Statistical and Computational Methods for GWAS

Dr. Hua Zhou

Epi 243: Molecular Epidemiology of Cancer

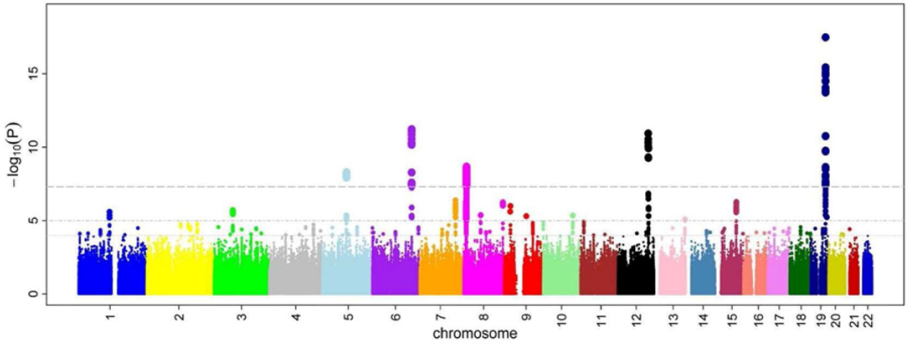
10am-11:50am, Feb 14, 2018

- Slides are available at `http://hua-zhou.github.io/teaching/epi243-2018winter/Epi243_2018fall_Zhou.pdf`
- To run the examples in this lecture, following steps:
  1. Download and install the Mendel software: `https://www.genetics.ucla.edu/software/mendel`
  2. Download and unzip the data for each example: Example 24a, Example 24b, Example 24c, Example 29a, Example 29b, Example 29c.

High-dimensional data

- It is an era of big data:
  Wall Street Journal, White House, McKinsey report, wiki, ...
- High throughput arrays for SNPs and next generation sequencing generate up to 10^6 predictors
- Screening this many predictors is problematic

Finding (a couple) needles in a huge haystack



Approaches to model selection

- Marginal test: check hay/needle one by one
- Continuous model selection by penalized regression (MCP or lasso): throw the whole haystack into water and pick up needles from bottom

Single marker (marginal) analysis

- Test each marker individually for association
  - Linear regression

$$y_i = \mu + \text{sex}_i \cdot \beta_1 + \text{age}_i \cdot \beta_2 + \text{SNP}_{ij} \cdot \beta_{\text{SNP}_j} + \varepsilon_i$$

  - Logistic regression

$$\text{logit}(E(y_i)) = \mu + \text{sex}_i \cdot \beta_1 + \text{age}_i \cdot \beta_2 + \text{SNP}_{ij} \cdot \beta_{\text{SNP}_j} + \varepsilon_i$$

- It is natural to incorporate covariates (age, gender, smoke, population substructure, ...) into the regression framework
- We obtain a p value for each marker

Challenges for marginal analysis: multiple testing

- In GWAS, we perform $500,000 \sim 1,000,000$ tests (many correlated)
- Using a significance level of $\alpha = 0.05$ for each test will give $25,000 \sim 50,000$ false positives!
- What to do?
  - Bonferroni correction to control the family-wise error rate (FWER)
  - Benjamini-Hochberg procedure to control the false discovery rate (FDR)

Bonferroni correction

Suppose we are testing m SNPs. In GWAS, $m = 10^5 \sim 10^6$.

	Null hypothesis is True (H_0)	Alternative hypothesis is True (H_1)	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- p_i the p-value of the i-th marker
- Bonferroni rule: reject all H_{0i}, i.e., declare association signal at SNP i, with $p_i \leq \frac{\alpha}{m}$
- Family-wise (Type I) Error Rate (FWER)
  $= \mathbf{P}$(make a type I error for at least one SNP)
  $= \mathbf{P}(V \geq 1) = \mathbf{P}(\cup_i E_i) \leq \sum_i \mathbf{P}(E_i) = m \cdot \alpha/m = \alpha$
- α/m usually on the order of 10^{-8}. That means a very stringent significance level for each SNP

False discovery rate (FDR)

	Null hypothesis is True (H_0)	Alternative hypothesis is True (H_1)	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- False discovery rate (FDR) = expected proportion of false positives among all declared significance = $\mathbf{E}[V/(V+S)]$
- Benjamini and Hochberg (1995) procedure
  - Let $p_{(1)} \leq \cdots \leq p_{(m)}$ be the sorted p-values
  - BH Rule: Reject all H_{0i} for $p_i \leq T$ with

$$T = \max\left\{p_{(i)} : p_{(i)} \leq \frac{i\alpha}{m}\right\}$$

- Theorem: Following BH procedure, then FDR $\leq \alpha$
- FDR is less conservative than Bonferroni correction

Fun fact: Benjamini and Hochberg (1995) is probably the most cited statistical paper (per year)

Paper	Citations	Per Year
Kaplan-Meier (Kaplan and Meier, 1958)	46886	808
EM (Dempster et al., 1977)	44050	1129
Cox model (Cox, 1972)	40920	930
Metropolis (Metropolis et al., 1953)	31284	497
FDR (Benjamini and Hochberg, 1995)	30975	1450
Unit root test (Dickey and Fuller, 1979)	18259	493
Lasso (Tibshirani, 1996)	15306	765
bootstrap (Efron, 1979)	12992	351
FFT (Cooley and Tukey, 1965)	11319	222
Gibbs sampler (Gelfand and Smith, 1990)	6531	251

Citation counts from Google Scholar on Feb 17, 2016.

