

Missing Data in Family-Based Genetic Association Studies

Shelley B. Bull and Juan Pablo Lewinger
University of Toronto

Background

Genetic association – Study designs
Case-parent (Trio) design and data
Missing data mechanisms

Current Methods

Original transmission/disequilibrium test (TDT)
FBAT (Family-based association test) methods
Some observations

A Likelihood-Ratio-Based Test of Association

Definition of the Test Statistic
Treatment of Missing Data
Evaluation and Conclusions

Genetic Association - Study Designs

“Outcome” is disease status = affected/unaffected
“Exposure” is candidate gene/marker genotype/alleles

Unrelated case-control association

- sensitive to population stratification or admixture, i.e.. confounding by ethnicity or population history
- arises when the sampled population consists of multiple subpopulations in which the disease prevalence and genotype frequencies differ among subpopulations

Family-based association

- less efficient than the unrelated case-control design
- immune to population stratification,
by conditioning on parental genotypes
- issues in dealing with incompletely observed or missing data in families, specifically missing parental genotypes

Case-parent (Trio) Design / Data

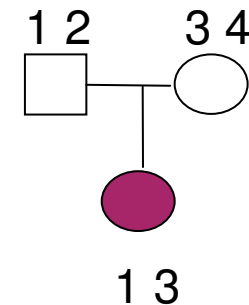
Ascertain (sample) on the child's disease status (phenotype): Ω

Two informative parents:

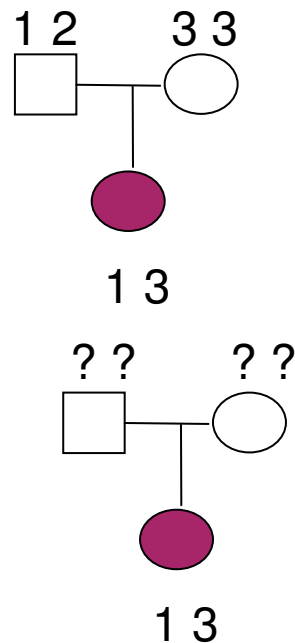
Mother transmits allele 3 to affected child

Under H_0 : $\text{pr}(\text{transmit } 3 \mid \Omega) = \text{pr}(\text{transmit } 4 \mid \Omega) = \frac{1}{2}$

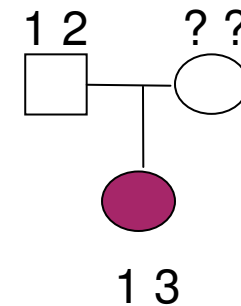
Under H_A : $\text{pr}(\text{transmit } 3 \mid \Omega) > \text{pr}(\text{transmit } 4 \mid \Omega)$



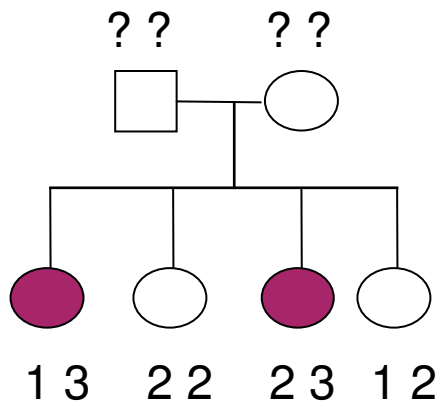
One uninformative parent:



One missing parent:



Both parents missing:



Missing Data Mechanisms

Issue: Conditioning event, i.e. the parental genotypes, is incompletely observed or unobserved

Missing at random:

- distribution of genotypes of the missing parents
(conditionally on genotypes of offspring, available parent),
is NOT different from parents with observed genotypes
- valid estimates of population genotype frequencies can be
estimated from the sampled parents (given ascertainment)

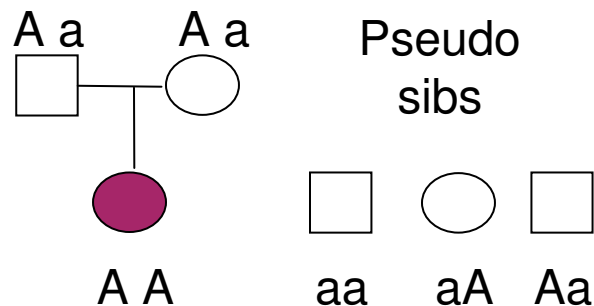
Informative missingness:

- whether a parent is missing depends on his/her genotype
at the locus of interest:
 - genotype is associated with early mortality from the disease of interest,
 - genotype is associated with a different disease leading to missingness,
 - propensity to be missing is correlated with genotype frequency
in sub-populations within the sample.

