Applied Statistical Genetics in R

for population-based association studies

PART I: Genetic data concepts and tests

Outline

1. Overview of population-based investigations
   - Types of studies
   - Genotype versus gene expression
   - Population versus family-based
2. Data examples, features and challenges
   - FAMuSS
   - HGDP
   - Virco
3. Linkage disequilibrium (LD)
   - Measuring LD with D' and r^2
   - LD and population stratification
4. Hardy-Weinberg equilibrium (HWE)
   - Measuring HWE with χ^2 and Fisher's exact tests

Overview of population-based investigations

- Types of studies
  - Candidate polymorphism studies
  - Candidate gene studies
  - Fine mapping studies
  - Genome-wide association studies

Population-based investigations

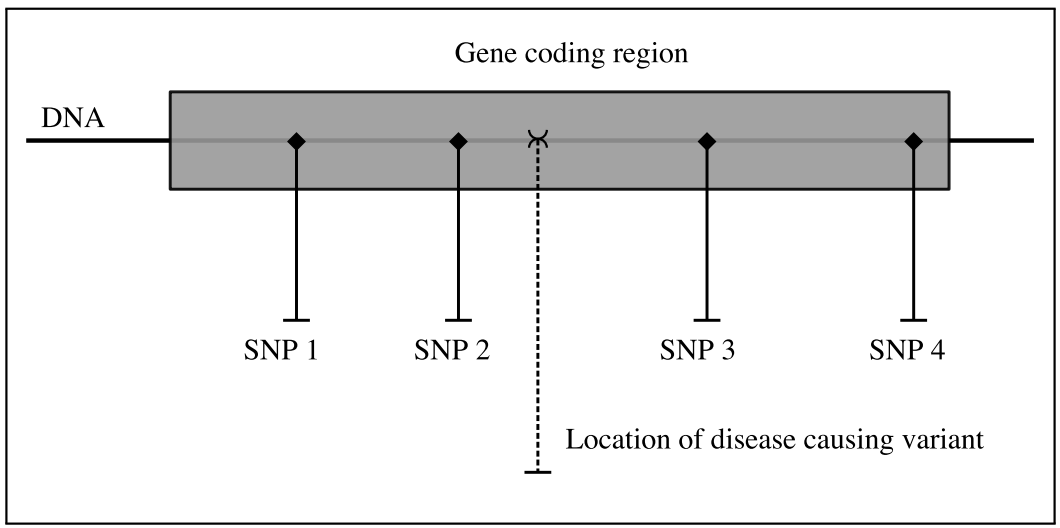
- Candidate polymorphism studies
  - Investigation of genotype-trait association.
  - Based on *a priori* hypothesis about functionality.
  - Goal is to determine whether SNP has *direct* influence on disease trait.

Definitions:

- **polymorphism**: genetic variant occurring in greater than 1% of a population
- **single nucleotide polymorphism (SNP)**: variant at a single site (base pair position) on the genome.

Population-based investigations

- Candidate gene studies
  - Involve multiple SNPs within a single gene.
  - Choice of SNPs depends on defined linkage disequilibrium (LD) blocks.
  - SNPs under study capture underlying genetic variability, though may not be causal.



Population-based investigations

- Fine mapping studies
  - Goal is to identify *location* (position) of disease causing variant.
  - Can obviate need for marker-based investigations, reducing variability in associated tests.
  - Not a focus of this tutorial.

Population-based investigations

- Genome-wide association (GWA) studies
  - Include whole and partial genome-wide scans ($100 - 1,000$ Kb segments of DNA).
  - Use of array chip (Affymetrix) technology.
  - Additional pre-processing for data error checks and quality control is needed.
  - Analysis is similar to candidate gene studies.
  - Computationally intensive $\longrightarrow$ R packages GenABEL and SNPassoc developed for this context.

Population-based investigations

- Genotype versus gene expression
  - Genetic **association** studies consider the relationship between genetic *sequence* information (genotype) and trait.
  - Gene **expression** ("microarray") studies aim to characterize associations among gene *products* (amount of RNA or proteins) and disease outcomes.
- Variable coding different
  - Genotype = 3-level (categorical) variable
  - Gene product = continuous variable
- Model of association can differ
  - Genotype (independent variable) $\longrightarrow$ Trait (dependent variable)
  - Trait (independent variable) $\longrightarrow$ Gene product (dependent variable).

Population-based investigations

- Genotype versus gene expression
  - Data set-up often different:
    - Association studies: individuals are rows, genes are represented by columns.
    - Expression studies: genes are in rows, individuals are represented by columns.
  - R packages developed for gene expression data not always easily applicable to data arising from genetic association studies.

