Multifactor Dimensionality Reduction for Detecting Epistasis

William S. Bush
Marylyn Ritchie, PhD
Center for Human Genetics Research
Vanderbilt University

Overview

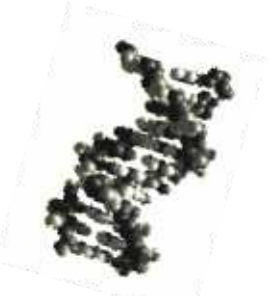
- The Data Structures of Genetics
- Epistasis
- Multifactor Dimensionality Reduction (MDR)
  - Power Analysis
  - Software and requirements
- Future Directions

Study of Proportions

Mendel's genetics paper published

Garrod coined the phrase "Inborn Errors of Metabolism"

Bateson, Lancefield, and Bridges discover "modifier" genes



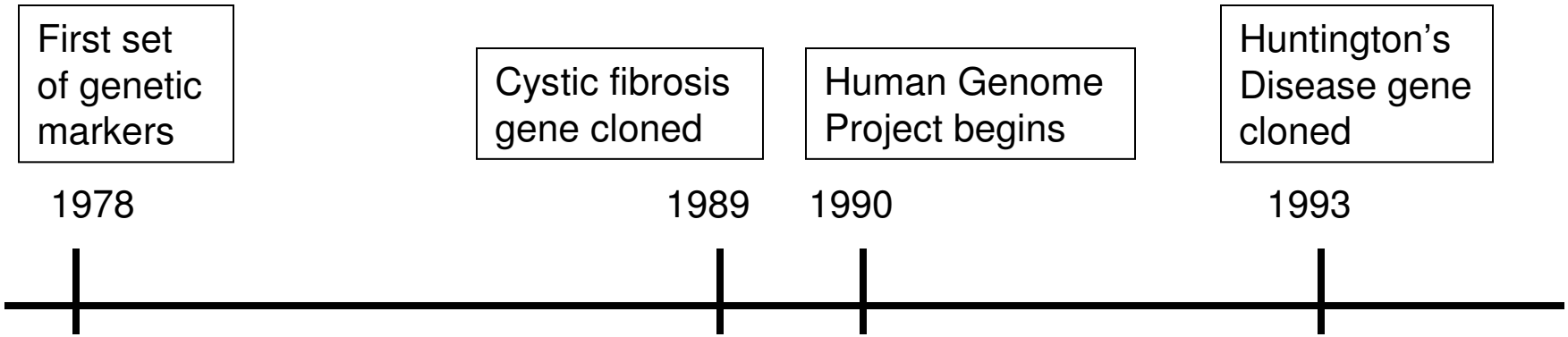
1865 — 1908 — 1910-1920 — 1953

DNA helix discovered

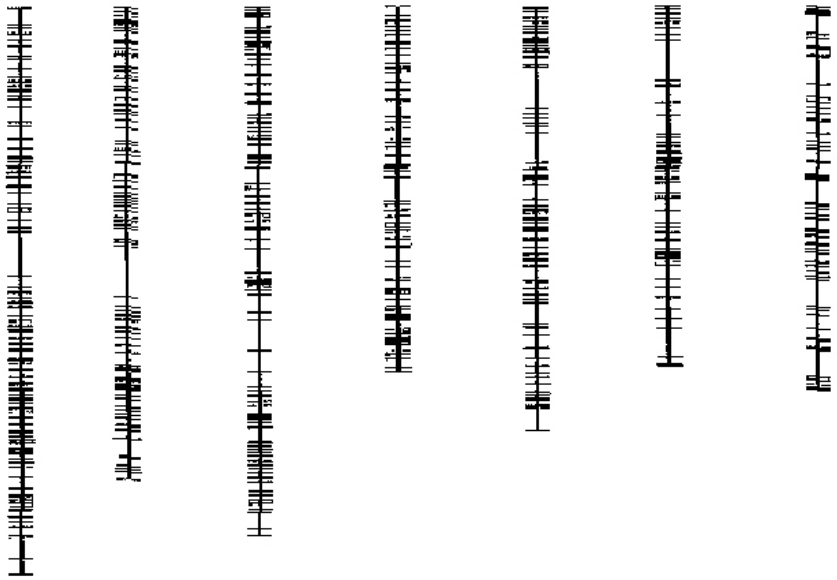
Data structure

	Experiment 1 Form of the Seed		Experiment 2 Color of the Albumen	
Plants	round	wrinkled	yellow	green
1	45	12	25	11
2	27	8	32	7
3	24	7	14	5
4	19	10	70	27
5	32	11	24	13
6	26	6	20	6
7	88	24	32	13
8	22	10	44	9
9	28	6	50	14
10	25	7	44	18

