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

Jean Monlong

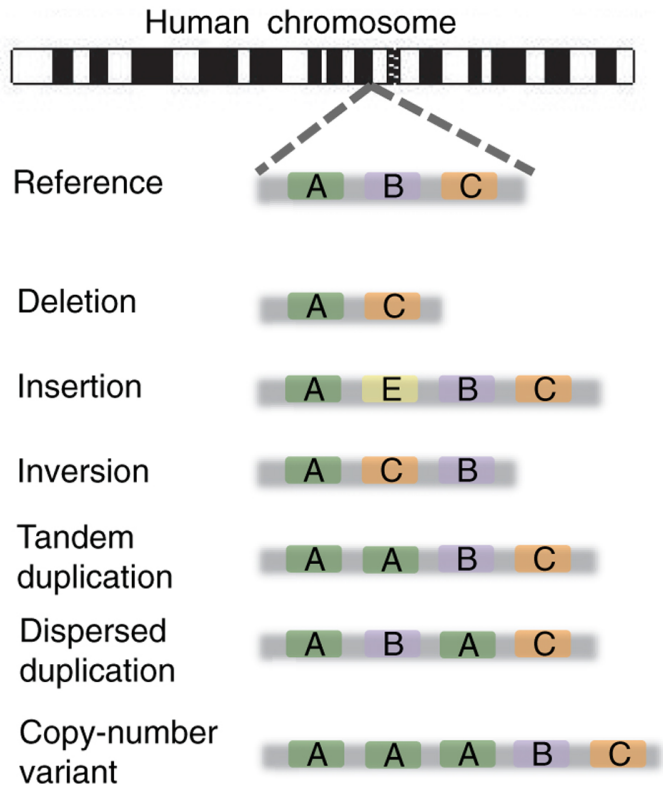
Guillaume Bourque's group

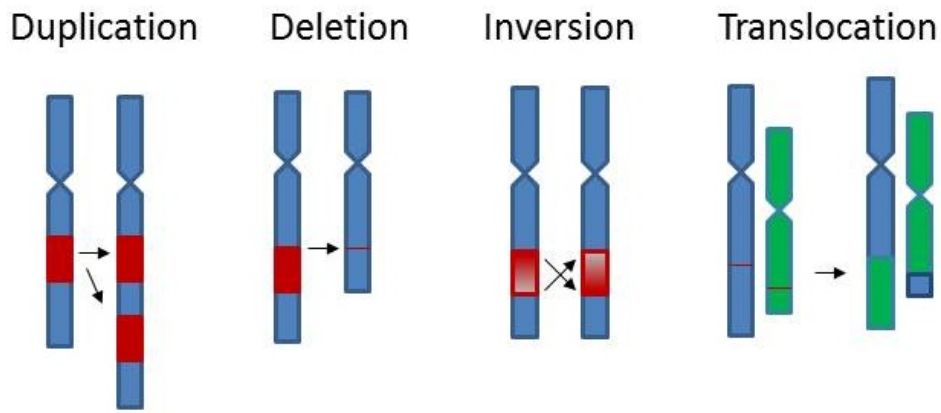
Genome Informatics - September 21-24, 2014

Human Genetics Dept.

Structural variation

Genetic variation involving more than 500bp.

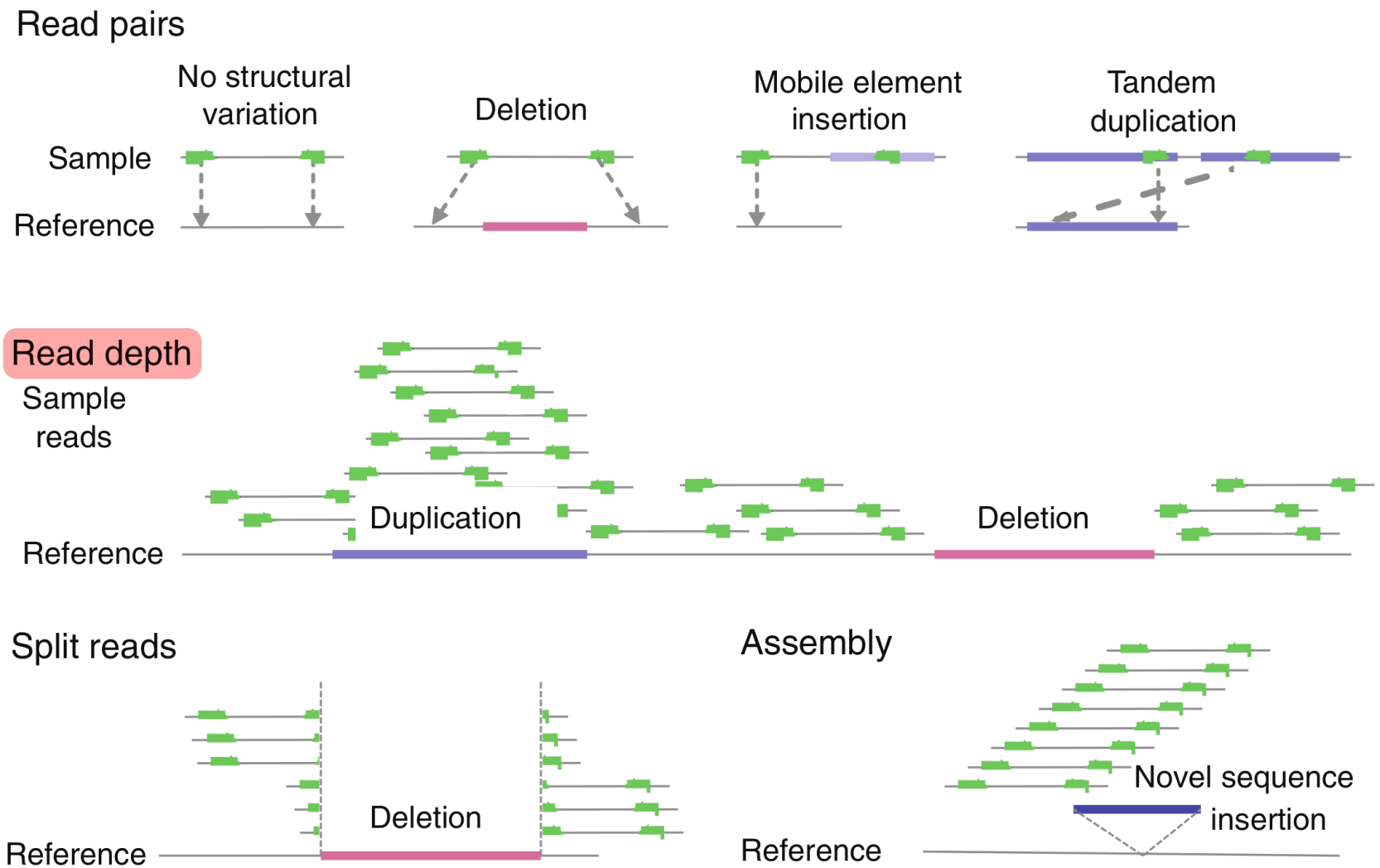




Baker 2012, Nature Methods. Raphael Lab, Brown University.

