Estimating Genome-wide Copy Number Using Allele Specific Mixture Models

Rafael A. Irizarry

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

Ok Steve, I get it. I'll work harder on Copy Number.

Rafael A. Irizarry

Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health

Acknowledgements

- Benilton Carvalho, JHU Biostat
- Wenyi Wang, UC Berkeley
- Terry Speed, UC Berkeley
- Shin Lin, UPenn
- Simon Cawley, Affymetrix
- Aravinda Chakravarti, JHU IGM
- Dan Arking, JHU IGM
- Dave Cutler, JHU IGM
- James MacDonald, Michigan
- Seth Falcon, Robert Gentleman and Bioconductor Team

Affymetrix SNP chip terminology

Genomic DNA

SNP

A

TAGCCATCGGTANGTACTCAATGAT

G

Perfect Match probe for Allele A **ATCGGTAGCCATTCATGAGTTACTA**

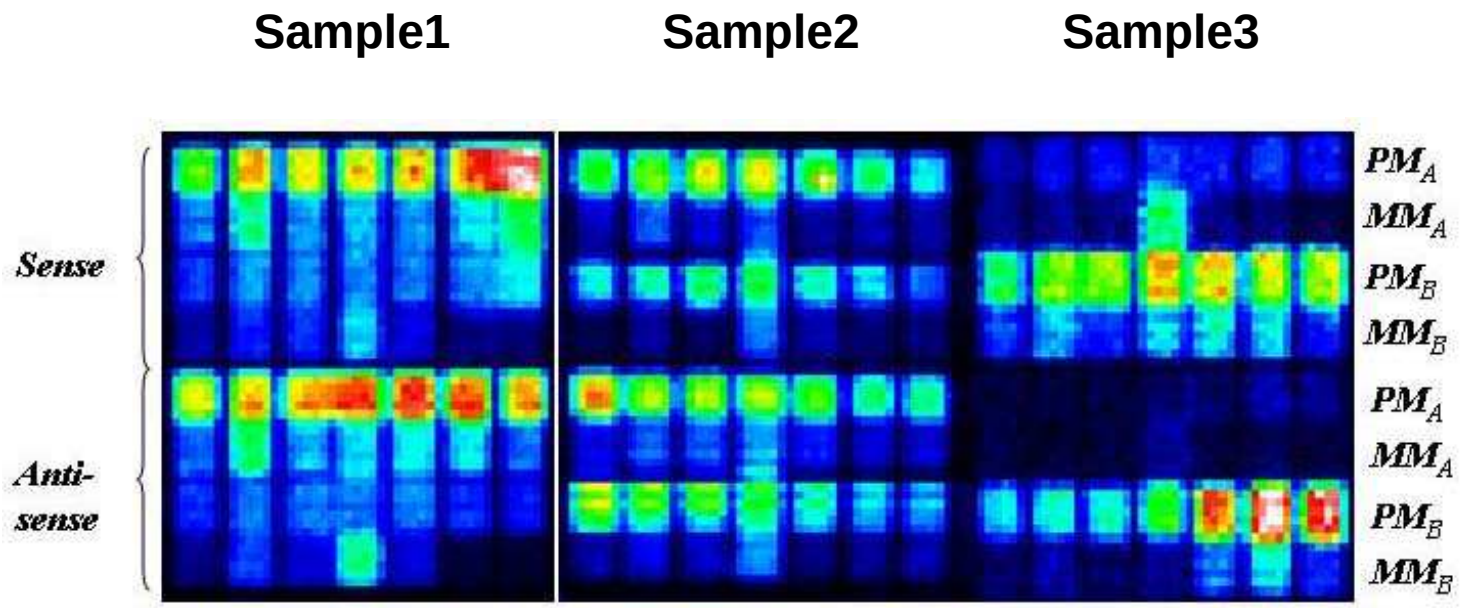
Perfect Match probe for Allele B **ATCGGTAGCCATCCATGAGTTACTA**

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?

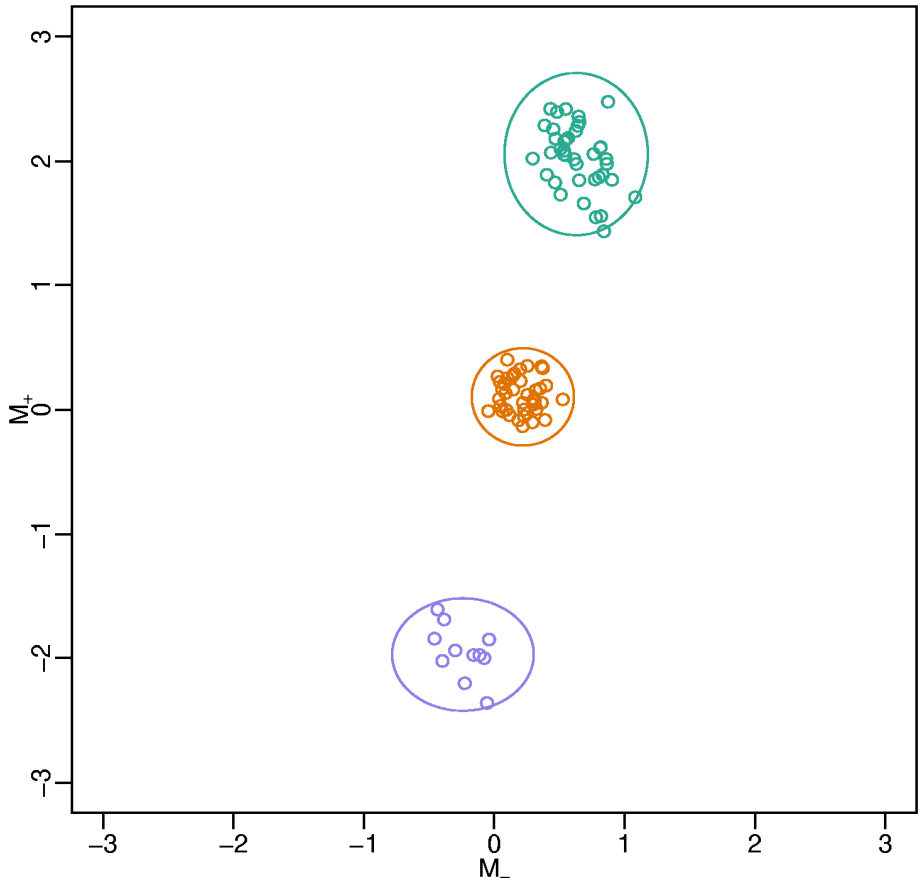
Probe Intensities

Fake (idealized) image for 3 samples on one SNP

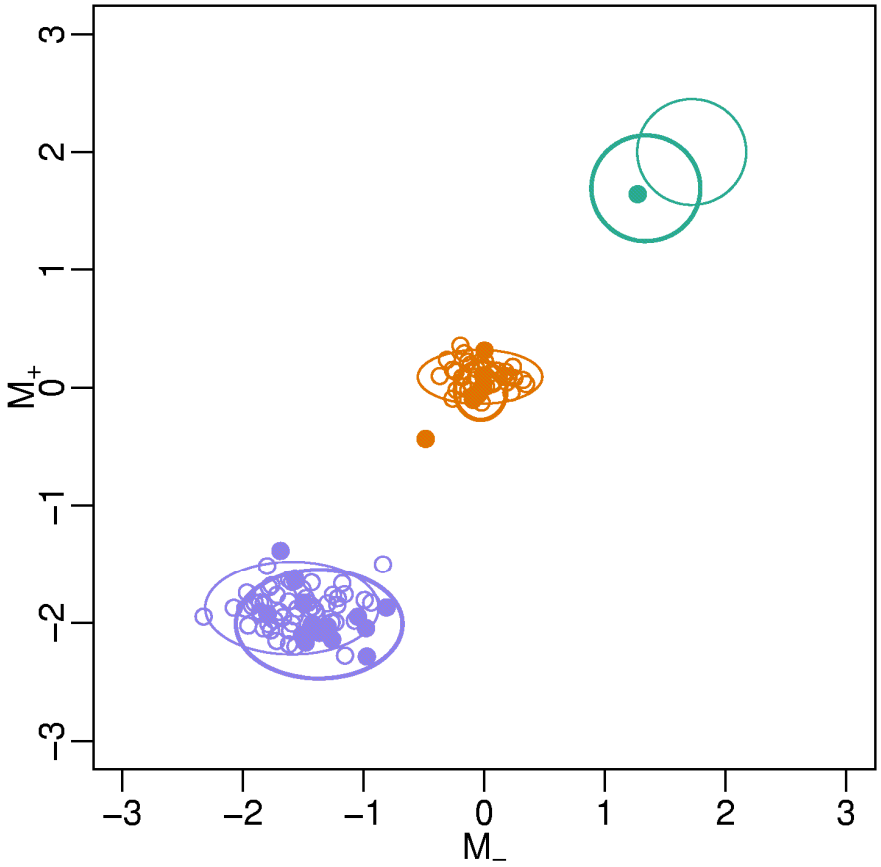


Fake, as the probes are not all adjacent on the chip
Idealized, as all the probes are high or low as they should be.

