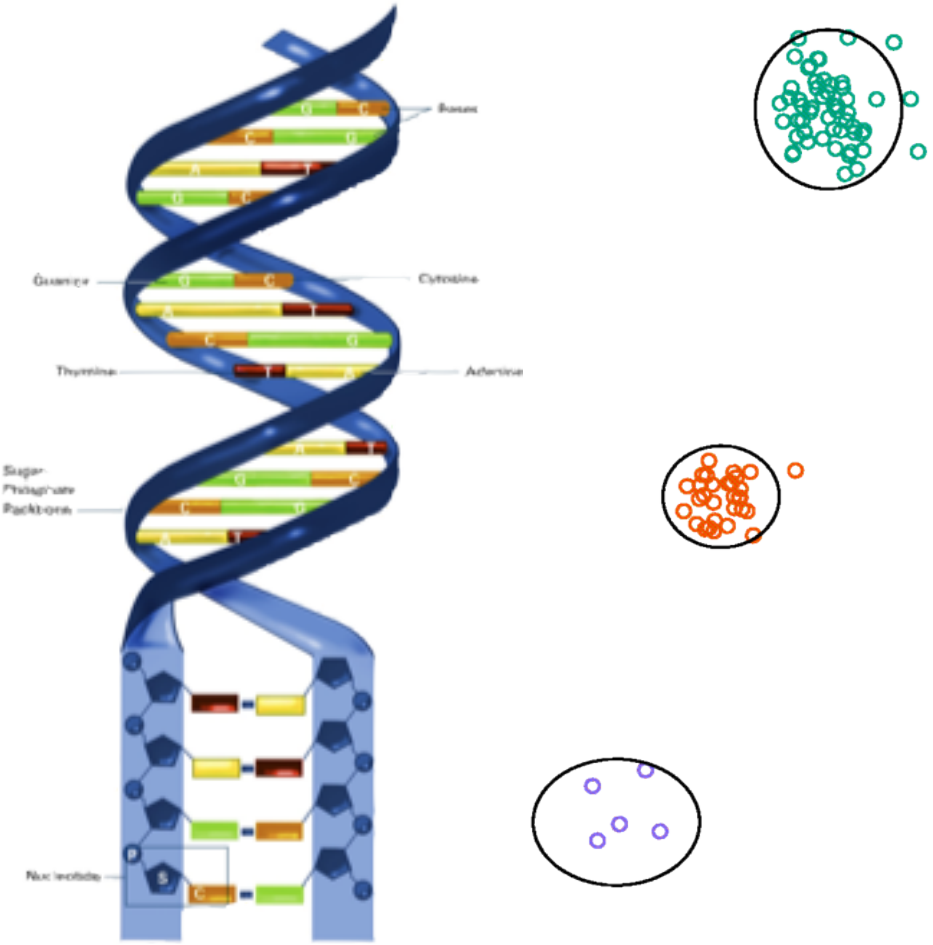
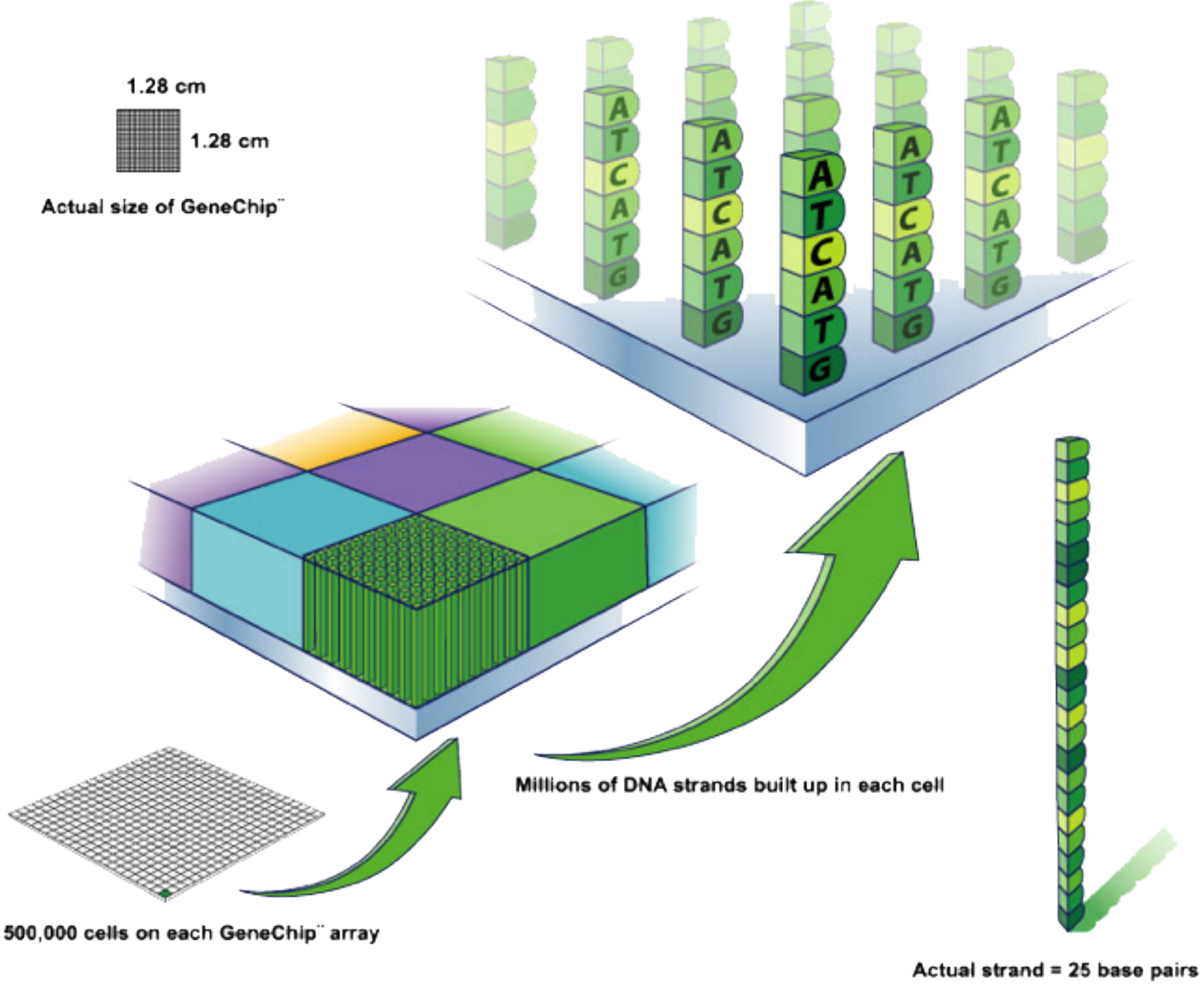
UNIVERSITY OF
CAMBRIDGE



SNP Analysis

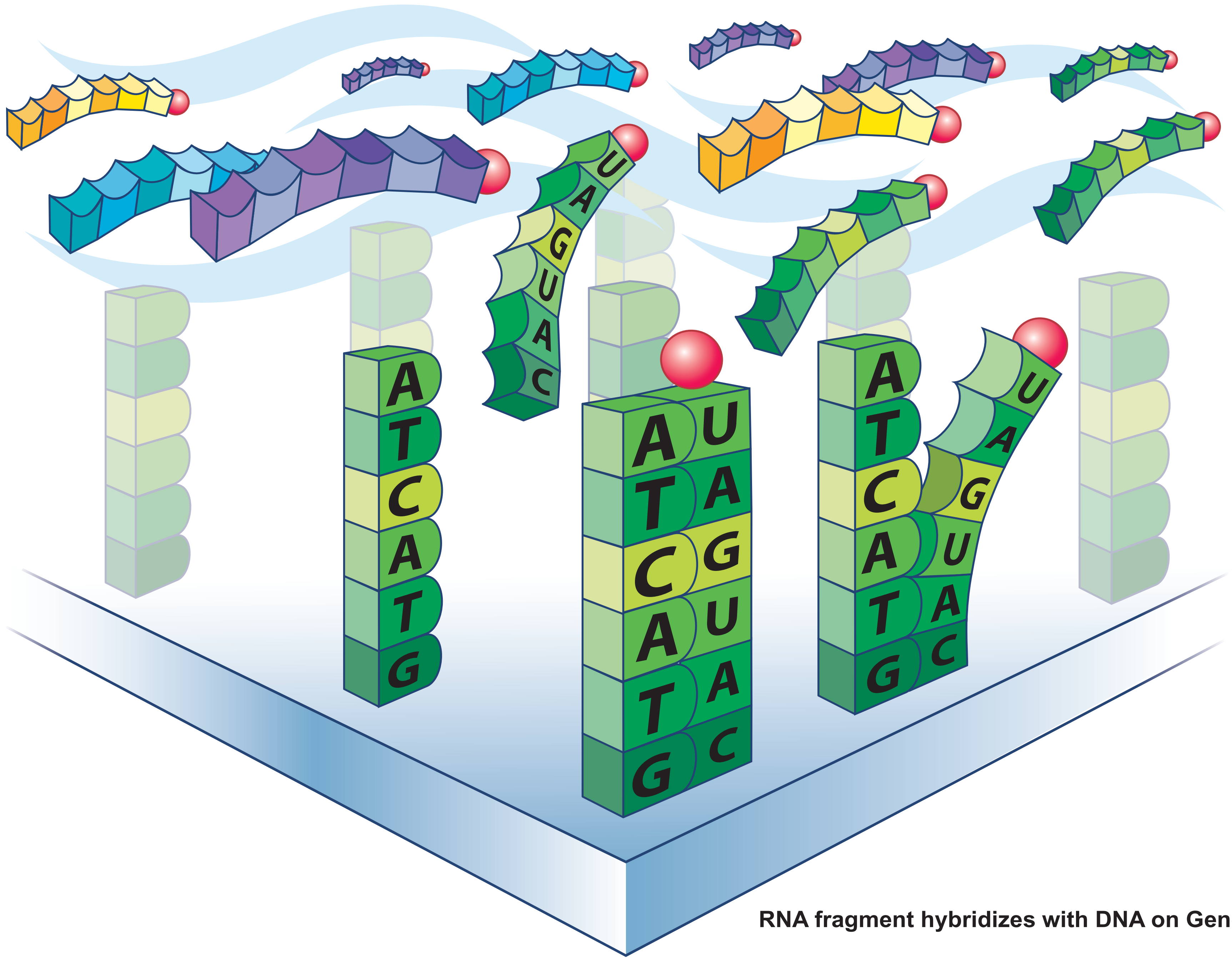
Benilton S Carvalho

Array Design


1.28 cm
1.28 cm
Actual size of GeneChip¨
A
T
C
A
T
G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip¨ array
Actual strand = 25 base pairs

