Regression-Based Linkage Analysis of General Pedigrees

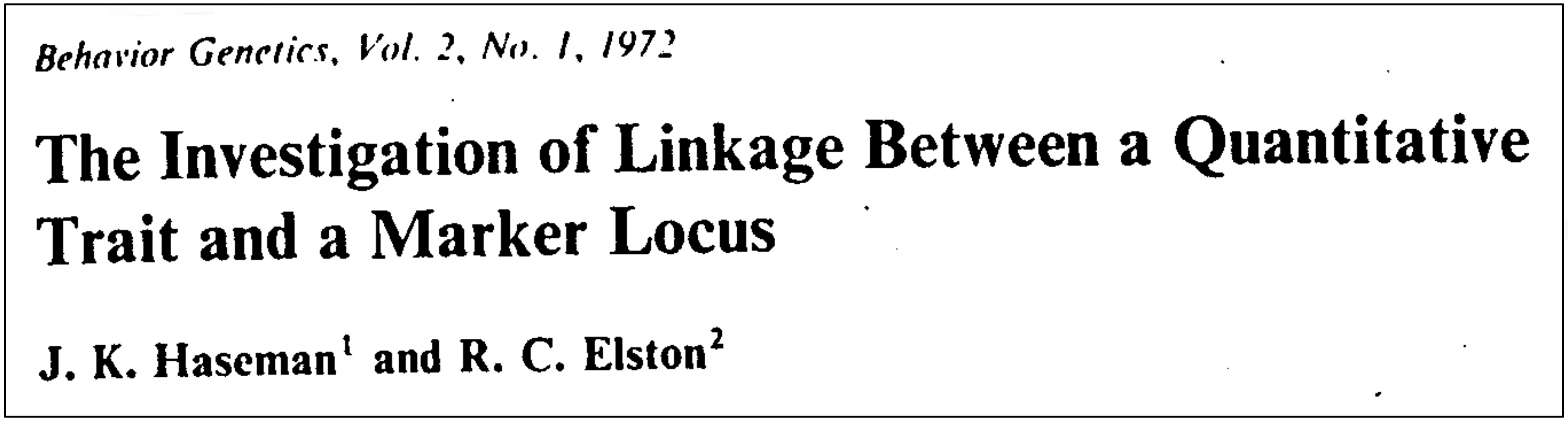
Pak Sham, Shaun Purcell,
Stacey Cherny, Gonçalo Abecasis

This Session

- Quantitative Trait Linkage Analysis
  - Variance Components
  - Haseman-Elston

- An improved regression based method
  - General pedigrees
  - Non-normal data

- Example application
  - PEDSTATS
  - MERLIN-REGRESS



Behavior Genetics, Vol. 2, No. 1, 1972

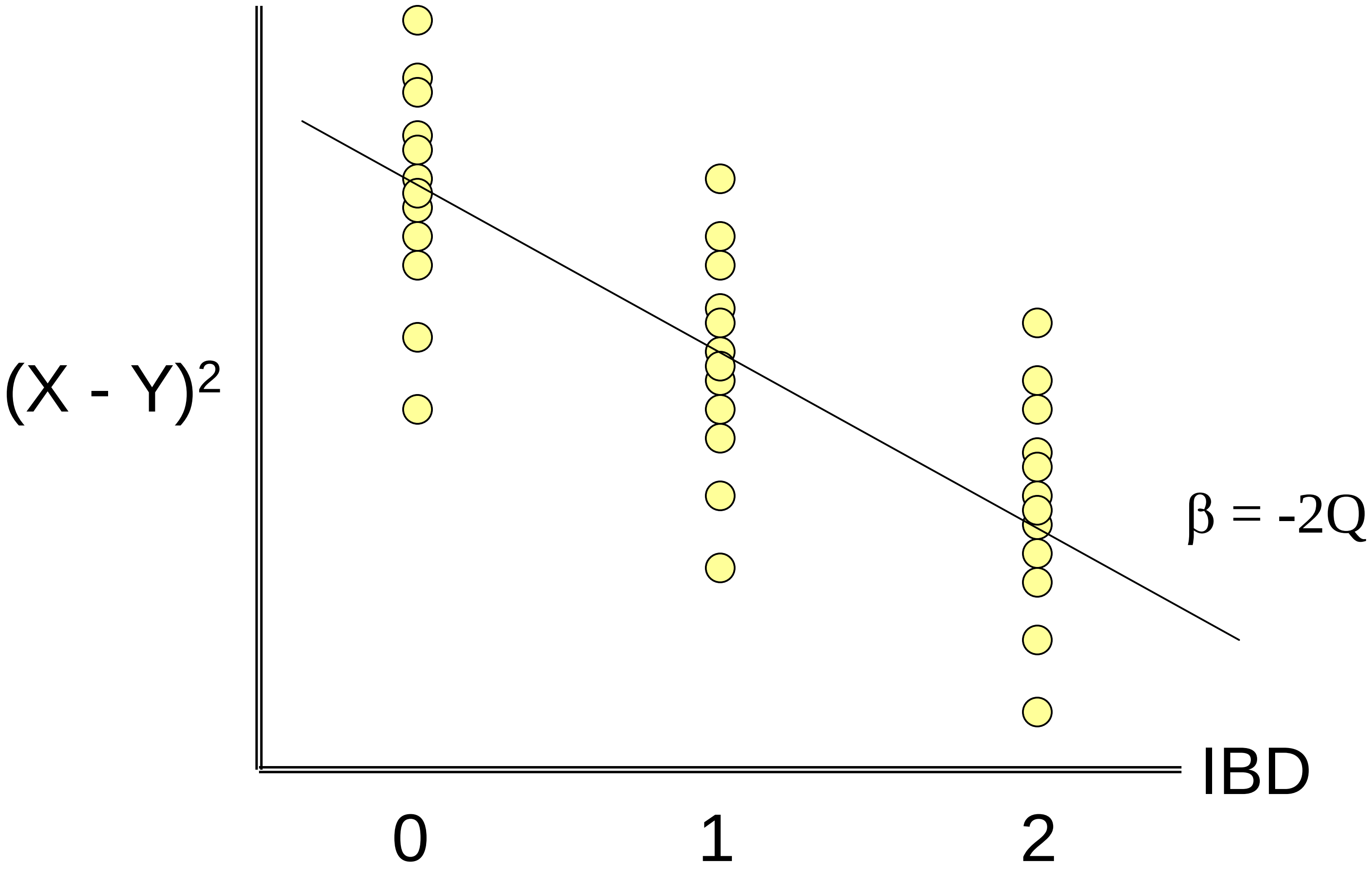
The Investigation of Linkage Between a Quantitative Trait and a Marker Locus

J. K. Haseman[1] and R. C. Elston[2]

- Simple regression-based method
  - squared pair trait difference
  - proportion of alleles shared identical by descent

$$(X - Y)^2 = 2(1 - r) - 2Q(\hat{\pi} - 0.5) + \varepsilon \qquad \text{(HE-SD)}$$

