Genetic Analysis of Ordinal Traits

Heping Zhang

Department of Epidemiology and Public Health

Yale University

Harvard University

September 21, 2006

YALE UNIVERSITY

SCHOOL OF MEDICINE

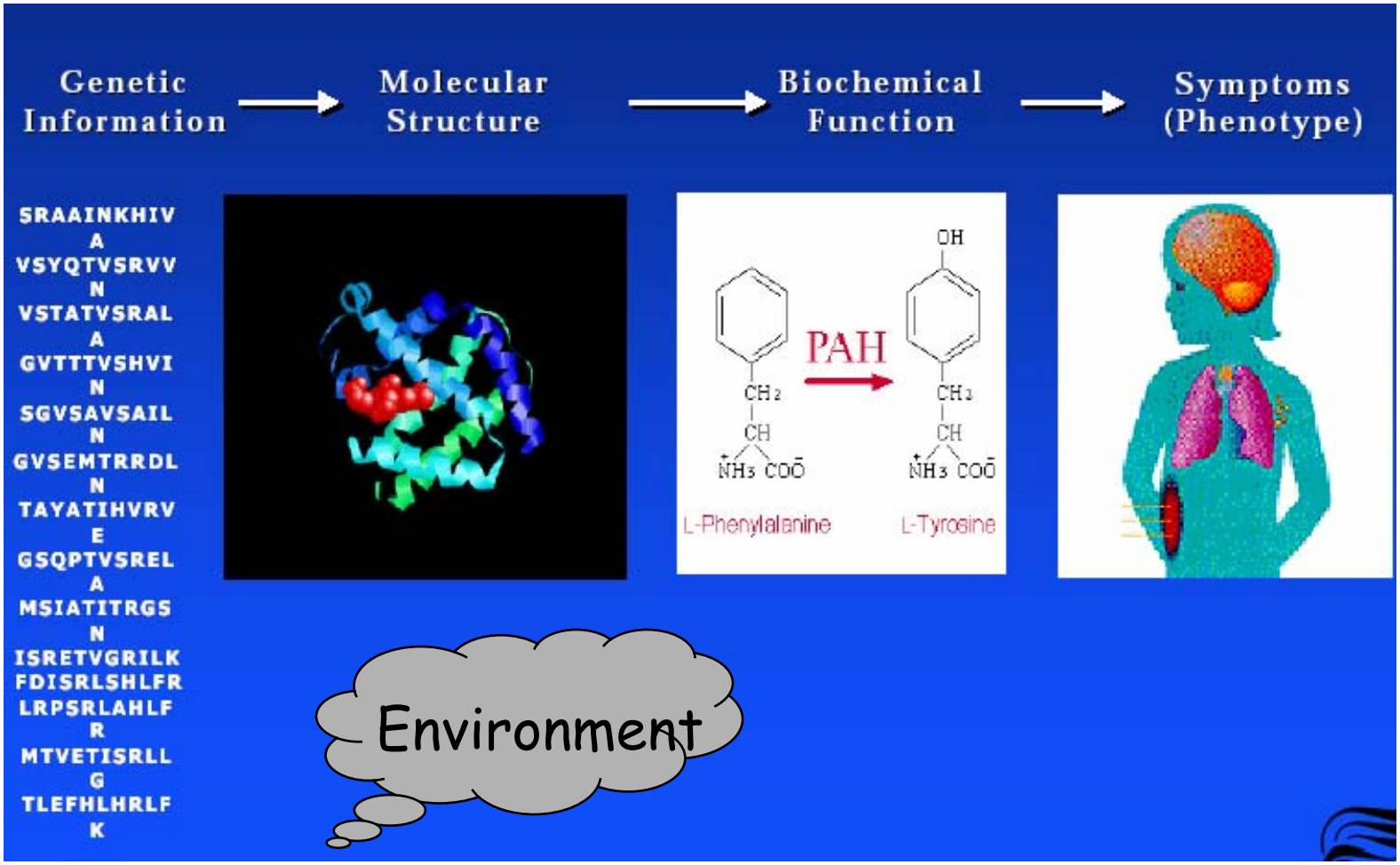
Outline

- Background
- Linkage Analysis
  - Model/Simulation/Application
  - Discussion/References
- Association Analysis
  - Model/Simulation/Application
  - Discussion/References

