BICF Nanocourse: Genome Analysis

Exome- and Genome Sequencing in Population-Based Studies of Complex Diseases/Traits

May 2, 2019

Julia Kozlitina, Ph.D.
McDermott Center for Human Growth and Development
UT Southwestern

Outline

- Introduction: population-based sequencing studies
- Sequencing-based study workflow:
  - Study design
  - Data preparation and QC
  - Single-variant association tests
    - Association tests for binary traits (case-control)
    - Association tests for quantitative traits
  - Gene- or region-based association tests
  - Summarizing and prioritizing results

Genetic Disorders

Mendelian (Monogenic) Diseases:

- Caused by mutation(s) in a single gene
- Follow Mendelian inheritance patterns in families (autosomal dominant, recessive, X-linked)
- Caused by rare mutations with *high penetrance* (i.e., having a mutation is sufficient to cause disease)
- Typically rare
- Examples: Sickle-cell disease, Cystic Fibrosis, Haemophilia, Huntington's disease

Complex (Multifactorial) Diseases/Traits:

- Influenced by variation in many genes often acting together with environmental factors
- Cluster in families, but do not follow Mendelian inheritance patterns
- Disease susceptibility influenced by many alleles, each having small effect
- More common
- Examples: Cancer, Type 2 Diabetes, Cardiovascular Disease

Exome sequencing as a tool for Mendelian disease gene discovery

Bamshad et al. Nature Reviews Genetics, volume 12, pages 745-755 (2011)

