BIOL 502 Population Genetics Spring 2017

Lecture 2 The Hardy-Weinberg Principle & Linkage Disequilibrium

Arun Sethuraman

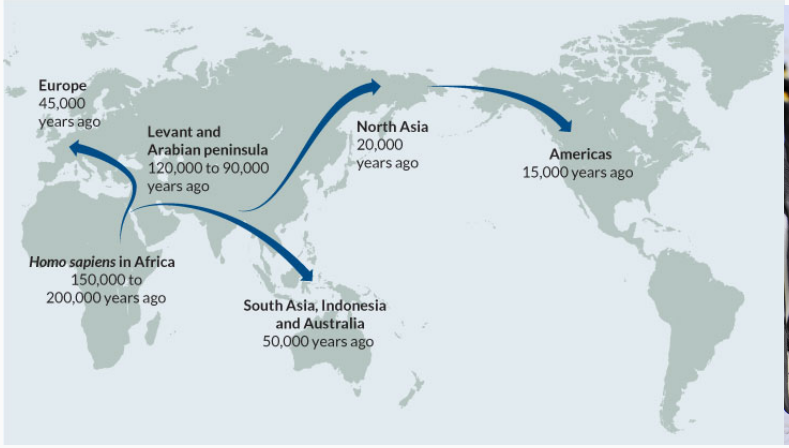
California State University San Marcos

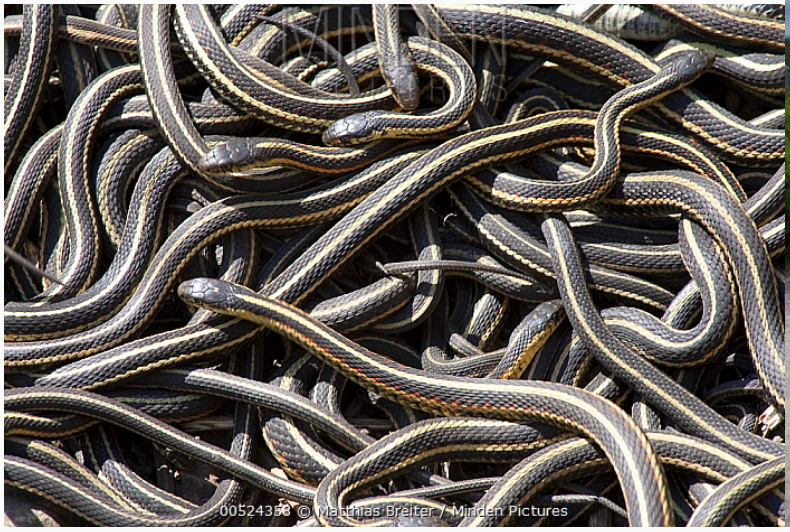
