Sib transmission and disequilibrium tests for linkage using multiple highly linked markers

Jinheum Kim

jinhkim@suwon.ac.kr

Department of Applied Statistics

University of Suwon

Contents

- Allele-based sib TDT: review
- Propose omnibus tests based on haplotype
- Simulation studies
- Concluding remarks

Association study

- Goal: test for association between genetic markers and disease-susceptibility genes related to a trait
- Sources: causal association, LD, confounding
- Two ways: population-based case-control study *or* family-based TDT
  $\rightarrow$ TDT: Not affected by population stratification unlike case-control study, *i.e.* free from a chance false positive
- Requirement for TDT: proband's marker genotype + parental marker genotypes

An example of spurious association

Sample	Pop'n 1 M_1	Pop'n 1 M_2	Pop'n 1 Total	Pop'n 2 M_1	Pop'n 2 M_2	Pop'n 2 Total
Case	9	1	10	25	25	50
Control	81	9	90	25	25	50
Total	90	10	100	50	50	100
	$\chi^2 = 0(1.000)$			$\chi^2 = 0(1.000)$		

$$\downarrow (1:1)$$

Sample	Combined M_1	Combined M_2	Combined Total
Case	34	26	60
Control	106	34	140
Total	140	60	200
	$\chi^2 = 7.26(0.007)$		

Sib TDT

- When does it need? late-onset diseases
  $\rightarrow$ 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 dist'n bet'n two population of the affected and the unaffected
  - (ii) two sibs' marker genotype not identical
    $\rightarrow$ if not, noninformative

Tests related to sib TDT

- Curtis(AHG, 1997)
- Boehnke & Langefeld (AJHG, 1998)
- Spielman & Ewens (AJHG, 1998)
- Hovath & Laird (AJHG, 1998)

Spielman & Ewens' Test

- With two-allele marker for simplicity
- Idea: compare the marker allele frequencies bet'n the affected and unaffected sibs
- Data structure, given $N_f^a, N_f^u, t_{f1}, t_{f2}, t_{f3}$

	Freq. of genotypes			
Affection status	M_1M_1	M_1M_2	M_2M_2	Total
Affected	x_{f1}	x_{f2}	x_{f3}	N_f^a
Unaffected	y_{f1}	y_{f2}	y_{f3}	N_f^u
Total	t_{f1}	t_{f2}	t_{f3}	N_f

Test statistic

- O_f=# of M_1 allele among the affected sibs within the sibship f
- $E_f = \mathsf{E}_0(O_f), V_f = \mathsf{Var}_0(O_f)$ 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 multi-hypergeometric distribution under H_0
- O_f is a linear combination of $\mathbf{x}_f$, *i.e.* $\mathbf{x}'_f\mathbf{c}$,
  $\mathbf{c} = (2, 1, 0)'$
  $\rightarrow A_f, V_f$ are calculable through the dist'n of $\mathbf{x}_f$

Why haplotype-based? But ...

- (Def'n) Haplotype set of alleles on a chromosome
- Many markers has been genotyped within a very short physical distance
- More informative
- Haplotype information is not usually available from genotype information
  $\rightarrow$ For example, when the number of heterozygous loci equals c, the number of possible haplotype pairs corresponds to 2^{c-1}

Notations

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

Data structure

- Data structure, given $N_f^a, N_f^u, \mathbf{t}'_f = (t_{f1}, \ldots, t_{fk})$

Affection status	Freq. of genotypes G_1	$\ldots$	G_k	Total
Affected	x_{f1}	$\ldots$	x_{fk}	N_f^a
Unaffected	y_{f1}	$\ldots$	y_{fk}	N_f^u
Total	t_{f1}	$\ldots$	t_{fk}	N_f

Reconstruction of data structure

- When the phases of genotype are resolved, r_{fh}, s_{fh} are deterministic
- Reconstruct l sub-tables based on haplotypes, $e.g.$, for haplotype h,

	Freq. of haplotype pairs			
Affection status	hh	$hk(k \neq h)$	$pq(p, q \neq h)$	Total
Affected				N_f^a
Unaffected				N_f^u
Total	r_{fh}	s_{fh}	$N_f - r_{fh} - s_{fh}$	N_f

Proposed test statistic

- Idea: apply Spielman & Ewens' test to the reconstructed table sequentially for each haplotype
- 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

Omnibus tests

- $T_1 = \text{max}_h |z_h|$
  $\rightarrow$ Bonferroni's correction needs for multiple tests
- $T_2 = (h-1)/h \sum_h z_h^2 \sim \chi^2_{h-1}$ asymptotically under H_0
  $\rightarrow$ Conservative; ignore dependency bet'n haplotypes among sibs within a sibship

Permutation 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 N 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 reconstruction table

- 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: sample frequency of haplotype $h, h = h_1, \ldots, h_l$
- Estimated column marginals in reconstructed table

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

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

where

$$D_g = \sum_{(s,t) \in \mathcal{H}_g} f_s f_t, \; w_{stg} = f_s f_t / D_g, \; s, t = h_1, \ldots, h_l$$