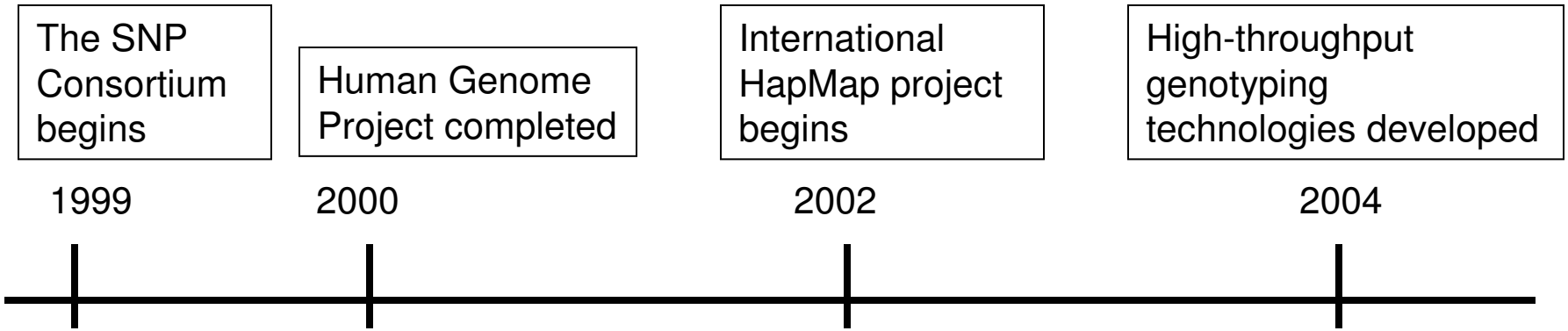
Study of Chromosomes



Data structure



Study of Common Variation



Data structure

```
0 2 1 2 2 1 1 2 2 2 2 1 1 0 1 2 0 1 0 2 1 1 2 1 2 1 0 0 1 2 1 0 1 2 0 1 2 2 1 0 1 1 2 0 1 1 2 1 2
0 1 1 0 2 2 2 1 1 0 2 1 1 2 0 1 0 1 1 1 0 2 0 1 2 1 2 1 1 0 2 1 1 0 1 1 2 1 2 1 2 1 0 1 1 0 1 1 2
0 2 2 2 2 0 0 0 2 2 1 2 2 1 1 2 0 1 2 0 1 1 1 2 1 2 2 1 1 0 0 0 0 2 2 2 2 1 2 1 1 0 1 2 1 0 2 2 0
1 1 1 2 2 2 2 1 2 0 2 2 1 2 2 1 2 1 0 0 0 2 0 0 0 1 2 2 1 1 2 2 2 0 1 1 0 1 0 2 2 2 2 0 1 2 1 1 0
1 1 1 1 2 1 1 0 0 2 1 1 0 2 1 2 1 2 2 2 0 2 1 2 2 0 1 1 0 1 1 1 2 1 2 1 2 2 0 2 2 0 1 2 0 2 1 1 2
1 1 0 1 2 1 2 1 2 1 1 1 2 2 1 0 2 2 1 2 1 1 2 1 2 1 1 2 0 0 1 1 1 1 1 1 2 1 2 2 0 0 0 2 0 1 1 2 2
0 1 1 0 0 0 2 1 2 1 1 2 0 2 1 1 1 1 0 0 0 2 0 1 2 2 1 1 1 1 1 0 2 2 2 2 2 2 1 1 1 2 1 1 0 1 2 0 2
0 1 0 2 1 1 2 1 0 0 1 2 1 1 2 2 0 2 2 2 0 0 1 0 1 1 2 2 1 2 2 1 1 2 0 1 2 2 2 2 0 0 2 1 1 1 1 0 0
0 1 1 0 1 1 1 0 0 1 1 1 2 2 1 2 2 2 2 2 1 1 2 1 1 1 2 1 1 1 1 1 1 2 1 0 1 0 1 1 1 2 1 1 1 1 2 1 1
1 1 2 2 1 1 0 1 0 1 0 1 0 1 0 2 1 1 0 1 1 1 1 0 2 2 1 0 2 2 1 1 1 1 1 1 1 1 1 1 0 1 2 2 0 0 2 1 1
0 1 1 1 1 1 2 2 0 1 2 0 2 1 1 2 1 0 1 0 1 2 2 1 2 2 2 2 1 2 0 1 1 1 1 1 0 1 1 1 2 2 1 0 0 2 1 2 2
1 2 2 1 2 1 0 0 1 2 0 0 2 1 1 0 0 2 1 1 0 1 1 0 2 2 1 1 2 2 0 1 2 0 2 1 1 2 2 2 1 0 2 0 1 1 1 2 2
```

William S. Bush

Epistasis

- **Epistasis** – two or more genes interacting in a non-additive manner to affect disease outcome; gene-gene interactions

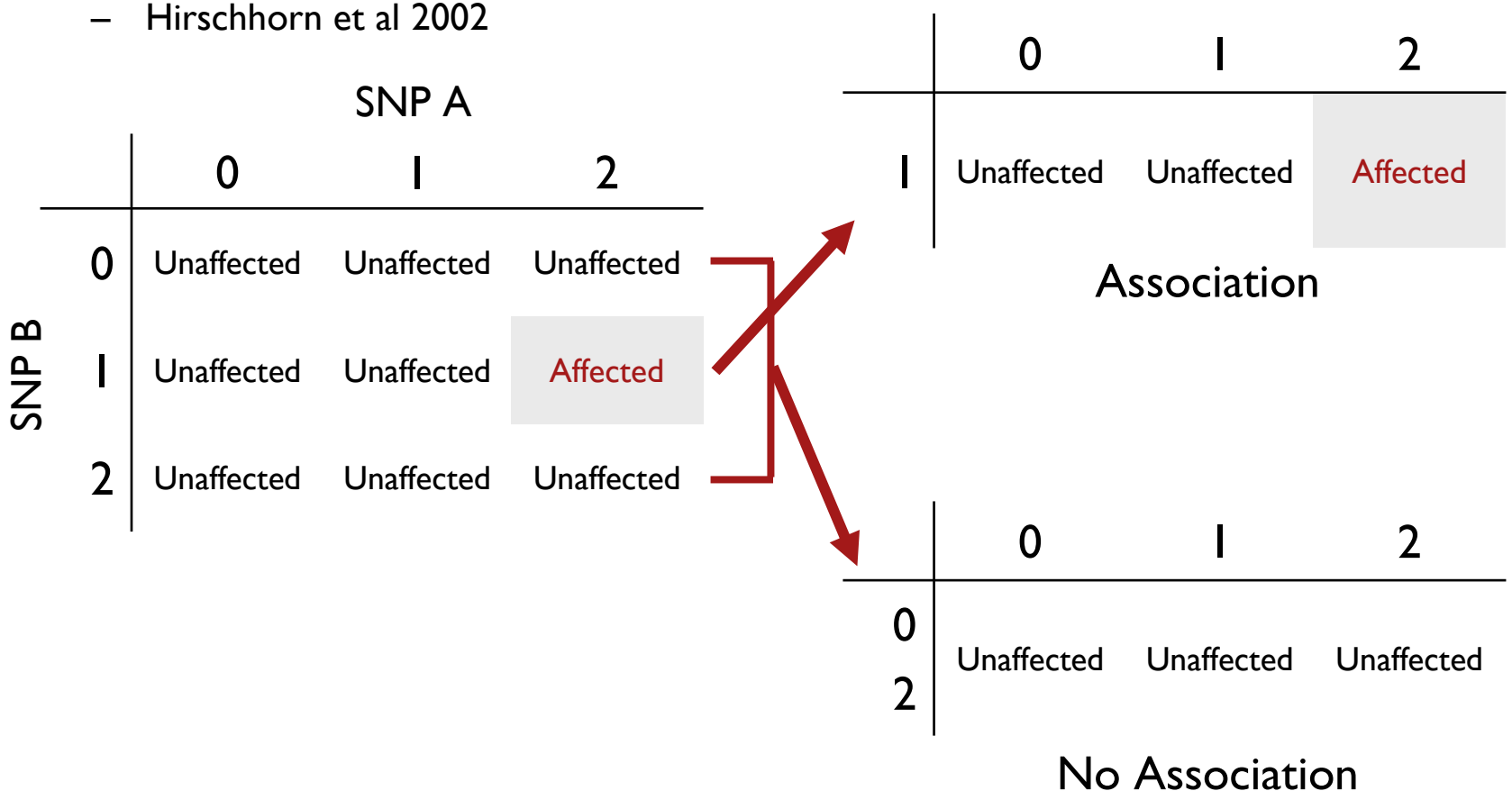
		SNP A		
		0	1	2
SNP B	0	0.01	0.01	0.01
	1	0.01	0.15	0.15
	2	0.01	0.15	0.50

		SNP A		
		0	1	2
SNP B	0	0.01	0.01	0.38
	1	0.01	0.18	0.01
	2	0.38	0.01	0.01

Frankel and Schork, *Nature Genetics* 14:371-373 (1996)

