# Biophysics 101

Section 5,

October 21, 2003

## Population Genetics

Tom Patterson

Slightly modified by Chih Long Liu

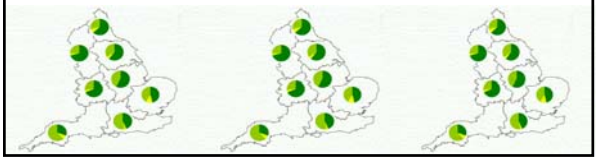
Studying genes in real populations.

Creating and analyzing mathematical models of genes in populations.

## Section 5 topics:

**•Allele distributions, Hardy-Weinberg, Chi-test, and Bonferroni correction.**

**•Recombination, linkage measurement, and haplotypes.**



## Announcements

- First half of PS2 now returned; 1 week appeal period
- Second half of PS2 should be returned around Fri/Sat (10/24-10/25)
- Final Project Idea submissions were due yesterday 10/21
  - Submissions will be compiled for the section and emailed to all of you or posted on the course website
  - Final project presentation videotaping option
- Final Project Proposals due Tues. 11/4
- PS3 Perl section is hard; start now if not already!

## Population Genetics Problems

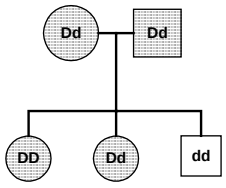
According to molecular anthropology's latest results, it seems that Basques settled in Europe with the first Homo Sapiens and that they lived side by side with Neanderthal men. They would thus be the most direct descendants of the Stone Age artists who, about 20000 years ago, have painted the Lascaux and Altamira caves : Basques may thus be the oldest West European population.

http://perso.club-internet.fr/mcteguy/baskhise.html

•An unusually large proportion of the Basque people have Rh$^-$ blood type. Was this caused by selection pressure, perhaps from the mountainous region the Basque live in, or is it the result of population isolation and random drift?

•Can modern genetic profiling help anthropologists understand the ancient migration history of the Basques?


## RHD genetics




•Lets illustrate some analytical population genetics principles with RHD.

•Lets assume RHD is caused by one, recessive allele.

Phenotypes

Rh- blood type

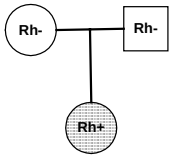
Rh+ blood type

Alleles

D: normal allele, active antigen.

d: mutant allele, inactive antigen

## What (we assume) won't happen:




…but why might this happen for RhD, or for phenotypes in general?

1. Back mutation (very rare).
2.
3.

## Some RhD biology

RefSeq **Summary:** The Rh blood group system is the second most clinically significant of the blood groups, second only to ABO. It is also the most polymorphic of the blood groups, with variations due to deletions, gene conversions, and missense mutations. The Rh blood group includes this gene which encodes the RhD protein and a second gene which encodes both the RhC and RhE antigens on a single polypeptide. The two genes are found in a cluster which includes a third unrelated gene on chromosome 1. The classification of Rh-positive and Rh-negative individuals is determined by the presence or absence of the highly immunogenic RhD protein on the surface of erythrocytes. Alternative splicing of this gene results in two transcript variants encoding two different isoforms.

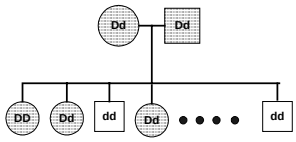
The function of the RhD antigen was first discovered in 1990 by doing an alignment of the RhD sequence against the yeast genome!

The yeast homolog is a membrane protein involved in the transport of ammonium (NH4+) ions, and this activity was later confirmed for the human RhD membrane protein in erythrocytes (red blood cells).

The RhD protein may allow erythrocytes to absorb toxic ammonium (NH4+) ions from body tissues and transport them to detoxifying organs.

RhD is non-essential, since Rh- individuals often lack it entirely.

## What if this family had 100 children….




What's the expected number of…

Rh+ phenotypes: _____

Rh- phenotypes: _____

DD genotypes: _____

Dd genotypes: _____

dd genotypes: _____

## In reverse…

Suppose 16% of a population is Rh-

What's the expected percentage of…

$Rh^+$ phenotypes: ___%

$Rh^-$ phenotypes: ___%

DD genotypes: ___%

Dd genotypes: ___%

dd genotypes: ___%

*Are we making any assumptions here?*

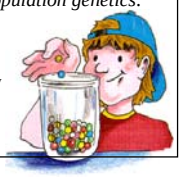
## Know the p's and q's of allele frequencies.

*A simple, and useful metaphor for alleles in population genetics:*

For a given marker, represent each allele by a different colored marble. Have everyone in the population add two marbles to a jar according to their genotype. If the alleles are uniformly distributed in the population:

•The frequency of an allele in the population should match the frequency of it's marble in the jar.