YALE UNIVERSITY

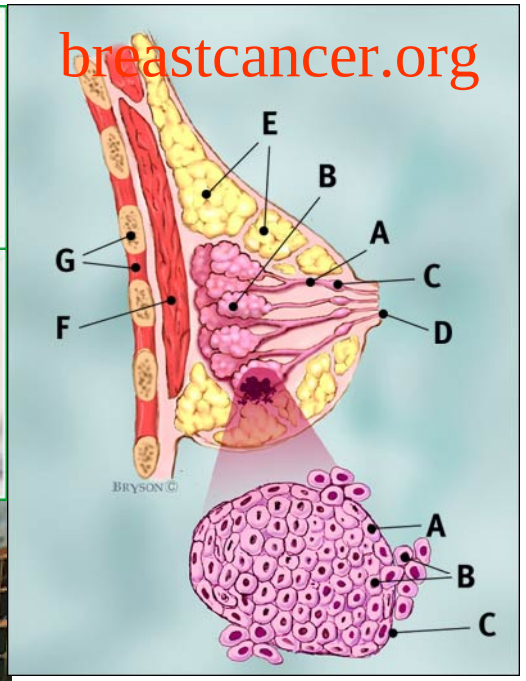
SCHOOL OF MEDICINE

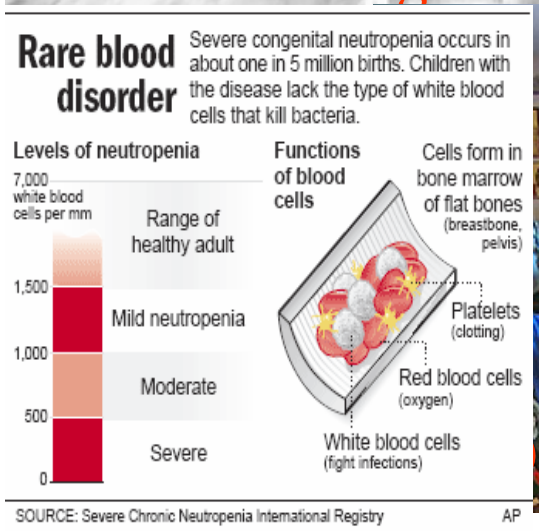
Genetics and Diseases



Dr. Doug Brutlag Lecture Syllabus “central paradigm” //www.s-star.org/

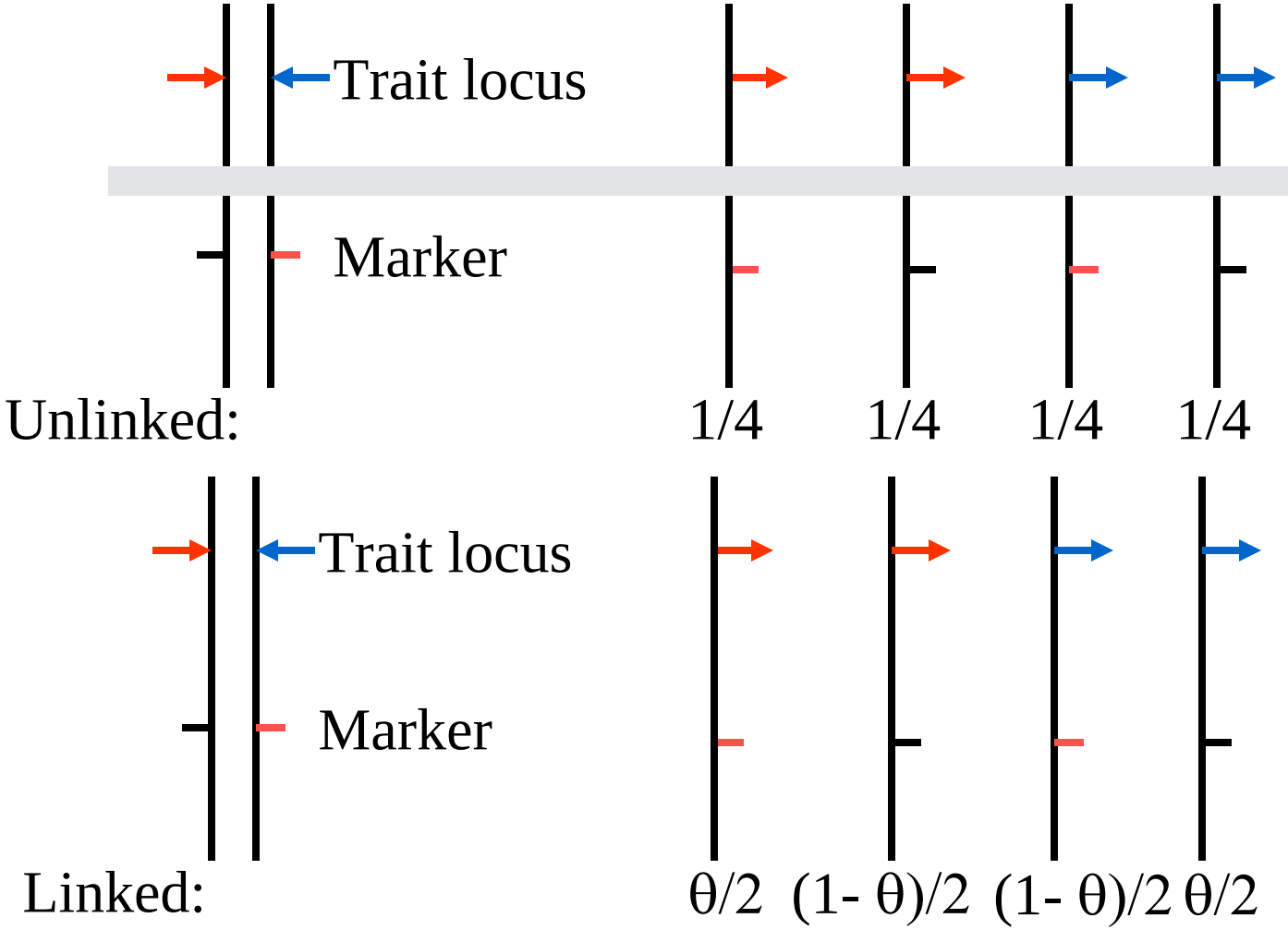
Human Conditions



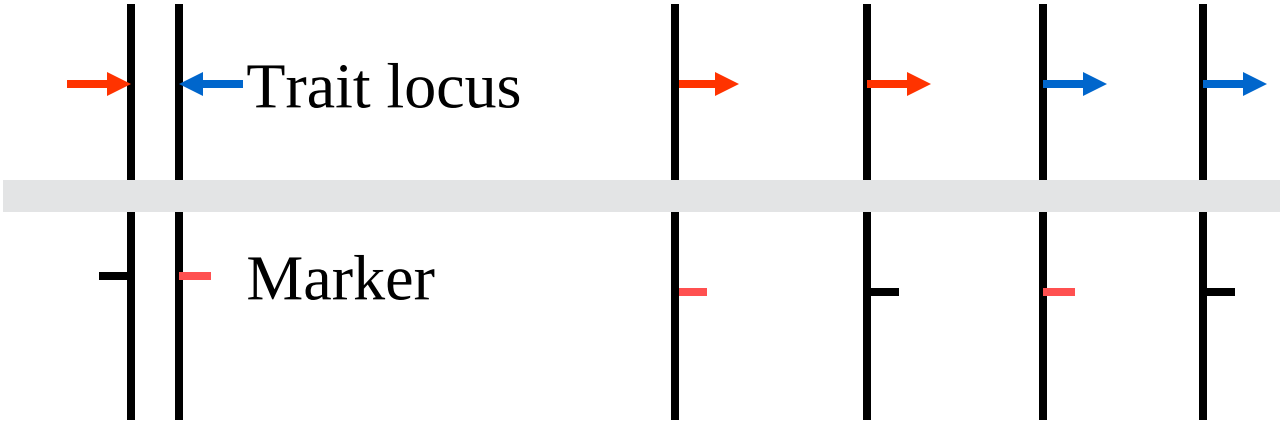


YALE UNIVERSITY
SCHOOL OF MEDICINE

Linkage - Recombination



Coefficient of Linkage Disequilibrium



Allele Frequency		Haplotype Frequency?	
→	d	→ , →	$d(1-m)$
→	$1-d$	→ , −	dm
−	m	− , →	$(1-d)(1-m)$
−	$1-m$	− , −	$(1-d)m$

Freq(→, −) - Freq(→)Freq(−)= δ

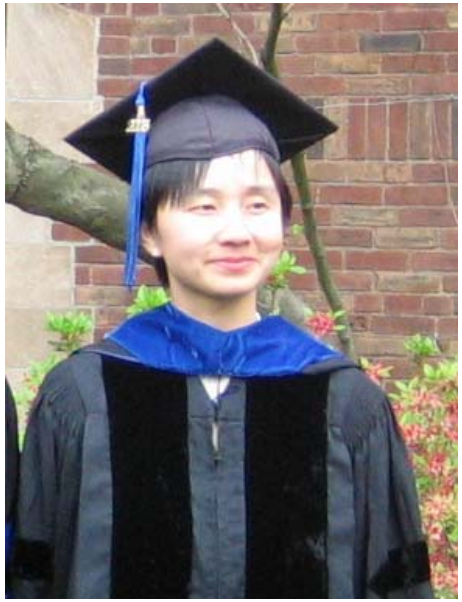
Null Hypothesis – Linkage Disequilibrium

The null hypothesis of haplotype relative risk (Falk and Rubinstein, 1987) being 1 is: $\delta(1-2\theta)=0$

TDT is to test for linkage in presence of association or test for association in presence of linkage (Spielman et al. 1993; Ewens and Spielman 1995).

