Population Genetics II
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)
Outline for this lecture

• Decay of heterozygosity, revisited briefly
• Mutation
• Coalescent
• Linkage disequilibrium
Recall:
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

Last time, we derived a recursion formula.
Decay of heterozygosity

Two ways to get 2 of the same allele:

- **Generation t**
  - Probability: $\frac{1}{2N}$

- **Generation t+1**
  - Same source in previous generation

- **Generation t**
  - Probability: $\left(1 - \frac{1}{2N}\right) \times G_t$

- **Generation t+1**

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

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

Genetic drift & decay of heterozygosity -> variation is removed from the population.

Mutation restores genetic variation.

Neutral theory: most of the DNA sequence differences within a population are due to neutral mutations.
Let $\mu =$ mutation rate to neutral alleles (per bp per generation) (sometimes $u$ stands in for $\mu$)

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

If now allow mutation:
After 1 round with mutation possible:

$$G_{t+1} = (1 - \mu)^2 \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \times G_t \right)$$

At equilibrium, $G_t = G_{t+1}$
Claim: at equilibrium, probability that 2 alleles drawn at random are identical is (essentially):

\[
\frac{1}{1 + 4N\mu}
\]

Proof:

\[
G_{eq} = (1 - \mu)^2 \left( \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) \times G_{eq} \right)
\]

\[
G_{eq} \left(1 - \left(1 - \frac{1}{2N}\right) \times (1 - \mu)^2\right) = \frac{1}{2N}(1 - \mu)^2
\]

\[
G_{eq} = \frac{\frac{1}{2N}(1 - \mu)^2}{\left(1 - \left(1 - \frac{1}{2N}\right) \times (1 - \mu)^2\right)} \quad \text{Use } (1-\mu)^2 \sim 1-2\mu
\]

\[
\approx \frac{1 - 2\mu}{2N \left(1 - \left(1 - \frac{1}{2N}\right) \times (1 - 2\mu)\right)} = \frac{1 - 2\mu}{(1 + 4N\mu - 2\mu)}
\]

\[
\approx \frac{1}{(1 + 4N\mu)} \quad \text{Use } 2\mu \text{ very small}
\]
Mutation

4Nμ comes up repeatedly in population genetics; often referred to as theta:

$$\theta = 4N\mu$$

(Different from recombination rate $\theta$!! Pop geneticists often use "r" for recombination rate.)

At equilibrium (with drift and mutation rate $\mu$), the probability that 2 alleles drawn at random are the same is

$$\frac{1}{(1+4N\mu)} = \frac{1}{1+\theta}$$

Expected heterozygosity at equilibrium:

$$H_{eq} = 1 - G_{eq} = 1 - \frac{1}{1+\theta} = \frac{\theta}{1+\theta}$$
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
2 sample (item) coalescent

- N = population size of diploid individuals
- n = sample size of haploid chromosomes
- MRCA = most recent common ancestor
- $T_2 = \text{coalescence time for 2 chromosomes}$
2 sample (item) coalescent

Prob that the time of MRCA is \( t \) generations ago:

\[
P(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right)
\]

Did not coalesce for first \( t-1 \) generations

Coalesced at \( t \)

Can approximate (as \( N \to \infty \))

\[
P(T_2 = t) = \left(\frac{1}{2N}\right) e^{-\left(\frac{1}{2N}\right)t}
\]

Geometric distribution \( p(x) = (1-p)^{x-1} p \)

Has expected value or mean \( 1/p \), so here \( E(T_2) = 2N \)
In “coalescent units” let’s $t' = t/2N$, then $E(T_2) = 1$
n-coalescent

Have \( \binom{n}{2} = \frac{n(n-1)}{2} \) possible pairs that could coalesce. Analogous to 2-item approximation:

\[
P(T_n = t) = \left(\frac{n}{2N}\right) e^{-\left(\frac{n}{2N}\right)t}
\]

\[
E(T_n) = \frac{2N}{\binom{n}{2}} = \frac{2N}{\frac{n(n-1)}{2}}
\]

