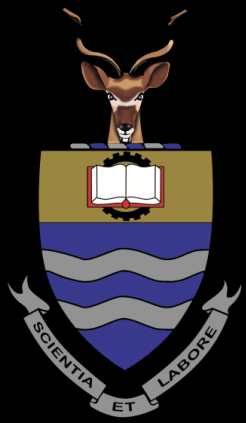
H3ABioNet

Pan African Bioinformatics Network for H3Africa



Introduction to Genome Wide Association Studies 2015
Sydney Brenner Institute for Molecular Bioscience
Shaun Aron

Raw data QC – Genotype Calling

Technical Bias

- Many sources of technical bias in a genotyping experiment
  - DNA sample quality and handling
  - Experimental conditions
  - Variations in genotyping microarray chip
  - Batch effects
- Not related to biology and may lead to confounding of call rates and genotyping

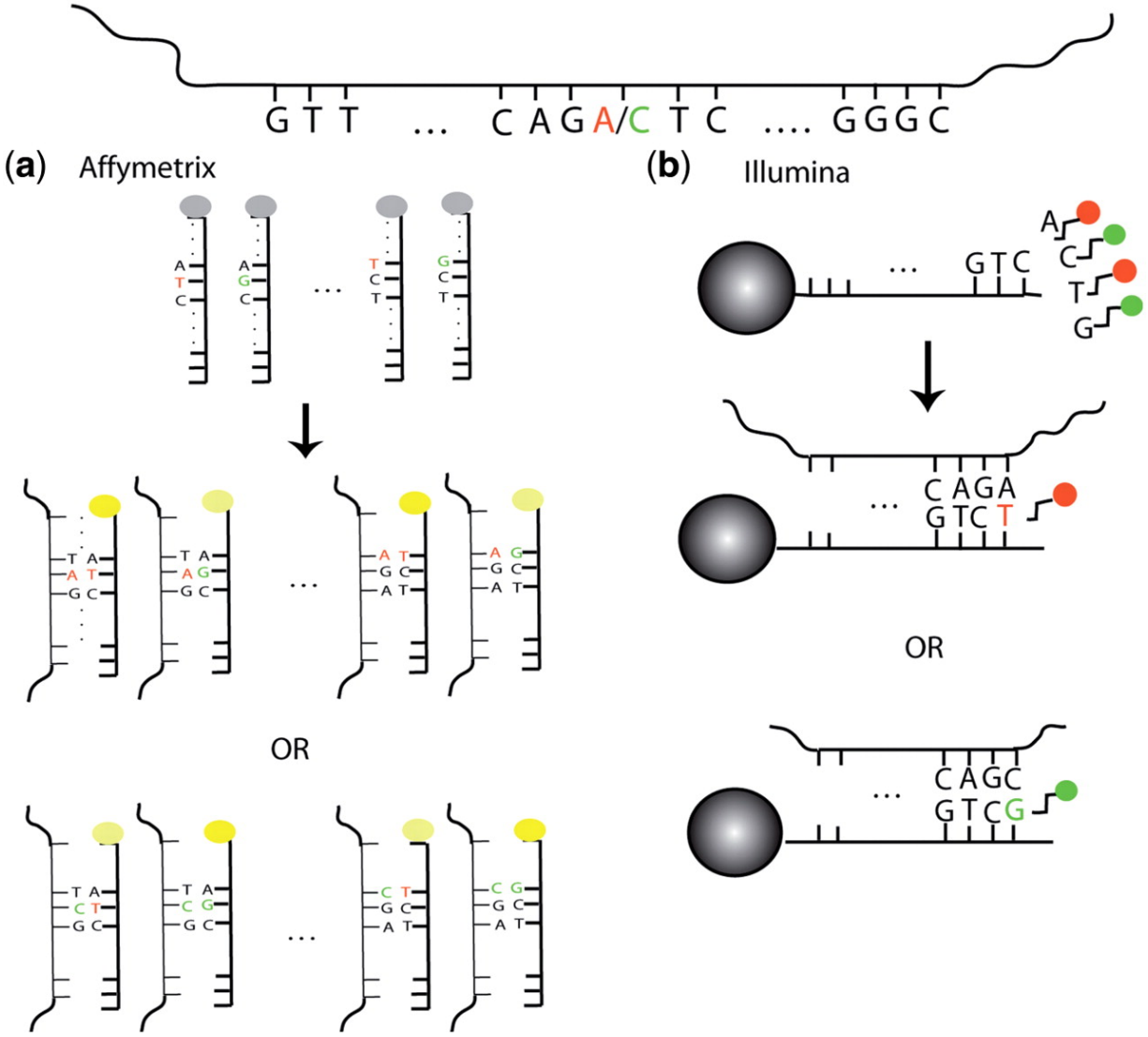
QC of raw intensity data

- First stage of analysis is the conversion of raw intensity values to genotypes
- Several further qc steps to remove technical and biological bias
- Focus of this lecture is on the qc of the raw intensity data and genotype calling
- Following lecture is based on the qc of genotype data