•The frequency of a genotype in the population should match the frequency of selecting its corresponding two marbles from the jar.



•Let p represent the frequency of $Rh^+$ marbles in the jar.

•Let q represent the frequency of $Rh^-$ marbles.

Then if the allele distribution in the population is uniform:

$Rh^+$ frequency: $p^2 + 2pq$

$Rh^-$ frequency: $q^2$

DD frequency: _____

Dd frequency: _____

dd frequency: _____

## Thus if 16% of the population is $Rh^-$…

And if the alleles were uniformly distributed in the population, a condition known as Hardy-Weinberg equilibrium…

q = p =

$Rh^+$ phenotypes: _____

$Rh^-$ phenotypes: _____

DD genotypes: ____

Dd genotypes: ____

dd genotypes: ____

## Suppose we genotype 1000 people, 160 of whom are Rh-

…and we get the following results:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Genotyping results:** | | | |
| 2 | | | | |
| 3 | **DD** | **Dd** | **dd** | **Total** |
| 4 | **397** | **443** | **160** | **1000** |
| 5 | | | | |
| 6 | | | | |
| 7 | **…but if we make our assumption, we should get:** | | | |
| 8 | | | | |
| 9 | **DD** | **Dd** | **dd** | **Total** |
| 10 | **360** | **480** | **160** | **1000** |
| 11 | | | | |
| 12 | **CHITEST(A4:B4,A10:B10):** | | **0.99%** | |
| 13 | **CHITEST(A4:C4,A10:C10):** | | **3.59%** | |
| 14 | | | | |
| 15 | | | | |

**Which of these formulas is correct?**

Either way, we have exceeded the 95% confidence threshold that our population is not in Hardy-Weinberg equilibrium for the D allele.

## What's the allele frequency?

*hint:*

p =

q =

## The H-W status of our RhD data, take 2

Expected DD =

Expected Dd =

Expected dd =

| | A | B | C | D |
|---|---|---|---|---|
| | Genotyping results: | | | |
| | DD | Dd | dd | Total |
| | 397 | 443 | 160 | 1000 |
| | | | | |
| | DD | Dd | dd | Total |
| | | | | 1000 |
| | CHITEST(A4:C4,A10:C10): | | | |

## Given the following facts, what would you expect the allele distribution of RhD to be like?

- RhD is a powerful antigen – a membrane protein in the walls of red blood cells that readily triggers the production of antibodies.
- $Rh^-$ blood can be given to $Rh^+$ subjects.
- If $Rh^+$ blood is given to $Rh^-$ subjects, a strong, often fatal immune reaction results.
- Maternal and fetal circulatory systems are well isolated from each other, but there is often enough exposure to fetal blood during delivery to trigger the formation of antibodies that can have an adverse effect on future pregnancies.

How might the above facts change the genotype frequencies in a population?

___ DD ___ Dd ___ dd

## We genotyped 100 markers on chromosome 1 near RhD in our subject pool, and upon (correct) H-W testing, we found four with significant ($\alpha < 5\%$) non-random allele distributions.

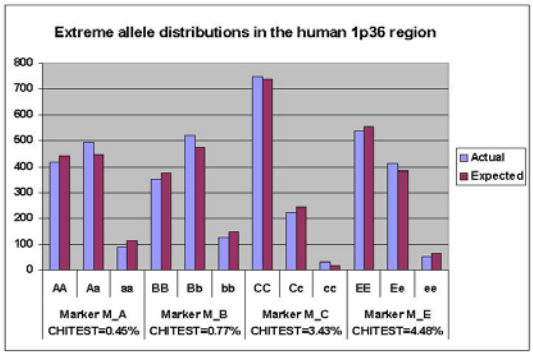
| Marker M_A | | | |
|---|---|---|---|
| Genotyping results: | | | |
| AA | Aa | aa | Total |
| 417 | 493 | 90 | 1000 |
| | | | |
| ...assuming H-W eq, we should get: | | | |
| AA | Aa | aa | Total |
| 440 | 447 | 113 | 1000 |
| | CHITEST: | 0.45% | |

| Marker M_B | | | |
|---|---|---|---|
| Genotyping results: | | | |
| BB | Bb | bb | Total |
| 353 | 521 | 126 | 1000 |
| | | | |
| ...assuming H-W eq, we should get: | | | |
| BB | Bb | bb | Total |
| 376 | 474 | 149 | 1000 |
| | CHITEST: | 0.77% | |

| Marker M_C | | | |
|---|---|---|---|
| Genotyping results: | | | |
| CC | Cc | cc | Total |
| 747 | 223 | 30 | 1000 |
| | | | |
| ...assuming H-W eq, we should get: | | | |
| CC | Cc | cc | Total |
| 737 | 243 | 20 | 1000 |
| | CHITEST: | 3.43% | |

| Marker M_E | | | |
|---|---|---|---|
| Genotyping results: | | | |
| EE | Ee | ee | Total |
| 537 | 412 | 51 | 1000 |
| | | | |
| ...assuming H-W eq, we should get: | | | |
| EE | Ee | ee | Total |
| 552 | 382 | 66 | 1000 |
| | CHITEST: | 4.48% | |

## Here are the results in chart form




…are you ready to publish???

