

# Detecting Epistatic Interactions Contributing to a Quantitative Trait: The Restricted Partition Method

Rob Culverhouse, PhD

Washington University in St. Louis, School of Medicine

May 28, 2004

# Single locus analog for our analyses: Measured Genotype

Quantitative trait analysis using unrelated individuals

- No notion of “affected” without placing a threshold
- For loci in linkage disequilibrium with trait locus, expect genotypes to have different mean trait values

	AA	Aa	aa
mean(trait)	34.5	12.2	41.5

# Epistasis

Genes interacting in a non-additive way

# Epistasis

Genes interacting in a non-additive way


## Examples:

- Triglyceride level (Nelson et al. 2001)
- Alzheimer disease (Zubenko et al. 2001)
- Breast cancer (Ritchie et al. 2001)

# Epistasis

Genes interacting in a non-additive way

## Examples:

- Triglyceride level (Nelson et al. 2001)
- Alzheimer disease (Zubenko et al. 2001)
- Breast cancer (Ritchie et al. 2001)
-  Drug effects (response and toxicity)

# Epistasis

Genes interacting in a non-additive way

Some possible consequences:

- Which is the “bad” allele may depend on genetic background or environmental exposure



# Epistasis

Genes interacting in a non-additive way

Some possible consequences:

- Which is the “bad” allele may depend on genetic background or environmental exposure
- “Importance” of a locus depends on allele freq.

# “Importance” of a locus depends on allele freq

## Fixed genetic model for TSC

	ApoE alleles			LDLR alleles	
	$p(\epsilon 2)$	$p(\epsilon 3)$	$p(\epsilon 4)$	$p(A_1)$	$p(A_2)$
Population 1	<b>0.08</b>	<b>0.77</b>	<b>0.15</b>	<b>0.22</b>	<b>0.78</b>
Population 2	<b>0.02</b>	<b>0.03</b>	<b>0.95</b>	<b>0.50</b>	<b>0.50</b>

# “Importance” of a locus depends on allele freq

## Fixed genetic model for TSC

	ApoE alleles			LDLR alleles	
	$p(\epsilon 2)$	$p(\epsilon 3)$	$p(\epsilon 4)$	$p(A_1)$	$p(A_2)$
Population 1	0.08	<b>0.77</b>	0.15	0.22	<b>0.78</b>
Population 2	0.02	0.03	<b>0.95</b>	0.50	<b>0.50</b>

# “Importance” of a locus depends on allele freq

## Fixed genetic model for TSC

	ApoE alleles			LDLR alleles	
	p( $\epsilon$ 2)	p( $\epsilon$ 3)	p( $\epsilon$ 4)	p(A <sub>1</sub> )	p(A <sub>2</sub> )
Population 1	0.08	0.77	0.15	0.22	0.78
Population 2	0.02	0.03	0.95	0.50	0.50

## % Variance explained

	ApoE	LDLR	ApoE x LDLR	total
Population 1	<b>41.0</b>	2.9	8.9	52.8
Population 2	3.7	<b>25.3</b>	2.0	31.1

# Epistasis

Genes interacting in a non-additive way

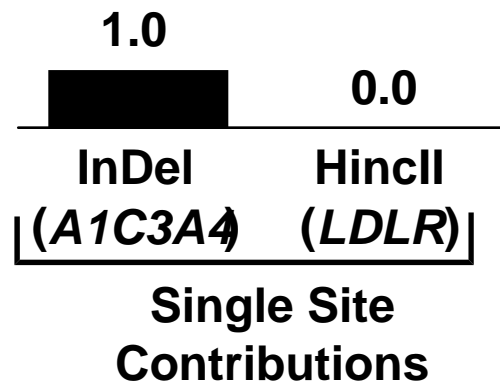
Some possible consequences:

- Which is the “bad” allele may depend on genetic background or environmental exposure
- “Importance” of a locus depends on allele freq.
- Contributing loci may only be noticed in a multilocus analysis

# Variability in Ln(Triglyceride) explained by Single locus vs Two locus analyses

Males, N =188

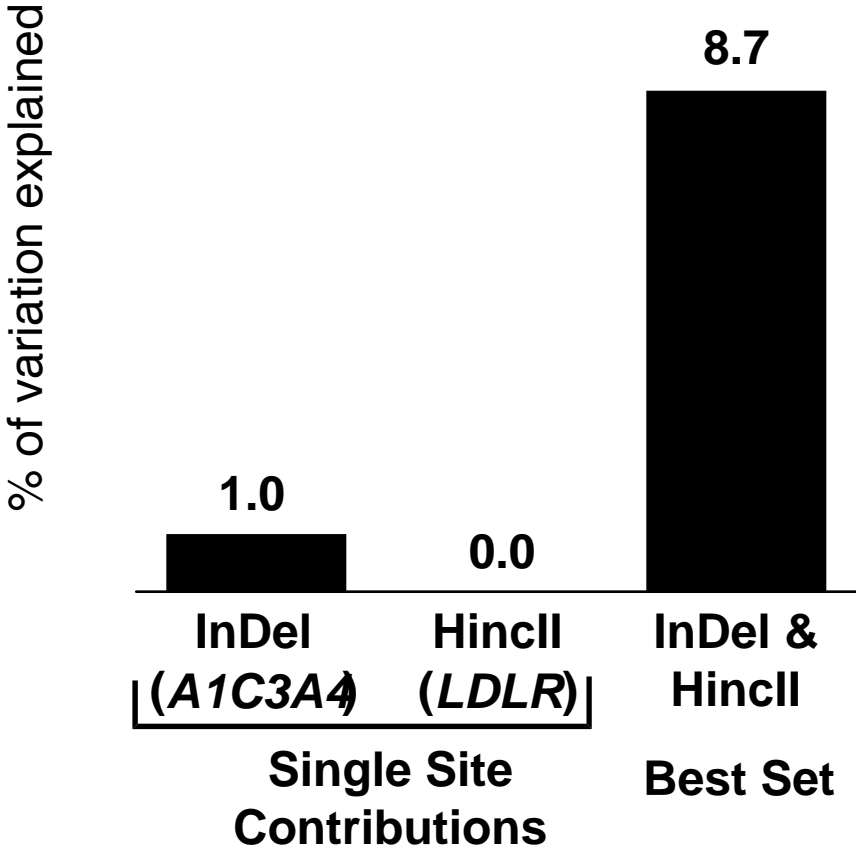
% of variation explained



(Nelson et al 2001)

# Variability in Ln(Triglyceride) explained by Single locus vs Two locus analyses

Males, N = 188



(Nelson et al 2001)

# Two Locus Epistatic Model

(a qualitative trait example)

	BB	Bb	bb	
AA	?	?	?	0.5
Aa	?	?	?	0.5
aa	?	?	?	0.5
	0.5	0.5	0.5	

$$p(A)=p(B)=0.5$$

Cell entries indicate probability of having disease

Analyzing these loci separately would give the impression that neither one contributes to the phenotype

# Two Locus Epistatic Model

(a qualitative trait example)

	<b>BB</b>	<b>Bb</b>	<b>bb</b>	
<b>AA</b>	?	?	?	<b>0.5</b>
<b>Aa</b>	?	?	?	<b>0.5</b>
<b>aa</b>	?	?	?	<b>0.5</b>
	<b>0.5</b>	<b>0.5</b>	<b>0.5</b>	

$$p(A)=p(B)=0.5$$

Cell entries indicate probability  
of having disease

Analyzing these loci separately would give the impression that  
neither one contributes to the phenotype

# Two Locus Epistatic Model

(a qualitative trait example)

