

Detection of SNP-SNP Interactions in Case-Parent Trios

Ingo Ruczinski

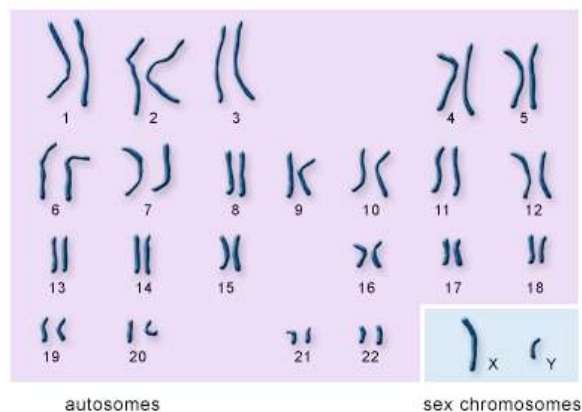
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

June 2, 2009

Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

Karyotypes

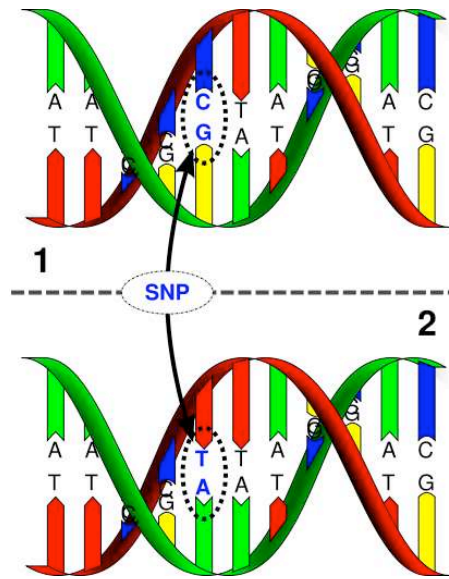


<http://ghr.nlm.nih.gov/>

Ingo Ruczinski

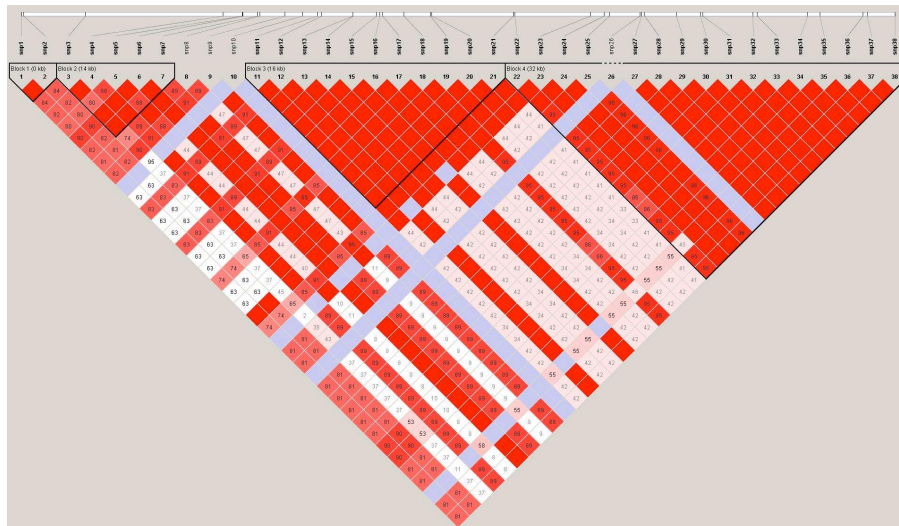
SNP-SNP Interactions in Case-Parent Trios

Single Nucleotide Polymorphisms



urgi.versailles.inra.fr

Haplotype Blocks



Kim and Dionne (2007)

SNP-SNP Interactions

Many statistical approaches have been used in the literature for the detection of SNP-SNP interactions:

- Regression with Higher Order Interactions
- Logic (Boolean) Regression and Monte Carlo Logic Regression
- Multifactor Dimensionality Reduction (MDR)
- Classification and Regression Trees (CART)
- Random Forests
- Boosting
- Multivariate Adaptive Regression Splines (MARS)
- Neural Networks

References:

Heidema AG et al (2006). *The Challenge for Genetic Epidemiologists: How to Analyze...* BMC Genetics 7: 23.
McKinney BA et al (2006). *Machine Learning for Detecting Gene-Gene Interactions...* Appl Bioinf 5(2): 77-88.
Musani SK (2007). *Detection of Gene x Gene Interactions in...* Hum Hered 63(2): 67-84.

Biological and Statistical Interactions

	BB	Bb	bb
AA			
Aa			
aa			

$$(\text{SNP}A^D \wedge \text{SNP}B^R) \vee (\text{SNP}B^D \wedge \text{SNP}A^R)$$

	BB	Bb	bb
AA			
Aa			
aa			

$$(\text{SNP}^{\text{A}^{\text{D}}} \wedge \text{SNP}^{\text{B}^{\text{R}}}) \vee (\text{SNP}^{\text{B}^{\text{D}}} \wedge \text{SNP}^{\text{A}^{\text{R}}})$$

→ Statistical interaction:

Deviation from additivity in a linear statistical model.

→ Epistasis:

Masking of phenotype expressed by one gene by the effects of another gene.

Reference: Moore JH (2005). *A Global View of Epistasis*. Nature Genetics, 37: 13-4.

Introduction

“Current methods for analyzing complex traits include analyzing and localizing disease loci one at a time. However, complex traits can be caused by the interaction of many loci, each with varying effect.”

“... patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, it becomes overwhelmingly computationally burdensome to analyze all 3-way, 4-way to N-way 'and' patterns, 'or' patterns, and combinations of loci.”

Reference: Lucek PR, Ott J (1997). *Neural Network Analysis of Complex Traits*. Genet Epi 14(6):1101-6.