Table of contents

Organization of Genomic Variation

Subpopulations/Demes

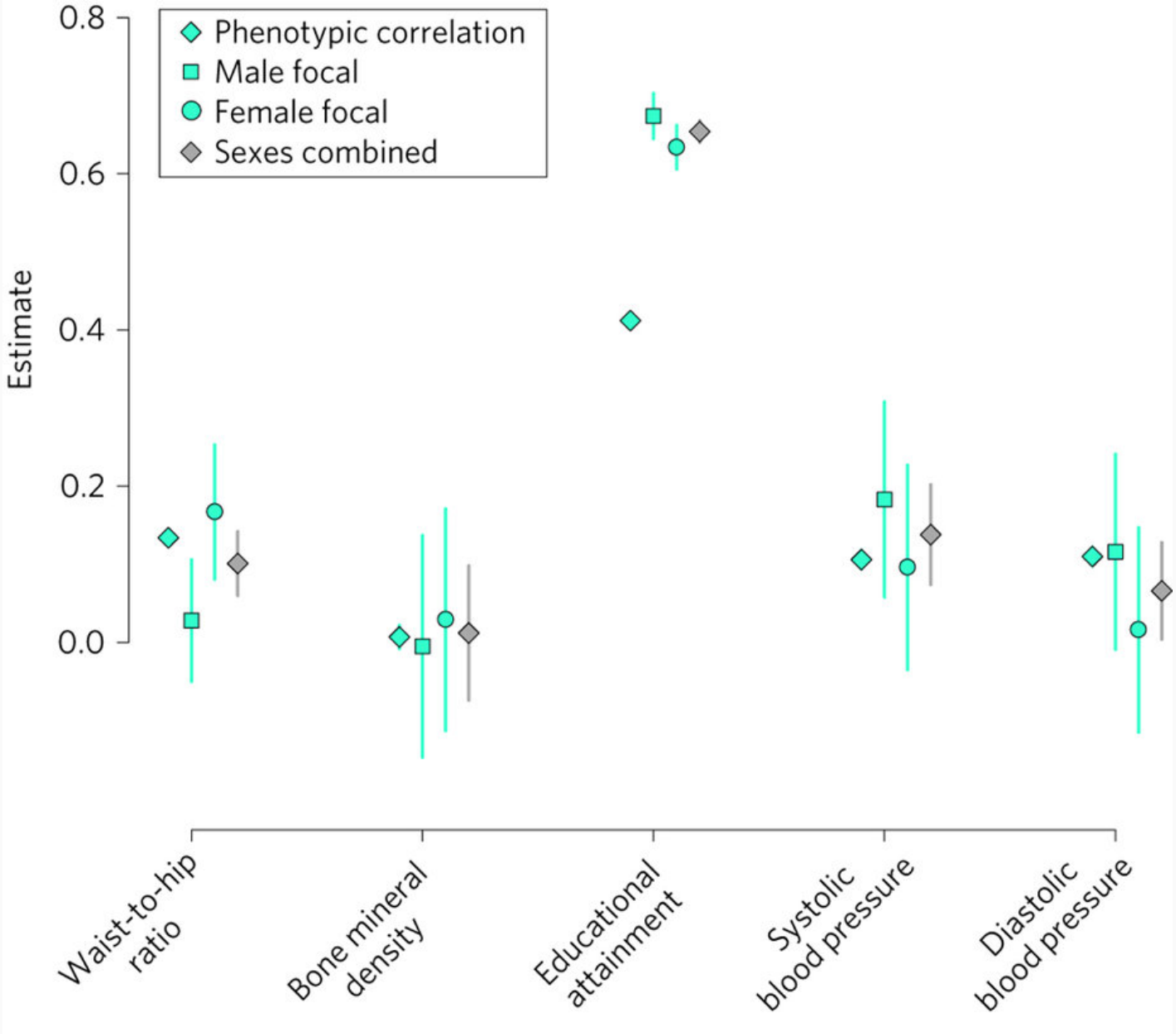
Individuals in a population are very rarely homogeneously distributed in space and time.