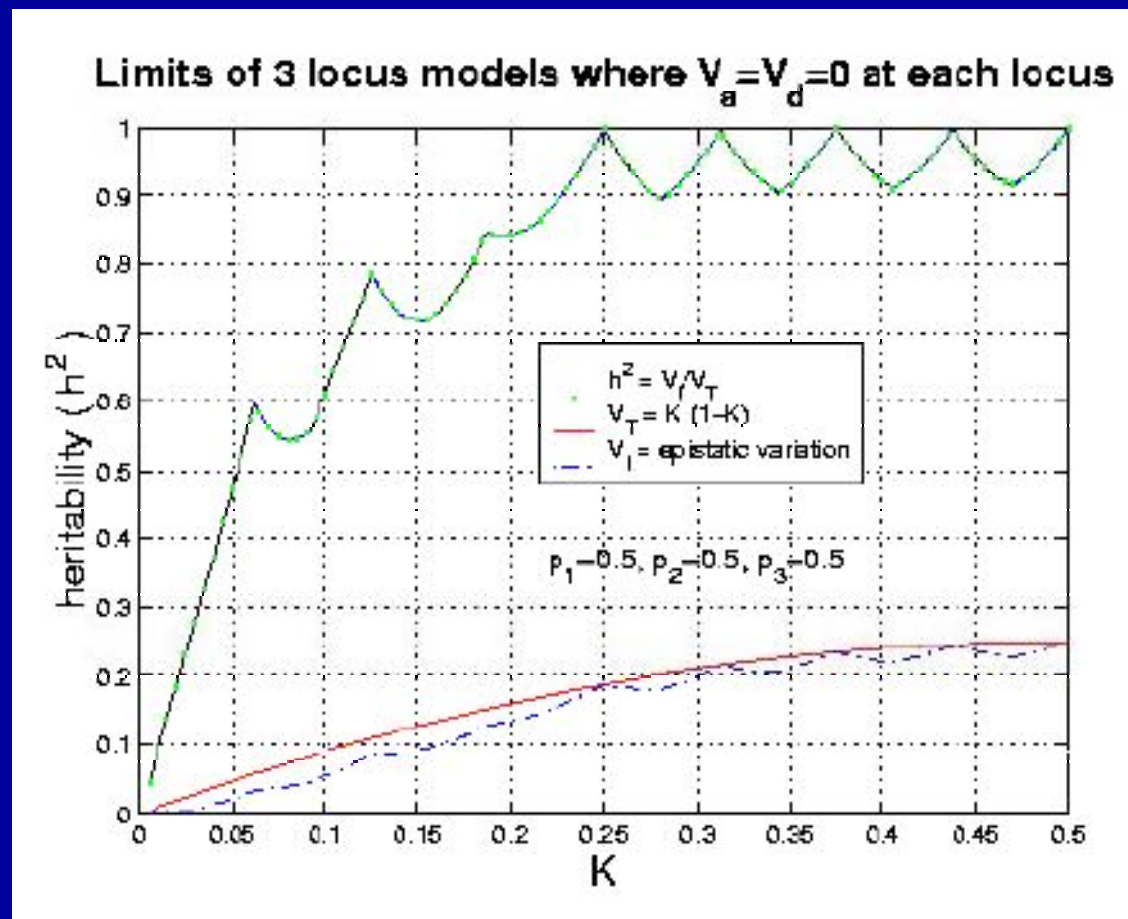
	BB	Bb	bb	
AA	1	0	1	0.5
Aa	0	1	0	0.5
aa	1	0	1	0.5
	0.5	0.5	0.5	

$$p(A)=p(B)=0.5$$

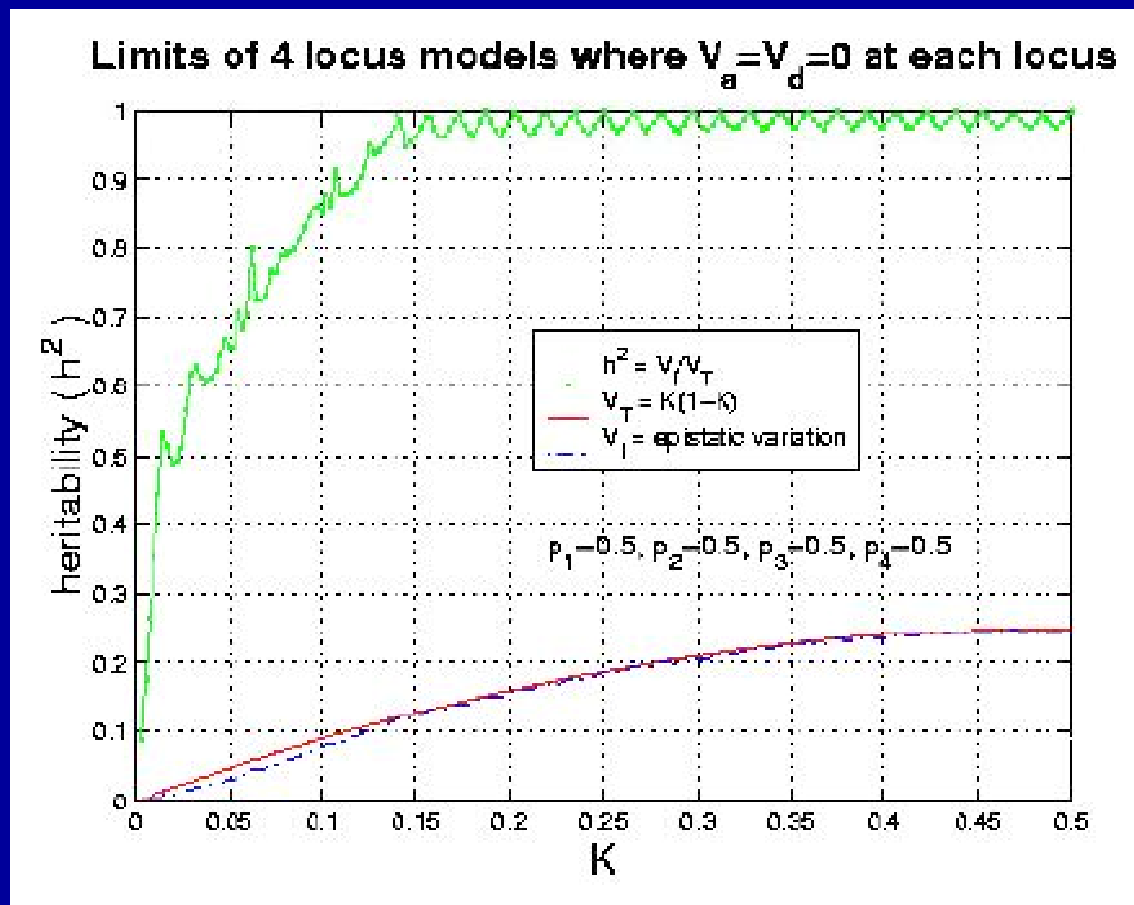
Cell entries indicate probability of having disease