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

SV detection using High-Throughput Sequencing

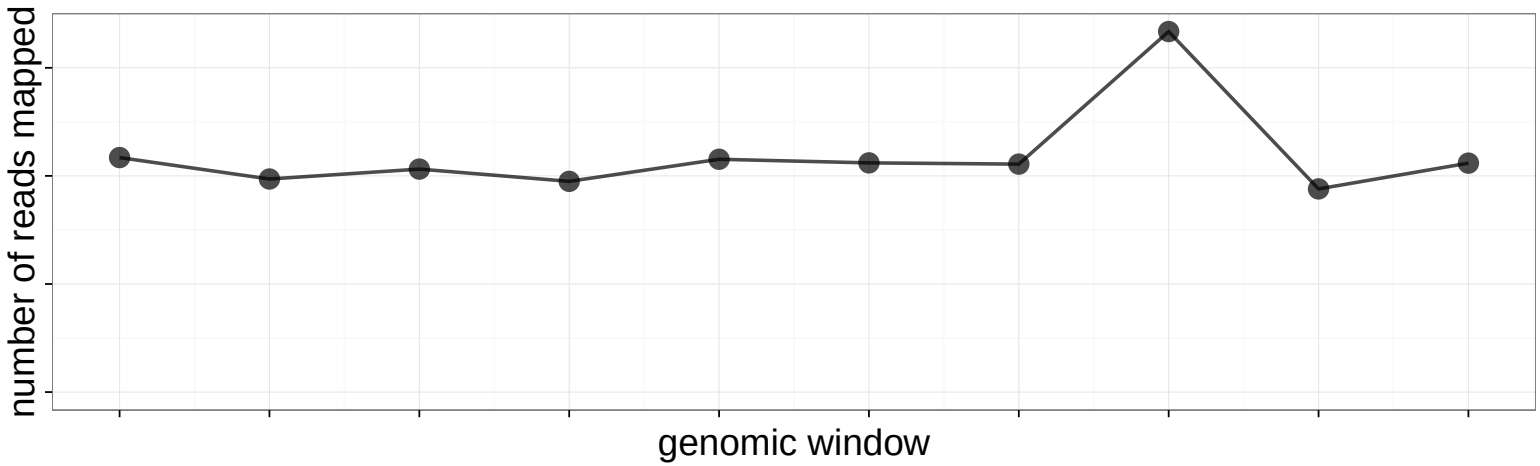


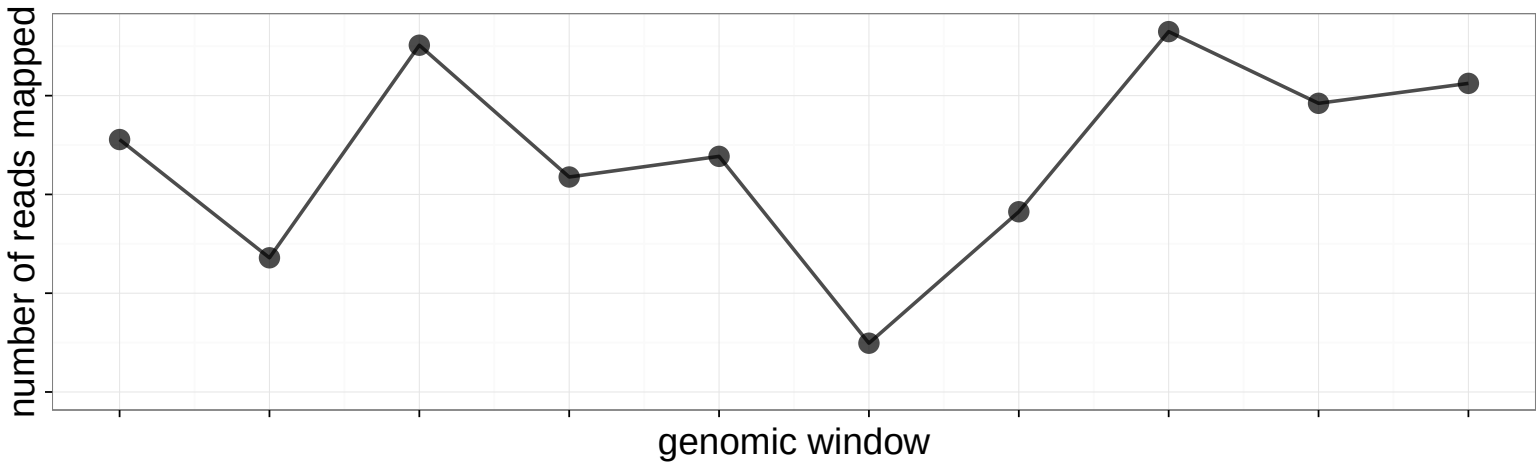
Baker 2012, Nature Methods.

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.



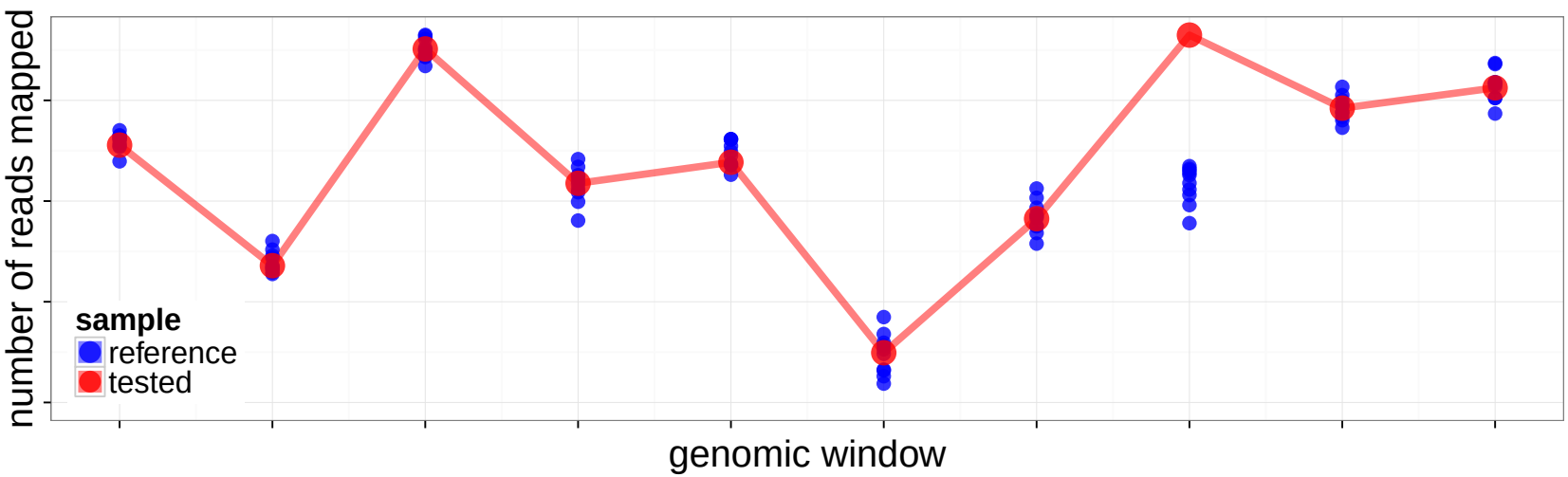


Objective

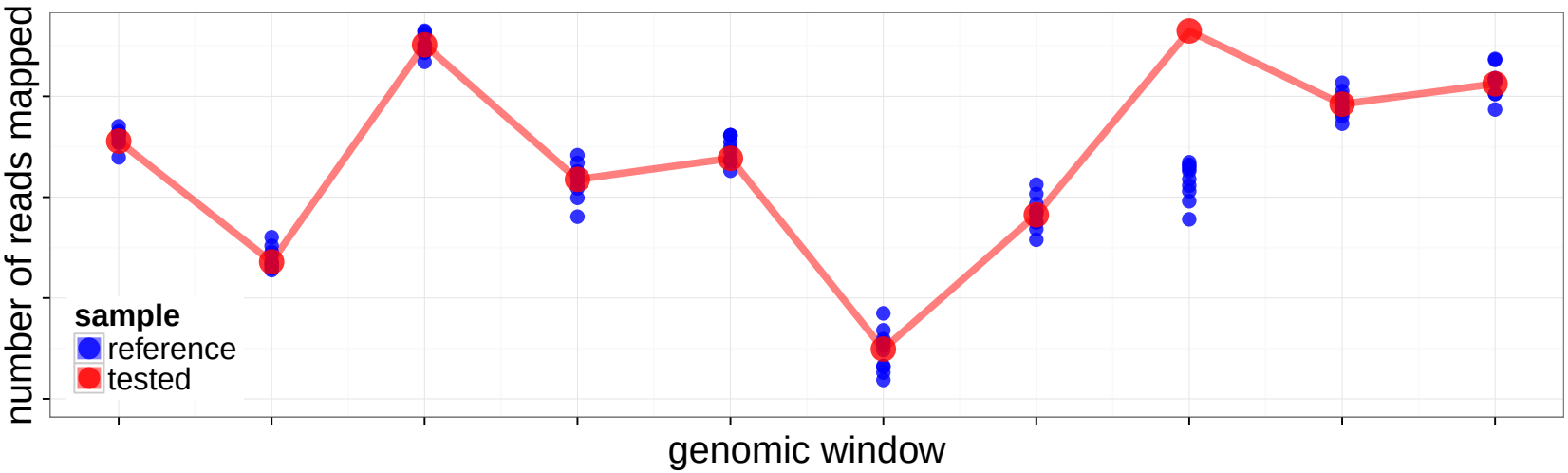
Test the entire genome, including low-mappability regions, and detect subtle abnormal coverage.

