Module 10 - LE vs. LD and Linkage vs. Association

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Outline

Part of Chapter 5 - Linkage Equilibrium vs. Linkage Disequilibrium and Linkage vs. Association

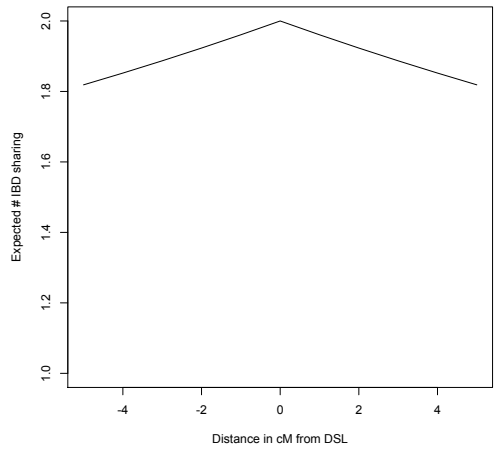
- The need for fine-mapping beyond linkage analysis
- Linkage equilibrium vs. Linkage Disequilibrium (LD)
- Different measures of LD: D, D' and r^2
- When does LD occur?
- The importance of LD in fine-mapping.
- Binary trait: would allele frequency differ between cases and controls at a marker merely linked to DSL but not in LD?
- Quantitative trait: would the mean value of the phenotype differ between the three genotypes at a marker merely linked to DSL but not in LD?
- Linkage vs. Association Summary: intra-family dependency vs. population-level dependency.

Linkage: Long-range Dependency I

- In linkage analysis, we typically place $\approx$ 500-1000 markers, $\approx$ 10cM-5cM apart, across the genome.

$$\theta_{\text{marker, DSL}} = \frac{1 - e^{-2t}}{2} \leq \frac{1 - e^{-2 \cdot \frac{0.1}{2}}}{2} = 0.048.$$

- So, if we have enough data, we will be able to find the linked marker(s), i.e. reject the null $H_0 : \theta = 0.5$ for the linked marker(s) using either parametric linkage analysis or allele-sharing method.



Linkage: Long-range Dependency II

- However, the linked marker(s) could be 5cM away from DSL. And 5cM $\approx$ 5 million bp, so we still need additional methods to determine a more precise location for the DLS.

- What if we add more markers to the genome to conduct linkage analysis?
- If we use linkage as a measure, a marker that is 5CM away is not statistically different from a marker that is 2 cM away from DSL.

- Map resolution by linkage: 5 cM $\approx$ 5 million bp.
  - A disease/trait locus D : shared by two affected sibs.
  - A marker locus M : 5 cM (r.f. $\theta \approx .05$) away from D.
  - Chance the marker also shared: $(1-\theta)^2 = .9$
  - Great news in terms of finding the approximate location of D through M, but not so great for fine mapping (the exact or more refined location of D): 5cM is roughly 5 million DNA bps to sift-through!

