GS01 0163
Analysis of Microarray Data

Keith Baggerly and Bradley Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

24 November 2009

Lecture 25: SNP Arrays

- SNPs, GWAS, HapMap project
- Affymetrix SNP Arrays

Single Nucleotide Polymorphisms (SNPs)

Two unrelated people share about 99.5% of their DNA sequence.

One of the most common differences is that at specific sites some people may have (for instance) a `G`, while others might have an `A`. These sites are called single nucleotide polymorphisms, or SNPs.

Each of the two bases that can occur at the SNP is called an allele. (A third or even fourth allele is possible, but very uncommon. We will ignore the possibility hereafter.)

By convention, the most common allele at each SNP is called `A` and the less common SNP is called `B`.

Genotypes

Since there are two copies of each chromosome (except for X and Y in males), there are three possible pairs of alleles for each SNP: AA, AB, and BB.

For each SNP, an individual's genotype is the specific combination of alleles that it possesses.

Haplotypes

Chromosomes are not inherited as indivisible units. Through a process known as recombination the descendant's chromosome contains segments of DNA taken randomly from the two different parent chromosomes.

Segments of DNA that are far apart are inherited independently, whereas segments that are close together tend to be inherited together.

Segments that are common to many people are called haplotypes. The distribution of haplotypes varies between populations and geographic regions.

Sequences of adjacent SNPs can be used to model haplotypes.

Genome-wide Association Studies

Many genome-wide association and gene-environment interaction studies are being undertaken in order to find genes associated with complex, heritable disorders, including cancer.

Since there are about 10 million common SNPs, testing every individual for all SNPs would have been extremely expensive. (NGS might change that.)

By identifying tag SNPs that uniquely identify the common haplotypes, much less testing is required.

It is estimated that about 300,000 to 600,000 tag SNPs contain most of the information about the patterns of genetic variation.

International HapMap Project

The International HapMap project is a recent, large-scale effort to *facilitate* GWAS studies:

- Phase 1: 269 samples, 1.1 M SNPs
- Phase 2: 270 samples, 3.9 M SNPs
- Phase 3: 1115 samples, 1.6 M SNPs

Phase 3 platforms:

- Illumina Human1M (by Wellcome Trust Sanger Institute)
- Affymetrix SNP 6.0 (by Broad Institute)

HapMap Phase 3 Samples

label	population sample	# samples	QC+ Draft 1
ASW*	African ancestry in SW USA	90	71
CEU*	North/West European in Utah	180	162
CHB	Han Chinese in Beijing	90	82
CHD	Chinese in Denver	100	70
GIH	Gujarati Indians in Houston	100	83
JPT	Japanese in Tokyo	91	82
LWK	Luhya in Webuye	100	83
MEX*	Mexican in LA	90	71
MKK*	Maasai in Kinyawa	180	171
TSI	Toscans in Italy	100	77
YRI*	Yoruba in Ibadan	180	163
		1301	1115

Data Access

Unlike the TCGA SNP data, the HapMap project data is available from http://hapmap.ncbi.nlm.nih.gov/downloads/raw_data/?N=D.

We will use the `CUPID.tgz` dataset in the `hapmap3_affy6.0` subdirectory. This archive contains 77 CEL files from the Affymetrix SNP 6.0 platform.

Affymetrix SNP Chips

Mapping 10K (1 array, 18 μm feature size)

Mapping 10K v2.0

Mapping 100K (2 arrays, 8 μm feature size)

Mapping 500K (250K Nsp and 250K Sty, 5 μm feature size)

SNP 5.0

SNP 6.0

Affymetrix SNP 6.0

More than 906,600 SNPs:

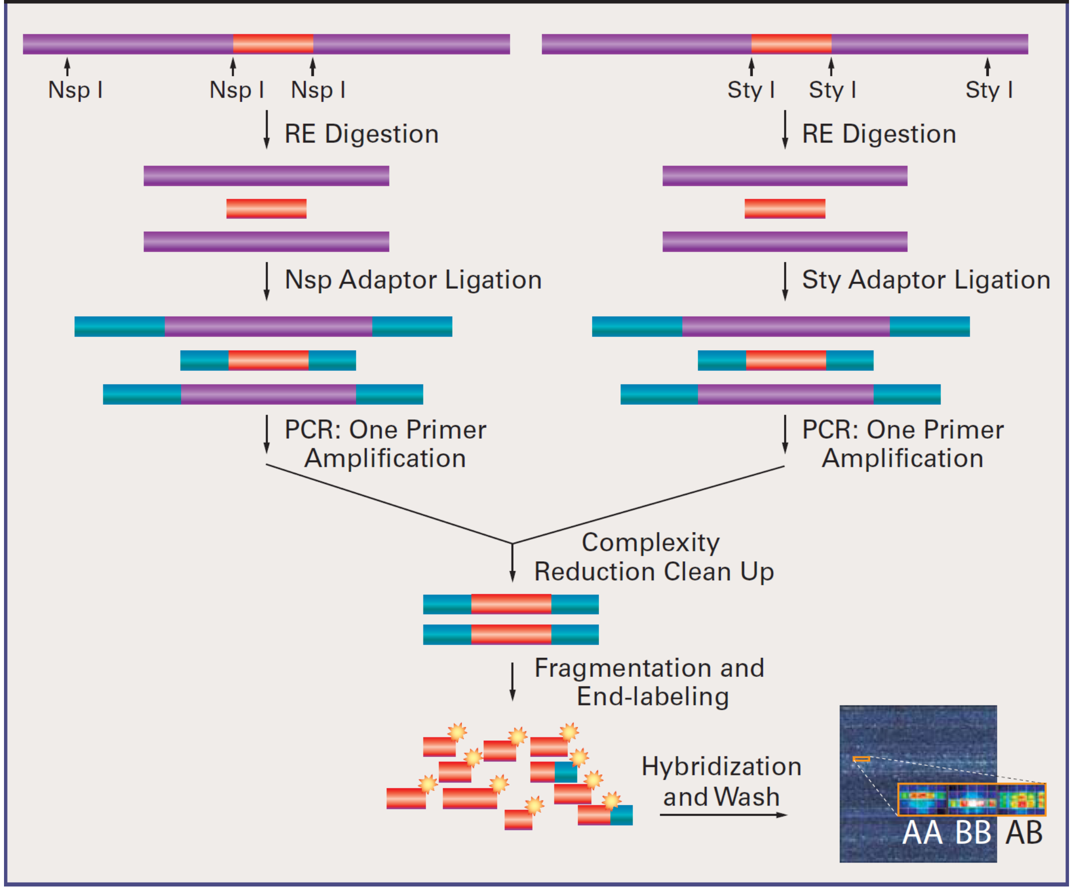
- Approx. 482,000 SNPs derived from previous generation arrays
- Additional tag SNPs from early phase of HapMap project

