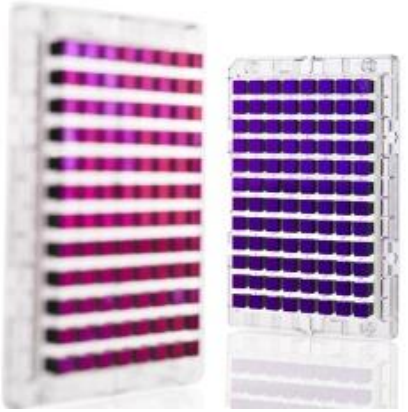
Genome-wide association studies (GWAS) – *still* continued!



Outline for GWAS

- Imputation
- Following up on hits
- eQTLs & Transcriptome-wide association studies (TWAS)
- Limitations, heritability, and “missing heritability”, Polygenic risk scores

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! (Groups ~3)

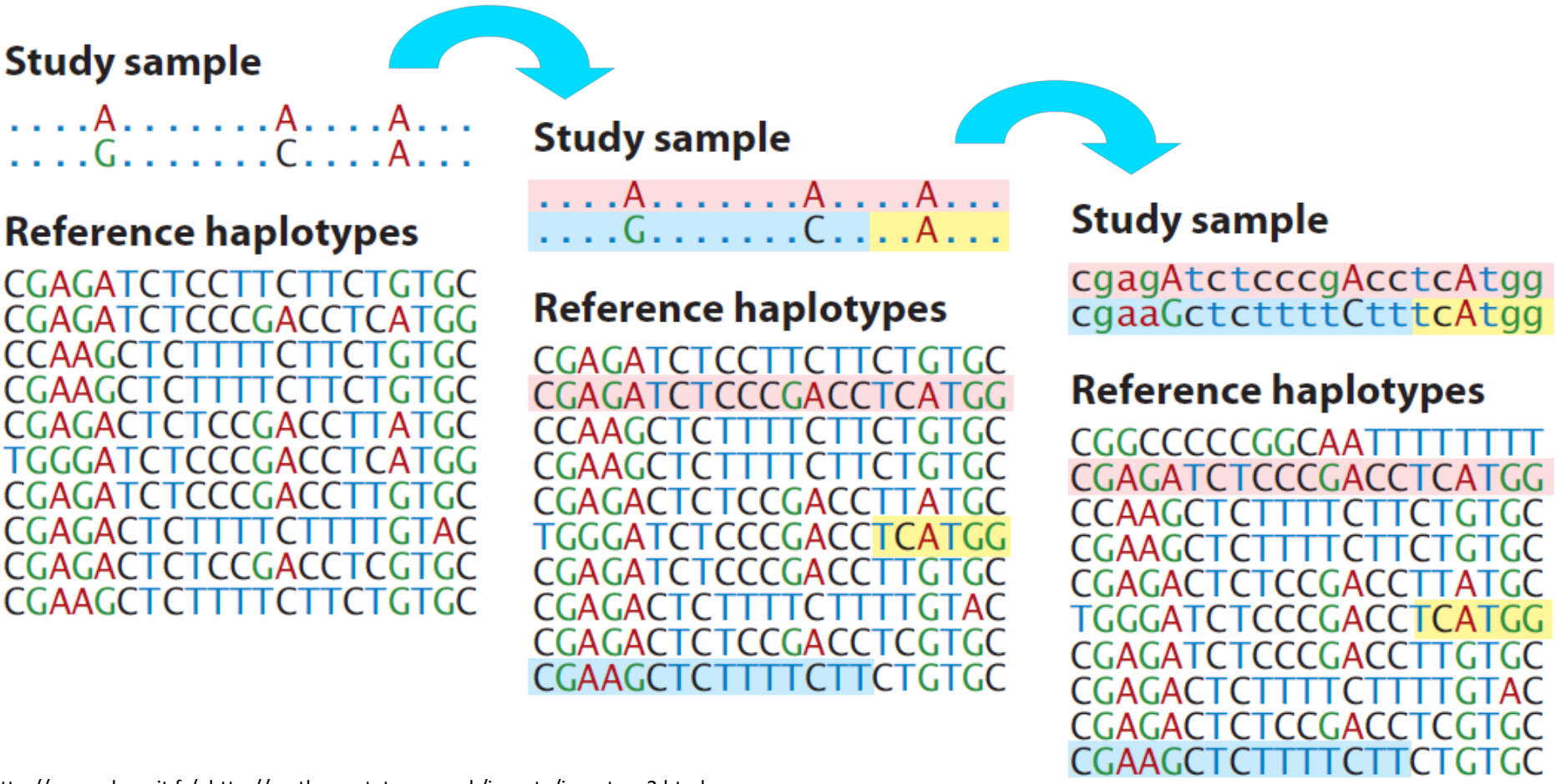
How would you fill in the remaining genotypes of this study sample individual, given the reference haplotypes?

