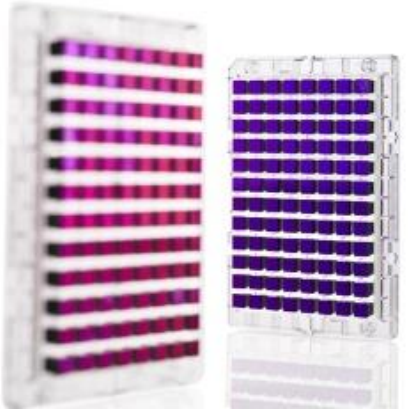
Genome-wide association studies (GWAS) – continued!



Outline for GWAS

- Analysis
  - QC [last lecture]
  - Prostate cancer example [last lecture]
  - Imputation
  - Replication & Meta-analysis
- More analysis
  - Locus characterization
  - eQTLs
  - Limitations, heritability, & “missing heritability”
  - Gene/pathway tests
  - Polygenic models

Recap: Manhattan plot example: Kaiser GWAS of Prostate Cancer

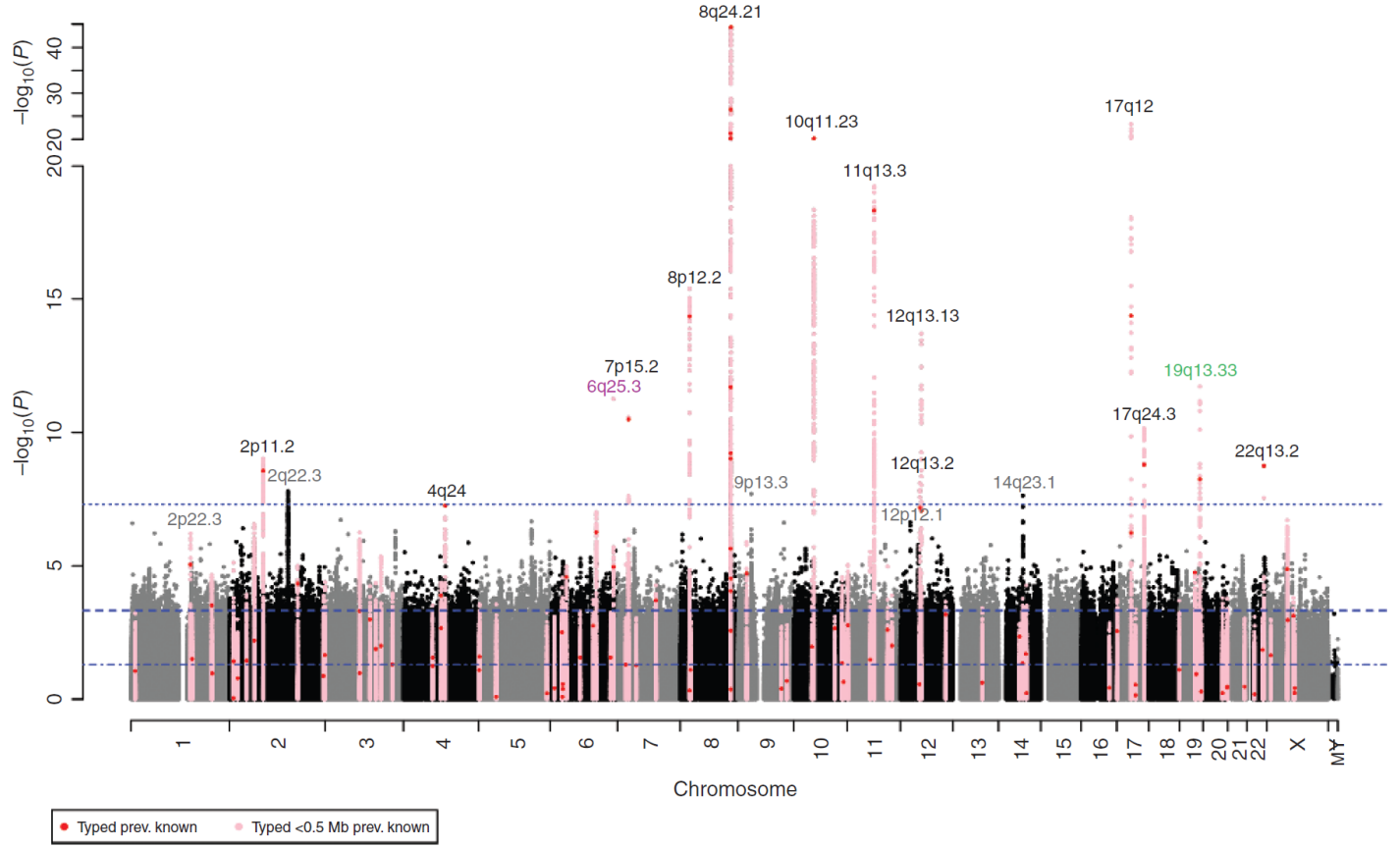
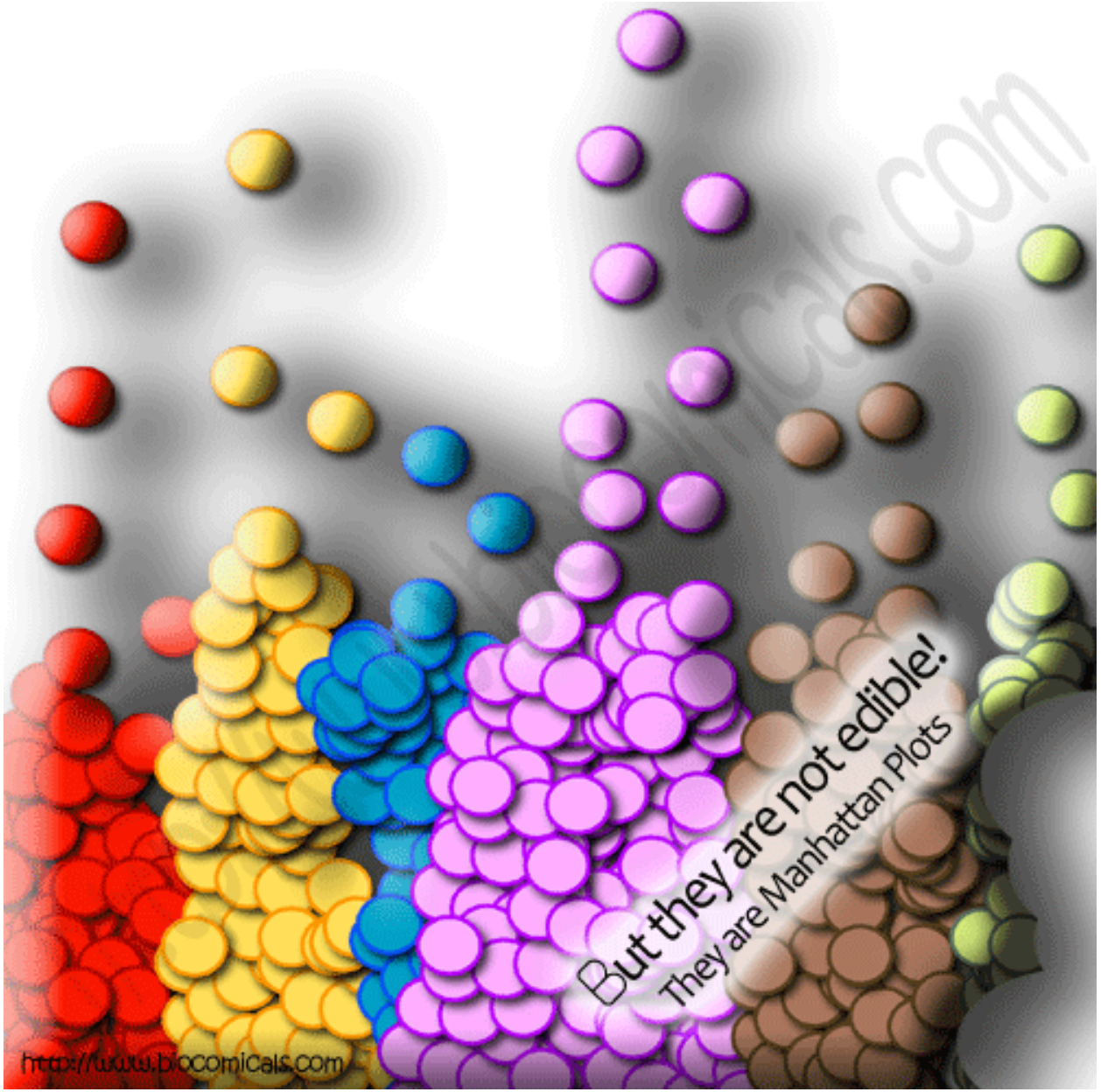


Figure 1. Results from a GWAS of prostate cancer in the KP population (8,399 cases and 38,745 controls), highlighting key chromosomal regions. *P* values are for variant associations with prostate cancer from transethnic fixed-effects meta-analysis of four race/ethnicity GWAS (non-Hispanic white, Latino, African-American, East Asian), each adjusted for age and ancestry principal components. Horizontal dashed lines indicate genome-wide statistical significance ($P < 5 \times 10^{-8}$), Bonferroni-corrected significance for 105 known prostate cancer risk variants ($P < 0.05/105$), and nominal significance ($P < 0.05$). Red points denote our findings for the 105 known risk SNPs, and pink indicates results for SNPs within a 0.5 Mb window around these SNPs. The new 6q25.3 indel rs4646284 is colored magenta, and the 19q13.33 prostate cancer or PSA SNP rs2659124 is noted in green. Those loci containing previously reported variants that we replicated at genome-wide significance are noted in black text. Loci with variants that were novel in the KP GWAS but failed to replicate are noted in gray text.

