Beyond single genes or proteins

Marylyn D Ritchie, PhD

Professor, Biochemistry and Molecular Biology

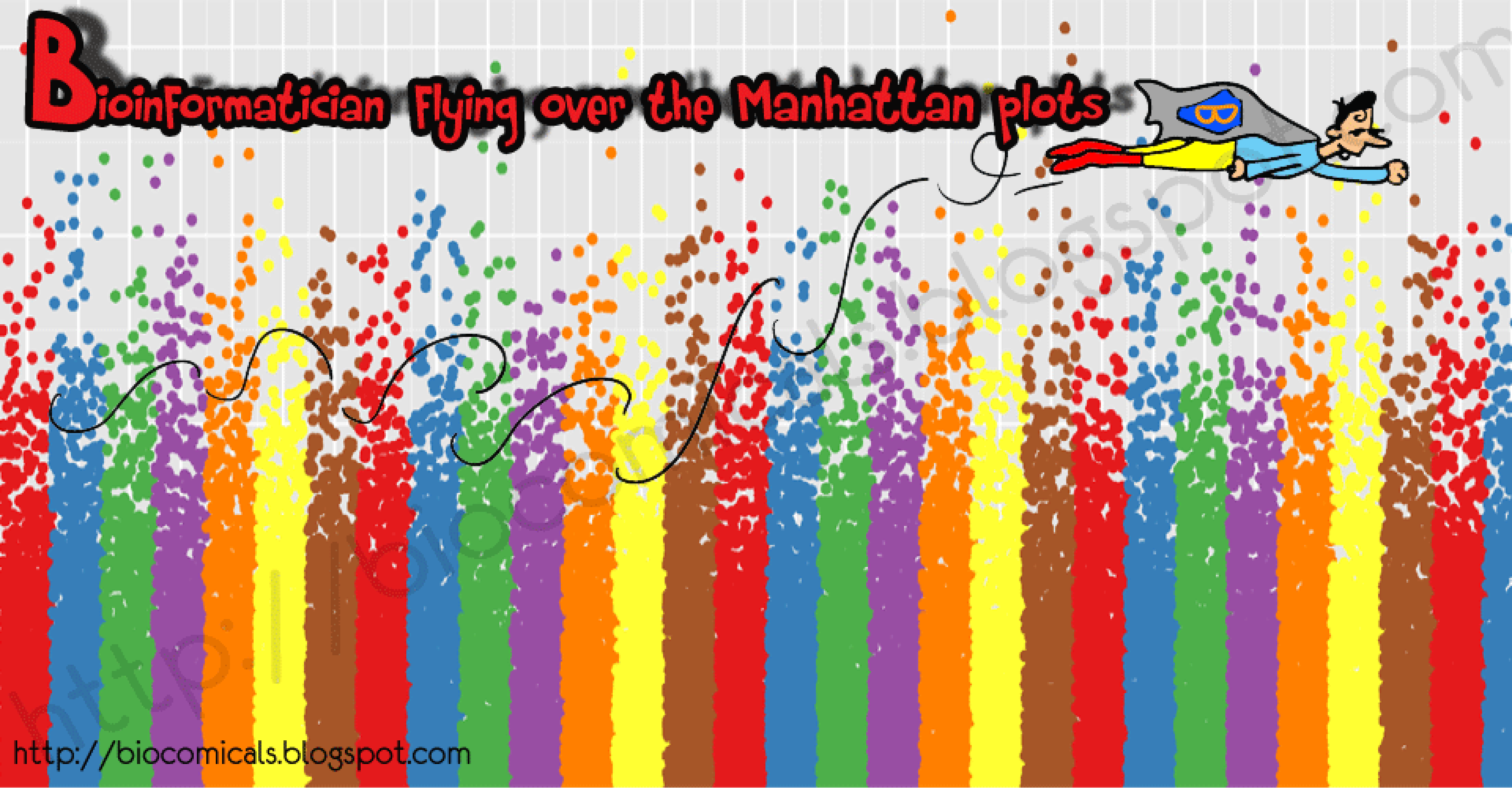
Director, Center for Systems Genomics

The Pennsylvania State University

Center for Systems Genomics
Pennsylvania State University

Ritchie Lab
Pennsylvania State University

Traditional Approach

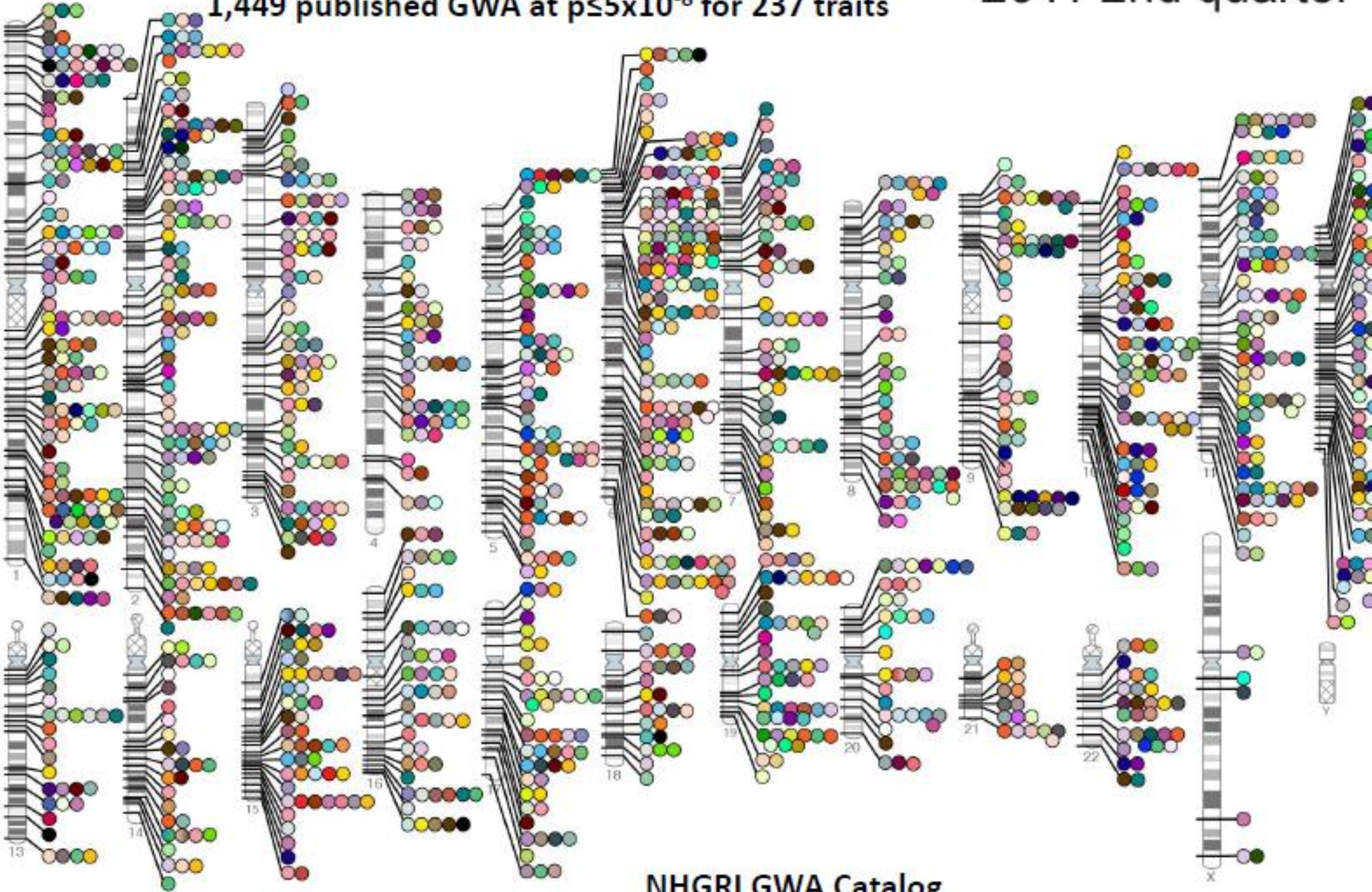

Bioinformatician Flying over the Manhattan plots
http://biocomicals.blogspot.com

Genome-wide Association Studies (GWAS)

- Technology has advanced rapidly creating many molecular genetic tools for data generation
- Hundreds of thousands to millions of markers
- Hundreds to thousands of individuals
  - Population based
  - Family based
- Whole genome sequencing is the new frontier c data generation
  - Increasing data at all levels of biological variation

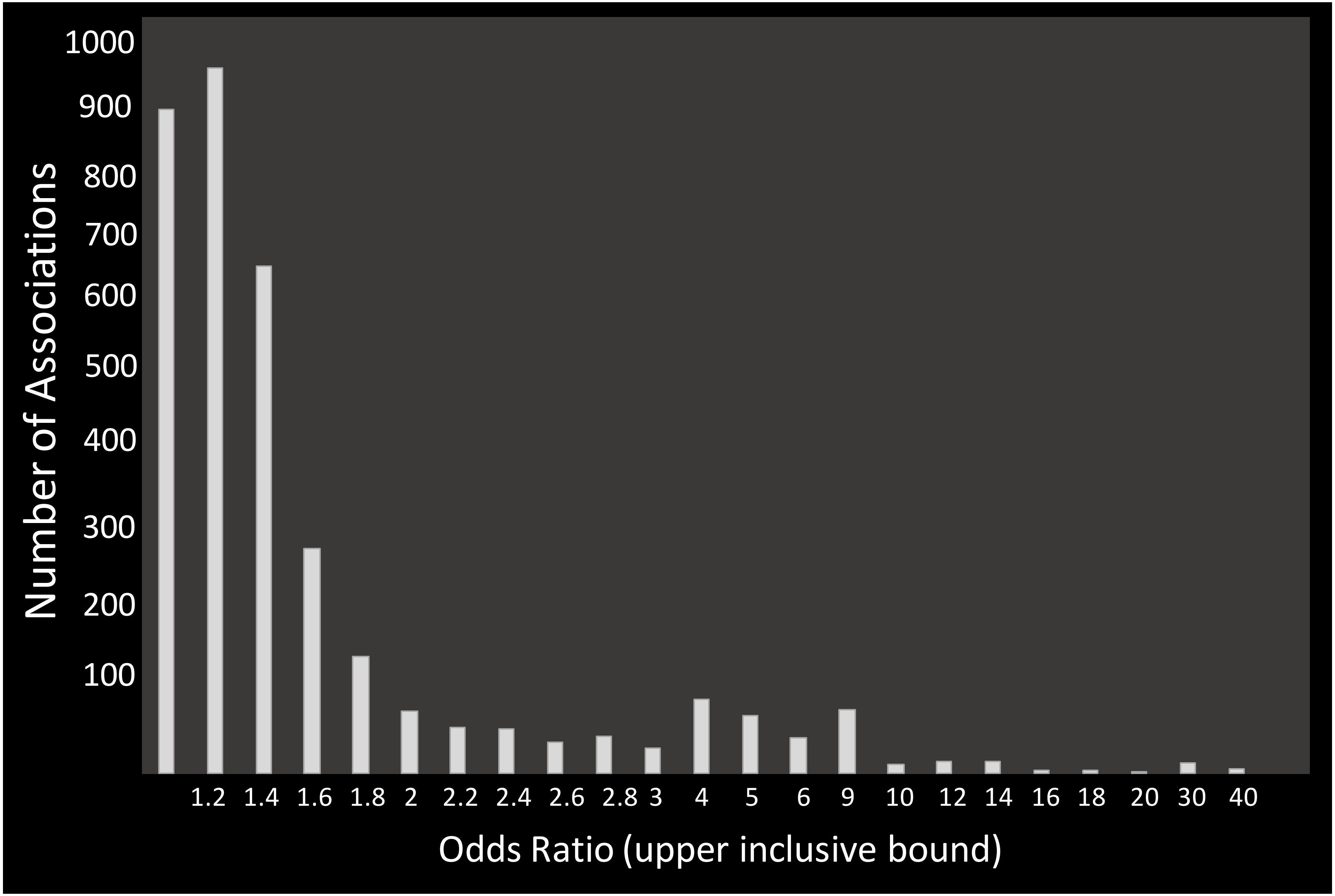
Published Genome-Wide Associations through 06/2011, 1,449 published GWA at $p \leq 5 \times 10^{-8}$ for 237 traits

2011 2nd quarter

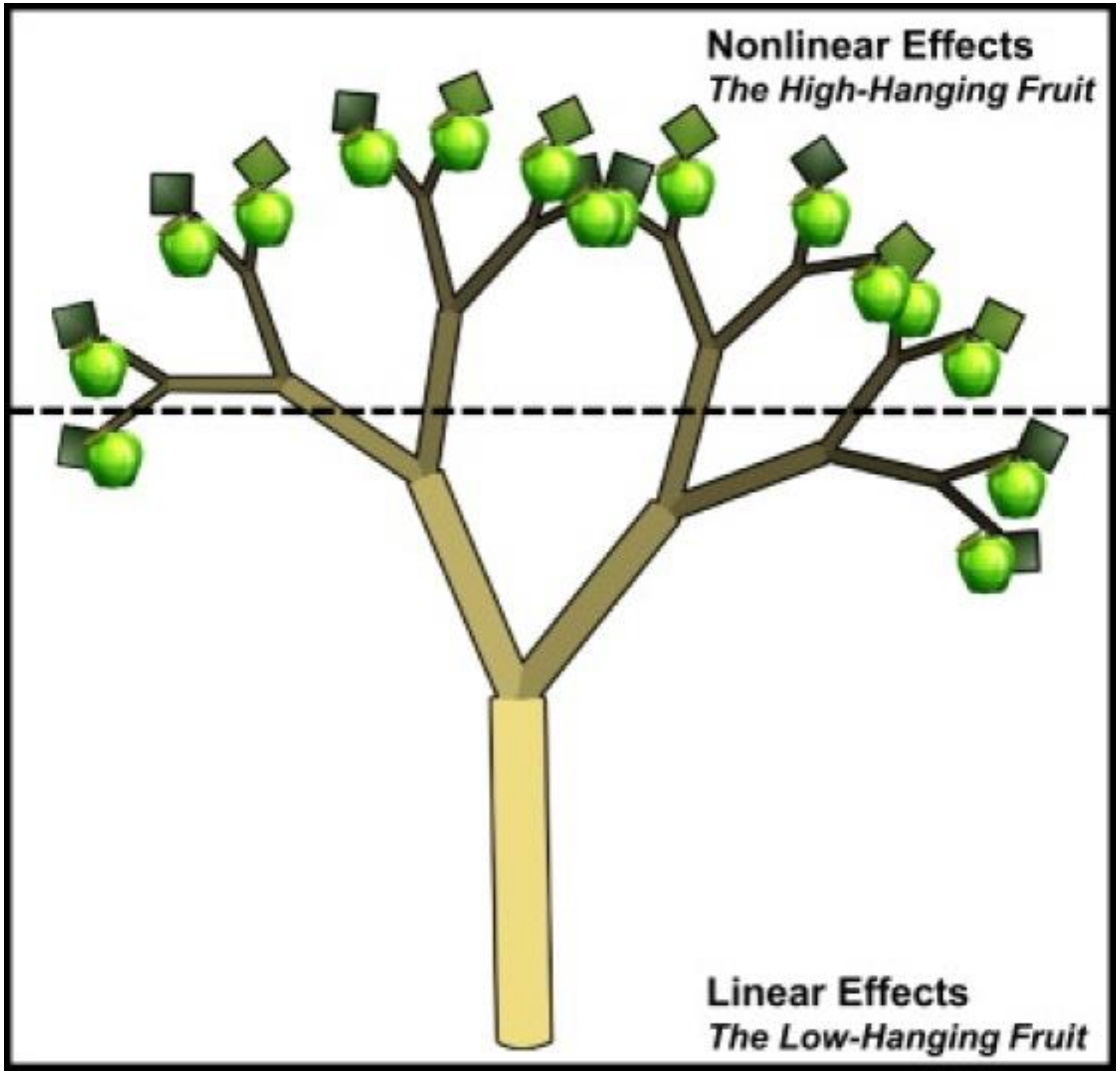


NHGRI GWA Catalog
www.genome.gov/GWAStudies

Distribution of Effects



Marylyn Ritchie, Jan 2014



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

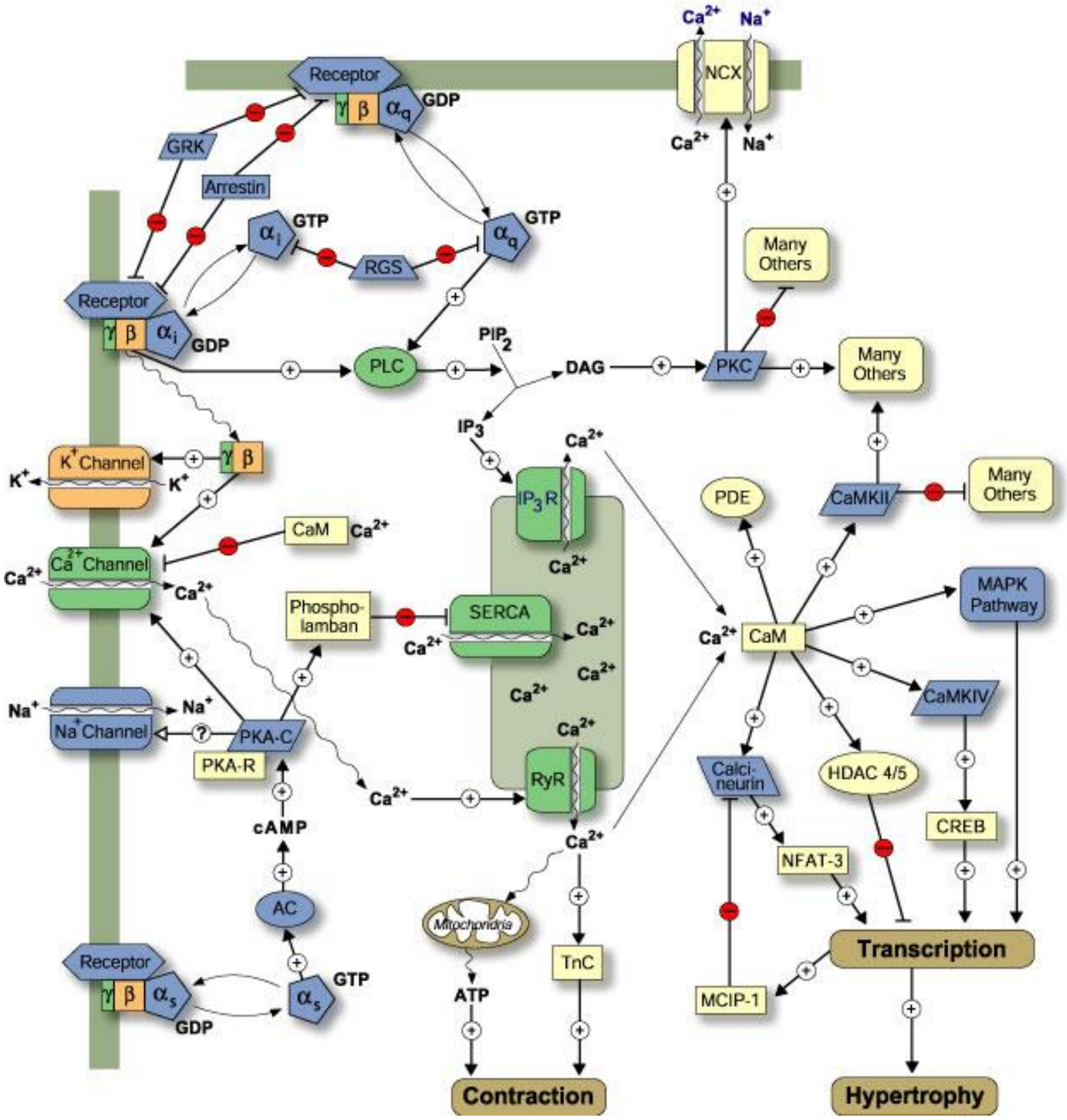
The case of the missing heritability

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.

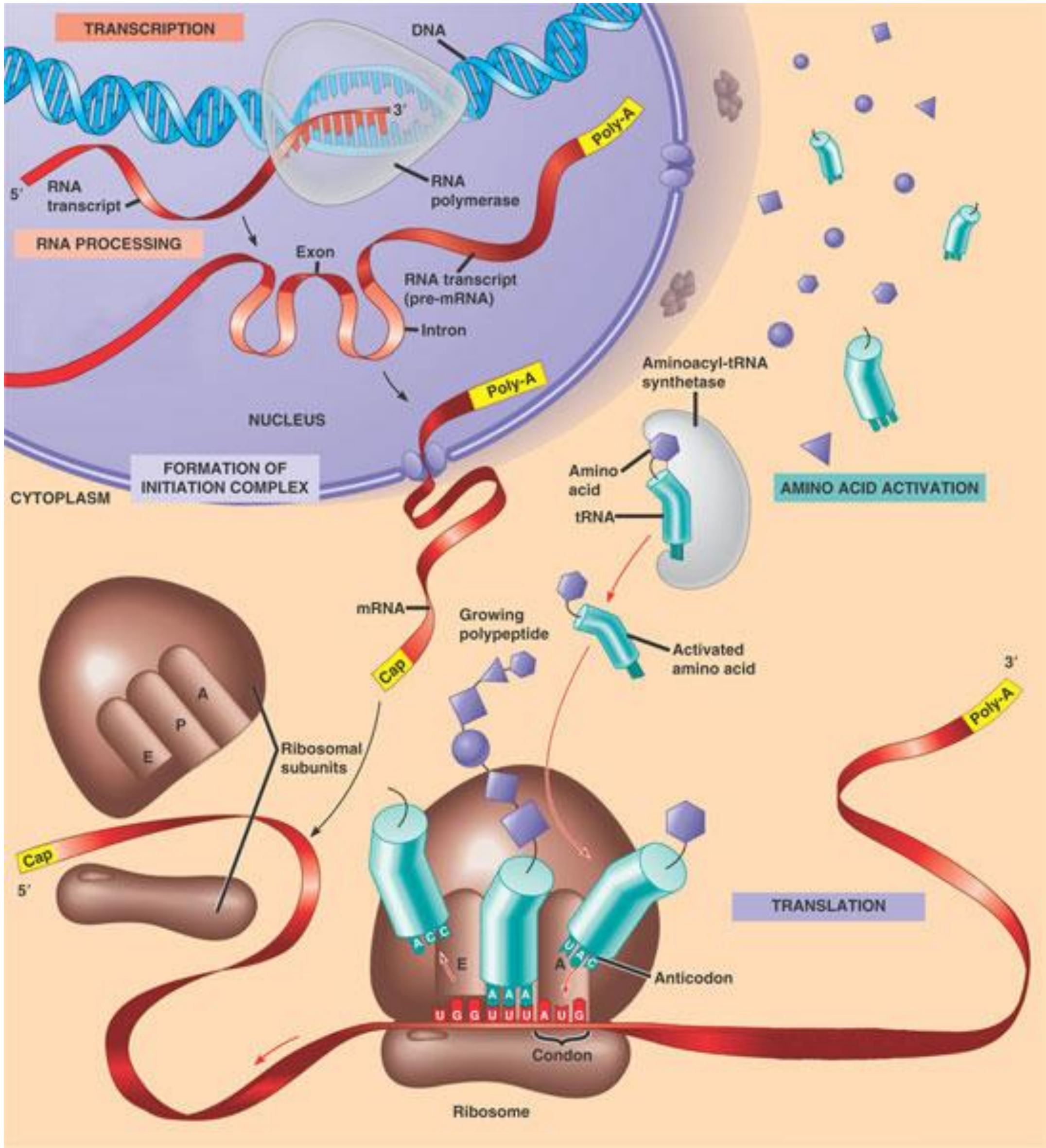
- Under our nose
- Out of sight
- In the architecture
- Underground networks
- Lost in diagnosis
- The great beyond

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

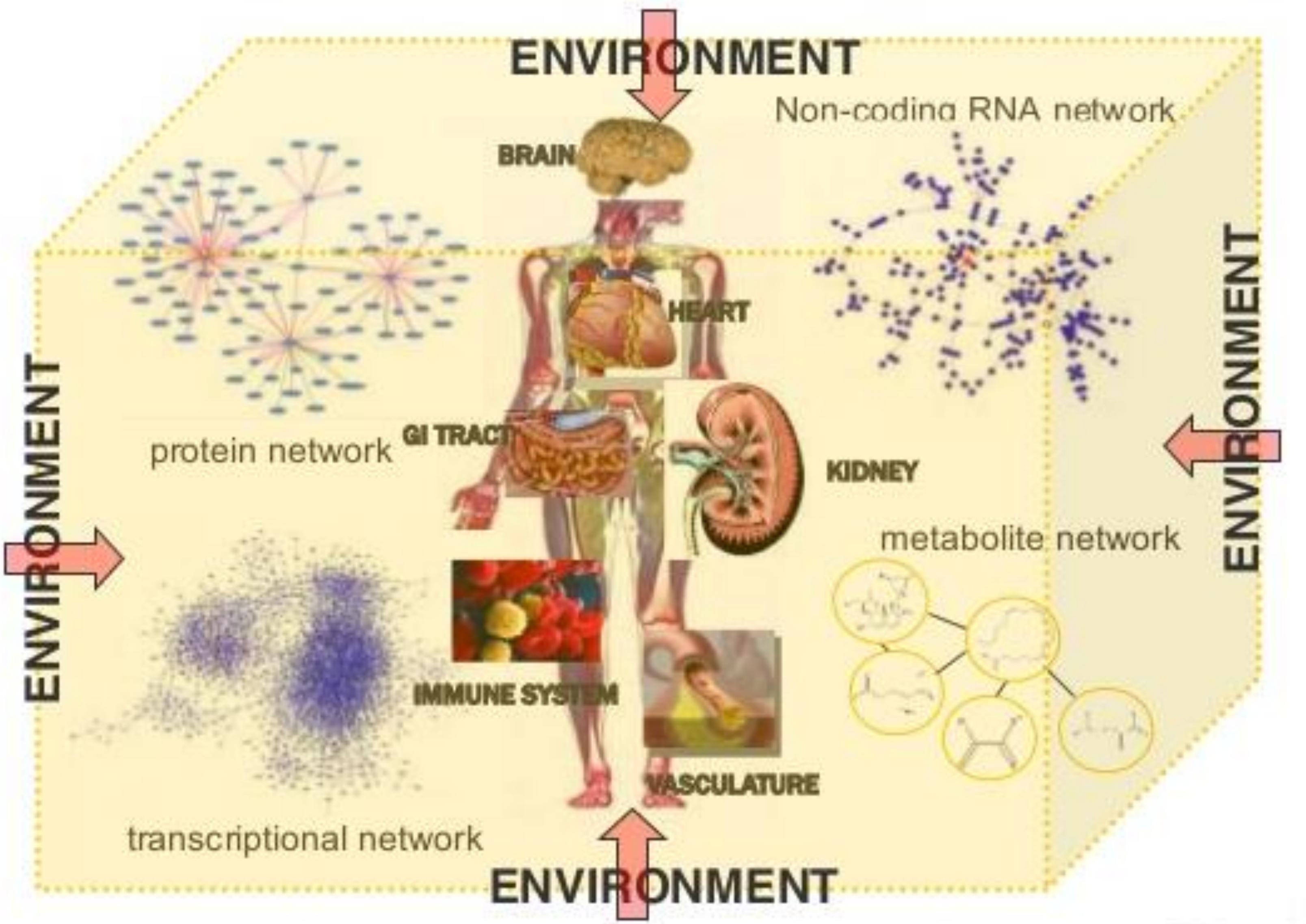
Biology is complex



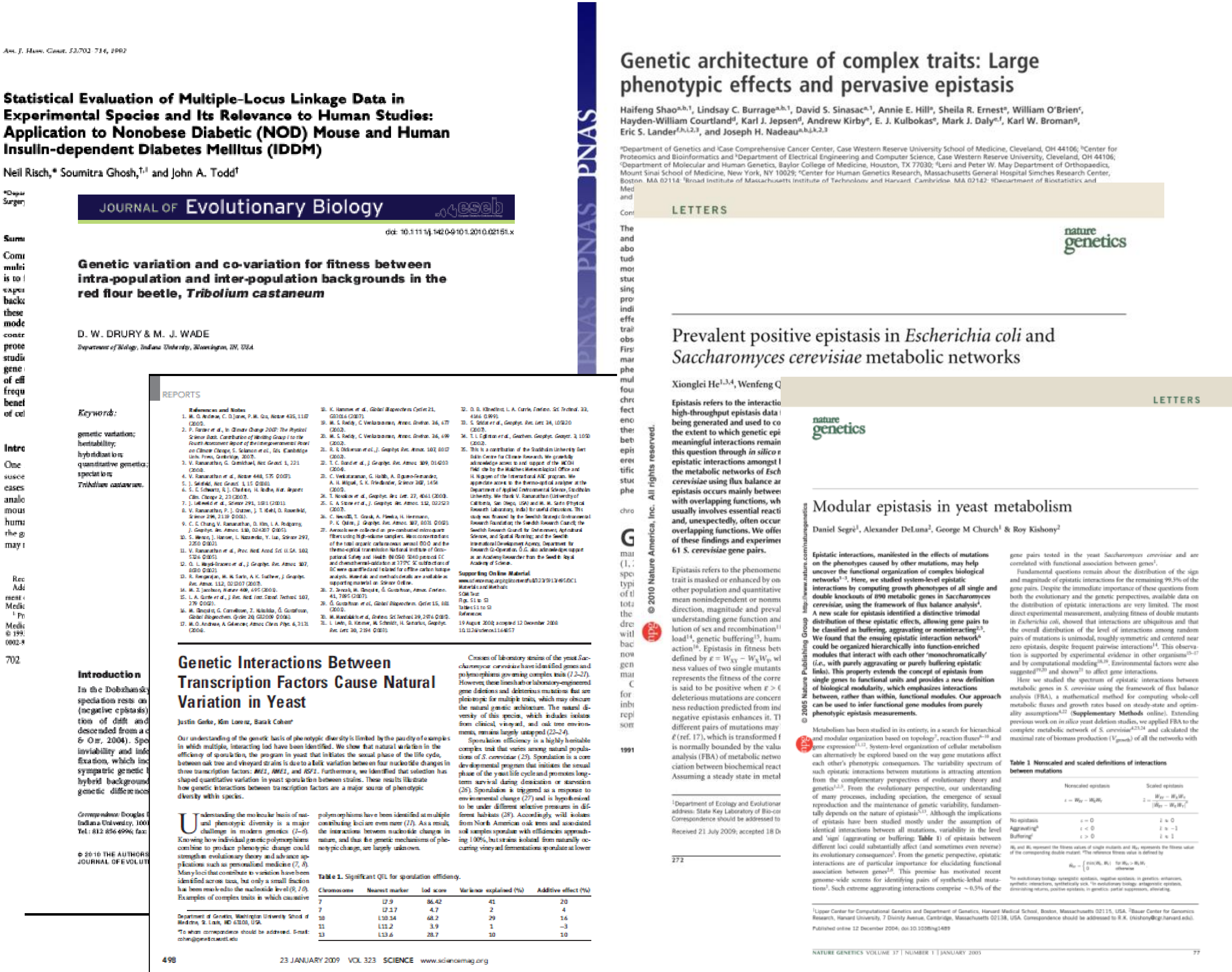
Molecular biology is complex



Biology is complex



Joel Dudley-SlideShare



Am. J. Hum. Genet. 52:702–714, 1992

Statistical Evaluation of Multiple-Locus Linkage Data in Experimental Species and Its Relevance to Human Studies: Application to Nonobese Diabetic (NOD) Mouse and Human Insulin-dependent Diabetes Mellitus (IDDM)

Neil Risch,* Soumitra Ghosh,†,‡ and John A. Todd†

JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2010.02151.x

Genetic variation and co-variation for fitness between intra-population and inter-population backgrounds in the red flour beetle, *Tribolium castaneum*

D. W. DRURY & M. J. WADE

Department of Biology, Indiana University, Bloomington, IN, USA

PNAS

Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis

Haifeng Shao[a,b,1], Lindsay C. Burrage[a,b,1], David S. Sinasac[a,1], Annie E. Hill[a], Sheila R. Ernest[a], William O'Brien[c], Hayden-William Courtland[d], Karl J. Jepsen[d], Andrew Kirby[e], E. J. Kulbokas[e], Mark J. Daly[e,f], Karl W. Broman[g], Eric S. Lander[f,h,i,2,3], and Joseph H. Nadeau[a,b,j,k,2,3]

REPORTS

Genetic Interactions Between Transcription Factors Cause Natural Variation in Yeast

Justin Gerke, Kim Lorenz, Barak Cohen*

Our understanding of the genetic basis of phenotypic diversity is limited by the paucity of examples in which multiple, interacting loci have been identified. We show that natural variation in the efficiency of sporulation, the program in yeast that initiates the sexual phase of the life cycle, between oak tree and vineyard strains is due to allelic variation between four nucleotide changes in three transcription factors: *IME1*, *RME1*, and *RSF1*. Furthermore, we identified that selection has shaped quantitative variation in yeast sporulation between strains. These results illustrate how genetic interactions between transcription factors are a major source of phenotypic diversity within species.

498 23 JANUARY 2009 VOL 323 SCIENCE www.sciencemag.org

LETTERS

nature genetics

Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks

Xionglei He[1,3,4], Wenfeng Q

© 2010 Nature America, Inc. All rights reserved.

LETTERS

nature genetics

Modular epistasis in yeast metabolism

Daniel Segrè[1], Alexander DeLuna[2], George M Church[1] & Roy Kishony[2]

Epistatic interactions, manifested in the effects of mutations on the phenotypes caused by other mutations, may help uncover the functional organization of complex biological networks[1–3]. Here, we studied system-level epistatic interactions by computing growth phenotypes of all single and double knockouts of 890 metabolic genes in *Saccharomyces cerevisiae*, using the framework of flux balance analysis[4]. A new scale for epistasis identified a distinctive trimodal distribution of these epistatic effects, allowing gene pairs to be classified as buffering, aggravating or noninteracting[2,5]. We found that the ensuing epistatic interaction network[6] could be organized hierarchically into function-enriched modules that interact with each other 'monochromatically' (*i.e.*, with purely aggravating or purely buffering epistatic links). This property extends the concept of epistasis from single genes to functional units and provides a new definition of biological modularity, which emphasizes interactions between, rather than within, functional modules. Our approach can be used to infer functional gene modules from purely phenotypic epistasis measurements.