- 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

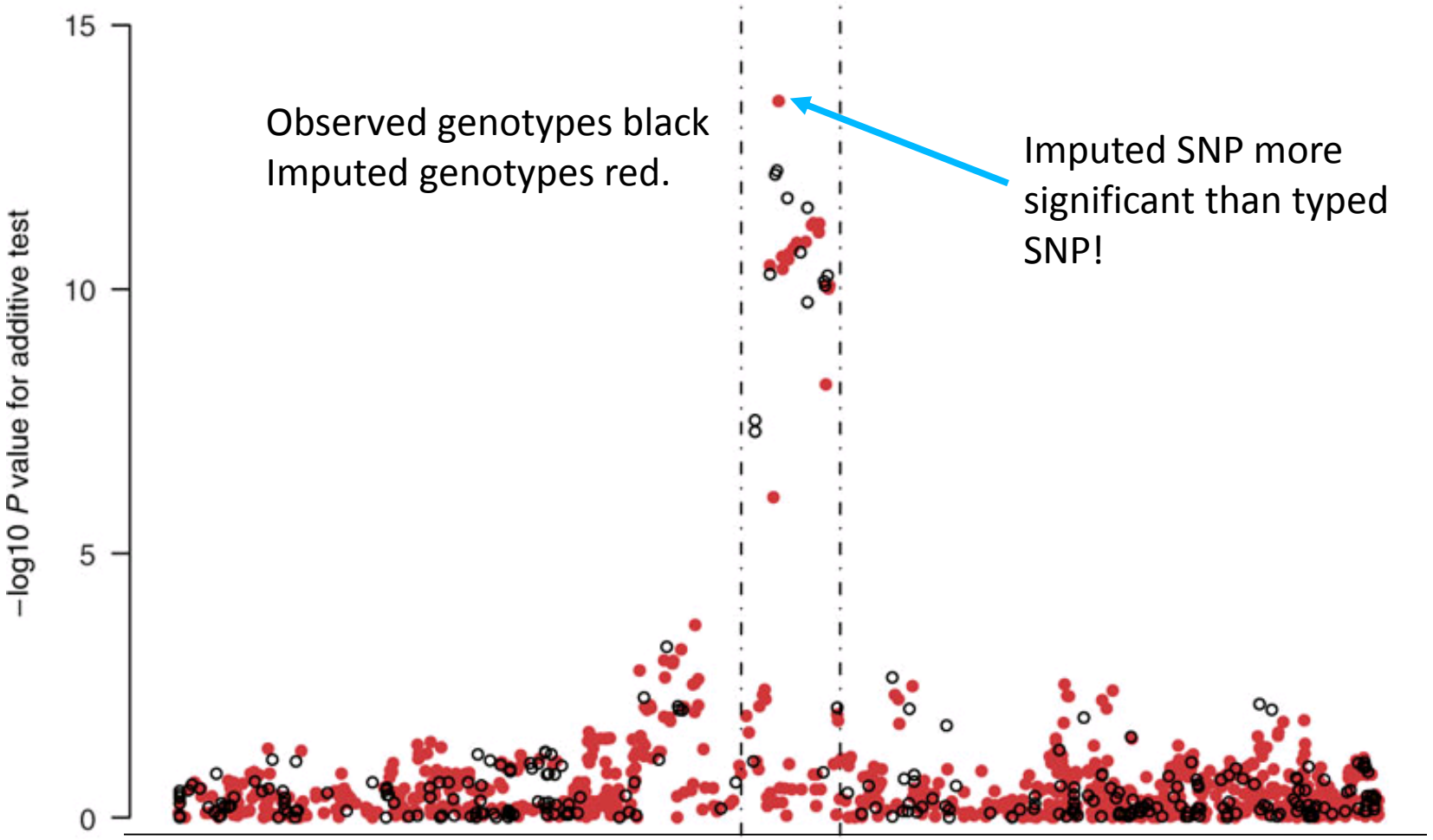


- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

Li et al., Ann Rev Human Genet, 2009

Imputation Application

TCF7L2 gene region & T2D from the WTCCC data



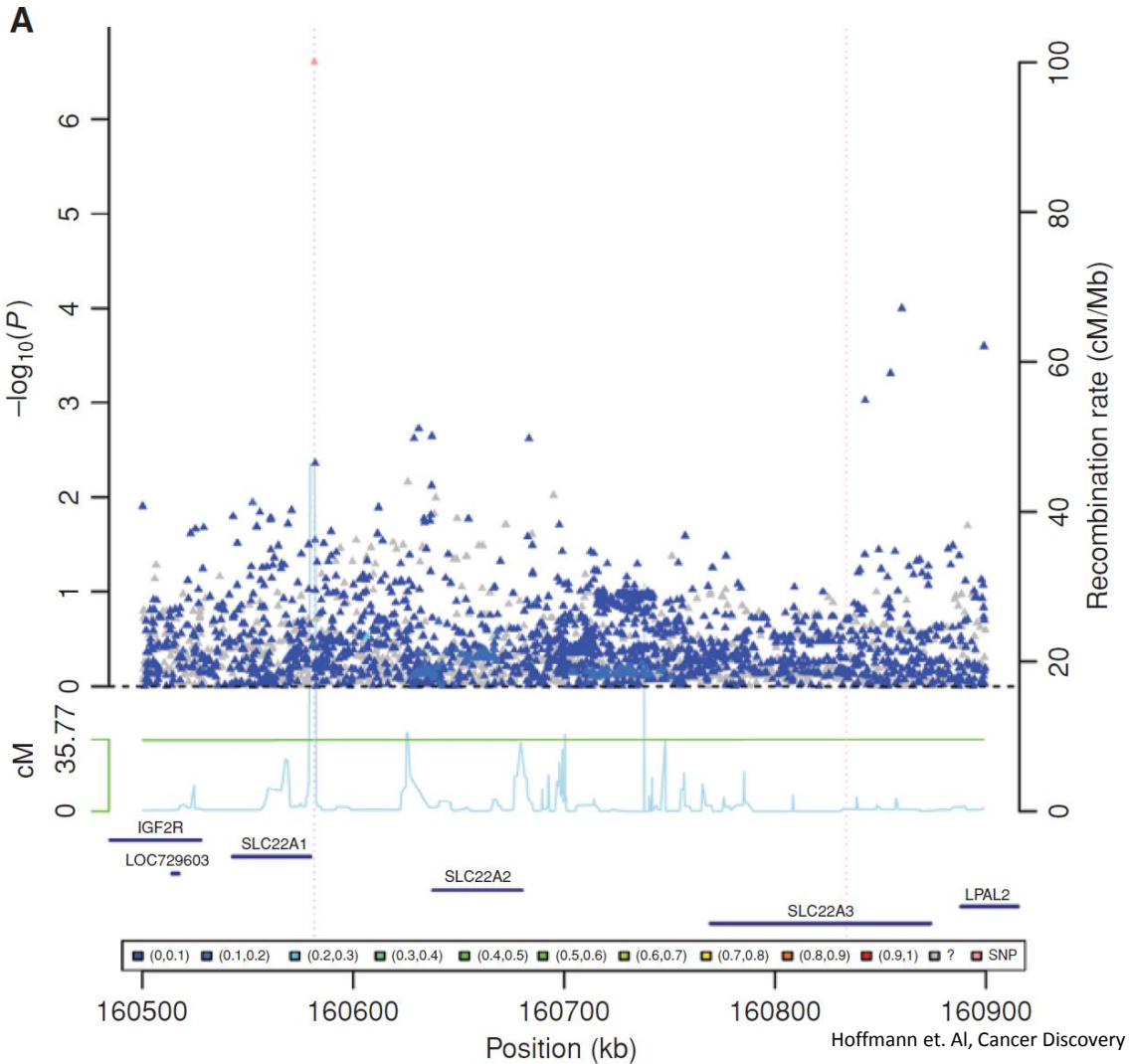
Marchini Nature Genetics2007

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

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

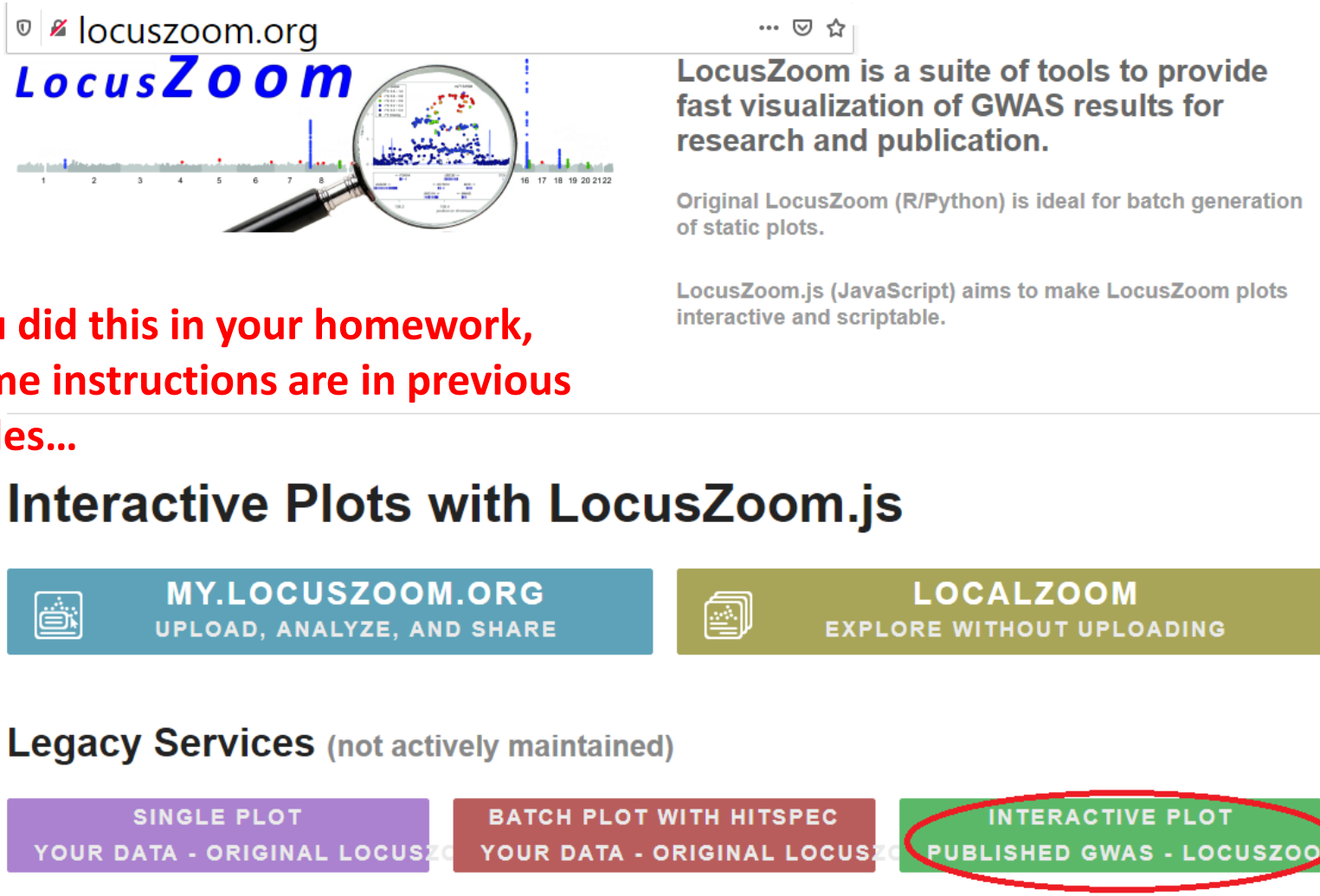
Figure 2. Regional fixed-effects meta-analysis plots from GWAS in KP population of the two risk variants that replicated: (**A**) 6q25.3 (rs4646284), (**B**) 19q13.33 (rs2659124). The color code for the points represents the r^2 of each SNP with the risk variant (ranges defined in the legend). The dotted 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