More than 946,000 probes for detecting copy number variation:

- 202,000 probes targetting 5,677 known regions of copy number variation
- more than 744,000 additional evenly spaced SNPs to enable detection of novel copy number variation

Affymetrix SNP 6.0 Assay

Figure 1: Overview of the Genome-Wide Human SNP Assay 5.0/6.0



Affymetrix Genomewide SNP 6.0 Datasheet

SNP Chip Design

The SNP chip's basic design is similar to that of expression arrays, in that an array of 25 bp oligonucleotide sequences (features) is laid across the surface of the chip. The sample's DNA is amplified, a marker is attached, and hybridized to the array. The array is scanned to quantify the relative amount of sample bound to each feature.

For SNPs, there is a pair of probes: one for each of the alleles.

For non-polymorphic CNV probes, there is just a single probe.

On early chips there were both PM and MM probes for each of the two alleles, making a quartet.

Early chips contained multiple quartets per SNP at different offsets (e.g. -4, -2, -1, 0, 1, 3, 4) to the SNP's location.

Recent chips just use two replicates of the PM pair that best distinguishes the two alleles.

Genotyping Algorithms

ABACUS

MPAM (Affy 10k)

DM (Affy 100k)

BRLMM (Affy 500k)

Birdseed (Affy SNP 6.0)

CRLMM

MPAM

Liu et al., Bioinformatics 19(18), 2003, 2397–2403.

Detection Filter:

Calculate discrimination score (DS) for each probe pair:

$$DS = (PM - MM)/(PM + MM)$$

$ds_i^{(sA)}$ is DS score of i'th probe pair for allele A on the sense strand.

$ds_i^{(tB)}$ is DS score of i'th probe pair for allele B on the antisense strand.

DS of A allele $d^{(sA)} = \mathrm{median}(d_i^{(sA)})$.

DS of SNP $d = \max(\min(d^{(sA)}, d^{(tA)}), \min(d^{(sB)}, d^{(tB)}))$.

Feature Extraction

Relative allele signal (RAS) for the ith probe quartet of the sense strand:

$$s_i^{(s)} = A_i^{(s)}/(A_i^{(s)} + B_i^{(s)})$$

where

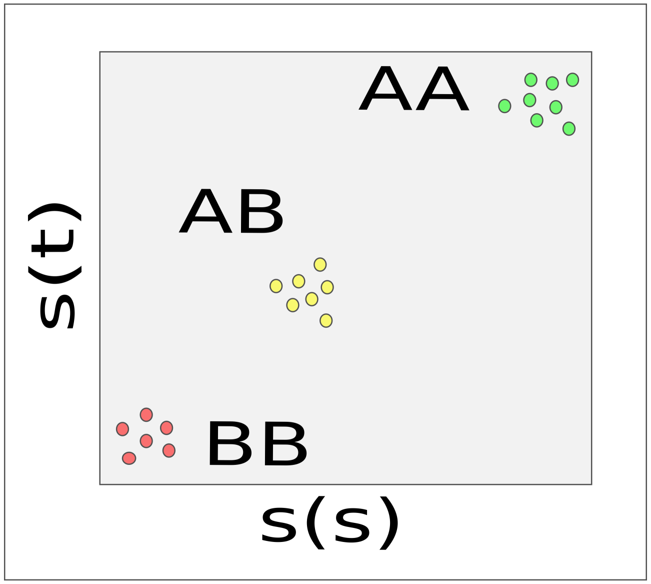
$$A_i^{(s)} = \max(PM_i^{(sA)} - MM_i^{(s)}, 0)$$
$$B_i^{(s)} = \max(PM_i^{(sB)} - MM_i^{(s)}, 0)$$
$$MM_i^{(s)} = (MM_i^{(sA)} + MM_i^{(sB)})/2$$

RAS for sense strand $s^{(s)} = \text{median}(s_i^{(s)})$

RAS for antisense strand $s^{(t)} = \text{median}(s_i^{(t)})$

Feature Space

The pair $(s^{(s)}, s^{(t)})$ is a point in a unit square feature space.



Points close to $(1, 1)$ should be AA.

Points close to $(0, 0)$ should be BB.

Points close to $(0.5, 0.5)$ should be AB.

Genotype Clusters are SNP Dependent

In real data, the locations, sizes, and shapes of the genotype clusters depend on the SNP concerned:

- affinity of target and probe depend on the sequence,
- cross hybridization.

Therefore, genotype cluster regions must be estimated for every SNP separately using a large training data set.

Classification with MPAM

For each SNP, use (modified) PAM to cluster features in training data into k groups.

