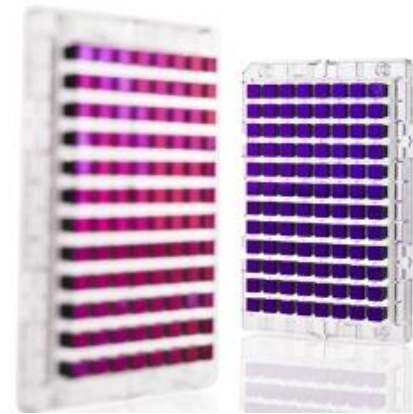


Genome-wide association studies (GWAS)



Further reading (not required)

- Nature reviews genetics:

<http://www.nature.com/nrg/series/gwas/index.html>

- McCarthy et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges
- Marchini and Howie. Genotype imputation for genome-wide association studies
- Ott. Family-based designs for genome-wide association studies.
- Gibson. Rare and common variants: twenty arguments.

Outline for GWAS

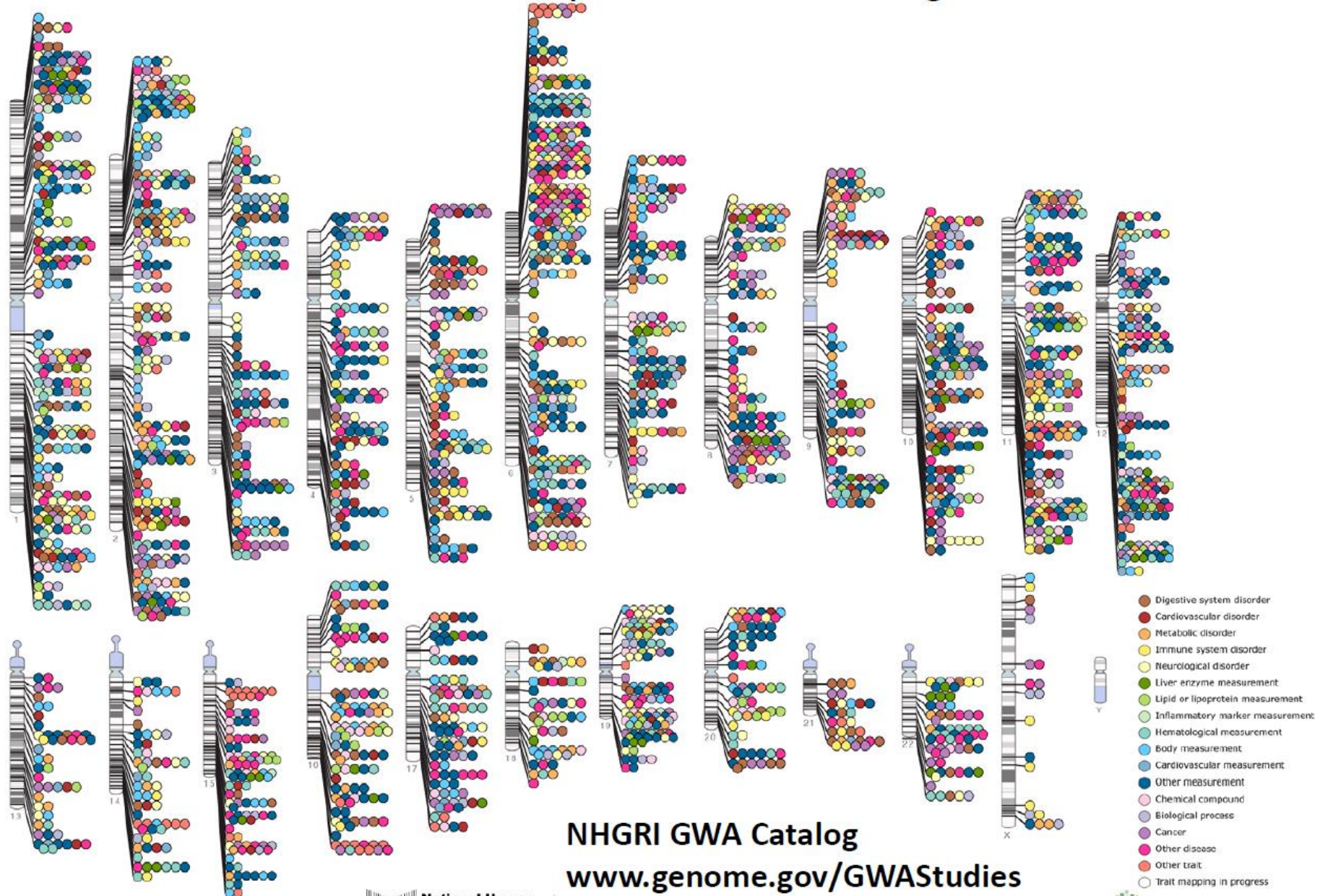
- Review / Overview
- Design
- Analysis
 - QC
 - Prostate cancer example
 - Imputation
 - Replication & Meta-analysis
- Advanced analysis
 - Limitations & “missing heritability”
 - Gene/pathway tests
 - Polygenic models

Outline for GWAS

- Review / Overview
- Design
- Analysis
 - QC
 - Prostate cancer example
 - Imputation
 - Replication & Meta-analysis
- Advanced analysis
 - Limitations & “missing heritability”
 - Gene/pathway tests
 - Polygenic models

Published Genome-Wide Associations through 07/2012

Published GWA at $p \leq 5 \times 10^{-8}$ for 18 trait categories



NHGRI GWA Catalog

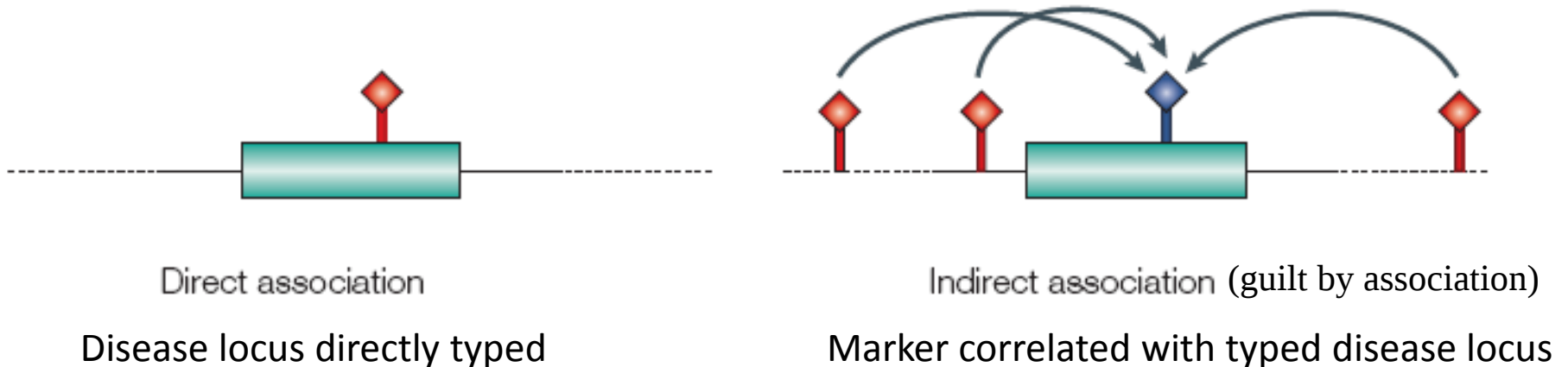
www.genome.gov/GWASStudies

www.ebi.ac.uk/fgpt/gwas/

Outline for GWAS

- Review / Overview
- Design
- Analysis
 - QC
 - Prostate cancer example
 - Imputation
 - Replication & Meta-analysis
- Advanced analysis
 - Limitations & “missing heritability”
 - Gene/pathway tests
 - Polygenic models

Design 1: Microarray – Genotype a subset of markers available



- Previously we talked about using this to “tag” candidate genes (only Genotype a subset of the SNPs).
- Same principal applies to tagging the whole genome.