In fact,  
the trait is completely determined by the 2-locus genotype

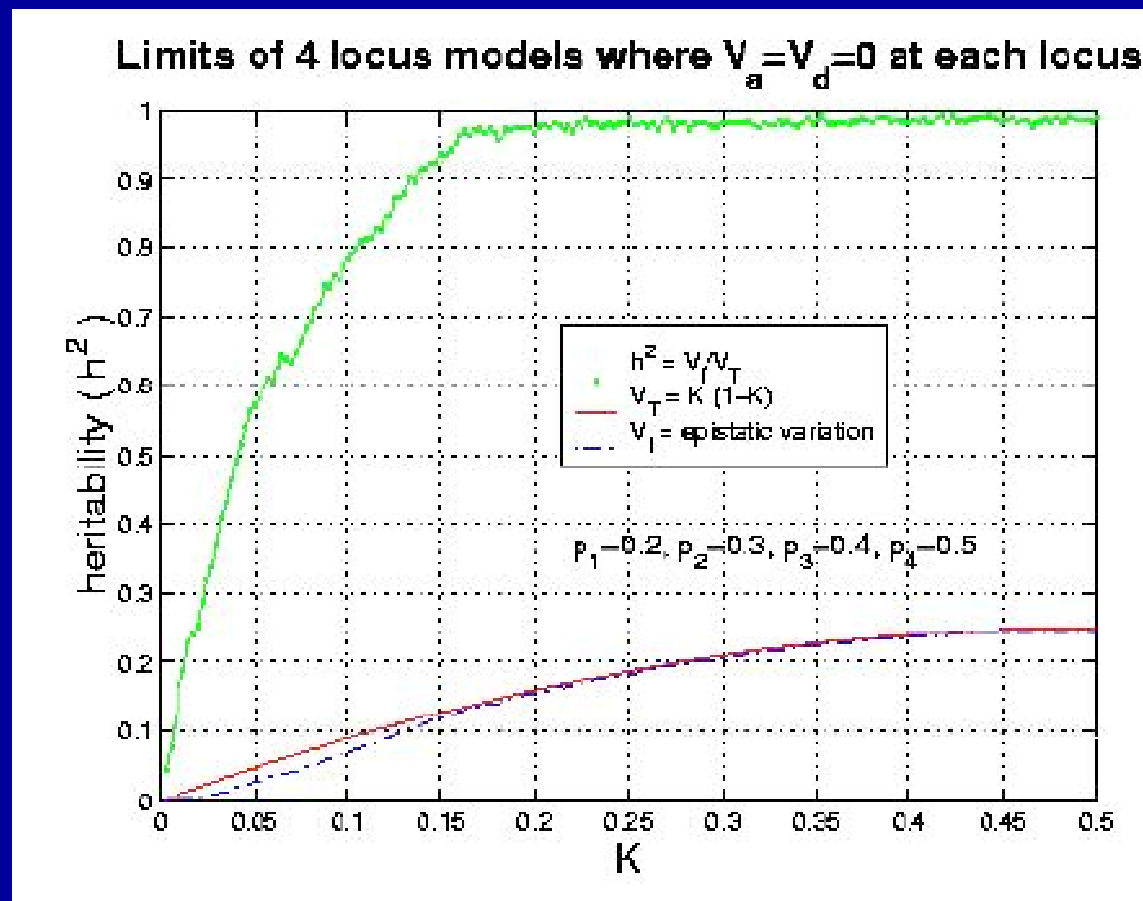
# Maximum Possible Heritability in Purely Epistatic (Qualitative) Models



# Maximum Possible Heritability in Purely Epistatic (Qualitative) Models



# Maximum Possible Heritability in Purely Epistatic (Qualitative) Models



# Testing for Epistasis contributing to quantitative traits

Basic Question:

Do subsets of multi-locus genotypes correspond to different mean trait values?

# Testing for Epistasis contributing to quantitative traits

Basic Question:

Do subsets of multi-locus genotypes correspond to different mean trait values?

Simplest approach:

F-test for difference in means between several groups

Drawbacks:

- Rejection of the null does not provide a model
- No measure of importance for the differences

# Combinatorial Partition Method

(Nelson et al. 2001)

Evaluates every partition a multilocus genotype matrix  
for the amount of phenotypic variation explained

## Advantages:

- Provides an epistatic model for further investigation
- Relates the partition to a measure of importance:  $R^2$

# Combinatorial Partition Method

(Nelson et al. 2001)

Evaluates every partition a multilocus genotype matrix for the amount of phenotypic variation explained

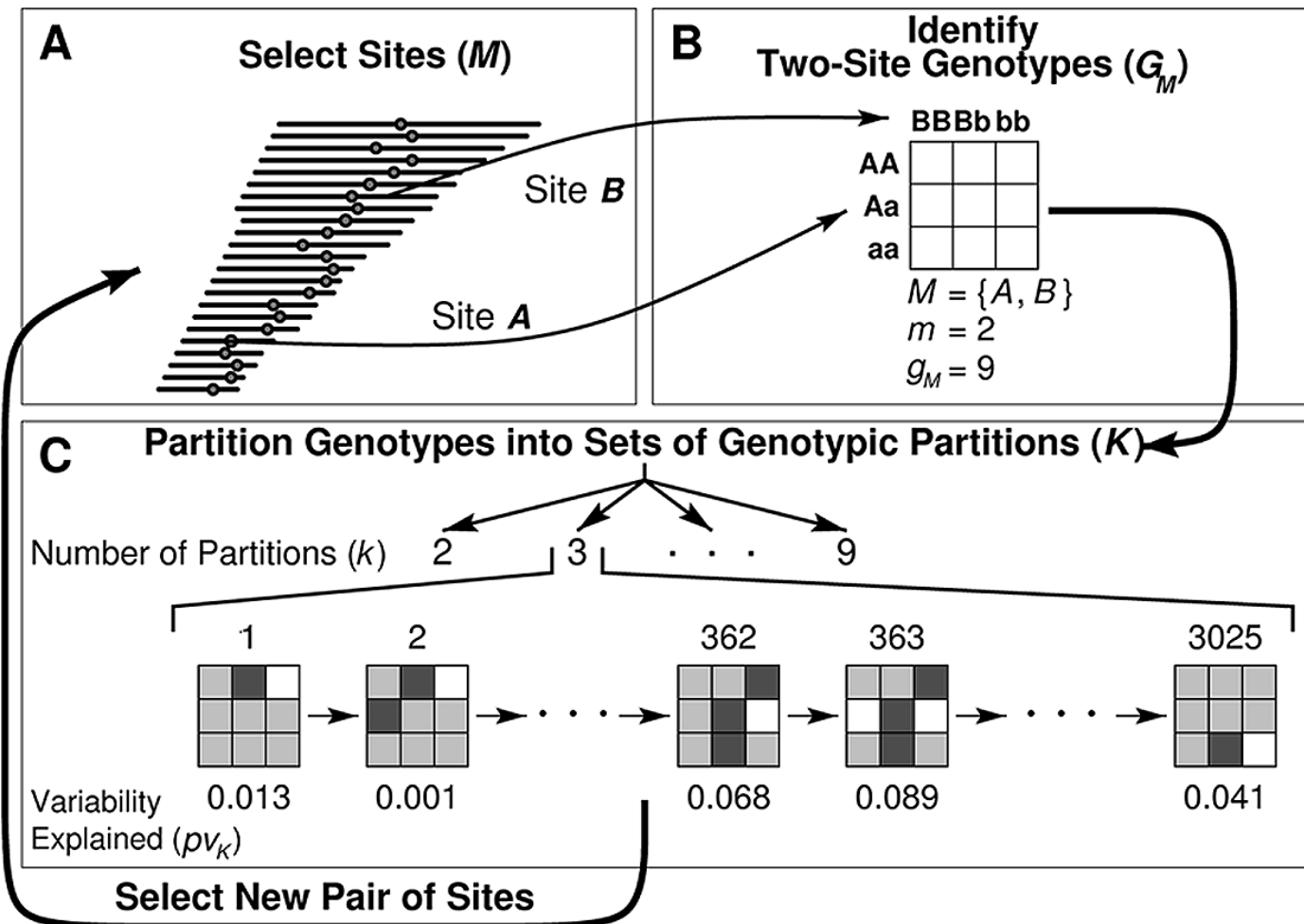
## Advantages:

- Provides an epistatic model for further investigation
- Relates the partition to a measure of importance:  $R^2$

## Drawbacks:

- Computation - (impractical for more than 2 loci)
- No easy way to assess statistical significance

# CPM algorithm for 2-locus analyses



CPM (Nelson *et al.* 2001. *Genome Research* 11:458-470)

Thanks to Taylor Maxwell

# Computations for CPM

Ways to partition  $g$  genotypes into  $K$  sets:  $S(g, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^g$

**21,146** partitions evaluated for each pair of bi-allelic candidate loci

Approximately  **$10^{21}$**  partitions for each combination of 3 loci

# Computations for CPM

Ways to partition  $g$  genotypes into  $K$  sets:  $S(g, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^g$