PopSV: Population-based approach

Use a set of reference experiments to detect abnormal patterns.



PopSV: Population-based approach

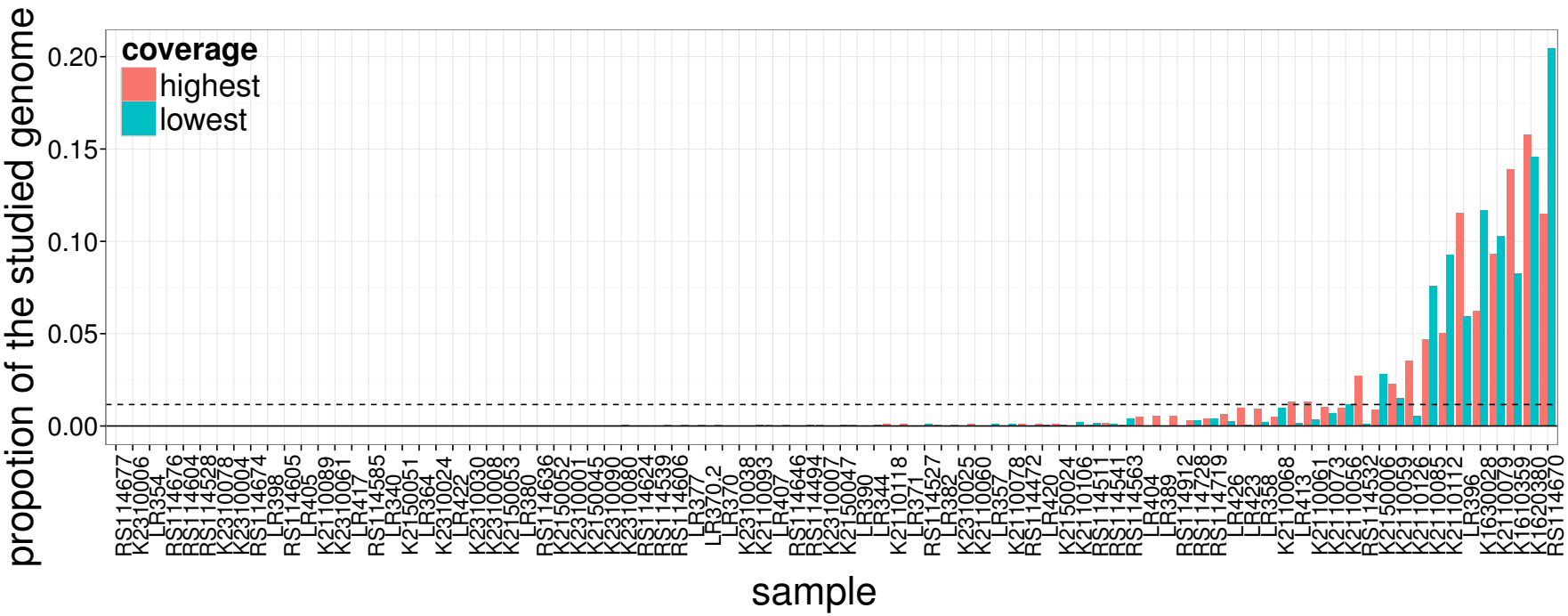


Workflow

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.

