An Information Theoretic Approach to Analysing GWAS Data

Daniel F. Schmidt and Enes Makalic

Centre for Molecular, Environmental, Genetic & Analytic (MEGA) Epidemiology
School of Population Health
University of Melbourne

Melbourne, June 19, 2009

Content

Problem

- Genome Wide Association Studies (GWAS)
- Determine if there exists an association between genes (genotype) and disease (phenotype)

Brief Biology Background

- Cells are the basic building blocks of organisms
- Deoxyribonucleic Acid (DNA)
  - Genes
  - Contains the genetic ‘code’
  - Inherited from parents
  - Governs the behaviour of cells
- A single-nucleotide polymorphism (SNP)
  - Genetic code at some position (locus) in the DNA

Statistical Model of GWAS

- Assumptions
  - We have ($m > 0$) SNPs and ($n > 0$) people
  - SNPs are independent
  - Grouping ; dominant w.r.t. to minor allele
- Notation
  - phenotypes $\mathbf{x} \in \{0,1\}^n$
  - genotypes $\mathbf{G} \in \{0,1\}^{(n \times m)}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & \dots & g_{1m} \\ g_{21} & g_{22} & \dots & g_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \dots & g_{nm} \end{bmatrix} \tag{1}$$

Standard Methodology, Part 1

- Test for association between the phenotype and each SNP independently
- Null hypothesis : no association between SNP and phenotype
  - Contingency table for SNP j and person $i = (1, \ldots, n)$

	$(g_{ij} = 0)$	$(g_{ij} = 1)$
$(x_i = 0)$	a	b
$(x_i = 1)$	c	d

  - Fisher's exact test or χ^2 test for independence
  - Significance level (commonly) $\alpha = 0.05$
  - m tests based on p-values $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$
- Combine p-values into a simultaneous test procedure

Standard Methodology, Part 2

- Control the Familywise Error Rate (FWER)
  - Probability of making at least one false discovery
  - Type I error rate : probability of incorrectly rejecting the 'null' hypothesis

$$\text{FWER} = P(V \geq 1) \tag{2}$$

  - Bonferonni correction

$$\alpha^* = \frac{\alpha}{m} \tag{3}$$

- Control the γ-FWER

$$P\left\{\text{FDP} > \gamma\right\} \leq \alpha \tag{4}$$

Standard Methodology, Part 3

- Control the False Discovery Rate (FDR)
  - False Discover Proportion (FDP)

$$\text{FDP} = \frac{\text{number of false rejections}}{\text{total number of rejections}} \tag{5}$$

  - Put a bound on FDR = E(FDP)
  - Sort p-values in ascending order
  - Let k be the largest i ($1 \leq i \leq m$) for which

$$\hat{p}_{(i)} \leq \left(\frac{i}{m}\right)\alpha \tag{6}$$

    - Reject hypotheses $H_{(i)}$, $(i = 1, \ldots, k)$

Content

Inductive Inference, Part 1

- Let θ denote the model and $\mathbf{y}^n$ the data
- Define
  - $I(\theta)$: description length of the model
  - $I(\mathbf{y}^n|\theta)$: description length of the data given the model
- Find a model that minimises total codelength $I(\mathbf{y}^n, \theta)$

$$\hat{\theta} = \arg\min_{\theta} \left\{ I(\theta) + I(\mathbf{y}^n|\theta) \right\} \tag{7}$$

- Measure of model and data complexity are expressed in the same units
- Examples :
  - Minimum Description Length (MDL)
  - Minimum Message Length (MML)

Inductive Inference, Part 2

- Normalized Maximum Likelihood (NML)
  - MDL formula for total codelength of data relative to a chosen model class

$$I_{\mathrm{NML}}(\mathbf{y}^n) = \log \underbrace{\int_{\mathbf{x}^n} p(\mathbf{x}^n|\hat{\theta}(\mathbf{x}^n))d\mathbf{x}^n}_{\text{Parametric Complexity}} - \log p(\mathbf{y}^n|\hat{\theta}(\mathbf{y}^n)) \quad (8)$$

  - where $\hat{\theta}(\cdot) \in \Theta$ is the maximum likelihood estimate of θ
- Asymptotically, NML approaches the Bayesian Information Criterion (BIC) as $n \rightarrow \infty$

$$I_{\mathrm{NML}}(\mathbf{y}^n) = -\log p(\mathbf{y}^n|\hat{\theta}(\mathbf{y}^n)) + \frac{k}{2}\log n + o(1) \quad (9)$$

Single Hypothesis Testing, Part 1

- We have K competing hypotheses $H_1, \ldots H_K$
- Find candidate model $\hat{k}$ such that

$$\hat{k} = \arg\min_k \left\{ I(k) + I_k(\mathbf{y}^n) \right\} \tag{10}$$

- $I_k(\mathbf{y}^n)$: codelength of each candidate model
- $I(k)$: preamble code stating which candidate model is used
  - Prior distribution over K candidate models
  - If K is small, or models are nested

$$I(k) = \log K \tag{11}$$

Single Hypothesis Testing, Example

- We observe a dataset $\mathbf{y}^n$ of n non-negative integers
- Two possible hypotheses ($K = 2$)
  - $\mathbf{y}^n \sim$ Poisson, $(k = P)$
  - $\mathbf{y}^n \sim$ Geometric, $(k = G)$
- Two universal models, $I_p(\mathbf{y}^n)$ and $I_g(\mathbf{y}^n)$
  - Preamble code, $I(P) = I(G) = \log 2$
- Perform the test

$$\hat{k} = \begin{cases} P & \text{if } I_p(\mathbf{y}^n) + \log 2 < I_g(\mathbf{y}^n) + \log 2 \\ G & \text{if } I_p(\mathbf{y}^n) + \log 2 > I_g(\mathbf{y}^n) + \log 2 \end{cases}$$

- Could easily extend to more hypotheses, i.e. negative-binomial, etc.

