GWAS and prior knowledge to uncover gene-gene interactions

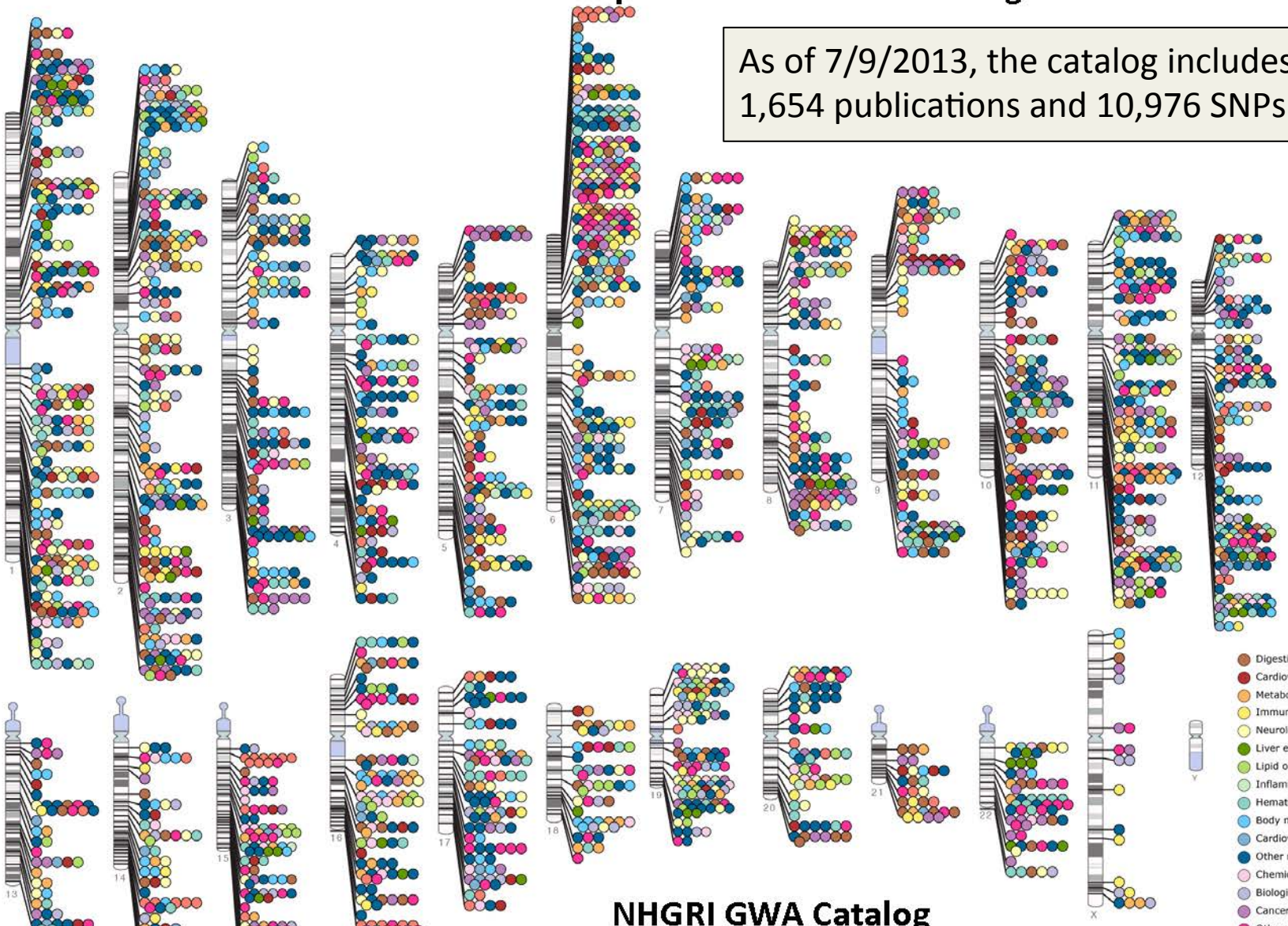
Marylyn D. Ritchie, PhD
Director, Center for Systems Genomics
The Pennsylvania State University
Biochemistry and Molecular Biology
July 18, 2013

PENNSTATE
1855

Published Genome-Wide Associations through 07/2012

Published GWA at $p \leq 5X10^{-8}$ for 18 trait categories

As of 7/9/2013, the catalog includes 1,654 publications and 10,976 SNPs.

