Genetic Association Analysis

Clement Ma

Sequence Analysis Workshop

December 11, 2014

Lecture slides adapted from Hyun Min Kang and Gonçalo Abecasis

Outline

- Introduction
- Data overview
- Analysis of common variants
- Analysis of low-frequency variants

Genetic Association Analysis

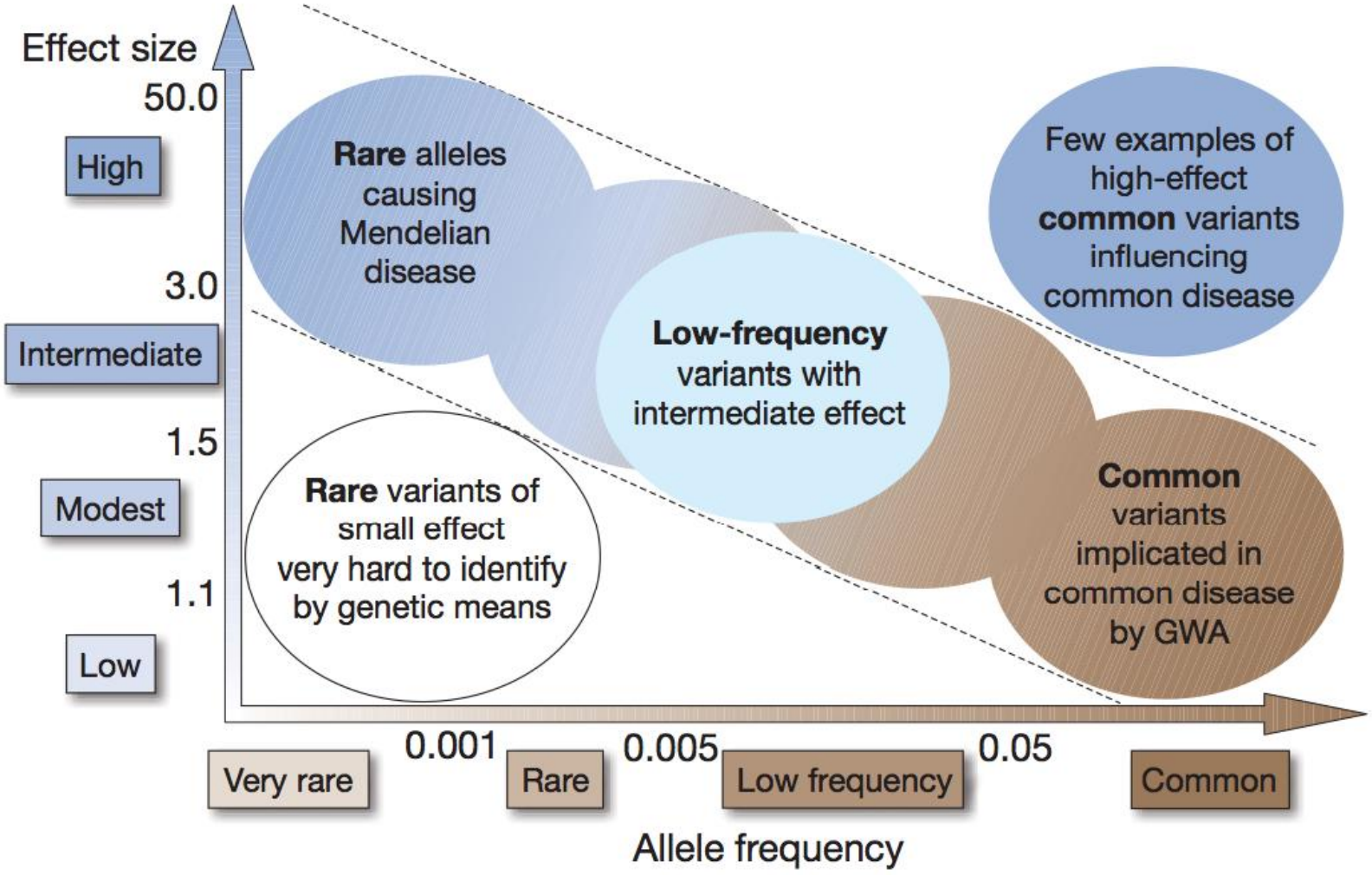
INTRODUCTION

Genetic association studies

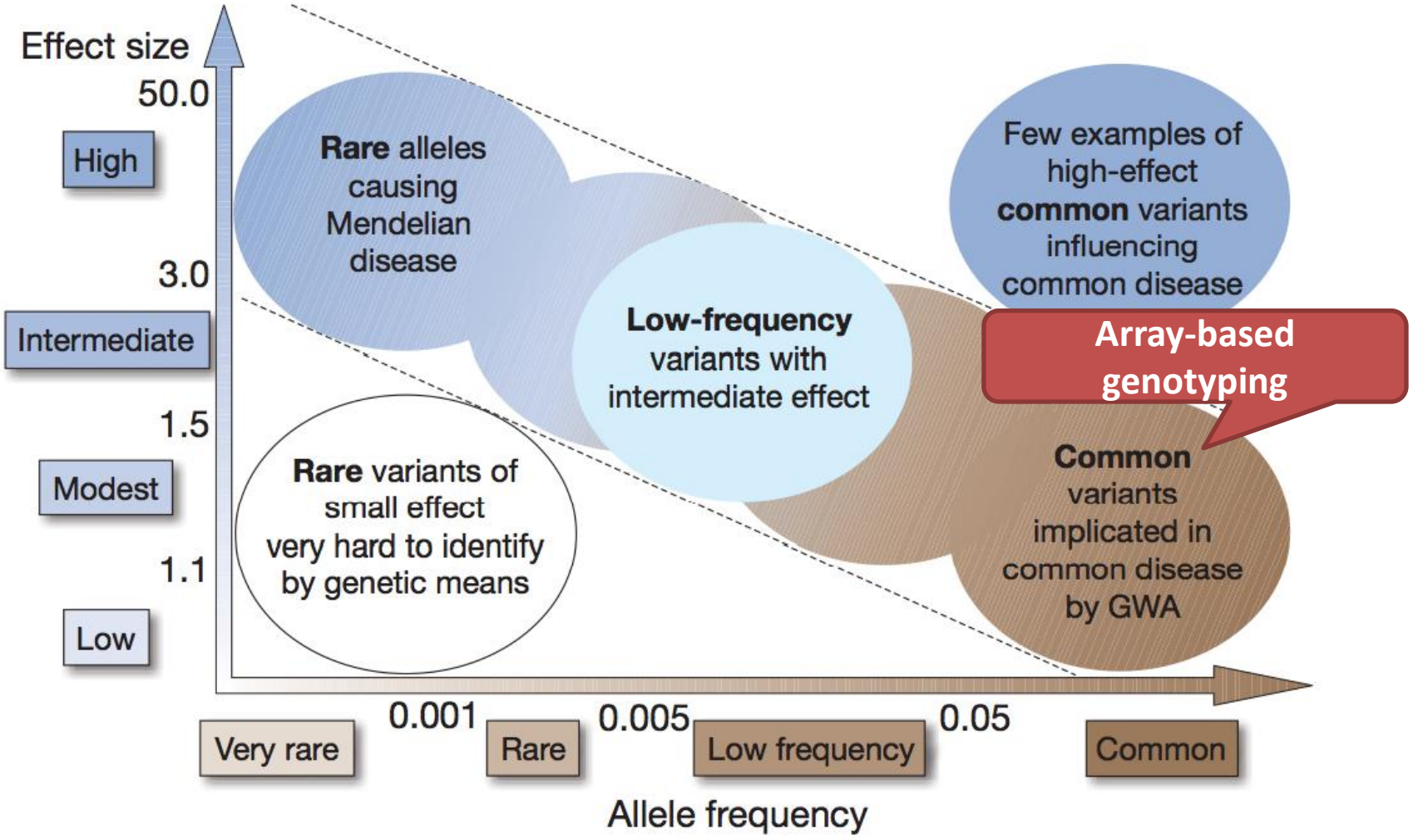
- Goal: Identify genetic variants associated with diseases and traits

- Why?
  - Improve understanding of genetic mechanisms underlying diseases and traits
  - Identify potential drug targets for new therapies
  - Screen individuals with high risk for disease