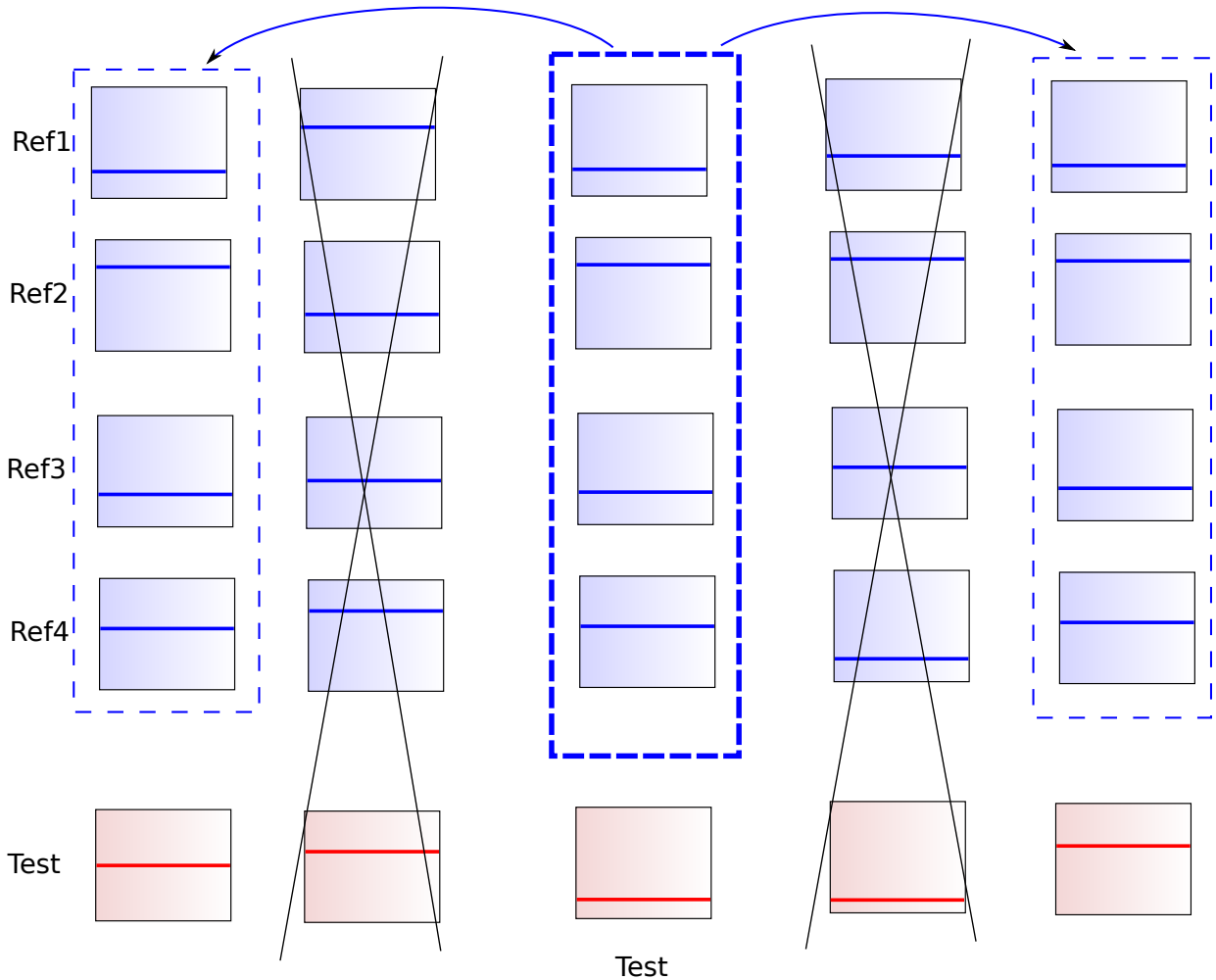
PopSV: importance of normalization

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



PopSV: importance of normalization

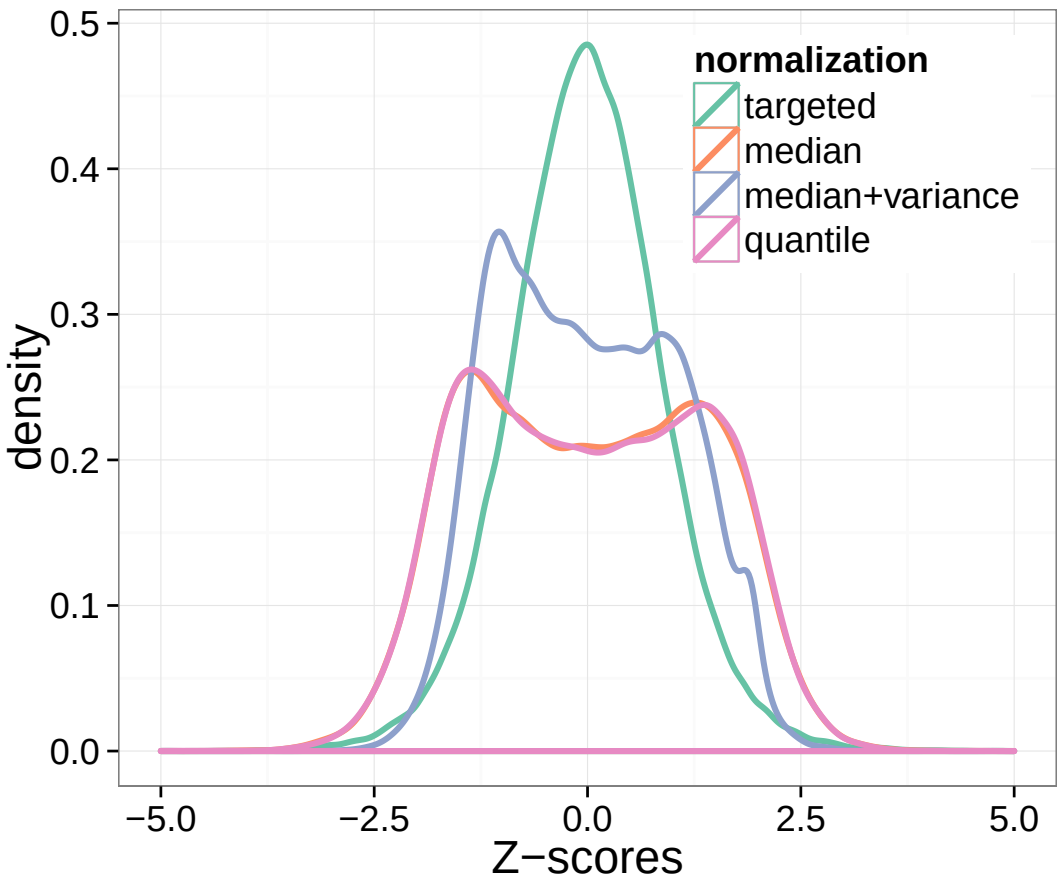
- PCA-based normalization (*Krumm*, 2012; *Boeva*, 2014).
- Targeted normalization: linear using a subset of the genome.



PopSV: Z-score and test

For a sample s:

- For each bin b: $z = \frac{BC_s^b - \overline{BC_{reference}^b}}{sd_{reference}^b}$
- $pv = \mathbb{P}(|z| \leq |Z|)$ with $Z \sim \mathcal{N}(0, \sigma)$ where σ is estimated from the z distribution across all bins.



Application

CageKid : Renal Cell Cancer

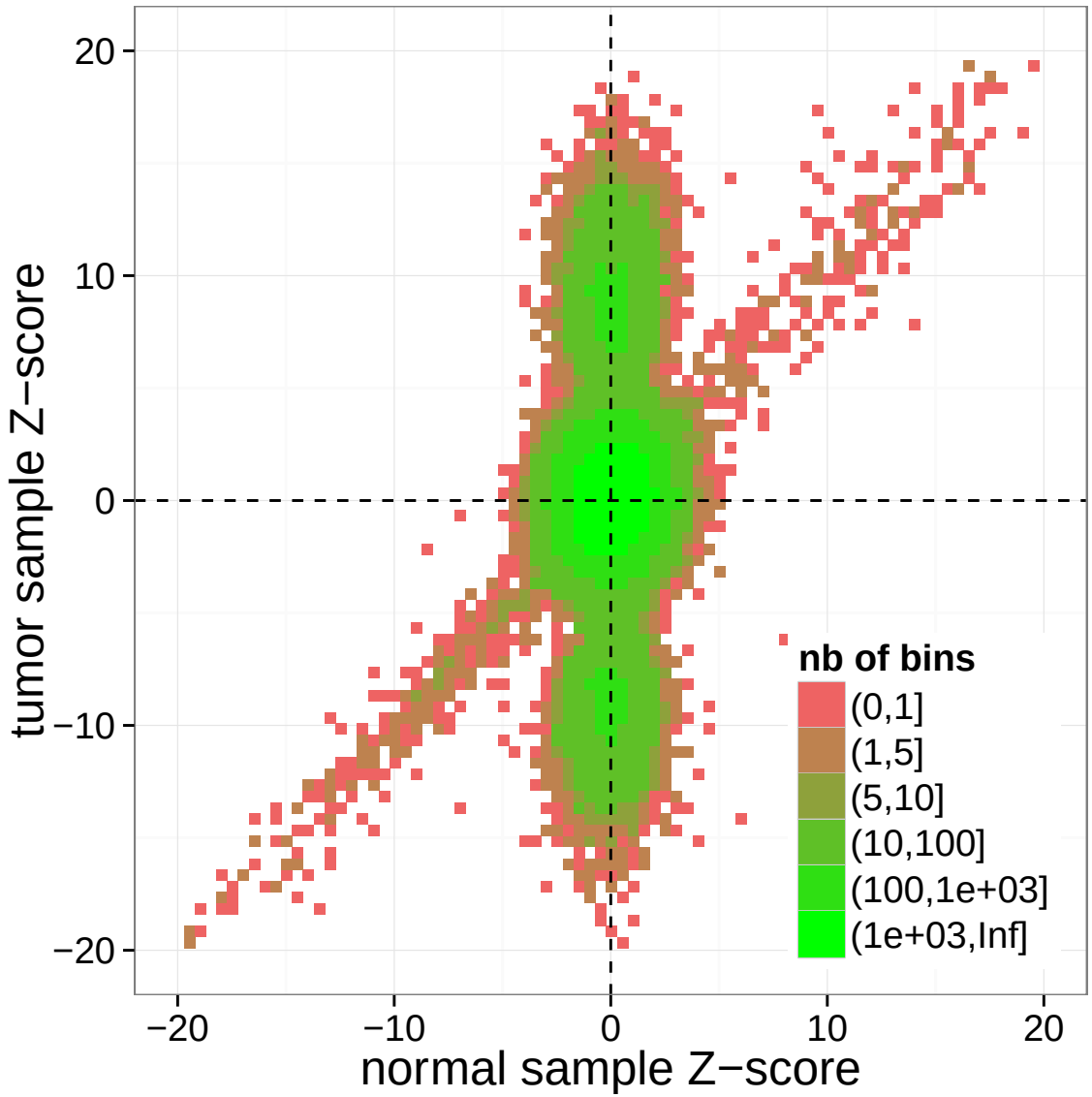
Whole-Genome Sequencing of 100 individuals, $\sim$ 40X coverage, Illumina paired-end 100bp, normal and tumor paired samples.

- Normal samples $\rightarrow$ reference samples.
- 2kb bins.