Population-based investigations

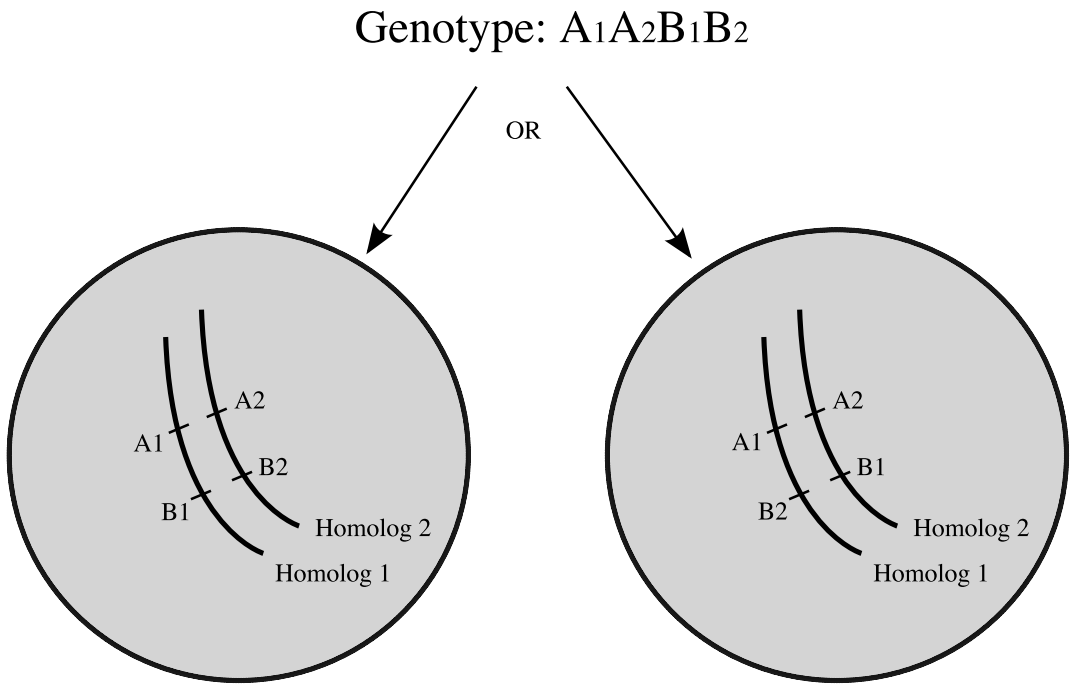
- Population versus family-based
  - **Population**-based investigations refer to studies involving unrelated individuals $\longrightarrow$ allelic phase is unobservable in population-based investigations.
  - **Family**-based studies involve data collected on multiple individuals within the same family unit $\longrightarrow$ require application of cluster data methods (e.g. GEE or mixed model) for appropriate inference. (These methods can also be relevant to population-based studies with repeated measures or other forms of clustering.)

Definitions:

- **allelic phase**: alignment of nucleotides on a single homolog.
- **homologous chromosomes**: chromosomes with potentially different alleles (genetic sequences) that carry information on the same trait (feature.)

Population-based investigations

- Population versus family-based
  - Example of ambiguous phase for A_1A_2, B_1B_2 genotype:



Data examples

1. Functional SNPs Associated with Muscle Size and Strength (FAMuSS) study
2. Human Genome Diversity Project (HGDP) study
3. Virco HIV sequence data

Data examples

1. Functional SNPs Associated with Muscle Size and Strength (FAMuSS) study
   - Goal: to identify genetic determinants of skeletal muscle size and strength before and after exercise training.
   - Includes $n = 1397$ college student volunteers.
   - Data on 227 SNPs across multiple genes collected.
   - Extensive clinical and demographic data available on subset (n=1035) of study participants.

Data examples

- Reading data into R:

```
> fms <- read.delim("http://people.umass.edu/foulkes/asg/
+ data/FMS_data.txt", header=T, sep="\t")
```

- Displaying data in table:

```
> install.packages("xtable")
> library(xtable)
> attach(fms)
> print(xtable(data.frame(id,actn3_r577x,actn3_rs540874,
+ actn3_rs1815739, Gender, Age, Race, NDRM.CH)[1:20,]))
```

Data examples

Table: Sample FAMuSS Data

	id	r577x	rs540874	rs1815739	Gender	Age	Race	NDRM.CH
1	FA-1801	CC	GG	CC	Female	27	Caucasian	40.00
2	FA-1802	CT	GA	TC	Male	36	Caucasian	25.00
3	FA-1803	CT	GA	TC	Female	24	Caucasian	40.00
4	FA-1804	CT	GA	TC	Female	40	Caucasian	125.00
5	FA-1805	CC	GG	CC	Female	32	Caucasian	40.00
6	FA-1806	CT	GA	TC	Female	24	Hispanic	75.00
7	FA-1807	TT	AA	TT	Female	30	Caucasian	100.00
8	FA-1808	CT	GA	TC				
9	FA-1809	CT	GA	TC	Female	28	Caucasian	57.10
10	FA-1810	CC	GG	CC	Male	27	Hispanic	33.30
11	FA-1811	CC	GG	CC				
12	FA-1812	CT	GA	TC	Female	30	Caucasian	20.00
13	FA-1813	CT	GA	TC	Female	20	Caucasian	25.00
14	FA-1814	CT	GA	TC	Female	23	African Am	100.00
15	FA-1815							
16	FA-1816	TT	GA	TC	Female	24	Caucasian	28.60
17	FA-1817	CT	GA	TC				
18	FA-1818	CT	GA	TC				
19	FA-1819	CT	GG	CC	Male	34	Caucasian	7.10
20	FA-1820	CC	GA	TC	Female	31	Caucasian	75.00

