Risk Prediction of Complex Disease

David Evans

University of
BRISTOL

Genetic Testing and Personalized Medicine

- The idea that diagnosis, preventative and therapeutic interventions are tailored to individuals based upon their genotypes

  Diagnosis -> Modification of risk

  Tailoring treatment options

- Predictive testing in the case of monogenic diseases has been used for years (1300+ tests available)

  Preventative strategies radical (PKU, Breast cancer)

- Is this possible also in complex diseases?

  Predictive utility of many different variants -> genomic profiling

  Environmental risk factors

		Reality		
		Diseased	Normal	
Test Outcome	Positive	True Positive	False Positive (Type I error)	Positive Predictive Value
	Negative	False Negative (Type II error)	True Negative	Negative Predictive Value
		Sensitivity	Specificity	

Sensitivity = P(T+ | D+)

A sensitivity of 100% means that the test recognises all sick people

“SNOUT”

Property of test itself

		Reality		
		Diseased	Normal	
Test Outcome	Positive	True Positive	False Positive (Type I error)	Positive Predictive Value
	Negative	False Negative (Type II error)	True Negative	Negative Predictive Value
		Sensitivity	Specificity	

Specificity = P(T- | D-)

A specificity of 100% means that the test identifies all healthy people as healthy

Positive results in a highly specific test is used to confirm disease

“SPIN”

Property of test itself

		Reality		
		Diseased	Normal	
Test Outcome	Positive	True Positive	False Positive (Type I error)	Positive Predictive Value
	Negative	False Negative (Type II error)	True Negative	Negative Predictive Value
		Sensitivity	Specificity	

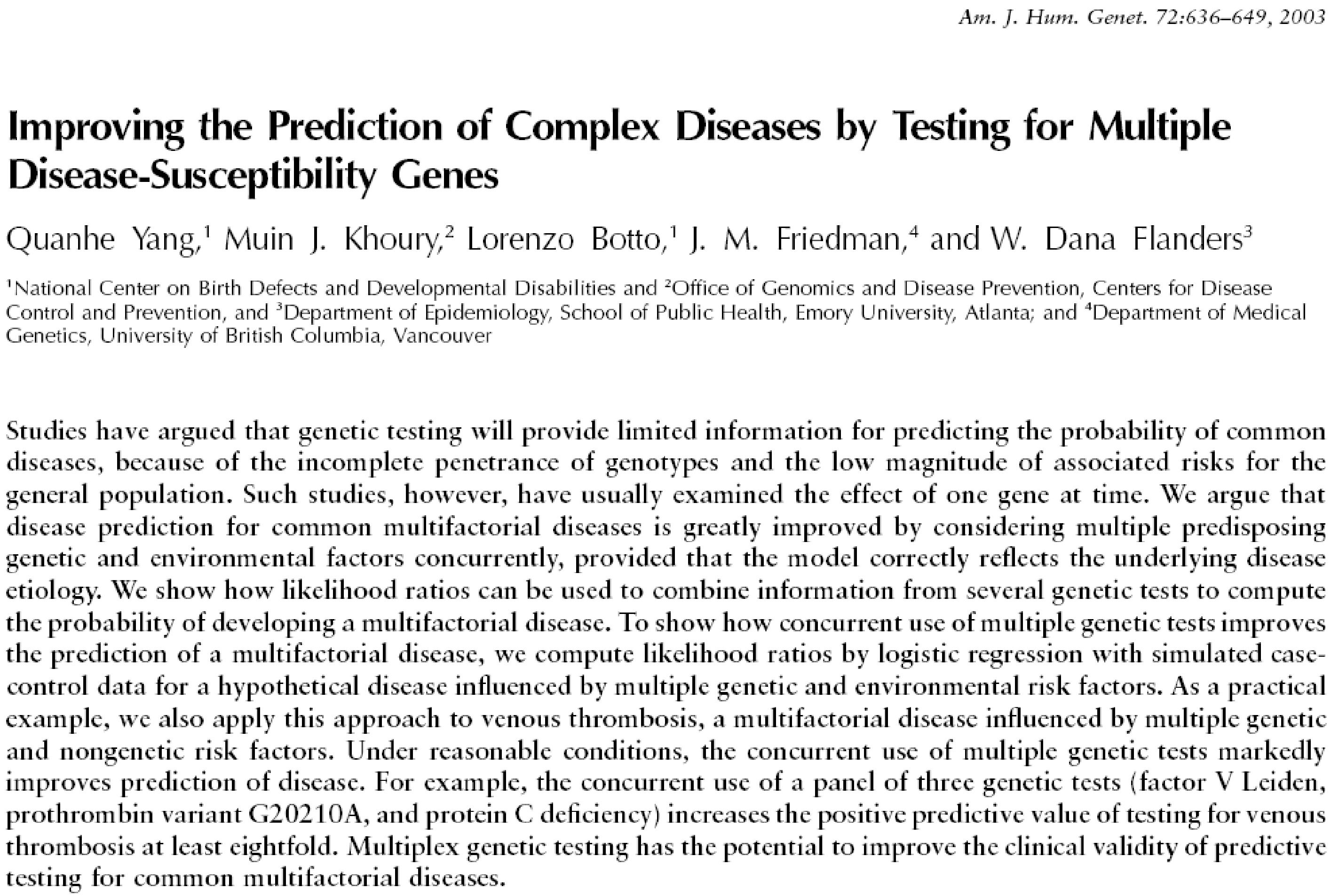
PPV = P(D+ | T+)

Depends on prevalence of disease

		Reality		
		Diseased	Normal	
Test Outcome	Positive	True Positive	False Positive (Type I error)	Positive Predictive Value
	Negative	False Negative (Type II error)	True Negative	Negative Predictive Value
		Sensitivity	Specificity	

NPV = P(D- | T-)

Depends on prevalence of disease



Am. J. Hum. Genet. 72:636–649, 2003

Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes

Quanhe Yang,[1] Muin J. Khoury,[2] Lorenzo Botto,[1] J. M. Friedman,[4] and W. Dana Flanders[3]

