Polygenic Risk Scores

David Evans

MRC Centre for Causal Analyses in Translational Epidemiology
The University Of Queensland AUSTRALIA
DIAMANTINA INSTITUTE
University of BRISTOL

Polygenic Risk Scores

- For explaining trait heritability and assessing genetic overlap between conditions
  - Not very good, better methods exist
- For “Mining the Phenome”
- For individual risk prediction

Calculating Genetic Risk Scores

- Unweighted Method

  For each individual:

  Score = sum(x_i)

  where x_i is the dosage for “risk” allele i

- Weighted Method

  For each individual:

  Score = sum(x_i) * log(OR_i)

  x_i = Number of risk alleles (=0,1,2) at SNP i

  OR_i = Estimated OR at SNP i from discovery set

Using PLINK to Calculate GRS

./plink --bfile mydata --score myprofile.raw

where myprofile.raw looks like:

SNPA	A	1.95
SNPB	C	2.04
SNPC	C	-0.98
SNPD	C	-0.24

Output looks like:

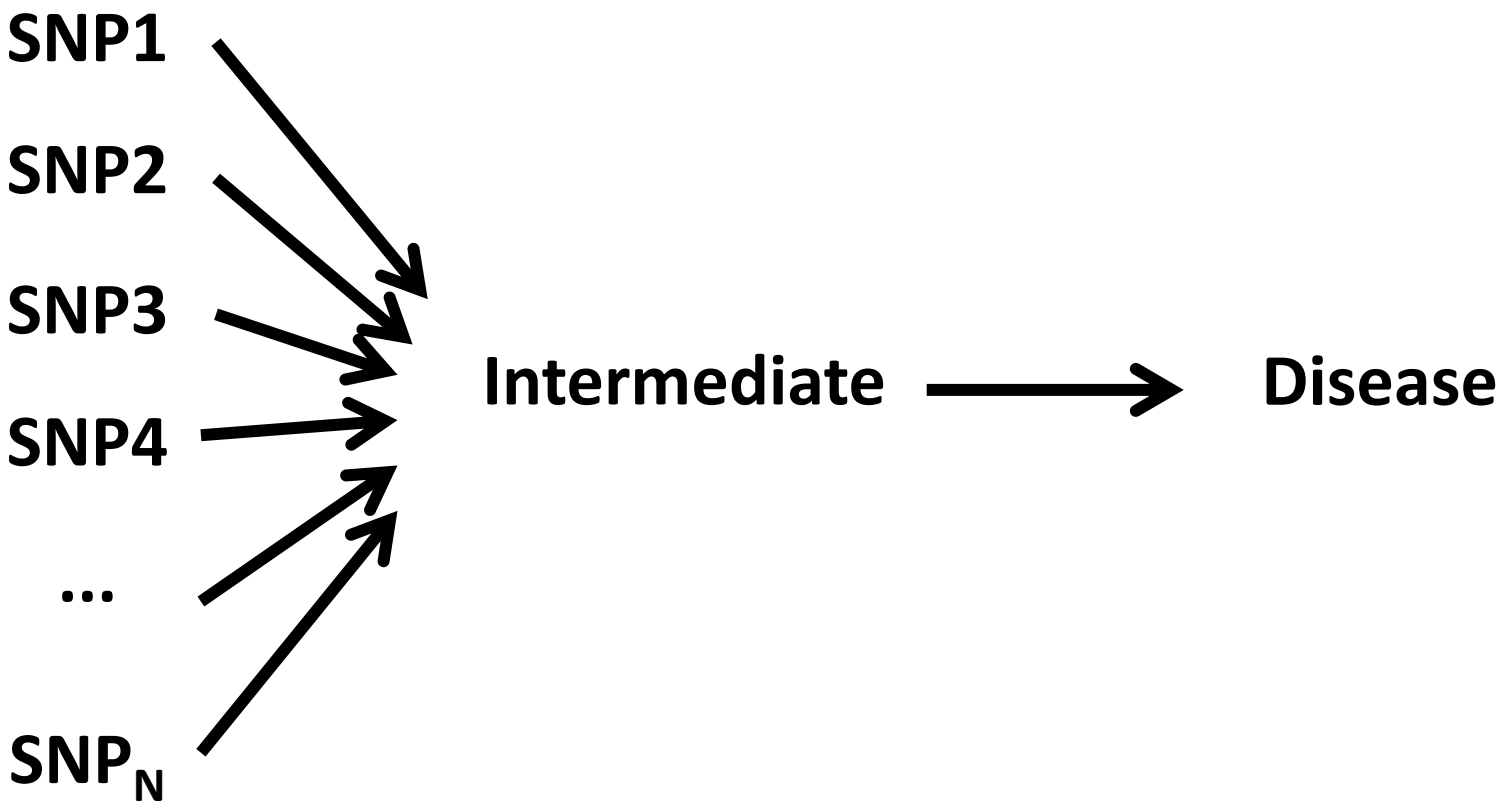
FID Family ID
IID Individual ID
PHENO Phenotype for that
CNT Number of non-missing SNPs used for scoring
CNT2 The number of named alleles
SCORE Total score for that individual

By default, if a genotype in the score is missing for a particular individual, then the expected value is imputed, i.e. based on the sample allele frequency

Mining the Phenome Using Allelic Scores

GWAS

- Typically association between one marker at a time
  - Implicates biological pathways
  - Implicates modifiable exposures
- Low power
- Invert paradigm to look at relationship between allelic scores indexing exposures/intermediates and disease


SNP1
SNP2
SNP3
SNP4
...
SNP$_N$
Intermediate
Disease

Potential Advantages

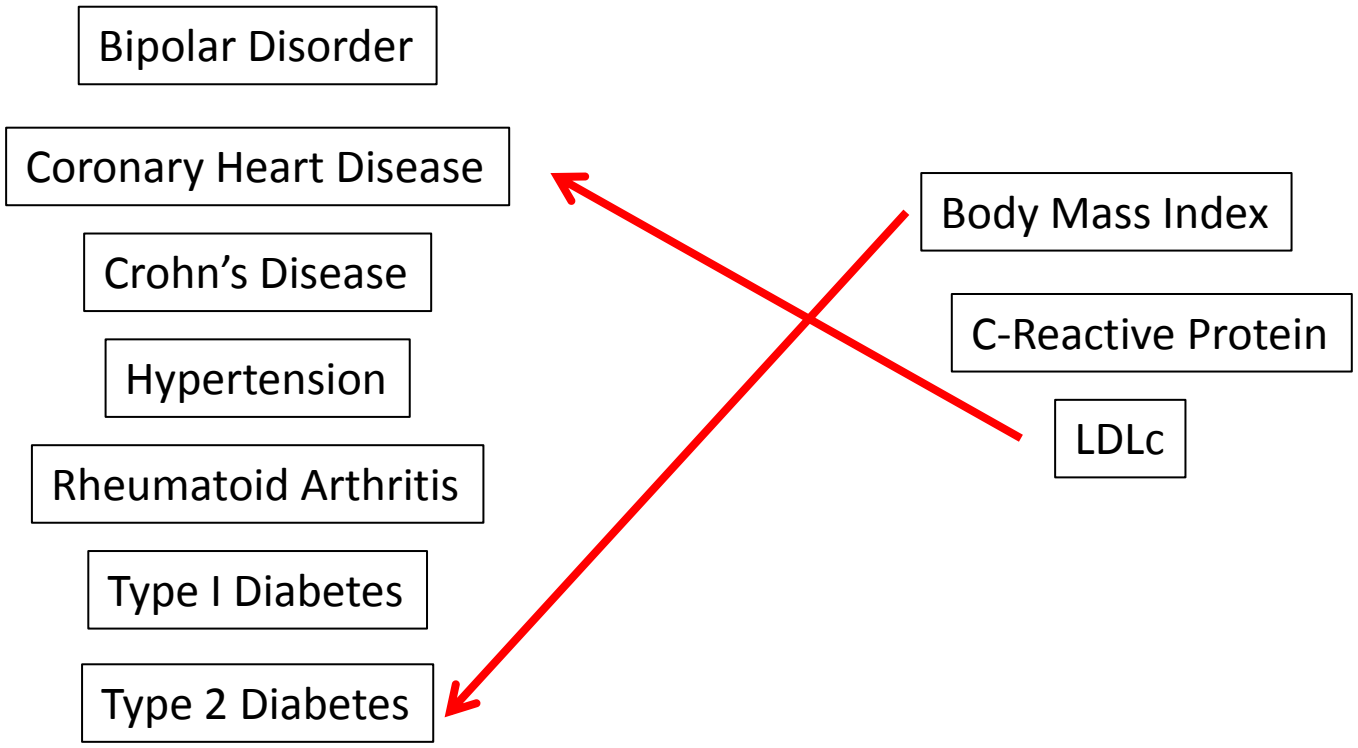
- Allelic scores have greater power than single variants
- Can be used to screen potential causal relationships between 1000s of intermediates/exposures and disease
  - No need to measure intermediate/exposure in disease collection of interest
  - Molecular phenotypes
- Can be used to screen relationships between intermediate/exposure and 100s of diseases
  - *Extremely* large datasets/consortia
- Where we don't know the genetic variants underlying the exposure/intermediate, could we use an anonymous genome-wide score?