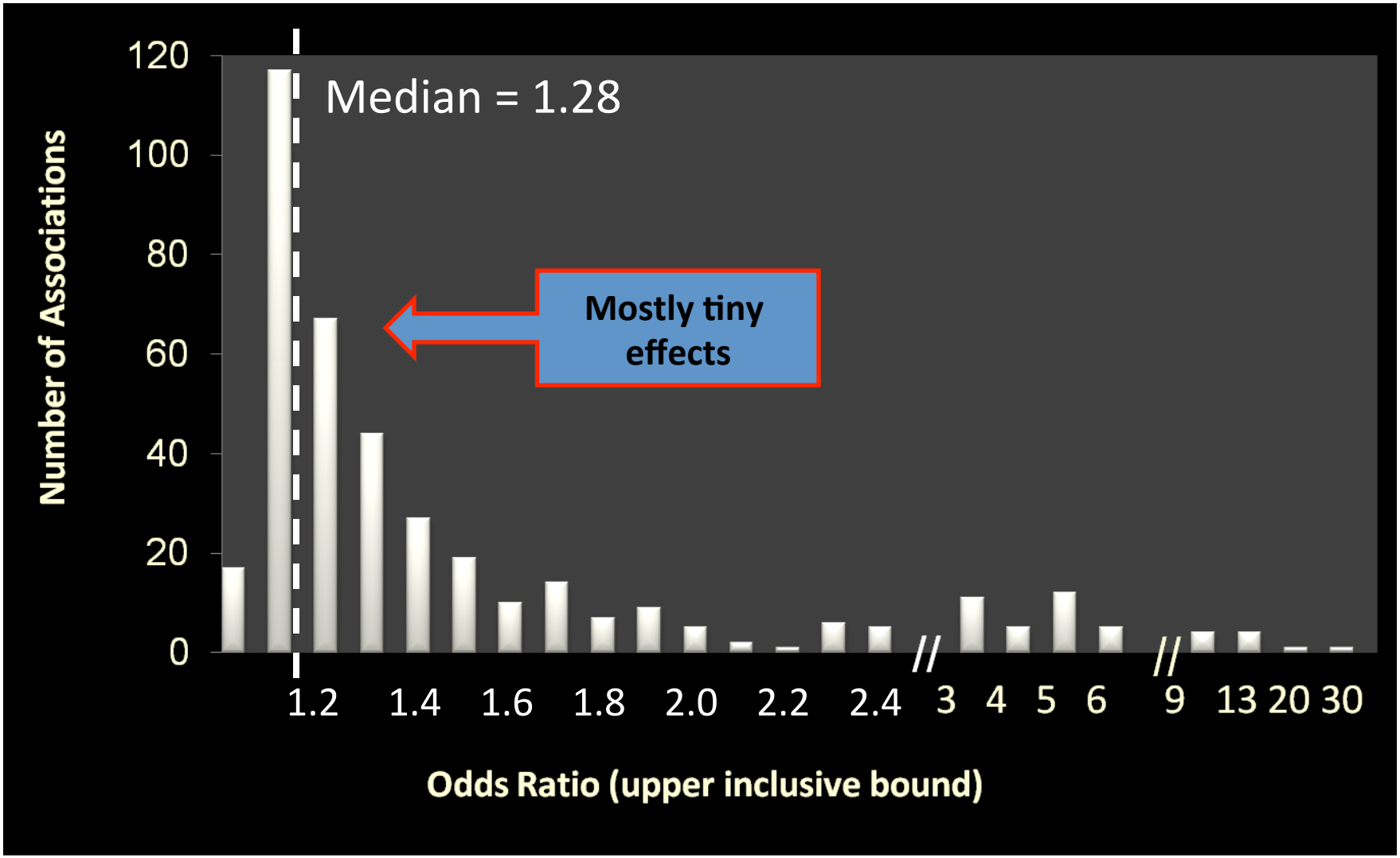


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

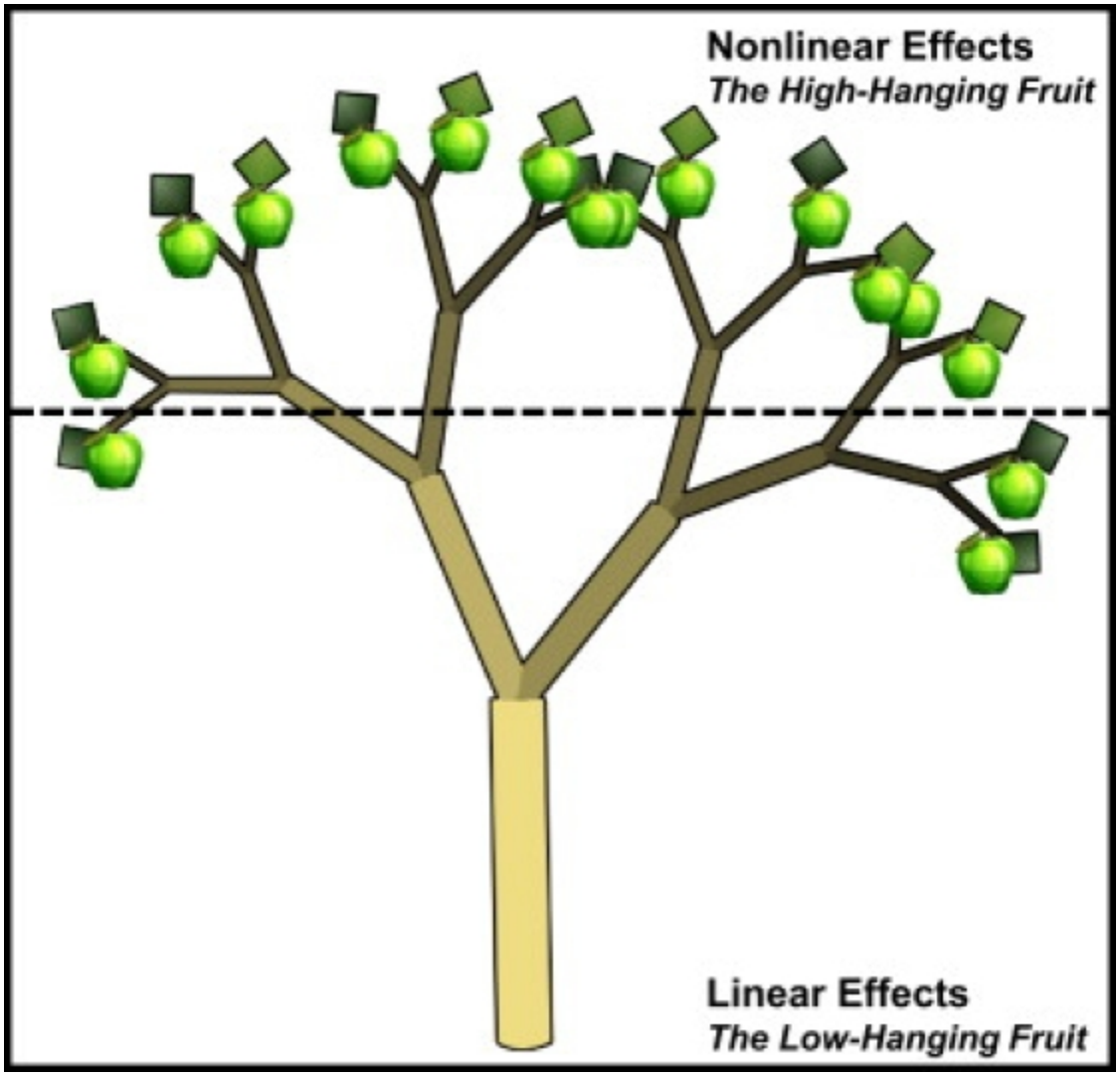
Distribution of Effects



PENNSTATE

Courtesy of Teri Manolio

Distribution of Effects



PENNSTATE

Moore and Williams. Am J Hum Genet. 2009; 85(3): 309–320

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Missing Heritability

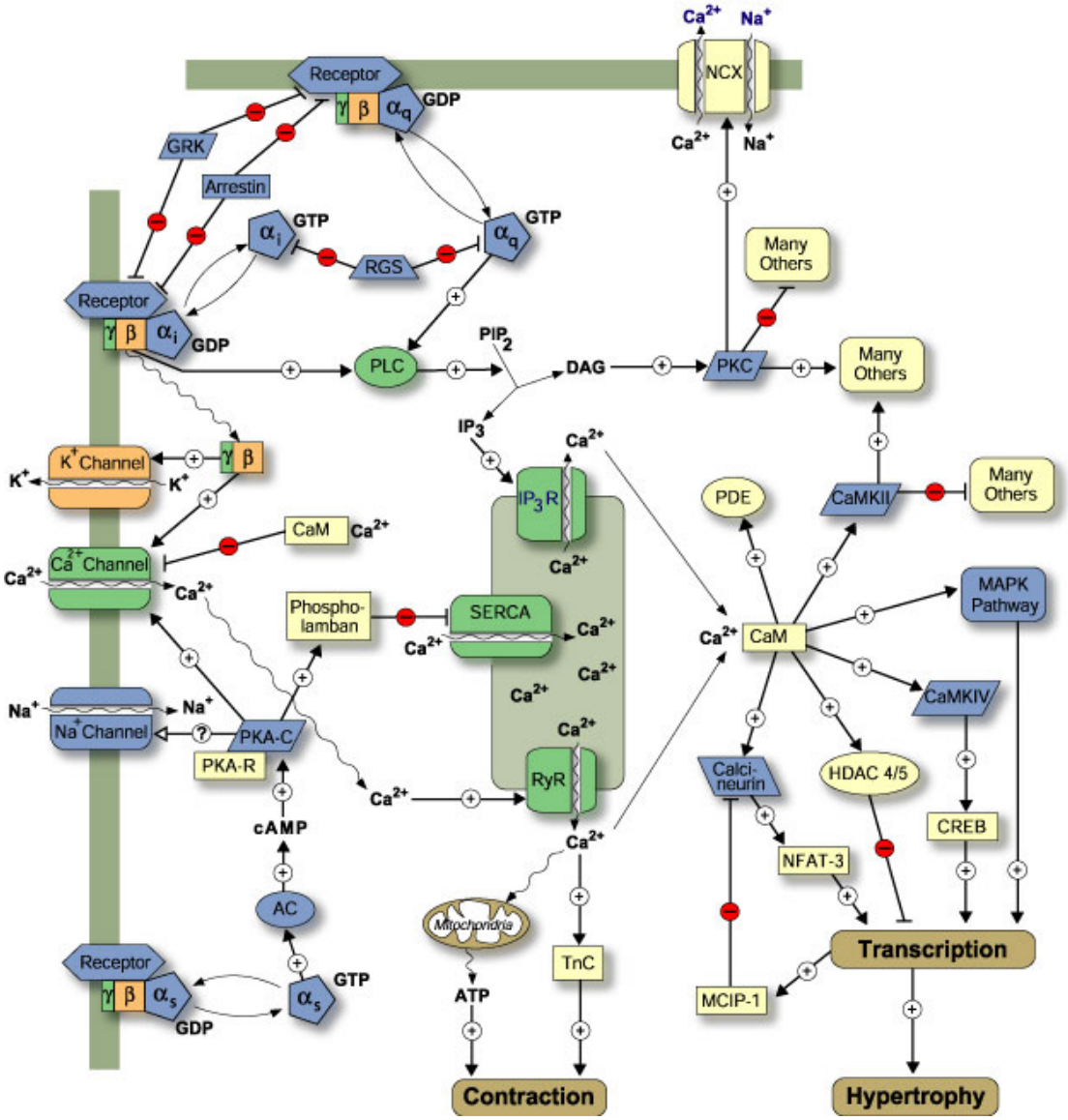
- Under our nose
- Out of sight
- In the architecture
- Underground networks