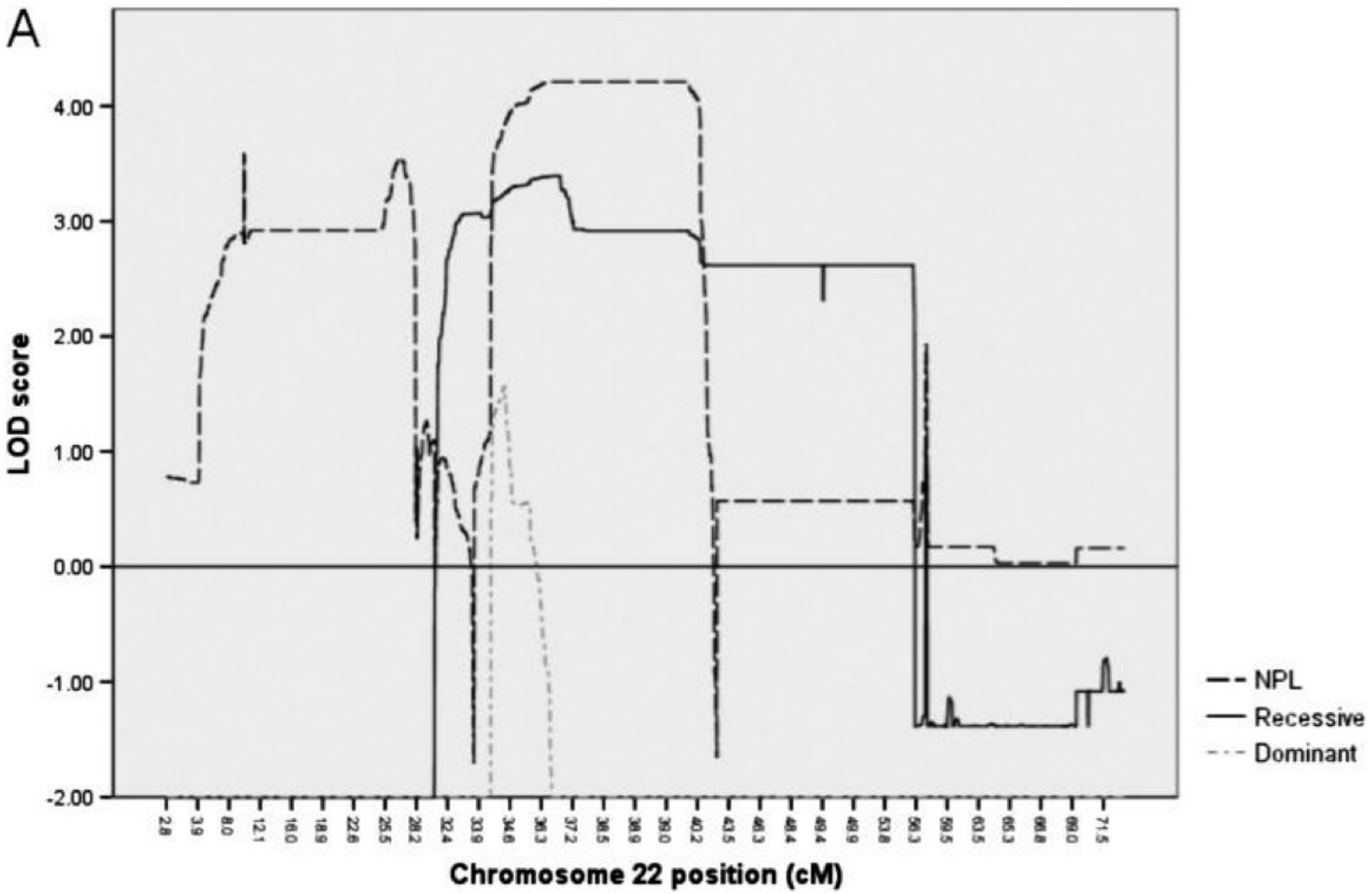
The Need for Fine-mapping: an Example I

- Linkage is not designed to distinguish between linked markers!

- E.g. Genome-wide Linkage Analysis of a Parkinsonian-Pyramidal Syndrome Pedigree by 500 K SNP Arrays
  - *On chromosome 22, a broad region extending from 4.17 to 28.18 cM was associated with an average LOD score of 2.90,*
  - *And another region extending* **from 34.40 to 41.93 cM (28,934,667 bp4,951,655 bp)** *was associated with an average LOD score of 4.08 (average $p = 0.00003$) (Figure 2A).*
  - *Within the latter region, the maximum predicted LOD score of 4.21 was observed in region extending from 36.58 to 39.98 cM.*

The Need for Fine-mapping: an Example II

Figure 2



The Need for Fine-mapping: an Example III

- Caution with using dense set of markers to perform linkage analysis.
- Ignoring Linkage Disequilibrium among Tightly Linked Markers Induces False-Positive Evidence of Linkage for Affected Sib Pair Analysis
  - *Most multipoint linkage programs assume linkage equilibrium among the markers being studied. The assumption is appropriate for the study of sparsely spaced markers with intermarker distances exceeding a few centimorgans*
  - *However, with recent advancements in high-throughput genotyping technology, much denser markers are available, and linkage disequilibrium (LD) may exist among the markers. Applying linkage analyses that assume linkage equilibrium to dense markers may lead to bias.*

Linkage Disequilibrium (LD) I

- **Linkage Equilibrium**: alleles at different loci in a haplotype are **independent**, **at the population level**.
- That is, a haplotype frequency is the product of the underlying allele frequencies:

$$freq(AB) = freq(A)freq(B), \;\; p_{AB} = p_A p_B$$

- **Linkage Disequilibrium (LD)** refers to the **lack of independence** (or allelic association) of alleles at loci on a haplotype randomly sampled from a **population**.

- Note that HWE refers the independent pairing of two alleles at one locus to form a genotype, while LE refers to the independent pairing of two (or more) alleles at two (or more) loci to form a haplotype.

LD measures - D I

LE: $freq(AB) = freq(A)freq(B),\ \ p_{AB} = p_A p_B$

	B	b	Total
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
Total	n_B	n_b	n

	B	b	Total
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	p_a
Total	p_B	p_b	1

- There are different (pair-wise) LD measures (D, D' and r^2).
- **The basic measure** D,

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

- What about using $p_{Ab} - p_A p_b$?

$$p_{Ab} - p_A p_b = p_A - p_{AB} - p_A(1 - p_B) = -(p_{AB} - p_A p_B) = -D.$$

- Similarly,

$$p_{aB} - p_a p_B = -D,\ \ p_{ab} - p_a p_b = D.$$

LD measures - D II

- We can also show that

$$D = p_{AB}p_{ab} - p_{Ab}p_{aB}.$$

$$p_{AB}p_{ab} = (p_Ap_B + D)(p_ap_b + D) = p_Ap_Bp_ap_b + p_Ap_BD + p_ap_bD + D^2$$

$$p_{Ab}p_{aB} = (p_Ap_b - D)(p_ap_B - D) = p_Ap_Bp_ap_b - p_Ap_bD - p_ap_BD + D^2$$

$$p_{AB}p_{ab} - p_{Ab}p_{aB} = D(p_Ap_B + p_ap_b + p_Ap_b + p_ap_B)$$
$$= D(p_A(p_B + p_b) + p_a(p_B + p_b)) = D(p_A + p_a) = D.$$

LD measures - D III

- What is the **range** of D?

$$p_{AB} = p_A p_B + D \geq 0 \Longrightarrow D > -p_A p_B,$$

$$p_{ab} = p_a p_b + D \geq 0 \Longrightarrow D > -p_a p_b,$$

$$p_{Ab} = p_A p_b - D \geq 0 \Longrightarrow D < p_A p_b,$$

$$p_{aB} = p_a p_B - D \geq 0 \Longrightarrow D < p_a p_B.$$

$$D_{max} = min(p_A p_b, p_a p_B),$$

$$D_{min} = max(-p_A p_B, -p_a p_b) = -min(p_A p_B, p_a p_b),$$

$$D_{min} \leq D \leq D_{max}$$

LD measures - D' I

- **The value D depends on the allele frequencies**, so it is not useful to do comparisons between pairs of loci.
- E.g. case 1 $p_A = 0.3, p_B = 0.3$.

