Regression Models for Linkage: Merlin Regress

Pak Sham, Shaun Purcell,
Stacey Cherny, Goncalo Abecasis

Problem with VC linkage analysis

- Maximum likelihood variance components linkage analysis
  - Powerful but
    - Prone to give false positives in selected samples or non-normal traits
    - Ascertainment correction may allow VC to be applied to selected samples
    - Conditioning on trait values deals with both selection and non-normality, but is computationally intensive in large pedigrees

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

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

Suitable for selected samples, and robust to normality
BUT: Less powerful than VC linkage analysis

Why is HE regression less powerful?

- Wright (1997), Drigalenko (1998)
  - phenotypic difference discards sib-pair QTL linkage information
  - squared pair trait **sum** (mean-corrected) provides extra information for linkage

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

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
  - cross-product (mean-corrected) (HE-CP)

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

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

Am. J. Hum. Genet. 68:1527–1532, 2001

Report

Equivalence between Haseman-Elston and Variance-Components Linkage Analyses for Sib Pairs

P. C. Sham and S. Purcell

Social, Genetic & Developmental Research Centre, Institute of Psychiatry, London

- Clarify the relative efficiencies of existing HE methods
- Demonstrate equivalence between a new HE method and variance components methods
- Show application to the selection and analysis of extreme, selected samples

NCPs for H-E regressions

Dependent	*Variance*	*NCP per sibpair*
$(X - Y)^2$	$8(1 - r)^2$	$\frac{Q^2 Var(\hat{\pi})}{2(1-r)^2}$
$(X + Y)^2$	$8(1 + r)^2$	$\frac{Q^2 Var(\hat{\pi})}{2(1+r)^2}$
XY	$1 + r^2$	$\frac{Q^2 Var(\hat{\pi})}{1+r^2}$

Combining into one regression

- New dependent variable :
  - a linear combination of
    - squared-sum
    - squared-difference
      - Inversely weighted by their variances:

$$\frac{(X+Y)^2}{(1+r)^2}-\frac{(X-Y)^2}{(1-r)^2}=-\frac{4r}{1-r^2}+\frac{4(1+r^2)}{(1-r^2)^2}Q(\hat{\pi}-0.5)+\varepsilon$$

Weighted H-E

$$NCP = Q^2 Var(\hat{\pi}) \frac{(1+r^2)}{(1-r^2)^2}$$

- A function of
  - square of QTL variance
  - marker informativeness
    - complete information: Var($\hat{\pi}$)=1/8
  - sibling correlation
- Equivalent to variance components
  - to second-order approximation
    - Rijsdijk *et al* (2000)