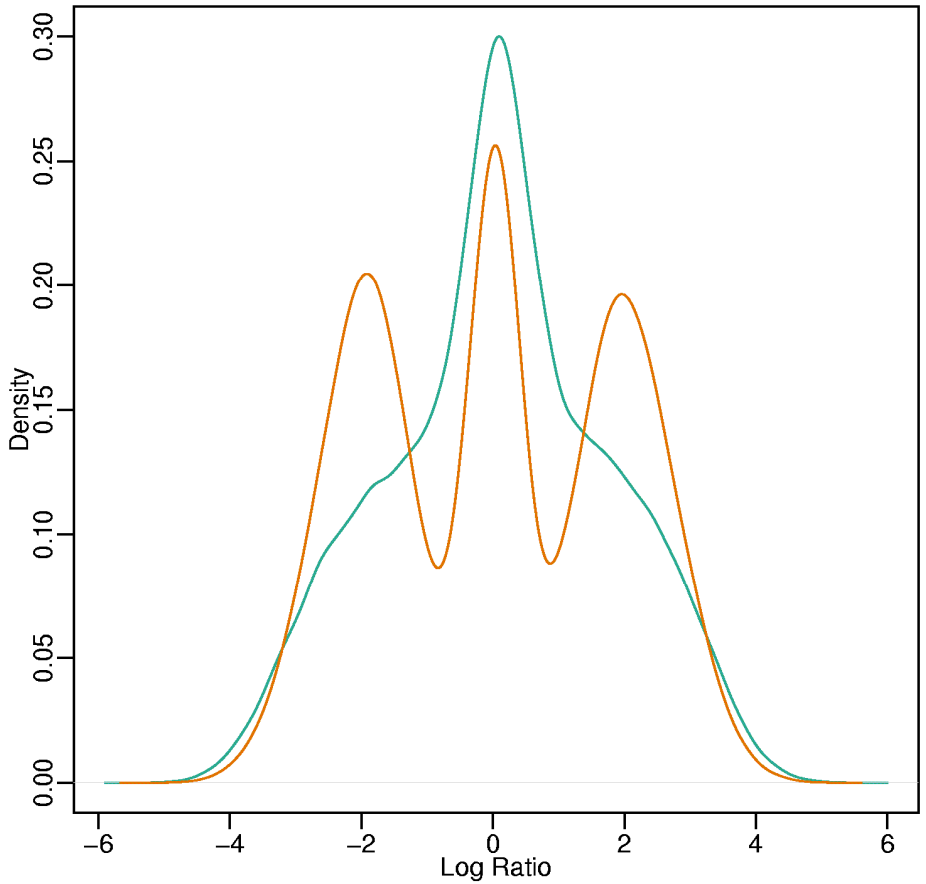
Fit model: define genotype regions and confidence measure



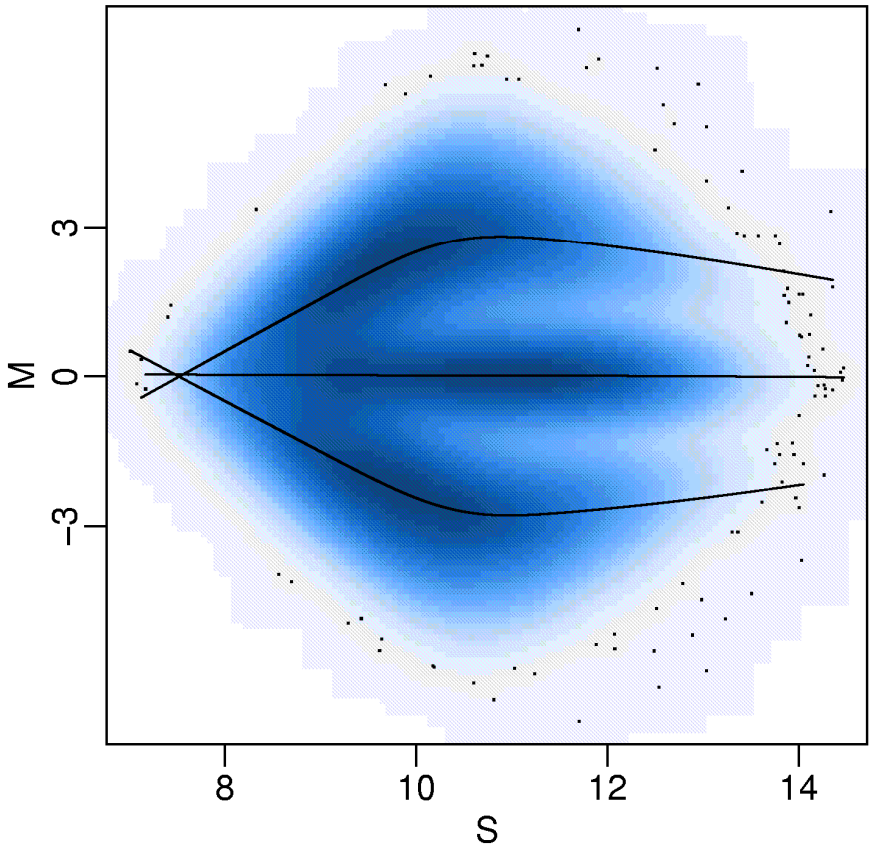
Low frequency alleles?



Different quality hybes



Intensity effect on M



Confidence measures

Figure 2A

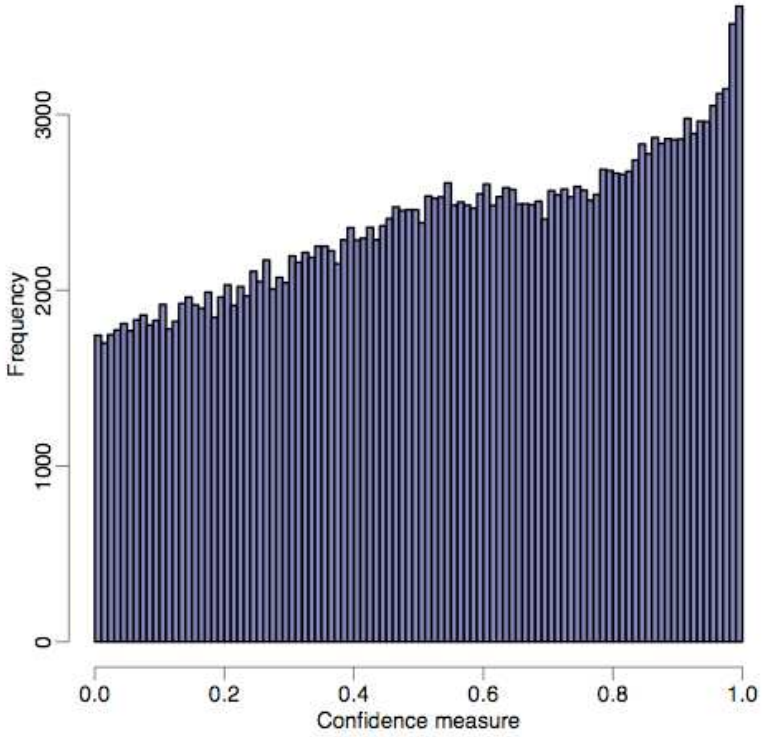
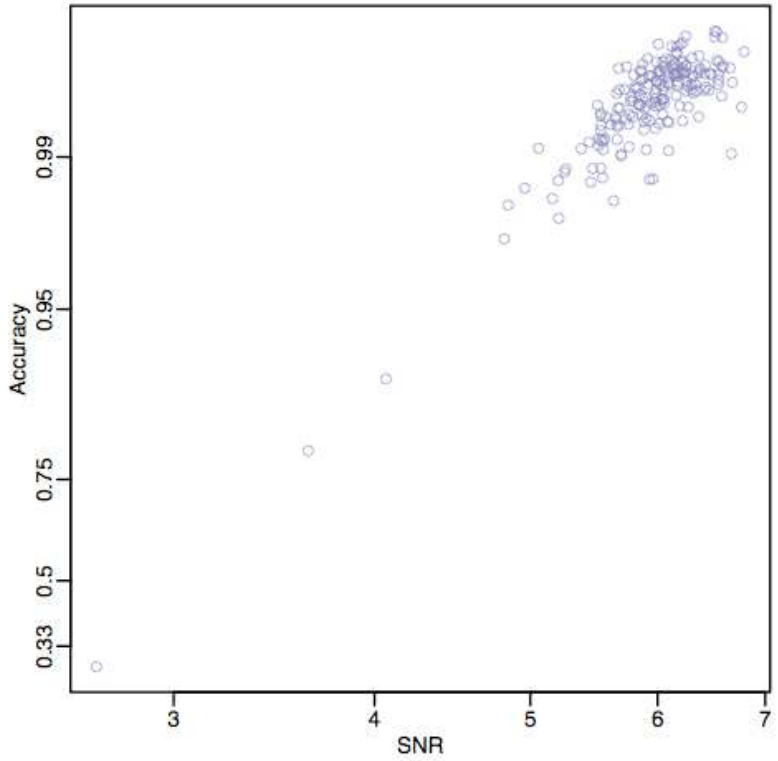
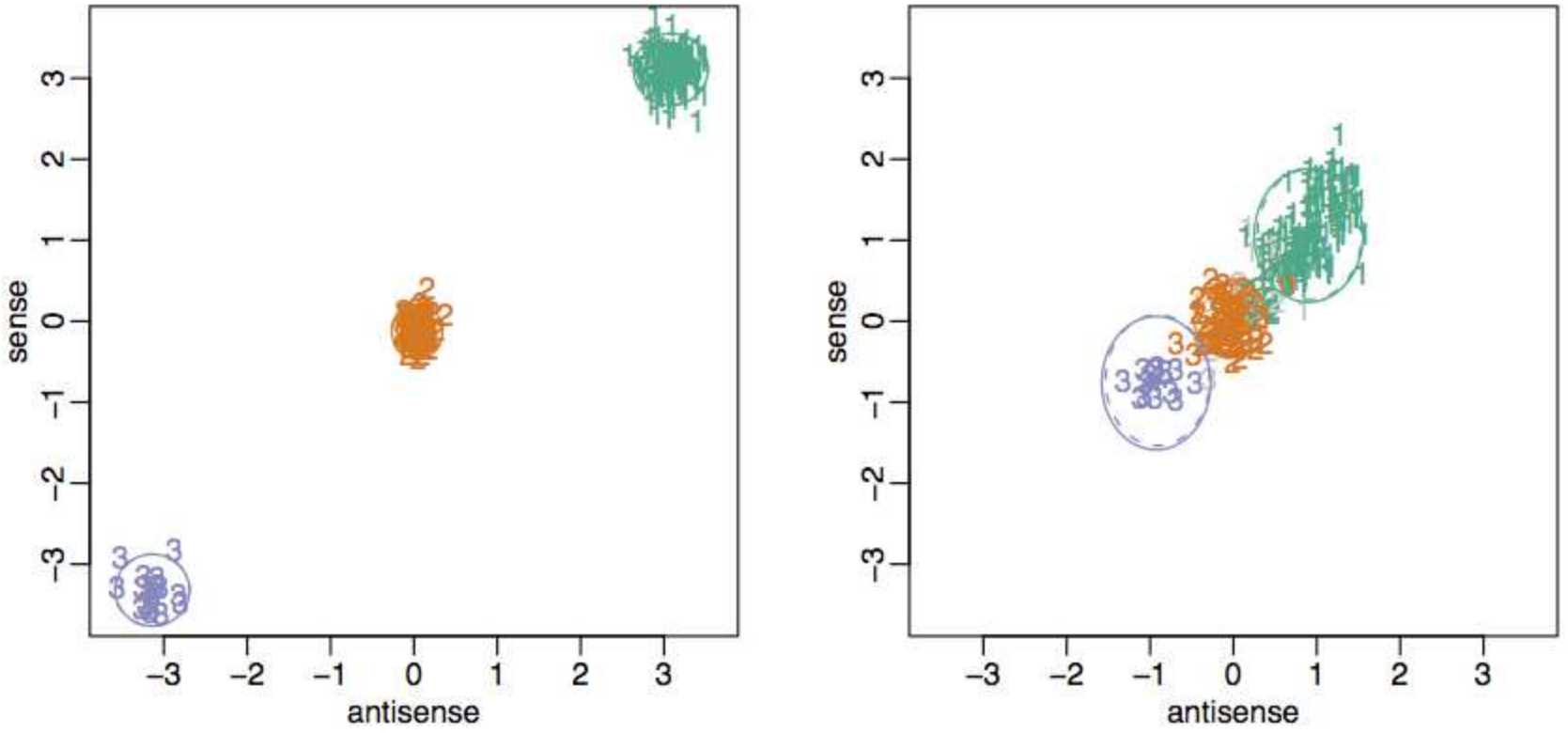