Multiple Hypothesis Testing, Part 1

- Generalise previous approach to testing m independent hypotheses
- Define $\hat{\mathbf{k}}^m = (\hat{k}_1, \ldots, \hat{k}_m)$ as the set of 'accepted' hypotheses

$$\hat{\mathbf{k}}^m = \arg\min_{\mathbf{k}^m} \left\{ \sum_{j=1}^{m} l(k_j) + l_{k_j}(\mathbf{y}_i^n) \right\} \tag{12}$$

- Preamble code for k_i is now important!
  - Uniform code expresses strong ignorance
  - Optimal only when all hypotheses are equally likely (unrealistic)

Multiple Hypothesis Testing, Example

- Now observe m datasets $\mathbf{y}_i^n$, $(i = 1, \ldots, m)$ of n non-negative integers
- Assume preamble codes, $I(P)$ and $I(G)$ exist
- For all $i = 1, \ldots, m$, we perform the tests

$$\hat{k}_i = \begin{cases} P & \text{if } I_p(\mathbf{y}_i^n) + I(P) < I_g(\mathbf{y}_i^n) + I(G) \\ G & \text{if } I_p(\mathbf{y}_i^n) + I(P) > I_g(\mathbf{y}_i^n) + I(G) \end{cases}$$

- Again, easily extends to more hypotheses, i.e. negative-binomial, etc.

Multiple Hypothesis Testing, Part 2

- Example, $m = 10$ and uniform code might yield

$$\mathbf{k}^m = (P, P, P, P, P, P, P, P, P, G)$$

  Not very close to uniform !
- Suitable preamble code for $\mathbf{k}^m$ must :
  - attain shorter codelengths than uniform code for most data
  - be invariant to relabeling of the hypotheses
  - be invariant to permutations of the set $\mathbf{k}^m$
- Solution : encode $\mathbf{k}^m$ as data generated by a K-nomial distribution

Multiple Hypothesis Testing, Part 3

- Total codelength for all data sets $\mathbf{Y}^m = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ and chosen hypotheses $\mathbf{k}^m$

$$\underbrace{-\sum_{i=1}^{m}\sum_{j=1}^{K}\mathcal{I}(k_i = j)\log\alpha_j}_{I_h(\mathbf{k})} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{K}\mathcal{I}(k_i = j)I_j(\mathbf{y}_i^{n_i})}_{I_h(\mathbf{Y}^m|\mathbf{k})} - \log\Gamma(K+1)$$

  where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are hypothesis frequencies
- This is termed ‘hard’ assignment
- Each data set encoded by only one of the K candidate models

Multiple Hypothesis Testing, Part 4

- Mixture modelling interpretation

$$\underbrace{-\sum_{i=1}^{m}\sum_{j=1}^{K} r_{ij} \log \frac{r_{ij}}{\alpha_j}}_{I_s(\mathbf{k})} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{K} r_{ij} I_j(\mathbf{y}_i^{n_i})}_{I_s(\mathbf{Y}^m|\mathbf{k})} - \log \Gamma(K+1)$$

- where r_{ij} are the posterior probabilities of the hypotheses

$$r_{ij} = \frac{\exp(-I_j(\mathbf{y}_i^{n_i}))\alpha_j}{\sum_{q=1}^{K} \exp(-I_q(\mathbf{y}_i^{n_i}))\alpha_q}$$

- Optimisation of α performed using the EM algorithm

Multiple Hypothesis Testing, Part 5

- For each $i = 1, \ldots, m$, perform the test

$$\hat{k}_i = \arg\max_j \left\{ r_{i,j} \right\}$$

- Alternatively, could do decision-theoretic analysis

$$\hat{k}_i = \arg\min_j \left\{ r_{i,j} L_{i,j} \right\}$$

with $L_{i,j}$ the loss incurred by accepting hypothesis j for dataset i

Content

NML Test for GWAS data

- Compress the data under two different models
  - Genotype and phenotype are independent

$$b_1(\theta, \mathbf{k}, \mathbf{n}) = \binom{n_1}{k_1} \theta^{k_1} (1-\theta)^{n_1-k_1} \binom{n_2}{k_2} \theta^{k_2} (1-\theta)^{n_2-k_2}$$

  - Phenotype depends on the genotype

$$b_2(\boldsymbol{\theta}, \mathbf{k}, \mathbf{n}) = \binom{n_1}{k_1} {\theta_1}^{k_1} (1-\theta_1)^{n_1-k_1} \binom{n_2}{k_2} {\theta_2}^{k_2} (1-\theta_2)^{n_2-k_2}$$

- Optimal hypothesis ?
  - Use the difference in codelength !