© 2005 Nature Publishing Group http://www.nature.com/naturegenetics

NATURE GENETICS VOLUME 37 | NUMBER 1 | JANUARY 2005

ARTHRITIS & RHEUMATISM
Vol. 62, No. 3, March 2010, pp 705–710
DOI 10.1002/art.27292
© 2010, American College of Rheumatology

Evidence of Epistasis Between *TNFRSF14* and *TNFRSF6B* Polymorphisms in Patients With Rheumatoid Arthritis

Obje
genes codi
factor recep
TNFRSF6B,
toid arthriti
susceptibili
been relate
tion, and m
tiation. Int
bind a com
pressed in
investigate
rs6684865 a
RA predisp
Meth
phisms was
ethnically
validate the
with 211 pa
ally studied
Resu

Supp
by Fondo de
PI070369. Ms I
from the Mini
work was sup
respectively. D
Dr. Urcelay i
Biomédica–He
¹New
PhD, Alfonso
Benjamín Fern
pital Clínico S
Pascual-Salced
Spain.
Addr
digenes, MSc,
C/Profesor Ma
hotmail.com.
Subm
form Novemb

Genes and Immunity (2010) 11, 351–356
© 2010 Macmillan Publishers Limited All rights reserved 1466-4879/10 $32.00
www.nature.com/gene

SHORT COMMUNICATION

Evidence of interaction of CARD8 rs2043211 with NALP3 rs35829419 in Crohn's disease

RL Roberts¹˒², RKG Topless¹,
¹Department of Biochemistry, Universi
Christchurch, New Zealand and ²Depa

*The location of CARD8 within an i
as a nuclear factor (NF)κB inhibit
association of the CARD8 loss-of-f
results. A recent study provided ev
function variants in nucleotide-bindi
NALP3. To confirm this interaction,
controls). We found that the prese
protective effect against Crohn's dis
odds ratio (OR)=0.66, 95% confide
that these genotype combinations p
excessive interleukin 1β.*
Genes and Immunity (2010) 11, 351

Keywords: *inflammasome; TUCA*

Introduction

The *CARD8* gene (also known
up-regulated CARD-containing a
is located within the inflamm
(IBD) 6 linkage region on chromo
expressed in both monocytes an
CARD8 interacts with the NACHT
(also known as cryopyrin, NLF
Apoptosis-associated Speck-like
CARD to form a caspase 1 activ
the NALP3 inflammasome, which
tion and secretion of interleukin
microbial challenge (Figure 1).¹˒² I
the inflammasome, CARD8 is als
nuclear factor (NF)κB, and it ha
'cross-talk' occurs between CARD
tide-binding oligomerization dom
also known as CARD15)²˒³ As ov
and dysregulation of NFκB are
disease (CD),⁴ it is not surprisi
considered an attractive candida
immediately following the identi
view was re-enforced by a st
significant protective effect of t

Correspondence: Dr RL Roberts, Dep
University of Otago, Dunedin 9054, New
E-mail: rebecca.roberts@stonebow.otago.ac.nz
Received 11 November 2009; revised 11
January 2010; published online 25 Febru

Genes and Immunity (2010) 11, 199–208
© 2010 Macmillan Publishers Limited All rights reserved 1466-4879/10 $32.00
www.nature.com/gene

ORIGINAL ARTICLE

Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis

FBS Briggs¹, PP Ramsay¹, E Madden¹, JM Norris², VM Holers³, TR Mikuls⁴, T Sokka⁵, MF Seldin⁶, PK Gregersen⁷, LA Criswell⁸ and LF Barcellos¹
¹Division of Epidemiology, School of Public Health, University of California, Berkeley, CA, USA; ²Department of Epidemiology, Colorado School of Public Health, University of Colorado, Denver, Aurora, CO, USA; ³Integrated Department in Immunology, University of Colorado School of Medicine, Aurora, CO, USA; ⁴Department of Internal Medicine and Omaha VA Medical Center, University of Nebraska Medical Center, Omaha, NE, USA; ⁵Department of Medicine, Jyväskylä Central Hospital, Jyväskylä, Finland; ⁶Rowe Program in Molecular Medicine and Human Genetics, University of California, Davis, CA, USA; ⁷Feinstein Institute for Medical Research, North Shore Long Island Jewish Health System, Manhasset, NY, USA and ⁸Department of Medicine, Rosalind Russell Medical Research Center for Arthritis, University of California, San Francisco, CA, USA

Investigating genetic interactions (epistasis) has proven difficult despite the recent advances of both laboratory methods and statistical developments. With no 'best' statistical approach available, combining several analytical methods may be optimal for detecting epistatic interactions. Using a multi-stage analysis that incorporated supervised machine learning and methods of association testing, we investigated epistatic interactions with a well-established genetic factor (PTPN22 1858T) in a complex autoimmune disease (rheumatoid arthritis (RA)). Our analysis consisted of four principal stages: Stage I (data reduction)—identifying candidate chromosomal regions in 292 affected sibling pairs, by predicting PTPN22 concordance using multipoint identity-by-descent probabilities and a supervised machine learning algorithm (Random Forests); Stage II (extension analysis)—testing detailed genetic data within candidate chromosomal regions for epistasis with PTPN22 1858T in 677 cases and 750 controls using logistic regression; Stage III (replication analysis)—confirmation of epistatic interactions in 947 cases and 1756 controls; Stage IV (combined analysis)—a pooled analysis including all 1624 RA cases and 2506 control subjects for final estimates of effect size. A total of seven replicating epistatic interactions were identified. SNP variants within CDH13, MYO3A, CEP72 and near WFDC1 showed significant evidence for interaction with PTPN22, affecting susceptibility to RA.
Genes and Immunity (2010) 11, 199–208; doi:10.1038/gene.2009.110; published online 21 January 2010

Keywords: *epistasis; rheumatoid arthritis; PTPN22; Random Forests*

Introduction

Genome-wide association studies, which provide the ability to simultaneously investigate hundreds of thousands of genetic markers in large numbers of individuals, have successfully led to the discovery of genetic risk factors with modest effects in several complex diseases, including autoimmune diseases.¹˒² Nevertheless, it is apparent that current approaches to genetic analysis, which include almost exclusively, marginal associations using a univariate approach, are not able to identify a substantial fraction of the genetic burden. This may reflect the involvement of rare variants, copy number variation, *gene × gene* interactions, *gene × environment* interactions and/or epigenetic mechanisms. Currently, there is no consensus regarding appropriate approaches for evaluating these components of complex diseases.

Investigating genetic or *gene × gene* interactions (also known as 'epistasis', where the action of one gene is modified by one or several other genes) has proven difficult, despite recent advances of both laboratory methods and statistical developments. For example, the 15th biennial Genetic Analysis Workshop (GAW15) investigated genetic interactions in rheumatoid arthritis (RA (MIM 180300)) using several data sets. A variety of statistical approaches were used to investigate epistasis in RA in both family and population-based data sets with varying genetic marker density; results varied greatly, showing that robust and comprehensive approaches not restricted by a small sample size or sparse data are necessary.³⁻⁸ With no 'best' statistical approach available, combining several analytical methods may be optimal for detecting epistatic interactions.⁹˒¹⁰ Here, we performed a comprehensive multi-stage genetic investigation with a replication analysis to reveal epistatic

Correspondence: Dr LF Barcellos, Division of Epidemiology, School of Public Health, University of California, 209 Hildebrand Hall, Berkeley, CA 94720, USA.
E-mail: barcellos@genepi.berkeley.edu
Received 12 October 2009; accepted 15 October 2009; published online 21 January 2010

Addiction Biology

GENETIC STUDY doi:10.1111/j.1369-1600.2009.00183.x

Interaction of *SLC6A4* and *DRD2* polymorphisms is associated with a history of delirium tremens

Victor M. Karpyak¹, Joanna M. Biernacka¹˒², Mark W. Vander Weg³˒⁴, Susanna R. Stevens², Julie M. Cunningham⁵, David A. Mrazek¹ & John L. Black¹˒⁵
Department of Psychiatry and Psychology, Mayo Clinic College of Medicine, Rochester, MN, USA¹, Department of Biostatistics, Mayo Clinic College of Medicine, Rochester, MN, USA², Center for Research in the Implementation of Innovative Strategies in Practice (CRIISP), VA Medical Center, Iowa City, IA, USA³, Department

Behav Genet (2010) 40:357–365
DOI 10.1007/s10519-009-9314-8

ORIGINAL RESEARCH

Gene–Gene Interaction Between COMT and MAOA Potentially Predicts the Intelligence of Attention-Deficit Hyperactivity Disorder Boys in China

Qiu-Jin Qi
Hao-Bo Zh
Ning Ji · L

Received: 4 M
© Springer S

Abstract
gene conta
affecting th
oxidase A (
(MAOA-uV
COMT an
degradation
cortical (PF
individual
modulated
studies he
(ADHD) a
tently show
transmitted
present stu
between C
affect the i
ADHD sub
MAOA inte
and perform
MAOA-uV
verbal IQ
Val158Met

Edited by Ste

Qiu-Jin Qian

Q.-J. Qian · L
L.-L. Guan ·
Institute of M
Ministry of H
e-mail: wang

S. V. Farone
Departments
SUNY Upst

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 5 2010, pages 694–695
doi:10.1093/bioinformatics/btq009

Genetics and population analysis Advance Access publication January 16, 2010

Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS

Casey S. Greene¹˒†, Nicholas A. Sinnott-Armstrong¹˒†, Daniel S. Himmelstein¹, Paul J. Park², Jason H. Moore¹˒* and Brent T. Harris²
¹Department of Genetics and ²Department of Pathology, Dartmouth Medical School, Lebanon, NH 03756, USA
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Epistasis, the presence of gene-gene interactions, has been hypothesized to be at the root of many common human diseases, but current genome-wide association studies largely ignore its role. Multifactor dimensionality reduction (MDR) is a powerful model-free method for detecting epistatic relationships between genes, but computational costs have made its application to genome-wide data difficult. Graphics processing units (GPUs), the hardware responsible for rendering computer games, are powerful parallel processors. Using GPUs to run MDR on a genome-wide dataset allows for statistically rigorous testing of epistasis.

Results: The implementation of MDR for GPUs (MDRGPU) includes core features of the widely used Java software package, MDR. This GPU implementation allows for large-scale analysis of epistasis at a dramatically lower cost than the standard CPU-based implementations. As a proof-of-concept, we applied this software to a genome-wide study of sporadic amyotrophic lateral sclerosis (ALS). We discovered a statistically significant two-SNP classifier and subsequently replicated the significance of these two SNPs in an independent study of ALS. MDRGPU makes the large-scale analysis of epistasis tractable and opens the door to statistically rigorous testing of interactions in genome-wide datasets.

Availability: MDRGPU is open source and available free of charge from http://www.sourceforge.net/projects/mdr.

Contact: jason.h.moore@dartmouth.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 16, 2009; revised on January 7, 2010; accepted on January 8, 2010

1 INTRODUCTION

Genome-wide association studies hold promise for the discovery of the genetic factors that underlie common human diseases (Hirschhorn and Daly, 2005; Wang *et al.*, 2005). Unfortunately this promise has largely not been realized (Shriner *et al.*, 2007; Williams *et al.*, 2007). It is thought that this failure could be due to epistasis, the role of gene–gene interactions, which has commonly been ignored in these studies. Powerful and model-free methods such as multifactor dimensionality reduction (MDR) have been developed (Ritchie *et al.*, 2001), but an exhaustive examination of even pair-wise interactions in a 550 000 SNP dataset would require the analysis of 1.5×10^{11} combinations. While an analysis of this scale is approachable with modern cluster computing, an analysis that includes permutation testing to assess the statistical significance of results remains infeasible with CPU-based approaches.

Rendering photo-realistic video games in real time is also computationally difficult. For video game graphics, specific hardware (the graphics processing unit or GPU) has been developed. The GPU is a massively parallel computing platform that can be adapted to some scientific tasks. We have previously shown that MDR is one of these tasks (Sinnott-Armstrong *et al.*, 2009). Here we provide software which makes practical the analysis of epistasis in genome-wide data through the use of GPUs and demonstrate its application to a genome-wide analysis of epistasis of sporadic amyotrophic lateral sclerosis (ALS).

2 METHODS

MDRGPU, a software tool capable of analyzing genome-wide data, is a Python implementation of MDR, which uses the PyCUDA library to run MDR on GPUs. MDRGPU 1.0 supports balanced accuracy, large datasets, execution across an arbitrary number of GPUs, permutation testing and the analysis of high-order interactions. It runs on GPUs which support CUDA (i.e. the NVIDIA GeForce 8800 series and higher). Parallel execution of one realization across multiple GPUs is supported with the pp library for Python. MDRGPU provides a command-line interface for scripted analysis.

The GPU architecture has various memory spaces available. MDRGPU uses the constant cache, global memory, shared memory and registers. Shared memory is used to store the intermediate case and control counts for each attribute combination and to store the number of true and false positives and negatives. The global memory is accessed directly to fetch attributes. The constant cache is used in MDRGPU to store the case–control status. Dataset sizes of greater than 65 536 attributes require splitting which is handled seamlessly by MDRGPU. This splitting does not cause linear slowdown; there is simply more overhead of launching, so datasets with large numbers of instances see less of a performance reduction than datasets with few instances. The largest number of addressable attributes is 4 billion requiring 4 GB RAM per instance. In order for the case–control status to be held in constant memory, there can be at most 16 384 instances.

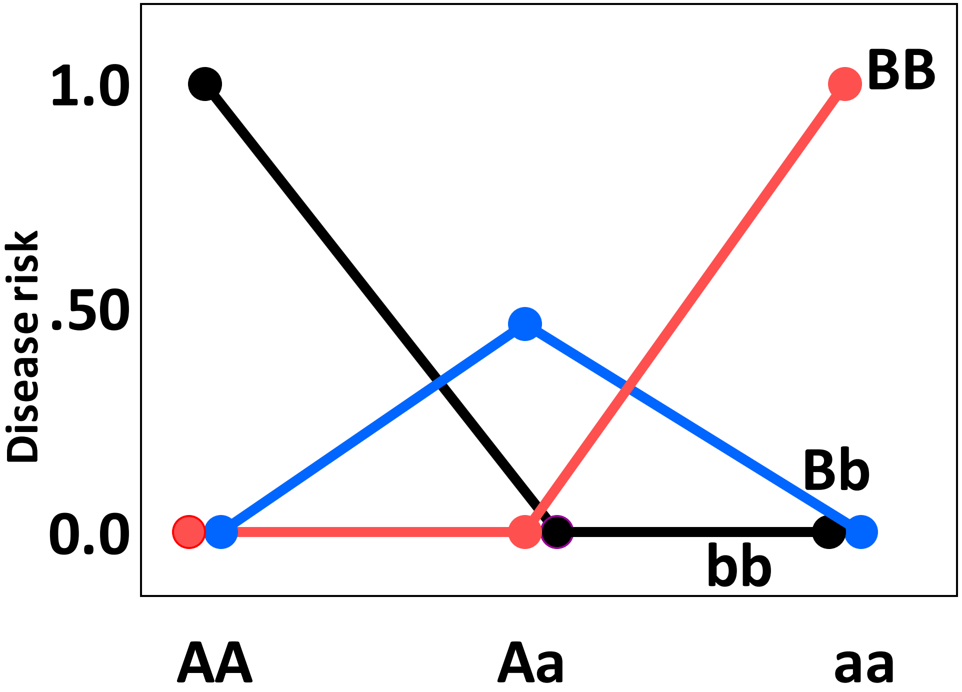
Our proof of concept analysis was performed on three GPU workstations (detailed in Supplementary Material S1). These systems contain three GeForce 295 cards, each of which contains two GPUs. For the first stage of this analysis, we used an ALS dataset from Schymick *et al.* (2007) as our detection dataset. This dataset was obtained from QUEUE at Coriell, but has since been moved to dbGaP. It contains 276 individuals with sporadic ALS and 271 control individuals. These individuals are genotyped at 555 352 SNPs using the Illumina Infinium II HumanHap550 SNP chip. We processed this dataset by removing SNPs with a minor allele frequency <0.2 or those

[*] To whom correspondence should be addressed.
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

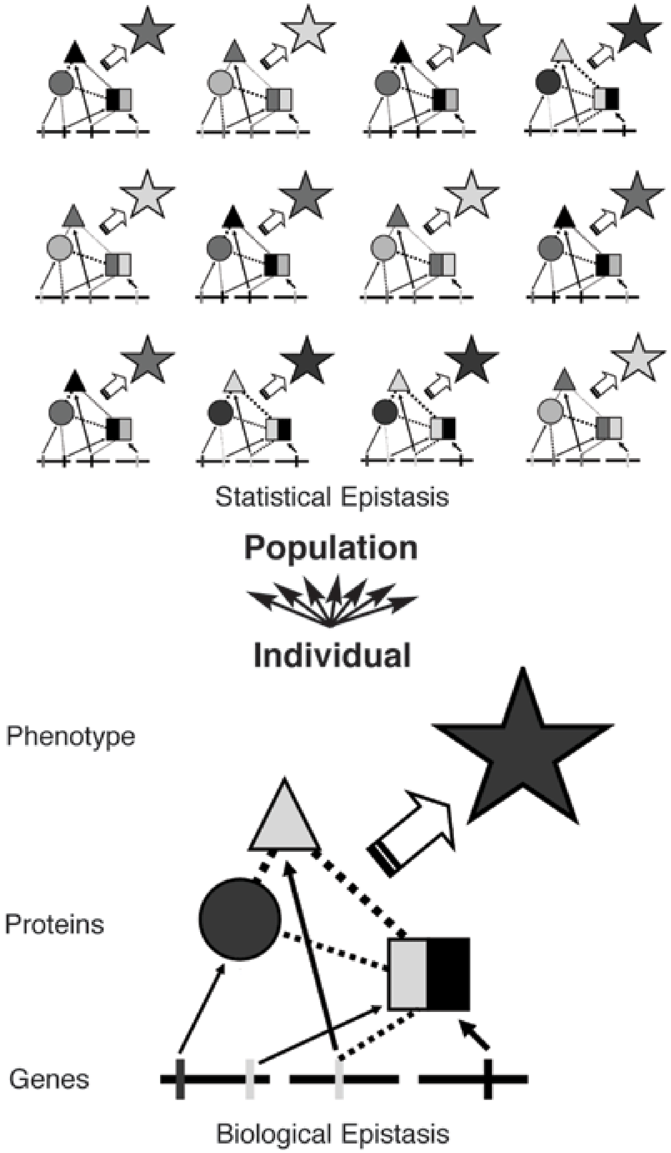
Epistasis

- 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

Epistasis is important because…

- Biologists believe bio-molecular interactions are very common
- Identifying "the gene" associated with common disease has not been as successful like it has for Mendelian disease
- Epistasis is detected when properly investigated
- Mendelian single-gene disorders are now being considered complex traits with gene-gene interactions (modifier genes)

- **Most people agree epistasis exists but the degree of independent main effect with epistasis versus interaction effects in the absence of statistically detectable main effects are a topic of controversy**

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

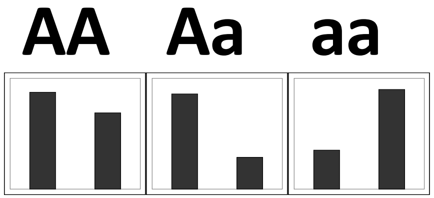
Genetic Epidemiology - Association Analysis

- 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

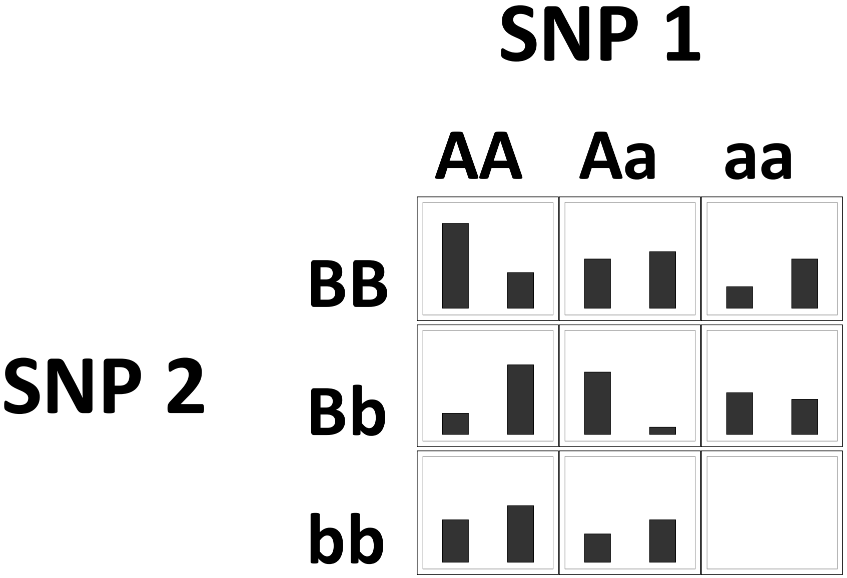
N = 100 50 Cases, 50 Controls

SNP 1



Curse of Dimensionality

N = 100 50 Cases, 50 Controls



SNP 1

AA Aa aa

SNP 2

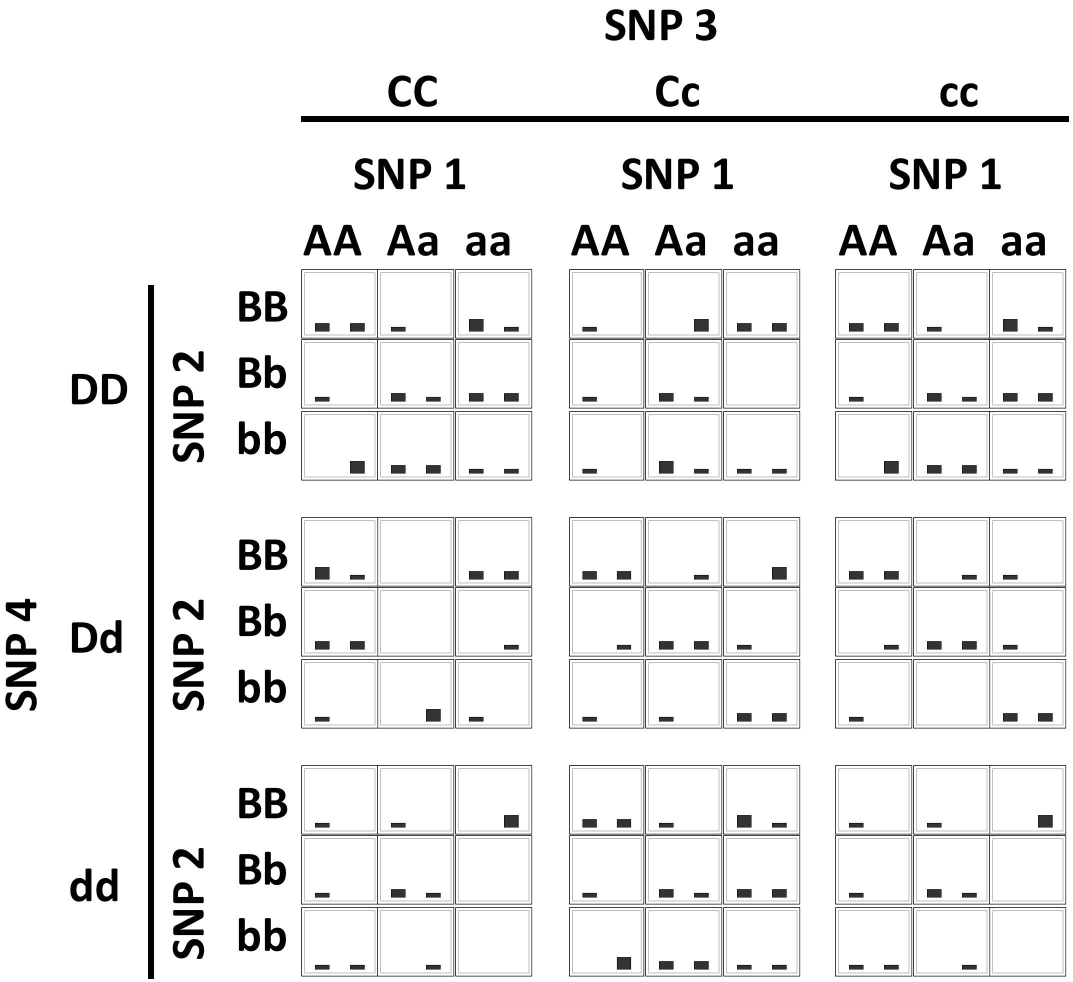
BB

Bb

bb

Curse of Dimensionality

N = 100 50 Cases, 50 Controls

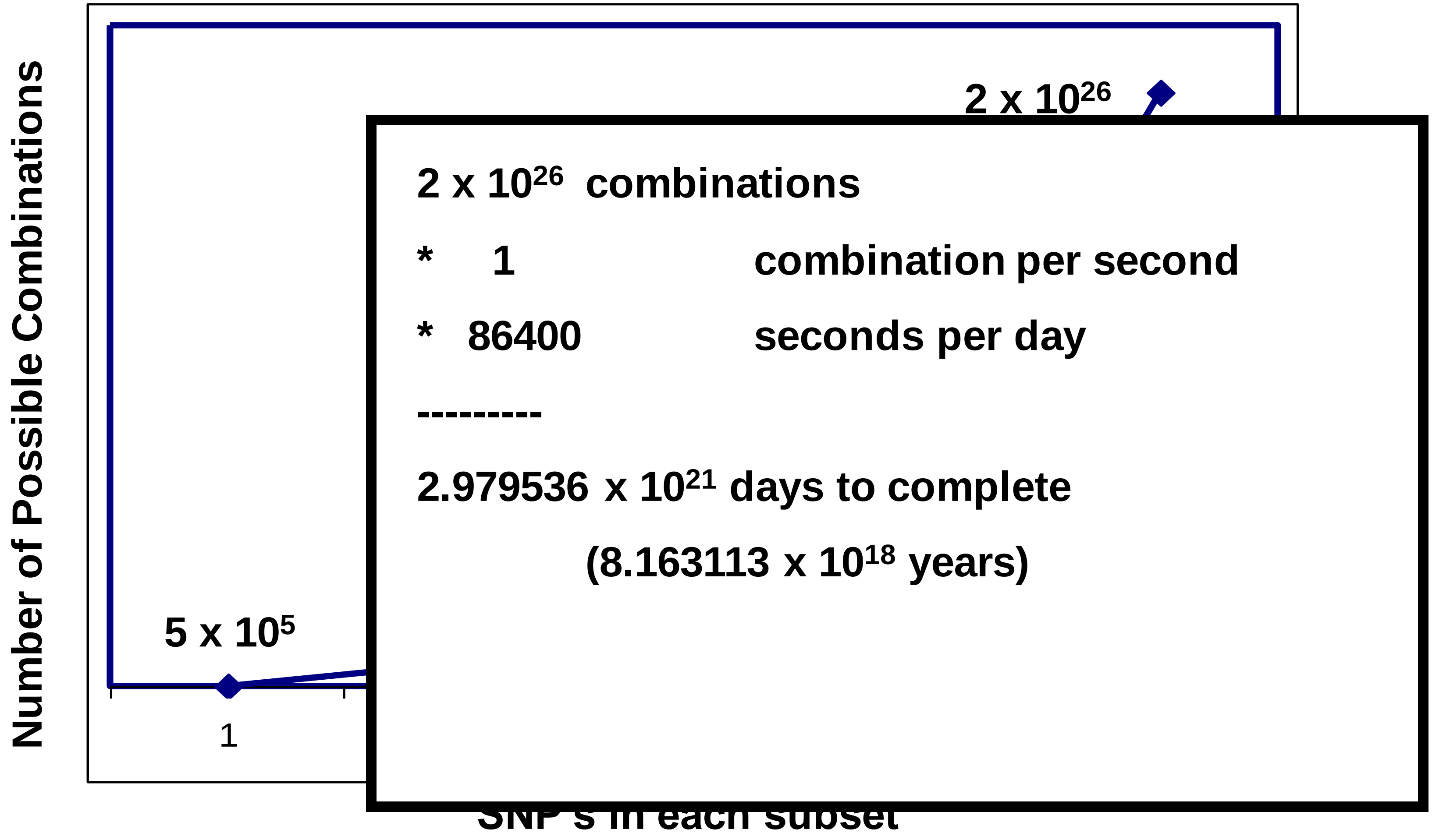


If interactions with minimal main effects are the norm rather than the exception, can we analyze all possible combinations of loci with traditional approaches to detect purely interaction effects ?

NO

How many combinations are there?

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



2 x 10^{26} combinations

*** 1 combination per second**

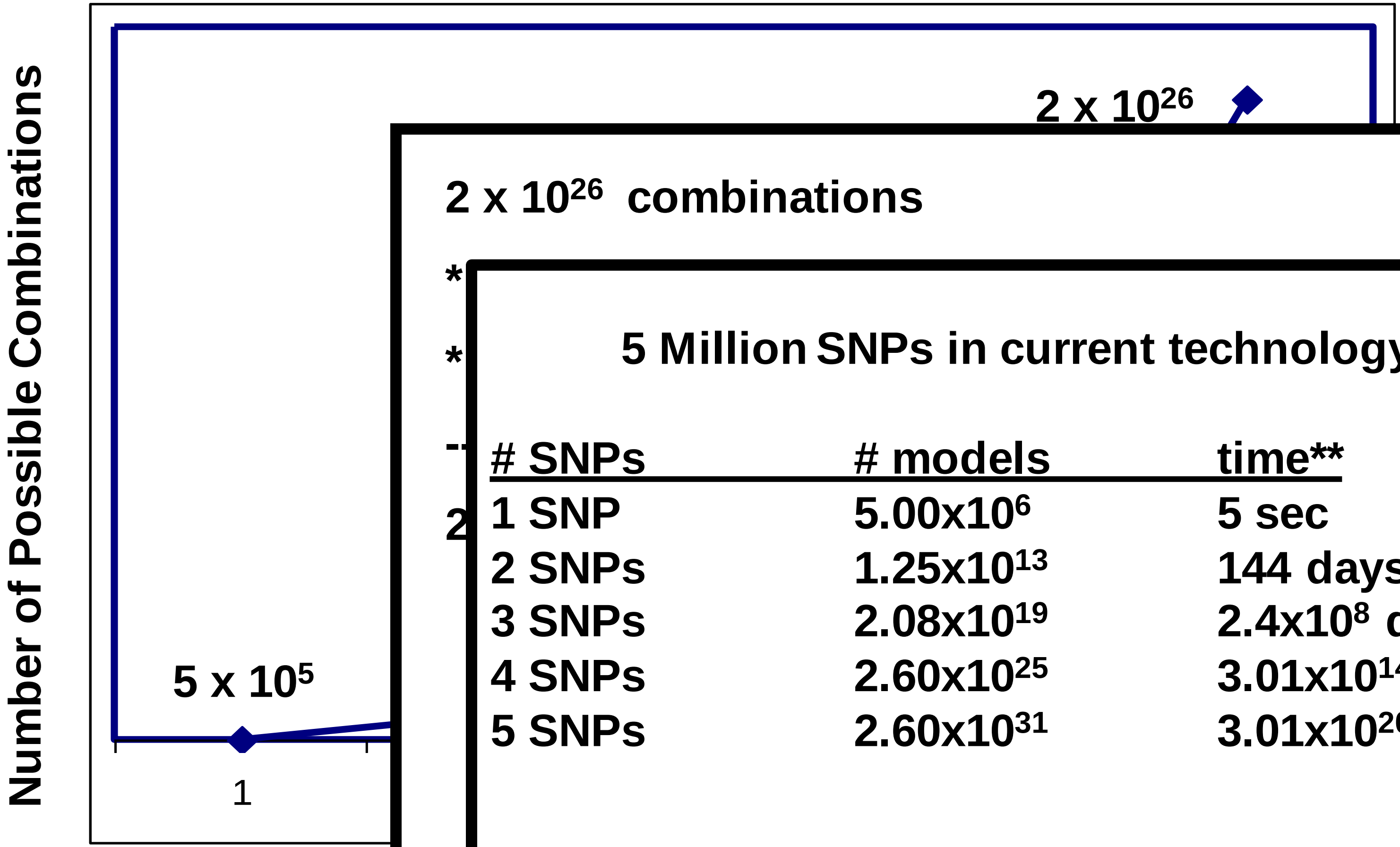
*** 86400 seconds per day**

2.979536 x 10^{21} days to complete

(8.163113 x 10^{18} years)

How many combinations are there?