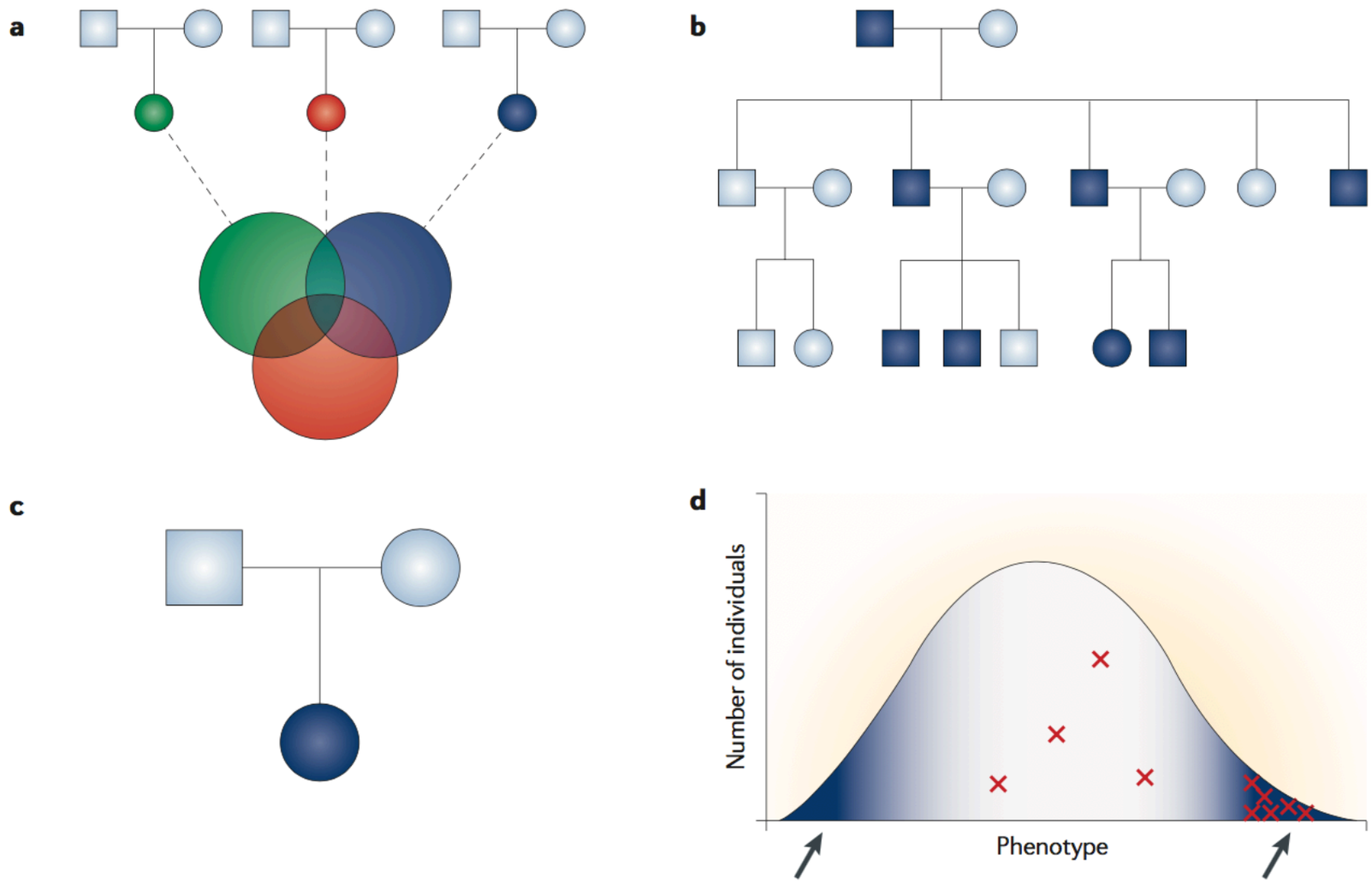


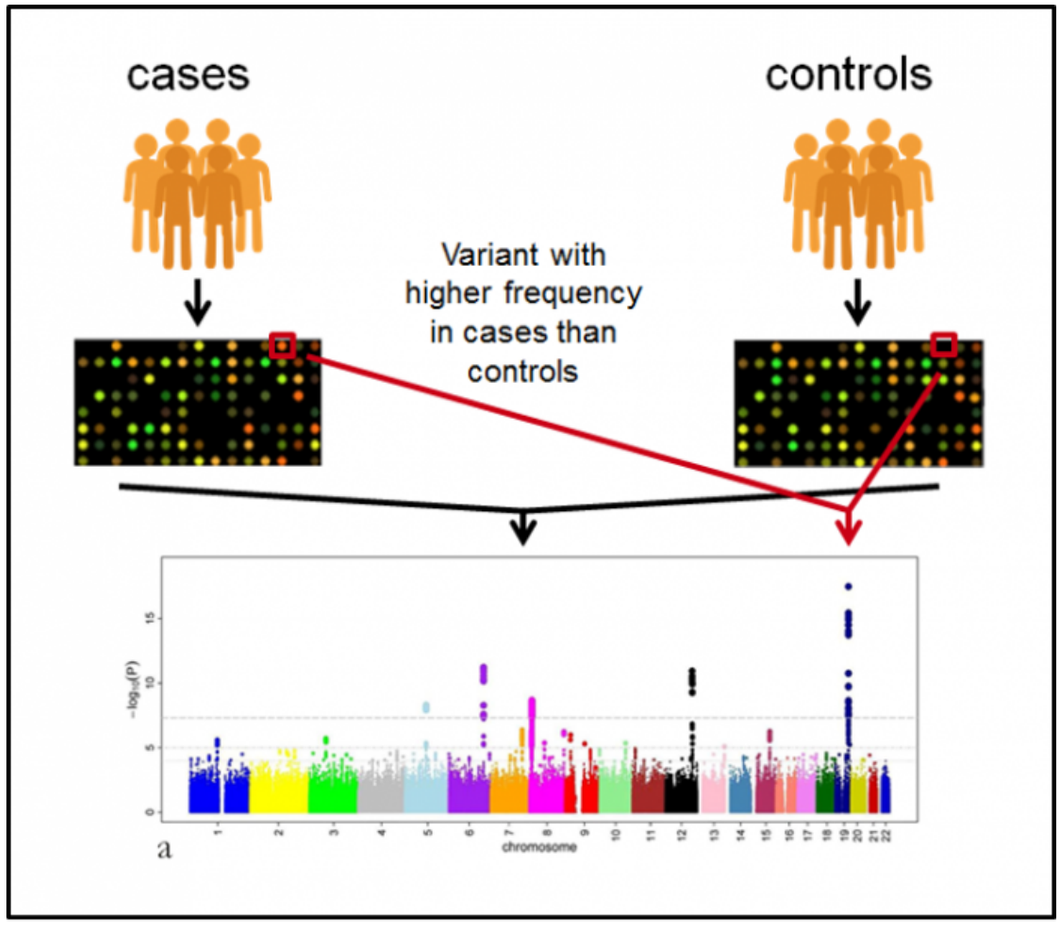
Figure 2 | Strategies for finding disease-causing rare variants using exome sequencing.

