Bioinformatics

Nicholas J. Schork, Ph.D.
The Scripps Research Institute and
The Scripps Translational Science Institute, La Jolla, CA

Doug Armand; http://www.gettyimages.com/detail/BB7295-001/Stone

SCRIPPS GENOMIC MEDICINE
A Collaboration of Scripps Health and The Scripps Research Institute

Scripps Translational Science Institute

The Scripps Research Institute

Whole Genome Sequencing Has Arrived…

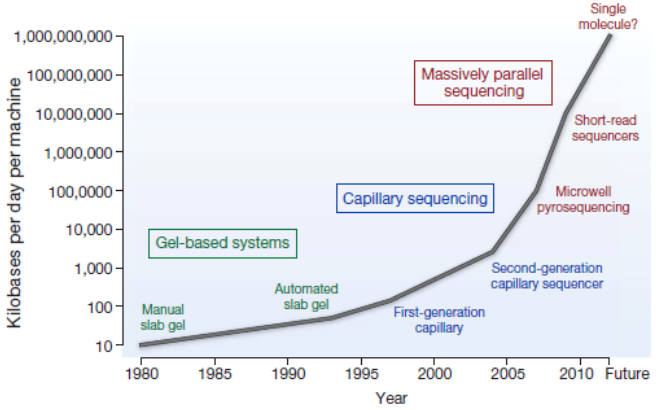
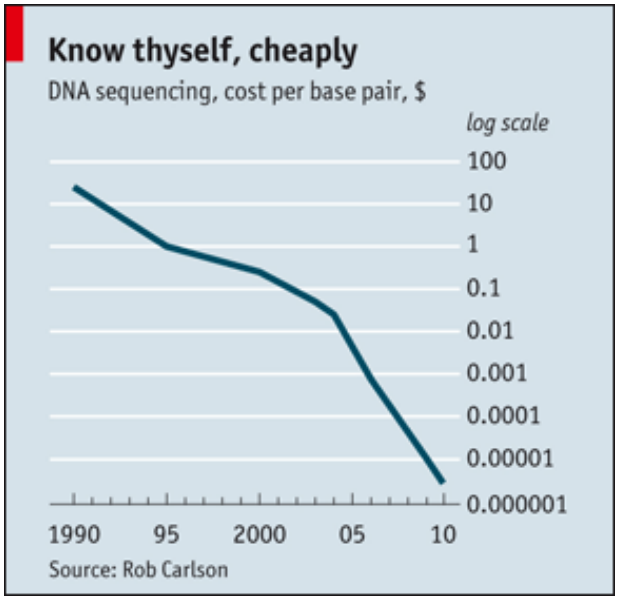
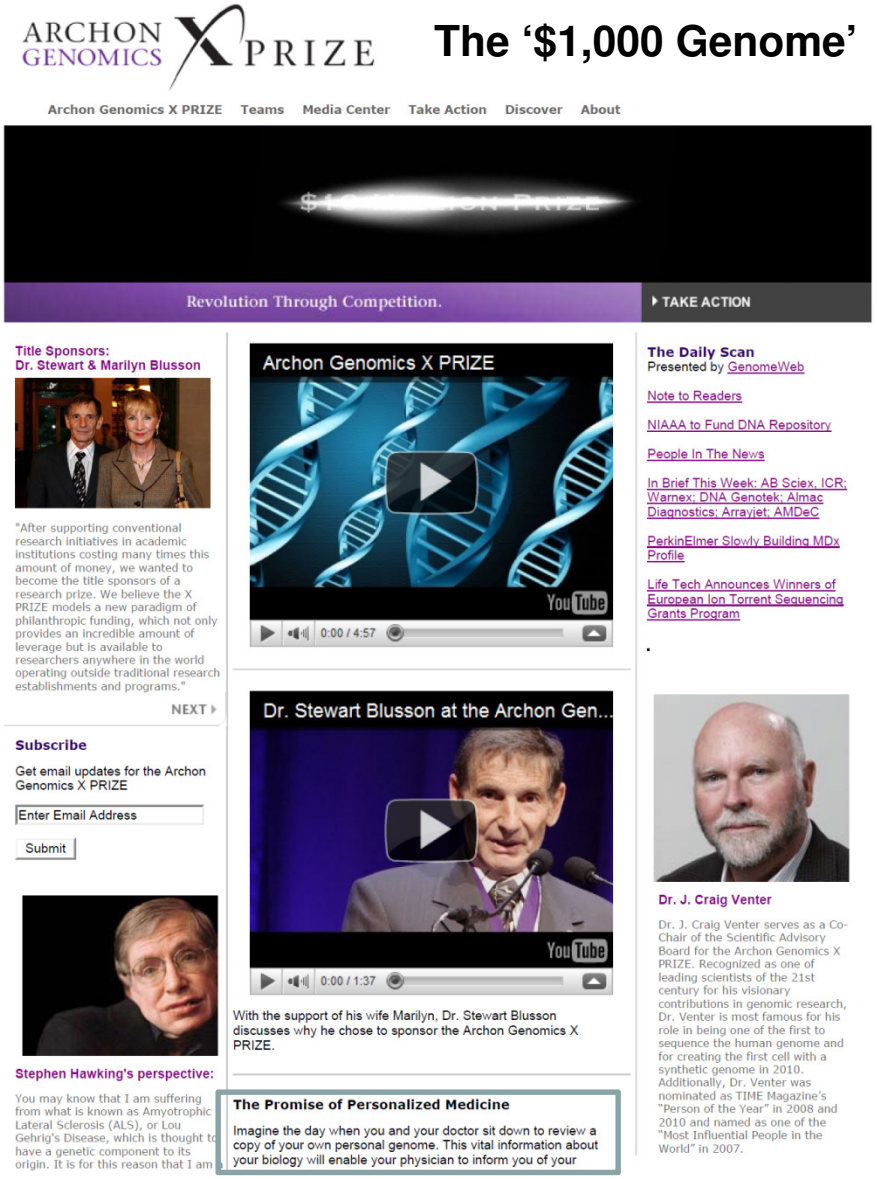


Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.





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."

CopenhagenGenomics 02/22/2011

Functional Annotations: *In Silico* Approaches

- Individual genomes harbor 3-5 million non-reference variants (>50,000 are novel)
- Some of these variants influence phenotypic expression, but which?
- Many variants may have been studied for function or associated with a trait
- Novel and rare variants require computational assessments for function

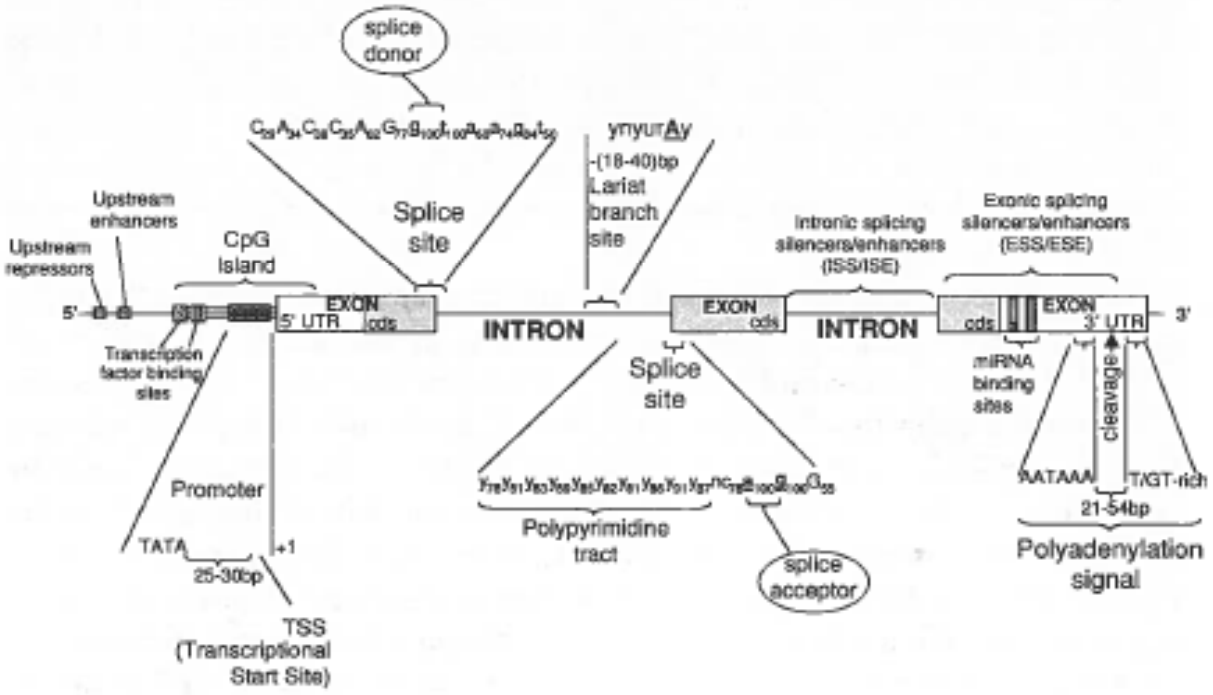


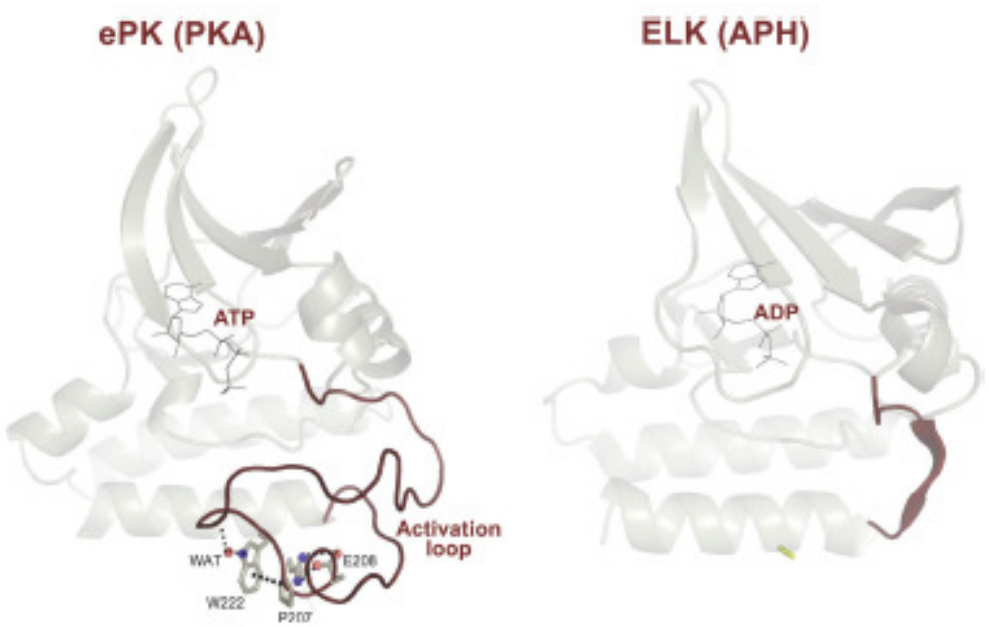
Figure 11.2 The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