## How to use the Bonferroni correction for multiple hypothesis testing.

In the RhD example, we used the standard 5% significance level. In other words, the chance of getting our extreme results due to sampling error is less than 5%. In the lingo of statistical testing, 5% is our alpha value.

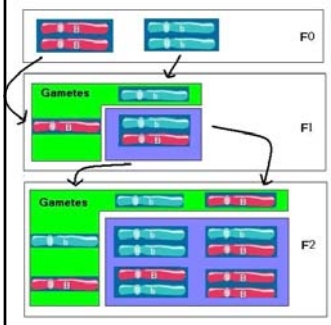
Since we sampled 100 times, chances are we will get a few random hits. The conservative way to adjust for multiple hypotheses, sample sets, or tests is to divide alpha by n, the number of tests:

$$\alpha_{new} = \alpha_{old}/n = 5\%/100 = 0.05\%$$

As you can see, none of our extreme alleles were extreme to that significance, so our findings are negative.

http://mathworld.wolfram.com/BonferroniCorrection.html

## One marker cross




Genotype frequencies for F2

| | B (50%) | b (50%) |
|---|---|---|
| B (50%) | BB(25%) | Bb (25) |
| b (50%) | Bb (25%) | bb (25%) |

## Two Marker Cross, No Linkage

Two markers, M_A and M_B. M_A is on chromosome 3 , and M_B is on chromosome 9.


## Two Markers, no linkage

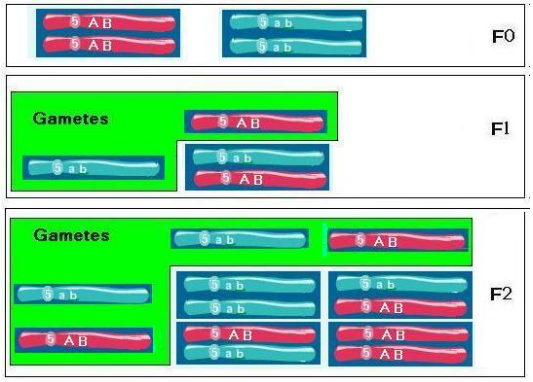
| | | Genotype | Gametes | | | | Progeny |
|---|---|---|---|---|---|---|---|
| F0 | Mother | AABB | AB 100% | | | | (F1) AaBb 100% |
| | Father | aabb | ab 100% | | | | |
| F1 | Mother | AaBb | AB 25% | Ab 25% | aB 25% | ab 25% | (F2) see below |
| | Father | AaBb | AB 25% | Ab 25% | aB 25% | ab 25% | |

| | AB (25%) | Ab (25%) | aB (25%) | ab (25%) |
|---|---|---|---|---|
| AB (25%) | AABB (6.25%) | AABb (6.25%) | AaBB (6.25%) | AaBb (6.25%) |
| Ab (25%) | AABb (6.25%) | AAbb (6.25%) | AaBb (6.25%) | Aabb (6.25%) |
| aB (25%) | AaBB (6.25%) | AaBb (6.25%) | aaBB (6.25%) | aaBb (6.25%) |
| ab (25%) | AaBb (6.25%) | Aabb (6.25%) | aaBb (6.25%) | aabb (6.25%) |

F2 Progeny

| AABB | AABb | AAbb | AaBB | AaBb | Aabb | aaBB | aaBb | aabb |
|---|---|---|---|---|---|---|---|---|
| 6.25% | 12.50% | 6.25% | 12.50% | 25.00% | 12.50% | 6.25% | 12.50% | 6.25% |

## Two markers, linked, no recombination




## Two Markers, complete linkage

| | | Genotype | Gametes | | Progeny |
|---|---|---|---|---|---|
| F0 | Mother | AABB | AB 100% | | (F1) AaBb 100% |
| | Father | aabb | ab 100% | | |
| F1 | Mother | AaBb | AB 50% | ab 50% | (F2) see below |
| | Father | AaBb | AB 50% | ab 50% | |

| | AB | ab |
|---|---|---|
| AB | AABB (25%) | AaBb (25%) |
| ab | AaBb (25%) | aabb (25%) |