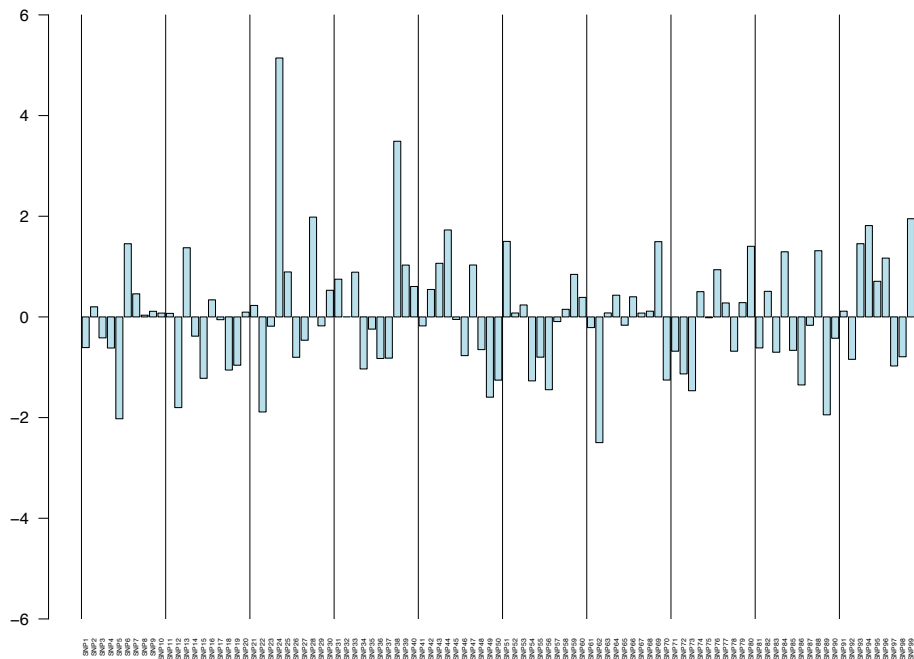
Introduction

- Assume we have a candidate gene study with 1000 cases of a certain disease, and 1000 controls. We typed 100 SNPs of interest, and we were lucky - the two causal SNPs were typed.
- Also assume that a double penetrance model describes the relationship between SNPs and phenotype: subjects with two variant alleles of either SNP 24 or SNP 80 and at least one variant allele of the other have higher odds of disease:

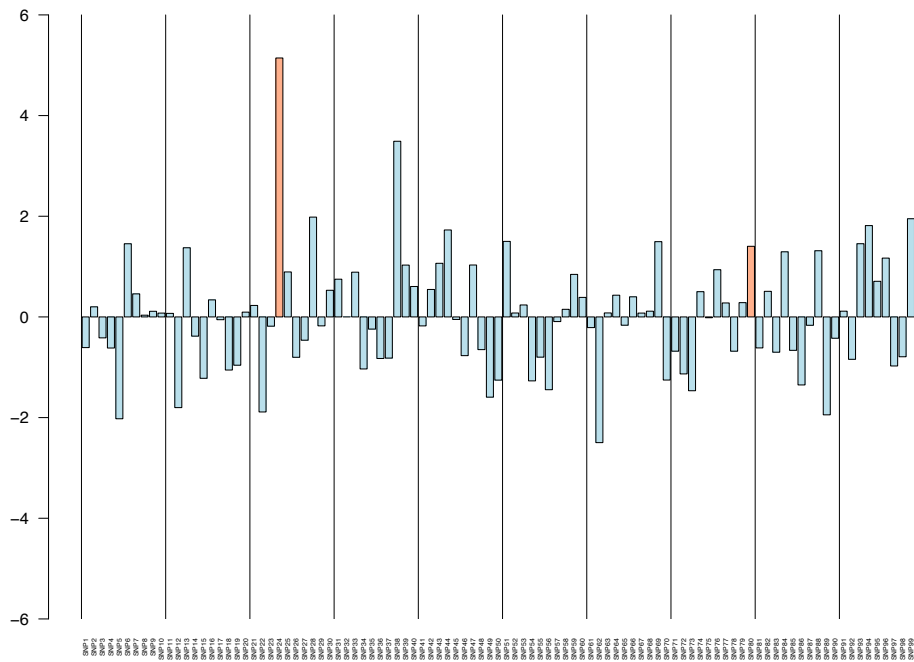
$$\text{logit}(p) = \alpha + \beta \times \text{Ind}\{(\text{SNP}24^D \wedge \text{SNP}80^R) \vee (\text{SNP}24^R \wedge \text{SNP}80^D)\}$$

- How could we detect this signal?

Introduction



Introduction



Logic Regression

- The predictors are the SNPs in dominant and recessive coding.

SNP X	X.R	X.D
AA	0	0
AT	0	1
TT	1	1

- Let X_1, \dots, X_k be binary (0/1) predictors, Y a response variable.
- Fit a model

$$g(E(Y)) = b_0 + \sum_{j=1}^t b_j \cdot L_j,$$

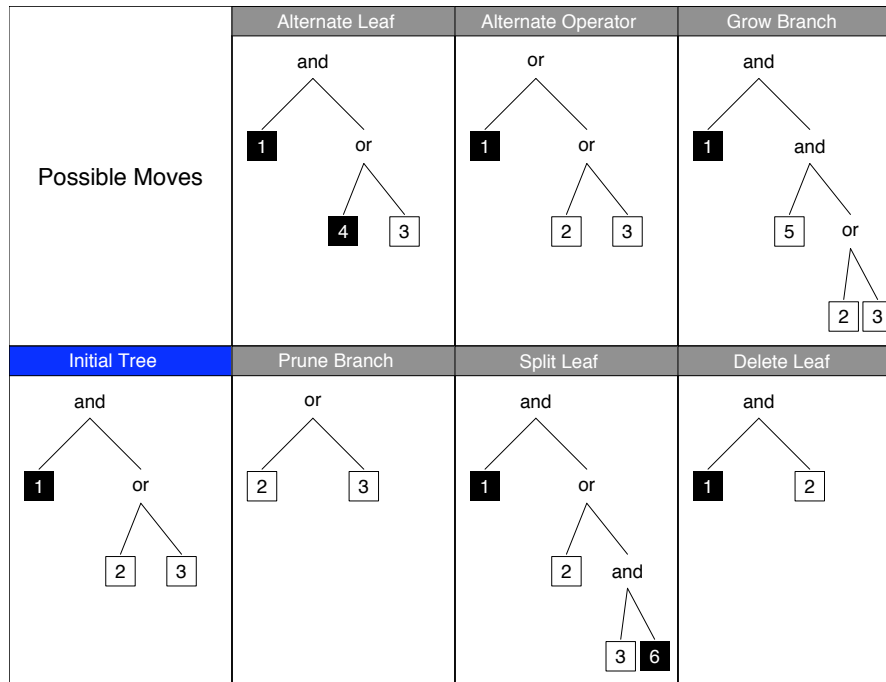
where L_j is a Boolean combination of the covariates, e.g.

$$L_j = (X_1 \vee X_2) \wedge X_4^c.$$

- Determine the logic terms L_j and estimate the b_j simultaneously.