In coalescence units,

\[
E(T_n) = \frac{1}{\binom{n}{2}} = \frac{1}{\frac{n!}{(n-2)!2!}} = \frac{2}{n(n-1)}
\]
n-coalescent

In coalescence units,

\[ E(T_n) = \frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)} \]

Mean elapsed time in coalescence units (*2N)

\[ E(T_{MRCA\ for\ n\ chromosomes}) = T_2 + T_3 + T_4 + \ldots + T_n = 2(1-1/n) \text{ coalescent units} \]
Linkage disequilibrium

• “Non-random associations between alleles at different loci”
• Contrast with HWE: HWE relates to alleles A and a at the same locus
• LD statistics quantify $Pr(AB$ haplotype) compares to $Pr(A)*Pr(B)$ at different loci
• Important in the design and interpretation of disease mapping studies
Mapping disease genes

• **Linkage**
  - Quantify co-segregation of trait and genotype in families

  ![Linkage Diagram](image)

  LOD score traditionally used to measure statistical evidence for linkage

• **Association**
  - Common design: case-control sample, analyzed for allele frequency differences

  ![Association Diagram](image)
Association in a case-control sample

Let $N = N_{\text{case}} + N_{\text{control}}$ (2N observations of alleles)

Most basic test for biallelic markers: compare allele frequencies in cases vs controls in a 2x2 table.

\[
\begin{array}{c|cc}
\text{case} & \text{ctrl} \\
\hline
A_1 & N_{11} & N_{12} \\
A_2 & N_{21} & N_{22} \\
\end{array}
\]

Chi-square with $n-1$ df ($n = \# \text{ of alleles}$)

\[
\chi^2 = \sum \frac{(obs - exp)^2}{exp}
\]
Association in a case-control sample

**Alternatives:** logistic regression

Let $P =$ probability of being a case.

$$\log\left(\frac{P}{1-P}\right) = a_0 + (a_1x_1 + \ldots + a_mx_m) + b_1G$$

$x_i$ are covariates (e.g., gender, age)

$G$ represents genotype (0, 1, or 2 copies of a specified allele)

(corresponds to a log-additive, that is, multiplicative model).

Statistical test: determine the improvement in fit when the genotype term is added. (Likelihood ratio chi-square).
Linkage disequilibrium

Note: the above tests should work great if the marker you genotyped is actually the disease locus.

What if the marker is "nearby" or "correlated" with the disease locus?

Here the concept of “linkage disequilibrium” (LD) comes in.

The International HapMap Project / 1000 Genomes
• **Goal:** determine the common patterns of DNA sequence variation (LD among SNPs) in human populations
• Identifies redundancy among SNPs for more efficient disease mapping and pharmacogenetics studies
Human DNA sequence variation

How to measure/describe "patterns" of DNA sequence variation?

How to use these patterns to find disease genes that affect phenotypes?
Alleles on the preserved "ancestral background" tend to be in LD
LD involves haplotype frequencies.

Focus on pair-wise LD, SNP markers

Genotypes do not necessarily determine haplotypes:

Consider 2-locus genotype $A_1 A_2 B_1 B_2$.

Two possible phases:

\[
\begin{array}{c|c|c|c|}
A_1 & A_2 & \ & \\\n\hline
B_1 & B_2 & \ & \\
\end{array}
\quad \begin{array}{c|c|c|c|}
A_1 & A_2 & \ & \\
\hline
B_2 & B_1 & \ & \\
\end{array}
\]
Linkage Disequilibrium (LD), aka allelic association

For two loci A and B:
LD is said to exist when alleles at A and B tend to co-occur on haplotypes in proportions different than would be expected under statistical independence.

