

Tutorial T2

Statistical Analysis of Haplotype-Disease Associations Using HAPSTAT

DANYU LIN

Department of Biostatistics

University of North Carolina

email: lin@bios.unc.edu

website: <http://www.bios.unc.edu/~lin>

ENAR Spring Meeting

Atlanta, March 12, 2007

PART I
METHODOLOGY

INTRODUCTION

Complex Diseases and Association Studies

DNA: linear sequence of 4 nucleotides (A, T, C, G)

Single Nucleotide Polymorphism (SNP): change of a single nucleotide in the DNA sequence

- millions
- 90% of all human genetic variation
- major impact on disease susceptibility
- biallelic

Human Genome Sequencing and HapMap Projects

Catalogue of SNPs and Genotyping Technologies

SNPs-Based Association Studies

Studies of Unrelated Individuals vs Family-Based Studies

Candidate-Gene vs Genomewide Studies



Single SNP Analysis: $G \in \{0, 1, 2\}$

Haplotype: specific combination of nucleotides at a series of (nearby) SNPs on the same chromosome

Why Haplotypes?

1. full information content
2. efficiency gain
 - linkage disequilibrium information from multiple markers
 - unmeasured causal variants
 - interaction of multiple variants on the same chromosome
3. data reduction
 - actual haplotypes \ll possible haplotypes
4. missing genotype data

(Unphased) Genotype: combination of 2 homologous haplotypes

Statistical Problem: sum of 2 ordered sequences of 0's and 1's

Haplotype Frequencies and Haplotype Assignments: Clark ('90); Excoffier & Slatkin ('95); Chiano & Clayton ('98); Stephens et al ('01); Niu et al ('02); Qin et al ('02); Marchini et al ('06)

Haplotype-Disease Association

Imputation Method (Lin & Huang '07)

- biased estimation of genetic effects
- reduced statistical power
- inflated type I error

“Phasing cases and controls together can lead to a bias towards the null hypothesis of no association and therefore a loss of power.

Conversely, phasing cases and controls separately can inflate type I error rates.” (Balding '06)

Proper Methods: Schaid et al ('02); Zhao et al ('03); Epstein & Satten ('03); Stram et al ('03); Lake et al ('03); Lin ('04); Spinka et al ('05); Lin & Zeng ('06); Lin et al ('05); Zeng et al ('06)

PRELIMINARIES

Notation

$M = \#$ SNPs

$K = 2^M$

- actual haplotypes \ll possible haplotypes

$h =$ haplotype = unique sequence of M numbers from $\{0, 1\}$

$h_k = k$ th possible haplotype ($k = 1, \dots, K$)

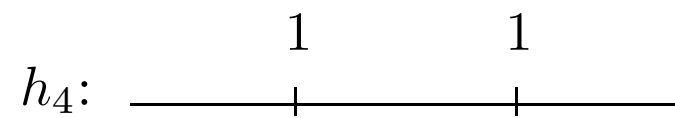
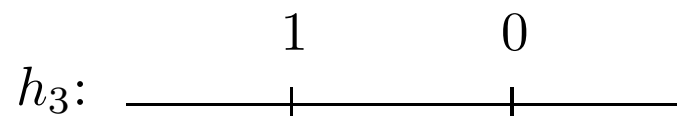
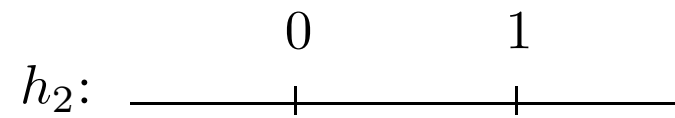
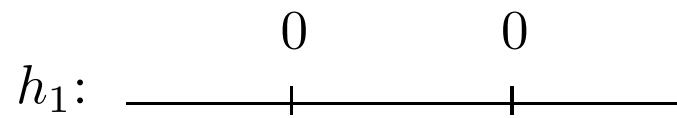
$H =$ haplotype pair

$G =$ genotype = ordered sequence of M numbers from $\{0, 1, 2\}$

- $H = (h_k, h_l) \implies G = h_k + h_l$
- $G \xrightarrow{\text{not}} H$ if heterozygous at > 1 SNP

$Y =$ phenotype

$X =$ covariates (environmental variables)



$$G = (2, 1) \implies H = (h_3, h_4)$$

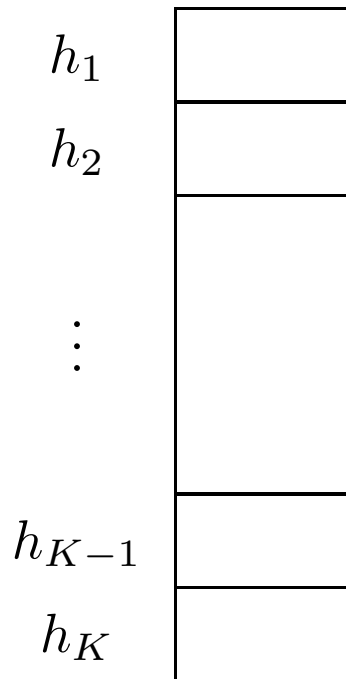
$$G = (1, 1) \implies H = (h_1, h_4) \text{ or } (h_2, h_3)$$

Association Models

$$P(Y|X, H; \theta)$$

Separate models: Compare a target haplotype h^* to all other haplotypes

Overall model: Compare multiple haplotypes to a reference simultaneously



Mode of inheritance:

Additive: Having two copies of the causal haplotype doubles the effect on the trait as compared to having one copy.