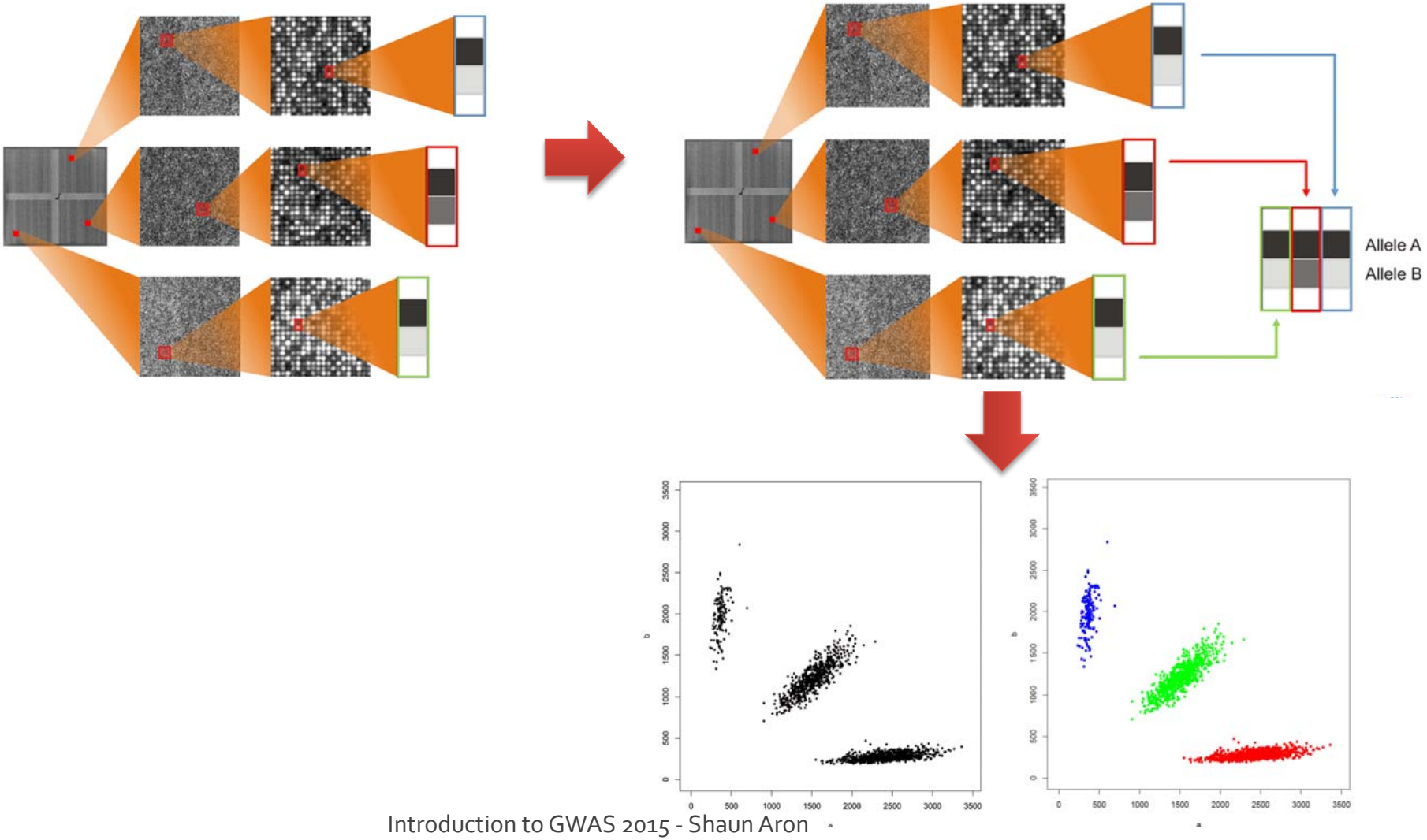
GWAS data

- Most service providers will provide you with either raw CEL (idat). files or genotype call files
- The raw CEL (idat) files contain the raw intensity values from the microarray chip
- The genotype file will contain the actual alleles called for each SNP
- Various methods available for calling genotypes from raw intensity files