Linkage Analysis

Use of latent variables



Rui Feng
Assistant Professor
University of Alabama

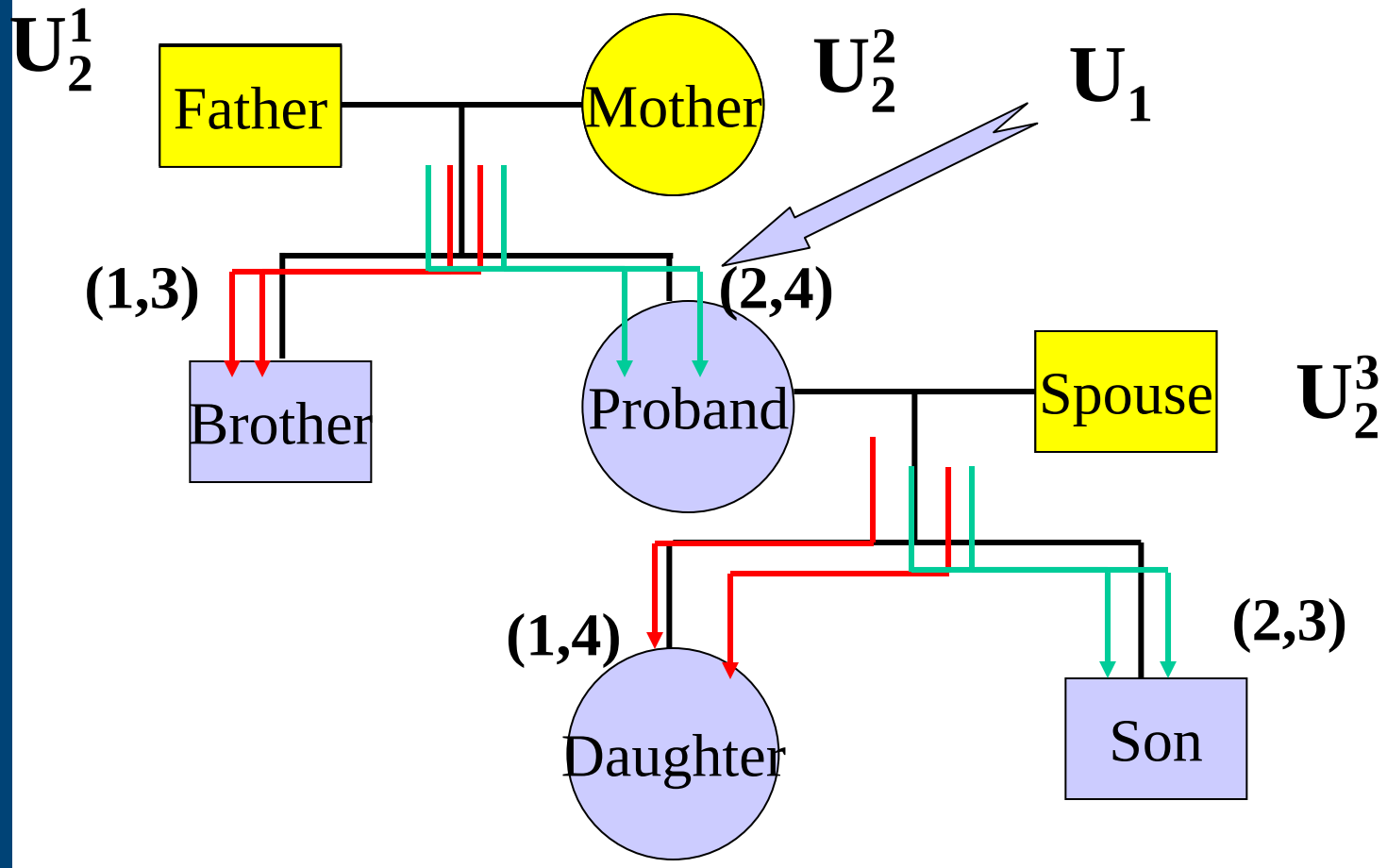
Hongtu Zhu
Associate Professor
University of North Carolina

YALE UNIVERSITY
SCHOOL OF MEDICINE

Inheritance Vector

(1,3,2,4,1,4,2,3)

1: Father's paternal allele
2: Father's maternal allele
3: Mother's paternal allele
4: Mother's maternal allele



YALE UNIVERSITY
SCHOOL OF MEDICINE

Model

$$P\{Y^i \mid U^i\} = \prod P\{Y_j^i \mid U^i\}$$

$$\text{logit}\ \ P\{Y_j^i = 0 \mid U^i\} = X_j^i \beta + \alpha_0 + \gamma a_j^i$$

$$\text{logit}\ \ P\{Y_j^i \le 1 \mid U^i\} = X_j^i \beta + \alpha_1 + \gamma a_j^i$$

$$a_j^i = (U_1^i, U_{2,2j-1}^i + U_{2,2j}^i, U_{2,2j-1}^i U_{2,2j}^i)^T$$

$$\gamma = (\gamma_1, \gamma_2, \gamma_3)$$

YALE UNIVERSITY
SCHOOL OF MEDICINE

Complete Likelihood $L_C(\theta, \gamma, \beta)$

$$\sum_{i=1}^{I}\left[\{U_1^i \log(p_1)+(1-U_1^i)\log(1-p_1)\}+\sum_{k=1}^{2m_i}U_{2,k}^i \log(p_2)+(2m_i-\sum_{k=1}^{2m_i}U_{2,k}^i)\log(1-p_2)\right]$$

$$+\sum_{i=1}^{I}\sum_{j=1}^{n_i}[I(Y_j^i=0)\log\pi_{0j}^i+I(Y_j^i=1)\log(\pi_{1j}^i-\pi_{0j}^i)+I(Y_j^i=2)\log(1-\pi_{1j}^i)],$$

where

$$\pi_{0j}^i=\frac{\exp(x_j^i\beta+\alpha_0+\gamma a_j^i)}{1+\exp(x_j^i\beta+\alpha_0+\gamma a_j^i)}$$

and

$$\pi_{1j}^i=\frac{\exp(x_j^i\beta+\alpha_1+\gamma a_j^i)}{1+\exp(x_j^i\beta+\alpha_1+\gamma a_j^i)}$$