**21,146** partitions evaluated for each pair of bi-allelic candidate loci

Approximately  **$10^{21}$**  partitions for each combination of 3 loci

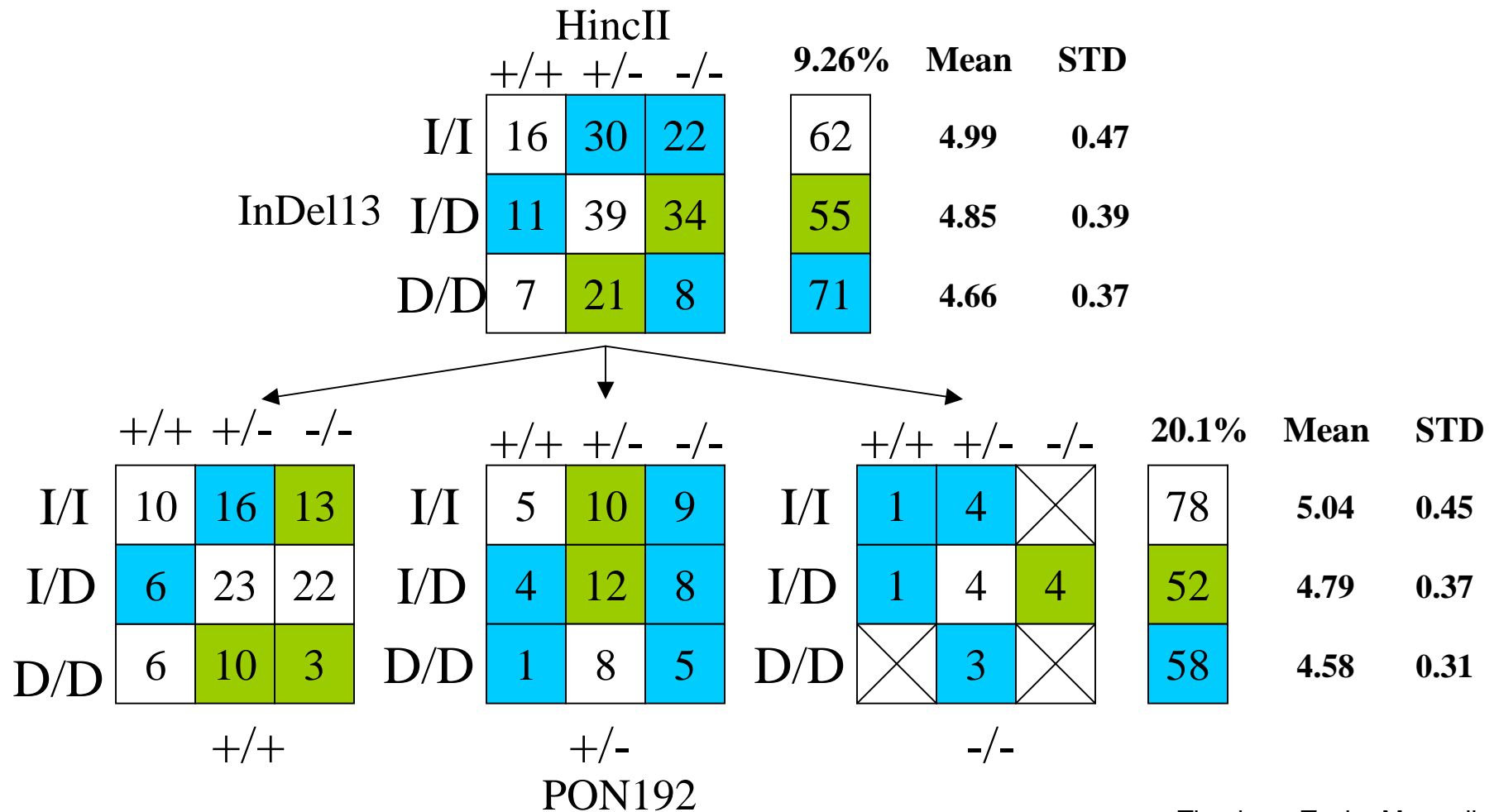
Evaluating 1 million partitions each second,

checking the partitions for the first three loci: 31 million years

# Why a 3-locus analysis might be good:

Serum Triglyceride

2-loci explain **9.3%** of the trait variation, 3-loci explain **20.1%**



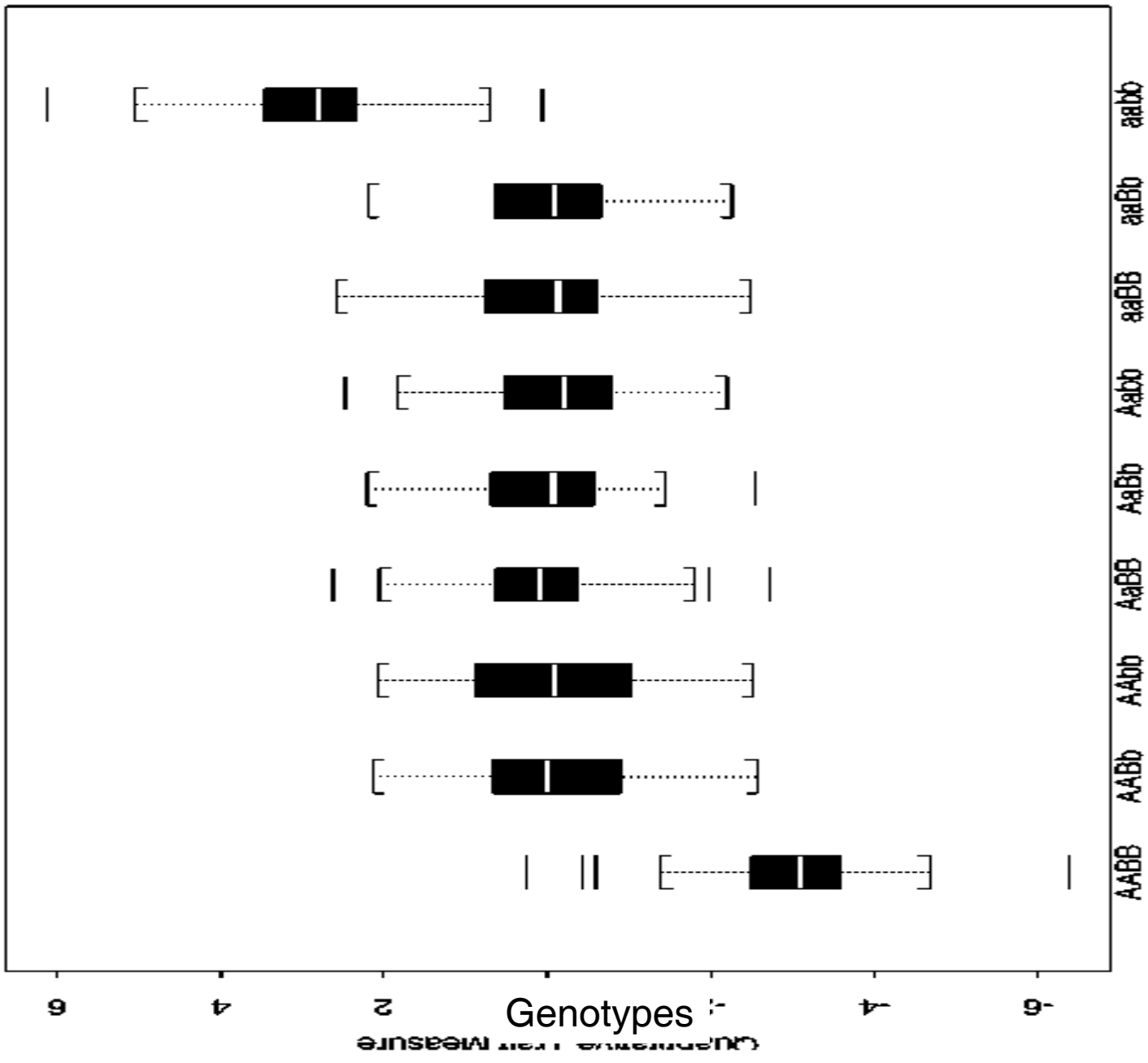
Thanks to Taylor Maxwell

# Observation

No partition that merges genotypes with widely differing means can be efficient at explaining the variation

This fact can be used to restrict the number of partitions evaluated

Quantitative Trait



# Restricted Partition Method

## Algorithm:

- Test cells for different means (using multiple comparison method)
- Merge two nearest groups (that are not significantly different)
- Iterate until groups all different or all cells are merged

If more than one group remains,  
evaluate model for variation explained ( $R^2$ )