Definition:

- **bi-allelic**: detectable presence of two alleles across a population

Data examples

2. The Human Genome Diversity Project (HGDP)
   - Goal: to document and characterize genetic variation in humans worldwide
   - Genetic and demographic data recorded on $n = 1064$ individuals across 27 countries.
   - Genotype data on 4 SNPs in the AKT1 gene available.

Data examples

- Reading data into R:

```
> HGDP.dat <- read.delim("http://people.umass.edu/foulkes/asg/
+ data/HGDP_AKT1.txt",header=T,sep="\t")
```

Table: Sample HGDP data

	ID	Population	Geographic.origin	Geographic.area	AKT1.C0756A
1	HGDP00980	Biaka Pygmies	Central African Republic	Central Africa	CA
2	HGDP01406	Bantu	Kenya	Central Africa	CA
3	HGDP01266	Mozabite	Algeria (Mzab)	Northern Africa	AA
4	HGDP01006	Karitiana	Brazil	South America	AA
5	HGDP01220	Daur	China	China	AA
6	HGDP01288	Han	China	China	AA
7	HGDP01246	Xibo	China	China	AA
8	HGDP00705	Colombian	Colombia	South America	AA
9	HGDP00706	Colombian	Colombia	South America	AA
10	HGDP00707	Colombian	Colombia	South America	AA
11	HGDP00708	Colombian	Colombia	South America	AA
12	HGDP00709	Colombian	Colombia	South America	AA
13	HGDP00710	Colombian	Colombia	South America	AA
14	HGDP00598	Druze	Israel (Carmel)	Israel	AA
15	HGDP00684	Palestinian	Israel (Central)	Israel	AA

Data examples

3. Virco^{TM} HIV sequence data
   - Goal: to characterize association of Human immunodeficiency virus (HIV) sequence information and drug-specific fold-resistance measures.
   - Data on $n = 1066$ viral isolates are available.
   - Fold-resistance on 8 protease (Pr) inhibitors (drugs) measured.
   - Protease sequence is comprised of 99 amino acids (categorical variables).

Data examples

- Reading data into R:

```
> virco <- read.csv("http://people.umass.edu/foulkes/asg/
+ data/Virco_data.csv", header=T, sep=",")
```

Table: Sample Virco data

	SeqID	IDV.Fold	P10	P63	P90	CompMutList
1	3852	14.20	I	P	M	L10I, M46I, L63P, G73CS, V77I, L90M, I93L
2	3865	13.50	I	P	M	L10I, R41K, K45R, M46I, L63P, A71V, G73S, ...
3	7430	16.70	I	P	M	L10I, I15V, K20M, E35D, M36I, I54V, R57K, ...
4	7459	3.00	I	P	M	L10I, L19Q, E35D, G48V, L63P, H69Y, A71T, ...
5	7460	7.00	-	-	-	K14R, I15V, V32I, M36I, M46I, V82A
6	7461	21.00	I	P	M	L10I, K20R, M36I, N37D, I54V, R57K, D60E, ...
7	7462	8.00	-	P	-	M36I, G48V, I54V, D60E, I62V, L63P, V82A
8	7463	100.00	I	-	M	L10I, I13V, M36I, N37D, G48V, I54V, D60E, ...
9	7464	18.00	-	P	-	V32I, M46I, L63P, V82A, I93L
10	7465	15.00	-	I	M	E34K, R41K, K43R, I54V, I62V, L63I, A71V, ...
11	7466	4.00	I	P	-	L10I, E35D, M36I, G48V, D60E, L63P, H69Y
12	7467	45.00	-	P	-	I13V, K14R, K20M, E35D, M36I, N37D, K45R, ...
13	15492	1.00	X	-	-	L10X, I15V, I50V, I62V, A71V, I72V, N83Z
14	15493	1.00	F	A	-	L10F, I13V, L33F, M46X, I50V, L63A, T74S, ...
15	15494	2.00	F	-	-	L10F, V32I, M46I, I47V, I62V

Data features

1. Genotype:
   - Observed genetic sequence (string of categorical variables).
   - Human studies – each element of the string is a pair of DNA bases (A, C, T, or G).
2. Trait:
   - Quantitative (continuous) measure of disease progression OR
   - Dichotomous (binary) indicator for presence of disease.
3. Covariates:
   - Clinical and demographic factors.
   - Potential confounders, effect modifiers and effect mediators.

