Genome-Wide Association Studies

Yuan Jiang
Statistics, Oregon State University

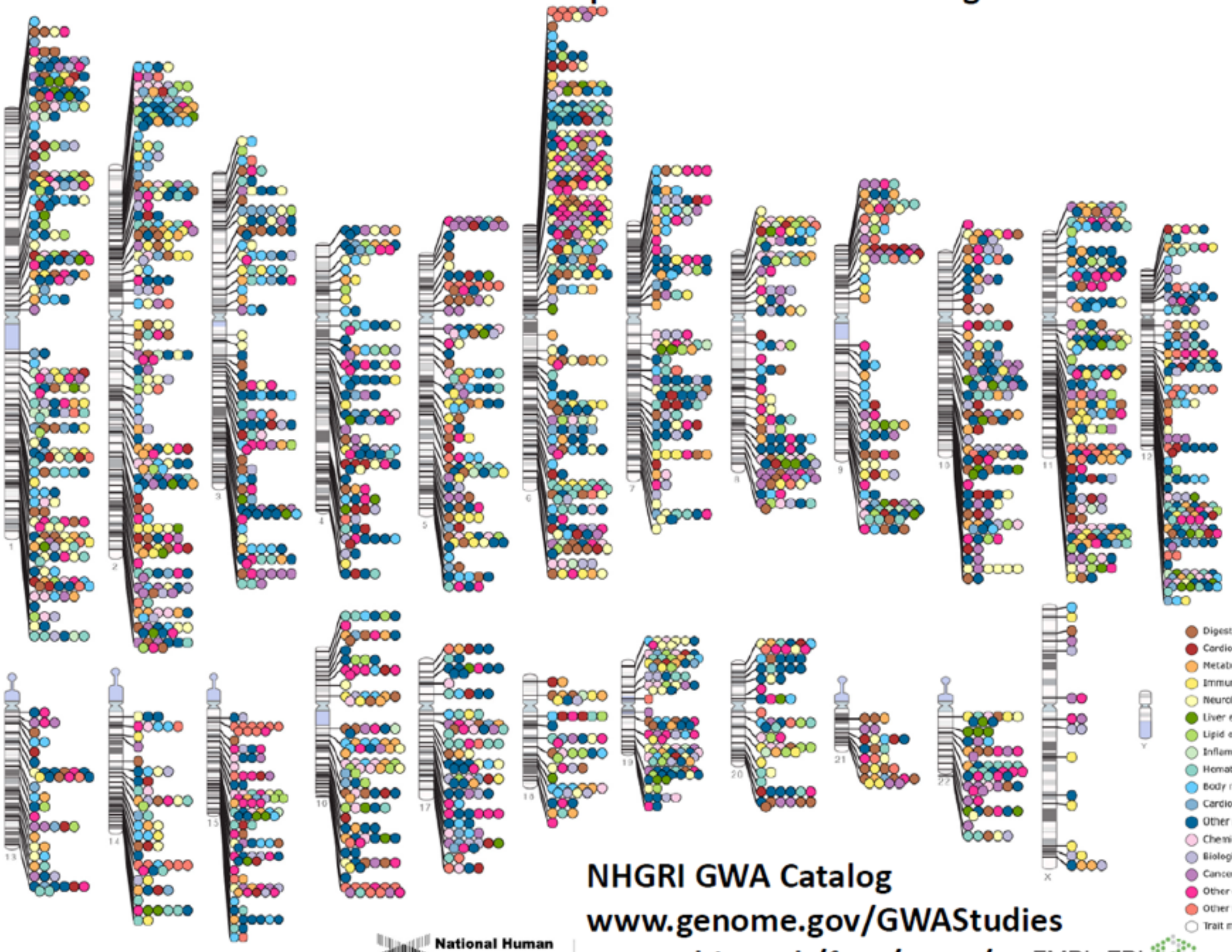
Outline

- Overview
- Concepts
- Analysis
- Discussions

What is GWAS?