BB Bb bb

AA

Aa

aa


BB Bb bb

AA

Aa

aa


BB Bb bb

AA

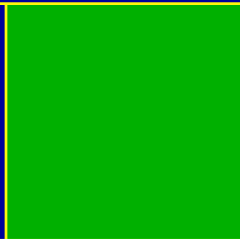
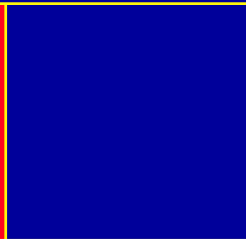
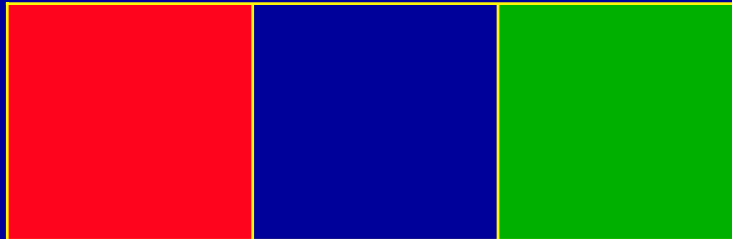
Aa

aa

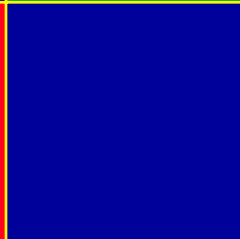
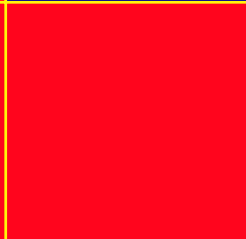
			BB
			Bb
			bb
AA			
Aa			
aa			

BB Bb bb

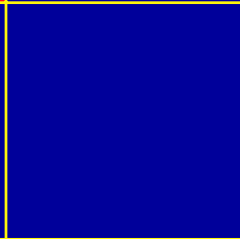
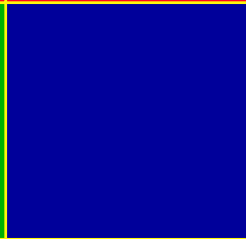
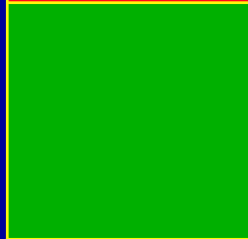
AA



Aa

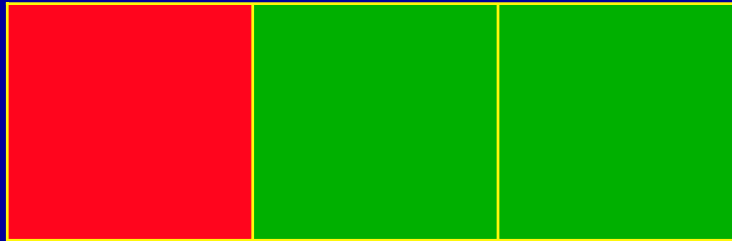


aa



BB Bb bb

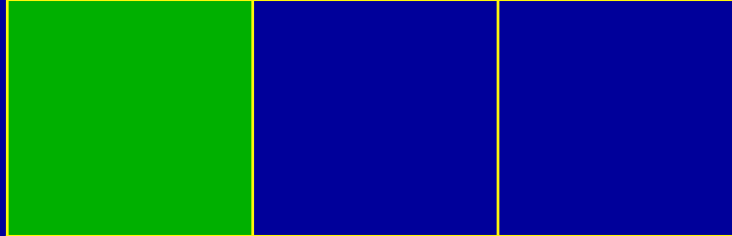
AA



Aa



aa



BB Bb bb

AA

Aa

aa

AA	BB	Bb	bb
Aa	BB	Bb	bb
aa	BB	Bb	bb

BB Bb bb

AA

Aa

aa

AA	Red	Green	Green
Aa	Red	Red	White
aa	Green	White	Red

# Computational Complexity for RPM

simultaneous  
loci analyzed

RPM

2	8 iterations to find the partition, one partition evaluated
3	26 iterations, one evaluation
4	80 iterations, one evaluation

# Computational Complexity for RPM

simultaneous  
loci analyzed

RPM

CPM

2	8 iterations to find the partition, one partition evaluated	21,146
3	26 iterations, one evaluation	$> 10^{21}$
4	80 iterations, one evaluation	$> 10^{88}$

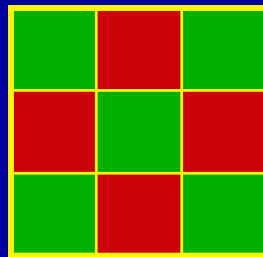
# What to do with the extra clock cycles?

Use permutation tests to obtain p-values for the results

# Testing the RPM

## Initial Simulations:

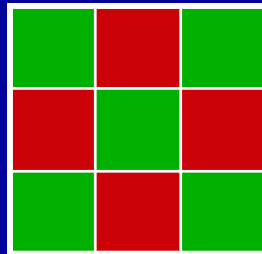
- A class of purely epistatic quantitative trait model
- 2 contributing and 8 unlinked loci simulated (allele freq = 0.5 for all)
  - Groups had different mean trait values =  $\mu_i$
  - Traits of individuals =  $\mu_i + \varepsilon$  ( $\varepsilon$  from  $N(0,1)$ )
  - 4 distances between the group means examined
  - 500 unrelated subjects each simulation



Checker board

# Testing the RPM

(Simulated Data - 1000 data sets, 500 individuals each)



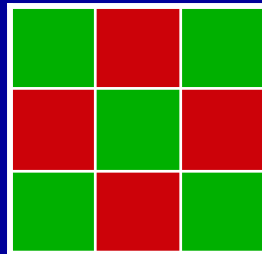
Contributing Loci

Other loci

sd	Model $R^2$	RPM $R^2$	TP%	FP%	$R^2 \neq 0$	TP %
0.25	0.015	0.024	9.7	90.0	0.014	37.8
0.5	0.059	0.066	51.4	40.2	0.014	35.8
1.0	0.200	0.209	79.3	1.1	0.015	38.3
2.0	0.500	0.508	77.9	0	0.014	37.6

# Testing the RPM

(Simulated Data)



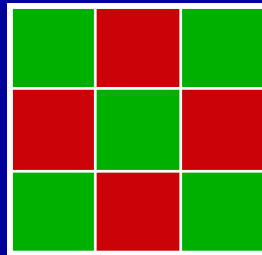
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	TP%	FP%	R <sup>2</sup> ≠ 0	TP %
<b>0.25</b>	0.015	0.024	9.7	90.0	0.014	37.8
<b>0.5</b>	0.059	0.066	51.4	40.2	0.014	35.8
<b>1.0</b>	0.200	0.209	79.3	1.1	0.015	38.3
<b>2.0</b>	0.500	0.508	77.9	0	0.014	37.6

# Testing the RPM

(Simulated Data)



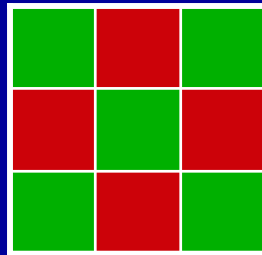
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	TP%	FP%	R <sup>2</sup> ≠ 0	TP %
0.25	<b>0.015</b>	0.024	9.7	90.0	0.014	37.8
0.5	<b>0.059</b>	0.066	51.4	40.2	0.014	35.8
1.0	<b>0.200</b>	0.209	79.3	1.1	0.015	38.3
2.0	<b>0.500</b>	0.508	77.9	0	0.014	37.6

# Testing the RPM

(Simulated Data)



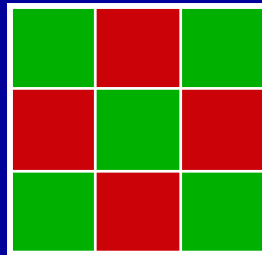
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	TP%	FP%	R <sup>2</sup> ≠ 0	TP %
0.25	0.015	<b>0.024</b>	9.7	90.0	0.014	37.8
0.5	0.059	<b>0.066</b>	51.4	40.2	0.014	35.8
1.0	0.200	<b>0.209</b>	79.3	1.1	0.015	38.3
2.0	0.500	<b>0.508</b>	77.9	0	0.014	37.6

