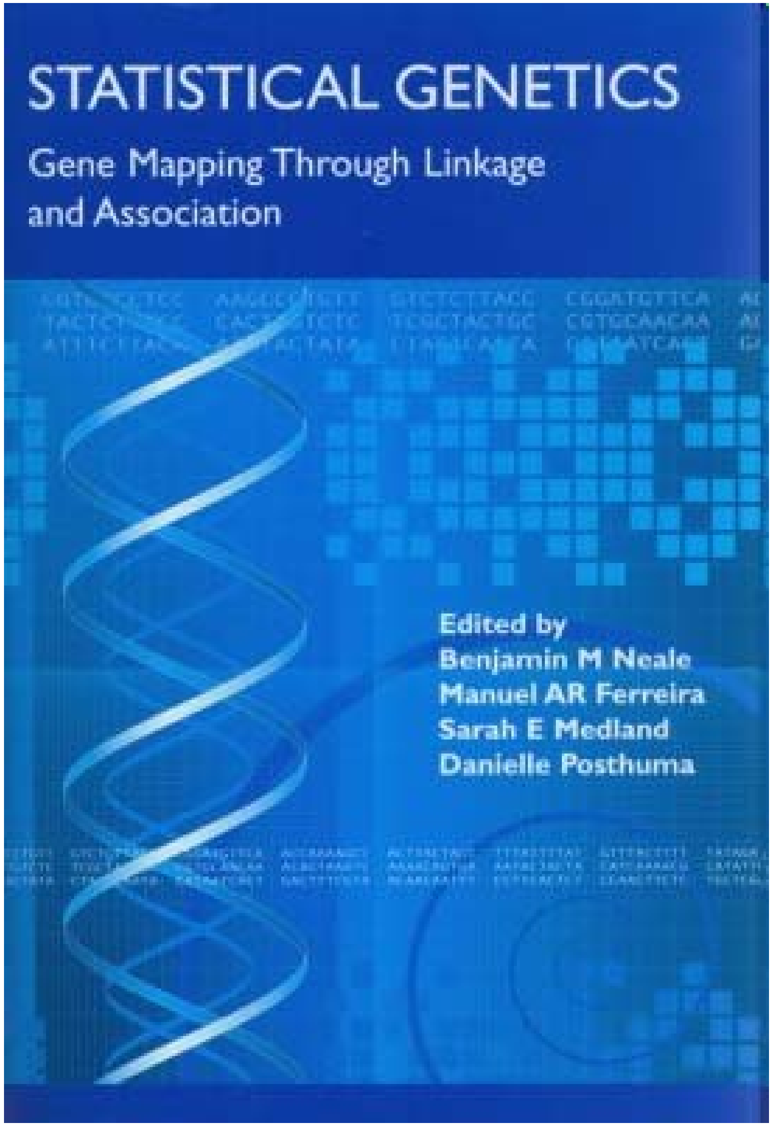


STATISTICAL GENETICS

Gene Mapping Through Linkage and Association

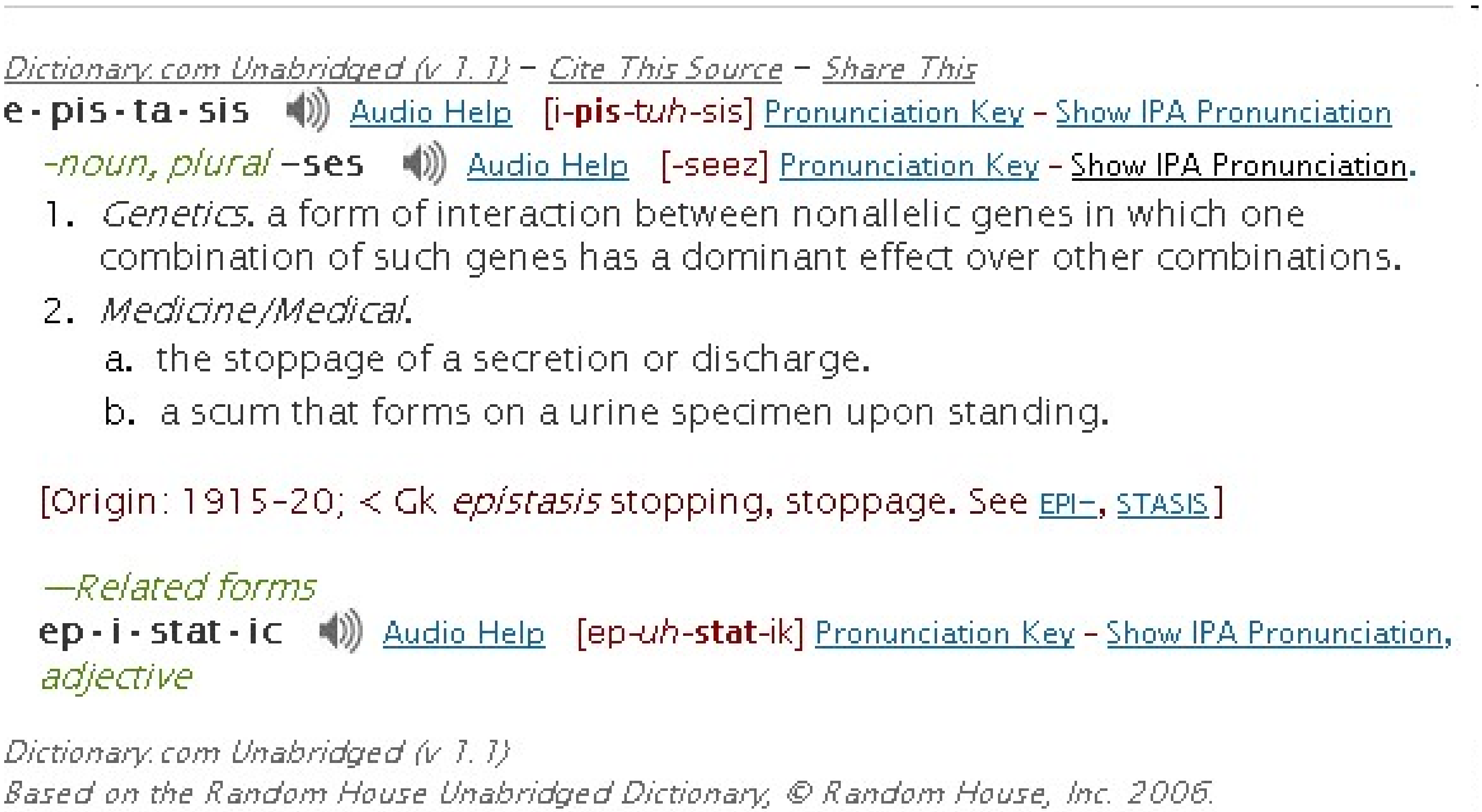
Edited by
Benjamin M Neale
Manuel AR Ferreira
Sarah E Medland
Danielle Posthuma

http://genemapping.org/

Epistasis in Association Studies

David Evans

University of
BRISTOL



Dictionary.com Unabridged (v 1.1) - *Cite This Source* - *Share This*

e·pis·ta·sis 🔊 Audio Help [i-**pis**-*tuh*-sis] Pronunciation Key - Show IPA Pronunciation

–noun, plural **–ses** 🔊 Audio Help [-seez] Pronunciation Key - Show IPA Pronunciation.

1. *Genetics.* a form of interaction between nonallelic genes in which one combination of such genes has a dominant effect over other combinations.
2. *Medicine/Medical.*
   a. the stoppage of a secretion or discharge.
   b. a scum that forms on a urine specimen upon standing.

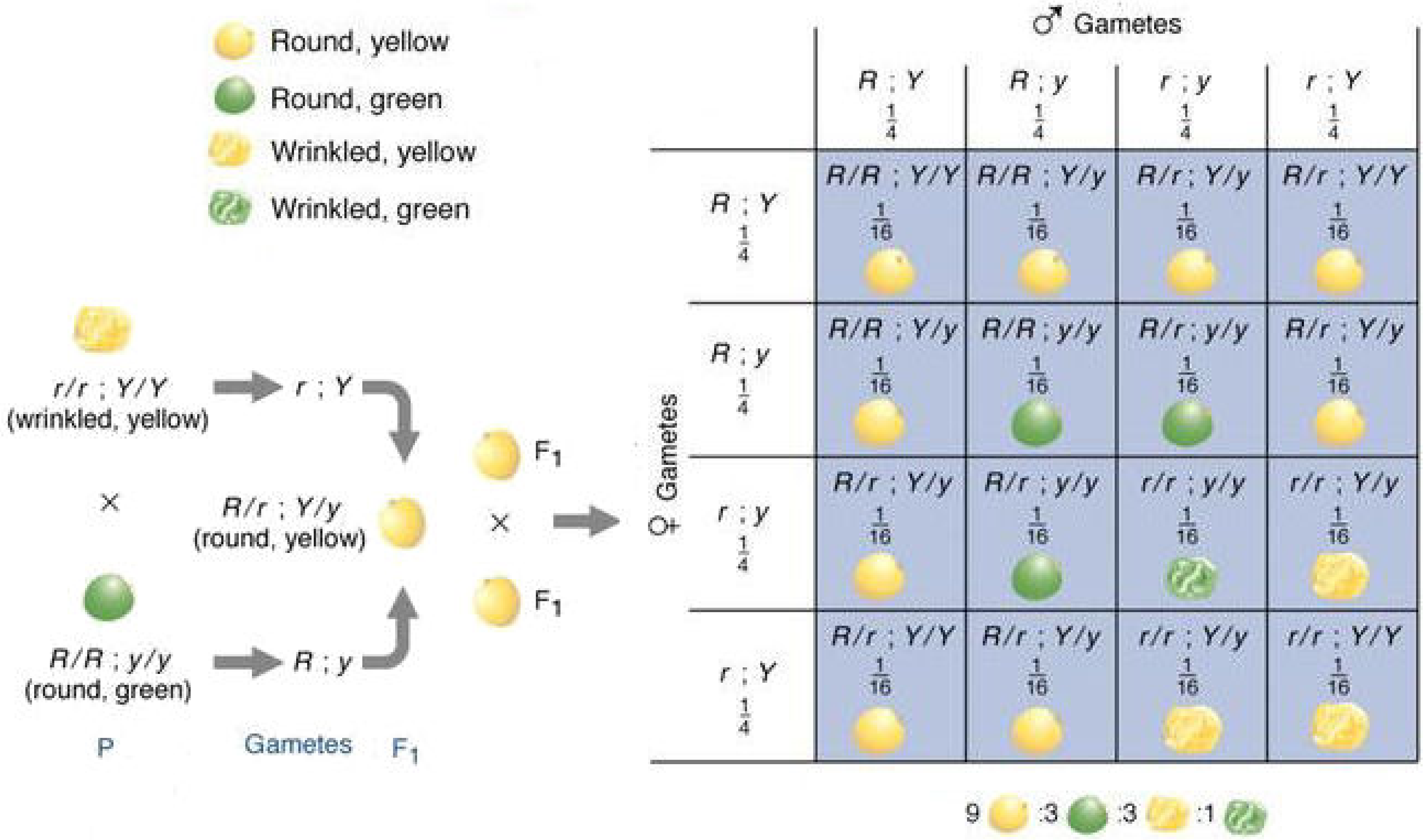
[Origin: 1915–20; < Gk *epistasis* stopping, stoppage. See EPI-, STASIS]

—Related forms

ep·i·stat·ic 🔊 Audio Help [ep-*uh*-**stat**-ik] Pronunciation Key - Show IPA Pronunciation, *adjective*

Dictionary.com Unabridged (v 1.1)
Based on the Random House Unabridged Dictionary, © Random House, Inc. 2006.