Reference: Ruczinski et al (2003). *Logic Regression*. J of Comp Graph Stat, 12(3): 475-511.

Logic Regression

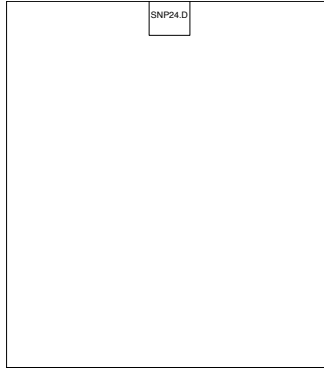


Logic Regression

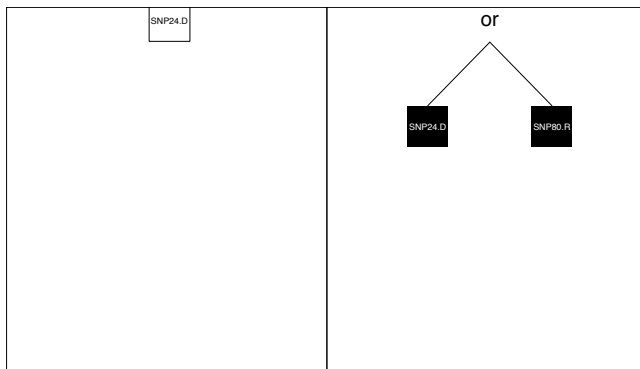
We try to fit the model $g(E(Y)) = b_0 + \sum_{j=1}^t b_j \cdot L_j$.

- Select a scoring function (RSS, log-likelihood, ...).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leaves in a tree.
- Initialize the model with $L_j = 0$ for all j .
- Carry out the Simulated Annealing Algorithm:
 - Propose a move.
 - Accept or reject the **move** depending on the scores and the temperature.