Dominant: Having one or two copies of the causal haplotype has the same effect on the trait.

Recessive: Only having two copies of the causal haplotype has an effect on the trait.

Co-dominant: Having two copies of the causal haplotype can have an arbitrarily different effect on the trait than having one copy.

Separate logistic regression models:

$$\begin{aligned} & \text{logit}P\{Y = 1|H = (h_k, h_l); \theta\} \\ = & \begin{cases} \alpha + \beta I(h_k = h_l = h^*) & \text{(recessive)} \\ \alpha + \beta\{I(h_k = h^*) + I(h_l = h^*) - I(h_k = h_l = h^*)\} & \text{(dominant)} \\ \alpha + \beta\{I(h_k = h^*) + I(h_l = h^*)\} & \text{(additive)} \\ \alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*) & \text{(codominant)} \end{cases} \end{aligned}$$

$$\begin{aligned} & \text{logit}P\{Y = 1|X = x, H = (h_k, h_l); \theta\} \\ & = \alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*) \\ & \quad + \beta'_3 x + \beta'_4\{I(h_k = h^*) + I(h_l = h^*)\}x + \beta'_5 I(h_k = h_l = h^*)x \end{aligned}$$

- h^* = target haplotype of interest
- $I(\mathcal{A})$ indicator function for event \mathcal{A}

Distribution of Halplotypes

$$\pi_{kl} = P\{H = (h_k, h_l)\} \quad (k, l = 1, \dots, K)$$

$$\pi_k = P(h = h_k) \quad (k = 1, \dots, K)$$

Hardy-Weinberg equilibrium:

$$\pi_{kl} = \pi_k \pi_l \quad (k, l = 1, \dots, K)$$

Hardy-Weinberg disequilibrium:

$$\pi_{kl} = \begin{cases} \pi_k^2 + \rho \pi_k (1 - \pi_k) \\ (1 - \rho) \pi_k \pi_l \quad (k \neq l) \end{cases}$$

- $\rho =$ inbreeding coefficient
- $\rho = 0 \Rightarrow$ HWE
- $\rho > 0 \Rightarrow$ excess homozygosity
- $\rho < 0 \Rightarrow$ excess heterozygosity

Distribution of (Y, X, G)

$$P(Y, X, G)$$

$$= \sum_H P(Y, X, G, H)$$

$$= \sum_H P(Y|X, G, H)P(X|G, H)P(G|H)P(H)$$

$$= \sum_{H \in \mathcal{S}(G)} \{P(Y|X, H)P(H)\}P(X|G)$$

- $P(X|G, H) = P(X|G)$
- $\mathcal{S}(G) =$ set of H 's consistent with G

CROSS-SECTIONAL STUDIES

Data: (Y_i, X_i, G_i) ($i = 1, \dots, n$)

- Y can be discrete or continuous, univariate or multivariate

Model: $P(Y|X, H; \theta)$

- generalized linear models
- generalized linear mixed models

Likelihood:

$$\prod_{i=1}^n P(Y_i, X_i, G_i) \propto \prod_{i=1}^n \sum_{H \in \mathcal{S}(G_i)} P(Y_i|X_i, H; \theta) P(H; \gamma)$$

Computation: EM

Asymptotics: Consistency, normality, efficiency

Inference: Wald statistics, likelihood ratio statistics

CASE-CONTROL STUDIES

Data: $(X_i, G_i|Y_i)$ ($i = 1, \dots, n$)

Model: $P(Y|X, H; \theta)$

- logistic, probit, complementary log-log
- proportional odds, multivariate logistic, multivariate probit

Likelihood:

$$\begin{aligned} \prod_{i=1}^n P(X_i, G_i|Y_i) &= \prod_{i=1}^n \frac{P(Y_i, X_i, G_i)}{P(Y_i)} \\ &= \prod_{i=1}^n \frac{\sum_{H \in \mathcal{S}(G_i)} P(Y_i|X_i, H; \theta) P(H; \gamma) P(X_i|G_i)}{\sum_{x,g} \sum_{H \in \mathcal{S}(g)} P(Y_i|x, H; \theta) P(H; \gamma) P(X = x|G = g)} \end{aligned}$$

- infinite-dimensional nuisance parameter
- rare disease

Computation: Profile likelihood

- profile over a small number of parameters
- EM

Asymptotics: Consistency, normality, efficiency

- infinite-dimensional nuisance parameters

Inference: Wald statistics, likelihood ratio statistics

COHORT STUDIES

Data: $(Y_i, \Delta_i, \bar{X}_i(Y_i), G_i) \quad (i = 1, \dots, n)$

- T = time to disease
- C = censoring time
- $Y = \min(T, C)$ (observation time)
- $\Delta = I(T \leq C)$ (disease status)
- $\bar{X}(t)$ = covariate history by t