Stochastic Complexity of Binomial Models

- The parametric complexity of a Binomial model for n samples is

$$\mathrm{COMP}(n) = \log \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^{k} \left(\frac{n-k}{n}\right)^{n-k}$$

- The NML codelengths for the two hypotheses are then

$$\begin{aligned} I_1(\mathbf{k}, \mathbf{n}) &= -\log b_1(\hat{\theta}(\mathbf{k}, \mathbf{n}), \mathbf{k}, \mathbf{n}) + \mathrm{COMP}(n_1 + n_2) \\ I_2(\mathbf{k}, \mathbf{n}) &= -\log b_2(\hat{\boldsymbol{\theta}}(\mathbf{k}, \mathbf{n}), \mathbf{k}, \mathbf{n}) + \mathrm{COMP}(n_1) + \mathrm{COMP}(n_2) \end{aligned}$$

where

$$\hat{\theta}(\mathbf{k}, \mathbf{n}) = \frac{k_1 + k_2}{n_1 + n_2}, \; \hat{\boldsymbol{\theta}}(\mathbf{k}, \mathbf{n}) = \left(\frac{k_1}{n_1}, \frac{k_2}{n_2}\right)$$

Example

- Example Phenotype/Genotype data :

x	0	1	1	0	1	0	1	1	0	1	1	0
g	0	0	1	0	1	0	1	1	0	0	1	1

splits into

$$\mathbf{x}_0 = (0, 1, 0, 0, 0, 1), \quad \mathbf{x}_1 = (1, 1, 1, 1, 1, 0)$$

- so that $\mathbf{k} = (2, 5)$, $\mathbf{n} = (6, 6)$
- Codelengths : $I_1(\mathbf{k}, \mathbf{n}) = 9.7669$ vs $I_2(\mathbf{k}, \mathbf{n}) = 9.1791$
  - Suggests that the phenotype is dependent on the genotype

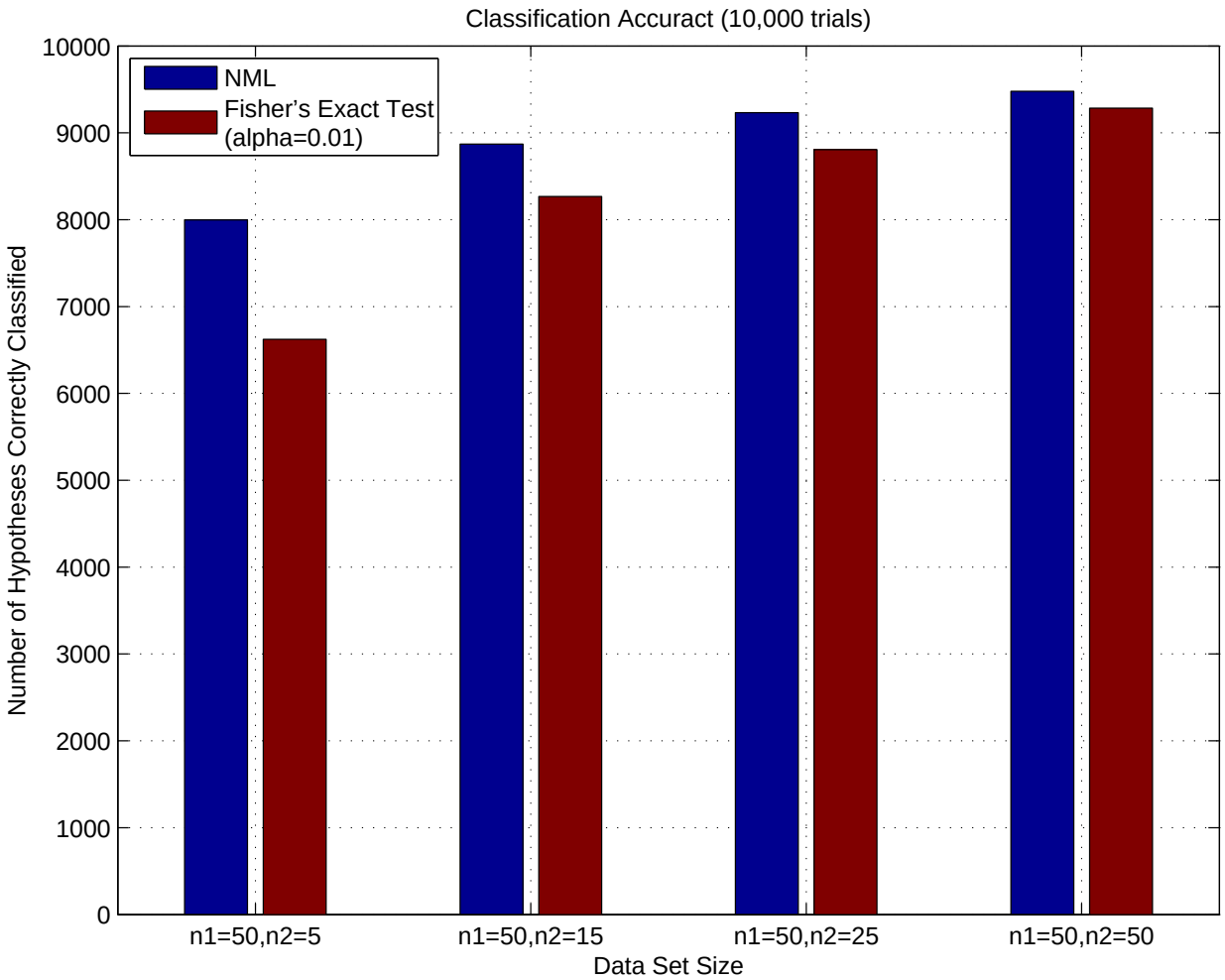
Content

Results (1)