Definitions:

- **Multi-locus genotype**: observed genotype across multiple SNPs or genes.
- **Haplotype**: specific combination of alleles that are in *alignment* on a single homolog and tend to be inherited together.
- **Diplotype**: pair of haplotypes, one inherited from each parental genome.
- **Zygosity**: comparative genetic make-up of two homologous chromosomes:
  - **homozygous**: two observed bases/alleles are the same (both variant/minor allele or both wildtype/major allele.)
  - **heterozygous**: two observed bases/alleles are different.

Data features

- **Example 1** (Identifying the minor allele and its frequency):
  - Suppose we are interested in determining the minor allele for the SNP labeled `actn3_rs540874` in the FAMuSS data.
  - Determine number of observations with each genotype:

```
> attach(fms)
> GenoCount <- summary(actn3_rs540874)
> GenoCount

actn3_rs540874
 AA  GA  GG
226 595 395
```

(Note: n=181 individuals are missing this genotype.)

Data features

- (example continued)
  - Determine sample size and calculate genotype frequencies:

```
> NumbObs <- sum(!is.na(actn3_rs540874))
> GenoFreq <- as.vector(GenoCount/NumbObs)
> GenoFreq
```

```
[1] 0.1858553 0.4893092 0.3248355
```

  - Determine allele frequencies:

```
> FreqA <- (2*GenoFreq[1] + GenoFreq[2])/2
> FreqA  # minor allele frequency
```

```
[1] 0.4305099
```

```
> FreqG <- (GenoFreq[2] + 2*GenoFreq[3])/2
> FreqG  # major allele frequency
```

```
[1] 0.5694901
```

  - We see from the output that A represents the minor allele with frequency $= 0.43$.

Data features

- (example continued)
  - Alternatively, we can use the genetics package:

```
> install.packages("genetics")
> library(genetics)
> Geno <- genotype(actn3_rs540874, sep="")
> summary(Geno)
Number of samples typed: 1216 (87%)

Allele Frequency: (2 alleles)
   Count Proportion
G   1385       0.57
A   1047       0.43
NA   362         NA

Genotype Frequency:
    Count Proportion
G/G   395       0.32
G/A   595       0.49
A/A   226       0.19
NA    181         NA

Heterozygosity (Hu)  = 0.4905439
Poly. Inf. Content   = 0.3701245
```

Analytic challenges

1. Pre-processing: genetic data concepts and tests (Part 1)
2. Multiplicity: inflation of type-1 error due to multiple testing (Part 2)
3. Ambiguous phase: unobservable haplotypes (Part 3)
4. High-dimensionality: complex, uncharacterized relationships among genetic markers (Part 4)

Linkage disequilibrium (LD)

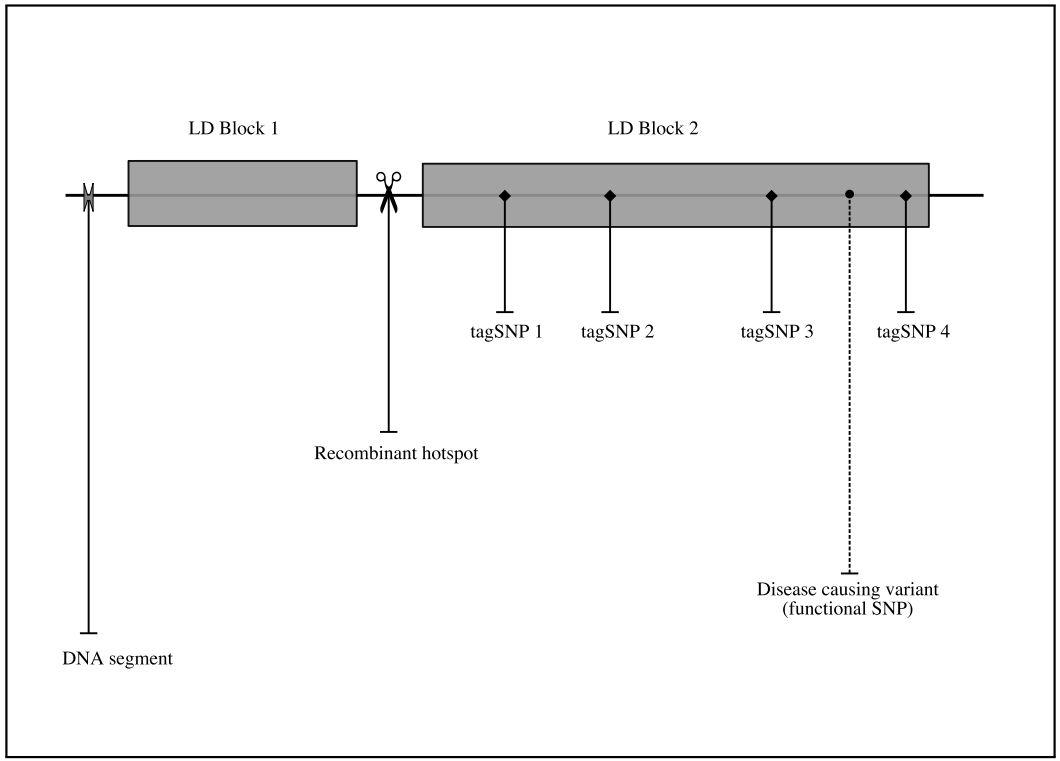
- SNPs under study may not directly cause the disease
- SNPs are correlated with disease because they are in *linkage disequilibrium* (LD) with true disease causing (functional) variant.
- Commonly measured by D' and r^2.