Technology behind a GWAS Microarray

Target prep

Amplify

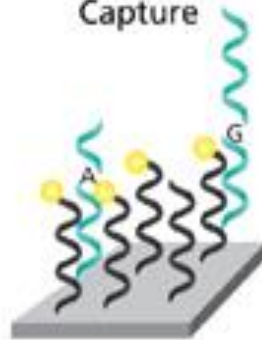


Fragment



Hybridization

Capture



+

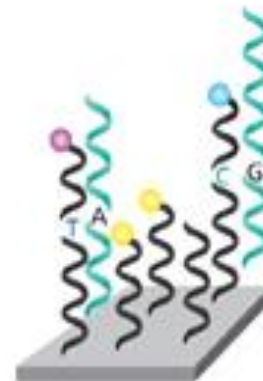
Label



Labeled solution probe
Sample binds to
array

Ligation

Differentiate



Labeled probes
bind to sample,
differentiating
between the two
alleles

Signal amplification

Stain and image

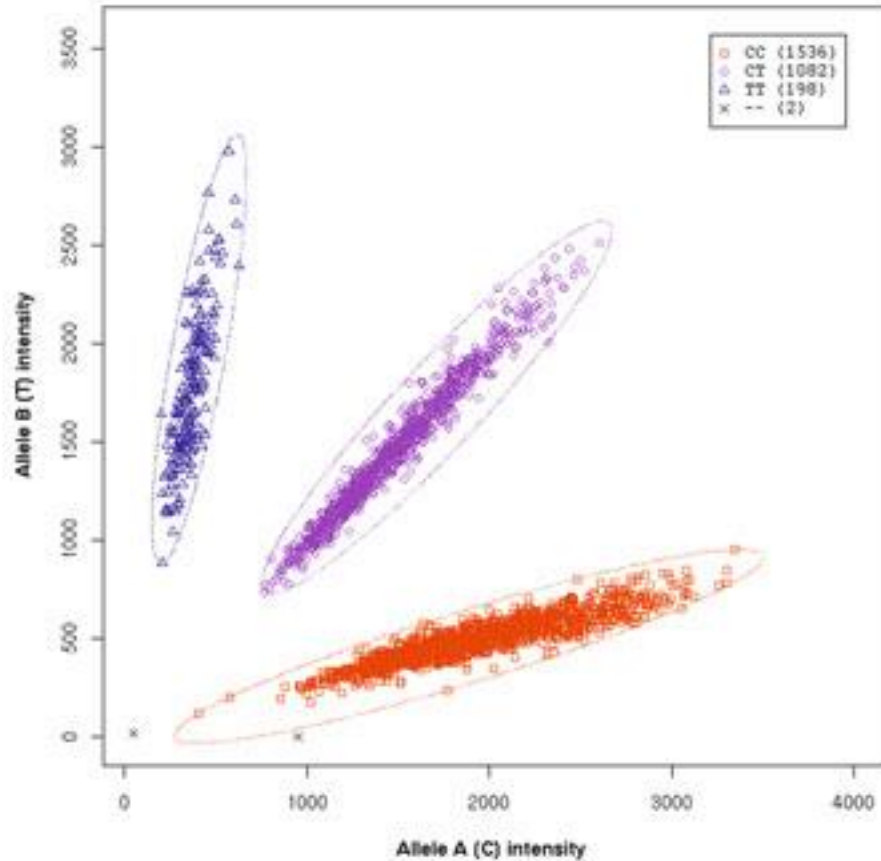


Make it bright
enough then
measure intensity of
array

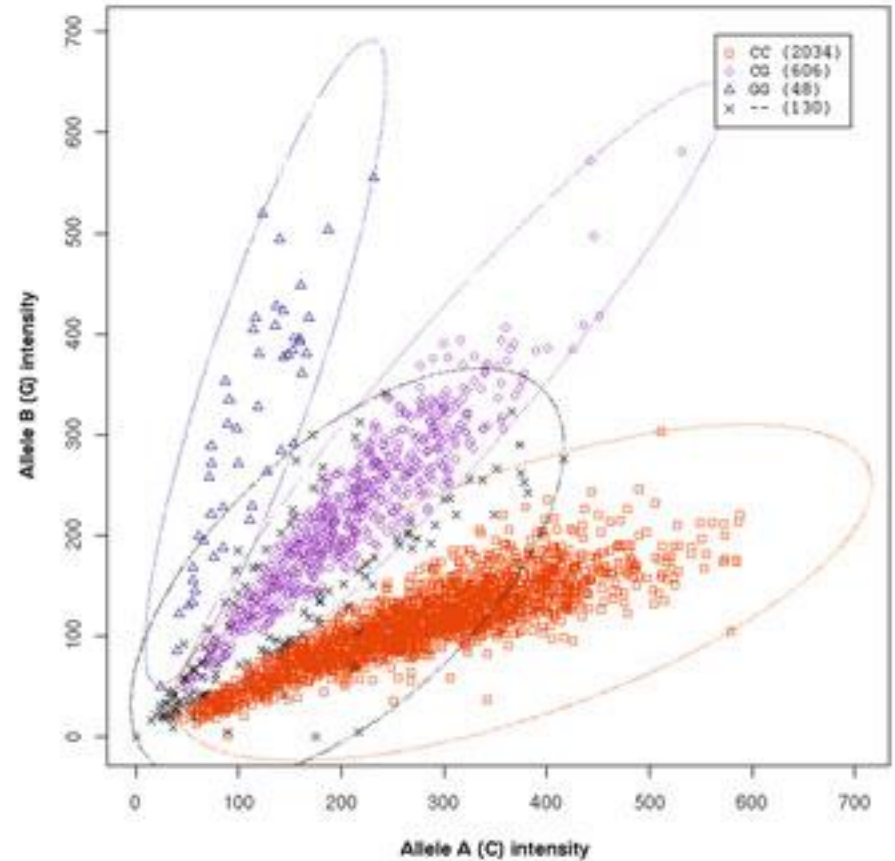
Assay ~ 0.7 - 5M SNPs (keeps increasing)

Affymetrix, <http://www.affymetrix.com>

Raw microarray data: Genotype calls

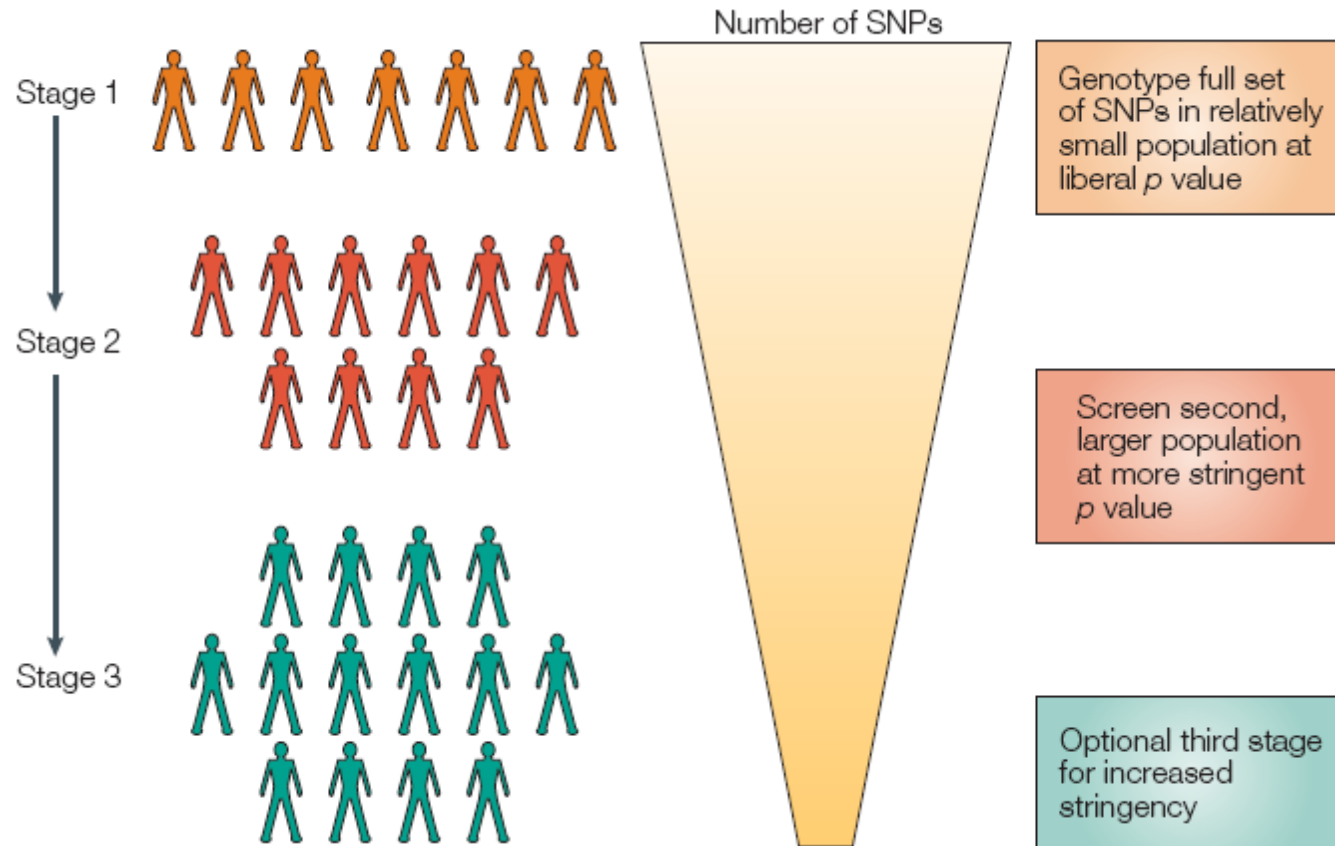


Good calls!



Bad calls!

Genome-wide Association Studies (GWAS)



Design 2: Next generation sequencing

- Genotype essentially “all” SNPs in genome
 - A bunch of random reads all over the genome are assembled to produce the genome
 - Some regions of the genome have more reads than others; can choose how much you want to sequence (5x coverage 1000 Genomes, 60-80x Complete Genomics, ...)
- More expensive
- Extreme phenotypes, since can't do everyone?
- More in “Next Generation Sequencing” lecture

Design choices

- **GWAS Microarray**

- Only assay SNPs designed into array (0.7-5 million)
- Cheaper (so more subjects tradeoff)

GWAS Sequencing

“De novo” discovery
(particularly good for rare variants)

More expensive (but costs are falling) (many less subjects)

Need much more expansive IT support

Lots of interesting interpretation problems (field rapidly evolving)

Can also do a hybrid design, more later in the lecture.

Design choices

- **Exome Microarray**

- Only assay SNPs designed into array (~300K+custom); in **exons only** and that could affect protein coding function
- Cheapest (so many more subjects)

Exome Sequencing

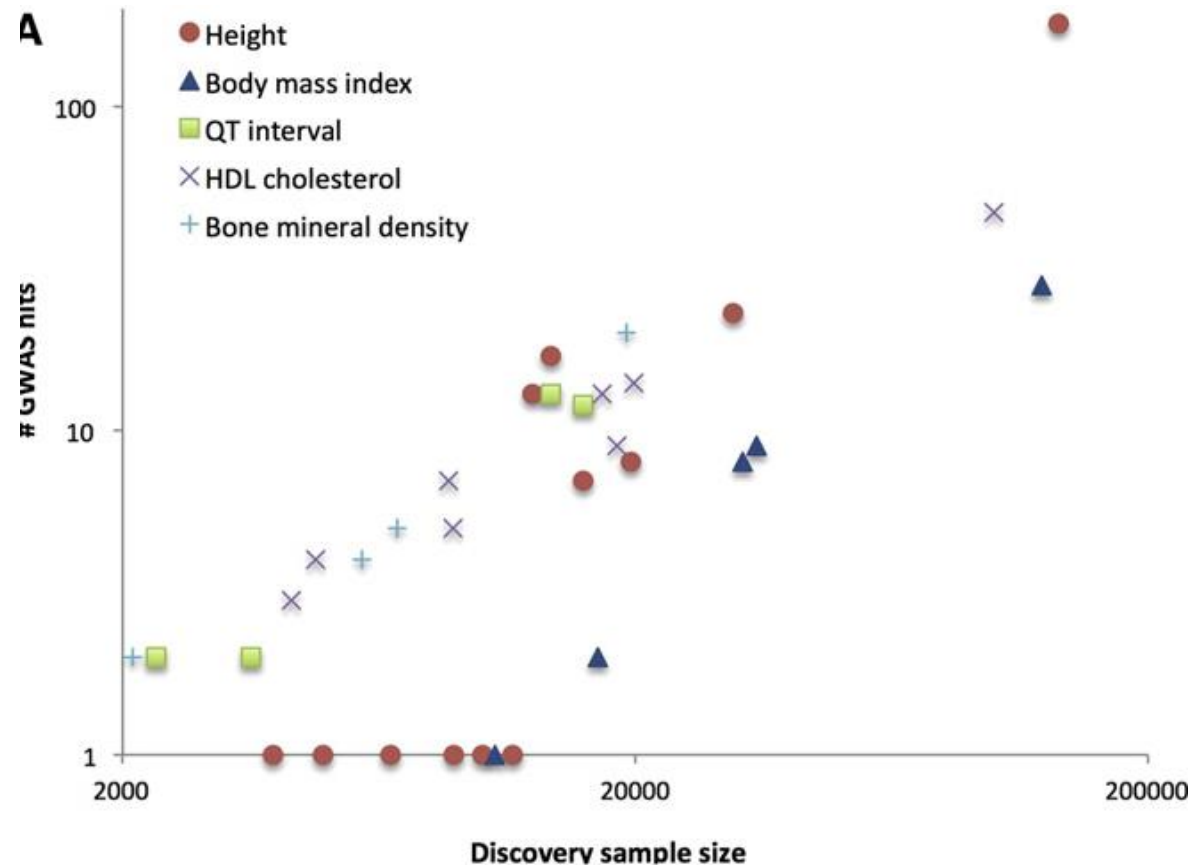
“De novo” discovery
(particularly good for rare variants); **%age of exons only**
(first step is a pull down that does not capture all exons)