Read-Depth measure - 2 strategies

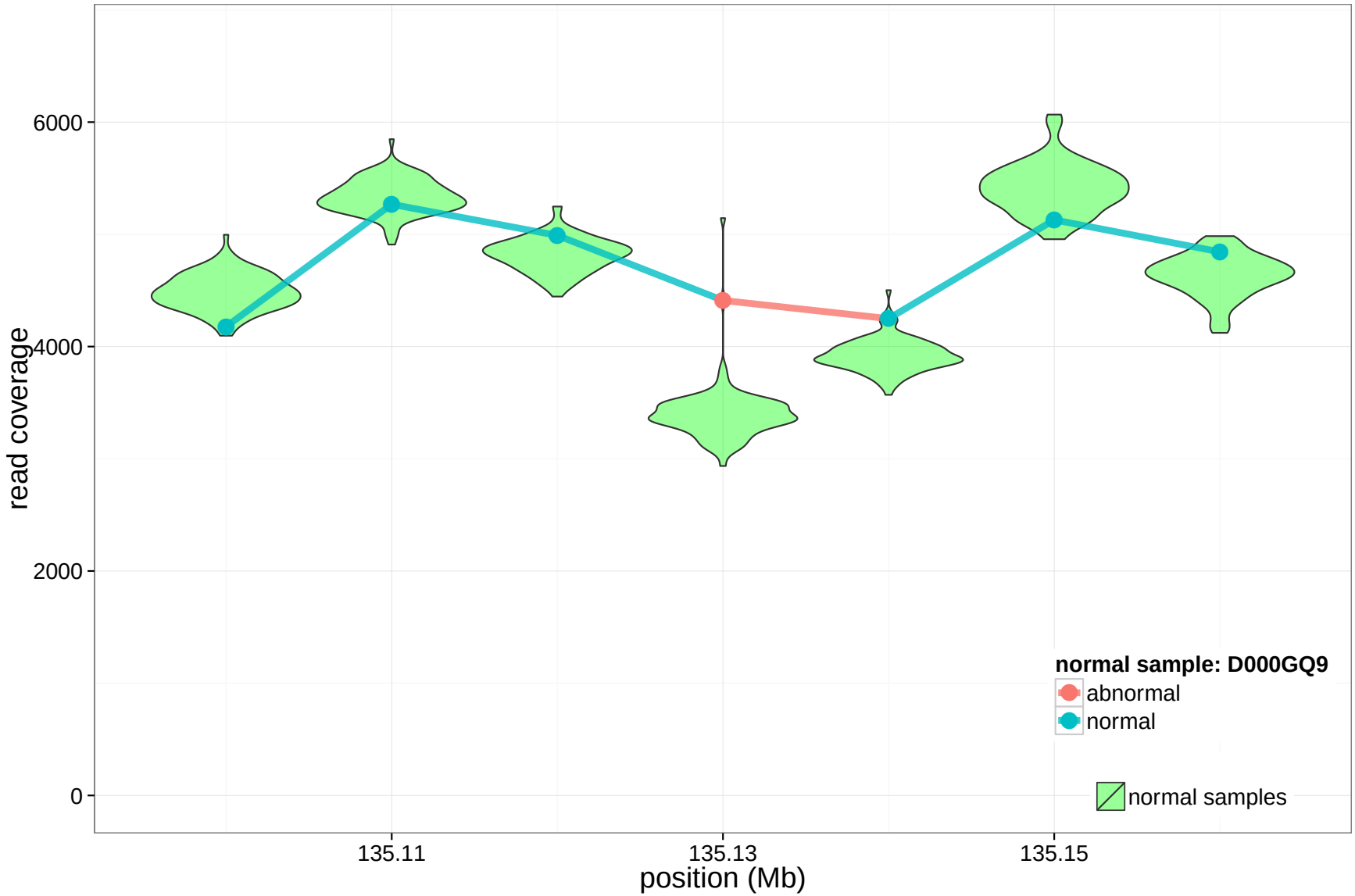
- **concordant reads**: only properly paired and mapped read pairs.
- **discordant reads**: improperly mapped read pairs or low mapping quality.

Using concordant reads



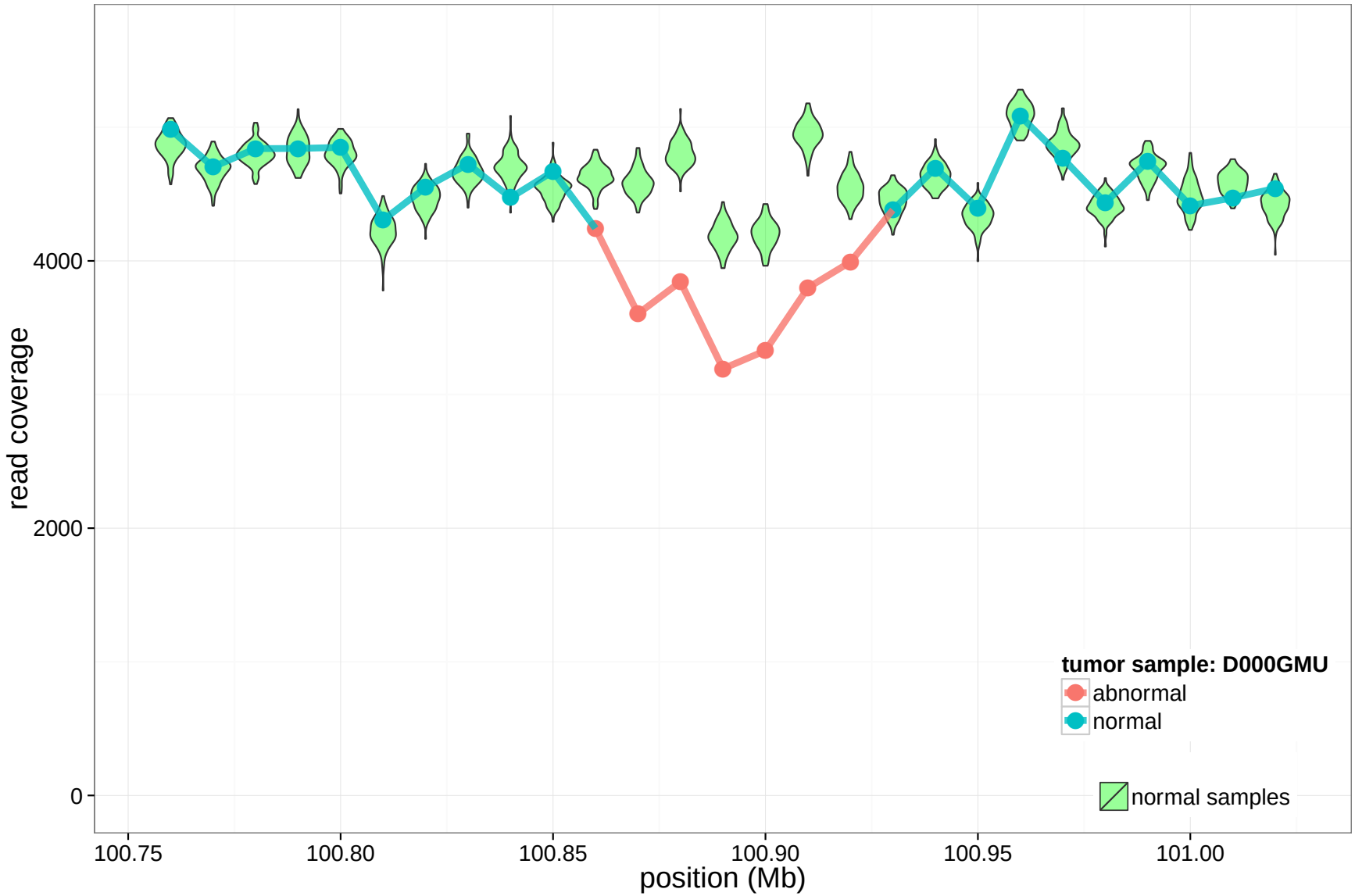
"funky snowman" plot

Example: Telomeric region



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

Example: Partial tumoral event

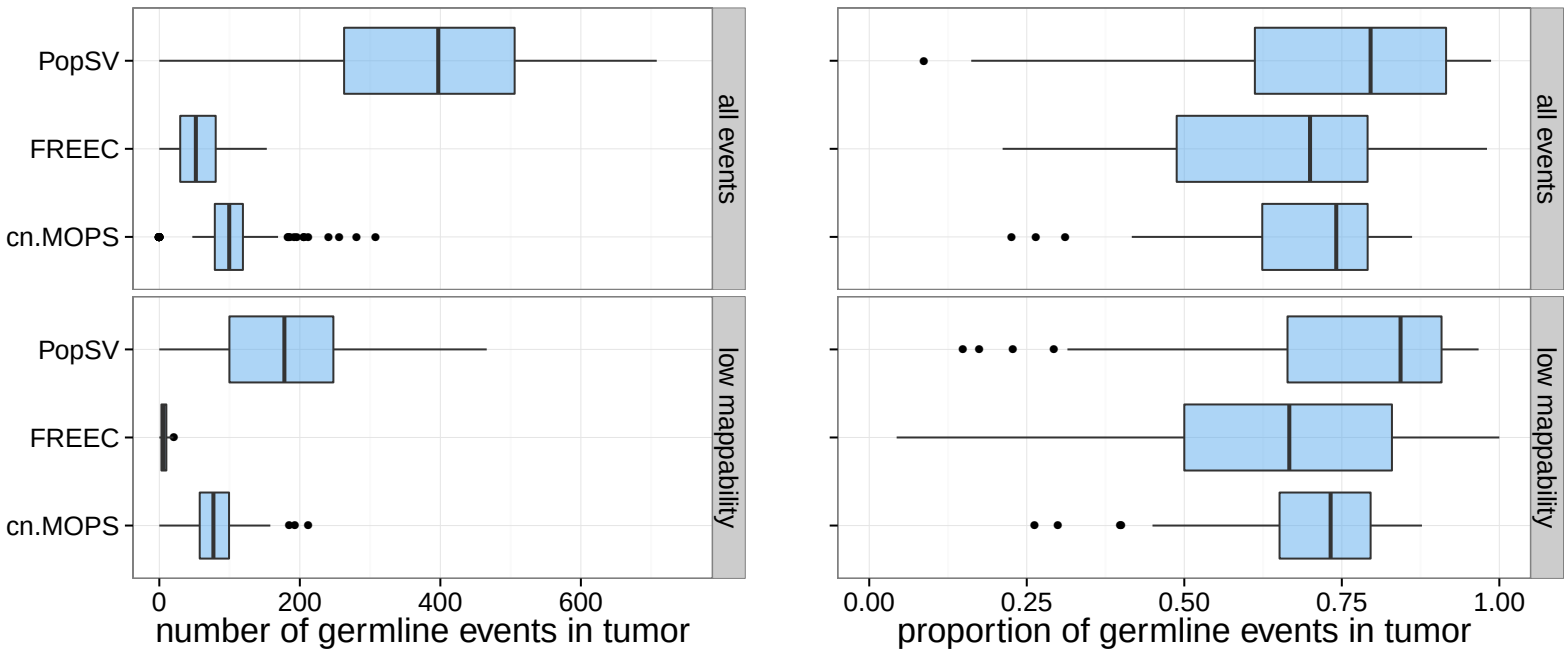


Chr.1, overlapping CDC14A gene (cell division cycle), not detected by other approaches.