Allen et al (2003)
Kistner & Weinberg (2004)
Chen (2004)

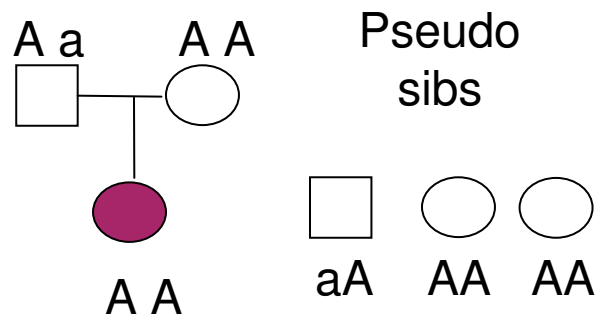
Original TDT for a Biallelic Marker

Two heterozygous parents:



		Transmitted	
		A	a
Not Transmitted	A	0	0
	a	2	0

One heterozygous parent:



		Transmitted	
		A	a
Not Transmitted	A	1	0
	a	1	0

Sum over all families:

b = # heterozygous parents transmit A
 c = # heterozygous parents transmit a

Original TDT for a Biallelic Marker

Sum over all N families:

Test statistic is: $T = (b - c)^2 / (b + c) \sim \text{asymptotic } \chi^2 (1 \text{ df})$

- Analogous to a matched case-control pair design with allele as the exposure, leading to McNemar's test

More generally: using all 3 pseudo-sibs corresponds to a likelihood of the conditional logistic form, leading to a score test.

Properties:

- Valid type I error under arbitrary parental genotype distributions and population stratification
- Analysis that ignores families with missing parents retains validity even under “informative missingness”
- Test for linkage of a marker locus to a disease locus (θ = recombination distance) in the presence of association between marker and disease-gene alleles (δ is allelic association / linkage disequilibrium)
- Power depends on level of allelic association between marker and disease loci

FBAT (Family-based Association) Methods

General framework for constructing valid tests under general mechanisms of genotype missingness

Specification of test statistics:

$$T = \sum_{i,j} f(G_{i,j})h(Y_{i,j})$$

Laird et al (2000)

$h(Y_{ij})$ is a function of phenotype, eg. 1=affected, 0=unaffected

$f(G_{ij})$ is defined by genotype, eg. # of 'A' alleles

Distribution of T :

Conditional on parental genotypes and observed traits

Under the null hypothesis of no linkage (H_0),

- offspring genotypes and all phenotypes are conditionally independent,
- the permutation distribution of offspring genotype values

follows Mendel's law of segregation.

Kaplan et al (1997)

For missing parents,

- cannot condition on unobserved parental genotypes,
- condition on the minimal sufficient statistics (under H_0)
for the parental genotypes. Rabinowitz and Laird (2000)
- distribution now depends on the offspring genotypes.

Some Observations

- most model specifications focus on conditional log-linear models and genetic relative risk/association parameters, and do not explicitly consider conventional genetic linkage parameters such as allele frequencies, penetrance, and genetic distance
- relatively little explicit attention given to ideas of “missing at random” and “informative missingness”
- in some cases, some missing data treatments can lead to loss of validity in the presence of population stratification, eg. parental reconstruction methods
- variation in the extent to which genotype and phenotype information from the entire nuclear family is used eg. TDT does not use information on
 - family structure
 - affected status of parents
 - unaffected offspring
 - families with two homozygous parents
- recent interest in methods that will retrieve this information

