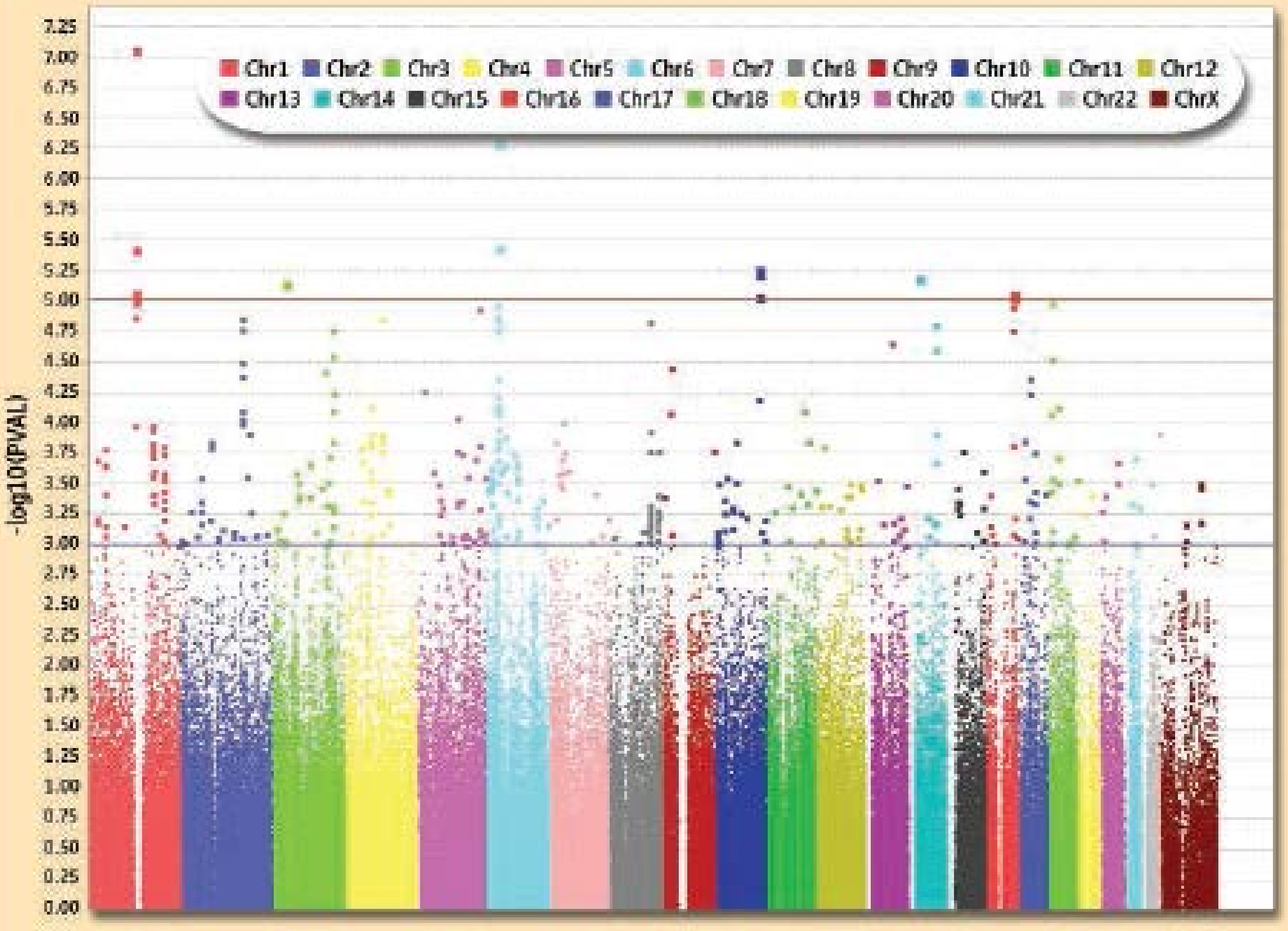
Statistical Methods for Next Generation Sequencing Data

Nicholas J. Schork, Ph.D.
J. Craig Venter Institute, La Jolla, CA &
The University of California, San Diego, La Jolla, CA

1. Background: The limits of the contemporary GWAS
2. Analysis of rare variants in sequencing studies
3. Predicting the functional effect of variants
4. Population genetic analysis of rare variants
5. The human ‘diplome’ and the need to phase
6. ‘Filtering’ strategies for identifying causal variants

J. Craig Venter™
INSTITUTE

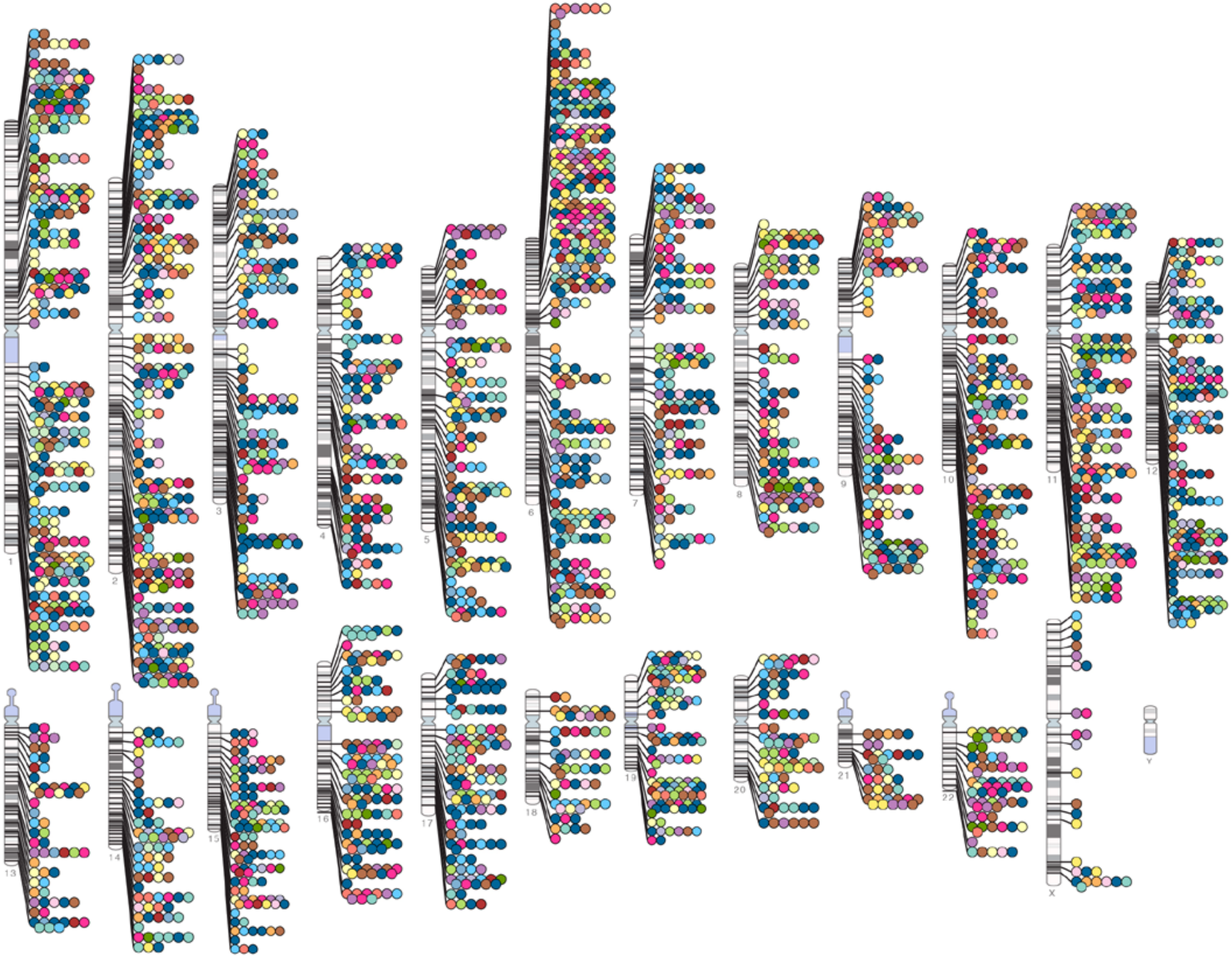
Genome Wide Association Studies (GWAS): Common SNPs

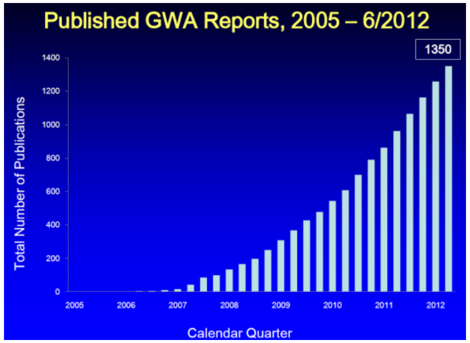


Rising to the top. In a genome-wide association study for type 2 diabetes, 386,731 genetic markers, shown here by chromosome, pop up. Those above the higher line appeared to be significantly associated with disease.

Published Genome-Wide Associations through 6/2012

(GWAS hits at $p \leq 5 \times 10^{-8}$ for 17 trait categories; Individual Chromosomal Locations)





- ● Digestive system disease
- ● Cardiovascular disease
- ● Metabolic disease
- ● Immune system disease
- ● Nervous system disease
- ● Liver enzyme measurement
- ● Lipid or lipoprotein measurement
- ● Inflammatory marker measurement
- ● Hematological measurement
- ● Body measurement
- ● Cardiovascular measurement
- ● Other measurement
- ● Response to drug
- ● Biological process
- ● Cancer
- ● Other disease
- ● Other trait

NHGRI GWA Catalog (www.genome.gov/GWAStudies)

The Limitations of Standard GWA Study Paradigms

- GWAS focusing on **common variations** have resulted in unequivocal statistical associations

- Associated genes have, on average, very small effects on disease (Odds Ratios of ~1.2-1.4)

- Collectively, the variations typically explain a very small fraction of the disease burden in the population (e.g., 4-10%)

- How can contemporary GWA study paradigms be extended, complemented or replaced to advance the identification and characterization of genetic factors contributing to disease? **Detect Rare variations?**

NEWS FEATURE PERSONAL GENOMES NATURE|Vol 456|6 November 2008

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.

If you want to predict how tall your children might one day be, a good bet would be to look in the mirror, and at your mate. Studies going back almost a century have estimated that height is 80–90% heritable. So if 29 centimetres separate the tallest 5% of a population from the shortest, then genetics would account for as many as 27 of them[1].

This year, three groups of researchers[2–4] scoured the genomes of huge populations (the largest study[4] looked at more than 30,000 people) for genetic variants associated with the height differences. More than 40 turned up.

But there was a problem: the variants had tiny effects. Altogether, they accounted for little more than 5% of height's heritability — just 6 centimetres by the calculations above. Even though these genome-wide association studies (GWAS) turned up dozens of variants, they did "very little of the prediction that you would do just by asking people how tall their parents are", says Joel Hirschhorn at the Broad Institute in Cambridge, Massachusetts, who led one of the studies[2].

Height isn't the only trait in which genes have gone missing, nor is it the most important. Studies looking at similarities between identical and fraternal twins estimate heritability at more than 90% for autism[5] and more than 80% for schizophrenia[6]. And genetics makes a major contribution to disorders such as obesity, diabetes and heart disease. GWAS, one of the most celebrated techniques of the past five years, promised to deliver many of the genes involved (see 'Where's the reward?', page 20). And to some extent they have, identifying more than 400 genetic variants that contribute to a variety of traits and common diseases. But even when dozens of genes have been linked to a trait, both the individual and cumulative effects are disappointingly small and nowhere near enough to explain earlier estimates of heritability. "It is the big topic in the genetics of common disease right now," says Francis Collins, former head of the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. The unexpected results left researchers at a point "where we all had to scratch our heads and say, 'Huh?'", he says.

Although flummoxed by this missing heritability, geneticists remain optimistic that they can find more of it. "These are very early days, and there are things that are doable in the next year or two that may well explain another sizeable chunk of heritability," says Hirschhorn. So where might it be hiding?

18

© 2008 Macmillan Publishers Limited. All rights reserved

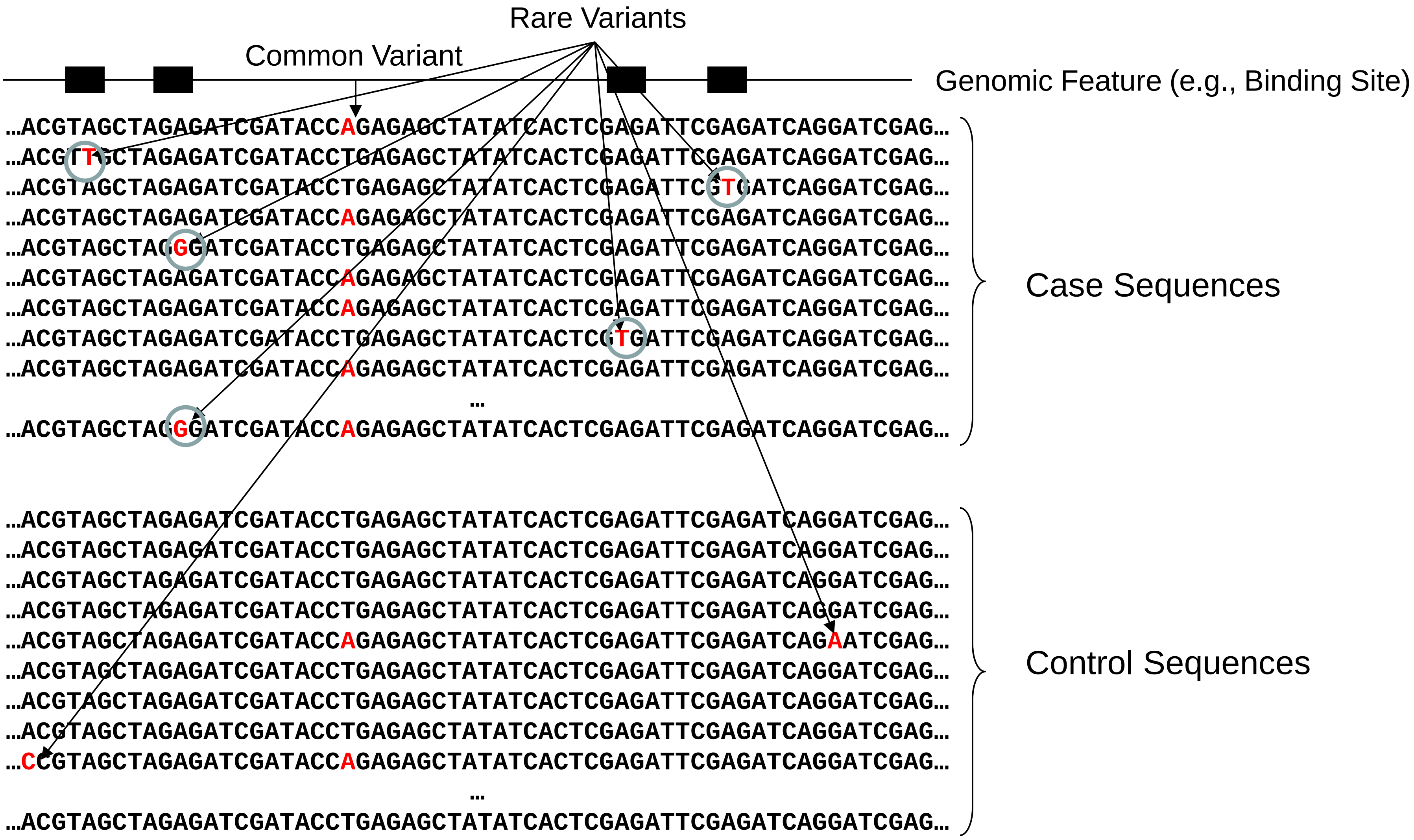
Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

‘Collapsing’ Rare Variations Based on Functional ‘Features’

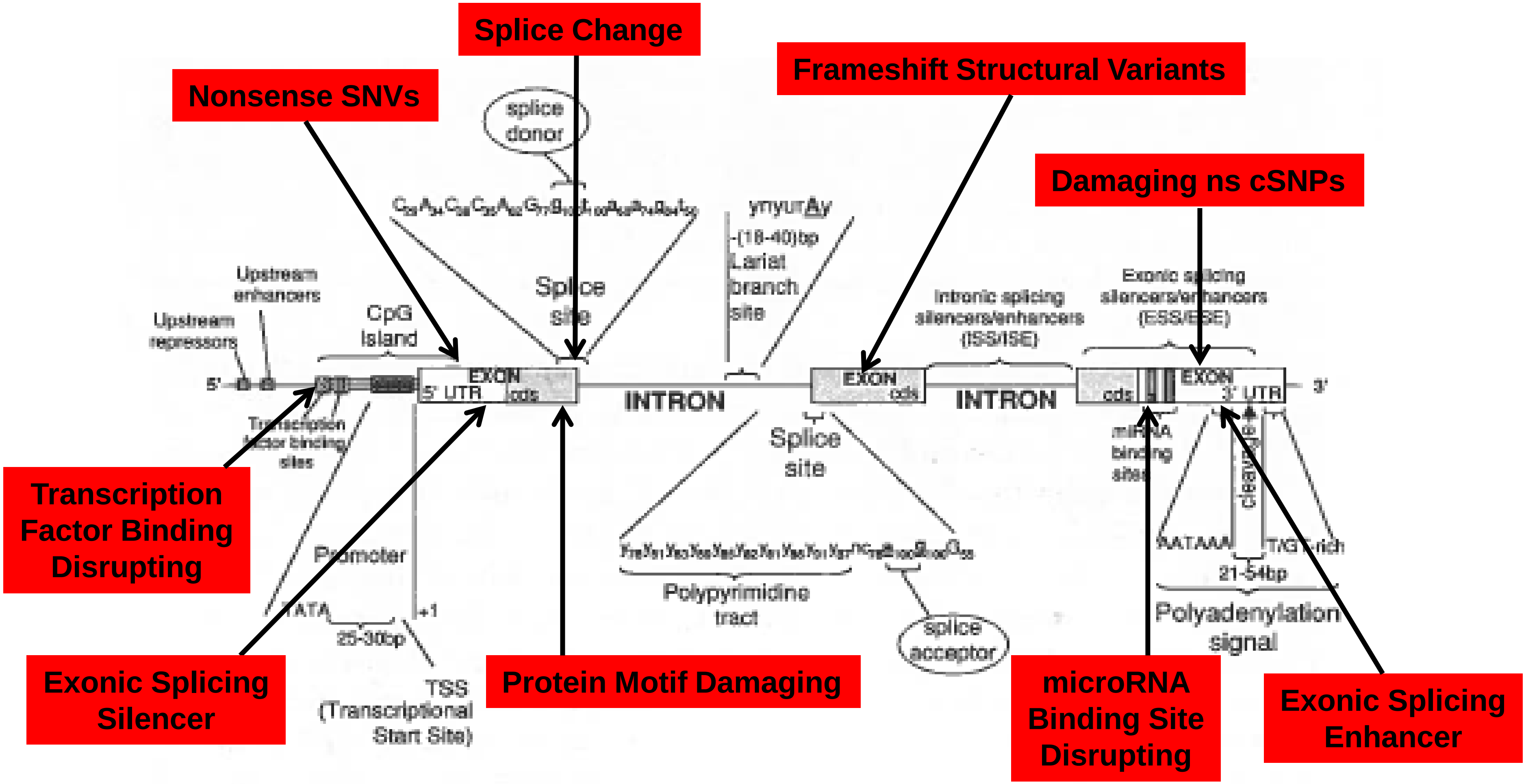


…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTTGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGTGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGGGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGTGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…
…ACGTAGCTAGGGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…

…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGAATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…CCGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…

Basic Intuition: Compare the *Collective* Frequency of Variants Between, e.g., Groups

Functional Annotations: *Bioinformatic* Predictions



Splice Change

Nonsense SNVs

Frameshift Structural Variants

Damaging ns cSNPs

Transcription Factor Binding Disrupting

Exonic Splicing Silencer

Protein Motif Damaging

microRNA Binding Site Disrupting

Exonic Splicing Enhancer

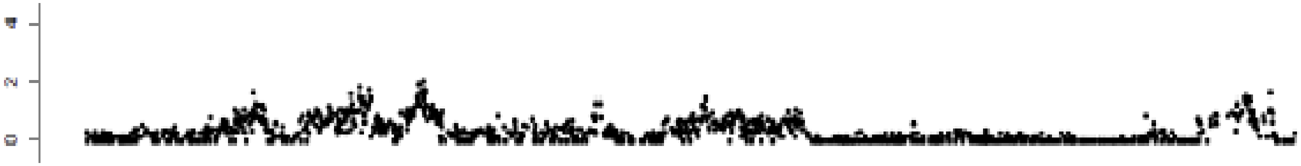
Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

Plumpton and Barnes. “Predictive Functional Analysis of Polymorphisms: An Overview.” in Bioinformatics for Geneticists. Wiley, 2007

We have developed methodology and tools for comprehensive bioinformatic WGS annotation

(Schork, Torkamani and colleagues: Bioinformatics 2008, 2009; Cancer Research (2009), Nat Gen Rev (2010), Genomics (2011))

Defined Region(s) vs. Moving Window Analyses

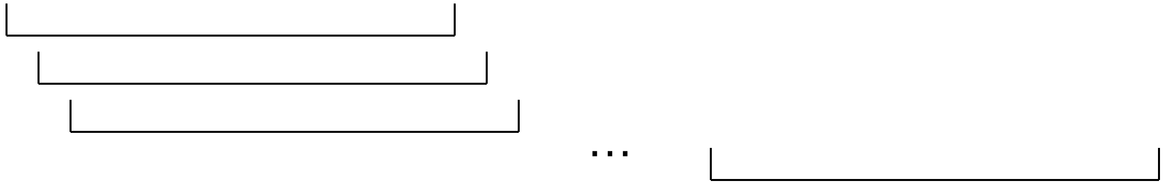


…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGTGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGGGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…
…ACGTAGCTAGGGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…

Case Sequences

…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGAATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…CCGTAGCTAGAGATCGATACCAGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…
…
…ACGTAGCTAGAGATCGATACCTGAGAGCTATATCACTCGAGATTCGAGATCAGGATCGAG…

Control Sequences



Multiple ‘Driver’ Tumor Mutations in the Same Gene/Protein

Torkamani, Verkhivker, Schork. *Cancer Letters*. 2008

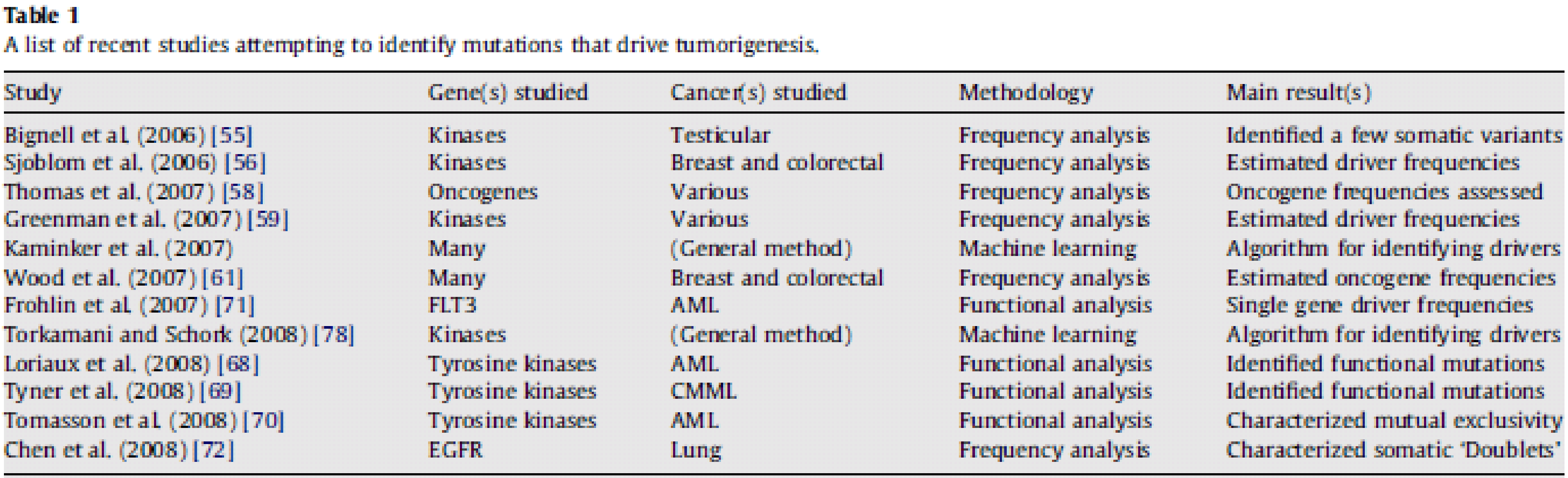
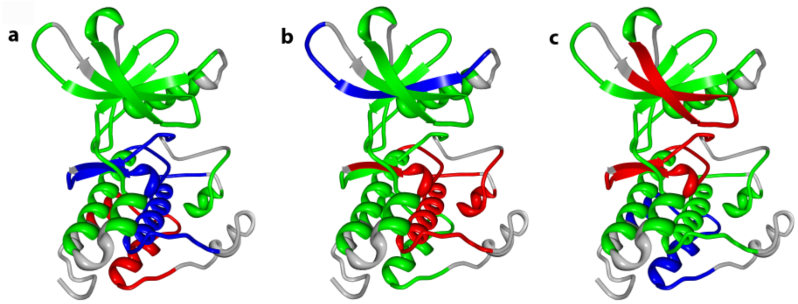


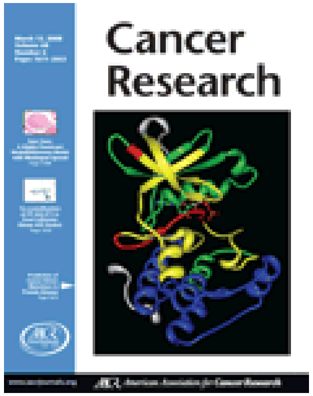
Table 1
A list of recent studies attempting to identify mutations that drive tumorigenesis.

