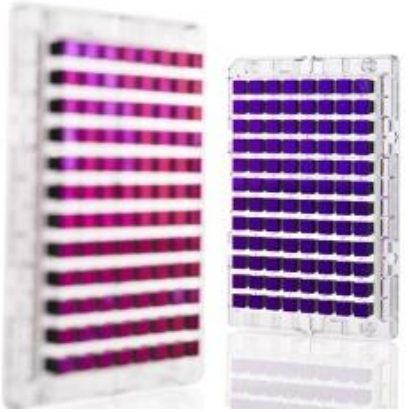
Genome-wide association studies (GWAS) – *still* continued!

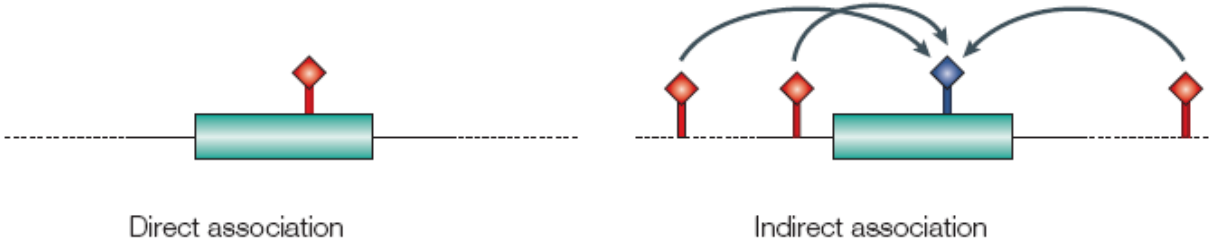


Outline for GWAS

- Limitations, heritability, and “missing heritability”
- Gene/pathways/polygenic models
- Pleiotropy
- Genetic risk scores
- Transcriptome-wide association studies (TWAS)

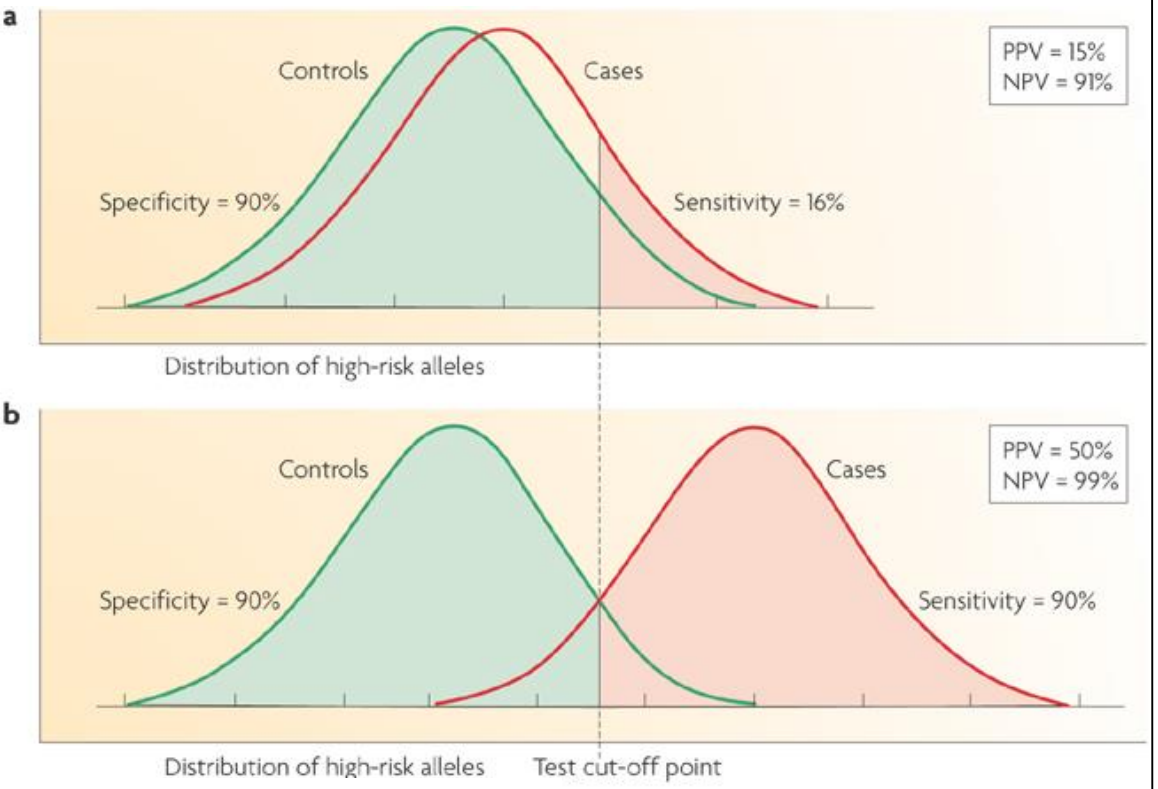
Limitations of GWAS

- Not very predictive
- Explain little heritability, generally, but improving
- Focus on common variation, but getting rarer
- Many associated variants are not causal



Limitations of GWAS

- Not very predictive



Witte, Nat Rev Genet 2009

Example:
AUC for Breast Cancer Risk

58%: Gail model (# first degree relatives w bc, age menarche, age first live birth, number of previous biopsies) + age, study, entry year
58.9%: SNPs
61.8%: Combined

Wacholder et al., NEJM 2010

What does it mean to explain little of the heritability, and where is it?

Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits[a]	h^2 All GWAS SNPs[b]
Type 1 diabetes	0.9[98]	0.6[99,c]	0.3[12]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]	
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]	0.2[14]
Crohn's disease	0.6–0.8[103]	0.1[11]	0.4[12]
Ulcerative colitis	0.5[103]	0.05[12]	
Multiple sclerosis	0.3–0.8[104]	0.1[45]	
Ankylosing spondylitis	>0.90[105]	0.2[106]	
Rheumatoid arthritis	0.6[107]		
Schizophrenia	0.7–0.8[108]	0.01[79]	0.3[109]
Bipolar disorder	0.6–0.7[108]	0.02[79]	0.4[12]
Breast cancer	0.3[110]	0.08[111]	
Von Willebrand factor	0.66–0.75[112,113]	0.13[114]	0.25[14]
Height	0.8[115,116]	0.1[13]	0.5[13,14]
Bone mineral density	0.6-0.8[117]	0.05[118]	
QT interval	0.37–0.60[119,120]	0.07[121]	0.2[14]
HDL cholesterol	0.5[122]	0.1[57]	
Platelet count	0.8[123]	0.05–0.1[58]	

[a] Proportion of phenotypic variance or variance in liability explained by genome-wide-significant and validated SNPs. For a number of diseases, other parameters were reported, and these were converted and approximated to the scale of total variation explained. Blank cells indicate that these parameters have not been reported in the literature.

[b] Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.

[c] Includes pre-GWAS loci with large effects.

Visscher, AJHG 2011

Class exercise: Where is it?

