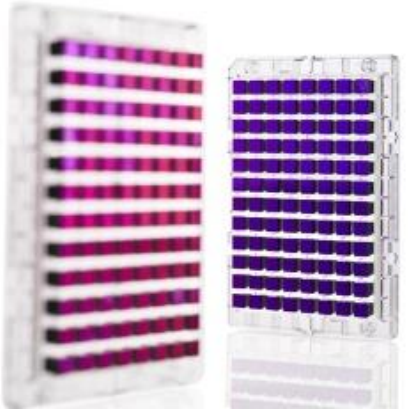
Genome-wide association studies (GWAS) – continued!



Outline for GWAS

- Analysis
  - QC
  - Analysis approach
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis
- GWAS Limitations

Why QC data?

nature medicine

Retraction | Published: 08 October 2019

Retraction Note: CCR5-Δ32 is deleterious in the homozygous state in humans

Xinzhu Wei & Rasmus Nielsen

Nature Medicine **25**, 1796(2019) | Cite this article

7795 Accesses | **2** Citations | **211** Altmetric | Metrics

The original article was published on 03 June 2019

Retraction to: *Nature Medicine* https://doi.org/10.1038/s41591-019-0459-6, Published online 3 June 2019.

The authors are retracting this Brief Communication to correct the scientific literature. Arising from exchanges with David Reich and colleagues, they have been made aware of a genotyping calling bias in the underlying UK Biobank data from which the main results of the study were drawn. Further analyses confirmed that the central finding of the study – that homozygous CCR5-Δ32 mutation is associated with increased mortality in the UK Biobank – is a result of this technical artifact. Because the main conclusion of

HIV resistance

QC Steps: Class exercise!

- Can you think about potential QC steps you might take at an individual and SNP level basis? Which one first?
- What tests have we talked about in this course?
- Special chromosomes?
- (Could families be problematic? useful?)

QC Steps

- "garbage rises to the top"
- Filter SNPs and Individuals
  - MAF (e.g., 1%, but very sample size dependent!)
  - Low call rates (means genotype probably didn't work correctly)

•Test for HWE among controls & within ethnic groups. Use conservative alpha-level (10^{-6}, e.g., but opinions vary) (again, to remove artifacts)

•Check for relatedness.

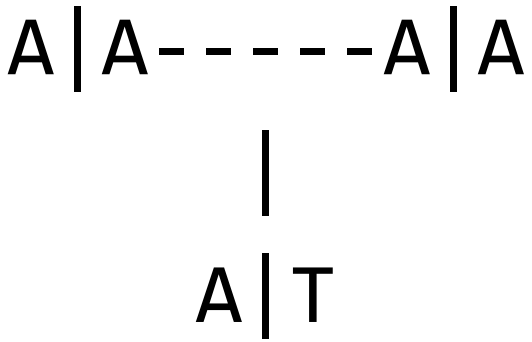
  - MZ twins, or accidentally genotyped same sample twice?
  - Remove first degree, or account for.

•Check genotype gender (mislabeled samples)

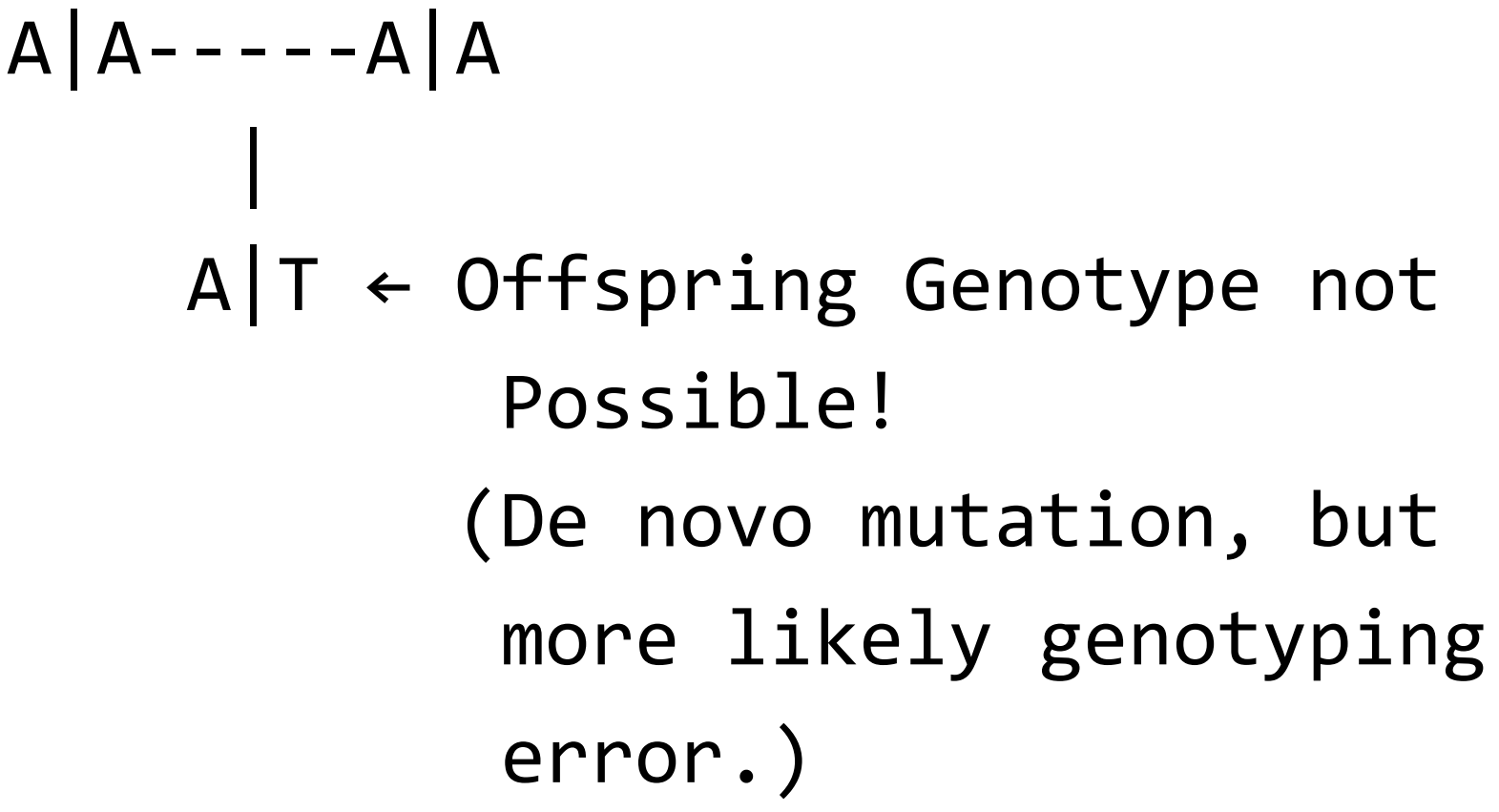
•Filter Mendelian inhertance (family-based, or potentially cryptics, if large enough sample)

•plink sofware: pngu.mgh.harvard.edu/~purcell/plink/reference.shtml

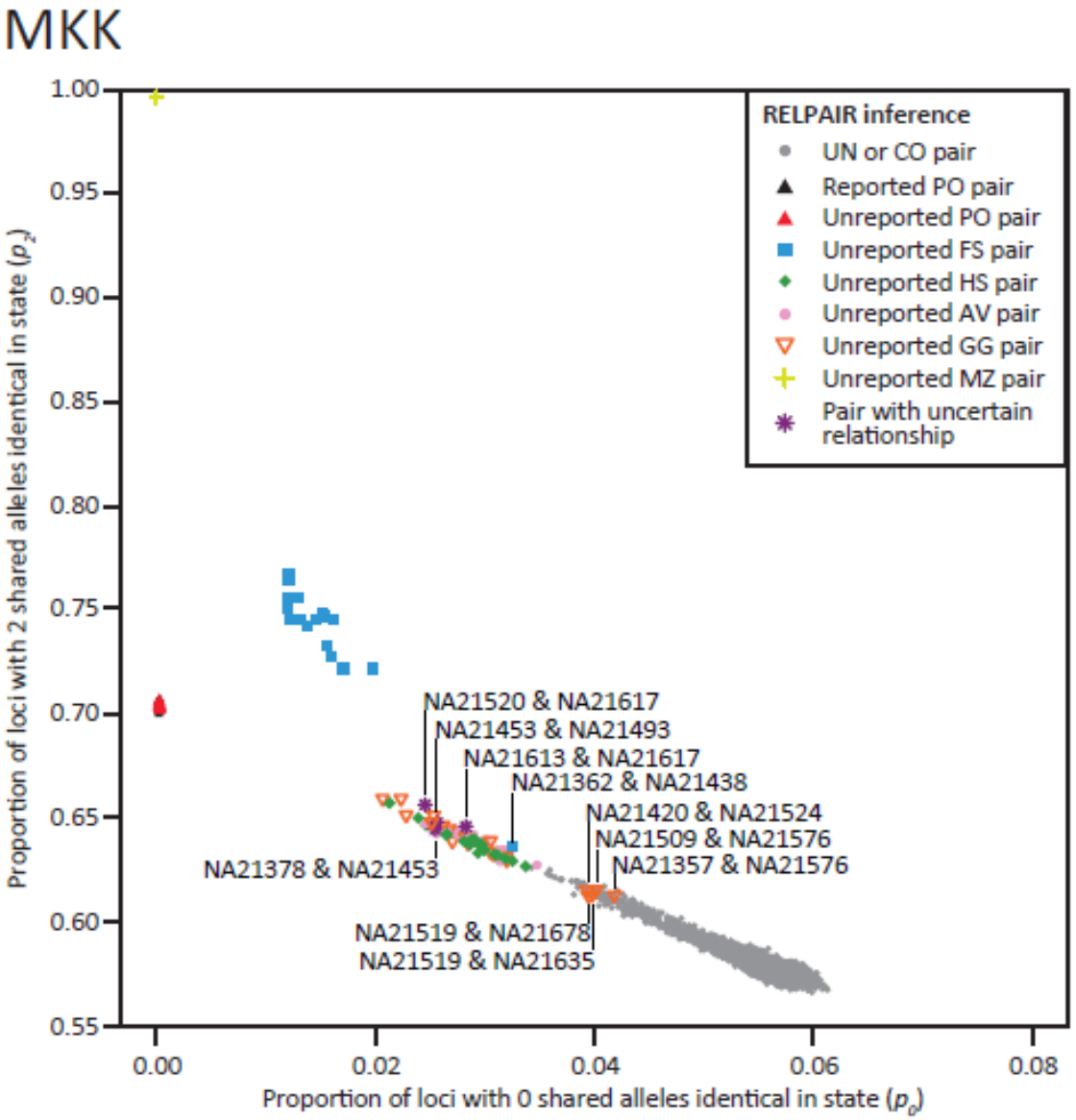
Filter for Mendelian inheritance



Filter for Mendelian inheritance



Check for relatedness, e.g., HapMap



- Take overall average of all SNPs of how many alleles are shared.
- E.g., parent-offspring never shares zero alleles.
- HapMap was supposed to be unrelateds (this was a bit of a surprise!)

