

15

Linkage Disequilibrium and Haplotype Inference

15.1: Learning Objectives

By the end of this chapter, the reader should be able to define linkage disequilibrium (LD), compute the standard LD coefficients D , D' and r^2 , relate LD decay to physical distance and recombination, and describe haplotype inference methods (Clark's algorithm, the EM algorithm, and Bayesian/PHASE-type methods).

15.2: What is Linkage Disequilibrium?

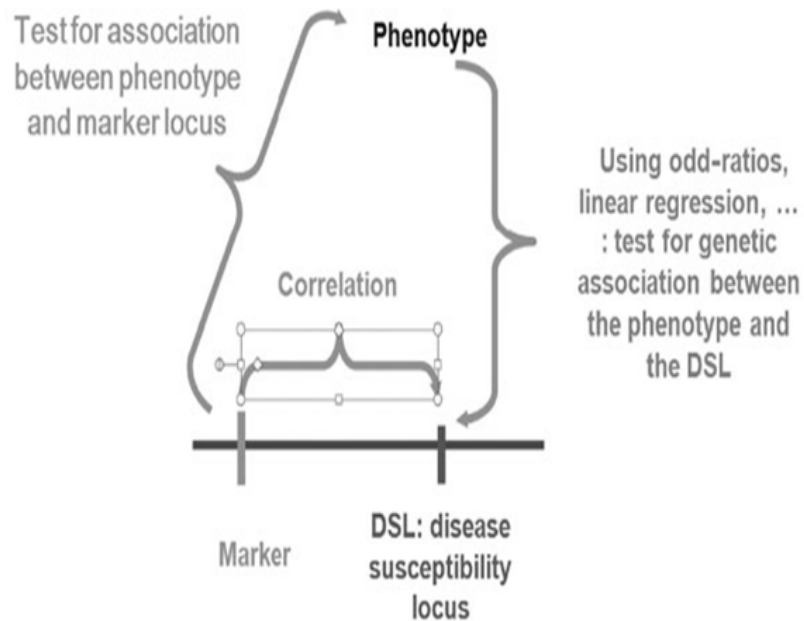
Two loci are said to be in **linkage equilibrium** if the alleles at the two loci occur together in gametes independently of one another — i.e., the haplotype frequency equals the product of the individual allele frequencies. **Linkage disequilibrium (LD)** is the non-random, statistical *association* between alleles at two (typically linked) loci, i.e., a departure from this independence, observed more often than chance in a population.

Consider two bi-allelic loci: locus 1 with alleles A/a (frequencies p_A , p_a) and locus 2 with alleles B/b (frequencies p_B , p_b). If haplotype AB has population frequency h_{AB} , and there were no LD, we would expect

$h_{AB} = p_A \times p_B$. The **disequilibrium coefficient**, D , measures the departure from this expectation:

$$D = h_{AB} - p_A p_B$$

Equivalently, $D = h_{AB} h_{ab} - h_{Ab} h_{aB}$ (the “cross-product” difference among the four haplotype frequencies). $D = 0$ corresponds to linkage equilibrium; $D \neq 0$ indicates LD (either positive, $D > 0$, meaning A and B co-occur more than expected, or negative, $D < 0$).



The logic of association testing via LD: a genotyped marker locus (M) that is in strong LD with an unobserved disease-susceptibility locus (DSL) can reveal a statistical association between the marker and the phenotype, even though the marker itself has no direct biological effect — the correlation is the only reason a marker-based association test (Chapter 11) can succeed without genotyping the causal variant directly.

15.3: Why LD Arises and Decays

LD between two loci commonly arises from: (i) physical proximity, so that a new mutation at one locus initially arises on a single ancestral haplotype background and remains associated with nearby alleles until recombination breaks the association; (ii) population admixture (mixing of two populations with different allele frequencies produces LD genome-wide, even between unlinked loci, analogous to the Wahlund effect of Chapter 3 — of particular relevance to admixed African populations); (iii) selection favouring particular allele combinations; and (iv) genetic drift, especially in small or recently founded populations.

LD **decays** across generations due to recombination (Chapter 8): each generation, a fraction θ (the recombination fraction between the two loci) of gametes is recombinant, breaking down the non-random haplotype association. It can be shown that:

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

where D_t is the disequilibrium after t generations of random mating, starting from initial disequilibrium D_0 . LD therefore decays geometrically with generation number, at a rate governed by the recombination fraction — tightly linked loci (small θ) retain LD over many more generations than loosely linked loci (large θ , up to $\theta=0.5$ for unlinked loci, for which LD decays fastest, in a single generation under strict random mating, absent other forces sustaining it such as admixture).

