# Efficient calculation of empirical p-values for genome wide linkage through weighted mixtures

Sarah E Medland, Eric J Schmitt,
Bradley T Webb, Po-Hsiu Kuo,
Michael C Neale
Virginia Institute for Psychiatric and Behavioral Genetics

# Standard approaches to evaluating significance

- Nominal p-values based on (presumed) asymptotic null distributions
- Empirical p-values from simulation or ‘gene-dropping’
- Empirical p-values from permutation

# Standard approaches to evaluating significance

- Nominal p-values
  - Pros: computation free
  - Cons: unrealistic expectations of data lead to decreased accuracy
- Empirical p-values
  - Pros: increased accuracy;
    explicit correction for the data distributions
  - Cons: computationally intensive;
    require a degree of programming skill (or access to a programmer)

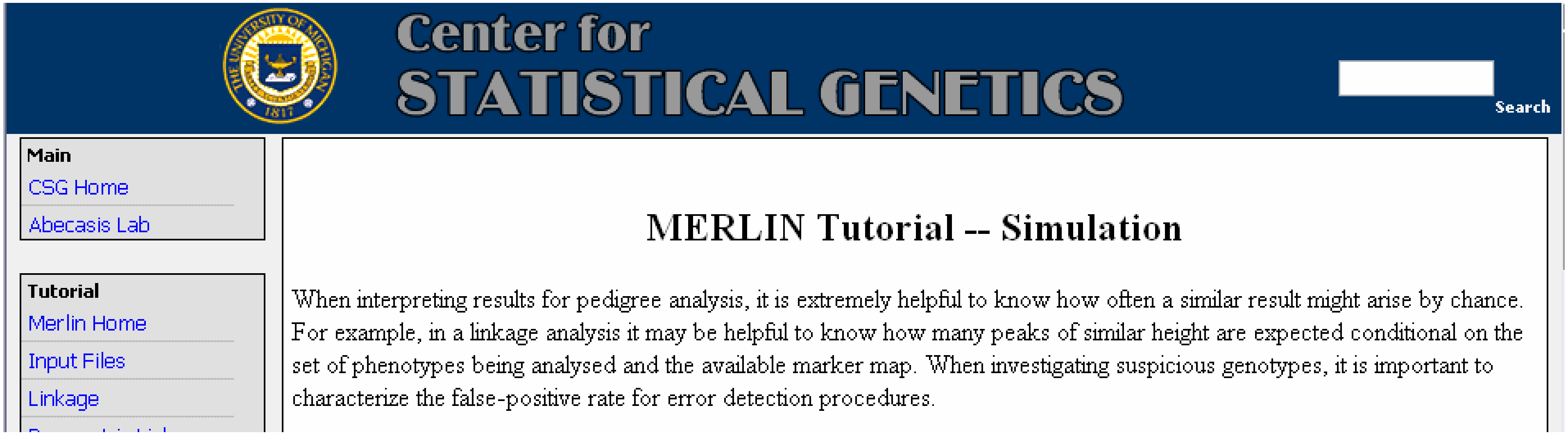
# Gene-dropping vs permutation

- Both produce asymptotically unbiased estimates of significance

  Ott, 1989; Churchill & Doerge, 1994

- Gene-dropping is implemented in the software most commonly used to analyze human data

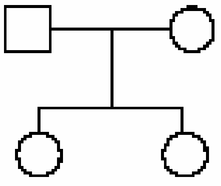
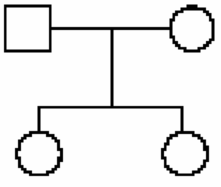
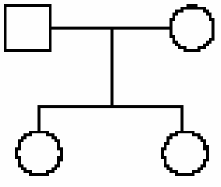
# Gene-dropping simulation



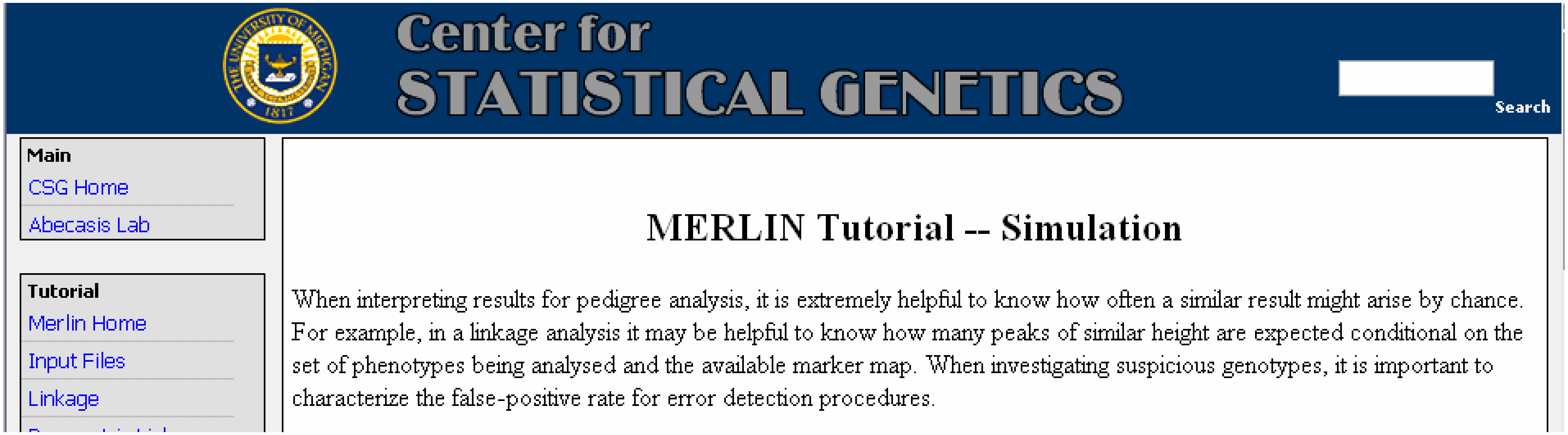
## MERLIN Tutorial -- Simulation

When interpreting results for pedigree analysis, it is extremely helpful to know how often a similar result might arise by chance. For example, in a linkage analysis it may be helpful to know how many peaks of similar height are expected conditional on the set of phenotypes being analysed and the available marker map. When investigating suspicious genotypes, it is important to characterize the false-positive rate for error detection procedures.