Validation and benchmark

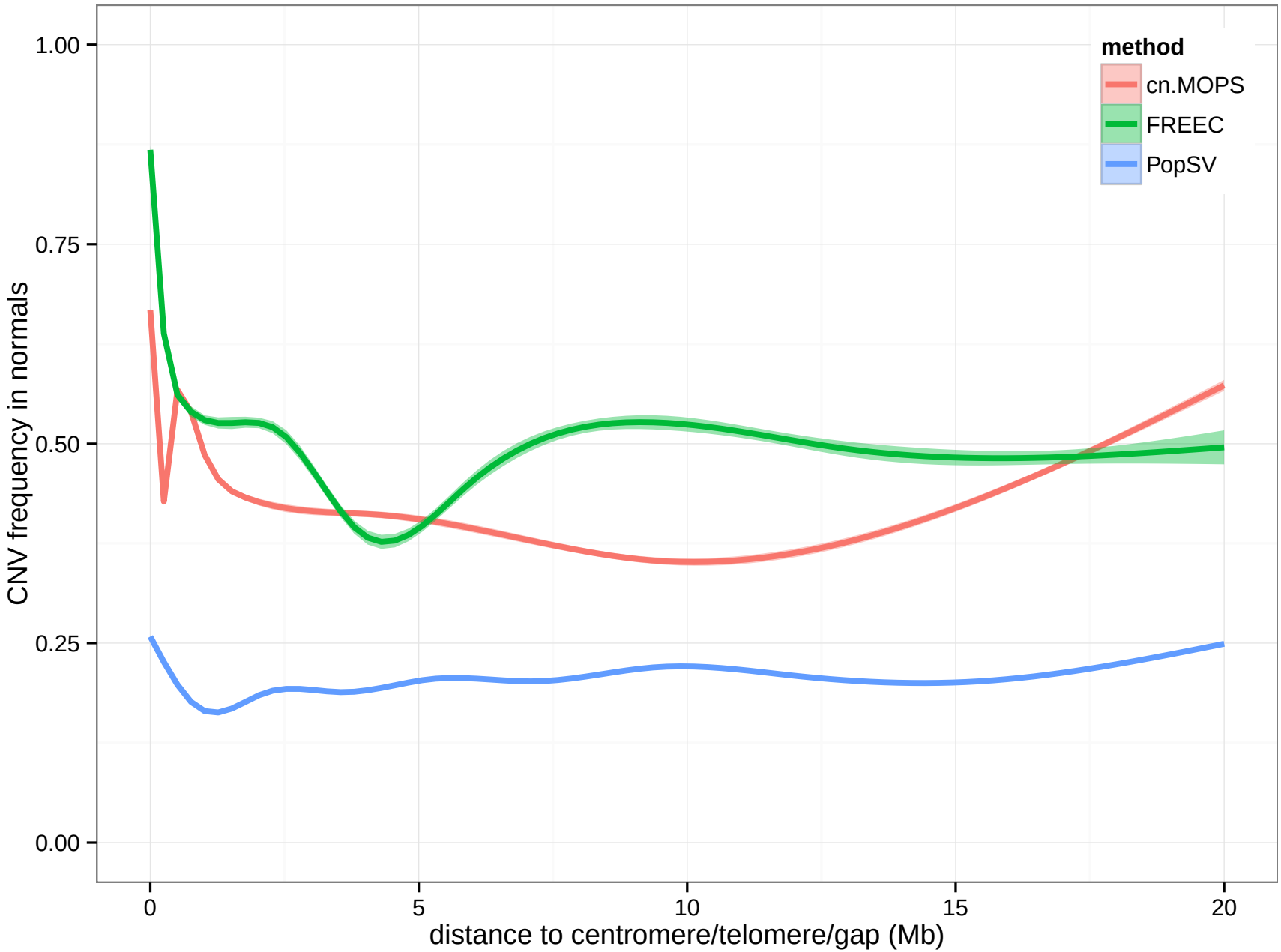
- Germline events detected in tumor samples ?
- Consistent with SNP-array calls ?
- Twin dataset: consistent with the pedigree ?

Germline events detected in tumor samples



PopSV detected more consistent calls than other methods with similar specificity.

Centromere/telomere/gap and systematic errors



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.

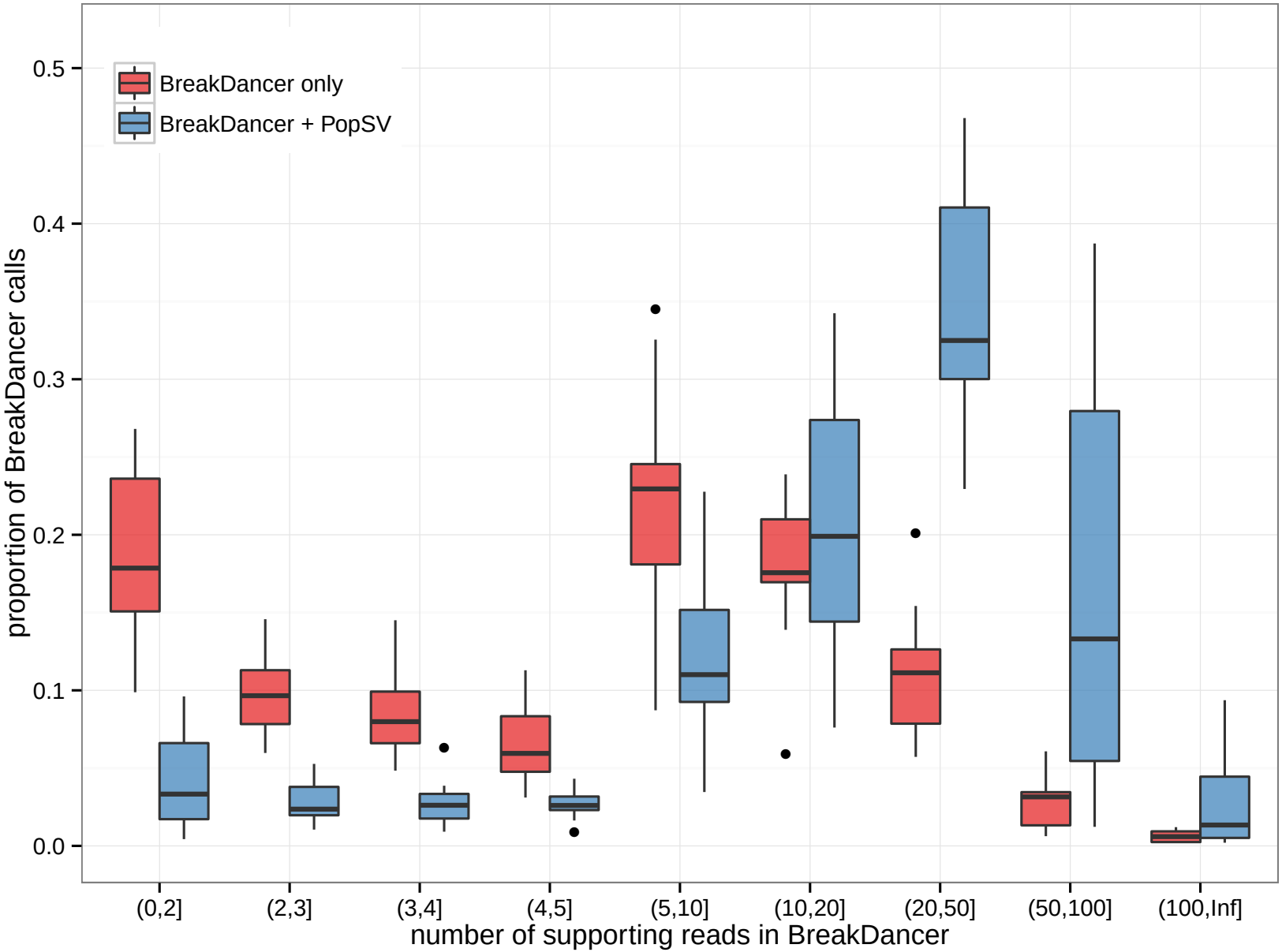
Usage

PopSV flags abnormal regions for further characterization using orthogonal approaches.