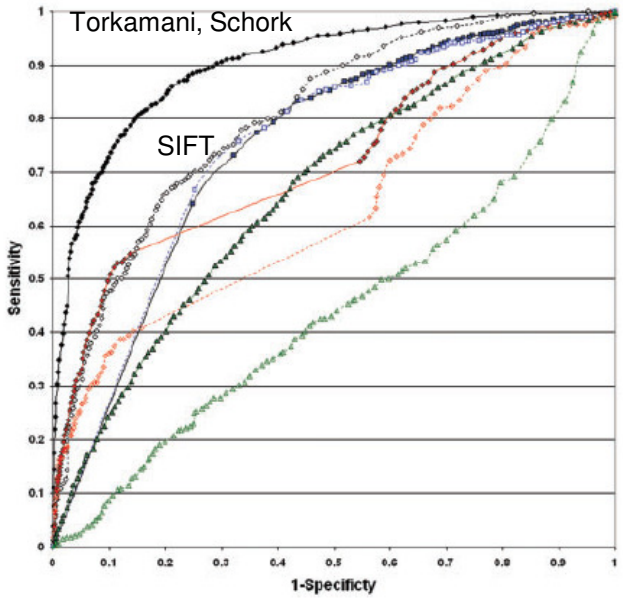
Plumpton and Barnes. "Predictive Functional Analysis of Polymorphisms: An Overview." in Bioinformatics for Geneticists. Wiley, 2007

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

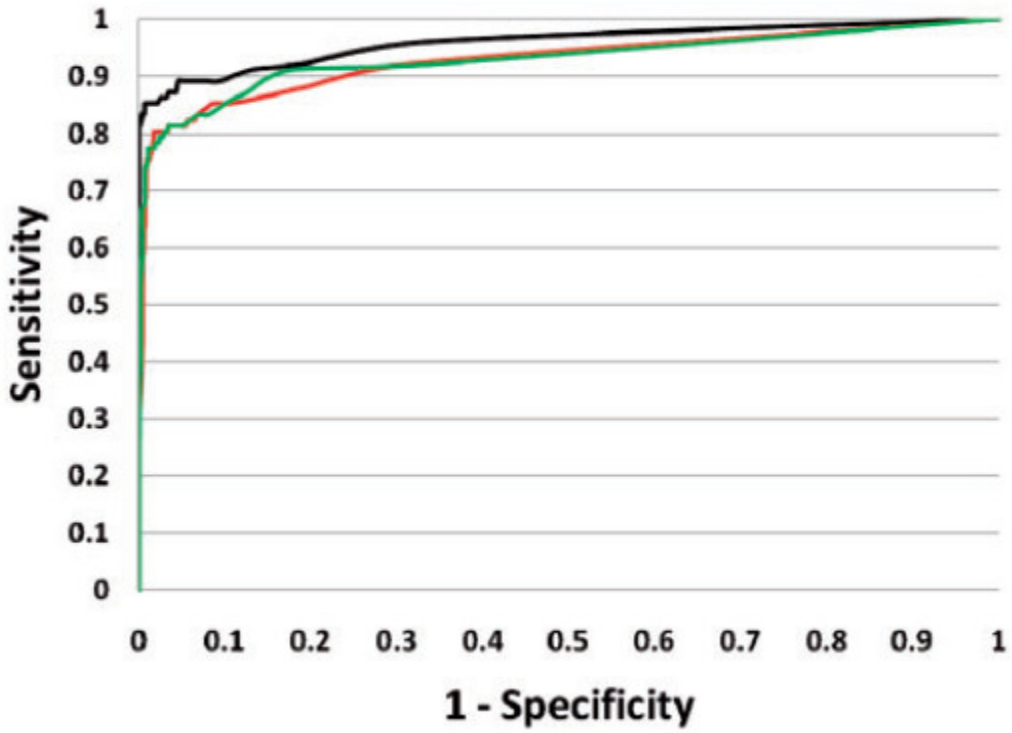
Ali Torkamani[1] and Nicholas J. Schork[2,*]

- **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



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

Filters to Identify Causative Variants in Single Genomes

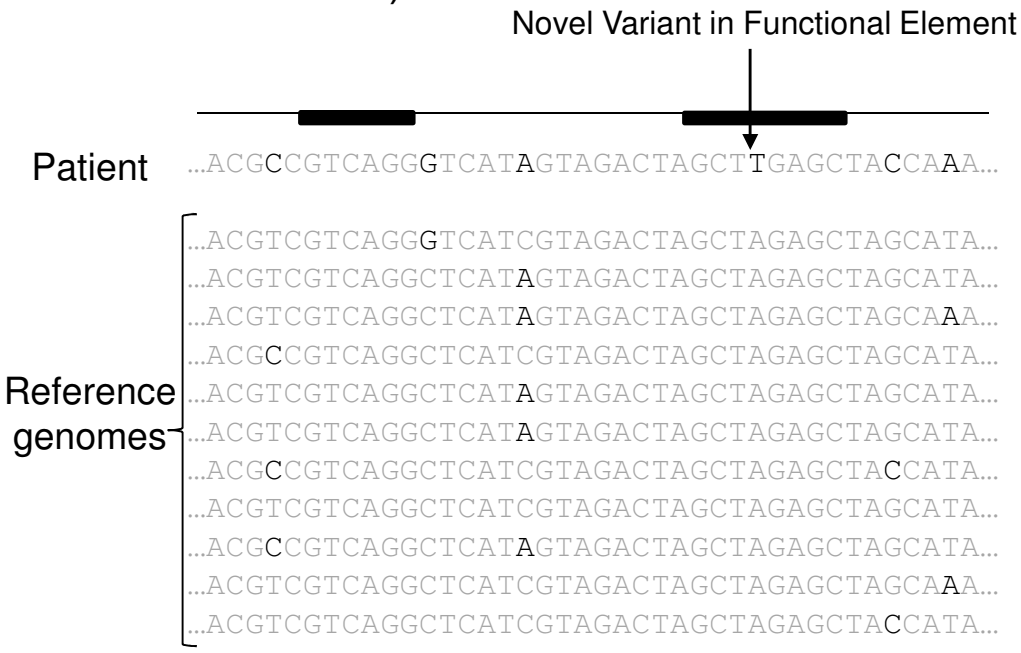
Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency

Alan F. Rope,[1] Kai Wang,[2,19] Rune Evjenth,[3] Jinchuan Xing,[4] Jennifer J. Johnston,[5] Jeffrey J. Swensen,[6,7] W. Evan Johnson,[8] Barry Moore,[4] Chad D. Huff,[4] Lynne M. Bird,[9] John C. Carey,[1] John M. Opitz,[1,4,6,10,11] Cathy A. Stevens,[12] Tao Jiang,[13,14] Christa Schank,[8] Heidi Deborah Fain,[15] Reid Robison,[15] Brian Dalley,[16] Steven Chin,[6] Sarah T. South,[1,7] Theodore J. Pysher,[6] Lynn B. Jorde,[4] Hakon Hakonarson,[2] Johan R. Lillehaug,[3] Leslie G. Biesecker,[5] Mark Yandell,[4] Thomas Arnesen,[3,17] and Gholson J. Lyon[15,18,20,*]

Table 4. Summary of the Filtering Procedure and Candidate Genes with VAAST

SNV-Calling Pipeline	GATK	Samtools	GNUMAP
III-4 (total SNVs)	1546	1499	2168
III-4 (nsSNVs)	146	114	155
VAAST candidate genes (NAA10 ranking)	4 (3)	3 (2)	5 (2)
Present in III-4 and mother II-2 (nsSNVs)	122	107	116
VAAST candidate genes (NAA10 ranking)	3 (2)	2 (1)	2 (2)
Present in III-4, mother II-2, and grandmother I-2 (nsSNVs)	115	95	104
VAAST candidate genes (NAA10 ranking)	2 (1)	2 (1)	1 (1)
Present in III-4, II-2, and I-2, absent in brother III-2 and uncle II-8 (nsSNVs)	8	6	8
VAAST candidate genes (NAA10 ranking)	1 (1)	1 (1)	2 (1)