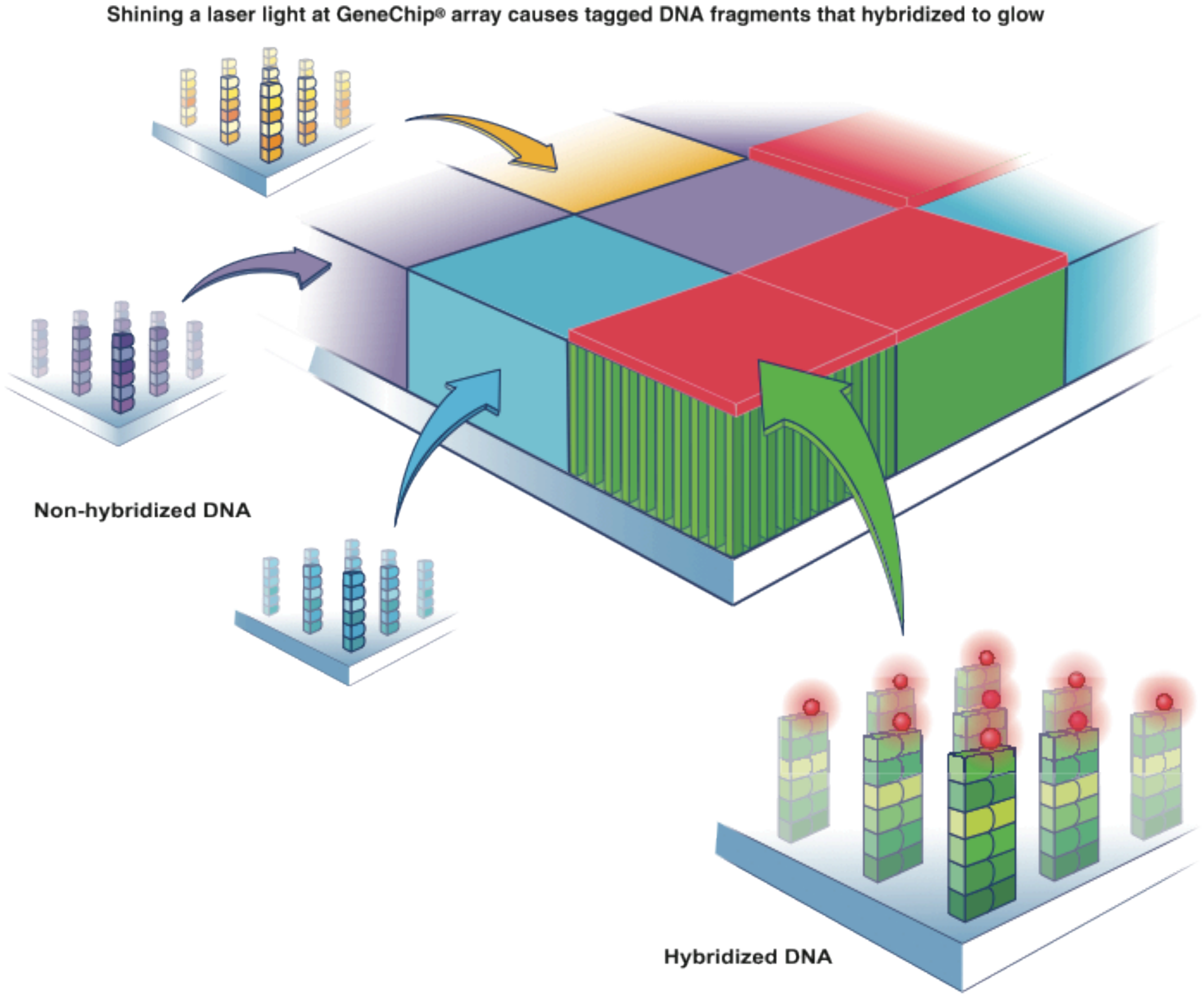
Hybridization

RNA fragments with fluorescent tags from sample to be tested

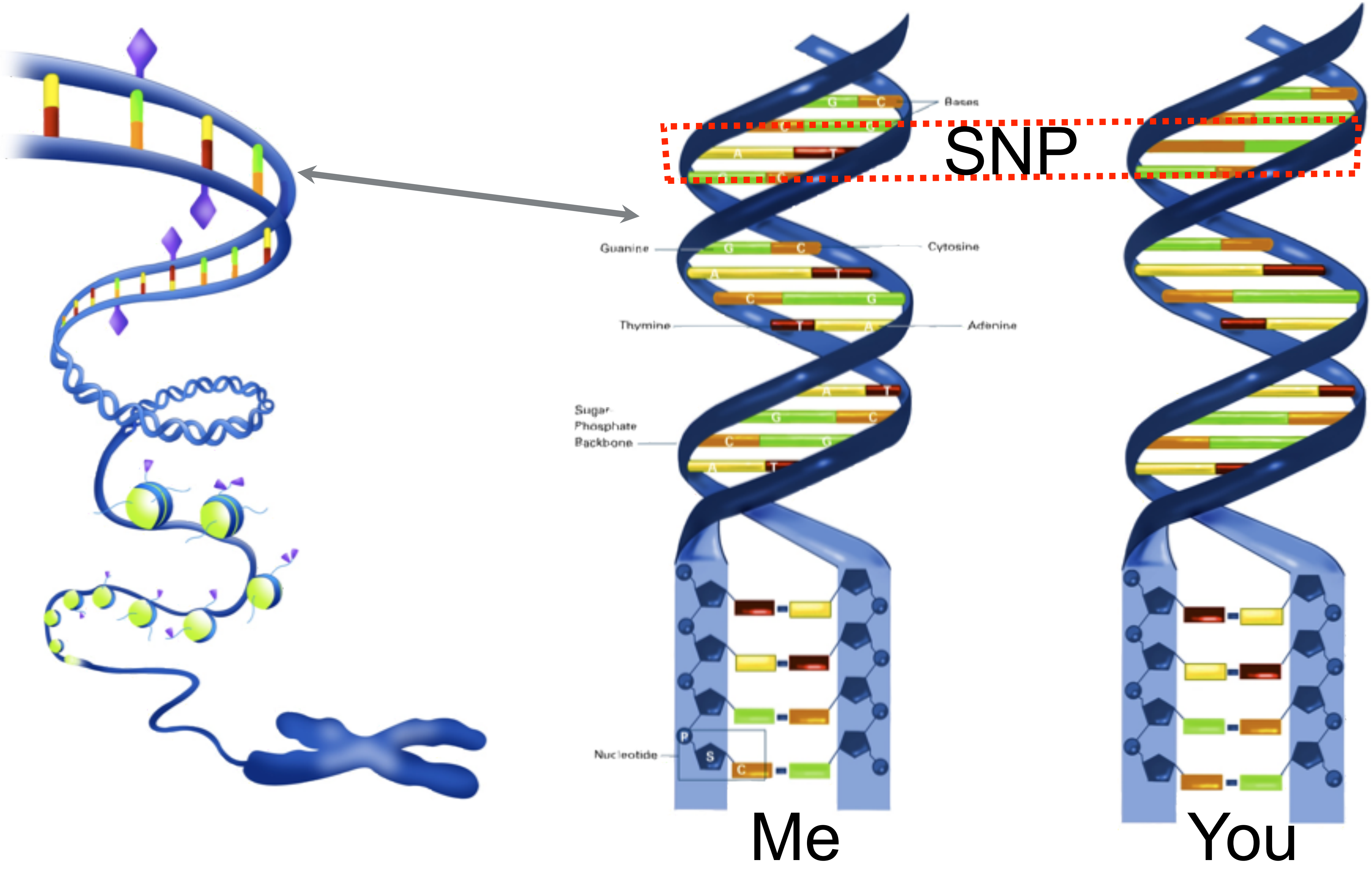


RNA fragment hybridizes with DNA on GeneChip® array

Scanning


Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

What is a SNP?



SNP

Me

You

Courtesy: NIH

SNP’s

- DNA sequence variations;
- Prevalence > 1%;
- Responsible for ~90% of all genetic variation;
- In average, at every 100-300 bp;

SNP’s and Disease Associations

- Yasuda *et. al.* show that each allele C in rs2237892 increases the odds of type 2 diabetes by 1.40 times compared to TT;
- Ferreira *et. al.* show that each T in rs4948418 increases the odds of bipolar disease by 1.45 times compared to CC;

- Nature Genetics, Aug 17, 2008.

Genotype accurately at high-density using oligonucleotide microarrays!