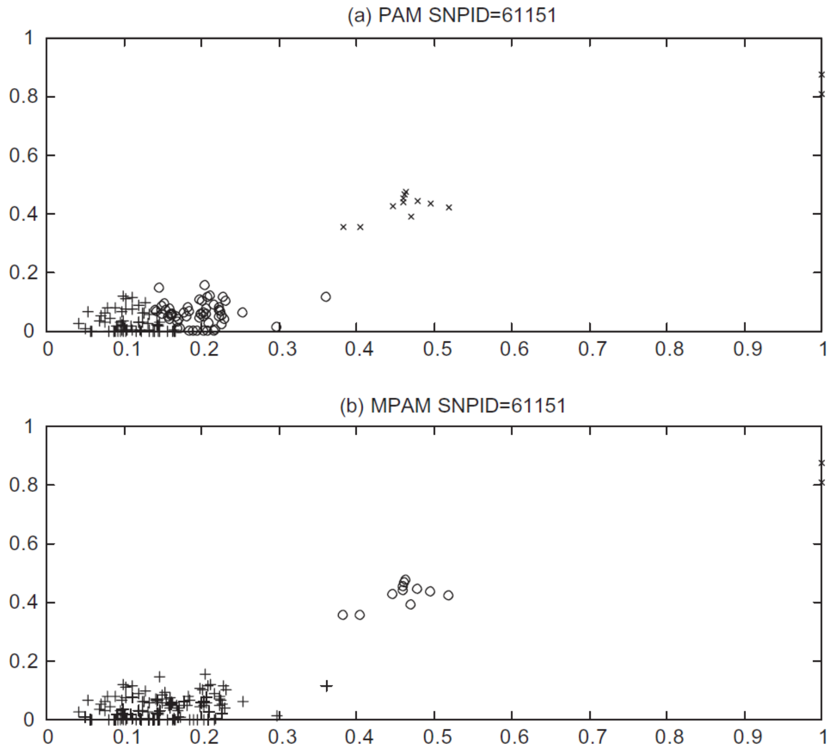
For genotyping, we expect 1 to 3 groups.

Assign genotypes based on median coordinates of the clusters.

Use average silhouette width to determine quality of the classification.

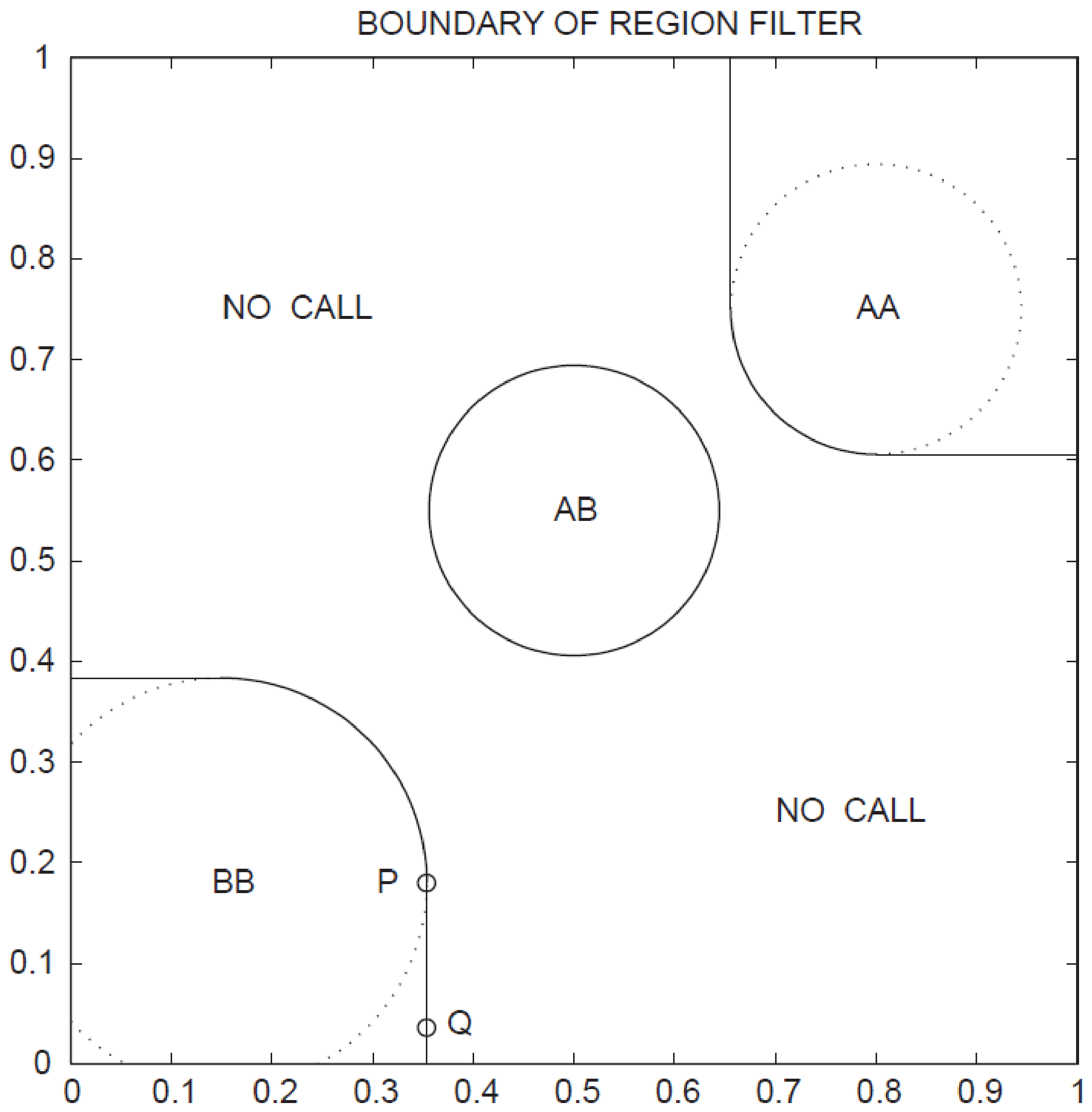
PAM fails with very different allele frequencies

With very different cluster sizes, PAM tends to split largest cluster:



PAM was modified to penalize small between-group distances.