YALE UNIVERSITY

SCHOOL OF MEDICINE

EM Algorithm

E-Step:

$$Q(p,\gamma,\beta;p^{(0)},\gamma^{(0)},\beta^{(0)}) = E\{\log[L_C(p,\gamma,\beta)]\,|\,Y,p^{(0)},\gamma^{(0)},\beta^{(0)}\}$$

M-Step: Maximizing

$$Q(p,\gamma,\beta;p^{(0)},\gamma^{(0)},\beta^{(0)})$$

and producing an updated

$$(p^{(1)},\gamma^{(1)},\beta^{(1)}).$$

YALE UNIVERSITY

SCHOOL OF MEDICINE

Test for Linkage

H_0: $\gamma_2 = 0$ (no linkage at locus t)

$$LR(p_2) = \frac{\max_{\omega,\gamma_2} \prod_i L_*^i(\omega, \gamma_2, p_2)}{\max_{\omega} \prod_i L_*^i(\omega, \gamma_2 \mid \gamma_2 = 0)}$$

p_2 is the frequency of the underlying disease allele.

Comparison with GENEHUNTER/ALLEGRO

Simulation Experiment Design

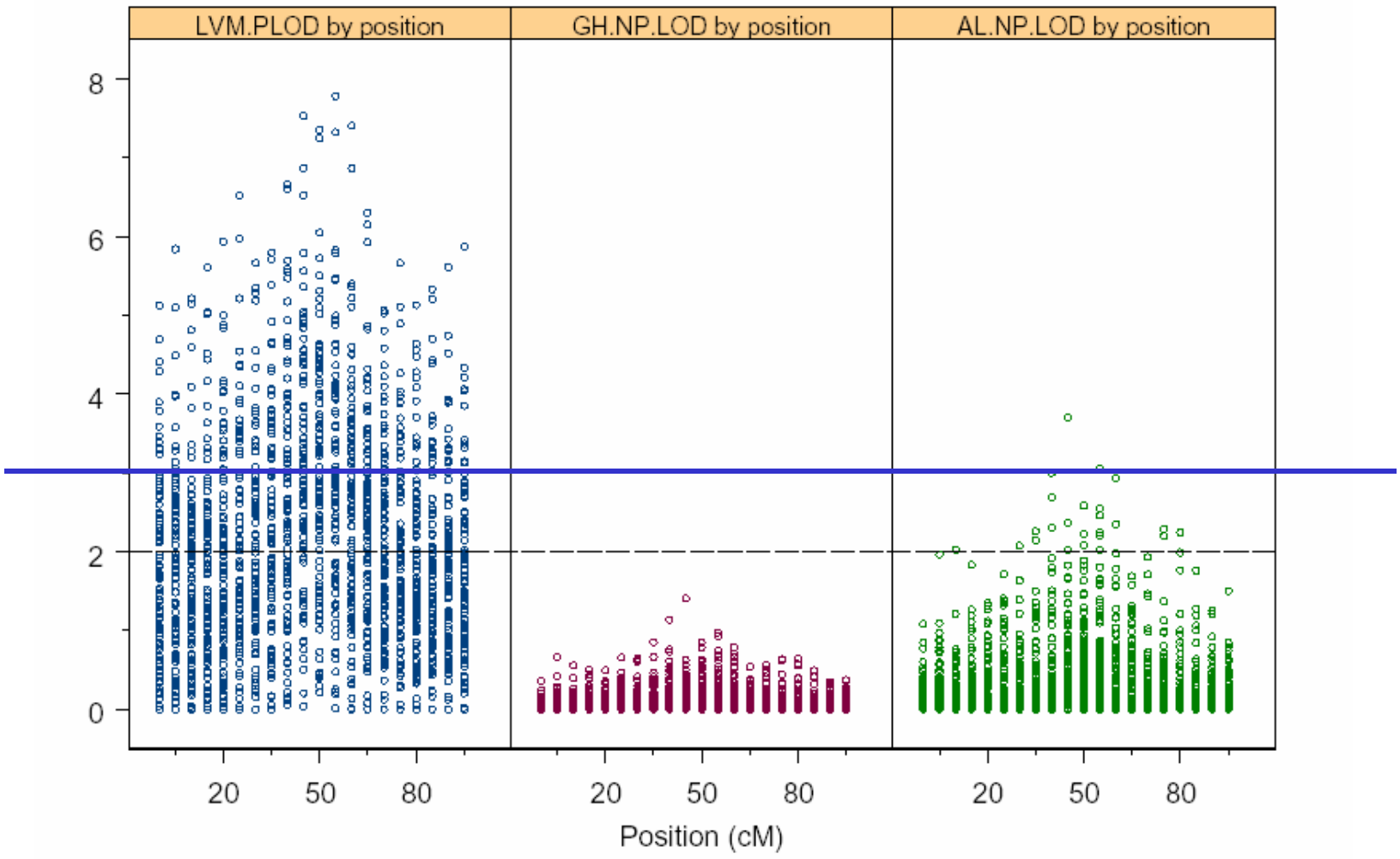
10 markers, 5 cM apart, with 10 alleles each, except that the 3rd one is the di-allelic disease gene.

200 families; 5 members;

$\theta_2 = 0.3$, $\alpha_0 = -2.0$, $\alpha_1 = -1.0$, $\gamma_2 = 2.0$

100 Replications

Power Comparison



Hoarding

Hoarding is a component of obsessive-compulsive disorder. It is the excessive collection and retention of things or animals until they interfere with day-to-day functions such as home, health, family, work and social life.

http://www.sciencentral.com

YALE UNIVERSITY
SCHOOL OF MEDICINE

Linkage Analyses for Hoarding

Dataset:

We used data from 223 individuals in 51 families with 77 sib pairs.

Genotypes are allele sizes from 370 markers on 22 chromosomes.

