유전자-질병 관련성 및 유전체-환경 상호작용

연세대학교 보건대학원 교수 지선하

1. 유전역학에서 사용되는 기본 이론

가. Hardy-Weinberg Equilibrium

일반 인구집단에서 유전자-질병과의 관련성(association)을 연구할 때 가장 먼저 평가해야 하는 것이 유전자 분포의 하디-와인버그 평형이다. 유전자형의 하디-와인버그 평형이란, 인구집단에 외적요인이 작용하지 않는다면 유전자와 유전자형 빈도는 변하지 않고 평형을 이루게 됨을 의미한다. 일반적으로 관련성 연구에서 가장 많이 사용하는 연구설계인 환자-대조군 연구의 예를 들면 다음과 같다. 우선적으로 대조군에서 유전자의 하디-와인버그 평형을 이루는지를 파악하게 된다.

			Male gametes	
		Allele	A	a
	Allele	Freq.	p	q
Female gametes	A	p	AA p^2	Aa pq
	a	q	aA qp	aa q^2

▶ 대수적 설명 : 대립유전자 A의 대립유전자 빈도를 p라 하고 대립유전자 a의 대립유전자 빈도를 q라 하면, p + q = 1 이 된다. 하디-와인버그 평형상태에서는 AA 유전자형 빈도가 p^2, Aa의 빈도는 $2pq$, aa의 빈도는 q^2가 된다. 세 유전형의 빈도를 합하면 1이 된다 ($p^2+2pq+q^2=1$).

▶ 예) 사람의 정상적 피부와 눈 색깔을 나타내는 대립유전자인 A는 알비노증을 나타내는 a에 대해 우성이다. 즉, AA와 Aa인 사람은 정상색을 지니지만 aa를 지닌 사람은 알비노증이다. 알비노증은 20,000명에 한 명 꼴로 나타난다. 이러한 상황에서 대립유전자의 유전자 빈도를 추정하고자 한다. 하디-와인버그 평형을 가정한다면, aa의 빈도는 q^2이므로 $q^2=1/20000$ 이다. 따라서 $q=1/141$ 이고 우성대립유전자 A의 빈도는 $p=1-q=140/141$ 가 된다.

■ 하디-와인버그 법칙의 조건

① 교배는 무작위적으로 이뤄져야 한다. ② 돌연변이는 생기지 않는다. ③ 이입과 이출이 없다. ④ 대립유전자는 멘델의 제1법칙에 따라 분리되어야 한다. ⑤ 표본집단의 크기가 크다. ⑥ 개체군에는 선택이 작용하지 않는다.

■ Test Algorithm

H0 : HWE , $P(AA)=p^2$, $P(Aa)=2pq$, $P(aa)=q^2$

$$\sum \frac{(\text{관찰값}-\text{기대값})^2}{\text{기대값}} \sim X^2_{\text{자유도}}$$

자유도 = (유전형의 개수) - (추정할 모수의 개수) - 1

■ 예 제

	개 수		빈 도	
	관찰값	기대값	관찰값	기대값
AA	298	294.3063 ②	0.2980	0.2943 ①
Aa	489	496.3875	0.4890	0.4964
aa	213	209.3063	0.2130	0.2093
total	1000	1000	1	1

대립형질의 빈도(allele frequency)

p = (2×298 + 489)/(2×1000) = 0.5425

q = (489 + 2×213)/(2×1000) = 0.4575

유전형의 빈도(gene frequency)

p(AA) = $p^2=(0.5425)^2=0.2943$ ------ ①

p(Aa) = $2pq=2\times(0.5425)\times(0.4575)=0.4964$

p(aa) = $q^2=(0.4575)^2=0.2093$

기대값(expected freqeuncy)

AA = p(AA)×1000 = 294.3064 ------ ②

Aa = p(Aa)×1000 = 496.3875

aa = p(aa)×1000 = 209.3063

⇒ 검정통계량

$$\chi^2=\frac{(298-294.3063)^2}{294.3063}+\frac{(489-496.3875)^2}{496.3875}+\frac{(213-209.3063)^2}{209.3063}=0.2215$$

자유도 =3 - 1 - 1 = 1

∴ 자유도 1인 카이제곱분포에 근거한 p값이 0.6379 이므로 관찰된 값은 HWE 상태라고 할 수 있다.