A Likelihood-Ratio-Based Test of Association

Objective

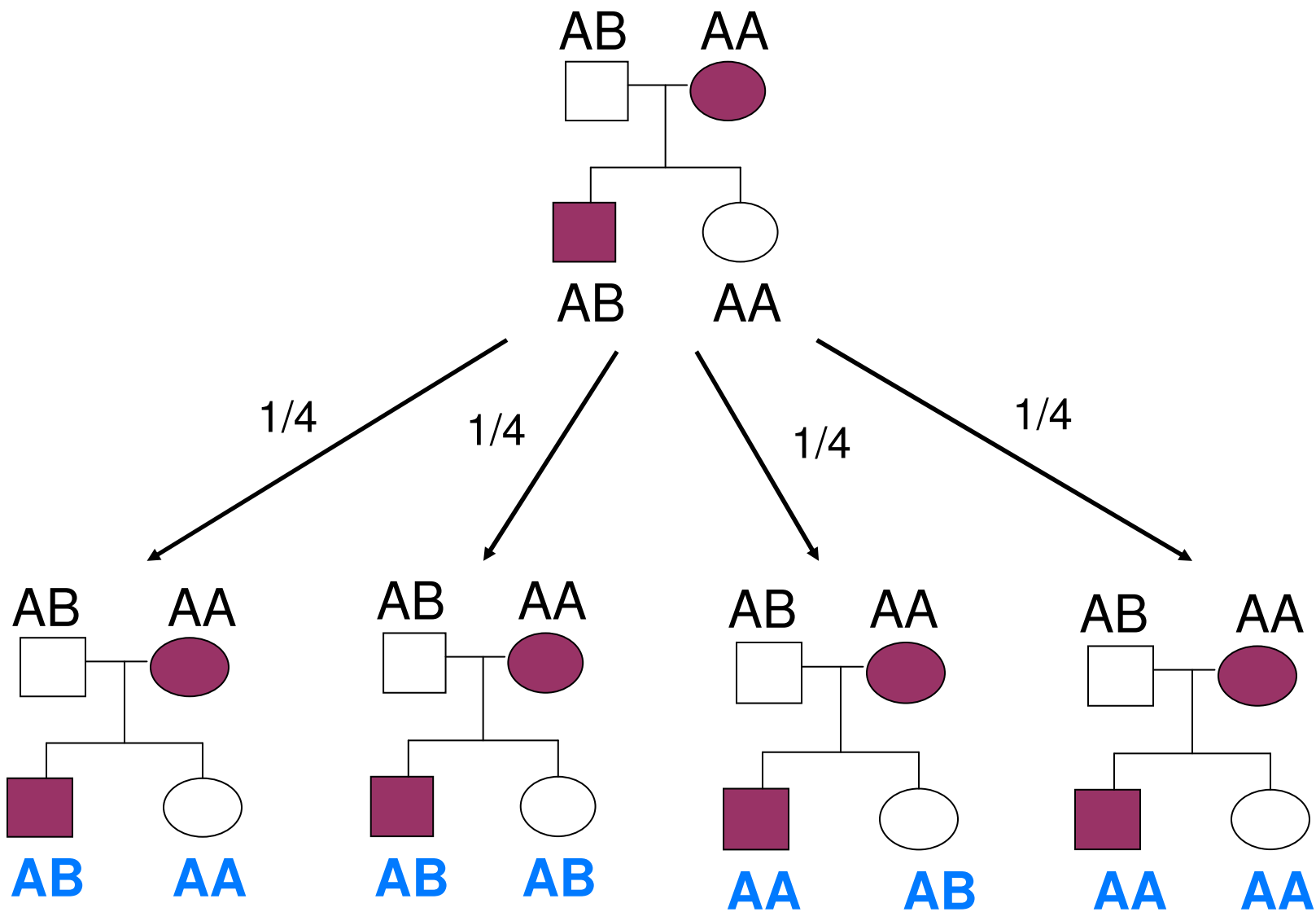
Construct a test of association that:

- Retains **immunity to population stratification**
- Makes **efficient** use of all family information available.
- Can be applied with any pattern of **missing genotypes**.

Conditional framework of Rabinowitz and Laird

- Immunity to population stratification obtained by conditioning on parental genotypes and all phenotypes:
 - Under null, children's genotypes and all phenotypes are conditionally independent given the parental genotypes.
 - Conditional distribution completely characterized by Mendel's law of segregation.

$$P_{H_0}(G_c | G_p, Y) = P_{H_0}(G_c | G_p) = 2^{-k} (G)$$



Formally

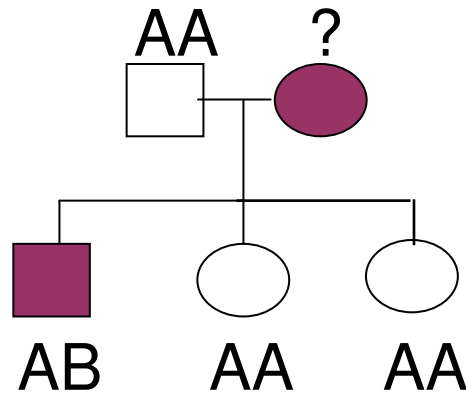
- $S=(G_p, Y)=(\text{Parental genotypes and all phenotypes})$ constitute a **sufficient statistic** for the null hypothesis of no linkage.
- Given an appropriate test statistic, $T=T(G,Y)$, compare $t_{\text{obs}}=T(g_{\text{obs}}, y_{\text{obs}})$ with the reference distribution

$$P_{H_0}(T \mid G_p, Y) = P_{H_0}(T \mid G_p)$$

Missing parental genotypes

- Cannot condition on parental genotypes.
- However, a sufficient statistic for the null hypothesis still exists.
- It also depends now on children's genotypes.

Example 1



Condition on: observed phenotypes, one parent missing, one parent *AA*, ***at least*** and one child *AB*, and ***at least*** one child *AA*.

AB, AA, AA \longrightarrow $1/6$

AA, AB, AA \longrightarrow $1/6$

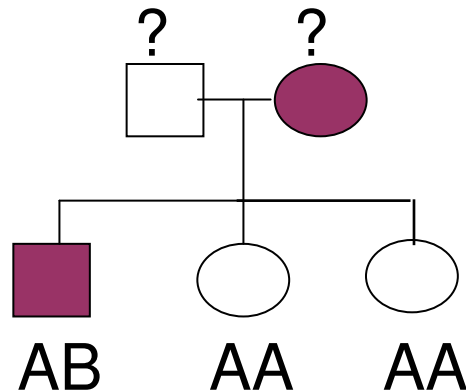
AA, AA, AB \longrightarrow $1/6$

AB, AB, AA \longrightarrow $1/6$

AB, AA, AB \longrightarrow $1/6$

AA, AB, AB \longrightarrow $1/6$

Example 2



Condition on: observed phenotypes, both parents missing, *exactly* one child AB, and *exactly* 2 children AA.