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



2 x 10^26 combinations

*

*

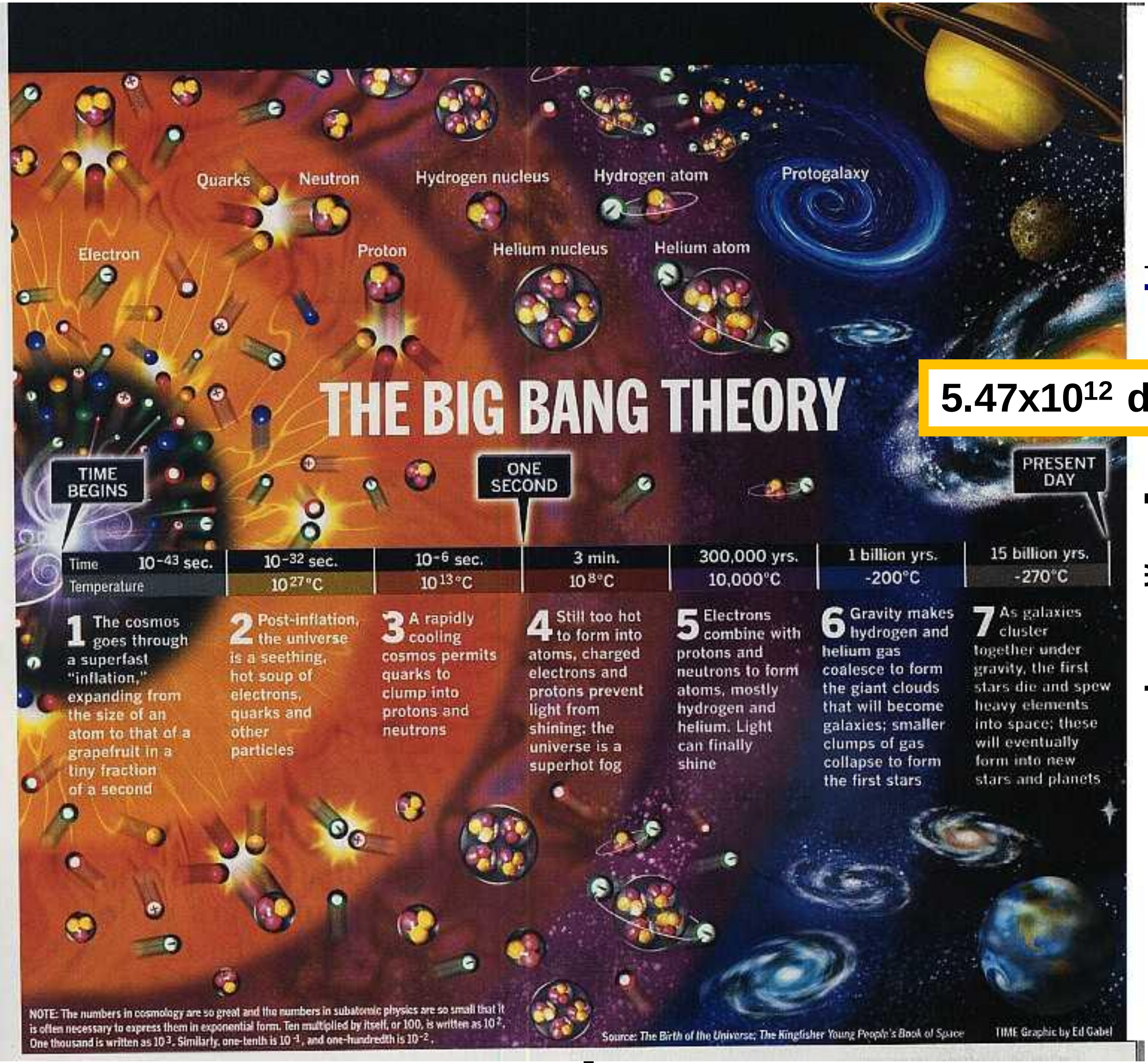
--

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



THE BIG BANG THEORY

Time	10^{-43} sec.	10^{-32} sec.	10^{-6} sec.	3 min.	300,000 yrs.	1 billion yrs.	15 billion yrs.
Temperature		10^{27}°C	10^{13}°C	10^{8}°C	10,000°C	-200°C	-270°C

1 The cosmos goes through a superfast "inflation," expanding from the size of an atom to that of a grapefruit in a tiny fraction of a second

2 Post-inflation, the universe is a seething, hot soup of electrons, quarks and other particles

3 A rapidly cooling cosmos permits quarks to clump into protons and neutrons

4 Still too hot to form into atoms, charged electrons and protons prevent light from shining; the universe is a superhot fog

5 Electrons combine with protons and neutrons to form atoms, mostly hydrogen and helium. Light can finally shine

6 Gravity makes hydrogen and helium gas coalesce to form the giant clouds that will become galaxies; smaller clumps of gas collapse to form the first stars

7 As galaxies cluster together under gravity, the first stars die and spew heavy elements into space; these will eventually form into new stars and planets

NOTE: The numbers in cosmology are so great and the numbers in subatomic physics are so small that it is often necessary to express them in exponential form. Ten multiplied by itself, or 100, is written as 10^2. One thousand is written as 10^3. Similarly, one-tenth is 10^{-1}, and one-hundredth is 10^{-2}.

Source: The Birth of the Universe; The Kingfisher Young People's Book of Space

TIME Graphic by Ed Gabel

10^{26}

5.47x10^{12} days

ent technology

time**

5 sec

144 days

2.4x10^{8} days

3.01x10^{14} days

3.01x10^{20} days

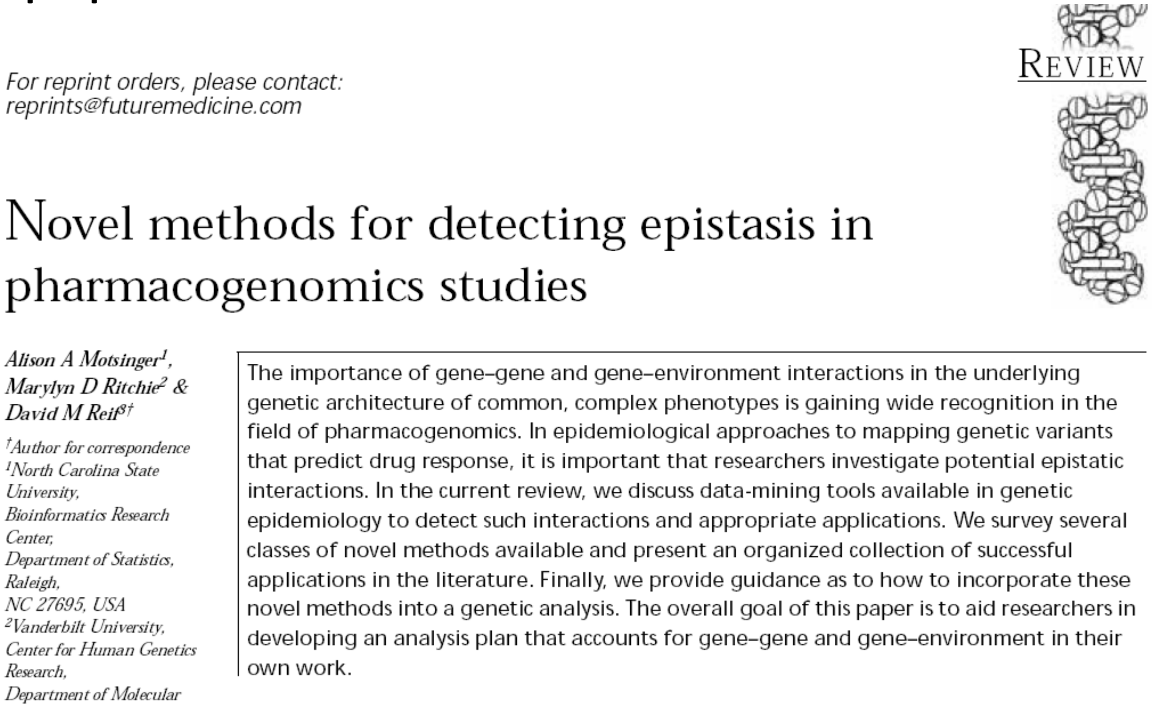
**assuming 1 CPU that performs 1 million tests per second

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*[3†]

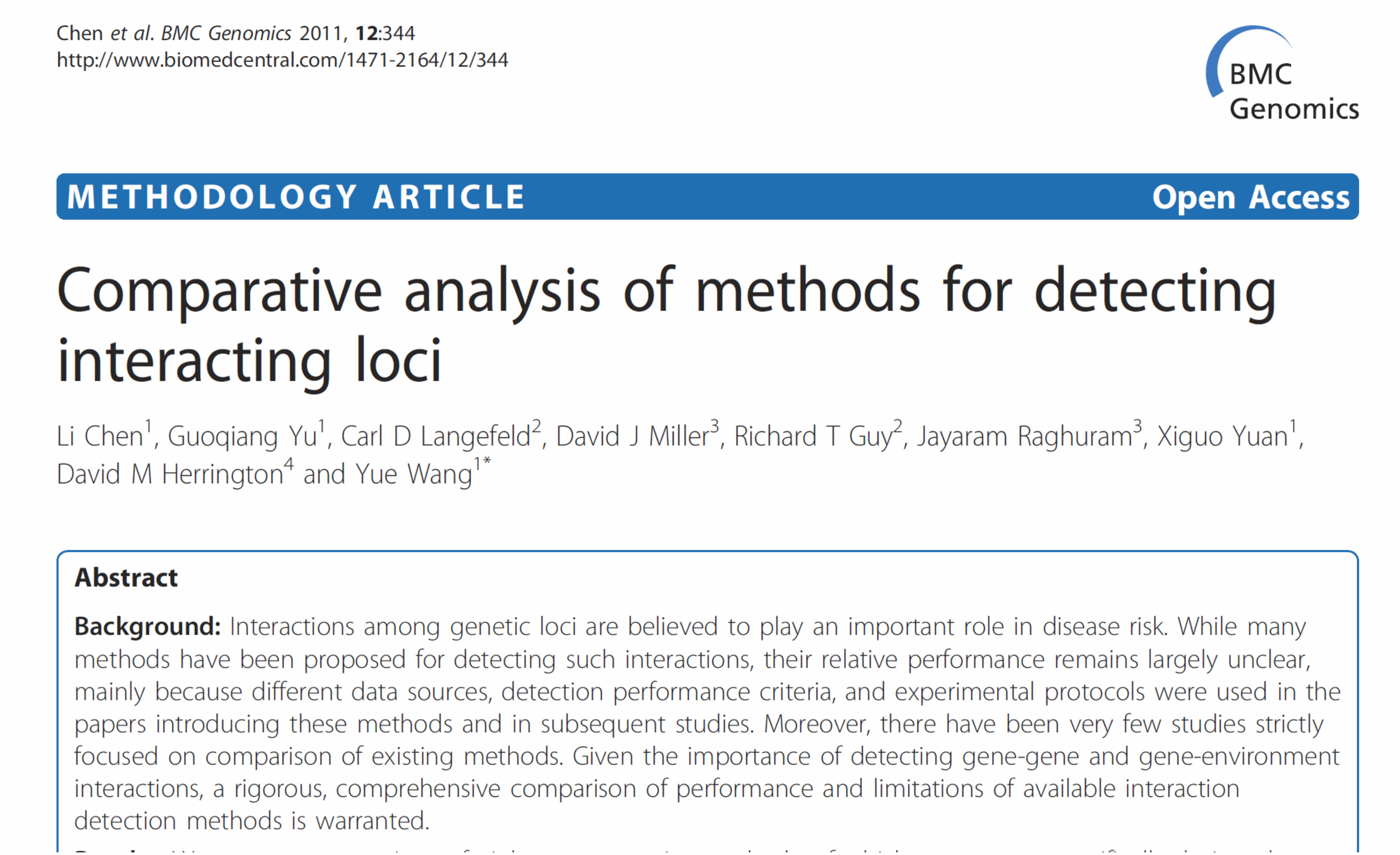
†Author for correspondence
[1]North Carolina State University, Bioinformatics Research Center, Department of Statistics, Raleigh, NC 27695, USA
[2]Vanderbilt University, Center for Human Genetics Research, Department of Molecular

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

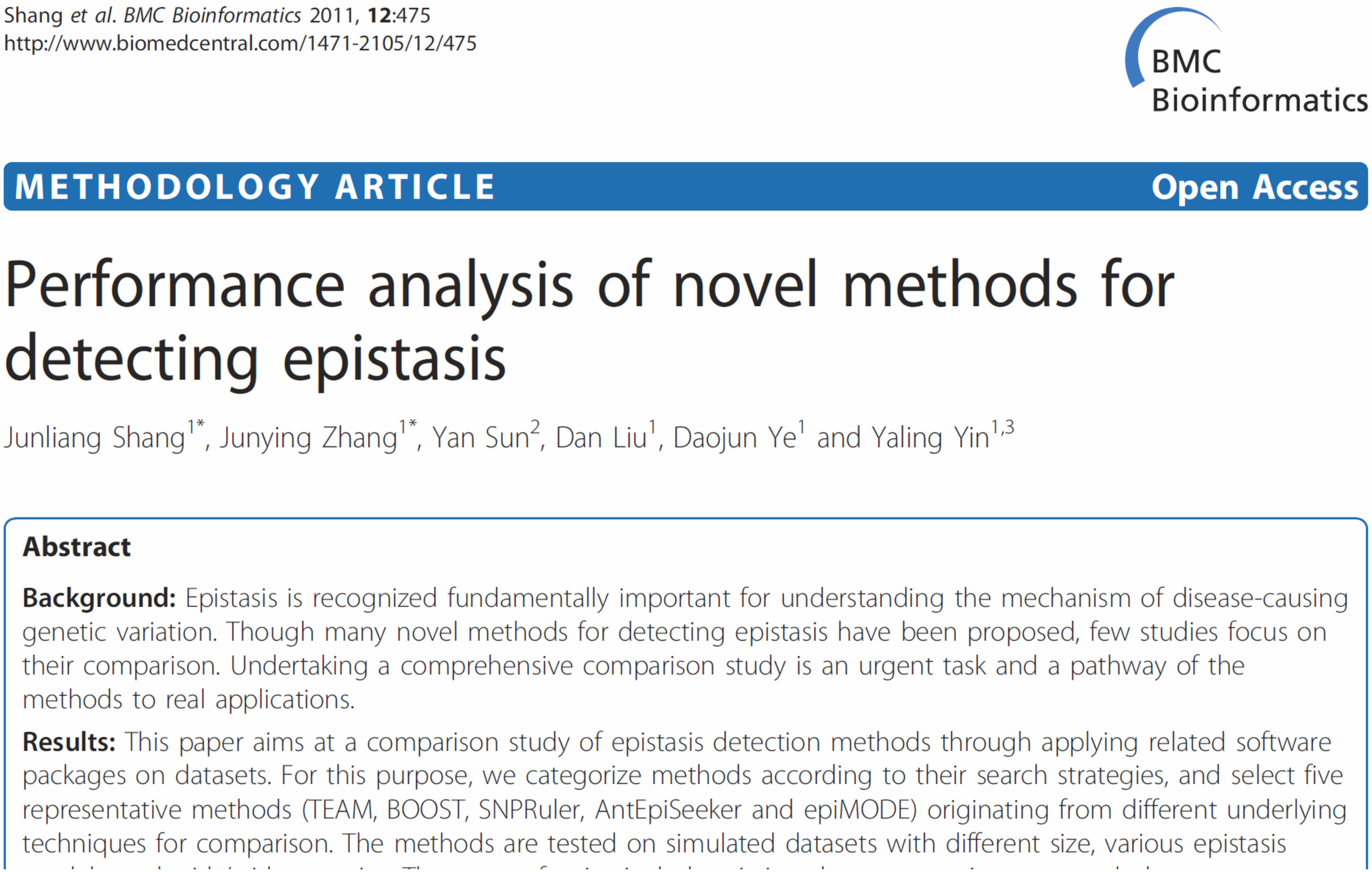
Comparative analysis of methods for detecting interacting loci

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

Complex Fitness Landscape

Fitness

Model

Epistasis in GWAS Data

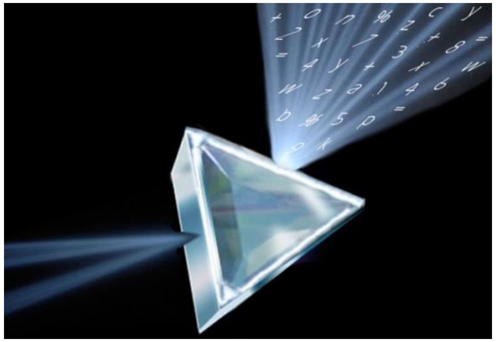
- ~~Exhaustive evaluation~~
- Evaluate interactions in top hits from single-SNP analysis
- Use prior biological knowledge to evaluate specific combinations – “Candidate Epistasis”

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

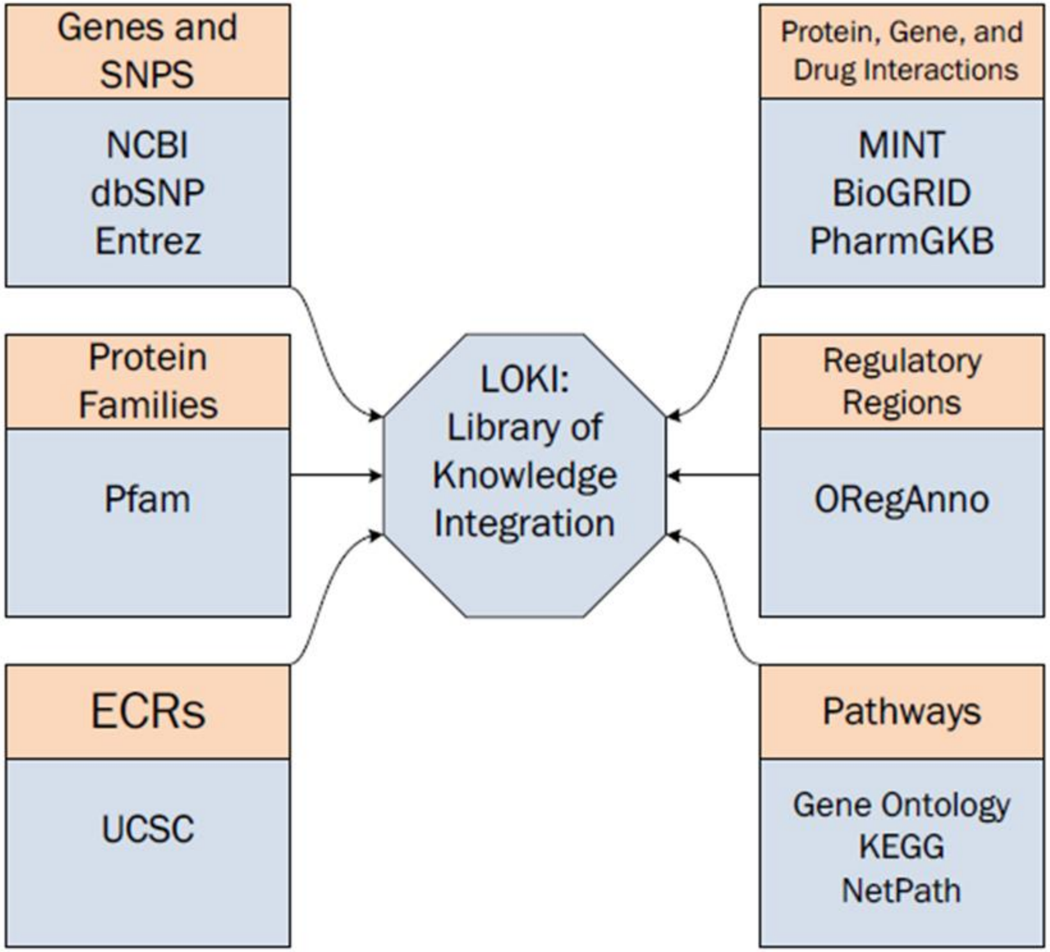
The Biofilter

- 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).

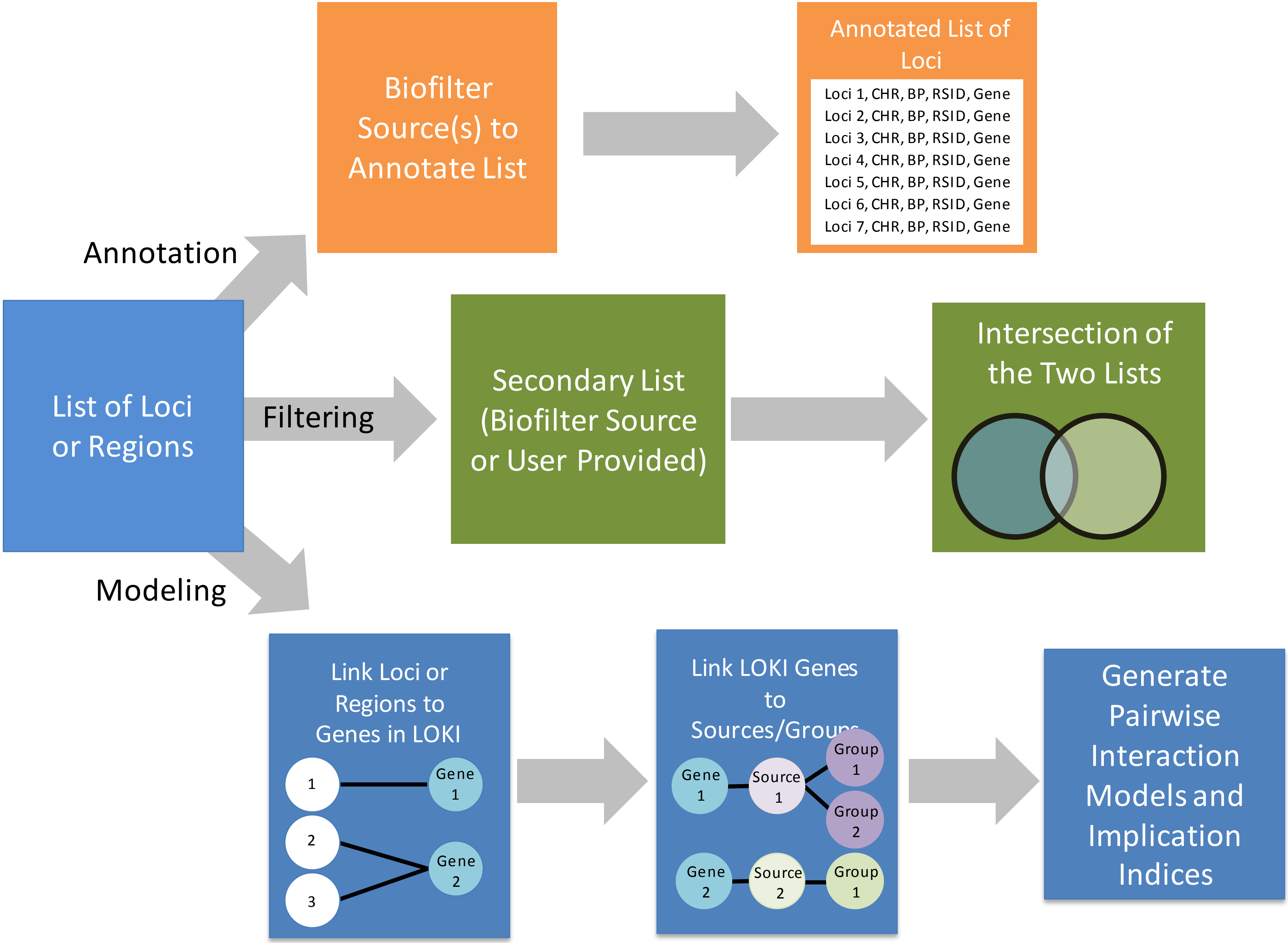
LOKI: Library of Knowledge Integration



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).

The Biofilter

- Method described: Bush et al. 2009 *Pacific Symposium on Biocomputing, Pendergrass et al, BioData Mining,2013*
  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



List of Loci or Regions

Annotation

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

Filtering

Secondary List (Biofilter Source or User Provided)

Intersection of the Two Lists

Modeling

Link Loci or Regions to Genes in LOKI

Link LOKI Genes to Sources/Groups

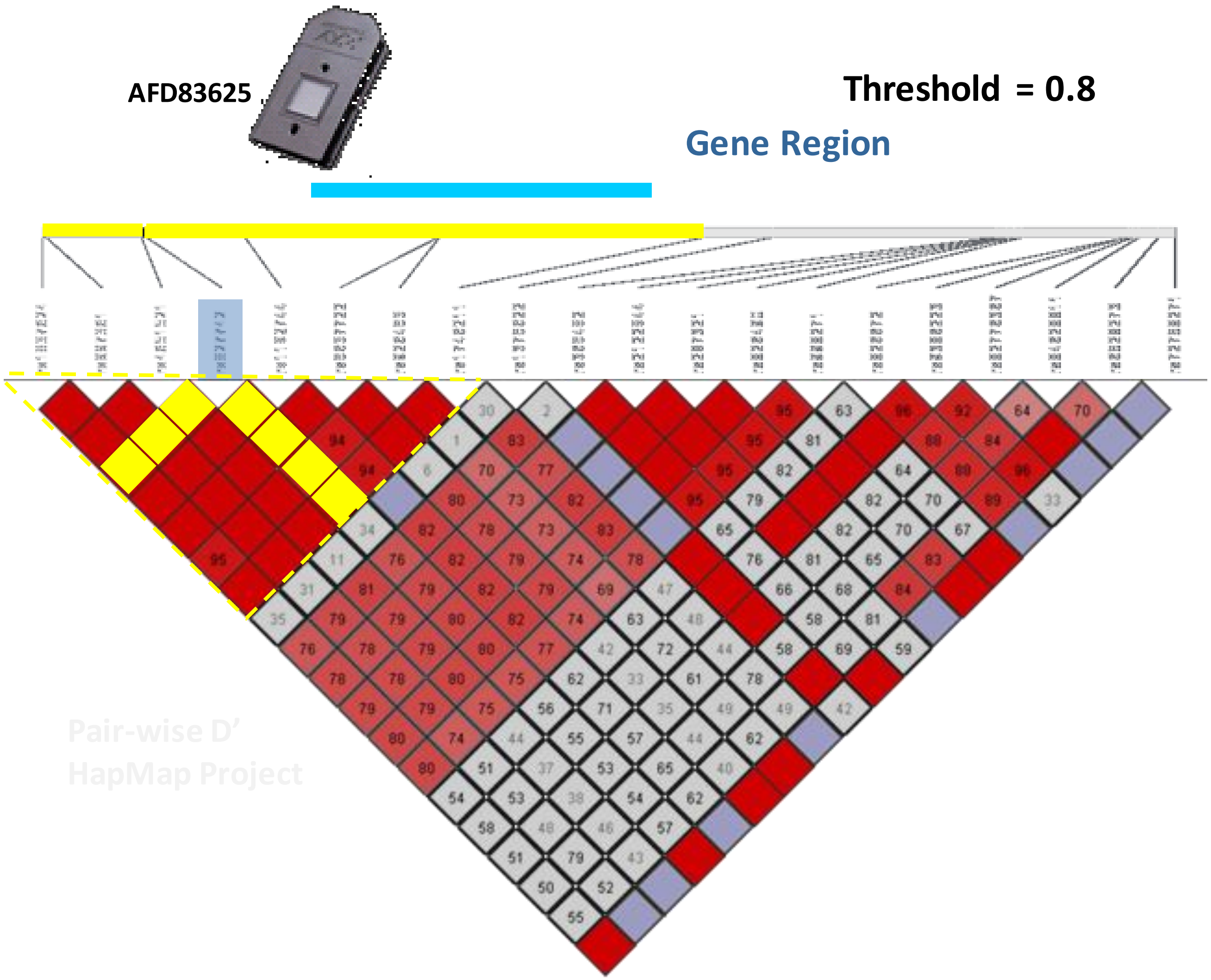
Generate Pairwise Interaction Models and Implication Indices

Candidate Epistasis Analysis of GWAS

Four Step Process

1. **Relate SNPs to Genes**
2. **Relate genes to one another**
3. **Generate multi-SNP models using this information**
4. **Evaluate the multi-SNP models using statistical technique**

Relate SNPs to Genes



LD-Spline: Mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium. Bush WS, Chen G, Torstenson ES, Ritchie MD. BioData Min. 2009 Dec 3;2(1):7

Using Biofilter: Prioritizing Analysis

Candidate Gene/Regions

- Previous Linkage Regions
- Differential Gene Expression
- Candidate Pathways
- Known biology
- …

Candidate Epistasis

- KEGG (Pathways)
- DIP (Protein-protein interactions)
- PFAM (Protein families)
- GO (Gene Ontology)
- Reactome (Pathways)
- Netpath (Signal transduction)
- …

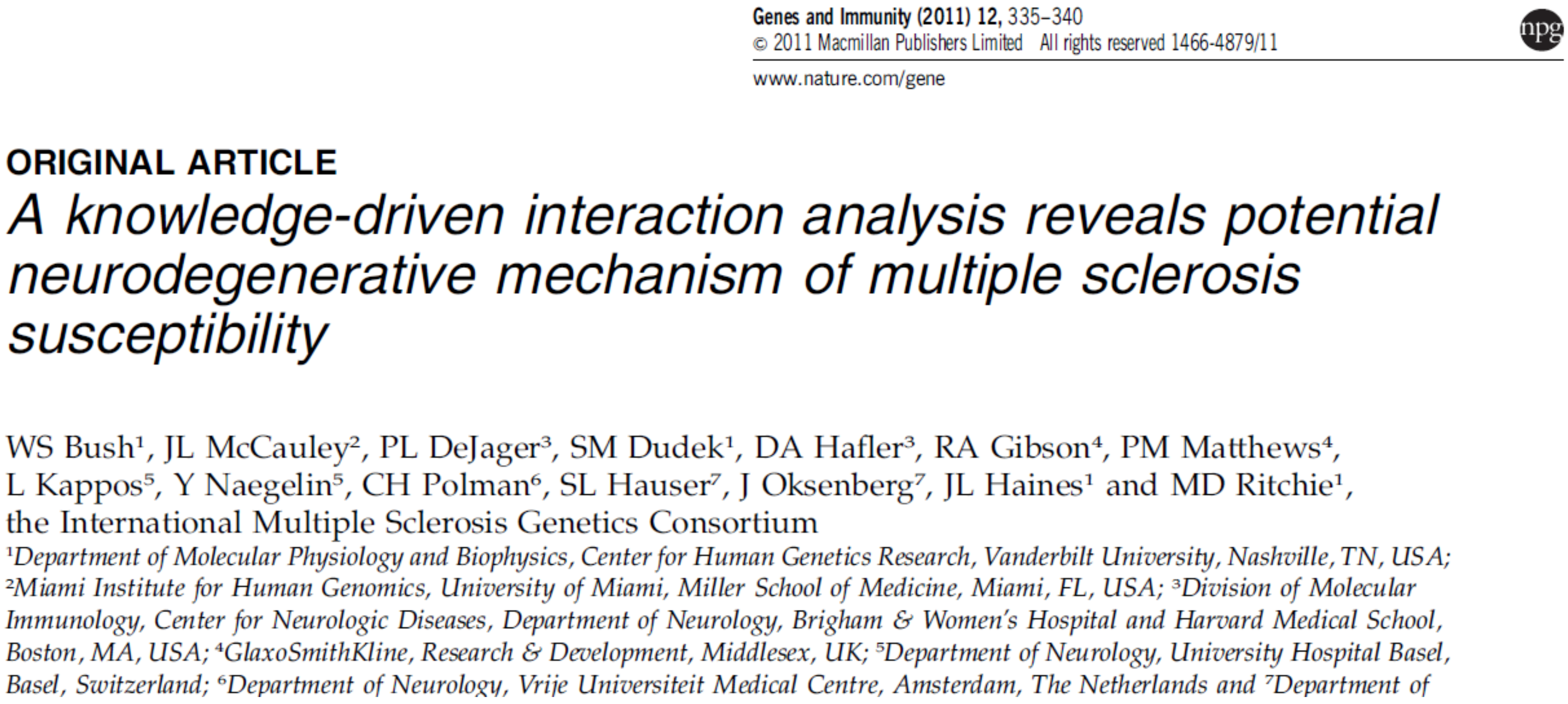
Candidate Approaches

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



Genes and Immunity (2011) 12, 335–340
© 2011 Macmillan Publishers Limited All rights reserved 1466-4879/11
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[3], SM Dudek[1], DA Hafler[3], RA Gibson[4], PM Matthews[4], L Kappos[5], Y Naegelin[5], CH Polman[6], SL Hauser[7], J Oksenberg[7], JL Haines[1] and MD Ritchie[1], the International Multiple Sclerosis Genetics Consortium

[1]*Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA;* [2]*Miami Institute for Human Genomics, University of Miami, Miller School of Medicine, Miami, FL, USA;* [3]*Division of Molecular Immunology, Center for Neurologic Diseases, Department of Neurology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, USA;* [4]*GlaxoSmithKline, Research & Development, Middlesex, UK;* [5]*Department of Neurology, University Hospital Basel, Basel, Switzerland;* [6]*Department of Neurology, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands and* [7]*Department of*

- **930 trio families from US and UK (IMSGC)**
- **Genotyped on Affymetrix 500K array**
  - **Post QC ~300,000 SNPs**

- **Reduction of search space from 53 billion models to 20 million models but this could be reduced further**

Full Model $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Reduced Model $\beta_1 x_1 + \beta_2 x_2$

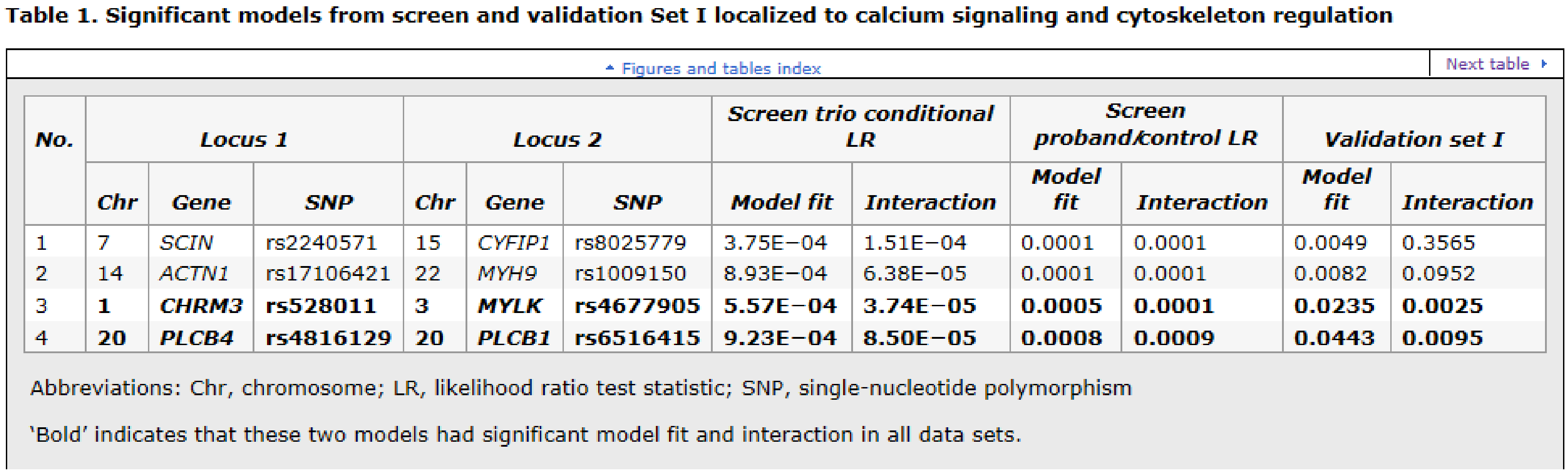

Figures and tables index
Next table

Table 1. Significant models from screen and validation Set I localized to calcium signaling and cytoskeleton regulation

No.	Locus 1			Locus 2			Screen trio conditional LR		Screen proband/control LR		Validation set I	
	Chr	Gene	SNP	Chr	Gene	SNP	Model fit	Interaction	Model fit	Interaction	Model fit	Interaction
1	7	*SCIN*	rs2240571	15	*CYFIP1*	rs8025779	3.75E−04	1.51E−04	0.0001	0.0001	0.0049	0.3565
2	14	*ACTN1*	rs17106421	22	*MYH9*	rs1009150	8.93E−04	6.38E−05	0.0001	0.0001	0.0082	0.0952
3	**1**	***CHRM3***	**rs528011**	**3**	***MYLK***	**rs4677905**	**5.57E−04**	**3.74E−05**	**0.0005**	**0.0001**	**0.0235**	**0.0025**
4	**20**	***PLCB4***	**rs4816129**	**20**	***PLCB1***	**rs6516415**	**9.23E−04**	**8.50E−05**	**0.0008**	**0.0009**	**0.0443**	**0.0095**

Abbreviations: Chr, chromosome; LR, likelihood ratio test statistic; SNP, single-nucleotide polymorphism

'Bold' indicates that these two models had significant model fit and interaction in all data sets.