F2 Progeny

| AABB | AaBb | aabb |
|---|---|---|
| 25.00% | 50.00% | 25.00% |

## Two Markers, linked, with recombination.

## Two Markers, linkage, 1% recombination

| | | Genotype | Gametes | | | | Progeny |
|---|---|---|---|---|---|---|---|
| F0 | Mother | AABB | AB 100% | | | | (F1) AaBb 100% |
| | Father | aabb | ab 100% | | | | |
| F1 | Mother | AaBb | AB 49% | Ab 1% | aB 1% | ab 49% | (F2) see below |
| | Father | AaBb | AB 49% | Ab 1% | aB 1% | ab 49% | |

| | AB (49%) | Ab (1%) | aB (1%) | ab (49%) |
|---|---|---|---|---|
| AB (49%) | AABB (24%) | AABb (0.49%) | AaBB (0.49%) | AaBb (24%) |
| Ab (1%) | AABb (0.49%) | AAbb (0.01%) | AaBb (0.01%) | Aabb (0.49%) |
| aB (1%) | AaBB (0.49%) | AaBb (0.01%) | aaBB (0.01%) | aaBb (0.49%) |
| ab (49%) | AaBb (24%) | Aabb (0.49%) | aaBb (0.49%) | aabb (24%) |

F2 Progeny

| AABB | AABb | AAbb | AaBB | AaBb | Aabb | aaBB | aaBb | aabb |
|---|---|---|---|---|---|---|---|---|
| 24.01% | 0.98% | 0.01% | 0.98% | 48.02% | 0.98% | 0.01% | 0.98% | 24.01% |

## Measuring Linkage Disequilibrium

Consider two Markers M_A, and M_B, each with two neutral alleles (A,a) and (B,b). If the alleles are uniformly distributed in the population:

| | allele | A | a |
|---|---|---|---|
| allele | freq. | pA | 1-pA |
| B | pB | pApB | (1-pA)pB |
| b | 1-pB | pA(1-pB) | (1-pA)(1-pB) |

Single allele frequencies
Joint allele frequecies

Often, there is a surplus or deficit of (AB, ab) or (Ab, aB) genotypes, measured as:

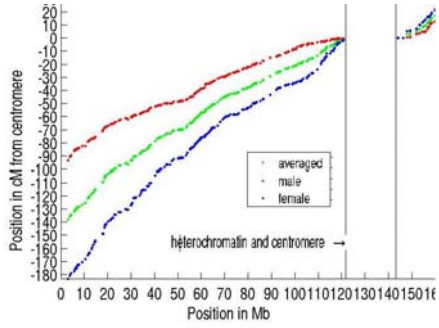
$$D = \tfrac{1}{2}(p_{AB} + p_{ab} - p_{Ab} - p_{aB})$$

The magnitude of the "linkage disequiblibrium" or LD is often masked by low allele frequencies. The measure R is an attempt to correct for that:

$$R = \frac{D}{\sqrt{p_A p_B (1-p_A)(1-p_B)}}$$

## Human chromosome 1: linkage vs. physical distance

According to this chart, is the recombination rate greater for males or females?



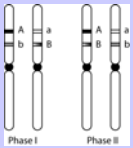

http://www.molgen.mpg.de/~service/projects/humangen/goldenPath/mapPlots/Jun24/male_female/

## Haplotype and Phase, Revisited

**haplotype** – from "haploid genotype," a set of linked DNA changes along a chromosome.

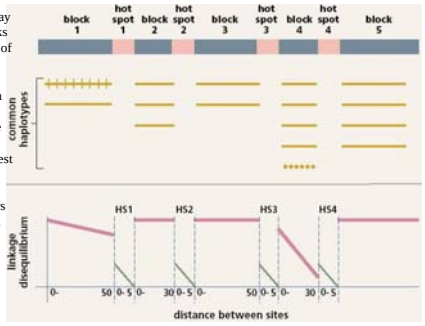
**cis and trans linkage** – in Phase I, alleles Ab and aB are cis-linked (coupling), while they are trans-linked (repulsion) in Phase II.

Phase I and Phase II describe phase of linkage of the alleles of two genes that lie on the same chromosome.




## Haplotype blocks

There is an effort underway to exploit haplotype blocks to improve the efficiency of genetic testing in disease research. Large, distinct haplotype blocks, that can be identified with few marker alleles will reduce the amount of genomic analysis, since one need test only enough markers to determine the haplotype, and the remaining markers will follow automatically.




http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v29/n2/full/ng1001-109.html&filetype=pdf

## Next Week

- Clustering