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

Jean Monlong

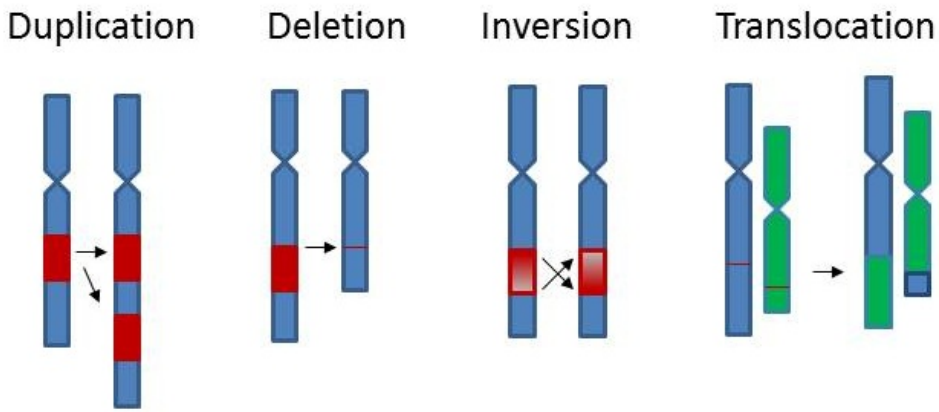
Guillaume Bourque's group

Canadian Human and Statistical Genetics Meeting
April 19, 2015

McGill
Human Genetics Dept.

Structural variation

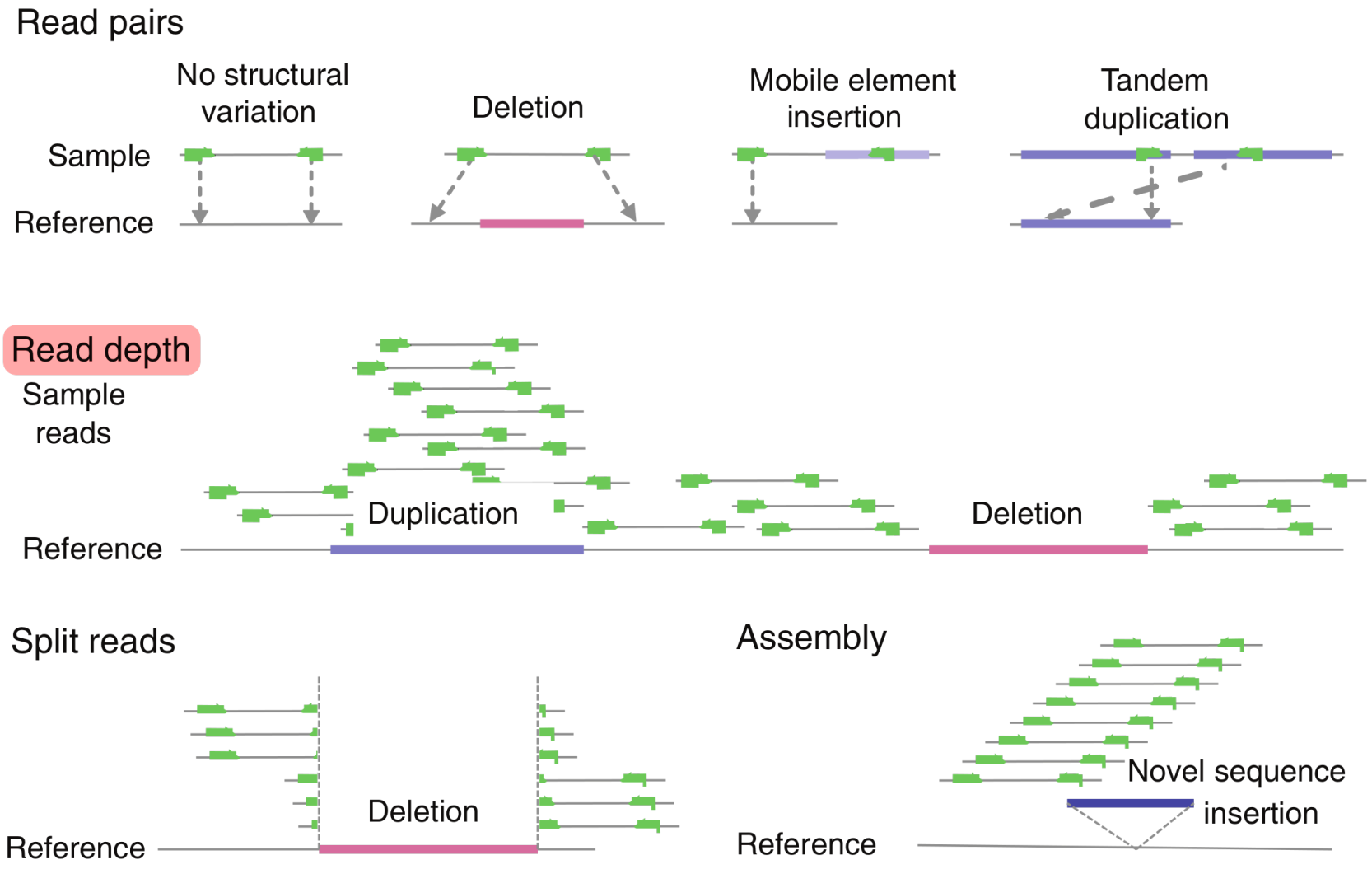
Genetic variation involving more than 500bp.



Raphael Lab, Brown University.

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

SV detection using High-Throughput Sequencing

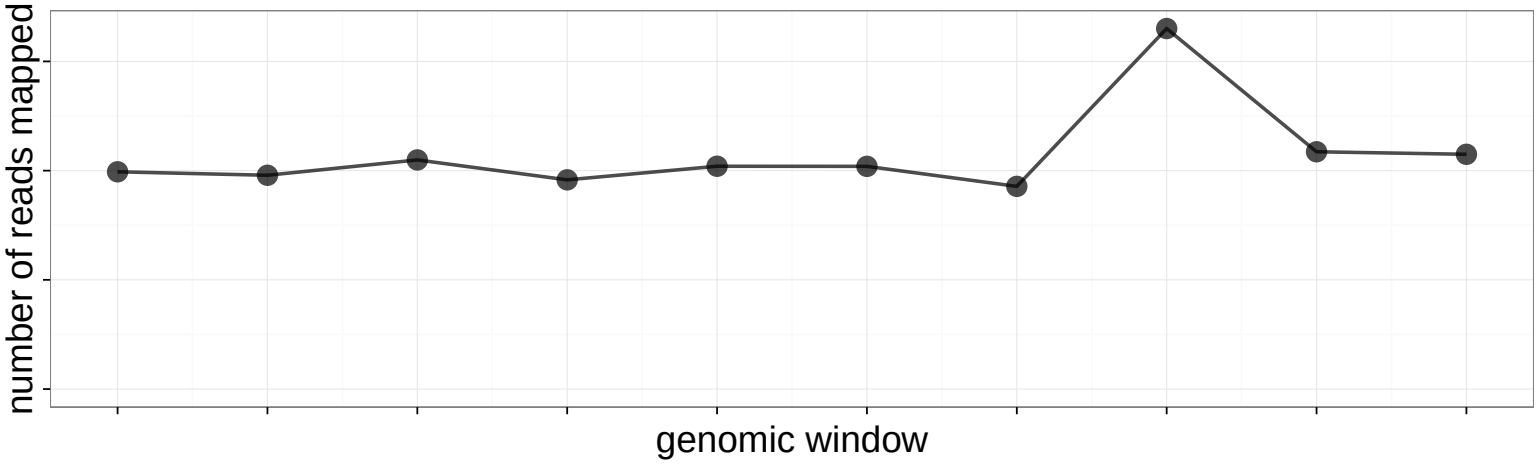


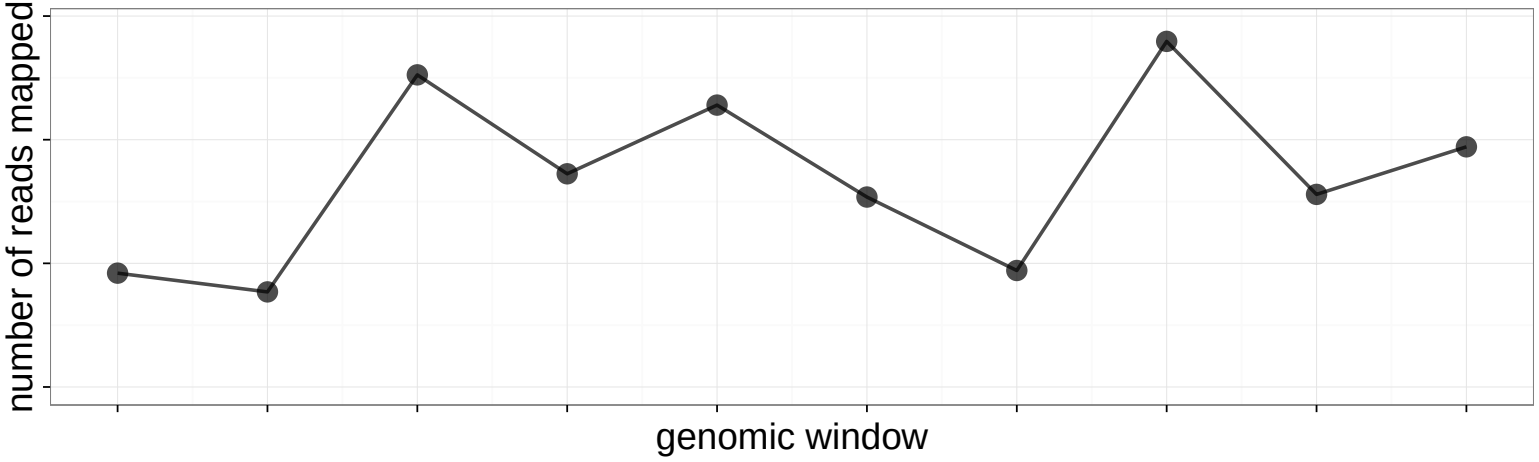
Baker 2012, Nature Methods.

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.



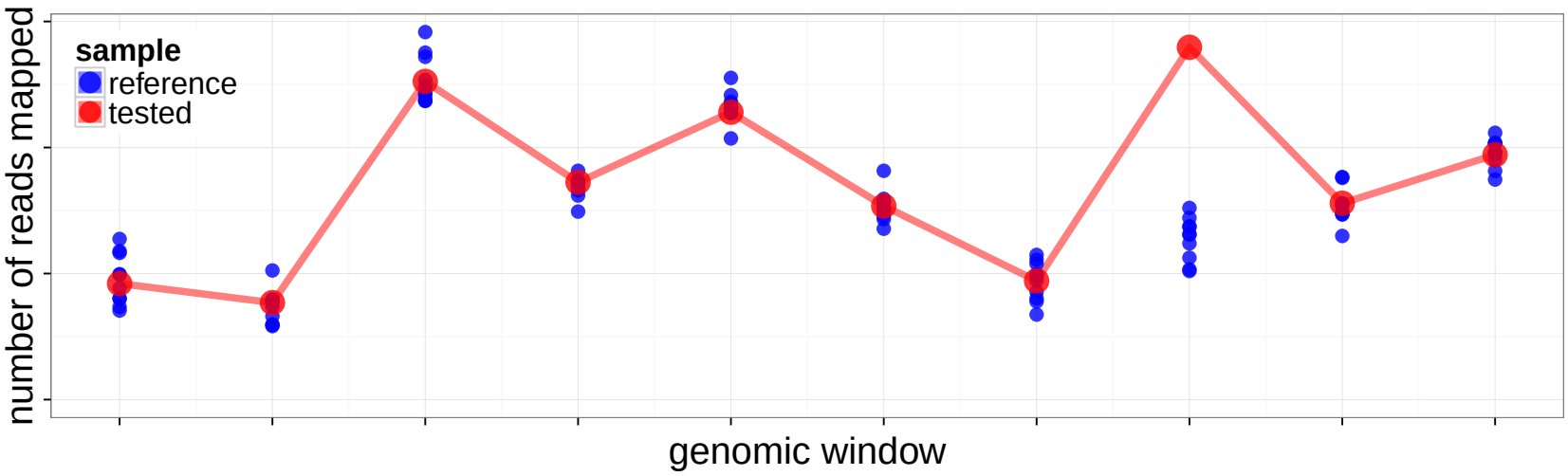


Objective

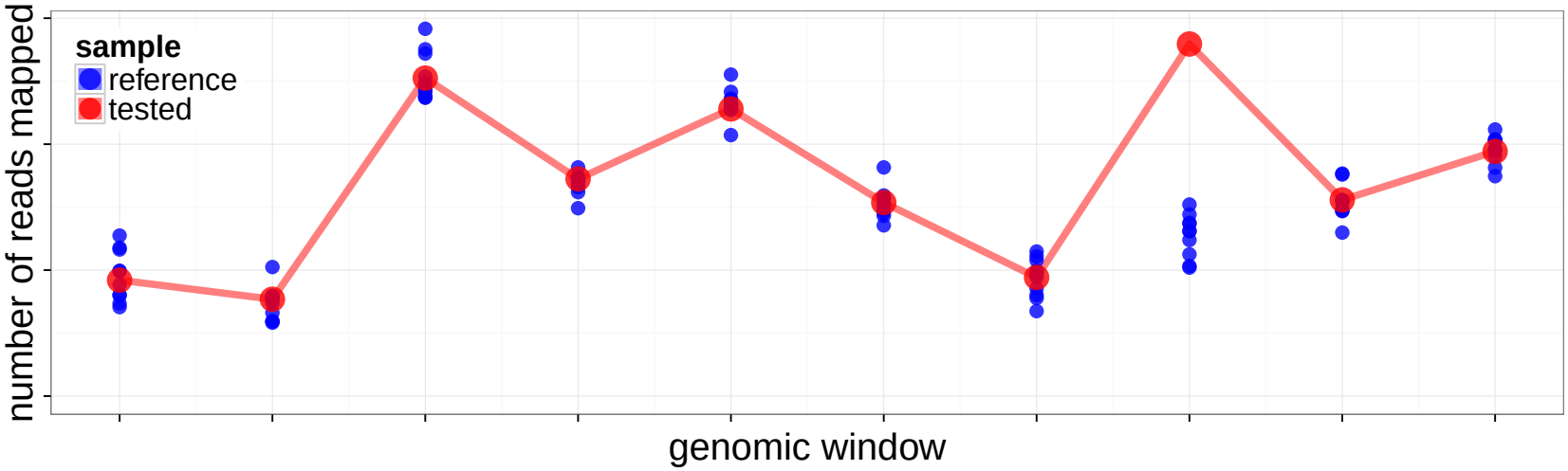
Test the entire genome, including low-mappability regions, and detect subtle abnormal coverage.