Exome-/genome- sequencing for identifying complex disease-variant associations

- A widely used tool to identify genetic loci influencing complex disease phenotypes is a genome-wide association study (GWAS) -

- GWAS systematically evaluate common variants (minor allele frequency [MAF] >5%) across the genomes of many *unrelated* individuals to find genetic variants *associated* with a particular disease.

- Traditionally, based on genotyping arrays including 500K-2.5M SNPs



- Decreasing DNA sequencing costs now make it feasible to perform association studies based on exome- or whole-genome sequencing (WGS), that can assess both common and rare variants

https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/

Examples of exome/genome sequencing studies for complex diseases

- NHLBI Exome Sequencing Project (ESP) - has performed whole exome sequencing of >6,500 individuals (with different diseases) to study genetic contributions to the risk of several heart, lung and blood phenotypes.

- T2D-GENES: Deep whole-exome sequencing in 10,000 people from five ethnicities to discover how variation in the protein-coding portion of the genome contributes to type 2 diabetes risk.

- UK10K Project - exome sequencing in 6,000 individuals ascertained for various diseases and whole-genome sequencing in 4,000 healthy individuals with detailed phenotyping

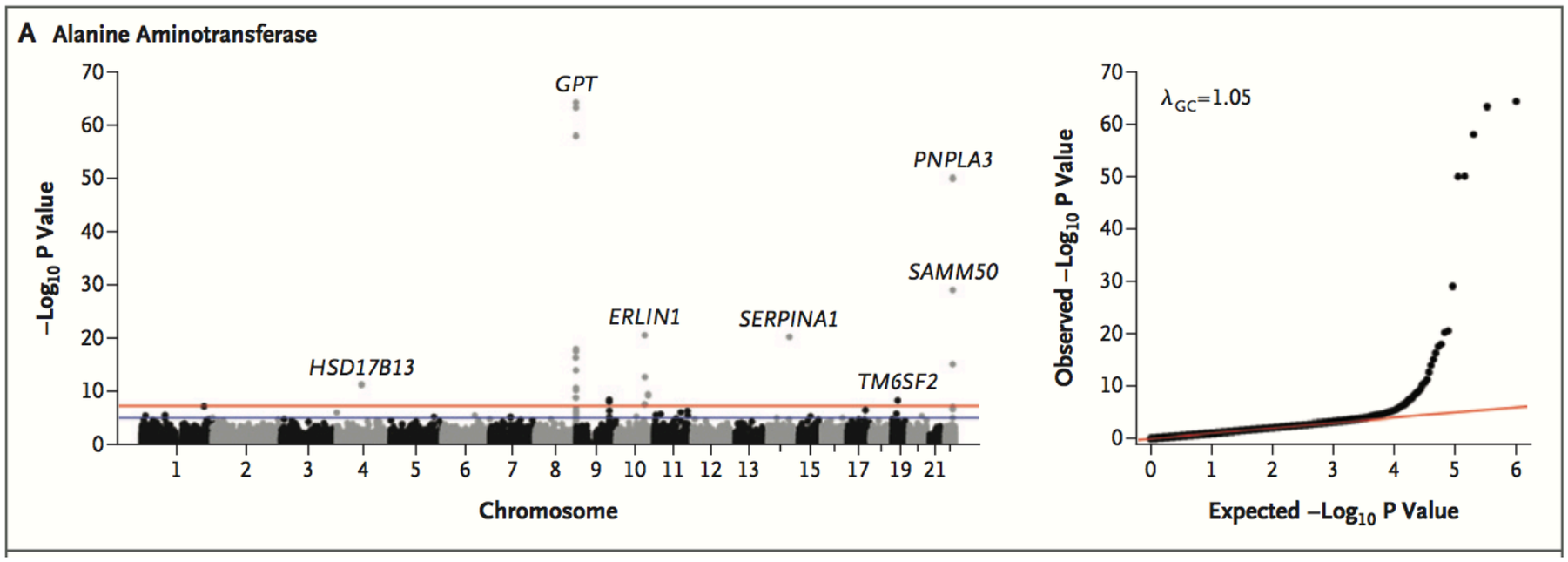
- UK Biobank - recently released exome sequence data on 50K individuals

- Pakistan Risk of Myocardial Infarction Study (PROMIS) - sequenced the protein-coding regions of 10,503 adult participants to understand the determinants of cardiometabolic diseases in individuals from South Asia

Examples of findings from exome/genome sequencing studies for complex diseases

A Protein-Truncating *HSD17B13* Variant and Protection from Chronic Liver Disease

Abul-Husn et al. N Engl J Med 2018;378:1096-106.