- Lost in diagnosis
- The great beyond

PENNSTATE

1855

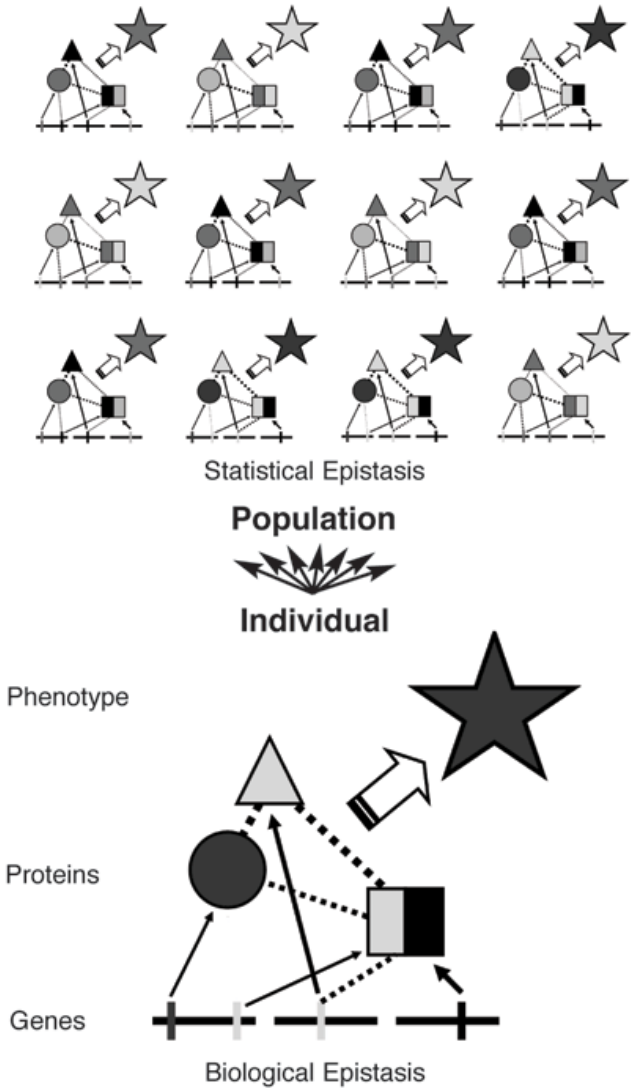
Maher, B. *Nature* 2008; 456:18-21.

Biology is complex



PENNSTATE
1855

Statistical vs. biological epistasis



PENNSTATE

Moore and Williams, BioEssays 27:637–646, 2005

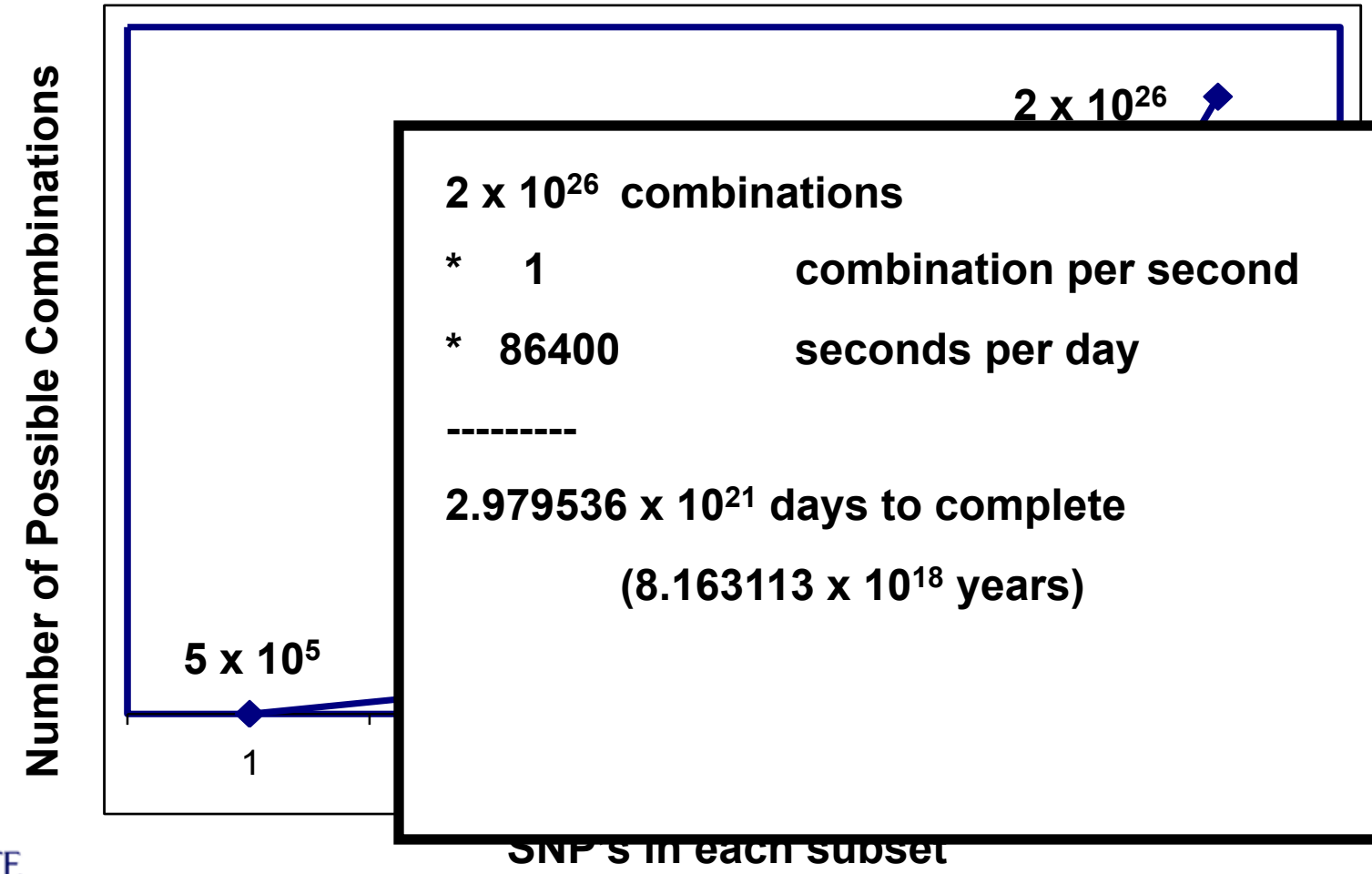
If interactions with minimal main effects are the norm rather than the exception, can we analyze all possible combinations of loci with traditional approaches to detect purely interaction effects ?

