

Empirical Comparison of a General Approach for Analyzing Sibling-Pair Data

Fredrick R. Schumacher¹, David V. Conti², John S. Witte³

¹ Dept. Epi & Biostats, Case Western Reserve University

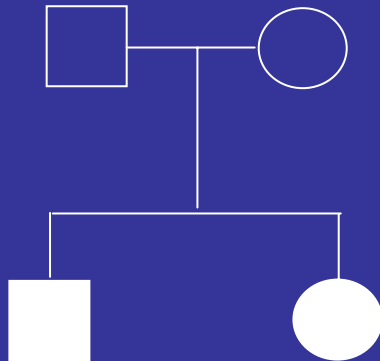
² Dept. of Prev. Med., USC

³ Dept. of Epi & Biostats, UCSF

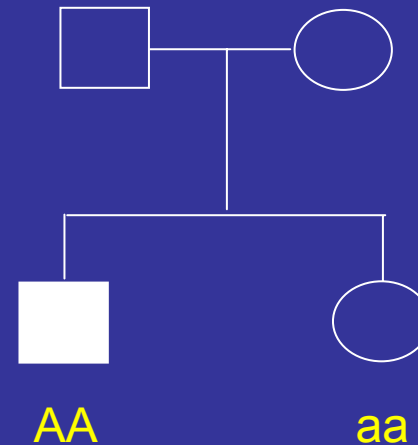
Background

- Sibling-based designs are commonly used for both linkage and association studies

Linkage



Association



Goals

- Contrast the measures of disease risk from two statistical methods on a real dataset
- Empirically compare two statistical methods on simulated data

Statistical Method (1)

Conditional Logistic Regression (CLR)

- Standard epidemiologic tool for the analysis of matched data (i.e. 1-1, 1-2, N-M)
- Implemented in many commonly used statistical packages (i.e. STATA, SAS “proc PHREG”, S-Plus “coxph”)
- Valid for testing association in the presence of linkage with “minimal” sib-ships
- CLR can be applied to sib-ships of arbitrary size, but not valid if residual correlation exists
 - An estimated grouped jackknife variance can be used in the presence of sibling correlation
- Easy to incorporate covariates, GxG, or GxE interactions

Statistical Method (1)

CLR

$$l(\beta) = \prod_{i=1}^I \Pr(D_i | n_i, Z_i) = \prod_{i=1}^I \frac{\prod_{j \in D_i} \exp(Z'_{ij} \beta)}{\sum_{S \in C_i} \prod_{j \in S} \exp(Z'_{ij} \beta)}$$

$i = 1, \dots, I$ denote sibship

D_i = the set of affected siblings in sibship

n_i = the number of affected sibs

M_i = marker genotypes in the i^{th} sibship

Z_i = coding for the marker genotypes

C_i = set of all possible subsets for which n_i -affected sibs are sampled from the i^{th} sibship

β = the log-odds of disease

Statistical Method (2)

Within and Between Orthogonal Decomposition approach (WB)

- Uses data with or without parental information
- Capable of handling sib-pairs or sib-ships of various sizes
- Controls for spurious associations due to population stratification and admixture
- Uses information from concordant and discordant pairs
- The genotype score can be expressed as orthogonal components

Statistical Method (2)

WB

$$\hat{y}_{ij} = \mu + \beta_b b_i + \beta_w w_{ij}$$

b_i : between-family component of g_{ij}

w_{ij} : within-family component of g_{ij}

The between- and within-family components are defined as

$$b_i = \begin{cases} \frac{\sum_j g_{ij}}{n_i} & \text{if parental genotypes are unknown} \\ \frac{g_{iF} + g_{iM}}{2} & \text{if parental genotypes are known} \end{cases}$$