AB,AA,AA \longrightarrow 1/3

AA,AB,AA \longrightarrow 1/3

AA,AA,AB \longrightarrow 1/3

Formally

- S =(phenotypes, observed parental genotypes, pattern of missingness, and a function of the children's genotypes) constitute a **sufficient statistic** for the null hypothesis of no linkage.
- Given an appropriate test statistic, $T=T(X)$, compare $t_{\text{obs}}=T(X_{\text{obs}})$ with the reference distribution

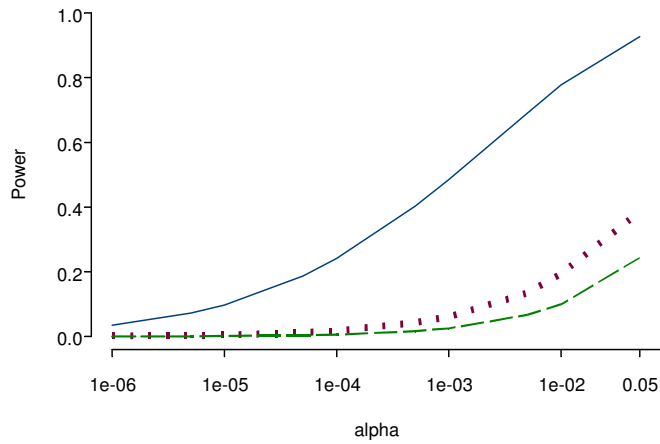
$$P_{H_0}(T \mid S)$$

FBAT vs. TDT

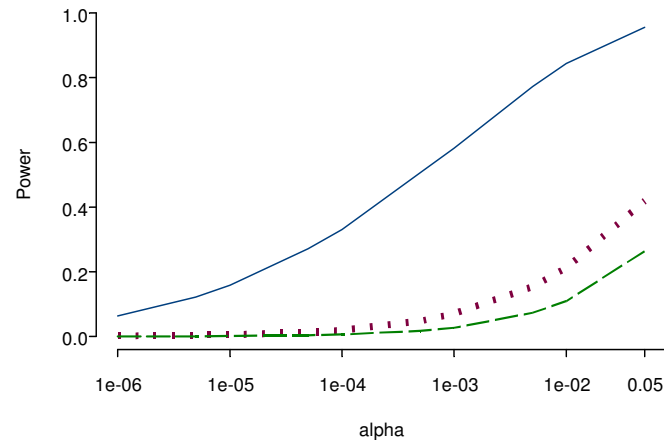
300 families: 1/3 complete, 1/3 one parent missing and 1/3 both parents missing

Dominant model

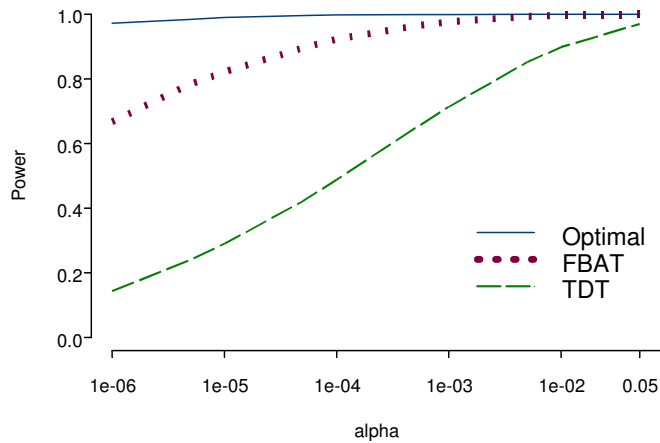
$\psi=0.5$
 $q=0.1$



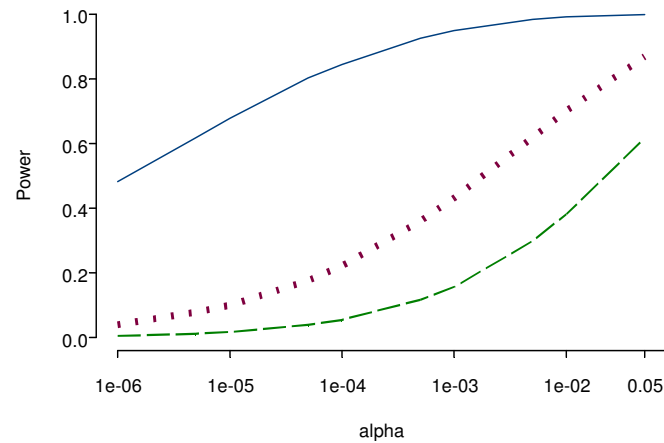
$\psi=0.5$
 $q=0.5$



$\psi=0.9$
 $q=0.1$



$\psi=0.9$
 $q=0.5$



Alternative Choice of Test Statistic

- Based on the standard parametric two point linkage model that incorporates allelic association parameters:

$$\theta, f_0, f_1, f_2, p, q, \psi$$

- Most powerful conditional test against fixed alternative ω is based on the conditional likelihood ratio statistic:

$$\frac{\Pr_{\omega}(\mathbf{X} \mid \mathbf{S})}{\Pr_{H_0}(\mathbf{X} \mid \mathbf{S})}$$

- Good power is wanted for all alternatives defined by the parametric model.
- Estimate parameters

$$\eta = (f_0, f_1, f_2, p, q, \psi)$$

based on the likelihood

$$L(\eta) = \Pr(S \mid Y_A; \eta)$$

- Segregation analysis using traits and founder genotypes.

