IntraTumor Heterogeneity and Cancer Evolution

S. Cenk Sahinalp

Bogaz’da Yaz Okulu 2018
1. **Algorithmic infrastructure for genomics**
   - mapping, indexing, compression of big omic data to accommodate 250M human genomes to be sequenced by 2030
2. **Interpretation of genomic sequence data to resolve the sequence composition of repetitive genomic loci (e.g. immunoglobulin heavy locus, pharmacogenes)**
3. **Large scale (expressed) genomic alteration detection in heterogeneous tumor samples and tumor evolution modeling**
4. **Cancer network discovery and (rare) cancer driver prioritization**
5. **The role of lncRNA based regulation in tumor emergence or progression**
Integrative inference of (sub)clonal tumor evolution from bulk and single-cell sequencing data
Clonal theory of cancer evolution

Computational problems in intro-tumor heterogeneity

1. Number of distinct cancer cell populations
2. For each population set of mutations it harbors
3. Tumor purity and cancer cell fraction of each population
4. Tumor evolutionary tree

<table>
<thead>
<tr>
<th>1. Number of distinct cancer cell populations</th>
<th>2. For each population set of mutations it harbors</th>
<th>3. Tumor purity and cancer cell fraction of each population</th>
<th>4. Tumor evolutionary tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><img src="mutation-symbols.png" alt="Mutation symbols" /></td>
<td><img src="cell-symbols.png" alt="Cell symbols" /></td>
<td><img src="tumor-purity.png" alt="Tumor purity and cell fraction" /></td>
<td><img src="evolutionary-tree.png" alt="Evolutionary tree" /></td>
</tr>
</tbody>
</table>
Clonal theory of cancer evolution

\[ \star = \{M_1, M_2, \ldots \} \]

Set of mutations providing selective advantage

Nowell PC, Science 1976
Tree of tumor evolution

Healthy cells

First cancer cells

Subclonal population

Cancer-driving mutations

Subclonal mutations

Healthy cells

15%

First cancer cells

10%

Subclonal population

35%

20%

20%

Set of mutations providing selective advantage

\[ \{M_1, M_2, \ldots \} \]
Deciphering tumor evolution and subclonal composition

1. Sequencing
2. Mutation calling
3. Tree inference
Studying tumor evolution by the use of bulk sequencing data

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Var. reads</th>
<th>Ref. reads</th>
<th>(Avg) CN</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>820</td>
<td>1100</td>
<td>2</td>
</tr>
<tr>
<td>M2</td>
<td>324</td>
<td>3700</td>
<td>4</td>
</tr>
<tr>
<td>M3</td>
<td>1215</td>
<td>2800</td>
<td>3</td>
</tr>
</tbody>
</table>

sequencing and mutation calling
Studying tumor evolution by the use of bulk sequencing data

