

Module 9 - Basic Concepts of Allele-Sharing Method (Non-Parametric Linkage Analysis)

(Fundamentals of) Statistical Genetics

Lei Sun

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

Appendix A - Basic Concepts of (Non-Parametric) Linkage Analysis or Allele-Sharing Method

- ➡ From simple Mendelian disease to complex traits
- ➡ Challenges of full-parametric linkage analysis for complex traits
- ➡ Revisit IBD sharing and IBD distribution calculation examples
- ➡ Rationale and principle of the allele-sharing method
- ➡ Expected IBD sharing by a pair of affected sibs
 - ◆ at the DSL
 - ◆ at a marker unlinked with DSL
 - ◆ at a marker linked with DSL
- ➡ IBD transition probability
- ➡ Affected sib-pair (ASP) method
- ➡ Additional considerations: multiple testing for genome-wide scans, use of other relative pairs, incorporating important covariates, analyzing QTL etc.
- ➡ Missing inheritance and IBD sharing information and its implications

GENEHUNTER - see other notes

Simple Mendelian Diseases

- ➡ Trait rare and severe.
- ➡ Single gene locus model, often with full penetrance and no phenocopies.
- ➡ Limited covariates effects (e.g. age, environmental factors).
- ➡ Classical modes of inheritance, such as dominant, recessive.
- ➡ Could find a simple model that adequately explains the disease inheritance pattern observed in a pedigree.
- ➡ Full-parametric linkage analysis is suitable: inference of θ between a known marker locus and the unknown DSL.

$$\Phi = (p, f_0, f_1, f_2, \theta)$$

$$L(\Phi) = P(Y_{\text{phenotype}}, G_{\text{genotype}} | R_{\text{relation/pedigree}}, M_{\text{inheritance model}})$$

$$H_0 : L_{\tilde{\Phi}}, \tilde{\Phi} = (\tilde{p}, \tilde{f}_0, \tilde{f}_1, \tilde{f}_2, 1/2)$$

$$H_1 : L_{\hat{\Phi}}, \hat{\Phi} = (\hat{p}, \hat{f}_0, \hat{f}_1, \hat{f}_2, \hat{\theta})$$

$$LOD = \log_{10} \frac{L_{\hat{\Phi}}}{L_{\tilde{\Phi}}}$$

Complex Traits I

- ➡ Many factors may complicate the determination of a suitable model.
- ➡ **Environmental/Covariate factors** such as diet, smoking status etc.
- ➡ Variable age-at-onset: the chance of being affected given a genotype is age related.
- ➡ **Heterogeneity**
 - ◆ Allelic heterogeneity: different mutations (at-risk alleles) inside one gene (e.g. over 1,000 mutations for the CFTR gene for Cystic Fibrosis).
 - ◆ Locus (genetic) heterogeneity: different genes/loci; the same phenotype may be caused by different genes.
 - ◆ Clinical (phenotypic) heterogeneity: the same genotype but different phenotype.
- ➡ **Interaction**: both $G \times G$ and $G \times E$, and higher-order interaction as well.
- ➡ Variable expression of disease (severity and types of a disease); phenotype definition (affected or not) may not be clear.

Complex Traits II

- ➡ Other transmission mechanism (irregular segregation), e.g. Transmission Ratio Distortion (TRD) and Imprinting (at-risk alleles are preferentially received from the father or the mother.)
- ➡ Examples of complex traits
 - ◆ Common disorders of childhood and adulthood: diabetes, asthma, heart disease, cancers.
 - ◆ Psychiatric disorders and behavioural traits: schizophrenia, autism spectrum, bipolar.
 - ◆ Quantitative variation: height, cholesterol level, blood pressure/hypertension.

Challenges of FP Linkage Analysis for Complex Traits

- ➡ **The appropriate mode of inheritance that adequately explains the disease pattern may not be clear.**

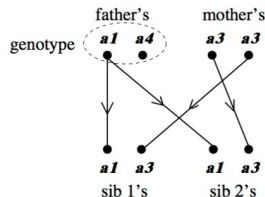
- ➡ Need many parameters!

The conventional four parameters, (p, f_0, f_1, f_2) , needed for the analysis of a simple Mendelian disease is obviously not enough to represent the causal system of a complex disease.

A complete model would need many parameters to model all the factors, such as the joint effect of multiple genes and effects of covariates.

- ➡ **An inference procedure loses its efficiency if there are too many parameters!** So it may not be desirable to use a complex model even if we could find one!

- ➡ An alternative approach: non-parametric linkage analysis, also known (and more suitably called) allele-sharing method!



➡ IBS: a set of alleles are said to be **Identical By State** if they are the same allelic type, e.g.

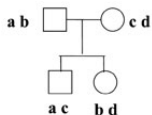
- ◆ a1 of sib 1 and a1 of sib 2
- ◆ a3 of sib 1 and a3 of sib 2

➡ IBD: a set of alleles are said to be **Identical By Descent** if they were inherited from the same ancestral allele (same origin), e.g.

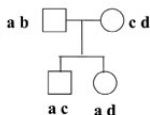
- ◆ a1 of sib 1 and a1 of sib 2.

