Measuring copy number from high-throughput SNP arrays

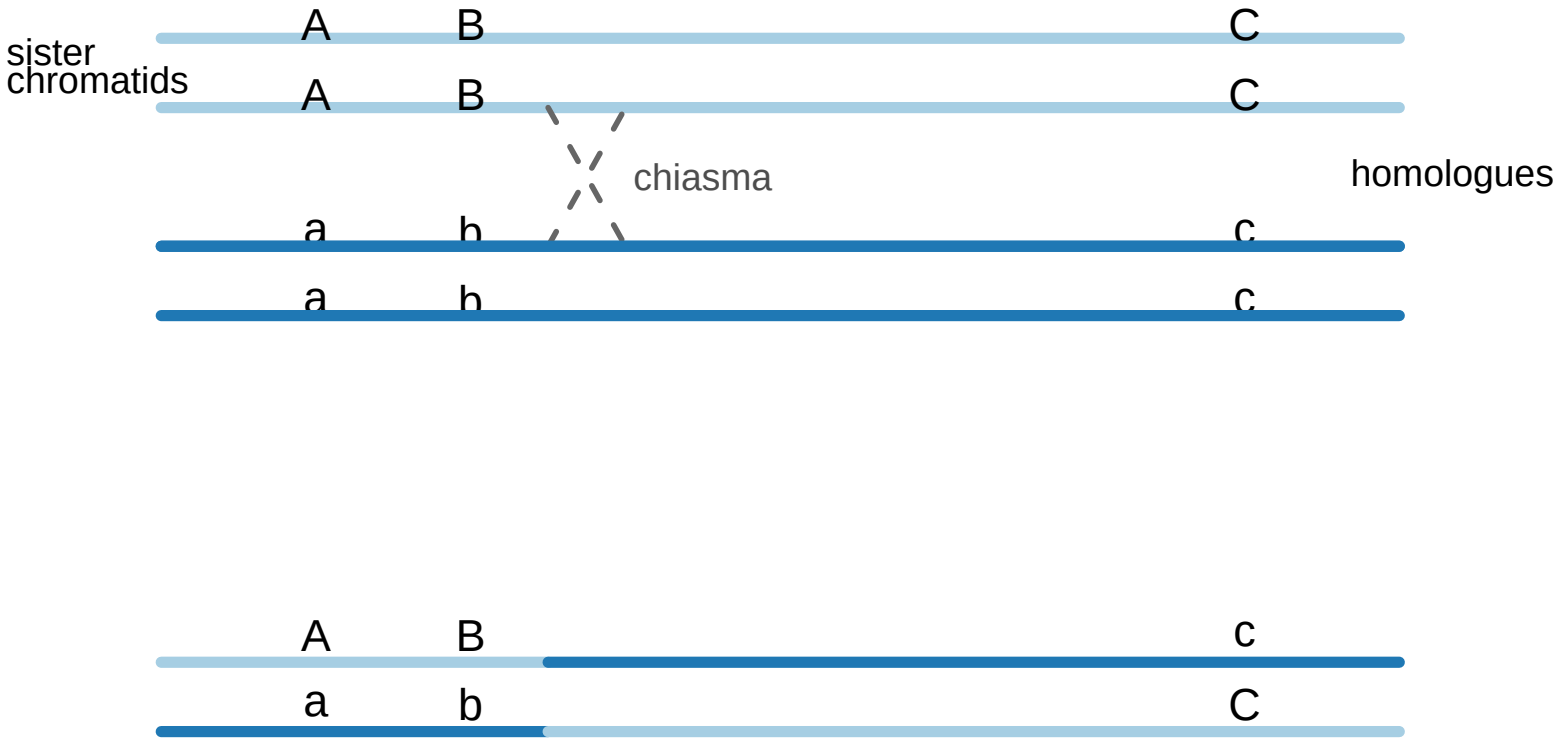
Rob Scharpf

Oncology Biostatistics
Johns Hopkins University

March 24, 2011

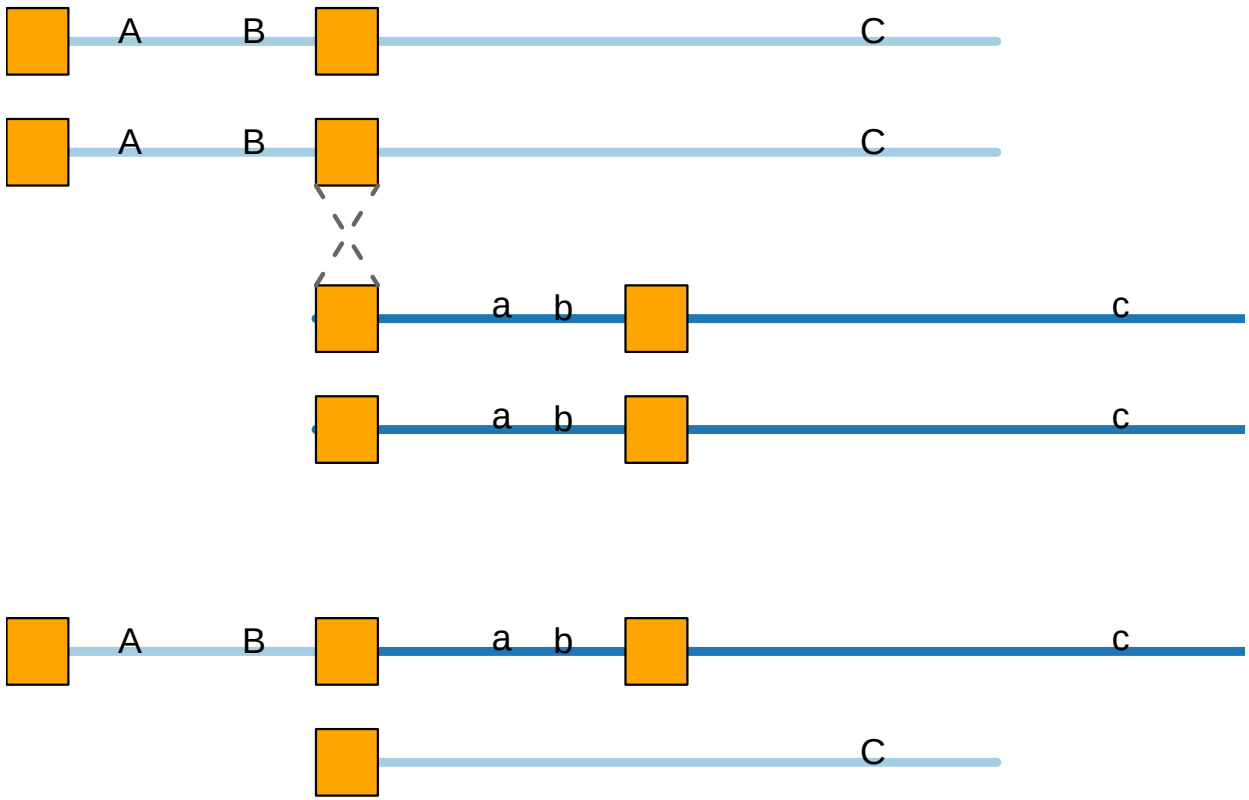
Normal recombination

Meiosis



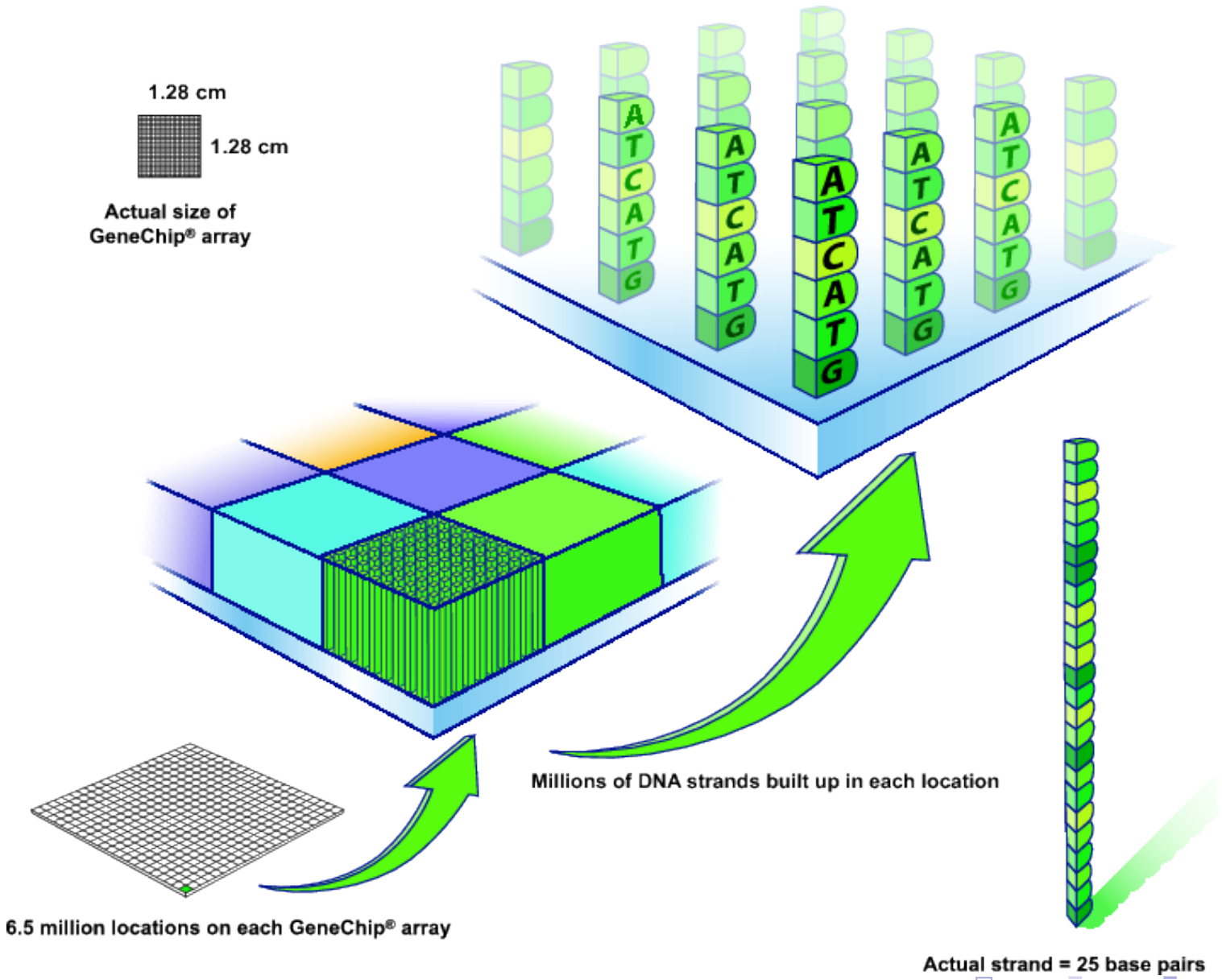
Recombination causing CNV

Meiosis



E. Eichler: `http://www.nature.com/scitable/content/How-crossing-over-can-generate-copy-number-2499179`

Affymetrix SNP chip



Affymetrix SNP chip terminology

SNP