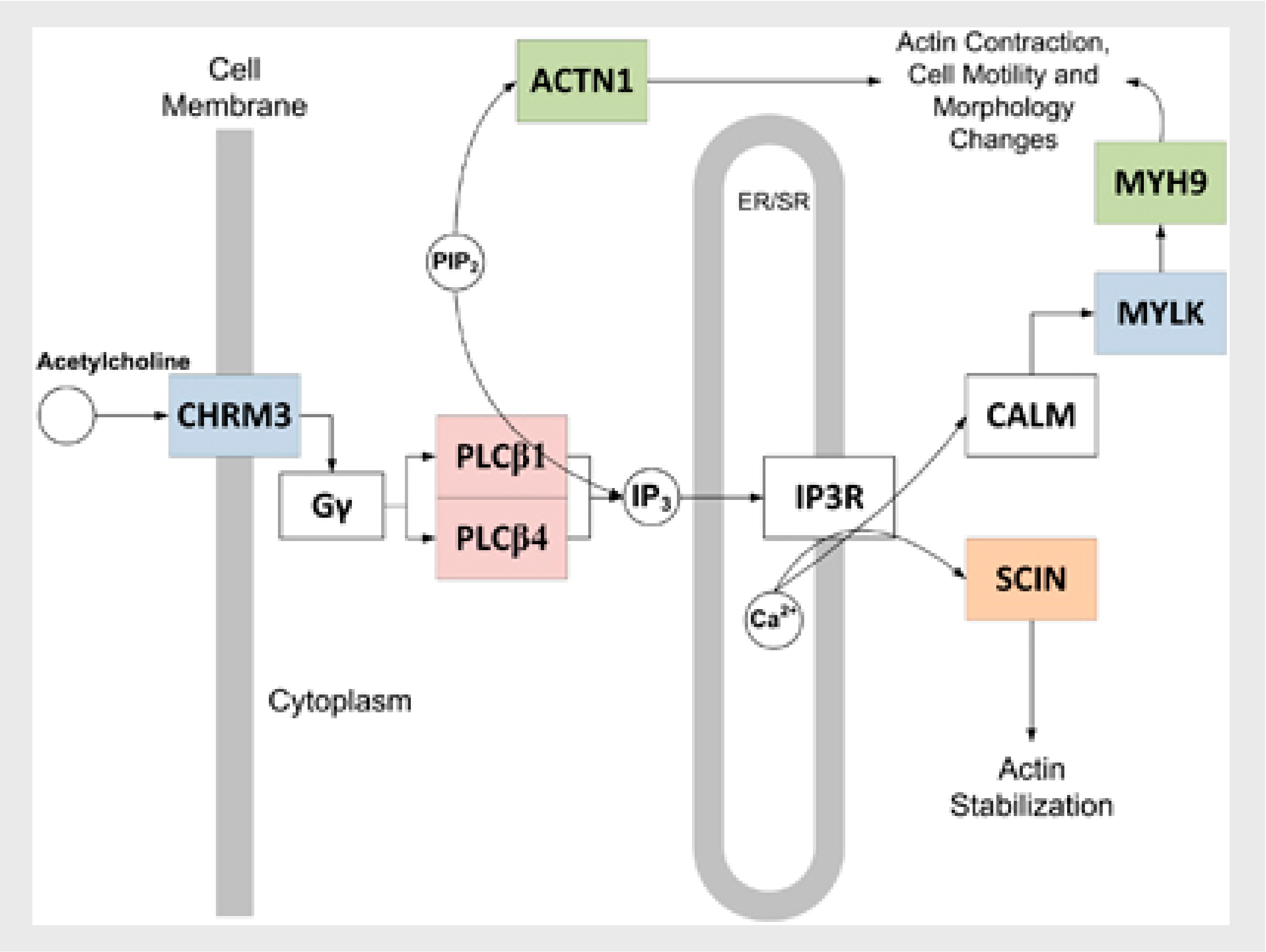
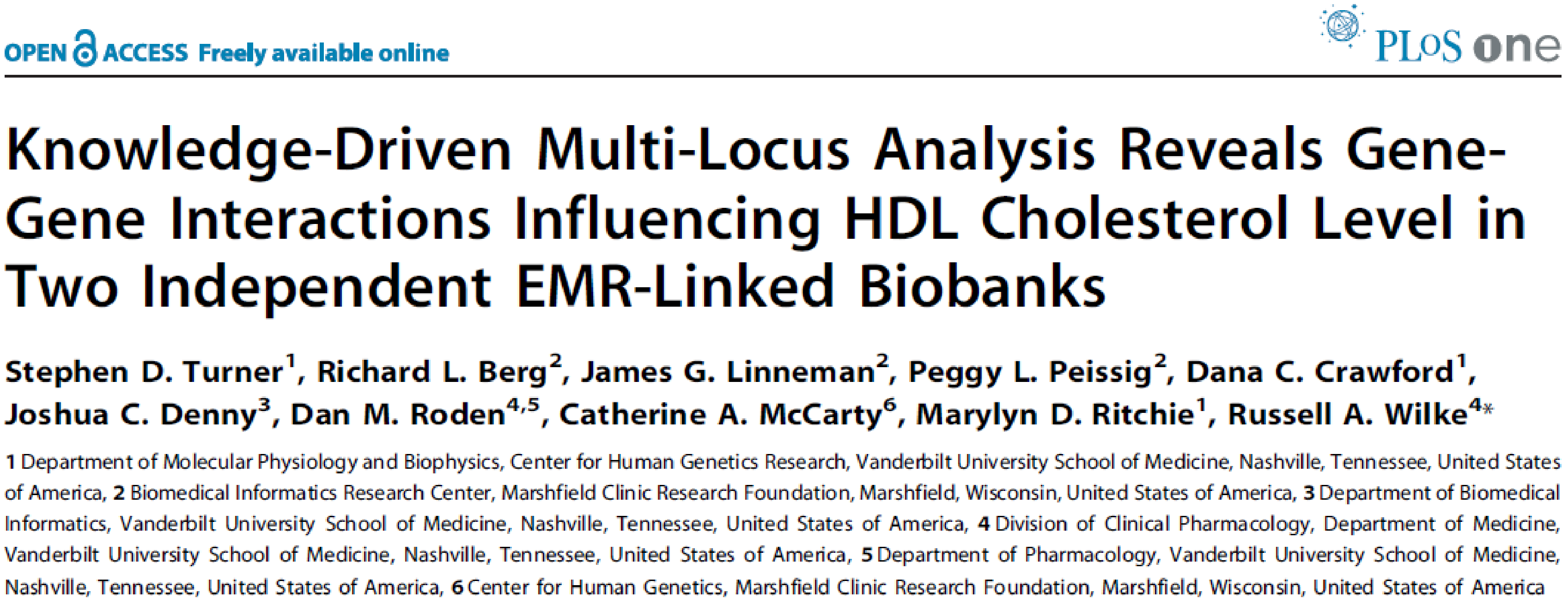

Cell Membrane
ACTN1
Actin Contraction, Cell Motility and Morphology Changes
MYH9
ER/SR
PIP$_2$
MYLK
Acetylcholine
CHRM3
CALM
PLCβ1
Gγ
IP$_3$
IP3R
PLCβ4
SCIN
Ca^{2+}
Cytoplasm
Actin Stabilization

Figure 1



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

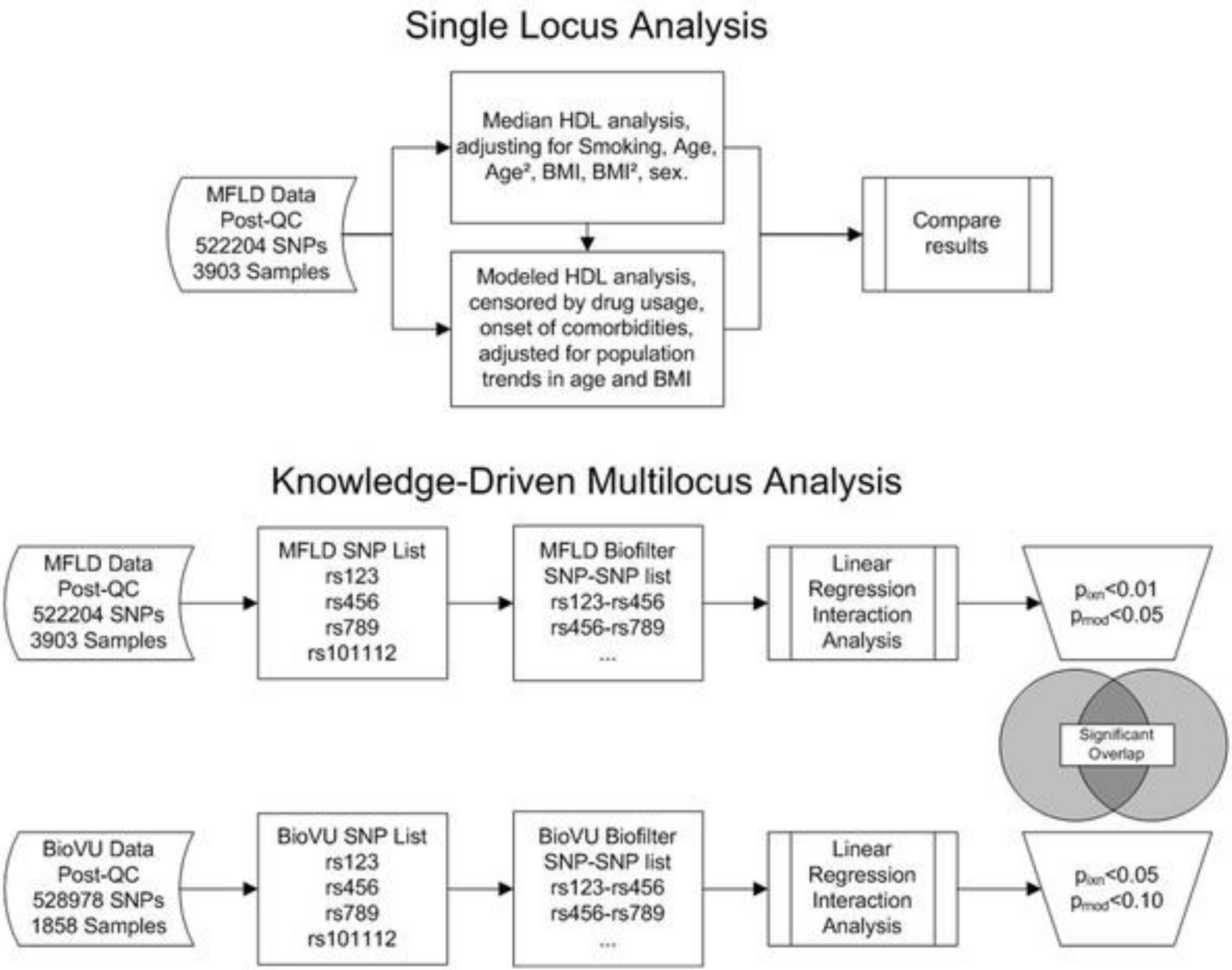

Single Locus Analysis
MFLD Data
Post-QC
522204 SNPs
3903 Samples
Median HDL analysis, adjusting for Smoking, Age, Age², BMI, BMI², sex.
Modeled HDL analysis, censored by drug usage, onset of comorbidities, adjusted for population trends in age and BMI
Compare results
Knowledge-Driven Multilocus Analysis
MFLD Data
Post-QC
522204 SNPs
3903 Samples
MFLD SNP List
rs123
rs456
rs789
rs101112
MFLD Biofilter SNP-SNP list
rs123-rs456
rs456-rs789
...
Linear Regression Interaction Analysis
$p_{int}<0.01$
$p_{mod}<0.05$
Significant Overlap
BioVU Data
Post-QC
528978 SNPs
1858 Samples
BioVU SNP List
rs123
rs456
rs789
rs101112
BioVU Biofilter SNP-SNP list
rs123-rs456
rs456-rs789
...
Linear Regression Interaction Analysis
$p_{int}<0.05$
$p_{mod}<0.10$

Figure 1

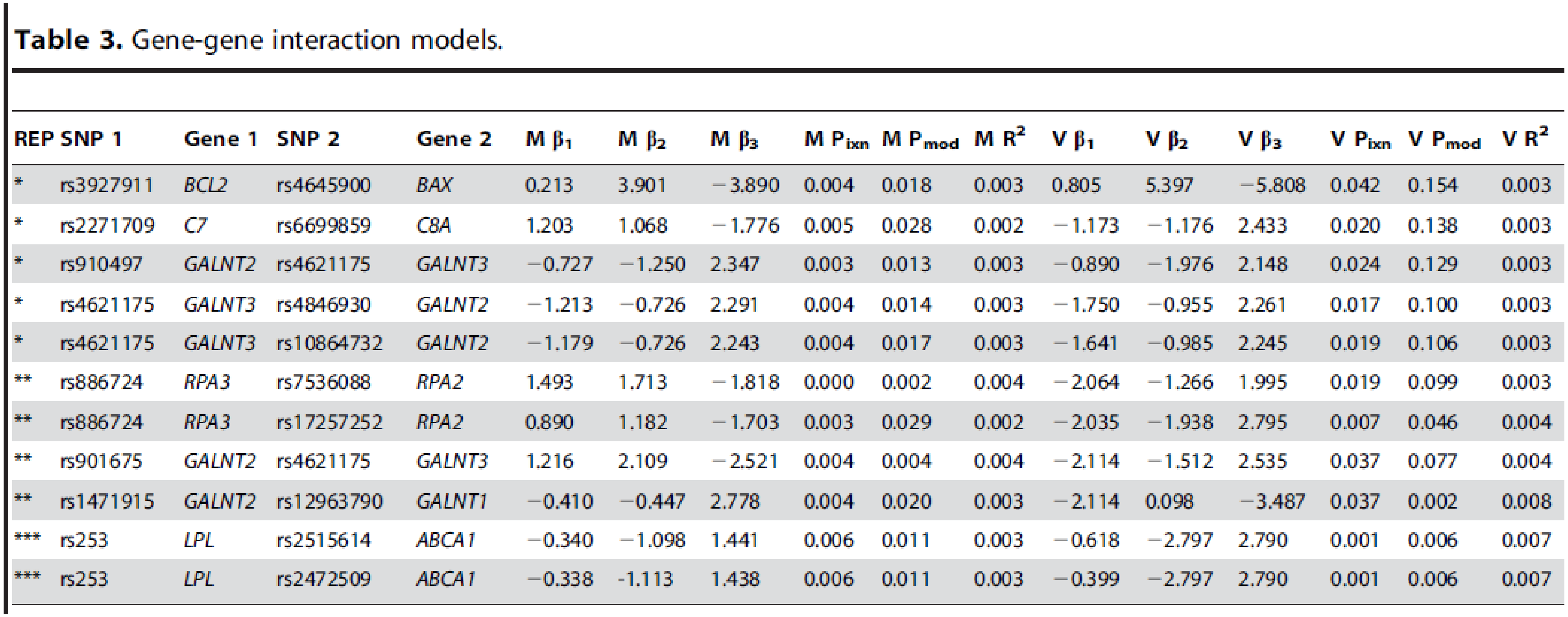


Table 3. Gene-gene interaction models.

REP	SNP 1	Gene 1	SNP 2	Gene 2	M β_1	M β_2	M β_3	M P_{ixn}	M P_{mod}	M R^2	V β_1	V β_2	V β_3	V P_{ixn}	V P_{mod}	V R^2
*	rs3927911	*BCL2*	rs4645900	*BAX*	0.213	3.901	−3.890	0.004	0.018	0.003	0.805	5.397	−5.808	0.042	0.154	0.003
*	rs2271709	*C7*	rs6699859	*C8A*	1.203	1.068	−1.776	0.005	0.028	0.002	−1.173	−1.176	2.433	0.020	0.138	0.003
*	rs910497	*GALNT2*	rs4621175	*GALNT3*	−0.727	−1.250	2.347	0.003	0.013	0.003	−0.890	−1.976	2.148	0.024	0.129	0.003
*	rs4621175	*GALNT3*	rs4846930	*GALNT2*	−1.213	−0.726	2.291	0.004	0.014	0.003	−1.750	−0.955	2.261	0.017	0.100	0.003
*	rs4621175	*GALNT3*	rs10864732	*GALNT2*	−1.179	−0.726	2.243	0.004	0.017	0.003	−1.641	−0.985	2.245	0.019	0.106	0.003
**	rs886724	*RPA3*	rs7536088	*RPA2*	1.493	1.713	−1.818	0.000	0.002	0.004	−2.064	−1.266	1.995	0.019	0.099	0.003
**	rs886724	*RPA3*	rs17257252	*RPA2*	0.890	1.182	−1.703	0.003	0.029	0.002	−2.035	−1.938	2.795	0.007	0.046	0.004
**	rs901675	*GALNT2*	rs4621175	*GALNT3*	1.216	2.109	−2.521	0.004	0.004	0.004	−2.114	−1.512	2.535	0.037	0.077	0.004
**	rs1471915	*GALNT2*	rs12963790	*GALNT1*	−0.410	−0.447	2.778	0.004	0.020	0.003	−2.114	0.098	−3.487	0.037	0.002	0.008
***	rs253	*LPL*	rs2515614	*ABCA1*	−0.340	−1.098	1.441	0.006	0.011	0.003	−0.618	−2.797	2.790	0.001	0.006	0.007
***	rs253	*LPL*	rs2472509	*ABCA1*	−0.338	-1.113	1.438	0.006	0.011	0.003	−0.399	−2.797	2.790	0.001	0.006	0.007

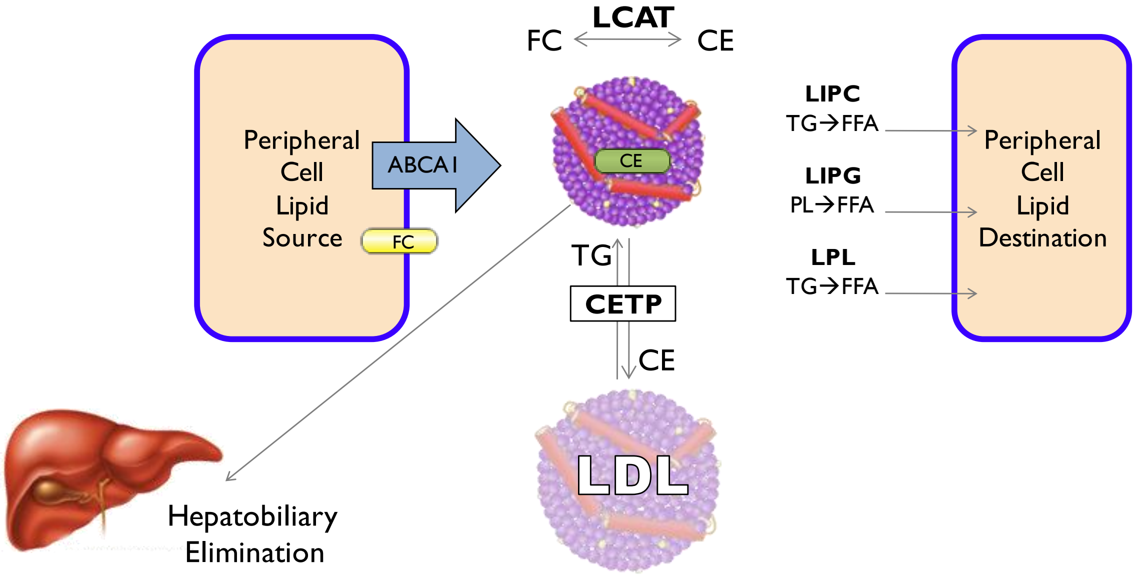
- **Tested 22,769 two-SNP models in Marshfield (discovery).**
  - **11 significant (p_{int}<0.01, p_{anova}<0.05)**
- **Tested 11 two-SNP models in BioVU (replication).**
  - **6 marginally significant (p_{int}<0.05, p_{anova}<0.10).**
  - **2 had consistent direction for all three βs.**

Application of the Biofilter:
HDL - eMERGE

- Main effects of each SNP in each dataset reduce HDL.
- Interaction effect coefficient is positive
  - Joint effect is nonlinear
  - Epistasis – heterogeneity, antagonism, negative epistasis
  - This kind of effect also seen in 4/5 sig. GxG interactions in IDDM (Barrett et al. 2009 *Nature Genetics*)

SNP 1	**Gene 1**	**SNP 2**	**Gene 2**	**MF β_1**	**MF β_2**	**MF β_3**	**MF P**	**BioVU β_1**	**BioVU β_2**	**BioVU β_3**	**BioVU P**
rs253	LPL	rs2515614	ABCA1	-	-	+	0.006	-	-	+	0.001
rs253	LPL	rs2472509	ABCA1	-	-	+	0.006	-	-	+	0.001

Turner et al, PLoS ONE 2011.



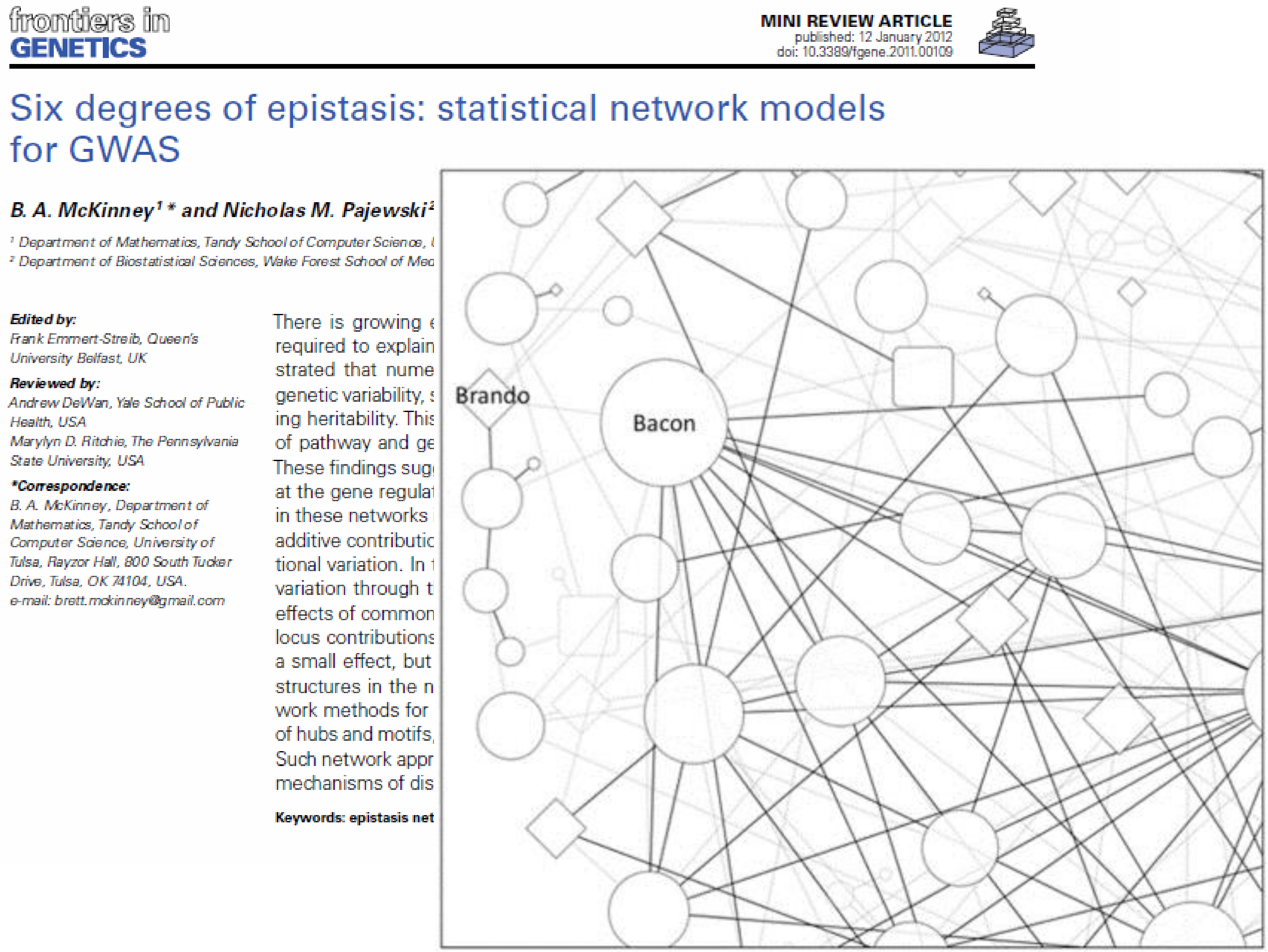
- LPL mediates the release of FFA and TG from HDL particles.
- ABCA1 shuttles free cholesterol into HDL particles during intravascular remodeling.

SNP 1	Gene 1	SNP 2	Gene 2	MF β_1	MF β_2	MF β_3	MF P	BioVU β_1	BioVU β_2	BioVU β_3	BioVU P
rs253	LPL	rs2515614	ABCA1	-	-	+	0.006	-	-	+	0.001
rs253	LPL	rs2472509	ABCA1	-	-	+	0.006	-	-	+	0.001

Turner et al, PLoS ONE 2011.

Pathway Analysis
Biofilter

Beyond simple epistasis models….



frontiers in GENETICS

MINI REVIEW ARTICLE
published: 12 January 2012
doi: 10.3389/fgene.2011.00109

Six degrees of epistasis: statistical network models for GWAS

B. A. McKinney[1] * and Nicholas M. Pajewski[2]

[1] Department of Mathematics, Tandy School of Computer Science, U
[2] Department of Biostatistical Sciences, Wake Forest School of Med

Edited by:
Frank Emmert-Streib, Queen's University Belfast, UK

Reviewed by:
Andrew DeWan, Yale School of Public Health, USA
Marylyn D. Ritchie, The Pennsylvania State University, USA

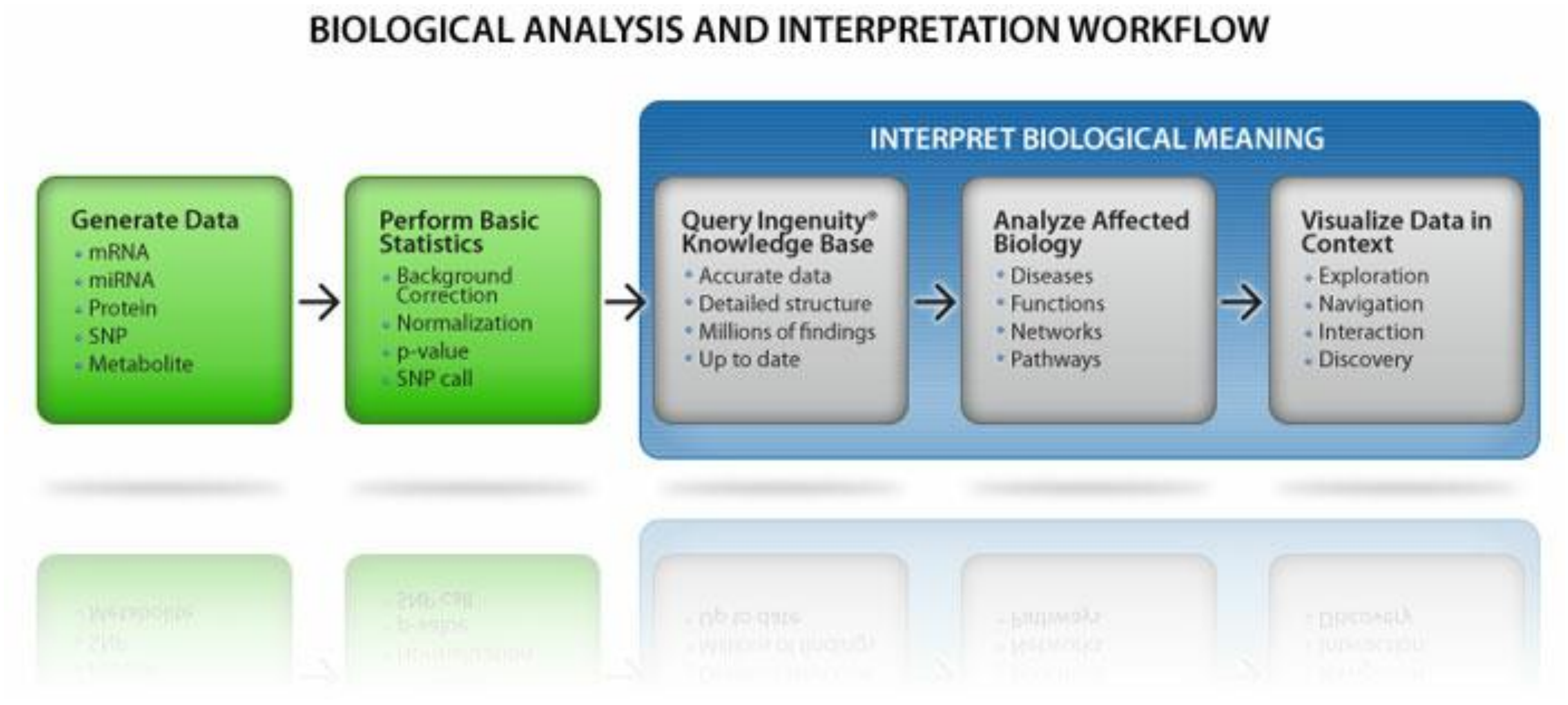
***Correspondence:**
B. A. McKinney, Department of Mathematics, Tandy School of Computer Science, University of Tulsa, Rayzor Hall, 800 South Tucker Drive, Tulsa, OK 74104, USA.
e-mail: brett.mckinney@gmail.com

There is growing e
required to explain
strated that nume
genetic variability, s
ing heritability. This
of pathway and ge
These findings sug
at the gene regulat
in these networks
additive contributio
tional variation. In t
variation through t
effects of common
locus contributions
a small effect, but
structures in the n
work methods for
of hubs and motifs,
Such network appr
mechanisms of dis

Keywords: epistasis net

Pathway Analysis Approaches

- Ingenuity systems pathway analysis
  - IPA www.ingenuity.com (free trial)



BIOLOGICAL ANALYSIS AND INTERPRETATION WORKFLOW

Generate Data
- mRNA
- miRNA
- Protein
- SNP
- Metabolite

→

Perform Basic Statistics
- Background Correction
- Normalization
- p-value
- SNP call

→

INTERPRET BIOLOGICAL MEANING

Query Ingenuity® Knowledge Base
- Accurate data
- Detailed structure
- Millions of findings
- Up to date

→

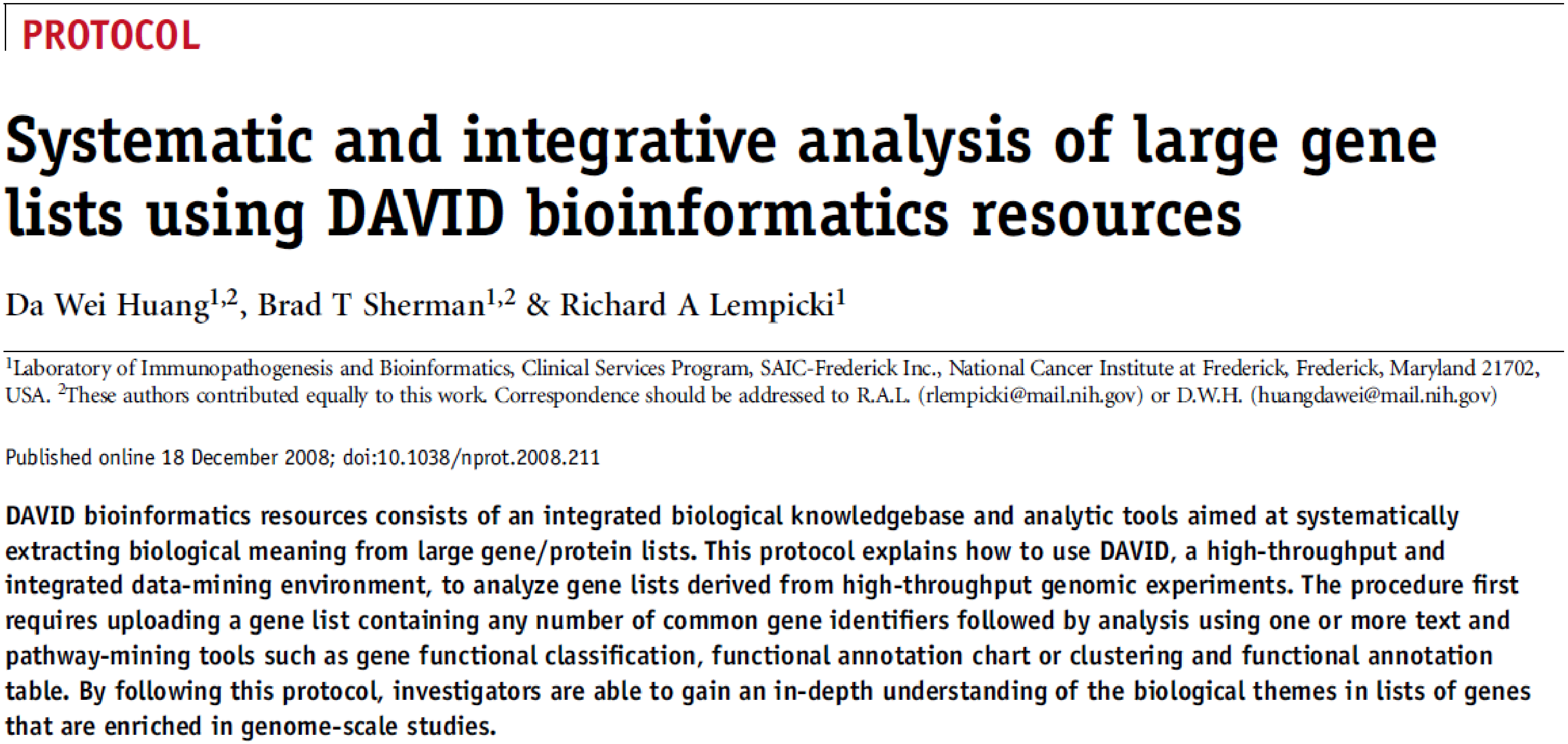
Analyze Affected Biology
- Diseases
- Functions
- Networks
- Pathways