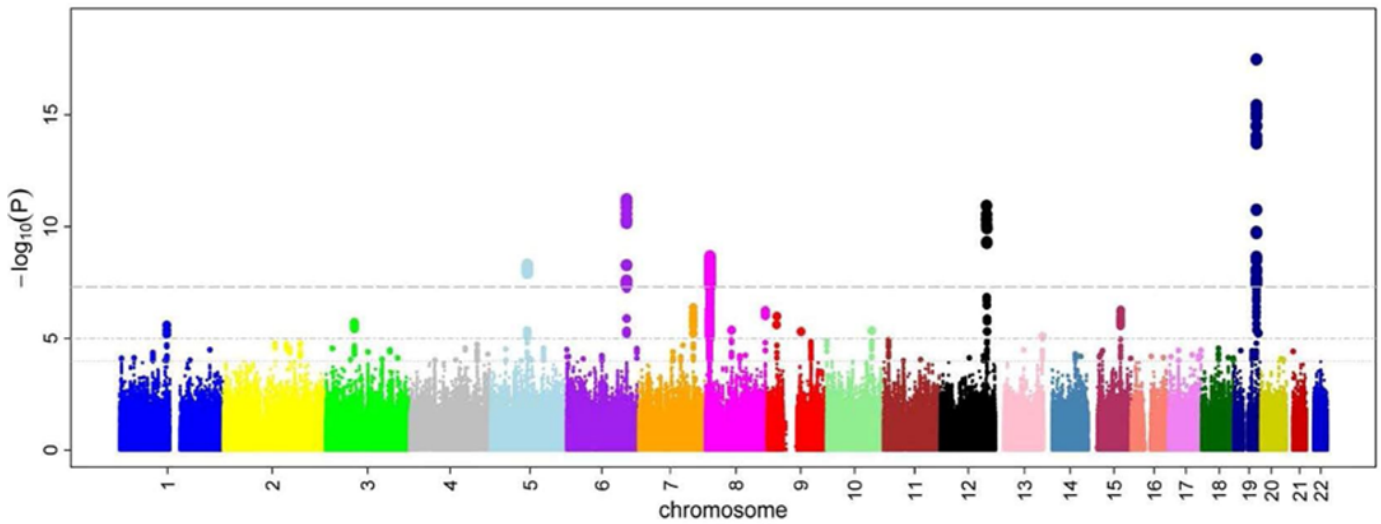
Issues with marginal tests

- Both Bonferroni correction and BH procedure select SNPs with top p values
- This might *not* be the best thing to do
- Suppose you are taking a test on Epi 243 with 100 topics and unfortunately you know none of them. But you are allowed to bring a few consultants. Your friend Ken knows 85 topics, Eric knows 75 topics, ... If two consultants are allowed, who do you choose?
  - marginal test: choose the two most knowledgeable friends Ken and Eric (but what if what Eric knows is already known by Ken?)
  - stage-wise selection: choose Ken first and then next most knowledgeable friend *in what Ken doesn't know*
  - best subset selection: go over all possible pairs to find the most knowledgeable combination. There might be two friends who only know 50% of the topics each but they perfectly complement each other!

Similarly, to prioritize SNPs for diagnosis, disease subtyping, predicting prognosis, explaining trait variation, ..., we can

- select top SNPs by ranking their p values, or
- select top SNPs by best subset regression or its heuristics.

The second approach is likely to select a group of SNPs that complement each other in prediction or explanatory power.



Continuous model selection

- Best subset selection is too computationally expensive. In GWAS with 10^6 SNPs, examine all k combinations requires fitting $\binom{10^6}{k}$ regressions!
- A recent trend: heuristic best subset selection by penalized regression (lasso and variants)
- Continuous model selection method puts all potential predictors together in a regression framework
  - Quantitative trait: linear regression
  - Case-control study: logistic regression
- It provides a unified framework to consider genetic variants, environmental factors, interactions and so on *simultaneously*

Linear regression (for quantitative traits)

- Multiple linear regression solves the least squares problem

$$\min_{\mu,\beta} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

  where y_i is the response for case i, x_{ij} is the value of predictor j for case i, β_j is the regression coefficient corresponding to predictor j, and μ is the intercept
- We assume most of β_j are 0 and want to find those non-zero ones
- Best subset selection is computationally prohibitive
- When sample size n is smaller than the number of predictors p, the least square problem is underdetermined. (Measure p things by taking just n observations.)

Lasso penalized regression

- Tibshirani (1996) and Donoho and Johnstone (1994) introduced the lasso penalty to regularize estimation in underdetermined regression problems
- In lasso penalized regression, we minimize the modified sum of squares

$$f(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Purpose of penalized regression

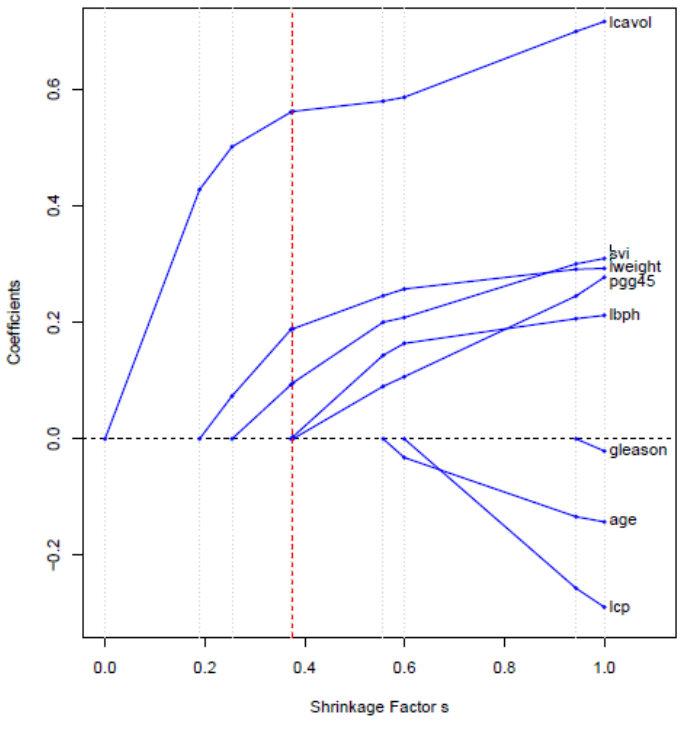
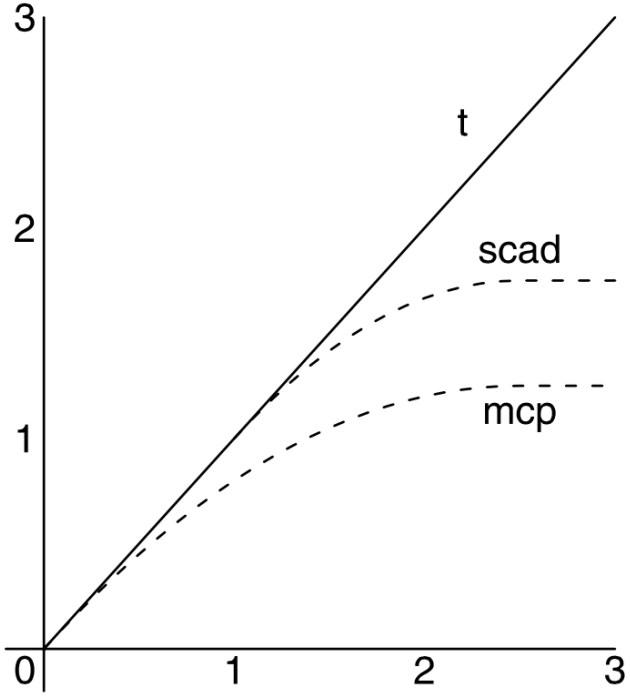