Theorem 15.1 (Geometric Decay of Linkage Disequilibrium)

Under random mating with recombination fraction θ between two loci, and no forces (drift, selection, admixture) regenerating disequilibrium, the disequilibrium coefficient after t generations is $D_t = (1 - \theta)^t D_0$.

Proof. Consider the four haplotype frequencies $h_{AB}, h_{Ab}, h_{aB}, h_{ab}$ in generation t , with $D_t = h_{AB,t}h_{ab,t} - h_{Ab,t}h_{aB,t}$. In forming generation $t + 1$'s gametes, a fraction $(1 - \theta)$ of gametes are non-recombinant (inheriting the parental haplotype intact) and a fraction θ are recombinant (formed by combining the allele at locus 1 from one parental haplotype with the allele at locus 2 from the other). The frequency of haplotype AB in generation $t + 1$ is therefore

$$h_{AB,t+1} = (1 - \theta)h_{AB,t} + \theta p_A p_B,$$

since a recombinant AB gamete requires drawing the A allele (frequency p_A) and the B allele (frequency p_B) independently from the population's separate allele pools (recombination re-randomises the combination). Substituting into $D_{t+1} = h_{AB,t+1} - p_A p_B$ (the equivalent single-term definition of D , Section 15.2):

$$\begin{aligned} D_{t+1} &= (1 - \theta)h_{AB,t} + \theta p_A p_B - p_A p_B \\ &= (1 - \theta)h_{AB,t} - (1 - \theta)p_A p_B \\ &= (1 - \theta)(h_{AB,t} - p_A p_B) = (1 - \theta)D_t, \end{aligned}$$

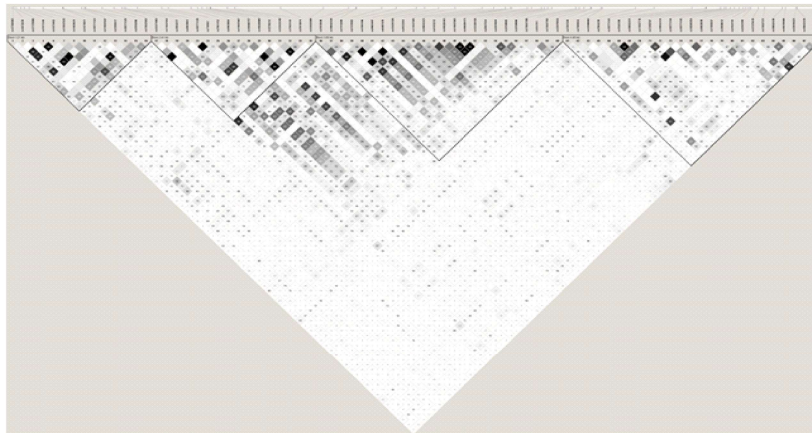
using the fact that allele frequencies p_A, p_B themselves do not change under recombination alone (Chapter 3, recombination reshuffles haplotypes but not allele frequencies). This is a first-order linear recurrence with solution $D_t = (1 - \theta)^t D_0$ by direct iteration (or induction: the base case $t = 0$ is trivial, and the inductive step is exactly the recurrence just derived).

■

Worked Example 15.2 (LD decay between a malaria-resistance marker and a linked SNP, Niger Delta population)

A population geneticist studies two SNPs 50 kb apart, in strong initial LD ($D_0 = 0.08$) in a founding population, flanking a locus associated with malaria resistance in a Niger Delta community. If the recombination fraction between the two SNPs is $\theta = 0.02$ per generation (typical for markers at this physical distance, Section 15.6), Theorem 15.1 predicts the disequilibrium after 25 generations (roughly 500–625 years, at 20–25 years per generation) to be $D_{25} = (1 - 0.02)^{25} \times 0.08 = (0.98)^{25} \times 0.08 \approx 0.603 \times 0.08 \approx 0.0483$ — LD has decayed to about 60% of its initial value, illustrating why closely linked markers can retain population-level LD, and hence remain useful for association mapping (Chapter 11), over many centuries, whereas the same calculation with a more loosely linked marker (say $\theta = 0.20$) would give $(0.80)^{25} \approx 0.0038$, essentially complete decay — the quantitative basis for the tag-SNP selection strategy of Chapter 12.

15.4: Normalised Measures of LD: D^2 and r^2



A triangular LD heatmap: pairwise r^2 between nearby SNPs, the standard visual summary of a region's LD structure.

