INSTITUT DE CARDIOLOGIE DE MONTRÉAL

Université de Montréal

Across and down in large SNP studies: the MAX test vs SAS PROC CASECONTROL

Dana Aeschliman

Statistical Genetics Research Group

Montreal Heart Institute Research Centre

ACATGGGGCC
$\mu+A_1X_1+D_1Z_1+$
$A_2X_2+D_2Z_2+I_{aa}$
$X_1X_2+I_{ad}X_1Z_2G$
GAAATTGCAT

Statistical Genetics Research Group

At the Montreal Heart Institute & Université de Montréal

Introduction

What’s a SNP?

Person A: …ACGG**G**TAG…

…ACGG**G**TAG…

Person B: …ACGGCTAG…

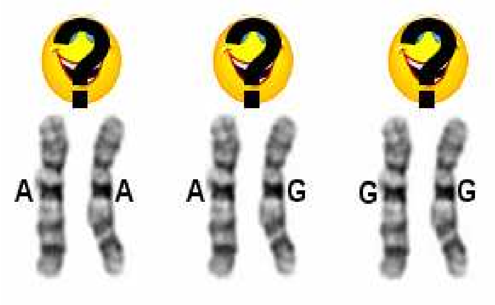
…ACGG**G**TAG…

Assessing differences in Single Nucleotide Polymorphisms (SNP) composition between Cases and Controls has become a very popular way of searching for genetic determinants of phenotypes.

Genomic studies with 1000’s of SNPs

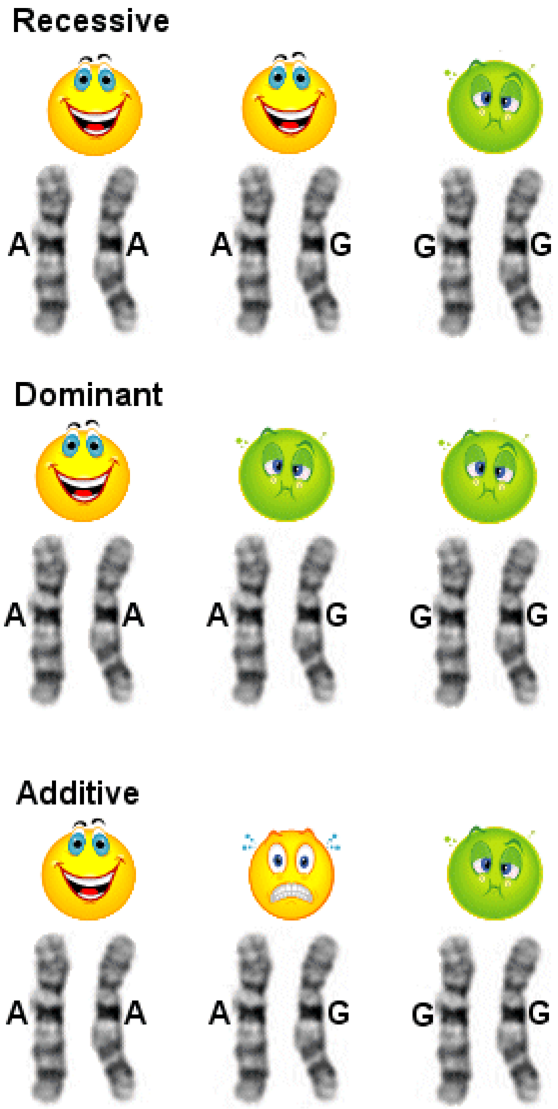
Have both **“across”** and **“down”** aspects of the **multiple testing** problem

The “across” aspect comes from **the inheritance models which need to be investigated**



The “down” aspect comes from the fact that **many thousands of SNPs may be tested**

Inheritance modeling



SAS tests :

- **Genotype test** : most powerful of the 3 for the recessive mode

	AA	AG	GG
Case			
Ctrl			

- **Allelic test** : most powerful of the 3 for dominant mode*

	A	G
Case		
Ctrl		

- **Armitage trend test**: most powerful of the 3 for additive mode

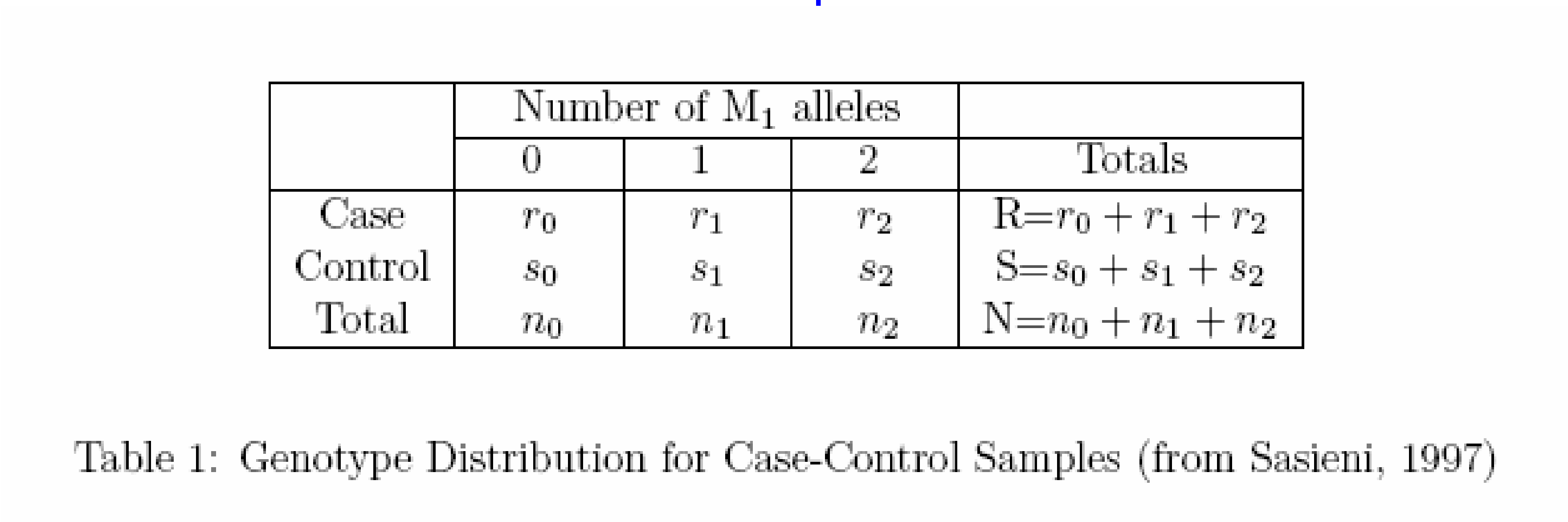
* Is not serological test (Sasieni,1997, **Biometrics**)

Outline of project

The MAX test explores recessive, additive, and dominant models while producing one P-value for each SNP.

The MAX-maxT algorithm. We treat the "down" aspect of the problem using **Westfall and Young's (1993) Algorithm 4.1**, which gives p-values corrected for the correlation between SNPs.

The data for a specific SNP



	Number of M_1 alleles			
	0	1	2	Totals
Case	r_0	r_1	r_2	$R = r_0 + r_1 + r_2$
Control	s_0	s_1	s_2	$S = s_0 + s_1 + s_2$
Total	n_0	n_1	n_2	$N = n_0 + n_1 + n_2$

Table 1: Genotype Distribution for Case-Control Samples (from Sasieni, 1997)

- R cases, $\boldsymbol{r}$=(r_0, r_1, r_2)

- S controls, $\boldsymbol{s}$=(s_0, s_1, s_2)

- Under H_0, $\boldsymbol{r}$ and $\boldsymbol{s}$ are multinomial with success probability vector **p**=(p_0, p_1, p_2) and thus $E(Sr_i\text{-}Rs_i)=0$, for any *i*.