Published Genome-Wide Associations through 07/2012

Published GWA at p≤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

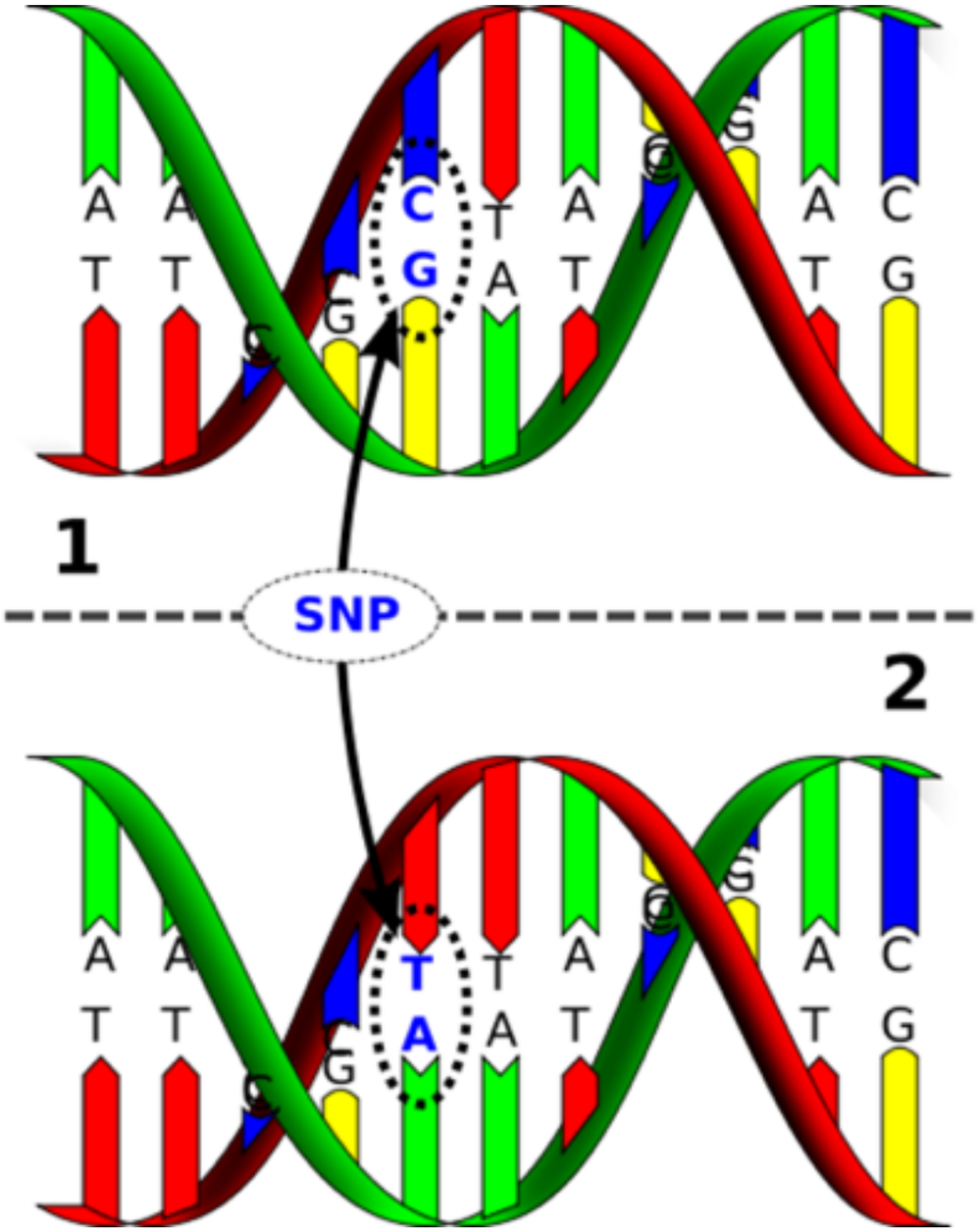
Single Nucleotide Polymorphisms

- A single nucleotide — A, T, C or G — in the genome differs between members of a biological species

- All common SNPs have only two alleles:

  AAGC**C**TA
  AAGC**T**TA

- dbSNP is a SNP database from the National Center for Biotechnology Information (NCBI)

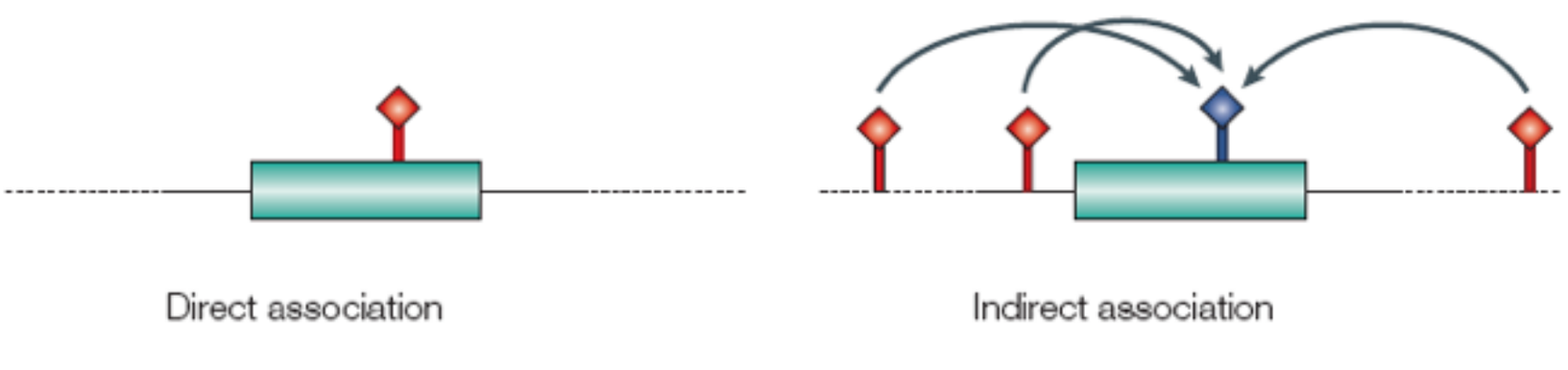


Chromosome	Length(bp)	All SNPs		TSC SNPs	
		SNPs	kb per SNP	SNPs	kb per SNP
1	214,066,000	129,931	1.65	75,166	2.85
2	222,889,000	103,664	2.15	76,985	2.90
3	186,938,000	93,140	2.01	63,669	2.94
4	169,035,000	84,426	2.00	65,719	2.57
5	170,954,000	117,882	1.45	63,545	2.69
6	165,022,000	96,317	1.71	53,797	3.07
7	149,414,000	71,752	2.08	42,327	3.53
8	125,148,000	57,834	2.16	42,653	2.93
9	107,440,000	62,013	1.73	43,020	2.50
10	127,894,000	61,298	2.09	42,466	3.01
11	129,193,000	84,663	1.53	47,621	2.71
12	125,198,000	59,245	2.11	38,136	3.28
13	93,711,000	53,093	1.77	35,745	2.62
14	89,344,000	44,112	2.03	29,746	3.00
15	73,467,000	37,814	1.94	26,524	2.77
16	74,037,000	38,735	1.91	23,328	3.17
17	73,367,000	34,621	2.12	19,396	3.78
18	73,078,000	45,135	1.62	27,028	2.70
19	56,044,000	25,676	2.18	11,185	5.01
20	63,317,000	29,478	2.15	17,051	3.71
21	33,824,000	20,916	1.62	9,103	3.72
22	33,786,000	28,410	1.19	11,056	3.06
X	131,245,000	34,842	3.77	20,400	6.43
Y	21,753,000	4,193	5.19	1,784	12.19
RefSeq	15,696,674	14,534			1.08
Totals	2,710,164,000	1,419,190	1.91	887,450	3.05

Common vs. Rare

- Within a population, SNPs can be assigned a minor allele frequency (MAF)
- A SNP allele that is common in one geographical or ethnic group may be much rarer in another.
- Common SNP: MAF > 1%
- Rare SNP: MAF < 1%