# Testing the RPM

(Simulated Data)



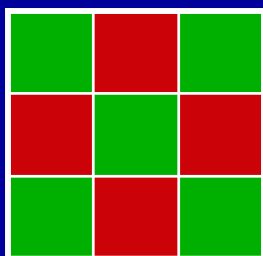
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	<b>TP%</b>	FP%	R <sup>2</sup> ≠ 0	TP %
0.25	0.015	0.024	<b>9.7</b>	90.0	0.014	37.8
0.5	0.059	0.066	<b>51.4</b>	40.2	0.014	35.8
1.0	0.200	0.209	<b>79.3</b>	1.1	0.015	38.3
2.0	0.500	0.508	<b>77.9</b>	0	0.014	37.6

# Testing the RPM

(Simulated Data)



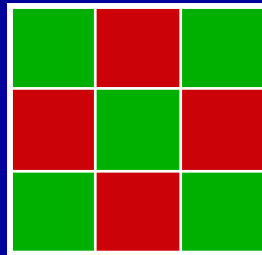
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	TP%	FP%	R <sup>2</sup> ≠ 0	TP %
0.25	0.015	0.024	9.7	<b>90.0</b>	0.014	37.8
0.5	0.059	0.066	51.4	<b>40.2</b>	0.014	35.8
1.0	0.200	0.209	79.3	<b>1.1</b>	0.015	38.3
2.0	0.500	0.508	77.9	<b>0</b>	0.014	37.6

# Testing the RPM

(Simulated Data)



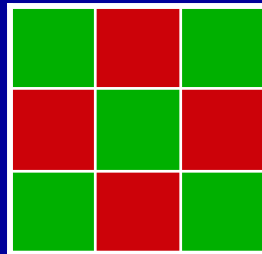
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	TP%	FP%	R <sup>2</sup> ≠ 0	TP %
0.25	0.015	0.024	9.7	90.0	<b>0.014</b>	<b>37.8</b>
0.5	0.059	0.066	51.4	40.2	<b>0.014</b>	<b>35.8</b>
1.0	0.200	0.209	79.3	1.1	<b>0.015</b>	<b>38.3</b>
2.0	0.500	0.508	77.9	0	<b>0.014</b>	<b>37.6</b>

# Testing the RPM

(Simulated Data)



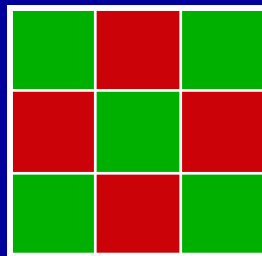
## Contributing Loci

## Other loci

sd	Model R <sup>2</sup>	RPM R <sup>2</sup>	TP%	FP%	R <sup>2</sup> ≠ 0	TP %
0.25	0.015	0.024	9.7	90.0	0.014	<b>37.8</b>
0.5	0.059	0.066	51.4	40.2	0.014	<b>35.8</b>
1.0	0.200	0.209	79.3	1.1	0.015	<b>38.3</b>
2.0	0.500	0.508	77.9	0	0.014	<b>37.6</b>

# Power tests for the RPM

(100 data sets, 10 loci, 5000 permutations/locus pair)



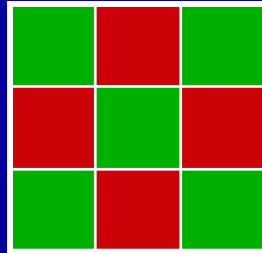
Contributing Loci  
(Power)

Other loci  
(False Positives)

sd	$R^2$	Contributing Loci (Power)		Other loci (False Positives)		
		$p_c < 0.05$	$p_c < 0.01$	$p_u < 0.05$	$p_c < 0.05$	$p_c < 0.01$
0.25	0.015	8%	7%	256 (5.8%)	6 (0.14%)	5 (0.11%)
0.5	0.059	87%	81%	204 (4.6%)	2 (0.05%)	1 (0.02%)
1.0	0.200	100%	100%	259 (5.8%)	5 (0.11%)	1 (0.02%)
2.0	0.500	100%	100%	230 (5.2%)	2 (0.04%)	1 (0.02%)

# Power tests for the RPM

(100 data sets, 10 loci, 5000 permutations/locus pair)



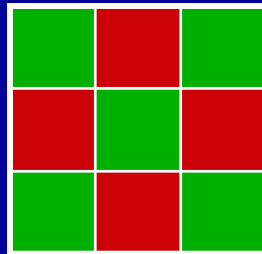
Contributing Loci  
(Power)

Other loci  
(False Positives)

sd	R <sup>2</sup>	Contributing Loci (Power)		Other loci (False Positives)		
		p <sub>c</sub> < 0.05	p <sub>c</sub> < 0.01	p <sub>u</sub> < 0.05	p <sub>c</sub> < 0.05	p <sub>c</sub> < 0.01
0.25	0.015	8%	7%	256 (5.8%)	6 (0.14%)	5 (0.11%)
0.5	0.059	87%	81%	204 (4.6%)	2 (0.05%)	1 (0.02%)
1.0	0.200	100%	100%	259 (5.8%)	5 (0.11%)	1 (0.02%)
2.0	0.500	100%	100%	230 (5.2%)	2 (0.04%)	1 (0.02%)

# Power tests for the RPM

(100 data sets, 10 loci, 5000 permutations/locus pair)



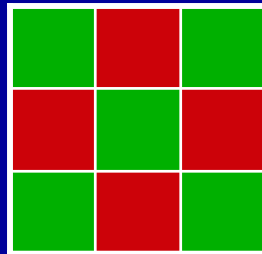
Contributing Loci  
(Power)

Other loci  
(False Positives)

sd	R <sup>2</sup>	Contributing Loci (Power)		Other loci (False Positives)		
		p <sub>c</sub> < 0.05	p <sub>c</sub> < 0.01	p <sub>u</sub> < 0.05	p <sub>c</sub> < 0.05	p <sub>c</sub> < 0.01
0.25	0.015	8%	7%	256 (5.8%)	6 (0.14%)	5 (0.11%)
0.5	0.059	87%	81%	204 (4.6%)	2 (0.05%)	1 (0.02%)
1.0	0.200	100%	100%	259 (5.8%)	5 (0.11%)	1 (0.02%)
2.0	0.500	100%	100%	230 (5.2%)	2 (0.04%)	1 (0.02%)

# Power tests for the RPM

(100 data sets, 10 loci, 5000 permutations/locus pair)



Contributing Loci  
(Power)

Other loci  
(False Positives)

sd	$R^2$	Contributing Loci (Power)		Other loci (False Positives)		
		$p_c < 0.05$	$p_c < 0.01$	$p_u < 0.05$	$p_c < 0.05$	$p_c < 0.01$
0.25	0.015	8%	7%	256 (5.8%)	6 (0.14%)	5 (0.11%)
0.5	0.059	87%	81%	204 (4.6%)	2 (0.05%)	1 (0.02%)
1.0	0.200	100%	100%	259 (5.8%)	5 (0.11%)	1 (0.02%)
2.0	0.500	100%	100%	230 (5.2%)	2 (0.04%)	1 (0.02%)

# Unequal Allele Frequency Models

(100 data sets each, N=500, 5000 permutation/locus pair)

Examined epistatic models with various  $R^2$ : 0.05, 0.10, 0.30

Contributing loci allele frequencies

.5 .3 .1

.5	■	■	■
.3	■	■	■
.1	■	■	■

Non-contributing loci allele frequencies

.5 .4 .3 .2 .1

.5	■	■	■	■	■
.4	■	■	■	■	■
.3	■	■	■	■	■
.2	■	■	■	■	■

# Unequal Allele Frequency Models

(100 data sets each, N=500, 5000 permutation/locus pair)

Examined epistatic models with various  $R^2$ : 0.05, 0.10, 0.30

Results for  $R^2 = 0.05$

Contributing Loci  
(power)

