Statistical Tests for X Chromosome Association Study with Simulations

Jian Wang

July 10, 2012

Statistical Tests

- Zheng G, et al. 2007. Testing association for markers on the X chromosome. Genetic Epidemiology 31:834-843
  - NOT assuming X-inactivation (Wrong!)
  - Six different association tests: combinations of allele-based test and genotype-based test in males and females.
  - Use the minimum p value of the six test statistics
  - Need to adjust for the correlation between the test statistics

(i) Female (genotype)

	BB	AB	AA	Total
Cases	r_{f0}	r_{f1}	r_{f2}	r_f
Controls	s_{f0}	s_{f1}	s_{f2}	s_f
Total	n_{f0}	n_{f1}	n_{f2}	n_f

(ii) Male (allele)

	B	*A*	Total
Cases	r_{m0}	r_{m1}	r_m
Controls	s_{m0}	s_{m1}	s_m
Total	n_{m0}	n_{m1}	n_m

(iii) Female (allele)

	B	*A*	Total
Cases	$2r_{f0}+r_{f1}$	$r_{f1}+2r_{f2}$	$2r_f$
Controls	$2s_{f0}+s_{f1}$	$s_{f1}+2s_{f2}$	$2s_f$
Total	$2n_{f0}+n_{f1}$	$n_{f1}+2n_{f\,2}$	$2n_f$

(iv) Male+female (allele)

	B	*A*	*Total*
Cases	$2r_{f0}+r_{f1}+r_{m0}$	$2r_{f2}+r_{f1}+r_{m1}$	$2r_f+r_m$
Controls	$2s_{f0}+s_{f1}+s_{m0}$	$2s_{f2}+s_{f1}+s_{m1}$	$2s_f+s_m$
Total	$2n_{f0}+n_{f1}+n_{m0}$	$2n_{f2}+n_{f1}+n_{m1}$	$2n_f+n_m$

Allele-counting method?

- Count alleles in a 2*2 table and compare the allele counts between cases and controls.
- Assumes that the effect of 1 copy of a variant allele on phenotype is the same in males as in females
  - Males have half impact on the results of this test as females.
- Assuming Hardy-Weinberg proportion in females.
  - Would have wrong size if HWP fails.

Statistical Tests

- Clayton D. 2008. Testing for association on the X chromosome. Biostatistics 9:593-600
  - Assuming X-inactivation

- Plinks: commonly used
  - NOT assuming X-inactivation
  - Using females only
  - Using whole sample with sex as a covariate
    - genotypes of male are coded as (0,1) for *a* and *A*, respectively
    - genotypes of female are coded as (0,1,2) for *aa*, *aA* and *AA*, respectively

Clayton's Approaches

- Assumptions:
  - X-inactivation
  - Allele frequencies are same in males and females
  - NOT assume Hardy-Weinberg proportion in females
- Statistical tests:
  - 1 degree-of-freedom test ✓
  - 2 degree-of-freedom test (less powerful)

1 df Test for Autosomal Loci

- Consider an autosomal diallelic locus with genotypes aa, Aa and AA coded as A_i=(0,1,2), respectively, and a phenotype Y_i, i=1,2...N.
- Score test for testing for an additive effect of this locus on phenotype is given as the genotype-phenotype covariance:

$$U_A = \sum_{i=1}^{N} (Y_i - \overline{Y}) A_i,$$

- $\overline{Y}$ is the mean of Y in the sample.
- Assume the variance V_A is constant for all A_i
- Consider the distribution $\Pr(A_i|Y_i)$

1 df Test for Autosomal Loci

- The statistic is asymptotically normally distributed under null hypothesis with mean 0 and variance

$$\mathrm{Var}(U_A) = V_A \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

- where V_A is the variance of genotype A_i:

$$\widehat{V_A} = \frac{1}{N-1} \sum_{i=1}^{N} (A_i - \overline{A})^2$$

- The ratio $(U_A)^2 / \widehat{\mathrm{Var}(U_A)}$ is asymptotically distributed as chi-squared on 1df.
- Applicable for a quantitative phenotype.

