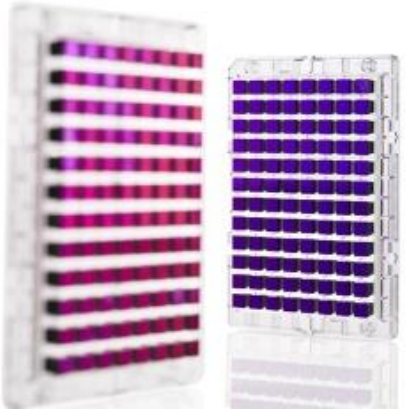
Genome-wide association studies (GWAS)



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

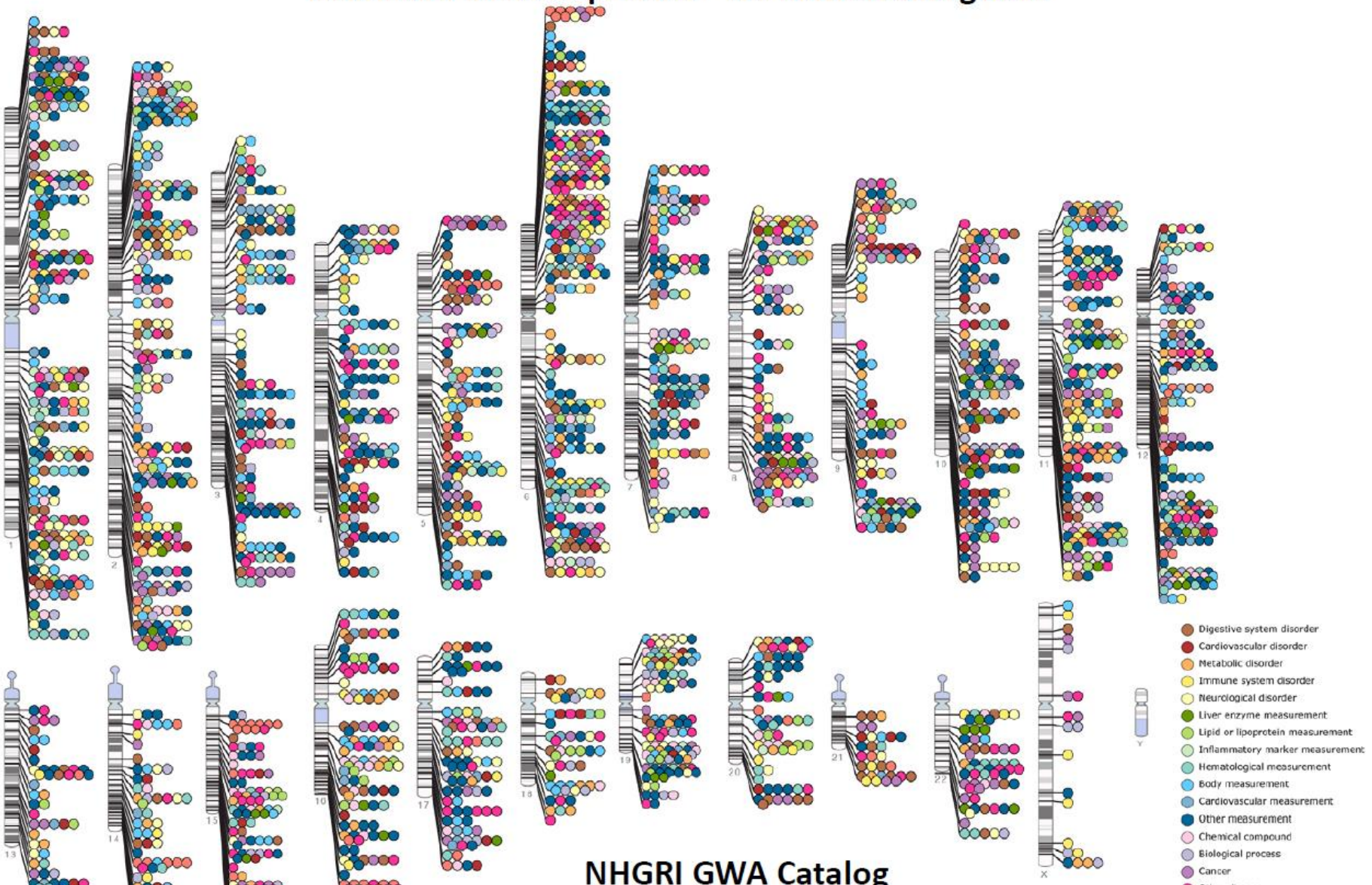
- Review / Overview
- Design
- Analysis
  - QC
  - Prostate cancer example
  - Imputation
  - Replication & Meta-analysis
- Advanced analysis
  - Limitations & “missing heritability”
  - Gene/pathway tests
  - Polygenic models

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 $p \leq 5X10^{-8}$ for 18 trait categories



NHGRI GWA Catalog

www.genome.gov/GWAStudies

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

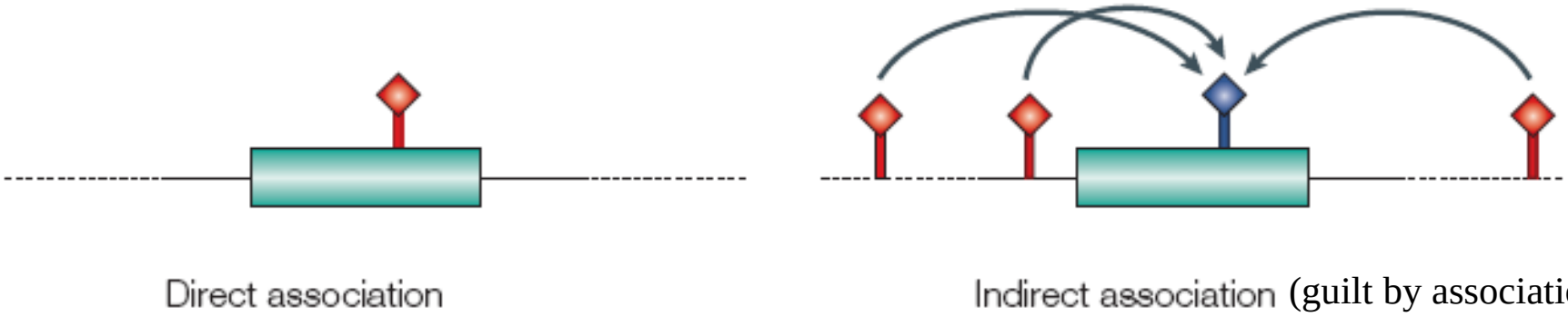
National Human Genome Research Institute

EMBL-EBI

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



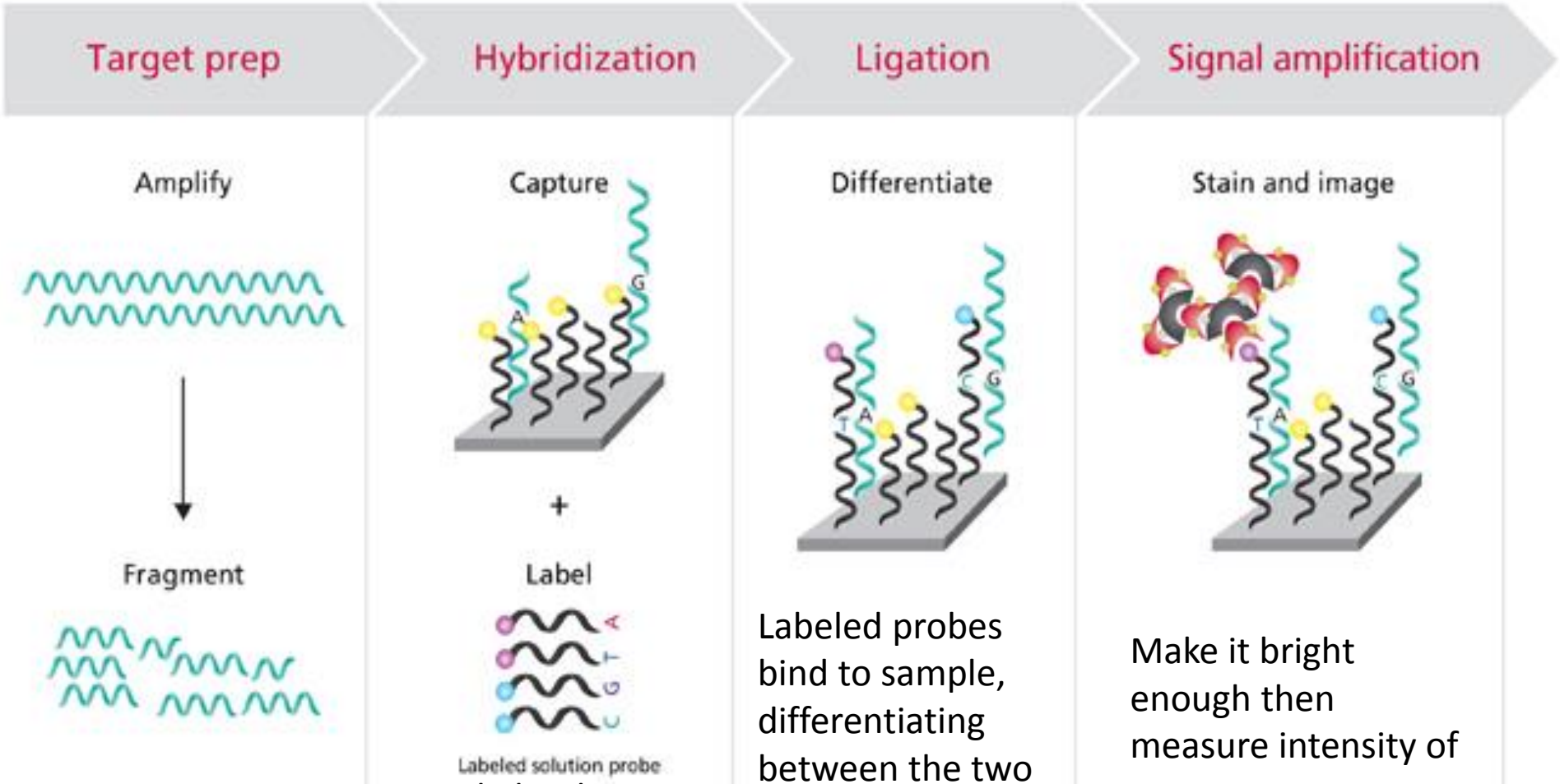
Disease locus directly typed

Marker correlated with typed disease locus

- 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.