# Gene dropping/Simulation

| Ped | Observed | | Sim1 | | Sim2 | | Sim3 | |
|---|---|---|---|---|---|---|---|---|
|  | 1/2 | 3/4 | 4/3 | 2/3 | 3/4 | 4/4 | 1/3 | 4/4 |
| | 1/3 | 2/4 | 3/3 | 2/3 | 4/4 | 3/4 | 1/4 | 1/4 |
| $\hat{\pi}$ | 0 | | .5 | | .25 | | .75 | |
|  | 1/3 | 4/4 | 1/4 | 3/4 | 1/4 | 3/4 | 2/2 | 4/4 |
| | 1/4 | 3/4 | 3/4 | 3/4 | 1/3 | 4/4 | 2/4 | 2/4 |
| $\hat{\pi}$ | .25 | | 1 | | 0 | | .5 | |
|  | 1/3 | 2/4 | 4/1 | 1/2 | 4/2 | 1/1 | 1/2 | 1/4 |
| | 1/4 | 1/4 | 2/4 | 1/4 | 1/2 | 1/2 | 1/1 | 1/1 |
| $\hat{\pi}$ | 1 | | .5 | | .25 | | 1 | |

# Gene-dropping simulation



MERLIN Tutorial -- Simulation

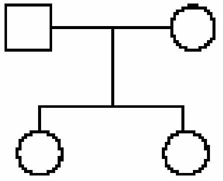
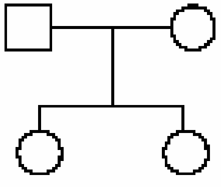
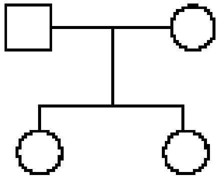
When interpreting results for pedigree analysis, it is extremely helpful to know how often a similar result might arise by chance. For example, in a linkage analysis it may be helpful to know how many peaks of similar height are expected conditional on the set of phenotypes being analysed and the available marker map. When investigating suspicious genotypes, it is important to characterize the false-positive rate for error detection procedures.

```
prompt> merlin -d c1.dat -m c1.map -p c1.ped
--vc --simul --reruns 1000 -r 1234 --save
```

# Gene-dropping simulation

- Extract the LOD or chi-square from the output of these null replicates
- Add in your observed LOD or chi-square
- Sort the file and calculate the probability of the observed using the simulated null distribution

# Permutation

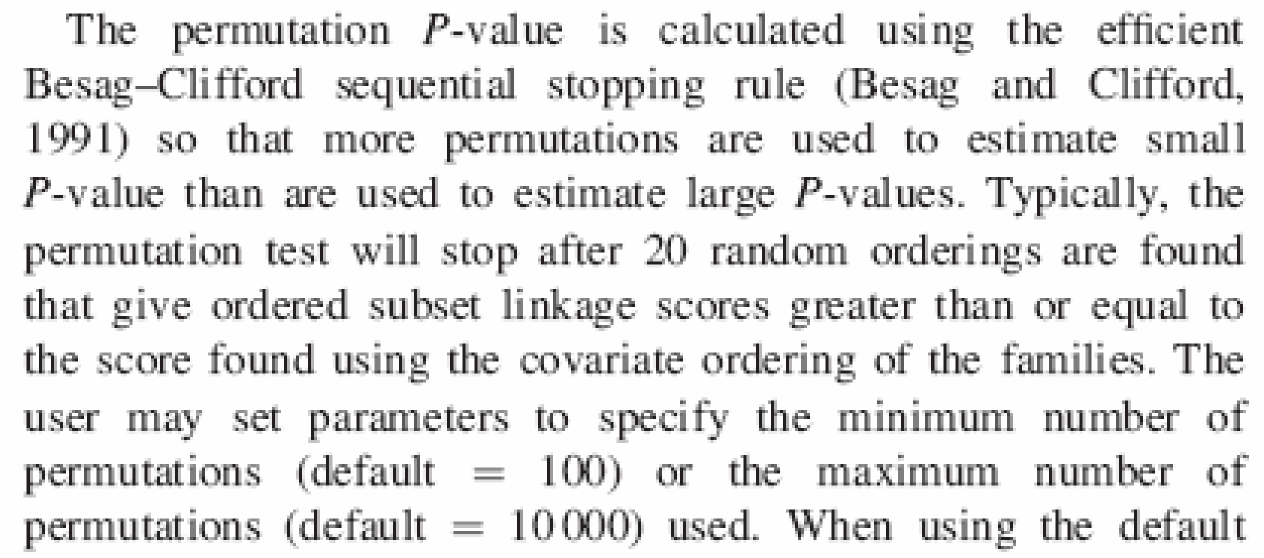
| Ped | Observed | P1 | P2 | P3 |
|---|---|---|---|---|
|  | 1/2 3/4<br>1/3 2/4<br>$\hat{\pi}$ 0 | .25 | 1 | 0 |
|  | 1/3 4/4<br>1/4 3/4<br>$\hat{\pi}$ .25 | 1 | 0 | 1 |
|  | 1/3 2/4<br>1/4 1/4<br>$\hat{\pi}$ 1 | 0 | .25 | .25 |

# Improving the efficiency of empirical p-value estimation

- Sequential stopping rules
  - Less simulations for lower LOD scores

    **Besag & Clifford (1991) Biometrika 78 p301**
  - Implementation in FLOSS

    Browning (2006) Bioinformatics Applications Note, 22 p512

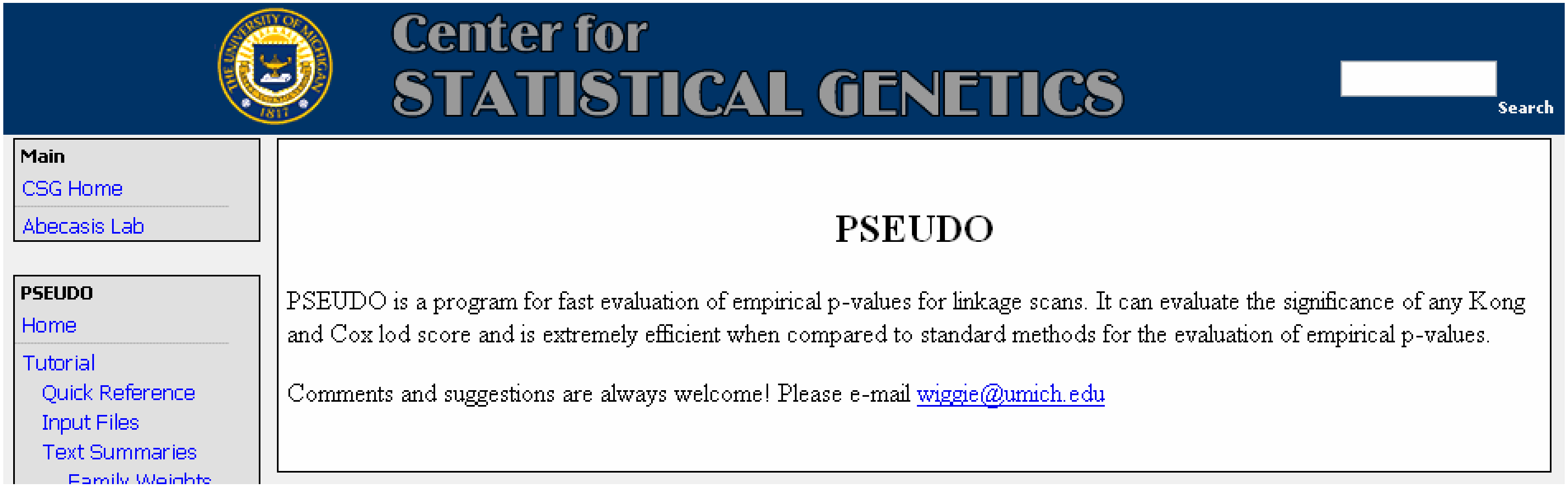


The permutation *P*-value is calculated using the efficient Besag–Clifford sequential stopping rule (Besag and Clifford, 1991) so that more permutations are used to estimate small *P*-value than are used to estimate large *P*-values. Typically, the permutation test will stop after 20 random orderings are found that give ordered subset linkage scores greater than or equal to the score found using the covariate ordering of the families. The user may set parameters to specify the minimum number of permutations (default = 100) or the maximum number of permutations (default = 10 000) used. When using the default

