GeneSpring GX 11
New Analysis Features:
Genome-Wide Association Studies

Michael Janis
Application Scientist
Agilent Technologies
Introduction
Overview of GWAS analysis

a) Genomic Variations in the form of SNPs and CNVs exist in the human genome.

- Single nucleotide polymorphisms (SNPs):
  - DNA sequence variations that occur when a single nucleotide in genome is altered
  - Variation must occur in at least 1% of the population to be considered a SNP
  - Occur every 100 to 300 bases

- Identification of SNPs with significant association to disease can point to regions of human genome where disease-causing genes may reside

- SNPs may be in direct association with genes that cause disease or may be indirectly associated by linkage disequilibrium
Introduction
Overview of GWAS analysis

a) Genomic Variations in the form of SNPs and CNVs exist in the human genome.
b) These variations could be pathogenic, benign or confer disease susceptibility.
c) SNPs are usually bi-allelic, with possible genotype calls of AA, BB or AB. Within a population, SNPs can be assigned a MAF.
d) Disease risk can be associated primarily either with the dosage of a gene, or with any single allele.
Introduction
Overview of GWAS analysis

a) Genomic Variations in the form of SNPs and CNVs exist in the human genome.
b) These variations could be pathogenic, benign or confer disease susceptibility.
c) SNPs are usually bi-allelic, with possible genotype calls of AA, BB or AB. Within a population, SNPs can be assigned a MAF.
d) Disease risk can be associated primarily either with the dosage of a gene, or with any single allele.

Genome-Wide Association Studies are studies in which a dense set of SNPs across the genome is genotyped to survey the most common genetic variation for a role in disease.
Support for genome-wide association studies
Experimental designs and supported technologies

• Support for population-based case-control studies
  – Categorical phenotypes where case and control designators can be made – *i.e.* disease vs. disease-free

• Support for other population-based studies
  – Quantitative phenotypes where no case and control designations can be made – *i.e.* height

• Support for Affymetrix and Illumina whole-human genotyping arrays

SNP-based

Agilent Technologies
Support for genome-wide association studies
Supported array technologies

**Affymetrix**
- 100K (50K Xba, 50K Hind)*
- 500K (250K Nsp, 250K Sty)
- SNP 5.0
- SNP 6.0
- Technology available from server

*You need at least 6 distinct cel files for Hind or Xba experiments, and 6 file pairs for combined experiments

**Custom**
- Illumina format file input – these could be genotype calls from any other source

**Illumina**
- All SNP arrays
- GeneSpring GX plugin for GenomeStudio must be used to export data in the correct format
- Technology created during experiment creation
Overview of a genome-wide association study analysis

Components of a GWAS:

a) Sample selection
b) Genotyping
c) Statistical tests for associations between the SNPs and the trait.
d) Replication of the identified trait in an independent population.
Overview of a genome-wide association study analysis
GeneSpring Analysis Steps

Components of a GWAS:

a) Sample selection
b) Genotyping
c) Statistical tests for associations between the SNPs and the trait.
d) Replication of identified trait in an independent population.

- Experiment Creation/Genotype Calling
- QC
- Filters
- Stratification Correction
- Statistical Analysis
- Biological Contextualization
Genome-wide association studies
Workflow in GeneSpring GX

Genotype Calling

QC

Filters

Stratification Correction
Statistical Analysis

SNP Tagging
SNP Regression
HTR and LD Analysis
## Genotype calling
On supported array platforms

<table>
<thead>
<tr>
<th>Technology</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix Mapping 50K Xba240</td>
<td>BRLMM</td>
</tr>
<tr>
<td>Affymetrix Mapping 50K Hind240</td>
<td>BRLMM</td>
</tr>
<tr>
<td>100K (50K Xba240 and 50K Hind240)</td>
<td>BRLMM</td>
</tr>
<tr>
<td>Affymetrix Mapping 250K Nsp</td>
<td>Birdseed</td>
</tr>
<tr>
<td>Affymetrix Mapping 250K Sty</td>
<td>Birdseed</td>
</tr>
<tr>
<td>500K (250K Nsp and 250K Sty)</td>
<td>Birdseed</td>
</tr>
<tr>
<td>Affymetrix GenomeWide SNP5</td>
<td>Birdseed</td>
</tr>
<tr>
<td>Affymetrix GenomeWide SNP6</td>
<td>Birdseed</td>
</tr>
<tr>
<td>Illumina (all)</td>
<td>None; genotype calls generated in GenomeStudio</td>
</tr>
<tr>
<td>Custom (Illumina Format)</td>
<td>None; genotype calls are present in the input file</td>
</tr>
</tbody>
</table>
Quality Control
Of samples and SNPs