1 df Test for X Chromosome Loci

- X-inactivation:
  - A female will have approximately half her cells with 1 copy active while the remainder of her cells have the other copy activated.
  - Males should be equivalent to homozygous females.
- Coding for X loci:
  - Males: $A_i = (0,2)$ for a and A alleles respectively
  - Females: $A_i = (0,1,2)$ for aa, Aa and AA genotypes respectively.

1 df Test for X Chromosome Loci

- If the allele frequencies are same in males and females
  - The expectation of *A* is equal in males and females
  - The expectation of U_A will remain 0
- The variance of *A* differs between males and females
  - Variance is *2p(1-p)* in females and *4p(1-p)* in males, where p is the minor allele frequency.

$$\hat{V}_M = 4p(1-p)$$

$$\hat{V}_F = 1/(F-1)\sum_{i=1}^{F}(A_i - \overline{A})^2$$

$$\widehat{V} = \widehat{V_F}\sum_{i=1}^{F}(Y_i - \overline{Y})^2 + \widehat{V_M}\sum_{i=F+1}^{N}(Y_i - \overline{Y})^2$$

- The same 1df chi-squared test can be conducted.

Stratified Tests

- If the allele frequencies are NOT same in males and females
  - Stratified score test in the analysis
    - The test statistic and its estimated variance are calculated separately in each stratum
    - Sum the test statistics and variances over strata
    - Same 1 df chi-sq test can be performed
    - The different stratum contributions would need to be weighted appropriately
  - If the sex and phenotype has strong association, this will result in loss of power
  - However, in the simulation studies, we find that there is no significant power loss

Conditional on Genotype

- Consider the distribution $\Pr(Y_i|A_i)$
  - $U_A = \sum_{i=1}^{N}(Y_i - \overline{Y})A_i, \ \ \mathrm{Var}(U_A) = V_Y \sum_{i=1}^{N}(A_i - \overline{A})^2$
  - $V_Y = \sum_{i=1}^{N}(Y_i - \overline{Y})^2/(N-1)$
  - An identical asymptotic test
  - Regress Y on A in a GLM and test for regression coefficient of A using a Huber-White estimate for the variance-covariance matrix of coefficients
  - The stratified version of the test would be obtained by additionally including the stratifying factor in the GLM

R Package for Clayton's Approaches

- R package: whole-genome association studies
- http://www.bioconductor.org/packages/2.10/bioc/html/snpStats.html
- Library: snpStats
- It can use the plink binary ped file as input
- Incorporate the information of "diploid" for female genotype
- An example:

```
genobed<-"xchrom_17.bed"
genobim<-"xchrom_17.bim"
genofam<-"xchrom_17.fam"
data<-read.plink(genobed,genobim,genofam)
data1<-data
D<-data$fam$sex==2
data1$genotype<-new("XSnpMatrix",data$genotype,diploid=D)
re1<-single.snp.tests(data1$fam$affected,data=data1$fam,snp.data=data1$genotype)
re2<-single.snp.tests(data1$fam$affected,data1$fam$sex,data=data1$fam,snp.data=data1$genotype)
```

R Package for Clayton's Approaches

		Not adjust for sex					Adjust for sex		
	N	Chi.square d.1.df	Chi.square d.2.df	P.1df	P.2df	Chi.square d.1.df	Chi.square d.2.df	P.1df	P.2df
rs3120733	3480	7.680666	8.94684	0.005582	0.011408	8.095423	9.361597	0.004438	0.009272
rs525161	3481	12.38907	14.60077	0.000432	0.000675	11.75987	13.97157	0.000605	0.000925
rs559165	3480	11.23412	11.65569	0.000803	0.002944	11.21578	11.63735	0.000811	0.002972
rs28576503	3482	12.98154	13.78788	0.000315	0.001014	12.90062	13.70695	0.000328	0.001056
rs28719866	3482	11.11458	11.40087	0.000857	0.003345	11.02953	11.31582	0.000897	0.00349
rs28444648	3480	10.15823	10.47012	0.001437	0.005326	9.91551	10.2274	0.001639	0.006014
rs28599172	3467	10.5431	10.54374	0.001166	0.005134	11.97719	11.97784	0.000539	0.002506
rs5940510	3481	7.719249	9.57082	0.005464	0.008351	7.962648	9.814218	0.004775	0.007394
rs12557310	3480	13.14358	13.2422	0.000289	0.001332	13.44997	13.54858	0.000245	0.001143
rs1454268	3482	9.265828	11.56982	0.002335	0.003074	9.689129	11.99312	0.001854	0.002487
rs1084451	3482	8.886967	10.98022	0.002872	0.004127	9.270666	11.36392	0.002329	0.003407
rs5940536	3469	13.228	14.87472	0.000276	0.000589	11.50504	13.15176	0.000694	0.001394
rs5983743	3481	9.826408	9.888733	0.00172	0.007123	9.764787	9.827113	0.001779	0.007346
rs5940540	3477	10.39686	10.87112	0.001262	0.004359	11.0602	11.53446	0.000882	0.003128
rs5940403	3476	11.62007	11.92036	0.000652	0.002579	11.63626	11.93654	0.000647	0.002559
rs676920	3482	10.94061	12.35371	0.000941	0.002077	11.10087	12.51396	0.000863	0.001917
rs553678	3475	12.41785	12.93499	0.000425	0.001553	12.54723	13.06436	0.000397	0.001456