Other Loci Combined  
(false positives)

Allele Freq		$p_c < 0.05$	$p_c < 0.01$	$p_u < 0.05$	$p_c < 0.05$	$p_c < 0.01$
0.5	0.5	0.78	0.68	59 (5.4%)	1	0
	0.3	1.00	0.99	62 (5.6%)	3	1
	0.1	1.00	1.00	49 (4.5%)	1	1
0.3	0.3	0.85	0.71	55 (5.0%)	1	0
	0.1	1.00	1.00	46 (4.2%)	0	0
0.1	0.1	0.71	0.64	68 (6.2%)	1	0

# Unequal Allele Frequency Models

(100 data sets each, N=500, 5000 permutation/locus pair)  
Examined epistatic models with various  $R^2$ : 0.05, 0.10, 0.30

Results for  $R^2 = 0.05$

Contributing Loci (power)				Other Loci Combined (false positives)		
Allele Freq		$p_c < 0.05$	$p_c < 0.01$	$p_u < 0.05$	$p_c < 0.05$	$p_c < 0.01$
0.5	0.5	0.78	0.68	59 (5.4%)	1	0
	0.3	1.00	0.99	62 (5.6%)	3	1
	0.1	1.00	1.00	49 (4.5%)	1	1
0.3	0.3	0.85	0.71	55 (5.0%)	1	0
	0.1	1.00	1.00	46 (4.2%)	0	0
0.1	0.1	0.71	0.64	68 (6.2%)	1	0

# Unequal Allele Frequency Models

(100 data sets each, N=500, 5000 permutation/locus pair)

Examined epistatic models with various  $R^2$ : 0.05, 0.10, 0.30

Results for  $R^2 = 0.05$

Contributing Loci  
(power)

Other Loci Combined  
(false positives)

Allele Freq		$p_c < 0.05$	$p_c < 0.01$	$p_u < 0.05$	$p_c < 0.05$	$p_c < 0.01$
0.5	0.5	0.78	0.68	59 (5.4%)	1	0
	0.3	1.00	0.99	62 (5.6%)	3	1
	0.1	1.00	1.00	49 (4.5%)	1	1
0.3	0.3	0.85	0.71	55 (5.0%)	1	0
	0.1	1.00	1.00	46 (4.2%)	0	0
0.1	0.1	0.71	0.64	68 (6.2%)	1	0

# Applying the RPM to real data

## Etoposide metabolism data:

- Etoposide is a commonly used anticancer agent with a broad range of anti-tumor activity.
- Data Provided by the St. Jude Children's Research Hospital

# Applying the RPM to real data

## Etoposide metabolism data:

- Etoposide is a commonly used anticancer agent with a broad range of anti-tumor activity.
- Data Provided by the St. Jude Children's Research Hospital
- Phenotypes: 2 pharmacokinetic assessments of etoposide metabolism
- Predictor covariates: Genotypes from 8 candidate loci, Race, Sex  
(Data: genotypes and phenotypes of 102 individuals)

# Applying the RPM to real data

## Etoposide metabolism data:

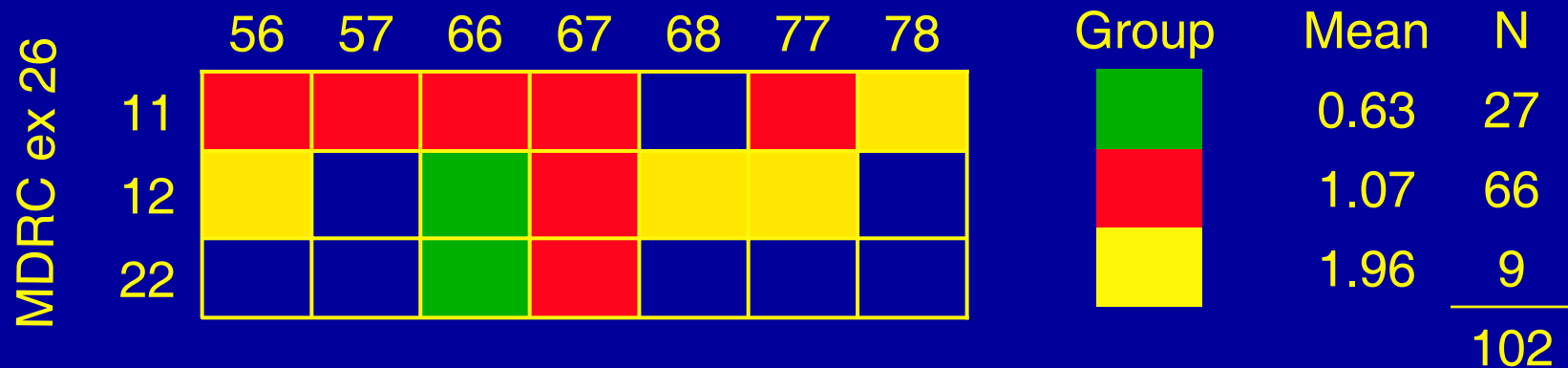
- Etoposide is a commonly used anticancer agent with a broad range of anti-tumor activity.
- Data Provided by the St. Jude Children's Research Hospital
- Phenotypes: 2 pharmacokinetic assessments of etoposide metabolism
- Predictor covariates: Genotypes from 8 candidate loci, Race, Sex  
(Data: genotypes and phenotypes of 102 individuals)
- None of the predictors were significant in univariate analyses

# Etoposide Metabolism

First analysis: p-values corrected for  $2 \times C(10,2) = 90$  comparisons

## Result for Trait 2 (AUC)

UGT1A1 genotype



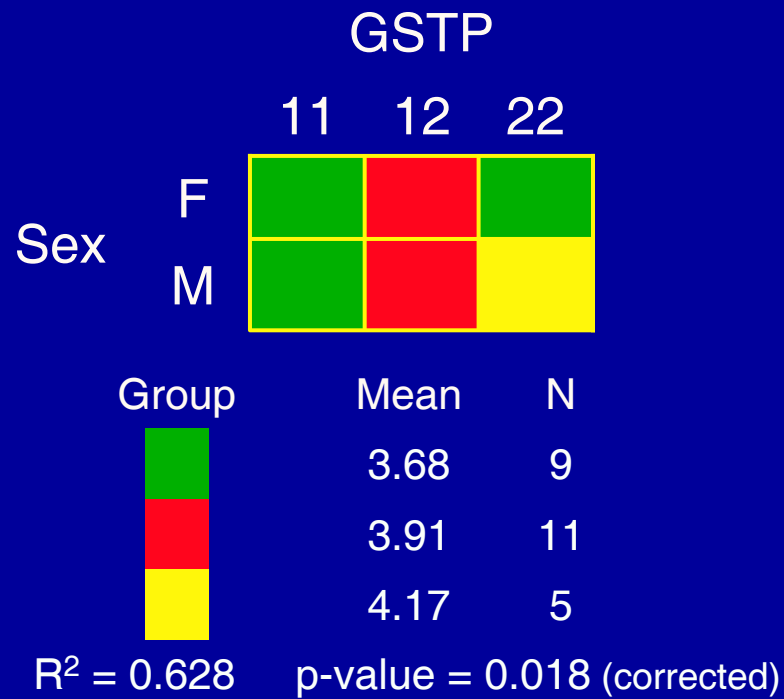
$R^2 = 0.266$

p-value = 0.045 (corrected)

