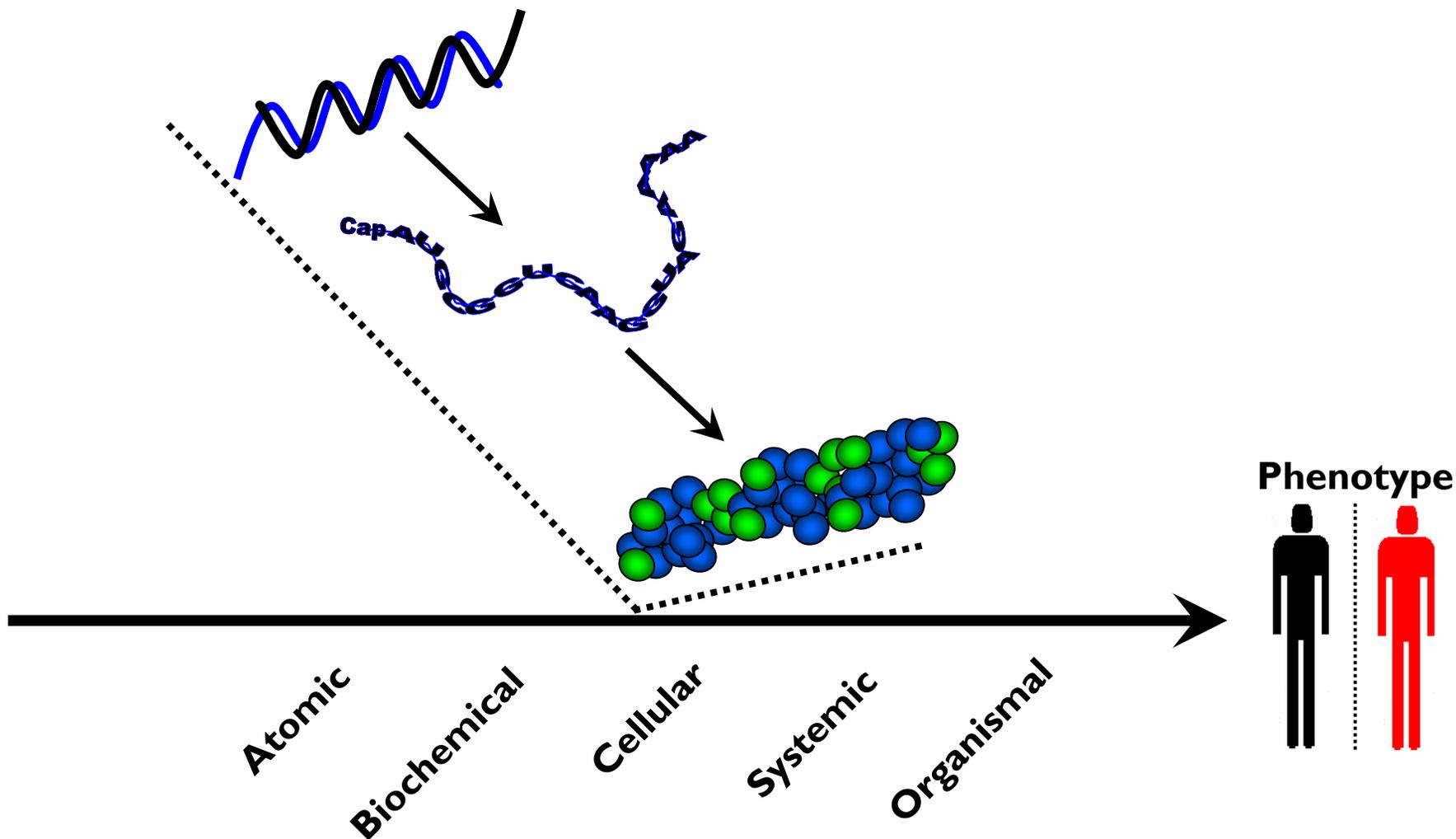


Detection and characterization of gene-gene and gene- environment interactions

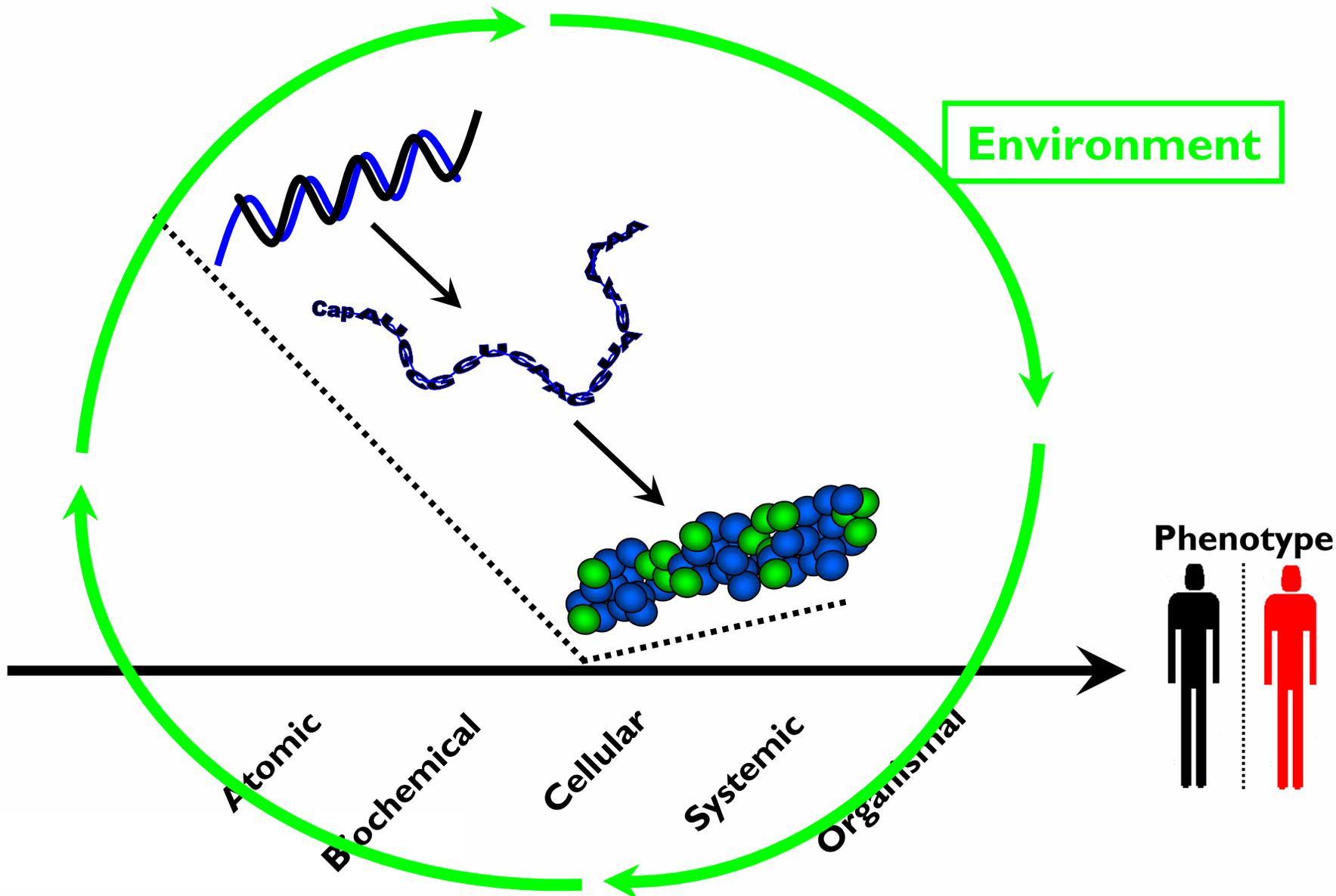
David Reif, Ph.D.



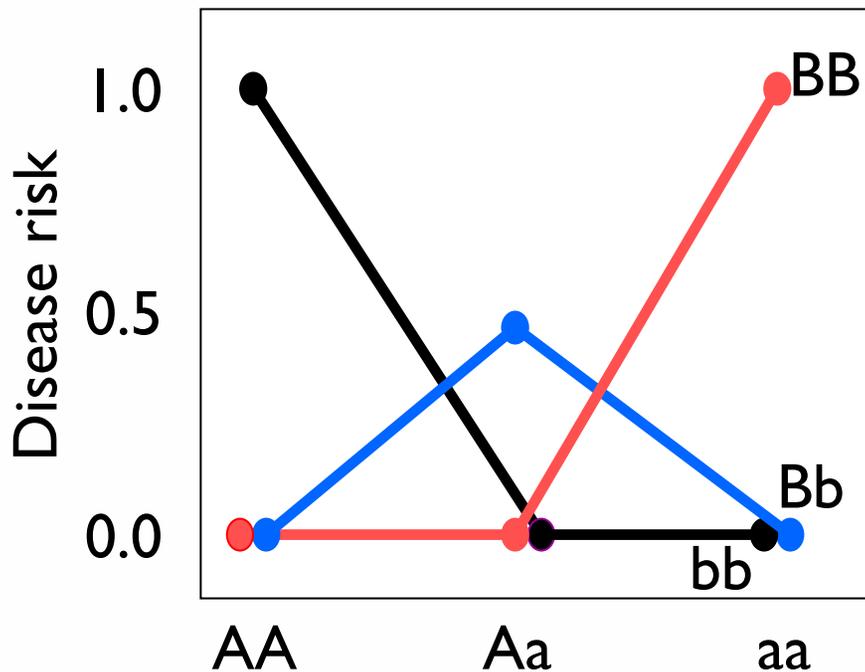
We analyze only a slice of the information related to complex phenotypes