Hoffmann et. Al, Cancer Discovery 2015


But they are not edible!
They are Manhattan Plots
http://www.biocomicals.com

Replication, Replication, Replication

- To replicate:
  - Association test for replication sample significant at 0.05/{Number of SNPs replicating} alpha level
  - Same genetic model (e.g. additive, dominant)
  - Same direction
  - Sufficient sample size for replication!
    - [e.g., do a power calculation]
- Non-replications not necessarily a false positive
  - LD structures, different populations (e.g., flip-flop)
  - covariates, phenotype definition, underpowered
  - Winner's curse

Meta-analysis

- Combine multiple studies to increase power
- Either combine p-values (Fisher's test),
- or coefficient estimates + standard error (*much* better)

Replication & Meta-analysis

Table 2. Prostate cancer genome-wide association study results for new risk indel rs4646284 at 6q25.3 and for risk SNP rs2659124 at 19q13.33

A

Variant	Risk allele	Ref. allele	Group	Risk AF	No conditioning		Conditioning on other 6q25.3 SNPs		
					OR (95% CI)	*P*	OR (95% CI)	*P*	r^2_{info}
Indel: rs4646284	TG (indel)	T	KP NH white	0.298	1.17 (1.12–1.23)	6.7×10^{-11}	1.16 (1.10–1.23)	3.3×10^{-7}	0.91
			KP Latino	0.250	1.13 (0.94–1.36)	0.20	1.06 (0.85–1.31)	0.61	0.89
			KP Asian	0.324	1.03 (0.84–1.27)	0.76	0.91 (0.68–1.23)	0.56	0.89
			KP Af-Amer	0.365	1.18 (1.02–1.37)	0.029	1.21 (1.03–1.42)	0.020	0.85
			KP Meta FE	—	1.17 (1.12–1.22)	5.5×10^{-12}	1.15 (1.09–1.21)	8.5×10^{-8}	—
			KP Meta RE	—	1.17 (1.12–1.22)	5.5×10^{-12}	1.15 (1.07–1.22)	7.7×10^{-5}	—
			MEC	0.361	1.13 (1.03–1.25)	0.0094	1.13 (1.03–1.25)	0.014	0.81
			PEGASUS	0.278	1.27 (1.17–1.37)	1.4×10^{-8}	1.20 (1.09–1.32)	0.00024	0.80
			Overall Meta FE	—	1.18 (1.14–1.22)	1.0×10^{-19}	1.16 (1.11–1.21)	5.4×10^{-12}	—
			Overall Meta RE	—	1.18 (1.13–1.23)	9.8×10^{-17}	1.16 (1.11–1.21)	5.4×10^{-12}	—

Hoffmann et. Al, Cancer Discovery 2015`

Beyond Genotyped SNPs: Imputation of SNP Genotypes

- Combine data from different platforms (e.g., Affy & Illumina) (for replication / meta-analysis).
- Estimate unmeasured or missing genotypes.
- Based on measured SNPs and external info (e.g., haplotype structure of HapMap).
- Increase GWAS power (impute and analyze all), e.g. Sick sinus syndrome, most significant was 1000 Genomes imputed SNP (Holm et al., Nature Genetics, 2011)
- HapMap as reference, now 1000 Genomes Project; UK10K, Haplotype reference consortium?
- Hybrid sequencing/genotyping approaches

Generalization of Tag SNPs: Imputed SNPs using a Reference Panel

- Generalization of pairwise r^2 of tagSNPs to multi-marker correlation

Study sample

....A.....A....A...
....G.....C....A...

Reference haplotypes

CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTCGTGC
CGAAGCTCTTTTCTTCTGTGC

Class Exercise!

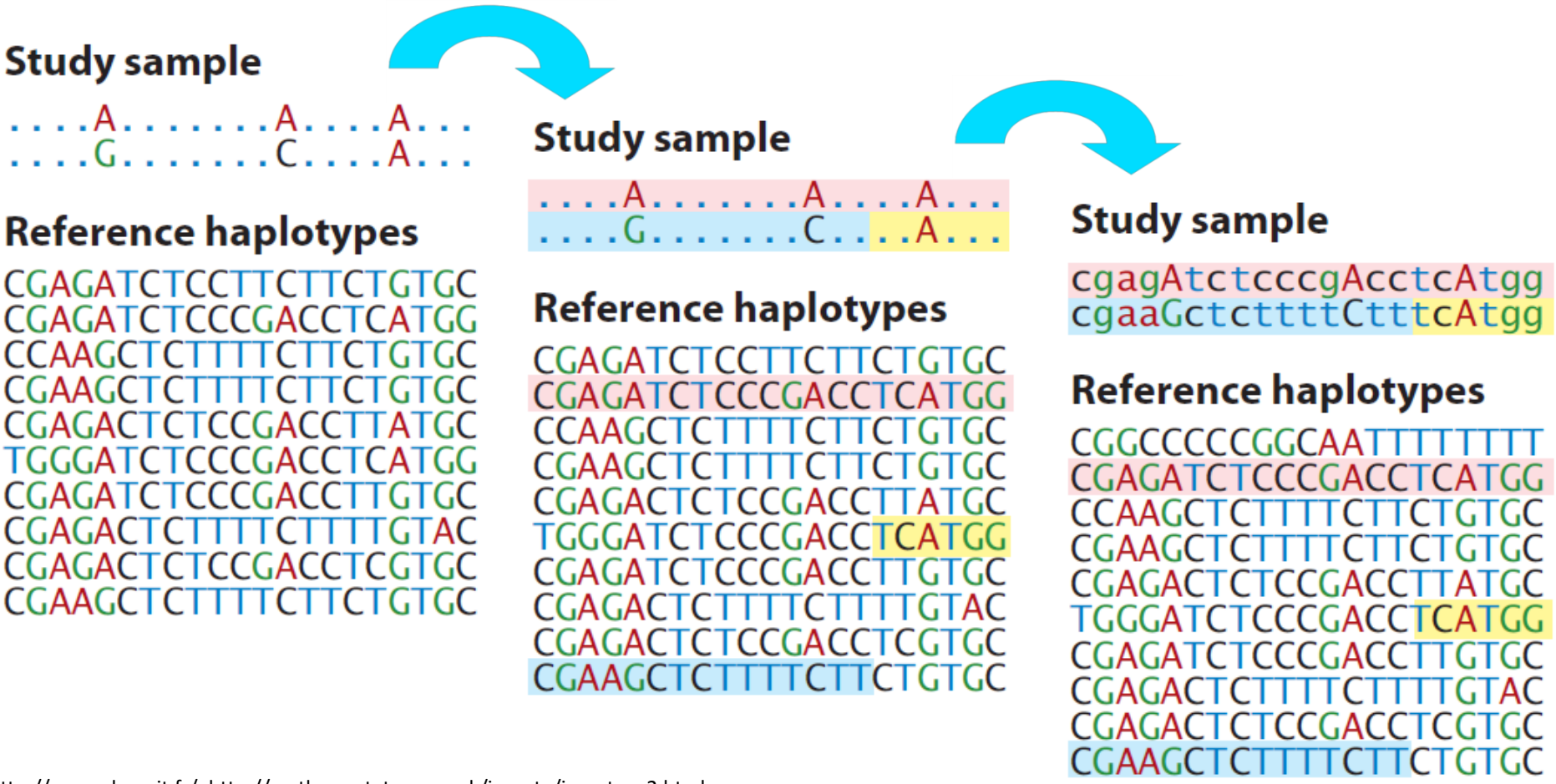
How would you fill in the remaining genotypes of this study sample individual, given the reference haplotypes?

- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

Li et al., Ann Rev Human Genet, 2009

Generalization of Tag SNPs: Imputed SNPs using a Reference Panel

- Generalization of pairwise r^2 of tagSNPs to multi-marker correlation

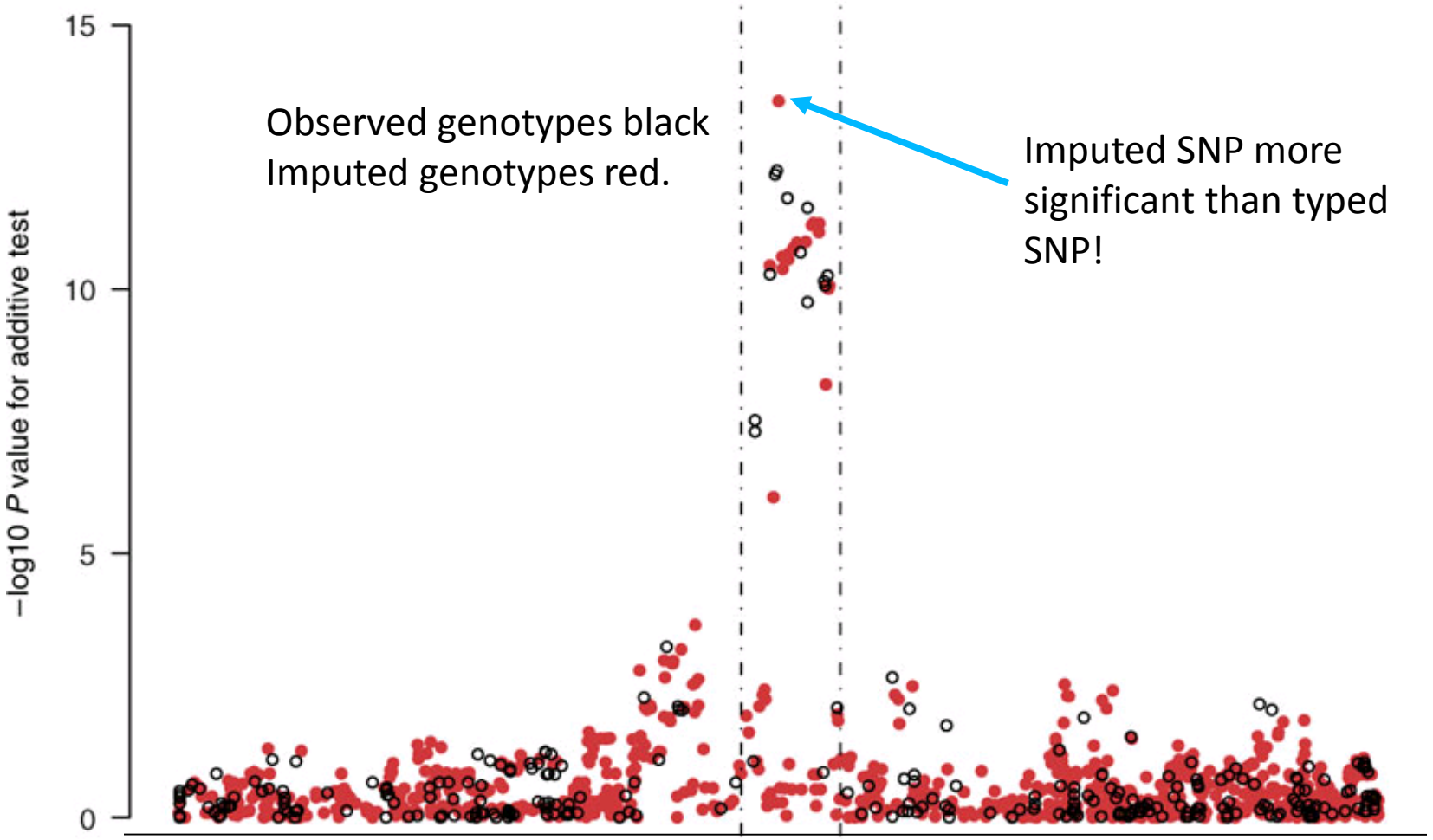


- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

Li et al., Ann Rev Human Genet, 2009

Imputation Application

TCF7L2 gene region & T2D from the WTCCC data



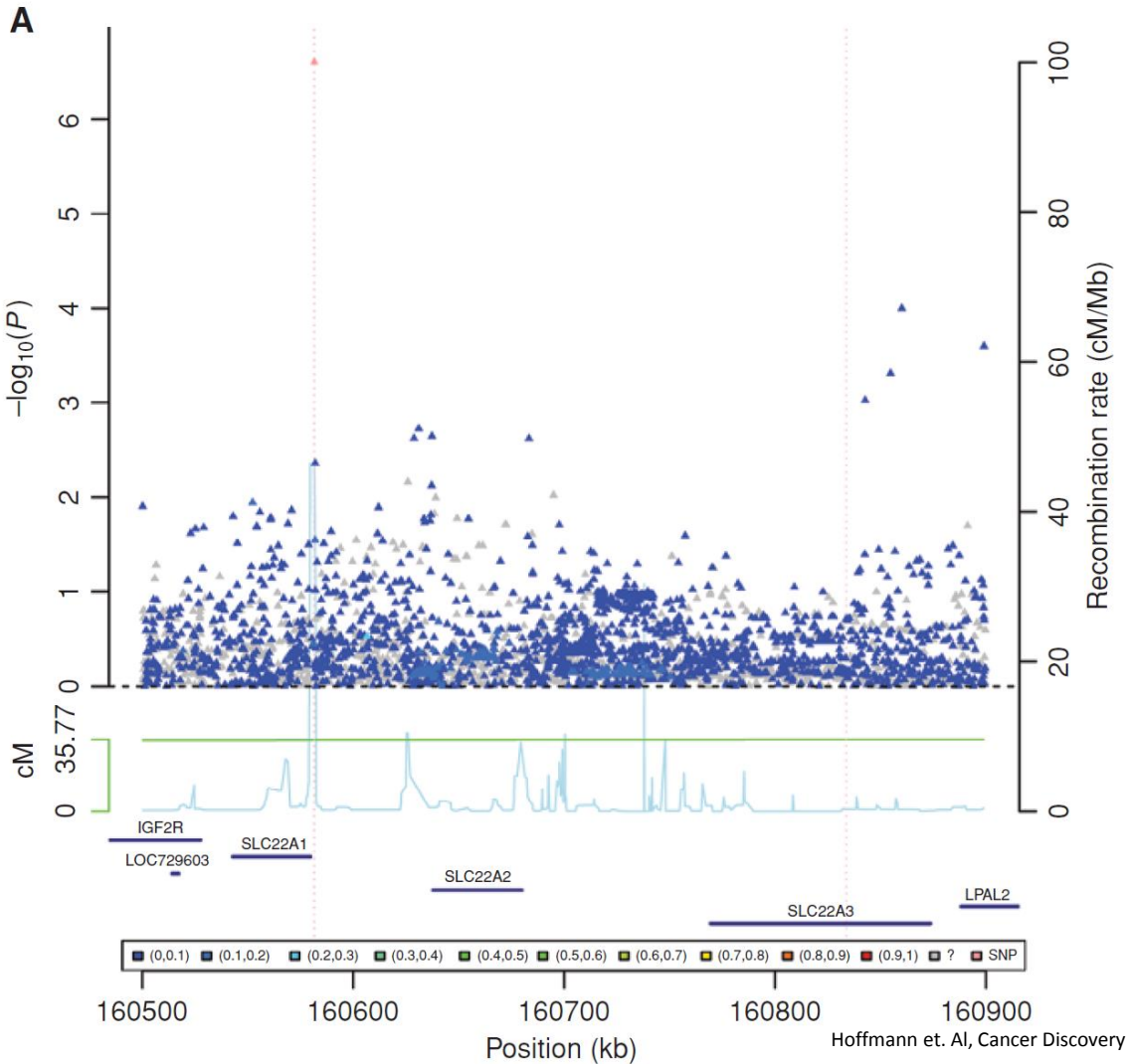
Marchini Nature Genetics2007

http://www.stats.ox.ac.uk/~marchini/#software

Imputation Application #2 – Do not always see a jet of points, even for common SNP…

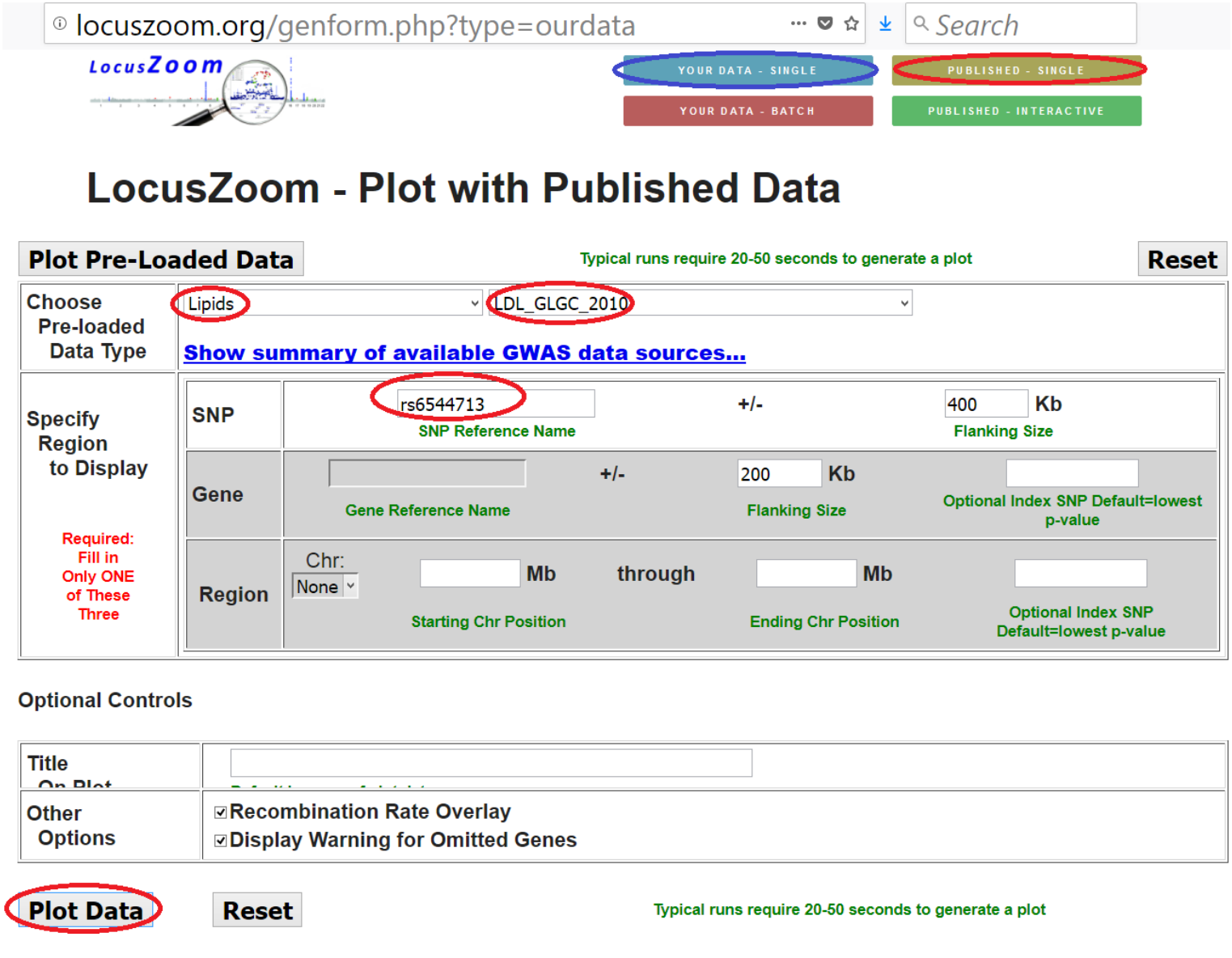
Figure 2. Regional fixed-effects meta-analysis plots from GWAS in KP population of the two risk variants that replicated: (**A**) 6q25.3 (rs4646284), (**B**) 19q13.33 (rs2659124). The color code for the points represents the r^2 of each SNP with the risk variant (ranges defined in the legend). The dotted vertical lines pass through the risk variants and the other significant SNPs in the region, on which analyses were conditioned.

(Same SNP as 4 slides ago)

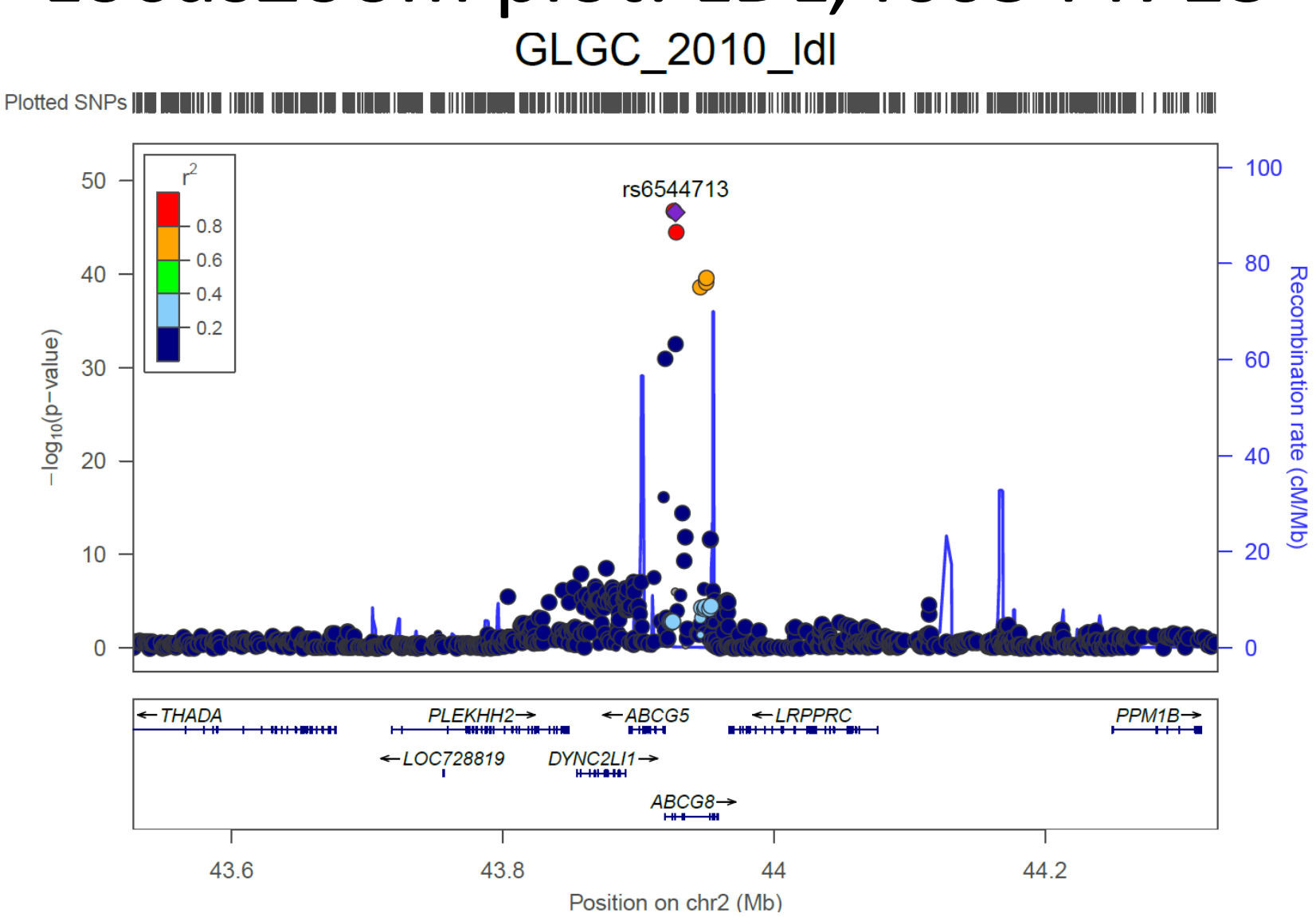


Hoffmann et. Al, Cancer Discovery 2015`

LocusZoom plot: LDL, rs6544713



LocusZoom plot: LDL, rs6544713



Imputation QC – r_{info}^2 value

- Similar to pairwise tagging r^2 value, but multi-marker r^2 value
  - More than pairwise r^2, why it can beat pairwise tagging r^2 value
- Rules of thumb
  - $r^2>0.8$: very good
  - $0.5<r^2<0.8$: not as good, maybe include
  - $0.3<r^2<0.5$: not great, maybe include
  - $r^2<0.3$: probably exclude from analysis

Imputation example: How rare can you go?