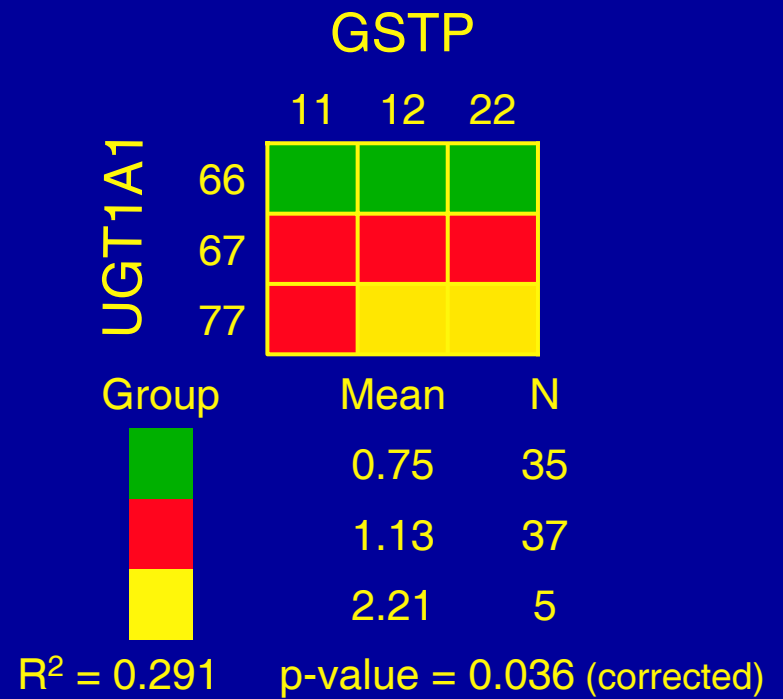
# Etoposide Metabolism

Second analysis: 4 Subpopulations: AA, CA, Male, Female  
 p-values corrected for a total of 378 tests (including the original 90)

Trait 1 (clearance) (AA)



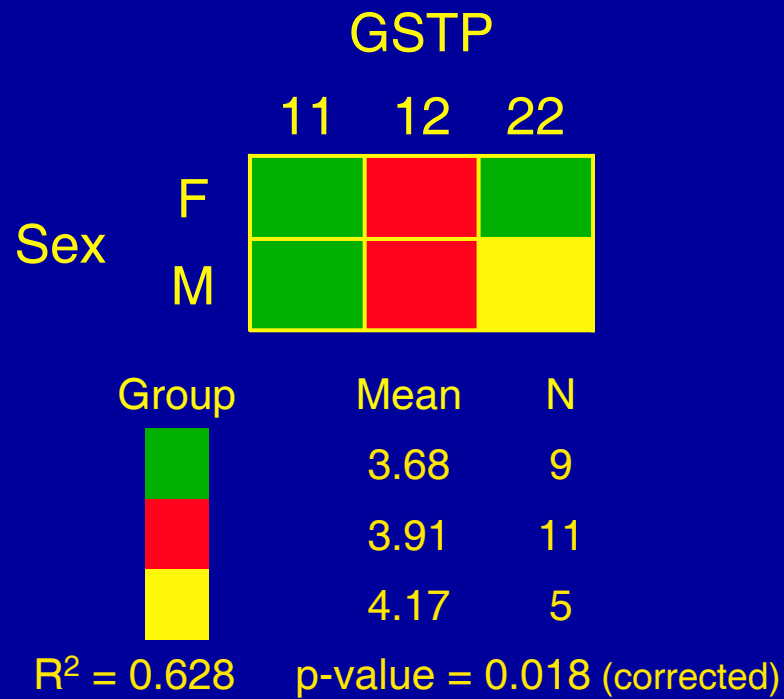
Trait 2 (AUC) (CA)



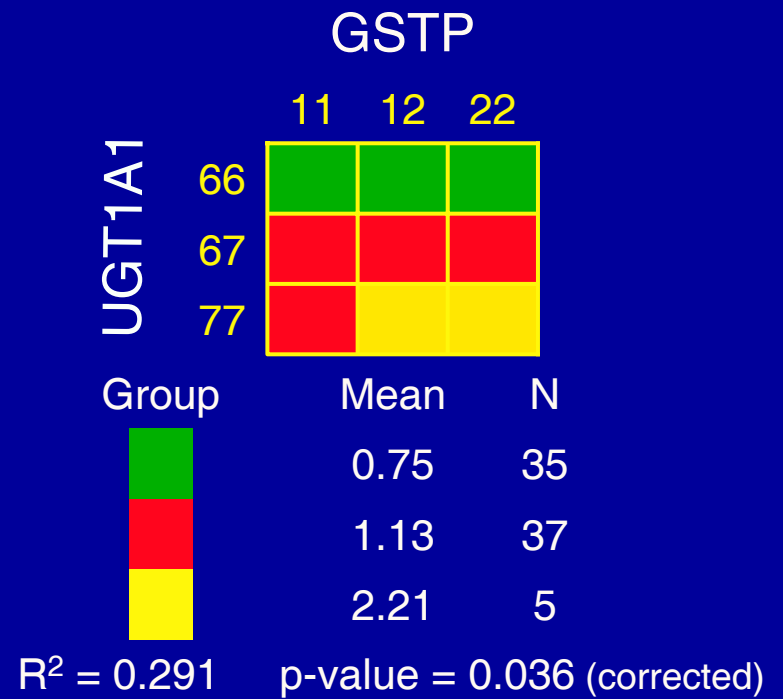
# Etoposide Metabolism

Second analysis: 4 Subpopulations: AA, CA, Male, Female  
 p-values corrected for a total of 378 tests (including the original 90)

Trait 1 (clearance) (AA)



Trait 2 (AUC) (CA)



# Continuing work

- Further testing:
  - Models with 3 and 4 contributing loci
  - Effect of model misspecification
  - Greater number of simulations for robustness
  - Varying the merging parameters (now merges if  $p > 0.05$ )
- Applying to real data (including gene x environment interactions)
- Adapting the method for qualitative traits
- Difficulties to address:
  - Computation time for permutation tests
  - Multiple testing correction (FDR?)
  - Robustness (cross validation?)

For more information

Detecting Epistatic Interactions Contributing to  
Quantitative Traits

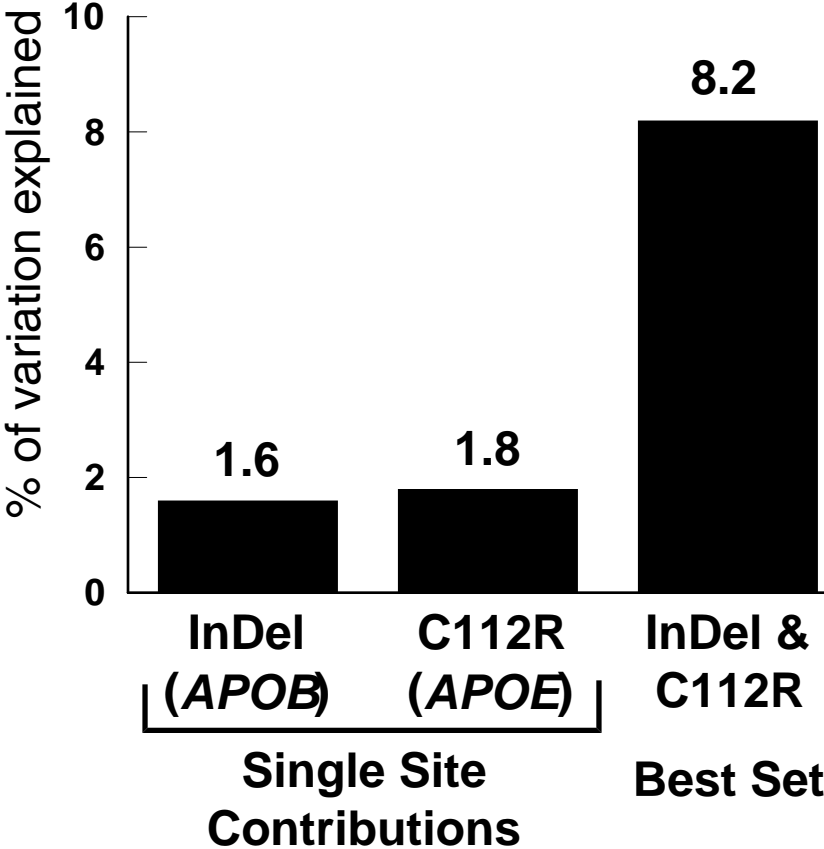
Robert Culverhouse, Tsvika Klein, and William Shannon

Online in Genetic Epidemiology



# Variability in Ln(Triglyceride) explained by Single locus vs Two locus analyses

Females, n=241



(Nelson et al 2001)