# Improving the efficiency of empirical p-value estimation

- Replicate Pool Method
  - Run a small number of simulations/permutations saving the per family contributions
  - Resample from the ‘pool’ of null replicates

    Terwilliger & Ott, 1992; Song et al, 2004; Zou et al, 1995; Wigginton & Abecais, 2006



PSEUDO

PSEUDO is a program for fast evaluation of empirical p-values for linkage scans. It can evaluate the significance of any Kong and Cox lod score and is extremely efficient when compared to standard methods for the evaluation of empirical p-values.

Comments and suggestions are always welcome! Please e-mail wiggie@umich.edu

$\hat{\pi}$

# Proposed 'Weighted Mixture' Method

- Significance values derived by permutation depend on distributions of the allele sharing statistic: $\hat{\pi}$
    - Any 2 loci with identical $\hat{\pi}$ distributions *will* yield identical empirical p-values
- There is relatively little variation in the distribution of $\hat{\pi}$ across loci

# Proposed 'Weighted Mixture' Method

- Hypothesis: it is possible to approximate the distribution of $\hat{\pi}$ at a given locus *x* by creating a weighted mixture of *l* loci
- If so, the p-value obtained from the weighted mixture should be a good approximation of the p-value obtained by traditional permutation

# Proposed 'Weighted Mixture' Method

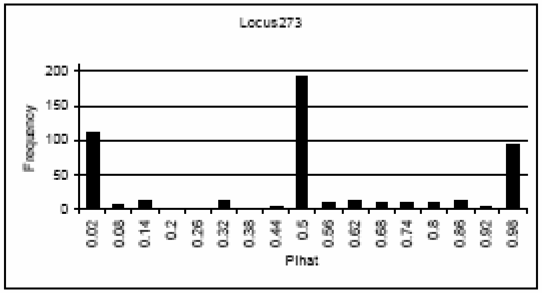
- Five part process implemented in R
  - Bin the $\hat{\pi}$ distribution
    - We used 51 bins ($\hat{\pi}$/50 rounded to the nearest integer)

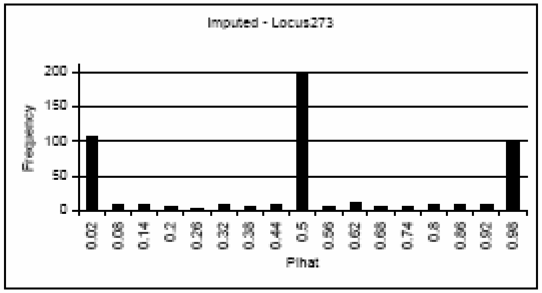
# Proposed ‘Weighted Mixture’ Method

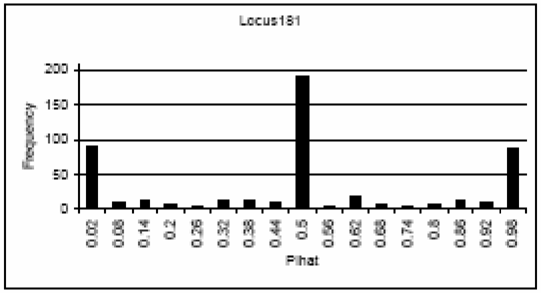
- Five part process implemented in R
  - Bin the $\hat{\pi}$ distribution
  - Identify a pool of ‘modal’ distributions
    - ❑We experimented with systematic and random identification of the distributions
    - ❑Best results obtained using the Bioconductor package GENEFINDER
    - ❑Bin frequencies were entered as an array and the 5 most similar distributions were identified using a Euclidean distance metric
    - ❑We experimented using pools of the 50, 20 and 10 most commonly identified distributions

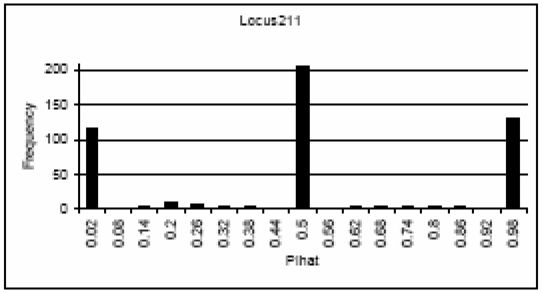
# Proposed 'Weighted Mixture' Method

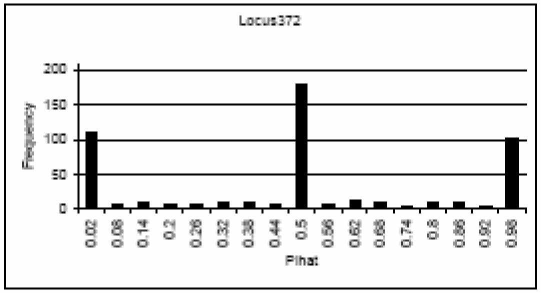
- Five part process implemented in R
  - Bin the $\hat{\pi}$ distribution
  - Identify a pool of 'modal' distributions
  - Obtain mixture weights (*w*)
    - Simplified multivariate regression using weighted least squares using modal distributions as predictors
    - Estimate the distribution of each locus in turn, recording the regression weights

