Population Genetics I
Genomics: Bio5488, Spring 2022

Sheng Chih (Peter) Jin, Ph.D.
Assistant Professor of Genetics and Pediatrics
Washington University School of Medicine
jin810@wustl.edu

(with thanks to Nancy Saccone, Don Conrad and slides from past years)
What is population genetics?

• Very broadly: The science of genetic variation in populations of organisms

• Origin, amount, frequency, distribution of this variation in space and time

• Focus on human population genetics
Why study (human) population genetics?

• Demographic inference
  - Our DNA records history of humankind: population size changes, migrations, etc.

• Functional inference
  - E.g., alleles deleterious to "fitness" are unlikely to be common

• Complex disease
  - What analysis approaches are appropriate
  - What methods can leverage population history

• Especially now in the era of sequencing
  - You will have your own genome sequence
Outline for today

• Hardy-Weinberg

• Forward models: Wright Fisher model

• Decay of heterozygosity

• Backward models: Coalescent
Hardy-Weinberg Law

Goal: describe the relationship of allele and genotype frequencies in a population that follows Mendel's law of segregation

Under certain assumptions, an equilibrium is maintained
Some preliminaries:
Locus with 2 alleles: $A_1, A_2$

$p = \Pr(A_1)$
$q = \Pr(A_2)$

$p + q = 1$

If we know the genotype frequencies in a population sample
- $\Pr(A_1A_1)$, $\Pr(A_1A_2)$, $\Pr(A_2A_2)$ - we can calculate the allele frequencies:

\[
p = \Pr(A_1) = 1 \times \Pr(A_1A_1) + 0.5 \times \Pr(A_1A_2) + 0 \times \Pr(A_2A_2)
\]

\[
q = 1 - p = 0 \times \Pr(A_1A_1) + 0.5 \times \Pr(A_1A_2) + 1 \times \Pr(A_2A_2)
\]

What about the reverse?
Hardy-Weinberg Law

What about the reverse? If we know the allele frequencies, can we calculate the genotype frequencies?

In general, allele frequencies do not uniquely determine genotype frequencies:

\[ p = 0.5, \ q = 0.5 \] can correspond to:

\[
\begin{align*}
Pr(A_1A_1) &= 0.25 & \text{OR} & \quad Pr(A_1A_1) &= 0.5 \\
Pr(A_1A_2) &= 0.5 & \quad \quad \quad & \quad Pr(A_1A_2) &= 0 \\
Pr(A_2A_2) &= 0.25 & \quad \quad \quad & \quad Pr(A_2A_2) &= 0.5
\end{align*}
\]
Under certain assumptions, **Hardy-Weinberg equilibrium** holds: that is, the probabilities of the three genotypes ($A_1A_1$, $A_1A_2$, $A_2A_2$) are $p^2$, $2pq$, and $q^2$, respectively.

**HW law (under the assumption below):**
1) HWE is established after 1 generation of random mating.
2) HWE, once established, is maintained over the generations.

**Assumptions:**

a. (Infinitely) large population
b. Discrete generations
c. **Random mating**
d. No selection, no migration, no mutation
e. Equal initial genotype frequencies in the two sexes

Note also we take genotypes to be unordered.
**Hardy-Weinberg equilibrium**

To see 2): Suppose HWE holds. Then

<table>
<thead>
<tr>
<th>Paternal</th>
<th>Maternal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A_1$ ($p$)</td>
</tr>
<tr>
<td>$A_1$ ($p$)</td>
<td>$A_1A_1$ ($p^2$)</td>
</tr>
<tr>
<td>$A_2$ ($q$)</td>
<td>$A_2A_1$ ($qp$)</td>
</tr>
</tbody>
</table>

**Why is $p$ the prob of a parent transmitting $A_1$?**

<table>
<thead>
<tr>
<th>Parental genotype</th>
<th>Prob of having this genotype</th>
<th>Prob of transmitting ‘A1’</th>
<th>Joint probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1$</td>
<td>$p^2$</td>
<td>1</td>
<td>$p^2$</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$2pq$</td>
<td>0.5</td>
<td>$pq$</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$q^2$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>$q^2$</strong></td>
<td><strong>0</strong></td>
<td><strong>$p^2 + pq = p(p+q) = p$</strong></td>
</tr>
</tbody>
</table>

Because HWE holds in current pop
Hardy-Weinberg equilibrium

To see 1): HWE is established after 1 generation of random mating:

Notations:
Let $p = \Pr(A_1) = \text{allele freq of } A_1 \text{ in parental generation}$
$q = \Pr(A_2) = \text{allele freq of } A_2 \text{ in parental generation}$

$\Pr(A_iA_j) = \text{frequency of genotype } A_iA_j \text{ in initial (parental) generation}$
(we don’t assume these equal $p^2$, $2pq$, $q^2$ for $A_1A_1$, $A_1A_2$, $A_2A_2$)

$\Pr_1$ is notation for probability (frequency) in the next (offspring) generation
To see 1):
Given genotype frequencies in a (parental) population, Mendelian principles of segregation dictate the probability distribution for the genotypes in the offspring population.

<table>
<thead>
<tr>
<th>Father</th>
<th>Mother</th>
<th>Prob that Offspr is $A_1A_1$</th>
<th>Prob that Offspr is $A_1A_2$</th>
<th>Prob that Offspr is $A_2A_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1$</td>
<td>$A_1A_1$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_1$</td>
<td>$A_1A_2$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_1$</td>
<td>$A_2A_2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$A_1A_1$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$A_1A_2$</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$A_2A_2$</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$A_1A_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$A_1A_2$</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$A_2A_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
To see 1):

<table>
<thead>
<tr>
<th>Father</th>
<th>Mother</th>
<th>Prob that Offspr is $A_1A_1$</th>
<th>Prob that Offspr is $A_1A_2$</th>
<th>Prob that Offspr is $A_2A_2$</th>
<th>Freq of this mating type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1$</td>
<td>$A_1A_1$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$\Pr(A_1A_1)\Pr(A_1A_1)$</td>
</tr>
<tr>
<td>$A_1A_1$</td>
<td>$A_1A_2$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>$\Pr(A_1A_1)\Pr(A_1A_2)$</td>
</tr>
<tr>
<td>$A_1A_1$</td>
<td>$A_2A_2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$\Pr(A_1A_1)\Pr(A_2A_2)$</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$A_1A_1$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>$\Pr(A_1A_2)\Pr(A_1A_1)$</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$A_1A_2$</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
<td>$\Pr(A_1A_2)\Pr(A_1A_2)$</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>$A_2A_2$</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>$\Pr(A_1A_2)\Pr(A_2A_2)$</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$A_1A_1$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$\Pr(A_2A_2)\Pr(A_1A_1)$</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$A_1A_2$</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>$\Pr(A_2A_2)\Pr(A_1A_2)$</td>
</tr>
<tr>
<td>$A_2A_2$</td>
<td>$A_2A_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$\Pr(A_2A_2)\Pr(A_2A_2)$</td>
</tr>
</tbody>
</table>

In offspring, genotype probabilities are:

\[
\Pr_1(A_1A_1) = 1 \cdot \Pr(A_1A_1)\Pr(A_1A_1) + 0.5\Pr(A_1A_1)\Pr(A_1A_2) + 0.5 \Pr(A_1A_2)\Pr(A_1A_1) + 0.25 \Pr(A_1A_2)\Pr(A_1A_2)
\]

\[
= [\Pr(A_1A_1) + 0.5\Pr(A_1A_2)]^2
\]

\[
\Pr_1(A_2A_2) = \ldots = [\Pr(A_2A_2) + 0.5\Pr(A_1A_2)]^2
\]

\[
\Pr_1(A_1A_2) = \ldots = 2[\Pr(A_1A_1) + 0.5\Pr(A_1A_2)][\Pr(A_2A_2) + 0.5\Pr(A_1A_2)]
\]
In offspring,
\[ \Pr_1(A_1A_1) = [\Pr(A_1A_1) + 0.5\Pr(A_1A_2)]^2 \]
\[ \Pr_1(A_2A_2) = ... = [\Pr(A_2A_2) + 0.5\Pr(A_1A_2)]^2 \]
\[ \Pr_1(A_1A_2) = ... = 2[\Pr(A_1A_1) + 0.5\Pr(A_1A_2)][\Pr(A_2A_2) + 0.5\Pr(A_1A_2)] \]
Note also that in the parents' generation,
\[ p = \Pr(A_1) = \Pr(A_1A_1) + 0.5\Pr(A_1A_2) \]
\[ q = \Pr(A_2) = \Pr(A_2A_2) + 0.5\Pr(A_1A_2) \]
Thus,
\[ \Pr_1(A_1A_1) = [\Pr(A_1)]^2 = p^2 \]
\[ \Pr_1(A_2A_2) = [\Pr(A_2)]^2 = q^2 \]
\[ \Pr_1(A_1A_2) = 2\Pr(A_1)\Pr(A_2) = 2pq \]
And, again, in the offspring generation
\[ p_1 = \Pr_1(A_1) = \Pr_1(A_1A_1) + 0.5\Pr_1(A_1A_2) = p^2 + 0.5(2pq) = p(p+q) = p \]
\[ q_1 = \Pr_1(A_2) = \Pr_1(A_2A_2) + 0.5\Pr_1(A_1A_2) = q \]
So the *allele frequencies* are unchanged in the next generation; thus
\[ \Pr_1(A_1A_1) = p_1^2, \Pr_1(A_2A_2) = q_1^2, \Pr_1(A_1A_2) = 2p_1q_1. \]
Testing for HWE