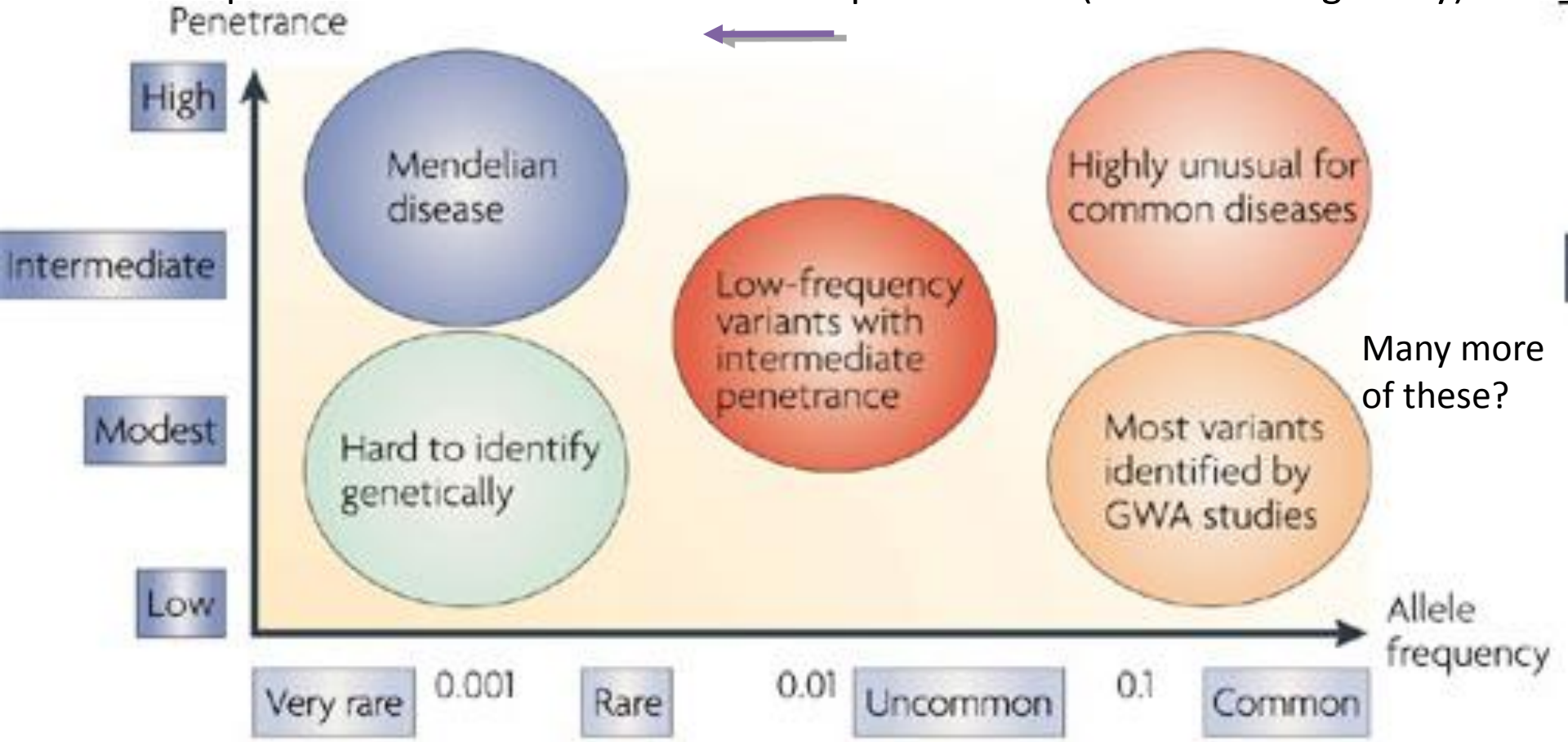
- Can GWAS find more?
- Are our analysis methods not creative enough?
- Are we missing genetic things?

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.

Where's the heritability?

Common disease rare variant (CDRV) hypothesis: diseases due to multiple rare variants with intermediate penetrances (allelic heterogeneity)



See: NEJM, April 30, 2009

Nature Reviews | Genetics
McCarthy et al., 2008

Will GWAS results explain more heritability?

Possibly, if…

1. Causal SNPs not yet detected due to power issues – weak effects
   - Previous GWAS designed/powered for MAF>0.05/0.10; lower MAFs for newer arrays and larger reference panels
2. More complicated modeling necessary, e.g. gene-gene interaction (epistasis), gene-environment interaction, etc.

How much is left to discover? "Array" heritability

- Compute heritability using **all** SNPs
  - Be careful with QC, esp. case-control
- If you adjust for the known hits, how much remains yet to be found?
- Restricted to narrow sense heritability
- E.g.,

Lecture 2:

"Broad sense": $H^2=\sigma^2_G/\sigma^2_P=(\sigma^2_A + \sigma^2_D)/\sigma^2_P$

"Narrow sense": $h^2 = \sigma^2_A/\sigma^2_P$

http://www.complextraitgenomics.com/software/gcta/

nature genetics

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

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

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

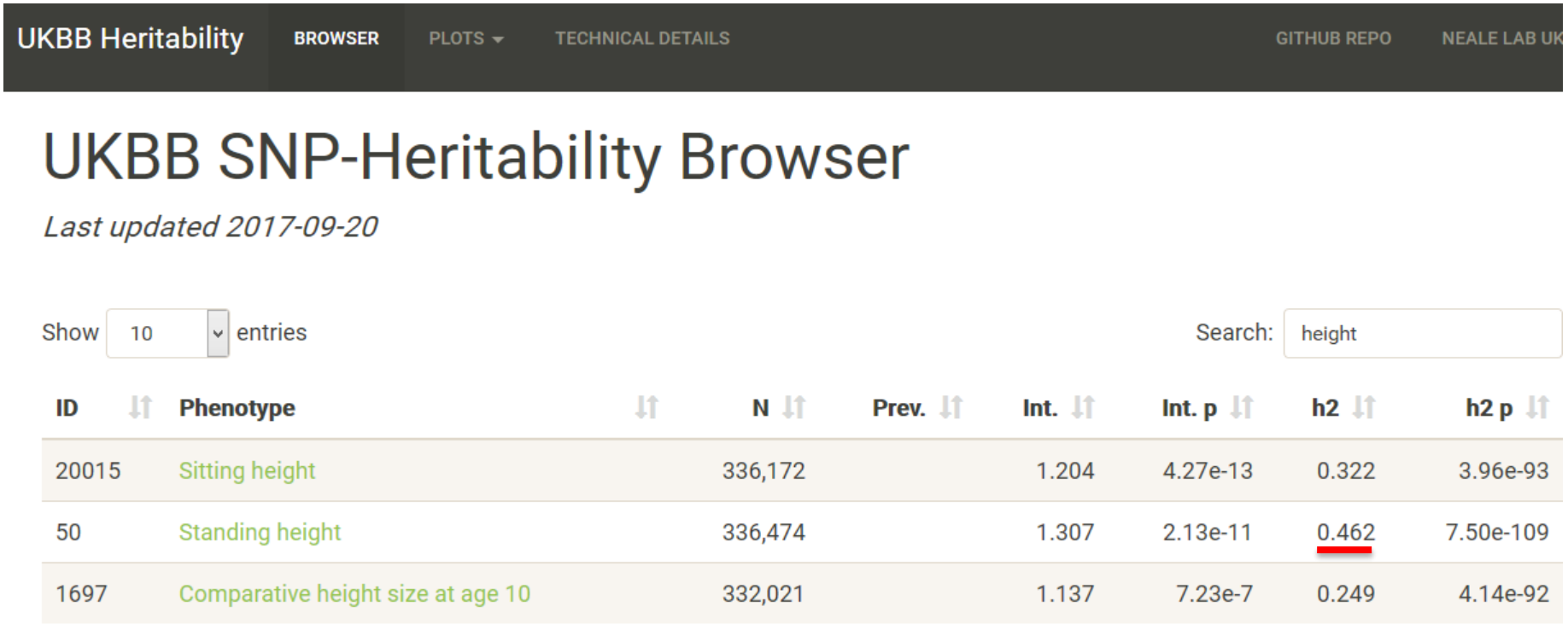
GWASs in human populations have discovered hundreds of SNPs significantly associated with complex traits[1,2], yet for any one of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for instance, occur if causal variants have lower minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of each to the heritability of height in humans as a model complex trait.