Part I
Creating the genotyping algorithm

- Carvalho et al. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics (2007) vol. 8 (2) pp. 485-99
- Lin et al. Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays. Genome Biol (2008) vol. 9 (4) pp. R63

SNP Chip

- Genomic unit of interest: SNP;
- Intensities are observed for a list of SNP’s on both alleles (A and B) often on two directions (sense and antisense);

$$\begin{aligned} M &= \log \frac{\theta_A}{\theta_B} \\ &= \log \theta_A - \log \theta_B \\ S &= \frac{\log \theta_A + \log \theta_B}{2} \end{aligned}$$

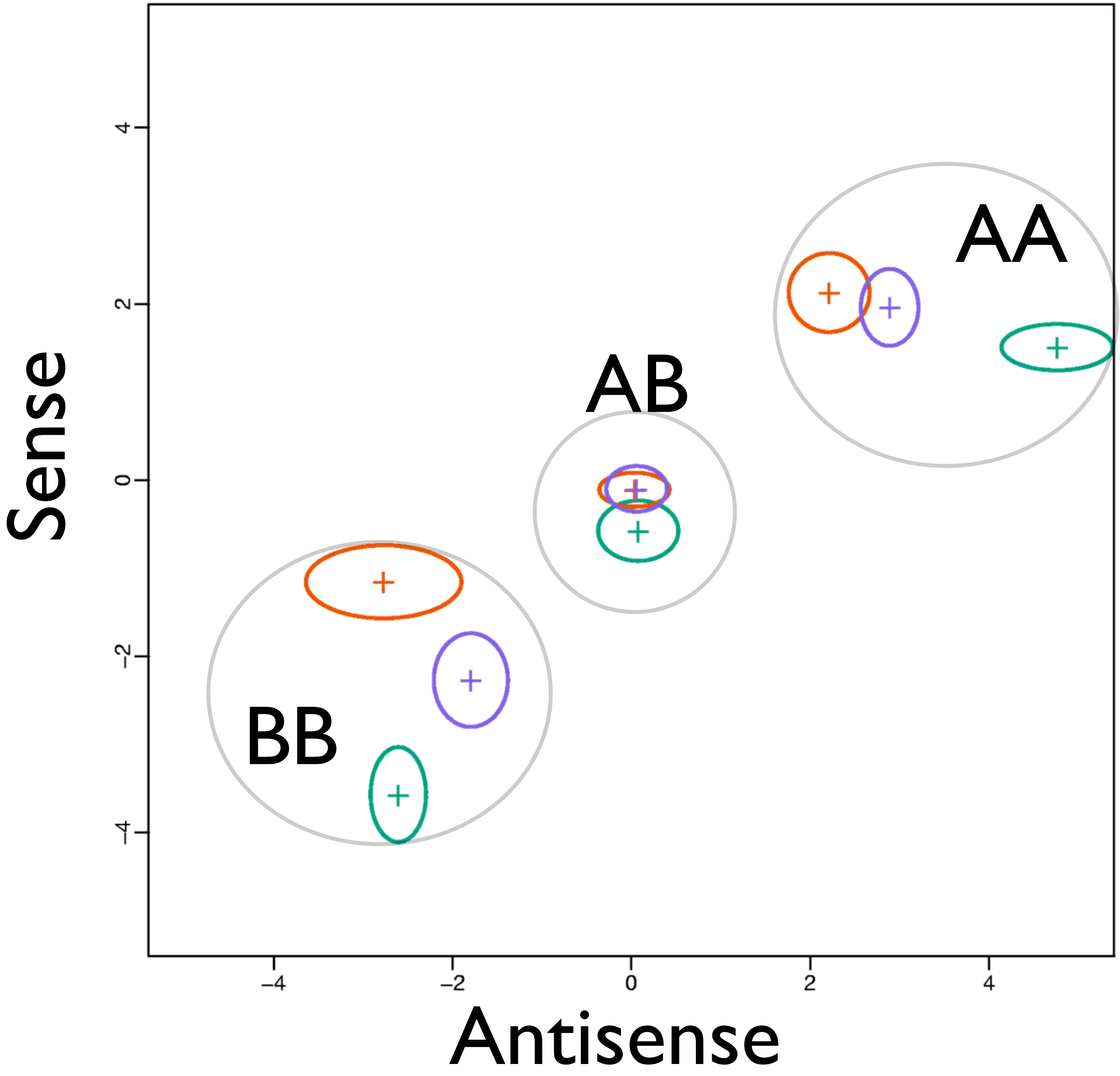
Naïve Genotyping with Log-Ratio

$$\begin{aligned} M &= \log\theta_A - \log\theta_B \\ M &> K \rightarrow \mathrm{AA} \\ -K \leq M &\leq K \rightarrow \mathrm{AB} \\ M &< -K \rightarrow \mathrm{BB} \end{aligned}$$

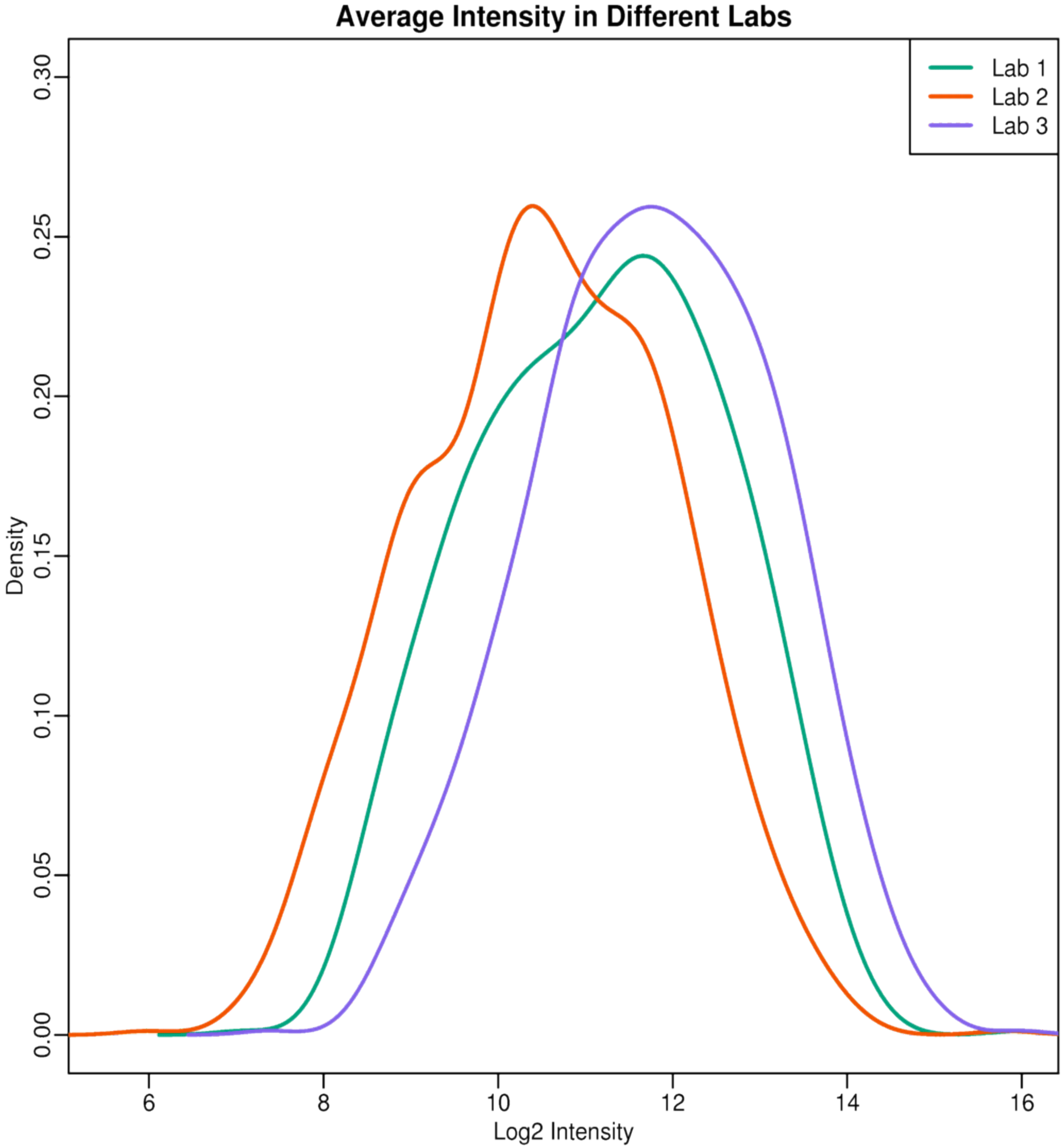
HapMap Dataset

- 270 subjects (1000+ on Phase 3);
- Different ethnicities (eg. CEU, CHB, JPT, YRI);
- Gold-standard genotypes publicly available;
- Samples available in different SNP platforms;