→

Visualize Data in Context
- Exploration
- Navigation
- Interaction
- Discovery

Pathway Analysis Approaches

- Database for Annotation, Visualization and Integrated Discovery (DAVID)
  - provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes



PROTOCOL

Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources

Da Wei Huang[1,2], Brad T Sherman[1,2] & Richard A Lempicki[1]

[1]Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to R.A.L. (rlempicki@mail.nih.gov) or D.W.H. (huangdawei@mail.nih.gov)

Published online 18 December 2008; doi:10.1038/nprot.2008.211

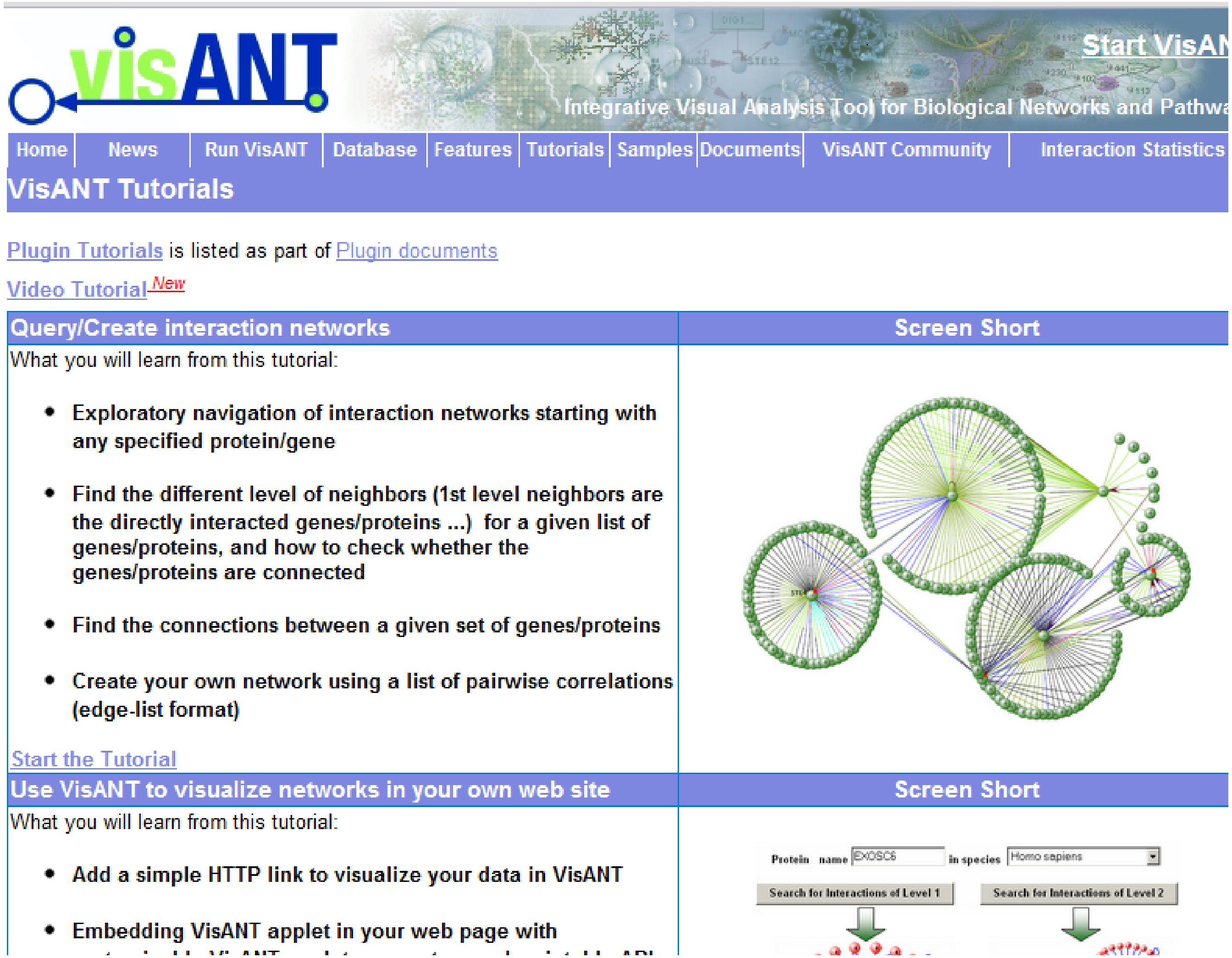
DAVID bioinformatics resources consists of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists. This protocol explains how to use DAVID, a high-throughput and integrated data-mining environment, to analyze gene lists derived from high-throughput genomic experiments. The procedure first requires uploading a gene list containing any number of common gene identifiers followed by analysis using one or more text and pathway-mining tools such as gene functional classification, functional annotation chart or clustering and functional annotation table. By following this protocol, investigators are able to gain an in-depth understanding of the biological themes in lists of genes that are enriched in genome-scale studies.

- **Step-by-step instructions for using DAVID**

DAVID tools are able to…

1. Identify enriched biological themes, particularly GO terms
2. Discover enriched functional-related gene groups
3. Cluster redundant annotation terms
4. Visualize genes on BioCarta & KEGG pathway maps
5. Display related many-genes-to-many-terms on 2-D view
6. Search for other functionally related genes not in the list
7. List interacting proteins
8. Explore gene names in batch
9. Link gene-disease associations
10. Highlight protein functional domains and motifs
11. Redirect to related literatures
12. Convert gene identifiers from one type to another
13. And more

Pathway Analysis Approaches



visANT

Start VisANT

Integrative Visual Analysis Tool for Biological Networks and Pathwa

Home | News | Run VisANT | Database | Features | Tutorials | Samples | Documents | VisANT Community | Interaction Statistics

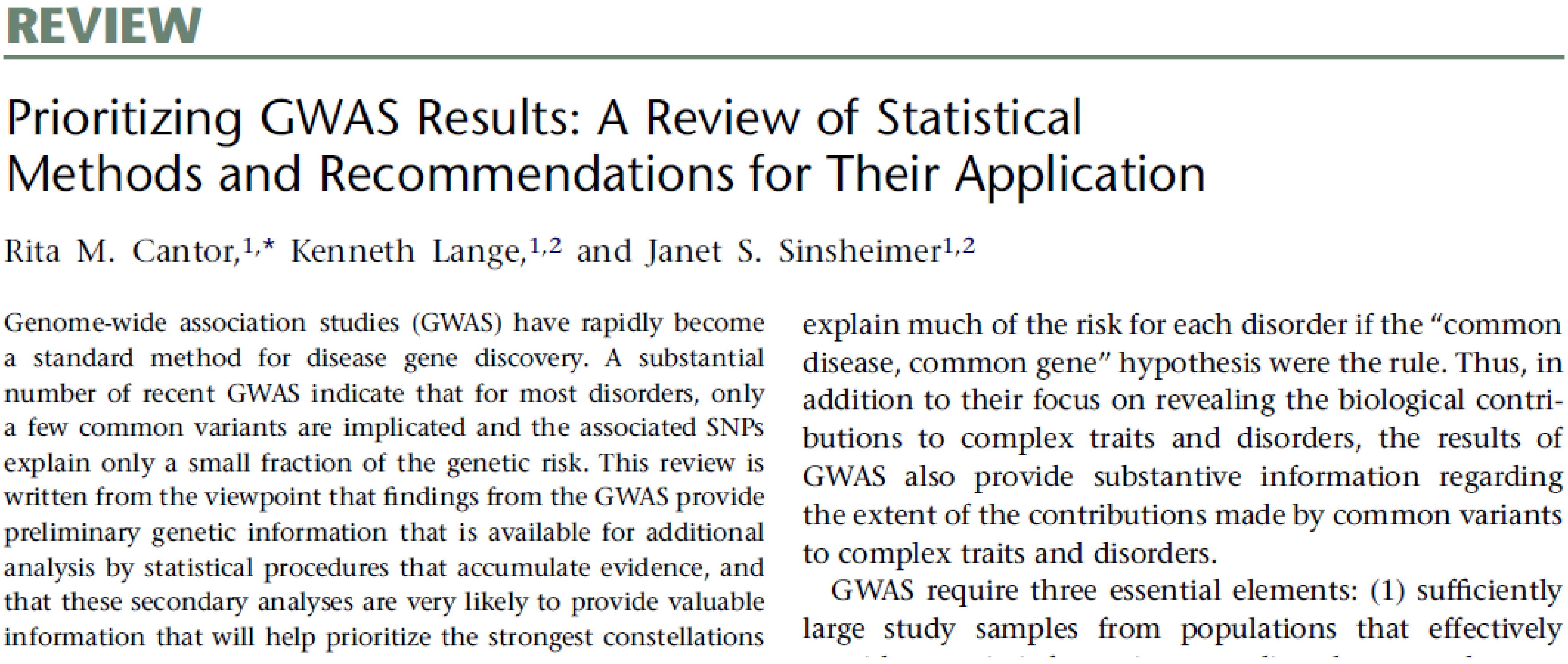
VisANT Tutorials

Plugin Tutorials is listed as part of Plugin documents

Video Tutorial *New*

Query/Create interaction networks	Screen Short
What you will learn from this tutorial: • Exploratory navigation of interaction networks starting with any specified protein/gene • Find the different level of neighbors (1st level neighbors are the directly interacted genes/proteins ...) for a given list of genes/proteins, and how to check whether the genes/proteins are connected • Find the connections between a given set of genes/proteins • Create your own network using a list of pairwise correlations (edge-list format) Start the Tutorial	
Use VisANT to visualize networks in your own web site	**Screen Short**
What you will learn from this tutorial: • Add a simple HTTP link to visualize your data in VisANT • Embedding VisANT applet in your web page with	Protein name EXOSC6 in species Homo sapiens Search for Interactions of Level 1 Search for Interactions of Level 2

Pathway Analysis Approaches



REVIEW

Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application

Rita M. Cantor,[1,*] Kenneth Lange,[1,2] and Janet S. Sinsheimer[1,2]

Genome-wide association studies (GWAS) have rapidly become a standard method for disease gene discovery. A substantial number of recent GWAS indicate that for most disorders, only a few common variants are implicated and the associated SNPs explain only a small fraction of the genetic risk. This review is written from the viewpoint that findings from the GWAS provide preliminary genetic information that is available for additional analysis by statistical procedures that accumulate evidence, and that these secondary analyses are very likely to provide valuable information that will help prioritize the strongest constellations

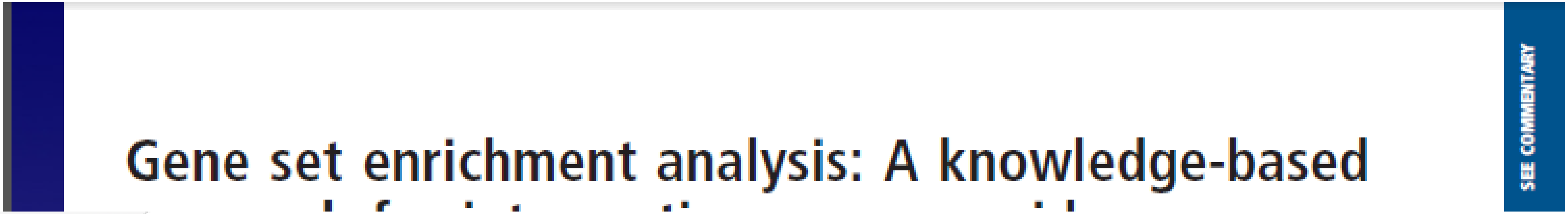
explain much of the risk for each disorder if the "common disease, common gene" hypothesis were the rule. Thus, in addition to their focus on revealing the biological contributions to complex traits and disorders, the results of GWAS also provide substantive information regarding the extent of the contributions made by common variants to complex traits and disorders.

GWAS require three essential elements: (1) sufficiently large study samples from populations that effectively

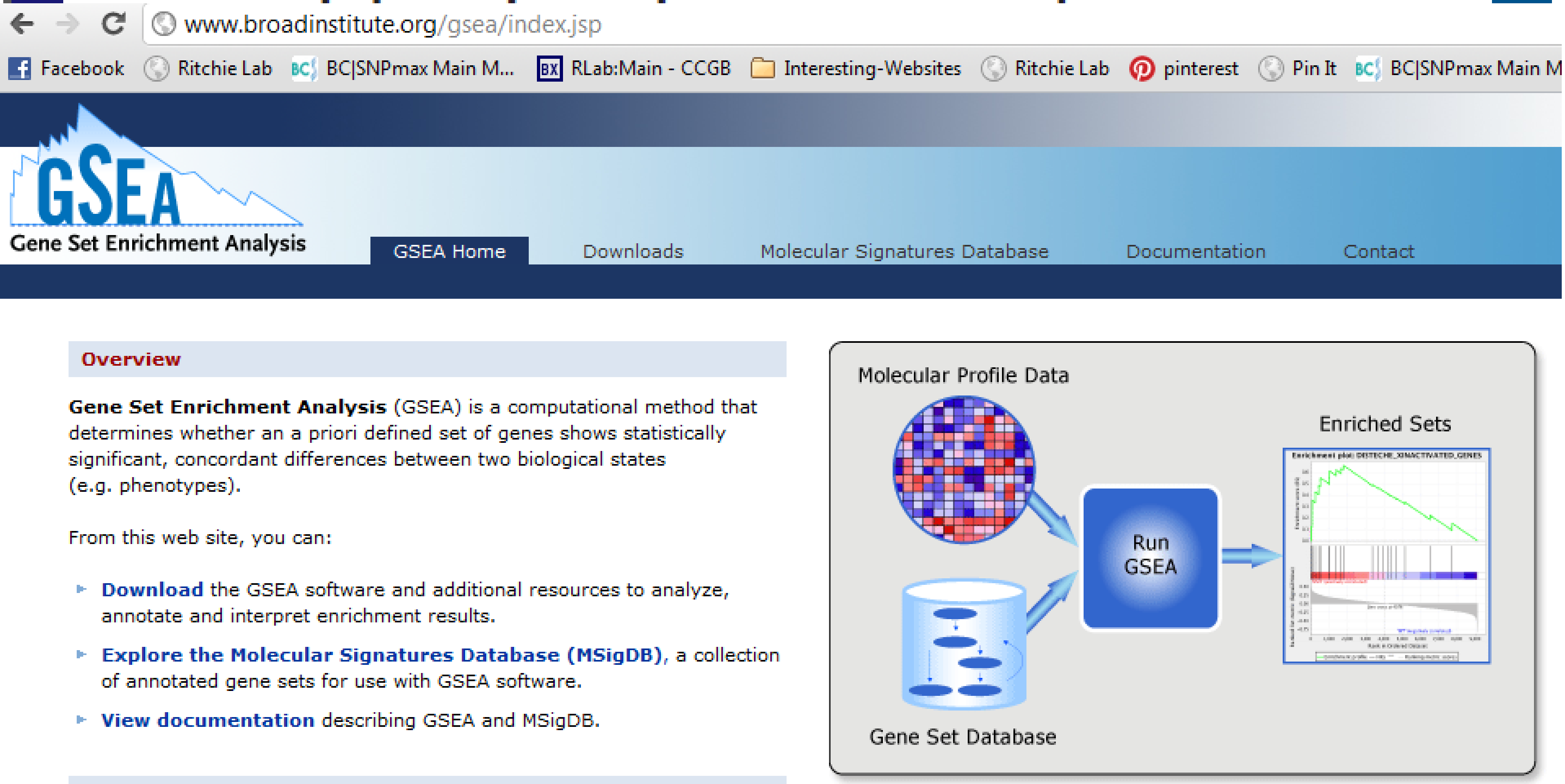
Alternative pathway approaches

- Multiple pathway approaches in development
    - Gene set enrichment analysis (GSEA)
    - INTERSNP
    - PATH
    - Prioritizer
    - and many more…..
- Many use overlapping sources of data
- All have strengths and weaknesses

Alternative pathway approaches



Gene set enrichment analysis: A knowledge-based



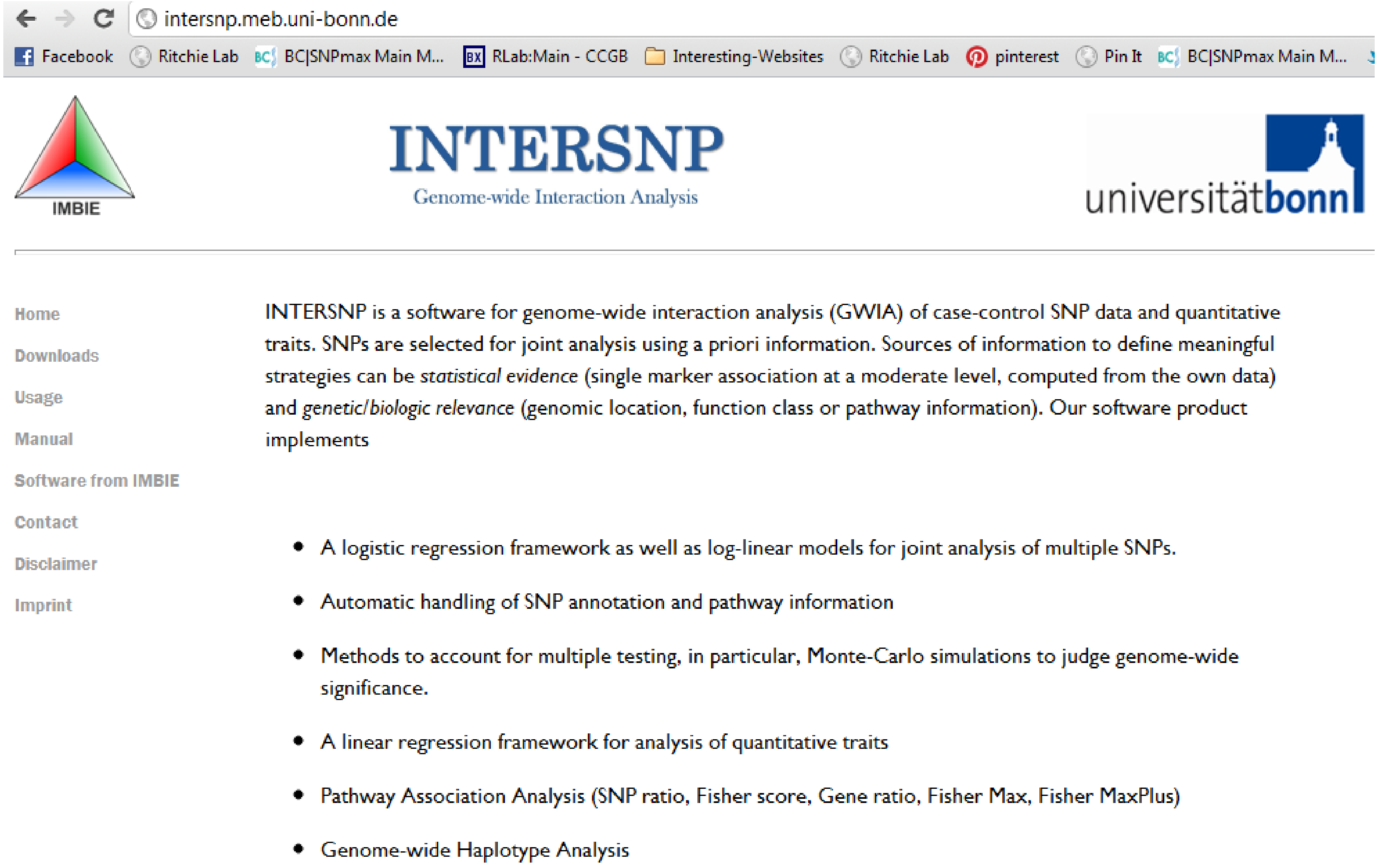
Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- **View documentation** describing GSEA and MSigDB.

Alternative pathway approaches



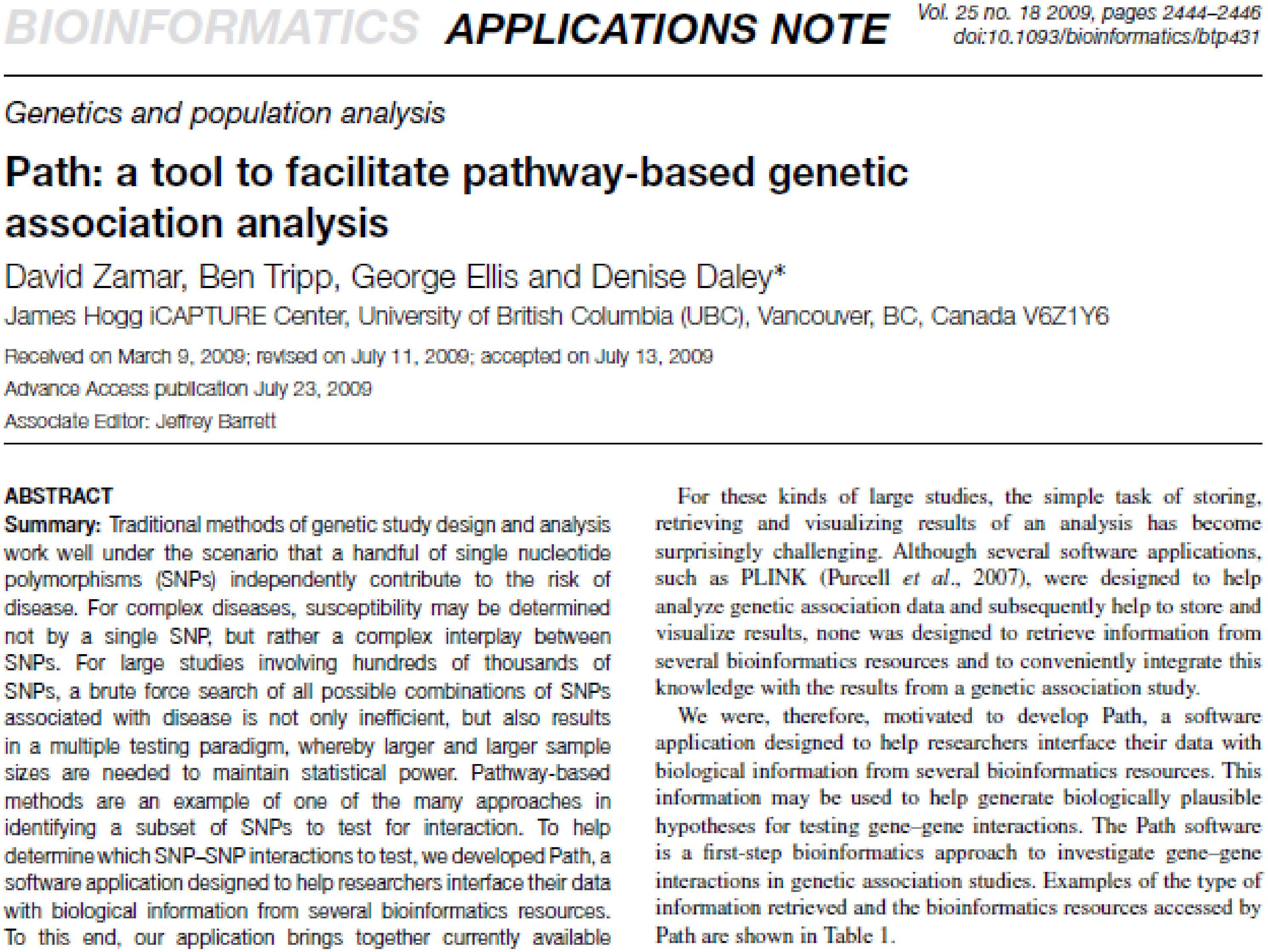
INTERSNP

Genome-wide Interaction Analysis

INTERSNP is a software for genome-wide interaction analysis (GWIA) of case-control SNP data and quantitative traits. SNPs are selected for joint analysis using a priori information. Sources of information to define meaningful strategies can be *statistical evidence* (single marker association at a moderate level, computed from the own data) and *genetic/biologic relevance* (genomic location, function class or pathway information). Our software product implements

- A logistic regression framework as well as log-linear models for joint analysis of multiple SNPs.
- Automatic handling of SNP annotation and pathway information
- Methods to account for multiple testing, in particular, Monte-Carlo simulations to judge genome-wide significance.
- A linear regression framework for analysis of quantitative traits
- Pathway Association Analysis (SNP ratio, Fisher score, Gene ratio, Fisher Max, Fisher MaxPlus)
- Genome-wide Haplotype Analysis

Alternative pathway approaches



BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 18 2009, pages 2444–2446
doi:10.1093/bioinformatics/btp431

Genetics and population analysis

Path: a tool to facilitate pathway-based genetic association analysis

David Zamar, Ben Tripp, George Ellis and Denise Daley*
James Hogg iCAPTURE Center, University of British Columbia (UBC), Vancouver, BC, Canada V6Z1Y6

Received on March 9, 2009; revised on July 11, 2009; accepted on July 13, 2009
Advance Access publication July 23, 2009
Associate Editor: Jeffrey Barrett

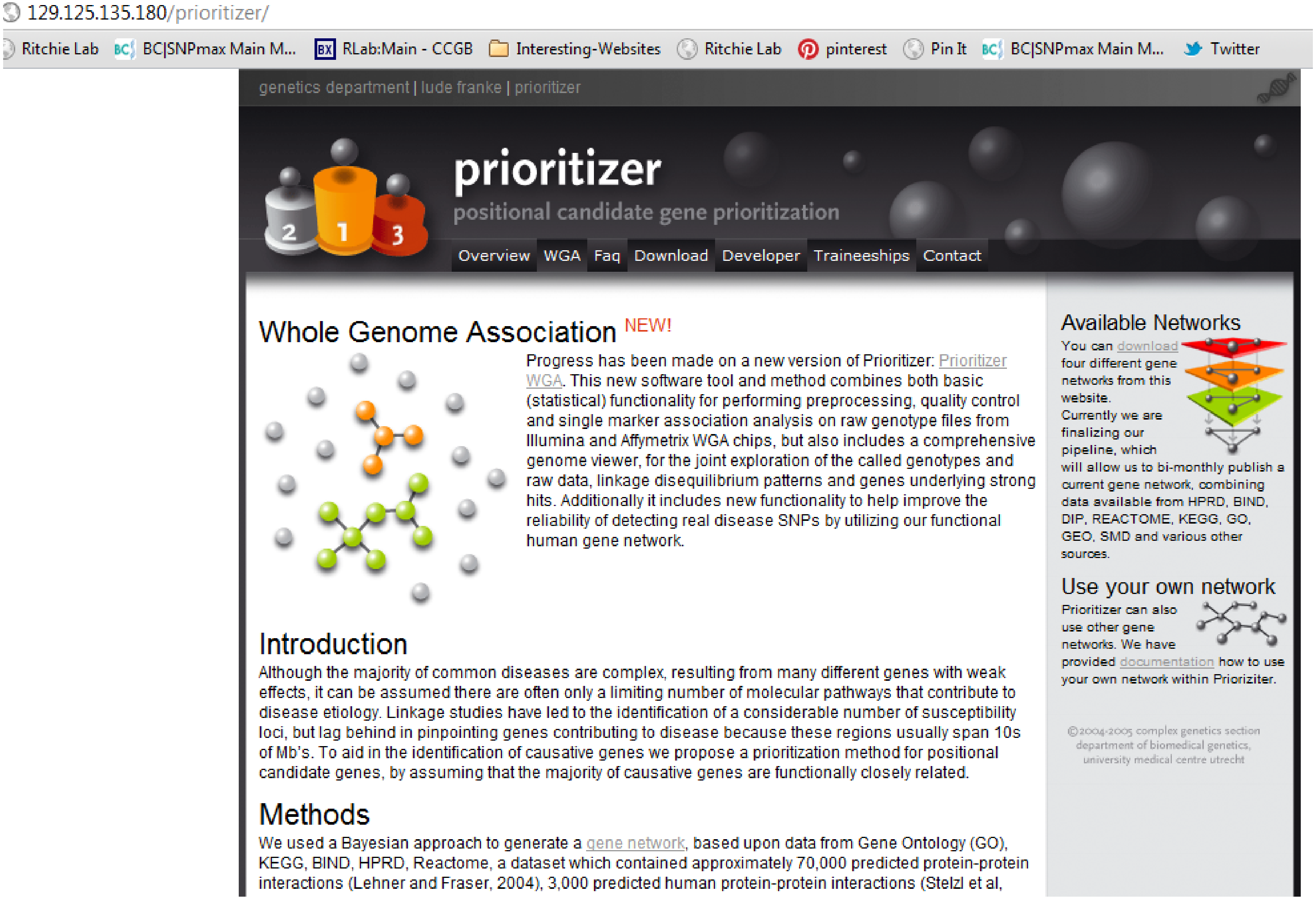
ABSTRACT

Summary: Traditional methods of genetic study design and analysis work well under the scenario that a handful of single nucleotide polymorphisms (SNPs) independently contribute to the risk of disease. For complex diseases, susceptibility may be determined not by a single SNP, but rather a complex interplay between SNPs. For large studies involving hundreds of thousands of SNPs, a brute force search of all possible combinations of SNPs associated with disease is not only inefficient, but also results in a multiple testing paradigm, whereby larger and larger sample sizes are needed to maintain statistical power. Pathway-based methods are an example of one of the many approaches in identifying a subset of SNPs to test for interaction. To help determine which SNP–SNP interactions to test, we developed Path, a software application designed to help researchers interface their data with biological information from several bioinformatics resources. To this end, our application brings together currently available

For these kinds of large studies, the simple task of storing, retrieving and visualizing results of an analysis has become surprisingly challenging. Although several software applications, such as PLINK (Purcell *et al.*, 2007), were designed to help analyze genetic association data and subsequently help to store and visualize results, none was designed to retrieve information from several bioinformatics resources and to conveniently integrate this knowledge with the results from a genetic association study.

We were, therefore, motivated to develop Path, a software application designed to help researchers interface their data with biological information from several bioinformatics resources. This information may be used to help generate biologically plausible hypotheses for testing gene–gene interactions. The Path software is a first-step bioinformatics approach to investigate gene–gene interactions in genetic association studies. Examples of the type of information retrieved and the bioinformatics resources accessed by Path are shown in Table 1.

Alternative pathway approaches



prioritizer

positional candidate gene prioritization

Whole Genome Association NEW!

Progress has been made on a new version of Prioritizer: Prioritizer WGA. This new software tool and method combines both basic (statistical) functionality for performing preprocessing, quality control and single marker association analysis on raw genotype files from Illumina and Affymetrix WGA chips, but also includes a comprehensive genome viewer, for the joint exploration of the called genotypes and raw data, linkage disequilibrium patterns and genes underlying strong hits. Additionally it includes new functionality to help improve the reliability of detecting real disease SNPs by utilizing our functional human gene network.

Introduction

Although the majority of common diseases are complex, resulting from many different genes with weak effects, it can be assumed there are often only a limiting number of molecular pathways that contribute to disease etiology. Linkage studies have led to the identification of a considerable number of susceptibility loci, but lag behind in pinpointing genes contributing to disease because these regions usually span 10s of Mb's. To aid in the identification of causative genes we propose a prioritization method for positional candidate genes, by assuming that the majority of causative genes are functionally closely related.

Methods

We used a Bayesian approach to generate a gene network, based upon data from Gene Ontology (GO), KEGG, BIND, HPRD, Reactome, a dataset which contained approximately 70,000 predicted protein-protein interactions (Lehner and Fraser, 2004), 3,000 predicted human protein-protein interactions (Stelzl et al,

Available Networks

You can download four different gene networks from this website. Currently we are finalizing our pipeline, which will allow us to bi-monthly publish a current gene network, combining data available from HPRD, BIND, DIP, REACTOME, KEGG, GO, GEO, SMD and various other sources.

