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

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

1  0.09  0.36  0.45  0.25
DNA sequencing of tumors

Bulk DNA Sequencing ($)

Single-cell DNA Sequencing ($$$)

Cancer Cell Fractions

Solution Space
DNA sequencing of tumors

Bulk DNA Sequencing ($)

Cancer Cell Fractions

<table>
<thead>
<tr>
<th>Cell Fraction</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1</td>
</tr>
<tr>
<td>T2</td>
<td>0.09</td>
</tr>
<tr>
<td>T3</td>
<td>0.36</td>
</tr>
<tr>
<td>T4</td>
<td>0.45</td>
</tr>
<tr>
<td>T5</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Single-cell DNA Sequencing ($$$)

Solution Space
DNA sequencing of tumors

Bulk DNA Sequencing ($)

Cancer Cell Fractions

<table>
<thead>
<tr>
<th>Fraction</th>
<th>1</th>
<th>0.09</th>
<th>0.36</th>
<th>0.45</th>
<th>0.25</th>
</tr>
</thead>
</table>

Single-cell DNA Sequencing ($$$)

Solution Space

<table>
<thead>
<tr>
<th>Cell</th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>c6</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c2</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>c3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

False Negative
### Phylogeny inference from DNA sequencing

<table>
<thead>
<tr>
<th>Method</th>
<th>Bulk Sequencing Data</th>
<th>Single-cell Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCITE</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>[Jahn et al., 2016]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OncoNEM</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>[Ross &amp; Markowetz, 2017]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPhyR</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>[El-Kebir, 2018]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SiCloneFit</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>[Zafar et al., 2019]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhiSCS</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>[Malikic et al., 2019a]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-SCITE</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>[Malikic et al. 2019b]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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.

**Input Parameters**

- **# of cells to sequence**
  - 1
  - 0.09
  - 0.36
  - 0.45
  - 0.25

**Cancer Cell Fractions**

- $T_1$
- $T_2$
- $T_3$

**PhyDOSE**

**# of cells to sequence**
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 $f$
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 $\gamma$.

<table>
<thead>
<tr>
<th>Cancer Cell Fractions $\mathbf{f}$</th>
<th>Confidence Level $\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0.09 0.36 0.45 0.25</td>
<td>$\gamma = 0.95$</td>
</tr>
</tbody>
</table>
**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 $T$ among $\mathcal{T}$ with probability at least $\gamma$.

<table>
<thead>
<tr>
<th>Cancer Cell Fractions $\mathbf{f}$</th>
<th>Confidence Level</th>
<th>$\gamma = 0.95$</th>
<th>$k^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 0.09, 0.36, 0.45, 0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Solving the SCS-PC

$T_1$ $T_3$ $T_2$

True phylogeny unknown

$T$
**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 \( \mathbf{f} \) and confidence level \( \gamma \),

\[
\begin{array}{c|c|c|c|c|c}
\text{Cancer Cell Fractions} & 1 & 0.09 & 0.36 & 0.45 & 0.25 \\
\hline
\text{Confidence Level} & \gamma = 0.95
\end{array}
\]
**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 \( \mathbf{f} \) 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 \( \gamma \).

<table>
<thead>
<tr>
<th>Cancer Cell Fractions ( \mathbf{f} )</th>
<th>Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0.09 0.36 0.45 0.25</td>
<td>( \gamma = 0.95 )</td>
</tr>
</tbody>
</table>

\[ k^* = \arg \min_k P(\text{Success} \mid T, \mathcal{T}, k, \mathbf{f}) \geq \gamma \]
What is a successful experiment given T?

Cancer Cell Fractions $f$

$T$

SCOPIT

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

Cancer Cell Fractions $f$

$1 \quad 0.09 \quad 0.36 \quad 0.45 \quad 0.25$

$T$

$6$ cells

$k$

Clonal Prevalence $u$

$0.1 \quad 0.09 \quad 0.36 \quad 0.2 \quad 0.25$

$p$

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

$SCOPIT$

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

Cancer Cell Fractions $f$

$T$

6 cells

$k$

Clonal Prevalence $u$

$p$

Success $\sim Mult(p, k)$

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

Cancer Cell Fractions $f$

$T$ -> 6 cells

Clonal Prevalence $u$

$p$

Success $\sim$ Mult($p$, $k$)

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

SCOPIT
[Davis et al. 2019]
**Key idea:** distinguishing feature

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

Success is defined as observing a distinguishing feature.

\[ T = T_1 \]
Probabilistic model

Cancer Cell Fractions $f$

$\begin{array}{l}
1 \quad 0.09 \\
0.36 \\
0.45 \\
0.25 \\
\end{array}$

$T$

3 cells

$\begin{array}{l}
? \\
? \\
? \\
\end{array}$

Clonal Prevalence $u$

$\begin{array}{l}
0.09 \\
0.36 \\
0.55 \\
\end{array}$

$p$

$\begin{array}{l}
0 \\
0 \\
3 \\
\end{array}$

$\begin{array}{l}
0 \\
1 \\
2 \\
\end{array}$

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

$\text{Clonal Prevalence } u$

|   | 0.09 | 0.36 | 0.55 |

<table>
<thead>
<tr>
<th></th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Power calculation for fixed tree $T$

Confidence Level

$\gamma = 0.95$

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

Cancer Cell Fractions $f$

| 1 | 0.09 | 0.36 | 0.45 | 0.25 |

Clonal Prevalence $u$

| 0.09 | 0.36 | 0.55 |

$p$
Power calculation for fixed tree $T$

Cancer Cell Fractions $f$

| 1 | 0.09 | 0.36 | 0.45 | 0.25 |

$T$

Confidence Level $\gamma = 0.95$

$\text{? cells}$

Clonal Prevalence $u$

0.09 0.36 0.55

$p = \text{?}$

$k^* = 32$

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

$k$ 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

Taking the maximum yields and upper bound

\[ k^* = 32 \]

\[ k^*_1 = 32 \]
\[ k^*_3 = 32 \]
\[ k^*_2 = 4 \]

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

Taking the maximum yields and upper bound

\[ k^* = 32 \]

Account for multiple distinguishing features

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

Adjust for false negatives
T-SCS-PC is NP-hard by reduction from Set Cover

Lemma: Let \((\mathcal{J}, T_0, f, \gamma = \epsilon)\) be the \(T\text{-SCS-PC}\) 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

\[
\Pr(Y_{k^*} \mid u(T_0, f)) \geq \gamma
\]
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
- $\gamma = 0.95$
Acute myeloid leukemia (AML) cohort

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 $k^*$. 

Number of Cells

Seq. $k^*(\gamma = 0.75)$ $k^*(\gamma = 0.95)$
PhyDOSE-IT and phydoser R package

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)