Direct vs. Indirect

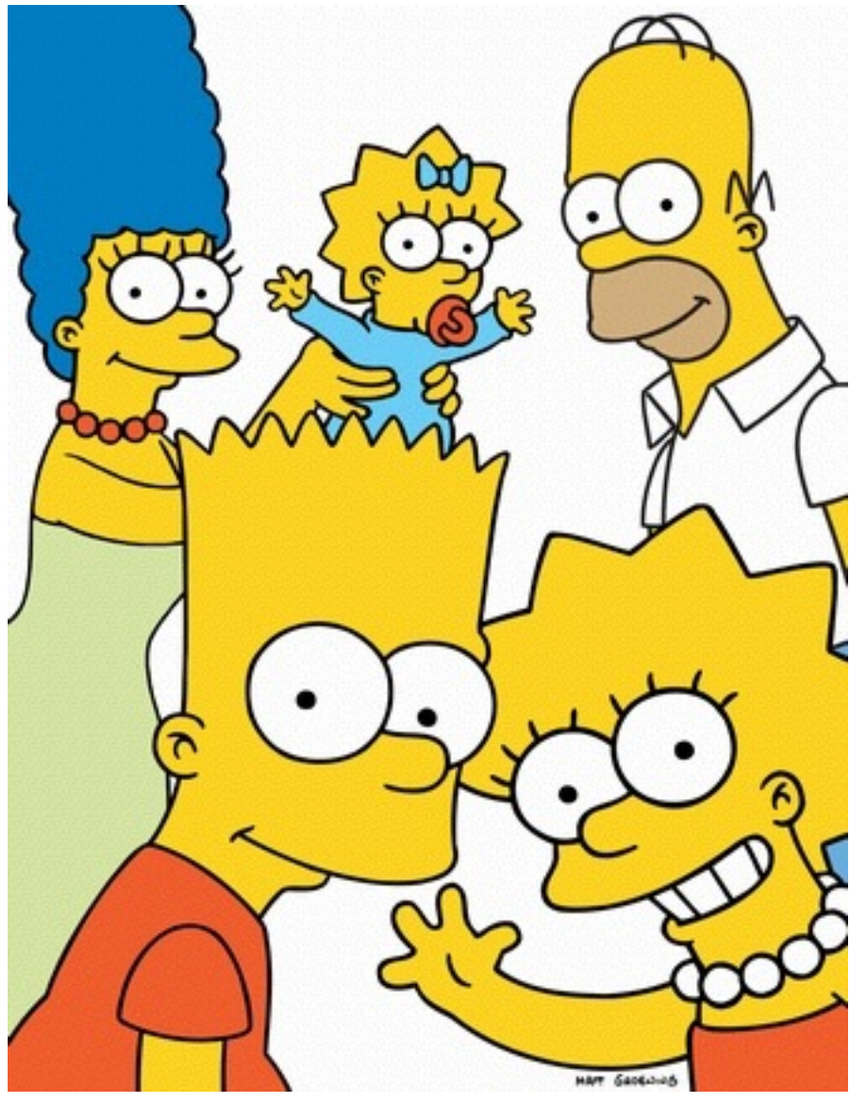


Direct association

Indirect association

- Left Panel: disease locus directly typed
- Right Panel: markers correlated with disease locus

Population vs. Families



Ongoing Biggest Studies

- GWAS Microarray: 100,000 People in the Kaiser RPGEH, still to be analyzed
- Sequencing: 1000 Genomes Project, UK10K
- Exome Sequencing: GO ESP (12,031 subjects, for exome microarray design)

Quality Control

- “Garbage rises to the top!”
- Filter SNPs and Individuals
  - MAF (e.g., >1% but sample size dependent)
  - Low call rates (genotyping did not work correctly)
  - Test for Hardy–Weinberg equilibrium within controls & ethnic groups
    - p-value (e.g., > 10^(-3) but opinions vary)
  - Check for relatedness
    - MZ twins, or accidentally genotyped same sample twice?
  - Check genotype gender
  - Filter Mendelian inheritance

Hardy-Weinberg Equilibrium

Seven assumptions underlying HWE:

- organisms are diploid
- only sexual reproduction occurs
- generations are non overlapping
- mating is random
- population size is infinitely large
- allele frequencies are equal in the sexes
- there is no migration, mutation or selection

Hardy-Weinberg Equilibrium

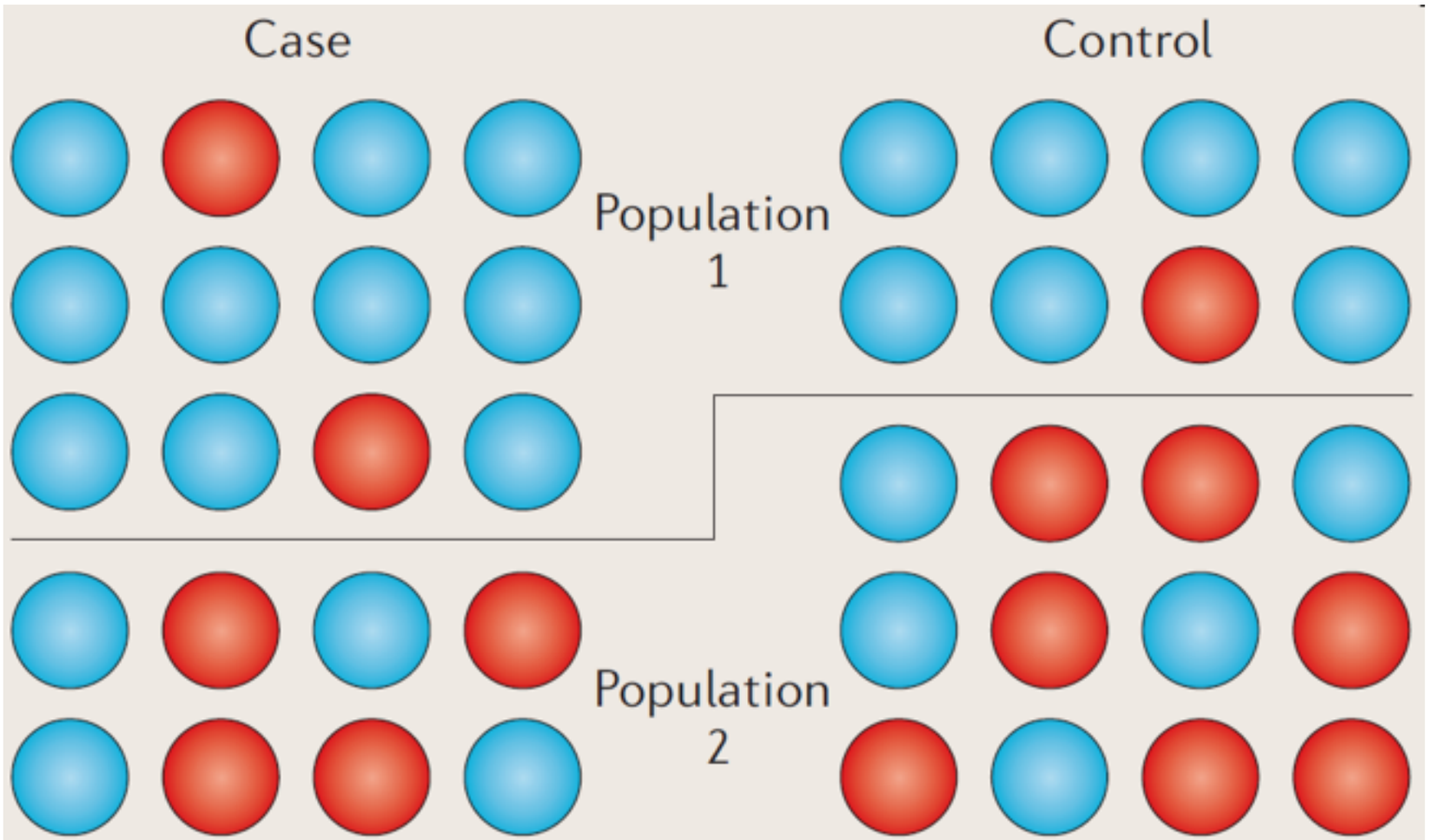
- Statistically, these assumptions guarantee the independence between the two alleles A and a in a population
- $P(A) = p$, $P(a) = q$, with $p + q = 1$
- Then, $P(AA) = p^2$, $P(Aa \text{ or } aA) = 2pq$, $P(aa) = q^2$
- Pearson Chi-Squared test