Post-call filter to exclude bad calls



Dynamic Model-based Algorithm (DM)

Di et al., Bioinformatics 21(9), 2005, 1958–1963.

Introduced to overcome perceived deficiences in MPAM:

- Large number of training samples required to observe all three phenotypes and build accurate empirical models.
- SNPs with low minor allele frequency are difficult to model accurately.
- Requires manual inspection of selected SNPs.
- Not flexible enough to accommodate additional improvements.

DM

DM:

- aggregates multiple SNP quartets into SNP genotype call and confidence metric
- stratifies states into four models: Null, A, AB, and B.
- uses a one-sided Wilcoxon signed rank test to produce four p-values, one for each model

DM also includes methods for SNP screening and probe reduction, and enables SNP screening using a relatively small sample set.

Four Models

Four states for each quartet:

Null: No probe brighter than the rest. All assumed to be background.

A or B: Only PM probe for A (or B) is bright. Other 3 probes assumed to be background.

AB: Both PM probes are bright. Two MM probes assumed to be background.

Which state is most likely?

Likelihood Models

Assume all probes in a quartet are independent, and signal intensities are i.i.d. normal random variables.

For each probe in a quartet ($x = 1, 2, 3, 4$):

μ_x	mean
σ_x^2	variance
n_x	number of pixels
$\hat{\mu}_x$	estimated mean assuming model m
$\hat{\sigma}_x^2$	estimated variance assuming model m

Log likelihood given by

$$L(m) = -\frac{1}{2}\sum_{x=1}^{4} n_x \left[\ln(2\pi\hat{\sigma}_x^2) + \frac{\sigma_x^2 + (\mu_x - \hat{\mu}_x)^2}{\hat{\sigma}_x^2}\right]$$

Estimated mean and variance for Null model

For the Null model, all probes are background and evenly distibuted, hence:

$$\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = \hat{\mu}_4 = \frac{\sum_{x=1}^{4} n_x \mu_x}{\sum_{x=1}^{4} n_x}$$

$$\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_3^2 = \hat{\sigma}_4^2 = \frac{\sum_{x=1}^{4} n_x [\sigma_x^2 + \mu_x^2]}{\sum_{x=1}^{4} n_x} - \hat{\mu}^2$$

Estimated mean and variance for model A

For model A, the PM for A is foreground, the other three are background:

$$\hat{\mu}_1 = \mu_1$$
$$\hat{\sigma}_1^2 = \sigma_1^2$$

$$\hat{\mu}_2 = \hat{\mu}_3 = \hat{\mu}_4 = \frac{\sum_{x \neq 1} n_x \mu_x}{\sum_{x \neq 1} n_x}$$
$$\hat{\sigma}_2^2 = \hat{\sigma}_3^2 = \hat{\sigma}_4^2 = \frac{\sum_{x \neq 1} n_x [\sigma_x^2 + \mu_x^2]}{\sum_{x \neq 1} n_x} - \hat{\mu}^2$$

Similarly for model B.

Estimated mean and variance for model AB

For model AB, the PMs for A and B are foreground, the two MM are background:

$$\hat{\mu}_1 = \hat{\mu}_3 = \frac{n_1\mu_1 + n_3\mu_3}{n_1 + n_3}$$

$$\hat{\sigma}_1^2 = \hat{\sigma}_3^2 = \frac{n_1[\sigma_1^2 + (\hat{\mu}_1 - \mu_1)^2] + n_3[\sigma_3^2 + (\hat{\mu}_3 - \mu_3)^2]}{n_1 + n_3}$$

$$\hat{\mu}_2 = \hat{\mu}_4 = \frac{n_2\mu_2 + n_4\mu_4}{n_2 + n_4}$$

$$\hat{\sigma}_2^2 = \hat{\sigma}_4^2 = \frac{n_2[\sigma_2^2 + (\hat{\mu}_2 - \mu_2)^2] + n_4[\sigma_4^2 + (\hat{\mu}_4 - \mu_4)^2]}{n_2 + n_4}$$

SNP Level Aggregation

Different probe quartets might support different states. Need to aggregate robustly over all probe quartets.

Score for model m:

$$S(m) = L(m) - \max\{L(k), k = 1, 2, 3, 4, k \neq m\}$$

For each model m generate a vector of the scores for all n quartets in the SNP:

$$V_m = \{S_1(m), S_2(m), \ldots, S_n(m)\}, m = \mathrm{Null, A, AB, B}$$

Finding most likely model

Use Wilcoxon signed rank test to evaluate support for each model across all probe quartets.

For all 4 models, apply to hypotheses $H_0 : \text{median}(S_i(m)) = 0$ versus $H_1 : \text{median}(S_i(m)) > 0$ to obtain 4 p-values.

The least p-value, provided it is below a threshold, determines the genotype call and the p-value is the confidence of that call.

If the least p-value exceeds the threshold, or the best model is null, the final genotype is no-call.

BRLMM

BRLMM

- performs multiple chip analysis: simultaneous estimation of probe effects and allele signals for each SNP.
- estimates genotypes by a multiple-sample classification, borrowing information from other SNPs as necessary to make better predictions.

BRLMM makes weaker assumptions about probe behavior, making it more robust on real-world data.