Definition:

- **Linkage disequilibrium (LD)**: an association in the alleles present at each of two sites on a genome.

Linkage disequilibrium (LD)

Figure: Illustration of LD blocks and associated tag SNPs



Linkage disequilibrium (LD)

- Suppose A and a are the possible alleles at Site 1, B and b are the alleles at Site 2 and p_A, p_a, p_B and p_b denote the corresponding population frequencies.
- Under independence, $pr(AB) = p_A p_B$.

Table: Expected allele distribution under independence

		Site 2		
		B	b	
Site 1	A	$n_{11} = Np_A p_B$	$n_{12} = Np_A p_b$	$n_{1.} = Np_A$
	a	$n_{21} = Np_a p_B$	$n_{22} = Np_a p_b$	$n_{2.} = Np_a$
		$n_{.1} = Np_B$	$n_{.2} = Np_b$	$N^* = 2n$

*N homologs across n individuals in our sample.

Linkage disequilibrium (LD)

- Under independence (linkage equilibrium), we expect the observed data to support the numbers above.
- If Sites 1 and 2 are associated, the numbers will deviate by an amount represented by the scalar D:

Table: Expected allele distribution under linkage disequilibrium

		Site 2		
		B	b	
Site 1	A	$n_{11} = N(p_A p_B + D)$	$n_{12} = N(p_A p_b - D)$	$n_{1.}$
	a	$n_{21} = N(p_a p_B - D)$	$n_{22} = N(p_a p_b + D)$	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$N = 2n$

Linkage disequilibrium (LD)

- We can write: $D = p_{AB} - p_A p_B$ and estimate p_{AB} using an expectation maximization (EM) algorithm.
- EM approach (or alternative) is needed since in practice, the number of homologs with A and B alleles (Np_{AB}) is not observed.
- A rescaled value of D takes into account the constraints on cell counts:

$$D' = \frac{|D|}{D_{max}}$$

where D_{max} represents the upper bound on D and is given by:

$$D_{max} = \begin{cases} min\left(p_{A_1} p_{B_2}, p_{A_2} p_{B_1}\right) & D > 0 \\ min\left(p_{A_1} p_{B_1}, p_{A_2} p_{B_2}\right) & D < 0 \end{cases}$$

Linkage disequilibrium (LD)

- **Example 2** (Measuring LD using D'):
  - Suppose we aim to calculate LD as measured by D' for two SNPs within the gene alpha-actinin 3 (actn3) based on FAMuSS data.
  - Create genotype objects:

```
> attach(fms)
> actn3_r577x[1:10]

 [1] CC CT CT CT CC CT TT CT CT CC
Levels:  CC CT TT

> actn3_rs540874[1:10]

 [1] GG GA GA GA GG GA AA GA GA GG
Levels:  AA GA GG

> library(genetics)
> Actn3Snp1 <- genotype(actn3_r577x,sep="")
> Actn3Snp2 <- genotype(actn3_rs540874,sep="")
```

Linkage disequilibrium (LD)

- (example continued)
  - This yields:

```
> Actn3Snp1[1:10]

 [1] "C/C" "C/T" "C/T" "C/T" "C/C" "C/T" "T/T" "C/T" "C/T" "C/C"
Alleles: C T

> class(Actn3Snp1)

[1] "genotype" "factor"

> LD(Actn3Snp1,Actn3Snp2)$"D'"

[1] 0.8858385
```

  - Result of $D' = 0.89$ suggests that there is a high degree of LD between SNPs labeled `r577x` and `rs540874` within the `actn3` gene.