Law of Independent Assortment



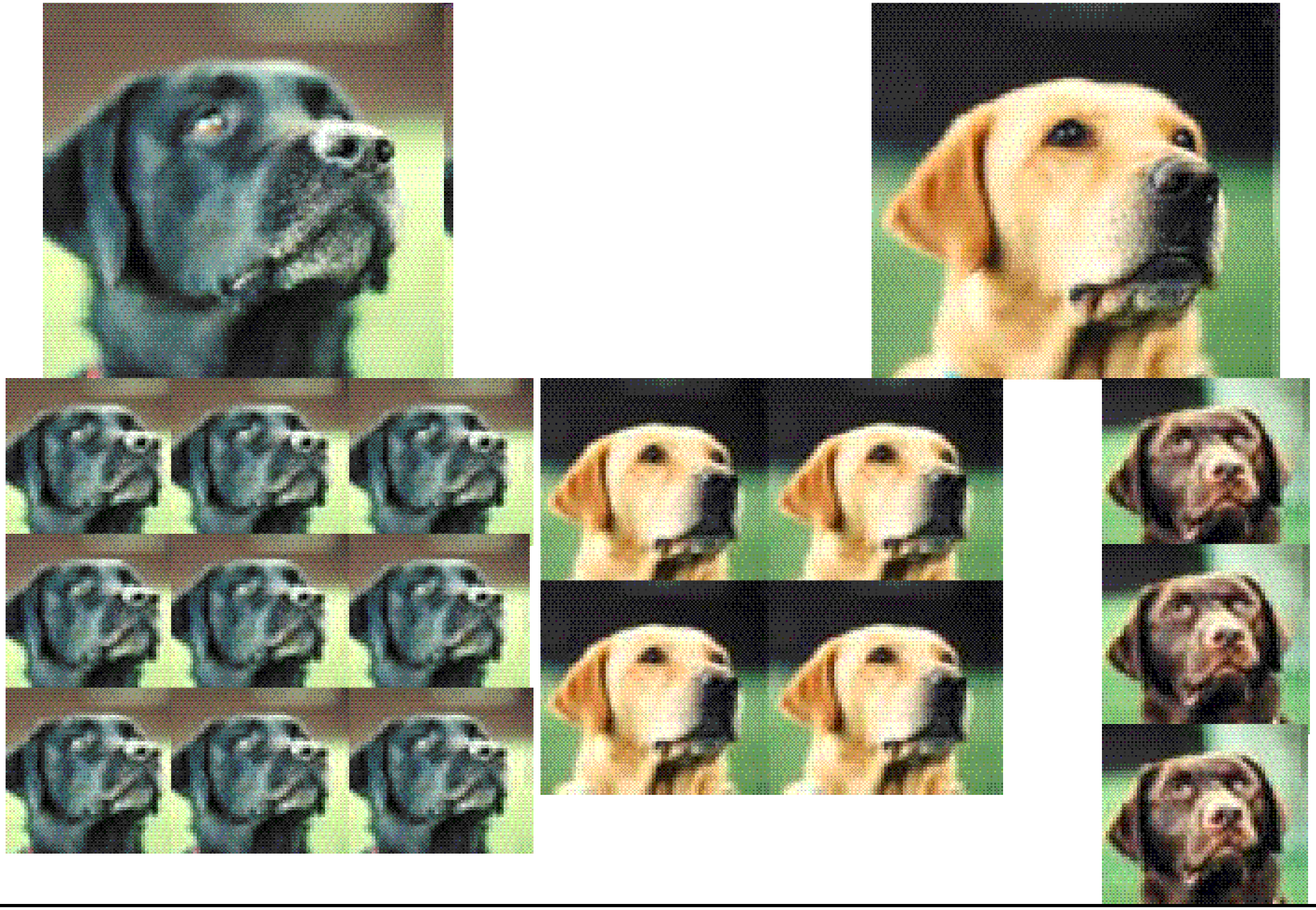
♀ Gametes	♂ Gametes			
	R ; Y $\frac{1}{4}$	R ; y $\frac{1}{4}$	r ; y $\frac{1}{4}$	r ; Y $\frac{1}{4}$
R ; Y $\frac{1}{4}$	R/R ; Y/Y $\frac{1}{16}$	R/R ; Y/y $\frac{1}{16}$	R/r ; Y/y $\frac{1}{16}$	R/r ; Y/Y $\frac{1}{16}$
R ; y $\frac{1}{4}$	R/R ; Y/y $\frac{1}{16}$	R/R ; y/y $\frac{1}{16}$	R/r ; y/y $\frac{1}{16}$	R/r ; Y/y $\frac{1}{16}$
r ; y $\frac{1}{4}$	R/r ; Y/y $\frac{1}{16}$	R/r ; y/y $\frac{1}{16}$	r/r ; y/y $\frac{1}{16}$	r/r ; Y/y $\frac{1}{16}$
r ; Y $\frac{1}{4}$	R/r ; Y/Y $\frac{1}{16}$	R/r ; Y/y $\frac{1}{16}$	r/r ; Y/y $\frac{1}{16}$	r/r ; Y/Y $\frac{1}{16}$

Biological Epistasis

- Bateson (1909) “a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect…”

- Phenotypic differences among individuals with various genotypes at one locus depend on their genotypes at other loci
  - Does NOT depend on allele frequency

Epistasis in the labrador retriever dog

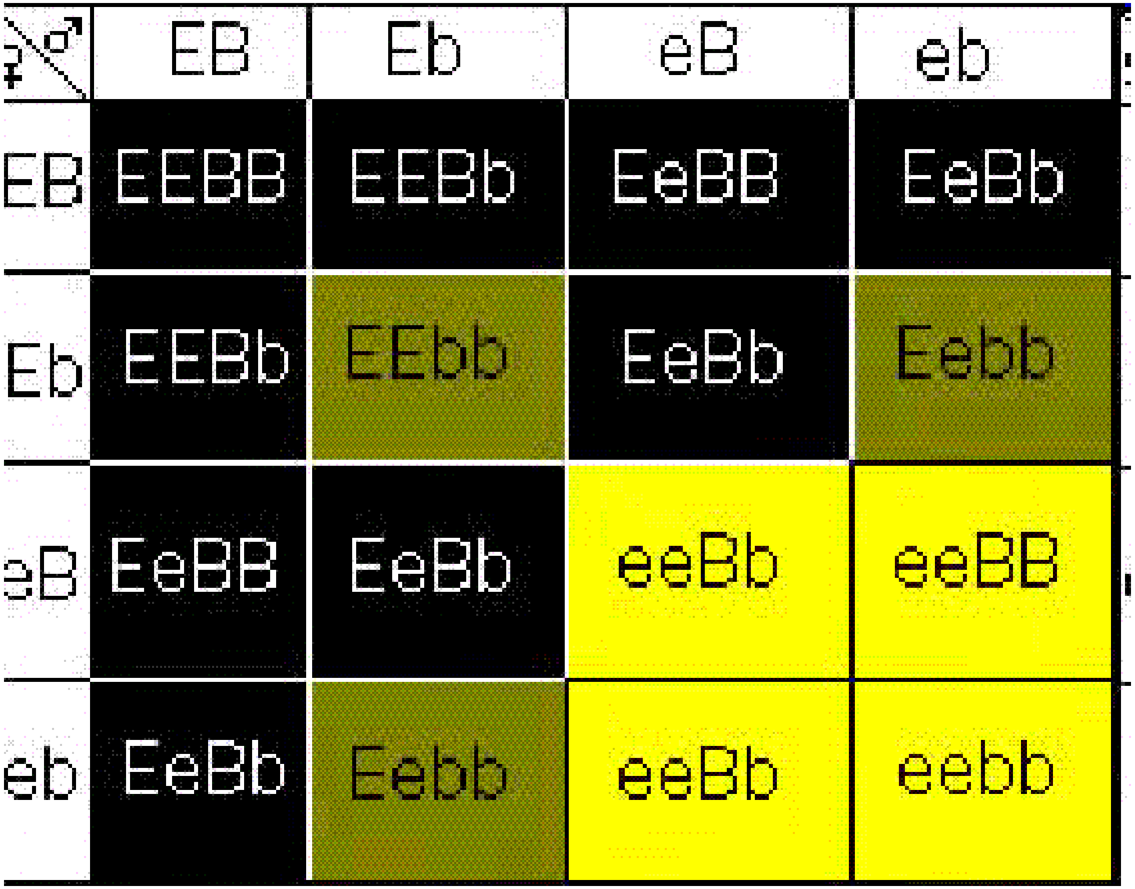


Recessive Epistasis

B = Black

b = brown

E/e = gold locus



♀ ♂	EB	Eb	eB	eb
EB	EEBB	EEBb	EeBB	EeBb
Eb	EEBb	EEbb	EeBb	Eebb
eB	EeBB	EeBb	eeBb	eeBB
eb	EeBb	Eebb	eeBb	eebb

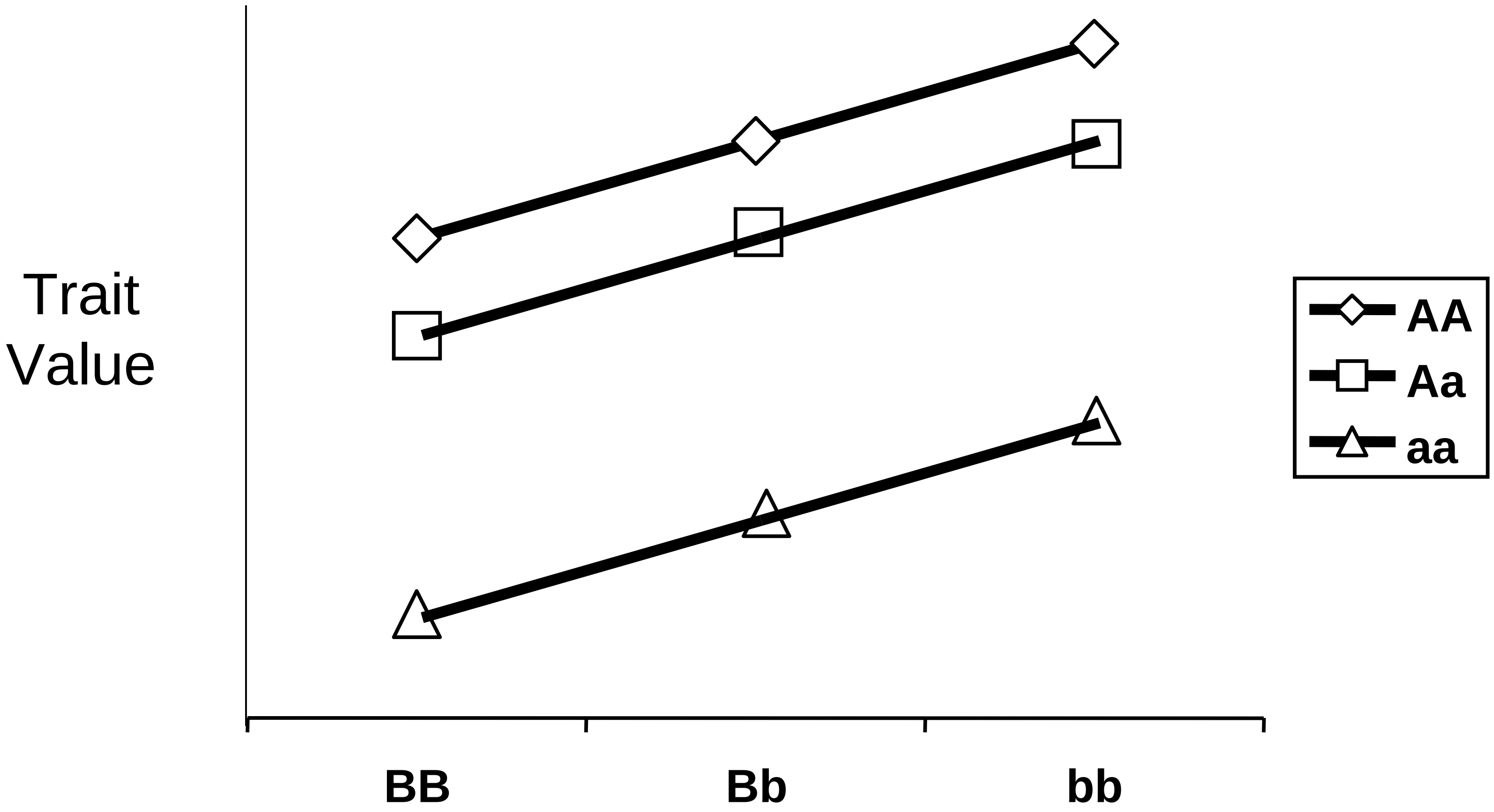
Statistical Epistasis

- Deviation of multilocus genotypic values from the additive combination of the single locus components
  - Close to statistical concept of interaction
  - Depends on allele frequencies
  - Population specific
  - May be scale dependent

