Robust Computational Tools for Multiple Testing With Genetic Association Studies

Bill Welbourn
Department of Mathematics and Statistics
Utah State University

February 16, 2012

Outline

- Overview of Genome-wide Association Studies (GWAS).
- Utility of permutation tests for multiple hypothesis testing (MHT) in GWAS.
- Efficient parallel computing approach for permutation tests in GWAS: the GPER algorithm.

A GWAS

- A conventional case-control GWAS:
  - Focus lies upon locating genetic risk factors leading to disease incidence, for some well-defined phenotype.
  - Sample some number of diseased (cases) and non-diseased (controls) individuals.
  - Genotype these subjects for a common set of genetic mutations, called single-nucleotide polymorphisms (SNPs).
  - For each marker (i.e., SNP), carry out the test of the null hypothesis of no association between genotype and disease status.
  - A statistically significant association could indicate close proximity (relative to the marker) of a disease susceptible locus within the human genome.

A Single Nucleotide Polymorphism (SNP)[1]

[1] Photo taken from `http://www.mdsupport.org/images/geneticsexplained2.jpg` on February 13, 2012.

Controlling the Family-wise Type I Error Rate

- Many definitions of the Type I error rate in multiple hypothesis testing.
- For GWAS, control of the Family-wise Type I error rate (FWER) – the probability of committing at least one Type I error under $\mathcal{H}_0$ (the complete null hypothesis) – is a popular approach.
- The Bonferroni multiple testing procedure (MTP) – for m markers, at the α nominal level in the FWER compare pointwise p-values to the fraction α/m – is by far the most exploited procedure for control of the FWER in GWAS.
  - Appropriate for tests of mutually independent null hypotheses.
  - Tends to be conservative when testing correlated null hypotheses.

Resampling Accounts for Correlated Null Hypotheses

- SNPs for GWAS are chosen so that they are as independent as possible, but correlations between them nevertheless arise.
- Permutation MTPs, such as the maxT and minP, fully account for the correlation among the null hypotheses, resulting in higher statistical power for control of the FWER.
- Permutation minP and maxT are computationally challenging, and although recommended in theory, they are seldom used in practice.

PLINK: A Popular Software to Analyze GWAS Data

- Developed by Shaun Purcell at the Center for Human Genetic Research, Massachusetts General Hospital, and the Broad Institute of Harvard & MIT, with support of others[2].
- Provides a wide-range of utilities for GWAS data, including: data management; quality control; basic tests of association; and multiple hypothesis testing (MHT) correction by way of the maxT MTP.

[2] http://pngu.mgh.harvard.edu/~purcell/plink/

PLINK (Cont'd)

- Implementation of the maxT MTP takes considerable time to process within PLINK.
  - Example: 1K maxT permutations upon 45 168 SNPs of a Bipolar Disorder GWAS data set of $n_1 = 1001$ cases and $n_0 = 1034$ controls takes about 48 minutes upon current desktop computer technology[3].
  - Extrapolating this result out to the entire GWAS sample of $\sim$ 770K SNPs, approximately 55 days of computing would be required to perform 100K maxT permutations.

[3] Computer system comprised of an Intel Core i7 920 processor, operating at 3 GHz clock speed.

Parallel Computing Upon the GPU

- We propose a parallel computing approach using NVIDIA's CUDA architecture, denoted as the GPER[4] algorithm.
  - Architecture introduced by NVIDIA in November 2006.
  - CUDA enabled GPUs have hundreds of cores that can collectively run thousands of computing threads.
- We interface with CUDA C, an extension to the C programming language, which allows the user to program in parallel upon the GPU.
- CUDA is especially well-suited for problems involving binary operations (e.g., sums, products, single-level conditional statements).

[4] Named from the acronym GPU and the word permutation, emphasizing the utility of the graphics processing unit (GPU) in the algorithm.

Application

- We applied $R = 102\,480^{\dagger}$ maxT permutations within GPER upon the entire $m = 769\,672$ SNP panel for a GWAS investigating Bipolar disorder.
  - **GPER completed this task in under 110 minutes.**
- We also applied $R = 102\,480$ maxT permutations against the $m = 45\,168$ SNP loci, discussed earlier within the presentation.
  - GPER completed this task in 6 minutes and 20 seconds. Extrapolating the result of the PLINK run, **GPER is more than 770x times faster than PLINK** upon the desktop computer system used for benchmarking.

[†] GPER performs blocks of 2^{10} permutations. Here, $R = 100 \times 2^{10}$.

Performance Benchmarking

- We simulated marker panels of size $m = 40\text{K}$ for various sample sizes (n) and balancing characteristics ($n_1 : n_0$) of GWAS samples. $R = 1\text{K}$ and $R = 10\,240$ data permutations for PLINK and GPER, respectively.

Computational Time (Minutes)

n	Cases (n_1)	Controls (n_0)	GPER	PLINK	Speedup
2000	1000	1000	0.6	43.0	**785x**
	900	1100	0.5	42.3	**830x**
	800	1200	0.5	42.1	**890x**
4000	2000	2000	1.2	89.7	**775x**
	1800	2200	1.1	88.8	**840x**
	1600	2400	1.0	87.1	**910x**

Historical Revisions Using Bipolar GWAS Data

NVIDIA GeForce GTX 260 GPU:

- December 2009 – Speedup of about 19x over PLINK.
- February 2010 – Speedup of about 40x over PLINK.