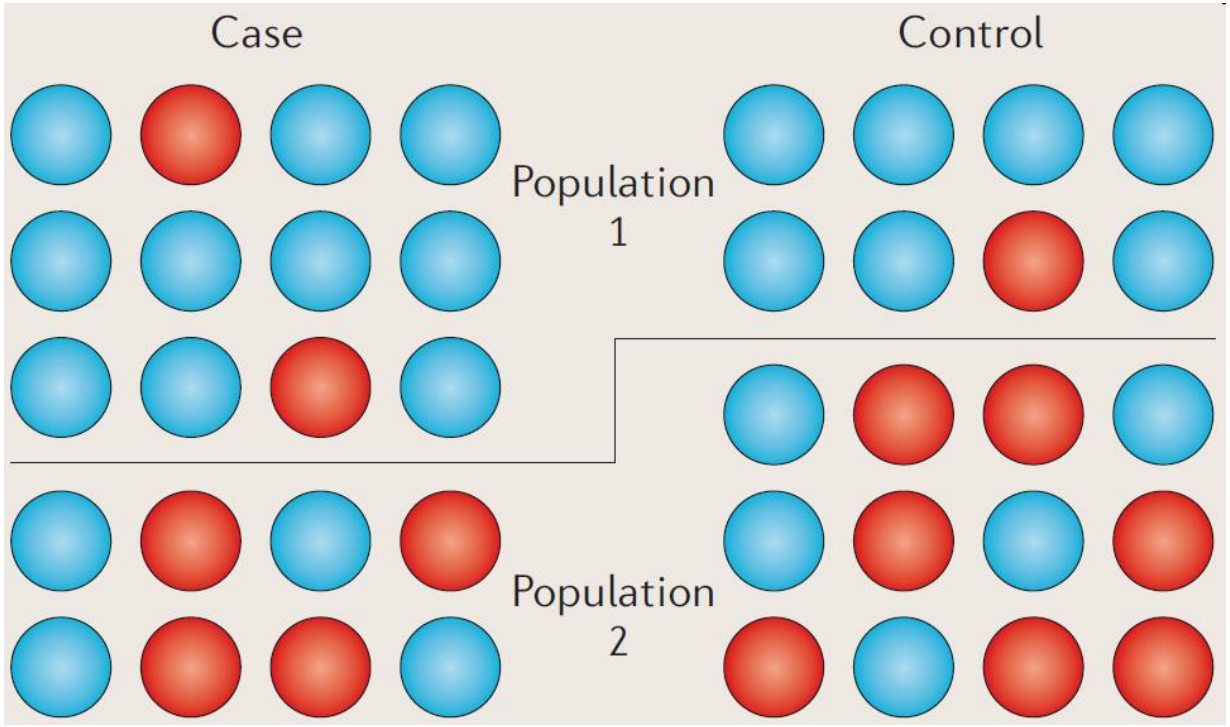
Pemberton et al., AJHG 2010

GWAS analysis

- Most common approach: Model each SNP separately
- Additive coding of SNP most common, just a covariate in a regression framework
- Dichotomous phenotype: logistic regression
- Continuous phenotype: linear regression
- Correct for multiple comparisons
  - e.g., Bonferroni, 1 million gives $\alpha=5x10^{-8}$
  - [Often use $5x10^{-8}$ for larger GWAS arrays (or imputed SNPs, more later) also, SNPs are correlated, so Bonferroni is a little too conservative…]
- Adjust for potential population stratification (details next slides…)
  - principal components (PC's), on best performing SNPs
  - software usually does LD filter (e.g., Eigensoft)

Recap of population substructure

- Two populations have different disease frequency, and different allele frequency (Recall Pima & Papago example).
- Association picks up they are different populations!



Balding, Nature Reviews Genetics 2010

How to handle population substructure?

- Any ideas?

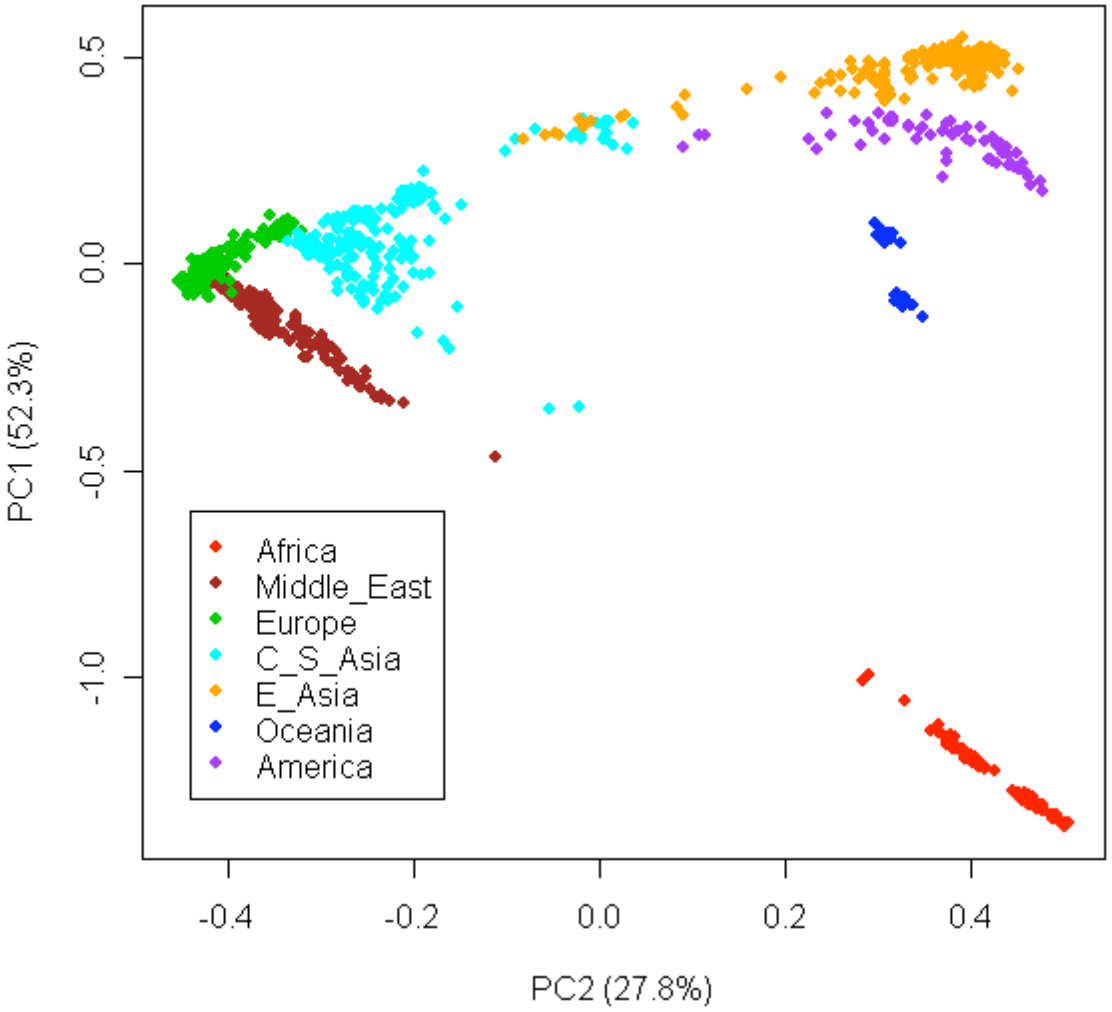
How to handle population substructure?

- Hint:
  - Majority of allele frequency differences are due to ancestry…?

How to handle population substructure?

- Homogeneous group?
  - But what if we don't want to?
- Indicator variable for each group?
  - Is that really practical?
- Majority of allele frequency differences are due to ancestry…?
  - Data reduction techniques…?
    - Some sort of continuous adjustment for how much ancestry…?

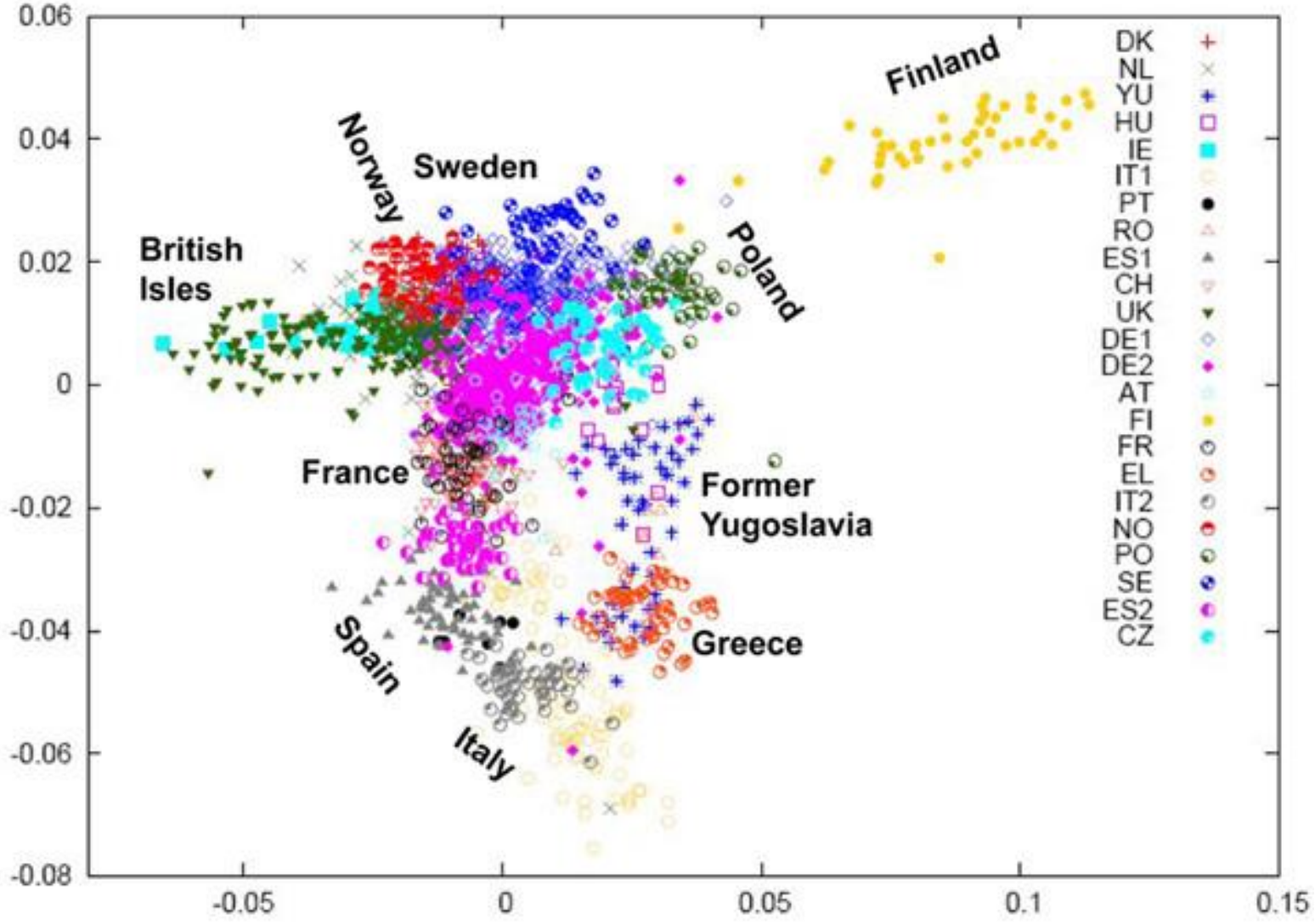
Control for race/ethnicity/ancestry: Principal Components (PCs) as covariates in regression model



- Principal component analysis is, in general, a data reduction technique
- Each PC is a linear combination of all SNPs
- First PC accounts for as much of the variability in the data as possible
- Subsequent PCs are orthogonal (uncorrelated) with previous
- In GWAS, it captures ancestry clines