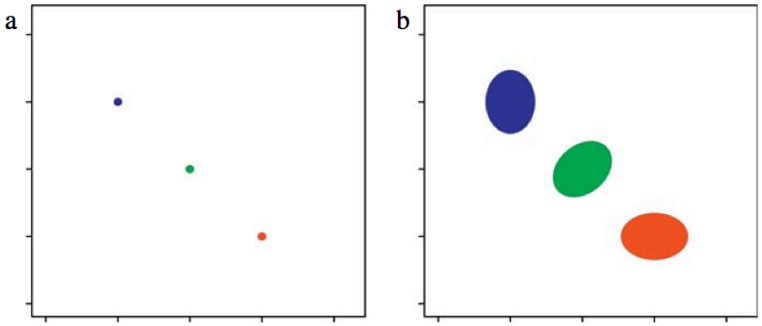
Affymetrix vs Illumina

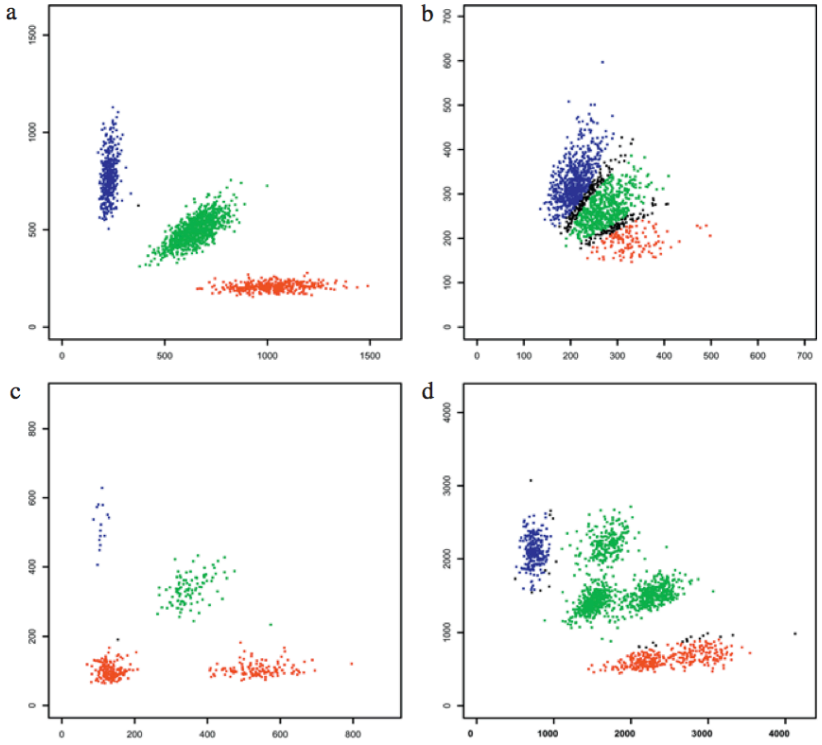


Genotype calling



Genotype calling





Data files

- .CEL(idat) files – Cell Intensity File
  - Includes intensity values, standard deviation of intensity, number of pixels used to calculate intensity value
  - A flag to indicate an outlier as determined by the algorithm used to measure intensity values
  - Most Illumina platforms include proprietary software that calls genotypes from idat files

Preprocessing

- The main goal of SNP array preprocessing is to summarise intensity values into quantities that can be used to discriminate genotype classes with some measure of confidence
  - The first step is to measure the how well the intensity values separate into distinct clusters
  - Then use the SNPs that cluster well to call the genotypes associated with each of the clusters

Genotype Calling

- There are several different algorithms designed to normalise and analyse the raw intensities measures to arrive at inferred genotypes
- These methods have been developed and improved together with the advancement and improvement of the microarray chips

Genotype calling algorithms

AFFYMETRIX

- Modified Partitioning Around Medoids (MPAM)
- Dynamic Model (DM)
- The Robust Linear Model with Mahalanobis distance classifier (RLMM)
- Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM)
- Birdseed