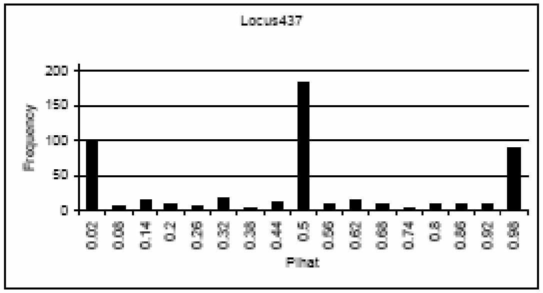


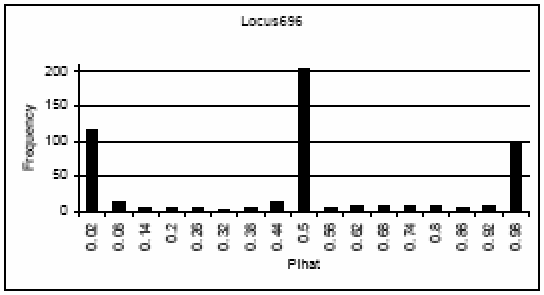


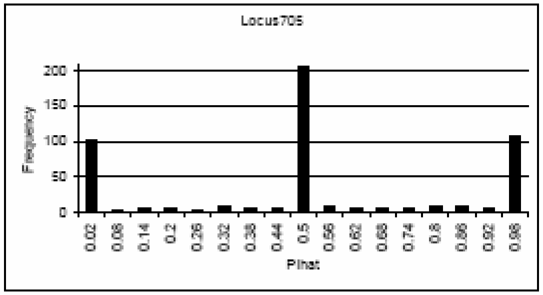












Locus 273 imputed = .28*Locus181+.01*Locus211+.10*Locus372+.06*Locus437+.37*Locus696+.18*Locus705

# Proposed 'Weighted Mixture' Method

- Five part process implemented in R
  - Bin the $\hat{\pi}$ distribution
  - Identify a pool of 'modal' distributions
  - Obtain mixture weights (*w*)
  - Permute the modal loci
    - ❑ n=5000
    - ❑ Test statistic $\chi^2$ retained for each permutation

# Proposed ‘Weighted Mixture’ Method

- Five part process implemented in R
  - Bin the $\hat{\pi}$ distribution
  - Identify a pool of ‘modal’ distributions
  - Obtain mixture weights (*w*)
  - Permute the modal loci
  - Weighted bootstrapping of  the modal test statistics
    - ❑Compile composite test statistic distributions
    - ❑Weighted drawing of $w_i$*5000 random test statistics from each of the *I* modal test statistic distributions
    - ❑Average significance value from 100 replicates

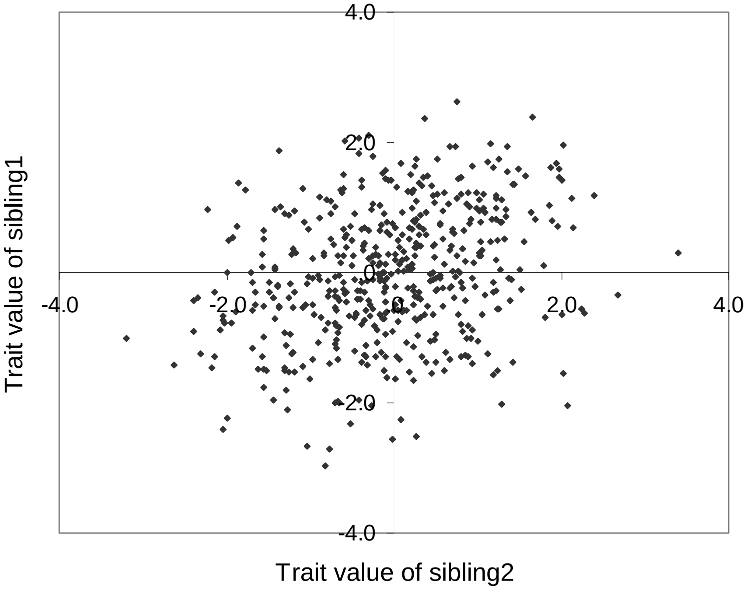
# Simulations

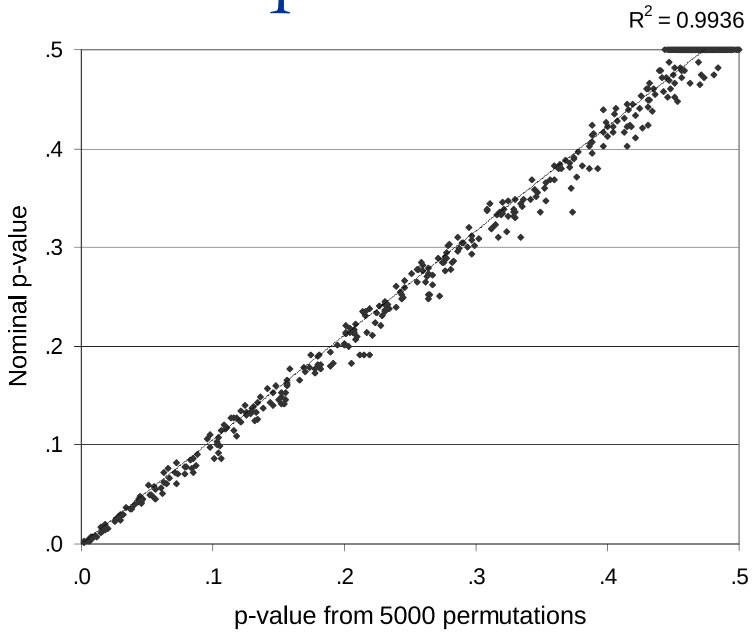
- Simulated genotypes for 500 families
  - 2 parents and 2 offspring
- Map based on the Irish affected sib-pair study of alcohol dependence (Prescott et al, 2006; Kuo, submitted)
  - 1020 autosomal markers (deCODE panel)
  - Average 4cM spacing
- 3 causal and unmeasured bi-allelic loci – on different chromosomes