Epistasis

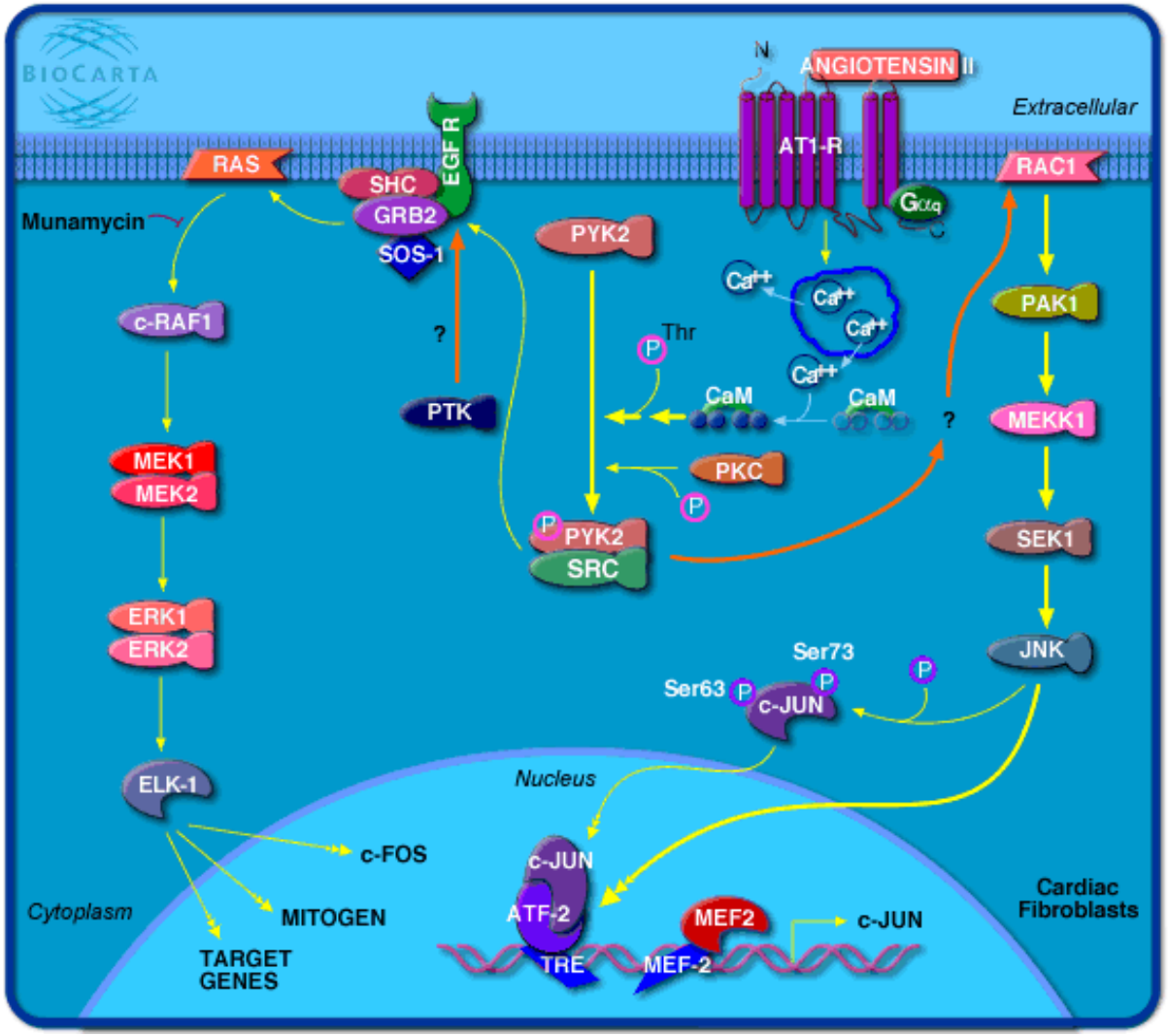
- Single locus studies do not replicate
  - Hirschhorn et al 2002



Epistasis

Renin-Angiotenisin System

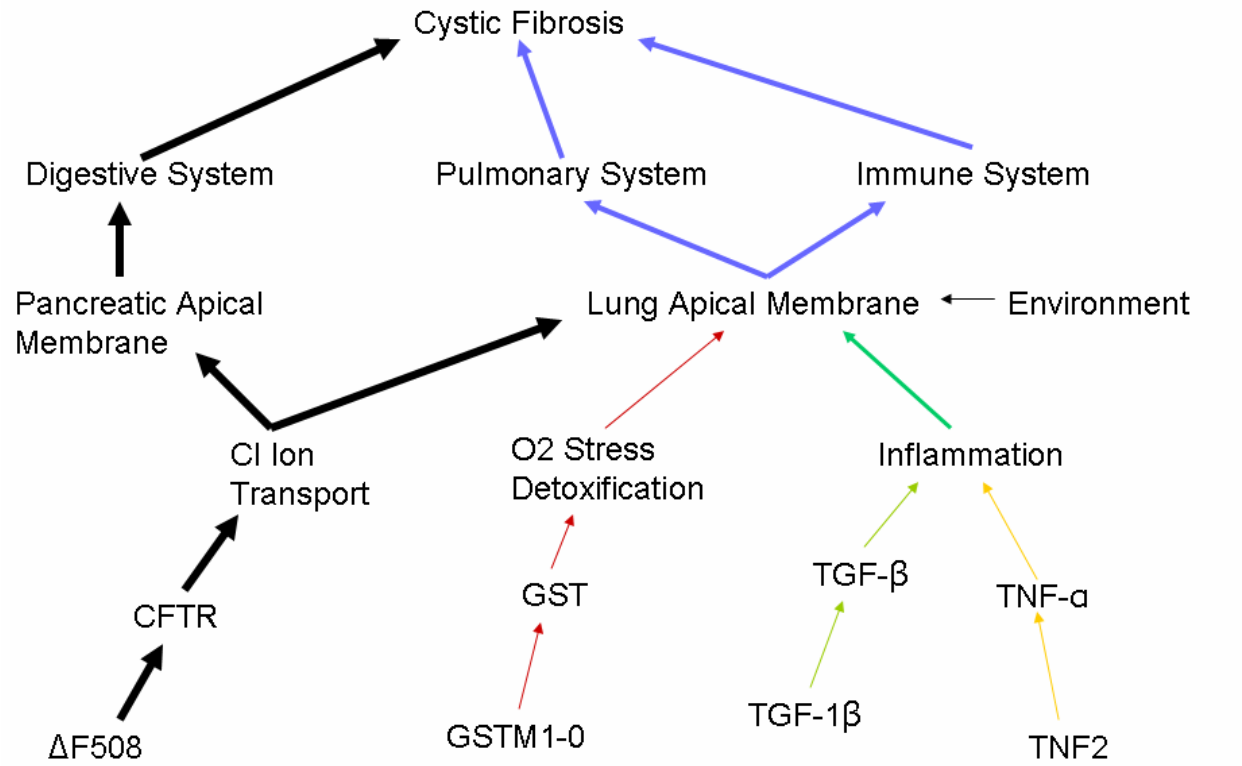
Biocarta



William S. Bush

Epistasis

- Mendelian single-gene disorders are now being considered complex traits with gene-gene interactions (modifier genes)