- Single hypothesis testing
- Simulation process
  1. Choose hypothesis (same, different) randomly with $1/2$ probability
  2. Sample binomial parameter(s) uniformly form $[0, 1]$
  3. Draw two samples of size n_1, n_2 from either the same binomial or the two different binomials
  4. Present samples to criteria and ask them to nominate a hypothesis
  5. Score the choices with 0-1 loss
- Repeat many times ...

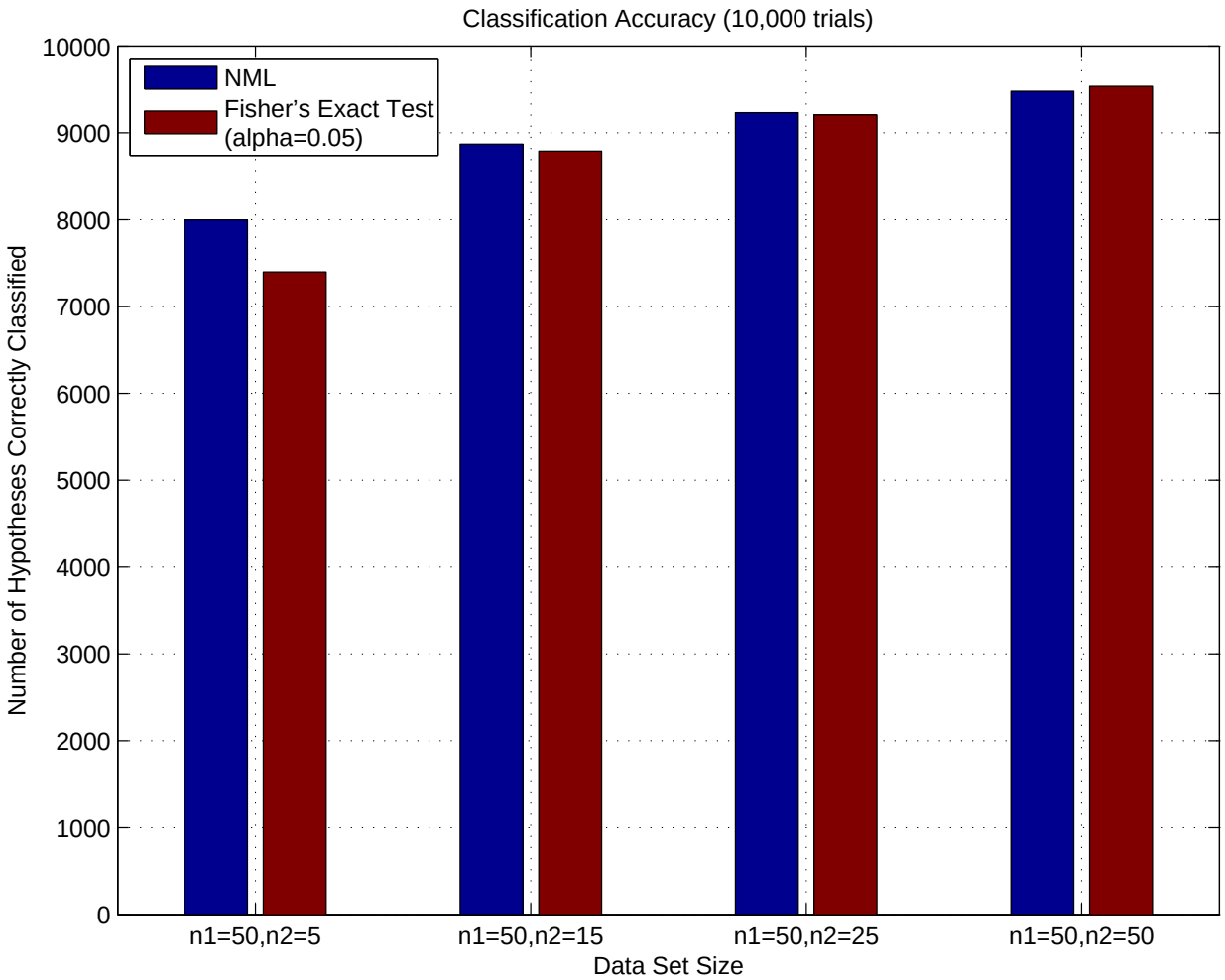
Results (2)

