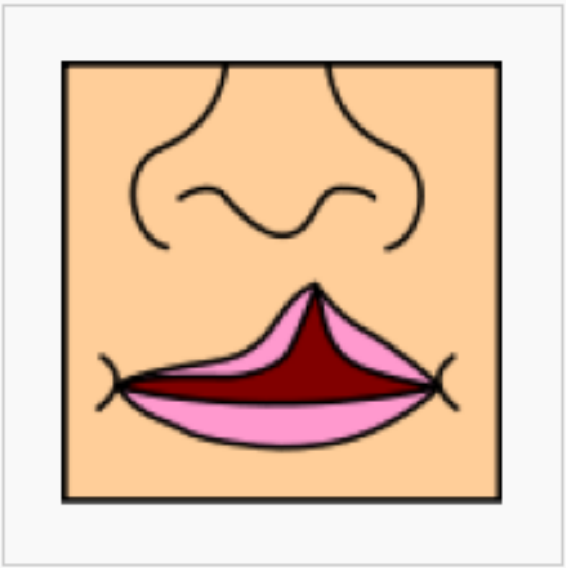
Association of Denovo Copy Number Variants with Oral Clefts

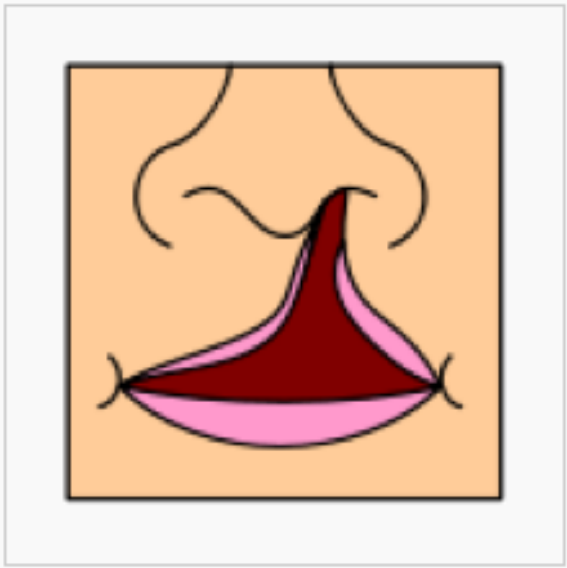
Rob Scharpf

June 24, 2012

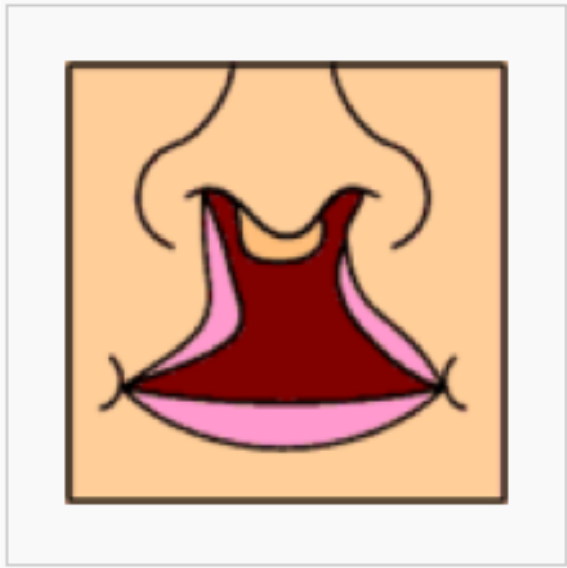
Cleft lip



Unilateral incomplete



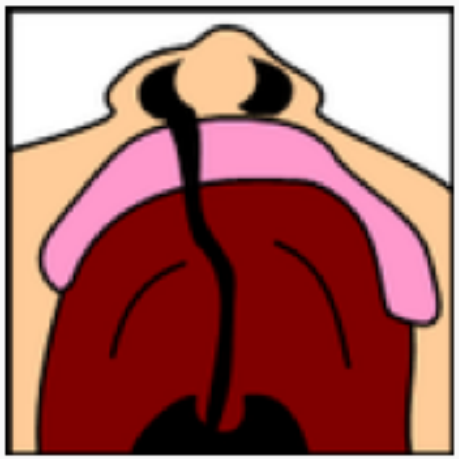
Unilateral complete



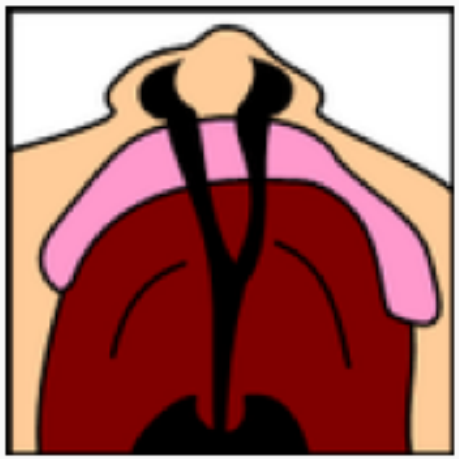
Bilateral complete

Cleft palate, cleft lip and palate

Incomplete cleft palate



Unilateral complete lip and palate



Bilateral complete lip and palate

Goal

To identify structural variants contributing to oral cleft:

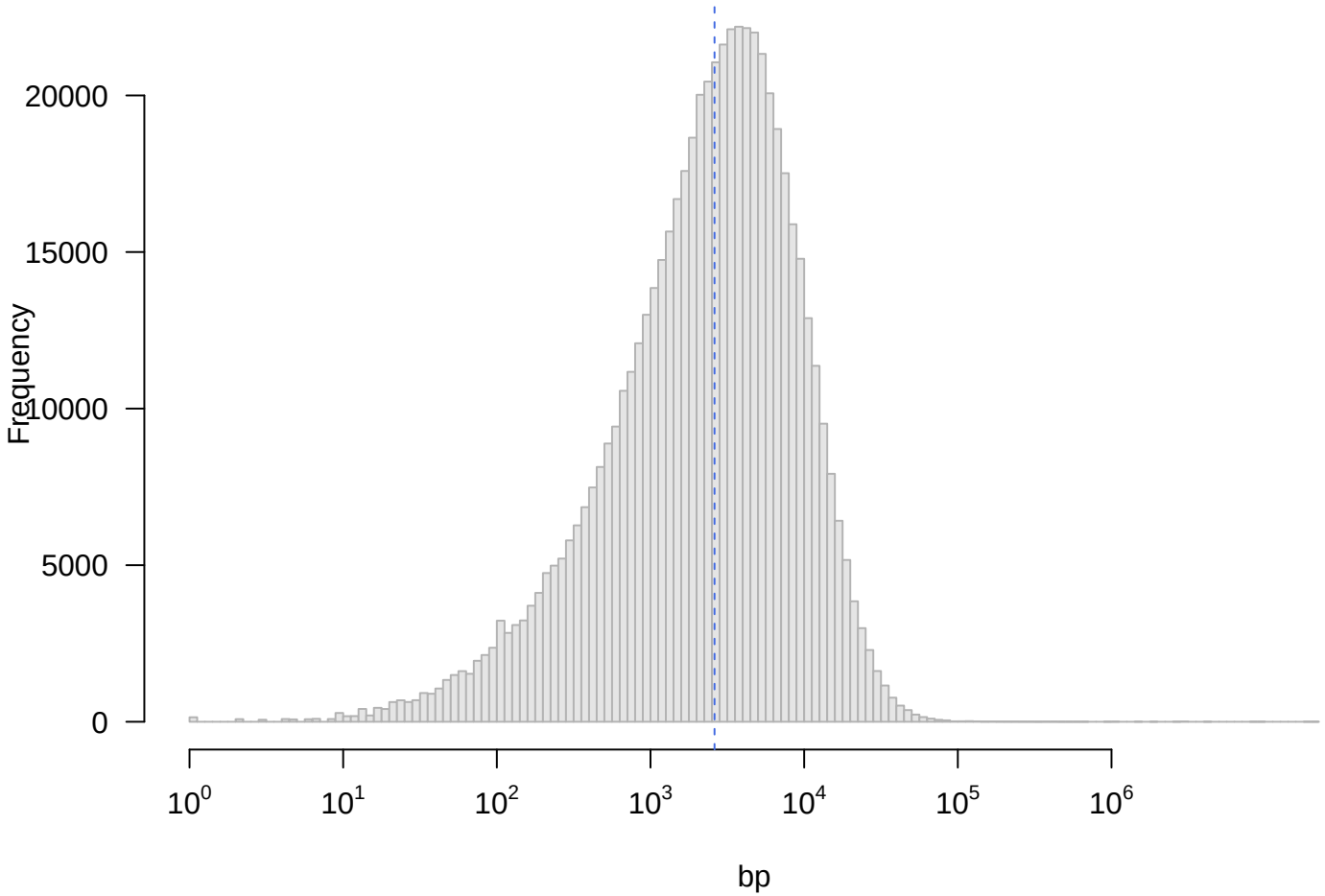
- denovo CNV discovery (by-sample)
- compare denovo frequency among oral cleft offspring to *similar* controls

Sample size

	Cleft	Controls
Discovery	2082	752
WGA	120	99
MAD $>$ 0.3	212	138
non-EA	1090	-
n	660	515

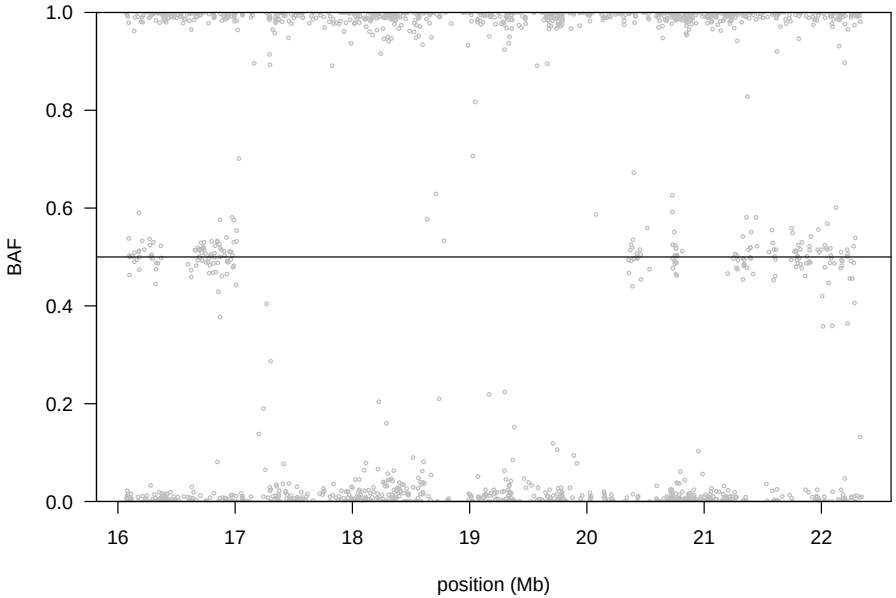
- cell frequencies are the number of Father-Mother-Offspring trios
- The controls are trios with dental caries.