YALE UNIVERSITY

SCHOOL OF MEDICINE

Linkage Evidence for Hoarding

	p-values in analysis of.				
	Parametric		Nonparametric		
Marker (location in cM)	LMV	GH	LMV	GH	AL
4q34-35:					
DS42431 (163.26)	.006	.101	.001	.120	.156
D4S2417 (169.00)	.005	.072	.0009	.154	.192
D4S408 (182.13)	.012	.063	.006	.068	.091
D4S1652 (195.14)	.003	.040	.004	.126	.136
5q35.2-35.3:					
D5S1471 (172.13)	.003	.122	.001	.560	.563
D5S1456 (174.80)	.002	.139	.0003	.628	.640
D5SMfd154 (182.89)	.0006	.095	.00006	.299	.299
D5S408 (195.49)	.0002	.030	.00001	.133	.100
17q25:					
D17S1301 (99.39)	.005	.066	.0005	.052	.024
D17S784 (116.23)	.002	.034	.0006	.019	.007

YALE UNIVERSITY
SCHOOL OF MEDICINE

Challenges

Computation: Convergence and speed.
Concept: Inheritance vectors vs identity by decent.
Theory: Asymptotic distributions of the test statistics.

YALE UNIVERSITY
SCHOOL OF MEDICINE

References

• Zhang, Feng, and Zhu. A latent variable model of segregation analysis for ordinal trait. JASA, 98: 1023-1034, 2003.

• Zhu and Zhang. Hypothesis testing in mixture regression models. JRSS-B, 66: 3-16, 2004.

• Feng, Leckman, and Zhang. A latent variable model for linkage analysis of ordinal traits and a genome-wide scan of hoarding. PNAS, 101: 16739-16744, 2004 .

• Zhu and Zhang. Generalized score test of homogeneity for mixed effects models. Revised for Ann Statist, 34: 1545–1569, 2006.

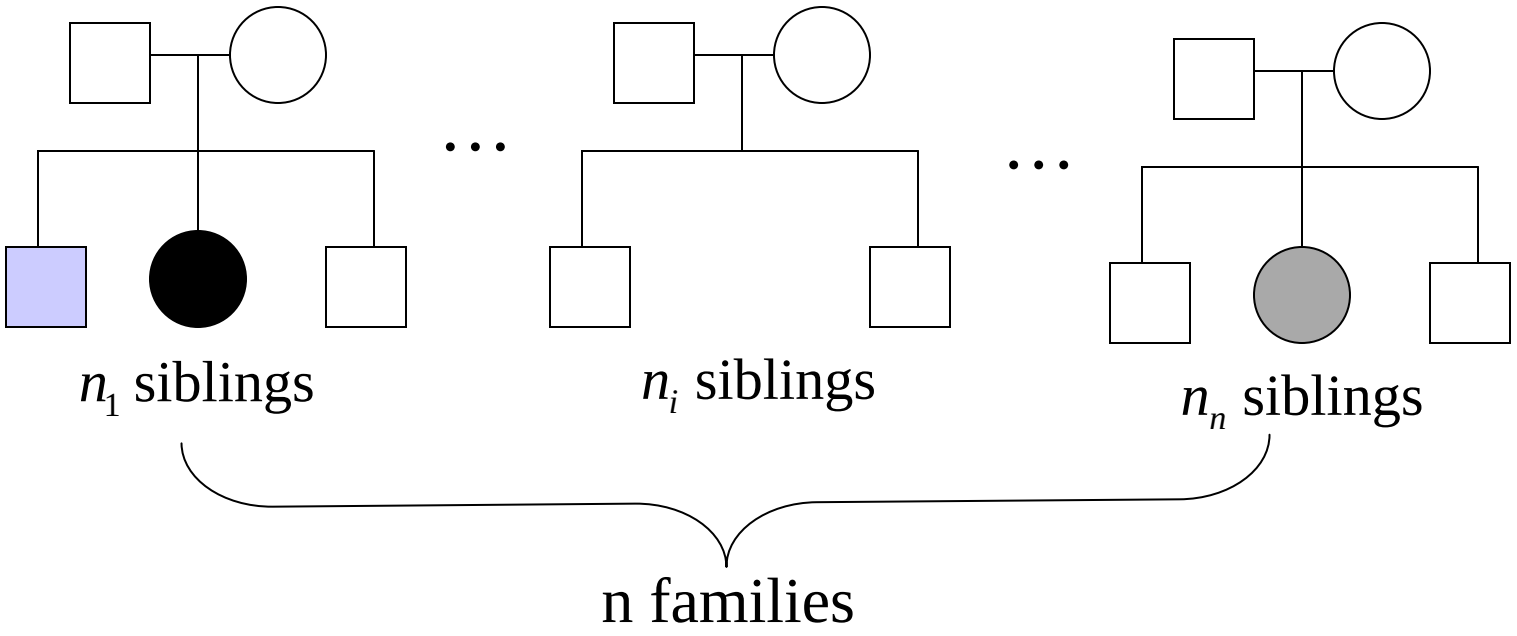
YALE UNIVERSITY

SCHOOL OF MEDICINE

Association Analysis

Xueqin Wang and Yuanqing Ye

Data Structure



Transmission/Disequilibrium Test (TDT)

- Eliminate the confounding effects caused by population stratification/admixture, and other factors
- A McNemar's test

Further Developments

- Q-TDT proposed by Allison (1997)
- Q-TDT further investigated by Rabinowitz (1997)
- S-TDT (Spielman and Ewens 1998)
- FBAT (Lunetta et al. 2000; Rabinowitz and Laird 2000)
- Many other extensions

General Test Statistic

Assume that there are n nuclear families. In the i^{th} family, there are n_i siblings, i=1,…, n. For the j^{th} child in the i^{th} family, the trait value is y_{ij} , the covariates is z_{ij} and the genotype is g_{ij} . X_{ij} is the number of allele A in the genotype g_{ij} . The association test statistic can be constructed as follows:

$$T = \sum_{i=1}^{n} T_i = \sum_{i=1}^{n} \sum_{j=1}^{n_i} W_{ij} X_{ij},$$

where W_{ij} is a weight function of y_{ij} and z_{ij}.

Model and Method

- Di-allelic maker with possible alleles A and a.
- Assume that there is a trait increasing allele , and

we use to denote the wild type allele(s)

- Consider a trait taking values in ordinal responses

$1,\ldots, K$.

$$\text{Model}: \ \text{logit}(P(Y \le k \mid g)) = \alpha_k + \beta I(g) + \alpha' z,$$
$$k = 1,\ldots, K-1,$$

where α_k's are level parameters, and β is genetic effect.

$I(g)$ is the number of copies of allele D in genotype g.