Used exome sequence data and electronic health records from 46,544 participants in the DiscovEHR human genetics study, along with several replication cohorts, to discover variants associated with chronic liver disease.

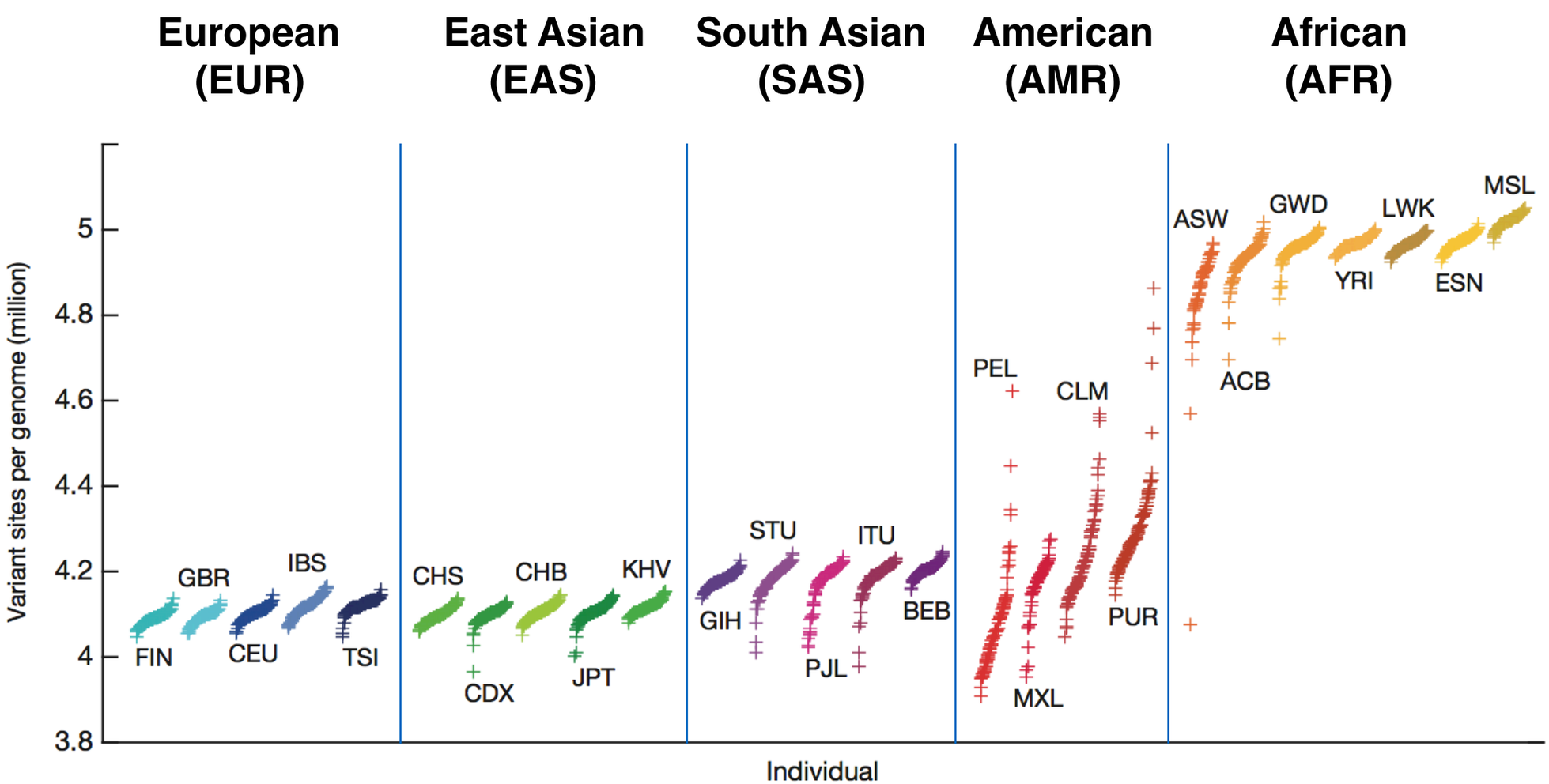
Human Genetic Variation

- A typical genome differs from the reference human genome at 4.1 million to 5.0 million sites