- Quality Control on Samples
  - Filter Samples by their genotype call rate
  - Sample outlier detection using EIGENSTRAT
  - Metrics from Birdseed (Only for Affy data)

- Quality Control on SNPs
  - Filter SNPs by their genotype call rate
  - Filter by violation of Hardy-Weinberg Equilibrium (HWE)
  - Filter by Minor Allele Frequency (MAF)
Quality Control
On samples

Filter Samples by Missing Values
Birdseed Report—Only for Affymetrix samples
EIGENSTRAT filter on samples
Quality Control
Filter samples by missing values

Samples with high proportion of missing values above the cut-off are highlighted.
Samples with high missing values can be excluded.

Input Parameters:
Specify a cutoff proportion below. Samples for which the proportion of missing genotype calls exceeds this cutoff will be identified and displayed for potential exclusion from further analysis.

Cutoff for Missingness: 0.05

Output views:
Samples for which the proportion of missing genotype calls exceeds the specified cutoff are highlighted in the spreadsheet below. Highlighted samples can be excluded from further analysis by clicking on the "Include/Exclude Samples" button. Additional samples can be manually selected on the spreadsheet for exclusion using Control-Click. Samples excluded earlier can be included back using the "Include/Exclude Samples" button.
Quality Control
Filter samples by missing values

Exclusion of samples should be done with caution, since removing samples at a later stage would remove the EIGENSTRAT Corrected and Allele Frequency data.

Note:
If you Include/Exclude samples from the experiment, then the "EIGENSTRAT Corrected" and "Allele Frequency" data are deleted from the experiment if they exist.
## Quality Control

Birdseed report (Affymetrix Samples)

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Chromosome</th>
<th>Gender</th>
<th>Homozygous rate</th>
<th>Call Rate</th>
<th>Heterozygous rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM374529 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374530 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374531 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374532 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374533 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374534 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374535 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
<tr>
<td>GSM374536 CEL</td>
<td>1</td>
<td>male</td>
<td>0.38044</td>
<td>0.13847</td>
<td>0.86153</td>
</tr>
</tbody>
</table>

Computed Gender

Homozygous rate

Call Rate

Heterozygous rate
Quality Control
EIGENSTRAT filter on samples

Population stratification refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of allele with disease.

Bias introduced into the analysis due to stratification can be handled by EIGENSTRAT.
Quality Control
Sample stratification
Quality Control
EIGENSTRAT filter on samples

Identifies samples which are outliers due to stratification
Outliers can then be excluded from analysis
Quality Control
EIGENSTRAT filter on samples

1. Identification of Principal Component which has the highest Eigen Value

2. Significance of that component is determined by the TW p-value.

3. Examination of the z-scores in the selected component to identify stratification.

4. A z-score of 4 or 5 is considered significant stratification and such sample should be excluded from analysis.
Quality Control
Filters for SNPs

Quality Control on SNPs
Filter SNPs by Missing Values
Filter SNPs by Differential Missingness
Filter SNPs by divergence from HWE
Filter by Minor Allele Frequency (MAF)
Quality Control
Filter SNPs by missing values

Removes SNPs which show missing genotype calls for an unacceptably high proportion of samples.
Quality Control
Identify SNPs with differential missingness

Deleted SNPs would give missing genotype calls.

This filter establishes correlation between missing SNP and disease state, using a Chi-Square test.

Retains the SNPs with a p-value below the specified cutoff.
Quality Control
Filter SNPs by HWE $p$-value

Hardy-Weinberg equilibrium (HWE) is measured for each individual SNP.

HWE represents the population distribution of allele and genotype frequencies such that the distribution is stable from generation to generation.

SNPs showing a strong deviation from the HWE are identified using a Chi-Square test and can be removed.

Many tests assume HWE or perform optimally when HWE is followed.