Hirschhorn & Daly, Nat Rev Genet 2005

Technology behind a GWAS Microarray



Sample binds to array

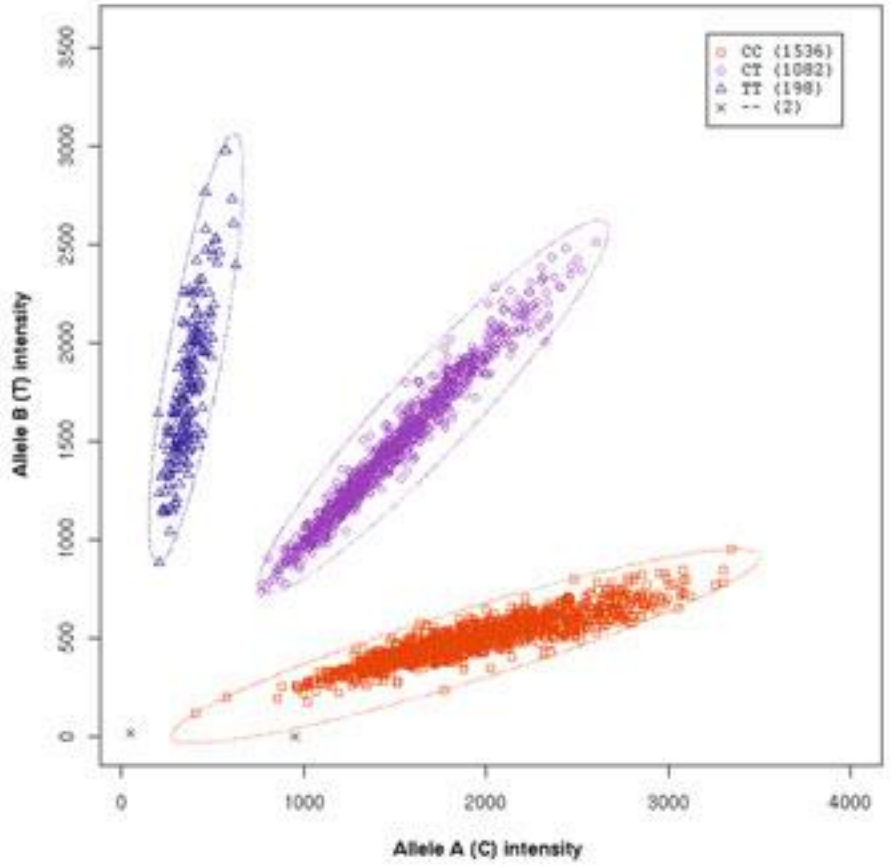
Labeled probes bind to sample, differentiating between the two alleles

Make it bright enough then measure intensity of array

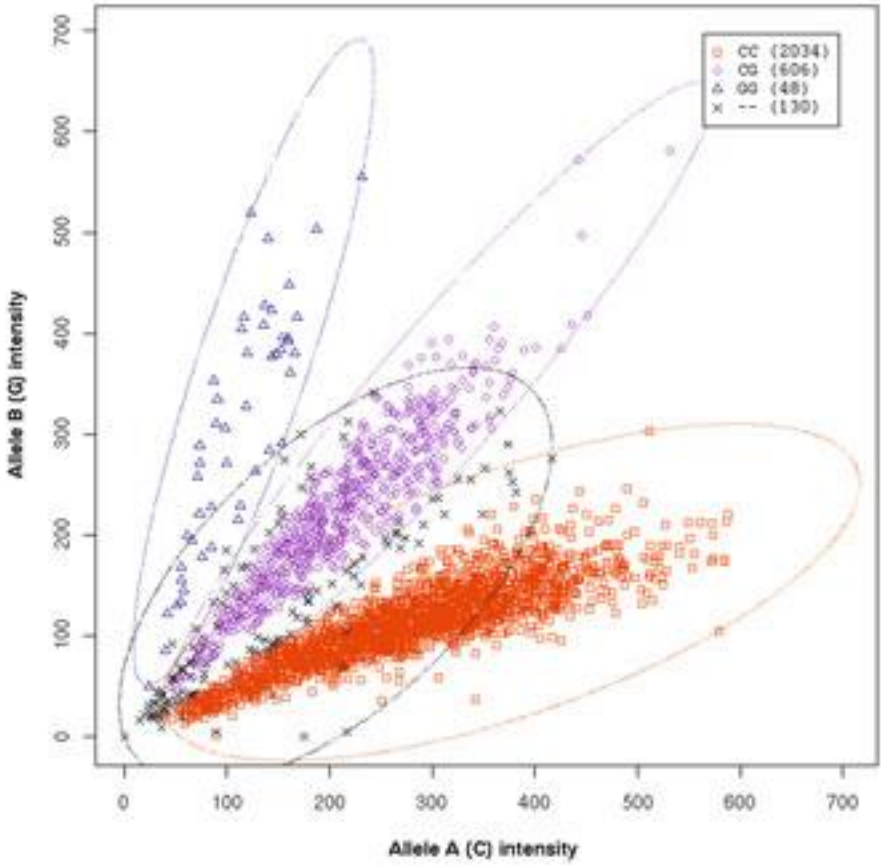
Assay ~ 0.7 - 5M SNPs (keeps increasing)

Affymetrix, http://www.affymetrix.com

Raw microarray data: Genotype calls

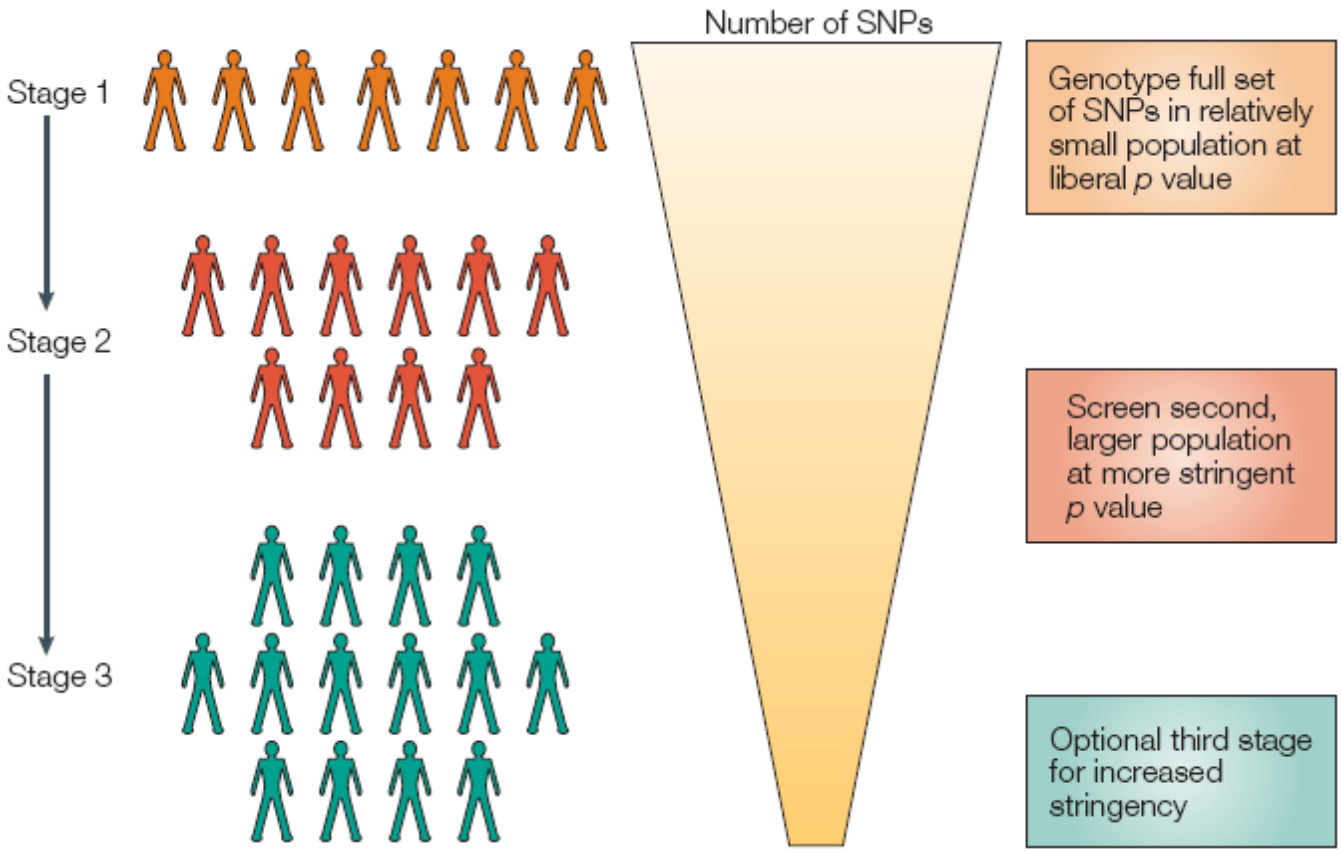


Good calls!



Bad calls!