Proof of Principal Study - Method

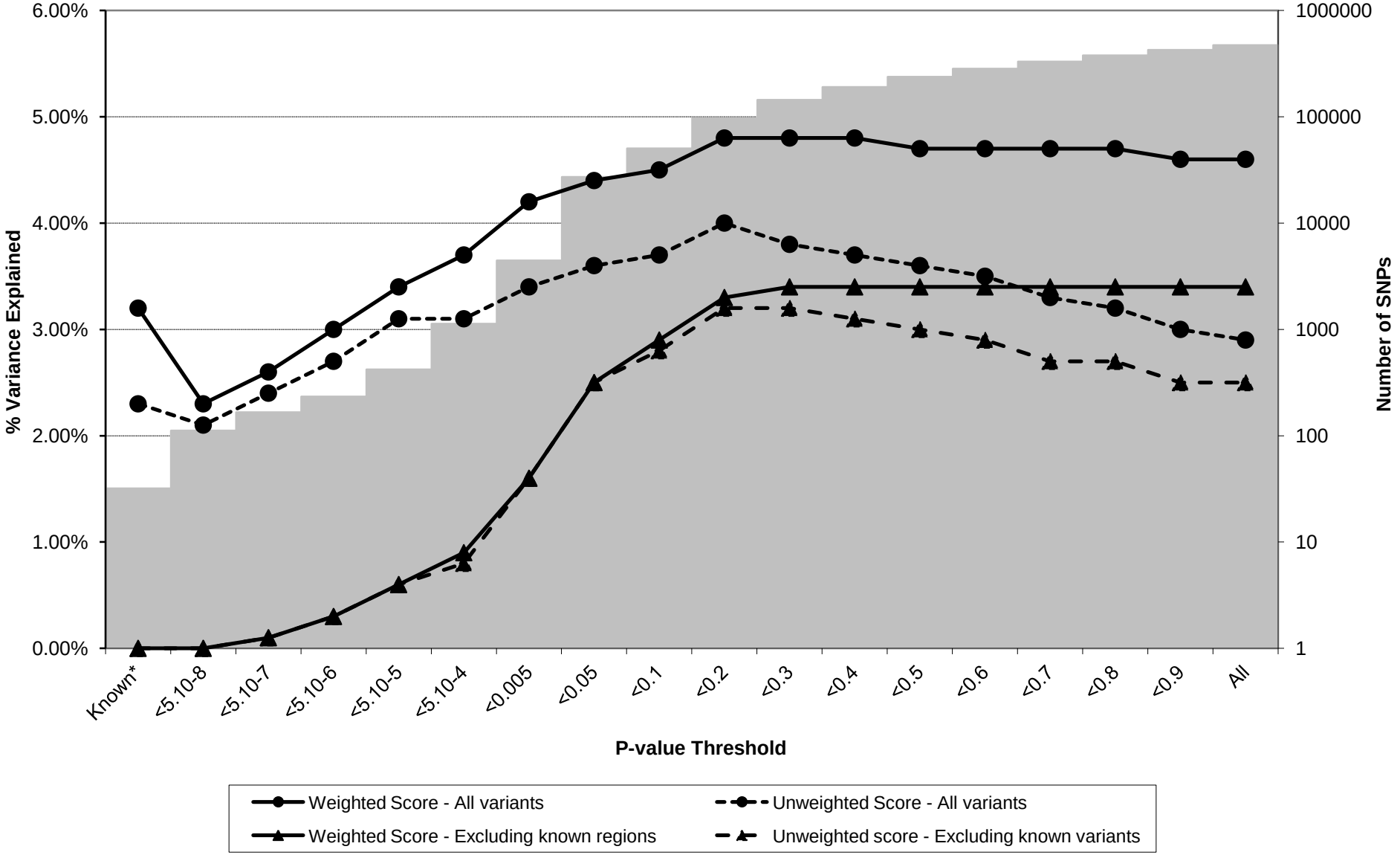
- Take results from GWAS meta-analyses
  - BMI (Speliotes et al. 2010)
  - CRP (Dehghan et al. 2009)
  - LDLc (Teslovich et al. 2010)
- Construct allelic scores in ALSPAC kids and mums
- Correlate allelic scores with case control status in WTCCC1

Wellcome Trust Case-Control Consortium

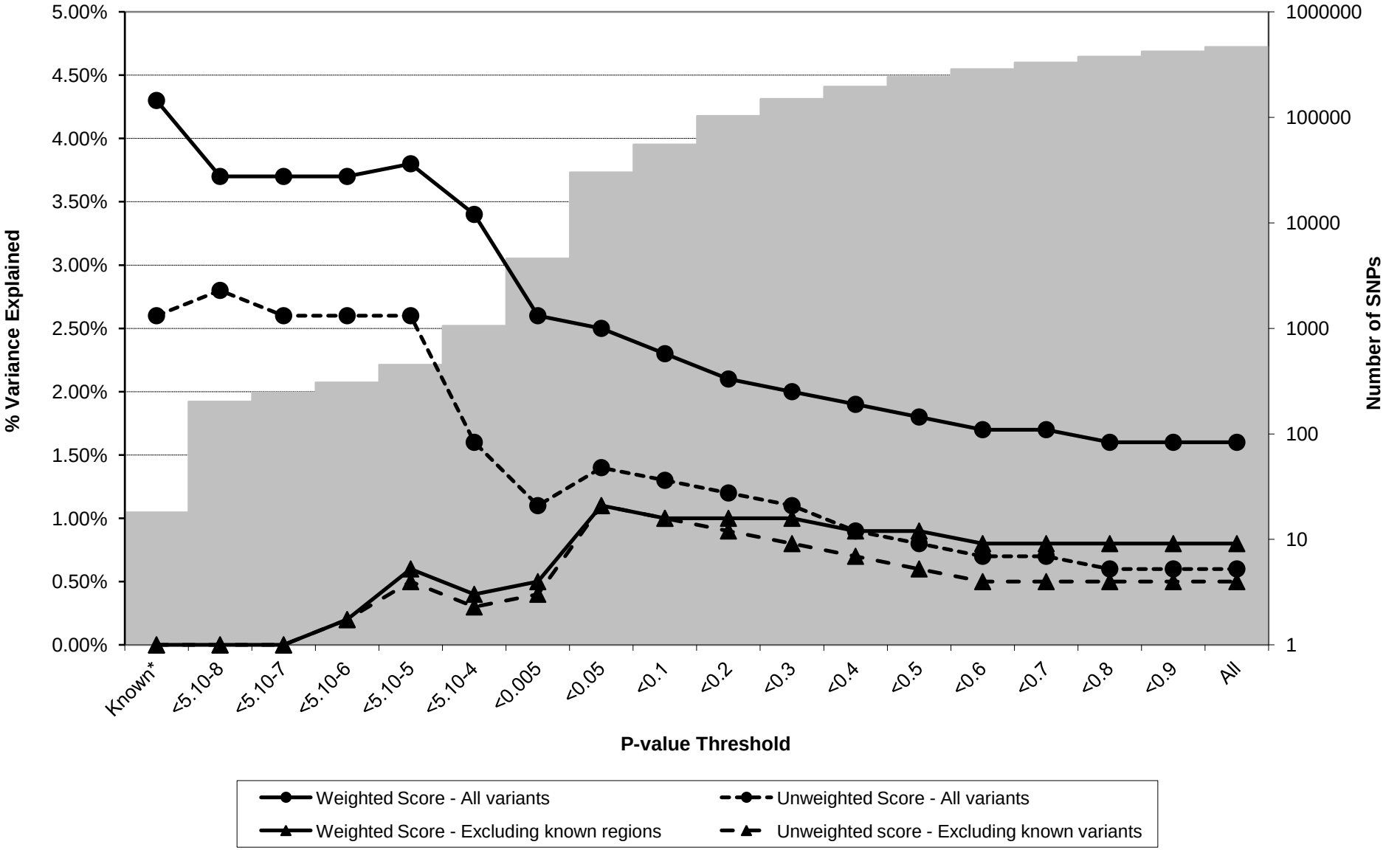
Genome-Wide Association Across Major Human Diseases



Body Mass Index



C-Reactive Protein



Intermediate-Disease Association WTCCC

Table 1. Association between case-control status in the WTCCC and either a weighted genome-wide score consisting of all SNPs across the genome ("GW Score"), a weighted allelic score consisting of highly significant SNPs ($p<5\times10^{-8}$) from known regions only ("Known"), or a weighted genome-wide score consisting of all SNPs across the genome with SNPs from known regions removed from its construction ("Complement").