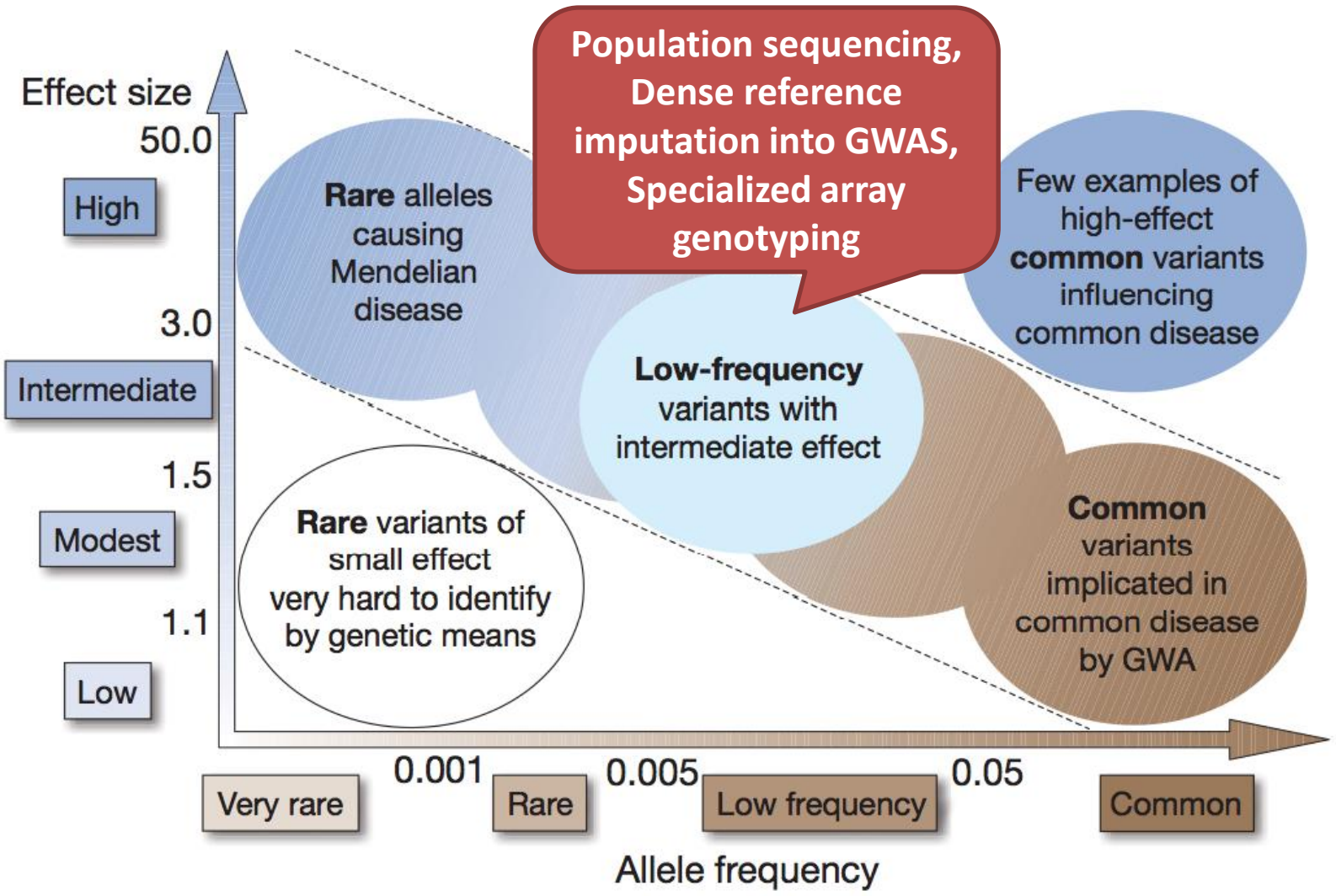
Genetic architecture of complex traits



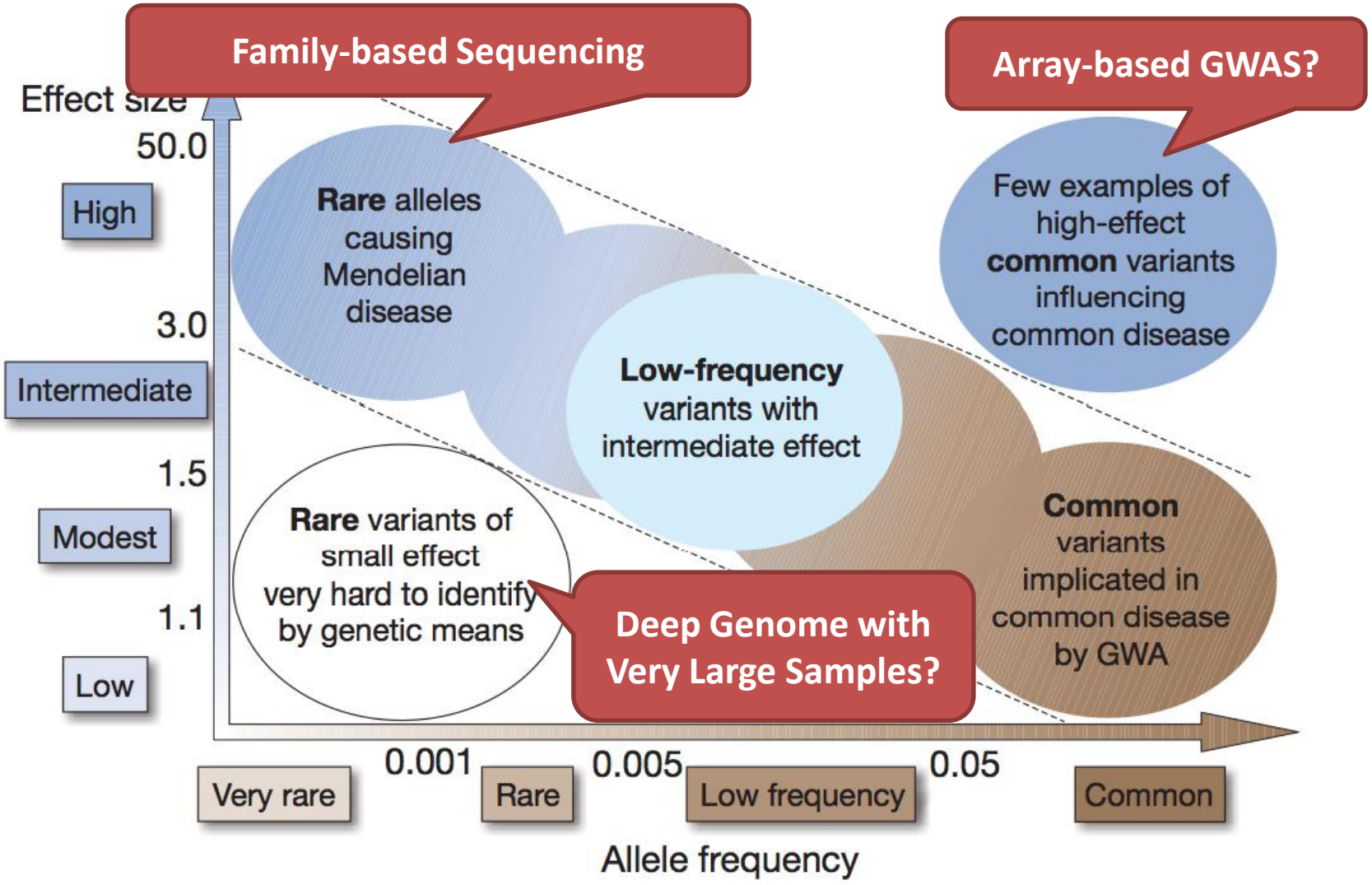
Genetic architecture of complex traits



Genetic architecture of complex traits

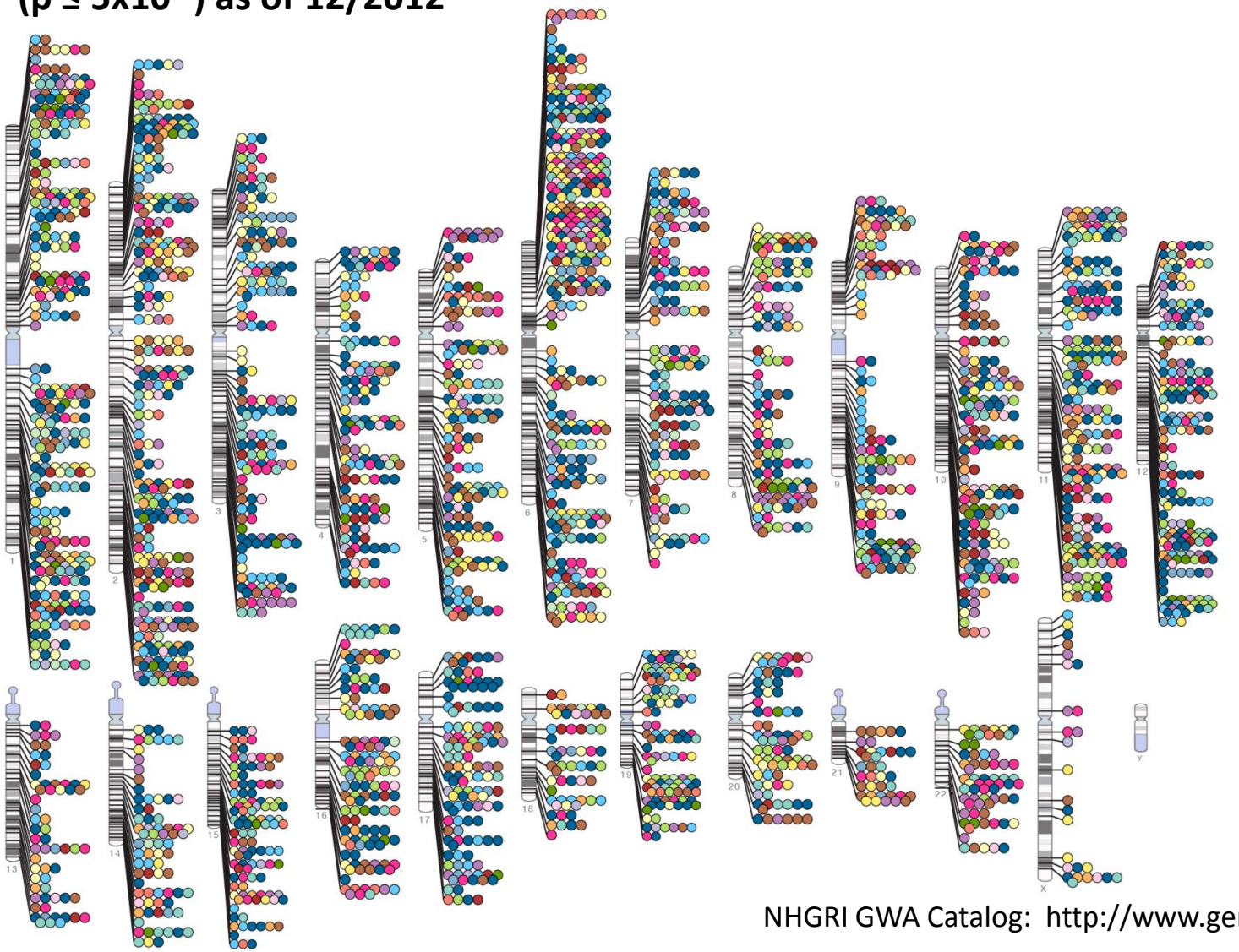


Genetic architecture of complex traits



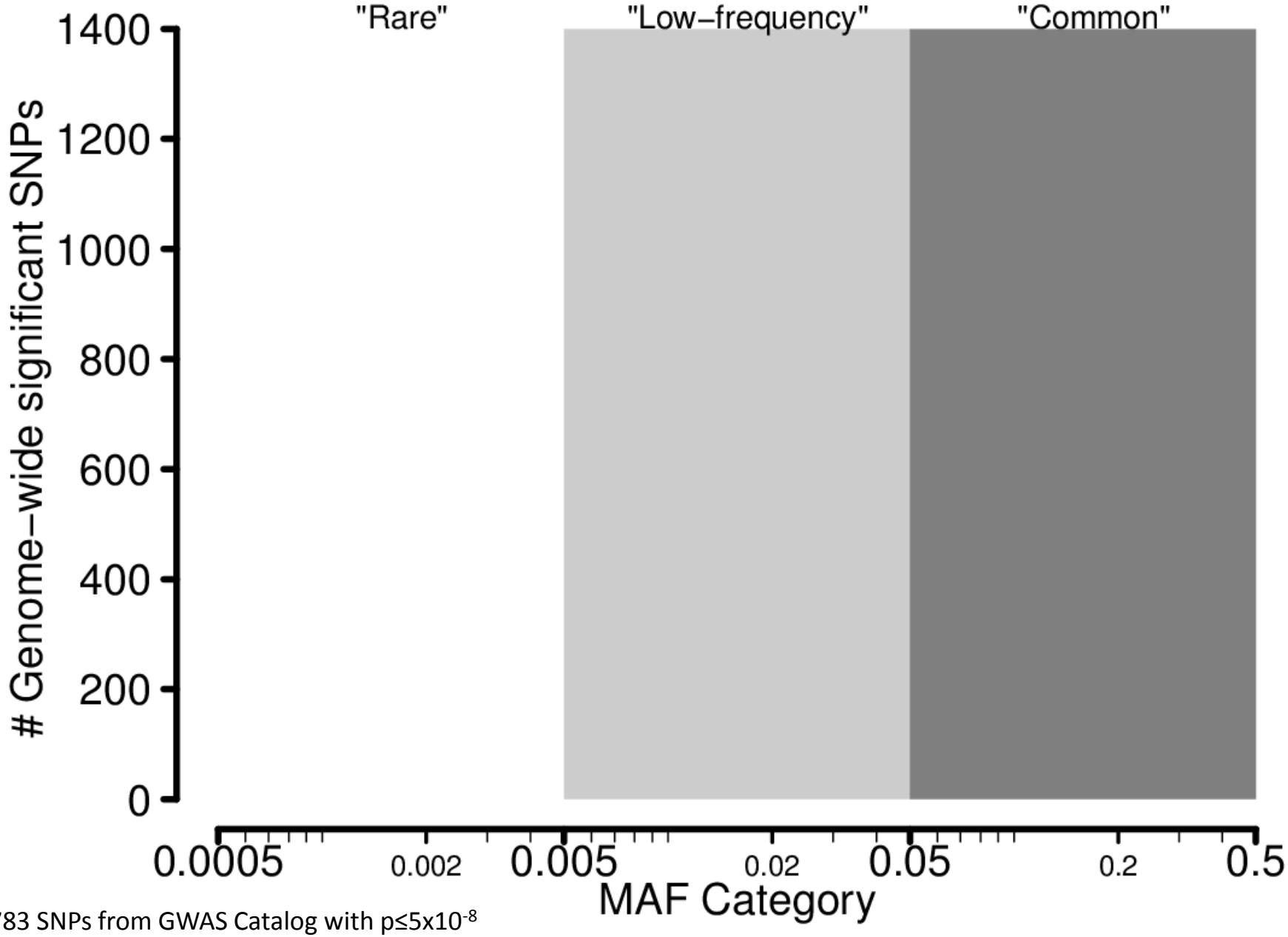
Genotype array-based GWAS identified thousands of associated variants

Published G-W significant associations ($p \leq 5x10^{-8}$) as of 12/2012



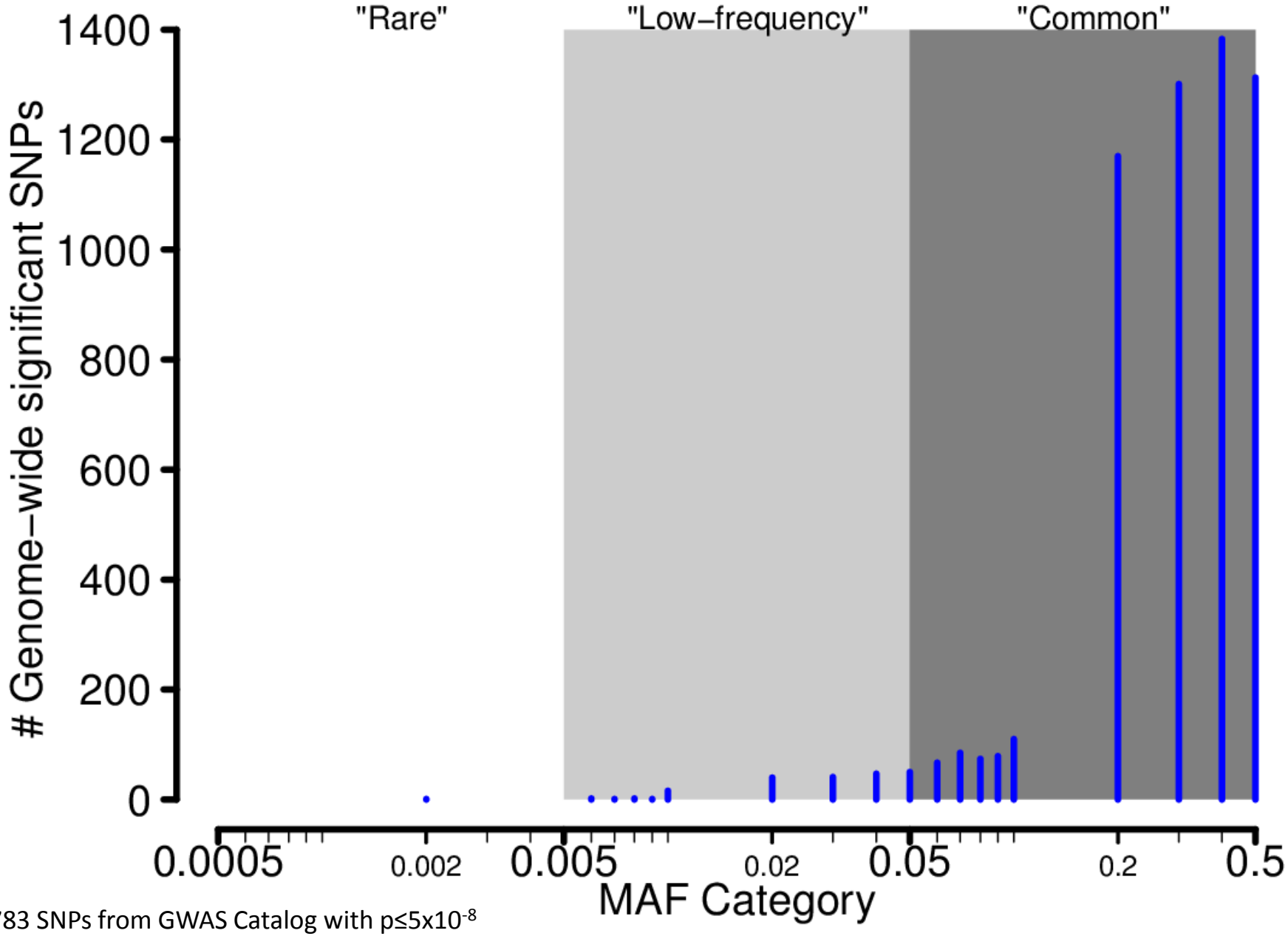
NHGRI GWA Catalog: http://www.genome.gov/GWAStudies/

Genome-wide significant SNPs by MAF



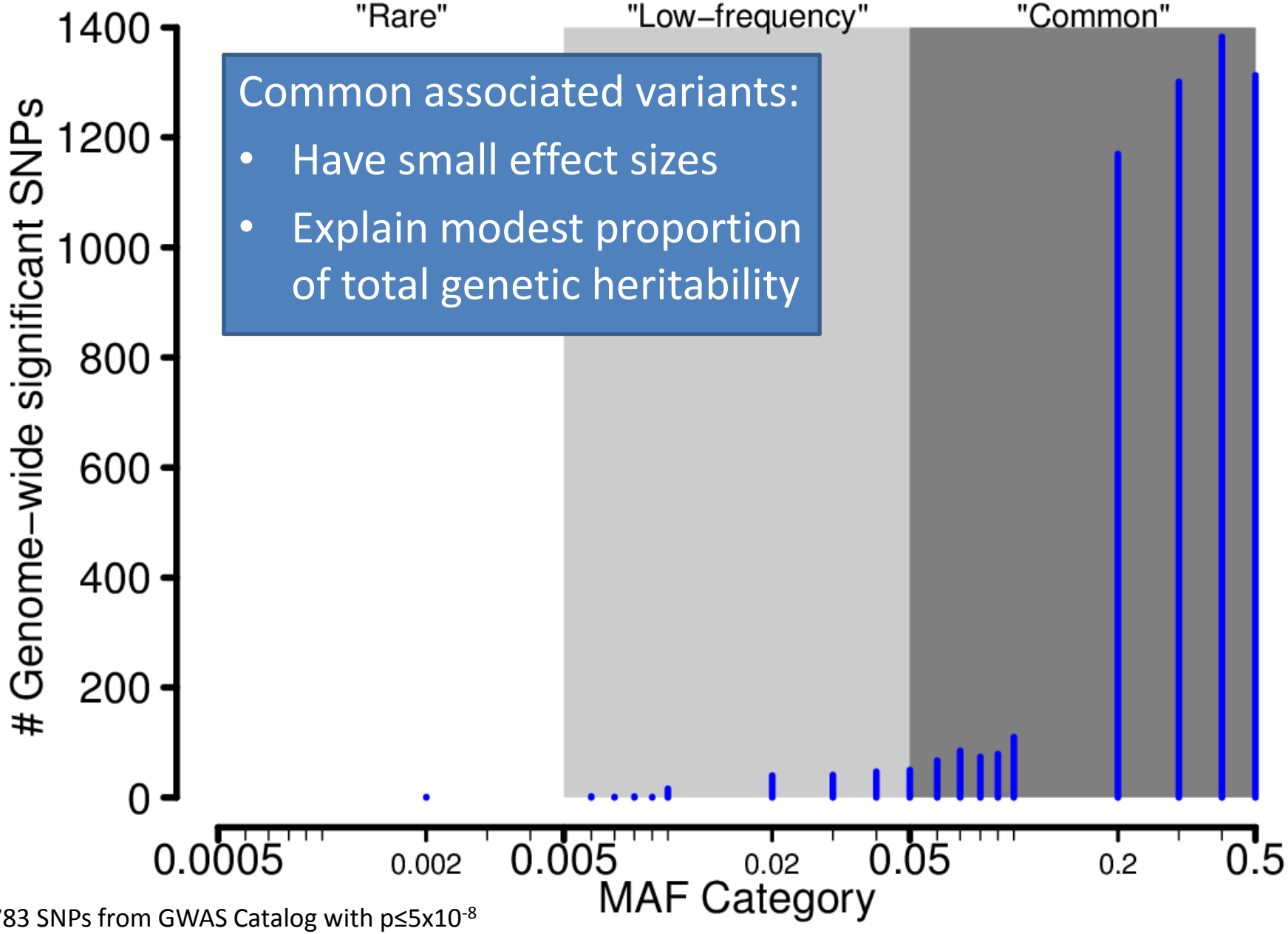
5,783 SNPs from GWAS Catalog with $p \leq 5 \times 10^{-8}$

Genome-wide significant SNPs by MAF



5,783 SNPs from GWAS Catalog with $p \leq 5 \times 10^{-8}$

Genome-wide significant SNPs by MAF

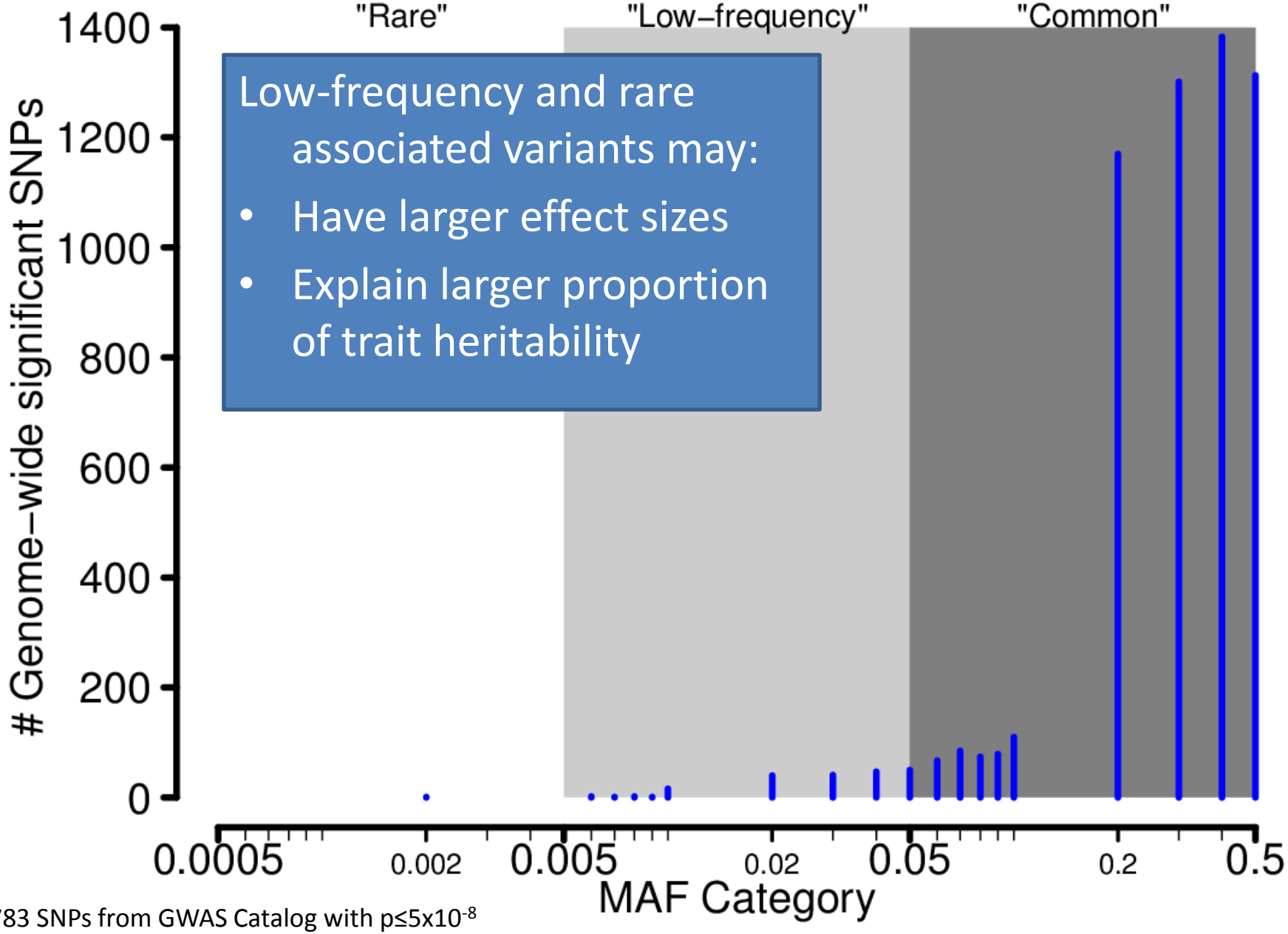


Common associated variants:

- Have small effect sizes
- Explain modest proportion of total genetic heritability

5,783 SNPs from GWAS Catalog with $p \leq 5 \times 10^{-8}$

Genome-wide significant SNPs by MAF



Low-frequency and rare associated variants may:

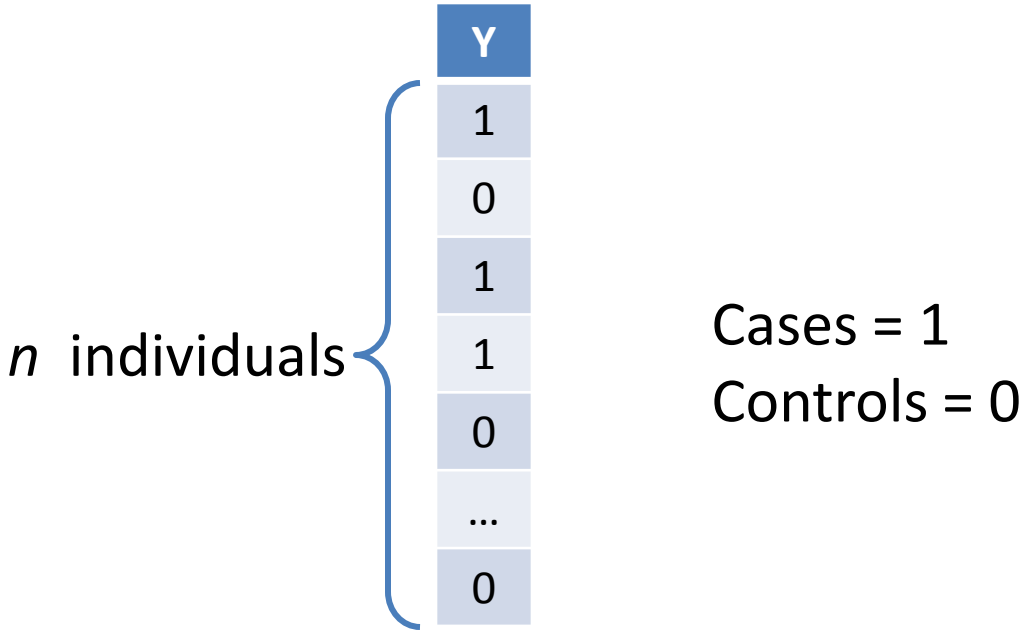
- Have larger effect sizes
- Explain larger proportion of trait heritability

5,783 SNPs from GWAS Catalog with $p \leq 5 \times 10^{-8}$

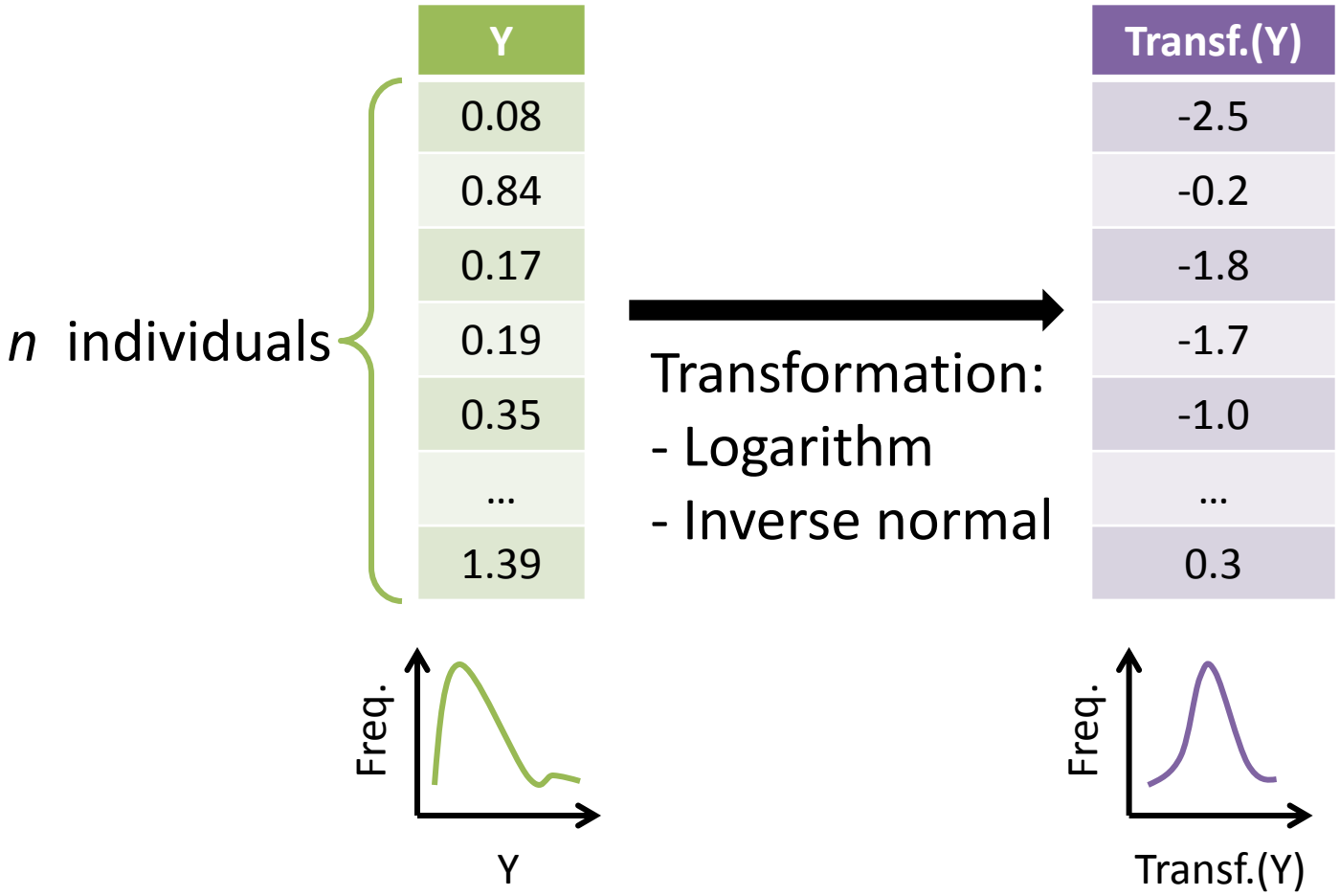
Genetic Association Analysis

DATA OVERVIEW

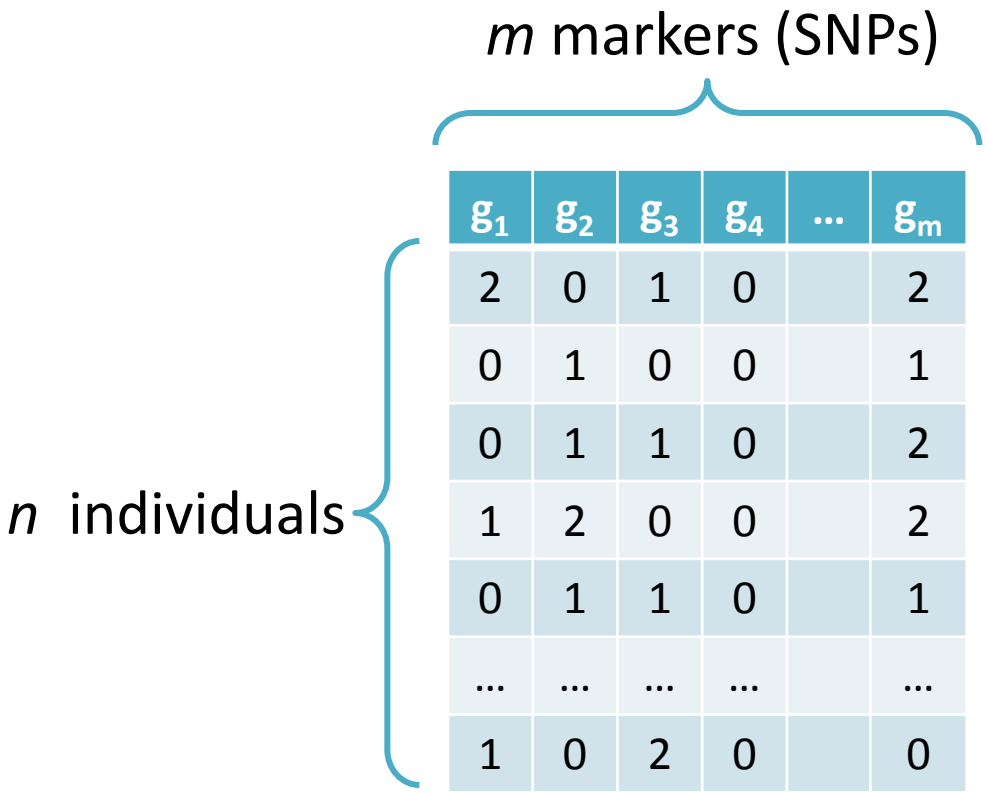
Phenotypes: binary trait



Phenotypes: quantitative trait (QT)



Genotypes: hard genotypes



Aside

Genotype imputation

- **Goal:** to increase power by using previously genotyped GWAS samples
- **Problem:** GWAS samples genotyped at fewer or different variant sites
- **Method**: Use genotype imputation to fill in missing genotypes

(Marchini et. al., *Nat. Genet.*, 2007; Li et. al. *Genet. Epidemol.*, 2009)

Aside

Using genotype imputation to fill in missing genotypes

1. **Starting Data**

Genotyped sample

. . **C** . . **G** . **C** .

Reference haplotypes

A G A T C T C C T

A G C T C T C A T

A G A T C G C C T

A G A T C T A C T

Aside

Using genotype imputation to fill in missing genotypes

2. Identify shared regions of chromosome

Genotyped sample

.	.	**C**	.	.	**G**	.	**C**	.

Reference haplotypes

A	G	A	T	C	T	C	C	T
A	G	**C**	T	C	T	C	A	T
A	G	A	T	C	**G**	C	**C**	T
A	G	A	T	C	T	A	C	T

Aside

Using genotype imputation to fill in missing genotypes

3. Fill in missing genotypes

Genotyped sample

A	G	**C**	T	C	**G**	C	**C**	T

Reference haplotypes

A	G	A	T	C	T	C	C	T
A	G	**C**	T	C	T	C	A	T
A	G	A	T	C	**G**	C	**C**	T
A	G	A	T	C	T	A	C	T

Genotypes: imputed dosages

m markers (SNPs)

n individuals

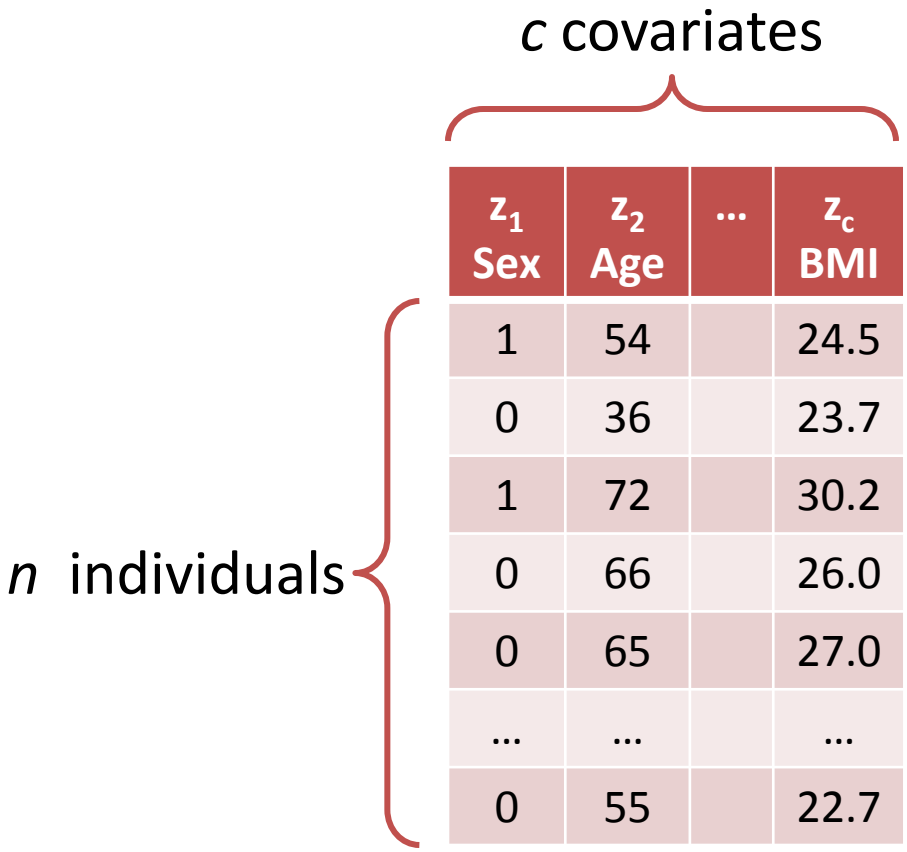
g_1	g_2	g_3	g_4	…	g_m
1.99	0.21	0.98	0.01	…	2.00
0.00	1.4	0.00	0.00	…	1.00
0.01	0.8	1.00	0.00	…	2.00
1.34	1.6	0.03	0.00	…	1.99
0.4	0.89	1.00	0.03	…	0.99
…	…	…	…	…	…
1.01	0.34	2.00	0.00	…	0.01

Imputation Quality Score

r^2_1	r^2_2	r^2_3	r^2_4	…	r^2_m
0.7	0.4	0.98	0.99	…	0.97

(Marchini et. al., *Nat. Genet.*, 2007; Li et. al. *Genet. Epidemol.*, 2009)

Additional covariates



Study individuals: relatedness and population structure

- Unrelated individuals

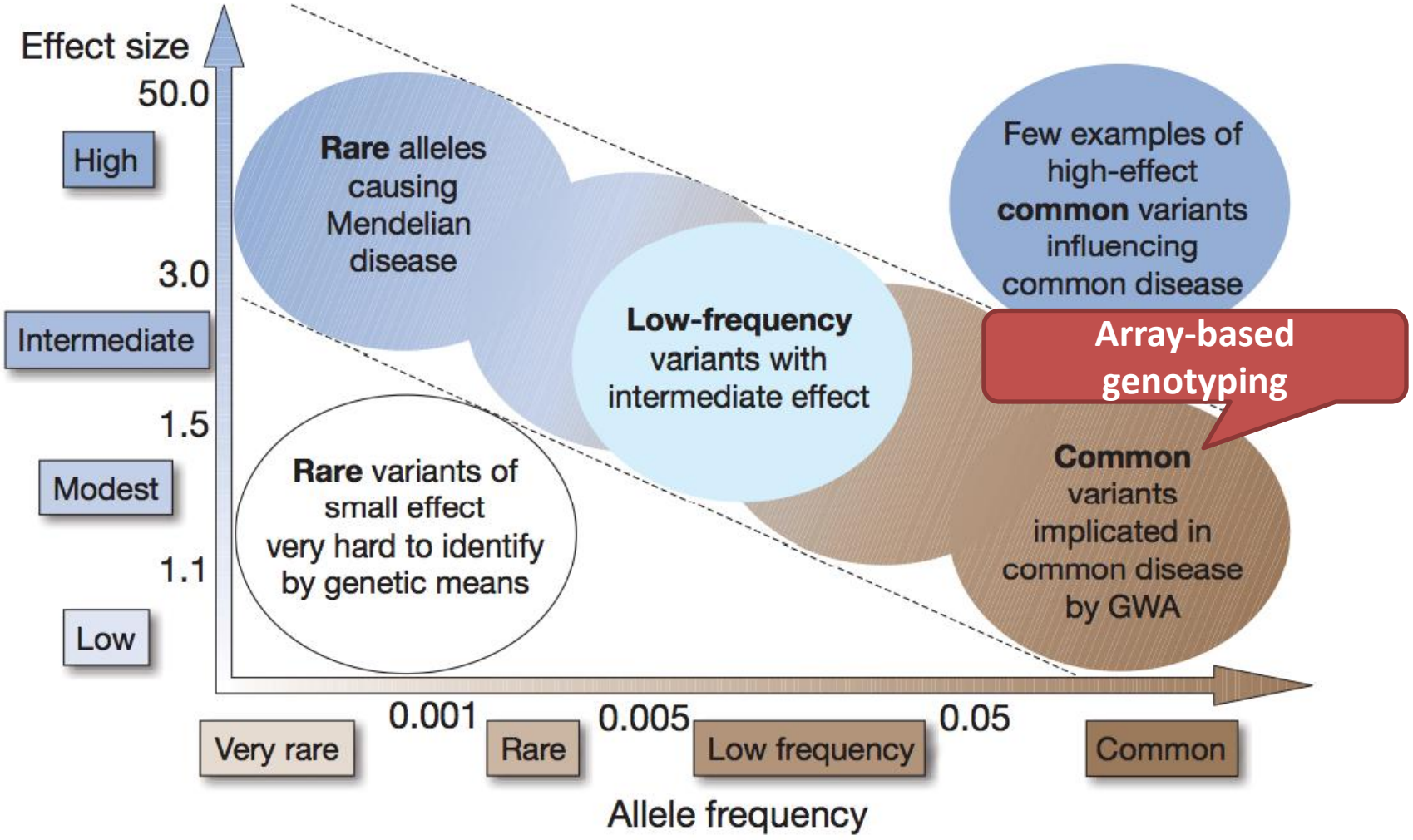
- Related individuals
  - Identify any relationships between individuals

