Global patterns of copy number variation in humans from a population-based analysis.

ICHG Kyoto

Jean Monlong
April 5, 2016

Bourque Lab

McGill University Human Genetics Dept.

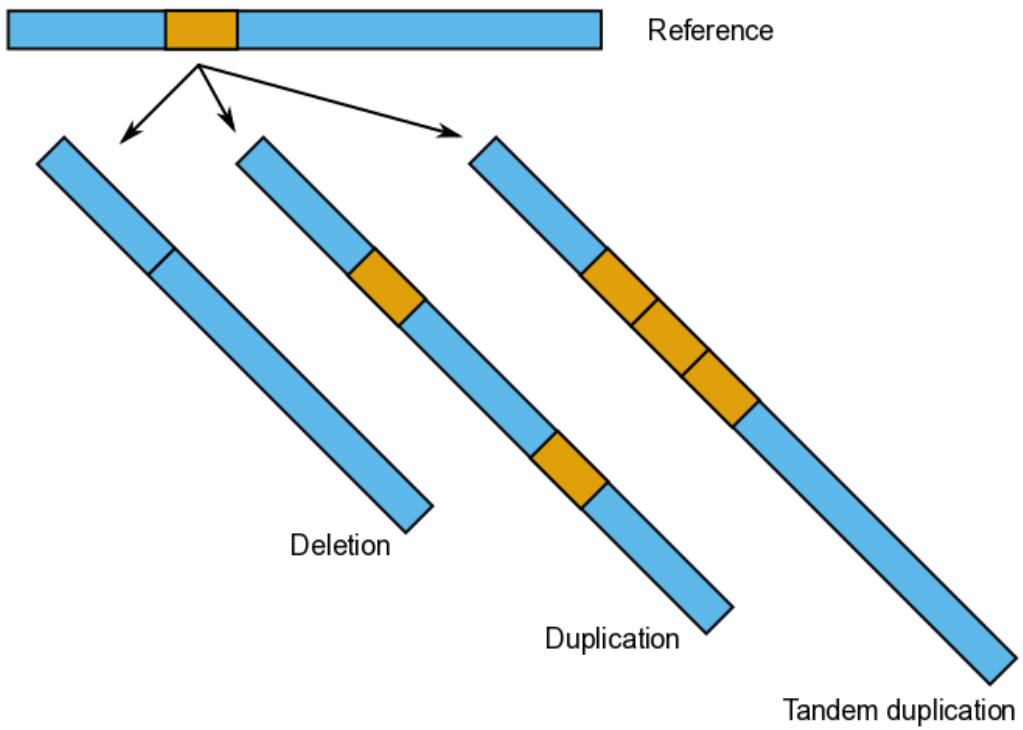
Disclosure Information

I have no financial relationships to disclose

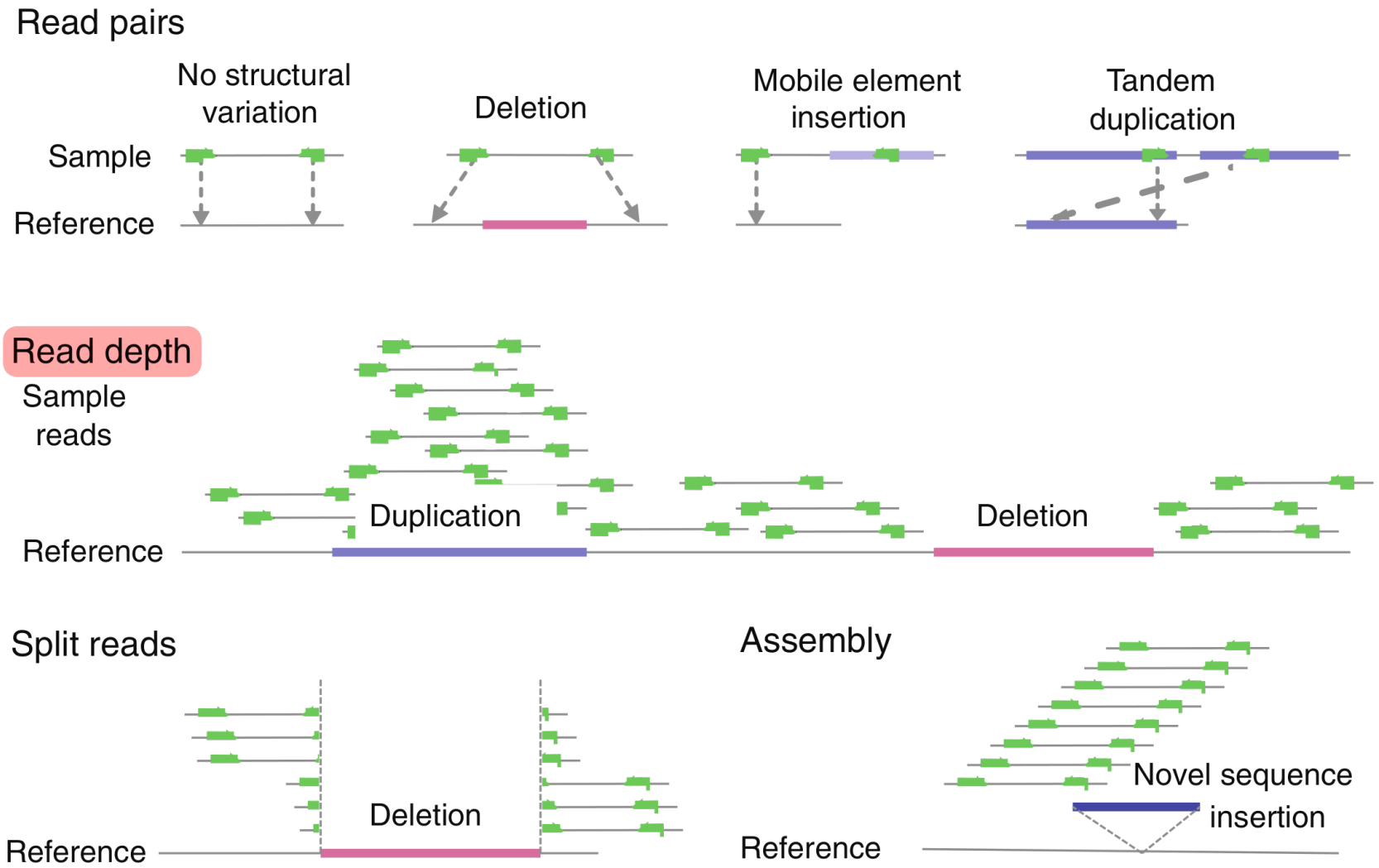
Copy-Number Variation

Copy Number Variation (CNV)

Imbalanced genetic variation involving more than 500bp.



CNV detection from High-Throughput Sequencing



Baker 2012, Nature Methods.

Low-mappability regions

- Repeat-rich regions, centromeres, telomeres.
- $\sim$13% of the human genome.

Low-mappability regions

- Repeat-rich regions, centromeres, telomeres.
- ∼13% of the human genome.

- More prone to CNV.
  - Enriched in **Segmental Duplications** (Sharp *Annual Review* 2006).
  - **Short Tandem Repeats** highly polymorphic (Warbuton *BMC Genomics* 2008).
  - **Transposons** involved in CNV formation (Sen *AJHG* 2006).

Low-mappability regions

- Repeat-rich regions, centromeres, telomeres.
- ∼13% of the human genome.

- More prone to CNV.
  - Enriched in **Segmental Duplications** (Sharp *Annual Review* 2006).
  - **Short Tandem Repeats** highly polymorphic (Warbuton *BMC Genomics* 2008).
  - **Transposons** involved in CNV formation (Sen *AJHG* 2006).

- Involved in phenotype and disease.
  - Short Tandem Repeats and **gene expression** (Gymrek *Nat. Genetics* 2016).
  - Repeats CNV involved in ∼**30 genetic disorders** (Mirkin *Nature* 2007).
  - Retrotransposition in **cancer** (Lee *Science* 2012).