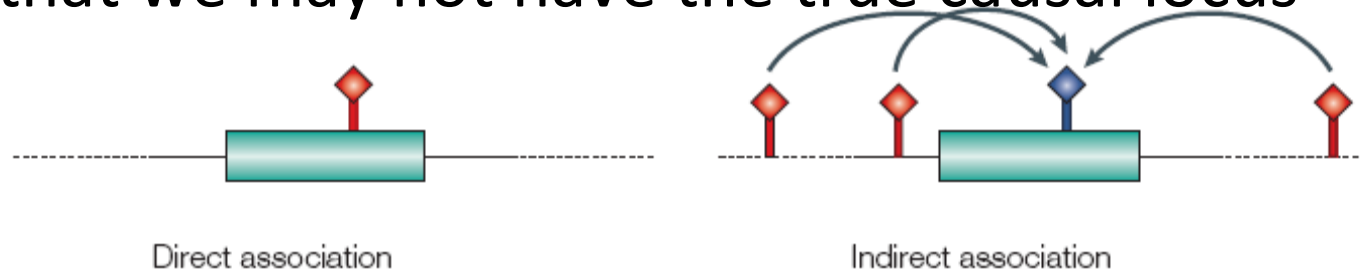
You did this in your homework, some instructions are in previous slides…

Outline for GWAS

- Imputation
- Following up on hits
- eQTLs & Transcriptome-wide association studies (TWAS)
- Limitations, heritability, and “missing heritability”, Polygenic risk scores

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

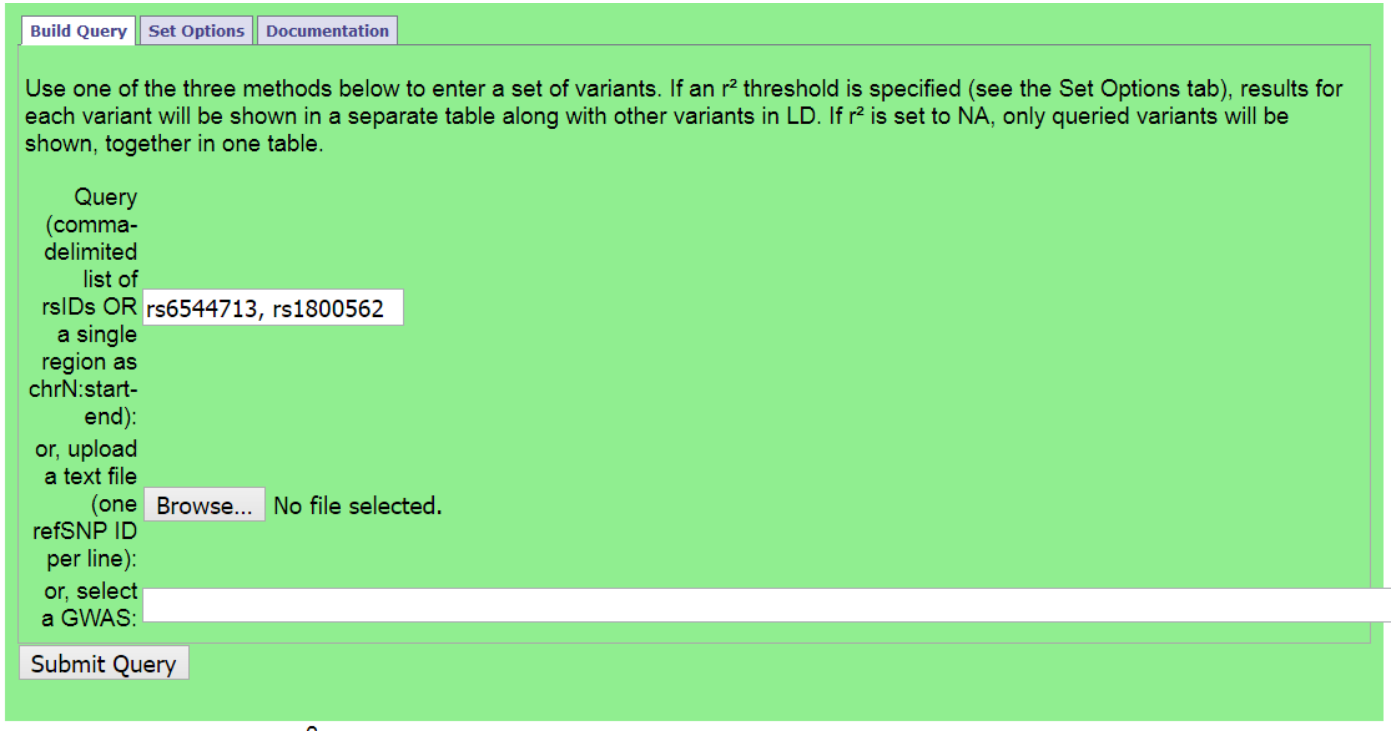
- Recall that we may not have the true causal locus



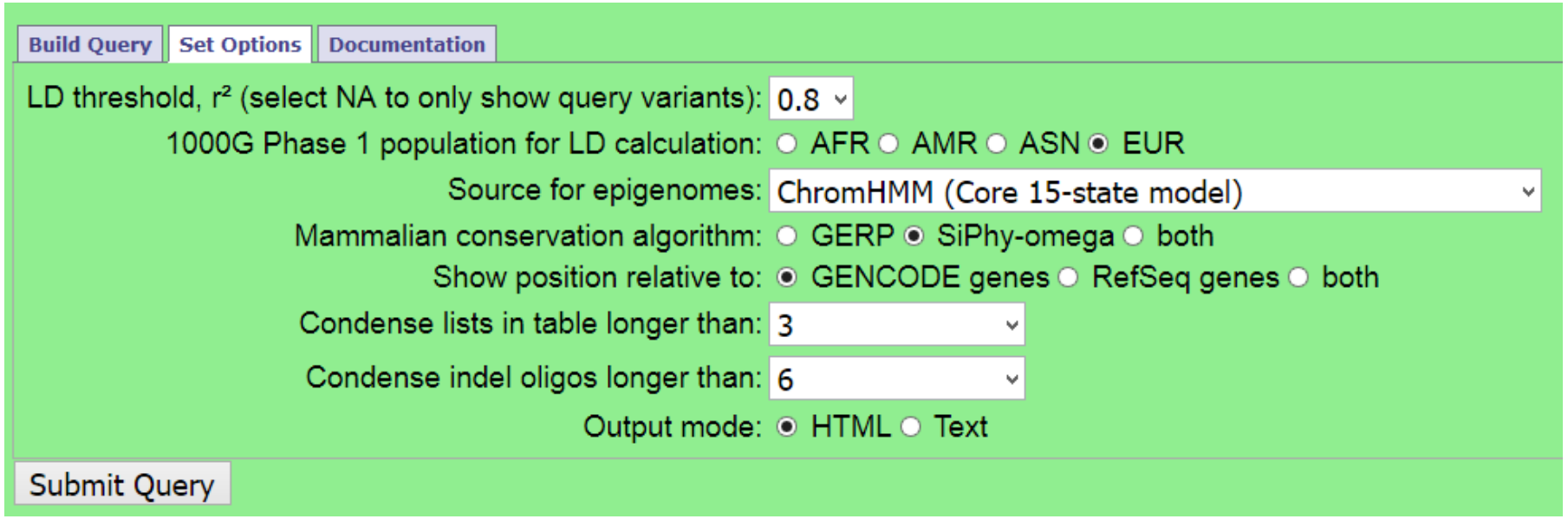
- Fine mapping vs. imputed genotypes…
- Tie-in to biology
- A lot of work has been done to characterize SNPs and prioritize what should be followed up on
- Functional studies

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:

(Talk about eQTLs later today)

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?