Genomic DNA: TACATAGCCATCGGTA**A**/**G**GTACTCAATGATGATA

PM probe for Allele A: ATCGGTAGCCATTCATGAGTTACTA

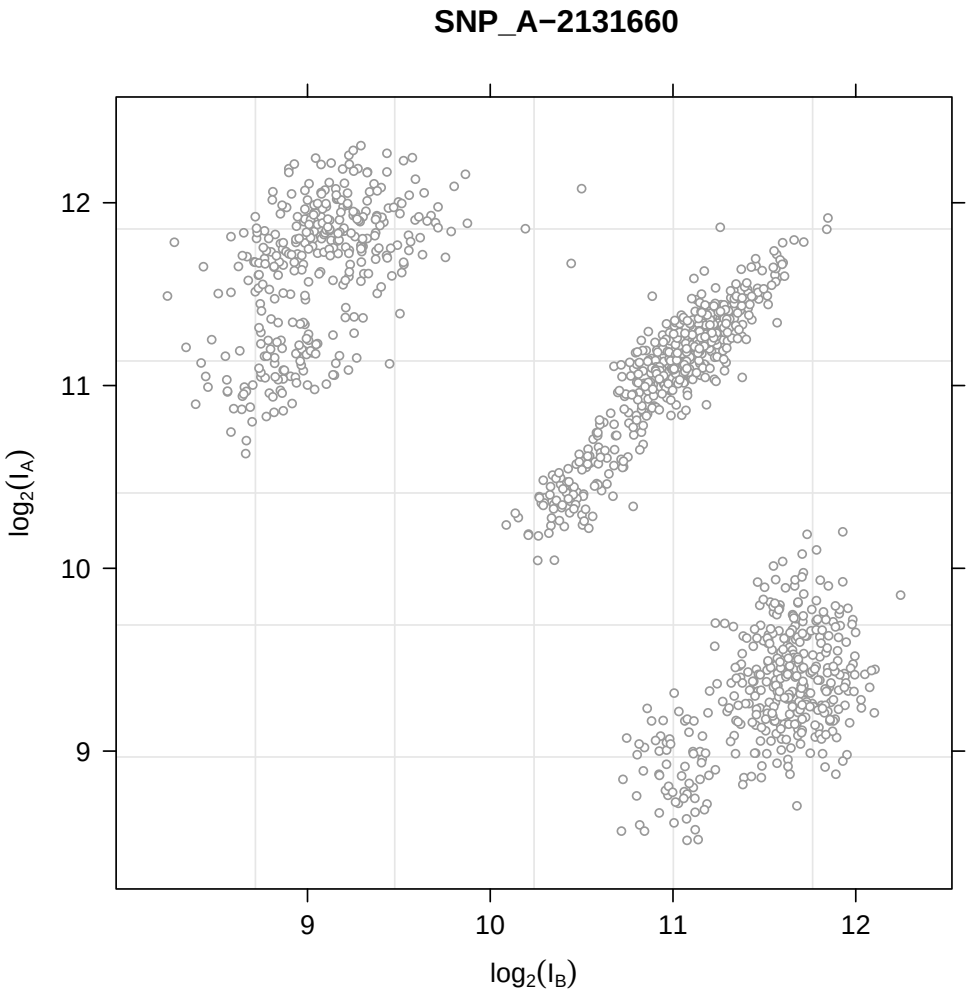
PM probe for Allele B: ATCGGTAGCCATCCATGAGTTACTA

Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:

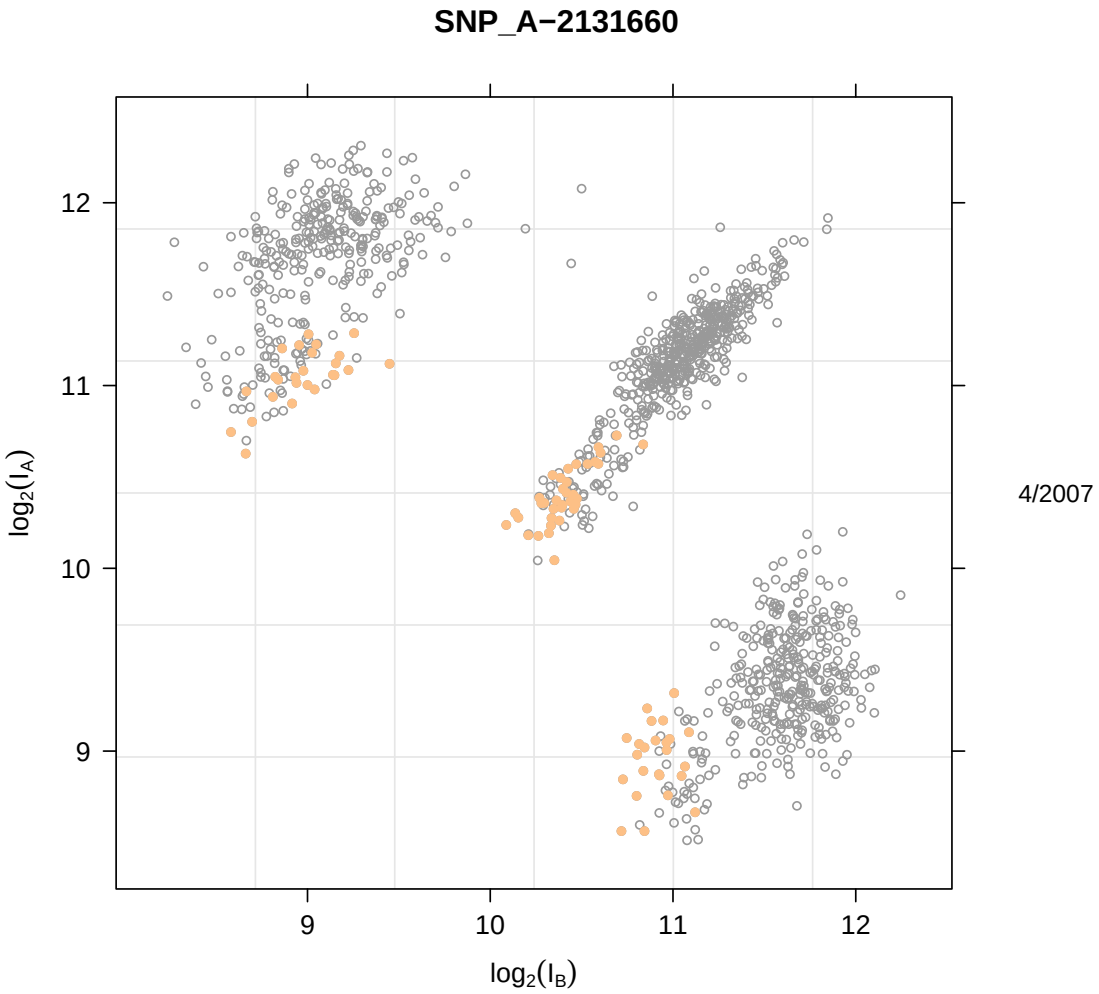
Is a person **AA** , **AG** or **GG** at this Single Nucleotide Polymorphism?

HapMap data for one SNP

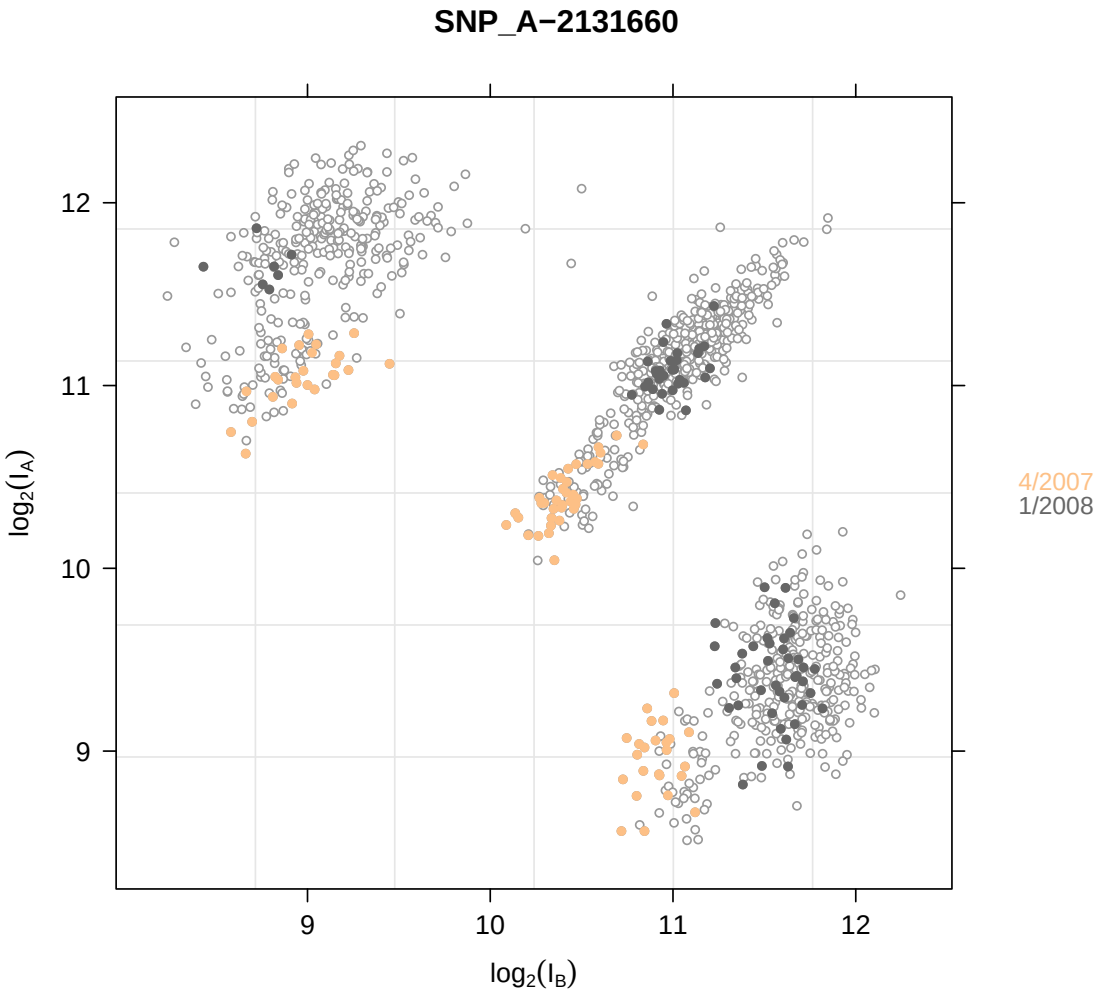
People are a mixture of AA, AB, and BB genotypes



