Cancer Phylogeny Inference Using Multi-Sample Somatic Variants

Victoria Popic
PhD Student in Batzoglou Lab
Stanford University
Intra-Tumor Heterogeneity
Intra-Tumor Heterogeneity
Multi-Sample Sequencing Studies

Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing
Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horvath, M.Math., James Larkin, M.D., Ph.D., David
Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing
Marco Gerlinger, Stuart Horvath

Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing
Li Ding, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S.

Genome evolution during progression to breast cancer

Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling.
LICHeE*: Fast and Scalable Inference of Multi-Sample Cancer Lineages

Genome Biology, 2015

* Initially called “SMuTH” (’13-’14)

Raheleh Salari  Rob West  Arend Sidow
**Input Data**

Single Nucleotide Variants (SNVs)

<table>
<thead>
<tr>
<th>#chr</th>
<th>position</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>1 184306474</td>
<td>A/G HMCN1</td>
</tr>
<tr>
<td>$M_2$</td>
<td>1 18534005</td>
<td>C/A IGSF21</td>
</tr>
<tr>
<td>$M_3$</td>
<td>1 110456920</td>
<td>G/A UBL4B</td>
</tr>
</tbody>
</table>

... $M_N$ 10 26503064 C/G MYO3A

**Variant allele frequencies (VAFs) per sample**

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>...</th>
<th>$S_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.1</td>
<td>0.25</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.4</td>
<td>0.4</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**Note:** In general, the method can handle any type of variant given its cell prevalence (CP) values in each sample.
Mutations do not recur independently in different cells
⇒ cells sharing the same mutation must have inherited it from a common ancestral cell
Perfect Phylogeny Model: Constraints

Three SNV Ordering Constraints:

1. a mutation present in a given set of samples cannot be a successor of a mutation present in a smaller subset of these samples
2. a mutation cannot have a VAF higher than that of its predecessor mutation (except due to CNVs)
3. the sum of the VAFs of mutations disjointly present in distinct subclones cannot exceed the VAF of a common predecessor mutation present in these subclones

(1) M1: $S_1 \ S_3 \ S_4$

(2) M1: 0.2 VAF

(3) M1: 0.3 VAF

M2: $S_1 \ S_2 \ S_3 \ S_4$

M2: 0.4 VAF

M2: 0.2 VAF

M3: 0.3 VAF
Perfect Phylogeny Model: Constraints

Three SNV Ordering Constraints:
1. a mutation present in a given set of samples cannot be a successor of a mutation present in a smaller subset of these samples
2. a mutation cannot have a VAF higher than that of its predecessor mutation (except due to CNVs)
3. the sum of the VAFs of mutations disjointly present in distinct subclones cannot exceed the VAF of a common predecessor mutation present in these subclones

**Goal:** find all lineage trees that satisfy the above three constraints
LICHeE’s Problem Formulation

DAG encoding all pairwise valid precedence relationships – evolutionary constraint network
LICHeE’s Problem Formulation

True lineage tree will be a spanning tree of this DAG

DAG encoding all pairwise valid precedence relationships – *evolutionary constraint network*

→ search for all lineage trees that satisfy constraint (3)
**Given:** SSNV multi-sample VAFs

**Algorithm steps:**
1. Grouping and clustering SSNVs
2. Evolutionary Constraint Network Construction
3. Lineage Tree Search and Ranking
1. Grouping and clustering SSNVs
   - presence patterns across samples
   - VAF similarity
Presence Patterns Across Samples

Germline SNV  SSNVs (present in subsets of samples, not in lymph)

VAF

Greyzone

Binary Profile 1 1 1 1 1 0 1 0 1 1 0 0 1 1 0 0 0 0 1
Presence Patterns Across Samples

Germline SNV

SSNVs (present in subsets of samples, not in lymph)

VAF

Greyzone

Binary Profile

SSNV Groups
VAF-Based Clustering

Mean \((0.35, 0.3, 0.2)\)

\((0.2, 0.2, 0.15)\)

\((0.2, 0.15)\)

\((0.1)\)
2. **Evolutionary Constraint Network Construction**

- encodes whether a given cluster of SSNVs could have preceded another
- valid lineage trees are embedded in this network
Evolutionary Constraint Network
3. **Lineage Tree Search and Ranking**

- search for spanning trees satisfying VAF constraints within an error margin (extension of Gabow and Myers’78)
- top tree minimizes the squared deviation from the cluster centroids
Lineage Tree Search

Sample composition:

- **Lymph**
- Sample 1
- Sample 2
- Sample 3
- Sample 4

VAF constraint violation

VAF values:
- (0.35, 0.3, 0.2)
- (0.2, 0.15)
- (0.1, 0.2, 0.1)
- (0.2, 0.2, 0.15)
- (0.1)
RESULTS
ccRCC Study by Gerlinger et al. (2014)

8 patients, 587 SNVs

ccRCC Study by Gerlinger et al. (2014)
ccRCC Study by Gerlinger et al (2014)
ccRCC Study by Gerlinger et al (2014)

RMH004

VT  R10  R2  R4  R8  R3

GL

18  14  3  13  10  6  2  17  12

0.06  0.17

R2  R4  R10  R3

MSH6

ATM

PTEN

SMARCA4

VHL

PBRM1

ARID1A

R4  R8

R2
HGSC Study by Bashashati et al. (2013)

19 tumors, 6 patients, 340 SNVs

HGSC Study by Bashashati et al. (2013)
LICHeE Runtime DEMO Movie
Acknowledgements
Raheleh Salari
Iman Hajirasouliha
Dorna Kashef-Haghighi
Daniel Newburger
Robert West
Arend Sidow
Serafim Batzoglou

Thank You