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

(Talk about eQTLs later today)

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	

Maybe?

- 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

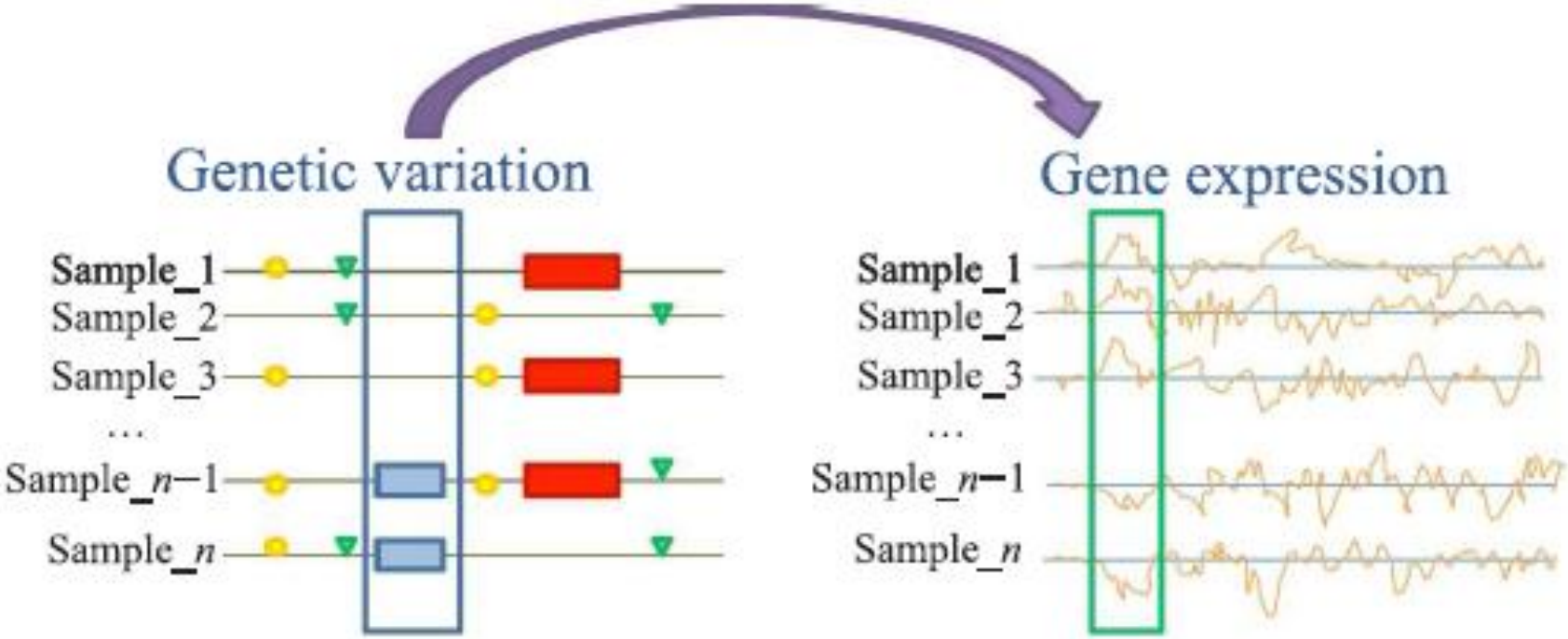
chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
2	43845437	0.96	0.99	rs4299376	G	T	0.85	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

For later… note that multiple variants are an eQTL…

Outline for GWAS

- Imputation
- Following up on hits
- eQTLs & Transcriptome-wide association studies (TWAS)
- Limitations, heritability, and “missing heritability”, Polygenic risk scores

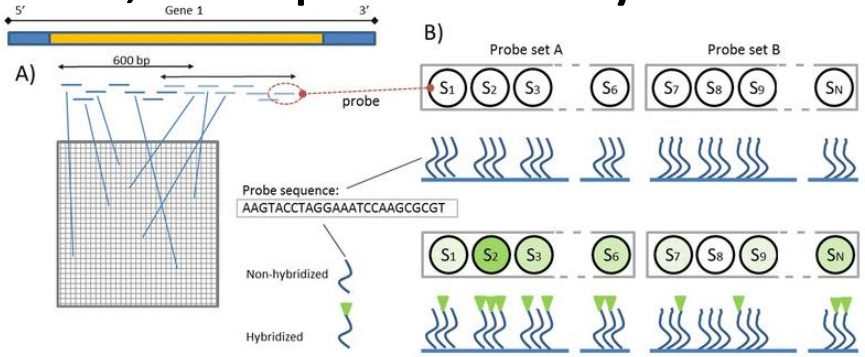
What is an eQTL?



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

Measuring the expression of a gene

- Note: *by tissue*
- Reverse-transcribe RNA (cDNA, complementary DNA), then
  - Microarray
    - Multiple fixed probesets
  - RNA Seq
    - More expensive but…
    - Better dynamic range
    - Better at picking up low expression
- A lot of QC, not enough time to cover



Microarray schematics. **a** Probes corresponding to the characteristic fragments of a given gene are placed in different locations across the array; **b** Single probes are arranged in sets corresponding to the same region of the gene. DNA that hybridizes to the probe can be detected using a fluorescent reporter system. Increasing the number of probes to which cDNA hybridizes correctly, increases the contrast between this probe set (*probe set A*) and any other probe set (*probe set B*)

Jaksik et al. Microarray experiments and factors wihich affect their reliability.

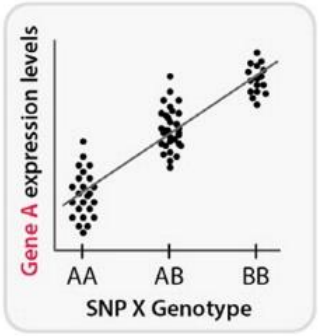
What is an eQTL?

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



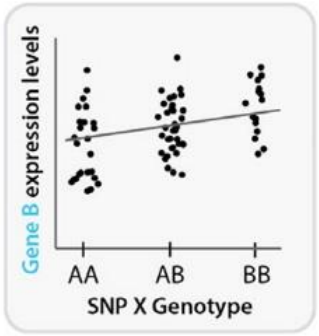
Altered **Protein A** levels, effect on the binding to the transcription factor binding sites of downstream genes

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?