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

Data from a GWAS that are collected to detect statistical associations between SNPs and complex traits are usually analyzed by testing each SNP individually for an association with the trait. To account for the

UK Biobank h^2 for height

- https://nealelab.github.io/UKBB_ldsc/



ID	Phenotype	N	Prev.	Int.	Int. p	h2	h2 p
20015	Sitting height	336,172		1.204	4.27e-13	0.322	3.96e-93
50	Standing height	336,474		1.307	2.13e-11	0.462	7.50e-109
1697	Comparative height size at age 10	332,021		1.137	7.23e-7	0.249	4.14e-92

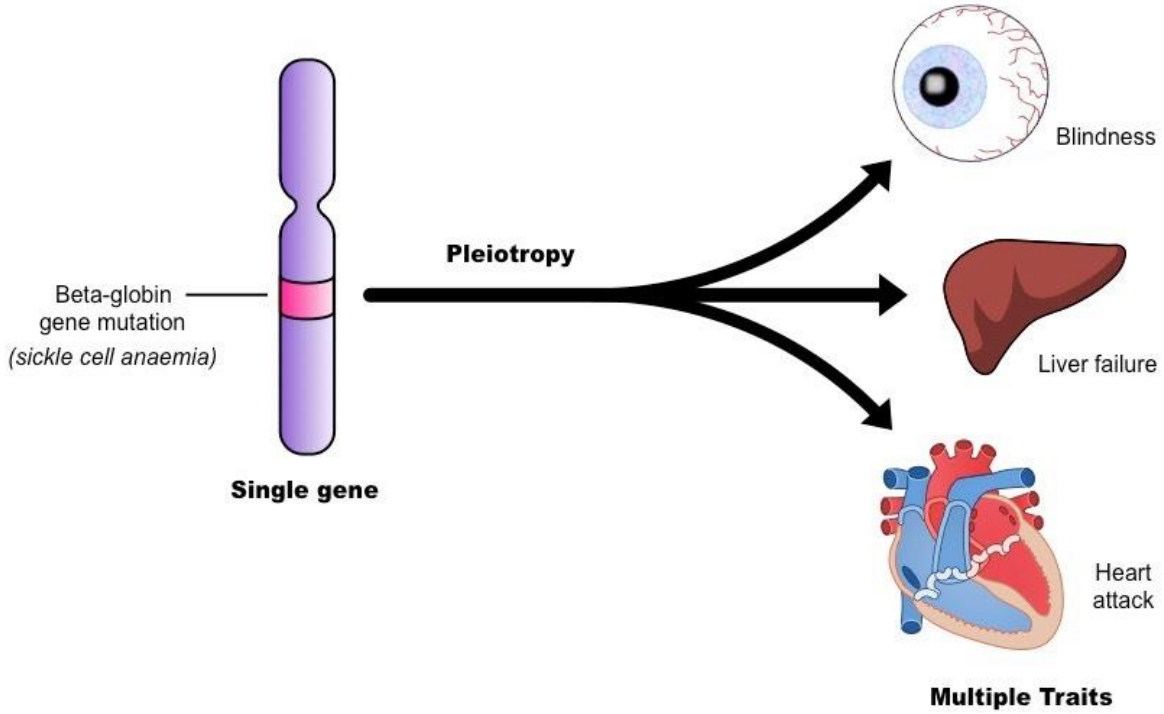
Recall our pleiotropy example: HOXB13 G84E multi-cancer association

Table 1. Pleiotropic effect of G84E mutation on risk of cancer.

Cancer[a]	# Case carriers[b]	# Cases[b]	Frequency[c]	Odds Ratio[d] (95% CI)	Corrected Odds Ratio[e] (95% CI)	P-value[d]
Kidney	5	303	0.94%	2.32 (0.94, 5.76)	3.32 (0.89, 9.99)	0.034
Bladder	5	335	0.85%	2.05 (0.81, 5.22)	2.84 (0.70, 8.94)	0.065
Non-Hodgkin's Lymphoma	10	846	0.67%	1.81 (0.98, 3.36)	2.42 (0.96, 5.52)	0.029
Melanoma	15	1301	0.66%	1.50 (0.89, 2.55)	1.88 (0.81, 3.95)	0.066
Pancreas	2	149	0.77%	1.49 (0.28, 7.83)	1.86 (0.18, 13.70)	0.32
Breast	38	3183	0.68%	1.48 (1.04, 2.10)	1.84 (1.07, 3.16)	0.014
Endometrium	6	578	0.59%	1.44 (0.64, 3.24)	1.77 (0.49, 5.22)	0.19
Colon	11	1119	0.56%	1.42 (0.77, 2.60)	1.74 (0.65, 4.06)	0.13
Thyroid	3	217	0.79%	1.35 (0.35, 5.22)	1.61 (0.23, 8.81)	0.33
Ovary	2	198	0.58%	1.32 (0.32, 5.49)	1.56 (0.20, 9.26)	0.35
Multiple Myeloma	1	135	0.42%	1.26 (0.20, 7.92)	1.46 (0.12, 13.75)	0.40
Lung	5	667	0.43%	1.10 (0.44, 2.72)	1.18 (0.30, 4.22)	0.42
Oral	2	283	0.40%	0.98 (0.23, 4.10)	0.97 (0.14, 6.78)	NA
Lymphocytic Leukemia	1	218	0.26%	0.53 (0.05, 5.24)	0.39 (0.03, 9.06)	NA
Any[f]	123	10493	0.67%	1.36 (1.13, 1.63)	1.63 (1.22, 2.29)	5.8×10^{-4}
Any single cancer[g]	106	9532	0.63%	1.41 (1.14, 1.75)	1.72 (1.23, 2.51)	8.9×10^{-4}
Two or more cancers[h]	17	961	1.01%	2.21 (1.35, 3.61)	3.12 (1.60, 6.14)	0.0011