IBD vs IBS Example

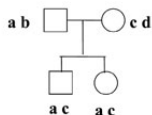
Figure A.1 of the Textbook



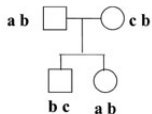
IBS= 0
IBD= 0



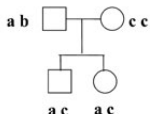
IBS= 1
IBD= 1



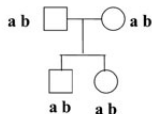
IBS= 2
IBD= 2



IBS= 1
IBD= 0



IBS= 2
IBD= 1 or 2



IBS= 2
IBD= 0 or 2

IBD Distribution

$$\vec{p} = (p_0, p_1, p_2)$$

- ▀ $p_i = P(i \text{ number of alleles are shared IBD by a pair of individuals})$.
- ▀ Can be used to **summarize and measure pairwise relationships**, though not a one-to-one map, e.g.

a MZ-twin pair:	$\vec{p} = (0, 0, 1)$
a full sib pair:	$\vec{p} = (1/4, 1/2, 1/4)$
a half-sib pair:	$\vec{p} = (1/2, 1/2, 0)$
a first-cousin pair:	$\vec{p} = (3/4, 1/4, 0)$
a unrelated pair:	$\vec{p} = (1, 0, 0)$

- ▀ Recall that these IBD distributions are NOT conditional on any observed genotype data. They simply measure the probability that genetic material at a randomly selected locus from the genome to have common ancestry origin between two individuals.
- ▀ It is a characteristic for a population sample of specific relationship type, but NOT for one specific pair of interest.

IBD Distribution Calculation Examples I

E.g. half-sib pair, $\vec{p} = (1/2, 1/2, 0)$

- Let's assume the half-sib pair has the same father
- Let's assume father's genotype is $(a1, a4)$.
- It's 'impossible' to share 2 alleles IBD, since there is no common origin from the two mother's side (unless the two mothers are relatives as well), so

$$p_2 = 0.$$

$$p_1 = P(1 \text{ IBD}) = P(\text{both inherited } a1) + P(\text{both inherited } a4) = \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{2} = \frac{1}{2}.$$

$$p_0 = P(0 \text{ IBD}) = P(\text{sib1 inherited } a1 \text{ and sib2 inherited } a4) \\ + P(\text{sib1 inherited } a4 \text{ and sib2 inherited } a1) = \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{2} = \frac{1}{2}.$$

$$(\text{ or } p_0 = 1 - p_1 - p_2 = 1 - \frac{1}{2} - 0 = \frac{1}{2})$$

- Note that the calculation does not depend on the specific genotype or how we code the genotype as long as we keep track the two origins/alleles. The two alleles could be the same allelic type (e.g. $a3$ and $a3$).

IBD Distribution Calculation Examples II

E.g. full-sib pair, $\vec{p} = (1/4, 1/2, 1/4)$.

- Use the fact that full-sib pair is 'independent' sum of two half-sib pairs in terms of IBD sharing (Mendel's segregation from father to offsprings is independent of that from mother to offsprings.)

$$\begin{aligned} p_2 &= P(2 \text{ IBD}) = P(1 \text{ IBD from father side and 1 IBD from mother side}) \\ &= P(1 \text{ IBD from father side})P(1 \text{ IBD from mother side}) = \frac{1}{2} \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

$$\begin{aligned} p_0 &= P(0 \text{ IBD}) = P(0 \text{ IBD from father side and 0 IBD from mother side}) \\ &= P(0 \text{ IBD from father side})P(0 \text{ IBD from mother side}) = \frac{1}{2} \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

$$\begin{aligned} p_1 &= P(1 \text{ IBD}) = P(1 \text{ IBD from father side and 0 IBD from mother side}) \\ &\quad + P(0 \text{ IBD from father side and 1 IBD from mother side}) \\ &= P(1 \text{ IBD from father side})P(0 \text{ IBD from mother side}) \\ &= P(0 \text{ IBD from father side})P(1 \text{ IBD from mother side}) \\ &= \frac{1}{2} \frac{1}{2} + \frac{1}{2} \frac{1}{2} = \frac{1}{2}. \end{aligned}$$












$$\text{(or } p_1 = 1 - p_0 - p_2 = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2} \text{)}$$

IBD Distribution Calculation Examples III

- ▶ We emphasize here again that calculations of IBD distributions do NOT conditional on any observed genotype data. It is a characteristic for a population sample of specific relationship type, e.g. all full-sib pairs.
- ▶ However, there are situations we might be interested in the IBD information for a particular pair.
- ▶ Assume genotypes of a pair of individuals are $g_1 = (g_{1,1}, g_{1,2})$, $g_2 = (g_{2,1}, g_{2,2})$, e.g. $g_1 = (a1, a3)$ and $g_2 = (a1, a3)$.
- ▶ Note that

$$P(0 \text{ IBD} \mid (a1, a3)(a1, a3) \text{ for a sib pair}) \neq 1/4!$$

- ▶ And this probability will further change if we add genotypes of the parents or genotypes of the sib pair from linked markers!