PopSV approach

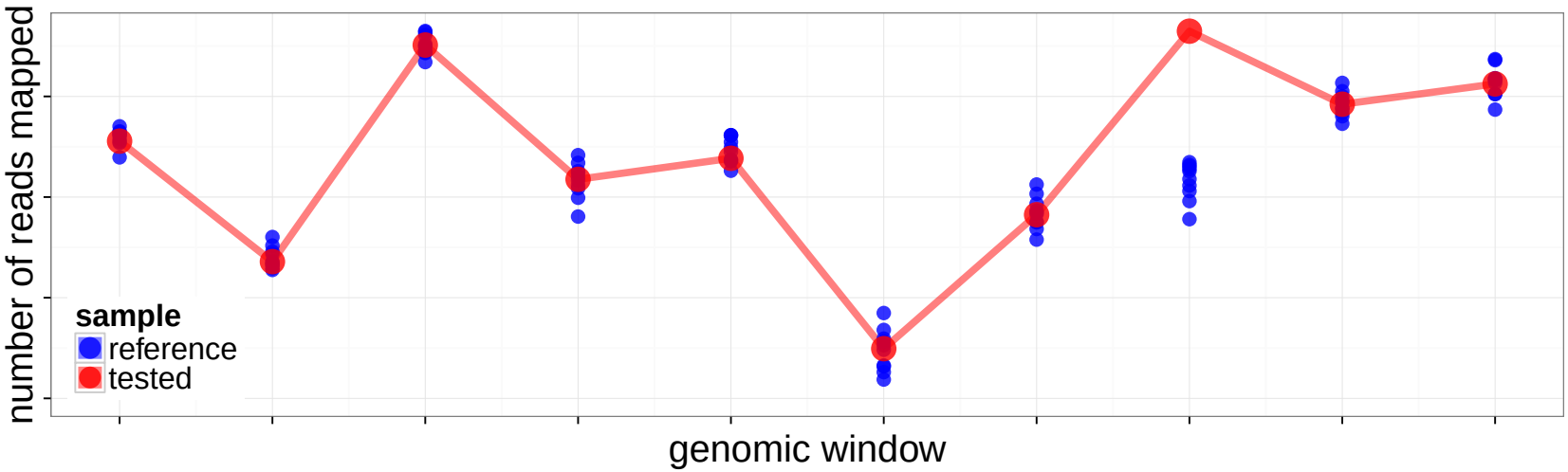
PopSV approach

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.



Benchmark and validation

Existing methods

FREEC LASSO-based segmentation; GC and mappability correction.

cn.MOPS Multi-sample Bayesian-based segmentation.

Whole-Genome Sequencing data

- 45 samples, including **10 twin families** (i.e 2 twins + 2 parents).
- 95 pairs of **normal/tumor samples** from Renal Cell Carcinoma (CageKid).

Benchmark and validation

- **Replication** in the **twins**.
- Concordance with pedigree.

- Replication in the **paired tumor**.

- Concordance of **different bin sizes**
- **PCR** validation.

- **Overall** performance and in **different repeat context**.

Validation conclusions

- PopSV detects **3-5x more variants**.
- **Wider genomic range**.
- **Robust** across challenging regions:
  - Low-coverage.
  - Segmental duplications.
  - DNA satellites.
  - Short tandem repeats
  - GC-rich/poor.
- **Resolution** down to half the bin size.

CNV patterns in normal genomes

CNV in normal genomes

640 normal genomes

- 45 samples from the Twin study (∼40X)
- 95 normal samples from Renal Cell Carcinoma (∼54X).
- 500 unrelated samples from GoNL (∼14X).

CNV in normal genomes

640 normal genomes

- 45 samples from the Twin study (∼40X)
- 95 normal samples from Renal Cell Carcinoma (∼54X).
- 500 unrelated samples from GoNL (∼14X).

Where are CNVs located ?

In Centromere ? Telomere ? Segmental duplication ?
DNA satellites ? Short tandem repeats ?
Transposable Elements ? Exons ? Promoters ?

CNV in normal genomes

640 normal genomes

- 45 samples from the Twin study (∼40X)
- 95 normal samples from Renal Cell Carcinoma (∼54X).
- 500 unrelated samples from GoNL (∼14X).

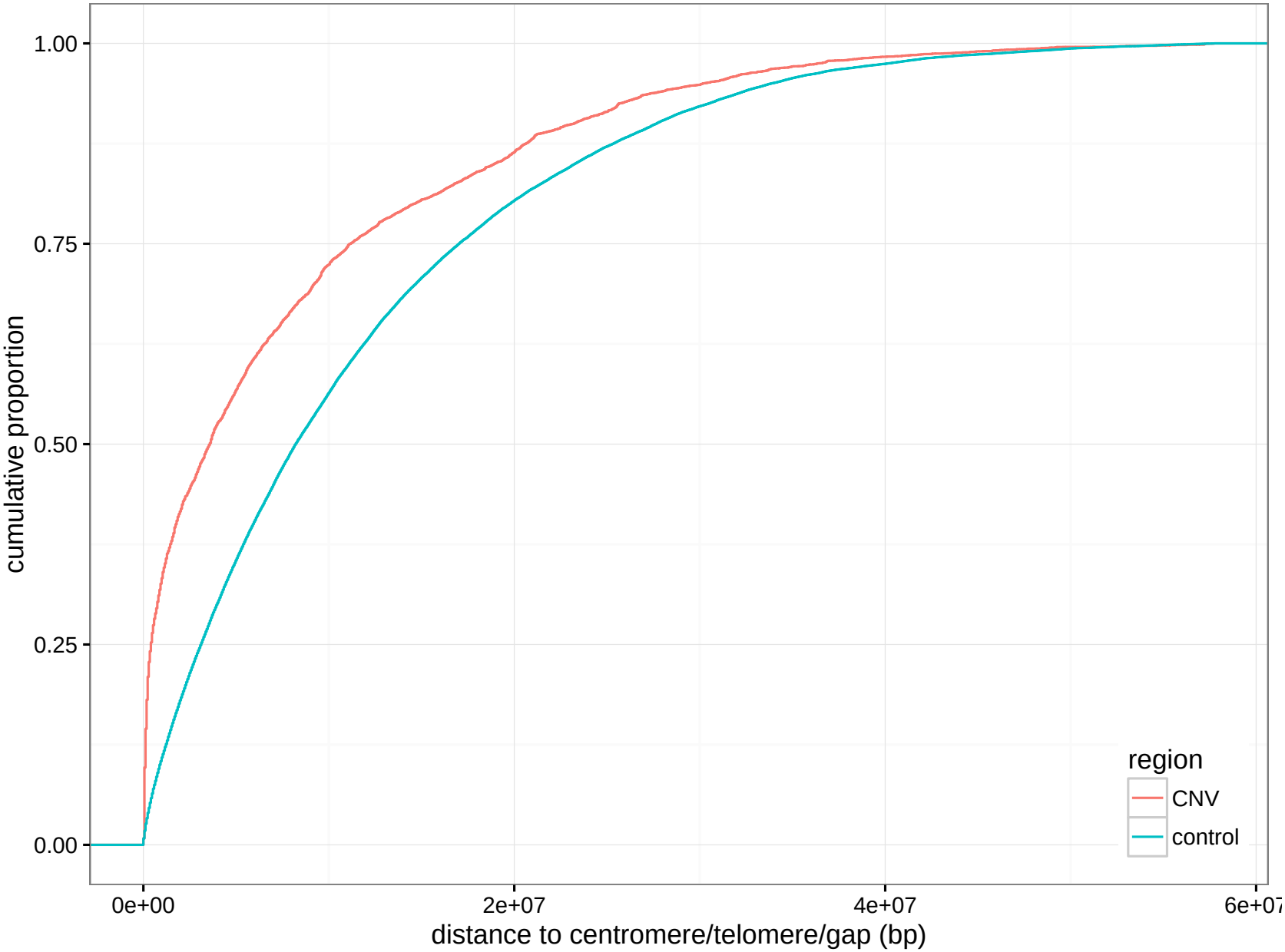
Where are CNVs located ?

In Centromere ? Telomere ? Segmental duplication ?
DNA satellites ? Short tandem repeats ?
Transposable Elements ? Exons ? Promoters ?