- Li et al., Science 2008

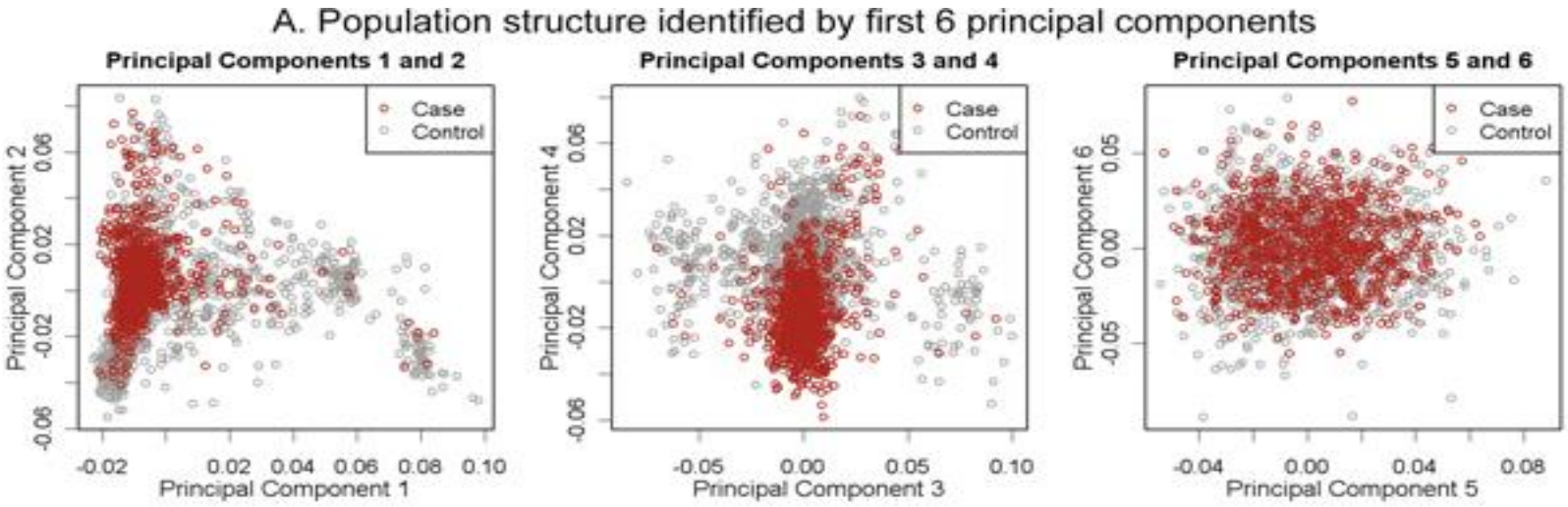
PCs pick up fine population structure

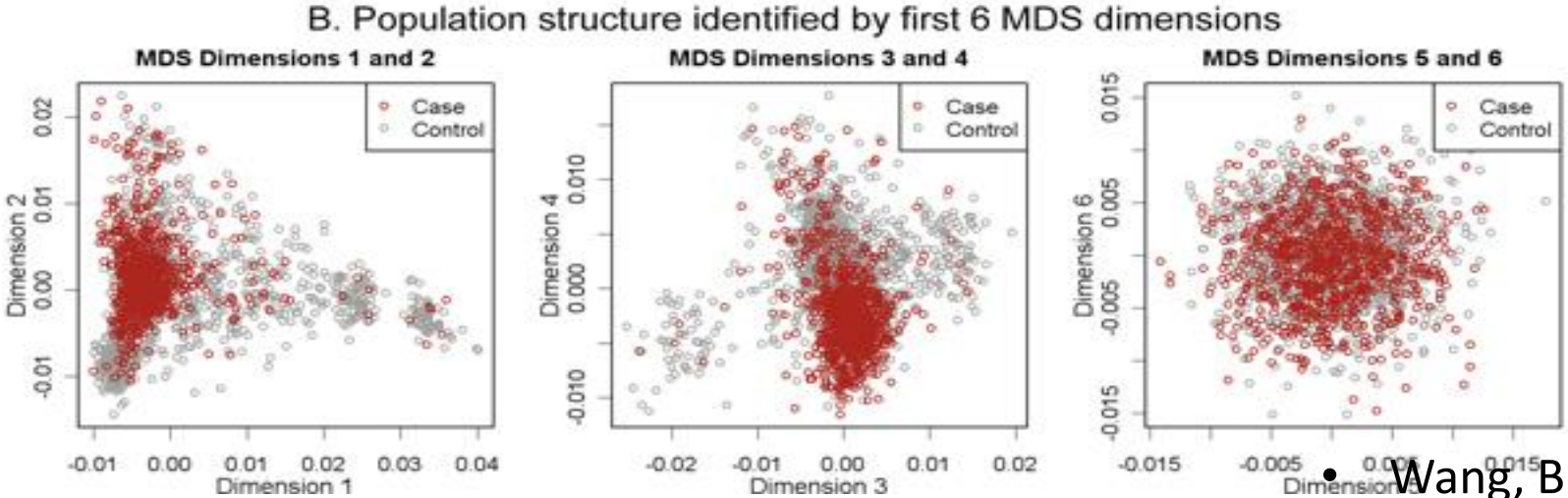


- Razib, Current Biology 2008

Adjusting for PC's

Do not want separate cluster for cases as controls! – Why want individuals from similar population for controls (and randomization). [These look reasonable, but if see a separation…]





- Wang, BMC Proc 2009

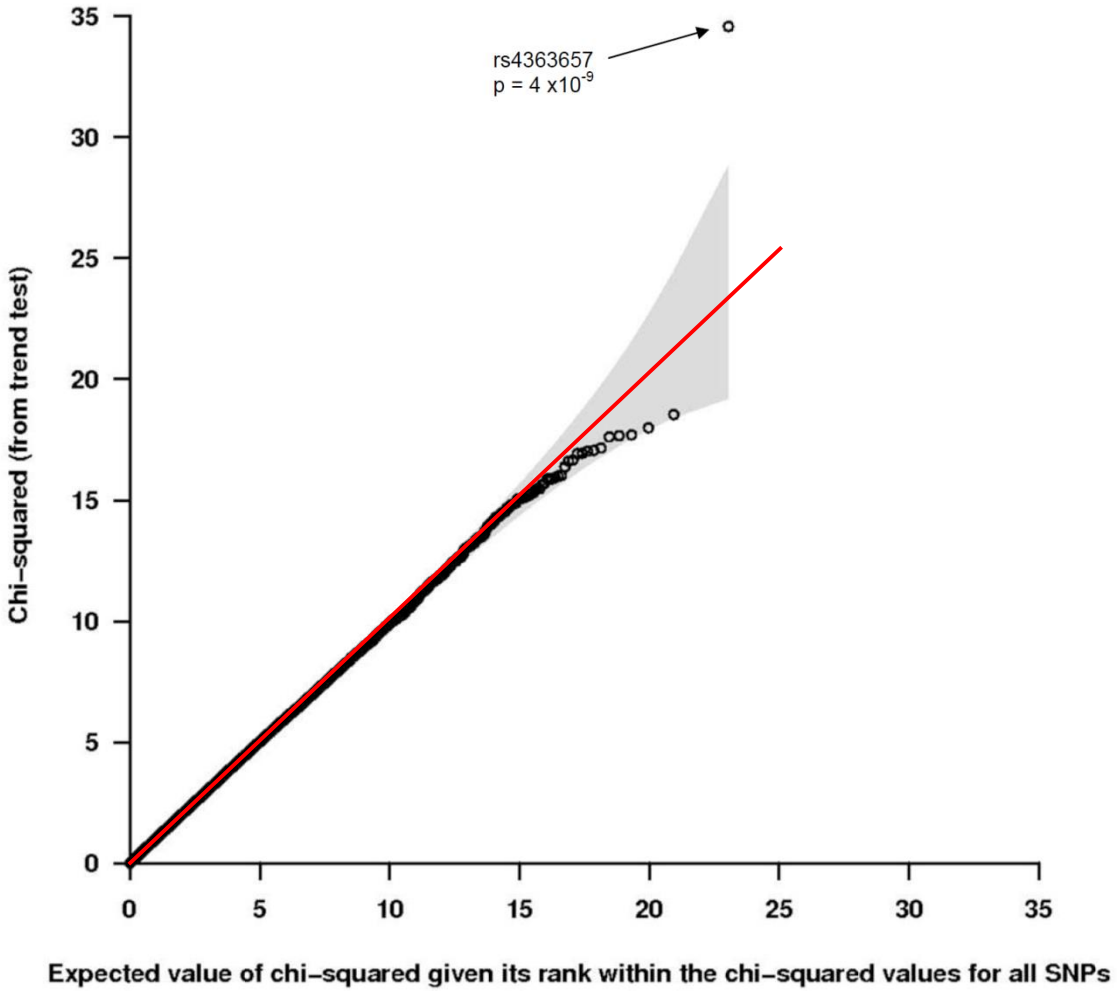
Instead of adjusting for PCs, adjust for the rest of the genome (mixed models)

- (1) Increase power by jointly modelling all variants, and (2) adjust for population structure
- Adjust for genetic relatedness between all pairs of individuals (kinship matrix)
  - This is huge, need approximations to handle this in modern cohorts.
  - In practice -- leave one chromosome out analysis
    - E.g., chromosome 1, adjust for kinship based on chromosome 2-22
    - Don’t really want to adjust for variants immediately nearby (LD, proximal contamination)
- Software: Emmax, GEMMA, BOLT-LMM (fast)

After run analysis

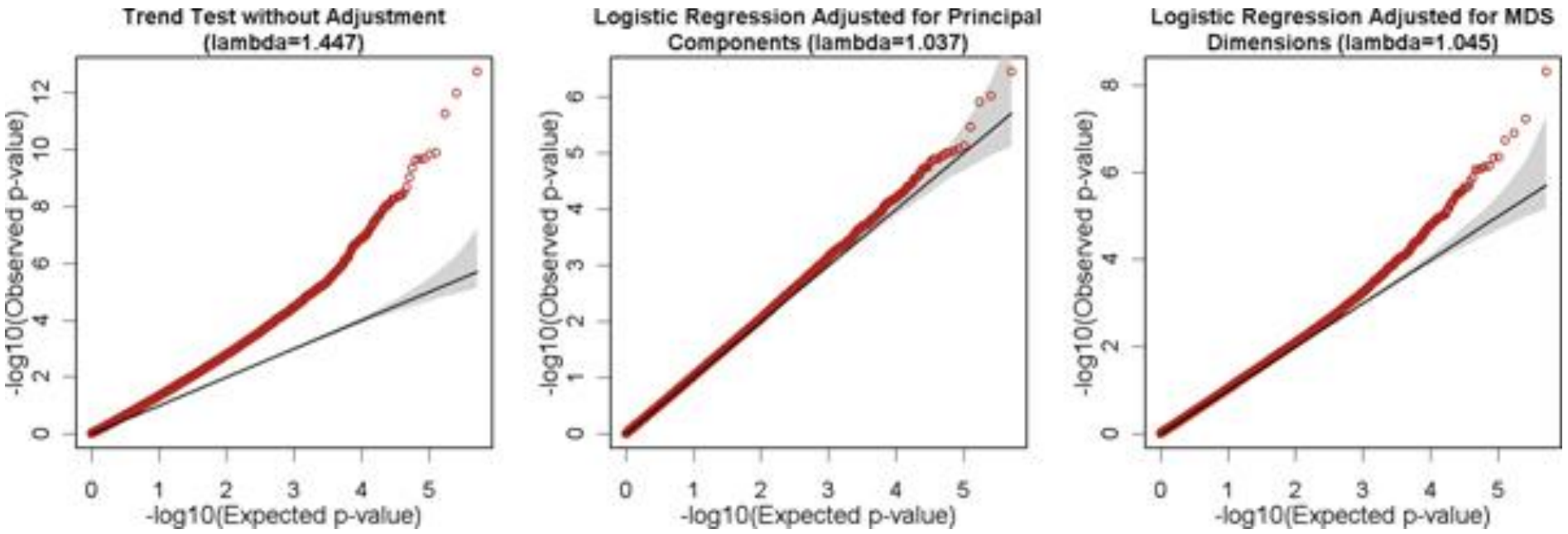
- Assess the fit with Q-Q plot
- Look at results in Manhattan plots
- **Replication** of hits in a separate cohort

Quantile-quantile (Q-Q) plot review



- Most SNPs are on the line (i.e., null), but want a few hits off the line (true significant associations!)
- This isn't really true any more with polygenic traits and large populations – so much of the genome *is* associated…

Q-Q plots and PC adjustment recap



- Exactly on line if no signal at all.
- If don't adjust for PC's, then p-values are all inflated artificially.
- Adjusting for PC's fixes the problem.