BRLMM Approach

1. Normalize probe intensities and estimate allele signals for each SNP

2. Use DM to make an initial guess at each SNPs genotype

3. Select SNPs containing a minimum number of all three genotypes

4. Transform allele signal estimates into a better behaved 2D space

5. Use selected SNPs to estimate a prior distribution of typical cluster centers and variance-covariance matrices

6. Re-evaluate each SNP, combining initial genotype guesses with prior information in an ad-hoc Bayesian procedure to derive a posterior estimate of cluster centers and variances

7. Determine genotype and confidence score for each observation based on its Mahalanobis distance from the three cluster centers

Normalization and Probe Intensities

Use quantile normalization at the feature level.

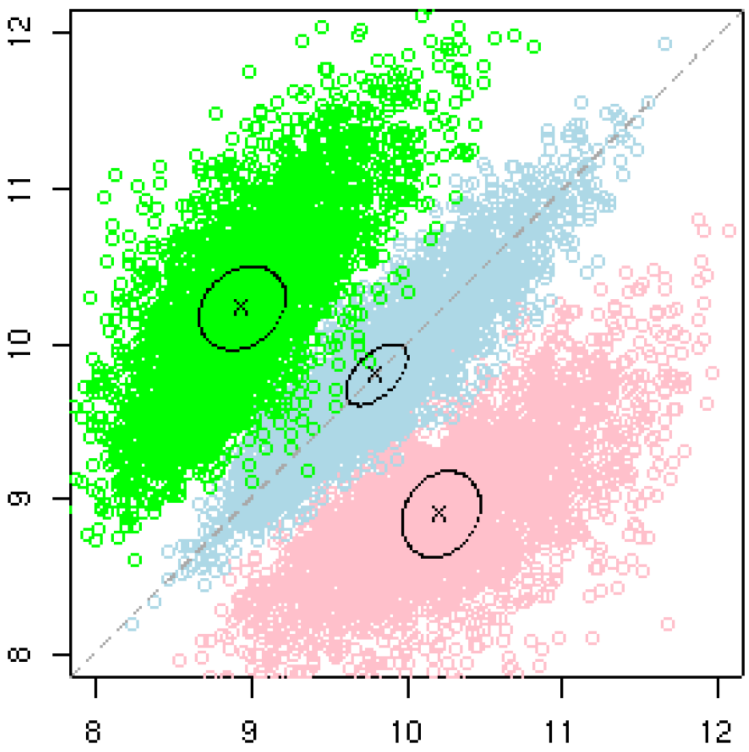
No background correction used. (For most fragments containing SNPs, target levels are well above background.)

Use log-scale transformation for the PM intensities.

Use median polish to fit feature effects to the data and obtain a signal.

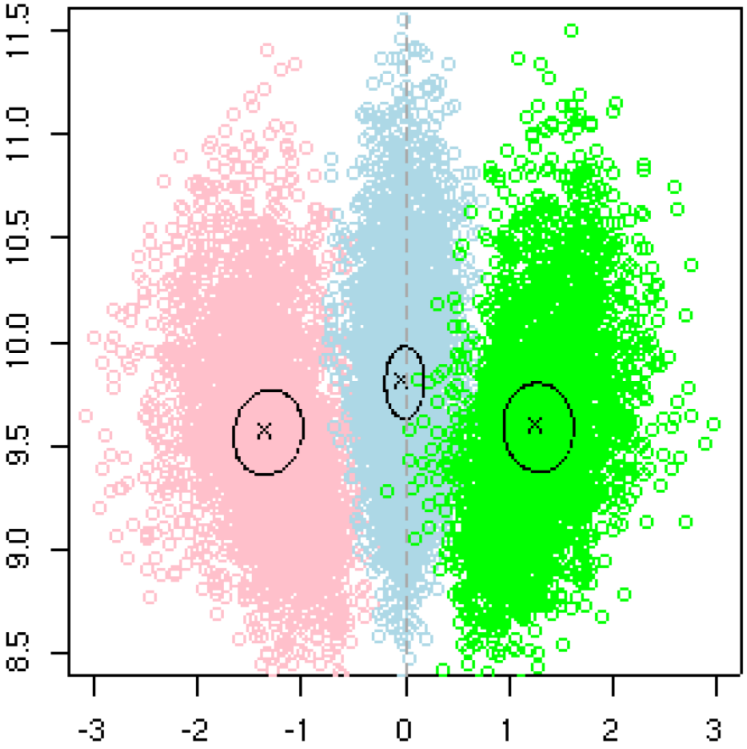
Summarize probes into two values, representing A and B signals.

Clustering Space Transformation



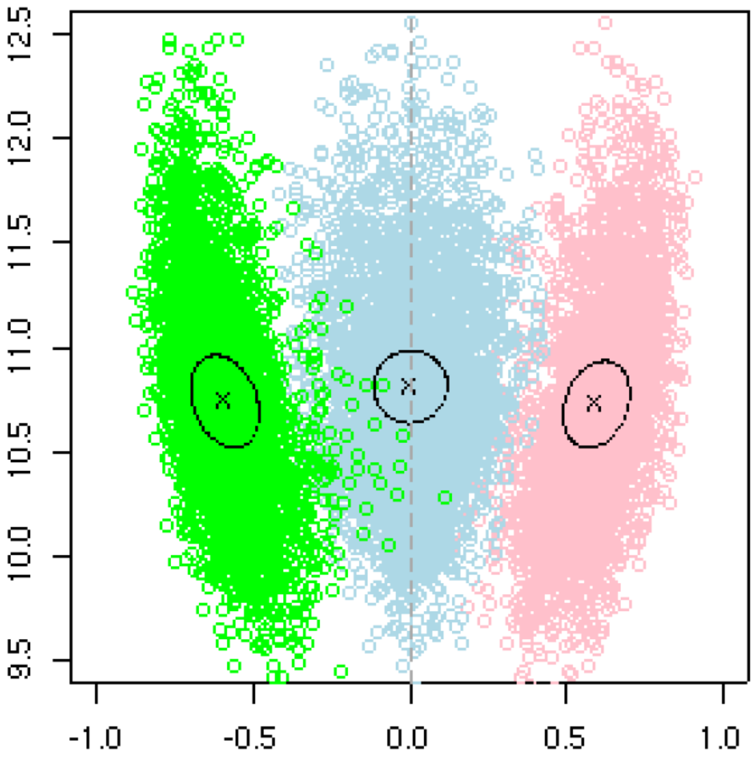
Clustering Space Transformation

MA transformation isolates most of the difference between genotypes onto the M axis, but artificially makes homozygous clusters more broadly variable than heterozygous clusters.