- Different ways the epistatic values/variance components can be calculated:
  - Hierarchical ANOVA (Sham, 1998)
  - Method of Contrasts (Cockerham, 1954)
  - Using partial derivatives of the population mean (Kojima, 1959; Tiwari & Elston, 1997)

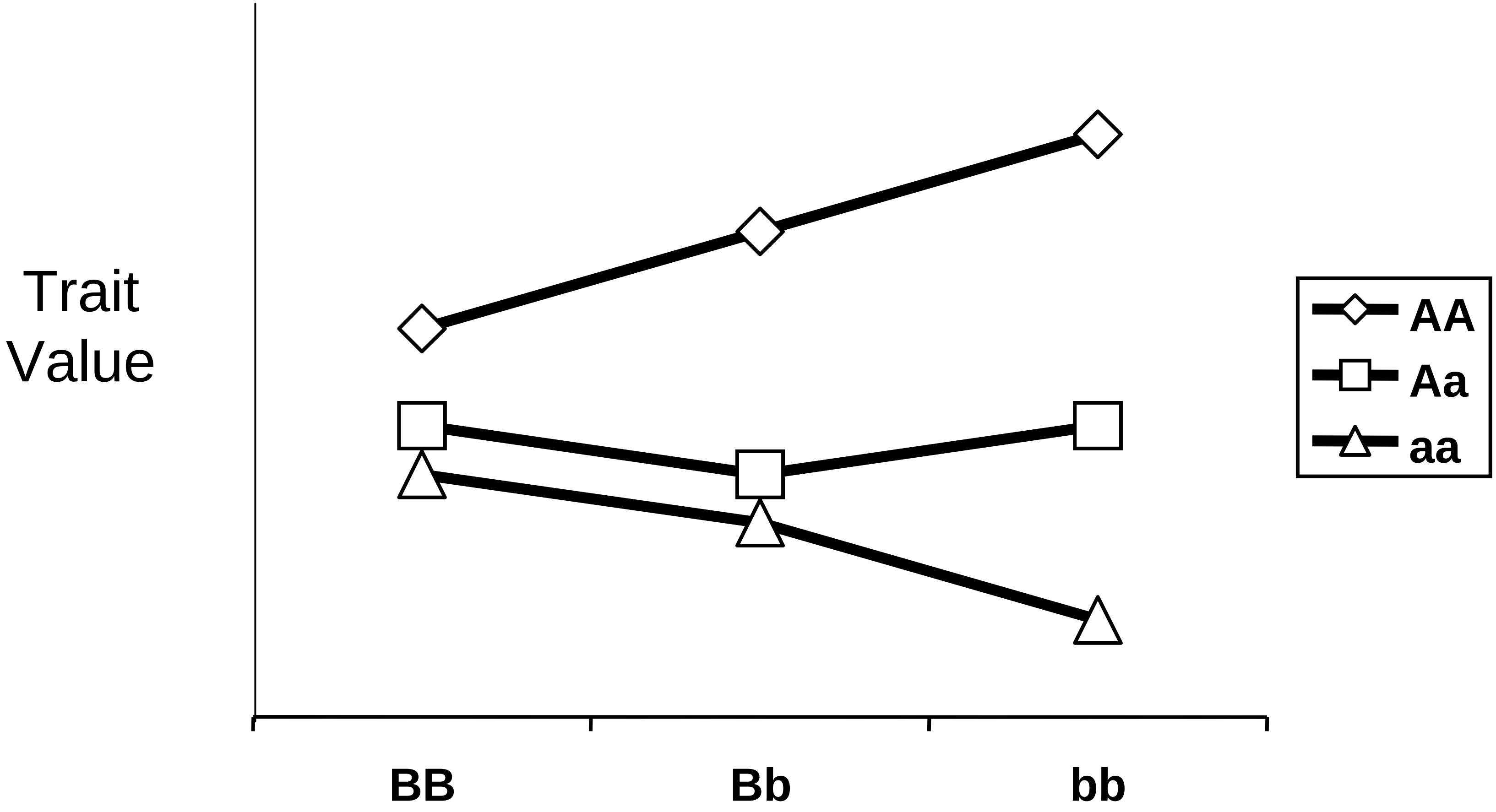
No Epistasis

- Effect at locus B independent of effect at Locus A



Epistasis Two Loci

- Locus A modifies the effect at locus B



Genetic Variance – One Locus

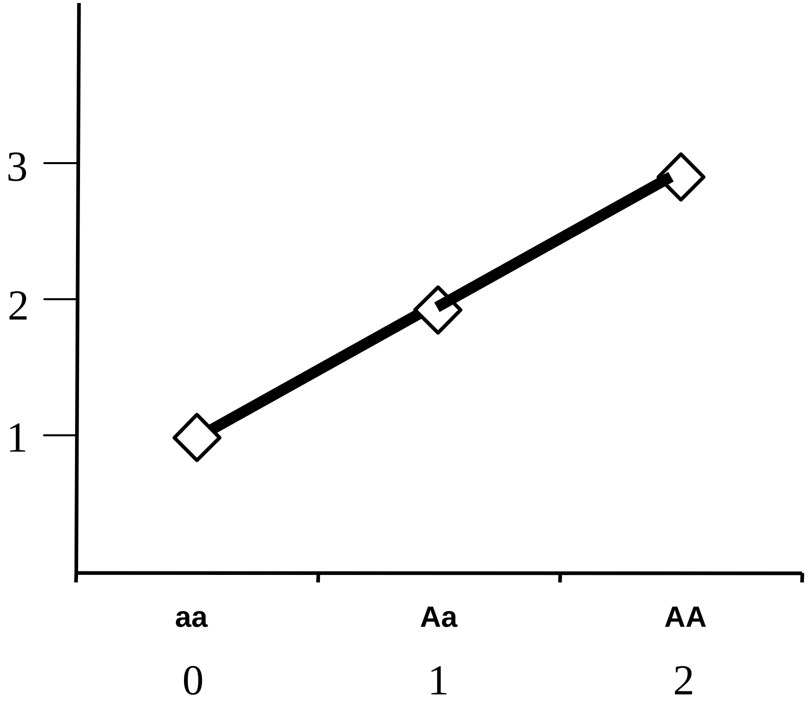
	AA	Aa	aa
Genotypic Mean *Frequency*	Y_{AA} f_{AA}	Y_{Aa} f_{Aa}	Y_{aa} f_{aa}

$$\mu = \sum_{i=AA,Aa,aa} f_i Y_i \qquad \sigma^2_{A+D} = \sum_{i=AA,Aa,aa} f_i (Y_i - \mu)^2$$

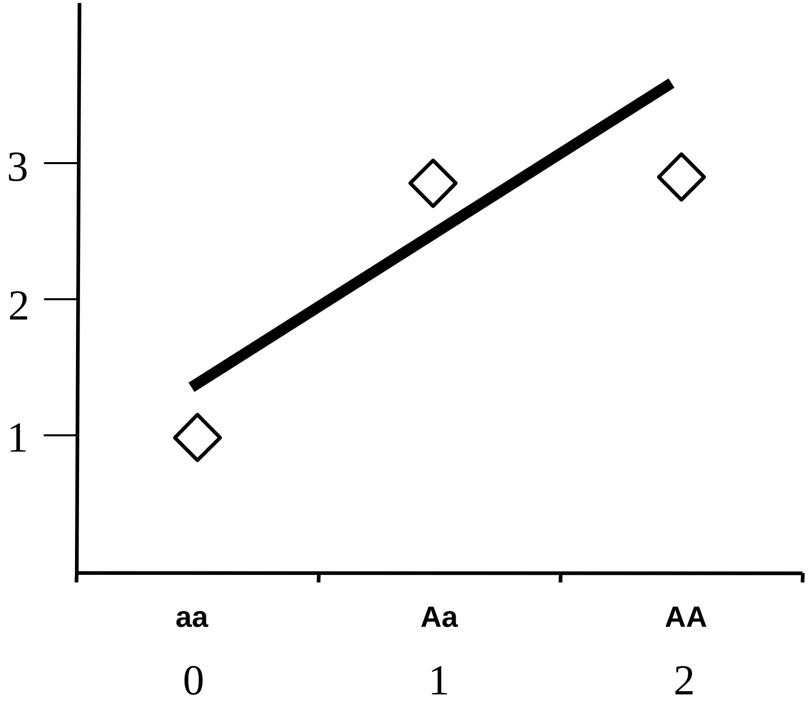
$$\sigma^2_A = 2 f_A f_a [f_A (Y_{AA} - Y_{Aa}) + f_a (Y_{Aa} - Y_{aa})]^2$$

$$\sigma^2_D = f^2_A f^2_a (Y_{AA} - 2Y_{Aa} - Y_{aa})^2$$

Partitioning the Variance One Locus



Additive Model



Dominant Model

- Additive Genetic Variance is variance explained by regression
- Dominance variance is residual variance not explained by regression

Least Squares Regression One Locus

- Represent genotypes of each individual by indicator variables:

	Additive Coefficient	Additive and Dom Coefficient	
Genotype	X_1	X_1	Z_1
aa	-1	-1	-½
Aa	0	0	½
AA	1	1	-½

$$Y = \mu + a_1 X_1 + d_1 Z_1 + \varepsilon$$

- Fit by least squares (or maximum likelihood)
- Can provide tests of significance
- Partitions data into variance components