Illumina 610 quad: Distance between markers

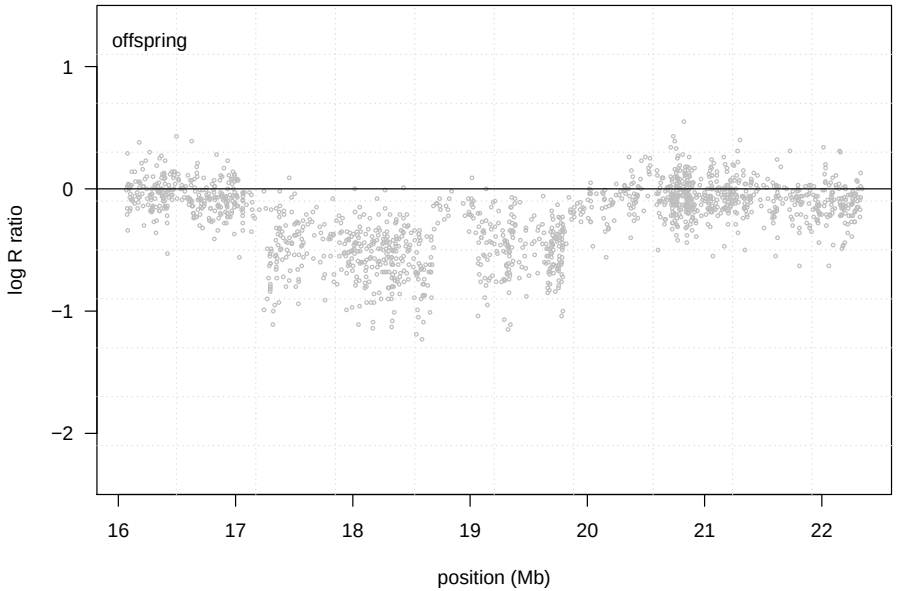


Low level summaries

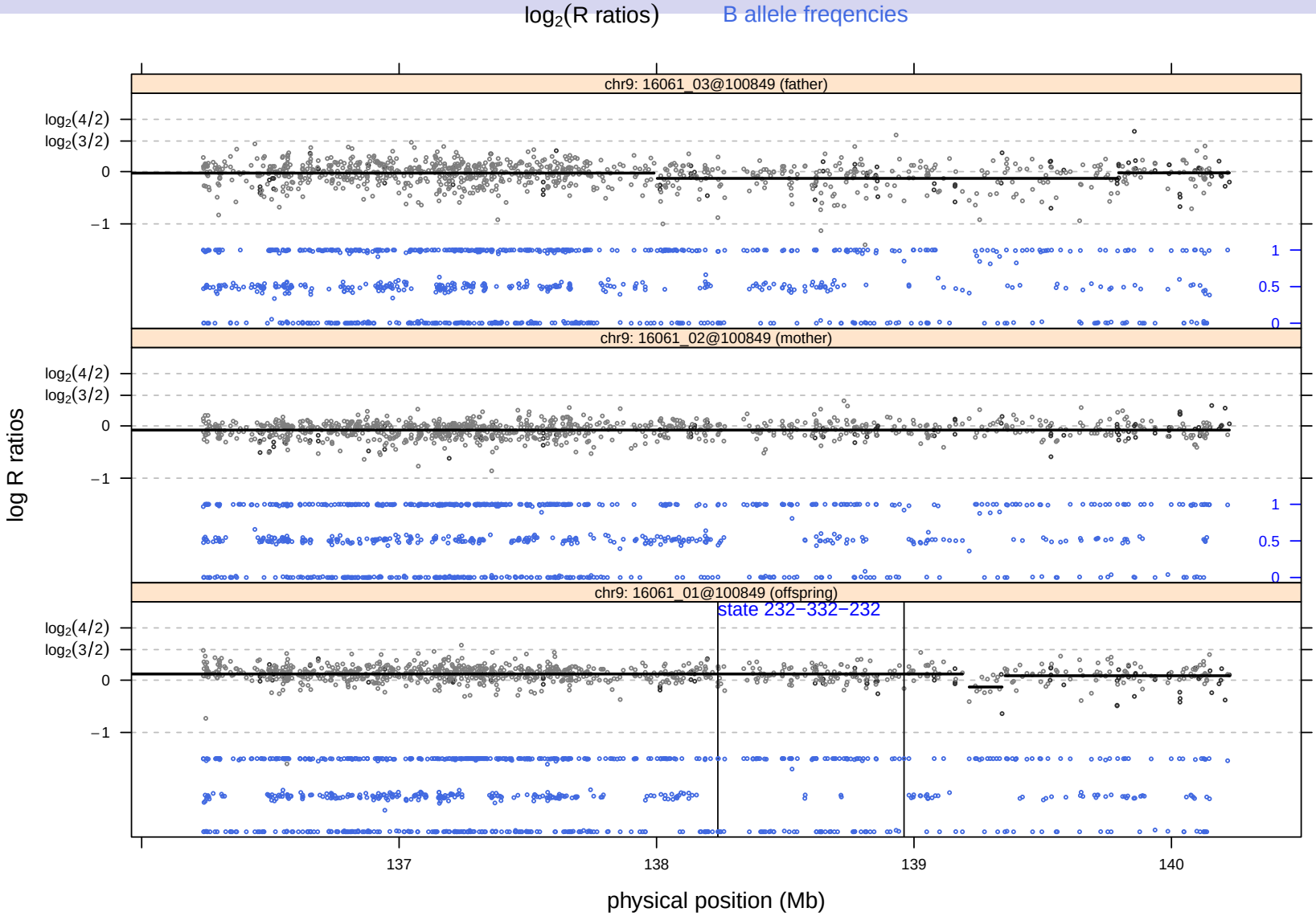
B allele frequencies



$\log_2$ R ratios

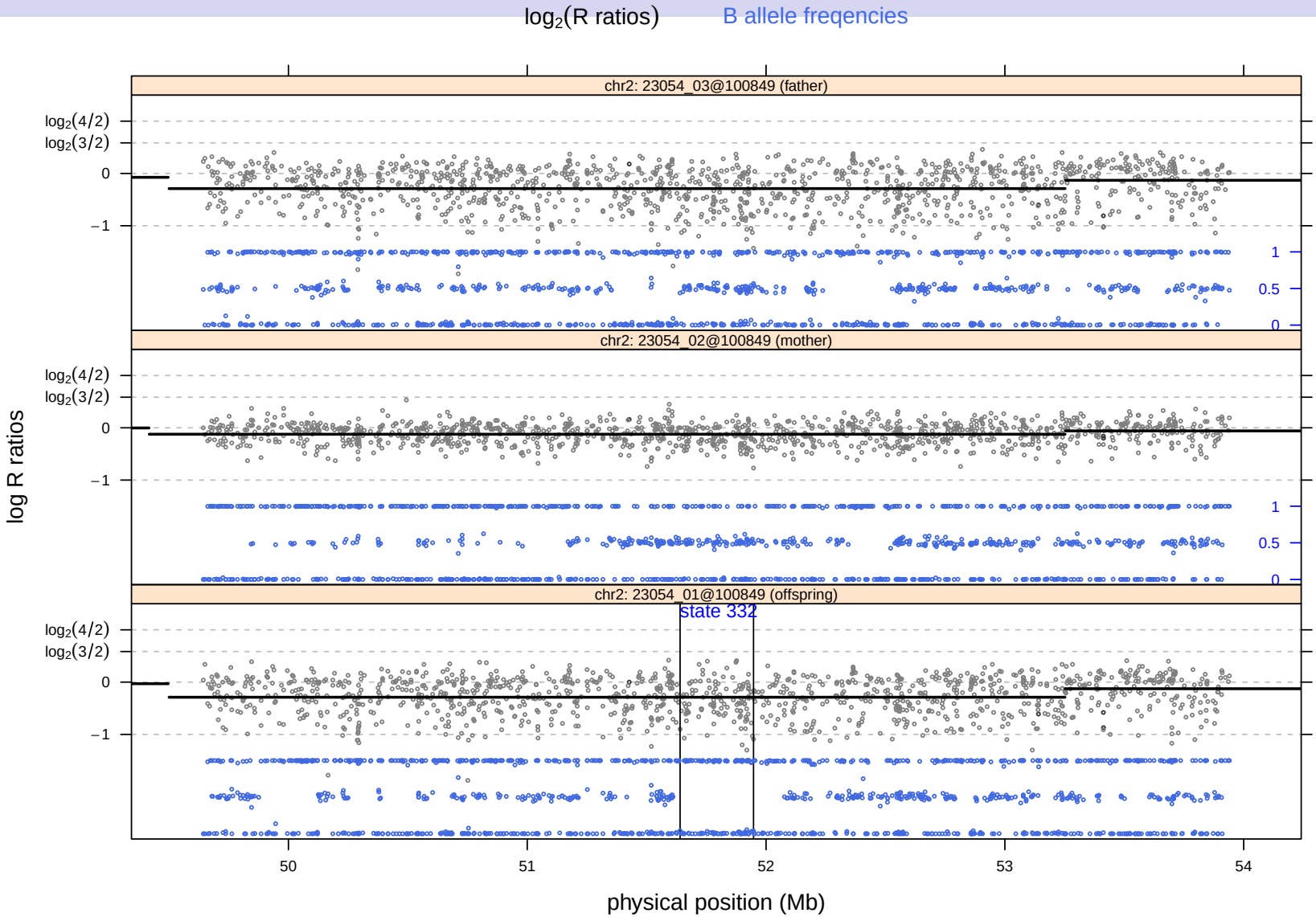


Data for case-parent trio: false positive



PennCNV joint HMM: '332' is denovo hemizygous deletion

False positives



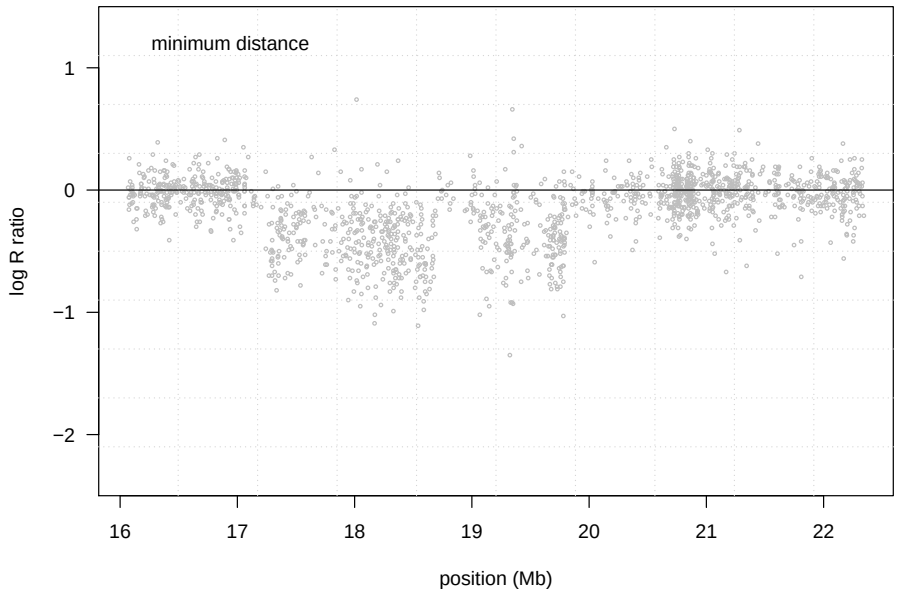
Computational characteristics

- The joint HMM is $\approx$ 3 hours per trio
- 3 * 2082 trios $\approx$ 6,246 CPU hours

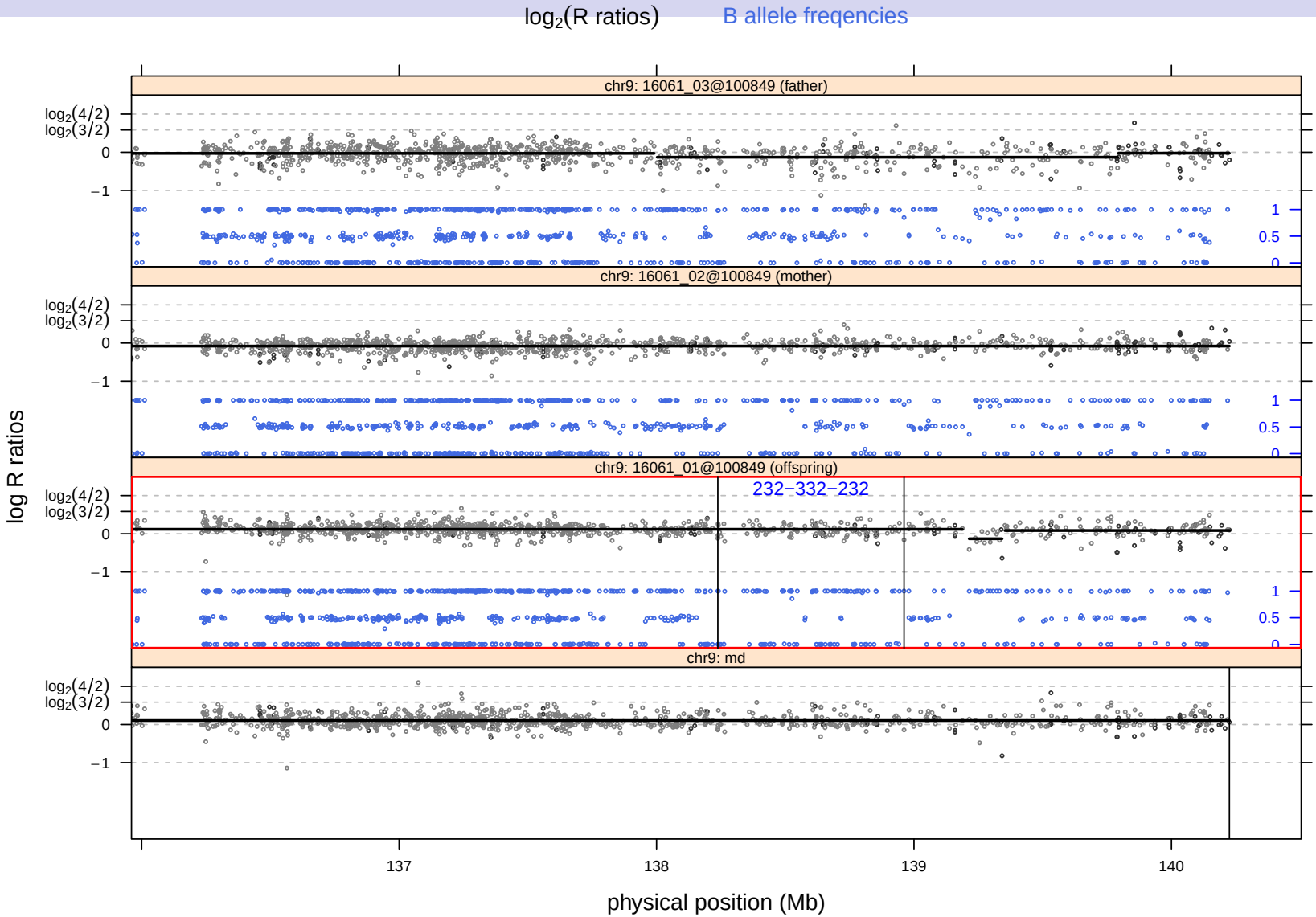
```
$ qstat -u hschwend
job-ID   name       user        state submit/start at
-----------------------------------------------------
 371289 Top10.sh   hschwend     r     10/30/2009 00:12:21
 371291 reasons.sh hschwend     r     10/30/2009 00:14:39
 371292 for.sh     hschwend     r     10/30/2009 00:14:39
 371293 working.sh hschwend     r     10/30/2009 00:15:09
 371294 late.sh    hschwend     r     10/30/2009 00:15:09
 371295 and.sh     hschwend     r     10/30/2009 00:15:09
 371296 not.sh     hschwend     r     10/30/2009 00:15:09
 371297 going.sh   hschwend     r     10/30/2009 00:15:25
 371298 to.sh      hschwend     r     10/30/2009 00:15:25
 371299 bed.sh     hschwend     r     10/30/2009 00:15:25
```