Study	Gene(s) studied	Cancer(s) studied	Methodology	Main result(s)
Bignell et al. (2006) [55]	Kinases	Testicular	Frequency analysis	Identified a few somatic variants
Sjoblom et al. (2006) [56]	Kinases	Breast and colorectal	Frequency analysis	Estimated driver frequencies
Thomas et al. (2007) [58]	Oncogenes	Various	Frequency analysis	Oncogene frequencies assessed
Greenman et al. (2007) [59]	Kinases	Various	Frequency analysis	Estimated driver frequencies
Kaminker et al. (2007)	Many	(General method)	Machine learning	Algorithm for identifying drivers
Wood et al. (2007) [61]	Many	Breast and colorectal	Frequency analysis	Estimated oncogene frequencies
Frohlin et al. (2007) [71]	FLT3	AML	Functional analysis	Single gene driver frequencies
Torkamani and Schork (2008) [78]	Kinases	(General method)	Machine learning	Algorithm for identifying drivers
Loriaux et al. (2008) [68]	Tyrosine kinases	AML	Functional analysis	Identified functional mutations
Tyner et al. (2008) [69]	Tyrosine kinases	CMML	Functional analysis	Identified functional mutations
Tomasson et al. (2008) [70]	Tyrosine kinases	AML	Functional analysis	Characterized mutual exclusivity
Chen et al. (2008) [72]	EGFR	Lung	Frequency analysis	Characterized somatic 'Doublets'

Neutral SNPs Congenital Disease SNPs Tumor SNPs

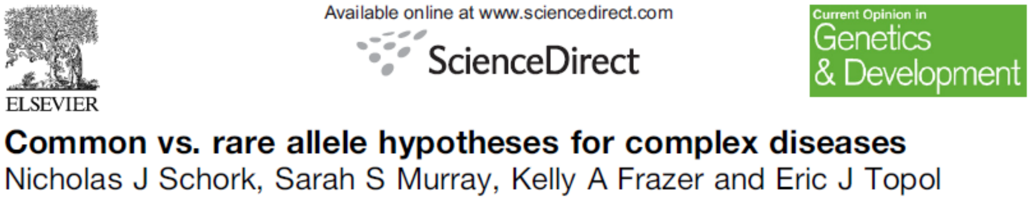


Random, Deficit, Enrichment



Torkamani Schork. *Cancer Research*. 68; 2008

Collections of ‘Causally Associated’ Rare Germline Variants



Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Genetics & Development

Common vs. rare allele hypotheses for complex diseases

Nicholas J Schork, Sarah S Murray, Kelly A Frazer and Eric J Topol

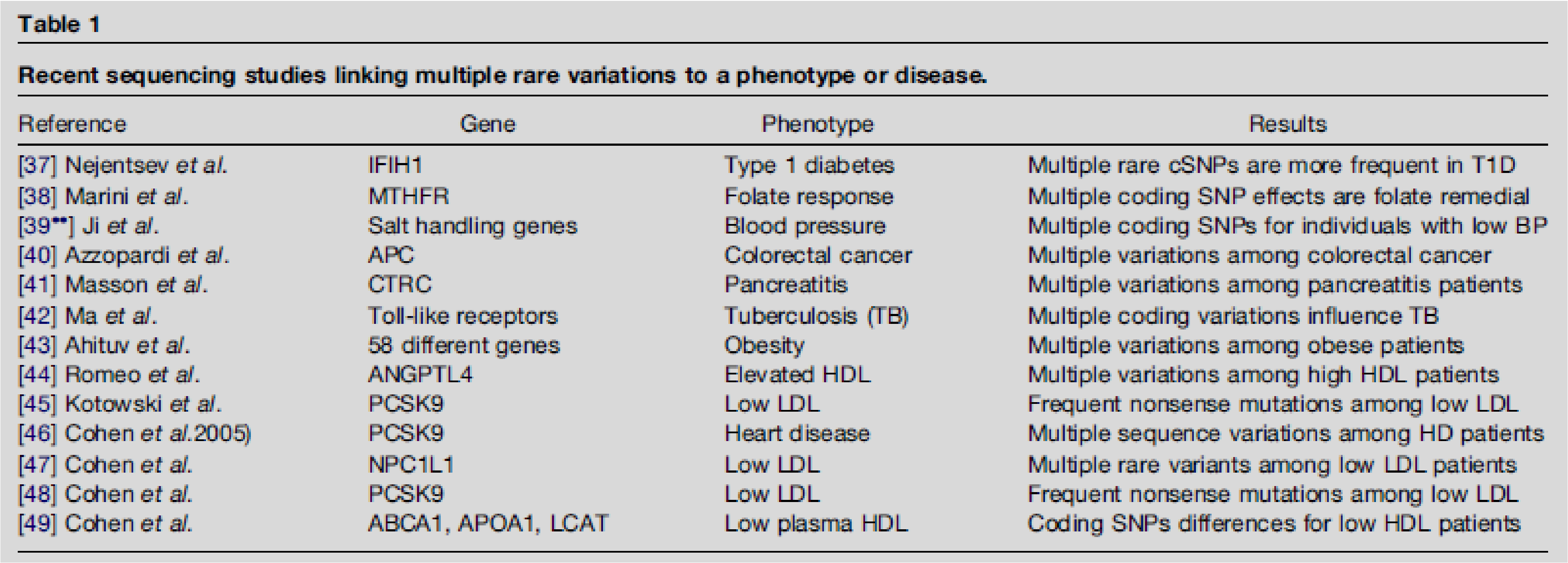


Table 1

Recent sequencing studies linking multiple rare variations to a phenotype or disease.

Reference	Gene	Phenotype	Results
[37] Nejentsev *et al.*	IFIH1	Type 1 diabetes	Multiple rare cSNPs are more frequent in T1D
[38] Marini *et al.*	MTHFR	Folate response	Multiple coding SNP effects are folate remedial
[39••] Ji *et al.*	Salt handling genes	Blood pressure	Multiple coding SNPs for individuals with low BP
[40] Azzopardi *et al.*	APC	Colorectal cancer	Multiple variations among colorectal cancer
[41] Masson *et al.*	CTRC	Pancreatitis	Multiple variations among pancreatitis patients
[42] Ma *et al.*	Toll-like receptors	Tuberculosis (TB)	Multiple coding variations influence TB
[43] Ahituv *et al.*	58 different genes	Obesity	Multiple variations among obese patients
[44] Romeo *et al.*	ANGPTL4	Elevated HDL	Multiple variations among high HDL patients
[45] Kotowski *et al.*	PCSK9	Low LDL	Frequent nonsense mutations among low LDL
[46] Cohen *et al.*2005)	PCSK9	Heart disease	Multiple sequence variations among HD patients
[47] Cohen *et al.*	NPC1L1	Low LDL	Multiple rare variants among low LDL patients
[48] Cohen *et al.*	PCSK9	Low LDL	Frequent nonsense mutations among low LDL
[49] Cohen *et al.*	ABCA1, APOA1, LCAT	Low plasma HDL	Coding SNPs differences for low HDL patients

- 1000 Genomes Project (www.1000genomes.org)

Whole Genome Sequencing Has Arrived…

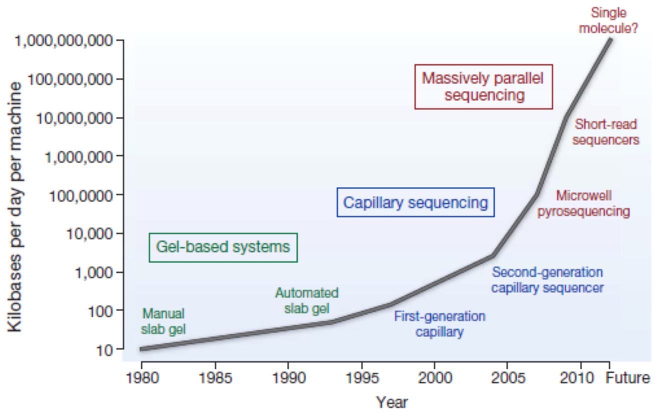
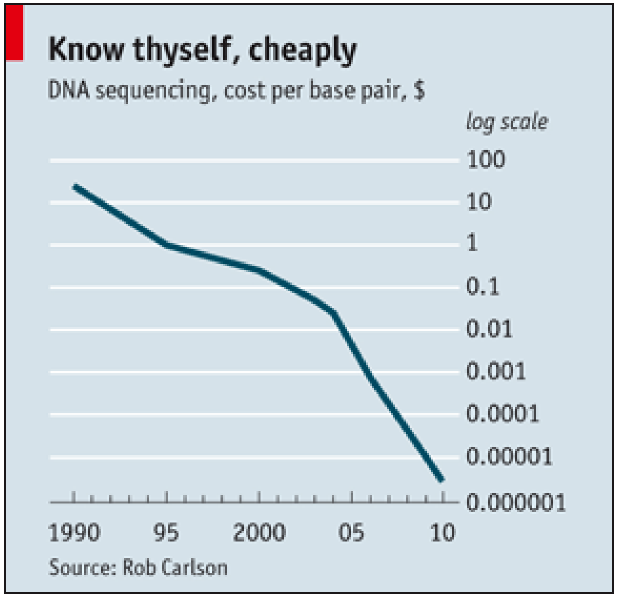
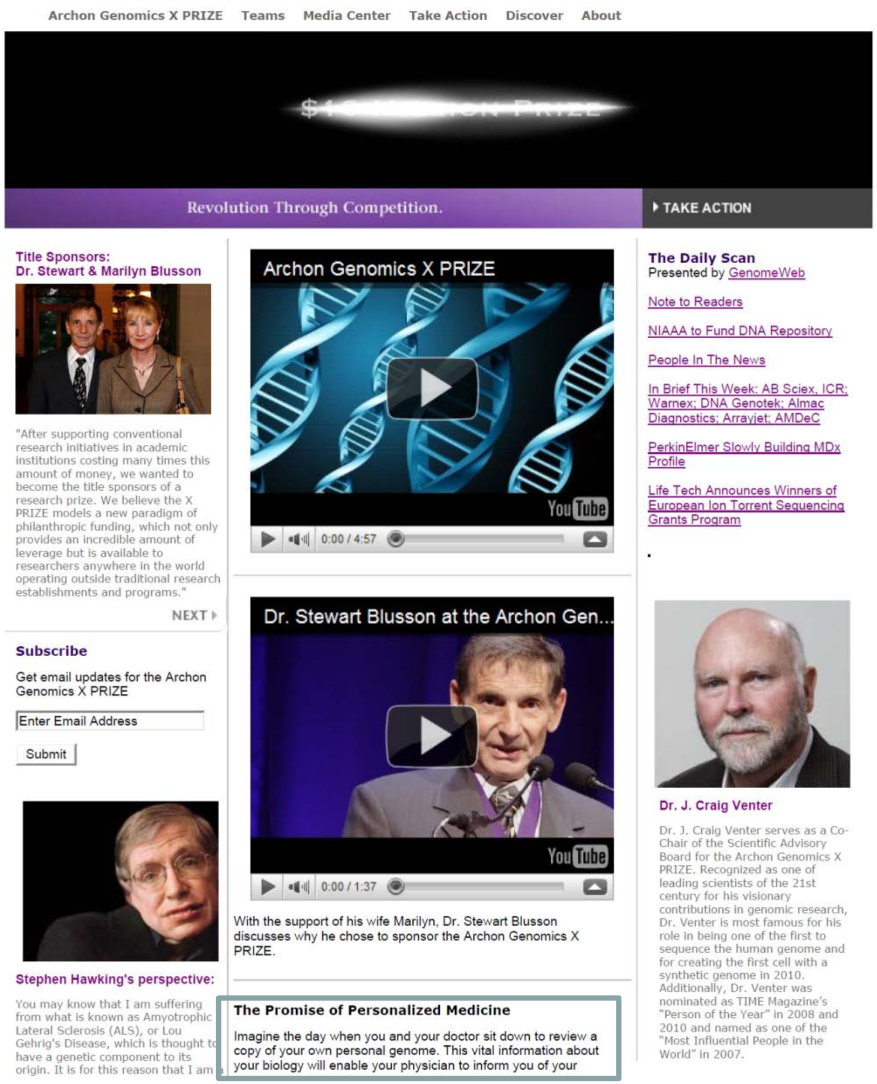


Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.



ARCHON GENOMICS X PRIZE

The '$1,000 Genome'



**Title Sponsors:
Dr. Stewart & Marilyn Blusson**

"After supporting conventional research initiatives in academic institutions costing many times this amount of money, we wanted to become the title sponsors of a research prize. We believe the X PRIZE models a new paradigm of philanthropic funding, which not only provides an incredible amount of leverage but is available to researchers anywhere in the world operating outside traditional research establishments and programs."

NEXT ▸

Subscribe

Get email updates for the Archon Genomics X PRIZE

Enter Email Address

Submit

Stephen Hawking's perspective:

You may know that I am suffering from what is known as Amyotrophic Lateral Sclerosis (ALS), or Lou Gehrig's Disease, which is thought to have a genetic component to its origin. It is for this reason that I am a

With the support of his wife Marilyn, Dr. Stewart Blusson discusses why he chose to sponsor the Archon Genomics X PRIZE.

The Promise of Personalized Medicine

Imagine the day when you and your doctor sit down to review a copy of your own personal genome. This vital information about your biology will enable your physician to inform you of your

The Daily Scan
Presented by GenomeWeb

Note to Readers

NIAAA to Fund DNA Repository

People In The News

In Brief This Week: AB Sciex, ICR; Warnex; DNA Genotek; Almac Diagnostics; Arrayjet; AMDeC

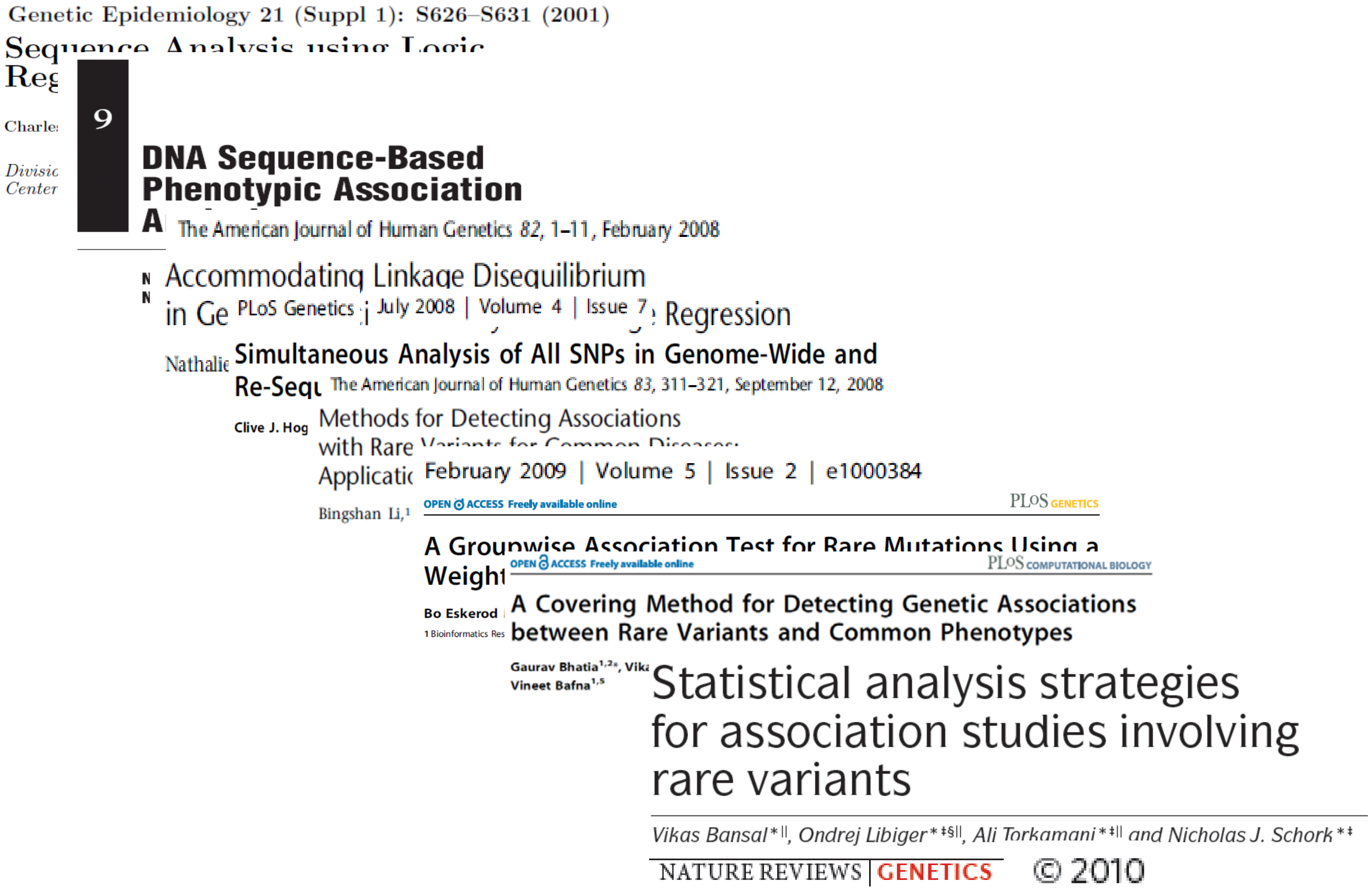
PerkinElmer Slowly Building MDx Profile

Life Tech Announces Winners of European Ion Torrent Sequencing Grants Program

Dr. J. Craig Venter

Dr. J. Craig Venter serves as a Co-Chair of the Scientific Advisory Board for the Archon Genomics X PRIZE. Recognized as one of leading scientists of the 21st century for his visionary contributions in genomic research, Dr. Venter is most famous for his role in being one of the first to sequence the human genome and for creating the first cell with a synthetic genome in 2010. Additionally, Dr. Venter was nominated as TIME Magazine's "Person of the Year" in 2008 and 2010 and named as one of the "Most Influential People in the World" in 2007.

Multilocus Association Studies with DNA Sequencing Data



Genetic Epidemiology 21 (Suppl 1): S626–S631 (2001)

Sequence Analysis using Logic
Reg

Charle

Divisic
Center

9

DNA Sequence-Based
Phenotypic Association
A

N
N

The American Journal of Human Genetics 82, 1–11, February 2008

Accommodating Linkage Disequilibrium
in Ge   Regression

Nathalie

PLoS Genetics | July 2008 | Volume 4 | Issue 7

Simultaneous Analysis of All SNPs in Genome-Wide and
Re-Seq

Clive J. Hog

The American Journal of Human Genetics 83, 311–321, September 12, 2008

Methods for Detecting Associations
with Rare Variants for Common Diseases:
Applicati

Bingshan Li,[1]

February 2009 | Volume 5 | Issue 2 | e1000384

OPEN ACCESS Freely available online   PLOS GENETICS

A Groupwise Association Test for Rare Mutations Using a
Weight

Bo Eskerod

1 Bioinformatics Res

OPEN ACCESS Freely available online   PLOS COMPUTATIONAL BIOLOGY

A Covering Method for Detecting Genetic Associations
between Rare Variants and Common Phenotypes

Gaurav Bhatia[1,2]*, Vika
Vineet Bafna[1,5]

Statistical analysis strategies
for association studies involving
rare variants

Vikas Bansal‖, Ondrej Libiger*‡§‖, Ali Torkamani*‡‖ and Nicholas J. Schork*‡*

NATURE REVIEWS | GENETICS   © 2010

Other Methods

European Journal of Human Genetics (2006) 14, 1037–1043
© 2006 Nature Publishing Group All rights reserved 1018-4813/06 $30.00
www.nature.com/ejhg

ARTICLE

A summary statistic approach to sequence variation in noncoding regions of six schizophrenia-associated gene loci

Jane Winantea[1,4], My N
Peter Propping[1], Marku

The American Journal of Human Genetics *85*, 427–446, October 9, 2009

ARTICLE

Rare, Evolutionarily Unlikely Missense Substitutions in *ATM* Confer Increased Risk of Breast Cancer

Sean V. Tavtigian,[1,12] Peter J. Oefner,[2,12] Davit Babikyan,[1] Anne Hartmann,[2] Sue Healey,[3]
Florence Le Calvez-Kelm,[1] Fabienne Lesueur,[1] Graham B. Byrnes,[1]
Nathalie Forey,[1] Corinna F
Sandrine McKay-Chopin,[1]
David C. Whiteman,[3] Aus
Kathleen Cuningham Foun
(kConFab),[6] Suleeporn San
Esther M. John,[10,11] and (

OPEN ACCESS Freely available online
October 2010 | Volume 6 | Issue 10 |
PLoS GENETICS

A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactio

Dajiang J. Liu[1,2], Suzanne

1 Department of Molecular and Human
Houston, Texas, United States of America

OPEN ACCESS Freely available online
November 2010 | Volume 6 | Issue 11
PLoS GENETICS

An Evolutionary Framework for Association Testing in Resequencing Studies

C. Ryan King[1*], Paul J. Rathouz[1,2], D

The American Journal of Human Genetics *87*, 604–617, November 12, 2010

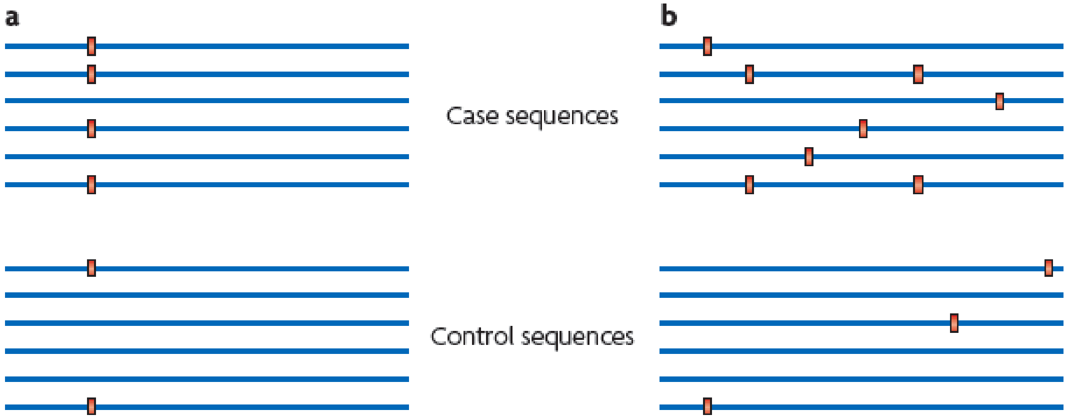
ARTICLE

Extending Rare-Variant Testing Strategies: Analysis of Noncoding Sequence and Imputed Genotypes

Matthew Zawistowski,[1,2] Shyam Gopalakrishnan,[1,2] Jun Ding,[1,2] Yun Li,[3,4] Sara Grimm,[5]
and Sebastian Zöllner[1,2,6,7,*]

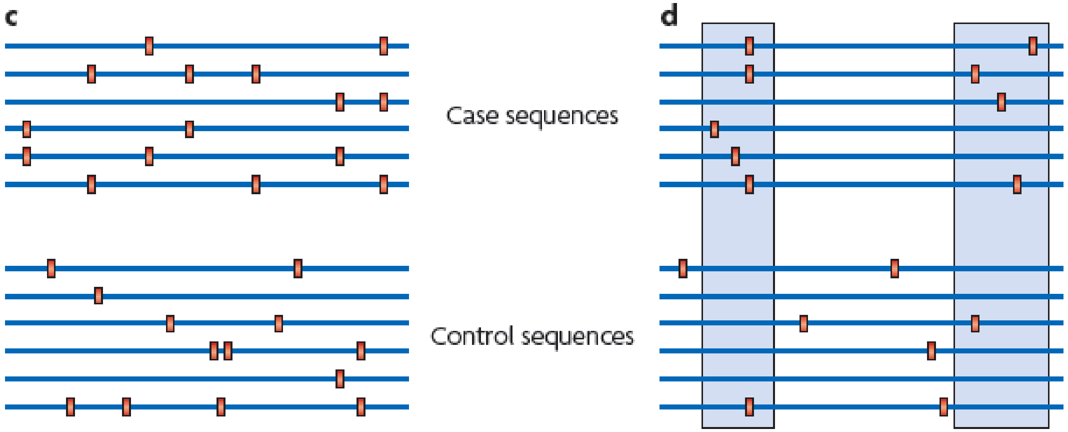
The ‘Anna Karenina’ or ‘Extreme Allelic Heterogeneity’ (EAH) Rare Variant Setting vs. Other Settings

Most studied: ‘Extreme Allelic Heterogeneity’ (EAH) setting. 'Happy families are all alike; every unhappy family is unhappy in its own way.‘ Leo Tolstoy, *Anna Karenina*



Common Variant

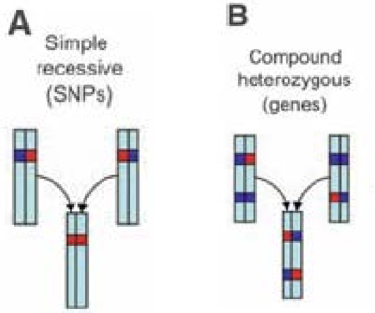
EAH

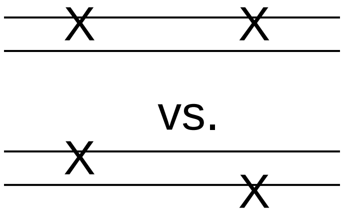


Synergistic Effects

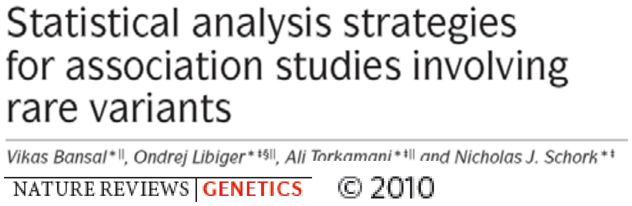
Region Specific EAH

Roach et al. Science (2010)





Compound Heterozygosity



Statistical analysis strategies for association studies involving rare variants

Vikas Bansal*||, Ondrej Libiger*‡§||, Ali Torkamani*‡|| and Nicholas J. Schork*‡

NATURE REVIEWS | GENETICS © 2010

Approaches for the Analysis of Collections of Rare Variants

Summary Statistics

- Leverages, e.g., weighted averages, sample diversity measures, sample distances between groups, etc. at the group summary level

Sequence Similarity and Diversity Measures

- Compare the nucleotide content of an individual's sequence against all other individuals and look for patterns among/between, e.g., cases and controls

Regression Methods

- Phenotype is the dependent and individual variants, collections of variants, non-genetic factors, and interaction terms as independent/predictor variable

Phase-Dependent Models (Compound Heterozygosity)

- Requires phase information and contrasting cis/trans effect models.