More expensive than
microarrays, less expensive
than gwas sequencing

Need more expansive IT
support, but not as much as
whole genome

Size matters

- Large # SNPs tested – multiple comparisons
- Very small effect sizes
- Tag SNP, rather than actual SNP
- Consortia



Biggest studies...

- GWAS Microarray:
 - US: 100,000 people in the Kaiser RPGEH (Hoffmann et al., Genomics, 2011a&b)
 - UK: 500,000 people in process
- Sequencing: 1000 Genomes Project, UK10K
- Exome Sequencing: GO ESP (12,031 subjects, for exome microarray design)

Outline for GWAS

- Review / Overview
- Design
- Analysis
 - QC
 - Prostate cancer example
 - Imputation
 - Replication & Meta-analysis
- Advanced analysis
 - Limitations & “missing heritability”
 - Gene/pathway tests
 - Polygenic models

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 (mislabelled samples)
- Filter Mendelian inheritance (family-based, or potentially cryptics, if large enough sample)
- plink software: pngu.mgh.harvard.edu/~purcell/plink/reference.shtml

Filter for Mendelian inheritance

A | A - - - - A | A

|

A | T

Filter for Mendelian inheritance

A | A - - - - A | A

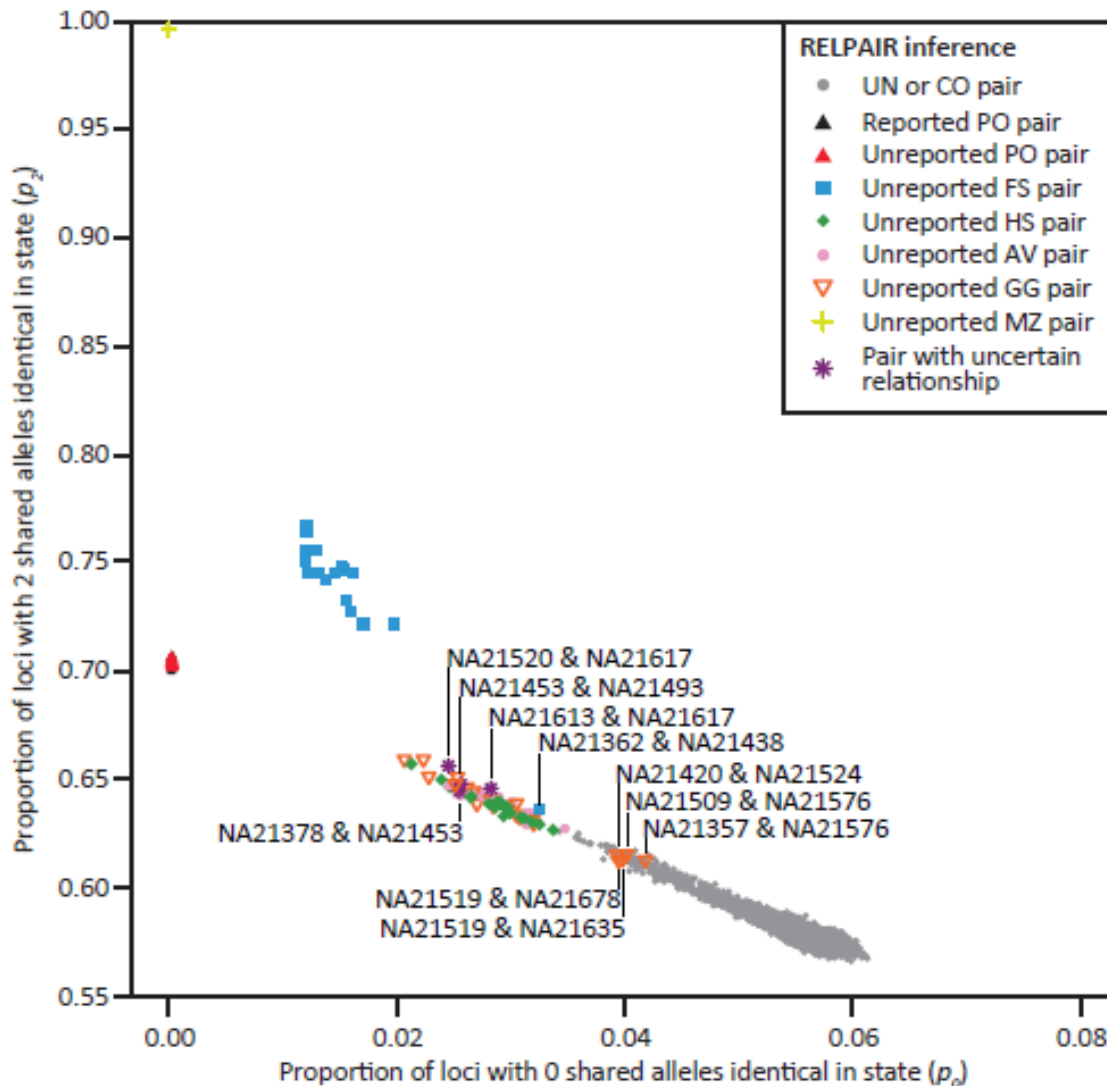
|

A | T ← Offspring Genotype not
Possible!

(De novo mutation, but
more likely genotyping
error.)

Check for relatedness, e.g., HapMap

MKK



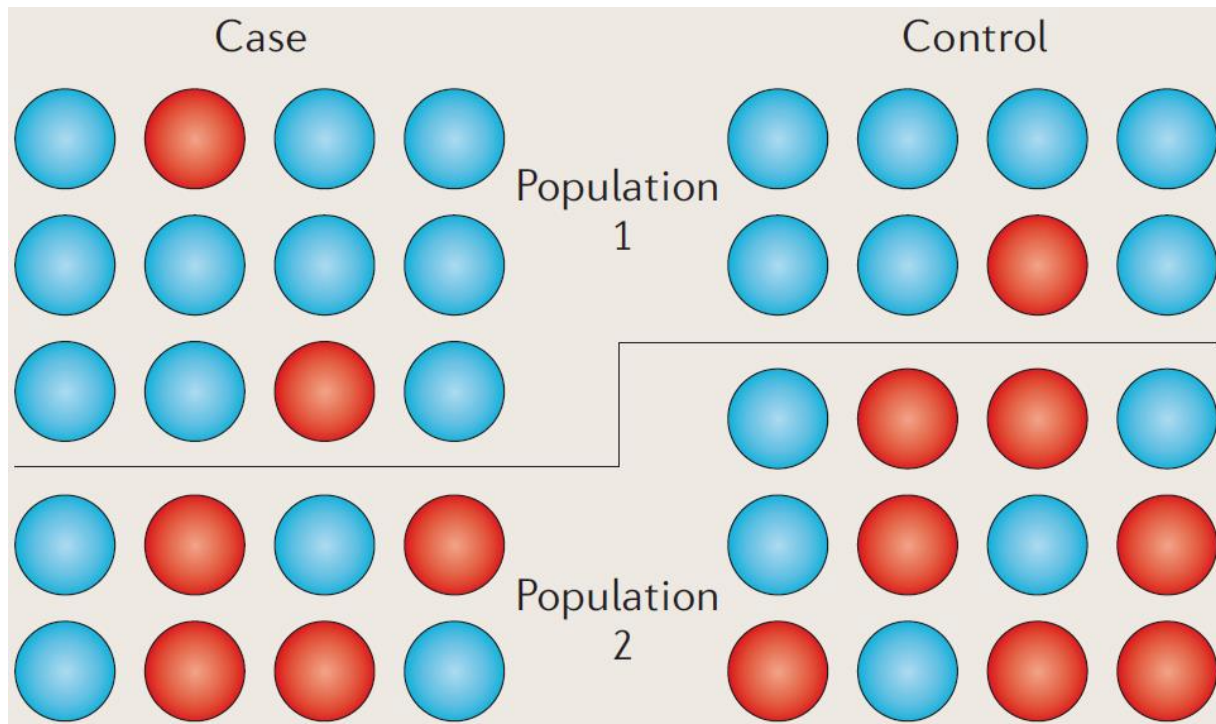
- Take overall average of all SNPs of how many alleles are shared.
- E.g., parent-offspring never shares zero alleles.
- HapMap was supposed to be unrelateds (this was a bit of a surprise!)