$$D_{max} = min(p_A p_b, p_a p_B) = min(0.3 \times 0.7, 0.7 \times 0.3) = 0.21$$

$$D_{min} = -min(p_A p_B, p_a p_b) = -min(0.3 \times 0.3, 0.7 \times 0.7) = -0.09$$

- E.g. case 2 $p_A = 0.3, p_B = 0.1$.

$$D_{max} = min(p_A p_b, p_a p_B) = min(0.3 \times 0.9, 0.7 \times 0.1) = 0.07$$

$$D_{min} = -min(p_A p_B, p_a p_b) = -min(0.3 \times 0.1, 0.7 \times 0.9) = -0.03$$

- **Normalize** D so that $0 \leq D' \leq 1$:

$$D' = \frac{D}{D_{max}}, \text{ if } D > 0,$$

$$D' = \frac{D}{D_{min}}, \text{ if } D < 0.$$

LD measures - D' II

- $D' = 1$ **is called complete LD.**
- e.g. E.g. case 2 $p_A = 0.3, p_B = 0.1$ and $p_{AB} = 0.1$.

$$D = p_{AB} - p_A p_B = 0.1 - 0.3 \times 0.1 = 0.07.$$

$$D' = D/D_{max} = 0.07/0.07 = 1.$$

- $D' = 1$ **implies at least one of the cell counts is zero, i.e. at least one of the 4 haplotypes was not observed.**
- e.g. E.g. case 2 $p_A = 0.3, p_B = 0.1$ and $p_{AB} = 0.1$.

$$p_{aB} = p_B - p_{AB} = 0.1 - 0.1 = 0.$$

LD measures - D' III

- **If allele frequencies are similar, high D' implies the markers are good surrogates for each other.**
- E.g. case 3 $p_A = 0.1, p_B = 0.1$ and $p_{AB} = 0.1$. In this case, $D' = 1$ and A and B are surrogate to each other!
- E.g. case 2 $p_A = 0.3, p_B = 0.1$ and $p_{AB} = 0.1$ where $D' = 1$. Although allele B always goes with allele A in the same haplotype (i.e. knowing B implies A), because allele frequencies are not similar, Allele A could also go with allele b (i..e. knowing A does NOT imply B).

LD measures - r^2 I

- Another normalized LD measure: $0 \leq r^2 \leq 1$.

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} = \frac{D^2}{p_A p_a p_B p_b}.$$

- As a measure of loss in efficiency when one marker is replaced with another marker in an association study
  Linkage Disequilibrium in Humans: Models and Data.
- $N_{marker} = \frac{1}{r^2} N_{DSL}$: *In other words, to achieve (approximately) the same power at the marker locus as is achieved at the susceptibility locus, the sample size must be increased by a factor of* $\frac{1}{r^2}$.

- $r^2 = 1$ implies the two markers are statistically identical: if we know the allele at one locus, then we can predict perfectly the allele at the other locus ($r^2 = 1$ only if a pair of diagonal cells equal to zero, i.e. $p_{AB} = p_{ab} = 0$ or $p_{Ab} = p_{aB} = 0$.

LD measures - r^2 II

- Connection with Pearson's correlation coefficient ρ.
  - Let $X = 0$ or 1 for alleles a and A, $P(X = 1) = p_A$, allele frequency of A.
  - Let $Y = 0$ or 1 for alleles b and B, $P(Y = 1) = p_B$, allele frequency of B.
  - Let $P(XY = 1) = p_{AB}$, the frequency of haplotype AB (i.e. when both X and Y are $= 1$).
  - It is easy to show that

$$E(X) = p_A, \quad Var(X) = E(X^2) - (E(X))^2 = p_A - p_A^2 = p_A p_a.$$

$$E(Y) = p_B, \quad Var(Y) = p_B p_b.$$

$$E(XY) = p_{AB}.$$

  - Therefore

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}} = \frac{p_{AB} - p_A p_B}{\sqrt{p_A p_a p_B p_b}} = r.$$

LD measures - D' vs. r^2 I

- For simplicity, we assume $D = p_{AB} - p_A p_B > 0$,

$$D' = \frac{D}{min(p_A p_b, p_a p_B)}, \quad r = \frac{D}{\sqrt{p_A p_a p_B p_b}}$$

p_A	p_B	p_{AB}	D	D'	r
0.5	0.5	0.3	0.05	0.20	0.20
0.5	0.5	0.4	0.15	0.60	0.60
0.5	0.5	0.5	0.25	1	1
0.3	0.3	0.2	0.11	0.52	0.52
0.3	0.3	0.3	0.21	1	1
0.5	0.3	0.3	0.15	1	0.65
0.4	0.3	0.3	0.18	1	0.80
0.3	0.3	0.3	0.21	1	1
0.3	0.05	0.05	0.035	1	0.35
0.3	0.01	0.01	0.007	1	0.15
0.1	0.05	0.05	0.045	1	0.69

LD measures - D' vs. r^2 II

- If $p_A = p_B$, then

$$min(p_A p_b, p_a p_B) = p_A p_a = \sqrt{p_A p_a p_B p_b} \Longrightarrow D' = r.$$

- If $p_{AB} = p_B$, and without loss of generality, we also assume that $p_A > p_B$, then

$$D' = \frac{p_{AB} - p_A p_B}{min(p_A p_b, p_a p_B)} = \frac{p_B(1 - p_A)}{p_a p_B} \equiv 1.$$

$$r = \frac{p_{AB} - p_A p_B}{\sqrt{p_A p_a p_B p_b}} = \frac{p_B(1 - p_A)}{\sqrt{p_A p_a p_B p_b}} = \sqrt{\frac{p_a}{p_A}}\sqrt{\frac{p_B}{p_b}}.$$

  - If $p_{AB} = p_B$ AND $p_A = p_B$, then

$$r = \sqrt{\frac{p_a}{p_A}}\sqrt{\frac{p_B}{p_b}} \equiv 1.$$

  - However, this is not true if $p_A \neq p_B$.

