Haplotype-based Association Study Between ACE gene and Hypertension

Jinheum Kim[†,1], Chung Mo Nam[2], Dae Ryong Kang[2], Il Suh[2], Yun Kyung Lee[3]

[†]jinhkim@suwon.ac.kr

[1]Department of Applied Statistics, University of Suwon

[2]Department of Preventive Medicine and Public Health, Yonsei University College of Medicine

[3]Graduate School of Public Health, Yonsei University

Contents

- Terminology
- Single-locus-based association & linkage study
- Haplotype-based association & linkage study
- Case study
- Simulation studies
- Summary

Terminology

- Gene *DNA segment that codes for a functional unit*
- Locus *location of a gene on a chromosome*
- Allele *different forms of a gene that occupy the same locus*
- Linkage *closeness between loci on a chromosome*
- Linkage disequilibrium *non-random association of alleles*
- Polymorphism *difference in DNA sequence among individuals*
- SNPs (Single Nucleotide Polymorphisms)

Single-locus-based approaches

- Population-based approach
  - Compare marker-allele distribution in affected and unaffected individuals
  - Always do NOT provide linkage between the marker and a susceptibility gene for the disease due to confounding factors such as *population stratification or population admixture*

Single-locus-based approaches

- Family-based approach
  - Parental genetic information is required
  - Compare # of parents who *either* do *or* transmit the associated marker allele to affected children
  - Immune to population stratification due to within-family comparisons

Prob. structure *of* TDT table

- Given a marker with M_1, M_2 and a disease gene with D_1, D_2,

		Non-transmitted allele	
		M_1	M_2
Transmitted	M_1	$q^2 + q\delta/p$	$q(1-q) + (1-\theta-q)\delta/p$
allele	M_2	$q(1-q) + (\theta-q)\delta/p$	$(1-q)^2 - (1-q)\delta/p$

where

$$
\begin{aligned}
p &= P(D_1), \quad q = P(M_1), \\
\delta &= P(M_1 D_1) - P(M_1)P(D_1), \\
\theta &= \text{recombination fraction between marker and disease loci}
\end{aligned}
$$

- Detect simultaneously linkage and association
- Only heterozygous parents are informative

Why haplotype-based? But . . .

- Haplotype *set of alleles on a chromosome*
- Many markers are genotyped within a very short physical distance
- More informative

- Haplotype information is not usually available from genotype information
- *Eg.* when # of heterozygous loci=c, # of possible haplotype pairs=2^{c-1}

Hypothetical family trios

- Two samples

Sample	Father ($\mathcal{F}$)	Mother ($\mathcal{M}$)	Child ($\mathcal{C}$)
1	12/12/11	22/12/22	12/12/12
2	12/22/11	22/12/22	12/22/12

Deducing haplotypes

- Probable haplotype pairs

Sample		$\mathcal{F}$	$\mathcal{M}$	$\mathcal{C}$
1		{111,221}	{222,212}	{111,222}
	or	{121,211}	{212,222}	{121,212}
2		{121,221}	{212,222}	{121,212}

- In Sample 1, $\mathcal{F}$'s haplotype uncertainty exists!
- In Sample 2, haplotypes of parents are deducible!

Previous researches

- Lazzeroni & Lange (1998, Hum Hered): multiple test
- Wilson (1997, AHG), Clayton & Jones (1999, AJHG): Discard families with ambiguous haplotypes
- Clayton (1999, AJHG): Likelihood-based but *not* robust to population admixture
- Zhao *et al.* (2000, AJHG)
  - How to resolve haplotype ambiguity? *Allocate a conditional probability to each of haplotype group corresponding to a set of genotypes*

Zhao *et al.* (2000, AJHG)

- Explicit formula
  For each g, estimate the number of families : $\mathcal{F}$ with $\{H_i, H_j\}$ transmits H_i and $\mathcal{M}$ with $\{H_k, H_l\}$ transmits H_k

$$\hat{t}_g^{ik,jl} = n_g \times \frac{h_i h_j h_k h_l}{\sum_g h_{i^s} h_{j^s} h_{k^s} h_{l^s}},$$

  where $\{i^s k^s, j^s l^s\} \in g$: haplotype group compatible with the set of genotypes g

Table for TDT

- Transmission/non-transmission table
  Construct $\hat{T} = (\hat{t}_{\gamma\delta})$

$$\hat{t}_{\gamma\delta} = \sum_g \left\{ \sum_k \sum_l \hat{t}_g^{\gamma k, \delta l} + \sum_i \sum_j \hat{t}_g^{i\gamma, j\delta} \right\}$$