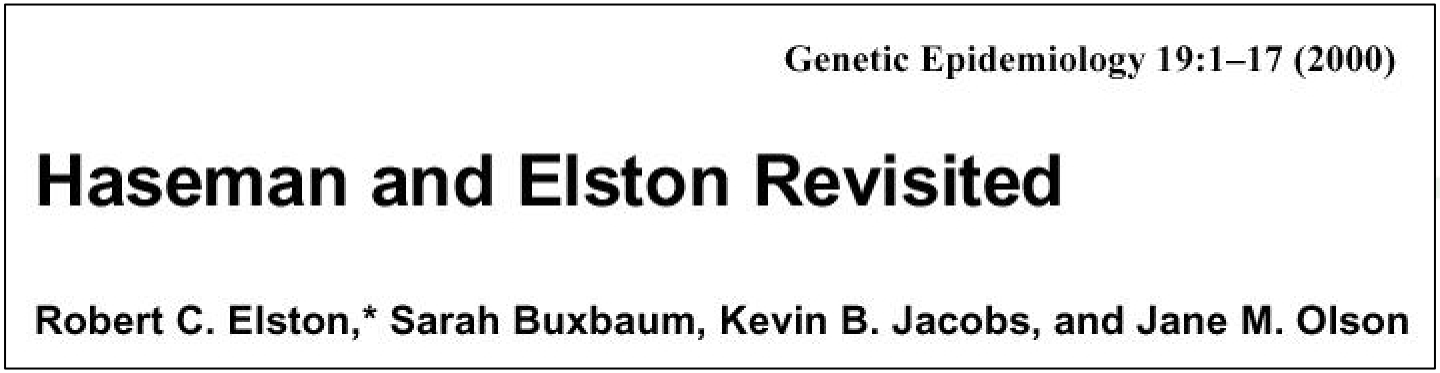
Haseman-Elston regression



Sums versus differences

- Wright (1997), Drigalenko (1998)
  - phenotypic difference discards sib-pair QTL linkage information
  - squared pair trait **sum** provides extra information for linkage
    - independent of information from HE-SD

$$(X + Y)^2 = 2(1 + r) + 2Q(\hat{\pi} - 0.5) + \varepsilon \qquad \text{(HE-SS)}$$



Genetic Epidemiology 19:1–17 (2000)

Haseman and Elston Revisited

Robert C. Elston,* Sarah Buxbaum, Kevin B. Jacobs, and Jane M. Olson

- New dependent variable to increase power
  - mean corrected cross-product (HE-CP)

$$XY = \tfrac{1}{4}\left((X+Y)^2 - (X-Y)^2\right)$$

- But this was found to be less powerful than original HE when sib correlation is high

Variance Components Analysis

$$\Omega = \begin{bmatrix} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \hat{\pi}_{marker}\sigma_a^2 + 2\varphi\sigma_g^2 \\ \hat{\pi}_{marker}\sigma_a^2 + 2\varphi\sigma_g^2 & \sigma_a^2 + \sigma_g^2 + \sigma_e^2 \end{bmatrix}$$

Where,

φ is the kinship coefficient for the two individuals

$\hat{\pi}_{marker}$ is the IBD sharing proportion

Likelihood function

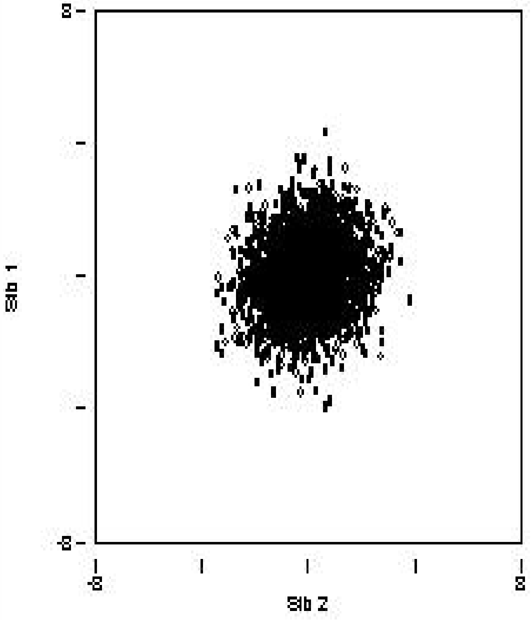
$$L = \prod_i \sum_{j=0,1,2} Z_{ij} (2\pi)^{-1} |\Omega_{IBD=j}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}_{IBD=j}^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

$$\approx \prod_i (2\pi)^{-1} |\Omega^*|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{*-1}(\mathbf{y}-\boldsymbol{\mu})}$$

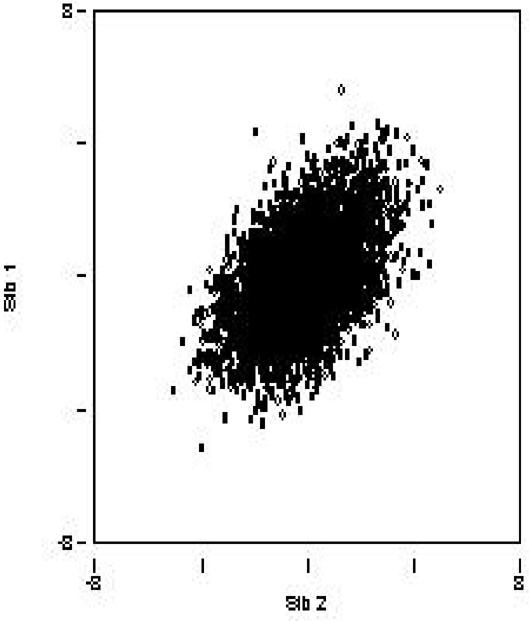
$Z_{ij} = P(IBD_i = j \,|\, \text{marker data})$ IBD sharing probabilities