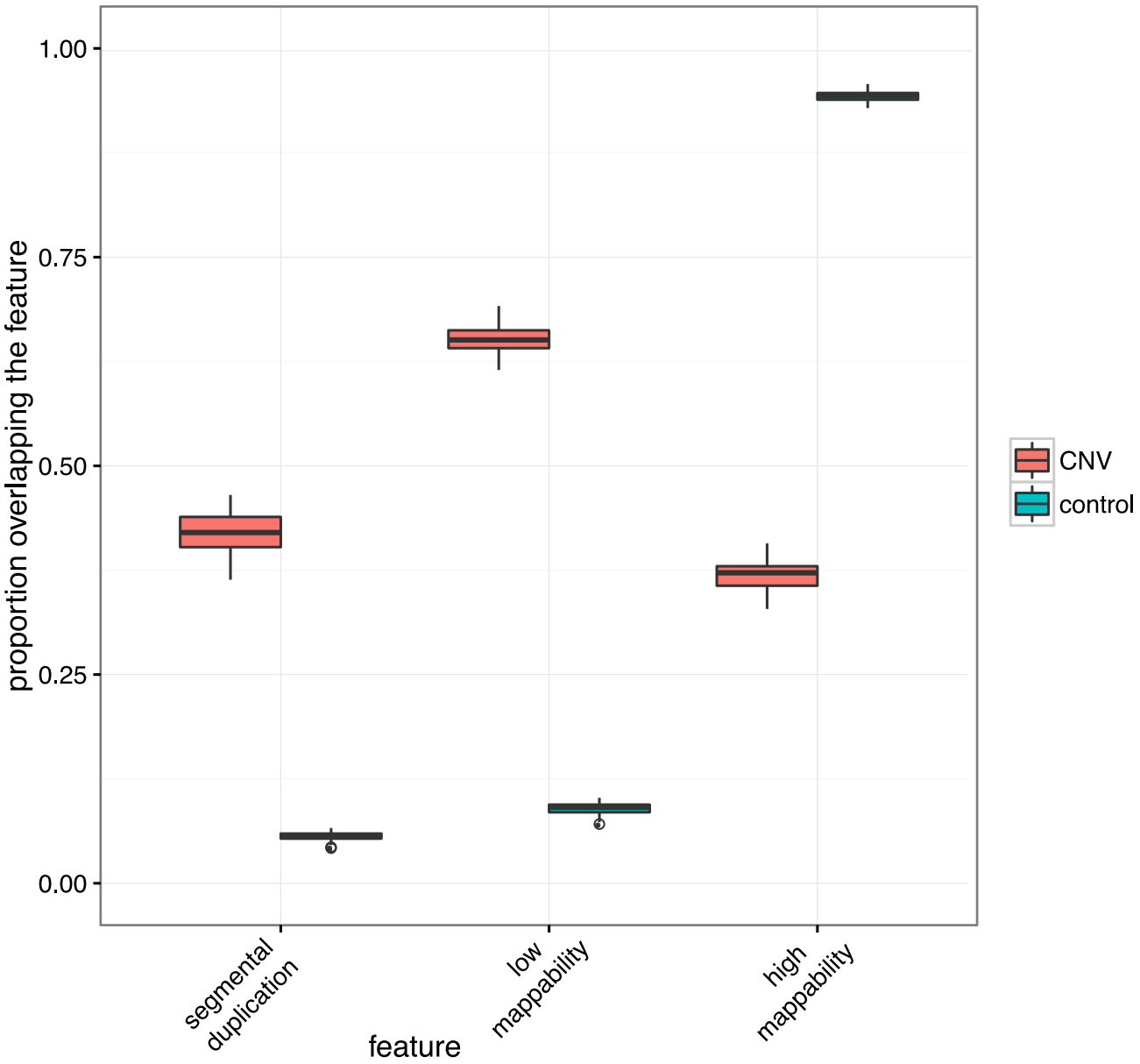
Control regions

- Same size distribution.
- Randomly distributed.

Enriched close to Centromere/Telomere/Gap (CTG)



Enriched in SD and low-coverage regions



Going further

1. Control for the SD and CTG patterns.
2. Look at other repeat classes.

Control regions

- Randomly distributed.
- Same size distribution.

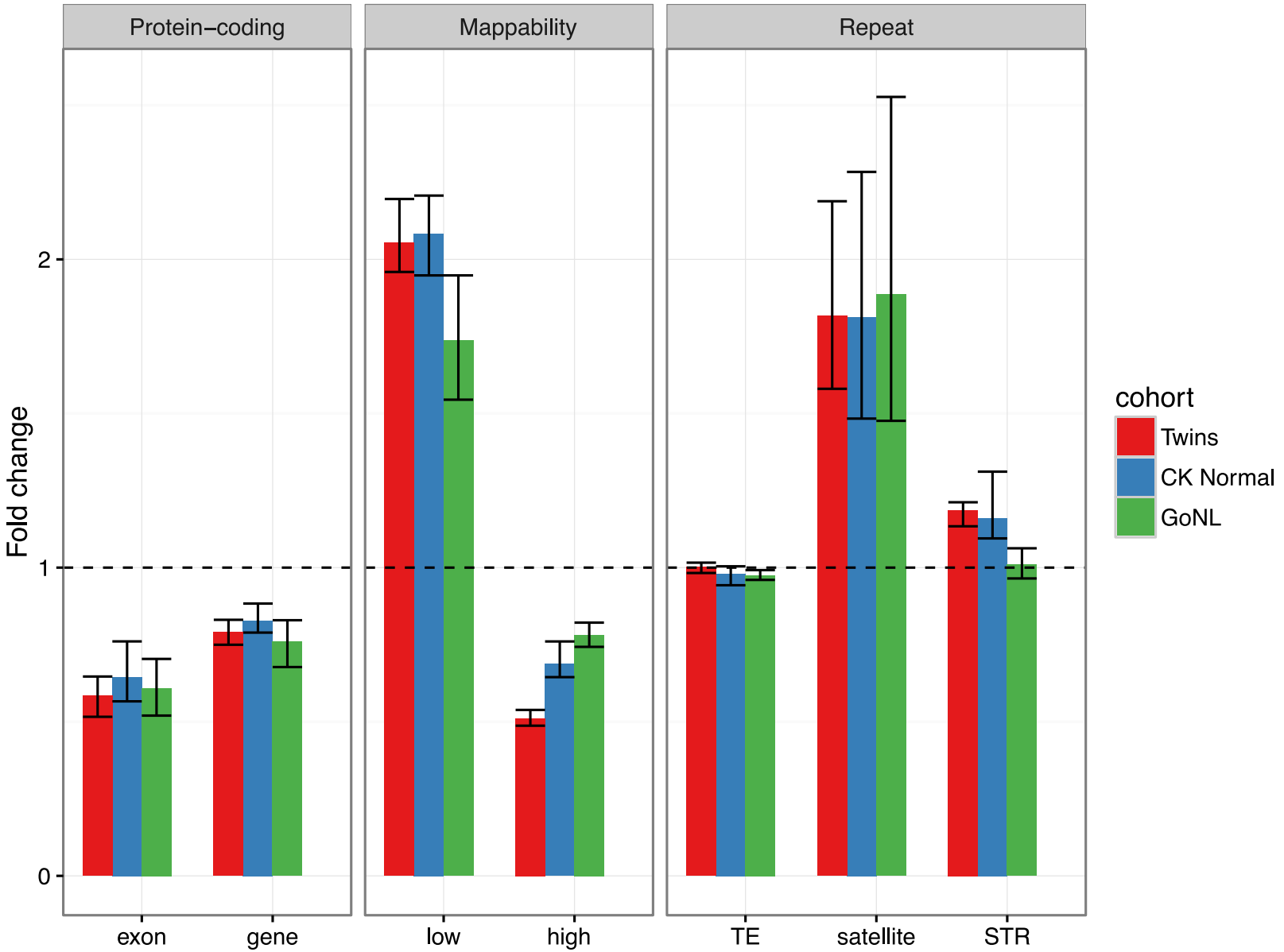
Going further

1. Control for the SD and CTG patterns.
2. Look at other repeat classes.

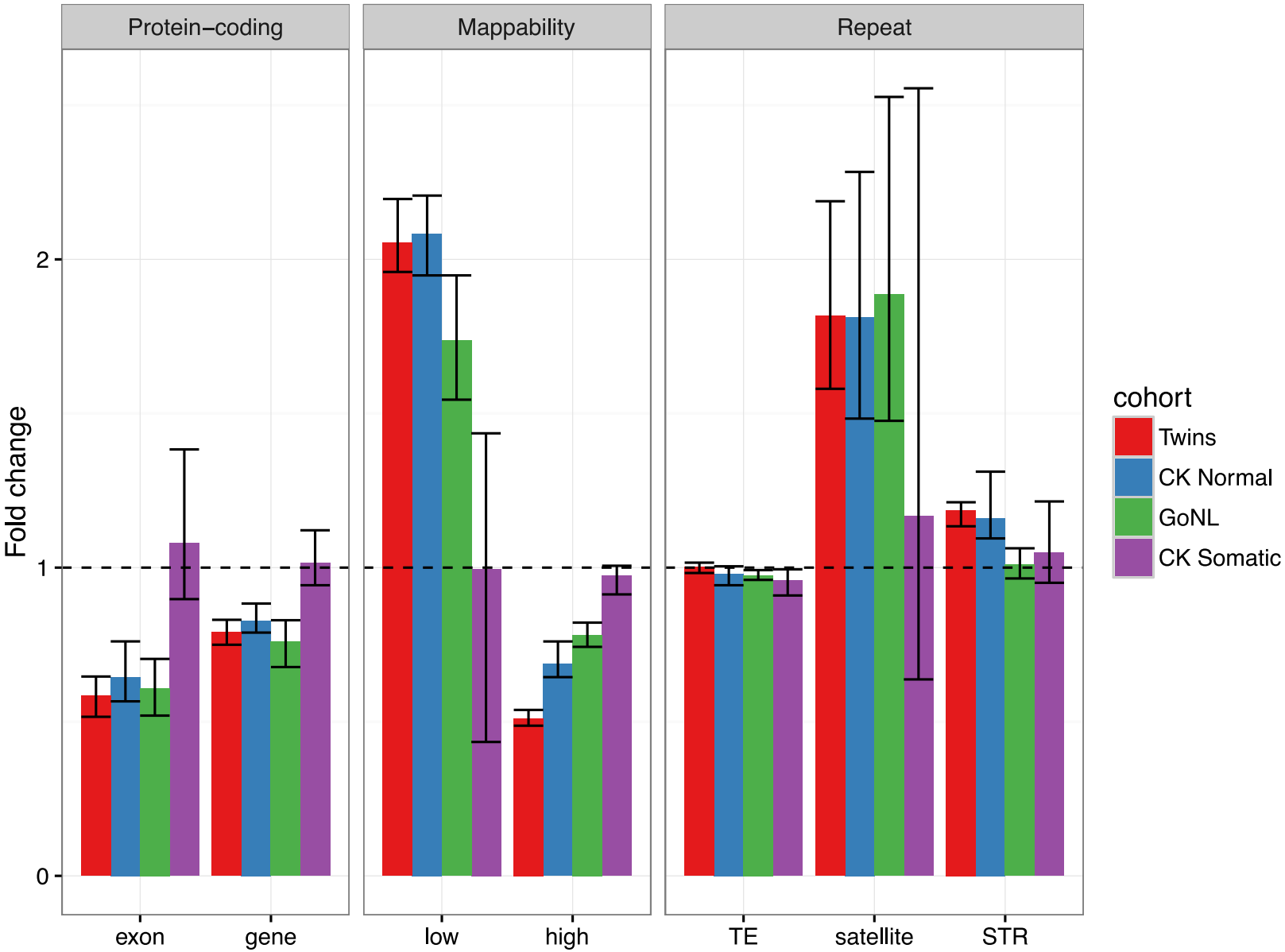
Control regions

- Randomly distributed.
- Same size distribution.
- Same proportion **overlapping a segmental duplication**.
- Similar **distance to CTG**.