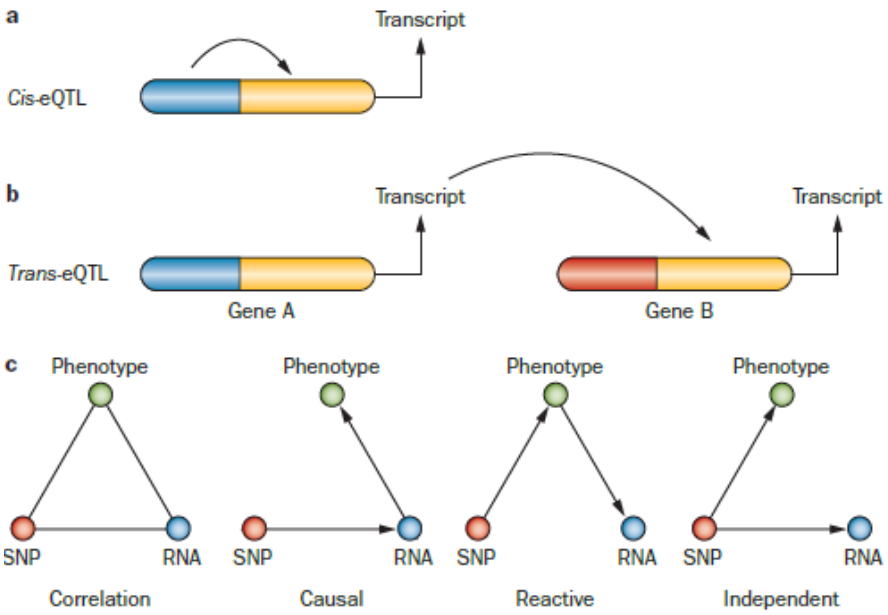
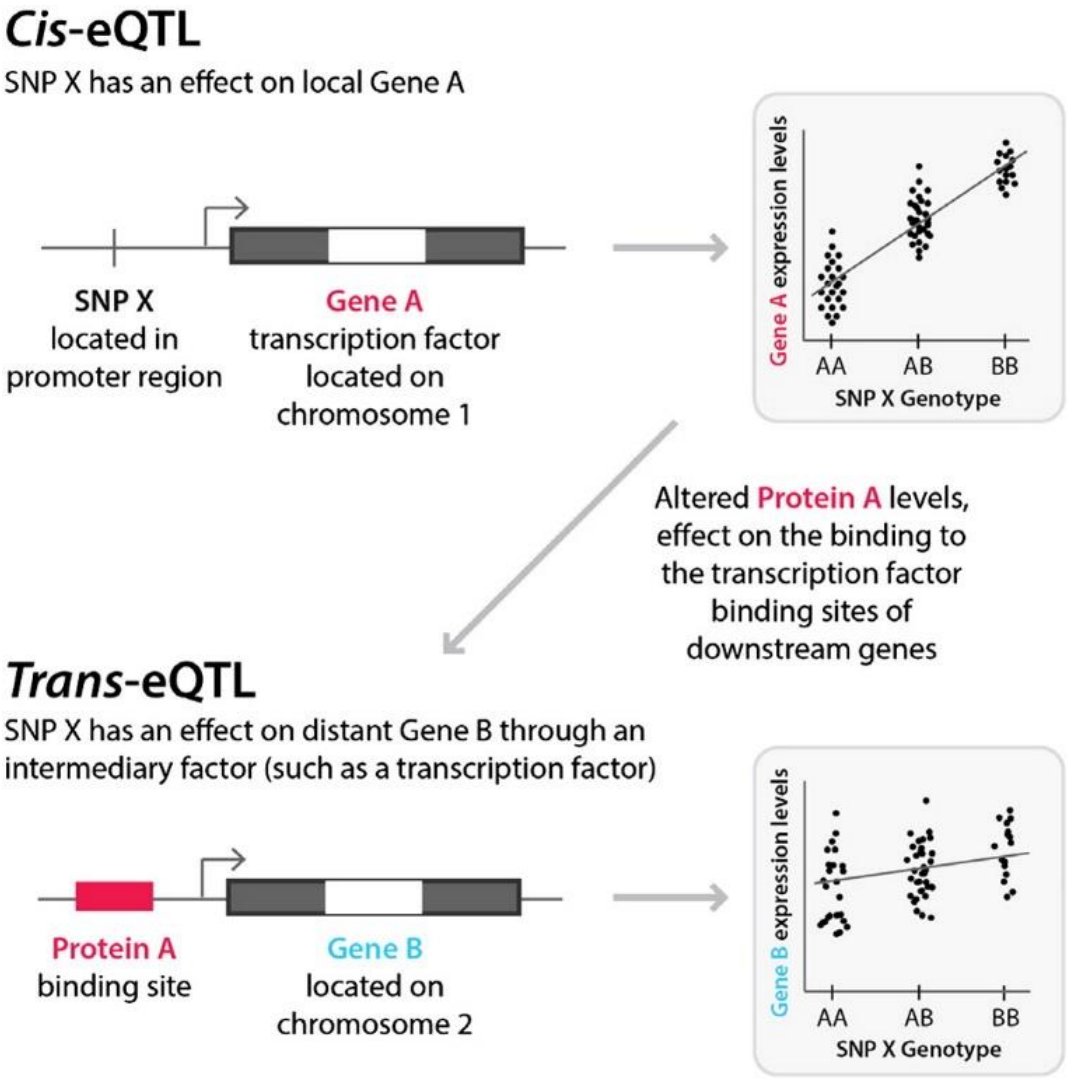


Figure 2 | Principles of eQTL analysis. Loci that control transcript levels are referred to as eQTL. Typically, eQTL analysis of humans or mice reveals thousands of eQTL, which are classified as either 'local' or 'distant', depending on the distance of the locus from the gene that an eQTL regulates. **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?

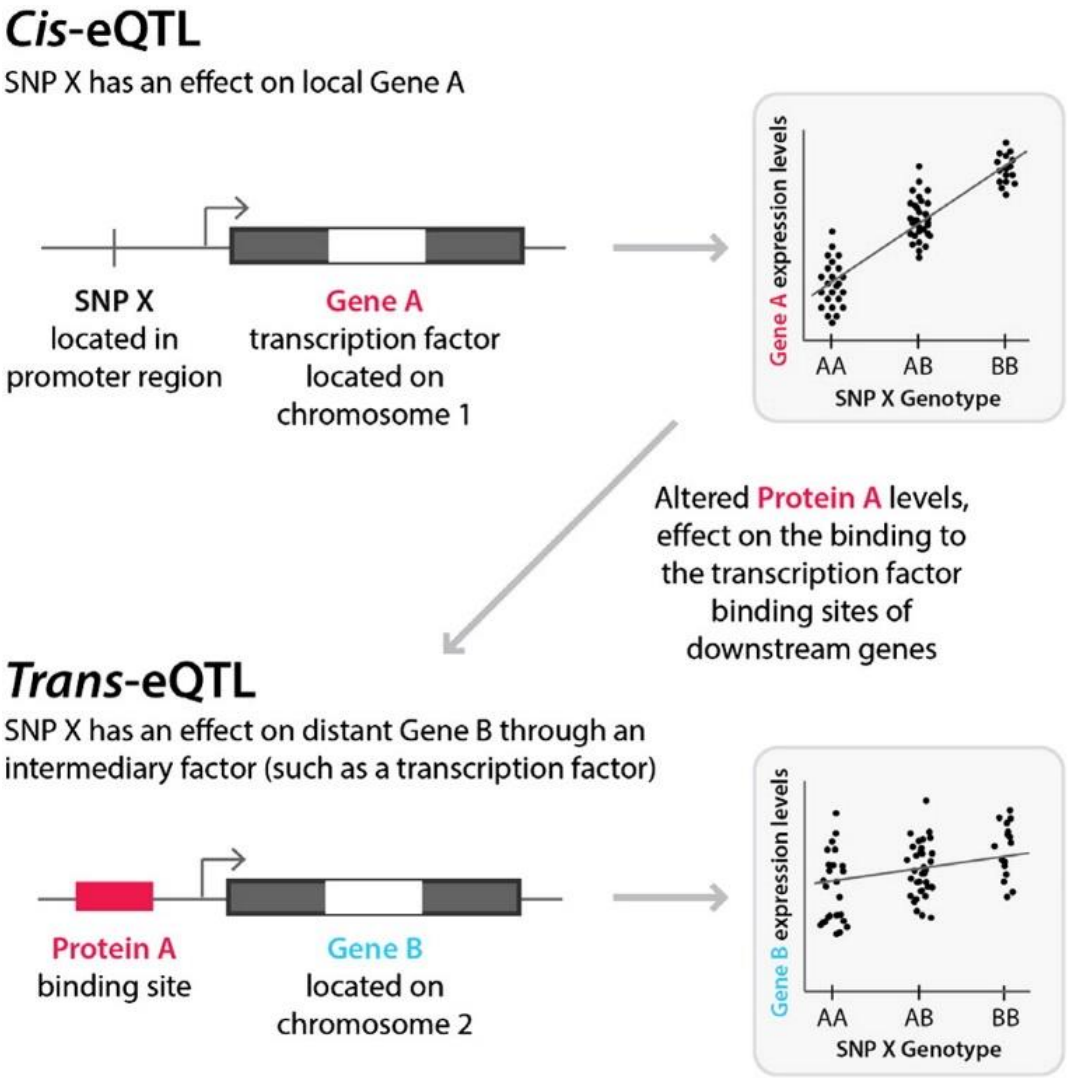


<u>CLASS EXERCISE</u>

- 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?