When a SNP has the same genotype across all the samples, HWE filter automatically sets the $p$-value to 0; the SNP clearly violates HWE and should always be filtered out.
Quality Control
Filter SNPs by HWE $p$-value

At least 2 groups are needed in the interpretation

Parameters to be excluded

SNPs which are not deviating significantly from HWE
Quality Control
Filter SNPs by minor allele frequency (MAF)

Filters out SNPs with Minor Allele Frequency less than a specified value.

MAF is calculated on the given data, thus the B allele from Affymetrix arrays is not automatically considered to be the minor allele.

Since huge allelic imbalances could lead to wrong results, MAF filter is recommended before most of the analysis.
Analysis
Different ways of evaluating association in GeneSpring GX

Which SNPs are associated with the expression of the phenotype being studied?
Analysis  
Test for Association

Key assumption is that observed differences in allele frequencies are due to disease rather than differences in background population between cases and controls

• Draw case and control samples from the same population
• If not, population stratification correction must be applied before performing statistical tests for association
Population Structure or Stratification refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease.
Population Stratification
Different ways of evaluating association in GeneSpring GX

What are the effects of uncorrected population stratification?

- Spurious associations can result due to population structure.

- Any SNP with allele proportions which differ between the subgroup and the general population will be wrongly associated with case or control status.

- Some over represented alleles could actually be causal, but would be swamped by other SNPs that do not have a causal role.
Stratification Correction
Different ways of evaluating association in GeneSpring GX

1. Eigenstrat:
   - Eigenstrat modifies the genotype as well as phenotype.
   - Does not work well if sample size is small
   - Applied BEFORE any statistical tests

2. Genomic Control:
   - Modification of Chi-Square test
   - In stratified samples, the test statistics gets inflated by a constant factor.
     Genomic control measures this factor and downscales all test statistics
Modes of Inheritance

Describes the relationship between the phenotype and the phenotype-associated SNP
Choose the applicable mode of inheritance.

Eignestrat Corrected data will be generated for the Mode of Inheritance chosen.
Stratification Correction on Samples
EIGENSTRAT Correction

Choose only the most significant PC to do the correction

Top PCs display changed
Stratification Correction on Samples
EIGENSTRAT Correction

Corrected datasets for each mode of inheritance can be created and saved.

SNPwise variation: Calculates the difference between the uncorrected and corrected datasets for each SNP and quantifies the correction.

Samplewise variation: Calculates the difference between the uncorrected and corrected datasets for each sample and quantifies the correction.

Corrected along the most significant PC.
Statistical Tests
For Association

Statistical tests:
1. Pearson’s Chi-Square
2. Fisher’s exact
3. Cochran-Armitage
4. Chi-Square correlation

MTC:
1. Benjamini-Yekutieli (BY)
2. Bonferroni
3. Benjamini-Hochberg (BH)
4. Storey’s q-value

Chi-Square correlation is the only test which can run on the Eigenstrat corrected data.
Statistical Tests
For Association

- Genomic Control or EIGENSTRATE can be applied to correct for population stratification
- Specify models of inheritance: Additive, Dominant, Recessive, Custom
- Trait types
  - Nominal
  - Ordinal
SNP Tagging
Using haplotype blocks

A **Tag** SNP is a SNP that is in strong LD with multiple other SNPs such that it can serve as a proxy for these other SNPs on large-scale genotyping platforms.

- Association between 2 alleles located near each other on a chromosome, such that they are inherited together more frequently than expected by chance is called LD.
- The genome is comprised of regions of strong LD, also known as haplotype blocks, interspersed by presumed recombination hot spots or regions of low LD.
- All SNPs with an LD block need not be genotyped, instead the same information can be obtained using Tag SNPs.
- Knowing genotype of Tag SNP allows you to predict the genotype of SNP at the linked loci nearby.
SNP Tagging
Using haplotype blocks

**Tagged** SNP is SNP a SNP that you do not have to genotype because its state can be inferred by one or a combination of tag SNPs

Advantages of using Tag SNPs

- Reduce cost of genotyping by eliminating redundancy in dataset
  - When Tag SNPs are used to reduce genotyping effort in association studies, it is important to know how much power is lost as well as how much power is gained if same number of randomly chosen SNPs were used instead
- Decrease multiple testing issue
SNP Tagging
Using haplotype blocks

Calson’s Greedy algorithm used for identification of Tag SNP

MAF cutoff: eliminates SNPs where allele frequency is too low and will contribute noise to tagging

LD cutoff: $R^2$ values range from 0-1. The square of Pearson’s coefficient between two sets of genotype calls across samples. Measure of correlation between the 2 SNPs.