YALE UNIVERSITY
SCHOOL OF MEDICINE

Conditional Likelihood

The conditional likelihood for the sibling's genotypes given the observed sibling's phenotypes, covariates and parental genotype in the *i*th family :

$$P\{M_i \mid y_i, z_i, M_i^P\} = P\{M_i, M_i^P\} \sum_g \frac{P\{y_i \mid g, z_i\}}{P\{y_i, M_i^P \mid z_i\}} P\{g \mid M_i\}.$$

Score Function

The score function

$$\frac{\partial}{\partial \beta} \log(P\{M_i \mid y_i, z_i, M_i^P\})|_{\beta=0}$$

$$= \frac{\delta}{P\{A\}(1-P\{A\})} \sum_j [1-\gamma(y_{ij})-\gamma(y_{ij}-1)](X_{ij}-2P\{A\})$$

YALE UNIVERSITY
SCHOOL OF MEDICINE

Score Statistic

The score function under the null hypothesis is $T - E(T \mid Y, M^P)$, where

$$T = \sum_{i=1}^{n} \sum_{j=1}^{n_i} w(y_{ij}, z_{ij}) X_{ij},$$

$$w(k, z) = 1 - \hat{\gamma}(k, z) - \hat{\gamma}(k-1, z)$$

$$\hat{\gamma}(k, z) = \frac{\exp(\hat{\lambda}_k + \hat{\alpha}' z)}{1 + \exp(\hat{\lambda}_k + \hat{\alpha}' z)}, k = 1, \cdots K-1$$

$$\hat{\gamma}(0, z) \equiv 0 \quad \hat{\gamma}(K, z) \equiv 1$$

Expectation and Variance

Following the idea of Rabinowitz and laird (2000), we can compute or estimate the conditional expectation and the conditional variance given the observed trait values under null hypothesis in the following three cases:

(a) both parental marker information is available;

(b) only one of parental marker information is available; and

(c) none of parental marker information is available.

Expectation and Variance

$$E\{T \mid Y\} = \sum_{i,j} [R^{+}(y_{ij}) - R^{-}(y_{ij})] E\{X_{ij}\}$$

$$Var\{T \mid Y\} = \sum_{i} \{ \sum_{j} [R^{+}(y_{ij}) - R^{-}(y_{ij})]^2 Var\{X_{ij}\}$$

$$+ 2 \sum_{k>j} [R^{+}(y_{ij}) - R^{-}(y_{ij})][R^{+}(y_{ik}) - R^{-}(y_{ik})] Cov(X_{ij}, X_{ik}) \}$$

YALE UNIVERSITY
SCHOOL OF MEDICINE

Both Parents Genotyped

When both parents' genotypes are observed, the children's genotypes are conditionally independent.

Parental Genotypes	Expectation	Variance
(AA, AA)	2	0
(AA, Aa)	3/2	1/4
(AA, aa)	1	0
(Aa, Aa)	1	1/2
(Aa, aa)	1/2	1/4
(aa, aa)	0	0

One Parent Genotyped

Parental Genotype	Children's Possible Genotypes	Cond. Probability			Joint Conditional Genotype Distribution of Two Sibs
		AA	Aa	aa	
AA	{AA}	1			P{AA, AA}=1
	{Aa}		1		P{Aa, Aa}=1
	{AA, Aa}	1/2	1/2		P{AA, Aa}=$2^{n-2}/(2^n-2)$ P{AA, AA}=P(Aa, Aa}=$(2^{n-2}-1)/(2^n-2)$
aa	{Aa}		1		P{Aa, Aa}=1
	{aa}			1	P{aa, aa}=1
	{Aa, aa}				P{AA, Aa}=$2^{n-2}/(2^n-2)$ P{Aa, Aa}=P{aa, aa}=$(2^{n-2}-1)/(2^n-2)$
Aa	{AA}	1			P{AA, AA}=1
	{Aa}		1		P{Aa, Aa}=1
	{aa}			1	P{aa, aa}=1
	{AA, Aa}	n_{AA}/n	n_{Aa}/n		P{AA, AA}=$n_{AA}(n_{AA}-1)/n(n-1)$ P(Aa, Aa}=$n_{Aa}(n_{Aa}-1)/n(n-1)$ P{AA, Aa}=$n_{AA}n_{Aa}/n(n-1)$

One Parent Genotyped (continued)

Parental Genotype	Children's Possible Genotypes	Cond. Probability			Joint Conditional Genotype Distribution of Two Sibs
		AA	Aa	aa	
Aa	{Aa, aa}		n_{Aa}/n	n_{aa}/n	P{Aa, Aa}=$n_{Aa}(n_{Aa}-1)/n(n-1)$ P(aa, aa}= $n_{aa}(n_{aa}-1)/n(n-1)$ P{Aa, aa}=$n_{Aa}n_{aa}/n(n-1)$
	{AA,aa}	$\frac{4^{n-1}-3^{n-1}}{4^n-2\cdot3^n+2^n}$	$\frac{2\cdot4^{n-1}-4\cdot3^{n-1}+2^n}{4^n-2\cdot3^n+2^n}$	$\frac{4^{n-1}-3^{n-1}}{4^n-2\cdot3^n+2^n}$	P{AA, AA}=P{aa, aa}= P{AA, Aa}/2= P{Aa, aa}/2 = $(4^{n-2}-3^{n-2})/(4^n-2\cdot3^n+2^n)$ P{AA, aa}= $\frac{4^{n-2}}{4^n-2\cdot3^n+2^n}$ P{Aa, Aa}= $\frac{4^{n-1}-8\cdot3^{n-2}+2^n}{4^n-2\cdot3^n+2^n}$
	{AA, Aa, aa}				

$n_{AA,}$ $n_{Aa,}$ and n_{aa} is the number of children with genotype AA, Aa, and aa, respectively.