Genome-wide Assocation 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

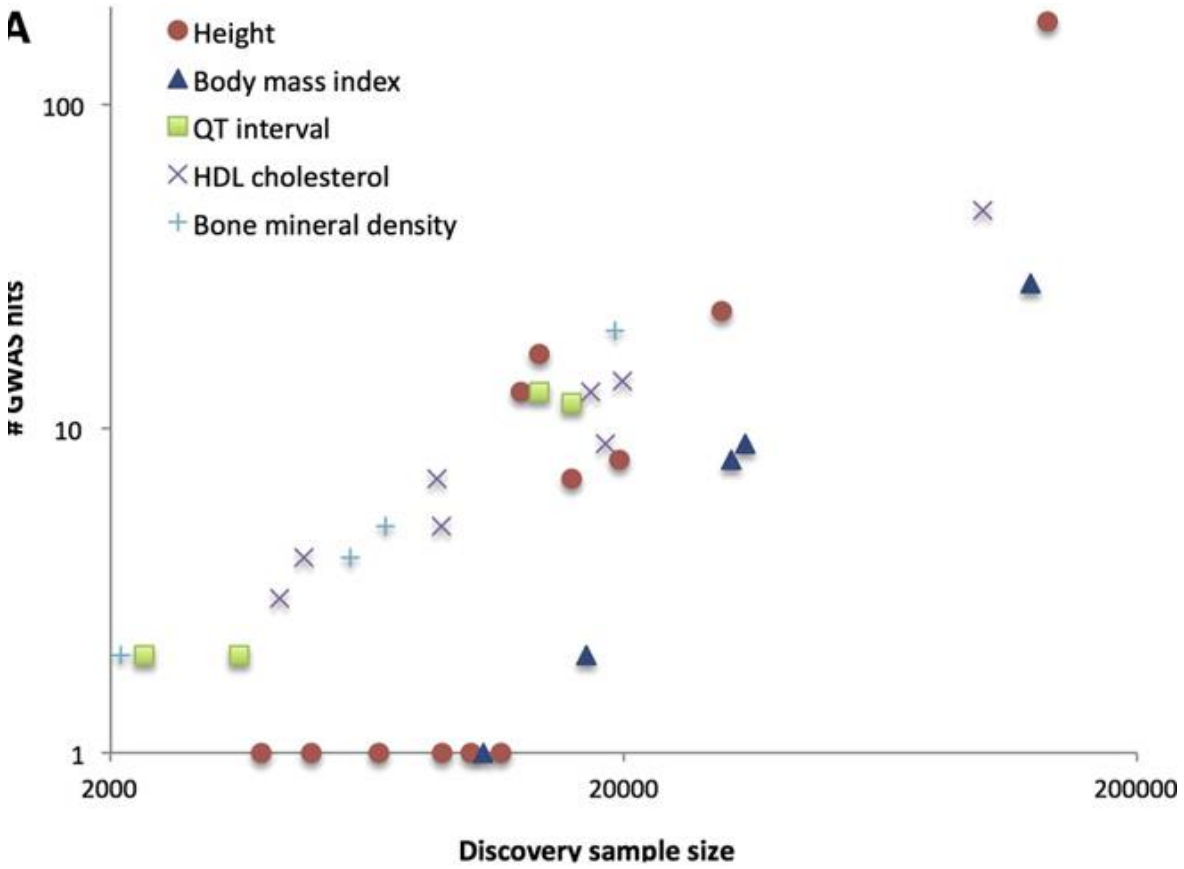
"De novo" discovery (particularly good for rare variants); **%age of exons only** (first step is a pull down that does not capture all exons)

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
- Consortiums



Visscher, AJHG 2012

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.

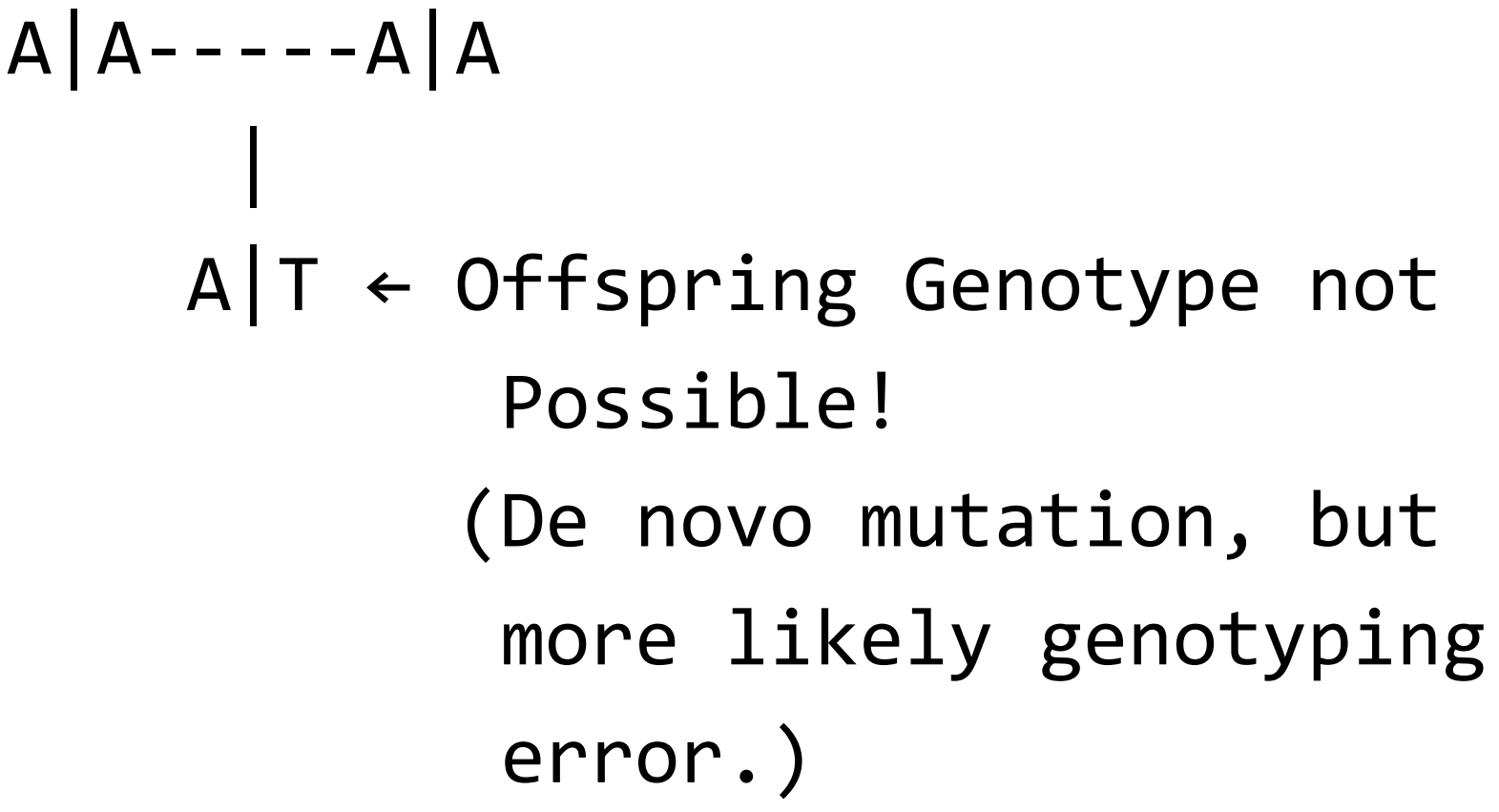
  - MZ twins, or accidentally genotyped same sample twice?
  - Remove first degree, or account for.

•Check genotype gender (mislabeled samples)

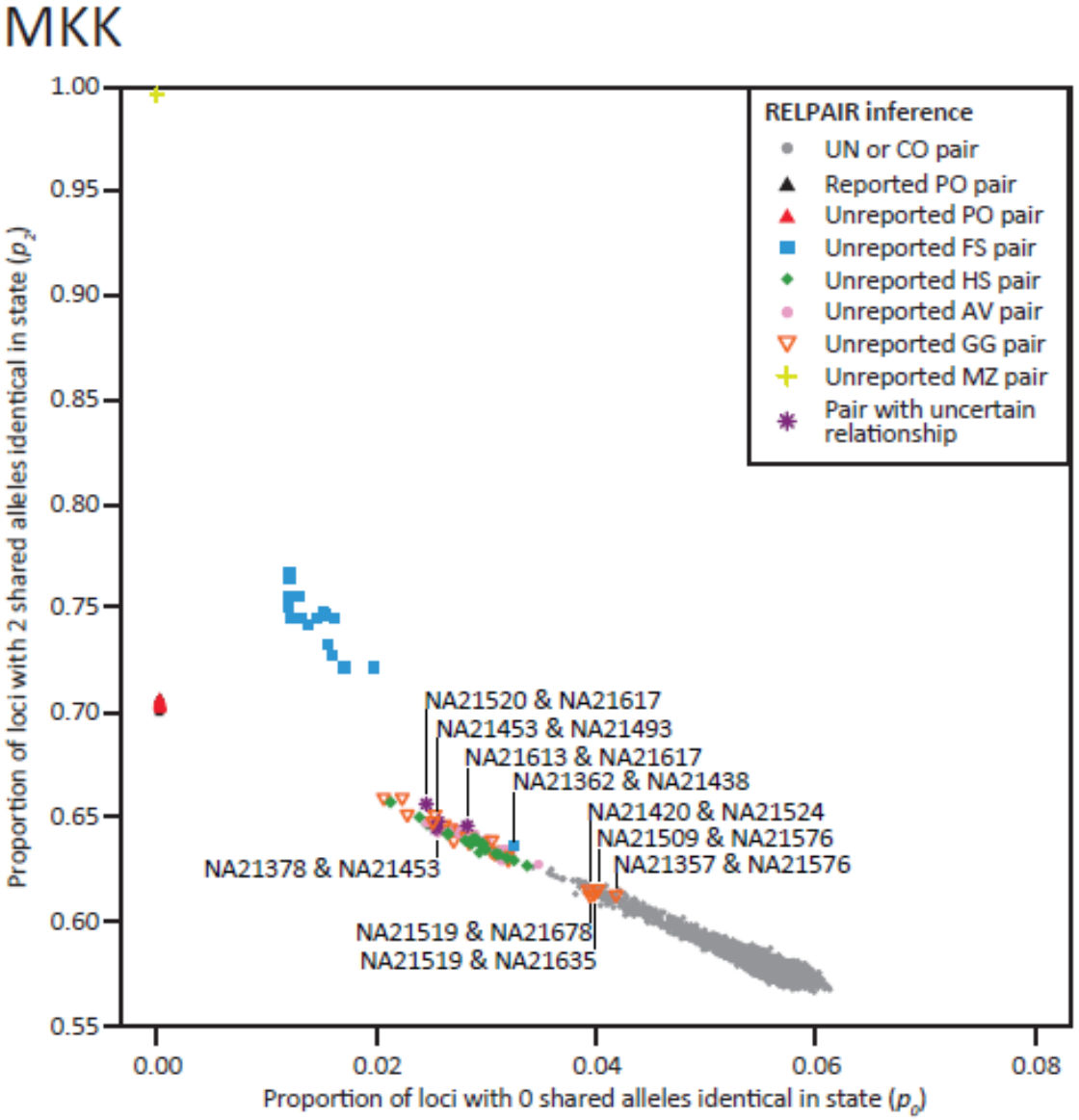
•Filter Mendelian inhertance (family-based, or potentially cryptics, if large enough sample)

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

Filter for Mendelian inheritance



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!)