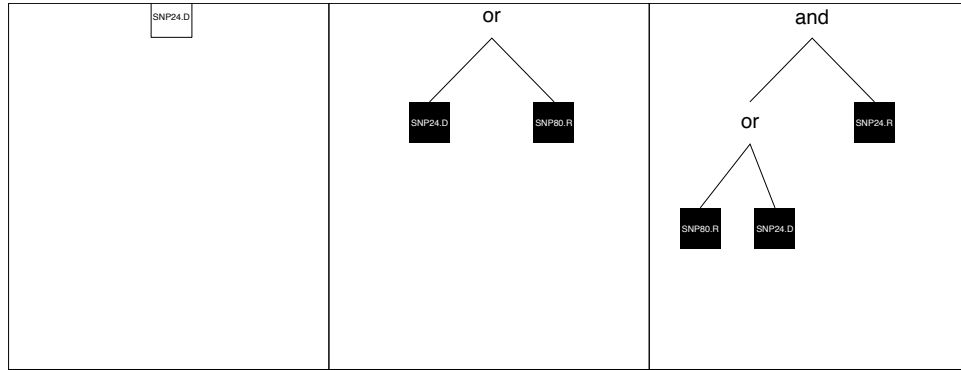
Logic Regression



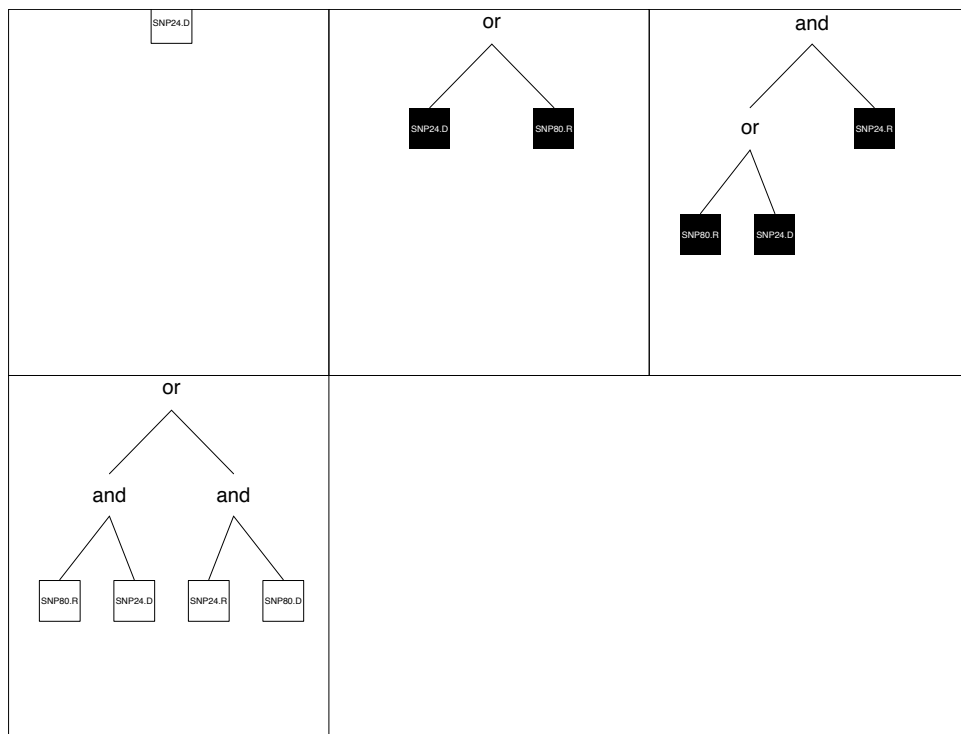
Logic Regression



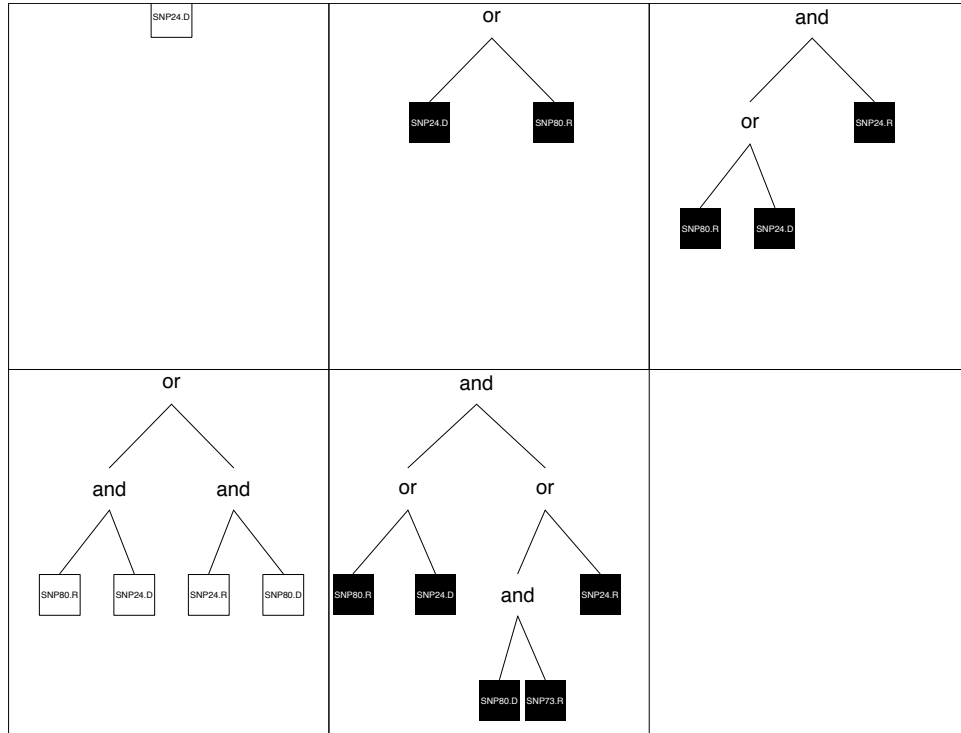
Logic Regression



Logic Regression



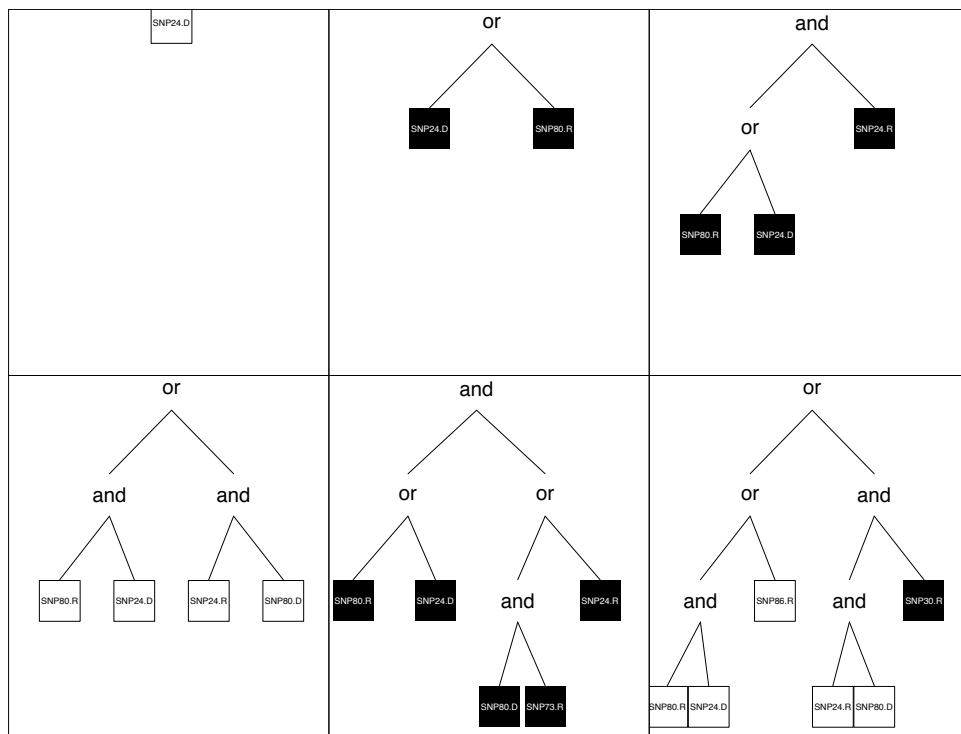
Logic Regression



Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

Logic Regression



Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

We implemented two flavors for the required model selection. Both approaches require a definition of [model size](#).

- 1 Cross-validation
This is most applicable when prediction is the main objective, i. e. not necessarily the best option in SNP association studies.
- 2 Permutation tests
This is a test for association, i. e. the preferred test in SNP association studies. The model size is chosen via a sequence of hypothesis tests.

Schizophrenia

- Diagnosis and classification of Schizophrenia (SZ) depends on the observation of disease related symptoms: delusion, hallucination, bizarre behavior, disorganized speech, . . .
- About 1% prevalence worldwide, equal in males and females.
- Schizoaffective disorder (SZA) is closely related to SZ.
- Over 2000 papers on linkage and association studies.
- About 150 genes implicated, most on chromosomes 1, 6, and 22. Most findings are somewhat inconclusive, though.
- Most studies are single marker analysis only, very few address SNP-SNP interactions.

Schizophrenia Study

- Case-parent trios of Ashkenazi Jewish descents.
- Diagnosis of SZ and SZA based on DSM IV.
- Dense coverage of 64 candidate genes.
- 375 SNPs on 11 chromosomes genotyped for 312 trios.
- Original analysis through single marker TDT.

Goal:

Explore SNP-SNP interactions for association with SZ and SZA.

Reference: Fallin et al (2005). *Bipolar I disorder and schizophrenia...* Am J Hum Genet 77(6):918-36.

TDT - Allelic

The transmission disequilibrium test measures the over-transmission of an allele from parents to affected offsprings. For a set of n parents with alleles 1 and 2 at a genetic locus, each parent can be summarized by the transmitted and the non-transmitted allele:

		Non-TA		Σ
		1	2	
TA	1	a	b	a + b
	2	c	d	c + d
Σ		a + c	b + d	2n

Under the null of no association, $\frac{(b-c)^2}{b+c} \sim \chi_1^2$

Only the heterozygous parents contribute information!

TDT - Genotypic

Assume that at a certain locus the father has alleles 11 and the mother has alleles 12. The four *Mendelian children* thus have alleles 11, 12, 11, and 12.

Assume the affected proband has genotype 11.