$\Omega^* = \sum_{j=0,1,2} Z_{ij} \Omega_{IBD=j}$ "Expected" Ω

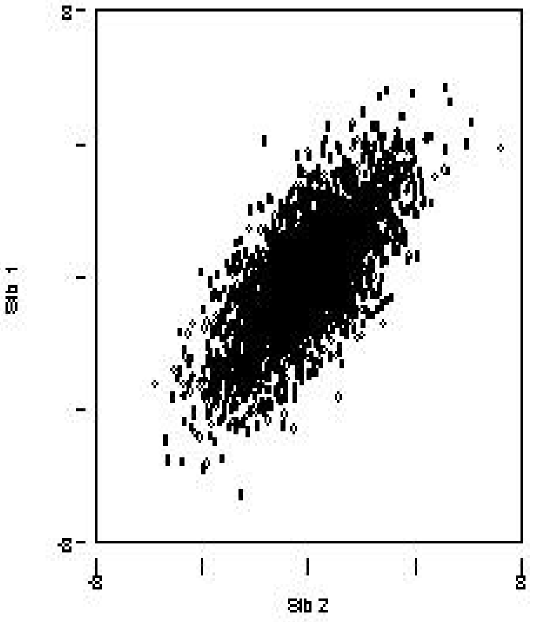
Linkage



IBD 0

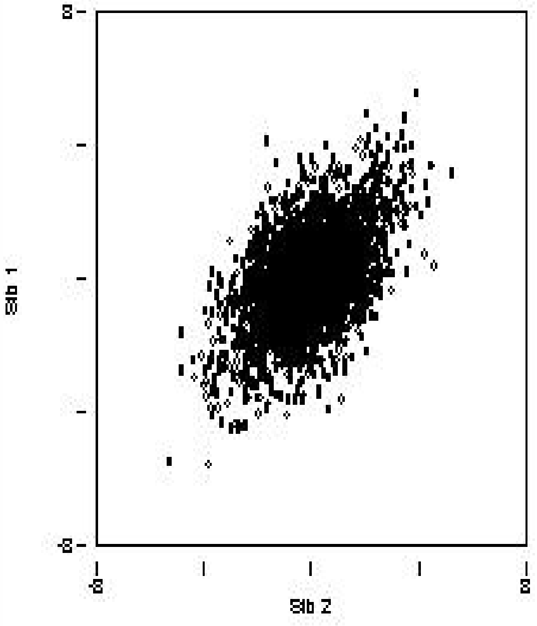


IBD 1

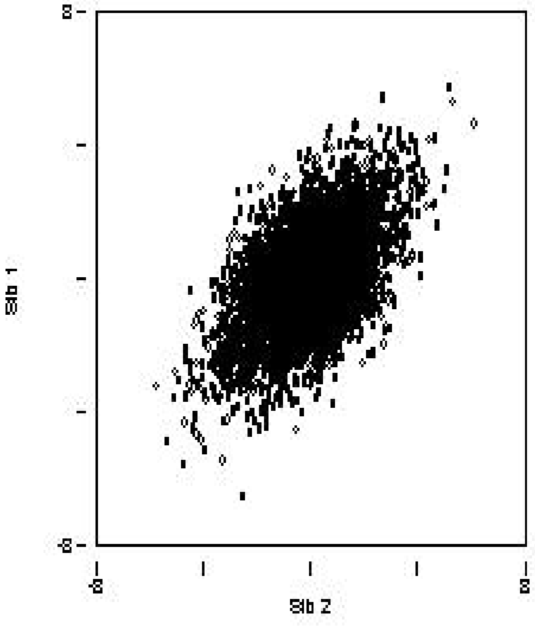


IBD 2

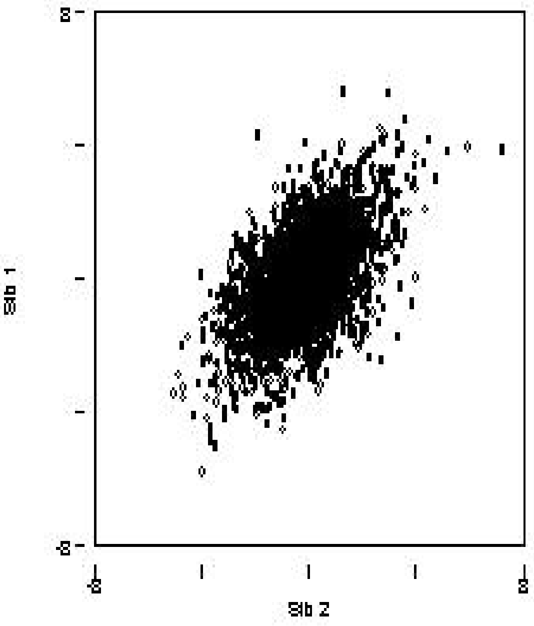
No Linkage



IBD 0



IBD 1



IBD 2

The Problem

- Maximum likelihood variance components linkage analysis
  - Powerful (Fulker & Cherny 1996) but
    - Not robust in selected samples or non-normal traits
    - Conditioning on trait values (Sham et al 2000) improves robustness but is computationally challenging
- Haseman-Elston regression
  - More robust but
    - Less powerful
    - Applicable only to sib pairs

Aim

- To develop a regression-based method that
  - Has same power as maximum likelihood variance components, for sib pair data
  - Will generalise to general pedigrees

Extension to General Pedigrees

- Multivariate Regression Model
- Weighted Least Squares Estimation
- Weight matrix based on IBD information

Switching Variables

- To obtain unbiased estimates in selected samples
  - Dependent variables = IBD
  - Independent variables = Trait

Dependent Variables

- Estimated IBD sharing of all pairs of relatives
- Example:

$$\hat{\mathbf{\Pi}} = \begin{bmatrix} \hat{\pi}_{12} \\ \hat{\pi}_{13} \\ \hat{\pi}_{14} \\ \hat{\pi}_{23} \\ \hat{\pi}_{24} \\ \hat{\pi}_{34} \end{bmatrix}$$

Independent Variables

- Squares and cross-products
  - (equivalent to non-redundant squared sums and differences)
- Example

$$\mathbf{Y} = \begin{bmatrix} x_1x_2 \\ x_1x_3 \\ x_1x_4 \\ x_2x_3 \\ x_2x_4 \\ x_3x_4 \\ x_1x_1 \\ x_2x_2 \\ x_3x_3 \\ x_4x_4 \end{bmatrix}$$

Covariance Matrices

Dependent $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}$

Obtained from prior (p) and posterior (q)
IBD distribution given marker genotypes

$$Cov_I(\hat{\pi}_{ij},\hat{\pi}_{kl}) = \left(\sum p\pi_{ij}\pi_{kl} - \tilde{\pi}_{ij}\tilde{\pi}_{kl}\right) - \left(\sum q\pi_{ij}\pi_{kl} - \hat{\pi}_{ij}\hat{\pi}_{kl}\right)$$

Covariance Matrices

Independent $\Sigma_{\mathbf{Y}}$

Obtained from properties of multivariate normal distribution, under specified mean, variance and correlations

$$E(X_i X_j X_k X_l) = r_{ij} r_{kl} + r_{ik} r_{jl} + r_{il} r_{jk}$$

Assuming the trait has mean zero and variance one. Calculating this matrix requires the correlation between the different relative pairs to be known.