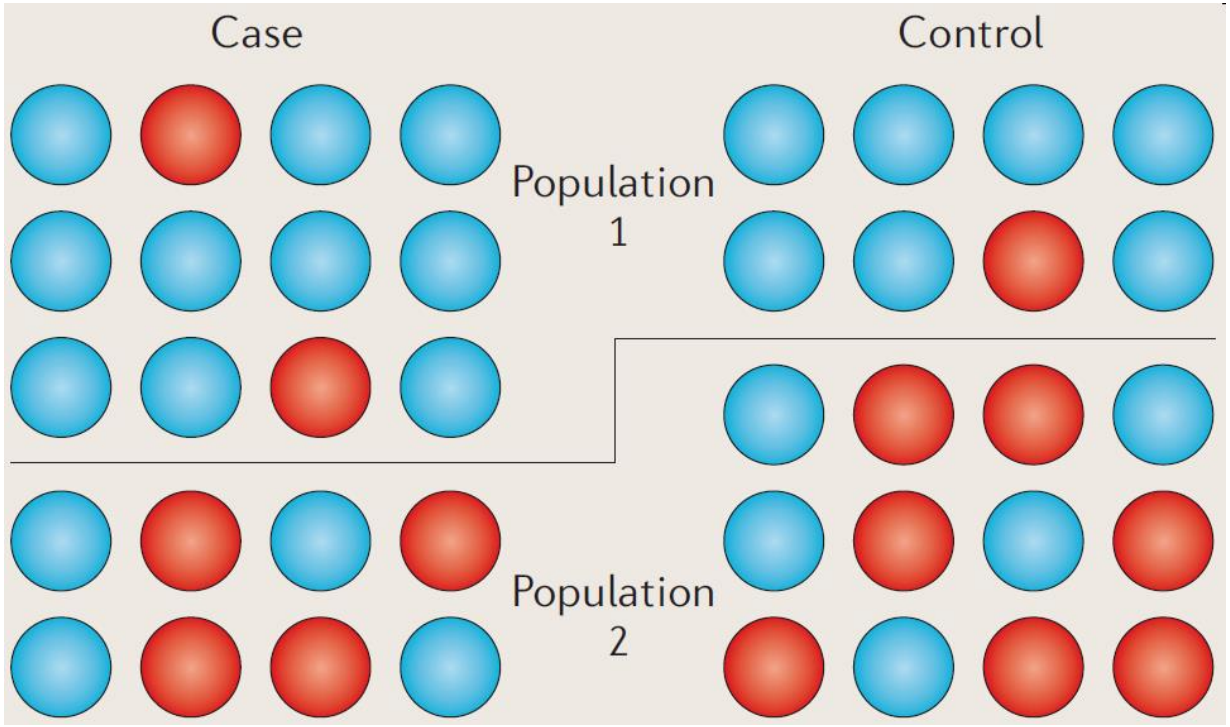
Pemberton et al., AJHG 2010

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 $\alpha=5x10^{-8}$
  - more next time
- Adjust for potential population stratification
  - principal components (PC's), on best performing SNPs
  - software usually does LD filter (e.g., Eigensoft)

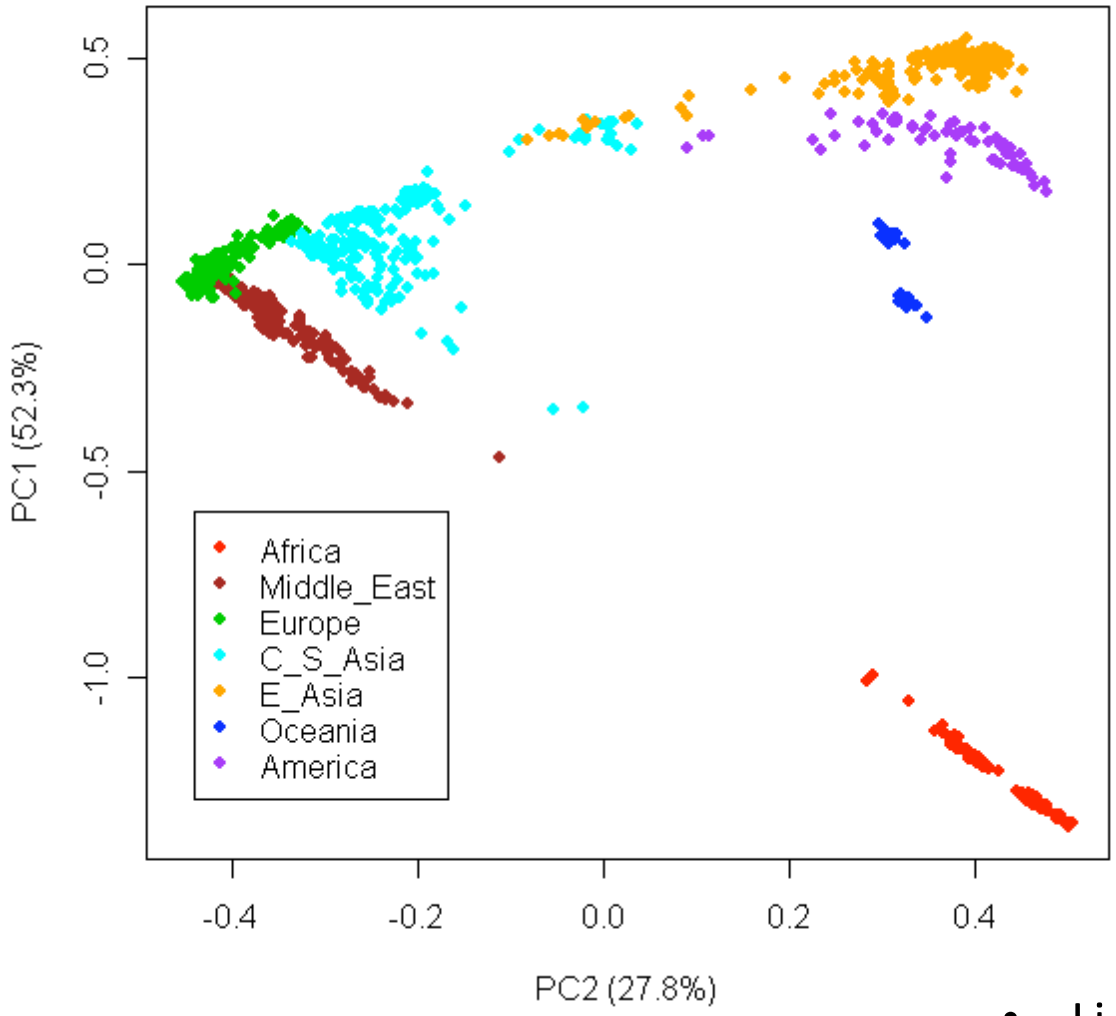
Recap of population substructure

- Two populations have different disease frequency, and different allele frequency.
- Association picks up they are different populations!