	BMI						CRP						LDLc					
	GW Score		Known		Complement		GW Score		Known		Complement		GW Score		Known		Complement	
	Dir	P	Dir	P	Dir	P	Dir	P value	Dir	P value	Dir	P	Dir	P value	Dir	P value	Dir	P
BD	−	0.051	−	0.62	−	0.026	+	0.37	+	0.11	+	0.96	−	0.049	−	0.88	−	0.059
CHD	+	0.37	+	0.17	+	0.57	+	0.028	+	0.80	+	0.079	+	1.7×10^{-3}	+	9.2×10^{-3}	+	0.049
HT	−	0.76	−	0.58	+	0.76	+	0.20	+	0.23	+	0.53	−	0.011	−	0.75	−	0.012
CD	−	0.97	+	0.90	+	0.99	+	2.9×10^{-4}	+	0.051	+	0.011	−	0.73	−	0.76	−	0.71
RA	−	0.18	+	0.15	−	0.085	+	0.17	+	0.028	+	0.69	−	0.26	−	0.25	−	0.50
T1D	−	0.97	+	0.77	+	0.85	+	0.020	+	0.15	+	0.033	−	0.018	+	0.58	−	0.20
T2D	+	$<2\times10^{-16}$	+	4.3×10^{-7}	+	1.8×10^{-12}	+	7.6×10^{-8}	+	0.50	+	2.1×10^{-7}	+	0.66	−	0.12	+	0.48

See Tables S1 through S3 for a complete list of results.

BD = Bipolar Disorder; CHD = Coronary Heart Disease; HT = Hypertension; CD = Crohn's Disease; RA = Rheumatoid Arthritis; T1D = Type 1 Diabetes; T2D = Type 2 Diabetes.

Dir = Direction of effect; P = P value.

doi:10.1371/journal.pgen.1003919.t001

Conclusions

- Possible to identify potentially causal relationships using approach
- In theory could scale the approach up to examine thousands of intermediate phenotypes
- Genome-wide allelic scores lack specificity

OPEN ACCESS Freely available online

PLOS | GENETICS

Mining the Human Phenome Using Allelic Scores That Index Biological Intermediates

David M. Evans[1,2,3]*, Marie Jo A. Brion[1,2,4,5¤], Lavinia Paternoster[1,2], John P. Kemp[1,2], George McMahon[1,2], Marcus Munafò[6], John B. Whitfield[7], Sarah E. Medland[7], Grant W. Montgomery[7], The GIANT consortium[¶], The CRP consortium[¶], The TAG Consortium[¶], Nicholas J. Timpson[1,2], Beate St. Pourcain[1,2], Debbie A. Lawlor[1,2], Nicholas G. Martin[7], Abbas Dehghan[8], Joel Hirschhorn[4,9,10], George Davey Smith[1,2]

1 MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom, 2 School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, 3 University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland, Australia, 4 Broad Institute at MIT and Harvard, Cambridge, Massachusetts, United States of America, 5 Queensland Brain Institute, University of Queensland, Brisbane, Australia, 6 UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol, United Kingdom, 7 QIMR Berghofer Medical Research Institute, Brisbane, Australia, 8 Department of Epidemiology, Erasmus Medical Centre, Rotterdam, The Netherlands, 9 Division of Genetics and Endocrinology and Program in Genomics, Children's Hospital, Boston, Massachusetts, United States of America, 10 Department of Genetics, Harvard Medical School, Boston, Massachusetts

Abstract

It is common practice in genome-wide association studies (GWAS) to focus on the relationship between disease risk and genetic variants one marker at a time. When relevant genes are identified it is often possible to implicate biological intermediates and pathways likely to be involved in disease aetiology. However, single genetic variants typically explain small amounts of disease risk. Our idea is to construct allelic scores that explain greater proportions of the variance in biological intermediates, and subsequently use these scores to data mine GWAS. To investigate the approach's properties, we indexed three biological intermediates where the results of large GWAS meta-analyses were available: body mass index, C-reactive protein and low density lipoprotein levels. We generated allelic scores in the Avon Longitudinal Study of Parents and Children, and in publicly available data from the first Wellcome Trust Case Control Consortium. We compared the explanatory ability of allelic scores in terms of their capacity to proxy for the intermediate of interest, and the extent to which they associated with disease. We found that allelic scores derived from known variants and allelic scores derived from hundreds of thousands of genetic markers explained significant portions of the variance in biological intermediates of interest, and many of these scores showed expected correlations with disease. Genome-wide allelic scores however tended to lack specificity suggesting that they should be used with caution and perhaps only to proxy biological intermediates for which there are no known individual variants. Power calculations confirm the feasibility of extending our strategy to the analysis of tens of thousands of molecular phenotypes in large genome-wide meta-analyses. We conclude that our method represents a simple way in which potentially tens of thousands of molecular phenotypes could be screened for causal relationships with disease without having to expensively measure these variables in individual disease collections.

Citation: Evans DM, Brion MJA, Paternoster L, Kemp JP, McMahon G, et al. (2013) Mining the Human Phenome Using Allelic Scores That Index Biological Intermediates. PLoS Genet 9(10): e1003919. doi:10.1371/journal.pgen.1003919

Editor: Takashi Gojobori, National Institute of Genetics, Japan

Received: March 30, 2013; **Accepted:** September 12, 2013; **Published:** October 31, 2013

Copyright: © 2013 Evans et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The UK Medical Research Council (grant 74882), the Wellcome Trust (grant 076467) and the University of Bristol provide core support for ALSPAC. We thank 23andMe for funding the genotyping of the ALSPAC children's sample. Funding to pay the Open Access publication charge for this article was provided by the Wellcome Trust. MJAB was funded by the Wellcome Trust (grant 085553) and Leducq Foundation. JPK is funded by a Wellcome Trust grant (WT093431MA). AD is supported by NWO grant (veni, 916.12.154) and the EUR Fellowship. The QIMR Twin-Family Studies were supported by NIH grants (AA07535, AA07728, AA13320, AA13321, AA13326, AA14041, AA11998, AA17688, DA018854, DA019951); by grants from the Australian National Health and Medical Research Council (241944, 339462, 389927, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 552485, 552498); from the Australian Research Council (A7960034, A79906588, A79801419, DP0770096, DP0212016, DP0343921); and the FP6 EuroSTRESS Project (QLG2-CT-2002-01254). Genotyping was partially supported by grant AA13320 to the late Richard Todd, PhD, MD. GWM is supported by the National Health and Medical Research Council (NHMRC) Fellowship Scheme. The research leading to this work has received funding from the EU 7th Framework Programme under grant agreement number 247640, GEnoCoDE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dave.evans@bristol.ac.uk

¤ These authors contributed equally to this work.

¶ Membership of the GIANT consortium, the CRP consortium, and the TAG Consortium is provided in the Supporting Information.

Introduction

It is common practice within genome-wide association studies (GWAS) and their meta-analyses to focus on the relationship between disease risk and single nucleotide polymorphisms (SNPs) one genetic variant at a time. This strategy is often very informative in terms of identifying biological intermediates and/or pathways likely to be important in disease pathogenesis. For example, the association between coronary heart disease and genetic variants located within genes regulating levels of low