Bansal, Libiger, Torkamani, Schork. *Nature Reviews Genetics*. November 2010

Sanofi/Scripps Study: Gene Sequence Variation and Obesity

- 298 Individuals (148 morbidly obese; 150 controls)

- Two endocannabinoid genes sequenced using Illumina GA (FAAH; MGLL)

- Standard assembly for SNP identification (60x coverage; 3 reads per variant)

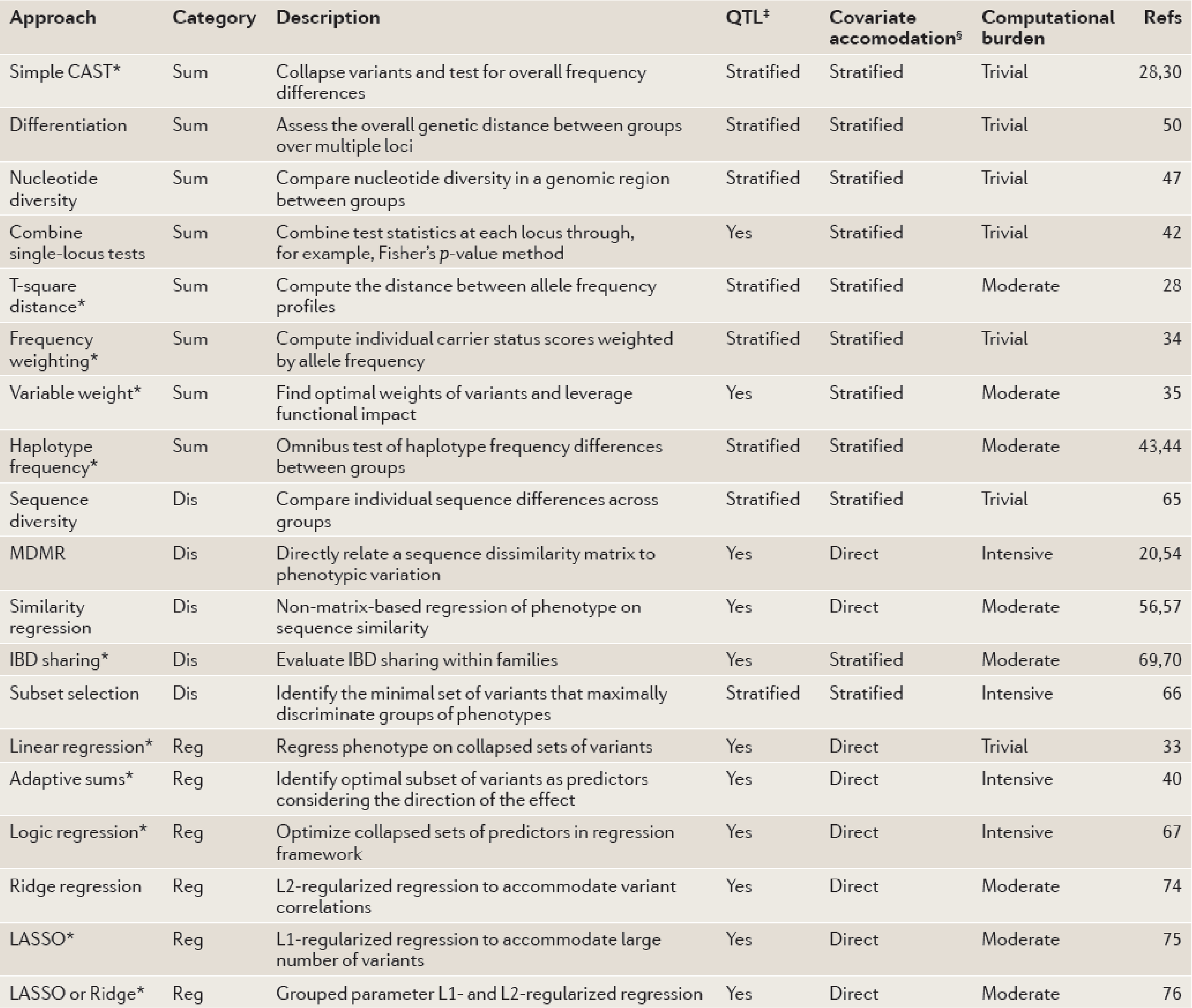
- 242 variants identified in FAAH (many novel and rare): 31 kb of sequence

- 1232 variants identified in MGLL (many novel and rare): 157 kb of sequence

- FAAH: located on chromosome 1p33, known to hydrolize anandamide (AEA), and other fatty acid amides

- MGLL: located on chromosome 3q21.3, a presynaptic enzyme that hydrolyzes 2-arachidonoylglycerol (2-AG), the most abundant endocannabinoid found in the brain

Harismendy et al. Genome Biol. 2010 Nov 30;11(11):R118. PMID: 21118518
Bansal et al. Pac Symp Biocomput. 2011:76-87. PMID: 21121035



	Approach	Category	Description	QTL[‡]	Covariate accomodation[§]	Computational burden	Refs
	Simple CAST*	Sum	Collapse variants and test for overall frequency differences	Stratified	Stratified	Trivial	28,30
→	Differentiation	Sum	Assess the overall genetic distance between groups over multiple loci	Stratified	Stratified	Trivial	50
	Nucleotide diversity	Sum	Compare nucleotide diversity in a genomic region between groups	Stratified	Stratified	Trivial	47
→	Combine single-locus tests	Sum	Combine test statistics at each locus through, for example, Fisher's *p*-value method	Yes	Stratified	Trivial	42
→	T-square distance*	Sum	Compute the distance between allele frequency profiles	Stratified	Stratified	Moderate	28
→	Frequency weighting*	Sum	Compute individual carrier status scores weighted by allele frequency	Stratified	Stratified	Trivial	34
	Variable weight*	Sum	Find optimal weights of variants and leverage functional impact	Yes	Stratified	Moderate	35
→	Haplotype frequency*	Sum	Omnibus test of haplotype frequency differences between groups	Stratified	Stratified	Moderate	43,44
→	Sequence diversity	Dis	Compare individual sequence differences across groups	Stratified	Stratified	Trivial	65
→	MDMR	Dis	Directly relate a sequence dissimilarity matrix to phenotypic variation	Yes	Direct	Intensive	20,54
	Similarity regression	Dis	Non-matrix-based regression of phenotype on sequence similarity	Yes	Direct	Moderate	56,57
	IBD sharing*	Dis	Evaluate IBD sharing within families	Yes	Stratified	Moderate	69,70
→	Subset selection	Dis	Identify the minimal set of variants that maximally discriminate groups of phenotypes	Stratified	Stratified	Intensive	66
	Linear regression*	Reg	Regress phenotype on collapsed sets of variants	Yes	Direct	Trivial	33
	Adaptive sums*	Reg	Identify optimal subset of variants as predictors considering the direction of the effect	Yes	Direct	Intensive	40
→	Logic regression*	Reg	Optimize collapsed sets of predictors in regression framework	Yes	Direct	Intensive	67
→	Ridge regression	Reg	L2-regularized regression to accommodate variant correlations	Yes	Direct	Moderate	74
	LASSO*	Reg	L1-regularized regression to accommodate large number of variants	Yes	Direct	Moderate	75
	LASSO or Ridge*	Reg	Grouped parameter L1- and L2-regularized regression	Yes	Direct	Moderate	76

Bansal et al. Nature Reviews: Genetics (2010)

Multiple Variant Effects May Shaping Gene Function

- **Extreme Heterogeneity** (Li and Leal 2008)
- **Additive/Cumulative** (Morris and Zeggini 2010)
- **Synergy/Combinations** (Wessel and Schork 2006; Schork et al. 2008)
- **Opposing Rare Allele Effects** (Han and Pan 2010)
- **Common + Rare** (Madsen and Browning 2009; Han and Pan 2010)
- **Compound Heterozygosity** (?)

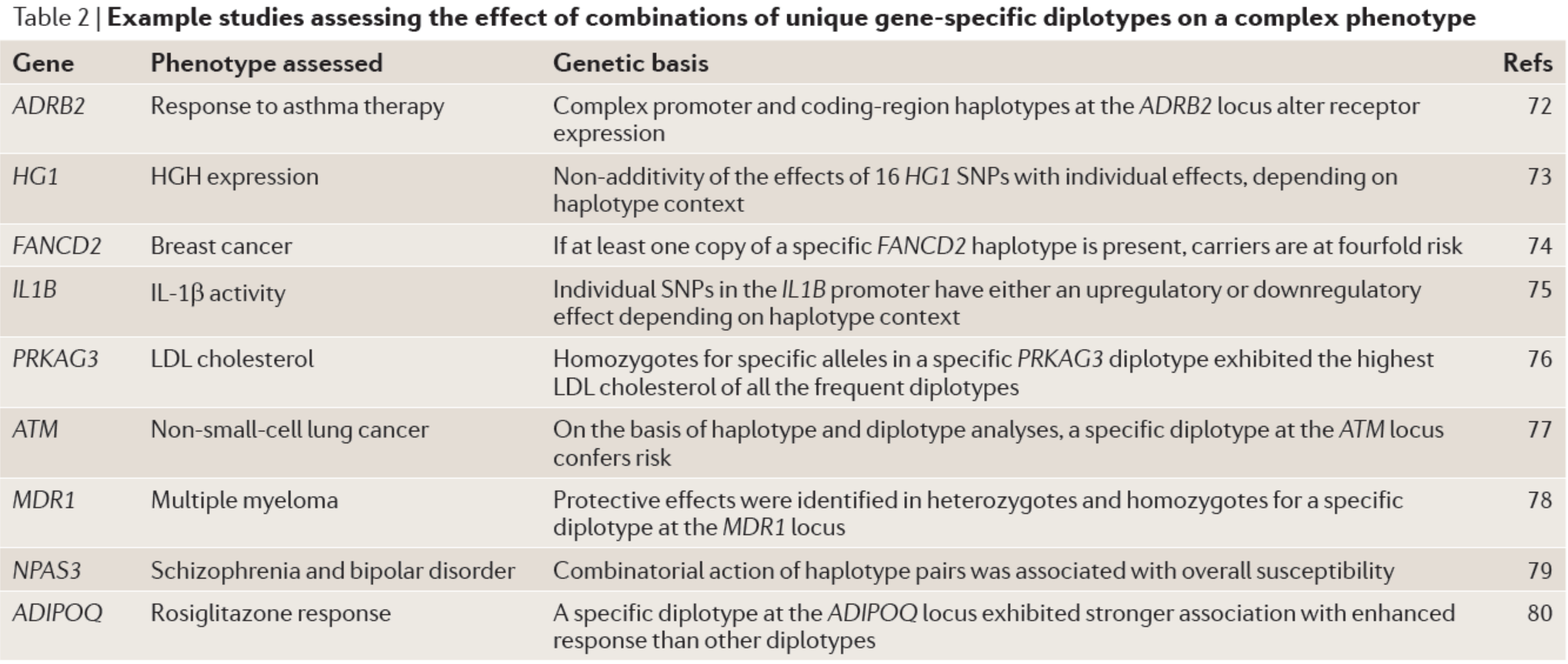


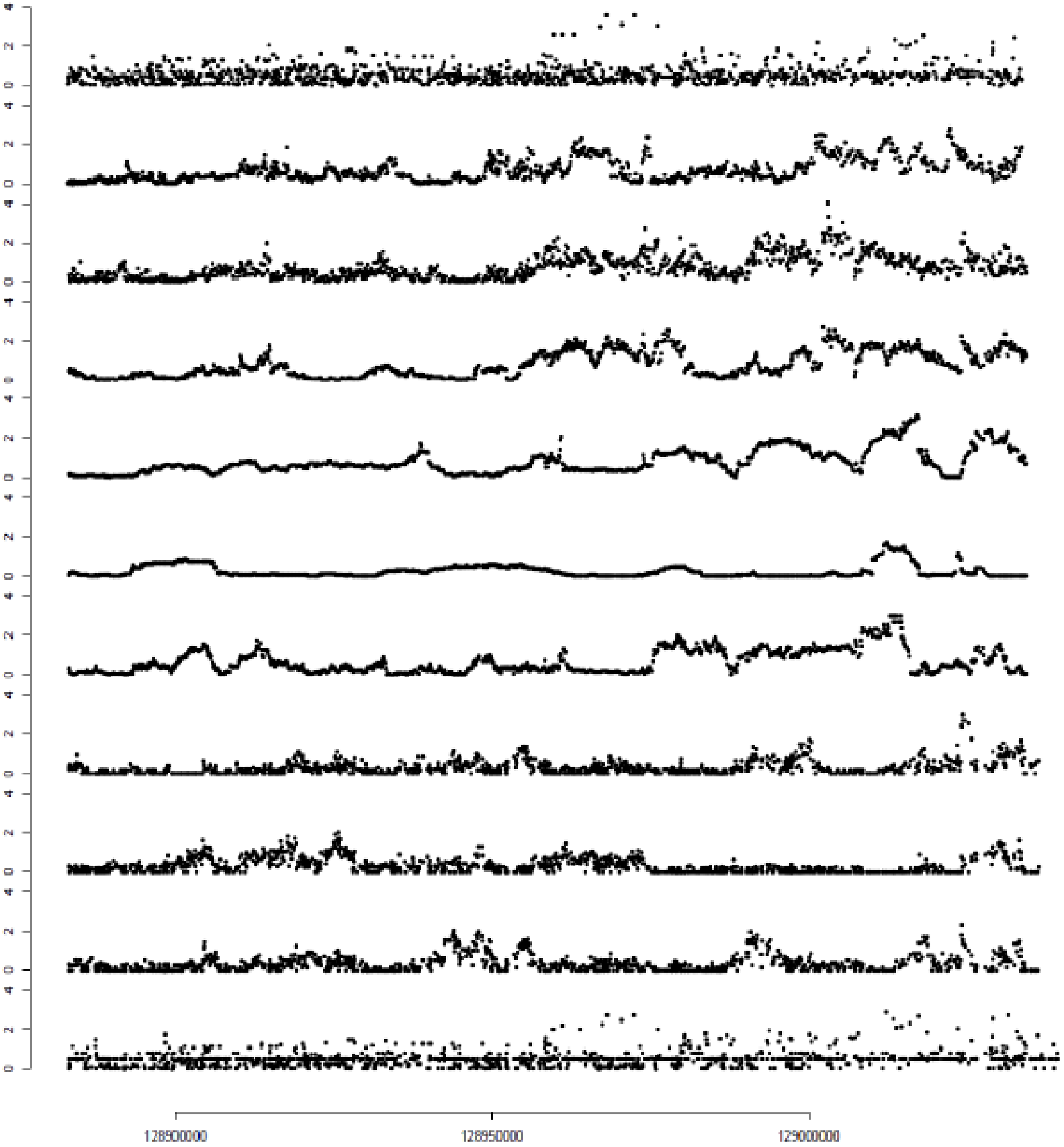
Table 2 | **Example studies assessing the effect of combinations of unique gene-specific diplotypes on a complex phenotype**

Gene	Phenotype assessed	Genetic basis	Refs
ADRB2	Response to asthma therapy	Complex promoter and coding-region haplotypes at the *ADRB2* locus alter receptor expression	72
HG1	HGH expression	Non-additivity of the effects of 16 *HG1* SNPs with individual effects, depending on haplotype context	73
FANCD2	Breast cancer	If at least one copy of a specific *FANCD2* haplotype is present, carriers are at fourfold risk	74
IL1B	IL-1β activity	Individual SNPs in the *IL1B* promoter have either an upregulatory or downregulatory effect depending on haplotype context	75
PRKAG3	LDL cholesterol	Homozygotes for specific alleles in a specific *PRKAG3* diplotype exhibited the highest LDL cholesterol of all the frequent diplotypes	76
ATM	Non-small-cell lung cancer	On the basis of haplotype and diplotype analyses, a specific diplotype at the *ATM* locus confers risk	77
MDR1	Multiple myeloma	Protective effects were identified in heterozygotes and homozygotes for a specific diplotype at the *MDR1* locus	78
NPAS3	Schizophrenia and bipolar disorder	Combinatorial action of haplotype pairs was associated with overall susceptibility	79
ADIPOQ	Rosiglitazone response	A specific diplotype at the *ADIPOQ* locus exhibited stronger association with enhanced response than other diplotypes	80

HGH, human growth hormone; IL-1β, interleukin-1β; LDL, low-density lipoprotein.

Tewhey et al. 2011

Different Methods Applied to the MGLL Gene



Set Method (Hoh and Ott 2003)

Omnibus Haplotype (Fallin et al. 2001)

Logic regression (Kooperberg et al. 2001)

Ridge regression (Malo et al. 2008)

Sequence similarity (Nievergelt et al. 2007)

Diversity (Jost 2007)

Distance Dispersion (Anderson 2006)

Subset selection (Bhatia et al. 2010)

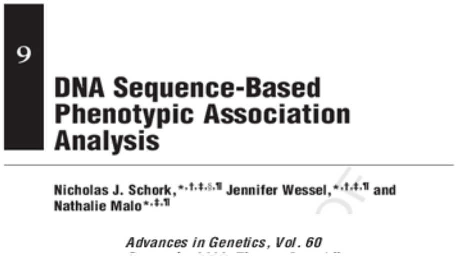
Weighted average (Madsen et al. 2009)

Hotelling’s T-square (Li and Leal 2008)

Fisher’s exact, single locus test

Bansal et al. PSB 2011

Distance-Based Sequence Analysis for Associations: Simple Nucleotide-Level Identity-By-State Similarity Matrix

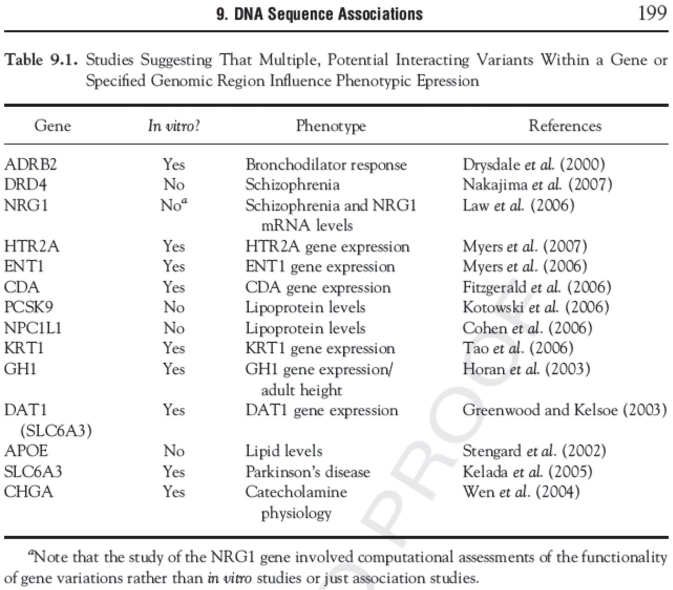


9

DNA Sequence-Based Phenotypic Association Analysis

Nicholas J. Schork,[*,†,‡,§,¶] Jennifer Wessel,[*,†,‡,¶] and Nathalie Malo[*,‡,¶]

Advances in Genetics, Vol. 60



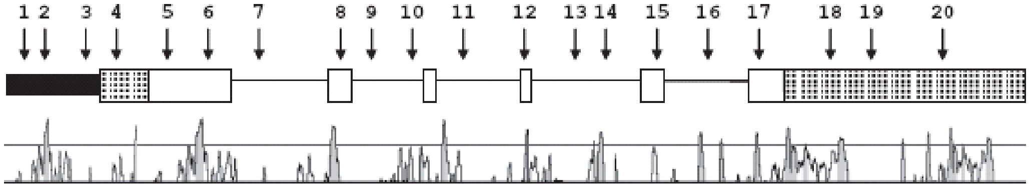
9. DNA Sequence Associations 199

Table 9.1. Studies Suggesting That Multiple, Potential Interacting Variants Within a Gene or Specified Genomic Region Influence Phenotypic Expression

Gene	*In vitro?*	Phenotype	References
ADRB2	Yes	Bronchodilator response	Drysdale *et al.* (2000)
DRD4	No	Schizophrenia	Nakajima *et al.* (2007)
NRG1	No[a]	Schizophrenia and NRG1 mRNA levels	Law *et al.* (2006)
HTR2A	Yes	HTR2A gene expression	Myers *et al.* (2007)
ENT1	Yes	ENT1 gene expression	Myers *et al.* (2006)
CDA	Yes	CDA gene expression	Fitzgerald *et al.* (2006)
PCSK9	No	Lipoprotein levels	Kotowski *et al.* (2006)
NPC1L1	No	Lipoprotein levels	Cohen *et al.* (2006)
KRT1	Yes	KRT1 gene expression	Tao *et al.* (2006)
GH1	Yes	GH1 gene expression/ adult height	Horan *et al.* (2003)
DAT1 (SLC6A3)	Yes	DAT1 gene expression	Greenwood and Kelsoe (2003)
APOE	No	Lipid levels	Stengard *et al.* (2002)
SLC6A3	Yes	Parkinson's disease	Kelada *et al.* (2005)
CHGA	Yes	Catecholamine physiology	Wen *et al.* (2004)

[a] Note that the study of the NRG1 gene involved computational assessments of the functionality of gene variations rather than *in vitro* studies or just association studies.

Sequence Diversity/Similarity Measure Approach



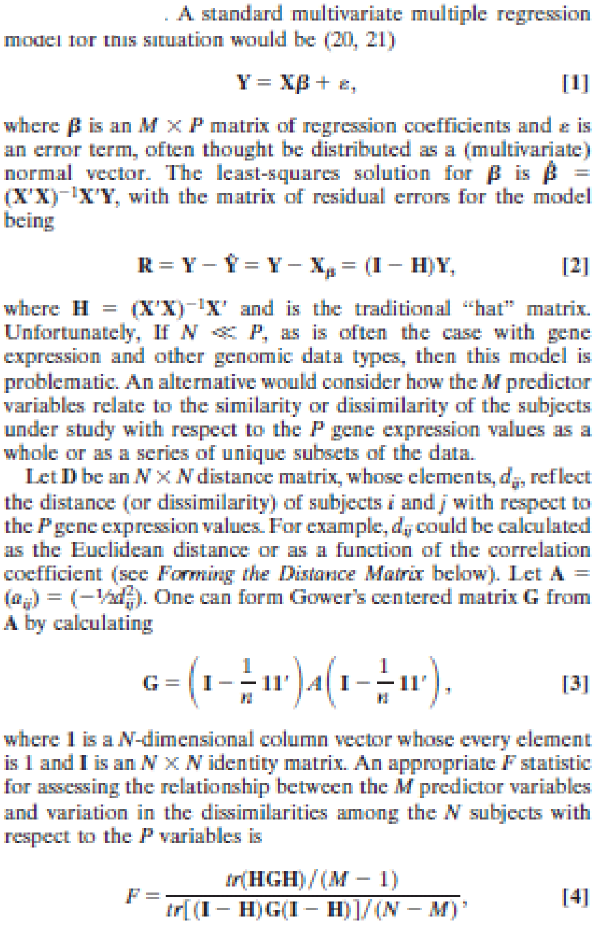
```
A.T..C..T...G..-...C......T.T...A..---...G...G..T..GC...T...A...---..C..GCT......   C1
A.C..C..T...G..A...C......T.G...A..ACT...C...G..T..GC...C...A...---..G..GCTCGT...   C2
A.T..C..T...G..-...C......C.G...A..ACT...G...A..T..-C...C...G...---..C..GCT......   C3

G.C..C..G...A..-...C......C.T...G..---...C...A..A..--...T...A...---..C..GCTCGTCGT   D1
A.C..A..G...G..A...T......T.T...G..ACT...G...G..A..-C...T...G...AAA..C..GCTCGTCGT   D2
G.C..A..T...G..A...C......C.T...G..---...C...G..T..-C...T...G...AAA..C..GCTCGT...   D3
```

Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet Epidemiol. 2011 [Epub ahead of print]; PMID:21308765

- ‘Distance’ measure is important and may impact inferences…
- Weighting schemes can be used to leverage information about positions
- Nucleotide sharing assumes **alignments** are perfect and capture structural variations
- Nucleotide sharing does not consider multinucleotide variations as single variations
- Take a ‘window’ of the genome, analyze it, and move to a new window…

Relating Variation in Similarity to Outcomes: MDMR/GAMOVA



. A standard multivariate multiple regression model for this situation would be (20, 21)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad [1]$$

where $\boldsymbol{\beta}$ is an $M \times P$ matrix of regression coefficients and ε is an error term, often thought be distributed as a (multivariate) normal vector. The least-squares solution for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, with the matrix of residual errors for the model being

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}_{\beta} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \qquad [2]$$

where $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and is the traditional "hat" matrix. Unfortunately, if $N \ll P$, as is often the case with gene expression and other genomic data types, then this model is problematic. An alternative would consider how the M predictor variables relate to the similarity or dissimilarity of the subjects under study with respect to the P gene expression values as a whole or as a series of unique subsets of the data.

Let $\mathbf{D}$ be an $N \times N$ distance matrix, whose elements, d_{ij}, reflect the distance (or dissimilarity) of subjects i and j with respect to the P gene expression values. For example, d_{ij} could be calculated as the Euclidean distance or as a function of the correlation coefficient (see *Forming the Distance Matrix* below). Let $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$. One can form Gower's centered matrix $\mathbf{G}$ from $\mathbf{A}$ by calculating

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)A\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right), \qquad [3]$$

where $\mathbf{1}$ is a N-dimensional column vector whose every element is 1 and $\mathbf{I}$ is an $N \times N$ identity matrix. An appropriate F statistic for assessing the relationship between the M predictor variables and variation in the dissimilarities among the N subjects with respect to the P variables is

$$F = \frac{tr(\mathbf{HGH})/(M-1)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(N-M)}, \qquad [4]$$

Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables

Matthew A. Zapala* and Nicholas

ARTICLE

Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis

Jennifer Wessel and Nicholas J. Schork

Large-scale, multilocus genetic association studies require powerful and appropriate statistical-analysis tools that are designed to relate genotype and

OPEN ACCESS Freely available online

PLoS GENETICS

Generalized Analysis of Molecular Variance

Caroline M. Nievergelt[1,3,4,5], Ondrej Libiger[1,3,4], Nicholas J. Schork[1,2,3,4,5,6*]

1 Department of Psychiatry, University of California at San Diego, La Jolla, California, United States of America, 2 Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, California, United States of America, 3 Rebecca and John Moores UCSD Cancer Center, University of California at San Diego, La Jolla, California, United States of America, 4 The Center for Human Genetics and Genomics, University of California at San Diego, La Jolla, California, United States of America, 5 The Stein Institute for Research on Aging, University of California at San Diego, La Jolla, California, United States of America, 6 Scripps Genomic Medicine and Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California, United States of America

Many studies in the fields of genetic epidemiology and applied population genetics are predicated on, or require, an assessment of the genetic background diversity of the individuals chosen for study. A number of strategies have been developed for assessing genetic background diversity. These strategies typically focus on genotype data collected on the individuals in the study, based on a panel of DNA markers. However, many of these strategies are either rooted in cluster analysis techniques, and hence suffer from problems inherent to the assignment of the biological and statistical meaning to resulting clusters, or have formulations that do not permit easy and intuitive extensions. We describe a very general approach to the problem of assessing genetic background diversity that extends the analysis of molecular variance (AMOVA) strategy introduced by Excoffier and colleagues some time ago. As in the original AMOVA strategy, the proposed approach, termed generalized AMOVA (GAMOVA), requires a genetic similarity matrix constructed from the allelic profiles of individuals under study and/or allele frequency summaries of the populations from which the individuals have been sampled. The proposed strategy can be used to either estimate the fraction of genetic variation explained by grouping factors such as country of origin, race, or ethnicity, or to quantify the strength of the relationship of the observed genetic background variation to quantitative measures collected on the subjects, such as blood pressure levels or anthropometric measures. Since the formulation of our test statistic is rooted in multivariate linear models, sets of variables can be related to genetic background in multiple regression-like contexts. GAMOVA can also be used to complement graphical representations of genetic diversity such as tree diagrams (dendrograms) or heatmaps. We examine features, advantages, and power of the proposed procedure and showcase its flexibility by using it to analyze a wide variety of published data sets, including data from the Human Genome Diversity Project, classical anthropometry data collected by Howells, and the International HapMap Project.

