Genome Wide Assocation Studies (GWAS)

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

April 23, 2012

Illumina Genotyping

Illumina's genotype technology uses Tag SNPs to genotype millions of markers simultaneously. A tag SNP is a representative single nucleotide polymorphism (SNP) in a region of the genome with high linkage disequilibrium (the non-random association of alleles at two or more loci). It is possible to identify genetic variation without genotyping every SNP in a chromosomal region.

Illumina Genotyping

Chips

Omni Whole-Genome Arrays

BeadChip	Array Format	Markers per Sample
HumanOmni5-Quad	4	~ 4.3 million
HumanOmni2.5S	8	~ 2.5 million
HumanOmni2.8-8	8	~ 2.5 million
HumanOmni1S	8	~ 1.25 million
HumanOmni1-Quad	4	~ 1.1 million
HumanOmniExpress	12	~ 700,000
HumanCytoSNP-12	12	~ 300,000

Species: Human, Mouse, Rat, Bovine, Equine, Porcine, Ovine, Canine

Preprocessing of Data

Data begins as intensity values for each sample and each snp (2-color, one for each candidate allele).
Data is then:

- normalized
- converted to polor coordinates
- genotype clusters are generated
- GenCall score is calculated
- final genotype is determined

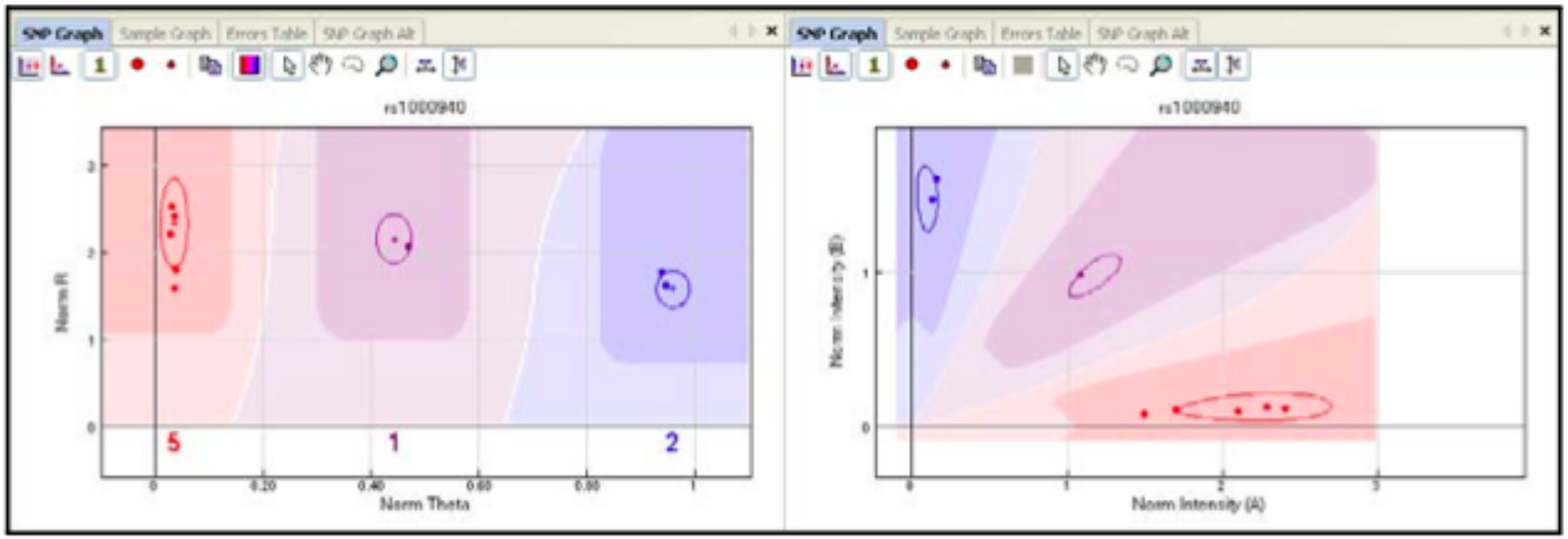
Cartesian and Polar Coordinates

Cartesian coordinates use the X-axis to represent the intensity of the A allele and the Y-axis to represent the intensity of the B allele.

Polar coordinates use the X-axis to represent normalized theta (the angle deviation from pure A signal, where 0 represents pure A signal and 1.0 represents pure B signal), and the Y-axis to represent the distance of the point to the origin (normalized R).

For the Radius (R), the Manhattan distance ($A + B$) is used rather than the Euclidian distance ($sqrt(A * A + B * B)$).