Maximum distance a non-tag SNP can be away from a tag SNP
### SNP Tagging
**Using haplotype blocks**

#### SNP Tagging (Step 2 of 4)

**Entities passed after Minor Allele Frequency Filter**

<table>
<thead>
<tr>
<th>Name</th>
<th>Minor Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs123456</td>
<td>0.29</td>
</tr>
<tr>
<td>rs789012</td>
<td>0.28</td>
</tr>
<tr>
<td>rs345678</td>
<td>0.27</td>
</tr>
</tbody>
</table>

- **SNPs which have passed the MAF**

#### SNP Tagging (Step 3 of 4)

Here, each row contains a tag SNP in the left column, and the non-tag SNPs (which have passed the MAF filter) which are *represented* by this tag SNP in the right column. Each of the SNPs in the right column has an r-square greater than the LD cutoff with the tag SNP in the left column. The right column SNPs for a row are collectively the 'bin' for the corresponding tag SNP.

You can change the LD cutoff below, which will change the bin results.

#### Tagged SNPs

<table>
<thead>
<tr>
<th>Name</th>
<th>Bin</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10489588</td>
<td>rs10489588, rs110925, rs4587514, rs10492941, rs112213, rs4648462</td>
</tr>
<tr>
<td>rs10752719</td>
<td>rs10752719, rs11081857, rs2887275, rs10492942, rs10492945, rs4131573</td>
</tr>
<tr>
<td>rs1120120</td>
<td>rs1120120, rs7510849, rs304893136, rs10492945, rs10492948, rs1986022</td>
</tr>
<tr>
<td>rs2353169</td>
<td>rs2353169, rs4587514, rs49294935, rs1524909, rs1186332, rs48485335</td>
</tr>
<tr>
<td>rs3789491</td>
<td>rs3789491, rs10492945, rs3789527, rs278020, rs278021, rs1196904</td>
</tr>
<tr>
<td>rs27825</td>
<td>rs27825, rs428274, rs428274, rs428271, rs10489448, rs10489449</td>
</tr>
<tr>
<td>rs10492966</td>
<td>rs10492966, rs10492965, rs10492964, rs10492963, rs1049436, rs2268172</td>
</tr>
<tr>
<td>rs10492978</td>
<td>rs10492978, rs10492979, rs249037, rs249037</td>
</tr>
<tr>
<td>rs4614455</td>
<td>rs4614455, rs4614455, rs4614455, rs4614455, rs4614455</td>
</tr>
<tr>
<td>rs4614458</td>
<td>rs4614458, rs4614458, rs4614458, rs4614458, rs4614458</td>
</tr>
<tr>
<td>rs1049453</td>
<td>rs1049453, rs1049453, rs1049453, rs1049453, rs1049453</td>
</tr>
<tr>
<td>rs556258</td>
<td>rs556258, rs10492987, rs848208, rs1089455, rs296619, rs10495959</td>
</tr>
<tr>
<td>rs10495998</td>
<td>rs10495998, rs10495998, rs10495998, rs10495998, rs10495998</td>
</tr>
<tr>
<td>rs881247</td>
<td>rs881247, rs328704, rs328704, rs10492997, rs10492997, rs10492997</td>
</tr>
<tr>
<td>rs2888820</td>
<td>rs2888820, rs1107887, rs4422957, rs1076824, rs1076824, rs1076824</td>
</tr>
<tr>
<td>rs465548</td>
<td>rs465548, rs465548, rs465548, rs465548, rs465548</td>
</tr>
<tr>
<td>rs2952170</td>
<td>rs2952170, rs10492997, rs692214, rs910253, rs910253, rs910253</td>
</tr>
<tr>
<td>rs6677472</td>
<td>rs6677472, rs6677472, rs6677472, rs6677472, rs6677472</td>
</tr>
<tr>
<td>rs250</td>
<td>rs250, rs10492916, rs10492931, rs1320596, rs669681, rs10492916</td>
</tr>
<tr>
<td>rs398</td>
<td>rs398, rs463363, rs463363, rs463363, rs463363, rs463363</td>
</tr>
<tr>
<td>rs10492019</td>
<td>rs10492019, rs10492019, rs10492019, rs10492019, rs10492019</td>
</tr>
<tr>
<td>rs97834</td>
<td>rs97834, rs2985334, rs2985334, rs2985334, rs2985334</td>
</tr>
</tbody>
</table>