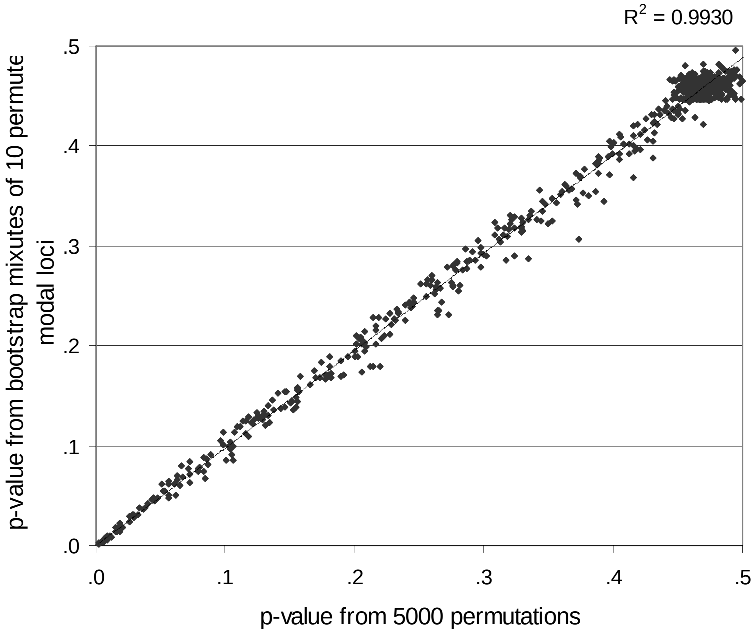
# Simulations

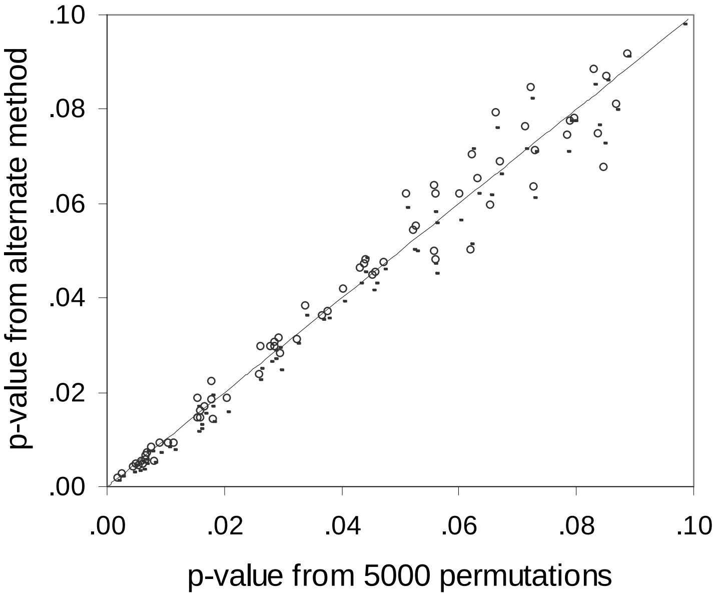
- Phenotypic data simulated under 7 conditions
    1. Unlinked normally distributed quantitative trait
    2. Normally distributed quantitative trait
    3. Highly skewed non-normal qualitative trait
    4. Normally distributed quantitative trait  with EDAC sampling
    5. Binary trait with 20% prevalence
    6. Bivariate normally distributed quantitative trait
    7. Bivariate skewed quantitative trait