[1]National Center on Birth Defects and Developmental Disabilities and [2]Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, and [3]Department of Epidemiology, School of Public Health, Emory University, Atlanta; and [4]Department of Medical Genetics, University of British Columbia, Vancouver

Studies have argued that genetic testing will provide limited information for predicting the probability of common diseases, because of the incomplete penetrance of genotypes and the low magnitude of associated risks for the general population. Such studies, however, have usually examined the effect of one gene at time. We argue that disease prediction for common multifactorial diseases is greatly improved by considering multiple predisposing genetic and environmental factors concurrently, provided that the model correctly reflects the underlying disease etiology. We show how likelihood ratios can be used to combine information from several genetic tests to compute the probability of developing a multifactorial disease. To show how concurrent use of multiple genetic tests improves the prediction of a multifactorial disease, we compute likelihood ratios by logistic regression with simulated case-control data for a hypothetical disease influenced by multiple genetic and environmental risk factors. As a practical example, we also apply this approach to venous thrombosis, a multifactorial disease influenced by multiple genetic and nongenetic risk factors. Under reasonable conditions, the concurrent use of multiple genetic tests markedly improves prediction of disease. For example, the concurrent use of a panel of three genetic tests (factor V Leiden, prothrombin variant G20210A, and protein C deficiency) increases the positive predictive value of testing for venous thrombosis at least eightfold. Multiplex genetic testing has the potential to improve the clinical validity of predictive testing for common multifactorial diseases.

▷ Likelihood Ratio

$$LR(G) = \frac{P(G|D)}{P(G|\overline{D})}$$

$$LR(G) = LR(g_1)LR(g_2)\ldots LR(g_n)$$

$$\ln\left[\frac{P(D|G)}{P(\overline{D}|G)}\right] = \alpha_{\text{pop}} + \beta G^{T} . \tag{B1}$$

Applying Bayes's theorem, we have

$$\ln\left[\frac{P(D|G)}{P(\overline{D}|G)}\right] = \ln\left[\frac{P(G|D)P(D)}{P(G|\overline{D})P(\overline{D})}\right] = \ln\left[\frac{P(G|D)}{P(G|\overline{D})}\right] + \ln\left[\frac{P(D)}{P(\overline{D})}\right] .$$

Therefore, the likelihood ratio is

$$\ln LR_{\text{pop}}(G) = \ln\frac{N_{\overline{D}}}{N_D} + \ln\frac{P(D|G)}{P(\overline{D}|G)} = \ln\frac{N_{\overline{D}}}{N_D} + \alpha_{\text{pop}} + \beta G^{T}\,(\text{from eq. [B1]}) = \alpha^{*}_{\text{pop}} + \beta G^{T} ,$$

where α_{pop} is the intercept term in the population logistic model (background disease risk), N_D is the number of people in the population who develop the disease, $N_{\overline{D}}$ is the number of people in the population who do not develop the disease, $P(D) = N_D|(N_D + N_{\overline{D}})$, and $\alpha^{*}_{\text{pop}} = \alpha_{\text{pop}} + \ln(N_{\overline{D}}|N_D)$ (Albert 1982).

To prove the validity of estimating likelihood ratio from a case-control study, we introduce the dummy variable S to indicate whether an individual is selected for the case-control sample and denote the sampling fraction as $f_1 = P(S = 1|D)$ and $f_0 = P(S = 1|\overline{D})$. It is essential that the risk odds ratio in the case-control study estimates the risk ratio and the probability of being selected for a sample is independent of genotype in both those with and without the disease—that is, $P(S = 1|D,G) = P(S = 1|D)$ and $P(S = 1|\overline{D},G) = P(S = 1|\overline{D})$. We can compute the probability of disease, given a particular set of genetic test results, using a logistic model for the sample as

$$\ln\left[\frac{P(D|G,S = 1)}{P(\overline{D}|G,S = 1)}\right] = \ln\left[\frac{P(D|G)P(S = 1|D)/P(S|G)}{P(\overline{D}|G)P(S = 1|\overline{D})/P(S|G)}\right] = \ln\left[\frac{P(D|G)}{P(\overline{D}|G)}\right] + \ln\left(\frac{f_1}{f_0}\right) \tag{B2}$$

after cancellation of the denominator. Substitution of equation (B1) into equation (B2) gives

$$\ln\left[\frac{P(D|G,S = 1)}{P(\overline{D}|G,S = 1)}\right] = \alpha_{\text{pop}} + \beta G^{T} + \ln\left(\frac{f_1}{f_0}\right) = \alpha_{\text{CC}} + \beta G^{T} , \tag{B3}$$

where $\alpha_{\text{CC}} = \alpha_{\text{pop}} + \ln(f_1/f_0)$. Thus, the logistic model continues to apply in the sample with the same β coefficient but with an adjusted $\alpha^{*} = \alpha_{\text{pop}} + \ln(f_1/f_0)$ (Breslow et al. 1980).

Similar to the derivation of likelihood ratio estimated using logistic regression in the population, the likelihood ratio in the case-control study population is found to be

$$\boxed{\ln LR_{\text{CC}}(G) = \ln\frac{N_{\text{CO}}}{N_{\text{CA}}} + \ln\frac{P(D|G)}{P(\overline{D}|G)} = \ln\frac{N_{\text{CO}}}{N_{\text{CA}}} + \alpha_{\text{CC}} + \beta G^{T} ,} \tag{B4}$$

where $\alpha_{\text{cc}} = \alpha_{\text{pop}} + \ln(f_1/f_0)$ is the intercept term estimated from a case-control study, as shown in equation (B3). Because

$$\ln\left(\frac{f_1}{f_0}\right) = \ln\left(\frac{N_{\text{CA}}/N_D}{N_{\text{CO}}/N_{\overline{D}}}\right) = \ln\left(\frac{N_{\overline{D}}}{N_D}\right) - \ln\left(\frac{N_{\text{CO}}}{N_{\text{CA}}}\right) , \tag{B5}$$

substitution of equation (B5) into equation (B4) gives $\ln LR_{\text{pop}}(G) = \ln LR_{\text{cc}}(G)$.

