Significance of gene-gene interactions (epistasis)

PSB 2015 Tutorial

Marylyn D Ritchie, PhD

Director, Biomedical and Translational Informatics, Geisinger Clinic

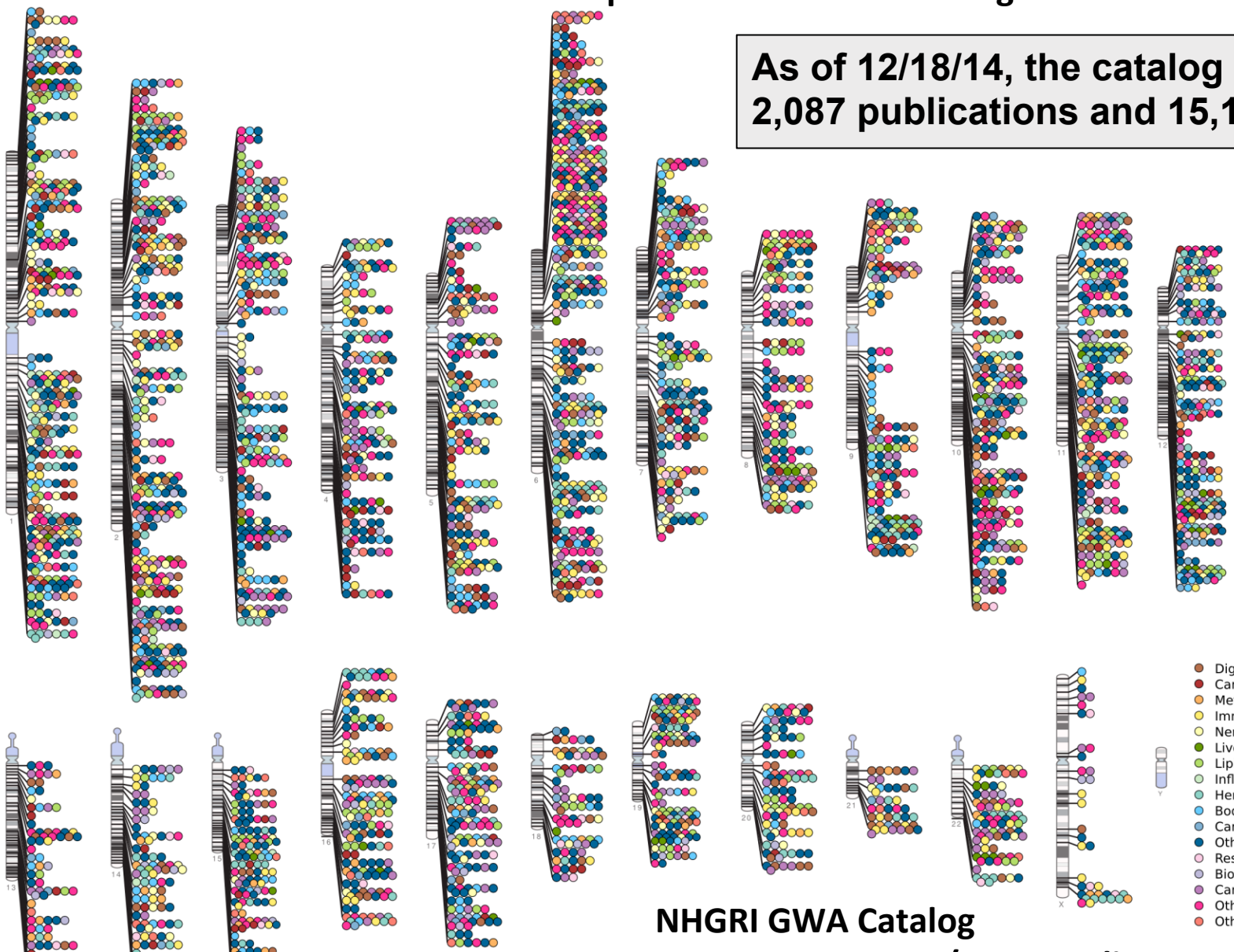
Professor, Biochemistry and Molecular Biology, The Pennsylvania State University

PENNSTATE 1855

GEISINGER
REDEFINING BOUNDARIES®

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

As of 12/18/14, the catalog includes 2,087 publications and 15,176 SNPs.