Use your own network

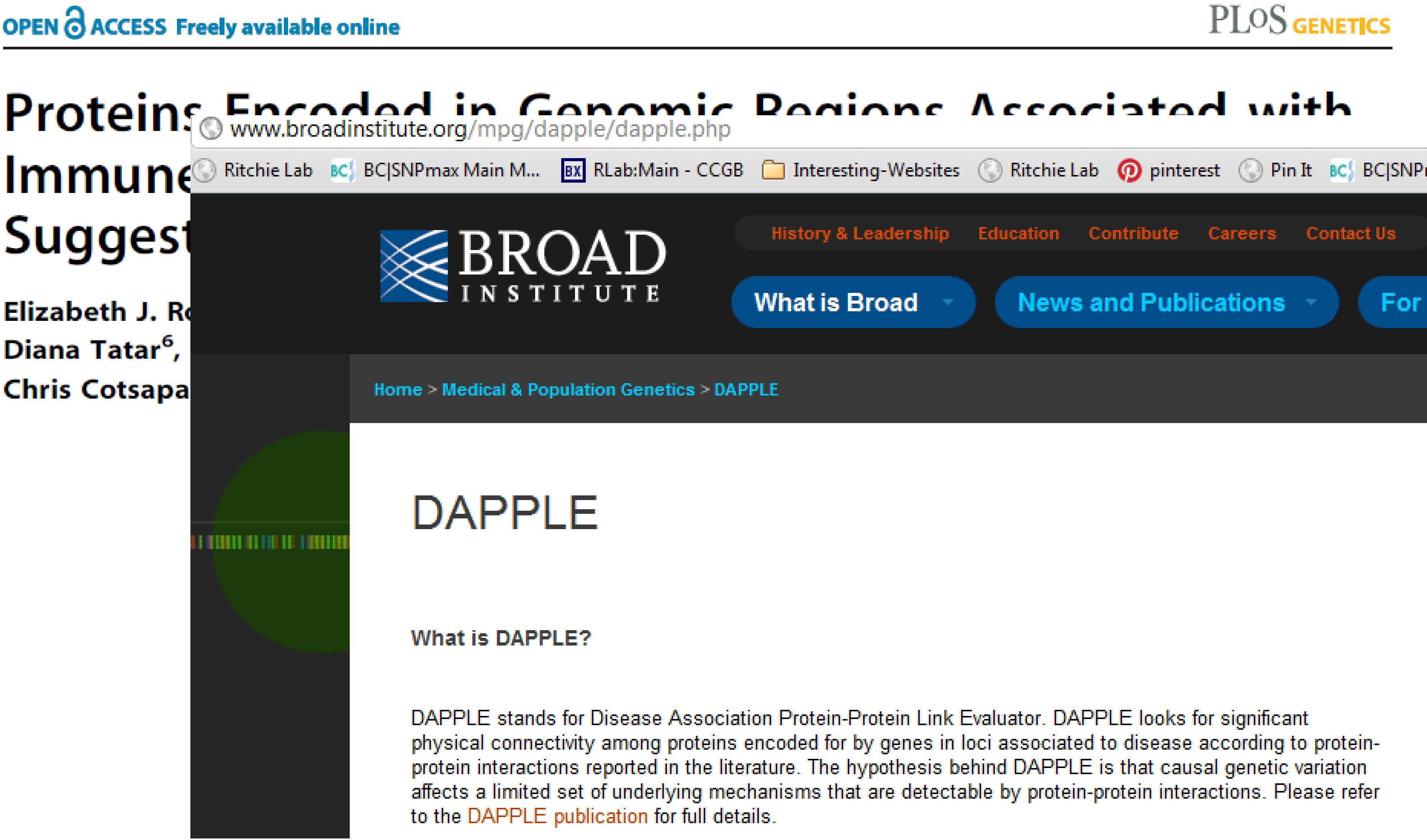
Prioritizer can also use other gene networks. We have provided documentation how to use your own network within Prioriziter.

©2004-2005 complex genetics section
department of biomedical genetics,
university medical centre utrecht

Alternative knowledge base approaches

- Protein-protein interaction databases
- Gene ontology
- Function-based GWAS
  - Using eQTL information
- Text mining applications
  - Textspresso
  - GRAIL

Gene based



OPEN ACCESS Freely available online

PLOS GENETICS

Proteins Encoded in Genomic Regions Associated with

Immune

Suggest

Elizabeth J. R

Diana Tatar6,

Chris Cotsapa

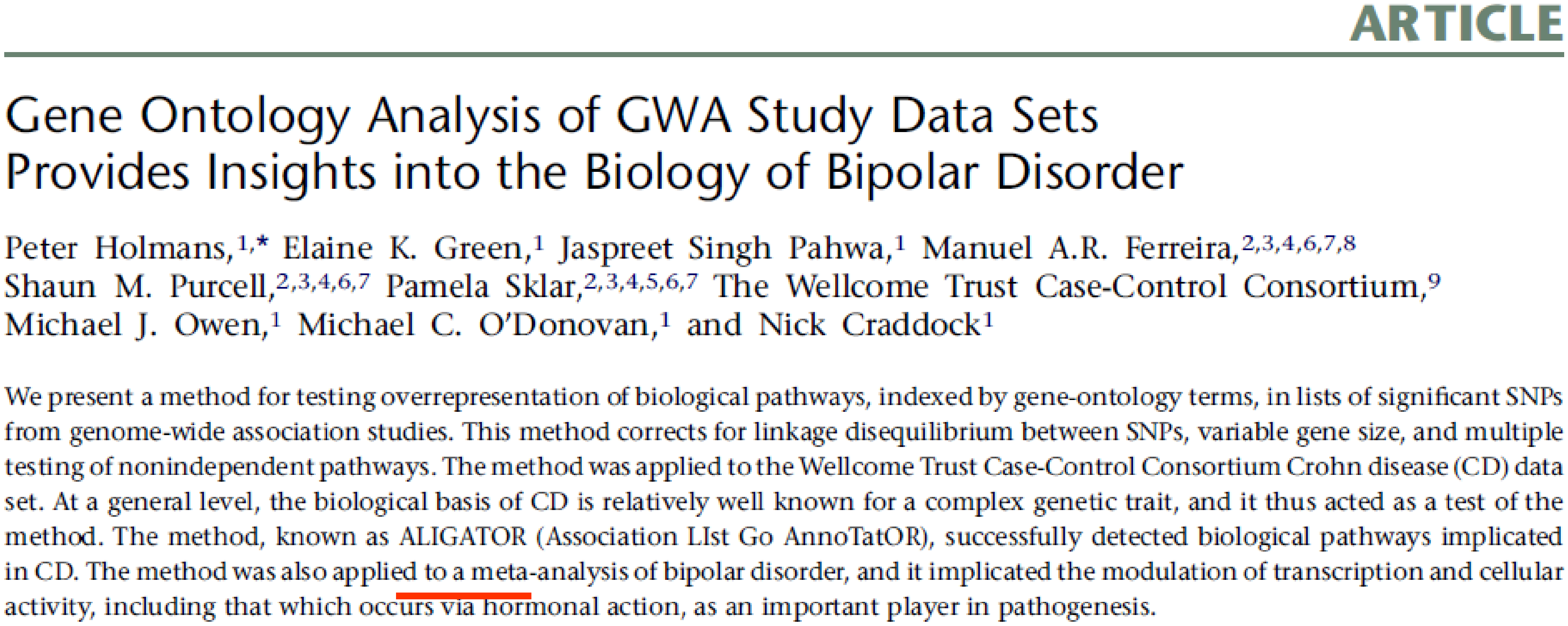
Home > Medical & Population Genetics > DAPPLE

DAPPLE

What is DAPPLE?

DAPPLE stands for Disease Association Protein-Protein Link Evaluator. DAPPLE looks for significant physical connectivity among proteins encoded for by genes in loci associated to disease according to protein-protein interactions reported in the literature. The hypothesis behind DAPPLE is that causal genetic variation affects a limited set of underlying mechanisms that are detectable by protein-protein interactions. Please refer to the DAPPLE publication for full details.

Gene based analysis



ARTICLE

Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder

Peter Holmans,[1,*] Elaine K. Green,[1] Jaspreet Singh Pahwa,[1] Manuel A.R. Ferreira,[2,3,4,6,7,8] Shaun M. Purcell,[2,3,4,6,7] Pamela Sklar,[2,3,4,5,6,7] The Wellcome Trust Case-Control Consortium,[9] Michael J. Owen,[1] Michael C. O'Donovan,[1] and Nick Craddock[1]

We present a method for testing overrepresentation of biological pathways, indexed by gene-ontology terms, in lists of significant SNPs from genome-wide association studies. This method corrects for linkage disequilibrium between SNPs, variable gene size, and multiple testing of nonindependent pathways. The method was applied to the Wellcome Trust Case-Control Consortium Crohn disease (CD) data set. At a general level, the biological basis of CD is relatively well known for a complex genetic trait, and it thus acted as a test of the method. The method, known as ALIGATOR (Association LIst Go AnnoTatOR), successfully detected biological pathways implicated in CD. The method was also applied to a meta-analysis of bipolar disorder, and it implicated the modulation of transcription and cellular activity, including that which occurs via hormonal action, as an important player in pathogenesis.

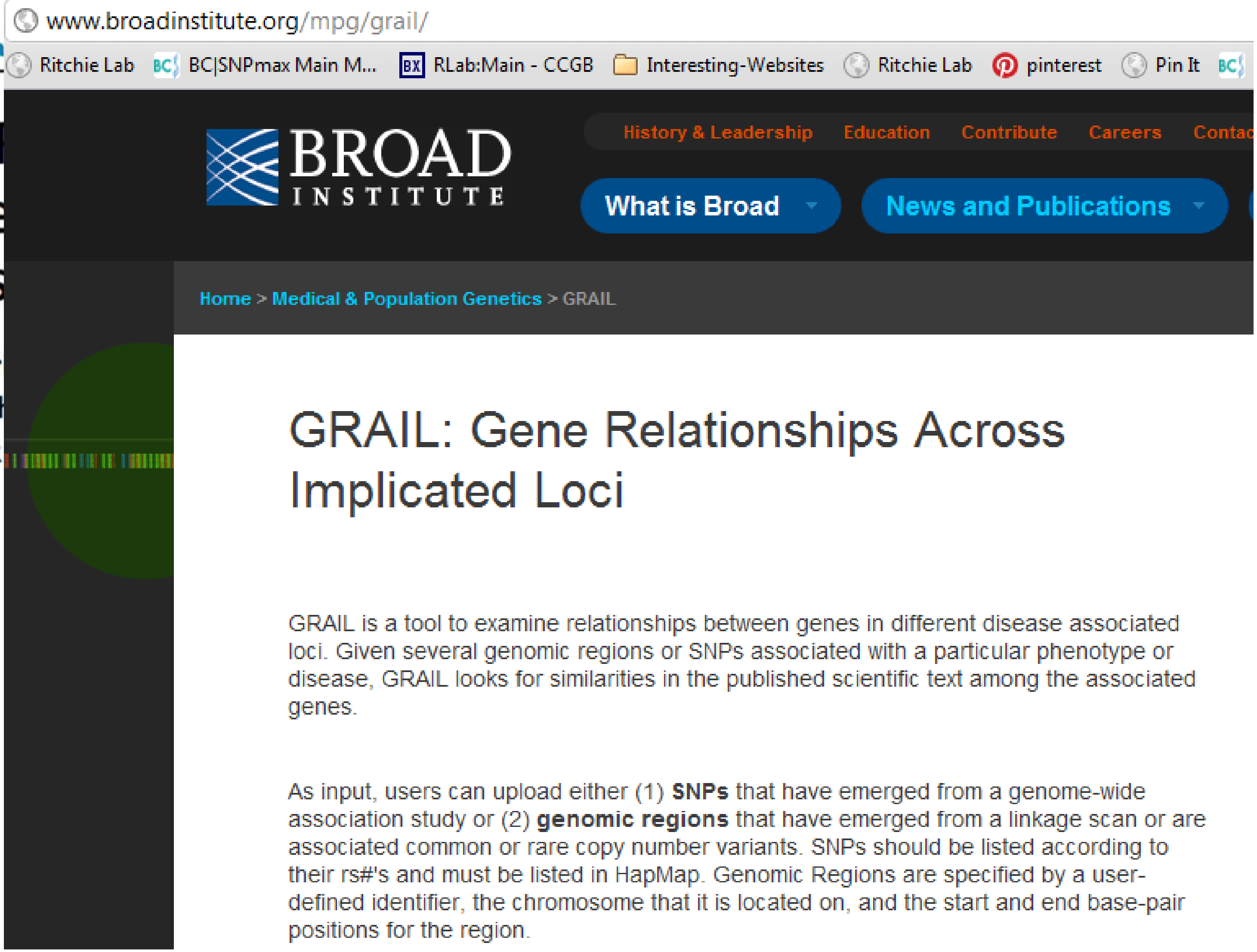
http://x004.psycm.uwcm.ac.uk/~peter/

Gene based analysis

OPEN ACCESS Freely availa

Identifying
Regions: Pre
Associations

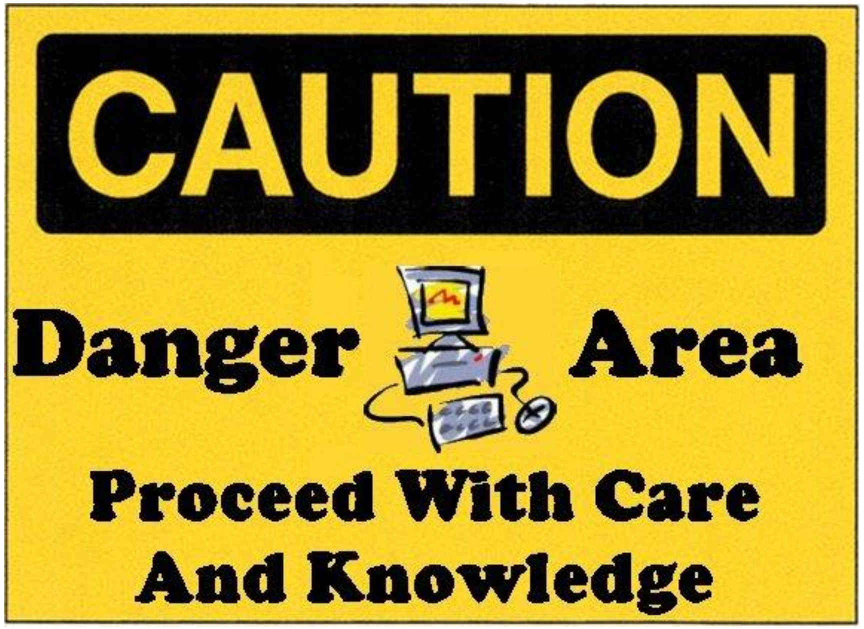
Soumya Raychaudhur
International Schizoph
Scolnick^{2,8,10}, Ramnik



GRAIL: Gene Relationships Across Implicated Loci

GRAIL is a tool to examine relationships between genes in different disease associated loci. Given several genomic regions or SNPs associated with a particular phenotype or disease, GRAIL looks for similarities in the published scientific text among the associated genes.

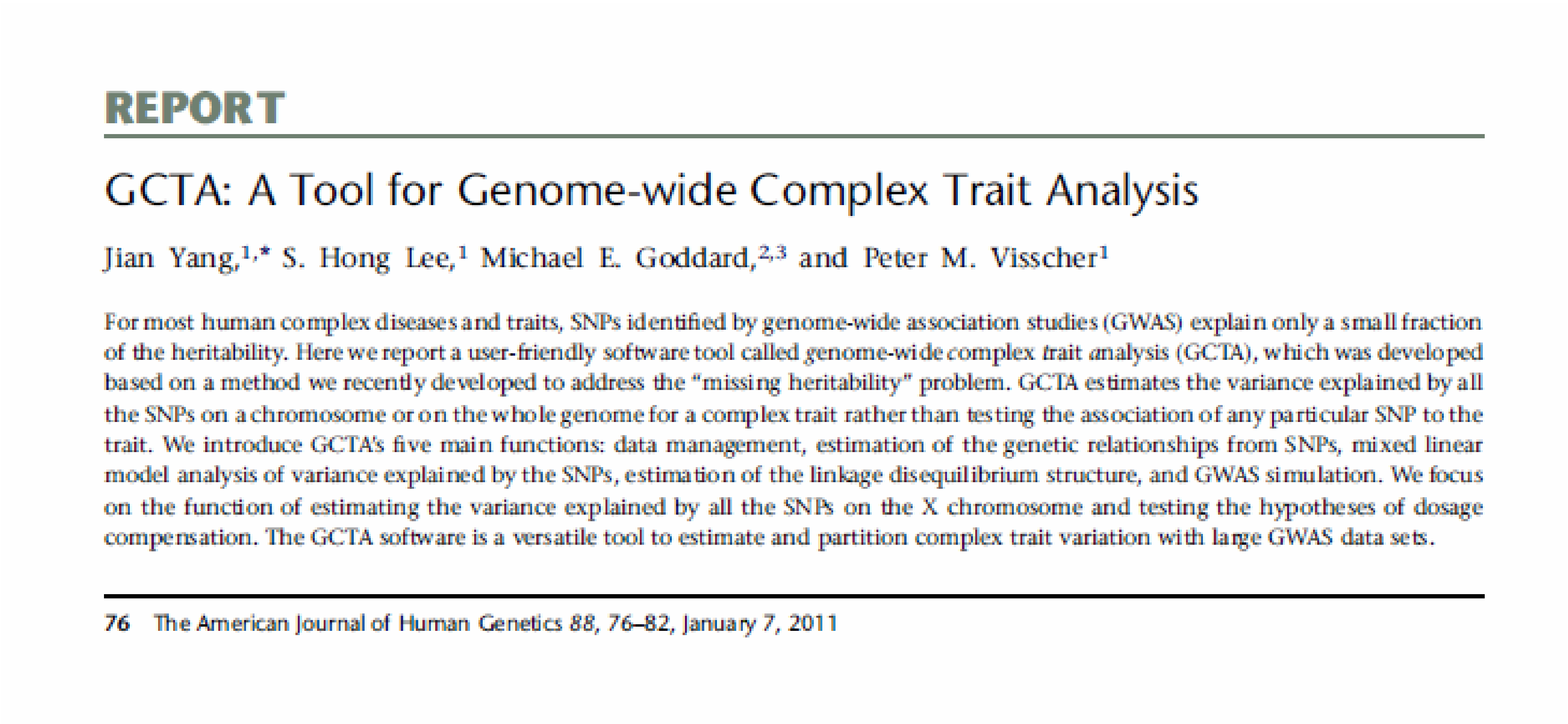
As input, users can upload either (1) **SNPs** that have emerged from a genome-wide association study or (2) **genomic regions** that have emerged from a linkage scan or are associated common or rare copy number variants. SNPs should be listed according to their rs#'s and must be listed in HapMap. Genomic Regions are specified by a user-defined identifier, the chromosome that it is located on, and the start and end base-pair positions for the region.



- Interpretation
  - Easy to create a story
- Size of gene/pathway
  - More likely to have significant results by chance if they are bigger
  - Use methods that perform permutation testing to account for gene/pathway size

Polygenic Modeling
Pathway Analysis
Biofilter

Polygenic modeling (En Masse)



REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

For most human complex diseases and traits, SNPs identified by genome-wide association studies (GWAS) explain only a small fraction of the heritability. Here we report a user-friendly software tool called *g*enome-wide *c*omplex *t*rait *a*nalysis (GCTA), which was developed based on a method we recently developed to address the "missing heritability" problem. GCTA estimates the variance explained by all the SNPs on a chromosome or on the whole genome for a complex trait rather than testing the association of any particular SNP to the trait. We introduce GCTA's five main functions: data management, estimation of the genetic relationships from SNPs, mixed linear model analysis of variance explained by the SNPs, estimation of the linkage disequilibrium structure, and GWAS simulation. We focus on the function of estimating the variance explained by all the SNPs on the X chromosome and testing the hypotheses of dosage compensation. The GCTA software is a versatile tool to estimate and partition complex trait variation with large GWAS data sets.

76 The American Journal of Human Genetics *88*, 76–82, January 7, 2011

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Schizophrenia is a severe mental disorder with a lifetime risk of about 1%, characterized by hallucinations, delusions and cognitive deficits, with heritability estimated at up to 80%[1,2]. We performed a genome-wide association study of 3,322 European individuals with schizophrenia and 3,587 controls. Here we show, using two analytic approaches, the extent to which common genetic variation underlies the risk of schizophrenia. First, we implicate the major histocompatibility complex. Second, we provide molecular genetic evidence for a substantial polygenic component to the risk of schizophrenia involving thousands of common alleles of very small effect. We show that this component also contributes to the risk of bipolar disorder, but not to several non-psychiatric diseases.

We genotyped the International Schizophrenia Consortium (ISC) case-control sample for up to ~1 million single nucleotide polymorphisms (SNPs), augmented by imputed common HapMap SNPs. In the genome-wide association study (GWAS; genomic control $\lambda_{GC} = 1.09$; Supplementary Table 1 and Supplementary Figs 1–3), the most associated genotyped SNP ($P = 3.4 \times 10^{-7}$) was located in the first intron of myosin XVIIIB (*MYO18B*) on chromosome 22. The second strongest association comprised more than 450 SNPs on chromosome 6p spanning the major histocompatibility complex (MHC; Fig. 1). There is some evidence for between-site heterogeneity in both allele frequencies and odds ratios (Table 1). We observed associations consistent with previous reports in the 22q11.2 deletion region and *ZNF804A* (ref. 3) (Supplementary Table 2, Supplementary Fig. 2 and section 5 and 6 in Supplementary Information).

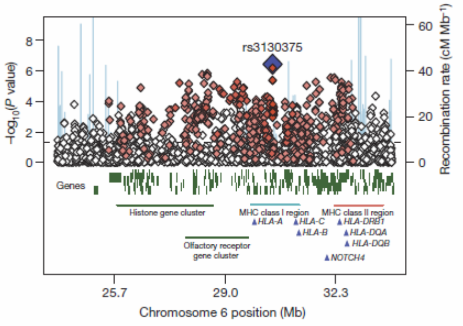


Figure 1 | Association results across the MHC region. Results are shown as $-\log_{10}$(P value) for genotyped SNPs. The most associated SNP is shown as a blue diamond. The colour of the remaining markers reflects r^2 with rs3130375, light pink, $r^2 > 0.1$, red, $r^2 > 0.8$. The recombination rate from the CEU HapMap (second y axis) is plotted in light blue.

The best imputed SNP, which reached genome-wide significance (rs3130297, $P = 4.79 \times 10^{-8}$, T allele odds ratio = 0.747, minor allele frequency (MAF) = 0.114, 32.3 megabases (Mb)), was also in the MHC, 7 kilobases (kb) from *NOTCH4*, a gene with previously reported associations with schizophrenia[4]. We imputed classical human leukocyte antigen (HLA) alleles; six were significant at $P < 10^{-3}$, found on the ancestral European haplotype[5] (Table 1, Supplementary Table 3 and section 3 in Supplementary Information). However, it was not possible to ascribe the association to a specific HLA allele, haplotype or region (Supplementary Table 3 and Supplementary Fig. 4).

We exchanged GWAS summary results with the Molecular Genetics of Schizophrenia (MGS) and SGENE consortia for genotyped SNPs with $P < 10^{-3}$. There were 8,008 cases and 19,077 controls of European descent in the combined sample (see refs 6, 7 and section 7 in Supplementary Information). Our top genotyped MHC SNP (rs3130375) had $P = 0.086$ and $P = 0.14$ in MGS and SGENE, respectively. Considering the combined results for genotyped and imputed SNPs across the MHC region more broadly, rs13194053 had a genome-wide significant combined $P = 9.5 \times 10^{-9}$ (ISC, MGS and SGENE: $P = 3 \times 10^{-4}$, 1×10^{-2} and 1×10^{-4}, respectively; C allele

Table 1 | MHC association for the most significant genotyped SNP rs3130375

a MHC association for rs3130375 by sample

		Frequency (rs3130375, A allele)		
Sample	Ancestry	Cases	Controls	*P* value
University of Aberdeen	Scottish	0.132	0.168	0.0060
University of Edinburgh	Scottish	0.137	0.135	0.8930
University College London*	British	0.132	0.143	0.4836
Trinity College Dublin	Irish	0.110	0.170	0.0012
Cardiff University	Bulgarian	0.077	0.084	0.5602
Portuguese Island Collection	Portuguese	0.048	0.061	0.3510
Karolinska Institutet (5.0)	Swedish	0.043	0.119	0.0004
Karolinska Institutet (6.0)	Swedish	0.089	0.142	0.0040

b MHC association for classical HLA alleles with $P < 1 \times 10^{-3}$

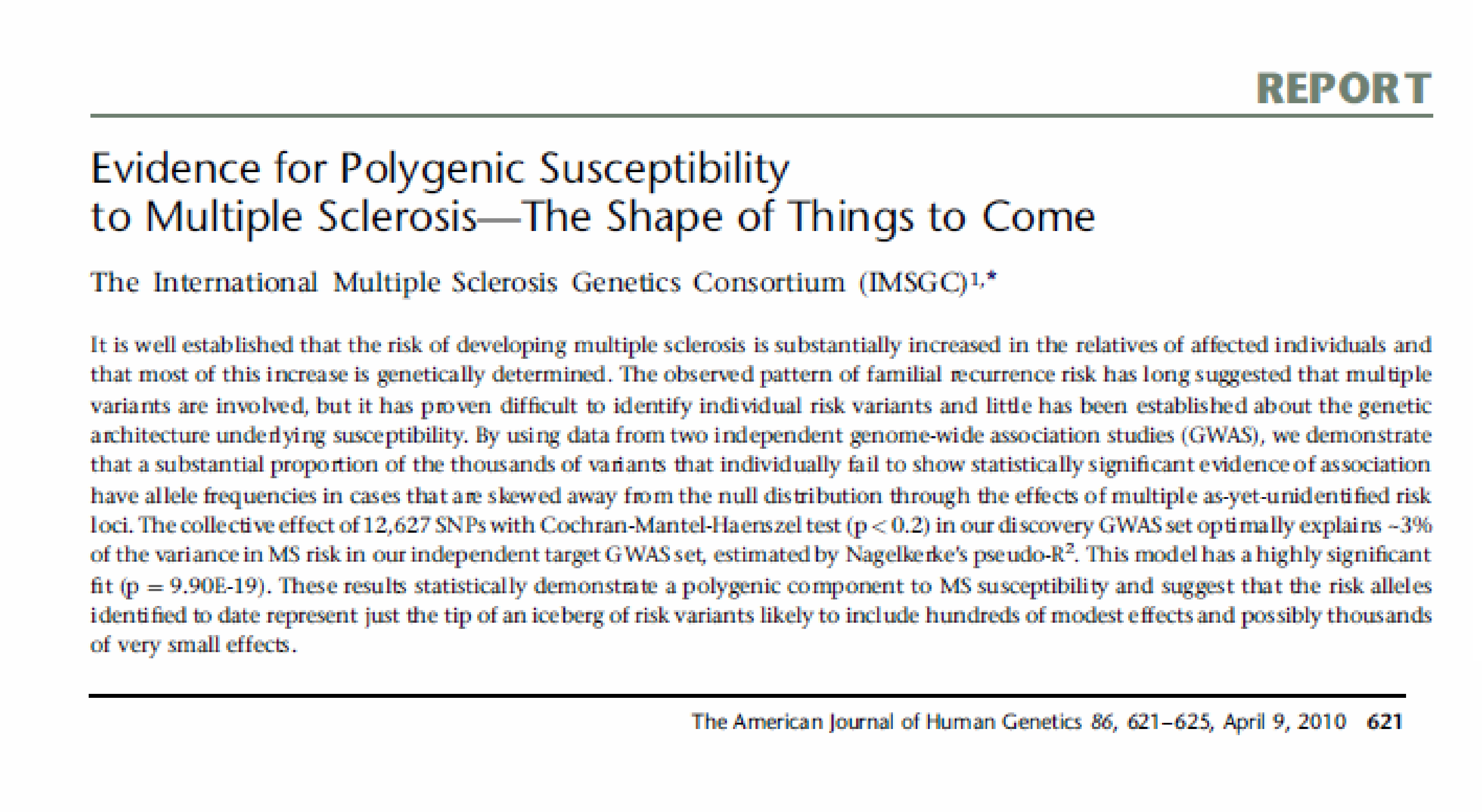
HLA allele	Frequency†	Odds ratio	*P* value
*HLA-A*0101*	0.103	0.785	4×10^{-5}
*HLA-C*0701*	0.113	0.778	5×10^{-5}
*HLA-B*0801*	0.068	0.757	3×10^{-5}
*HLA-DRB*0301*	0.121	0.768	3×10^{-6}
*HLA-DQB*0201*	0.210	0.857	4×10^{-4}
*HLA-DQA*0501*	0.205	0.798	6×10^{-7}

Total sample Cochran–Mantel–Haenszel $P = 4 \times 10^{-7}$; Breslow–Day heterogeneity test $P = 0.012$ (d.f. = 6).

* SNP failed genotyping quality control in UCL. Allele frequency for UCL based on imputed genotypes.

† Frequency is estimated population frequency.

*Lists of authors and their affiliations appear at the end of the paper.



REPORT

Evidence for Polygenic Susceptibility to Multiple Sclerosis—The Shape of Things to Come

The International Multiple Sclerosis Genetics Consortium (IMSGC)[1,*]

It is well established that the risk of developing multiple sclerosis is substantially increased in the relatives of affected individuals and that most of this increase is genetically determined. The observed pattern of familial recurrence risk has long suggested that multiple variants are involved, but it has proven difficult to identify individual risk variants and little has been established about the genetic architecture underlying susceptibility. By using data from two independent genome-wide association studies (GWAS), we demonstrate that a substantial proportion of the thousands of variants that individually fail to show statistically significant evidence of association have allele frequencies in cases that are skewed away from the null distribution through the effects of multiple as-yet-unidentified risk loci. The collective effect of 12,627 SNPs with Cochran-Mantel-Haenszel test ($p < 0.2$) in our discovery GWAS set optimally explains ~3% of the variance in MS risk in our independent target GWAS set, estimated by Nagelkerke's pseudo-R^2. This model has a highly significant fit ($p = 9.90E-19$). These results statistically demonstrate a polygenic component to MS susceptibility and suggest that the risk alleles identified to date represent just the tip of an iceberg of risk variants likely to include hundreds of modest effects and possibly thousands of very small effects.

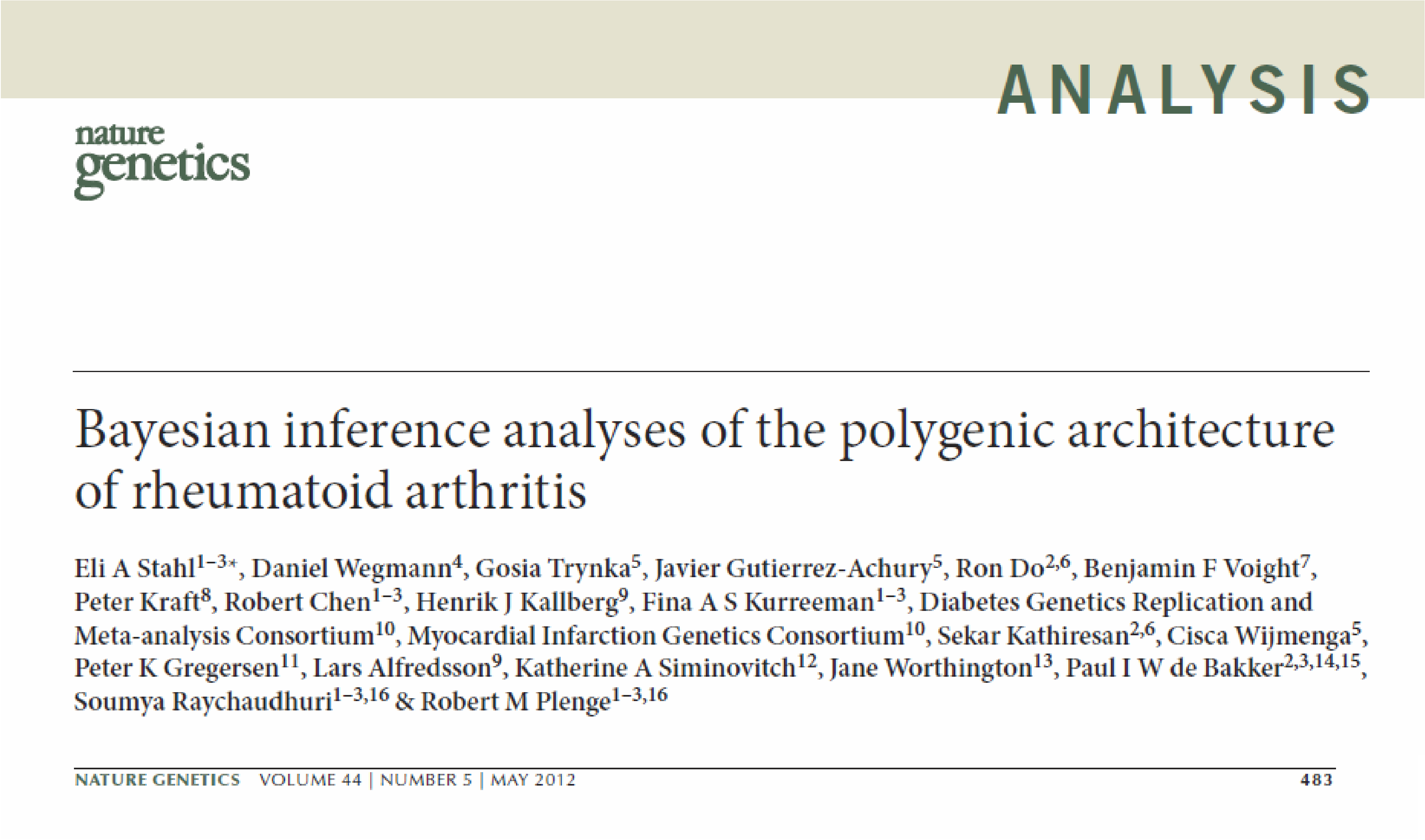
OMICS A Journal of Integrative Biology
Volume 15, Number 6, 2011
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2010.0090

Polygenic Modeling of Genome-Wide Association Studies: An Application to Prostate and Breast Cancer

John S. Witte and Thomas J. Hoffmann

Abstract

Genome-wide association studies (GWAS) have successfully detected and replicated associations with numerous diseases, including cancers of the prostate and breast. These findings are helping clarify the genomic basis of such diseases, but appear to explain little of disease heritability. This limitation might reflect the focus of conventional GWAS on a small set of the most statistically significant associations with disease. More information might be obtained by analyzing GWAS using a polygenic model, which allows for the possibility that thousands of genetic variants could impact disease. Furthermore, there may exist common polygenic effects between potentially related phenotypes (e.g., prostate and breast cancer). Here we present and apply a polygenic model to GWAS of prostate and breast cancer. Our results indicate that the polygenic model can explain an increasing—albeit low—amount of heritability for both of these cancers, even when excluding the most statistically significant associations. In addition, nonaggressive prostate cancer and breast cancer appear to share a common polygenic model, potentially reflecting a similar underlying biology. This supports the further development and application of polygenic models to genomic data.