Cartesian and Polar Coordinates



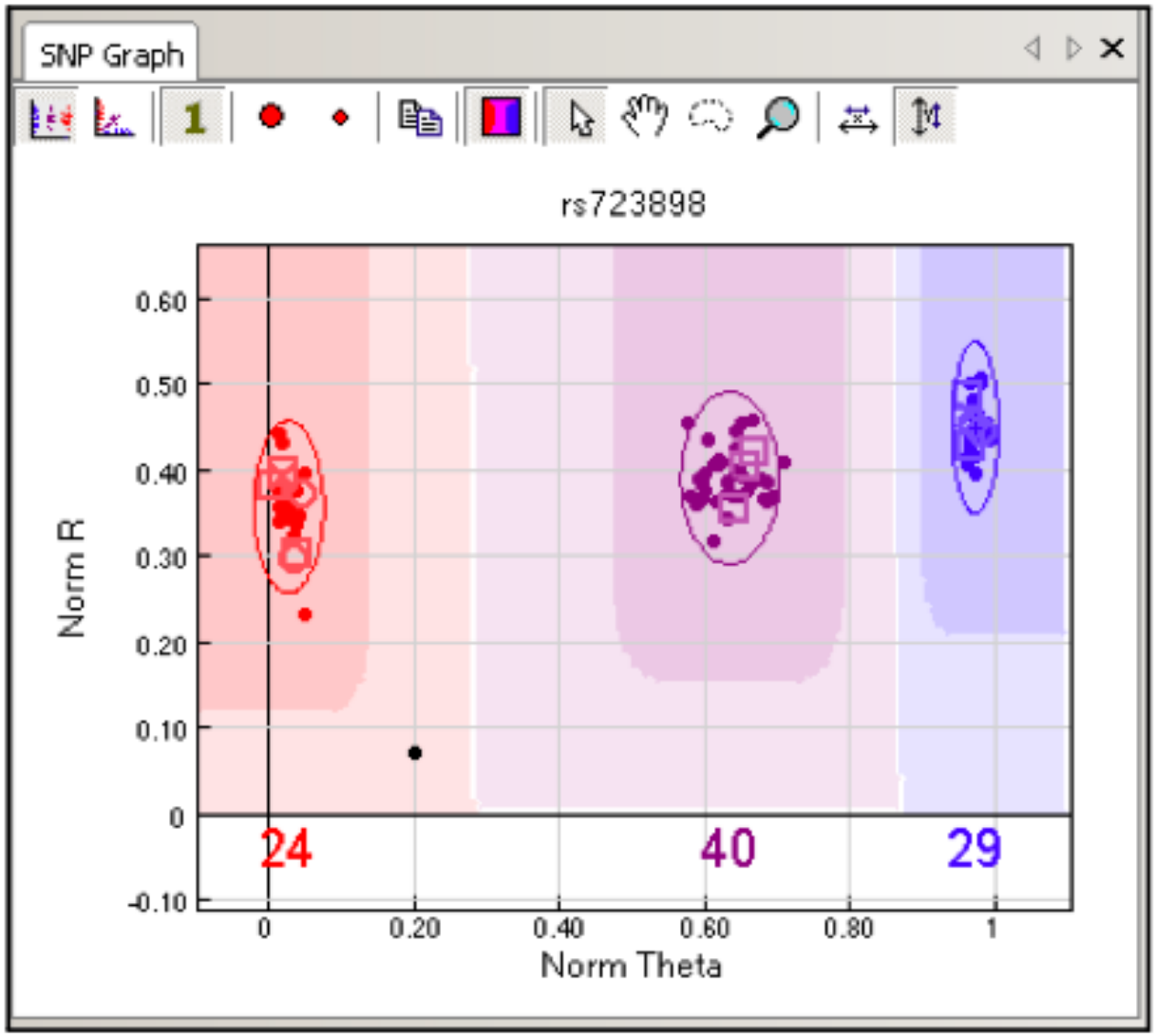
Clustering

Given a population of samples that exhibit three genotypes for every SNP, a clustering algorithm is used to determine the cluster positions of the genotypes. If certain SNPs have one or two clusters that lack representation, the algorithm will estimate the missing cluster positions.

The lower the minor allele frequency, the more samples are required to achieve representation of all clusters. A population of 100 or more samples is typically recommended. The cluster

ovals represent the location of the clusters with two standard deviations.

Clusters


SNP Graph
rs723898
Norm R
0.60
0.50
0.40
0.30
0.20
0.10
0
-0.10
24
40
29
0
0.20
0.40
0.60
0.80
1
Norm Theta

GenCall Score I

GenCall Score is a quality metric that indicates the reliability of each genotype call. The GenCall Score is a value between 0 and 1 assigned to every called genotype. Genotypes with lower GenCall scores are located further from the center of a cluster and have a lower reliability.

GenCall Scores are calculated using information from the clustering of the samples. To get a GenCall Score, each SNP is evaluated based on the following characteristics of the clusters:

- angle
- dispersion overlap
- intensity

GenCall Score II

There is no global interpretation of a GenCall Score, as the score depends on the clustering of your samples at each SNP, which is affected by many different variables including the quality of the samples and the loci.

Final genotype for a sample and SNP pair is then the cluster the sample resides in a whether the sample meets the GenCall Score cutoff.
Illumina recommends that you use a GenCall Score cutoff of 0.15 for Infinium products and 0.25 for GoldenGate products.

Data Format


```
[Header]
GSGT Version 1.9.4
Processing Date 2/13/2012 12:06 AM
Content BovineHD_B.bpm
Num SNPs 777962
Total SNPs 777962
Num Samples 1588
Total Samples 1588
[Data]
SNP Name Sample ID Allele1-Forward Allele2-Forward Allele1-Top Allele2-Top
Allele1-AB Allele2-AB GC Score X Y
ARS-BFGL-NGS-106614 2891 C C C C B B 0.4913 0.041 1.125
ARS-BFGL-NGS-89367 2891 T G A C A B 0.8079 0.757 0.661
BovineHD3100000001 2891 T T A A A A 0.4954 1.297 0.052
BovineHD3100000002 2891 A A A A A A 0.9381 1.225 0.022
BovineHD3100000003 2891 C C G G B B 0.4787 0.012 0.669
```

Cleaning the data

Sample cleaning

- missing genotypes (ie greater than 10%)
- if parents were also genotyped, mendelian inheritance
- Excessive autosomal homozygosity
- Sex misassignment (X heterozygosity)

Marker cleaning

- Minor allele frequency (MAF) < 0.01
- Missing values (ie $> 5\%$)
- Hardy Weinberg expectation (p-value $< 10^{-7}$ in founders)

Batch QA

Check for batch effects

- MAF by batch
- MAF by plate
- MAF by sample type

software

PLINK
http://pngu.mgh.harvard.edu/ purcell/plink/ R packages

http://cran.fhcrc.org/web/views/Genetics.html