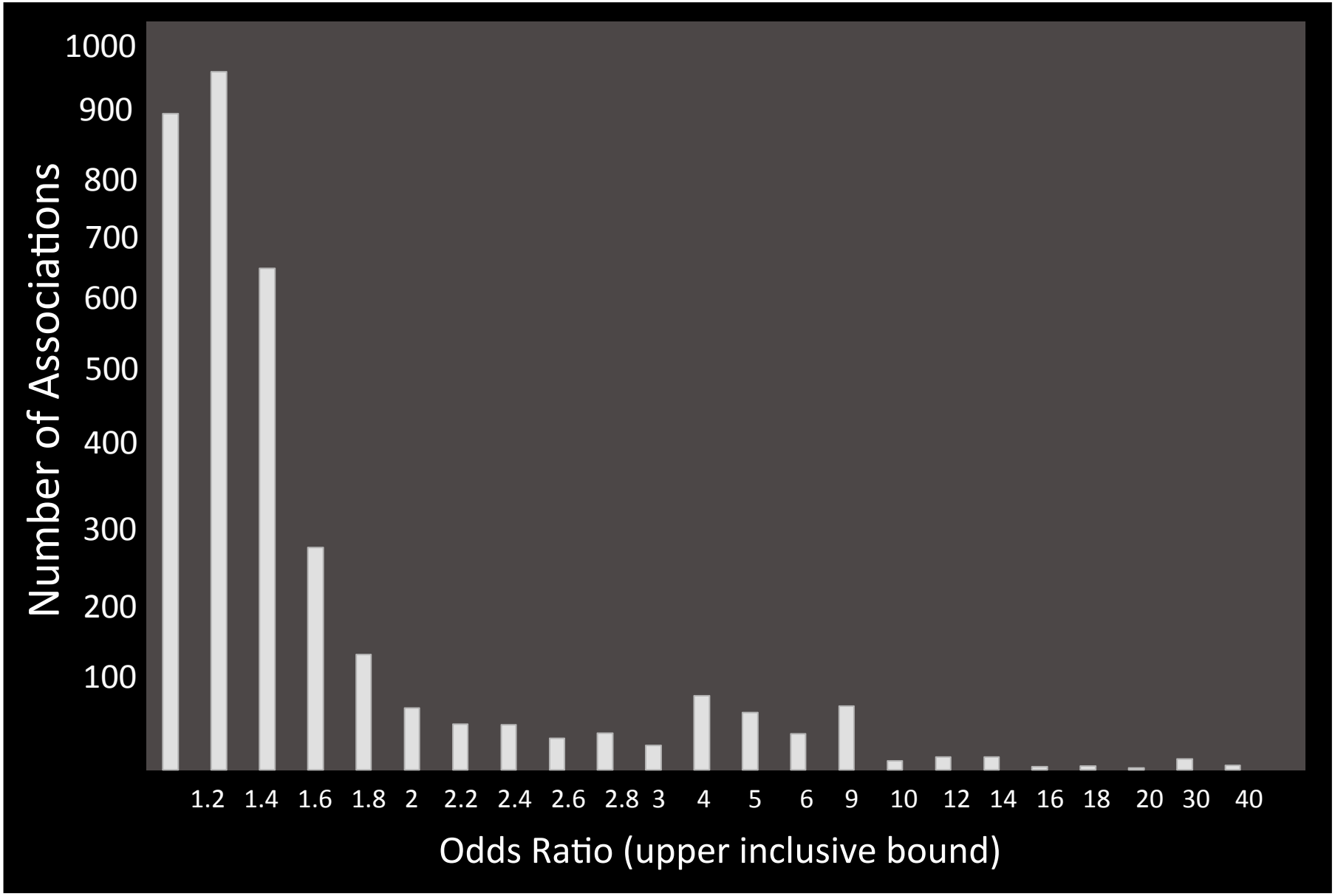


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

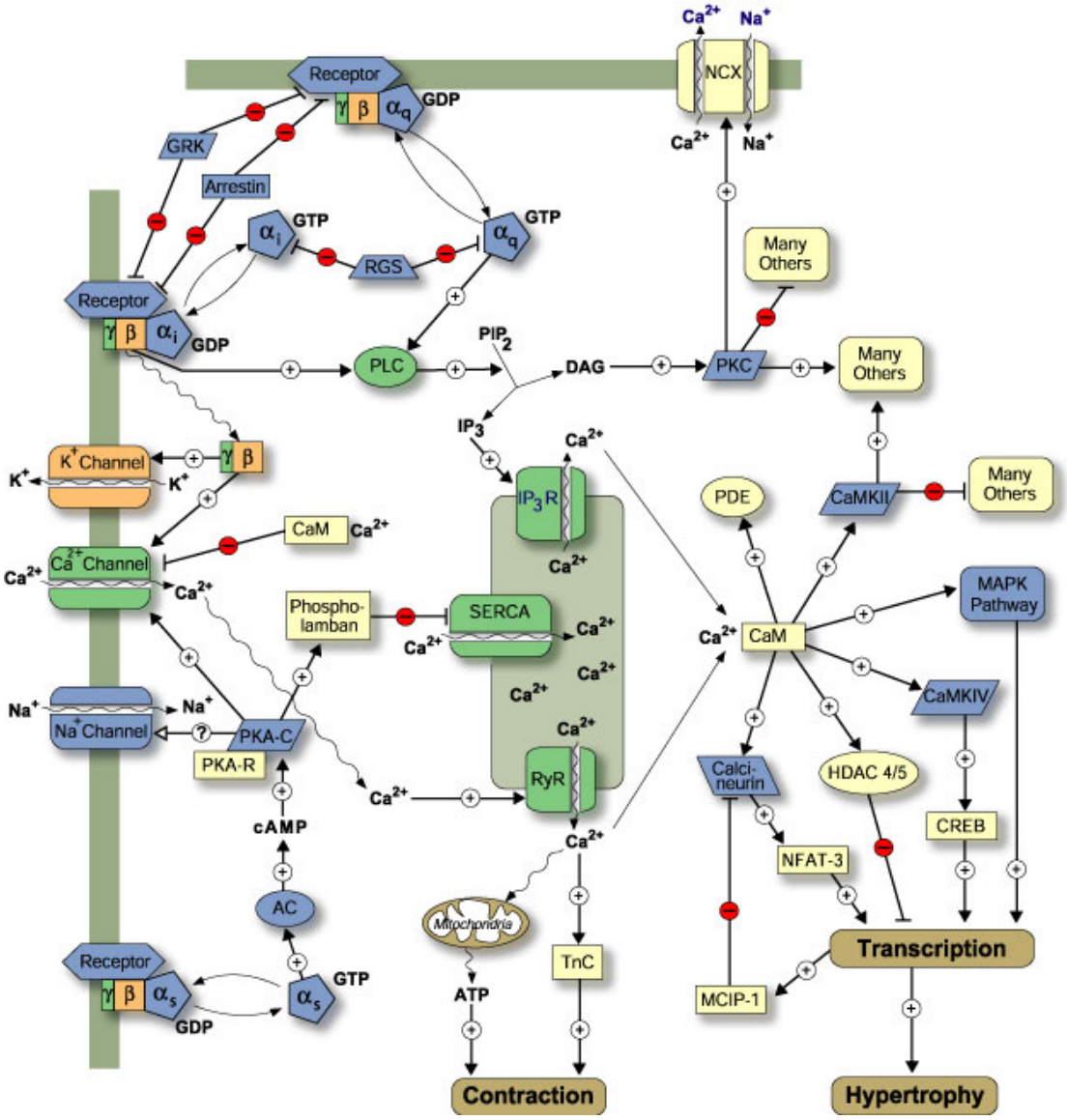
EMBL-EBI

Distribution of Effects



Marylyn Ritchie, Jan 2014

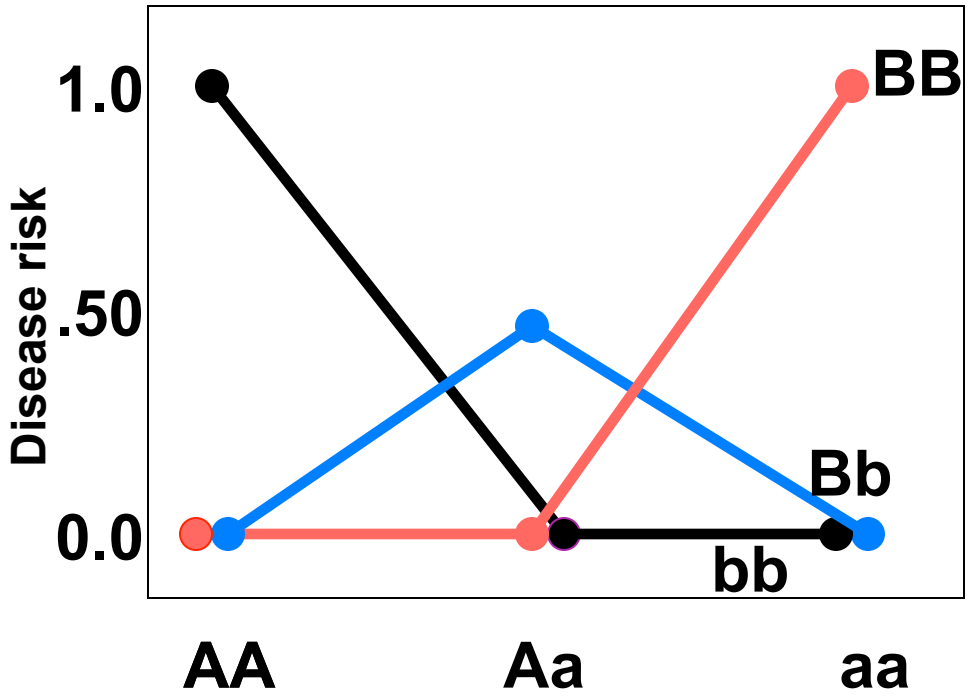
Biology is complex



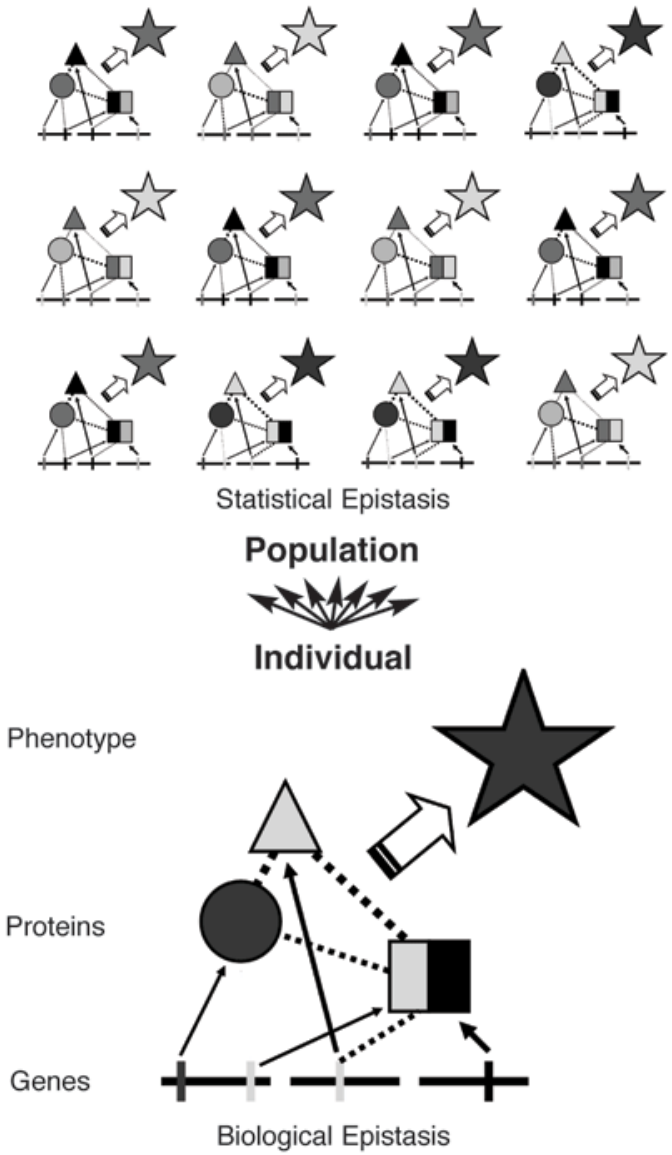
Epistasis

- Epistasis – two or more genes interacting in a non-additive manner to confer disease risk; gene-gene interactions