Citation: Nievergelt CM, Libiger O, Schork NJ (2007) Generalized analysis of molecular variance. PLoS Genet 3(4): e51. doi:10.1371/journal.pgen.0030051

Introduction

Genetic and genetic epidemiologic studies involving large numbers of individuals and/or populations are being pursued more and more often as a result of the development of high-throughput genotyping technologies and the creation of genotype data repositories such as the dbSNP (http://www.ncbi.nlm.nih.gov/SNP) and the International HapMap Project databases (http://www.hapmap.org).

population genetic studies. In addition, it has been shown that, due to phenomena such as varying degrees of admixture and/or cryptic relatedness in the study population, ignoring genetic background in epidemiologic studies testing associations between particular genetic variations and a phenotype can result in false positive and false negative results [9–19], which underscores the importance of genetic background analysis even in very simple genetic association studies.

PLoS Genetics | www.plosgenetics.org 0467 April 2007 | Volume 3 | Issue 4 | e51

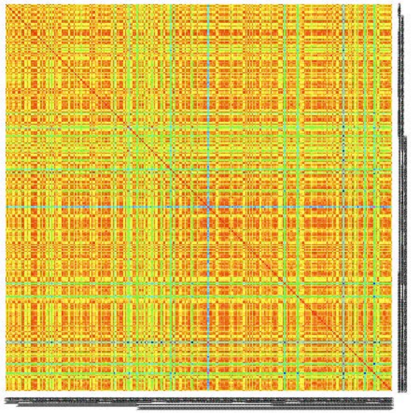
No *a priori* clustering or data reduction: test of predictors and variation in matrix

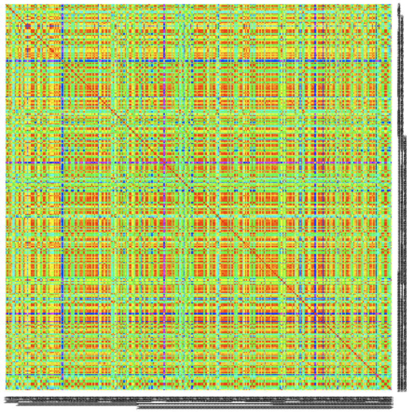
GAMOVA based association analysis with sequence data

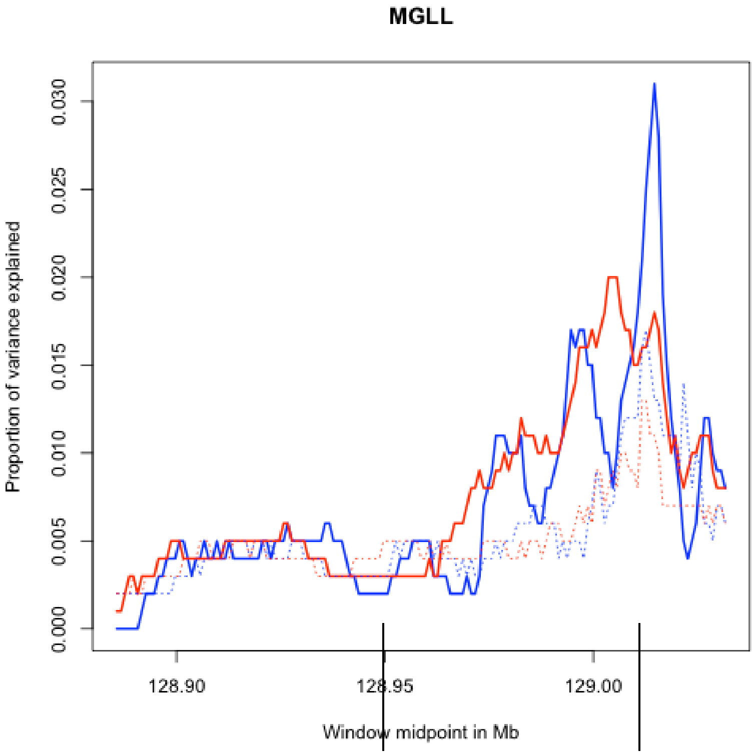
Wessel and Schork, AJHG (2006); Schork et al. Adv Gen (2008);

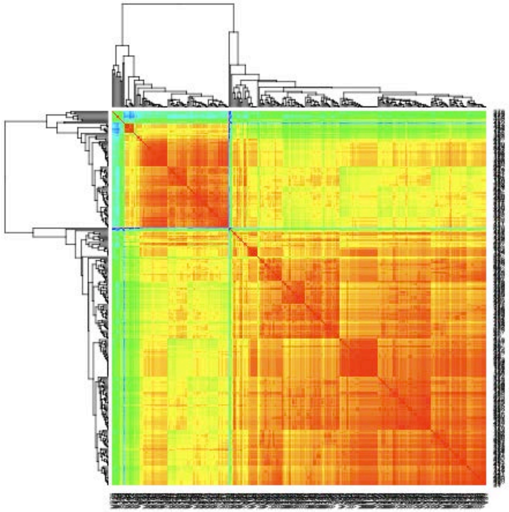
Ordered by BMI

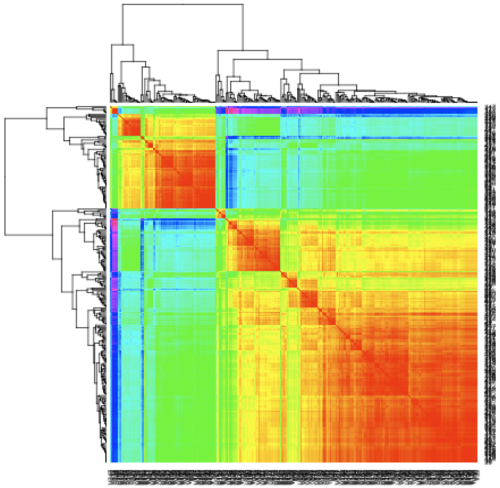
Ordered by similarity











Similarity Approach (Synergy)

Diversity Methods: Summary Measures vs. Comparing Individual Sequences

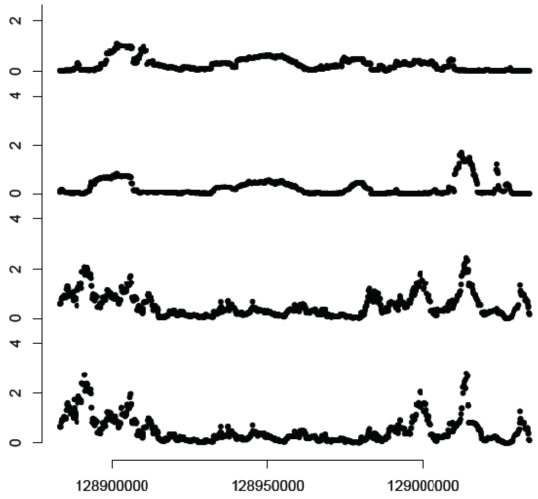
Molecular Ecology (2008) 17, 4015–4026 doi: 10.1111/j.1365-294X.2008.03887.x

G_{ST} and its relatives do not measure differentiation

LOU JOST
Via Runtun, Baños, Tungurahua, Ecuador

$$\Delta = \left(\sum_{i=1}^{k} p_i^{\lambda}\right)^{(1/(1-\lambda)}$$

Figure B.2. Window-based association analysis for the MGLL gene assuming a diversity statistic with different exponents based on the work of Jost (2007). The λ values used to construct the graphs are, from the bottom panel to the top panel: 0.2 0.5, 2.0, and 4.0.

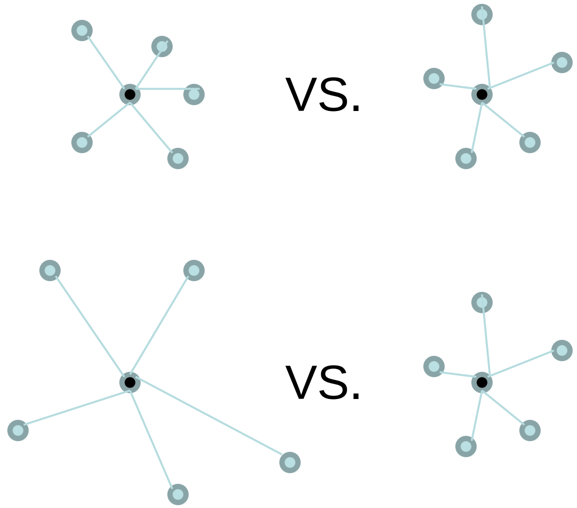


Summary Measure Approach

Biometrics 62, 245–253
March 2006 DOI: 10.1111/j.1541-0420.2005.00440.x

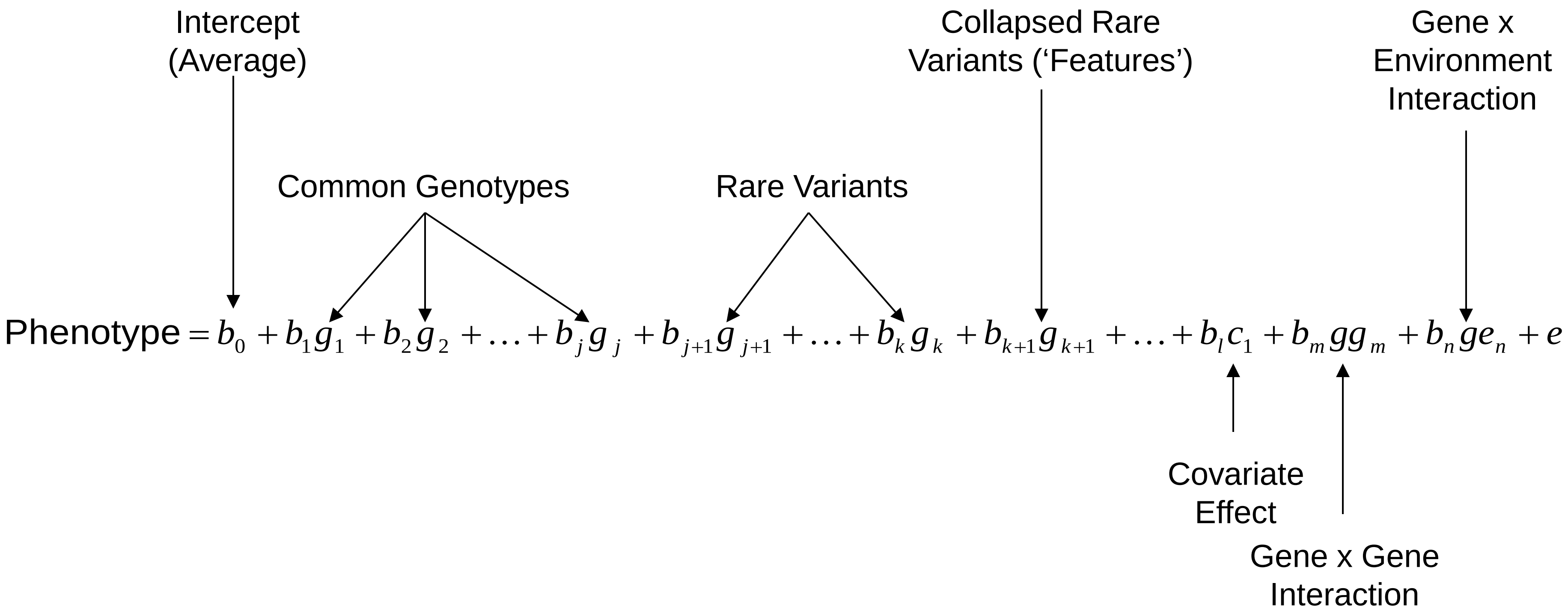
Distance-Based Tests for Homogeneity of Multivariate Dispersions

Marti J. Anderson
Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
email: mja@stat.auckland.ac.nz



Sequence Diversity/Similarity Measure Approach

Multilocus Regression for Sequence-Based Associations



Intercept (Average)

Collapsed Rare Variants (‘Features’)

Gene x Environment Interaction

Common Genotypes

Rare Variants

$$\text{Phenotype} = b_0 + b_1 g_1 + b_2 g_2 + \ldots + b_j g_j + b_{j+1} g_{j+1} + \ldots + b_k g_k + b_{k+1} g_{k+1} + \ldots + b_l c_1 + b_m gg_m + b_n ge_n + e$$

Covariate Effect

Gene x Gene Interaction

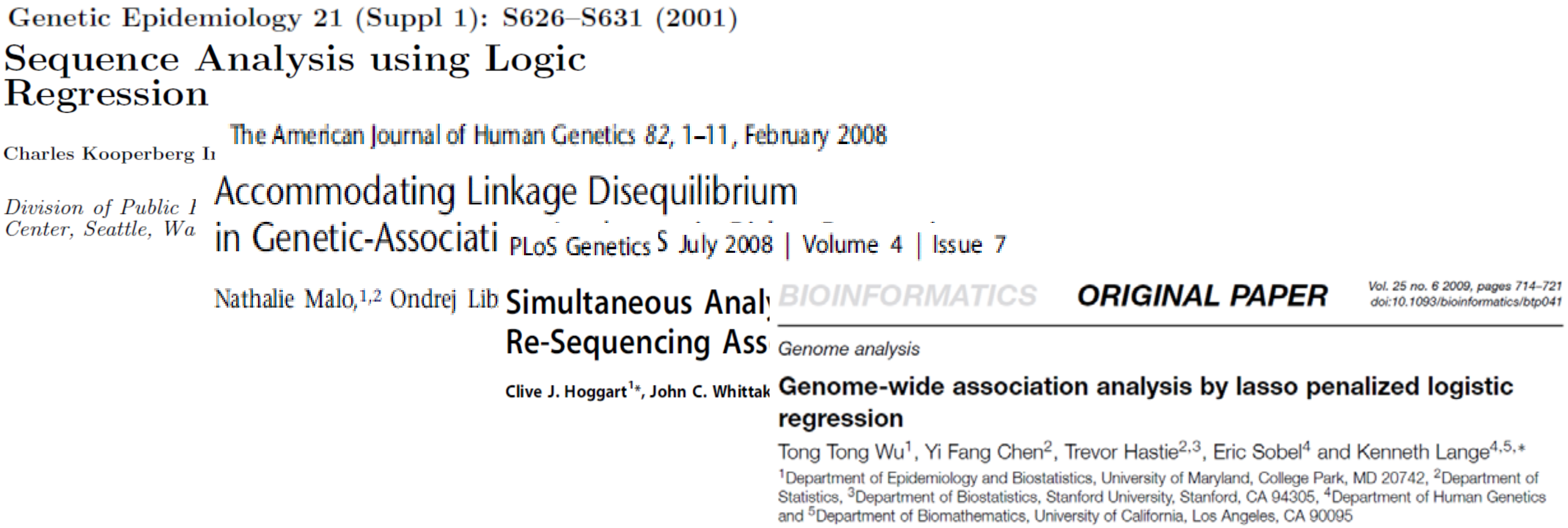
Problem 1: There will likely be many more ‘predictors’ than subjects

Problem 2: Collinearity between predictors (due to LD or by definition)

Solution?: Some form of regularization or shrinkage: $(\hat{\alpha}, \hat{\beta}) = \arg\min\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\}$ subject to $\sum_j |\beta_j| \leqslant t.$

Regression Method Approach (Stepwise, LASSO, Ridge, etc.)

Regression-Based Multilocus Association Analysis



Genetic Epidemiology 21 (Suppl 1): S626–S631 (2001)

Sequence Analysis using Logic Regression

Charles Kooperberg I

Division of Public 1
Center, Seattle, Wa

The American Journal of Human Genetics *82*, 1–11, February 2008

Accommodating Linkage Disequilibrium
in Genetic-Associati

Nathalie Malo,[1,2] Ondrej Lib

PLoS Genetics | July 2008 | Volume 4 | Issue 7

Simultaneous Anal
Re-Sequencing Ass

Clive J. Hoggart[1*], John C. Whittak

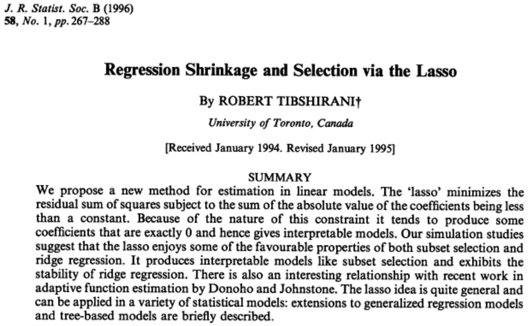
BIOINFORMATICS **ORIGINAL PAPER** *Vol. 25 no. 6 2009, pages 714–721*
doi:10.1093/bioinformatics/btp041

Genome analysis

Genome-wide association analysis by lasso penalized logistic regression

Tong Tong Wu[1], Yi Fang Chen[2], Trevor Hastie[2,3], Eric Sobel[4] and Kenneth Lange[4,5,*]

[1]Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, [2]Department of Statistics, [3]Department of Biostatistics, Stanford University, Stanford, CA 94305, [4]Department of Human Genetics and [5]Department of Biomathematics, University of California, Los Angeles, CA 90095



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

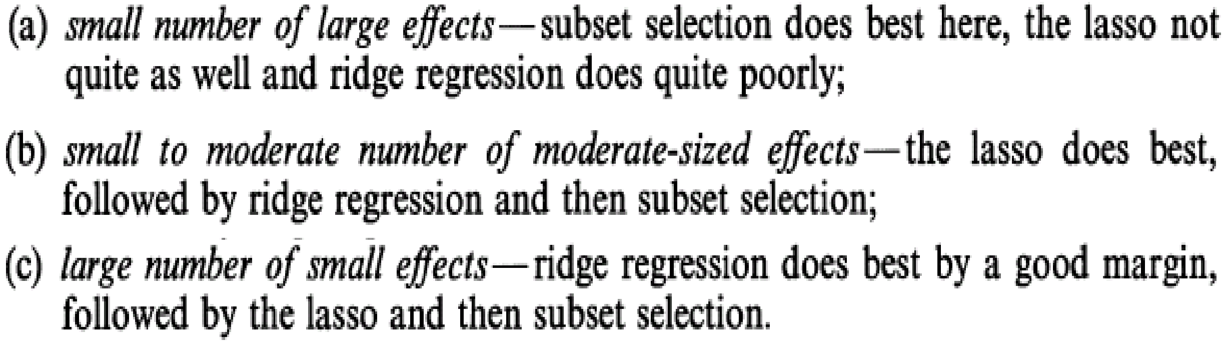
By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.



(a) *small number of large effects*—subset selection does best here, the lasso not quite as well and ridge regression does quite poorly;

(b) *small to moderate number of moderate-sized effects*—the lasso does best, followed by ridge regression and then subset selection;

(c) *large number of small effects*—ridge regression does best by a good margin, followed by the lasso and then subset selection.

- **Problem**: a researcher won't know *a priori* which situation represents the truth…

Genomic Features with Collapsed Variations

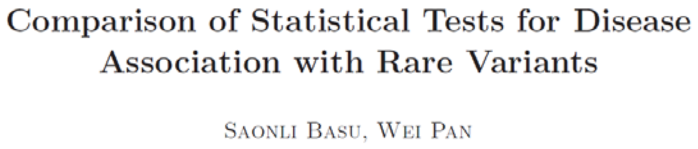
Table 2. P-values for association for each analysis method for specific sets of collapsed variations in the MGLL Gene

Different Procedures

			FAAH		
	NS	**H3K27**	**TFBS**	**FOX2**	**Amidase**
# of variants	5	29	4	14	5
Dispersion (Dis)	0.59	0.05	0.77	0.99	0.61
Diversity (Div)	0.43	0.42	0.81	0.33	0.46
MDMR Similarity (Sim)	0.19	0.21	0.05	0.14	0.41
Li & Leal (LL)	0.60	0.03	0.60	1.00	0.50
Subset Selection (SS)	1.00	0.01	0.60	0.75	0.60
Madsen & Browning (MB)	1.00	0.01	0.33	1.00	0.75
Logic Regression (LR)	0.23	0.18	0.39	0.22	0.48
Ridge Regresssion (RR)	0.35	0.09	0.06	0.33	0.54
PLINK Haplotype (Phap)	NA	0.92	NA	0.34	0.61
PLINK Set Analysis (Pset)	1.00	1.00	0.02	1.00	1.00
			MGLL		
	NS	**H3K27**	**TFBS**	**FOX2**	**Amidase**
# of variants	9	100	11	3	0
Dispersion	0.28	0.99	0.02	0.72	NA
Diversity	0.77	0.65	0.73	0.64	NA
MDMR	0.81	0.07	0.67	0.29	NA
Li & Leal	1.00	1.00	1.00	0.75	NA
SubsetSelection	0.60	0.43	1.00	1.00	NA
Madsen & Browning	0.75	0.30	0.02	0.20	NA
Logic Regression	0.35	0.67	0.02	0.49	NA
Ridge Reg.	0.71	0.50	0.01	0.61	NA
PLINK Haplotype	NA	0.81	0.07	NA	NA
PLINK Set Analysis	1.00	0.43	0.05	1.00	NA

Bansal et al. PSB 2011

Simulation-based Comparison of Methods



Comparison of Statistical Tests for Disease Association with Rare Variants

Saonli Basu, Wei Pan

http://www.biostat.umn.edu/~weip/paper/RV2.pdf

- Simulate a wide variety of settings: with LD, with opposite effect variants, with neutral variants, etc.

- Fit a number of different methods

- The Kernel Machine Regression (KMR) which was shown to be equivalent to GAMOVA/MDMR similarity-based method was one of the most consistently best performers

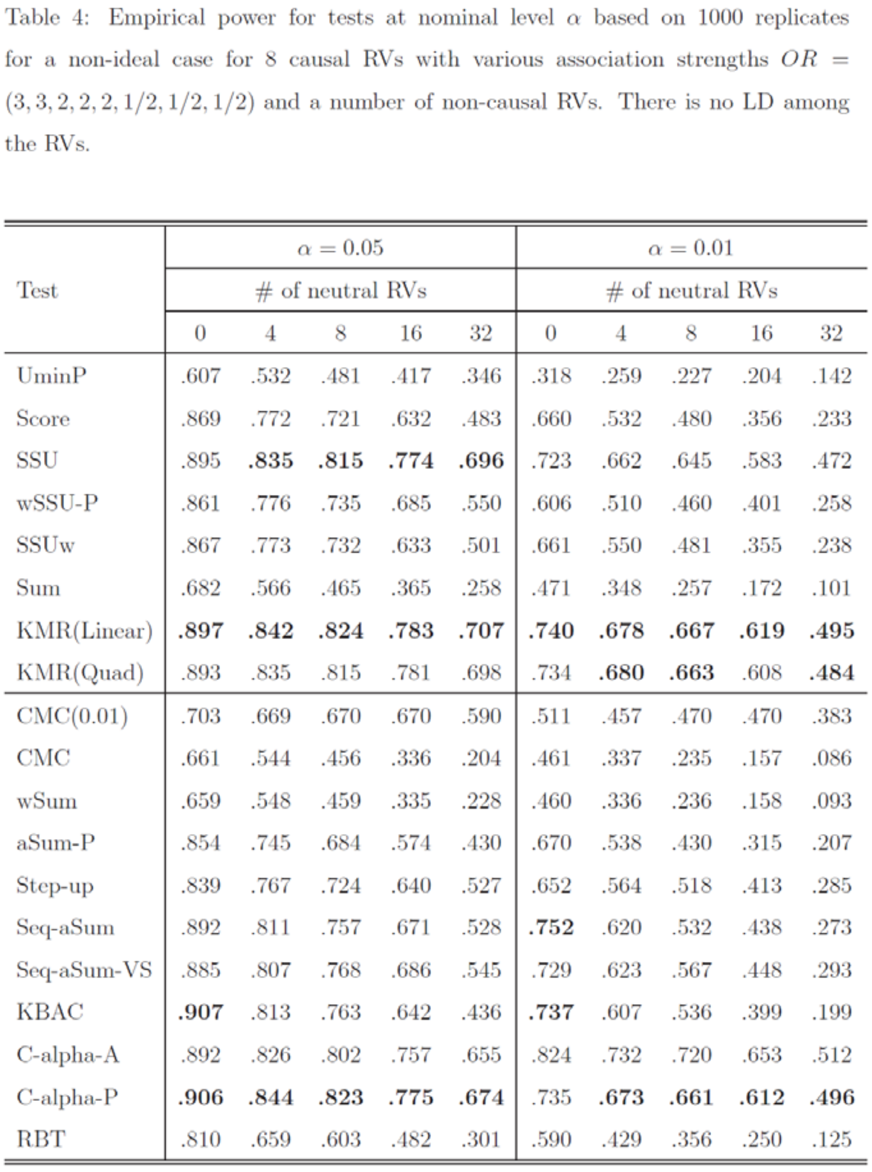


Table 4: Empirical power for tests at nominal level α based on 1000 replicates for a non-ideal case for 8 causal RVs with various association strengths $OR = (3, 3, 2, 2, 2, 1/2, 1/2, 1/2)$ and a number of non-causal RVs. There is no LD among the RVs.