Discordant versus concordant reads

- Heterogeneous coverage $\Rightarrow$ 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

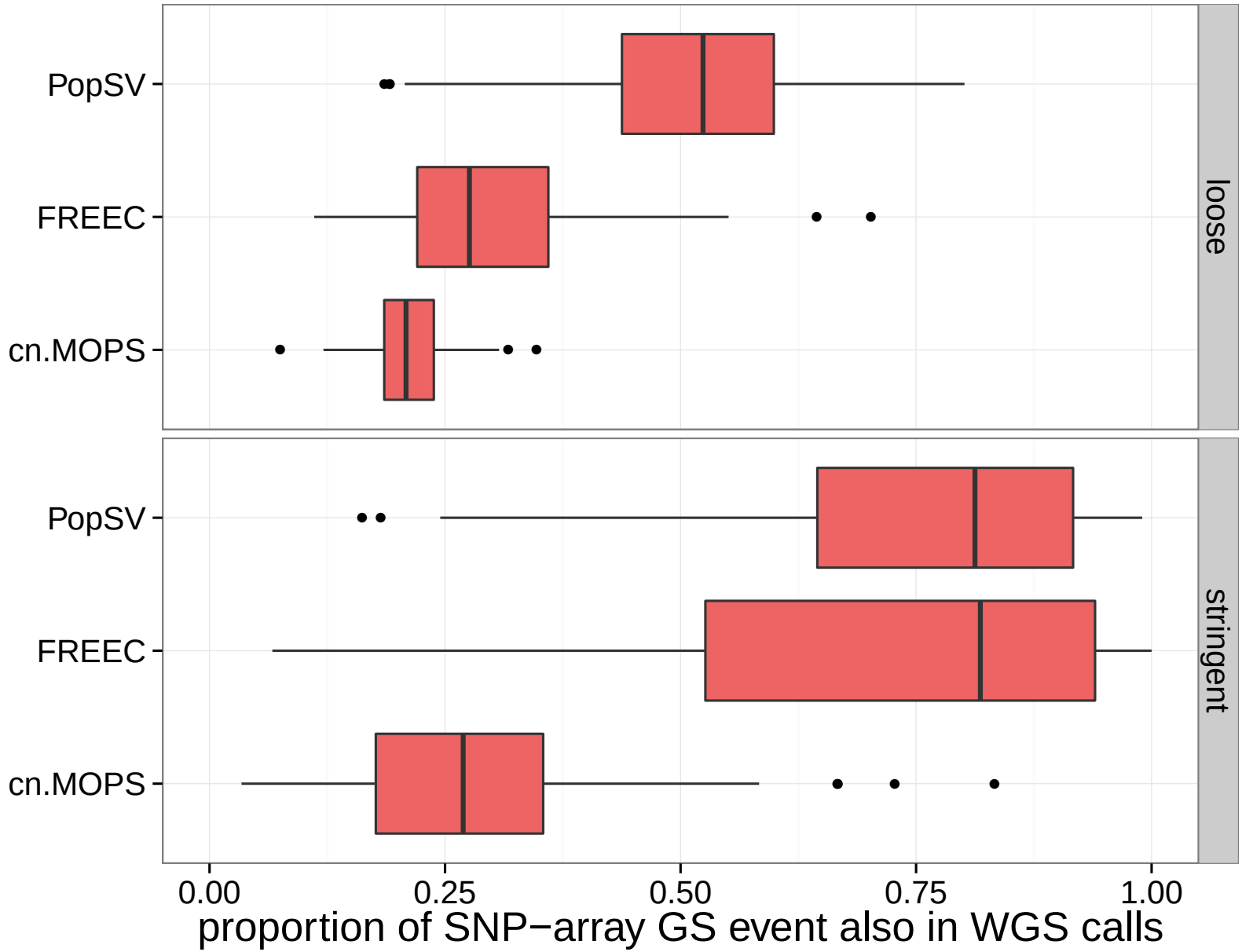
- **Guillaume Bourque**
- Mathieu Bourgey
- Louis Letourneau
- Francois Lefebvre
- Eric Audemard
- Toby Hocking
- Simon Girard
- Simon Gravel
- Mathieu Blanchette
- Mehran Karimzadeh Reghbati

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

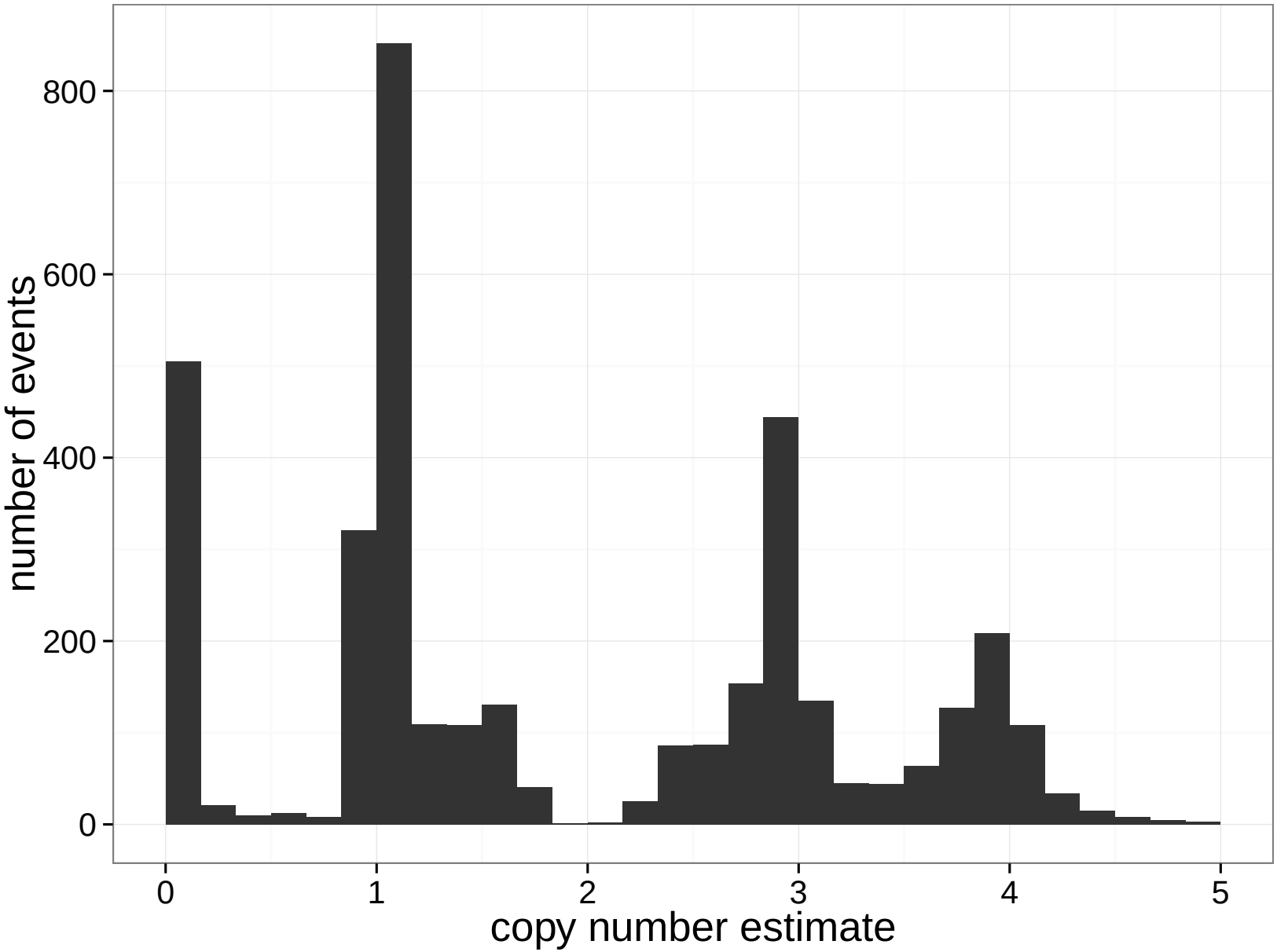
GenomeQuébec

Thank You !