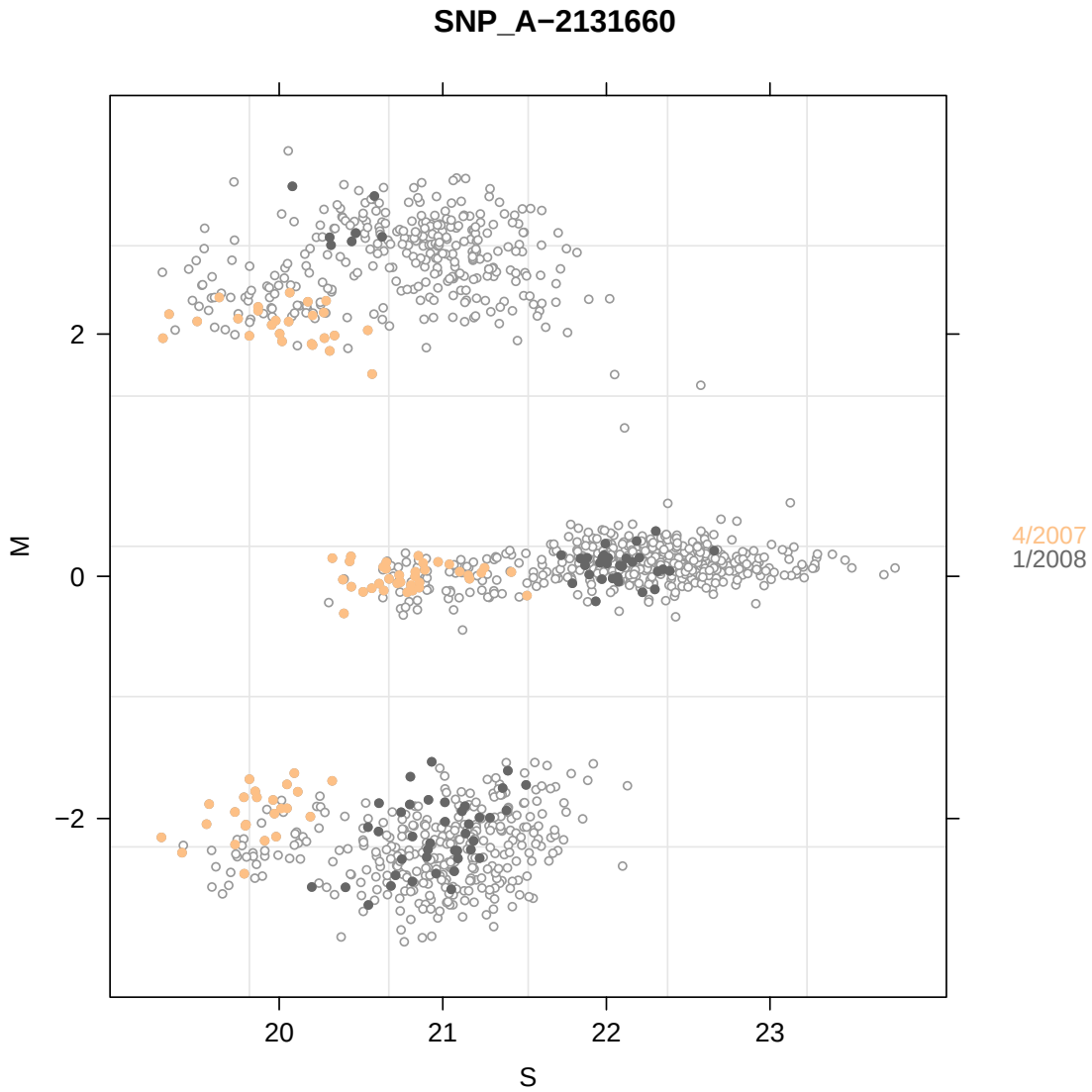
Batch effects



Batch effects



The log-ratio (M) is more stable



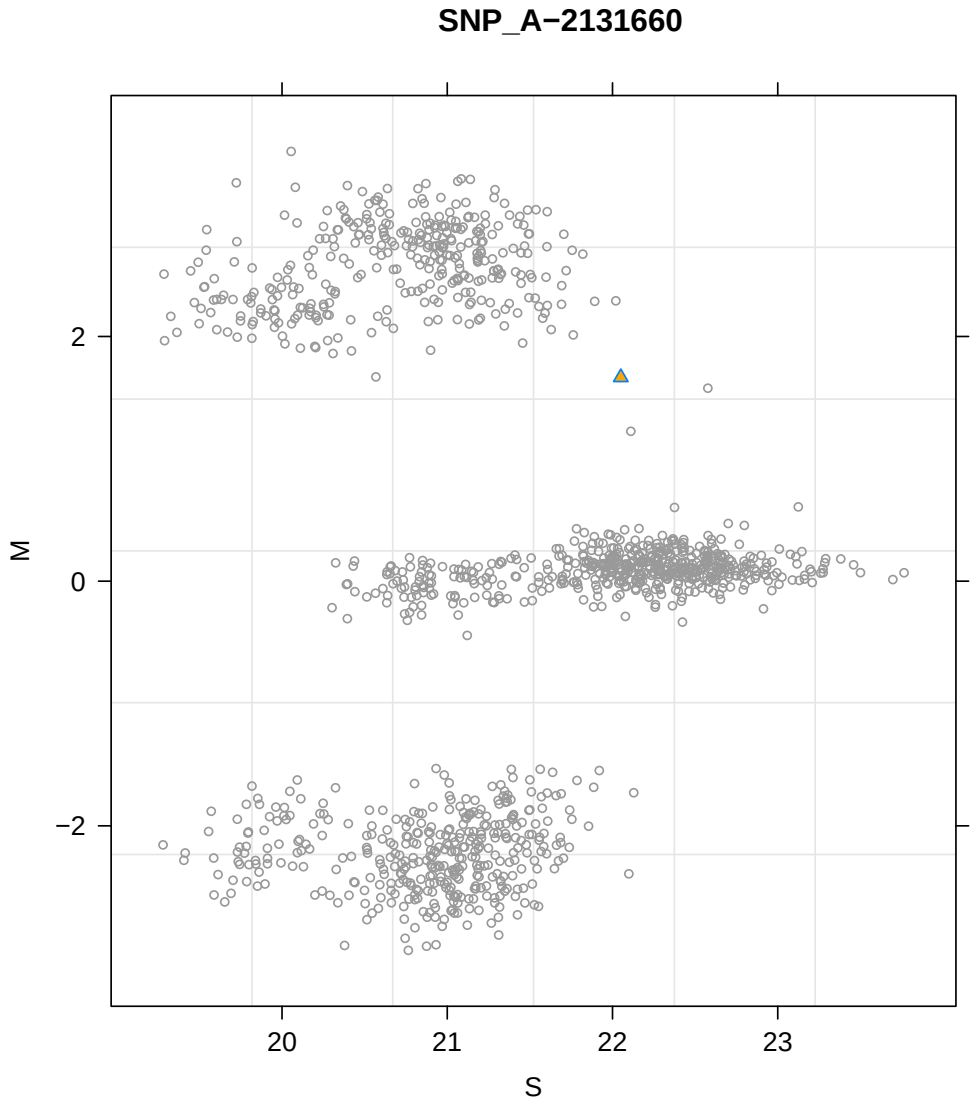
Copy number estimation

Leek 2010: *Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study*

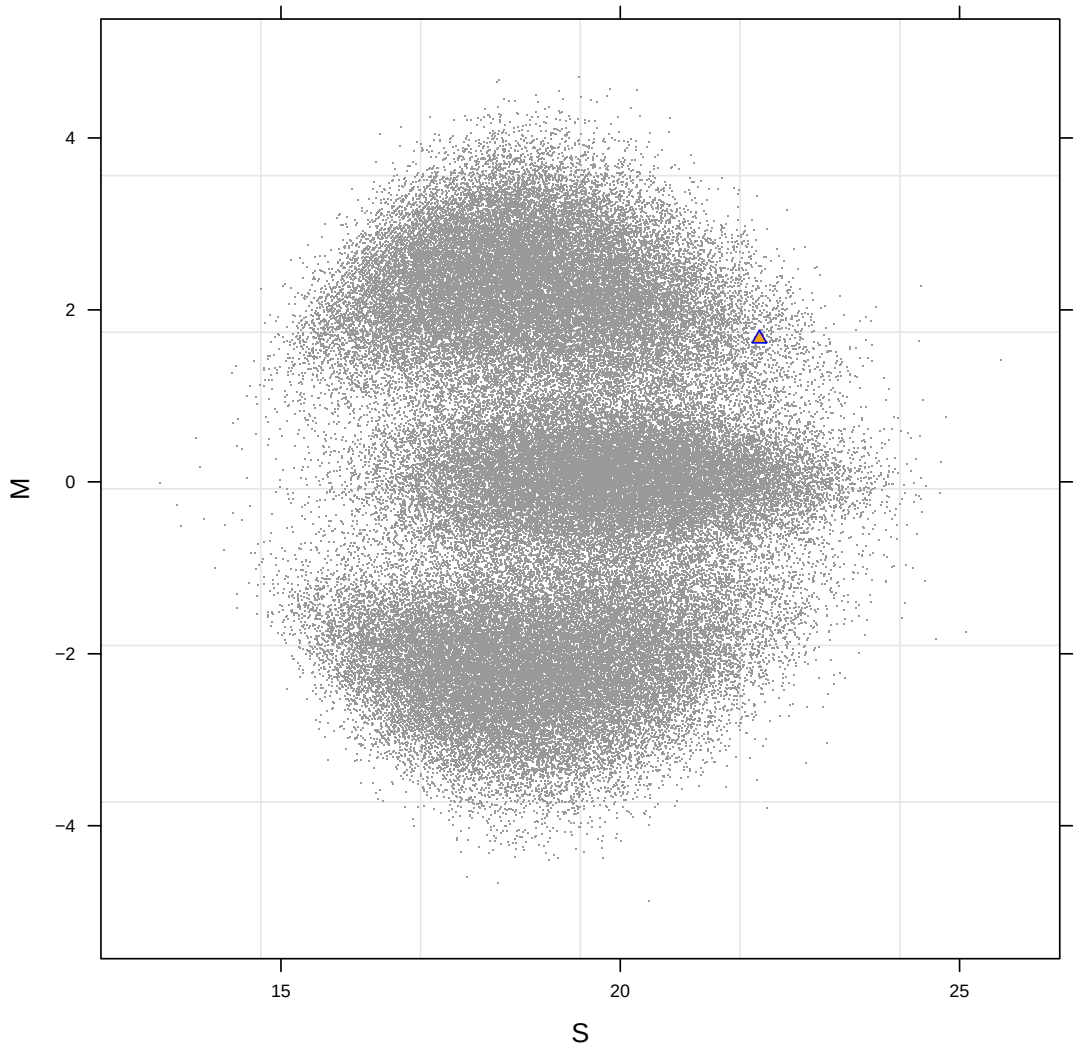
- There is an aliasing of signal strength S and batch effects
- Genotypes estimated from the log-ratios, M, are robust to batch effects.