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

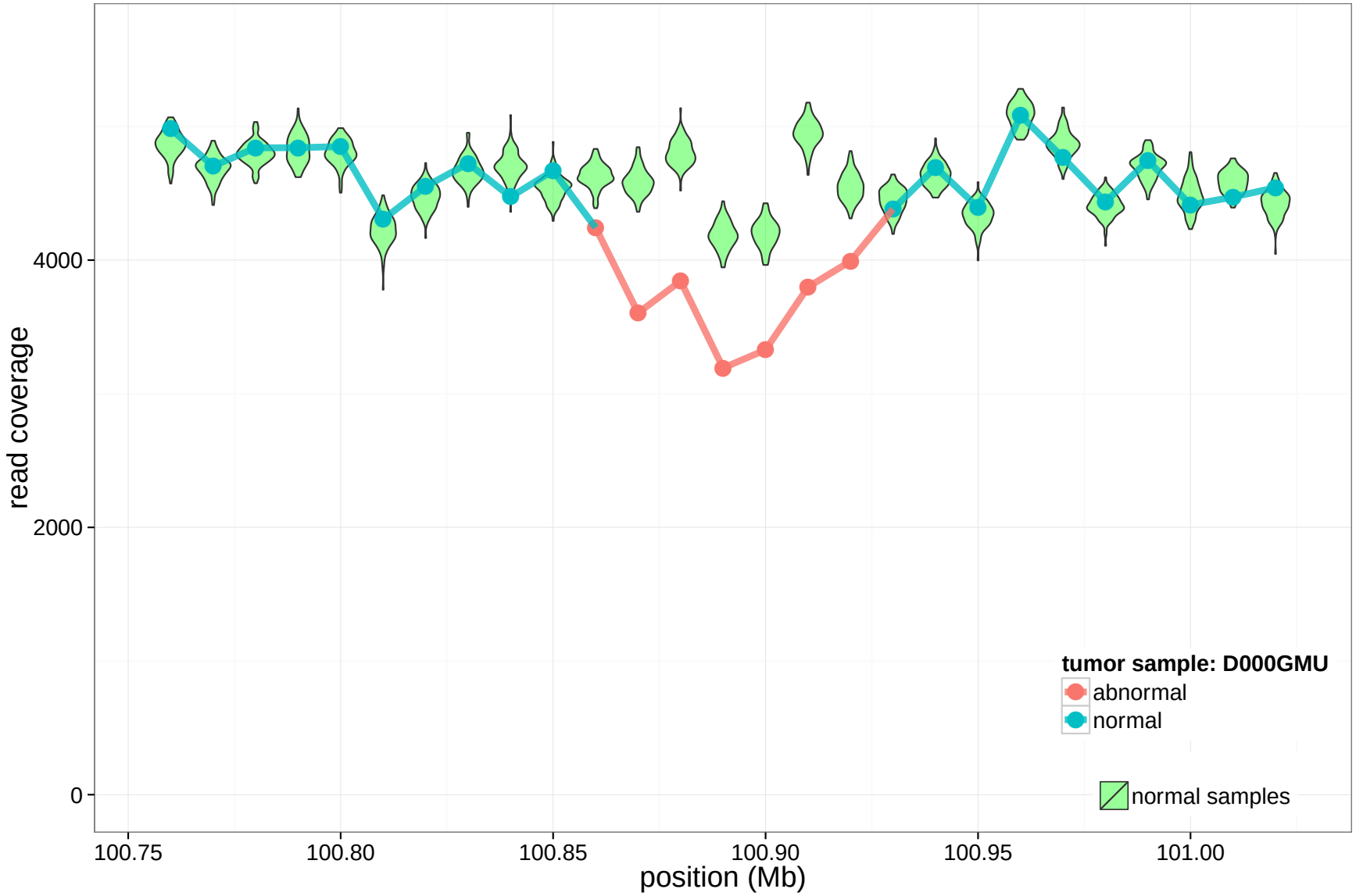
- Whole-Genome Sequencing of 100 individuals.
- Normal and tumor paired samples.
- Reference samples : normal samples.

Twin family dataset

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

$\sim$ *40X coverage, Illumina paired-end 100bp*

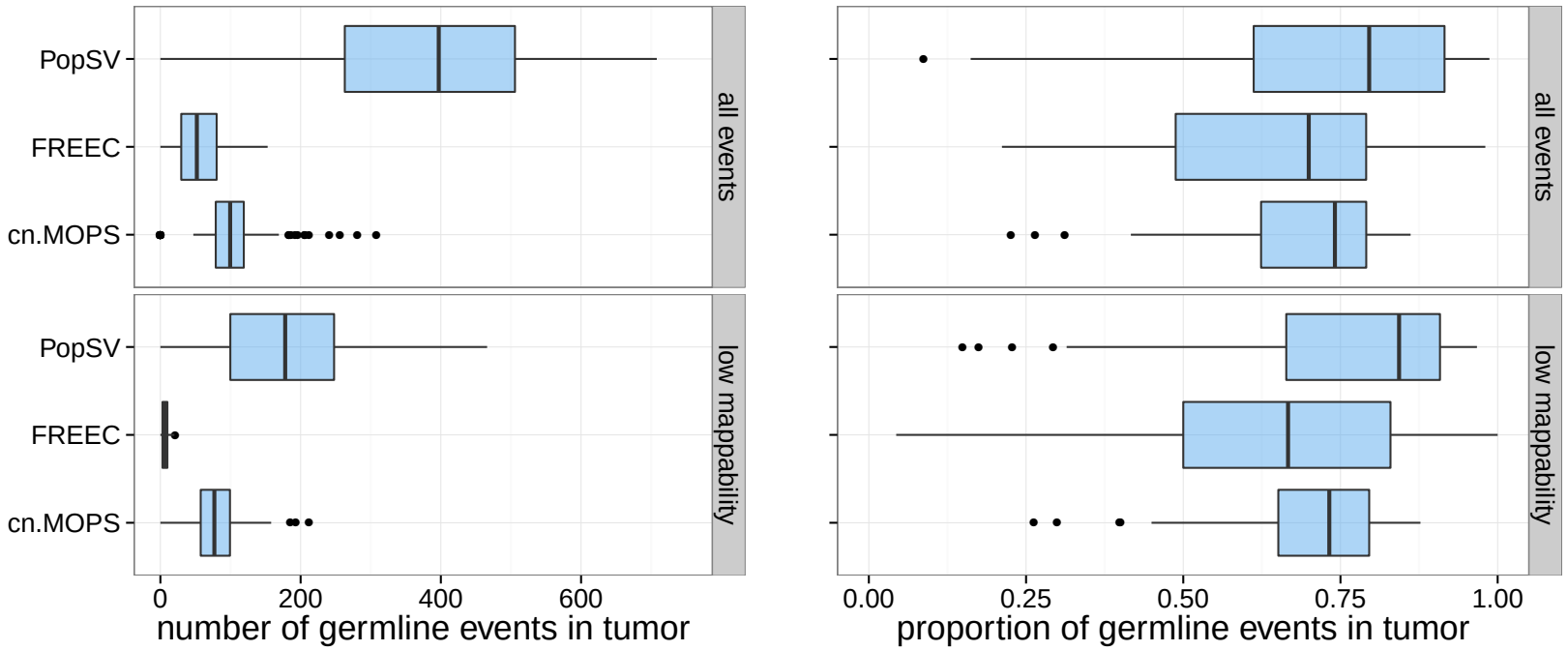
Example : Partial signal supporting tumoral deletion



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

Evaluating PopSV performance

Germline events detected in tumor samples



Results

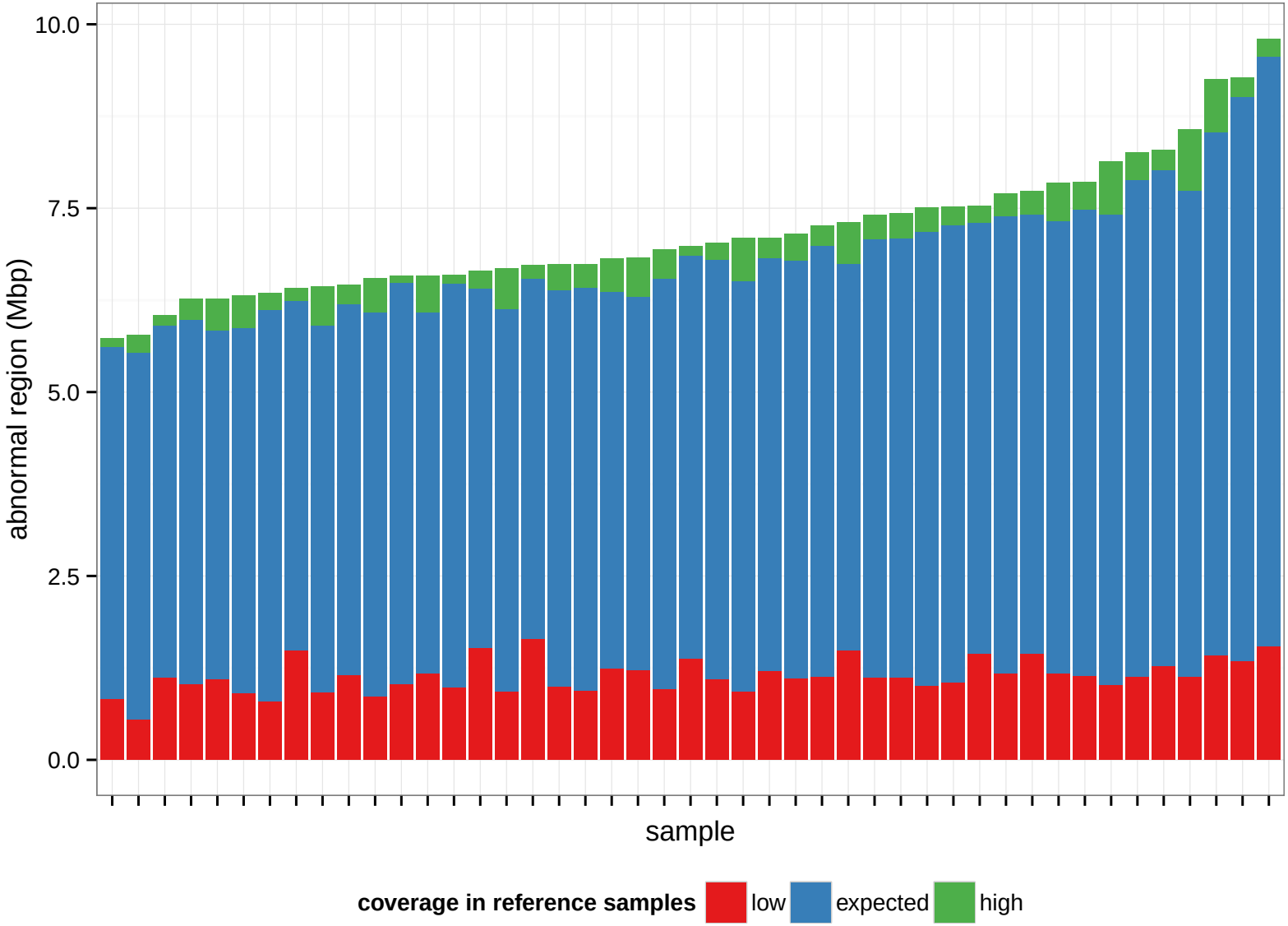
PopSV detected more consistent calls than other methods with similar specificity.

Other validation and benchmark

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

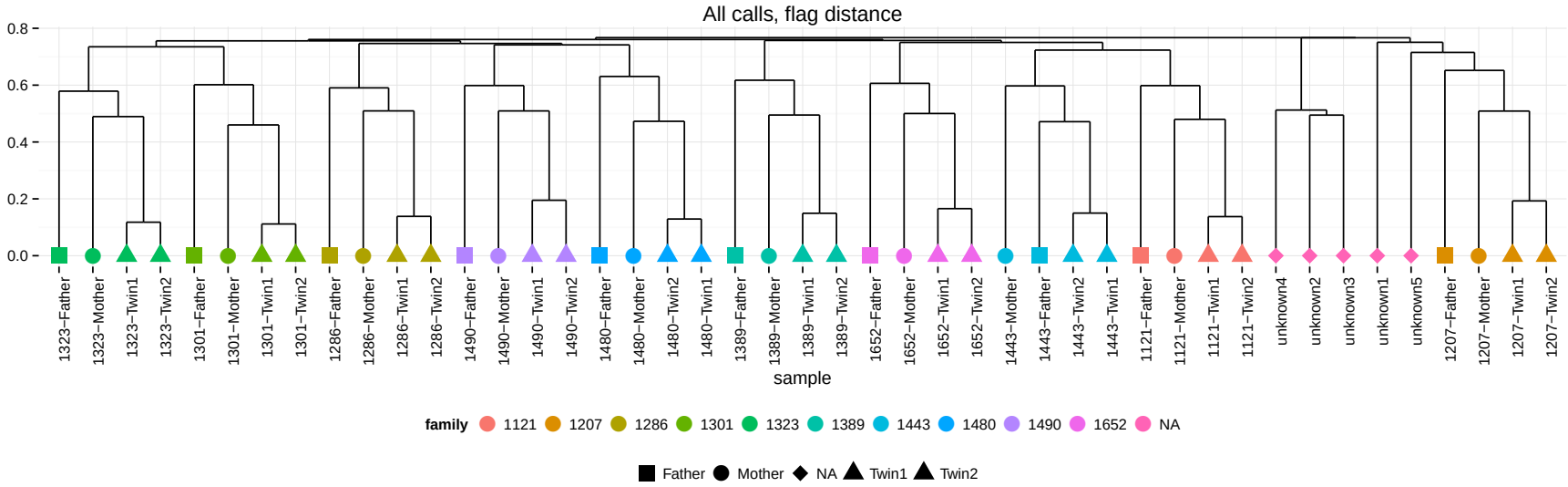
For more details/discussion come see **poster 30** tomorrow !