GWAS analysis

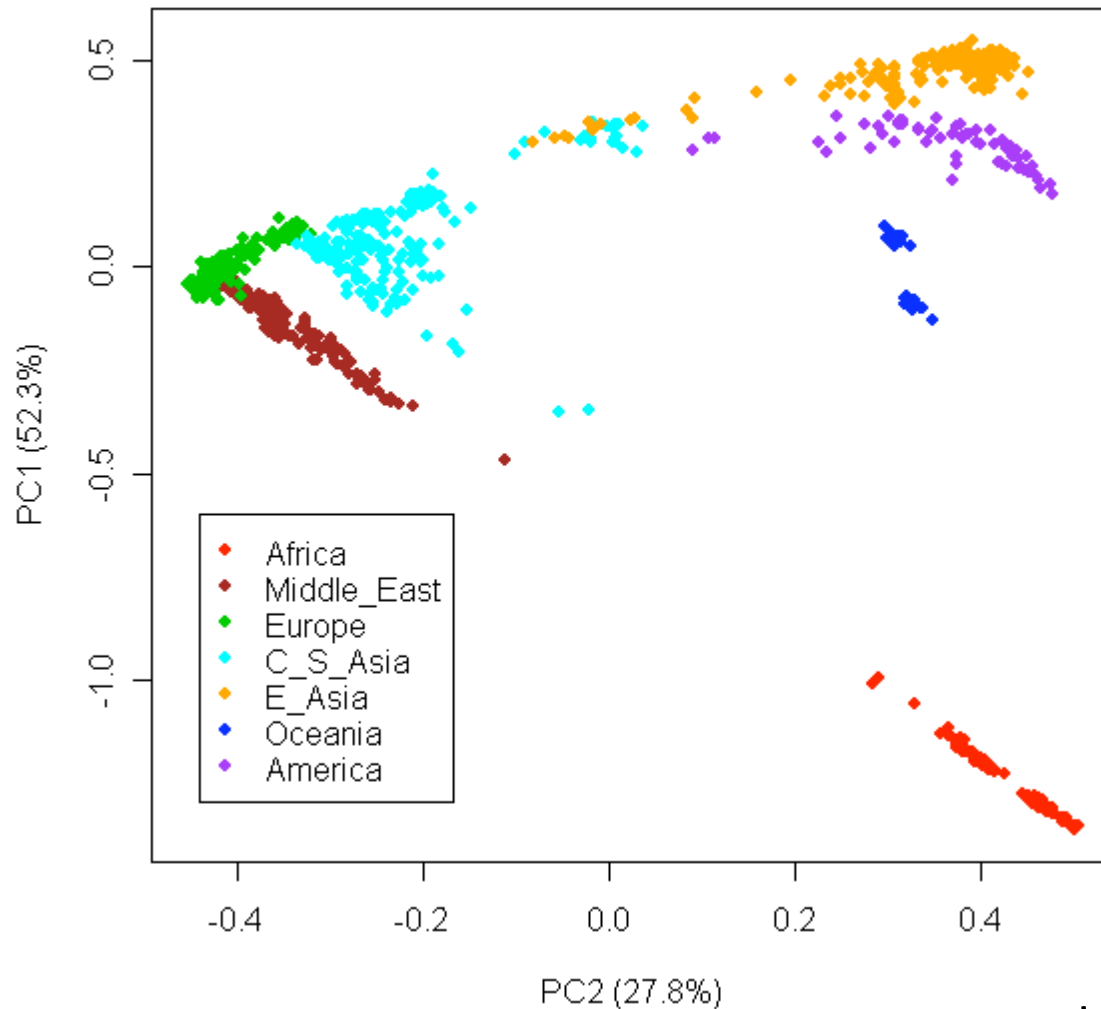
- Most common approach: Model each SNP separately
- Additive coding of SNP most common, just a covariate in a regression framework
- Dichotomous phenotype: logistic regression
- Continuous phenotype: linear regression
- Correct for multiple comparisons
 - e.g., Bonferroni, 1 million gives $\alpha=5 \times 10^{-8}$
 - [Often use 5×10^{-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
 - principal components (PC's), on best performing SNPs
 - software usually does LD filter (e.g., Eigensoft)

Recap of population substructure

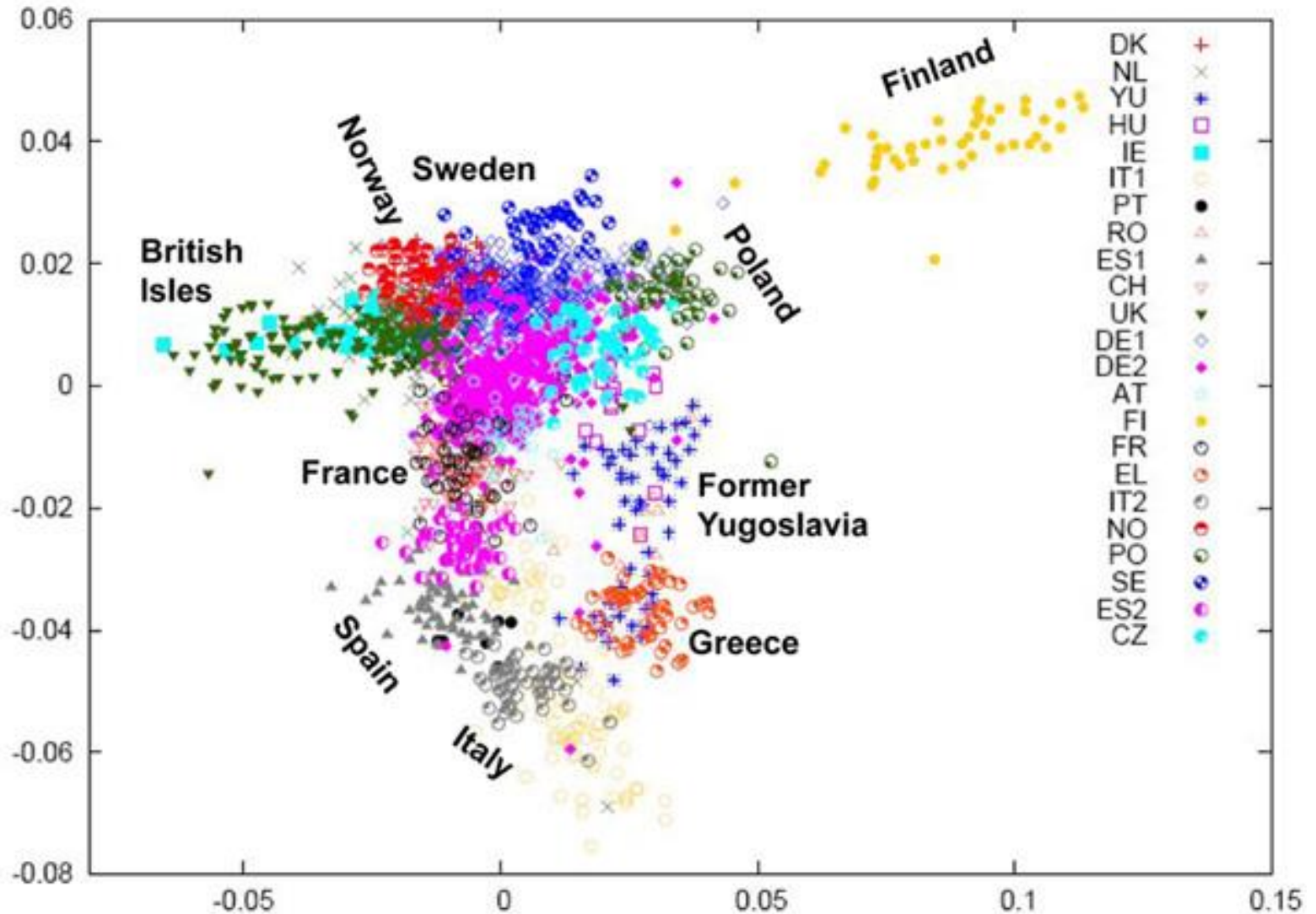
- Two populations have different disease frequency, and different allele frequency (Recall Pima & Papago example).
- Association picks up they are different populations!



