Structural Variants I
CS 294, Fall 2012

Jasmine Nirody and Rachel Wang

Biophysics and Statistics, UC Berkeley

11/20/2012
Types of structural variation (SV): deletions, insertions (novel or mobile elements), inversions, translocations, tandem duplications, etc.

Associations with genetic diseases and cancer

Traditional techniques can detect SV of size $> 1$ kb; NGS allows for finer resolution ($> 50$ bp)

Difficulties:
- Short reads
- SVs tend to occur in repetitive regions of the genome
- Wide spectrum of type and size
Paired-End Sequence Reads

- Two ends of the same DNA segment are sequenced. The lengths of the segments follow a tight distribution.
- “Discordant” pairs: unexpected mapping span, position and/or orientation.
- Locations of break points:
  - Split-read (Pindel)
  - Read pair (BreakDancer, VariationHunter)
“Pattern growth” approach (details on board)
Pindel: Results

D 321 chr1 56173830 56174202 Supports: 15
AAGAGTTGGTTAGTTAGAAATATAGGGgccc<311>ataggACAAGGTTACAAGGAATG3C7GAAGGAGAGGATG
  GAGTTATAGAAATATAGG ACAAGGTTACAAGGGAATG + 56173670
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGGAA + 56173677
  CAGTTTATAGAAATATAGG ACAAGGTTACAAGGAATG + 56173681
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGG + 56173687
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGGAAATG + 56173690
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGGAA + 56173695
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGG + 56173697
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGGAA + 56173700
  GGGTTTATAGAAATATAGG ACAAGGTTACAAGGAAATG + 56173710
  TGGTTTATAGAAATATAGG ACAAGGTTACAAGGAATG - 56174339
  TGGTTTATAGAAATATAGG ACAAGGTTACAAGGAATG - 56174356
Pindel: Results

(a) Max_D_Size = 10
(b) Max_D_Size = 100
(c) Max_D_Size = 1000
(d) Max_D_Size = 10,000
(e) Max_D_Size = 100,000
(f) Max_D_Size = 1000,000
Pindel: Results

![Graphs showing Pindel results](image-url)
Pindel: Results

(a) Runtime (seconds) vs. Maximum deletion size
(b) Max memory (MB) vs. Maximum deletion size

- Blue: Loading reference and reads
- Red: Mining break points
- Green: Sort and output
- Black: Total CPU time
BreakDancer: Methods Overview

(a) Mapping parameters
- Paired-end reads
- Mapping
- Structural variants
- (v) Compute confidence scores
- (iv) Structural variation position, type, size and number of anomalous mapped read pairs
- Detection parameters
- (i) Genome-wide tally of anomalous read pairs
- (ii) Search for anomalous regions
- (iii) Identify interconnected clusters

(b) Deletion | Insertion | Inversion | Intrachromosomal translocation | Interchromosomal translocation

(more details on board)
BreakDancer: Results

The graphs show the results for different physical coverage levels with various read lengths. The x-axis represents the physical coverage (x), and the y-axis shows the percentage of reads correctly identified. The graphs compare different read lengths: 20bp, 40bp, 60bp, 80bp, and 100bp. Each curve represents a different coverage level, illustrating how the accuracy of read identification changes with increasing coverage and read length.
BreakDancer: Results

The graph illustrates the True Positive Rate (TPR) and False Positive Rate (FPR) for different methods as a function of physical coverage (fold). The methods include TPR BDMax Q30, FPR BDMax Q30, TPR analytic, TPR detectable, TPR BDMini, FPR BDMini, TPR BD all, and FPR BD all. The TPR generally increases with increasing physical coverage, while the FPR decreases.
BreakDancer: Results

The graph shows the performance of BreakDancer with varying separation thresholds. Different lines represent different metrics and parameter combinations, as indicated in the legend.
VariationHunter: Methods

(Details on board)
Simulation results

- Imposed set of insertions and deletions in chromosomes 1 and 22 in one of the reference genomes from the Genome Reference Consortium and random SNPs uniformly with a rate 0.1% per base

- **VariationHunter-SC** (weighted)
  - Deletions: 64% and 55% of deletions (> 200 bp) found in chromosomes 1 and 22; false positive rate 10%
  - Insertions: 50% of insertions (50-80 bp) found; false positive rate 22%

- **VariationHunter-Pr**
  - Deletions: 67% and 60% of deletions (> 200 bp) found in chromosomes 1 and 22; false positive rate 13%
  - Insertions: 55% of insertions (50-80 bp) found; false positive rate 30%
VariationHunter: Results

Table 1. Comparison of SV detected in the Illumina paired-end read library generated from the genome of NA18507 with the validated sites of variation using a fosmid-based approach from the same individual.

| Validation type | VariationHunter-SC | | VariationHunter-Pr | | Bentley et al. (2008) |
|-----------------|--------------------|----------------|------------------------|------------------------|
| Validation type |       |       |       |       |       |       |       |       |       |       |
| Deletion        | S 92  | L 143 | S 8959 | L 57  | S 7599 | L 85  | S 8537 | L 58  | S 5704 | L 49  |
| Inversion       | S 13  | L 82  | S 504  | L 2   | S 433  | L 25  | S 181  | L 11  | S NA   | L NA  |

We require that at least 50% of either the validated or predicted deletion interval be covered to call an overlap. Inversions are considered to be captured if there is any intersection between the validated and predicted interval. The original study with the Illumina data does not report the inversion calls, primarily because inversions were usually flanked by repeat sequences that were mostly missed by unique sequence mapping (Bentley et al. 2008).

*Validation types (sample [S] and locus-level [L] validation) remapped to human genome build 36.

Pred., predicted; Capt., captured; NA, not available.
### Table 2. Comparison of small indels ($\leq$100bp) detected in the Illumina paired-end read library generated from the genome of NA18507 with the DIP sites predicted by fosmid end-sequence mapping (Kidd et al. 2008) from the same individual

<table>
<thead>
<tr>
<th></th>
<th>VariationHunter-SC ($\leq$100 bp)</th>
<th>VariationHunter-Pr ($\leq$100 bp)</th>
<th>Bentley et al. (2008) ($\leq$ 100 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weighted</td>
<td>Unweighted</td>
<td>Weighted</td>
</tr>
<tr>
<td></td>
<td>Pred.</td>
<td>$\in$ DIP</td>
<td>Pred.</td>
</tr>
<tr>
<td>Deletion</td>
<td>2200</td>
<td>233</td>
<td>1885</td>
</tr>
<tr>
<td>Insertion</td>
<td>5575</td>
<td>135</td>
<td>3772</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Build 35 coordinates of the DIP intervals were converted to build 36 coordinates using the UCSC liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). We require that at least 1 bp of either the validated or predicted deletion interval be covered to call an overlap. The original study with the Illumina data does not report insertion calls detected with matepair analysis (Bentley et al. 2008). Pred., predicted; DIP, deletion/insertion polymorphism; NA, not available.
VariationHunter: Results

NA18507 Deletion Size Histogram

- AluY
- L1Hs