- G84E example talked about in Austin pg. 62
- MAF 0.0034 in European ancestry. r^2=0.57.
- In general, a moving target, still unknown how rare is possible

PLOS | GENETICS

RESEARCH ARTICLE

Imputation of the Rare *HOXB13* G84E Mutation and Cancer Risk in a Large Population-Based Cohort

Thomas J. Hoffmann[1,2], Lori C. Sakoda[3], Ling Shen[3], Eric Jorgenson[3], Laurel A. Habel[3], Jinghua Liu[2], Mark N. Kvale[2], Maryam M. Asgari[3], Yambazi Banda[2], Douglas Corley[3], Lawrence H. Kushi[3], Charles P. Quesenberry Jr.[3], Catherine Schaefer[3], Stephen K. Van Den Eeden[3,4], Neil Risch[1,2,3], John S. Witte[1,2,4]*

1 Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America, **2** Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, **3** Division of Research, Kaiser Permanente, Northern California, Oakland, California, United States of America, **4** Department of Urology, University of California San Francisco, San Francisco, California, United States of America

HOXB13 G84E (continued)

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

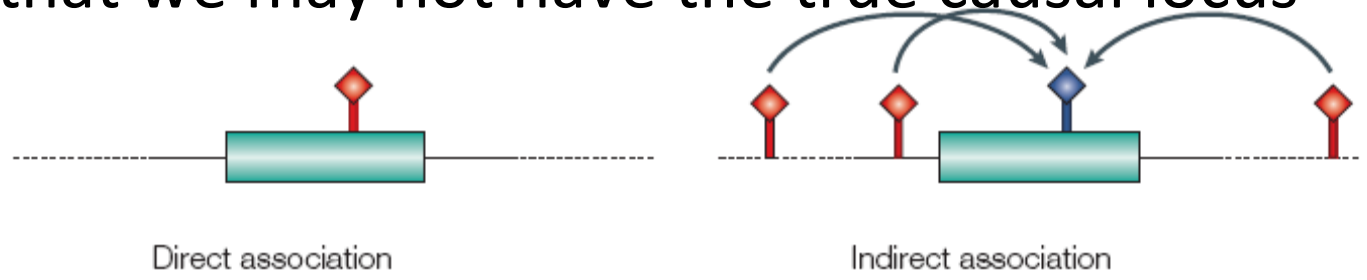
Cancer[a]	# Case carriers[b]	# Cases[b]	Frequency[c]	Odds Ratio[d] (95% CI)	Corrected Odds Ratio[e] (95% CI)	P-value[d]
Kidney	5	303	0.94%	2.32 (0.94, 5.76)	3.32 (0.89, 9.99)	0.034
Bladder	5	335	0.85%	2.05 (0.81, 5.22)	2.84 (0.70, 8.94)	0.065
Non-Hodgkin's Lymphoma	10	846	0.67%	1.81 (0.98, 3.36)	2.42 (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

Outline for GWAS

- Analysis
  - QC [last lecture]
  - Prostate cancer example [last lecture]
  - Imputation
  - Replication & Meta-analysis
- More analysis
  - Locus characterization
  - eQTLs
  - Limitations, heritability, & “missing heritability”
  - Gene/pathway tests
  - Polygenic models

What after we've discovered (and replicated) a novel variant?

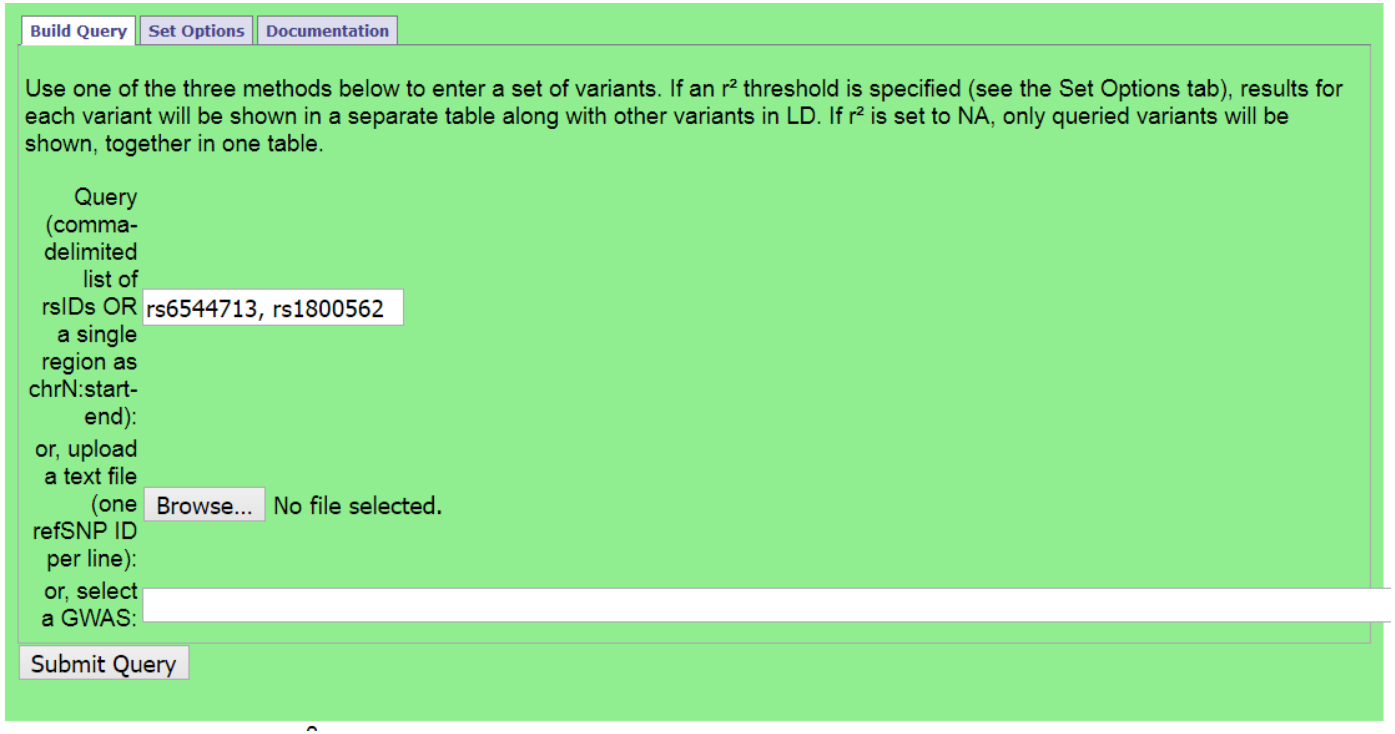
- Recall that we may not have the true causal locus



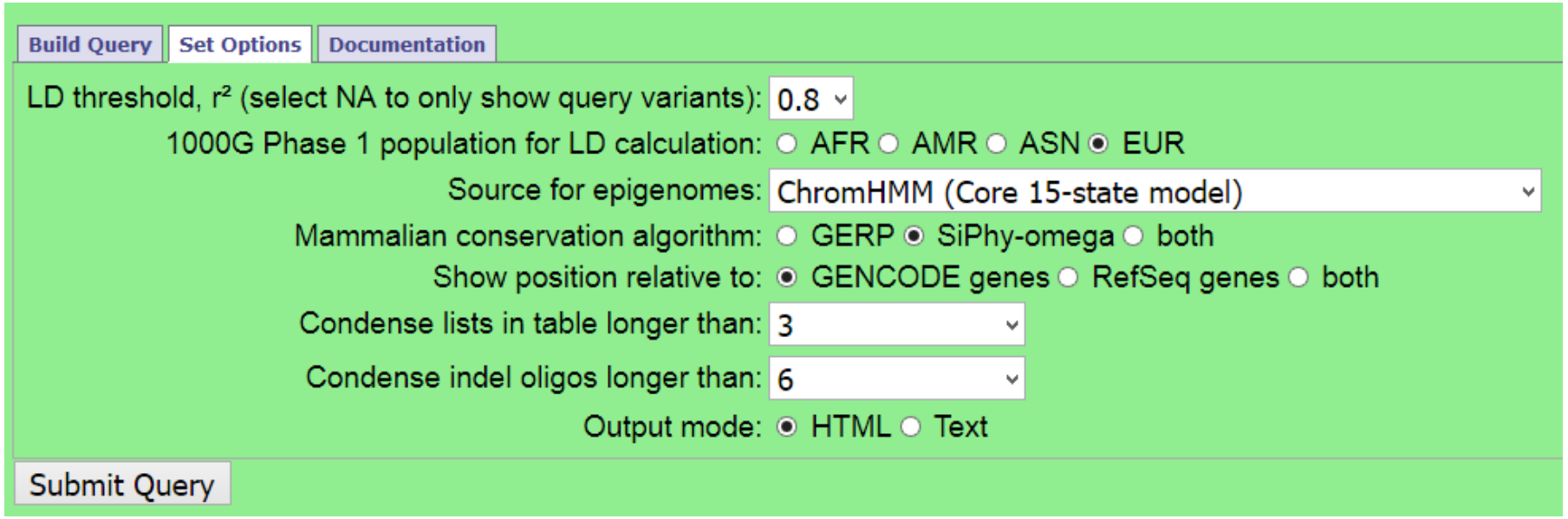
- Fine mapping vs. imputed genotypes
- Discovery not enough these days, want a tie-in to biology
- A lot of work has been done to characterize SNPs and prioritize what should be followed up on
- Functional studies

Refining GWAS SNPs: Haploreg

- http://archive.broadinstitute.org/mammals/haploreg/haploreg.php



Haploreg further options (r^2, 1000 Genomes population)



Which one is causal / biological mechanism?