Pleiotropy

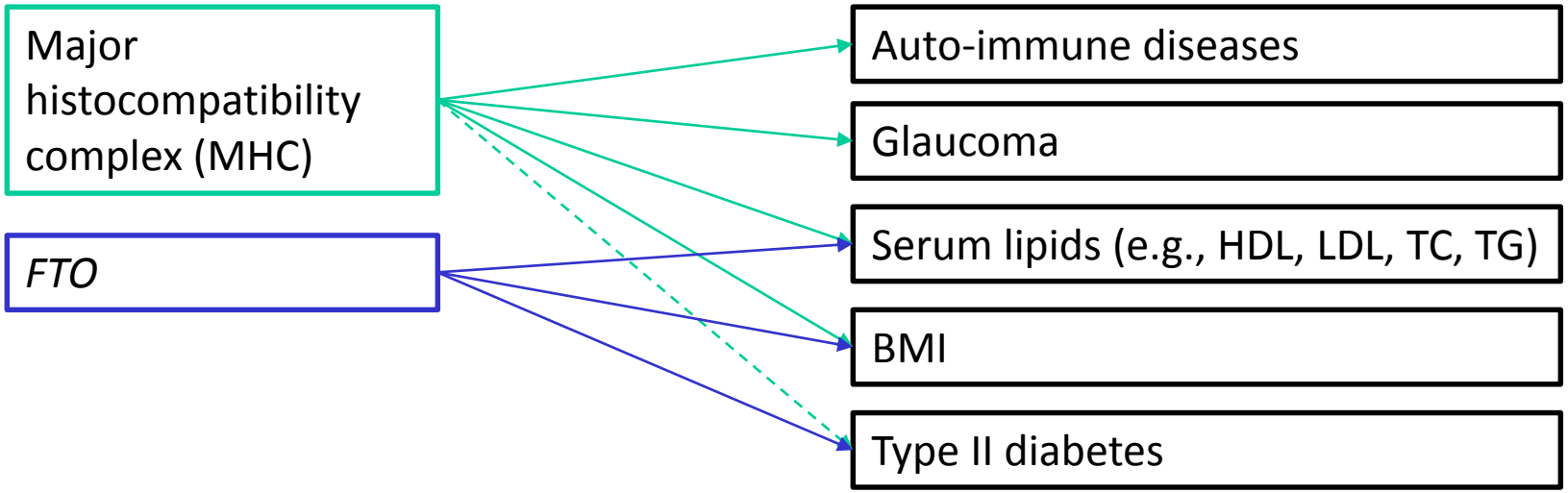
- Pleiotropy: one gene, multiple traits



http://ib.bioninja.com.au/_Media/pleiotropy_med.jpeg

Pleiotropy

- But it's more like each gene affects lots of traits



Shared genetic basis?

- How much of a shared genetic basis do traits have?

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 19 2012, pages 2540–2542
doi:10.1093/bioinformatics/bts474

Genetics and population analysis

Advance Access publication July 26, 2012

Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood

S.H. Lee[1,*], J. Yang[2], M.E. Goddard[3], P.M. Visscher[1,2] and N.R. Wray[1]

[1]The University of Queensland, Queensland Brain Institute, Brisbane, QLD 4072, [2]The University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, QLD 4102 and [3]Department of Agriculture and Food Systems, University of Melbourne, VIC 3010, Melbourne, Australia

Shared cancer example

Table 1. GWAS estimates of cancer heritability on the liability scale*

Cancer	h_l^2 (95% CI)	h_r^2 (95% CI)
Bladder	0.123 (0.086 to 0.160)	0.112 (0.075 to 0.148)
Breast (ER-)	0.096 (0 to 0.199)	0.079 (0 to 0.181)
Endometrium	0.178 (0.085 to 0.270)	0.177 (0.085 to 0.270)
Esophagus†	0.381 (0.174 to 0.588)	0.370 (0.164 to 0.577)
Glioma	0.046 (0 to 0.116)	0.036 (0 to 0.106)
Kidney	0.147 (0.023 to 0.270)	0.136 (0.012 to 0.259)
Lung		
Asian†,‡	0.121 (0.064 to 0.177)	0.102 (0.046 to 0.159)
European	0.206 (0.142 to 0.271)	0.189 (0.125 to 0.253)
Lymphoma		
CLL	0.220 (0.162 to 0.278)	0.164 (0.105 to 0.222)
DLBCL	0.092 (0.038 to 0.145)	0.088 (0.035 to 0.142)
Osteosarcoma	0.159 (0.079 to 0.239)	0.158 (0.078 to 0.238)
Pancreas	0.098 (0.037 to 0.160)	0.084 (0.023 to 0.145)
Prostate		
Overall	0.378 (0.244 to 0.513)	0.285 (0.151 to 0.419)
Nonadvanced stage	0.351 (0.211 to 0.491)	0.259 (0.120 to 0.399)
Advanced stage	0.232 (0.157 to 0.307)	0.193 (0.119 to 0.268)
Stomach† (noncardia)	0.253 (0 to 0.522)	0.243 (0 to 0.512)
Testes	0.299 (0.084 to 0.513)	0.199 (0 to 0.415)

* h_l^2 (95% confidence interval [CI]) is the estimated heritability on the liability scale (95% CI) using all qualifying single-nucleotide polymorphisms (SNPs), while h_r^2 is the heritability after removing SNPs within 250 kb of a previous genome-wide association study hit. CI = confidence interval; CLL = chronic lymphocytic leukemia; DLBCL = diffuse large B-cell lymphoma; ER = estrogen receptor; GWAS = genome-wide association study.
† Asian population.
‡ Nonsmoking women.