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$. *A vertical line is drawn at* $s = 0.36$, *the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso*

1. The lasso penalty $\lambda \sum_j |\beta_j|$ shrinks each β_j toward the origin and tends to discourage models with large number of irrelevant predictors. Thus, the procedure performs continuous model selection
2. The positive tuning constant λ is chosen by cross-validation, or tweaked to capture a fixed number of predictors. Smaller values of λ admit more predictors

MCP penalty



1. The Lasso penalty may over-shrink large regression coefficients and incur false positives
2. The MCP penalty (see left plot) puts lesser penalty on large regression coefficients and reduce bias in the penalized estimates
3. Since version 14.4, Mendel uses MCP penalty by default. To force the use of the lasso penalty, use the keyword `LASSO_PENALTY = True` in the control file

Algorithm for fitting lasso – coordinate descent

- Idea of coordinate descent: update parameters β_j one by one
- Fu (1998) and Daubechies et al. (2004) suggest coordinate descent for lasso penalized ℓ_2 regression. For inexplicable reasons, they did not follow up their theory with numerical confirmation for highly undetermined problems
- Friedman et al. (2007) and Wu and Lange (2008) show that the coordinate descent regression is incredibly quick for both ℓ_1 and ℓ_2 regression
- One can find 5 significant predictors out of 50,000 potential predictors for 200 cases in tenths of a second on a desktop computer

Logistic regression (for binary traits)

The loglikelihood in logistic regression is

$$\begin{aligned} L(\theta) &= \sum_i [y_i \ln \pi_i(\beta) + (1 - y_i) \ln[1 - \pi_i(\beta)] \\ \pi_i(\beta) &= \frac{e^{x_i^t \beta}}{1 + e^{x_i^t \beta}}. \end{aligned}$$

Straightforward calculations show

$$\begin{aligned} \nabla f(\beta) &= \sum_i [y_i - \pi_i(\beta)] x_i \\ d^2 f(\beta) &= -\sum_i \pi_i(\beta)[1 - \pi_i(\beta)] x_i x_i^t. \end{aligned}$$

The loglikelihood $L(\beta)$ is therefore concave.

Coordinate ascent in lasso penalized logistic regression

$$\max_{\theta} L(\theta) - \lambda \sum_{j=1}^{p} |\beta_j|$$

1. In coordinate ascent, we update one parameter at a time by Newton's method. For any parameter β_j, the penalized loglikelihood is differentiable on the half-intervals $(-\infty, 0]$ and $[0, \infty)$. In fact, most parameters never budge from 0.
2. Cyclic coordinate ascent continues until the objective function changes little and parameter estimates are stable.
3. Mendel adjusts λ to give a fixed number of predictors by a strategy of bracketing and bisection. Thus, we search on both λ and the regression coefficients.

Extension to sequence data (rare variants)

Dark matter in the missing heritability

Table 1. Estimates of heritability and number of loci for several complex traits.

▲ Figures & Tables index

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

*Residual is after adjustment for age, gender, diabetes.

- Structure variation: copy number variations (CNVs, insertions and deletions), copy neutral variations (inversions and translocations)
- Gene-by-gene and gene-by-environment interactions
- Epigenetic effects
- Rare variants

Strategies for dealing with rare variants

- Increase sample size ($$$)
- Grouping: define biologically meaningful groups such as genes, pathways, ... and do marginal group test or group-wise penalized regression
- Bring in biological information

Incorporate group information in continuous model selection

Group lasso (Yuan and Lin, 2006) performs continuous model selection at the group level.

- For linear regression, we minimize

$$f(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu - x_i^t\beta)^2 + \lambda\sum_{G}\|\beta_G\|_2,$$

  where β_G is the subvector of coefficients in group G and $\|\beta_G\|_2 = \sqrt{\beta_{G1}^2 + \cdots + \beta_{Gk}^2}$

- For logistic regression, we maximize

$$f(\theta) = L(\theta) - \lambda\sum_{G}\|\beta_G\|_2.$$

- Still no selection within the groups ...

Group+Lasso Penalized Regression

- For quantitative traits, we minimize

$$f(\theta) = \frac{1}{2}\sum_{i=1}^{q}(y_i - \mu - x_i^t\beta)^2 + \lambda_L\sum_{j=1}^{p}|\beta_j| + \lambda_E\sum_{G}\|\beta_G\|_2,$$

  where β_G is the subvector of coefficients in group G and $\|\beta_G\|_2 = \sqrt{\beta_{G1}^2 + \cdots + \beta_{Gk}^2}$.

- For case/control studies, we maximize

$$f(\theta) = L(\theta) - \lambda_L\sum_{j=1}^{p}|\beta_j| - \lambda_E\sum_{G}\|\beta_G\|_2.$$

- Little change in the algorithm: coordinate descent/ascent.