ARRAY

- 10K
- 100K
- 500K
- Human Genome-wide SNP6.0

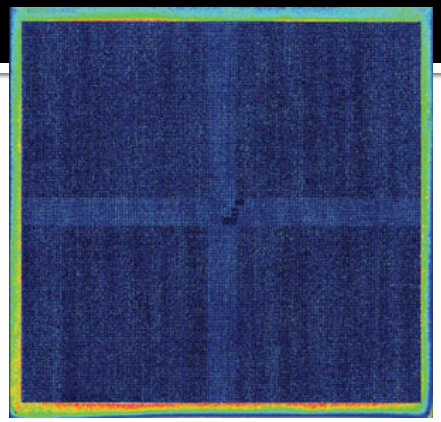
Genotype calling algorithms

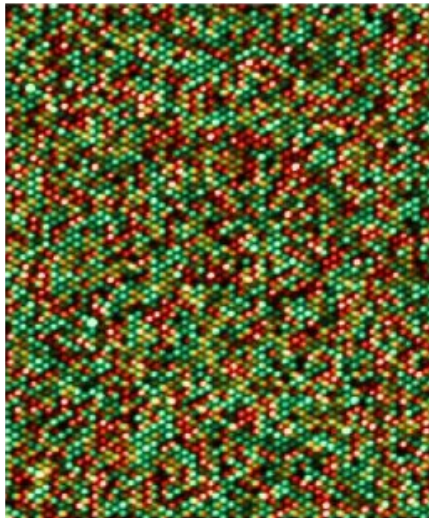
ILLUMINA

- GenCall – proprietary software in GenomeStudio
- GenoSNP
- Illuminus

- Newer algorithms
  - M3
  - zCall
  - OptiCall

Genotype calling





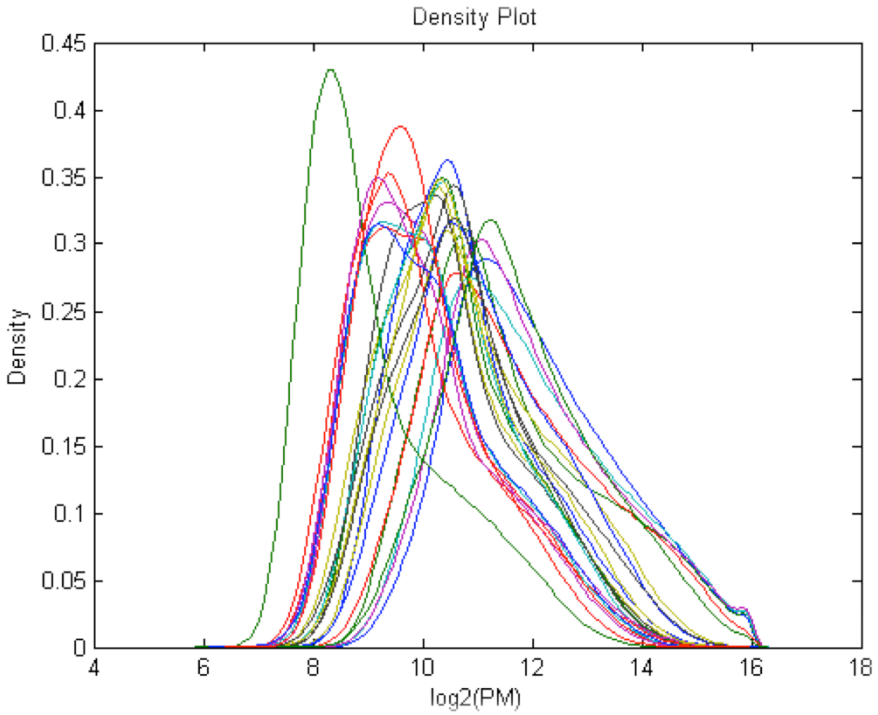
Algorithm	Insitute	Reference
Birdseed	Affymetrix/Broad	Korn et.al 2008 Nat Gen 40:1253-1260
BRLMM	Affymetrix	Cawley et al. 2006
CHIAMO	WTCCC	WTCCC 2007 Nature 447:661 -78
CRLMM	John Hopkins University	Carvalho et al. 2007 Biostatistics 8:485-99
GEL	University of Chicago	Nicolae et al. Bioinformatics 22:1942-7
JAPL	Wellcome Trust, Sanger Institute	Plagnol et al. 2007 PLoS Genetics 3:e74
OptiCall	Wellcome Trust, Sanger Insitute	Shah et al. 2012 Bioinformatics 28:1598-1603

