Biostatistics 666
Statistical Models in Human Genetics

Gonçalo Abecasis
goncalo@umich.edu

Today

- Course Logistics
- Genetics: The Basics
- Hardy Weinberg Equilibrium

Course Logistics

Scheduling

Office Hours

Class Notes

Grading

Course Objective

- Provide an understanding of statistical models used in gene mapping studies
- Survey commonly used algorithms and procedures in genetic analysis

Course Notes

- We will not be using a textbook
  - Extremely important to attend class, and ask questions as needed!
- Copies of slides and additional content available online at
  - http://genome.sph.umich.edu/wiki/666

Assessment

- In most years, grades is combination of:
  - 6-10 home work assignments (40%)
  - 2 in-class written assessments (60%)
- This year, there is an opportunity to try something different …

Potential Group Projects

- Evaluate strategies for selecting cases and controls for inclusion in a genetic study
- Evaluate standard analysis tools for next generation sequence data
- Systematically review design and analysis features of successful sequencing studies

Academic Integrity

- All assignments you submit for evaluation must represent your own work.
- If you copy or paraphrase other work, you must clearly mark these sections and indicate sources.
- Cheating, plagiarism and aiding and abetting these acts constitutes academic misconduct and is a serious offense.
- See also the School policy on academic conduct.

Academic Integrity

- All assignments are made on an individual basis – the solutions you provide must be your own work.
- Cheating, plagiarism and aiding and abetting these acts constitutes academic misconduct and is a serious offense.
- See also the School policy on academic conduct.

In a set of assignments or exams that is broadly identical, each will be scored as zero and referred to Department or School.

Scheduling

- We will try to start classes at 8:30 sharp.
  - Due to prior commitments, I may have to miss several lectures and starting sharply on time should allow us to make up for any lost time.

Office Hours

- Please cross out times for which you are unavailable in the sheet going around
- Room 4614
  School of Public Health Tower

Course Contents

Brief Overview

Genetic Mapping

"Compares the inheritance pattern of a trait with the inheritance pattern of chromosomal regions"

Positional Cloning

"Allows one to find where a gene is, without knowing what it is."

Some of the Topics Covered

- Maximum Likelihood
- Modeling Genes in Populations
- Relating Genes to Phenotypes

Modeling Genes in Populations

- Hardy Weinberg Equilibrium
- Linkage Disequilibrium
- The Coalescent
- Methods for Haplotyping
- Methods for Handling Short Read Sequence Data
- Segregation of Variants in Pedigrees

Modeling Relationship Between Genes and Phenotypes

- Introduction to Genetic Linkage Analysis
- Genetic Wide Association Testing
- Modeling Population Structure
- Evaluating the Consequences of Rare Variants

3 Common Questions

- Are there genetic variants influencing a trait?
  - Epidemiological studies
- Where are those variants located?
  - Linkage analysis
- What are those variants?
  - Association analysis

Is a trait genetic?

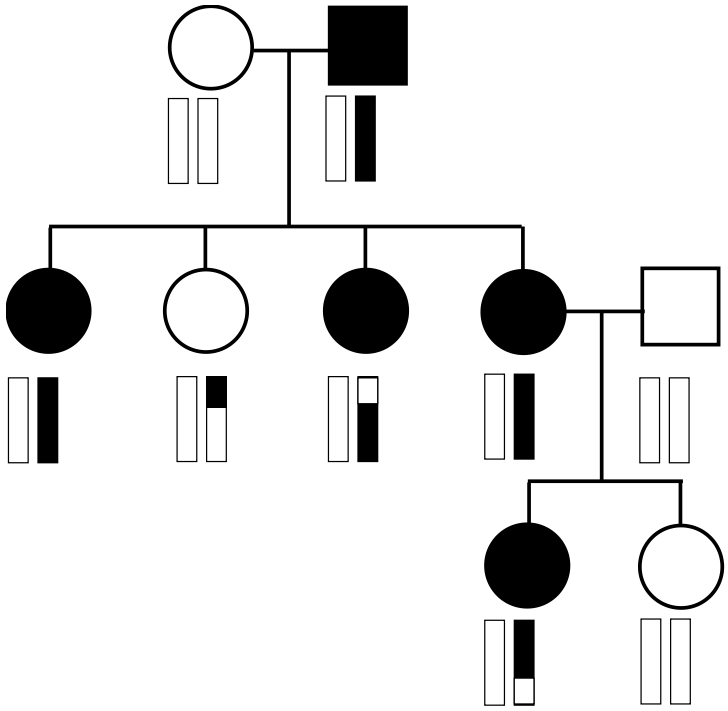
- Examine distribution of trait in the population and among relatives