We analyze only a slice of the information related to complex phenotypes



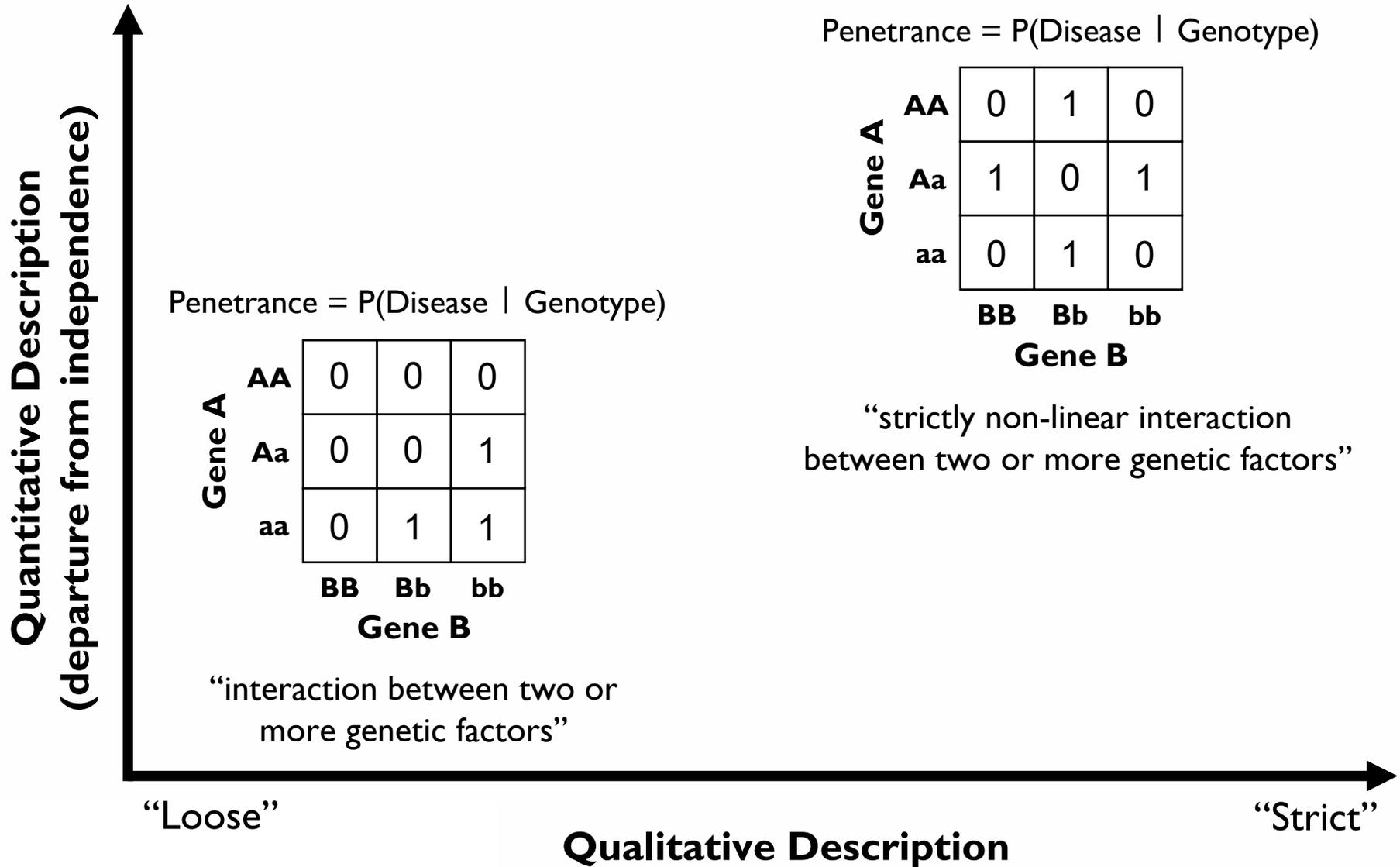
Epistasis in human disease



“standing upon”
(i.e. one gene masks the effect of another)
[Bateson (1909)]

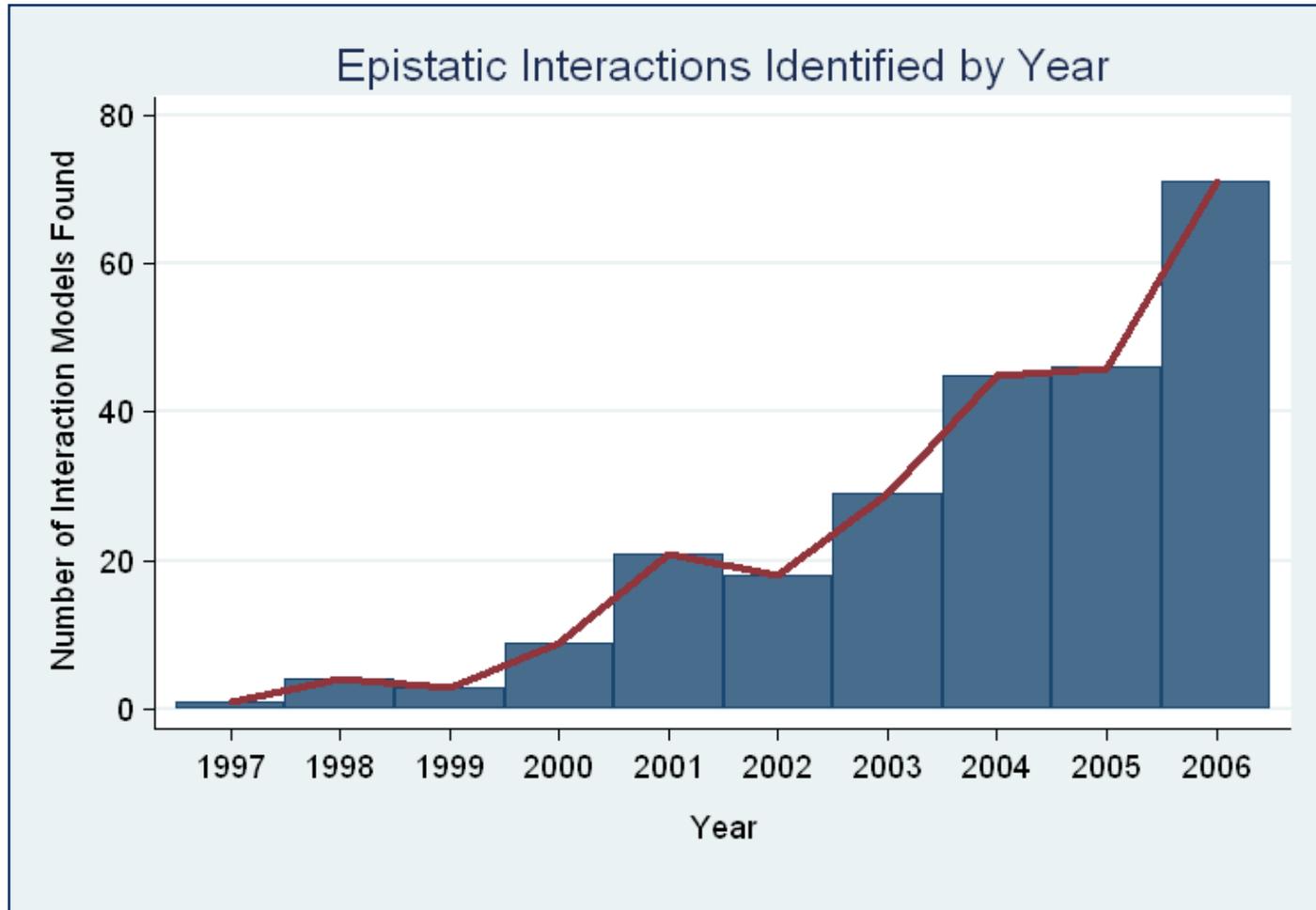
“interaction between genes”
[Cordell (2002)]

Degrees of epistasis



“gene-gene interactions are commonly found when properly investigated”

[Moore (2003)]



[Motsinger, Reif, Ritchie (2007)]

Novel approaches for detecting and characterizing interactions

Detection:

Multifactor Dimensionality Reduction (MDR)

Random Forests™

Restricted Partition Method (RPM)

Grammatical Evolution Neural Networks (GENN)

Symbolic Discriminant Analysis (SDA)

Multi-stage approaches:

 Focused Interaction Testing Framework (FITF)

 Set Association

 Joint permutation and filtering approaches

Characterization:

Logistic Regression

Interaction Dendrograms and Diagrams

Alternative solution representations (e.g. Decision Trees)

Expert Knowledge:

 Pathway inference/analysis

 Natural Language Processing (NLP) mining of literature

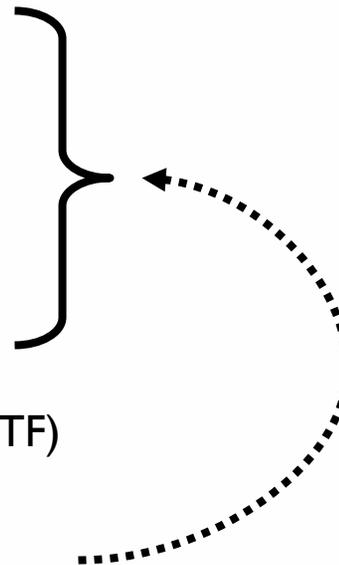
Novel approaches for detecting and characterizing interactions

Detection:

Multifactor Dimensionality Reduction (MDR)
Random Forests™
Restricted Partition Method (RPM)
Grammatical Evolution Neural Networks (GENN)
Symbolic Discriminant Analysis (SDA)

Multi-stage approaches:

Focused Interaction Testing Framework (FITF)
Set Association
Joint permutation and filtering approaches