NO

PENNSTATE

How many combinations are there?

- ~500,000 SNPs to span the genome (HapMap)



2×10^{26} combinations

*** 1 combination per second**

*** 86400 seconds per day**

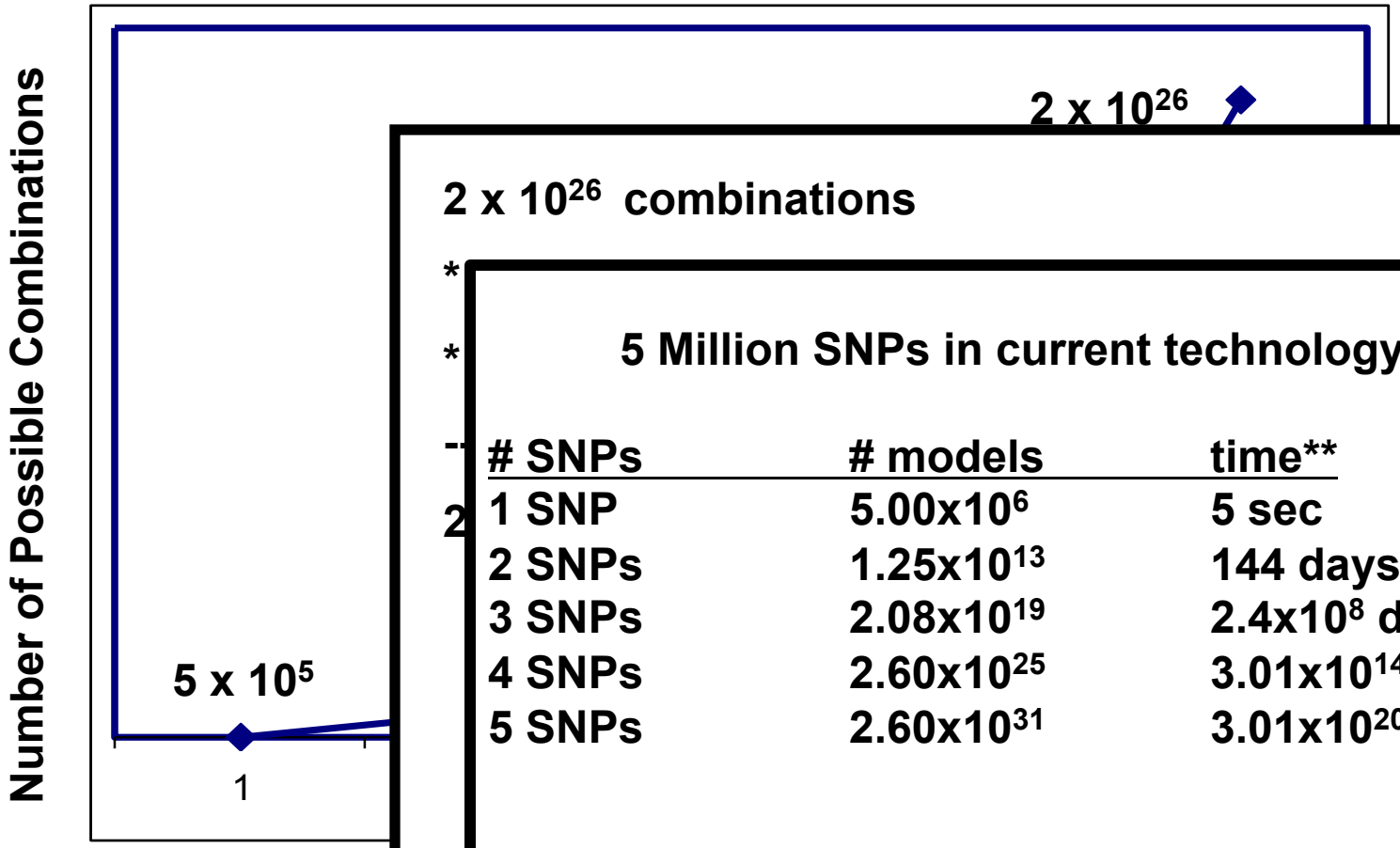
2.979536×10^{21} days to complete

(8.163113×10^{18} years)

PENNSTATE

How many combinations are there?

- ~500,000 SNPs to span the genome (HapMap)