Genetic Variance – Two Loci

	AA	Aa	aa
BB	$\mathrm{Y}_{\mathrm{AABB}}$ f_{AABB}	$\mathrm{Y}_{\mathrm{AaBB}}$ f_{AaBB}	$\mathrm{Y}_{\mathrm{aaBB}}$ f_{aaBB}
Bb	$\mathrm{Y}_{\mathrm{AABb}}$ f_{AABb}	$\mathrm{Y}_{\mathrm{AaBb}}$ f_{AaBb}	$\mathrm{Y}_{\mathrm{aaBb}}$ f_{aaBb}
bb	$\mathrm{Y}_{\mathrm{AAbb}}$ f_{AAbb}	$\mathrm{Y}_{\mathrm{Aabb}}$ f_{Aabb}	$\mathrm{Y}_{\mathrm{aabb}}$ f_{aabb}

$$A = \{AA, Aa, aa\}$$

$$B = \{BB, Bb, bb\}$$

$$\mu = \sum_{i \in A} \sum_{j \in B} f_{ij} Y_{ij}$$

$$\sigma_G^2 = \sum_{i = A} \sum_{j \in B} f_{ij} (Y_{ij} - \mu)^2$$

$$\sigma_I^2 = \sigma_G^2 - \sigma_{A+D}^2$$

Components of Variance for a Two Locus Model

Additive genetic variance Locus 1

Additive genetic variance Locus 2

Dominance genetic variance Locus 1

Dominance genetic variance Locus 2

Additive x Additive genetic variance

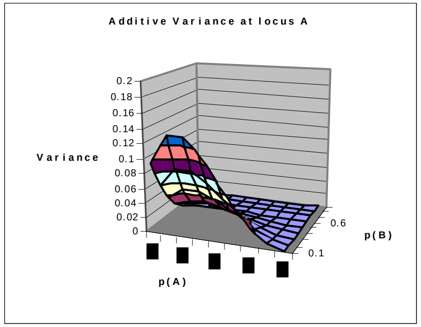
Additive x Dominance genetic variance

Dominance x Additive genetic variance

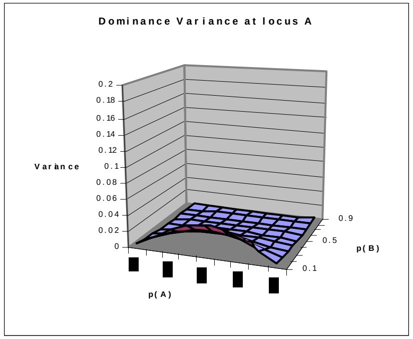
Dominance x Dominance genetic variance

"Epistatic" Variance

Additive Variance A

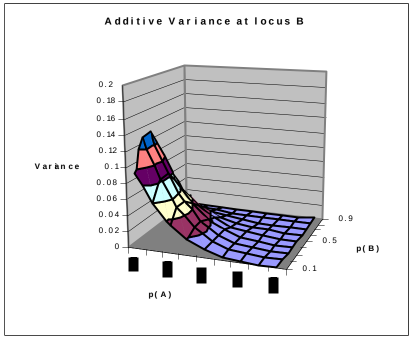


Dominance Variance A

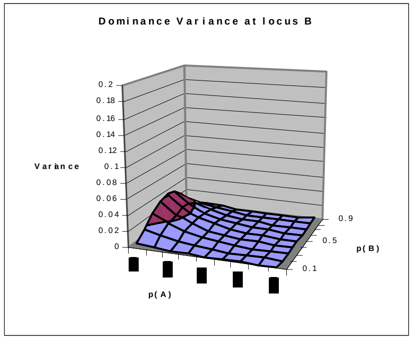


	BB	Bb	bb
AA	0	0	0
Aa	0	0	0
aa	0	0	1

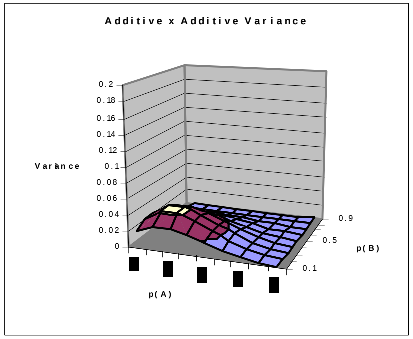
Additive Variance B



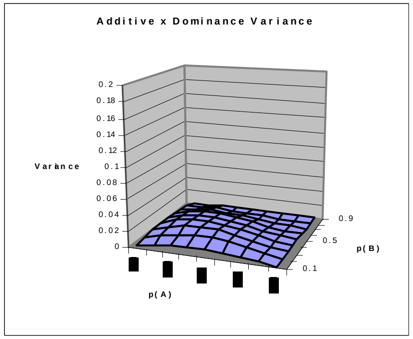
Dominance Variance B



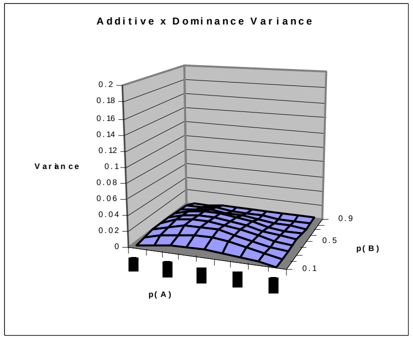
Additive x Additive



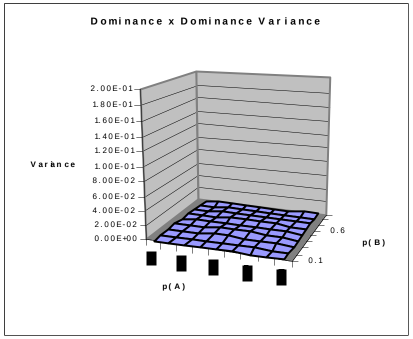
Additive x Dominant



Dominant x Additive



Dominant x Dominant



Least Squares Regression Two Loci

$$Y = \mu + a_1 X_1 + a_2 X_2 + d_1 Z_1 + d_2 Z_2 + i_{aa} w_{aa} + i_{ad} w_{ad} + i_{da} w_{da} + i_{dd} w_{dd} + \varepsilon$$

$$x_1 = \begin{cases} 1 \text{ if } AA \\ 0 \text{ if } Aa \\ -1 \text{ if } aa \end{cases} \qquad x_2 = \begin{cases} 1 \text{ if } BB \\ 0 \text{ if } Bb \\ -1 \text{ if } bb \end{cases}$$

$$z_1 = \begin{cases} \tfrac{1}{2} \text{ if } Aa \\ -\tfrac{1}{2} \text{ otherwise} \end{cases} \qquad z_2 = \begin{cases} \tfrac{1}{2} \text{ if } Bb \\ -\tfrac{1}{2} \text{ otherwise} \end{cases}$$

$$w_{aa} = x_1 \times x_2 \qquad w_{ad} = x_1 \times z_2 \qquad w_{da} = z_1 \times x_2 \qquad w_{dd} = z_1 \times z_2$$

- Can also formulate using logistic regression for dichotomous traits

Why model Epistasis in GWAS?

- Epistasis is important in model organisms

  c.f. Studies in Guinea fowl, yeast, drosophila

- Single locus tests will not always detect epistasis !!!

- Modeling epistasis may improve power to detect loci (?)

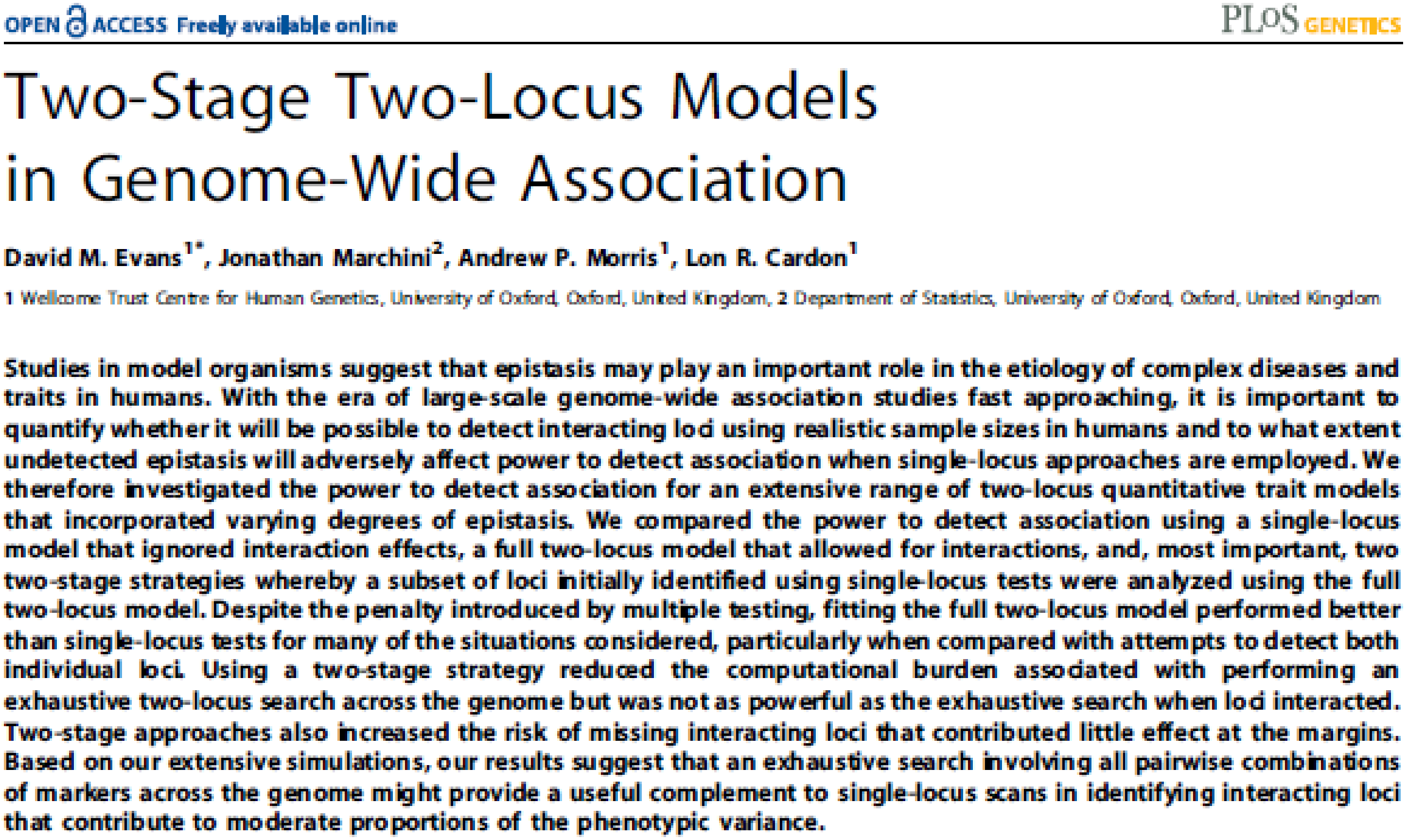
- Epistasis is ignored in human studies

  -Main effects hard enough to find!

  -Multiple testing problem: e.g. 100,000 markers gives a cutoff of $p = 1 \times 10^{-11}$!!!