Genotypes robust to batch effects

Need to model the dependency of M on S

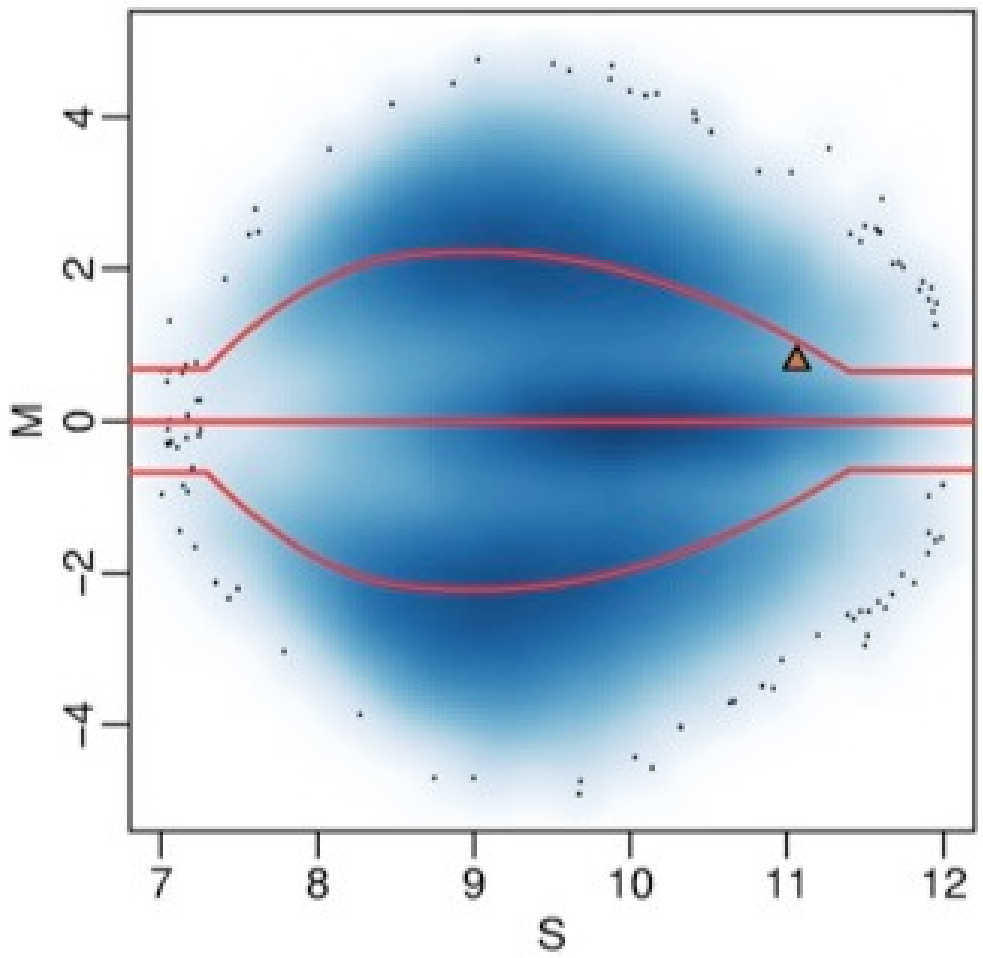


M-S relationship can be estimated from each sample



Estimate the M-S relationship for each sample

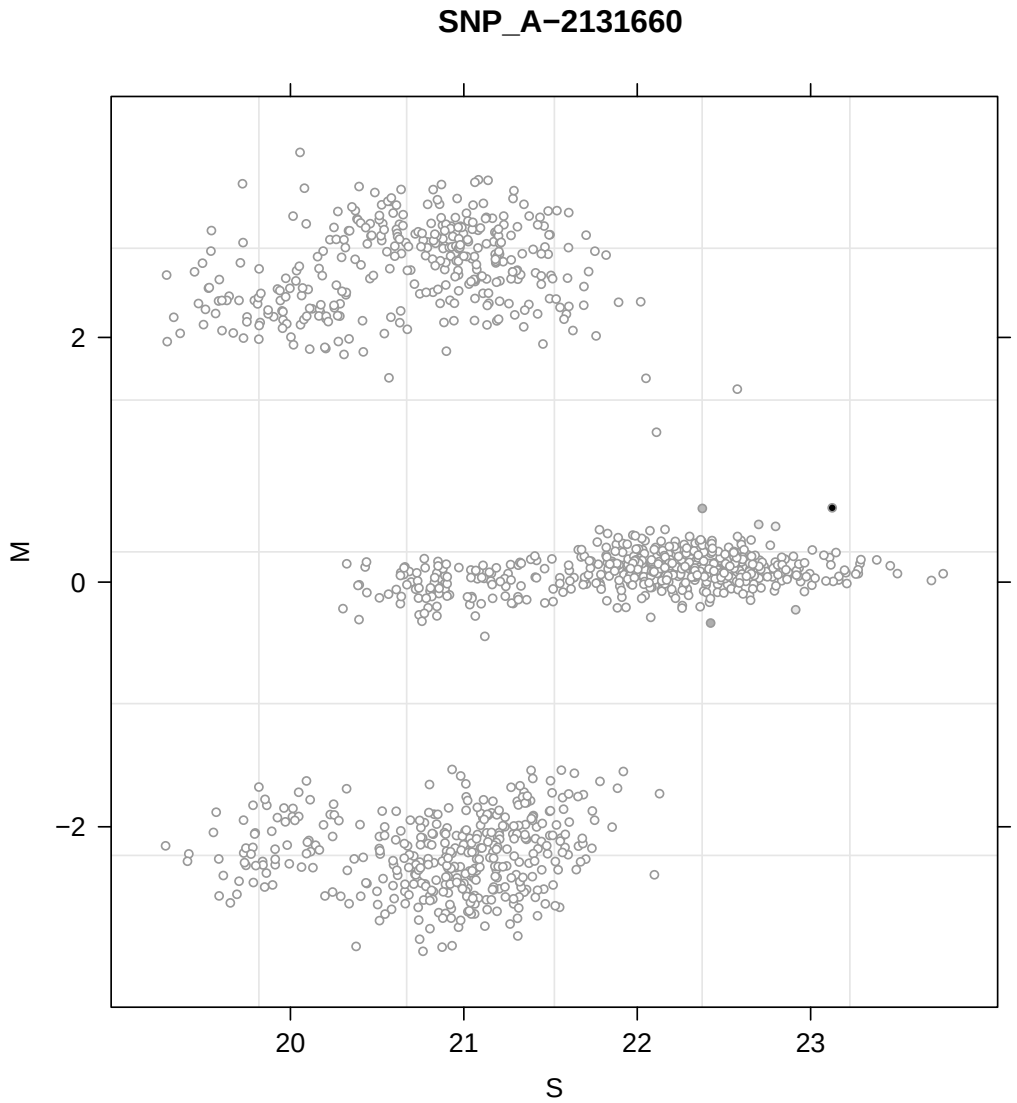
Fit a mixture model



Carvalho 2010: $[M_{ijk}|\mu_{ig}, \lambda_{ijg}] = f_{jkg}(S_{ijk}) + \mu_{ig} + \lambda_{ijg} + \sigma_{ig}\epsilon_{ijkg}$