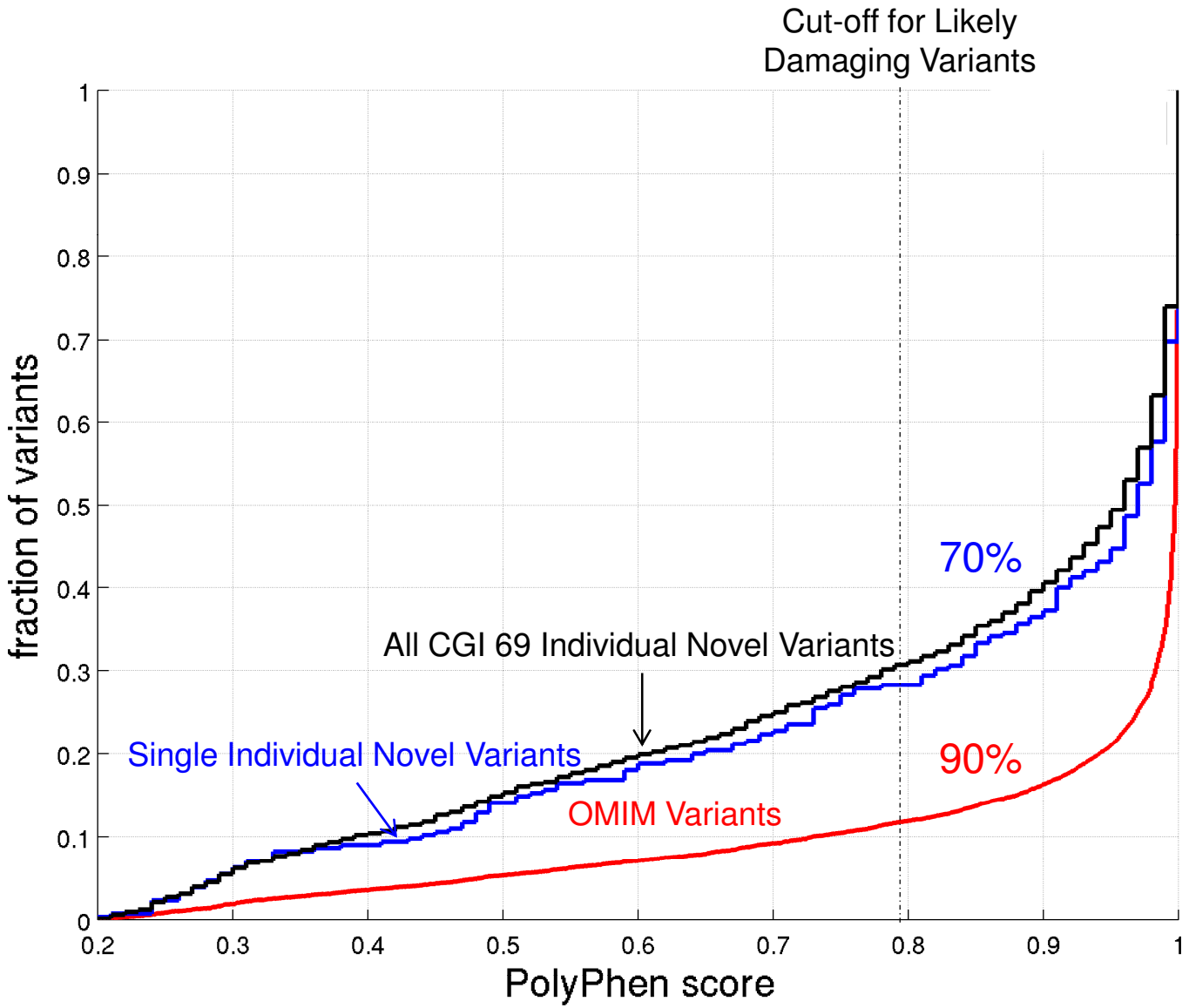
- Genome analysis can be used to identify causative mutations for idiopathic conditions
- Novel variants are likely to be causative for idiopathic conditions
- Annotations and predicted functional effects can help prioritize variants (especially for novel variants)



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

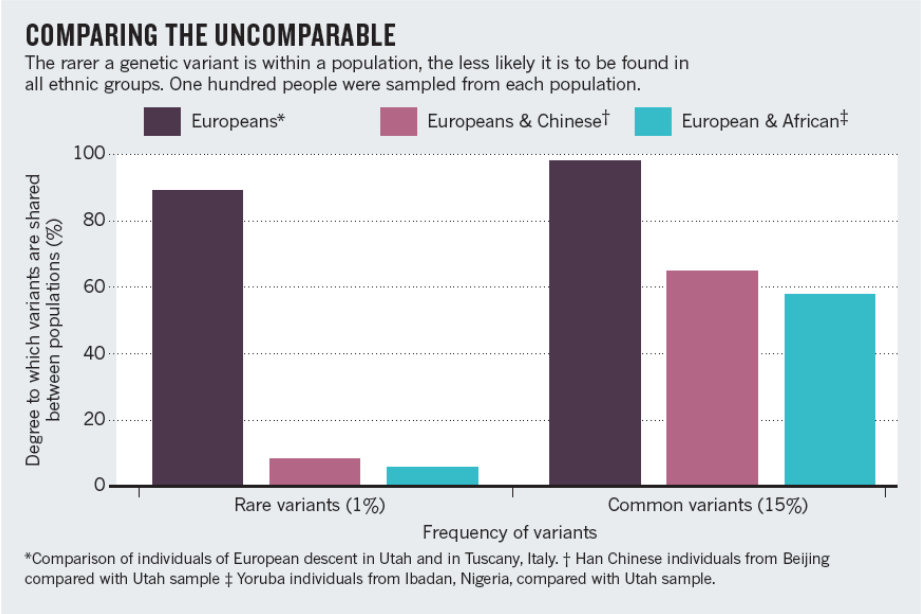
Many Novel Variants on an Individual's Genome Predicted to be Functional



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**.



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

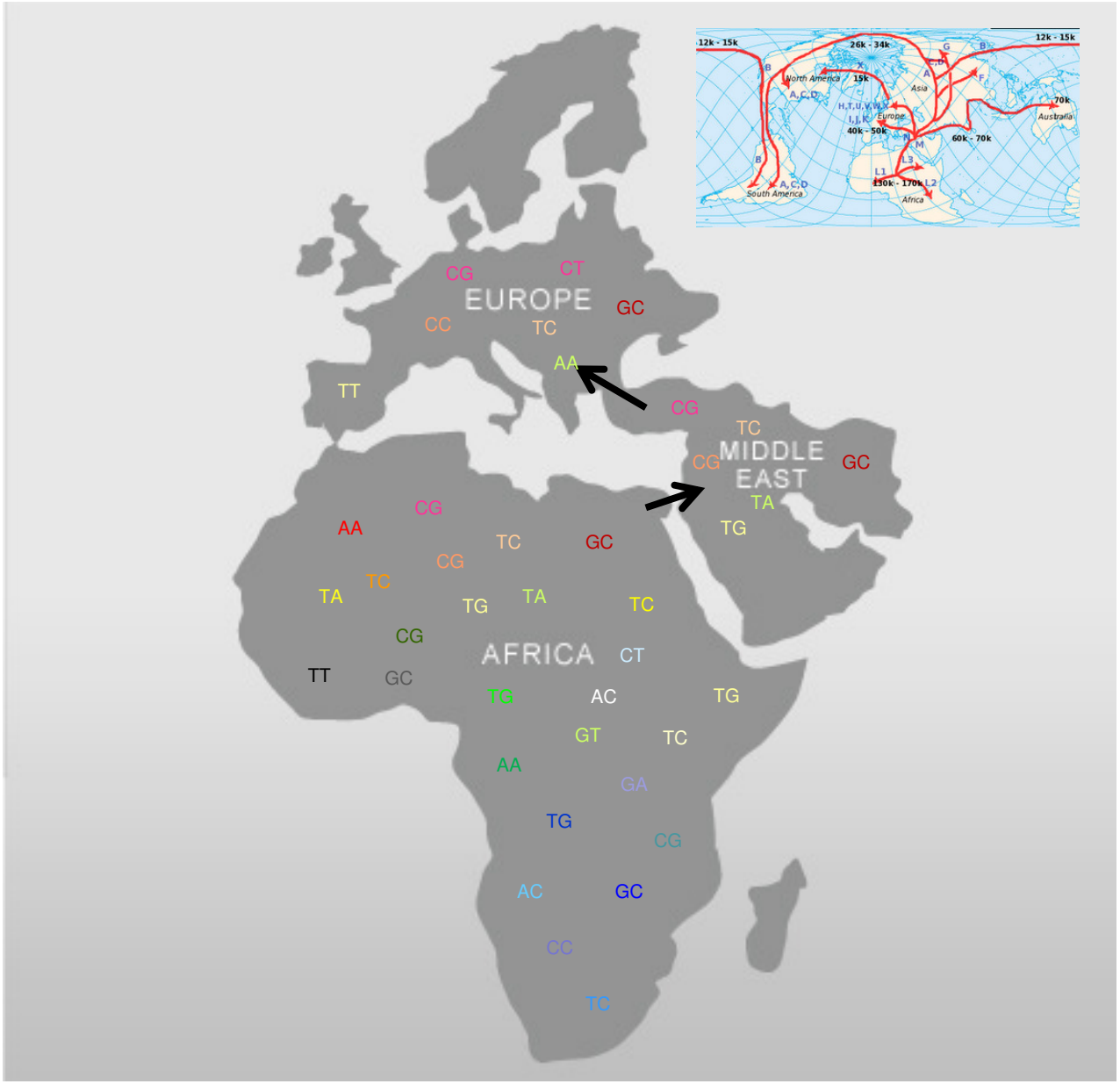
- greater diversity
- selection has washed away some older deleterious alleles
- less homozygosity for older deleterious alleles