Estimation

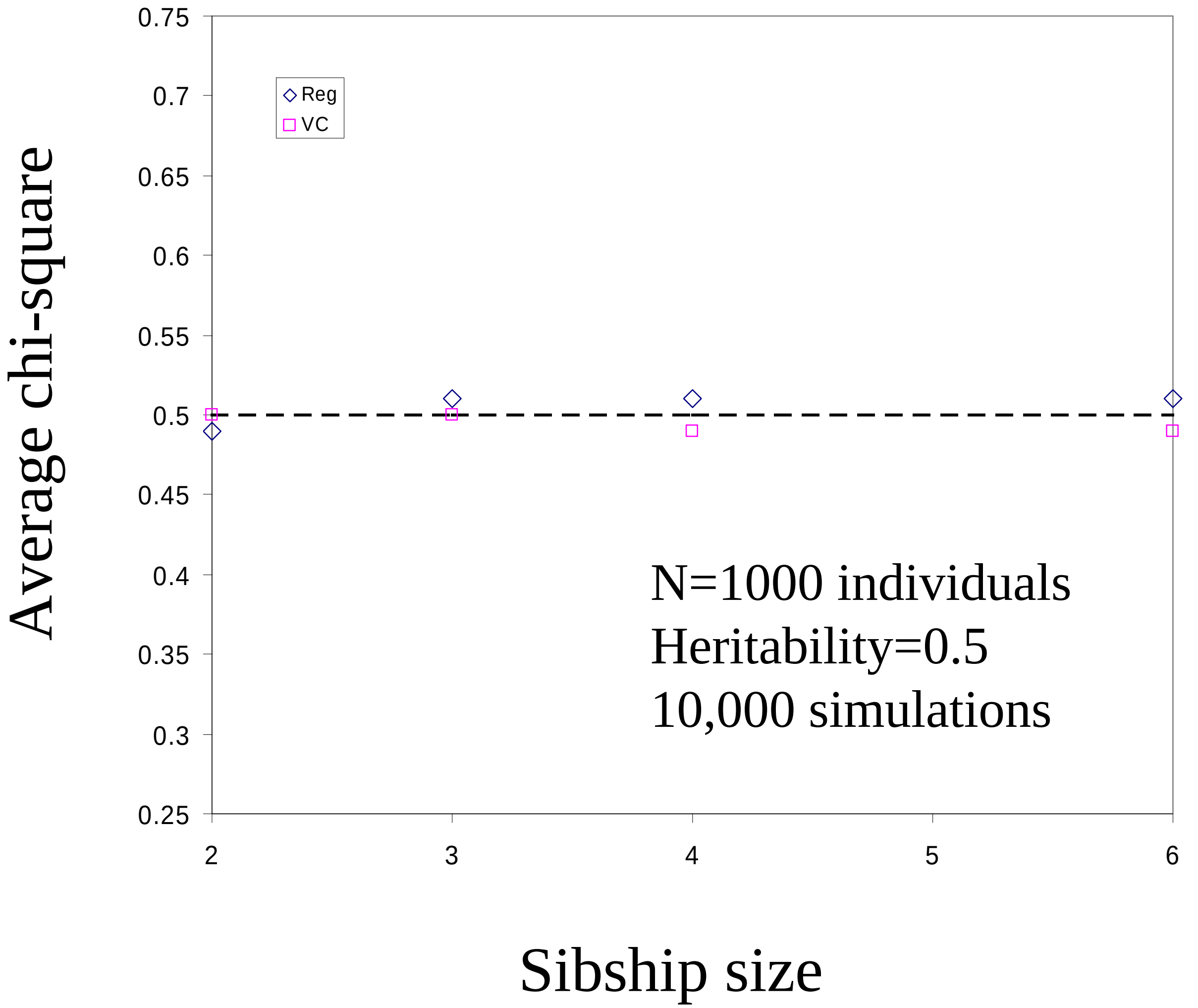
For a family, regression model is

$$\hat{\mathbf{\Pi}}_{\mathbf{C}} = \mathbf{Q}\mathbf{\Sigma}_{\hat{\mathbf{\Pi}}}\mathbf{H}\mathbf{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_{\mathbf{C}} + \boldsymbol{\varepsilon}$$

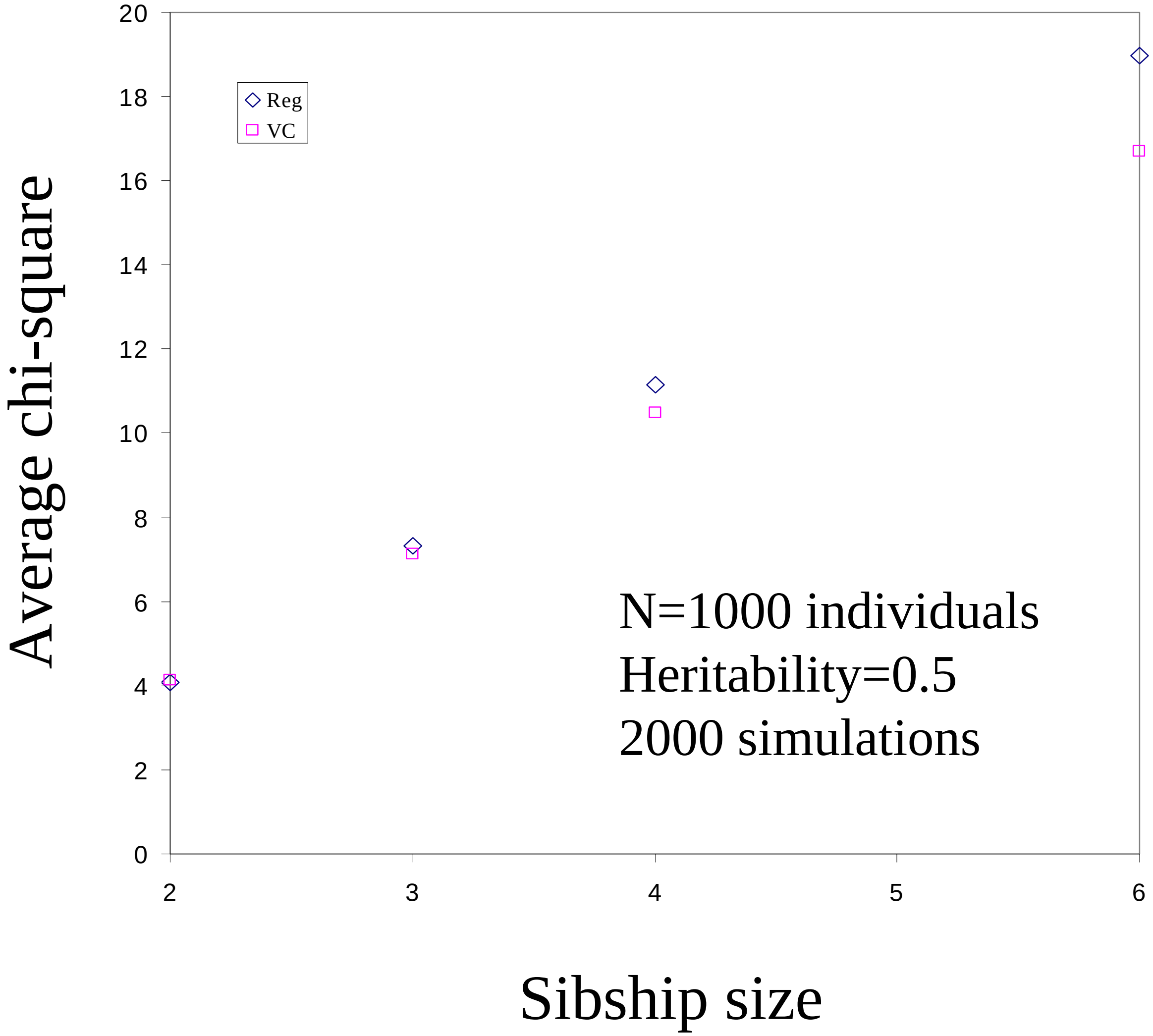
Estimate Q by weighted least squares, and obtain sampling variance, family by family

Combine estimates across families, inversely weighted by their variance, to give overall estimate, and its sampling variance