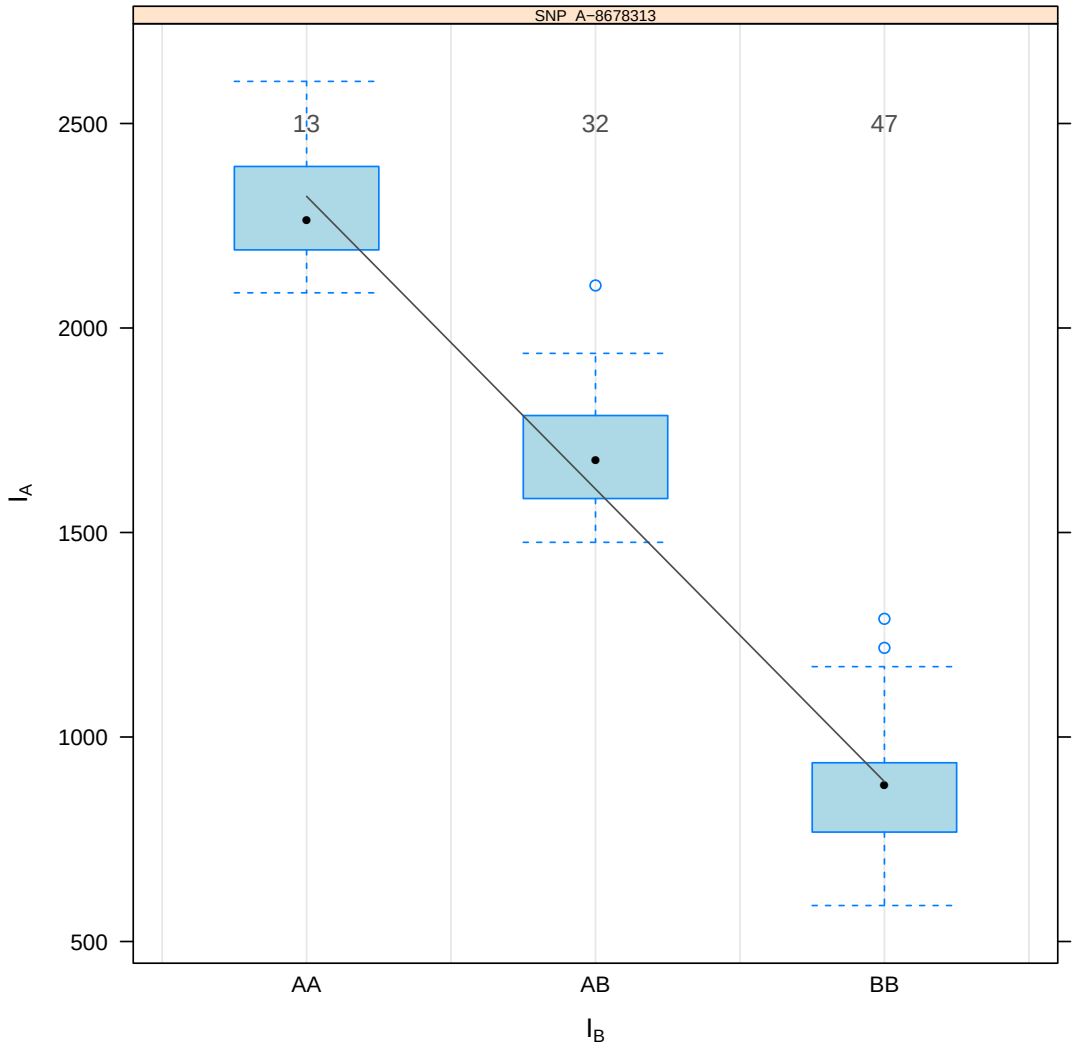
Estimated genotypes

Confidence scores



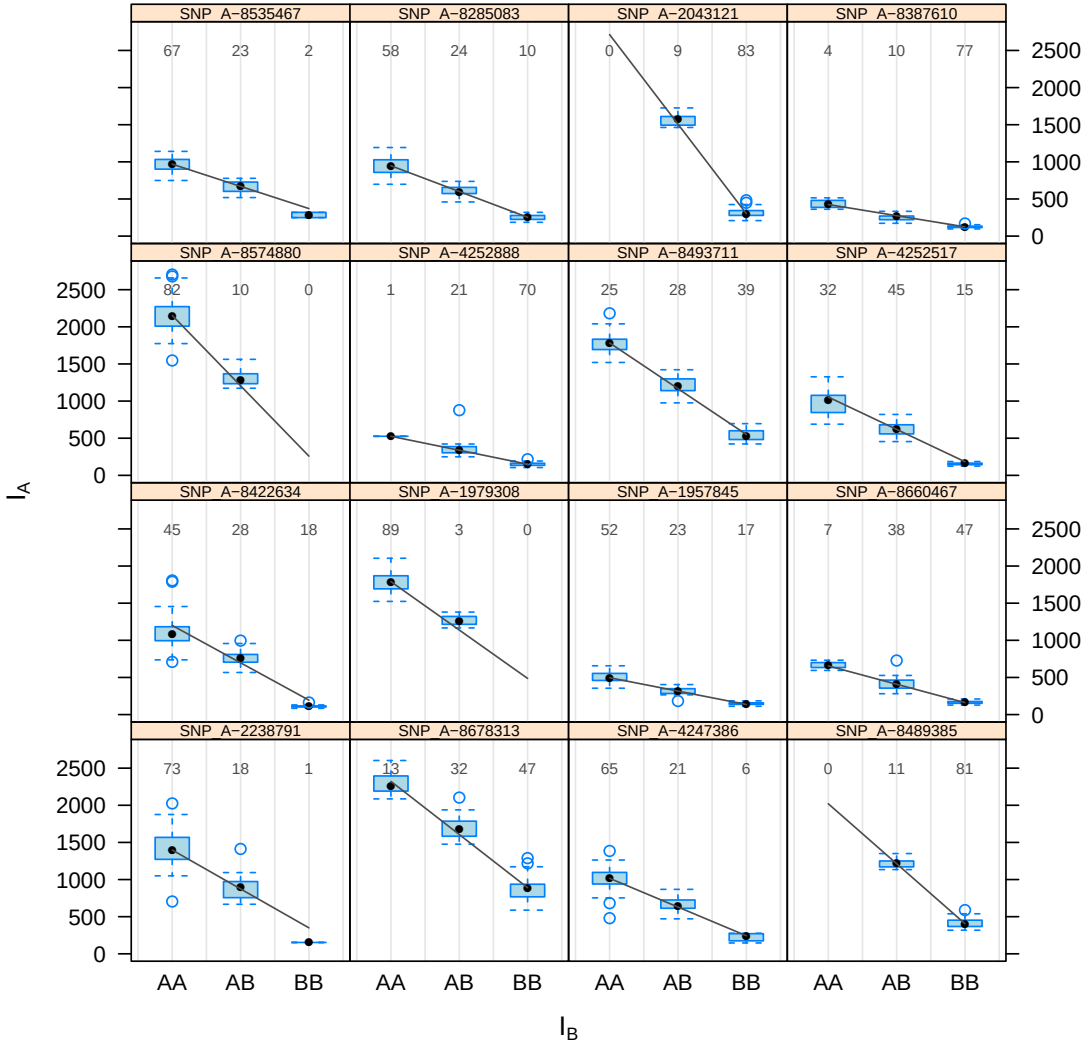
Use genotypes to guide copy number estimation

Model for allele A



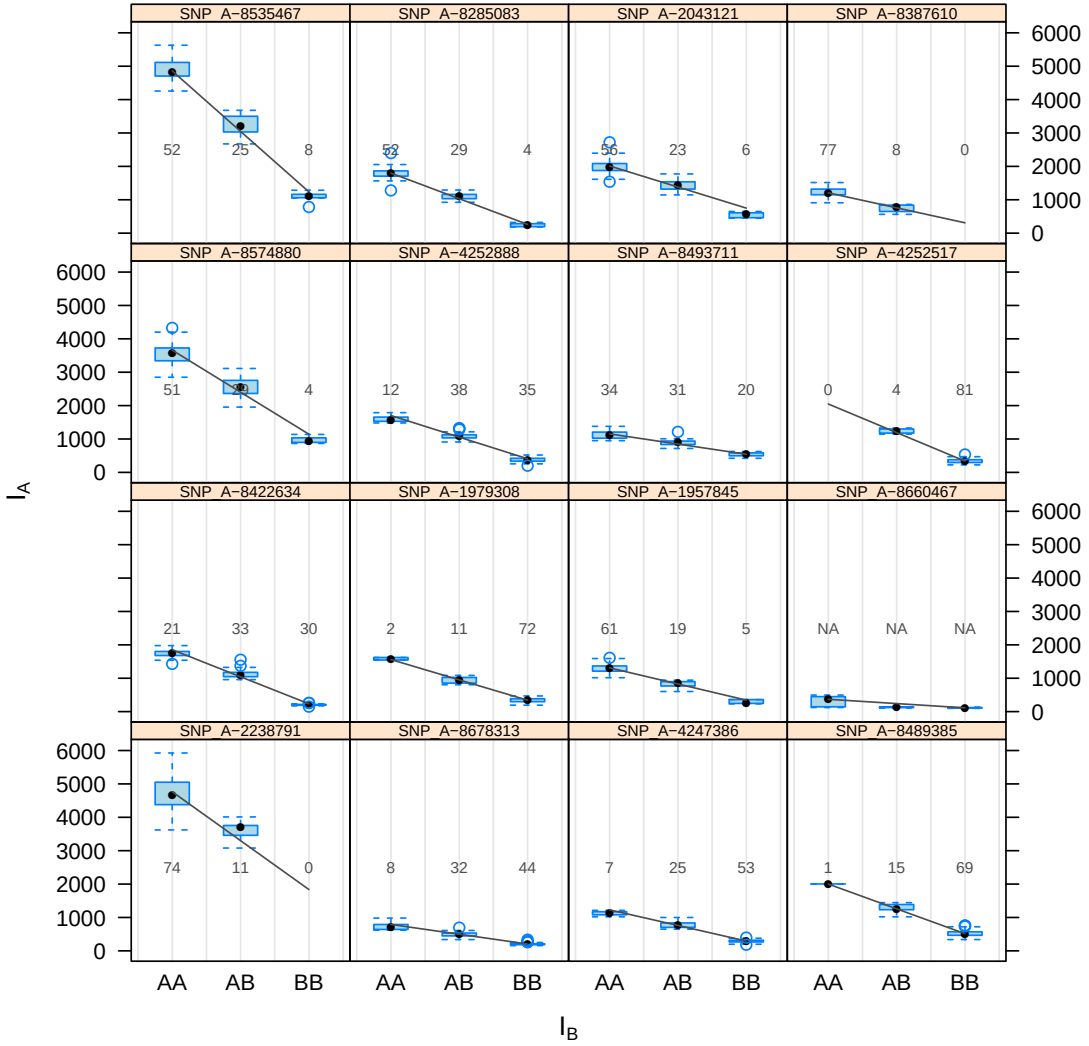
16 randomly selected SNPs

Model for allele A



16 randomly selected SNPs

Model for allele A



Model for copy number

$$
\begin{aligned}
\left[\ I_{ijkl}\ \right] &= \left[\left(\ \text{Optical}_{ijl} + \text{Nonspecific}_{ijl}\ \right) \times \left(\ \delta_{ijkl}\ \right)\right] + \\
&\quad \left[\ \text{Specific}_{ijl} \ \times \ \varepsilon_{ijkl}\ \right] \\
&\equiv \left[\ \nu_{ijl} \ \times \ \delta_{ijkl}\ \right] + \left[\ \phi_{ijl} c_{ijkl} \ \times \ \varepsilon_{ijkl}\ \right] \text{ for } l \in \{A, B\}.
\end{aligned}
$$