The minimum distance

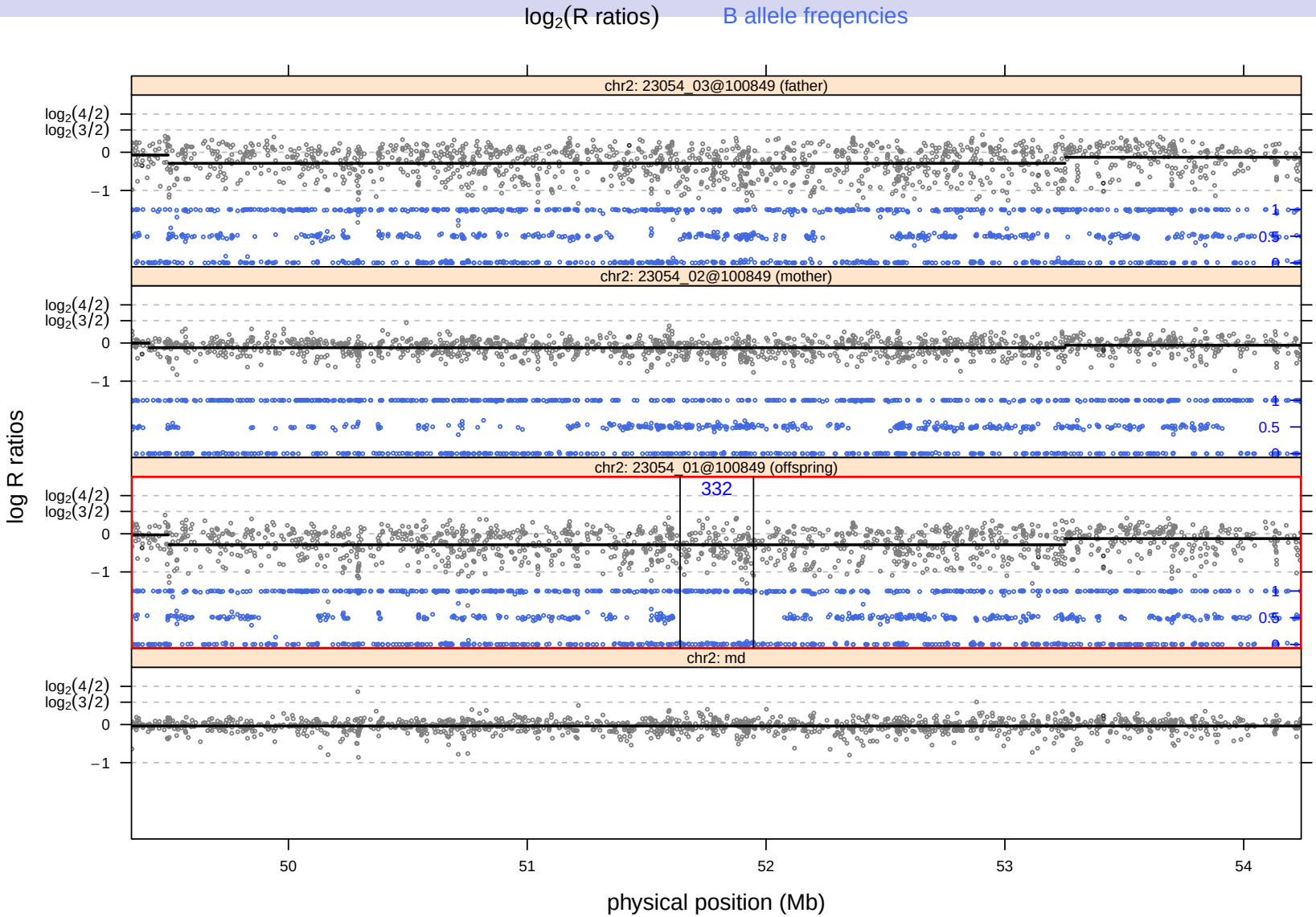
$$\text{minimum distance} \equiv (r_{\mathrm{O}} - r_{\mathrm{F}}) \times \mathsf{I}_{[|r_{\mathrm{O}} - r_{\mathrm{F}}| > |r_{\mathrm{O}} - r_{\mathrm{M}}|]} + (r_{\mathrm{O}} - r_{\mathrm{M}}) \times \mathsf{I}_{[|r_{\mathrm{O}} - r_{\mathrm{F}}| \leq |r_{\mathrm{O}} - r_{\mathrm{M}}|]}$$



False positives



False positives



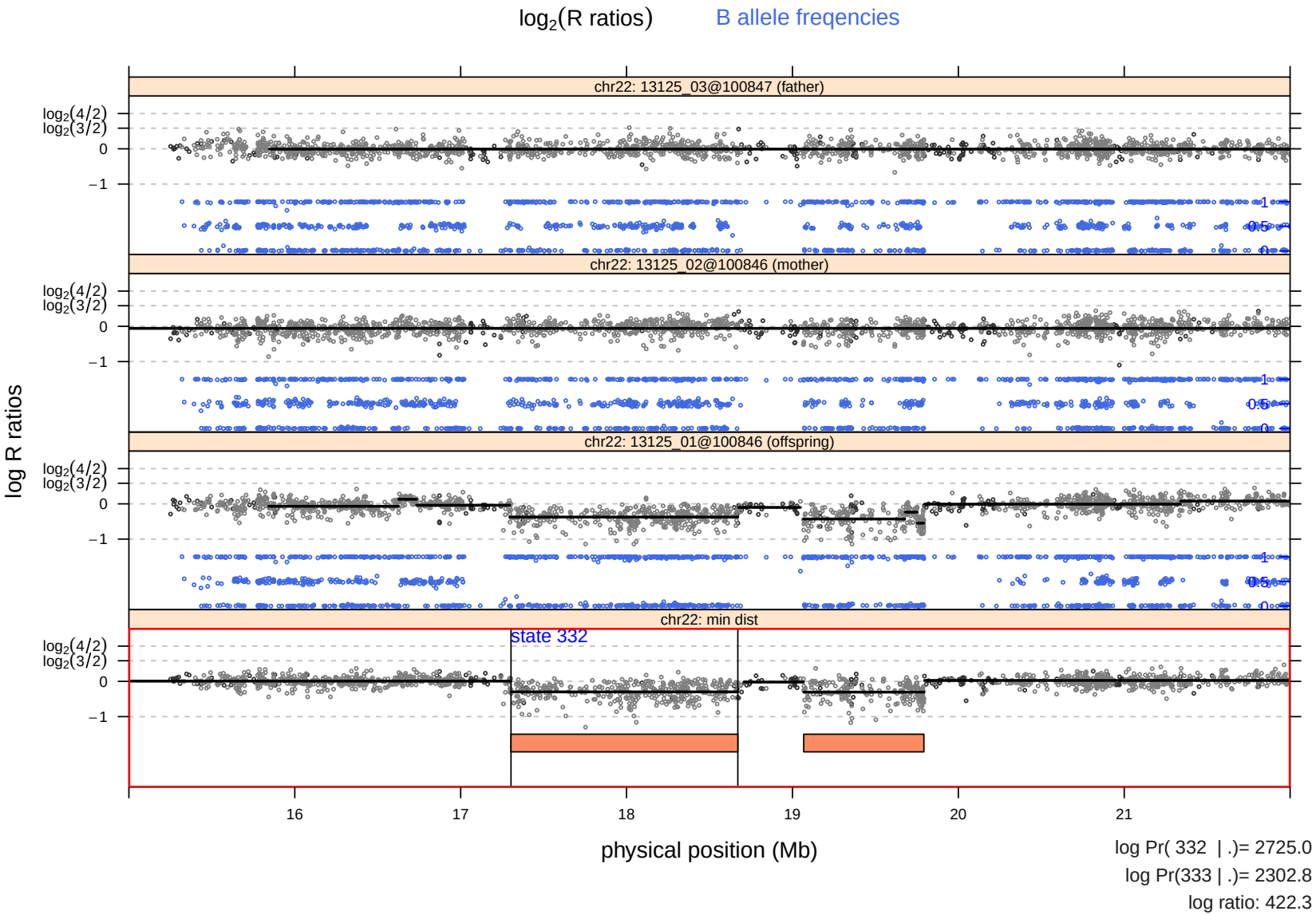
Algorithm

1. Calculate the minimum distance
2. Segment the minimum distance
3. Posterior classification

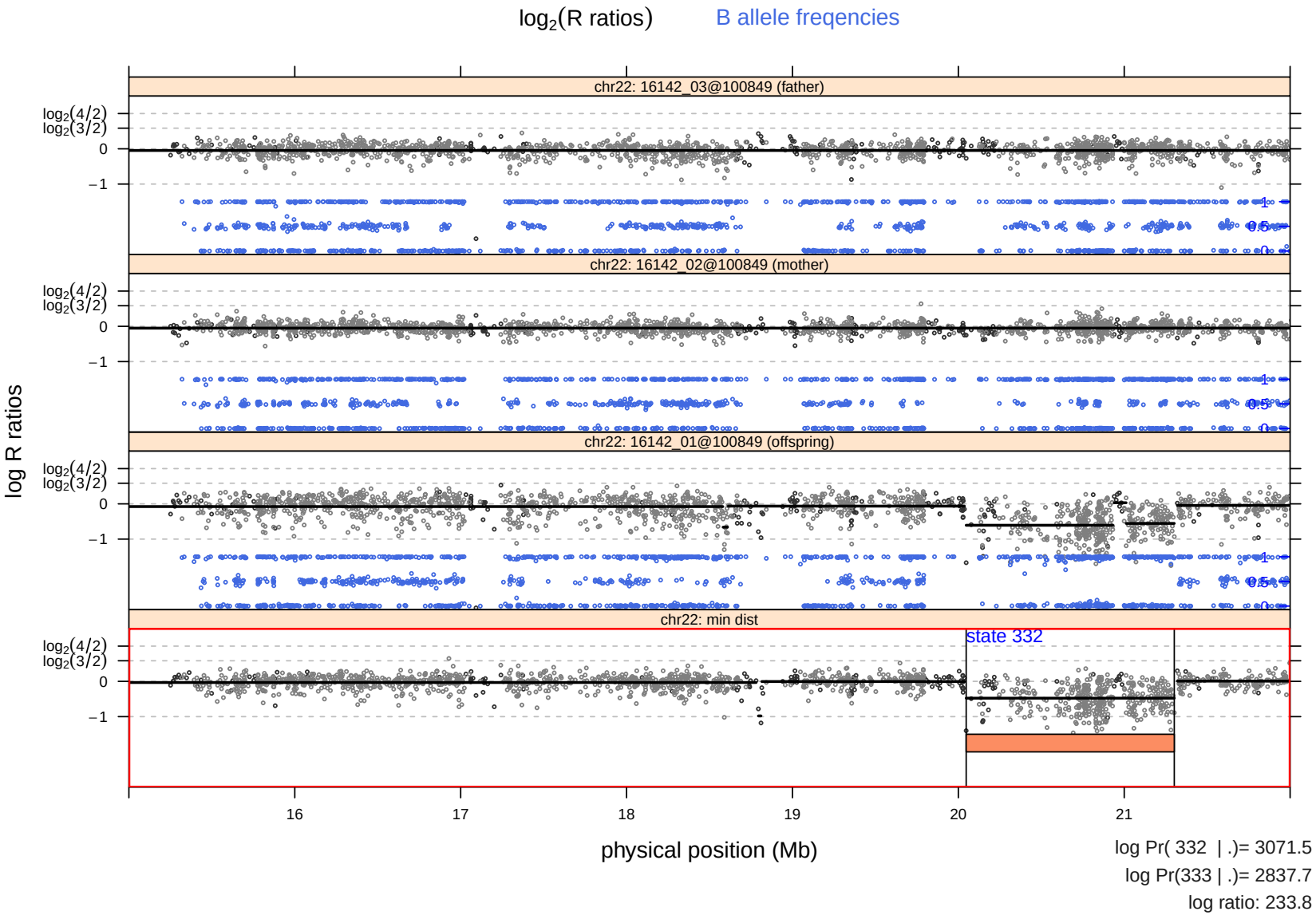
Computational characteristics

- 591 CPU hours (versus 6,246)
- When multiple CPUs are detected, the MinimumDistance R package parallelizes by chromosome
  - system time $\approx$ 1 day