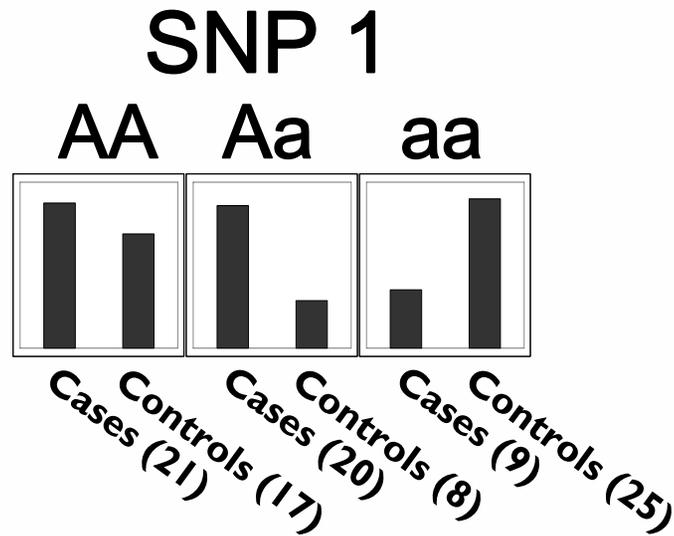
Characterization:

Logistic Regression
Interaction Dendrograms and Diagrams
Alternative solution representations (e.g. Decision Trees)
Expert Knowledge:

Pathway inference/analysis
Natural Language Processing (NLP) mining of literature

Curse of dimensionality

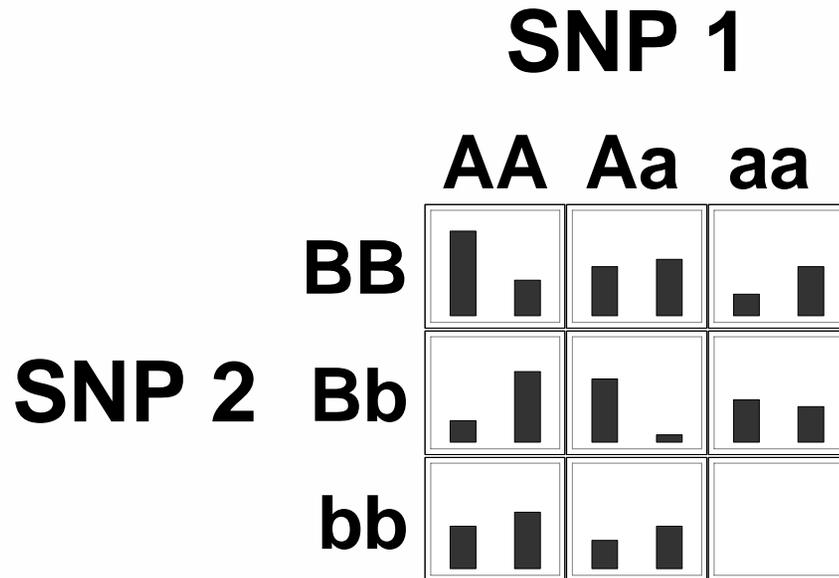
[Bellman (1961)]



$N = 100$
Cases = 50
Controls = 50

Curse of dimensionality

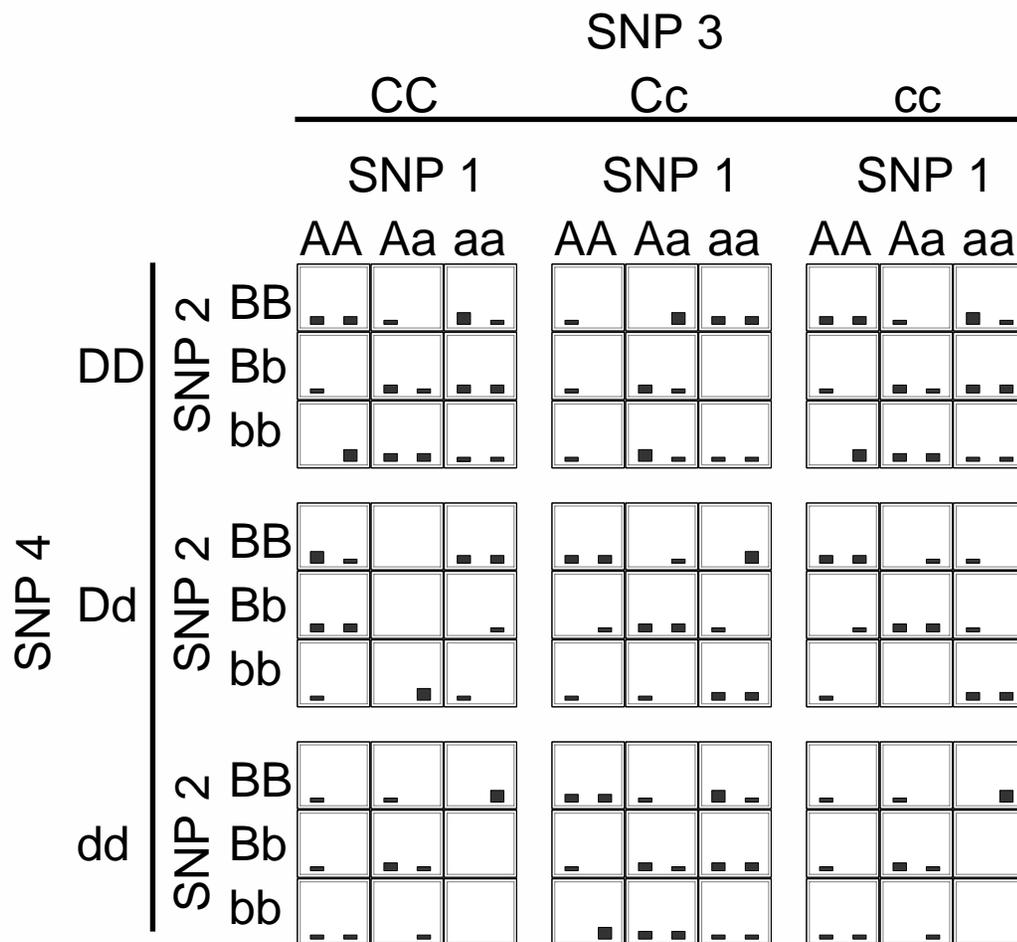
[Bellman (1961)]



$N = 100$
Cases = 50
Controls = 50

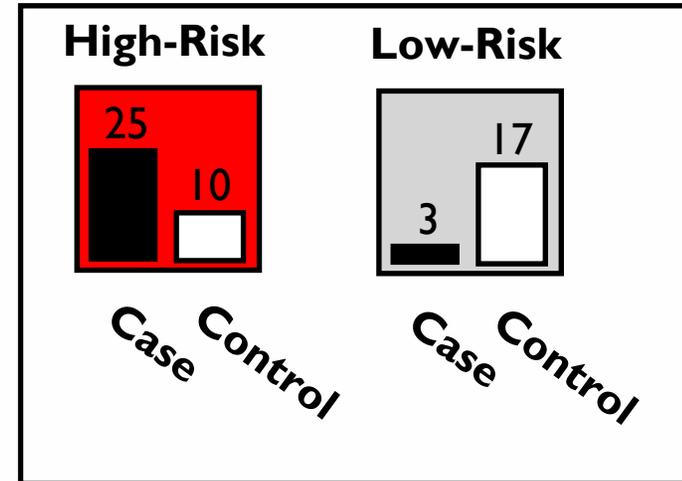
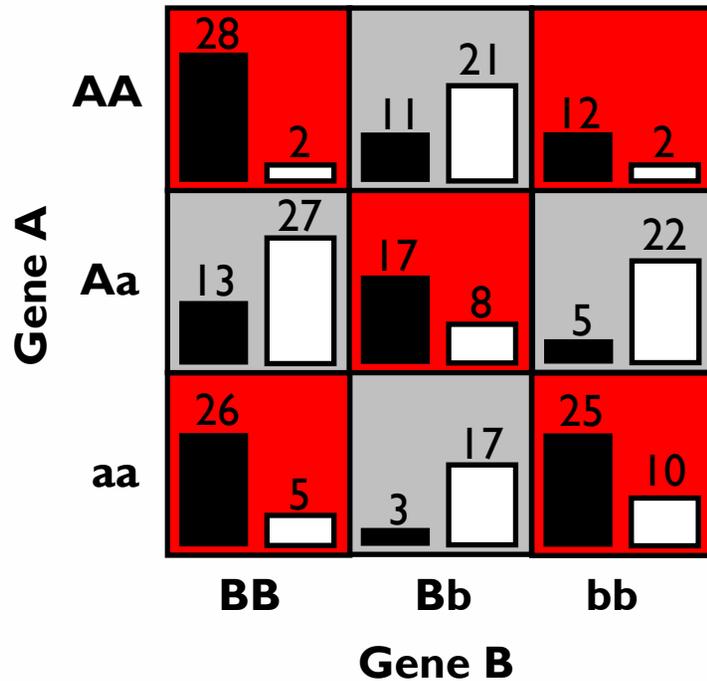
Curse of dimensionality

[Bellman (1961)]



$N = 100$
Cases = 50
Controls = 50

Multifactor Dimensionality Reduction (MDR)



Collapses combinations of attributes (e.g. two genetic factors) into
High-Risk/Low-Risk