$$w_{ij} = g_{ij} - b_i$$

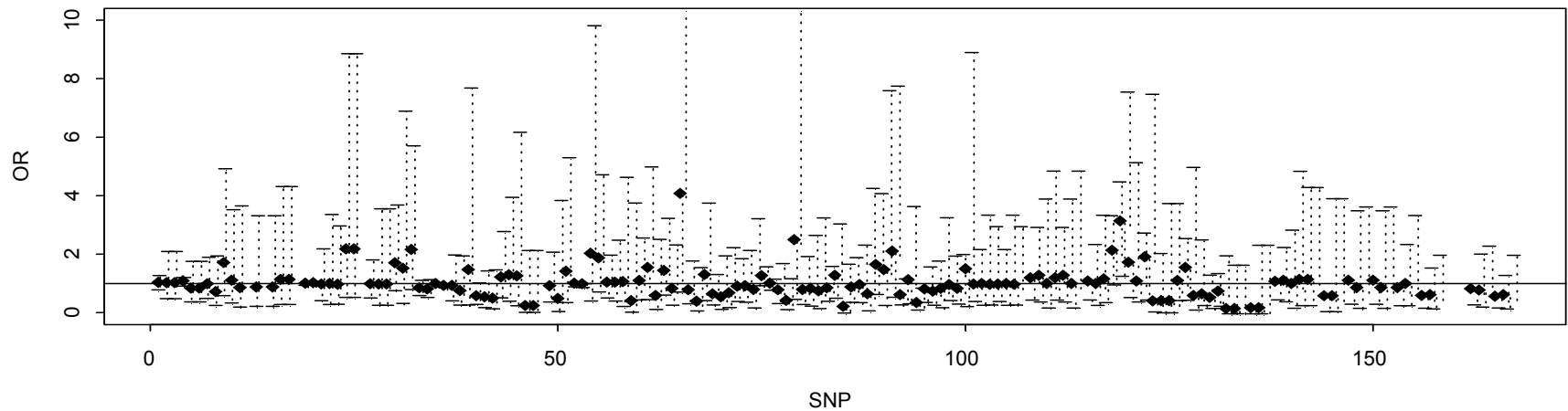
Dataset (1)

Prostate Cancer Genetic Epidemiologic (CaP Genes) Study

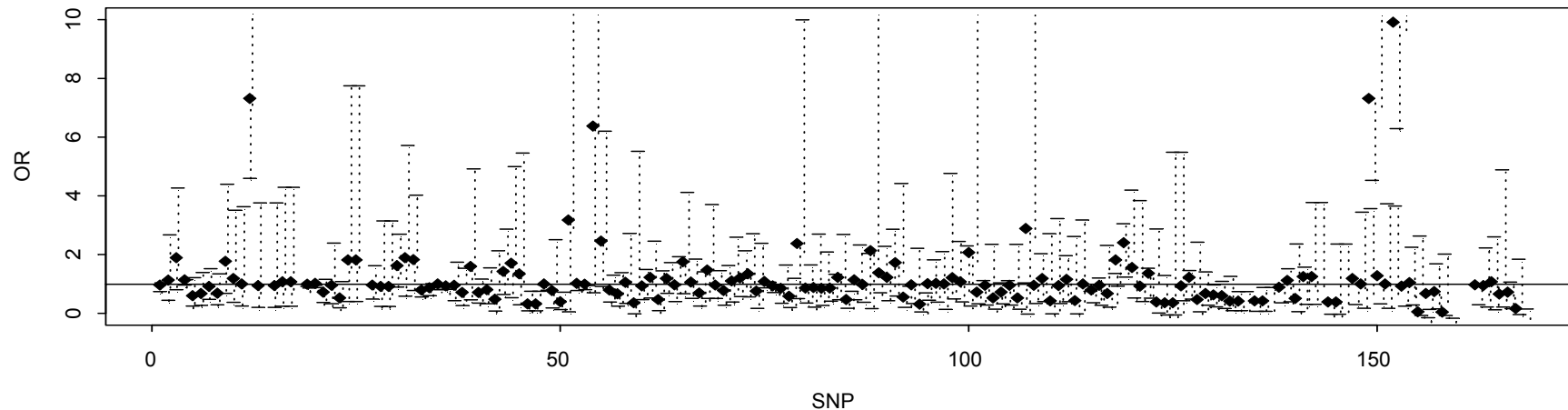
- Prevalent cancer cases were recruited from three Cleveland area hospitals and one Detroit area hospital from January 1998 to January 2001
- Inclusion Criteria
 - Histological confirmed prostate cancer
 - At least one brother, affected or unaffected, willing to participate
 - If a younger brother was a control, they must be within 8 years of the case's diagnostic age
 - Control status confirmed by a PSA screening test

CaP Genes Results (1)