NVIDIA GeForce GTX 470 GPU:

- May 2010 – Speedup of about 175x over PLINK.
- April 2011 – Speedup of about 360x over PLINK.
- September 2011 – Speedup of about 770x over PLINK.

NVIDIA GeForce GTX 580 GPU is projected to increase GPER performance by as much as 40% over the NVIDIA GeForce GTX 470 GPU.

Outline

- Dependence of multiple testing distribution on sampling design.
- Corrected permutation tests in GWAS.
- Application.

The Test Statistics Null Distribution

- The test statistics null distribution (Q_0) – the distribution which defines the critical region for the test statistic at locus j (T_j), $j = 1, \ldots, m$, under $\mathcal{H}_0$ – is arguably the most vital component to the MHT problem.
- An incorrect choice of Q_0 could lead to control in the Type I error rate at a level other than that intended.
- Correct identification of Q_0 has not been correctly handled within the GWAS literature.

Asymptotic Distributional Assumption for T_j Under $\mathcal{H}_0$

- When testing hundreds of thousands of null hypotheses, the test statistic rejection region for T_j will call for its corresponding null hypothesis to be rejected for large values in $|T_j|$.
- To avoid the computational burden of the maxT and minP MTPs, conventional GWAS practice is to approximate the distribution of T_j by its underlying asymptotic distribution under $\mathcal{H}_0$, $\tilde{Q}_0$.
- However, $\tilde{Q}_0$ is a *continuous* distribution, whereas Q_0 is in fact *discrete*.
  - Once a random sample has been drawn, the margins of the 2×3 contingency table for SNP locus j – cross-classifying phenotype and genotype – are fixed.
  - Thus, there is only a finite number of realizations for T_j.

Identically Distributed Assumption for T_j Under $\mathcal{H}_0$

- Conventional MHT methods in GWAS consider both margins of each 2×3 table as being fixed.
- In this regard, the distribution of T_j under $\mathcal{H}_0$ will depend upon the marginal values for its corresponding 2×3 table.
- However, if the T_j are not identically distributed under $\mathcal{H}_0$, application of the maxT MTP can lead to an unbalanced multiplicity correction [Dudoit et al. (2003)].
- There is no compelling reason to systematically favor some null hypotheses over others in GWAS.

Identically Distributed Assumption for T_j Under $\mathcal{H}_0$

2×3 Contingency Tables for Two Loci of a Bipolar GWAS

		Number of copies of minor allele			Totals
		0	1	2	
Locus 1	Cases	261	487	247	995
	Controls	309	513	203	1025
	Totals	570	1000	450	2020
Locus 2	Cases	954	8	0	962
	Controls	1004	4	0	1008
	Totals	1958	12	0	1970

- The distribution of T_j is likely very different between these two loci under $\mathcal{H}_0$. As a result, the maxT MTP can lead to an unbalanced multiplicity adjustment.

Genotype Data are Random for a Case-Control Study

- When sampling from the population, the phenotypes are fixed – row margin of the 2×3 table is fixed; genotype data are random – rows of the table are independent random trinomials.

2×3 Contingency Table for Locus j – Random Column Margin

	Number of copies of minor allele			Totals
	0	1	2	
Cases	X_{j10}	X_{j11}	X_{j12}	n_1
Controls	X_{j00}	X_{j01}	X_{j02}	n_0
Totals	X_{j0}	X_{j1}	X_{j2}	n

- Accounting for the randomness in the genotype data leads to a richer support for T_j, when compared to freezing the values of the column margin.