Kuipers et al., *BBA-Reviews on Cancer*, 2017

<table>
<thead>
<tr>
<th>Software</th>
<th>Year</th>
<th>Reference</th>
<th>Phylogeny</th>
<th>Multiple samples</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrAp</td>
<td>2013</td>
<td>[37]</td>
<td>Y</td>
<td>N</td>
<td>Exhaustive search</td>
</tr>
<tr>
<td>Clonial</td>
<td>2014</td>
<td>[31]</td>
<td>N</td>
<td>Y</td>
<td>Binomial / EM</td>
</tr>
<tr>
<td>PyClone</td>
<td>2014</td>
<td>[38]</td>
<td>N</td>
<td>Y</td>
<td>Dirichlet process, beta-binomial / MCMC</td>
</tr>
<tr>
<td>RecBTP</td>
<td>2014</td>
<td>[39]</td>
<td>Y</td>
<td>N</td>
<td>Approximation algorithm</td>
</tr>
<tr>
<td>SciClone</td>
<td>2014</td>
<td>[40]</td>
<td>N</td>
<td>N</td>
<td>Beta mixture model</td>
</tr>
<tr>
<td>AnceTree</td>
<td>2015</td>
<td>[41]</td>
<td>Y</td>
<td>Y</td>
<td>Optimisation / MILP</td>
</tr>
<tr>
<td>CITUP</td>
<td>2015</td>
<td>[42]</td>
<td>Y</td>
<td>Y</td>
<td>Optimisation / QIP</td>
</tr>
<tr>
<td>LICHeE</td>
<td>2015</td>
<td>[43]</td>
<td>Y</td>
<td>Y</td>
<td>Heuristic</td>
</tr>
<tr>
<td>BayClone</td>
<td>2015</td>
<td>[44]</td>
<td>N</td>
<td>Y</td>
<td>Gibbs sampling / Metropolis-Hastings</td>
</tr>
<tr>
<td>CTISingle</td>
<td>2016</td>
<td>[45]</td>
<td>Y</td>
<td>N</td>
<td>Dirichlet process, beta-binomial / MCMC</td>
</tr>
<tr>
<td>Clue</td>
<td>2016</td>
<td>[46]</td>
<td>Y</td>
<td>Y</td>
<td>Metropolis-coupled MCMC</td>
</tr>
<tr>
<td>CHAT</td>
<td>2014</td>
<td>[54]</td>
<td>N</td>
<td>N</td>
<td>Dirichlet process Gaussian mixture model / MCMC</td>
</tr>
<tr>
<td>CloneHD</td>
<td>2014</td>
<td>[55]</td>
<td>N</td>
<td>Y</td>
<td>HMM / local optimisation</td>
</tr>
<tr>
<td>SubcloneSeeker</td>
<td>2014</td>
<td>[56]</td>
<td>Y</td>
<td>Y</td>
<td>Exhaustive enumeration</td>
</tr>
<tr>
<td>PhyloWGS</td>
<td>2015</td>
<td>[58]</td>
<td>Y</td>
<td>Y</td>
<td>Tree-structured stick-breaking / MCMC</td>
</tr>
<tr>
<td>SCHISM</td>
<td>2015</td>
<td>[57]</td>
<td>Y</td>
<td>Y</td>
<td>Likelihood ratio tests / genetic algorithm</td>
</tr>
<tr>
<td>SPRUCE</td>
<td>2016</td>
<td>[59]</td>
<td>Y</td>
<td>Y</td>
<td>Exhaustive enumeration</td>
</tr>
<tr>
<td>CANOPY</td>
<td>2016</td>
<td>[60]</td>
<td>Y</td>
<td>Y</td>
<td>MCMC</td>
</tr>
</tbody>
</table>
Studying tumor evolution by the use of bulk sequencing data

Kuipers et al., BBA-Reviews on Cancer, 2017
**CTPsingle: Clustering of mutations based on read counts**

- \( M_i \): heterozygous SNV from diploid region
- \( t_i \): total number of reads covering genomic position of \( M_i \)
  \[ \Rightarrow \text{total number of cells in the sample} \sim \frac{t_i}{2} \]
- \( v_i \): total number of reads supporting \( M_i \)
  \[ \Rightarrow \text{total number of cells harboring} \ M_i \text{ is} \sim v_i \]

Expected fraction of cells harboring \( M_i \)

\[ \frac{v_i}{t_i} = \frac{2v_i}{t_i} = 2 \cdot VAF(M_i) \]

**THE MAIN ASSUMPTIONS:**

1. Mutations having similar \( 2 \cdot VAF(M_i) \) occur for the first time at the same cellular population.
2. The existence of clusters of mutations with similar \( 2 \cdot VAFs \).

<table>
<thead>
<tr>
<th>ClusterID</th>
<th>Frequency</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>10%</td>
<td>⭐️</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>30%</td>
<td>🌟</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>55%</td>
<td>🌟</td>
</tr>
</tbody>
</table>

**Graph:**
- **X-axis:** \( 100 \cdot 2 \cdot VAF \)
- **Y-axis:** number of mutations with value of \( 100 \cdot 2 \cdot VAF \) equal to the corresponding value at x-axis
Clustering ambiguity: subclones with similar cellular prevalence

1. During the clustering step mutations emerging at subclonal populations with similar cellular prevalence are clustered together
2. Inaccurate clustering influences the inference of subclonal prevalences and tree of tumor evolution
Phylogeny inference ambiguity: multiple equally likely trees

1. Linear (chain) topology is usually among solutions
2. In many cases, in addition to linear topology, we also have other solutions with score equal to 0.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>10%</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>30%</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>55%</td>
</tr>
</tbody>
</table>
12 patients sequenced at three time points of interest

- Baseline, On-Treatment (12-weeks), and Progression

- Sensitively obtain mutation calls (through SiNVICT)