- Remark
  - $\hat{T}$: symmetrical in sense of $E(\hat{t}_{\gamma\delta}) = E(\hat{t}_{\delta\gamma})$
  - Adapt the marginal homogeneity test to test of linkage

Test statistic

- TDT
  Use Spielman & Ewens (1996, AJHG)'s multiallelic TDT for $\hat{T}$, i.e.,

$$T_{s\&e} = \frac{h-1}{h} \sum_{\gamma=1}^{h} \frac{(\hat{t}_{\gamma\cdot} - \hat{t}_{\cdot\gamma})^2}{\hat{t}_{\gamma\cdot} + \hat{t}_{\cdot\gamma} - 2\hat{t}_{\gamma\gamma}}$$

- Remark
  - $T_{s\&e}$ follows the chi-square distribution with $df = h - 1$ asymptotically? *No except $h = 2$*
  - Why? *Sham (1997, AJHG) & Lazzeroni and Lange (1998, Hum Hered) + dependency between cell counts*

Empirical p-value

- Randomization process
  Randomly assign to each affected offspring, with equal chance, *either* the observed genotypes at all sites *or* the non-transmitted genotypes at all sites
- Two samples (*continued*)

Sample		$\mathcal{F}$	$\mathcal{M}$	$\mathcal{C}$
1		12/12/11	22/12/22	12/12/12
	or	12/12/11	22/12/22	22/12/12
2		12/22/11	22/12/22	12/12/12
	or	12/22/11	22/12/22	22/22/12

Another approaches

- Two perspectives
  - Develop other test statistics
  - Find a way to reduce the number of compatible haplotype groups
- Adapt score test *or* LR test for testing the marginal homogeneity in paired comparison studies
- Use an unaffected sibling *or* an affected one if any

Score and LR test

- Stuart (1955, BKA)

$$T_s = \Delta' \Sigma^{-1} \Delta,$$

where $\Delta = (\hat{t}_{1\cdot} - \hat{t}_{\cdot 1}, \ldots, \hat{t}_{h-1\cdot} - \hat{t}_{\cdot h-1})$ and

$$\Sigma = (\sigma_{ij}) = \begin{cases} \hat{t}_{i\cdot} + \hat{t}_{\cdot i} - 2\hat{t}_{ii}, & i = j \\ -(\hat{t}_{ij} + \hat{t}_{ji}), & i \neq j \end{cases}$$

- Bradley & Terry (1952, BKA)

$$T_{b\&t} = 2(\log L_1 - \log L_0)$$

through a logistic model such as

$$\log(p_{ij}/p_{ji}) = b_i - b_j$$

Utilizing sibship

- Sample 1 (*continued*)
  - If another sibling $\in$ $\{12/11/12, 22/22/12, 12/22/12, 22/11/12\}$, informative!
  - if *not*, i.e, $\in \{12/12/12, 22/12/12\}$, non-informative!

Case study

- 783 students who were aged of 15 at Kangwha-eup in 1995 were monitored up to 1997 every year
- Phenotype: High BP
- Case group: students experienced at least once SBP>130mmHG *or* DBP>85mmHg
  Control group: selected from the student having lowest BP sequentially
- Trio: students whose parents' genotypes are available among students in Case group
- Case:Control=101:176; 40 trios
- 4 SNPs, A-240T, T-93C, ID, G2350A of ACE in region 17q23

Results *for* Kangwha data

- Estimated haplotype frequencies

	Control		Case		
	Bayesian	EM	Bayesian	EM	Trio
IATG†	0.003	0.003	-	-	-
IATA	-	-	-	-	0.006
IACG	0.017	0.016	0.010	0.010	0.006
IACA	0.594	0.595	0.604	0.604	0.537
ITTG	-	0.002	-	-	-
ITTA	0.034	0.033	0.015	0.015	0.031
ITCG	-	-	0.005	0.005	-
ITCA	-	-	0.010	0.010	0.006
DATG	-	-	0.005	0.005	-
DATA	-	-	-	-	-
DACG	0.011	0.012	-	-	-
DACA	0.006	0.006	-	-	-
DTTG	0.298	0.297	0.337	0.336	0.413
DTTA	0.026	0.026	0.010	0.010	-
DTCG	0.011	0.011	0.005	0.005	-
DTCA	-	-	-	-	-