The MAX test (Freidlin and Zheng 2002, Zheng and Gastwirth 2006) builds on the ideas of Armitage (1955), Sasieni (1997), and Slager and Schaid (2001).

Armitage's trend test statistic :

$$T(x_1) = \frac{\sum_{i=0}^{2} x_i(Sr_i - Rs_i)}{\sqrt{\mathrm{var}\left[\sum_{i=0}^{2} x_i(Sr_i - Rs_i)\right]}} \sim \mathrm{N}(0,1).$$

- Recessive coding: $x_0=0$, $x_1=0$, $x_2=1$.
- Additive coding: $x_0=0$, $x_1=0.5$, $x_2=1$.
- Dominant coding: $x_0=0$, $x_1=1$, $x_2=1$. *

* Is serological test (Sasieni,1997, **Biometrics**)

The MAX test makes use of the multivariate normal (MVN) vector ($T(0)$, $T(0.5)$, $T(1)$).

Under H_0, the correlation matrix of this MVNRV is known. Thus, a null sample can be drawn.

The MAX test compares

$$T_{\mathrm{MAX}}=\mathrm{MAX}(|T(0)|, |T(0.5)|, |T(1)|)$$

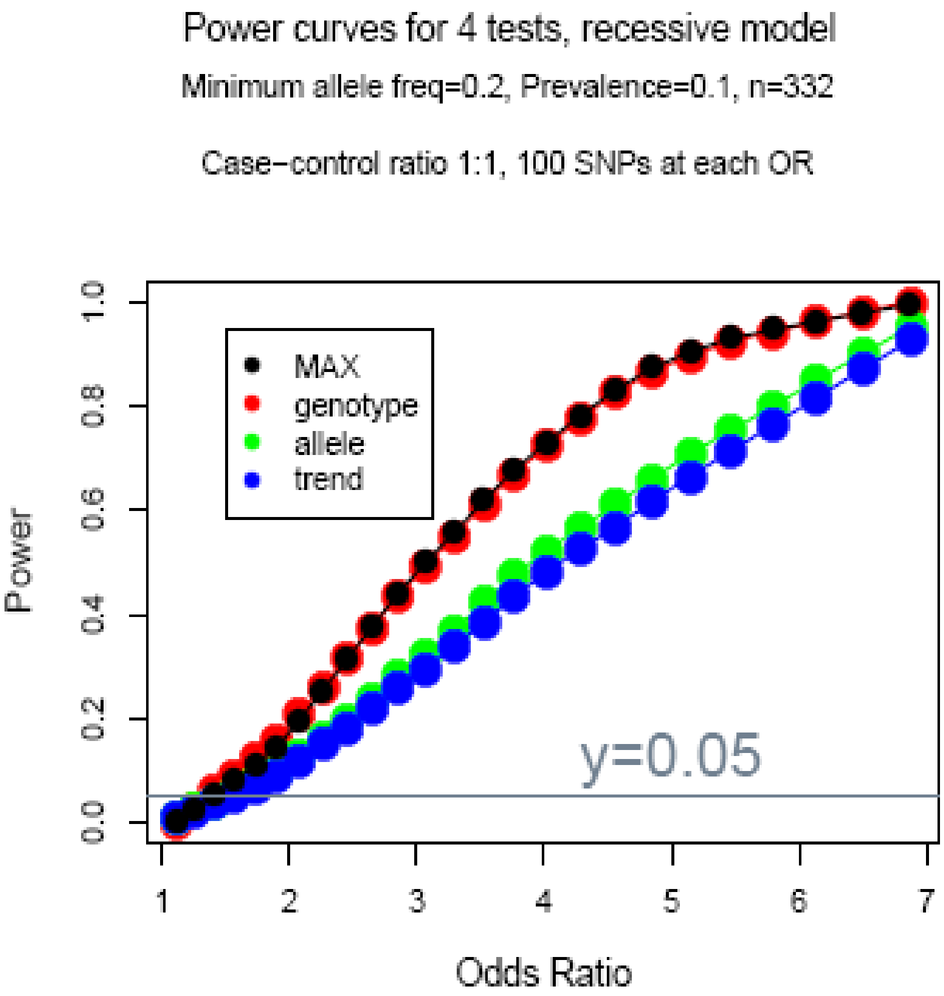
to the RVs from the null sample.

We use a small sample correction and allow the accuracy of the estimation of the P-value to depend on the magnitude of the P-value.

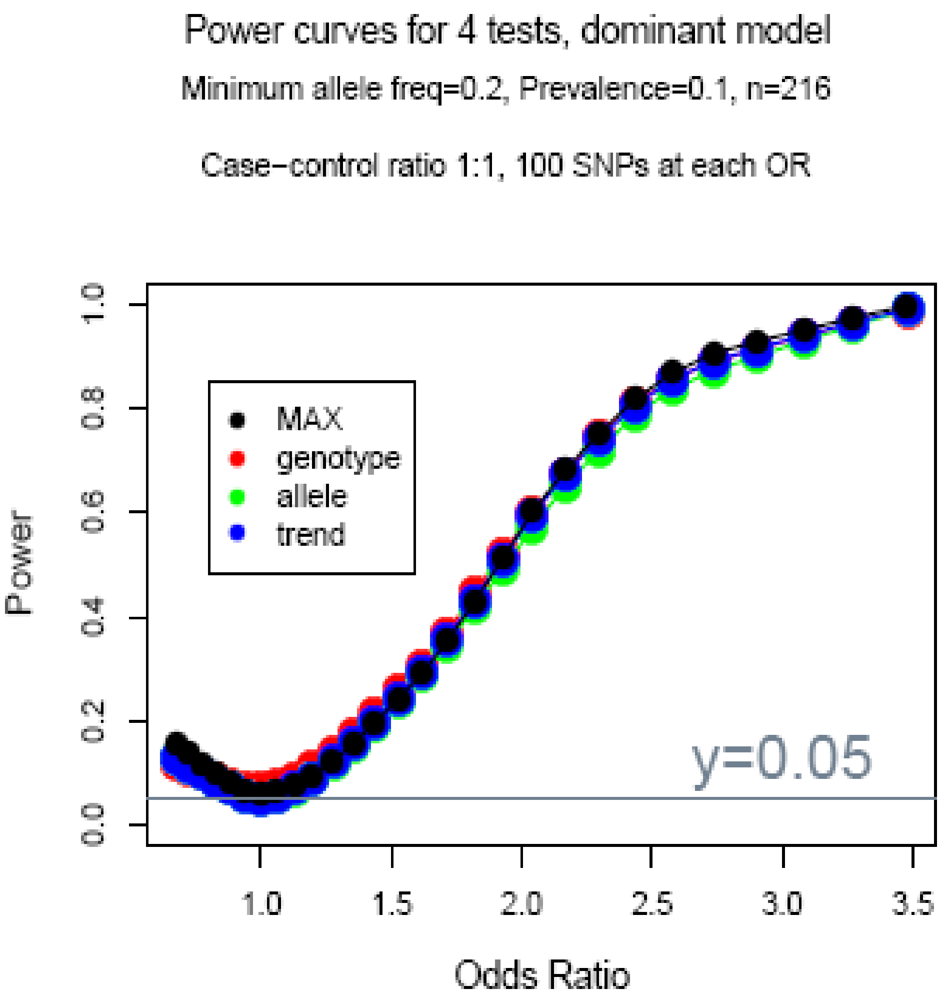
Results of comparing our implementation of the MAX test to SAS PROC CASECONTROL

Sample sizes and odds ratios suggested by Zheng and Gastwirth (2006) Table II.