Ex: consider locus with alleles a and A. Suppose in a sample of 550 individuals, we observe 200 aa, 300 Aa, 50 AA. Is this consistent with HWE?

Answer: use chi-squared test with 1 degree of freedom:

\[ \chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \]

N = total number of alleles observed = 2*550 = 1100
Pr(a) = (200*2 + 300)/1,100 = 7/11
Pr(A) = (300 + 50*2)/1,100 = 4/11

**Expected genotype counts under HWE:**
N(aa) = \((49/121)*550 = 222.727\)
N(Aa) = \(2*(28/121)*550 = 254.545\)
N(AA) = \((16/121)*550 = 72.727\)

**Observed genotype:**
\(200\)
\(300\)
\(50\)

\[ \chi^2 = \frac{[(49 \times 550/121) - 200]^2}{222.727} + \frac{[(56 \times 550/121) - 300]^2}{254.545} + \frac{[(16 \times 550/121) - 50]^2}{72.727} \]

\[ = 2.319 + 8.11689 + 7.102 = 17.54 \rightarrow p-value \approx 0.00002816 = 2.816 \times 10^{-5} \]
Testing for HWE

chi-squared test with 1 degree of freedom:

3.841 is the critical value for $\alpha = 0.05$.

6.635 is the critical value for $\alpha = 0.01$.

Checking for HWE can detect genotyping errors (e.g., deviation from HWE b/c too few heterozygotes)
Another use of HWE: Estimating allele frequencies for a disease gene

Brief intro to genetic disease models

Suppose a disease is caused by a single major gene with two alleles: wild type (+) and mutant (d). Let T denote the presence of the trait or disease.

One way to specify a disease model: use parameters for penetrances, gene frequencies, population prevalence.

Penetrances for each of the possible genotypes at the gene are denoted by:

\[ f_{++} = \Pr(T \mid ++) \]
\[ f_{+d} = \Pr(T \mid +d) \]
\[ f_{dd} = \Pr(T \mid dd) \]
Classical models of disease

**Classical autosomal dominant inheritance** (no phenocopies, fully penetrant).

Penetrance table:

<table>
<thead>
<tr>
<th></th>
<th>++</th>
<th>+d</th>
<th>dd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Classical autosomal recessive inheritance** (no phenocopies, fully penetrant).

Penetrance table:

<table>
<thead>
<tr>
<th></th>
<th>++</th>
<th>+d</th>
<th>dd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Can also have penetrances between 0 and 1
Using HWE to estimate disease allele frequencies

If random mating is a reasonable assumption, HWE provides a way to estimate allele frequencies for (pure) Mendelian dominant or recessive diseases, given the prevalence of a disease/trait.

Pure Dominant: Prevalence = \( f_{++}p^2 + 2f_{+d}pq + f_{dd}q^2 = 2pq + q^2 \)

Pure Recessive: Prevalence = \( f_{++}p^2 + 2f_{+d}pq + f_{dd}q^2 = q^2 \)

Ex: Consider a recessive disease in which a random sample of 40,000 individuals yields 4 cases. What is a good estimate of the frequency of the disease allele ("disease gene")?
Assuming HWE, prevalence of the disease = \( Pr(dd) = q^2 \).
Thus \( q = \sqrt{\text{prevalence}} = \sqrt{4/40,000} = 0.01 \).