An illustrative simulation study

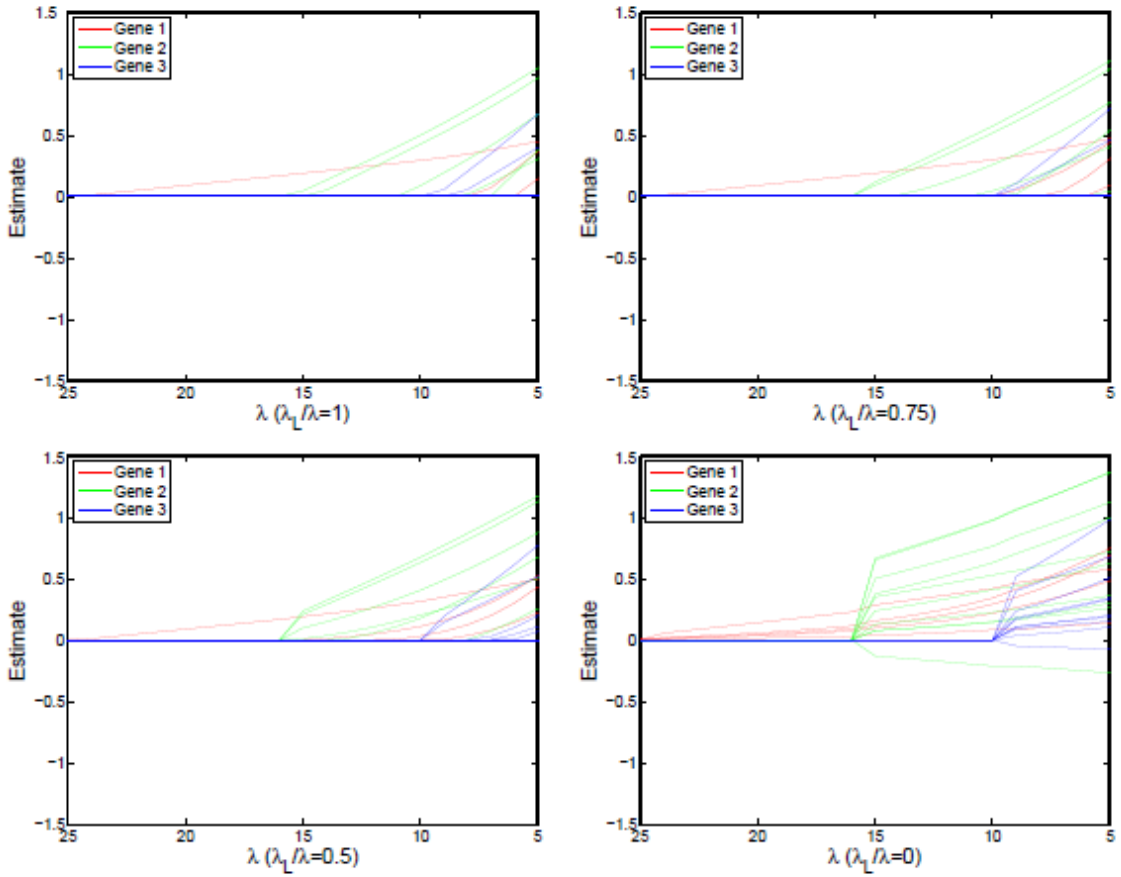


Fig. 1. A simulation example with 500 cases and 500 controls. There are three genes. Gene 1 (red) contains one common causal variant (MAF 10% and RR 1.2) and four neutral rare variants. Gene 2 (green) contains five causal rare variants (MAF 1% and RR 5) and five neutral rare variants. Gene 3 (blue) contains ten neutral rare variants. All neutral rare variants have MAF 1% and RR 1. The wild-type penetrance f_0 is set at 0.01. The pure lasso penalty ($\lambda_L/\lambda = 1$) picks up significant variants (common and rare) sequentially. The pure group penalty ($\lambda_L/\lambda = 0$) picks up the genes (groups) 1, 2, and 3 sequentially. The mixed group plus lasso penalty ($\lambda_L/\lambda = 0.75$ or 0.50) achieves a good compromise between the two.

Incorporate biological information in continuous model selection

Weighted Group+Lasso Penalized Regression

- For quantitative traits, we minimize

$$f(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu - x_i^t\beta)^2 + \lambda_L\sum_{j=1}^{p} w_j|\beta_j| + \lambda_E\sum_{G} w_G\|\beta_G\|_2$$

- For case/control studies, we maximize

$$f(\theta) = L(\theta) - \lambda_L\sum_{j=1}^{p} w_j|\beta_j| - \lambda_E\sum_{G} w_G\|\beta_G\|_2$$

- w_j reflects available biological information of the variant (such as scores from SIFT prediction, linkage trace, etc)
- Larger w_j discourages selection of j-th variant
- Little change in the algorithm: coordinate descent/ascent (fast)

References

- General reference for Mendel:
  Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013) Mendel: The Swiss army knife of genetic analysis programs, *Bioinformatics* 29(12):1568-1570.
- Lasso for GWAS
  Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009). Genomewide association analysis by lasso penalized logistic regression. *Bioinformatics* 25:714-721.
- Algorithm for lasso:
  Wu TT, Lange K (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* 2:224-244.
- Group lasso for GWAS:
  Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26:2375-2382.
- Weighted (group) lasso for GWAS:
  Zhou H, Alexander D, Sehl M, Sinsheimer J, Sobel E, Lange K (2011) Penalized regression for genome-wide association screening of sequence data. *Pacific Symposium on Biocomputing 2011* 106-117.