Genotype	p(D)
AABB	0.0
AABb	0.0
AAbb	1.0
AaBB	0.0
AaBb	.50
Aabb	0.0
aaBB	1.0
aaBb	0.0
aabb	0.0

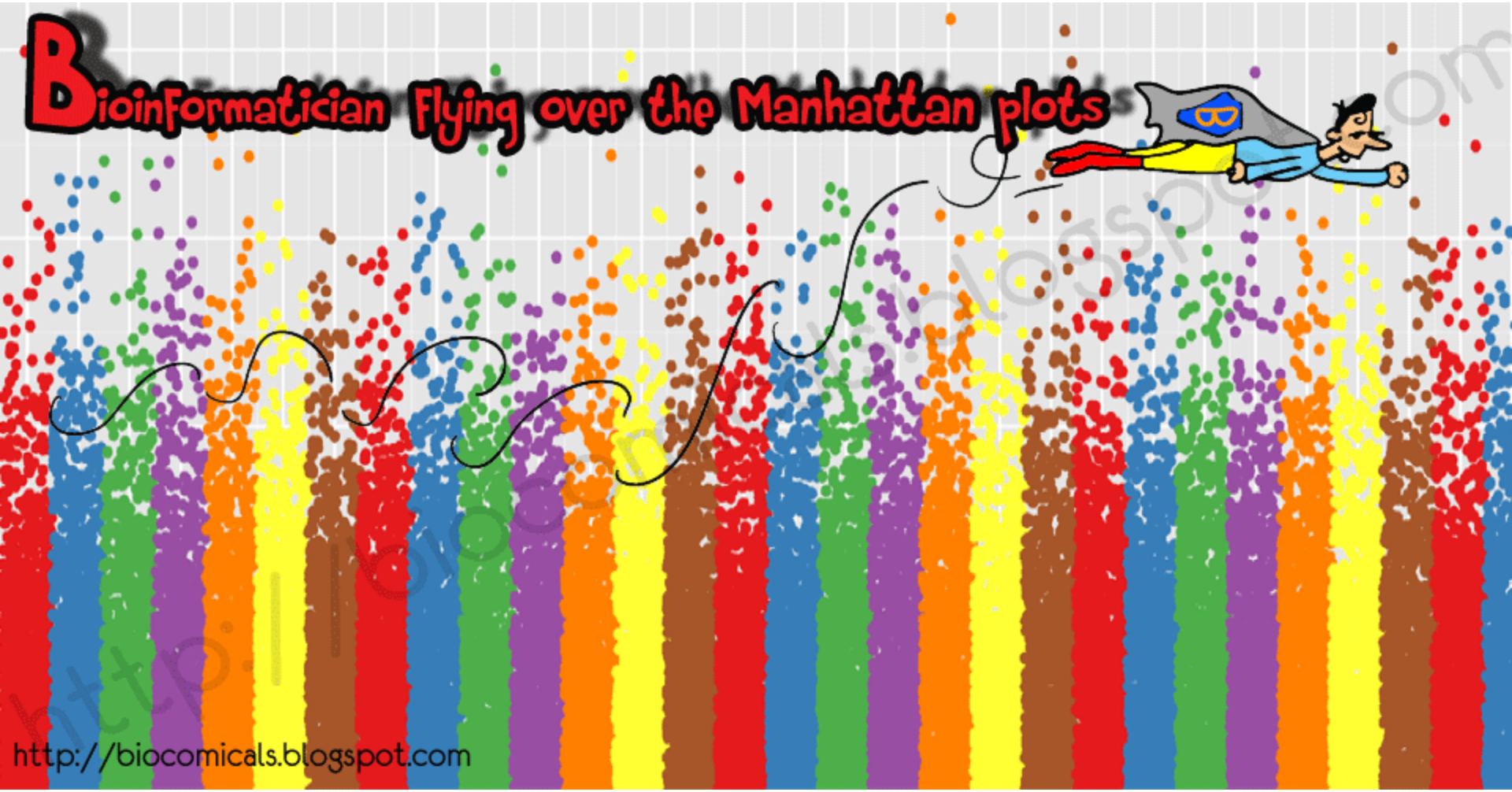


Statistical Epistasis vs. Biological Epistasis



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

Traditional Approach



Traditional Statistical Approaches

Genetic Epidemiology - 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 Approaches