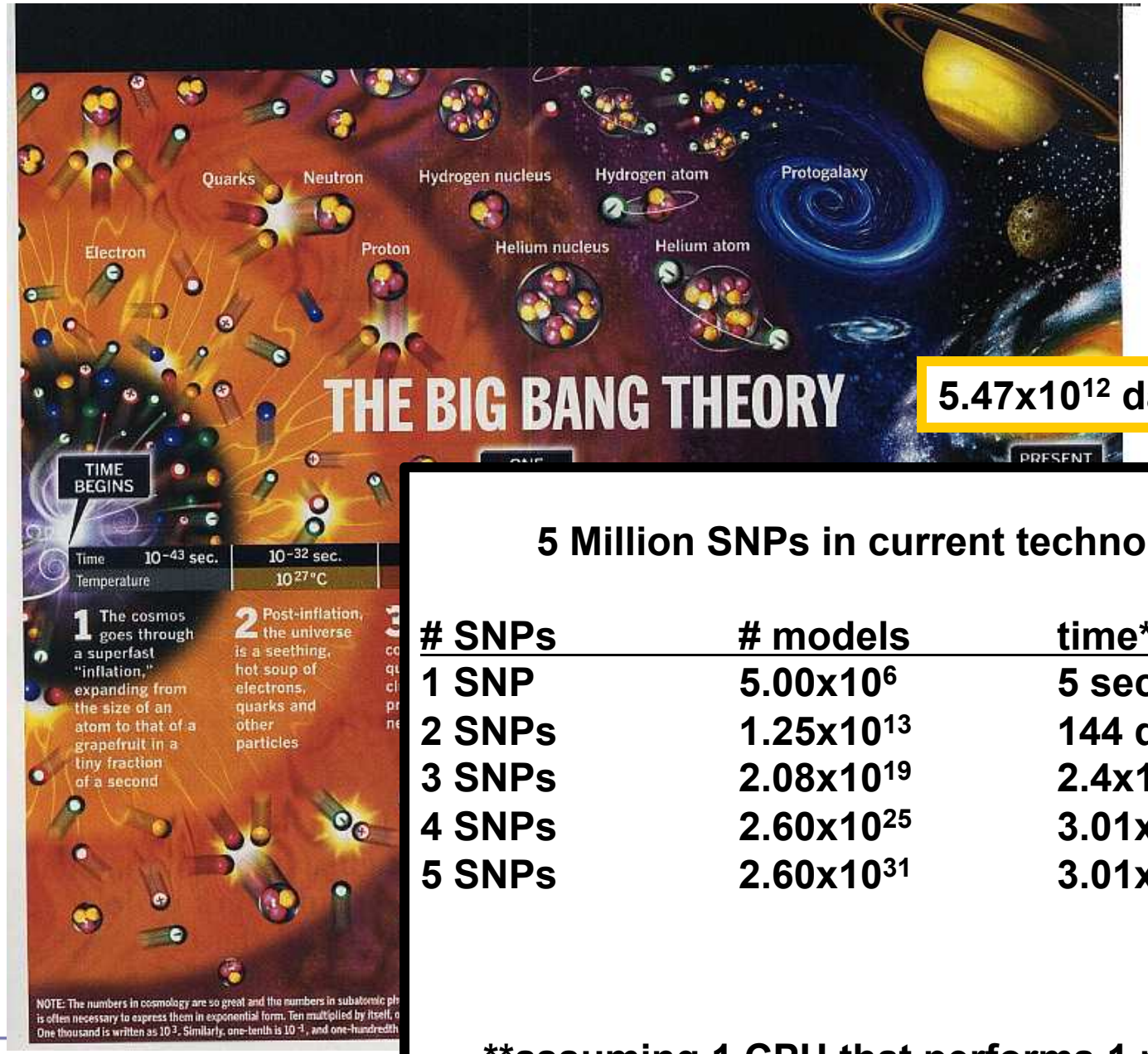
2 x 10^{26} combinations

5 Million SNPs in current technology

# SNPs	# models	time**
1 SNP	5.00×10^{6}	5 sec
2 SNPs	1.25×10^{13}	144 days
3 SNPs	2.08×10^{19}	2.4×10^{8} days
4 SNPs	2.60×10^{25}	3.01×10^{14} days
5 SNPs	2.60×10^{31}	3.01×10^{20} days

****assuming 1 CPU that performs 1 million tests per second**

PENNSTATE



5.47×10^{12} days

5 Million SNPs in current technology

# SNPs	# models	time**
1 SNP	5.00×10^{6}	5 sec
2 SNPs	1.25×10^{13}	144 days
3 SNPs	2.08×10^{19}	2.4×10^{8} days
4 SNPs	2.60×10^{25}	3.01×10^{14} days
5 SNPs	2.60×10^{31}	3.01×10^{20} days

****assuming 1 CPU that performs 1 million tests per second**

PENNSTATE
1855

Epistasis Analysis in GWAS data

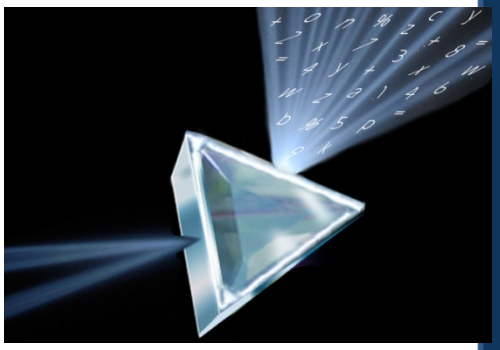
- Exhaustive evaluation
- Evaluate interactions in top hits from single-SNP analysis
- Use prior biological knowledge to evaluate specific combinations – "Candidate Epistasis"

PENNSTATE

Bush WS, Dudek SM, **Ritchie MD**. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing*, 368-79 (2009).

The Biofilter

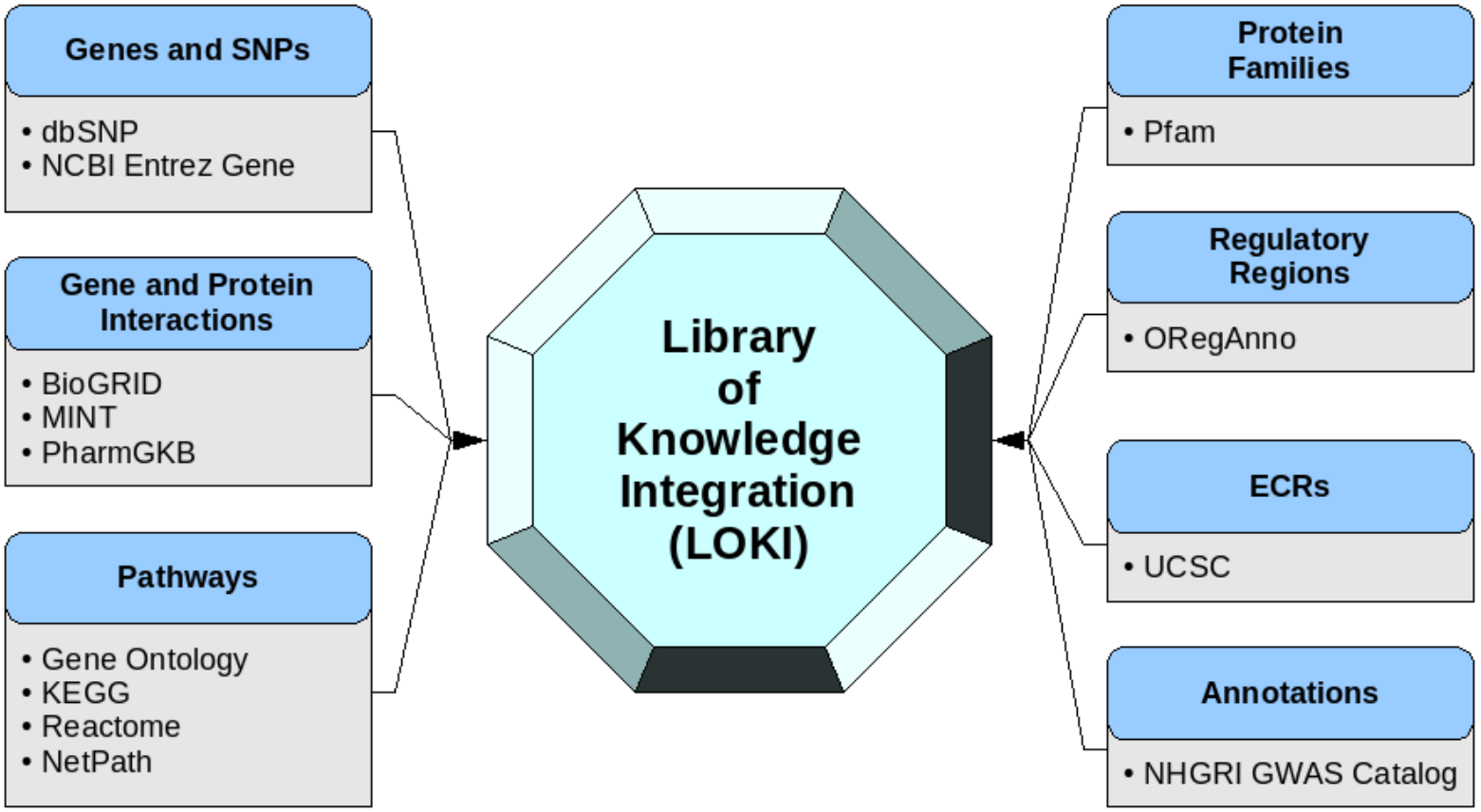
- Use publicly available databases to establish relationships between gene-products
- Suggestions of biological epistasis between genes
- Integrating information from the genome, transcriptome, and proteome into analysis