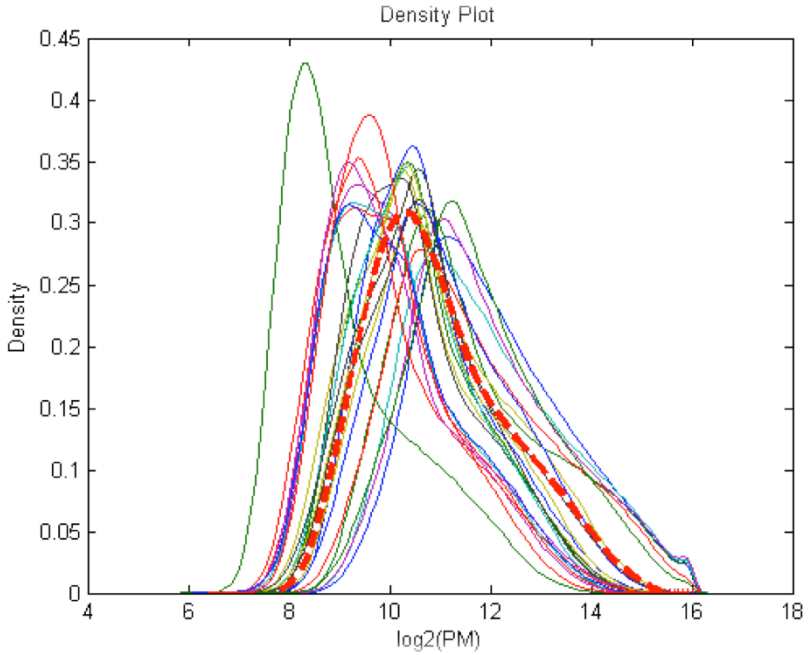
Genotype calling algorithms

- First normalises intensities either across arrays or within a single array
- Makes use of a training dataset or test dataset to model expected location of clusters
- Then fits the signals from the experiment data to the clusters and scores them based on distance from the cluster for the genotype call

Quantile Normalisation

- Quantile normalisation aims to correct the technical bias across arrays by adjusting probe intensity distributions
- Approach is to replace the *n*th highest probe intensity value of each array with the mean of the *n*th highest probe intensity values across all arrays.
- End result is to ensure that an arrays highest-intensity probe has the same values across arrays as does the second highest etc.


Density Plot
Density
log2(PM)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
4
6
8
10
12
14
16
18


Density Plot
Density
log2(PM)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
4
6
8
10
12
14
16
18

Human Genome-Wide SNP6.0

- The most recent version of the Affymetrix 6.0 array only makes use of PM probes largely driven by the development of more complex calling algorithms
- Each SNP on the array is interrogated by 6 or 8 PM probes-3 or 4 replicates of the same probe for each of the 2 alleles
- In addition the array also contains about 1 million copy number probes
- This new array design led to the development of the latest genotype calling algorithm Birdseed

Birdseed

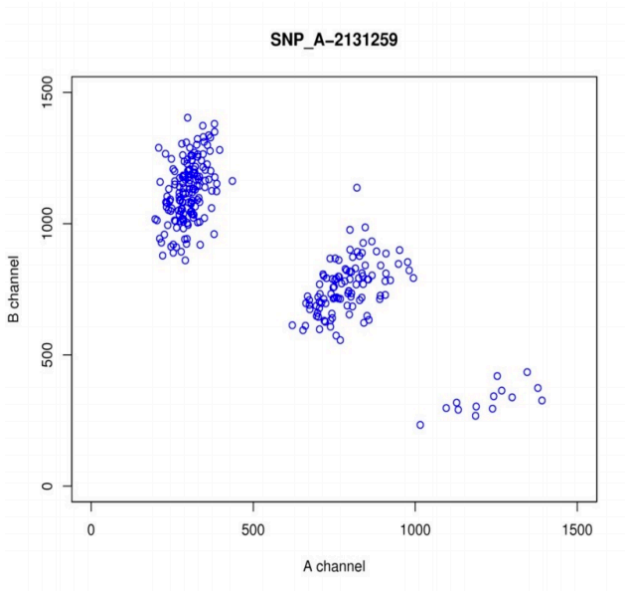
- Birdseed was developed by the Broad institute and selected by Affymetrix as the choice of algorithm for genotype calling on the SNP 6.0 array
- Birdseed is an improvement on the BLRMM algorithm designed specifically for the human SNP 6.0 array

Birdseed

- High level overview
  - Building models from training data for every single SNP on the array
  - Genotype SNPs on sample data using those models

Birdseed – Phase I

- Builds models of all SNPs by using a training dataset (HapMap)