- E.g. Inflammatory Bowel Disease (Crohn's)
  - General population
    - 1-3 cases per 1,000 individuals
  - Twins of affected individuals
    - 44% of monozygotic twins also have Crohn's
    - 3.8% of dizygotic twins also have Crohn's

Where are those genes?

- Find genetic markers that co-segregate with disease

- A portion of chromosome 16 co-segregates with Crohn's in many families



What are those genes?

- Identify genetic variants that are associated with disease…

- E.g. Mutations which disrupt NOD2 are much more common in Crohn's patients

	Crohn's	Controls
Arg702Trp:	11%	4%
Gly908Arg:	4%	2%
Leu1007fs	8%	4%

A Very Short Primer on Genetic Variation

Biostatistics 666

DNA – Information Store

- Encodes the information required for cells and organisms to function and produce new cells and organisms.

- DNA variation is responsible for many individual differences, some of which are medically important.

Human Genome

- Multiple chromosomes
  - 22 autosomes
    - Present in 2 copies per individual
    - One maternally and one paternally inherited copy
  - 1 pair of sex chromosomes
    - Females have two X chromosomes
    - Males have one X chromosome and one Y chromosome
- Total of ~3 x 10^9 bases (each A, C, T or G)

Inheritance of DNA

- Through recombination, a new “DNA string” is formed by combining two parental DNA strings
- Thus, each chromosome we carry is a mosaic of the two chromosomes carried by our parents
- Only a small number of changeovers between the two parental chromosomes
  - On average ~1 per Morgan (~10^8 bases)
- Copying of DNA sequences is imperfect and, for typical sequences, the error rate is about 1 per 10^8 bases copied

Human Variation

- Every chromosome is unique …
- … but when two chromosomes are compared most of their sequence is identical
- About 1 per 1,000 bases differs between pairs of human chromosomes

DNA Sequences That Vary…

- Genes (protein coding sequences, which total <2% of all DNA)
  - ~20,000-25,000 in humans
- Pseudogenes
  - Ancient genes, inactivated through mutation
- Promoters and Enhancers
  - Sequences which control gene expression
- Repeat DNA
  - Often more variable than other types of sequences
  - Historically useful for tracking DNA through families or populations
- Packaging sequences, “spacer” DNA, etc.

Hardy Weinberg Equilibrium

Biostatistics 666

Remainder of this Lecture …

- Properties of alleles in a population

- Allele frequencies
- Genotypes frequencies
- Hardy-Weinberg equilibrium

Alleles

- Alternative forms of a particular sequence
- Each allele has a frequency, which is the proportion of chromosomes of that type in the population

Allele Frequency Notation

- For two alleles
  - Usually labeled p and $q = 1 - p$
- For more than 2 alleles
  - Usually labeled p_A, p_B, p_C ...
  - … subscripts A, B and C indicate allele name

Genotype

- The pair of alleles carried by an individual
  - If there are **n** alternative alleles …
  - … there will be **n(n+1)/2** possible genotypes
- **Homozygous Genotype**
  - Genotype where the two alleles are in the same state
- **Heterozygous Genotypes**
  - Genotype where the two alleles are in different states

Genotype Frequencies

- Since alleles occur in pairs, these are a useful descriptor of genetic data …
- However, in any non-trivial study we might have a lot of frequencies to estimate ...
- $p_{AA}, p_{AB}, p_{AC}, \ldots p_{BB}, p_{BC}, \ldots p_{CC}$ …

The simple part …

- Genotype frequencies lead to allele frequencies…
- For example, for two alleles:
  - $p_A = p_{AA} + ½\ p_{AB}$
  - $p_B = p_{BB} + ½\ p_{AB}$
- Fortunately, the reverse is also possible!