PENNSTATE

Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing*, 368-79 (2009).

LOKI: Library of Knowledge Integration



PENNSTATE

Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing*, 368-79 (2009).

The Biofilter

- Method described: Bush et al. 2009 *Pacific Symposium on Biocomputing*
- Applications
  - Multiple Sclerosis
    - Bush et al. 2009 *ASHG* talk, 2011 *Genes & Immunity*
  - HDL
    - Turner et al. 2010 *ASHG* Talk, 2011 *PLoS ONE*
  - HIV Pharmacogenomics
    - Grady et al. 2010 *ASHG* poster, 2011 *Pacific Symposium on Biocomputing*
  - Lipid traits
    - Holzinger et al. in preparation
  - BMI
    - Verma et al., in preparation
  - Cataracts
    - Hall et al., in preparation

PENNSTATE

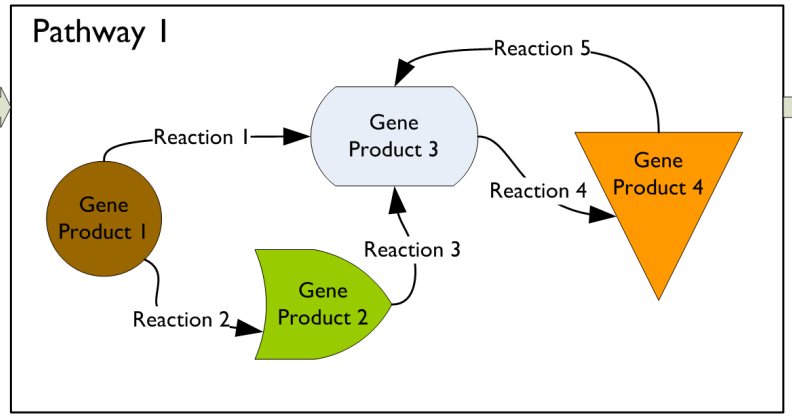
Using Biofilter: GWAS Annotation

Are there biological relationships between significant results?

Single Locus Statistical Results

SNP 1, Rs101841, p = 0.000163
SNP 2, Rs182645, p = 0.000268
SNP 3, Rs23876, p = 0.00324
SNP 4, Rs378645, p = 0.004354
SNP 5, Rs37564, p = 0.02341
SNP 6, Rs8751, p = 0.03412
SNP 7, Rs86745, p = 0.03685
SNP 8, Rs41254, p = 0.04675

Biofilter Analysis



Annotated Statistical Results

Results in the Same Gene
SNP 3, Rs23876, p = 0.00324
SNP 6, Rs8751, p = 0.03412

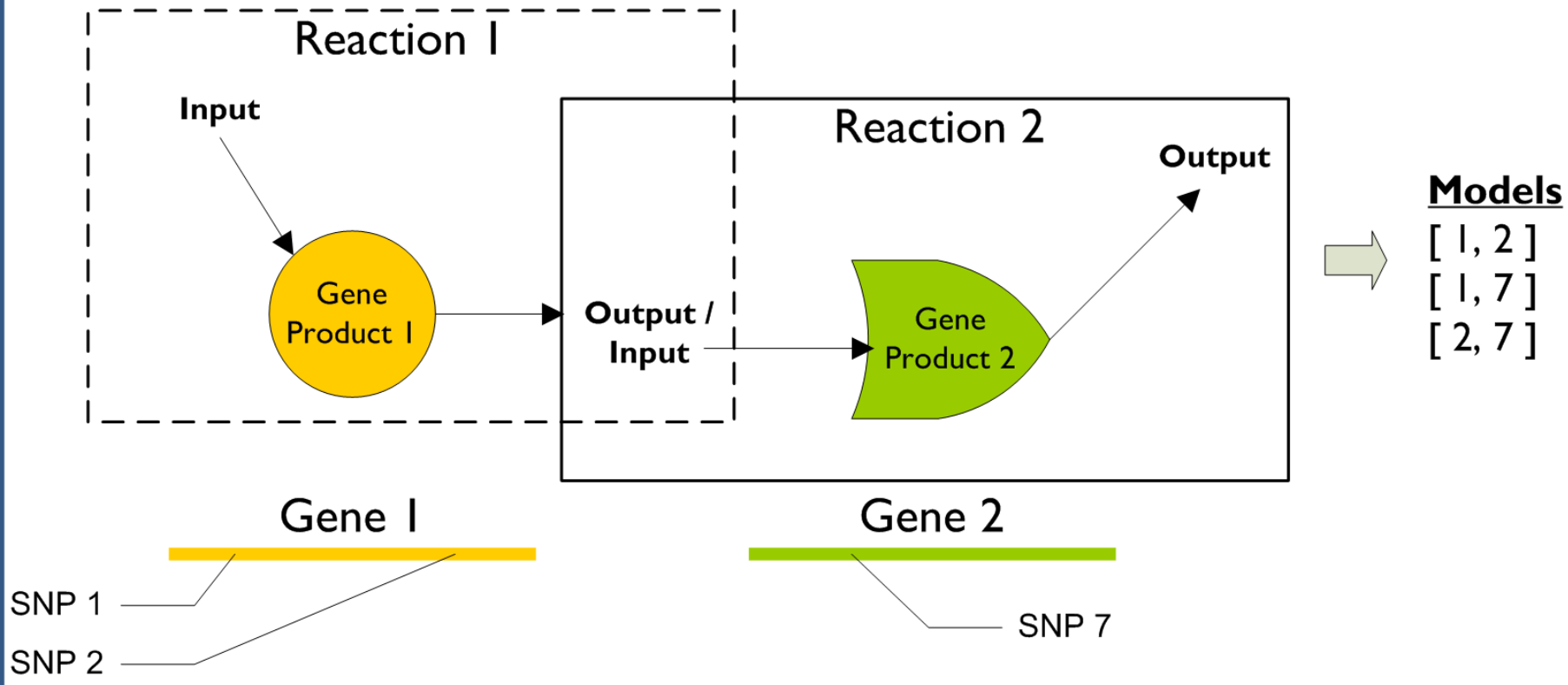
Results in the Same Pathway
SNP 2, Rs182645, p = 0.000268
SNP 3, Rs23876, p = 0.00324
SNP 6, Rs8751, p = 0.03412
SNP 7, Rs86745, p = 0.03685