Wang, BMC Proc 2009

Manhattan plot example: Kaiser GWAS of Prostate Cancer

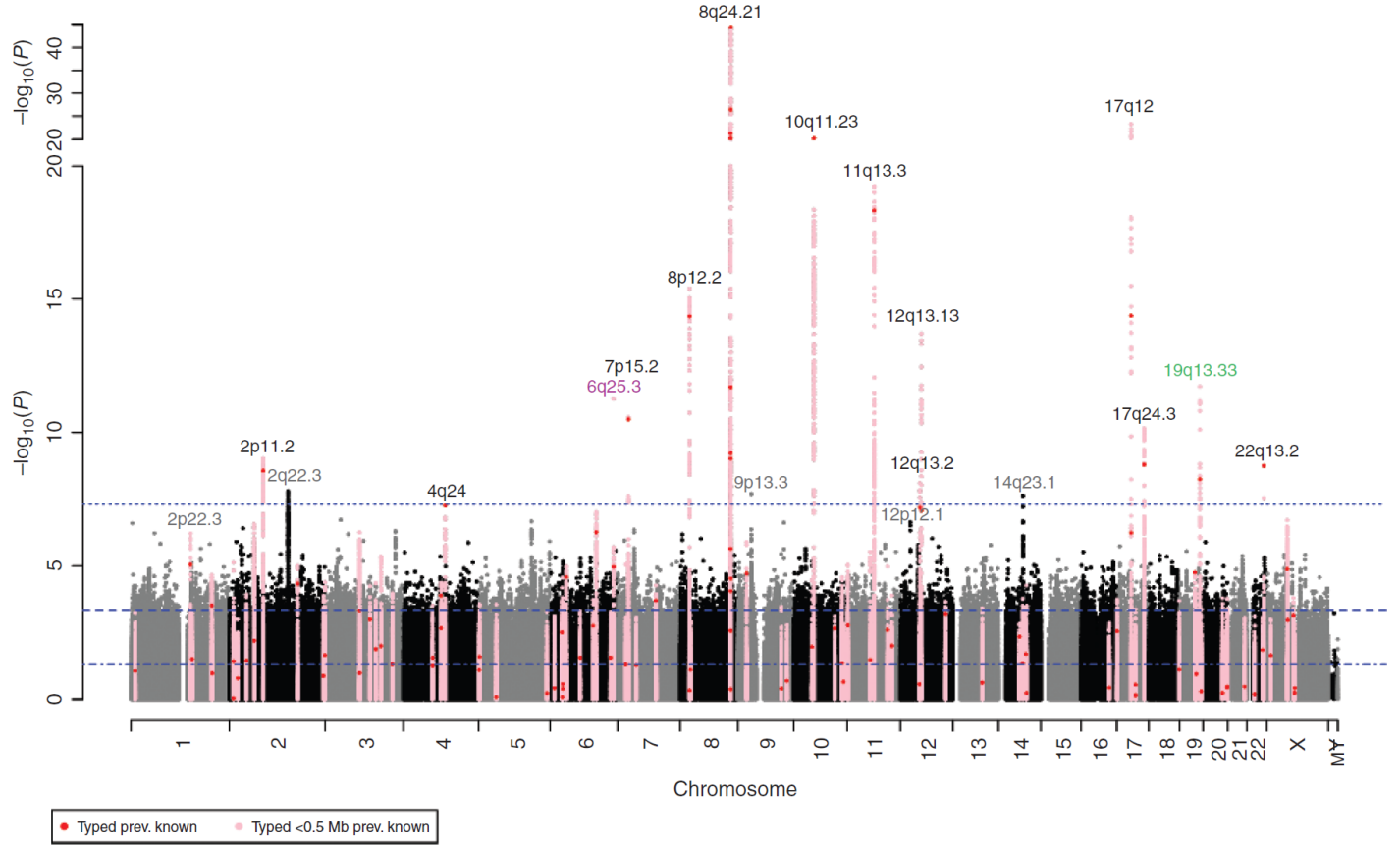
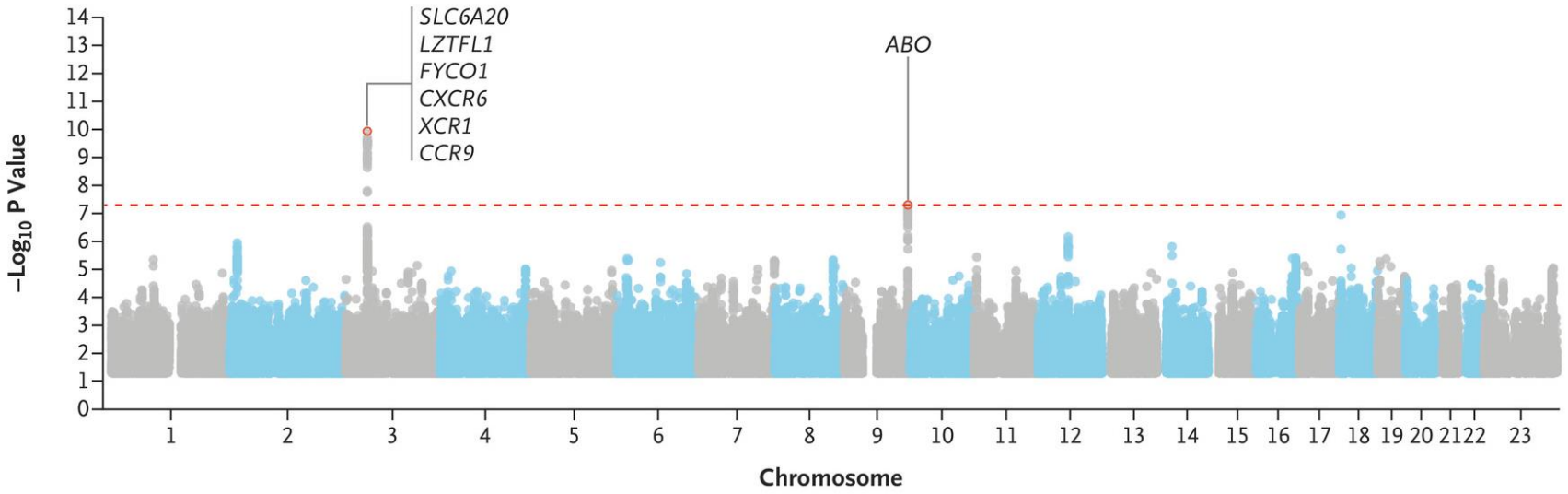


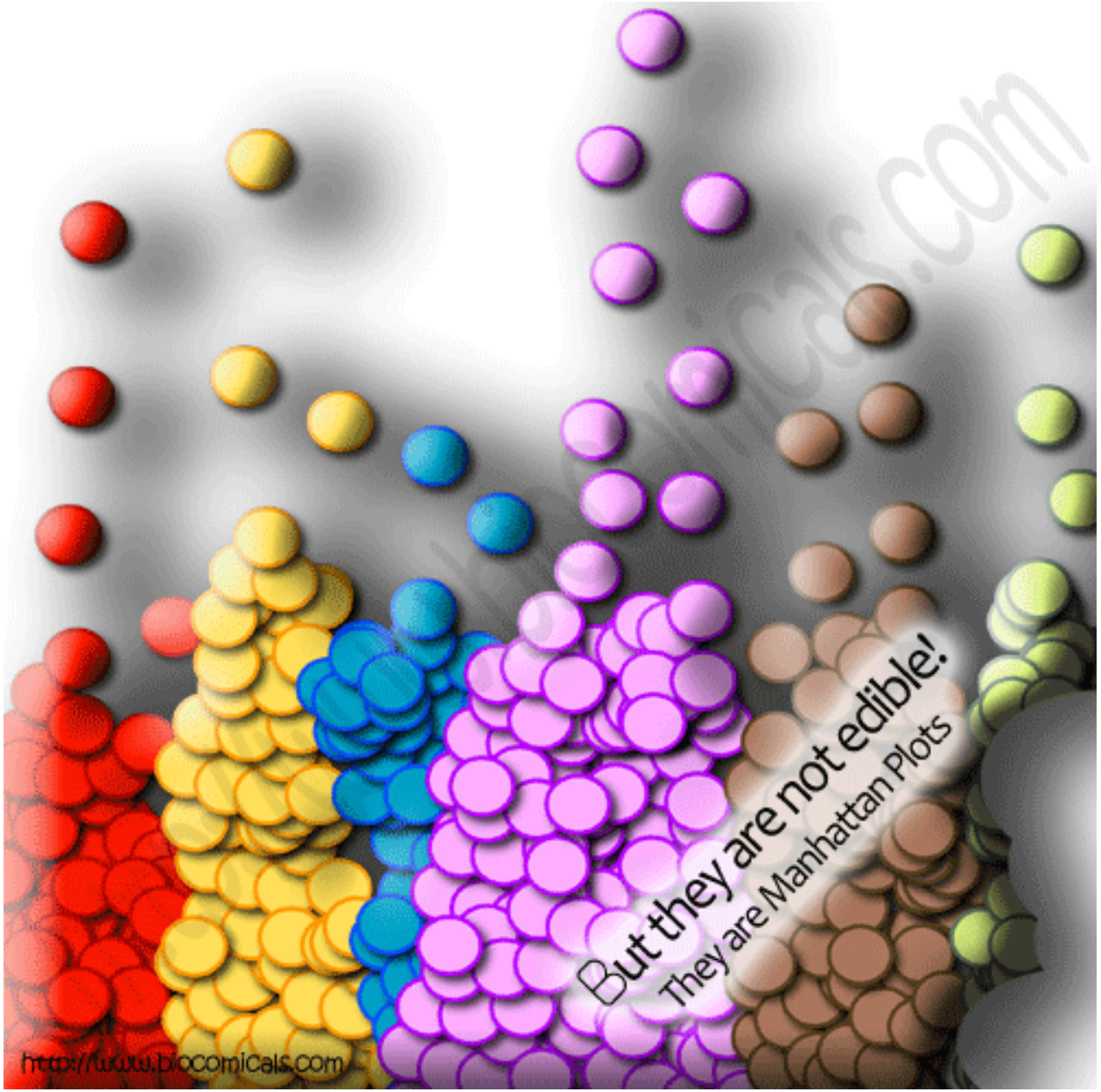
Figure 1. Results from a GWAS of prostate cancer in the KP population (8,399 cases and 38,745 controls), highlighting key chromosomal regions. *P* values are for variant associations with prostate cancer from transethnic fixed-effects meta-analysis of four race/ethnicity GWAS (non-Hispanic white, Latino, 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

Manhattan plot example: Covid-19



The severe covid-19 GWAS group. Genomewide association study of severe covid-19 with respiratory failure. NEJM, 2020; 383:1522-1534.


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

Replication, Replication, Replication

- To replicate in a separate cohort:
  - 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`

Replication and Meta-analysis

Table 2. Susceptibility Loci Associated with Severe Covid-19 with Respiratory Failure.*

Chromosome and Analysis	Meta-analysis		Italian Panel				Spanish Panel			
	P Value	Odds Ratio (95% CI)	P Value	Odds Ratio (95% CI)	Allele Frequency		P Value	Odds Ratio (95% CI)	Allele Frequency	
					patient	control			patient	control
3p21.31†										
Main analysis	1.15×10^{-10}	1.77 (1.48–2.11)	1.98×10^{-7}	1.74 (1.27–2.38)	0.14	0.09	1.32×10^{-4}	1.85 (1.50–2.28)	0.09	0.05
Analysis corrected for age and sex	9.46×10^{-12}	2.11 (1.70–2.61)	7.02×10^{-8}	1.95 (1.53–2.48)	0.14	0.09	1.17×10^{-5}	2.79 (1.76–4.42)	0.09	0.05
9q34.2‡										
Main analysis	4.95×10^{-8}	1.32 (1.20–1.47)	2.90×10^{-6}	1.37 (1.20–1.57)	0.42	0.35	3.55×10^{-3}	1.26 (1.08–1.48)	0.42	0.35
Analysis corrected for age and sex	5.35×10^{-7}	1.39 (1.22–1.59)	5.31×10^{-5}	1.37 (1.17–1.60)	0.42	0.35	2.81×10^{-3}	1.45 (1.13–1.84)	0.42	0.35

* The meta-analysis included 1610 patients and 2205 control participants; the Italian analysis, 835 and 1255, respectively; and the Spanish analysis, 775 and 950, respectively. Allele frequencies of the minor or risk allele (see below) are shown among the patients and the control participants. All the association test statistics were adjusted for the top 10 principal components from the principal-component analysis. Two analyses were performed: a main analysis, which was corrected for 10 principal components, and an analysis that was corrected for age and sex in addition to 10 principal components. In the analyses that were corrected for age and sex, 25 control participants were excluded from the Spanish analysis and the meta-analysis because of missing covariate data. The P values and corresponding odds ratios and 95% confidence intervals (CIs) are shown with respect to the minor allele. Association results for the recessive and heterozygous models for both meta-analyses (main and corrected for age and sex) are shown in Supplementary Appendix 3. Covid-19 denotes coronavirus disease 2019.

† For chromosome 3p21.31, the association boundaries for each index single-nucleotide polymorphism (SNP; see the Supplementary Methods section), with the genomic positions retrieved from genome build hg38, were chr3:45800446 through 46135604. The Single Nucleotide Polymorphism database (dbSNP) identifier was rs11385942 (the rs identifier from the National Center for Biotechnology Information, rs11385942, is annotated as chr3:45834968 through 45834969:AAA:AA in dbSNP, version 153, and as chr3:45834967:GA:G in the TransOmics for Precision Medicine [TOPMed] imputation reference panel). The SNP rs11385942 was imputed according to TOPMed with high confidence (TOPMed estimated imputation accuracy, $R^2=0.94$ and $R^2=0.95$ for the Italian and Spanish panels, respectively) (Supplementary Appendix 2).The minor or risk allele was GA, and the major allele was G. The key genes (i.e., the candidate genes in the region) were *SLC6A20*, *LZTFL1*, *FYCO1*, *CXCR6*, *XCR1*, and *CCR9*.

‡ For chromosome 9q34.2, the association boundaries for each index SNP, with the genomic positions retrieved from genome build hg38, were chr9:133257521 through 133279871. The SNP rs657152 was genotyped according to the Global Screening Array (GSA) in the Italian and Spanish panels (Supplementary Appendix 2). The minor or risk allele was A, and the major allele was C. The key gene was *ABO*.

The severe covid-19 GWAS group. Genomewide association study of severe covid-19 with respiratory failure. NEJM, 2020; 383:1522-1534.