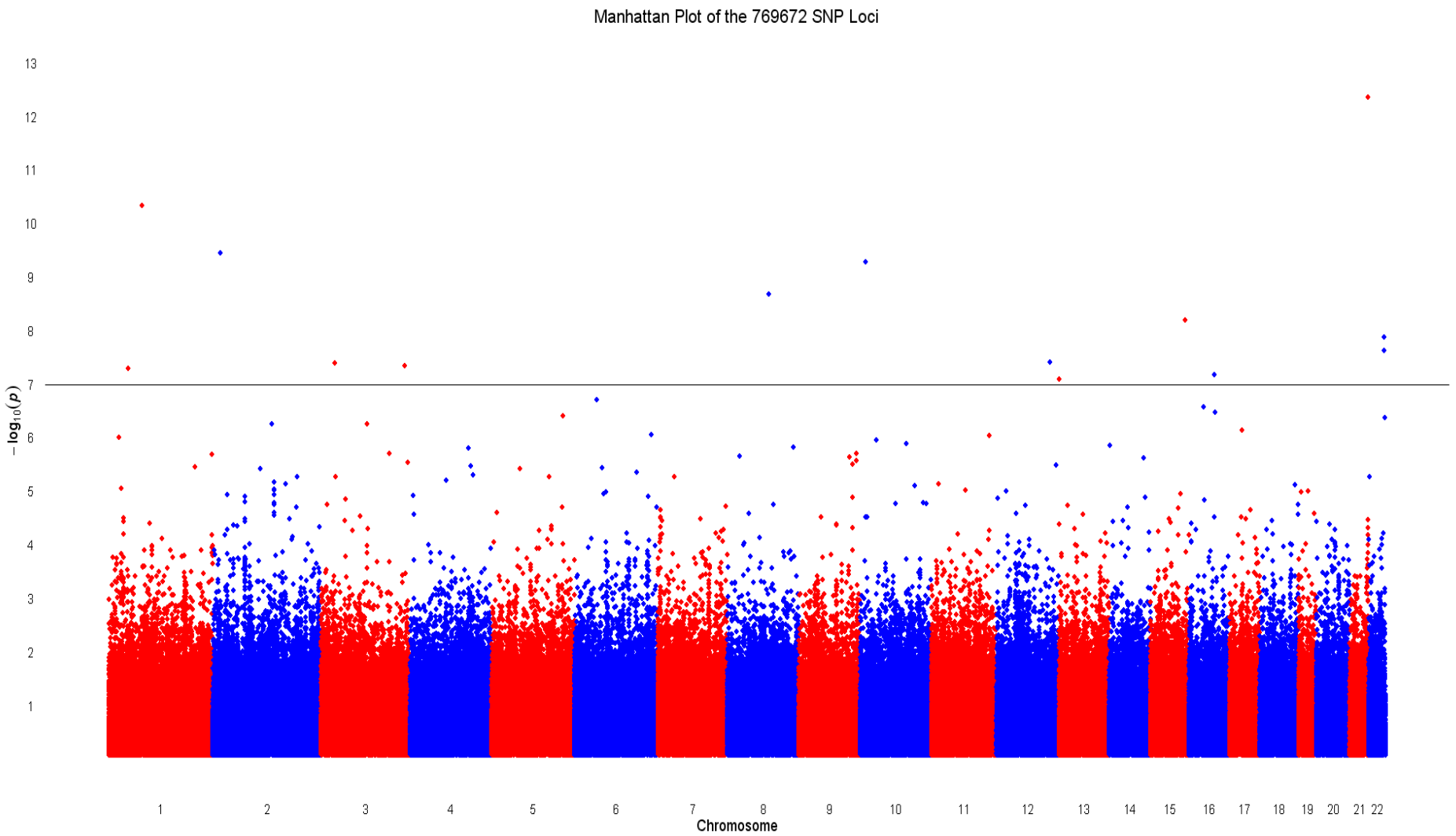
Motivation – Proposal

- Reliance upon the χ^2_1 distribution for the Cochran-Armitage trend test (CATT) statistic under $\mathcal{H}_0$ can lead to improper control of the FWER in GWAS.
  - We have demonstrated through simulation that the maxT and minP MTPs are not immune to this – unbalanced multiplicity adjustment.
- We propose using the exact unconditional distribution for the CATT statistic within GWAS (denoted Q_0^*).
  - Utility within the minP MTP leads to high statistical power and proper control of the FWER.
  - For a balanced GWAS, correct identification of Q_0 for the CATT statistic has the ability to boost statistical power to detect associations at loci possessing a rare variant allele.

Problems – Resolutions

- The parameter vectors for the random trinomials under $\mathcal{H}_0$ become a nuisance.
  - We propose estimating the appropriate parameters under $\mathcal{H}_0$ at their MLEs. In this regard, computed *p*-values are not exact, but rather approximate known as *bootstrap p*-values.
- Utilization of Q_0^* in practice within the minP MTP presents a very difficult computational problem.
  - Projected to take **thousands of years** to analyze GWAS data using this approach upon the personal computer.
  - We have developed a CUDA-based algorithm to rapidly estimate the exact unconditional Type I error of the CATT statistic under $\mathcal{H}_0$ from a *truncated* unconditional reference set.
  - We propose utility of *p*-value lookup tables within the minP MTP.

Manhattan Plot of Pointwise *P*-values Under Q_0^*



Statistically Significant Markers at the 5% FWER[6]

		Adjusted P-value		Risk Estimates	
CHR	t_j	maxT ($\tilde{Q}_0$)	minP (Q_0^*)	OR	95% CI
21	43.5	< 0.001	< 0.001	47.6	(6.5, 346.0)
1	36.6	0.001	< 0.001	21.6	(5.2, 89.9)
2	33.3	0.003	< 0.001	19.9	(4.8, 82.9)
10	35.7	0.001	< 0.001	13.6	(4.2, 44.2)
8	28.6	0.035	0.001	58.0	(3.5, 952.7)
15	30.3	0.016	0.003	8.9	(3.5, 22.6)
22	27.9	0.051	0.007	0.06	(0.01, 0.25)
22	27.0	0.079	0.010	16.6	(4.0, 69.6)
3	26.2	0.116	0.015	16.1	(3.8, 67.7)
12	25.2	0.182	0.020	28.6	(3.9, 211.0)
3	25.5	0.165	0.021	15.8	(3.8, 66.3)
1	26.6	0.094	0.026	8.1	(3.2, 20.7)
16	25.2	0.198	0.027	0.04	(0.01, 0.26)
13	24.3	0.138	0.040	8.9	(3.1, 20.2)

[6] $B = 10.2\text{K}/102.4\text{K}$ maxT/minP Permutations w/in GPER.