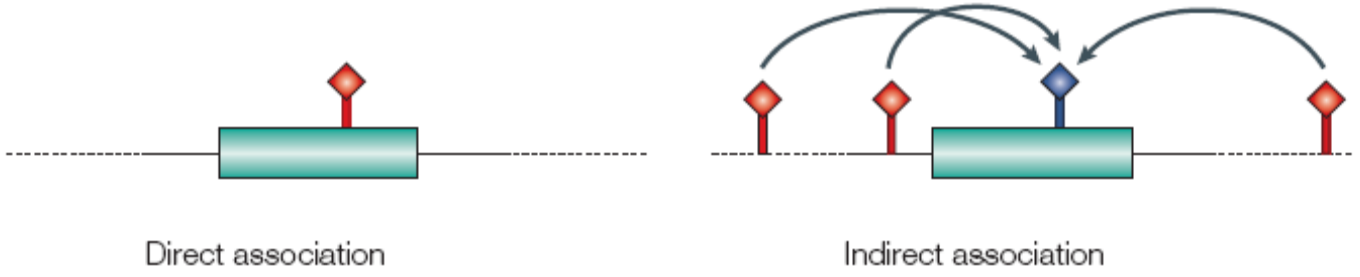
CLASS EXERCISE

- 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? LD?
  - 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?

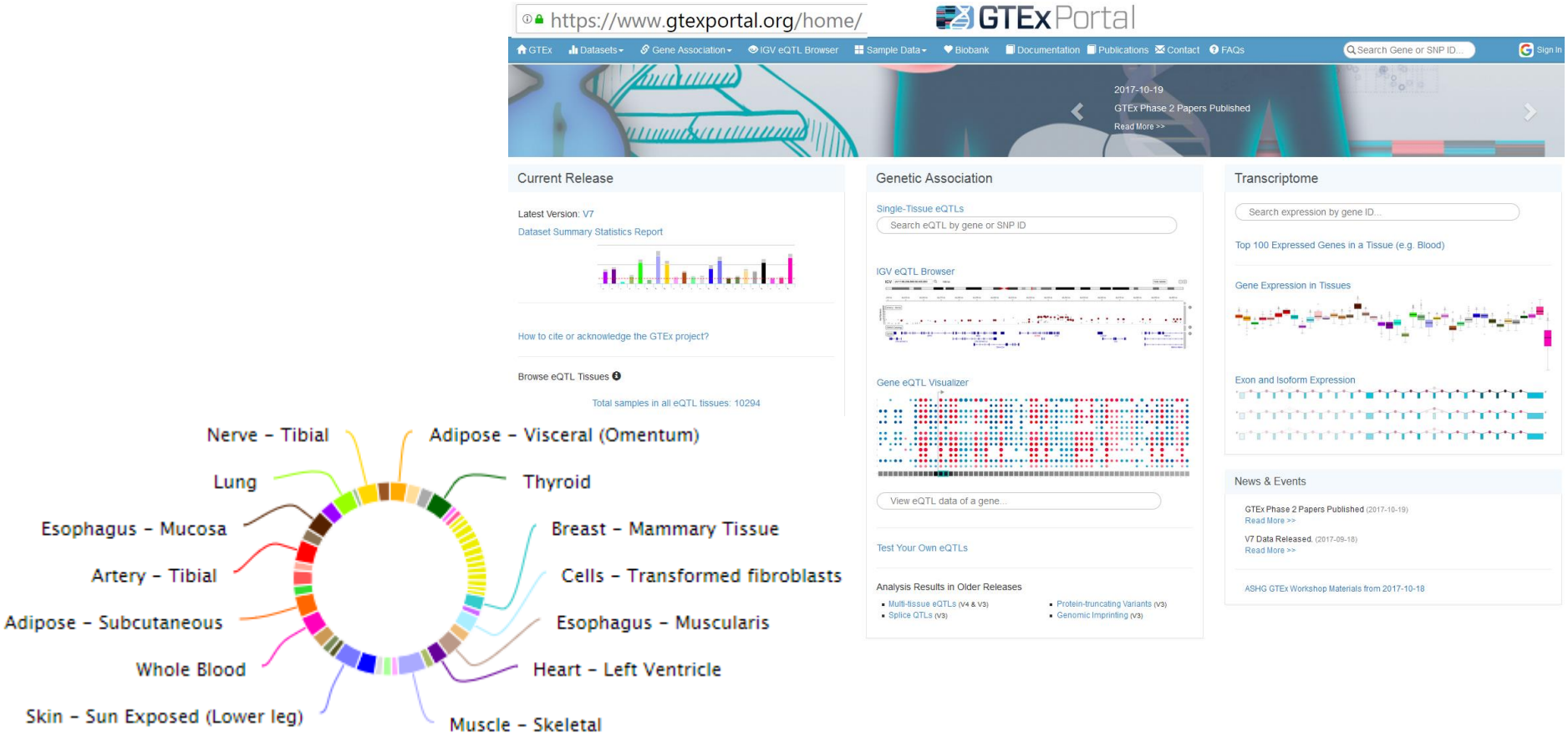
- The expression of a gene can be viewed as a quantitative trait.
- Test the SNPs around the gene for association with the expression of the gene.
- Does the SNP affect expression?
- ***The same problem still applies, don't know causal locus!***



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

cis eQTLs: GTeX

- Database of SNP associations with gene expression in humans



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?