  -Computational problems

  -Storage problems



OPEN ACCESS Freely available online

PLoS GENETICS

Two-Stage Two-Locus Models in Genome-Wide Association

David M. Evans[1*], Jonathan Marchini[2], Andrew P. Morris[1], Lon R. Cardon[1]

1 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, 2 Department of Statistics, University of Oxford, Oxford, United Kingdom

Studies in model organisms suggest that epistasis may play an important role in the etiology of complex diseases and traits in humans. With the era of large-scale genome-wide association studies fast approaching, it is important to quantify whether it will be possible to detect interacting loci using realistic sample sizes in humans and to what extent undetected epistasis will adversely affect power to detect association when single-locus approaches are employed. We therefore investigated the power to detect association for an extensive range of two-locus quantitative trait models that incorporated varying degrees of epistasis. We compared the power to detect association using a single-locus model that ignored interaction effects, a full two-locus model that allowed for interactions, and, most important, two two-stage strategies whereby a subset of loci initially identified using single-locus tests were analyzed using the full two-locus model. Despite the penalty introduced by multiple testing, fitting the full two-locus model performed better than single-locus tests for many of the situations considered, particularly when compared with attempts to detect both individual loci. Using a two-stage strategy reduced the computational burden associated with performing an exhaustive two-locus search across the genome but was not as powerful as the exhaustive search when loci interacted. Two-stage approaches also increased the risk of missing interacting loci that contributed little effect at the margins. Based on our extensive simulations, our results suggest that an exhaustive search involving all pairwise combinations of markers across the genome might provide a useful complement to single-locus scans in identifying interacting loci that contribute to moderate proportions of the phenotypic variance.

Epistasis in GWAS?

- What are the consequences of fitting two locus models when epistasis is absent?

  -“If it isn’t there, what happens if we go looking for it?”

- What are the consequences of fitting two locus models when epistasis is present?

  -“If it IS there, what happens if we go looking for it?”

- What are the consequences of fitting single locus models when epistasis is present?

  -“If it’s there, what happens when we ignore it?”

METHOD

▷ Simulate quantitative variable

-BOTH loci combined are responsible for 10%, 5%, 2% or 1% of total variance

-500, 1000 or 2000 individuals

-Assume 100,000 markers across the genome

-Perfect LD between marker and trait locus

-Comprehensive range of allele frequencies at both loci

-50 different models incorporating varying degrees of epistasis

MA	M1	M2	M3	M5	M7	M10	M11
1 ¾ ½ ¾ ½ ¼ ½ ¼ 0	0 0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 1 1	0 0 0 0 0 0 1 0 1	0 0 0 0 0 0 1 1 1	0 0 0 0 0 1 0 1 0	0 0 0 0 0 1 0 1 1

M12	M13	M14	M15	M16	M17	M18
0 0 0 0 0 1 1 0 0	0 0 0 0 0 1 1 0 1	0 0 0 0 0 1 1 1 0	0 0 0 0 0 1 1 1 1	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 1	0 0 0 0 1 0 0 1 0

M19	M21	M23	M26	M27	M28	M29
0 0 0 0 1 0 0 1 1	0 0 0 0 1 0 1 0 1	0 0 0 0 1 0 1 1 1	0 0 0 0 1 1 0 1 0	0 0 0 0 1 1 0 1 1	0 0 0 0 1 1 1 0 0	0 0 0 0 1 1 1 0 1

M30	M40	M41	M42	M43	M45	M56
0 0 0 0 1 1 1 1 0	0 0 0 1 0 1 0 0 0	0 0 0 1 0 1 0 0 1	0 0 0 1 0 1 0 1 0	0 0 0 1 0 1 0 1 1	0 0 0 1 0 1 1 0 1	0 0 0 1 1 1 0 0 0

M57	M58	M59	M61	M68	M69	M70
0 0 0 1 1 1 0 0 1	0 0 0 1 1 1 0 1 0	0 0 0 1 1 1 0 1 1	0 0 0 1 1 1 1 0 1	0 0 1 0 0 0 1 0 0	0 0 1 0 0 0 1 0 1	0 0 1 0 0 0 1 1 0

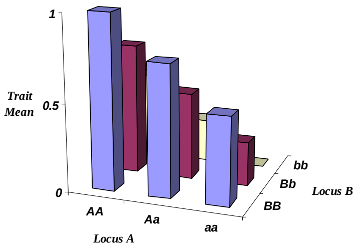
M78	M84	M85	M86	M94	M97	M98
0 0 1 0 0 1 1 1 0	0 0 1 0 1 0 1 0 0	0 0 1 0 1 0 1 0 1	0 0 1 0 1 0 1 1 0	0 0 1 0 1 1 1 1 0	0 0 1 1 0 0 0 0 1	0 0 1 1 0 0 0 1 0

M99	M101	M106	M108	M113	M114	M170	M186
0 0 1 1 0 0 0 1 1	0 0 1 1 0 0 1 0 1	0 0 1 1 0 1 0 1 0	0 0 1 1 0 1 1 0 0	0 0 1 1 1 0 0 0 1	0 0 1 1 1 0 0 1 0	0 1 0 1 0 1 0 1 0	0 1 0 1 1 1 0 1 0

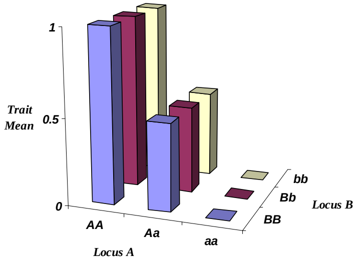
Evans et al. (2006)
PLOS Genet

CONTROL

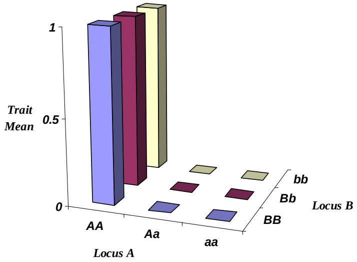
Additive



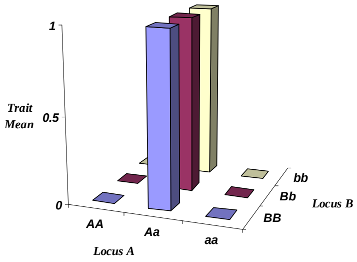
Additive Locus A



Dominant Locus A

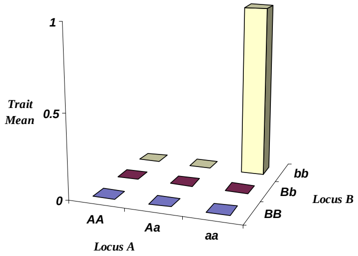


Heterozygote Advantage Locus A

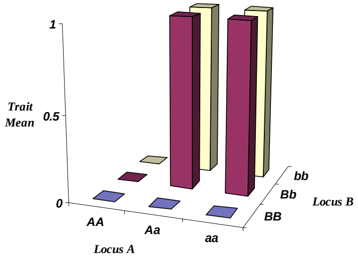


LESS FREAKY

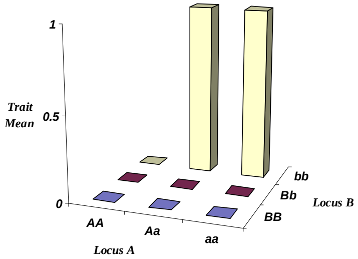
Complimentary Gene Action



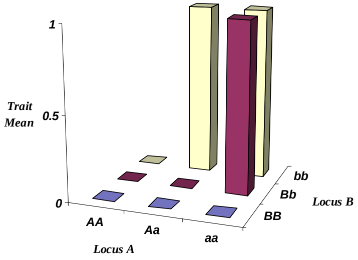
Dominant x Dominant Complimentary Gene Action



Dominant x Recessive Complimentary Gene Action

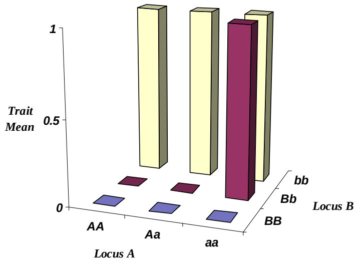


Threshold Model

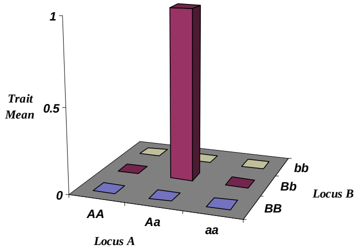


WAY FREAKY

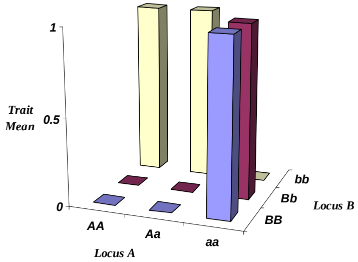
Modifying Effects



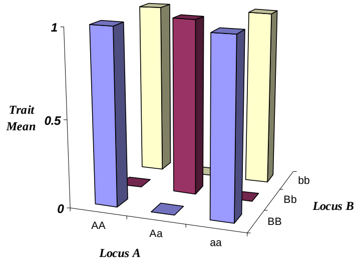
D x D Epistasis



XOR



Chequer Board



Simulating Genetic Data

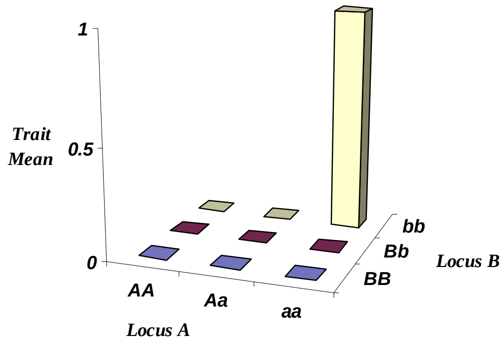
- Allele frequencies

-Allele frequencies at each locus will determine frequency of genotypes

$$p_A = 0.8 \quad p_a = 0.2 \quad \rightarrow \quad p_{AA} = 0.64 \quad p_{Aa} = 0.32 \quad p_{aa} = 0.04$$

- Genetic Model

-Need to specify trait means for each genotype combination

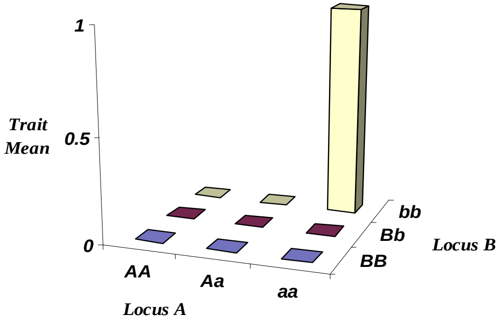

Complimentary Gene Action

-A random normal deviate can then be placed on these means to simulate the action of the environment

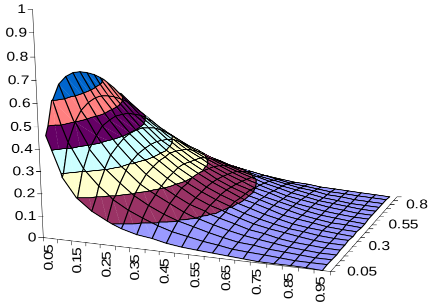
- Each combination of parameters will result in a unique variance profile

Simulating Genetic Data

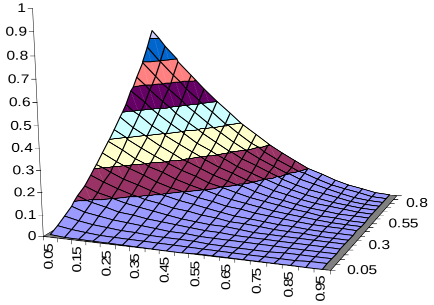
Complimentary Gene Action



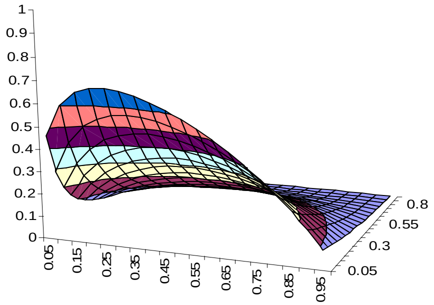
Additive Variance Locus 1



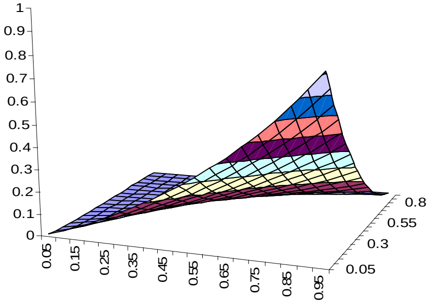
Dominance Variance Locus 1



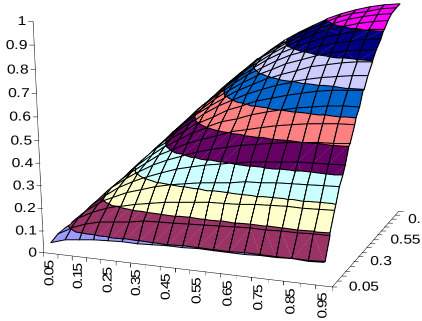
Additive Variance Locus 2



Dominance Variance Locus 2



Epistatic Variance Components



(Variance components shown for complimentary gene action model)