Controlling for SD and distance to CTG



Controlling for SD and distance to CTG



Controlling for SD and distance to CTG

- **Satellites** enrichment driven by **ALR/Alpha**, **(GAATG)n/(CATTC)n** families.
- **Short Tandem Repeats**
  - Enrichment **distributed** across families...
  - ... but **stronger for larger STR**.
- Transposable elements (TE):
  - **SVA** class enriched.
  - Expected: *L1HS*, *L1PA2* to *L1PA5*.
  - Surprises: *HERVH*, *LTR38*, *LTR4*.

Repeat CNVs and protein-coding genes

Set	CNVs	Genes with CNVs		
		Exon	+ Promoter	+ Intron
All CNVs	91733	7206	11341	13259
Low coverage	26888	682	1151	1977
Extremely low coverage	10010	**347**	465	521
STR	4286	45	286	748
Satellite	1822	2	21	33
TE	20491	164	1747	3998
STR/Satellite/TE	22313	**166**	1760	4014

Repeat CNV: more than 90% of the CNV is annotated as repeat.

Conclusion

Summary

- PopSV
  - uses reference samples.
  - detects more CNVs.
  - is robust across the entire genome.

Summary

- PopSV
  - uses reference samples.
  - detects more CNVs.
  - is robust across the entire genome.

- In normal genomes:
  - CNVs enriched in **low coverage regions**.
  - Specific enrichment in **satellites, simple repeats, TEs**.

  - **Not** due to segmental duplication enrichment.
  - **Replicated** across datasets but **different from somatic** patterns.

  - Some CNVs in low coverage regions or repeats **hit exonic sequence**.

- **Guillaume Bourque**
- Mathieu Bourgey
- Louis Letourneau
- Francois Lefebvre
- Eric Audemard
- Toby Hocking
- Simon Girard