† Denote haplotypes corresponding to pair of (I/D, A-240T, T-93C, G2350A)

Results *for* Kangwha data

- P-values of association and linkage test

Method		Case-Control	Trio
Single locus	I/D	0.988	0.217/0.280†
	A-240T	0.899	0.132/0.175
	T-93C	0.852	0.140/0.185
	G2350A	0.828	0.170/0.223
Haplotype		0.184§	0.152/0.095‡

† Normal approximation and Yates' continuity correction

‡ Zaho *et al.* (2000, AJHG)'s and score tests

§ Zhao, Curtis and Sham (2000, Hum Hered)'s χ^2 test

Simulation studies

- Design parameters
  - # of loci=3
  - Types of haplotype frequencies

Type	Pop.	Frequencies *of* $(H_1, H_2, H_3, H_4, H_5, H_6, H_7, H_8)$
1	1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	(0.490, 0.000, 0.210, 0.000, 0.210, 0.000, 0.090, 0.000)
2	1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
3	1	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)
	2	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)

  - Disease susceptible haplotypes=H_7, H_8
  - RR=1(level); 1.2, 1.6, 2.0, 4.0, 6.0(power)
  - # of subjects in each population=200
  - Case:Control=1:2
  - # of replication=200; # of resampling=100

Empirical levels

Approach	Conf.†	Single locus			Haplotype	
		locus 1	locus 2	locus 3	LR§ ($T_{s\&e}$)	T_s
Population-	I	0.040	0.045	0.030	0.070	
based	II	0.030	0.050	0.270	0.170	
	III	0.065	0.040	0.025	0.045	NA
	IV	0.055	0.045	0.060	0.080	
	V	0.055	0.055	0.070	0.115	
	VI	0.055	0.070	0.050	0.080	
Family-	I	0.030	0.040	0.035	0.055	0.060
based	II	0.015	0.035	0.045	0.045	0.050
	III	0.015	0.025	0.015	0.050	0.050
	IV	0.055	0.010	0.045	0.075	0.060
	V	0.035	0.055	0.070	0.080	0.075
	VI	0.045	0.050	0.040	0.035	0.040

§ Zhao, Curtis and Sham (2000, Hum Hered)'s χ^2 test

Empirical powers *for* Conf. I

		Single locus			Haplotype	
Approach	RR	locus 1	locus 2	locus 3	LR($T_{s\&e}$)	T_s
Population-	1.2	0.060	0.070	0.055	0.085	
based	1.6	0.080	0.075	0.045	0.140	
	2.0	0.155	0.125	0.035	0.220	NA
	3.0	0.370	0.340	0.065	0.555	
	4.0	0.665	0.640	0.050	0.820	
	6.0	0.975	0.975	0.035	0.995	
Family-	1.2	0.030	0.050	0.020	0.055	0.065
based	1.6	0.050	0.040	0.010	0.100	0.100
	2.0	0.090	0.115	0.025	0.115	0.120
	3.0	0.230	0.195	0.040	0.315	0.300
	4.0	0.415	0.405	0.040	0.615	0.595
	6.0	0.760	0.705	0.025	0.960	0.940

Empirical powers *for* Conf. V

		Single locus			Haplotype	
Approach	RR	locus 1	locus 2	locus 3	LR($T_{s\&e}$)	T_s
Population-	1.2	0.040	0.065	0.030	0.090	
based	1.6	0.115	0.105	0.035	0.185	
	2.0	0.225	0.245	0.090	0.300	NA
	3.0	0.645	0.675	0.055	0.760	
	4.0	0.940	0.940	0.030	0.990	
	6.0	1.000	1.000	0.030	1.000	
Family-	1.2	0.035	0.060	0.030	0.050	0.060
based	1.6	0.070	0.065	0.025	0.110	0.095
	2.0	0.110	0.130	0.045	0.170	0.155
	3.0	0.340	0.365	0.040	0.485	0.495
	4.0	0.630	0.595	0.045	0.810	0.795
	6.0	0.920	0.930	0.050	0.985	0.980

Summary

- Investigate single-locus-based and haplotype-based association and linkage tests
- The first two high-frequency haplotypes are IACA and DTTG in Kangwha data
- Hypertension is NOT linked with the makers on ACE gene with Kangwha data.
- . . . And in the future . . .
  - how to reduce the haplotype uncertainty
  - how to include observations with only one parent *or* with only sibship
  - how to combine all types of observations into one test statistic

Thank you