A Scatterplot of OR (95% CI) for 60 SNPs from CLR

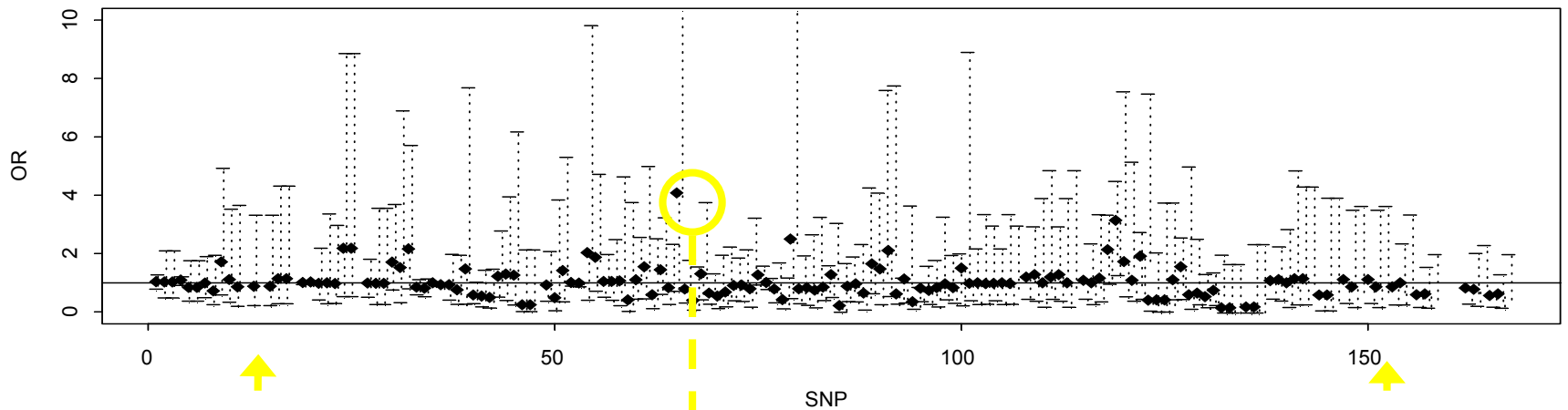


A Scatterplot of OR (95% CI) for 60 SNPs from WB

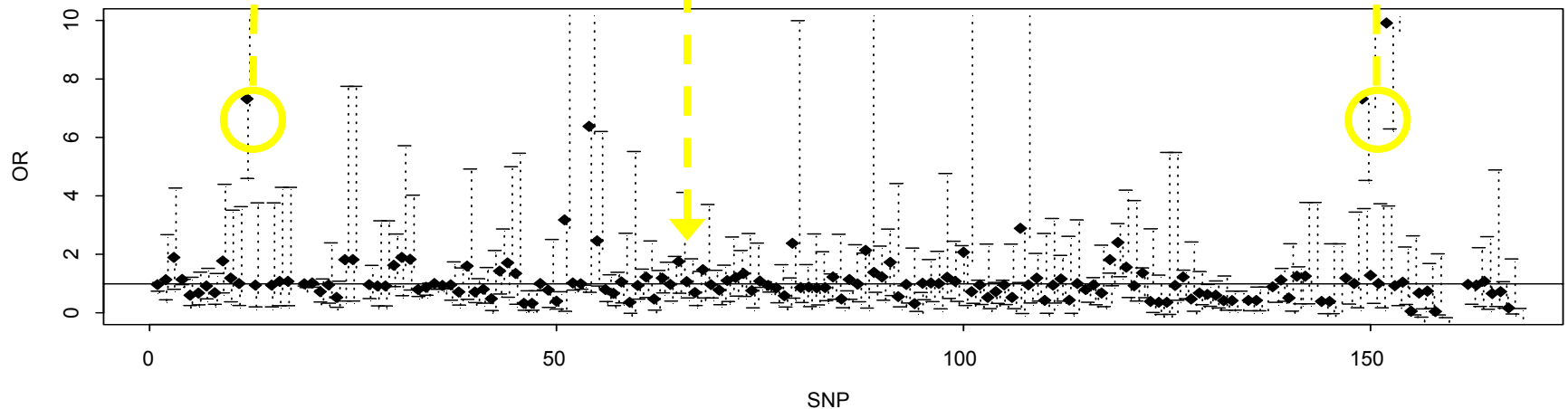


CaP Genes Results (1)