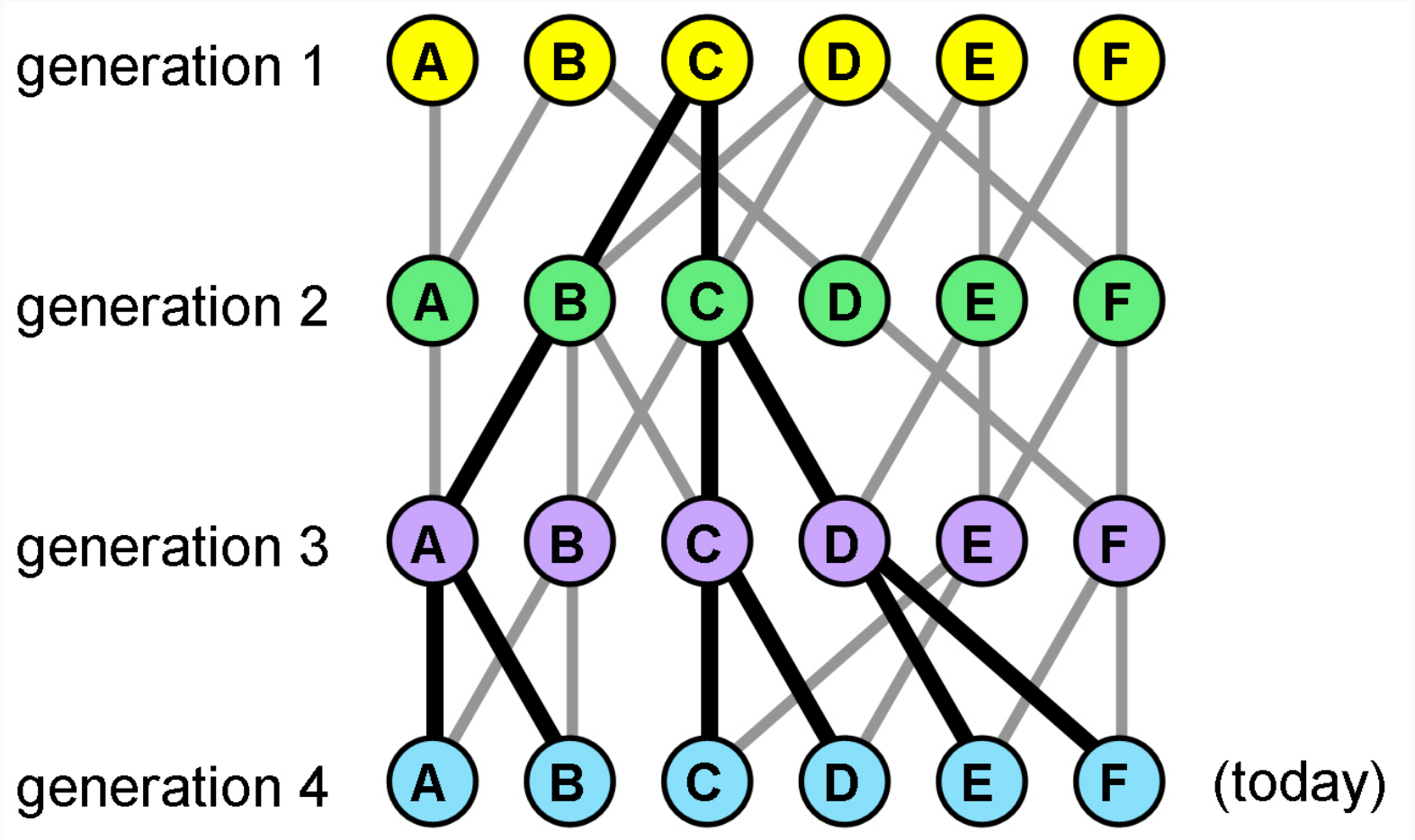
Random Mating

Genetic evidence of assortative mating in humans, Robinson et al. 2017, Nature Human Behavior



[2]

Non-overlapping Generations



Hardy-Weinberg Principle

Assumptions

- Diploidy
- Sexual reproduction
- Non-overlapping generations
- Bi-allelic locus
- Identical allele frequencies among males and females
- Random mating
- Large population size (in theory, infinite)
- No migration, or negligible
- No mutation
- No selection or differential fitness effects

Bi-allelic Case

- Assume two alleles A and a at a genomic locus - i.e. biallelic locus
- `Event`: drawing an offspring's genotype from the population with A or a gametes (sperm and egg)
- `Random Variable`: $X=$ genotype of offspring
- `Sample Space` for $X = \{AA, Aa, aa\}$
- $F =$ paternal allele
- $M =$ maternal allele
- `Sample Space` for $F, M = \{A, a\}$

What is HWE?

- A state of equilibrium after one generation of random mating in an ideal population
- `Expected` offspring genotype frequencies ($f(AA), f(Aa), f(a)$), or heterozygosities ($2pq$), or homozygosities ($p^2 + q^2$).
- Allele frequencies of offspring generation are equal to the allele frequencies in parental generation
- If `observed` genotype or allele frequencies are different from `expected`, population is said to be *evolving*
- Recall - *evolution* is descent with modification.

HWE Problem

`Observed`

$p = 0.25, q = 0.75$

`Expected` After one generation of random mating:

- Genotype frequency of AA
- Genotype frequency of Aa
- Genotype frequency of aa
- Allele frequency of A
- Allele frequency of a

Case of random mating of Genotypes

- Same case as above, but let's assume that the `Event` is drawing the offspring genotype from a population when the parents can have either AA, Aa, aa genotypes.
- So mating events are: *AAxAA*, *AAxAa*, *AAxaa*, *AaxAa*, *Aaxaa* and *aaxaa*.
- Let's derive HWE proportions (i.e. genotype and allele frequencies after one generation of random mating).