- Population structure
  - Individuals are from different populations

Genetic Association Analysis

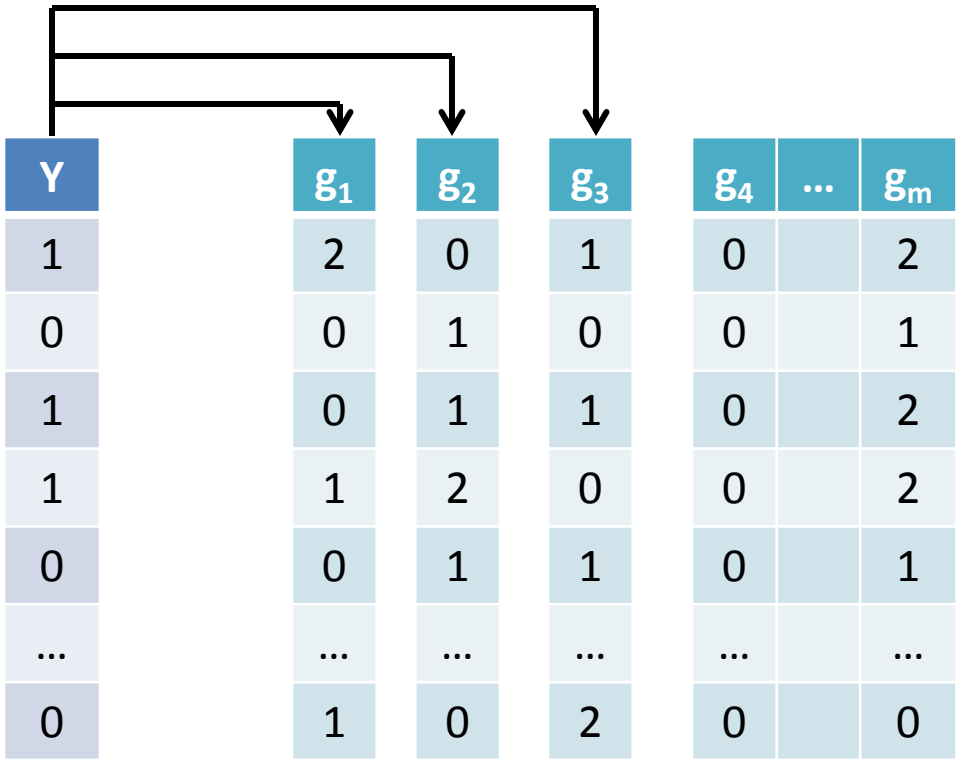
ANALYSIS OF COMMON VARIANTS

Genetic architecture of complex traits



Single variant analysis

Test each variant for association with outcome

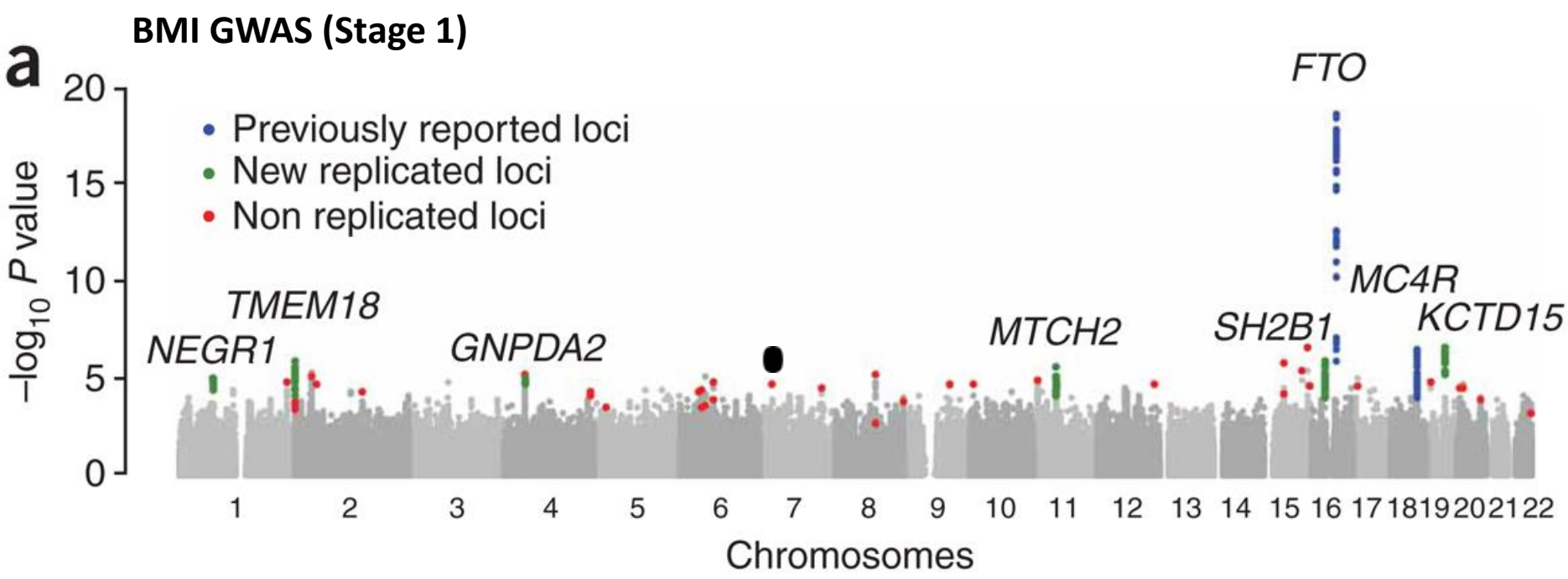


Y	g_1	g_2	g_3	g_4	…	g_m
1	2	0	1	0		2
0	0	1	0	0		1
1	0	1	1	0		2
1	1	2	0	0		2
0	0	1	1	0		1
…	…	…	…	…		…
0	1	0	2	0		0

Analysis methods

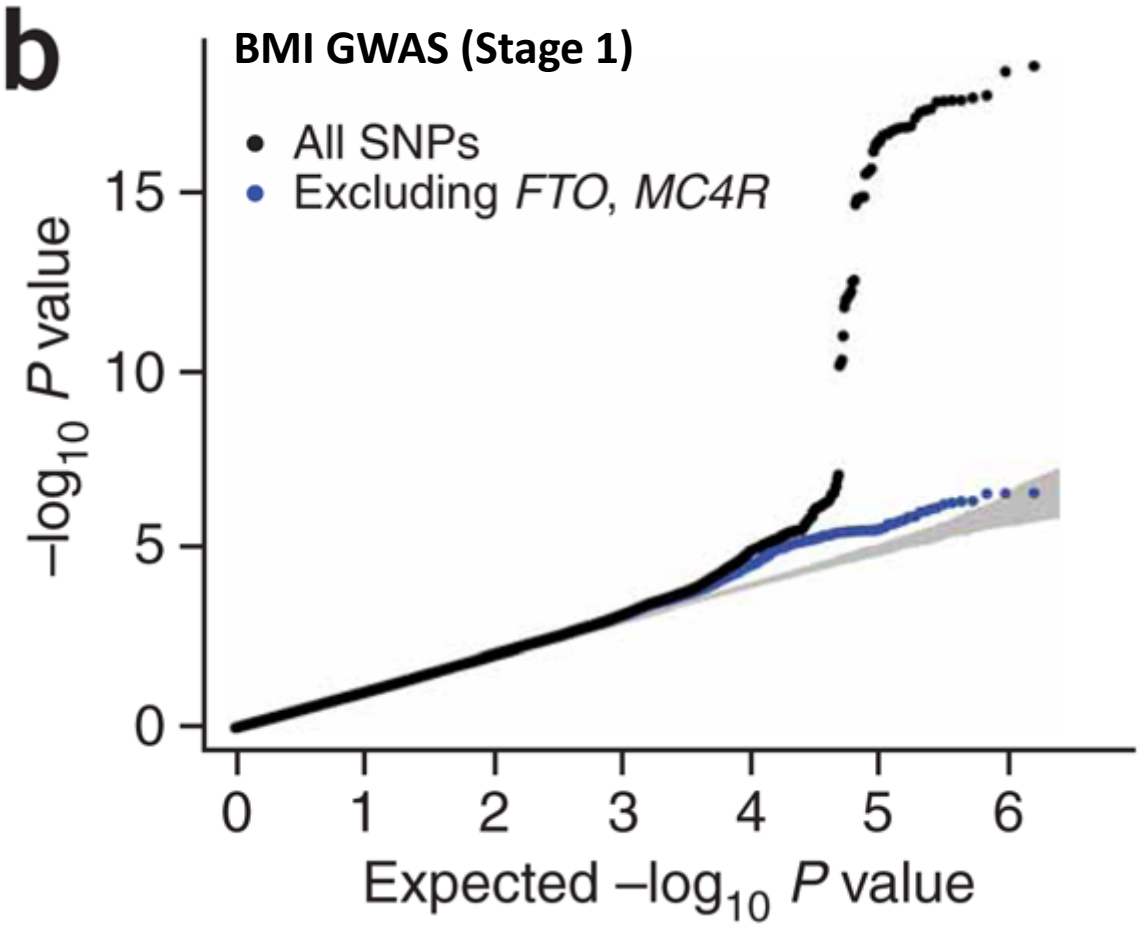
- Binary traits
  - Contingency table tests cannot adjust for covariates
    - Chi-square Test
    - Cochran-Armitage Trend Test
    - Fisher's Exact Test
  - Logistic regression can account for covariates
- Quantitative traits
  - Linear regression

Visualizing results: Manhattan Plot



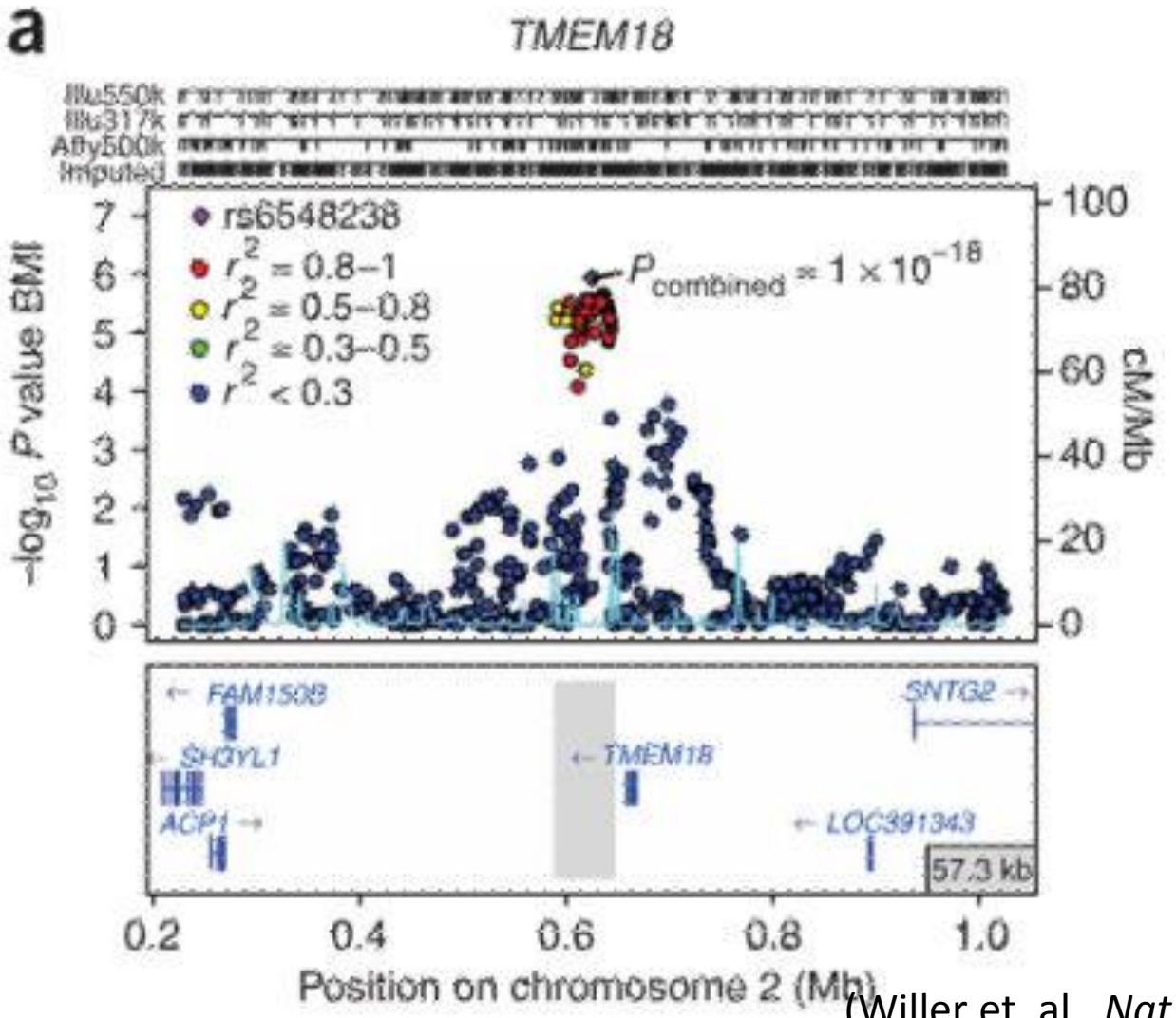
(Willer et. al., *Nat. Genet.*, 2009)

Visualizing results: quantile-quantile (QQ) plot)



(Willer et. al., *Nat. Genet.*, 2009)

Visualizing results: regional plot



(Willer et. al., *Nat. Genet.*, 2009)

Sources of association

- Causal association
  - Genetic marker alleles influence susceptibility

best

- Linkage disequilibrium
  - Genetic marker alleles associated with other nearby alleles that influence susceptibility

useful

- Population stratification
  - Genetic marker is unrelated to disease alleles

misleading

Example of spurious association due to population stratification

Population 1

	Allele 1	Allele 2
Affected	50 ($f_{1,Aff}$=0.2)	200
Unaffected	25 ($f_{1,Unaff}$=0.2)	100

χ^2 = 0.00 p-value = 1.0

Population 2

	Allele 1	Allele 2
Affected	100 ($f_{1,Aff}$=0.8)	25
Unaffected	200 ($f_{1,Aff}$=0.8)	50

χ^2 = 0.00 p-value = 1.0

Combined

	Allele 1	Allele 2
Affected	150 ($f_{1,Aff}$=0.4)	225
Unaffected	225 ($f_{1,Aff}$=0.6)	150

χ^2 = 29.2 p-value = 6.5×10^{-8}

The stratification problem happens..

- If..
  - Phenotypes differ between populations
  - and allele frequencies have drifted apart
- Then..
  - Unlinked markers exhibit association
  - Not very useful for gene mapping!
- For example, Glaucoma has prevalence of ~2% in elderly Caucasians, but ~8% in African-Americans

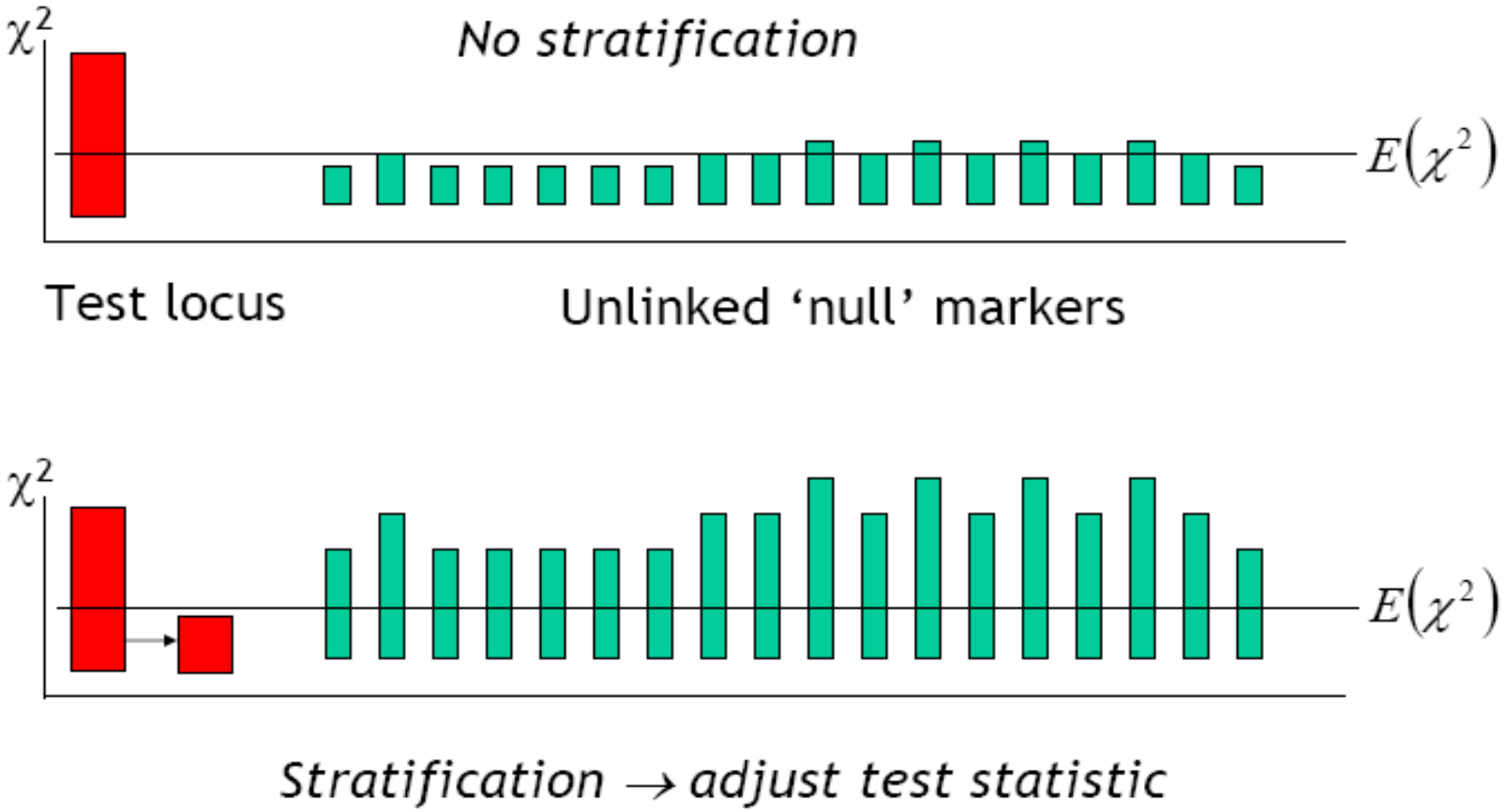
Possible solutions for population stratification