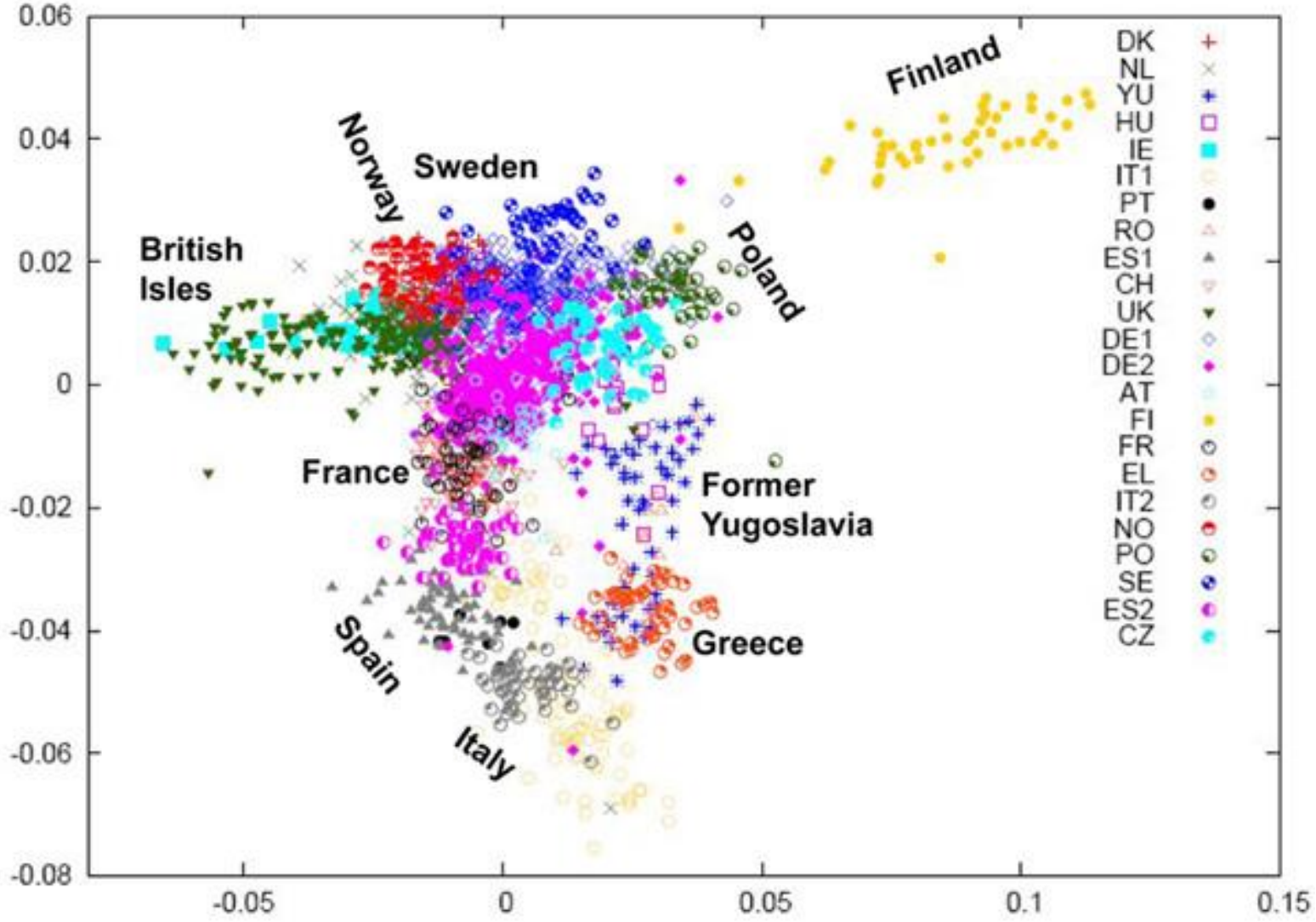
Balding, Nature Reviews Genetics 2010

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



- Li et al., Science 2008

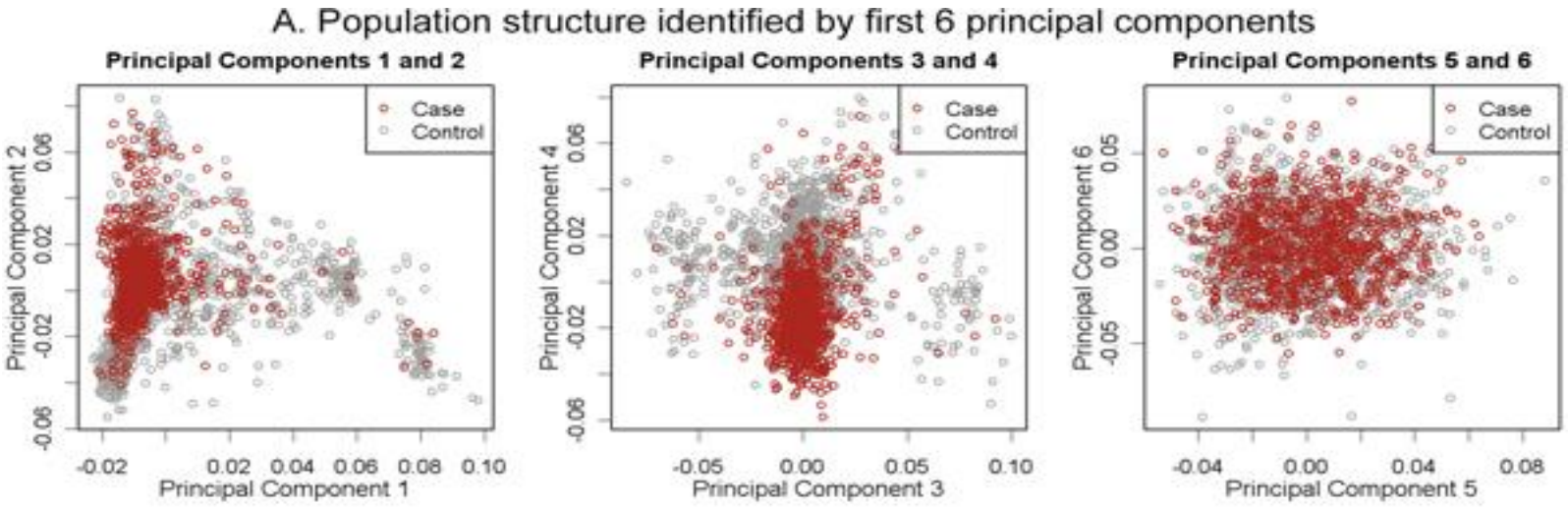
PCs pick up fine population structure

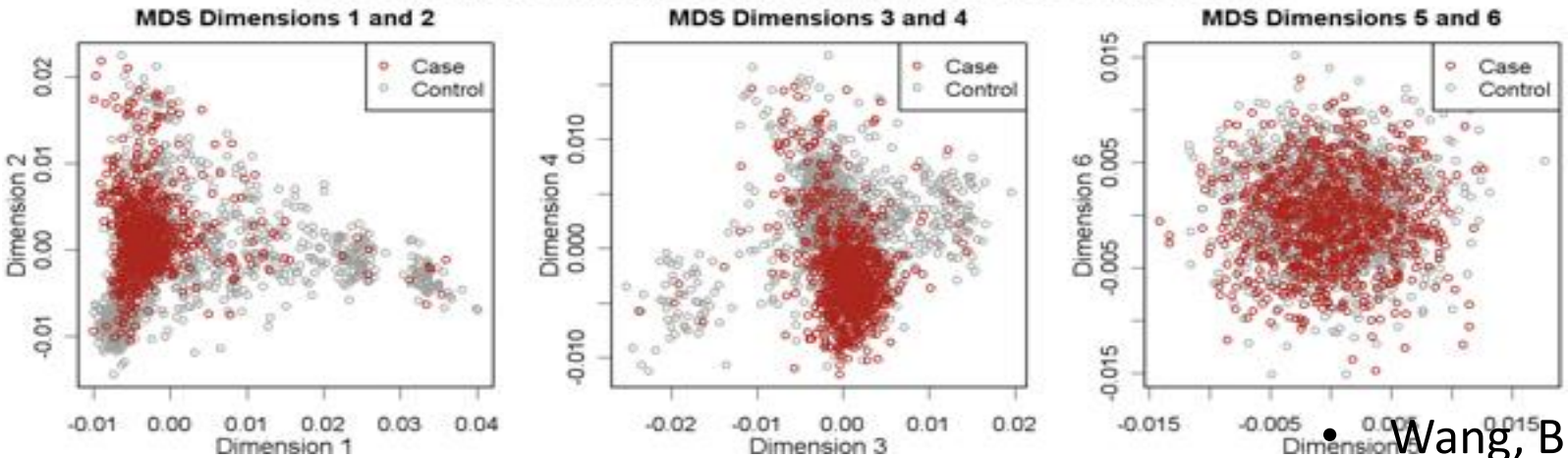


- Razib, Current Biology 2008

Adjusting for PC's

Do not want separate cluster for cases as controls! – Why want individuals from similar population for controls.



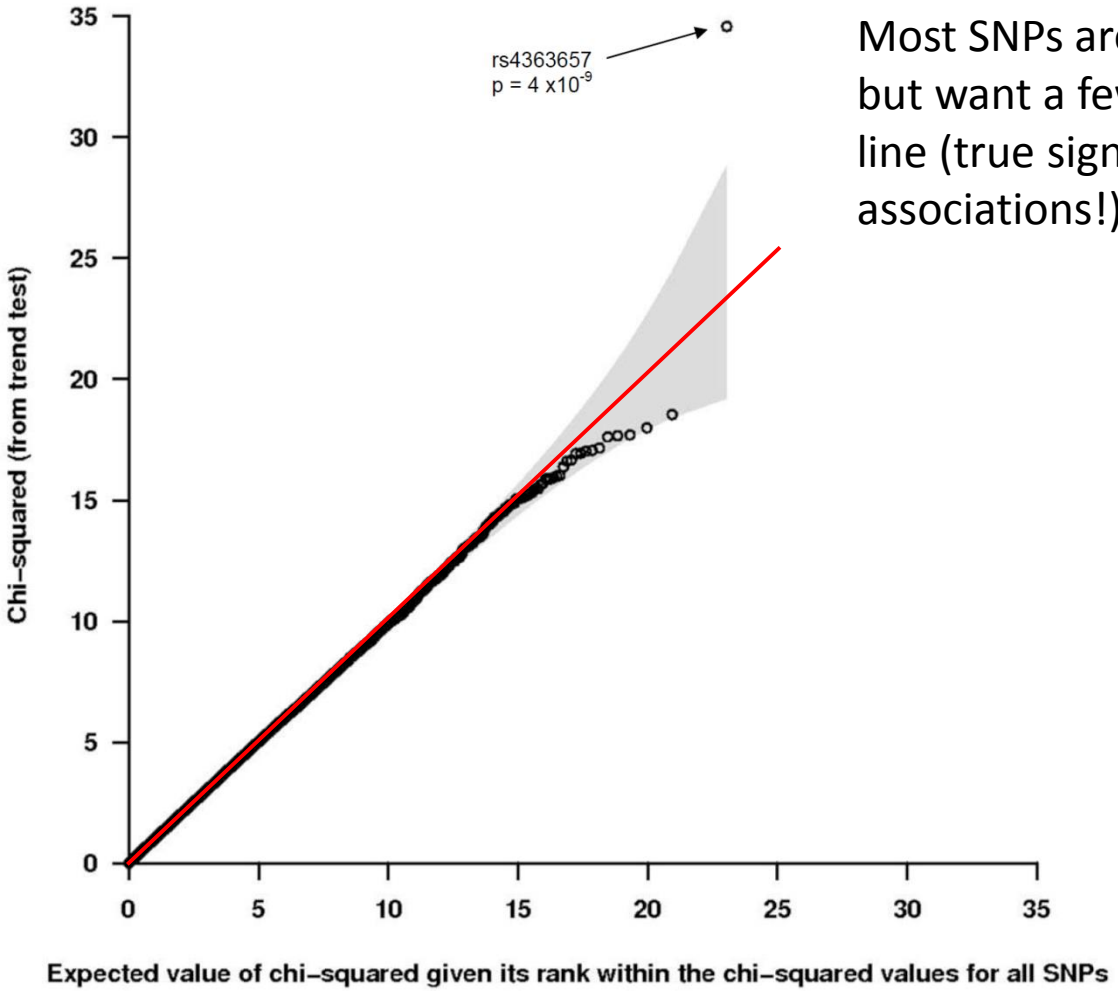


- Wang, BMC Proc 2009

After run analysis

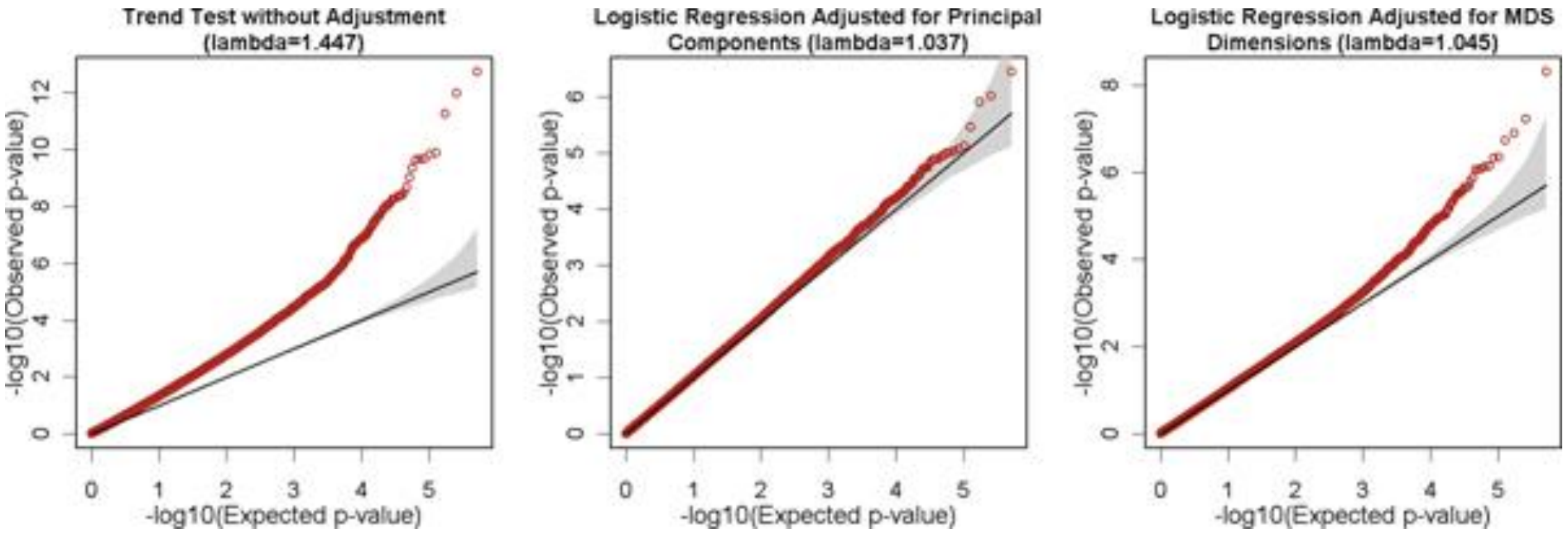
- Assess the fit with QQ-plot
- Look at results in Manhattan plots
- **Replication** of hits in a separate cohort