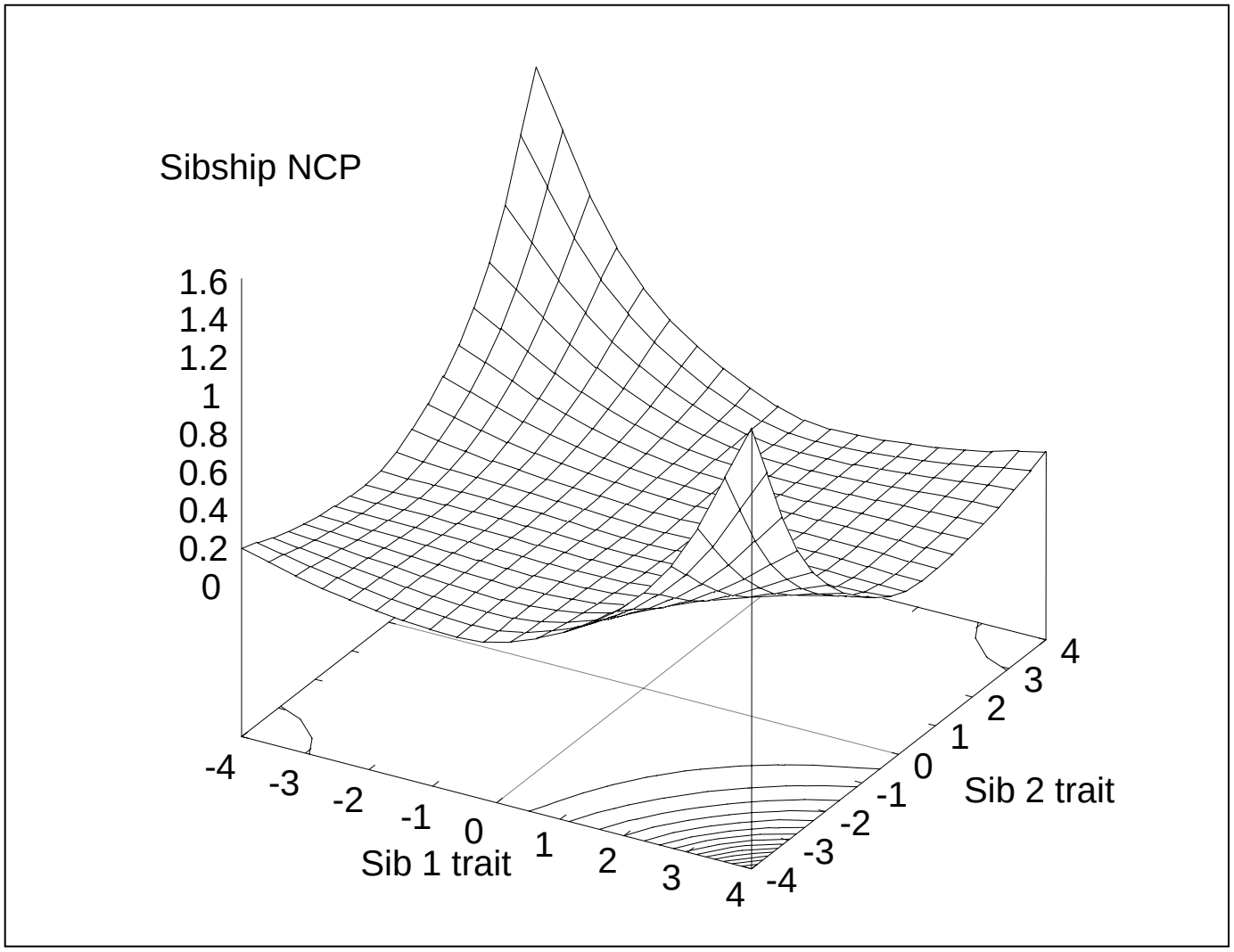
Sample selection

- A sib-pairs' squared mean-corrected DV is proportional to its expected NCP

$$E(NCP|trait) \propto \left(\frac{(X+Y)^2}{(1+r)^2} - \frac{(X-Y)^2}{(1-r)^2} + \frac{4r}{1-r^2} \right)^2$$

- Equivalent to variance-components based selection scheme
  - Purcell *et al*, (2000)

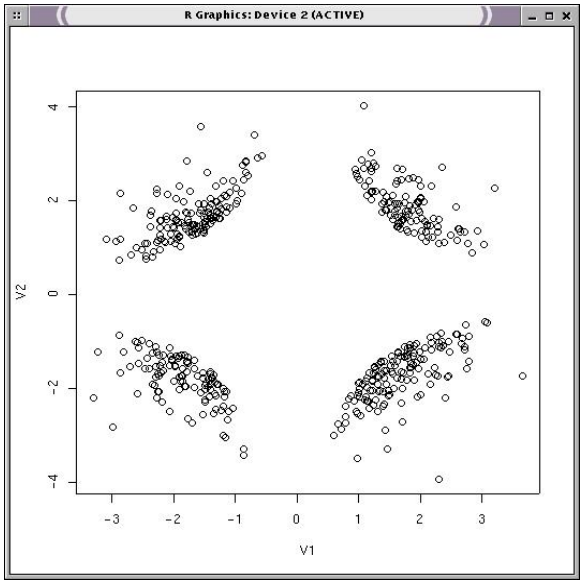
Information for linkage



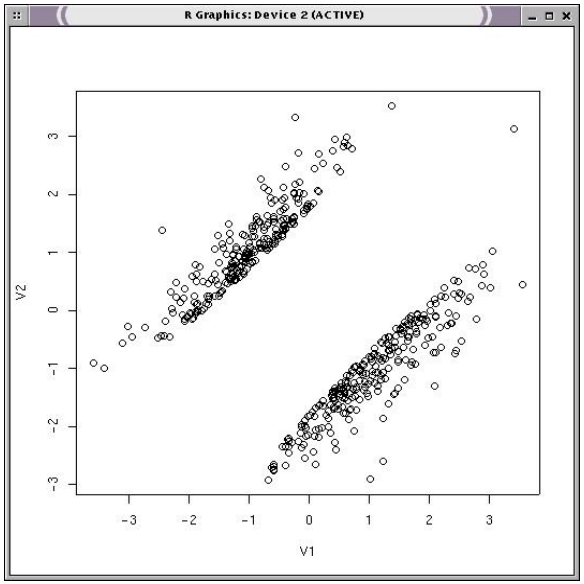
Selected samples

- 5% most informative pairs selected

$r = 0.05$



$r = 0.60$



Am. J. Hum. Genet. 71:238–253, 2002

Powerful Regression-Based Quantitative-Trait Linkage Analysis of General Pedigrees

Pak C. Sham,[1] Shaun Purcell,[1] Stacey S. Cherny,[1,2] and Gonçalo R. Abecasis[3]

[1]Institute of Psychiatry, King's College, London; [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford; and [3]Center for Statistical Genetics, University of Michigan, Ann Arbor