Genetic Epidemiology - Association Analysis

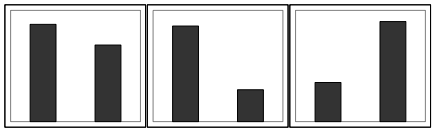
- Logistic Regression
  - Small sample size can result in biased estimates of regression coefficients and can result in spurious associations (Concato et al. 1993)
  - Need at least 10 cases or controls per independent variable to have enough statistical power (Peduzzi et al. 1996)
  - Curse of dimensionality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

Curse of Dimensionality

N = 100 **50 Cases, 50 Controls**

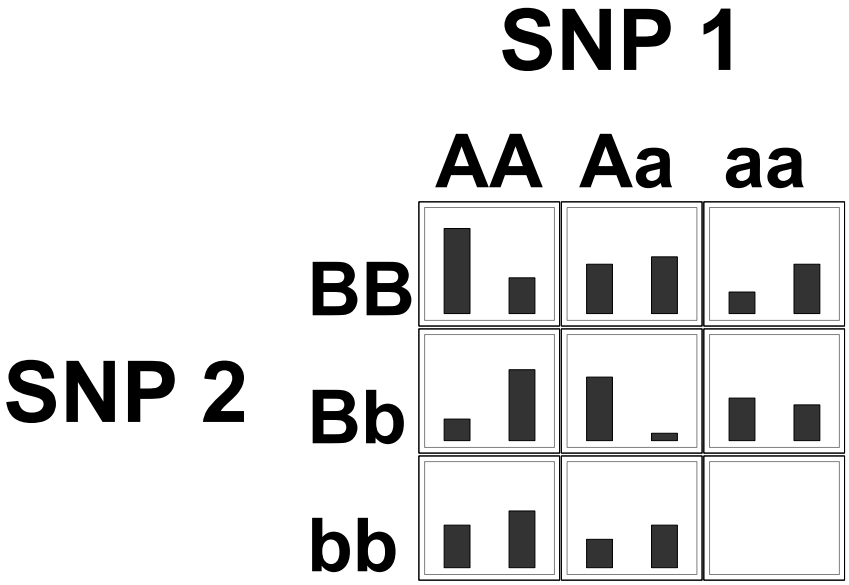
SNP 1

AA Aa aa



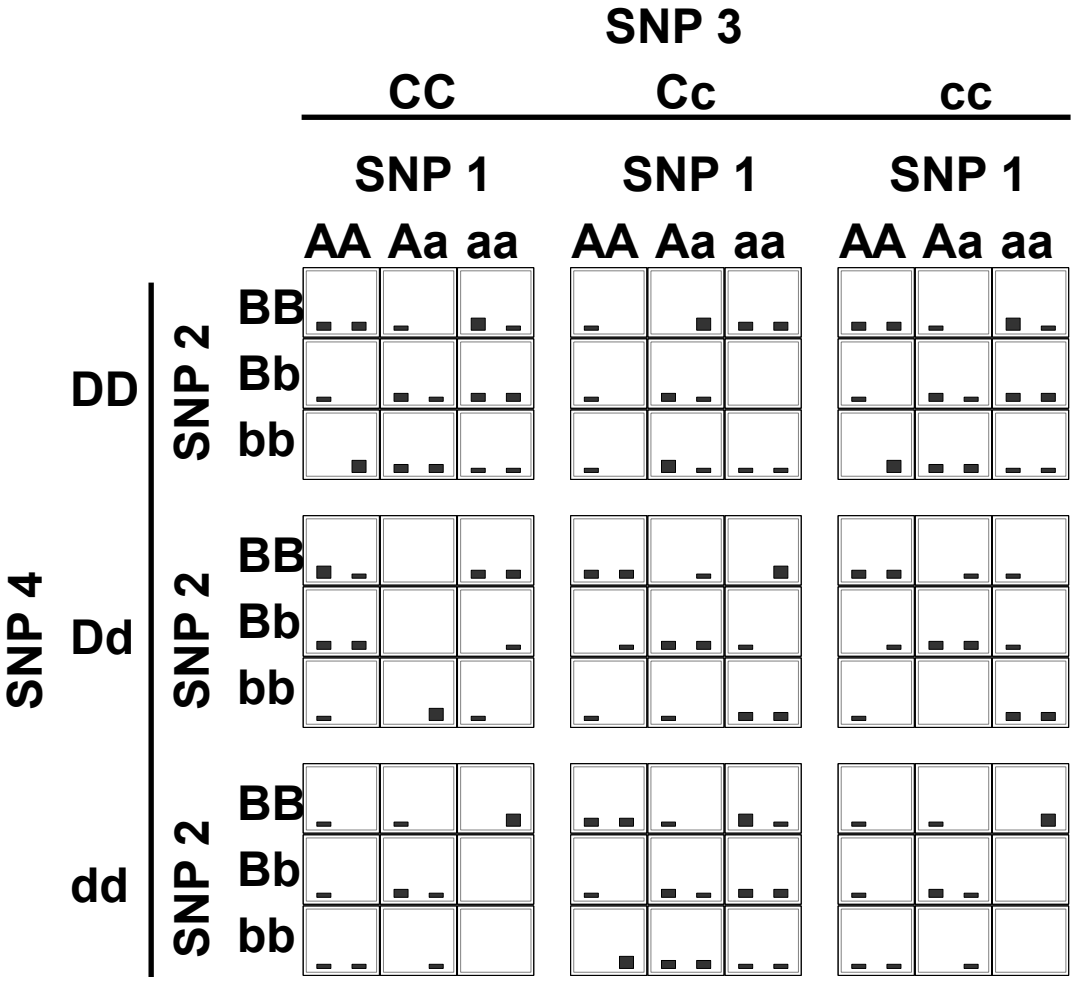
Curse of Dimensionality

N = 100 **50 Cases, 50 Controls**



Curse of Dimensionality

N = 100 **50 Cases, 50 Controls**

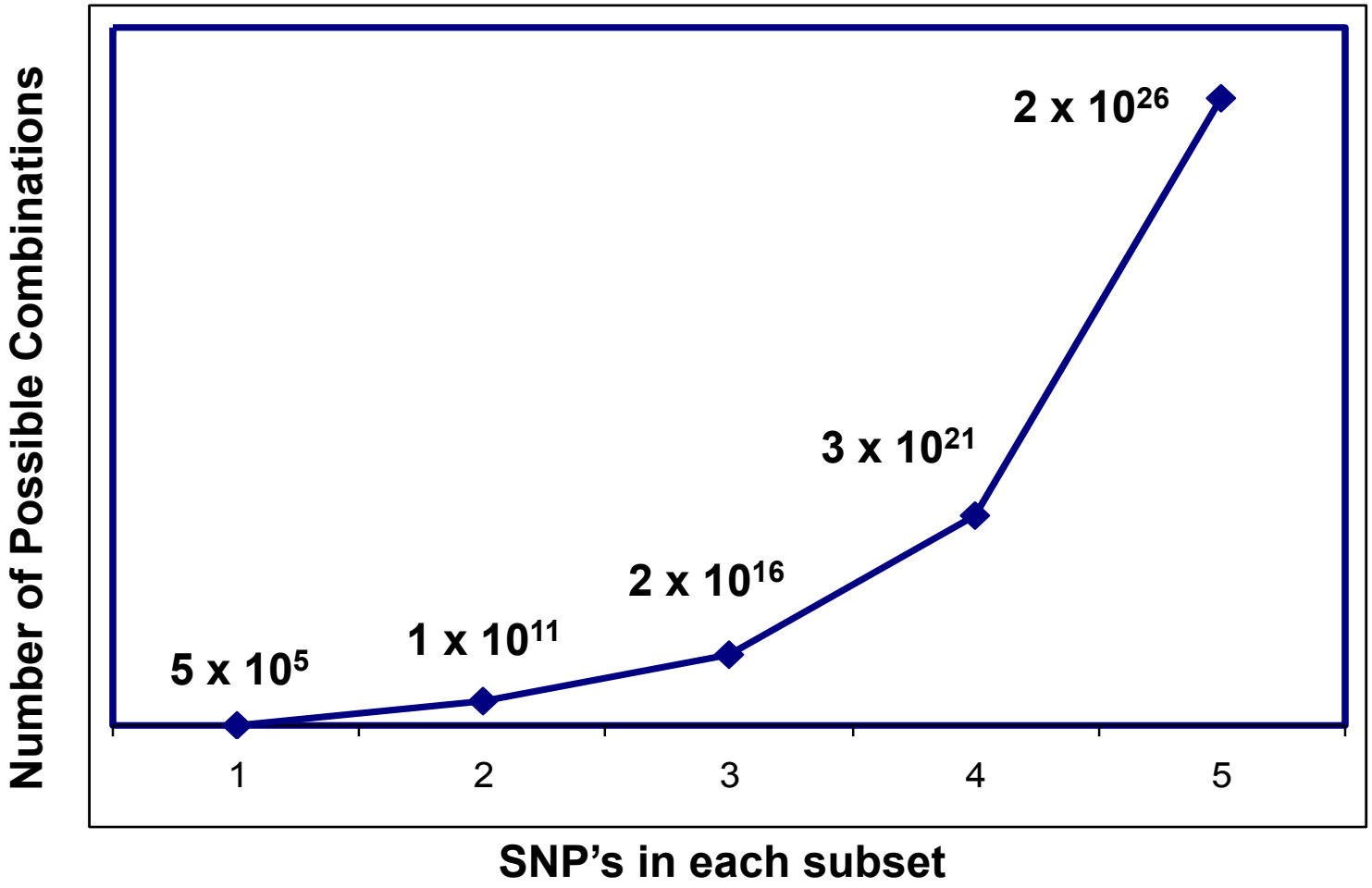


If interactions with minimal main effects are the norm rather than the exception, can we analyze all possible combinations of loci with traditional approaches to detect purely interaction effects ?

NO

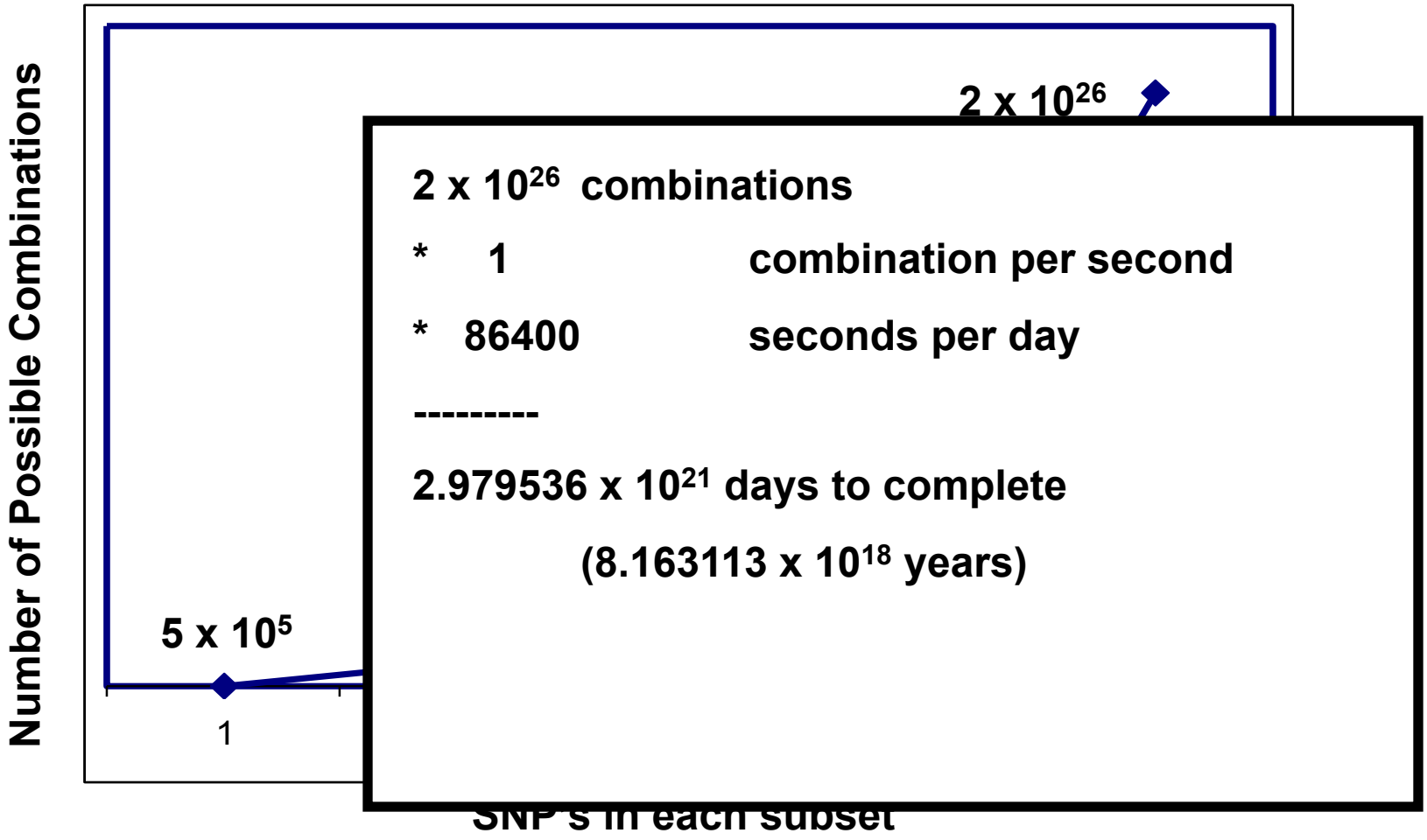
How many combinations are there?