LD measures - D' vs. r^2 III

- In particular, if A is a common allele ($p_A \approx p_a$), and B is a rare allele ($p_B = \epsilon$)

$$r \approx \sqrt{p_B}, \quad r^2 \approx p_B = \epsilon$$

- So, the popular r^2 measure encounters challenges in the context of rare variants analysis!

- Which measure is better? A comparison of linkage disequilibrium measures for fine-scale mapping
- D' and r^2 have different biological interpretations: D' measures only recombinational history, r^2 summarizes both recombinational and mutational history.
- The amount of LD across the genome is not uniform and depends on genealogy of a population sample, mutation, selection, admixture between populations, etc.
- How to estimate LD when phase is unknown?
- What happen if there are more than two alleles at a locus: LD measures of multi-allelic markers. Hedrick, 1987.

Example of Different LD Measures

Textbook Table *5.4 shows a comparison of the measures of LD on a fictitious sample of 100 chromosomes.*

Table 5.4 Measures of linkage disequilibrium

		B Locus		
	A Locus	B	b	Row Total
	A	43	27	70
	a	2	28	30
Column Total		45	55	100

$$D = (43 - 70 * 45/100)/100 = 0.115$$

$$D_{\max} = \min(70 * 55, 30 * 45)/10,000 = 0.135$$

$$D' = 0.115/0.135 = 0.8581$$

$$r = 0.115/\sqrt{0.7 * 0.3 * 0.55 * 0.45} = .5044$$

$$\hat{p}_A = 70/100 = 0.7, \quad \hat{p}_B = 45/100 = 0.45, \quad \hat{p}_{AB} = 43/100 = 0.43$$

Example of LD I

- CFTR gene (exon 10) for Cystical Fibrosis: mutation ΔF508 (A) with population frequency $p_A = 2\%$.
- A nearby marker (intron 8): allele 9T (B) with population frequency $p_B = 11\%$.
- ΔF508 and 9T in strong linkage disequilibrium: all chromosomes with ΔF508 have 9T allele.
- Counts (frequencies) for a total of 204 Caucasian CF patients (Kiesewetter et al. 1993, courtesy of Andrew Paterson).

	9T (B)	not 9T (b)	Total
ΔF508 (A)	143 (70%)	0 (0%)	143 (70%)
not ΔF508 (a)	14 (7%)	47 (23%)	61 (30%)
Total	157 (77%)	47 (23%)	204

- Using the above data, can we calculate the LD measure as

$$\hat{D} = \hat{p}_{AB} - p_A p_B = 0.7 - 0.02 \cdot 0.11 = 0.7 - 0.077?$$

Example of LD II

- **Wrong**: $p_{AB} = 0.7$ is the estimated haplotype frequency of (A,B)=(ΔF508, 9T) in the CF population, while $p_A = 2\%$ and $p_B = 11\%$ are allele frequency in the generation population.

- If we consider the population of CF patients only

$$\hat{D} = \hat{p}_{AB} - \hat{p}_A \hat{p}_B = 0.7 - 0.7 \cdot 0.77 = 0.7 - 0.54.$$

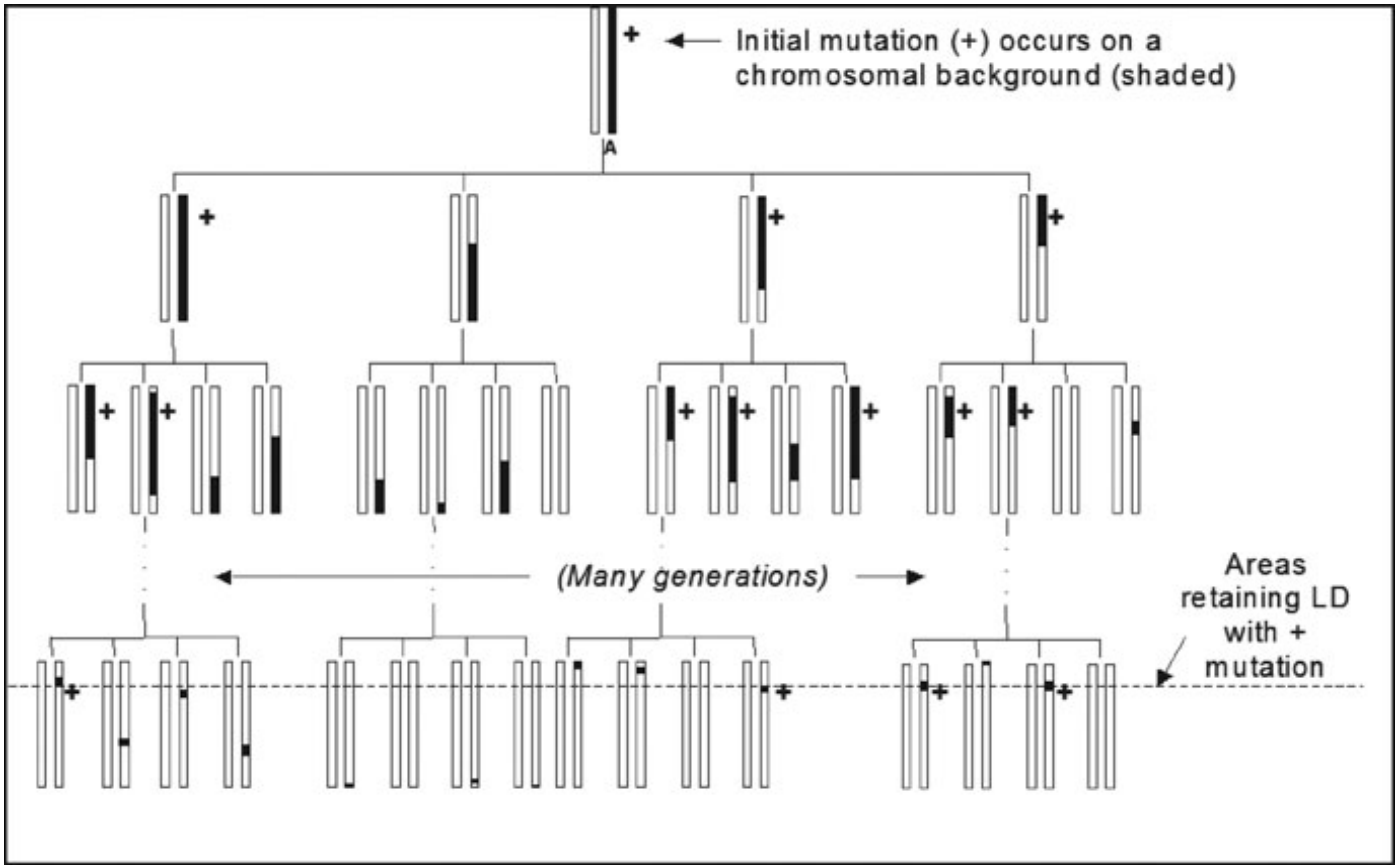
When Does LD Occur? Linked Loci $\neq$ LD I