Middle East

- only migrant genotypes represented
- early bottleneck created

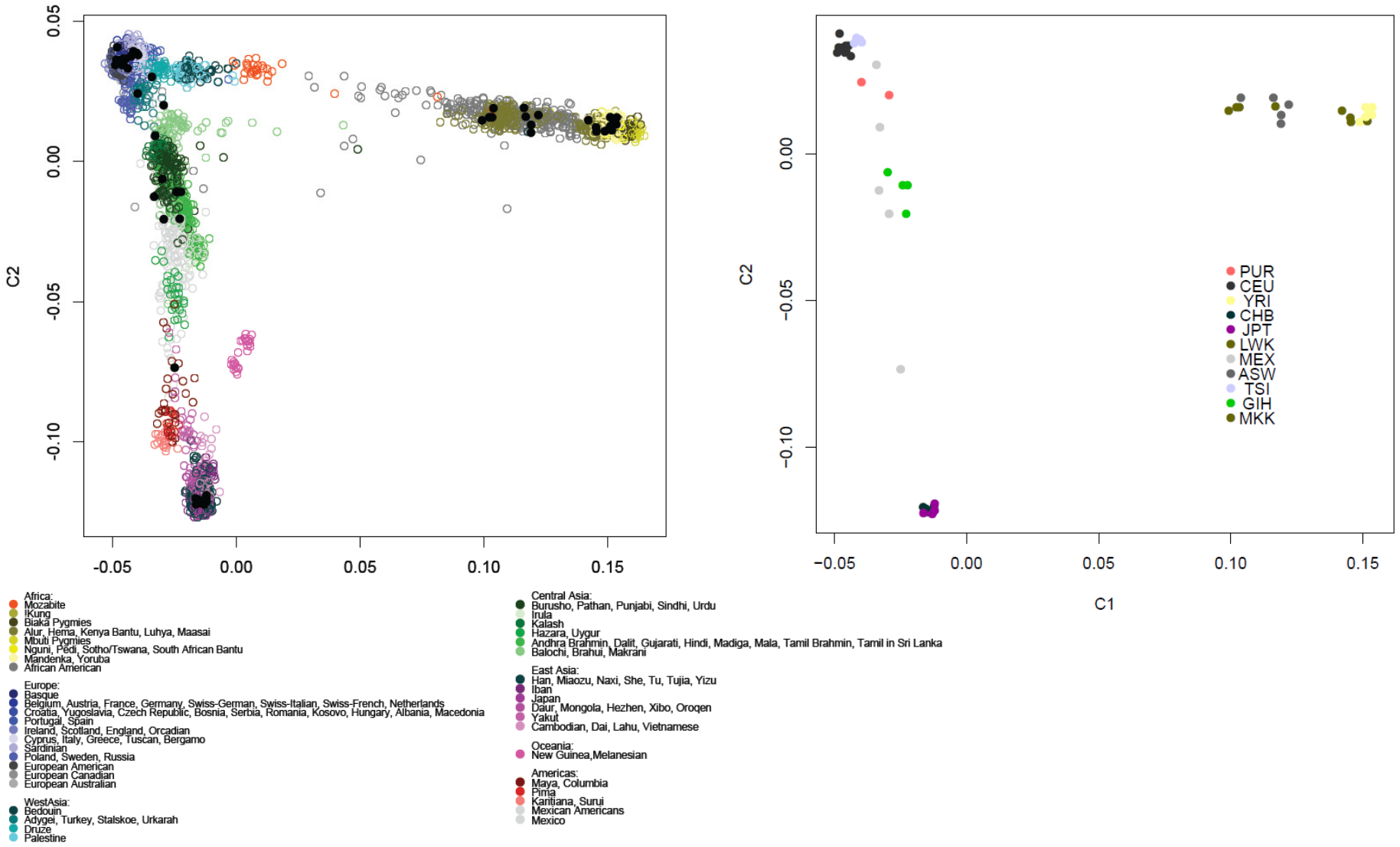
Europe

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

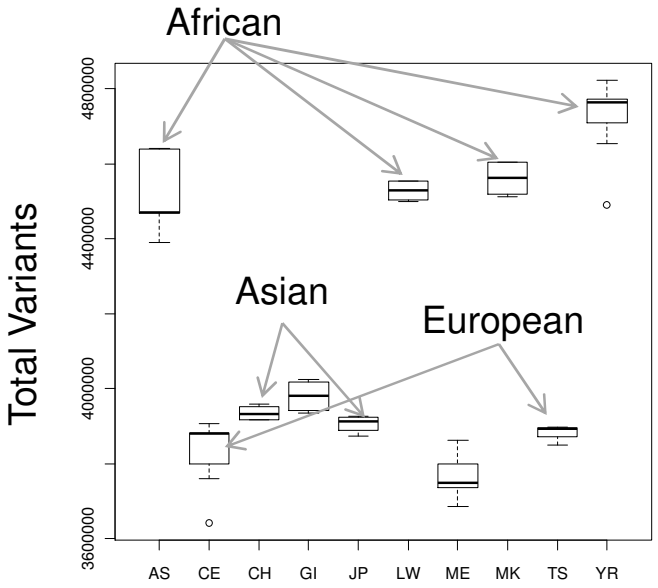


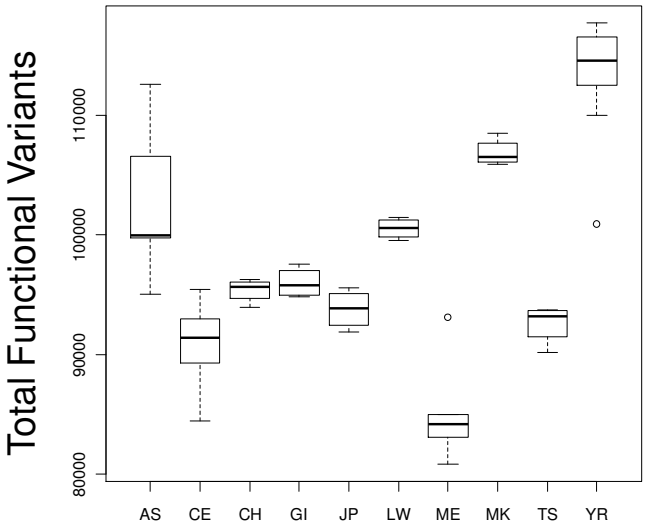
Lohmueller et al. Nature. 2008 451:994-7

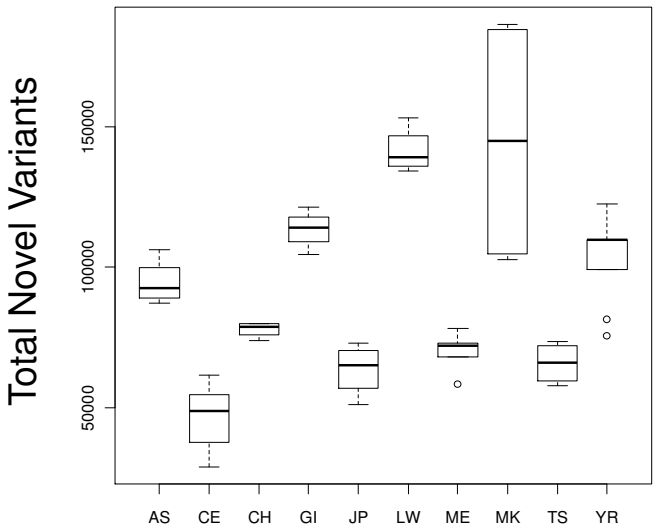
52 Unrelated Individual Whole Genome Variants (CGI)

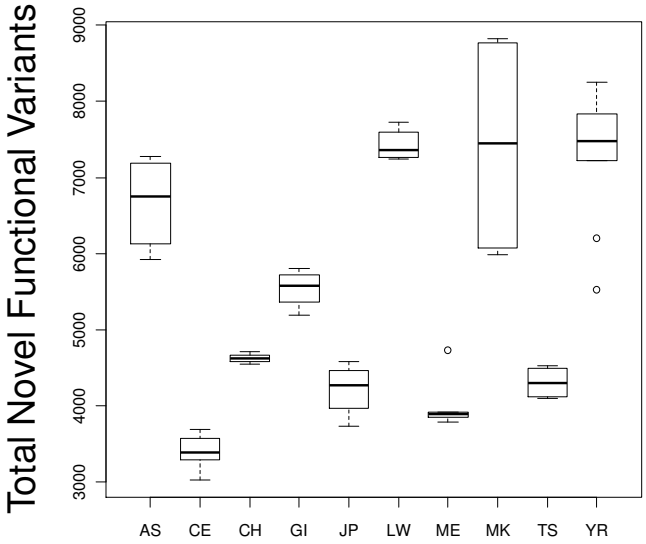


52 Unrelated Individual Whole Genome Variants (CGI)

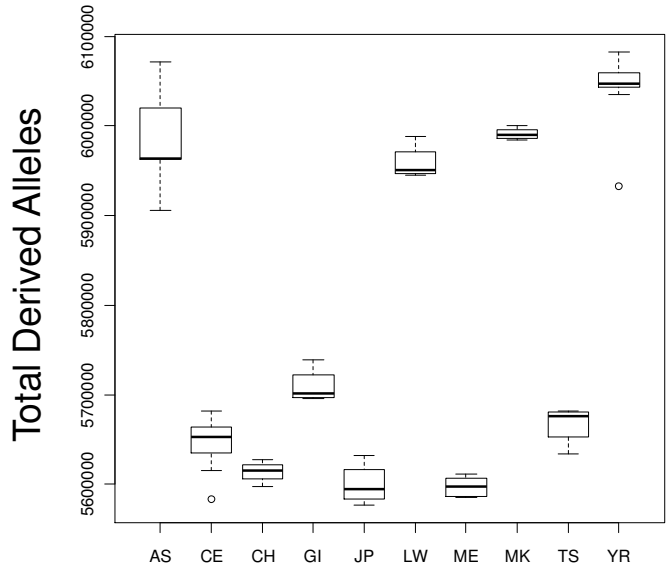




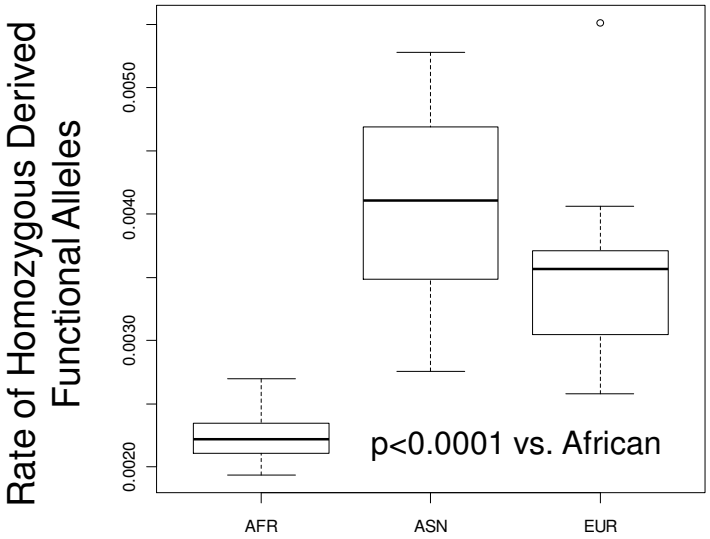


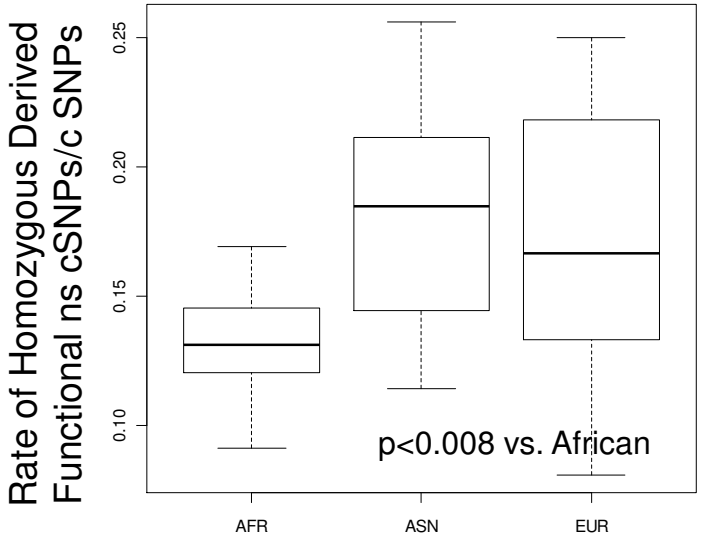


52 CGI Genomes: Evolutionary 'Derived' Alleles



- The standard Human 'Reference' genome is limited for population comparisons, since it is made up of contemporary European DNA
- We determined the 'ancestral' vs. 'derived' alleles from chimp (Jan 2006 Wash U, build 2, version 1) and/or macaque genomes (Jan 2006, JCVI version 1, edition 4)
- Many loci on the human genome are not completely fixed for the derived allele (i.e., some possess ancestral allele)
- We find evidence substantiating Lohmueller et al. (Nature. 2008 451:994-7) but *with individual whole genome* data