Hardy-Weinberg Equilibrium

- Random union of gametes
- Relationship described in 1908
  - Hardy, British mathematician
  - Weinberg, German physician
- Shows **n** allele frequencies determine **n(n+1)/2** genotype frequencies
  - Large populations

Required Assumptions

- Diploid, sexual organism
  - Non-overlapping generations
- Autosomal locus
- Large population
- Random mating
- Equal genotype frequencies among sexes
- Absence of natural selection

Random Mating: Mating Type Frequencies

Mating	Frequency
$A_1A_1 * A_1A_1$	
$A_1A_1 * A_1A_2$	
$A_1A_1 * A_2A_2$	
$A_1A_2 * A_1A_2$	
$A_1A_2 * A_2A_2$	
$A_2A_2 * A_2A_2$	
Total	1.0

Mendelian Segregation: Offspring Genotype Frequencies

		Offspring		
Mating	**Frequency**	**A_1A_1**	**A_1A_2**	**A_2A_2**
$A_1A_1 * A_1A_1$				
$A_1A_1 * A_1A_2$				
$A_1A_1 * A_2A_2$				
$A_1A_2 * A_1A_2$				
$A_1A_2 * A_2A_2$				
$A_2A_2 * A_2A_2$				

And now…

$$
\begin{aligned}
p'_{11} &= p_{11}^2 + p_{11}p_{12} + \tfrac{1}{4}\,p_{12}^2 \\
&= (p_{11} + \tfrac{1}{2}\,p_{12})^2 \\
&= p_1^2 \\
p'_{22} &= p_{22}^2 + p_{22}p_{12} + \tfrac{1}{4}\,p_{12}^2 \\
&= (p_{22} + \tfrac{1}{2}\,p_{12})^2 \\
&= p_2^2 \\
p'_{12} &= 2\,p_{11}p_{22} + p_{11}p_{12} + p_{12}p_{22} + \tfrac{1}{2}\,p_{12}^2 \\
&= 2(p_{11} + \tfrac{1}{2}\,p_{12})(p_{22} + \tfrac{1}{2}\,p_{12}) \\
&= 2\,p_1 p_2
\end{aligned}
$$

Conclusion

- Genotype frequencies are function of allele frequencies
  - Equilibrium reached in one generation
  - Independent of initial genotype frequencies
  - Random mating, etc. required
- Conform to binomial expansion
  - $(p_1 + p_2)^2 = p_1^2 + 2p_1p_2 + p_2^2$

Simple HWE Exercise

- If the defective alleles of the cystic fibrosis (CFTR) gene have a cumulative frequency of 1/50 what is:
  - The proportion of carriers in the population? (These are individuals with one defective allele)
  - The proportion of affected children at birth? (These are individuals with two defective alleles)

A few more notes…

- Extends to multiple alleles
  - Expand $(p_1 + p_2 + p_3 + \cdot \cdot \cdot + p_k)^2$
- Frequency of A/A homozygotes is p_A^2
- Frequency of A/B heterozygotes is $2p_Ap_B$
- Holds in almost all human populations
  - Little inbreeding (typical F = ~0.005)

Something to think about…

- Why would inbreeding matter?

Checking Hardy-Weinberg Equilibrium

- A common first step in *any* genetic study is to verify that the data conforms to Hardy-Weinberg equilibrium

- Deviations can occur due to:
  - Systematic errors in genotyping,
  - Unexpected population structure,
  - Presence of homologous regions in the genome,
  - Association with trait in case-control studies.

- Which of these causes would you expect to increase the proportion of heterozygotes?