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)

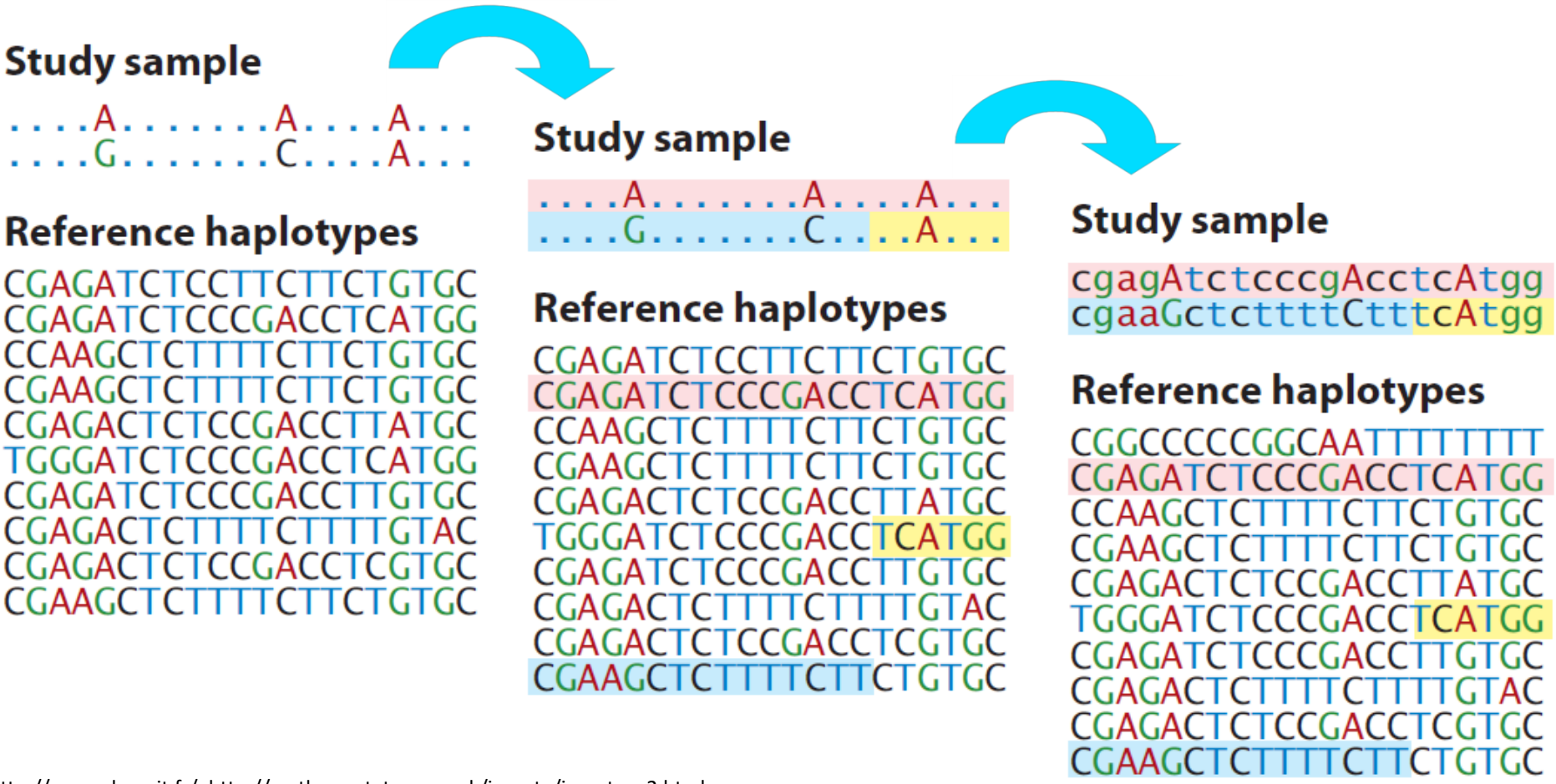
How would you fill in the remaining genotypes of this study sample individual, given the reference haplotypes?