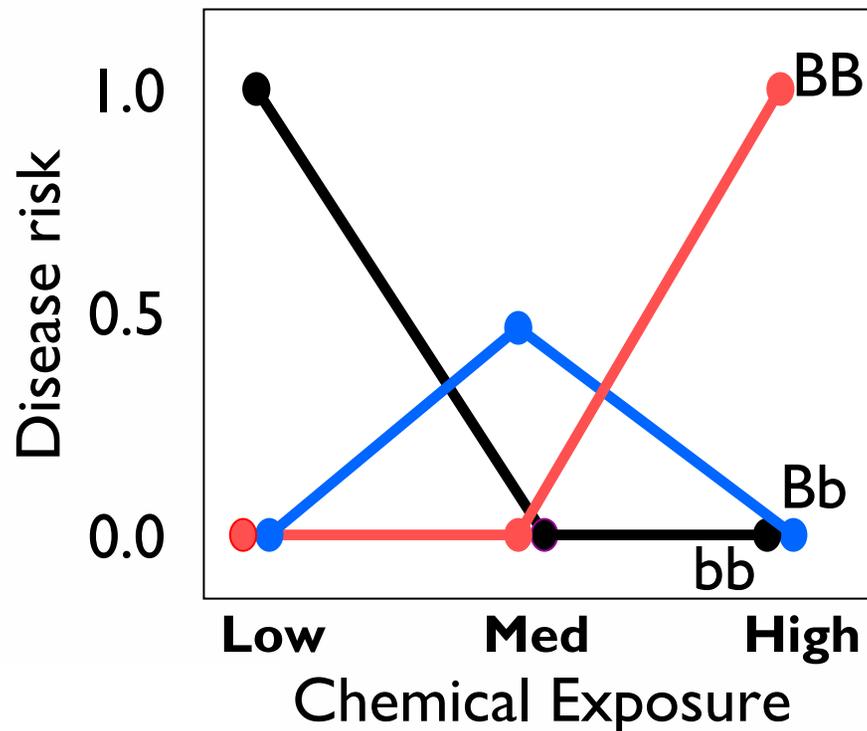
Ecogenetics

“genetic determinants that dictate susceptibility to environmentally influenced adverse health effects”

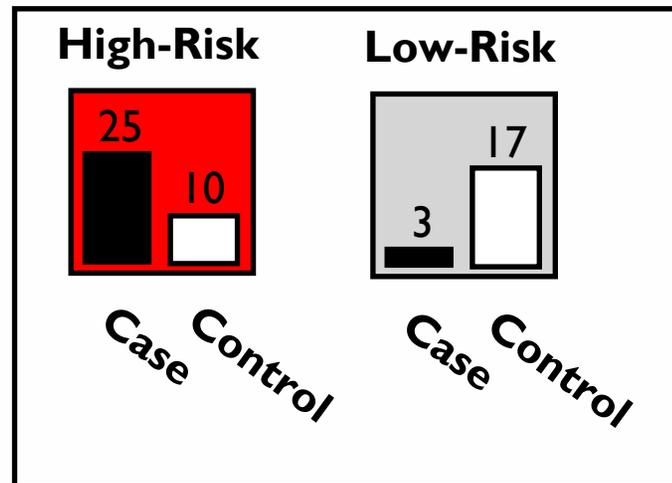
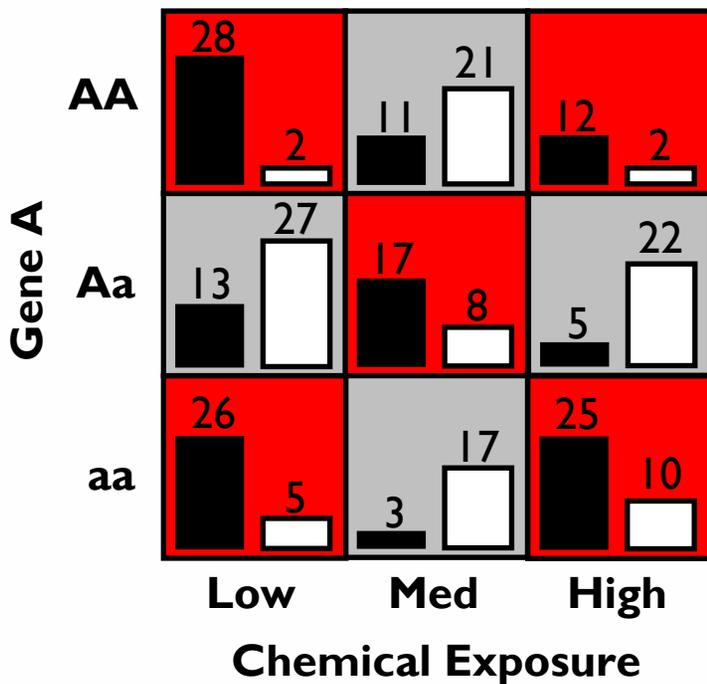
[Costa and Eaton (2006)]

“Genes load the gun. The environment pulls the trigger.”

[Bray (1998)]



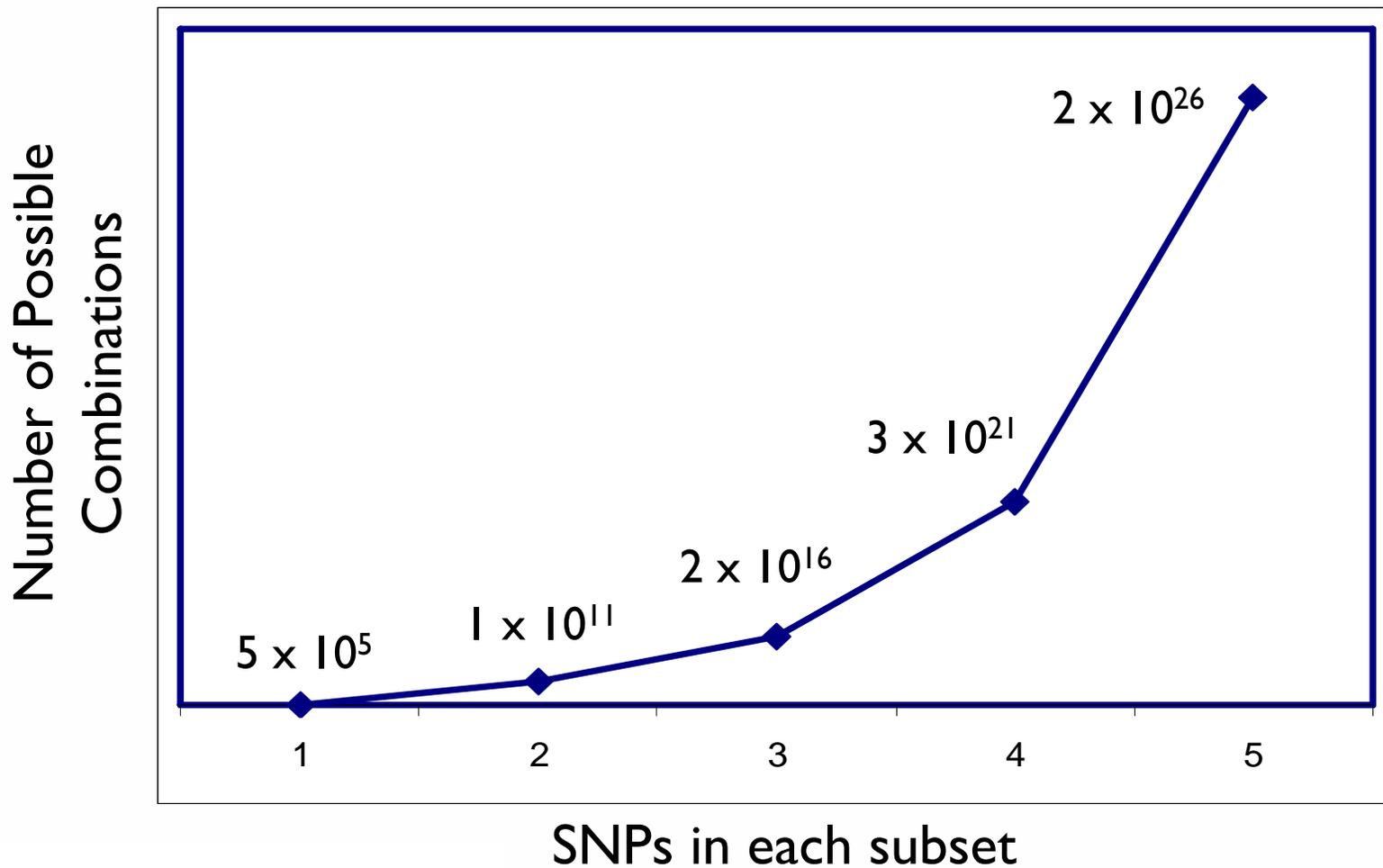
Multifactor Dimensionality Reduction (MDR)



Collapses combinations of attributes (e.g. genetic factor plus environmental factor) into High-Risk/Low-Risk