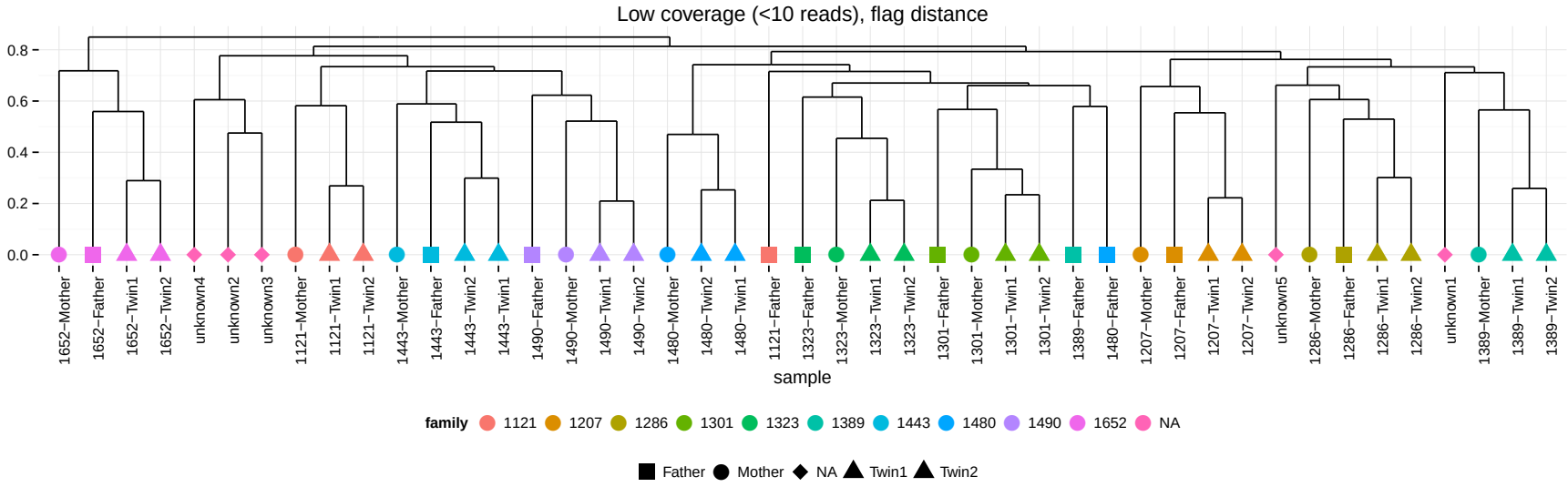
Twin dataset : PopSV on normal genomes



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

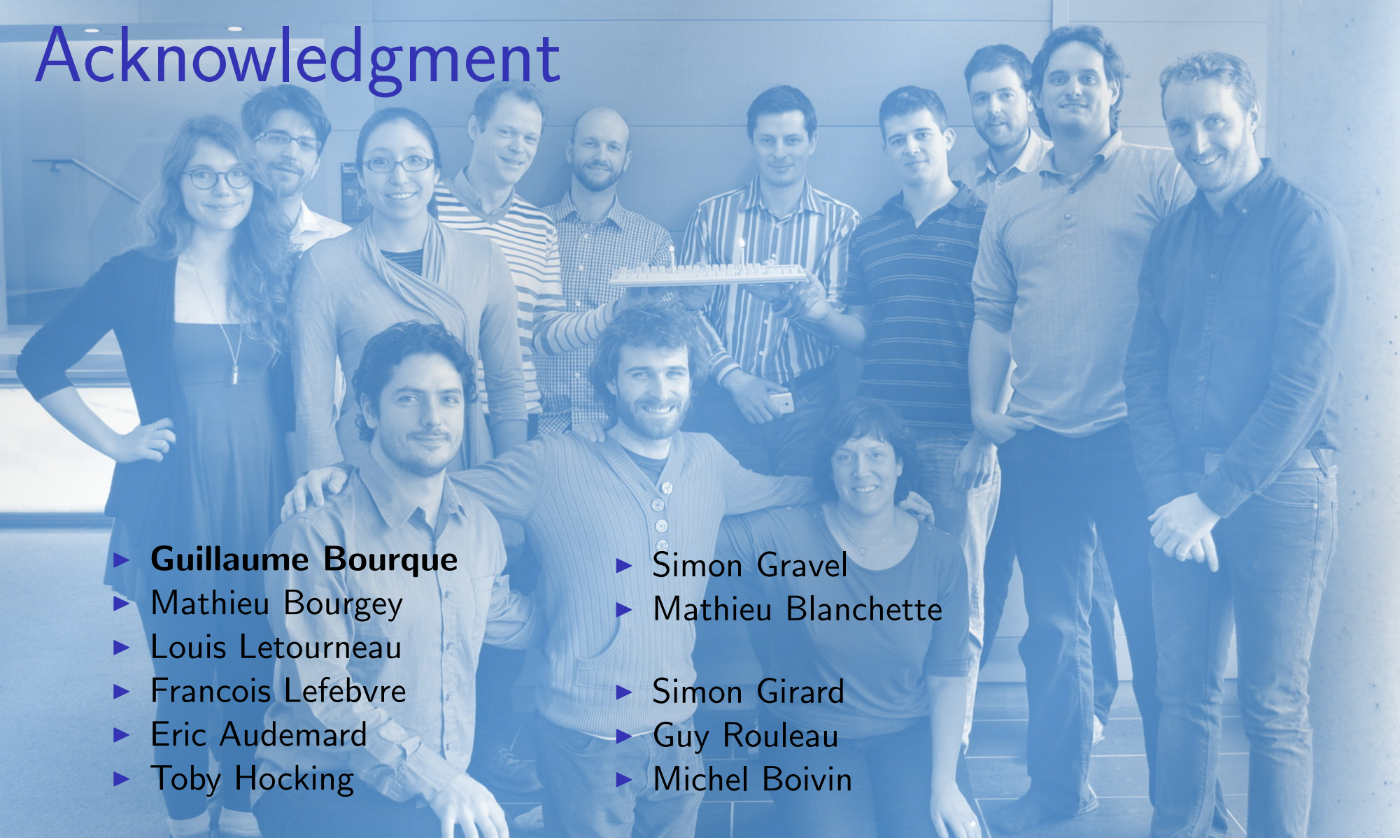
- variants in high and low mappability regions
- variants with partial signal (e.g. in tumors).

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

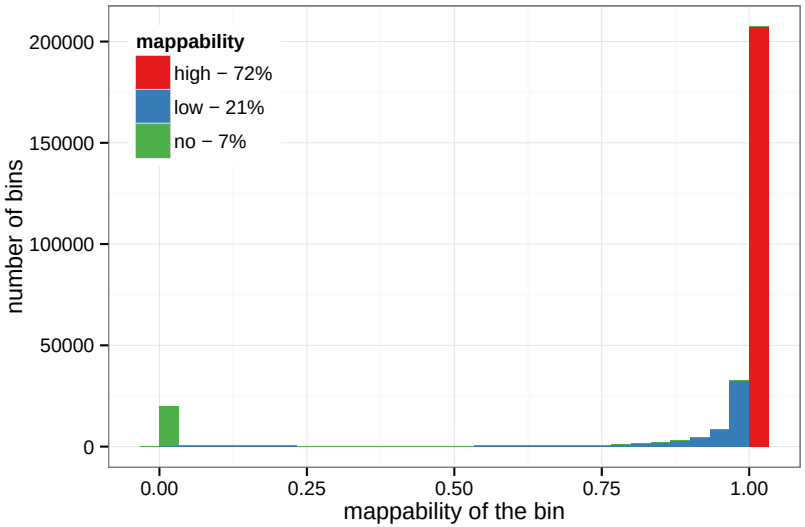
CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

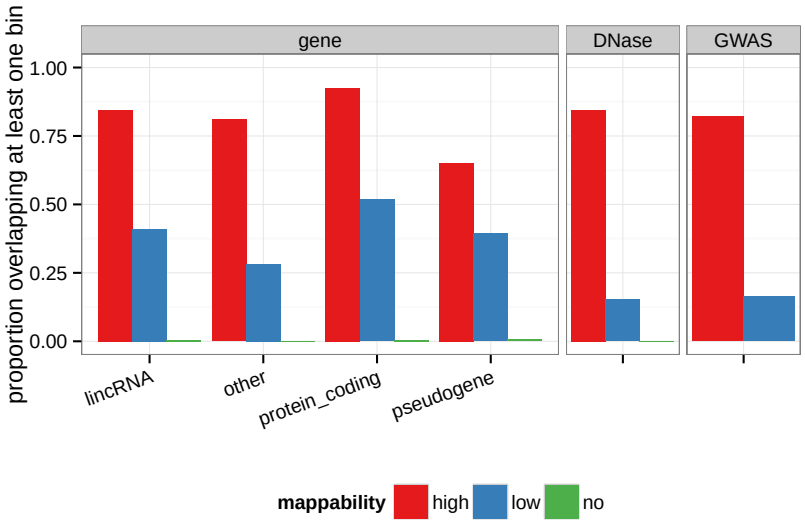
GenomeQuébec

NSERC CRSNG

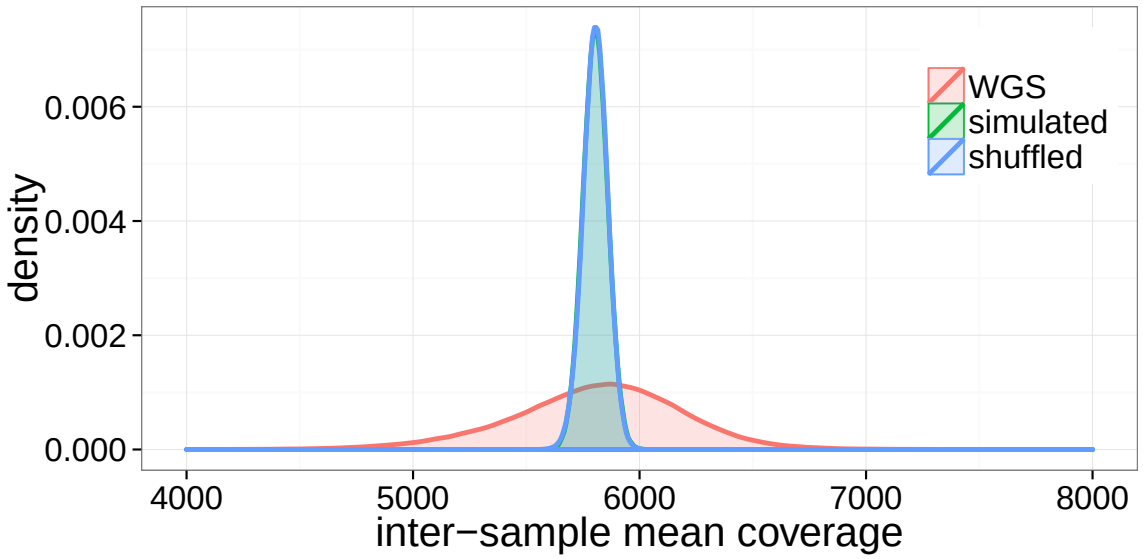
Thank You !

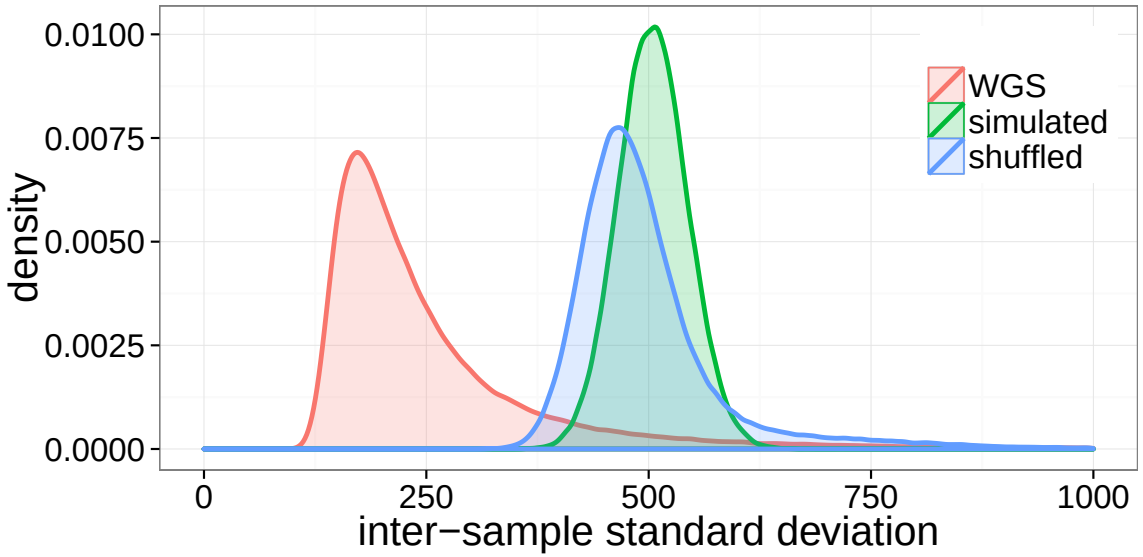
Low-mappability regions overlap functional elements