- Single hypothesis testing



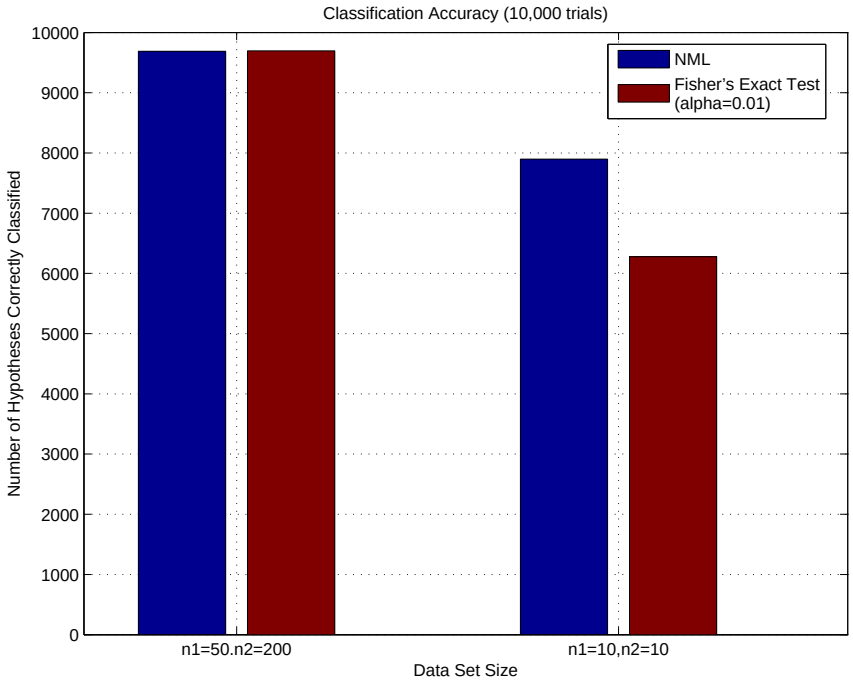
Results (3)

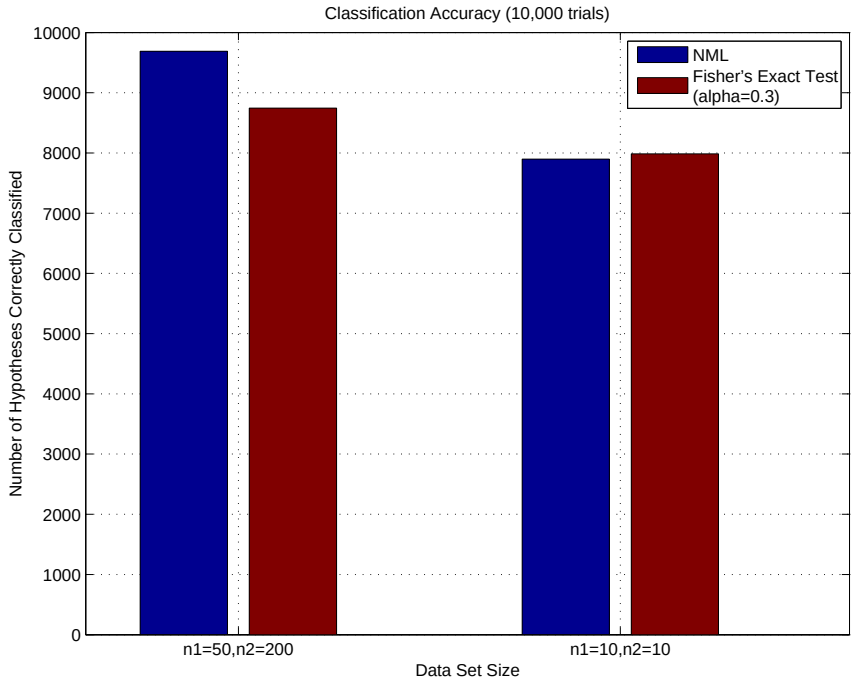
- Single hypothesis testing



Results (4)