Genomic Profiling

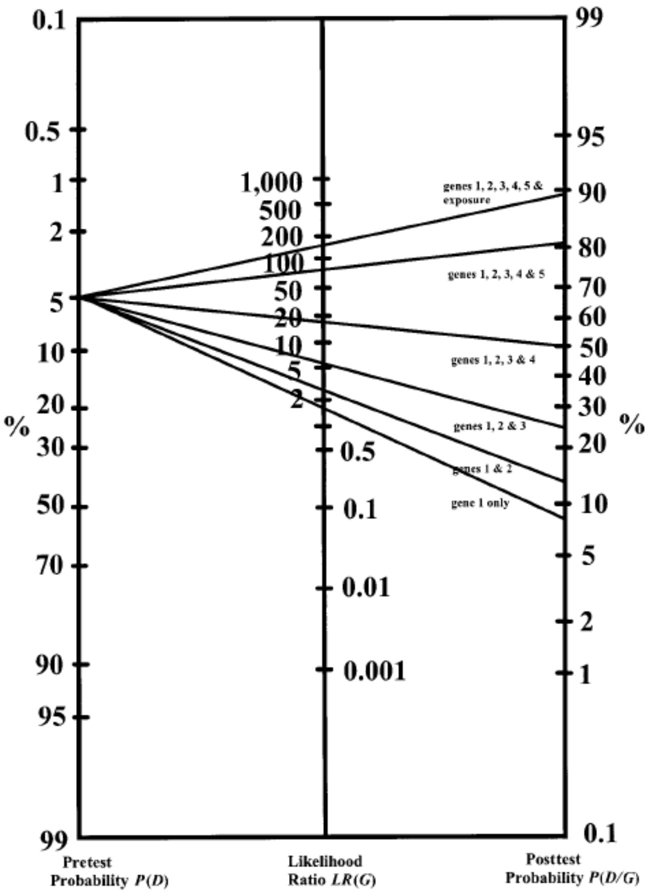
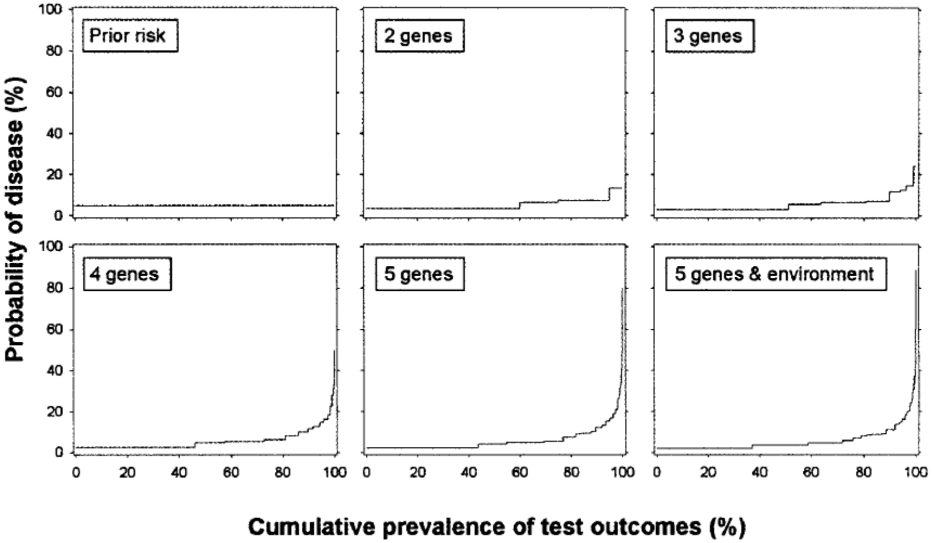


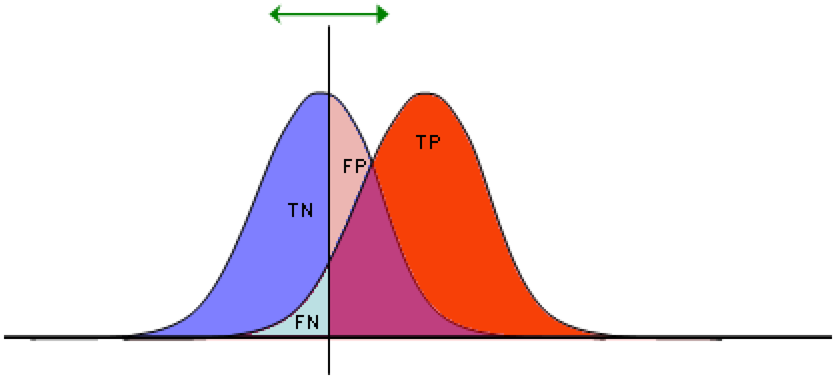
Figure 1 Power of a panel of genetic tests and exposure on predictability of the common disease (simulated data)

(from Yang et al. 2003 AJHG)



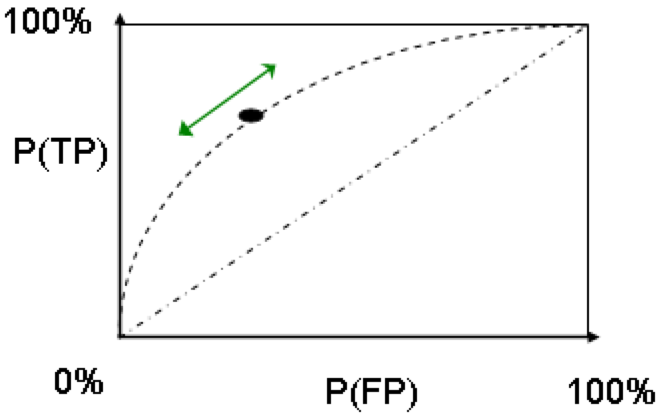
(from Janssens et al. 2004 AJHG)

