BioStat 710 HW3

Statistical Genetics and Genetics Epidemiology

Guangjian (Jeff) Du

Question 1. Explain why linkage disequilibrium is essential for association mapping with a limited number of markers.

Answer:
If all polymorphisms were independent at the population level, association studies would have to examine every one of them. Linkage disequilibrium makes tightly linked variants strongly correlated producing cost savings for association studies. Linkage Disequilibrium Enables Genetic Association Studies.

Question 2. Explain why association mapping is better at fine mapping disease loci than linkage analysis

Answer:
Linkage analysis, linkage study is the following: if a disease runs in a family, one could look for genetic markers that run exactly the same way in the family (from grandma, to dad, to me, for example). If we find one, we assume the gene that causes the disease is somewhere in the same area of the genome as the marker.
association mapping, gather some people with a disease and some people with out a disease, and look to see if a certain allele (or genotype) is present more often in the cases than the controls.

If the allele plays a role in causing the disease, or is correlated with a causal allele, it will have a higher frequency in the case population than the control population.

After a linkage study, one nominates "candidate genes" in the region under the linkage signal, and performs an association study on alleles in the genes. In this way, a specific gene, or even a specific allele, can be identified as playing a possible causal role in the disease. The resolution is much higher, but it was previously implausible to perform these sorts of studies on regions much larger than a couple genes.

However, with the HapMap and the technology to genotype hundreds of thousands of alleles in parallel, it's now possible to perform association studies on the level of the whole genome. This would essentially skip the step of a linkage scan. In contrast to linkage studies, association maping (association studies) can identify variants with relatively small individual contributions to disease risk.

Question 3. Explain how linkage disequilibrium originates in a population and what causes it to decay

Answer:
Origination, Linkage disequilibrium occurs when genotypes at the two loci are not independent of another. Alleles that exist today arose through ancient mutation events; One allele arose first, and then the other; Recombination generates new arrangements for ancestral alleles.

Decay, linkage disequilibrium decays each generation at a rate determined by the degree of recombination.
I.E. After t generations: $D^t_{AB} = (1-\theta)^t * D^0_{AB}$

TEXT 1. Briefly explain the concept of recombination (what is a recombinant chromosome?) and define the recombination fraction. What is the relationship between the recombination fraction and linkage?

Answer:
Recombination: the rearrangement of genetic material, especially by crossing over in chromosomes or by the artificial joining of segments of DNA from different organisms.
Recombinant chromosome: Crossing over occurs between prophase 1 and metaphase 1 and is the process where homologous chromosomes pair up with each other and exchange different segments of their genetic material to form recombinant chromosomes. It can also happen during mitotic division, which may result in loss of heterozygosity.
Recombination fraction: also called recombination frequency, between two loci is defined as the ratio of the number of recombined gametes to the total number of gametes produced.
Relationship between recombination fraction and linkage: Disease-mapping approaches that use the joint transmission of affection status and alleles at the marker locus to localize the disease gene are called Linkage Analysis. The term linkage refers to the failure of Mendels second Law of independent assortment; or more specifically, the situation where the recombination fraction parameter $\theta < 1$ between two loci. Two loci are said to be unlinked if $\theta = 1$, and correlate spondingly, Mendels second law of independent assortment holds.
Formally, to test for linkage, the null hypothesis is $H_0 : \theta = 1$ (or equivalently, no linkage) and the alternative hypothesis is $H_A : \theta < 1$ (or equivalently, linkage is present).

TEXT 3. Suppose a population of 2000 chromosomes; 1000 carry an A allele at a marker and 1000 carry a. Now suppose a disease mutation (+) arises on one chromosome bearing an A allele, and all the rest of the chromosomes have - at that location.

(a) What are the marginal frequencies at the marker and DSL?

Answer:

Marginal Frequencies at the markers:

$$Feq_A = 1000, \quad P_A = \frac{1000}{2000} = 0.5$$

$$Feq_a = 1000, \quad P_a = \frac{1000}{2000} = 0.5$$

Marginal Frequencies at the DSL:

$$Feq_+ = 1, \quad P_+ = \frac{1}{2000} = 1/2000$$

$$Feq_1 = 1999, \quad P_- = \frac{1999}{2000} = 1999/2000$$

*(b) Fill in the 2 * 2 table of marker and disease mutation haplotypes.*

HaploTypes Table

	Disease	Locus	
A Locus	+	-	Raw Total
A	1	999	1000
a	0	1000	1000
Col Total	1	1999	2000

The 2 * 2, marker and disease mutation HaploTypes Table.