Testing Hardy Weinberg Equilibrium

- Consider a sample of $2N$ alleles

- n_A alleles of type A
- n_B alleles of type B

- n_{AA} genotypes of type AA
- n_{AB} genotypes of type AB
- n_{BB} genotypes of type BB

Simple Approach

- Calculate allele frequencies and expected counts
- Construct chi-squared test statistic
- Convenient, but can be inaccurate, especially when one allele is rare

A Better Approach: Exact Test of Genotypic Proportions

- Iterate over all possible outcomes and sum probabilities of outcomes with equal or lesser probability
- One sided tests are also possible
- Approach analogous to Fisher's exact test for contingency tables

$$P_{HWE} = \sum_{n^*_{AB}} I\left[P(N_{AB} = n_{AB} \mid N, n_a) \geq P(N_{AB} = n^*_{AB} \mid N, n_a)\right] P(N_{AB} = n^*_{AB} \mid N, n_a)$$

Probability of n_{AB} Heterozygotes

- The number of potential arrangements with n_{AB} heterozygotes is $2^{n_{AB}} N!/(n_{AA}!\, n_{AB}!\, n_{BB})$
- The number of potential arrangements for 2N alleles is $(2N)!/(n_A!\, n_B!)$
- The ratio of these two quantities gives $P(N_{AB} = n_{AB} | N, n_A)$

Probability of n_{AB} Heterozygotes

- To reconstruct formula, first calculate:
  - Possible rearrangements for *2N* alleles
  - Possible rearrangements with n_{AB} heterozygotes

$$P(N_{AB} = n_{AB} \mid N, n_A) = \frac{2^{n_{AB}}\, N!}{n_{AA}!\, n_{AB}!\, n_{BB}!} \cdot \frac{n_A!\, n_B!}{(2N)!}$$

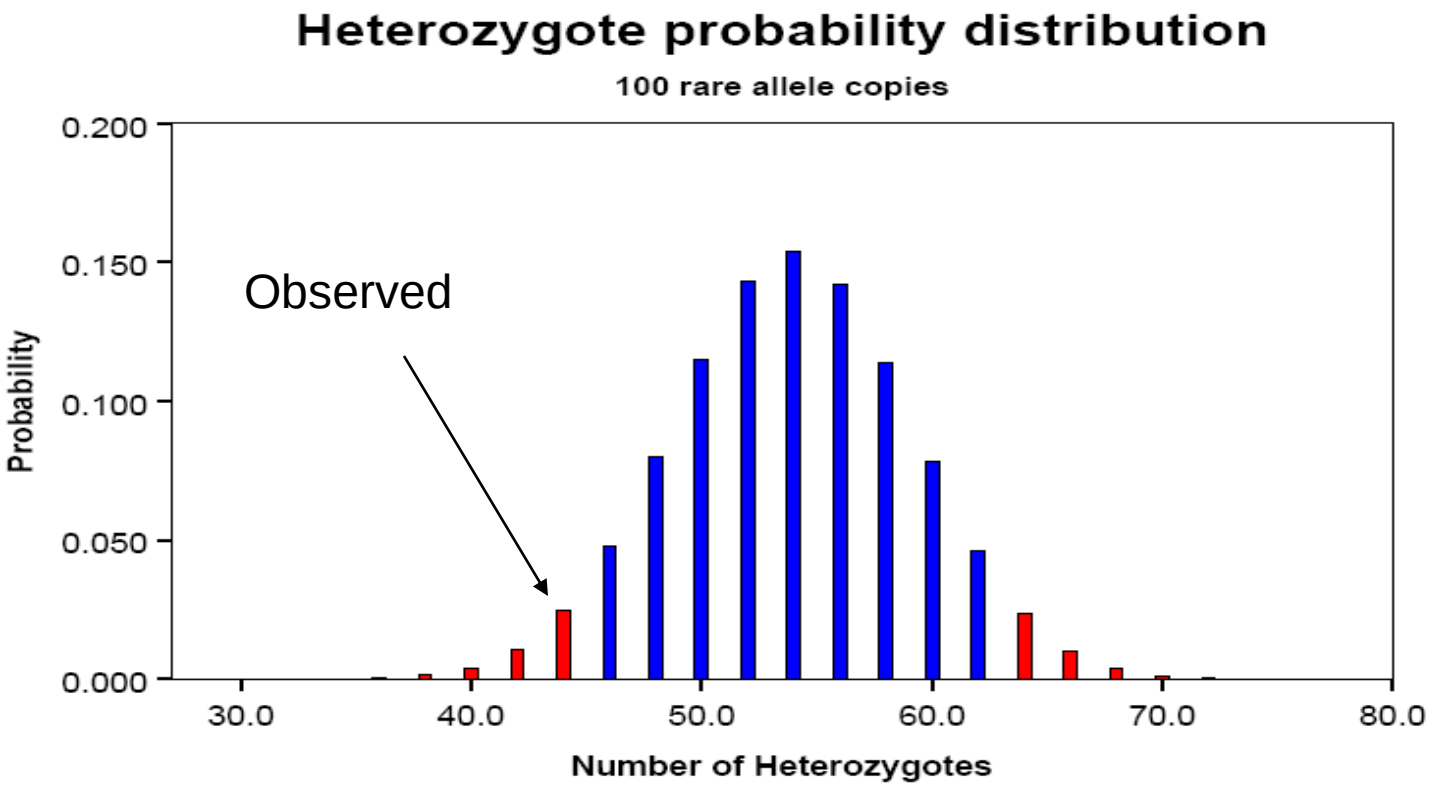
- Calculation can be carried out efficiently in recursive fashion