ROC Curves



TP	FP
FN	TN
1	1

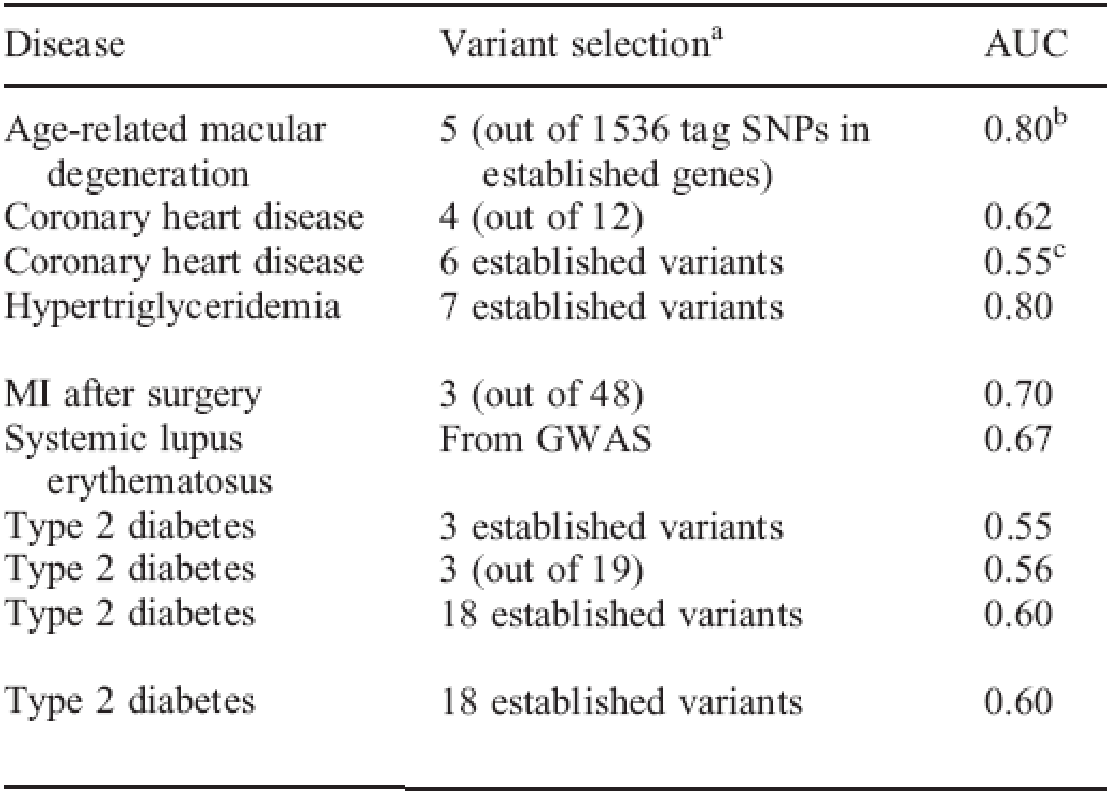
Sensitivity



1 - Specificity

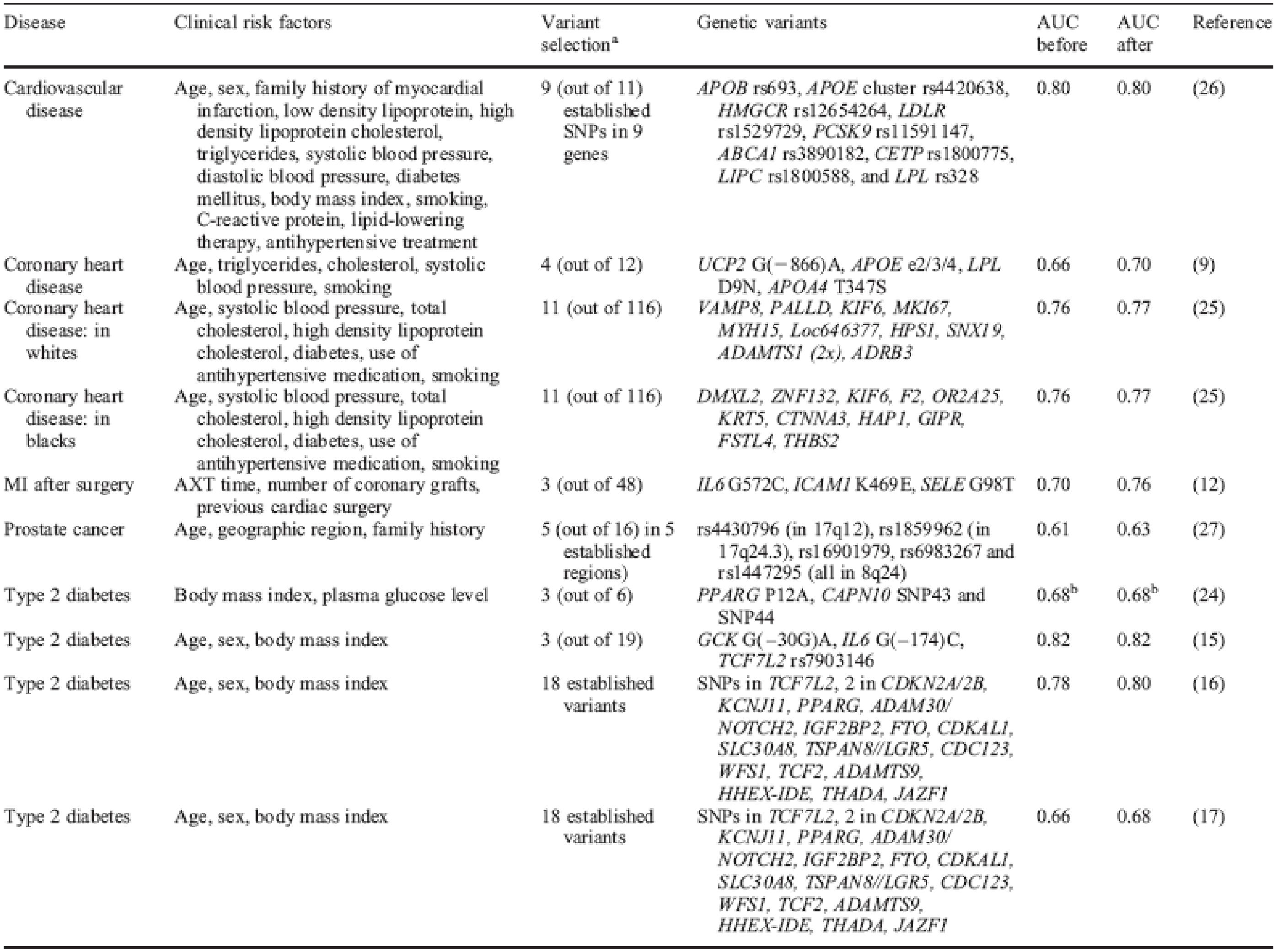
- Area under Curve (AUC) 0.5 - 1

Genomic Profiling



Disease	Variant selection[a]	AUC
Age-related macular degeneration	5 (out of 1536 tag SNPs in established genes)	0.80[b]
Coronary heart disease	4 (out of 12)	0.62
Coronary heart disease	6 established variants	0.55[c]
Hypertriglyceridemia	7 established variants	0.80
MI after surgery	3 (out of 48)	0.70
Systemic lupus erythematosus	From GWAS	0.67
Type 2 diabetes	3 established variants	0.55
Type 2 diabetes	3 (out of 19)	0.56
Type 2 diabetes	18 established variants	0.60
Type 2 diabetes	18 established variants	0.60

Janssens & van Duijn (2008) HMG



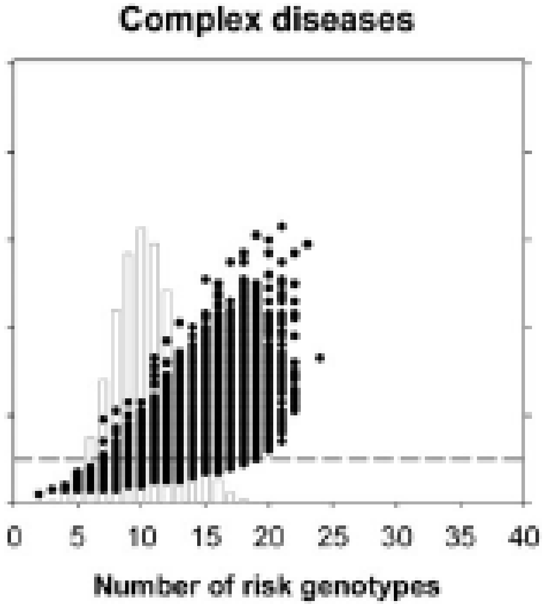
Disease	Clinical risk factors	Variant selection[a]	Genetic variants	AUC before	AUC after	Reference
Cardiovascular disease	Age, sex, family history of myocardial infarction, low density lipoprotein, high density lipoprotein cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure, diabetes mellitus, body mass index, smoking, C-reactive protein, lipid-lowering therapy, antihypertensive treatment	9 (out of 11) established SNPs in 9 genes	*APOB* rs693, *APOE* cluster rs4420638, *HMGCR* rs12654264, *LDLR* rs1529729, *PCSK9* rs11591147, *ABCA1* rs3890182, *CETP* rs1800775, *LIPC* rs1800588, and *LPL* rs328	0.80	0.80	(26)
Coronary heart disease	Age, triglycerides, cholesterol, systolic blood pressure, smoking	4 (out of 12)	*UCP2* G(−866)A, *APOE* e2/3/4, *LPL* D9N, *APOA4* T347S	0.66	0.70	(9)
