Tests for Linkage and/or Association Between Hypertension and Angiotensinogen(AGT) Gene Based on Haplotypes

Jinheum Kim

`jinhkim@suwon.ac.kr`

Department of Applied Statistics

University of Suwon

Contents

- Allele-based sib TDT: Review
- Haplotype-based sib TDTs
- A real example
- Concluding remarks

Sib TDT

- When does it need? Diseases with late age of onset
  $\Rightarrow$ Possibly parental data not available
  $\Rightarrow$ Sibling's data available instead
- Minimum requirements
  - (i) At least one unaffected sib additionally
    $\Rightarrow$ Able to compare the marker distribution between two population of the affected and the unaffected
  - (ii) Two sibs' marker genotypes not identical
    $\Rightarrow$ If not, noninformative
- TDT *vs.* sib TDT
  - Controls: parental *vs.* unaffected sib(s)

Spielman & Ewens' Test (AJHG, 1998)

- With two-allele marker for simplicity
- Idea: compare the marker allele frequencies between the affected and unaffected sibs
- O_f=# of M_1 allele among the affected sibs within the sibship f
- $E_f = \mathrm{E}_0(O_f), V_f = \mathrm{Var}_0(O_f)$ under H_0 : no linkage
- $z^2 = \left(\sum_f O_f - \sum_f E_f\right)^2 / \sum_f V_f \sim \chi_1^2$ asymptotically under H_0

Remarks

- A kind of stratified statistic to adjust the confounding factor which is the varying genotype frequencies from sibship to sibship
- $\mathbf{x}'_f = (x_{f1}, x_{f2}, x_{f3})$ follows a conditionally multivariate hypergeometric distribution under H_0, where $x_{f1}, x_{f2},$ and x_{f3} are, respectively, the number of affected sibs who have genotypes $M_1M_1, M_1M_2,$ and M_2M_2
- O_f is a linear combination of $\mathbf{x}_f$, *i.e.* $\mathbf{c}'\mathbf{x}_f$,
  $\mathbf{c}' = (2, 1, 0)$
  $\Rightarrow E_f$ & V_f are calculable through the distribution of $\mathbf{x}_f$

Remarks

- Explicit forms of E_f & V_f
  - $N_f^a(N_f^u)$=# of affected(unaffected)sibs within the sibship f
  - r_f=# of sibs who are of genotype M_1M_1 within the sibship f
  - s_f=# of sibs who are of genotype M_1M_2 within the sibship f
  - Null mean

$$E_f = (2r_f + s_f)\frac{N_f^a}{N_f}, \quad N_f = N_f^a + N_f^u$$

  - Null variance

$$V_f = [4r_f(N_f - r_f - s_f) + s_f(N_f - s_f)]\frac{N_f^a N_f^u}{N_f^2(N_f - 1)}$$

Notations

- $G_1, \ldots, G_k (k = 3^c)$: distinct genotypes in case 2-allele markers at c loci,
- $h_1, \ldots, h_l (l = 2^c)$: distinct haplotypes
- x_{fg}, t_{fg}: # of the affected sibs and total sibs with genotype G_g within the fth sibship, $f = 1, \ldots, F; g = 1, \ldots, k$
- r_{fh}, s_{fh}: # of sibs having haplotype pairs hh and $hm (m \neq h)$ within the fth sibship, $f = 1, \ldots, F; h = h_1, \ldots, h_l$

Proposed test statistic

- Idea: apply Spielman & Ewens' test for each haplotype whenever the phases of genotype are resolved
  $\Rightarrow$ How does it possible? r_{fh} & s_{fh} are deterministic
- O_{fh}: # of haplotype h in the affected sibs within the fth sibship, $f = 1, \ldots, F; h = h_1, \ldots, h_l$
- $E_{fh} = \mathrm{E}_0(O_{fh}), V_{fh} = \mathrm{Var}_0(O_{fh})$ under $H_0 :$ no linkage
- For each h,
  $$z_h^2 = \left(\sum_f O_{fh} - \sum_f E_{fh}\right)^2 / \sum_f V_{fh} \sim \chi_1^2$$
  asymptotically under H_0

Two omnibus tests

- $T_1 = \text{max}_{i=1,\ldots,l} |z_{h_i}|$
  $\Rightarrow$ Need Bonferroni's correction for multiple tests
  $\Rightarrow$ Use Permutation test
- $T_2 = (l-1)/l \sum_{i=1}^{l} z_{h_i}^2 \sim \chi_{l-1}^2$ asymptotically under H_0
  $\Rightarrow$ Conservative
  $\Rightarrow$ Why? Ignore dependency between haplotypes among sibs within a sibship