	AFR	AMR	EAS	EUR	SAS
Samples	661	347	504	503	489
Mean coverage	8.2	7.6	7.7	7.4	8.0
SNPs	4.31M	3.64M	3.55M	3.53M	3.6M
Indels	625k	557k	546k	546k	556k
Large deletions	1.1k	949	940	939	947
CNVs	170	153	158	157	165
Nonsynonymous	12.2K	10.4K	10.2K	10.2K	10.3k
Filtered LoF	182	152	153	149	151

The 1000 Genomes Project Consortium, 2015

Human Genetic Variation



The 1000 Genomes Project Consortium, 2015

Loss-of-Function Variants in Human Protein-Coding Genes

- Human genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated.

Variant type	1000G low-coverage average per individual		
	CEU	CHB+JPT	YRI
Stop	26.2 (5.2)	27.4 (6.9)	37.2 (6.3)
Splice	11.2 (1.9)	13.2 (2.5)	13.7 (1.9)
Frameshift indel	38.2 (9.2)	36.2 (9.0)	44.0 (8.0)
Large deletion	28.3 (6.2)	26.7 (5.9)	26.6 (5.5)
Total	103.9 (22.5)	103.5 (24.3)	121.5 (21.7)

Table 1. Numbers of LoF variants before and after filtering. Total numbers of candidate LoF variants and average number of LoF sites per individual (homozygous sites in parentheses) are shown for each LoF class. For large deletions, numbers represent total number of genes predicted to be inactivated.

MacArthur et al., Science, Vol. 335 (2012), p. 823

How many variants to expect from exome sequencing?

Table 1 | **Mean number of coding variants in two populations**

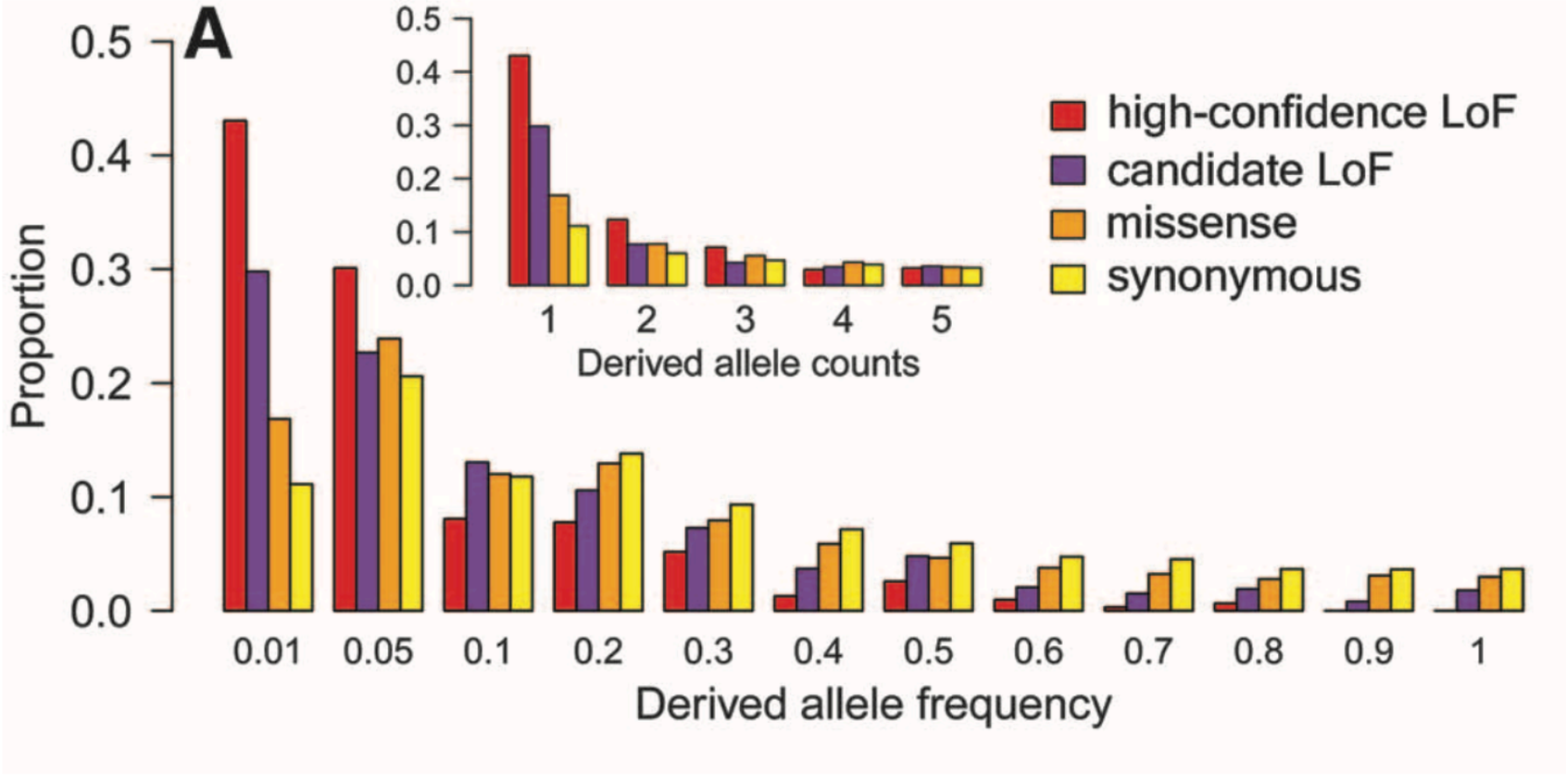
Variant type	Mean number of variants (± sd) in African Americans	Mean number of variants (± sd) in European Americans
Novel variants		
Missense	303 (± 32)	192 (± 21)
Nonsense	5 (± 2)	5 (± 2)
Synonymous	209 (± 26)	109 (± 16)
Splice	2 (± 1)	2 (± 1)
Total	520 (± 53)	307 (± 33)
Non-novel variants		
Missense	10,828 (± 342)	9,319 (± 233)
Nonsense	98 (± 8)	89 (± 6)
Synonymous	12,567 (± 416)	10,536 (± 280)
Splice	36 (± 4)	32 (± 3)
Total	23,529 (± 751)	19,976 (± 505)
Total variants		
Missense	11,131 (± 364)	9,511 (± 244)
Nonsense	103(± 8)	93 (± 6)
Synonymous	12,776 (± 434)	10,645 (± 286)
Splice	38 (± 5)	34 (± 4)
Total	24,049 (± 791)	20,283 (± 523)