# Results: Normally distributed quantitative trait

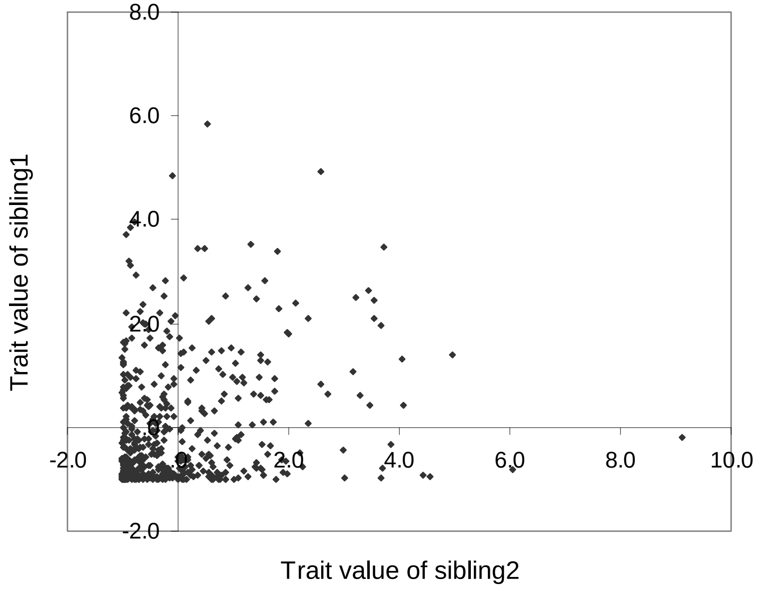


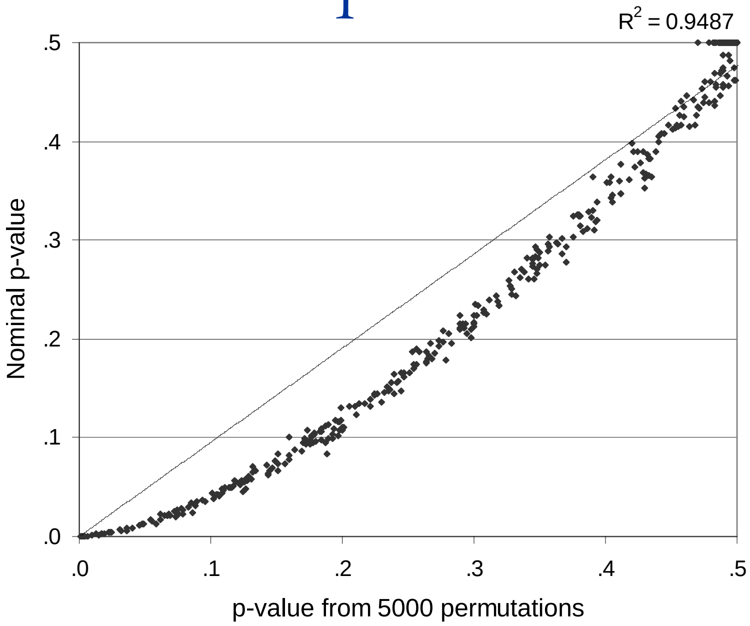


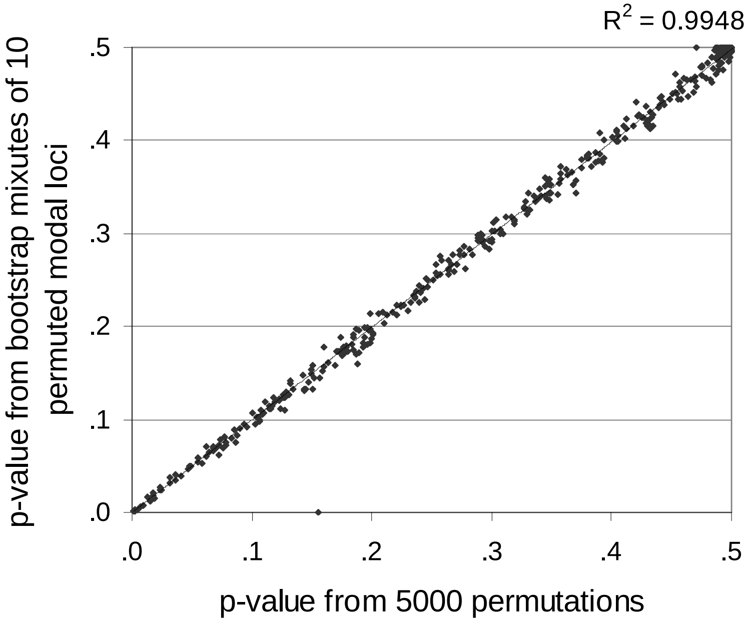


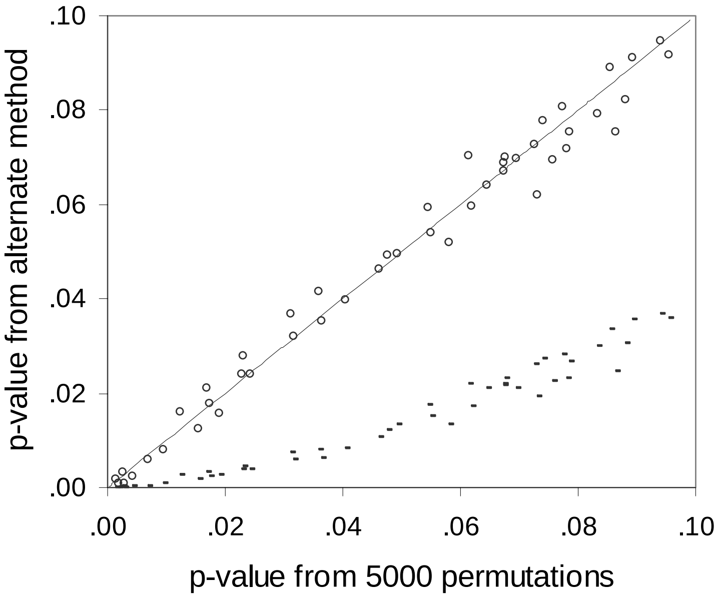


# Results: Highly skewed non-normal qualitative trait

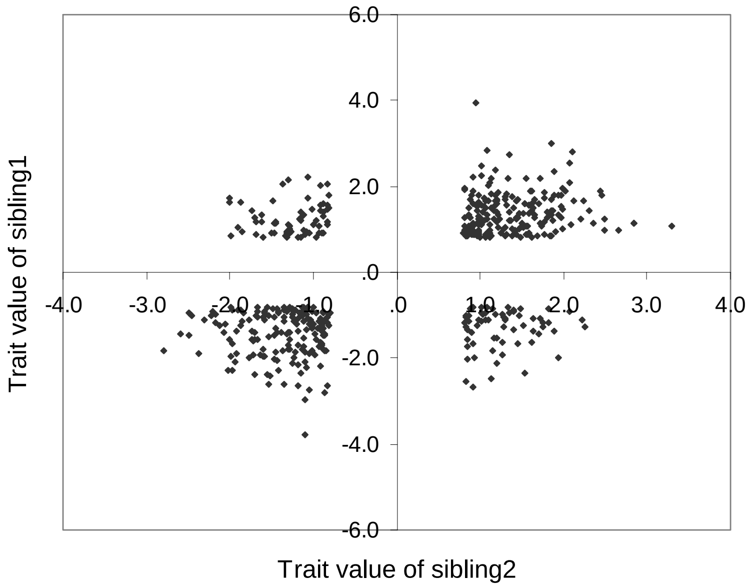


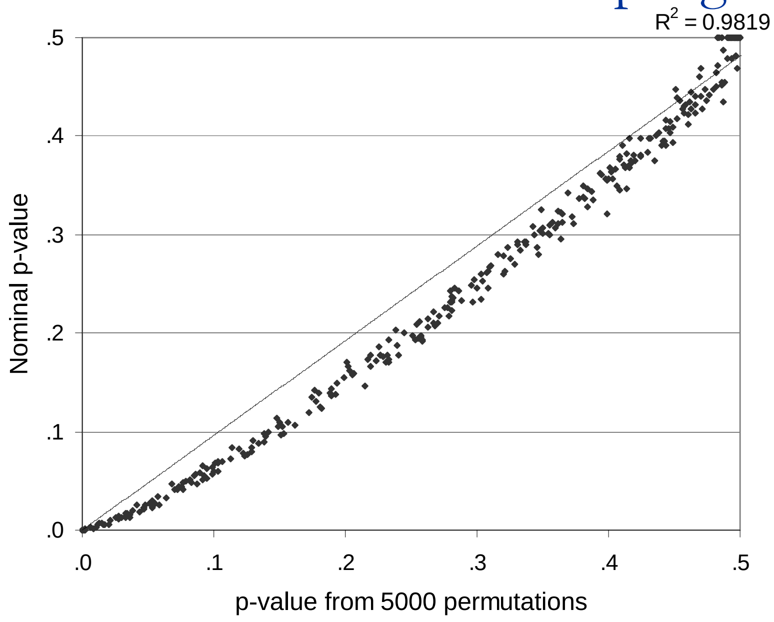


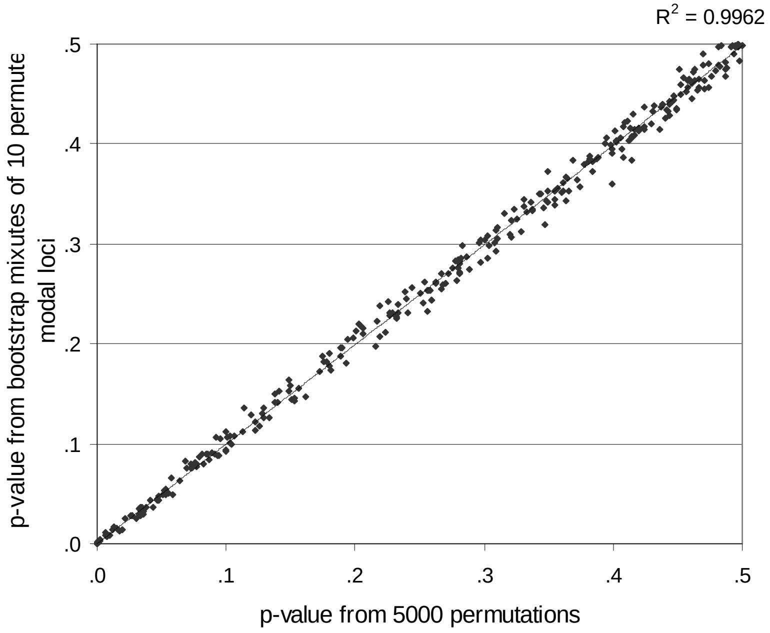


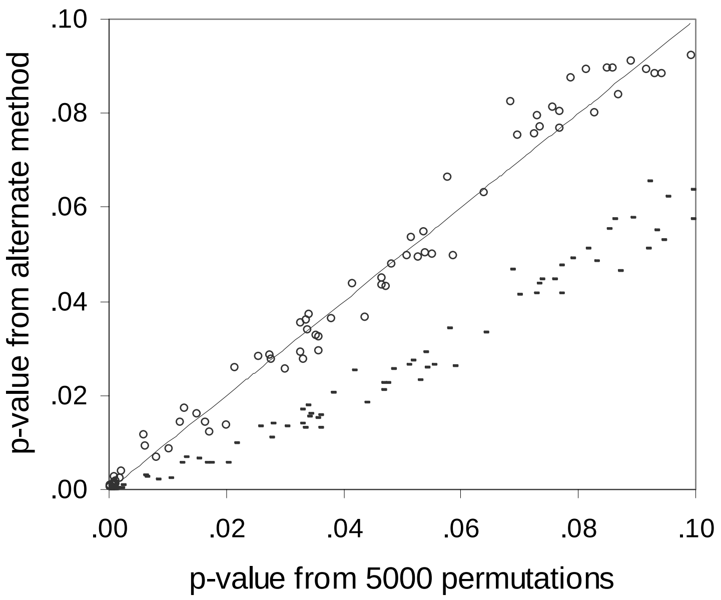


# Results: Normally distributed quantitative trait EDAC sampling

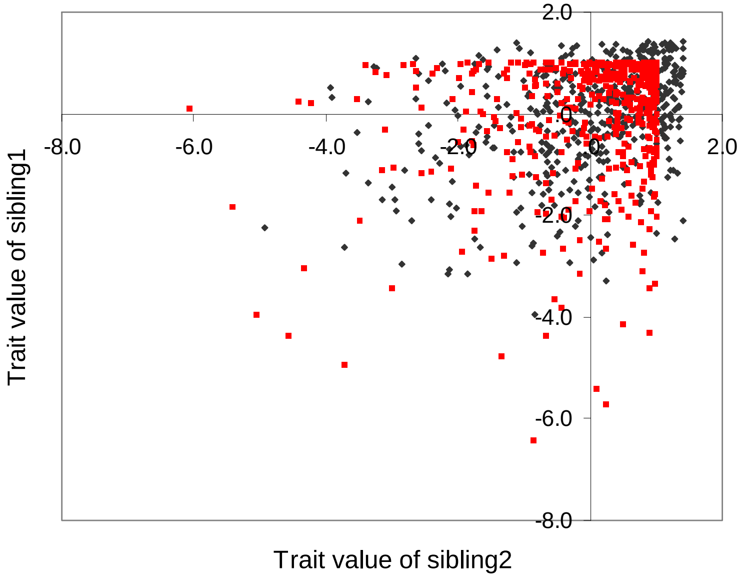