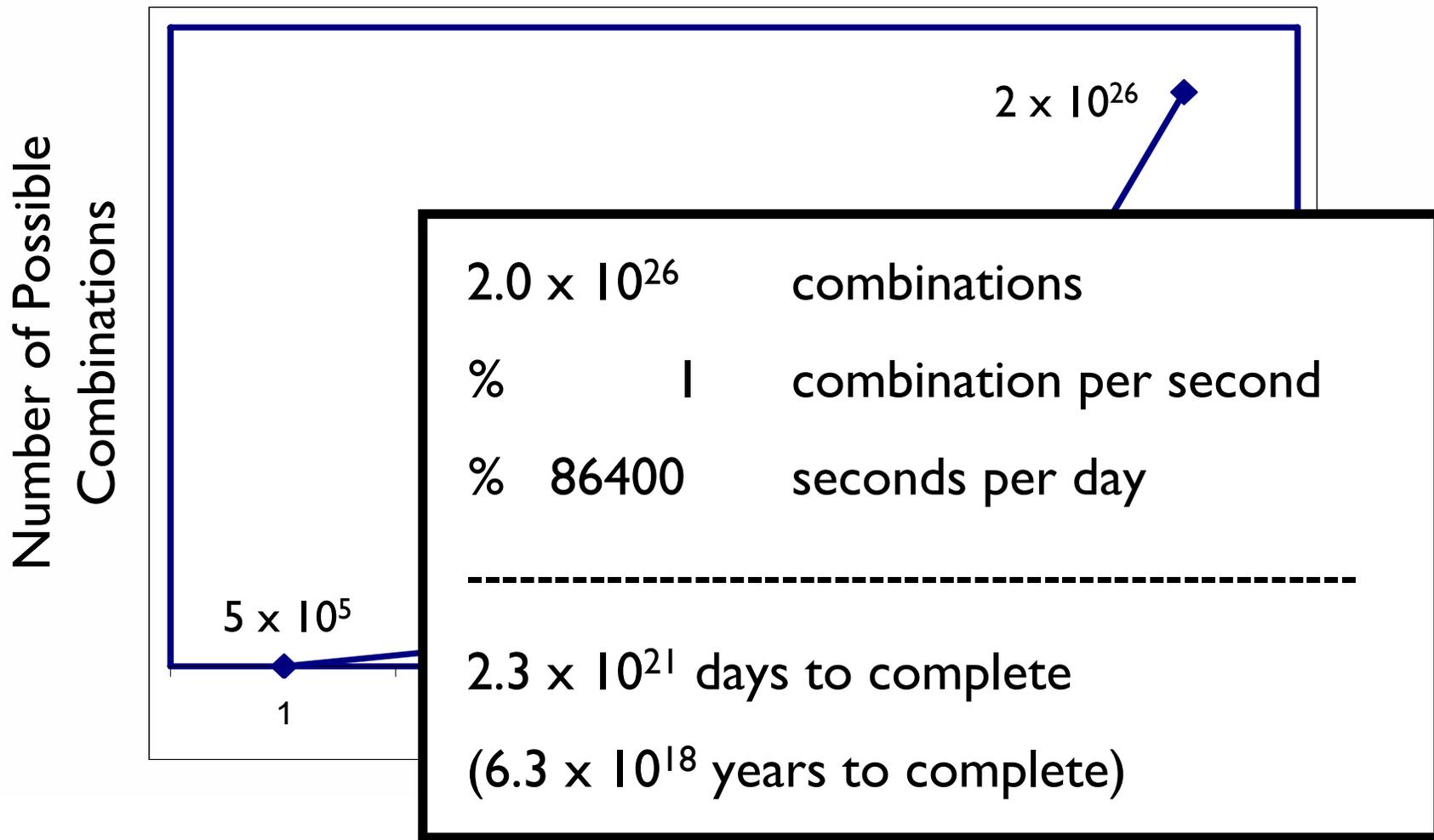
Exploding combinatorics

(For a genome-wide study including 500,000 SNPs)

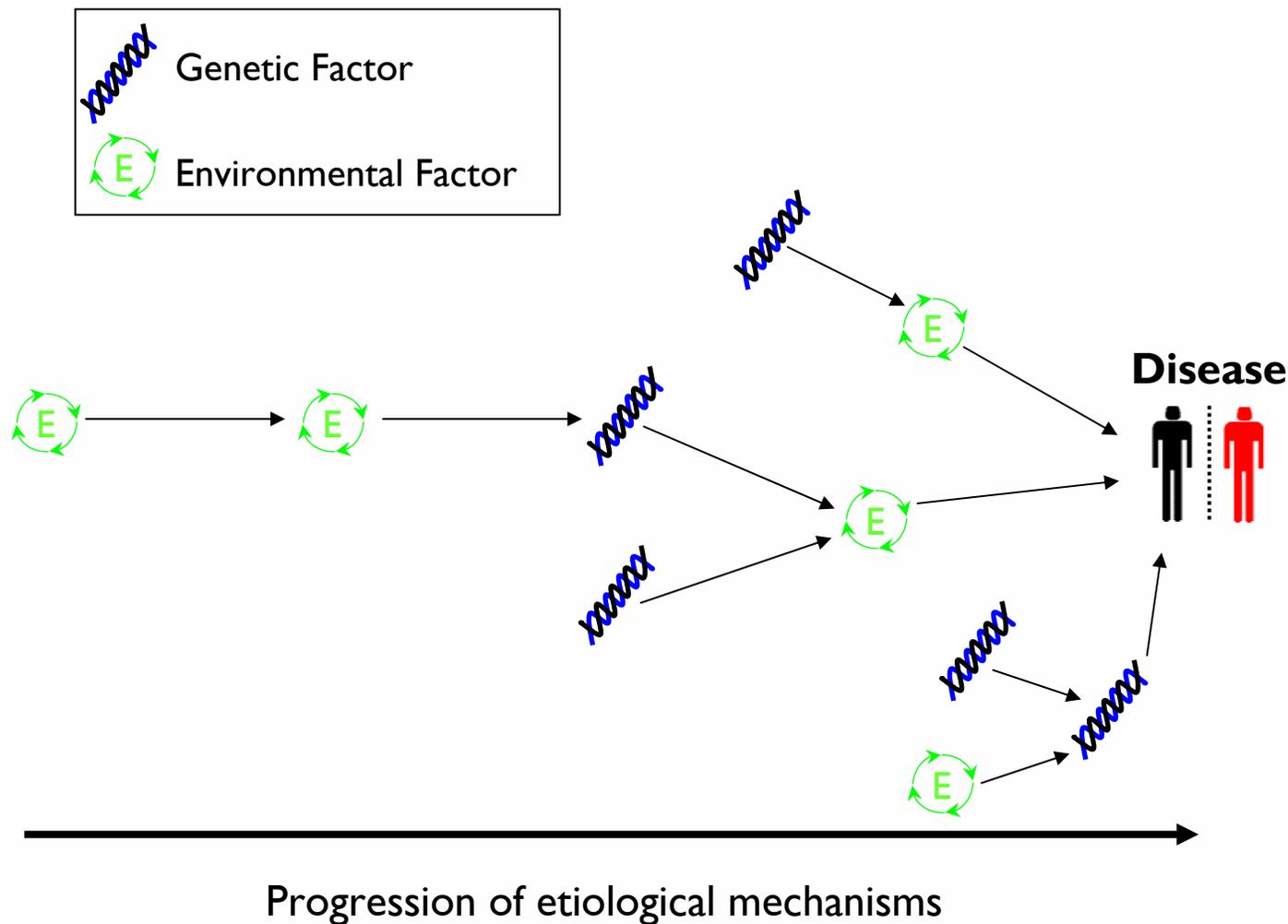


Exploding combinatorics

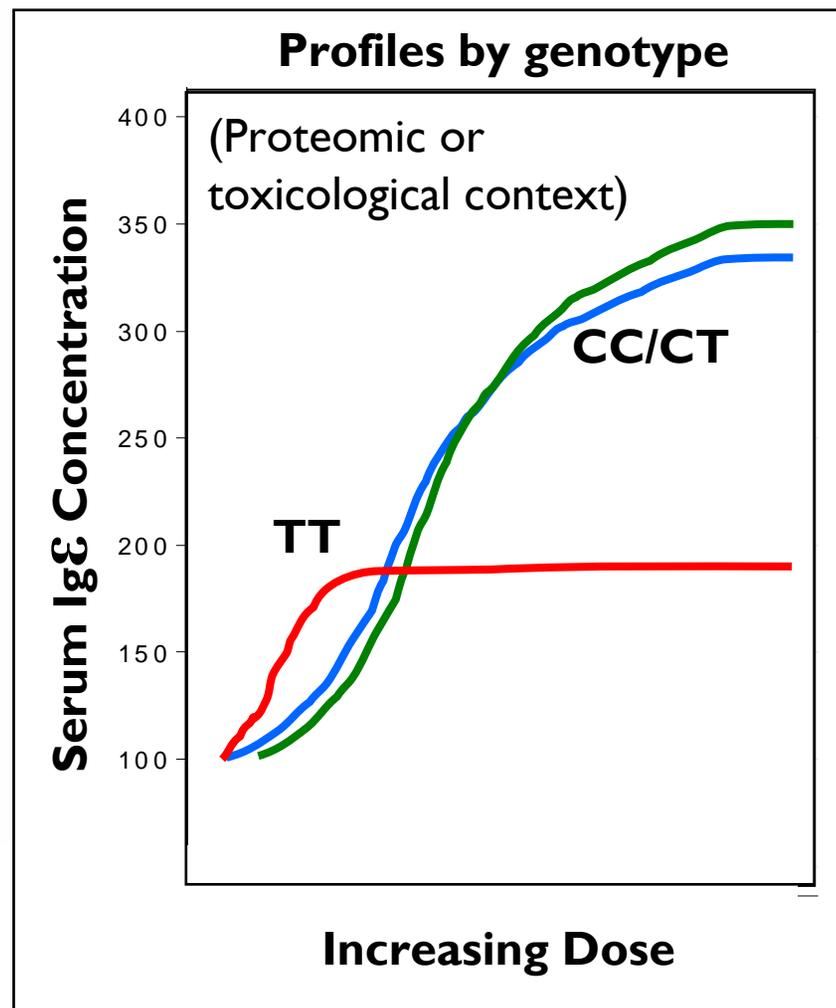
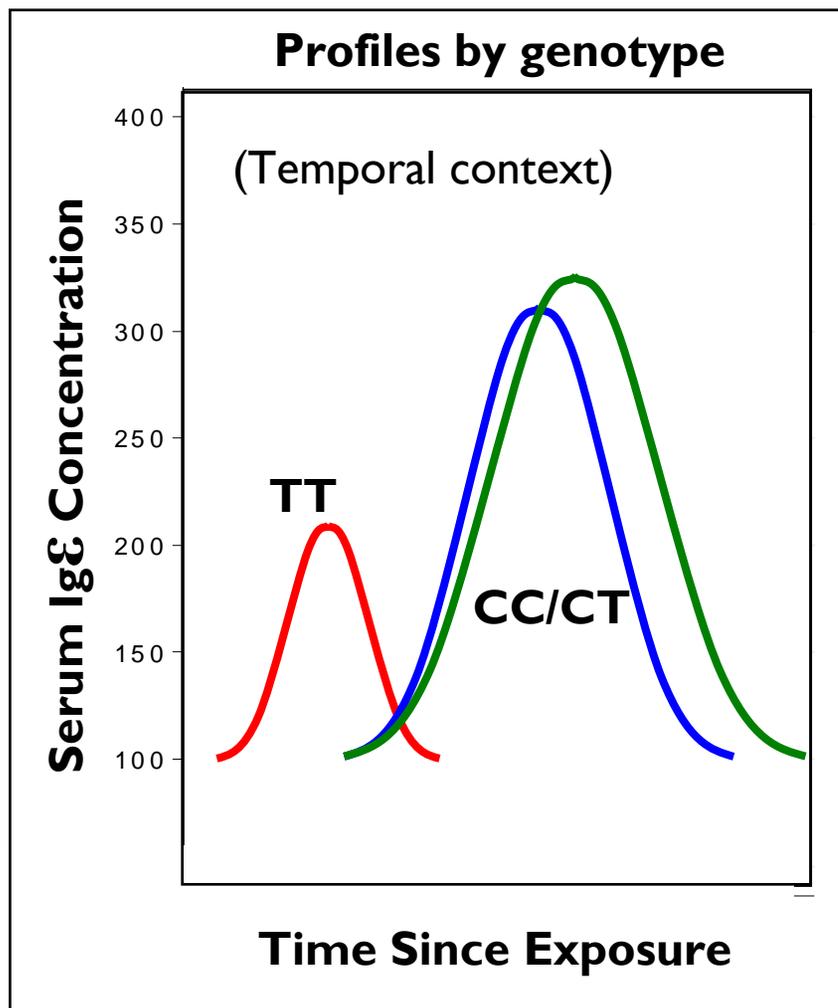
(For a genome-wide study including 500,000 SNPs)