Testing for Hardy Weinberg Equilibrium

χ^2 test

Recall

$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$ where *observed* and *expected* refer to the observed and expected *numbers* in any genotypic class and $\sum$ denotes that the values are summed over all genotypic classes.

Degrees of Freedom

df = Number of classes of data - Number of parameters estimated from the data - 1

Problem 2.3

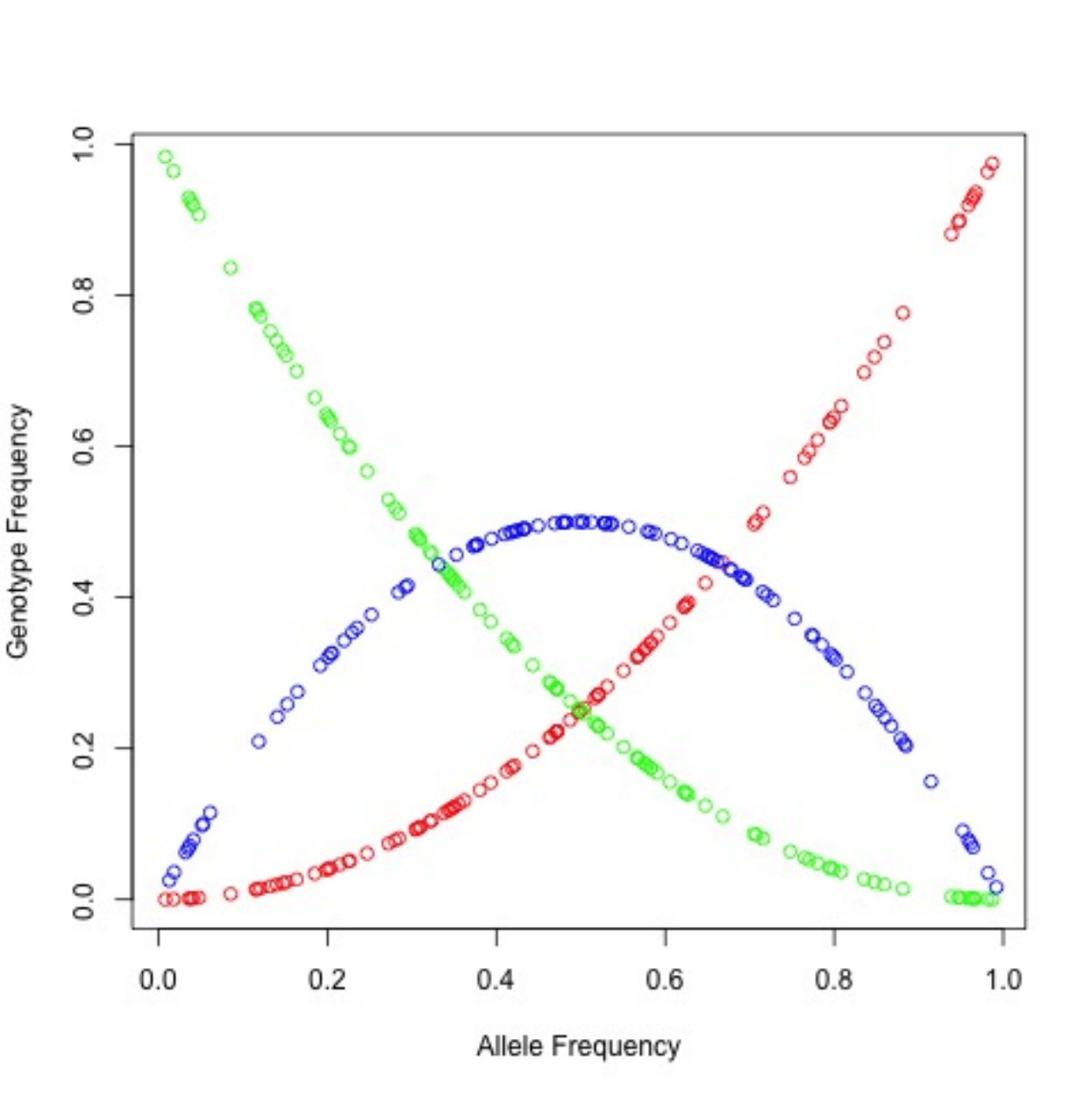
The table shows observed numbers of AA, Aa, aa genotypes in samples of size 100 from each of four populations. Calculate the chi-square value of goodness of fit to Hardy-Weinberg proportions and the associated P vale from each sample. For which samples can the hypothesis of HW-proportions be rejected?