Figure 2B

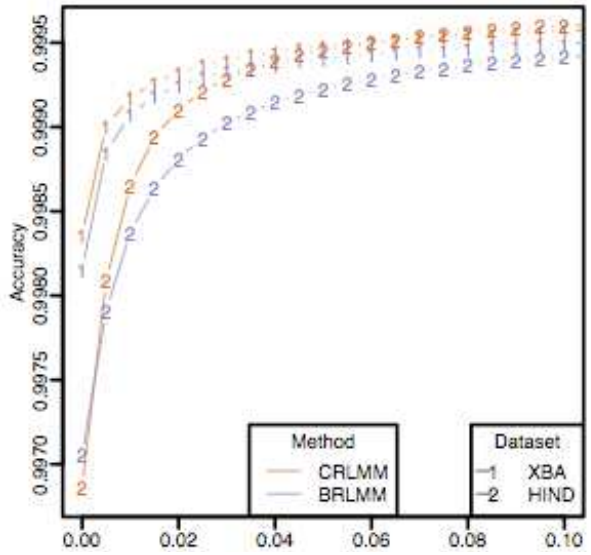


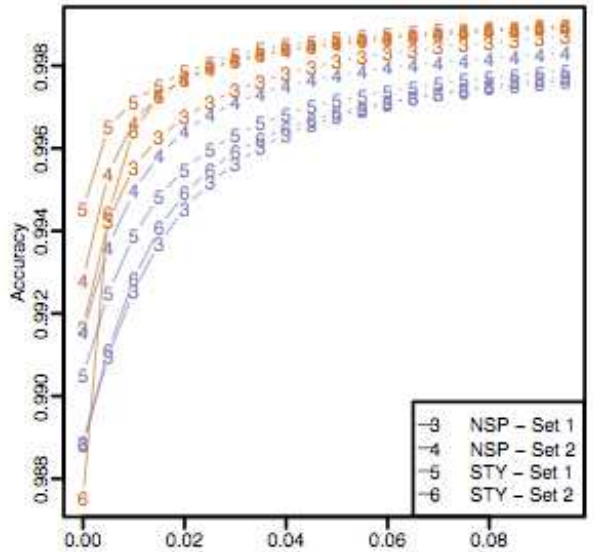
All arrays not equal!