- 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

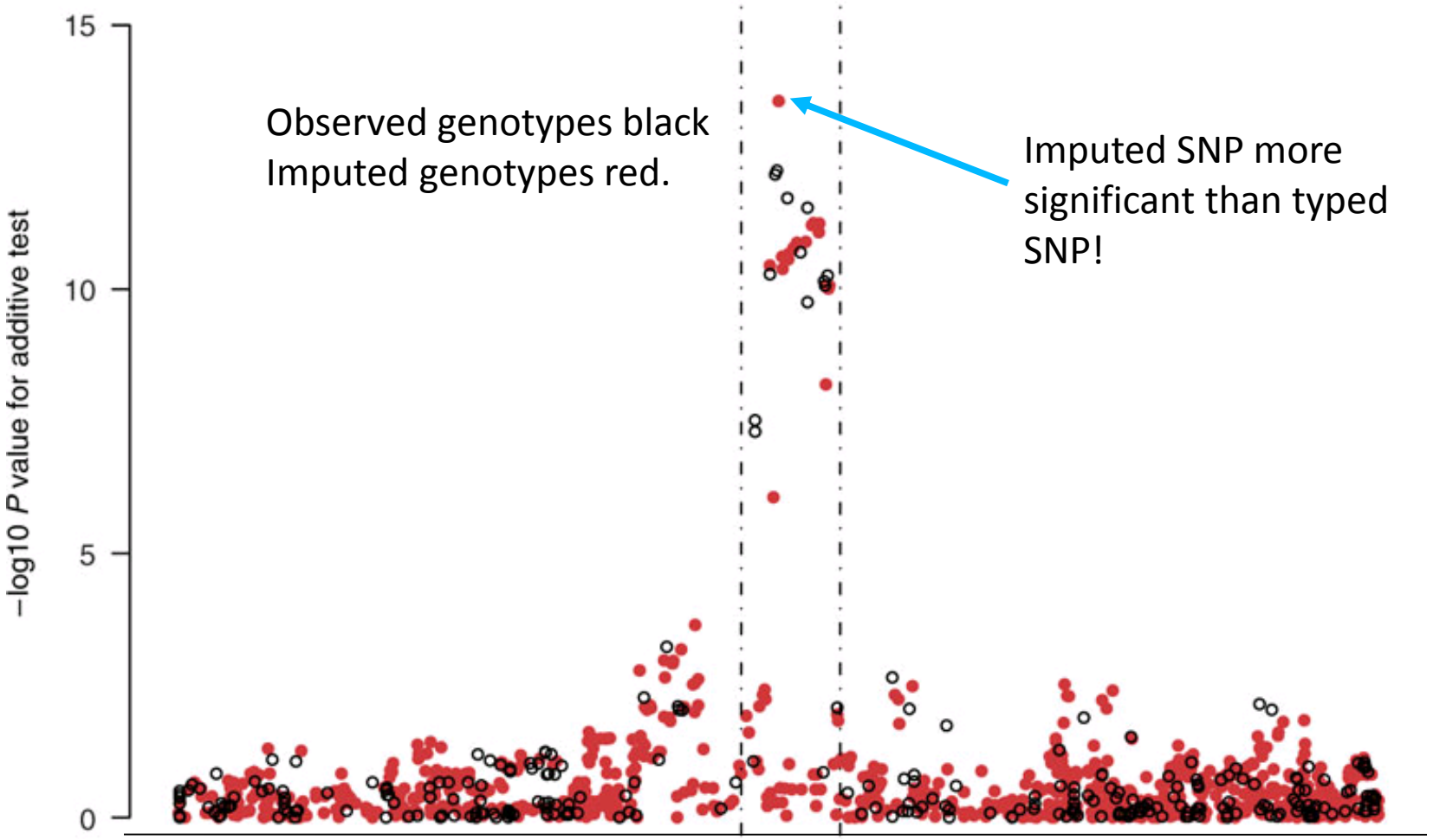


- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

Li et al., Ann Rev Human Genet, 2009

Imputation Application

TCF7L2 gene region & T2D from the WTCCC data



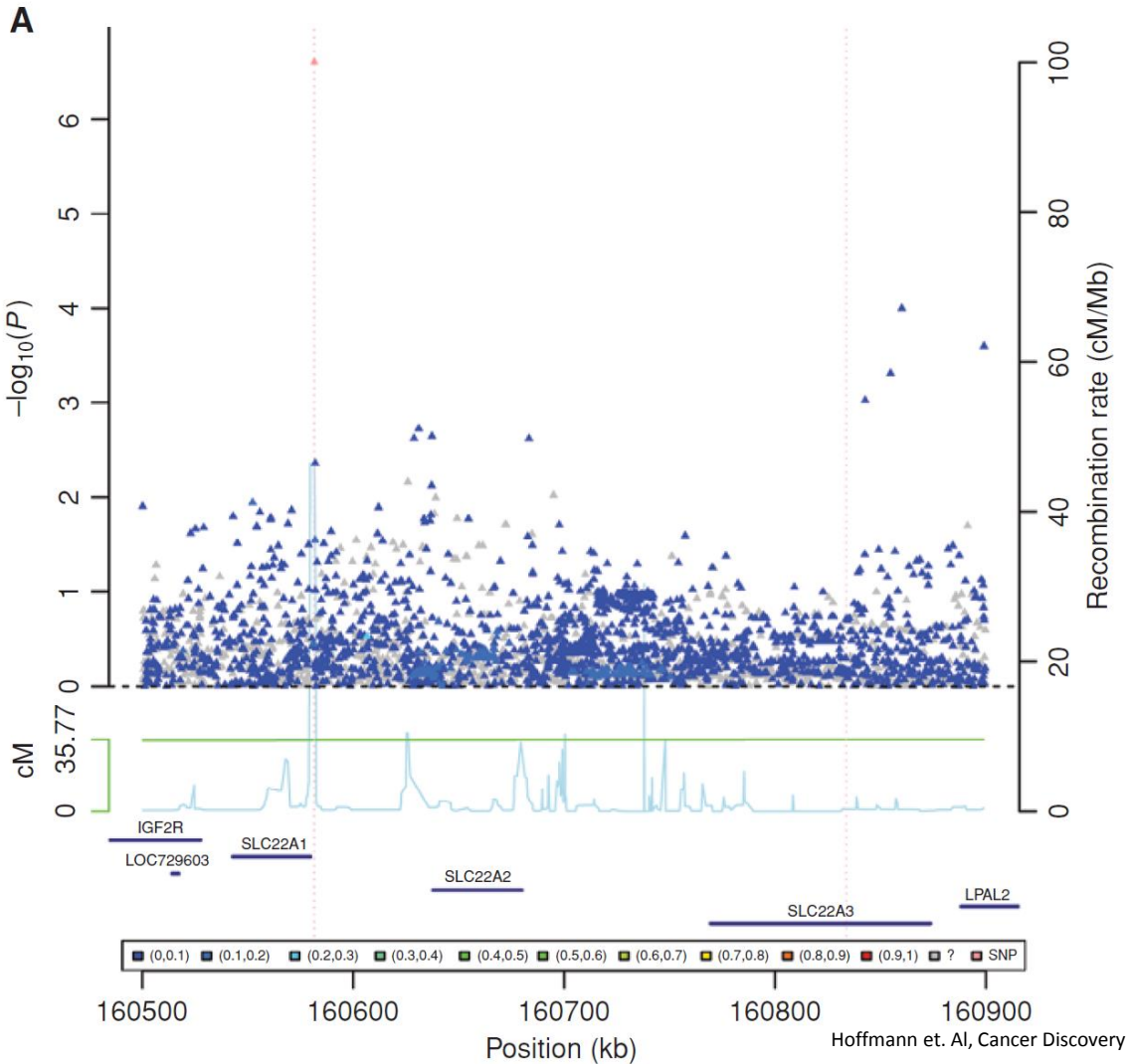
Marchini Nature Genetics2007

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

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

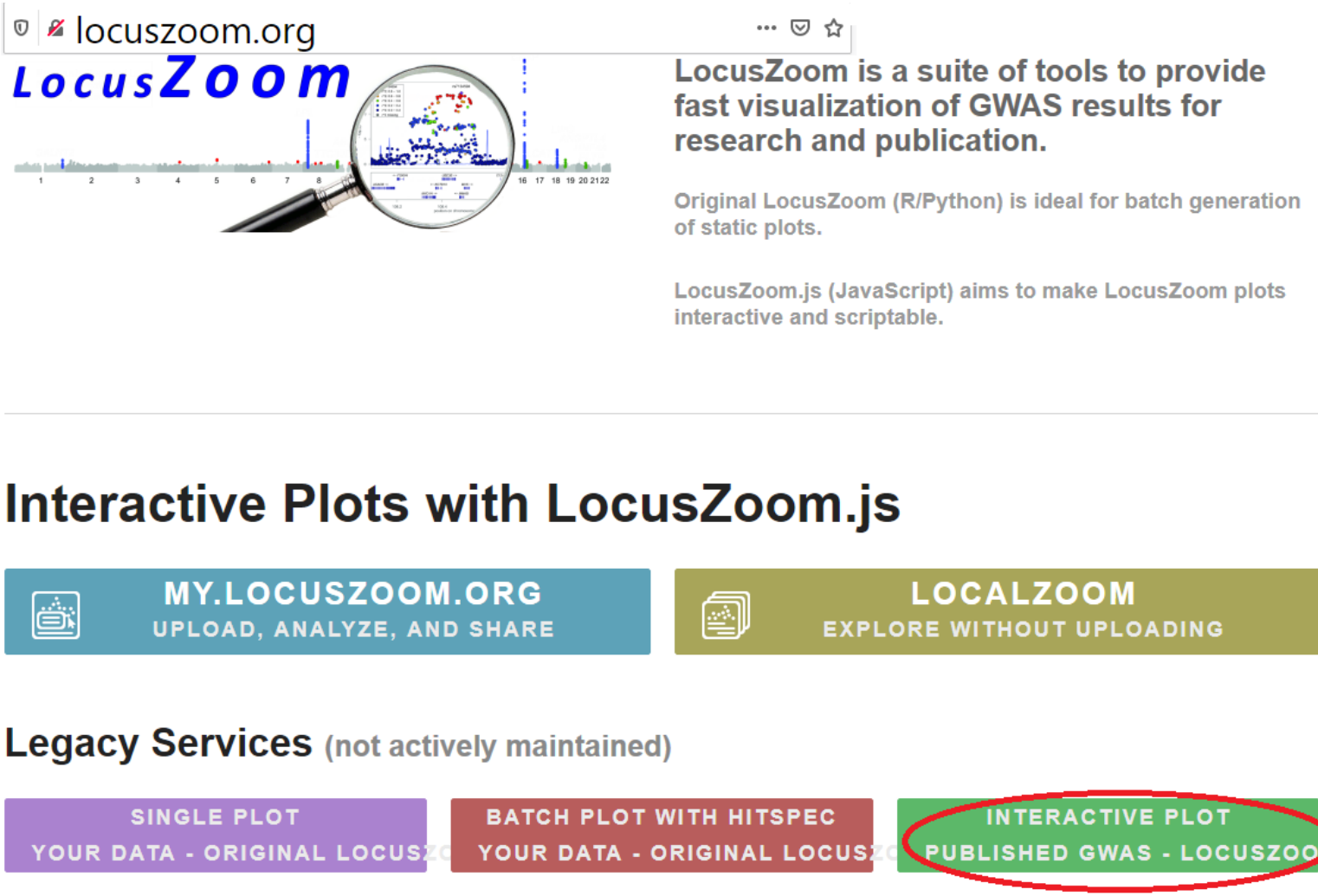
Figure 2. Regional fixed-effects meta-analysis plots from GWAS in KP population of the two risk variants that replicated: (**A**) 6q25.3 (rs4646284), (**B**) 19q13.33 (rs2659124). The color code for the points represents the r^2 of each SNP with the risk variant (ranges defined in the legend). The dotted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

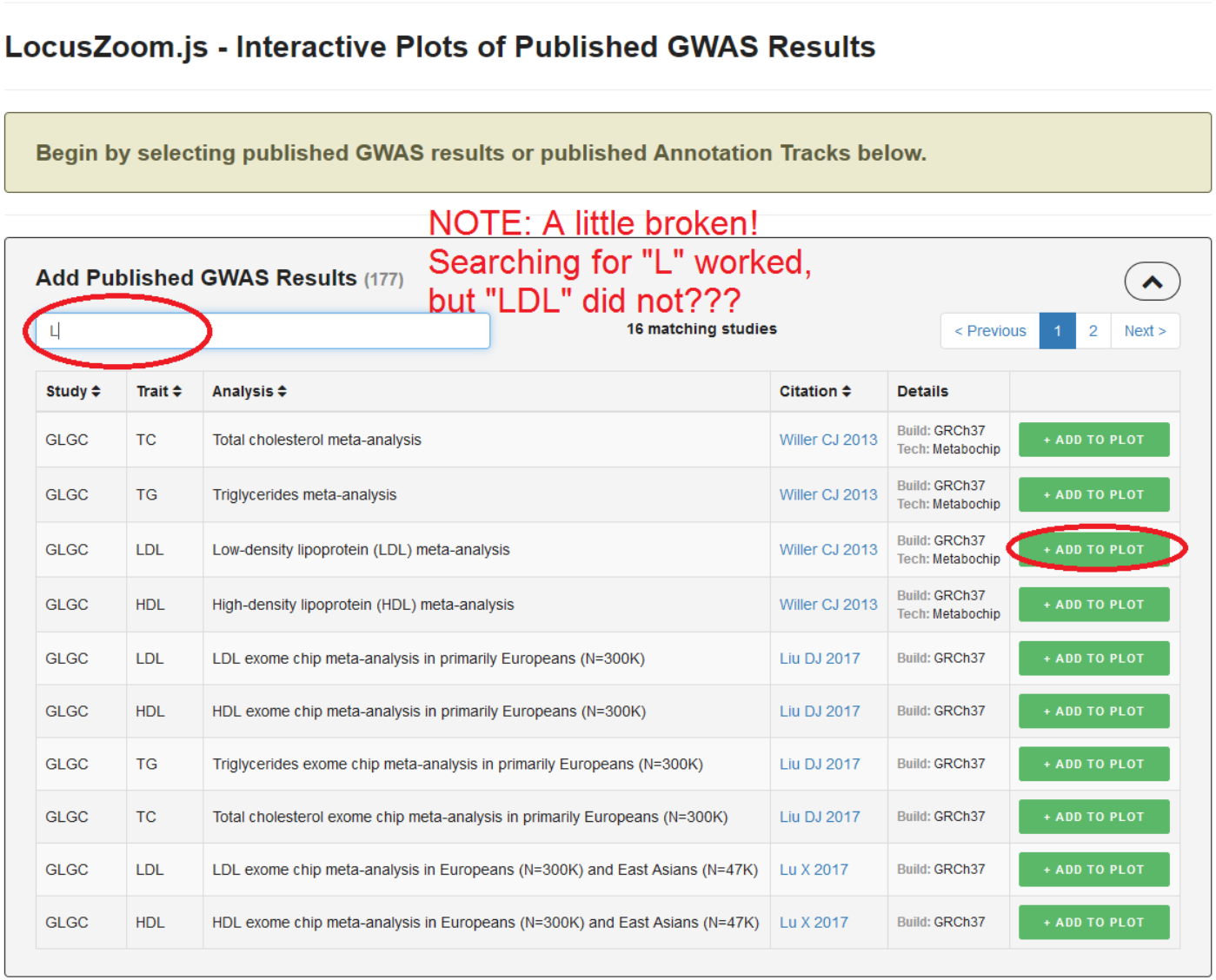


Hoffmann et. Al, Cancer Discovery 2015`

LocusZoom plot: LDL, rs6544713



LocusZoom plot: LDL, rs6544713



LocusZoom.js - Interactive Plots of Published GWAS Results

Begin by selecting published GWAS results or published Annotation Tracks below.

NOTE: A little broken!
Searching for "L" worked,
but "LDL" did not???

Add Published GWAS Results (177)