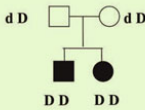
Pedigree	Relationship
	MZ-twin
	parent-offspring
	full-sib
	half-sib+first-cousin
	half-sib
	grandparent-grandchild
	avuncular
	first-cousin
	half-avuncular
	half-first-cousin
	unrelated

Relationship type (notation)	Distribution of IBD Sharing			Kinship coefficient, ϕ
	p_0	p_1	p_2	
MZ-twin (MZ)	0	0	1	0.5
Parent-offspring (PO)	0	1	0	0.25
Full-sib (FS)	0.25	0.5	0.25	0.25
Half-sib + first-cousin (HSFC)	0.375	0.5	0.125	0.1875
Half-sib (HS)	0.5	0.5	0	0.125
Grandparent-grandchild (GPC)	0.5	0.5	0	0.125
Avuncular (AV)	0.5	0.5	0	0.125
First-cousin (FC)	0.75	0.25	0	0.0625
Half-avuncular (HAV)	0.75	0.25	0	0.0625
Half-first-cousin (HFC)	0.875	0.125	0	0.03125
Unrelated (UN)	1	0	0	0

Detecting pedigree relationship errors In Statistical Human Genetics: Methods and Protocols. Elston R, Satagopan J and Sun S (editors), Human Press, Inc. Springer, pp.25 to 46.

Rationale of the Allele-Sharing Method I

- ➡ Consider a simple example of rare recessive disease.



Assumptions:

- Recessive mode of inheritance
- Rare disease allele $D \Rightarrow$ parents are heterozygous
- Fully penetrant and no phenocopies
penetrance function: $f_0 = f_1 = 0$ and $f_2 = 1$

\Rightarrow Affected offspring must carry "DD"

IBD=0:	$p_0=0$
IBD=1:	$p_1=0$
IBD=2:	$p_2=1$

In general, p_0 , p_1 and p_2 depend on the allele frequencies of D , the mode of inheritance and the penetrance function.

- ➡ IBD distribution for **affected sib pairs at DSL** is

$$\vec{p} = (0, 0, 1).$$

Rationale of the Allele-Sharing Method II

- ➡ The expected number of alleles shared IBD by affected sib pairs at the DSL is then

$$E(\# \text{ IBD at DSL} | \text{affected sibs}) = 1p_1 + 2p_2 = 2.$$

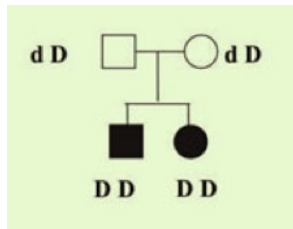
- ➡ What is the IBD sharing at a marker N that has nothing to do with the disease (no linked to DSL), say on a different chromosome from DSL? The IBD distribution should follow the null distribution of $\vec{p} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, and

$$E(\# \text{ IBD at } N | \text{affected sibs, 2 IBD at DSL}) = 1p_1 + 2p_2 = 1.$$

- ➡ So, if we gather a large number of affected sib pairs, then we should
 - ◆ observe **Excess IBD Sharing** at the DSL
 - ◆ but not at markers that are not linked to DSL!
- ➡ BUT, we don't know the location of DSL!

Rationale of the Allele-Sharing Method III

- ▶ We do have data for a large number of markers spaced across the genome.
- ▶ What is the IBD sharing at a marker M that is linked with DSL at θ ?
- ▶ Note that the allele notation of (M_1 , M_2 , M_3 , M_4) is arbitrary, they can be the same allelic type and will not affect the probability calculation. All we need is to keep track the **origin** of each transmitted allele.

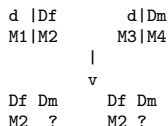


father		mother
d Df		d Dm
M1 M2		M3 M4
	v	
sib 1		sib 2
Df Dm		Df Dm
? ?		? ?

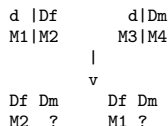
Rationale of the Allele-Sharing Method IV

Consider alleles inherited from father's side

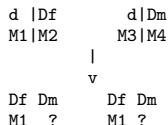
- ◆ M2 1 IBD with probability: $(1 - \theta)^2$.



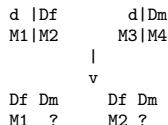
- ◆ 0 IBD with probability: $(1 - \theta)\theta$.



- ◆ M1 1 IBD with probability: θ^2 .



- ◆ 0 IBD with probability: $\theta(1 - \theta)$.



Rationale of the Allele-Sharing Method V

➡ In summary, from father's side

$$\begin{array}{ll} 1 \text{ IBD with probability} & \theta^2 + (1 - \theta)^2 = \phi \\ 0 \text{ IBD with probability} & 2\theta(1 - \theta) = \psi = 1 - \phi \end{array}$$

➡ Meiosis of the two parents are independent of each other, so together

2 IBD	1 IBD AND 1 IBD	$(\theta^2 + (1 - \theta)^2)^2$	ϕ^2
1 IBD	1 IBD AND 0 IBD	$(\theta^2 + (1 - \theta)^2) \cdot 2\theta(1 - \theta)$	$2\phi\psi$
	OR		
	0 IBD AND 1 IBD	$2\theta(1 - \theta) \cdot (\theta^2 + (1 - \theta)^2)$	
0 IBD	0 IBD AND 0 IBD	$(2\theta(1 - \theta))^2$	ψ^2

