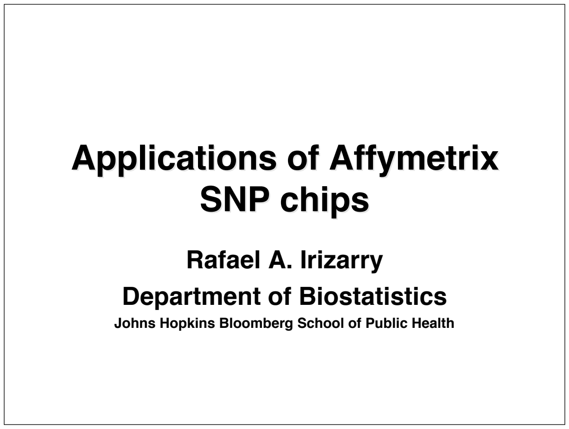


Applications of Affymetrix SNP chips

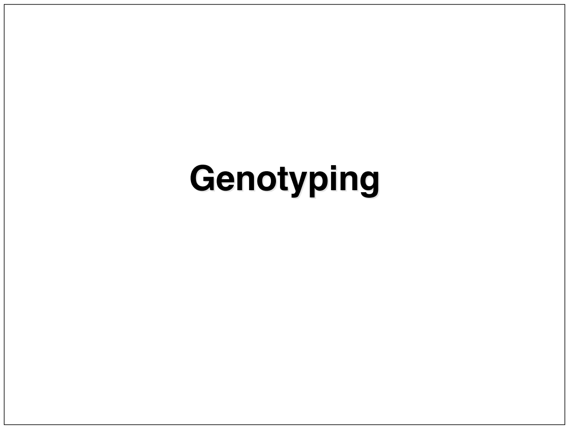
Rafael A. Irizarry

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
- Seth Falcon, Robert Gentleman and Bioconductor Team



Genotyping

What are SNPs?

SNP

A

Genomic DNA: **TAGCCATCGGTANGTACTCAATGAT**

G

A person can be AA , AG or GG

Affymetrix SNP chip terminology

SNP

A

Genomic DNA: **TACATAGCCATCGGTANGTACTCAATGATGATA**

G

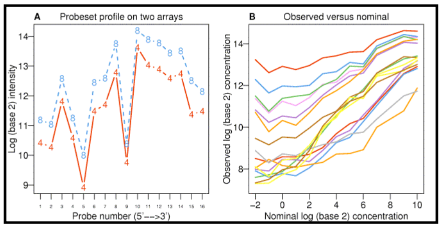
PM probe for Allele A: **ATCGGTAGCCATTCATGAGTTACTA**

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?

Probe effect



Affymetrix SNP chip terminology

SNP

Genomic DNA: TACATAGCCATCGGTA$\overset{A}{\underset{G}{N}}$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?

Affymetrix SNP chip terminology

SNP

Genomic DNA: TACATAGCCATCGGTA$\overset{A}{\underset{G}{N}}$GTACTCAATGATGATA

PM probe for Allele A: **GTAGCCATTCATGAGTTACTACTCT**

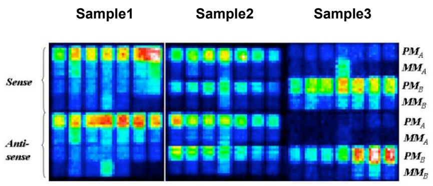
PM probe for Allele B: **GTAGCCATCCATGAGTTACTACTCT**

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.

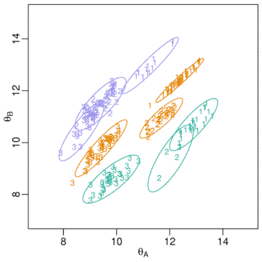
Notation

- Once we are done with first part of preprocessing we have the following:

θ_A and θ_B proportional to log of the amount of fragments from allele A and B respectively

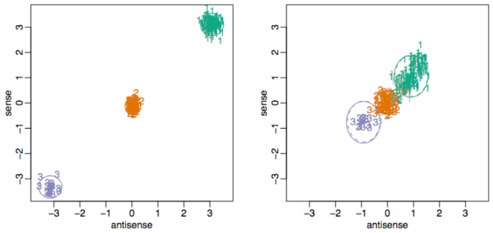
In principal these can only be (log of) 0, x, or 2x, but we know better than to believe this.. In fact we know not to expect the same cut-off to work for all SNPs