- Use likelihood ratio statistic:

$$\exp(T) = \frac{\Pr \hat{\omega} (\mathbf{X} | \mathbf{S})}{\Pr_{H_0} (\mathbf{X} | \mathbf{S})}$$

where

$$\omega = (\theta = 0, \hat{\eta})$$

- T can be computed if there are missing data
assuming data are missing at random.

Performance

- Simulation study
 - Compare power of LR test to power of commonly used tests such as TDT and FBAT.
 - Compare power of LR test to maximum power attainable.

Simulation Design

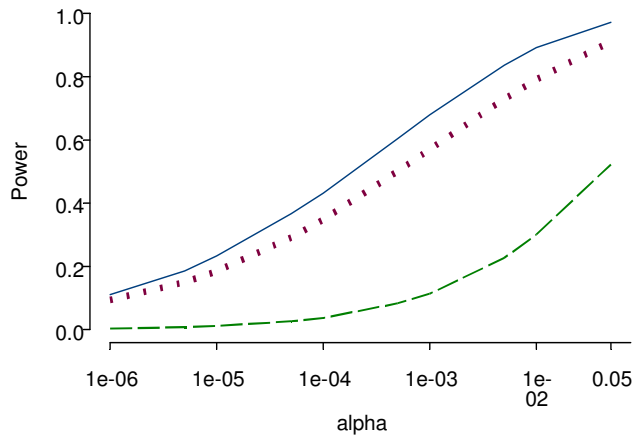
- Range of scenarios with prevalence $\approx 1\%$
 - Common **dominant** disease
 - Common **recessive** disease
 - Common **additive** disease
- Other parameters
 - Recombination fraction: $\theta=0.001, 0.01$
 - Allelic association: $\psi=10, 50$ and 90%
 - marker allele frequency: $q=0.1, 0.5$
- Sample sizes: **150, 300, 600** families
- Ascertainment: **Complete, single**