Clustering Space Transformation

Transformed Contrast transformation can be used to balance the variability in homozygous and heterozygous clusters.



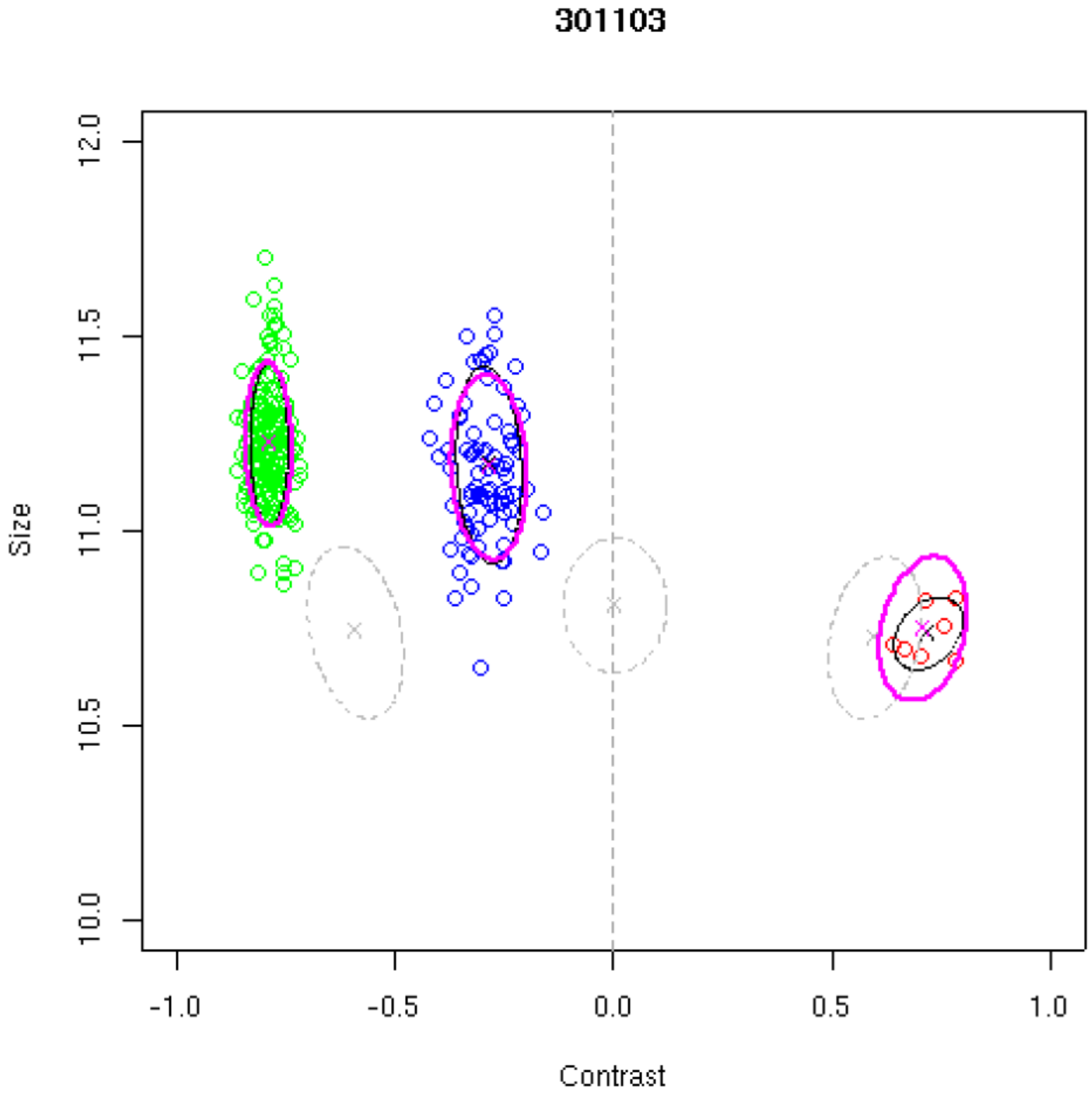
Genotyping

Model three clusters (for AA, AB, BB) defining cluster centers and covariance matrices.