(c) What are D and D' for this table?

Answer:

$$D = \frac{1 - 1000 * \frac{1}{1999+1}}{2000} = 1/4000$$

$$D_{max} = \frac{min(1000 * 1, 1000 * 1999)}{2000^2} = 1/4000$$

$$D' = \frac{D}{D_{max}} = \frac{1/4000}{1/4000} = 1$$

(d) What is the correlation between the marker and DSL?

Answer:

$$r = \frac{D}{\sqrt{P_A * P_B * P_a * P_b}} = \frac{1/4000}{\sqrt{\frac{1000}{2000} * \frac{1000}{2000} * \frac{1}{2000} * \frac{1999}{2000}}} = \frac{1/4000}{\frac{\sqrt{1999}}{4000}}$$

$$r = \frac{1}{\sqrt{1999}} = 0.0223$$

(e) Repeat the questions above, now assuming only 100 chromosomes, one mutation on the same haplotype as an A allele, and a 50/50 split of A and a alleles.
Answer (e)-1

Marginal Frequencies at the markers:

$$Feq_A = 50, \quad P_A = \frac{50}{100} = 0.5$$

$$Feq_a = 50, \quad P_a = \frac{50}{100} = 0.5$$

Marginal Frequencies at the DSL:

$$Feq_+ = 1, \quad P_+ = \frac{1}{100} = 0.01$$

$$Feq_1 = 99, \quad P_- = \frac{99}{100} = 0.99$$

Answer (e)-2

HaploTypes Table

	Disease	Locus	
A Locus	+	-	Raw Total
A	1	49	50
a	0	50	50
Col Total	1	99	100

Answer (e)-3

$$D = \frac{1 - 50 * \frac{1}{99+1}}{100} = 1/200$$

$$D_{max} = \frac{min(50 * 1, 50 * 99)}{100^2} = 1/200$$

$$D' = \frac{D}{D_{max}} = \frac{1/200}{1/200} = 1$$

Answer (e)-4

$$r = \frac{D}{\sqrt{P_A * P_B * P_a * P_b}} = \frac{1/200}{\sqrt{\frac{50}{100} * \frac{50}{100} * \frac{1}{100} * \frac{99}{100}}} = \frac{1/200}{\frac{\sqrt{99}}{2 * 100}}$$

$$r = \frac{1}{\sqrt{99}} = 0.10$$

(f) What is the predicted value of D after 10 rounds of random mating if θ = 0.4? if θ = 0.01?

Answer:
According to Equation 5.3 on textbook page 82:
$D_t = (1 - \theta)^t * D_0$

In case I:
$\theta = 0.4$

$D_0 = 1/4000$

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

$D_{10} = (1 - 0.4)^{10}$

$D_{10} = 1.5 * 10^{-6}$

In case I:
$\theta = 0.01$

$D_0 = 1/4000$

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

$D_{10} = (1 - 0.01)^{10}$

$D_{10} = 2.3 * 10^{-4}$

TEXT 5. Show that the maximum value of the correlation (+1) between two biallelic loci is reached when the marginal allele frequencies at locus are the same P(A) = P(B), and the two off-diagonal cells of the 2 * 2 table are zero. What are the corresponding requirements for an r of 1?

Answer:

HaploTypes Table I

	Disease	Locus	
A Locus	B	b	Raw Total
A	P_{AB}	0	P_A
a	0	P_{ab}	P_a
Col Total	P_B	P_b	Total = 1

Here, according to the Haplotypes table, we could have:
$P_{AB} = P_A = P_B = p$
$P_{ab} = P_a = P_b = (1 - p)$

Replace $P_{AB}, P_A, and\ P_B\ with\ p$

Replace $P_{ab}, P_a, and\ P_b\ with\ (1 - p)$

HaploTypes Table II

	Disease	Locus	
A Locus	B	b	Raw Total
A	p	0	p
a	0	(1-p)	(1-p)
Col Total	p	(1-p)	1

$$D = P_{AB} - P_A * P_B$$
$$r = \frac{D}{\sqrt{P_A * P_B * P_a * P_b}}$$
$$r = \frac{p - p * p}{\sqrt{p * p * (1 - p) * (1 - p)}}$$
$$r = \frac{p(1 - p)}{p(1 - p)}$$
$$r = 1$$

TEXT 6. Show that the maximum value of D' is 1 when any cell of the 2 * 2 table is zero.

