

Population-based detection of Structural Variants in normal and aberrant genomes.

Jean Monlong, PhD2

Guillaume Bourque's group

Research Day - June 5, 2014



McGill

Human Genetics Dept.

What is structural variation ?

Genetic variation involving more than 500bp.



Reference



Deletion



Insertion



Inversion



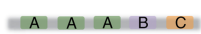
Tandem duplication



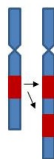
Dispersed duplication



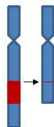
Copy-number variant



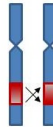
Duplication



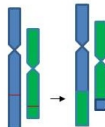
Deletion



Inversion



Translocation



Baker 2012, Nature Methods. Raphael Lab, Brown University.

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

Why is it important ?

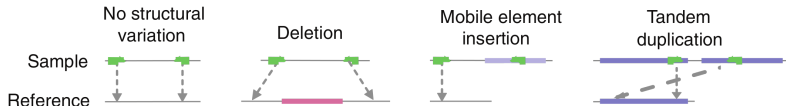
- ▶ Major role in evolution.
- ▶ Population Genetics: widespread variation across humans.
- ▶ Association with diseases and cancer.

SV detection using High-Throughput Sequencing

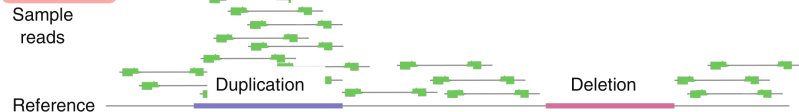
- ▶ Sample is sequenced.
- ▶ Reads are mapped to the reference genome.
- ▶ Unexpected patterns could be explain by presence of SVs.

SV detection using High-Throughput Sequencing

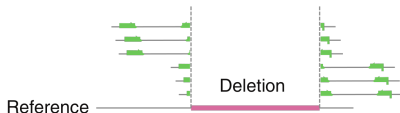
Read pairs



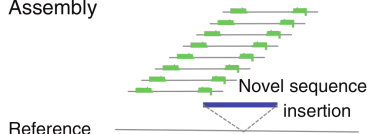
Read depth



Split reads



Assembly

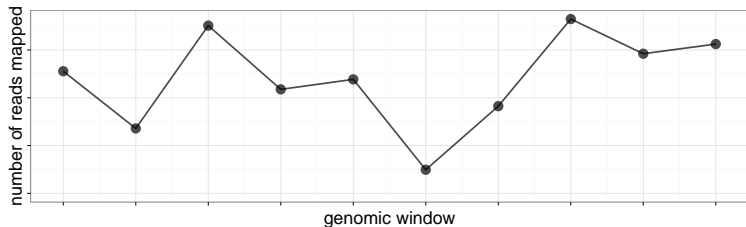
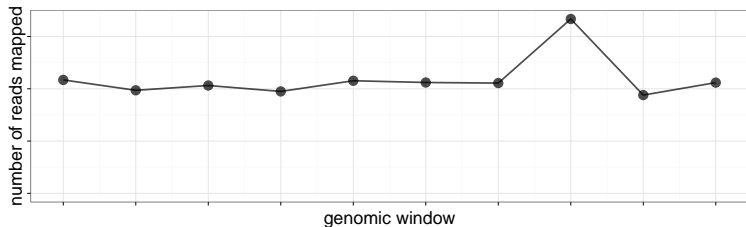


Baker 2012, Nature Methods.

Limitation

Low mappability

- ▶ Noisy or reduced signal in repeat-rich regions, centromeres, telomeres.
- ▶ Unpredictable segmentation → reduced sensitivity/specificity.
- ▶ Filtering problematic regions reduces the genome range tested.

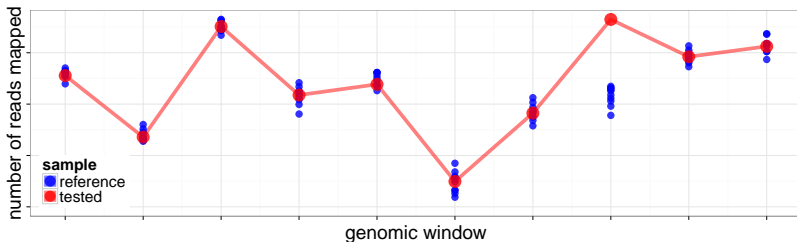


Objective

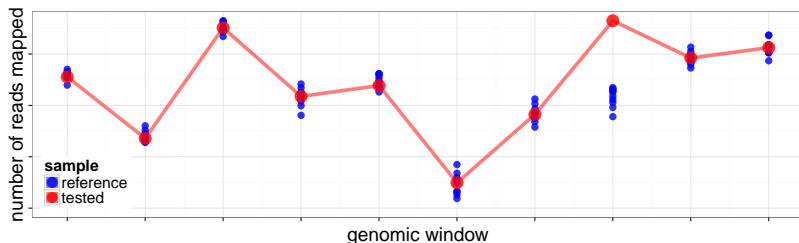
Test the entire genome, including low-mappability regions, and detect subtle abnormal coverage.