Coronary heart disease: in whites	Age, systolic blood pressure, total cholesterol, high density lipoprotein cholesterol, diabetes, use of antihypertensive medication, smoking	11 (out of 116)	*VAMP8, PALLD, KIF6, MKI67, MYH15, Loc646377, HPS1, SNX19, ADAMTS1 (2x), ADRB3*	0.76	0.77	(25)
Coronary heart disease: in blacks	Age, systolic blood pressure, total cholesterol, high density lipoprotein cholesterol, diabetes, use of antihypertensive medication, smoking	11 (out of 116)	*DMXL2, ZNF132, KIF6, F2, OR2A25, KRT5, CTNNA3, HAP1, GIPR, FSTL4, THBS2*	0.76	0.77	(25)
MI after surgery	AXT time, number of coronary grafts, previous cardiac surgery	3 (out of 48)	*IL6* G572C, *ICAM1* K469E, *SELE* G98T	0.70	0.76	(12)
Prostate cancer	Age, geographic region, family history	5 (out of 16) in 5 established regions)	rs4430796 (in 17q12), rs1859962 (in 17q24.3), rs16901979, rs6983267 and rs1447295 (all in 8q24)	0.61	0.63	(27)
Type 2 diabetes	Body mass index, plasma glucose level	3 (out of 6)	*PPARG* P12A, *CAPN10* SNP43 and SNP44	0.68[b]	0.68[b]	(24)
Type 2 diabetes	Age, sex, body mass index	3 (out of 19)	*GCK* G(−30G)A, *IL6* G(−174)C, *TCF7L2* rs7903146	0.82	0.82	(15)
Type 2 diabetes	Age, sex, body mass index	18 established variants	SNPs in *TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/ NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX-IDE, THADA, JAZF1*	0.78	0.80	(16)
Type 2 diabetes	Age, sex, body mass index	18 established variants	SNPs in *TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/ NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX-IDE, THADA, JAZF1*	0.66	0.68	(17)