Linkage disequilibrium (LD)

- **Example 3** (Measuring pairwise LD for a group of SNPs):
  - Now suppose we are interested in calculating pairwise LD for a group of SNPs within the actn3 gene:
  - Create dataframe:

```
> Actn3Snp3 <- genotype(actn3_rs1815739,sep="")
> Actn3Snp4 <- genotype(actn3_1671064,sep="")
> Actn3AllSnps <- data.frame(Actn3Snp1,Actn3Snp2,Actn3Snp3,Actn3Snp4)
```

  - A matrix of pairwise LD measures is then given by the upper triangular elements of the D' matrix:

```
> LD(Actn3AllSnps)$"D'"
```

	Actn3Snp1	Actn3Snp2	Actn3Snp3	Actn3Snp4
Actn3Snp1	NA	0.8858385	0.9266828	0.8932708
Actn3Snp2	NA	NA	0.9737162	0.9556019
Actn3Snp3	NA	NA	NA	0.9575870
Actn3Snp4	NA	NA	NA	NA

Linkage disequilibrium (LD)

- A related measure of LD is r^2, given by:

$$r^2 = \chi_1^2/N = \frac{D^2}{p_A p_B p_A p_B}$$

where χ^2 is the usual Pearson's test statistic:

$$\chi_1^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Notably, the difference between D' and r^2 rests in the type of adjustment made to the scalar D.

Linkage disequilibrium (LD)

- **Example 4** (Measuring LD based on r^2 and the χ^2-statistic):
  - Suppose that we are again interested in measuring LD between the SNPs `r577x` and `rs540874` within the `actn3` gene based on the FAMuSS data.
  - Again using the LD() function:

```
> attach(fms)
> library(genetics)
> Actn3Snp1 <- genotype(actn3_r577x,sep="")
> Actn3Snp2 <- genotype(actn3_rs540874,sep="")
> LD(Actn3Snp1,Actn3Snp2)$"R^2"

[1] 0.6179236
```

  - Result is based on the $n = 725$ individuals with complete data on both SNPs.
  - Corresponding χ^2-statistic is $0.6179236 * 725 * 2 = 896$.

Linkage disequilibrium (LD)

- (example continued)
  - The `LD()` function also returns this statistic and a corresponding "p-value", as seen below with the complete function output:

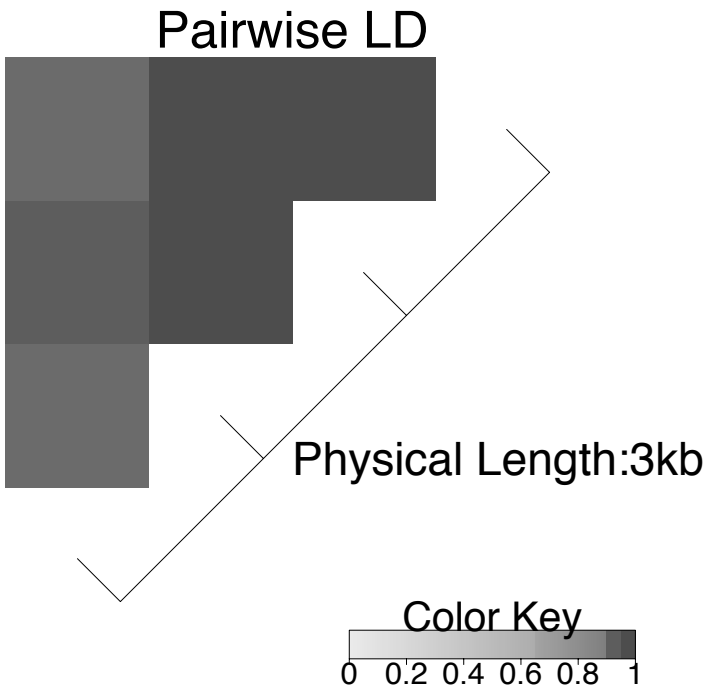
```
> LD(Actn3Snp1,Actn3Snp2)

Pairwise LD
-----------
                    D         D'       Corr
Estimates: 0.1945726 0.8858385 0.7860811

               X^2 P-value   N*
LD Test: 895.9891       0 725

(*note: N in this output is represented by n in above notation.)
```

  - Caution in interpretation as test: (1) Generated based on 2 observations per person (correlated data) and (2) Cell counts are estimated (introducing additional variability.)

Linkage disequilibrium (LD)

- **Example 5** (Plotting pairwise LD):

```
> install.packages("LDheatmap")
> library(LDheatmap)
> LDheatmap(Actn3AllSnps,
+ LDmeasure="D'")
```



Linkage disequilibrium (LD)

- Alternative R packages / functions:
  - SNPassoc: LD()
  - gap: LD22()
  - mapLD: mapLD()

Linkage disequilibrium (LD)

- Population stratification can lead to erroneous conclusions about the presence of LD between two SNPs.
- Commonly referred to as "Simpson's paradox" in biostatistics.