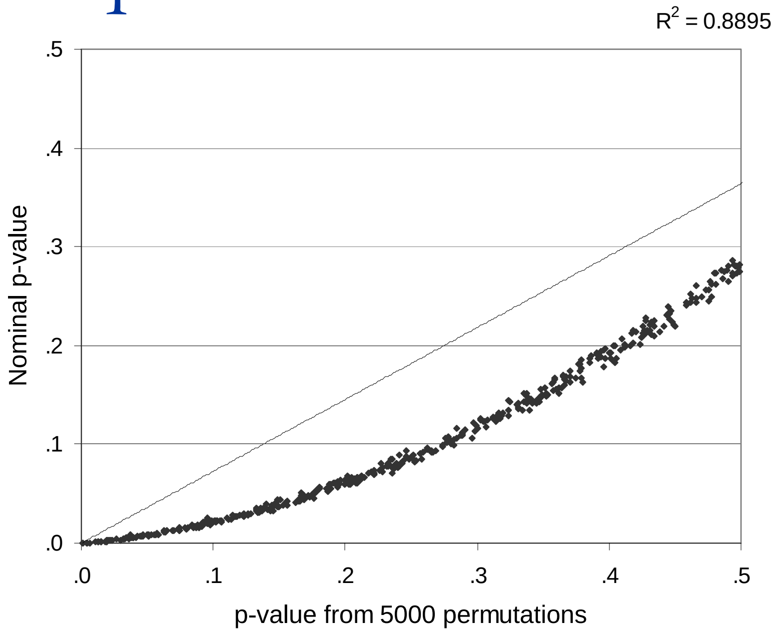


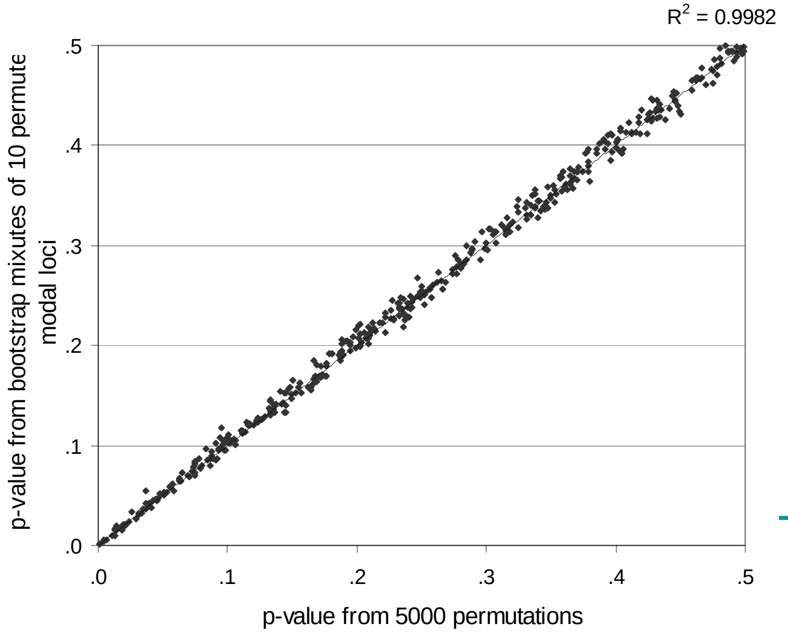


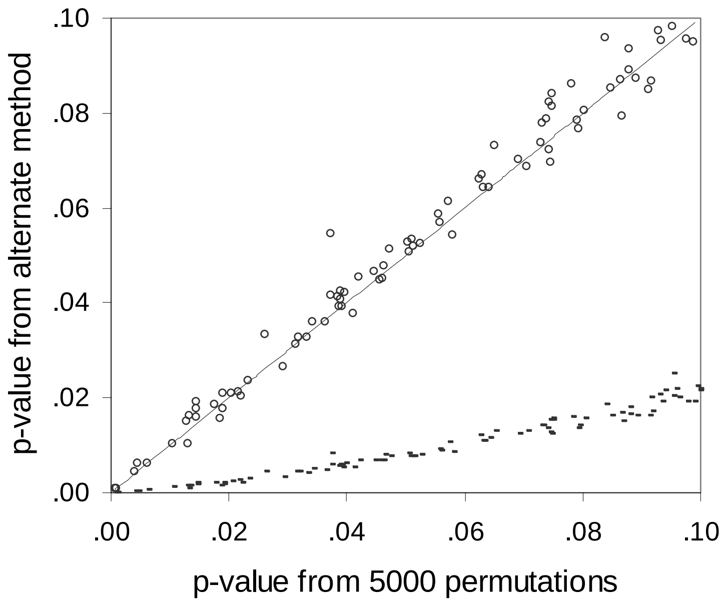


# Results: Bivariate skewed quantitative trait









# Results

- Weighted mixtures gave good approximations of empirical p-values

| | Mean absolute deviation* | Variance |
|---|---|---|
| Univariates | | |
| Normal | 0.0112 | 1.119E-04 |
| Skew | 0.0075 | 3.586E-05 |
| EDAC | 0.0079 | 3.863E-05 |
| Binary | 0.0081 | 4.182E-05 |
| Bivariates | | |
| Normal | 0.0059 | 2.217E-05 |
| Skew | 0.0065 | 8.317E-05 |