Statistically Significant Markers at the 5% FWER[6]

		Adjusted *P*-value		Risk Estimates	
CHR	t_j	maxT ($\tilde{Q}_0$)	minP (Q_0^*)	OR	95% CI
21	43.5	< 0.001	< 0.001	47.6	(6.5, 346.0)
1	36.6	0.001	< 0.001	21.6	(5.2, 89.9)
2	33.3	0.003	< 0.001	19.9	(4.8, 82.9)
10	35.7	0.001	< 0.001	13.6	(4.2, 44.2)
8	28.6	0.035	0.001	58.0	(3.5, 952.7)
15	30.3	0.016	0.003	8.9	(3.5, 22.6)
22	**27.9**	**0.051**	**0.007**	**0.06**	**(0.01, 0.25)**
22	**27.0**	**0.079**	**0.010**	**16.6**	**(4.0, 69.6)**
3	**26.2**	**0.116**	**0.015**	**16.1**	**(3.8, 67.7)**
12	**25.2**	**0.182**	**0.020**	**28.6**	**(3.9, 211.0)**
3	**25.5**	**0.165**	**0.021**	**15.8**	**(3.8, 66.3)**
1	**26.6**	**0.094**	**0.026**	**8.1**	**(3.2, 20.7)**
16	**25.2**	**0.198**	**0.027**	**0.04**	**(0.01, 0.26)**
13	**24.3**	**0.138**	**0.040**	**8.9**	**(3.1, 20.2)**

[6] $B = 10.2\text{K}/102.4\text{K}$ maxT/minP Permutations w/in GPER.

Outline

- Explain why we study GxE interactions. Highlight the conventional approach to detecting this type of interaction in genetic association studies, and the challenges imposed with this approach in a multiple testing setting.
- Outline our proposed method for detecting gene-environment interaction.
- Illustrate implementation of our proposed methodology for a binary environmental factor: in a general context; and, by way of empirical data.

Why Do We Study GxE Interactions?

- They can illuminate fundamental biological mechanisms involved within disease etiology.
- They can be important for risk prediction and for evaluating the benefit of changes in modifiable environmental exposures.
- Failure to adequately account for GxE interaction in a genetic analysis can *mask the effects* of both genetic and environmental factors.
- Can lead to a better understanding of the complete etiology of disease, inclusive of both distinct and interacting pathways comprised of genetic and environmental factors.

How Do We Detect GxE Interactions?

- Setup:
  - Let D be an indicator of disease status. Assume that we have a sample of cases ($D = 1$) and [unrelated] controls ($D = 0$).
  - We consider a binary environmental factor, with E an indicator random variable for exposure.
  - Assume m SNP markers, genotyped upon the study subjects.
- If G denotes some genetic model of inheritance (GMI; e.g., additive, dominant) of the genotypes, one might consider a model for a given SNP of the form

$$\text{logit}\left(\Pr\left(D = 1 | G, E\right)\right) = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE. \qquad (1)$$

- A standard approach to test for GxE interaction would be to perform a 1-df test of $H_0 : \beta_{ge} = 0$ upon this model for each SNP.

An Induced Multiple Testing Problem

- In conducting the test of $H_0 : \beta_{ge} = 0$ upon the model (1) across SNP loci, a multiple testing problem is introduced.
- We could perform a Bonferroni correction to control the FWER. However, this approach is too conservative, insofar as the test statistics are most certainly correlated.
- Alternatively, we might consider a permutation approach. However, strong control of the FWER is not guaranteed [Pollard and van der Laan (2004); Dudoit and van der Laan (2008)].
- Testing several GMIs across loci exacerbates the multiplicity problem.

Approach

- We propose adapting the SNP-SNP interaction framework (denoted LPCV) of Boulesteix et al. (2007) to include tests of GxE interaction.
- To do this, we consider *logical patterns* in G (coded 0, 1, or 2) and E of the form, for example,

$$L_A = (G \in \{0, 1\}) \wedge (E = 1),$$

  of which may bring about higher or lower risk of developing a particular complex disease when compared to some alternative logical pattern (denoted L_B).
- In particular, we consider some q-fold collection of logical patterns, $\{L_{A_l}, L_{B_l}\}_{l=1,\ldots,q}$.

Overview of the LPCV Approach

1. For each $l = 1, \ldots, q$, we collect the subset of data pertaining to subjects satisfying either of the patterns, L_{A_l} or L_{B_l}.

2. Dichotomize the subset of data in accordance with L_{A_l}, say.

3. The test of the null hypothesis of no association between disease status (D) and the dichotomized indicator random variable – pertaining to subject membership to pattern L_{A_l}, say – is carried out, yielding a chi-square test statistic.

4. The maximum of these test statistics is selected.

5. Control of the FWER is by way of the distribution of the maximum chi-square test statistic under the complete null hypothesis.

Detecting GxE Interaction by MaxT

- We propose control over the FWER by way of the permutation-based maxT procedure of Westfall and Young (1993).
- Our approach simultaneously assesses: GxE interaction, corresponding with each of the dominant and recessive GMIs; the main effect upon the environmental factor; and, the main genetic effect, corresponding with each of the dominant and recessive GMIs.
- The maxT approach corrects for the multiple testing problem across genetic markers.
- We have leveraged the data management tools of GPER in our development of efficient computational tools.

Proposed Logical Patterns for a Binary E

I	Effect	Logical Pattern (L_{A_I})
1	GxE	$(G = 0) \wedge (E = 0)$
2		$(G = 0) \wedge (E = 1)$
3		$(G \in \{0, 1\}) \wedge (E = 0)$
4		$(G \in \{0, 1\}) \wedge (E = 1)$
5		$(G \in \{1, 2\}) \wedge (E = 0)$
6		$(G \in \{1, 2\}) \wedge (E = 1)$
7		$(G = 2) \wedge (E = 0)$
8		$(G = 2) \wedge (E = 1)$
9	G	$(G = 0) \wedge (E \in \{0, 1\})$ (Dominant GMI)
10		$(G \in \{0, 1\}) \wedge (E \in \{0, 1\})$ (Recessive GMI)
$q = 11$	E	$(G \in \{0, 1, 2\}) \wedge (E = 1)$

- For a binary environmental factor, we consider L_{B_I} to be the complement of L_{A_I}, for all $I = 1, \ldots, q = 11$.