Individual Risk Prediction

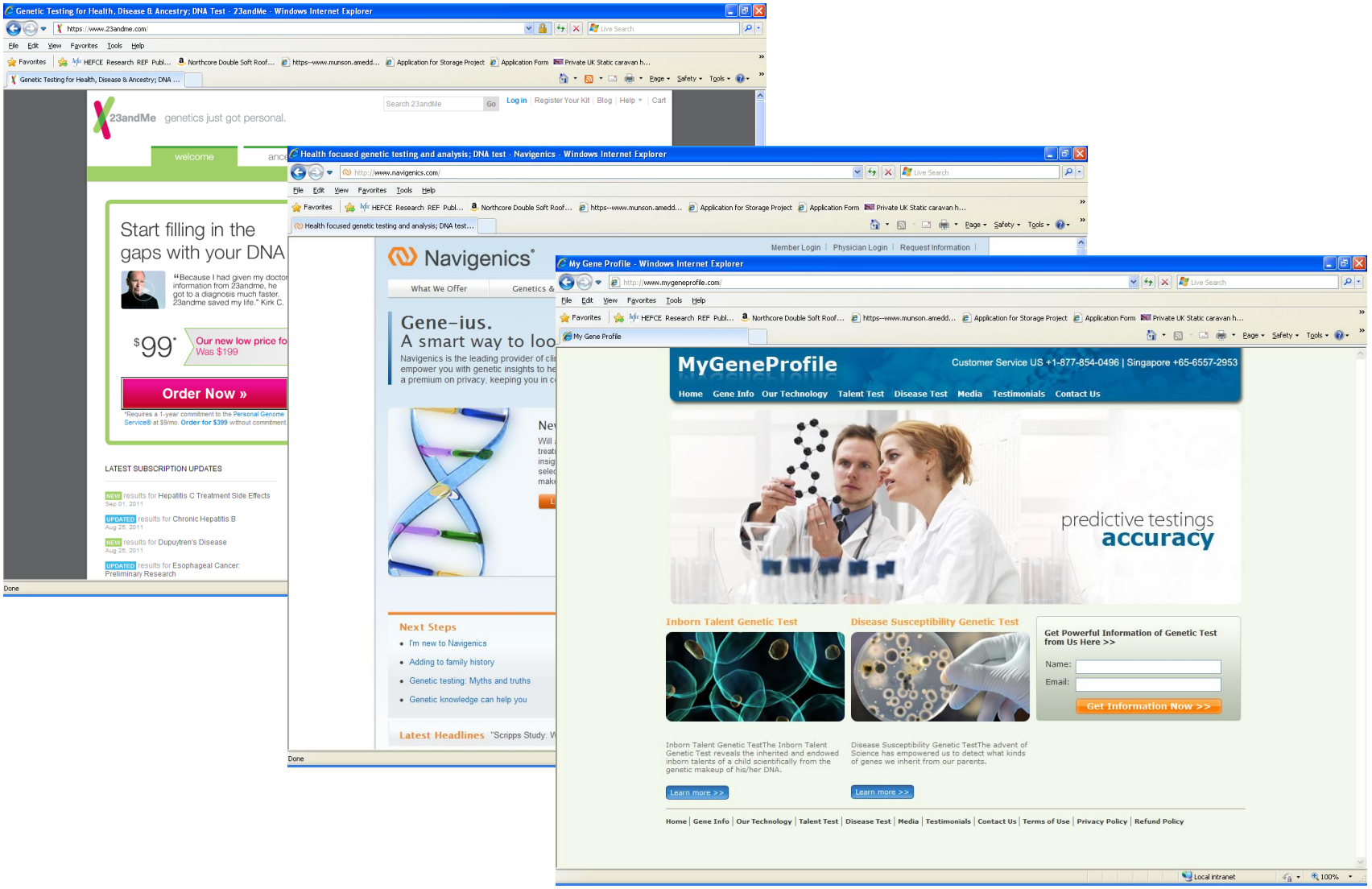
Case Study: Patient X

Age: *40*

Sex: *Male*

Ethnicity: *American*

Direct to Consumer Genetic Testing



My Home
Inbox

My Health
Disease Risk
Carrier Status
Drug Response
Traits
Health Labs

My Ancestry
Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs

Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
Genome Sharing

23andWe
Research Surveys (24)
Research Snippets
Research Initiatives
Research Discoveries

disease risk

Share my health results with family and friends

Show results for [redacted]

See new and recently updated reports »

23andWe Discoveries were made possible by 23andMe members who took surveys.

Elevated Risk

Name	Confidence	Your Risk	Avg. Risk	Compared to Average
Lung Cancer	★★★★	11.6%	8.5%	1.37x
Rheumatoid Arthritis	★★★★	3.2%	2.4%	1.36x
Esophageal Squamous Cell Carcinoma (ESCC)	★★★★	0.4%	0.4%	1.21x
Stomach Cancer (Gastric Cardia Adenocarcinoma)	★★★★	0.3%	0.2%	1.22x
Primary Biliary Cirrhosis new	★★★★	0.1%	0.1%	1.43x
Chronic Lymphocytic Leukemia	★★★			
Keloid	★★★			
Osteoarthritis	★★★			
Primary Biliary Cirrhosis: Preliminary Research update	★★★			
Squamous Cell Carcinoma	★★★			
Stroke	★★★			
Chronic Obstructive Pulmonary Disease (COPD)	★★			
Developmental Dyslexia	★★			
Gout	★★			
Sjögren's Syndrome	★★			

Decreased Risk

Name	Confidence	Your Risk	Avg. Risk	Compared to Average
Psoriasis	★★★★	7.1%	11.4%	0.62x
Alzheimer's Disease	★★★★	4.9%	7.2%	0.69x
Age-related Macular Degeneration	★★★★	4.0%	6.5%	0.60x
Melanoma	★★★★	2.2%	2.9%	0.75x
Restless Legs Syndrome	★★★★	0.9%	2.0%	0.44x
Type 1 Diabetes	★★★★	0.6%	1.0%	0.56x
Crohn's Disease	★★★★	0.4%	0.5%	0.69x
Multiple Sclerosis	★★★★	0.2%	0.3%	0.69x
Exfoliation Glaucoma	★★★★	0.2%	0.7%	0.22x
Celiac Disease	★★★★	0.05%	0.12%	0.41x
Atopic Dermatitis	★★★			
Atrial Fibrillation: Preliminary Research	★★★			
Basal Cell Carcinoma	★★★			
Bipolar Disorder: Preliminary Research	★★★			
Breast Cancer Risk Modifiers	★★★			
Cluster Headaches	★★★			
Kidney Cancer	★★★			
Kidney Disease	★★★			
Neuroblastoma	★★★			
Paget's Disease of Bone	★★★			
Pancreatic cancer	★★★			
Progressive Supranuclear Palsy update	★★★			
Back Pain	★★			
Creutzfeldt-Jakob Disease	★★			

Typical Risk

Name	Confidence	Your Risk	Avg. Risk	Compared to Average
Obesity	★★★★	54.2%	63.9%	0.85x
Coronary Heart Disease	★★★★	53.3%	46.8%	1.14x
Type 2 Diabetes	★★★★	27.2%	25.7%	1.06x
Atrial Fibrillation	★★★★	23.0%	27.2%	0.85x
Prostate Cancer ♂	★★★★	16.4%	17.8%	0.92x
Venous Thromboembolism	★★★★	11.9%	12.3%	0.96x
Gallstones	★★★★	6.2%	7.0%	0.88x
Colorectal Cancer	★★★★	5.9%	5.6%	1.05x
Chronic Kidney Disease	★★★★	3.0%	3.4%	0.87x
Parkinson's Disease	★★★★	1.4%	1.6%	0.85x
Ulcerative Colitis	★★★★	0.9%	0.8%	1.15x
Bipolar Disorder	★★★★	0.10%	0.10%	0.94x
Scleroderma (Limited Cutaneous Type)	★★★★	0.05%	0.07%	0.80x
Breast Cancer ♀	★★★★	0.00%	0.00%	1.00x
Lupus (Systemic Lupus Erythematosus) ♀	★★★★	0.00%	0.00%	1.00x
Alopecia Areata	★★★			
Ankylosing Spondylitis	★★★			
Asthma	★★★			
Behçet's Disease	★★★			
Brain Aneurysm	★★★			
Coronary Heart Disease: Preliminary Research	★★★			
Follicular Lymphoma	★★★			
Generalized Vitiligo	★★★			
High Blood Pressure (Hypertension) update	★★★			

My Home
Inbox

My Health
Disease Risk
Carrier Status
Drug Response
Traits
Health Labs

My Ancestry
Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs

Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
Genome Sharing

23andWe
Research Surveys (25)
Research Snippets
Research Initiatives
Research Discoveries

traits

Share my health results with family and friends

Show results for [redacted]

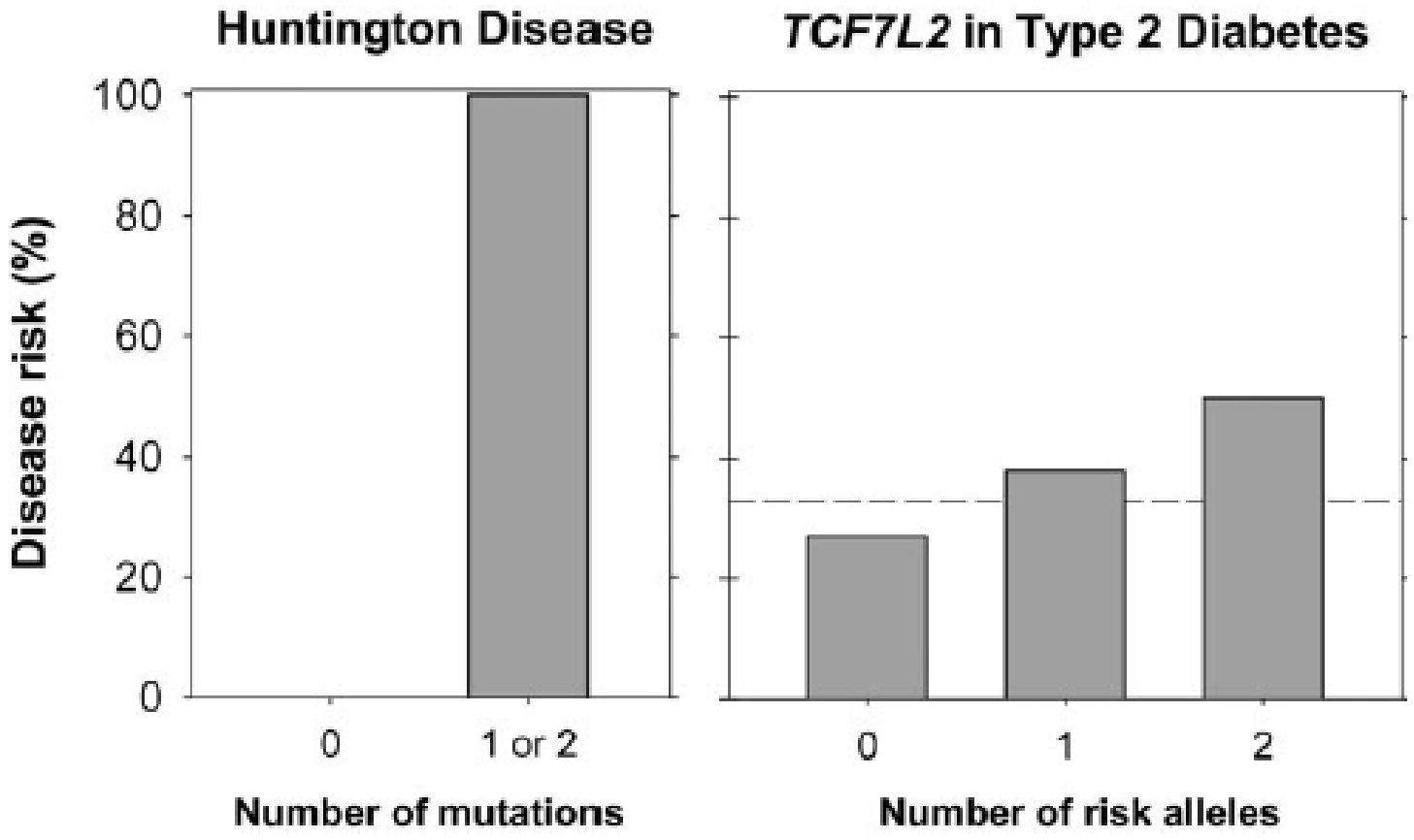
See new and recently updated reports »

23andWe Discoveries were made possible by 23andMe members who took surveys.