Traditional Statistical Approach

Association Analysis

- Typically one marker or SNP at a time to detect loci exhibiting main effects
- Follow-up with an analysis to detect interactions between the main effect loci
- Some studies attempt to detect pair-wise interactions even without main effects
- Higher dimensions are usually not possible with traditional methods

Traditional Statistical Approach

Association Analysis

- Logistic Regression
  - Small sample size can result in biased estimates of regression coefficients and can result in spurious associations (Cancato et al 1993)
  - Need at least 10 cases or controls per independent variable to have enough statistical power (Peduzzi et al 1996)
  - Curse of dimensionality is the problem (Bellman 1961)

Novel Computational Approaches

- Extensions to regression analysis
  - Automated Detection of Informative Combined Effects (DICE)
    - Tahri-Daizadeh N et al. *Genome Research* 13: 1952-1960 (2003)
  - Classification and Regression Trees (CART)/ Patterning and Recursive Partitioning (PRP)
    - Bastone L et al. *Human Heredity* 58: 82-92 (2004)
  - Logic Regression
    - Kooperberg et al. *Genetic Epidemiology* (S1): 626-631 (2001)
  - Penalized Logistic Regression
    - Zhu et al. *Biostatistics* 5: 427 - 443 2004
  - Multivariate adaptive regression splines (MARS)
    - Cook et al. Statistics in Medicine 23: 1439 – 1453 (2004)

Novel Computational Approaches

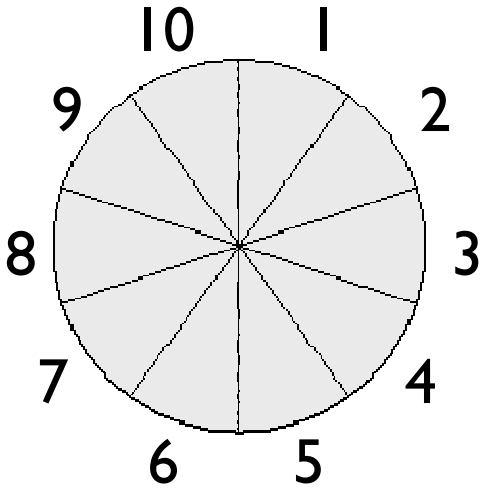
- Data reduction approaches
  - Combinatorial Partitioning Method (CPM)
    - Nelson et al., *Genome Research* 11:458-70 (2001)
  - Restricted Partitioning Method (RPM)
    - Culverhouse R, Klein T, Shannon W. *Genet Epidemiol.* 2004 27:141-52
  - Multifactor Dimensionality Reduction (MDR)
    - Ritchie et al., *American Journal of Human Genetics* 69:138-147 (2001)
  - Set Association
    - Hoh J, Wille A, and Ott J *Genome Research* 11: 2115-2119 (2001)

Multifactor Dimensionality Reduction (MDR)

Ritchie et al., *American Journal of Human Genetics* 69:138-147 (2001)
Ritchie MD, Hahn LW, and Moore JH, *Genetic Epidemiology* 24: 150-157. (2003)
Hahn LW, Ritchie MD, and Moore JH, *Bioinformatics* 19: 376-382. (2003)

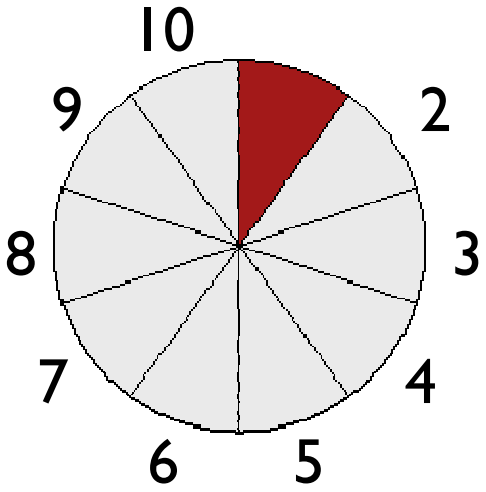
- Inspired by the Combinatorial Partitioning Method
- Model-free
- Non-parametric
- The goal of MDR is to detect gene-gene and gene-environment interactions in the presence and absence of main effects

Multifactor Dimensionality Reduction (MDR)



Affected Status	1	2	3	4	5	6	7	8	9	10
0	2	2	1	2	1	2	0	1	2	1
0	1	1	2	1	2	1	0	0	1	0
0	2	1	1	2	1	0	1	0	2	0
1	0	0	1	0	2	1	2	1	1	2
1	1	0	0	1	0	0	1	2	1	1
1	0	2	2	0	1	0	1	1	1	0
0	1	0	1	1	1	2	2	2	2	1
0	2	2	1	1	0	1	0	2	2	1
0	1	1	2	2	1	0	1	0	1	2
0	2	2	1	2	1	2	0	1	2	1
0	1	0	1	1	1	2	1	2	1	1
0	1	0	1	0	1	2	2	0	2	2
1	2	1	2	1	2	2	1	0	1	0
0	1	0	1	1	1	2	2	2	2	1

Multifactor Dimensionality Reduction (MDR)

