3. Linkage Disequilibrium (LD)

Iván Galván-Femenía [1], Jan Graffelman [1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

November 17, 2016

ivan.galvan@upc.edu

Master in Innovation and Research in Informatics (MIRI)

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Masters in Computer Science and Engineering

Contents

LD

- LD: an association between the alleles at different sites in the genome.
- Maybe a consequence of the physical closeness of the sites, but not necessarily so.
- LD is an important concept in disease-marker association studies.

Linkage Disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE)

- Both concepts refer to association between alleles
- HWE refers to association between alleles at the same locus (within one marker)
- LD refers to association between alleles at different loci (between markers)

Measures of LD

- D (deviation from independence)
- Lewontin's $D' = \frac{D}{D_{max}}$
- R^2
- χ^2 statistic of a contingency table
- p-value in a chi-square test or in an exact test
- ...

Haplotype

- A **haplotype** is a combination of alleles at adjacent loci on a chromosome that are transmitted together to the next generation.
- In practice, a haplotype often refers to a set of SNPs on a single chromosome that are statistically associated.
- A haplotype map of the human genome has been constructed (www.hapmap.org).

LD

- Consider a population of n individuals
- Consider two sites (two bi-allelic markers)
- One marker with alleles A and a, and one marker with alleles B and b.
- Allele frequencies p_A; p_a; p_B and p_b.
- Expected probabilities of each haplotype under independence:

		SNP2		
		B	b	
SNP1	A	$p_A p_B$	$p_A p_b$	p_A
	a	$p_a p_B$	$p_a p_b$	p_a
		p_B	p_b	1

LD

Observed probabilities of each haplotype in presence of LD

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

where D can be:

$$D = p_{AB} - p_A p_B,$$
$$D = -p_{Ab} + p_A p_b,$$
$$D = -p_{aB} + p_a p_B,$$
$$D = p_{ab} - p_a p_b$$

if D $\approx$ 0 : no LD

How to compute LD?

- $D = p_{AB} - p_A p_B$
- p_A and p_B can be estimated by the sample allele frequencies $\hat{p}_A$ and $\hat{p}_B$.
- However, p_{AB} is unknown.
- Because we have data at the genotype level, and p_{AB} is at the haplotype level.

The data

Observed genotype data

		SNP2		
		BB	Bb	bb
SNP1	AA	n_{AABB}	n_{AABb}	n_{AAbb}
	Aa	n_{AaBB}	n_{AaBb}	n_{Aabb}
	aa	n_{aaBB}	n_{aaBb}	n_{aabb}

- This data can be considered a sample from a multinomial distribution with 9 categories, where the probability of each of the 9 categories ultimately depends on the four haplotype probabilities p_{AB}; p_{Ab}; p_{aB} and p_{ab}.
- We will use a maximum likelihood (ML) approach

ML estimation

$$\boldsymbol{\theta} = (p_{AB}, p_{Ab}, p_{aB}, p_{ab}), \qquad \boldsymbol{x} = (n_{AABB}, n_{AABb}, \ldots, n_{aabb})$$

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{n!}{n_{AABB}! \cdot n_{aabb}!} \cdot (p_{AB}^2)^{n_{AABB}} \cdots (p_{ab}^2)^{n_{aabb}}$$

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = C + 2n_{AABB}\ln(p_{AB}) + \cdots + 2n_{aabb}\ln(p_{ab})$$

- The problem can be reparametrized in terms of p_A; p_B and p_{AB}
- Because $p_A = p_{AB} + p_{Ab}$, $p_B = p_{AB} + p_{aB}$ and $p_{AB} = 1 - p_{Ab} - p_{aB} - p_{ab}$
- Setting $\frac{\partial l}{\partial \theta} = 0$, no closed form solution can be found.
- We maximize the likelihood by a Newton-Raphson algorithm.
- Alternatively the expectation-maximization (EM) algorithm may be used.

Example data set

- Data from the FAMuSS (Functional SNPs Associated with Muscle Size and Strength) study (Foulkes, 2009)
- $n = 1397$ individuals and 225 SNPs
- Muscle performance variables

Computing LD in R

```
> fms <- read.delim("http://www.stat-gen.org/book.e1/data/FMS_data.txt",header=T,sep="\t")
> n <- nrow(fms)
> p <- ncol(fms)
> print(n)
[1] 1397
> print(p)
[1] 347
> 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
> Actn3Snp1 <- genotype(actn3_r577x,sep="")
> Actn3Snp2 <- genotype(actn3_rs540874,sep="")
> out <- LD(Actn3Snp1,Actn3Snp2)
> class(out)
[1] "LD"
> attributes(out)
$names
[1] "call" "D" "D'" "r" "R^2" "n" "X^2"
[8] "P-value"
$class
[1] "LD"
> out$D
[1] 0.1945726
> out$"D'"
[1] 0.8858385
```

ML estimation

It.	$l(P_{AB}, P_A, P_B \| x)$	P_{AB}	P_A	P_B
0	-1471.8874	0.0100000	0.508276	0.434483
1	-1469.9878	0.0438867	0.503479	0.429587
2	-1460.8970	0.0375485	0.514644	0.441162
3	-1459.0183	0.0297541	0.514183	0.440727
4	-1458.2618	0.0288494	0.508727	0.435198
5	-1458.0022	0.0263196	0.509216	0.435692
6	-1457.9928	0.0257361	0.507443	0.433847
7	-1457.9840	0.0251530	0.509738	0.432716
8	-1457.9716	0.0253836	0.508019	0.434685
9	-1457.9709	0.0257321	0.507963	0.434594
10	-1457.9696	0.0256473	0.508296	0.434473
11	-1457.9696	0.0256113	0.508247	0.434500
12	-1457.9696	0.0256208	0.508278	0.434481
13	-1457.9696	0.0256212	0.508276	0.434483