Quantile-quantile (QQ) plot



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

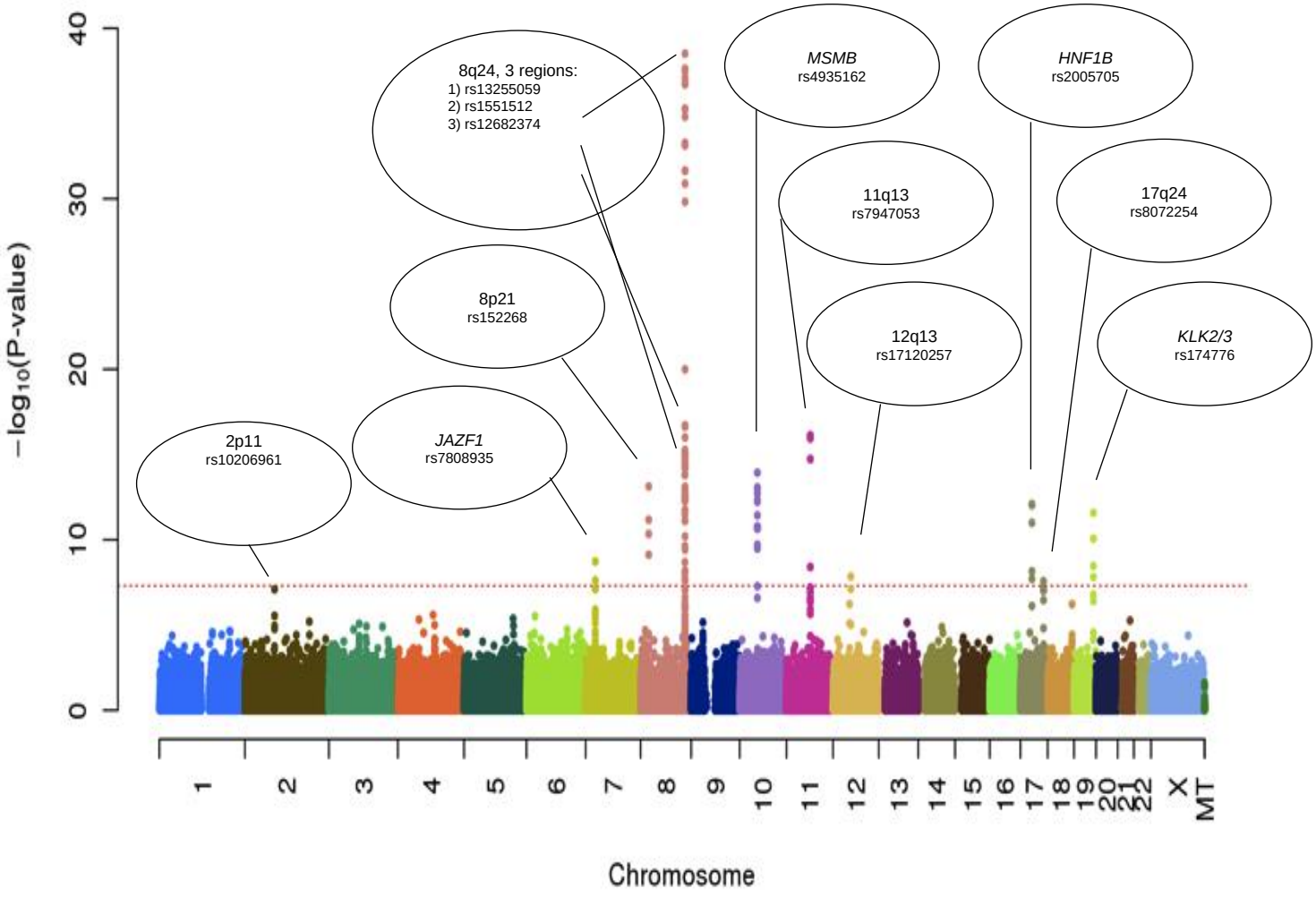
QQ-plots and PC adjustment recap

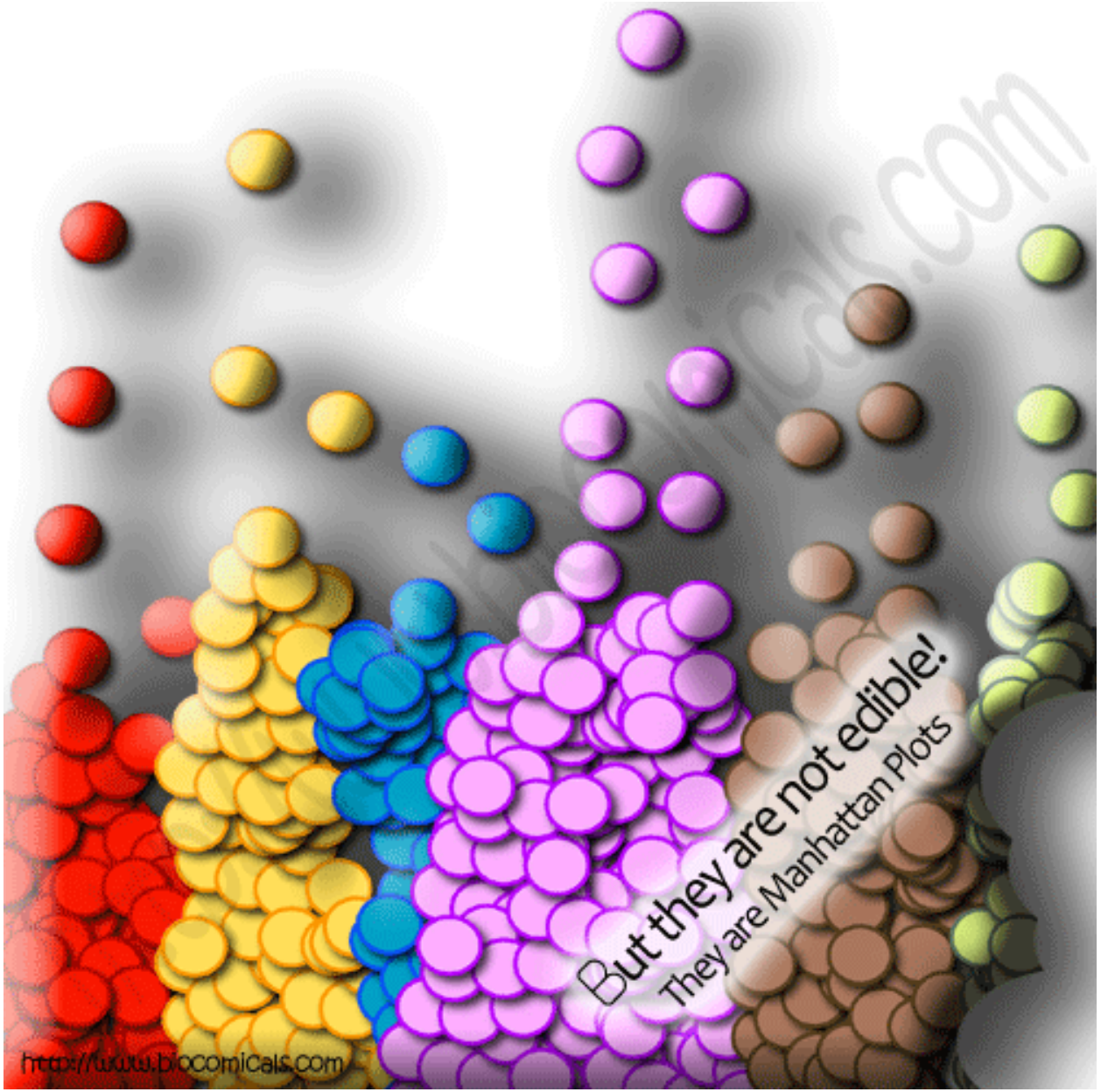


- Exactly on line if no signal at all.
- If don't adjust for PC's, then p-values are all inflated artificially.
- Adjusting for PC's fixes the problem.

Wang, BMC Proc 2009

Manhattan plot example: Kaiser GWAS of Prostate Cancer




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
- 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

Biological, clinical and population relevance of 95 loci for blood lipids

Vol 466 | 5 August 2010 | doi:10.1038/nature09270

A list of authors and their affiliations appears at the end of the paper.

Table 1 | Meta-analysis of plasma lipid concentrations in >100,000 individuals of European descent.