Total number of variants identified (15k - 25k) depends on ethnicity as well as on exome capture method and targeted exome size

Bamshad MJ, Nat Rev Genet (2011)

Allele frequency distribution

- Most variants are rare, especially coding ones

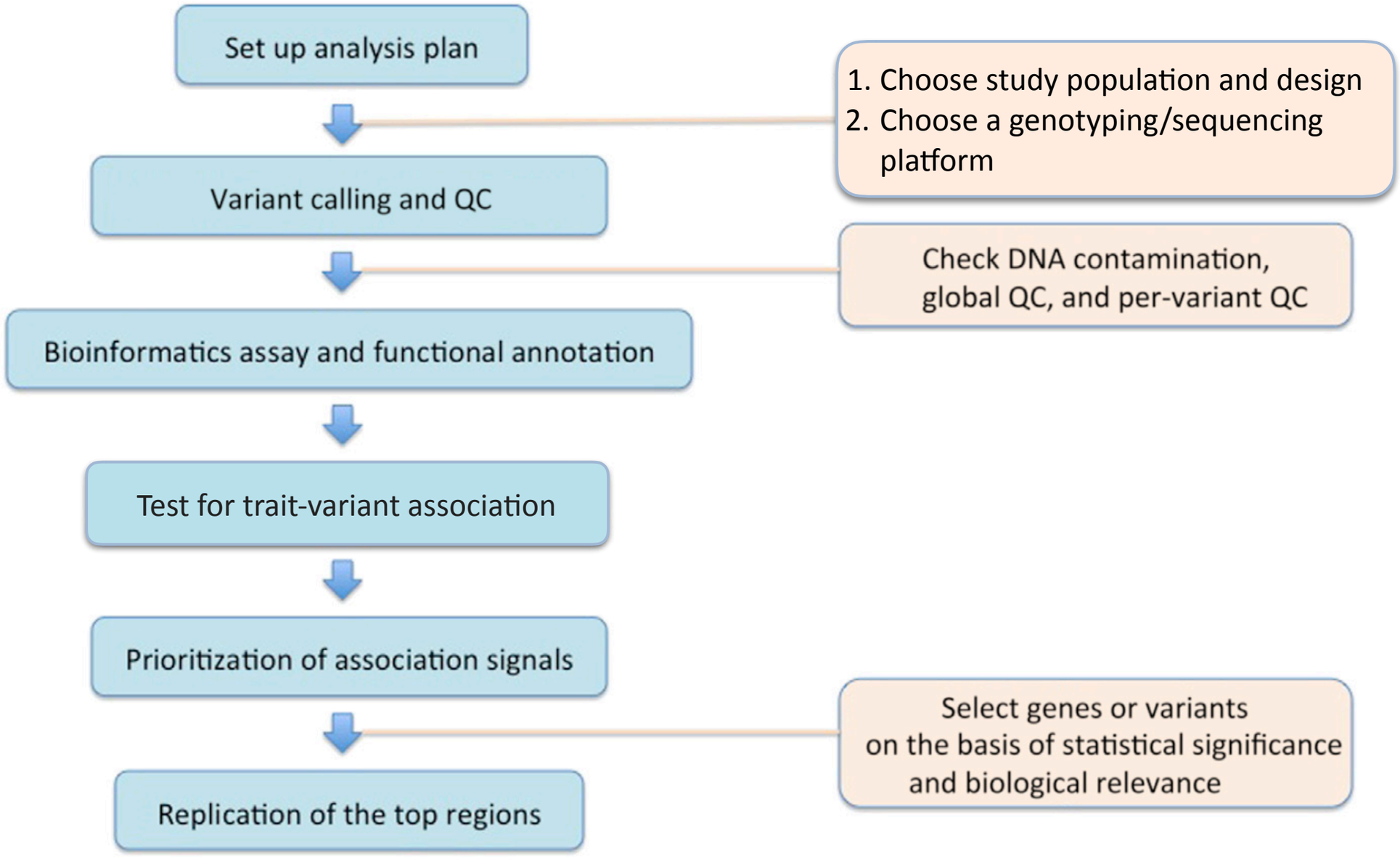


MacArthur et al., Science, Vol. 335 (2012), p. 823

Outline

- Introduction: population-based sequencing studies
- Sequencing-based study workflow:
  - Study design
  - Data preparation and QC
  - Single-variant association tests
    - Association tests for binary traits (case-control)
    - Association tests for quantitative traits
  - Gene- or region-based association tests
  - Summarizing and prioritizing results

Data-Processing and Analysis Flow Chart for Sequencing-Based Association Studies



Adapted from Lee et al., American Journal of Human Genetics 95, 5-23, July 3, 2014

Population-Based Association Studies

- Association studies for complex traits (both array-based and sequencing-based) rely on classical epidemiological designs:
    - **Case-control study**: recruit a group of unrelated individuals with a disease (*cases*) and without the disease (*controls*), and compare the frequency of alleles (or genotypes) at each genetic variant between cases and controls (using, e.g., chi-square tests or logistic regression)
    - **Cohort and cross-sectional studies**: recruit a sample of individuals from the general population, determine the genotype(s) and measure one or more phenotypes; use standard statistical methods to test for an association between genotype and phenotype (e.g., linear or logistic regression)
- **Advantages:** power, relatively easy to recruit participants (compared to family studies), can use existing cohorts (e.g., Framingham study)
- **Disadvantages**: prone to biases, confounding (e.g., population stratification)

Design of Association Studies

Main goals of study design:

- **Minimize bias -**

  If there is no association between a genetic variant and disease, statistical tests should not reject the null hypothesis (produce a significant result) more often than expected by chance