Recursion for $P(N_{AB} = n_{AB} \mid N, n_A)$

- $$P(N_{AB} = n_{AB} + 2 | N, n_A) = P(N_{AB} = n_{AB} | N, n_A) \frac{4 n_{AA} n_{BB}}{(n_{AB} + 1)(n_{AB} + 2)}$$

- $$P(N_{AB} = n_{AB} - 2 | N, n_A) = P(N_{AB} = n_{AB} | N, n_A) \frac{n_{AB}(n_{AB} - 1)}{4(n_{AA} + 1)(n_{BB} + 1)}$$

Exact Test


Heterozygote probability distribution
100 rare allele copies
0.200
0.150
0.100
0.050
0.000
Probability
Observed
30.0
40.0
50.0
60.0
70.0
80.0
Number of Heterozygotes

Comparison of Test Statistics

Possible Sample Configurations and Their Probabilities for a Sample of 100 Individuals and 21 Minor-Allele Copies Are Tabulated

No. of Heterozygotes (n_{AB})	Probability[a]	χ^2 Test P	Exact Test P Values		
			P_{HWE}	P_{high}	P_{low}
5	<.000001	<.000001[b]	<.000001[b]	1.000000	<.000001[b]
7	.000001	<.000001[b]	.000001[b]	1.000000	.000001[b]
9	.000047	<.000001[b]	.000048[b]	.999999	.000048[b]
11	.000870	.000039[b]	.000919[b]	.999952	.000919[b]
13	.009375	.002228[b]	.010293[b]	.999081	.010293[b]
15	.059283	.045180[b]	.069576	.989707	.069576
17	.214465	.342972	.284042	.930424	.284042
19	.406355	.906529	1.000000	.715958	.690396
21	.309604	.244336	.593645	.309604	1.000000

Note.—The probability of observing each possible outcome is given, together with the corresponding P values for tests of HWE based on the χ^2 statistic and on the exact test statistics P_{HWE}, P_{low}, and P_{high} (described in the main text).

[a] $P(n_{AB}|N = 100, n_A = 21)$.

[b] Configurations that would be rejected at the significance level $\alpha = 0.05$.

Comparison of Type I Error Rates

Actual Error Rates for the χ^2 Test Statistic and the P_{HWE} Test Statistic for Nominal Significance Level $\alpha = 0.01$ or 0.001