- Patrick Cossette
- Guy Rouleau
- Caroline Meloche

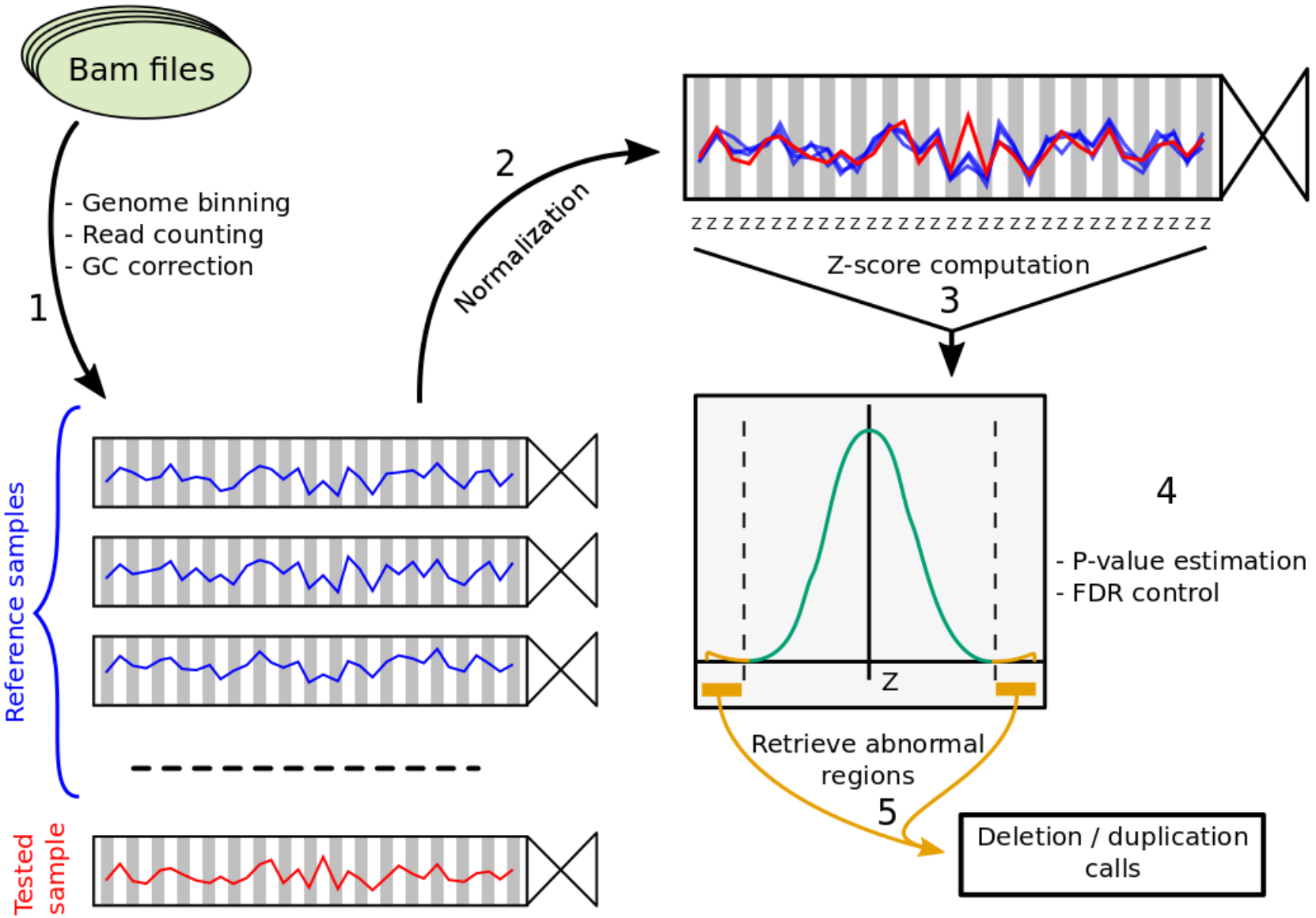
- Simon Gravel
- Mathieu Blanchette

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

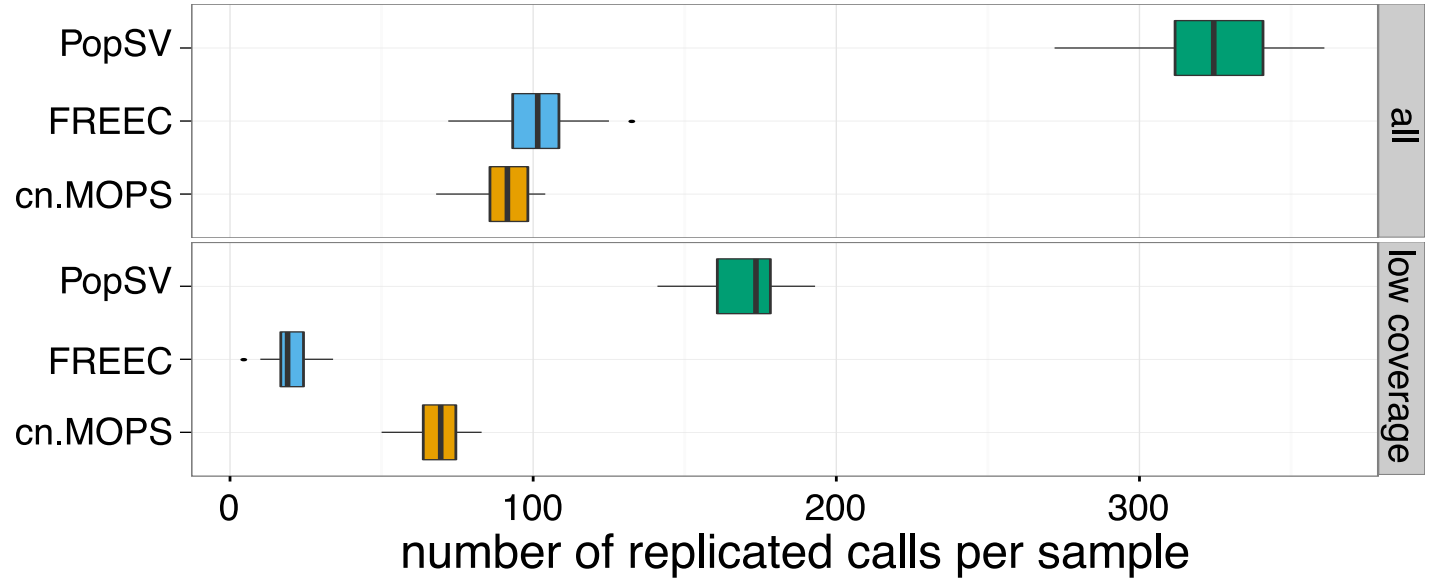
GenomeQuébec

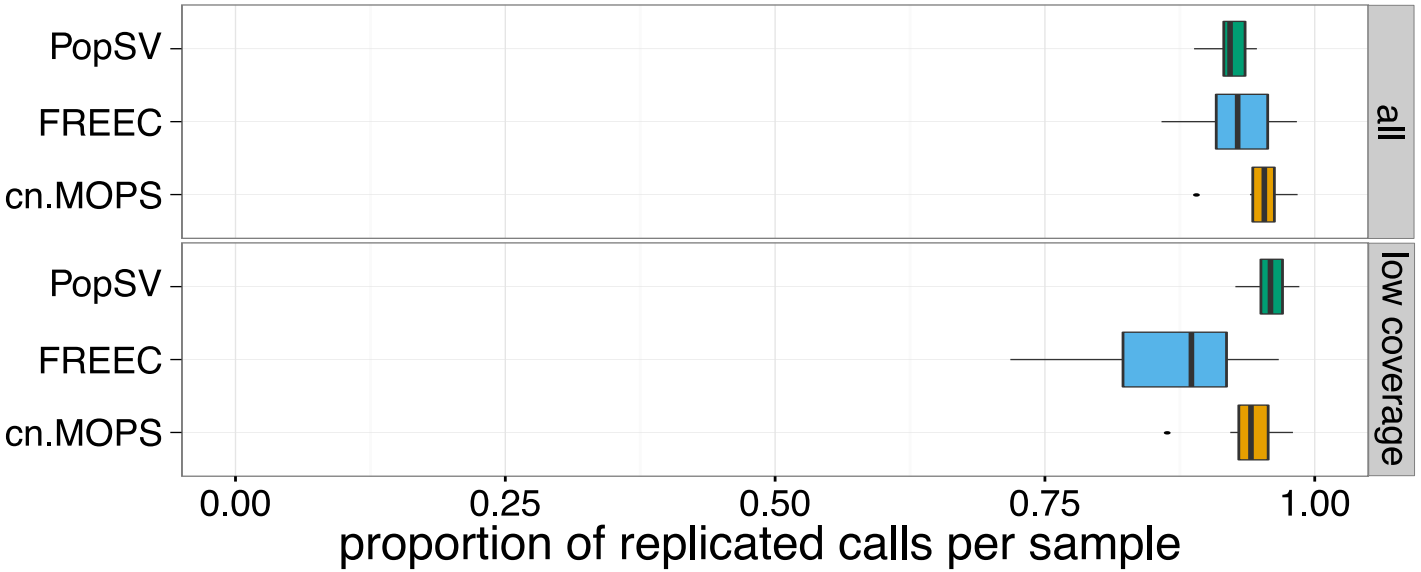
NSERC CRSNG

Workflow

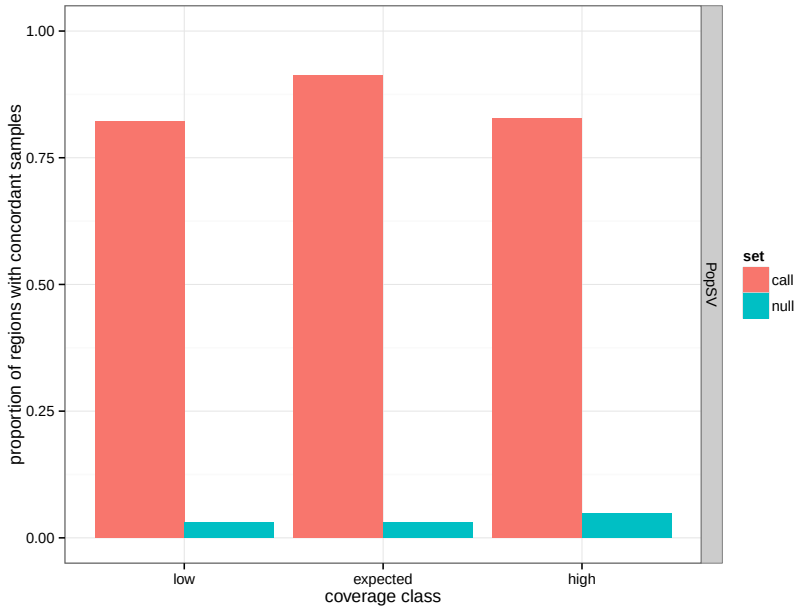

Bam files
1
- Genome binning
- Read counting
- GC correction
2
Normalization
Z-score computation
3
Reference samples
Tested sample
4
- P-value estimation
- FDR control
Z
Retrieve abnormal regions
5
Deletion / duplication calls