- Avoid stratification by design
  - Collect a better matched sample by ancestry
  - Use family-based controls
    - E.g. apply Transmission Disequilibrium Test (TDT)

- Analyze association by population groups
  - Using self reported ethnicity or genetic markers
  - Carry out association analysis within each group

- Account for inflated false-positive rate
  1. Apply genomic control
  2. Adjust for population principal components
  3. Variance component model for family-based association test

Genomic control



(Figure courtesy Shaun Purcell, Harvard, and Pak Sham, HKU)

Genomic inflation factor

- Compute χ^2 statistic for each marker
- Genomic inflation factor (λ)

$$\lambda = \frac{\text{Median Observed } \chi^2}{\text{Median Expected } \chi^2}$$

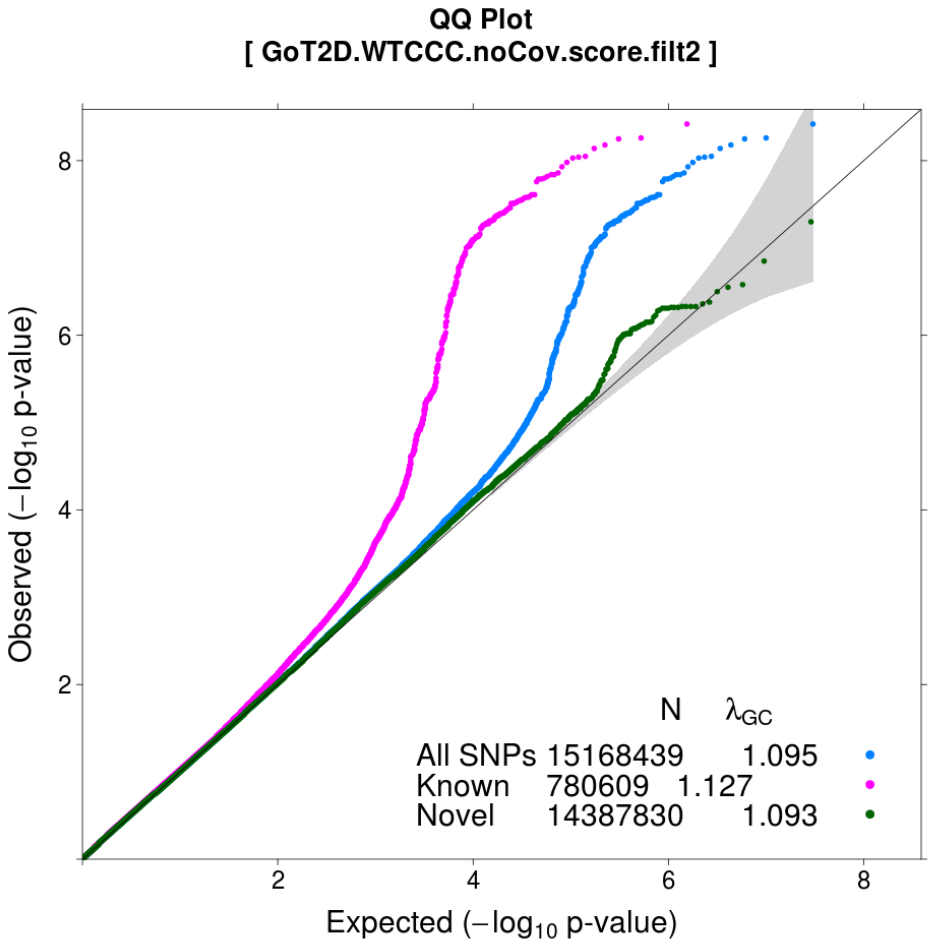
  - Median expected $\chi^2 = 0.456$
    - Why use median vs. mean?
- Adjust statistic at candidate markers
  - Replace χ^2_{biased} with $\chi^2_{fair} = \chi^2_{biased}/\lambda$
  - Should be $\lambda \geq 1$
    - Why?

(Devlin & Roeder, *Biometrics*, 1999)

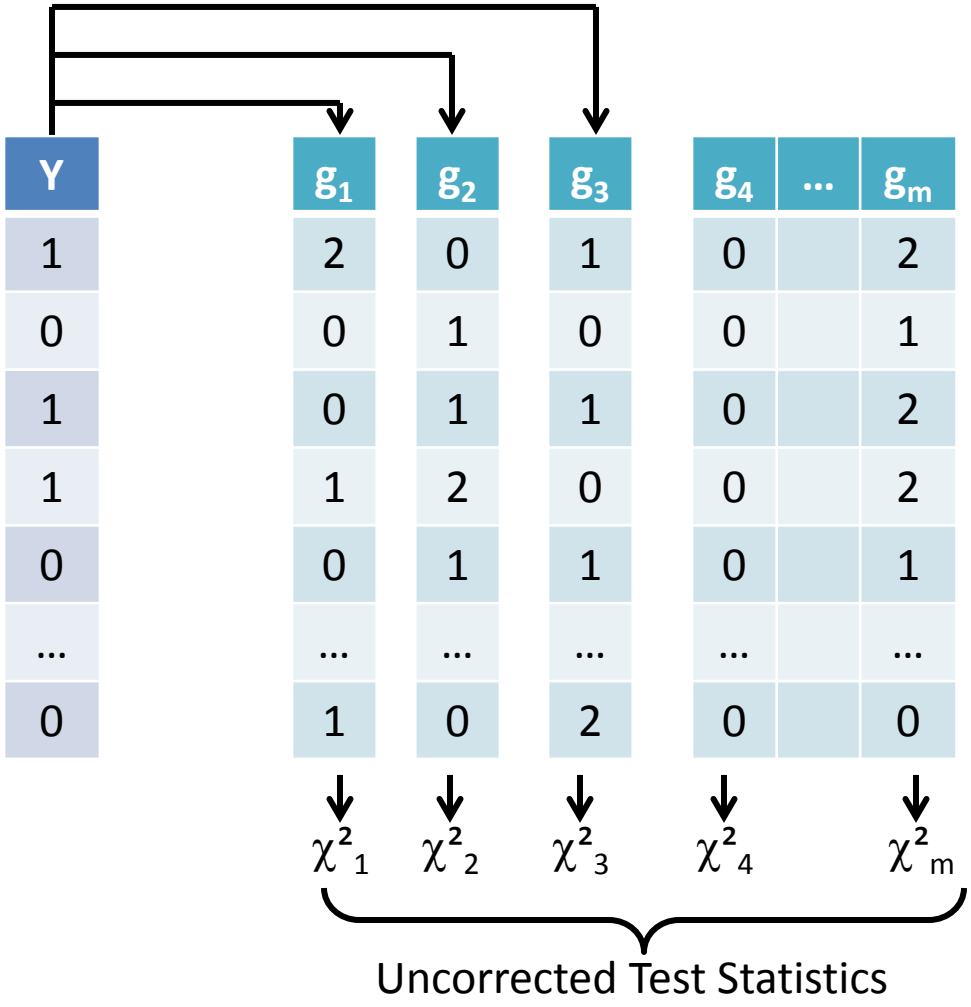
QQ plots: a useful diagnostic

- **Data:** WTCCC Study
- **Phenotype:** T2D status
- **Genotypes:** imputed using GoT2D reference
- **Analysis:** logistic regression

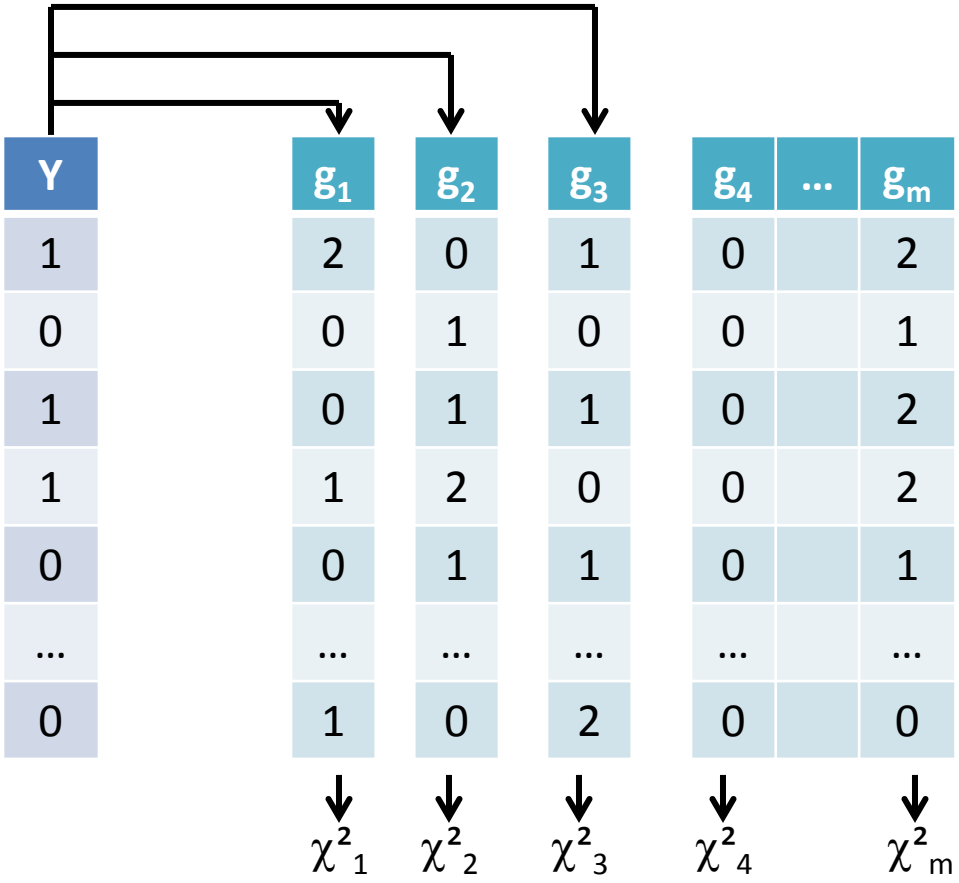
- Classify SNPs as within or outside Known (+/-1Mb) T2D loci
- For all SNPs, $\lambda = 1.095$
  - Some population stratification
- For Known SNPs, $\lambda = 1.127$
  - Very inflated, but under alternative hypothesis



Genomic control example



Genomic control example



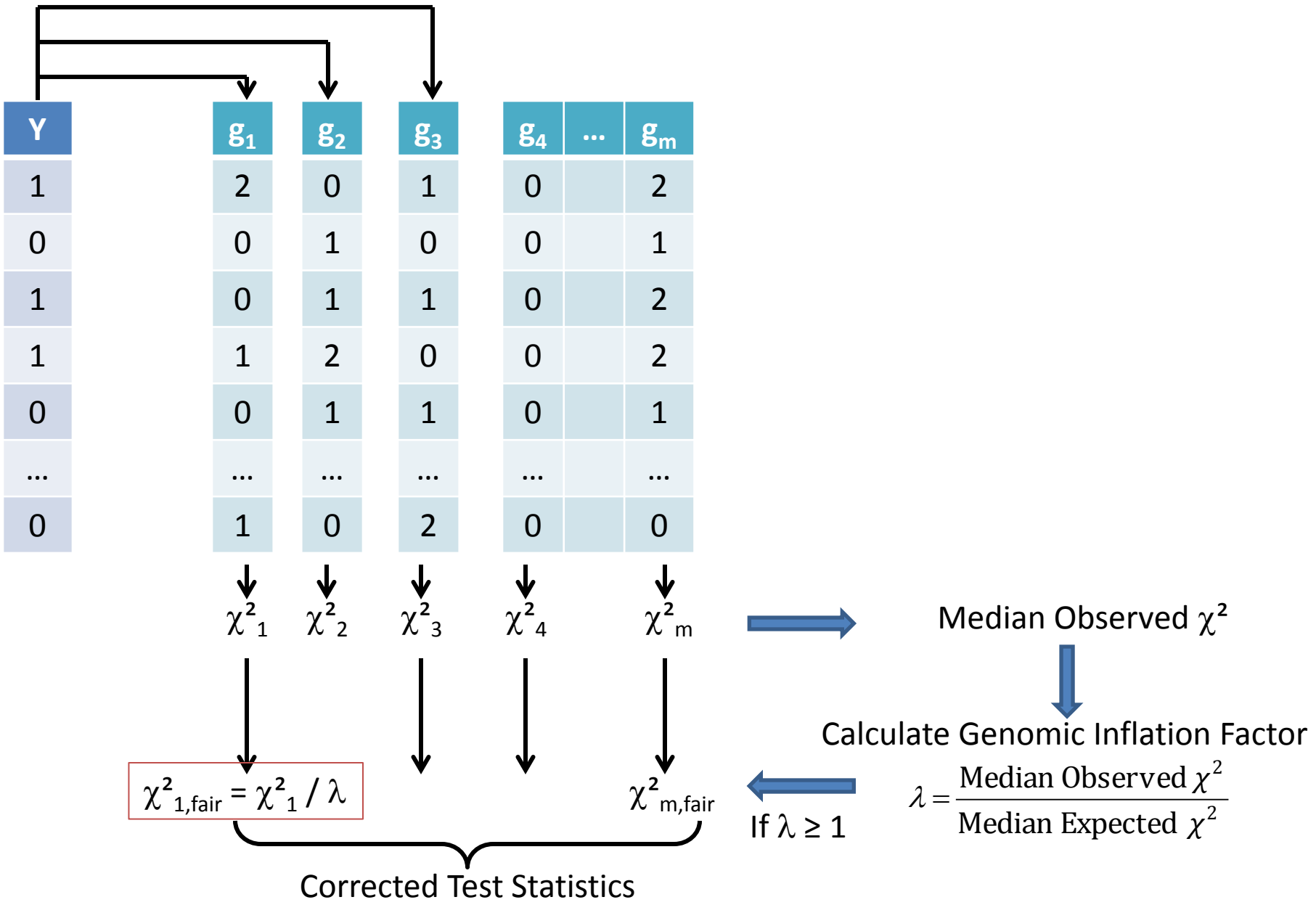
Y	g_1	g_2	g_3	g_4	…	g_m
1	2	0	1	0		2
0	0	1	0	0		1
1	0	1	1	0		2
1	1	2	0	0		2
0	0	1	1	0		1
…	…	…	…	…		…
0	1	0	2	0		0

Median Observed χ^2

Calculate Genomic Inflation Factor

$$\lambda = \frac{\text{Median Observed } \chi^2}{\text{Median Expected } \chi^2}$$

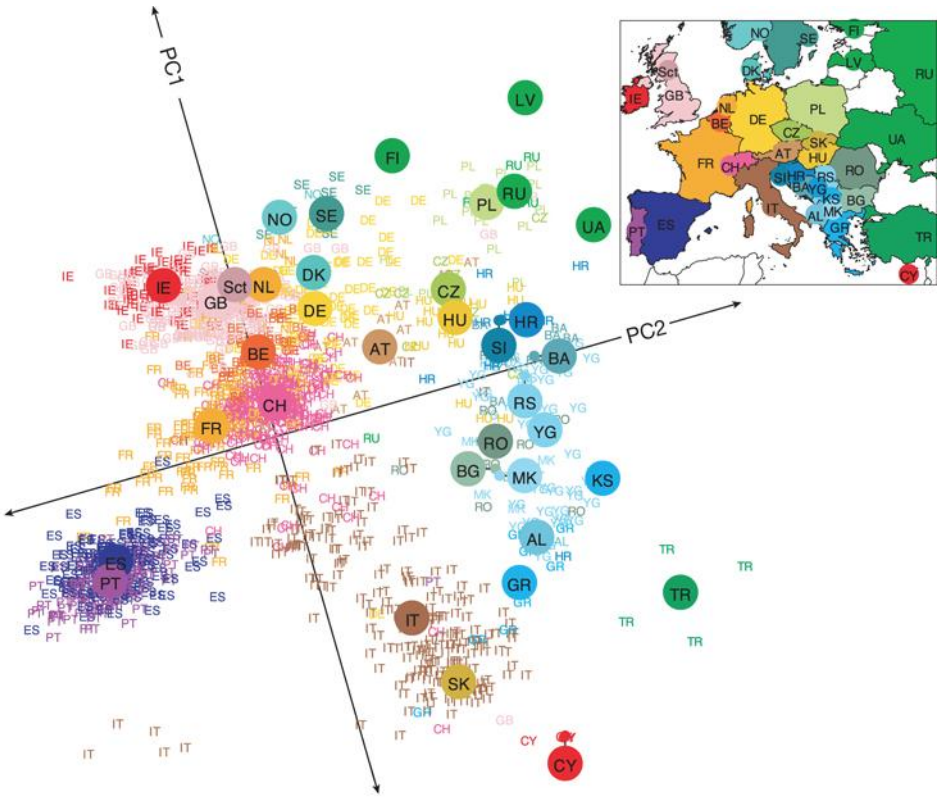
Genomic control example



Principal components analysis (PCA)