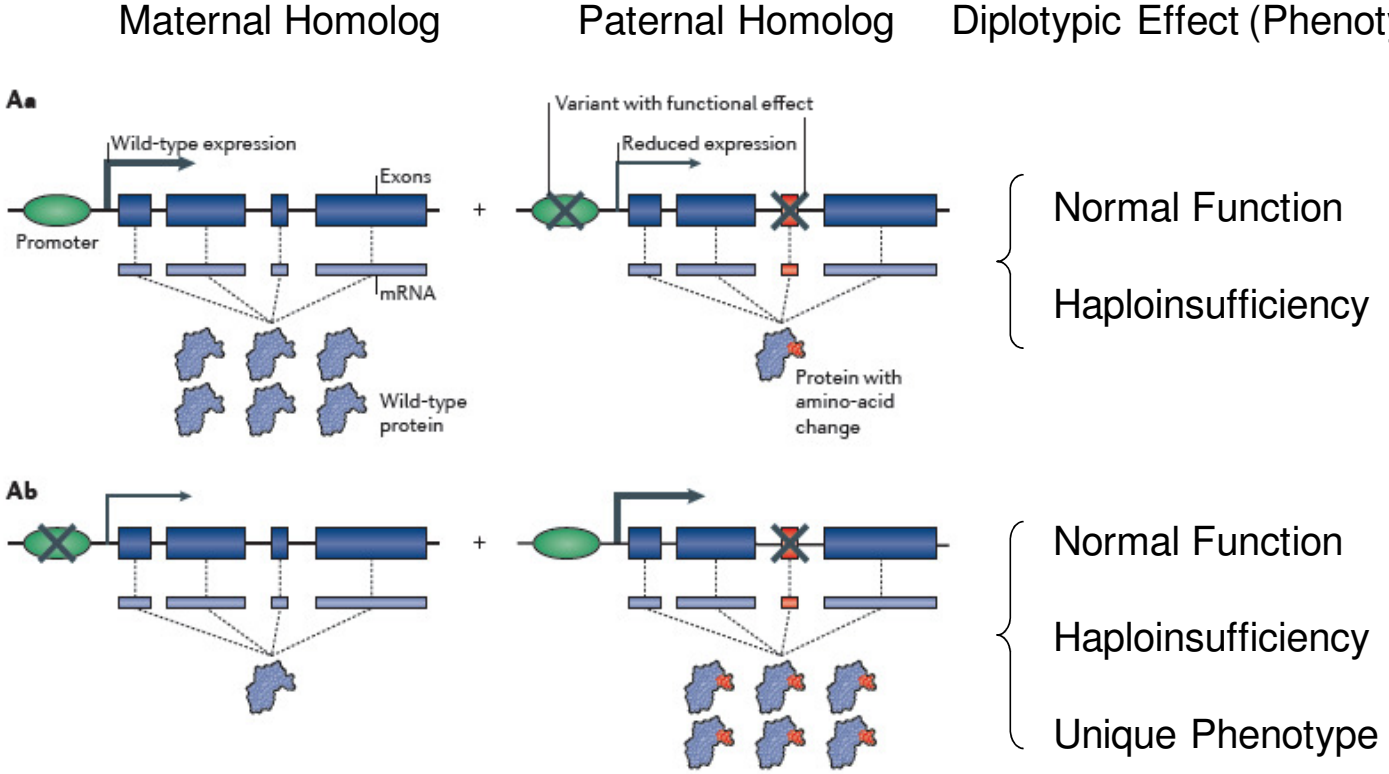
Nature Reviews Genetics | AOP, published online 8 February 2011;

OPINION

The importance of phase information for human genomics

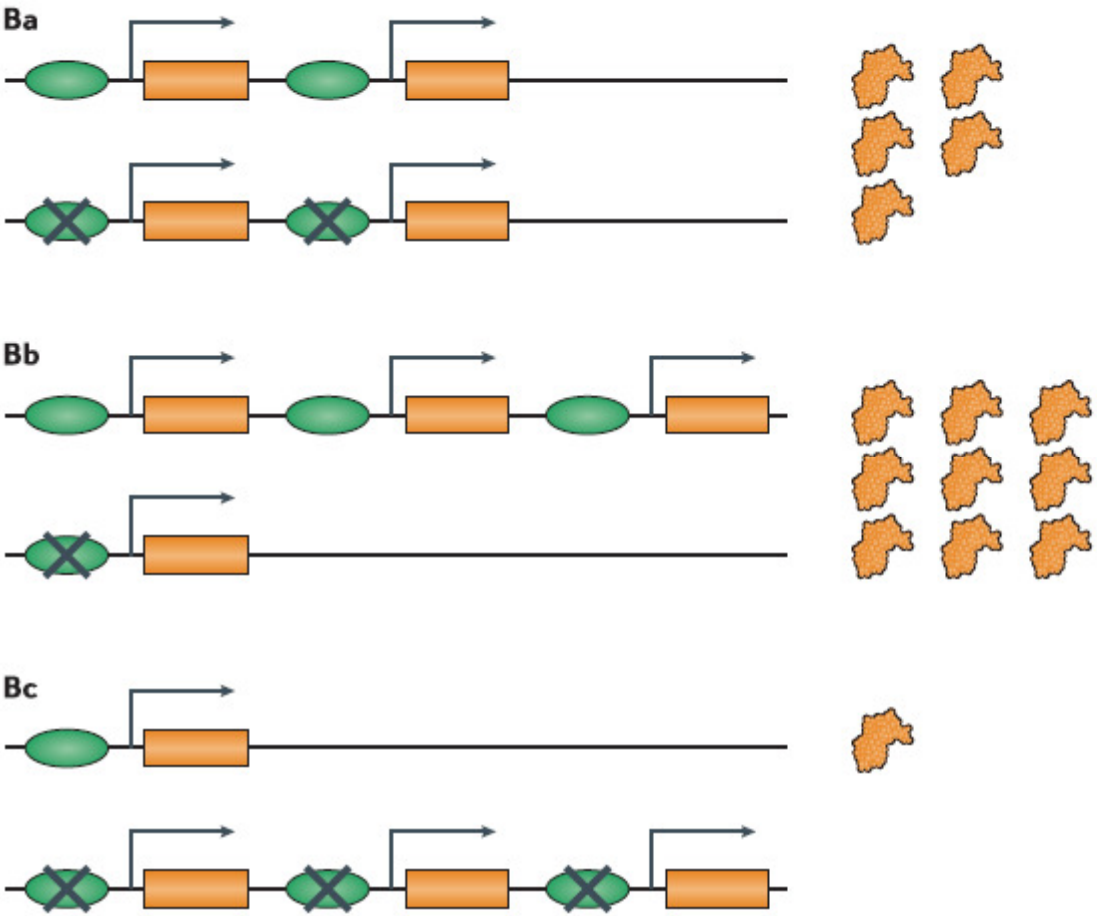
Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol and Nicholas J. Schork

- 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



Tewhey et al. (2011)

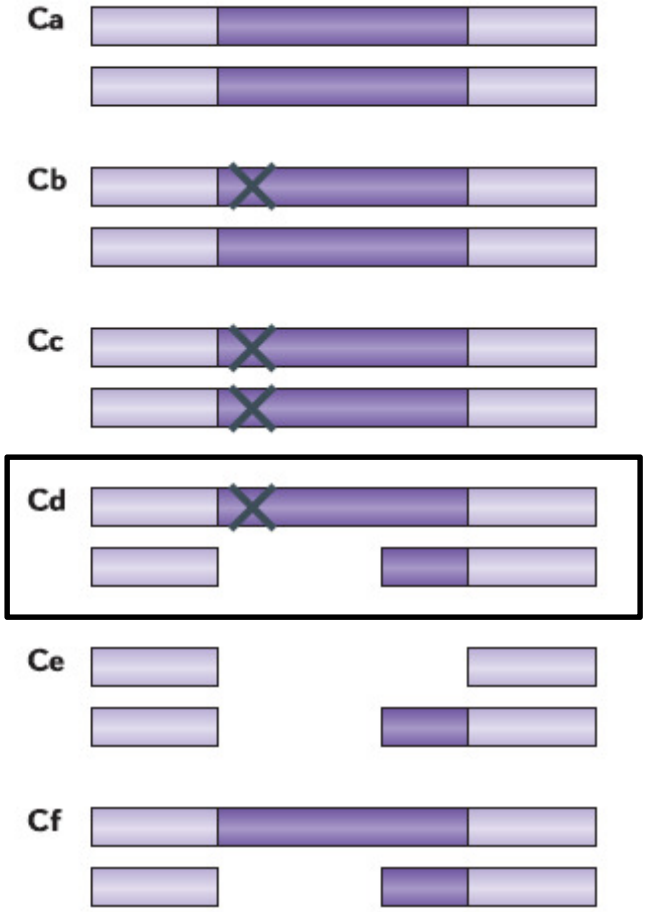
4 Gene Copies but 3 Different Scenarios



Copy Number Variations

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



'Unmasking' via Deletions

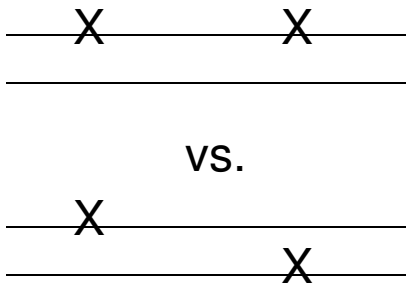
Tewhey et al. (2011)

Compound Heterozygosity

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,[1]* Gustavo Glusman,[1]* Arian F. A. Smit,[1]* Chad D. Huff,[1,2]* Robert Hubley,[1] Paul T. Shannon,[1] Lee Rowen,[1] Krishna P. Pant,[3] Nathan Goodman,[1] Michael Bamshad,[4] Jay Shendure,[5] Radoje Drmanac,[3] Lynn B. Jorde,[2] Leroy Hood,[1]† David J. Galas[1]†

30 APRIL 2010 VOL 328 **SCIENCE**



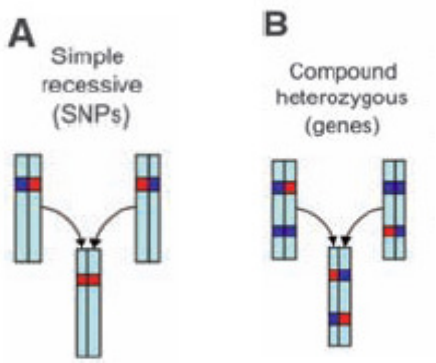


Table 1 | **Example clinical conditions and disorders influenced by compound heterozygosity in single genes**

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.