Name	Confidence ▲	Outcome
Alcohol Flush Reaction	★★★★	Does Not Flush
Bitter Taste Perception	★★★★	Unlikely to Taste
Earwax Type	★★★★	Wet
Eye Color	★★★★	Likely Brown
Hair Curl	★★★★	Straighter Hair on Average
Lactose Intolerance	★★★★	Likely Tolerant
Malaria Resistance (Duffy Antigen)	★★★★	Not Resistant
Male Pattern Baldness ♂	★★★★	Increased Odds
Muscle Performance	★★★★	Likely Sprinter
Non-ABO Blood Groups	★★★★	See Report
Norovirus Resistance	★★★★	Not Resistant
Resistance to HIV/AIDS	★★★★	Not Resistant
Smoking Behavior	★★★★	If a Smoker, Likely to Smoke More
Adiponectin Levels	★★★	See Report
Asparagus Metabolite Detection	★★★	Typical Odds of Detecting
Birth Weight	★★★	See Report
Blood Glucose	★★★	5.18 mmol/L on Average
Breastfeeding and IQ	★★★	See Report

Is it Going to Work? / Is there anything to be worried about?

Most Variants are of Small Effect



Janssens & van Duijn (2008) HMG

What Else Could We Do?

Genomic Profiling

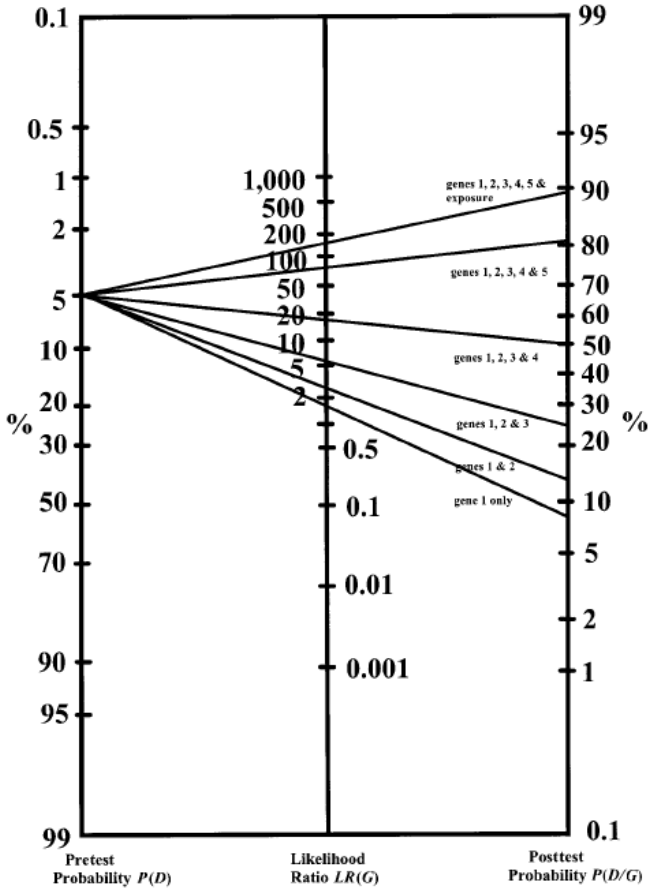
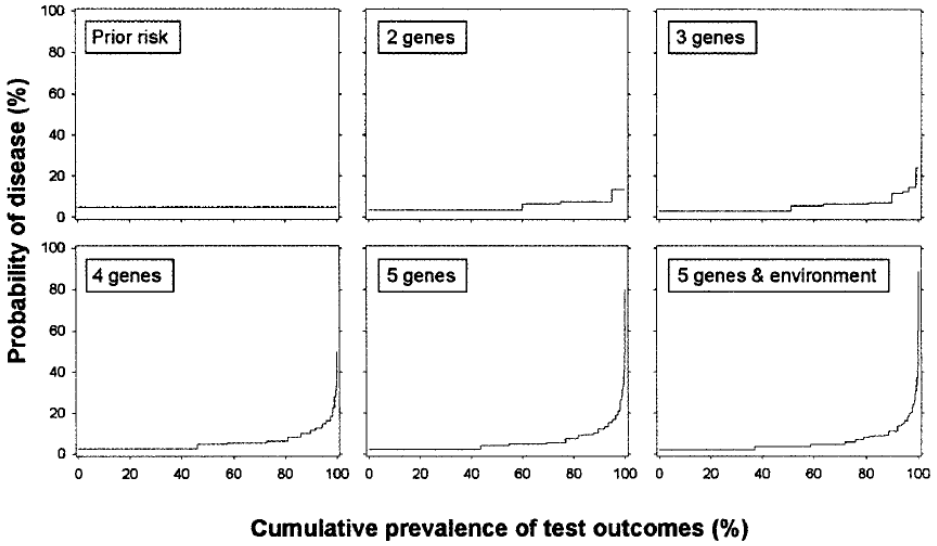


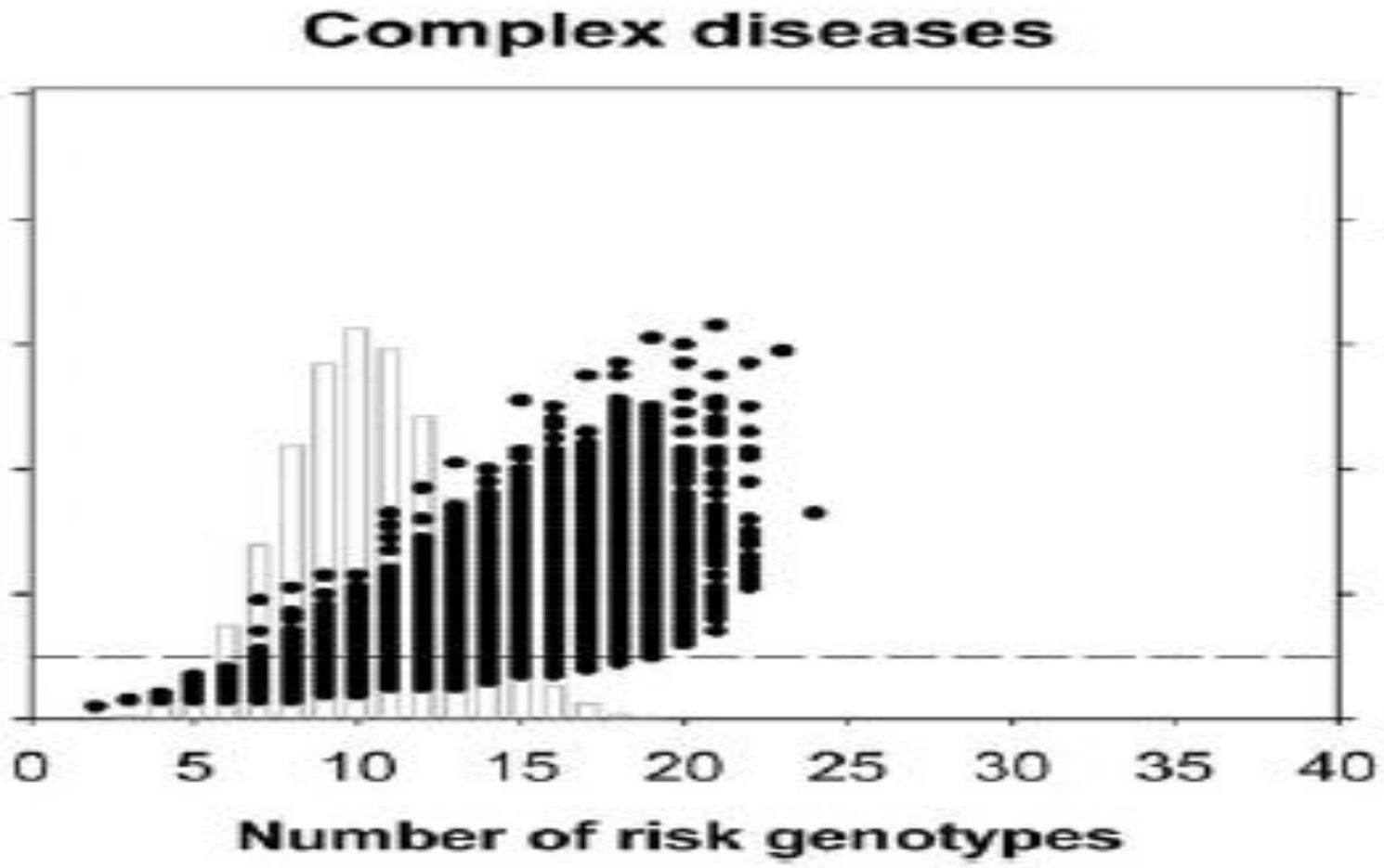
Figure 1 Power of a panel of genetic tests and exposure on predictability of the common disease (simulated data)