The three *Pseudo controls* then have the genotypes 11, 12, and 12.

	Y	X
Affected proband	1	11
Pseudo control #1	0	11
Pseudo control #2	0	12
Pseudo control #3	0	12

We can use conditional logistic regression to analyze the data!

Outline

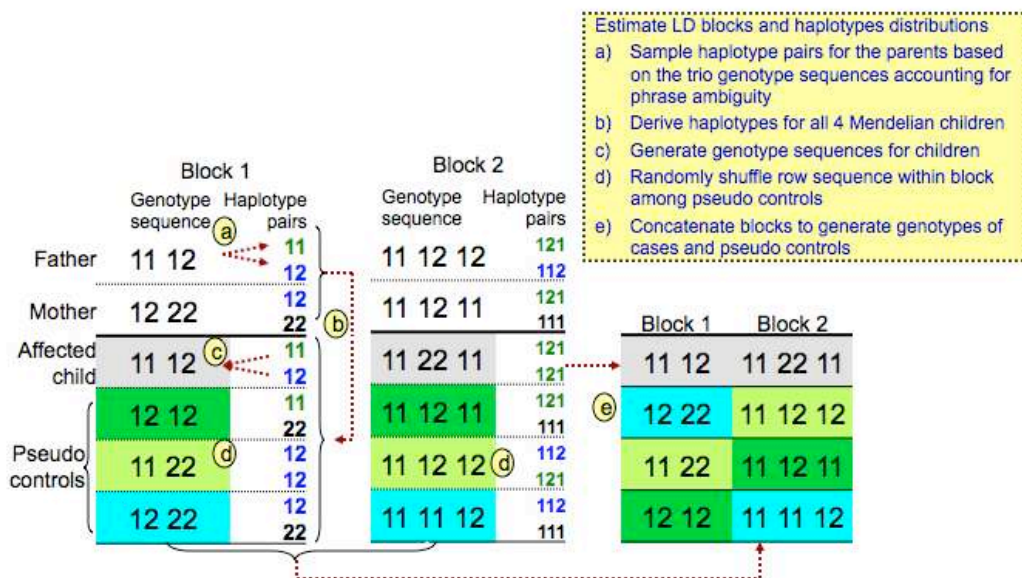
- 1 We extended the logic regression methodology to analyze trios with affected probands.
- 2 We developed an R package to
 - a) impute missing genotypes, using a haplotype-based approach that employs mating tables, and to
 - b) simulate case-parent trios with disease risk dependent on SNP-SNP interactions.
- 3 We carried out simulation studies to verify the validity of the methodology and the software.
- 4 We analyzed the data from the study of Schizophrenia among Ashkenazi Jewish families.

Trio Logic Regression

The rough idea is as follows:

- Pseudo controls are generated from the trio data, taking the LD block structure into account.
- Missing data are handled using haplotype-based imputation.
- The conditional logistic likelihood is used in logic regression to assess differences in cases and pseudo controls.

Trio Logic Regression



Trio Logic Regression

The steps in more detail:

- 1 Estimate the haplotype blocks and the haplotype frequencies using the parents' genotypes.
- 2 For each block and each trio, sample haplotype pairs for the parents and the offspring consistent with the observed genotypes in the trio, allowing for missing data.
- 3 Generate the probands genotype data from the haplotypes that were passed from the parents.
- 4 For each block and each trio, generate genotypes for three pseudo-controls (PC1, PC2, PC3) using the parents' haplotypes that were not passed to the proband. The assignment to PC1, PC2, and PC3 is random.
- 5 Assemble three pseudo-controls for each trio by augmenting the genotypes from the blocks.
- 6 For each locus, translate the genotype data into two binary variables in dominant and recessive coding.

Trio Logic Regression

The conditional logistic likelihood is

$$L(\beta) = \prod_{i=1}^N \left(\frac{\exp(X_{i0}\beta)}{\exp(X_{i0}\beta) + \sum_{m=1}^3 \exp(X_{im}\beta)} \right)$$

In this setting, i refers to a trio, X_{i0} refers to the exposure of the proband in trio i , and (X_{i1}, X_{i2}, X_{i3}) are the exposures of the 3 pseudo-controls in trio i .

→ Does the likelihood always have a maximum?

Trio Logic Regression

Number of trios where k pseudo-controls match the proband				
$k \rightarrow$	0	1	2	3
$X_{i0} = 0$	a_0	b_0	c_0	d_0
$X_{i0} = 1$	a_1	b_1	c_1	d_1

Likelihood contributions for $X_{i0} = 0$				
	a_0	b_0	c_0	d_0
X_{i0}	0	0	0	0
X_{i1}	1	0	0	0
X_{i2}	1	1	0	0
X_{i3}	1	1	1	0
$L_i(\beta)$	$\frac{1}{1+3 \exp(\beta)}$	$\frac{1}{2+2 \exp(\beta)}$	$\frac{1}{3+\exp(\beta)}$	$\frac{1}{4}$

Likelihood contributions for $X_{i0} = 1$				
	a_1	b_1	c_1	d_1
X_{i0}	1	1	1	1
X_{i1}	0	1	1	1
X_{i2}	0	0	1	1
X_{i3}	0	0	0	1
$L_i(\beta)$	$\frac{\exp(\beta)}{\exp(\beta)+3}$	$\frac{\exp(\beta)}{2 \exp(\beta)+2}$	$\frac{\exp(\beta)}{3 \exp(\beta)+1}$	$\frac{1}{4}$

Trio Logic Regression

The log-likelihood is:

$$\begin{aligned}
 \log(L(\beta)) = & a_1 \times \{\beta - \log(\exp(\beta) + 3)\} + \\
 & b_1 \times \{\beta - \log(2 \exp(\beta) + 2)\} + \\
 & c_1 \times \{\beta - \log(3 \exp(\beta) + 1)\} + \\
 & \frac{d_1}{4} - \\
 & a_0 \times \log\{1 + 3 \exp(\beta)\} - \\
 & b_0 \times \log\{2 + 2 \exp(\beta)\} - \\
 & c_0 \times \log\{3 + \exp(\beta)\} - \\
 & \frac{d_0}{4}
 \end{aligned}$$