Rationale of the Allele-Sharing Method VI

➡ So we have

$$p_0 = P(0 \text{ IBD at M} | \text{affected sibs, 2 IBD at DSL}) = \psi^2$$

$$p_1 = P(1 \text{ IBD at M} | \text{affected sibs, 2 IBD at DSL}) = 2\phi\psi$$

$$p_2 = P(2 \text{ IBD at M} | \text{affected sibs, 2 IBD at DSL}) = \phi^2$$

$$\phi = \theta^2 + (1 - \theta)^2, \quad \psi = 1 - \phi = 2\theta(1 - \theta)$$

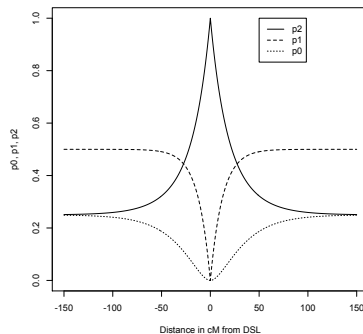
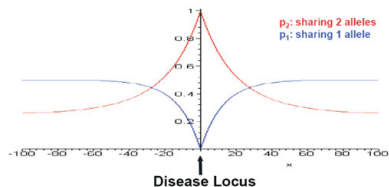
$$\theta = \frac{1 - e^{-2t}}{2}$$

t is the distance between M and DSL in Morgan.

Rationale of the Allele-Sharing Method VII

R codes

➡ This is the basis of Figure A.4.

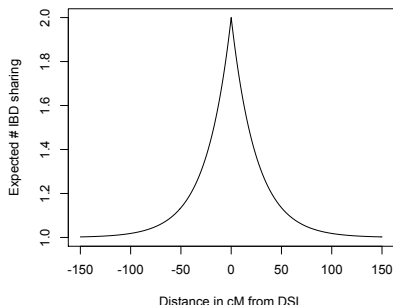


Rationale of the Allele-Sharing Method VIII

- ➡ The expected number of alleles shared IBD at the linked marker locus M:

$$\begin{aligned} E(\# \text{ IBD at M} | \text{affected sibs, 2 IBD at DSL}) &= 1 \cdot p_1 + 2 \cdot p_2 \\ &= 2\phi(1 - \phi) + 2\phi^2 > 1 \end{aligned}$$

- ➡ **The amount of oversharing increases as M gets closer to DSL!**



Additional Notes on the Rationale

- ➡ Note that the decay plot is conditional on the allele sharing at the DSL itself is 2. More generally,

$$E(\# \text{ IBD at M} | \text{affected sibs})$$

$$\begin{aligned} &= E(\# \text{ IBD at M} | \text{affected sibs}, 2 \text{ IBD at DSL})P(2 \text{ IBD at DSL} | \text{affected sibs}) \\ &+ E(\# \text{ IBD at M} | \text{affected sibs}, 1 \text{ IBD at DSL})P(1 \text{ IBD at DSL} | \text{affected sibs}) \\ &+ E(\# \text{ IBD at M} | \text{affected sibs}, 0 \text{ IBD at DSL})P(0 \text{ IBD at DSL} | \text{affected sibs}) \end{aligned}$$

- ➡ The marginal IBD distribution at the DLS depends on the disease model, i.e. penetrance f_i and frequency of the mutation p .

$$p_i = P(i \text{ IBD at DSL} | \text{affected sibs}) = \text{function of}(p, f_0, f_1, f_2).$$

- ➡ The calculation of $E(\# \text{ IBD at M} | \text{affected sibs}, i \text{ IBD at DSL})$ for $i = 0$ and 1 follows what we have learned for $i = 2$, but we do need the IBD transition probability $p_{ij} = P(j \text{ IBD at locus 2} | i \text{ IBD at locus 1})$ for a sib pair.

IBD Transition Probability I

$p_{ij} = P(j \text{ IBD at locus 2} | i \text{ IBD at locus 1})$ for a **full-sib pair**

IBD state Locus 1	IBD state Locus 2		
	0	1	2
0	ϕ^2	$2\phi\psi$	ψ^2
1	$\phi\psi$	$\phi^2 + \psi^2$	$\phi\psi$
2	ψ^2	$2\phi\psi$	ϕ^2

➡ Note that this is essentially Table A.3 of the Textbook with a slightly different notation.

$$\phi = \theta^2 + (1 - \theta)^2, \quad \psi = 1 - \phi = 2\theta(1 - \theta)$$

$$\theta = \frac{1 - e^{-2t}}{2}$$

t is the distance between M and DSL in Morgan.

IBD Transition Probability II

- ➡ Note that this **transition probability is relationship specific**, i.e. a different relationship type will have a different transition probability table. E.g. half-sib pairs.

IBD state Locus 1	IBD state Locus 2	
	0	1
0	ϕ	ψ
1	ψ	ϕ

- ➡ Finally, note that (related to Markov property of the inheritance process)

$$P(j \text{ IBD at locus 2} | i \text{ IBD at locus 1, both sibs are affected})$$

is the same as

$$P(j \text{ IBD at locus 2} | i \text{ IBD at the locus 1, any full-sib pair}).$$

- ➡ The phenotype information affects the calculation of the marginal IBD distribution for DSL or any marker linked to the DSL,
 $P(i \text{ IBD at a locus} | \text{both sibs are affected}) \neq P(i \text{ IBD at a locus} | \text{any full-sib pair}),$
but does not affect the IBD transition probability.