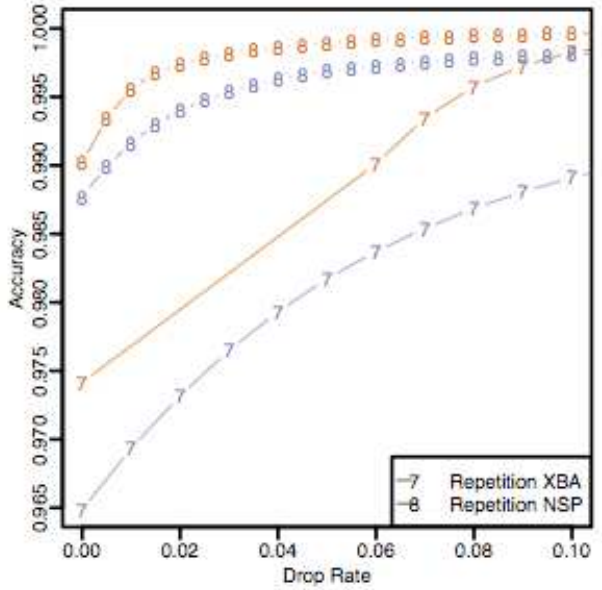
All SNPs not equal

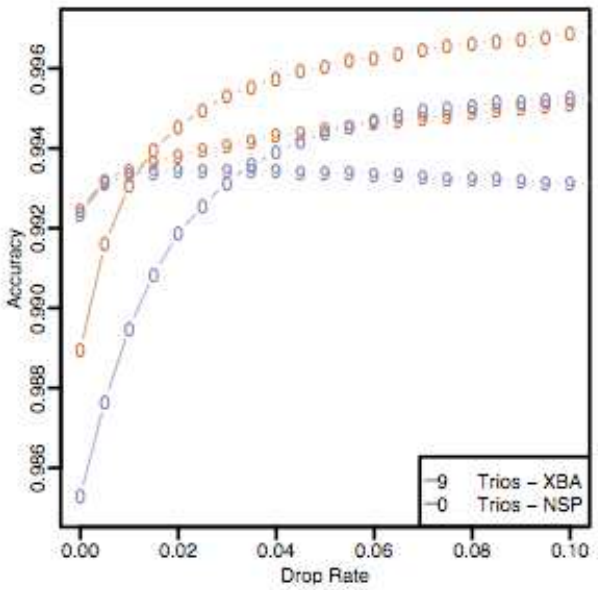


Accuracy versus Drop Rate

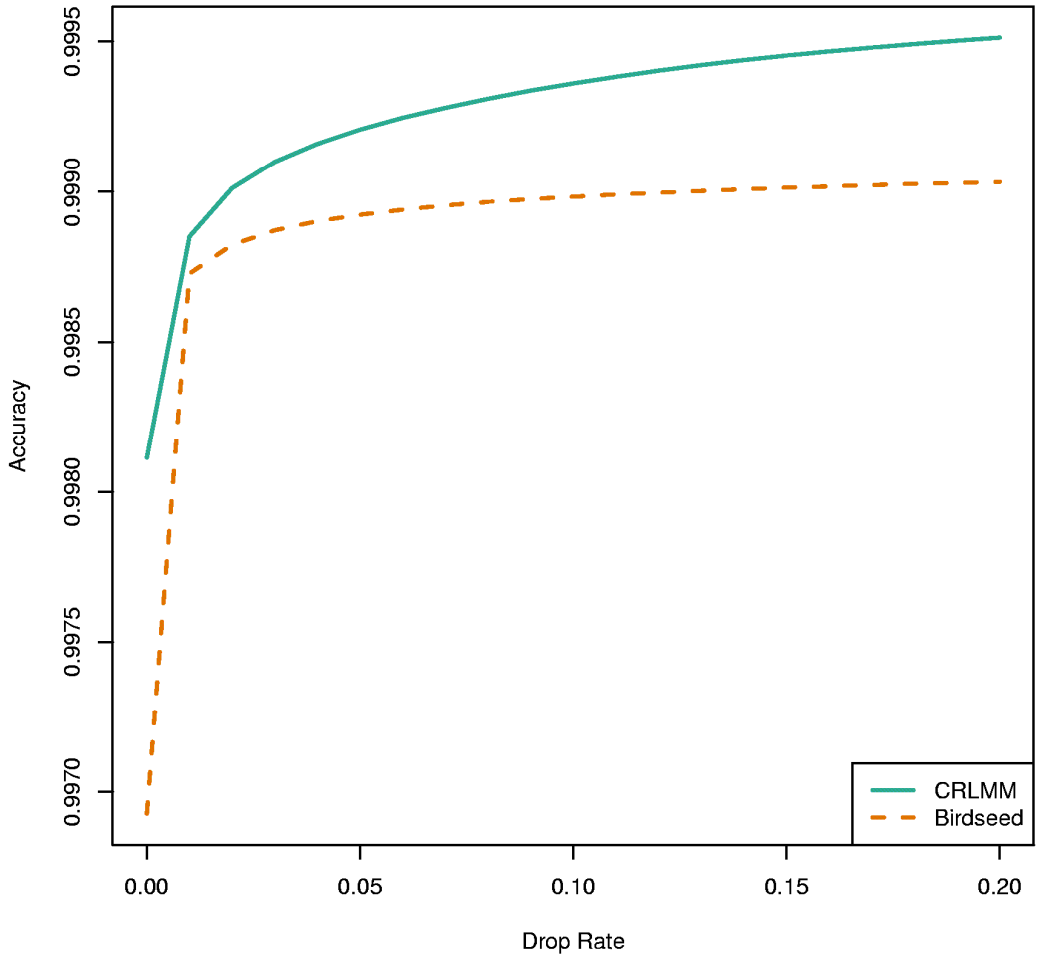








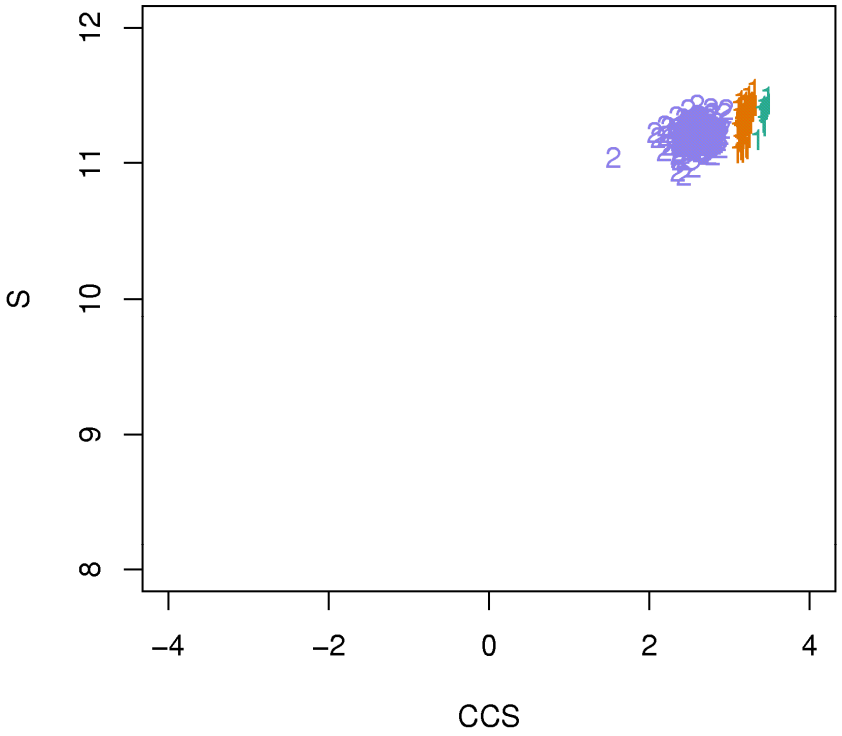
CRLMM v birdseed



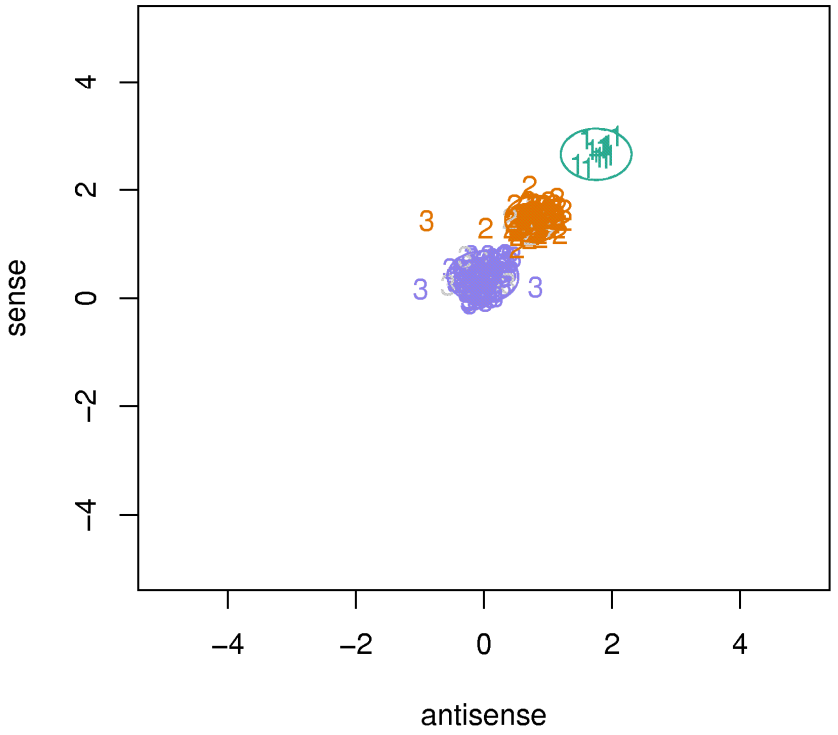
Why CRLMM better?