Genetic Model Assumption

- Clayton's approach considered X-inactivation, but does not model other X chromosome features.
- We are trying to model X-inactivation, skewness and escaping in X-inactivation

	Autosomal			
	Additive	**Dominant**	**Recessive**	
AA	2	1	1	
Aa	1	1	0	
aa	0	0	0	
	X Chromosome (Female)			
	Additive	**Dominant**	**Recessive**	**Escaping**
AA	1	1	1	2
Aa	0.5	1	0	1
aa	0	0	0	0
	X Chromosome (Male)			
A	1	1	1	1
a	0	0	0	0

- Random X-inactivation: additive
- More than half cells have copy with risk allele active: dominant
- Less than half cells have copy with risk allele active: recessive
- Escaping: the effect of Aa in female is same as the one of A in male (plink)

Gendrel and Heard. Fifty years of X-inactivation research. *Development* 138(23):5049–5055

Possible Approach

- For each SNP on X chromosome, we apply four different coding schemes and conduct regression on each coding scheme
- Four correlated statistical tests, so we need to adjust for multiple testing
  - Take the minimum of the 4 p values
  - FDR
  - The resulting 4 p values are adjusted for multiple testing using the approach proposed by Conneely and Boehnke (AJHG, 2007): pACT
- Compare the results with plink and Clayton's 1df approach