Power of LR vs. FBAT

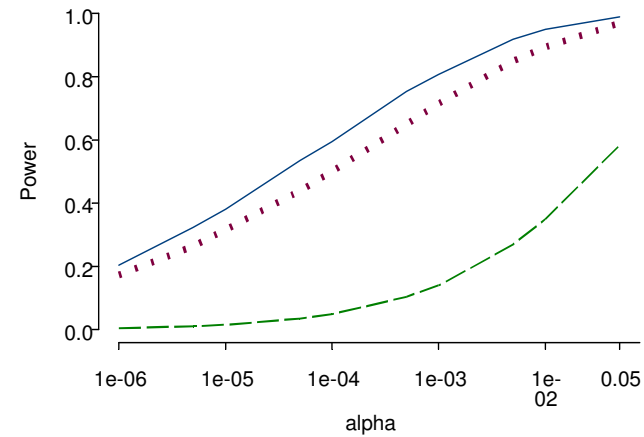
300 families. Complete data

Dominant model

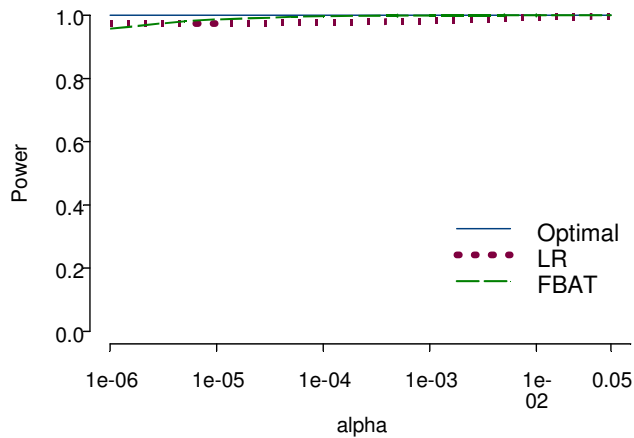
$\psi=0.5$
 $q=0.1$



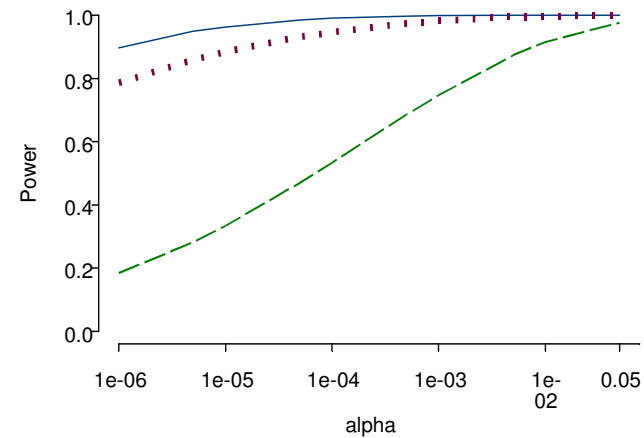
$\psi=0.5$
 $q=0.5$



$\psi=0.9$
 $q=0.1$



$\psi=0.9$
 $q=0.5$

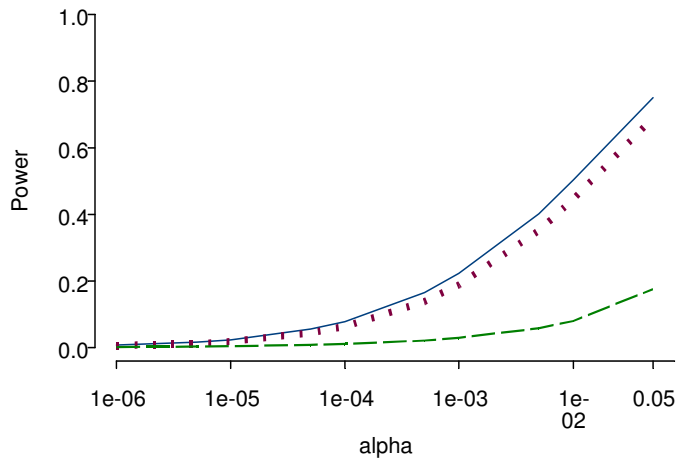


Power of LR vs. FBAT

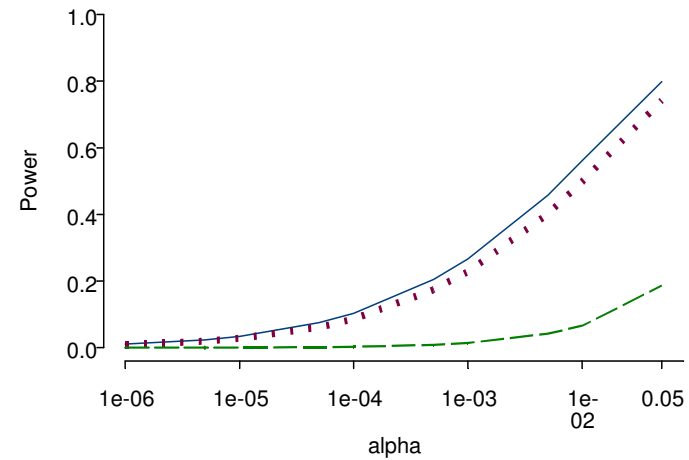
300 families: Both parents missing missing data