Janssens & van Duijn (2008) HMG

- Genetic variants appear to add little to traditional risk factors

- Some genetic variants might influence intermediate risk factors

Problems

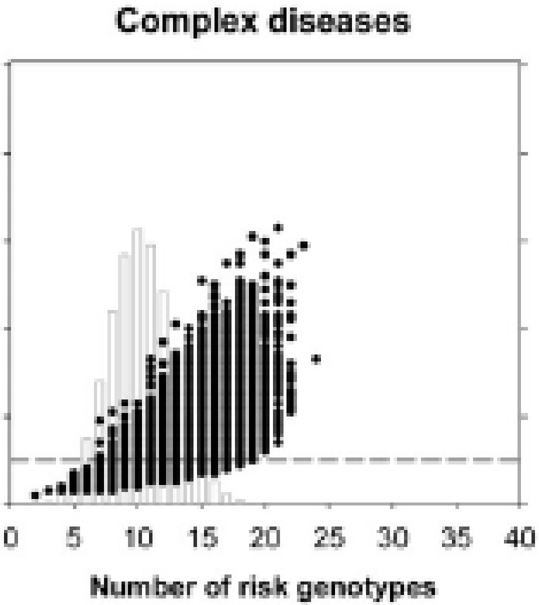


Janssens & van Duijn (2008) HMG

- Most individuals have disease risks only slightly higher or lower than the population average

- Substantial variation in disease risk may be seen between individuals with the same number of risk genotypes resulting from differences in effect sizes between risk genotypes

Problems



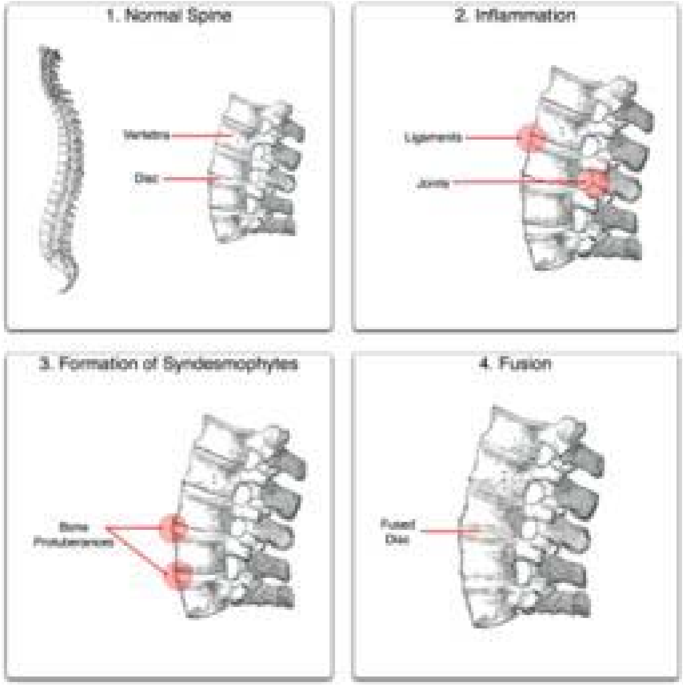
Janssens & van Duijn (2008) HMG

- Knowledge of increased risk may not be useful

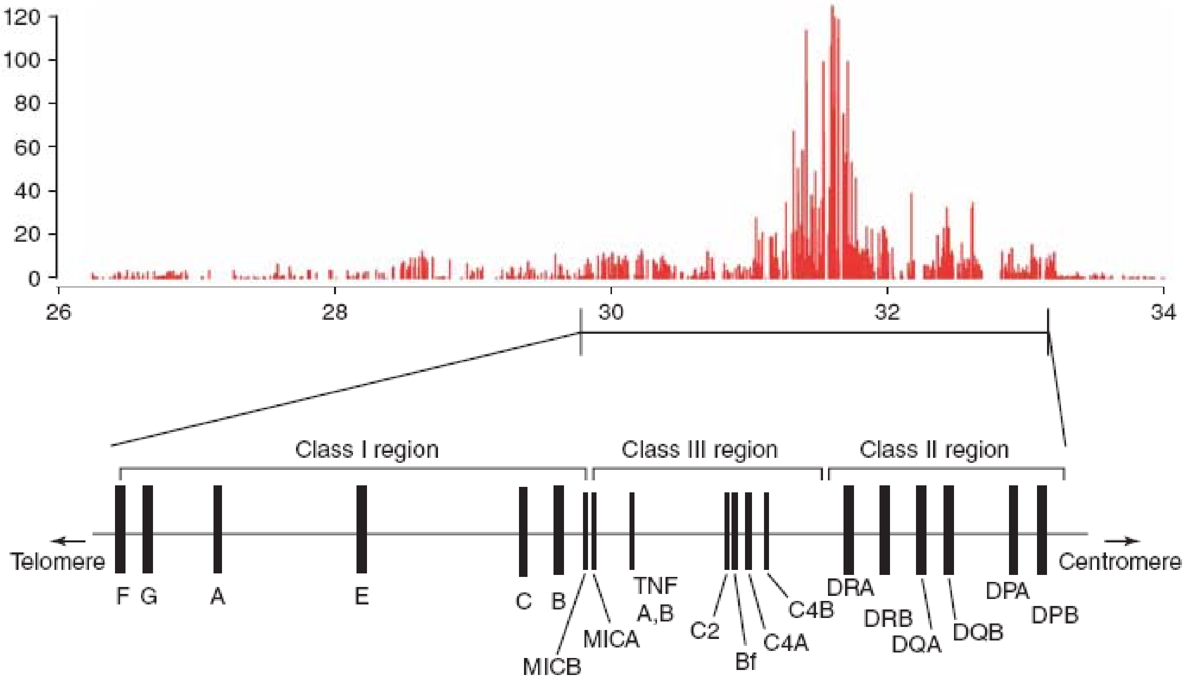
- Predictive value of genetic tests are limited by their heritability

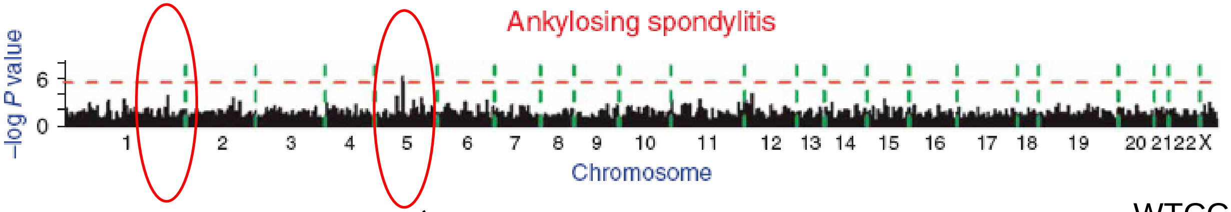
- Can we do better than just asking a first degree relative?