Summarized Data at a SNP Locus

- If $X = 1 + G + 3E \in \{1, \ldots, 6\}$, cross-classification of the case-control data at some SNP locus can be summarized by a 2×6 contingency table.

Cross-Classification of D and X

	Value of X						Totals
	1	2	3	4	5	6	
Cases	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	r_1
Controls	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	r_2
Totals	c_1	c_2	c_3	c_4	c_5	c_6	n

- For example, $X = 1$ for subjects with $E = 0$ and $G = 0$.

Dichotomize Data

- Consider the logical pattern $L_{A_1} = \underbrace{(G = 0) \wedge (E = 0)}_{X=1}$ for example. Then, we have

Collapsed 2×6 Table

	Logical Pattern		Totals
	L_{A_1}	L_{B_1}	
Cases	x_{11}	$\sum_{j \in \{2,\ldots,6\}} x_{1j}$	r_1
Controls	x_{21}	$\sum_{j \in \{2,\ldots,6\}} x_{2j}$	r_2
Totals	c_1	$\sum_{j \in \{2,\ldots,6\}} c_j$	n

Conduct Chi-Square Tests

- If W_l denotes the indicator random variable with success/failure defined by subject membership to logical pattern L_{A_l}/L_{B_l}, we consider testing the hypotheses

$$\begin{aligned} H_0^{(l)} &: \text{“No association between } D \text{ and } W_l\text{”} \\ H_A^{(l)} &: \text{“Risk of disease depends on level of } W_l\text{”} \end{aligned} \tag{2}$$

- To test these hypotheses, we use the Pearson chi-square test (PCT). If Z_l^2 denotes a random value for the corresponding test statistic, then under $H_0^{(l)}$, $Z_l^2 \dot{\sim} \chi_1^2$.

Multiple Testing Correction

- If

$$Z^2_{\max} = \max\left\{Z^2_1, \ldots, Z^2_q\right\},$$

denotes the maximum chi-square test statistic under $\mathcal{H}_0 = \cap_{l=1}^{q} H_0^{(l)}$, the maxT adjusted p-value for p_l is given by

$$\tilde{p}_l = \Pr\left(Z^2_{\max} \geq z_l^2 | \mathcal{H}_0\right) = \text{FWER}, \tag{3}$$

where the final equality holds assuming $\mathcal{H}_0$ is in fact true.

- Thus, at the α level of the FWER, we reject $H_0^{(l)}$ whenever $\tilde{p}_l \leq \alpha$.
- Computing (3) requires knowledge of the underlying distribution of the test statistic $Z^2_{\max}$ under $\mathcal{H}_0$.

A Permutation Approach for Control Over the FWER

- Boulsteix et al. (2007) utilize the PDF of a Multivariate Normal Distribution (MVN) to approximate adjusted *p*-values.
  - Asymptotic approximation could be poor for rare allele frequencies and/or rare population prevalence of exposure.
  - Cannot [directly] adjust for multiple testing of GxE interaction across loci.
- Conditional on the observed data, we propose multiple testing correction by way of the permutation null distribution of $Z^2_{\max}$.
  - Does not approximate the null distribution, resulting in proper control over the FWER.
  - Adjusts for multiple testing of GxE interaction across loci.

Introduction

- A candidate pathway, consisting of genes involved in modulating reactive oxygen species (ROS; chemically reactive molecules carrying oxygen) was constructed over 4 genes: eosinophil peroxidase (*EPX*); myeloperoxidase (*MPO*); hypoxia-inducible factor-1A (*HIF1A*); and nitric oxide synthase (*NOS2A*).
- Lifestyle factors of interest included use of aspirin (NSAIDs) and cigarette smoking, each dichotomized into recent/non-recent categories.
- Our interest lied not solely upon the main effects of these candidate genes and lifestyle (environmental) factors, but more importantly on their synergistic effect towards the risk of cancer.

Materials and Methods

- 1555 cases of colon cancer; 1956 healthy controls.
- A total of 29 biallelic SNP markers: 8 markers for *EPX*; 2 markers for *MPO*; 4 markers for *HIF1A*; and 15 markers for *NOS2A*.
- GxE interaction was assessed in two ways: (a) by GEM – using 100K data permutations; and, (b) by testing – using a 1-df LRT – the null hypothesis $H_0 : \beta_{ge} = 0$ for the model:

$$\text{logit}\left(\Pr\left(D = 1|G, E\right)\right) = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE,$$

  under each of the dominant and recessive GMIs (LRTI).
- Adjustment for multiple testing at gene level. We applied the pACT approach of Conneely and Boehnke (2007) for LRTI approach.