How to formally measure LD between alleles at 2 loci?
Linkage disequilibrium

Example: Consider 2 SNPs:
SNP 1: A 50% C 50%
SNP 2: A 50% G 50%

4 possible haplotypes:

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
<th>Expected freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>0.5 * 0.5</td>
</tr>
</tbody>
</table>

But perhaps in your sample you observe only the following:

A A C C A T A T C ... C G A T T ...

and

A A C C C T A T C ... C A A T T ...

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>
To measure LD between alleles at 2 biallelic loci

<table>
<thead>
<tr>
<th>Locus A</th>
<th>Locus B</th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁, A₂</td>
<td>B₁, B₂</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Given 2N haplotypes:

Haplotype freq for AᵢBⱼ is

\[ h_{ij} = \frac{n_{ij}}{2N} \]

Compare \( h_{ij} \) to the frequency expected under no association:

\[ p_{A₁}p_{B₁} = \left( \frac{n_{11} + n_{12}}{2N} \right) \left( \frac{n_{11} + n_{21}}{2N} \right) \]

Define the disequilibrium coefficient:

\[ D = h_{11} - p_{A₁}p_{B₁} \]
Notes:

1. \( D = h_{11} - p_{A_1}p_{B_1} = h_{22} - p_{A_2}p_{B_2} \)

2. Choice of allele labeling may affect sign but not absolute value of \( D \).
Common LD measures

Disequilibrium coefficient:
\[ D = h_{11} - p_{A1} p_{B1} \]

Normalized disequilibrium coefficient:
\[ D' = \frac{D}{|D|_{\text{max}}}, \text{ where} \]

\[ |D|_{\text{max}} = \begin{cases} 
\min(p_{A1}p_{B2}, p_{A2}p_{B1}) & \text{if } D > 0 \\
\min(p_{A1}p_{B1}, p_{A2}p_{B2}) & \text{if } D < 0 
\end{cases} \]

Range of D' is [-1,1]

Correlation coefficient:
\[ r^2 = \frac{D^2}{(p_{A1}p_{A2}p_{B1}p_{B2})} \]
LD measures

$|D'|$ is 1 when the alleles of the two markers are as correlated as they can be, given the allele frequencies of the co-occurring alleles.

The range of $r^2$ depends on the marker allele frequencies.

$r^2$ equals 1 if and only if (1) the MAFs at the two loci match and (2) the minor alleles always co-occur.

$D'$: useful for identifying regions of reduced recombination.

$r^2$: useful for identifying markers that are good predictors of allelic status at other markers.
Describing empirical LD patterns

Figure 2. Linkage disequilibrium across the single nucleotide polymorphisms (SNPs) genotyped in and around ACN9. D' is illustrated by shading, with darker shades indicating higher D'. The $r^2$ is indicated by the number inside the shaded block.

Haploview output
Dick et al., 2007
Reference populations - and their LD/haplotype patterns - are used to design “tag SNPs”, impute un-typed variants

• 1000 Genomes Project (1000G)
• Haplotype Reference Consortium (HRC)
• Previously: HapMap: Phase I 2003, Phase II 2007, Phase III
Using LD in study design

The International HapMap Project, Nature 2003s
A popular LD tag method:
• "$r^2$ bin tags" (Carlson et al., 2004): greedy algorithm that identifies bins of SNPs such that at least one member of each bin has $r^2 > T$ (threshold) with all bin members.
• Note: bin members are not necessarily contiguous

LD patterns inform the design of SNP genotyping arrays, selection of "tag" SNPs
Thus knowledge of LD patterns is important for disease gene mapping.

**Note:** tight linkage between two loci will tend to maintain LD.
Decay of LD

After $k$ generations, disequilibrium decays according to

$$D_k = (1 - \theta)^k D_0$$

where $\theta = $ the recombination fraction (assuming random mating).