Mendel Software Example 24a, 24b, 24c)

A GWAS data set

- 2200 individiuals
- 10,000 SNPs
- Trait values are simulated from loci
  - `rs2256412`
  - `rs1935681`
  - their interaction

Exercises:

- **Example 24a**: quantitative trait (linear regression), only marginal analysis, no interaction analysis
- **Example 24b**: quantitative trait (linear regression), base to all interaction analysis
- **Example 24c**: binary trait (logistic regression), penalized regression using MCP, interaction analysis

Running penalized regression in Mendel

Some keywords for penalized regression in Mendel Option 24

- `PENALIZED_REGRESSION = True` turns on penalized regression
- `PENALIZED_INTERACTION = True` turns on interaction analysis using penalized regression
- `DESIRED_PREDICTORS = 10 :: PENALIZED` sets the number desired SNPs from penalized regression
- `PREDICTOR_PENALTY_PROPORTION = 0.9` tips the balance between individual predictor penalty λ_L and group penalty λ_G

$$\lambda_L = \lambda p, \quad \lambda_E = \lambda(1-p)$$

`PREDICTOR_PENALTY_PROPORTION = 1`: purely variant analysis
`PREDICTOR_PENALTY_PROPORTION = 0`: purely group analysis

- `UNIFORM_WEIGHTS = true` uses weights $w_j \equiv 1$
  `UNIFORM_WEIGHTS = false` uses weights $w_j = \sqrt{4p(1-p)}$ (less penalty on rare variants)

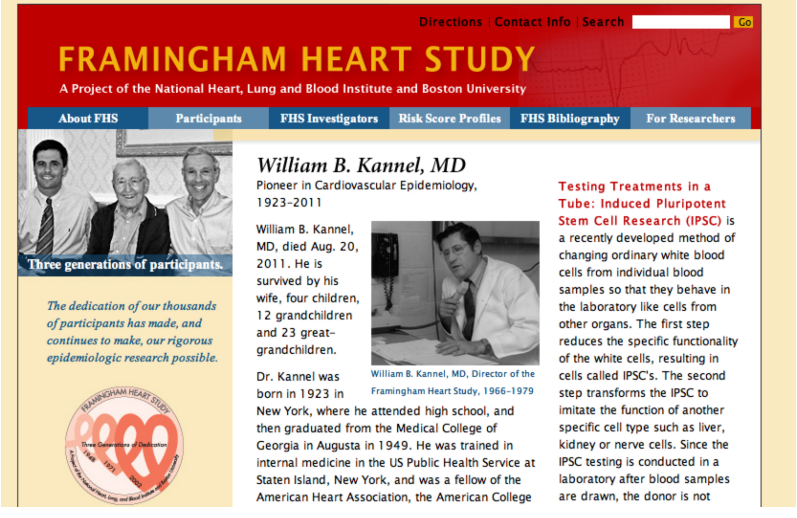
GWAS analysis based on pedigree data

- Background
- Statistical model for QTL analysis
- Mendel implementation
- Using Option 29 for Pedigree-GWAS
- Using Option 29 for adjusting ethnic admixture (cryptic relatedness) in population GWAS or GWAS with lost/suspicious pedigrees

When do I use Mendel Option 29?

Appropriate Problems and Data Sets

- Quantitative (multivariate) trait(s): HDL, weight, height, ...
- Genetic Data: GWAS-scale SNP data
- Study individuals can be pedigrees, unrelateds, or mixture of pedigrees/unrelateds
- Example: Framingham Heart Study, San Antonio Heart Study, large-scale population GWAS, ...



Implementation in Mendel Option 29

- Deal with multivariate traits with possible missingness
- Robust to outliers (use t-distribution)
- Fast (critical for genome-wide QTL analysis)
- Run a sequence of univariate trait analysis on the same genotype data (useful for large scale eQTL analysis with $10^3 \sim 10^5$ expression traits)

Other tools in Mendel for pedigree analysis

- Option 17: Gene dropping: simulate pedigree genotypes from founders
- Option 28: Trait simulation from generalized linear models (linear, logistic, Poisson) and variance component model (pedigrees)
- ...

Side-by-side comparison with Fast-LMM and GEMMA

	Mendel	FastLMM	GEMMA
Multi-threaded operation	Yes	Yes	No
Can estimate kinships via SNPs	Yes	Yes	Yes
Imports & exports kinship estimates	Yes	Yes	Yes
Allows retained co-variates	Yes	Yes	Yes
Allows linear constraints on co-variates	Yes	No	No
Can use either LRT or score test	Yes	No	Yes*
Allows multivariate trait	Yes	No	Yes
Allows missingness in multivariate traits	Yes	N/A	No
Can perform multiple univariate analyses	Yes	No	No
Allows > 2 variance components	Yes	No	No
Analyzes X-linked loci	Yes	No	No
Automatic SNP filtering on MAF	Yes	No	Yes
Allows non-additive SNP models	Yes	No	No
Detects outlier pedigrees	Yes	No	No
Detects outlier individuals	Yes	No	No
Can simulate genotype/phenotype data	Yes	No	No
Reads in fractional genotype values	No	Yes	Yes

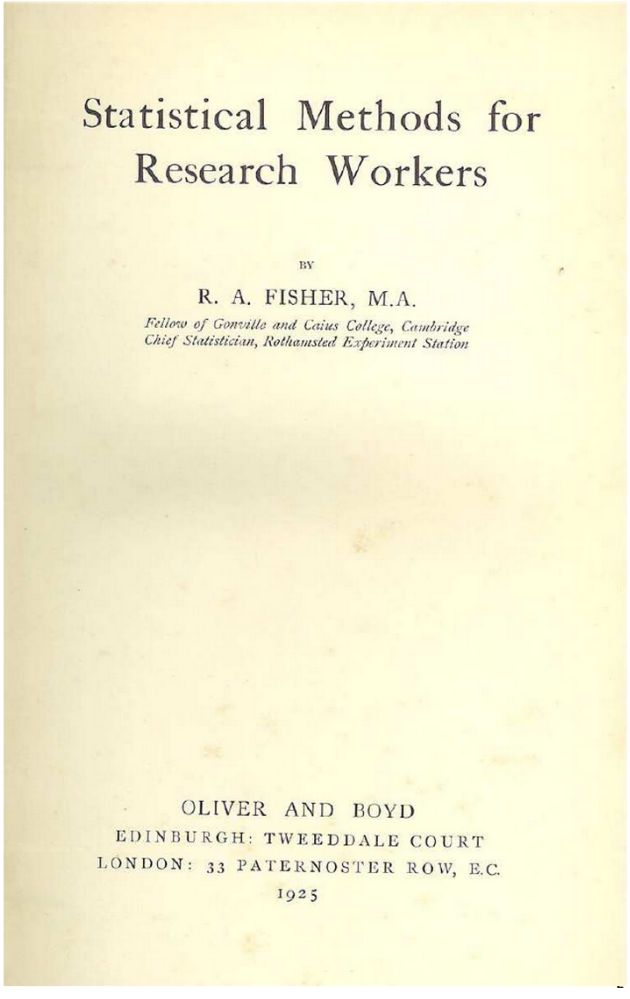
*GEMMA can use the likelihood ratio, score, or Wald test.