(from Yang et al. 2003 AJHG)



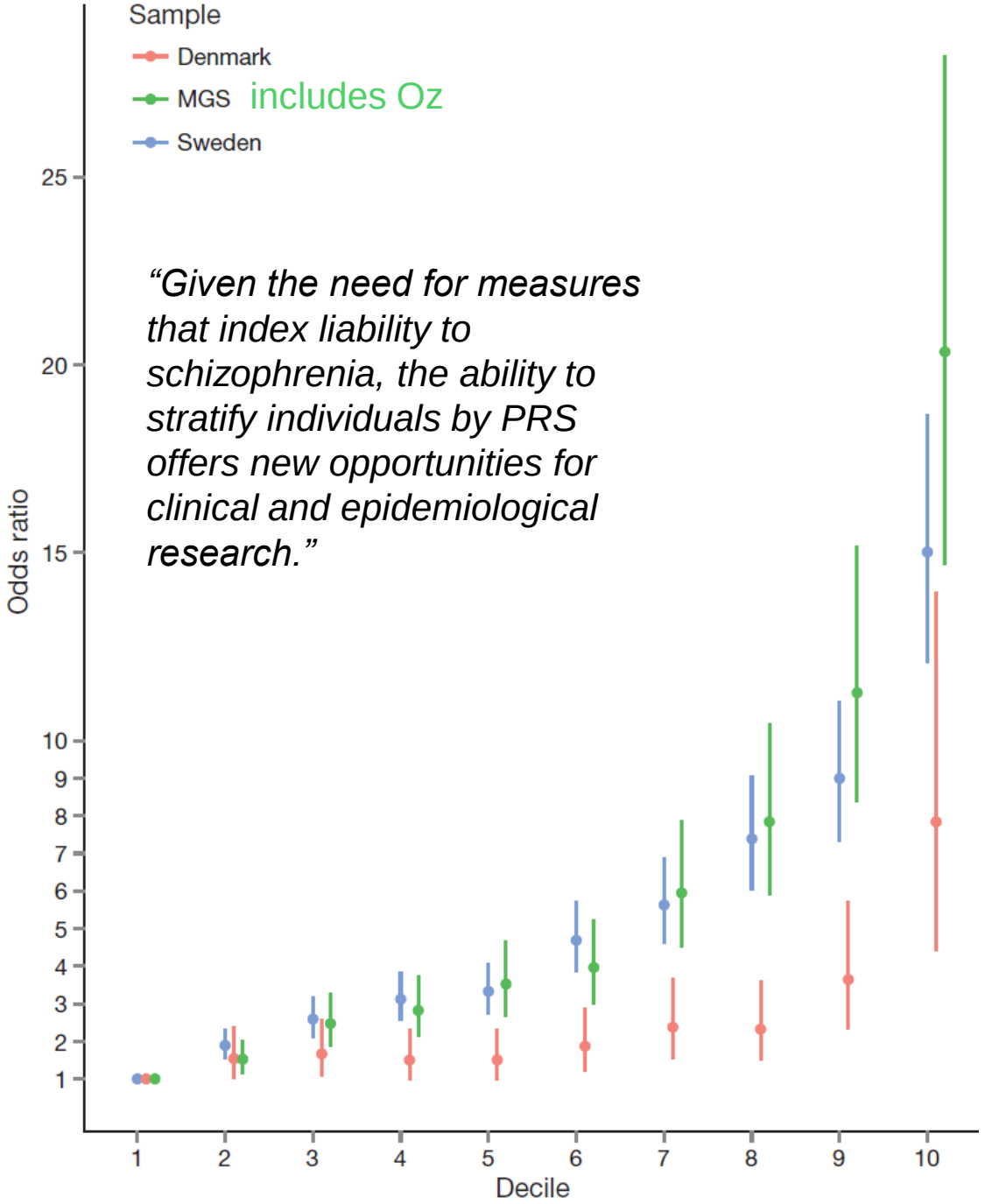
(from Janssens et al. 2004 AJHG)

Most Individuals Carry a Mixture of High and Low Risk Genotypes



Janssens & van Duijn (2008) HMG

Odds ratio for schizophrenia by polygenic risk score (PRS) decile in the Sweden, Denmark, and Molecular Genetics of Schizophrenia (US+Oz) studies.



"Given the need for measures that index liability to schizophrenia, the ability to stratify individuals by PRS offers new opportunities for clinical and epidemiological research."

Risk alleles and weights were derived from 'leave one out' analyses in which those samples were excluded from the GWAS meta-analysis. The threshold for selecting risk alleles was $P<0.05$.

The Predictive Utility of Genetic Variants is Limited By Heritability

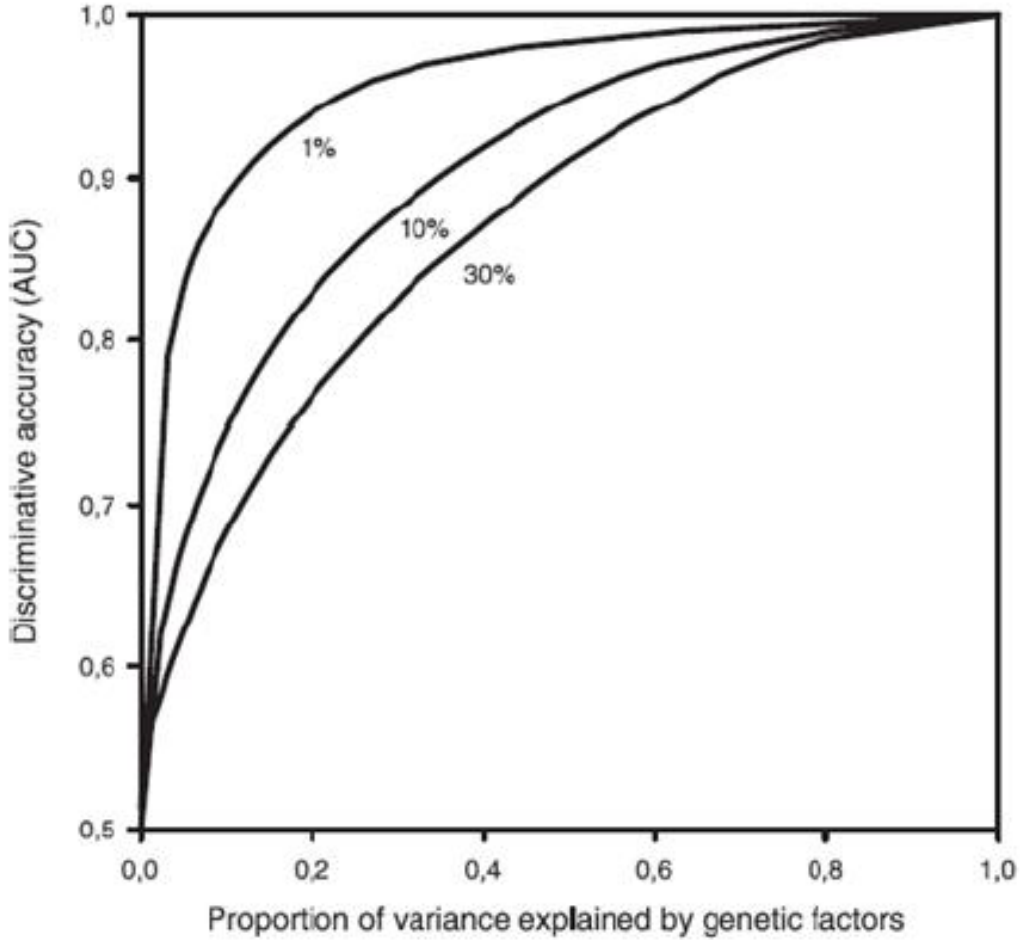
Table 1 Heritability estimates of various complex diseases and traits

Disease or trait	Heritability	Reference
Eye color	> 99%	[18]
Type 1 diabetes	88%	[19]
Schizophrenia	81%	[20]
Alzheimer's disease	79%	[21]
Height	70-87% (m), 68-85% (v)	[22]
Obesity	65-84% (m), 64-79% (w)	[23]
Smoking persistence	59% (m), 46% (w)	[24]
Anorexia nervosa	56%	[25]
Rheumatoid arthritis	53-65%	[26]
Panic disorder	43%	[27]
Prostate cancer	42%	[28]
Migraine	40-50%	[29]
Heart attack	38% (m), 57% (w)	[30]
Smoking initiation	37% (m), 55% (w)	[24]
Depression	37%	[31]
Colorectal cancer	35%	[28]
Anxiety disorder	32%	[27]
Homosexuality	30% (m), 50-60% (w)	[32]
Breast cancer	27%	[28]
Type 2 diabetes	26%	[33]
Lung cancer	26%	[28]
Happiness	22% (m), 41% (w)	[34]

Heritability and frequency estimates are obtained from published studies and meta-analyses. m = men, w = women.