Definition:

- **Population stratification**: presence of multiple subgroups between which there is minimal mating or gene transfer

Linkage disequilibrium (LD)

- **Example 6** (Population substructure and LD)

Table: Haplotype distribution assuming linkage equilibrium and varying dominant allele frequencies

Population 1		Site 2 B	Site 2 b	
Site 1	A	$200 * 0.8^2 = 128$	$200 * 0.8 * 0.2 = 32$	160
	a	$200 * 0.8 * 0.2 = 32$	$200 * 0.2^2 = 8$	40
		160	40	$N = 200$

(a) Assuming dominant allele frequencies of 0.8

Population 2		Site 2 B	Site 2 b	
Site 1	A	$200 * 0.2^2 = 8$	$200 * 0.8 * 0.2 = 32$	40
	a	$200 * 0.8 * 0.2 = 32$	$200 * 0.8^2 = 128$	160
		40	160	$N = 200$

(b) Assuming dominant allele frequencies of 0.2

Linkage disequilibrium (LD)

Table: Apparent LD in the presence of population stratification

Populations 1 & 2 Combined		Site 2 B	Site 2 b	
Site 1	A	$128 + 8 = 136$	$32 + 32 = 64$	200
	a	$32 + 32 = 64$	$8 + 128 = 136$	200
		200	200	$N = 400$

- (example continued)
  - Calculating observed and expected counts

```
> ObsCount <- matrix(c(136,64,64,136),2)
> ObsCount
     [,1] [,2]
[1,]  136   64
[2,]   64   136
> ExpCount <- chisq.test(ObsCount)$expected
> ExpCount
     [,1] [,2]
[1,]  100  100
[2,]  100  100
```

Linkage disequilibrium (LD)

- (example continued)
  - Taking the absolute difference between the observed and expected counts and dividing by N yields $D = 36/400 = 0.09$.
  - Based on the observed cell counts of the combined populations, we have $p_A = p_a = p_B = p_b = 0.5$
  - This yields $D_{max} = 0.25$ and $D' = 0.09/0.25 = 0.36$.
  - Caution: combining data across the two populations and not accounting for the resulting substructure in our analysis leads to the incorrect conclusion that there is mild pairwise LD.

Hardy-Weinberg equilibrium (HWE)

- Tests of HWE – Pearson's χ^2 and Fisher's exact
- Departures from HWE can be result of (1) association, (2) population stratification/admixture or (3) data errors.
- Use of genomic controls can help decipher reason for departures.

Definition:

- **Hardy Weinberg Equilibrium (HWE)**:
  1. allele frequencies are constant within a population over generations
  2. independence of alleles at a single site between two homologous chromosomes
- **Population admixture**: setting in which mating occurs between two populations for which the allele frequencies differ.

Hardy-Weinberg equilibrium (HWE)

Table: Genotype counts for two homologous chromosomes

		Homolog 2 A	Homolog 2 a	
Homolog 1	A	n_{11}	n_{12}	$n_{1.}$
	a	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

- The expected counts corresponding to these three observed counts, n_{11}, $n_{12}^* = n_{21} + n_{12}$, n_{22} are given by:

$$E_{11} = Np_{A_1}^2$$
$$E_{12} = 2Np_{A_1}(1 - p_{A_1})$$
$$E_{22} = N(1 - p_{A_1})^2$$

Hardy-Weinberg equilibrium (HWE)

- Let $p_{A_1} = (2n_{11} + n_{12}^*)/(2N)$.
- The χ^2-test statistic is then constructed in the usual way as:

$$\chi_1^2 = \sum_{(i,j)\in\mathcal{C}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where the summation is over the set $\mathcal{C}$ of three observed cells.

Hardy-Weinberg equilibrium (HWE)

- **Example 7** (Testing for HWE using Pearson's χ^2 test):
  - Suppose we are interested in testing for HWE for the SNP labeled `AKT1.C0756A.` in the HGDP data.
  - Calculate observed and expected counts to get χ^2 statistic:

```
> attach(HGDP.dat)
> Akt1Snp1 <- AKT1.C0756A.
> ObsCount <- table(Akt1Snp1)
> Nobs <- sum(ObsCount)
> ObsCount

Akt1Snp1
 AA  CA  CC
 48 291 724

> FreqC <- (2 * ObsCount[3] + ObsCount[2])/(2*Nobs)
> ExpCount <- c(Nobs*(1-FreqC)^2, 2*Nobs*FreqC*(1-FreqC),
+ Nobs*FreqC^2)
> ExpCount

[1]  35.22319 316.55362 711.22319

> ChiSqStat <- sum((ObsCount - ExpCount)^2/ExpCount)
> ChiSqStat

[1] 6.926975
```

Hardy-Weinberg equilibrium (HWE)

- (example continued)
  - This statistic has a χ^2 distribution with a single degree of freedom. The quantile corresponding to $1-\alpha$ where $\alpha = 0.05$ is given by:

```
> qchisq(1-0.05,df=1)

[1] 3.841459
```

  - Since $6.93 > 3.84$, based on this sample we would reject the null hypothesis of HWE at this SNP locus and conclude instead that the alleles on the two homologous chromosomes are associated with one another.