Efficiency of Mendel Option 29

San Antonio Family Heart Study (SAFHS)

Program	Trait	Analyzed Samples	Analyzed SNPs	RunTime (min:sec)	RAM (GB)
MENDEL default		1357	935,392	1:51	1.2
MENDEL all-pairs	HDL_1	1357	935,392	7:49	1.2
FAST-LMM		1397	941,546	76:11	30.0
GEMMA		1397	919,050	206:54	0.4
MENDEL default		818	935,392	1:33	1.1
MENDEL all-pairs	HDL_2	818	935,392	3:25	1.1
FAST-LMM		840	934,216	49:44	18.0
GEMMA		840	914,051	180:21	0.3
MENDEL default		914	935,392	1:38	1.1
MENDEL all-pairs	HDL_3	914	935,392	3:54	1.1
FAST-LMM		939	937,208	54:58	20.0
GEMMA		939	918,626	182:26	0.3
MENDEL default	HDL_{Joint}	1388	935,392	4:08	1.2
MENDEL all-pairs	with	1388	935,392	83:24	1.2
FAST-LMM	constrained			Not Available	
GEMMA	covariates			Not Available	
MENDEL default	HDL_{Joint}	1388	935,392	3:49	1.2
MENDEL all-pairs	without	1388	935,392	80:04	1.2
FAST-LMM	constrained			Not Available	
GEMMA	covariates	712	912,318	630:37	0.6

Variance Component Model for QTL Analysis


Statistical Methods for Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

- First explicit use in genetics dates back at least to R.A. Fisher (1925)
- Modern forms: linear mixed model (LMM), similarity regression, least squares kernel machines, ...
- Pedigree-GWAS poses serious computational challenge

Variance Component Model for QTL Analysis

- $Y = (Y_1, \ldots, Y_n)^t$ denotes the trait of n individuals and is modeled as a multivariate normal distribution

$$Y \sim \text{Normal}(\nu, \Omega)$$

  Mendel allows Y_i to be multivariate traits (pleiotropy)
- Mean: $\nu = A\beta$, where $A \in \mathbb{R}^{n \times p}$ is the design matrix for p covariates (age, sex, smoke, PCs, ...)
- Variance components: $\Omega \in \mathbb{R}^{n \times n}$ is modeled as

$$\Omega = 2\sigma_a^2 \Phi + \sigma_d^2 \Delta_7 + \sigma_h^2 H + \sigma_e^2 I$$

  with variance components
  - Φ: additive genetic effects. Mendel can use either
    - theoretical kinship: calculated from input pedigree information, or
    - empirical kinship: estimated from dense SNP data
  - Δ_7: dominance genetic effects
  - H: household effects $h_{ij} = 1\{i \text{ and } j \text{ are in the same household}\}$
  - I: random (environmental) errors

MLE for parameter estimation

- Given observed trait vector $Y = y$, the log-likelihood is

$$L(\theta) = \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln \det\Omega - \frac{1}{2}(y - A\beta)^t \Omega^{-1}(y - A\beta),$$

  where $\theta = (\beta_1, \ldots, \beta_p, \sigma_1^2, \ldots, \sigma_r^2)$.
- We estimate mean fixed effects $\beta_1, \ldots, \beta_p$ and variance components $\sigma_1^2, \ldots, \sigma_r^2$ by maximum likelihood estimation (MLE)
- Algorithm: Fisher's scoring method (roughly scales with the cube of the largest pedigree size)

Testing a SNP

How to test a genotyped SNP for association?

- Treat the SNP as an extra fixed effect and append genotypes to the design matrix A
- Estimate its corresponding fixed effect β_{p+1}
- $\beta_{p+1} \neq 0$ implies potential association of that SNP
- Formal statistical test is needed here

Triad of hypothesis testing

- Want to test H_0: $\beta_{p+1} = 0$ (SNP has no effect) vs H_a: $\beta_{p+1} \neq 0$ (SNP has effect)
- Three standard asymptotic tests to use
  - Likelihood ratio test (LRT)
  - Score test
  - Wald's test
- Which one to use?