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:

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

Independent Variables

- Squares and cross-products (mean-corrected)
  - (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

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

Handles imperfect marker information

Covariance Matrices

Independent $\boldsymbol{\Sigma}_{\mathbf{Y}}$

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

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

Estimation

For a family, regression model is

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

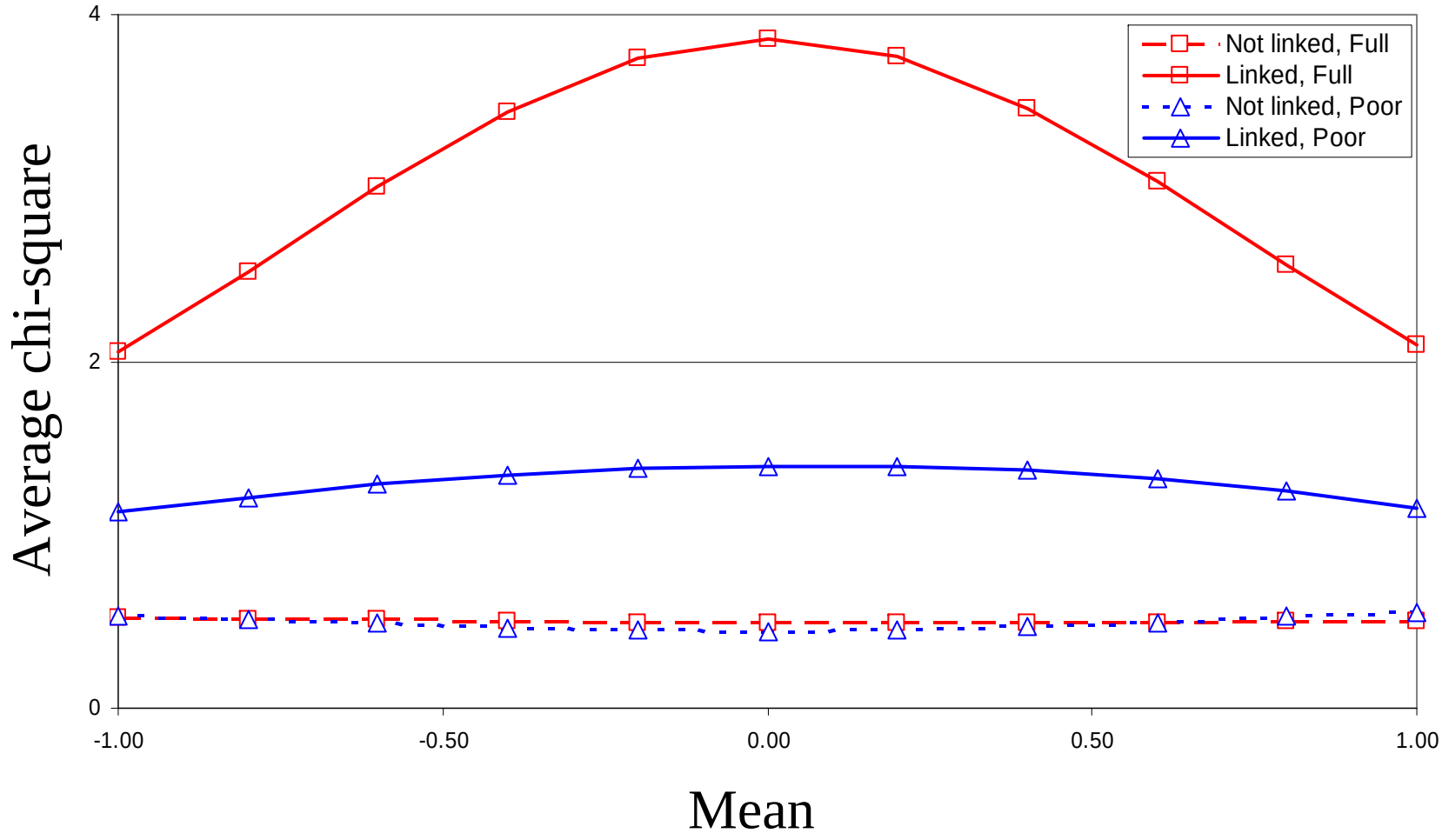
Estimate Q by weighted least squares, and obtain sampling variance, family by family

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

Implementation

- MERLIN-REGRESS
- Requires pedigree (.ped), data (.dat) and map (.map) files as input
- Key parameters:
  - --mean, --variance
    - Used to standardize trait
  - --heritability
    - Use to predict correlations between relatives

Mis-specification of the mean, 500 random sib pairs, 20% QTL



MERLIN-REGRESS Features

- Identifies informative families
  - --rankFamilies
- Provides measure for marker information content at each location
- Option for analyzing repeated measurements
  - --testRetest
- 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