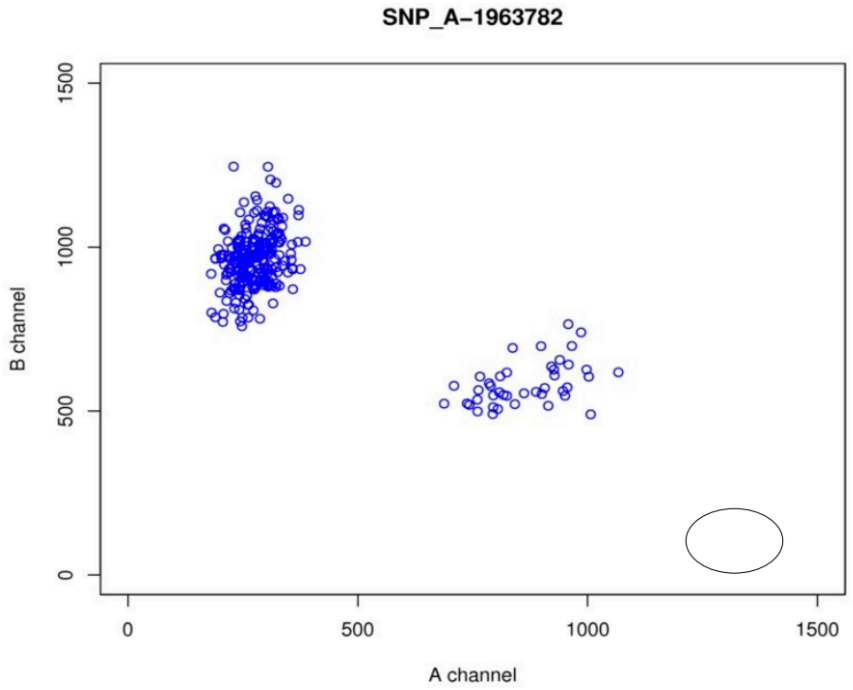


Each SNP can be thought of as a bird. The wingtips are AA and BB, the body is AB. Birds are computed for all SNPs.

AA: 1.1671 0.3133 0.0108 0.0039 0.0028 14
AB: 0.7499 0.7224 0.0056 0.0034 0.0089 102
BB: 0.2852 1.0713 0.0018 0.0019 0.0125 154

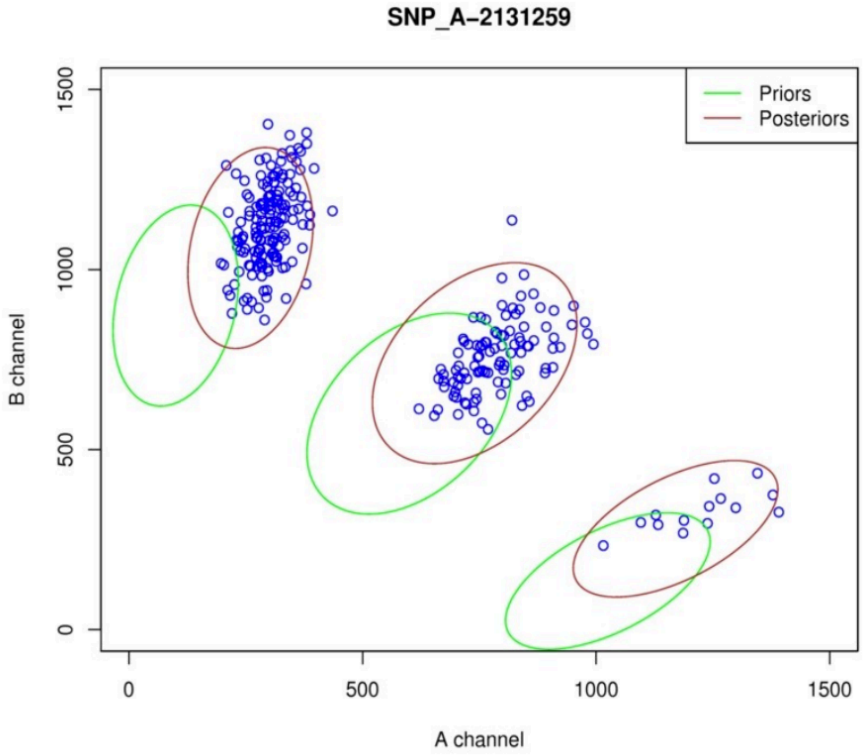
Birdseed – Phase I

- Since not all clusters are present in training data, Birdseed estimates cluster centers and covariance matrices



Birdseed –Phase II

- Birdseed uses a highly customised EM algorithm using the SNP-specific bird as the "seed" and clusters the SNPs
- Birdseed provides a confidence score for every genotype it makes (best 0-1 worst)
- Based on the call's proximity to its cluster