A Scatterplot of OR (95% CI) for 60 SNPs from CLR

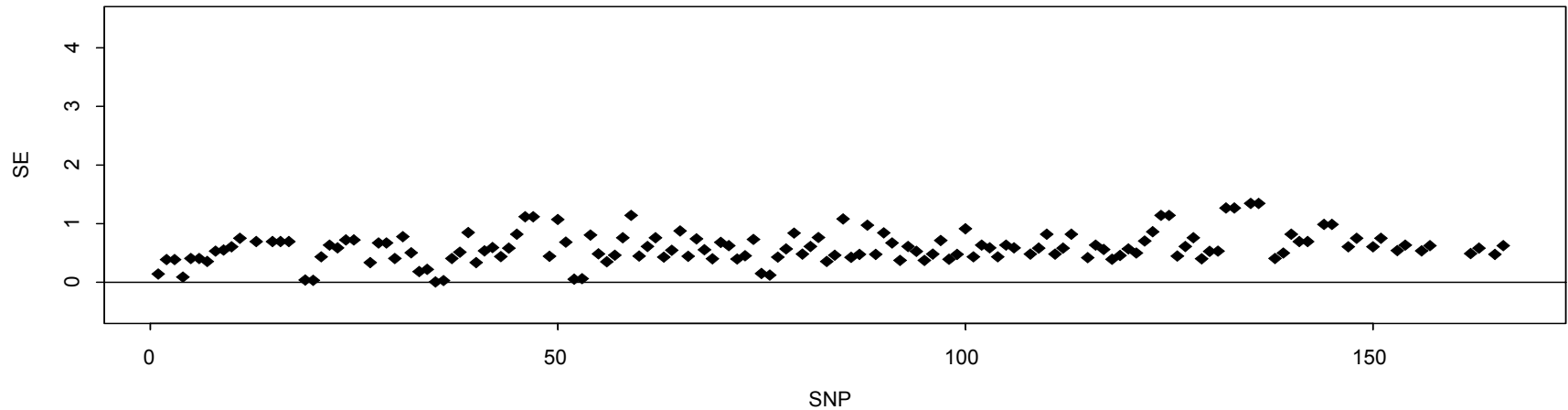


A Scatterplot of OR (95% CI) for 60 SNPs from WB

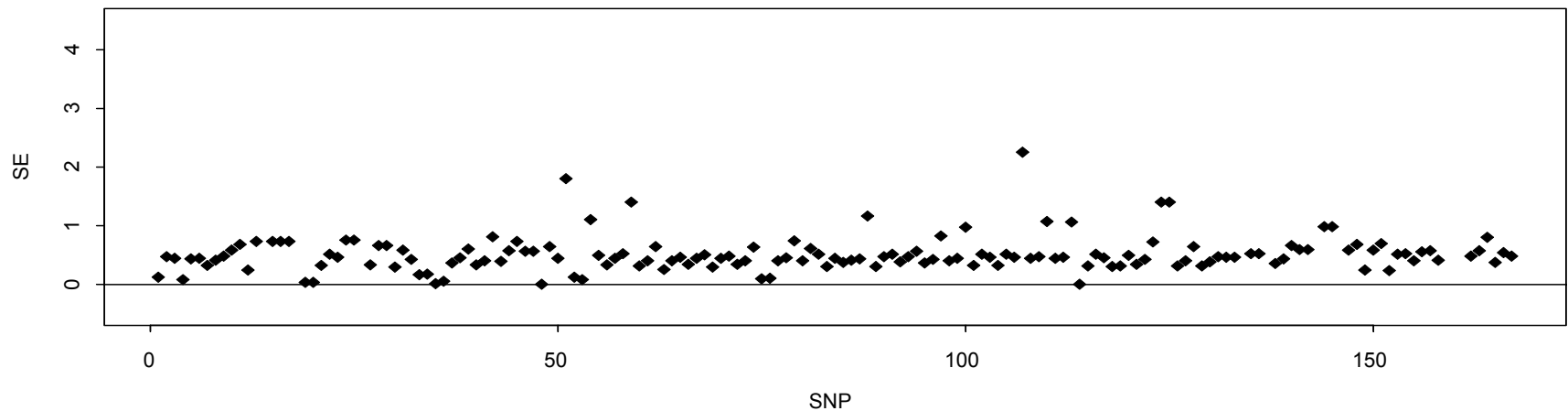


CaP Genes Results (2)

A Scatterplot of SE for 60 SNPs from CLR



A Scatterplot of SE for 60 SNPs from WB



CaP Genes Results (3)

Table 2. The marginal and joint distribution of odds ratios calculated from the CLR and WB models for 60 SNPs in the CaP Genes study