Pop	AA	Aa	aa
(a)	8	53	39
(b)	9	61	30

Testing/Simulating HWE in R

```
install.packages("HardyWeinberg") #Installs package
library(HardyWeinberg) #Load library
#Chisq Test of HWE
x<-c(MM=298,MN=489,NN=213) #Defines genotype vector
HW.test<-HWChisq(x, verbose=TRUE) #Chisq test
#Simulate a population in HWE
m <- 100 # number of markers
n <- 100 # sample size
X1<-HWData(m,n,exactequilibrium=TRUE) #Simulate
pA<-(2*X1[,1]+X1[,2])/(2*n) #A allele freq
pB<-1-pA #B allele freq
pAA<-X1[,1]/100 #AA genotype freq
pAB<-X1[,2]/100 #AB genotype freq
pBB<-X1[,3]/100 #BB genotype freq
plot(pA,pAA,xlab="Allele Frequency"
,ylab="Genotype Frequency",col="red")
points(pB,pAB,col="blue")
points(pA,pBB,col="green")
```

Simulating HWE in R



Try at home

Simulating populations out of HWE in R

You will use `R` to simulate 3 populations that are NOT in HWE. Subsequently, make plots of the allele frequency distributions versus the genotype frequencies.

Read

Read more about the HardyWeinberg package here:`https://cran.r-project.org/web/packages/HardyWeinberg/HardyWeinberg.pdf`

Sample Size Problems

Caution 1

If sample sizes are too small, i.e. if the allele frequencies are too small, correspondingly, expected genotype frequencies and numbers will also be small.

Rule of thumb

As a rule of thumb, if number of observed individuals for a particular class are less than 5, choose an exact test of HWE over the χ^2 test

`http://www.biostathandbook.com/small.html`

Caution 2

Always compute χ^2 values based on numbers of genotypes, and not on frequencies. Also, the number of degrees of freedom is different, because of the number of estimable parameters in a HWE test.

Small Sample Size Problem

Say, we have 4 diploid individuals, with observed genotypes AA, Aa, aa. If they have 4 A alleles, and 4 a alleles among them, what genotype configurations are possible?

Permutation probabilities - Weir 1996

Assume a biallelic locus, with two alleles A and a. Under the HWE hypothesis, the probability of the observed set of genotypic counts n_{AA}, n_{Aa}, n_{aa} in a sample size of n is:

$$Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(p_A^2)^{n_{AA}}(2p_Ap_a)^{n_{Aa}}(p_a^2)^{n_{aa}} \quad (1)$$

whereas the allele counts n_A and n_a are binomially distributed if HWE holds:

$$Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a} \quad (2)$$

Combining these, the conditional probability can be computed as:

$$Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a) = \frac{Pr(n_{AA}, n_{Aa}, n_{aa} \& n_A, n_a)}{Pr(n_A, n_a)} \quad (3)$$

Exact/Permutation Test

Rewriting this:

$$Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a) = \frac{Pr(n_{AA}, n_{Aa}, n_{aa})}{Pr(n_A, n_a)} = \frac{n! n_A! n_a! 2^{n_{Aa}}}{n_{AA}! n_{Aa}! n_{aa}! (2n)!} \quad (4)$$

Problem

Now let's go back to the problem of having 4 individuals, with 4 *A* alleles, and 4 *a* alleles. The different configurations of genotypes that are possible for these observed allelic combinations are:

1. (0, 4, 0), Pr = 0.2286
2. (1, 2, 1), Pr = 0.6857
3. (2, 0, 2), Pr = 0.0857

Rearranging these,

1. (2, 0, 2), Pr = 0.0857, Cumulative Pr = 0.0857
2. (0, 4, 0), Pr = 0.2286, Cumulative Pr = 0.3143
3. (1, 2, 1), Pr = 0.6857, Cumulative Pr = 1.0000

Cumulative probability corresponds to the *P* value of observing a fit as bad (or worse) than the sample configuration. So (2, 0, 2) would have a HWE rejection with a P-value of 0.0857, whereas (1, 2, 1) would imply that HWE would be rejected at the *P* value of 1.0.