Big Shifts

BRLMM

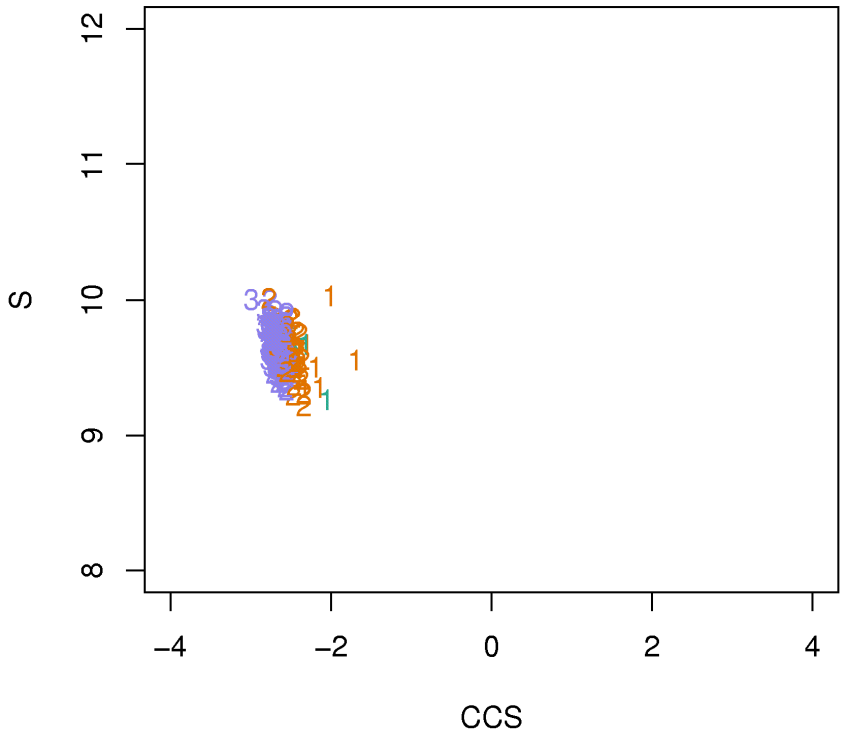


CRLMM

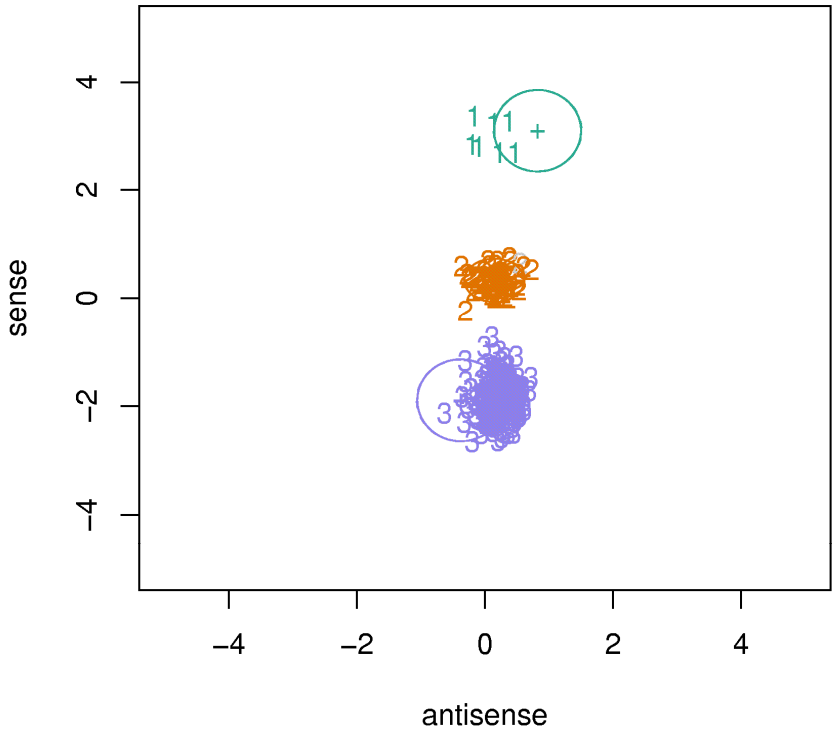


“Room for improvement” Probes

BRLMM



CRLMM



Improved Accuracy Prediction

Figure 3A

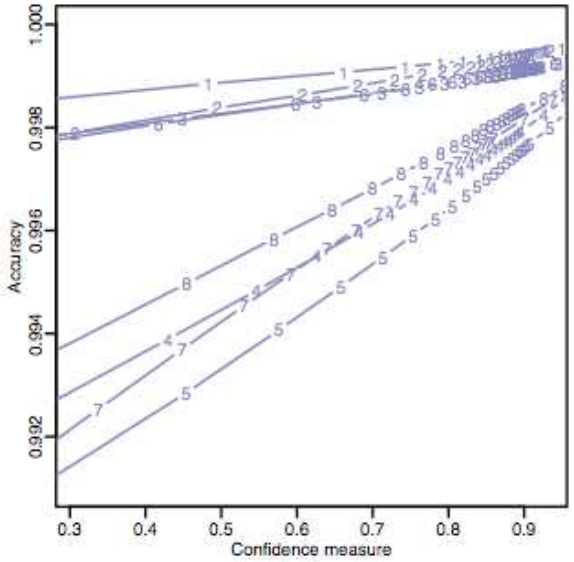


Figure 3B

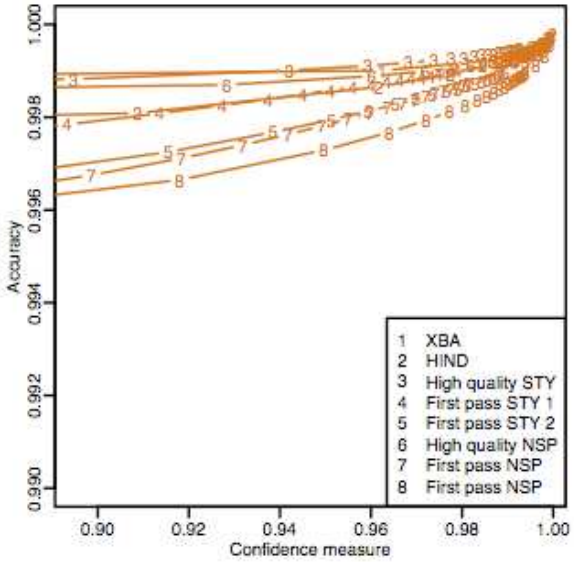
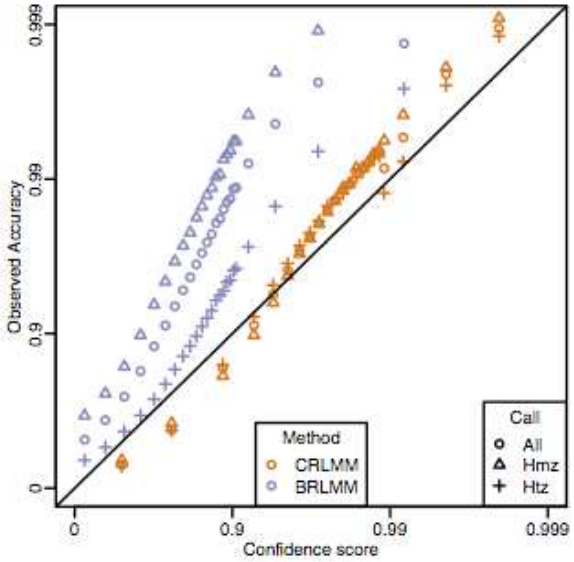
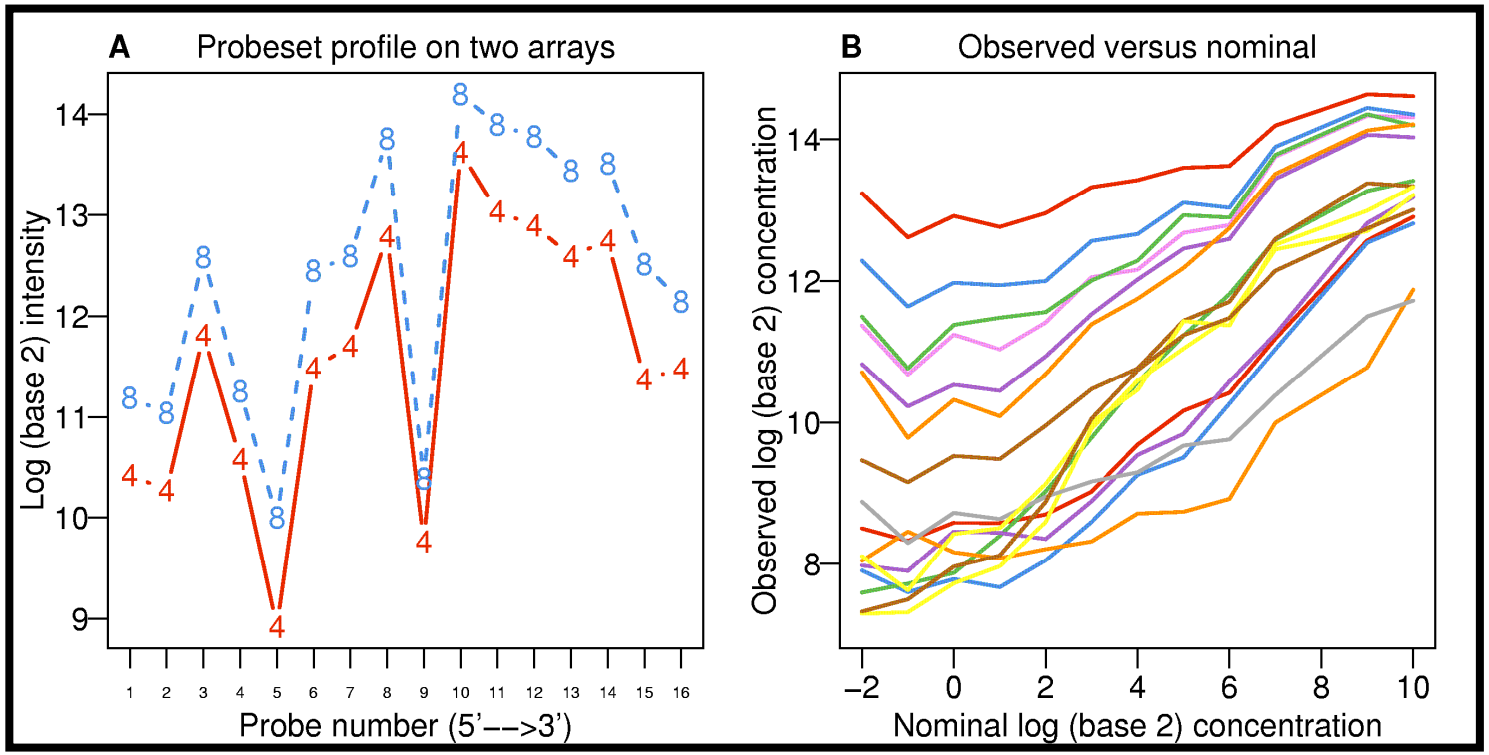