For dominant models, can solve quadratic equation

\[ ax^2 + bx + c = 0 \]

\[ \text{has solutions} \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]
Using HWE to estimate disease allele frequencies

If dominant model for a disease with prevalence K, with A the dominant allele

Genotype:    aa  Aa  AA
Penetrance:  0    1    1

Note this is equivalent to “a” being the allele for a recessive trait of "not having the disease":

Genotype:    aa  Aa  AA
Penetrance:  1    0    0
(for NOT having disease)
Prevalence of NOT having disease is 1-K.

So have a recessive trait with allele frequency \( p_a = 1 - p_A \), prevalence 1-K:
use the formula: \( 1 - K = f_{AA} p_A^2 + 2f_{Aa} p_A p_a + f_{aa} p_a^2 = p_a^2 \)
Then \( p_A = 1 - p_a \)
Wright-Fisher Model

In real life, populations are not infinite, and allele and genotype frequencies do change over time.

How do we extend from HWE to account for finite population, key processes?
Wright-Fisher Model

Assumptions:
• Two allele system
• N diploid individuals in each generation
• 2N gametes
• Random mating, no selection, no mutation
• Discrete generations

Each generation, the new population is made by sampling with replacement from the previous generation
Let’s play a few rounds of this game

G1
R  G  G  G  R  R  R  G

G2
R  G  R  G  G  G  G  R

G3
G  G  G  G  G  G  G  G

Fixation!
The game is faster by computer!

Fixation!

I = 400
A = 200
R = 100
G = 100

I = Number of Generations
A = Population size (gametes)
G = Count of the G allele
R = Count of the R allele
Again:

Fixation!

I = 400
A = 200
R = 100
G = 100

Fixation!

I = 400
A = 200
R = 100
G = 100
Let’s investigate this phenomenon:

Change population size

Change allele frequencies
Smaller population size....

I = 40
A = 20
R = 10
G = 10
Larger population size….
Larger population size....

$I = 1000$
$A = 2000$
$R = 1000$
$G = 1000$
Initial red allele frequency > green allele frequency

I = 400
A = 200
R = 150
G = 50
Initial red allele frequency > green allele frequency
Initial red allele frequency > green allele frequency

\[ I = 400 \]
\[ A = 200 \]
\[ R = 150 \]
\[ G = 50 \]
Each generation, the new population is made by **sampling with replacement** from the previous generation.

Let: \( P_t = \text{freq (A) among gametes} \)  
\( P_{t+1} = \ldots \)  
\( n_{t+1} = \text{count of (A)} \)  

Then: \( n_{t+1} \sim \text{Binomial}(P_t, 2N) \)

\[
\Pr(n_{t+1} = m) = \binom{2N}{m} p_t^m (1-p_t)^{2N-m}
\]

\[
E(p_{t+1}) = P_t \\
\text{Var}(p_{t+1}) = \frac{p_t(1-p_t)}{2N}
\]

Implications: sampling variance ("genetic drift") is dependent on population size. Allele frequency is a random sequence of numbers: \( p_1, p_2, p_3, \ldots \). Eventually \( p = 1 \) or \( p = 0 \). Stay “fixed” until new mutation.
Sampling with replacement

- Some alleles pass on no copies to the next generation, while some pass on more than one.
Decay of heterozygosity

Define $G_t = \text{homozygosity at generation } t$

= probability that a random draw of 2 chromosomes from the pop results in 2 of the same allele

Define $H_t = 1 - G_t = \text{heterozygosity at generation } t$

= probability that a random draw of 2 chromosomes from the pop results in 2 different alleles

Under Wright-Fisher assumptions, what happens to $H_t$ (or $G_t$) over time?
Decay of heterozygosity

Two ways to get 2 of the same allele:

Generation 0

Probability: \( \frac{1}{2N} \)

Generation 1

Same source in previous generation

Generation 0

Probability: \( \left(1 - \frac{1}{2N}\right) \times G_0 \)

Generation 1

\[ G_1 = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \times G_0 \]
Decay of heterozygosity

Two ways to get 2 of the same allele:

1. Probability in Generation t:
   \[ P(t) = \frac{1}{2N} \]

2. Probability in Generation t+1:
   \[ P(t+1) = \left( 1 - \frac{1}{2N} \right) \times G_t \]

\[ G_{t+1} = \frac{1}{2N} + \left( 1 - \frac{1}{2N} \right) \times G_t \]
Decay of heterozygosity: Proof

\[ G_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \times G_t \]

\[ H_{t+1} = 1 - G_{t+1} = 1 - \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \times G_t\right) \]

\[ = 1 - \frac{1}{2N} - G_t + \frac{1}{2N} \times G_t \]

\[ = (1 - G_t) - \frac{1}{2N} (1 - G_t) \]

\[ = (1 - \frac{1}{2N}) H_t \]

\[ = \left(1 - \frac{1}{2N}\right)^2 H_{t-1} = \ldots = \left(1 - \frac{1}{2N}\right)^{t+1} H_0 \]

Therefore \( H_t = \left(1 - \frac{1}{2N}\right)^t \times H_0 \)
Insights from $H_t = \left(1 - \frac{1}{2N}\right)^t \times H_0$

Half-life of $H$: at what $t$ is $H_t = \frac{1}{2} \times H_0$?

$$\frac{1}{2} = \left(1 - \frac{1}{2N}\right)^t$$

$$\ln\left(\frac{1}{2}\right) = t \times \ln\left(1 - \frac{1}{2N}\right)$$

$$t = \frac{-\ln(2)}{\ln\left(1 - \frac{1}{2N}\right)}$$

Use approximation $\ln(1 + x) \approx x$

$$t \approx 2N\ln(2) = 1.39 \times N$$

Larger $N$ corresponds to longer half-life

For $N = 10^4$, $t = (2,0000)\times\ln(2) = 1.39\times10^4$
Insights from $H_t = \left(1 - \frac{1}{2N}\right)^t \ast H_0$

Half-life of $H$: $t = 1.39 \text{ N}$

This indicates that even in a large population, eventually every allele will have descended from a single allele in the founding population! At a locus, all but 1 allele will have "died off" (been lost)

(Remember: assuming no selection, no mutation)

Half-life is larger for larger populations
Darwin versus Drift

Let's add selection to our model.

Need to account for differing fitness conferred by differing genotypes.
Define relative fitness for each possible individual
e.g.
Fitness RR = 1
Fitness RG = 1.1
Fitness GG = 2

Now the rules account for fitness:
Pick an individual with probability proportional to the fitness of their genotype.

Here, GG is twice as likely to be chosen as RR.
Now choose 1 chromosome to put into the next generation.
What relative fitness should we select? (Calibrate the scale)

Conserved elements < 0.01% increase in fitness
Darwin versus Drift

$I = 100$
$A = 100$
$R = 99$
$G = 1$
$fG = 2*R$

$I =$ # generations
$A =$ # gametes
$R =$ count of R allele
$G =$ count of G allele
Darwin versus Drift

\[ I = 100 \]
\[ A = 100 \]
\[ R = 99 \]
\[ G = 1 \]
\[ fG = 3 \times fR \]
Darwin versus Drift

\[ I = 100 \]
\[ A = 100 \]
\[ R = 99 \]
\[ G = 1 \]
\[ fG = 3*fR \]
Darwin versus drift

I = 100
A = 2000
R = 1999
G = 1
fG = 3*fR
Some surprising results!

Survival of the fittest? Survival of the luckiest?

Sometimes drift can overcome selection (Depends on allele frequency, population size)

Most new advantageous mutations are NOT fixed!
Chance can play a large role in determining which polymorphisms are fixed in a population.

(Not necessarily obvious)

Amount of variation at a locus, and fate of individual alleles, depends on mutation-selection-drift balance.
Examined 131,060 Icelanders born after 1972
Compared with expectations from Wright Fisher model

- Considerable effect of genetic drift, even with rapid population expansion rather than constant population size
Sampling with Replacement

- Some alleles pass on no copies to the next generation, while some pass on more than one.
Sampling with Replacement

- Some alleles pass on no copies to the next generation, while some pass on more than one.
The coalescent process

- "backward in time" process
- Lineage of alleles in a sample traced backward in time to their common ancestor allele
- Genealogies are unobserved, but can be estimated
- Conceptual framework for population genetic inference: mutation, recombination, demographic history
- Kingman, Tajima, Hudson
Decay of heterozygosity

Two ways to get 2 of the same allele:

- Probability in previous generation:
  \[ \frac{1}{2N} \]

- Probability in generation t:
  \[ \left( 1 - \frac{1}{2N} \right) \times G_t \]

\[ G_{t+1} = \frac{1}{2N} + \left( 1 - \frac{1}{2N} \right) \times G_t \]

Probability that 2 allele copies at generation t+1 coalesce in previous generation t!