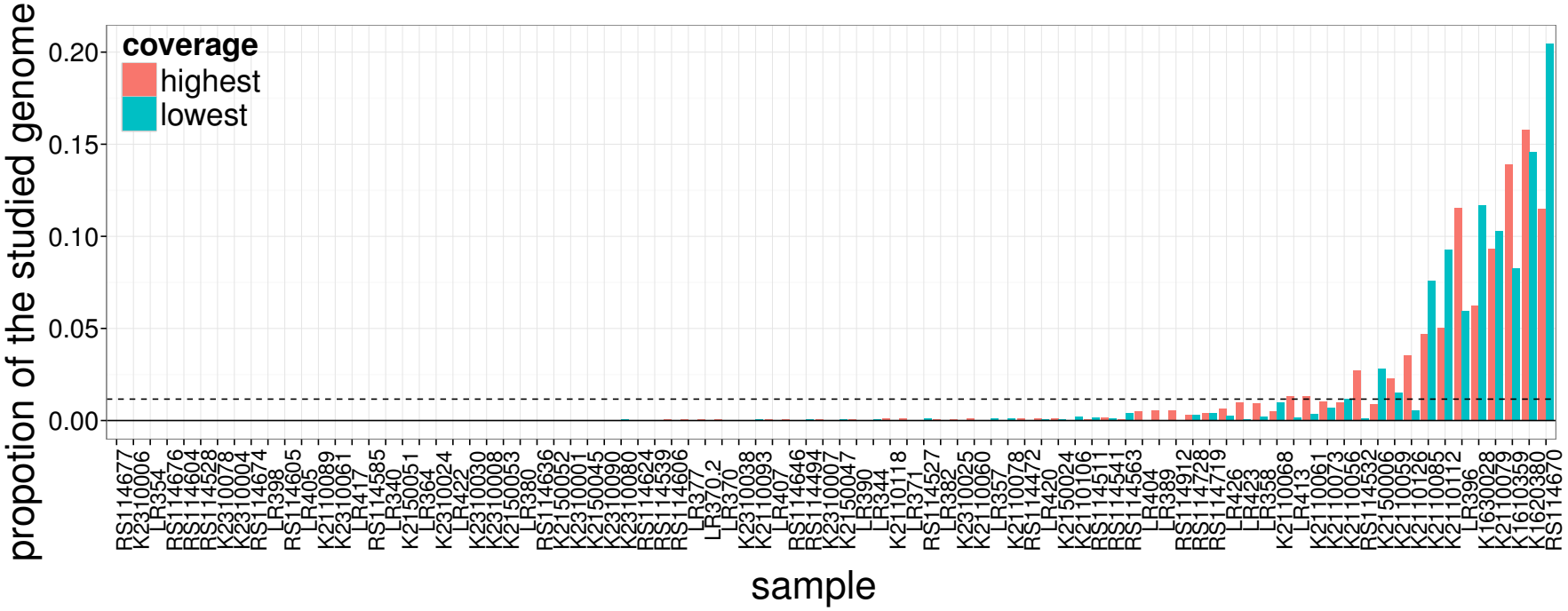
Unknown technical bias





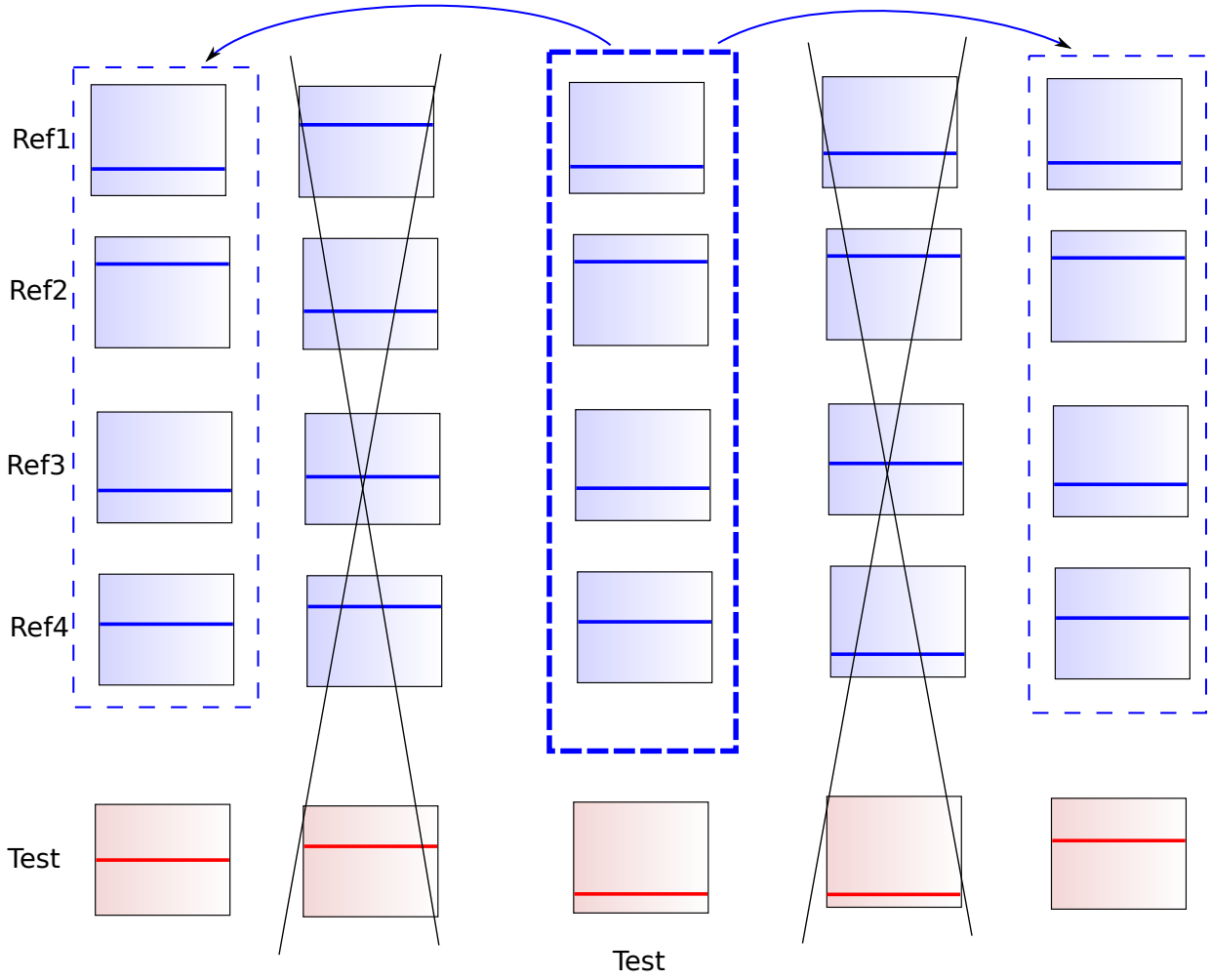
PopSV: importance of normalization

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



PopSV: importance of normalization

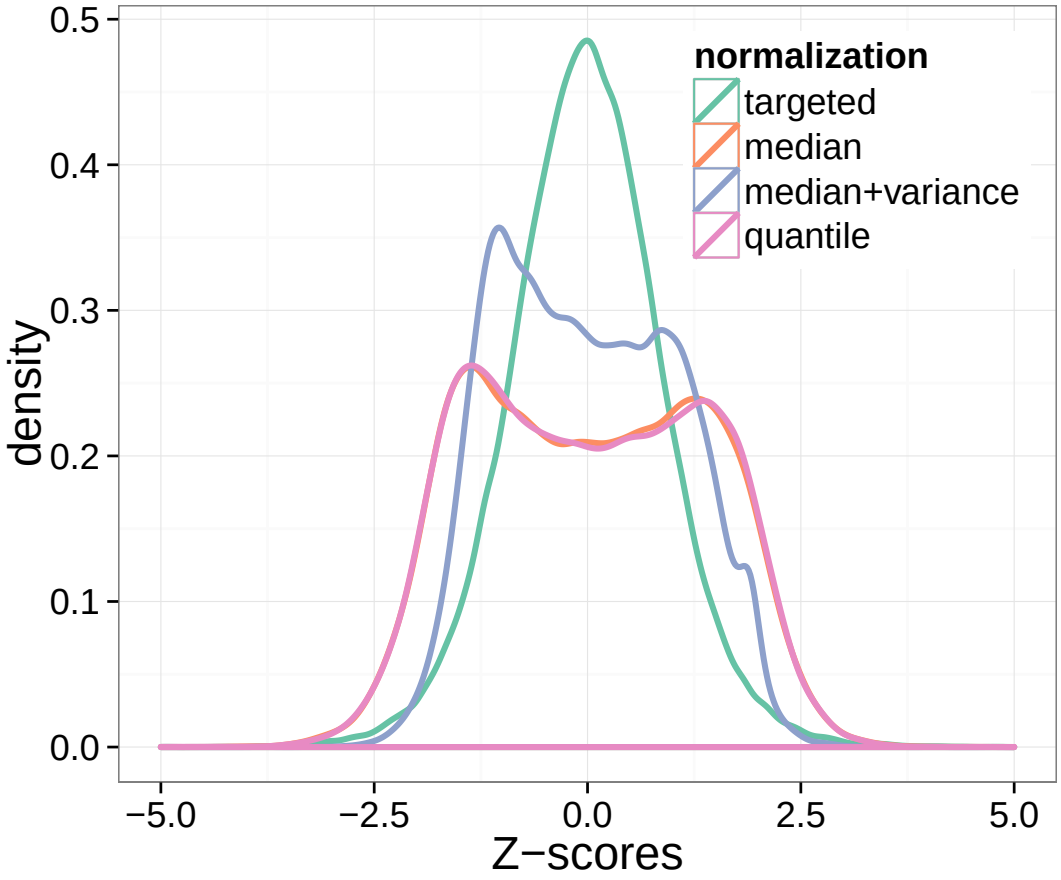
- PCA-based normalization (*Krumm*, 2012; *Boeva*, 2014).
- Targeted normalization: linear using a subset of the genome.



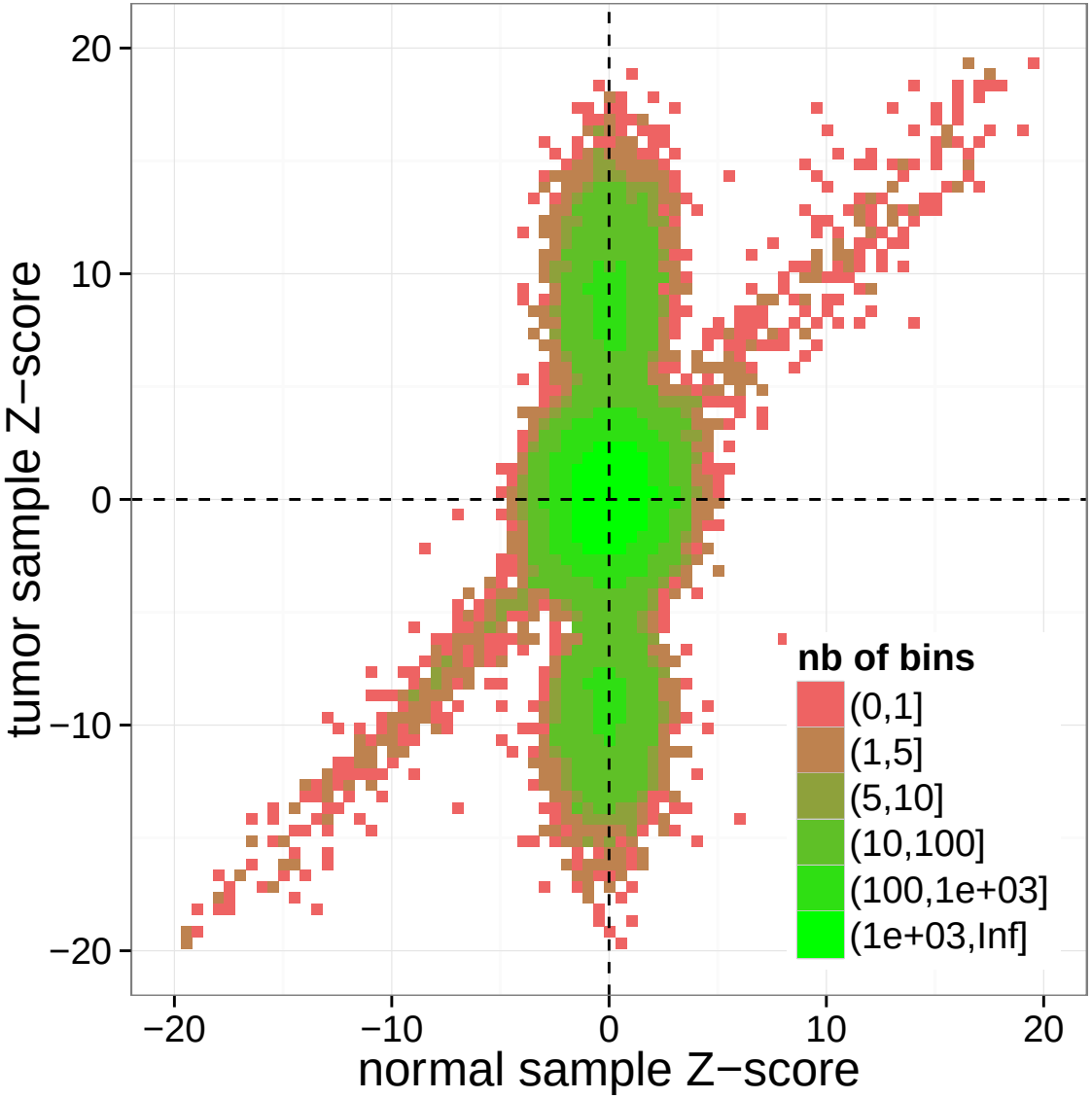
PopSV: Z-score and test

For a sample s:

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

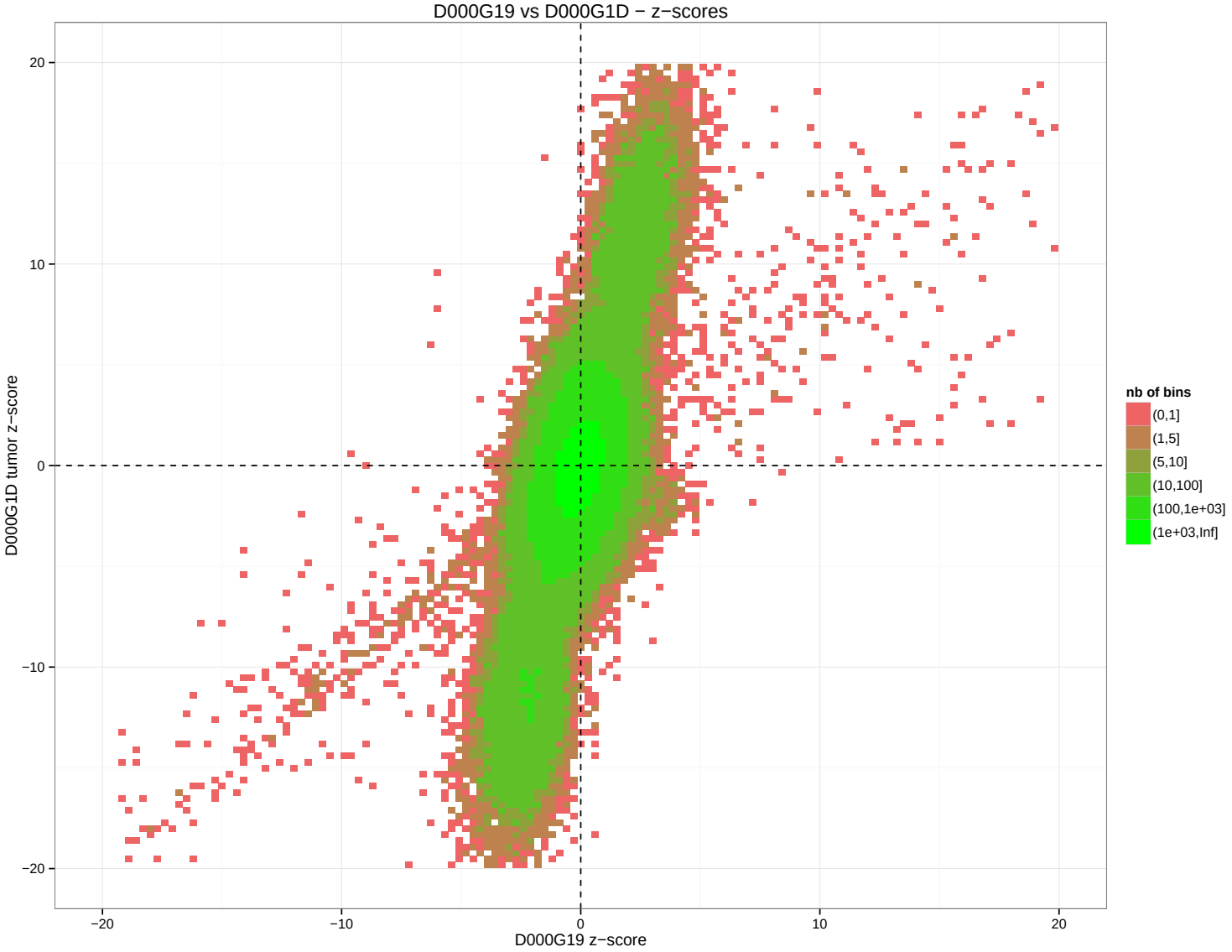


Z-scores : Normal versus Tumor

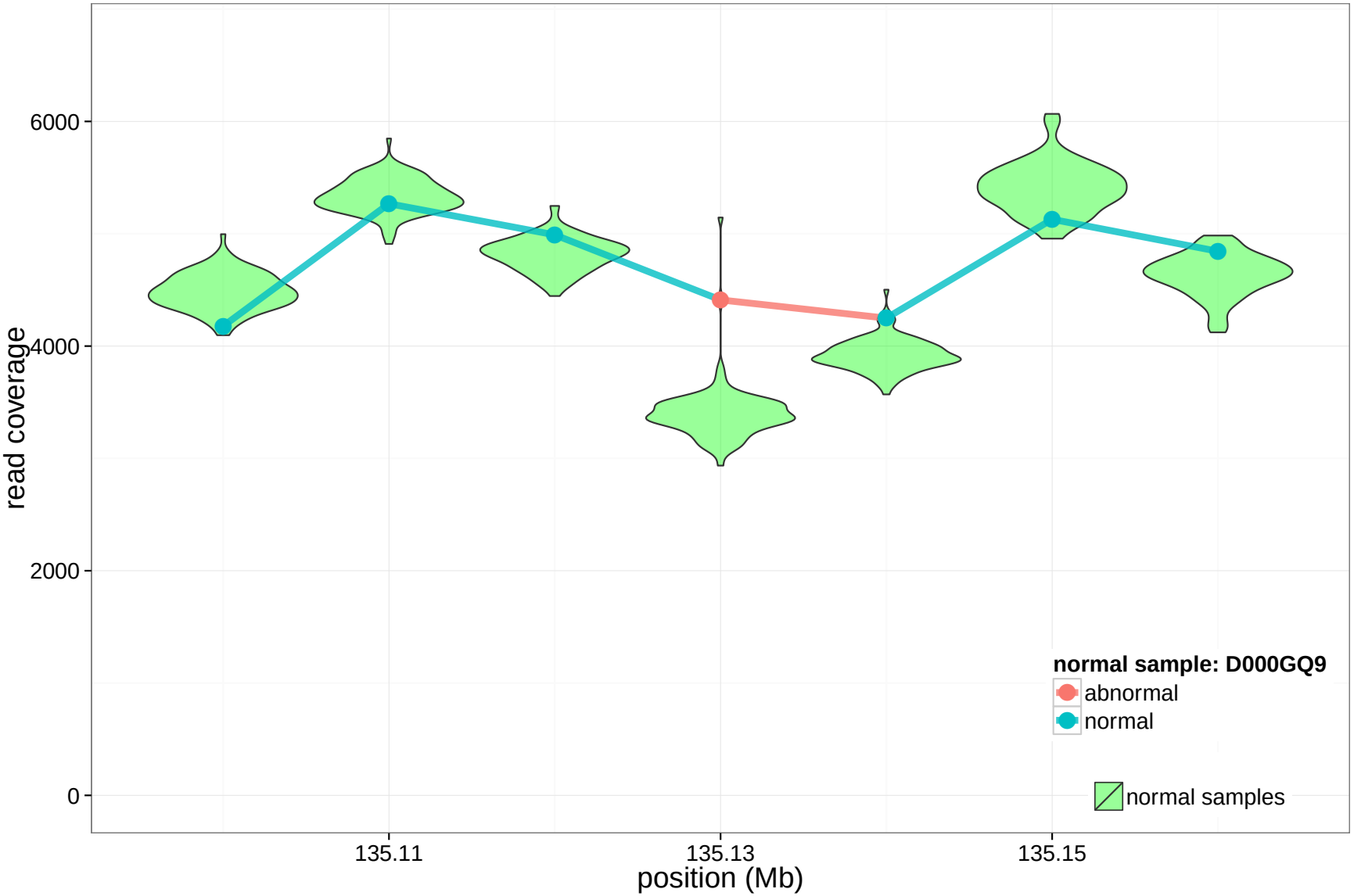


"funky snowman" plot

Z-scores : contamination detection

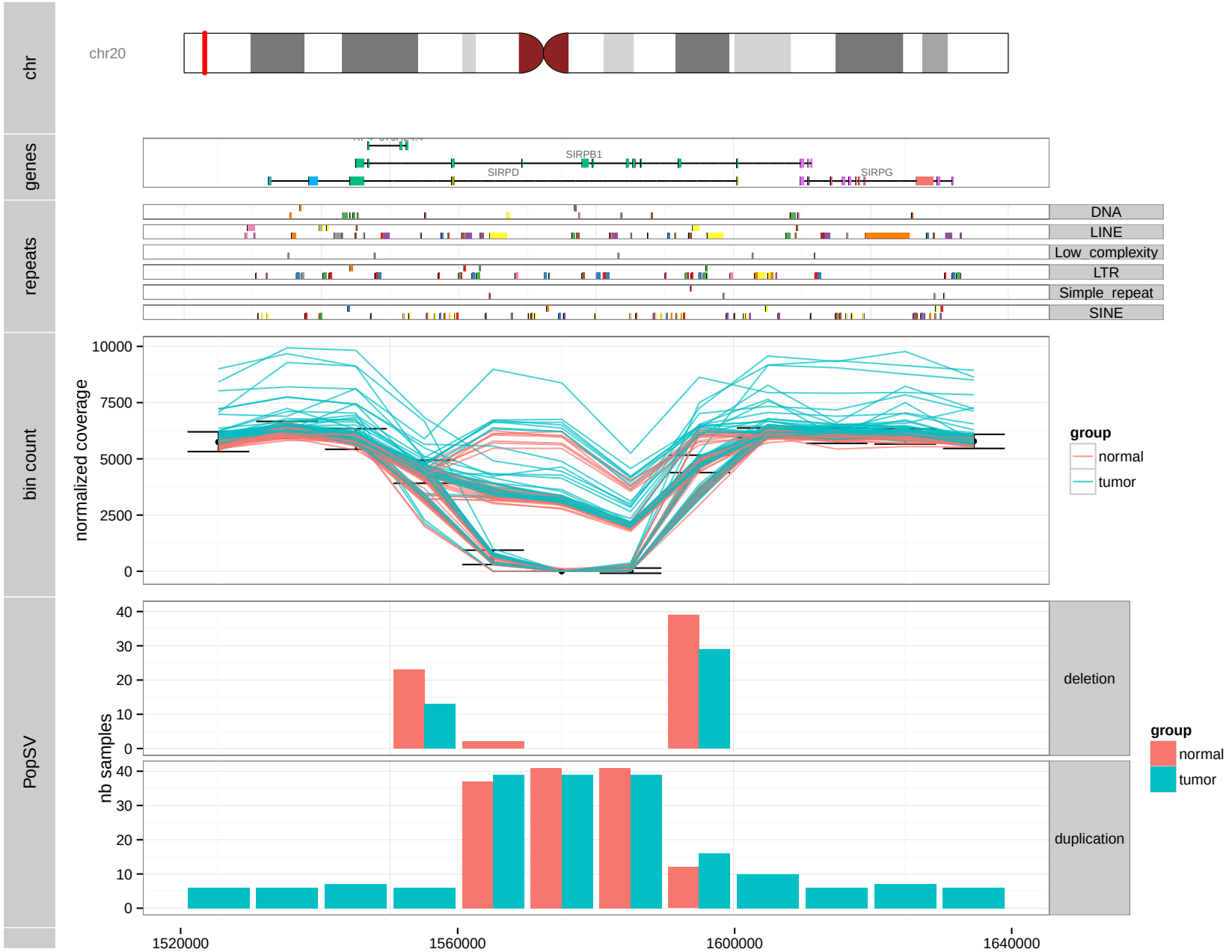


Example: Telomeric region

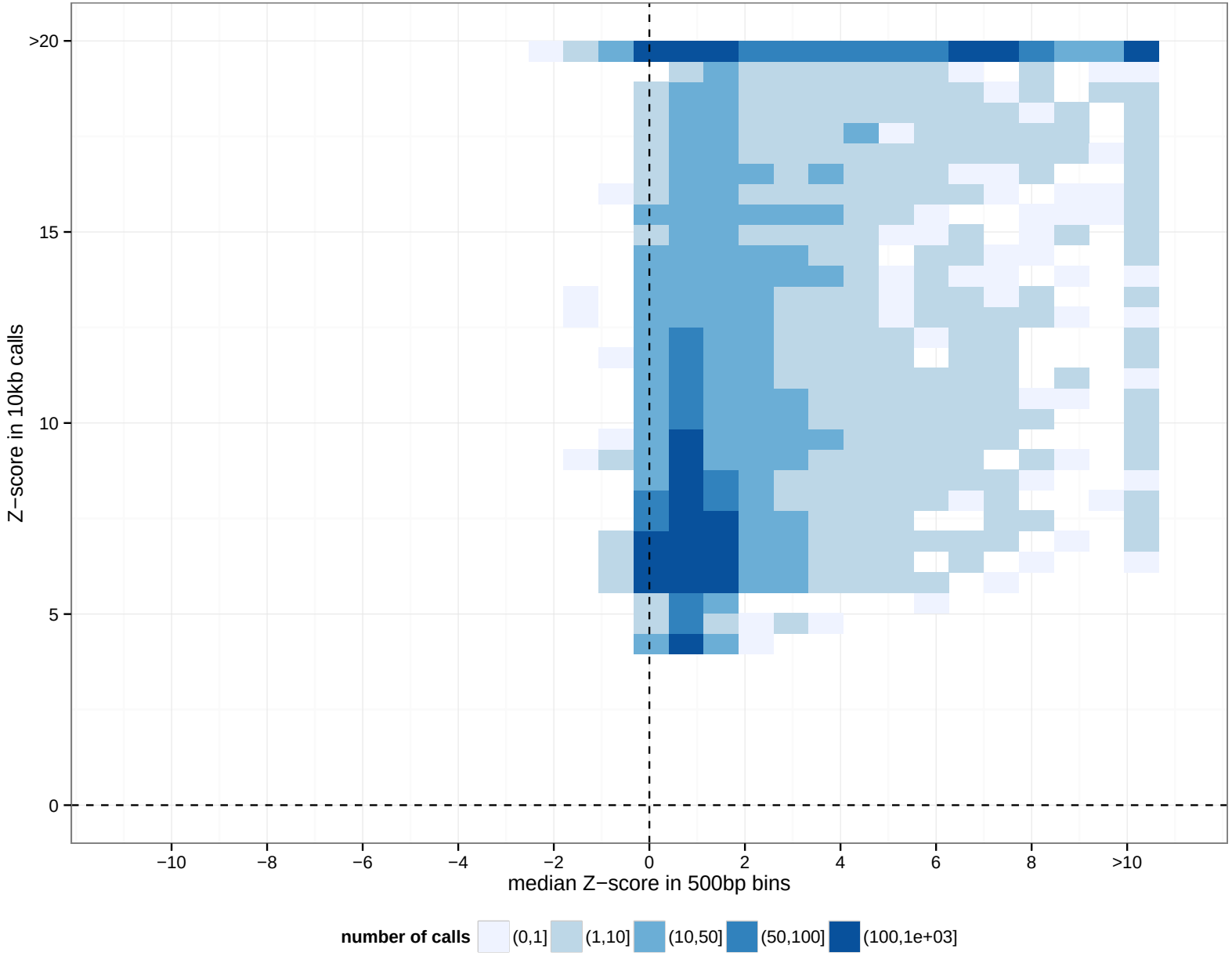


Chr.10, overlapping genes (PRAP1, CALY), not detected by other approaches.