After convergence:

$p_{AB} = 0.0256212; \quad p_{Ab} = p_A p_{AB} = 0.4826544; \quad p_{aB} = p_B - p_{AB} = 0.408862;$

$p_{ab} = 1 - p_{Ab} - p_{aB} - p_{AB} = 0.08286239$

$$D = -0.1952159$$

D'

- $-0.25 \leq D \leq +0.25$
- D' is an attempt to standarize D

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

$$D_{max} = \begin{cases} min(p_A p_b, p_a p_B) & D > 0 \\ min(p_A p_B, p_a p_b) & D < 0 \end{cases}$$

- $-1 \leq D' \leq 1$
- $D' \approx 0$: low LD
- $|D'|$ close to 1 : high LD.

R^2 and χ^2 statistic

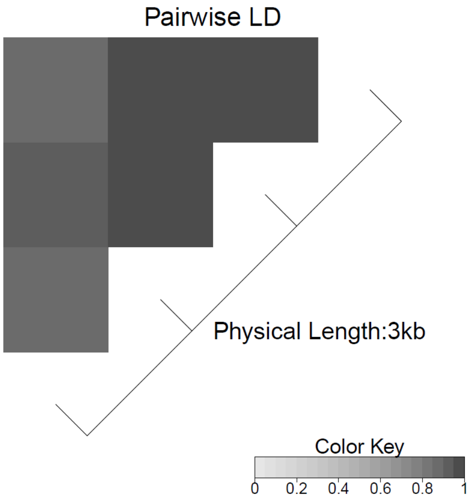
- The genotype data can be recoded as indicator data, creating indicators for the carriers of the A and B allele (AA=0, AB=1, BB=2).
- R^2 is the squared correlation between these indicators.
- R^2 is related to the χ^2 statistic of a 2x2 contingency table: $R^2 = \chi^2$ / (2n).
- The χ^2 is related to D:

$$R^2 = \chi^2/(2n) = \frac{D^2}{p_A p_B p_a p_b}$$

LD heatmap: graphics for LD with many SNPs

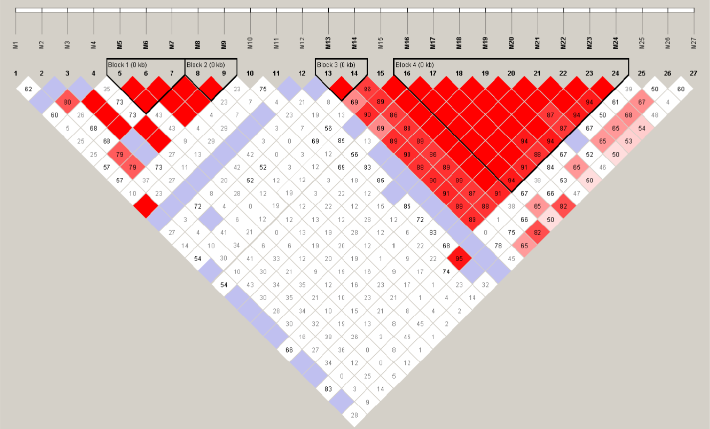
```
> install.packages("LDheatmap")
> library(LDheatmap)
> Actn3Snp1 <- genotype(actn3_r577x,sep="")
> Actn3Snp2 <- genotype(actn3_rs540874,sep="")
> Actn3Snp3 <- genotype(actn3_rs1815739,sep="")
> Actn3Snp4 <- genotype(actn3_1671064,sep="")
> ActnAll <- data.frame(Actn3Snp1,Actn3Snp2,Actn3Snp3,Actn3Snp4)
> LD(ActnAll)$"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
> LDheatmap(ActnAll,LDmeasure="D’")
```

LD heatmap

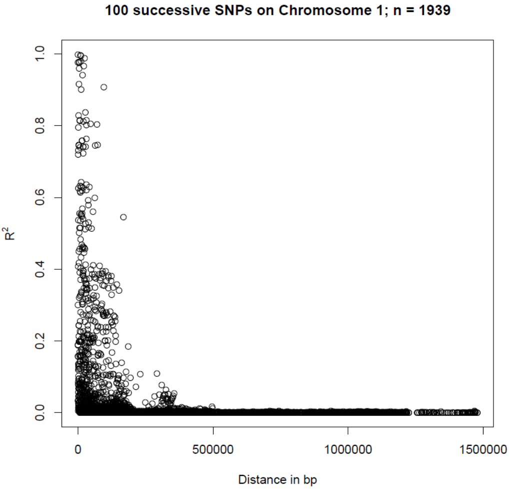


Another Heatmap (HaploView software)

100 (successive) SNPs on chromosome 1 of a sample of 45 individuals from a Chinese population of the HapMap project (www.hapmap.org), **27** remaining after removing monomorphics.



LD and physical distance



Computer exercises

1. Install the R packages `genetics`, `HardyWeinberg` and `LDheatmap`.
2. Load the database http://www-eio.upc.es/jan/data/bsg/CHBChr2-2000.rda
3. Calculate the statistics D; D'; R^2 and χ^2 for SNPs 12 and 13. Interpret the results.
4. Repeate the exercise 3 for SNPs 12 and 1000.
5. Select the first 100 SNPs from the database that have complete information (no missings).
6. Compute 4 matrices of association statistics, for D; D'; R^2 and χ^2 respectively.
7. Extract the subdiagonal part of each matrix into a vector.
8. Make a scatterplot matrix of the 4 association statistics. Are they related?
9. Make an LDheatmap for each of the four association statistics. Are the results similar?