Disease association and haplotypes

Previously, case-control and family-based approaches for association analysis of a single marker locus with the disease were presented. How to analyze a set of closely linked marker loci?

Possibilities:

1. single marker analysis for each marker
2. haplotype analysis

Problem with haplotype analysis:

Current high-throughput genotyping technologies provide single marker genotypes, but not the pair of haplotypes for an individual ($\rightarrow$ problem of unknown phase). Therefore, statistical methods are required for the determination of phase.

Haplotype estimation: unrelated individuals

Consider a sample of n unrelated individuals and let G_i be the unphased multi-locus genotype of the i-th individual. Let $j \mid k = G_i$ if the haplotypes j and k are compatible with G_i. Let $C_{G_i} = \{(j,k) : j \mid k = G_i\}$ be the set of all pairs of haplotypes being compatible with G_i. Then, assuming HWE for haplotypes, the likelihood of the sample is given by

$$L(h_1, \ldots, h_s \mid G_1, \ldots G_n) = \prod_{i=1}^{n} \sum_{(j,k) \in C_{G_i}} h_j \cdot h_k,$$

where h_j and h_k denote the frequencies of haplotypes j and k.

Haplotype estimation: unrelated individuals

For a sequence of haplotypes $j_1, \ldots, j_m$, let $z^r_{j_1,\ldots,j_m}$ denote the number of occurrences of haplotype r in this sequence, i.e.,

$$z^r_{j_1,\ldots,j_m} = \mid \{s : j_s = r\} \mid .$$

The expectation maximization (EM) recursion is

$$h_r^{(t+1)} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{(j,k) \in C_{G_i}} z^r_{j,k} \cdot \frac{h_j^{(t)} \cdot h_k^{(t)}}{\sum_{(j',k') \in C_{G_i}} h_{j'}^{(t)} \cdot h_{k'}^{(t)}}$$

Haplotype association: unrelated individuals

Having obtained the MLEs $\widehat{h}_r$, the conditional probability $w_i^{(j,k)}$ that the i-th individual possesses the (phased) multi-locus genotype (j,k) is estimated by

$$\widehat{w}_i^{(j,k)} = \begin{cases} \dfrac{\widehat{h}_j \cdot \widehat{h}_k}{\sum\limits_{(j',k') \in C_{G_i}} \widehat{h}_{j'} \cdot \widehat{h}_{k'}} & \text{if} \quad (j,k) \in C_{G_i} \\ 0 & \text{if} \quad (j,k) \notin C_{G_i} \end{cases}$$

With $\widehat{w}_i^{(j,k)}$, a table of haplotype counts in cases and controls similar to CC/9 is constructed and the χ^2 test statistic for $2 \times s$ tables is calculated.

Haplotype association: unrelated individuals

Example: Assume two diallelic loci $\{A, a\}$ and $\{B, b\}$, i.e., four haplotypes:

$AB(=1)$, $Ab(=2)$, $aB(=3)$, $ab(=4)$.

Further assume that an affected individual possesses genotype Aa at the first locus and genotype Bb at the second locus. Thus, there are two possibilities of (phased) multi-locus genotypes for this individual:

$AB/ab(=1/4)$ and $Ab/aB(=2/3)$.

Now assume that on the basis of estimated haplotype frequencies, $\widehat{w}_i^{(1,4)} = 0.4$ and $\widehat{w}_i^{(2,3)} = 0.6$. Then, the contribution of this individual to the 2×4 table is as follows:

Group	Haplotype			
	$1(AB)$	$2(Ab)$	$3(aB)$	$4(ab)$
Cases	0.4	0.6	0.6	0.4
Controls				

Haplotype association: unrelated individuals

To assess the significance of the χ^2 test statistic obtained from the $2 \times t$ table, it is mandatory to apply an appropriate simulation procedure: In each replicate of this simulation, a sample is constructed in which the case/control status of each individual is randomly permuted, under the restriction that the number of cases as well as the number of controls remains unchanged. The P value assigned to the χ^2 statistic calculated for the real data is the fraction of simulation replicates resulting in a greater or equal test statistic (c.f. FBA/14).

Haplotype estimation: case-parent triads

Consider a sample of n case-parent triads. Let G_i^f, G_i^m, and G_i^c denote the unphased multi-locus genotype of the father, mother, and child in the i-th case-parent triad. An (ordered) quadruple (j, k, u, v) of haplotypes is called a haplotype explanation and is said to be compatible with the genotype configuration $G_i = (G_i^f, G_i^m, G_i^c)$ if $j \mid k = G_i^f$, $u \mid v = G_i^m$, and $j \mid u = G_i^c$. Thus, the set C_{G_i} of haplotype explanations which are compatible with G_i is

$$C_{G_i} = \{(j, k, u, v) : j \mid k = G_i^f, u \mid v = G_i^m, j \mid u = G_i^c\}.$$

Then, assuming HWE for haplotypes, the likelihood of the sample is given by

$$L(h_1, \ldots, h_s \mid G_1, \ldots G_n) = \prod_{i=1}^{n} \sum_{(j,k,u,v) \in C_{G_i}} h_j \cdot h_k \cdot h_u \cdot h_v.$$

Haplotype estimation: case-parent triads

The expectation maximization (EM) recursion is

$$h_r^{(t+1)} = \frac{1}{4n} \sum_{i=1}^{n} \sum_{(j,k,u,v) \in C_{G_i}} z_{j,k,u,v}^{r} \cdot \frac{h_j^{(t)} \cdot h_k^{(t)} \cdot h_u^{(t)} \cdot h_v^{(t)}}{\sum_{(j',k',u',v') \in C_{G_i}} h_{j'}^{(t)} \cdot h_{k'}^{(t)} \cdot h_{u'}^{(t)} \cdot h_{v'}^{(t)}}$$

With only slight modifications, MLEs of haplotype frequencies can also be obtained from samples of general nuclear families (i.e., arbitrary number of children).