Recessive model



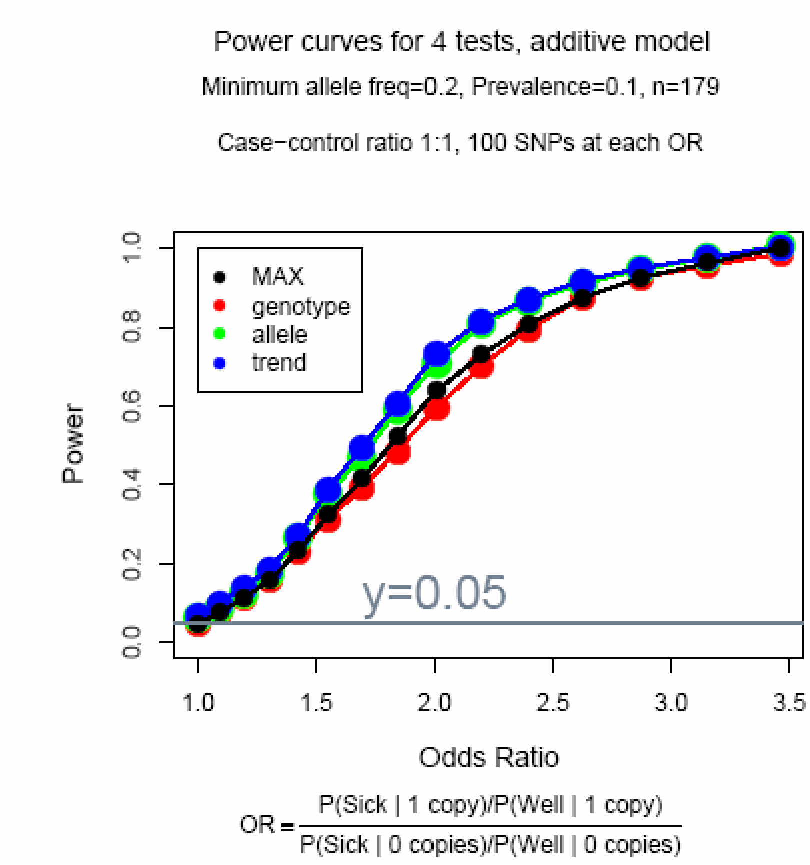
Dominant model

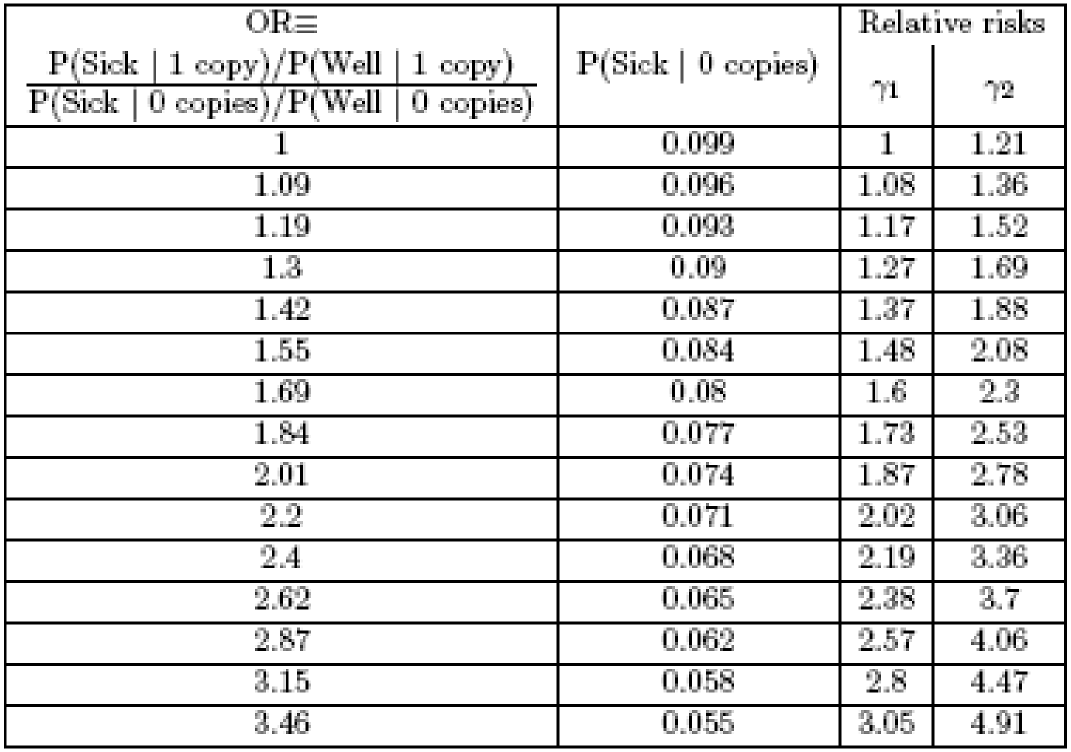


Results of comparing our implementation the MAX test to SAS PROC CASECONTROL

Additive model

Sample sizes and relative risks suggested by Zheng and Gastwirth (2006) Table II.





OR≡ $\frac{P(\text{Sick} \mid 1\text{ copy})/P(\text{Well} \mid 1\text{ copy})}{P(\text{Sick} \mid 0\text{ copies})/P(\text{Well} \mid 0\text{ copies})}$	P(Sick | 0 copies)	Relative risks	
		γ_1	γ_2
1	0.099	1	1.21
1.09	0.096	1.08	1.36
1.19	0.093	1.17	1.52
1.3	0.09	1.27	1.69
1.42	0.087	1.37	1.88
1.55	0.084	1.48	2.08
1.69	0.08	1.6	2.3
1.84	0.077	1.73	2.53
2.01	0.074	1.87	2.78
2.2	0.071	2.02	3.06
2.4	0.068	2.19	3.36
2.62	0.065	2.38	3.7
2.87	0.062	2.57	4.06
3.15	0.058	2.8	4.47
3.46	0.055	3.05	4.91

Table 2: Parameters of various additive models

We propose to correct our P-values for the multitude of tests by using Algorithm 4.1 of Westfall and Young (1993, p.66).

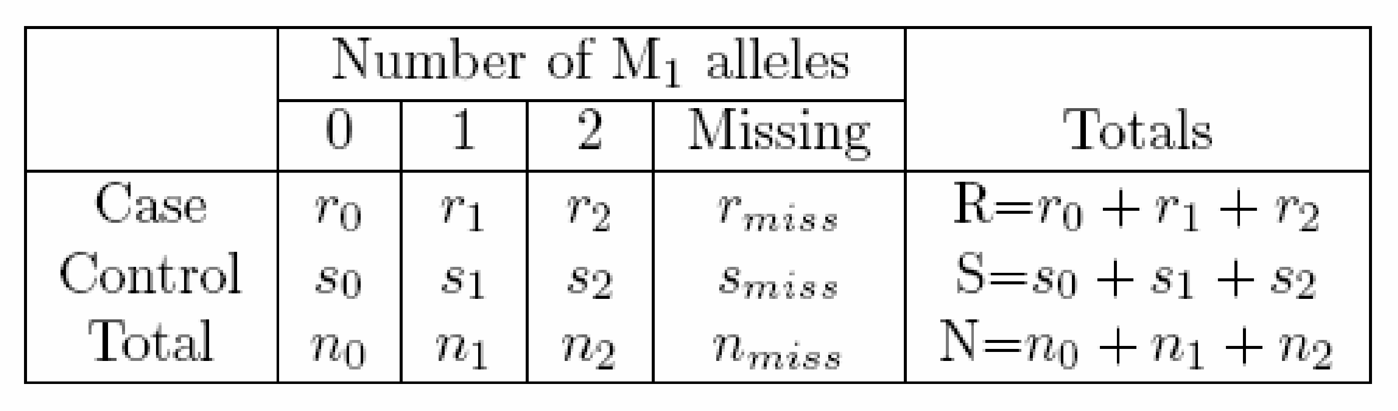
This is the maxT algorithm (Ge et al. TEST, 2003).

Under H_0 and HWE, the correlation between components of the MAX test depends only on the Minimum Allele Frequency (MAF).