Illumina Platform

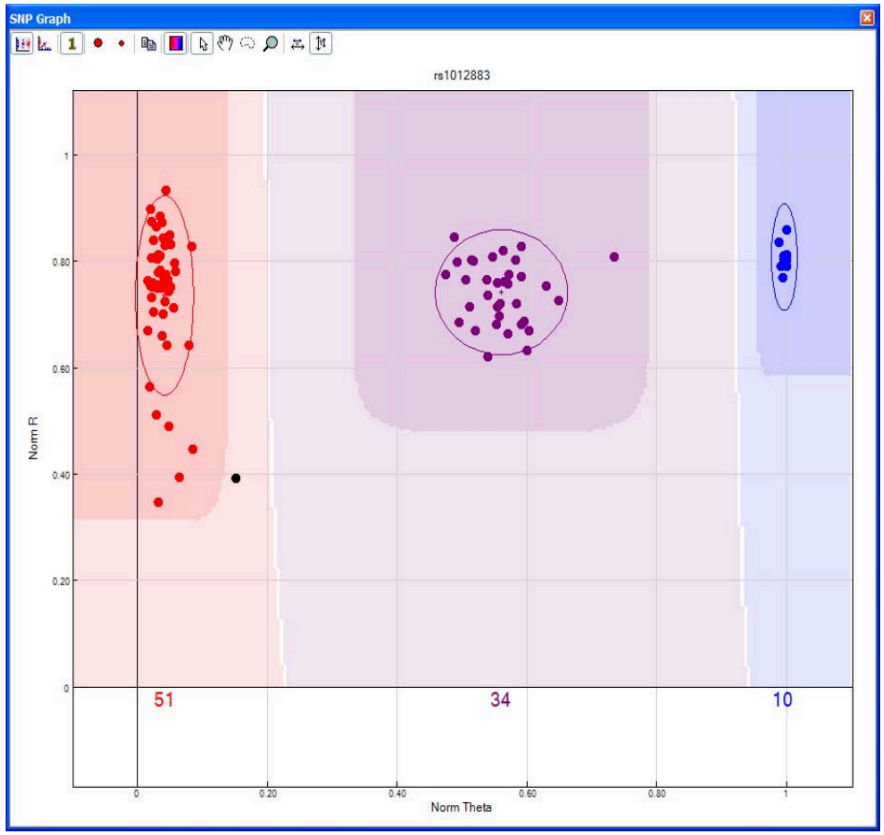
- Like Affymetrix, Illumina platform has gradually increased in capacity
- However, from a data analysis perspective , the array data output format has remained relatively consistent
- One raw measurement for allele A and one raw measurement for allele B at each SNP
- Less algorithmic developments

Illumina Genotyping

- Normalisation of intensities – outlier removal and background estimation – multiple array approach
- Normalised intensities are summarised, such that each SNP is assigned a pair of values corresponding to each allele
- This pair represents the allele intensities and is transformed into polar coordinates

Illumina Genotyping

- The genotyping call is then made based on a reference set of samples
- Similar approach to the BRLMM method
- Option to only make calls based on sample data or training dataset



Genotype Calling QC

- Metrics generated by genotype calling algorithms can be used to identify samples with poor calling rates
- Indication of some technical bias or sample preparation problem
- Samples can be flagged or removed

Affymetrix Power Tools (APT)

Affymetrix Power Tools (apt)

- Set of command line tools developed to process Affymetrix microarray data
- Tools for expression, SNP and CNV arrays
- Start with CEL files for your experiment
- Can carry out genotyping using: BRLMM and Birdseed – dependent on chip used
- Preprocessing step for assessment of data quality

Quality control of SNP6.0 array

- First step is to assess how well allele intensities separate into three clusters
- With previous microarray chips this could be assessed using MM probes
- Essentially DM method
- SNP6.0 array contains a set of 3022 QC probes developed for this assessment
- Did not perform well with samples with bad DNA quality

Quality control of SNP6.0

- DM measures the consistency of PM and MM intensities within each SNP with four possible genotypes (Null, AA, AB, BB)
- DM does not measure the degree to which PM match intensities cluster by genotype
- In high quality samples the A and B allele probe intensities will display three clusters
- In poor quality samples, these clusters will merge

Apt-geno-qc

- Improved method over DM based on PM probe intensities
- Generates two main values
  - Quality call rate (QCR)
  - Contrast quality control (CQC)
- Initially QCR was used only, but was found to underperform in problematic samples
- CQC is now adopted as the more accurate measure of data intensity quality

Apt-geno-qc

- Also makes a gender call for each sample based on the copy number probes on the A and Y chromosome
- Generates a gender estimate based on ratio of chrY to chrX average copy number probe intensities
- Also generated during the genotype calling step
- Can be used to check for sex concordance

CQC

- Measures the separation of allele intensities into three clusters in "contrast space"
- Contrast space is a projection of the 2D allele intensity space into an informative single dimension.
- Based on the Allelic Contrast equation

CQC

$$AllelicContrast = \frac{(A - B)}{(A + B)}$$

- For each SNP, A and B are the medians of the intensities for the replicate probes for the A and B alleles
- Allelic contrast values range from -1 for the ideal BB genotype to +1 for the ideal AA genotype and zero for the ideal AB genotype

CQC

- A high quality sample produces three genotype peaks, separated by two valleys in the histogram of the contrast values
- Poorly resolved clusters mean that genotyping is likely to fail and the sample will have a low call rate

