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?**

Genomic DNA: TAGCCATCGGTAGTACTCAATGAT

A person can be AA, AG or GG

---

**Affymetrix SNP chip terminology**

Genomic DNA: TACATAGCCATCGGTAGTACTCAATGATGATA

PM probe for Allele A: ATCGTAGCCATCATGAGTTACTA

PM probe for Allele B: ATCGTAGCCATCCATGAGTTACTA

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**

![Probe effect graph]

A: Observed and expected signal response
B: Observed and expected signal response
Affymetrix SNP chip terminology

Genomic DNA: TACATAGCCATCGGTANGTACTCAATGATGATA
PM probe for Allele A: ATCGGATAGCCATTCTAGCTGTTACTA
PM probe for Allele B: ATCGGATAGCCATTCTAGCTGTTACTA

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

Genomic DNA: TACATAGCCATCGGTANGTACTCAATGATGATA
PM probe for Allele A: GTAGCCATCATGAGTTACTACTCT
PM probe for Allele B: GTAGCCATCATGAGTTACTACTCT

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

Sample1 Sample2 Sample3

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:

\[ \theta_A \text{ and } \theta_B \text{ proportional to log of the amount of fragments from allele A and B respectively} \]

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

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:
\[ I_{ij} = \beta + \sigma \exp(\phi + \epsilon_{ij}) \]
Some Current Approaches


We use genotype calls?

Example: Mixture Models
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_i)^2 / Var(\varepsilon) as measure of quality

**Preprocessing model motivates genotype algorithm**

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

- Array denoted with j
- Shift in cluster center denoted with m
- Assume m are bivariate 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