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



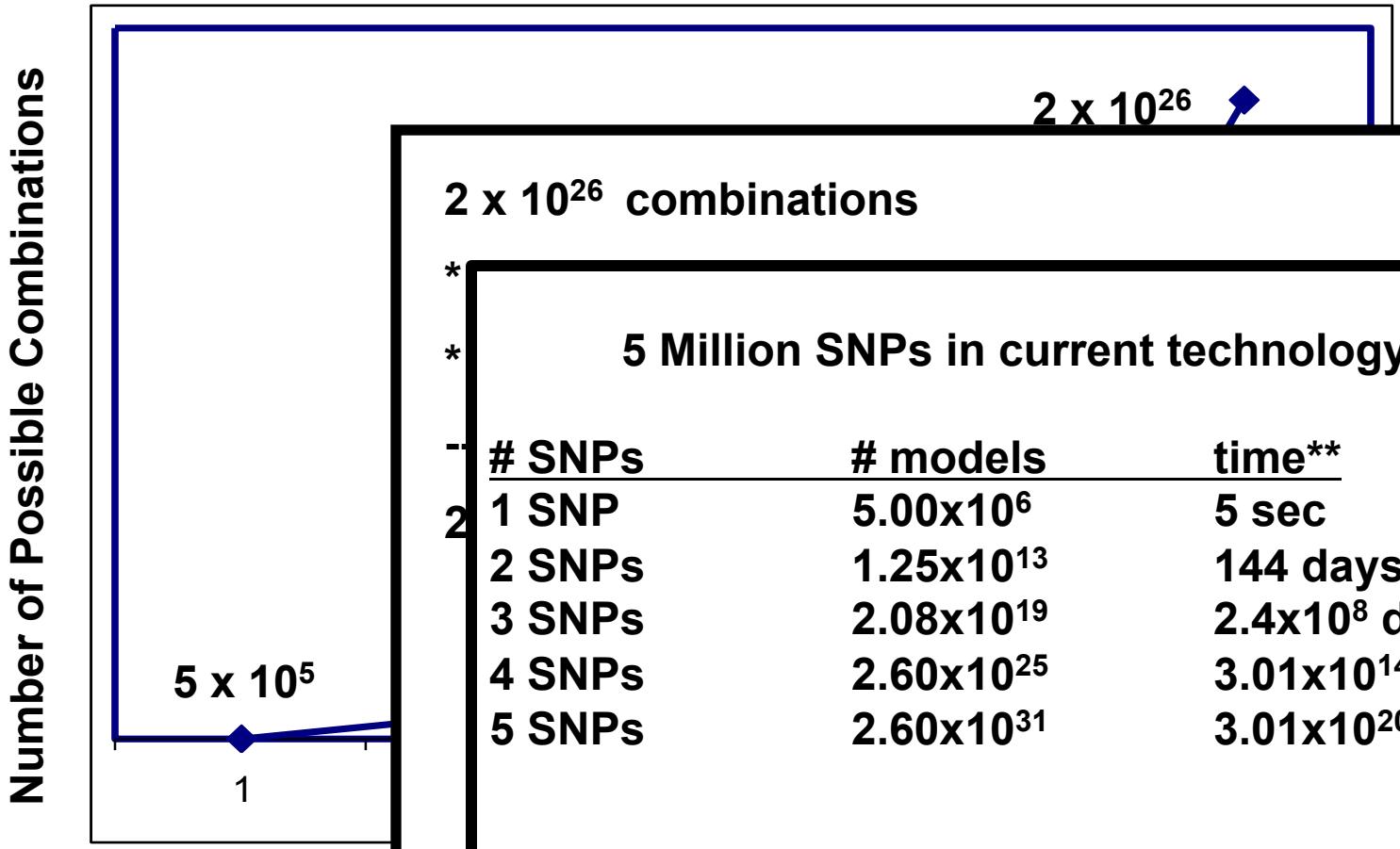
How many combinations are there?

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



How many combinations are there?