Filter for Mendelian Inheritance

- Offspring genotypes follow the Mendelian rule
  - Parents: AA & AA, Child: AT
  - It can be a mutation, but more likely a genotyping error

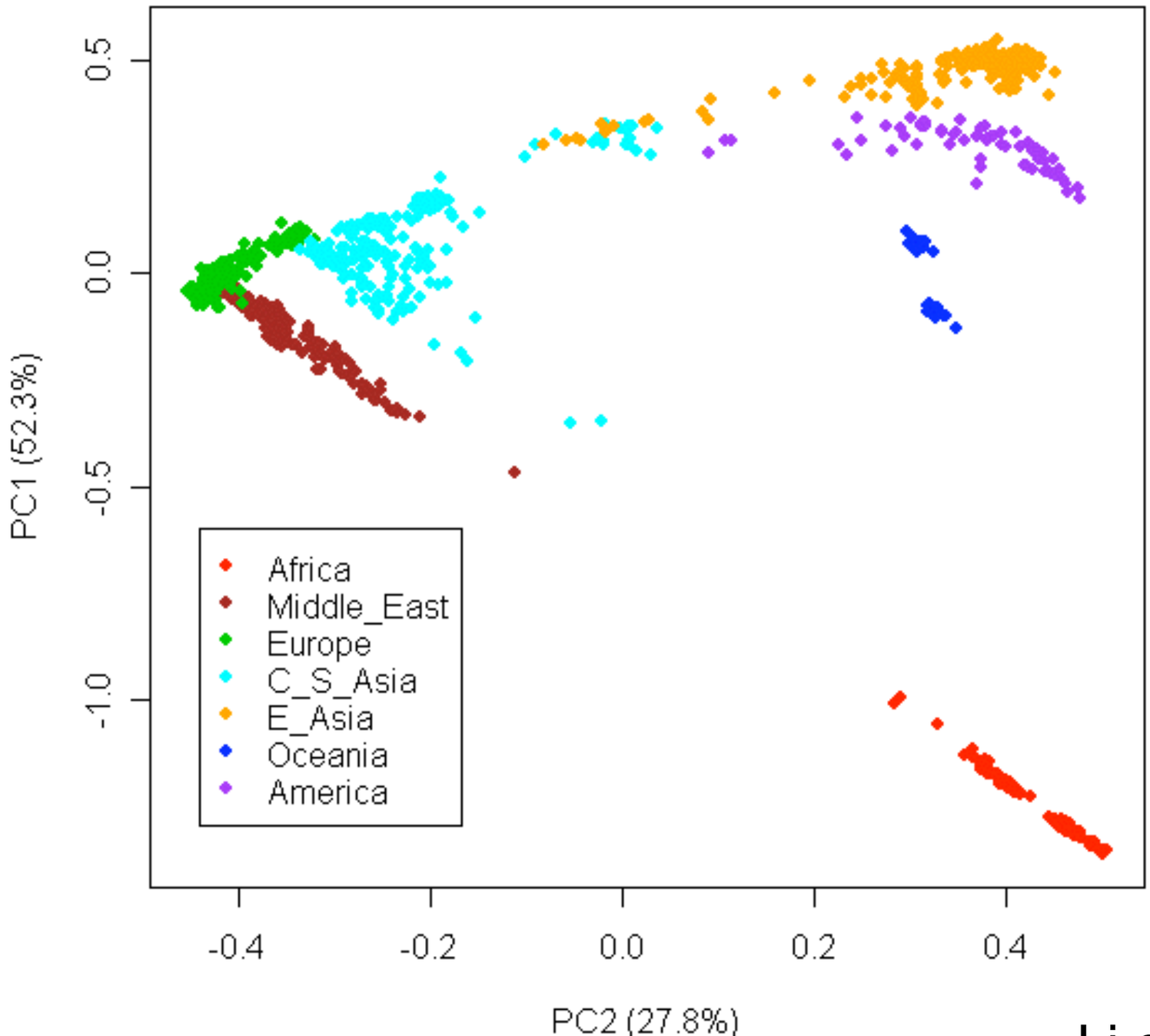
Population Substructure

- Cases disproportionately represent population 1
- Any SNP with allele proportions that differ between two populations will be associated (spurious association)