Locus	Chr	Lead SNP	Lead trait	Other traits	Alleles/MAF	Effect size	*P*	eQTL	CAD	Ethnic
LDLRAP1	1	rs12027135	TC	LDL	T/A/0.45	−1.22	4×10^{-11}	Y		+++?
PABPC4	1	rs4660293	HDL		A/G/0.23	−0.48	4×10^{-10}	Y		++++
PCSK9	1	rs2479409	LDL	TC	A/G/0.30	+2.01	2×10^{-28}			++++
ANGPTL3	1	rs2131925	TG	TC, LDL	T/G/0.32	−4.94	9×10^{-43}	Y		++++
EVI5	1	rs7515577	TC		A/C/0.21	−1.18	3×10^{-8}			+++?
SORT1	1	rs629301	LDL	TC	T/G/0.22	−5.65	1×10^{-170}	Y	Y	++++
ZNF648	1	rs1689800	HDL		A/G/0.35	−0.47	3×10^{-10}			+++−

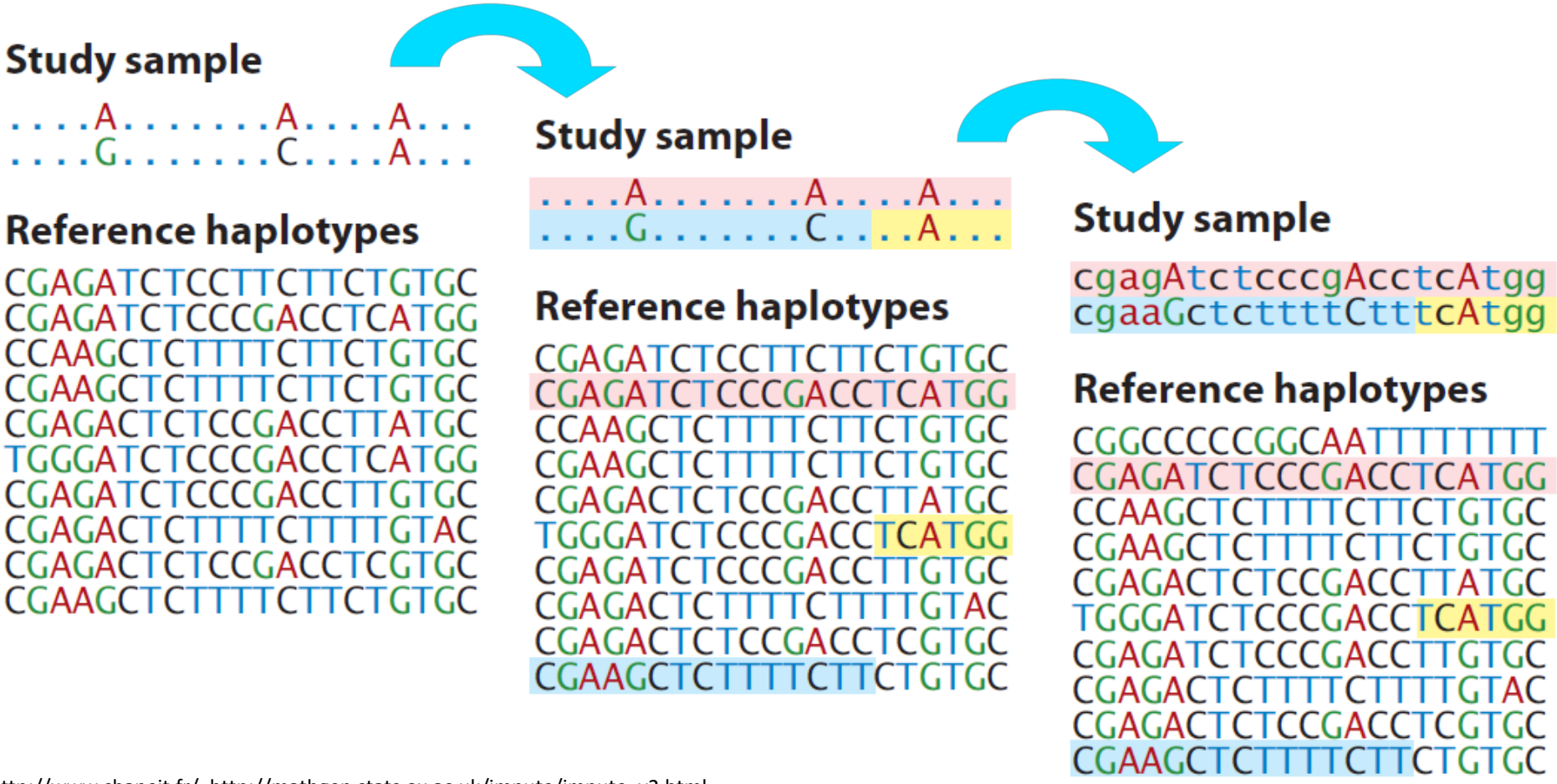
The gene name listed in 'Locus' column is either a plausible biological candidate gene in the locus or the nearest annotated gene to the lead SNP. Listed in 'Lead trait' column is the lipid trait with best *P*-value among all four traits. Listed in 'Other traits' are additional lipid traits with $P < 5 \times 10^{-8}$. Listed in 'Alleles/MAF' column are: major allele, minor allele and minor allele frequency (MAF) within the combined cohorts included in this meta-analysis (alleles designated with respect to the '+' strand; Supplementary Table 2). Numbers in 'Effect size' column are in mg dl^{-1} for the lead trait, modelled as an additive effect of the minor allele. *P*-values are listed for the lead traits. In the 'eQTL' column, 'Y' indicates that lead SNP has an eQTL with at least one gene within 500 kb with $P < 5 \times 10^{-8}$ in at least one the three tissues tested (liver, omental fat, subcutaneous fat). In the 'CAD' column, 'Y' indicates that the lead SNP meets the pre-specified statistical significance threshold of $P < 0.001$ for association with CAD and being concordant between the direction of lipid effect and the change in CAD risk. In the 'Ethnic' column, '+' indicates concordant effect on lead trait of the variant between the primary meta-analysis cohort and the European or non-European group, '−' indicates discordant effect on lead trait, and '?' indicates data not available for the group; in order, the ethnic groups are European, East Asian, South Asian and African American (Supplementary Table 11).
Chr, chromosome.

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?
- 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

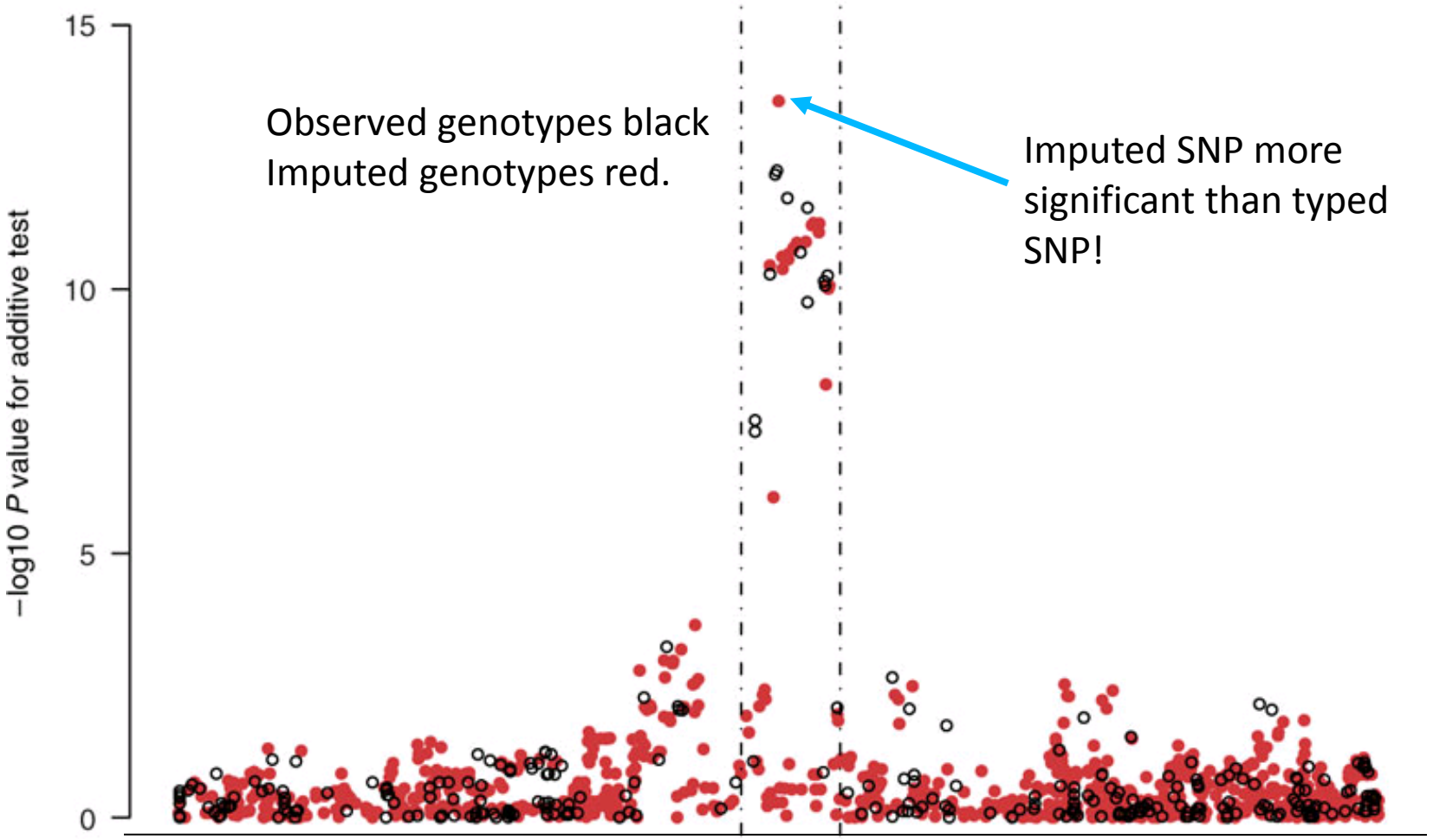


- http://www.shapeit.fr/, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://faculty.washington.edu/browning/beagle/beagle.html
- http://www.sph.umich.edu/csg/abecasis/MACH/download/

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 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.