Trio Logic Regression

The first derivative of the log-likelihood is:

$$\begin{aligned}\frac{\partial \log(L(\beta))}{\partial \beta} &= a_1 - (a_1 + c_0) \times \frac{\exp(\beta)}{\exp(\beta) + 3} + \\ & b_1 - (b_1 + b_0) \times \frac{2 \exp(\beta)}{2 \exp(\beta) + 2} + \\ & c_1 - (c_1 + a_0) \times \frac{3 \exp(\beta)}{3 \exp(\beta) + 1}\end{aligned}$$

The log-likelihood is monotonically decreasing in β . Further:

$$\lim_{\beta \rightarrow -\infty} \frac{\partial \log(L(\beta))}{\partial \beta} = a_1 + b_1 + c_1 \geq 0$$

and

$$\lim_{\beta \rightarrow +\infty} \frac{\partial \log(L(\beta))}{\partial \beta} = -(a_0 + b_0 + c_0) \leq 0$$

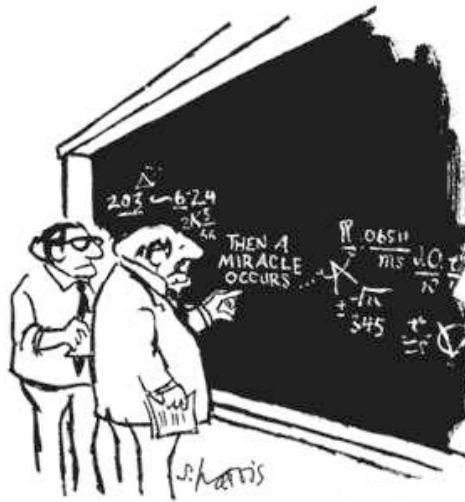
Trio Logic Regression

Since the derivative of the log likelihood function is a continuous function in β , it follows that the likelihood function has a maximum unless $a_1 + b_1 + c_1 = 0$ or $a_0 + b_0 + c_0 = 0$.

In other words, for either the exposure or the non-exposure groups, all pseudo-controls are equal to the respective probands.

→ This might occur in a particular sample where the number of trios is small and the (or some of the) SNPs contributing to the exposure definition have extremely low minor allele frequency, however, it is not plausible that such a separation would exist in the population in truth.

Missing Data

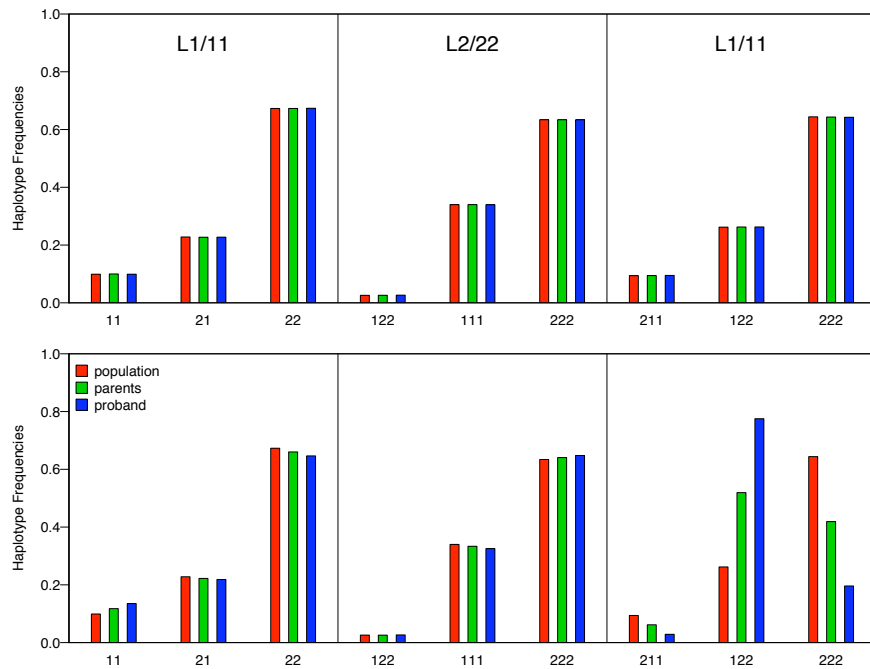


"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Simulation

	G			
1	$SNP_{13 1}^R$			
2	$SNP_{4 2}^R \wedge SNP_{5 3}^R$			
3	$(SNP_{1 1}^R \wedge \overline{SNP_{6 2}^D}) \vee SNP_{13 1}^R$			
4	$(SNP_{2 3}^R \wedge \overline{SNP_{5 2}^D}) \vee SNP_{7 3}^R$			
5	$\overline{(SNP_{4 1}^D \wedge SNP_{8 2}^R)} \vee (SNP_{5 1}^R \wedge \overline{SNP_{6 1}^D})$			
6	$((SNP_{4 2}^R \vee SNP_{7 2}^R) \wedge SNP_{8 3}^R) \vee (SNP_{9 2}^R \wedge SNP_{6 1}^R)$			
7	$(SNP_{3 1}^R \wedge SNP_{12 2}^R) \vee (SNP_{5 4}^R \wedge SNP_{15 1}^R) \vee (SNP_{9 2}^R \wedge SNP_{8 3}^R)$			
	$P(G)$	Haplotypes / block	# of haplotypes	# mating table rows
1	0.069	3	3	21
2	0.053	8, 5	40	$\approx 10^5$
3	0.073	3, 3, 3	27	$\approx 10^4$
4	0.060	4, 5, 8	160	$\approx 10^7$
5	0.057	8, 5, 5, 3	600	$\approx 10^{10}$
6	0.063	8, 8, 5, 3, 3	2,880	$\approx 10^{12}$
7	0.066	3, 4, 5, 5, 3, 5	4,500	$\approx 10^{13}$

Simulation



Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

Simulation

- When using conditional logistic regression to compare cases and pseudo-controls, the expected value of the parameter estimates is not the logs odds ratio β , but the log relative risk (Schaid 1996).

$$\begin{aligned}
 RR &= \frac{P(D|I_G = 1)}{P(D|I_G = 0)} = \frac{\exp(\alpha + \beta)/(1 + \exp(\alpha + \beta))}{\exp(\alpha)/(1 + \exp(\alpha))} \\
 &= \exp(\beta) \times \left(\frac{1 + \exp(\alpha + \beta)}{1 + \exp(\alpha)} \right)^{-1}
 \end{aligned}$$