		CLR ^a			
		OR ^c ≤0.95	0.95≤OR≤1.05	OR≥1.05	
WB ^b	OR≤0.95	37.3% (79) ^d	5.6% (0)	1.7% (0)	44.6% (134)
	0.95≤OR≤1.05	3.3% (0)	10.1% (9)	5.6% (0)	19.0% (14)
	OR≥1.05	1.7% (0)	3.7% (0)	31.0% (30)	36.4% (59)
		42.3% (104)	19.4% (11)	38.3% (76)	100%

^a Conditional logistic regression

^b Within- and Between-approach

^c Odds ratio

^d (#) of statistically significant (alpha=0.05) OR

Dataset (2)

Simulated Data

- The Genometric Analysis Simulation Program (GASP) version 3.31 was used to simulate fixed sib-ships under different conditions
- Two hundred sib-ships were randomly drawn from one thousand simulated families. Each sib-ship contained at least one affected individual, but concordant pairs were allowed. A thousand replicates was performed for each scenario
- The allele frequency for the trait locus was fixed at 0.50 for all scenarios
- The population disease prevalence was set at 20%

Dataset (2)

Simulated Data

Table 1. Adjusted simulated parameters

Sib-ship size:	2, 3, or 4
Recombination fraction (θ):	0, 0.10, 0.25, 0.40, or 0.50
Marker 1 allele frequency (p):	0.90, 0.75, 0.50, 0.25, or 0.10

Population Admixture

One hundred families were randomly drawn from two different scenarios to form an “Admixture” population

Admixture1: PopA($\theta=0.10$, $p=0.10$) and PopB($\theta=0.40$, $p=0.75$)

Admixture2: PopA($\theta=0.25$, $p=0.90$) and PopB($\theta=0.50$, $p=0.25$)

Simulation Results (1)

Table 3. Type I error rates for the CLR and WB models with a significance level set at 0.05

Sibship Size	θ^a p^b	CLR					WB				
		0.00	0.10	0.25	0.40	0.50	0.00	0.01	0.25	0.40	0.50
N=2	0.9	0.046	0.036	0.028	0.040	0.044	0.072	0.072	0.060	0.048	0.052
	0.75	0.040	0.040	0.060	0.075	0.048	0.048	0.048	0.068	0.080	0.064
	0.5	0.056	0.056	0.060	0.048	0.052	0.068	0.068	0.080	0.052	0.060
	0.25	0.052	0.052	0.060	0.052	0.028	0.080	0.080	0.068	0.064	0.040
	0.1	0.040	0.040	0.036	0.063	0.060	0.064	0.064	0.060	0.092	0.072
N=3	0.9	0.060	0.052	0.044	0.028	0.028	0.064	0.052	0.056	0.032	0.036
	0.75	0.048	0.052	0.044	0.044	0.040	0.056	0.076	0.048	0.040	0.044
	0.5	0.067	0.052	0.052	0.063	0.060	0.060	0.048	0.044	0.068	0.072
	0.25	0.063	0.083	0.036	0.052	0.044	0.068	0.060	0.048	0.048	0.052
	0.1	0.067	0.080	0.052	0.060	0.052	0.060	0.072	0.052	0.072	0.072
N=4	0.9	0.060	0.040	0.056	0.048	0.044	0.048	0.048	0.044	0.056	0.052
	0.75	0.052	0.063	0.044	0.052	0.036	0.020	0.040	0.036	0.060	0.056
	0.5	0.048	0.024	0.036	0.040	0.048	0.028	0.020	0.036	0.068	0.048
	0.25	0.075	0.052	0.056	0.048	0.028	0.072	0.056	0.040	0.040	0.052
	0.1	0.060	0.075	0.040	0.063	0.056	0.084	0.088	0.060	0.060	0.032
Admixture1		0.040					0.048				
Admixture2		0.083					0.060				

^a Recombination fraction

^b Marker 1 allele frequency

Discussion

- In general, the type I errors were slightly lower for the CLR model compared to the WB model
- The WB model uses all of the available information from concordant and discordant pairs
- In the absence of population admixture, the between parameter in the WB model was not statistically significant (Likelihood ratio test)
- In the presence of population admixture, the between parameter was statistically significant in all of the simulations (data not shown)
- Future work will include power estimates and the distribution of family types (concordant vs. discordant) included in the analysis

Acknowledgements

NHLBI

Jane Olson- Training Grant PI

CaP Genes Study

John S. Witte (UCSF)- Advisor

David V. Conti (USC)

Nora Nock

Cristina Ferrazzano

Cheryl Thompson

Cleveland Clinic Lab

Graham Casey

Mine Cicek

Sarah Plummer