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



PCs pick up fine population structure

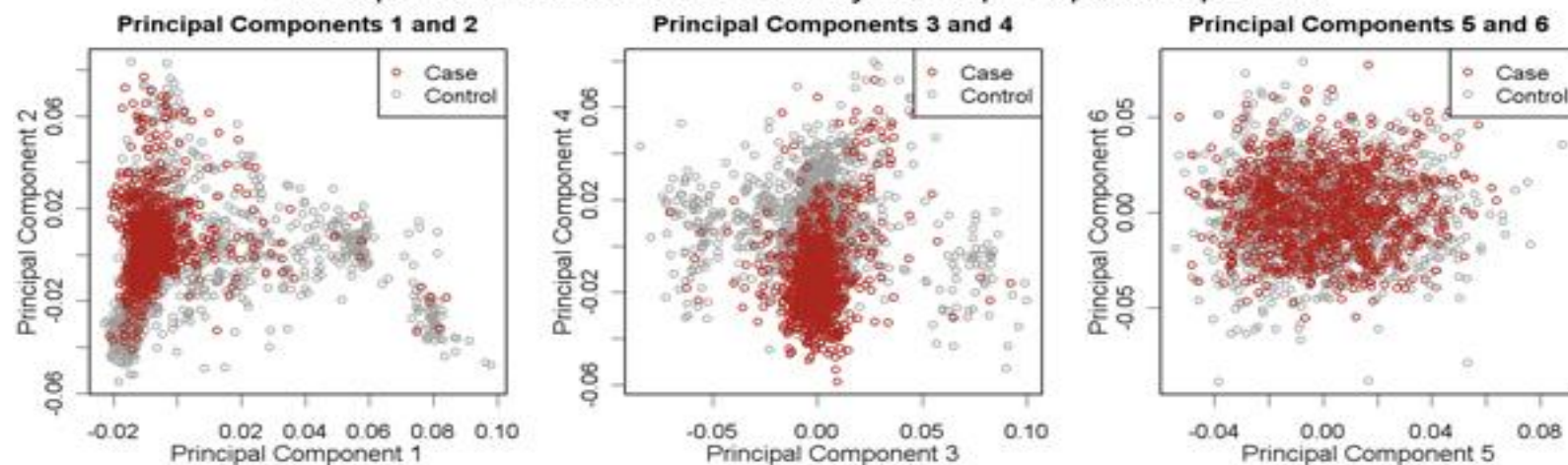


- Razib, Current Biology 2008

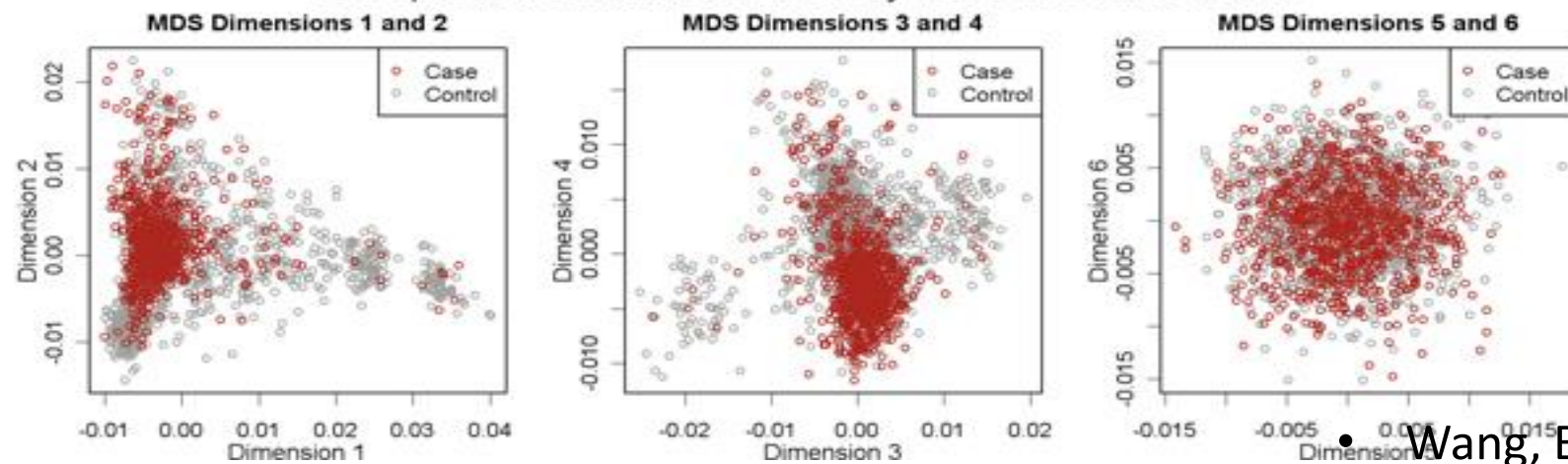
Adjusting for PC's

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...]

A. Population structure identified by first 6 principal components



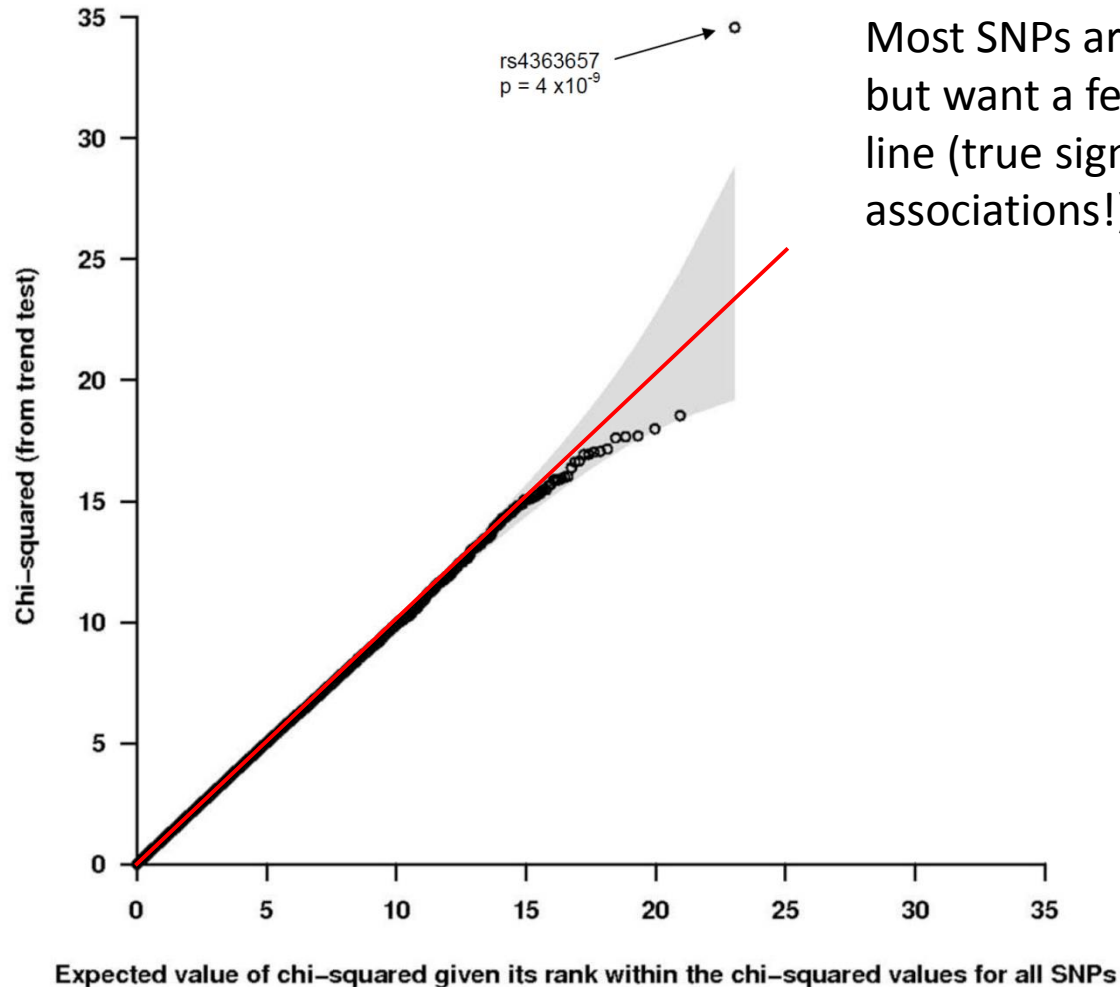
B. Population structure identified by first 6 MDS dimensions



After run analysis

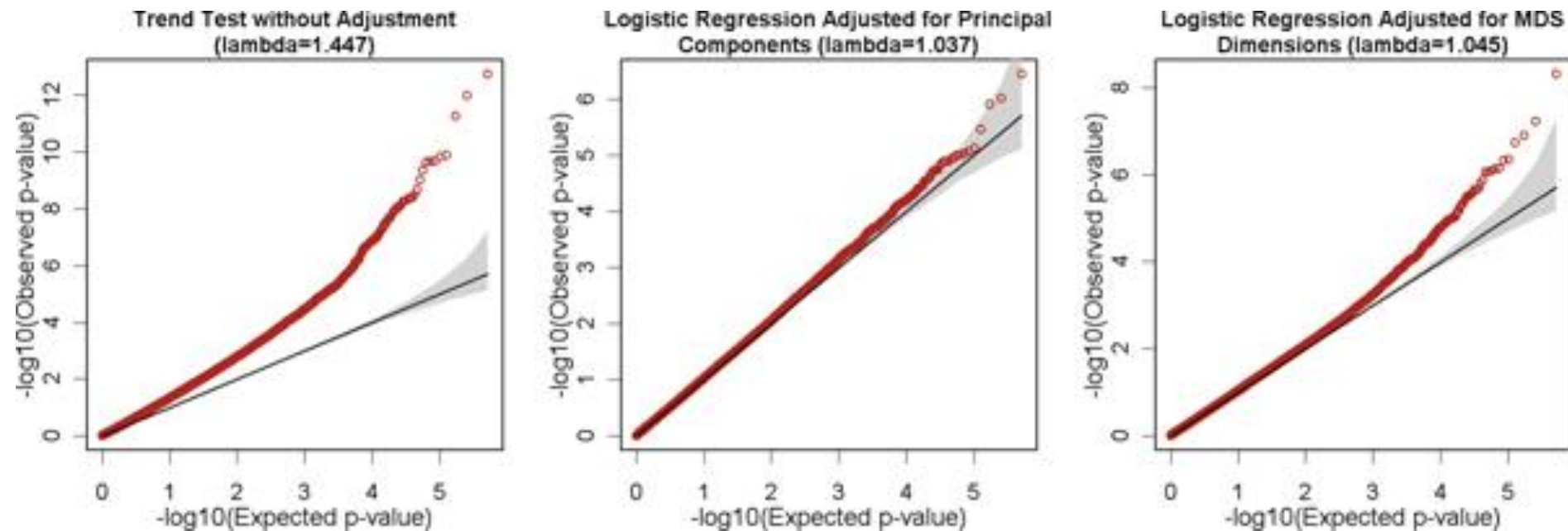
- Assess the fit with QQ-plot (saw last lecture)
- Look at results in Manhattan plots
- **Replication** of hits in a separate cohort

Quantile-quantile (QQ) plot review



Most SNPs are on the line, but want a few hits off the line (true significant associations!)

QQ-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.