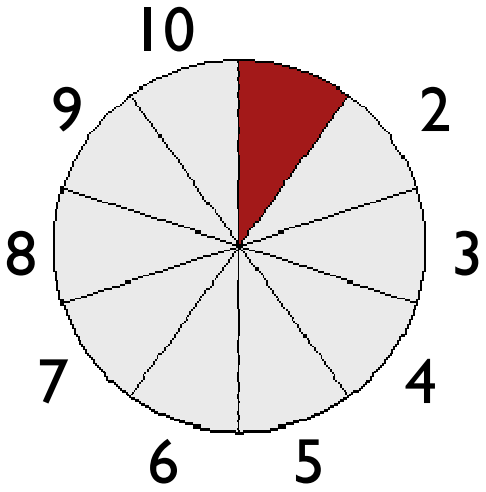


Variable Combinations

{ 1 2 }
{ 1 3 }
{ 1 4 }
{ 1 5 }
…

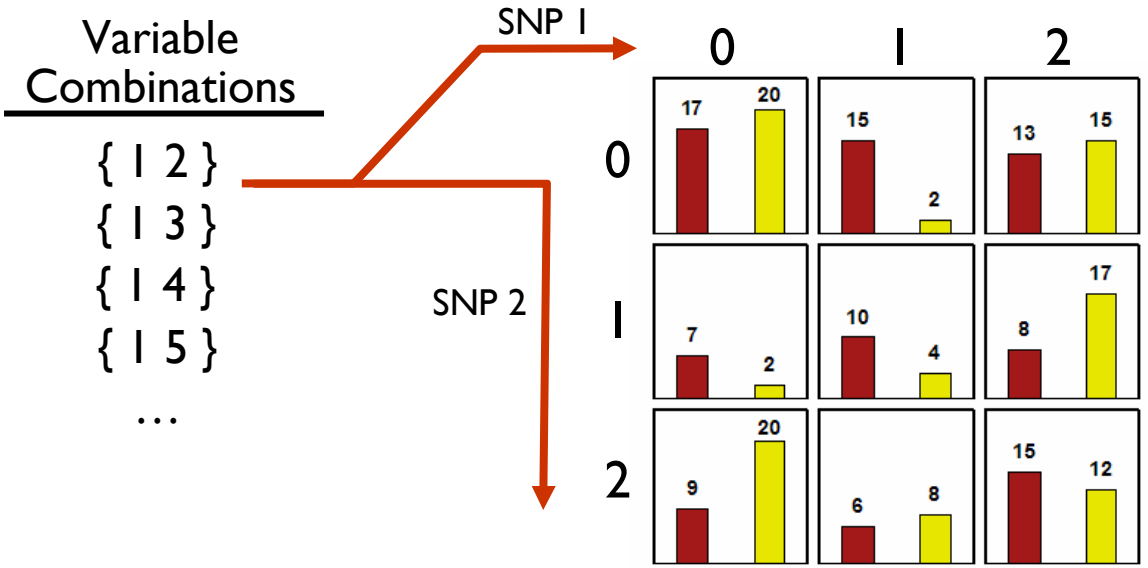
Affected Status	1	2	3	4	5	6	7	8	9	10
0	2	2	1	2	1	2	0	1	2	1
0	1	1	2	1	2	1	0	0	1	0
0	2	1	1	2	1	0	1	0	2	0
1	0	0	1	0	2	1	2	1	1	2
1	1	0	0	1	0	0	1	2	1	1

Multifactor Dimensionality Reduction (MDR)

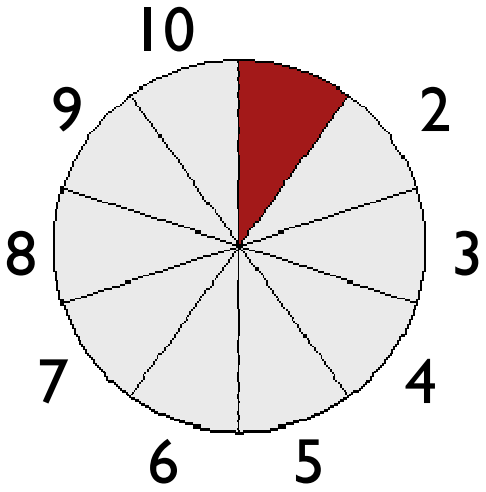


Variable Combinations

{ 1 2 }
{ 1 3 }
{ 1 4 }
{ 1 5 }
…

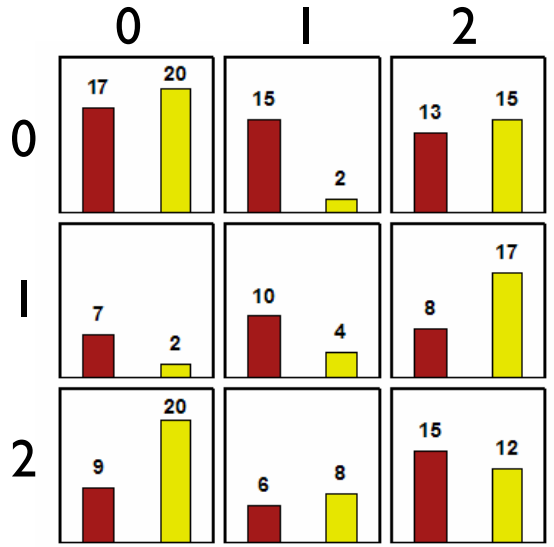


Multifactor Dimensionality Reduction (MDR)



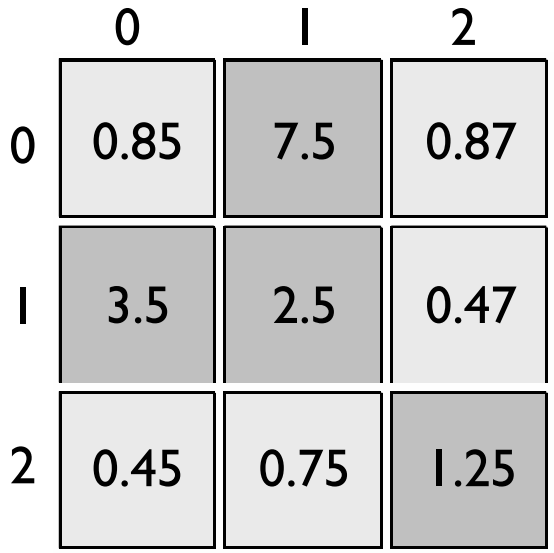
Variable Combinations

{ 1 2 }
{ 1 3 }
{ 1 4 }
{ 1 5 }
…

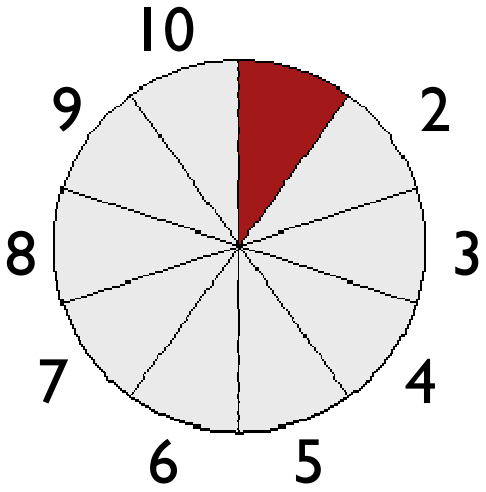


Models

Combination	Error
{ 1 2 }	**36.5%**

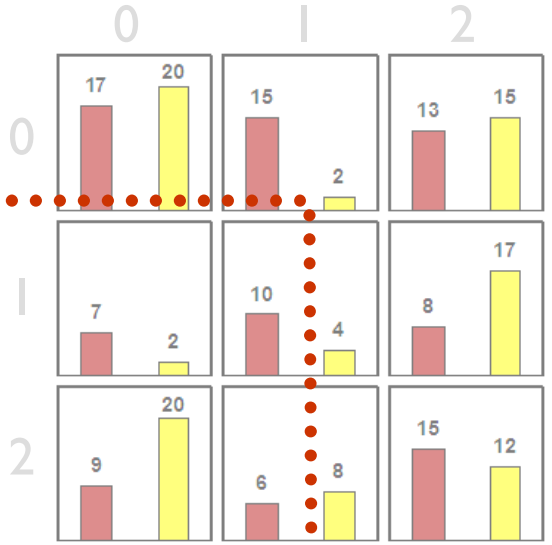


Multifactor Dimensionality Reduction (MDR)