No Parental Genotype

Children's Possible Genotypes	Cond. Probability			Joint Conditional Genotype Distribution of Two Sibs
	AA	Aa	aa	
{AA}	1			P{AA, AA}=1
{Aa}		1		P{Aa, Aa}=1
{aa}			1	P{aa, aa}=1
{AA, Aa}	n_{AA}/n	n_{Aa}/n		P{AA, AA}= $n_{AA}(n_{AA}-1)/n(n-1)$ P(Aa, Aa}=$n_{Aa}(n_{Aa}-1)/n(n-1)$ P{AA, Aa}= $n_{AA}n_{Aa}/n(n-1)$
{Aa, aa}		n_{Aa}/n	n_{aa}/n	P{Aa, Aa}=$n_{Aa}(n_{Aa}-1)/n(n-1)$ P(aa, aa}= $n_{aa}(n_{aa}-1)/n(n-1)$ P{Aa, aa}= $n_{Aa}n_{aa}/n(n-1)$
{AA,aa}	$\frac{4^{n-1}-3^{n-1}}{4^n-2\cdot 3^n+2^n}$	$\frac{2\cdot 4^{n-1}-4\cdot 3^{n-1}+2^n}{4^n-2\cdot 3^n+2^n}$	$\frac{4^{n-2}}{4^n-2\cdot 3^n+2^n}$	P{AA,AA}=P{aa,aa}=P{AA,Aa}/2= P{Aa, aa}/2=$(4^{n-2}-3^{n-2})/(4^n-2\cdot 3^n+2^n)$ P{AA, aa}=$(4^{n-1}-3^{n-1})/(4^n-2\cdot 3^n+2^n)$ P{Aa, Aa}=$(4^{n-1}-8\cdot 3^{n-2}+2^n)/(4^n-2\cdot 3^n+2^n)$
{AA, Aa, aa}				

YALE UNIVERSITY
SCHOOL OF MEDICINE

Simulation Studies

- Assess the type I error of our score test with respect to specific nominal levels (0.05, 0.01, and 0.0001) to validate the asymptotic behavior of the test statistic.
- Compare the power of our test with other test statistics.
- Choose the ordinal level K=3 or 4.

Simulation Design: No Association

- Generate the parent's genotypes via the haplotype frequencies
- Given the parental genotypes, generate the offspring genotype using 1cM between the two loci.(Allow parental genotypes missing)
- Conditional on the trait genotype, using the proportional odd model to generate the ordinal trait.

Simulation Design: Power

In the power calculation, we assume $P(D \mid A) = p_{DA}$, where $P(D \mid A)$ denotes the conditional probability that the disease allele equals to D given that the marker allele equals to A.

The haplotype frequencies with P{D}=P{A}=0.3 and $\delta = 0.11$.

Haplotype	Frequency
AD	0.2
Ad	0.1
aD	0.1
ad	0.6

Ordinal Traits Generated from a Proportional Odds Model

The ordinal traits are generated by the proportional odds model

$$\text{logit}[P(Y \le k \mid t)] = \alpha_k - t + z, k = 1,..., K-1,$$

where t is the number of copies of trait increasing alleles, z ~ N(1,2).

Possible values of t are 0, 1, and 2, and $\alpha_k = 1 + k$, for $k = 1,..., K-1$

20% of paternal and 20% maternal genotypes missing

YALE UNIVERSITY
SCHOOL OF MEDICINE

Type I Errors Based on 10,000 Replications – Test for Association in the Presence of Linkage

# F	K	Sig. level	OTDT		QTDT		TDT	
			Adjusted	NA	Adjusted	NA	Adjusted	NA
200	3	0.05	0.04856	0.0498	0.05035	0.04975	0.04817	0.05137
		0.01	0.0093	0.00978	0.00958	0.00975	0.00907	0.01046
		0.001	0.00066	0.0008	0.00083	0.00079	0.00083	0.00101
	4	0.05	0.04999	0.04974	0.04989	0.05001	0.05008	0.05065
		0.01	0.00902	0.00919	0.00927	0.0091	0.00955	0.00978
		0.001	0.00076	0.00067	0.00067	0.00066	0.0009	0.00098
400	3	0.05	0.04938	0.04975	0.05002	0.04961	0.04927	0.0501
		0.01	0.00925	0.00937	0.00895	0.00935	0.00978	0.00932
		0.001	0.00084	0.00072	0.00087	0.00072	0.00076	0.00099
	4	0.05	0.04951	0.04883	0.04944	0.04958	0.04908	0.05071
		0.01	0.00951	0.00953	0.00982	0.00956	0.0093	0.01024
		0.001	0.00086	0.0009	0.00076	0.00084	0.0009	0.00109

Powers Based on 10,000 Replications – Test for Association in the Presence of Linkage

# F	K	Sig. level	OTDT		QTDT		TDT	
			Adjusted	NA	Adjusted	NA	Adjusted	NA
200	3	0.05	*0.40669*	0.25212	0.23343	0.25197	0.1961	0.10151
		0.01	*0.18538*	0.09435	0.08415	0.09449	0.06541	0.02924
		0.001	*0.04689*	0.01931	0.01706	0.01944	0.01161	0.00452
	4	0.05	*0.4531*	0.25183	0.23544	0.25044	0.18439	0.10145
		0.01	*0.22012*	0.09456	0.08616	0.09406	0.06177	0.02838
		0.001	*0.05961*	0.01807	0.01644	0.01797	0.01018	0.0042
400	3	0.05	*0.69601*	0.45971	0.42656	0.45922	0.34709	0.1588
		0.01	*0.4486*	0.23251	0.20676	0.23257	0.15489	0.05063
		0.001	*0.18874*	0.06997	0.05936	0.07004	0.03835	0.00895
	4	0.05	*0.77043*	0.50638	0.46085	0.50401	0.35082	0.15774
		0.01	*0.5405*	0.26796	0.2323	0.2655	0.15561	0.05243
		0.001	*0.25723*	0.08623	0.07072	0.08542	0.04044	0.00915

YALE UNIVERSITY
SCHOOL OF MEDICINE

Performance for Quantitative Traits

Our test can serve as a unified test for any trait. For quantitative trait, the weights in our test are the functions of quantiles. Simulations show that our test is competitive with, but slightly less powerful than Q-TDT.