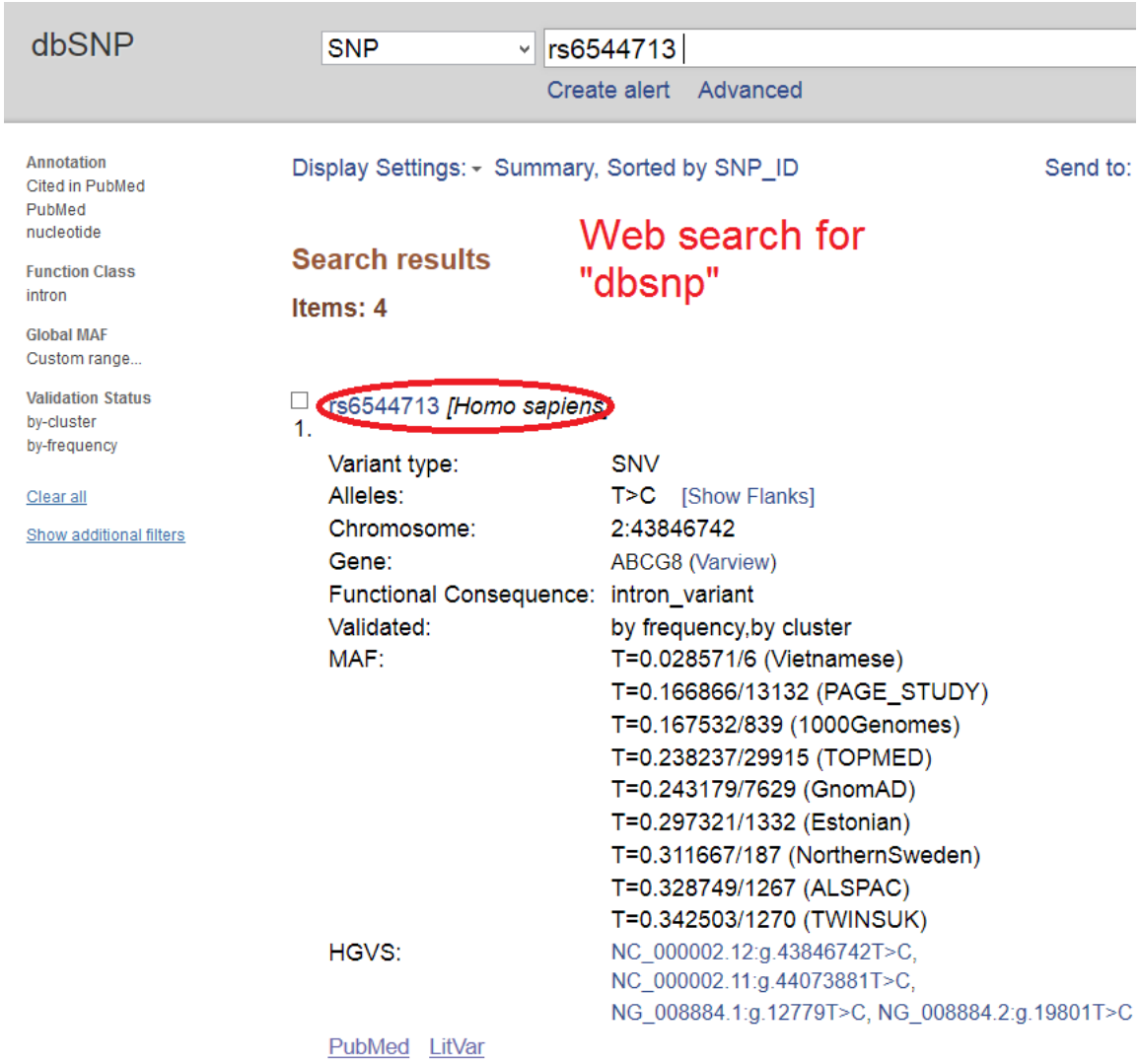
L

16 matching studies

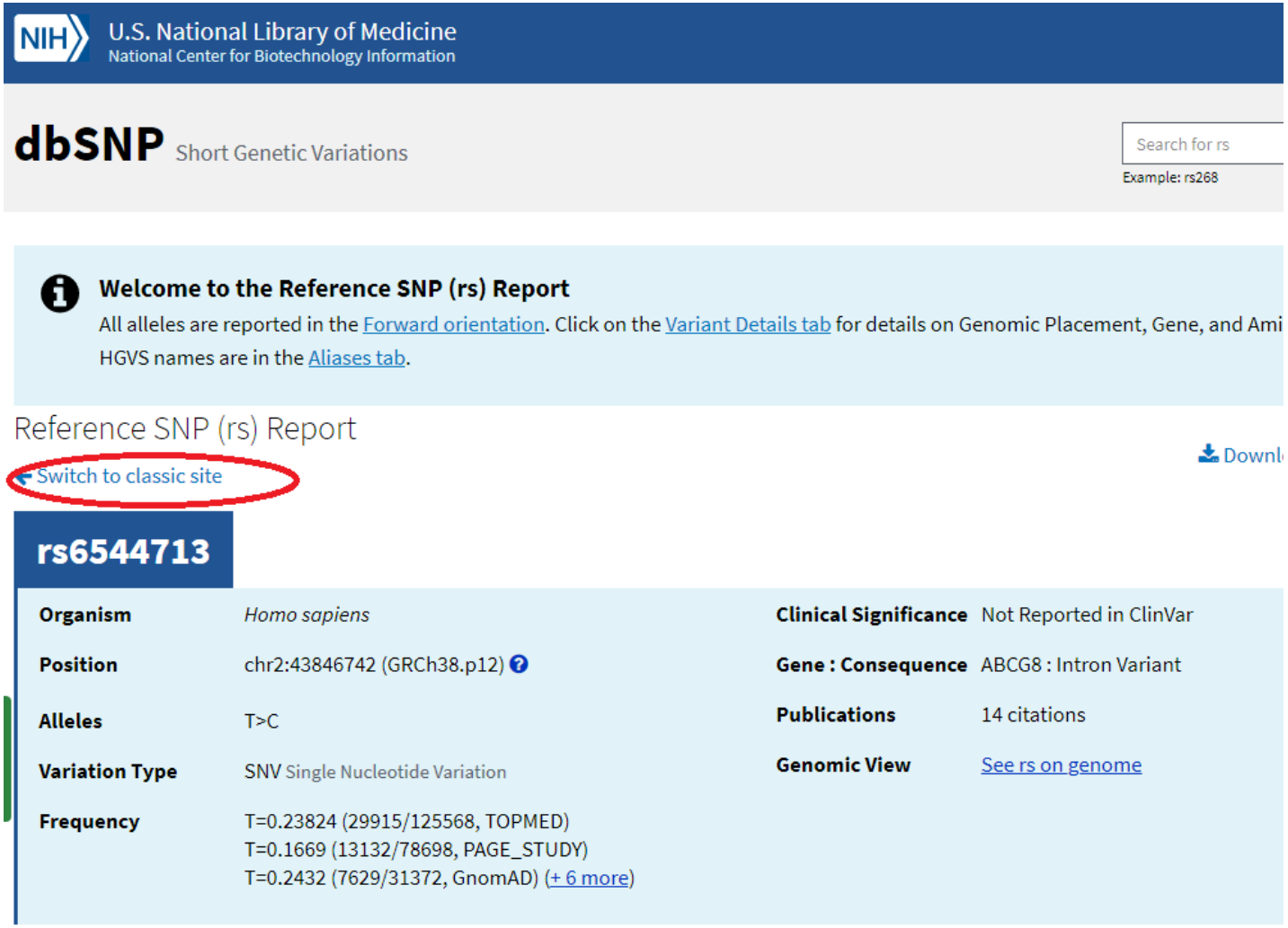
< Previous 1 2 Next >

Study	Trait	Analysis	Citation	Details	
GLGC	TC	Total cholesterol meta-analysis	Willer CJ 2013	Build: GRCh37 Tech: Metabochip	+ ADD TO PLOT
GLGC	TG	Triglycerides meta-analysis	Willer CJ 2013	Build: GRCh37 Tech: Metabochip	+ ADD TO PLOT
GLGC	LDL	Low-density lipoprotein (LDL) meta-analysis	Willer CJ 2013	Build: GRCh37 Tech: Metabochip	+ ADD TO PLOT
GLGC	HDL	High-density lipoprotein (HDL) meta-analysis	Willer CJ 2013	Build: GRCh37 Tech: Metabochip	+ ADD TO PLOT
GLGC	LDL	LDL exome chip meta-analysis in primarily Europeans (N=300K)	Liu DJ 2017	Build: GRCh37	+ ADD TO PLOT
GLGC	HDL	HDL exome chip meta-analysis in primarily Europeans (N=300K)	Liu DJ 2017	Build: GRCh37	+ ADD TO PLOT
GLGC	TG	Triglycerides exome chip meta-analysis in primarily Europeans (N=300K)	Liu DJ 2017	Build: GRCh37	+ ADD TO PLOT
GLGC	TC	Total cholesterol exome chip meta-analysis in primarily Europeans (N=300K)	Liu DJ 2017	Build: GRCh37	+ ADD TO PLOT
GLGC	LDL	LDL exome chip meta-analysis in Europeans (N=300K) and East Asians (N=47K)	Lu X 2017	Build: GRCh37	+ ADD TO PLOT
GLGC	HDL	HDL exome chip meta-analysis in Europeans (N=300K) and East Asians (N=47K)	Lu X 2017	Build: GRCh37	+ ADD TO PLOT

Very broken, need to get coordinates by hand!



Very broken, need to get coordinates by hand!



Very broken, need to get coordinates by hand!

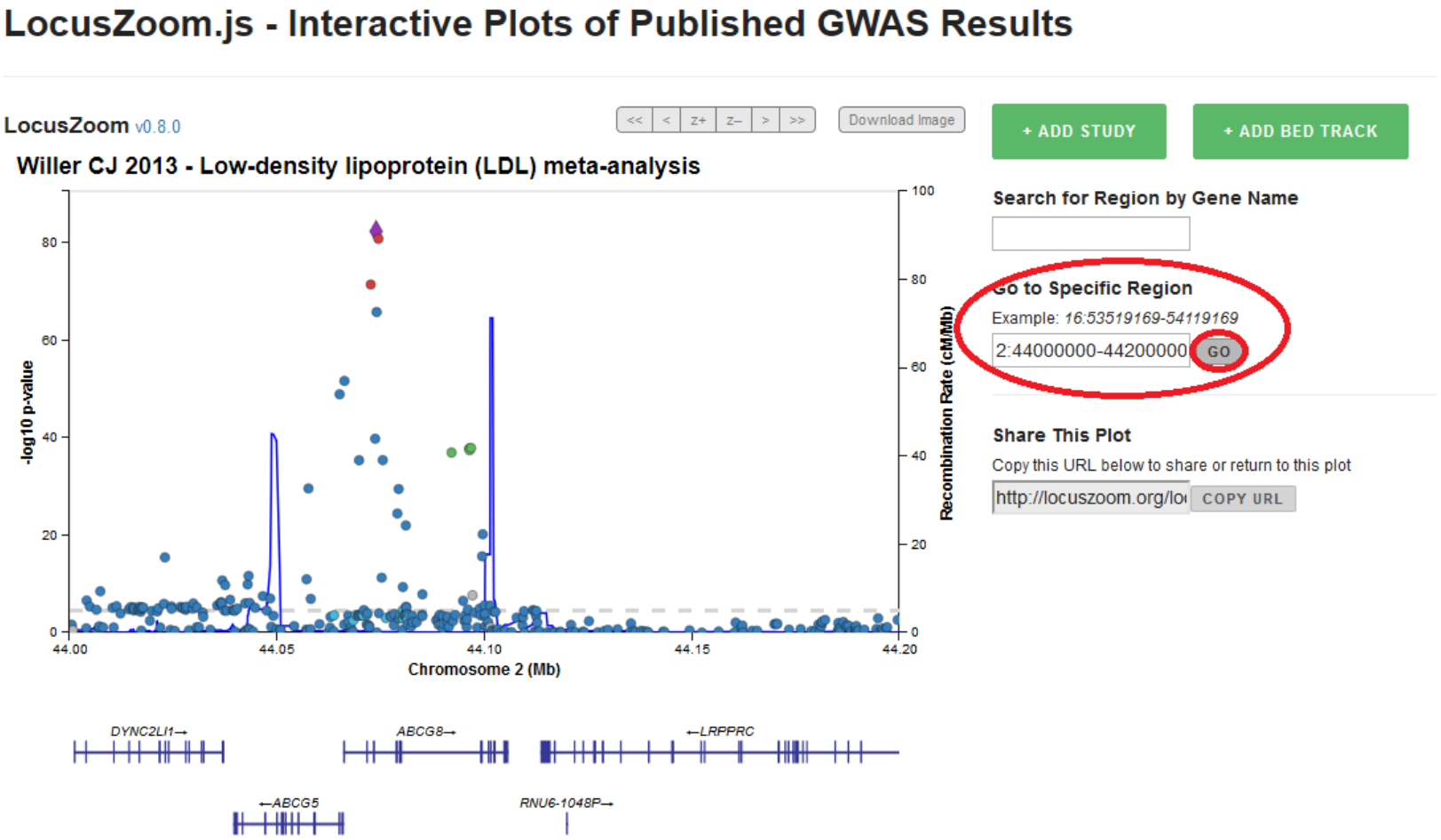
SNP Details are organized in the following sections:

GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

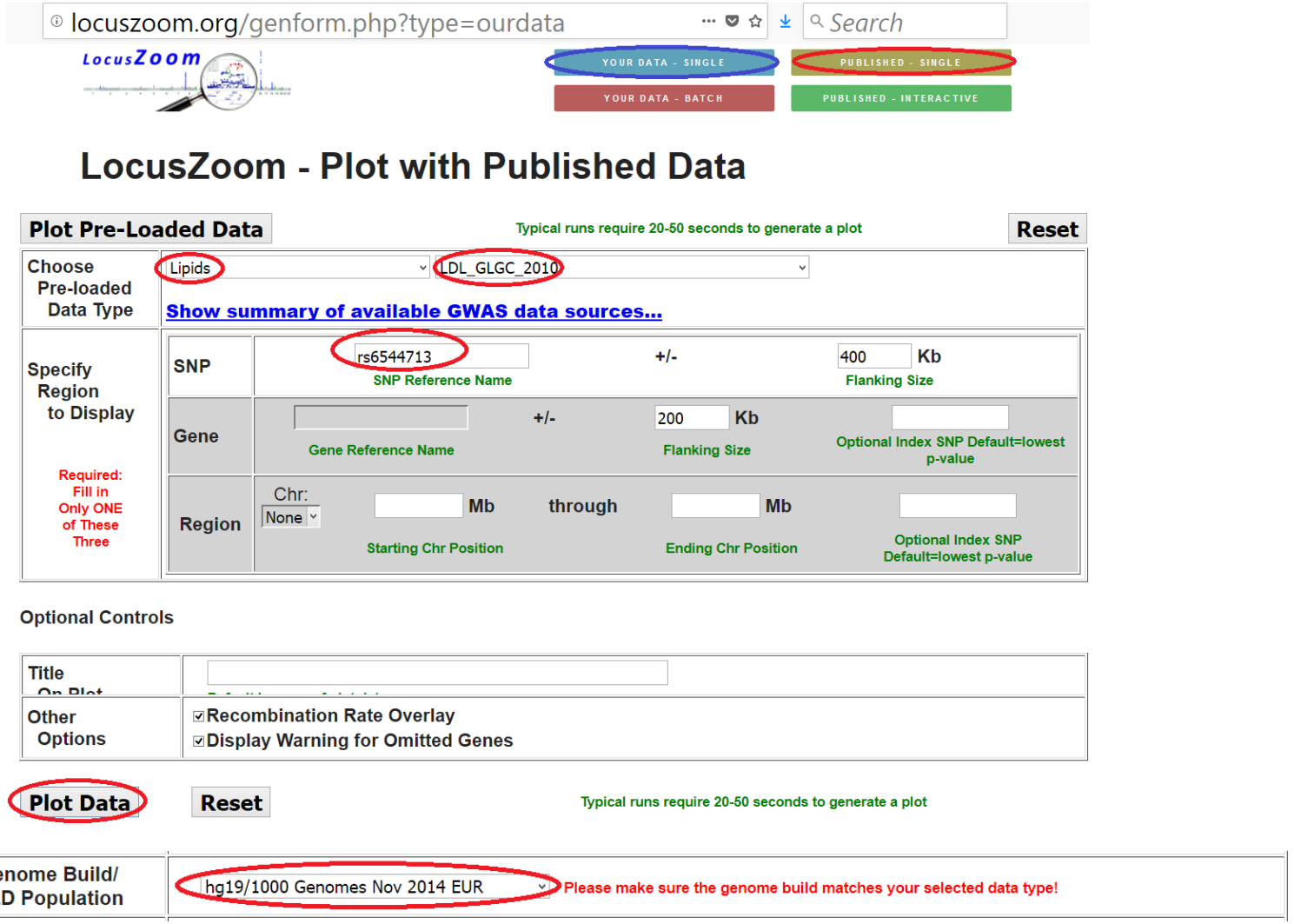
Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

Assembly	Annotation Release	Chr	Chr Pos	Contig
GRCh38.p7	108	2	43846742	NT_022184.16
GRCh37.p13	105	2	44073881	NT_022184.15

Very broken, need to get coordinates by hand!