- Don’t know for certain, can do some exploration:

Query SNP: rs1800562 and variants with r^2 >= 0.8

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
6	26072764	0.89	0.97	rs144861591	C	T	0.00	0.03	0.00	0.05						7 altered motifs			12 hits	15kb 5' of HFE	
6	26092913	1	1	rs1800562	G	A	0.00	0.02	0.00	0.05		BLD	4 tissues	BLD		5 altered motifs	32 hits	5 hits	20 hits	HFE	missense
6	26098246	1	1	rs79220007	T	C	0.00	0.02	0.00	0.05						NRSF			14 hits	HFE	
6	26123274	0.83	0.94	rs115740542	T	C	0.01	0.02	0.00	0.05		23 tissues		53 tissues	29 bound proteins	YY1			11 hits	HIST1H2BC	

- Likely will not know for certain, may prioritize for functional work

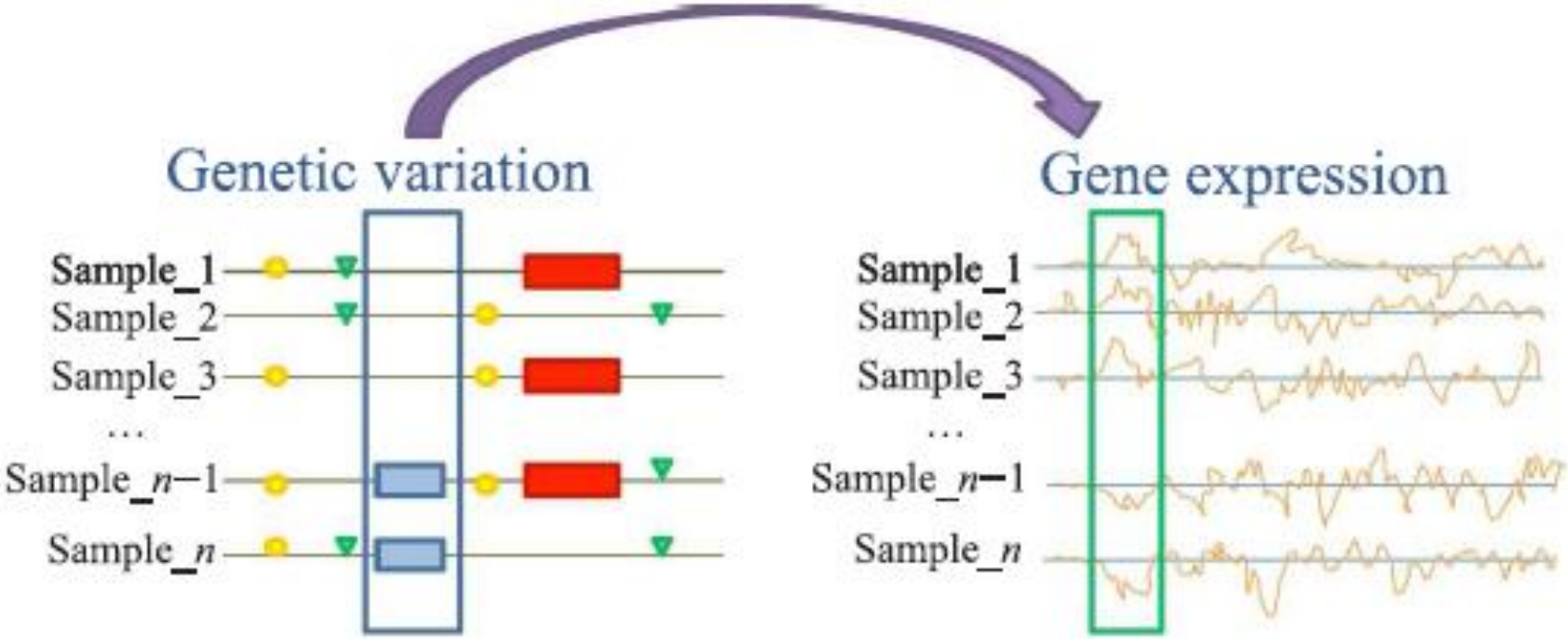
Which one is causal / biological mechanism?

- ?

Query SNP: rs6544713 and variants with r^2 >= 0.8

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
2	43845437	0.96	0.99	rs4299376	G	T	0.85	0.78	1.00	0.70			LIV, GI				5 hits	1 hit	1 hit	ABCG8	intronic
2	43846742	1	1	rs6544713	T	C	0.85	0.78	1.00	0.71			LIV, GI			Pou2f2	1 hit	2 hits	1 hit	ABCG8	intronic
2	43847292	0.99	1	rs4245791	C	T	0.87	0.78	1.00	0.70			LIV, GI	GI		7 altered motifs	2 hits	1 hit	1 hit	ABCG8	intronic
2	43864927	0.84	0.99	rs4603816	C	T	0.93	0.79	1.00	0.74						Cdx,Mef2,Nanog			1 hit	ABCG8	intronic
2	43869000	0.83	0.98	rs6740545	G	A	0.91	0.79	1.00	0.73						BCL				ABCG8	intronic
2	43869120	0.84	0.99	rs6544716	T	G	0.91	0.79	1.00	0.74										ABCG8	intronic
2	43869197	0.83	0.98	rs6755809	T	A	0.92	0.79	1.00	0.73									1 hit	ABCG8	intronic
2	43869263	0.83	0.98	rs6544717	G	A	0.97	0.81	1.00	0.73						Pax-5			1 hit	ABCG8	intronic
2	43869631	0.83	0.97	rs4952688	T	A	0.94	0.80	1.00	0.73						Zfp105			1 hit	ABCG8	intronic
2	43870228	0.82	0.98	rs4953026	C	T	0.91	0.79	1.00	0.74						CACD,TFII-I				ABCG8	intronic

What is an eQTL?



Tian Lu et al., Tsinghua Science and Technology, 2014, 19(6): 624-634.

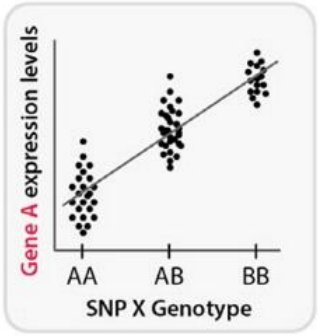
What is an eQTL?

Cis-eQTL

SNP X has an effect on local Gene A

SNP X located in promoter region

Gene A transcription factor located on chromosome 1



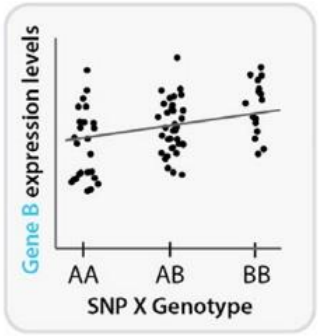
Altered **Protein A** levels, effect on the binding to the transcription factor binding sites of downstream genes

Trans-eQTL

SNP X has an effect on distant Gene B through an intermediary factor (such as a transcription factor)

Protein A binding site

Gene B located on chromosome 2



Harm-Jan Westra & Lude Franke, Bichimica et Biophysica Acta – Molecular Basis of Disease 1842(10):1892-1902

What is an eQTL?

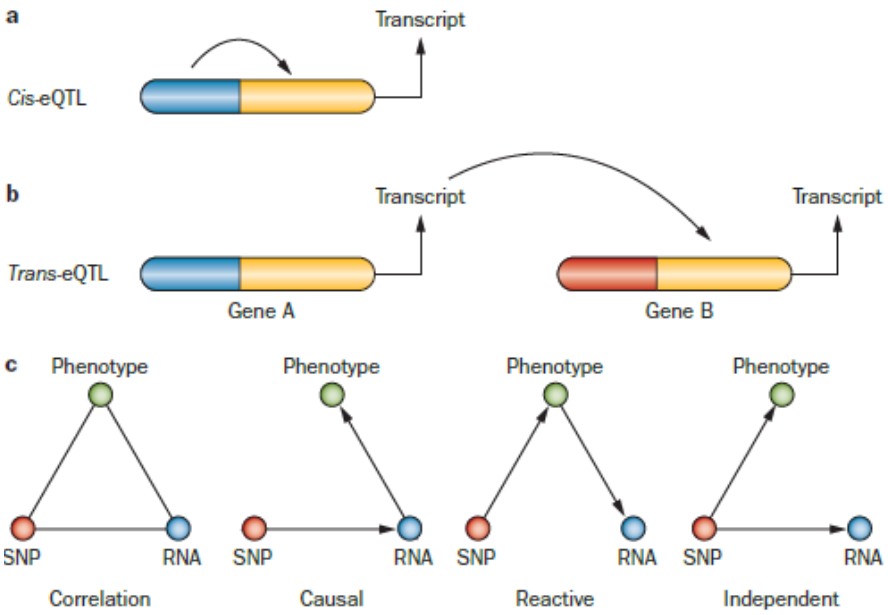
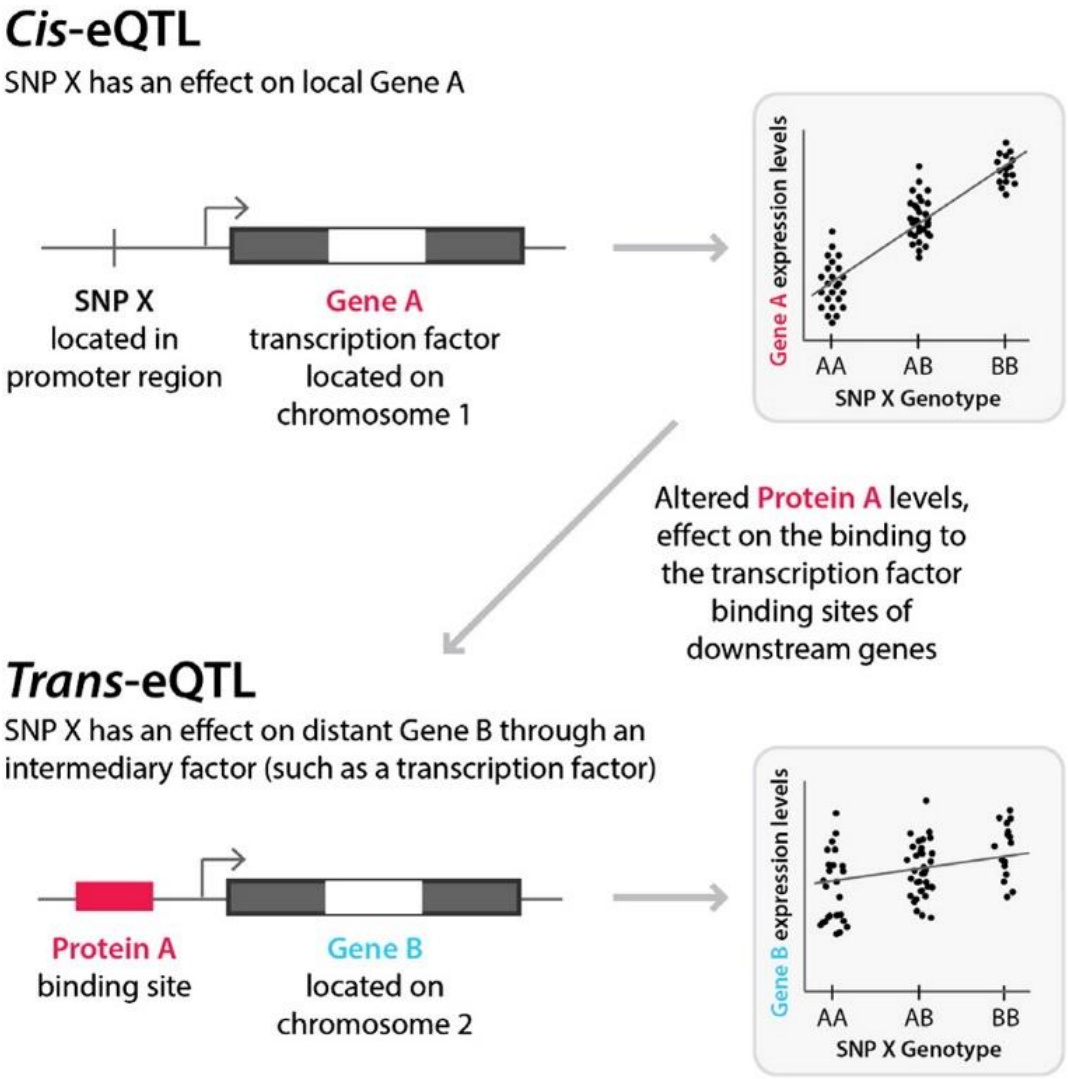


Figure 2 | Principles of eQTL analysis. Loci that control transcript levels are referred to as eQTL. Typically, eQTL analysis of humans or mice reveals thousands of eQTL, which are classified as either 'local' or 'distant', depending on the distance of the locus from the gene that an eQTL regulates. **a** | A local eQTL most likely acts in *cis*, meaning it affects the expression of a gene on the same chromosome (for example, an eQTL in a promoter only influences expression of the contiguous gene). **b** | Distal loci act in *trans*, affecting both copies of the genes they regulate (for example, an eQTL may encode a transcription factor that controls the expression of the regulated gene on both chromosomes). **c** | This panel illustrates possible causal relationships between a SNP, a transcript (RNA) and a physiological or pathologic trait (phenotype). Using the SNP as a 'causal anchor', causal relationships between the three can be modeled.[71] Abbreviations: eQTL, expression quantitative trait locus; SNP, single nucleotide polymorphism.