JNCI J Natl Cancer Inst (2015) 107(12): djv279
doi:10.1093/jnci/djv279

ARTICLE

Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for 13 Cancer Types

Shared cancer example

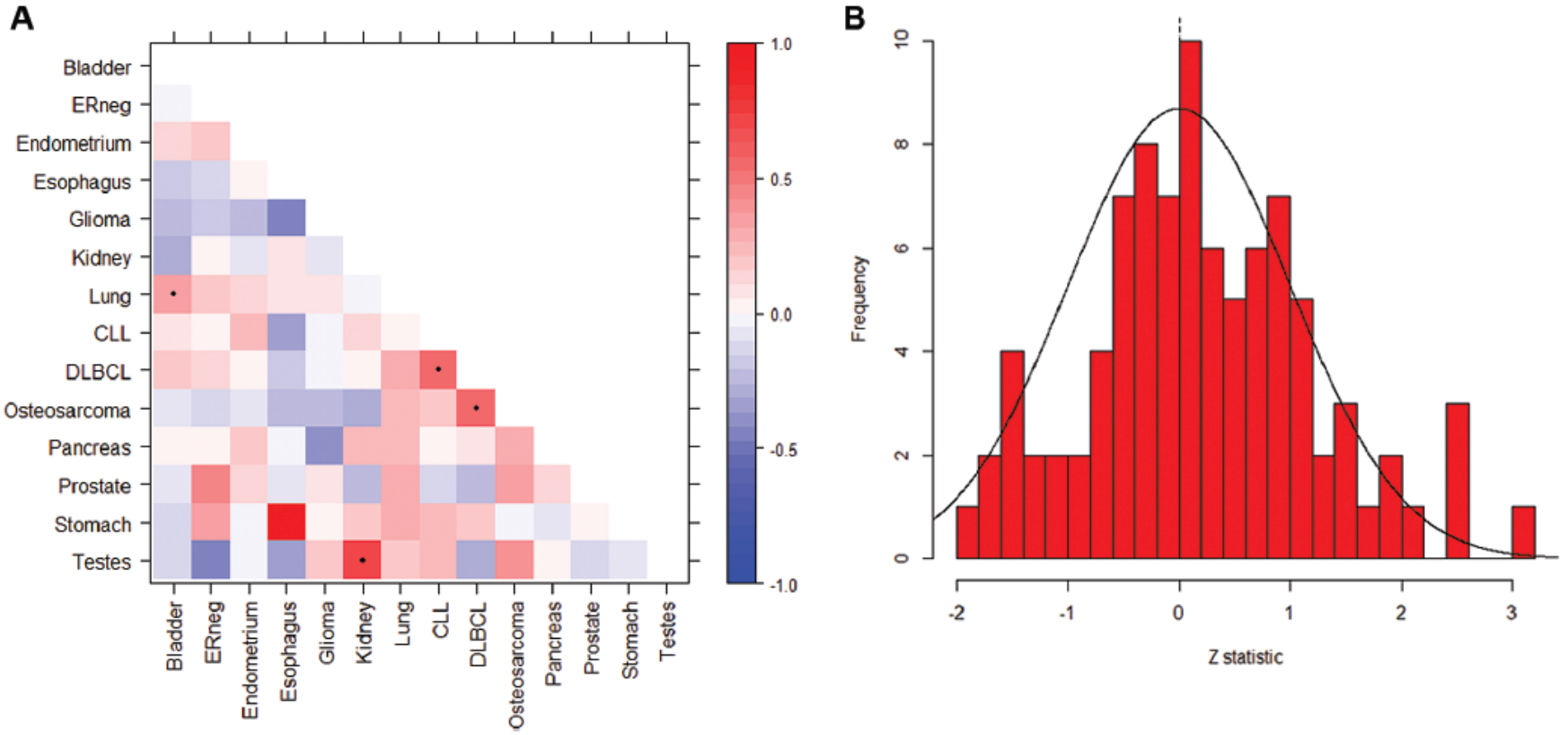


Figure 1. Genetic correlation of cancer pairs. **A)** The genetic correlation between cancer sites. **Dots** indicate $P < .01$. **B)** Distribution of the corresponding Z-statistics from testing the null hypotheses of no genetic correlations. **Black curve** illustrates the expected distribution under the null hypothesis of no genetic correlation. CLL = chronic lymphocytic leukemia; DLBCL = diffuse large B-cell lymphoma.

JNCI J Natl Cancer Inst (2015) 107(12): djv279
doi:10.1093/jnci/djv279

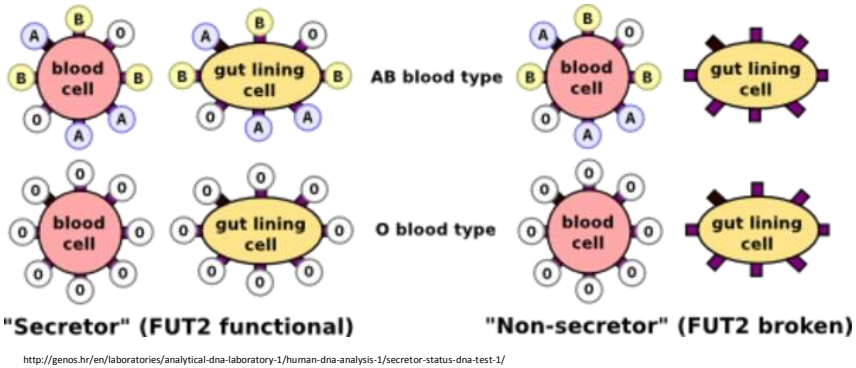
ARTICLE

Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for 13 Cancer Types

But, what about more complicated modeling? Epistasis

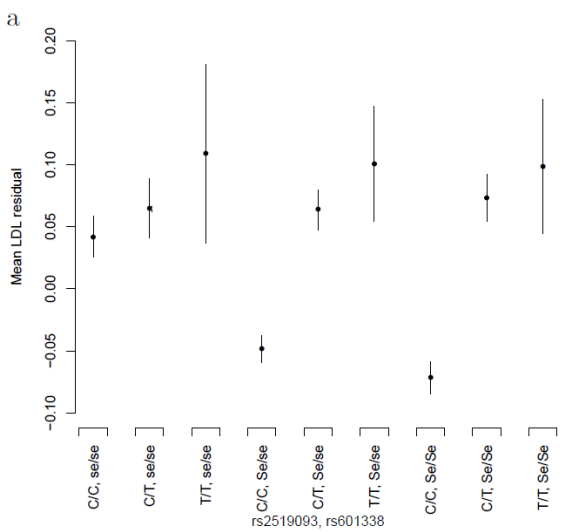
- Epistasis – snp x snp, or gene x gene interactions
- Very difficult to do genome-wide
  - Much less power to detect an interaction than a main effect in general
  - Multiple testing burden: Correct for 10^6 tests before? Now 10^{12} tests!!!
- Some argue not worth testing for an interaction unless a main effect…

Epistasis in serum lipids

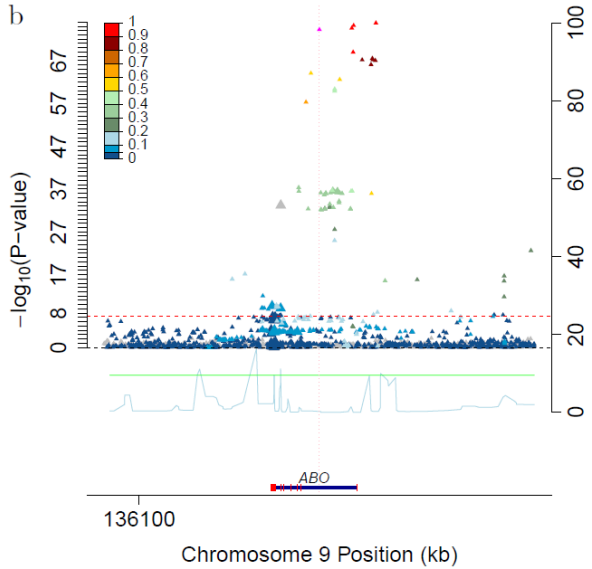


- Serum lipids example
  - Tested all pairs of known GWAS hit variants
  - Rs2519093 (*ABO*) x rs492602 (*FUT2*): Interaction $P_{LDL}=8.1x10^{-10}$
    - Rs492602 has $r^2=0.992$ with rs601338, a nonsense dominant variant determining *FUT2* non-secretor status

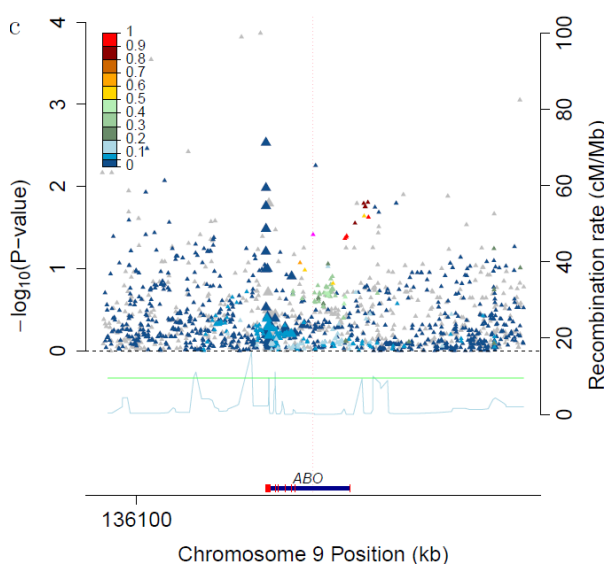
Residual LDL by genotype



ABO locus for secretors (Se/Se or Se/se)



ABO locus plot for non-secretors



Other things than SNPs

- Copy number variants (CNVs)
- Epigenetics, e.g., methylation, histone modification
- Gene expression (RNA levels) [eQTLs! we covered these]
- Proteomics (measures of proteins in cells)
- ...