IBD Sharing at the DSL Calculation I

Calculating the IBD sharing S for a pair of affected siblings given the disease model (p, f_0, f_1, f_2) .

$$p_i = P(S = i | \text{both sibs affected}), i = 0, 1, 2$$

- Let $G = (G1_1 G1_2, G2_1 G2_2)$ be the genotypes at the disease locus for a pair of full-sibs, where $G1_1$ and $G1_2$ are the two alleles for sib 1, and $G2_1$ and $G2_2$ are the two alleles for sib 2.
- For a two-allele marker/gene (alleles D and d), the 6 possible unordered genotypes are $g_1 = (DD, DD)$, $g_2 = (DD, Dd)$, $g_3 = (DD, dd)$, $g_4 = (Dd, Dd)$, $g_5 = (Dd, dd)$, and $g_6 = (dd, dd)$.

IBD Sharing at the DSL Calculation II

- ➡ Note that we assume

$$P(\text{both aff} | S = i, G) = P(\text{both aff} | G).$$

That is, the genotypes at the DSL fully determine probability of being affected. However, this assumption can be violated if there are multiple linked DSL.

- ➡ Also note that

$$P(\text{both aff} | G) = P(\text{sib 1 aff} | G_1 G_2) \cdot P(\text{sib 2 aff} | G_2 G_2).$$

Again, this assumption may be violated if the environmental factors are correlated with G .

IBD Sharing at the DSL Calculation III

- Finally, recall the conditional probabilities of genotypes of a pair of individuals given their IBD sharing, $P(G|S)$:

Unordered Genotype	IBD status		
	0	1	2
(ii, ii)	f_i^4	f_i^3	f_i^2
(ii, ij)	$4f_i^3 f_j$	$2f_i^2 f_j$	0
(ii, jj)	$2f_i^2 f_j^2$	0	0
(ii, jk)	$4f_i^2 f_j f_k$	0	0
(ij, ij)	$4f_i^2 f_j^2$	$f_i f_j (f_i + f_j)$	$2f_i f_j$
(ij, ik)	$8f_i^2 f_j f_k$	$2f_i f_j f_k$	0
(ij, kl)	$8f_i f_j f_k f_l$	0	0

- f_i, f_j, f_k and f_l are allele frequencies for alleles i, j, k and l .
- Genotypes are unordered within and between individuals, $(ij\ kl) = (lk\ ji) \dots$
- Note that the above conditional probabilities are **independent of the relationship of the pair**.
- Thompson (1975). [The estimation of pairwise relationships](#)
- In the case that there are only two alleles, the table can be simplified a bit.

IBD Sharing at the DSL Calculation IV

Now, we can do the calculation.

$$\begin{aligned} P(S = i | \text{both aff}) &= \frac{P(S = i, \text{both aff})}{P(\text{both aff})} \\ &= \frac{\sum_{G=g_1, \dots, g_6} P(S = i, G, \text{both aff})}{\sum_{j=0,1,2} \sum_{G=g_1, \dots, g_6} P(S = j, G, \text{both aff})} \\ &= \frac{\sum_{G=g_1, \dots, g_6} P(\text{both aff} | S = i, G) P(G | S = i) P(S = i)}{\sum_{j=0,1,2} \sum_{G=g_1, \dots, g_6} P(\text{both aff} | S = j, G) P(G | S = j) P(S = j)} \\ &= \frac{\sum_{G=g_1, \dots, g_6} P(\text{both aff} | G) P(G | S = i) P(S = i)}{\sum_{j=0,1,2} \sum_{G=g_1, \dots, g_6} P(\text{both aff} | G) P(G | S = j) P(S = j)} \\ &= \frac{\sum_{G=g_1, \dots, g_6} P(\text{sib 1 aff} | G_1 G_2) P(\text{sib 2 aff} | G_2 G_2) P(G | S = i) P(S = i)}{\sum_{j=0,1,2} \sum_{G=g_1, \dots, g_6} P(\text{sib 1 aff} | G_1 G_2) P(\text{sib 2 aff} | G_2 G_2) P(G | S = j) P(S = j)} \end{aligned}$$

IBD Sharing at the DSL Calculation V

- ◆ Note that $P(\text{aff}|G_1G_2)$ is specified by the penetrance, e.g.

$$f_2 = P(\text{aff}|DD) = 0.6, f_1 = 0.4, f_0 = 0.1.$$

- ◆ $P(G|S = j)$ is given in the table of conditional probability of genotypes of a pair of individuals given their IBD sharing, e.g. $f_D = 0.1$ and

$$P(DD \ Dd|S = 1) = 2f_D^2f_d = 2(0.1)^2(1 - 0.1) = 0.018.$$

- ◆ $P(S = i)$ follow the null IBD distribution, e.g. for a pair of sibs,

$$(P(S = 0), P(S = 1), P(S = 2)) = (p_0, p_1, p_2) = (1/4, 1/2, 1/4),$$

because the affection status of the sib pair is not present in the formula.