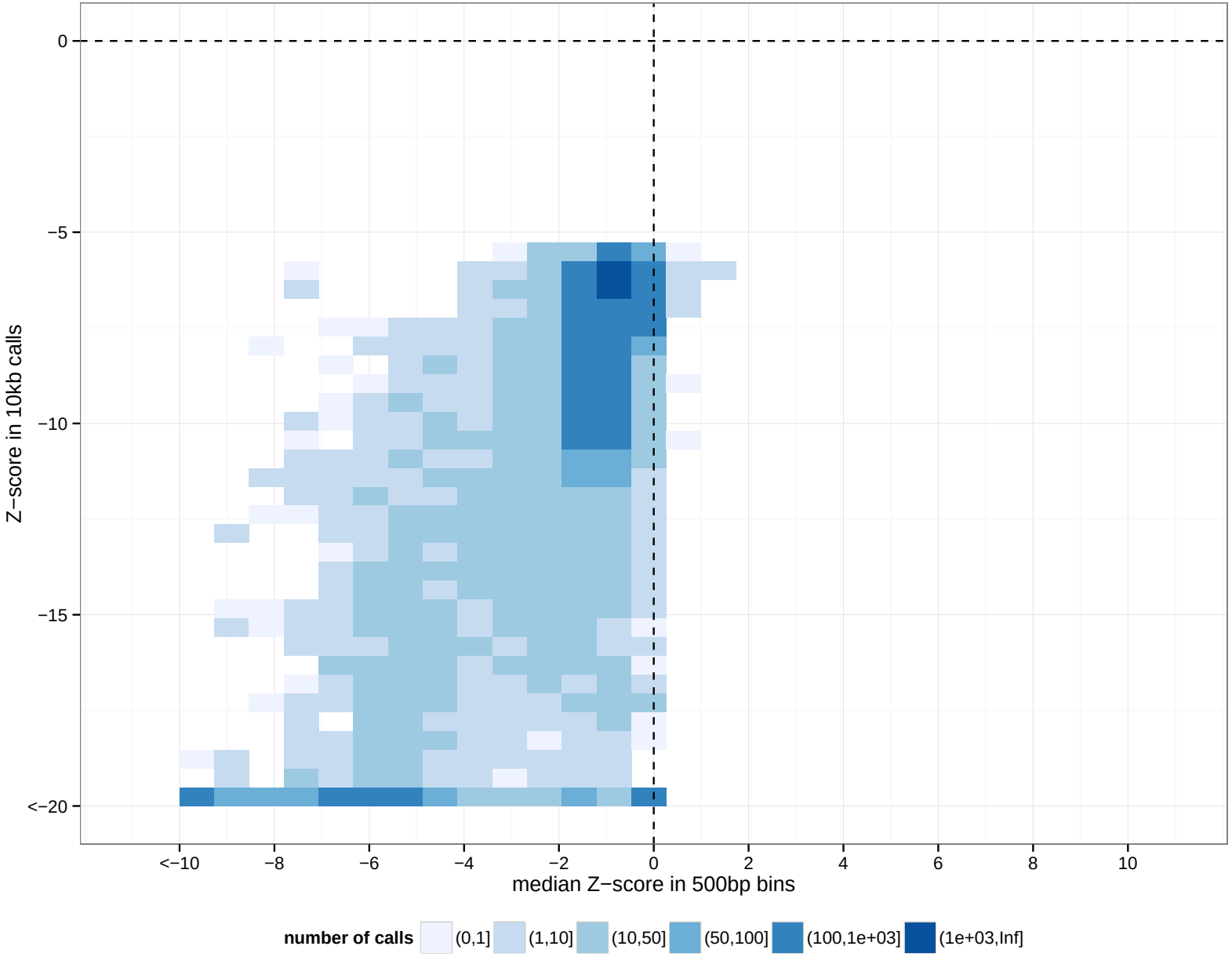
Example: NAHR candidate



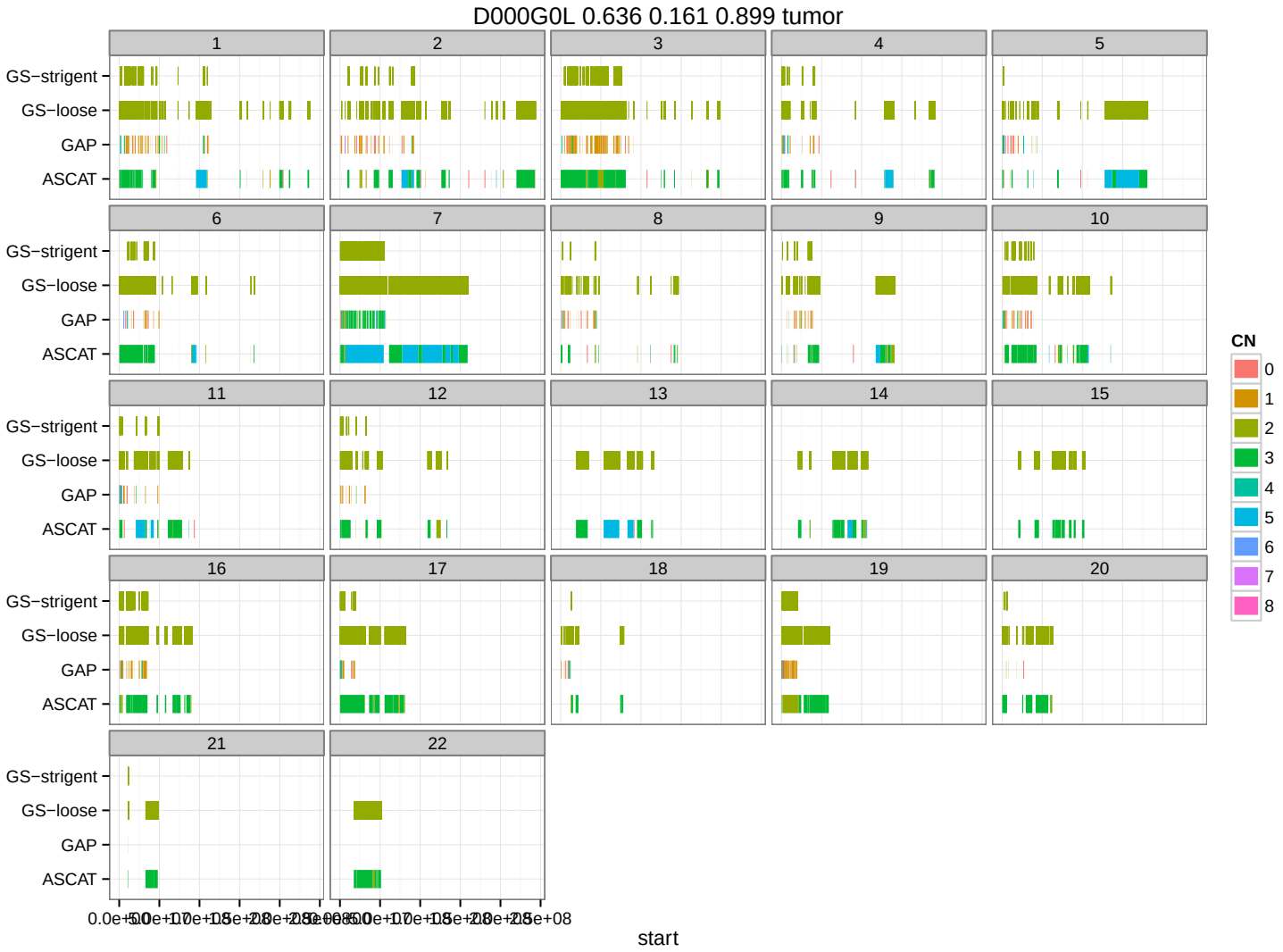
500bp Z-scores within 10kb calls



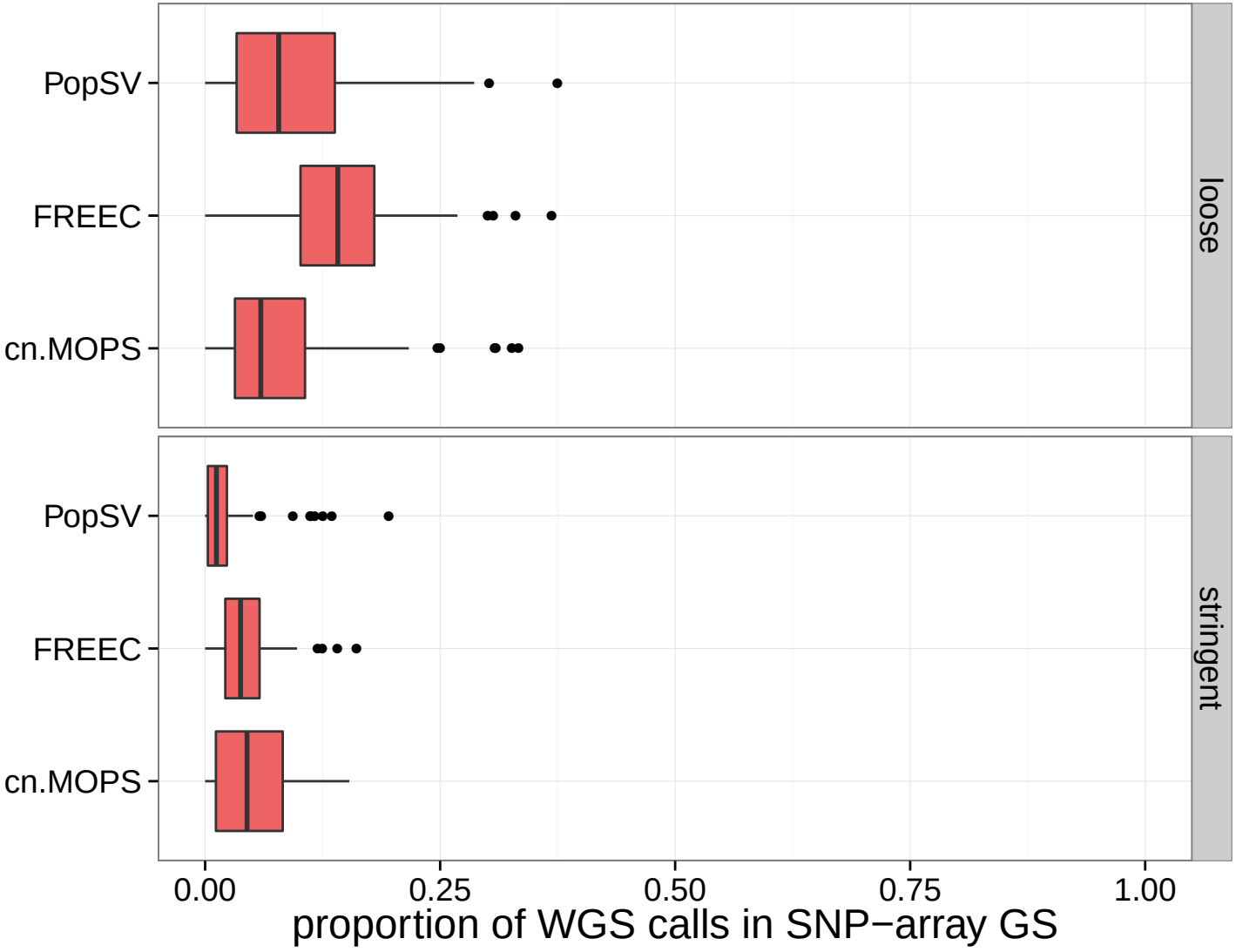
500bp Z-scores within 10kb calls



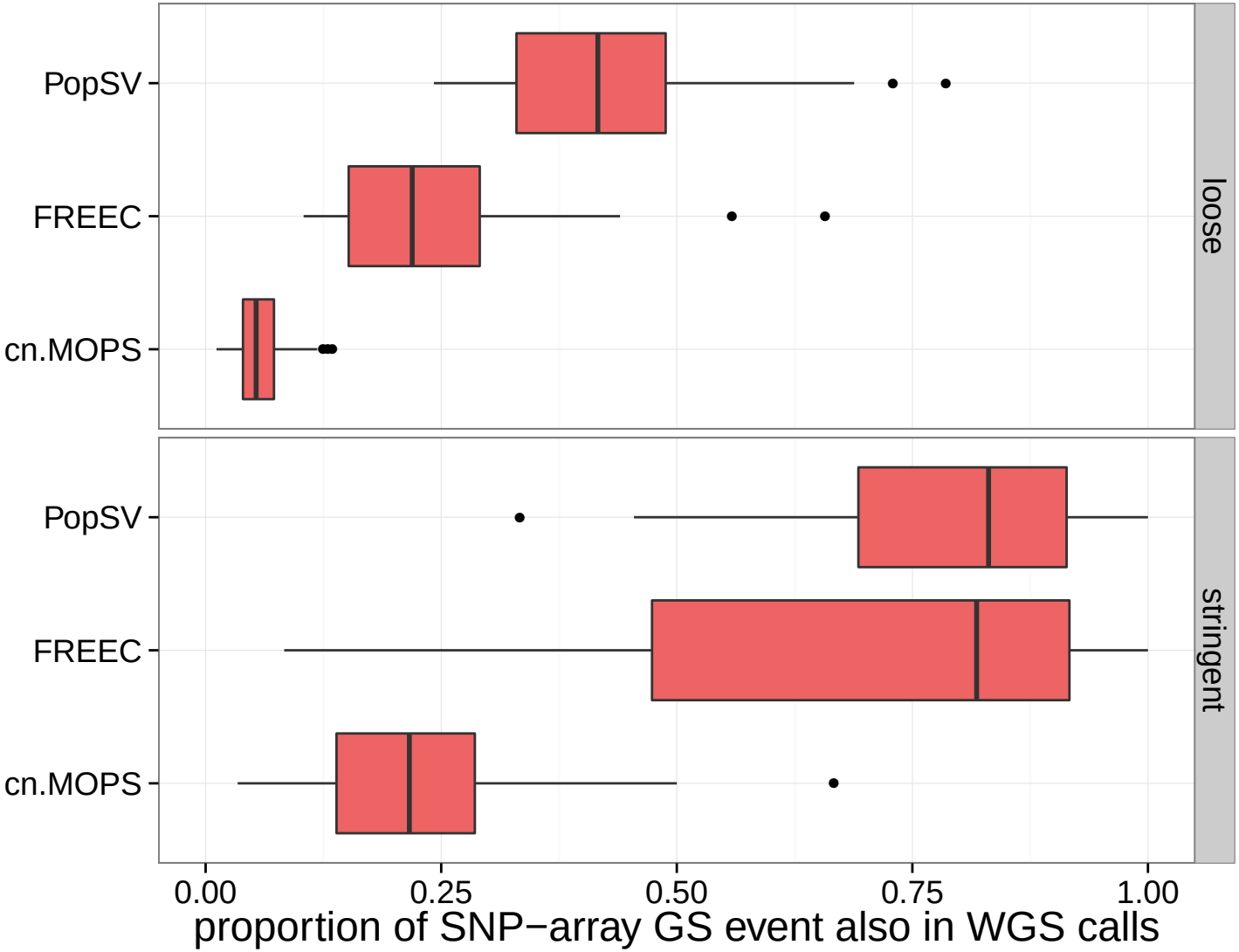
SNP array methods concordance



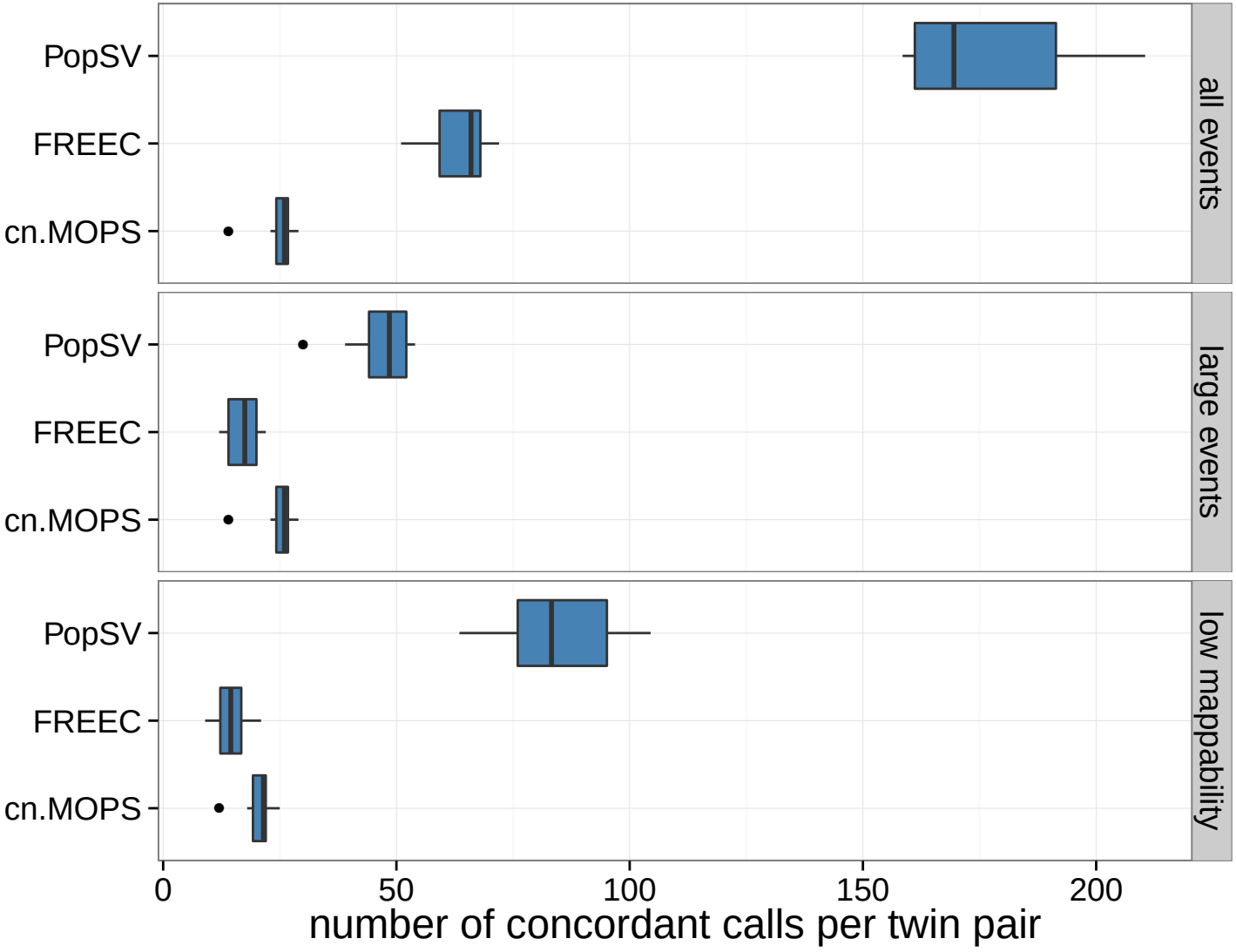
SNP array concordance