Manhattan plot example: Kaiser GWAS of Prostate Cancer

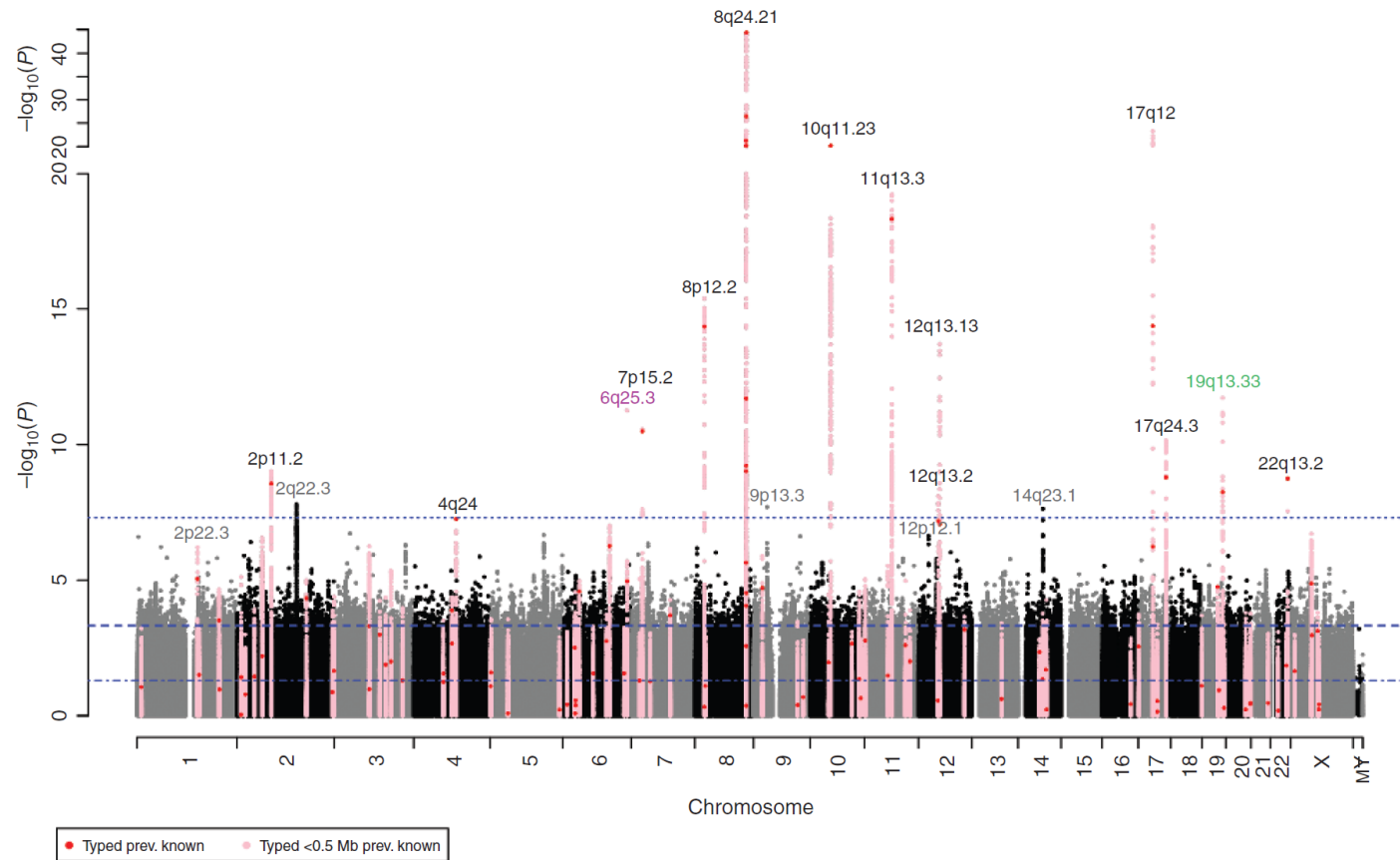
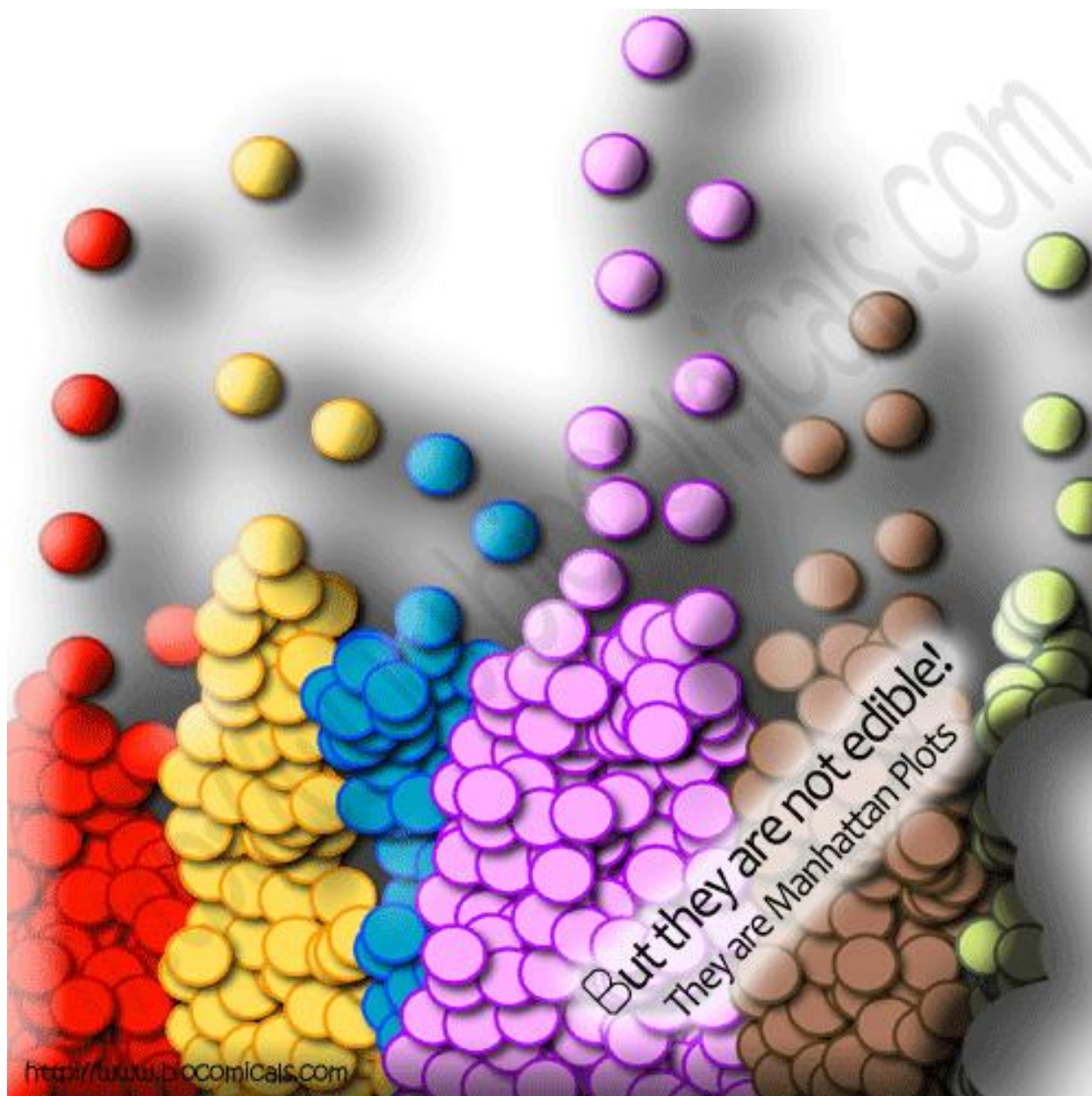


Figure 1. Results from a GWAS of prostate cancer in the KP population (8,399 cases and 38,745 controls), highlighting key chromosomal regions. P values are for variant associations with prostate cancer from transethnic fixed-effects meta-analysis of four race/ethnicity GWAS (non-Hispanic white, Latino,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



Replication, Replication, Replication

- To replicate:
 - Association test for replication sample significant at $0.05/\{\text{Number of SNPs replicating}\}$ 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

Meta-analysis

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

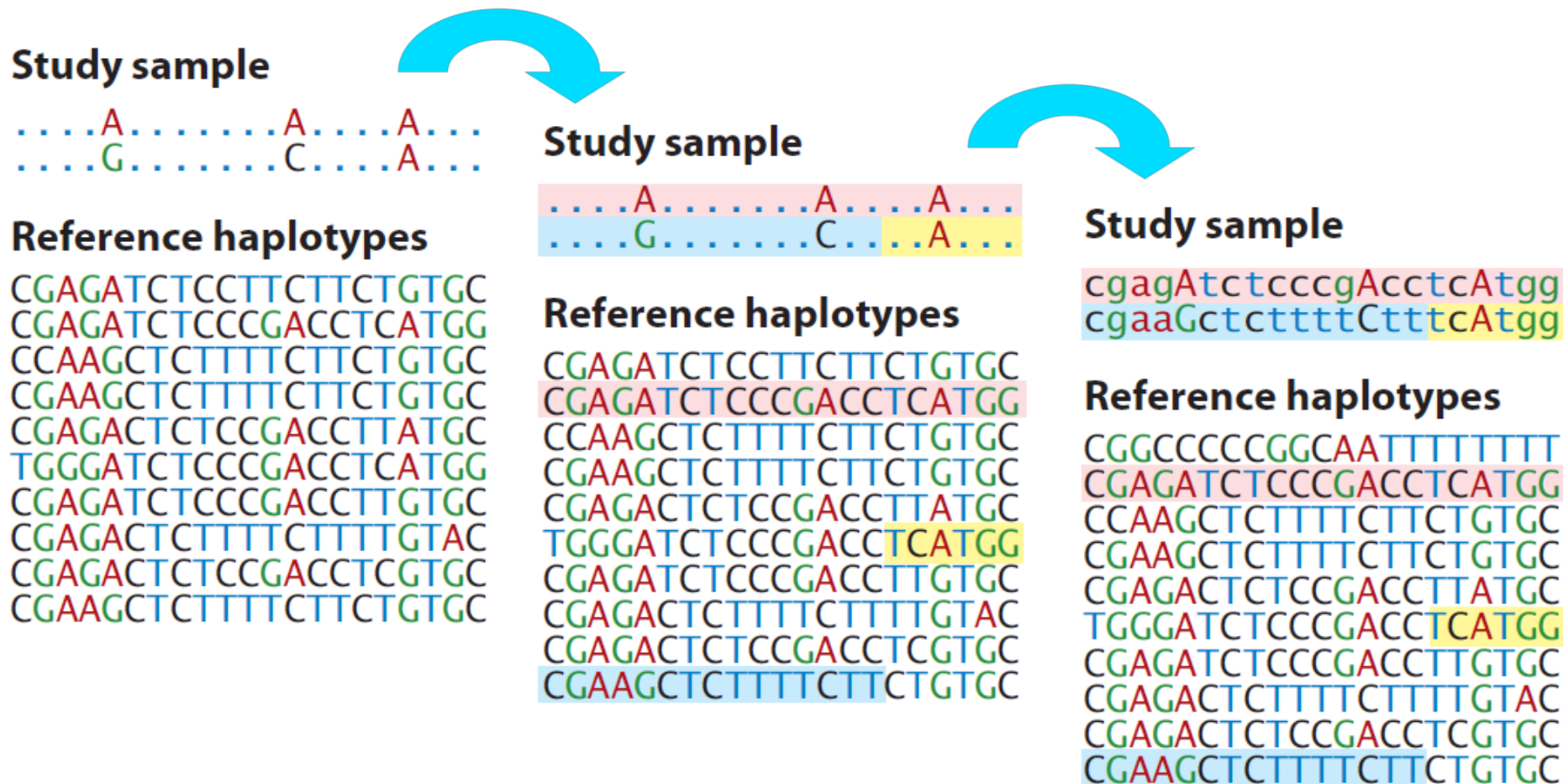
Variant	Risk allele	Ref. allele	Group	Risk AF	No conditioning		Conditioning on other 6q25.3 SNPs		
					OR (95% CI)	P	OR (95% CI)	P	r ² _{info}
Indel: rs4646284	TG (indel)	T	KP NH white	0.298	1.17 (1.12–1.23)	6.7 × 10 ^{−11}	1.16 (1.10–1.23)	3.3 × 10 ^{−7}	0.91
			KP Latino	0.250	1.13 (0.94–1.36)	0.20	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}	—