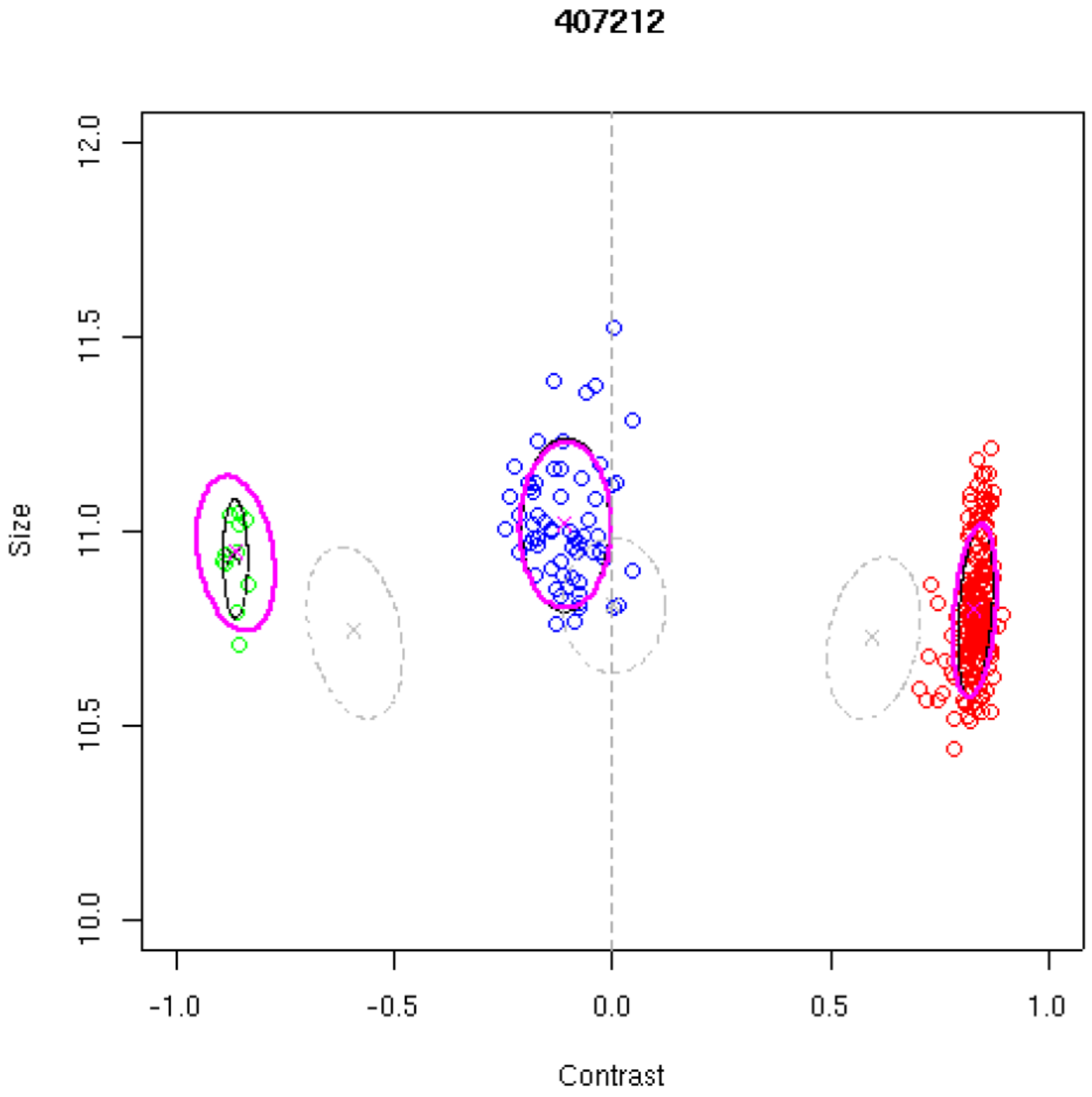
Determine distance from each test point to each cluster center using Mahalanobis distance (which takes into account variation and covariation in the cluster along each axis).

Confidence assigned to call is d_1/d_2 where d_1 is the smallest distance and d_2 is the second smallest distance.

Example Clustering



Example Clustering



CRLMM

Carvalho et al. developed preprocessing method that removes the bulk of the batch effect.

This permits the use of HapMap as training data.

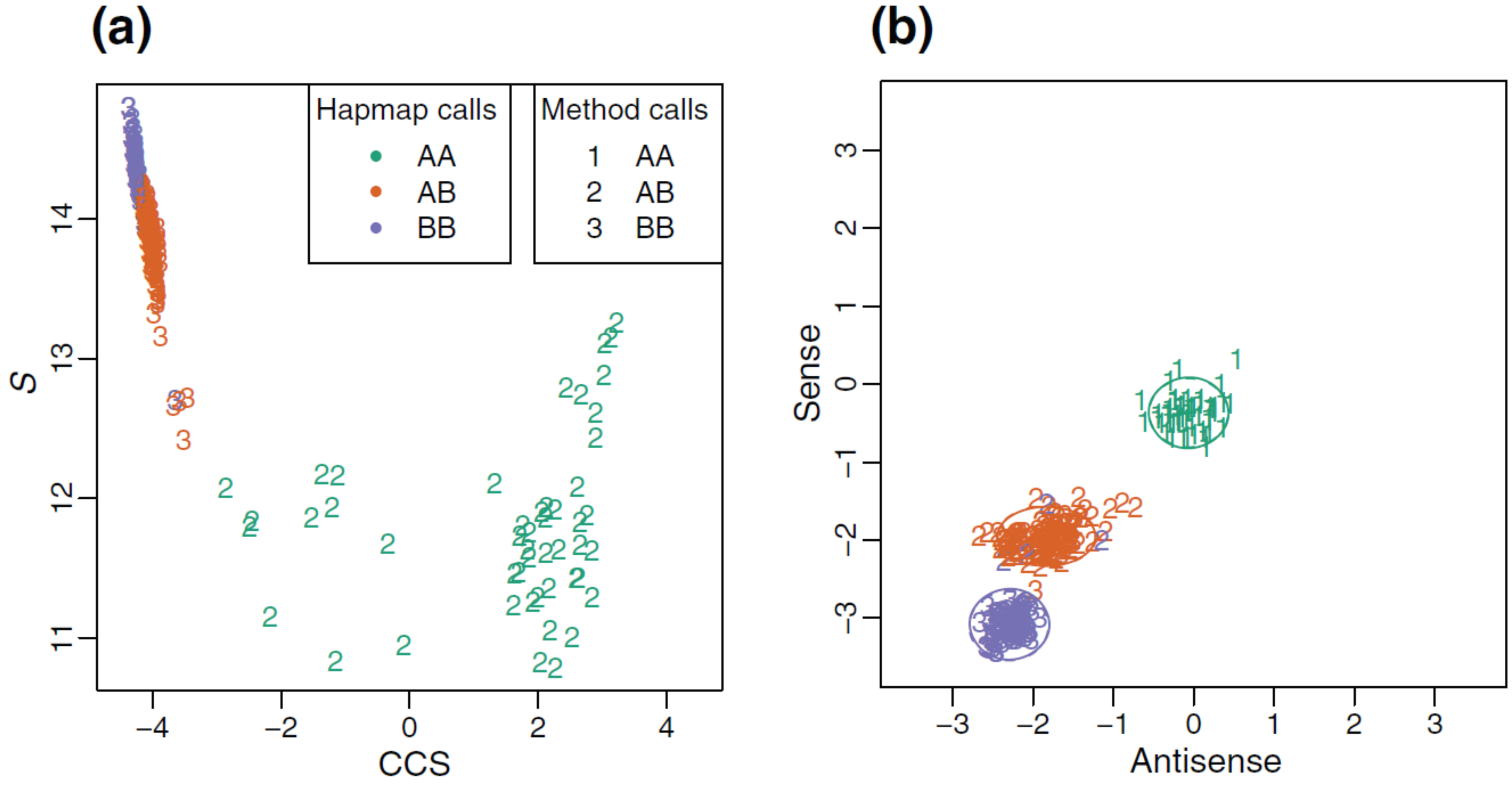
Summarize probes similar to RMA.