Tewhey et al. (2011)

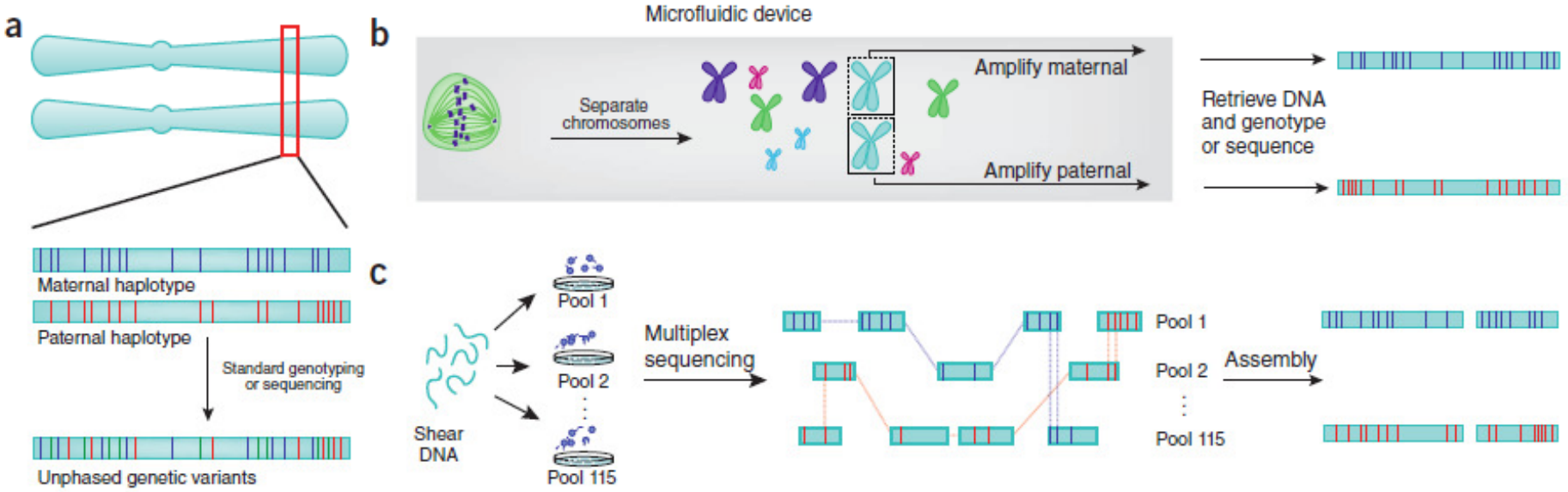
Biological Settings for ‘Diplomics’ Phenomena

Settings Involving Sequence Variations

- Allele Specific Expression (ASE)
- Allele Specific Methylation
- Cell-Specific Monoallelic Expression
- Parent-of-Origin Effects
- Dosage Compensation

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

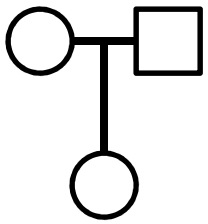
Vikas Bansal, Ryan Tewhey, Eric J. Topol & Nicholas J. Schork

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

Analysis of the Full Genomes of a Trio

STSI-1m STSI-1f



STSI-1

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)*

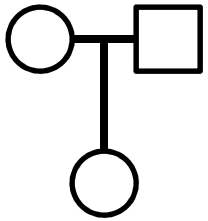
Individual	Seq (Gb)	SNVs	Novel	Ins	Novel	Del	Novel
Child (STSI-1)	121.9	3163286	210730	145411	56028	156147	61544
Mother (STSI-1m)	137.2	3229588	216800	155150	59506	166060	64507
Father (STSI-1f)	138.4	3236815	216996	157779	60310	169006	65139
Combined	-	4469443	419783	268714	125258	295595	135390

- Sequencing and variant calling by Complete Genomics, Inc.
- In house phasing algorithms + functional annotations of all variants
- Initial analyses: determine frequency of potential compound heterozygosity

Torkamani et al. (in review)

Phase and Heterozygosity

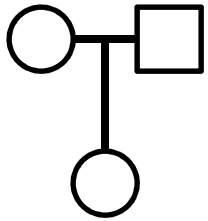
STSI-1m STSI-1f



STSI-1

Mend	Mat	Pat	dbSNP vs. Novel
PASS	T	T	dbsnp:rs3013105
PASS	G	G	dbsnp:rs3013106
PASS	T	T	novel
PASS	-	-	novel
PASS	T	T	dbsnp:rs11586932
PASS	T	T	dbsnp:rs3013107
PASS	G	G	dbsnp:rs55957512
PASS	A	A	novel
PASS	G	G	novel
PASS	A	A	dbsnp:rs12044360
PASS	C	C	novel
PASS	C	C	dbsnp:rs7524833
PASS	T	A	dbsnp:rs7534598
FAIL	?	?	dbsnp:rs7518195
PASS	?	?	dbsnp:rs2999889
PASS	?	?	dbsnp:rs2924853
PASS	?	?	dbsnp:rs2924852
PASS	G	A	dbsnp:rs164951
PASS	G	A	novel
PASS	G	G	dbsnp:rs354587
PASS	T	T	novel
FAIL	?	?	dbsnp:rs354577
PASS	G	A	novel
PASS	A	A	dbsnp:rs2999899

STSI-1m STSI-1f



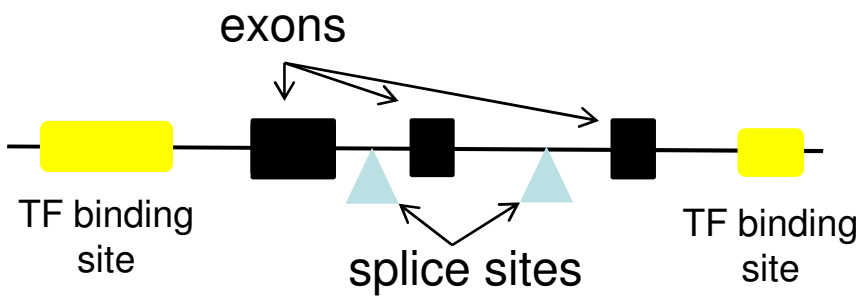
STSI-1

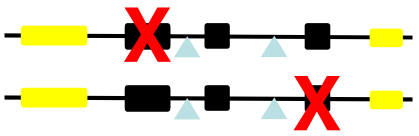
Mend	Mat	Pat	dbSNP vs. Novel
PASS	A	T	dbsnp:rs1203679
PASS	A	T	dbsnp:rs2076324
PASS	A	G	dbsnp:rs1203678
PASS	CCT	-	dbsnp:rs59028030
PASS	G	A	dbsnp:rs1203651
PASS	A	G	dbsnp:rs1203649
PASS	C	T	dbsnp:rs1203648
FAIL	?	?	dbsnp:rs3841673
PASS	C	T	dbsnp:rs979932
PASS	T	C	dbsnp:rs2235522
PASS	G	G	dbsnp:rs1203646
PASS	G	A	dbsnp:rs11580688
PASS	G	T	dbsnp:rs1203645
PASS	G	C	dbsnp:rs1980471
PASS	T	G	dbsnp:rs1203643
PASS	A	G	dbsnp:rs16852976
PASS	G	C	novel
PASS	T	C	dbsnp:rs1203639
PASS	T	A	dbsnp:rs11807320
PASS	T	C	dbsnp:rs11802460
PASS	G	A	dbsnp:rs2281170
PASS	A	G	dbsnp:rs2281169
PASS	TG	TG	dbsnp:rs4025941
PASS	G	G	dbsnp:rs1015370

Heterozygous Sites

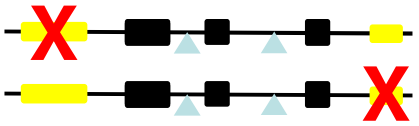
Torkamani et al. (in review)