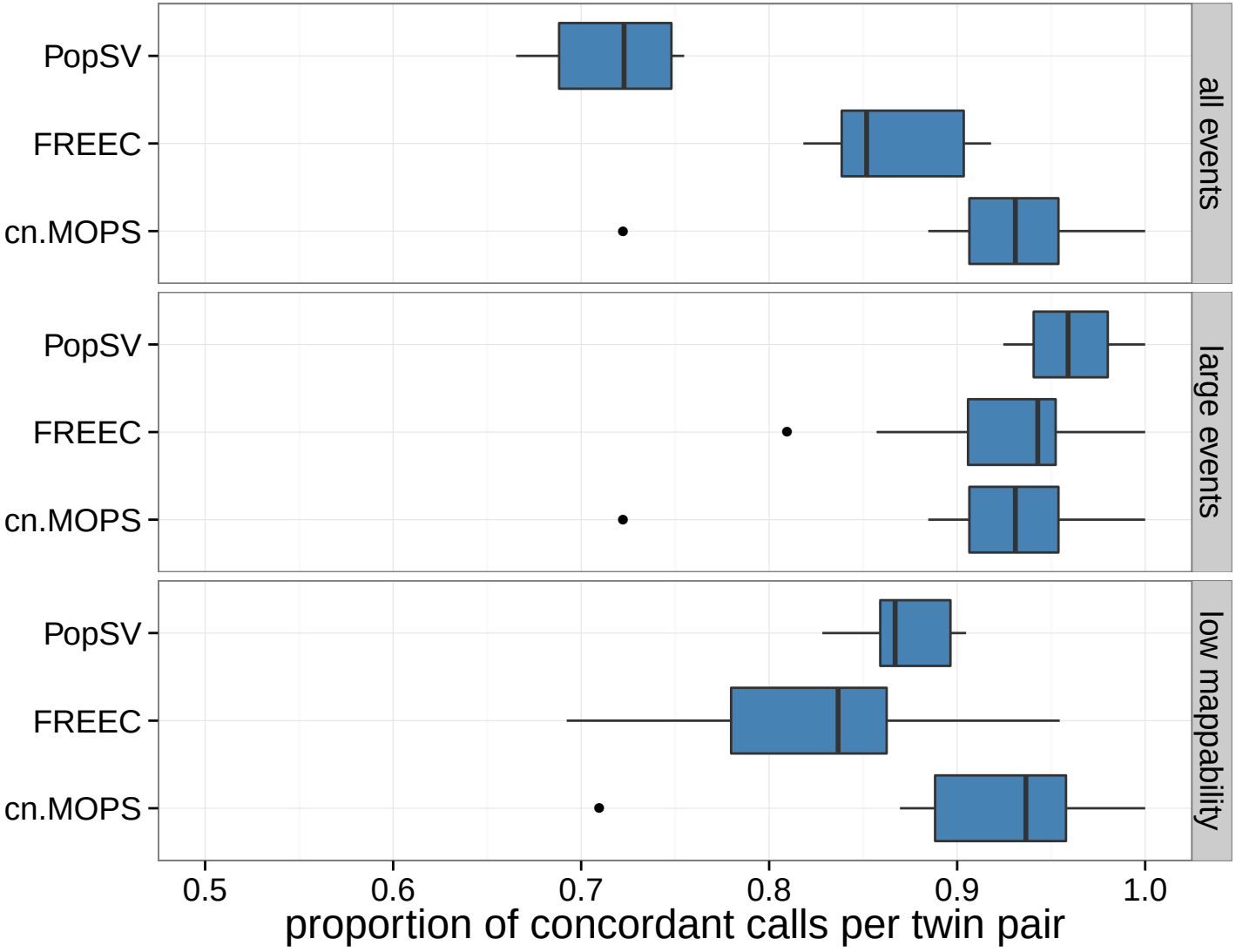
SNP array concordance



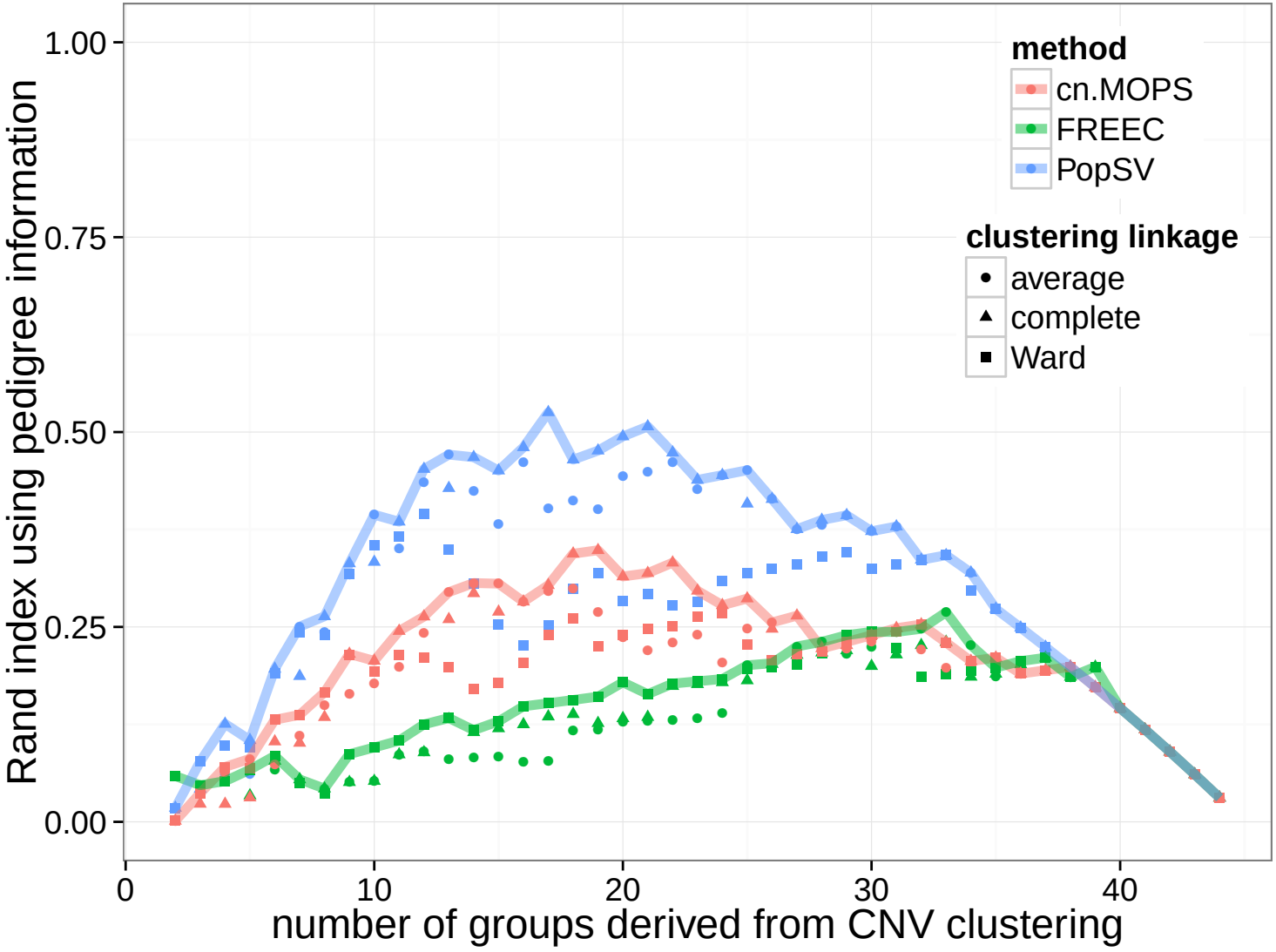
Twin concordance



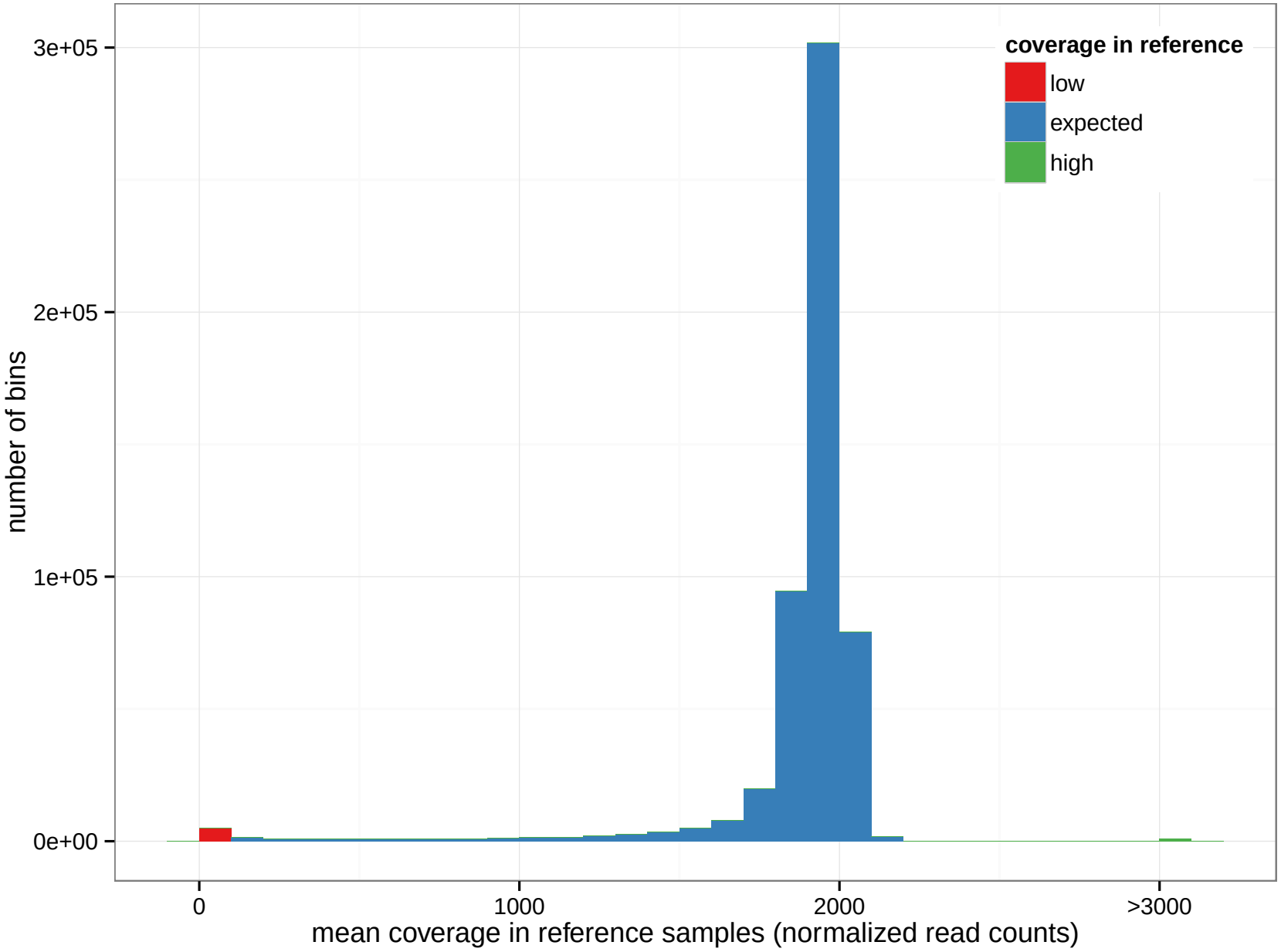
Twin concordance



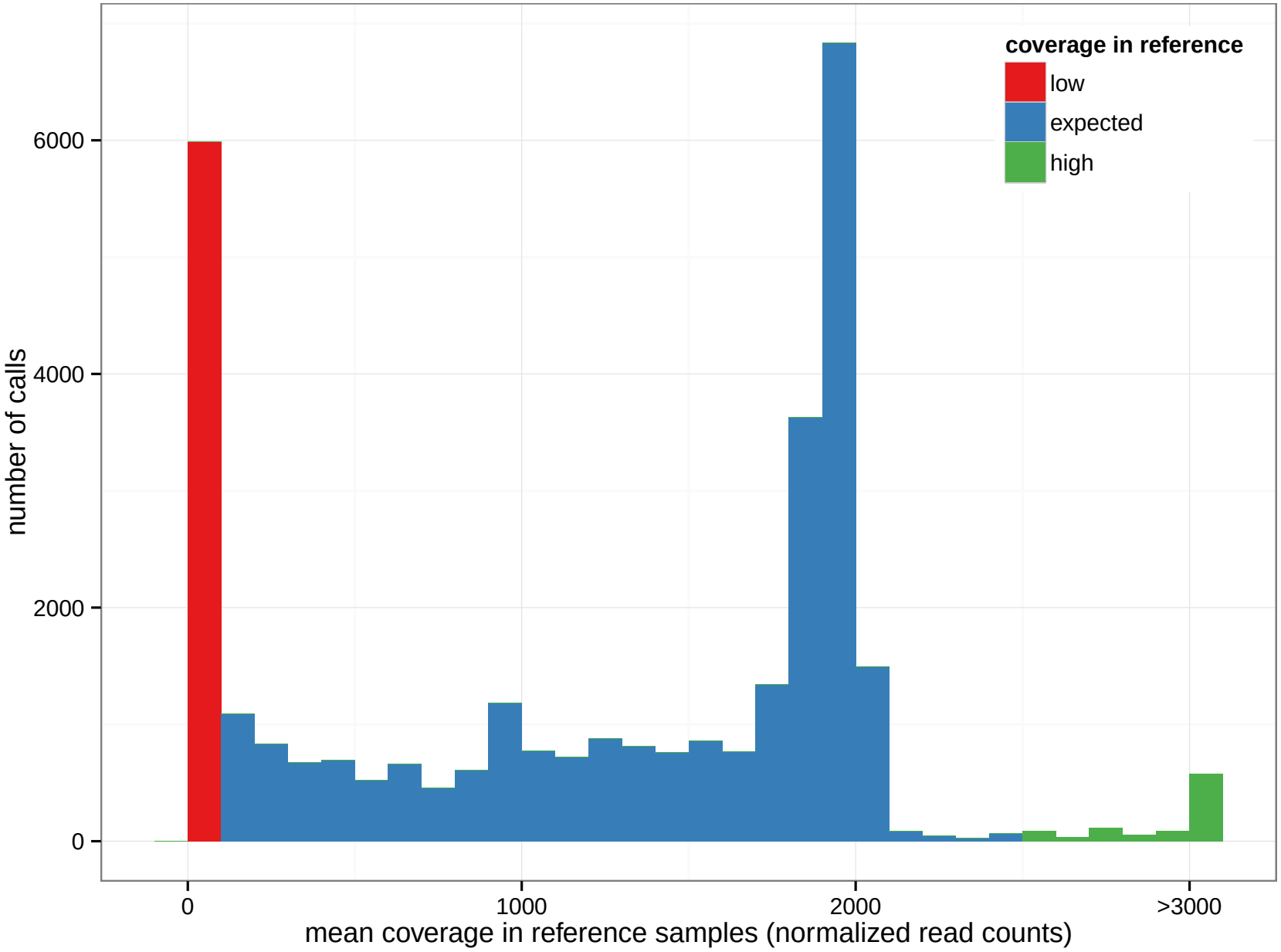
Twins and clustering quality



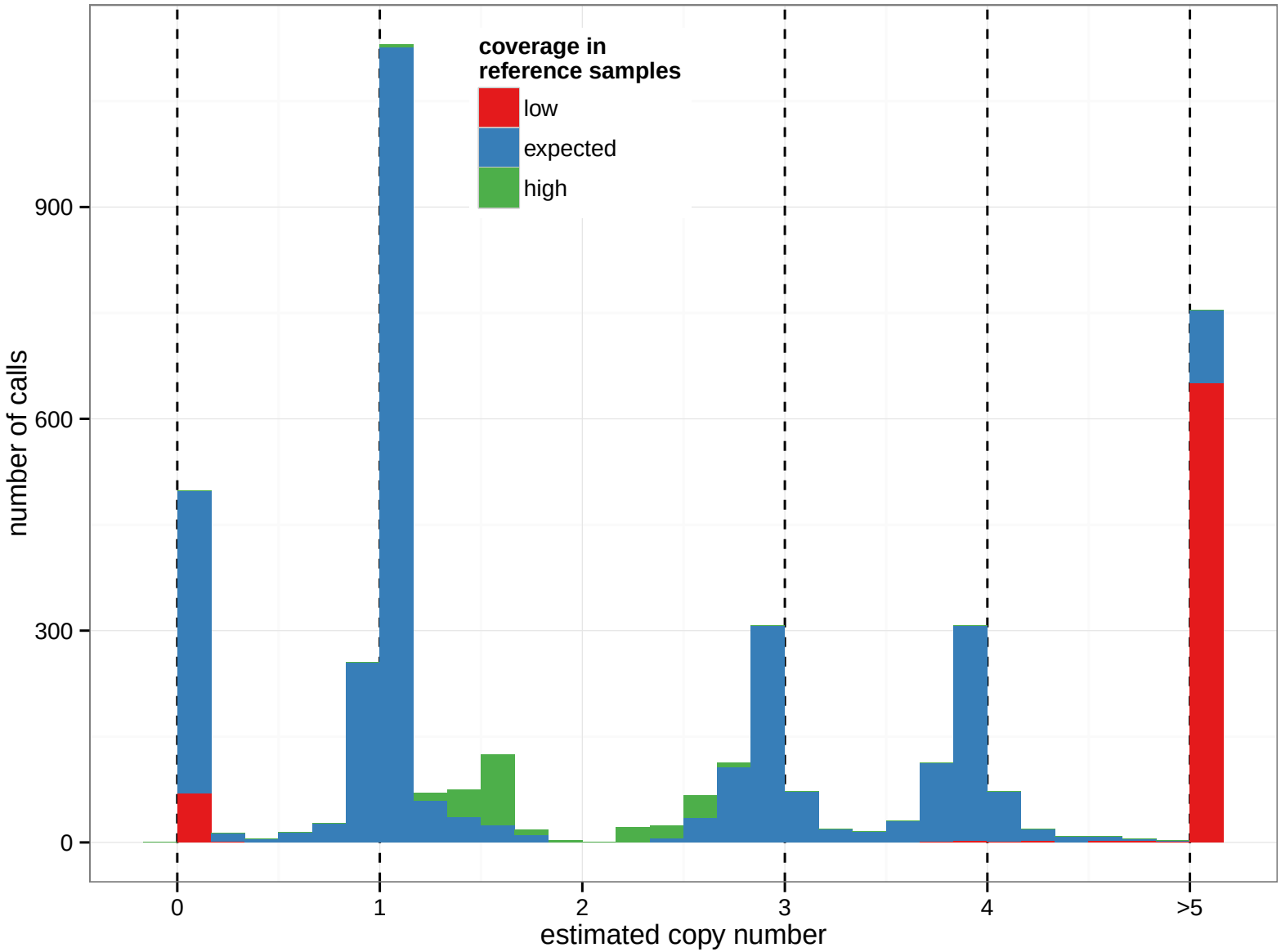
Many variants in low coverage regions



Many variants in low coverage regions