Ankylosing Spondylitis



- Auto-immune arthritis resulting in fusion of vertebrae
- Prevalence of 0.4% in Caucasians. More common in men.
- Often associated with psoriasis, IBD and uveitis
- Ed Sullivan, Mike Atherton



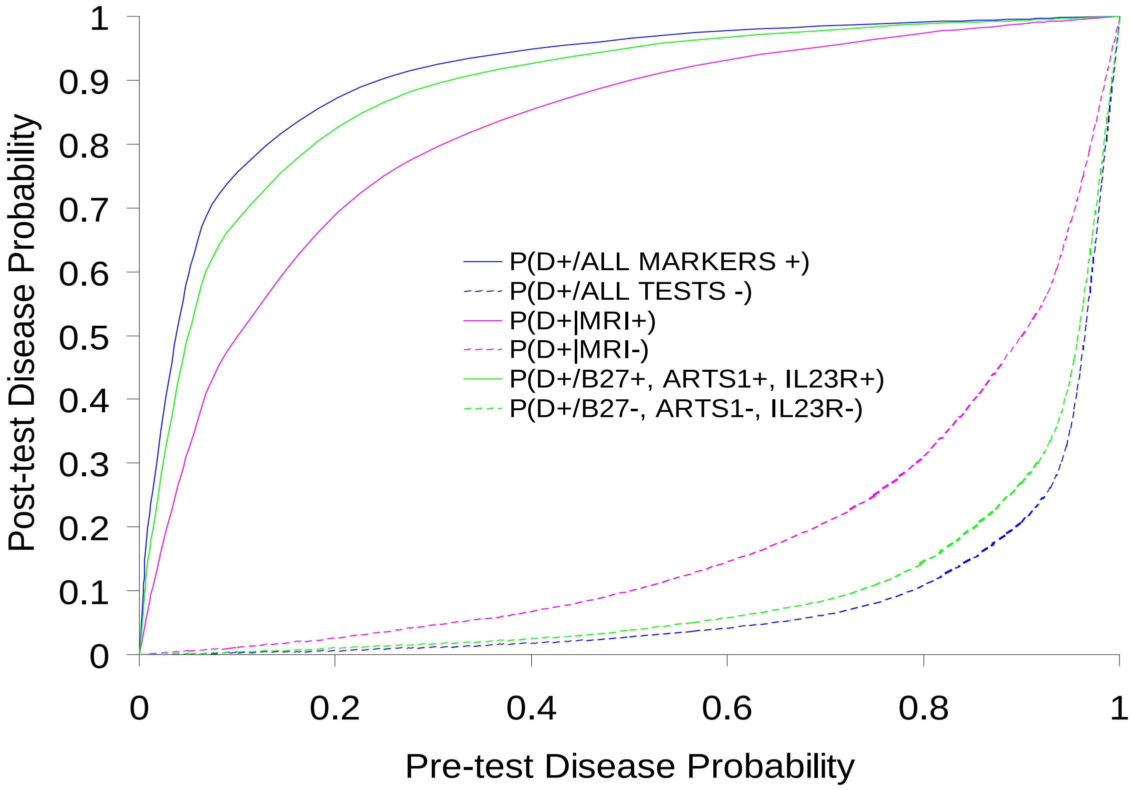


WTCCC (2007) *Nat Genet*

IL23-R

ARTS-1

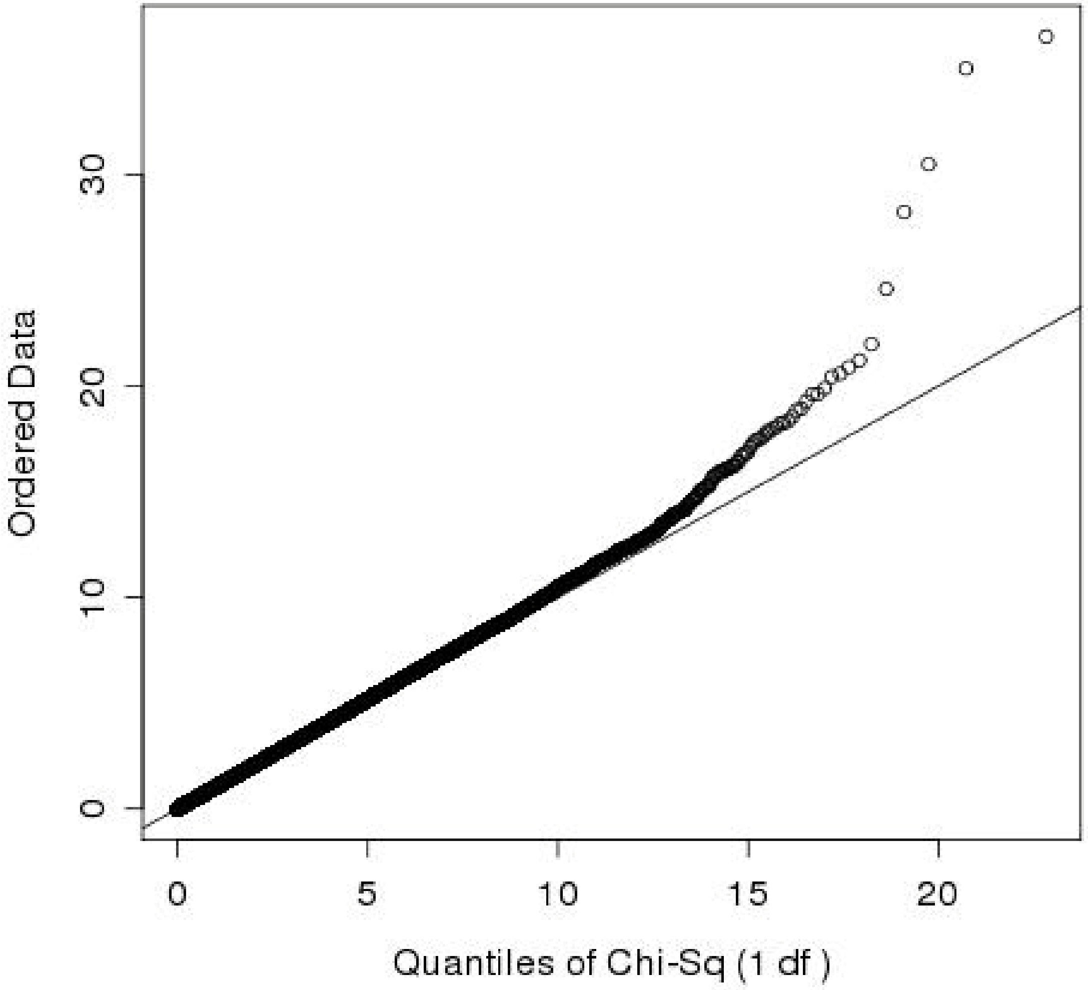
Ankylosing Spondylitis



(Brown & Evans, in prep)

- Prevalence of B27+, ARTS1+,IL23R+ is 2.4%
- Prevalence of B27-, ARTS1-, IL23R- is 19%

Genome-wide Prediction?



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

THE INDEPENDENT

Ireland, €1 70p

Thursday 7 June 2007

www.independent.co.uk

Tracey Emin

Exclusive: How I created the show of my life

PLUS YOUR CHANCE TO OWN A LIMITED-EDITION ARTWORK IN EXTRA

Bipolar disorder

Also known as manic depression, it affects 100 million people around the world

Coronary heart disease

The most frequent cause of death in Britain, with 100,000 victims every year. By 2020, it will be the biggest killer in the world

Hypertension

High blood pressure affects 16 million people in Britain. Can lead to stroke, heart disease and kidney failure

Rheumatoid arthritis

Nearly 400,000 people in Britain are afflicted with this auto-immune disease of the joints

Type 1 diabetes

Diabetic condition in which sufferers have to inject insulin. Affects 350,000 people in UK

Crohn's disease

Up to 60,000 people are affected by this debilitating bowel condition which can cause distress and pain for a lifetime

Type 2 diabetes

Almost 2 million Britons are affected by this late-onset disease, which is linked with the growing obesity epidemic

THE GENETIC REVOLUTION

DISCOVERY OF GENES RESPONSIBLE FOR SEVEN OF THE MOST COMMON ILLNESSES OFFERS HOPE TO MILLIONS OF SUFFERERS

FULL STORY, PAGE 2

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)

Methods

- “Training set”

  90% of cases and controls

  Run test of association in training set

  Select a set of nominally associated SNPs according to a threshold (α = 0.8, 0.5, 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001)

- “Prediction set”

  Apply prediction method to prediction set (10% of cases and controls)

- Cross validation

  Do ten times, record mean AUC and range of AUCs

Methods

- “Log Odds 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