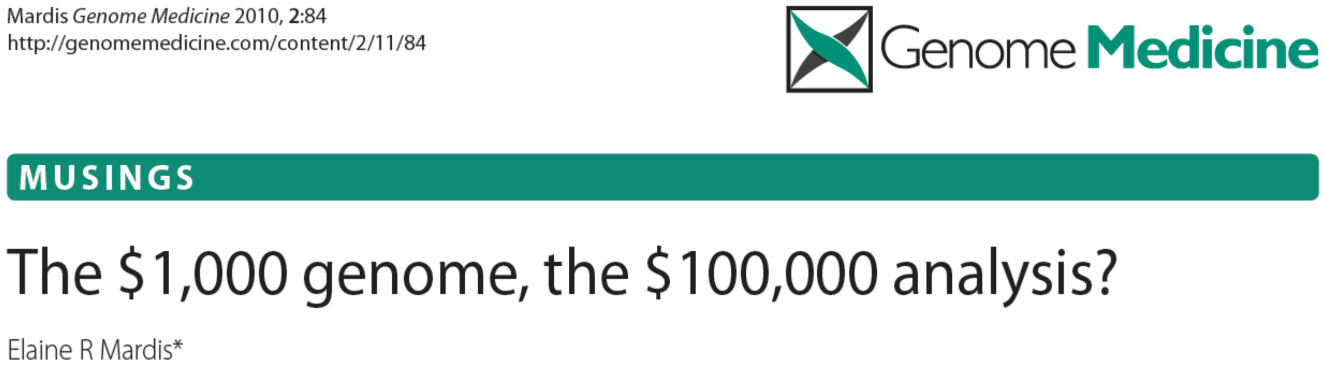
Test	$\alpha = 0.05$					$\alpha = 0.01$				
	# of neutral RVs					# of neutral RVs				
	0	4	8	16	32	0	4	8	16	32
UminP	.607	.532	.481	.417	.346	.318	.259	.227	.204	.142
Score	.869	.772	.721	.632	.483	.660	.532	.480	.356	.233
SSU	.895	**.835**	**.815**	**.774**	**.696**	.723	.662	.645	.583	.472
wSSU-P	.861	.776	.735	.685	.550	.606	.510	.460	.401	.258
SSUw	.867	.773	.732	.633	.501	.661	.550	.481	.355	.238
Sum	.682	.566	.465	.365	.258	.471	.348	.257	.172	.101
KMR(Linear)	**.897**	**.842**	**.824**	**.783**	**.707**	**.740**	**.678**	**.667**	**.619**	**.495**
KMR(Quad)	.893	.835	.815	.781	.698	.734	**.680**	**.663**	.608	**.484**
CMC(0.01)	.703	.669	.670	.670	.590	.511	.457	.470	.470	.383
CMC	.661	.544	.456	.336	.204	.461	.337	.235	.157	.086
wSum	.659	.548	.459	.335	.228	.460	.336	.236	.158	.093
aSum-P	.854	.745	.684	.574	.430	.670	.538	.430	.315	.207
Step-up	.839	.767	.724	.640	.527	.652	.564	.518	.413	.285
Seq-aSum	.892	.811	.757	.671	.528	**.752**	.620	.532	.438	.273
Seq-aSum-VS	.885	.807	.768	.686	.545	.729	.623	.567	.448	.293
KBAC	**.907**	.813	.763	.642	.436	**.737**	.607	.536	.399	.199
C-alpha-A	.892	.826	.802	.757	.655	.824	.732	.720	.653	.512
C-alpha-P	**.906**	**.844**	**.823**	**.775**	**.674**	.735	**.673**	**.661**	**.612**	**.496**
RBT	.810	.659	.603	.482	.301	.590	.429	.356	.250	.125

Additional Issues with Rare Variant Analysis

- Sequencing and Genotyping Errors
- Phasing and Diplotypic Effects
- Stratification
- The Use of *In Silico* Controls (e.g., 1000 Genomes Data)
- Moving Window vs. Annotation-Based Analyses
- Imputation
- Multiple Comparisons

- **Properties of Methods in Different Scenarios!**

Interpreting Genetic Variation is *THE* Issue…



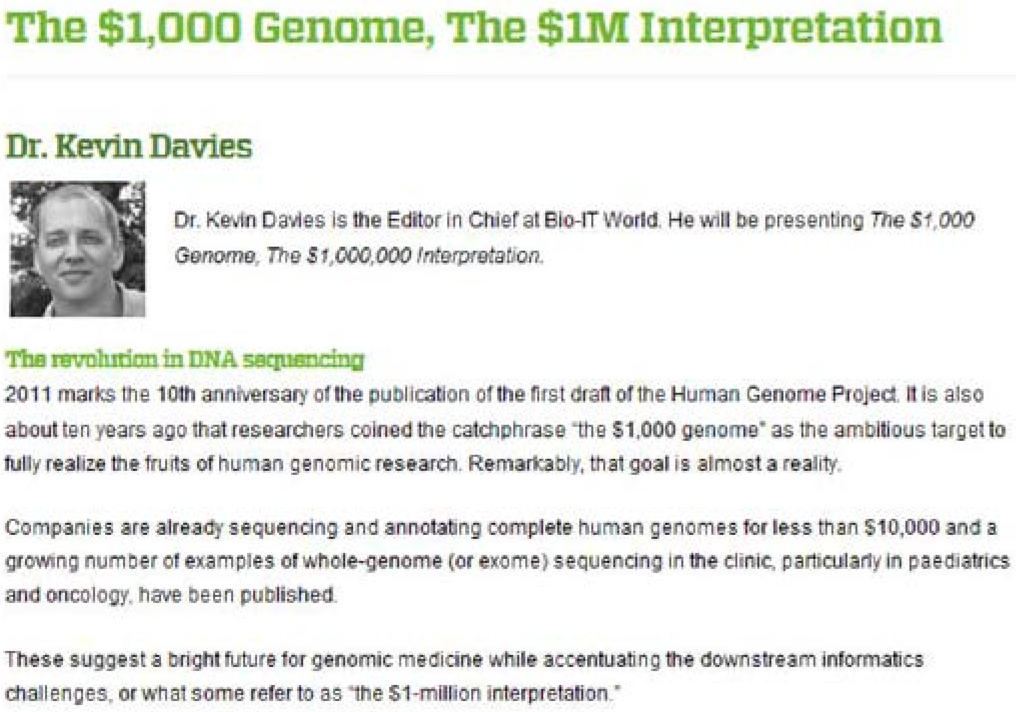
Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome **Medicine**

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*



The $1,000 Genome, The $1M Interpretation

Dr. Kevin Davies

Dr. Kevin Davies is the Editor in Chief at Bio-IT World. He will be presenting *The $1,000 Genome, The $1,000,000 Interpretation.*

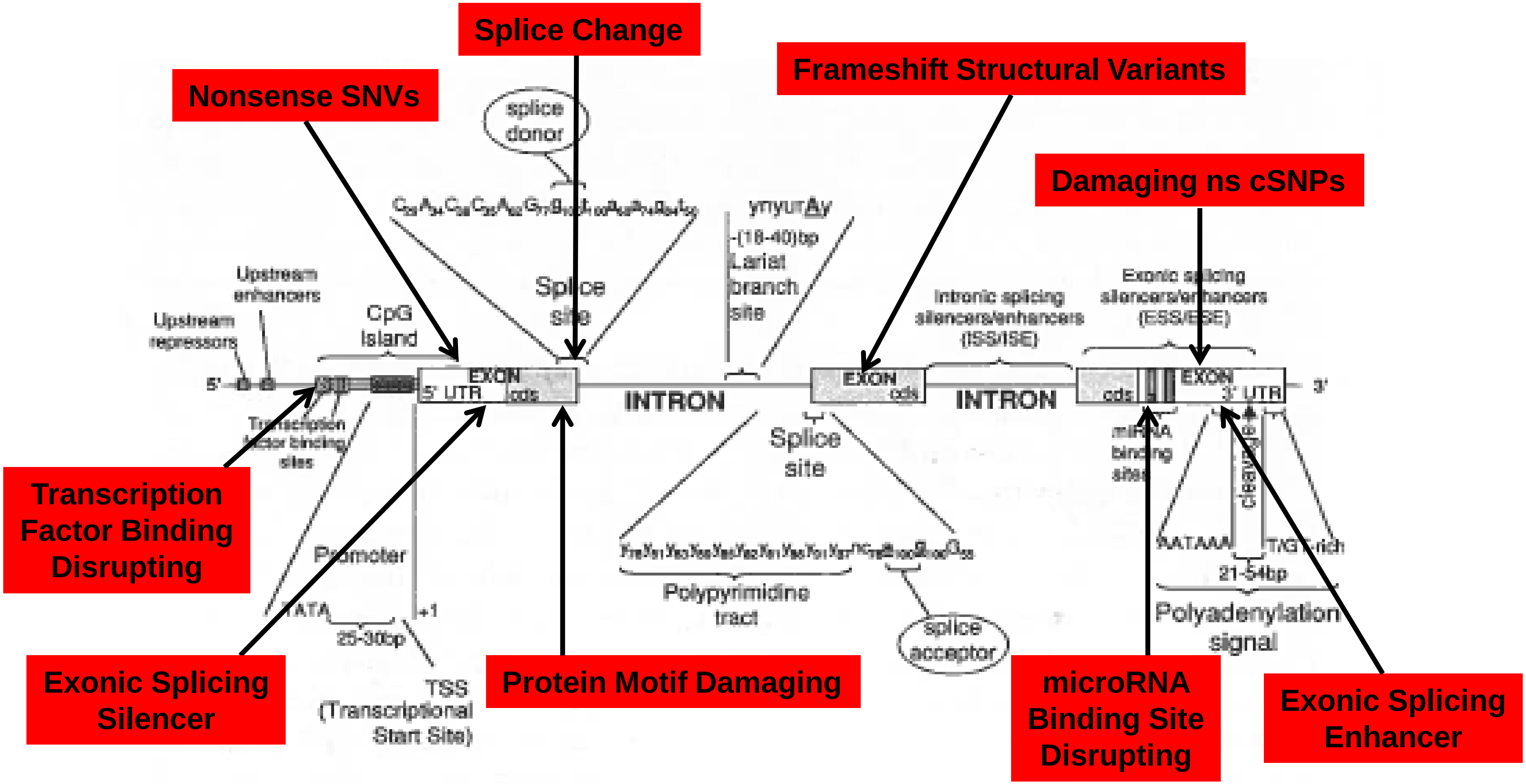
The revolution in DNA sequencing

2011 marks the 10th anniversary of the publication of the first draft of the Human Genome Project. It is also about ten years ago that researchers coined the catchphrase "the $1,000 genome" as the ambitious target to fully realize the fruits of human genomic research. Remarkably, that goal is almost a reality.

Companies are already sequencing and annotating complete human genomes for less than $10,000 and a growing number of examples of whole-genome (or exome) sequencing in the clinic, particularly in paediatrics and oncology, have been published.

These suggest a bright future for genomic medicine while accentuating the downstream informatics challenges, or what some refer to as "the $1-million interpretation."

Functional Annotations: *Bioinformatic* Predictions



Splice Change

Nonsense SNVs

Frameshift Structural Variants

Damaging ns cSNPs

Transcription Factor Binding Disrupting

Exonic Splicing Silencer

Protein Motif Damaging

microRNA Binding Site Disrupting

Exonic Splicing Enhancer

Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

Plumpton and Barnes. “Predictive Functional Analysis of Polymorphisms: An Overview.” in Bioinformatics for Geneticists. Wiley, 2007

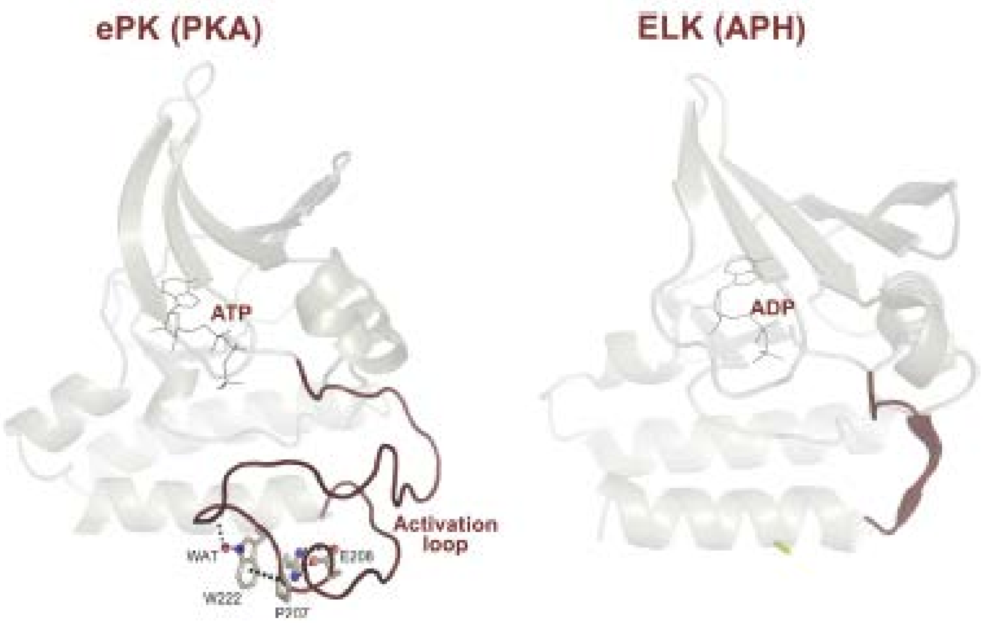
We have developed methodology and tools for comprehensive bioinformatic WGS annot

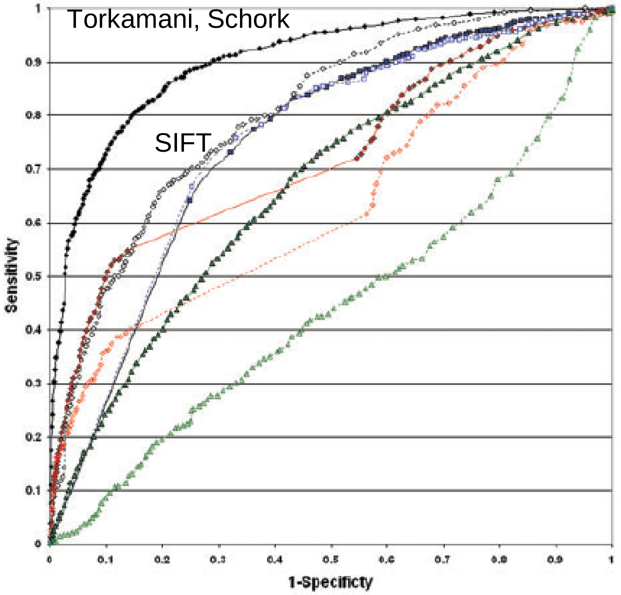
(Schork, Torkamani and colleagues: Bioinformatics 2008, 2009; Cancer Research (2009), Nat Gen Rev (2010), Genomics (2011))

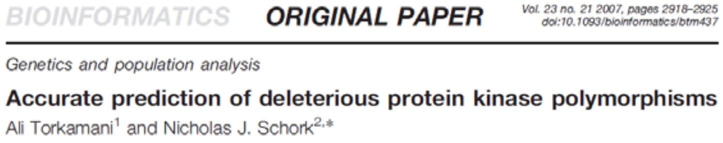
Functional Annotations: The Limits of Conservation

Torkamani, Kannan, Taylor, Schork. *PNAS* 105:9011-9016; 2008

Positions (residues/amino acids) of ~1000 disease causing variants in kinase proteins contrasted with the positions of ~1000 kinase variants not known to cause disease







BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2918–2925
doi:10.1093/bioinformatics/btm437

Genetics and population analysis

Accurate prediction of deleterious protein kinase polymorphisms

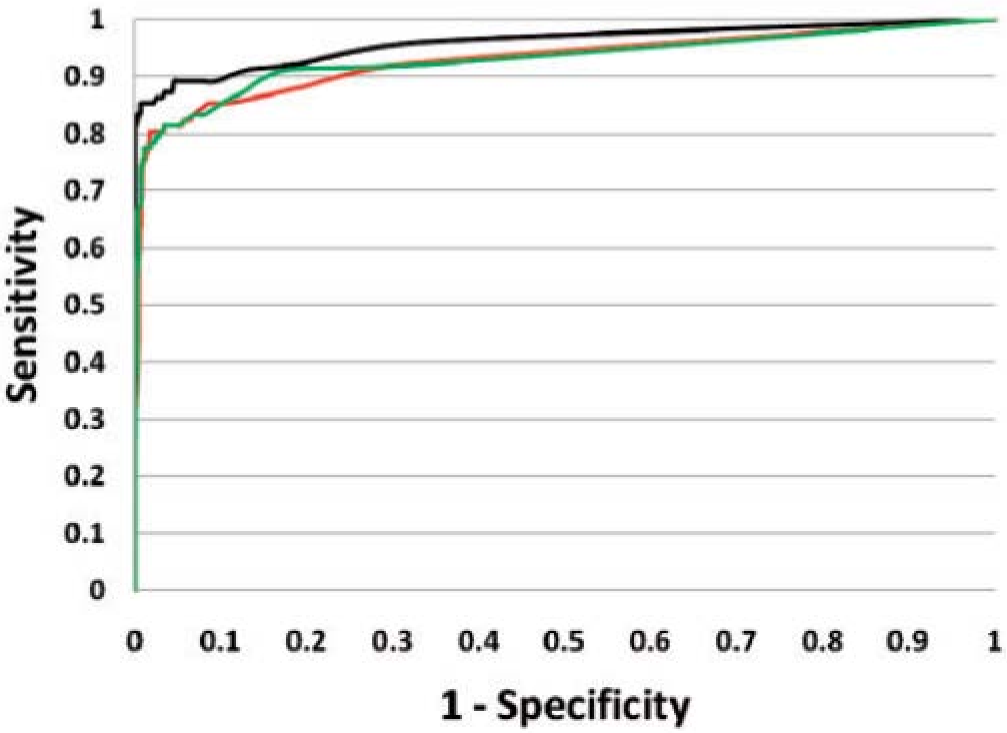
Ali Torkamani[1] and Nicholas J. Schork[2,*]

- **Review**: Lahiry, Torkamani, Schork, Hegele. *Nature Reviews Genetics* 11; 2010
- **Cancer Predictions**: Torkamani, Schork. *Cancer Research* 68; 2008

Functional Annotations: Non-Coding Regions

Torkamani and Schork. *Bioinformatics* 24(16):1787-92; 2008

ENCODE features of the positions of 102 known disease-causing variants contrasted with the positions of 1049 non-disease-causing



Welcome! Please login

Scripps Genome ADVISER | Annotation and Distributed Variant Interpretation Server

| Home | About | Annotate | Results | Tools | Forum | FAQs | Contact |

Start annotating your genomic variants today!

Recent News:

- [illegible] 2013: The story of Lilly Grossman and how SG ADVISER and Dr. Torkamani helped find her disease was featured in the article "[illegible]"
- [illegible] 2013: SG ADVISER was featured in Dr. Torkamani's talk "Finding the Causative Mutation" at the Future of Genomic Medicine VI conference
- [illegible] 2013: SG-ADVISER receives special mention in CLARITY clinical genome interpretation challenge!
- [illegible] 2013: Happy New Year! Welcome to SG Adviser. Please create an account, upload a file for annotation. Use our forum and/or contact us
- [illegible] 2012: Scripps Genome ADVISER is being changed!

CTSA

Scripps

SCRIPPS GENOMIC MEDICINE

A Collaboration of Scripps Health and The Scripps Research Institute

Scripps Translational Science Institute
[illegible]
La Jolla, CA 92037

http://genomics.scripps.edu/ADVISER/Home.jsp

Some features non-assay dependent; e.g., proximity to a TF start or end site

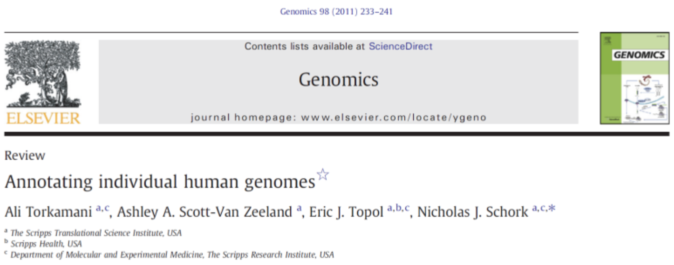
Functional Predictions of Variants in Public Databases

Variant Types	CGI 69	1000 Genomes	dbSNP (130)	HGM
Total number of variants:	**7300345**	**12052647**	**7463633**	**48836**
Total SNPs:	3721410	10462071	3803614	48836
Total Insertions:	1381717	590109	2116683	0
Total Deletions:	1534599	1000467	1144309	0
Total rearrangements:	662619	0	399027	0
Nonsense SNPs:	429	1267	2506	10544
Frameshift Structural Variants:	3716	4911	18127	0
Insertions:	1675	3348	10552	0
Deletions:	1636	1563	7053	0
Rearrangements:	405	0	522	0
Splicing Change Variants:	3021	1630	3833	118
Probably Damaging nscSNPs:	6202	20614	24893	28441
Possibly Damaging nscSNPs:	3061	10130	12189	4145
Protein motif damaging Variants:	4215	8773	20550	21436
TFBS Disrupting Variants:	5274	2749	3590	1
miRNA-BS Disrupting Variants:	555	1412	1233	75
ESE-BS Disrupting Variants:	3917	8177	11410	4738
ESS-BS Disrupting Variants:	2057	3168	4507	1357
Total Likely Functional Variants:	**26775**	**49890**	**75983**	**44412**
Rate of Likely Functional Variants:	0.004	0.004	0.010	0.909

Tools for *In Silico* Functional Prediction of Variants

- Model actual biophysical processes (e.g., protein structure, TF binding)

- Build classifiers using sequence information about the variants



Genomics 98 (2011) 233–241

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

ELSEVIER

GENOMICS

Review

Annotating individual human genomes☆

Ali Torkamani [a,c], Ashley A. Scott-Van Zeeland [a], Eric J. Topol [a,b,c], Nicholas J. Schork [a,c,*]

[a] The Scripps Translational Science Institute, USA
[b] Scripps Health, USA
[c] Department of Molecular and Experimental Medicine, The Scripps Research Institute, USA

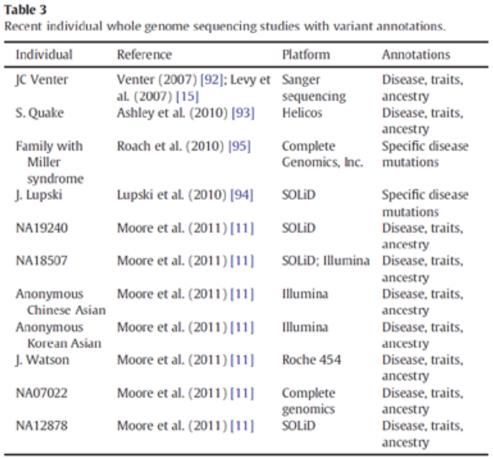
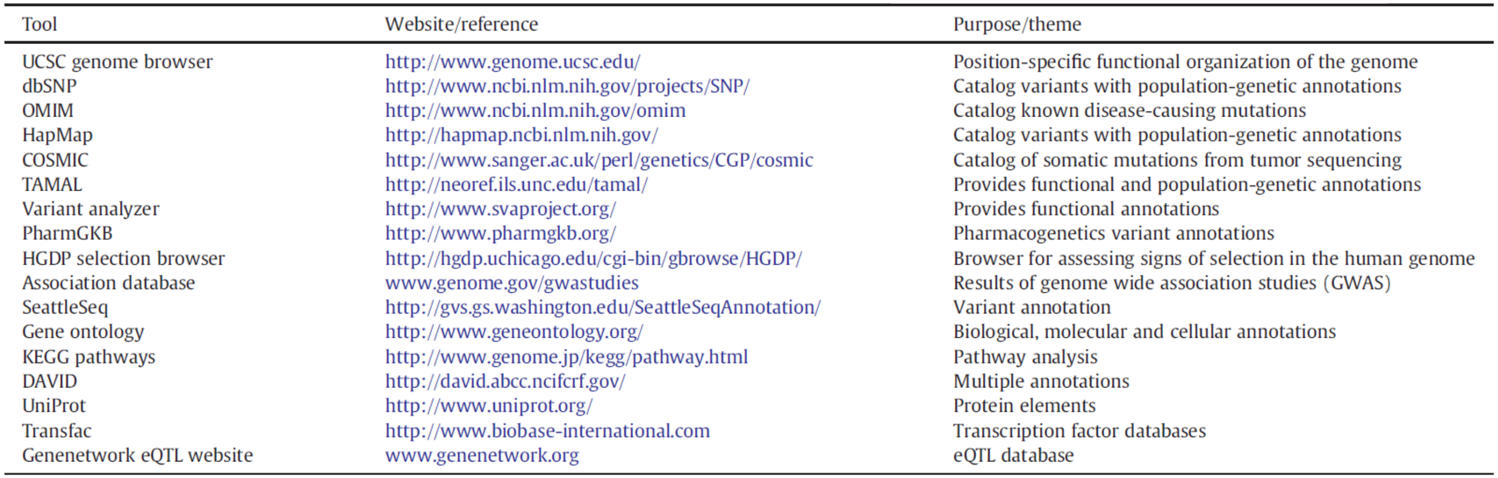


Table 3
Recent individual whole genome sequencing studies with variant annotations.

Individual	Reference	Platform	Annotations
JC Venter	Venter (2007) [92]; Levy et al. (2007) [15]	Sanger sequencing	Disease, traits, ancestry
S. Quake	Ashley et al. (2010) [93]	Helicos	Disease, traits, ancestry
Family with Miller syndrome	Roach et al. (2010) [95]	Complete Genomics, Inc.	Specific disease mutations
J. Lupski	Lupski et al. (2010) [94]	SOLiD	Specific disease mutations
NA19240	Moore et al. (2011) [11]	SOLiD	Disease, traits, ancestry
NA18507	Moore et al. (2011) [11]	SOLiD; Illumina	Disease, traits, ancestry
Anonymous Chinese Asian	Moore et al. (2011) [11]	Illumina	Disease, traits, ancestry
Anonymous Korean Asian	Moore et al. (2011) [11]	Illumina	Disease, traits, ancestry
J. Watson	Moore et al. (2011) [11]	Roche 454	Disease, traits, ancestry
NA07022	Moore et al. (2011) [11]	Complete genomics	Disease, traits, ancestry
NA12878	Moore et al. (2011) [11]	SOLiD	Disease, traits, ancestry

Table 1
Example tools for human variant annotations.