Example: G - SNP rs10853004 and E - NSAID Use

Cross-Classification of D and X at This Locus

	Value of X						Totals
	1	2	3	4	5	6	
Cases	509	447	97	220	213	52	1538
Controls	505	497	133	389	336	79	1939
Totals	1014	944	230	609	549	131	3477

Example: G - SNP rs10853004 and E - NSAID Use

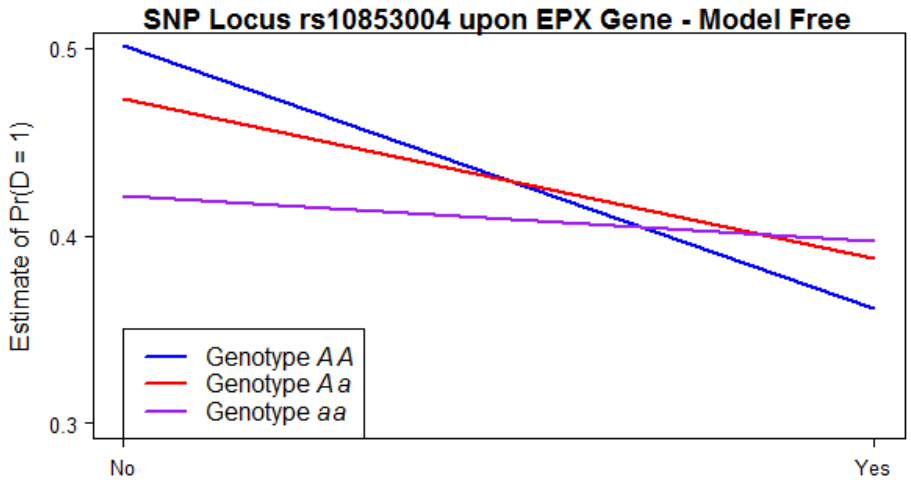
- If A/a denote the major/minor alleles at the locus, then:

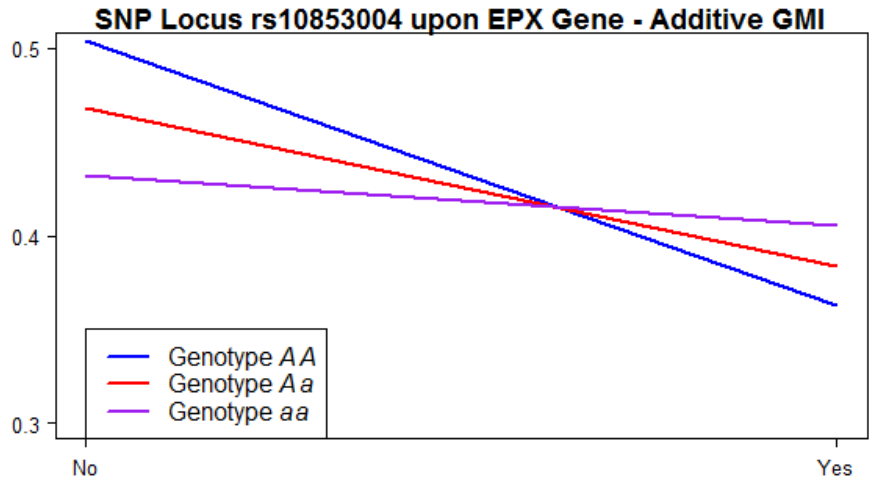
Stratified Genotype Odds Ratios

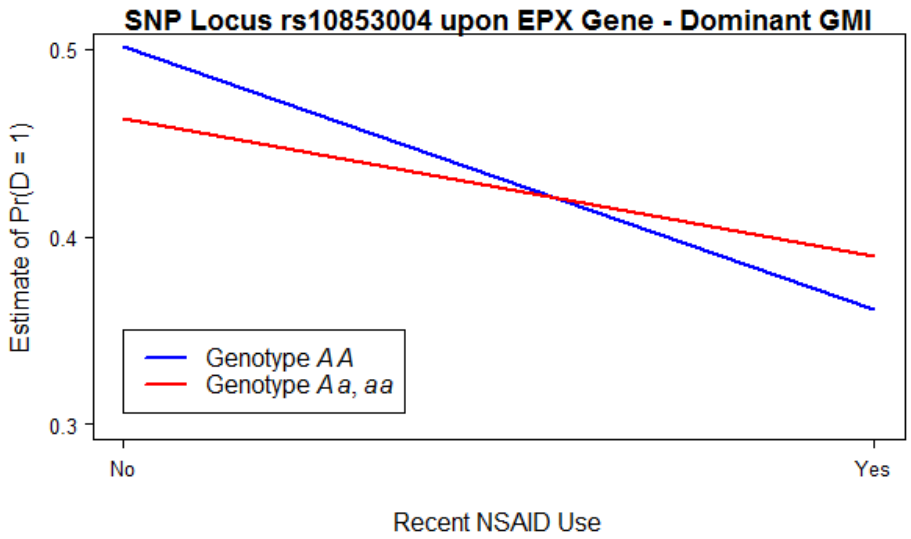
	Non-recent NSAID Use	Recent NSAID Use
$OR_{Aa:AA}$	0.89	1.12
$OR_{aa:Aa}$	0.81	1.04
$OR_{aa:AA}$	0.72	1.16