METHOD

- Quantify power to detect association

  -10,000 simulations for each combination of parameters

- Single locus test of association

  -Power to detect BOTH loci

  -Power to detect EITHER locus

- Two locus test of association

  -Power to detect BOTH loci

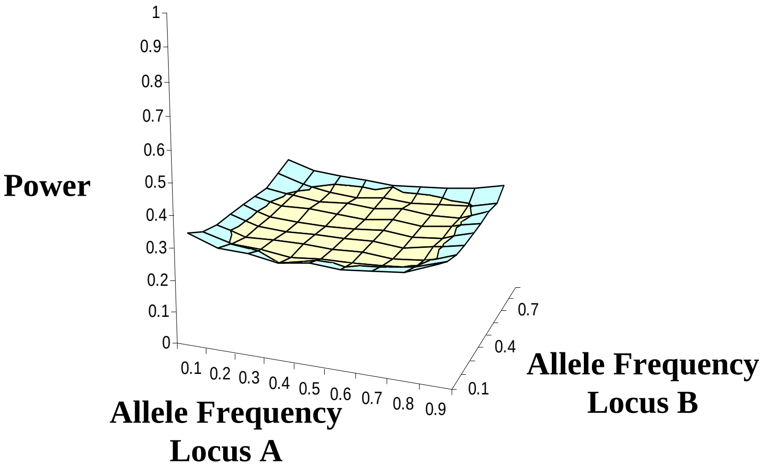
  -Different from power to detect the epistatic variance component explicitly

- All models fit via maximum likelihood

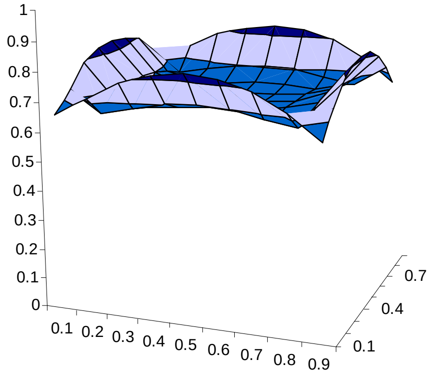
  -Significance assessed by minus two log-likelihood chi-square

Epistasis Isn’t There…

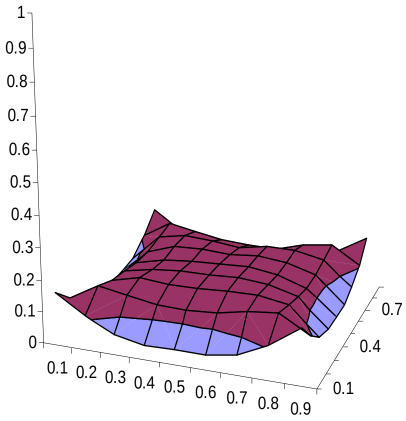
Two Locus Test



Power to Detect EITHER Locus
Single Locus Test



Power to Detect BOTH Loci
Single Locus Test



(Simulations represent 500 individuals, 10% genetic variance, Additive Model)

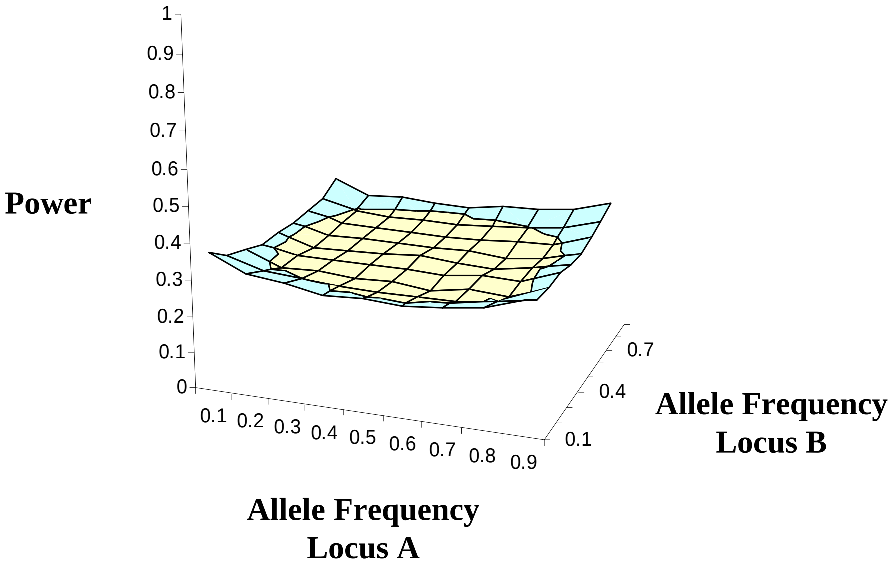
- The power to detect EITHER locus is greater using the single locus test when epistasis is NOT present

- The power to detect BOTH loci is actually less using single locus tests even when epistasis is NOT present

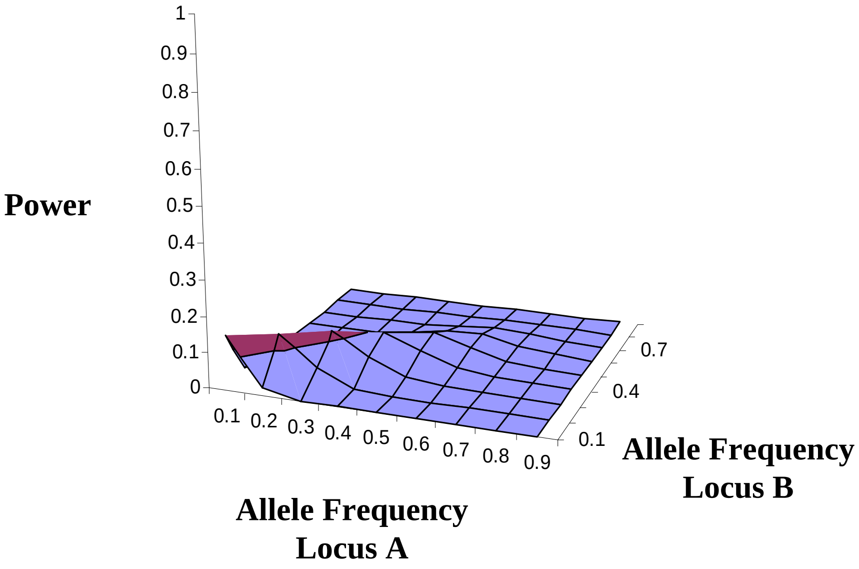
Epistasis IS There…

- The power to detect BOTH loci is always better using the two locus test when epistasis is present

Two Locus Test



Power to Detect BOTH Loci
SINGLE Locus Tests

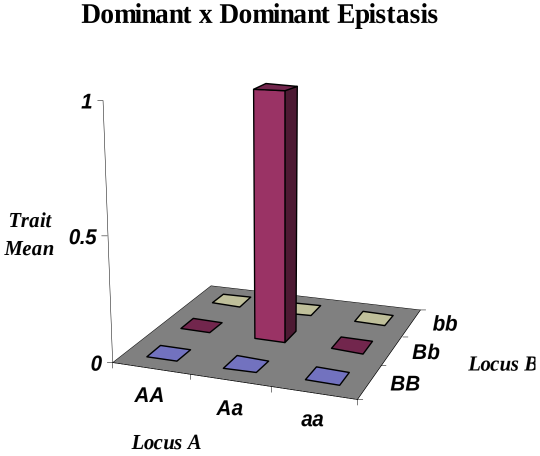


(Simulations represent 500 individuals, 10% genetic variance, Dominant x Dominant Complimentary Gene Action Model)

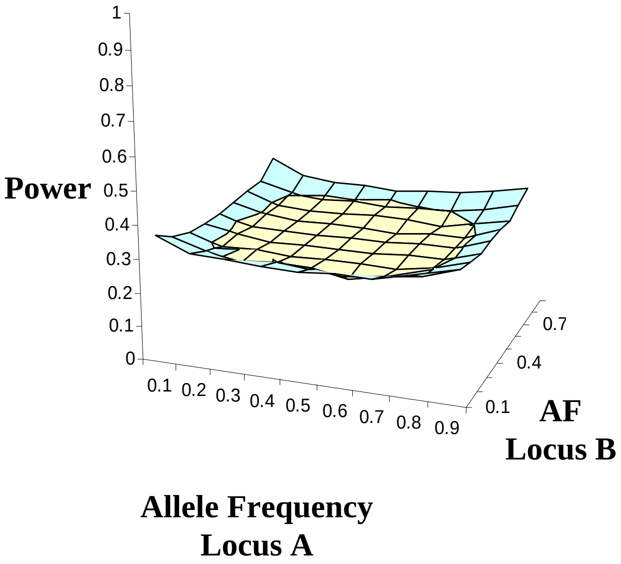
Power To Detect EITHER Locus

- When the model is EXTREME, the power of the two locus test is often better than the power to detect EITHER locus using single locus tests

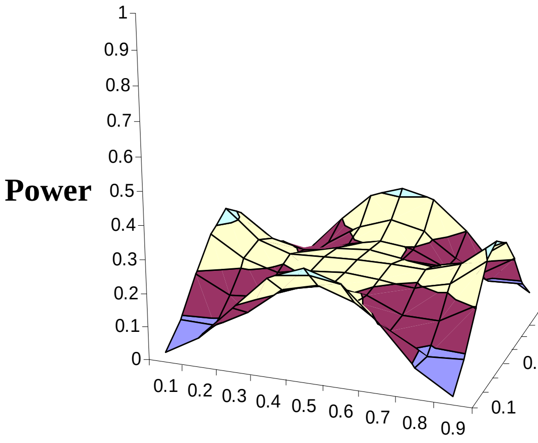
Model



Two Locus Test



Power to detect EITHER locus SINGLE Locus Tests



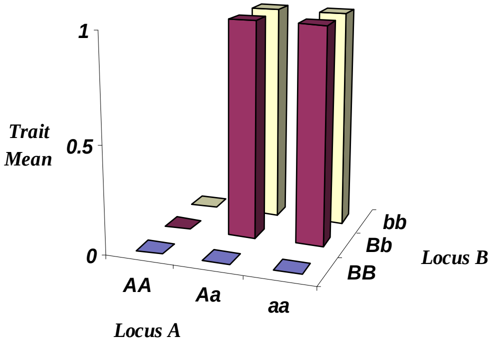
(500 individuals, 10% genetic variance)