Average chi-squared statistics: fully informative marker NOT linked to 20% QTL

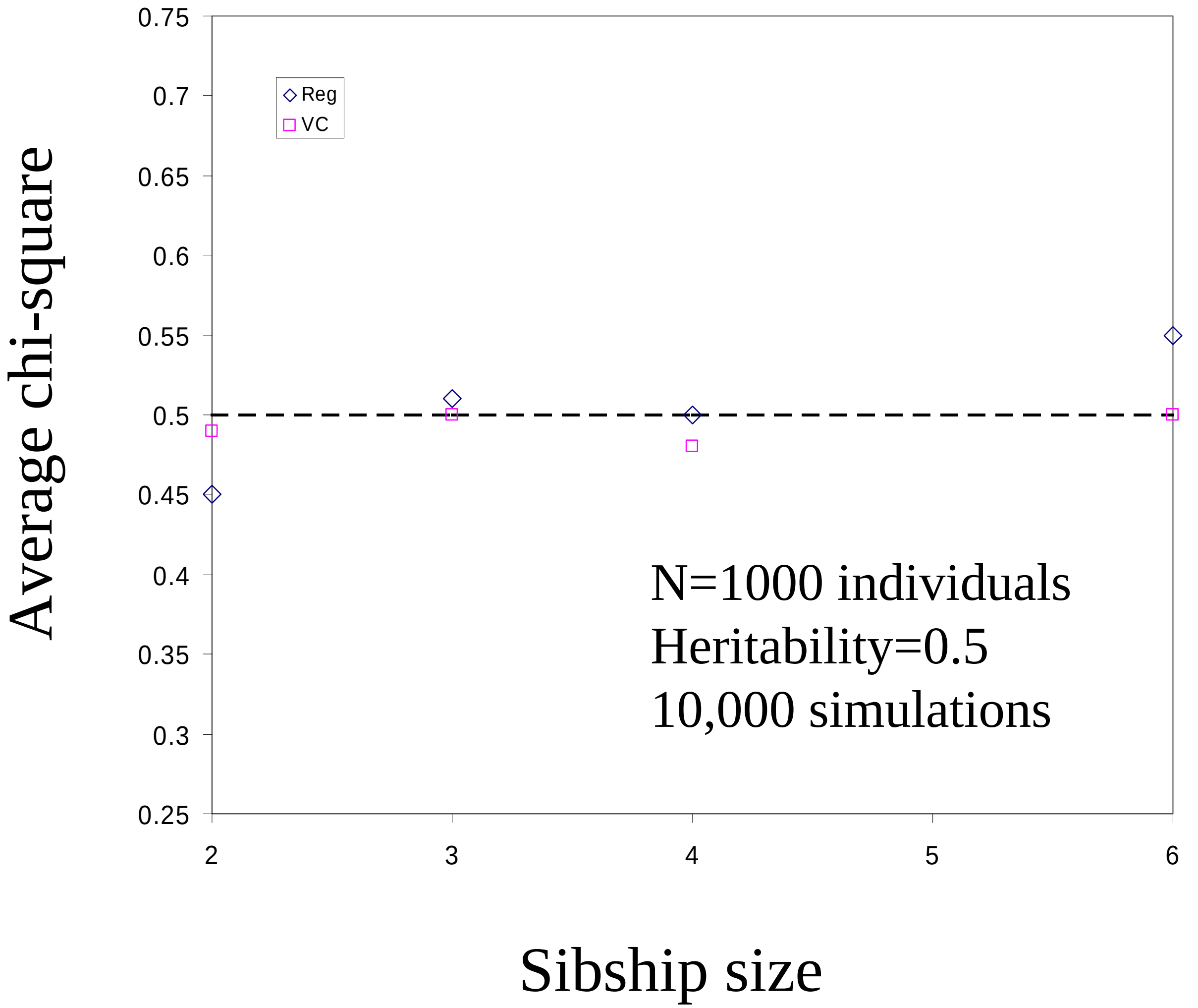


Average chi-squared statistics: fully informative marker linked to 20% QTL



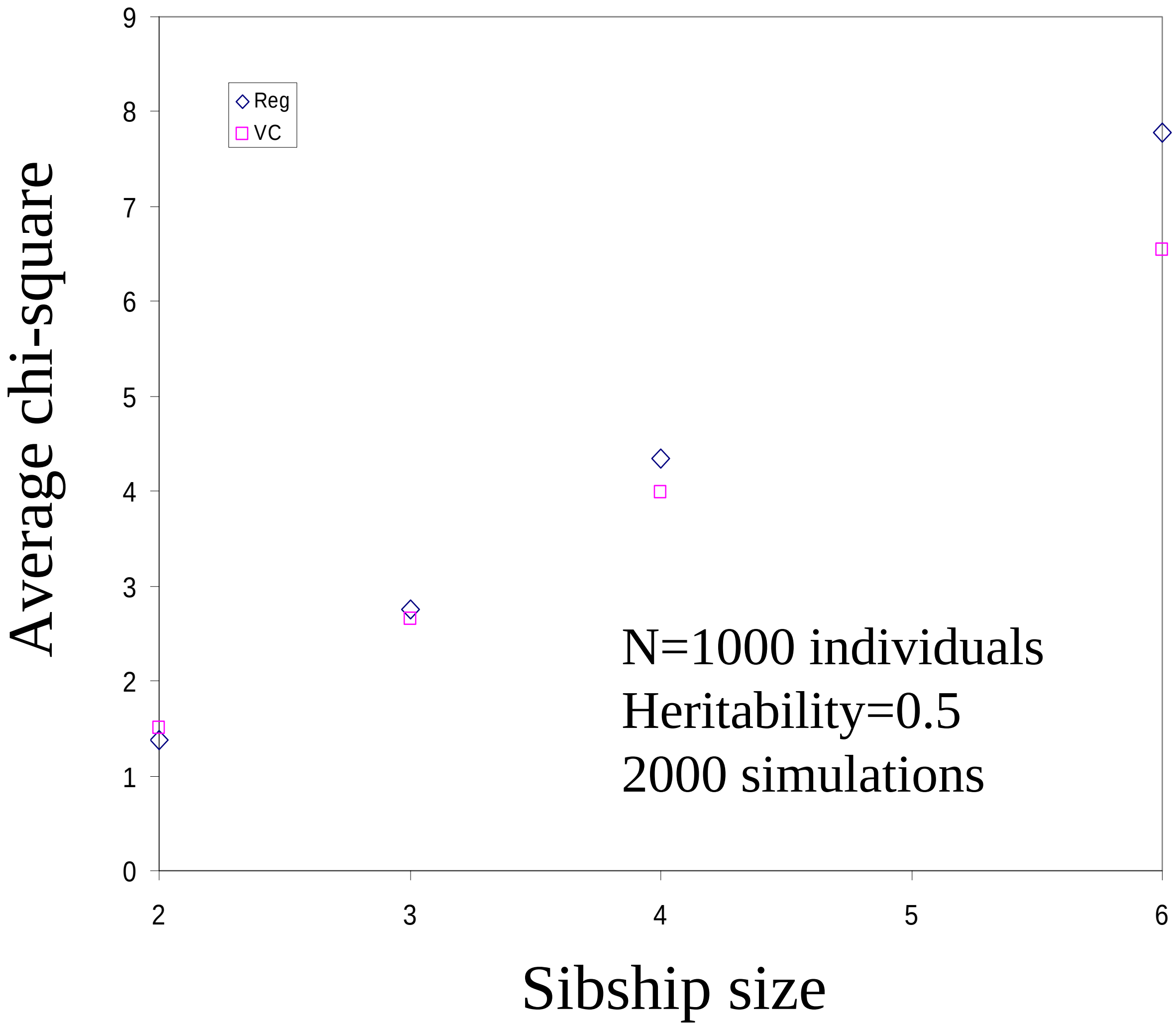
N=1000 individuals
Heritability=0.5
2000 simulations