- Single hypothesis testing



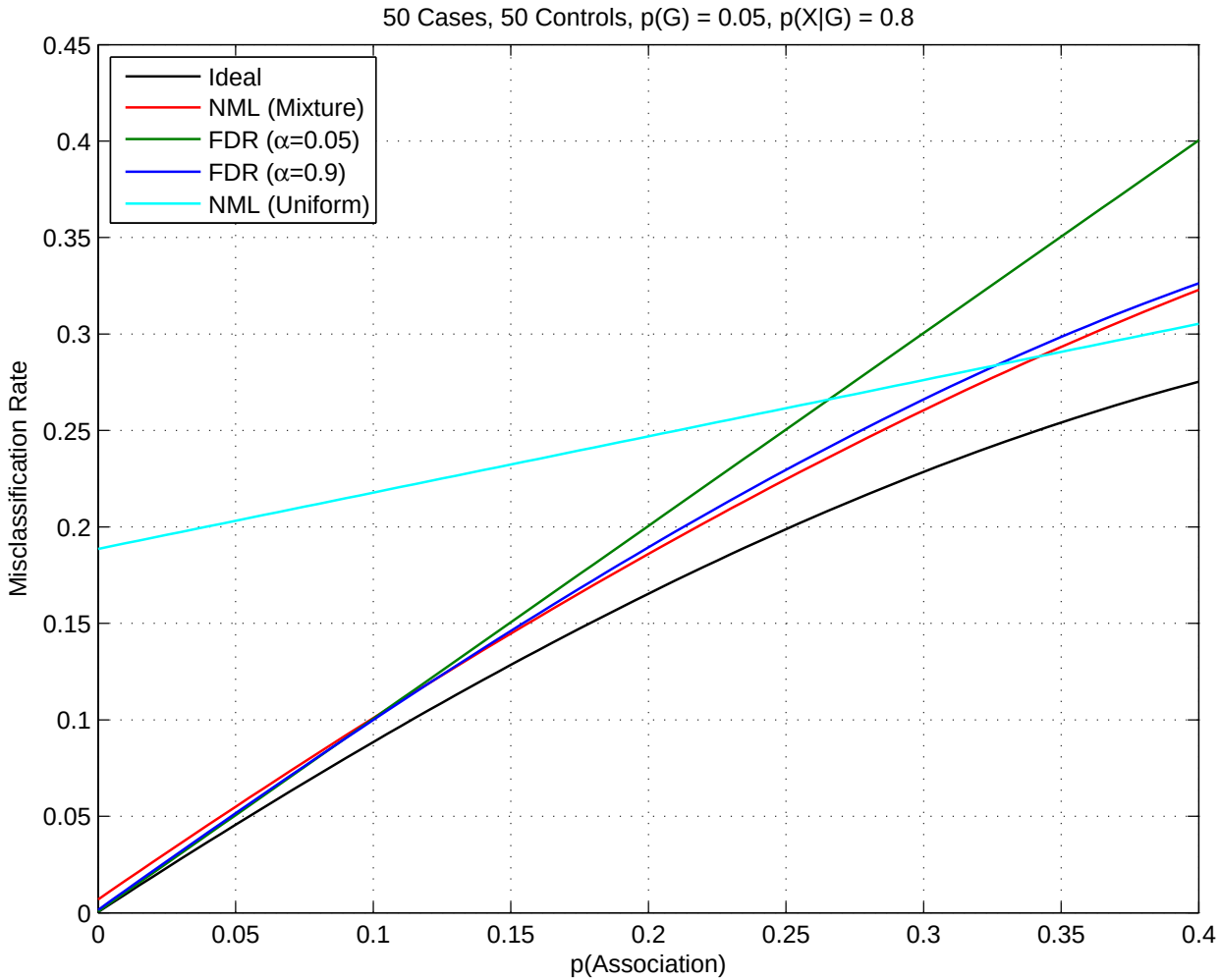


Results (5)

- Simulated GWAS Data variables
  - Number of Cases and Controls
  - Number of SNPs
  - Probability of Mutation, $p(G)$
  - Penetrance, $p(X|G)$
  - Probability of association, $p(Association)$

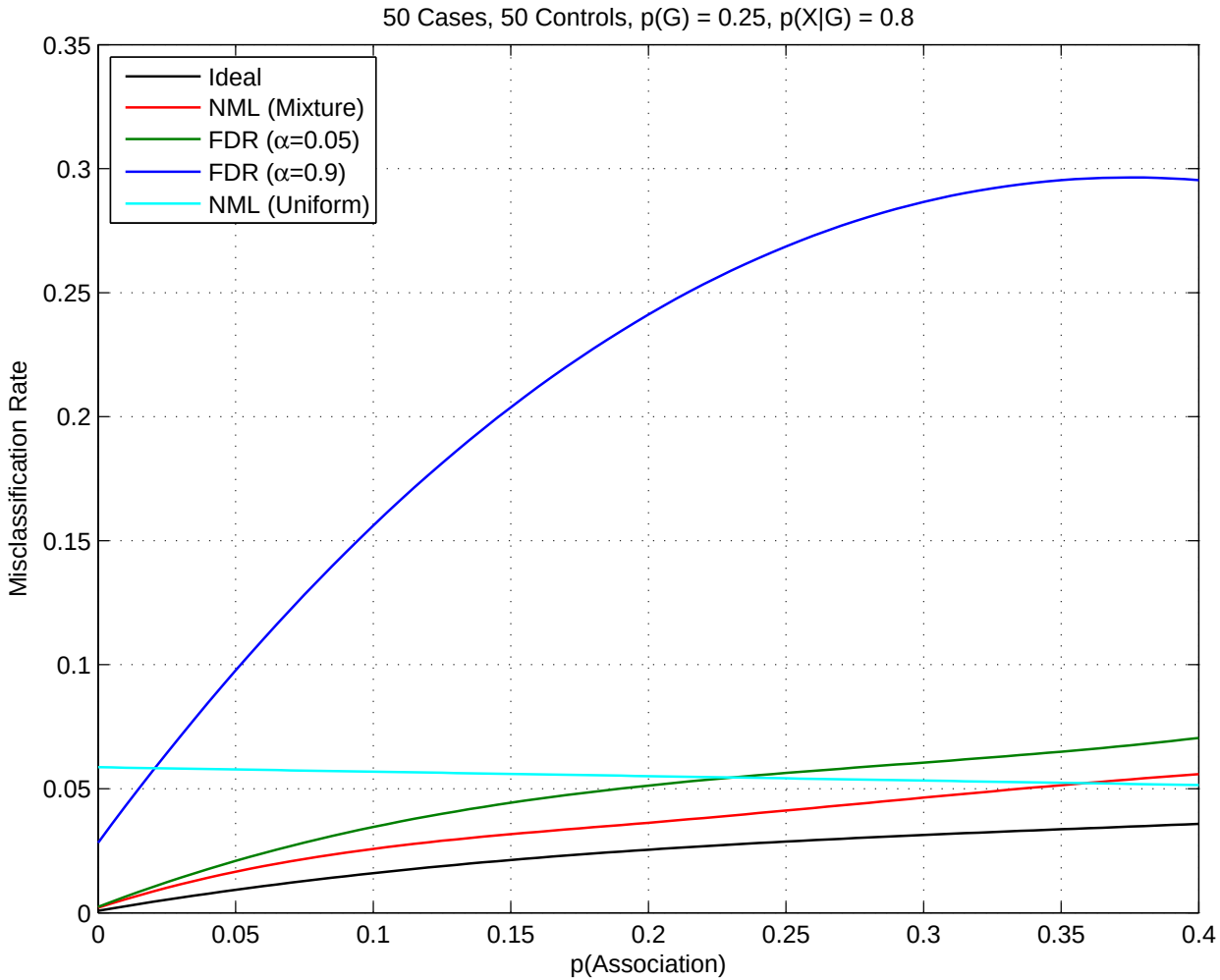
Results (6)