E.g. of cis-eQTL – transcription factor

W. Robb MacLellan et al, Nature Reviews Cardiology, 9:172-184

What is an eQTL?

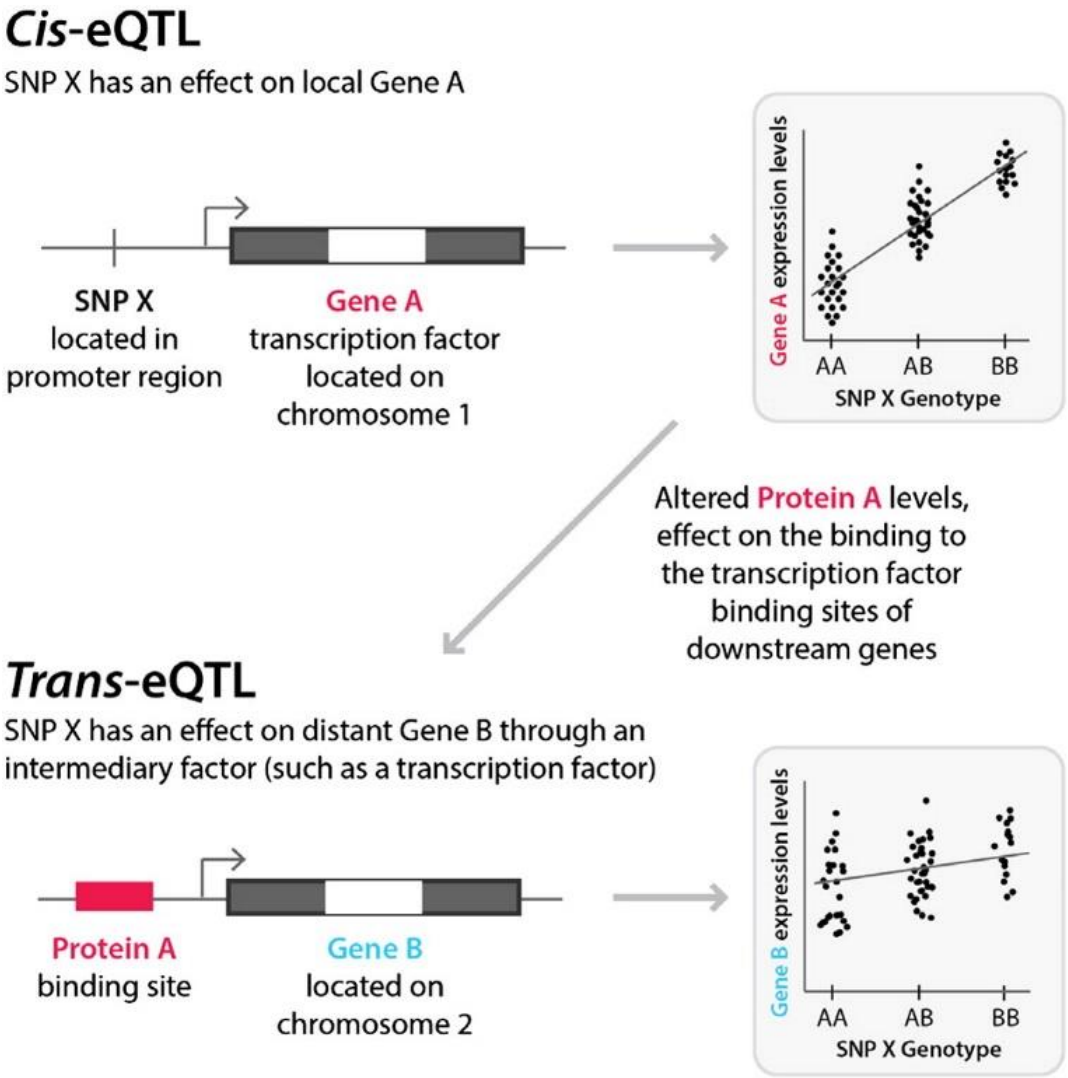


<u>CLASS EXERCISE</u>

- Design an eQTL study!
- (Assume that the expression of each gene has been quantified. Or sort of. This is expensive.)

Harm-Jan Westra & Lude Franke, Bichimica et Biophysica Acta – Molecular Basis of Disease 1842(10):1892-1902

What is an eQTL?



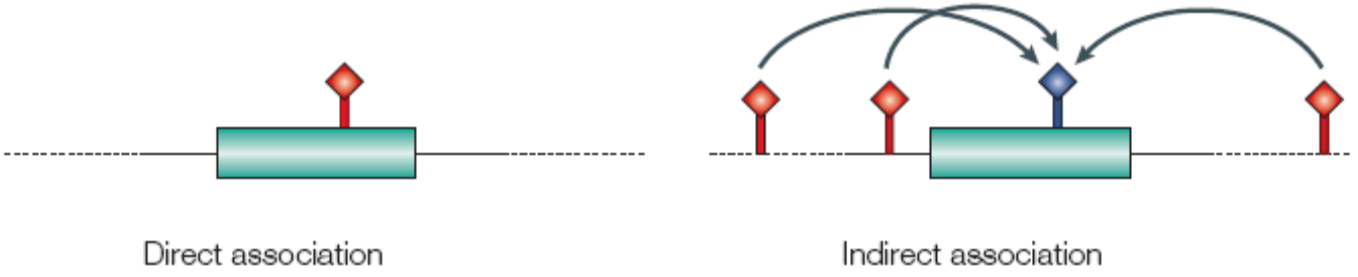
CLASS EXERCISE

- Design an eQTL study!
- (Assume that the expression of each gene has been quantified. Or sort of. This is expensive.)
  - What kind of genotyping method?
  - Advantages / Disadvantages to testing only cis, only trans, both? How many tests?
  - How far up/downstream of each gene?

Harm-Jan Westra & Lude Franke, Bichimica et Biophysica Acta – Molecular Basis of Disease 1842(10):1892-1902

What is an eQTL?

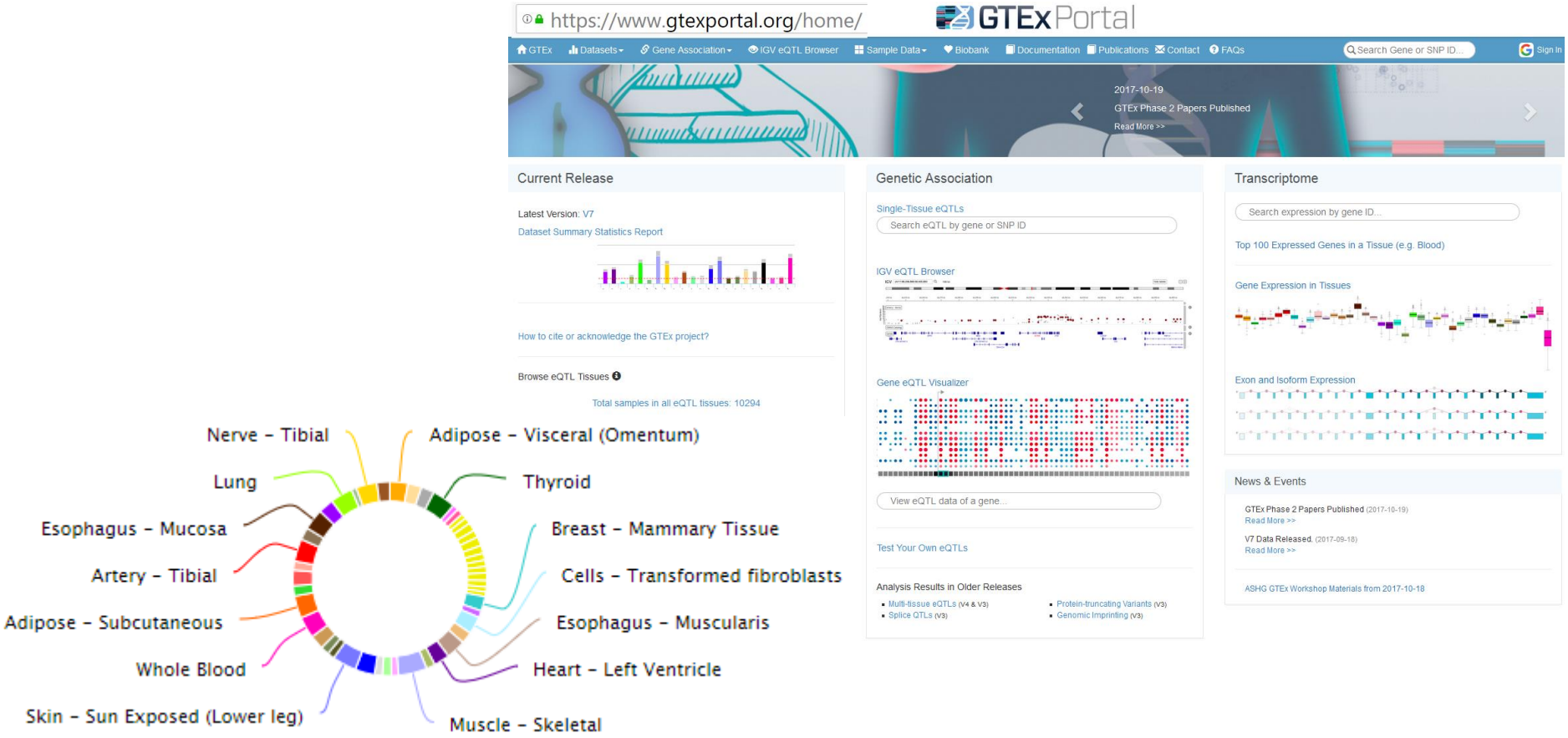
- The expression of a gene can be viewed as a quantitative trait.
- Test the SNPs around the gene for association with the expression of the gene.
- Does the SNP affect expression?
- The same problem still applies, don't know causal locus!



- Might differentiate between which gene expression is being altered, if between or amongst more than one.