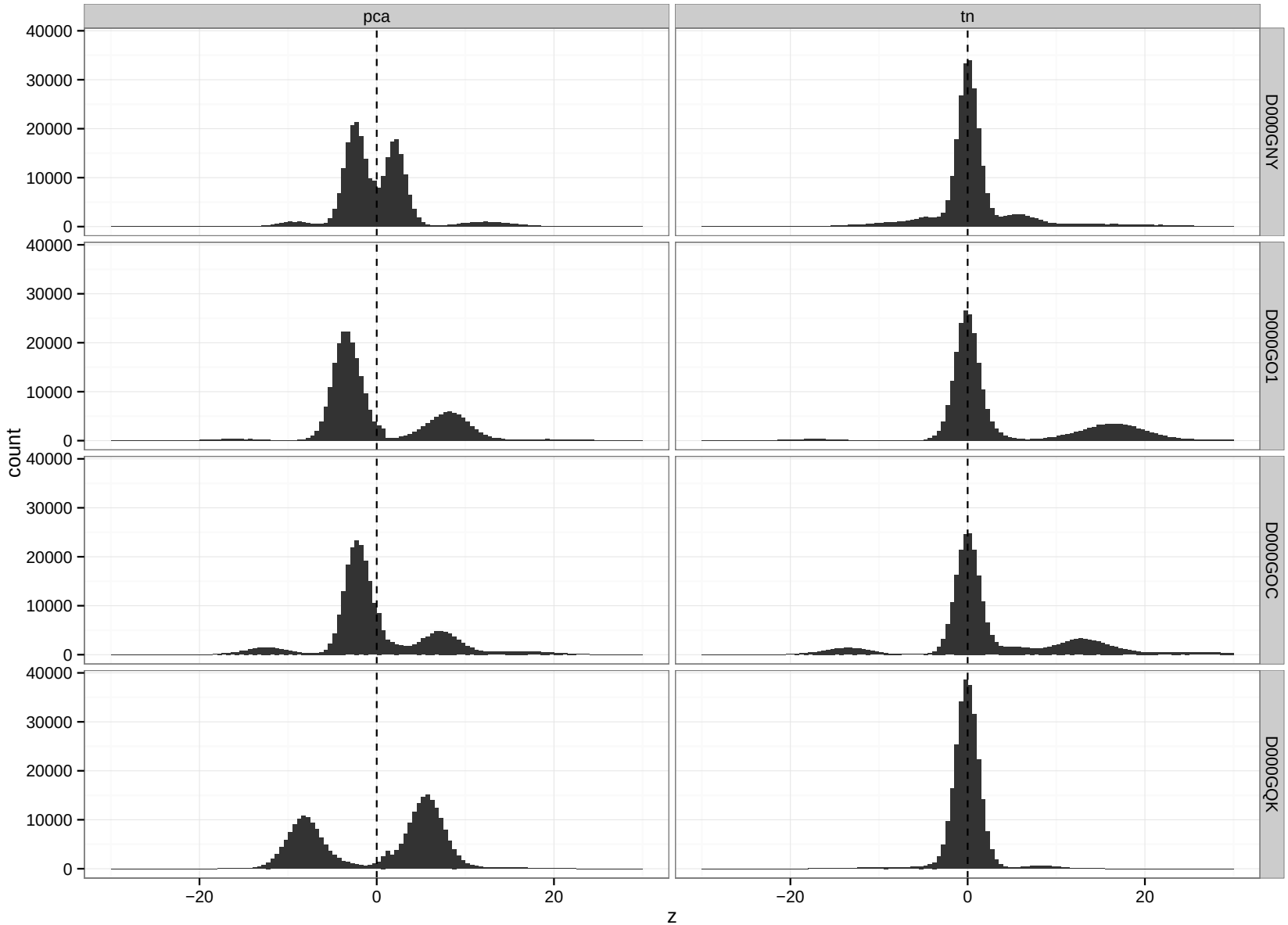
SNP-array concordance



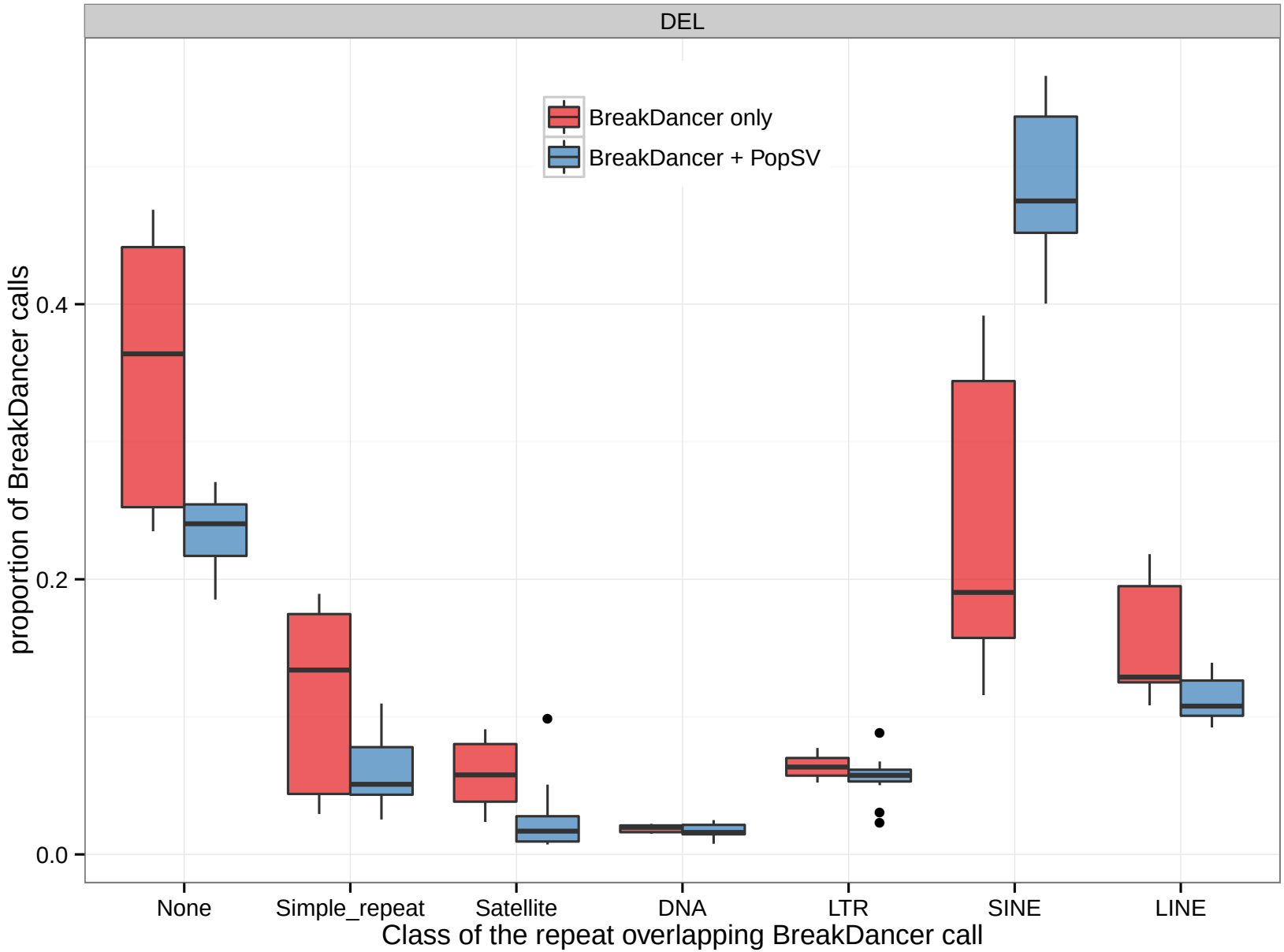
Copy-number distribution



PCA vs Targeted normalization in tumor samples



PopSV and BreakDancer



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