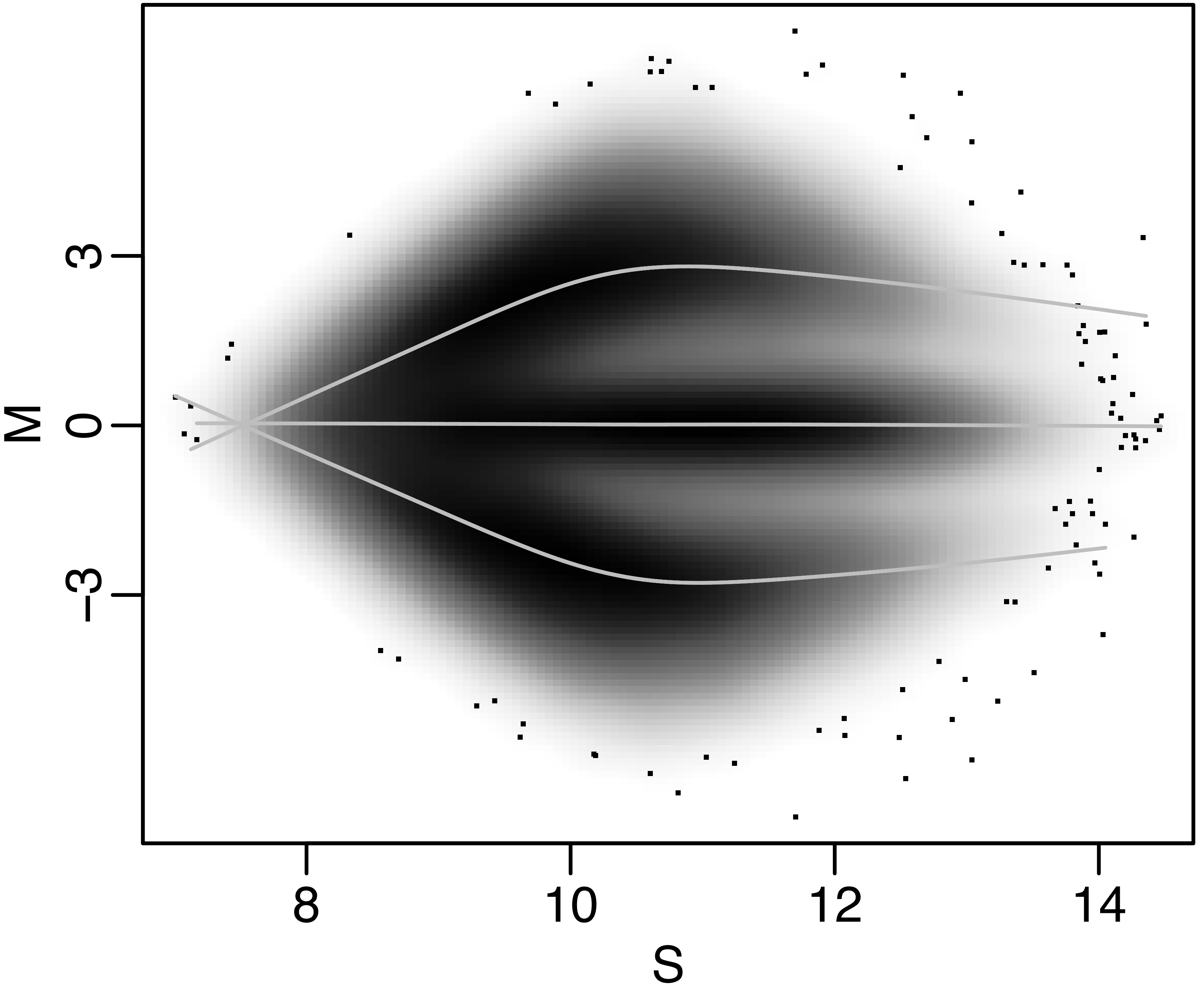
Learning from HapMap



Different Distributions

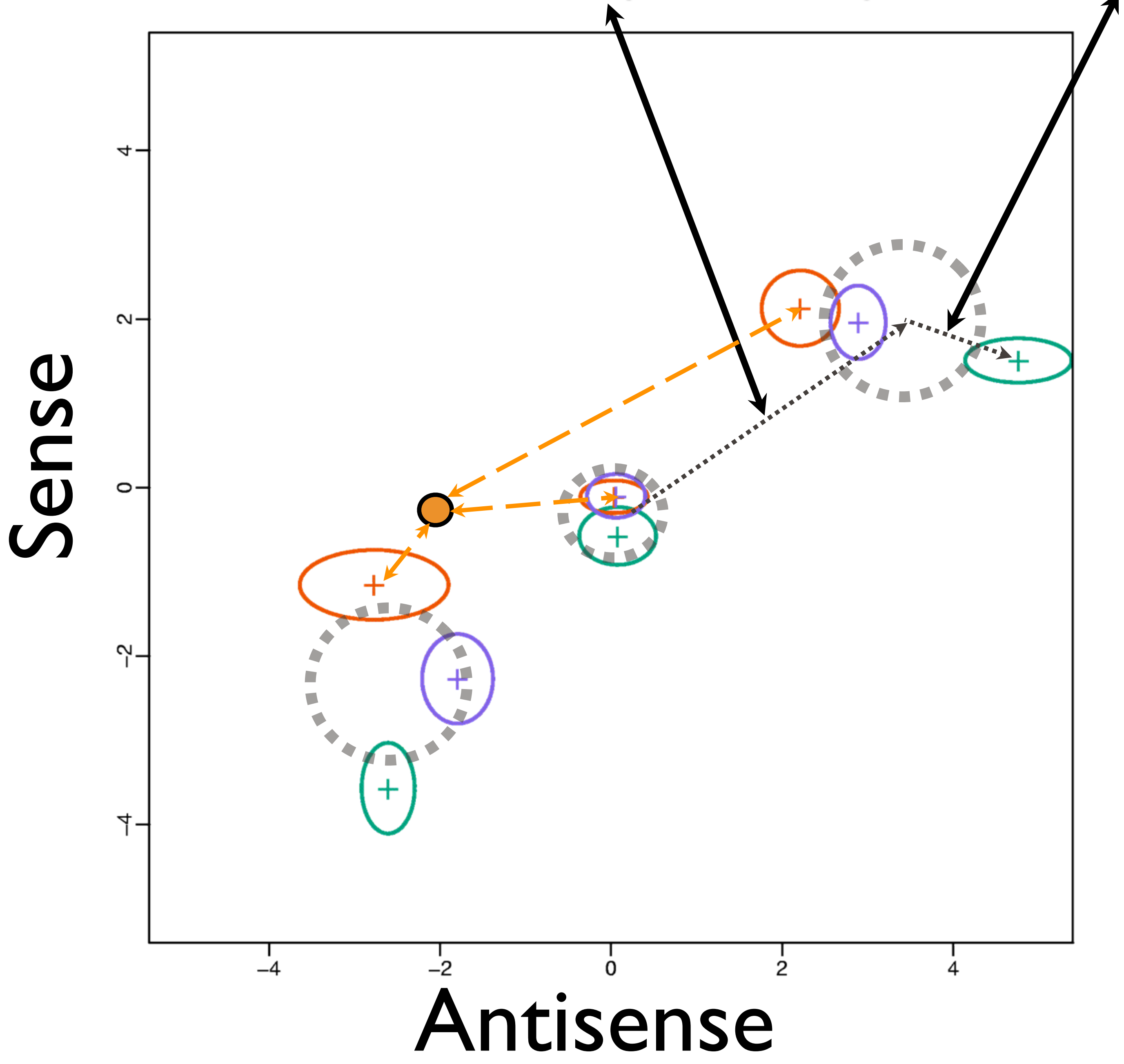


Log-ratio and Strength

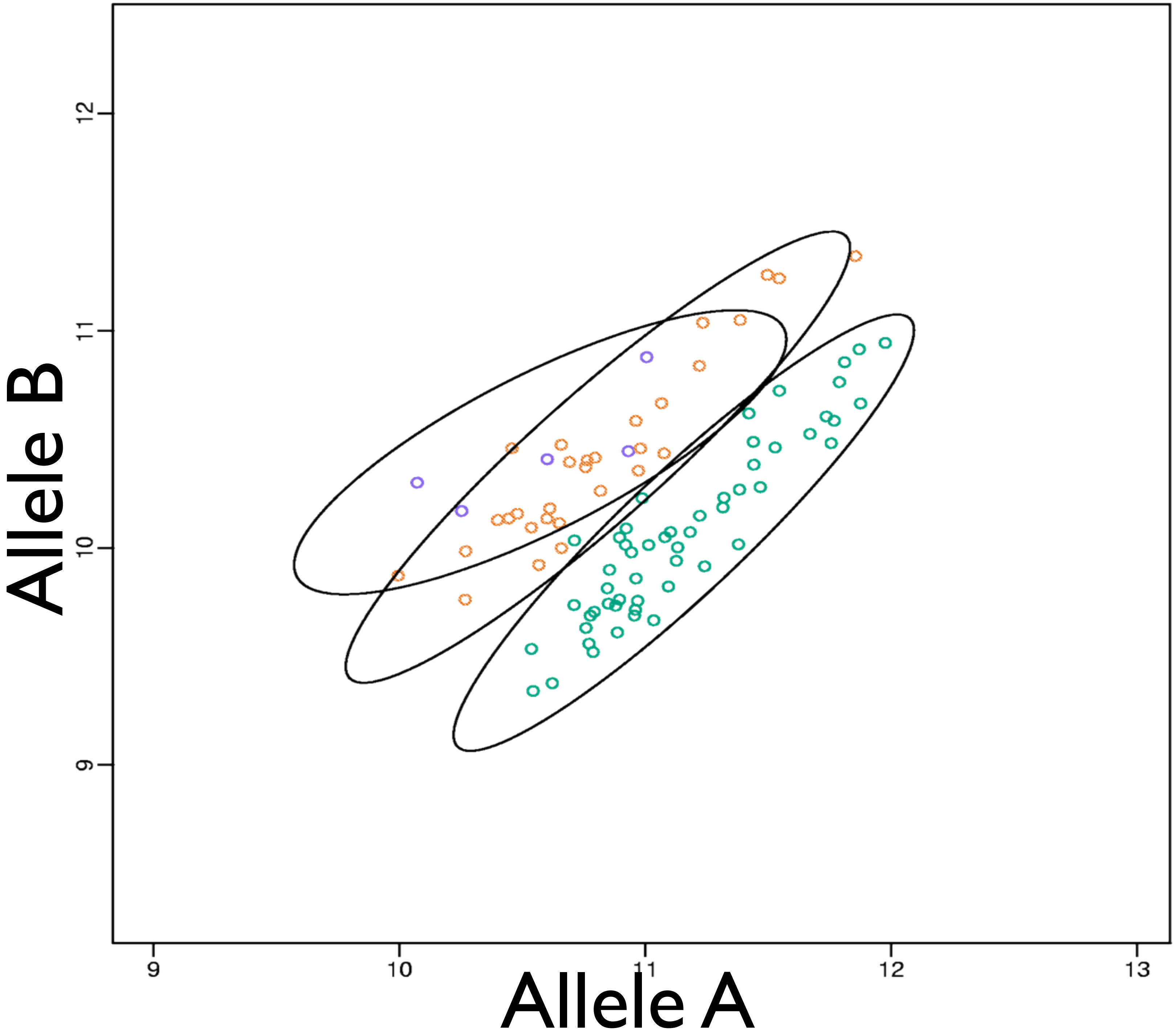


Model Used by CRLMM

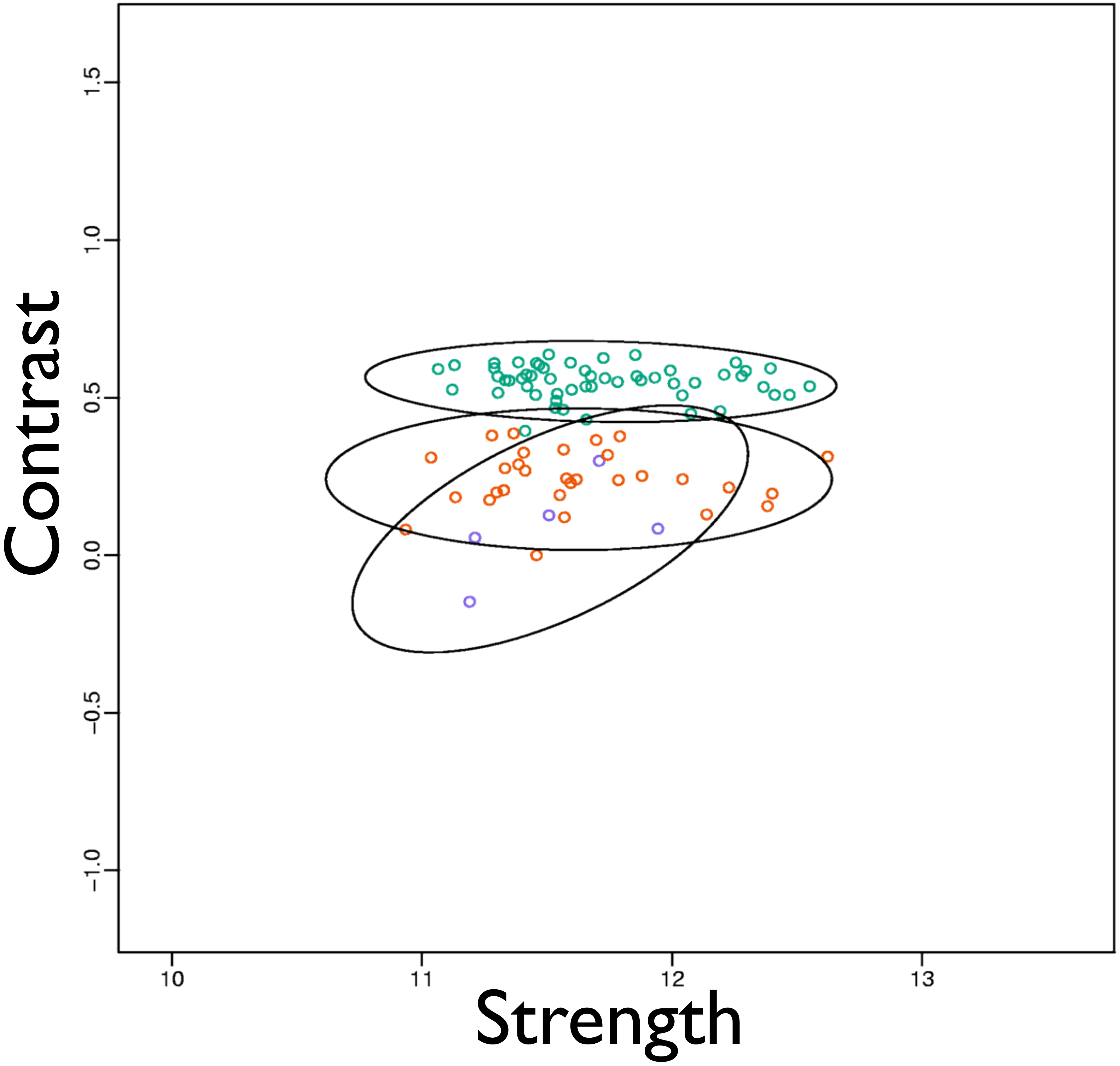
$$[M_{i,j,s}|Z_{i,j} = k, m_{i,k,s}] = f_{j,k}(\mathbf{X}_{i,j,s}) + m_{i,k,s} + \epsilon_{i,j,k,s}$$