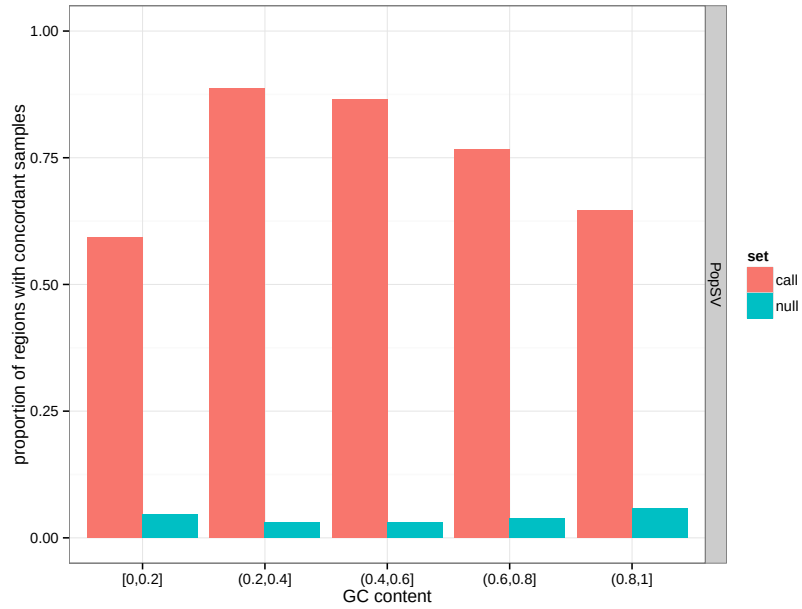
Replication in twins

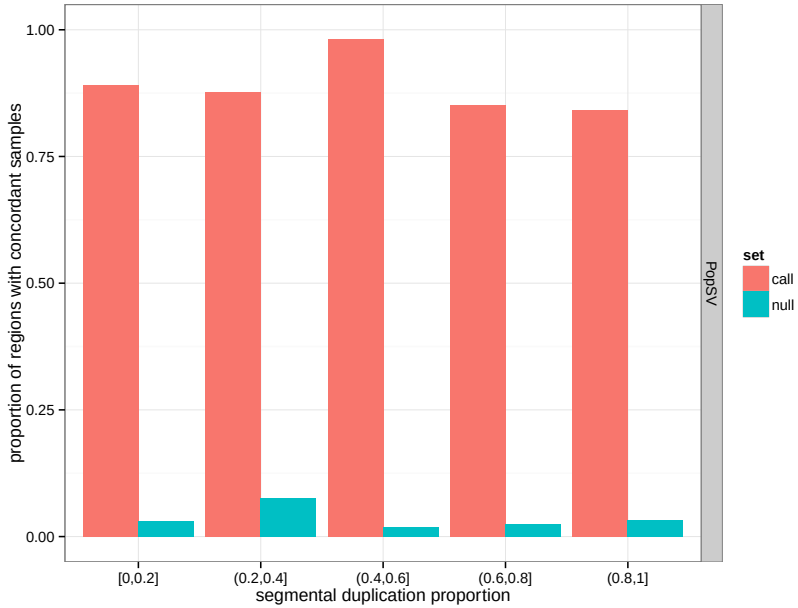


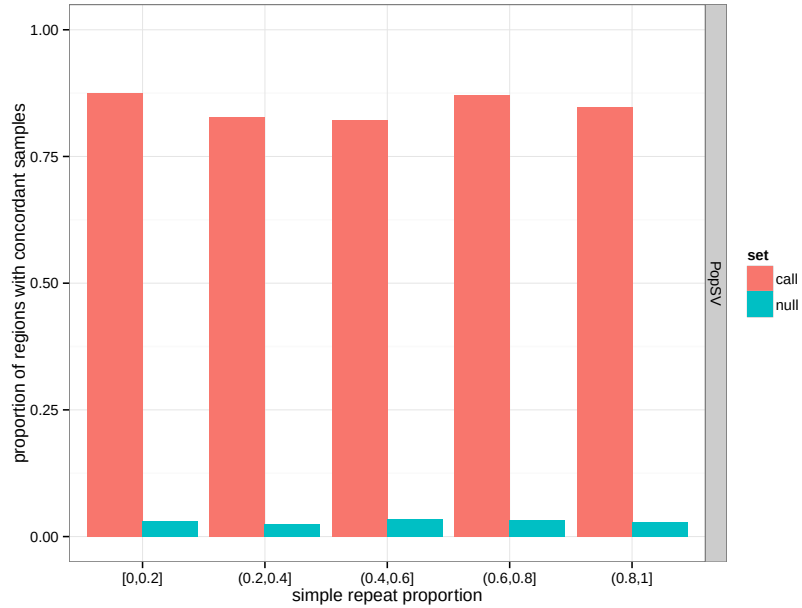


Robust across challenging regions