Because D depends on the allele frequencies at the two loci (and so cannot be directly compared across marker pairs with different allele frequencies), two standardised measures are in universal use:

D2 (Lewontin's D_2):

$$D' = D/D_{\max}, D_{\max} = \min(p_A p_b, p_a p_B) \text{ if } D > 0; \min(p_A p_B, p_a p_b) \text{ if } D < 0$$

D' ranges from -1 to 1 (often reported as $|D'|$, from 0 to 1); $|D'| = 1$ (“complete LD”) means no evidence of historical recombination between the loci within the sample’s ancestry — a useful qualitative summary but sensitive to small sample sizes, especially for rare alleles.

r^2 (the squared correlation coefficient):

$$r^2 = D^2 / (p_A p_a p_B p_b)$$

r^2 ranges from 0 to 1 and has a direct statistical interpretation: it is the squared Pearson correlation between the two loci’s allele indicator variables, and is the more relevant measure for association-study design, because the sample size needed to detect an association at a marker in LD with an (untyped) causal variant, relative to typing the causal variant directly, scales approximately as $1/r^2$.

15.5 Worked Example 15.1

Two SNPs are genotyped in a Nigerian population sample. Estimated haplotype frequencies: $h_{AB} = 0.30$, $h_{Ab} = 0.10$, $h_{aB} = 0.15$, $h_{ab} = 0.45$ (these sum to 1.00). Allele frequencies: $p_A = h_{AB} + h_{Ab} = 0.40$; $p_a = 0.60$; $p_B = h_{AB} + h_{aB} = 0.45$; $p_b = 0.55$.

(a) Compute D.

$$D = h_{AB} - p_A p_B = 0.30 - (0.40)(0.45) = 0.30 - 0.18 = 0.12$$

(b) Compute D'. Since $D > 0$, $D_{\max} = \min(p_A p_b, p_a p_B) = \min[(0.40)(0.55), (0.60)(0.45)] = \min(0.22, 0.27) = 0.22$.

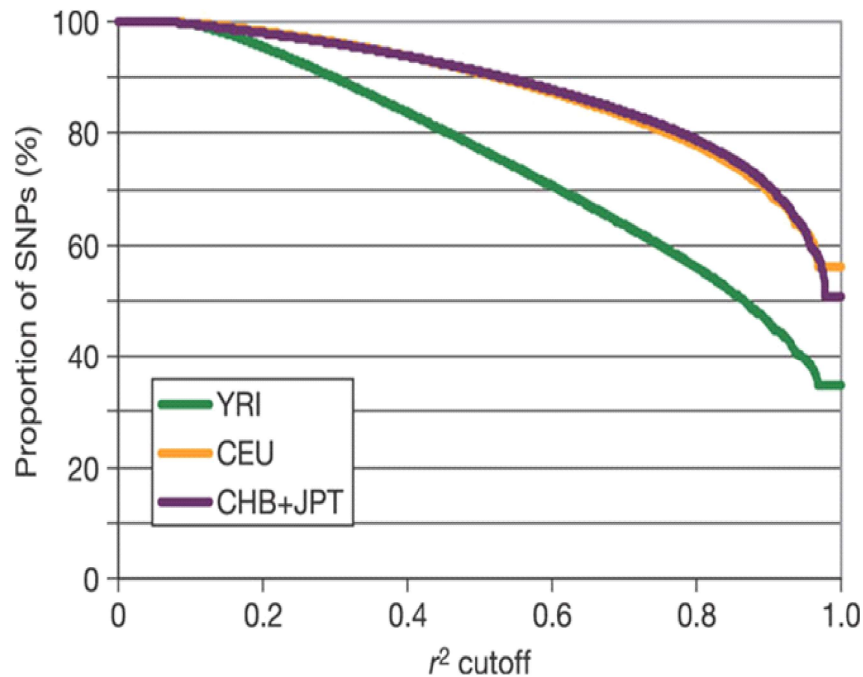
$$D' = 0.12/0.22 = 0.545$$

(c) Compute r^2 .

$$r^2 = (0.12)^2 / [(0.40)(0.60)(0.45)(0.55)] = 0.0144/0.0594 = 0.242$$

(d) Interpret. $D' = 0.545$ suggests moderate LD (some, but not complete, historical recombination between the loci); $r^2 = 0.242$ indicates that a study wishing to detect an association through this marker (locus 1, say), if the true causal variant is locus 2, would need roughly $1/0.242 \approx 4.1$ times the sample size that would be required if the causal variant itself were typed directly — an important practical consideration in designing Nigerian genetic studies that rely on commercially available SNP genotyping arrays, which may capture African-population LD patterns less completely than they capture LD patterns in the (largely European-ancestry) populations used to design many arrays.