Ways to enrich weak signals?: Gene/pathway-based tests

- Various ways of collapsing the genotype information in multiple genes
  - Less multiple comparison adjustment
- logit (Prob(y=1| **x, c**)) = $\alpha + \beta\mathbf{x} + \gamma\mathbf{c}$

  e.g.=$\alpha+\beta_1 x_1+\ldots+\beta_{12} x_{12}$+{PC's, other covariates}
  - y: disease status)
  - **x** is a vector of genotypes (e.g., a gene, or a pathway)
  - **c** is a vector of covariates
- H_0: $\beta=\mathbf{0}$
- Rare variant "burden" tests are along these lines
  - Make a lot of assumptions

Pathways - how to define?

- Many websites / companies provide ‘dynamic’ graphic models of molecular and biochemical pathways.
- Examples: GO (Gene Ontology), KEGG, BioCarta, Reactome
- May be interested in potential joint and/or interaction effects of multiple genes in one pathway.

Other Extreme: Polygenic Models

- Many weak associations combine to risk?
- Score model (use all GWAS SNPs):

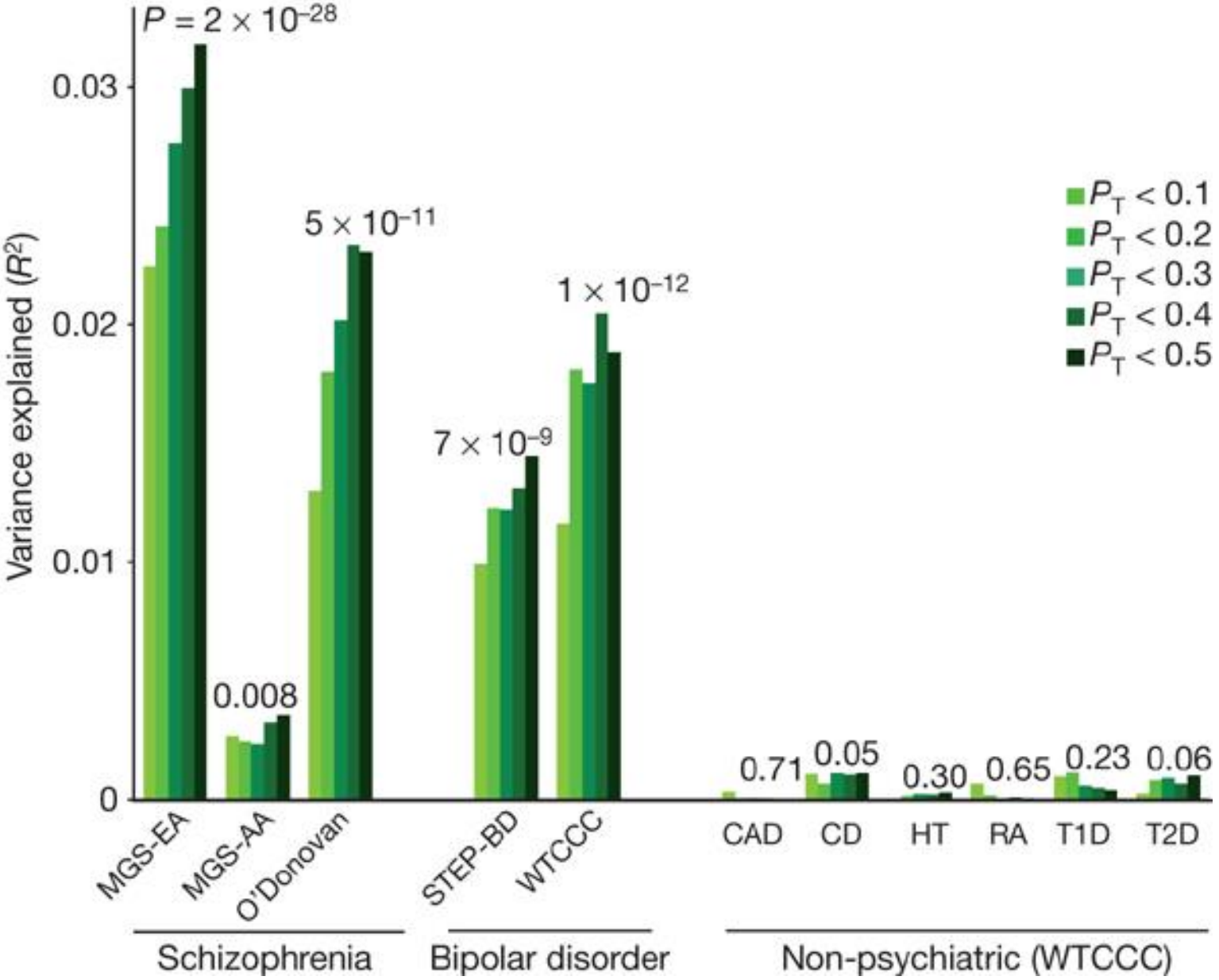
$$x_j = \frac{\sum_{i=1}^{m} \ln\left(OR_i\right) \times SNP_{ij}}{m}$$

where

  - $\ln(OR_i)$ = 'score' for SNP_i from 'discovery' sample
  - SNP_{ij} = # of alleles (0,1,2) for SNP_i, person j in 'validation' sample.
  - Large number of SNPs (m)
- x_j associated with disease?
- ASIDE: Genetic risk score **same** idea, but only on ***known hits***