- Therefore

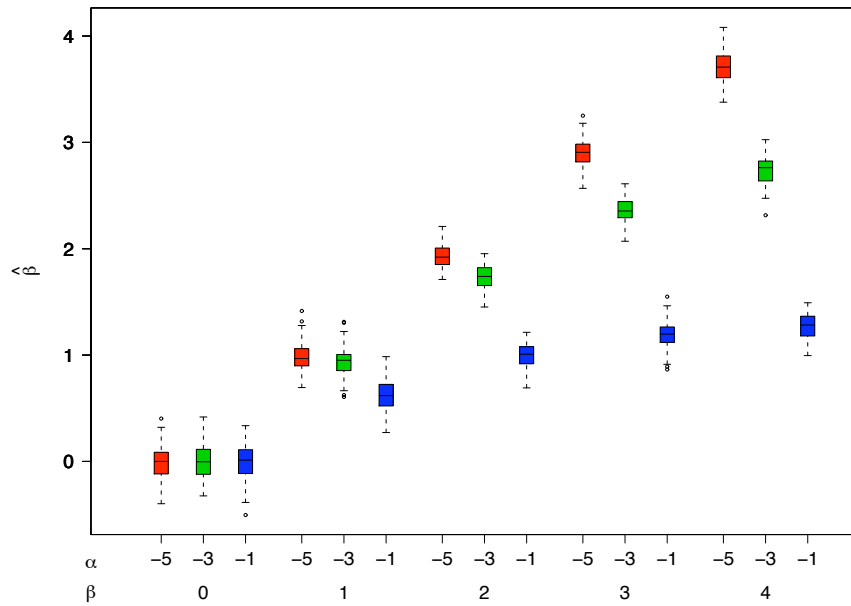
$$\log(RR) = \beta - \log \left(\frac{1 + \exp(\alpha + \beta)}{1 + \exp(\alpha)} \right)$$

Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

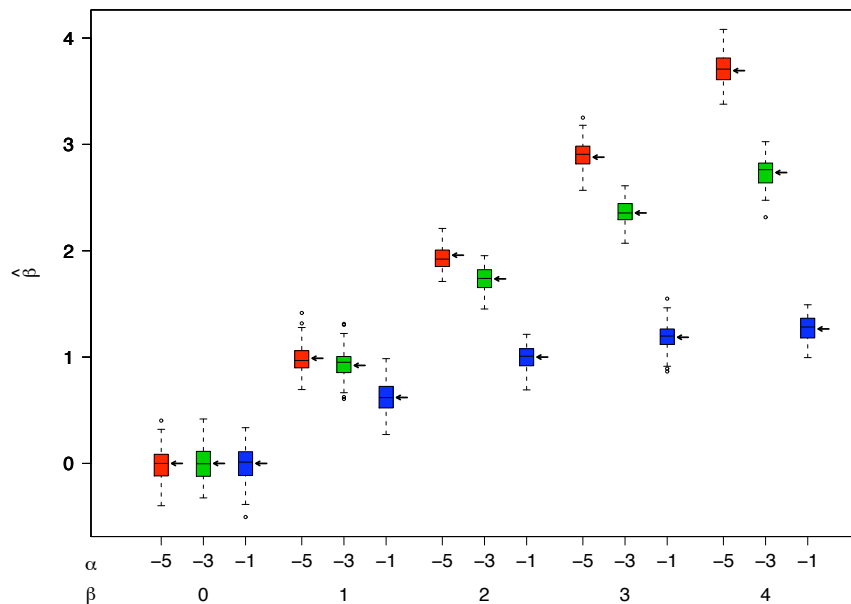
Simulation

$$\log(RR) = \beta - \log\left(\frac{1 + \exp(\alpha + \beta)}{1 + \exp(\alpha)}\right)$$



Simulation

$$\log(RR) = \beta - \log\left(\frac{1 + \exp(\alpha + \beta)}{1 + \exp(\alpha)}\right)$$



Software

- 1 The trio logic regression methods are implemented as an augmentation in the logic regression R package `LogicReg`.
- 2 The R package `trio` contains functions to generate logic regression input from pedigree or genotype files, to check for Mendelian errors, to impute missing data, and to simulate case-parent trios.
- 3 A software vignette is also available.

Set up data for trio logic regression or simulate trio data for high order SNP-SNP interaction



Documentation for package 'trio' version 1.0

Help Pages

[trio](#)
[trio.check](#)
[trio.sim](#)

Generate Trio Data Format Suitable for Trio Logic Regression
Check Case-Parent Trio Data for Mendelian Errors
Simulate Case-Parent Trios with Population Disease Risk Dependent on SNP-SNP Interaction

Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

Results

	Logic model	$\exp(\hat{\beta})$
1	$0.67 \times I_{\{302^D\}}$	1.94
2	$0.89 \times I_{\{302^D \vee 166^D\}}$	2.43
3	$1.15 \times I_{\{302^D \vee 166^D \vee 148^D\}}$	3.14
4	$1.30 \times I_{\{302^D \vee 166^D \vee 148^D \vee 368^R\}}$	3.65

SNP 302	Chromosome 12	NOS1	3782219
SNP 166	Chromosome 8	CHRN3	1530848
SNP 148	Chromosome 8	PNOC	3735736
SNP 368	Chromosome 22	COMT	740603

Ingo Ruczinski

SNP-SNP Interactions in Case-Parent Trios

Results

	Logic model	$\exp(\hat{\beta})$
1	$0.67 \times I_{\{\overline{302^D}\}}$	1.94
2	$0.89 \times I_{\{\overline{302^D} \vee 166^D\}}$	2.43
3	$1.15 \times I_{\{\overline{302^D} \vee 166^D \vee 148^D\}}$	3.14
4	$1.30 \times I_{\{\overline{302^D} \vee 166^D \vee 148^D \vee 368^R\}}$	3.65

SNP 302	Chromosome 12	NOS1	3782219
SNP 166	Chromosome 8	CHRN3	1530848
SNP 148	Chromosome 8	PNOC	3735736
SNP 368	Chromosome 22	COMT	740603