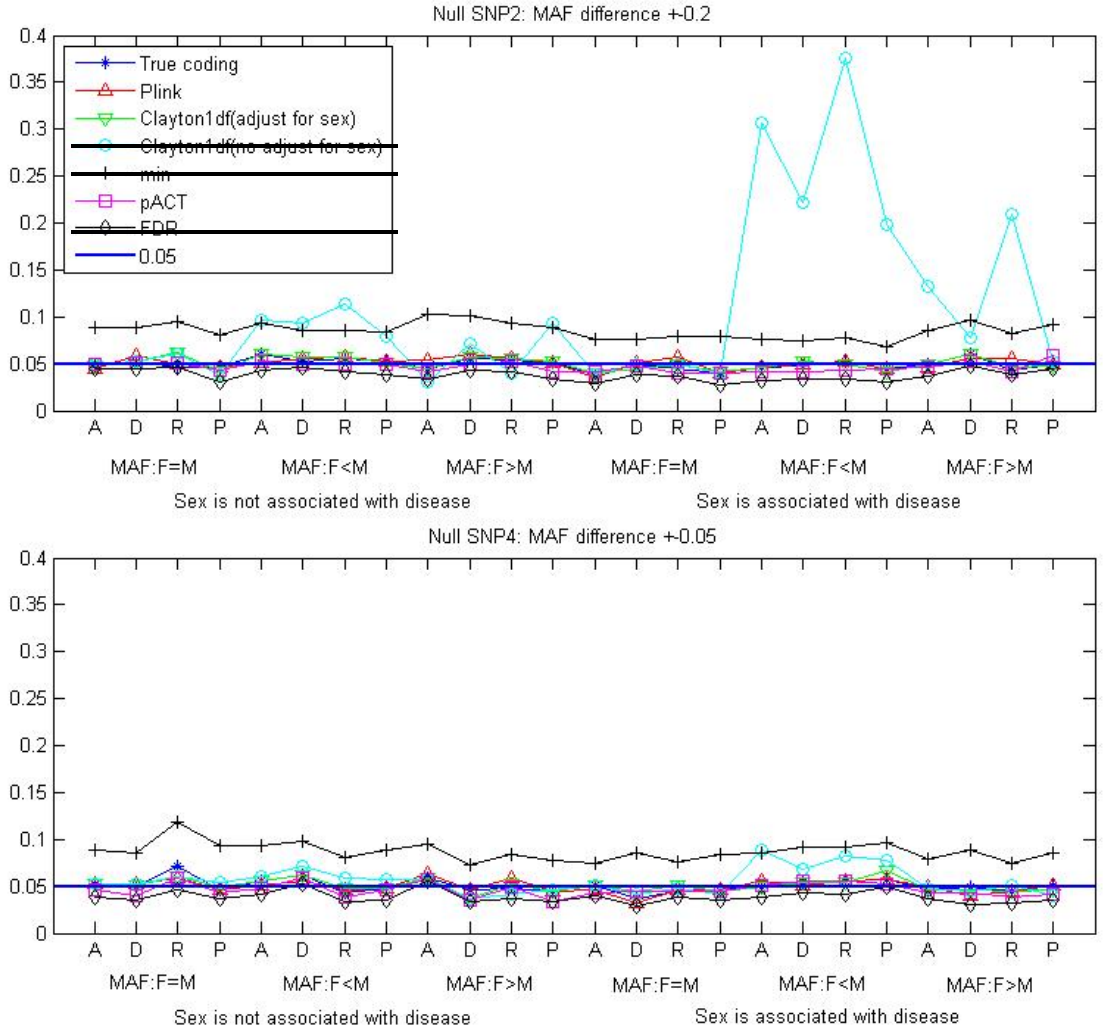
Simulation Studies

- Consider 4 di-allelic SNPs, X_1, X_2, X_3 and X_4 and a binary phenotype Y =(0,1)

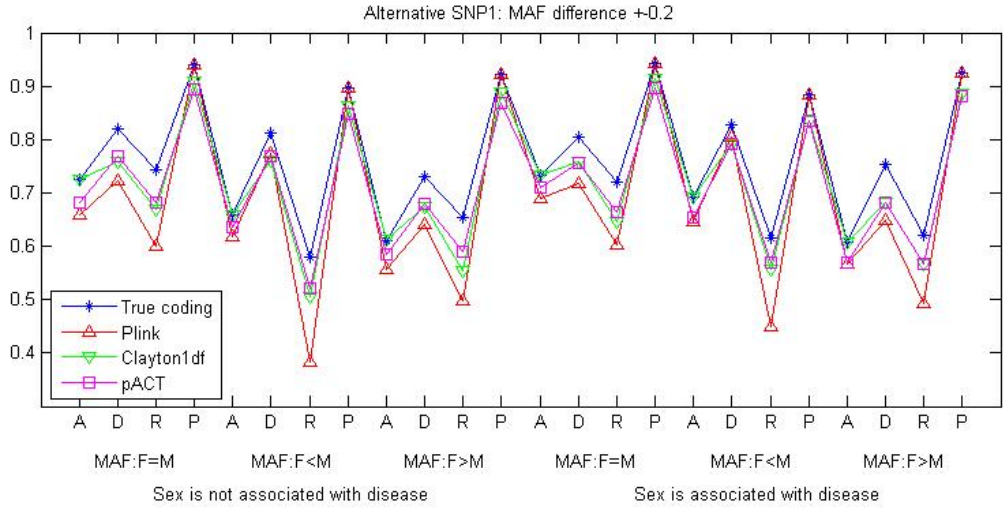
- Logit(Pr(Y=1|X_1,X_2, X_3,X_4, Sex)) $= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5$Sex