- For each mutation, check:
  - whether treatment has eliminated subclones,
  - whether new and more aggressive subclones emerged
Time Series Liquid Biopsy Data Can Help
Studying tumor evolution by the use of single-cell sequencing (SCS) data

### Mutational Matrix

<table>
<thead>
<tr>
<th>mut\cell</th>
<th>cell1</th>
<th>cell2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1</td>
<td>0</td>
<td>---</td>
</tr>
<tr>
<td>M2</td>
<td>0</td>
<td>1</td>
<td>---</td>
</tr>
<tr>
<td>M3</td>
<td>1</td>
<td>NA</td>
<td>---</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Studying tumor evolution by the use of single-cell sequencing (SCS) data

<table>
<thead>
<tr>
<th>Name of method</th>
<th>Authors</th>
<th>Journal</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>Kim and Simon</td>
<td>BMC Bioinformatics</td>
<td>2013</td>
</tr>
<tr>
<td>BitPhylogeny</td>
<td>Yuan et al.</td>
<td>Genome Biology</td>
<td>2015</td>
</tr>
<tr>
<td>OncoNEM</td>
<td>Ross &amp; Markowetz</td>
<td>Genome Biology</td>
<td>2016</td>
</tr>
<tr>
<td>SCITE</td>
<td>Jahn et al.</td>
<td>Genome Biology</td>
<td>2016</td>
</tr>
<tr>
<td>SiFit</td>
<td>Zafar et al.</td>
<td>Genome Biology</td>
<td>2017</td>
</tr>
</tbody>
</table>
Single-cell sequencing data

**SINGLE-CELL ISOLATION**
- Doublets
- Subclones with zero sampled cells
- Non-uniform sampling
**Single-cell sequencing data**

**SINGLE-CELL ISOLATION**
- Doublets
- Subclones with zero sampled cells
- Non-uniform sampling

**DNA AMPLIFICATION**
- Amplification errors (false positives)
- Unequal amplification (false negatives, NAs)

**DNA SEQUENCING AND MUTATION CALLING**
- False positives

<table>
<thead>
<tr>
<th></th>
<th>(m) cells</th>
<th>(\text{mut.})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_1)</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>(M_2)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(M_3)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(M_4)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
</tbody>
</table>

\(\bigstar = [M_2, M_3]\)
\(\bigstar = [M_4, M_5]\)
**Single-cell sequencing data**

**SINGLE-CELL ISOLATION**
- Doublets
- Subclones with zero sampled cells
- Non-uniform sampling

**DNA AMPLIFICATION**
- Amplification errors (false positives)
- Unequal amplification (false negatives, NAs)

**DNA SEQUENCING AND MUTATION CALLING**
- False positives

**Main types of noise in SCS data**
1. False positive (FP) usually \( \leq 10^{-5} \)
2. False negative (FN) in the range 0.1 – 0.3
3. Missing entries (NA) varies between 0.05-0.50
4. Doublets varies between 0 and 0.30
Tree inference by the use of SCS data - overview

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
<th>C₄</th>
<th>C₅</th>
<th>C₆</th>
<th>C₇</th>
<th>C₈</th>
<th>C₉</th>
</tr>
</thead>
<tbody>
<tr>
<td>M₁</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M₂</td>
<td>NA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M₃</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M₄</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M₅</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M₆</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M₇</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M₈</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M₉</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M₁₀</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$D_{n \times m}$ – single-cell data mutation matrix

$n$ – number of mutations

$m$ – number of cells

$T$, $\theta$ – tree topology

$\theta = (\alpha, \beta)$

$\alpha$ – false positive rate

$\beta$ – false negative rate
Strengths and weaknesses of SCS data

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>c6</th>
<th>c7</th>
<th>c8</th>
<th>c9</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>M2</td>
<td>NA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M8</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Advantages of combining bulk and SCS data

Bulk data

True tree

Clustering

Inferred solution

Single-cell data

Identifying branching helps in resolving clustering ambiguity
Many of the multiple topologies that are equally likely based on the bulk data have very low support from SC data.