*[permutation p-value – weighted mixture p-value]

# Conclusions

- The proposed method produces close approximations of traditional empirical pvalues
- Appears robust to phenotypic distribution problems and suitable for multivariate analyses

# Conclusions

- Advantages
  - Requires fewer analyses than other efficient methods
    - For a genome wide linkage scan at 3000 locations
      - Traditional permutation/gene dropping (5000 replicates): 15,000,000 analyses
      - Sequential stopping: Scan and phenotype specific
      - Replicate pool method (100 replicates): 300,000 analyses
      - Weighted mixture approach: 50,000 analyses

# Conclusions

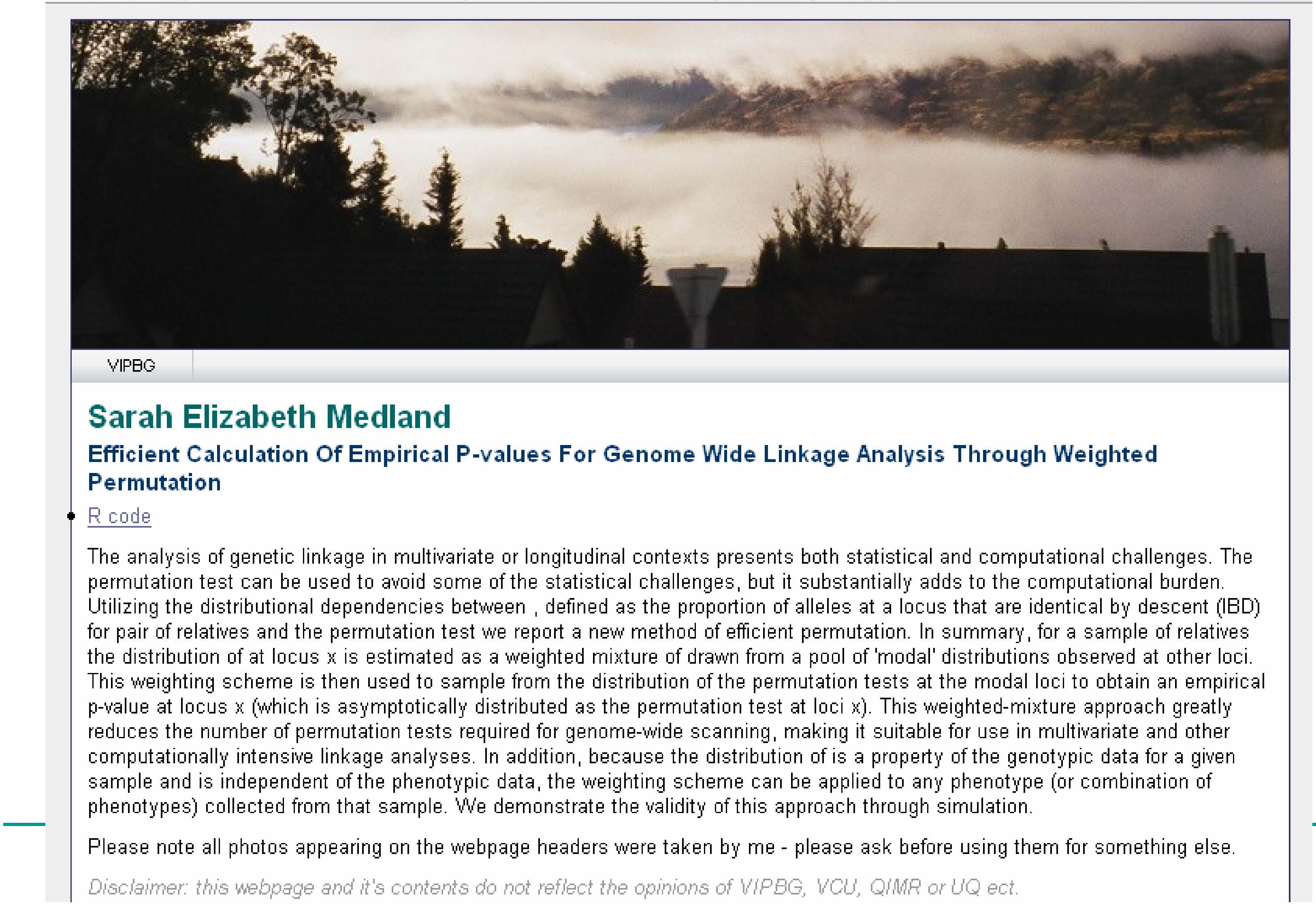
- Advantages
  - Modal weights are a property of the genotypic data & are transferable to any trait (or combinations of traits) analyzed using that genotypic data set.
    - Assuming MCAR/MAR missingness

# Conclusions

- Disadvantages
    - The variance of the weighted mixture p-values will vary across loci as a function of mixture weights
        - Suggestion: Use the weighted mixture method to obtain approximate p-values and also permute the peak markers
    - This method will be difficult to implement in situations where permutation test are difficult to implement
        - Complex arbitrary pedigrees & affected sib-pair studies

# http://www.vipbg.vcu.edu/~sarahme/permute.html

Medland, Schmitt, Webb, Kuo, Neale (submitted) Efficient calculation of empirical p-values for genome wide linkage analysis through weighted permutation



## Sarah Elizabeth Medland

### Efficient Calculation Of Empirical P-values For Genome Wide Linkage Analysis Through Weighted Permutation

- R code

The analysis of genetic linkage in multivariate or longitudinal contexts presents both statistical and computational challenges. The permutation test can be used to avoid some of the statistical challenges, but it substantially adds to the computational burden. Utilizing the distributional dependencies between , defined as the proportion of alleles at a locus that are identical by descent (IBD) for pair of relatives and the permutation test we report a new method of efficient permutation. In summary, for a sample of relatives the distribution of at locus x is estimated as a weighted mixture of drawn from a pool of 'modal' distributions observed at other loci. This weighting scheme is then used to sample from the distribution of the permutation tests at the modal loci to obtain an empirical p-value at locus x (which is asymptotically distributed as the permutation test at loci x). This weighted-mixture approach greatly reduces the number of permutation tests required for genome-wide scanning, making it suitable for use in multivariate and other computationally intensive linkage analyses. In addition, because the distribution of is a property of the genotypic data for a given sample and is independent of the phenotypic data, the weighting scheme can be applied to any phenotype (or combination of phenotypes) collected from that sample. We demonstrate the validity of this approach through simulation.

Please note all photos appearing on the webpage headers were taken by me - please ask before using them for something else.

*Disclaimer: this webpage and it's contents do not reflect the opinions of VIPBG, VCU, QIMR or UQ ect.*