IBD Sharing at the DSL Calculation VI

- Results are (it's easier to write a program to do the calculation):

$$P(\text{both sibs affected}) = 0.032870,$$

$$P(S_D = 0, \text{both sibs affected}) = 0.006320,$$

$$P(S_D = 1, \text{both sibs affected}) = 0.016425,$$

$$P(S_D = 2, \text{both sibs affected}) = 0.010125.$$

$$P(S_D = 0 | \text{both sibs affected}) = 0.192279 (< 1/4),$$

$$P(S_D = 1 | \text{both sibs affected}) = 0.499692 (\approx 1/2), \text{ and}$$

$$P(S_D = 2 | \text{both sibs affected}) = 0.308029 (> 1/4).$$

$$E[S_D | \text{both sibs affected}]$$

$$= 1 \cdot P(S_D = 1 | \text{both sibs affected}) + 2 \cdot P(S_D = 2 | \text{both sibs affected}) = 1.115751$$

Principle of the Allele-Sharing Method

- ➡ A pair of relatives who are similar w.r.t. the trait phenotype, say two sibs who are both affected, are likely to have the same disease allele, at the DSL D , inherited from a particular parent. Thus, affected relative pairs tend to have **excess IBD sharing at the DSL D locus**.
- ➡ If a marker locus M is linked to D , then the affected relative pair also tend to have IBD oversharing at the M locus because of **co-segregation due to linkage**.
- ➡ If a marker locus M is not linked to gene locus D , then there should be no such correlation of allele sharings at the two loci because of **independent segregation among unlinked loci**.
- ➡ So, allele-sharing method looks for markers that have IBD oversharing, among affected relatives, than what expected under the null hypothesis of linkage! Significant oversharing is an indication of linkage.

Affected Sib-Pair (ASP) Method I

- ➡ Collected n families, each family with 2 affected offsprings.
- ➡ Genotype a large number of genetic markers throughout the genome for each individual.
- ➡ At each marker locus, infer the number of alleles shared IBD by the i_{th} affected sib pair,

$$S_i.$$

If a marker is highly polymorphic, IBD status may be inferred with certainty based on the genotype data of the parents and offsprings. If not, posterior probability has be calculated (discussion later).

- ➡ For a given marker, the null hypothesis is:

H_0 : the marker is not linked to the DSL

Affected Sib-Pair (ASP) Method II

- Standardize the IBD sharing statistic S_i by its null mean and sd:

$$Z_i = \frac{S_i - \mu_i}{\sigma_i},$$

- Based on the null IBD distribution for a sib pair,
 $(p_0, p_1, p_2) = (1/4, 1/2, 1/4)$:

$$\mu = E_0(S_i) = 1 \cdot p_1 + 2 \cdot p_2 = 1,$$

$$\sigma^2 = \text{Var}_0(S_i) = \frac{1}{2}.$$

- Note that μ_i and σ_i are independent of markers/loci, as they are calculated independent of data. Values of μ_i and σ_i are fully determined by the pedigree structure or relationship type.

Affected Sib-Pair (ASP) Method III

- Linear combination across families to aggregate the data to obtain Non-Parametric Linkage (NPL) score.

$$NPL = Z = \frac{\sum_{i=1}^n Z_i}{\sqrt{n}}.$$

- If n is large, by CLT and under the null of no linkage,

$$H_0 : Z \sim N(0, 1).$$

- For each of genotyped markers, perform the corresponding hypothesis test, and significant oversharing at some marker(s) or the peak(s) is an indication of linkage to the DSL.

Multiple Hypothesis Testing Issue

- ➡ Because we are doing a genome-wide scan of many markers, similar to that for the full-parametric linkage analysis, the multiple hypothesis testing issue occur.
- ➡ A p-value of 2×10^{-5} corresponds approximately to a genome-wide adjusted p-value of 0.05.
- ➡ Lander ES, Kruglyak L (1995). [Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results](#). Nature Genetics 11:241-247.

Additional Considerations - Other Types of Relative Pairs I

- ➡ We can use the same principle, even though the n collected families may have different types of affected relative pairs.

- ➡ In normalization:

$$Z_i = \frac{S_i - \mu_i}{\sigma_i}$$

$\mu_i = E_0[S_i]$, and $\sigma_i = \sqrt{\text{Var}_0(S_i)}$ would be i_{th} family specific depending on the relationship of the affected relative pair.

- ➡ E.g. in the case of a first-cousin pair, based on the null IBD distribution:
 $(p_0, p_1, p_2) = (3/4, 1/4, 0)$:

$$\mu = 1/4$$

$$\sigma = \sqrt{3/16}.$$

Additional Considerations - Other Types of Relative Pairs II

- ➡ In linear combination:

$$Z = \frac{\sum_{i=1}^n w_i Z_i}{\sqrt{\sum_{i=1}^n w_i^2}}$$

w_i would be the weighting factor for the i_{th} relative pair,

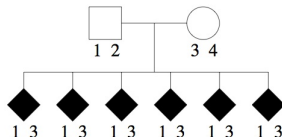
e.g. $w_i = 1$, $w_i = \sigma_i$, or $w_i = \sqrt{\sigma_i}$.

- ➡ Previously in the ASP design, we used $w_i = 1$ for all i .
- ➡ What are the optimal weights? **Optimal allele-sharing statistics for genetic mapping using affected relatives.**

Additional Considerations - Multiple Affected I

If there are multiple affecteds in a family, how to we summarize the IBD sharing?