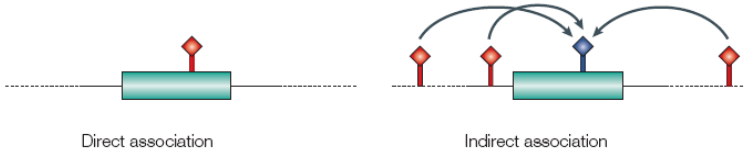
cis eQTLs: GTeX

- Database of SNP associations with gene expression in humans



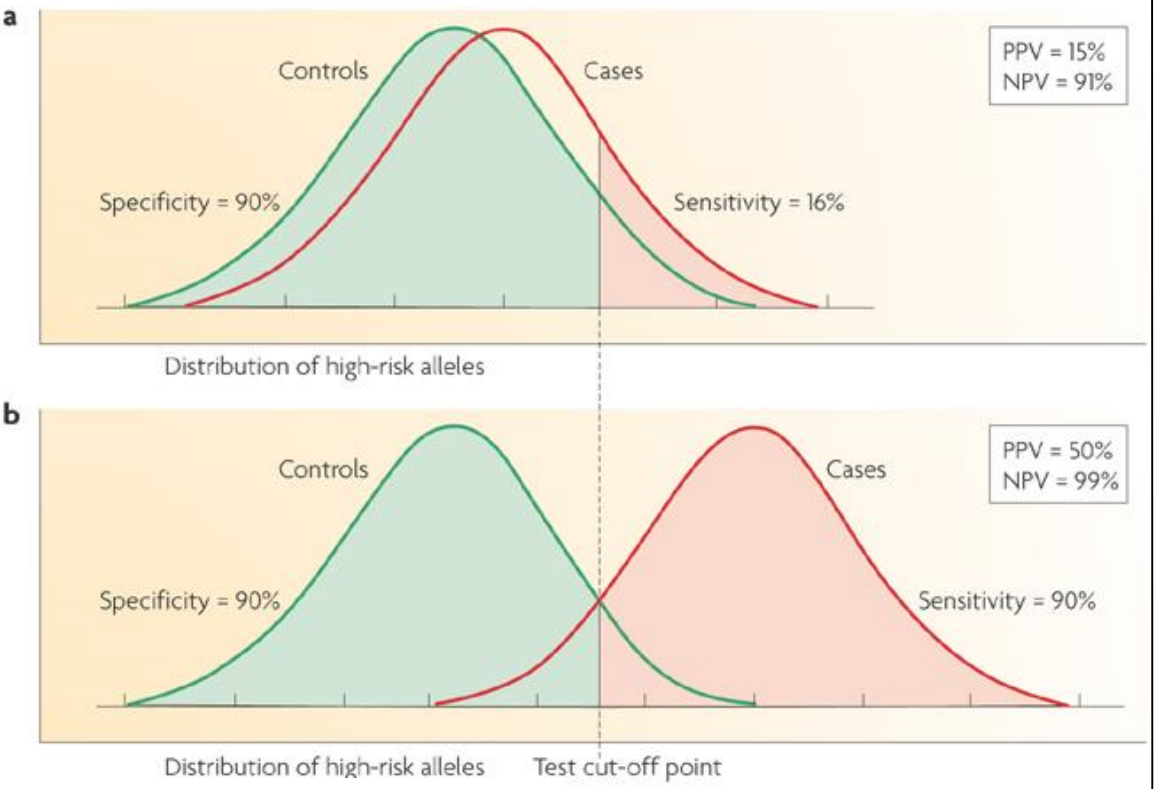
Limitations of GWAS

- 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

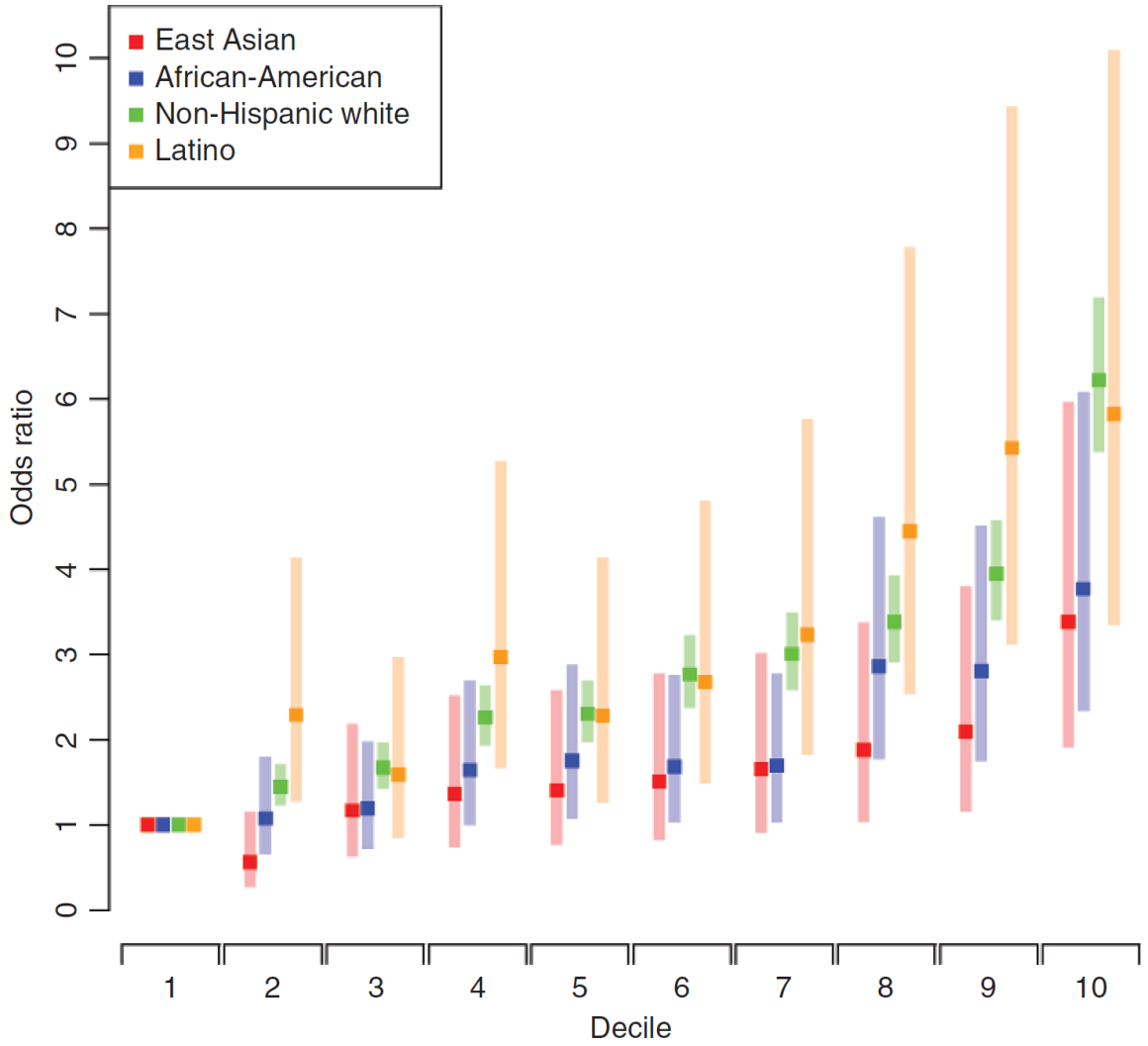
Example:
AUC for Breast Cancer Risk

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

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.



Hoffmann et. Al, Cancer Discovery 2015

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.

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.

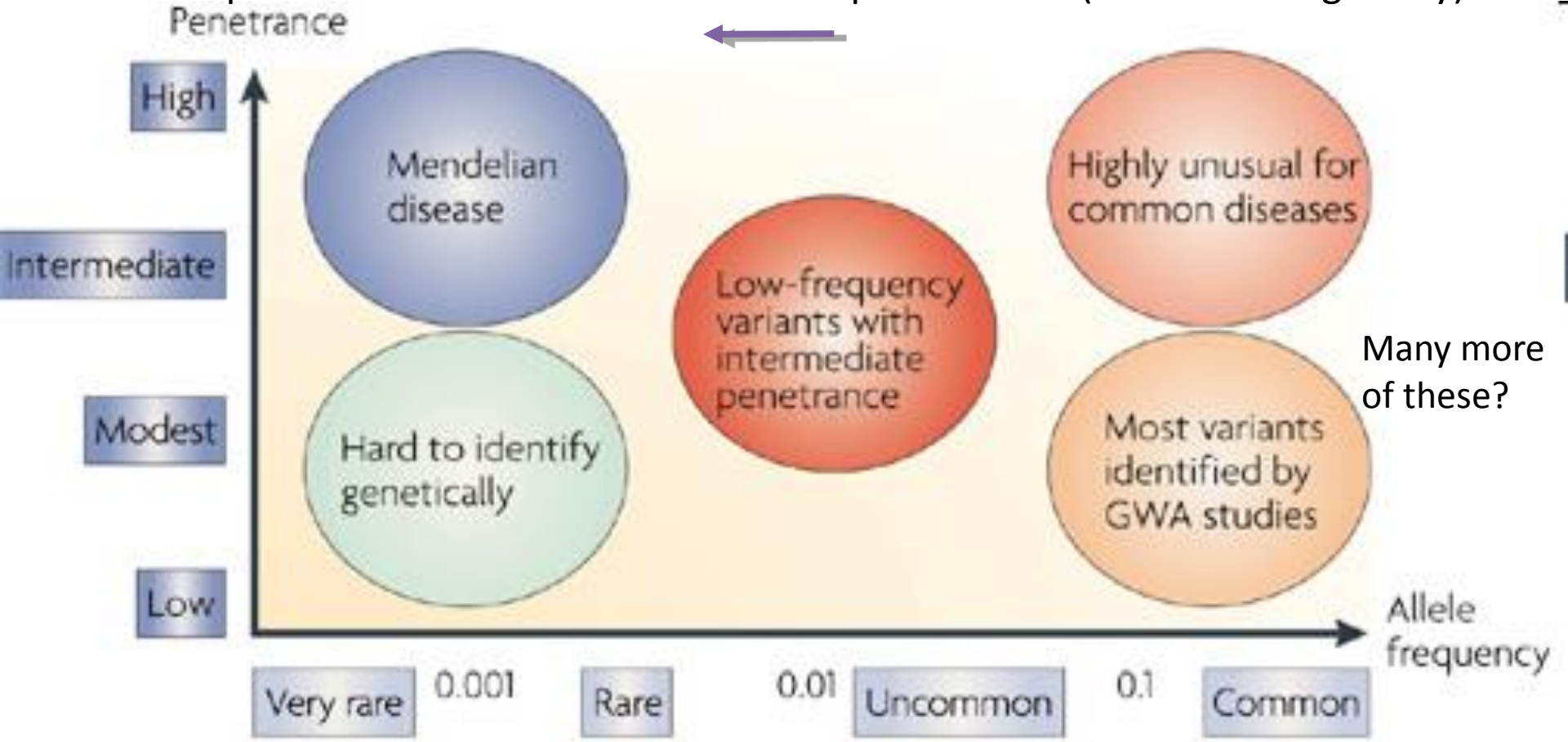
[b] Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.

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

Visscher, AJHG 2011

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.