Triad of hypothesis testing

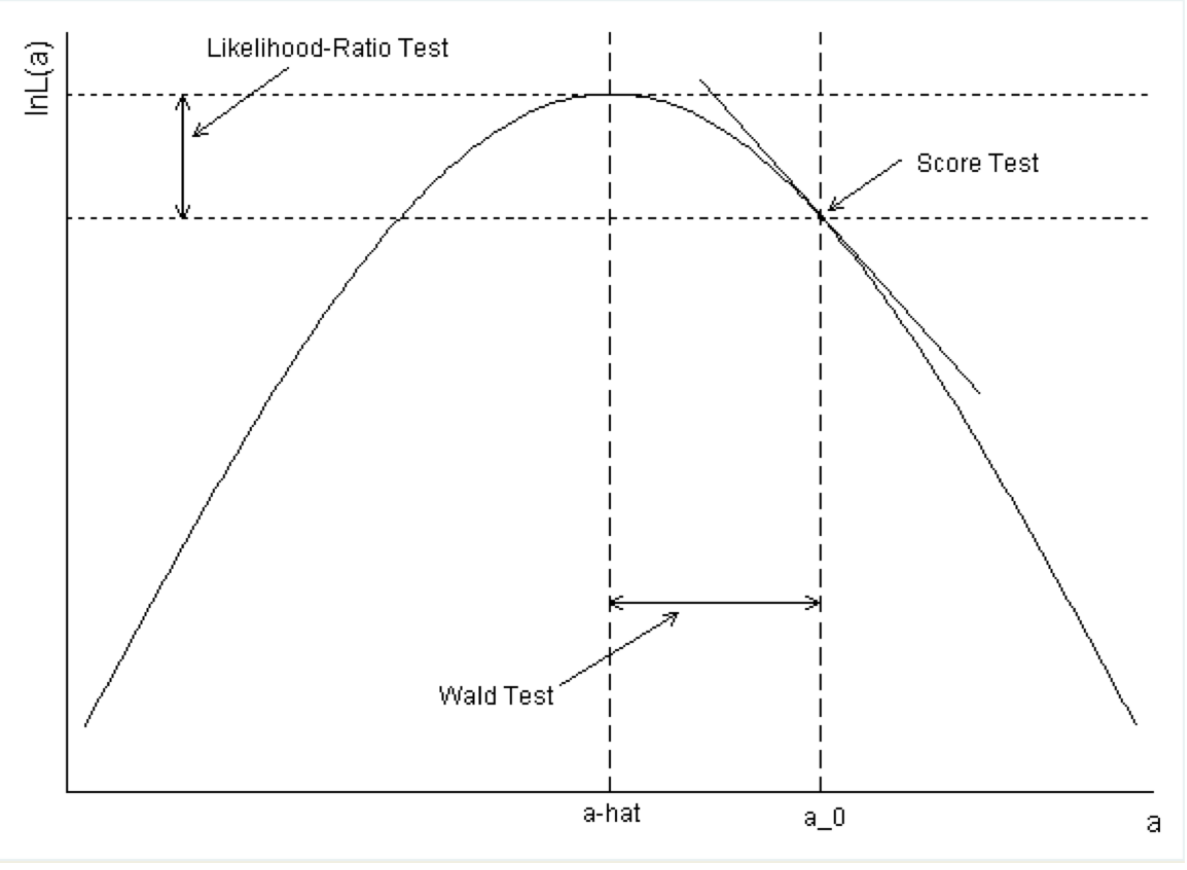
- Wald's test
  - Test statistic $z = \hat{\beta}_{p+1}/\text{SE}$
  - With large sample size n, z^2 is approximately χ^2_1
- LRT test
  - Find MLE under both null and alternative models
  - Compare the two likelihoods $\Lambda = \ell_0/\ell_a$
  - With large sample size n, $-2\ln\Lambda$ is approximately χ^2_1
- Score test
  - Test statistic

$$S = \left. \frac{[\partial L(\beta_{p+1})/\partial \beta_{p+1}]^2}{-E[\partial^2 L(\beta_{p+1})/\partial \beta^2_{p+1}]} \right|_{\beta_{p+1}=0}$$

  - With large sample size n, S is approximately χ^2_1

LRT-Wald-Score Triad

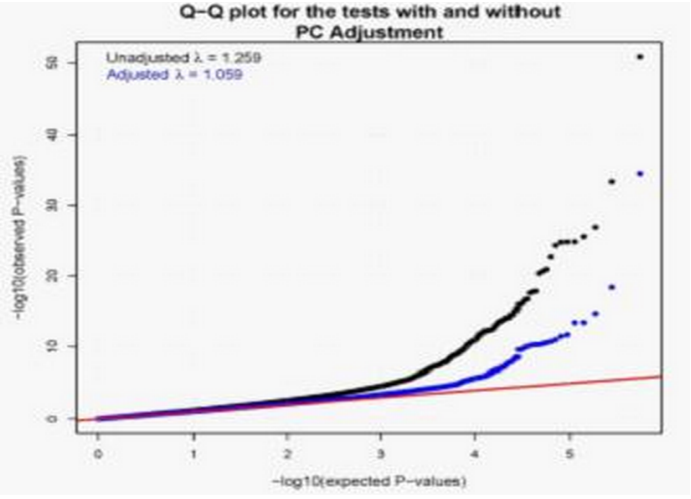
H_0: $\beta_{p+1} = 0$ (SNP has no effect) vs H_a: $\beta_{p+1} \neq 0$ (SNP has effect)



Triad of hypothesis testing

- Facts
  - LRT/score/Wald are asymptotically equivalent – with very large samples, they have same type I error and power
  - At small to medium sample sizes, LRT is generally more powerful than score/Wald
  - For LRT and Wald, MLE has to be performed one SNP at a time. That means $10^5 \sim 10^6$ optimizations for GWAS!
  - Forming the score $u(\beta) = \partial L(\beta)/\partial\beta$ for each SNP is cheap compared to fitting the corresponding alternative model
- Mendel Option 20 implements LRT
- Strategy in Mendel Option 29
  - Score test for screening (ranking SNPs by score p-values).
  - LRT to boost power
    Mendel Option 29 performs LRT only for top SNPs (set by user)

Using Mendel Option 29 for Population GWAS



- Hidden relatedness in large scale population GWAS causes many false positives
- Adjustment by PCs, ancestry estimates, ... is a must for publishing GWAS in high-profile journals
- Variance component model (linear mixed model) is another natural (and powerful) way to handle cryptic relatedness
- How? Estimate the "kinship coefficient matrix" from dense SNP data and fit a pedigree GWAS using the pseudo-pedigrees

Specifying the additive genetic component Φ

The keyword `KINSHIP_SOURCE` specifies the form of the additive genetic component Φ. It can take following values

- `KINSHIP_SOURCE = pedigree_structure`: use theoretical kinship calculated from the pedigrees listed in the input files
- `KINSHIP_SOURCE = SNPs_within_pedigrees` (default): estimate kinship between pairs of individuals within each pedigree listed in the input files
- `KINSHIP_SOURCE = SNPs_using_everyone`: estimate kinship between all pairs of individuals, ignoring pedigree structures in input files

Specifying the additive genetic component Φ

The keyword `KINSHIP_METHOD` specifies the method for estimating Φ for SNPs. It can take following values

- `KINSHIP_METHOD = GRM` (default): genetic relationship matrix method

$$\hat{\phi}_{ij} = \frac{1}{2S} \sum_{k=1}^{S} \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1-p_k)}$$