ISC / Purcell et al. *Nature* 2009

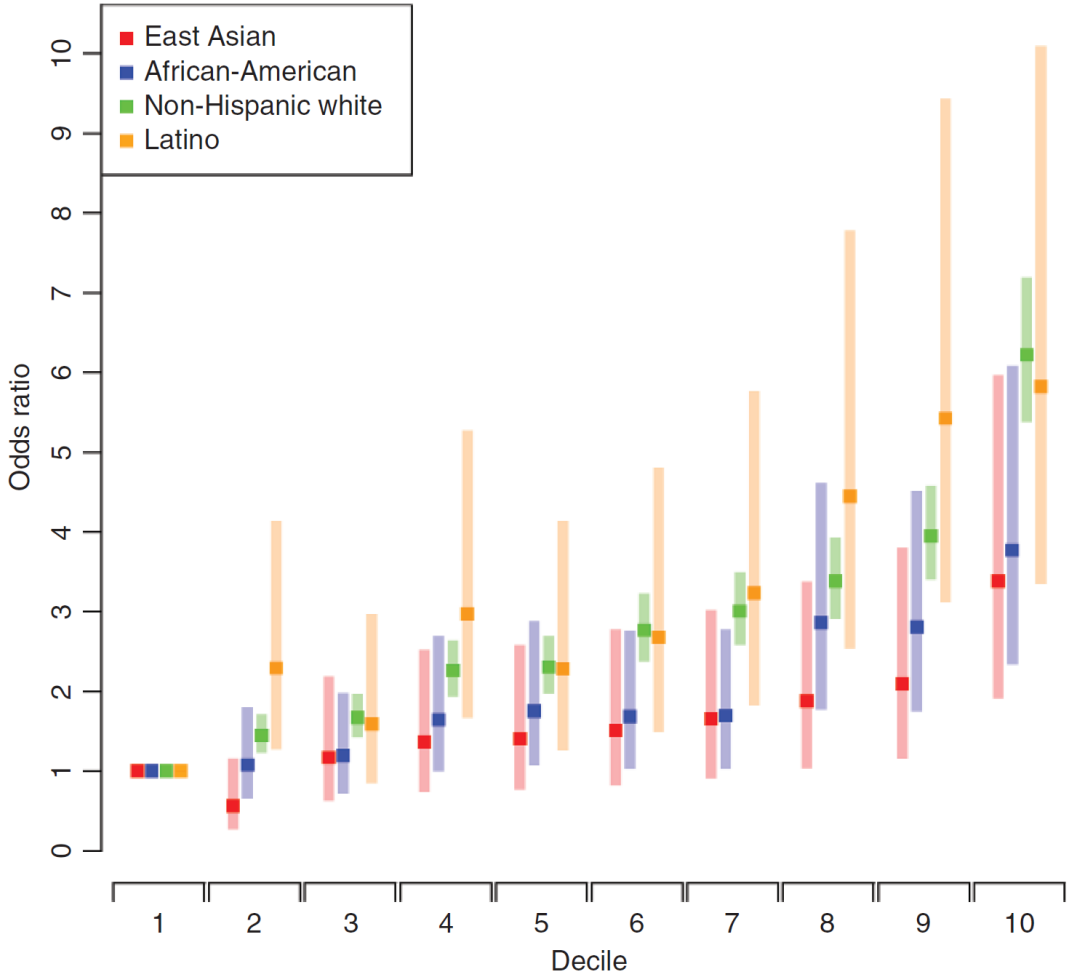
Application of Model



Purcell / ISC et al. *Nature* 2009

Genetic Risk Score (GRS) - Risk variant tails -> More predictive?

Figure 4. Impact of increasing deciles of risk profile scores based on 105 known risk SNPs across different race/ethnicity groups. Risk scores were generated by combining the 105 SNPs into a single score, applying the logs odds ratio estimated by previous studies to our KP study population genotype data. Odds ratios for effect of risk profile scores on prostate cancers calculated within each decile of the scores, using the lowest decile of risk profile scores as the referent category. The non-Hispanic white and Latino groups had substantially higher ORs than the African-American group, and the East Asian group always had the lowest ORs.



i.e., β_j

$$GRS_j = \frac{\sum_{i=1}^{m} \ln(OR_i) \times SNP_{ij}}{m}$$

Be careful of SNPs in LD… e.g., don't want to include same SNP effectively twice if reported in two studies.

Hoffmann et. Al, Cancer Discovery 2015

Transcriptome-Wide Association Studies (TWAS)

- We have used reference panel sequence data to fill in missing genotypes.
- We have talked about eQTLs, testing for correlation between SNPs and the genotype expression
  - But what about predicting/imputing what the expression data was based on the genotypes?
    - And then test that for correlation with the trait instead?
    - Guided gene-based test, if you will

Class Exercise!: Express yourself (or at least someone else)!

Reference population

Person	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	Expression
Ref-001	AA	AA	TT	TT	TT	0.98
Ref-002	AA	AA	TC	TG	TG	0.90
Ref-003	GG	GG	TC	TG	TG	0.10
Ref-004	GG	GG	CC	TT	TG	0.12
Ref-005	GG	GG	CC	TT	GG	0.05

How is this similar/different to genotype imputation?

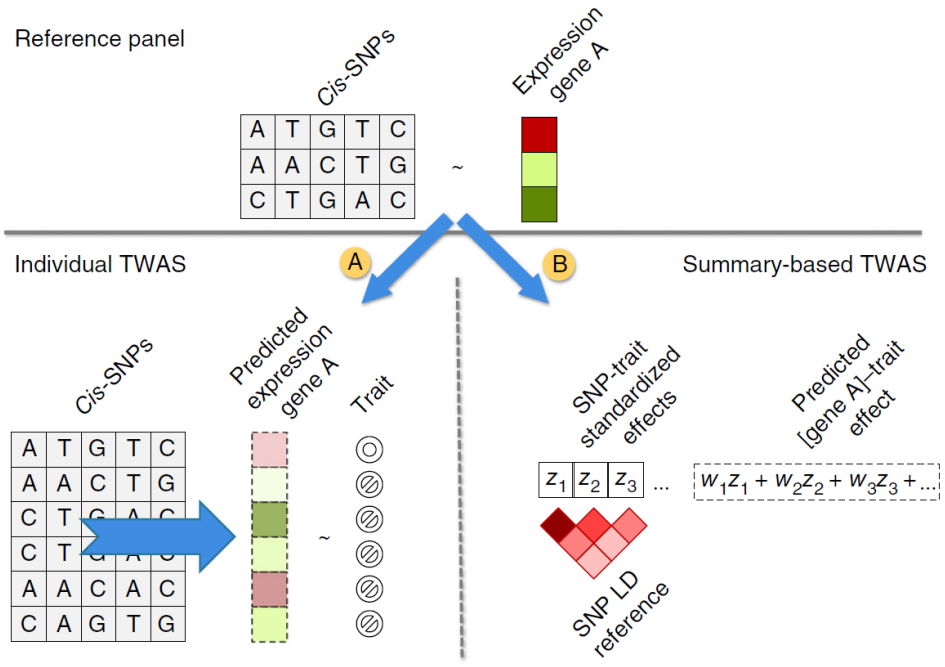
Target population

Person	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	Expression
Target-001	AA	AA	TT	TT	TT	?
Target-002	GG	GC	CC	CC	GG	?

TWAS – cis expression

- GTEx eQTL data
- Model that predicts expression data using the GWAS SNPs (instead of testing for association between those SNPs and expression)
- Test that expression for association with trait
- Repeat for each of the 44 GTEx tissues
- Software Predixcan

Figure 1 Schematic of the TWAS approach. Top, estimate gene expression effect sizes in the reference panel either directly (eQTL), modeling all SNPs (BLUP), or modeling SNPs and effect sizes (BSLMM). Path A: predict expression directly for genotyped samples using the effect sizes from the reference panel and measure the association between predicted expression and a trait. Path B: indirectly estimate association between predicted expression and a trait as the weighted linear combination of SNP-trait standardized effect sizes while accounting for LD among SNPs.