Power to Detect EITHER Locus

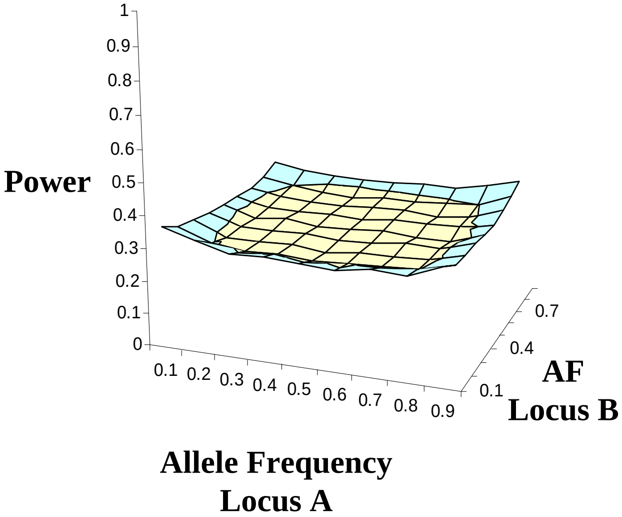
- When the model is less extreme, the power to detect EITHER locus using the single locus tests is often better than the two locus test

Model

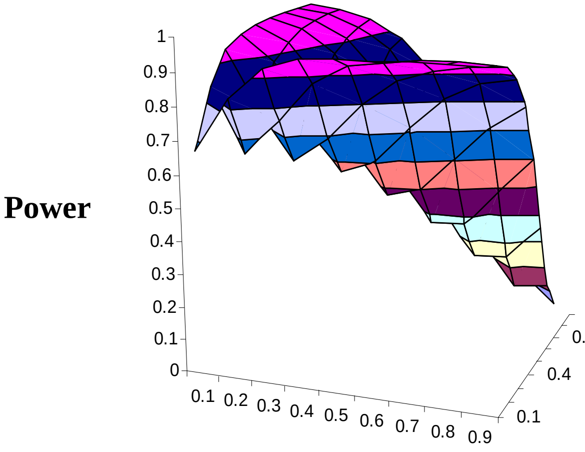
Dominant x Dominant Complimentary Gene Action Model



Two Locus Test



Power to detect EITHER locus SINGLE Locus Tests



(500 individuals, 10% genetic variance)

- The power to detect BOTH loci is always better fitting the two locus model regardless of the underlying model

- There are situations where fitting the full two-locus model will reveal effects which are not identified using single-locus methodology

- Multiple testing doesn’t kill you as much as you think!!!

Exhaustive or Two Stage Strategy?

- Idea: Two Stage Strategies

  -Test a subset of ${}^{100,000}C_2$ comparisons

  -Which comparisons are chosen depends on their performance in the single locus tests

- PLUS: Reduce cost due to multiple testing

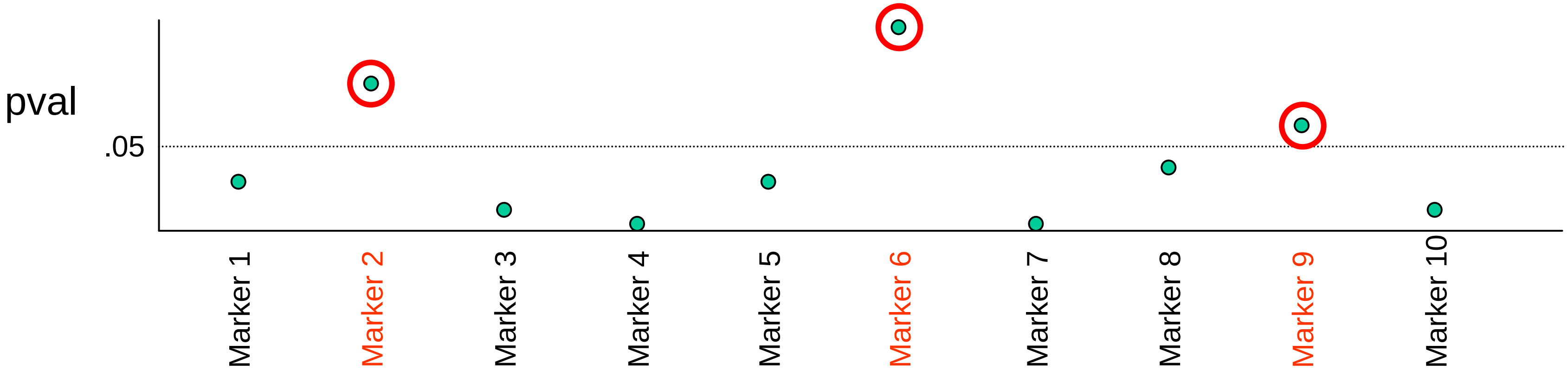
- MINUS: Throw away some comparisons which would be significant in the two locus test, yet are not significant in the single locus tests

Two-stage Models

▷ Strategy One

-Only markers which pass first stage threshold in the single locus analysis are tested

-All pair-wise combinations of these markers are tested



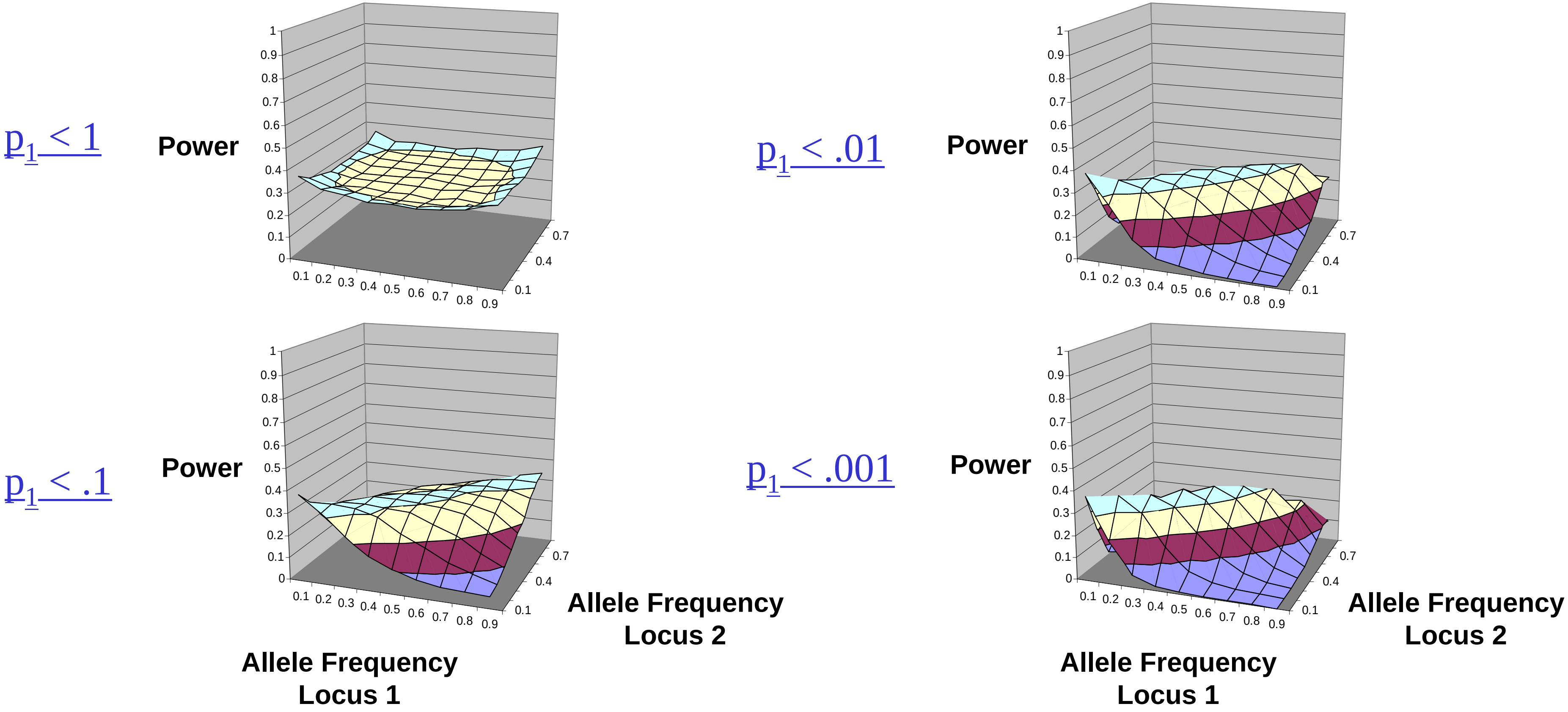
-Three comparisons:

-Marker 2 vs Marker 6

-Marker 2 vs Marker 9

-Marker 6 vs Marker 9

Two Stage Procedure: Strategy One



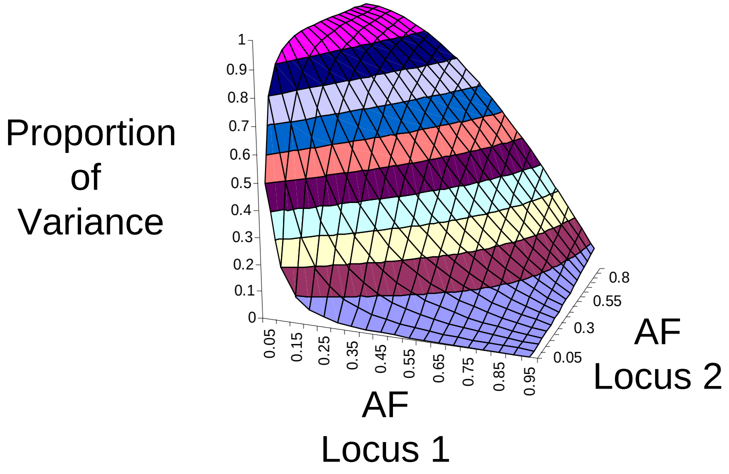
(Dominant x Dominant Complimentary Gene Action Model, 500 individuals, 10% genetic variance)

- As the first stage threshold becomes more stringent, the power to detect both loci DECREASES for the majority of the parameter space

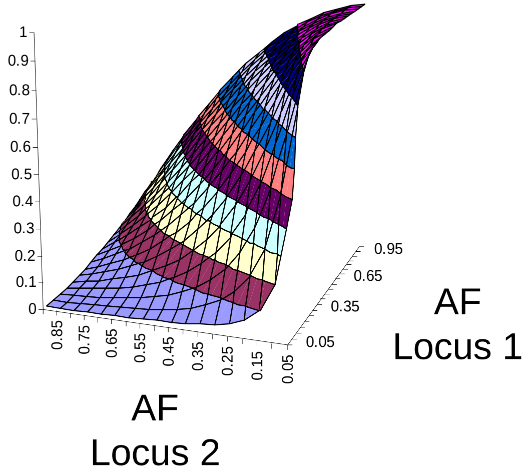
Why Does Strategy One Perform Poorly?