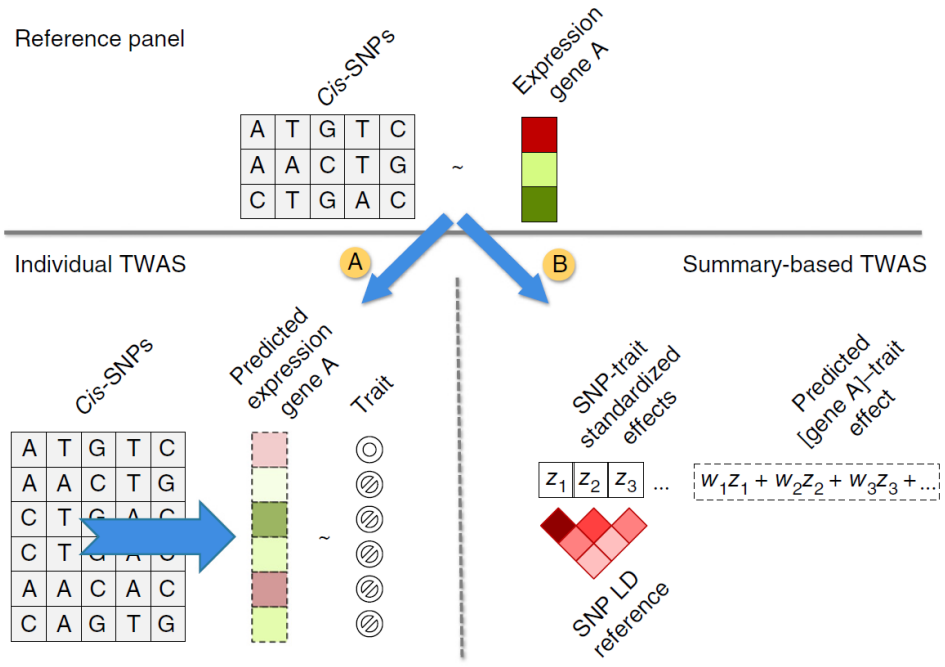
Target population

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

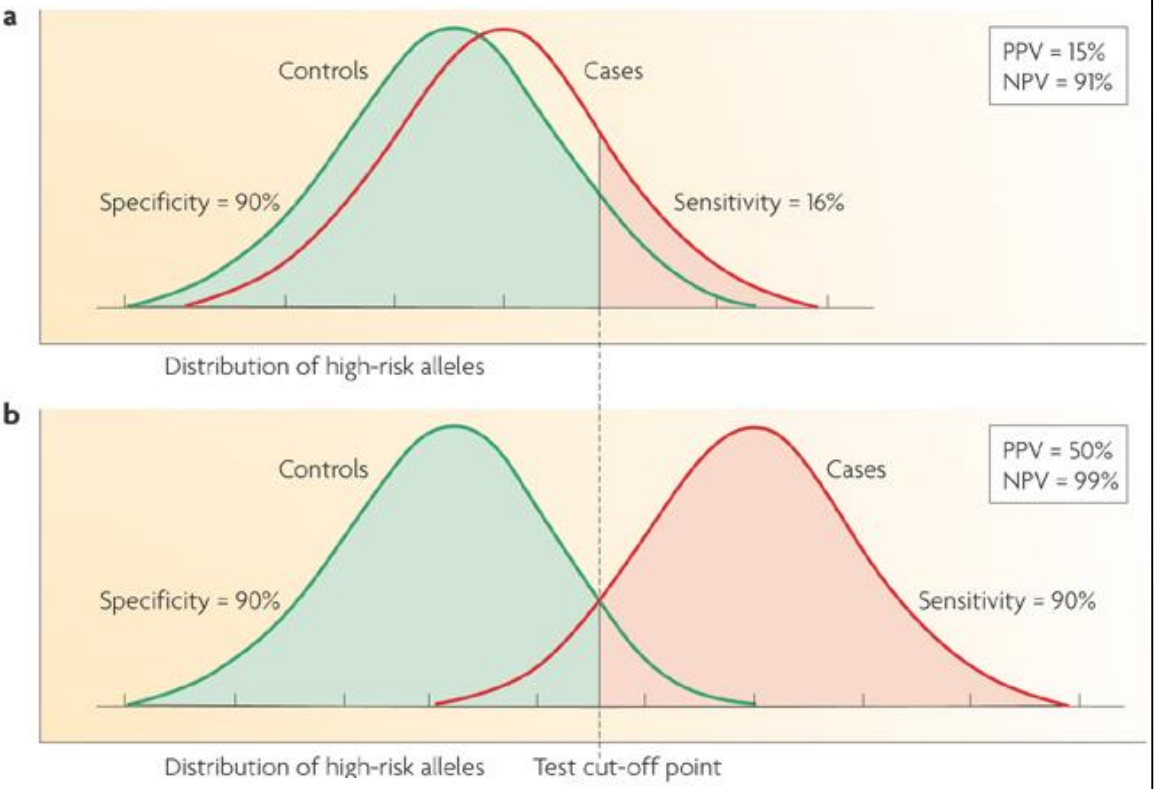
GWAS	Training expression	Gene	Chr.	Locus start	Locus end	*P* value			
						TWAS	Permuted	Best *cis*-SNP	
LOCKE.BMI	Omnibus	*INO80E*	16	29,506,615	30,517,114	3×10^{-9}	3×10^{-7}	3×10^{-6}	*
LOCKE.BMI	Omnibus	*FTSJ3*	17	61,396,793	62,407,372	1×10^{-7}	5×10^{-6}	9×10^{-7}	*
LOCKE.BMI	Omnibus	*PAM*	5	101,589,685	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}	*

Outline for GWAS

- Imputation
- Following up on hits
- eQTLs & Transcriptome-wide association studies (TWAS)
- Limitations, heritability, and “missing heritability”, Polygenic risk scores

Limitations of GWAS

- Not very predictive



Witte, Nat Rev Genet 2009

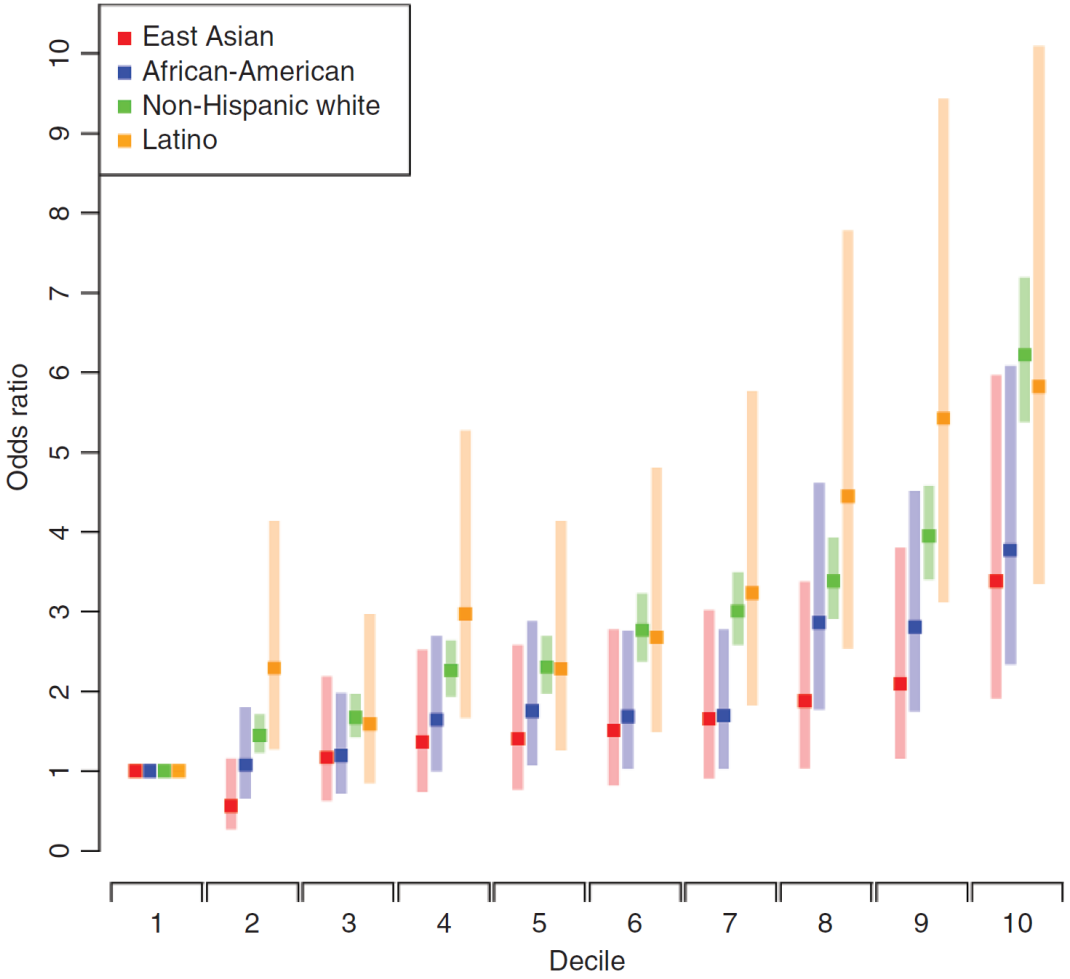
Example:
AUC for Breast Cancer Risk

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

Genetic Risk Score (GRS) – Overall genetic risk; tails show more sep.?

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

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.

Revised height variance explained (really not very predictive, or?)

Hum Mol Genet. 2018 Oct 15;27(20):3641-3649. doi: 10.1093/hmg/ddy271.

Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry.

Yengo L[1], Sidorenko J[1,2], Kemper KE[1], Zheng Z[1], Wood AR[3], Weedon MN[3], Frayling TM[3], Hirschhorn J[4], Yang J[1,5], Visscher PM[1,5]; GIANT Consortium.

Author information

Abstract

Recent genome-wide association studies (GWAS) of height and body mass index (BMI) in ~250000 European participants have led to the discovery of ~700 and ~100 nearly independent single nucleotide polymorphisms (SNPs) associated with these traits, respectively. Here we combine summary statistics from those two studies with GWAS of height and BMI performed in ~450000 UK Biobank participants of European ancestry. Overall, our combined GWAS meta-analysis reaches N ~700000 individuals and substantially increases the number of GWAS signals associated with these traits. We identified 3290 and 941 near-independent SNPs associated with height and BMI, respectively (at a revised genome-wide significance threshold of $P < 1 \times 10\text{-}8$), including 1185 height-associated SNPs and 751 BMI-associated SNPs located within loci not previously identified by these two GWAS. The near-independent genome-wide significant SNPs explain ~24.6% of the variance of height and ~6.0% of the variance of BMI in an independent sample from the Health and Retirement Study (HRS). Correlations between polygenic scores based upon these SNPs with actual height and BMI in HRS participants were ~0.44 and ~0.22, respectively. From analyses of integrating GWAS and expression quantitative trait loci (eQTL) data by summary-data-based Mendelian randomization, we identified an enrichment of eQTLs among lead height and BMI signals, prioritizing 610 and 138 genes, respectively. Our study demonstrates that, as previously predicted, increasing GWAS sample sizes continues to deliver, by the discovery of new loci, increasing prediction accuracy and providing additional data to achieve deeper insight into complex trait biology. All summary statistics are made available for follow-up studies.