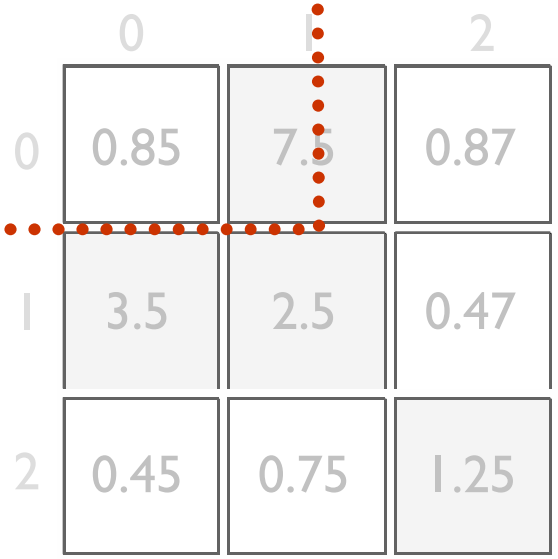
Variable Combinations

{ 1 2 }
{ 1 3 }
{ 1 4 }
{ 1 5 }
…



Models

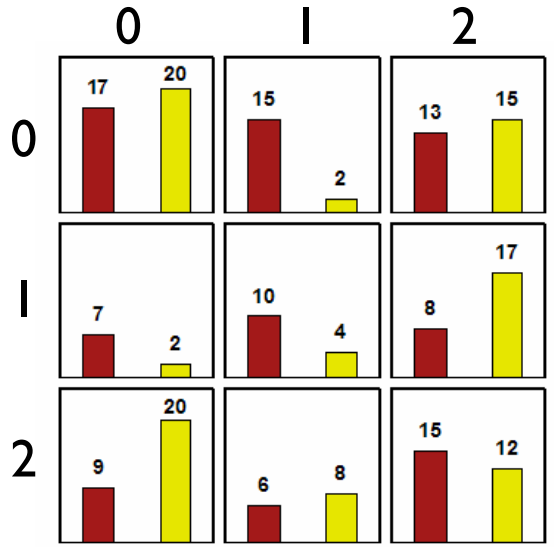
Combination	Error
{ 1 2 }	**36.5%**
{ 1 3 }	42.7%
{ 1 4 }	39.2%
{ 1 5 }	46.1%
…	
{ 3 7 }	38.0%



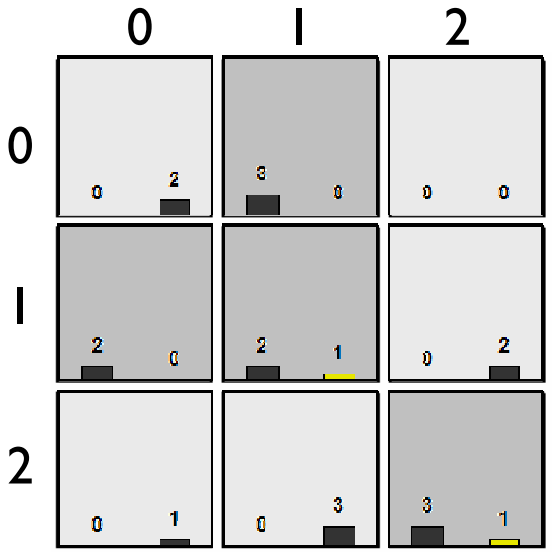
Multifactor Dimensionality Reduction (MDR)

Variable Combinations

{ 1 2 }
{ 1 3 }
{ 1 4 }
{ 1 5 }
…

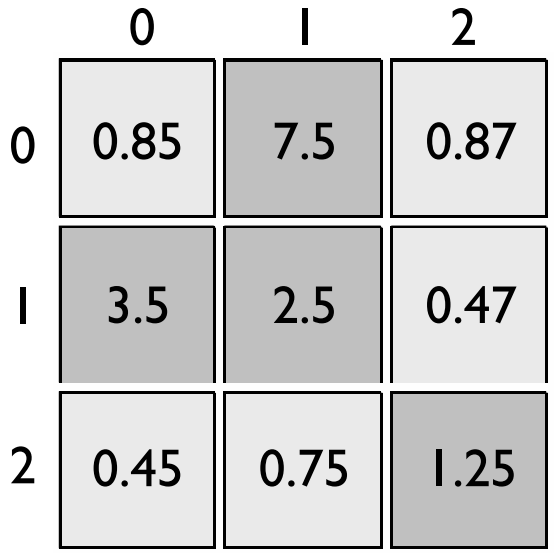


Combination	Prediction Error
{ 1 2 }	20.0%



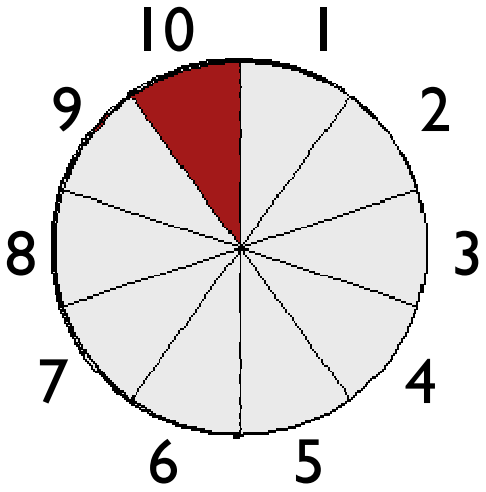
Models

Combination	Error
{ 1 2 }	36.5%
{ 1 3 }	42.7%
{ 1 4 }	39.2%
{ 1 5 }	46.1%
…	
{ 3 7 }	38.0%



William S. Bush

Multifactor Dimensionality Reduction (MDR)

