being the true tree and solve SCS-PC

\[ T = T_1 \]

**SCS Power Calculation for Phylogeny** $T$

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

\[
\begin{array}{c|c|c|c|c}
\text{Cancer Cell Fractions} & 1 & 0.09 & 0.36 & 0.45 & 0.25 \\
\hline
\text{Confidence Level} & \gamma = 0.95 \\
\end{array}
\]
Key idea: condition on each tree being the true tree and solve SCS-PC

\[ T = T_1 \]

**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 \( f \) and confidence level \( \gamma \), find the minimum number \( k^* \) of single cells needed such that the probability of a successful SCS experiment is greater than or equal to \( \gamma \).

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

Confidence Level

\[ \gamma = 0.95 \]

\[ k^* \]
What is a successful experiment given $T$?

Cancer Cell Fractions $f$

$\begin{bmatrix} 1 & 0.09 & 0.36 & 0.45 & 0.25 \end{bmatrix}$

$T$

SCOPIT
[Davis et al. 2019]
What is a successful experiment given T?

Cancer Cell Fractions $f$

|   | 0.09 | 0.36 | 0.45 | 0.25 |
|---|------|------|------|------|

$T$

6 cells

Clonal Prevalence $u$

|   | 0.1  | 0.09 | 0.36 | 0.2  | 0.25 |
|---|------|------|------|------|------|

$\mathbf{k}$

$\mathbf{p}$

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

SCOPIT

[Davis et al. 2019]
What is a successful experiment given T?

Cancer Cell Fractions $f$

$T$

$\begin{array}{cc}
1 & 0.09 \\
0.36 & 0.45 \\
0.25 & \\
\end{array}$

6 cells

$\begin{array}{cc}
? & \cdots \\
? & \\
\end{array}$

Clonal Prevalence $u$

$\begin{array}{cc}
0.1 & 0.09 \\
0.36 & 0.2 \\
0.25 & \\
\end{array}$

$k$

$p$

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

$\begin{array}{cccccccc}
6 & 0 & 0 & 0 & 0 & 0 & \\
5 & 1 & 0 & 0 & 0 & 0 & \\
\vdots & & & & & & \\
1 & 2 & 1 & 1 & 1 & 1 & \\
2 & 1 & 1 & 1 & 1 & 1 & \\
\end{array}$

SCOPIT

[Davis et al. 2019]
What is a successful experiment given \( T \)?

\[ \text{Cancer Cell Fractions } \mathbf{f} \]

\[ \begin{array}{c}
1 & 0.09 & 0.36 & 0.45 & 0.25 \\
\end{array} \]

\[ \downarrow \]

\[ T \]

\[ \begin{array}{c}
\text{6 cells} \\
\text{k} \\
\end{array} \]

\[ \downarrow \]

\[ \text{Clonal Prevalence } \mathbf{u} \]

\[ \begin{array}{c}
0.1 & 0.09 & 0.36 & 0.2 & 0.25 \\
\end{array} \]

\[ \downarrow \]

\[ p \]

\[ \text{Success } \sim \text{Mult}(p, k) \]

\[ \begin{array}{c}
6 & 0 & 0 & 0 & 0 \\
5 & 1 & 0 & 0 & 0 \\
\vdots \\
1 & 2 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 1 \\
\end{array} \]

\[ \text{SCOPIT} \]

[Davis et al. 2019]

But we don’t always need to observe all clones for a successful experiment!
Key idea: distinguishing feature

\[ T = T_1 \]
**Key idea:** distinguishing feature

Success is defined as observing a distinguishing feature.
Probabilistic model

Cancer Cell Fractions $f$

$0.09$ $0.36$ $0.45$ $0.25$

Clonal Prevalence $u$

$0.09$ $0.36$ $0.55$

Success is defined as observing a distinguishing feature.
Probabilistic model

Cancer Cell Fractions $f$

1 0.09 0.36 0.45 0.25

Success is defined as observing a distinguishing feature.

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

Clonal Prevalence $u$

$\begin{array}{ccc}
0.09 & 0.36 & 0.55 \\
\end{array}$

$p$

$\begin{array}{ccc}
0 & 0 & 3 \\
0 & 1 & 2 \\
1 & 1 & 1 \\
1 & 2 & 0 \\
2 & 1 & 0 \\
\end{array}$

Probability $p = 0.149$

Success is defined as observing a distinguishing feature.
Power calculation for fixed tree $T$

Cancer Cell Fractions $f$

- $1$  $0.09$  $0.36$  $0.45$  $0.25$

Confidence Level

$\gamma = 0.95$

? cells

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

Clonal Prevalence $u$

- $0.09$  $0.36$  $0.55$
Power calculation for fixed tree $T$

Cancer Cell Fractions $\mathbf{f}$

\[ 1 \quad 0.09 \quad 0.36 \quad 0.45 \quad 0.25 \]

Clonal Prevalence $\mathbf{u}$

\[ 0.09 \quad 0.36 \quad 0.55 \]

Confidence Level

$\gamma = 0.95$

$\mathbf{p} = \{ ? \ldots ? \}$

$\mathbf{k}^* = 32$ is the solution to the T-SCS-PC problem.

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

$k$ prob.

| $k$ | prob. |
|-----|-------|
| 3   | 0.15  |
| 4   | 0.25  |
| ... |       |
| 15  | 0.75  |
| ... |       |
| 32  | 0.95  |
Solving the SCS-PC

Taking the maximum yields and upper bound

\[ k^* = 32 \]

\( k^* = 32 \)

\( k^* = 32 \)

\( k^* = 4 \)

\( k^* = 32 \) is the solution to the SCS-PC problem.
Solving the SCS-PC

Taking the maximum yields and upper bound

$k^* = 32$

Adjust for false negatives

$k^* = 32$

$k^* = 32$

$k^* = 4$

Account for multiple distinguishing features

$k^* = 32$ is the solution to the SCS-PC problem.
**T-SCS-PC** is NP-hard by reduction from Set Cover

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

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

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

- 100 replications
- SCOPIT comparison
- SPhyR phylogeny inference
- $\gamma = 0.95$
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 $k^*$ 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)