- Simulated GWAS Test
  - $p(G) = 0.05, p(X|G) = 0.8$



Results (7)

- Simulated GWAS Test
  - $p(G) = 0.25, p(X|G) = 0.8$



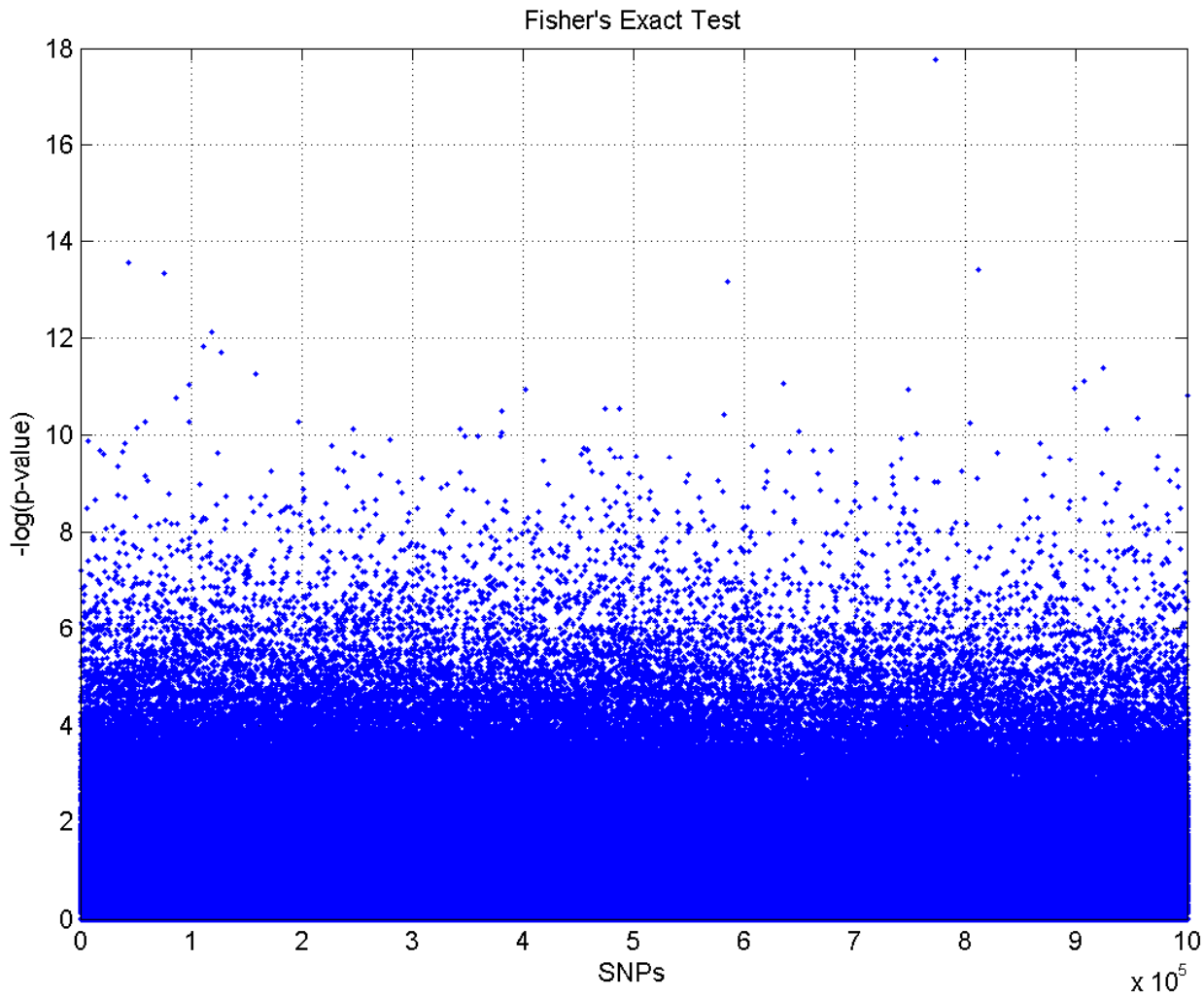
Results (8)

