Structural Variations

02-710 Computational Genomics
Seyoung Kim
Genomic Rearrangements/
Structural Variations (SVs)

• 1 Kb to several Mb in size

courtesy of Tobias Rausch (EMBL)
Genomic Rearrangements/Structural Variations (SVs)

- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication

courtesy of Tobias Rausch (EMBL)
Genomic Rearrangements/Structural Variations (SVs)

- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication
- Insertion

![Diagram of genomic rearrangements](image)

courtesy of Tobias Rausch (EMBL)
Genomic Rearrangements/ Structural Variations (SVs)

- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication
- Insertion, Inversion

![Diagram of genomic rearrangements]

Deletion

```
α □ □ γ
```

Duplication

```
α □ □ □ □ γ
```

Insertion

```
α □ □ δ □ γ
```

Inversion

```
γ □ □ □ α
```

Courtesy of Tobias Rausch (EMBL)
Genomic Rearrangements/
Structural Variations

• 1 Kb to several Mb in size
• Copy number variants
  – Deletion
  – Duplication
• Insertion, Inversion, Translocation
• Either neutral or non-neutral in function
• Non-neutral mechanisms
  – Disrupting genes
  – Creating fusion genes
  – Copy number changes of dosage-sensitive genes

courtesy of Tobias Rausch (EMBL)
Computational Methods for Detecting Genomic Rearrangements

Mate-pair or paired-end mapping abnormalities → Split-Read alignments → Read depth signals

Reference

courtesy of Tobias Rausch (EMBL)
Paired-end Sequencing

![Diagram of Paired-end Sequencing]

- Genomic segment
- Cut many times at random (Shotgun)
- Get two reads from each segment

Adopted from http://www.cs.utoronto.ca/~brudno/csc2431w10/2431_lec1.ppt
Definition of Coverage

Length of genomic segment: \( L \)
Number of reads: \( n \)
Length of each read: \( l \)

Definition: Coverage \( C = \frac{n l}{L} \)
Depth of Coverage and Physical Coverage

- Single-end sequencing
- Paired-end sequencing
- Paired-end sequencing
Insert Size for a Mate Pair

For each fragment

Let’s assume we know the insert size distribution.
Paired-end data

- Paired-end NGS (insert size distribution known due to fragment size selection)
<table>
<thead>
<tr>
<th>Mate-pair or paired-end mapping abnormalities</th>
<th>Read-depth signals</th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
</table>

- **Reference sequence**
- **Newly sequenced Genome**
- **Insertion**

**Sequenced mate pairs mapped to the reference genome**

- **Reference sequence (known)**
- **Newly sequenced Genome (unknown)**

**Insertion**

courtesy of Tobias Rausch (EMBL)
Mate-pair or paired-end mapping abnormalities

<table>
<thead>
<tr>
<th></th>
<th>Read-depth signals</th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
</table>

---

Insertion

Reference Sequence

Newly Sequenced Genome

---

Deletion

Reference Sequence

Newly Sequenced Genome

---

courtesy of Tobias Rausch (EMBL)
Mate-pair or paired-end mapping abnormalities

<table>
<thead>
<tr>
<th>Insertion</th>
<th>Read-depth signals</th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequenced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insertion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequenced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Insertions ↔ Deletions
<table>
<thead>
<tr>
<th>Mate-pair or paired-end mapping abnormalities</th>
<th>Read-depth signals</th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
</table>

**Insertion**
- Reference Sequence
- Newly Sequenced Genome
- Insertion

**Deletion**
- Reference Sequence
- Newly Sequenced Genome
- Deletion

**Inversion**
- Reference Sequence
- Newly Sequenced Genome
- Inversion

*Courtesy of Tobias Rausch (EMBL)*
Mate-pair or paired-end mapping abnormalities

<table>
<thead>
<tr>
<th>Read-depth signals</th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
</table>

![Diagram](image.png)

courtesy of Tobias Rausch (EMBL)
courtesy of Tobias Rausch (EMBL)
Mate-pair or paired-end mapping abnormalities

<table>
<thead>
<tr>
<th>Read-depth signals</th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
</table>

- **Tandem Duplication**
  - Reference Sequence
  - Newly Sequenced Genome
  - Tandem Duplication

- **Translocation**
  - Reference Sequences
  - Newly Sequenced Genome
  - chrA
  - chrB
  - Translocation

- **Large Insertion**
  - Reference Sequence
  - Newly Sequenced Genome
  - Large Insertion

Courtesy of Tobias Rausch (EMBL)
Computational Methods for Detecting Genomic Rearrangements

Mate-pair or paired-end mapping abnormalities

Read depth signals

Split-Read alignments

Reference

courtesy of Tobias Rausch (EMBL)
<table>
<thead>
<tr>
<th>Mate-pair or paired-end mapping abnormalities</th>
<th><strong>Read-depth signals</strong></th>
<th>Split-read alignments</th>
<th>Local assembly</th>
</tr>
</thead>
</table>

courtesy of Tobias Rausch (EMBL)
Mate-pair or paired-end mapping abnormalities

Read-depth signals

Split-read alignments

Local assembly

1 Copy 1 Copy 0 Copy 2 Copy 2 Copy

courtesy of Tobias Rausch (EMBL)

Chiang et al. (2009)
- Down-Syndrom
  - Partial Trisomie 21

\[
\log_2 \frac{\# \text{Reads}_{\text{Disease}}}{\# \text{Reads}_{\text{Normal}}}
\]

courtesy of Tobias Rausch (EMBL)
Detecting Genetic Polymorphisms from Shotgun Sequencing
Computational Methods for Detecting Genomic Rearrangements

Mate-pair or paired-end mapping abnormalities

Read depth signals

Split-Read alignments

Reference

courtesy of Tobias Rausch (EMBL)
With reads of length 40-100 bps are we able to find the exact breakpoint of a structural variation?

Yes – using split-read mapping
With reads of length 40-100 bps are we able to find the exact breakpoint of a structural variation?

Yes – using anchored split-read mapping

Mappable read mate provides anchor to narrow down search space

Medvedev et al. (2009)
The Pindel algorithm (Deletions)

Ye et al. (2009)
The Pindel algorithm (Real Data)

Ye et al. (2009)
The Pindel algorithm for complex variants

a) large deletion
b) tandem duplication
c) inversion
d-f) same as a-c with non-template sequence (yellow part)
Acknowledgements

• Tobias Rausch (EMBL)
• Kai Ye (Leiden University Medical Center)
• Anne-Katrin Emde (Freie Universität Berlin)