Are MAX stats from MVN’s with different correlation structures comparable?

In implementing Algorithm 4.1, we use permutation resampling of a data matrix where each row represents a subject.

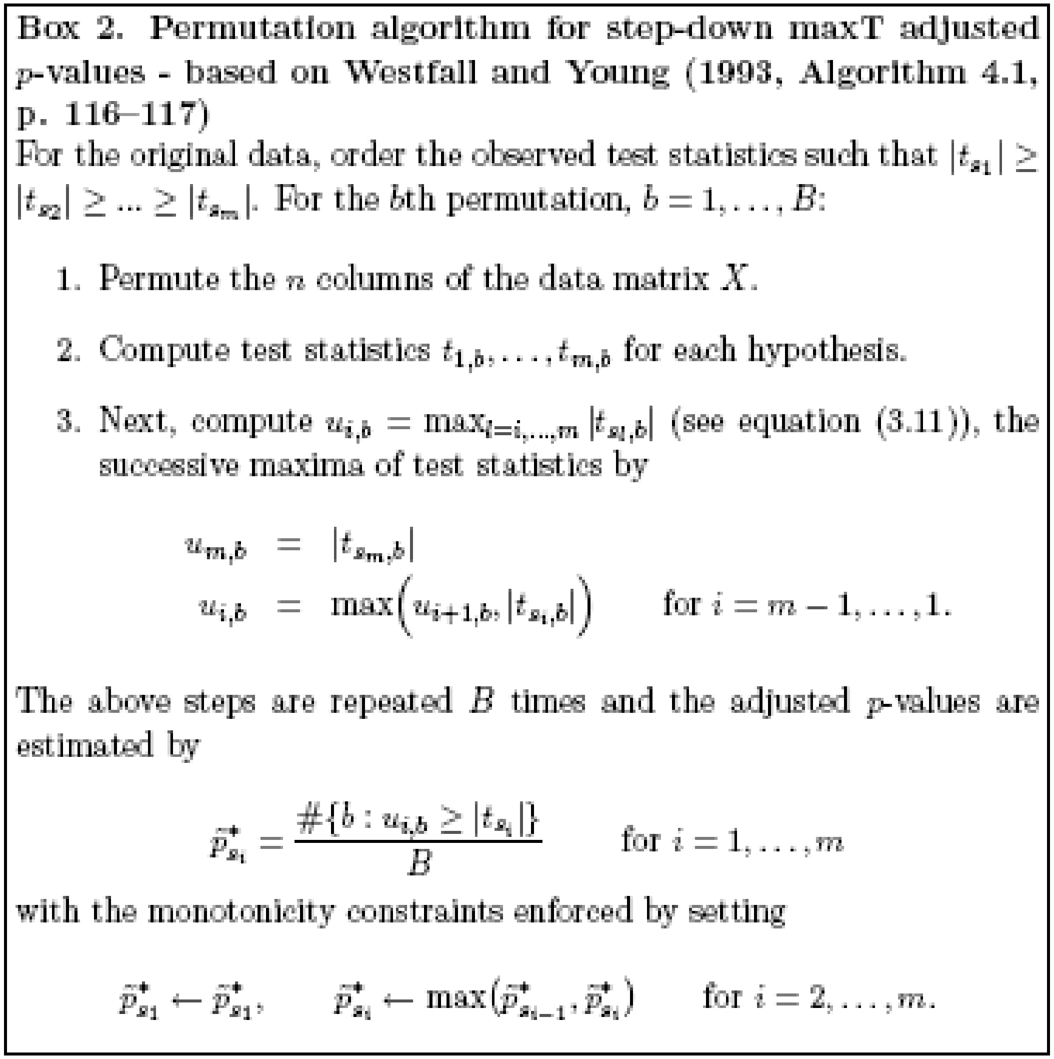
The data for a specific SNP under permutation resampling with missing data



	Number of M_1 alleles				Totals
	0	1	2	Missing	
Case	r_0	r_1	r_2	r_{miss}	R=$r_0+r_1+r_2$
Control	s_0	s_1	s_2	s_{miss}	S=$s_0+s_1+s_2$
Total	n_0	n_1	n_2	n_{miss}	N=$n_0+n_1+n_2$

When data is missing, R and S may change when the Case/Ctrl Status vector is resampled.

Once the MAX stats have been produced, the remaining portion of Algorithm 4.1 can be done in only 74 lines in SAS!



Box 2. Permutation algorithm for step-down maxT adjusted p-values - based on Westfall and Young (1993, Algorithm 4.1, p. 116–117)

For the original data, order the observed test statistics such that $|t_{s_1}| \geq |t_{s2}| \geq ... \geq |t_{s_m}|$. For the bth permutation, $b = 1, \ldots, B$:

1. Permute the n columns of the data matrix X.
2. Compute test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis.
3. Next, compute $u_{i,b} = \max_{l=i,\ldots,m} |t_{s_l,b}|$ (see equation (3.11)), the successive maxima of test statistics by

$$\begin{aligned} u_{m,b} &= |t_{s_m,b}| \\ u_{i,b} &= \max\Big(u_{i+1,b}, |t_{s_i,b}|\Big) \qquad \text{for } i = m-1, \ldots, 1. \end{aligned}$$

The above steps are repeated B times and the adjusted p-values are estimated by

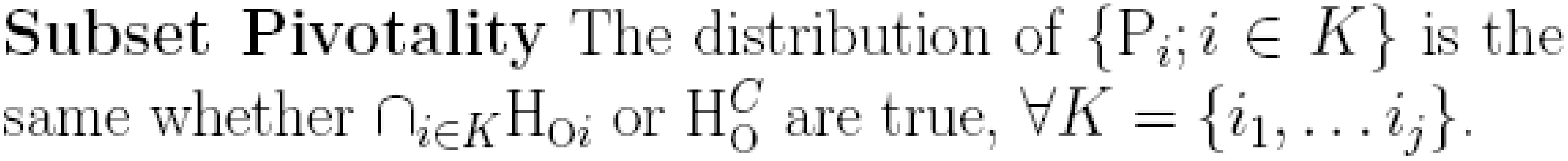
$$\tilde{p}^*_{s_i} = \frac{\#\{b : u_{i,b} \geq |t_{s_i}|\}}{B} \qquad \text{for } i = 1, \ldots, m$$

with the monotonicity constraints enforced by setting

$$\tilde{p}^*_{s_1} \leftarrow \tilde{p}^*_{s_1}, \qquad \tilde{p}^*_{s_i} \leftarrow \max(\tilde{p}^*_{s_{i-1}}, \tilde{p}^*_{s_i}) \qquad \text{for } i = 2, \ldots, m.$$

From Ge et al., **Test** (2003), Vol. 12, No. 1, pp. 1-77

Can we claim that we are strongly controlling the FWER? Does subset pivotality hold?



Subset Pivotality The distribution of $\{\mathrm{P}_i; i \in K\}$ is the same whether $\cap_{i \in K} \mathrm{H}_{\mathrm{O}i}$ or $\mathrm{H}_{\mathrm{O}}^{C}$ are true, $\forall K = \{i_1, \ldots i_j\}$.

We think subset pivotality holds as F_o(MAX_i, MAX_j) shouldn’t be affected by F_o(MAX_k), $k \neq i$ and $k \neq j$.

Computational speed

The speed of our implementation of the MAX test is fine for an impatient user.

The performance of our MAX-maxT algorithm is built to allow flexibility and parallelization.

In 2 different ways, we did 10,000 permutations of the Affection Status vector and calculated a new set of MAX statistics for each permutation for a data set of 1,400 subjects with 503 SNPs. By splitting up the job of generating null MAX stats onto 4 processes, we saw a time decrease of about 1/2.