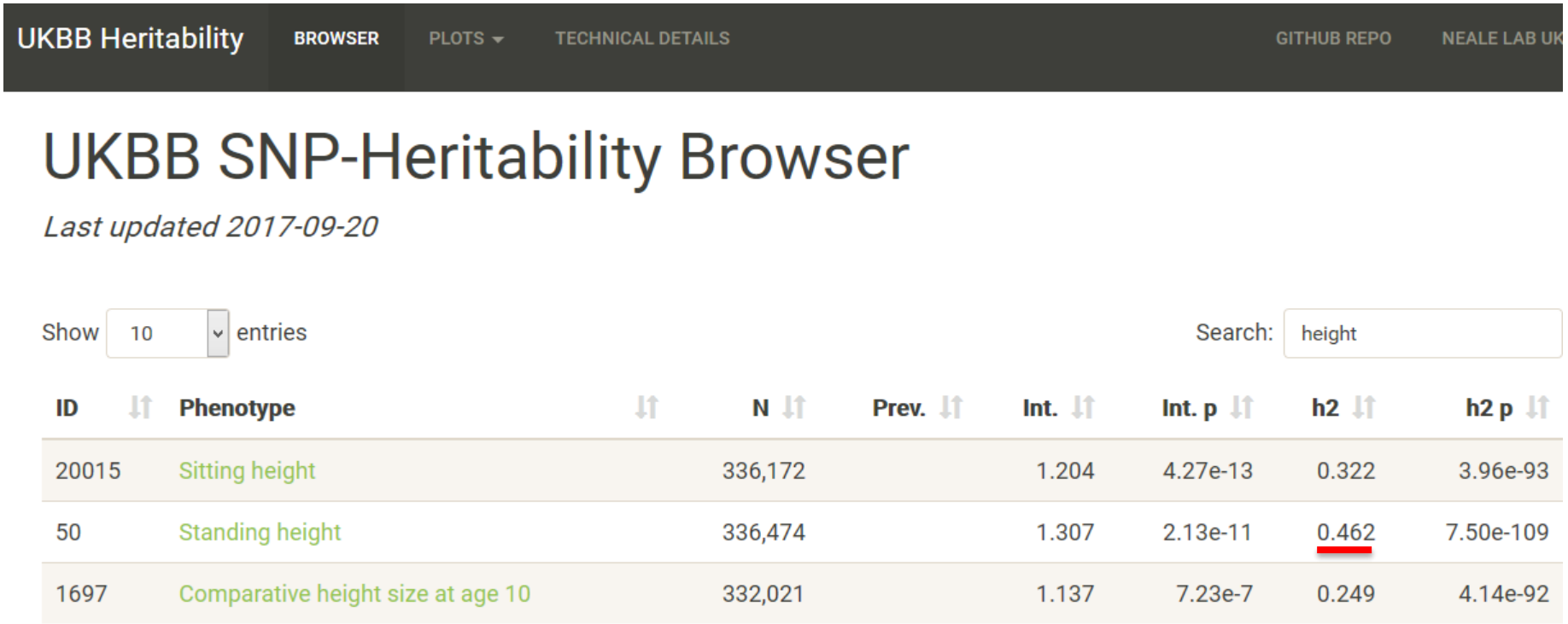
GWASs in human populations have discovered hundreds of SNPs significantly associated with complex traits[1,2], yet for any one of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have 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/



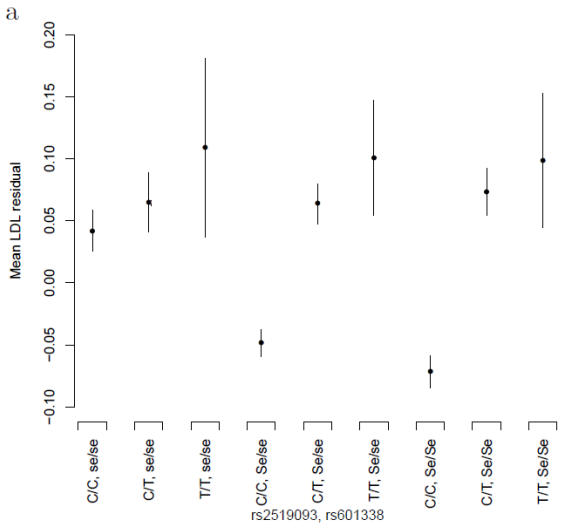
ID	Phenotype	N	Prev.	Int.	Int. p	h2	h2 p
20015	Sitting height	336,172		1.204	4.27e-13	0.322	3.96e-93
50	Standing height	336,474		1.307	2.13e-11	0.462	7.50e-109
1697	Comparative height size at age 10	332,021		1.137	7.23e-7	0.249	4.14e-92

Other sources - Epistasis

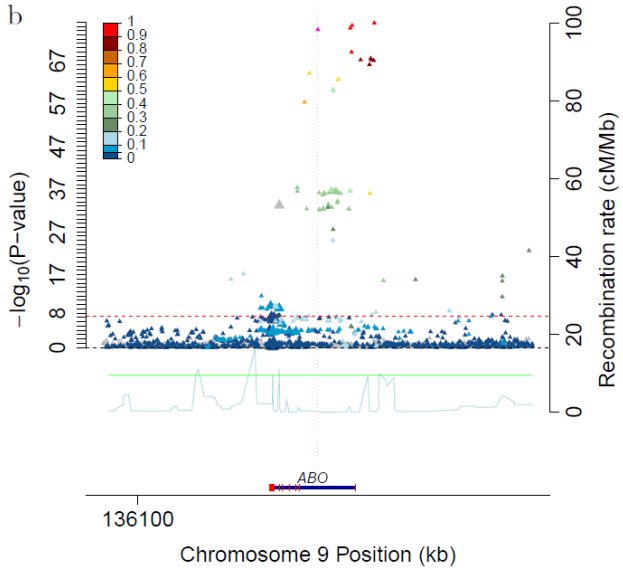
Serum lipids: *ABO* x *FUT2* interaction

- 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

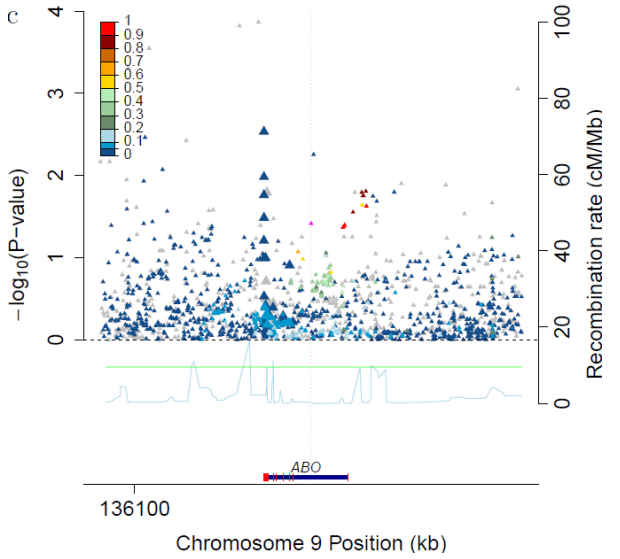
Residual LDL by genotype



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)
- 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):

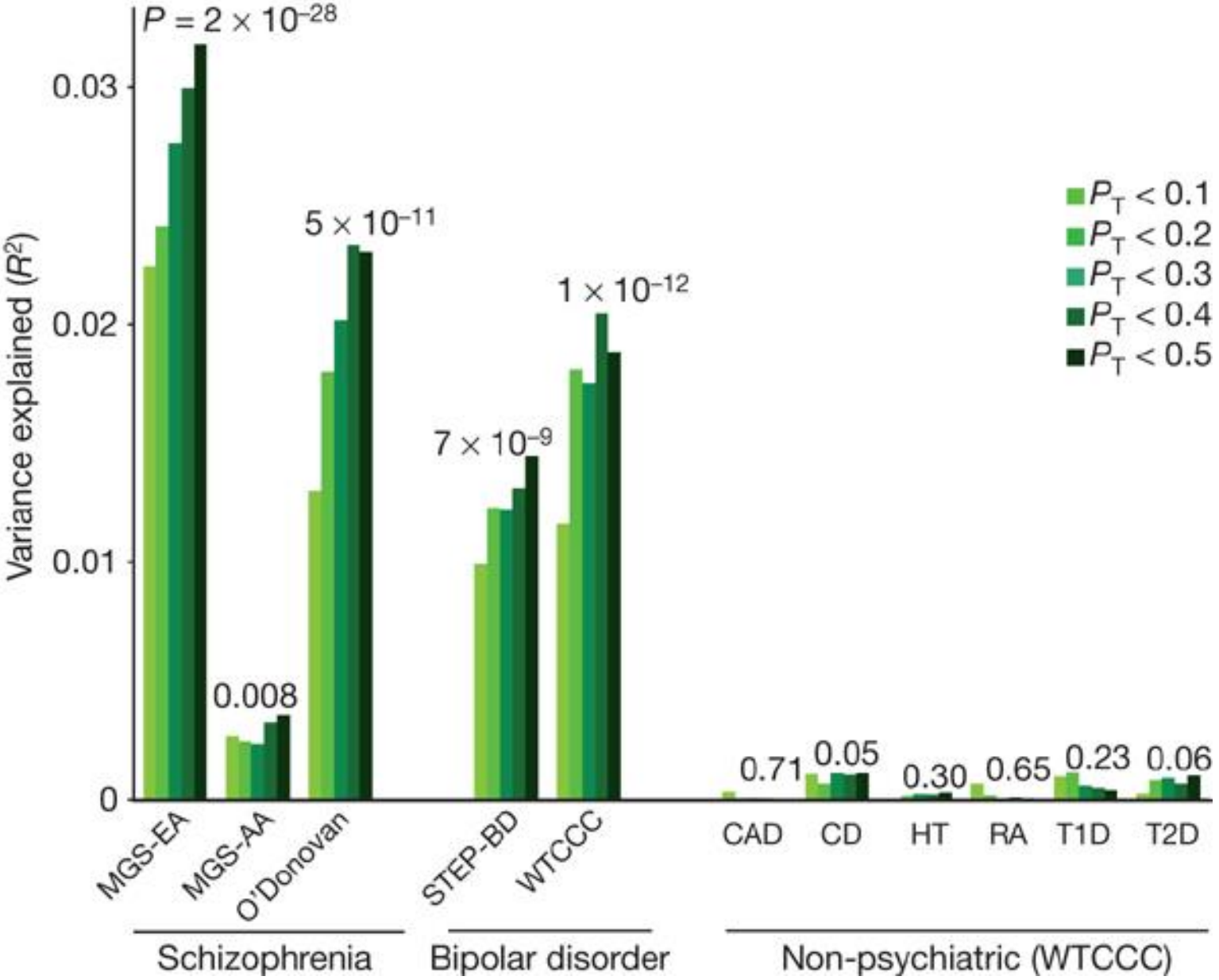
$$x_j = \frac{\sum_{i=1}^{m} \ln\left(OR_i\right) \times SNP_{ij}}{m}$$

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 this idea, but only on ***known hits***

ISC / Purcell et al. *Nature* 2009

Application of Model



Purcell / ISC et al. *Nature* 2009