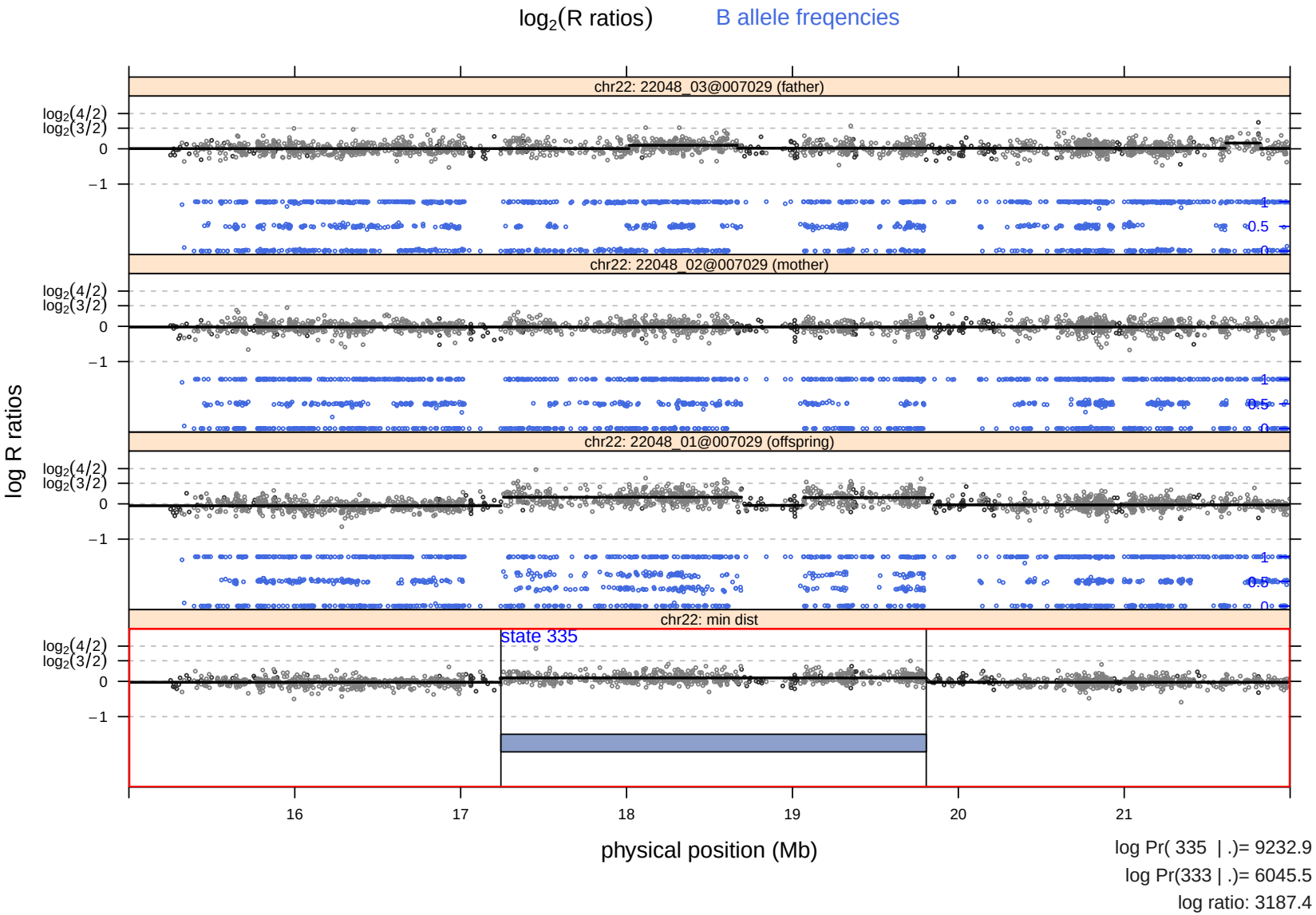
Denovo CNV discovery from arrays



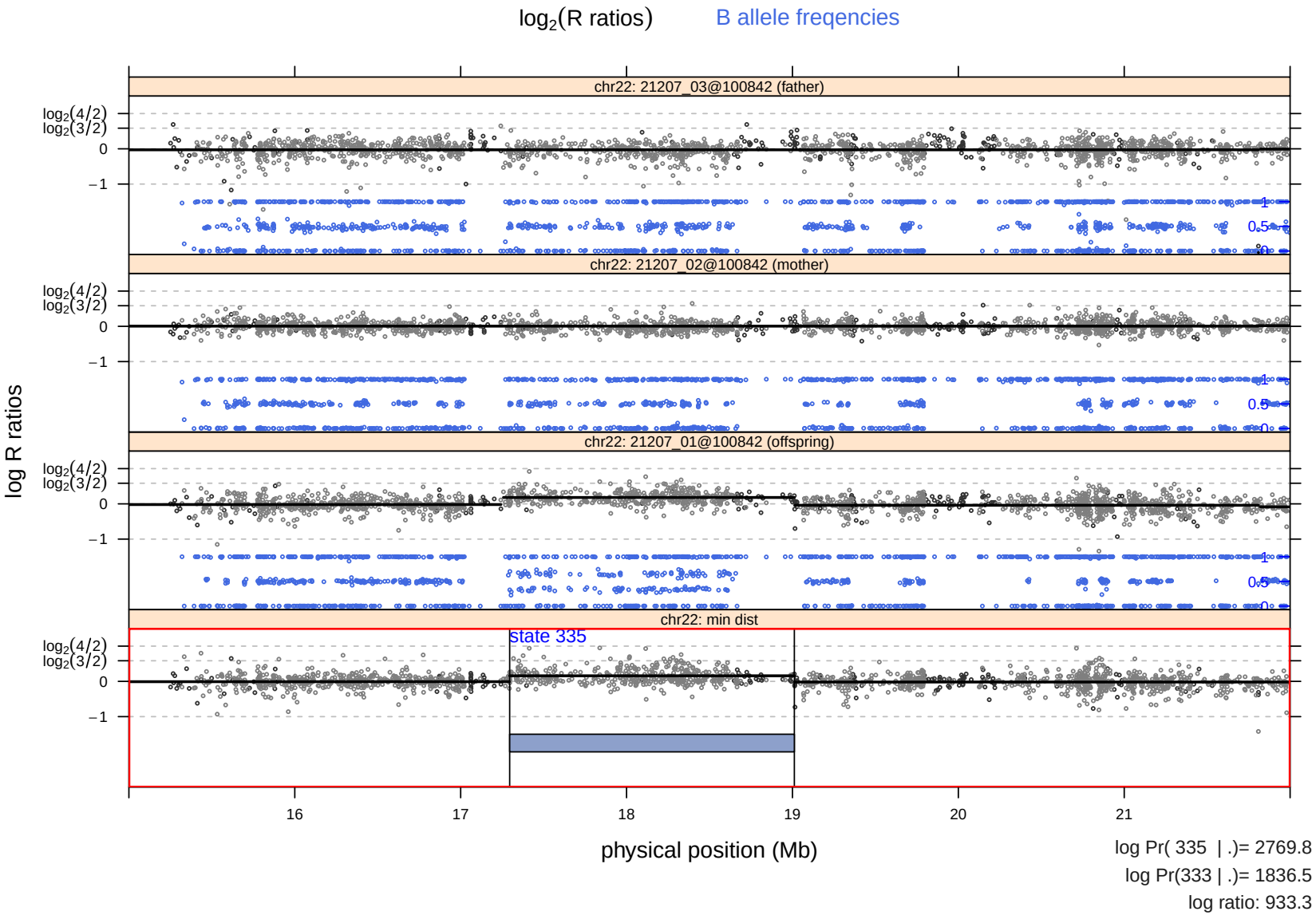
Denovo CNV discovery from arrays



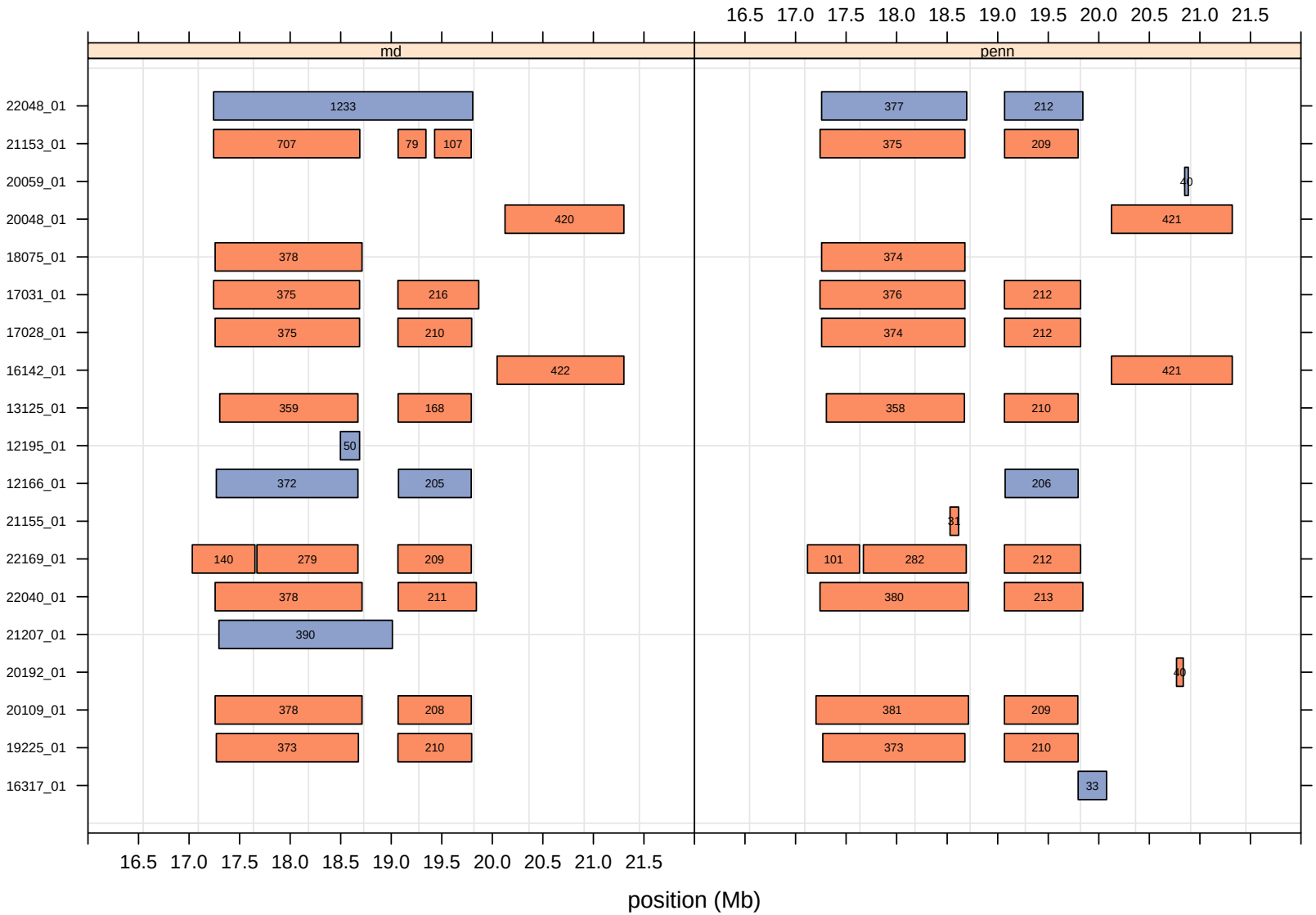
Denovo CNV discovery from arrays



Denovo CNV discovery from arrays



Denovo CNV in the DiGeorge region

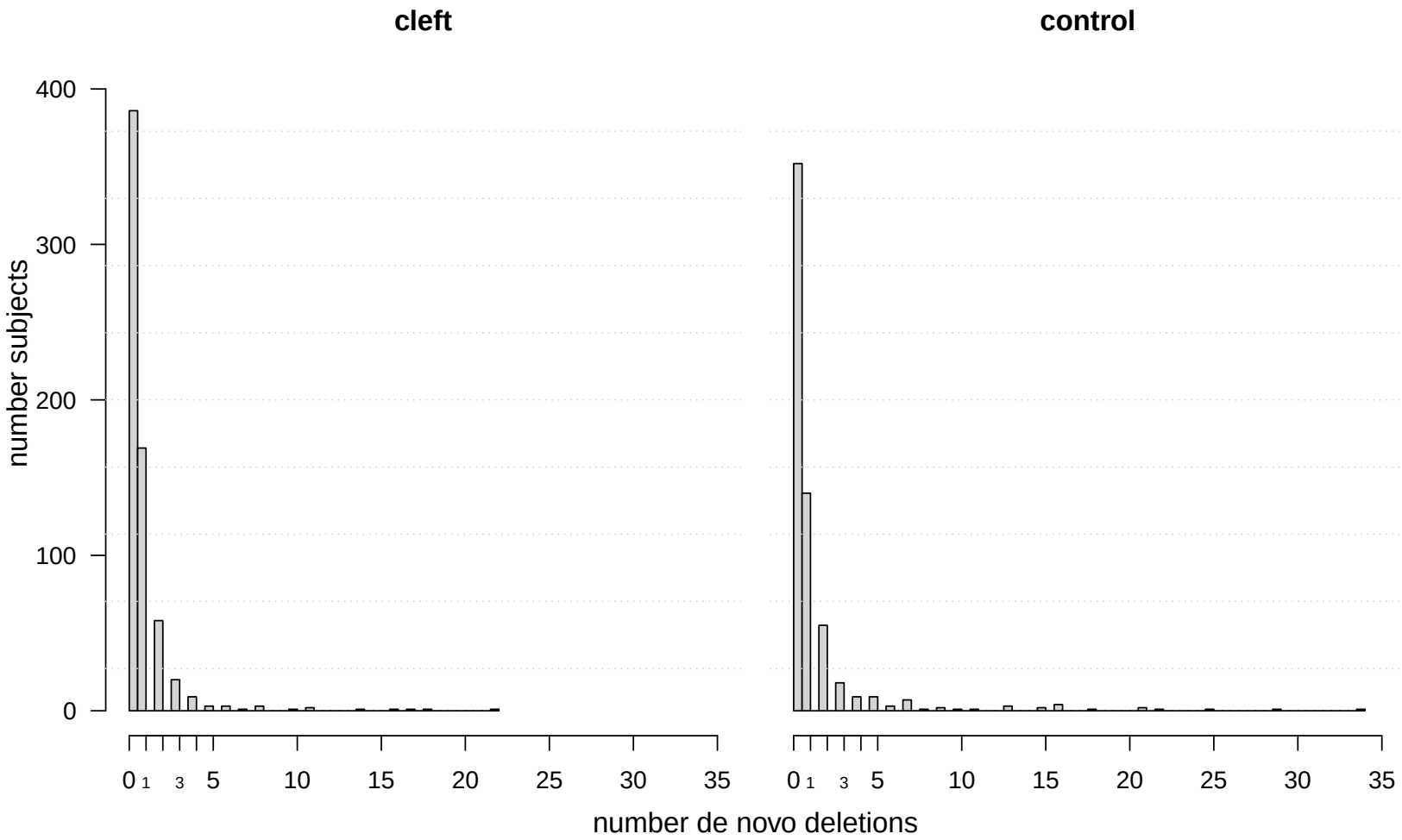


Sample size

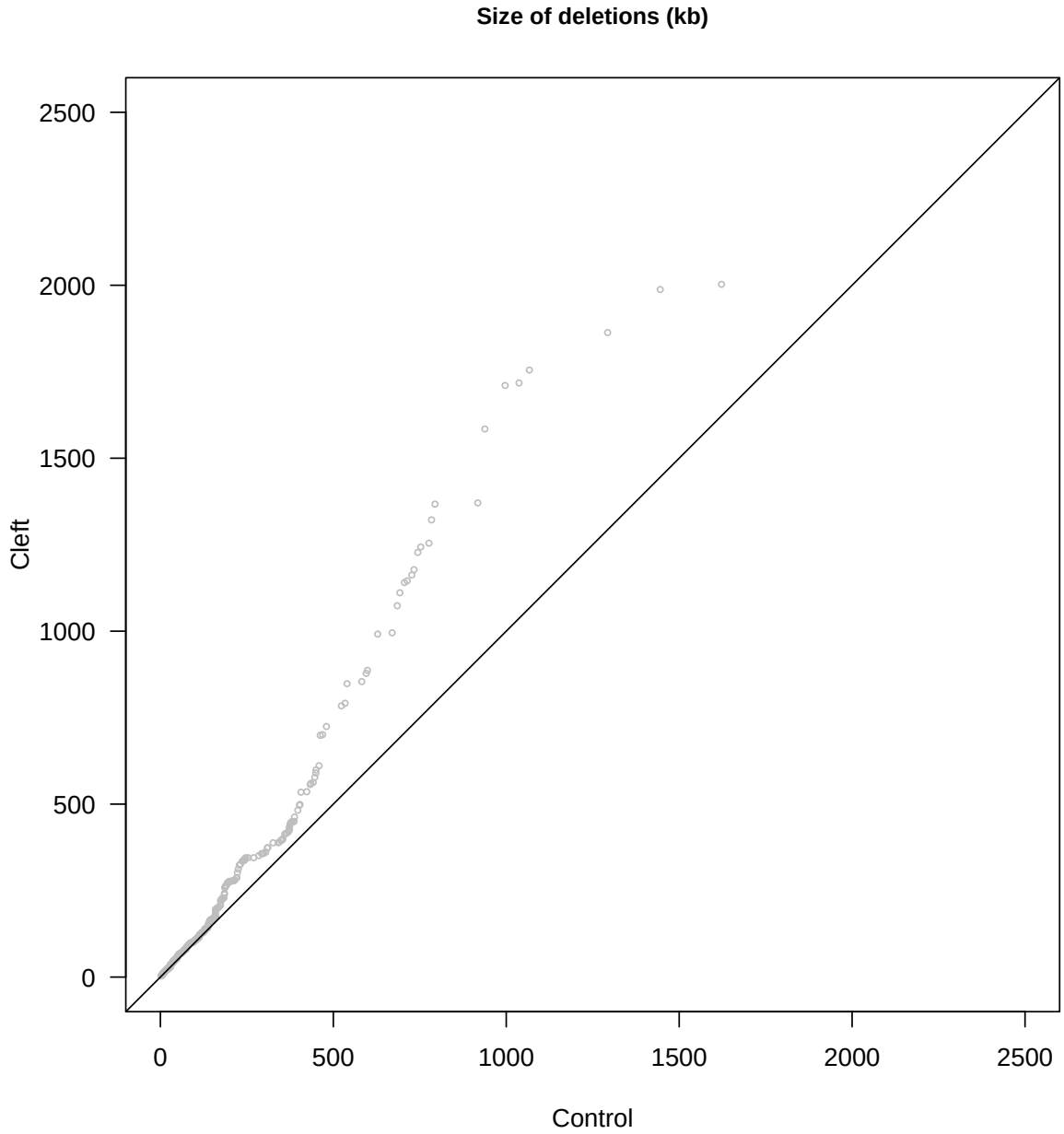
	Cleft	Controls
Discovery	2082	752
WGA	120	99
MAD $>$ 0.3	212	138
non-EA	1090	-
n	660	515

- cell frequencies are the number of Father-Mother-Offspring trios
- The controls are trios with dental caries.