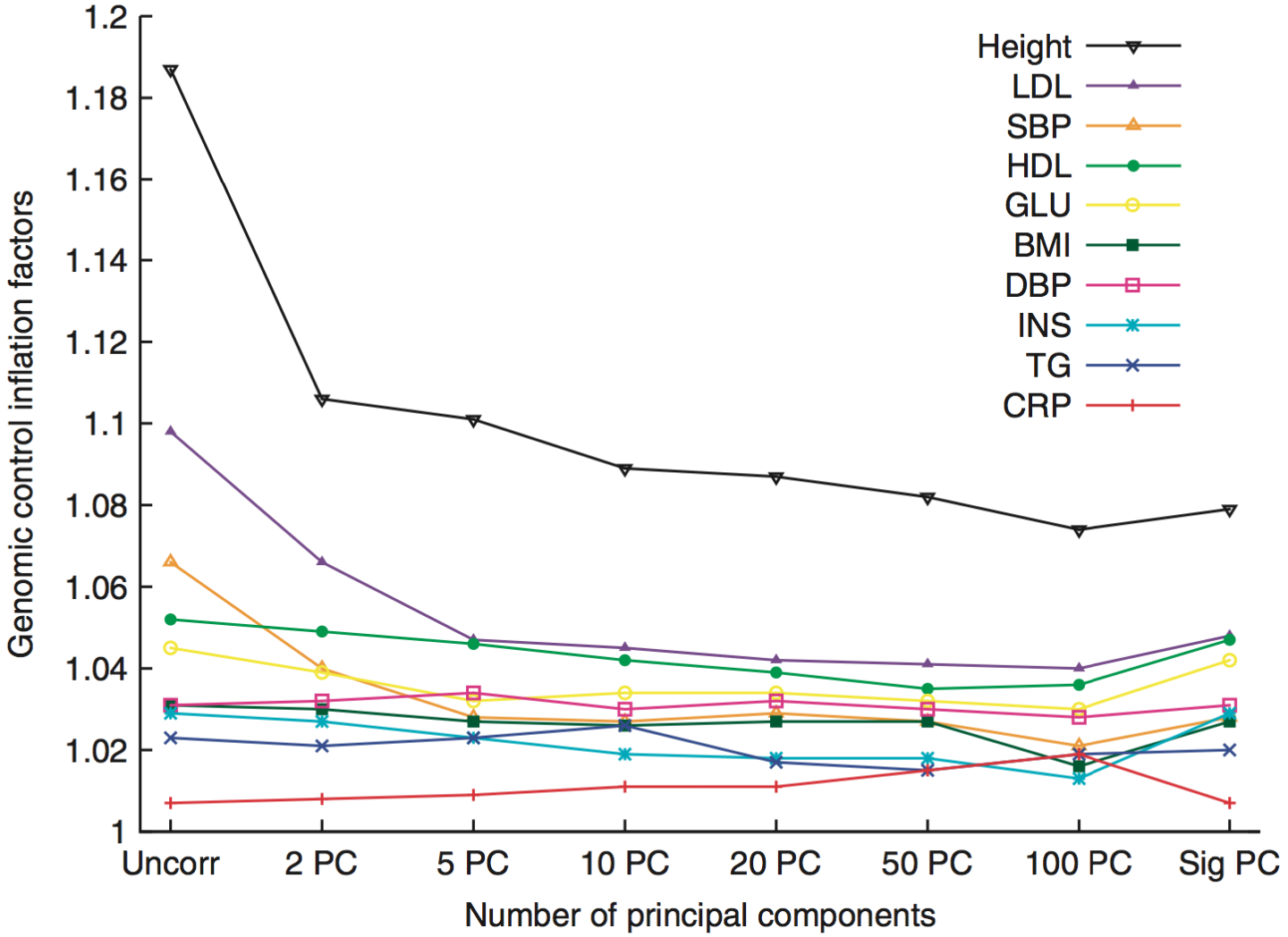
- Use PCA to determine "axes of genotype variation" for a selected set of genotypes
  - Principal components mirror European geography
- Include PC's as covariates in regression model to adjust for stratification

(Figure from Novembre et. al., *Nature,* 2008)



(Price et. al., *Nat. Genet.*, 2006)

Correcting for population structure using principal components



(Kang et. al., *Nat. Genet.*, 2010)

Variance component model for family-based association test

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$$

Variance component model for family-based association test

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$$

- Family-based analysis assumes phenotypes are correlated with relatives' phenotypes

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma_g^2 K + \sigma_e^2 I)$$ K_{ij} : kinship coefficient

Variance component model for family-based association test

- Population-based analysis assumes uncorrelated phenotypes between individuals under the null

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$$

- Family-based analysis assumes phenotypes are correlated with relatives' phenotypes

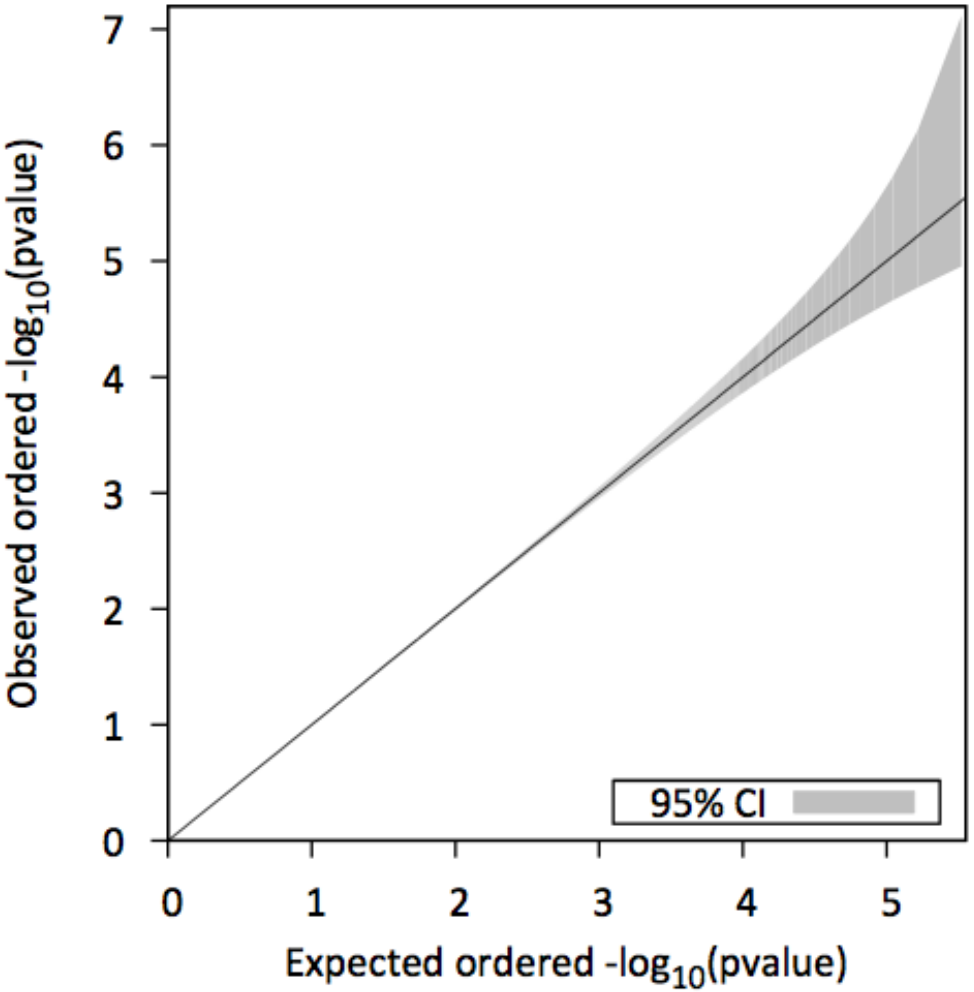
$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma_g^2 K + \sigma_e^2 I)$$ K_{ij} : kinship coefficient

- Similar model for population-based analysis to account for distant relationship inferred from dense SNP arrays

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma_g^2 \hat{K} + \sigma_e^2 I)$$ $\hat{K}_{ij}$: marker-based kinship coeff.

Kang HM et al, Nat Genet (2010) 42:348-54

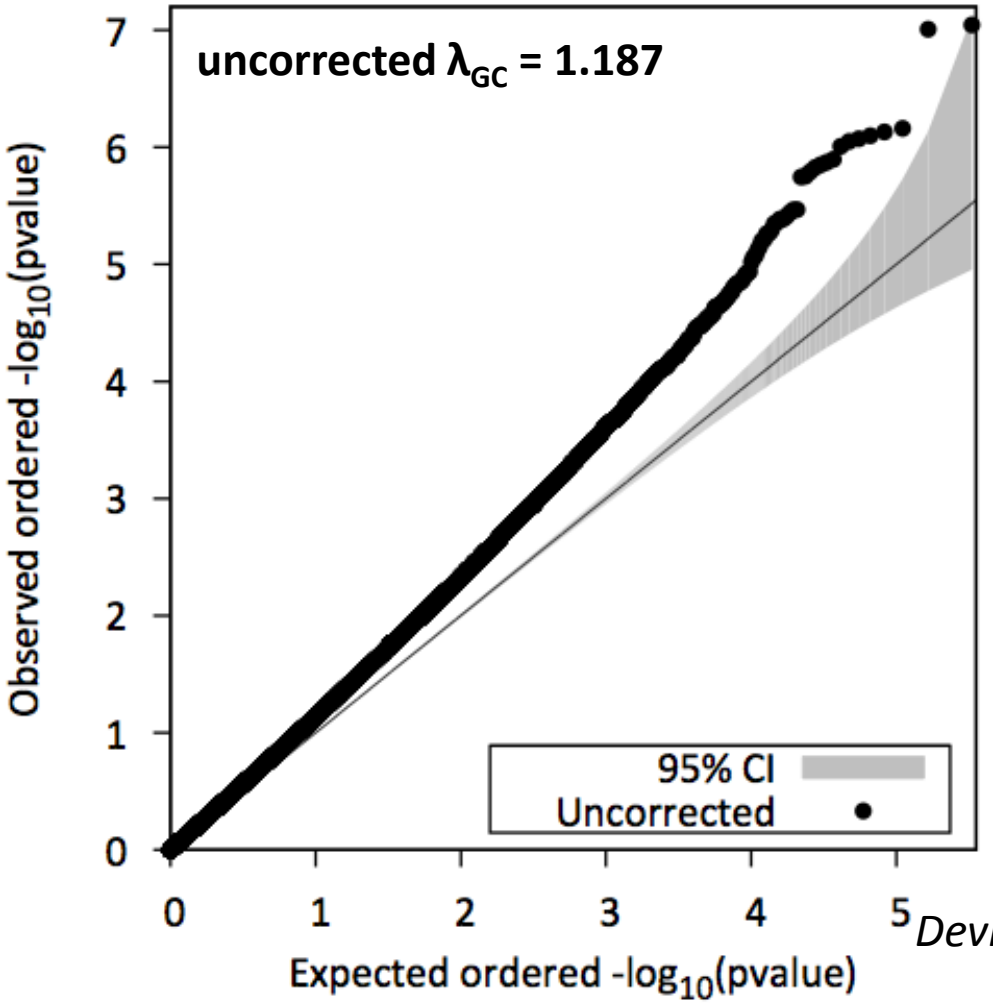
Genome-wide association of human height



- NFBC 1966 birth cohort
  - *Sabatti et al, Nat Genet (2008) 41:35-46*
- Illumina 370,000 SNPs
- 5,326 unrelated individuals

Uncorrected analysis

- Overdispersion of test statistics -

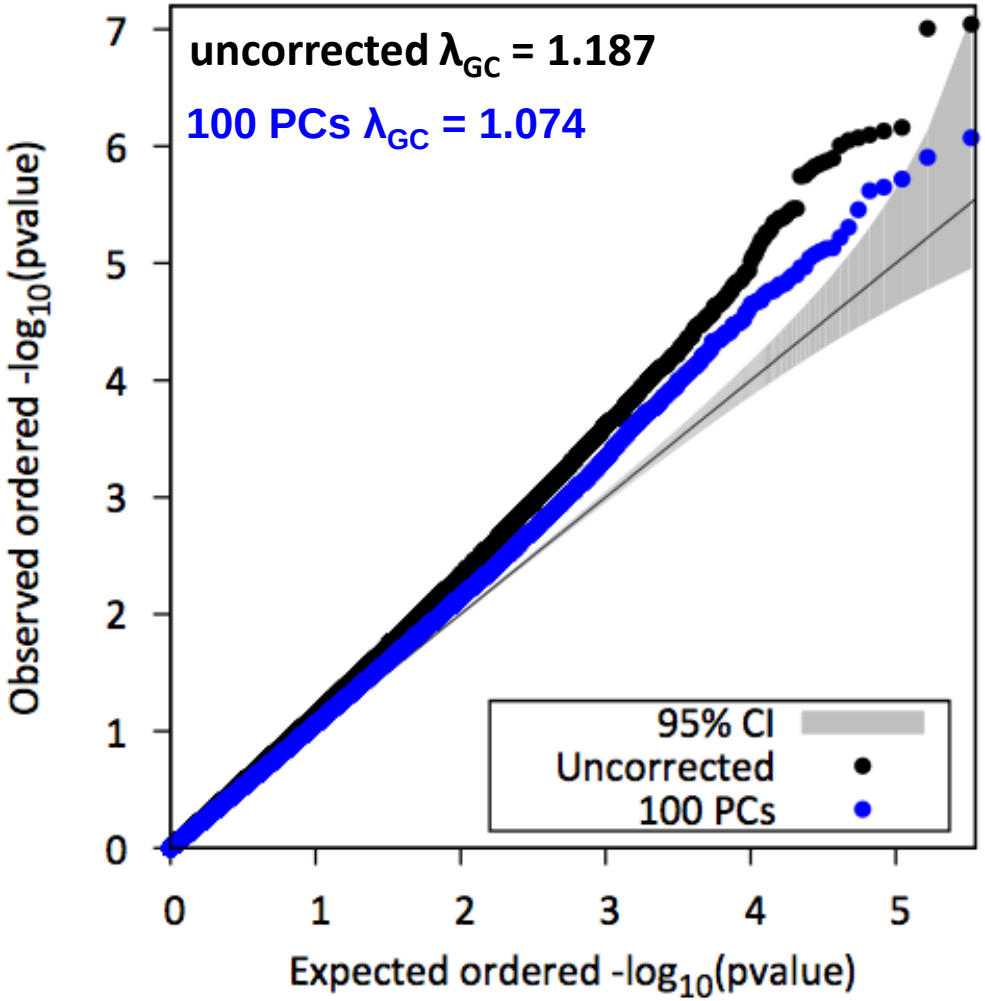


$$\lambda_{GC} = \frac{\text{median}\{T_1, T_2, \cdots, T_n\}}{\mathbf{E}[\text{median}\{T\}]}$$

Devlin & Roeder Biometrics (1999) 55:997-1004

Conditioning on principal components

- Overdispersion still exists -



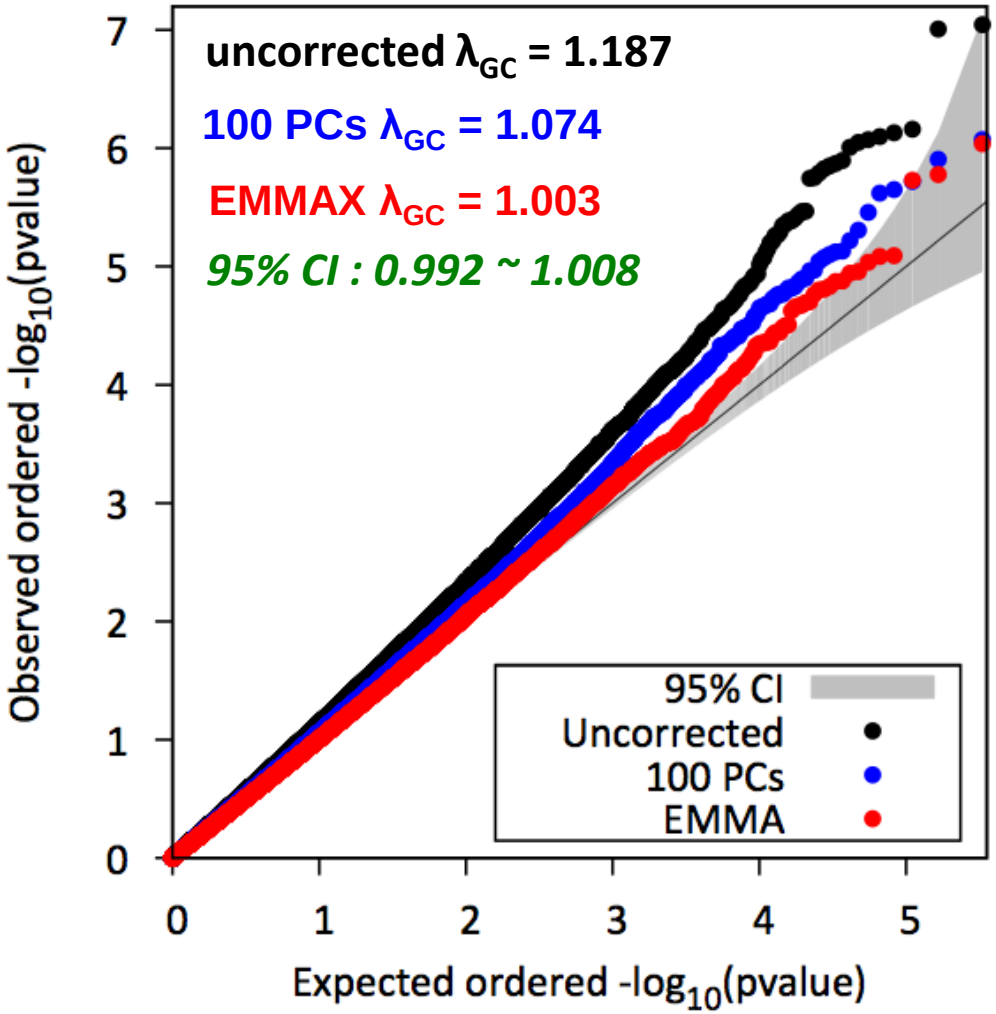
$$\mathbf{y} = \mu + \mathbf{x}\beta + G\gamma + \mathbf{e}$$

- G is top k(=100) eigenvectors of kinship matrix K
- λ_{GC} from 1.187 to 1.074
- λ_{GC} is still substantially higher than expected
- Corrects for population structure, but not hidden relatedness

Price AL et al, Nat Genet (2006) 38:904-909

Variance component model

- Overdispersion resolved -



$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{x}\beta + \mathbf{u} + \mathbf{e} \\ \mathrm{Var}(\mathbf{u}) &= \sigma_g^2 K \\ \mathrm{Var}(\mathbf{e}) &= \sigma_e^2 I \end{aligned}$$