Sample and Minor-Allele Count	$\alpha = 0.01$[a]		$\alpha = 0.001$[a]	
	χ^2	P_{HWE}	χ^2	P_{HWE}
$N = 1{,}000$				
1–100	.0208[b] (.0208)[b]	.0039 (.0039)	.0088[b] (.0088)[b]	.0004 (.0004)
101–200	.0100 (.0154)[b]	.0065 (.0052)	.0017[b] (.0053)[b]	.0006 (.0005)
201–400	.0097 (.0126)[b]	.0083 (.0067)	.0010 (.0032)[b]	.0008 (.0006)
401–1,000	.0100 (.0110)[b]	.0090 (.0081)	.0010 (.0018)[b]	.0009 (.0008)
$N = 100$				
1–10	.0292[b] (.0292)[b]	.0024 (.0024)	.0114[b] (.0114)[b]	.0001 (.0001)
11–20	.0191[b] (.0242)[b]	.0035 (.0030)	.0035[b] (.0074)[b]	.0003 (.0002)
21–40	.0083 (.0162)[b]	.0037 (.0033)	.0016[b] (.0045)[b]	.0004 (.0003)
41–100	.0099 (.0124)[b]	.0072 (.0057)	.0009 (.0023)[b]	.0006 (.0005)

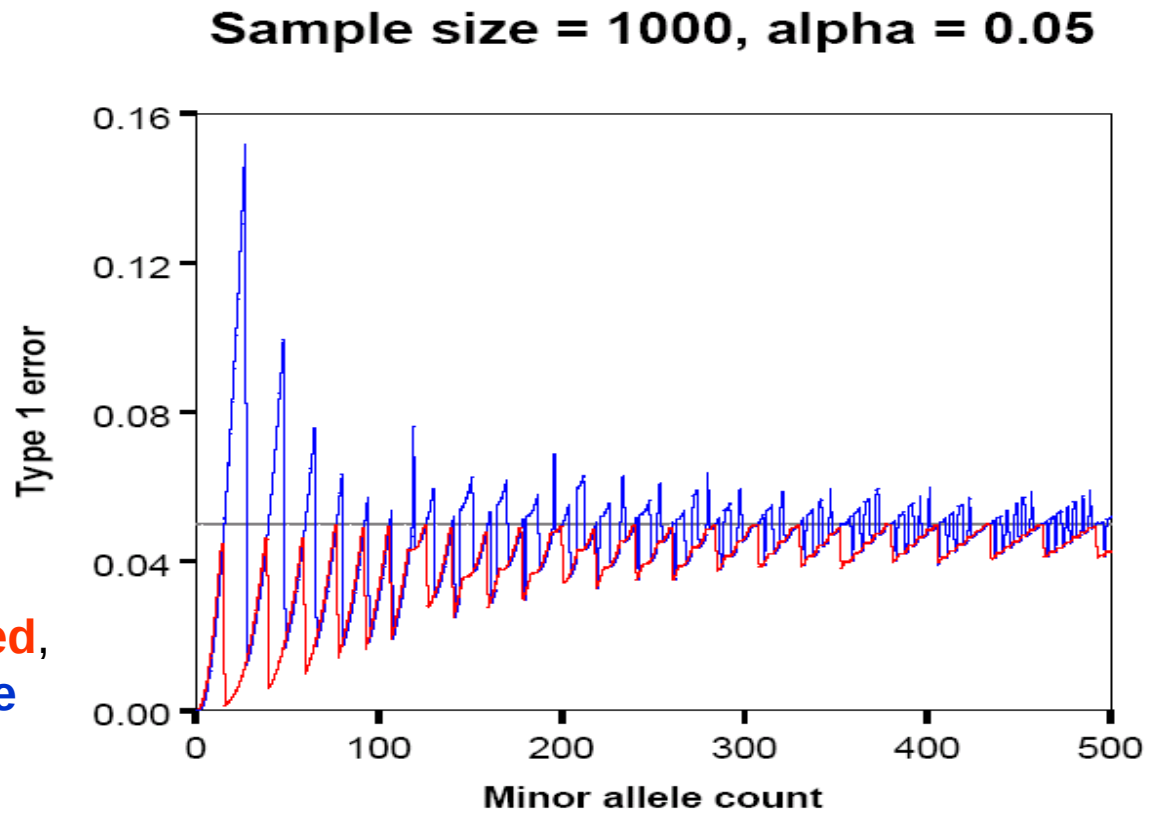
Note.—Results are tabulated for samples of 100 and 1,000 individuals and represent simple averages for each range of minor-allele counts.