ANALYSIS

nature genetics

Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis

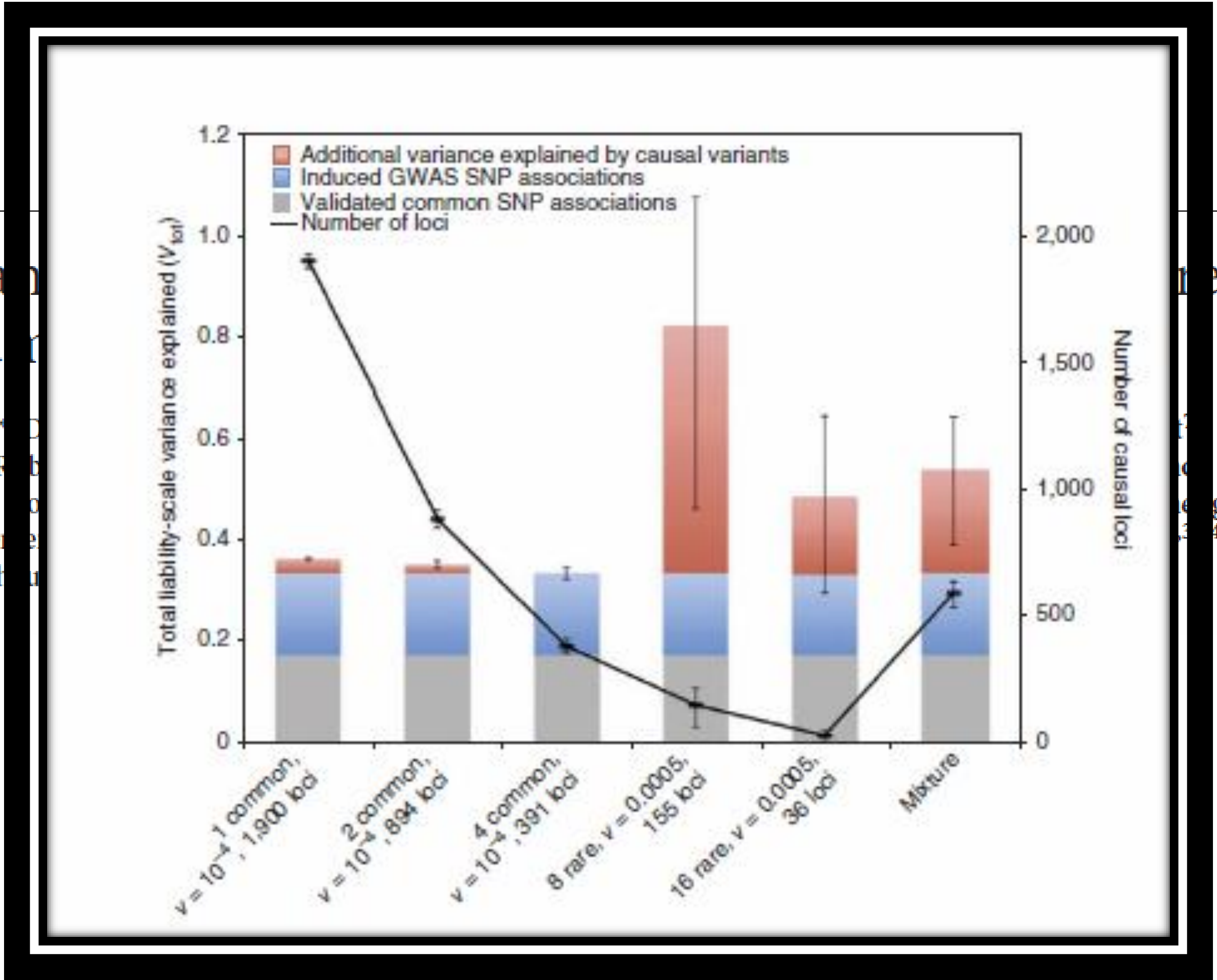
Eli A Stahl[1–3,*], Daniel Wegmann[4], Gosia Trynka[5], Javier Gutierrez-Achury[5], Ron Do[2,6], Benjamin F Voight[7], Peter Kraft[8], Robert Chen[1–3], Henrik J Kallberg[9], Fina A S Kurreeman[1–3], Diabetes Genetics Replication and Meta-analysis Consortium[10], Myocardial Infarction Genetics Consortium[10], Sekar Kathiresan[2,6], Cisca Wijmenga[5], Peter K Gregersen[11], Lars Alfredsson[9], Katherine A Siminovitch[12], Jane Worthington[13], Paul I W de Bakker[2,3,14,15], Soumya Raychaudhuri[1–3,16] & Robert M Plenge[1–3,16]

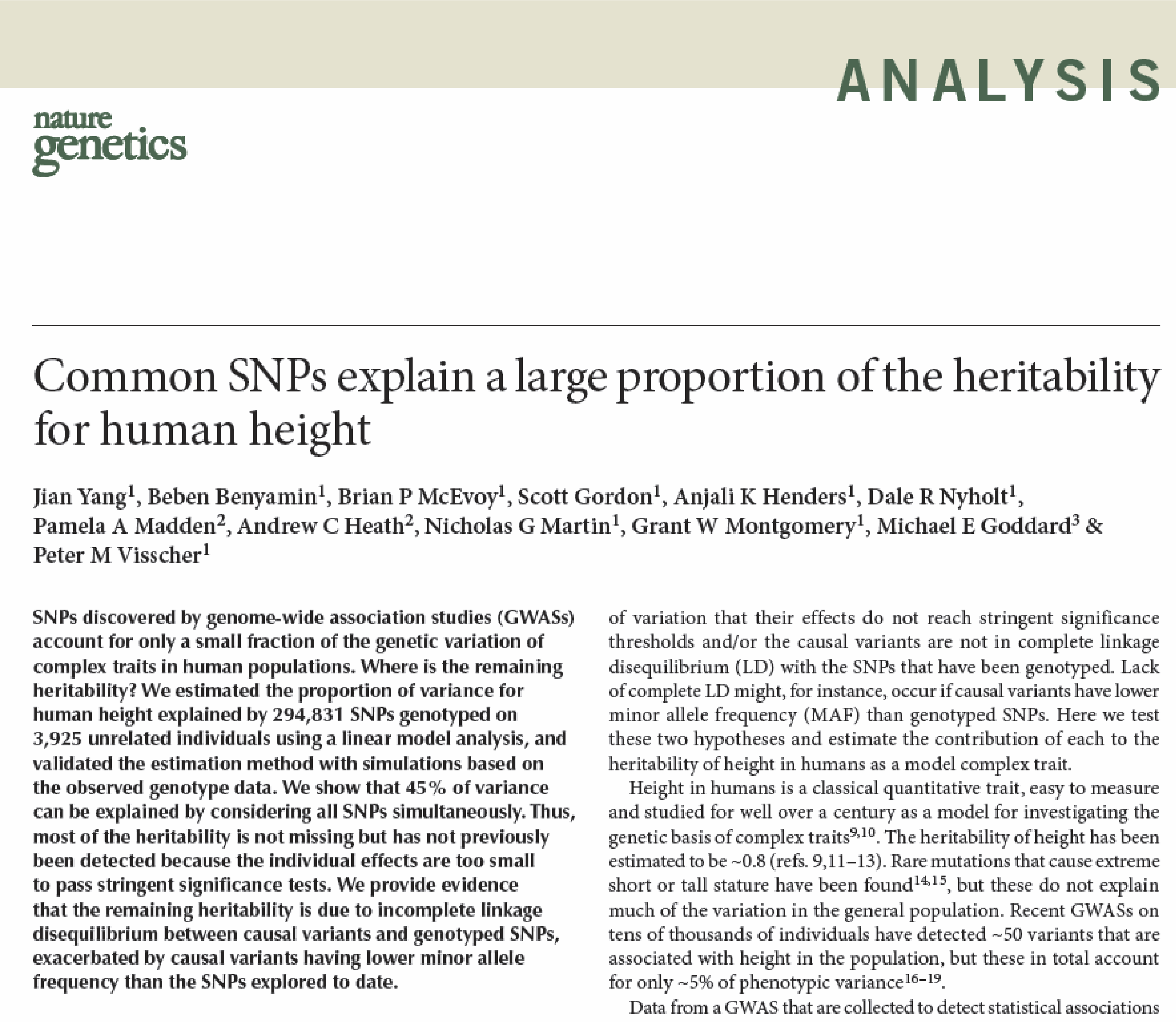
ANALYSIS

nature genetics

Bayesia re
# of rheu

Eli A Stahl[1–3]
Peter Kraft[8], R
Meta-analysis ga[5],
Peter K Grege ,4,15,
Soumya Raych





ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

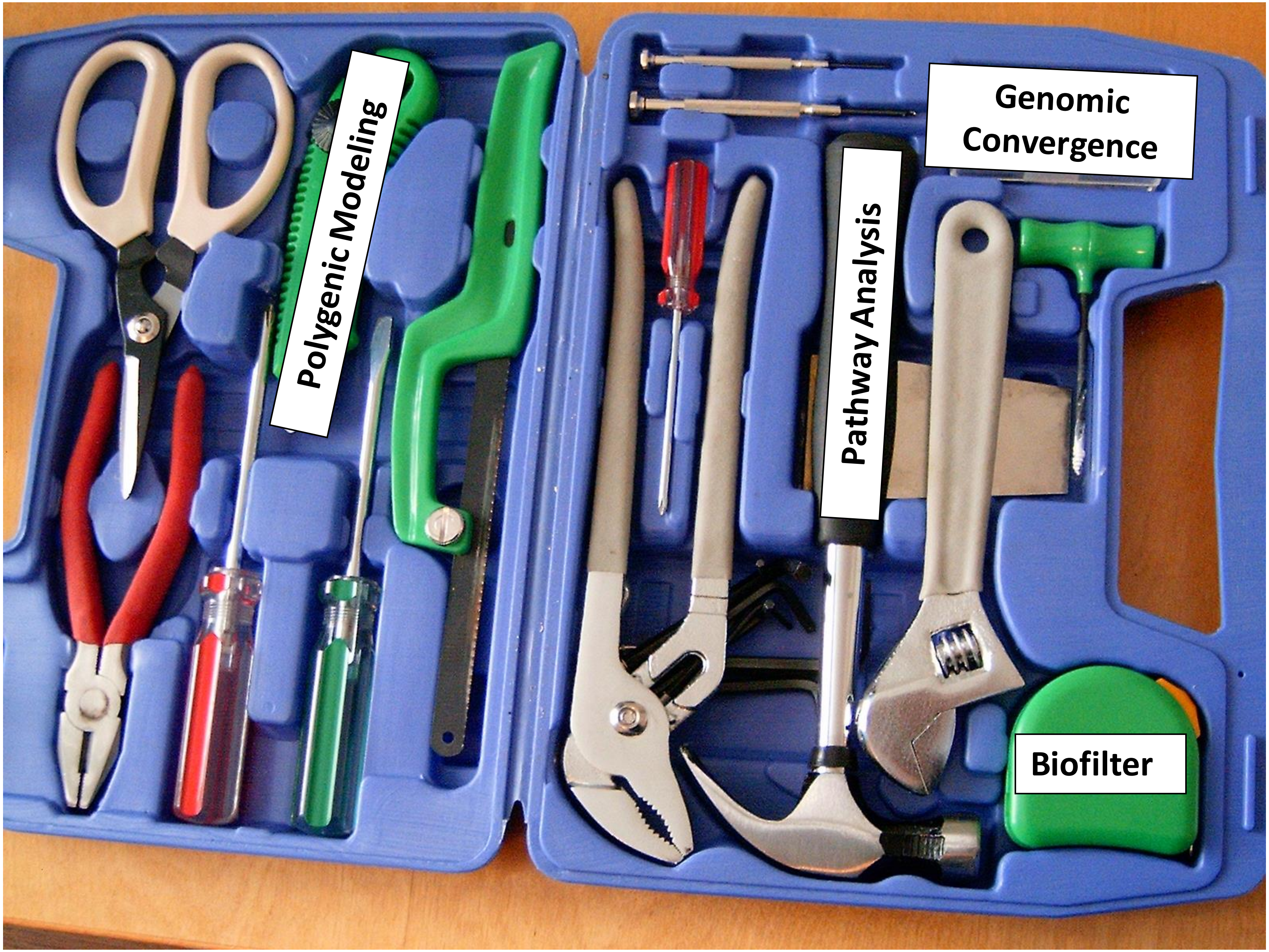
Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

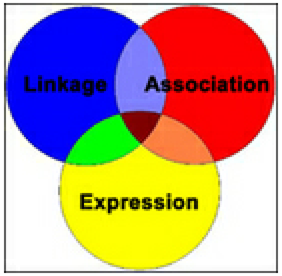
of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for instance, occur if causal variants have lower minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of each to the heritability of height in humans as a model complex trait.

Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits[9,10]. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance[16–19].

Data from a GWAS that are collected to detect statistical associations

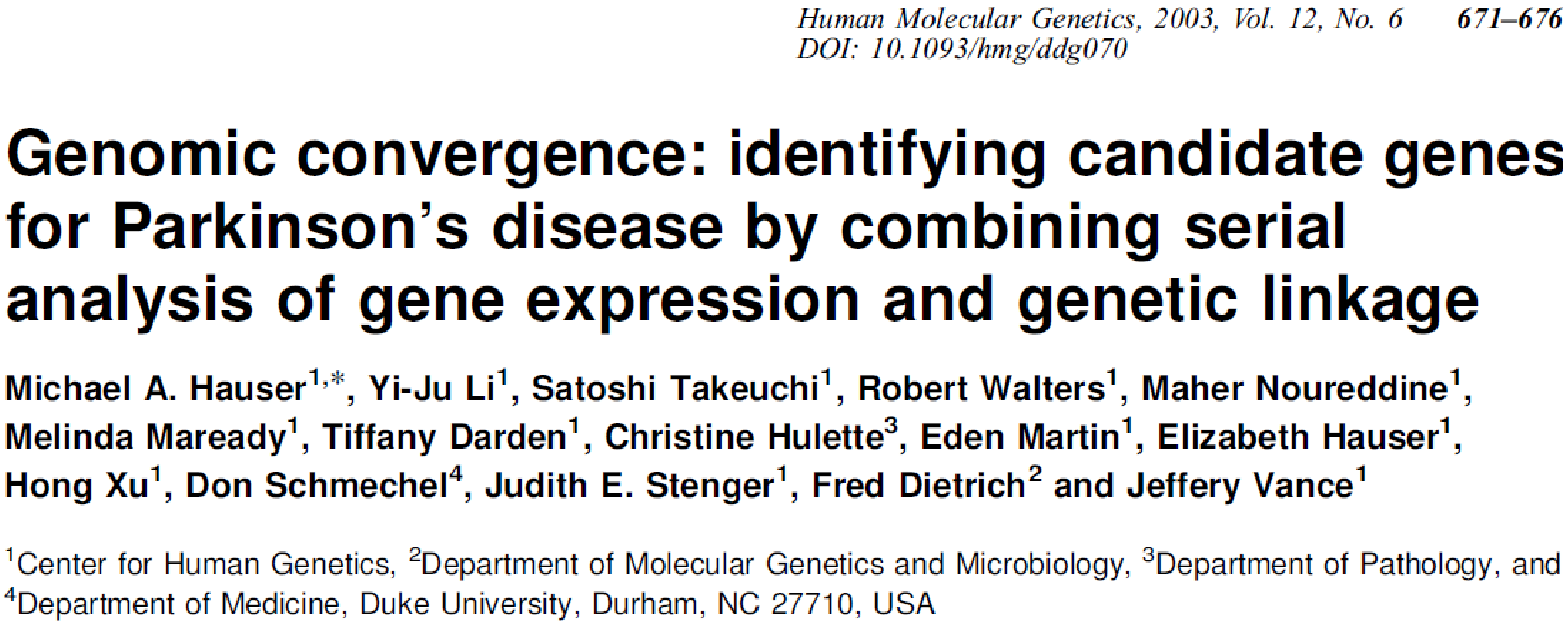

Polygenic Modeling
Genomic Convergence
Pathway Analysis
Biofilter

Genomic Convergence



- Multifactor approach that combines different kinds of genetic data
- Identify and prioritize susceptibility genes for complex traits
- Assumption
  - Regions of the genome that harbor susceptibility genes will show evidence of linkage, association, and/or differential gene expression

Genomic Convergence



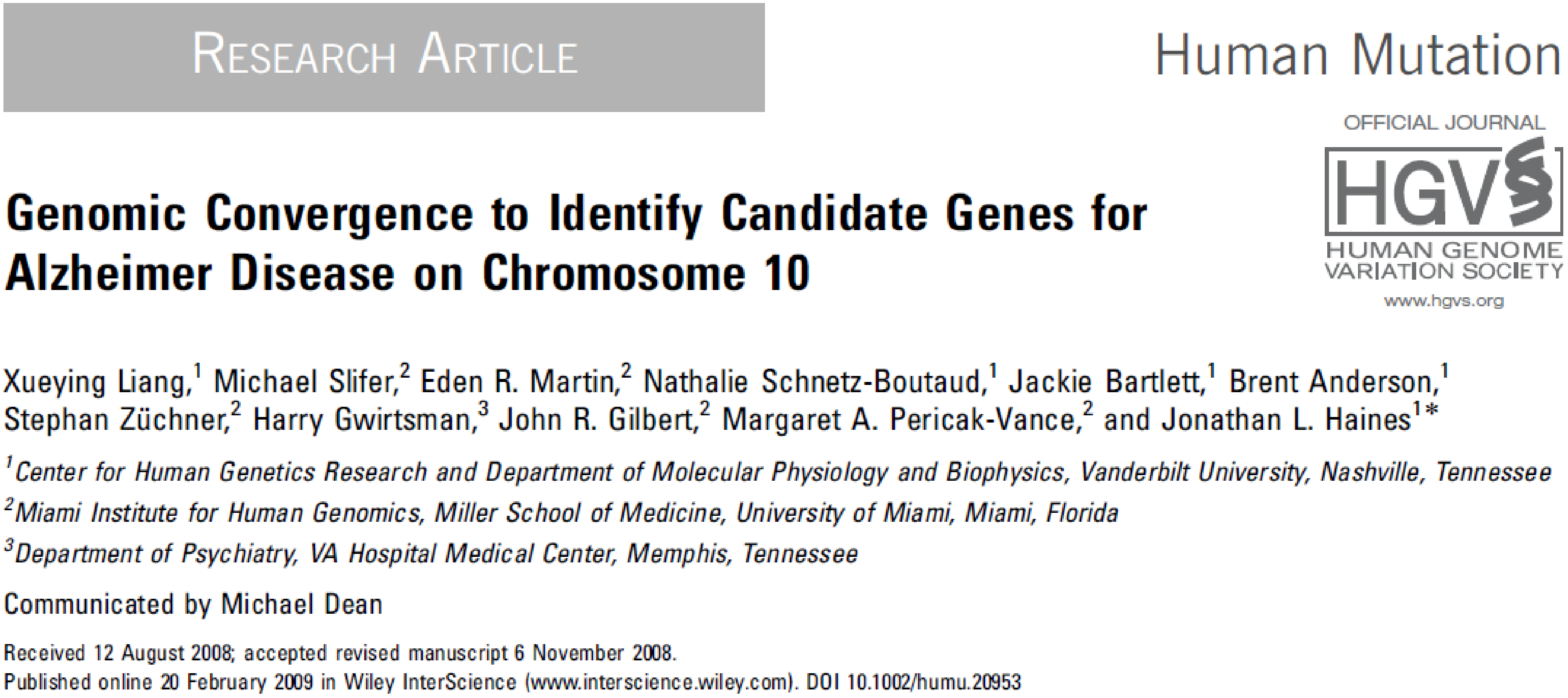
Human Molecular Genetics, 2003, Vol. 12, No. 6 ***671–676***
DOI: 10.1093/hmg/ddg070

Genomic convergence: identifying candidate genes for Parkinson’s disease by combining serial analysis of gene expression and genetic linkage

Michael A. Hauser[1,*], Yi-Ju Li[1], Satoshi Takeuchi[1], Robert Walters[1], Maher Noureddine[1], Melinda Maready[1], Tiffany Darden[1], Christine Hulette[3], Eden Martin[1], Elizabeth Hauser[1], Hong Xu[1], Don Schmechel[4], Judith E. Stenger[1], Fred Dietrich[2] and Jeffery Vance[1]

[1]Center for Human Genetics, [2]Department of Molecular Genetics and Microbiology, [3]Department of Pathology, and [4]Department of Medicine, Duke University, Durham, NC 27710, USA

Genomic Convergence



RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Genomic Convergence to Identify Candidate Genes for Alzheimer Disease on Chromosome 10

Xueying Liang,[1] Michael Slifer,[2] Eden R. Martin,[2] Nathalie Schnetz-Boutaud,[1] Jackie Bartlett,[1] Brent Anderson,[1] Stephan Züchner,[2] Harry Gwirtsman,[3] John R. Gilbert,[2] Margaret A. Pericak-Vance,[2] and Jonathan L. Haines[1]*

[1] *Center for Human Genetics Research and Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee*

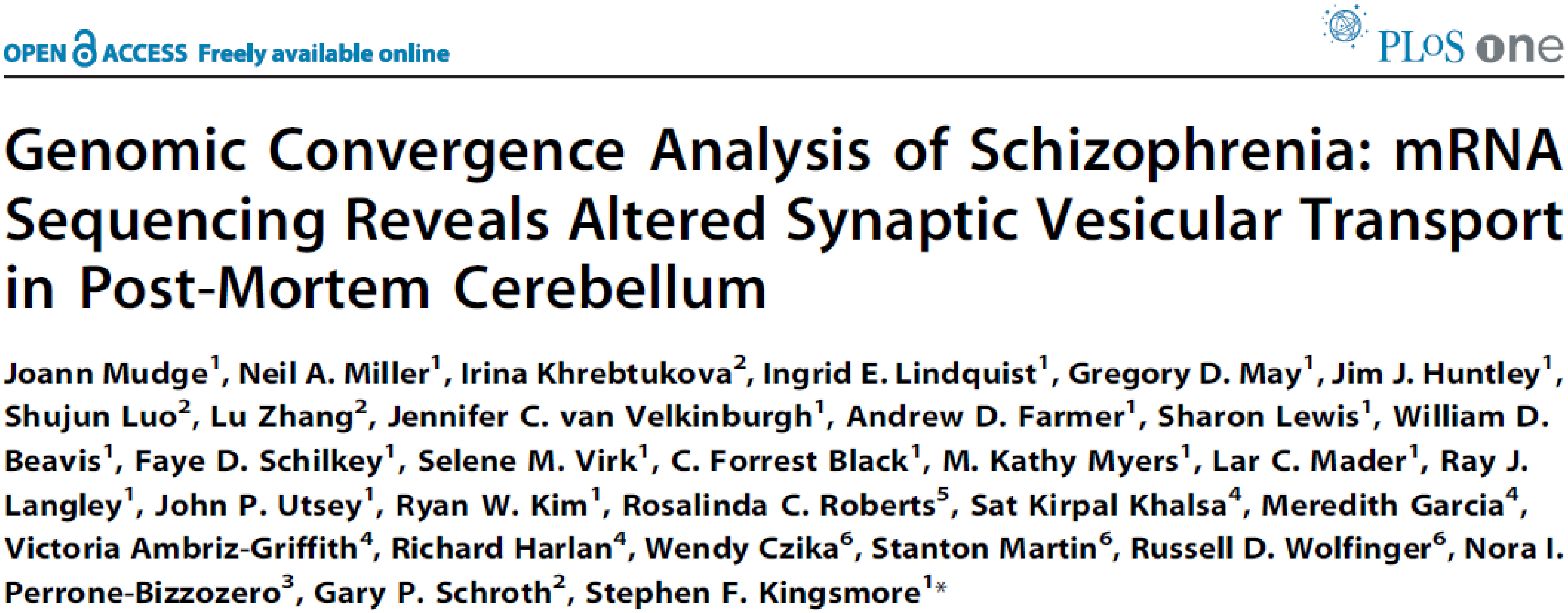
[2] *Miami Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, Florida*

[3] *Department of Psychiatry, VA Hospital Medical Center, Memphis, Tennessee*

Communicated by Michael Dean

Received 12 August 2008; accepted revised manuscript 6 November 2008.
Published online 20 February 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.20953

Genomic Convergence

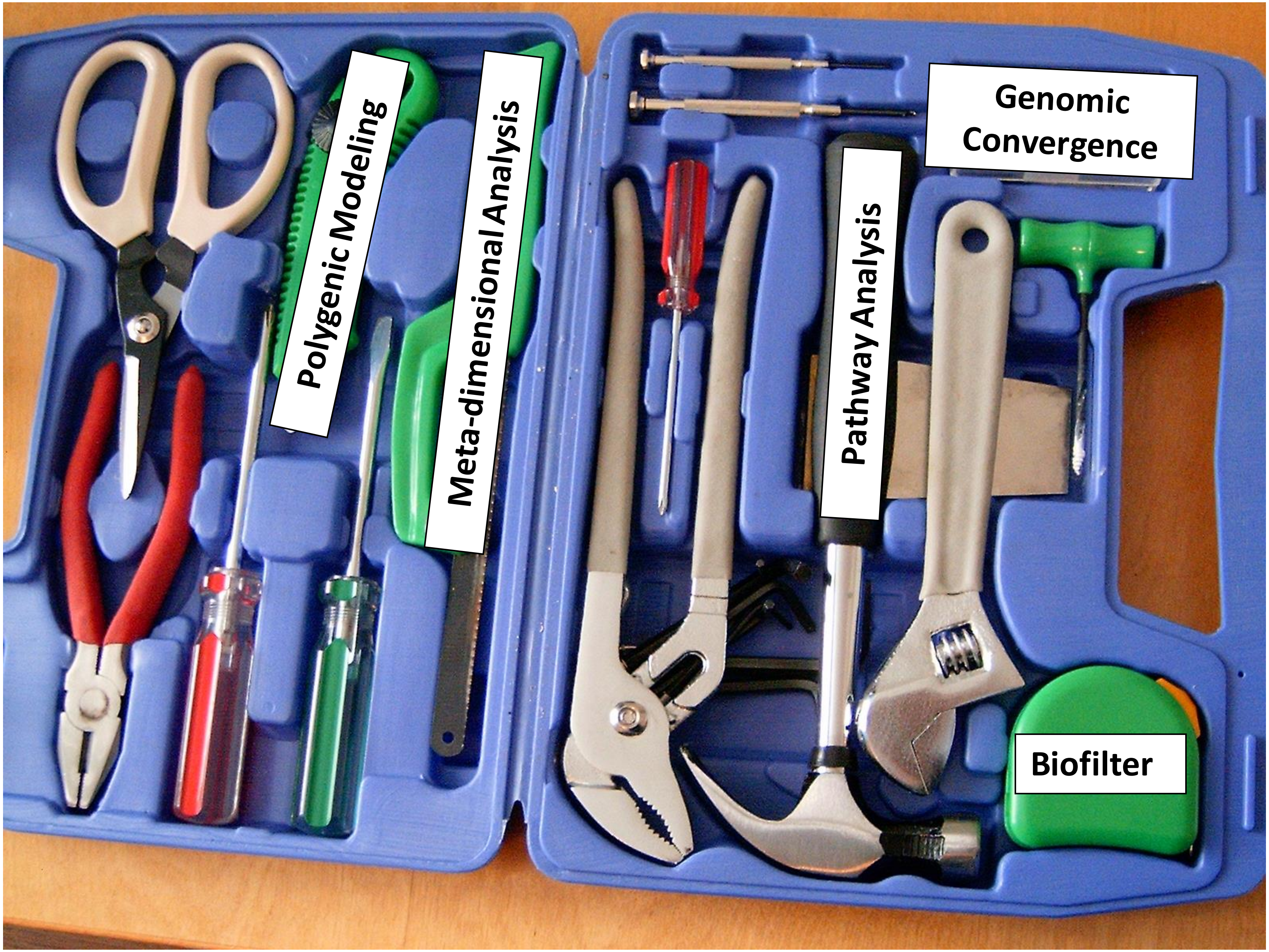


OPEN ACCESS Freely available online

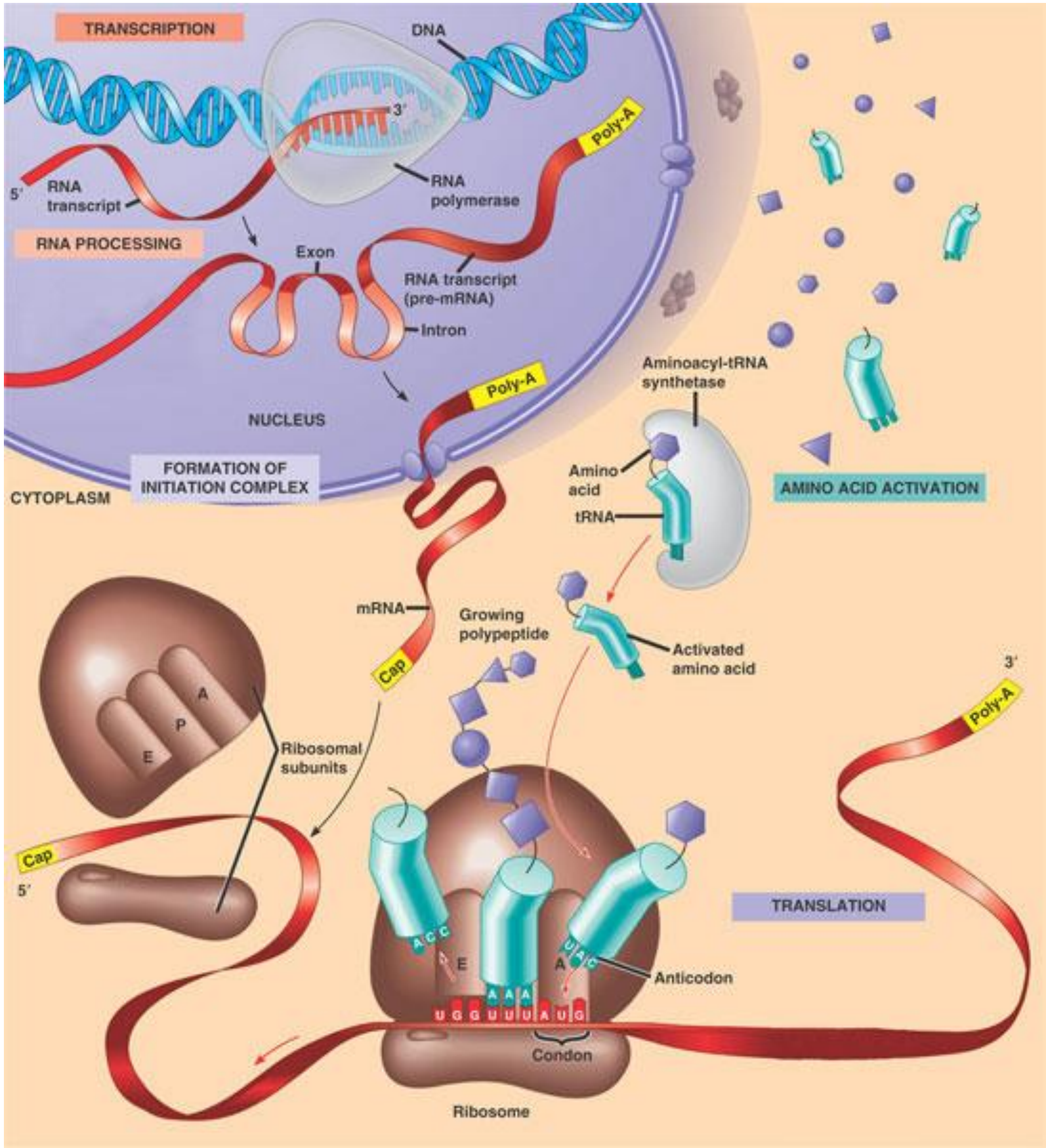
PLoS one

Genomic Convergence Analysis of Schizophrenia: mRNA Sequencing Reveals Altered Synaptic Vesicular Transport in Post-Mortem Cerebellum

Joann Mudge[1], Neil A. Miller[1], Irina Khrebtukova[2], Ingrid E. Lindquist[1], Gregory D. May[1], Jim J. Huntley[1], Shujun Luo[2], Lu Zhang[2], Jennifer C. van Velkinburgh[1], Andrew D. Farmer[1], Sharon Lewis[1], William D. Beavis[1], Faye D. Schilkey[1], Selene M. Virk[1], C. Forrest Black[1], M. Kathy Myers[1], Lar C. Mader[1], Ray J. Langley[1], John P. Utsey[1], Ryan W. Kim[1], Rosalinda C. Roberts[5], Sat Kirpal Khalsa[4], Meredith Garcia[4], Victoria Ambriz-Griffith[4], Richard Harlan[4], Wendy Czika[6], Stanton Martin[6], Russell D. Wolfinger[6], Nora I. Perrone-Bizzozero[3], Gary P. Schroth[2], Stephen F. Kingsmore[1]*


Polygenic Modeling
Meta-dimensional Analysis
Genomic Convergence
Pathway Analysis
Biofilter