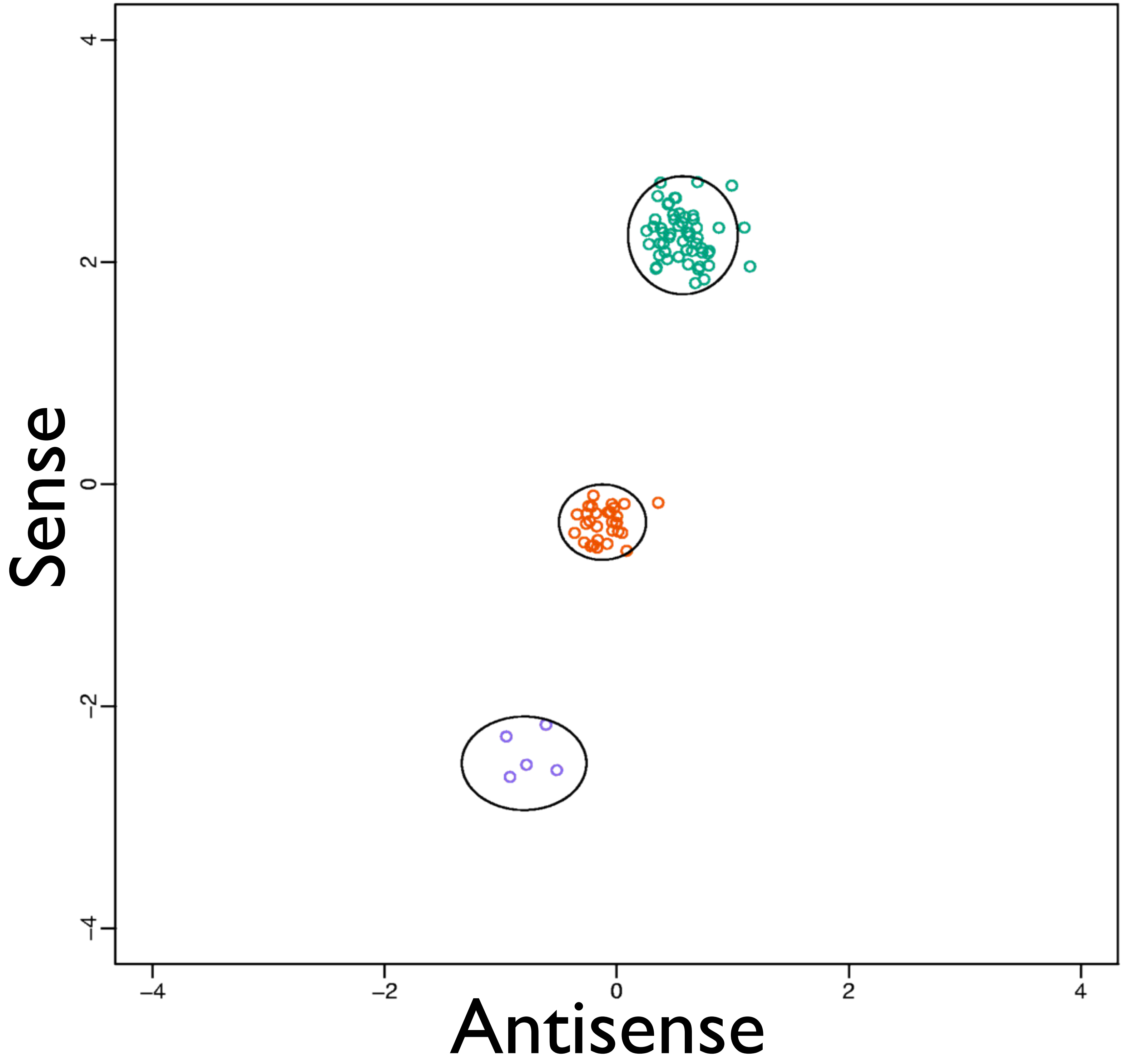
RLMM and Strand Discrimination



BRLMM and Strand Discrimination



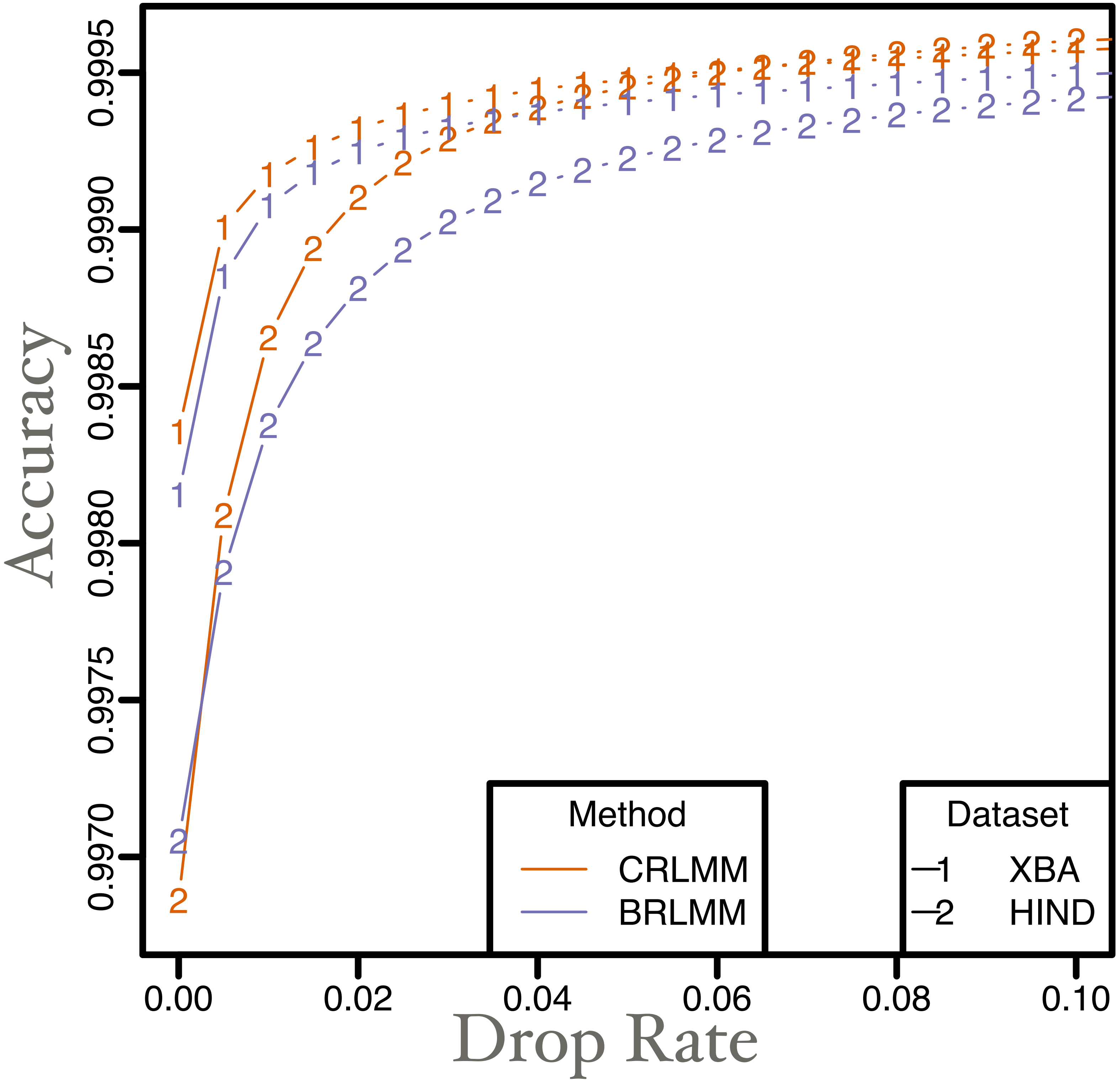
CRLMM and Strand Discrimination



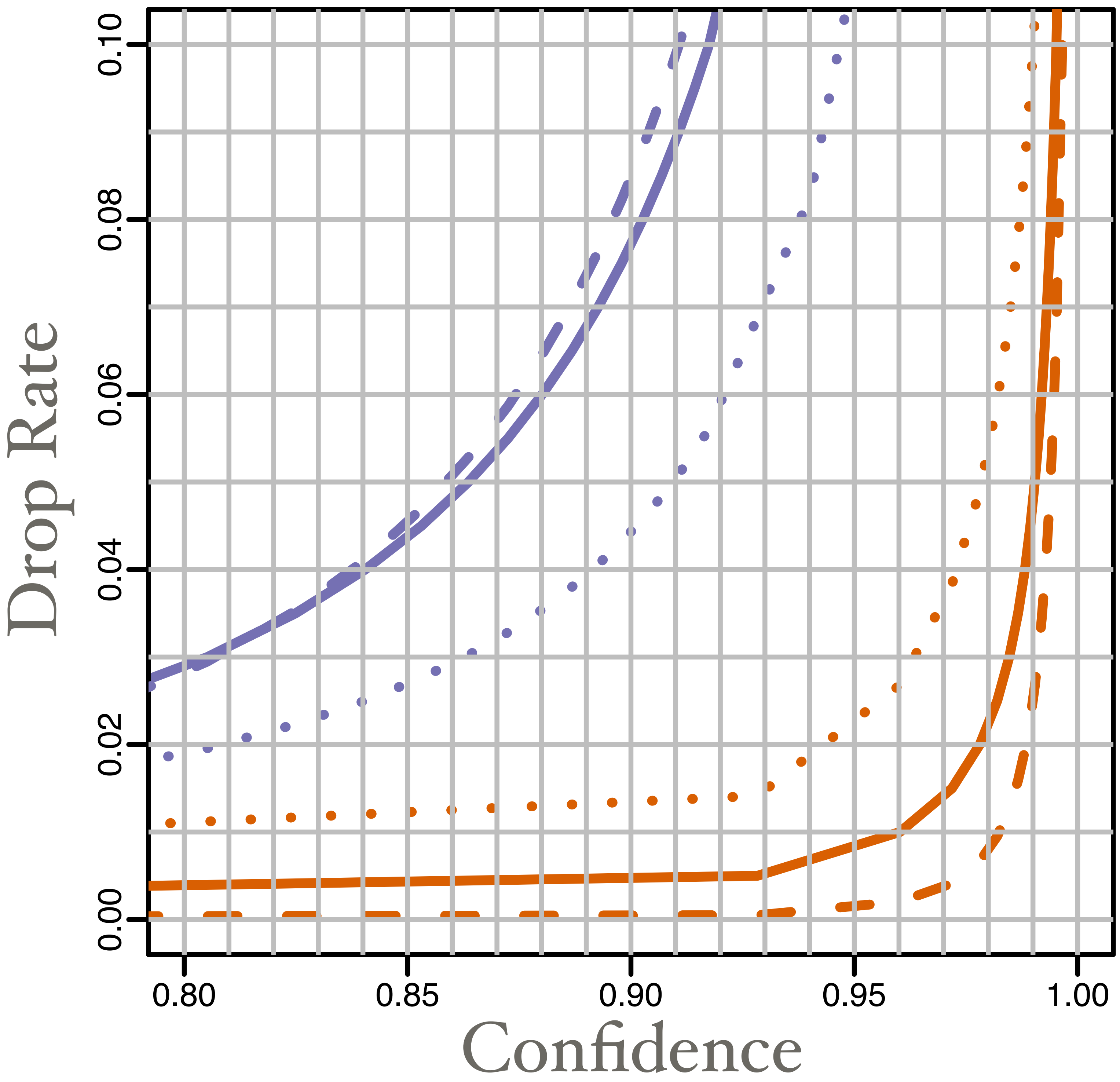
Assessment of a Genotyping Algorithm

- Accuracy - algorithm’s calls compared to the gold-standard calls;
- Observed drop-rate when filtering;
- Confidence scores not associated to genotypes;