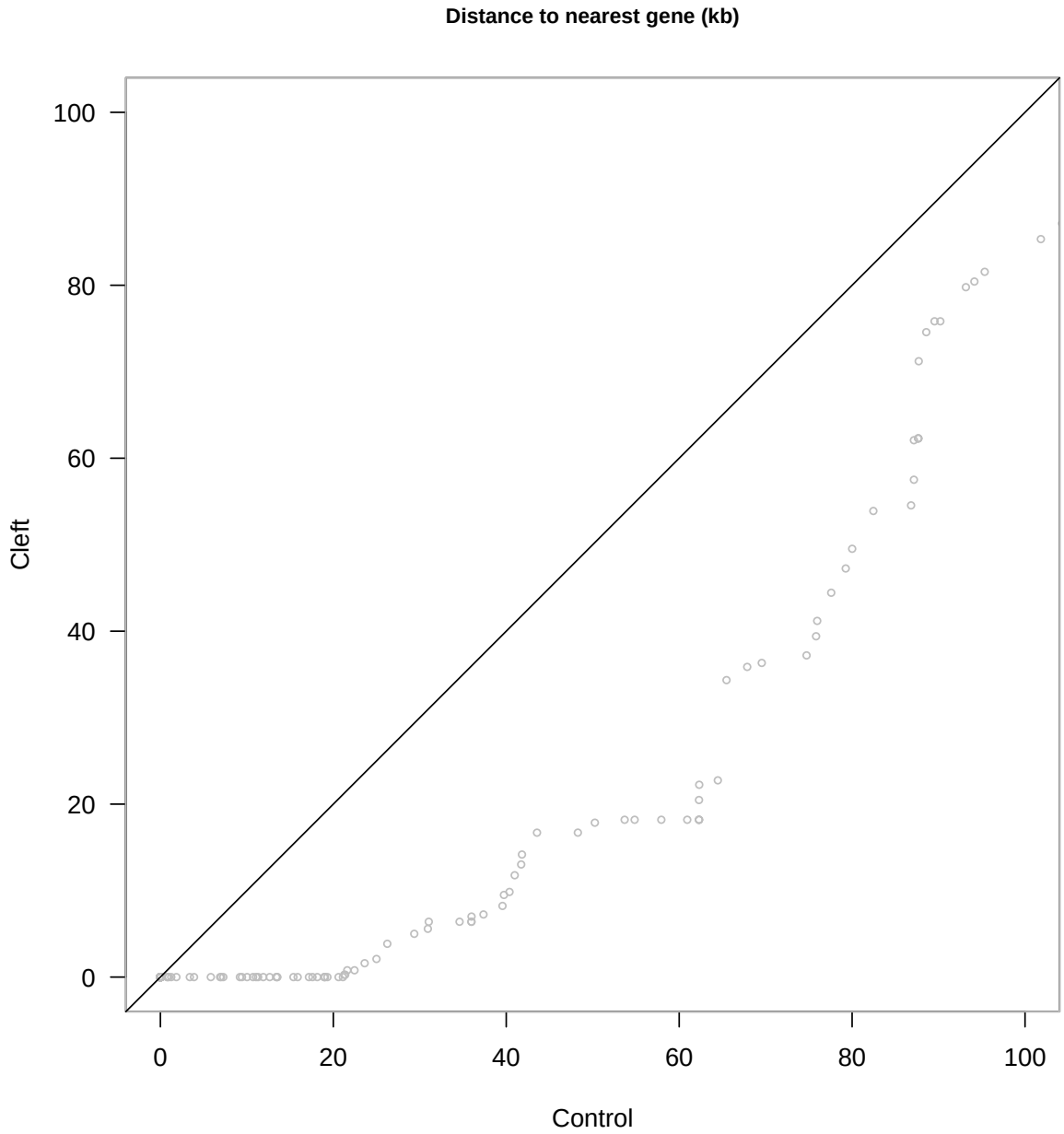
Frequency of denovo deletions



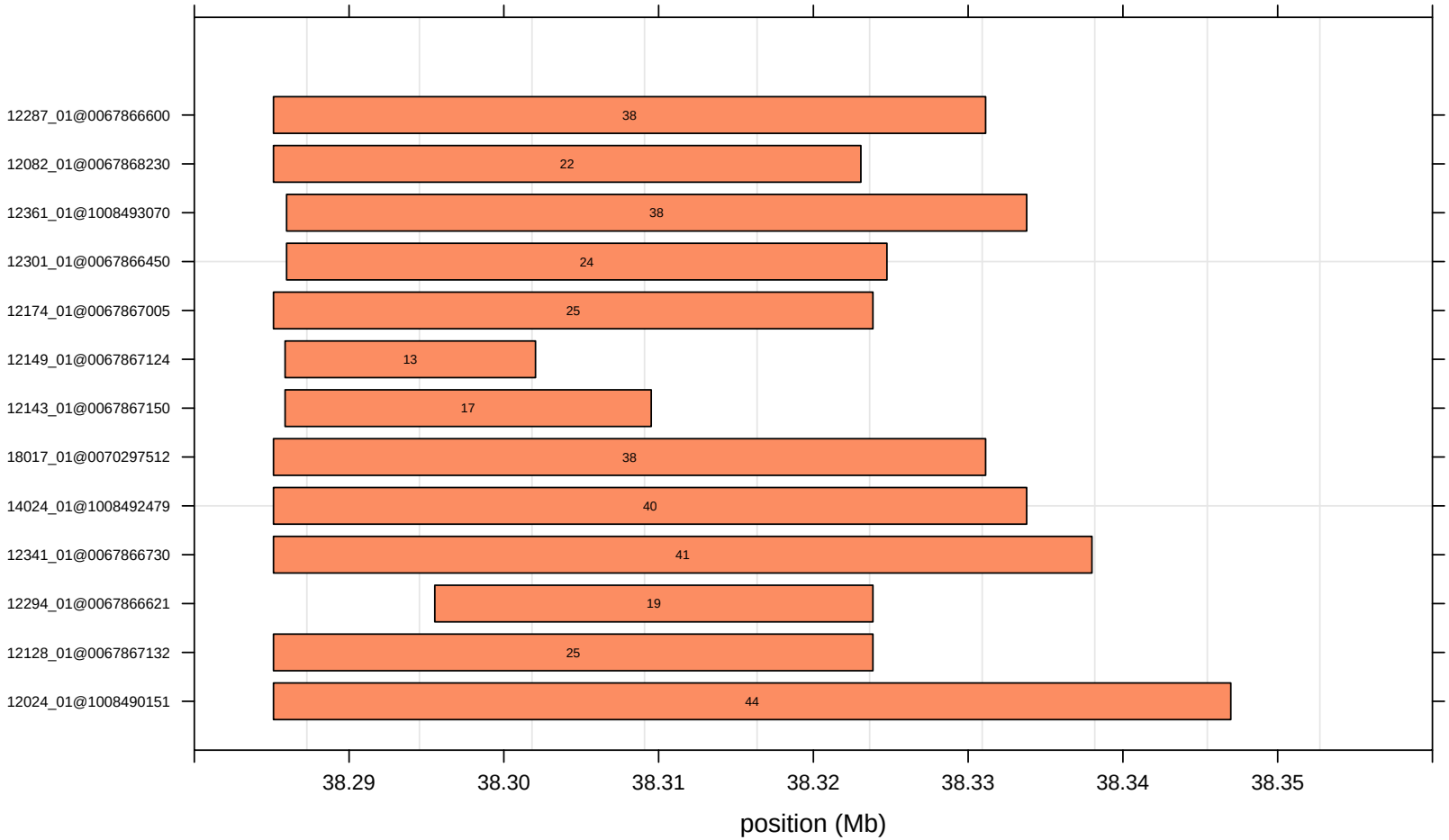
Size of deletions



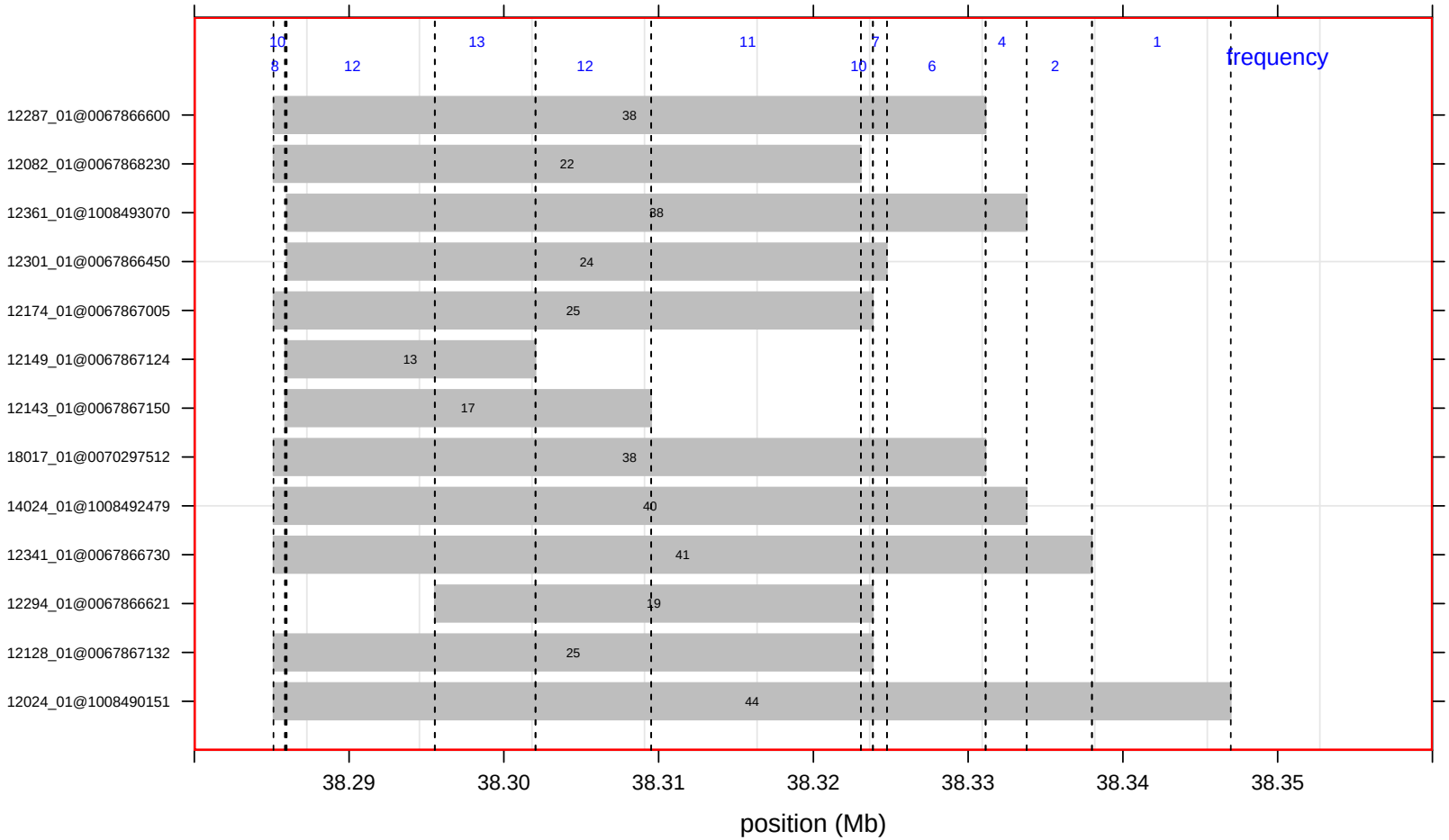
Distance to nearest gene



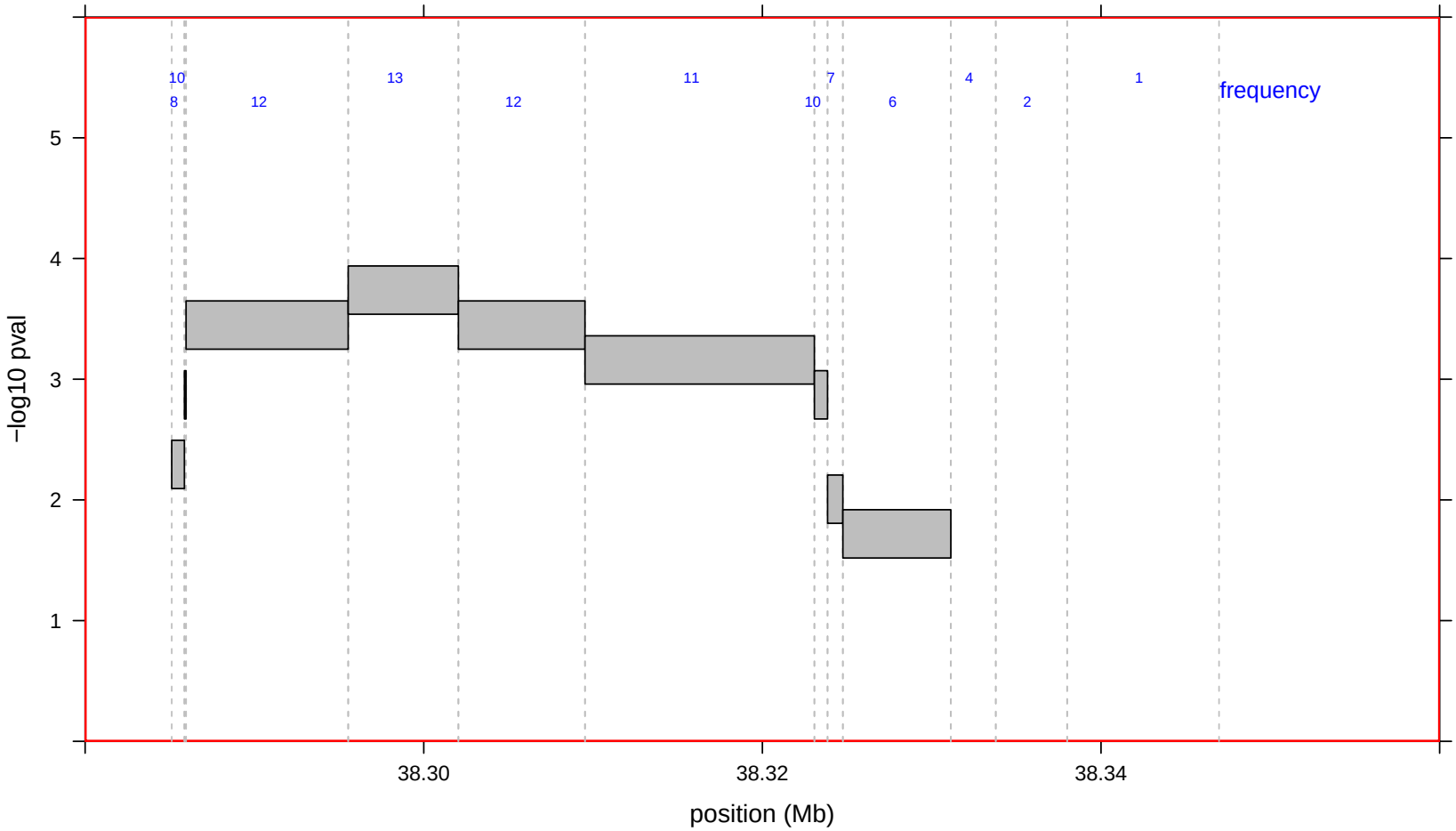
Association of de novo deletions with oral cleft



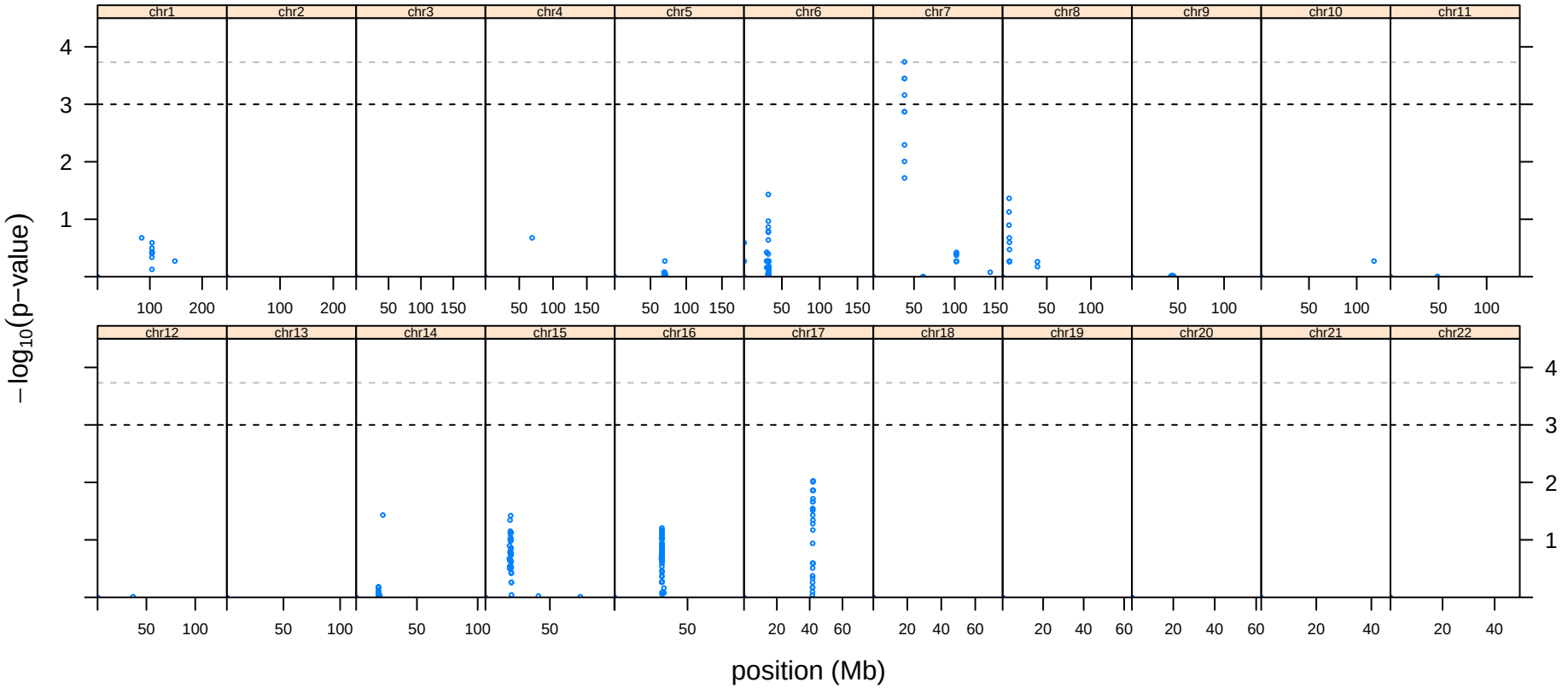
Association of de novo deletions with oral cleft