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



2 x 10^{26} combinations

*

*

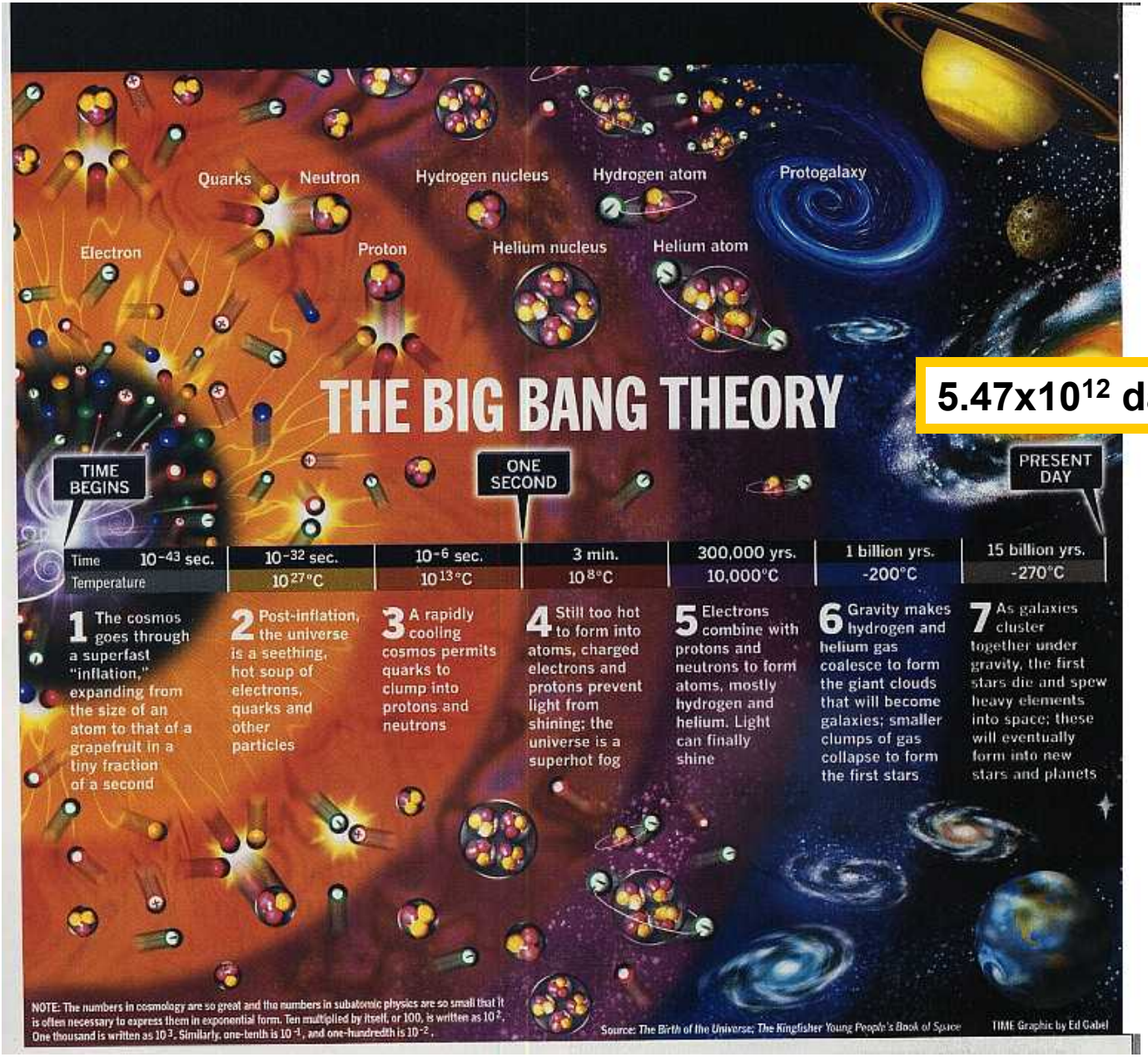
--

2

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



5.47×10^{12} days

Traditional Approach

- Advantages
  - Computationally feasible
  - Easy to interpret
- Disadvantages
  - Genes must have large main effects
  - Difficult to detect genes if interactions with other genetic and environmental factors are important
  - CANNOT do an exhaustive search

New Statistical Approaches

- Review paper

For reprint orders, please contact:
reprints@futuremedicine.com

REVIEW

Novel methods for detecting epistasis in pharmacogenomics studies

Alison A Motsinger[1], Marylyn D Ritchie[2] & David M Reif[8†]

†Author for correspondence
[1]North Carolina State University, Bioinformatics Research Center, Department of Statistics, Raleigh, NC 27695, USA
[2]Vanderbilt University, Center for Human Genetics Research, Department of Molecular

The importance of gene–gene and gene–environment interactions in the underlying genetic architecture of common, complex phenotypes is gaining wide recognition in the field of pharmacogenomics. In epidemiological approaches to mapping genetic variants that predict drug response, it is important that researchers investigate potential epistatic interactions. In the current review, we discuss data-mining tools available in genetic epidemiology to detect such interactions and appropriate applications. We survey several classes of novel methods available and present an organized collection of successful applications in the literature. Finally, we provide guidance as to how to incorporate these novel methods into a genetic analysis. The overall goal of this paper is to aid researchers in developing an analysis plan that accounts for gene–gene and gene–environment in their own work.

  - Pharmacogenomics. 2007 8(9) :1229-41.
- Reviews approximately 40 methods developed to detect gene-gene and gene-environment interactions

New Statistical Approaches

Chen *et al. BMC Genomics* 2011, **12**:344
http://www.biomedcentral.com/1471-2164/12/344

BMC Genomics

METHODOLOGY ARTICLE Open Access

Comparative analysis of methods for detecting interacting loci

Li Chen[1], Guoqiang Yu[1], Carl D Langefeld[2], David J Miller[3], Richard T Guy[2], Jayaram Raghuram[3], Xiguo Yuan[1], David M Herrington[4] and Yue Wang[1*]

Abstract

Background: Interactions among genetic loci are believed to play an important role in disease risk. While many methods have been proposed for detecting such interactions, their relative performance remains largely unclear, mainly because different data sources, detection performance criteria, and experimental protocols were used in the papers introducing these methods and in subsequent studies. Moreover, there have been very few studies strictly focused on comparison of existing methods. Given the importance of detecting gene-gene and gene-environment interactions, a rigorous, comprehensive comparison of performance and limitations of available interaction detection methods is warranted.

New Statistical Approaches

Shang *et al. BMC Bioinformatics* 2011, **12**:475
http://www.biomedcentral.com/1471-2105/12/475

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Performance analysis of novel methods for detecting epistasis

Junliang Shang[1*], Junying Zhang[1*], Yan Sun[2], Dan Liu[1], Daojun Ye[1] and Yaling Yin[1,3]

Abstract

Background: Epistasis is recognized fundamentally important for understanding the mechanism of disease-causing genetic variation. Though many novel methods for detecting epistasis have been proposed, few studies focus on their comparison. Undertaking a comprehensive comparison study is an urgent task and a pathway of the methods to real applications.

Results: This paper aims at a comparison study of epistasis detection methods through applying related software packages on datasets. For this purpose, we categorize methods according to their search strategies, and select five representative methods (TEAM, BOOST, SNPRuler, AntEpiSeeker and epiMODE) originating from different underlying techniques for comparison. The methods are tested on simulated datasets with different size, various epistasis

Simple Fitness Landscape

Fitness

Model

Complex Fitness Landscape

Fitnes

Model

Epistasis in GWAS Data

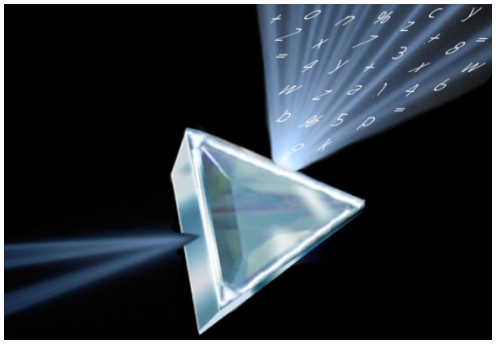
- ~~Exhaustive evaluation~~
- Evaluate interactions in top hits from single-SNP analysis
- Use prior biological knowledge to evaluate specific combinations – “Candidate Epistasis”

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* 2004 May 27;429(6990):446-52.

Goal: to build biologically plausible models of gene-gene interactions to test for association using an automated bioinformatics tool based on biological features