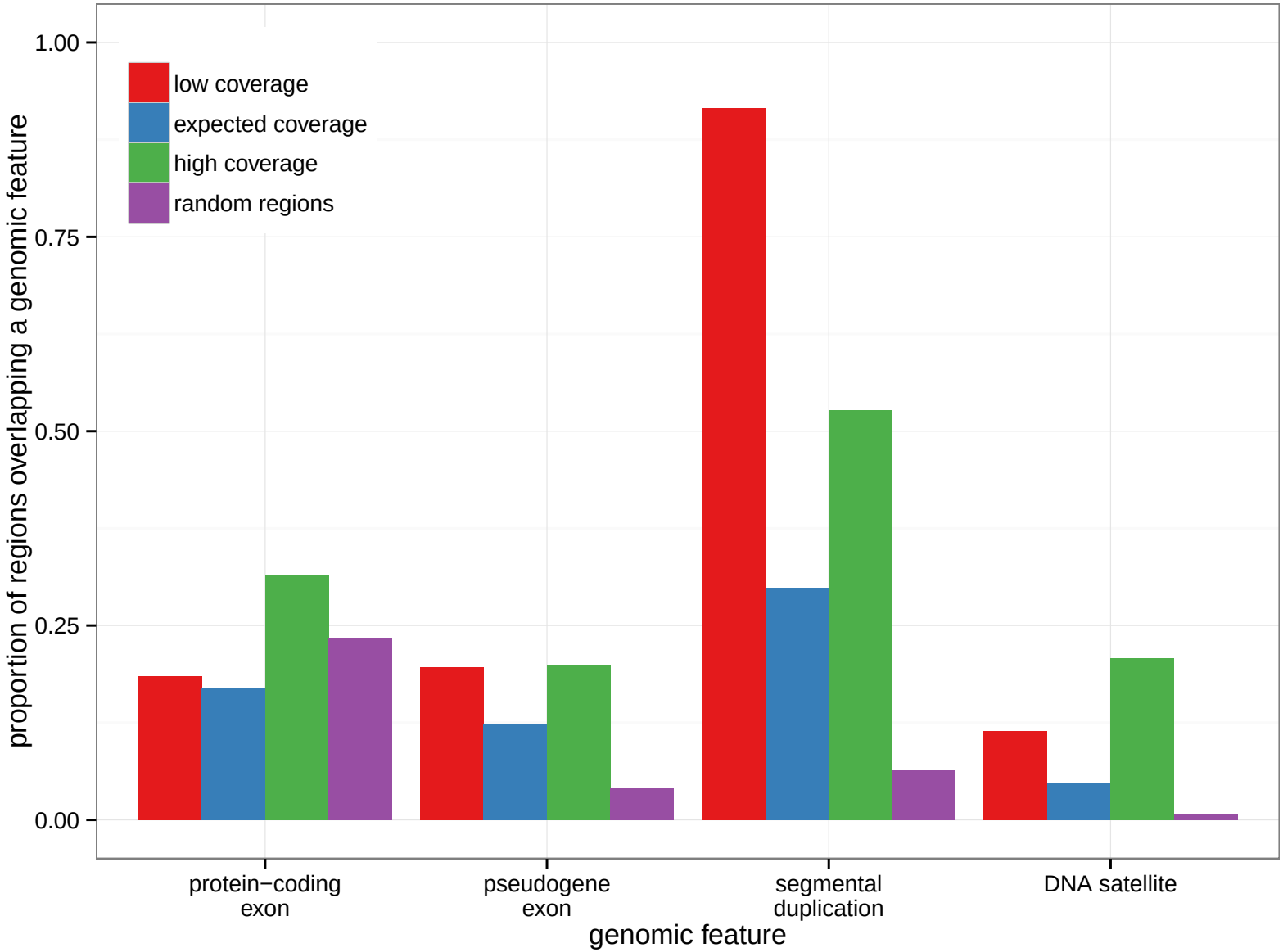


Twins dataset : copy number estimation

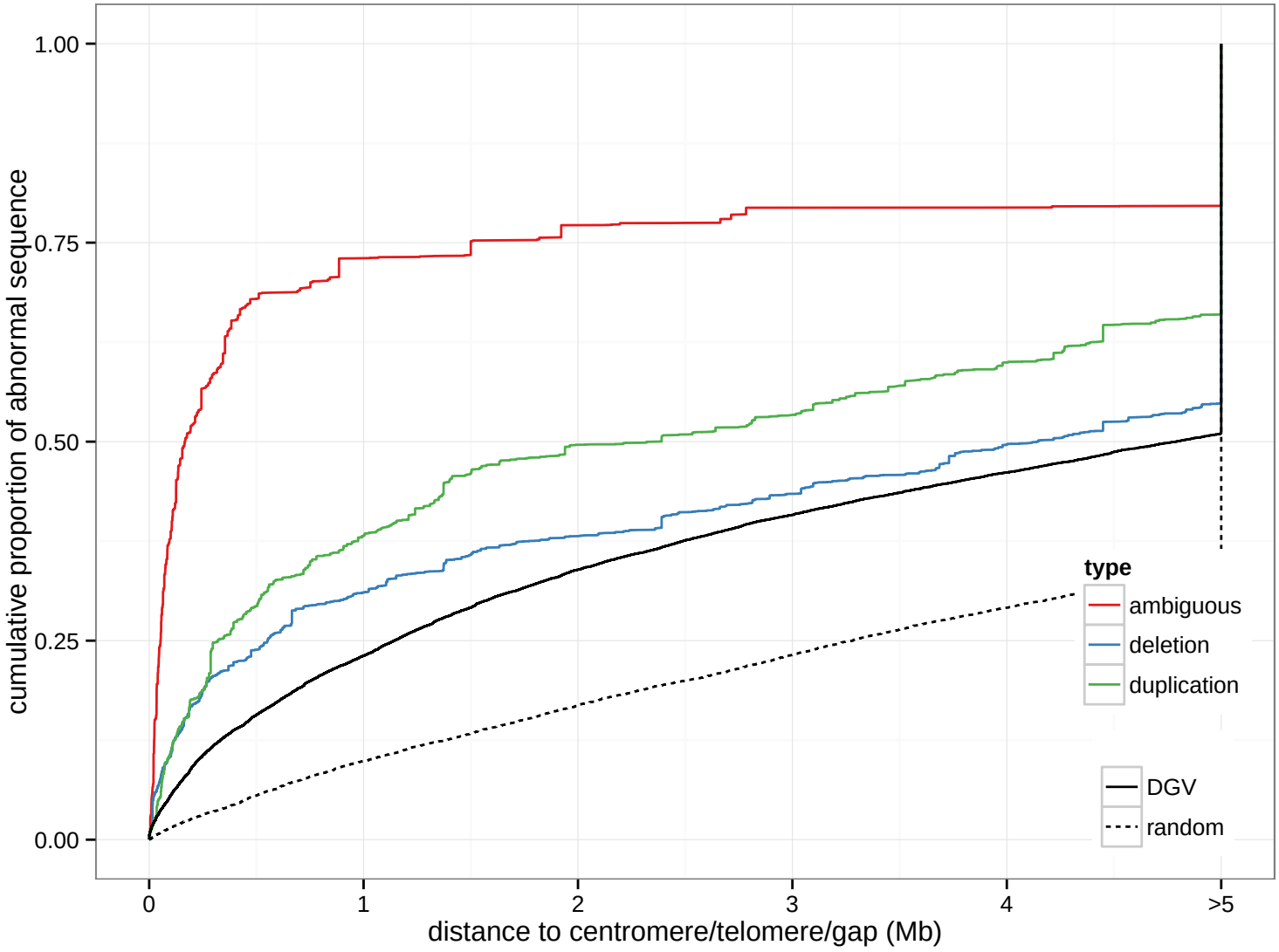


estimated copy number $= 2 \times$ *coverage in sample / coverage in reference samples*

Many calls in segmental duplications but also in genes



Distance to centromere/telomere/gaps



More SV detected near centromere/telomere/gaps.

Mappability