나. Linkage Disequilibrium

유전자들이 독립적으로 배합되어 있으면 연쇄평형(linkage equilibrium, LE) 상태라 하고, 그렇지 못한 경우를 연쇄비평형(linkage disequilibrium, LD)이라 한다. 연쇄평형 상태에서는 서로 다른 locus에 있는 유전자의 대립형질(allele)은 서로 독립적으로 타나나게 되고 따라서 haplotype의 빈도는 각 대립형질 빈도의 곱이 된다. 이 때, 유전자들은 HWE 임을 가정한다.

▶ 연쇄평형 상태에서 두 유전자 A와 B의 haplotype 빈도는 아래 표과 같이 나타난다.

			Alleles of A gene	
		Allele	A	a
	Allele	Freq.	p_1	p_2
Alleles of B gene	B	q_1	AB p_1q_1	aB p_2q_1
	b	q_2	Ab p_1q_2	ab p_2q_2

▶ 유전자 A가 대립형질 {A, a}로 구성되고 유전자 B가 대립형질 {B, b}로 구성되어있을 때, 대립형질 a의 빈도를 0.09, 대립형질 b의 빈도를 0.12라 한다. 연쇄평형상태라면 대립형질 a와 b를 동시에 가지고 있을 경우는 $0.09\times0.12=0.0108$ 가 된다. 그러나 실제 인구집단을 조사한 결과 대립형질 a와 b를 동시에 가지고 있는 경우가 0.07 로 높게 나타났다고 한다면, LD상태라고 할 수 있다.

■ Test Algorithm

(1) 인구집단을 대상으로 locus 1과 2에 대해 조사하면 다음과 같은 표로 정리할 수 있다. 두 유전자가 모두 hetero 인 경우, 4종류의 haplotype이 생성되게 된다.

즉, k5= $k5_{AB}+k5_{aB}+k5_{Ab}+k5_{ab}$.

Locus	Locus 1		
2	AA	Aa	aa
BB	k1	k2	k3
Bb	k4	k5	k6
bb	k7	k8	k9

(2) 관찰된 haplotype에 대한 빈도는 다음과 같이 정리된다.

	A	a
B	$2k1+k2+k4+k5_{AB}$	$k2+2k3+k6+k5_{aB}$
b	$k4+k6+2k7+k5_{Ab}$	$k6+k8+2k9+k5_{ab}$

(3) $k5_{AB}$, $k5_{aB}$, $k5_{Ab}$, $k5_{ab}$ 의 빈도를 안다면, LD에 대한 검정은 카이제곱분포에 근거하여 수행된다. 여기서 D는 결합척도를 나타낸다.

H0 : Linkage Equilibrium : $P(AB)-P(A)P(B)=D=0$

$$\sum \frac{(\text{관측값}-\text{기대값})^2}{\text{기대값}} \sim X^2_{\text{자유도}}$$

자유도 = (haplotype 개수) - (추정되는 모수의 개수) - 1

다. Haplotype의 의미와 필요성

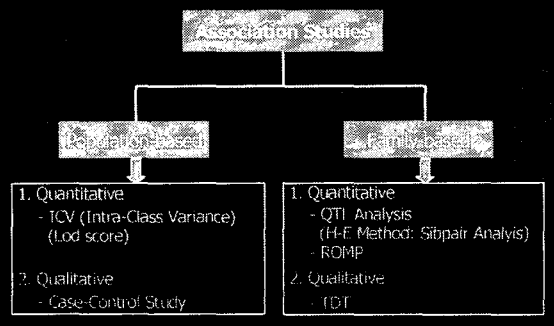
SNP(sing-nucleotide polymorphism)는 인간게놈에서 250~350bp가 발견되며 질병과 관련된 유전자의 돌연변이가 적기 때문에 유전자분석에 매우 유용하다. 최근에는 복합형질유전자(complex-trait gene)연구나 약물반응에 영향을 주는 유전자연구에서 각광받고 있다. 그러나 대량의 SNP자료는 유전자분석에 어려움을 준다. 단일 SNP가 지닌 정보가 상대적으로 매우 적고, 여러개의 SNP가 밀집되어 있는 유전자(gene)의 경우 단일 SNP는 나머지 SNP들과의 LD(linkage disequilibrium)정보가 무시되기 때문이다. 그러므로 인접한 SNP들의 정보를 결합한 haplotype을 이용한 유전자분석이 요구된다. 최근 발표된 논문에서도 haplotype이 SNP에 비해 정보량이 많고 LD정보도 포함하고 있어 haplotype에 의한 유전자 분석이 더욱 강력하고 로버스트함을 확인할 수 있다(Akey et al., 2001; Daly et al., 2001; Pritchard 2001).