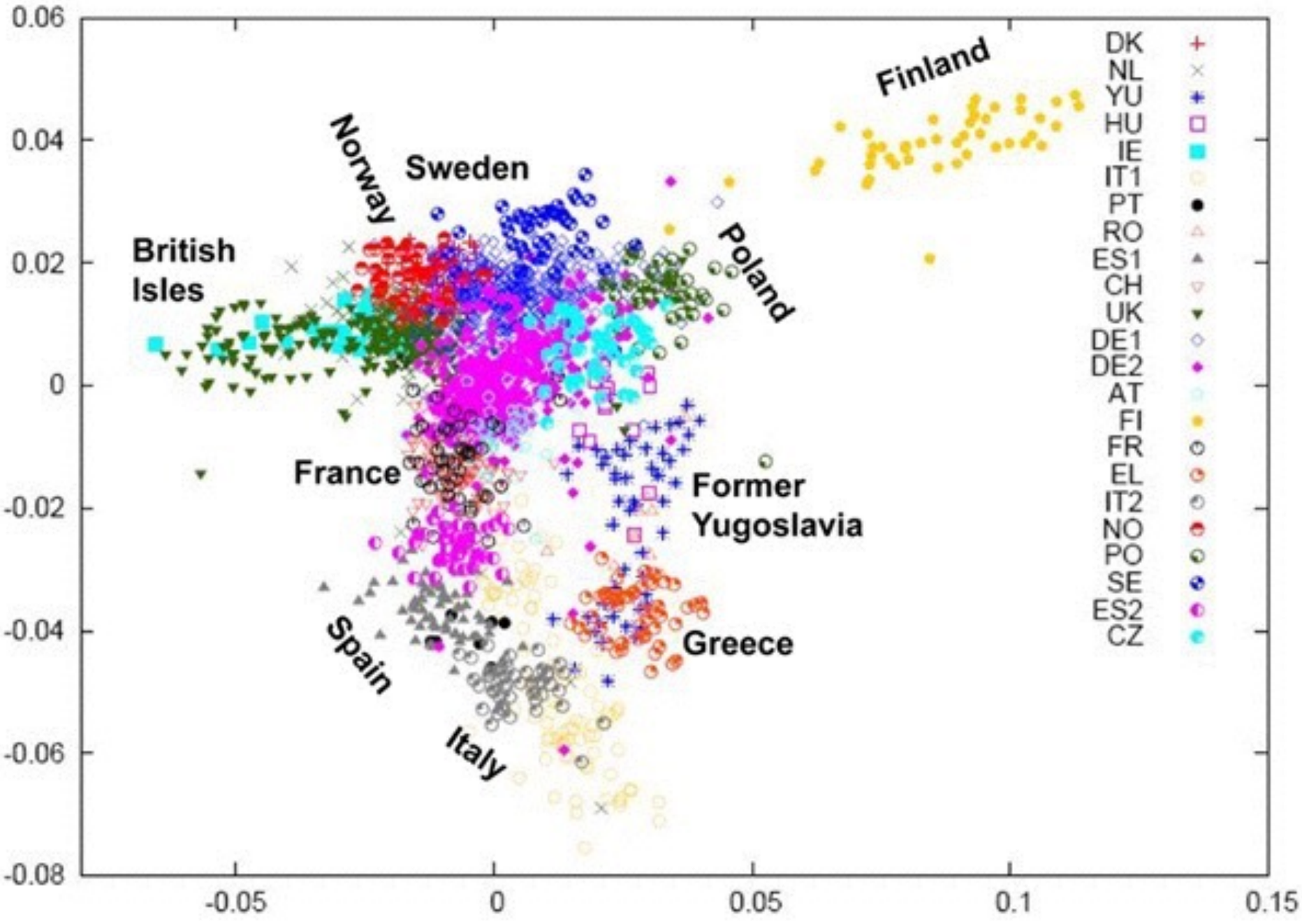
Balding, Nature Reviews Genetics 2010

Principal Components (PC's)



Li et al., Science 2008

Principal Components (PC's)



Razib, Current Biology 2008

GWAS Analysis

- Most common approach: single SNP analysis
- SNP with additive, dominate, and recessive codings
- Statistical models:
  - Continuous phenotype: linear regression
  - Dichotomous phenotype: Cochran–Armitage trend test
  - Dichotomous phenotype: logistic regression
- Adjust for population stratification, and other environmental factors (including them in the regression model)
- Multiple comparison correction

Dichotomous Phenotype

	aa	Aa	AA	Total
Control	20	20	20	60
Case	10	20	30	60
Total	30	40	50	120

SNP Codings

A: major allele, a: minor allele

- Additive: AA = 0, Aa = aA = 1, aa = 2
- Dominant: AA = 0, Aa = aA = aa = 1
- Recessive: AA = Aa = aA = 0, aa = 1

Cochran–Armitage Trend Test

	aa	Aa	AA	Total
Control	N_{11}	N_{12}	N_{13}	R_1
Case	N_{21}	N_{22}	N_{23}	R_2
Total	C_1	C_2	C_3	N

- Revised from Pearson Chi-Squared test to incorporate ordering in the effects
- Test statistic $T = \sum_i t_i (N_{1i} * R_2 - N_{2i} * R_1)$
  - t = (1,1,0) optimal for dominant coding
  - t = (0,1,1) optimal for recessive coding
  - t = (2,1,0) optimal for additive coding

Cochran–Armitage Trend Test

	aa	Aa	AA	Total
Control	20	20	20	60
Case	10	20	30	60
Total	30	40	50	120

- Pearson Chi-Squared test normalized statistic: 2
- Trend test normalized statistics:
  - Dominant:1.85
  - Recessive: -2.1
  - Additive: 2.3
- Trend test is more powerful than Pearson

Regression Models

- Linear regression for continuous phenotype
- Logistic regression for dichotomous phenotype
- Poisson regression for count phenotype
- Ordinal logistic regression for ordinary phenotype
- And so on…

Hypothesis Tests

- A statistical test procedure is comparable to a criminal trial; a defendant is considered not guilty as long as his or her guilt is not proven.

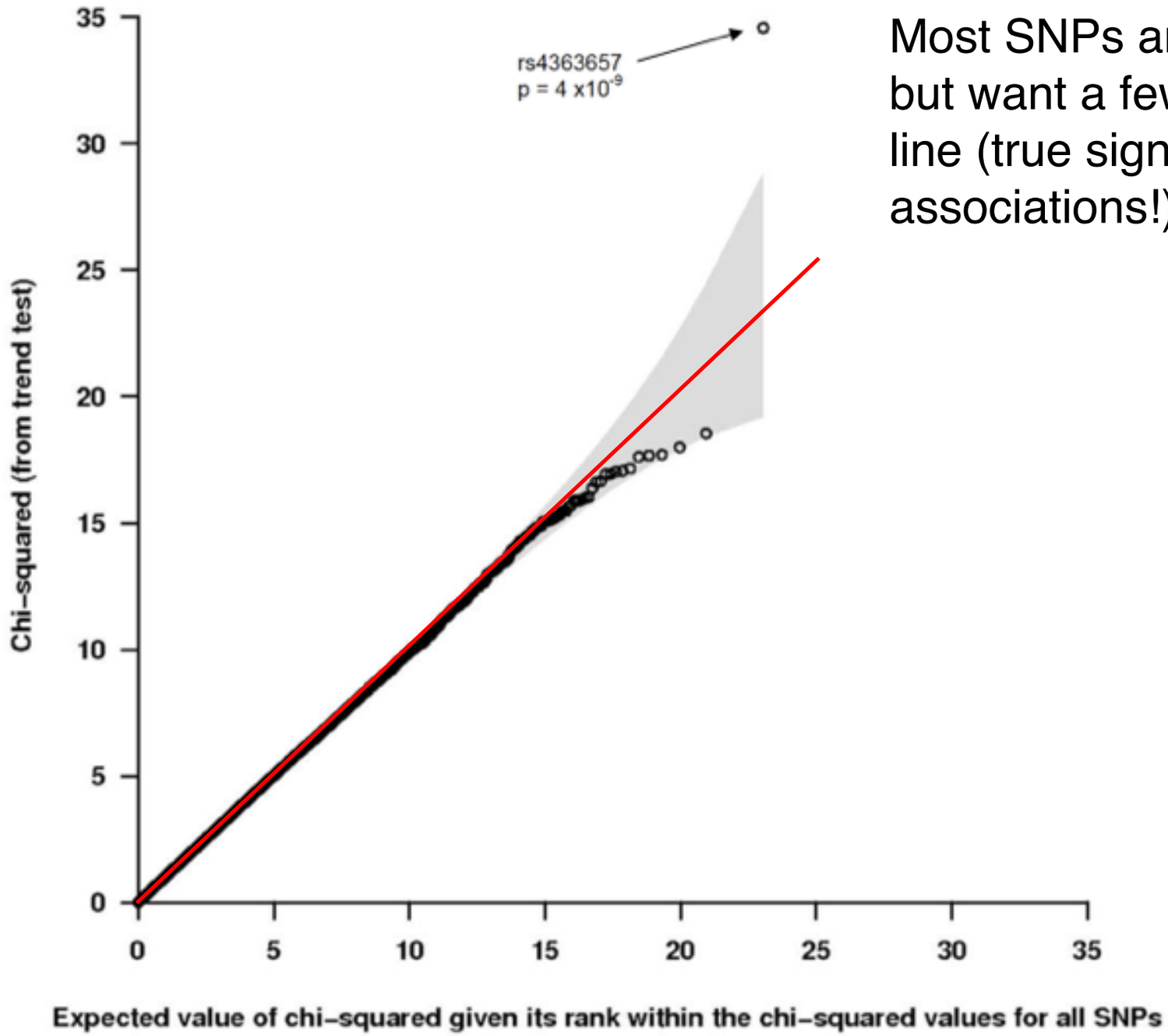
- The prosecutor tries to prove the guilt of the defendant