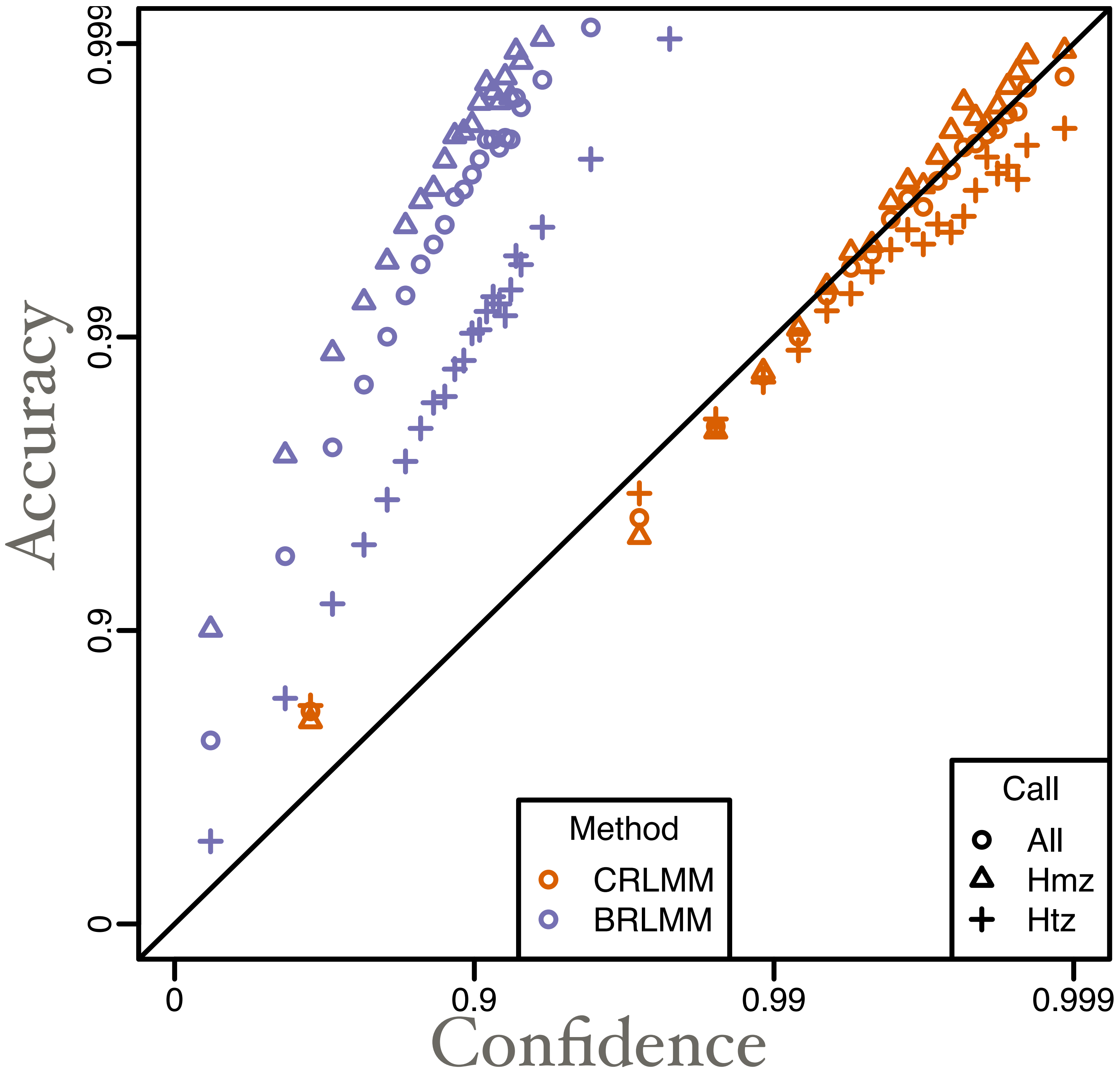
Accuracy vs. Drop Rate



Drop Rate vs. Confidence



Accuracy vs. Confidence



Summary

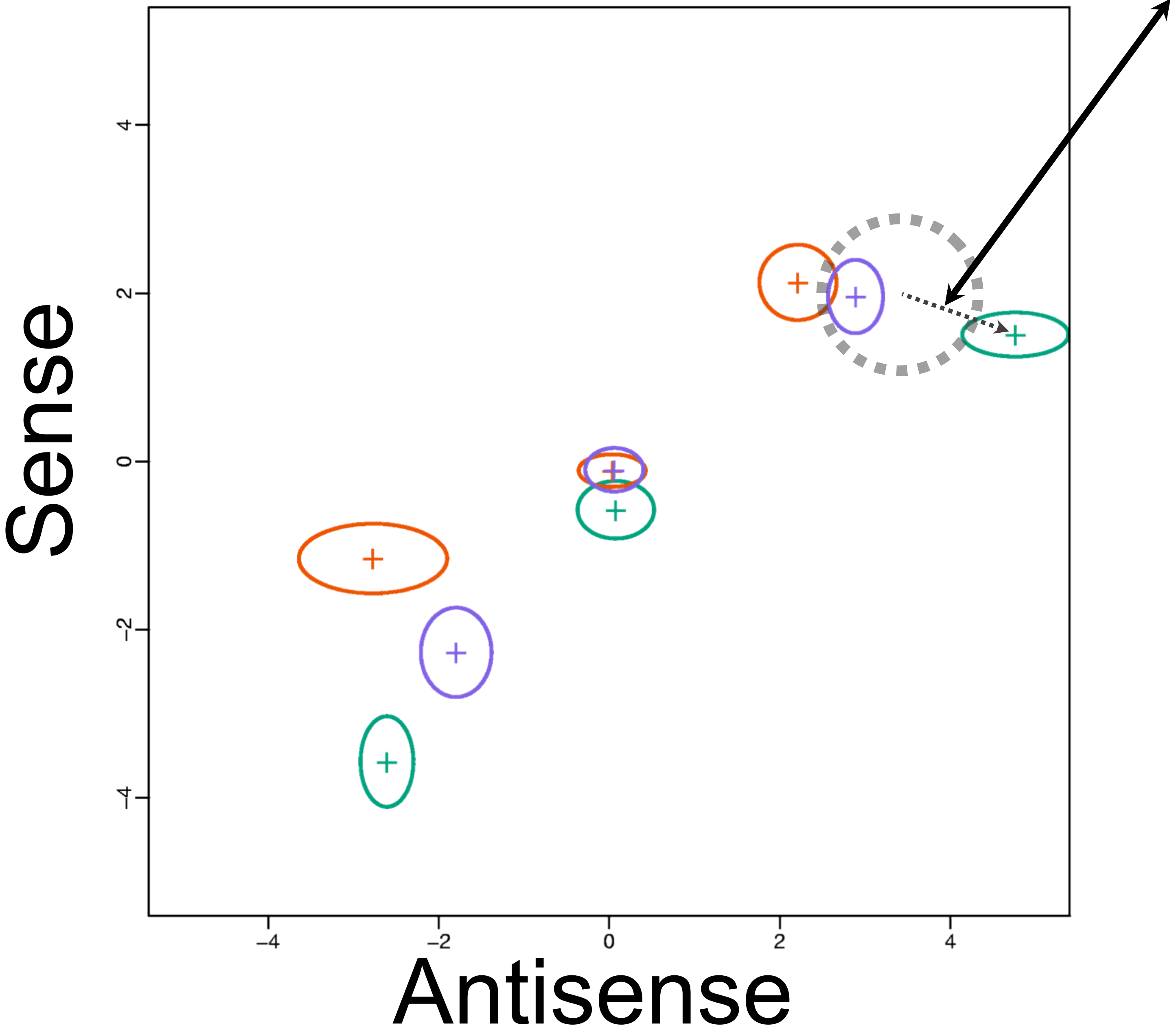
- CRLMM is a genotyping algorithm for the Affymetrix platform (50K, 250K, 500K and 1M);
- Outperforms standard tools:
  - HapMap: 50K, 250K, 1M;
  - Repetition: 250K, 1M;
- Freely available via BioConductor.

Part II
Improving the genotyping algorithm