Average chi-squared statistics: poorly informative marker NOT linked to 20% QTL



N=1000 individuals
Heritability=0.5
10,000 simulations

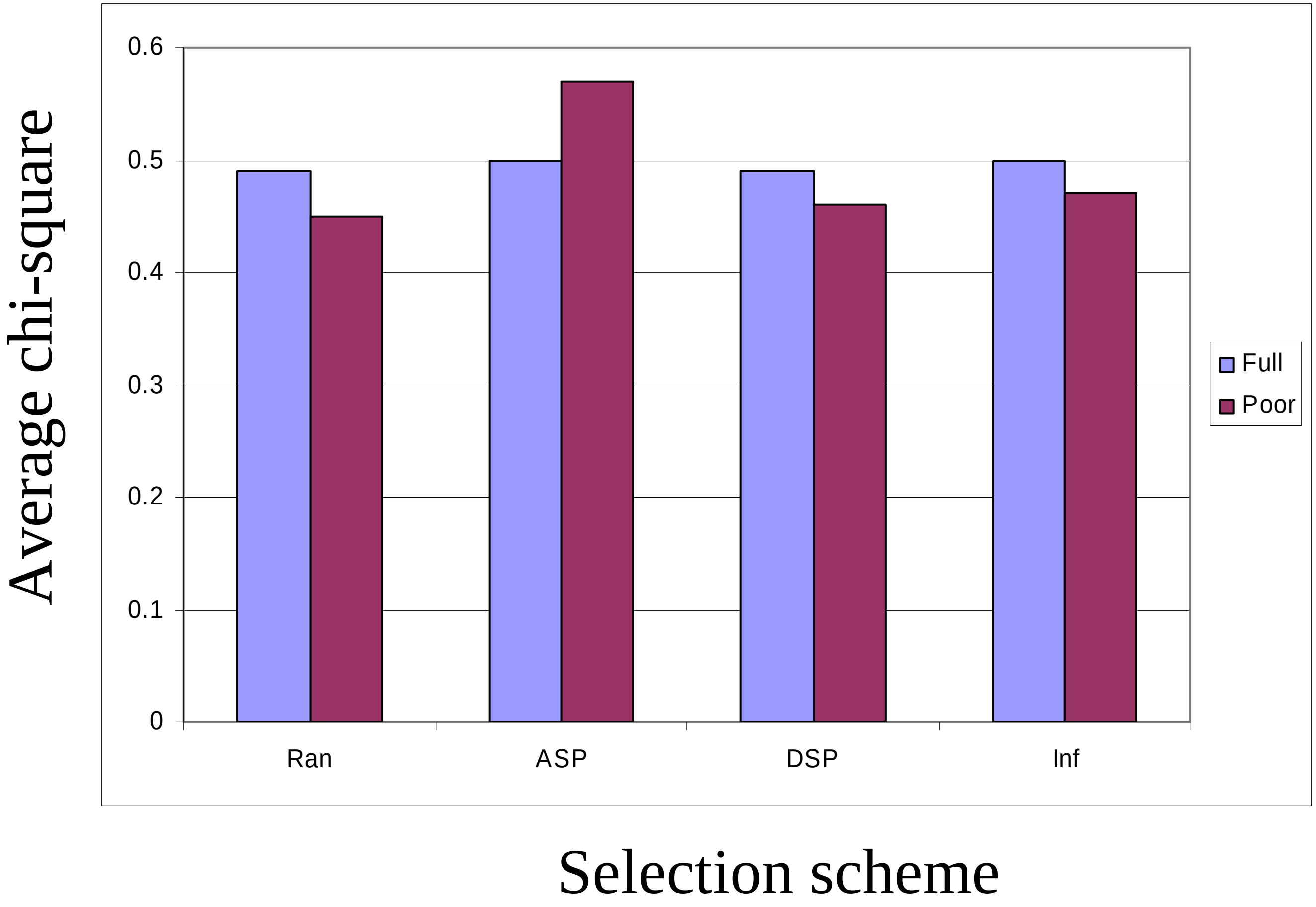
Average chi-squared statistics: poorly informative marker linked to 20% QTL



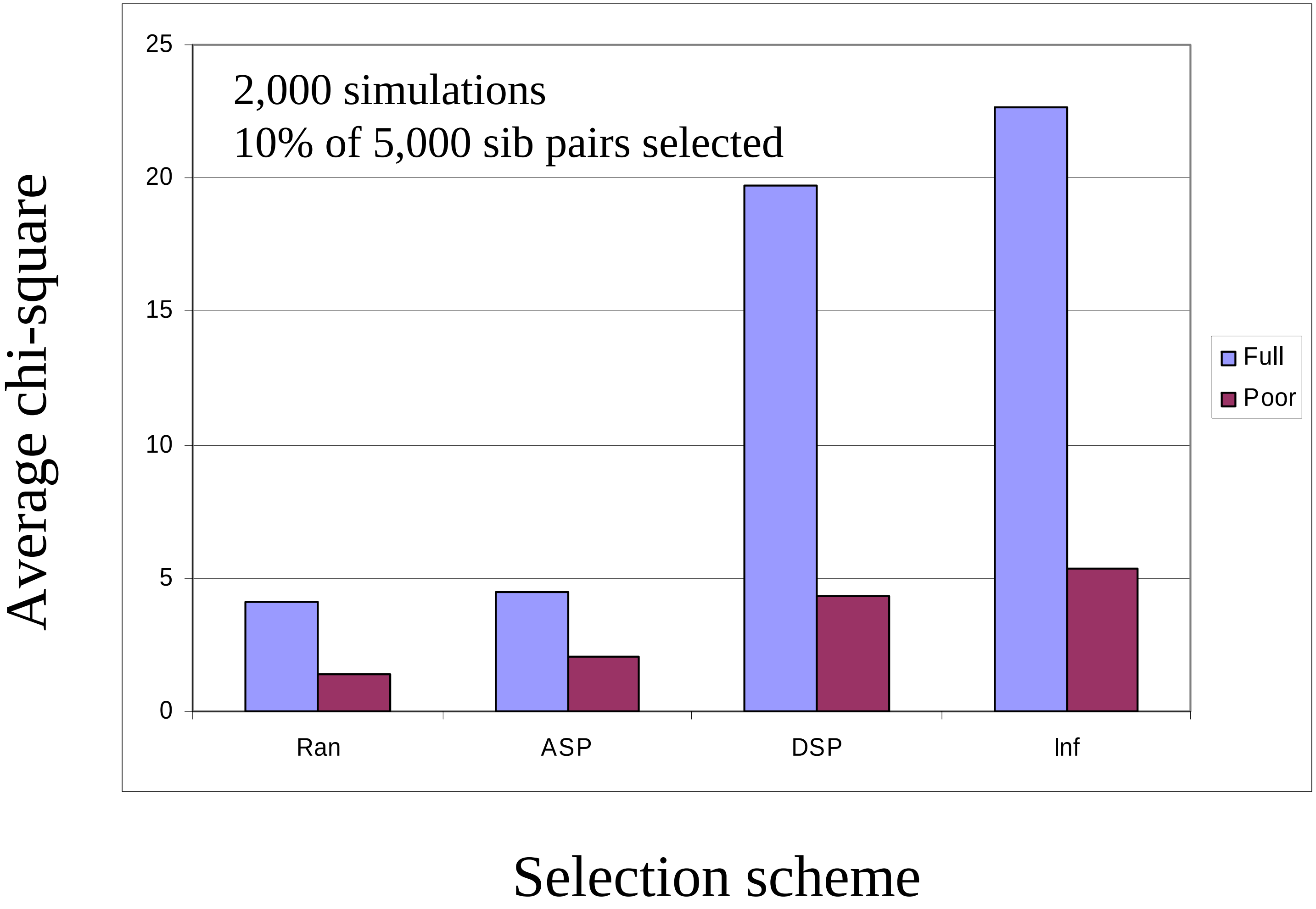
N=1000 individuals
Heritability=0.5
2000 simulations