Association of de novo deletions with oral cleft

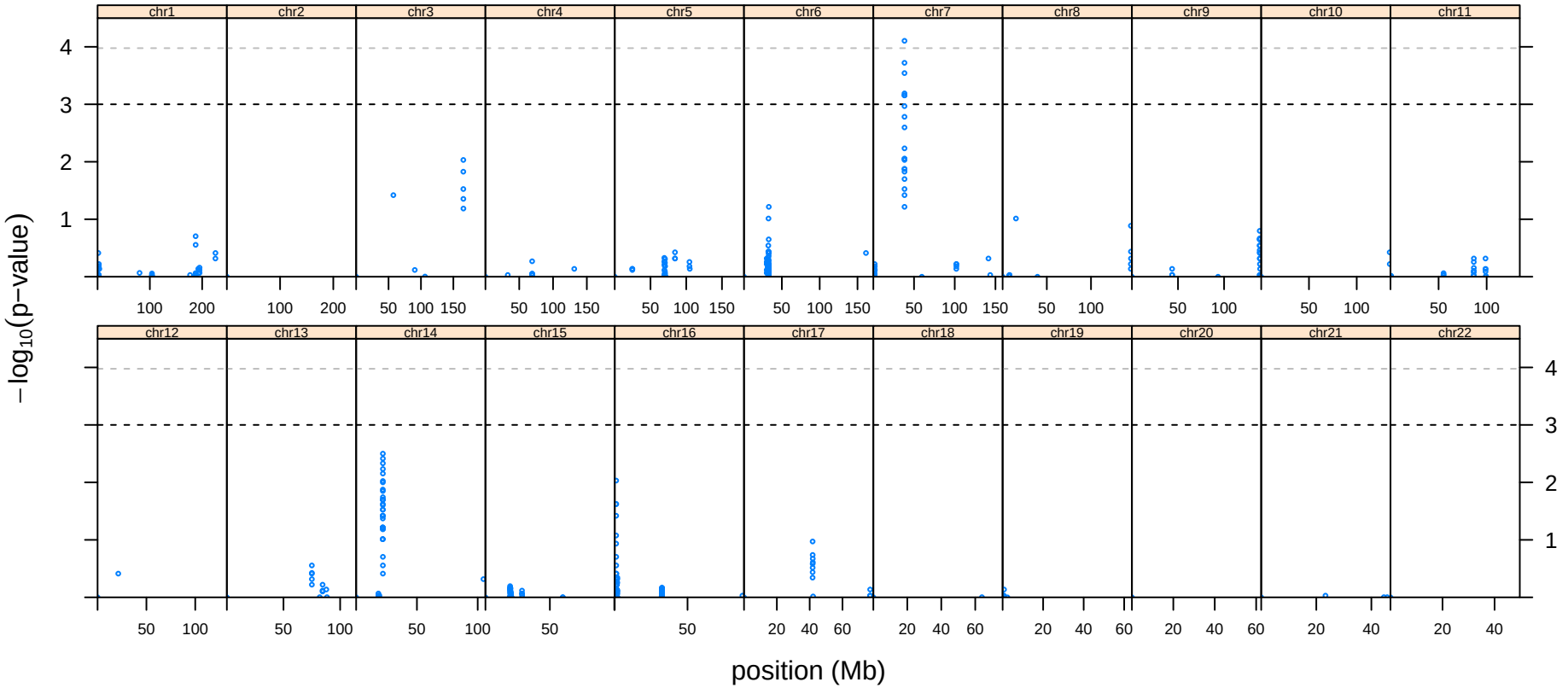


Association of cleft with denovo deletions



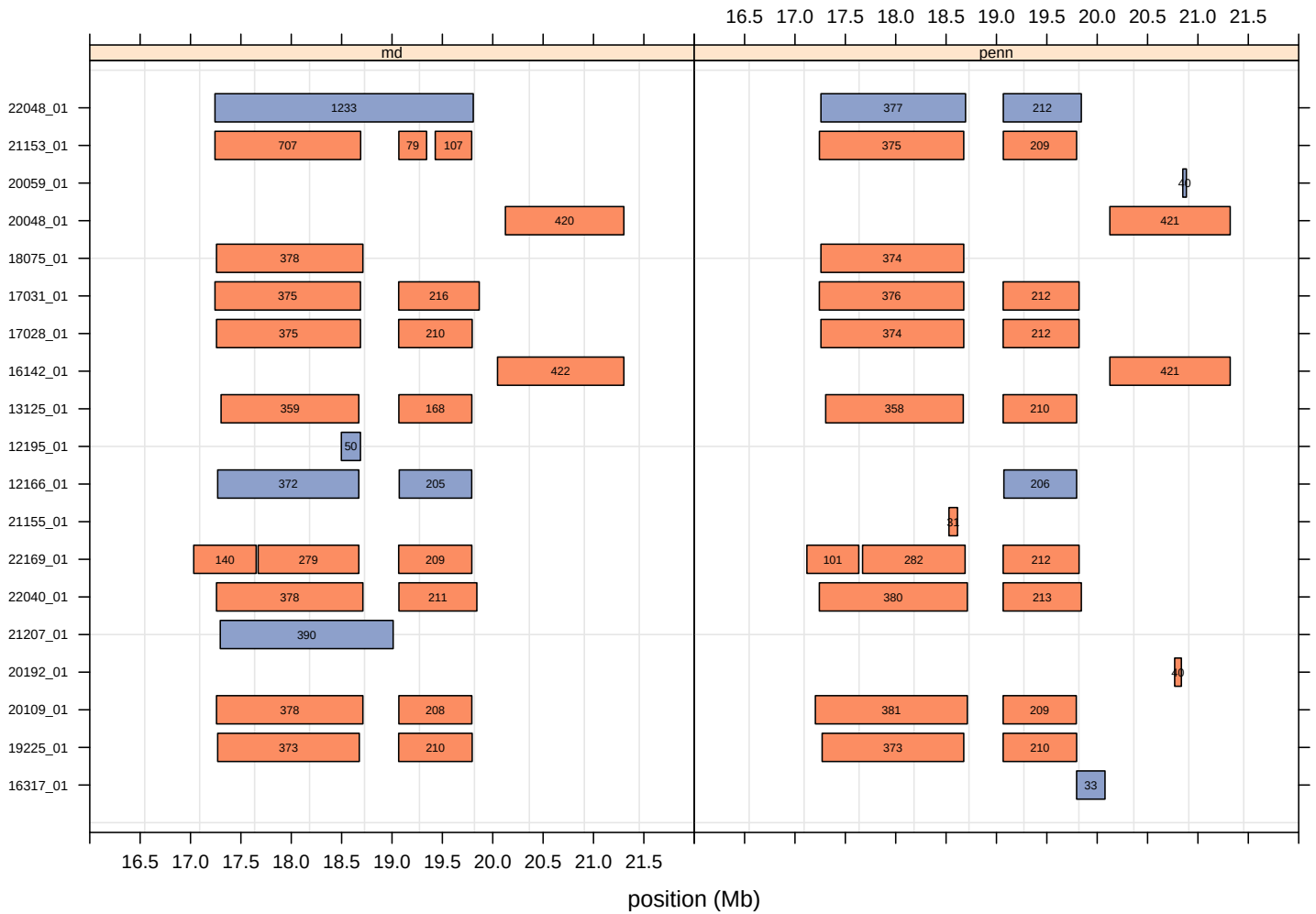
MinimumDistance

Association of cleft with denovo deletions

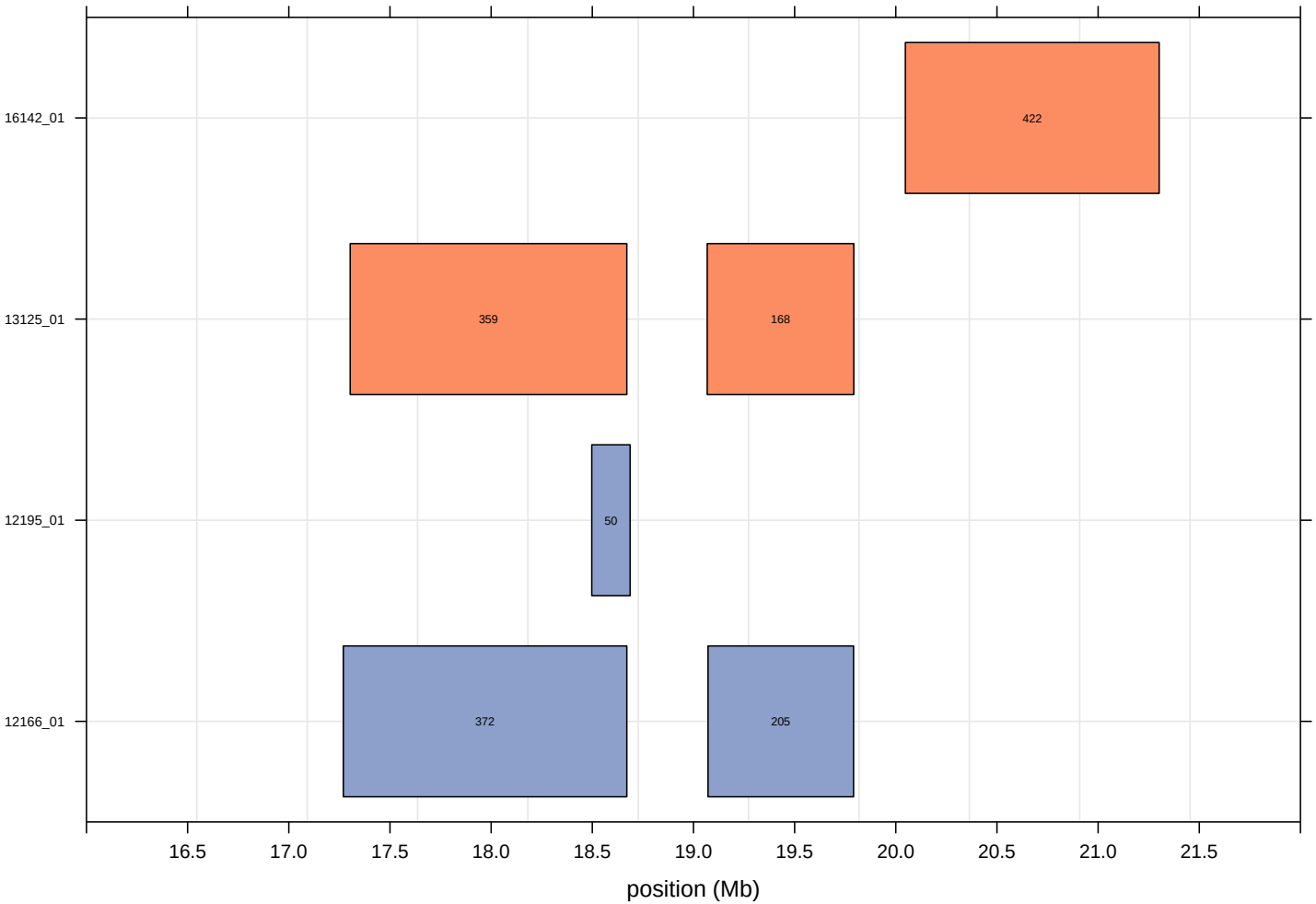


PennCNV

What about the DiGeorge region?

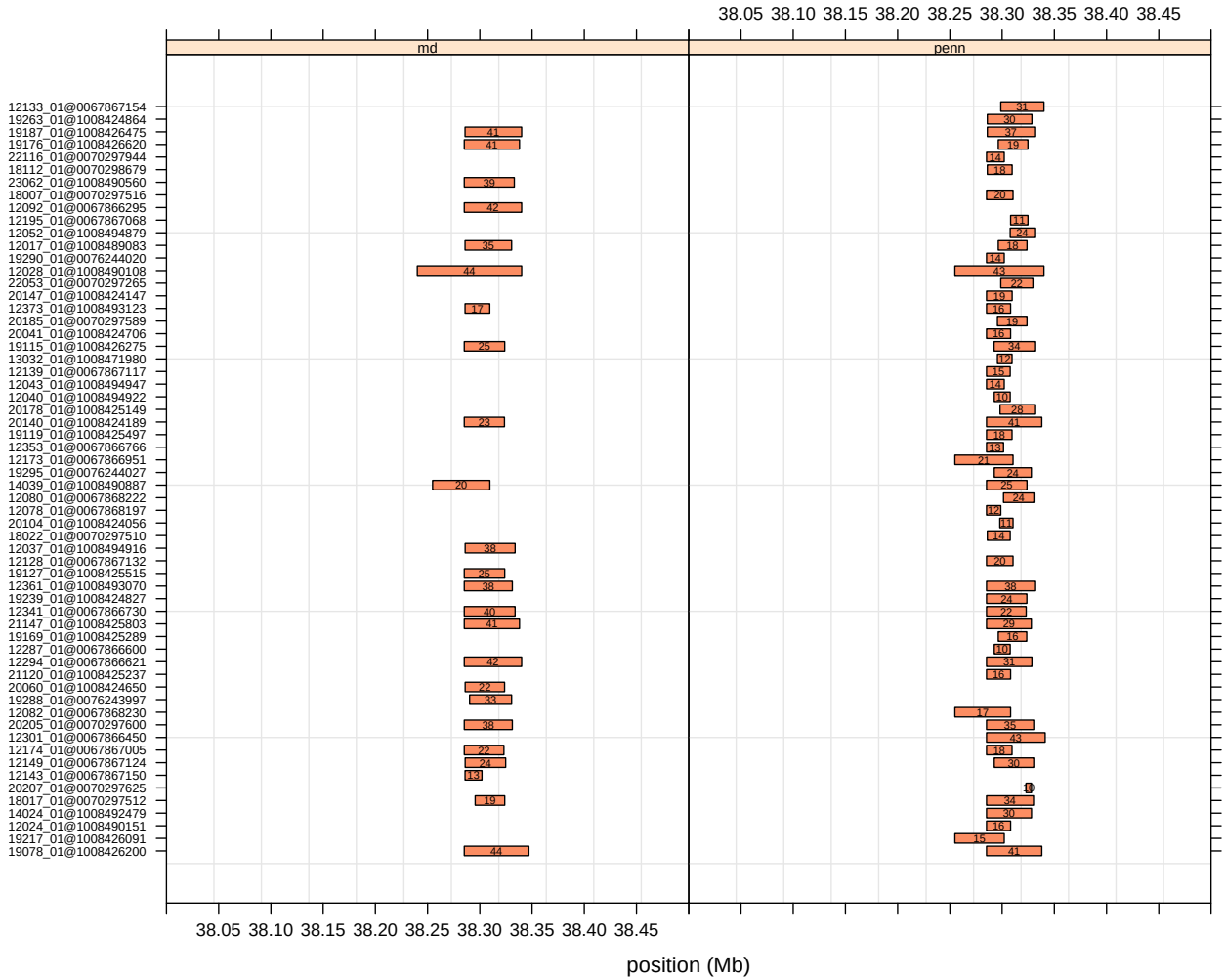


What about the DiGeorge region?