Janssens & van Duijn (2010) Investigative Genetics

The Predictive Utility of Genetic Variants is Limited By Heritability



Janssens & van Duijn (2010) Investigative Genetics

The Majority of Heritability for Most Diseases is Yet to Be Explained

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.

Maher (2009) Nature

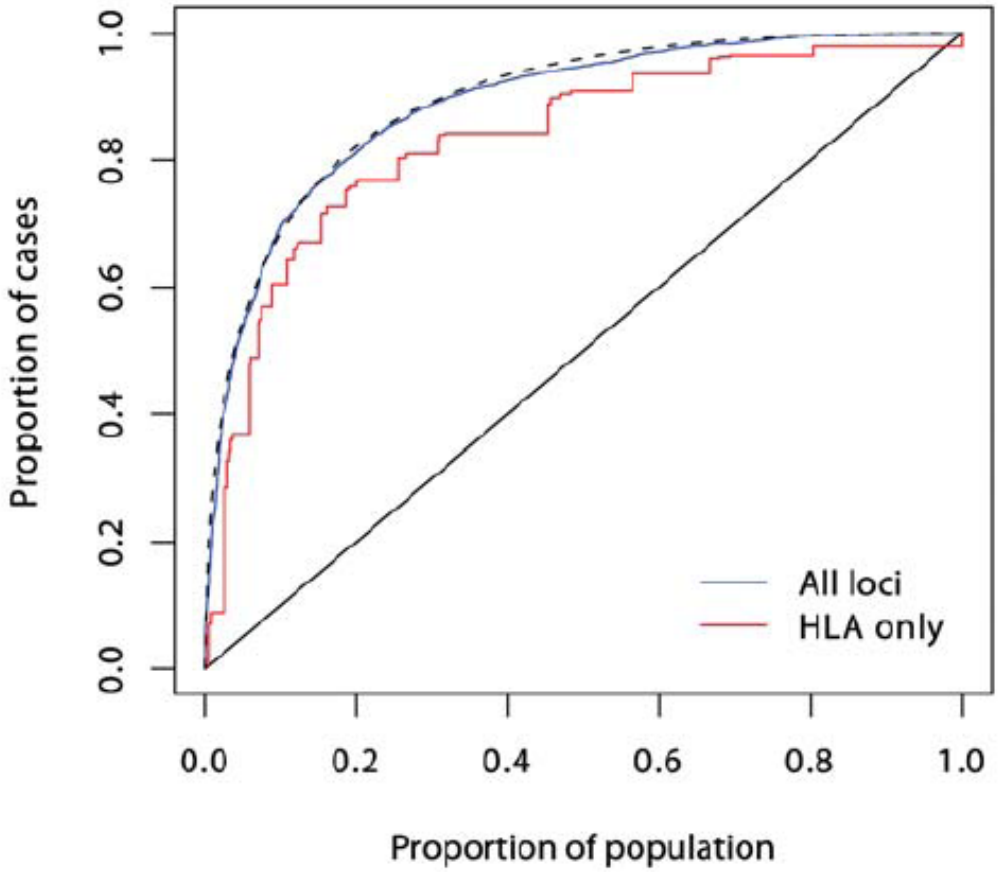
Typical AUC Values

Table 1. Recent studies on the prediction of complex diseases using multiple genes

Disease	Genetic variants	Variant selection[a]	AUC
Age-related macular degeneration	*CFH* Y402H, *CFH* rs1410996, *LOC387715* A69S, *C2-CFB*	5 (out of 1536 tag SNPs in established genes)	0.80[b]
Coronary heart disease	*UCP2* G(-866)A, *APOE* e2/3/4, *LPL* D9N, *APOA4* T347S	4 (out of 12)	0.62
Coronary heart disease	*AGT* T4072C, *ACE* I/D, *AGTR1* A1166C, *CYP11B2* C(-344)T, *ADD1* G614T, *GNB3* C825T	6 established variants	0.55[c]
Hypertriglyceridemia	*APOA5* S19W, *APOA5* T(-1131)C, *APOE e/3/4, GCKR* rs780094, *TRIB1* rs17321515, *TBL2/MLXIPL* rs17145738, *GALNT2* rs4846914	7 established variants	0.80
MI after surgery	*IL6* G572C, *ICAM1* K469E, *SELE* G98T	3 (out of 48)	0.70
Systemic lupus erythematosus	*PXK* rs6445975, *HLA* region rs3131379 and rs9275572, *IRF5/TNPO3* rs12537284, *KIAA1542* rs4963128, *ITGAM* rs9888739	From GWAS	0.67
Type 2 diabetes	*KCNJ11* G23L, *PPARG* P12A, *TCF7L2* rs7903146	3 established variants	0.55
Type 2 diabetes	*GCK* G(−30G)A, *IL6* G(−174)C, *TCF7L2* rs7903146	3 (out of 19)	0.56
Type 2 diabetes	*SNPs in TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX, THADA, JAZF1*	18 established variants	0.60
Type 2 diabetes	*SNPs in TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX, THADA, JAZF1*	18 established variants	0.60

Janssens & van Duijn (2008) HMG

GWAS' greatest success: T1D



Current known loci explain a λ_s of just under five, as compared with the value of 15 often quoted. However, it is likely that the latter figure is exaggerated, and the λ_s attributable to inheritance is likely to be less than ten. The heritability explained will be increased to some degree when the known regions are more fully studied, but the bulk of the remaining heritability is likely to be attributable to many small (or rare) effects, most of which are unlikely to be mapped. Thus, even for this highly heritable disease, the prediction achievable could fall some way short of that required for a targeted prevention strategy.

Figure 5. ROC curve prediction from all the SNPs listed in Supplementary Table 1 in Text S1 (in blue). The prediction curve using the six MHC SNPs alone is shown in red, and the dashed curve corresponds to a polygenic multiplicative model with $\lambda_s = 4.75$.

PLoS GENETICS
OPEN ACCESS Freely available online
Viewpoints

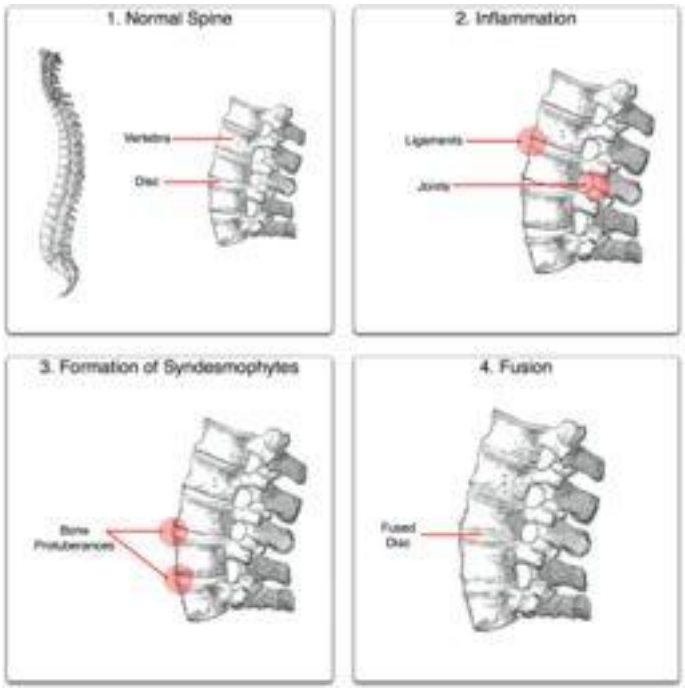
Prediction and Interaction in Complex Disease Genetics: Experience in Type 1 Diabetes

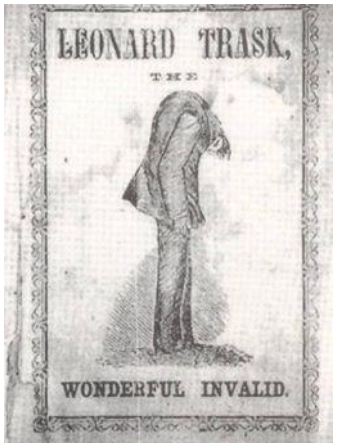
David G. Clayton*

Problems...

- Almost all the diagnostic utility is driven by the HLA
- What is the clinical utility?
- T1D is NOT the complex disease that has the highest AUC

Ankylosing Spondylitis





NIH Public Access
Author Manuscript
Nat Genet. Author manuscript; available in PMC 2013 May 01.

Published in final edited form as:
Nat Genet. ; 43(8): 761–767. doi:10.1038/ng.873.