- Many generations ago, a disease mutation $(+)$ occurred at locus D in a haplotype $\mathcal{H}$.
- $\mathcal{H}$ contains a unique set of alleles at linked marker loci $\{M_i\}$.
- Each M_i is θ_i away from D.
- During meiosis, recombination occur between M_i and D according to θ_i.
- Chance of the specific allele at the M_i locus and $+$ being co-inherited decreases with increasing θ_i and **number of generations**.
- After **1 generation**: $(1 - \theta_i)$.
- But it **accumulates through** n **generations**: $(1 - \theta_i)^n$.
- In current data, alleles at M_i may be **independent** of alleles at D, **at the population level.**
- Only a very small region around the disease locus is likely to retain the characteristic haplotype, $\mathcal{H}$, and remain in LD with the disease locus.

- **LD implies (tight) linkage, but linkage does not imply LD!**

When Does LD Occur? Linked Loci $\neq$ LD II

- A graphical illustration, Figure 5.1 of the Textbook



The Importance of LD in Gene Mapping Study I

- Consider a binary trait and a case-control sampling design.
- At the (unknown and to be discovered) DSL locus, the genotype frequency must differ between cases and controls, e.g. $p_{DD,case} > p_{DD,control}$.

- How about at a linked marker A?

$$p_{AA,case} = P(AA|case) = \frac{P(AA, case)}{P(case)}$$

$$= \frac{P(AA, dd, case) + P(AA, dD, case) + P(AA, DD, case)}{P(case, dd) + P(case, dD) + P(case, DD)}$$

$$= \frac{P(case|dd)P(AA, dd) + P(case|dd)P(AA, dD) + P(case|dd)P(AA, DD)}{P(case|dd)P(dd) + P(case|dD)P(dD) + P(case|DD)P(DD)}$$

$$= \frac{f_0 p_{Ad}^2 + f_1 2 p_{Ad} p_{AD} + f_2 p_{AD}^2}{f_0 p_d^2 + f_1 2 p_d p_D + f_2 p_D^2}$$

- Here we used HWE assumption for genotypes at a single locus as well as at multiple loci. Also $P(case|AA, dd) = P(case|dd)$ and similar for others since the genotype at the DSL locus determines the phenotype.

The Importance of LD in Gene Mapping Study II

- If this maker is in LE, that is, it is NOT in LD, then

$$p_{AA,case} = P(AA|case) = \frac{f_0 p_{Ad}^2 + f_1 2 p_{Ad} p_{AD} + f_2 p_{AD}^2}{f_0 p_d^2 + f_1 2 p_d p_D + f_2 p_D^2}$$

$$= \frac{f_0 p_A^2 p_d^2 + f_1 2 p_A p_d p_A p_D + f_2 p_A^2 p_D^2}{f_0 p_d^2 + f_1 2 p_d p_D + f_2 p_D^2} = p_A^2.$$

- We can perform similar calculations for other genotype groups and for control samples as well, e.g.

$$p_{AA,control} = p_A^2.$$

- So the **genotype frequency (hence allele frequency) for a linked marker but not in LD with the DSL do not differ between cases and controls!**
- **We will see a difference in genotype (and allele) frequency only if the marker is in LD with the DSL locus!** Since LD happens only for extremely tightly linked marker, we achieved our fine mapping goal!

The Importance of LD in Gene Mapping Study III

- Now let's look at a QTL example.