Effect of LD on P-values

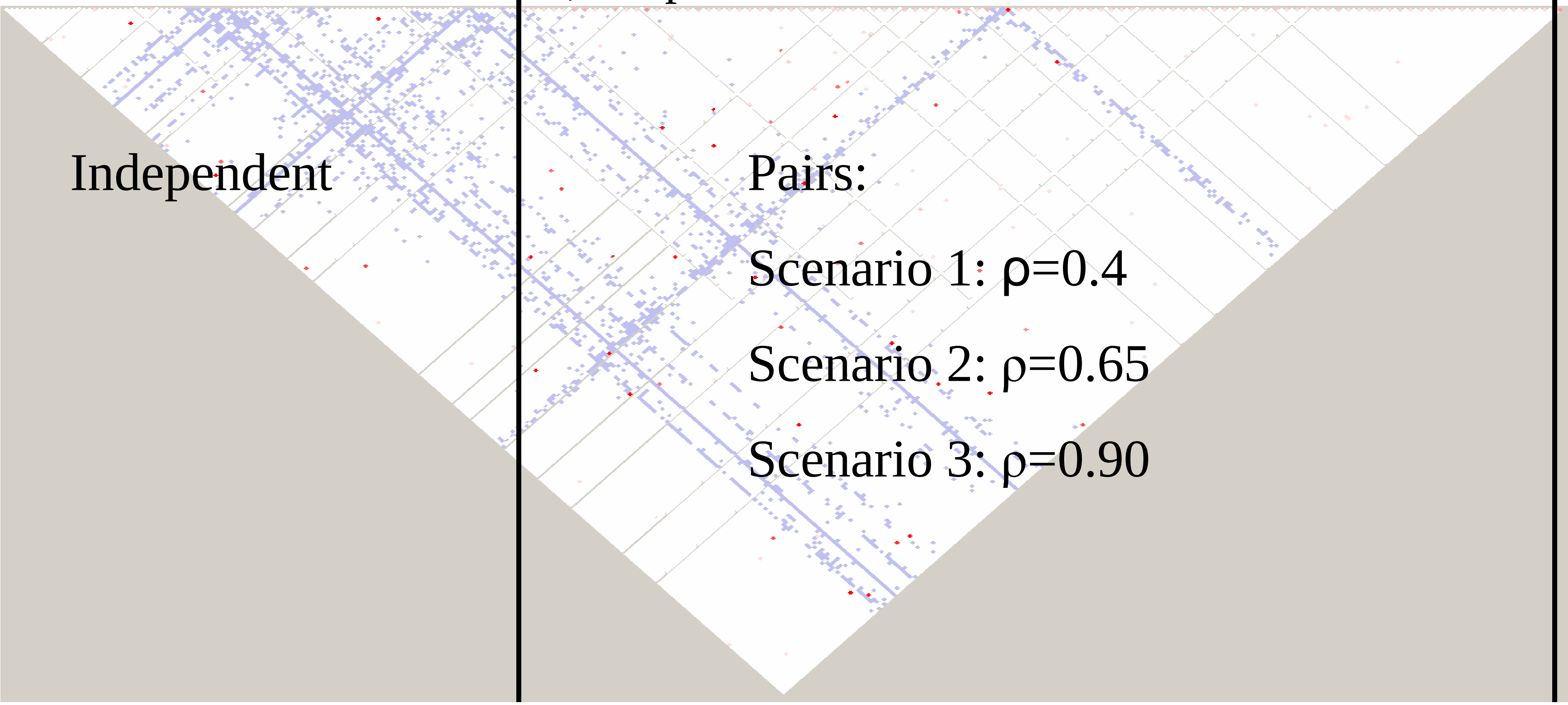
We test our implementation of Algorithm 4.1 with a small experiment.

300 data sets, 100 per scenario.

Each data set has 303 SNPs.

1,250 permutations/set.

Disease SNPs



Independent

Pairs:

