Population-based Detection of Structural Variants in Normal and Aberrant Genomes.

Jean Monlong

Guillaume Bourque’s group

Genome Informatics - September 21-24, 2014
Structural variation

Genetic variation involving more than 500bp.

![Diagram of human chromosome variations]

- **Reference**
- **Deletion**
- **Insertion**
- **Inversion**
- **Tandem duplication**
- **Dispersed duplication**
- **Copy-number variant**

**Structural Variant:** SV; **Copy Number Variation:** CNV.

SV detection using High-Throughput Sequencing

Read pairs

- No structural variation
- Deletion
- Mobile element insertion
- Tandem duplication

Read depth

- Sample reads
- Duplication
- Deletion

Split reads

- Deletion

Assembly

- Novel sequence insertion

Limitation

Low mappability

- Noisy or reduced signal in repeat-rich regions, centromeres, telomeres.
- Unpredictable segmentation $\rightarrow$ reduced sensitivity/specificity.
- Filtering problematic regions reduces the genome range tested.
**Objective**

Test the entire genome, including low-mappability regions, and detect subtle abnormal coverage.

**PopSV: Population-based approach**

Use a set of reference experiments to detect abnormal patterns.
PopSV: Population-based approach

Workflow

1. Genome is fragmented in bins.
2. Reads in each bin are counted, for each sample.
3. Normalization of the bin counts.
4. Each sample and each bin is tested for divergence from reference samples (Z-score).
5. P-value estimation and multiple test correction.
PopSV: importance of normalization

- Experiment-specific technical bias.
- Naive normalization (linear, quantile) is often not enough.
PopSV: importance of normalization

- Targeted normalization: linear using a subset of the genome.
PopSV: Z-score and test

For a sample $s$:

- For each bin $b$: $z = \frac{B_C^b - B_C^{b\text{reference}}}{sd^{b\text{reference}}}$

- $p_{\nu} = \mathbb{P}(|z| \leq |Z|)$ with $Z \sim \mathcal{N}(0, \sigma)$ where $\sigma$ is estimated from the $z$ distribution across all bins.
Application

**CageKid : Renal Cell Cancer**
Whole-Genome Sequencing of 100 individuals, ~ 40X coverage, Illumina paired-end 100bp, normal and tumor paired samples.

- Normal samples → reference samples.
- 2kb bins.

**Read-Depth measure - 2 strategies**
- **concordant reads**: only properly paired and mapped read pairs.
- **discordant reads**: improperly mapped read pairs or low mapping quality.
Using concordant reads

“funky snowman” plot
Example: Telomeric region

Chr.10, overlapping genes (PRAP1, CALY), not detected by other approaches.
Example: Partial tumoral event

Chr.1, overlapping CDC14A gene (cell division cycle), not detected by other approaches.
Validation and benchmark

- Germline events detected in tumor samples?
- Consistent with SNP-array calls?
- Twin dataset: consistent with the pedigree?

Germline events detected in tumor samples

PopSV detected more consistent calls than other methods with similar specificity.
Centromere/telomere/gap and systematic errors

![Graph showing CNV frequency in normals across different methods and distance to centromere/telomere/gap (Mb).](image)

- **method**: cn.MOPS, FREEC, PopSV

The graph illustrates the CNV frequency in normals across different methods as a function of distance to centromere/telomere/gap (Mb). The x-axis represents distance in Mbase pairs (Mb), and the y-axis represents CNV frequency in normals. The graph uses different colors to indicate the methods, with **red** for cn.MOPS, **green** for FREEC, and **blue** for PopSV.
PopSV using discordant reads

- Discordant reads support SVs.
- Goal: robust detection of an excess of discordant reads genome-wide.
- Challenging to estimate a background/expected model.

Usage
PopSV flags abnormal regions for further characterization using orthogonal approaches.

Discordant versus concordant reads
- Heterogeneous coverage ⇒ hybrid Poisson-Normal Z-score.
- Targeted normalization from PopSV on concordant reads.
PopSV and BreakDancer

BreakDancer: SV caller using paired-end mapping information (Chen, 2009).
Conclusion

PopSV: Robust and sensitive approach

- Superior to other Read-Depth methods.
- Wider range of the genome tested.
- Detection in low mappability regions and partial tumoral signal.

Work in progress

- More than an CNV caller.
  - Excess of discordant read pairs.
  - Combination with orthogonal approaches (PEM, Assembly).

- Custom binning: repeat annotation, Whole-Exome Sequencing.
Acknowledgment

- Guillaume Bourque
- Mathieu Bourgey
- Louis Letourneau
- Francois Lefebvre
- Eric Audemard
- Toby Hocking
- Simon Girard
- Simon Gravel
- Mathieu Blanchette
- Mehran Karimzadeh Reghbati
Thank You!
SNP-array concordance

proportion of SNP–array GS event also in WGS calls

<table>
<thead>
<tr>
<th>Method</th>
<th>Loose</th>
<th>Stringent</th>
</tr>
</thead>
<tbody>
<tr>
<td>PopSV</td>
<td><img src="image1" alt="Boxplot" /></td>
<td><img src="image2" alt="Boxplot" /></td>
</tr>
<tr>
<td>FREEC</td>
<td><img src="image3" alt="Boxplot" /></td>
<td><img src="image4" alt="Boxplot" /></td>
</tr>
<tr>
<td>cn.MOPS</td>
<td><img src="image5" alt="Boxplot" /></td>
<td><img src="image6" alt="Boxplot" /></td>
</tr>
</tbody>
</table>
Copy-number distribution

![Copy-number distribution graph]

- X-axis: Copy number estimate
- Y-axis: Number of events

The graph shows the distribution of copy number estimates, with peaks indicating the frequency of copy number events across different estimate values.
PCA vs Targeted normalization in tumor samples
PopSV and BreakDancer

BreakDancer: SV caller using paired-end mapping information (Chen, 2009).