#### Tag SNPs

- **Change Cutoff**

---

**Agilent Technologies**
SNP Regression
Associating the sample genotype with the phenotype

Regression used to identify confounding factors that may give false positive associations if not controlled for

- Independent variable: SNP genotypes
- Dependent variable: trait

Regression can be done using individual SNPs or multiple SNPs

- Qualitative Trait: multiple logistic regression
  - T-statistics p-values reported
- Quantitative Trait: multiple linear regression
  - F-statistics p-values reported
SNP Regression
Associating the sample genotype with the phenotype

Associates sample genotype with phenotype. Association could be of a single SNP or of multiple SNPs

Single SNP regression
Multiple SNP regression
SNP Regression
Associating the sample genotype with the phenotype

Runs multiple linear regression for Quantitative traits and multiple logistic regression for Categorical Traits
Haplotype Inference

Inferring haplotypes from genotype data is called “phasing”

Haplotypes can then be used to test for association to trait
Haplotype Inference

Inferring haplotypes from genotype data is called “phasing”

Haplotypes can then be used to test for association to trait

Multiple haplotypes can fit the genotype.

Haplotypes are clear in case of homozygous SNPs.

Phasing is ambiguous only when the individual is heterozygous at both loci.

Observed Genotype

Inferred Haplotypes

1Homozygous and 1 Heterozygous loci

Homozygous SNPs

Heterozygous SNPs
Haplotypes can then be used to test for association to trait.

Multiple haplotypes can fit the genotype.

Haplotypes are clear in case of homozygous SNPs.

Phasing is ambiguous only when the individual is heterozygous at both loci.

For a given block of $N$ independent bi-allelic SNPs, $2^N$ haplotypes are possible.
Haplotype Trend Regression

- Phasing is first done using HM to identify haplotypes
- Consecutive SNPs are taken for phasing
- Haplotypes with frequencies less than 0.01 are removed
- Remaining haplotypes are tested for association to trait using haplotype trend regression test
Analysis
Haplotype Trend Regression (HTR)

Allows you to identify if a haplotype block is associated with the phenotype
Analysis
Haplotype Trend Regression (HTR)

Only first SNP of haplotype block listed in Probe Set ID column

Quantitative Trait:
Multiple linear regression

Categorical trait:
Multiple logistic regression
Chi-Square statistics
Analysis
Haplotype Trend Regression (HTR)

Generates Haplotype frequencies.

Filters out haplotypes with frequencies <0.01

Regression testing and p-value for each Haploblock
Linkage Disequilibrium (LD)
And haplotype, tag SNPs again

Association of alleles located near each other on a chromosome such that they are inherited together more frequently than expected by chance is called LD.

LD refers to correlations among neighboring alleles, reflecting 'haplotypes' descended from single, ancestral chromosomes and varies between different ethnic groups.

The genome is comprised of regions of strong LD, also known as haplotype blocks, interspersed by presumed recombination hot spots or regions of low LD.

All SNPs with an LD block need not be genotyped, instead the same information can be obtained using Tag SNPs without great loss in accuracy.
Linkage Disequilibrium Plot

LD needs to be calculated first:
Analysis workflow link > LD Analysis

Visualize LD as $R^2$ or D’

Select SNPs with high LD in view and save as Entity List
Visualizing Genotyping Data in Genome Browser

Genotype and allele frequencies can be plotted in Genome Browser. Allele frequency must first be calculated: *Utility > Calculate Allele Frequency*
Biological Contextualization of Association Data
Simulated data for Association Studies

Published January 26, 2010

...it is demonstrated in a simulation study that even those signals that have been detected for common variants could, in principle, come from the effect of rare ones.
Case study of a genome-wide association study
Simulated Dataset in GeneSpring GX

Preselected "significant" SNPs:
   rs2300622 and rs6939678

50000 "background" SNPs.