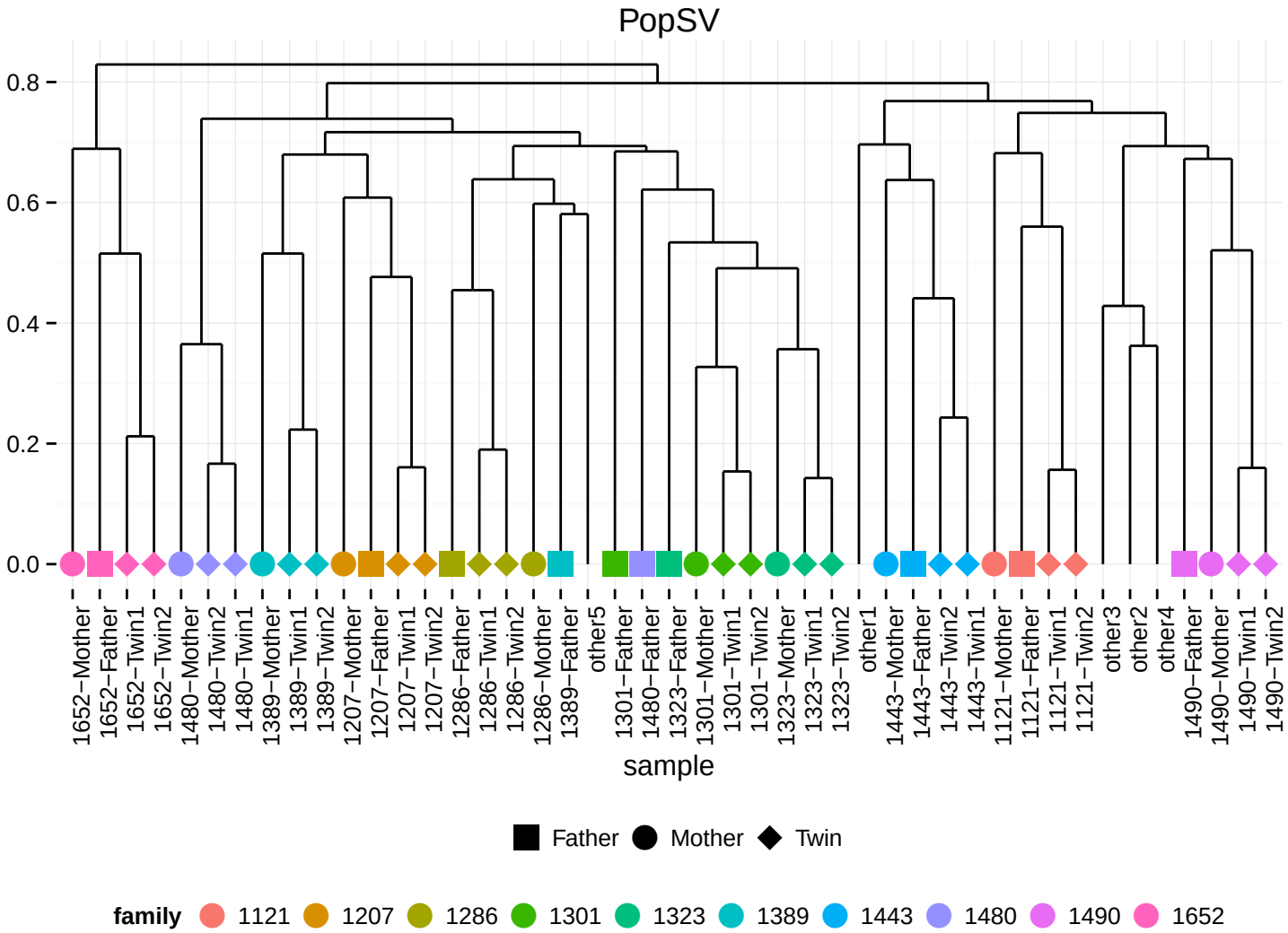






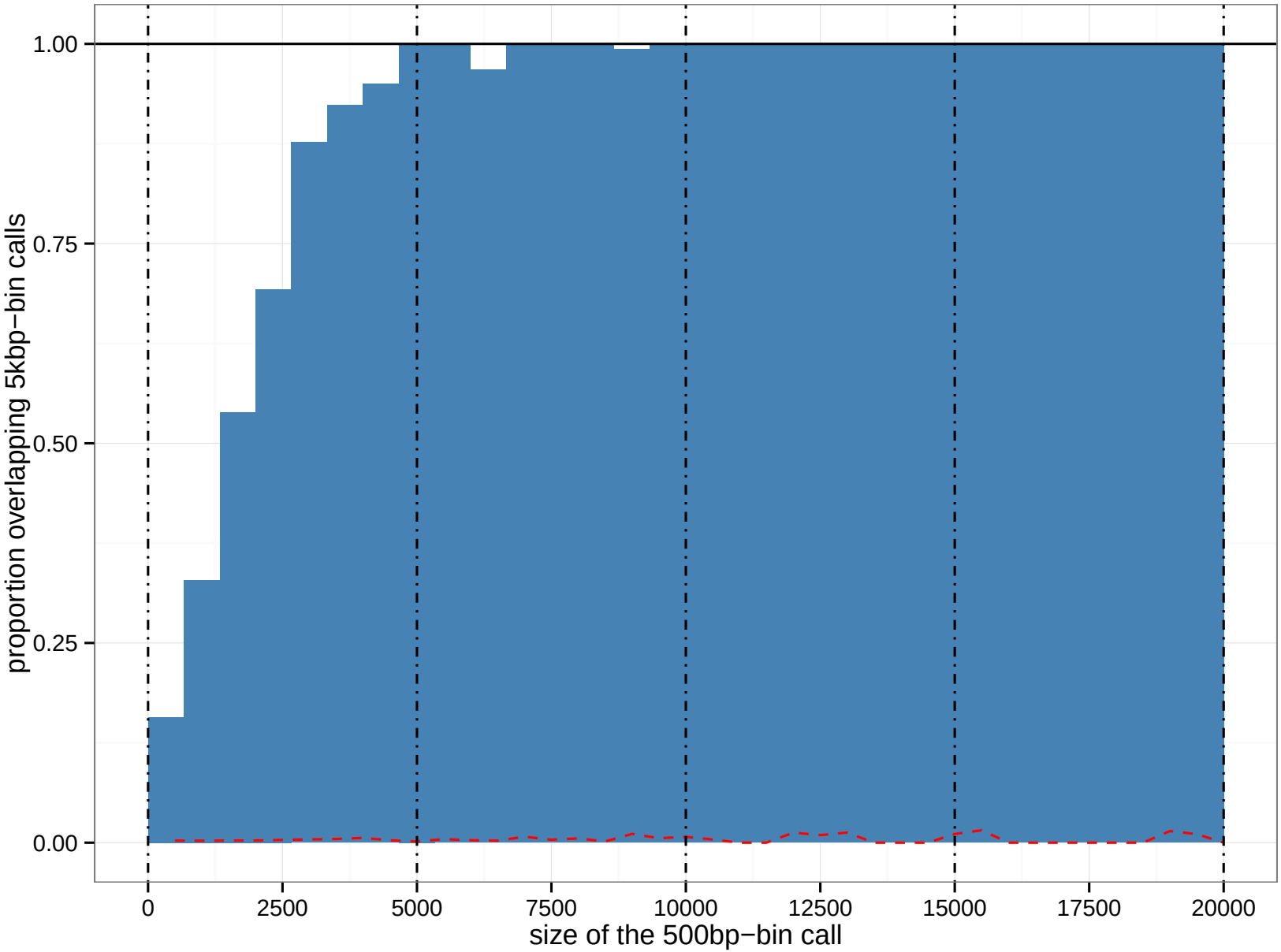


Robust across challenging regions

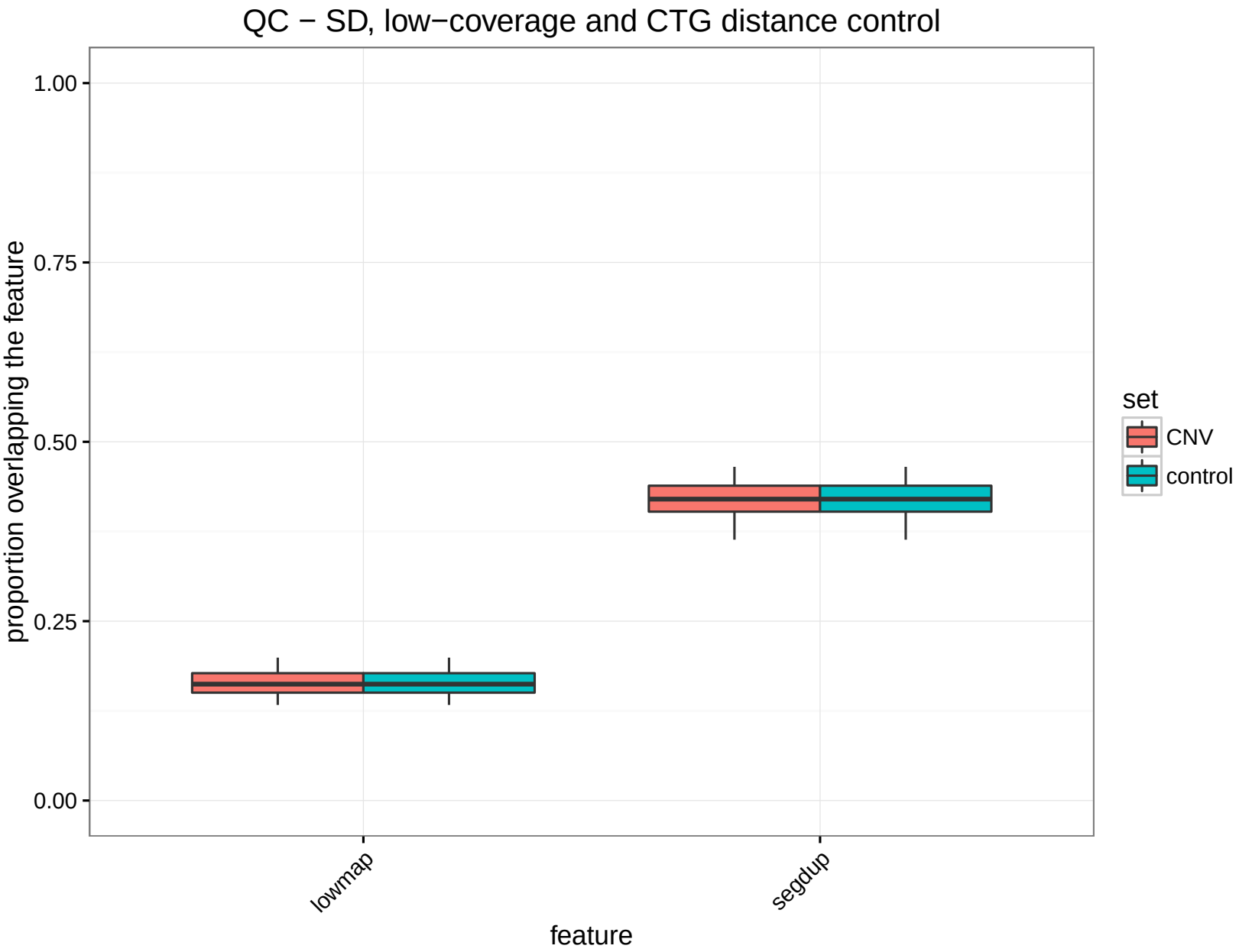


Using only CNVs in extremely low coverage regions !