**Advantages of combining bulk and SCS data**

**Bulk data**

- **True tree**
- **Possible set of optimal solutions**
- **Phylogenies which are optimal based on bulk data, but have low support from SC data**

**Single-cell data**

- **Inferred solution**
- **Identifying branching helps in resolving clustering ambiguity**
Advantages of combining bulk and SCS data

**Single-cell data**

- $15\%$ for $M_3$
- $10\%$ for $M_4$
- $35\%$ for $M_3$

Based on SC data $M_4$ is more likely to occur before $M_3$

**Bulk data**

- $2 \cdot VAF(M_3) = 0.85$
- $2 \cdot VAF(M_4) = 0.75$

From bulk data we have that it is very likely that $M_4$ does not occur before $M_3$
**B-SCITE – input and output**

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Variant reads</th>
<th>Reference reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 )</td>
<td>1100</td>
<td>2587</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>804</td>
<td>2710</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>537</td>
<td>3211</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
<th>( C_8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>NA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
</tbody>
</table>

**INPUT DATA REQUIREMENTS:**
- SNVs from regions not affected by CNAs
- Consider only mutations present in at least one single cell
- Targeted deep sequencing (\( \geq 1000x \)) of bulk sample

![B-SCITE Diagram]
\[ S_{\text{joint}}(T^*, \theta^*) = \arg\max_{(T, \theta)} [S_{SC}(T, \theta) + S_{bulk}(T)] \]
MCMC step

Current state \((T, \theta)\)

1. Propose \((T', \theta')\) state
First decide whether new \(T\) or new \(\theta\) is proposed
First decide whether new $T$ or new $\theta$ is proposed

**CASE 1: New $\theta$ is proposed**
- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$
  note that computation of bulk score is not required
MCMC step

Current state \((T, \theta)\)

1. Propose \((T', \theta')\) state

First decide whether new \(T\) or new \(\theta\) is proposed

**CASE 1: New \(\theta\) is proposed**
- \(T' = T\)
- \(\theta' = (\alpha', \beta')\) is proposed via simple Gaussian walk
- Compute \(S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')\)
  note that computation of bulk score is not required

**CASE 2: New \(T\) is proposed**
- \(T' = \) propose new mutation tree
An example of proposing new mutation tree

$T$

$T'$
1. Propose $(T', \theta')$ state

First decide whether new $T$ or new $\theta$ is proposed

**CASE 1: New $\theta$ is proposed**
- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{\text{joint}}(T', \theta') = S_{SC}(T', \theta') + S_{\text{bulk}}(T')$
  - Note that computation of bulk score is not required

**CASE 2: New $T$ is proposed**
- $T' = \text{propose new mutation tree}$
- $\theta' = \theta$
- Compute $S_{\text{joint}}(T', \theta') = S_{SC}(T', \theta') + S_{\text{bulk}}(T')$
**MCMC step**

1. **Propose \((T', \theta')\) state**

   First decide whether new \(T\) or new \(\theta\) is proposed.

   **CASE 1: New \(\theta\) is proposed**
   
   - \(T' = T\)
   - \(\theta' = (\alpha', \beta')\) is proposed via simple Gaussian walk
   - Compute \(S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')\)
     
     note that computation of bulk score is not required

   **CASE 2: New \(T\) is proposed**
   
   - \(T' = \) propose new mutation tree (steps described later)
   - \(\theta' = \theta\)
   - Compute \(S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')\)

2. **Accept or decline proposed \((T', \theta')\)**

   Accept the proposed \((T', \theta')\) with the probability

   \[
   \min \left\{ 1, \frac{q(T, \theta \mid T', \theta') P(T', \theta' \mid D)}{q(T', \theta' \mid T, \theta) P(T, \theta \mid D)} \right\}
   \]
1. Propose \((T', \theta')\) state

First decide whether new \(T\) or new \(\theta\) is proposed

**CASE 1: New \(\theta\) is proposed**
- \(T' = T\)
- \(\theta' = (\alpha', \beta')\) is proposed via simple Gaussian walk
- Compute \(S_{\text{joint}}(T', \theta') = S_{\text{SC}}(T', \theta') + S_{\text{bulk}}(T')\)
  note that computation of bulk score is not required

**CASE 2: New \(T\) is proposed**
- \(T' = \) propose new mutation tree (steps described later)
- \(\theta' = \theta\)
- Compute \(S_{\text{joint}}(T', \theta') = S_{\text{SC}}(T', \theta') + S_{\text{bulk}}(T')\)