- Only when there is enough charging evidence the defendant is convicted

Hypothesis Tests

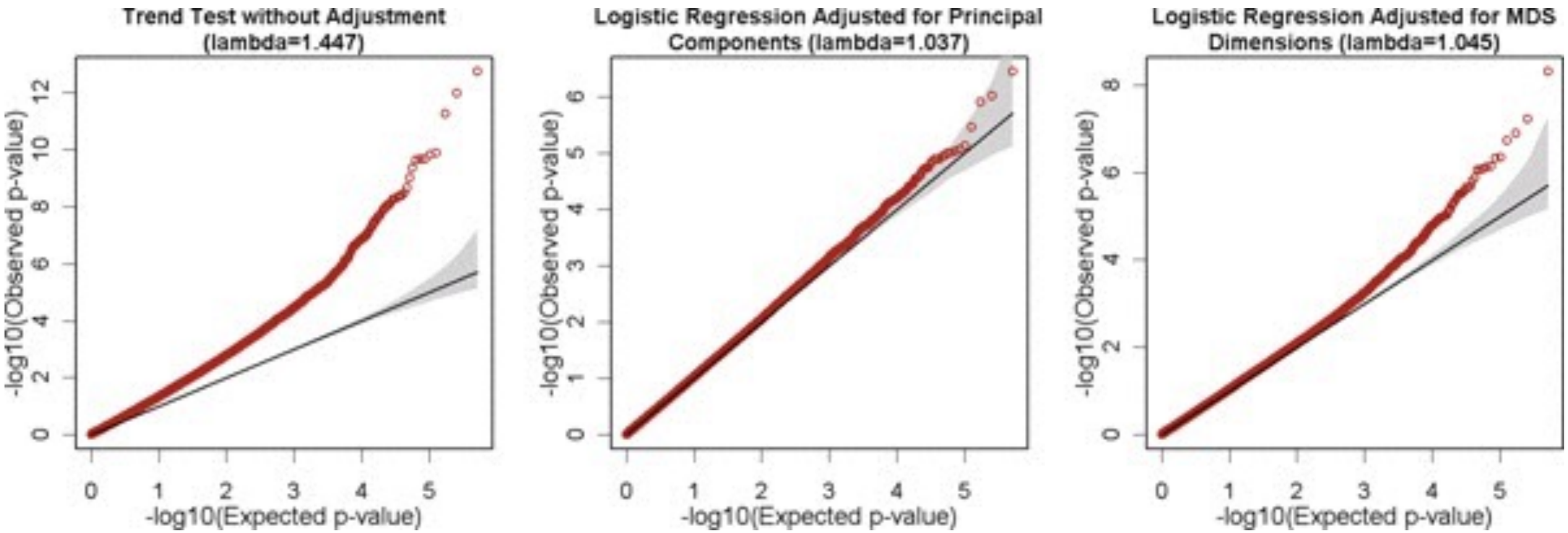
- For a single SNP, null hypothesis H_0: the SNP is not associated with the phenotype
- Alternative hypothesis H_1: the SNP is associated with the phenotype
- Test statistics are derived from the above tests, or regression models
- These tests give a *p*-value for each SNP