Figure 3C

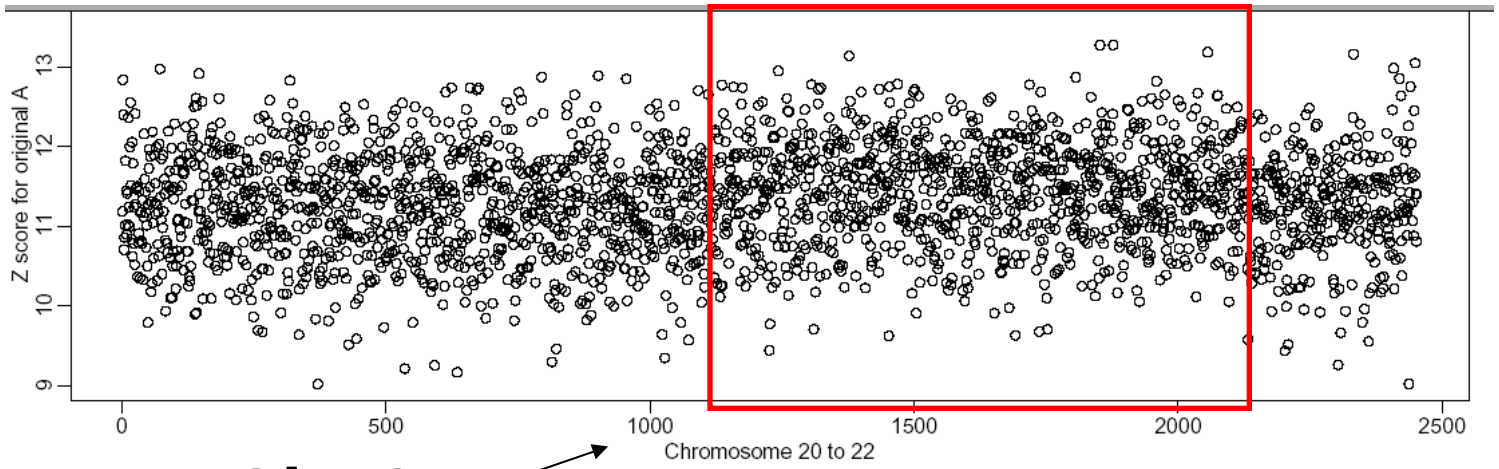


Copy Number

Probe effect



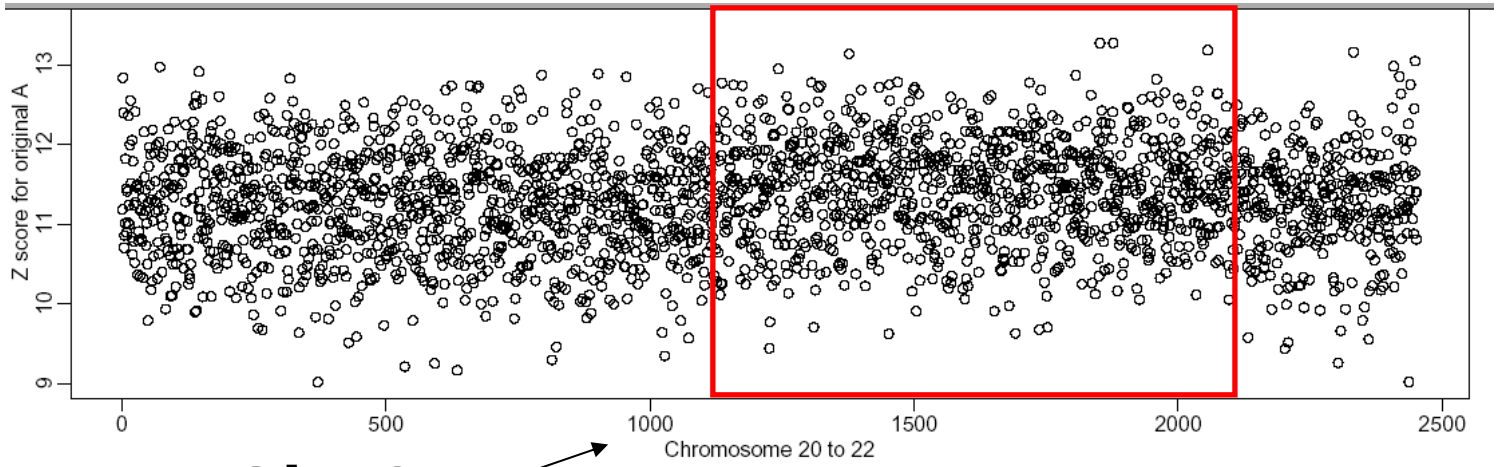
Copy Number



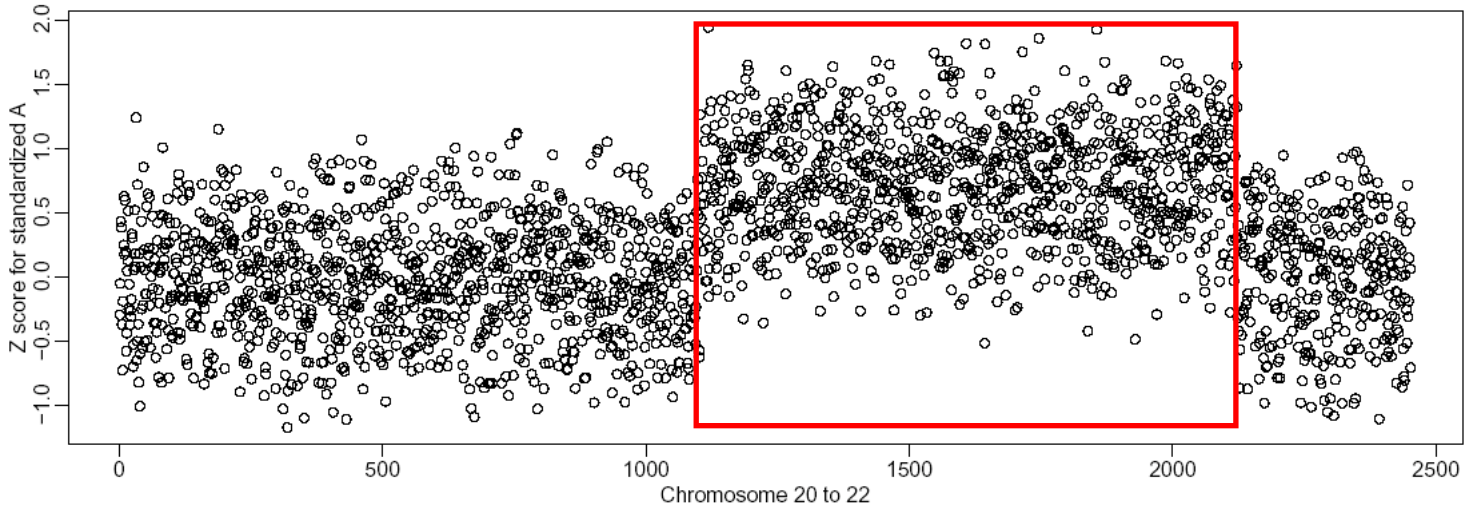
Chr 21

Now we want absolutes:
Probe effect a problem!

Copy Number



Chr 21



Statistical Problem

- A first step is to summarize probe intensities into single point estimates
- Regional (contiguous-point) copy number estimation
- Comparison across individuals

Model for Microarray Data

With expression arrays we see:

- Probe specific additive background noise
- Multiplicative probe effect
- Multiplicative measurement error

Wu et al., JASA (2004)

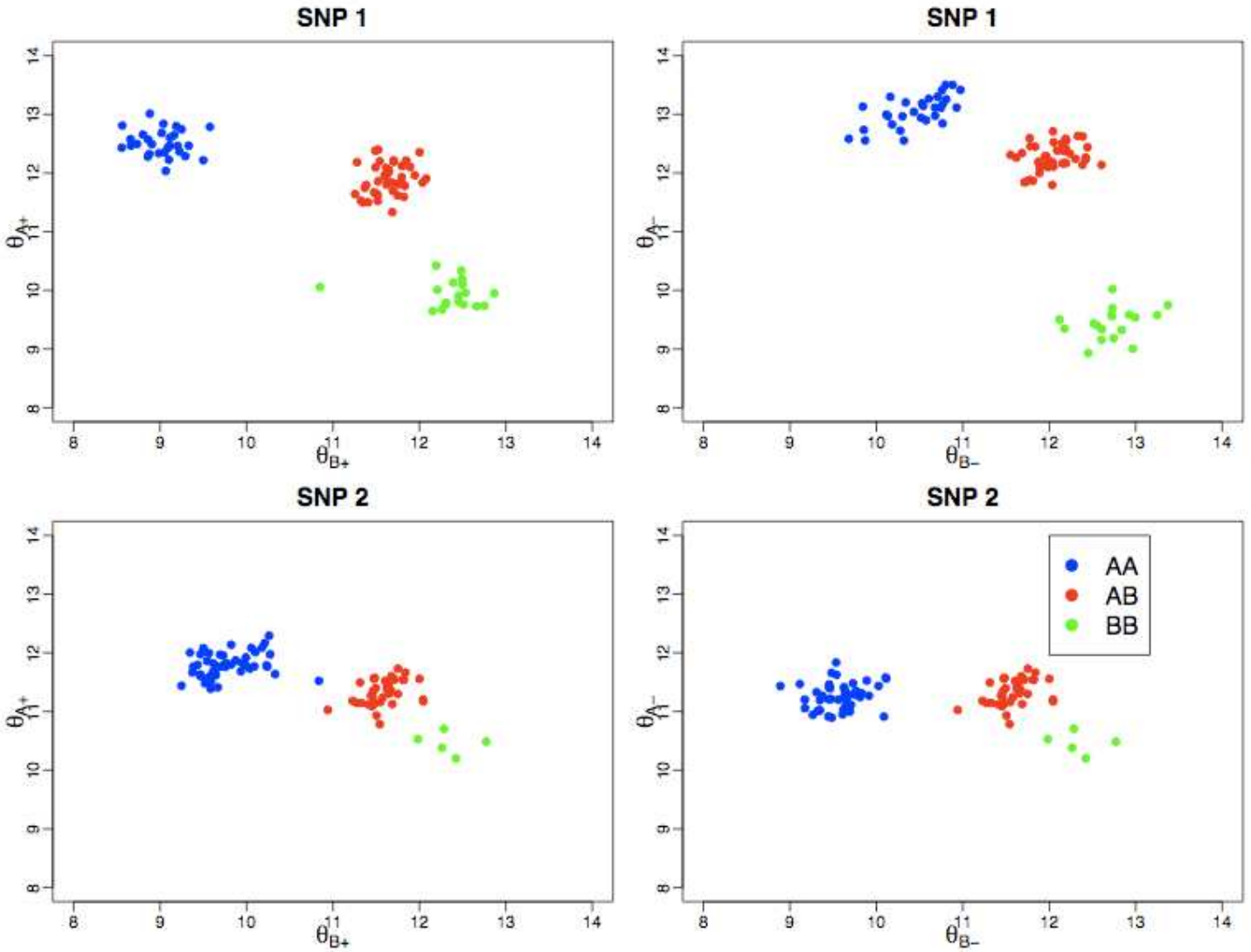
Model adapted for copy number applications:

$$I_{p,j} = \beta_p + C_p \exp(\phi_p + \varepsilon_{p,j})$$

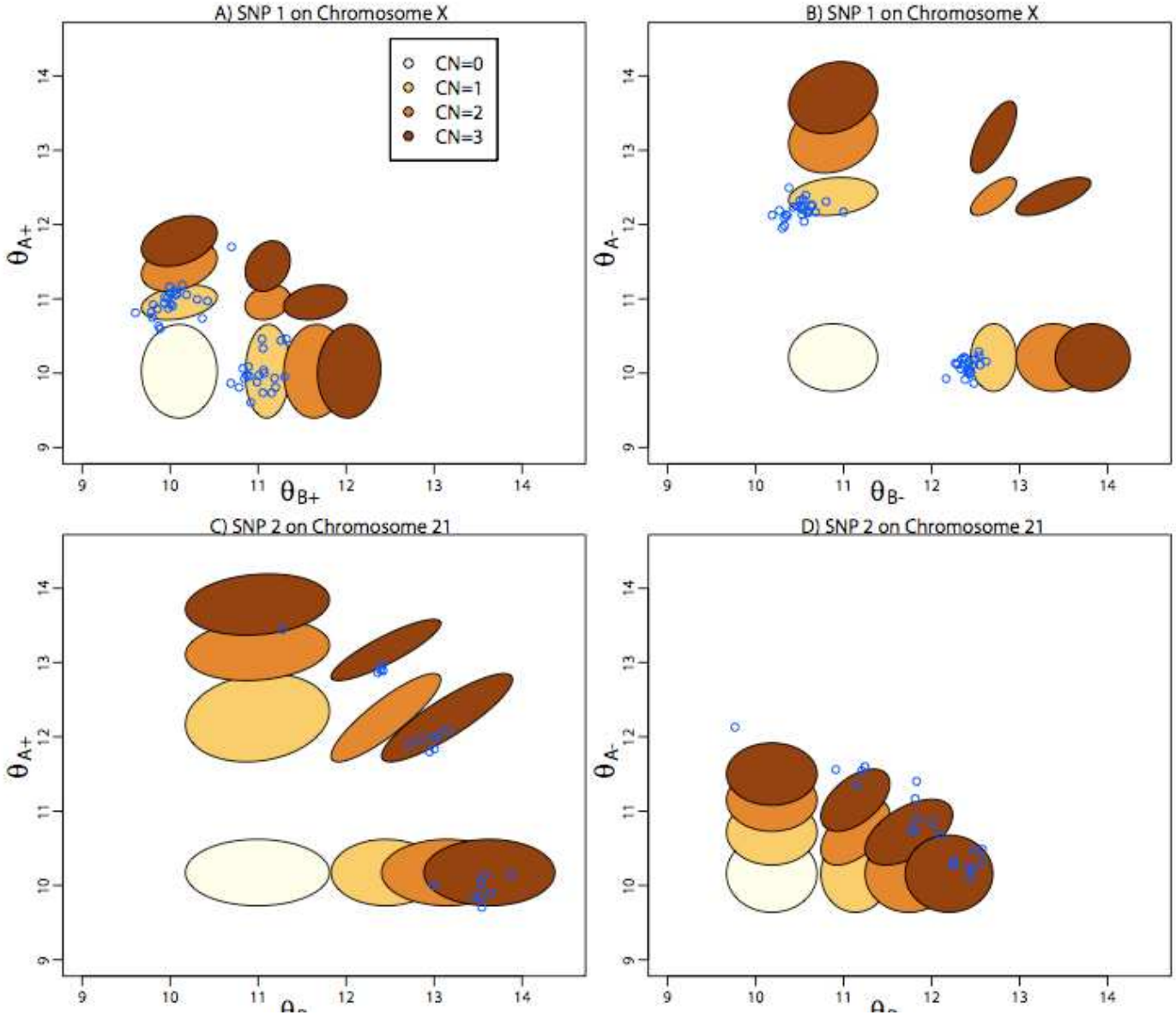
Some Current Approaches

- CNAT: Huang et al. Human Genomics (2004)
- CGAG: Nannya et al. Cancer Research (2005)
- GIM: Ishiwaka et al. Biochem Biophys Res Commun (2005)
- PLASQ: Laframboise et al. Biostatistics (2006)
- CARAT: Huang et al. BMC Bioinformatics (2006)

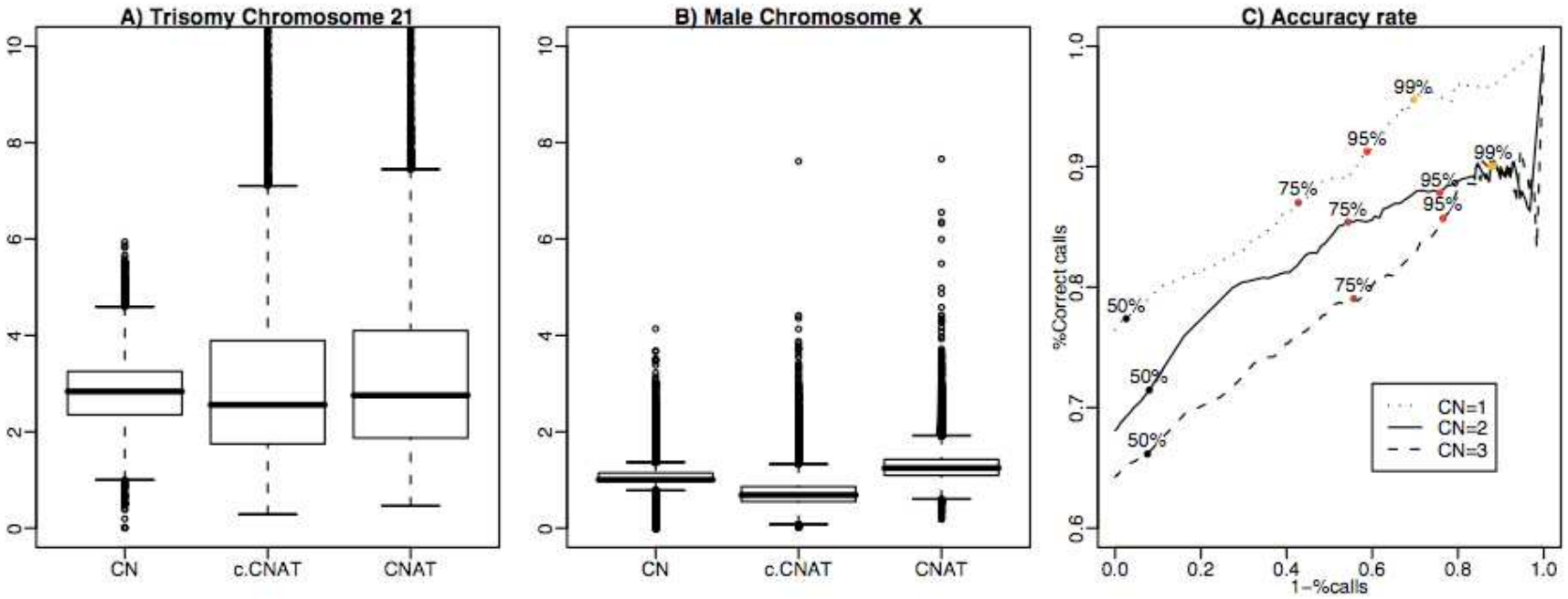
We use genotype calls?