Dominant model

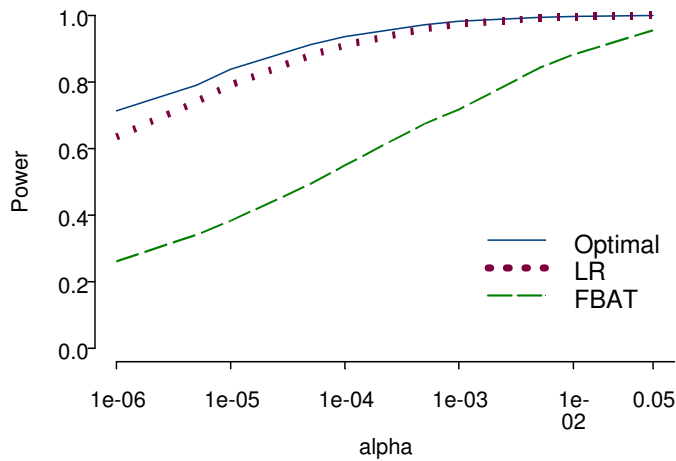
$\psi=0.5$
 $q=0.1$



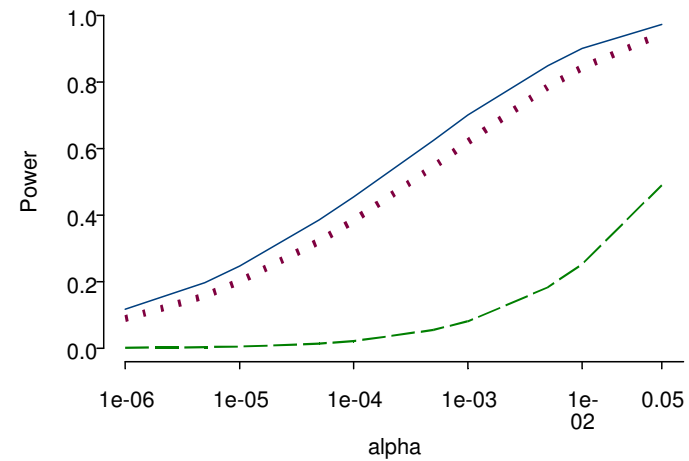
$\psi=0.5$
 $q=0.5$



$\psi=0.9$
 $q=0.1$



$\psi=0.9$
 $q=0.5$



Robustness

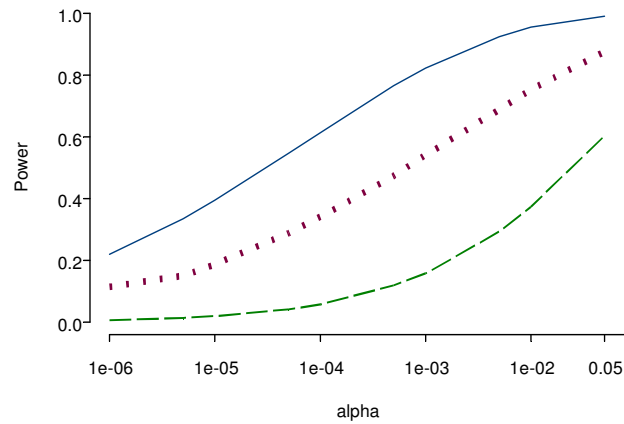
- For a range of disease scenarios with a mixture of two populations:
 - marker allele frequencies:
Population 1: $q_1 = 0.1$
Population 2: $q_2 = 0.5$
- Compare power between LR test and FBAT.

Power of LR vs. FBAT

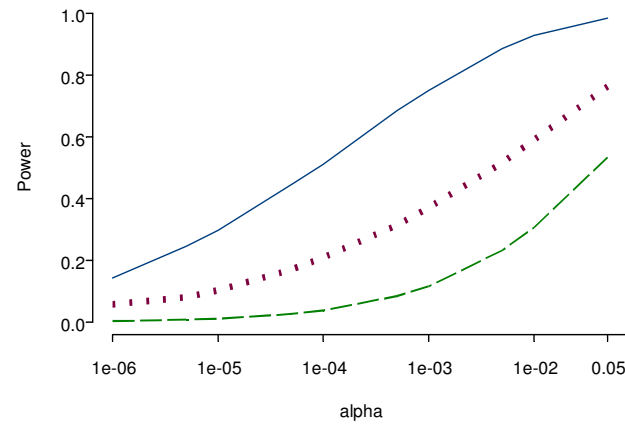
300 families: complete data

Dominant model. Mixture of two populations

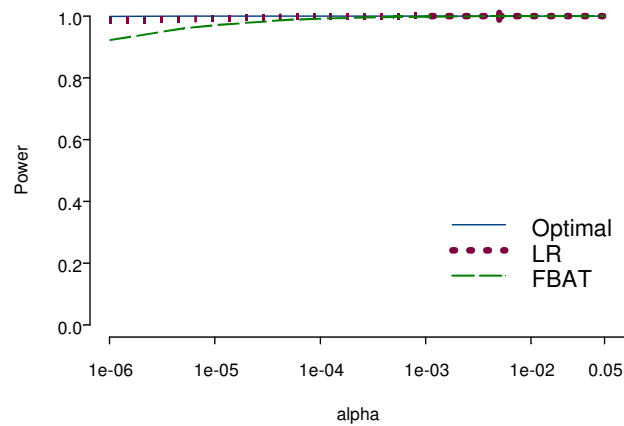
$\psi=0.5$
 $q_1=0.1$
 $q_2=0.2$



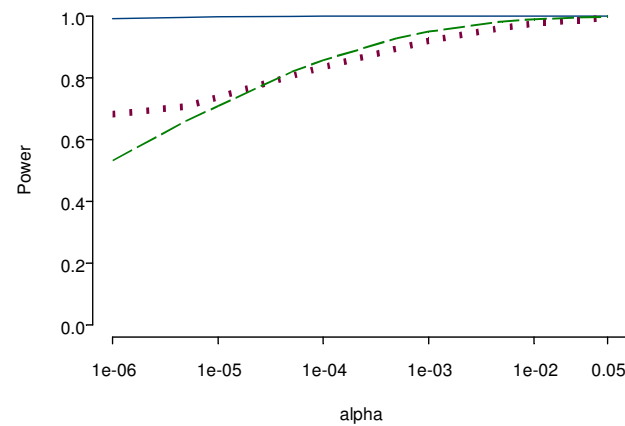
$\psi=0.5$
 $q_1=0.1$
 $q_2=0.5$



$\psi=0.9$
 $q_1=0.1$
 $q_2=0.2$



$\psi=0.9$
 $q_1=0.1$
 $q_2=0.5$



Conclusions

- Test more powerful than commonly used tests (TDT and FBAT) for all the scenarios considered under assumed model.
- Power always close to the theoretically maximum possible.
- Robust: power remains good under scenarios outside assumed model.

Future work

- Multiple alleles.
- More complex models.
- Quantitative, longitudinal and survival traits.
- Larger pedigrees.
- Multiple markers.