- `KINSHIP_METHOD = Mom`: method of moment method

$$\begin{aligned} \hat{\phi}_{ij} &= \frac{e_{ij} - \sum_{k=1}^{S}[p_k^2 + (1-p_k)^2]}{S - \sum_{k=1}^{S}[p_k^2 + (1-p_k)^2]}, \text{ where} \\ e_{ij} &= \frac{1}{4} \sum_{k=1}^{S} [x_{ik}x_{jk} + (2-x_{ik})(2-x_{jk})] \end{aligned}$$

How many SNPs to use for estimating Φ

The keyword `SNP_SAMPLING_INCREMENT` controls the SNP sampling frequency for estimating the kinship Φ. For example,

- `SNP_SAMPLING_INCREMENT = 5` (default): use 20% of SNPs
- `SNP_SAMPLING_INCREMENT = 1`: use all SNPs

The keywords `MINIMUM_SNPS_SAMPLED` (default value 5000) safeguard against using too few SNPs.

Summary

- Mendel Option 29 (PedGWAS) is an extremely efficient implementation of variance component model (mixed model) for QTL association mapping using general pedigrees
- Can handle multivariate traits with missingness
- A nice way to adjust for "cryptic relatedness" in large population GWAS studies

References for Pedigree-GWAS

- General reference for Mendel:
  Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013) Mendel: The Swiss army knife of genetic analysis programs, *Bioinformatics* 29(12):1568-1570.
- Fast pedigree GWAS using variance component model:
  Zhou H, Blangero J, Dyer TD, Chan KH, Sobel EM, Lange K (2016) Fast genome-wide QTL association mapping on pedigree and population data, to appear in *Genetic Epidemiology*.
- Estimating global and local kinship coefficient from dense SNP data:
  Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM (2011) Linkage analysis without defined pedigrees, *Genetic Epidemiology* 35:360–270.

Using Mendel Option 29

For this problem, we need

- Control file: specifies input/output files, analysis option
- Pedigree file: text-based SNP data can be put here or use binary SNP file for large GWAS data
- Definition file: disease and marker allele frequencies
- Map file: specifies the loci to use and the distance between them (distances are not important in this option)
- SNP binary/definition files (optional, but might be necessary for GWAS data)

Mendel Software Example 29a, 29b, 29c

Data description:

- 85 founders taken from 1000 Genome Project (unrelated Europeans)
- 253,141 SNPs on Chromosome 19
- Pedigree structures taken from Framingham Study
- Non-founders' genotypes simulated by gene dropping (using Option 17)
- 212 individuals in 27 pedigrees (sizes $1 \sim 36$)
- `simTrait` simulated from locus `rs10412915` (using Option 28)
- Covariates: `sex`

Example 29a: Control file

Mendel Control File:

```
ANALYSIS_OPTION = ped-GWAS
QUANTITATIVE_TRAIT = simTrait
PREDICTOR = SEX :: simTrait
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT
SNP_SAMPLING_INCREMENT = 5
KINSHIP_SOURCE = SNPs_within_pedigrees
OUTLIERS = True
```

Example 29a: Results

What do we observe in the output files?

- How many SNPs were used to estimate kinship?
- Are all SNPs tested?
- Is the major locus `rs10412915` the top SNP?
- Are there any other SNPs that pass Bonferroni threshold?
- Why some SNPs have exactly same regression estimates and p-values?
- Are there any pedigrees/individuals flagged as outlier?

Example 29a: A tweak

- What if we comment out `COVARIANCE_CLASS = ADDITIVE`?

```
ANALYSIS_OPTION = ped-GWAS
QUANTITATIVE_TRAIT = simTrait
PREDICTOR = SEX :: simTrait
!COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT
SNP_SAMPLING_INCREMENT = 5
KINSHIP_SOURCE = SNPs_within_pedigrees
OUTLIERS = True
```

- How are results different?

Example 29b: Control file

Mendel Control File:

```
ANALYSIS_OPTION = ped-GWAS
QUANTITATIVE_TRAIT = simTrait
PREDICTOR = SEX :: simTrait
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT
SNP_SAMPLING_INCREMENT = 1
KINSHIP_SOURCE = SNPs_using_everyone
```

Example 29b: Results

What do we observe in the `Summary` file?

- Is the major locus `rs10412915` still the top SNP?
- How many SNPs pass Bonferroni threshold?

Example 29c: PedGWAS on individual traits

- Same SNP genotypes and pedigree structure as in 16d and 16e and
- Bivariate traits `simTrait1` and `simTrait2` are simulated from the same locus `rs10412915`

Example 29c: Control file

Mendel Control File:

```
ANALYSIS_OPTION = ped-GWAS
QUANTITATIVE_TRAIT = simTrait1
QUANTITATIVE_TRAIT = simTrait2
PREDICTOR = SEX :: simTrait1
PREDICTOR = SEX :: simTrait2
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT
```

Example 29c: Results

What do we observe in the `Summary` file?

- How are the results for trait `simTrait1` and `simTrait2` displayed?
- Is the major locus `rs10412915` still the top SNP for trait `simTrait1` and `simTrait2`?

Example 29c: A tweak

- How do I test the multivariate trait (`simTrait1`, `simTrait2`) simultaneously?
- Set keyword `MULTIVARIATE_ANALYSIS = True`

```
ANALYSIS_OPTION = ped-GWAS
QUANTITATIVE_TRAIT = simTrait1
QUANTITATIVE_TRAIT = simTrait2
PREDICTOR = SEX :: simTrait1
PREDICTOR = SEX :: simTrait2
COVARIANCE_CLASS = ADDITIVE
COVARIANCE_CLASS = ENVIRONMENTAL
DESIRED_PREDICTORS = 10 :: LRT

MULTIVARIATE_ANALYSIS = True
```

- How are results different from 08d, 08e?

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Ser. B*, 57:289–300.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Soceity Series B.*, 39(1-38).

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.*, 74(366, part 1):427–431.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, 2(1):224–244.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67.