2. Accept or decline proposed \((T', \theta')\)

Accept the proposed \((T', \theta')\) with the probability

\[
\min \left\{ 1, \frac{q(T, \theta \mid T', \theta') P(T', \theta' \mid D)}{q(T', \theta' \mid T, \theta) P(T, \theta \mid D)} \right\}
\]
Mutation trees $\rightarrow$ clonal trees

Inferred mutation tree
Mutation trees → clonal trees

Inferred mutation tree
B-SCITE applied to Triple Negative BC

Ground truth inferred by Wang et al. 2014: on a TNBC specimen: three subclones inferred through hierarchical clustering of 374 mutations extracted from 144 single cell whole exome sequence data.
Results for TNBC Patient 1 from Wang et al. 2014: Input data consists of 72 single-cells and 18 mutations.
(a) Ground truth tree from Wang et al. 2014. (b) Tree obtained by SCS data only. (c) Tree reported by B-SCITE.
Results for ALL Patient 1 from Gawad et al. 2014: (a) trees obtained by clustering bulk-data read counts (coverage ~ 2000). (b) Tree obtained by SCS data only (c) Tree reported by B-SCITE. Input data consists of 111 single-cells and 20 mutations.
Conclusions

- Methods to infer clonal trees of evolution by the use of single (CTP-Single), multi-site (CITUP) bulk and integrated single-cell (B-SCITE) sequencing data
- Robust to the presence of various types of noise in both types of input data
- Achieve high accuracy, including on tumors consisting of tens of subclones
- Extend to the cases with multiple bulk-sequencing data
- Outperform existing methods on all measures of accuracy

More details: [https://www.biorxiv.org/content/early/2017/12/15/234914](https://www.biorxiv.org/content/early/2017/12/15/234914) (Malikic et al. RECOMB 2018)

CTP-Single & CITUP available at [https://github.com/nlgnndnmz/CTPsingle](https://github.com/nlgnndnmz/CTPsingle)

B-SCITE available at [https://github.com/smalikic/B-SCITE](https://github.com/smalikic/B-SCITE)
ReMixT:
Reconstructing Clone Specific Genomic Structure in Heterogeneous Tumor Samples via Bulk WGS
Segment Copy Number Change Evident in Whole Genome Sequencing Read Depths
Normal Contamination and Clonal Diversity Dilute the Signal of Copy Number Changes

**Tumour Sample**

- **Normal Cells**
- **Dominant Tumour Clone Cells**
- **Subdominant Tumour Clone Cells**

**Whole Genome Sequence Data**

- **Concordant Alignments**
- **Copy Number**
  - **Read Depth**
Joint Analysis to Increase Statistical Strength for Identifying Subclonal Copy Number Changes

**Tumour Sample**

- **Normal Cells**
- **Dominant Tumour Clone Cells**
- **Subdominant Tumour Clone Cells**

**Whole Genome Sequence Data**

- **Concordant Alignments**
- **Copy Number**
- **Read Depth**
- **Breakpoints**
- **Structure**
ReMixT: Probabilistic Genome Graph Model

- Allele and clone specific copy number model
  - HMM augmented with breakpoint dependencies
  - Unified state space
  - $O(K^2)$ transition calculations by exploiting symmetry
- Outlier modeling
- Allele uncertainty modelling
ReMixT: Reproducible Clonal Dynamics in Replicate Xenografts

- Cellular Frequency of Breakpoints recapitulates SNV clonal dynamics
- All breakpoints either ascend or descend between through successive passaging
ReMixT: Validation with SA501X3F Whole Genome Single Cell

- Single cell data validates subclonal segments associated with clone specific breakpoints
Conclusion

• ReMix-T simultaneously infers clone specific breakpoints and associated copy number alterations in a heterogeneous tumor sample from bulk sequencing data.

• Can predict breakpoint and associated subclonal frequencies

Current/Future Directions

- Exact solutions for SCS+Bulk HTS based on ILP and CSP instead of MCMC
- Perfect phylogeny with infinite sites model to be replaced with Dollo parsimony
- Clone specific SNV + SV + CNV composition of heterogeneous tumor samples from integrated bulk and single cell sequencing data
- The use of long read/single molecule sequencing technologies to associate two or more breakpoints for better structural inference
- Algorithms that can scale up to accommodate thousands of single cell WGS data
- Integration with liquid biopsy sequencing