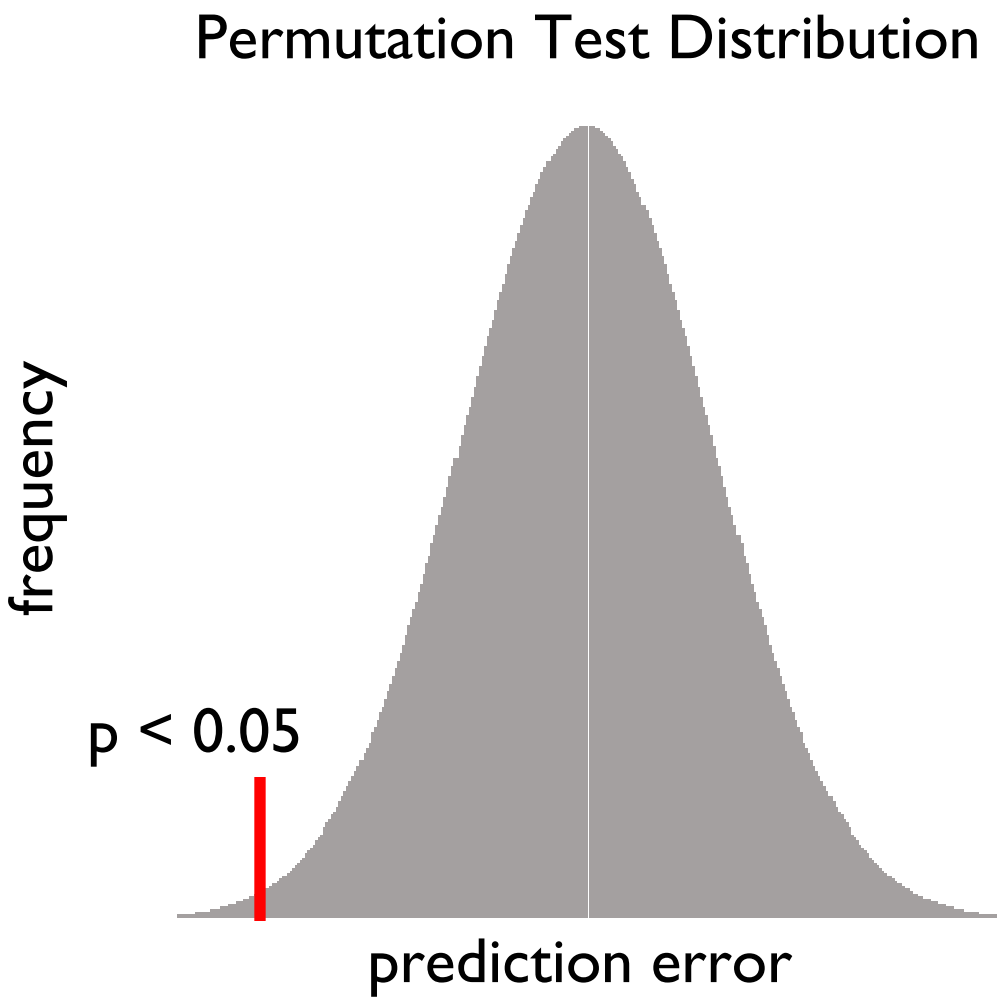


CV interval	Variables in Model	Classification error	Prediction error
1	{ 1 2 }	36.5	20.0
2	{ 1 2 }	35.4	37.41
3	{ 3 5 }	41.0	44.1
4	{ 1 2 }	34.7	28.67
5	{ 1 2 }	36.8	37.0
6	{ 1 2 }	38.2	37.2
7	{ 1 2 }	35.9	36.5
8	{ 1 2 }	36.1	28.1
9	{ 1 2 }	36.7	30.4
10	{ 1 2 }	37.1	32.5
CVC = 9		**AVG = 36.84**	**AVG = 33.19**

Multifactor Dimensionality Reduction (MDR)

Number of Variables	Variables in Model	Cross Validation Consistency	Classification Error Avg	Prediction Error Avg
1	{ 2 }	4	43.24	46.51
2	{ 1 2 }	9	36.84	33.19
3	{ 1 2 7 }	4	31.67	36.51
4	{ 1 2 7 9 }	3	28.32	42.67
5	{ 1 2 6 7 9 }	5	21.99	44.19

Multifactor Dimensionality Reduction (MDR)



MDR Power Study

Question : What is the power of MDR for detecting gene-gene interactions in the presence of common sources of error?

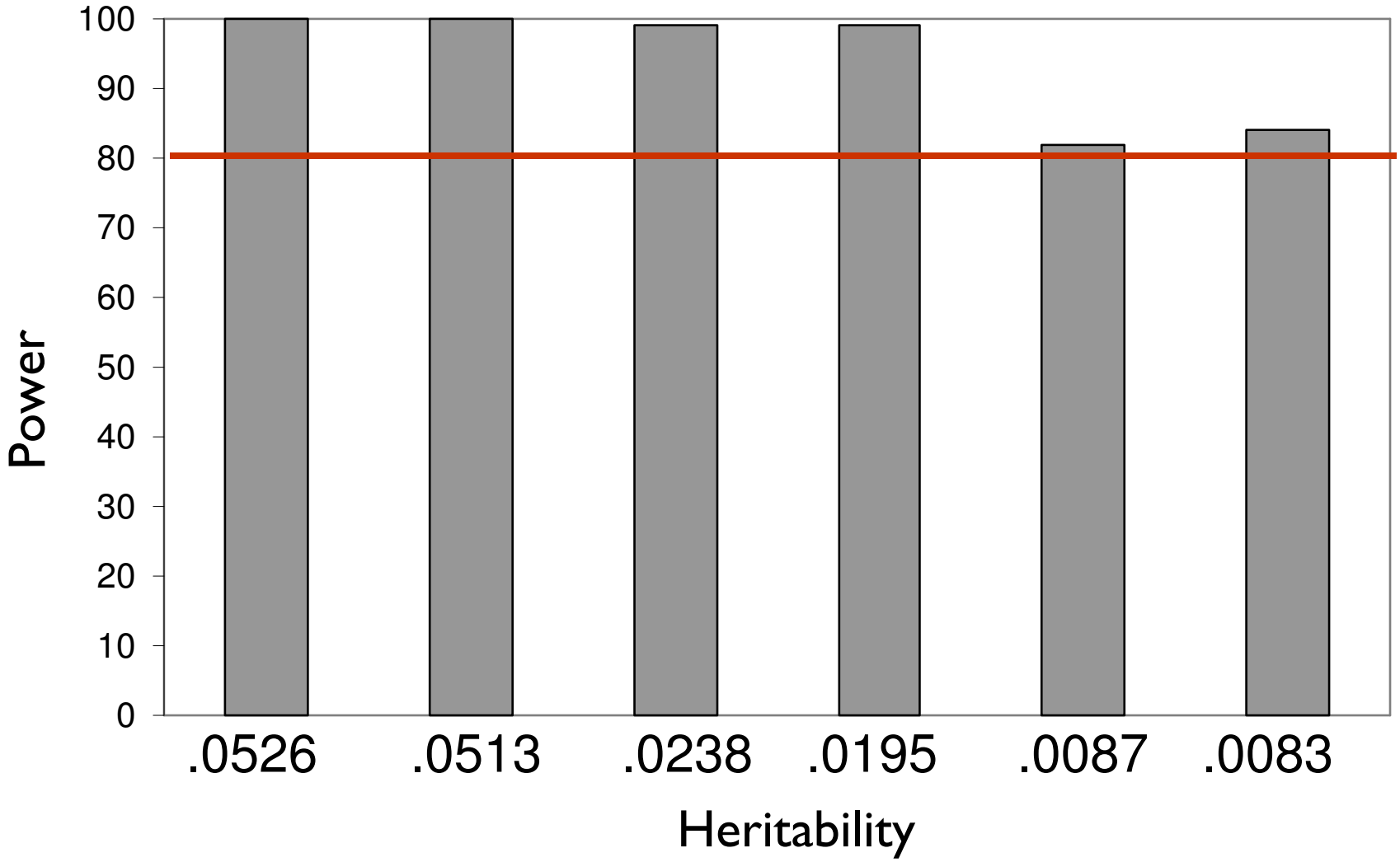
MDR Power Study

Ritchie MD, Hahn LW, and Moore JH, *Genetic Epidemiology* 24: 150-157. (2003)