… or is it omnigenic? …

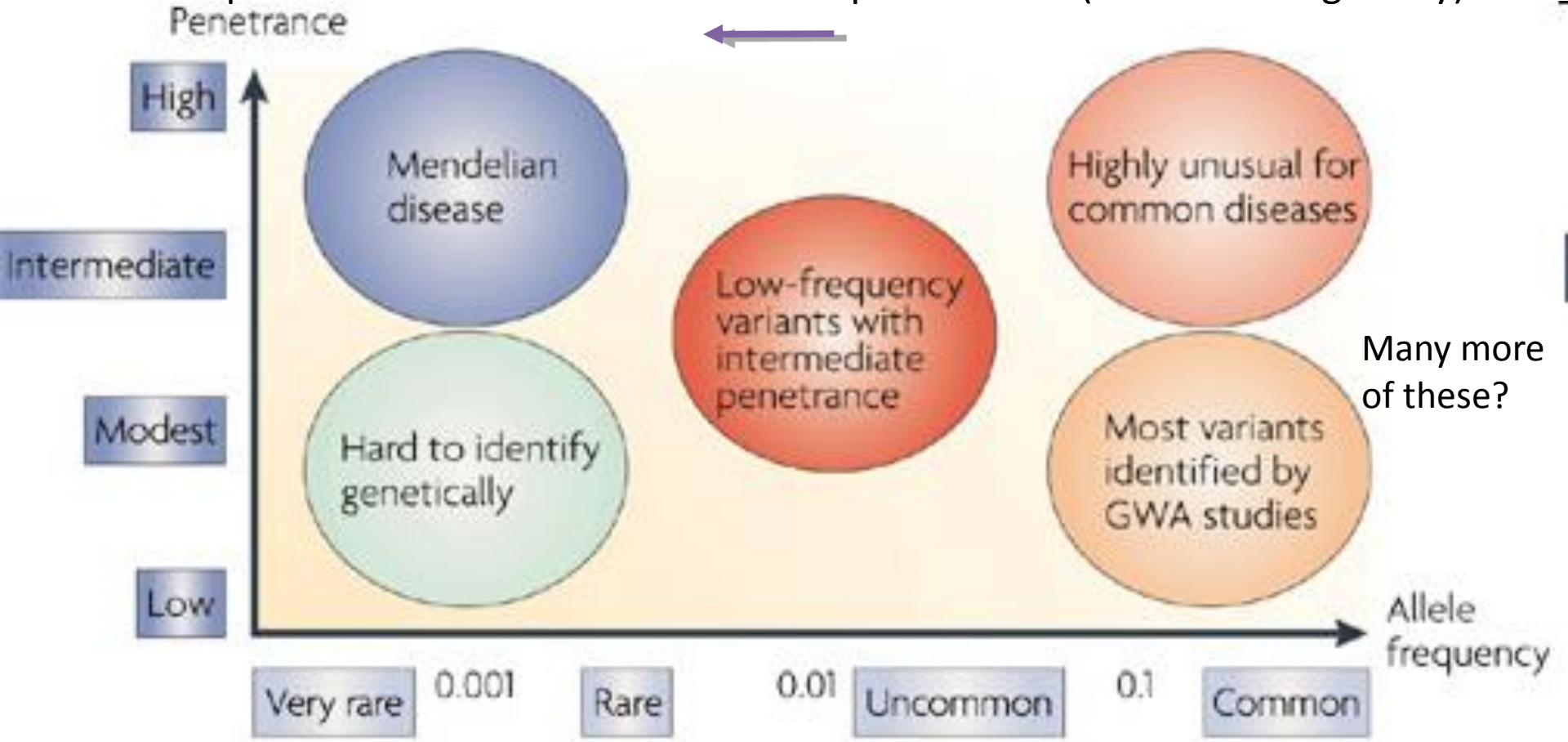
- Predict 100,000 variants with some effect on height…
- PRS variant that builds model based on *all* SNPS

An expanded view of complex traits: from polygenic to omnigenic

Evan A. Boyle,[1,†] Yang I. Li,[1,†] and Jonathan K. Pritchard[1,2,3,†]

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
   - Early 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. – *later lectures*

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

GCTA: http://www.complextraitgenomics.com/software/gcta/
LD Score (*can use summary statistics!*): https://github.com/bulik/ldsc

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.

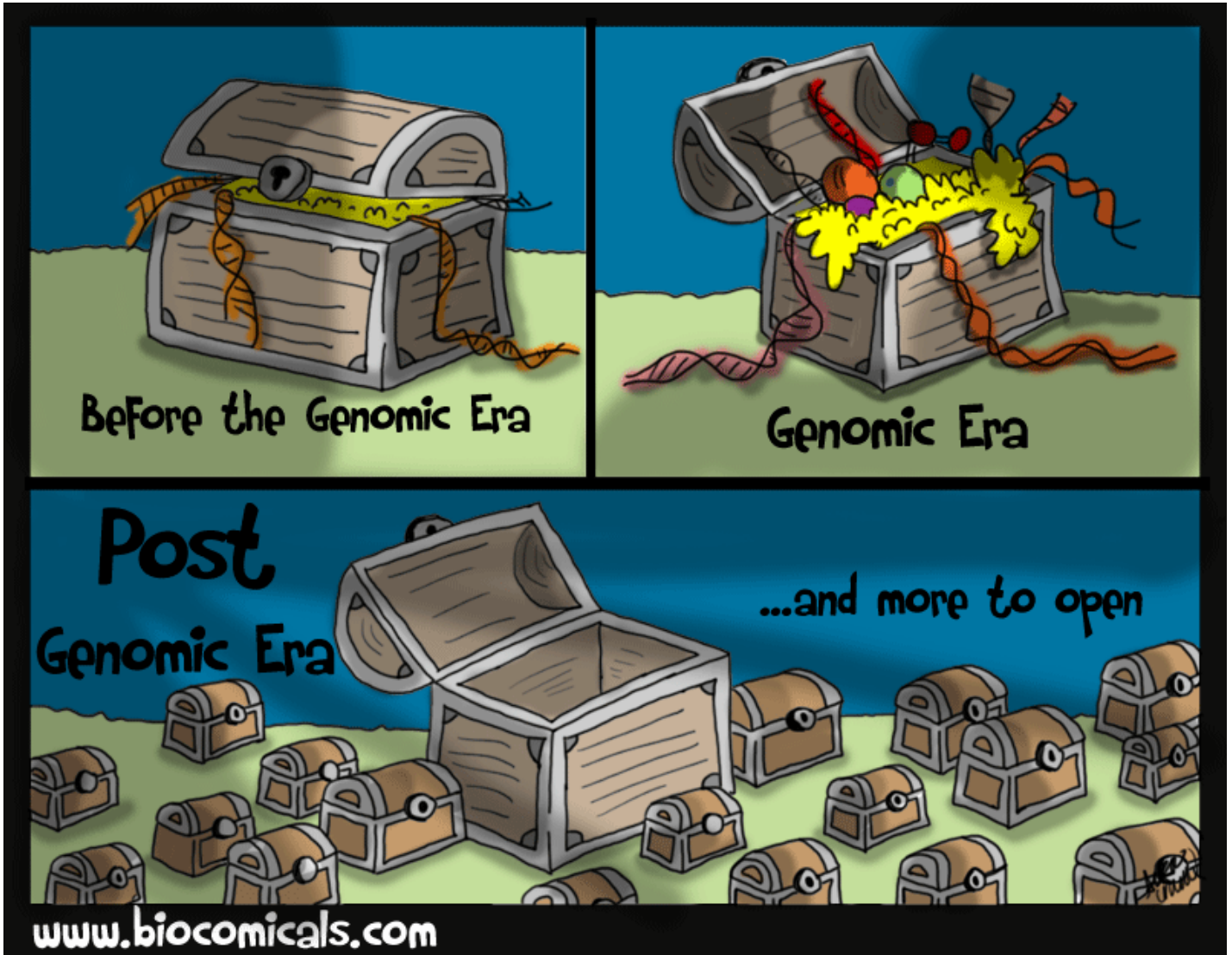
Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits[9,10]. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature 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

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)
- ...

Thanks!


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