Answer:

HaploTypes Table Originals

A Locus	Disease B	Locus b	Raw Total
A	P_{AB}	P_{Ab}	P_A
a	P_{aB}	P_{ab}	P_a
Col Total	P_B	P_b	Total = 1

Case I: $P_{AB} = 0$

We would get a new HaploTypes table:

HaploTypes Table Originals

A Locus	Disease B	Locus b	Raw Total
A	0	P_A	P_A
a	P_B	P_{ab}	P_a
Col Total	P_B	P_b	Total = 1

HERE

$P_b = P_A + P_{ab}$; $P_a = P_B + P_{ab}$

$D = P_{AB} - P_A * P_B$

The value of D is negative, so for D_{min}, we have:

$D_{min} = -min(P_A * P_B,\ P_a * P_b)$

Plug in P_a and P_b:

$D_{min} = -min(P_A * P_B,\ (P_A + P_{ab}) * (P_B + P_{ab}))$

$D_{min} = -(P_A * P_B)$

$$D' = \frac{D}{D_{min}} = \frac{0 - P_A * P_B}{-(P_A * P_B)} = 1$$

In this case, the maximum value of D' is 1.

Case II: $P_{ab} = 0$ We would get a new HaploTypes table:

HaploTypes Table Originals

	Disease	Locus	
A Locus	B	b	Raw Total
A	P_{AB}	P_{Ab}	P_A
a	P_{aB}	0	P_a
Col Total	P_B	P_b	Total = 1

HERE
$P_a = P_{aB}$; $P_b = P_{Ab}$
$P_A = P_{AB} + P_b$; $P_B = P_{AB} + P_a$
$D = P_{AB} - P_A * P_B = P_A - P_b - P_A * P_B = P_A * (1 - P_B) - P_b = P_A * P_b - P_b$
$D = P_b * (P_A - 1) = -P_a * P_b$
We have D value negative, $D_{min} = -min(P_A * P_B,\ P_a * P_b)$
Because $P_A * P_B = (P_{AB} + P_b) * (P_{AB} + P_a) > P_a * P_b$
We get $D_{min} = -(P_a * P_b)$
$D' = \frac{D}{D_{min}} = \frac{-P_a * P_b}{-P_a * P_b} = 1$
In this case, the maximum value of D' is 1.

Case III: $P_{Ab} = 0$ We would get a new HaploTypes table:

HaploTypes Table Originals

A Locus	Disease B	Locus b	Raw Total
A	P_{AB}	0	P_A
a	P_{aB}	P_{ab}	P_a
Col Total	P_B	P_b	Total = 1

HERE
$P_{AB} = P_A; \quad P_{ab} = P_b$
$D = P_{AB} - P_A * P_B = P_A - P_A * P_B = P_A * (1 - P_B)$
$D = P_A * P_b$

D value is positive, so we have:
$D_{min} = min(P_A * P_b,\ P_a * P_B)$
From the Haplotypes table, we could get:
$P_a = P_{aB} + P_{ab} = P_b + P_{aB}$
$P_B = P_{aB} + P_{AB} = P_A + P_{aB}$
So, we could get:
$P_a * P_B = (P_b + P_{aB}) * (P_A + P_{aB}) > P_A * P_b$
$D_{min} = P_A * P_b$
$D' = \frac{D}{D_{max}} = \frac{P_A * P_b}{P_A * P_b} = 1$
In this case, the maximum value of D' is 1.

Case IV: $P_{aB} = 0$ We would get a new HaploTypes table:

HaploTypes Table Originals

A Locus	Disease B	Locus b	Raw Total
A	P_{AB}	P_{Ab}	P_A
a	0	P_{ab}	P_a
Col Total	P_B	P_b	Total = 1

HERE
$P_{AB} = P_B; \quad P_{ab} = P_a$
$D = P_{AB} - P_A * P_B = P_B - P_A * P_B = P_B * (1 - P_A)$
$D = P_B * P_a$

D value is positive, so we have:
$D_{min} = min(P_A * P_b,\ P_a * P_B)$
From the Haplotypes table, we could get:
$P_b = P_{Ab} + P_{ab} = P_a + P_{AB}$
$P_A = P_{AB} + P_{Ab} = P_B + P_{Ab}$
So, we could get:
$P_A * P_b = (P_a + P_{AB}) * (P_B + P_{Ab}) > P_a * P_B$
$D_{min} = P_a * P_B$
$D' = \frac{D}{D_{max}} = \frac{P_a * P_B}{P_a * P_B} = 1$
In this case, the maximum value of D' is 1.