- Real GWAS Data
  - 1,072,820 SNPs
  - 200 cases, 200 controls

Fisher's Exact Test with FDR			Normalized Maximum Likelihood (NML)		
p-value	**RS**	**Chromosome**	**Posterior**	**RS**	**Chromosome**
1.9270e-008	rsXXX1	A	0.8700	rsXXX1	A
1.2715e-006	rsXXX2	A	0.0832	rsXXX3	B
1.4818e-006	rsXXX3	B	0.0777	rsXXX2	A
1.5926e-006	rsXXX4	B	0.0669	rsXXX4	B
1.8624e-006	rsXXX5	B	0.0597	rsXXX6	B

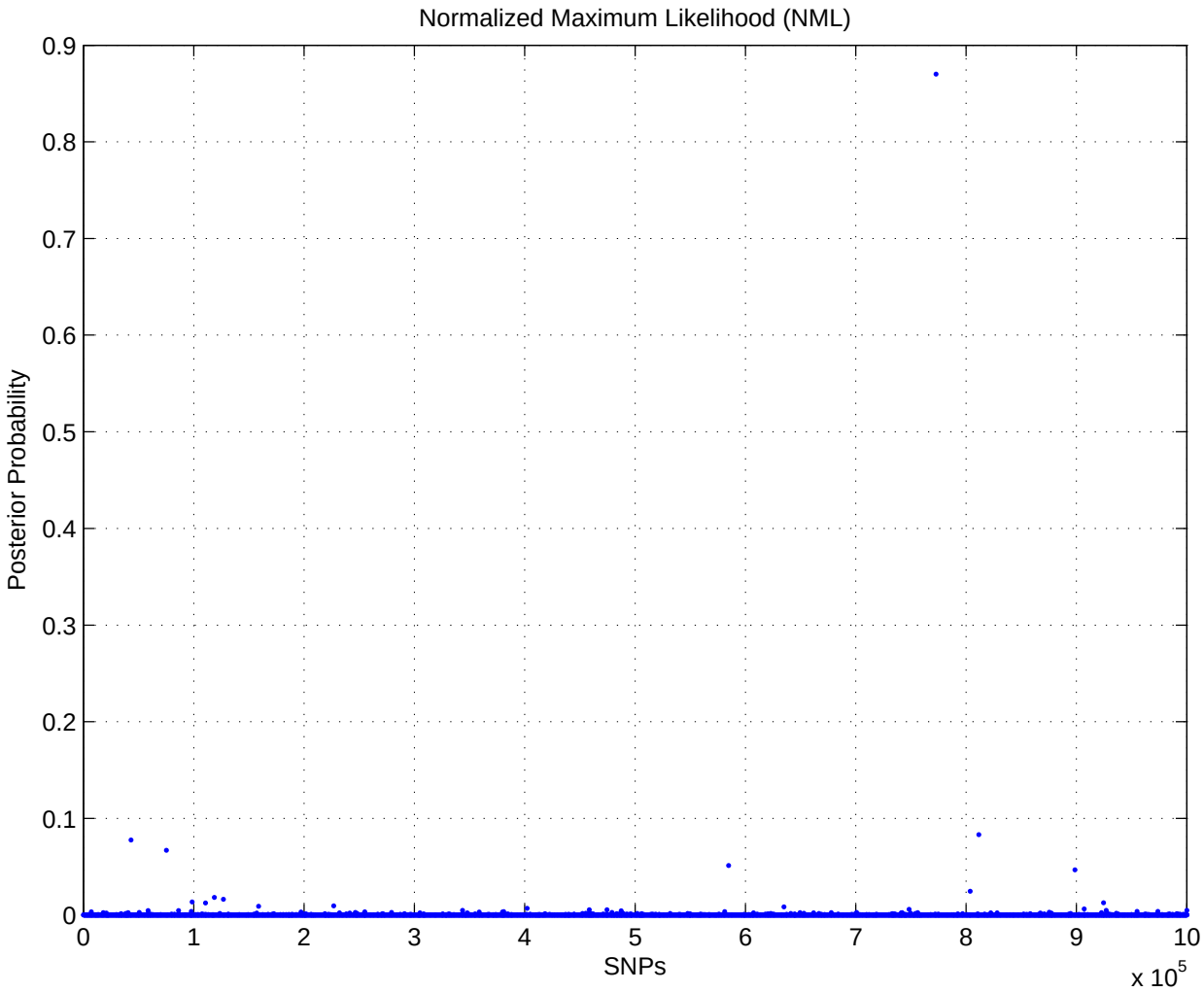
Results (9)

- Real GWAS Data



Results (10)

- Real GWAS Data



References

- Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society (Series B)*, 1995, 57, 289-300
- Grünwald, P. D. The Minimum Description Length Principle. *The MIT Press*, 2007
- Rissanen, J. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 1996, 42, 40-47
- Rissanen, J. Information and Complexity in Statistical Modeling. *Springer*, 2007
- Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 1978, 6, 461-464
- Wallace, C. S. & Freeman, P. R. Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 1987, 49, 240-252
- Wallace, C. S. Statistical and Inductive Inference by Minimum Message Length. *Springer*, 2005