Interaction between *ERAP1* and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility

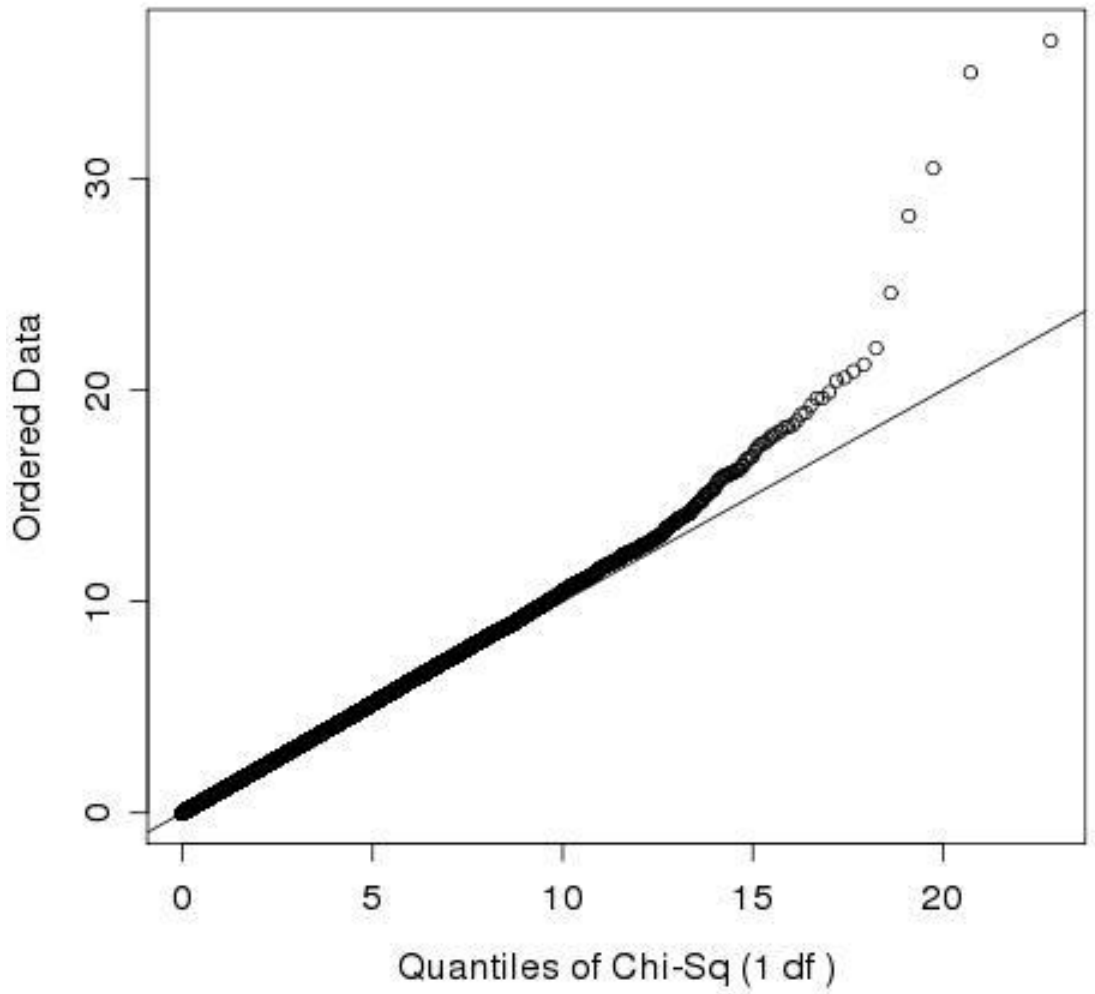
David M Evans[1,43], Chris C A Spencer[2,43], Jennifer J Pointon[3], Zhan Su[2], David Harvey[3],

- Auto-immune arthritis resulting in fusion of vertebrae
- Sensitivity and Specificity about 90% using HLA-B27
- Diagnostically useful
- A single SNP can tag HLA-B27 status with ~100% accuracy

What Else Could We Do?

Genome-wide Information…

Genome-wide Prediction?



Genome-wide Prediction?

Human Molecular Genetics, 2009, Vol. 18, No. 18 **3525–3531**
doi:10.1093/hmg/ddp295
Advance Access published on June 24, 2009

Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk

David M. Evans[1,*], Peter M. Visscher[2] and Naomi R. Wray[2]

[1]Department of Social Medicine, MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Bristol, UK and [2]Genetic Epidemiology and Queensland Statistical Genetics, Queensland Institute of Medical Research, Australia

Wellcome Trust Case-Control Consortium

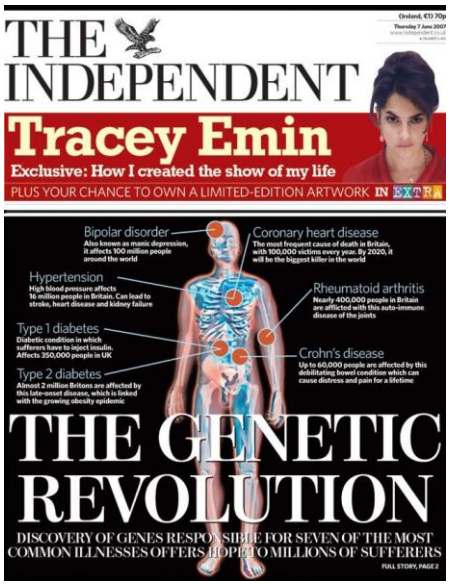
Genome-Wide Association Across Major Human Diseases

DESIGN

Collaboration amongst 26 UK disease investigators
2000 cases each from 7 diseases

GENOTYPING

Affymetrix 500k SNPs



CASES

1. Type 1 Diabetes
2. Type 2 Diabetes
3. Crohn's Disease
4. Coronary Heart Disease
5. Hypertension
6. Bipolar Disorder
7. Rheumatoid Arthritis

CONTROLS

1. UK Controls A (1,500 - 1958 BC)

Individual Risk Prediction

Table 2. Median AUC values for known variants and known variants plus genome-wide scores combined

Threshold	BD	CHD	CD	RA	T1D	T2D
Count method						
Known	0.549	0.572	0.769	0.701	0.784	0.666
0.8	0.657 (0.564)	0.624 (0.579)	0.782 (0.770)	0.716 (0.703)	0.793 (0.784)	0.702 (0.670)
0.5	0.671 (0.566)	0.619 (0.576)	0.780 (0.770)	0.718 (0.704)	0.794 (0.785)	0.670 (0.667)
0.1	0.651 (0.561)	0.593 (0.581)	0.771 (0.770)	0.718 (0.712)	0.787 (0.785)	0.690 (0.667)
0.05	0.656 (0.556)	0.589 (0.580)	0.770 (0.771)	0.715 (0.712)	0.787 (0.785)	0.686 (0.667)
0.01	0.608 (0.584)	0.608 (0.569)	0.770 (0.771)	0.716 (0.708)	0.788 (0.785)	0.669 (0.665)
0.001	0.563 (0.561)	0.597 (0.572)	0.770 (0.770)	0.710 (0.709)	0.786 (0.785)	0.668 (0.665)
0.0001	0.574 (0.561)	0.576 (0.576)	0.771 (0.770)	0.709 (0.709)	0.785 (0.787)	0.669 (0.669)
0.00001	0.561 (0.562)	0.579 (0.578)	0.770 (0.769)	0.703 (0.712)	0.785 (0.786)	0.669 (0.668)
Log odds method						
0.8	0.678 (0.572)	0.618 (0.585)	0.779 (0.770)	0.718 (0.708)	0.792 (0.786)	0.707 (0.668)
0.5	0.674 (0.566)	0.617 (0.580)	0.778 (0.770)	0.719 (0.709)	0.793 (0.786)	0.707 (0.666)
0.1	0.641 (0.562)	0.595 (0.583)	0.772 (0.770)	0.718 (0.715)	0.788 (0.785)	0.696 (0.667)
0.05	0.641 (0.562)	0.594 (0.579)	0.769 (0.771)	0.718 (0.715)	0.788 (0.786)	0.687 (0.667)
0.01	0.597 (0.579)	0.620 (0.573)	0.769 (0.772)	0.713 (0.711)	0.788 (0.785)	0.668 (0.666)
0.001	0.560 (0.563)	0.592 (0.576)	0.769 (0.770)	0.712 (0.714)	0.785 (0.784)	0.669 (0.667)
0.0001	0.569 (0.561)	0.577 (0.573)	0.770 (0.772)	0.710 (0.710)	0.784 (0.790)	0.667 (0.671)
0.00001	0.560 (0.562)	0.577 (0.581)	0.770 (0.770)	0.703 (0.713)	0.787 (0.785)	0.671 (0.673)

The first row displays the AUC achieved by using known variants only to discriminate case–control status. The values in the rows below this show the AUC achieved using known variant information combined with genome-wide scores. The values in plain font are the median AUC statistics produced when known variants plus nominally associated SNPs are used to discriminate case–control status for the same disease. The values in parenthesis are median AUC statistics produced when known variants for the disease of interest are combined with genome-wide scores derived from nominally associated bipolar SNPs (or coronary heart disease SNPs for bipolar cases). BD, bipolar disorder; CHD, coronary heart disease; HT, hypertension; CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type I diabetes; T2D, type II diabetes.

Is There A Future? Problems and Solutions (?)

- Individual risk prediction limited by heritability by definition
  - Can we include environmental predictors as well?
  - Only effective if not on causal path
- The majority of heritability for most traits is yet to be explained
  - Getting closer all the time
  - Polygenic / BLUP approaches
- You often do better by just looking at your parents
  - Including family information in test
- Results only relevant for a very small percentage of individuals
  - Population screening by WGS?
  - Include family information in the risk calculation?