Genes Harboring Compound Heterozygotes

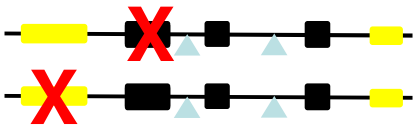




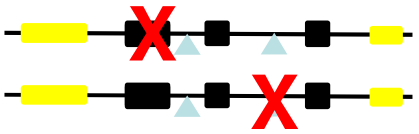
71/184



139/290



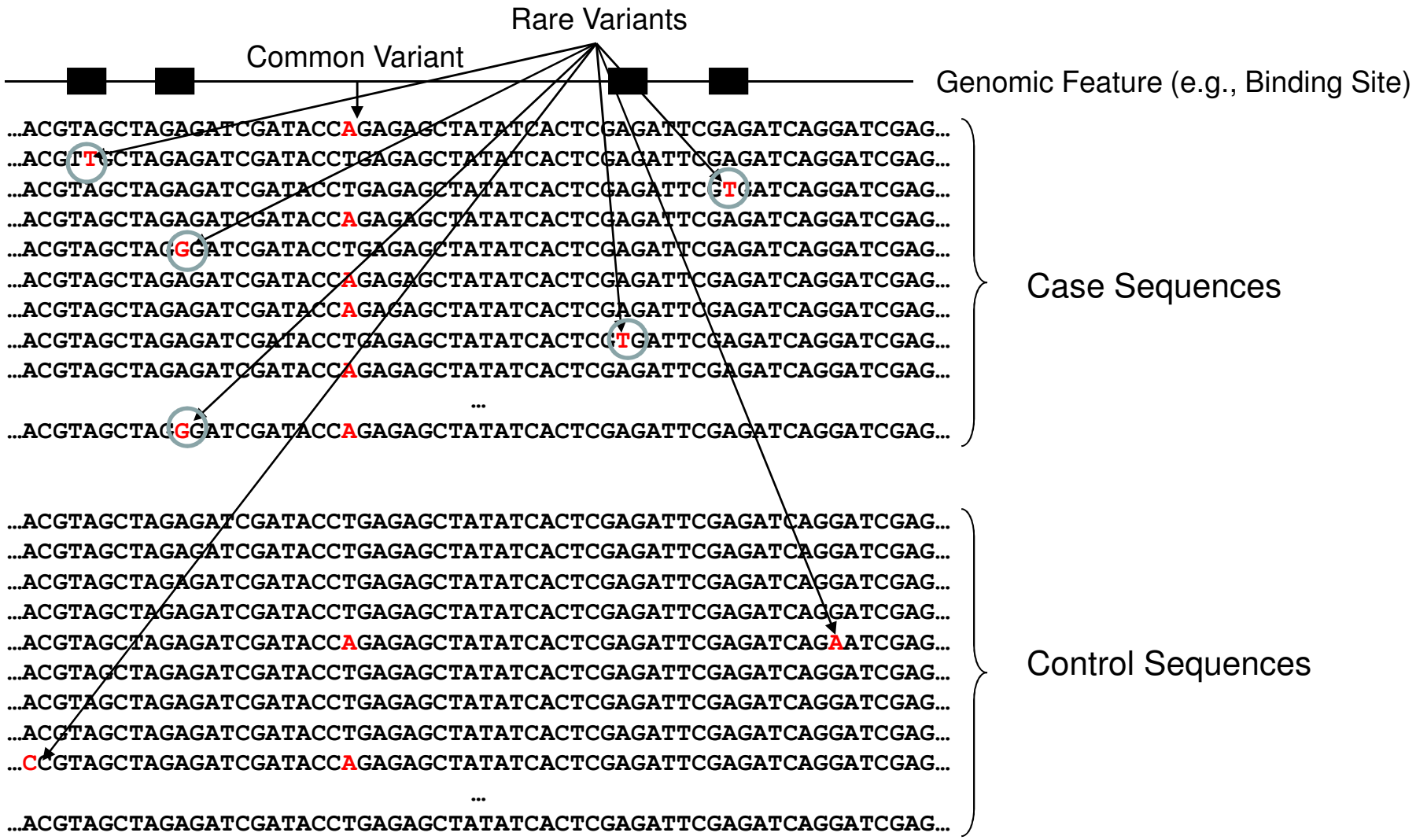
53/389



10/156

Torkamani et al. (in review)

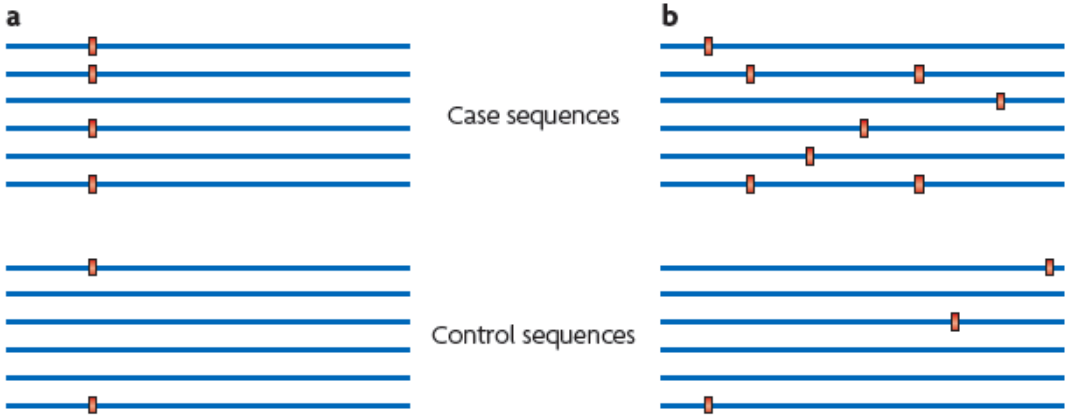
‘Collapsing’ Rare Variations Based on Functional ‘Features’



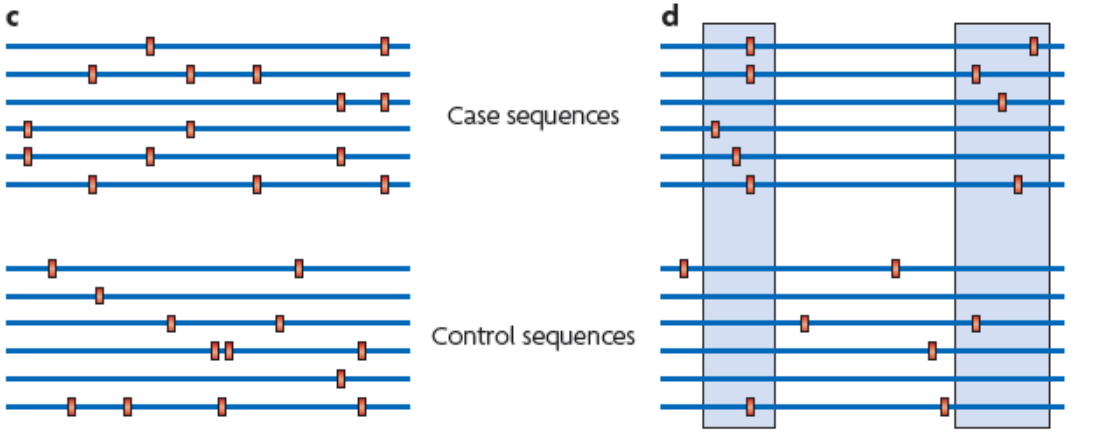
Basic Intuition: Compare the *Collective* Frequency of Variants Between, e.g., Groups

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*

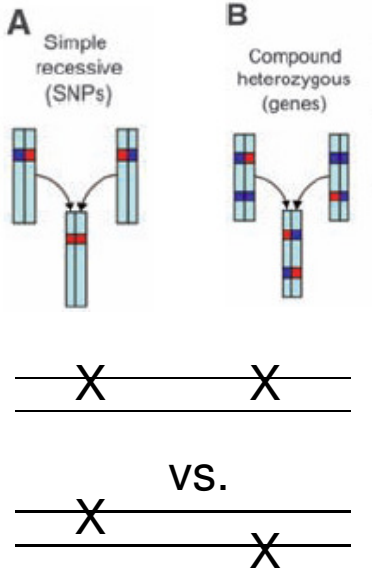


Common Variant **EAH**



Synergistic Effects **Region Specific EAH**

Roach et al. Science (2010)