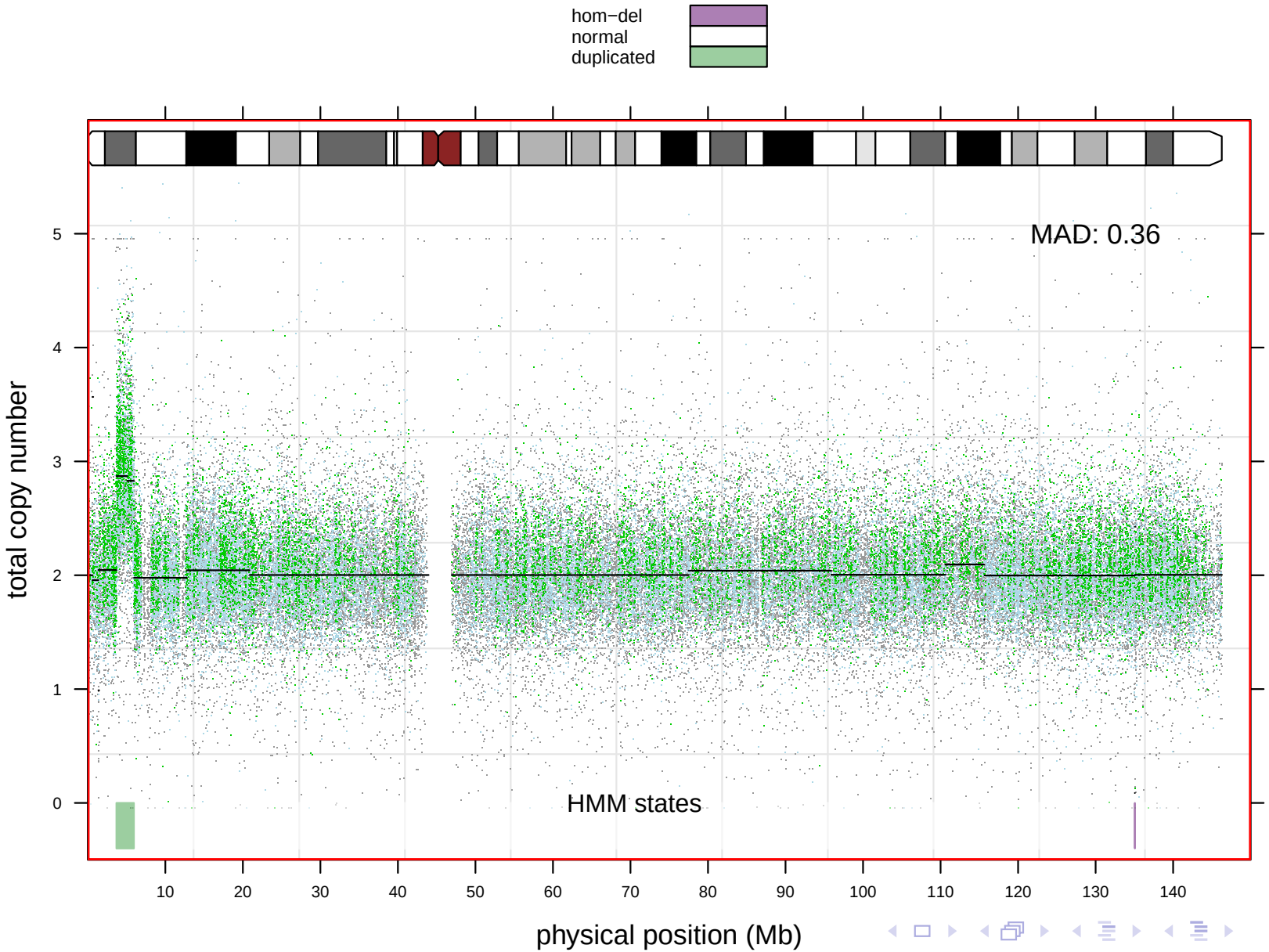
Estimation procedure:

- treat c_{ijkl} as known from the diallelic genotypes.
- for each allele, fit the linear model to robust measures of the within-genotype means

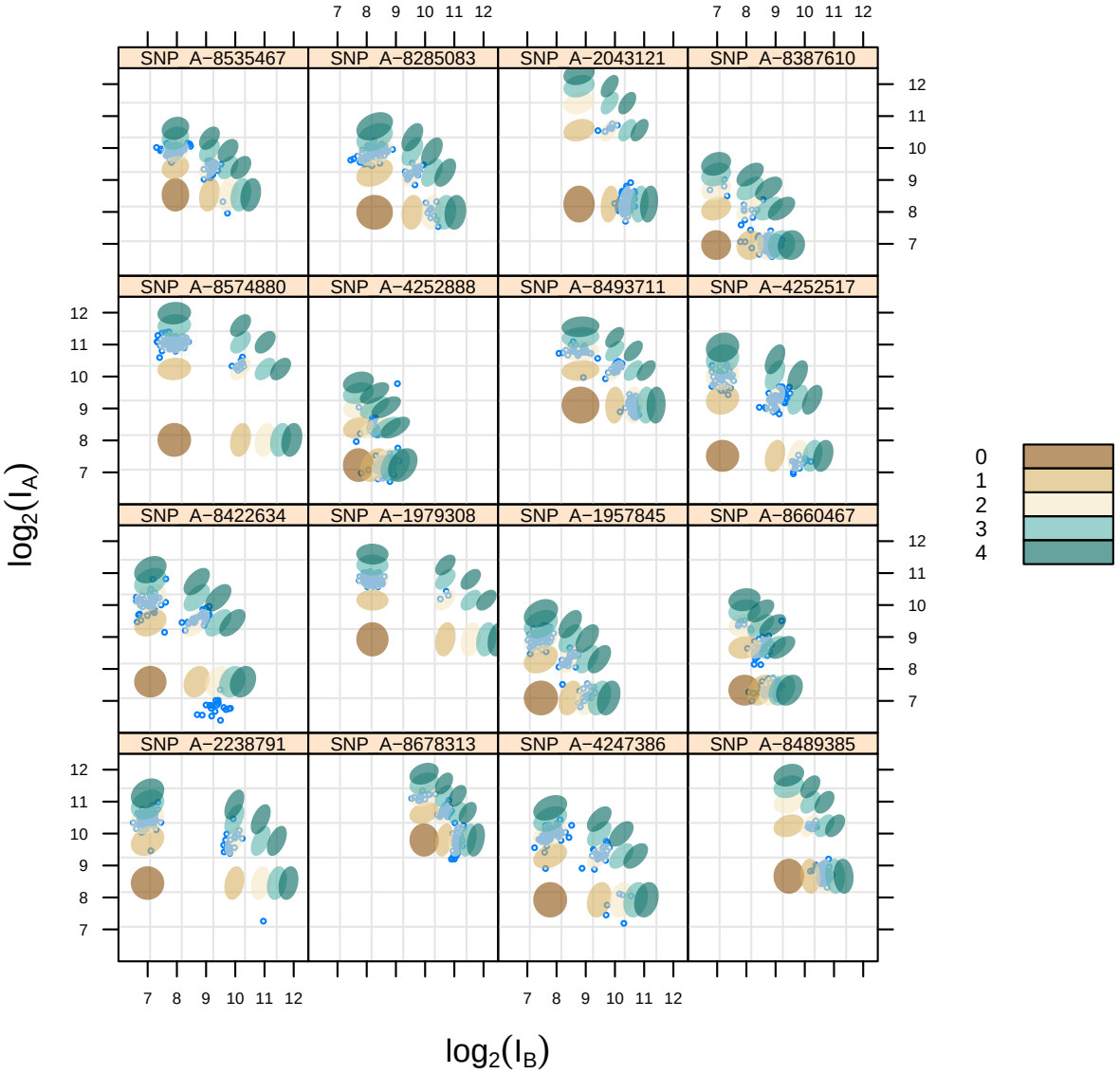
Raw copy number

$$\hat{c}_{ijkl} \quad = \quad \max\left\{ \frac{1}{\hat{\phi}_{ijl}} \left(I_{ijkl} - \hat{\nu}_{ijl} \right),\ 0 \right\} \text{ for } l \in \{A, B\}.$$