- “Count Method”

  For each individual:

  Score = sum(x_i)

Control Condition

- Differences between cases and controls might reflect undetected batch effects, population stratification and/or genotyping error

- These will inflate the apparent predictive ability of SNPs

- Predict a disease using SNPs derived from training sets of other diseases

- Would expect AUC ~ 0.5 in the absence of these factors

Bipolar Disorder

Threshold	Odds Method		Log Odds Method	
	Profiling	Control	Profiling	Control
p < .8	.653	.537	.668	.529
p < .5	.664	.527	.668	.531
p < .1	.646	.537	.636	.547
p < .05	.625	.537	.620	.537
p < .01	.570	.555	.567	.548
p < .001	.539	.534	.533	.527
p < .0001	.533	.518	.528	.520
p < .00001	.521	.525	.529	.521

Type I Diabetes

Threshold	Odds Method		Log Odds Method	
	Profiling	Control	Profiling	Control
p < .8	.620	.513	.721	.531
p < .5	.624	.515	.724	.518
p < .1	.637	.515	.743	.515
p < .05	.673	.537	.747	.526
p < .01	.697	.531	.749	.525
p < .001	.712	.544	.749	.545
p < .0001	.716	.540	.748	.534
p < .00001	.717	.540	.749	.533

Conclusions

- A genome-wide score provides significant (but not very good) discrimination between cases and controls

- Does this genome-wide score provide discriminative ability over and above that afforded by known variants?

Bipolar Disorder

Threshold	Odds Method		Log Odds Method	
	Profiling	Control	Known	All
Known	.549			
p < .8	.657	.564	.678	.572
p < .5	.671	.566	.674	.566
p < .1	.651	.561	.641	.562
p < .05	.656	.556	.641	.562
p < .01	.608	.584	.597	.579
p < .001	.563	.561	.560	.563
p < .0001	.574	.561	.569	.561
p < .00001	.561	.562	.560	.562

Type I Diabetes

Threshold	Odds Method		Log Odds Method	
	Known	All	Known	All
Known	.784			
p < .8	.793	.782	.792	.786
p < .5	.794	.785	.793	.786
p < .1	.787	.785	.788	.785
p < .05	.787	.785	.788	.786
p < .01	.788	.785	.788	.785
p < .001	.786	.785	.785	.784
p < .0001	.785	.787	.784	.790
p < .00001	.785	.786	.787	.785

Limitations

- Only additive relationships modelled

- Genotyping error, batch effects and/or population stratification in the cases group

Conclusions

- Currently genetic information of little diagnostic utility for (most) complex diseases

- A simple genome-wide score has discriminative ability and can add information over and above that afforded by known variants