Results with Biological Interaction
SNP 3, Rs23876, p = 0.00324
SNP 6, Rs8751, p = 0.03412
SNP 7, Rs86745, p = 0.03685

PENNSTATE
1855

Using Biofilter: Prioritizing Analysis

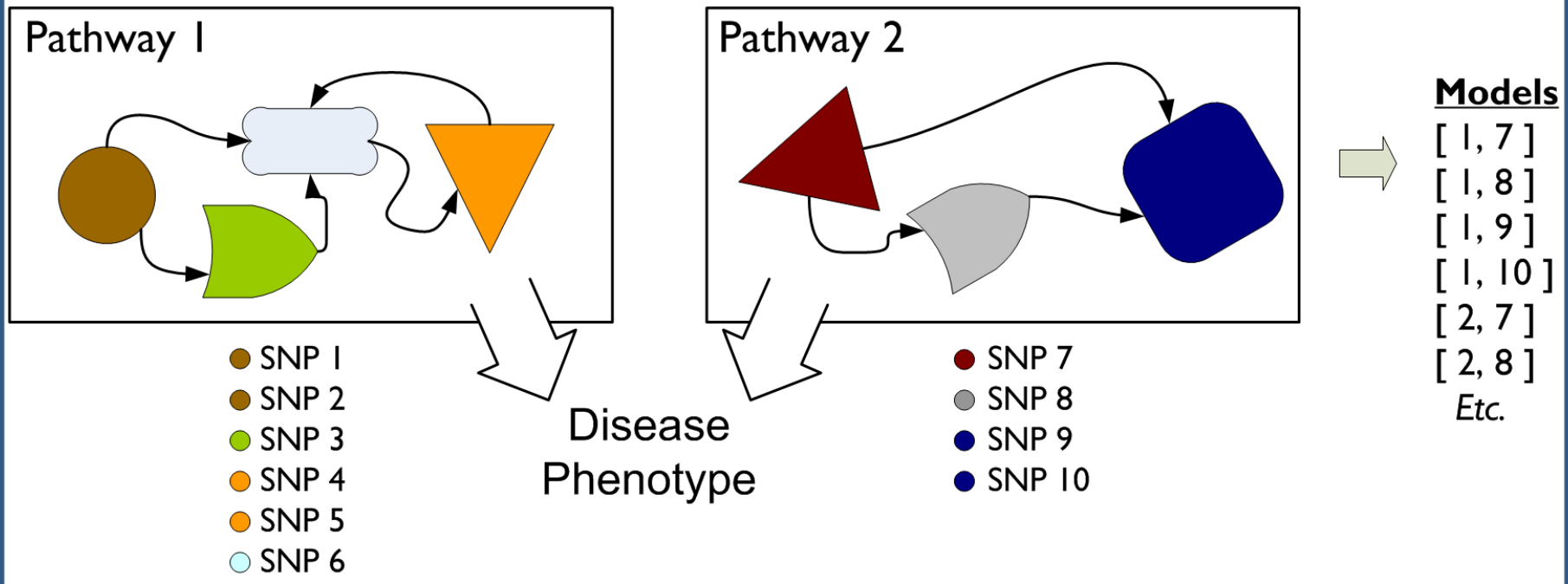
Is there epistasis in genes whose products interact either directly or through a metabolic intermediate?



PENNSTATE

Using Biofilter: Prioritizing Analysis

Is there epistasis between genes of two related pathways?



PENNSTATE

Candidate Approaches

Pros

- Smaller set of genes to explore
- Fewer statistical tests
- Results will have solid interpretations

Cons

- Limited by current state of knowledge
- Limitations of learning completely novel biology

PENNSTATE

Genes and Immunity (2011) 12, 335–340

www.nature.com/gene

ORIGINAL ARTICLE

A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility

WS Bush[1], JL McCauley[2], PL DeJager
L Kappos[5], Y Naegelin[5], CH Polman[6]
the International Multiple Sclerosis G
[1]*Department of Molecular Physiology and Biophys*
[2]*Miami Institute for Human Genomics, Universit*
Immunology, Center for Neurologic Diseases, Dep
Boston, MA, USA; [4]*GlaxoSmithKline, Research &*
Basel, Switzerland; [6]*Department of Neurology, Vr*

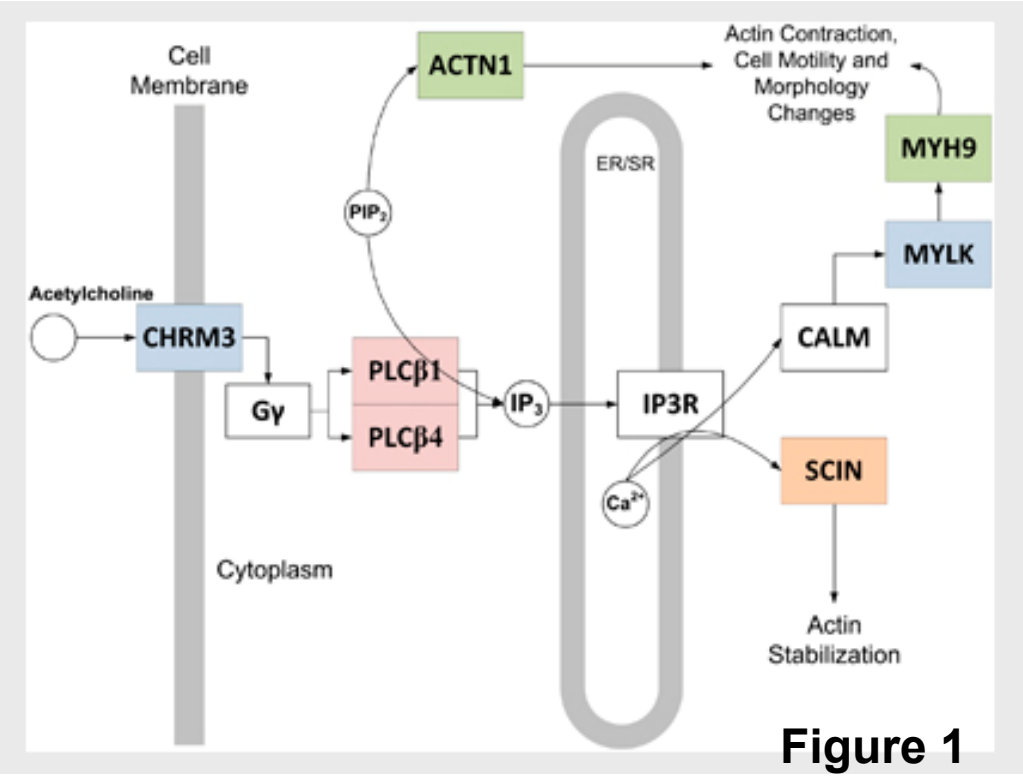


Figure 1

- 930 trio families
- Genotyped on Af
  - Post QC ~300,

PENNSTATE

OPEN ACCESS Freely available online

PLoS one

Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks

Stephen D. Turner[1], Richard L. Berg[2], James G. Linneman[2], Peggy L. Peissig[2], Dana C. Crawford[1], Joshua C. Denny[3], Dan M. Roden[4,5], Catherine A. McCarty[6], Marylyn D. Ritchie[1], Russell A. Wilke[4]*