■ Haplotype의 추정

SNP들에 대한 정보가 주어졌을 때, haplotype형태를 추정하는 방법은 크게 4가지로 Clark 알고리즘, EM(expectation maximization) 알고리즘, 유사베이즈(pseudo Bayesian) 알고리즘과 새로운 몬테카를로 접근방법(new MC)이 있다. Clark 알고리즘은 haplotype의 종류가 적을 때 가능한 방법으로 형태(phase)가 분명한 개인부터 시작하여 haplotype의 목록을 채워나간다. EM 알고리즘은 반복을 통해 haplotype의 빈도값을 추정하는 방법으로 다수의 SNP자료에 적합하다. 유사베이즈 알고리즘은 반복적인 확률표본 방법인 pseudo Gibbs sampler(PGS)를 수행한다. PGS는 확률탐색과 반복법이 혼합되어 앞서 두 알고리즘보다 효율적이다(Stephens et al., 2001b).

2. Association 연구

Association Studies



Contrasts between linkage analysis and association studies

Linkage analysis

- Based on concepts of cross-over
- Require family data to assess recombination in observed data
- Need large sample size for diseases with modest general effects

Source: Risch et al, Science, 1996

Contrasts between linkage analysis and association studies

Association studies

- Based on concepts of linkage disequilibrium
- Exploit consequences of recombination that has occurred between a mutation and a marker some generations ago
- Work on unrelated individuals
- More powerful statistically for modest genetic effects

Source: Risch et al, Science, 1996

Association Studies

Case-control studies based on the comparison of allele frequency of candidate genes from affected individuals (cases) and unaffected individuals (controls)

Candidate genes:

- Genes known to have "close biological relation" to the disease ("putative susceptibility genes")

Implications of possible association

The targeted allele is a cause of the diseases, or

The targeted allele is in linkage disequilibrium with the disease locus, or

Artifact of population admixture

"Population Stratification"

Nothing to do for disease

Some possible modifications

Perform in populations that are relatively "homogenous"

Use Unaffected sibling as controls

Use "internal" controls

- case-parental design
- Trio design (TDT)

3. 유전-환경 상호작용 연구

G X E Interaction

Designs for detecting GxE interaction

Interaction is deviation from the expected combined effects of genes (G) and environmental (E) risk factors

G X E Interaction

Different concepts for interaction

Biological (Ottman, 1996)

- G alone increases risk & E increases that
- G alone has no effect, but E changes that
- E alone has no effect, but G changes that
- Together G & E increases risk but not alone
- Individually G & E increase risk, but together it is far worse

Statistical

- Any deviation from predicted effects of G & E

G X E Interaction

Gene x Environmental Interaction

	Genotype -	Genotype +
Exposure -	I_{00}	I_{01}
Exposure +	I_{10}	I_{11}

I = Incidence of disease

G X E Interaction

Testing for G x E Interaction

Express I_{11} as a function of the other rates & adjust for baseline incidence I_{00}

Additive Model

- $I_{00} = I_{00} + I_{00} - I_{00}$
- $OR_{11} = OR_{10} + OR_{01} - 1$

Test Ho:

- OR int = $OR_{11}/(OR_{10} + OR_{01} - 1) = 1$

G X E Interaction

Testing for G x E Interaction

Multiplicative

- $I_{11} = I_{10} * I_{01}$
- $OR_{11} = OR_{10} * OR_{01}$

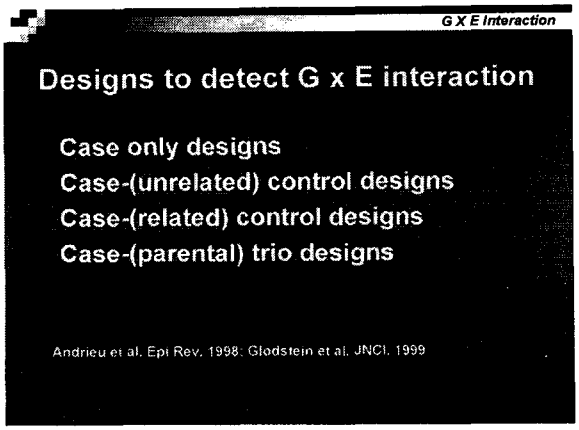
Test Ho:

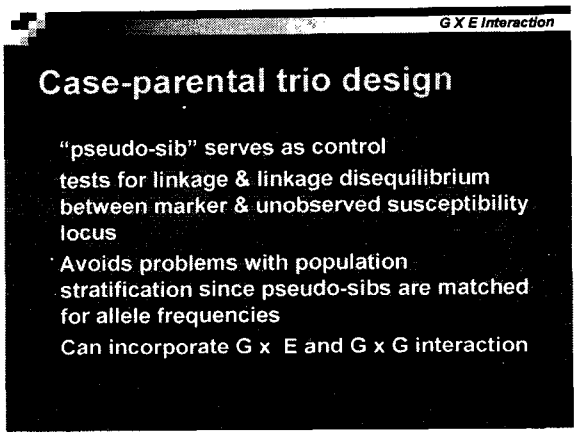
- OR int = $OR_{11}/(OR_{10} * OR_{01}) = 1$

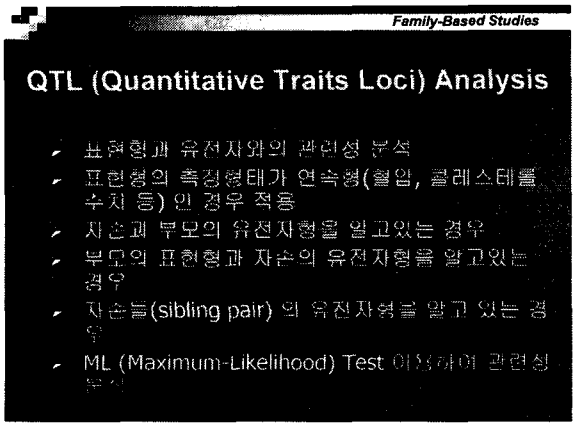
G X E Interaction

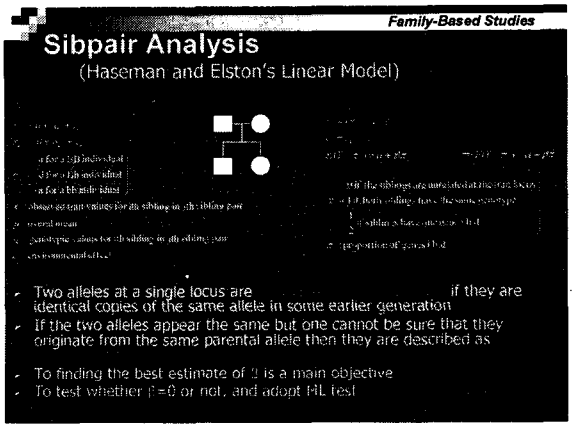
G x E in case-control design

Genotype	Exposure	Case	Control	OR	Information
Yes	Yes	a	b	ah/bg	Joint effect of G x E
Yes	No	c	d	ch/dg	Effect of G alone
No	Yes	e	f	eh/fg	Effect of E alone
No	No	g	h	1	Reference









■ 참고 문헌

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291-300

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-122

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232

Kwok PY, Gu Z (1999) Single nucleotide polymorphism libraries: why and how are we building them? Mol Med Today 5(12):538-43

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137