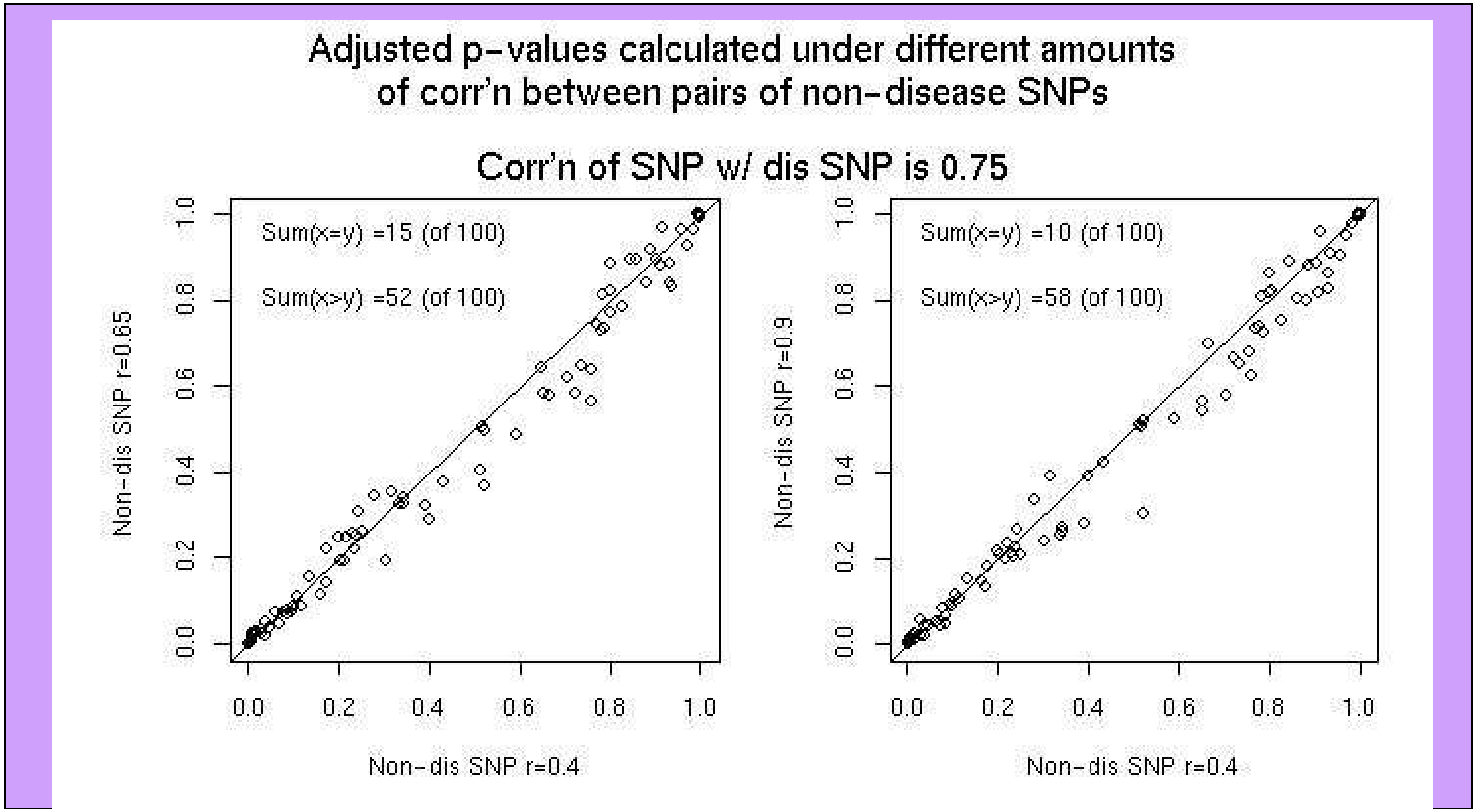
Scenario 1: ρ=0.4

Scenario 2: ρ=0.65

Scenario 3: ρ=0.90

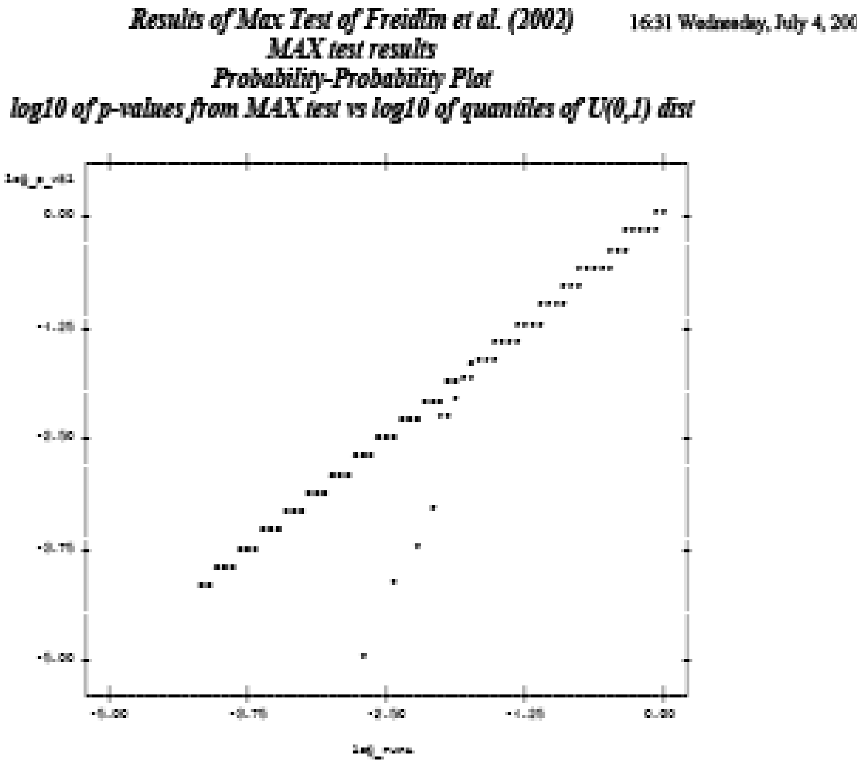
Results of effect of LD on maxT

Corr’n between pairs	Family wise error rate	True negative rate (per SNP)	True positive rate (per SNP)		
			Corr'n w/ disease SNP		
			0.75	**0.85**	**0.95**
0.4	0.07	0.9998	0.17	0.27	0.53
0.65	0.02	0.9999	0.18	0.26	0.55
0.9	0.05	0.9998	0.18	0.27	0.54



Discussion

- Caution in the presence of population structure (PS).
- With PS, we can have that the magnitude of the MAX statistic is correlated with the MAF.
- We include 2 graphics in our software that can help.

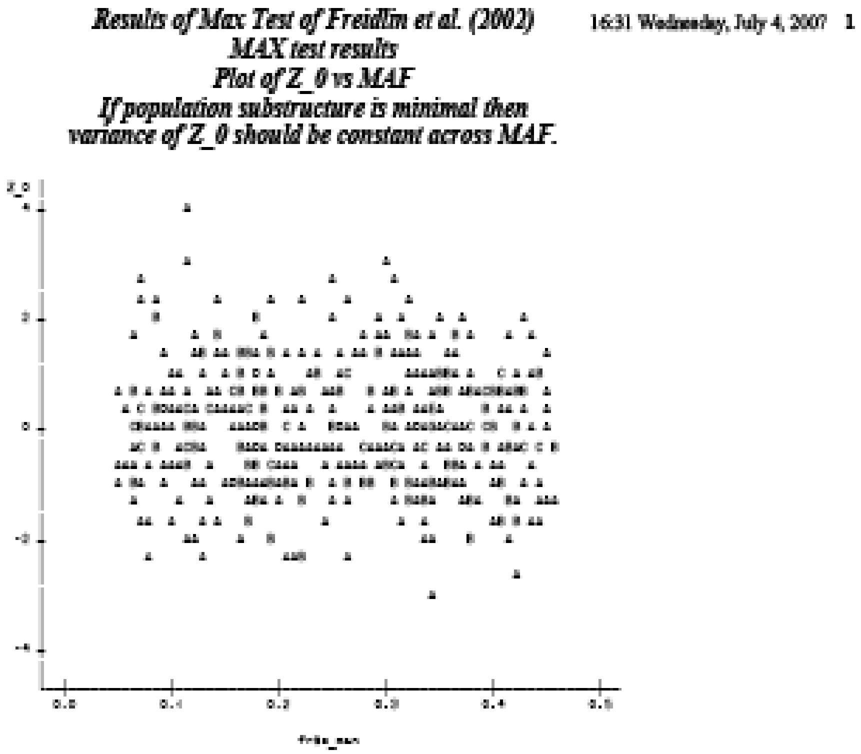


Results of Max Test of Freidlin et al. (2002) 16:31 Wednesday, July 4, 200

MAX test results

Probability-Probability Plot

log10 of p-values from MAX test vs log10 of quantiles of U(0,1) dist



Results of Max Test of Freidlin et al. (2002) 16:31 Wednesday, July 4, 2007 1

MAX test results

Plot of Z_0 vs MAF

If population substructure is minimal then variance of Z_0 should be constant across MAF.

Summary: A tool for SNP-phenotype association studies

- The MAX test investigates 3 inheritance models while yielding 1 p-value. We encode it in a SAS MACRO.
- The MAX-maxT test can be useful for producing corrected P-values which take into account the correlation structure of the data. We encode it in a SAS MACRO.

Acknowledgements

- Marie-Pierre Dubé
- Sylvie Provost, Amina Barhdadi

Thank you!

