Genes in Populations: Hardy Weinberg Equilibrium

Biostatistics 666

Previous Lecture: Primer In Genetics

- How information is stored in DNA
- How DNA is inherited
- Types of DNA variation
- Common designs for Genetic studies

Recommended Reading

- Lander and Schork (1994)
  Genetic Dissection of Complex Traits
  Science **265:**2037-48

- Paper was written >10 years, well before the human genome was sequenced

- Now, studies are facilitated by widespread availability of genetic sequence data
  (for an example, see http://genome.ucsc.edu/)

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?

Always Check Data Quality!

Results

Our analysis of the pedigree structures by means of the genotypes generated as part of the genome scan highlighted that, in each of the ethnic groups, there were individuals identified as males that were likely to be females (and vice versa), half siblings labeled as full siblings, and pedigree members that showed no relationship to their supposed pedigree. Given that not all of the parents were available for study, it was difficult to distinguish between parental errors and blood- or DNA-sample mixups. In summary, 24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree and were possibly blood-sample mixups. The percentages were consistent across all ethnic groups. In total, 212 individuals were removed from the pedigrees to eliminate these errors.

Study results are only as good as the data available and – as this example shows – things can go wrong!

Genomewide Search for Type 2 Diabetes Susceptibility Genes in Four American Populations

Margaret Gelder Ehm,[1] Maha C. Karnoub,[1] Hakan Sakul,[2,*] Kirby Gottschalk,[1] Donald C. Holt,[1] James L. Weber,[3] David Vaske,[3,†] David Briley,[1] Linda Briley,[1] Jan Kopf,[1] Patrick McMillen,[1] Quan Nguyen,[1] Melanie Reisman,[1] Eric H. Lai,[1] Geoff Joslyn,[2,‡] Nancy S. Shepherd,[1] Callum Bell,[2,§] Michael J. Wagner,[1] Daniel K. Burns,[1] and the American Diabetes Association GENNID Study Group[1]

Today ...

- 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
- **Homozygotes**
  - The two alleles are in the same state
- **Heterozygotes**
  - The two alleles are different

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:
  - $p_A = p_{AA} + ½\ p_{AB}$
  - $p_B = p_{BB} + ½\ p_{AB}$
- However, the reverse is also possible!

Hardy-Weinberg Equilibrium

- Random union of games
- 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?
  - The proportion of affected children at birth?

A few more notes…

- Extends to multiple alleles
  - Expand $(p_1 + p_2 + p_3 + \cdots + p_k)^2$
- 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 is analogous to Fisher's exact test for contingency tables