Results

		$\exp(\hat{\beta})$	$\hat{\beta}_{(se)}$	z	p
Marginal	$\overline{302^D}$	1.94	$0.67_{(0.16)}$	4.12	4e-05
	166^D	1.17	$0.16_{(0.22)}$	0.71	0.480
Logic	$\overline{302^D} \vee 166^D$	2.43	$0.89_{(0.18)}$	4.89	1e-06
Additive	$\overline{302^D}$	1.95	$0.67_{(0.16)}$	4.15	3e-05
	166^D	1.21	$0.19_{(0.22)}$	0.86	0.390
Additive	$\overline{302^D}$	2.46	$0.90_{(0.19)}$	4.86	1e-06
	166^D	2.52	$0.92_{(0.33)}$	2.77	0.006
	$\overline{302^D} : 166^D$	0.34	$-1.09_{(0.34)}$	-2.87	0.004

Results

		$\exp(\hat{\beta})$	$\hat{\beta}_{(se)}$	z	p
Marginal	$\overline{302^D}$	1.94	0.67 _(0.16)	4.12	4e-05
	166^D	1.17	0.16 _(0.22)	0.71	0.480
Logic	$\overline{302^D} \vee 166^D$	2.43	0.89 _(0.18)	4.89	1e-06
Additive	$\overline{302^D}$	1.95	0.67 _(0.16)	4.15	3e-05
	166^D	1.21	0.19 _(0.22)	0.86	0.390
Additive	$\overline{302^D}$	2.46	0.90 _(0.19)	4.86	1e-06
	166^D	2.52	0.92 _(0.33)	2.77	0.006
	$\overline{302^D} : 166^D$	0.34	-1.09 _(0.34)	-2.87	0.004

Results

		$\exp(\hat{\beta})$	$\hat{\beta}_{(se)}$	z	p
Marginal	$\overline{302^D}$	1.94	0.67 _(0.16)	4.12	4e-05
	166^D	1.17	0.16 _(0.22)	0.71	0.480
Logic	$\overline{302^D} \vee 166^D$	2.43	0.89 _(0.18)	4.89	1e-06
Additive	$\overline{302^D}$	1.95	0.67 _(0.16)	4.15	3e-05
	166^D	1.21	0.19 _(0.22)	0.86	0.390
Additive	$\overline{302^D}$	2.46	0.90 _(0.19)	4.86	1e-06
	166^D	2.52	0.92 _(0.33)	2.77	0.006
	$\overline{302^D} : 166^D$	0.34	-1.09 _(0.34)	-2.87	0.004

Results

$$\rightarrow 0.67 \times I_{\{\overline{302^D}\}} + 0.19 \times I_{\{166^D\}}$$

		$\overline{302^D}$		302^D		
166^D		0.67	0	0		0
		0.86	0.19	0.19		1
		0.86	0.19	0.19		2
		0	1	2		

main effects only

Results

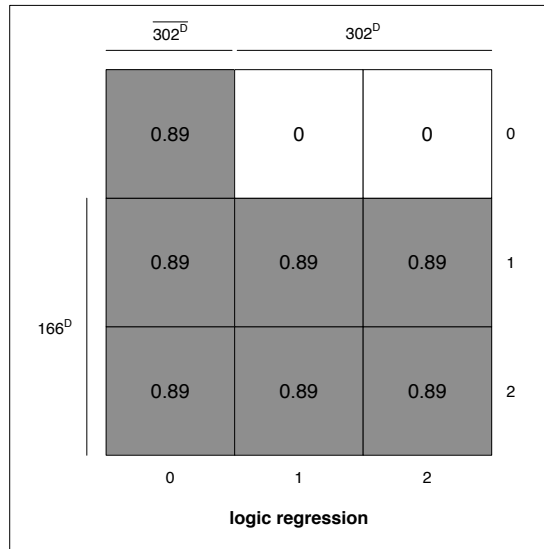
$$\rightarrow 0.90 \times I_{\{\overline{302^D}\}} + 0.92 \times I_{\{166^D\}} - 1.09 \times I_{\{\overline{302^D};166^D\}}$$

		$\overline{302^D}$		302^D		
166^D		0.9	0	0		0
		0.73	0.92	0.92		1
		0.73	0.92	0.92		2
		0	1	2		

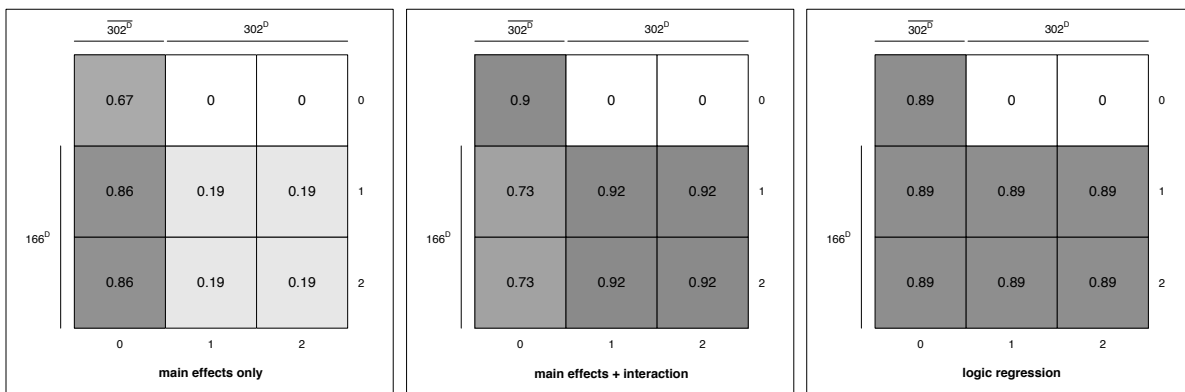
main effects + interaction

Results

$$\rightarrow 0.89 \times I_{\{302^D \vee 166^D\}}$$



Results



Results

		$\exp(\hat{\beta})$	$\hat{\beta}_{(se)}$	z	p
	$\overline{302^D}$	1.94	0.67 _(0.16)	4.12	4e-05
Marginal	166 ^D	1.17	0.16 _(0.22)	0.71	0.480
	148 ^D	1.54	0.43 _(0.25)	1.70	0.088
Logic	$\overline{302^D} \vee 166^D \vee 148^D$	3.14	1.15 _(0.20)	5.67	2e-08

Acknowledgments

Methods: Dani Fallin, Qing Li, Tom Louis.

Data: Ann Pulver.

Computing support: Marvin Newhouse, Jiong Yang.

Funding: NIH R01 DK061662, GM083084, HL090577, and a CTSA grant to the Johns Hopkins Medical Institutions.

<http://biostat.jhsph.edu/~iruczins/>