Permutation test procedure

- Step 0: calculate T, with value T_0, for the given data set
- Step 1: for each sibship, randomly permute affection status
- Step 2: calculate T on this pseudo-sample and determine whether it is more extreme than T_0
- Step 3: repeat steps 1 and 2 B times and estimate the P value as the proportion of times that T is more extreme than T_0

- Reference: Monks *et al.*(AJHG, 1998)

Haplotype reconstruction

- When required?
  - more than 2 heterozygous loci exist

- *In-silico* methods
  - Clark algorithm (Clark, MBE, 1990)
  - EM algorithm (Excoffier & Slatkin, MBE, 1995)
  - Gibbs sampling method (Stephens *et al.*, AJHG,2001)
  - Partition-ligation(Niu *et al*, AHJG, 2002)

Modified proposed tests

- When the phases of genotype are unresolved, r_{fh}, s_{fh} are probabilistic
- $\mathcal{H}_g$: set of all ordered haplotype pairs consistent with genotype $G_g, g = 1, \ldots, k$
- f_h: estimated frequency of haplotype $h, h = h_1, \ldots, h_l$
- $D_g = \Pr(G_g | f_h, h = h_1, \ldots, h_l) = \sum_{(s,t) \in \mathcal{H}_g} f_s f_t$ under random mating & HWE
- $w_{stg} = \Pr($Haplotype pair $= (s, t) | G_g) = f_s f_t / D_g$

Modified proposed tests

- Modified O_{fh}

$$\hat{O}_{fh} = 2\sum_{g=1}^{k} x_{fg}\{\sum_{(s,t)\in\mathcal{H}_g} w_{stg} I(s=h, t=h)\}$$

$$+\sum_{g=1}^{k} x_{fg}[\sum_{(s,t)\in\mathcal{H}_g} w_{stg}\{I(s=h, t=m, m\neq h)+I(s=m, t=h, m\neq h)\}]$$

- Modified r_{fh}

$$\hat{r}_{fh} = \sum_{g=1}^{k} t_{fg}\{\sum_{(s,t)\in\mathcal{H}_g} w_{stg} I(s=h, t=h)\}$$