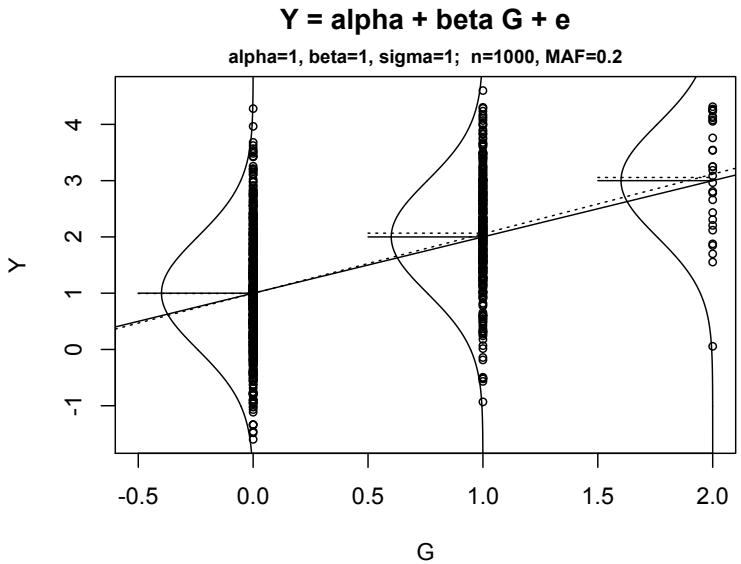
$$Y = \alpha + \beta G + e, \;\; e \sim N(0, \sigma^2).$$

- And this model implies that

$$(Y|G = dd) = (Y|G = 0) \sim N(\alpha, \sigma^2), \;\; E(Y|G = 0) = \mu_0 = \alpha,$$

$$(Y|G = dD) = (Y|G = 1) \sim N(\alpha + \beta, \sigma^2), \;\; E(Y|G = 1) = \mu_1 = \alpha + \beta,$$

$$(Y|G = DD) = (Y|G = 2) \sim N(\alpha + 2\beta, \sigma^2), \;\; E(Y|G = 2) = \mu_2 = \alpha + 2\beta.$$



The Importance of LD in Gene Mapping Study IV

- How about at a linked marker A? What is the mean value of Y among individuals with genotype AA?

$$E(Y|AA)$$

$$= E(Y|AA,dd)P(dd|AA) + E(Y|AA,dD)P(dD|AA) + E(Y|AA,DD)P(DD|AA)$$

$$= E(Y|AA,dd)\frac{P(dd,AA)}{P(AA)} + E(Y|AA,dD)\frac{P(dD,AA)}{P(AA)} + E(Y|AA,DD)\frac{P(DD,AA)}{P(AA)}$$

$$= \mu\frac{p_{Ad}^2}{p_A^2} + (\mu+\beta)\frac{2p_{Ad}p_{AD}}{p_A^2} + (\mu+2\beta)\frac{p_{AD}^2}{p_A^2}$$

- Again, if we assume this maker is in LE, that is, it is NOT in linkage disequilibrium, then

$$= \mu\frac{p_A^2p_d^2}{p_A^2} + (\mu+\beta)\frac{2p_Ap_dp_Ap_D}{p_A^2} + (\mu+2\beta)\frac{p_A^2p_D^2}{p_A^2}$$

$$= \mu p_d^2 + (\mu+\beta)2p_dp_D + (\mu+2\beta)p_D^2$$

The Importance of LD in Gene Mapping Study V

- This is the same as the mean value of Y at the population level:

$$E(Y) = E(Y|dd)p(dd) + E(Y|dD)p(dD) + E(Y|DD)p(DD)$$

$$= \mu p_d^2 + (\mu + \beta)2p_d p_D + (\mu + 2\beta)p_D^2$$

- We can perform similar calculations for other genotype groups and obtain

$$E(Y|AA) = E(Y|Aa) = E(Y|aa) = E(Y).$$

  as long as marker A is in LE with the DSL.

- So again, we will see a difference in phenotype Y mean value across the three genotypes (i.e. association between Y and a G of interest) only for a marker that is in LD with the DSL locus.

Linkage vs. Association I

- **Linkage is intra-familial**. Linkage studies co-segregation of alleles at two loci that are close to each other on the same chromosome from one generation to the next within families.

- **Association is a population property**. a more general term of non-independence. The goal of association analysis is determine whether a trait and a specific allele at some locus are associated at the population level.

- In linkage, different alleles at the same marker can segregate with the disease in different families.

  The particular alleles analyzed at two loci are not of interest in themselves and are used only as a tool for assessing the linkage properties of these loci.

- Association studies utilizing LD focus on a specific allele of a marker (allelic association). This allele tends to be on the same haplotype with the disease mutation allele (more frequently than what is expected under linkage equilibrium).

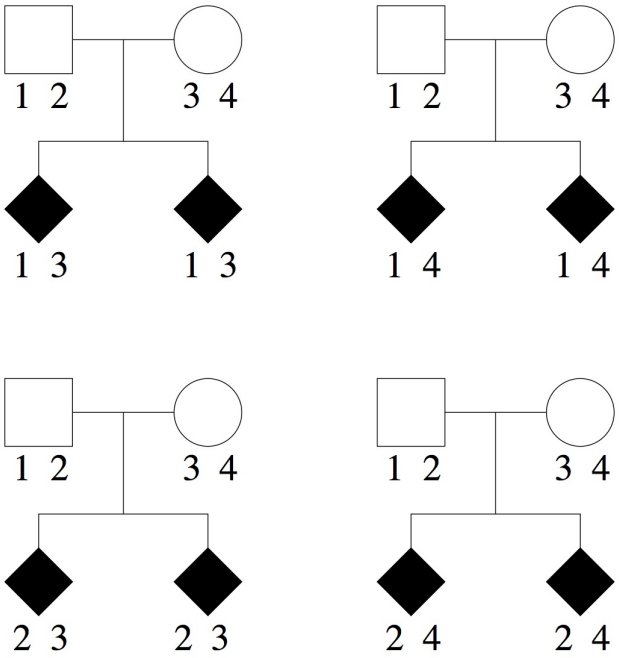
Linkage vs. Association II

The particular alleles are the subject of study, and in some cases (especially in the classical association studies) it is considered possible that they are themselves the cause of the phenotype.