It's not easy

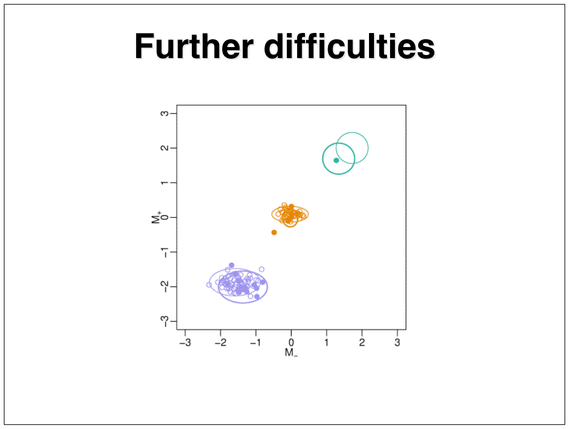


This picture shows that most the information is in the left right diagonal direction, i.e. in the log-ratios

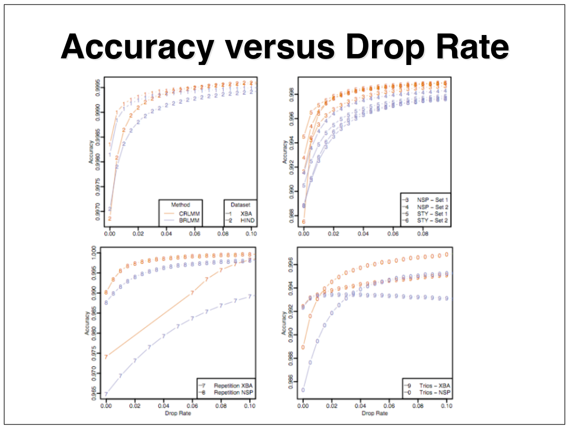
CRLMM



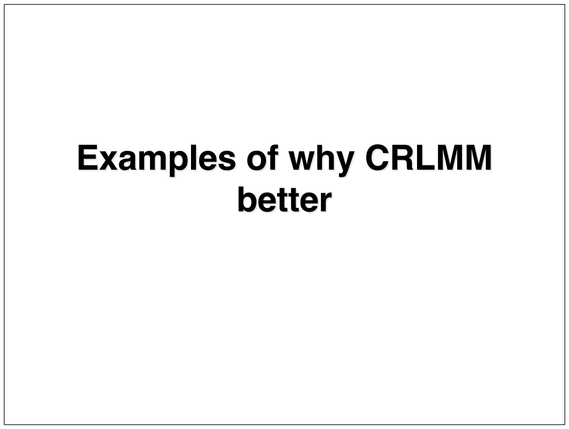
Carvalho et al. (2007) Biostatistics



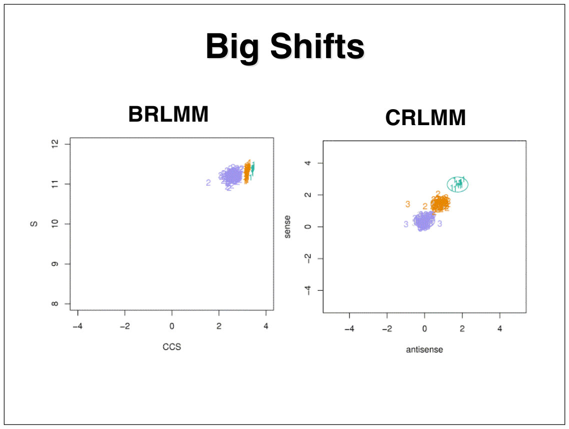
Further difficulties



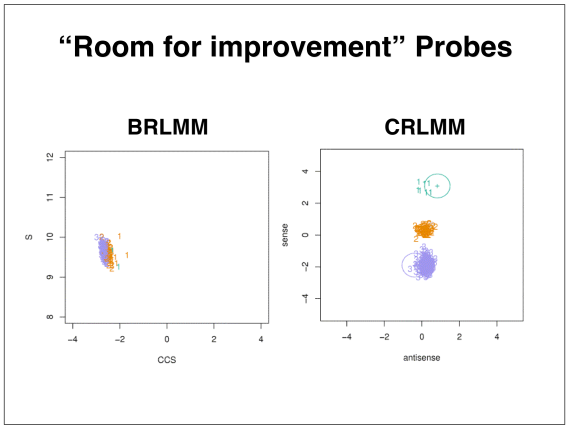
Accuracy versus Drop Rate



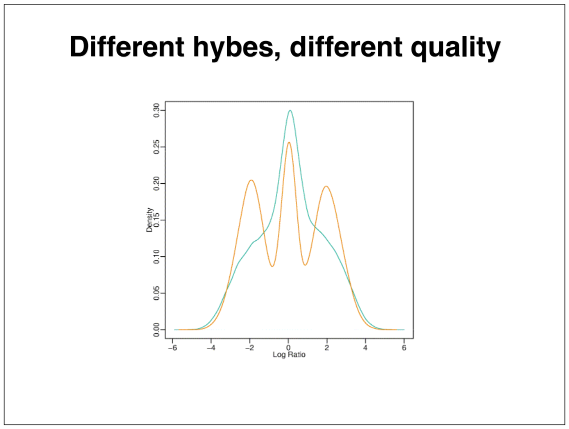