➡ e.g. five affected sibs in one family.



➡ S_{pair} : sum of the pairwise IBD sharing among affecteds.

➡ S_{all} : average number of non-trivial permutations that preserve an ordered set obtained by choosing one allele from each affected member of the pedigree.

S_{all} gives sharply increasing values as the number of affected individuals sharing a particular allele increase.

Additional Considerations - Multiple Affected II

➡ For the case shown in the graph

#sibs	S_pair	S_all
2	$2 \times 1 = 2$	$2/2^2 = 0.5$
3	$2 \times 3 = 6$	$16/2^3 = 2$
4	$2 \times 6 = 12$	$104/2^4 = 6.5$
5	$2 \times 10 = 20$	$688/2^5 = 21.5$
6	$2 \times 15 = 30$	$4976/2^6 = 77.75$

➡ A few notes on the allele-sharing scoring function, S .

- ◆ S is a function defined based on the IBD sharing among the affected individuals in a family at a given marker/locus.
- ◆ e.g. $S = \alpha \cdot p_1 + \beta \cdot p_2$ (and we have been using $S = 1 \cdot p_1 + 2 \cdot p_2$).
- ◆ In general, S can be any function that has a higher expected value under the alternative of linkage than under the null of no linkage.
- ◆ $0.25 \cdot p_1 + 1 \cdot p_2$: Strategies for mapping heterogeneous recessive traits by allele-sharing methods.
- ◆ $0.275 \cdot p_1 + 1 \cdot p_2$: Simple, robust linkage tests for affected sibs.

More Additional Considerations I

- Textbook page 197 used a likelihood approach for ASP design, testing the data is multinomial

$$H_0 : (p_0, p_1, p_2) = (1/4, 1/2, 1/4).$$

What are the disadvantages of this approach?

- Can we collect say discordant sib pairs (1 affected and 1 unaffected) and look for **undersharing**, or even both unaffected and look for oversharing?
- How to include unaffecteds if they have been collected?
- How do we incorporate important covariates (risk factors) such as age and smoking status?

Analysis of affected sib pairs, with covariates—with and without constraints.

More Additional Considerations II

- How to allow for heterogeneity in the analysis? (Power may be greatly reduced if locus/genetic heterogeneity is present in the data.)

Ordered subset analysis in genetic linkage mapping of complex traits.

- How to robustify the linkage analysis against various assumptions/uncertainties, e.g. HWE, Mendelian law of inheritance, and allele frequency estimation?

- How to joint analyze multiple disease loci? (Study the correlation of IBD sharing at two unlinked loci.)

Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans.

- How to estimate the recombination fraction θ between the putative DSL and the linked marker in this allele-sharing method context?

A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases.

Simultaneous localization of two linked disease susceptibility genes.

More Additional Considerations III

➡ How do we analyze quantitative traits?

- ◆ A Unified Haseman-Elston Method for Testing Linkage with Quantitative Traits.

The Haseman and Elston (H-E) method uses a simple linear regression to model the squared trait difference of sib pairs with the shared allele identical by descent (IBD) at marker locus for linkage testing.

- ◆ If a sib-pair is genetically similar (as measured by IBD) at the DSL, the sib-pair should also be phenotypically similar.
- ◆ If a sib-pair is genetically dissimilar at the DSL, the sib-pair should also be phenotypically dissimilar.
- ◆ Due to linkage, we can say the same for a marker M that is linked to the DSL!
- ◆ The allele-sharing method for binary traits used the exact same argument!

Missing IBD Sharing Information I

➡ In practice, IBD status may not be determined with certainty, e.g.

family 1		family 2	family 3
father	mother	father	mother
1 2	1 3	1 1	2 3
1 1	1 1	1 1	1 1
sib1	sib2	sib1	sib2
2 3	1 3	1 2	1 3
1 1	1 1	1 1	1 1
1 IBD w/p 1		0 IBD w/p 1/2	0 IBD w/p 1/4
		1 IBD w/p 1/2	1 IBD w/p 1/2
			2 IBD w/p 1/4
complete info.		partial info.	no info.

➡ If there were no genotypes for the parents, then there would be more uncertainty.

Missing IBD Sharing Information II

- ➡ The solution is to use the observed genotype data to calculate the posterior probability of the sharing statistic S , e.g. in ASP, calculate

$$E(S|\text{genotype data})$$

$$= 1 \cdot P(S = 1|\text{genotype data}) + 2 \cdot P(S = 2|\text{genotype data})$$

and replace S with $E[S|\text{genotype data}]$ in the calculation of the NPL score:

$$NPL = Z = \frac{\sum_{i=1}^n w_i Z_i}{\sqrt{\sum_{i=1}^n w_i^2}}$$

where

$$Z_i = \frac{E[S_i|\text{genotype data}_i] - \mu_i}{\sigma_i}, \text{ for the } i_{th} \text{ family}$$

Missing IBD Sharing Calculation I

- ➡ The key component is the inheritance vector inference.
- ➡ A Inheritance vector at a marker/locus contain all the inheritance information at that locus in a pedigree, and completely determine the IBD sharing at that locus, e.g.