Inferring CNV



Bivariate normal prediction regions

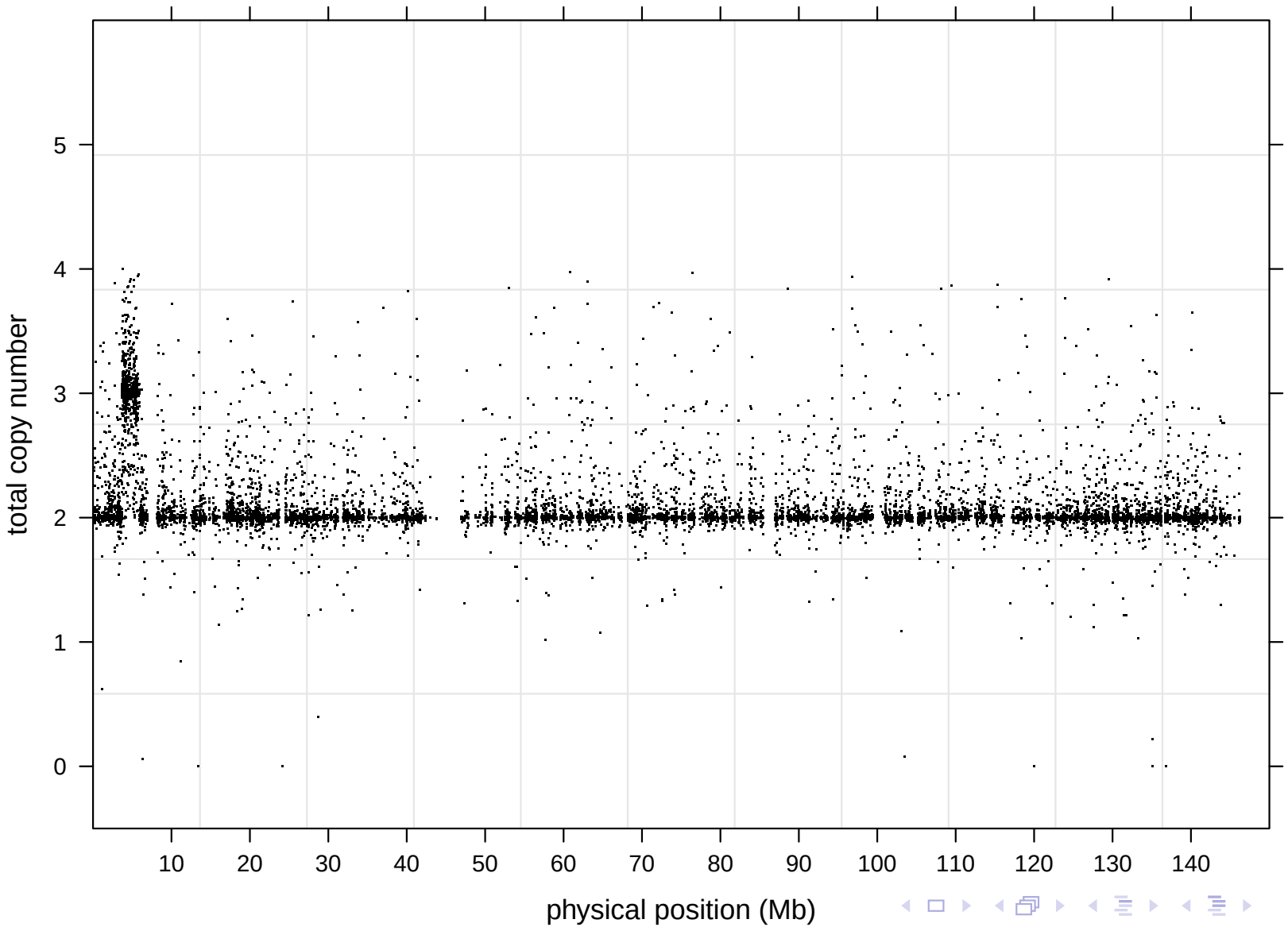


Ellipses shown for one batch

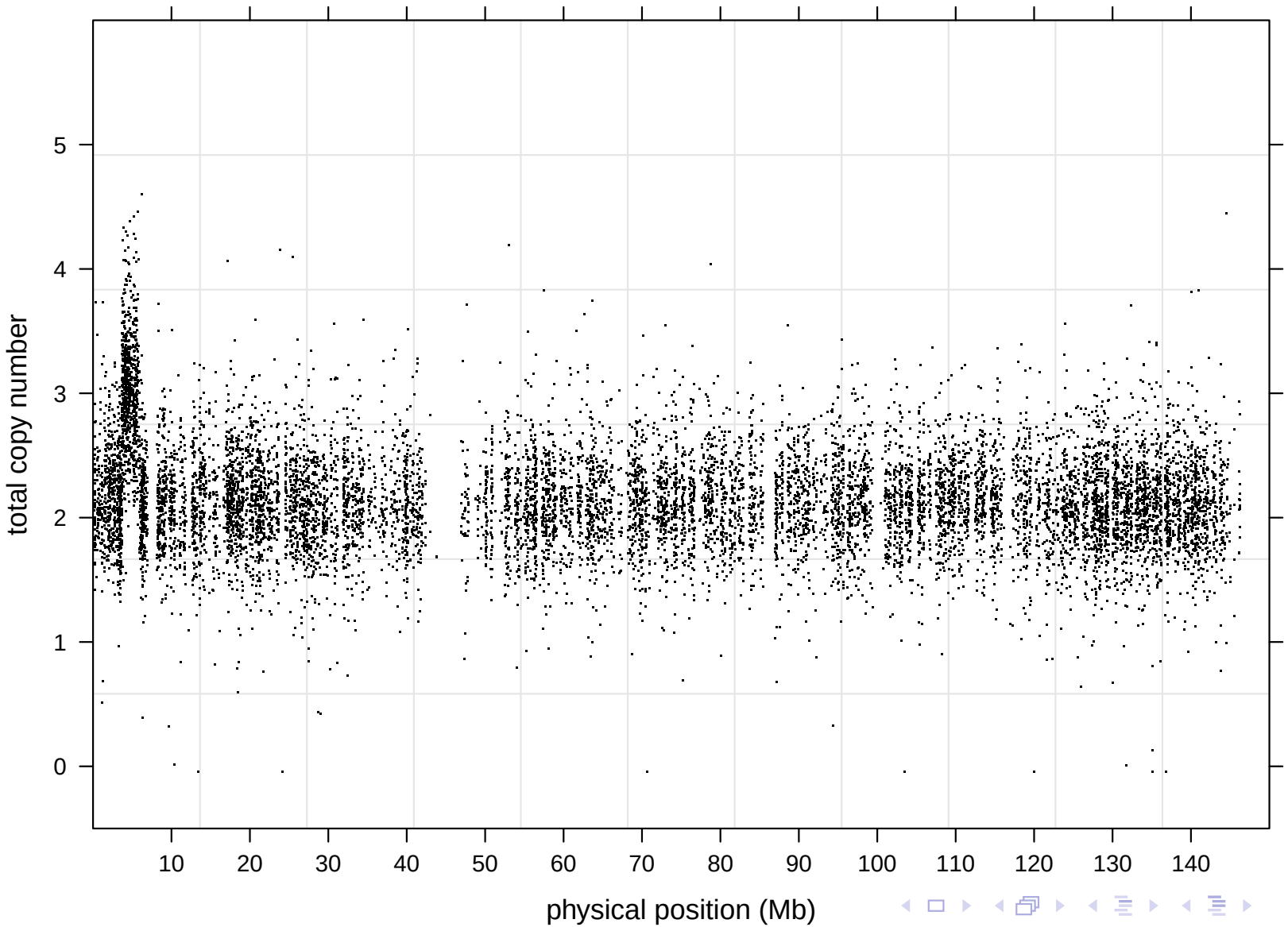
Bivariate normal prediction regions

$$\left[\begin{array}{c} \log_2(I_{ijkA}) \\ \log_2(I_{ijkB}) \end{array} \middle| \begin{array}{c} C_{ijkA} = c_A \\ C_{ijkB} = c_B \end{array} \right] \sim \mathrm{N}\left(\left[\begin{array}{c} \log_2(\nu_{ijA} + c_A\phi_{ijA}) \\ \log_2(\nu_{ijB} + c_B\phi_{ijB}) \end{array} \right], \boldsymbol{\Sigma}_{ij} \right).$$