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

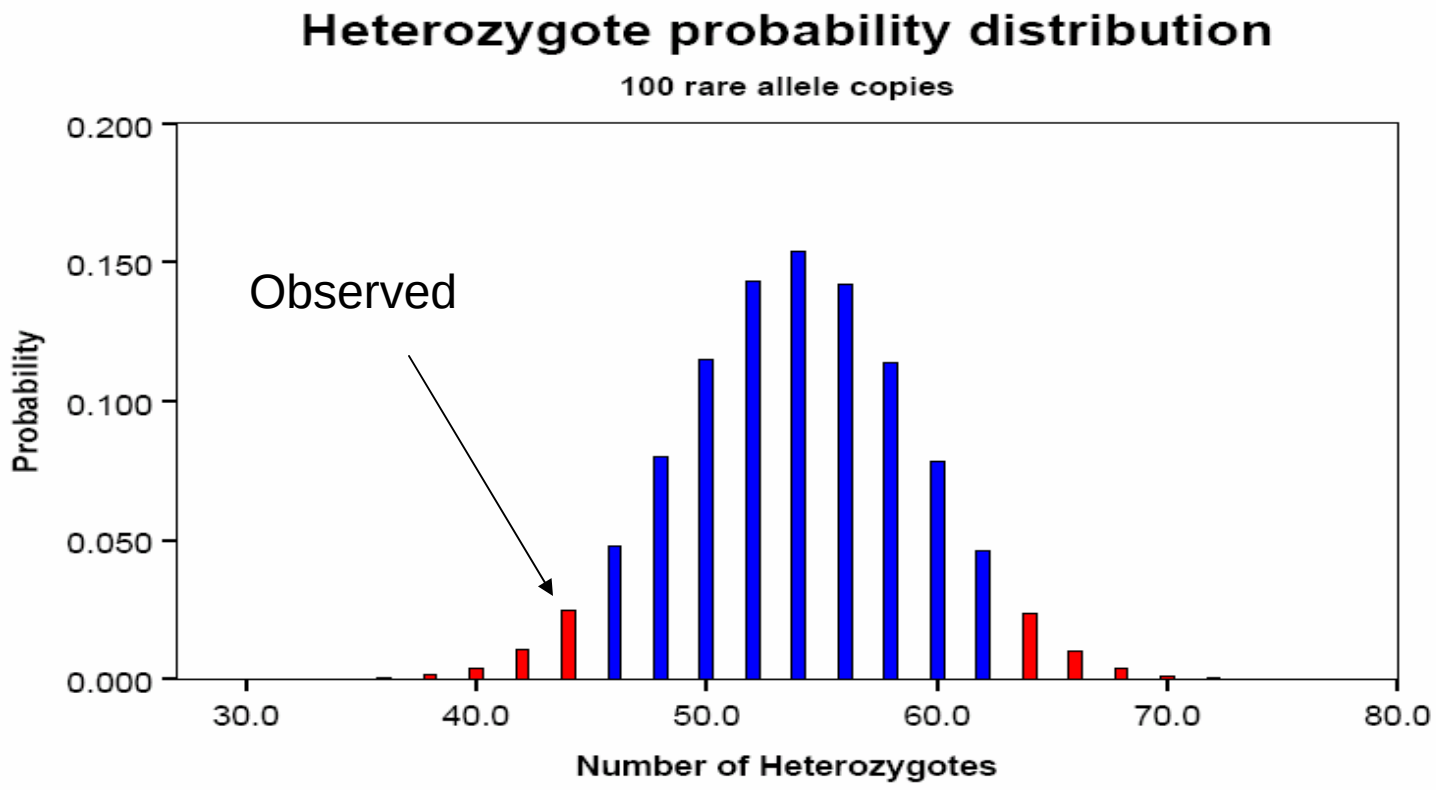
Probability of n_{AB} Heterozygotes

- 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

Exact Test


Heterozygote probability distribution
100 rare allele copies
Probability
0.200
0.150
0.100
0.050
0.000
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)

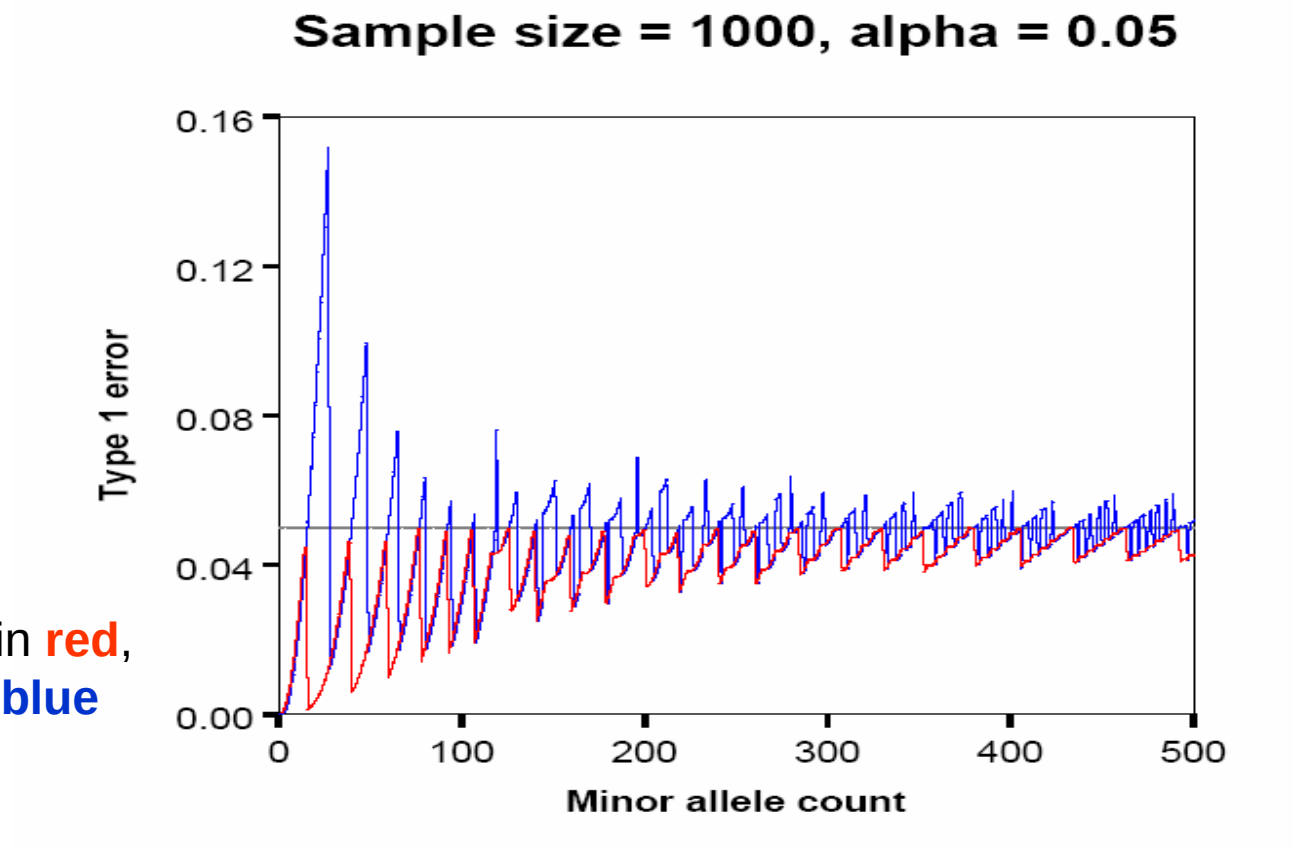
NOTE.—Results are tabulated for samples of 100 and 1,000 individuals and represent simple averages for each range of minor-allele counts.

[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**



Recommended Reading

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

Reading for Next Lecture

- Cardon and Bell (2001) Association study designs for complex diseases. *Nature Reviews Genetics* **2**:91-99

- Surveys important issues in analyzing population data.

- Defines linkage disequilibrium.