Average chi-squares:
selected sib pairs, NOT linked to 20% QTL

20,000 simulations
10% of 5,000 sib pairs selected

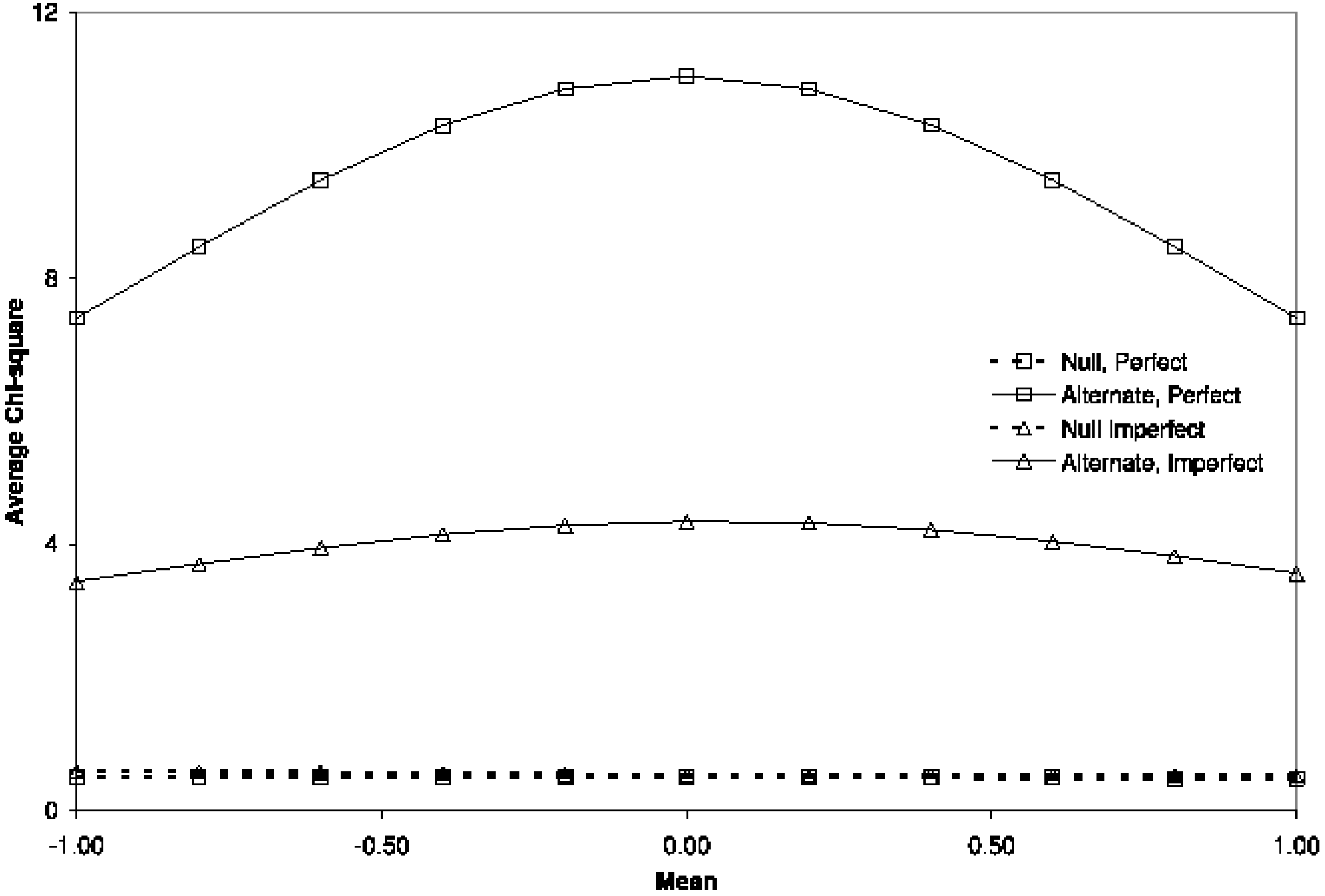


Average chi-squares: selected sib pairs, linkage to 20% QTL

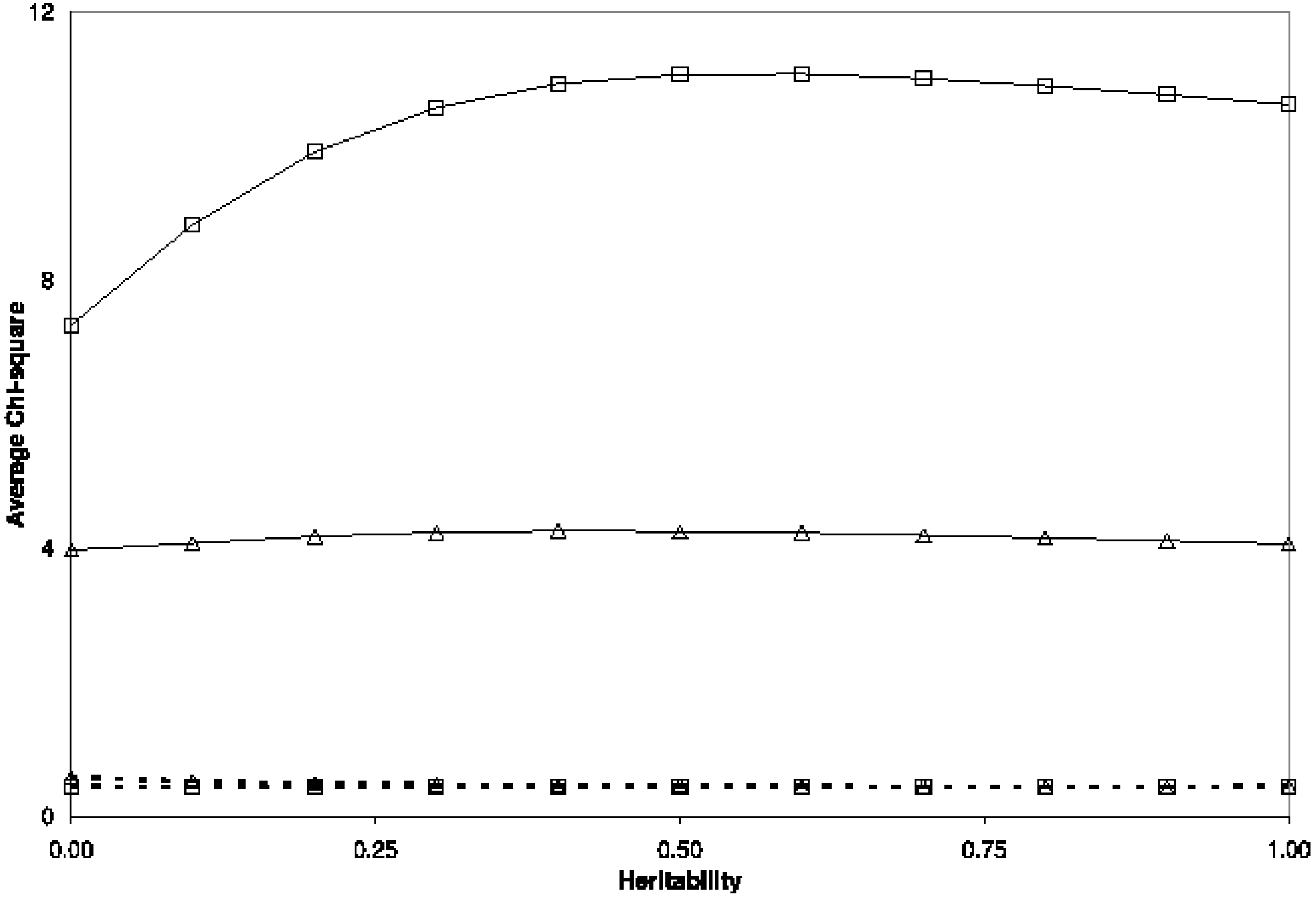


2,000 simulations
10% of 5,000 sib pairs selected

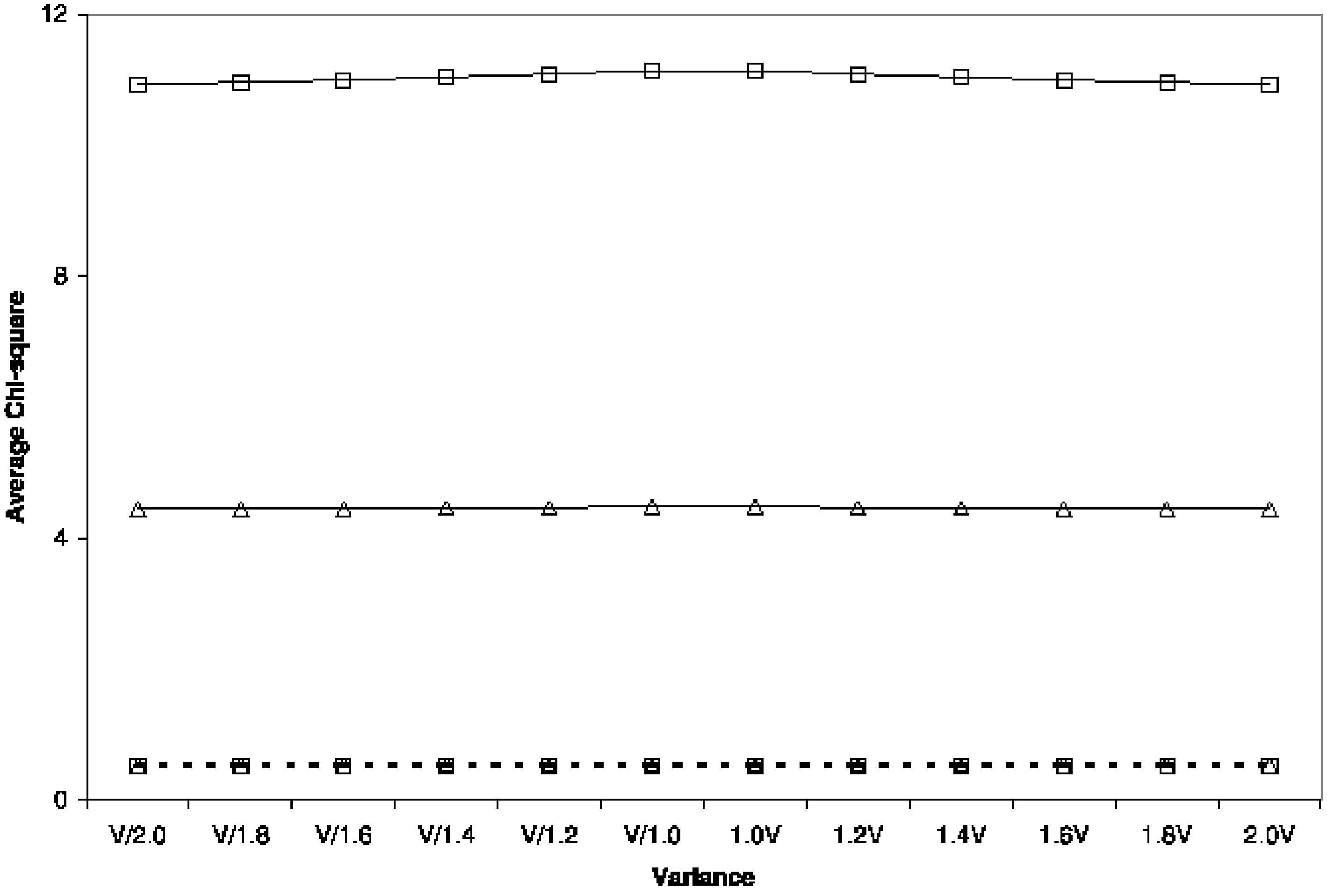
Mis-specification of the mean, 2000 random sib quads, 20% QTL



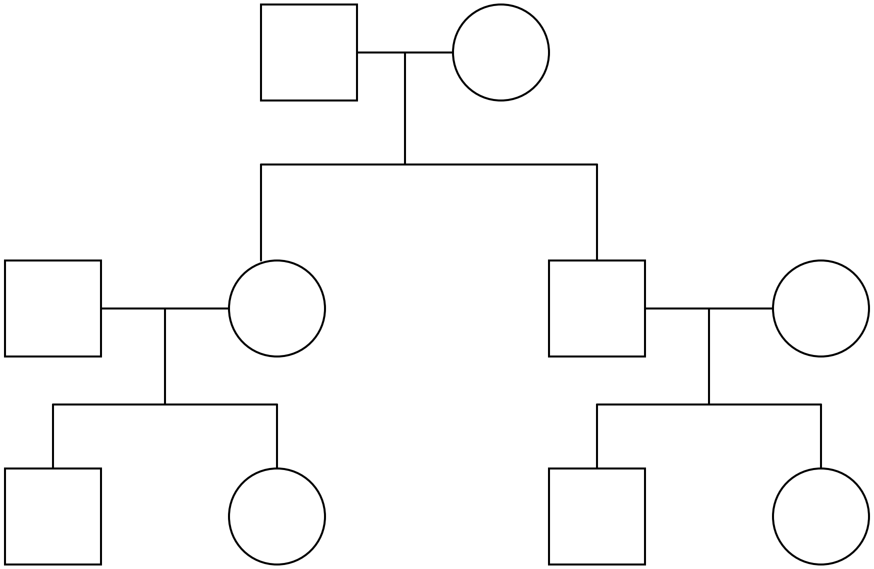
Mis-specification of the covariance, 2000 random sib quads, 20% QTL



Mis-specification of the variance, 2000 random sib quads, 20% QTL



Cousin pedigree



Average chi-squares for 200 cousin pedigrees, 20% QTL

	Poor marker information		Full marker information	
	REG	VC	REG	VC
Not linked	0.49	0.48	0.53	0.50
Linked	4.94	4.43	13.21	12.56

Conclusion

- The regression approach
  - can be extended to general pedigrees
  - is slightly more powerful than maximum likelihood variance components in large sibships
  - can handle imperfect IBD information
  - is easily applicable to selected samples
  - provides unbiased estimate of QTL variance
  - provides simple measure of family informativeness
  - is robust to minor deviation from normality
- But
  - assumes knowledge of mean, variance and covariances of trait distribution in population

Example Application: Angiotensin Converting Enzyme