http://www.statgen.org

ACATGGGGCC
$\mu+A_1X_1+D_1Z_1+$
$A_2X_2+D_2Z_2+I_{aa}$
$X_1X_2+I_{ad}X_1Z_2G$
GAAATTGCAT

Statistical Genetics Research Group

At the Montreal Heart Institute & Université de Montréal

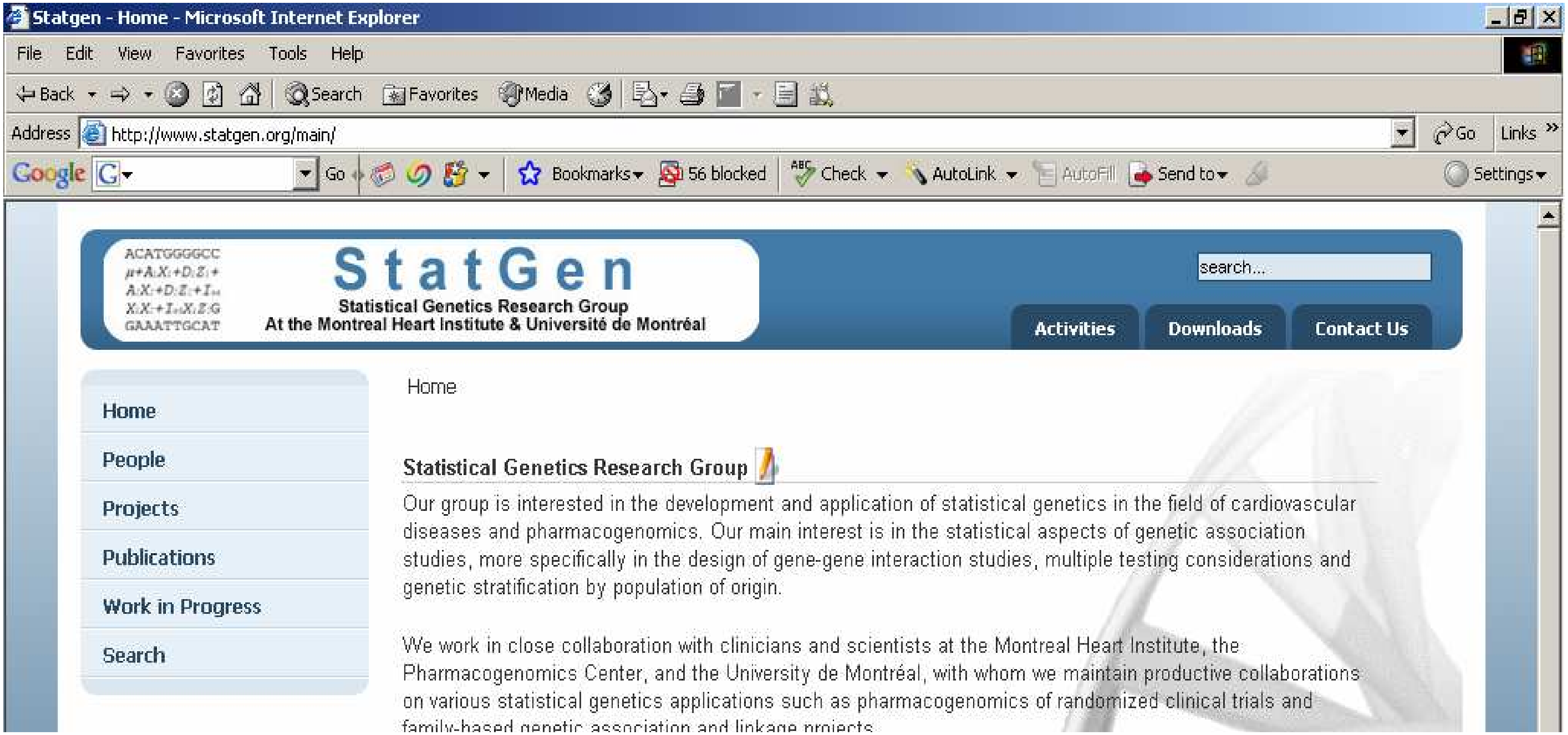
INSTITUT DE
CARDIOLOGIE
DE MONTRÉAL

Université
de Montréal

Both of our MACRO suites (MAX test and MAX-maxT) are available at:

http://www.statgen.org/

(Click on “Downloads”)


Statgen - Home - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.statgen.org/main/
Go Links
Google Go Bookmarks 56 blocked Check AutoLink AutoFill Send to Settings
ACATGGGGCC
GAAATTGCAT
StatGen
Statistical Genetics Research Group
At the Montreal Heart Institute & Université de Montréal
search...
Activities
Downloads
Contact Us
Home
People
Projects
Publications
Work in Progress
Search
Home

Statistical Genetics Research Group

Our group is interested in the development and application of statistical genetics in the field of cardiovascular diseases and pharmacogenomics. Our main interest is in the statistical aspects of genetic association studies, more specifically in the design of gene-gene interaction studies, multiple testing considerations and genetic stratification by population of origin.

We work in close collaboration with clinicians and scientists at the Montreal Heart Institute, the Pharmacogenomics Center, and the University de Montréal, with whom we maintain productive collaborations on various statistical genetics applications such as pharmacogenomics of randomized clinical trials and family-based genetic association and linkage projects.

MAX Test Parametric Bootstrap

One problem was finding the eigenvalues of the correlation matrices, solving a cubic, in order to do our parametric bootstrap. We solve this using some formulas from http://mathworld.wolfram.com/CubicFormula.html.

$$\begin{vmatrix} 1-\lambda & \rho_{0,0.5} & \rho_{0,1} \\ \rho_{0,0.5} & 1-\lambda & \rho_{0.5,1} \\ \rho_{0,1} & \rho_{0.5,1} & 1-\lambda \end{vmatrix} = 0.$$

$$(1-\lambda)^3 - (\rho^2_{0,0.5}\rho^2_{0,1}\rho^2_{0,0.5})(1-\lambda) = 2(\rho_{0,0.5}\rho_{0,1}\rho_{0,0.5}).$$

$$\begin{aligned} Q &= -(\rho^2_{0,0.5}\rho^2_{0,1}\rho^2_{0,0.5})/3, \\ R &= -(\rho_{0,0.5}\rho_{0,1}\rho_{0,0.5}), \\ \theta &= cos^{-1}\left(R/\sqrt{-Q^3}\right) \end{aligned}$$

$$\begin{aligned} \lambda_1 &= 1 - 2\sqrt{-Q}\cos\left(\frac{\theta}{3}\right) \\ \lambda_2 &= 1 - 2\sqrt{-Q}\cos\left(\frac{\theta+2\pi}{3}\right) \\ \lambda_3 &= 1 - 2\sqrt{-Q}\cos\left(\frac{\theta+4\pi}{3}\right) \end{aligned}$$

References

Zheng G, Gastwirth J. (2006) On estimation of the variance in Cochran–Armitage trend tests for genetic association using case–control studies. *Statistics in Medicine* **25,** 3150-3159.

Westfall, P.H. and Young, S.S. (1993) *Resampling-Based Multiple Testing*. John Wiley and Sons.

Whittemore A (2006) Population Structure in Genetic Association Studies. *Proceeding of the Joint Statistical Meeting 2006*

Computational speed

Using SAS 9.1 in a Gentoo Linux 2.6.18 operating system on a server with 4 core 2 duo processors, each a 64-bit Intel Xeon 3.00 GHz, we naively did 10,000 permutations of the Affection Status vector and calculated a new set of MAX statistics for each permutation for a data set of 1,400 patients with 503 SNPs. This took 13 hrs 26 mins.

On the same system, we did split "prepare" into 4 different processes and did the same job. This took 6 hrs 52 mins, including the time for the extra programming.

Why do we use permutation resampling rather than bootstrapping?

The variances of the components of the MAX test depend on the ratio of R to S:

$$\begin{aligned} U_{\text{dom}} &= S(R - r_0) - R(S - s_0) \\ \text{var}(U_{\text{dom}}) &= S^2 R p_0(1 - p_0) + R^2 S p_0(1 - p_0) \quad (\text{Under H}_o) \\ \text{var}(U_{\text{dom}}) &= RSNp_0(1 - p_0) \end{aligned}$$

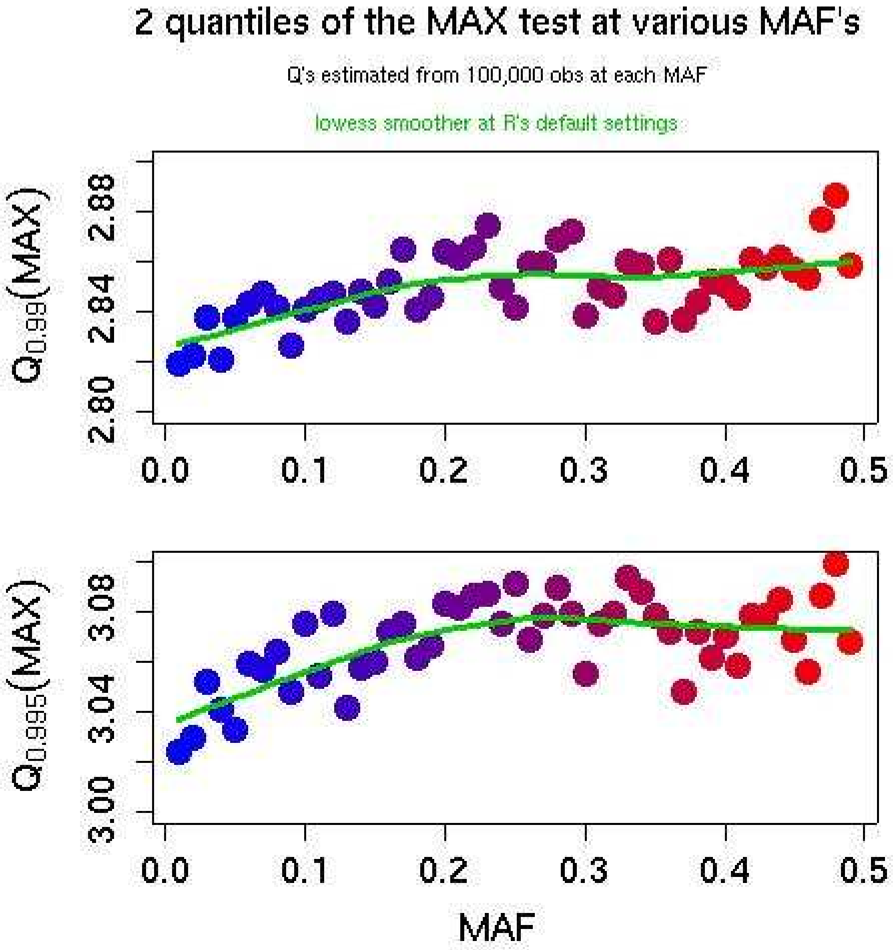
When there’s no missing data (in the entire data set), we don’t have to recompute the variances prior to scaling the components.