- Using **EMMAX** reduced λ_{GC} from 1.187 to 1.003
- λ_{GC} falls into 95% confidence intervals

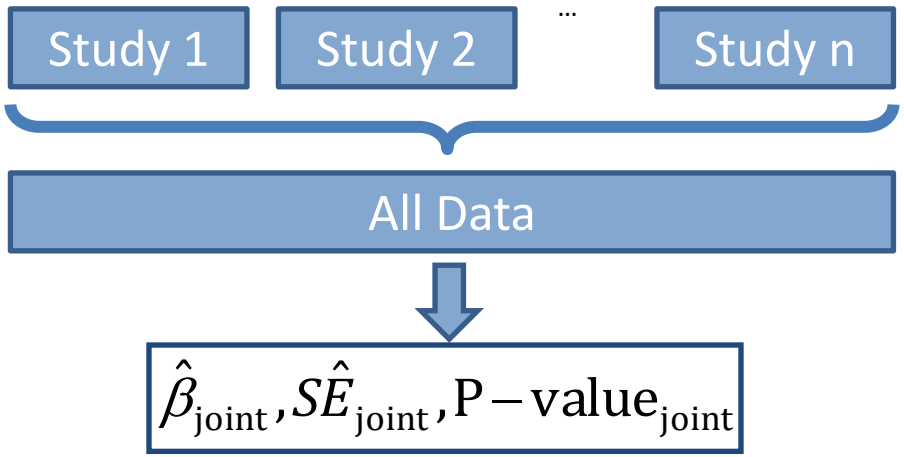
Multiple genetic association studies

- Most associated common variants have small effect sizes (e.g. odds ratios [OR] < 1.2)

- To increase power to detect small genetic effect sizes, combine information across studies using
  - Meta-analysis of study-level association results
  - Joint analysis of all individual-level data

Multiple studies: Data aggregation methods

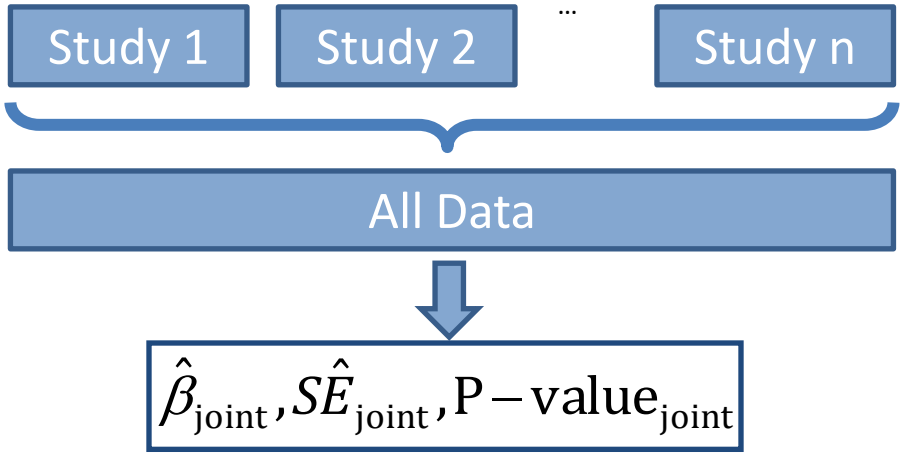
Joint analysis



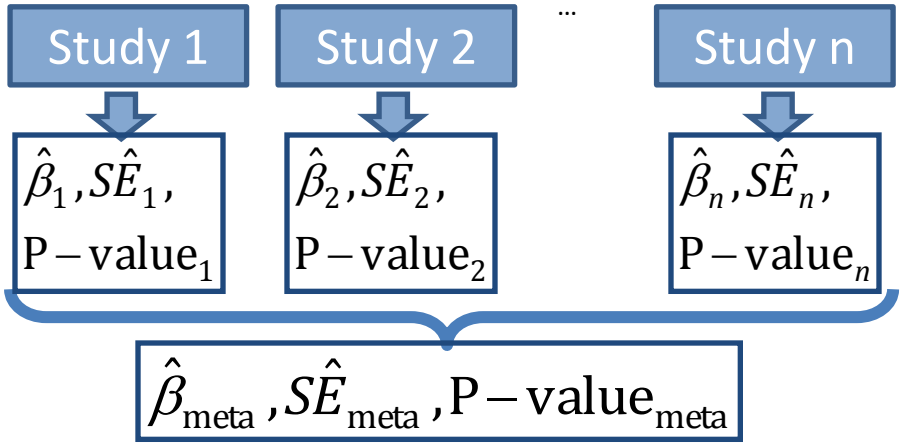
- Combine individual-level data and analyze jointly

Multiple studies: Data aggregation methods

Joint analysis



Meta-analysis



- Combine study-level association results using:
  - Inverse-variance weights
  - Sample-size weights

Joint vs. meta-analysis

- For common variants, both joint and meta-analysis are both well-calibrated, and have near-equivalent power
- Meta-analysis is more commonly used
  - Sharing individual-level data is difficult due to logistical and ethical restrictions
- Combining multiple studies is critical to increase power to detect small effect sizes

(Lin & Zeng, *Genet. Epidemiol.*, 2010)

Summary: analysis of common variants

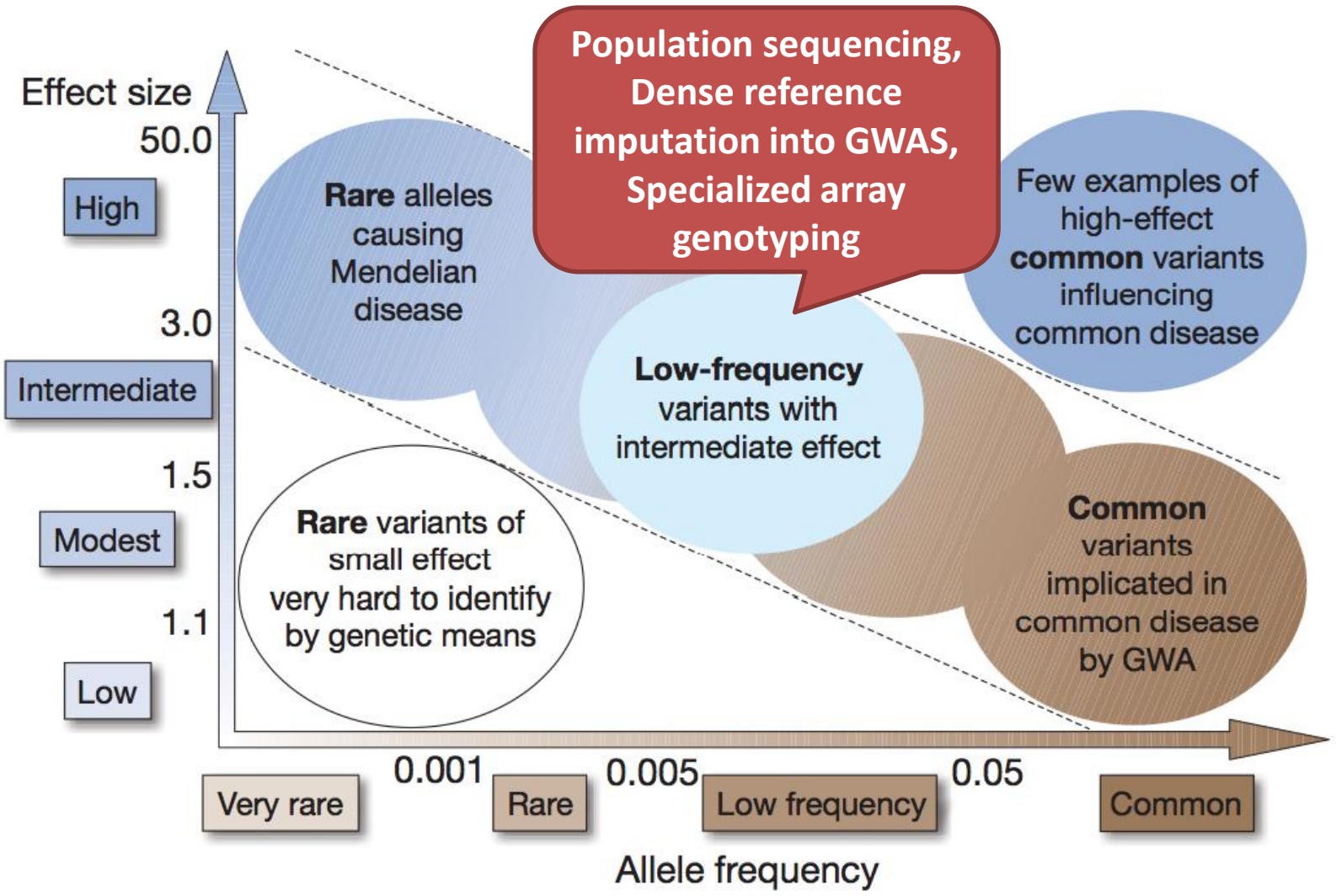
- Single variant analysis with regression-based methods have identified many trait-associated genetic variants

- Important to account for population structure and/or sample relatedness to avoid spurious association

Genetic Association Analysis

ANALYSIS OF LOW-FREQUENCY AND RARE VARIANTS

Genetic architecture of complex traits



Why study rare variants?

COMPLETE GENETIC ARCHITECTURE OF EACH TRAIT

- **Are there additional susceptibility loci to be found?**
- **What is the contribution of each identified locus to a trait?**
  - Sequencing, imputation and new arrays describe variation more fully
  - Rare variants are plentiful and should identify new susceptibility loci

UNDERSTAND FUNCTION LINKING EACH LOCUS TO A TRAIT

- **Do we have new targets for therapy?**
  What happens in gene knockouts?
  - Use sequencing to find rare human “knockout” alleles
  - Good: Results may be more clear than for animal studies
  - Bad: Naturally occurring knockout alleles are extremely rare

Why study rare variants?

COMPLETE GENETIC ARCHITECTURE OF EACH TRAIT

- Are there additional susceptibility loci to be found?
- Wh…?
  - …e fully
  - …y loci

Coding Variants Especially Useful!

UNDERSTAND FUNCTION LINKING EACH LOCUS TO A TRAIT

- **Do we have new targets for therapy?**
  What happens in gene knockouts?
  - Use sequencing to find rare human "knockout" alleles
  - Good: Results may be more clear than for animal studies
  - Bad: Naturally occurring knockout alleles are extremely rare

Lots of rare functional variants to discover

SET	# SNPs	Singletons	Doubletons	Tripletons	>3 Occurrences
Synonymous	270,263	128,319 (47%)	29,340 (11%)	13,129 (5%)	99,475 (37%)
Nonsynonymous	410,956	234,633 (57%)	46,740 (11%)	19,274 (5%)	110,309 (27%)
Nonsense	8,913	6,196 (70%)	926 (10%)	326 (4%)	1,465 (16%)
Non-Syn / Syn Ratio		1.8 to 1	1.6 to 1	1.4 to 1	1.1 to 1

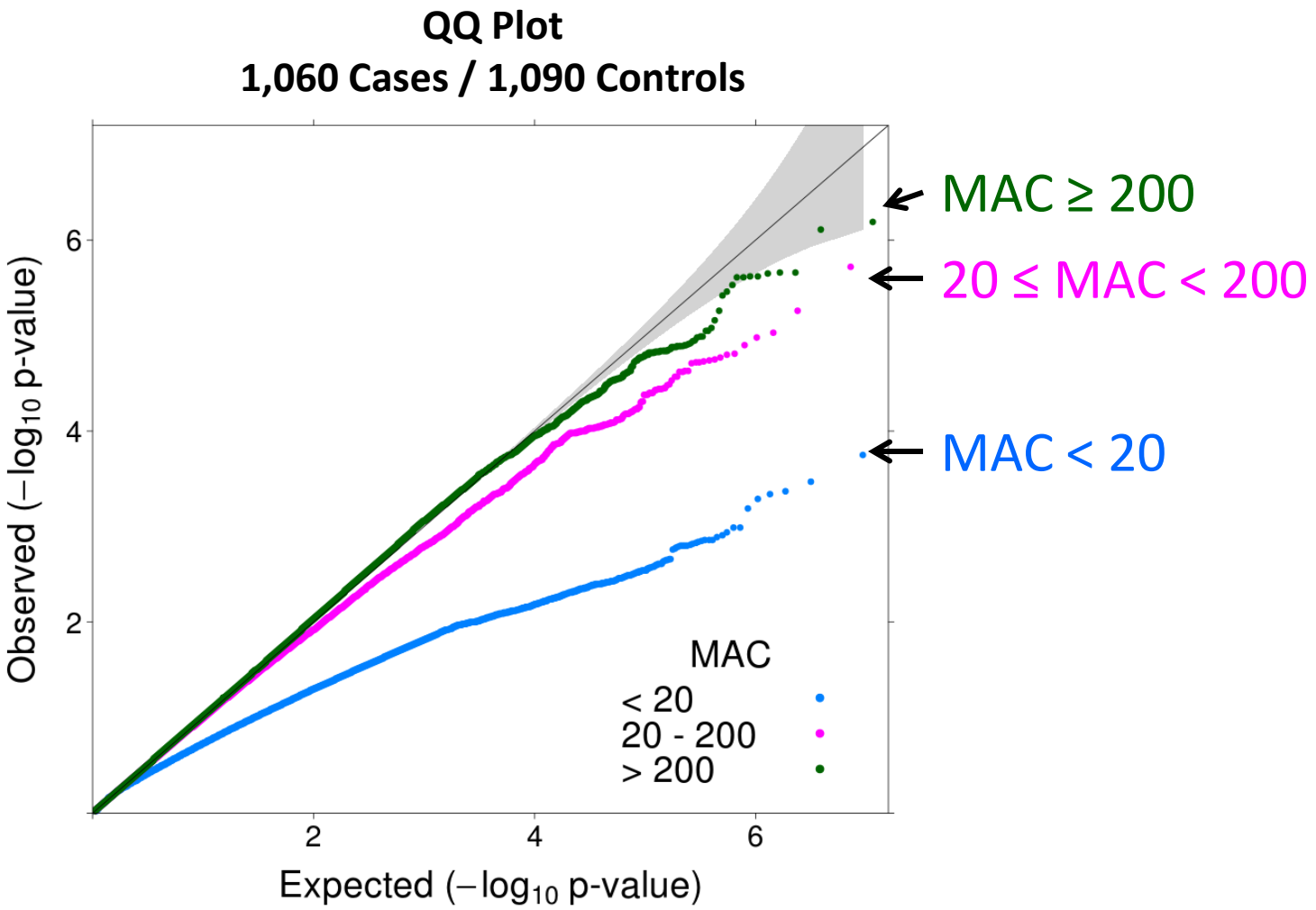
There is a very large reservoir of extremely rare, likely functional, coding variants.

NHLBI Exome Sequencing Project

Challenges for association testing of low-frequency variants

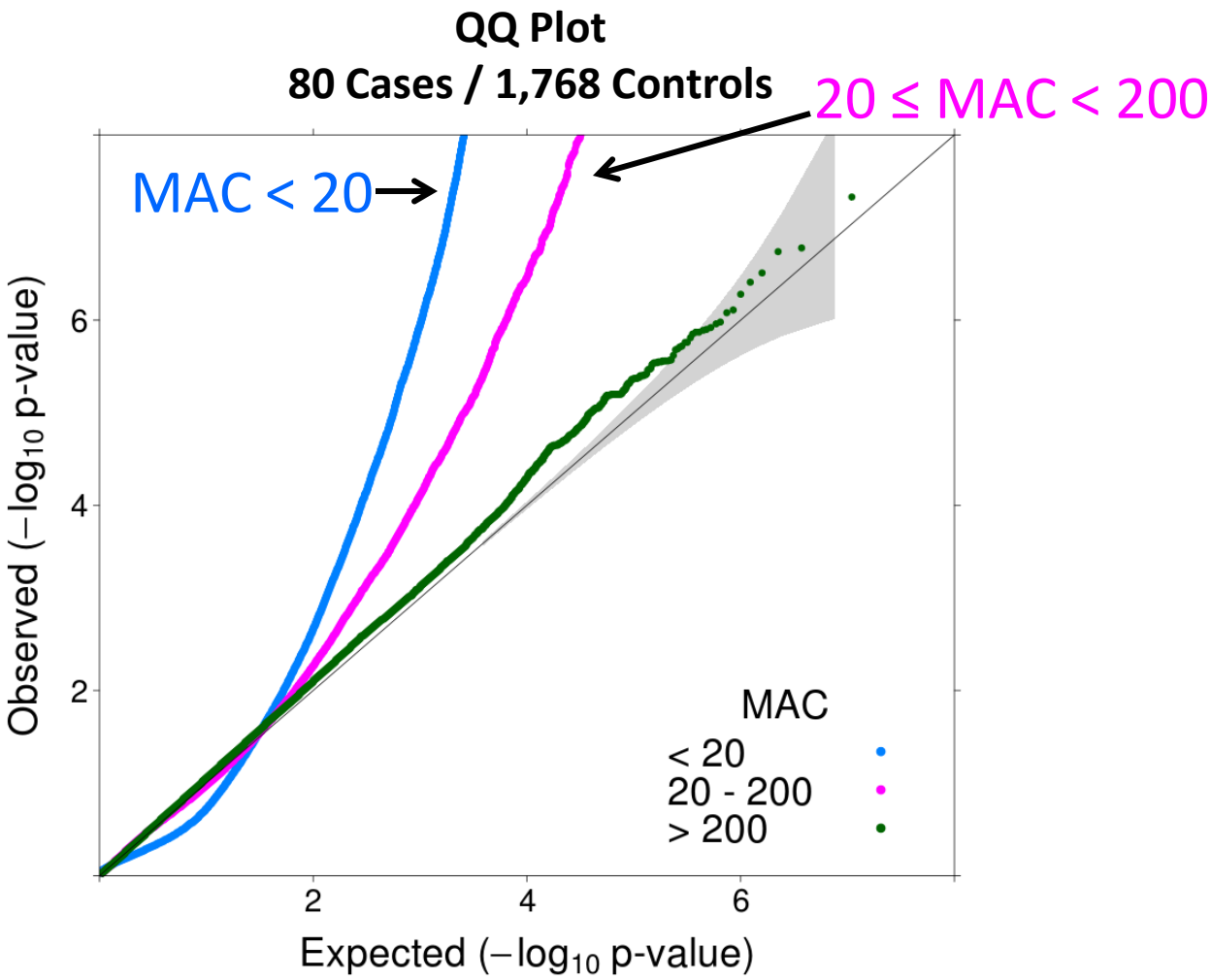
- Low minor allele count (MAC)
- Stringent $\alpha = 5x10^{-8}$ (multiple testing)
- For binary traits:
  - Unbalanced numbers of cases and controls (e.g. population-based studies)