15.6: LD and Physical Distance; Relevance to African Populations



LD decays faster with physical distance in the Yoruba (YRI) than in CEU or CHB+JPT samples — the empirical basis for denser marker panels in African association studies.

Because LD decays through recombination accumulated over many generations, and because recombination probability increases (roughly) with physical distance along a chromosome, LD generally decays with increasing physical distance between markers — but the *rate* of this decay differs substantially between populations, depending on their demographic history (effective population size, time since founding or admixture, and number of generations of random mating). African populations, including Nigerian populations, are known from genomic studies to have,

on average, the highest effective population sizes and the most ancient population history among major world populations, and consequently exhibit **more rapid LD decay** (LD extends over shorter physical distances) than populations of more recent common ancestry (e.g., European or East Asian populations descended from smaller founding groups). This has a direct practical consequence for Chapter 18 (GWAS): studies in Nigerian and other African populations generally require denser marker panels (more SNPs per unit physical distance) to achieve the same genomic coverage via LD as studies in populations with more extensive LD — a key motivation for African-specific genotyping array and reference panel development (e.g., the H3Africa initiative).

15.7: Haplotype Inference

Because standard genotyping technologies report unphased genotypes (i.e., for a heterozygote at two or more loci, it is not directly observed which alleles were inherited together on the same chromosome — the **phase** is ambiguous), **haplotype inference (phasing)** methods are needed to statistically reconstruct the most likely haplotypes underlying observed multi-locus genotype data.

- **Clark’s algorithm** is a parsimony-based method: it first resolves all unambiguous individuals (homozygous at all loci, or heterozygous at only one locus, whose haplotypes are directly determined), then iteratively resolves remaining ambiguous individuals by assuming, wherever possible, that their haplotypes are ones already observed/inferred elsewhere in the sample (minimising the total number of distinct haplotypes invoked) — a computationally simple but somewhat ad hoc, order-dependent approach.
- **The EM (Expectation–Maximization) algorithm** treats haplotype phase as missing data and iterates between an **E-step** (computing, for each ambiguous individual, the

expected/posterior probability of each compatible haplotype pair, given current estimated haplotype frequencies) and an **M-step** (re-estimating haplotype frequencies from these expected counts, via maximum likelihood, Chapter 10), converging to the maximum likelihood estimates of population haplotype frequencies.

- **Bayesian methods (e.g., the PHASE algorithm)** extend the EM approach by incorporating a coalescent-based prior (Chapter 16) over haplotypes, reflecting the biological expectation that haplotypes cluster into a genealogical tree structure, generally improving phasing accuracy, especially for rarer haplotypes, relative to EM alone — at greater computational cost. Software implementing PHASE-type algorithms remains widely used for haplotype inference in African population-genetic and disease-association studies, where reference haplotype panels are historically less complete than for European-ancestry populations.

Theorem 15.2 (Bounds on the Linkage Disequilibrium Coefficient r^2)

For two biallelic loci with allele frequencies p_A , p_B for the reference alleles, the squared correlation coefficient $r^2 = D^2 / [p_A(1-p_A) p_B(1-p_B)]$ satisfies $0 \leq r^2 \leq 1$, with $r^2 = 1$ if and only if the two loci are in complete association (only two of the four possible haplotypes occur, in exact proportion to the allele frequencies).

Proof. Since $D^2 \geq 0$ always, the lower bound $r^2 \geq 0$ is immediate. For the upper bound, note that r^2 is, by construction (Section 15.4), exactly the squared Pearson correlation coefficient between two binary indicator random variables — whether a randomly sampled haplotype carries the reference allele at locus A, and whether it carries the reference allele at locus B. The Cauchy-Schwarz inequality guarantees that any Pearson correlation coefficient lies in $[-1, 1]$,

reference allele at locus B. The Cauchy-Schwarz inequality guarantees that any Pearson correlation coefficient lies in $[-1,$

15.8: Chapter Summary

- Linkage disequilibrium is the non-random association between alleles at (usually linked) loci, measured by D , and standardised as D^2 or r^2 .
- LD decays geometrically with recombination fraction and generation number; African populations, including Nigerian populations, generally show faster LD decay due to larger, older effective population size.
- r^2 directly determines the sample-size penalty for indirect (marker-based) association mapping relative to direct typing of the causal variant.
- Haplotype phase, unobserved in standard genotype data, is inferred statistically via parsimony (Clark's algorithm), likelihood (EM algorithm), or Bayesian coalescent-based (PHASE) methods.

