SVJedi: Structural variation genotyping using long reads

Lolita Lecompte1, Pierre Peterlongo1, Dominique Lavenier1, Claire Lemaitre1
1Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

Introduction

Structural variations (SVs) are DNA segments that have been moved regardless another reference genome. The number of known SVs is increasing, especially with the development of third generation sequencing technologies, which produce long reads data.

Discovery vs. Genotyping

SV discovery

Given a reference genome, discover SVs between a sequencing sample and a reference genome.


SV genotyping

Given a set of already known variants for a species, assess the presence or absence of each variant in a given sequencing sample.

There is no known method for SV genotyping with long reads.

Method - SVJedi

1. Generating allele sequences
2. Alignment: Minimap2
3. Selection of informative alignments
   - Breakpoint overlap
   - Semi-global alignment
4. Genotype estimation

Results on a simulated dataset

Simulated dataset

- 1,000 deletions selected on the human chromosome 1 from dbVar (50 bp - 10k bp)
- Equally distributed: 0/0, 0/1, 1/1
- Long reads simulated with SimLoRD [1]
  16 % error rate, 20X sequence depth

SVJedi

<table>
<thead>
<tr>
<th>allele</th>
<th>0/0</th>
<th>0/1</th>
<th>1/1</th>
<th>./.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0</td>
<td>330</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0/1</td>
<td>19</td>
<td>305</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>1/1</td>
<td>3</td>
<td>10</td>
<td>311</td>
<td>9</td>
</tr>
</tbody>
</table>

59.85 % Precision

Sniffles [2]

<table>
<thead>
<tr>
<th>allele</th>
<th>0/0</th>
<th>0/1</th>
<th>1/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0</td>
<td>0</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>0/1</td>
<td>30</td>
<td>89</td>
<td>0</td>
</tr>
<tr>
<td>1/1</td>
<td>0</td>
<td>254</td>
<td>60</td>
</tr>
</tbody>
</table>

80.9 % of 1/1 deletions are wrongly genotyped as 0/1

Results on a real dataset

Dataset

- Sample: NA12878
- Data: Nanopore WGS consortium (37X)
- svclassify [3]: 2,677 high confidence deletions

svclassify

<table>
<thead>
<tr>
<th>allele</th>
<th>0/0</th>
<th>0/1</th>
<th>1/1</th>
<th>./.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0</td>
<td>194</td>
<td>1,289</td>
<td>202</td>
<td>28</td>
</tr>
<tr>
<td>0/1</td>
<td>8</td>
<td>75</td>
<td>877</td>
<td>4</td>
</tr>
</tbody>
</table>

SVJedi

<table>
<thead>
<tr>
<th>allele</th>
<th>0/0</th>
<th>0/1</th>
<th>1/1</th>
<th>./.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0</td>
<td>511</td>
<td>2,166</td>
<td>479</td>
<td></td>
</tr>
</tbody>
</table>

Identity: 81.89 %

Predicted genotypes: 98.81 %

Runtime: 2h35 (40 CPU)

Conclusion

- New method to genotype SVs using long reads data.
- Importance of dedicated genotyping methods.
- Currently only applied to deletions.
- Implementation: SVJedi

References