Carvalho et al. Quantifying uncertainty in genotype calls. Bioinformatics (2010) vol. 26 (2) pp. 242-9

Model Used by CRLMM

$$[M_{i,j,s}|Z_{i,j} = k, m_{i,k,s}] = f_{j,k}(\mathbf{X}_{i,j,s}) + \boxed{m_{i,k,s}} + \epsilon_{i,j,k,s}$$

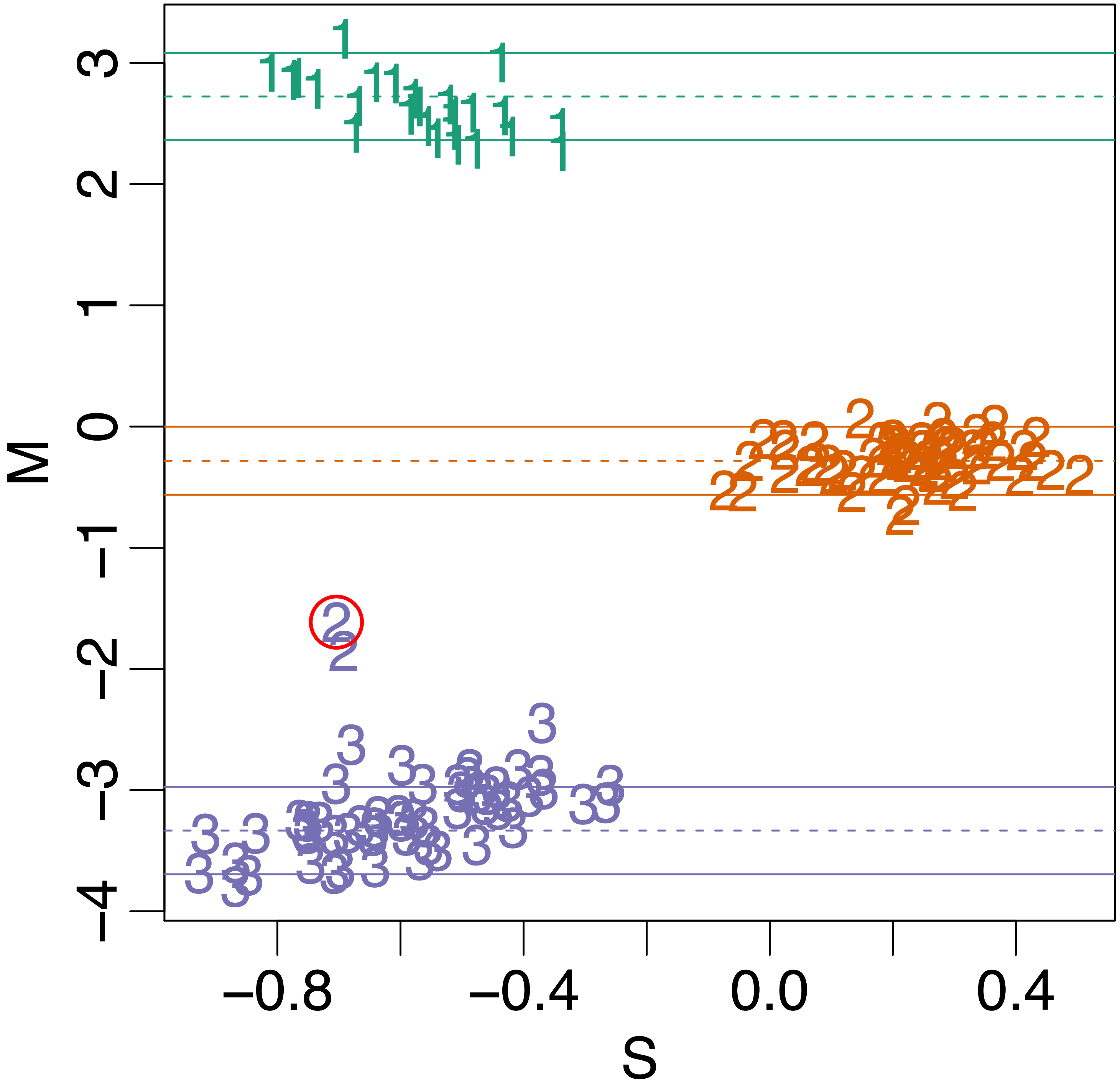


CRLMM Model

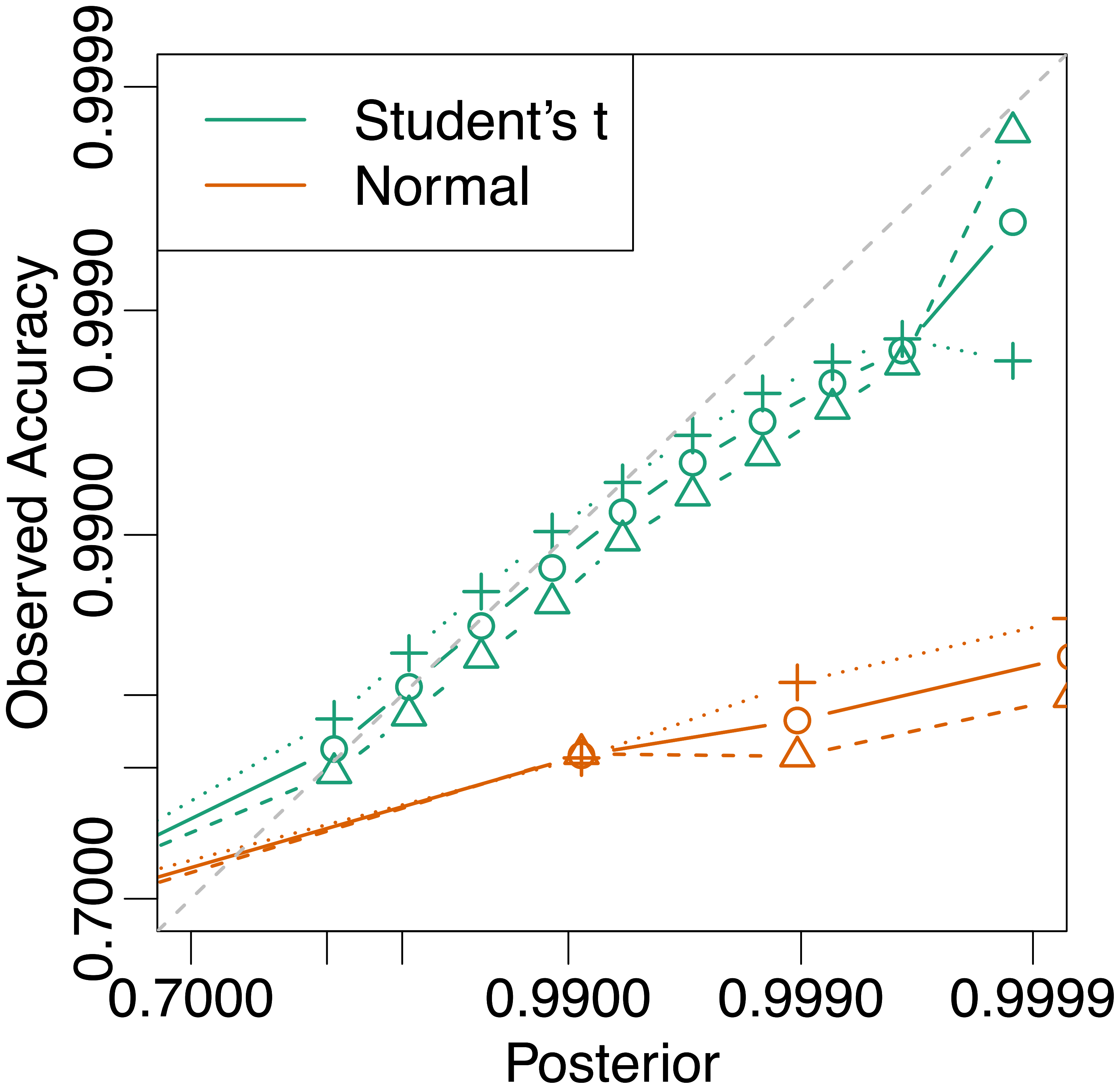
$$[M_{i,j,s} | Z_{i,j} = k, m_{i,k,s}] = f_{j,k}(\mathbf{X}_{i,j,s}) + m_{i,k,s} + \epsilon_{i,j,k,s}$$