Examples of why CRLMM better



Big Shifts

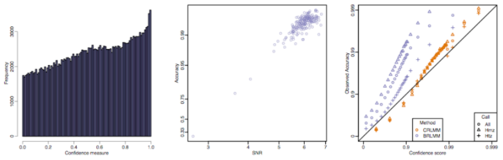


“Room for improvement” Probes



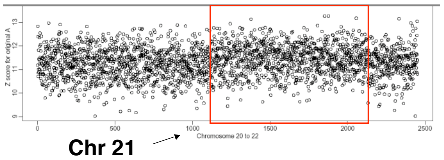
Different hybes, different quality

Bad Hybes



Copy Number

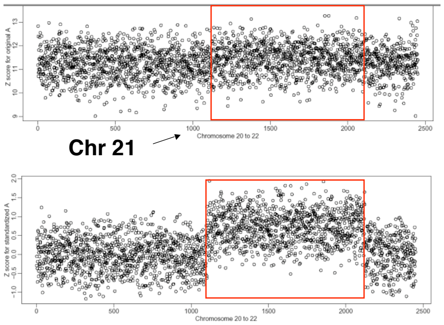
Copy Number



Chr 21

Now we want absolutes:
Probe effect a problem!

Copy Number



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)

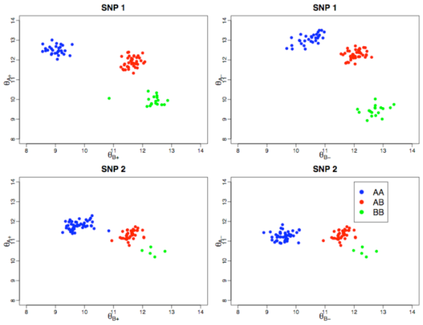
Model adapted for copy number applications:

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

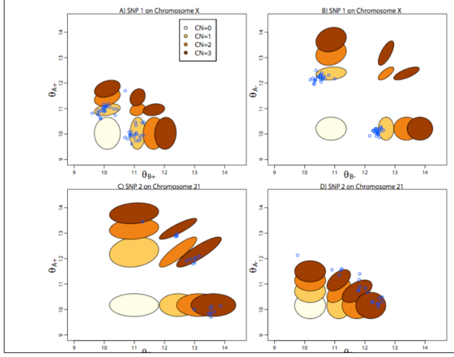
Some Current Approaches

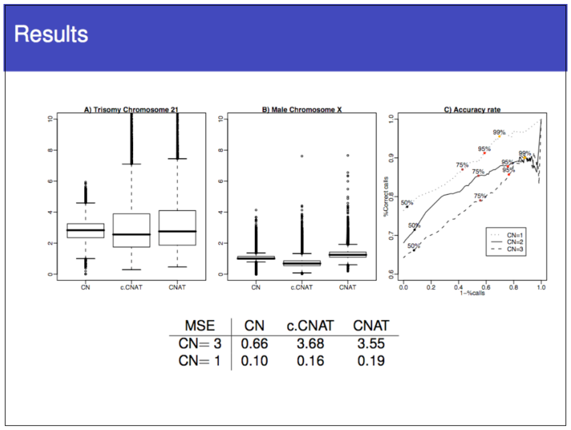
- CNAT: Huang et al. Human Genomics (2004)
- CGAG: Nannya et al. Cancer Research (2005)
- GIM: Ishiwaka et al. Biochem Biophys Res Commun (2005)
- PLASQ: Laframboise et al. Biostatistics (2006)
- CARAT: Huang et al. BMC Bioinformatics (2006)

We use genotype calls?



Example: Mixture Models





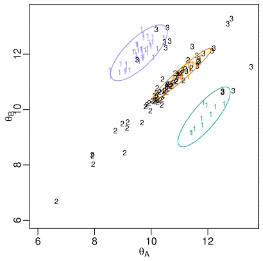
Results

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

Thanks!

Supplemental Slides

Lab Effect

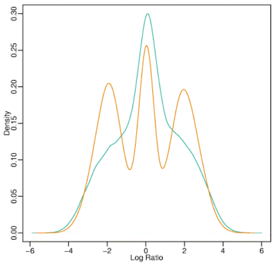


Why is this?

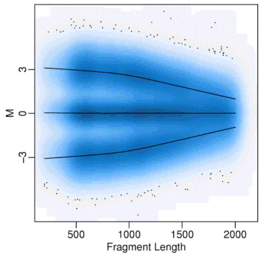
- Our guess is that the PCR step introduces a lot of SNP to SNP variation
- We have proxies for measuring PCR effect: fragment sequence and fragment length
- We can examine the fragment sequence via the probe sequence

Log-ratio biases persist

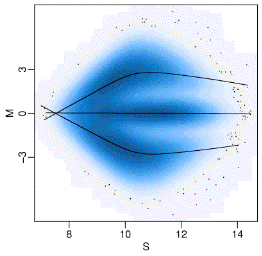
Different hybes, different quality



Length effect on M



Intensity effect on M



Normalization

- We normalize/summarize using RMA (no BG correction) after correcting for sequence and length effects on the log intensities
- We then examine log-ratios
- We keep sense and antisense separate

Use mixture model to fix this

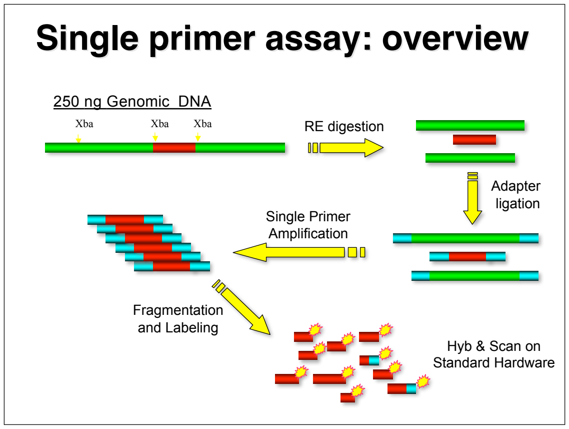
$$[M_i|Z_i = k] = f_k(X_i) + \varepsilon_{i,k}$$

- SNP denoted with I
- Z is true, so k = AA, AB or BB
- X are covariates that cause bias
- We later use SNR = Median(f_1)^2 / Var(ε) as measure of quality

Preprocessing model motivates genotype algorithm

$$[M_{i,j,s}|Z_{i,j} = k, m_{i,k,s}] = f_{j,k}(X_{i,j,s}) + m_{i,k,s} + \varepsilon_{i,j,k,s}.$$

•Array denoted with j
•Shift in cluster center denoted with m
•Assume m are bivairate normal with covariance V and the variance of the measurement error is inverse chi-squared
•Use training data to estimate
•Use empirical bayes approach for cases with few data points



Single primer assay: overview

250 ng Genomic DNA

Xba

Xba

Xba

RE digestion

Adapter ligation

Single Primer Amplification

Fragmentation and Labeling

Hyb & Scan on Standard Hardware