Extensions to HWE - Non-random Mating

Now consider a case, where there is non-random mating, especially an extreme case of inbreeding, where only $AA \times AA$, $Aa \times Aa$ and $aa \times aa$ matings occur at a single genetic locus that has two alleles A and a at frequencies p and q respectively. What will be the expected genotype frequencies and allele frequencies after one generation of non-random mating?

Non-Random Mating

Summary

1. F can be defined as the proportional reduction in heterozygosity due to non-random mating.
2. $F = \frac{H_{exp} - H_{obs}}{H_{exp}}$
3. Genotype frequencies change.
4. Allele frequencies remain the same, UNLESS there are fitness differences (we'll come back to this later!)

Extensions to HWE - Homework

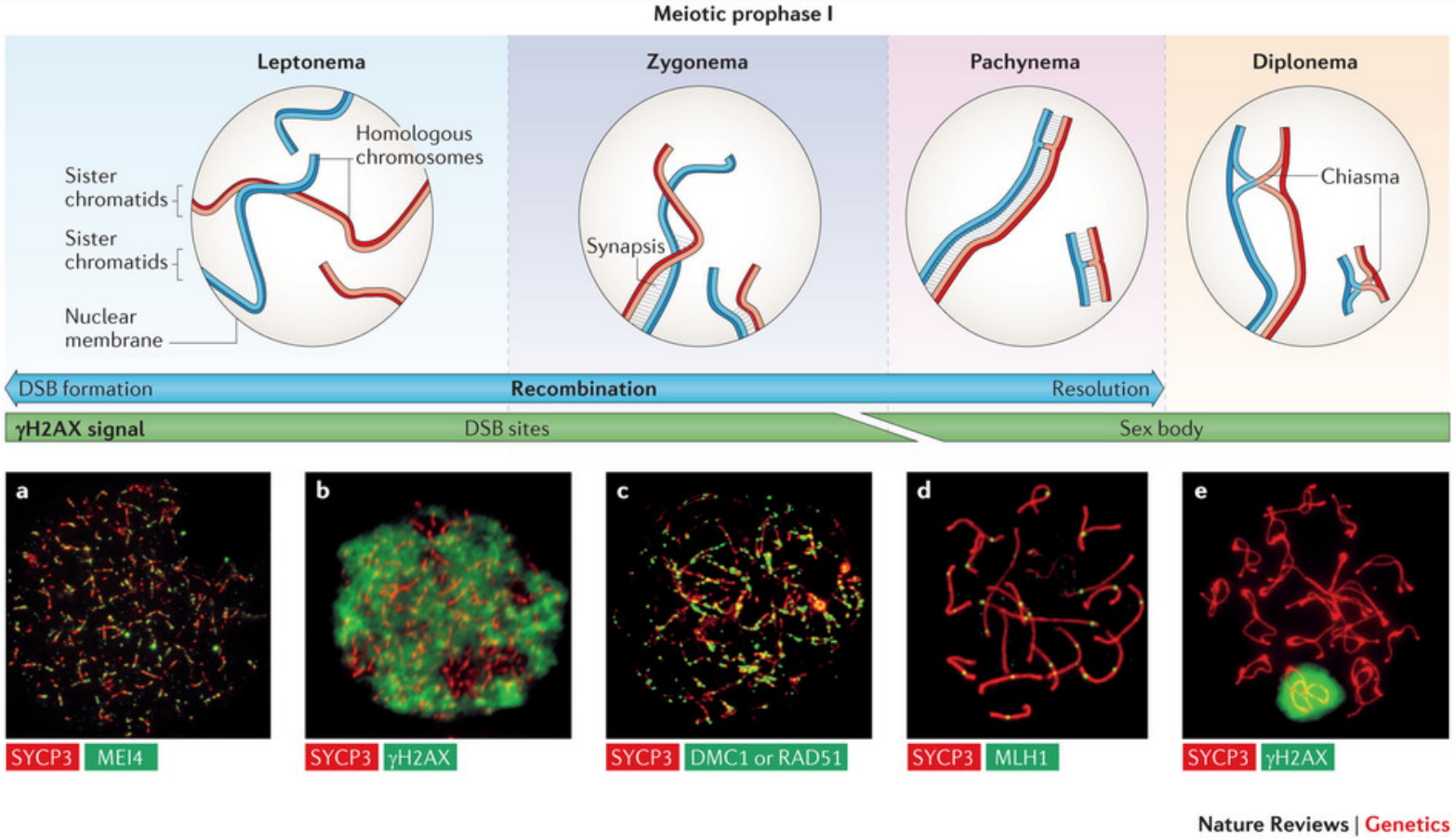
Derive these yourselves, with the help of the textbook