CQC

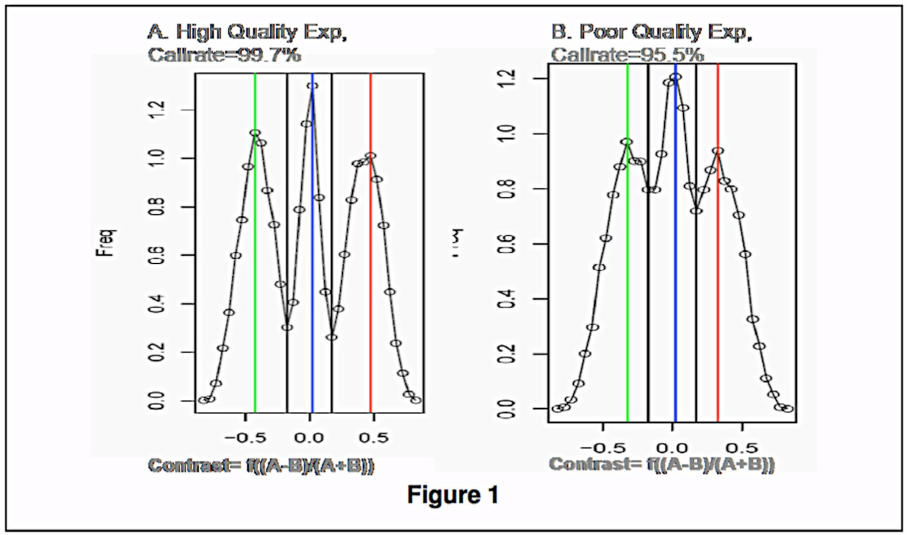
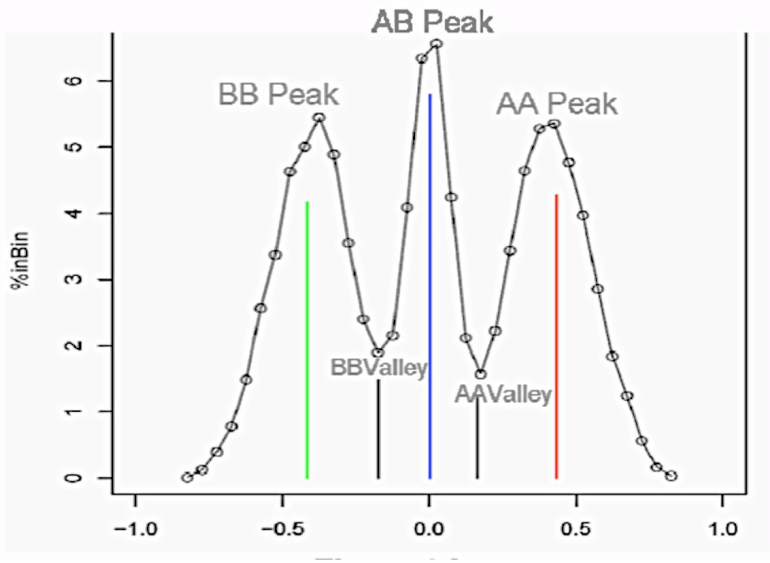

A. High Quality Exp,
Callrate=99.7%
B. Poor Quality Exp,
Callrate=95.5%
Freq
Contrast= f((A-B)/(A+B))
Contrast= f((A-B)/(A+B))

Figure 1


AB Peak
BB Peak
AA Peak
BBValley
AAValley
%inBin

CQC

- The difference between each homozygous peak and its associated valley is calculated and the least of these two values becomes the CQC value
- CQC values are generated for:
  - A random sampling of SNPs from the array
  - Sampling of SNPs only on NSP fragments
  - Sampling of SNPs only on STY fragments

CQC – Best practices for QC

- Best prediction of samples genotyping performance is the random CQC values
- Affymetrix recommends removal/re-running of samples which have a CQC<0.4
- If used in conjunction with QCR remove samples with QCR values less than 0.86
- Perform CQC in batches of samples
- Flag samples where more than 10% of samples in a batch do not pass CQC cutoff
- Flag batches where the mean passing CQC is<1.7

Apt-probeset-genotype

- Following QC of the CEL files and removal of poor quality samples genotype calling can be carried out on the samples
- Choice of algorithms
- Use Birdseed for SNP.6.0 data
- As previously mentioned Birdseed calls genotypes and generates a confidence score as well as an associated call rate for each SNP

Apt-probeset-genotype

- The algorithm produces three files:
  - birdseed-v2.calls.txt
    - Contains base calls for each SNP
  - birdseed-v2.confidence.txt
    - Contains the confidence values for the base calls
  - birdseed-v2.report.txt
    - Contains a summary of the base calling including call rate, heterozygosity rate etc.

Apt-probeset-genotype

- These files can be processed to remove samples with a call rate lower than 97%
- Genotype calling should be re-run once the low performing samples have been removed
- Further utilities are available to convert the Birdseed output files to Plink input files or you may write your own script to do this