1 Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **2** Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America, **3** Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **4** Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **5** Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **6** Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America

- eMERGE Genome-wide association study (Illumina 660)
- Phenotype: median HDL for anyone having 2+ HDL measurements in their EMR
- Marshfield PMRP n=3903
- Vanderbilt BioVU n=1858

Marshfield Clinic

Vanderbilt BioVU

PENNSTATE 1855

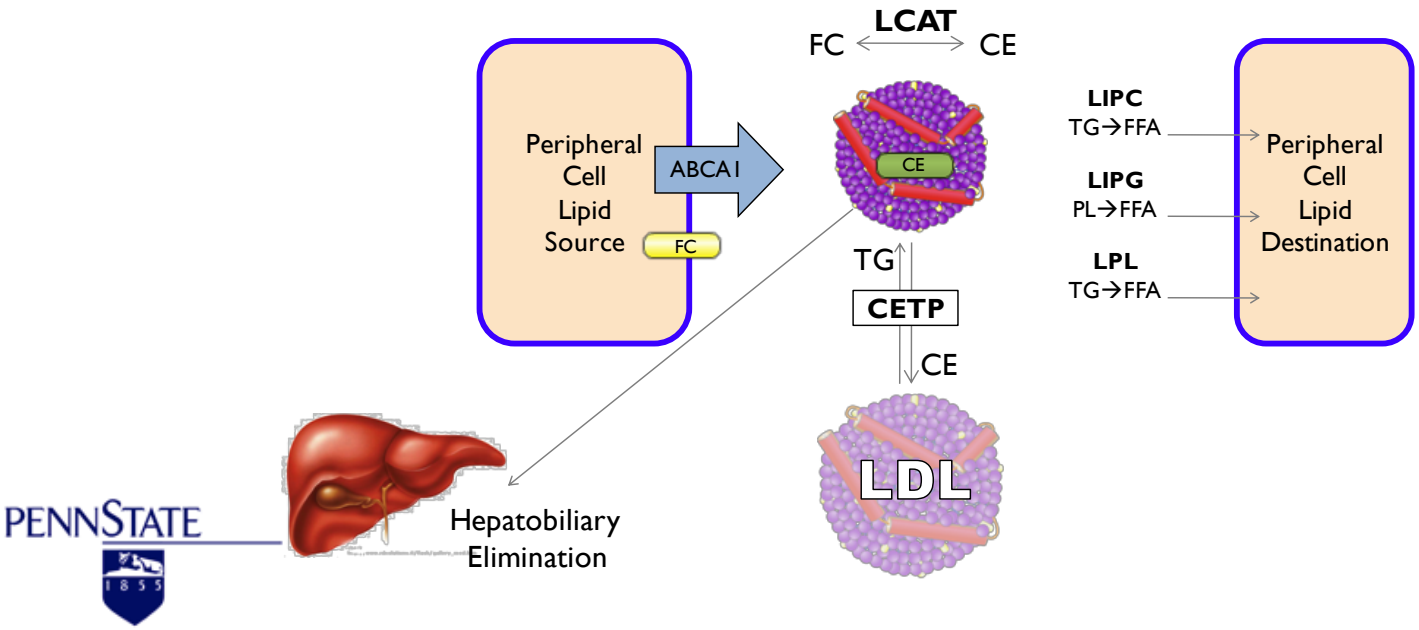
OPEN ACCESS Freely available online

PLoS one

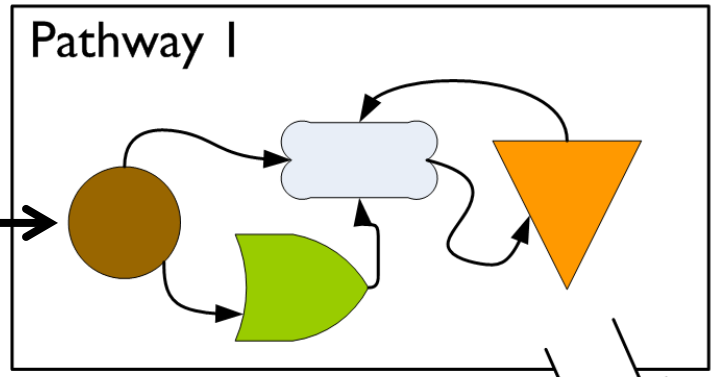
Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks

Stephen D. Turner[1], Richard L. Berg[2], James G. Linneman[2], Peggy L. Peissig[2], Dana C. Crawford[1], Joshua C. Denny[3], Dan M. Roden[4,5], Catherine A. McCarty[6], Marylyn D. Ritchie[1], Russell A. Wilke[4]*

1 Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **2** Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America, **3** Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **4** Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **5** Department of Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **6** Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America



Future Directions



2) Map SNPs –> gene -> pathway using Biofilter

1) SNPs from GWAS catalog for a particular disease-trait association

- SNP 1
- SNP 2
- SNP 3
- SNP 4
- SNP 5
- SNP 6

3) SNPs from KEGG, Reactome, or Netpath linked to SNPs from GWAS Catalog in LOKI

4) Exhaustive SNP-SNP models

SNP1 – SNP2
SNP1 – SNP3
SNP1 – SNP4
SNP1 – SNP5
. . .

PENNSTATE

Summary

- Biofilter is a bioinformatics application to annotate, filter, and construct gene-gene models for evaluation
- We have successfully used Biofilter in a number of genome-wide interaction analyses to identify replicating/confirmatory gene-gene models
- The GWAS catalog is an important and useful public database incorporated into LOKI – the knowledge base from which Biofilter draws its information

PENNSTATE

Future Directions

- Integrate more public databases into LOKI
  - Regulatory regions
  - Non-coding regions
- Develop additional filtering and model construction strategies based on specific hypotheses
- Develop a user-interface for ease of use

PENNSTATE

Acknowledgements

Ritchie Lab

Gretta Armstrong, project manager

Carrie Buchanan Moore, MD/PhD student*

Scott Dudek, software developer

Alex Frase, software developer*

Molly Hall, PhD student

Neerja Katiyar, PhD student*

Dokyoon Kim PhD, Postdoctoral fellow

Ruowang Li, PhD student

Sarah Pendergrass PhD, Research Associate*

Anurag Verma, Bioinformatics Programmer

Shefali Verma, Bioinformatics Analyst

John Wallace, software developer*

Dan Wolfe, bioinformatics research assistant*

HDL project - eMERGE

MS project - IMSGC

PENNSTATE 1855

* - working on Biofilter

Just because we have not found it yet, doesn't mean it's not there…..

www.genetic-programming.org

- marylyn.ritchie@psu.edu
- http://ritchielab.com

PENNSTATE