Experimental grouping:
Cases: Samples 1-50
Controls: Samples 51-100

Stratification:
Population A: Samples 1-35 and 51-75
Population B: Samples 36-50 and 76-100
Time for a demo

Menu Bar – Easy access to common functions

Navigator: Experiments Interpretations, and Entity Lists

Visualization Pane

Workflow: from experiment setup through analysis and biological contextualization
## Genome-wide association study

Benchmarking results in GeneSpring GX

<table>
<thead>
<tr>
<th>Num Samples</th>
<th>Config / Xmx</th>
<th>Expt Creation</th>
<th>EIGEN STRAT</th>
<th>Stat Tests</th>
<th>Tagging</th>
<th>Haplotyping</th>
<th>HTR</th>
<th>SNP Regression</th>
<th>MAF filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer ~546000 SNPs, 1145+1142 Illumina</td>
<td>AMD 64 bit 2 Quad Core Opteron 2.7GHz, 32 GB / 5GB</td>
<td>6hrs</td>
<td>1hr, 40mins</td>
<td>1 hr</td>
<td>1 hr</td>
<td>4 hrs</td>
<td>60 hrs</td>
<td>18 hrs</td>
<td>20 mins</td>
</tr>
<tr>
<td>Schizo 495172 SNPs, 741+751 Illumina</td>
<td>AMD 64 bit 2 Quad Core Opteron 2.7GHz, 32 GB / 10GB</td>
<td>4hrs</td>
<td>37min</td>
<td>-</td>
<td>45min</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&gt;25 mins</td>
</tr>
<tr>
<td>HapMap 180 v6 Samples</td>
<td>Intel Core2 Quad 2.4GHz, 32GB / 1.3 GB</td>
<td>2hr 30 mins</td>
<td>-</td>
<td>15 mins</td>
<td>1 hr</td>
<td>2hr 35 mins</td>
<td>-</td>
<td>70 mins</td>
<td>20 mins</td>
</tr>
</tbody>
</table>
Scalable solution that provides a collaborative space

- Trend towards larger, higher density datasets demands running simultaneous, computationally intensive tasks
- Integrative analysis likely requires expertise from scientists of different disciplines and groups
Thank you!

Customer support: Informatics_Support@agilent.com
Backup Slides
In Phase I, Birdseed builds models of all SNPs by using a training data set (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.

\[
\begin{align*}
AA: & \quad 1.1671 \quad 0.3133 \quad 0.0108 \quad 0.0039 \quad 0.0028 \quad 14 \\
AB: & \quad 0.7499 \quad 0.7224 \quad 0.0056 \quad 0.0034 \quad 0.0089 \quad 102 \\
BB: & \quad 0.2852 \quad 1.0713 \quad 0.0018 \quad 0.0019 \quad 0.0125 \quad 154
\end{align*}
\]

Finny Kuruvilla, MD PhD  
Broad Institute of Harvard and MIT  
Massachusetts General Hospital
Since not all the clusters are present in training data, birdseed estimates cluster centers and covariance matrices.
Birdseed can make highly accurate predictions because it has learned cluster morphology patterns by studying flocks of birds.
In Phase II Birdseed uses a highly customized EM algorithm using the SNP-specific bird as the “seed” (hence the name) & as cluster anchors

Josh Korn and Alec Wysocker
EIGENSTRAT Filter on Samples

- Use EIGENSTRAT to detect sample outliers
- Select significant Principal Component and outlying samples with 5 std dev
## Association Tests

<table>
<thead>
<tr>
<th></th>
<th>Allele A</th>
<th>Allele B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td><strong>A total</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>B total</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Allele frequencies in patients with disease are compared to those without disease.

Statistical test applied to determine whether these frequencies are different.
EIGENSTRAT Correction for Population Stratification

Chose significant Principal Component to perform correction

Chose model of inheritance for correction

- EIGENSTRAT correction must be performed for each model that you want to use for association testing (i.e. you cannot use EIGENSTRAT corrected data using Dominant Model and then chose Recessive for association testing)

If samples are removed from experiment after EIGENSTRAT population stratification correction, the correction has to be run again on the new sample set.