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

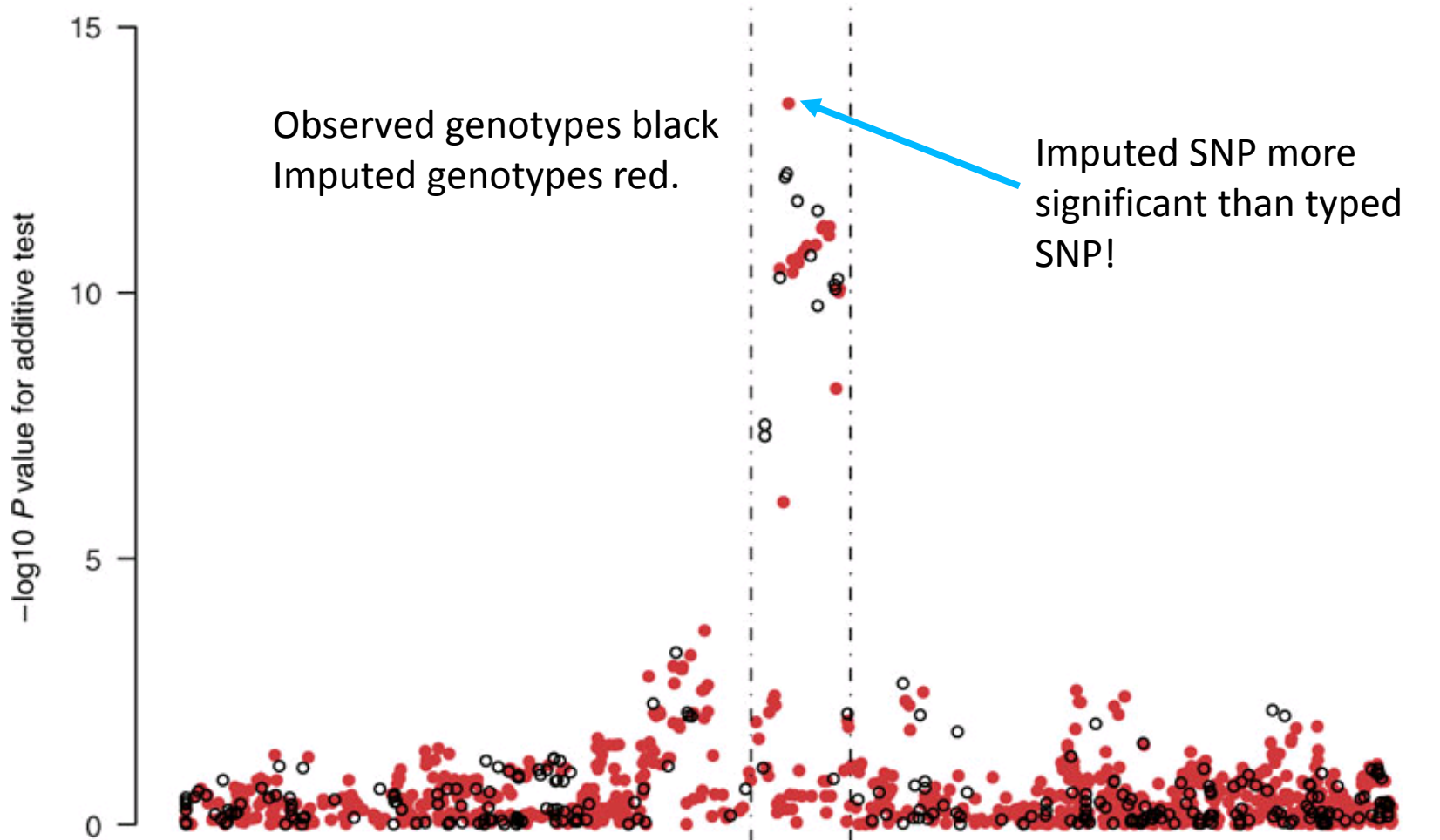
Generalization of Tag SNPs: Imputed SNPs using a Reference Panel

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



Imputation Application

TCF7L2 gene region & T2D from the WTCCC data



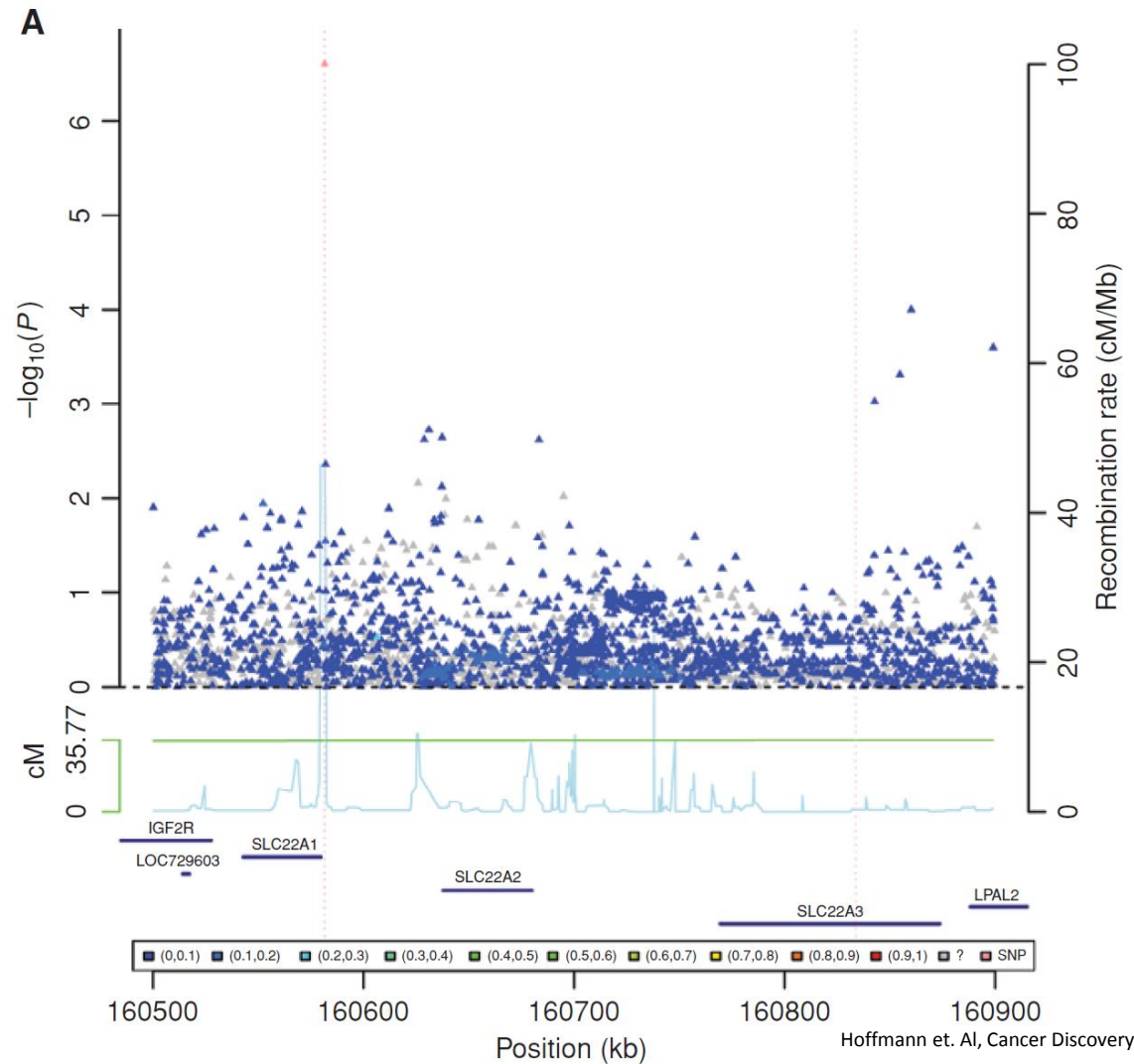
Marchini Nature Genetics 2007

<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
three slides ago)



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



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. Sakoda³, Ling Shen³, Eric Jorgenson³, Laurel A. Habel³, Jinghua Liu², Mark N. Kvale², Maryam M. Asgari³, Yambazi Banda², Douglas Corley³, Lawrence H. Kushi³, Charles P. Quesenberry Jr.³, Catherine Schaefer³, 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

- Review / Overview
- Design
- Analysis
 - QC
 - Prostate cancer example
 - Imputation
 - Replication & Meta-analysis
- Advanced analysis
 - Limitations & “missing heritability”
 - Gene/pathway tests
 - Polygenic models

Limitations of GWAS

- Not very predictive

Example:

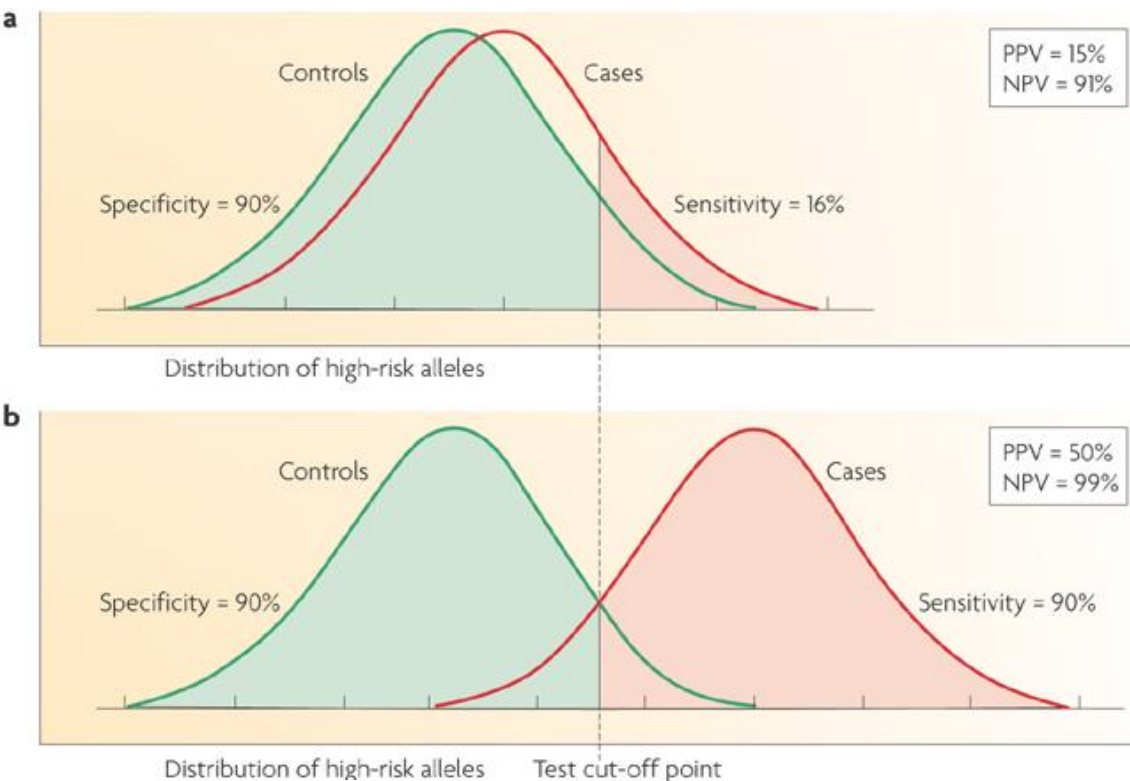
AUC for Breast Cancer Risk

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



Limitations of GWAS

- Not very predictive
- Explain little heritability
- Focus on common variation
- Many associated variants are not causal

Where's the heritability?