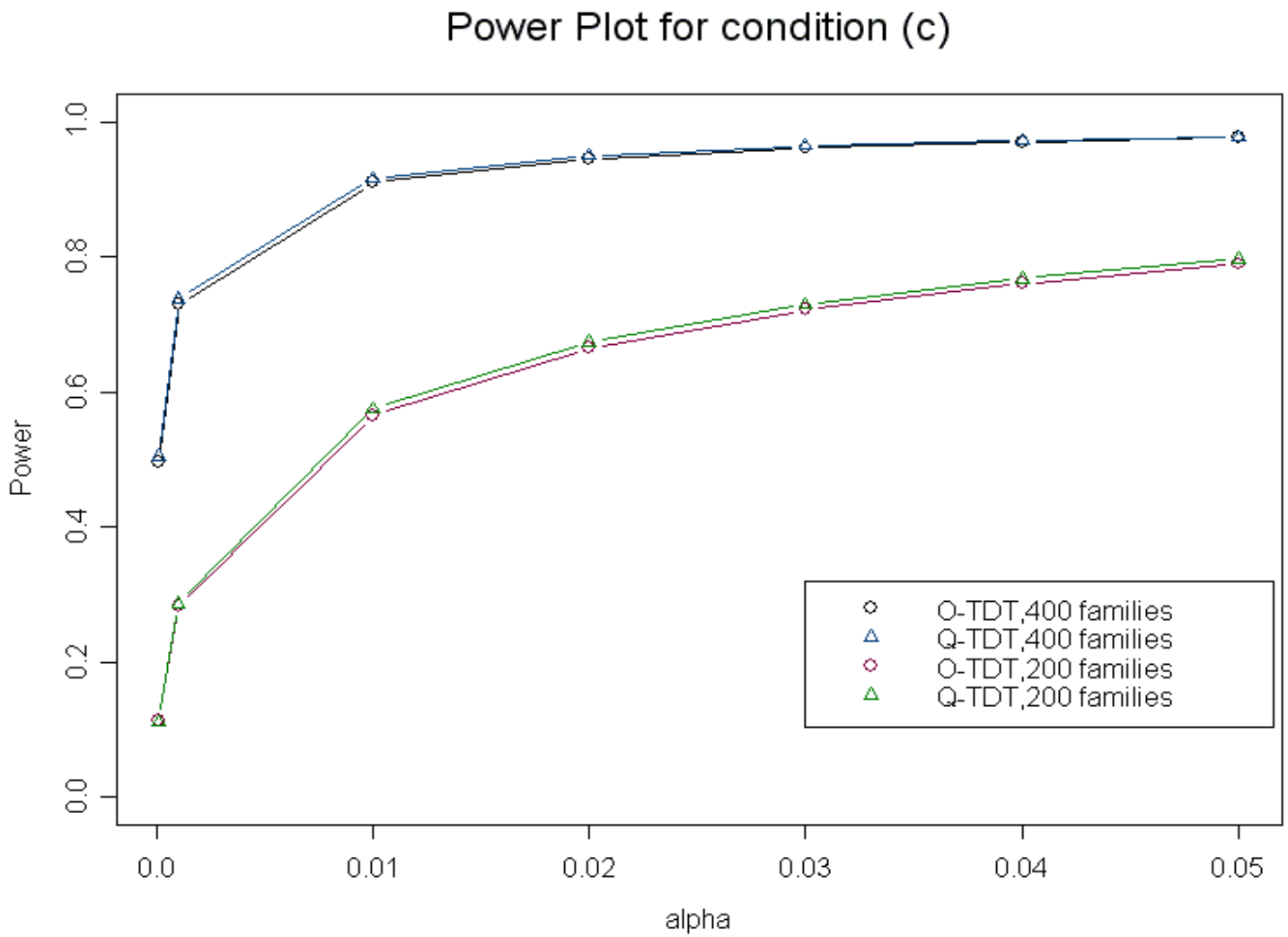
YALE UNIVERSITY
SCHOOL OF MEDICINE

Power: Quantitative Trait

Given the genotype at the trait locus, the quantitative trait follows the normal distribution with mean proportional to the number of the trait increasing allele and unit variance. Namely,

$$\mu_{dd} = 0,\ \mu_{Dd} = 1, \text{ and } \mu_{DD} = 2.$$

Figure: Power Comparison for QTL



YALE UNIVERSITY
SCHOOL OF MEDICINE

COGA data

- A large scale, multi-center study to map alcohol dependence susceptible genes.
- 143 families with 1614 individuals. 4720 SNPs from Illumina genotype data set.
- One ordinal trait with 4 levels was recorded (pure unaffected, never drank, unaffected with some symptoms, and affected).
- FBAT was also used for comparison

Results

SNP Markers That Are Significant at the 0.001 Level Based on O-TDT after Adjusting for Gender and Age					
SNP Markers	Chromosome	Physical location	P-values		Gene Names
			Gender and Age Adjusted	Un-adjusted	
rs1972373	14	18435498	0.000383	0.000165	
rs1571423	10	125256948	0.000456	0.000349	LOC440007
rs485874	1	18182512	0.000497	0.001006	
rs619	X	29916017	0.000548	0.077359	GK
rs718251	8	52437707	0.000671	0.010729	
rs1869907	15	38835904	0.000873	0.030668	

Results

SNP rs619 is from gene GK(glycerol kinase) in the chromosome region Xp21.3 (Baranzini et al. 1997, Fries et al. 1993, Pillers 1990). It is within a region in which Zhang et al. (2005) reported linkage to ALDX1 (p-value=0.004).

Results

SNP rs485874 is on chromosome 1. One of the genes tightly linked to this SNP is aldehyde dehydrogenase 4 family, member A1 (ALDH4). ALDH genes together with ADH enzymes play major role in the ethanol metabolism. Extensive studies have shown that ALDH2 is involved in the alcohol dependence (Crabb et al. 2004). Our study calls further attention to study the potential association between ALDH4 and alcohol dependence.

Results

SNP rs1869907 resides in a tightly linked region on chromosome 15 covering 27 known genes. One of genes in the region of SNP rs186907 is vacuolar protein sorting protein 18(VPS18) that mediates the vesicle trafficking, and a mutant of this gene has been found to be associated with liver disease (Sadler et al 2005).

Results

SNP rs1972373 is from chromosome 14 and it is tightly linked to a family of olfactory receptors genes, which can trigger smell signal to brain by reacting with odor molecules. How these genes are related to alcohol dependence warrants further investigation. Li et al. (2005) applied a rank-based association test using the age at onset of ALDX1 as the trait and detected significant association (p-value=0.0002) in the region of SNP rs1972373.

Results

SNP marker rs1571423 is in the chromosome region 10q26.13. It is between the two regions with linkage to ALDX1 identified by Zhang et. al (2005) on chromosome 10. Murray (2005) also reported linkage to ALDX1 at 10q26 using NPL multipoint analysis methods.

Discussion and Conclusion

- We propose a score test statistic for linkage analysis.
- Although it is derived from a proportional odds model for ordinal traits, power comparisons reveal that it can serve as a unified approach for dichotomous, quantitative, and ordinal traits.
- The score based Q-TDT test yields lower power than O-TDT for ordinal traits, but the difference ranges from a few to tens of percents, depending on the distribution of the ordinal traits.

References

• Zhang, Wang, and Ye. Detection of genes for ordinal traits in nuclear families and a unified approach for association studies. Genetics, 172: 693-699, 2006. .

• Wang, Ye, and Zhang. Family-based association tests for ordinal traits adjusting for covariates. Genetic Epidemiology, in press, 2006.

YALE UNIVERSITY

SCHOOL OF MEDICINE