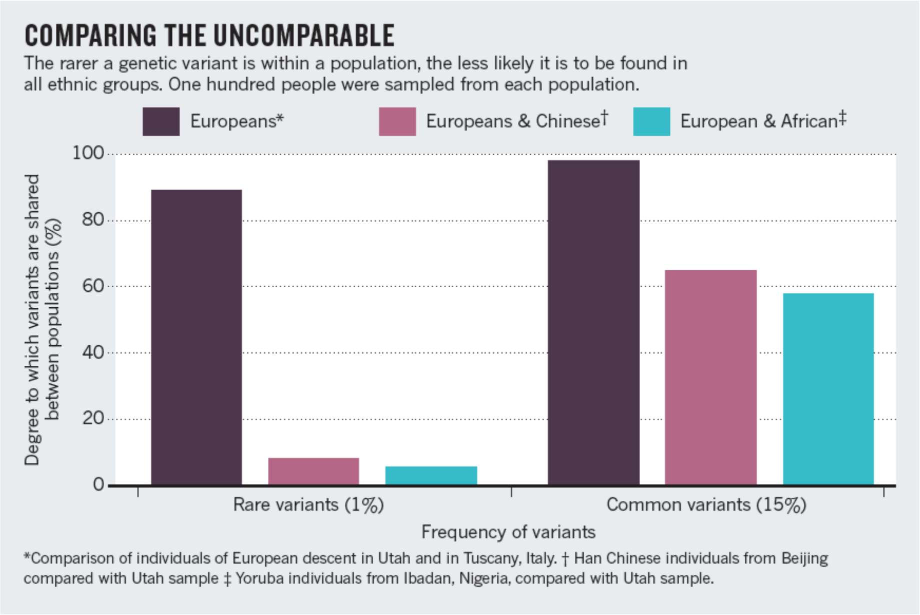
Tool	Website/reference	Purpose/theme
UCSC genome browser	http://www.genome.ucsc.edu/	Position-specific functional organization of the genome
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/	Catalog variants with population-genetic annotations
OMIM	http://www.ncbi.nlm.nih.gov/omim	Catalog known disease-causing mutations
HapMap	http://hapmap.ncbi.nlm.nih.gov/	Catalog variants with population-genetic annotations
COSMIC	http://www.sanger.ac.uk/perl/genetics/CGP/cosmic	Catalog of somatic mutations from tumor sequencing
TAMAL	http://neoref.ils.unc.edu/tamal/	Provides functional and population-genetic annotations
Variant analyzer	http://www.svaproject.org/	Provides functional annotations
PharmGKB	http://www.pharmgkb.org/	Pharmacogenetics variant annotations
HGDP selection browser	http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/	Browser for assessing signs of selection in the human genome
Association database	www.genome.gov/gwastudies	Results of genome wide association studies (GWAS)
SeattleSeq	http://gvs.gs.washington.edu/SeattleSeqAnnotation/	Variant annotation
Gene ontology	http://www.geneontology.org/	Biological, molecular and cellular annotations
KEGG pathways	http://www.genome.jp/kegg/pathway.html	Pathway analysis
DAVID	http://david.abcc.ncifcrf.gov/	Multiple annotations
UniProt	http://www.uniprot.org/	Protein elements
Transfac	http://www.biobase-international.com	Transcription factor databases
Genenetwork eQTL website	www.genenetwork.org	eQTL database

- Statistical RANKING algorithms are need to prioritize variants in a study

14 JULY 2011 | VOL 475 | NATURE | 163

Genomics for the world

Medical genomics has focused almost entirely on those of European descent. Other ethnic groups must be studied to ensure that more people benefit, say **Carlos D. Bustamante**, **Esteban González Burchard** and **Francisco M. De La Vega**.



COMPARING THE UNCOMPARABLE

The rarer a genetic variant is within a population, the less likely it is to be found in all ethnic groups. One hundred people were sampled from each population.

*Comparison of individuals of European descent in Utah and in Tuscany, Italy. † Han Chinese individuals from Beijing compared with Utah sample ‡ Yoruba individuals from Ibadan, Nigeria, compared with Utah sample.

Example Issues:

- Determining individual ancestry or locus/allele-specific ancestry
- Unmatched (based on ancestry) cases and controls in a GWAS-seq = false positives
- Reference panel for determining the ‘novelty’ of a variant involves different ancestry

Population-Level Phenomena and Global Diversity

Africa

- greater diversity
- selection has washed away some older deleterious alleles
- less homozygosity for older deleterious alleles

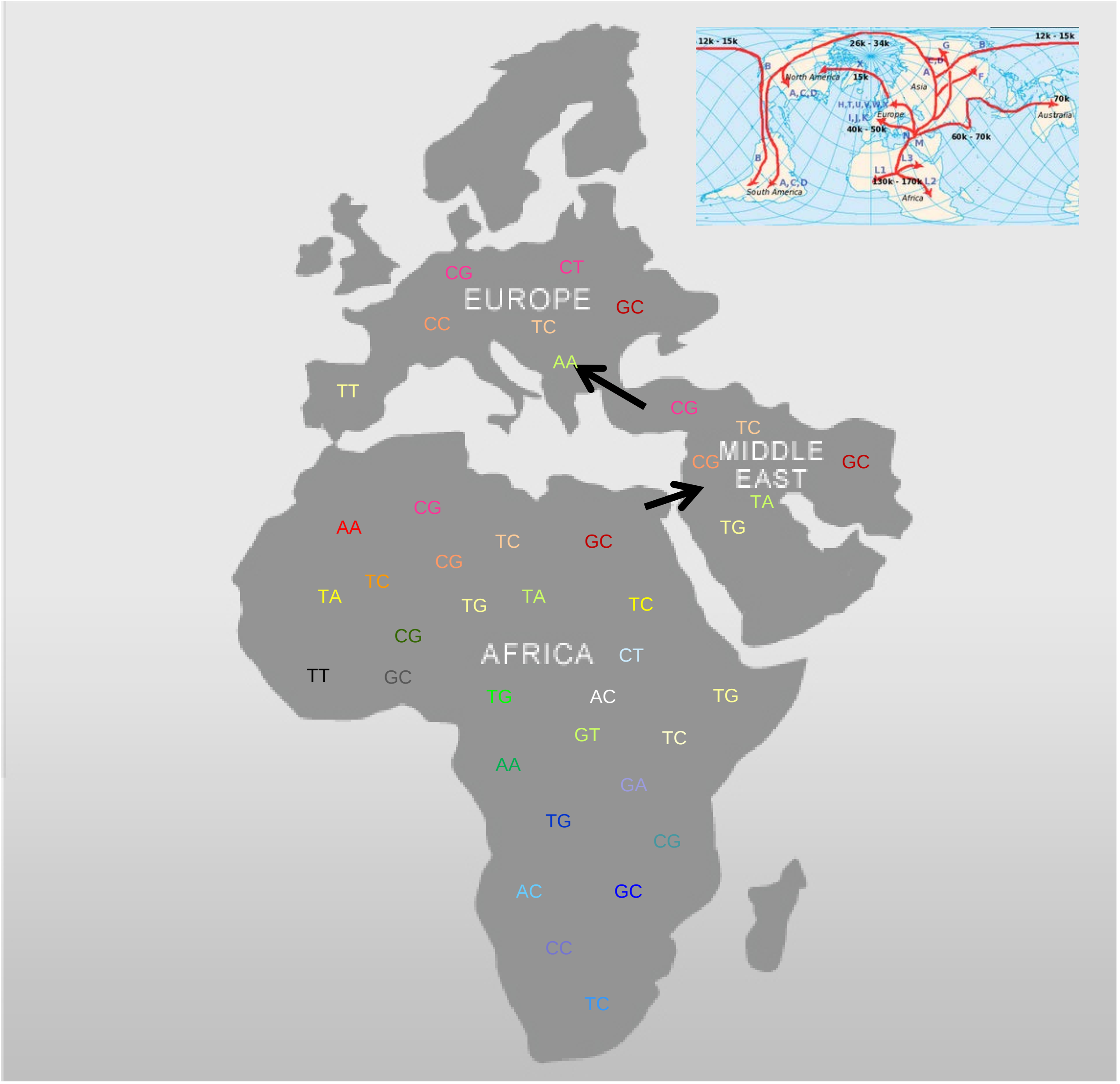
Middle East

- only migrant genotypes represented
- early bottleneck created

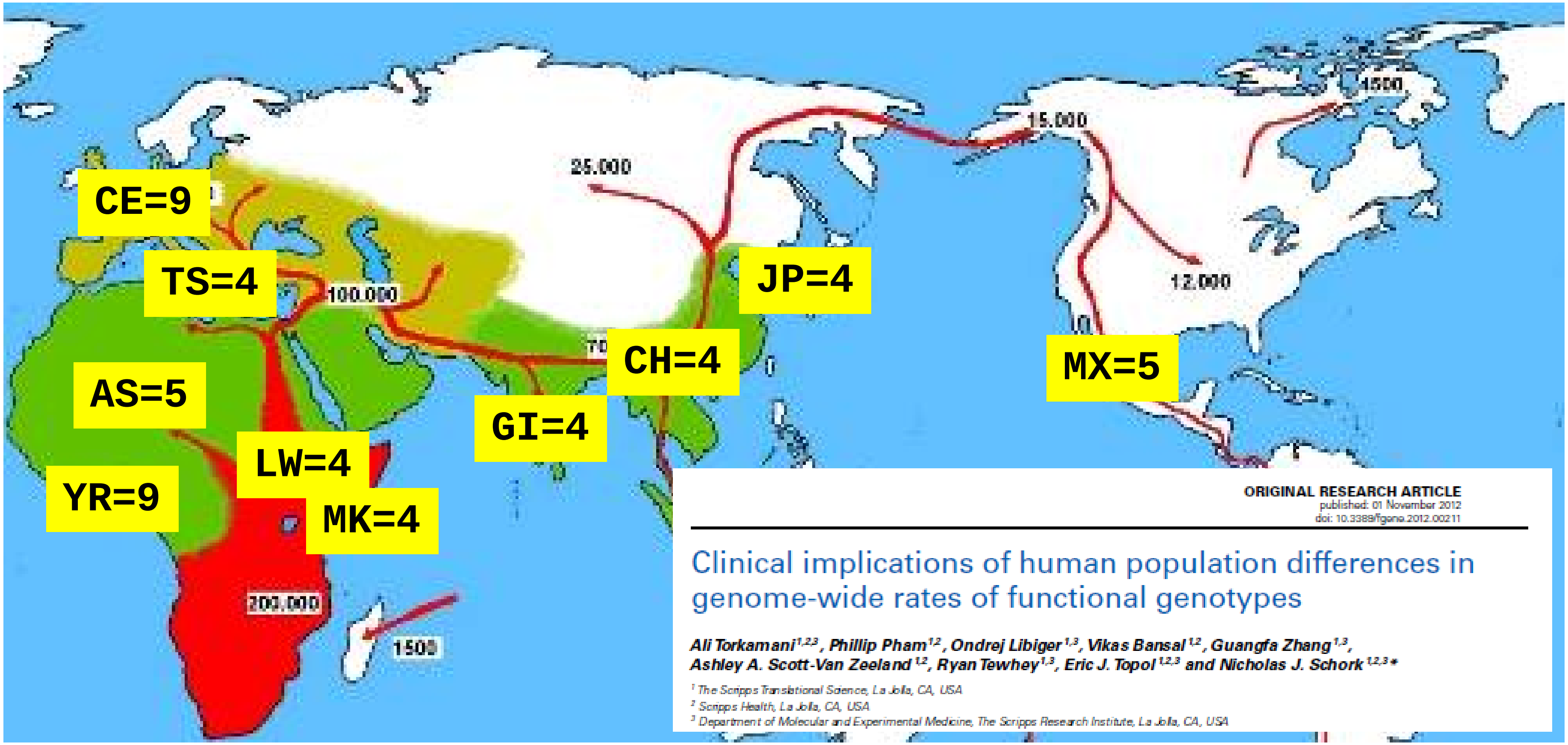
Europe

- only migrant genotypes represented
- not enough time for selection to wash away deleterious genotypes
- homozygosity for deleterious alleles is greater

Lohmueller et al. Nature. 2008 451:994-7



Available Whole Genome Sequences for Diversity Studies



ORIGINAL RESEARCH ARTICLE
published: 01 November 2012
doi: 10.3389/fgene.2012.00211

Clinical implications of human population differences in genome-wide rates of functional genotypes

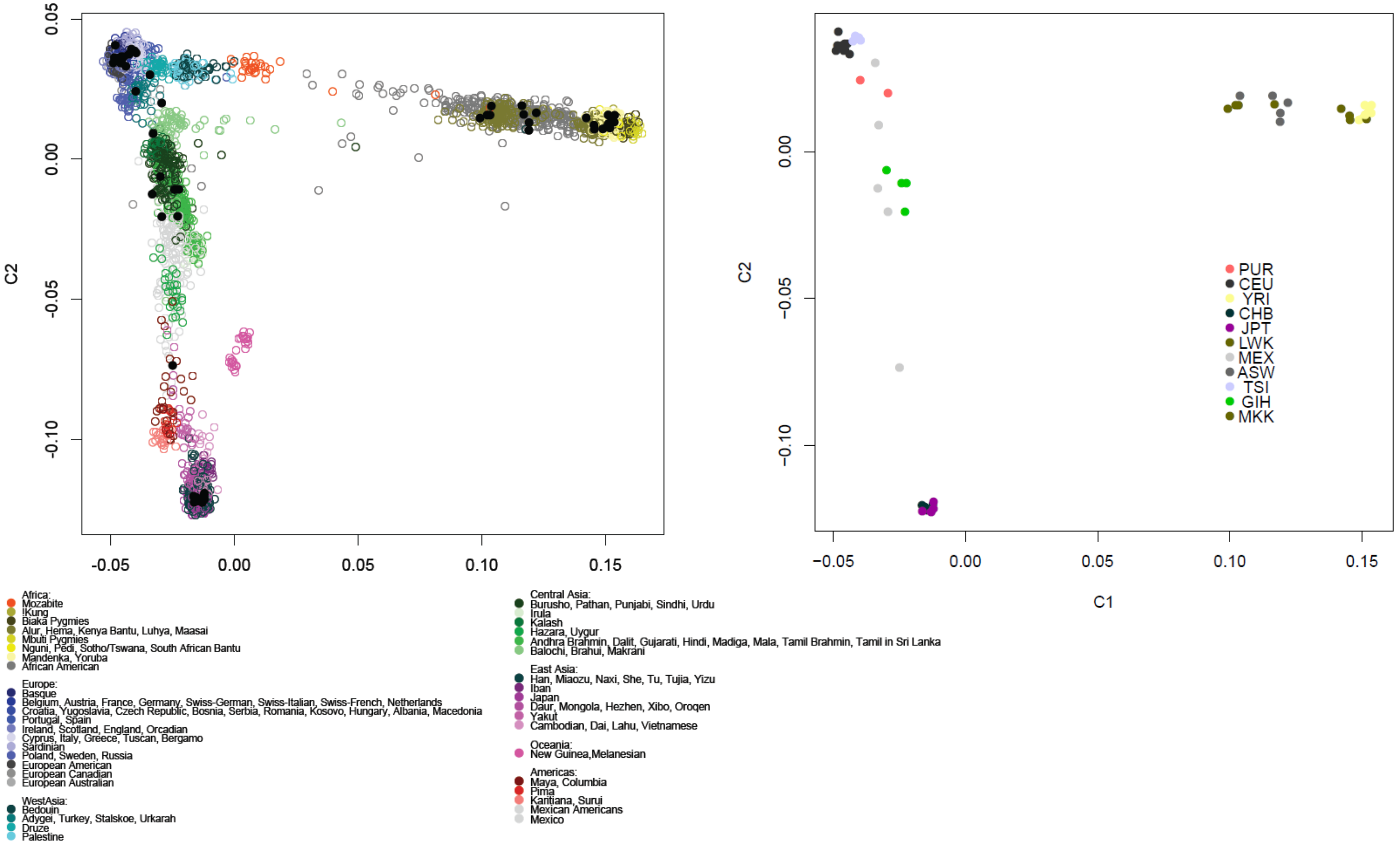
Ali Torkamani[1,2,3], Phillip Pham[1,2], Ondrej Libiger[1,3], Vikas Bansal[1,2], Guangfa Zhang[1,3], Ashley A. Scott-Van Zeeland[1,2], Ryan Tewhey[1,3], Eric J. Topol[1,2,3] and Nicholas J. Schork[1,2,3]*

[1] The Scripps Translational Science, La Jolla, CA, USA
[2] Scripps Health, La Jolla, CA, USA
[3] Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, USA

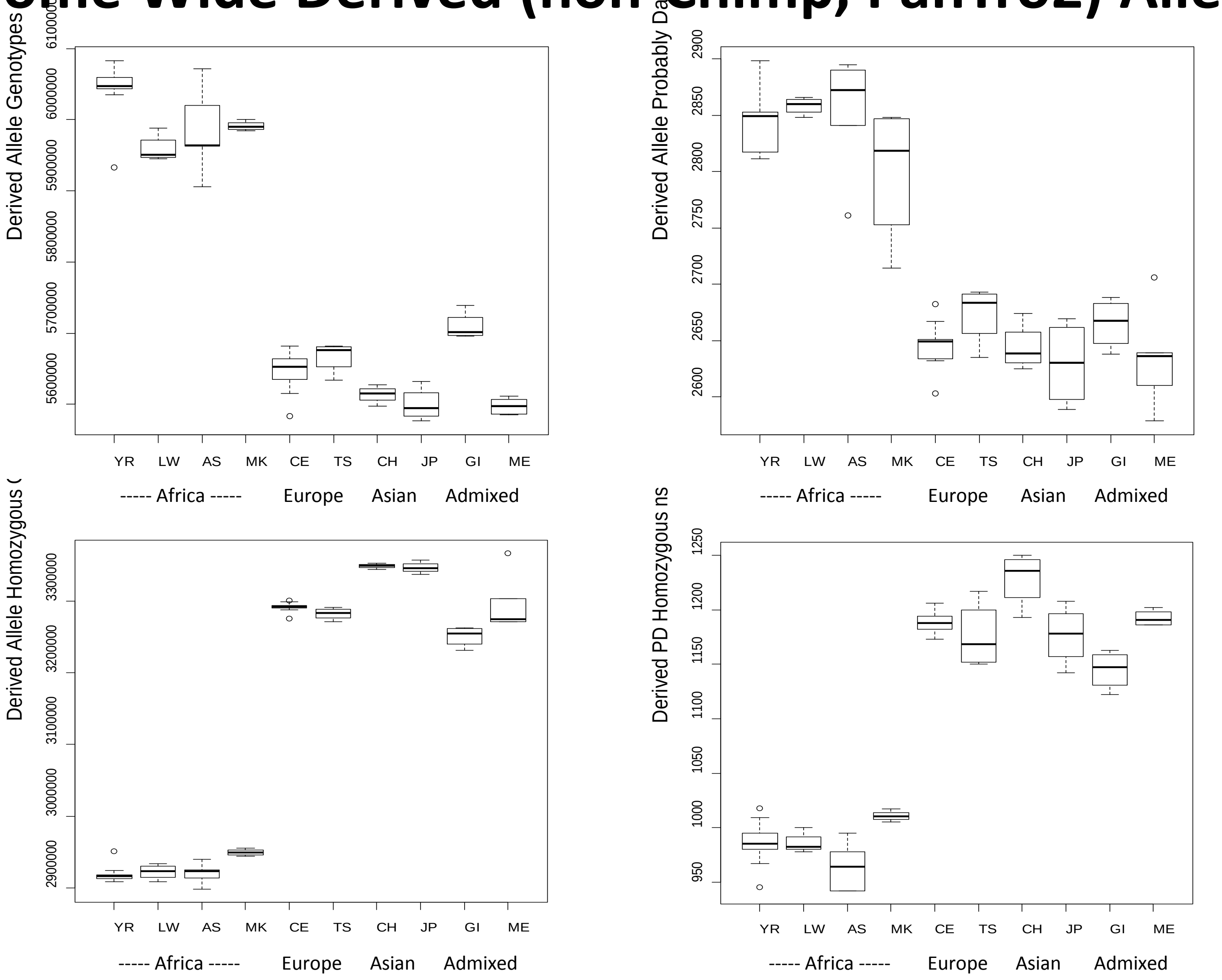
1. Identify all derived (i.e., non-chimp genome) alleles in each genome (30,000,000+)
2. Functionally characterize all variants (coding and non-coding) via bioinformatics analysis
3. Compare total number and rates per genome of functional variants across categories

4. Address the question of whether functional genomic diversity plagues ‘filtering’ strategy

52 Unrelated Individual Whole Genome Variants (CGI)



Genome Wide Derived (non-Chimp, PanTro2) Alleles



Historical bottlenecks, migrations, founder effects, random inbreeding, lack of time for selective pressure to operate, etc. have left an imprint on contemporary global standing variation and homozygosity in non-African populations on a WGS functional variant basis (extends the work of Bustamante et al.)

Population Specific Alleles (Unique to Each Population)

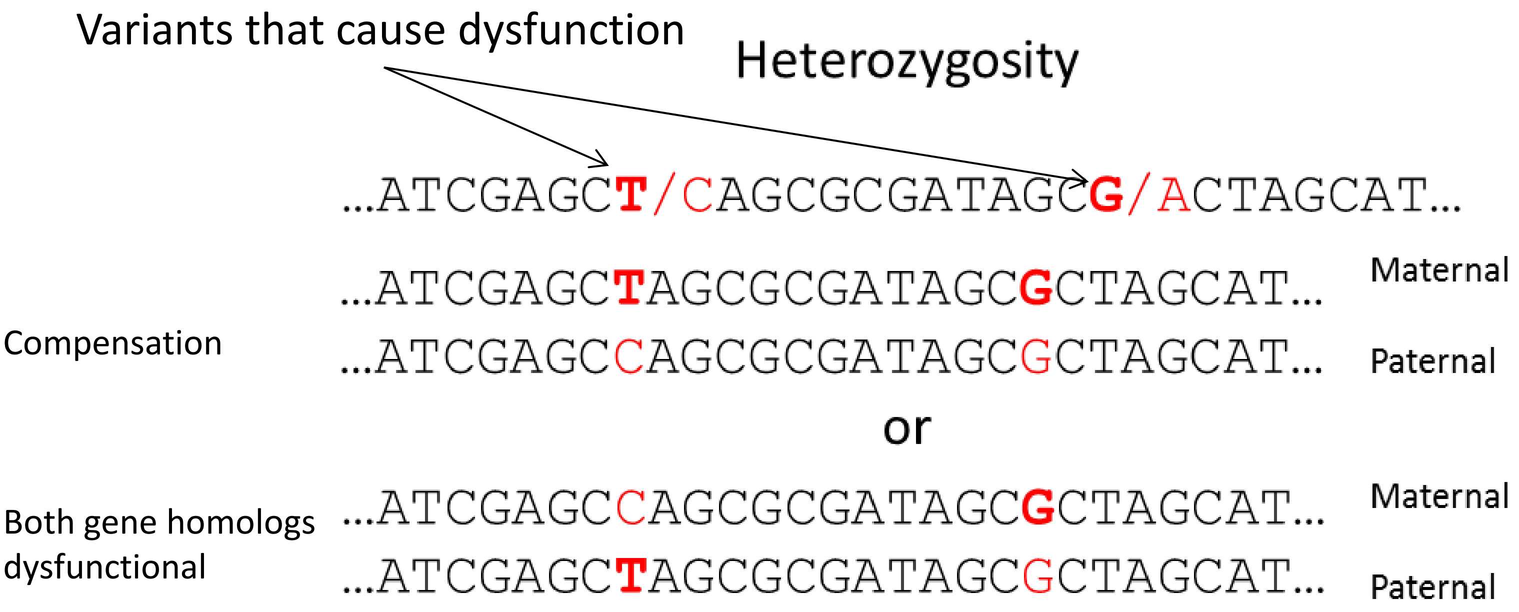
		Populations			z-test p-values		
Variant Type	**Label**	**AFR**	**EUR**	**ASN**	**AFR vs EUR**	**AFR vs ASN**	**EUR vs ASN**
Total number of variants:		7614850	2024886	1294731			
Nonsense SNPs rate	1	0.500	0.840	0.842	6.931E-09	6.329E-07	4.910E-01
Frameshift Structural Variants rate	2	1.663	3.008	2.989	1.597E-34	6.239E-25	4.621E-01
Frameshift Insertion rate	3	0.657	1.274	1.383	6.368E-19	1.089E-18	2.006E-01
Frameshift Deletion rate	4	0.879	1.417	1.352	3.877E-12	1.584E-07	3.102E-01
Frameshift Rearrangement rate	5	0.127	0.316	0.255	2.614E-09	2.228E-04	1.572E-01
Splicing Change Variants rate	6	1.707	2.514	2.379	4.655E-14	7.112E-08	2.223E-01
Probably Damaging nscSNPs rate	7	10.103	15.472	15.602	1.136E-91	4.578E-69	3.853E-01
Possibly Damaging nscSNPs rate	8	5.991	7.744	8.233	7.313E-19	3.064E-21	6.111E-02
Protein motif damaging Variants rate	9	4.104	6.311	6.581	2.612E-39	3.043E-35	1.726E-01
TFBS Disrupting Variants rate	10	2.793	4.173	4.063	7.493E-69	2.764E-42	1.785E-01
miRNA-BS Disrupting Variants rate	11	0.948	1.170	1.081	2.405E-03	7.715E-02	2.286E-01
ESE-BS Disrupting Variants rate	12	5.835	7.260	7.283	1.696E-13	2.840E-10	4.689E-01
ESS-BS Disrupting Variants rate	13	2.460	3.013	2.865	6.435E-06	3.539E-03	2.232E-01
Total Likely Functional Variant rate	14	23.718	34.906	35.436	8.999E-170	1.234E-132	2.128E-01

Frequencies of funct pop spec variants: Greater in non-Africans

Highly significant AFR vs. non-AFR

- The rate of novel functional variants (not just homozygous) is significantly higher in non-Africans
- The rate is uniformly higher across ALL functional classes, not just ns cSNPs
- Selection has had less time to ‘purifiy’ the European and Asian population (i.e., replicated Lohmuller et al.)

Diploidy and Compound Heterozygosity (CH)



Variants that cause dysfunction

Heterozygosity

...ATCGAGC**T**/CAGCGCGATAGC**G**/ACTAGCAT...