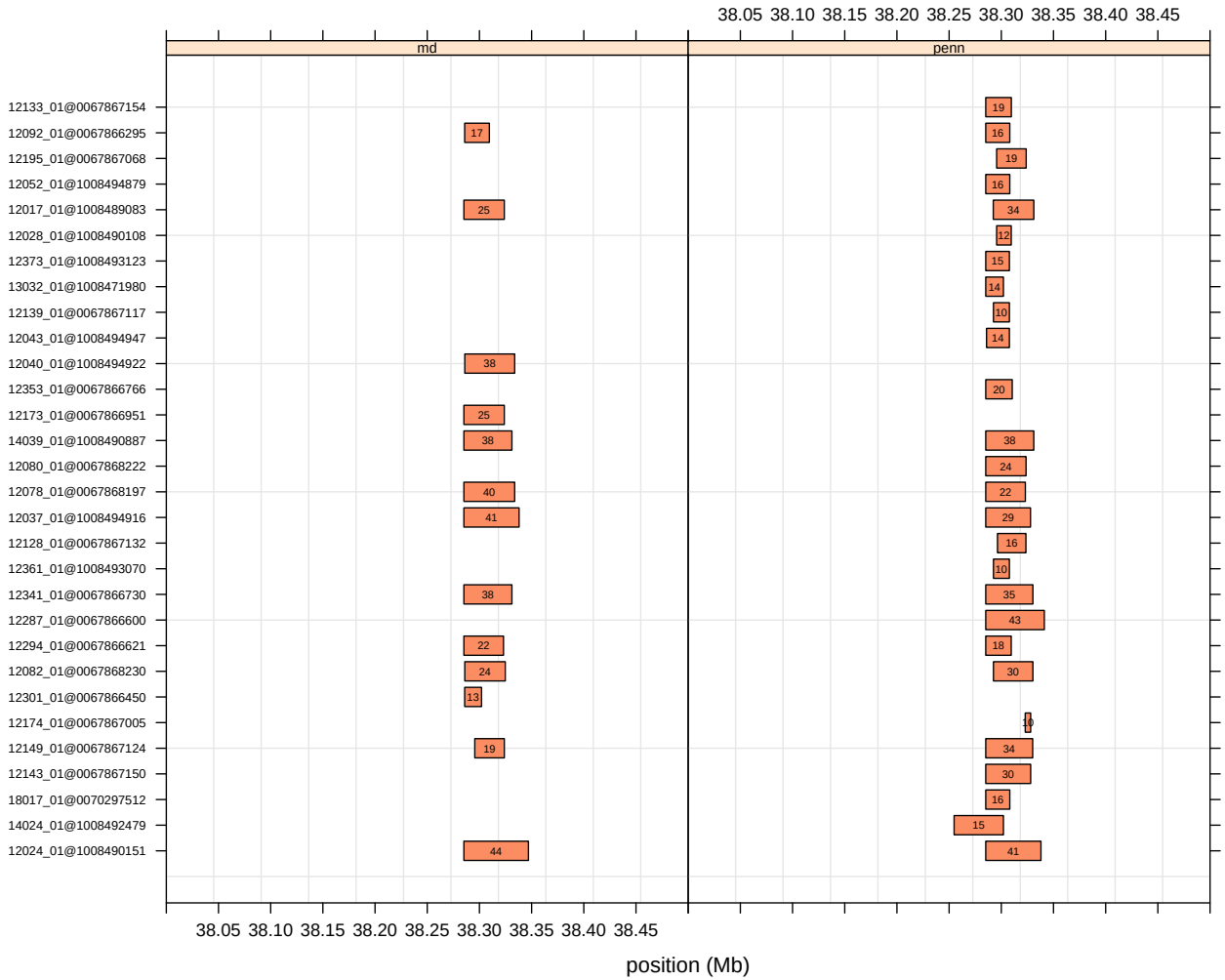


Only 2 de novo deletions among the offspring with European ancestry

Chromosome 7

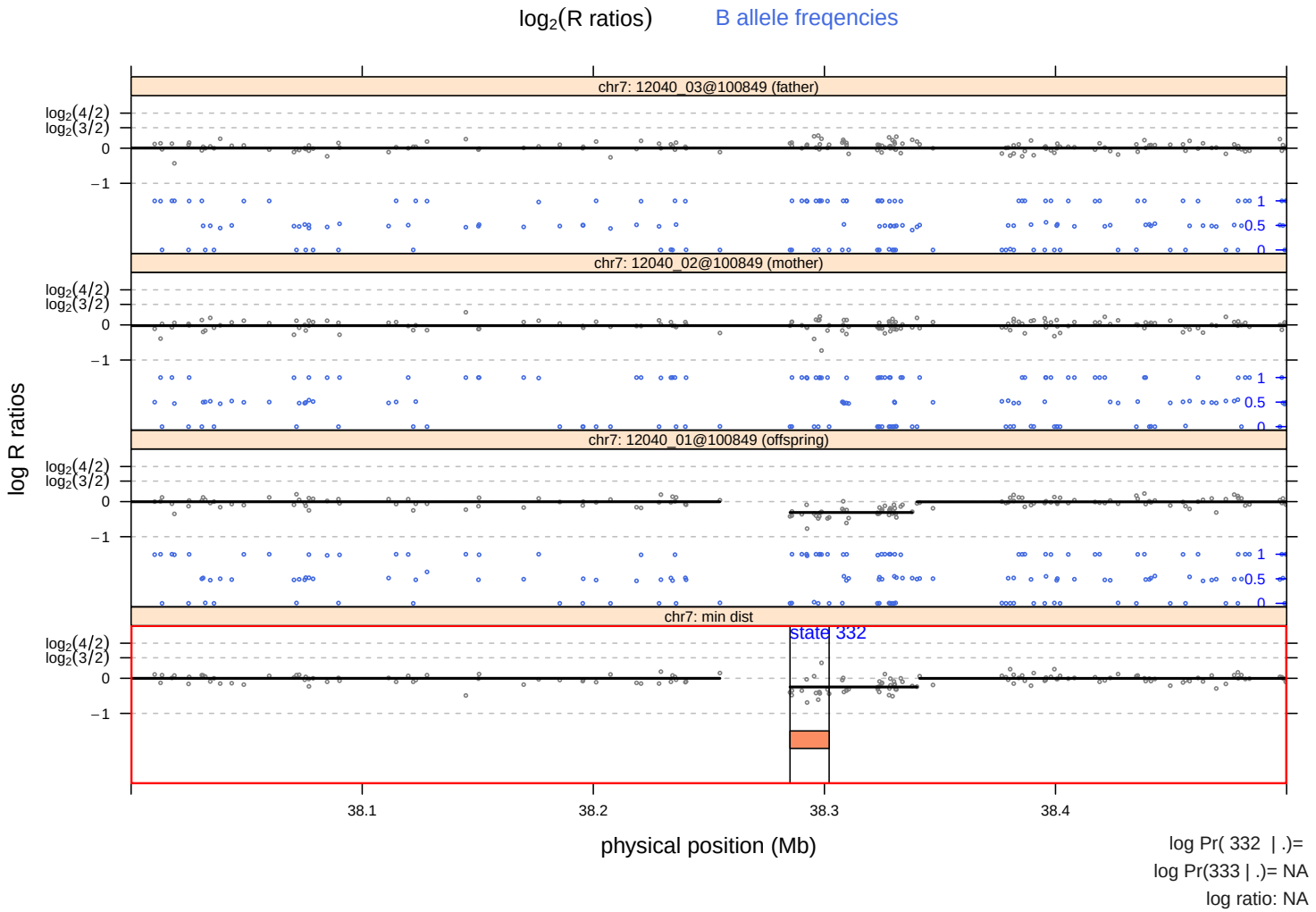


Chromosome 7



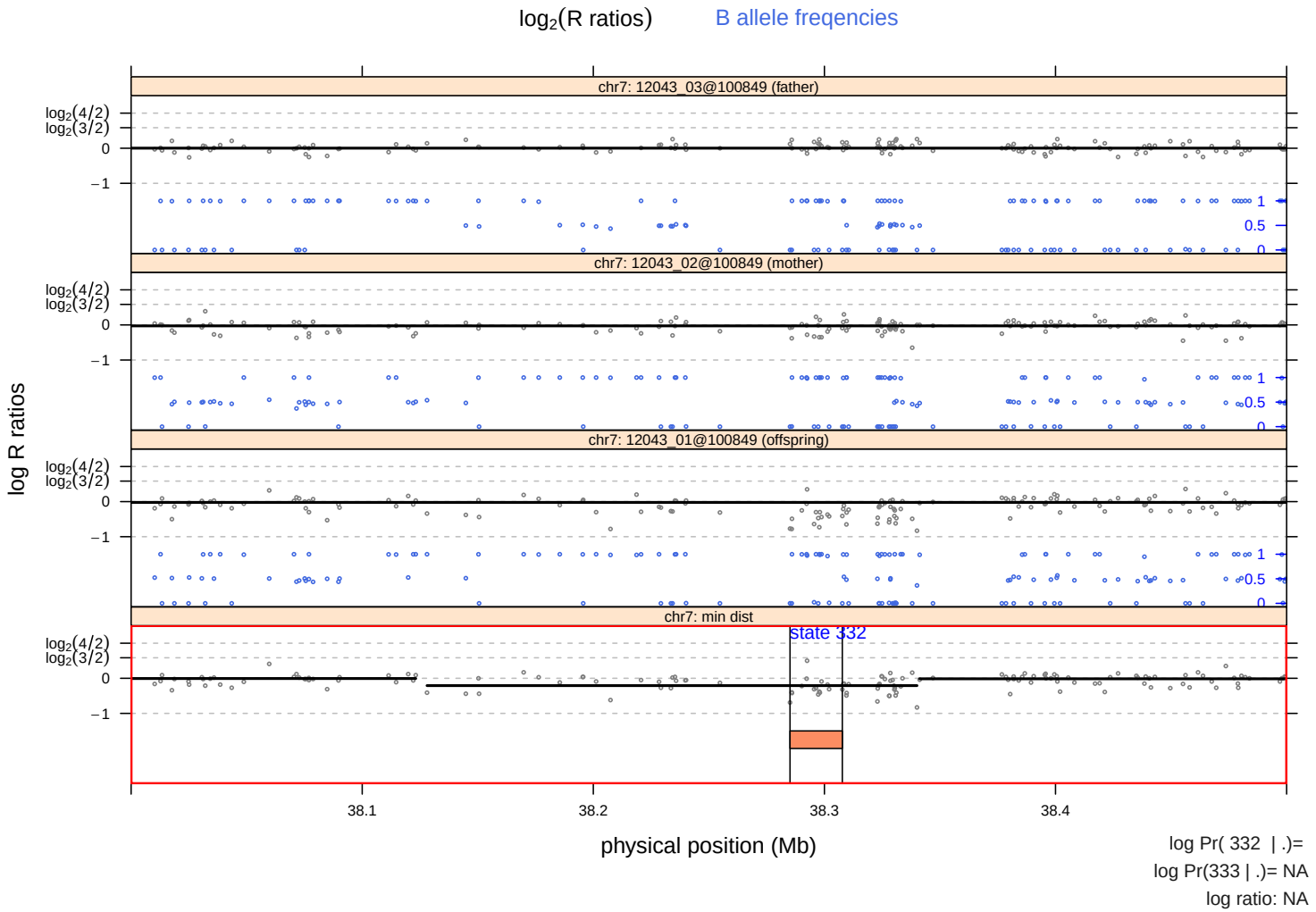
European ancestry

Chromosome 7



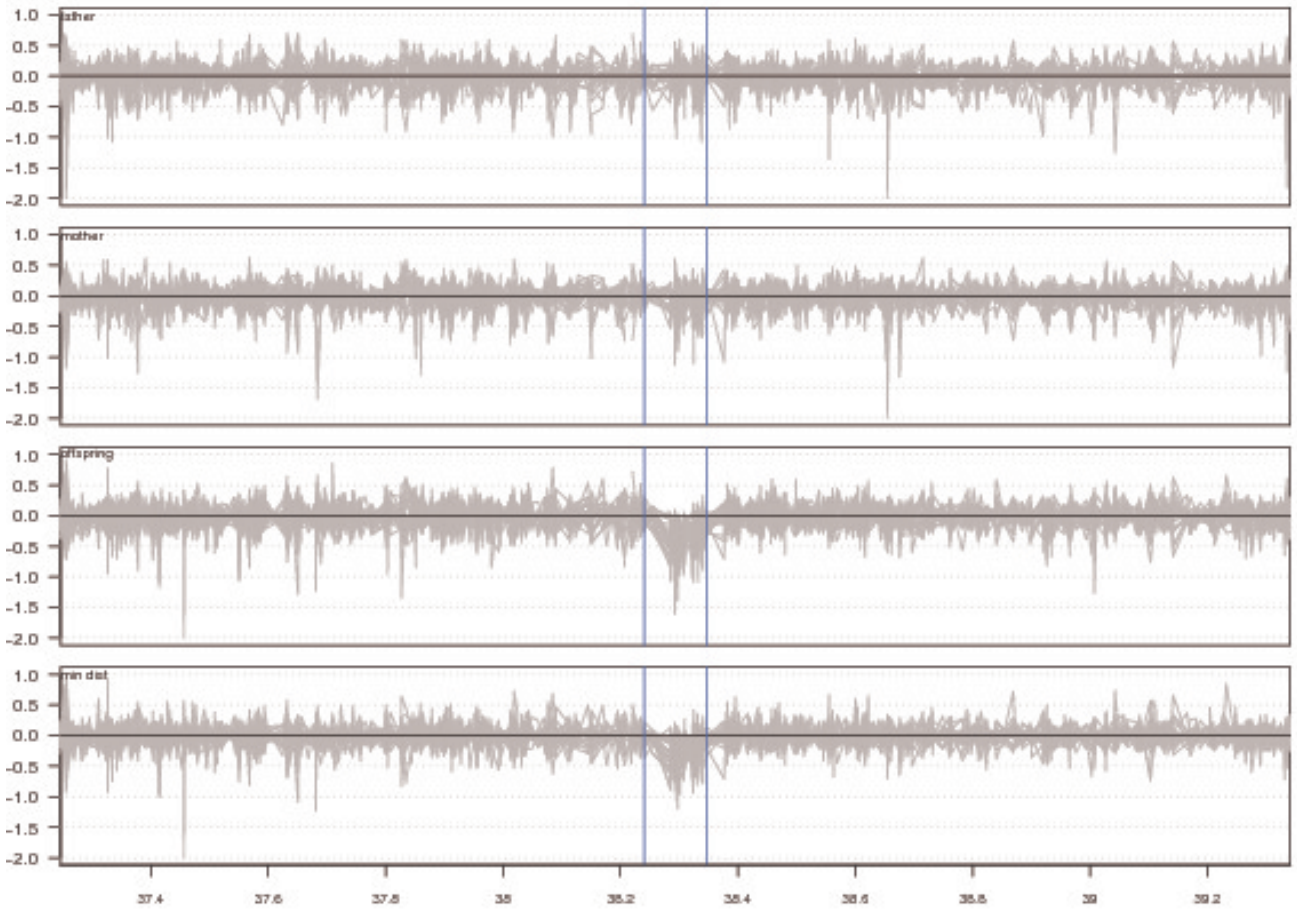
PennCNV only

Chromosome 7

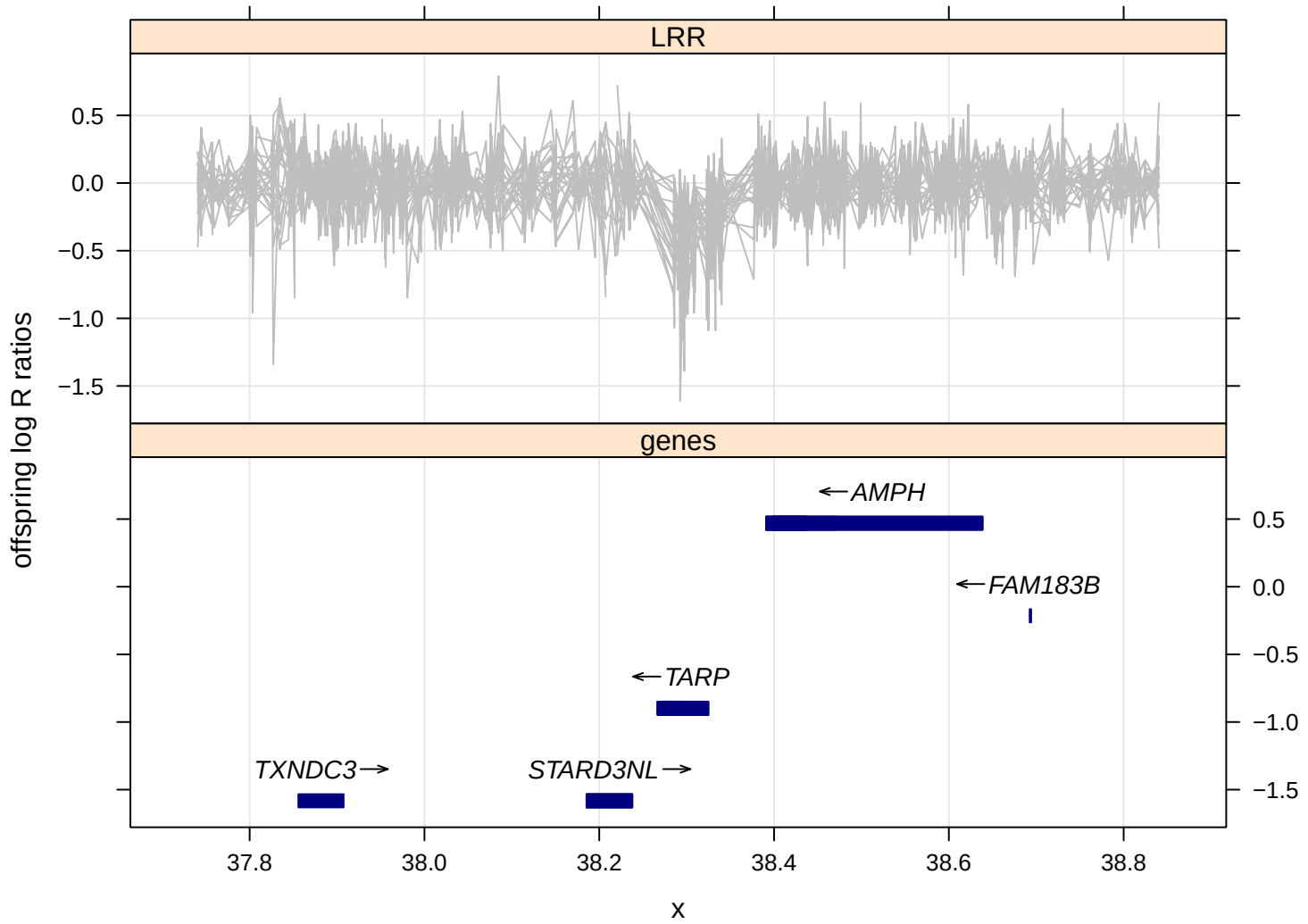


PennCNV only

Chromosome 7



Chromosome 7



Conclusions

- False positives are problematic using standard approaches for de novo CNV discovery
- Oral cleft offspring tend to have slightly bigger denovo deletions than in the controls, and the deletions tended to be closer to coding regions of the genome
- The association of the chromosome 7 region was statistically significant after adjusting for multiple testing.
- The association study used a small fraction of the available oral cleft data.

Acknowledgements

- Samuel Younkin
- Ingo Ruczinski
- Terri Beaty
- Holger Schwendl
- Alan Scott
- Jackie Hetmanski