Molecular biology is complex

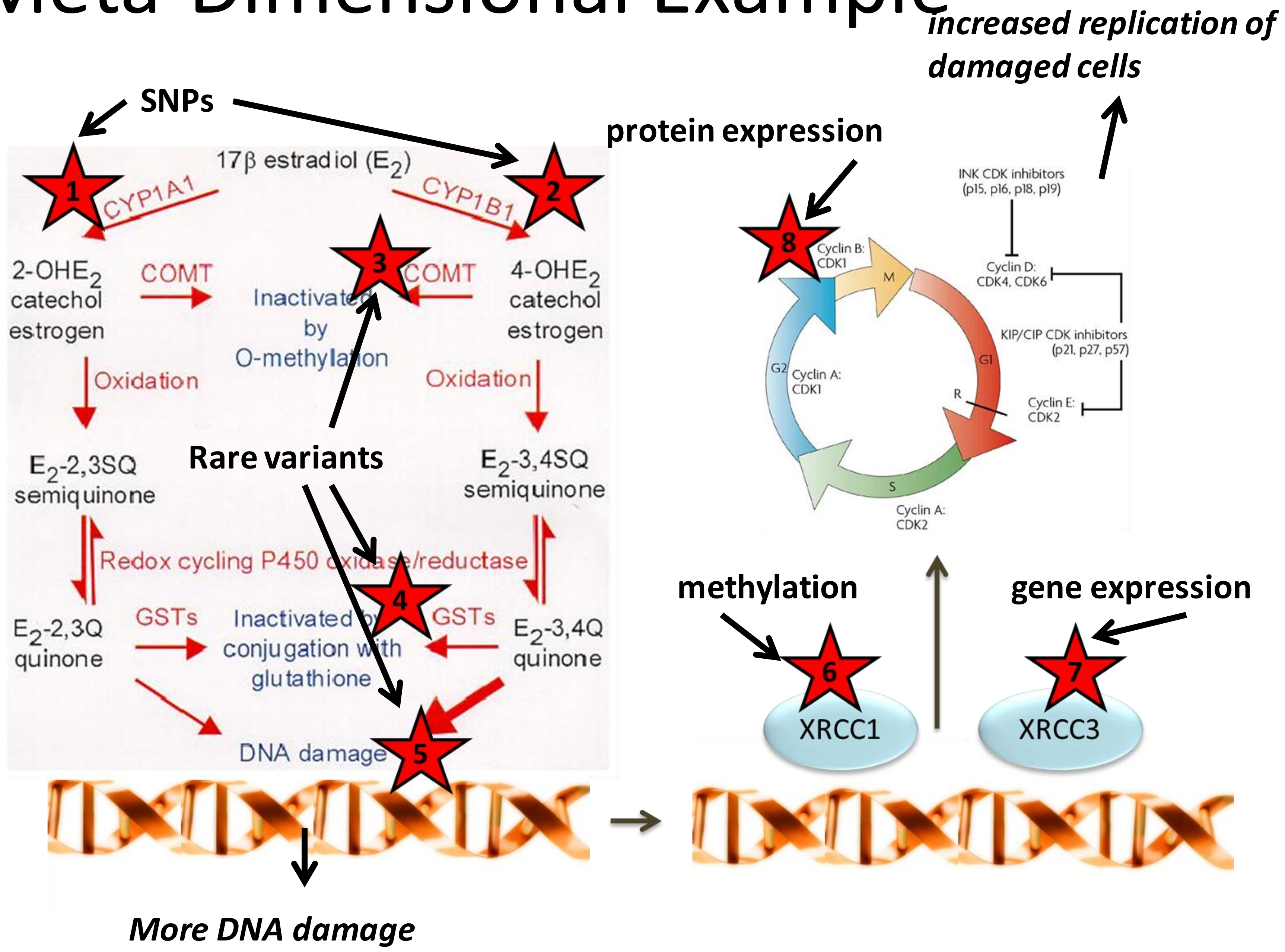


Meta-dimensional

- Meta- (from Greek: μετά = "after", "beyond", "with", "adjacent", "self") to indicate a concept which is an abstraction from another concept

- Meta-dimensional analysis of phenotypes
  - Abstracting from multiple data source
  - Abstracting from multiple data types
  - Abstracting from multiple data sets

Meta-Dimensional Example



SNPs

Rare variants

protein expression

methylation

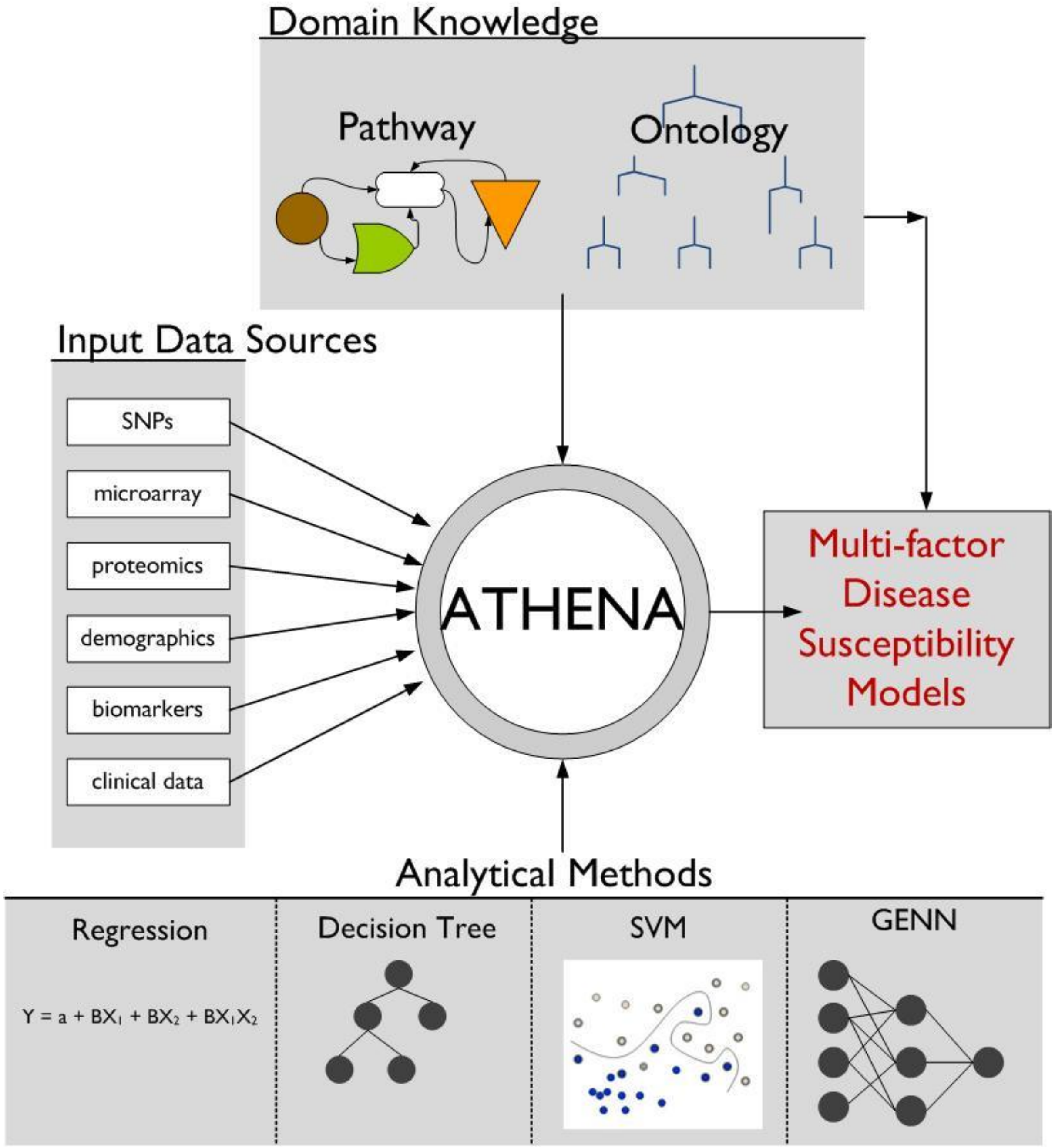
gene expression

increased replication of damaged cells

More DNA damage

ATHENA

- Analysis Tool for Heritable and Environmental Network Associations
  - Integrate genetic, environmental, and prior biological knowledge
  - Thorough data analysis
  - Combination of categorical and continuous independent and dependent variables



Domain Knowledge

Pathway

Ontology

Input Data Sources

SNPs

microarray

proteomics

demographics

biomarkers

clinical data

ATHENA

Multi-factor Disease Susceptibility Models

Analytical Methods

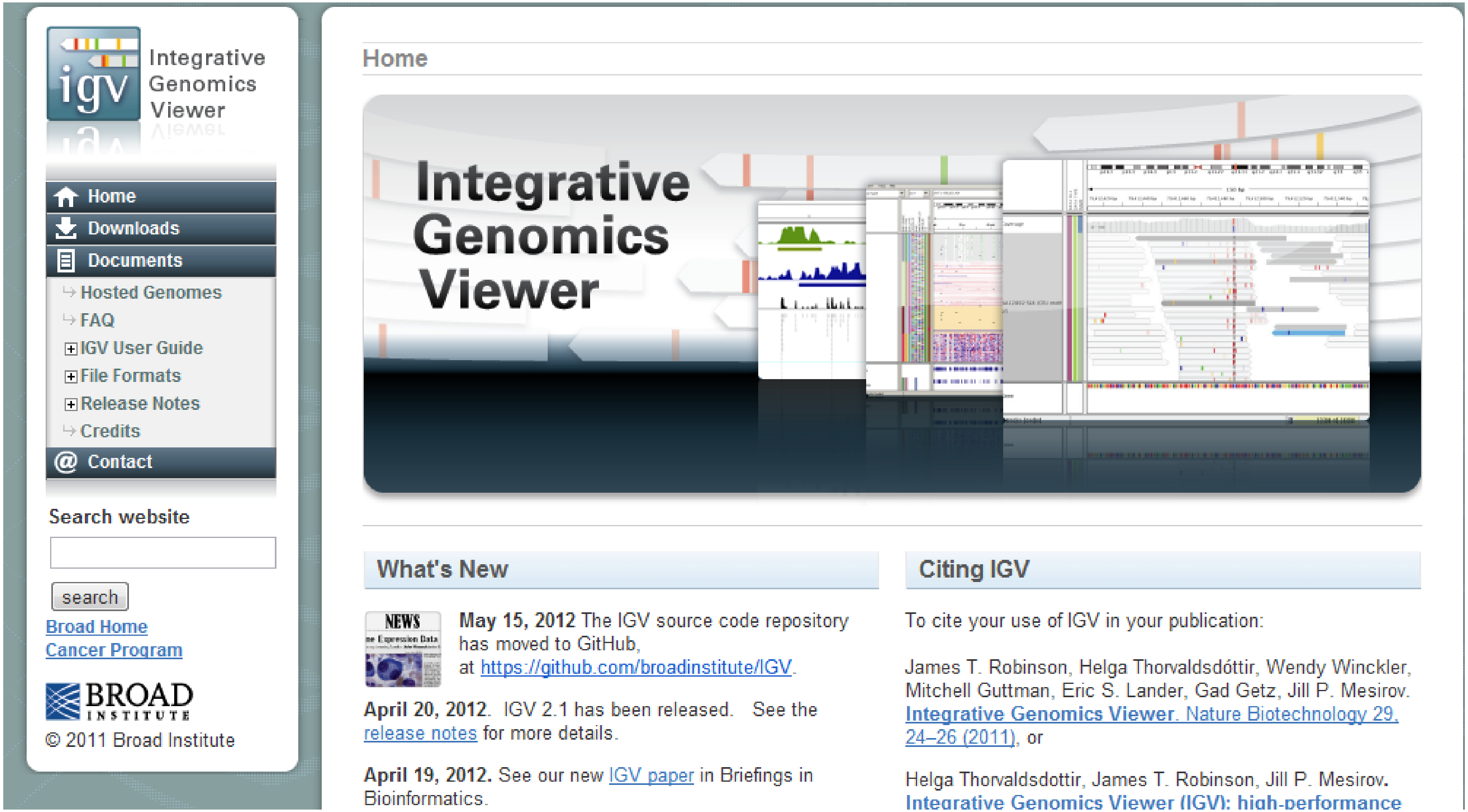
Regression

$Y = a + BX_1 + BX_2 + BX_1X_2$

Decision Tree

SVM

GENN



© 2011 Broad Institute

Home

What's New

May 15, 2012 The IGV source code repository has moved to GitHub, at https://github.com/broadinstitute/IGV.

April 20, 2012. IGV 2.1 has been released. See the release notes for more details.

April 19, 2012. See our new IGV paper in Briefings in Bioinformatics.

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer.** Nature Biotechnology 29, 24–26 (2011), or

Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance**

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

Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies

The current paradigm of human genetics research is to analyze variation of a single data type (i.e., DNA sequence or RNA levels) to detect genes and pathways that underlie complex traits such as disease state or drug response. While these studies have detected thousands of variations that associate with hundreds of complex phenotypes, much of the estimated heritability, or trait variability due to genetic factors, remain unexplained. We may be able to account for a portion of the missing heritability if we incorporate a systems biology approach into these analyses. Rapid technological advances will make it possible for scientists to explore this hypothesis via the generation of high-throughput omics data – transcriptomic, proteomic and methylomic to name a few. Analyzing this 'meta-dimensional' data will require clever statistical techniques that allow for the integration of qualitative and quantitative predictor variables. For this article, we examine two major categories of approaches for integrated data analysis, give examples of their use in experimental and *in silico* datasets, and assess the limitations of each method.

KEYWORDS: computational methods ▪ data integration ▪ pharmacogenomics ▪ systems biology

Emily R Holzinger[1,2]
& Marylyn D Ritchie*[2]

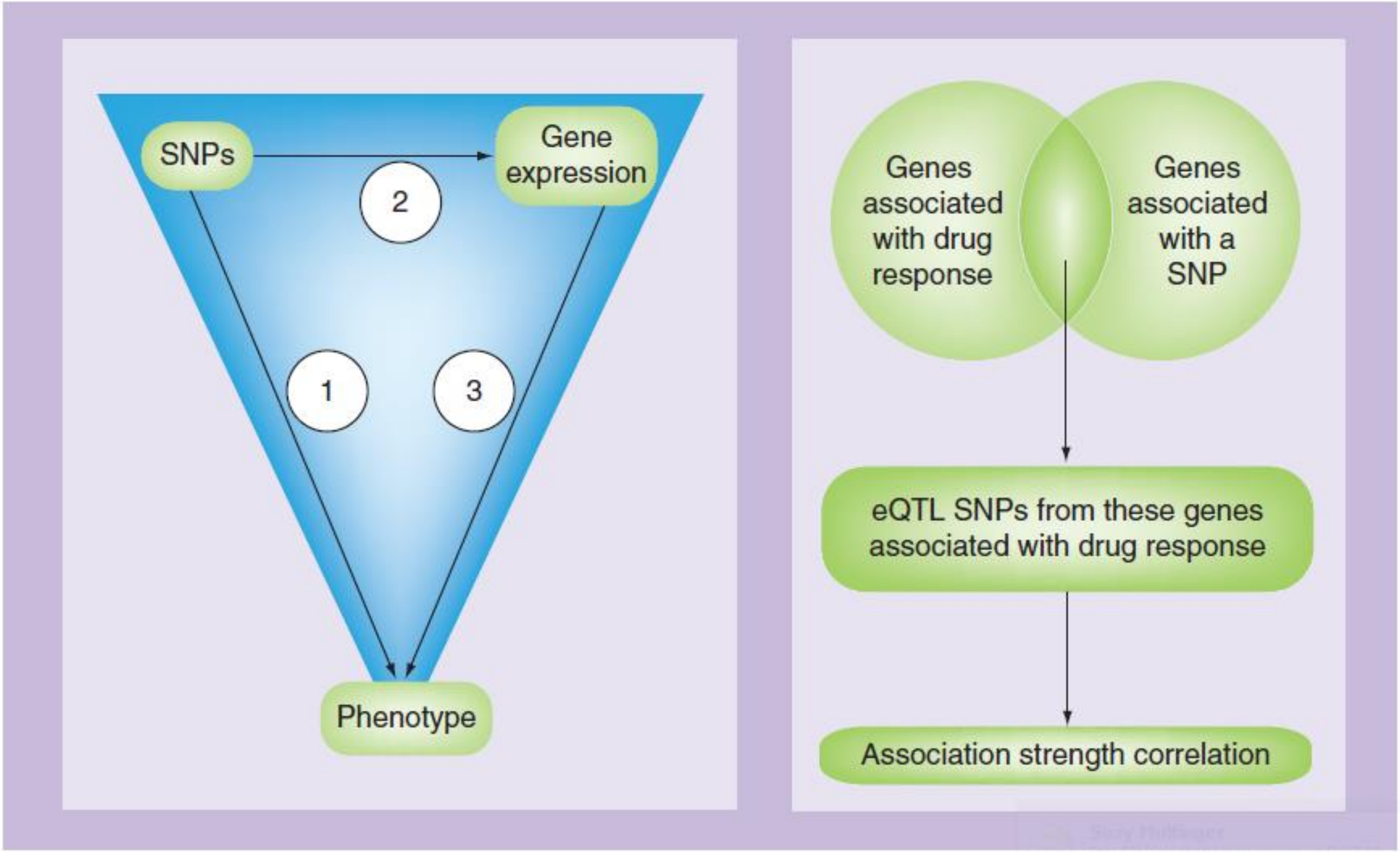

SNPs
Gene expression
2
1
3
Phenotype
Genes associated with drug response
Genes associated with a SNP
eQTL SNPs from these genes associated with drug response
Association strength correlation

Figure 1. Variations of the triangle method.
eQTL: Expression quantitative trait loci.

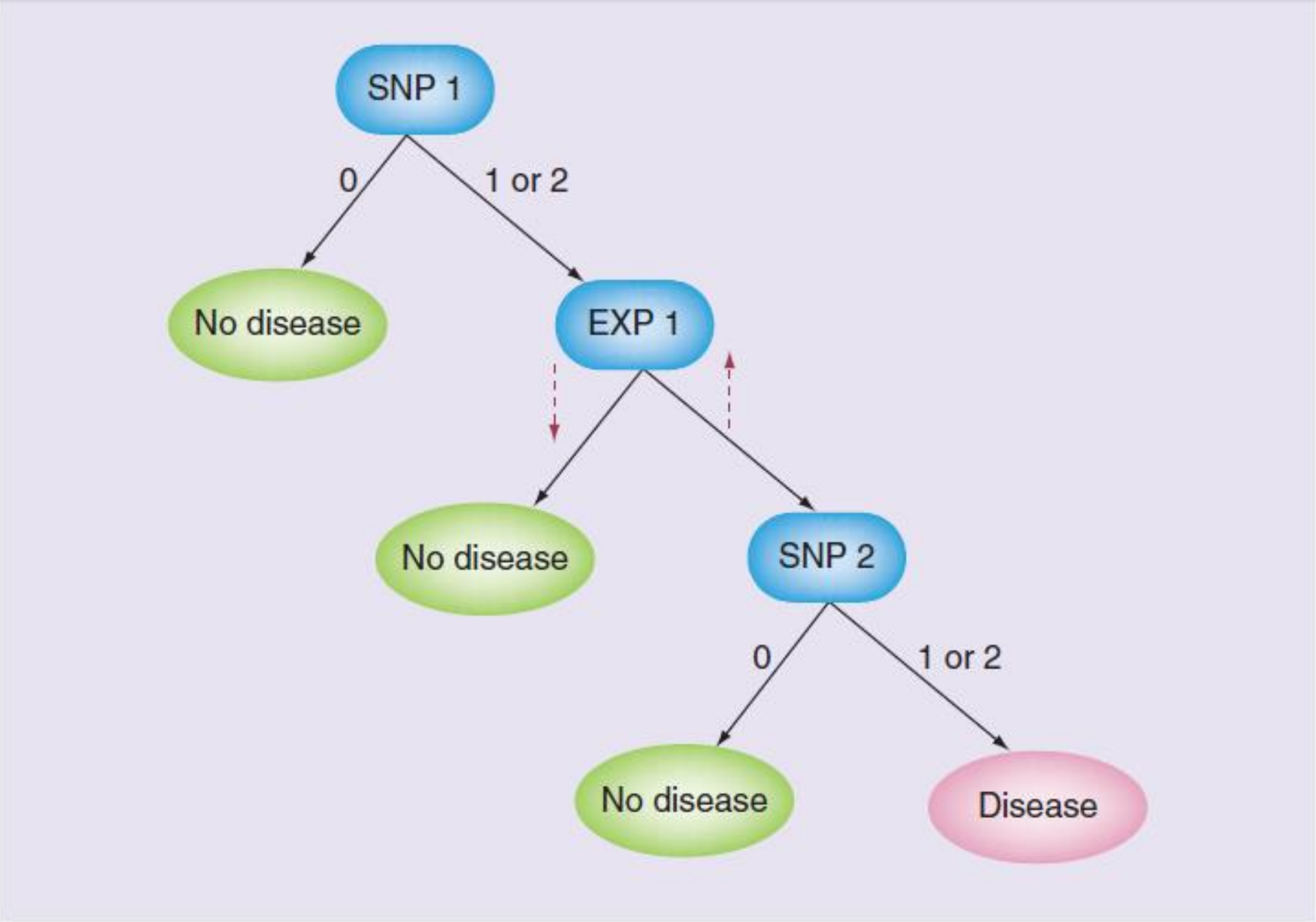


Figure 2. Decision tree example. For the SNP variables, the genotypes are represented as: 0: no minor alleles; 1: one minor allele; and 2: two minor alleles. The up and down dashed arrows indicate increased and decreased gene expression, respectively.

EXP: Gene expression.

Pharmacogenomics (2012) **13**(2), 213–222

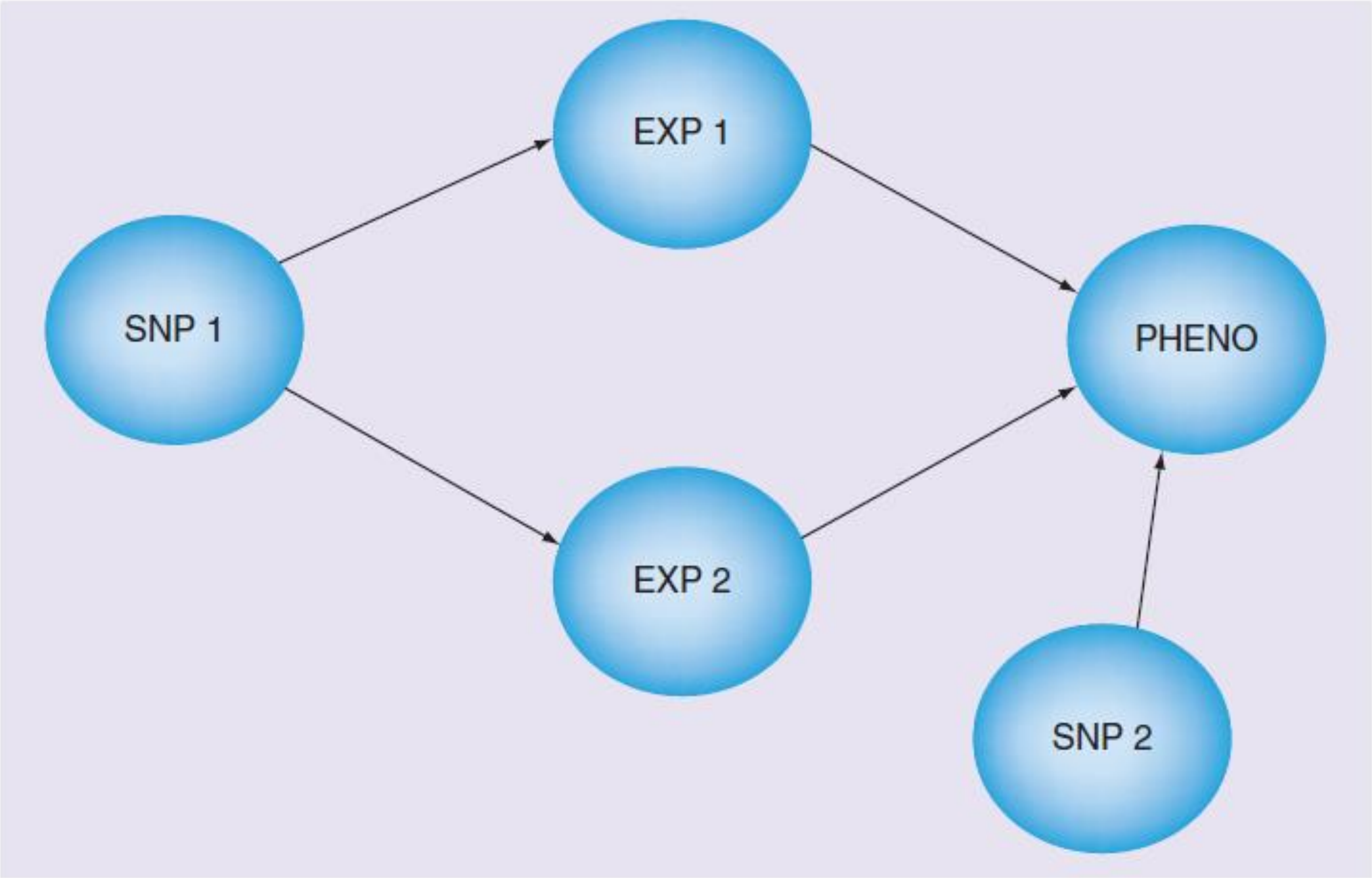


Figure 3. Bayesian network example with direct and indirect effects.
EXP: Gene expression; PHENO: Phenotype.

PERSONALIZED MEDICINE

NEW: Network-Enabled Wisdom in Biology, Medicine, and Health Care

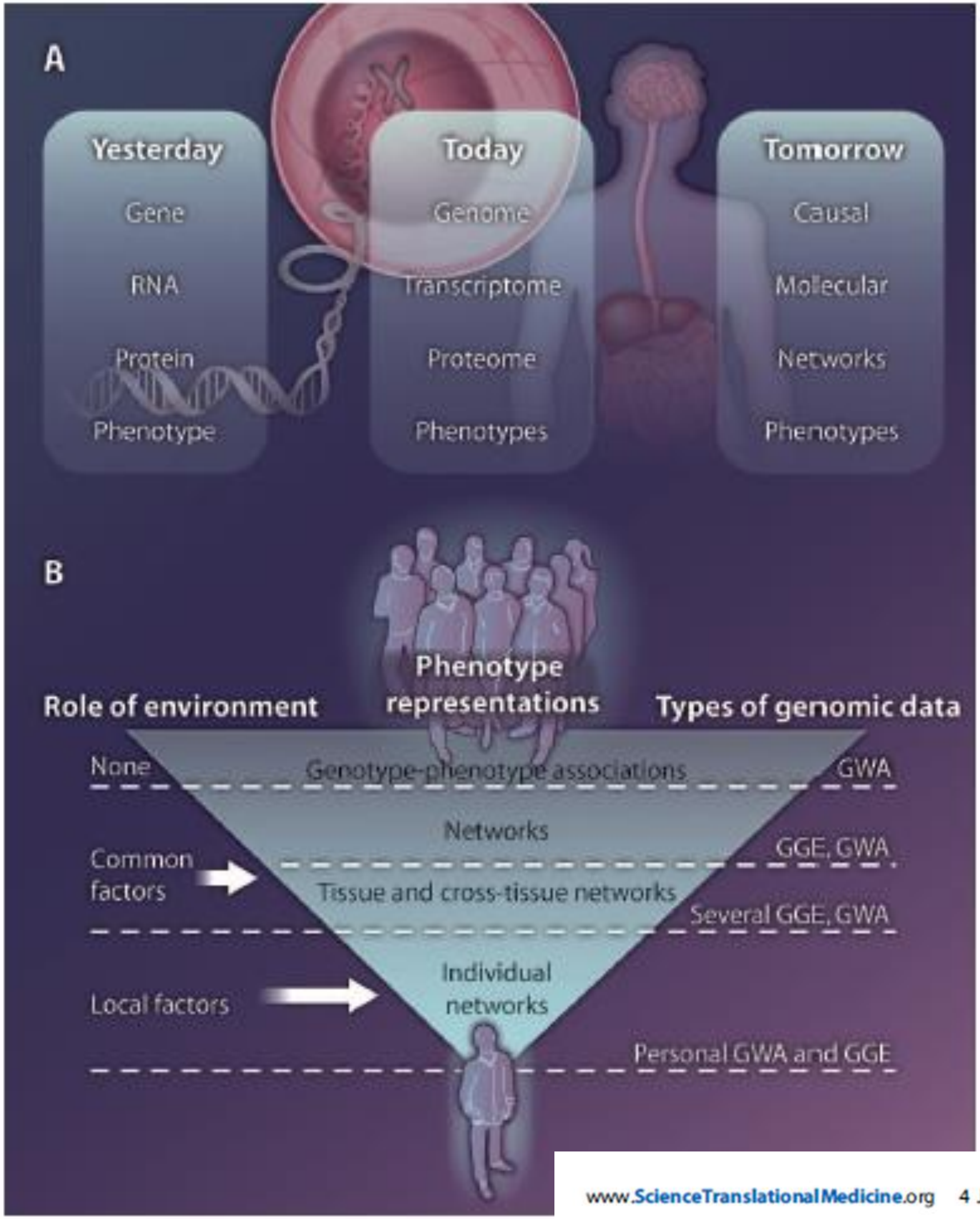
Eric E. Schadt[1] and Johan L. M. Björkegren[2,3,4]*

Complete repertoires of molecular activity in and between tissues provided by new high-dimensional "omics" technologies hold great promise for characterizing human physiology at all levels of biological hierarchies. The combined effects of genetic and environmental perturbations at any level of these hierarchies can lead to vicious cycles of pathology and complex systemic diseases. The challenge lies in extracting all relevant information from the rapidly increasing volumes of omics data and translating this information first into knowledge and ultimately into wisdom that can yield clinically actionable results. Here, we discuss how molecular networks are central to the implementation of this new biology in medicine and translation to preventive and personalized health care.

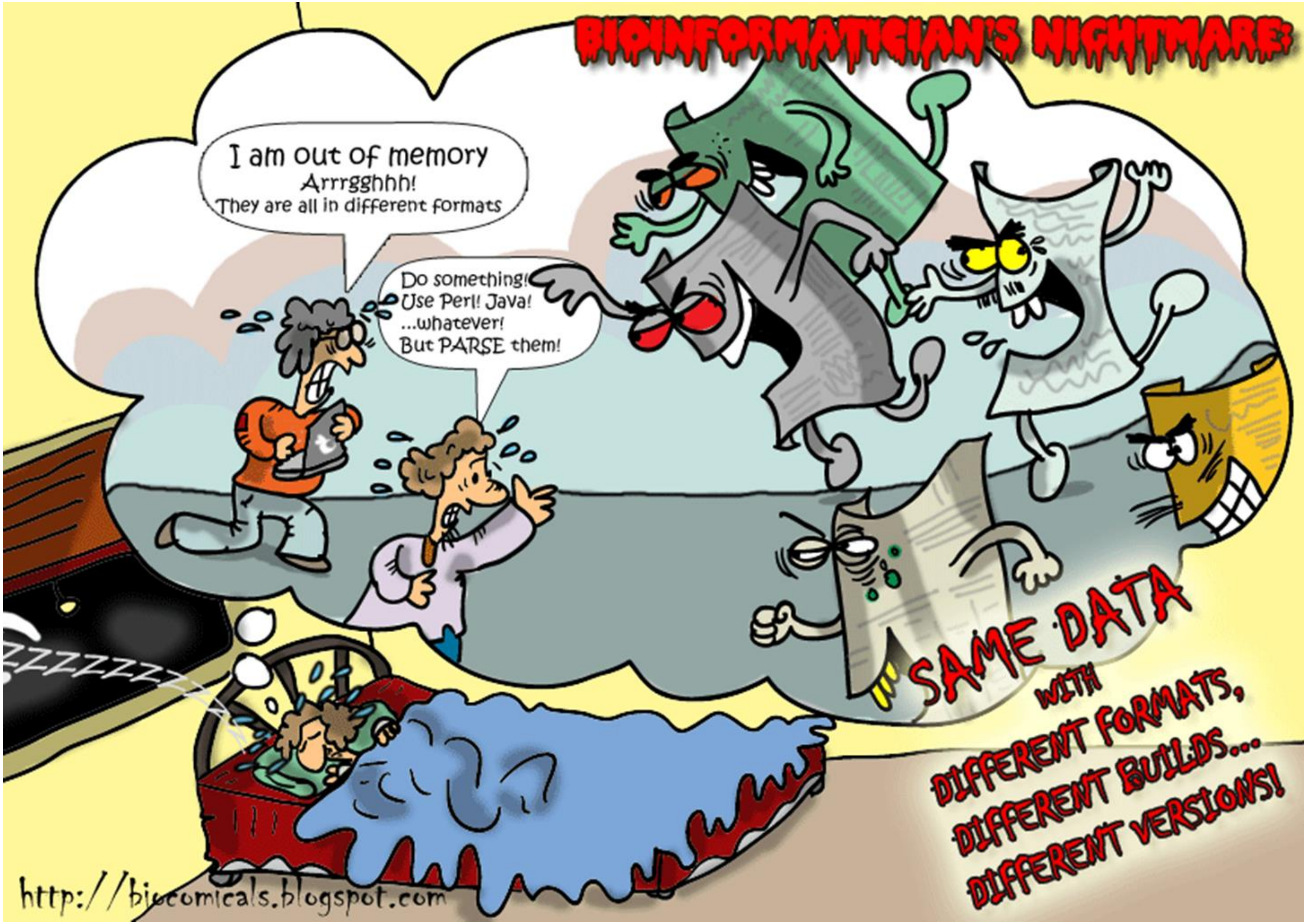
INTRODUCTION

Next-generation technologies that routinely measure biological parameters on a genome-wide scale ("omics" data)—such as DNA variations and epigenetic modifications, RNA and protein concentrations, and a variety of metabolites—are continuously being refined and offered at ever-decreasing costs. The resulting oceans of molecular data (moving quickly from the petabyte to exabyte scale or, even more scary, zetabyte—that's 21 zeros) cannot be deciphered with traditional mathematical analyses carried out on isolated computers. Nor is the traditional representation of biological processes as linear pathways sufficient to represent the hierarchy of levels of molecular and higher-order regulation, and the interplay that defines human physiology and individuals and their environment in ways that affect disease [questions remain as to the meaning of the disease associations observed in social networks (*1*)]. The architecture of biological networks shares similarities with well-studied ones in other disciplines, such as social and transportation networks. Like these large-scale information networks, molecular networks in biology are sparse and follow a power-law distribution in which most nodes have few interactions (say, one to three), whereas a smaller number, referred to as hub nodes, have many interactions (tens to hundreds or even thousands) (*2*) (Fig. 1).

Mapping the connectivity structure of networks (that is, the topology) is crucial for understanding how biological processes are defined at the molecular level, how they can be disrupted to cause disease, and how we


A
Yesterday
Gene
RNA
Protein
Phenotype
Today
Genome
Transcriptome
Proteome
Phenotypes
Tomorrow
Causal
Molecular
Networks
Phenotypes
B
Phenotype representations
Role of environment
Types of genomic data
None
Genotype-phenotype associations
GWA
Networks
GGE, GWA
Common factors
Tissue and cross-tissue networks
Several GGE, GWA
Local factors
Individual networks
Personal GWA and GGE

Summary


BIOINFORMATICIAN'S NIGHTMARE:
I am out of memory
Arrrgghhh!
They are all in different formats
Do something!
Use Perl! Java!
...whatever!
But PARSE them!
ZZZZZZZZZ
SAME DATA WITH DIFFERENT FORMATS, DIFFERENT BUILDS... DIFFERENT VERSIONS!
http://biocomicals.blogspot.com

Summary

Party with your
Data

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