- Simulated six different epistasis models in the absence and presence of sources of error
  - Genotyping Error (5%)
  - Phenocopy (50%)
  - Missing Data (5%)
  - Genetic Heterogeneity (50%)
- Evaluate the power of MDR to detect gene-gene interactions in the presence of error

MDR Power Study

Ritchie MD, Hahn LW, and Moore JH, *Genetic Epidemiology* 24: 150-157. (2003)



MDR Power Study

Ritchie MD, Hahn LW, and Moore JH, *Genetic Epidemiology* 24: 150-157. (2003)

SOURCE OF NOISE	MODEL 1	MODEL 2	MODEL 3	MODEL 4	MODEL 5	MODEL 6
NONE	100	100	99	99	82	84
GE	100	100	100	97	80	92
GH	3	41	2	3	4	4
PC	90	99	45	32	30	32
MS	100	100	99	97	82	87
GE + GH	4	41	2	3	4	6
GE + PC	94	99	41	48	28	33
GE + MS	100	100	98	98	74	84
GH + PC	0	1	0	0	0	0
GH + MS	5	38	0	2	4	6
PC + MS	96	99	42	43	14	16
GE + GH + PC	1	1	0	0	0	0
GE + GH + MS	6	34	2	1	3	7
GH + PC + MS	0	0	0	0	0	0
GE + PC + MS	94	100	48	42	18	16
GE + GH + PC + MS	0	1	0	1	0	0

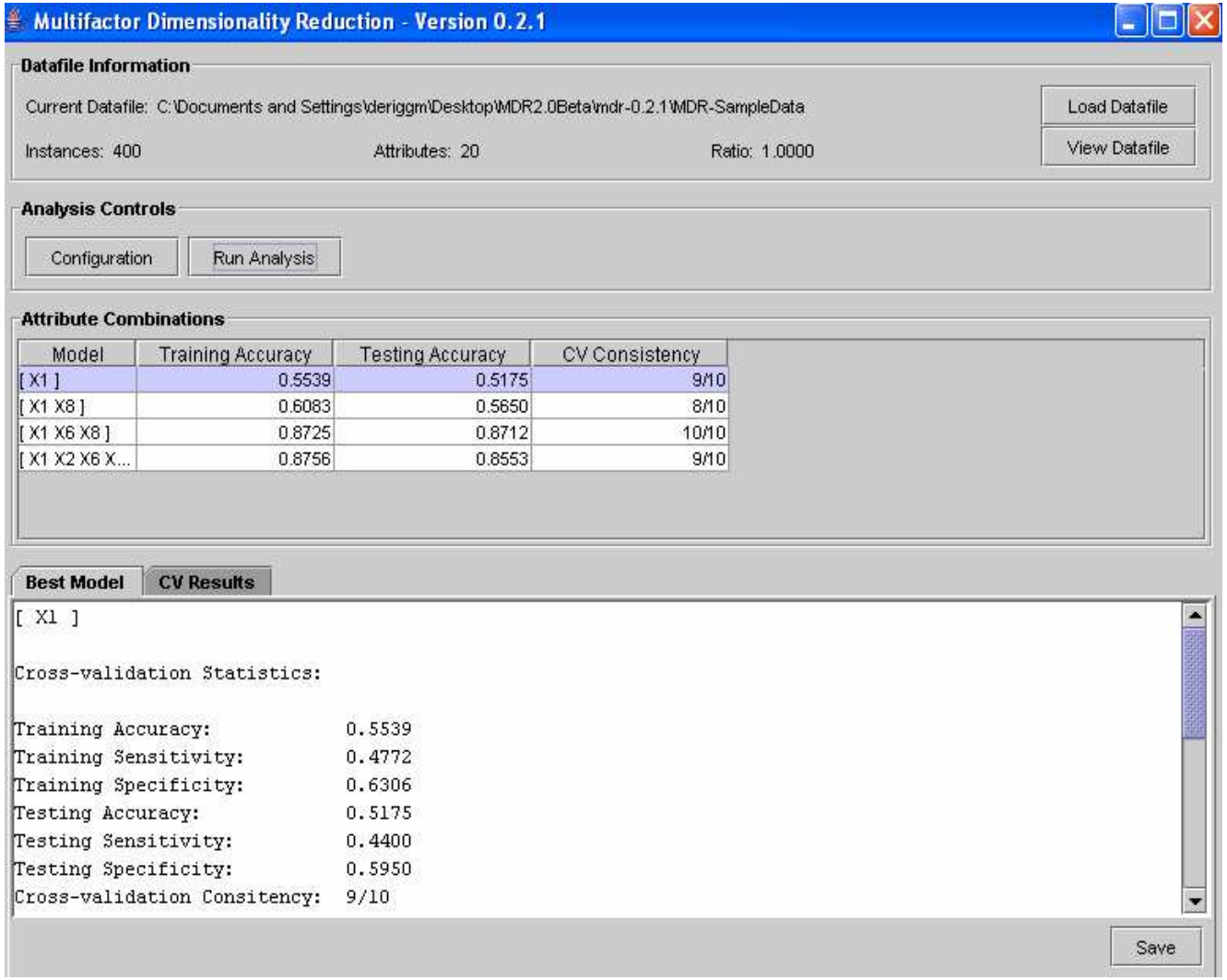
MDR Requirements

- Can detect main effects and interaction effects for discrete traits
- Sample size > 100 per group is optimal
  - Less than 100 individuals per group can result in biased prediction errors
- Dataset does not need to have a balanced number in each group
  - If more than 2:1 ratio, a sampling technique must be used
- No limit in number of individuals or variables
  - Though some configurations may be computationally infeasible

MDR Software

- Two distributions of software
  - http://www.epistasis.org/open-source-mdr-project.html
  - http://www.epistasis.org/weka-cg-project.html
- Open source
- JAVA – platform independent
- Modules included
  - Data manipulation
  - Permutation testing
- Command line JAVA version

MDR Software



MDR Software

- Parallel C++ version
  - Redesigned algorithm
  - Allows analysis of more variables or higher order interactions among variables
  - 2-way interaction analysis 50,000 completes in 40 hours on 24 Pentium 4 Processors
  - Currently developing methods to assign statistical significance
- Available via email only

Future Directions

- Comparison of MDR with other traditional methods on simulated data
  - logic regression, penalized logistic regression (CART and logistic regression completed)
- Simulation studies to detect interactions embedded in thousands of SNPs (genome-wide association)
- Simulations to investigate the effects of linkage disequilibrium between SNPs
- Implementation of bootstrapping to replace cross-validation

Acknowledgements

- Ritchie Lab
  - Marylyn Ritchie, PhD
  - Scott Dudek
  - Todd Edwards
  - Lance Hahn, PhD
  - Alison Motsinger
- Moore Lab
  - Jason H Moore, PhD
  - Nate Barney
  - Todd Holden
  - Bill White, MS

http://chgr.mc.vanderbilt.edu/ritchielab

marylyn.ritchie@vanderbilt.edu