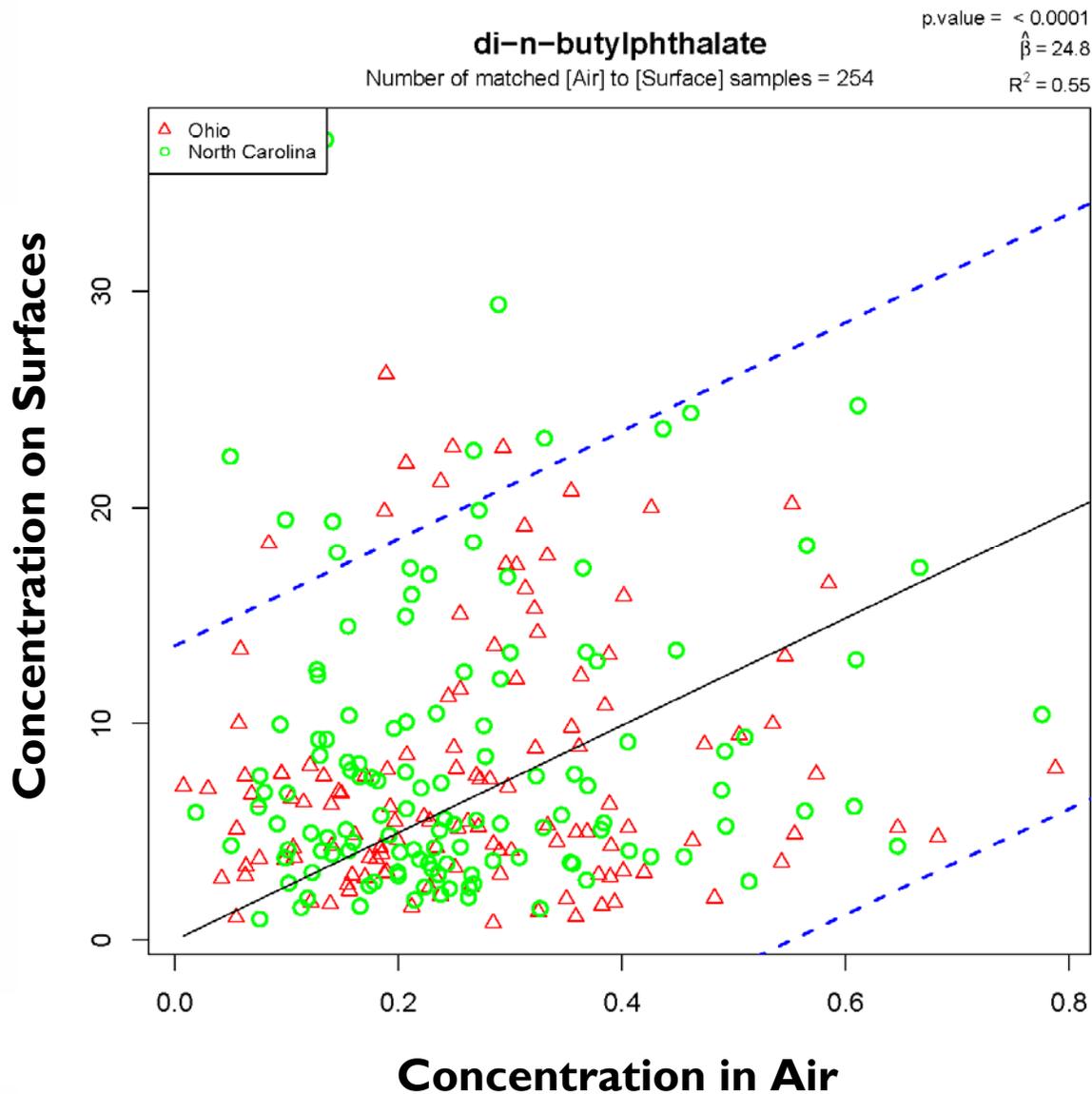
Complex diseases involve multiple etiopathological pathways



Gene-Environment interactions are context dependent

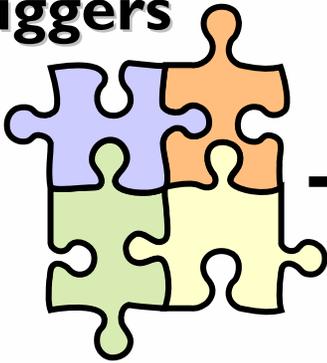


Measuring (characterizing) the environmental context

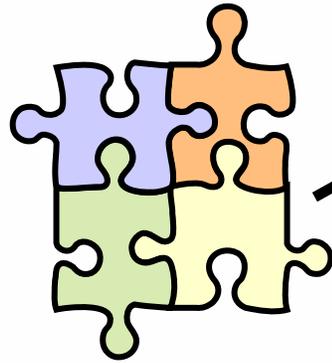
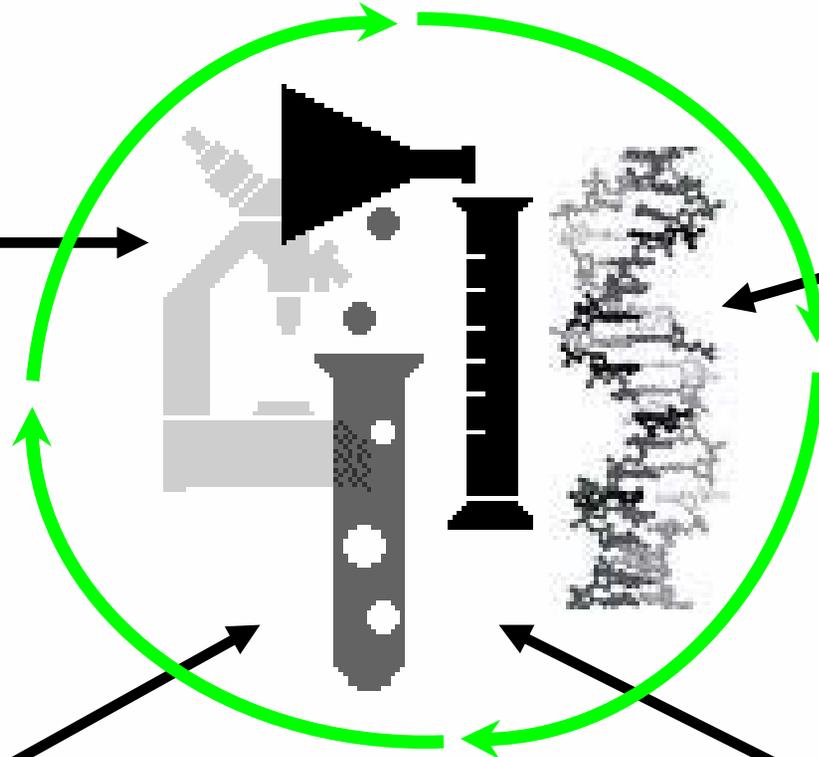
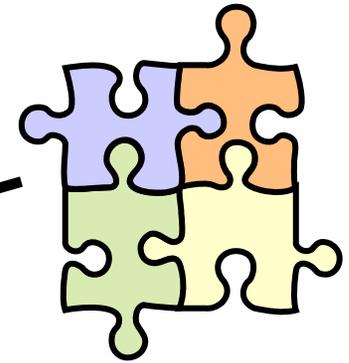


Asthma etiology

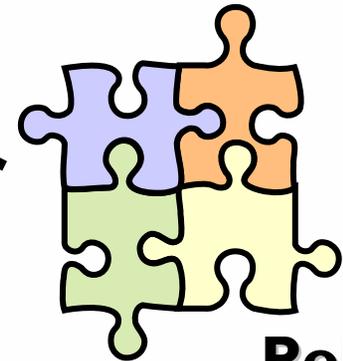
**Indoor
Triggers**



Genetics



Behavior

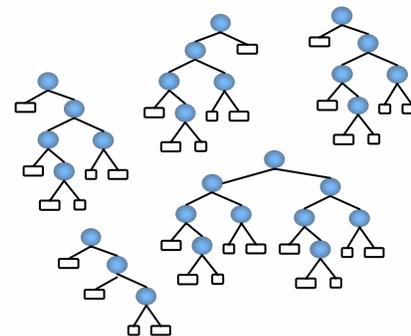


**Air
Pollution**

Example strategy for detecting and characterizing gene-environment interactions associated with asthma

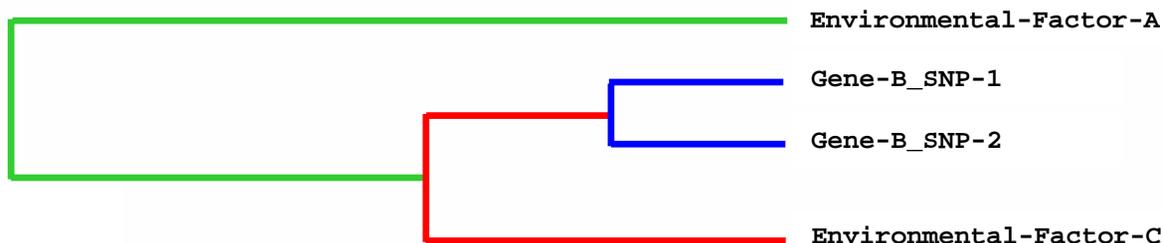
Detection:

Use Random Forests (RF) to identify genetic and/or environmental variables most associated with asthma.