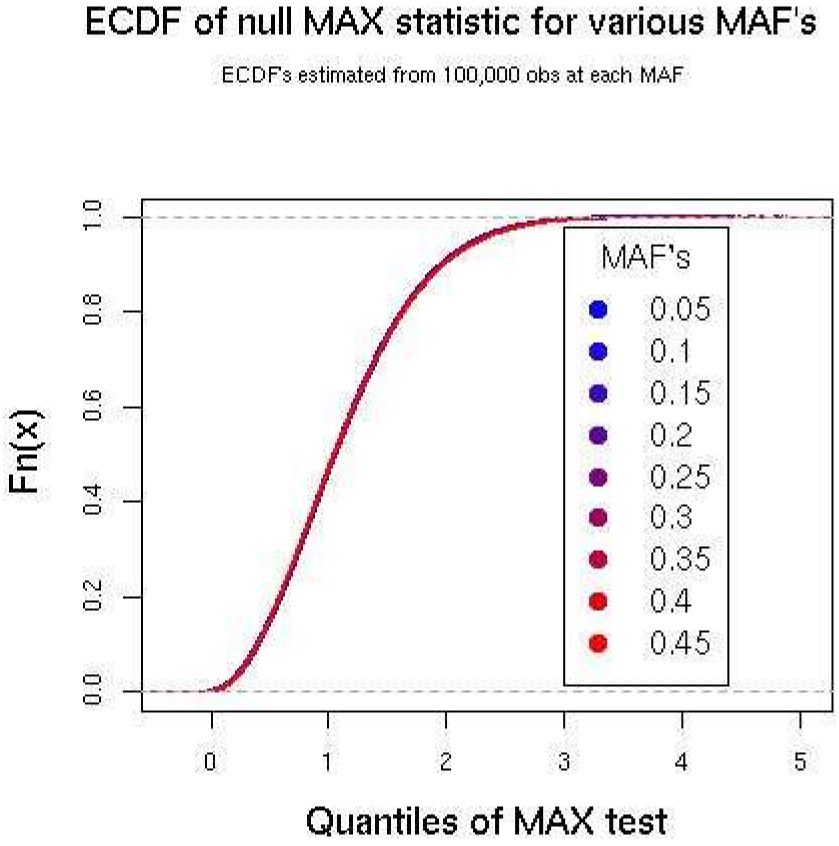
As well, the resampling code is simpler.

But the convergence to $\mathbf{G}_0$ would probably be faster under bootstrapping!

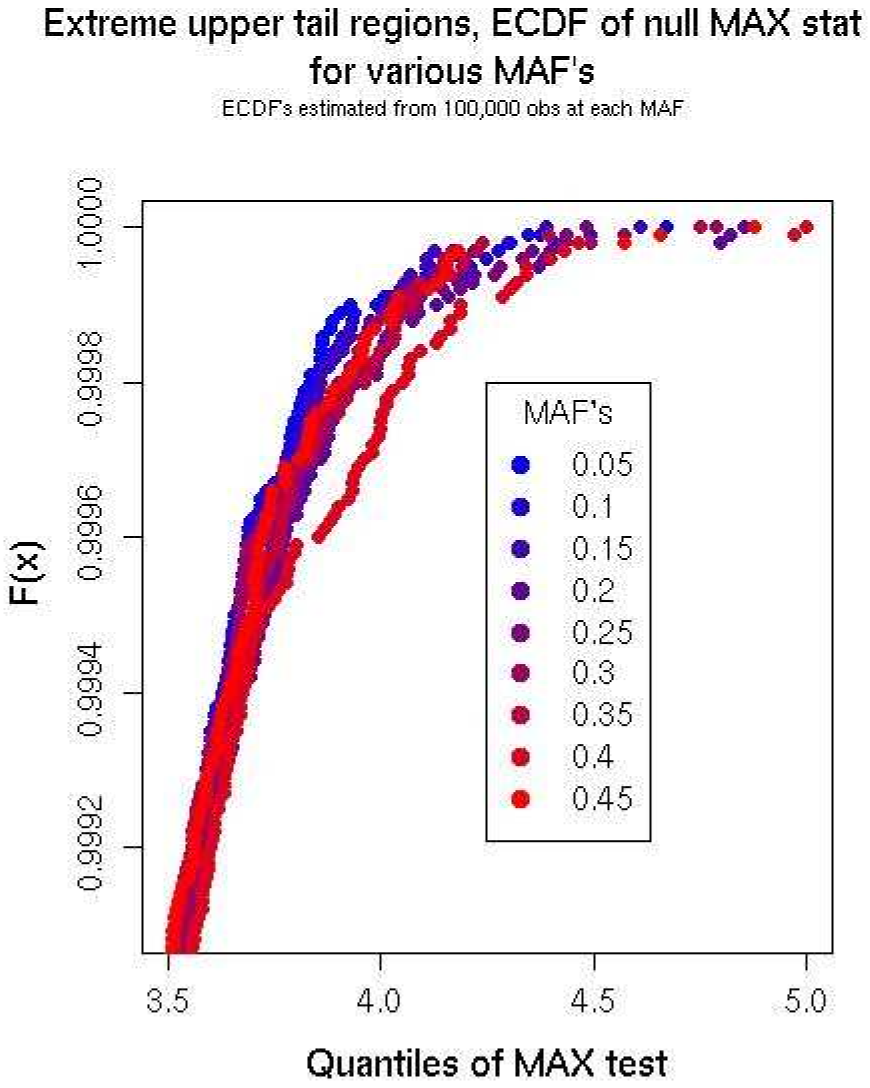
An area for further thought…

Distribution of the MAX Test statistic under various MAF’s





Distribution of the MAX Test statistic under various MAF’s



Effect of LD on P-values
(Another check of our implementation)

We produce 100 data sets of 4 SNPs with a disease vector. In each of these sets, we have 332 subjects, the disease prevalence is 0.1, the MAF of the disease SNP is 0.2, the mode of inheritance is recessive, the relative risk for having zero or one copy of the disease allele vs having 2 copies is about 4.54 so the corresponding OR is about 6.87. The 3 SNPs that are related to the disease only through the fact of being correlated with the disease allele are correlated with the disease allele at 0.75, 0.85 or 0.95. After the data set is generated we discard the SNP that is directly related to the disease.

We then produce 300 additional data sets, 100 under each of three different scenarios. In each scenario, we have 100 SNPs where each SNP is correlated with its neighbor at 0.05. These initial 100 are followed by 200 additional SNPs which are grouped into 100 distinct pairs. Thus, each additional data set at this stage has 300 SNPs. In the first scenario, the correlation within a pair is 0.4. In the second and third scenarios the within-pair correlation is respectively 0.65 and 0.9. In each of the 3 scenarios the first SNP in a pair is correlated with the second SNP of the previous pair at 0.05.

We then use each of the 100 data sets of 3 SNPs which are indirectly related to a disease vector three times, inserting it into one of the data sets from each scenario. In this way we finish with 300 data sets, each data set having 303 SNPs.