Model:

$$\lambda(t|H, X(t)) = \lambda_0(t)e^{\beta' Z(H, X(t))} \quad (\text{proportional hazards model})$$

- $Z(H, X(t))$ = specific function of H and $X(t)$

Likelihood

$$\prod_{i=1}^n \sum_{H \in \mathcal{S}(G_i)} \lambda(Y_i | H, X_i(Y_i))^{\Delta_i} \exp \left\{ - \int_0^{Y_i} \lambda(t | H, X_i(t)) dt \right\} P(H)$$

- nonparametric likelihood
- case-cohort and nested case-control designs

Computation:

- piece-wise constant $\lambda_0(\cdot)$
- EM

Asymptotics: Consistency, normality, efficiency

Variance Estimation: profile likelihood

Inference: Wald statistics, likelihood ratio statistics

CAROLINA BREAST CANCER STUDY

Population-based case-control study: 2311 cases, 2022 controls

3 SNPs in the XRCC1 gene

- codon 194: C→T
- codon 280: G→A
- codon 399: G→A
- 10% missing
- G_i missing $\implies \mathcal{S}(G_i)$ enlarged

Haplotype frequencies:

$$(CGG, CGA, CAG, TGG) = (0.62, 0.27, 0.07, 0.04)$$

Inbreeding coefficient = 0.04

Logistic regression: haplotype, age, race, smoking duration

Variable	Recessive	Dominant	Additive	Codominant model	
	Model	Model	Model	Additive	Recessive
Haplotype CGA	.34 (.17)	.20 (.11)	.20 (.08)	.17 (.11)	.08 (.24)
Dur1	.11 (.09)	.22 (.10)	.22 (.10)	.22 (.10)	—
Dur2	.01 (.11)	.01 (.13)	.01 (.13)	.01 (.13)	—
Dur3	.07 (.11)	.18 (.14)	.16 (.13)	.18 (.14)	—
Dur4	.32 (.09)	.37 (.11)	.37 (.11)	.37 (.11)	—
Hap*Dur1	-.36 (.22)	-.30 (.13)	-.25 (.10)	-.26 (.13)	.05 (.30)
Hap*Dur2	-.10 (.27)	-.02 (.16)	-.03 (.12)	.01 (.17)	-.10 (.38)
Hap*Dur3	-.10 (.28)	-.26 (.17)	-.17 (.12)	-.27 (.18)	.32 (.40)
Hap*Dur4	-.19 (.23)	-.15 (.14)	-.13 (.10)	-.13 (.14)	.01 (.32)
Race	.38 (.06)	.38 (.06)	.38 (.06)	.38 (.06)	.38 (.06)
Age	.04 (.01)	.04 (.01)	.04 (.01)	.04 (.01)	.04 (.01)

ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC) STUDY

Bi-ethnic cohort of 15,792 men and women aged 45-64 years

Etiology of atherosclerosis and other diseases

Common genetic polymorphisms and smoking

19 genes, ~ 6 SNPs/gene

Incident coronary heart disease

Case-cohort sampling

- stratified by age, gender and race
- 1008 cases, 927 controls

Haplotype	Frequency	Parameter	Estimate	St error	<i>p</i> -value
00110	.16	Main effect	.237	.105	.024
		Interaction	-.010	.119	.931
01001	.10	Main effect	-.295	.239	.218
		Interaction	.003	.273	.992
01100	.06	Main effect	.124	.243	.610
		Interaction	-.404	.276	.143
10110	.23	Main effect	-.078	.143	.585
		Interaction	.102	.171	.551
11001	.28	Main effect	.165	.146	.259
		Interaction	.048	.177	.786
11100	.15	Main effect	.029	.166	.863
		Interaction	-.136	.188	.469

Parameter	Estimate	St error	<i>p</i> -value
Haplotype 00110	.473	.288	.101
Haplotype 01100	.427	.377	.257
Haplotype 10110	.189	.255	.459
Haplotype 11001	.196	.409	.632
Haplotype 11100	.295	.338	.382
Smoking status	.703	.685	.305
00110× Smoking	-.097	.364	.790
01001× Smoking	-.468	.461	.310
01100× Smoking	-.179	.705	.800
10110× Smoking	-.098	.519	.851
11100× Smoking	-.178	.418	.670

REMARKS

Software: www.bios.unc.edu/~lin/hapstat

Other software:

- haplo.stats
- Chaplin

Model selection

- Akaike information criterion (AIC)

Missing SNP data

Multiple genes

Genomewide studies (Huang, Amos & Lin '07)

Selective genotyping for quantitative traits (Huang & Lin '07)

Family studies (Diao & Lin '07)

REFERENCES

Lin DY, Zeng D, Millikan R (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genet Epid* 29:299-312.

Lin DY, Zeng D (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *JASA* 101:89-118.

Zeng D, Lin DY, Avery CL, North KE, Bray MS (2006). Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics* 7:489-502.