Linkage Disequilibrium (LD)

Jan Graffelman[1]

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

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

jan.graffelman@upc.edu

November 19, 2019

Contents

1. Introduction to LD
2. LD statistics
3. Estimation of LD
4. Examples
5. Plink
6. Computer exercise

LD

- LD: an association between the alleles at different sites in the genome.
- The terms suggests this to be a consequence of the physical closeness of the sites, but this is 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 $D' = \frac{D}{D_{max}}$ ("standardization" of D)
- 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) on the same chromosome
- One marker with alleles A and a, and one marker with alleles B and b
- Four possible haplotypes: AB, Ab, aB and ab
- 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	$p_Ap_B + D$	$p_Ap_b - D$	p_A
	a	$p_ap_B - D$	$p_ap_b + D$	p_a
		p_B	p_b	1

$D = p_{AB} - p_Ap_B$ or $D = P_{AB}P_{ab} - P_{Ab}P_{aB}$

$D > 0$: known as "coupling"
$D < 0$: known as "repulsion"

How to compute D?

- $D = p_{AB} - p_A p_B$
- p_A and p_B can be estimated by the sample allele frequencies $\hat{p}_A$ and $\hat{p}_B$
- p_{AB} is unobserved and thus unknown
- 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 MN 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 approach

ML estimation

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

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

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

- The problem can be reparametrized in terms of p_A, p_B and P_{AB}
- (because $p_A = p_{AB} + p_{Ab}, p_B = p_{AB} + p_{aB}, p_{AB} = 1 - (p_{Ab} + p_{aB} + p_{ab})$)
- 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 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(file="c:/data/FMS_data.txt",header=TRUE,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.	$l(P_{AB}, P_A, P_B \mid x)$	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:

$$p_{AB} = 0.0256212, \; p_{Ab} = p_A - p_{AB} = 0.4826544, \; p_{aB} = p_B - p_{AB} = 0.408862, \; p_{ab} = 1 - (p_{AB} + p_{Ab} + p_{aB}) = 0.08286239$$

$$D = -0.1952159$$

D'

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

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

$$D_{max} = \begin{cases} \min\left(p_A p_b, p_a p_B\right) & D > 0 \quad \text{(coupling)} \\ \min\left(p_A p_B, p_a p_b\right) & D < 0 \quad \text{(repulsion)} \end{cases}$$

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

R^2 and χ^2 statistic

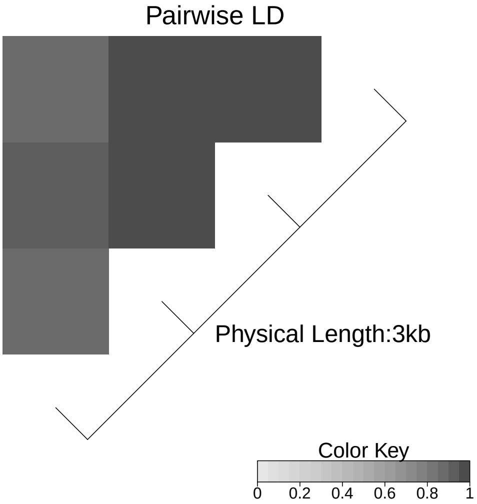
- The genotype data can be recoded as indicator data, creating indicators for the carriers of the A and B allele.
- R^2 is the squared correlation between these indicators.
- R^2 is related to the χ^2 statistic of a 2×2 contingency table: $R^2 = \chi^2/(2n)$.
- The χ^2 statistic is related to D