Hardy-Weinberg equilibrium (HWE)

- (example continued)
  - Alternatively, the `HWE.chisq()` function in the `genetics` package can be used to calculate this statistic:

```
> library(genetics)
> Akt1Snp1 <- genotype(AKT1.C0756A., sep="")
> HWE.chisq(Akt1Snp1)

        Pearson's Chi-squared test with simulated p-value
        (based on 10000 replicates)


data:  tab
X-squared = 6.927, df = NA, p-value = 0.007199
```

  - Same χ^2 statistic is returned with a corresponding $p = 0.0072$. Again, based on this result, we would reject the null hypothesis of HWE and conclude that there appears to be non-random mating.

Hardy-Weinberg equilibrium (HWE)

- (example continued)
  - In some instances. Fisher's exact test is more appropriate than the χ^2-test for assessing departures from HWE, e.g. expected cell counts are < 5 which occurs in the presence of rare alleles.

```
> attach(HGDP.dat)
> Akt1Snp1Maya <- AKT1.C0756A.[Population=="Maya"]
> ObsCount <- table(Akt1Snp1Maya)
> ObsCount

Akt1Snp1Maya
AA CA CC
 1  6 18

> Nobs <- sum(ObsCount)
> FreqC <- (2 * ObsCount[3] + ObsCount[2])/(2*Nobs)
> ExpCount <- c(Nobs*(1-FreqC)^2, 2*Nobs*FreqC*(1-FreqC),
+ Nobs*FreqC^2)
> ExpCount

[1]  0.64  6.72 17.64
```

Hardy-Weinberg equilibrium (HWE)

- (example continued)
  - Since the expected count for the first cell is less than 5, using Fisher's exact test to test for HWE is most appropriate. An exact probability of seeing the observed counts is given by `FisherP1` in the following code example:

```
> n11 <- ObsCount[3]
> n12 <- ObsCount[2]
> n22 <- ObsCount[1]
> n1 <- 2*n11+n12
> Num <- 2^n12 * factorial(Nobs)/prod(factorial(ObsCount))
> Denom <- factorial(2*Nobs) / (factorial(n1)*factorial(2*Nobs-n1))
> FisherP1 <- Num/Denom
> FisherP1

[1] 0.4011216
```

Hardy-Weinberg equilibrium (HWE)

- (example continued)
  - Alternatively, we can use the `HWE.exact()` function in the `genetics` package:

```
> library(genetics)
> Akt1Snp1Maya <- genotype(AKT1.C0756A.[Population=="Maya"], sep="")
> HWE.exact(Akt1Snp1Maya)

        Exact Test for Hardy-Weinberg Equilibrium

data:  Akt1Snp1Maya
N11 = 18, N12 = 6, N22 = 1, N1 = 42, N2 = 8, p-value = 0.4843
```

  - Based on this output, we see that the exact p-value is 0.4843 and we are unable to reject the null hypothesis that there is a departure from HWE in this population.

Hardy-Weinberg equilibrium (HWE)

- **Example 8** (HWE and geographic origin)
  - Consider role of geographic origin:

```
> attach(HGDP.dat)
> HWEGeoArea <- tapply(Akt1Snp1,INDEX=Geographic.area,HWE.chisq)
> HWEGeoArea$"Central Africa"


Pearson's Chi-squared test with simulated p-value (based on 10000
replicates)

data:  tab
X-squared = 0.2322, df = NA, p-value = 0.6649

> HWEGeoArea$"South America"

Pearson's Chi-squared test with simulated p-value (based on 10000
replicates)

data:  tab
X-squared = 27.2386, df = NA, p-value = 9.999e-05
```

Hardy-Weinberg equilibrium (HWE)

- Alternative R packages / functions:
  - SNPassoc: tableHWE()
  - HardyWeinberg: HWchisq(); HWTernaryPlot()
  - GenABEL: summary.snp.data(); descriptives.marker(); formetascore()
  - gap: hwe(); hwe.hardy()

Summary

- Topics covered:
  - Types of investigations
  - Sample data and features
  - Estimating linkage disequilbrium (LD) – D' and r^2.
  - Testing Hardy Weinberg equilibrium (HWE) – χ^2 and Fisher's exact.
- Useful R functions:
  - genetics: genotype(); LD(); HWE.chisq(); HWE.exact()
  - LDheatmap: LDheatmap()
  - SNPassoc: tableHWE(); LD()
  - HardyWeinberg: HWchisq(); ; HWTernaryPlot()
  - GenABEL: summary.snp.data(); descriptives.marker(); formetascore()
  - gap: hwe(); hwe.hardy(); LD22()
  - mapLD: mapLD()
  - generic: read.delim(); read.csv(); data.frame(); attach(); summary(); table(); matrix(); chisq.test(); sum(); prod(); factorial(); tapply().