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

Jean Monlong

Guillaume Bourque’s group

Canadian Human and Statistical Genetics Meeting
April 19, 2015

McGill
Human Genetics Dept.
Structural variation

Genetic variation involving more than 500bp.

- Duplication
- Deletion
- Inversion
- Translocation

Raphael Lab, Brown University.

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

Mappability issues

- Noisy or reduced signal in repeat-rich regions, centromeres, telomeres.
- Unpredictable segmentation → 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.
Application

CageKid : Renal Cell Carcinoma
- Whole-Genome Sequencing of 100 individuals.
- Normal and tumor paired samples.
- Reference samples : normal samples.

Twin family dataset
- Whole-Genome Sequencing of 45 individuals.
- 10 families (2 parents + 2 twins).
- Reference samples : all the samples.

~ 40X coverage, Illumina paired-end 100bp
Example: Partial signal supporting tumoral deletion

Chr.1, overlapping CDC14A gene (cell division cycle), not detected by other approaches.
Evaluating PopSV performance

Germline events detected in tumor samples

Results

PopSV detected **more consistent calls** than other methods with **similar specificity**.
Other validation and benchmark

- Consistent with SNP-array calls?
- Twin dataset: concordant between twins?
- Concordant calls when using different bin sizes?

For more details/discussion come see **poster 30** tomorrow!
**Twin dataset: PopSV on normal genomes**

16-fold enrichment in low coverage regions.
Twin dataset: PopSV calls in low coverage regions
Conclusion

PopSV has been applied to
- Whole-Genome Sequencing of normal genomes.
- Whole-Genome Sequencing of tumor genomes.
- Whole-Exome Sequencing data.

PopSV robustly detects
- variants in high and low mappability regions
- variants with partial signal (e.g. in tumors).

R package available at github.com/jmonlong/PopSV.

Future direction
- Other types of SVs as excess of discordant read pairs.
- Combination with orthogonal approaches (PEM, Assembly).
Acknowledgment

- Guillaume Bourque
- Mathieu Bourgey
- Louis Letourneau
- Francois Lefebvre
- Eric Audemard
- Toby Hocking
- Simon Gravel
- Mathieu Blanchette
- Simon Girard
- Guy Rouleau
- Michel Boivin
Thank You!
Low-mappability regions overlap functional elements

- mappability:
  - high: 72%
  - low: 21%
  - no: 7%

- number of bins:
  - 0
  - 50000
  - 100000
  - 150000
  - 200000

- mappability of the bin:
  - 0.00
  - 0.25
  - 0.50
  - 0.75
  - 1.00

- proportion overlapping at least one bin:
  - lincRNA
  - other
  - protein_coding
  - pseudogene

- mappability:
  - high
  - low
  - no
Unknown technical bias

![Graphs showing inter-sample mean coverage and standard deviation for WGS, simulated, and shuffled samples.](image)
PopSV: importance of normalization

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

- PCA-based normalization (Krumm, 2012; Boeva, 2014).
- Targeted normalization: linear using a subset of the genome.
PopSV: Z-score and test

For a sample $s$:

- For each bin $b$: $z = \frac{BC_s^b - BC_{\text{reference}}^b}{sd_{\text{reference}}^b}$

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

“funky snowman” plot
Z-scores: contamination detection
Example: Telomeric region

Chr.10, overlapping genes (PRAP1, CALY), not detected by other approaches.
Example: NAHR candidate

- **chr20**
- **genes**: RP4−576H24.4, SIRPB1, SIRPD, SIRPG
- **repeats**: DNA, LINE, Low_complexity, LTR, Simple_repeat, SINE
- **bin count**
- **PopSV**

### Coverage
- **normalized coverage**

### PopSV samples
- **nb samples**

### Deletion and Duplication
- **group**: normal, tumor
500bp Z-scores within 10kb calls
500bp Z-scores within 10kb calls

![Graph showing median Z-scores in 500bp bins and number of calls within different ranges. The x-axis represents the median Z-score in 500bp bins, and the y-axis represents the Z-score in 10kb calls. The graph uses different colors to indicate the number of calls: (0,1], (1,10], (10,50], (50,100], (100,1e+03], and (1e+03, Inf).]
SNP array methods concordance

D000G0L 0.636 0.161 0.899 tumor

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS−strict</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS−loose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS−strict</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS−loose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS−strict</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS−loose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS−strict</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS−loose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS−strict</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS−loose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCAT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CN
0
1
2
3
4
5
6
7
8

start
0.0e+00 5.0e+07 1.0e+08 1.5e+08 2.0e+08 2.5e+08
SNP array concordance

Proportion of SNP–array GS event also in WGS calls

- PopSV
- FREEC
- cn.MOPS

loose

stringent

0.00 0.25 0.50 0.75 1.00

proportion of SNP–array GS event also in WGS calls

29 / 14
Twin concordance

![Graph showing the number of concordant calls per twin pair for different methods (cn.MOPS, FREEC, PopSV) across all events, large events, and low mappability events. The x-axis represents the number of concordant calls, ranging from 0 to 200, and the y-axis represents the methods. The box plots show the distribution of concordant calls for each method in each event category.](image-url)
Twin concordance

The diagram shows the proportion of concordant calls per twin pair for different methods and event types. The x-axis represents the proportion of concordant calls, ranging from 0.5 to 1.0. The y-axis indicates different methods: PopSV, FREEC, and cn.MOPS. The data is categorized by event type, with separate sections for "all events," "large events," and "low mappability." The box plots illustrate the distribution of concordant calls for each method within these categories.
Twins and clustering quality

- Rand index using pedigree information

- Method:
  - cn.MOPS
  - FREEC
  - PopSV

- Clustering linkage:
  - average
  - complete
  - Ward

- Number of groups derived from CNV clustering

0.00 0.25 0.50 0.75 1.00

0 10 20 30 40
Many variants in low coverage regions

coverage in reference
- low
- expected
- high

mean coverage in reference samples (normalized read counts)

number of bins
Many variants in low coverage regions

![Graph showing number of calls against mean coverage in reference samples (normalized read counts). The x-axis represents the mean coverage in reference samples, ranging from 0 to >3000, and the y-axis represents the number of calls, ranging from 0 to 6000. The graph includes three categories: low, expected, and high coverage, with the 'low' category showing a peak at low coverage values.]
Twins dataset: copy number estimation

estimated copy number = \( 2 \times \frac{\text{coverage in sample}}{\text{coverage in reference samples}} \)

coverage in reference samples
- low
- expected
- high

number of calls
estimated copy number

0 1 2 3 4 >5

estimated copy number = \( 2 \times \frac{35}{14} \)
Many calls in segmental duplications but also in genes
Distance to centromere/telomere/gaps

More SV detected near centromere/telomere/gaps.
Mappability

![Graph of mappability with cumulative event proportion against proportion of mappable sequence. The graph shows different types of events: ambiguous, deletion, duplication, DGV, and random. Each type is represented by a different line color. The x-axis represents the proportion of mappable sequence, and the y-axis represents the cumulative event proportion. The graph illustrates how different types of events affect mappability across the sequence proportion.]