Characterization:

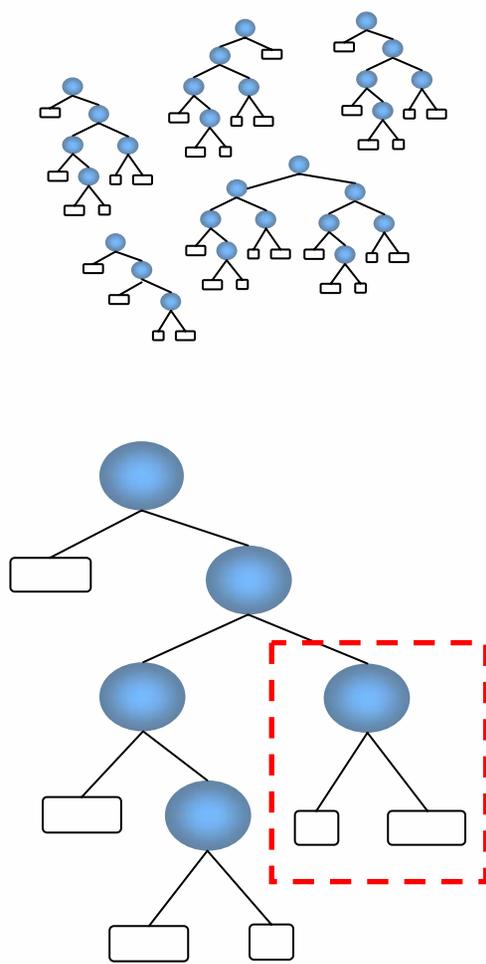
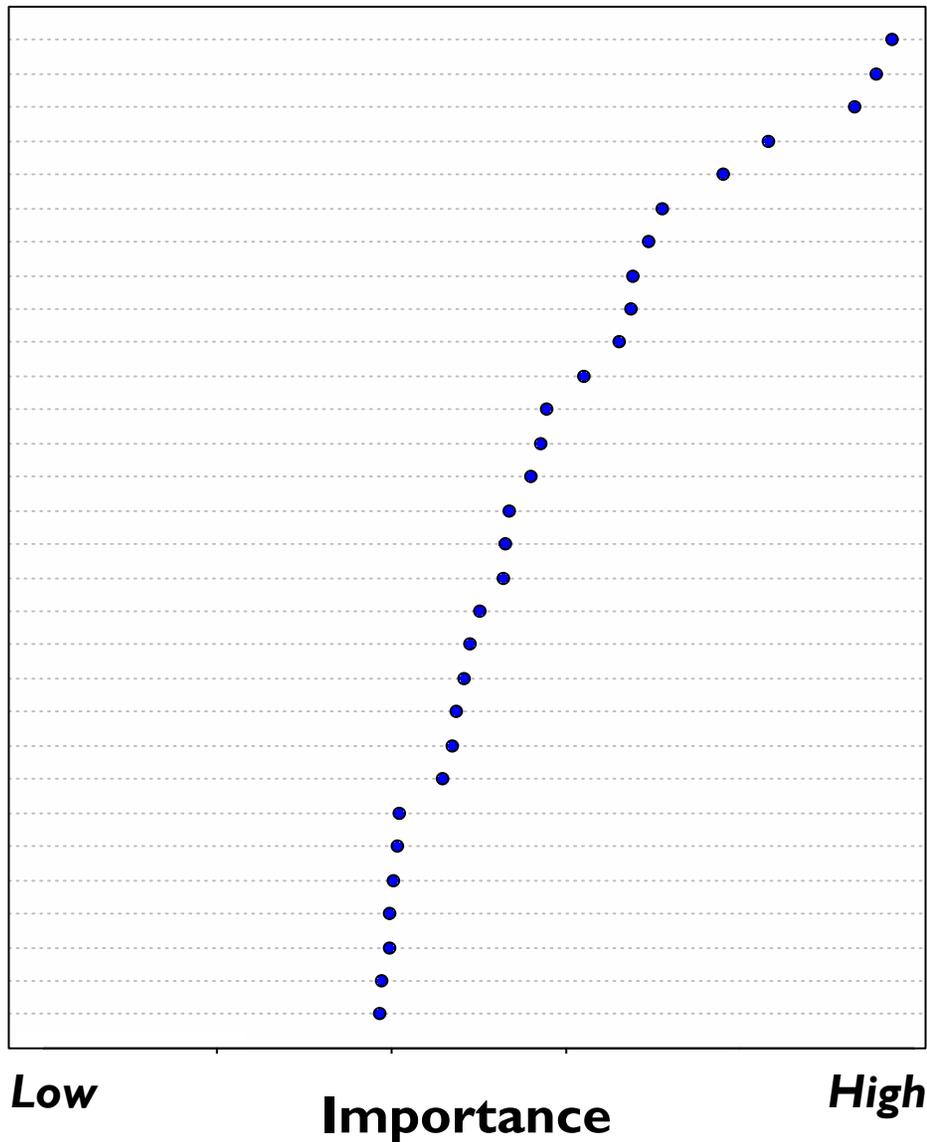
Use interaction dendrograms to characterize the nature of the interactions among the genetic variables and environmental variables most associated with asthma as identified by Random Forests.





Variable importance using RF on + data analyzed simultaneously

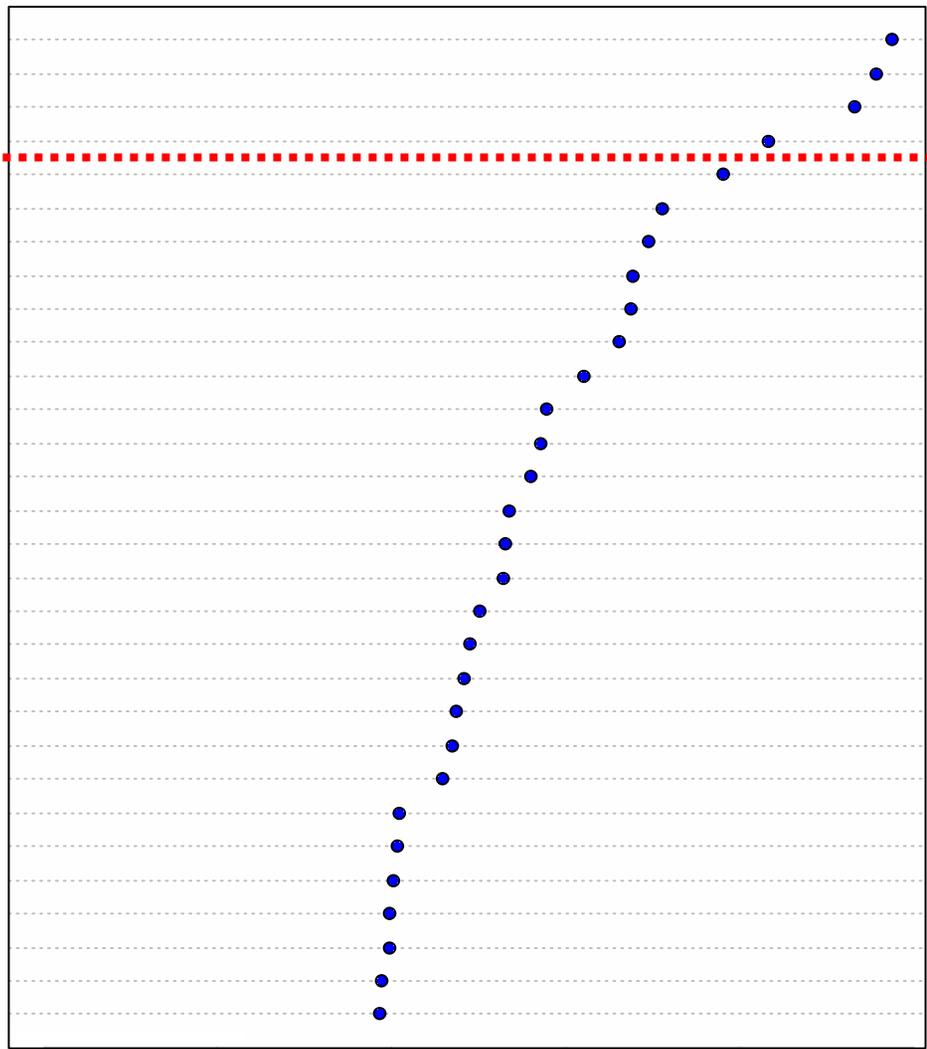
Gene-B_SNP-2
Gene-B_SNP-1
Environmental-Variable-A
Environmental-Variable-C
Environmental-Variable-K
Gene-X_SNP-1
Gene-J_SNP-4
Gene-Q_SNP-2
Environmental-Variable-M
Gene-N_SNP-1
Environmental-Variable-U
Environmental-Variable-W
Gene-J_SNP-2
Gene-J_SNP-1
Gene-Y_SNP-3
Environmental-Variable-H
Environmental-Variable-Z
Gene-A_SNP-1
Environmental-Variable-V
Gene-D_SNP-1
Environmental-Variable-E
Gene-U_SNP-1
Gene-K_SNP-6
Gene-L_SNP-3
Environmental-Variable-X
Gene-R_SNP-7
Environmental-Variable-Y
Environmental-Variable-I
Gene-Z_SNP-2
Gene-A_SNP-3