Missing IBD Sharing Calculation II

```

father x mother <- founder
(0) (1) (0) (1) <- indicator for origin
a1 a4 a3 a3 <- genotype

      |
      V

sib1      sib2 <- nonfounder
a1 a3 a4 a3 <- genotype
    
```

inheritance vector for nonfounders	prior prob.	posterior prob.	#IBD by the sib pair
0 0 0 0	1/16	0	
0 0 0 1	1/16	0	
0 0* 1 0*	1/16	1/4	1
0 0 1 1	1/16	1/4	0
0 1 0 0	1/16	0	
0 1 0 1	1/16	0	
0 1 1 0	1/16	1/4	0
0 1* 1 1*	1/16	1/4	1
1 0 0 0	1/16	0	
1 0 0 1	1/16	0	
1 0 1 0	1/16	0	
1 0 1 1	1/16	0	
1 1 0 0	1/16	0	
1 1 0 1	1/16	0	
1 1 1 0	1/16	0	
1 1 1 1	1/16	0	

$$E[S|\text{genotype data}] = 1*1/4 + 0*1/4 + 1*1/4 + 0*1/4 = 1/2$$

Missing IBD Sharing Calculation III

- ➡ The calculation of

$$P(S = i | \text{genotype data}), i = 0, 1, 2$$

can be quite complicated for multipoint analysis and large pedigree, because the genotype data include genotypes of ALL markers linked to the marker under study and from ALL family members

- ➡ Genotypes of linked markers and related individuals provide inheritance information for each other! Again
 - ◆ **Elston-Stewart Peeling algorithm.** Designed for large pedigrees, but not with large number of markers.
 - ◆ **Lander-Green Hidden Markov Model (HMM) algorithm.** Limited by the size of the pedigree, but computational time is linear w.r.t. the number of markers. More commonly used these days due to the availability of genotype data at a large number of markers.

Missing IBD Sharing Calculation IV

- ➡ GENEHUNTER relies on the HMM algorithm. It handles $2n - f < 16$ sized-pedigrees, where n and f are the numbers of nonfounders and founders in a pedigree.
- ➡ There are improvement (e.g. GENEHUNTER-PLUS, ALLEGRO), but there still is a limit in pedigree size.
- ➡ Besides “trimming” of one big pedigree to multiple smaller pedigrees, alternatives include SIMWALK and SIMWALK2: *uses Markov chain Monte Carlo (MCMC) and simulated annealing algorithms to perform these multipoint analyses.*
- ➡ Most of the genetics software are listed on this website: [GitHub](#).

Missing IBD Sharing Consequences I

- ➡ Recall that in the case of incomplete data, the test statistic:

$$NPL = Z = \frac{\sum_{i=1}^n w_i Z_i}{\sqrt{\sum_{i=1}^n w_i^2}},$$

where

$$Z_i = \frac{E[S_i | \text{data}] - \mu_i}{\sigma_i}, \text{ for the } i_{th} \text{ family}$$

$$\mu_i = E[S_i], \text{ and } \sigma_i = \sqrt{\text{Var}(S_i)}$$

- ➡ If n is large, CLT can be applied, and Z is approximately normally distributed. However, under the null hypothesis of no linkage, do we still have?

$$Z \sim N(0, 1) ?$$

Missing IBD Sharing Consequences II

- ➡ The short answer is that the variance is ≤ 1 , depending on the amount of missing information. Detailed reasoning and remedies are discussed in more advanced stat gene module.

$$E[X] = E[E[X|Y]],$$

$$\text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)].$$

- ➡ Kong and Cox (1997). [Allele-sharing models: LOD scores and accurate linkage tests.](#)
- ➡ Dan Nicolae (1999). Allele sharing models in gene mapping: a likelihood approach. Ph.D. thesis, Department of Statistics, University of Chicago.
- ➡ Nicolae and Kong (2004). [Measuring the Relative Information in Allele-Sharing Linkage Studies](#)

Full-parametric vs. Allele-sharing Linkage Analysis

- ➡ For single-locus high-penetrance traits, full parametric analysis on large pedigrees is generally more powerful. However, the power is reduced if an incorrect model is used.
- ➡ For many complex diseases, modes of inheritance are poorly understood, allele-sharing methods are more robust. However, the “optimal” scoring function depends on the underlying disease model.
- ➡ Allele-sharing methods are NOT fundamentally different from parametric linkage method. The robustness comes from the fact that the dimension of the parameter space is smaller.
- ➡ In fact, both methods need the SAME inheritance vector inference (e.g. the Lander-Green HMM algorithm) so that we can obtain the missing recombination information for the parametric θ linkage analysis, or the missing IBD information for the non-parametric allele-sharing linkage analysis.
- ➡ Kruglyak et al. (1996). [Parametric and nonparametric linkage analysis: a unified multipoint approach.](#)

➡ See separate notes

Exercises

- ▶ Chapter 6 Exercise 4(a)-(c).
- ▶ Chapter 6 Exercise 5.
- ▶ Chapter 6 Exercise 6.
- ▶ Chapter 6 Exercise 7(b)-(c).

What's Next

- ➡ Part of Chapter 5 - Linkage Disequilibrium
- ➡ Chapter 7 - The Basics of Genetic Association Analysis