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

LD heatmap: graphics for LD with many SNPs

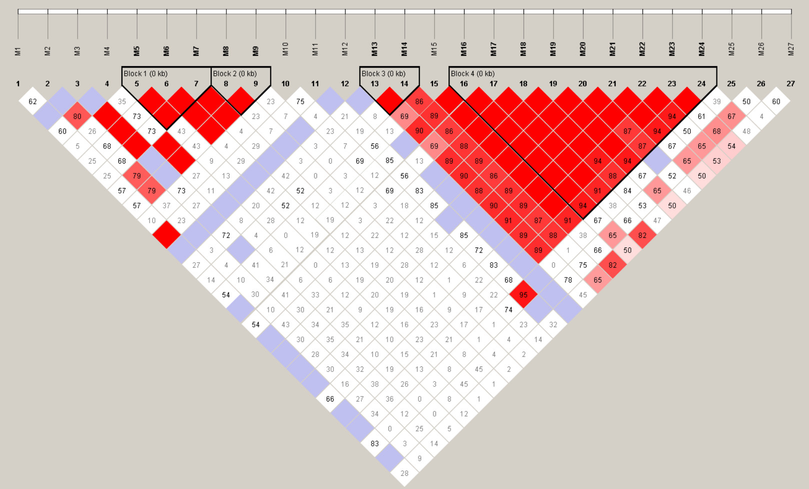
```
> install.packages("LDheatmap")
> library(LDheatmap)
> Actn3Snp1 <- genotype(actn3_r577x,sep="")
> Actn3Snp2 <- genotype(actn3_rs540874,sep="")
> Actn3Snp3 <- genotype(actn3_rs1815739,sep="")
> Actn3Snp4 <- genotype(actn3_1671064,sep="")
> ActnAll <- data.frame(Actn3Snp1,Actn3Snp2,Actn3Snp3,Actn3Snp4)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

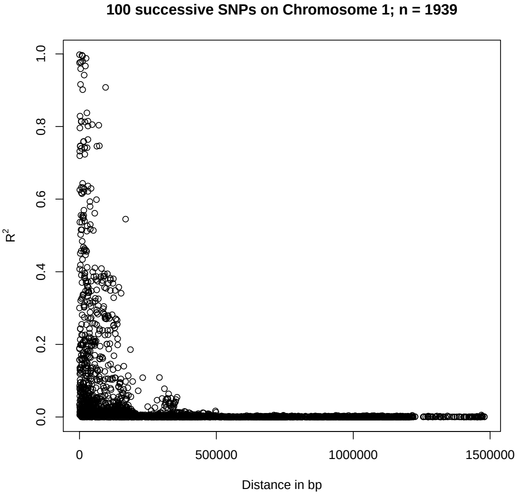


Another Heatmap (HaploView)

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

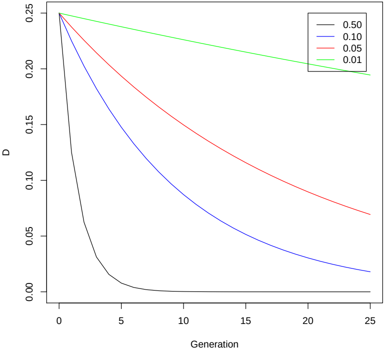


LD and recombination

- Let θ be the rate of recombination between two loci. $0 \leq \theta \leq 0.5$.
- Let D_0 be the degree of LD in generation 0. Then it can be shown that:

$$D_1 = (1 - \theta)D_0$$

$$D_t = (1 - \theta)^t D_0$$



The PLINK software (versions 1.9 or 2.0)

A widely used program in statistical genetics is the command-line based program PLINK

- Available at https://www.cog-genomics.org/plink2/
- maintained by Shaun Purcell and Christopher Chang
- Offers many options for
  - Data manipulation, format conversion.
  - Allele frequencies, missing data, Hardy-Weinberg equilibrium.
  - LD calculations, LD pruning.
  - Population substructure.
  - Kinship calculations.
  - Association analysis.
  - ...

Storage in plink data files

Often convenient to work with the triple of plink files:

- `.bed` file containing binary bi-allelic genetic variants
- `.bim` file containing information on the variants (identifier, chromosome, position, alleles)
- `.fam` containing sample (individual) information (sex, phenotype, family, father, mother)

A bit of PLINK (calling it from R)

```
#
# Generates .bim .fam and .bed files from a VCF file from the 1000 genomes project
#

plinkstring01 <- "plink --vcf ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf
                  --out DataChr22"
system(plinkstring01)

#
# Retain only variants without missing values
#

plinkstring02 <- "plink --bfile DataChr22 --geno 0 --make-bed --out DataChr22Nomissings"
system(plinkstring02)

#
# Retain only variants with a MAF about 0.05
#

plinkstring03 <- "plink --bfile DataChr22Nomissings --maf 0.05 --make-bed --out DataChr22NomissingsNoLowMa
system(plinkstring03)

#
# Filter on the Hardy-Weinberg p-value
#

plinkstring04 <- "plink --bfile DataChr22v3 --hwe 0.05 midp --make-bed --out DataChr22v4"
system(plinkstring04)
```

LD pruning

- It is often convenient to have independent genetic variants.
- Genetic variants that are physically close on a chromosome typically have high correlations.
- A subset of variants can be selected that is, at least approximately, independent.
- In practice, a window of fixed size is defined (in kb or as a variant number).
- An R^2 statistic can be calculated for each pair of variants in the window.
- Remove variants from the window until al remaining pairs of variants have $R^2 < t$, where t is some threshold.
- Shift the window along the chromosome, allowing for some overlap.
- The process is known as LD pruning.
- Easy to do in the PLINK software.

LD calculations in PLINK

```
#
# Compute pairwise LD statistics
#

plinkstring05 <- "plink --bfile DataChr22v4 --r2 --out Chr22LDstatistics"
system(plinkstring05)

#
# LD pruning
#

plinkstring06 <- "plink --bfile DataChr22v4 --indep-pairwise 50 5 0.2  --make-bed --out DataChr22v5"
system(plinkstring06)

plinkstring07 <- "plink --bfile DataChr22v5 --extract DataChr22v5.prune.in --make-bed --out DataChr22v6"
system(runstring07)
```

References

- Weir, B.S. (1996) *Genetic Data Analysis II*, Chapter 3, Sinauer Associates, Massachusetts.
- Foulkes, A.S. (2009) *Applied statistical genetics with R*. Springer.

Computer exercise

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