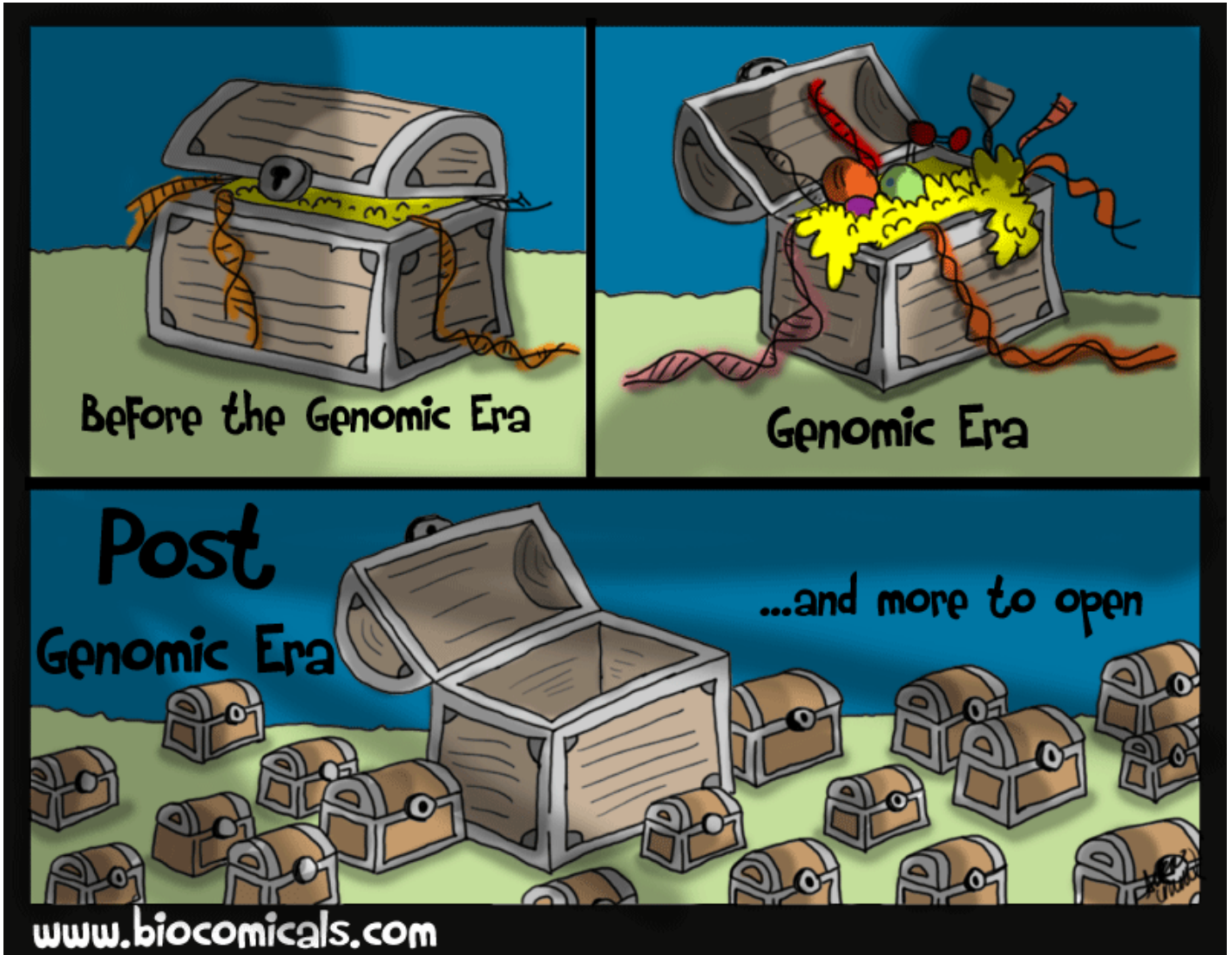


TWAS results

Table 1 TWAS significant genes with no known GWAS risk variants within 500 kb

GWAS	Training expression	Gene	Chr.	Locus start	Locus end	*P* value			
						TWAS	Permuted	Best *cis*-SNP	
LOCKE.BMI	Omnibus	*INO80E*	16	29,506,615	30,517,114	3×10^{-9}	3×10^{-7}	3×10^{-6}	*
LOCKE.BMI	Omnibus	*FTSJ3*	17	61,396,793	62,407,372	1×10^{-7}	5×10^{-6}	9×10^{-7}	*
LOCKE.BMI	Omnibus	*PAM*	5	101,589,685	102,866,809	3×10^{-7}	4×10^{-9}	3×10^{-3}	*
LOCKE.BMI	YFS	*GGNBP2*	17	34,400,737	35,446,278	1×10^{-6}	7×10^{-5}	3×10^{-6}	*
LOCKE.BMI	YFS	*MYO19*	17	34,351,477	35,399,284	3×10^{-6}	4×10^{-5}	3×10^{-6}	*
LOCKE.BMI	Omnibus	*OPRL1*	20	62,211,526	63,231,996	3×10^{-6}	9×10^{-5}	2×10^{-5}	*
LOCKE.BMI	NTR	*RABGAP1*	9	125,203,287	126,367,145	4×10^{-6}	3×10^{-5}	1×10^{-5}	*
LOCKE.BMI	YFS	*SMARCD2*	17	61,409,444	62,420,425	5×10^{-6}	9×10^{-5}	9×10^{-7}	*
LOCKE.BMI	METSIM	*AL049840.1*	14	103,677,607	104,679,149	5×10^{-6}	5×10^{-5}	1×10^{-7}	*
LOCKE.BMI	Omnibus	*LPAR2*	19	19,234,477	20,239,739	6×10^{-6}	3×10^{-5}	4×10^{-6}	*
WILLER.HDL	YFS	*MRPS18B*	6	30,085,486	31,094,172	1×10^{-7}	2×10^{-2}	4×10^{-6}	
WILLER.LDL	Omnibus	*PAM*	5	101,589,685	102,866,809	4×10^{-15}	9×10^{-12}	2×10^{-3}	*
WILLER.LDL	Omnibus	*ITIH4*	3	52,346,991	53,365,495	8×10^{-9}	4×10^{-6}	3×10^{-5}	*
WILLER.LDL	Omnibus	*WARS*	14	100,300,125	101,343,142	1×10^{-8}	6×10^{-7}	3×10^{-5}	*
WILLER.LDL	Omnibus	*MAN2C1*	15	75,148,133	76,160,971	1×10^{-8}	1×10^{-5}	6×10^{-5}	*
WILLER.LDL	YFS	*DHRS13*	17	26,724,799	27,730,089	6×10^{-7}	4×10^{-4}	2×10^{-6}	*
WILLER.LDL	YFS	*ERAL1*	17	26,681,956	27,688,085	8×10^{-7}	9×10^{-4}	2×10^{-6}	
WILLER.LDL	YFS	*HCG27*	6	30,665,537	31,671,745	2×10^{-6}	7×10^{-3}	7×10^{-8}	
WILLER.LDL	YFS	*VARS2*	6	30,376,019	31,394,236	3×10^{-6}	2×10^{-2}	1×10^{-5}	
WILLER.LDL	Omnibus	*PEX6*	6	42,431,608	43,446,958	5×10^{-6}	3×10^{-5}	4×10^{-4}	*
WILLER.LDL	Omnibus	*CSK*	15	74,574,398	75,595,539	6×10^{-6}	1×10^{-4}	1×10^{-5}	*
WILLER.TC	Omnibus	*PAM*	5	101,589,685	102,866,809	9×10^{-15}	3×10^{-13}	5×10^{-3}	*

Thanks!


Before the Genomic Era
Genomic Era
Post
Genomic Era
...and more to open
www.biocomicals.com