Simulation studies: design pars

- Types of haplotype frequencies

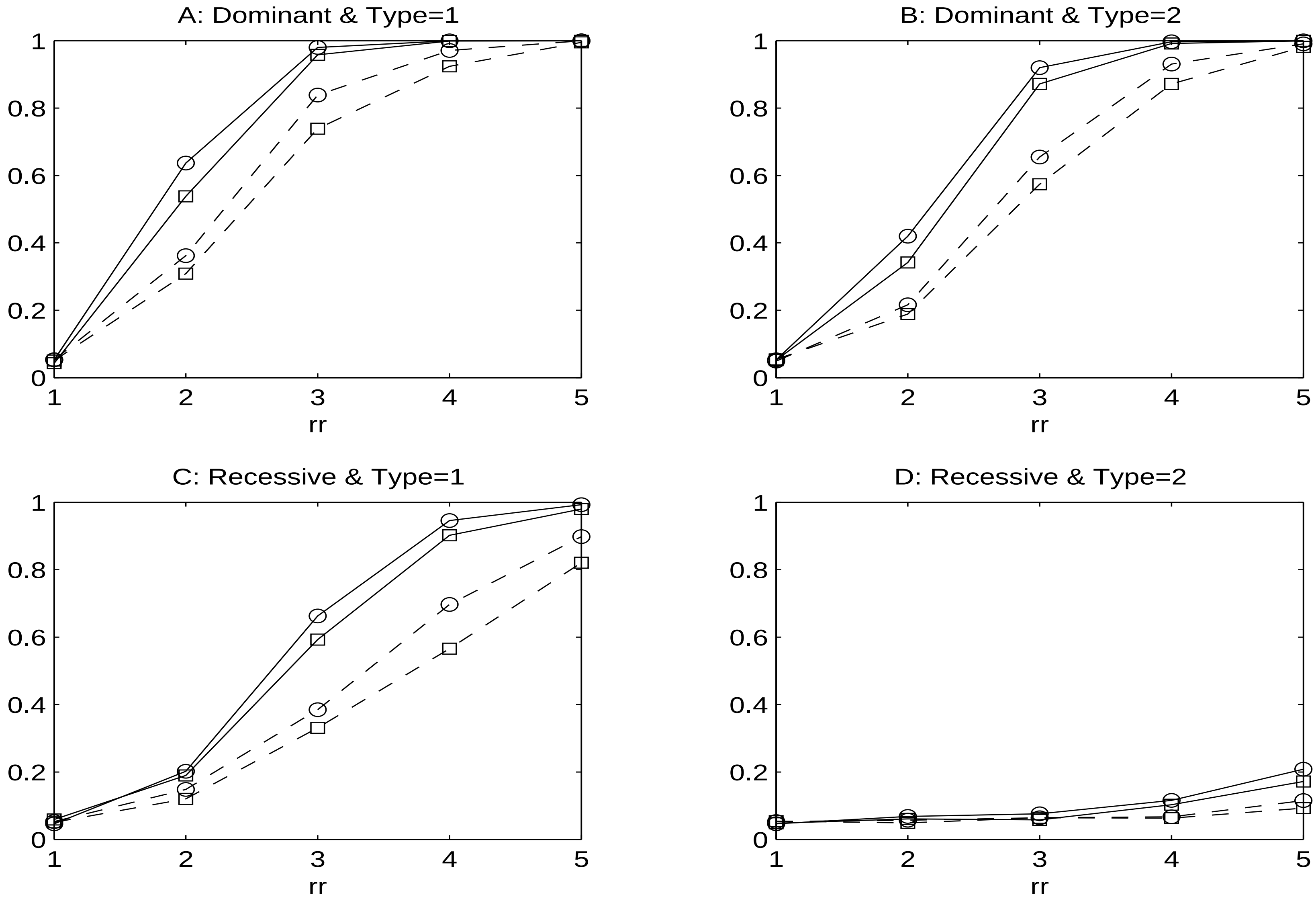
Type	Pop'n	Frequencies of $(h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8)$
I	1	(0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125)
	2	(0.250, 0.000, 0.250, 0.000, 0.250, 0.000, 0.250, 0.000)
II	1	(0.343, 0.147, 0.147, 0.063, 0.147, 0.063, 0.063, 0.027)
	2	(0.490, 0.000, 0.210, 0.000, 0.210, 0.000, 0.090, 0.000)

- # of sibship(F)=50, 100(level); 200(power)
- # of sibs within each sibship(s)=2, 5
- baseline risk=0.1 for pop'n 1; *br*= 0.2, 0.3, 0.4 for pop'n 2
- rr(power)=2, 3, 4, 5
- # of replication=1,000; # of permutation=300

Empirical levels

		F	50				100			
		T	T_1		T_2		T_1		T_2	
Type	br	s	2	5	2	5	2	5	2	5
I	2		.053	.052	.062	.056	.049	.054	.049	.056
	3		.060	.051	.054	.056	.069	.046	.072	.039
	4		.069	.053	.065	.044	.049	.058	.051	.053
II	2		.052	.054	.048	.054	.043	.055	.054	.050
	3		.046	.047	.051	.046	.053	.047	.058	.053
	4		.063	.038	.055	.041	.051	.051	.052	.043

Empirical powers



Concluding remarks

- Extend Spielman & Ewens' test based on haplotype instead of allele
- Modify reconstructed table with conditional probabilities due to haplotype uncertainty
- Robust to population admixture regardless of haplotype dist'n, br, and s
- T_2 is more powerful than T_1

- Alternative approaches: based on $2 \times k$ genotype table *or* $2 \times l$ halpotype table

Thank you.