Transform features into log ratio M and average log intensity S for both sense and antisense strands.

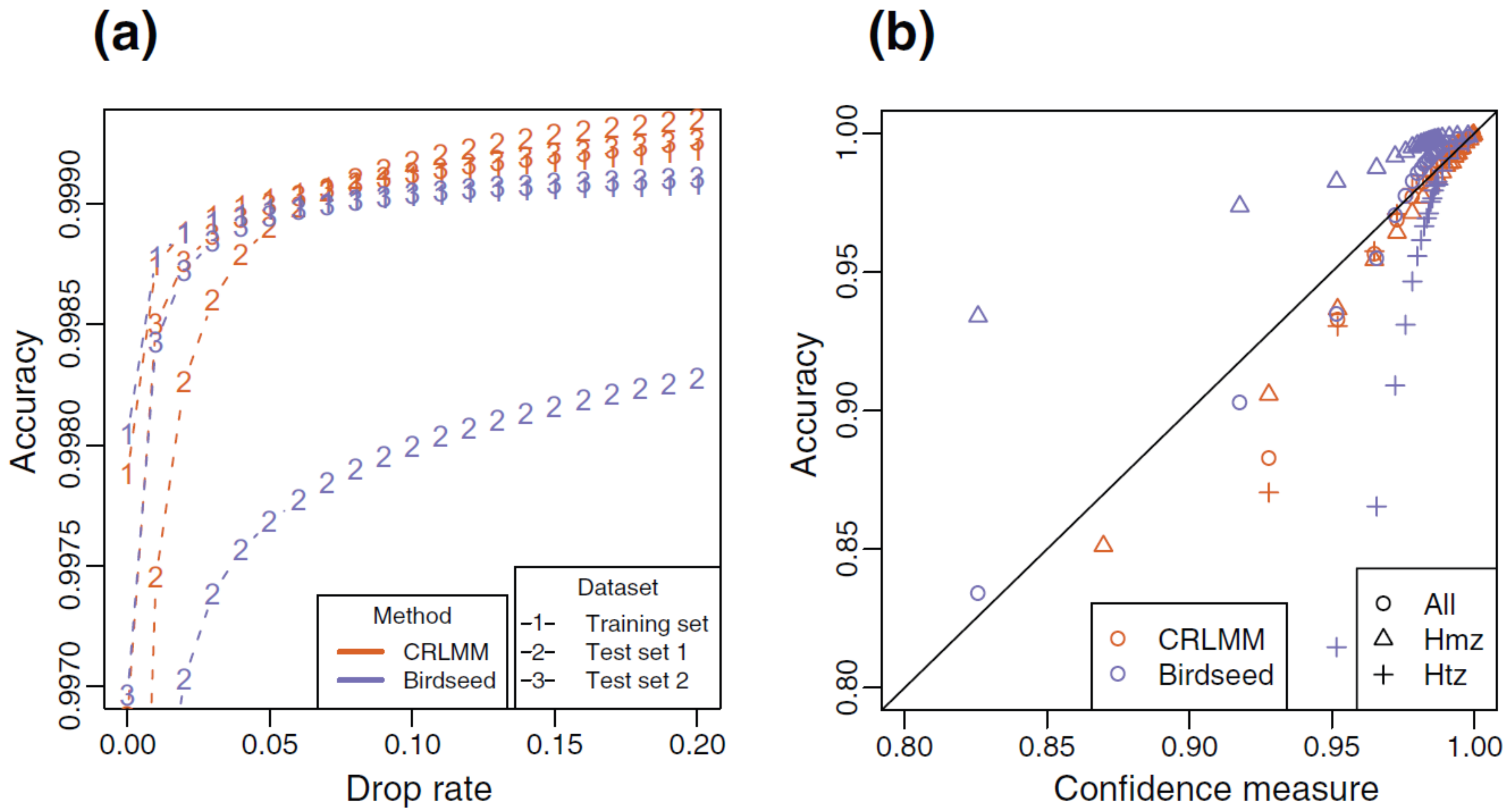
Use HapMap data (where available) to estimate priors on genotype regions.

Use genotype calls that achieve at least 99% concordance, recalculate genotype centers and scales, and recompute calls and log-likelihood ratios.

Example on which BRLMM does poorly



SNP Quality: CRLMM vs Birdseed



CRLMM Availability

CRLMM is available as a Bioconductor package for R.