- British population

- Circulating ACE levels
  - Normalized separately for males / females

- 10 di-allelic polymorphisms
  - 26 kb
  - Common
  - In strong linkage disequilibrium

- Keavney et al, HMG, 1998

Check The Data

- The input data is in three files:
  - keavney.dat
  - keavney.ped
  - keavney.map

- These are text files, so you can peek at their contents, using `more` or `notepad`

- A better way is to used `pedstats` …

Pedstats

- Checks contents of pedigree and data files

  ```
  pedstats –d keavney.dat –p keavney.ped
  ```

- Useful options:

--pairStatistics	Information about relative pairs
--pdf	Produce graphical summary
--hardyWeinberg	Check markers for HWE
--minGenos 1	Focus on genotyped individuals

- What did you learn about the sample?

Regression Analysis

- MERLIN-REGRESS

- Requires pedigree (.ped), data (.dat) and map (.map) file as input

- Key parameters:
  - --mean, --variance
    - Used to standardize trait
  - --heritability
    - Use to predicted correlation between relatives

- Heritability for ACE levels is about 0.60

MERLIN-REGRESS

- Identify informative families
  - --rankFamilies

- Customizing models for each trait
  - -t models.tbl
  - TRAIT, MEAN, VARIANCE, HERITABILITY in each row

- Convenient options for unselected samples:
  - --randomSample
  - --useCovariates
  - --inverseNormal

The End