Haplotype association: case-parent triads

Testing for association:

- Estimate haplotype frequencies and calculate weights $w_i^{(j,k,u,v)}$ for each haplotype explanation for the i-th family (similar to PC/4).
- Use these weights to construct the table of transmitted/non-transmitted haplotypes (c.f. FBA/7).
- Calculate the $\mathrm{TDT_{SE}}$ statistic (c.f. FBA/8) from this table of transmitted/non-transmitted haplotypes.
- Apply the simulation procedure described on FBA/13 to obtain the P value.

Power calculation

Power calculations are important for planning a linkage or association study.

The power of a study is the probability to detect as statistically significant a real effect of a given magnitude. Thus, power calculation requires to specify

- sample size
- kind and magnitude of the effect

Example:

Assume a sample of 400 affected sib pairs is available to test a marker locus for its linkage with the disease. What is the probability that this study results in a significant linkage result in case that in reality the marker locus is linked at $\theta = 0.1$ to a diallelic disease locus $\{D, d\}$, the disease allele frequency is $P(D) = 0.1$ and the genotype specific relative risks at the disease locus are $RR_{DD} = RR_{Dd} = 4$?

Power calculation

Additional specifications required for power calculations:

- marker characteristics (number of alleles and their frequencies)
- statistical test used to decide on linkage/association
- type I error rate

Methods for power calculation:

- Exact calculation
- Approximate calculation
- Simulation (by computer)

Power calculation: Exact calculation

Example:

- 400 affected sib pairs
- diallelic disease locus $\{D, d\}$ with $P(D) = 0.1$, $RR_{DD} = RR_{Dd} = 4$
- completely informative marker
- recombination fraction $\theta = 0.1$ between marker locus and disease locus
- statistical test: NPL score (c.f. NPL/24), i.e., $H_0 : \theta = 0.5$ is rejected for $n_2 > n_0 + \sqrt{n/2} \cdot u_{1-\alpha}$
- type I error rate $\alpha = 0.0001$

Under the additional assumption of a single locus disease model, these specifications imply that the distribution of IBD scores at the marker locus is $(z_2^M, z_1^M, z_0^M) = (0.317, 0.498, 0.185)$, c.f. NPL/6 and NPL/8.

Power calculation: Exact calculation

In this example, power can be obtained by summing up the probabilities of all samples that lead to the rejection of the null hypothesis, i.e.,

$$\text{Power} = \sum_{\{(n_0,n_1,n_2):n_2>n_0+\sqrt{n/2}\cdot u_{1-\alpha}\}} n! \cdot \prod_{i=0}^{2} \frac{(z_i^M)^{n_i}}{n_i!}$$

$$= \sum_{n_0=0}^{n} \binom{n}{n_0} \cdot (z_0^M)^{n_0} \cdot (1 - z_0^M)^{n-n_0}$$

$$\cdot \sum_{n_2>n_0+\sqrt{n/2}\cdot u_{1-\alpha}} \binom{n-n_0}{n_2} \cdot \left(\frac{z_2^M}{1-z_0^M}\right)^{n_2} \cdot \left(1 - \frac{z_2^M}{1-z_0^M}\right)^{n-n_0-n_2}$$

Example: Power=0.52

Power calculation: Approximate calculation

Example:

- 700 case-parents triads
- diallelic disease locus $\{D, d\}$ with $P(D) = 0.1$, $RR_{DD} = RR_{Dd} = 2$
- diallelic marker locus $\{A, a\}$ being in perfect linkage disequilibrium with the disease locus
- test: TDT
- type I error rate $\alpha = 10^{-7}$

Power calculation: Approximate calculation

Type	Parent 1	Parent 2	Child
1	AA	Aa	AA
	aa	Aa	Aa
2	AA	Aa	Aa
	aa	Aa	aa
3	Aa	Aa	AA
4	Aa	Aa	Aa
5	Aa	Aa	aa
6	other (uninformative families)		

Number of different samples is $\binom{n+5}{5}$, which for $n = 700$ is approximately

$1.4 \cdot 10^{12}$

$\Rightarrow$ Exact calculation is not feasible

Power calculation: Approximate calculation

However, the distribution of the TDT under the alternative can be approximated by the square of a normal distribution. The expectation and variance of this approximating normal distribution depend on the alternative but can easily calculated numerically. A power approximation is then obtained by calculating the tail probability of this distribution.

Example: Power=0.80

Power calculation: Simulation

Example:

A sample of pedigrees with the disease has been collected. The disease phenotype of all family members are already available. Prior to undertaking the typing of marker genotypes, it should be decided whether the collected pedigrees provide sufficient information to demonstrate linkage by parametric linkage analysis. Traits and marker characteristics as well as the recombination fraction between the marker and the disease are specified.

How to simulate the marker genotypes?

Power calculation: Simulation

Gene dropping:

1. Founder (marker and disease) genotypes are first simulated according to population frequencies.
2. The genes are “dropped” down the pedigree according Mendel’s laws.
3. Disease phenotypes are simulated from the disease genotypes according to the penetrances.
4. Simulated phenotypes are compared to the observed phenotypes. Simulations inconsistent with the observed phenotypes are rejected.

Drawback: Very inefficient for medium to large pedigrees

Better approach: Sample genotypes from the conditional distribution of the genotypes given disease phenotypes (SLINK, SIMLINK)