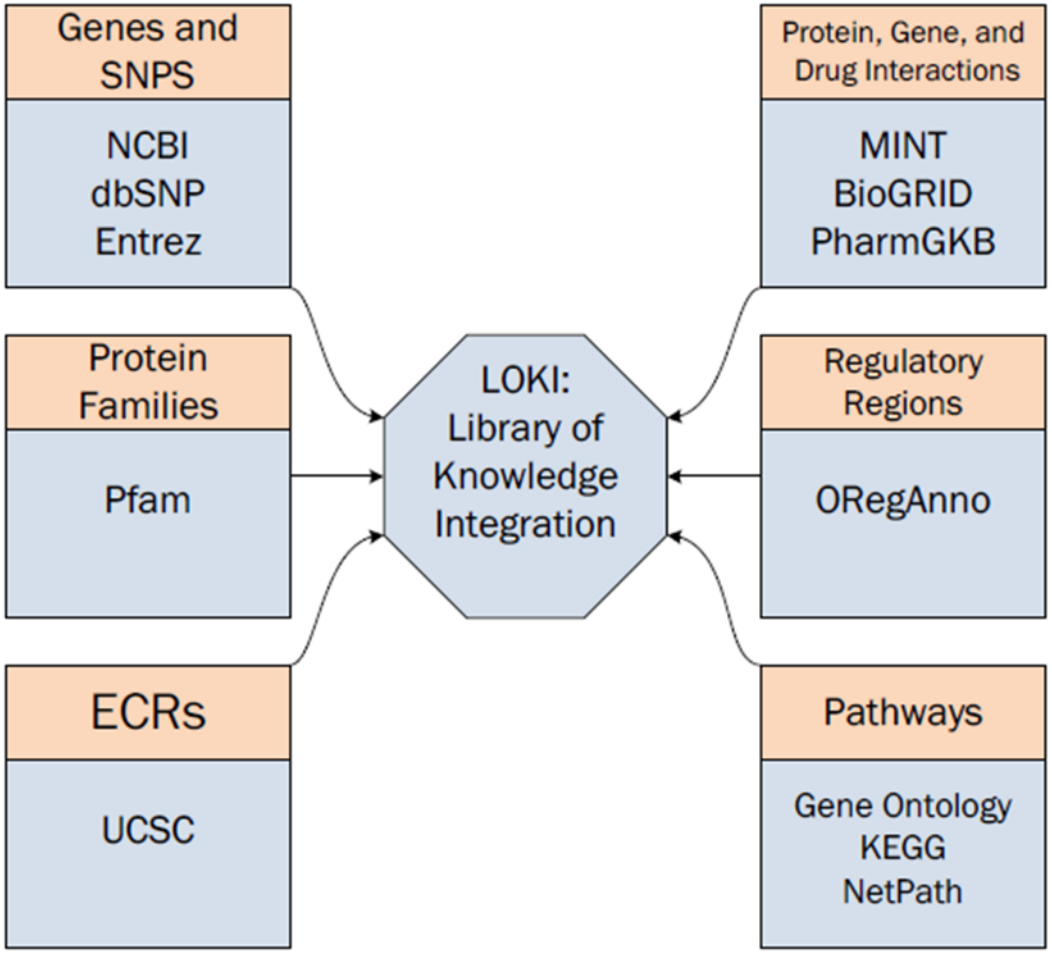
The Biofilter

- 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

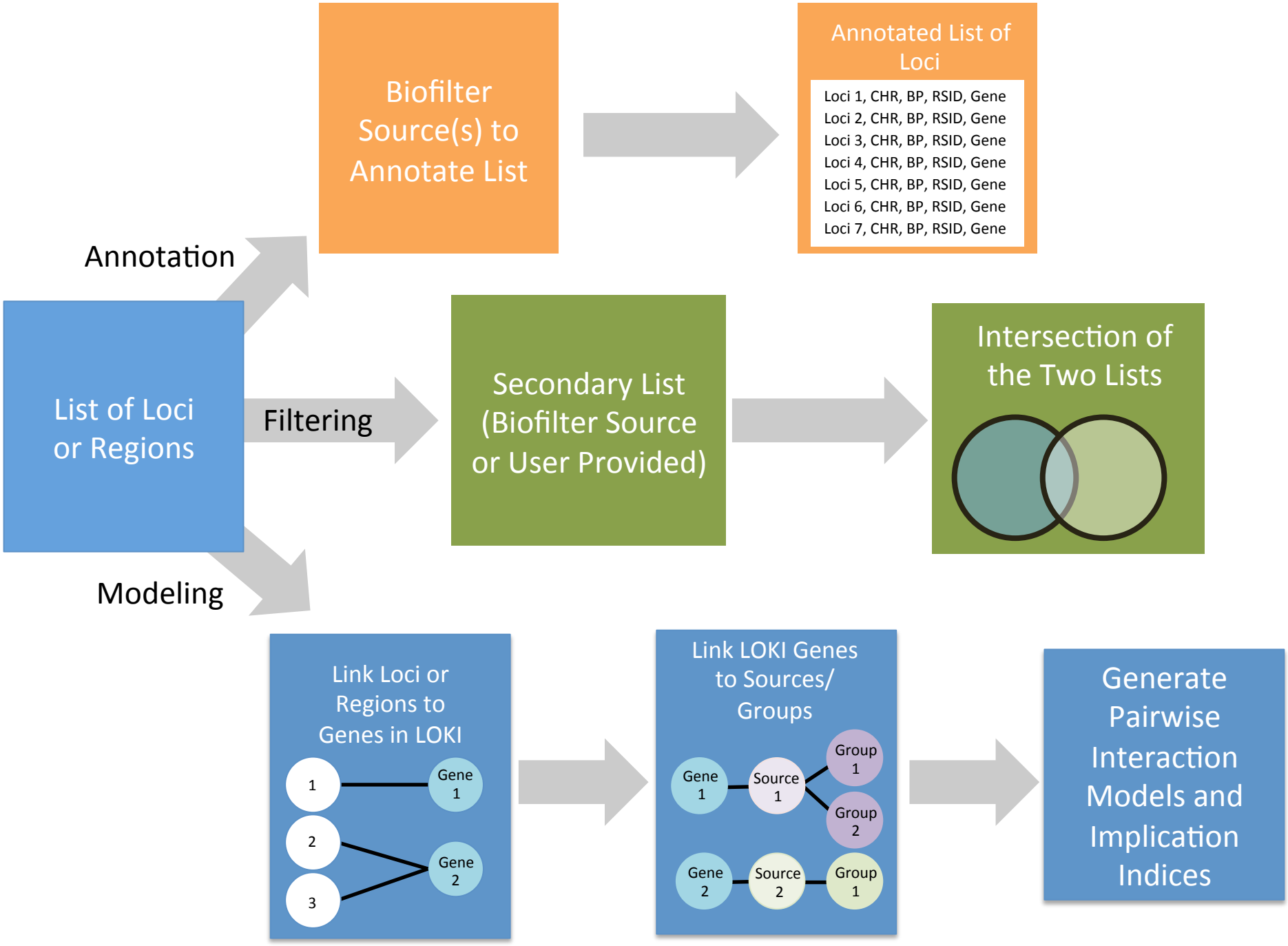


Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing*, 368-79 (2009).

LOKI: Library of Knowledge Integration



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).


Biofilter Source(s) to Annotate List
Annotated List of Loci
Loci 1, CHR, BP, RSID, Gene
Loci 2, CHR, BP, RSID, Gene
Loci 3, CHR, BP, RSID, Gene
Loci 4, CHR, BP, RSID, Gene
Loci 5, CHR, BP, RSID, Gene
Loci 6, CHR, BP, RSID, Gene
Loci 7, CHR, BP, RSID, Gene
Annotation
List of Loci or Regions
Filtering
Secondary List (Biofilter Source or User Provided)
Intersection of the Two Lists
Modeling
Link Loci or Regions to Genes in LOKI
1
2
3
Gene 1
Gene 2
Link LOKI Genes to Sources/ Groups
Gene 1
Source 1
Group 1
Group 2
Gene 2
Source 2
Group 1
Generate Pairwise Interaction Models and Implication Indices

Summary

- Gene-gene interactions are important components of complex trait genetic architecture
- Gene-gene interactions are challenging to detect:
  - Due to data sparseness in high dimensions
  - Due to the combinatorics of the search
  - Due to complexity
- Much research is ongoing to develop novel methods and strategies to address these issues