Quantile-Quantile Plot



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

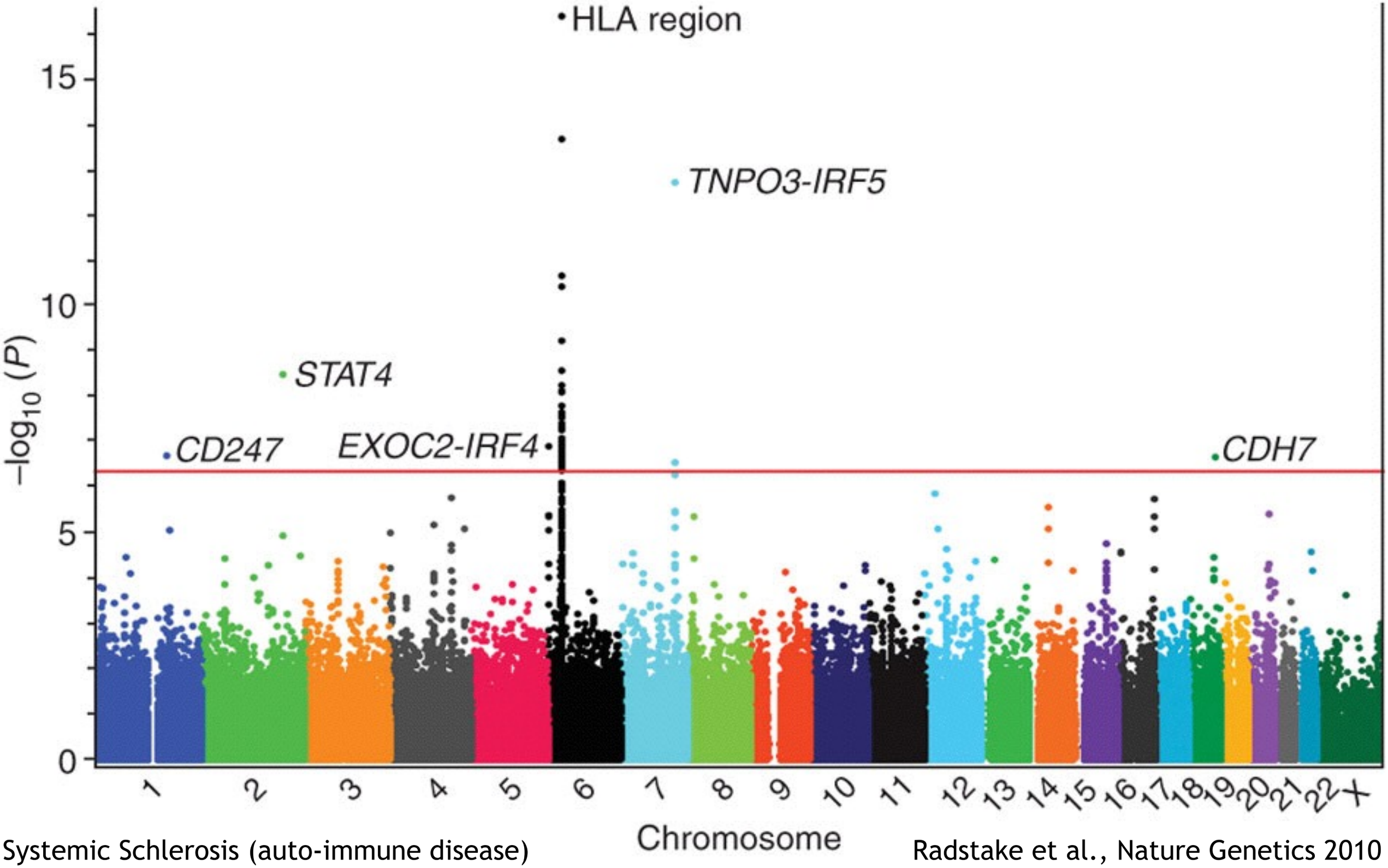
QQ-Plots and PC Adjustment



- 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



Systemic Schlerosis (auto-immune disease)

Radstake et al., Nature Genetics 2010

Multiple Testing

- Multiple testing occurs when one considers a set of hypothesis tests simultaneously

- By random chance alone, as the number of tests increases, it becomes more likely that at least one test will be significant

- If a test is performed at the 5% level, there is only a 5% chance of incorrectly rejecting the null hypothesis

- However, for 100 tests the expected number of incorrect rejections is 5

- Think about 1,000,000 SNPs and hypothesis tests!

Type I and II Errors

	H_0 is true	H_1 is true
Accept H_0	Right	Type II Error
Reject H_0	Type I Error	Right

- Type I Error: the probability that the null hypothesis H_0 is true, but is rejected
- Type II Error: the probability that the null hypothesis H_0 is false, but fails to be rejected
- Hypothesis test controls for Type I error at 0.05, and tries to minimize Type II error

Family-Wise Error Rate (FWER)

	Accept H_0	Reject H_0	Total
H_0 is true	U	V	m_0
H_1 is true	T	S	$m-m_0$
Total	m-R	R	m

- FWER = $P(V \geq 1)$, the probability that at least one false positive in the m tests
- We are controlling the number of Type I errors
- There are different ways to control FWER at 0.05

FWER Procedures

Single-Step Procedures:

- Bonferroni: reject H_{i0} if $P_i \leq 0.05/m$
- Sidak: reject H_{i0} if $P_i \leq 1-(1-0.05)^{1/m}$

Step-Wise Procedures:

- Order $P_1,\ldots,P_m$ from smallest to largest: $P_{(1)},\ldots,P_{(m)}$
- Step-down Holm: reject $H_{(i)0}$ if $P_{(h)} \leq 0.05/(m-h+1)$ for all $h \leq i$
- Step-up Hochberg: reject $H_{(i)0}$ if $P_{(h)} \leq 0.05/(m-h+1)$ for some $h \geq i$

False Discovery Rate (FDR)

	Accept H0	Reject H0	Total
H0 is true	U	V	m0
H1 is true	T	S	m-m0
Total	m-R	R	m

- FDR = E(V/R), is the expected value of the proportion of false positives among the rejected hypotheses
- We are controlling the proportion of Type I errors
- There are different ways to control FDR at 0.05

FDR Procedures

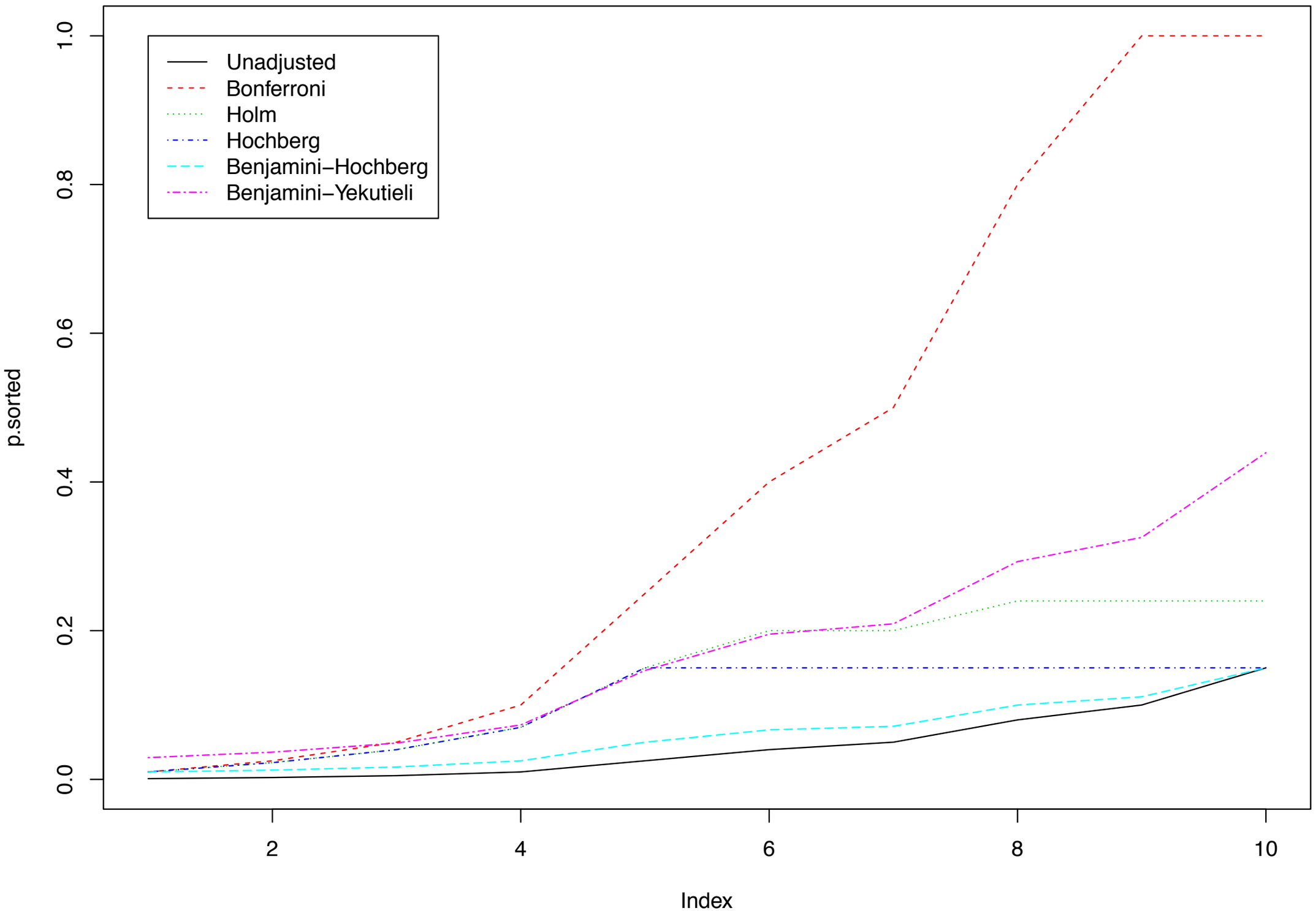
Benjamini-Hochberg procedure:

- A step-up procedure
- Reject $H_{(i)0}$ if there exists some $h \geq i$ such that $P_{(h)} \leq 0.05*h/m$

Other procedures:

- Benjamini-Yekutieli
- Many more

Adjusted *P*-values



Limitations of GWAS

- Not very predictive
- Explain little heritability
- Focus on common SNPs
- Many associated SNPs 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.

Visscher, AJHG 2011

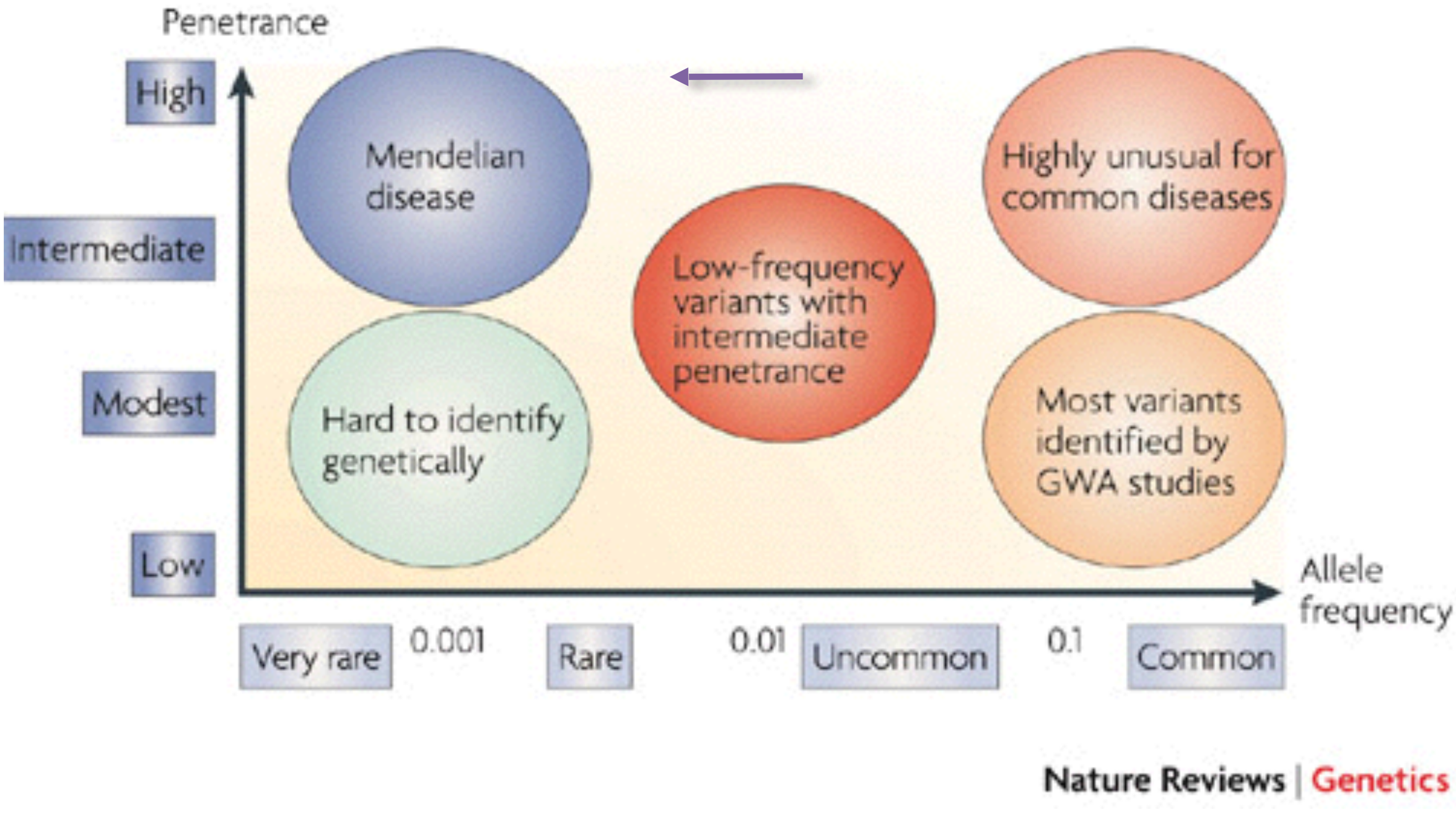
Will GWAS Explain More Heritability?

- Casual SNPs not yet detected due to power issue: weak effects

- More complicated modeling necessary, such as gene-gene interaction, gene-environment interaction, etc.

- But, many more researchers start to believe the Common disease rare variant (CDRV) hypothesis: diseases due to multiple rare SNPs with intermediate effects

Common Disease Rare Variants



See: NEJM, April 30, 2009

McCarthy et al., 2008

Rare Variants

- Most common approach: gene-based tests
- Single-SNP tests are lack of power
- Combine/Collapse multiple rare variants on a gene to a common variant
- Hypothesis tests are run on the gene level instead of SNP level
- Less multiple testing adjustment

Other Than SNPs

- Copy Number Variations
- Gene expression (RNA levels)
- Proteomics (protein levels)
- And so on…