- Tri-allelic loci
- X-linked loci

Linkage and Linkage Disequilibrium

Meiosis

Baudat et al. (2013)[1]



Linkage

- Genes on non-homologous chromosomes assort independently during meiosis
- Genes that are physically close to each other on a chromosome on the other hand are said to be "linked"
- Linked genes do not assort independently.
- Linkage is broken down due to homologous recombination (crossovers) during meiosis 1.
- The longer the evolutionary time-scale, the chance for recombination to break down chromosomal haplotypes.

Linkage and Recombination Rate

Consider two genetic loci, A and B, with alleles A, a and B, b respectively.

What are the expected genotypes at both loci in the population under HWE?

Thus A allele is independent of the a allele (due to Mendel's principle of segregation), and B allele is independent of the b allele, i.e. they are in *linkageequilibrium*.

But how about the A allele and the B allele?

Ergo...

Linkage Disequilibrium

The non-random association of alleles at different sites across a genome in a population.

Linkage and Recombination Rate

Remember, that gametes can be either "parental" or "recombinant" (remember BIOL 352?)

Assume that the population allele frequencies of A, B, a, b alleles are p_A, p_B, q_a, q_b respectively, such that $p_A + q_a = 1$ and $p_B + q_b = 1$.

Non-recombinants and Recombinants

Possible zygote genotypes are: AB, ab, Ab, aB, with the probabilities (or expected frequencies) $p_A \times p_B$, $q_a \times q_b$, $p_A \times q_b$, and $q_a \times p_B$ respectively.

Frequency of Recombination

Symbolized as r is the proportion of recombinant gametes produced by a double heterozygote, i.e. the probability of recombination between two genes = Pr(SCO1)+Pr(SCO2)+....

E.x. if AB/ab produces AB, ab, aB, Ab gametes in the proportions $0.38, 0.38, 0.12, 0.12$, then $r = 0.12 + 0.12 = 0.24$.

Questions

1) What would be the expected distribution of r with physical distance between two genes?

2) What is the range of r?

Expected Frequencies

In a population in linkage equilibrium, we would expect that:

$$P_{AB} = p_A \times p_B$$

$$P_{Ab} = p_A \times q_b$$

$$P_{aB} = q_a \times p_B$$

$$P_{ab} = q_a \times q_b$$

Such that $P_{AB} + P_{Ab} + P_{aB} + P_{ab} = 1$

Now let's derive the genotype frequencies if the population is in LD.

LD

$D_{AB} = p_{AB} - p_A \times p_B$

$D_{Ab} = p_{Ab} - p_A \times q_b$

$D_{aB} = p_{aB} - q_a \times p_B$

$D_{ab} = p_{ab} - q_a \times q_b$

Also, $D_{AB} = -D_{Ab}$, $D_{ab} = D_{AB}$, $D_{Ab} = D_{aB}$. So it's enough if we know D_{AB}, let's call this D.

So if $D = 0$, then the population is in linkage equilibrium, otherwise in LD.

Recall

Also, $p_{AB} = p_A p_B + D$

What is covariance?

Alternately, $r^2 = \frac{D}{p_A p_B q_a q_b}$, called the correlation.