- **Association: non-independence in the population**.
- **Linkage** always leads to a dependence, but for most pairs of loci that dependence is solely **intrafamilial**. There may not be association at the population level.
- **Linkage Equilibrium**: alleles at different loci in a haplotype are **independent**, **at the population level**.

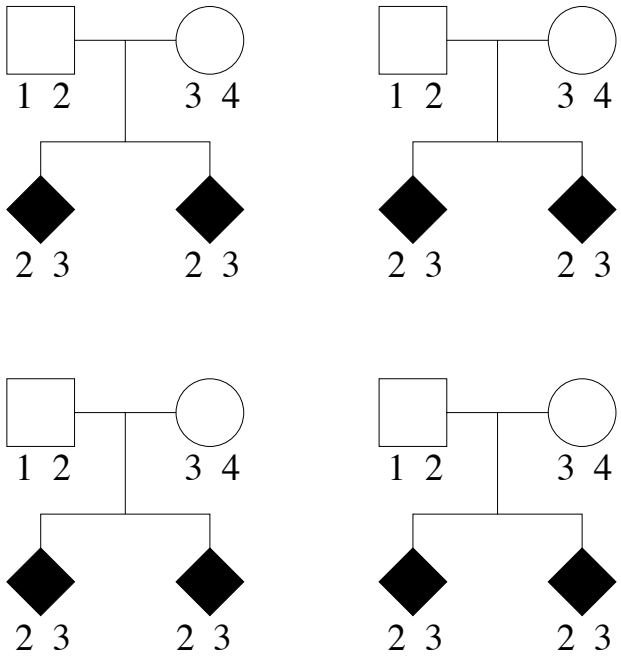
Linkage vs. Association Examples I

Linkage without Association



- All affected sib pairs share 2 IBD.
- But the specific allele shared or transmitted tend to differ between families.

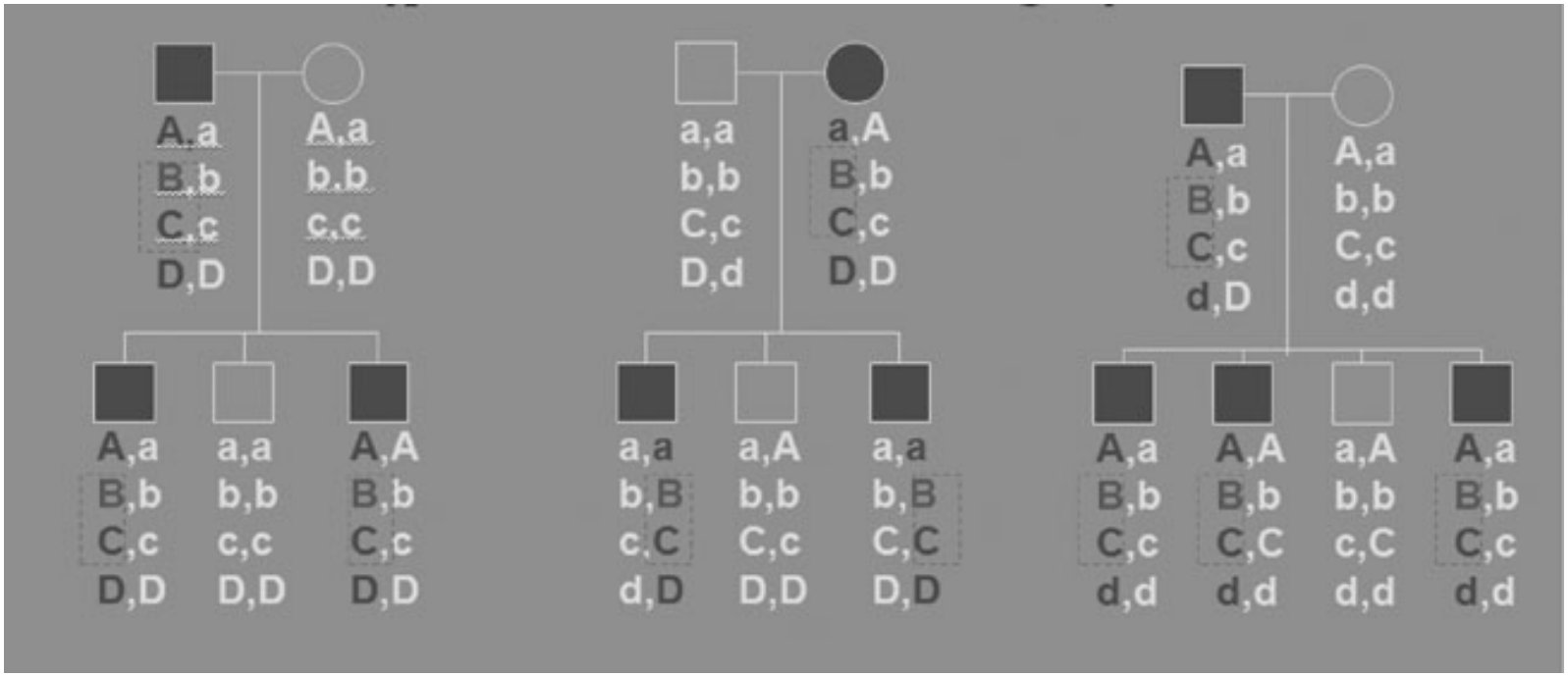
Linkage and Association



- All affected sib pairs share 2 IBD.
- And the specific allele shared or transmitted are consistent between families.

Linkage vs. Association Examples II

Figure 5.3 of the Textbook



A and D loci are linked, but only the B and C alleles are associated.