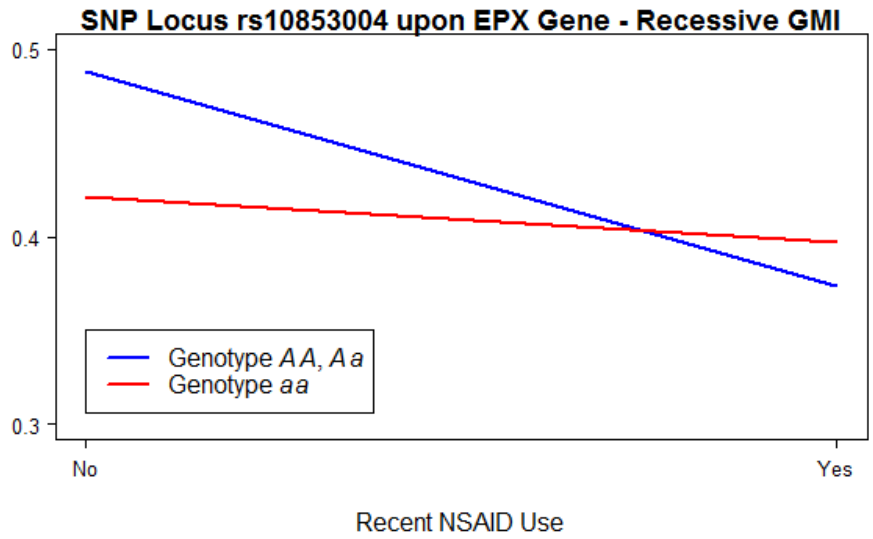
- Risk estimates between E strata are in opposite directions... **Cross-interaction** pattern of GxE interaction is apparent here.
- After testing both the dominant and recessive GMIs, **no statistically significant genetic main effect is apparent at this locus** ($p > 0.2$).

Example: G - SNP rs10853004 and E - NSAID Use









Example: G - SNP rs10853004 and E - NSAID Use

- Consider the logical pattern $L_{A_3} = (G \in \{0, 1\}) \wedge (E = 0)$. Corresponds to $X \in \{1, 2\}$.

Cross-Classification of D and X at This Locus

	Value of X						Totals
	1	2	3	4	5	6	
Cases	509	447	97	220	213	52	1538
Controls	505	497	133	389	336	79	1939
Totals	1014	944	230	609	549	131	3477

Example: G - SNP rs10853004 and E - NSAID Use

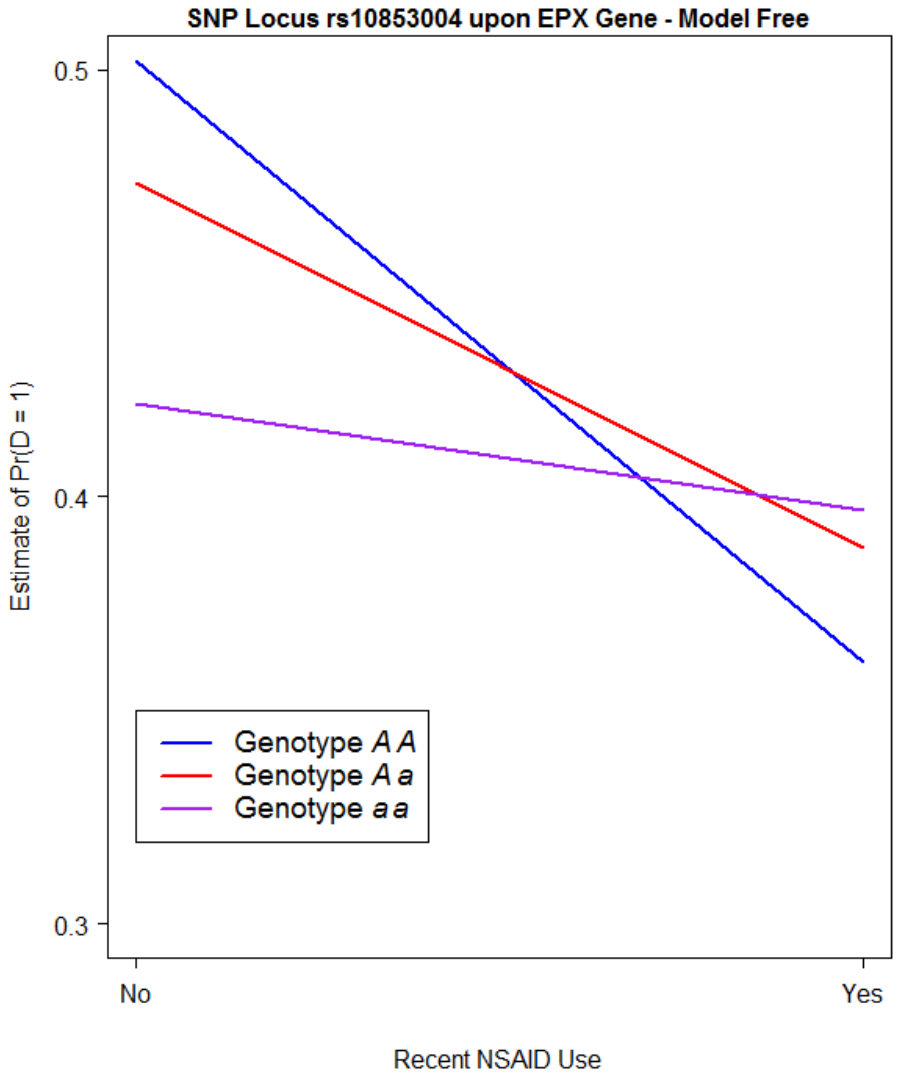
- Consider the logical pattern $L_{A_3} = (G \in \{0, 1\}) \wedge (E = 0)$. Corresponds to $X \in \{1, 2\}$.