**(DIFFERENT $\theta$!)

$$h_{11(1)} = (1 - \theta) h_{11(0)} + \theta p_{A1} p_{B1},$$

so at generation 1,

$$D = h_{11} - p_{A1} p_{B1} = (1 - \theta) (h_{11(0)} - p_{A1} p_{B1})$$
claim: $h_{ij(1)} = (1 - \theta) h_{ij(0)} + \theta p_{Ai} p_{Bj}$,

so at generation 1,
$D = h_{11} - p_{A1} p_{B1} = (1 - \theta) (h_{11(0)} - p_1 q_1) = (1 - \theta) D_0$

after k generations, get:  
$$D_k = (1 - \theta)^k D_0$$
Decay of LD

Therefore, after k generations

$$h_{ijk} - p_ip_j = (1 - \theta)^k (h_{ijo} - p_ip_j)$$

That is,

$$D_k = (1 - \theta)^k D_0$$

Disequilibrium decays by a factor of $(1 - \theta)$

Note: After 1 generation, genotype frequencies at a single locus are in equilibrium, haplotype frequencies are not!
Half-life of LD

How many generations (t) to reduce by $\frac{1}{2}$?

$$(1 - \theta)^t (h_{ij0} - p_iq_j) = \frac{1}{2} (h_{ij0} - p_iq_j)$$

$$(1 - \theta)^t = \frac{1}{2}$$

$$t \log (1 - \theta) = \log(0.5)$$

$$t = \log(0.5) / \log(1 - \theta)$$

e.g. if $\theta = 0.5$ (loci are unlinked) then

$$t = \log(0.5) / \log(0.5) = 1$$
Figure 4.1  Decay of linkage disequilibrium by generation.

c.f. Pak Sham, Statistics in Human Genetics, Chapter 4
LD is not a simple monotonic function of physical distance

From Taillon-Miller et al., Nat Genet 2000 (O=Xq25, □=Xq28)
LD is not necessarily a monotonic function of distance
Dawson et al., *Nature* 2002 (chromosome 22)
Where does LD come from?

Potential sources of LD:

1. Linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
Where does LD come from?

Potential sources of LD:

1. Linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
Genetic drift generates LD (|D| > 0)

- Via random changes in gamete frequencies
- Smaller isolates: slower decay of LD

**FIGURE 4.3**

The decay of linkage disequilibrium as measured by $r^2$ (Eq. 1.20) with recombination distance in several human populations. The populations labeled ALB, CAN, CAB, CAR, MON, and ROA represent small isolated villages from an Apennine valley in Northwestern Italy. Valley is the conglomerate of all of these isolates, VER represents the Italian population from the Veneto region, TSI the Italian population from Tuscany, CEU the European population, and YRI a large African population.

Potential sources of LD:

1. Linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
Suppose have loci A, B, C in that order.
Due to founder effect, suppose sample only 4 haplotypes out of the 8 possible:

A B C
1 1 1
1 2 1
2 1 2
2 2 2

Note:
A and B are in equilibrium
B and C are in equilibrium
A and C are in complete disequilibrium

Disequilibrium not necessarily related to distance!
Where does LD come from?

Potential sources of LD:

1. Linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification

At the appearance of the mutation, that allele occurs only on one haplotype background.
Where does LD come from?

Potential sources of LD:

1. Linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
An example of spurious association due to admixture/stratification:

<table>
<thead>
<tr>
<th></th>
<th>population 1</th>
<th></th>
<th>population 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>9</td>
<td>90</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
<td>100</td>
<td>50</td>
</tr>
</tbody>
</table>

chi-square = 0

combined

<table>
<thead>
<tr>
<th></th>
<th>34</th>
<th>26</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>106</td>
<td>34</td>
<td>140</td>
</tr>
</tbody>
</table>

chi-square = 7.26
p-value = 0.007
A first generation linkage disequilibrium map of chromosome 22
Dawson et al., Nature 2002

1504 SNPs analyzed in 2 distinct samples

Nature 2003: HapMap I (genome-wide)

Nature 2007

A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium*

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25–35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum $r^2$ of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum $r^2$ of up to 0.8 in African and up to 0.95 in