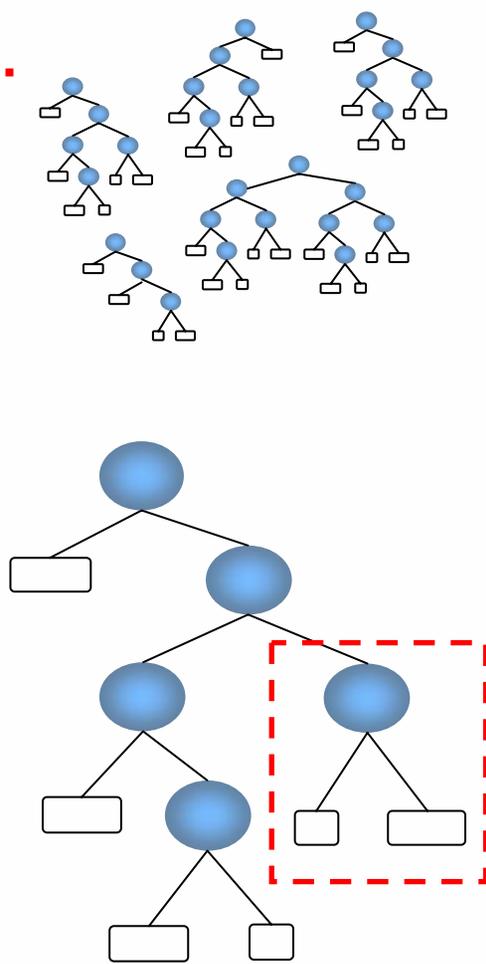


Variable importance using RF on + data analyzed simultaneously

Gene-B_SNP-2
Gene-B_SNP-1
Environmental-Variable-A
Environmental-Variable-C
Environmental-Variable-K
Gene-X_SNP-1
Gene-J_SNP-4
Gene-Q_SNP-2
Environmental-Variable-M
Gene-N_SNP-1
Environmental-Variable-U
Environmental-Variable-W
Gene-J_SNP-2
Gene-J_SNP-1
Gene-Y_SNP-3
Environmental-Variable-H
Environmental-Variable-Z
Gene-A_SNP-1
Environmental-Variable-V
Gene-D_SNP-1
Environmental-Variable-E
Gene-U_SNP-1
Gene-K_SNP-6
Gene-L_SNP-3
Environmental-Variable-X
Gene-R_SNP-7
Environmental-Variable-Y
Environmental-Variable-I
Gene-Z_SNP-2
Gene-A_SNP-3

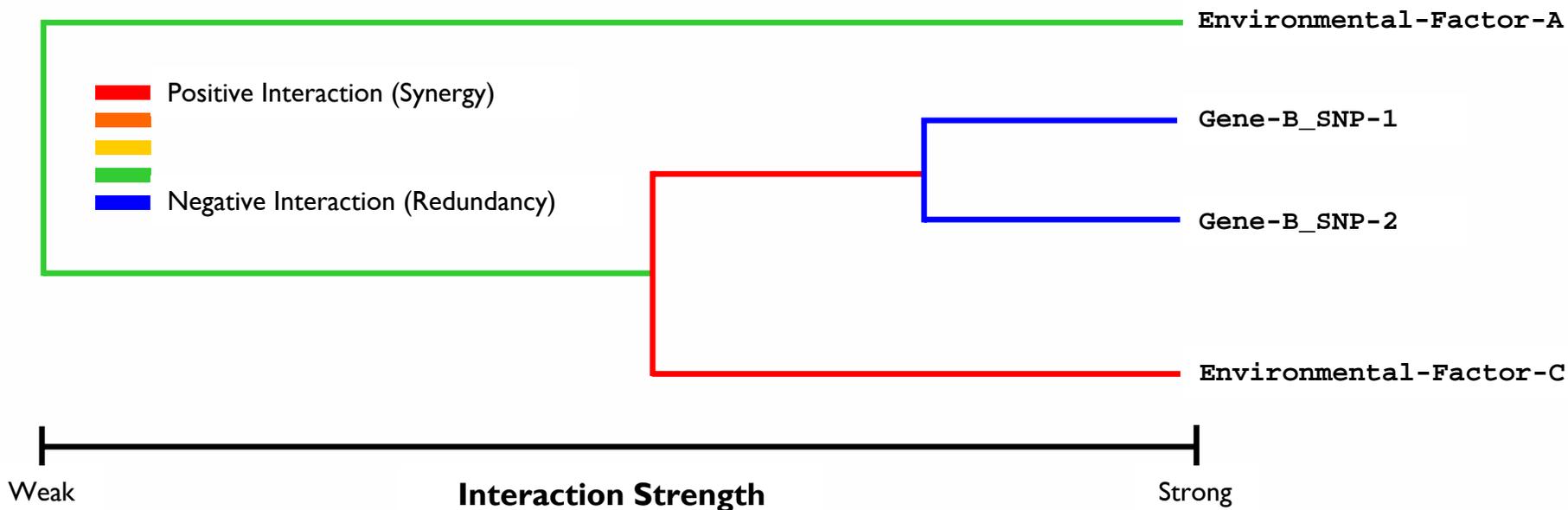


Low Importance High



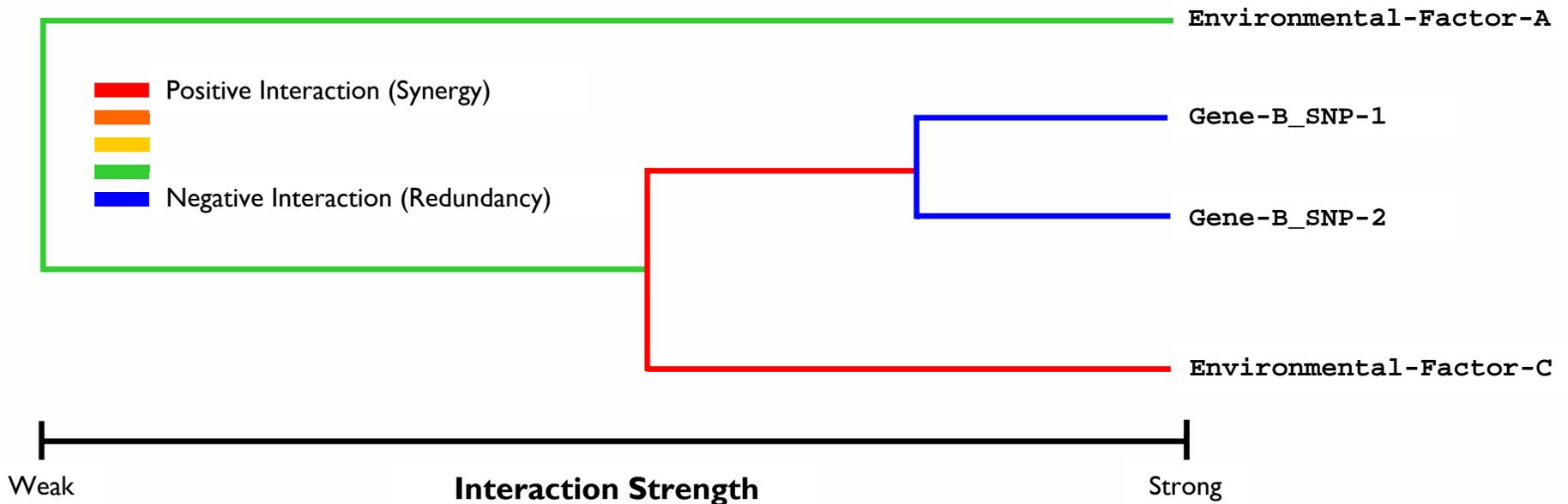
Characterizing interactions

Interaction dendograms for the genetic variables (SNPs) and environmental variables (indoor allergen measurements) most associated with asthma as identified by Random Forests.



Characterizing interactions

Interaction dendograms for the genetic variables (SNPs) and environmental variables (indoor allergen measurements) most associated with asthma as identified by Random Forests.



Interpretation: The redundancy between the two SNPs in Gene-B may be indicative of high *intra*genic LD (e.g. $r^2 > 0.9$).

Conclusions & Recommendations

- **Conclusions:**

- Given current analytical and computational power, study *design* is the major driver behind detection of G*G or G*E interactions.
 - Proper measurement of exposure variables
 - Proper characterization of endpoints

- **Recommendations:**

- Interdisciplinary science
 - Comprehensive studies include experts in multiple fields
- Both novel and traditional methods are valuable
 - Choice depends upon context
 - R (and related projects) allows facile implementation of new methods
 - GUIs prevalent for complex methods
 - “context independence” of methods
- Adopt a multifactorial mindset
 - Accept low-hanging fruit (univariate fruit is the sweetest of all), but explore interaction space

Acknowledgments

Alison Motsinger

(Vanderbilt University)

alison.motsinger@vanderbilt.edu

Elaine Cohen-Hubal

(U.S. EPA)

<http://www.epa.gov/comptox>

Jason Moore

(Dartmouth Medical School)

<http://epistasis.org>

Jane Gallagher

(U.S. EPA)

<http://www.epa.gov/NHEERL/hsd/>

Bill White

(Dartmouth Medical School)

<http://epistasis.org>

Brett McKinney

(University of Alabama-Birmingham)

<http://www.genetics.uab.edu/McKinneyLab>

Marylyn Ritchie

(Vanderbilt University)

<http://chgr.mc.vanderbilt.edu/content/ritchie>

John Little

(Virginia Polytechnic Institute)

<http://www.cee.vt.edu/people/little.html>

Aleks Jakulin

(Columbia)

<http://www.stat.columbia.edu/~jakulin>

John Wambaugh

(U.S. EPA)

<http://www.epa.gov/comptox>

DISCLAIMER: The contents of this presentation do not necessarily reflect EPA policy.