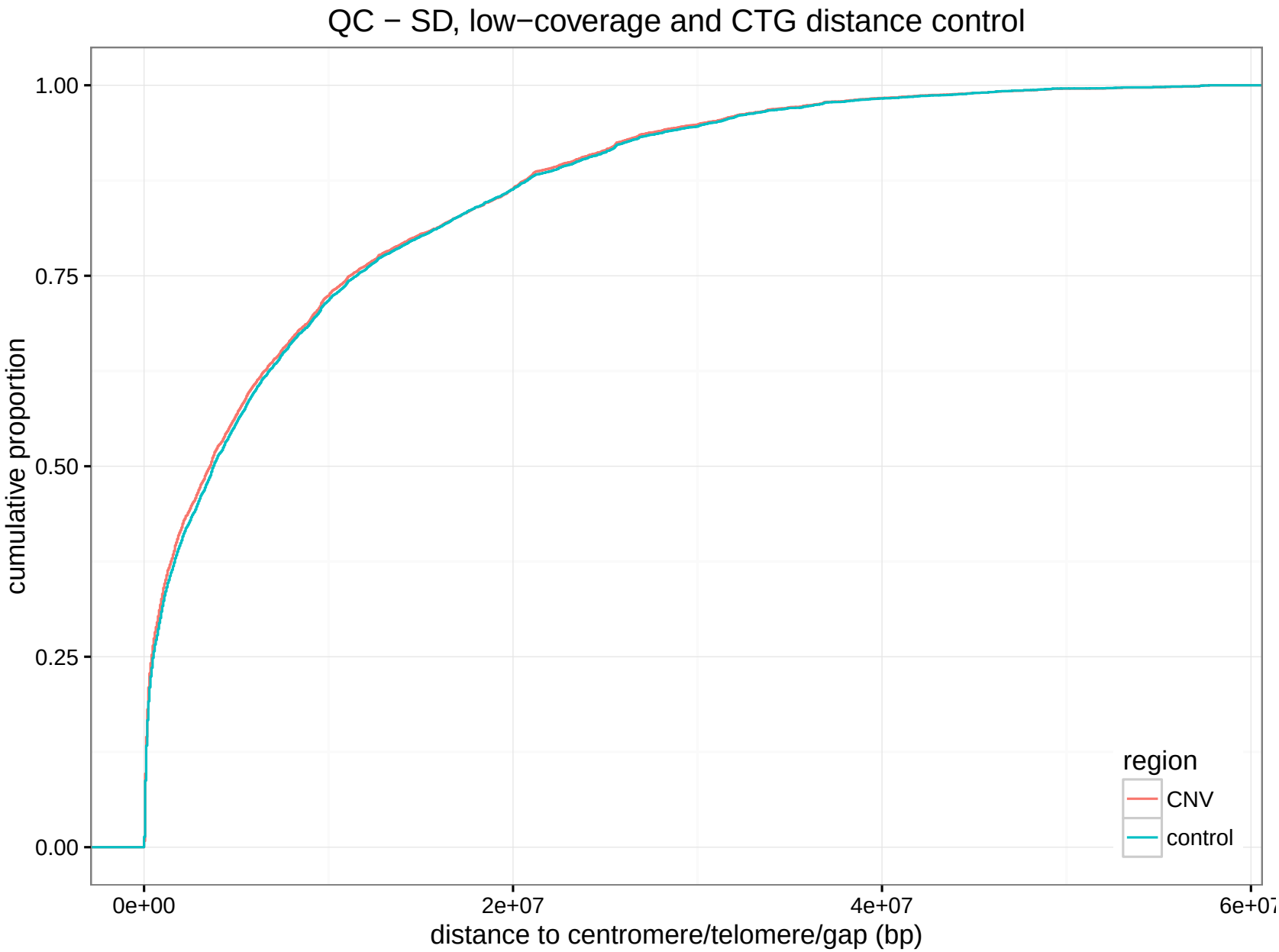
Resolution - 500 bp bins Vs 5 Kbp bins



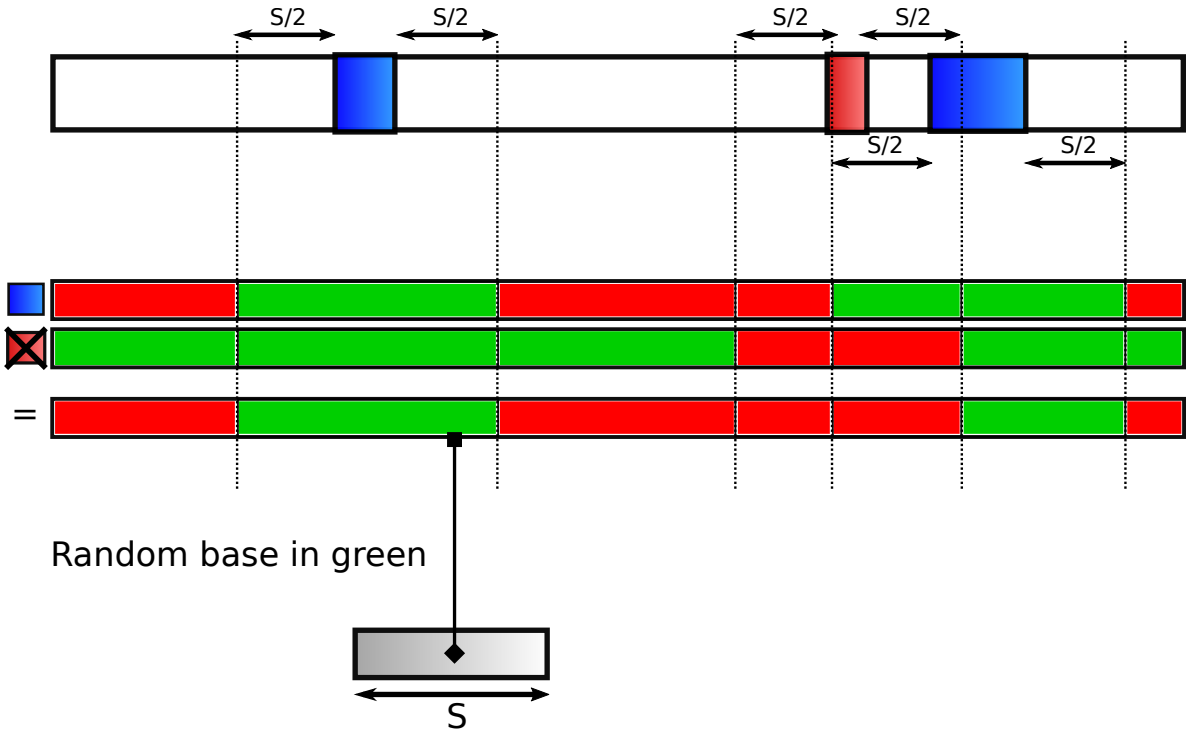
Control regions



Control regions

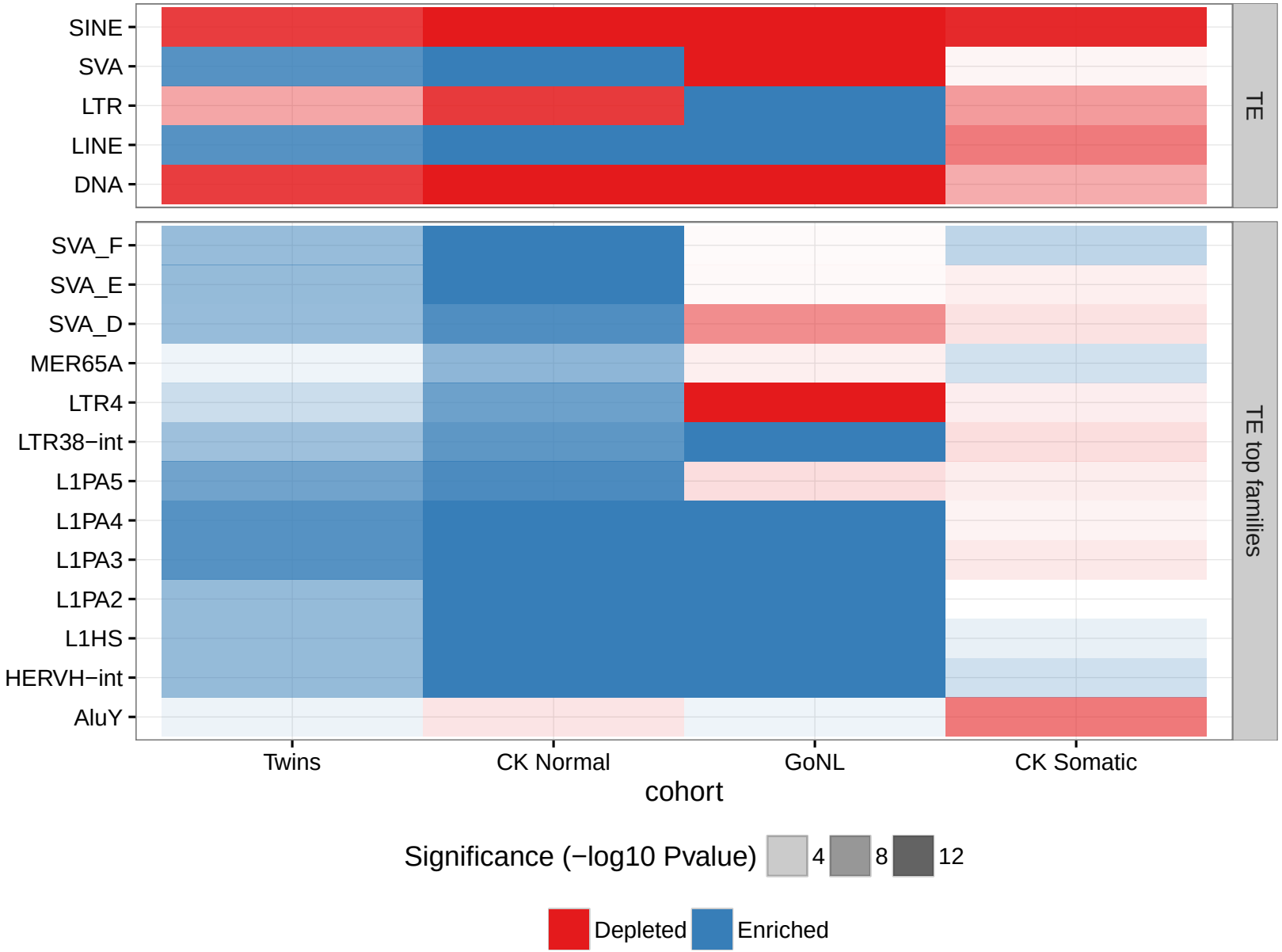


Control regions

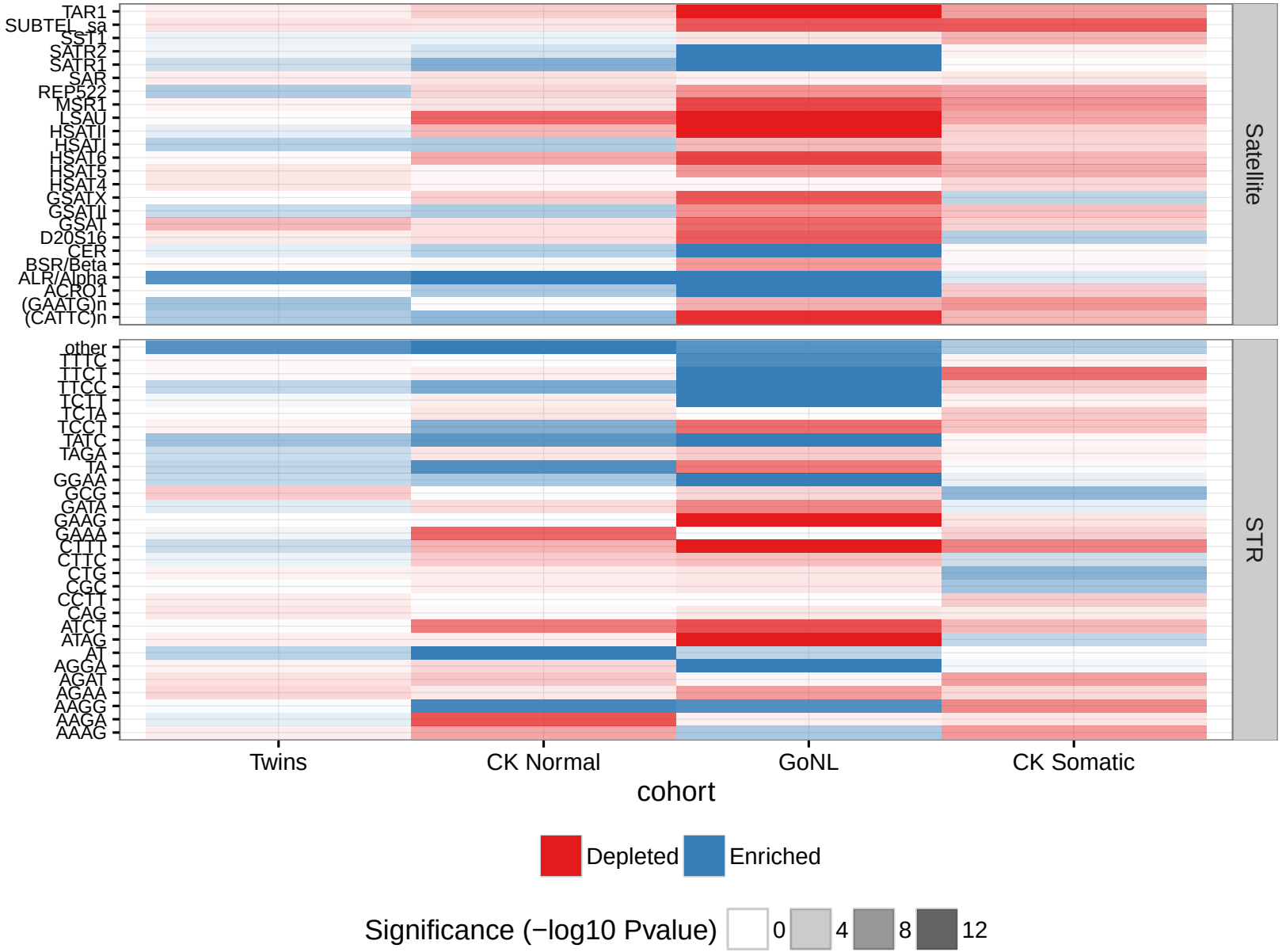


Random region of size S

Controlling for SD and distance to CTG



Controlling for SD and distance to CTG



Distance to CTG for somatic CNVs