Example: Mixture Models



Results



MSE	CN	c.CNAT	CNAT
CN= 3	0.66	3.68	3.55
CN= 1	0.10	0.16	0.19

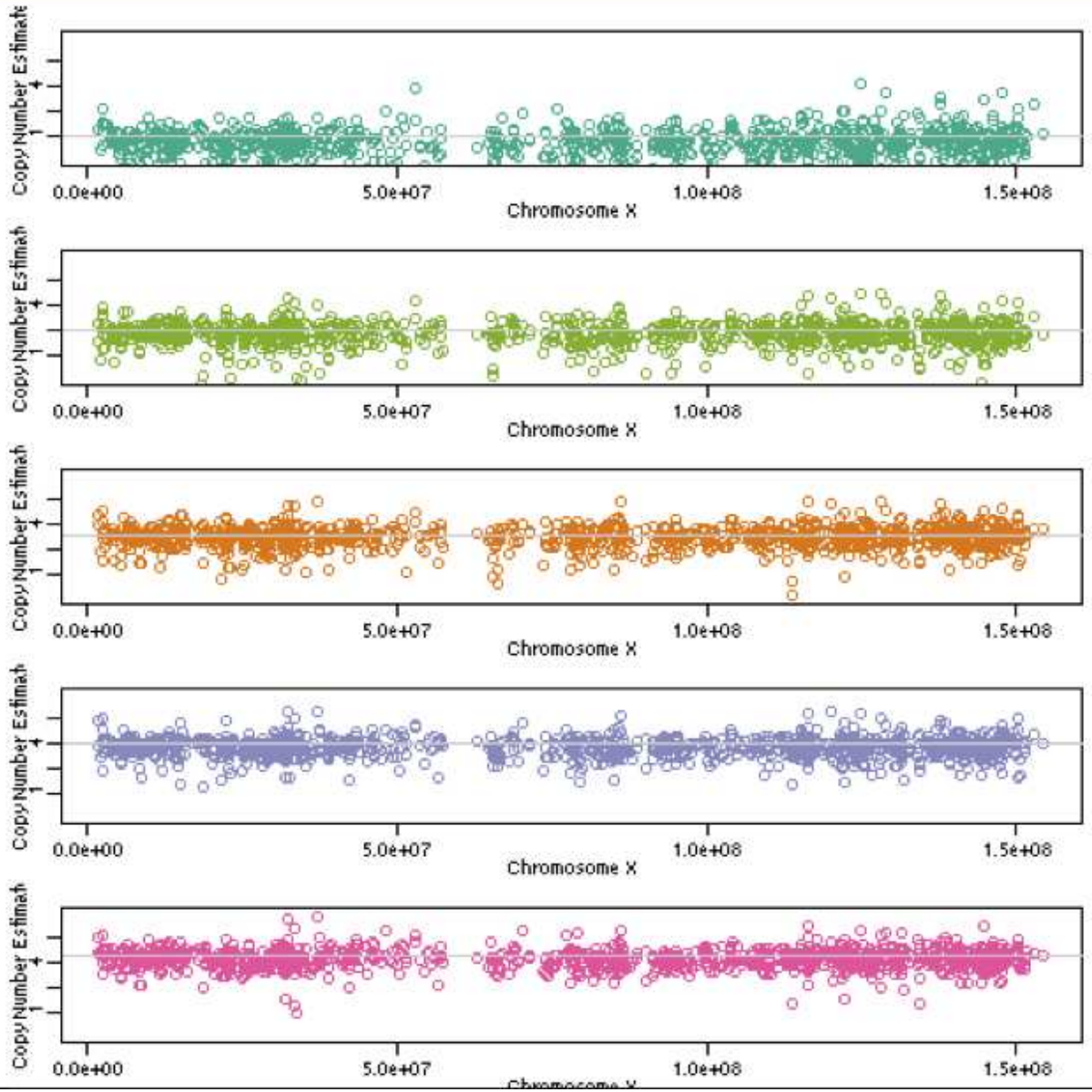
Inverse Regression

- Our model – SNP i, probe p

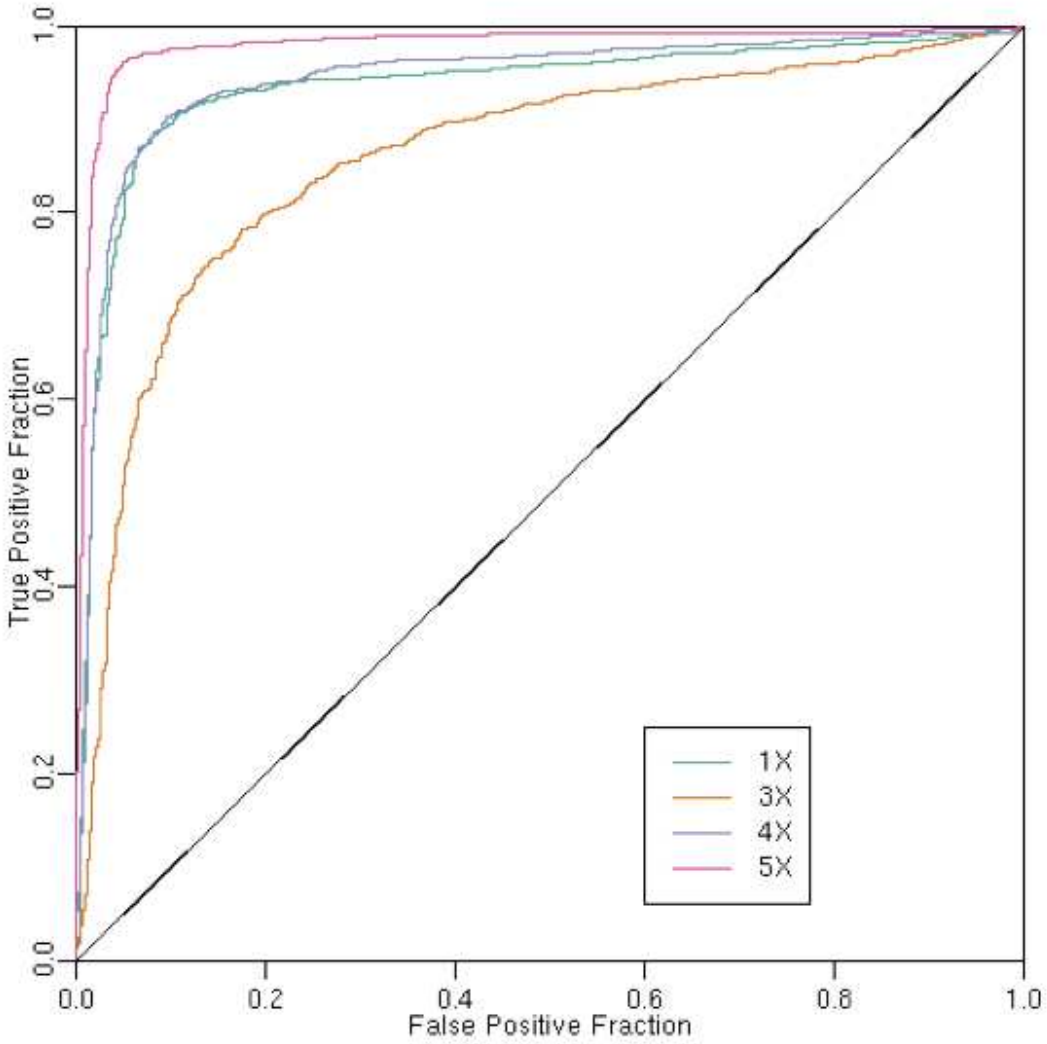
$$\begin{aligned} I_{A,j} &= \beta_A + \varepsilon_{1,A,j} + C_{A,j} \exp(\phi_A + \varepsilon_{2,A,j}) \\ I_{B,j} &= \beta_B + \varepsilon_{1,B,j} + C_{B,j} \exp(\phi_B + \varepsilon_{2,B,j}) \end{aligned}$$

- If $\log(I - \beta) < 1.96 * \sigma_{\varepsilon_1}$, then $C = 0$
- Else, $\log(C) = \log(I - \beta) - \phi \pm 1.96 * \sigma_{\varepsilon_2}$

Results – Chromosome X 1 – 5

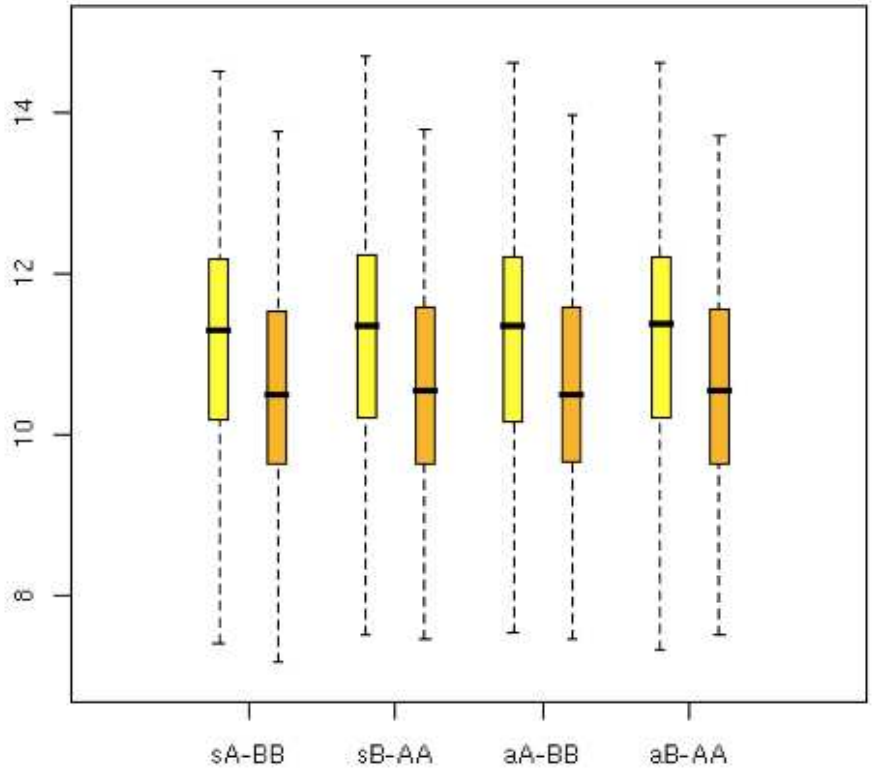


Results – Chromosome X 1 – 5



Cross-hybridizaton

- Assumption
  - Parameters for allele A only depend on C_A except for the correlation, and vice versa



Where is the gratuitous picture of Wolfgang?