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	

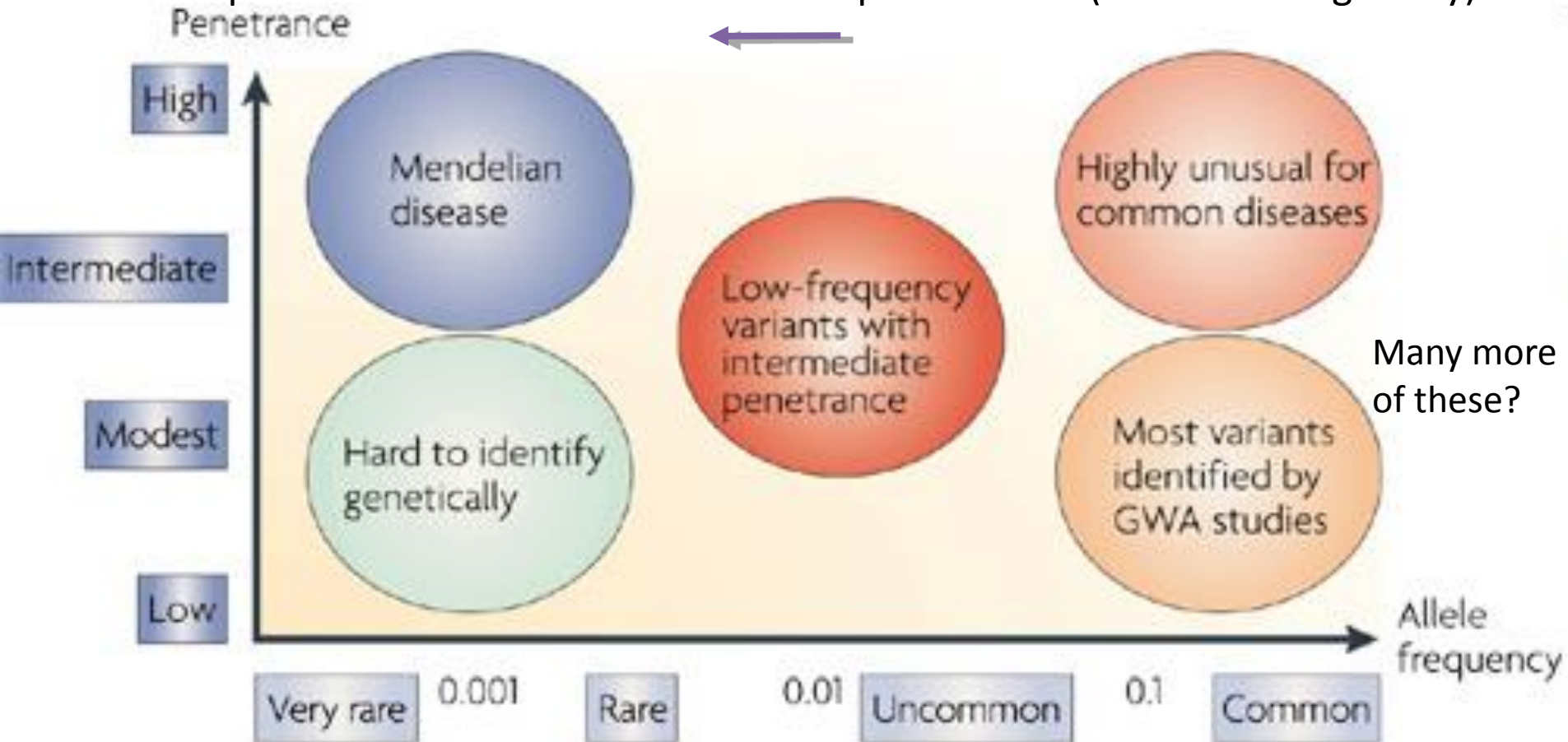
^a Proportion of phenotypic variance or variance in liability explained by genome-wide-significant and validated SNPs. For a number of diseases, other parameters were reported, and these were converted and approximated to the scale of total variation explained. Blank cells indicate that these parameters 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.

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

Will GWAS results explain more heritability?

Possibly, if...

1. Causal SNPs not yet detected due to power issues – weak effects
 - Previous GWAS designed/powerd 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, gene-environment interaction, etc.

“Chip” 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

Gene/pathway-based tests

- Various ways of collapsing the genotype information in multiple genes
 - Less multiple comparison adjustment
- $\text{logit}(\text{Prob}(y=1 | \mathbf{x}, \mathbf{c})) = \alpha + \beta\mathbf{x} + \gamma\mathbf{c}$
e.g. $= \alpha + \beta_1 x_1 + \dots + \beta_{12} x_{12} + \{\text{PC's, other covariates}\}$
 - y : disease status)
 - $\mathbf{x}$ is a vector of genotypes (e.g., a gene, or a pathway)
 - $\mathbf{c}$ is a vector of covariates
- $H_0: \beta=0$
- Rare variant “burden” tests are along these lines

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

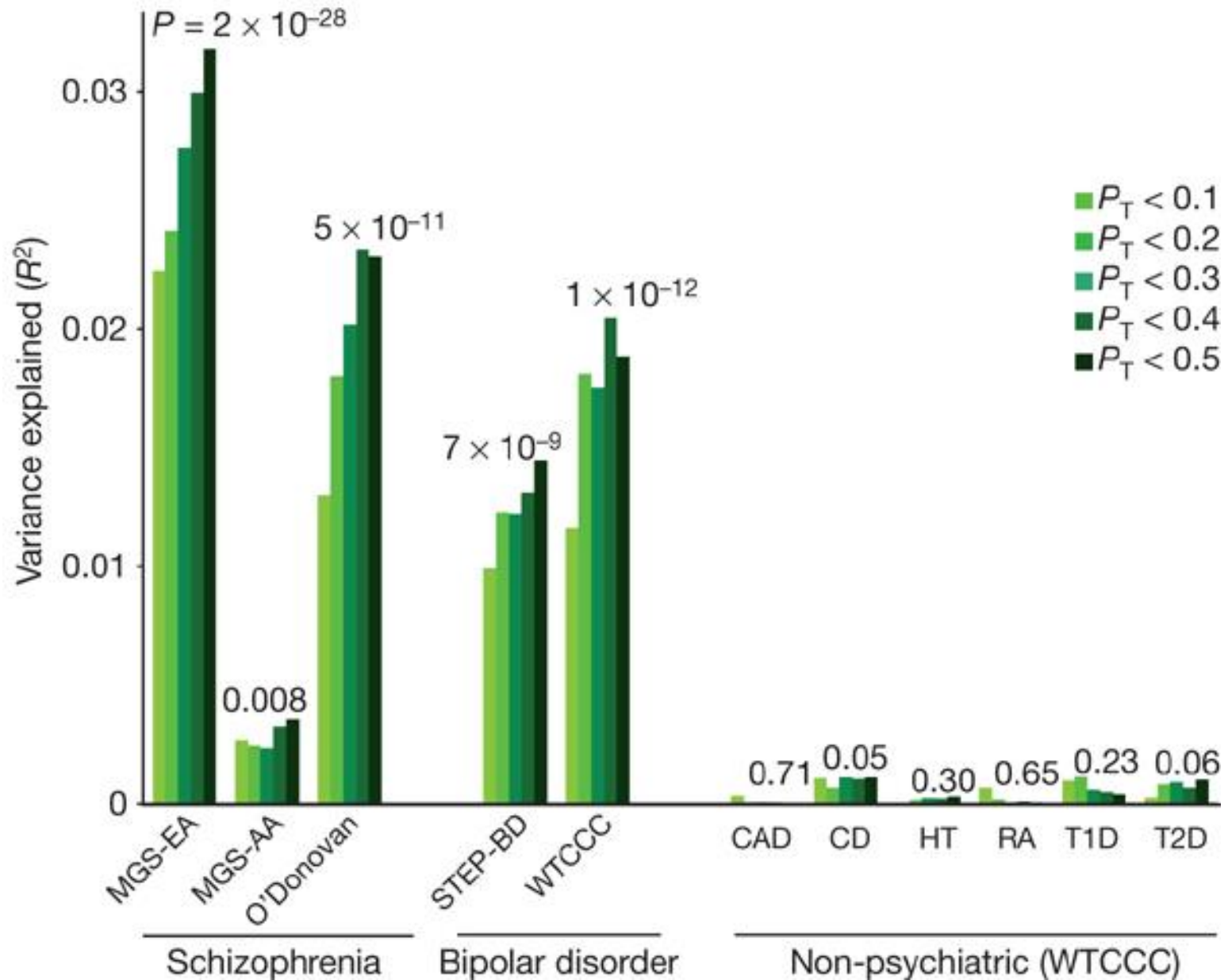
- Many weak associations combine to risk?
- Score model (use all GWAS SNPs):

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

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**

Application of Model



Other things than SNPs

- Copy number variants (CNVs)
- Epigenetics, e.g., methylation, histone modification
- Gene expression (RNA levels)
- Proteomics (measures of proteins in cells)
- ...