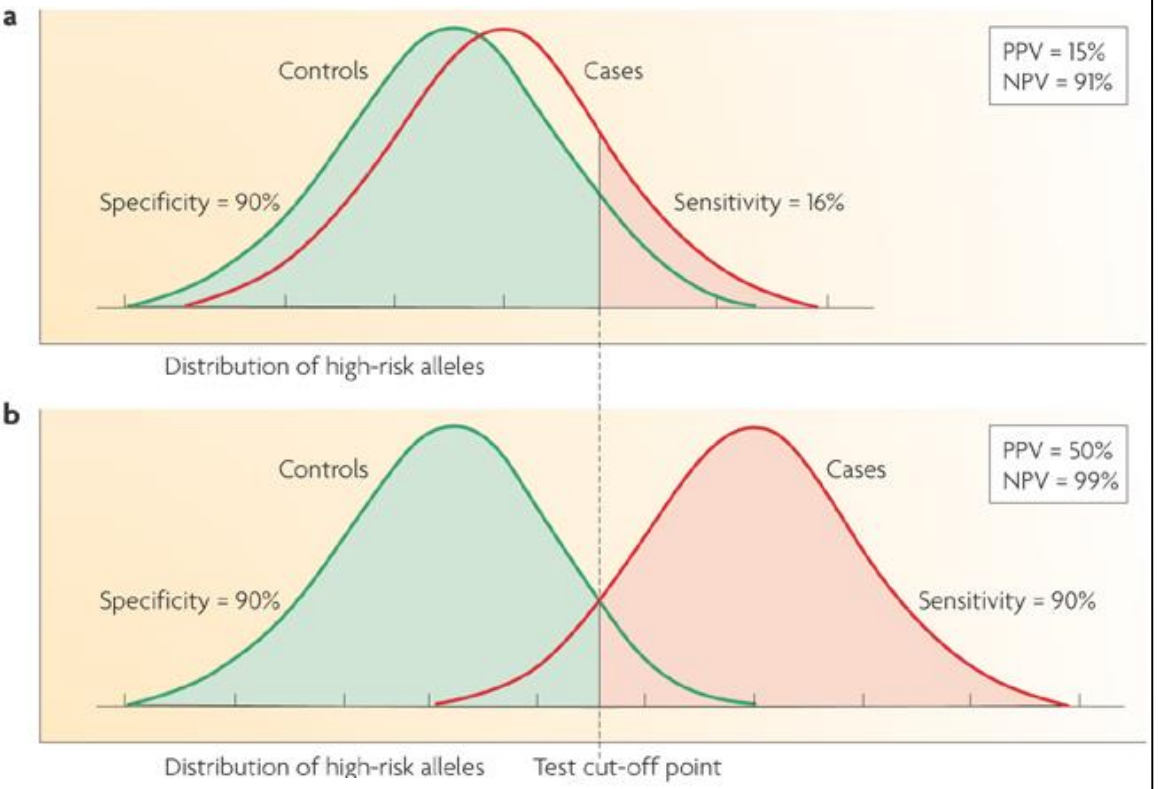
Cancer[a]	# Case carriers[b]	# Cases[b]	Frequency[c]	Odds Ratio[d] (95% CI)	Corrected Odds Ratio[e] (95% CI)	P-value[d]
Kidney	5	303	0.94%	2.32 (0.94, 5.76)	3.32 (0.89, 9.99)	0.034
Bladder	5	335	0.85%	2.05 (0.81, 5.22)	2.84 (0.70, 8.94)	0.065
Non-Hodgkin's Lymphoma	10	846	0.67%	1.81 (0.98, 3.36)	2.42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

- 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



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

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.

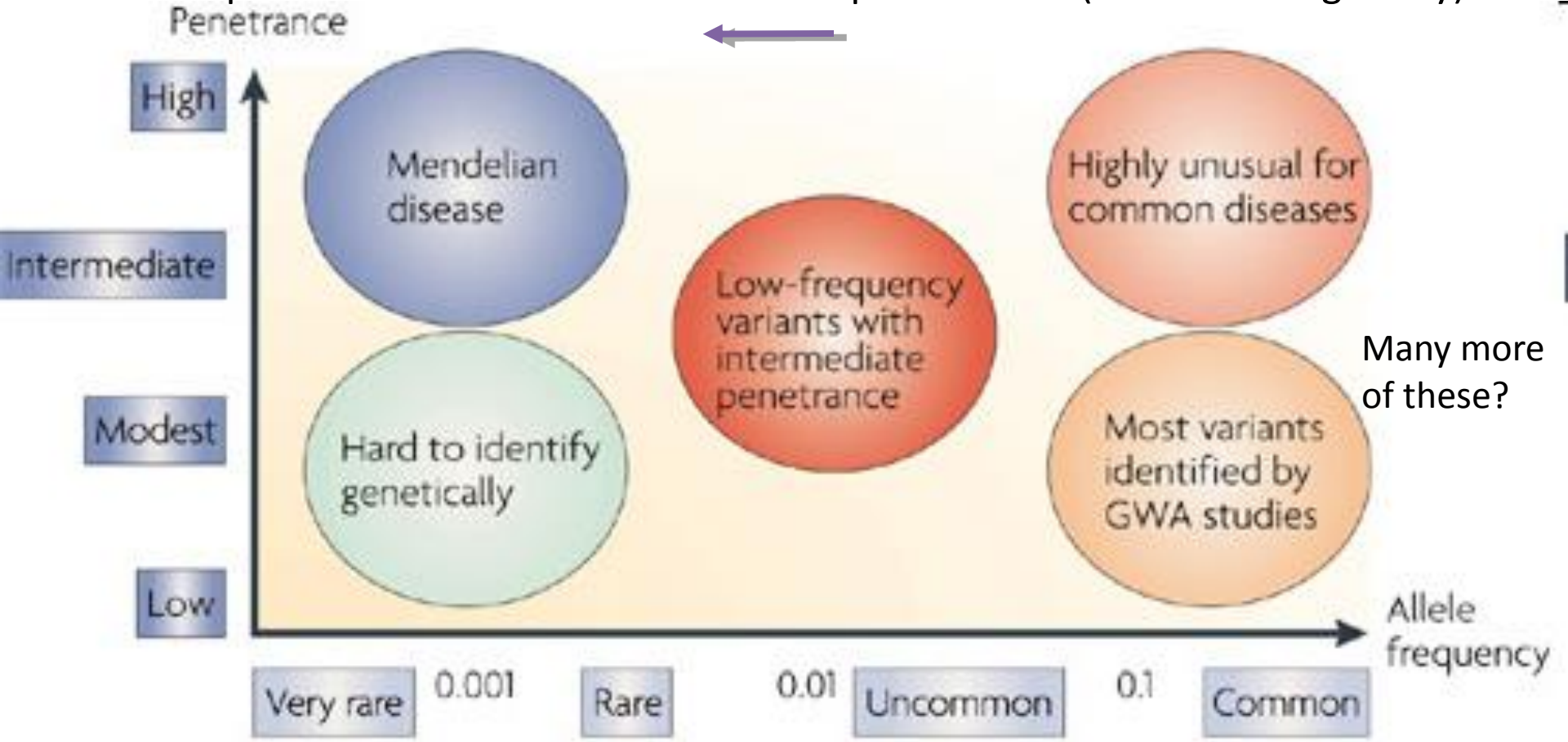
[b] Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.

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

Visccher, 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, 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

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

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: β=**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):

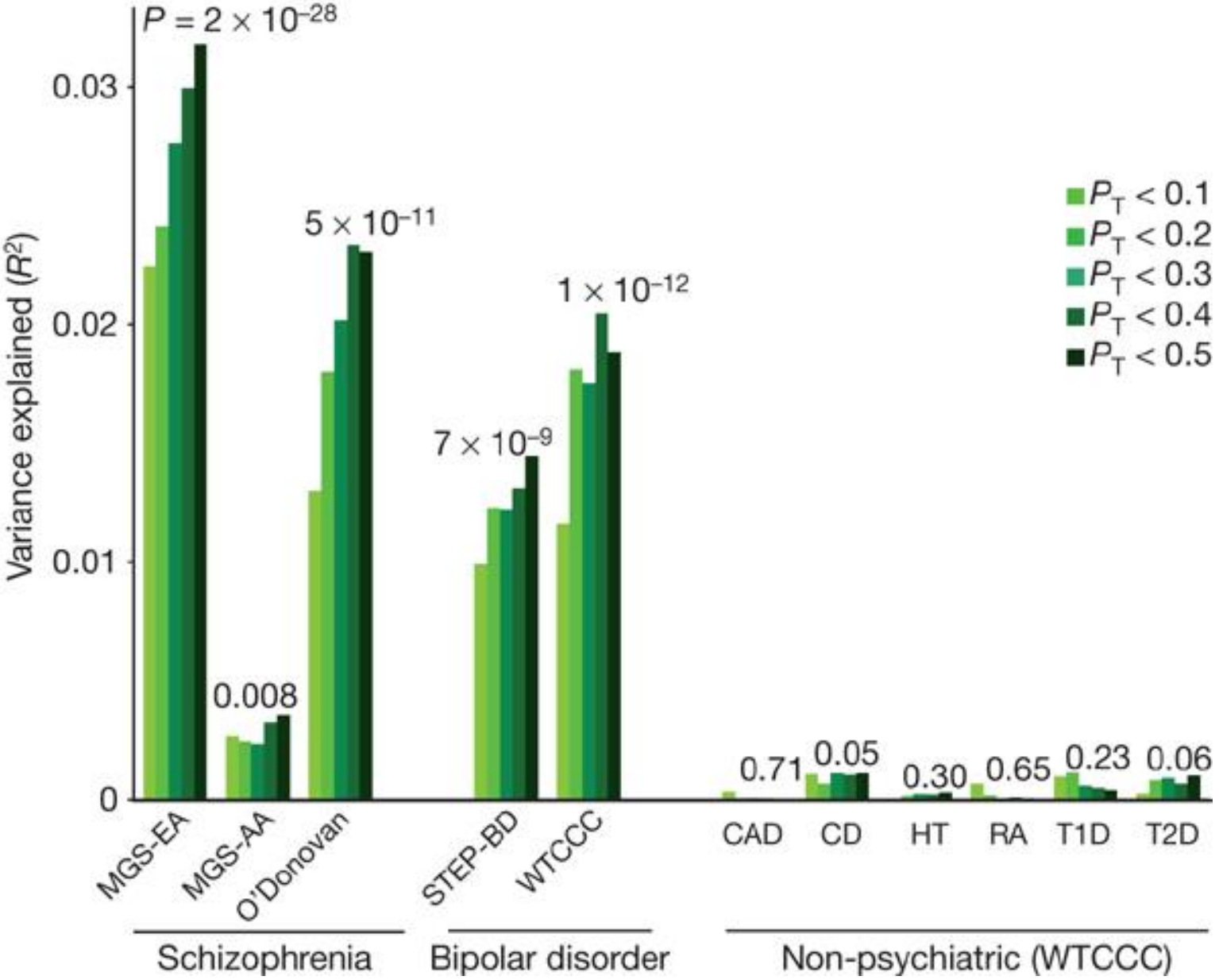
$$x_j = \frac{\sum_{i=1}^{m} \ln(OR_i) \times SNP_{ij}}{m}$$

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***

ISC / Purcell et al. *Nature* 2009

Application of Model



Purcell / ISC et al. *Nature* 2009

Other things than SNPs

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