LD over generations

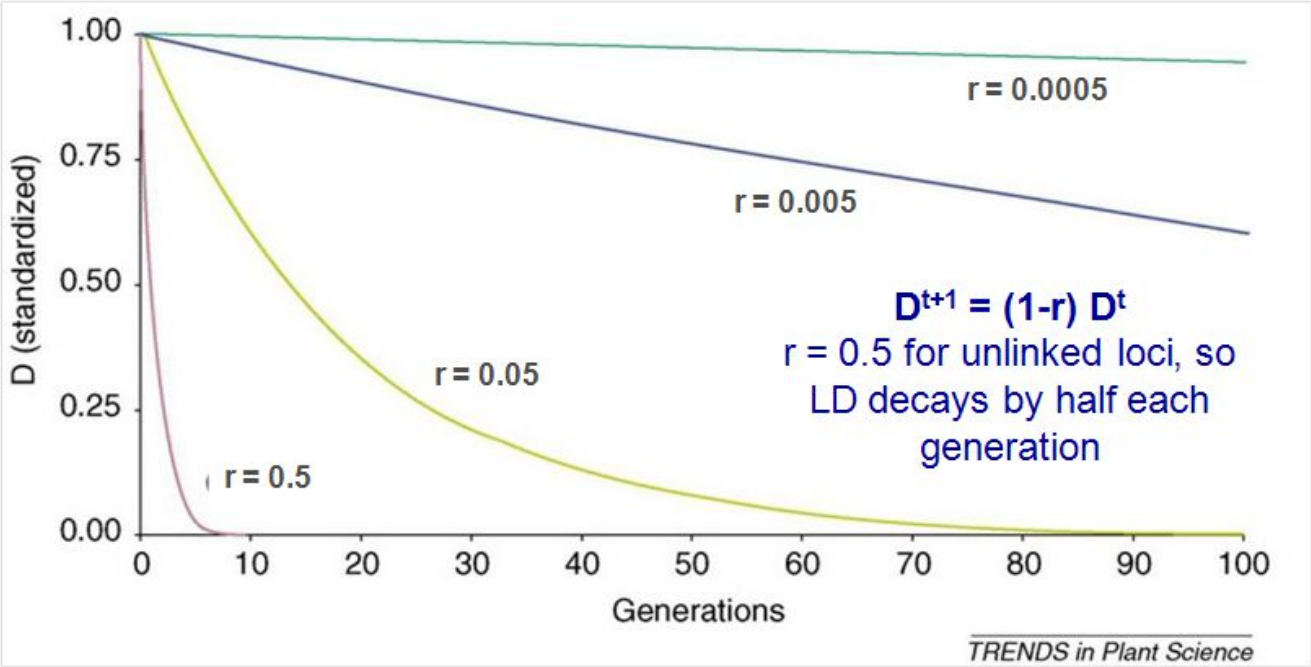
Now let's derive the general case of LD change over generations - assume that r is the recombination rate (frequency) as before.

Question

What is the expected probability of AB genotypes in the next generation as a function of recombination rates?

LD Decay

Rate of LD decay driven by recombination (r)



D is expressed in standardized units as D' or r^2

Mackay & Powell. 2007. TIPS 12: 57-63

χ^2 Test of LD

E.x. In a sample of 1000 British people, genotype counts at two genes indicated counts of 298 AA, 489 AG, 213 GG individuals for the first SNP, and 99 TT, 418 TC, and 483 CC for the second SNP. Compute if this population is in LD or not.

Causes of LD

- Mutation
- Admixture
- Recombination rate variation across the genome

Conclusion

Summary

- What are the two tenets of HWE?
- What is LD?
- Statistical tests for HWE and LE/LD

Questions?

VCF Files

VCF = Variant Call Format

`https://samtools.github.io/hts-specs/VCFv4.2.pdf`

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        N
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1
```

References I

F. Baudat, Y. Imai, and B. de Massy.
Meiotic recombination in mammals: localization and regulation.
Nature Reviews Genetics, 14(11):794–806, 2013.

M. G. A. A. E. V. D. C. M. B. M. W. J. P. A. A. B. P. Z. I. M. N. J. V. v. V.-O. H. S. T. L. C. S. G. I. o. A. T. G. c. S. E. M. N. G. M. P. K. E. M. W. G. I. M. M. K. E. N. J. Y. . P. M. V. Matthew R. Robinson, Aaron Kleinman.
Genetic evidence of assortative mating in humans.
Nature Human Behavior, 2017.