- For BOTH loci to be included in the second stage, Strategy One requires BOTH single locus tests to meet some threshold
- This threshold will not be met when the single locus variance is close to zero
- Therefore Strategy One will tend to fail whenever EITHER single locus component is close to zero

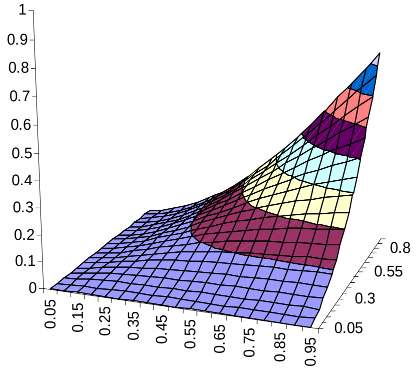
Variance Locus One



Variance Locus Two



Epistatic Variance

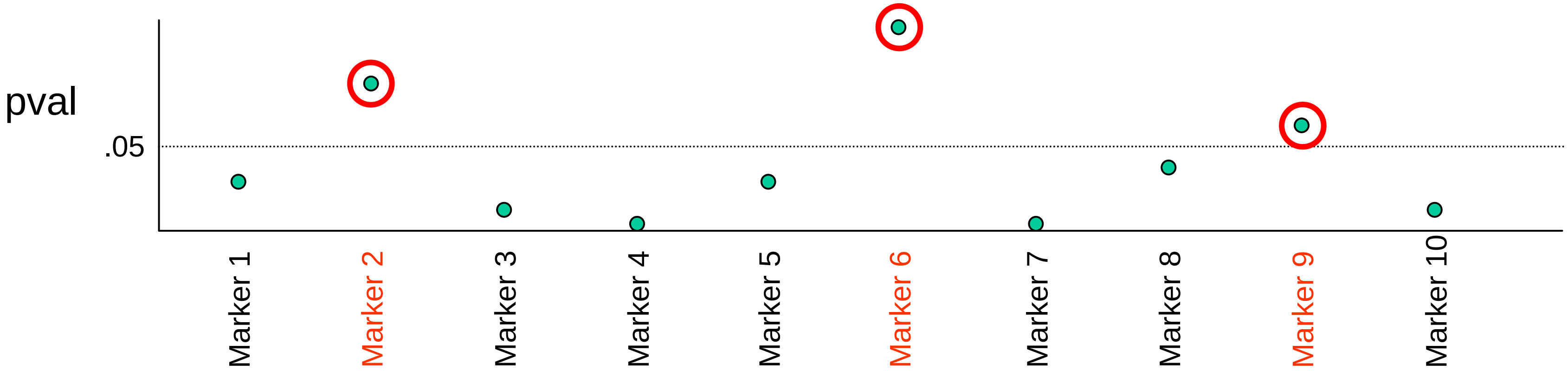


(Dominant x Dominant Complimentary Gene Action Model)

Two-stage Models

▷ Strategy Two

-Markers which pass first stage threshold are tested with ALL other markers



-24 comparisons:

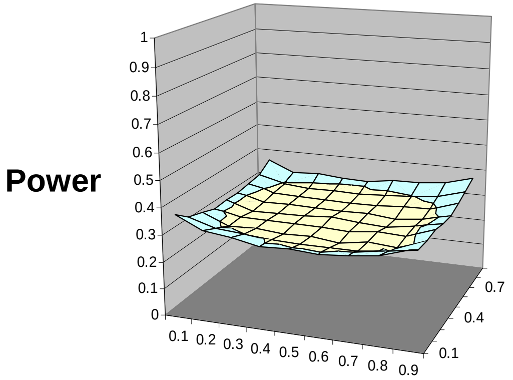
-Marker 2 vs Markers 1, 3, 4, 5, 6, 7, 8, 9, 10

-Marker 6 vs Marker 1, 3, 4, 5, 6, 7, 8, 9, 10

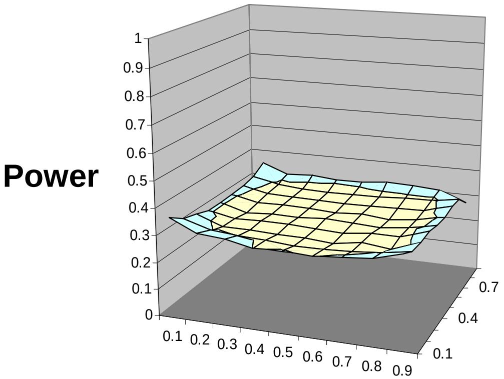
-Marker 9 vs Marker 1, 3, 4, 5, 7, 8, 10

Two-Stage Procedure: Strategy Two

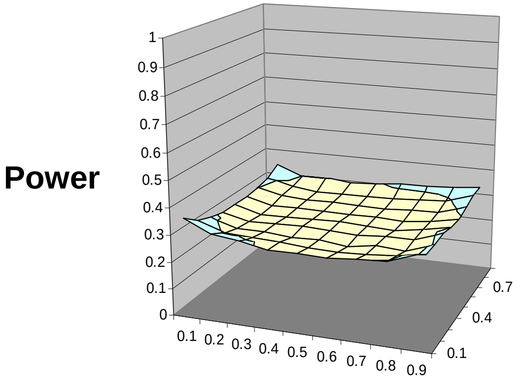
$p_1 < 1$



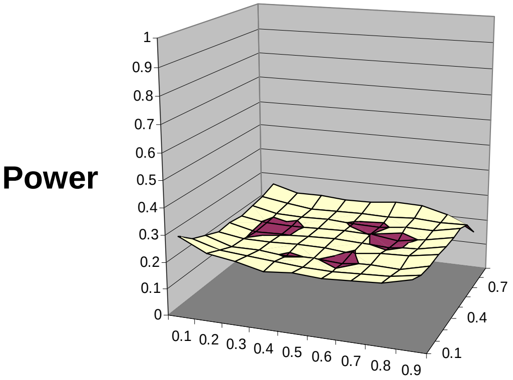
$p_1 < .001$



$p_1 < .01$



$p_1 < .0001$



(Dominant x Dominant Complimentary Gene Action Model, 500 individuals, 10% genetic variance)

- As the first stage threshold becomes more stringent, there is no increase in power
- There is a decrease in power at more stringent levels

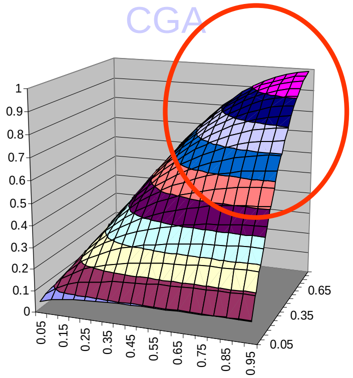
- Because of the need to condition on the first stage results being significant, there is no increased power in the second stage

- Since loci are included in the second stage if the variance is in EITHER single locus component, the strategy fails when the majority of variance is in the epistatic variance component or the first stage threshold is too severe

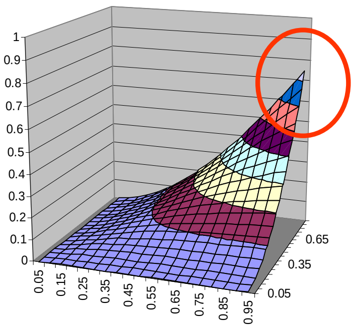
Less Extreme Models

Extreme Models

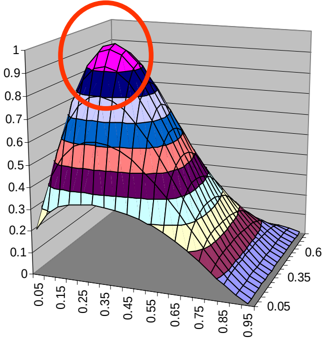
CGA



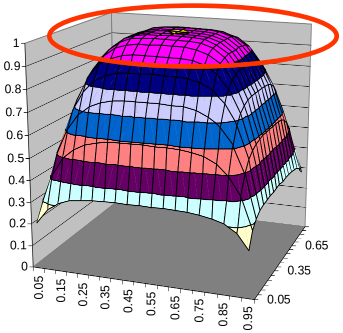
DxD CGA



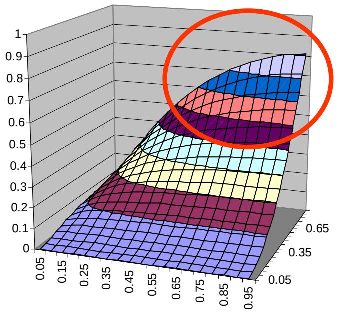
XOR



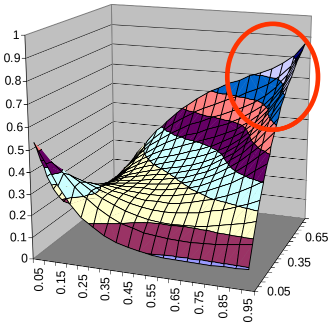
Chequer



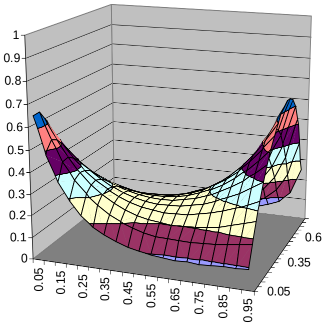
DxR CGA



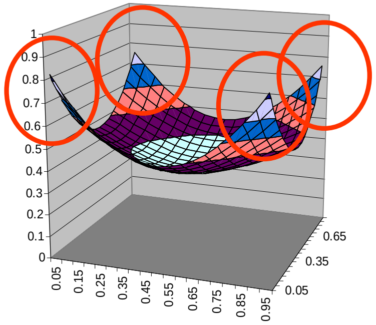
Threshold



ME



DxD



- Many simple looking models contain regions of their parameter space where loci would not be able to be identified using single locus analyses or two stage analyses

- An exhaustive search involving all pair-wise combinations of markers across the genome is superior to performing a two stage strategy

Conclusions

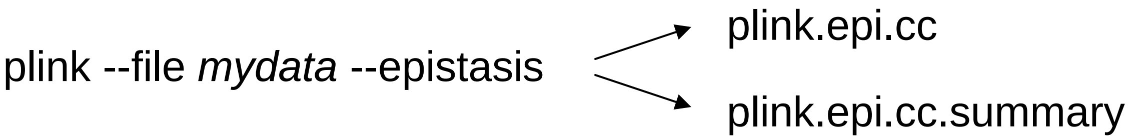
- An exhaustive search involving all pair-wise combinations of markers across the genome is superior to performing a two stage strategy
- Many simple looking models contain sizeable regions of their parameter space where loci would not be able to be identified using single locus analyses
- Despite the increased penalty due to multiple testing, it is possible to detect interacting loci which contribute to moderate proportions of the phenotypic variance with realistic sample sizes
- Is it worth incorporating epistasis in GWA?

Using PLINK to test for Epistasis

- Two tests of epistasis are implimented in PLINK

- Testing for additive x additive epistasis:

$$Y = \mu + a_1X_1 + a_2X_2 + i_{aa}w_{aa} + \varepsilon$$



plink --file *mydata* --epistasis

plink.epi.cc

plink.epi.cc.summary

- “Fast epistasis”:

```
plink --file mydata --fast-epistasis
```

- Possible to control output using the flags:

```
--epi1 0.0001

--epi2 0.0001
```

Practical

- Copy the files epistasis.ped and epistasis.map from H:/davide/LEUVEN2008

- Run single locus tests of association in this dataset

```
plink --file epistasis --assoc
```

- Run a scan for epistasis using the --fast-epistasis option

```
plink --file epistasis --fast-epistasis
```

- What are the two top interactions from this analysis?

- Are these loci flagged in the single locus analysis?

Practical

- Two locus results:

CHR1	SNP1	CHR2	SNP2	STAT	P
22	rs2014410	22	rs9607957	15.68	7.485e-005
22	rs2076672	22	rs2076109	15.72	7.348e-005

- Single locus results:

CHR	SNP	CHISQ	P	OR
22	rs2014410	0.02691	0.8697	1.01
22	rs9607957	3.333	0.06791	0.7318
22	rs2076672	3.071	0.07969	1.122
22	rs2076109	0.5072	0.4763	0.9587