Linkage vs. Association Examples III

- *It shows three markers with alleles A,a and C,c and D,d, and a DSL with disease allele B transmitted in three different families.*
- *All of the three markers are linked to the disease locus because the same alleles are being transmitted within families at all four locations.*
- Note that in each family, the black haplotype of the affected parent was transmitted without any recombination between the loci to the 7 affected offsprings, so strong evidence for linkage between all loci.
- **However**, *only the C allele is transmitted with the B allele in every family.*
- Note that at the parental generation, some parents have haplotype AB and some have aB, similarly for BD and Bd, but there is only halotype BC to start with!
- Linkage captures the co-segregation of a **given haplotype**, but association additionally captures the characteristics of the haplotype itself (i.e. the **marginal distribution of the haplotype)**
- *Thus the A and D loci are linked, but only the nd lleles will be associated with each other at the population level.*

Linkage vs. Association Examples IV

- How about a 'formal' linkage analysis? Let's consider linkage analysis between A and B (θ_{AB}, NPL_A), and between B and C (θ_{BC}, NPL_C),. Because haplotypes from the unaffected parents not informative due to homozygosity, for simplicity let's focus on the haplotypes from the affected parents. In addition, let's also focus on affected offsprings even though unaffected offsprings also provide linkage (and association) information.
  - Parametric linkage analysis
    - Between A and B loci: $r = 0$ out of $n = 7$ Note that the recombination events are counted for both AB and ab haplotypes, i..e linkage analysis not allelic specific.
    - Between B and C loci: $r = 0$ out of $n = 7$, the same as between A and B. However, in this case we actually only have BC haplotypes, but linkage analysis does not take into account this information.
    - Thus the (simplified) parametric linkage analysis leads to identical results for θ_{AB} and θ_{BC}.

Linkage vs. Association Examples V

- Non-parametric allele sharing linkage analysis
  - At locus A, we have all affected sib-pair (in fact half-sib pair since here we are only counting sharing of the allele inherited from one parent) share 1 allele IBD, substantially different from what's expected under the null. However, note again, the shared allele can be either A or a and linkage analysis does not differential that.
  - At locus C, we also have all affected pairs share 1 allele IBD, but in this case only the B allele, but linkage analysis does not care.
  - Thus the (simplified) non-parametric linkage analysis also leads to identical results for NPL_A and NPL_C.

- Association analysis.
  - At locus A among the 7 alleles of the affected, some are a and some are A.
  - At locus C among the 7 alleles of the affected are all C, but we do observe both c and C in the unaffected.
  - We need to formally compare the allele frequency observed in the cases with that in controls, but the message is clear here: association analysis is able to distinguish between locus A and and locus C, both are linked to the DSL B!

Association Analysis Overview I

- For **a quantitative trait** Y (typically assumed to be approximately normally distributed).
  - We have learned that the mean value of Y differ only at the DSL or markers that are in LD with the DSL:

$$E(Y|G = aa) \; vs. \; E(Y|G = Aa) \; vs. \; E(Y|G = AA).$$

  - Naturally, we can use the simple linear regression models to detect association between Y and G

$$Y = \alpha + \beta G + e, e \sim N(0, \sigma^2).$$

  - Depending on how we code G, we can have 1 d.f. additive (or dominant, recessive) model or 2 d.f. genotypic model. The 1.d.f. additive model is most commonly used model.

$G =$	aa	Aa	AA
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1
Genotypic	"0"	"1"	"2"

Association Analysis Overview II

$$Y = \alpha + \beta G + e, e \sim N(0, \sigma^2).$$

- The test of no association is equivalent to $H_0 : \beta = 0$.
- If we used 2 d.f. model, we can introduce dummy variables $D_1 = I(G = Aa)$ and $D_2 = I(G = AA)$, then $H_0 : \beta_1 = \beta_2 = 0$

$$Y = \alpha + \beta_1 D_1 + \beta_2 D_2 + e, e \sim N(0, \sigma^2).$$

- If we want to take into account any environmental factors and/or additional SNPs, then simply consider **multivariate linear regression models, possibly with interaction terms**.

$$Y = \alpha + \beta G + \gamma E + \delta GxE + e, e \sim N(0, \sigma^2).$$

Association Analysis Overview III

$$Y = \alpha + \beta G + \gamma E + \delta GxE + e, e \sim N(0, \sigma^2).$$

- From the prerequisite course, we have learned
  - Inference concerning the regression parameters: least squares estimation, hypothesis testing.
  - Interpretation of the parameters.
  - Inference from the estimated regression function: prediction.
  - Model checking: does the model fit?
  - Model selection: do we need all predictors? do we need interaction terms?
  - ....

Therefore, we will NOT spend more lecture time on this particular topic.

Association Analysis Overview IV

- For **a binary trait** Y, where $Y = 1$ denotes for being affected/cases, and $Y = 0$ for unaffected/controls.
  - We have learned that genotype (therefore allele) frequency differ between cases and controls only at the DLS or markers that are in LD with the DSL:

$$\{p_{aa,case}, p_{Aa,case}, p_{AA,case}\} \textit{ vs. } \{p_{aa,control}, p_{aa,control}, p_{aa,control}\}.$$

  - How do we detect association between Y and G?
  - Intuitively, we can **compare frequency differences between cases and controls**. We will study this.
  - But, can we **use the regression framework** to unify the analysis for QTL and binary traits, and to handle say environmental factors for binary traits? We will also study this.

Exercises

- Chapter 5 Exercise 3(a)-(e).
- Chapter 5 Exercise 4.
- Chapter 5 Exercise 5.

What's Next

- Chapter 7 - The Basics of (Population-based) Genetic Association Analysis