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

GWAS catalog (revisit)

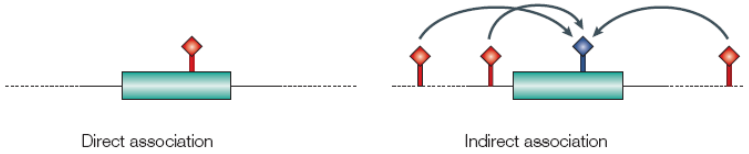
https://www.ebi.ac.uk/gwas/home

Outline for GWAS

- Analysis
  - QC
  - Analysis approach
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis
- GWAS Limitations

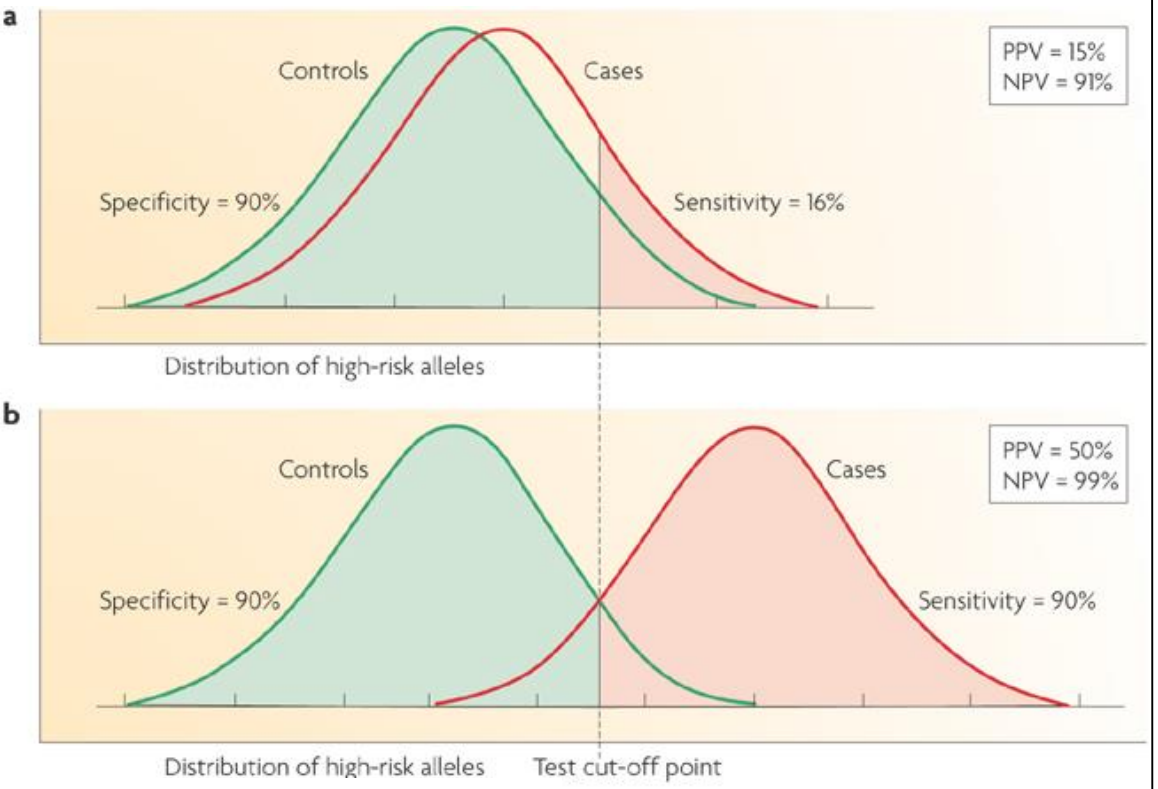
Limitations of GWAS

- 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

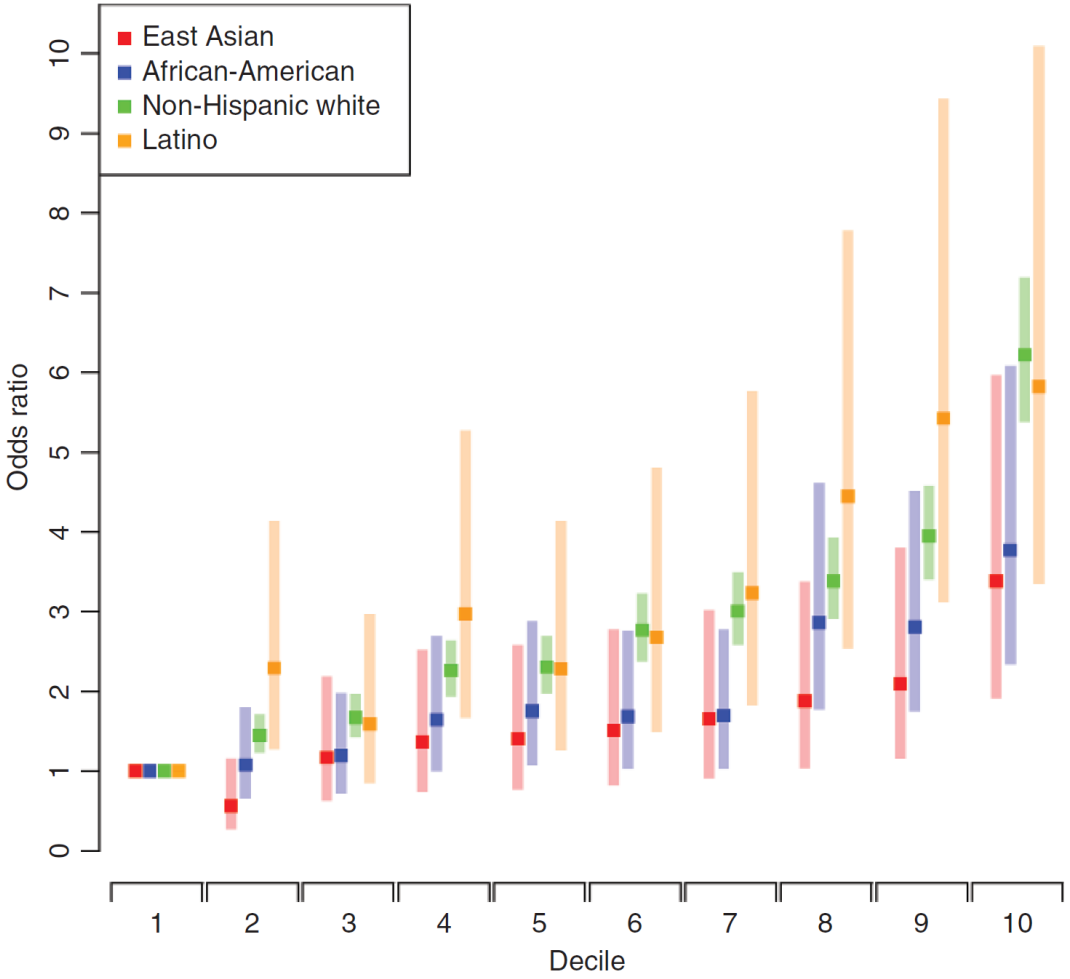
Example:
AUC for Breast Cancer Risk

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

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? …

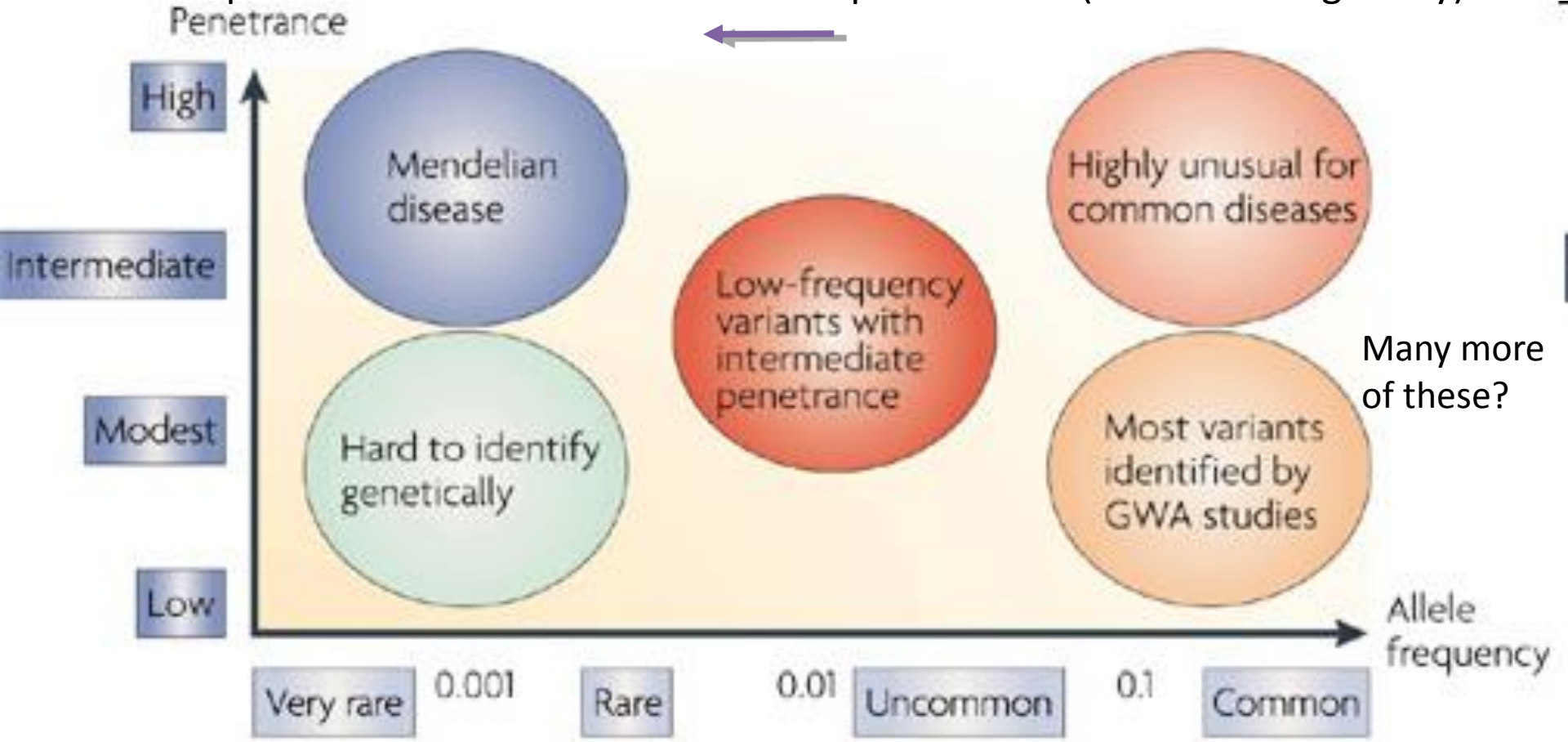
- Predict 100,000 variants with some effect on height…
- PRS variant that builds model based on *all* SNPS

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
   - 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

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)
- ...

But, what about more complicated modeling? Epistasis

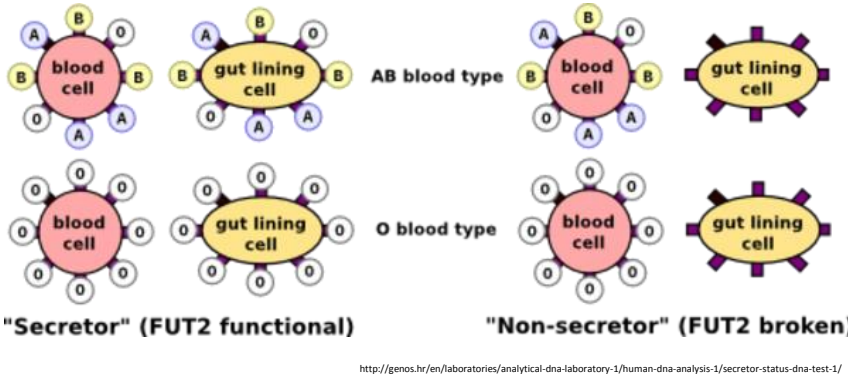
Epistasis – snp x snp, or gene x gene interactions

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



http://genos.hr/en/laboratories/analytical-dna-laboratory-1/human-dna-analysis-1/secretor-status-dna-test-1/

Serum lipids example

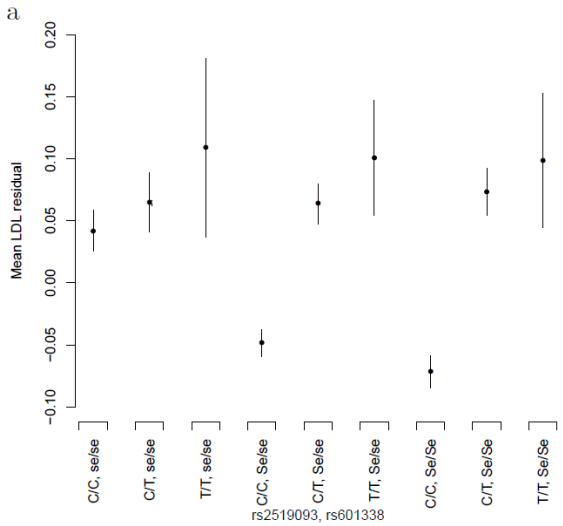
Tested all pairs of known GWAS hit variants

Rs2519093 (*ABO*) x rs492602 (*FUT2*): Interaction

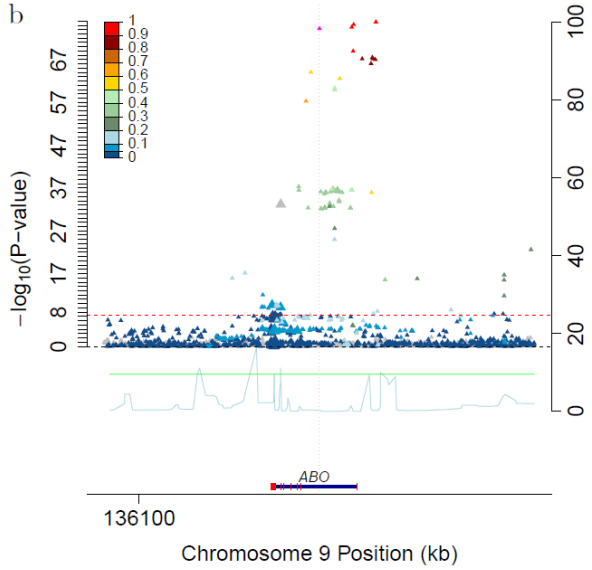
$P_{LDL}=8.1x10^{-10}$

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