- The location parameters for SNP's were estimated from the HapMap dataset;
- Assumes the SNP-specific shift is a fixed effect;
- SNP's with few observations on HapMap should have their confidences penalized somehow, and they don't;
- Error follows a Normal distribution;

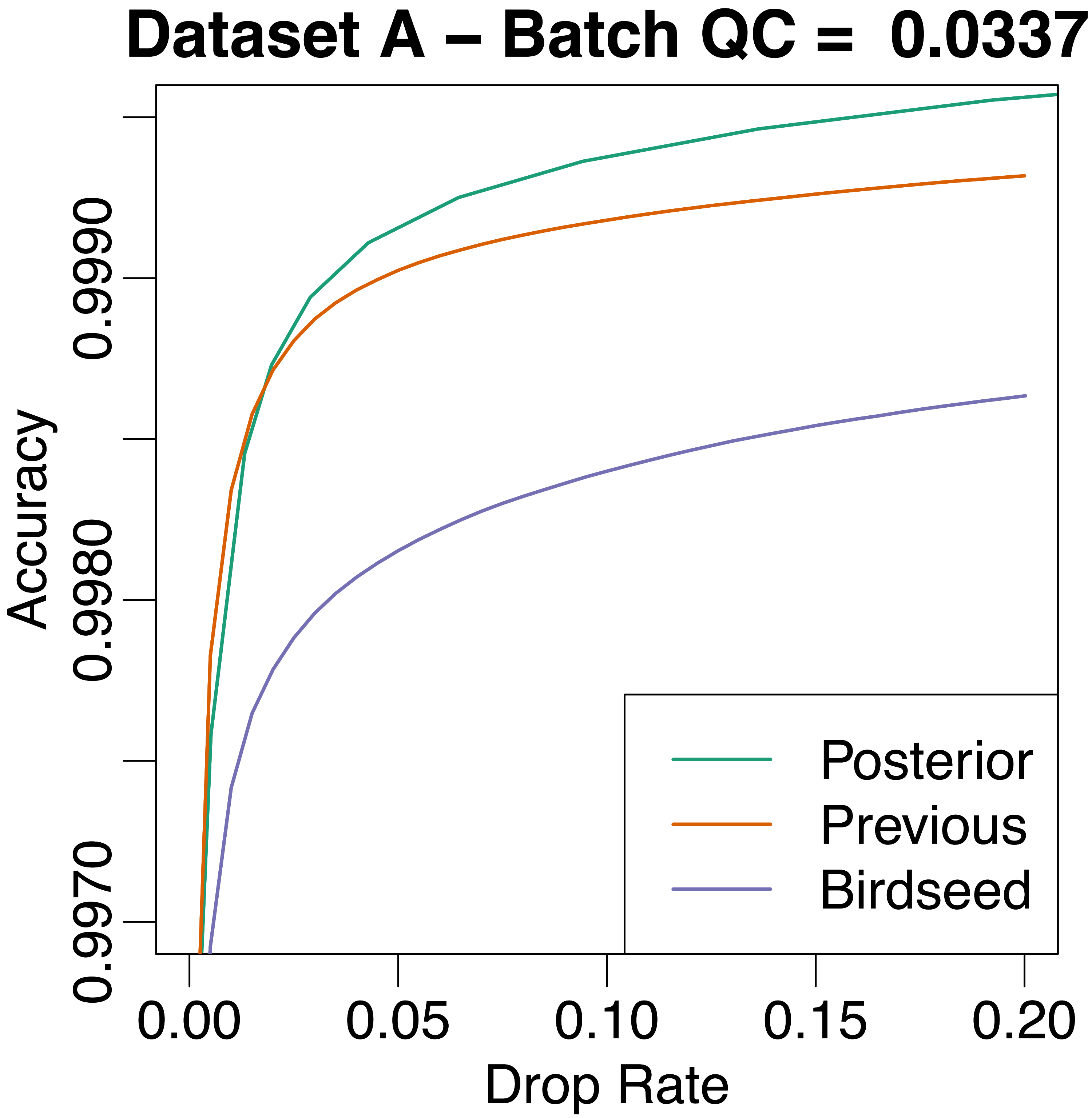
Observed Improvements



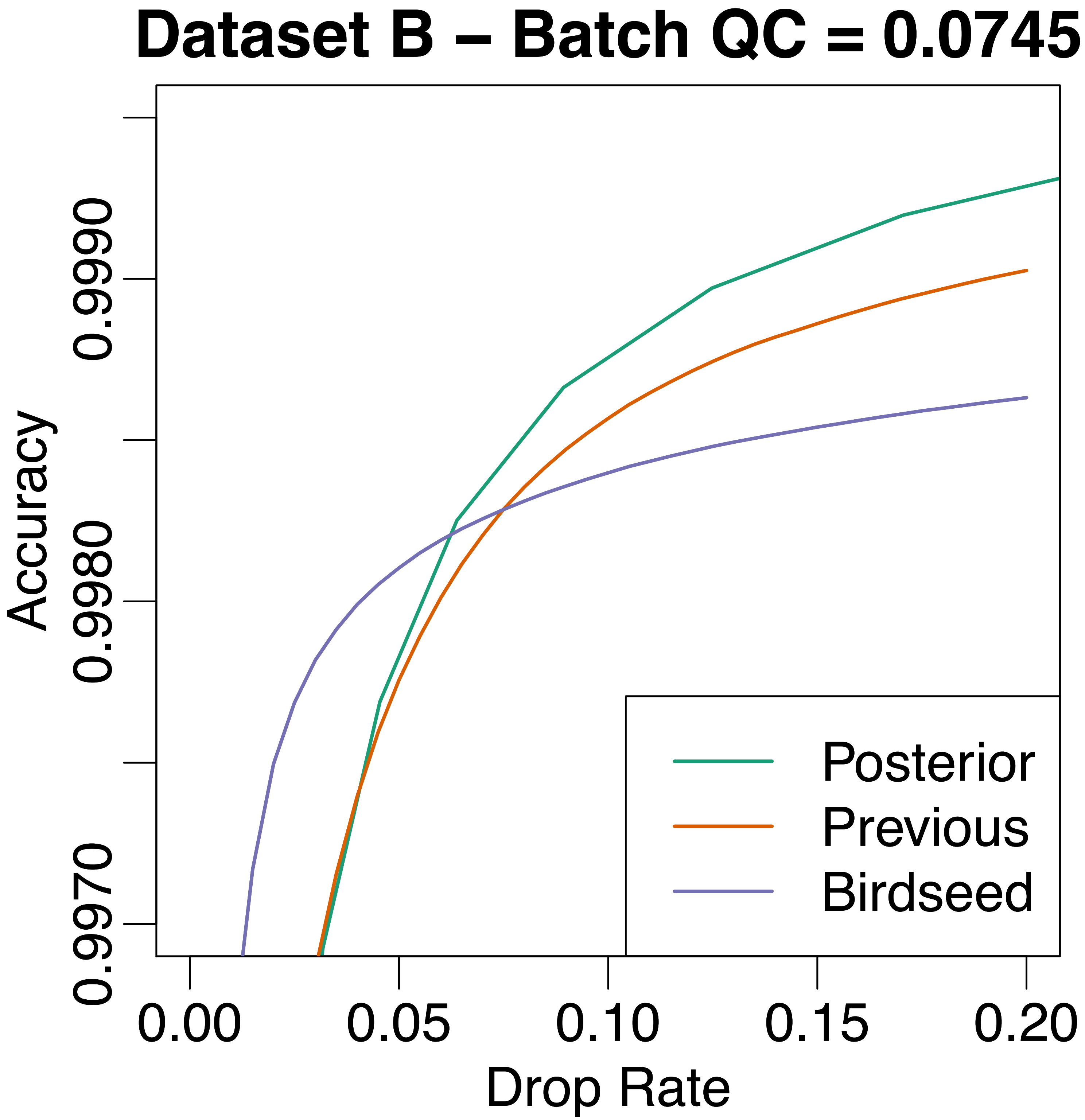
Observed Improvements



Observed Improvements



Observed Improvements



Summary

- New approach evaluated in different datasets;
- All samples are part of HapMap;
- Experiments performed in 7 laboratories;
- Posterior probabilities outperformed CRLMM;
- Sample size in training set was adequately accounted for;
- Also available for Illumina chips;