Compensation

...ATCGAGC**T**AGCGCGATAGC**G**CTAGCAT... Maternal

...ATCGAGCCAGCGCGATAGCGCTAGCAT... Paternal

or

Both gene homologs dysfunctional

...ATCGAGCCAGCGCGATAGC**G**CTAGCAT... Maternal

...ATCGAGC**T**AGCGCGATAGCGCTAGCAT... Paternal

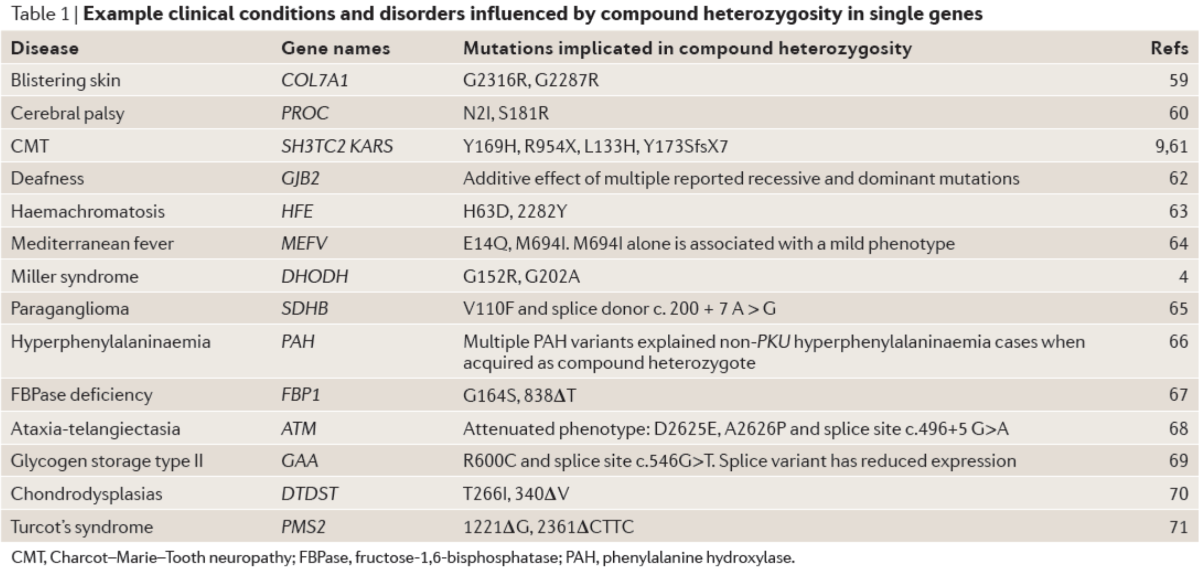


Table 1 | **Example clinical conditions and disorders influenced by compound heterozygosity in single genes**

Disease	Gene names	Mutations implicated in compound heterozygosity	Refs
Blistering skin	*COL7A1*	G2316R, G2287R	59
Cerebral palsy	*PROC*	N2I, S181R	60
CMT	*SH3TC2 KARS*	Y169H, R954X, L133H, Y173SfsX7	9,61
Deafness	*GJB2*	Additive effect of multiple reported recessive and dominant mutations	62
Haemachromatosis	*HFE*	H63D, 2282Y	63
Mediterranean fever	*MEFV*	E14Q, M694I. M694I alone is associated with a mild phenotype	64
Miller syndrome	*DHODH*	G152R, G202A	4
Paraganglioma	*SDHB*	V110F and splice donor c. 200 + 7 A > G	65
Hyperphenylalaninaemia	*PAH*	Multiple PAH variants explained non-*PKU* hyperphenylalaninaemia cases when acquired as compound heterozygote	66
FBPase deficiency	*FBP1*	G164S, 838ΔT	67
Ataxia-telangiectasia	*ATM*	Attenuated phenotype: D2625E, A2626P and splice site c.496+5 G>A	68
Glycogen storage type II	*GAA*	R600C and splice site c.546G>T. Splice variant has reduced expression	69
Chondrodysplasias	*DTDST*	T266I, 340ΔV	70
Turcot's syndrome	*PMS2*	1221ΔG, 2361ΔCTTC	71

CMT, Charcot–Marie–Tooth neuropathy; FBPase, fructose-1,6-bisphosphatase; PAH, phenylalanine hydroxylase.

Nature Reviews Genetics | AOP, published online 8 February 2011;

OPINION

The importance of phase information for human genomics

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork

Nature Reviews Genetics | AOP, published online 8 February 2011;

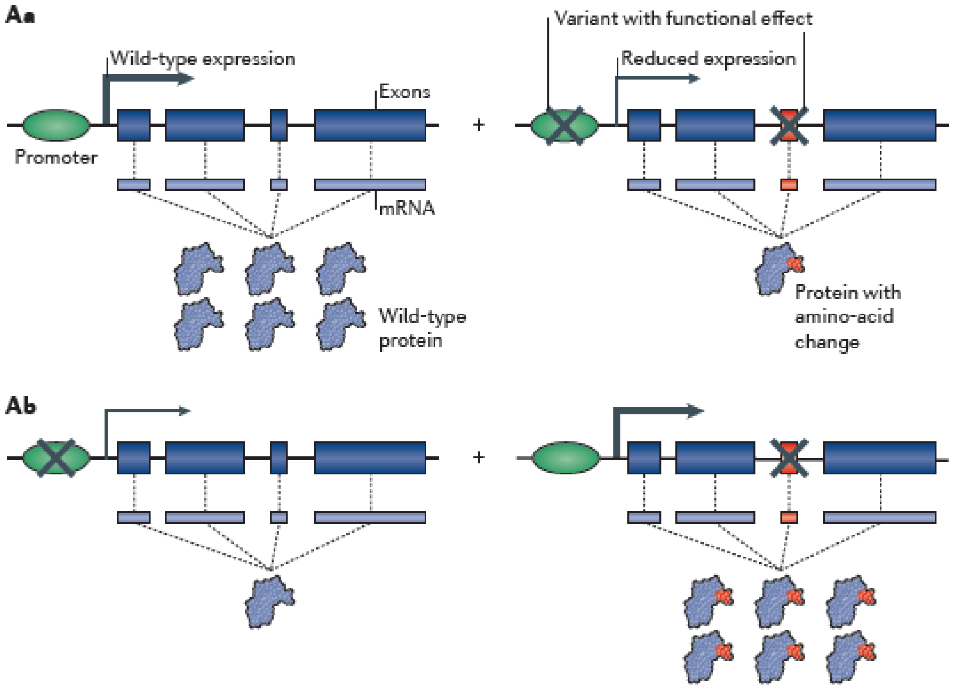
OPINION

The importance of phase information for human genomics

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork

- Can sense be made of the effect of multiple genic variations without knowing phase?
- Most studies simply tally the number of non-reference alleles at singular loci
- Determining phase is not trivial via population/*de novo* assembly algorithms

Maternal Homolog Paternal Homolog Diplotypic Effect (Phenotype)



Normal Function

Haploinsufficiency

Normal Function

Haploinsufficiency

Unique Phenotype

Tewhey et al. (2011)

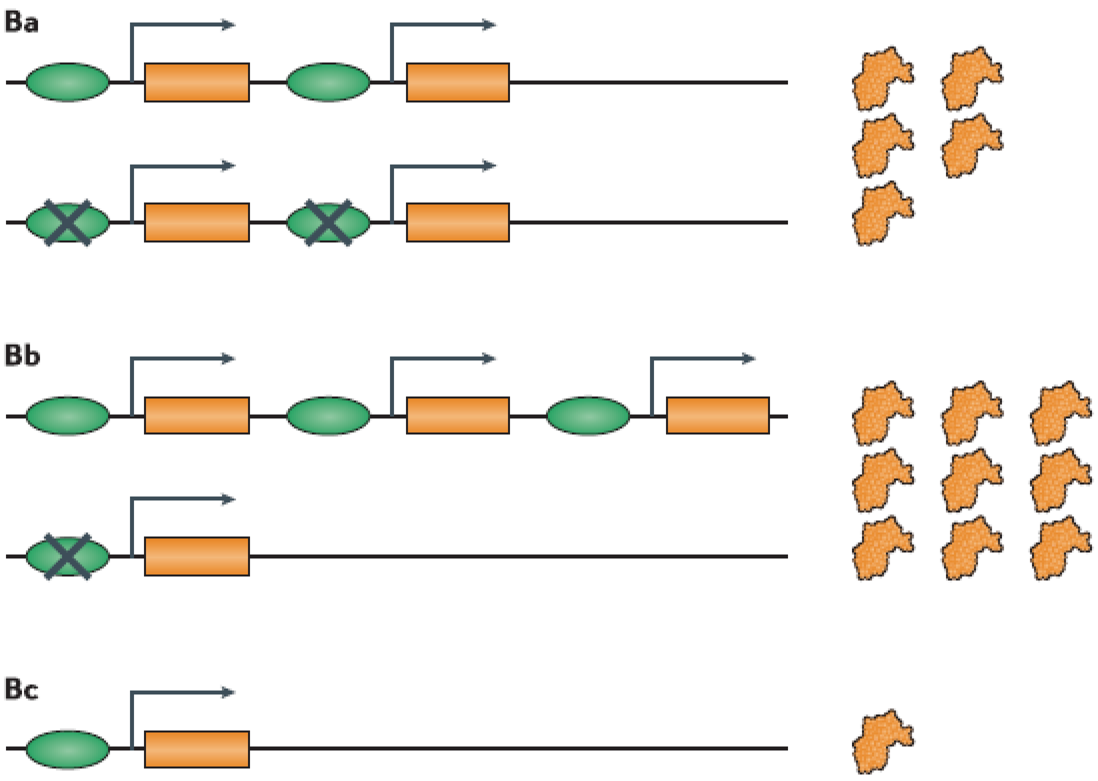
4 Gene Copies but 3 Different Scenarios

Uncovering the roles of rare variants in common disease through whole-genome sequencing

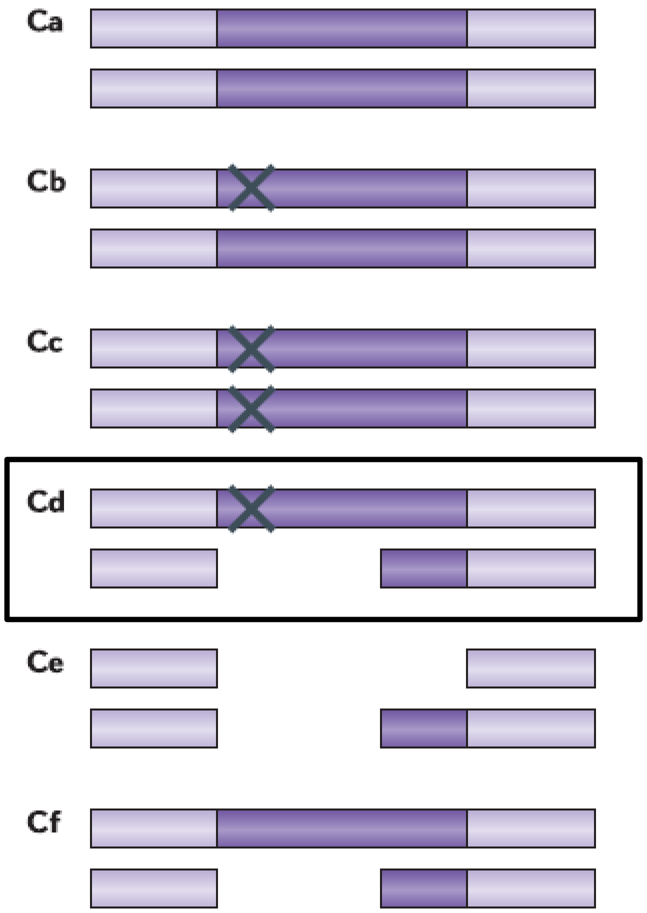
Elizabeth T. Cirulli and David B. Goldstein

VOLUME 11 | JUNE 2010 | 415

NATURE REVIEWS | GENETICS



Copy Number Variations



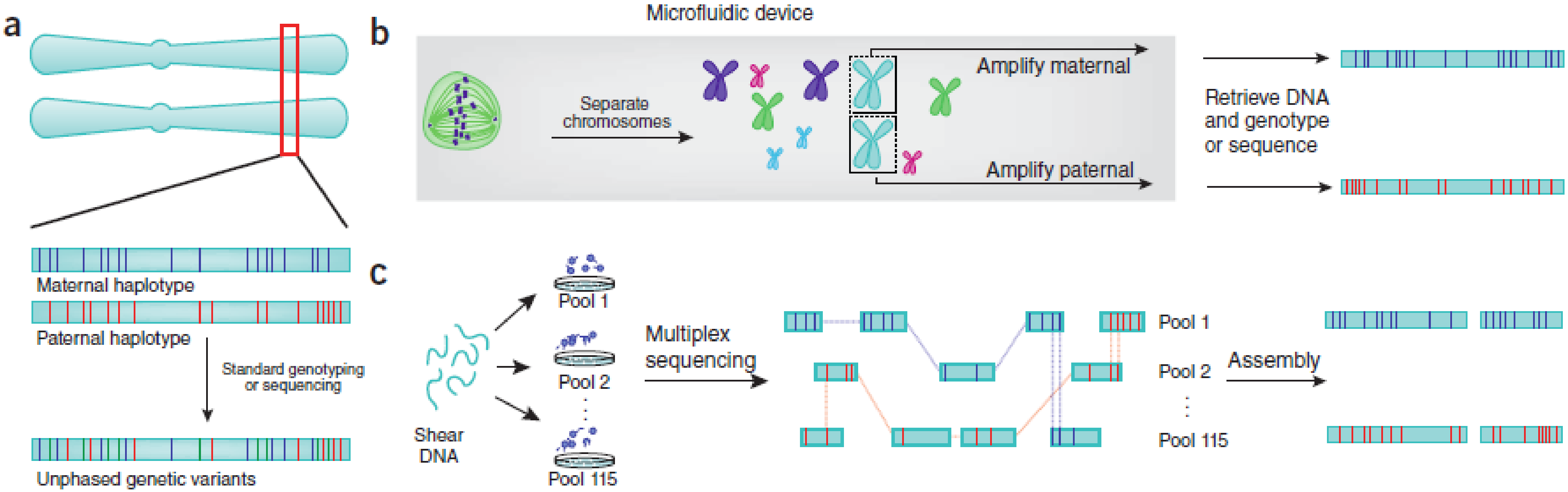
‘Unmasking’ via Deletions

Tewhey et al. (2011)

Phasing for Assessing ‘Diplomics’ Phenomena

Approaches to Resolving Phase

- Sequencing parents/relatives
- Population-based phasing (and imputation)
- Assembly of sequencing reads
- Separate chromosomes prior to sequencing



The next phase in human genetics

Vikas Bansal, Ryan Tewhey, Eric J. Topol & Nicholas J. Schork

Experimental haplotyping of whole genomes is now feasible, enabling new studies aimed at linking sequence variation to human phenotypes and disease susceptibility.^

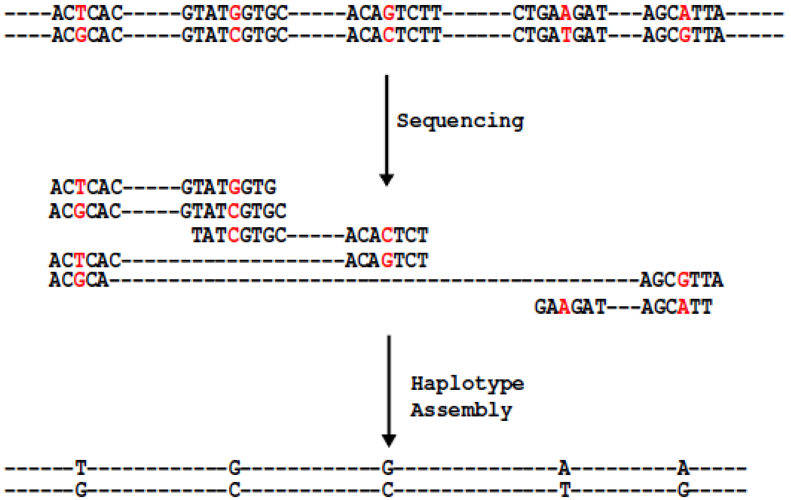
NATURE BIOTECHNOLOGY VOLUME 29 NUMBER 1 JANUARY 2011

NGS Assembly-Based Haplotyping and Phasing

BIOINFORMATICS

HapCUT: An Efficient and Accurate Algorithm for the Haplotype Assembly Problem

Vikas Bansal[1], Vineet Bafna,[1]



```
----ACTCAC-----GTATGGTGC-----ACAGTCTT------CTGAAGAT---AGCATTA-----
----ACGCAC-----GTATCGTGC-----ACACTCTT------CTGATGAT---AGCGTTA-----
```

Sequencing

```
    ACTCAC-----GTATGGTG
    ACGCAC-----GTATCGTGC
                TATCGTGC-----ACACTCT
    ACTCAC-------------------ACAGTCT
    ACGCA----------------------------------------------AGCGTTA
                                              GAAGAT---AGCATT
```

Haplotype Assembly

```
------T-----------G-----------G-------------A---------A-----
------G-----------C-----------C-------------T---------G-----
```

Correct phase

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

Switching Error

AAAAAAAAAAAAAAAAATTTTTTTTTTTTTTTTTTTTT

TTTTTTTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAA

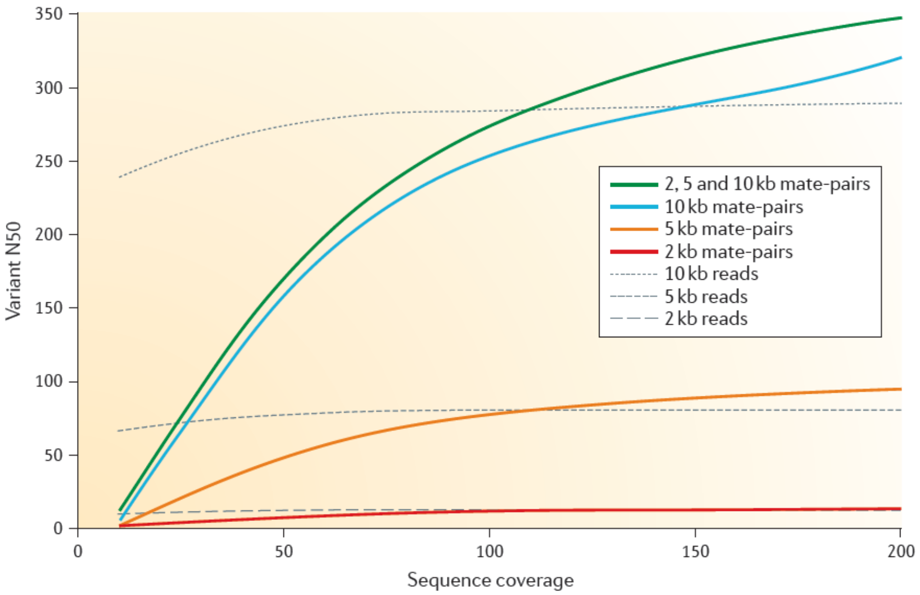


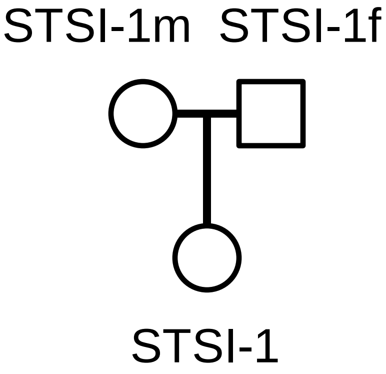
Figure 3 | **Phase reconstruction using mate-pair information.** Simulated 100 bp mate-pair read coverage of various depths (sequence (fold) coverage, *x*-axis) for chromosome 1 of a Yoruban individual. All simulations were done using SNP calls (for chromosome 1) for the Yoruban individual NA19240, obtained from the 1000 Genomes project (released December 2008). Paired-end reads were simulated with the starting position of one read, chosen consistently at random on the chromosome, and the insert length sampled from a normal distribution with a given mean insert length (2, 5 or 10 kb) and standard deviation equal to 10% of the mean. For each simulation experiment, we constructed a graph with nodes corresponding to the heterozygous SNPs and edges corresponding to reads that cover multiple variants. The N50 was calculated using the number of variants in each connected component of this graph that correspond to the phased haplotype blocks. The vN50 is defined as the point at which half of the heterozygous loci of the chromosome are contained in contigs with the vN50 or greater number of variants. Mate-pair libraries outperform reads of the same length because the size distribution of the insert consists of lengths greater than 10 kb, allowing for longer connections than are possible with single reads alone. The software used in the simulation studies is available from the Polymorphism Research Laboratory (see Further information).

The importance of phase information for human genomics

Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork

NATURE REVIEWS | GENETICS

Functional Variant Analysis of the Genomes of a Trio



COMPREHENSIVE ANNOTATION OF AN ENTIRE HUMAN DIPLOID GENOME
Ali Torkamani*, Vikas Bansal*, Ondrej Libiger, Phillip Pham, Ashley Van Zeeland,
Guangfa Zhang, Ryan Tewhey, Eric J. Topol, Nicholas J. Schork *(in review)*

Individual	Seq (Gb)	SNVs	Novel	Ins	Novel	Del	Novel
Child (STSI-1)	121.9	3163286	210730	145411	56028	156147	61544
Mother (STSI-1m)	137.2	3229588	216800	155150	59506	166060	64507
Father (STSI-1f)	138.4	3236815	216996	157779	60310	169006	65139
Combined	-	4469443	419783	268714	125258	295595	135390

- Sequencing and variant calling by Complete Genomics, Inc.

- In house phasing algorithms + **functional annotations of all variants**

- Primary analyses: catalog instances of potential functional compound heterozygosity

Torkamani et al. (in review)

Phasing and Analysis Approach

Phasing algorithm:

- Use Mendel’s laws to phase heterozygous variants

- For triply heterozygous variants, leverage population phasing/neighboring variants

- 4125865 phased SNVs (92%) and 348835 phased indels (87%)

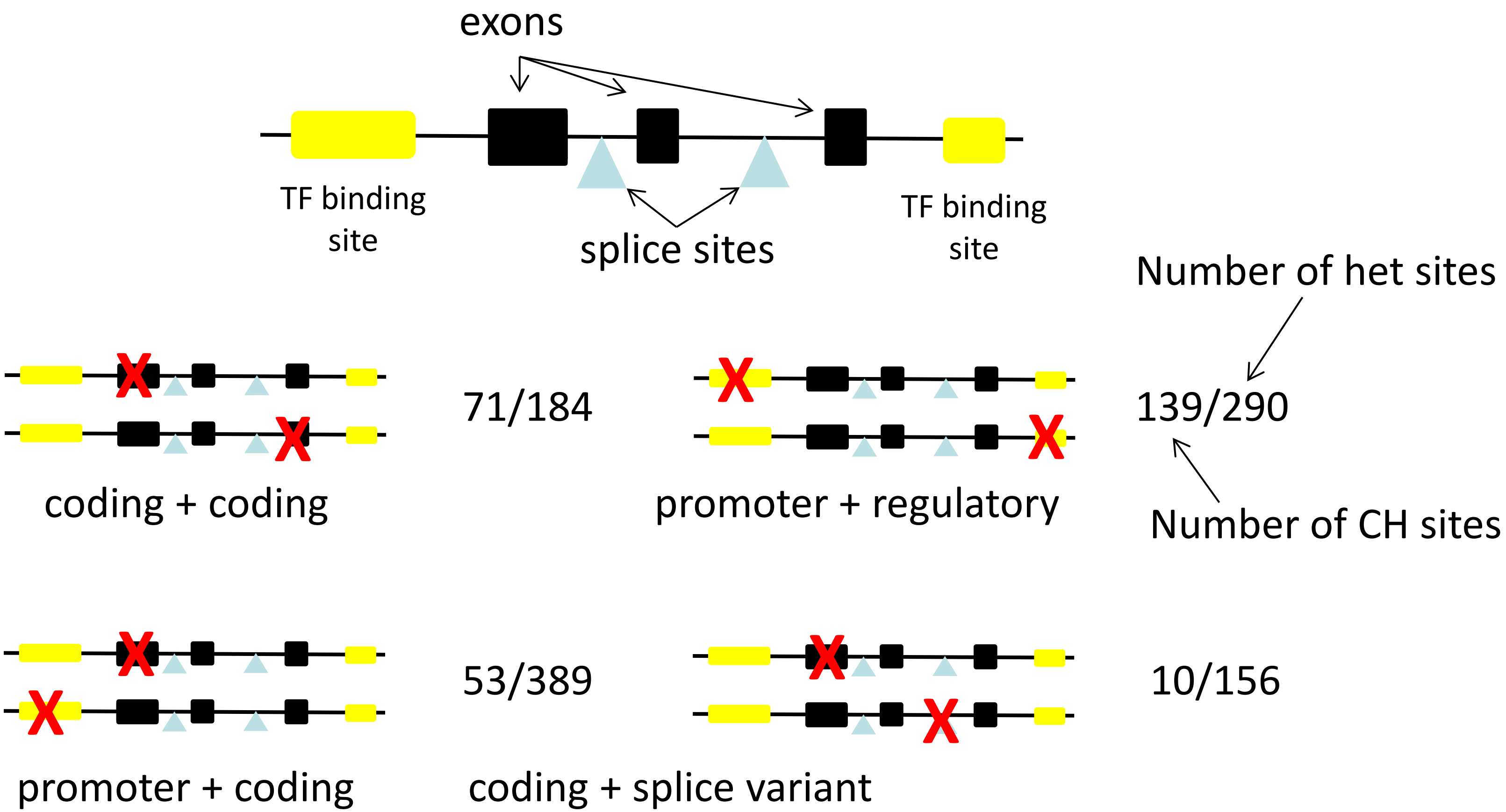
- Variants not in databases and de novo variants/sequencing errors can’t be phased

After phasing all variants:

1. Annotate positions of all variants (Human Genome hg18)
2. Predict likely functional effect of variants using bioinfomatics pipeline
3. Assign disease risk alleles from association study databases
4. Explore regions of high heterozygosity/nucleotide content differences between homologous chromosomes

Torkamani et al. (in review)

Genes Harboring Likely Functional CH Sites



exons

TF binding site

splice sites

TF binding site

Number of het sites

coding + coding

71/184

promoter + regulatory

139/290

Number of CH sites

promoter + coding

53/389

coding + splice variant

10/156

- Substantial number of potentially functionally significant CH sites in genomes
- RNA sequencing and eQTL studies are underway to assess these functionally

Torkamani et al. (in review)

DNA Sequencing Clinical Success Stories: Idiopathic Diseases

Nicholas Volker (PMID: 21173700)

The Beery Twins
(PMID: 21677200)

Madsen siblings; Miller Syndrome
(PMID: 20220176)

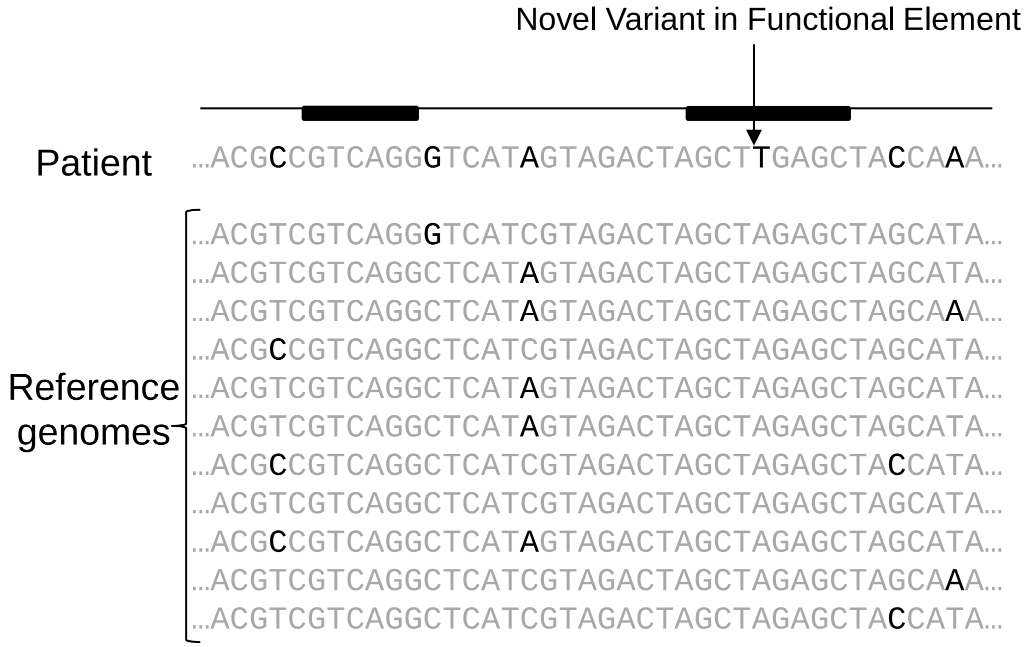
Dr. James Lupski (CMT)
(PMID: 20220177)

- Idiopathic conditions: defy conventional diagnostic categories, treatment unresponsive
- Sequencing the genomes of individuals with idiopathic conditions could shed light on origins
- Variants could be inherited in complex ways (e.g., compound heterozygotes) or be *de novo*
- Finding the pathogenic or causative variants among the many ‘candidates’ is problematic
- Strategies based on WGS, the use of reference genomes and bioinformatics tools exist

‘Filtering’ Strategies: Reference Genomes + Bioinformatics

Two reasonable(?) assumptions:

1. The pathogenic variant(s) is ‘novel’ (i.e., unique to the patient)
2. The effect of the variant is pronounced enough to be characterized bioinformatically

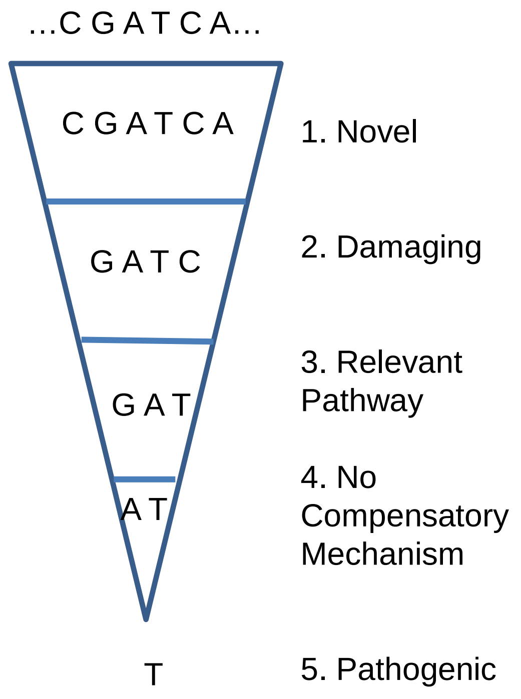


Novel Variant in Functional Element

Patient ...ACGCCGTCAGGGTCATAGTAGACTAGCTTGAGCTACCAAA...