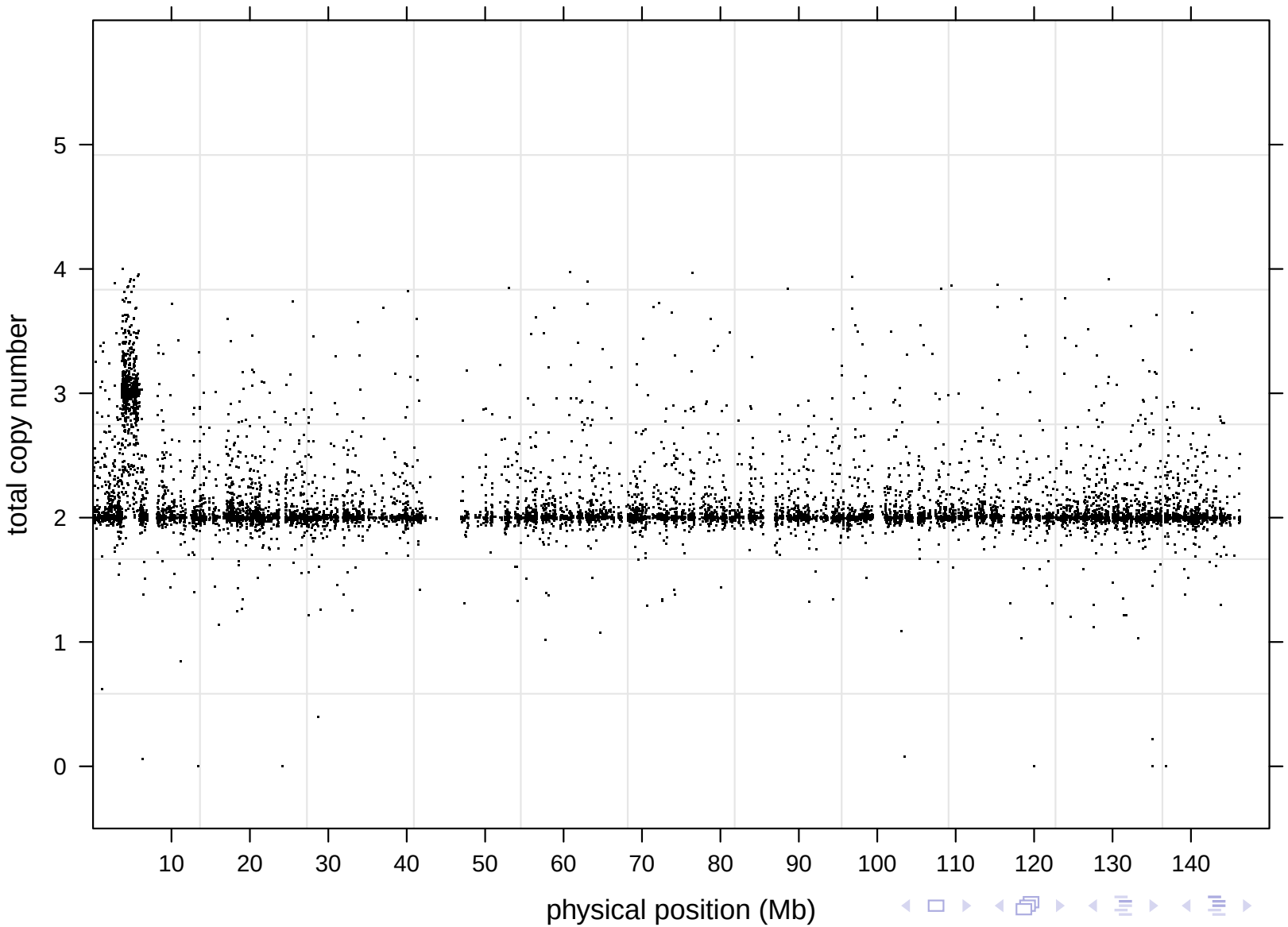
Posterior mean



Raw copy number

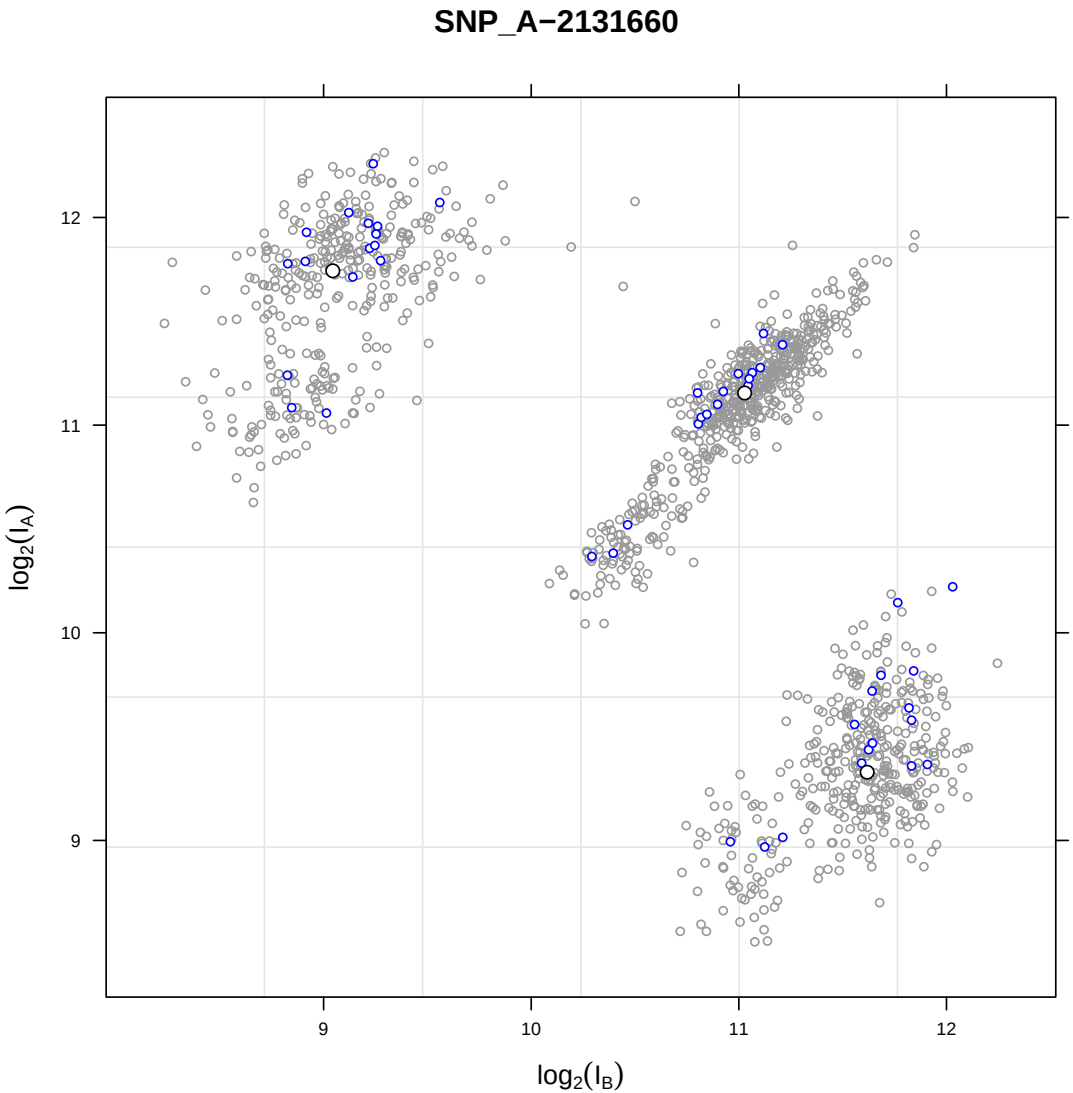


Posterior mean

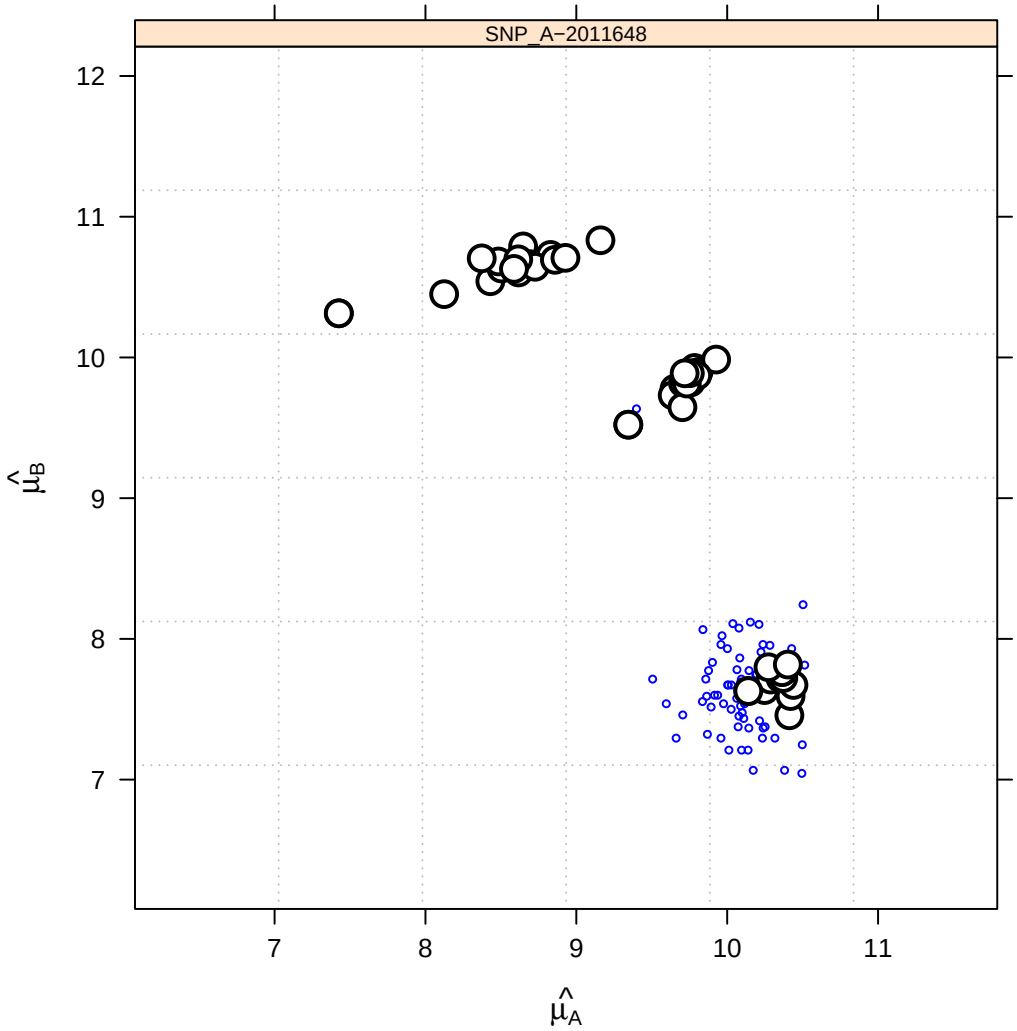


Multiple levels of variation

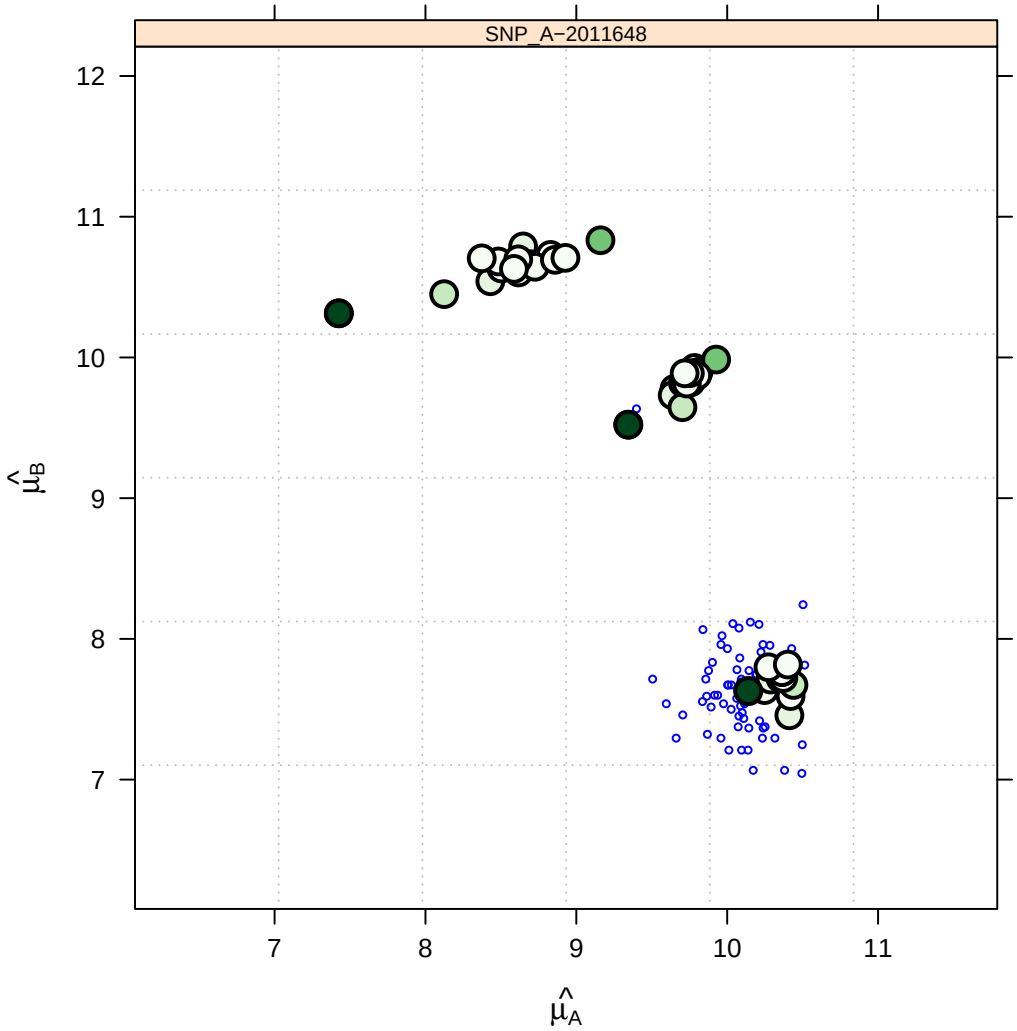
Normalized intensities



Principal components analysis of the batch summaries



Principal components analysis of the batch summaries



The batch-effect is consistent across SNPs.

Summary

- the total signal (A+B) in high-throughput SNP arrays is highly susceptible to batch differences
- genotypes based on log-ratios are robust to batch effects
- batches with largest effect on signal are reproducible across SNPs
- with appropriate experimental design, batch effects can be modeled as part of the estimation procedure for copy number
- `crlmm` compendium:
  `www.biostat.jhsph.edu:~/rscharpf/crlmmCompendium`

Acknowledgements

- Ingo Ruczinski
- Rafael Irizarry
- Benilton Cavalho
- Matthew Ritchie
- Jeff Leek