[a] The error rate for each bin is tabulated, followed by the cumulative error rate in parenthesis. The cumulative error rate is calculated by including each bin and all previous bins. For example, for a sample of size 1,000, when $\alpha = 0.001$, the type I error rate for the standard χ^2 test in a sample with 101–200 copies of the minor allele is 0.0017 and the cumulative error rate, corresponding to samples with 1–200 copies of the minor allele, is 0.0053.

[b] Exceeds nominal significance level.

Type I Error Rate Is Periodic!

Exact test in **red**,
χ^2 test in **blue**



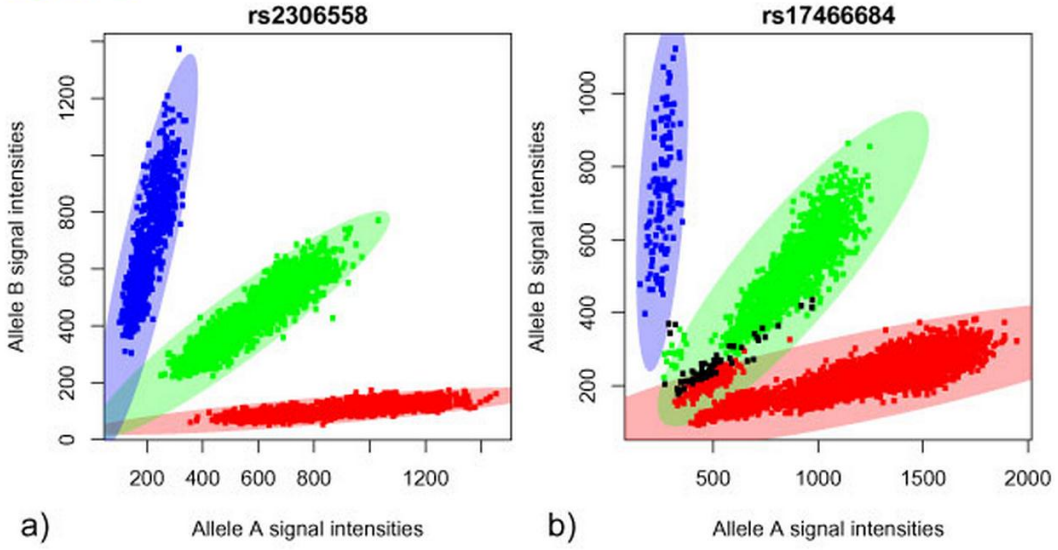
Poor Genotype Calling …

- … is the most common real life cause of deviations from HWE
- Most modern studies exclude markers that fail HWE tests …
- In genomewide association studies, thresholds of $p < 10^{-3}$ to 10^{-6} are common

Example of Good & Bad Genotype Calling

More Genotype Calling Examples

Figure 1.



Examples of cluster plots. Cluster plots for two SNPs. One spot corresponds to one sample. Samples with genotypes AA and BB are red and blue, respectively. Heterozygous samples are shown in green; samples with missing genotypes are black. The ellipses represent the cluster boundaries as computed by ACPA. a, A SNP with no samples in overlapping ellipses; b, red samples lie in the green ellipse. At the bottom of the green ellipse, samples have been erroneously classified as red samples.

Schillert *et al. BMC Proceedings* 2009 **3**(Suppl 7):S58

Summary

- Hardy Weinberg Equilibrium holds in most human population samples
- Deviations from HWE can indicate population structure or natural selection, but – most often – are genotyping artifacts
- Exact tests for Hardy Weinberg equilibrium provide accurate results and are typically recommended

Recommended Reading I

- Wigginton et al. (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**:887-93

Recommended Reading II

- An introduction to important issues in genetics:
  - Lander and Schork (1994) *Science* **265:**2037-48
- Paper was written >15 years ago, well before the human genome was sequenced
- Now, studies made easier by:
  - Availability of reference genome sequence
  - Much improved genotyping and sequencing technologies

Important Issues to Consider

- What are the specific challenges of genetic studies in humans?
- What are common strategies for improving the power of a genetic study?
- How can we combine different strategies to achieve cost-effective studies?