Logistic Wald test has low power* for low-frequency and rare variants in balanced studies



*Recently noted by Xing et al. (2012) *Ann Hum Genet* 76:168-77

Logistic score test is anti-conservative for low-frequency and rare variants in unbalanced studies



Recommended single marker tests for low-frequency variants

Binary Traits

- For balanced studies (case-control ratio < 3:2)
  - Use Firth bias-corrected*, or score logistic regression
  - Avoid Wald test (low power)
- For unbalanced studies (case-control ratio > 3:2)
  - Use Firth, likelihood ratio logistic regression
  - Avoid score test (inflated false positive rate)

Quantitative Traits

- Given normally-distributed QTs
  - Use any linear regression test

*(Firth, *Biometrika*, 1993)

(Ma et. al. *Genet. Epidemiol.*, 2013; Ma et. al., *in preparation*)

Limitations of single marker tests

- Single marker tests have low power for rare variants unless sample size very large
- For binary traits, variants require minimum MAC ≥ 26 to have p-values < $5x10^{-8}$:
  (No covariates; $N_{cases}=N_{ctrls}$)

	Cases	Ctrls
Genotype = AA	975	1000
Genotype = Aa	**25**	0
	1000	1000

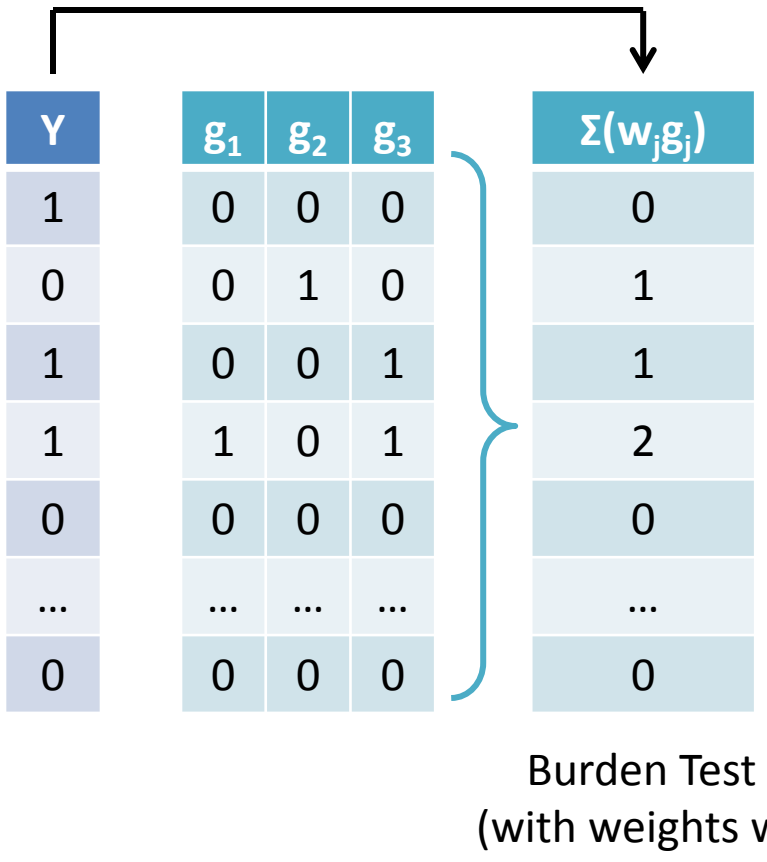
Fisher's Exact Test p = **$5.1x10^{-8}$**

	Cases	Ctrls
Genotype = AA	974	1000
Genotype = Aa	**26**	0
	1000	1000

Fisher's Exact Test p = **$2.5x10^{-8}$**

Gene-based tests

- Gene-based tests jointly analyze multiple rare variants in genetic region (e.g. gene)
- Increases power by:
  - Combining information across rare variants
  - Requiring less stringent α, e.g. $\alpha = 2.5 \times 10^{-6}$ for 20K genes



(Madsen & Browning, *PLoS Genet.*, 2009)

Selecting variants for gene-based tests

- If include variants of all frequencies, non-causal and common variants will dilute signal
- Commonly used filters or “masks”:
  - Include variants MAF ≤ 0.05 or 0.01
  - Weight variants by MAF
    - E.g. w_j ~ Beta(MAF, 1, 25)
  - Select variants based on functional annotation:
    - E.g. Protein Truncating Variants only, nonsynonymous, missense, etc.
- If mask is too restrictive, will reduce to single variant test, and no gain in power

Categories of aggregation tests

- **Burden tests** test association between (weighted) sum of rare alleles with disease or QT
  - CMC (Li & Leal, 2008), WSS (Madsen & Browning, 2009)

- **Dispersion tests** measure deviations from expected distribution
  - SKAT (Wu et al., 2011), C-alpha (Neale et al., 2011)

- **Combined tests** combine strengths of burden and dispersion tests
  - SKAT-O (Lee et al., 2012)

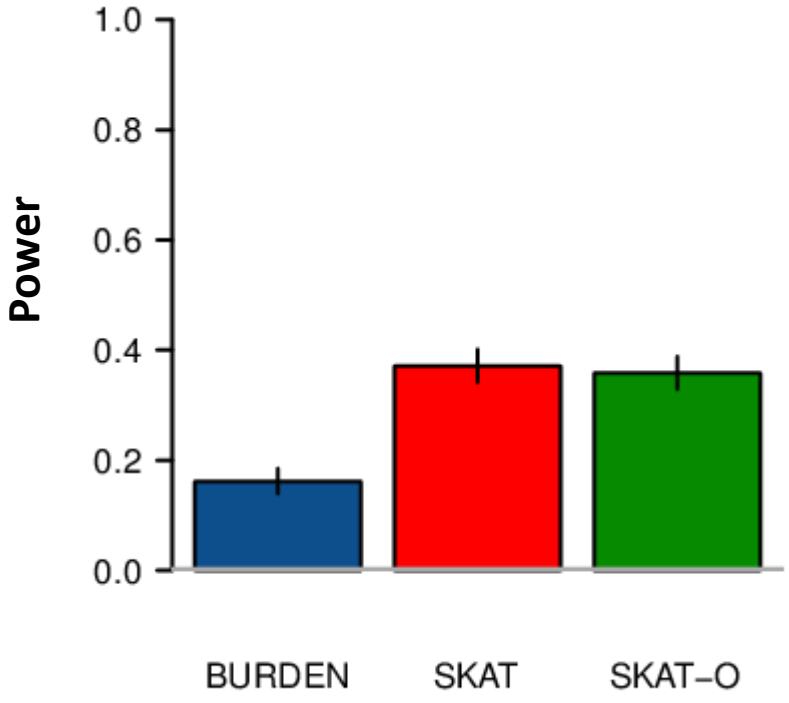
Power of gene-based tests

- Power of gene-based tests affected by the underlying (unknown) genetic architecture of the analyzed region:
  - Number of associated variants in region
  - Number of neutral variants diluting signals
  - Whether direction of effect is consistent within gene

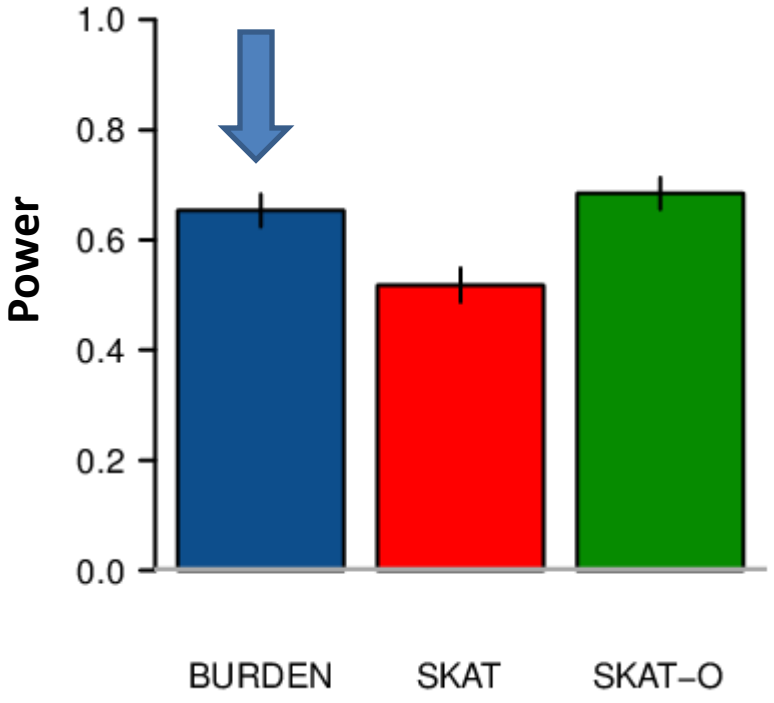
Power comparison

(All causal variants 100% deleterious)

10% Variants in Region are Causal



50% Variants in Region are Causal



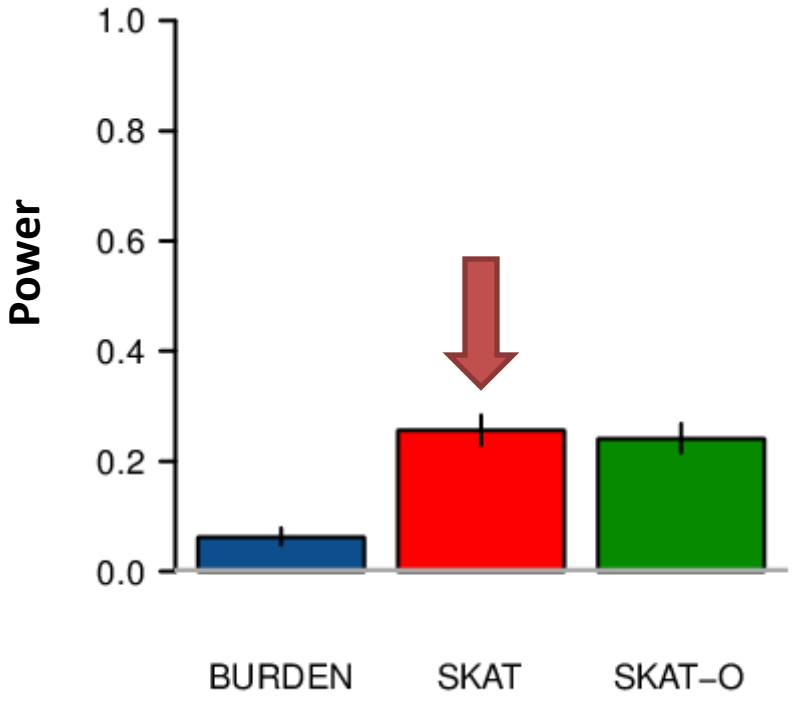
Burden is most powerful when there are many causal variants with same direction of effect

(Ma et al., *in preparation*)

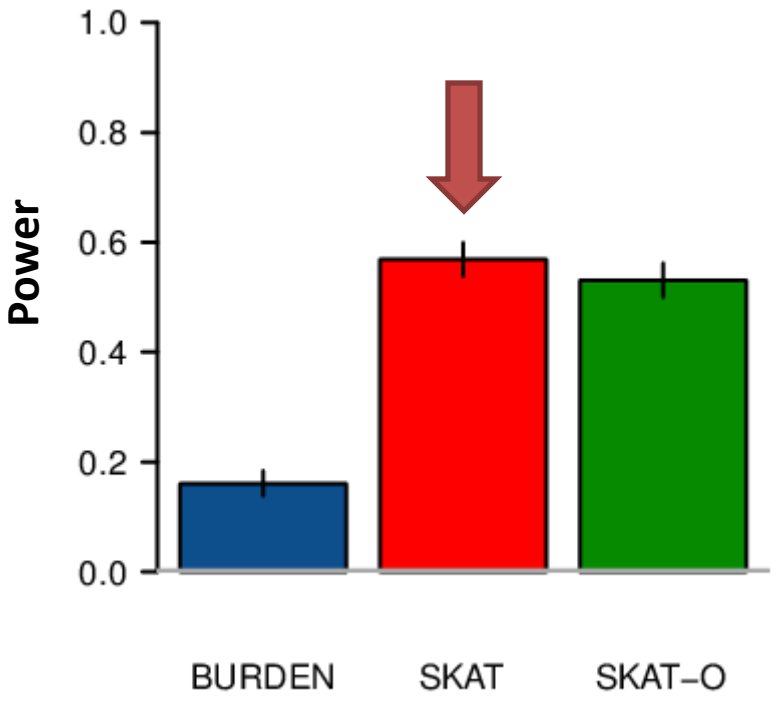
Power comparison

(Causal variants are 50% deleterious / 50% protective)

10% Variants in Region are Causal



50% Variants in Region are Causal



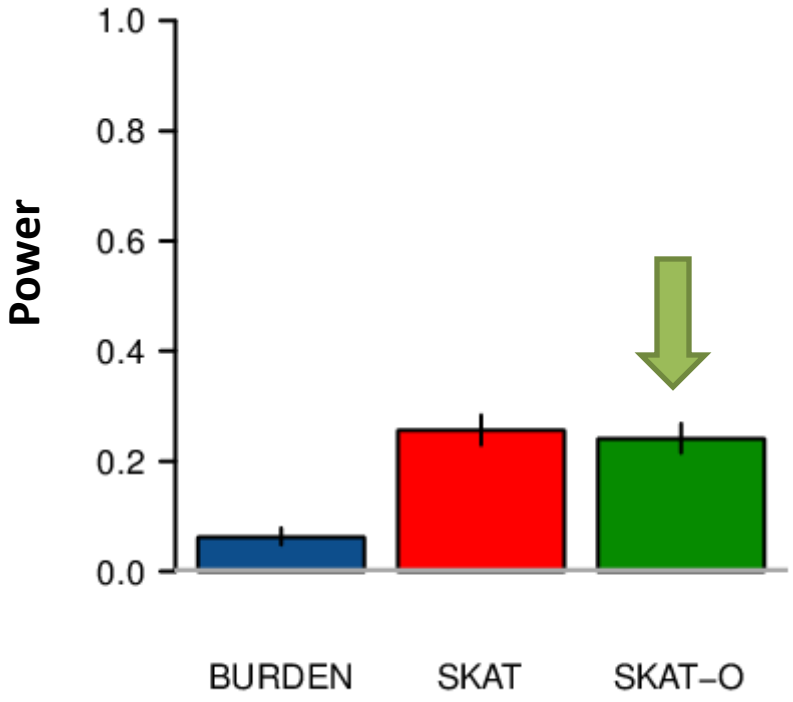
SKAT is most powerful when there are causal variants with opposite direction of effects

(Ma et al., *in preparation*)

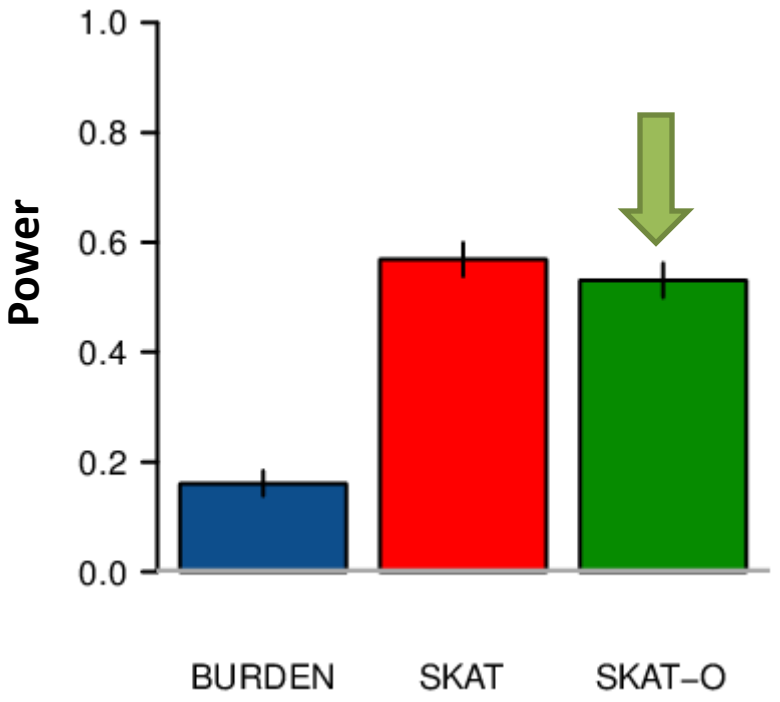
Power comparison

(Causal variants are 50% deleterious / 50% protective)

10% Variants in Region are Causal



50% Variants in Region are Causal



SKAT-O is generally powerful and robust for different genetic architectures

(Ma et al., *in preparation*)

Summary: analysis of low-frequency variants

- Single marker tests remain useful for low-frequency variants
  - Need to carefully select well-calibrated tests
- Gene-based tests can be more powerful for rare variants
  - Power generally determined by underlying genetic architecture