Reference genomes

...ACGTCGTCAGGGTCATCGTAGACTAGCTAGAGCTAGCATA...
...ACGTCGTCAGGCTCATAGTAGACTAGCTAGAGCTAGCATA...
...ACGTCGTCAGGCTCATAGTAGACTAGCTAGAGCTAGCAAA...
...ACGCCGTCAGGCTCATCGTAGACTAGCTAGAGCTAGCATA...
...ACGTCGTCAGGCTCATAGTAGACTAGCTAGAGCTAGCATA...
...ACGTCGTCAGGCTCATAGTAGACTAGCTAGAGCTAGCATA...
...ACGCCGTCAGGCTCATCGTAGACTAGCTAGAGCTACCATA...
...ACGTCGTCAGGCTCATCGTAGACTAGCTAGAGCTAGCATA...
...ACGCCGTCAGGCTCATAGTAGACTAGCTAGAGCTAGCATA...
...ACGTCGTCAGGCTCATCGTAGACTAGCTAGAGCTAGCAAA...
...ACGTCGTCAGGCTCATCGTAGACTAGCTAGAGCTACCATA...



...CGATCA...

CGATCA — 1. Novel

GATC — 2. Damaging

GAT — 3. Relevant Pathway

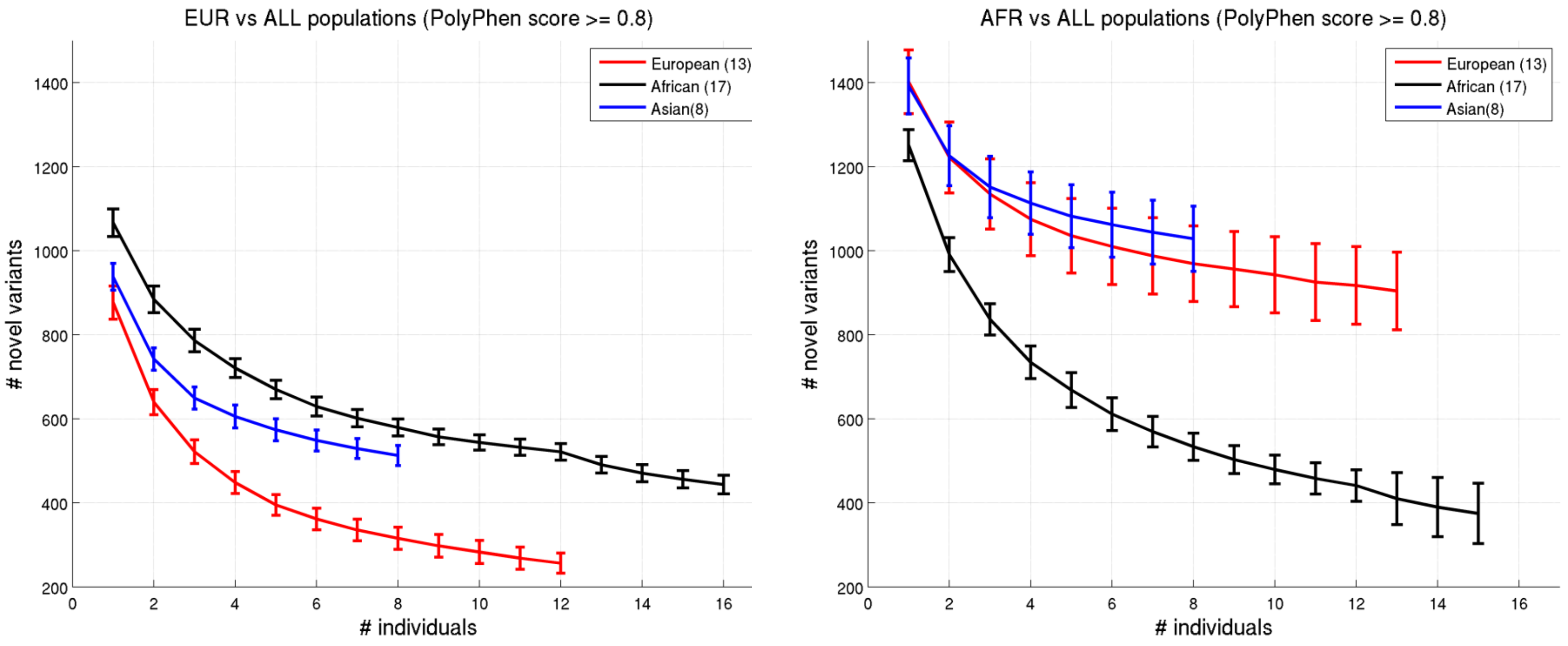
AT — 4. No Compensatory Mechanism

T — 5. Pathogenic

- What bioinformatic tools should be used for functionality? Does it make a difference?
- What reference populations for determining novelty should be used? Does it matter?

Filters to Identify Causative Variants in Single Genomes

- We 'implanted' known disease causative variants with Polyphen2 score > 0.8 in genomes
- Determined the observed number of novel functional variants with different reference



- Determining the novelty of a variant requires ancestry-appropriate reference genomes…
- This has implications for clinical studies as well as rare variant, GWAS-seq studies

Genetic Networks and Network Analysis

NATURE | VOL 411 | 3 MAY 2001

brief communications

Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.

H. Jeong*, S. P. Mason†, A.-L. Barabási*,
Z. N. Oltvai†

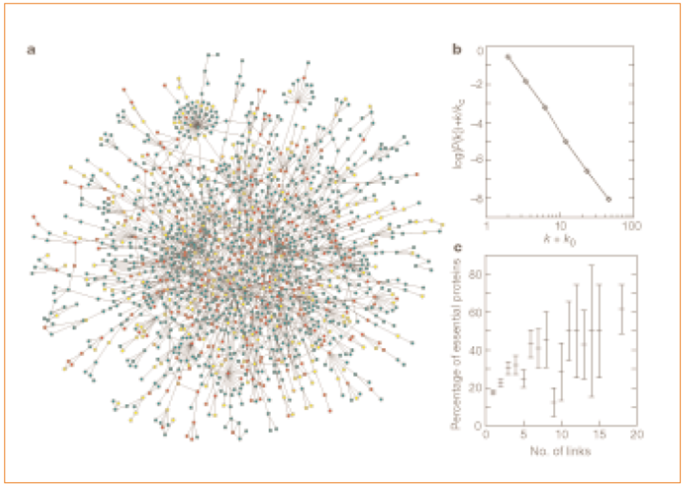


Figure 1 Characteristics of the yeast proteome. **a,** Map of protein–protein interactions. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown). **b,** Connectivity distribution *P*(*k*) of interacting yeast proteins, giving the probability that a given protein interacts with *k* other proteins. The exponential cut-off[6] indicates that the number of proteins with more than 20 interactions is slightly less than expected for pure scale-free networks. In the absence of data on the link directions, all interactions have been considered as bidirectional. The parameter controlling the short-length scale correction has value $k_0 \approx 1$. **c,** The fraction of essential proteins with exactly *k* links versus their connectivity, *k*, in the yeast proteome. The list of 1,572 mutants with known phenotypic profile was obtained from the Proteome database[13]. Detailed statistical analysis, including $r = 0.75$ for Pearson's linear correlation coefficient, demonstrates a positive correlation between lethality and connectivity. For additional details, see http://www.nd.edu/~networks/cell.

Cell *144*, March 18, 2011 ©2011

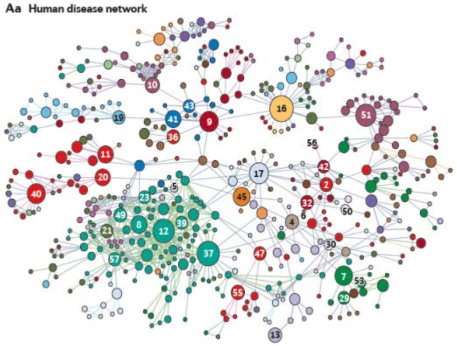
Interactome Networks and Human Disease

Marc Vidal,[1,2,*] Michael E. Cusick,[1,2] and Albert-László Barabási[1,3,4,*]

NATURE REVIEWS | GENETICS VOLUME 12 | JANUARY 2011

Network medicine: a network-based approach to human disease

Albert-László Barabási[*‡§], *Natali Gulbahce*[*‡||] *and Joseph Loscalzo*[§]

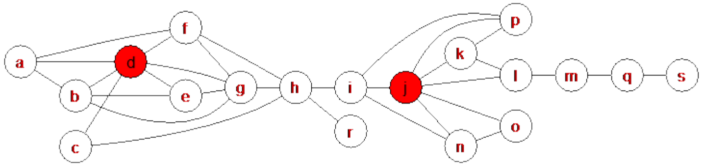


How can one leverage network information in drug matching algorithms?

Network Centrality Measures

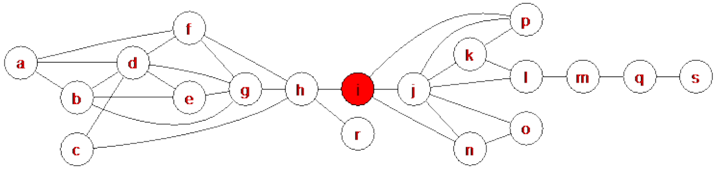
Degree Centrality

- Number of nodes connected to a given node
- How well a node is connected; direct influence



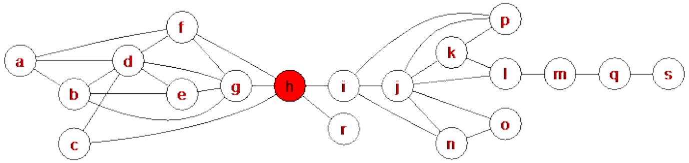
Closeness Centrality

- Sum of shortest distance (path) to all other nodes
- Inverse measure of centrality



Betweenness Centrality

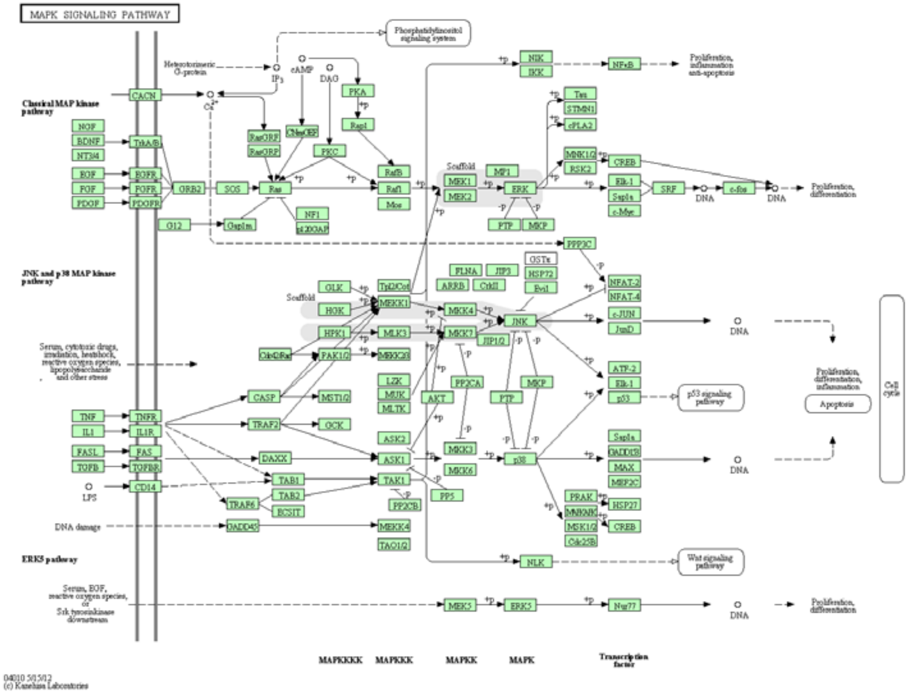
- Frequency that *node*=shortest path between 2 nodes
- Control of communication between other nodes



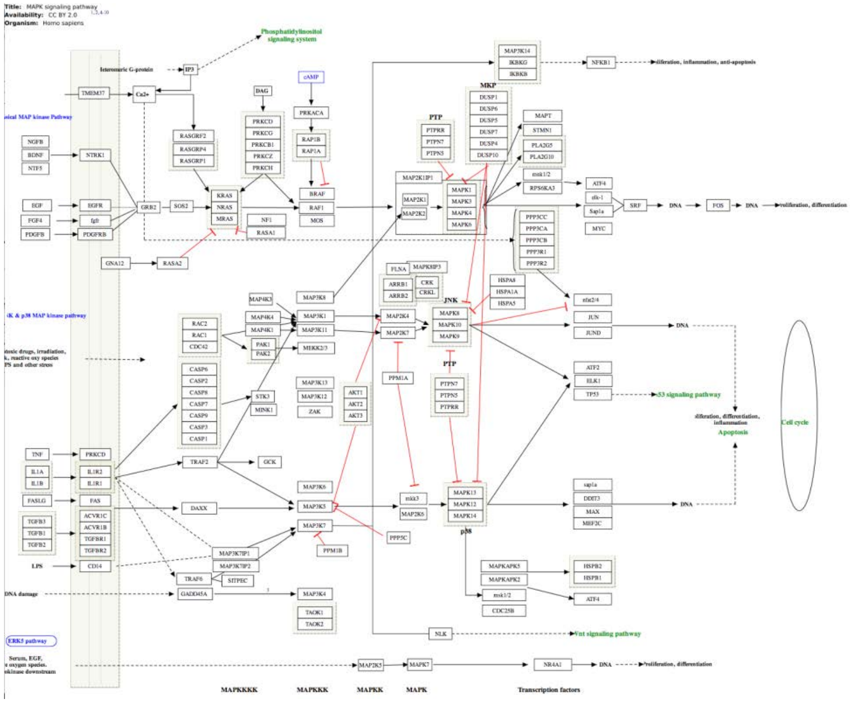
Many other measures of node’s importance in a network…

Whither Pathway Information?

- What source of pathway definitions?: e.g., KEGG vs. wikipathway
- How broad should Protein-Protein Interaction (PPI) networks be?



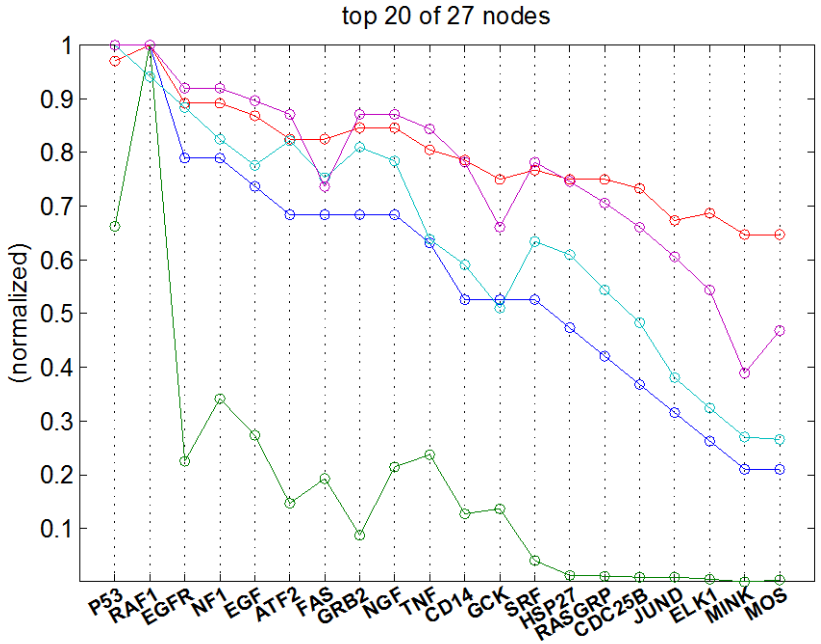
MAPK: KEGG



MAPK: WIKIPATHWAY

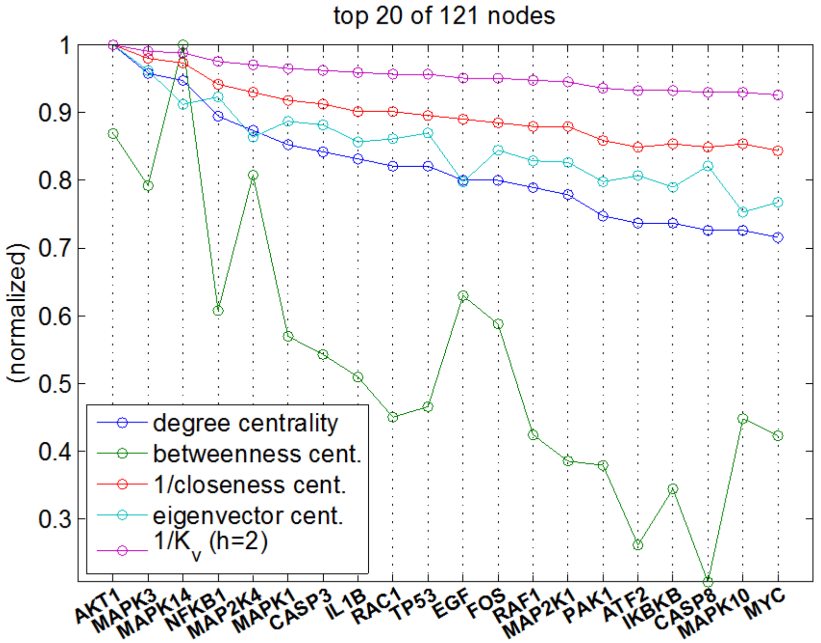
PPI Sub-Network of MAPK Pathway: High-ranking central nodes

KEGG



Network diameter: 3;
characteristic path length: 1.73

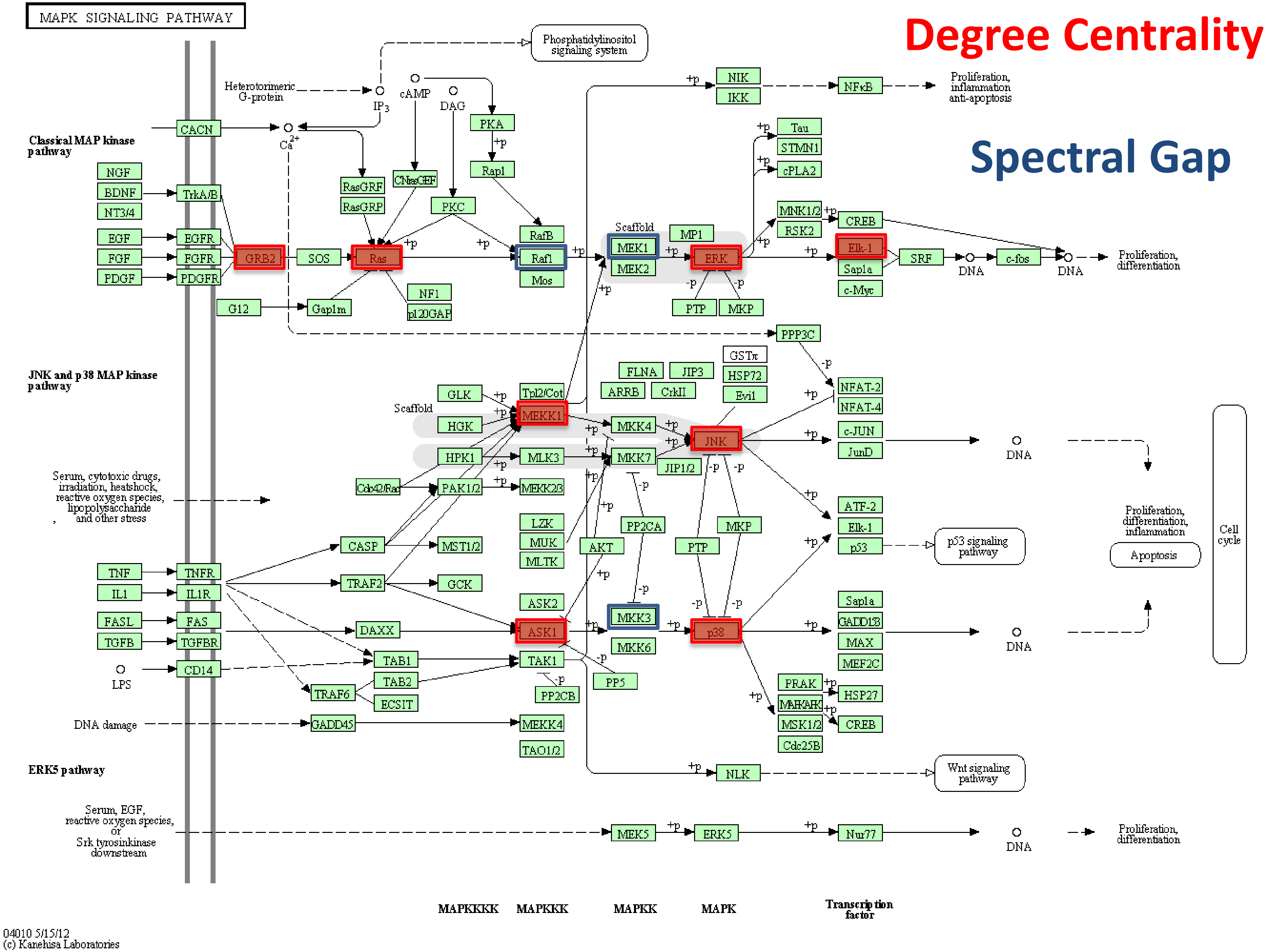
WIKIPATHWAY



Network diameter: 3;
characteristic path length: 1.69

Degree Centrality

Spectral Gap


MAPK SIGNALING PATHWAY
Phosphatidylinositol signaling system
Heterotorimeric G-protein
IP3
cAMP
DAG
CACN
Ca2+
PKA
+p
Rap1
Classical MAP kinase pathway
NGF
BDNF
NT3/4
TrkA/B
RasGRF
RasGRP
CNrasGEF
PKC
EGF
FGF
PDGF
EGFR
FGFR
PDGFR
GRB2
SOS
Ras
RafB
Raf1
Mos
Scaffold
MEK1
MEK2
MP1
ERK
-p
PTP
MKP
NF1
p120GAP
G12
Gap1m
NIK
IKK
NFκB
Proliferation, inflammation anti-apoptosis
Tau
STMN1
cPLA2
MNK1/2
RSK2
CREB
Elk-1
Sap1a
c-Myc
SRF
DNA
c-fos
DNA
Proliferation, differentiation
PPP3C
JNK and p38 MAP kinase pathway
GLK
Tpl2/Cot
Scaffold
MEKK1
HGK
HPK1
MLK3
MKK4
MKK7
JIP1/2
FLNA
JIP3
ARRB
CrkII
GSTπ
HSP72
Evi1
JNK
NFAT-2
NFAT-4
c-JUN
JunD
DNA
Serum, cytotoxic drugs, irradiation, heatshock, reactive oxygen species, lipopolysaccharide and other stress
Cdc42/Rac
PAK1/2
MEKK2/3
LZK
MUK
MLTK
PP2CA
AKT
PTP
MKP
ATF-2
Elk-1
p53
p53 signaling pathway
Proliferation, differentiation, inflammation
Apoptosis
Cell cycle
CASP
MST1/2
TNF
IL1
TNFR
IL1R
TRAF2
GCK
ASK2
FASL
TGFB
FAS
TGFBR
DAXX
ASK1
MKK3
MKK6
p38
Sap1a
GADD153
MAX
MEF2C
DNA
LPS
CD14
TAB1
TAB2
TAK1
PP5
PP2CB
TRAF6
ECSIT
PRAK
MAPKAPK
MSK1/2
Cdc25B
HSP27
CREB
DNA damage
GADD45
MEKK4
TAO1/2
ERK5 pathway
NLK
Wnt signaling pathway
Serum, EGF, reactive oxygen species, or Srk tyrosinkinase downstream
MEK5
ERK5
Nur77
DNA
Proliferation, differentiation
MAPKKKK
MAPKKK
MAPKK
MAPK
Transcription factor
04010 5/15/12
(c) Kanehisa Laboratories