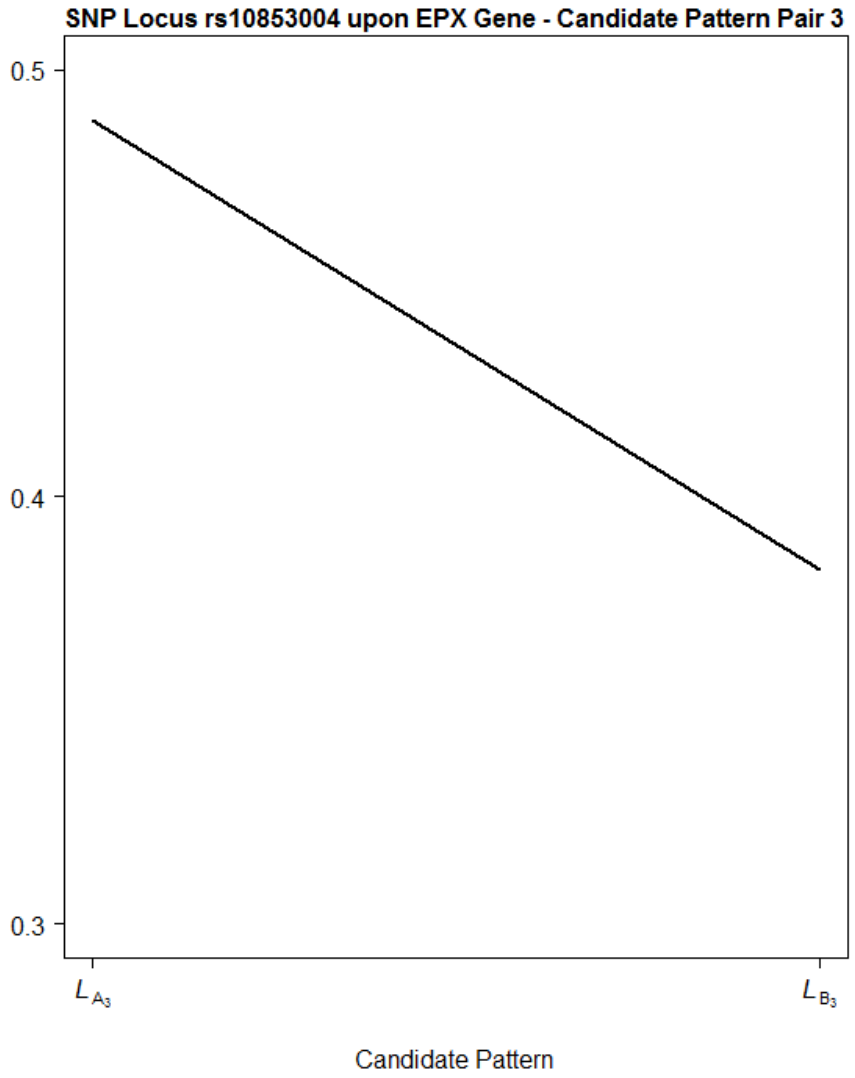
Collapsed 2×6 Table at This Locus

	Logical Pattern		Totals
	L_{A_3}	L_{B_3}	
Cases	956	582	1538
Controls	1002	937	1939
Totals	1958	1519	3477

- PCT statistic, $z_3^2 = 38.2$; $\tilde{p}_3 < 0.001$.

Example: G - SNP rs10853004 and E - NSAID Use





Results

- After multiple testing correction at the gene level, GEM detected statistically significant GxE interaction upon all 29 genetic markers with NSAID use in their synergistic effect toward risk of colon cancer – all adjusted *p*-values less than 10^{-3}.
- After multiple testing by way of pACT, no statistically significant GxE interactions were detected in applying the LRTI approach – all adjusted *p*-values greater than 0.2.

Challenges With the Analysis of Genetic Data

- The mapping of the human genome and the completion of the Human HapMap project over the past decade have significantly altered how research is conducted with respect to the genetic epidemiology of human disease.
- As laboratory techniques continue to improve and costs decrease, the volume of genetic data will inexorably rise, and robust tools for data management, statistical analysis, and computation will likewise need to keep pace.

Proposed Resolutions

- We have proposed two data management techniques and a parallel processing algorithm (named GPER), whose collective aim is to accelerate simulation of the permutation null distributions for the maxT and minP MTPs upon GWAS data.
- We extended these computational and data management tools, and proposed tools which enhance the statistical analysis governing the Cochran-Armitage trend test (CATT) statistic upon GWAS data.
  - In practice, these proposed enhancements introduce a rather profuse computational problem. We leveraged upon the GPU basis of the GPER algorithm and proposed a parallel processing approach to tackle this computational problem.

Proposed Resolutions (Cont'd)

- We extended the utility of the maxT MTP, adapting its control over the FWER when detecting markers involved in gene-environment interactions.
- We have proposed several tools for addressing the computational problems arising from adapting GEM in practice, including a data management tool analogous to that proposed for GWAS data. An `R` package is under development for our GEM approach.
- In the case of assessing a GxE interaction upon a single genetic marker and a binary environmental factor, we have proposed a network algorithm (NA) approach which produces exact conditional maxT adjusted *p*-values.

Limitations/Future Directions

- Our enhancements for GWAS propose estimating nuisance parameters at their MLEs. The corresponding *p*-values are approximate, called bootstrap *p*-values.
- We would like to implement the ability to account for confounding factors within GEM.
- The FWER may not be the ideal Type I error rate to control for GEM; FDR may result in more power to detect GxE interaction. Especially important, as genome-wide interaction studies are becoming common [Ober and Vercelli (2011)].
- GEM can only be assumed to control the FWER in the *weak* sense. However, our simulations indicate proper control of the FWER under several partial null hypothesis scenarios.

Questions?

Thank you.