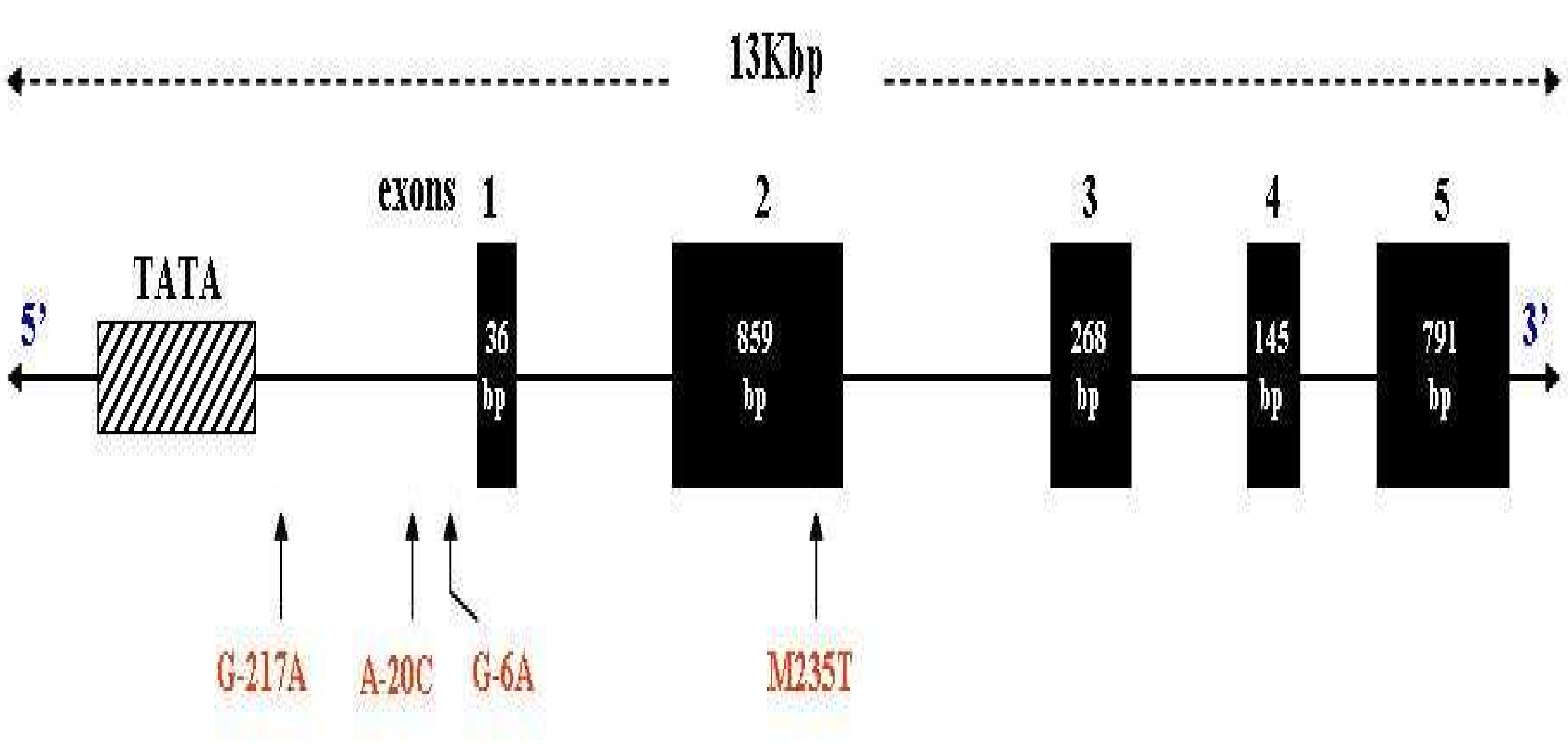
- Modified s_{fh}

$$\hat{s}_{fh} = \sum_{g=1}^{k} t_{fg}[\sum_{(s,t)\in\mathcal{H}_g} w_{stg}\{I(s=h, t=m, m\neq h)+I(s=m, t=h, m\neq h)\}]$$

A real example

- Data: 92 sibship adopted from Yonsei Cardiovascular Genome Center
- Phonotype: Hypertension
- Purpose: Test for linkage between AGT gene and hypertension
- Materials: 4 SNPs $\Rightarrow$ G-217A(s_1), A-20C(s_2), G-6A(s_3), M235T(s_4)
- Empirical $p-$values based on 10,000 times of permutation

Diagram of AGT gene



Schematic diagram of the human *AGT* gene illustrating the location of 4 diallelic polymorphisms (1q42-3)

Estimated haplotype frequencies

SNP(s)							
(s_1, s_2, s_3, s_4)		(s_1, s_3, s_4)		(s_3, s_4)		s_3	
Haplotype	f_h	Haplotype	f_h	Haplotype	f_h	Haplotype	f_h
AAAT	-	AAT	-	AT	-	A	0.8281
AAAC	0.1983	AAC	0.2031	AC	0.8281		
AAGT	-	AGT	-	GT	0.1641	G	0.1719
AAGC	-	AGC	-	GC	0.0078		
ACAT	-						
ACAC	-						
ACGT	-						
ACGC	-						
GAAT	-	GAT	-				
GAAC	0.4655	GAC	0.6250				
GAGT	0.1810	GGT	0.1641				
GAGC	0.0086	GGC	0.0078				
GCAT	-						
GCAC	0.1466						
GCGT	-						
GCGC	-						

Empirical p-values

			T_1		T_2	
#(SNPs)	SNPs	F	observed	$p-$value	observed	$p-$value
4	s_1, s_2, s_3, s_4	26(16)	1.342	0.538	4.382	0.420
3	s_1, s_2, s_3	26(15)	1.342	0.559	2.796	0.475
	s_1, s_2, s_4	28(17)	1.087	0.669	2.337	0.585
	s_1, s_3, s_4	28(15)	1.357	0.370	4.042	0.295
	s_2, s_3, s_4	26(11)	1.342	0.461	3.982	0.312
2	s_1, s_2	35(19)	0.662	0.836	0.621	0.795
	s_1, s_3	28(14)	1.286	0.424	2.400	0.403
	s_1, s_4	31(16)	1.302	0.373	2.460	0.373
	s_2, s_3	26(10)	1.342	0.479	2.300	0.304
	s_2, s_4	28(12)	1.087	0.525	1.859	0.502
	s_3, s_4	28(9)	1.357	0.240	3.124	0.236
1	s_1	38(13)	0.176	1.000	0.031	1.000
	s_2	35(12)	0.494	0.775	0.244	0.775
	s_3	28(7)	1.151	0.334	1.324	0.334
	s_4	44(11)	1.109	0.382	1.230	0.382

Concluding remarks

- Extend Spielman & Ewens’ test based on haplotype instead of allele
- Modify sib TDTs with conditional probabilities due to haplotype uncertainty
- Lack of efficient sample size, 7 to 16 sibship among 92 sibship
- More significant when using two SNPs, G-6A & M235T, than when using three(G-217A additionally) or all four SNPs
- Develop a test including covariances between haplotypes among sibs within a sibship

Thank you.