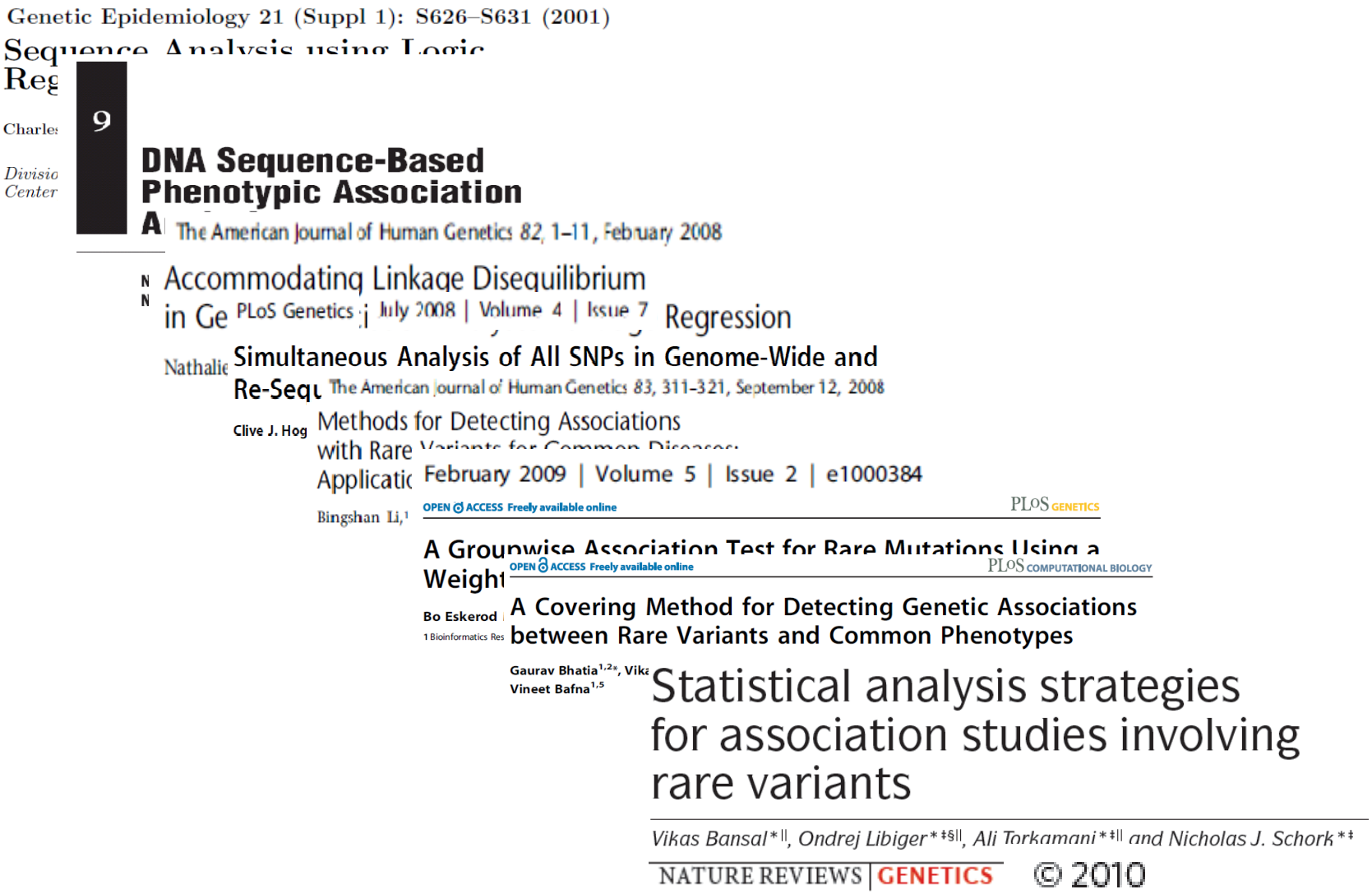
Compound Heterozygosity

Statistical analysis strategies for association studies involving rare variants

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

NATURE REVIEWS | GENETICS

Multilocus Association Studies with DNA Sequencing Data

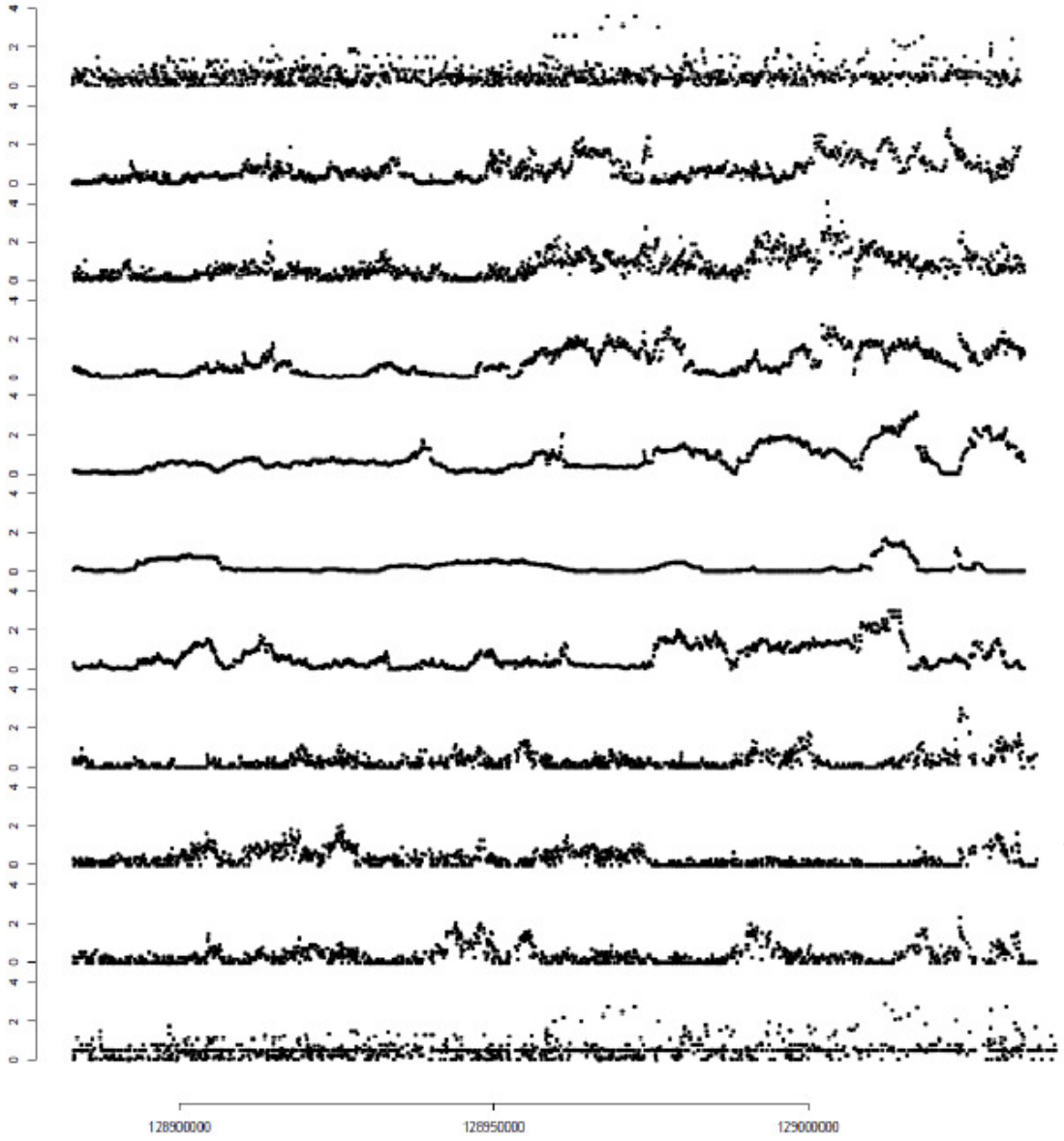


EAH, synergistic, ‘protective vs. deleterious,’ common + rare, additive, etc. models

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)

Different Methods=Different Assumptions=Different Results



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

MGLL gene region: 150 obese vs. non-obese; 5 kb window-based analysis

Bansal et al. PSB 2011

Genetic Risk Assessment for Disease Prevention:
Standard Markers vs. Genetic Markers

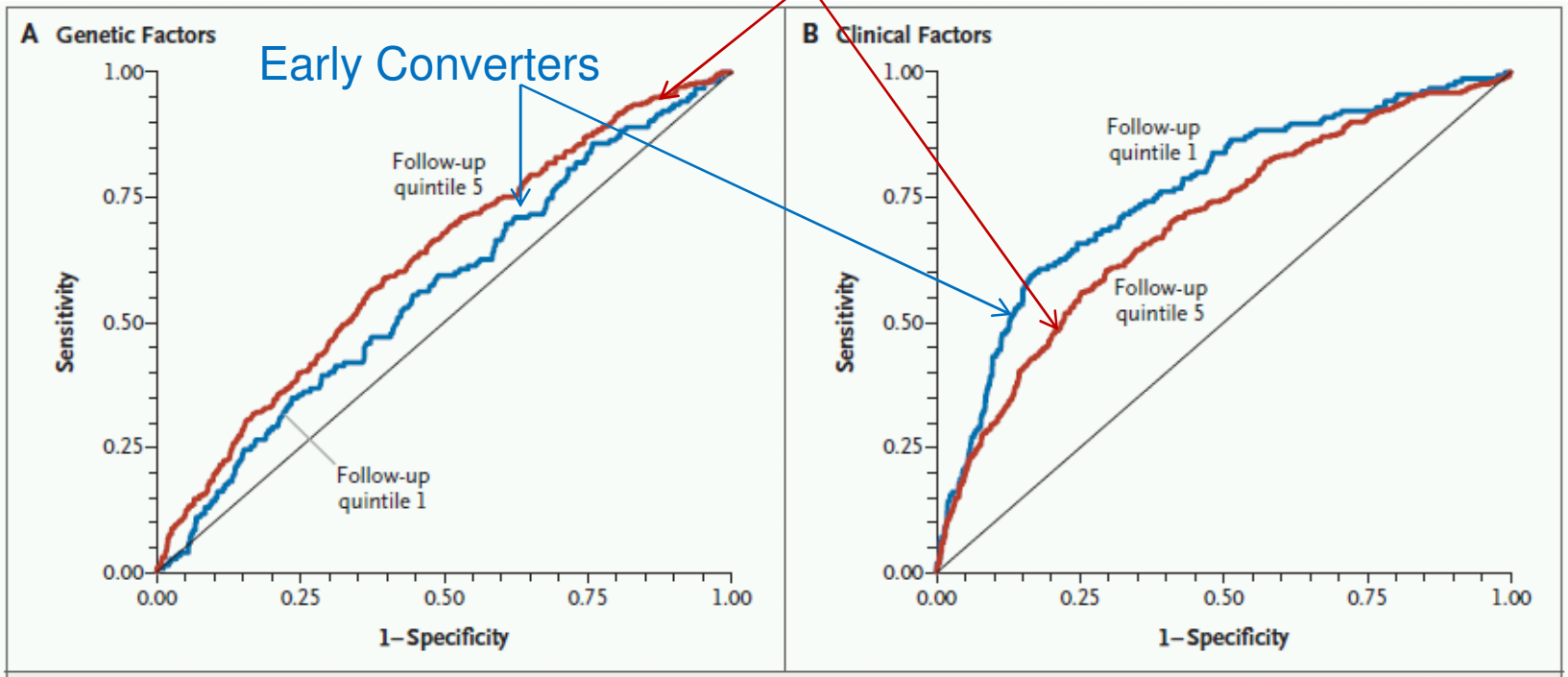


Figure 4. Area under the ROC Curve (C Statistic) for Clinical and Genetic Models Predicting Type 2 Diabetes, According to the Duration of Follow-up.

The effect of genetic risk factors increases with the duration of follow-up, with an area under the ROC curve (AUC) of 0.56 in quintile 1 (blue) and 0.62 in quintile 5 (red) (P=0.01) (Panel A), whereas the effect of clinical risk factors decreased with the duration of follow-up, with an AUC of 0.75 in quintile 1 and 0.67 in quintile 5 (P=0.01) (Panel B). The black line indicates reference values.

New England Journal of Medicine (2008) 359: 2220-2232

N Engl J Med. 2011 Feb 10;364(6):524-34

Effect of Direct-to-Consumer Genomewide Profiling to Assess Disease Risk

Cinnamon S. Bloss, Ph.D., Nicholas J. Schork, Ph.D., and Eric J. Topol, M.D.

CONCLUSIONS

In a selected sample of subjects who completed follow-up after undergoing consumer genomewide testing, such testing did not result in any measurable short-term changes in psychological health, diet or exercise behavior, or use of screening tests. Potential effects of this type of genetic testing on the population at large are not known.

- Recipients of genetic risk data did not seem to engage in healthier behaviors
- Risks may have not been large enough to be motivating?
- Sampling biases?
- What about the provision of non-genetic risk information?

- Social networks = motivation; ‘Quantified Self’ movement; ‘DIY Genomics’; etc.