Acknowledgements

Joanna Szyda

Ying Liu

Fang Xie

Lucia Mirea

David Tritchler

Paul Corey

David Andrews

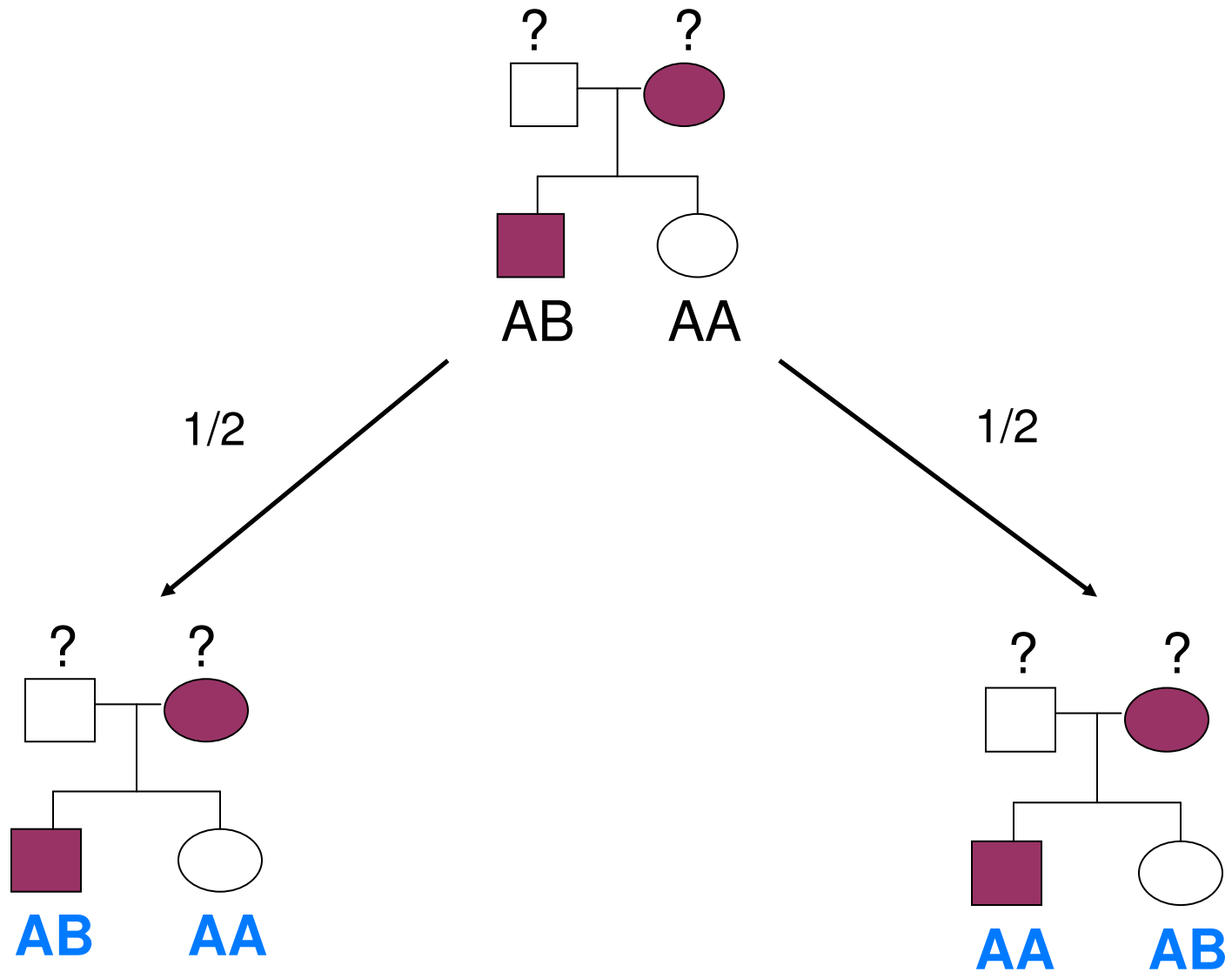
Lei Sun

NCE in Mathematics (MITACS)

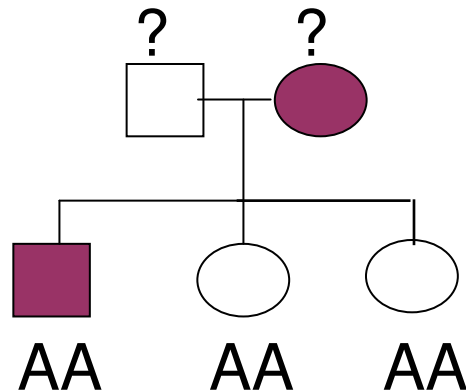
Canadian Institutes of Health Research

NSERC

Example 1



Example 4



Condition on: observed phenotypes, both parents missing, and ***exactly*** 3 children AA.

AB,AA,AA \longrightarrow 1