15.9: Computational Corner: Computing D , D' and r^2

R.

```
# Linkage disequilibrium: D, D', and r^2 from haplotype
# frequencies (Section 15.4)
# Two SNPs; haplotype counts among 200 chromosomes
hap_counts <- c(AB = 90, Ab = 30, aB = 20, ab = 60)
n <- sum(hap_counts)
pA <- (hap_counts["AB"] + hap_counts["Ab"]) / n
pB <- (hap_counts["AB"] + hap_counts["aB"]) / n
pAB <- hap_counts["AB"] / n

D <- pAB - pA * pB
Dmax <- if (D >= 0) min(pA*(1-pB), (1-pA)*pB)
else min(pA*pB, (1-pA)*(1-pB))
```

```

Dprime <- D / Dmax
r2 <- D^2 / (pA * (1 - pA) * pB * (1 - pB))
cat("D =", D, " D' =", Dprime, " r^2 =", r2, "\n")

# install.packages("genetics")
library(genetics)
gl <- genotype(c("A/A", "A/B", "B/B")) # illustrative use of
genetics package for LD()

```

Python.

```

hap_counts = {"AB": 90, "Ab": 30, "aB": 20, "ab": 60}
n = sum(hap_counts.values())
pA = (hap_counts["AB"] + hap_counts["Ab"]) / n
pB = (hap_counts["AB"] + hap_counts["aB"]) / n
pAB = hap_counts["AB"] / n

D = pAB - pA * pB
Dmax = min(pA*(1-pB), (1-pA)*pB) if D >= 0 else min(pA*pB,
(1-pA)*(1-pB))
Dprime = D / Dmax
r2 = D**2 / (pA * (1 - pA) * pB * (1 - pB))
print(D, Dprime, r2)

```

Both languages return the same D, D² and r² from Worked Example 15.1's haplotype counts, illustrating that computing LD statistics is a small, self-contained calculation that scales, in real genome-wide analyses (Chapter 18), to millions of SNP pairs computed via dedicated tools such as PLINK rather than by hand.

Mini-Project

Mini-Project 15: Computing LD Across Multiple Marker Pairs

Extend the two-SNP LD calculation of Section 15.9 to a small multi-marker haplotype block and visualise the resulting LD structure.

- Step 1.** Simulate haplotype data for 6 SNPs on the same chromosome with realistic decaying LD by physical distance (Section 15.3).
- Step 2.** Compute pairwise D' and r^2 for all 15 SNP pairs.
- Step 3.** Construct and plot an LD heatmap (triangular matrix) of the r^2 values.
- Step 4.** Identify the pair of SNPs with the highest r^2 and interpret what this implies for tagging one SNP with the other.
- Step 5.** Write a short paragraph on how this LD structure would affect the number of independent tests in a candidate-gene study of this region (linking back to Chapter 12).

Deliverable: *Simulation code, an LD heatmap, and a short written interpretation connecting LD structure to multiple-testing burden.*

15.10: Practice Questions

- Q1.** Two loci have haplotype frequencies $h_{AB} = 0.45$, $h_{Ab} = 0.05$, $h_{aB} = 0.05$, $h_{ab} = 0.45$. Compute D , D' , and r^2 , and comment on the strength of LD.
- Q2.** Explain why LD decays faster (over the same physical distance) in a Nigerian population sample than in a European population sample, in terms of population history.
- Q3.** A candidate causal SNP cannot be directly genotyped, but a nearby marker SNP with $r^2 = 0.5$ to it can be. If detecting

the causal SNP's effect directly would require 1,000 samples, approximately how many samples are needed using the marker SNP instead?

- Q4.** Briefly describe, in your own words, the difference in approach between Clark's algorithm and the EM algorithm for haplotype inference.
- Q5.** Explain why a marker locus that is only weakly correlated (small r^2) with an unobserved disease susceptibility locus provides a poor proxy for association testing, even if the marker is physically very close to the causal variant.
- Q6.** Using the R or Python code of Section 15.9, compute D , D^2 , and r^2 for a different haplotype dataset: $AB = 150$, $Ab = 10$, $aB = 10$, $ab = 150$ (out of 320 total haplotypes). Comment on what the near-maximal D^2 but the actual r^2 value together tell you about this marker pair.