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.

CageKid : Renal Cell Cancer

Whole-Genome Sequencing of 100 individuals, $\sim 40X$ coverage, Illumina paired-end 100bp, normal and tumor paired samples.

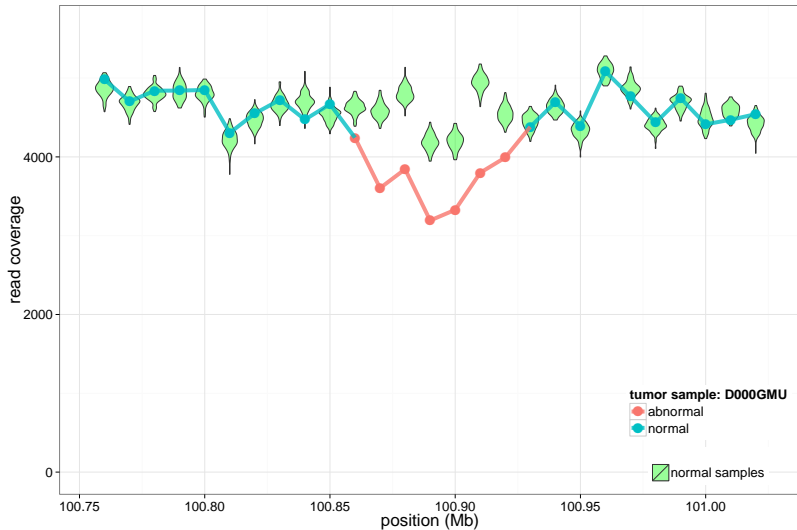
- ▶ Normal samples \rightarrow reference samples.
- ▶ 10kb bins.
- ▶ Only properly paired and mapped read pairs.

Validation and benchmark

- ▶ Germline events detected in tumor samples ?
- ▶ Concordant with SNP-array calls ?
- ▶ Twin dataset: concordant with the pedigree ?
- ▶ Concordant when using different bin sizes ?

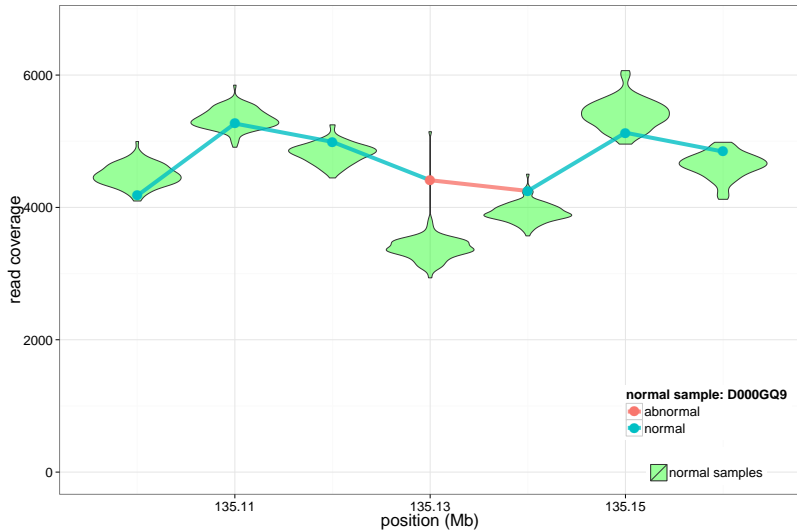
PopSV detected more concordant calls than other methods.

Example: Partial tumoral event



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

Example: Telomeric region



Chr.10, overlapping genes (PRAP1, CALY), not detected by other approaches.

PopSV flexibility

Custom binning: repeat annotation

- ▶ Increased resolution in regions of interest.
- ▶ Promising results: enrichment in centromere/telomere.

Counting discordant reads

- ▶ Detect excess of discordant reads.
- ▶ Promising results, including on repeats.

Conclusion

Robust and sensitive approach

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

Work in progress

- ▶ Explore results and application to other projects (e.g. Pan-Cancer Analysis of Whole Genome).
- ▶ Custom binning: repeat annotation, Whole-Exome Sequencing.
- ▶ More than an CNV caller.
 - ▶ 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
- ▶ Mehran Karimzadeh Reghbaty



CIHR IRSC

Canadian Institutes of
Health Research

Instituts de recherche
en santé du Canada



Genome Québec