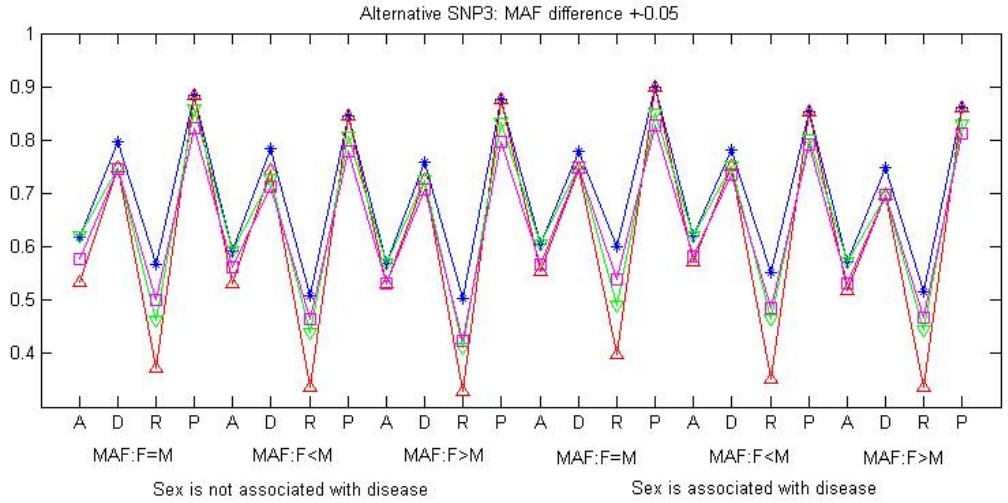
	OR					MAF				
	SNP1	SNP2	SNP3	SNP4	Sex	SNP1&2		SNP3&4		Model
b0	exp(b1)	exp(b2)	exp(b3)	exp(b4)	exp(b5)	F	M	F	M	
-2.55	1.3	1	1.3	1	1	0.4	0.4	0.25	0.25	ADD
-2.55	1.3	1	1.3	1	1	0.4	0.4	0.25	0.25	DOM
-2.55	1.3	1	1.3	1	1	0.4	0.4	0.25	0.25	REC
-2.55	1.3	1	1.3	1	1	0.4	0.4	0.25	0.25	PLINK
-2.55	1.3	1	1.3	1	1	0.2	0.4	0.2	0.25	ADD
-2.55	1.3	1	1.3	1	1	0.2	0.4	0.2	0.25	DOM
-2.55	1.3	1	1.3	1	1	0.2	0.4	0.2	0.25	REC
-2.55	1.3	1	1.3	1	1	0.2	0.4	0.2	0.25	PLINK
-2.55	1.3	1	1.3	1	1	0.4	0.2	0.25	0.2	ADD
-2.55	1.3	1	1.3	1	1	0.4	0.2	0.25	0.2	DOM
-2.55	1.3	1	1.3	1	1	0.4	0.2	0.25	0.2	REC
-2.55	1.3	1	1.3	1	1	0.4	0.2	0.25	0.2	PLINK
-2.95	1.3	1	1.3	1	1.5	0.4	0.4	0.25	0.25	ADD
-2.95	1.3	1	1.3	1	1.5	0.4	0.4	0.25	0.25	DOM
-2.95	1.3	1	1.3	1	1.5	0.4	0.4	0.25	0.25	REC
-2.95	1.3	1	1.3	1	1.5	0.4	0.4	0.25	0.25	PLINK
-2.95	1.3	1	1.3	1	1.5	0.2	0.4	0.2	0.25	ADD
-2.95	1.3	1	1.3	1	1.5	0.2	0.4	0.2	0.25	DOM
-2.95	1.3	1	1.3	1	1.5	0.2	0.4	0.2	0.25	REC
-2.95	1.3	1	1.3	1	1.5	0.2	0.4	0.2	0.25	PLINK
-2.95	1.3	1	1.3	1	1.5	0.4	0.2	0.25	0.2	ADD
-2.95	1.3	1	1.3	1	1.5	0.4	0.2	0.25	0.2	DOM
-2.95	1.3	1	1.3	1	1.5	0.4	0.2	0.25	0.2	REC
-2.95	1.3	1	1.3	1	1.5	0.4	0.2	0.25	0.2	PLINK

Simulation Results: Type I Errors



Simulation Results: Powers





Future Work

- Better understanding of the biological mechanism for X chromosome
- Other approaches to simulate data which could better represent X chromosome
- Bayesian approaches to incorporate information of the X chromosome features, such as skewness
- Thanks!