- **Maximize power -**

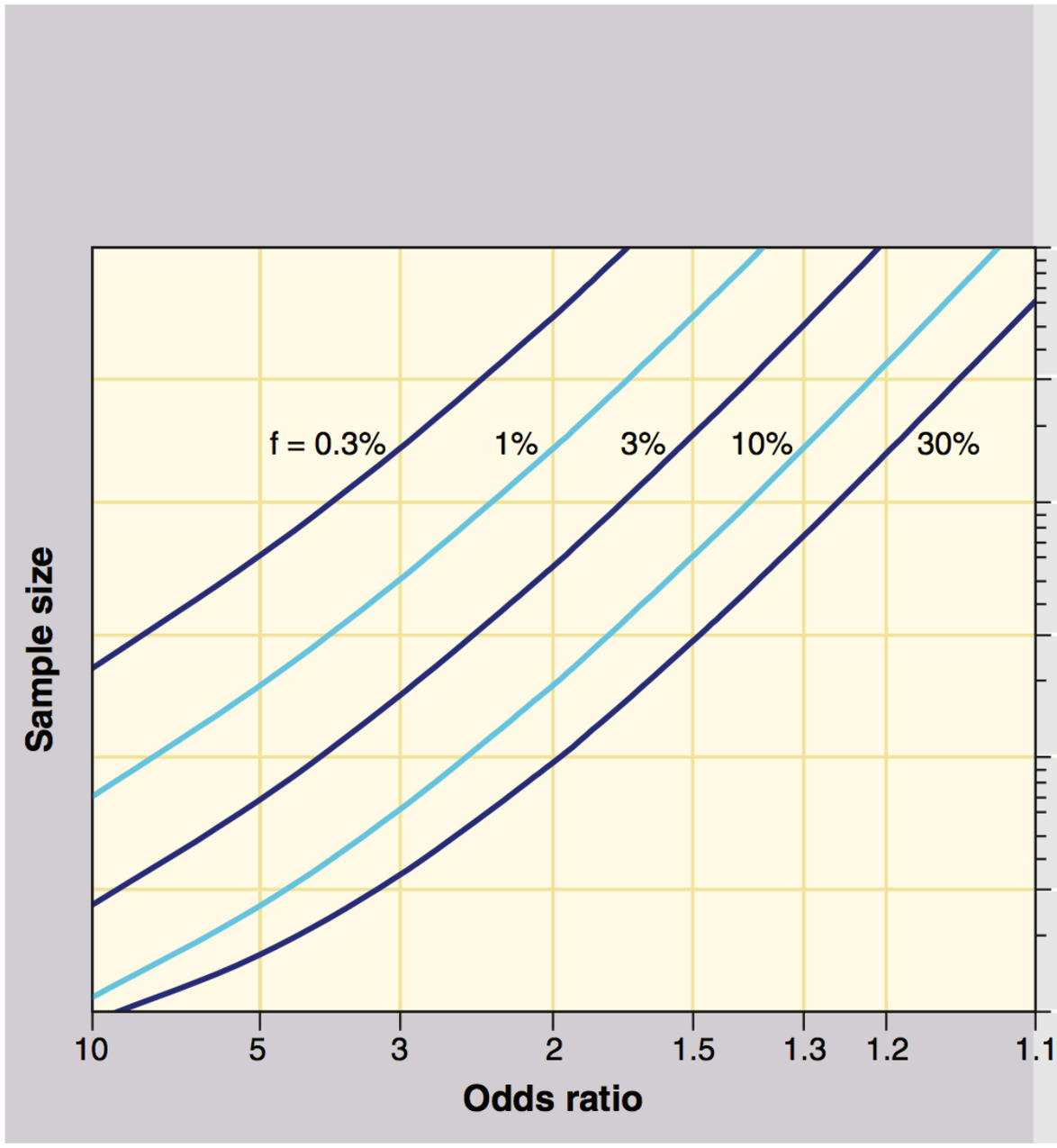
  If a variant is truly associated with the trait, statistical tests should have a good chance to detect the effect (i.e., reject the null hypothesis of no association).

Design Strategies for Sequencing Studies

- Large numbers of cases and controls needed to detect association with rare variants
  - For example, observing a single allele with a 0.5% or 0.05% frequency with 99% probability requires sequencing at least 460 or 4,600 individuals, respectively.
  - With an odds ratio (OR) = 1.4, the sample sizes required to achieve 80% power are 6,400, 54,000, and 540,000 for a MAF = 0.1, 0.01, and 0.001, respectively, if one assumes 5% disease prevalence and a significance level of 5×10^{-8}.
- Power can be improved by sampling individuals at the extremes of the phenotype distribution, since those are likely to be enriched for causal variants
- Using population controls (e.g., data from ExAC or UK Biobank) can improve power, but can introduce bias (batch effects) if the sequencing platforms and capture methods differ between cases and controls

Adapted from Lee et al., AJHG (2014), 95, pp. 5-23

Sample Sizes Required for Association Studies



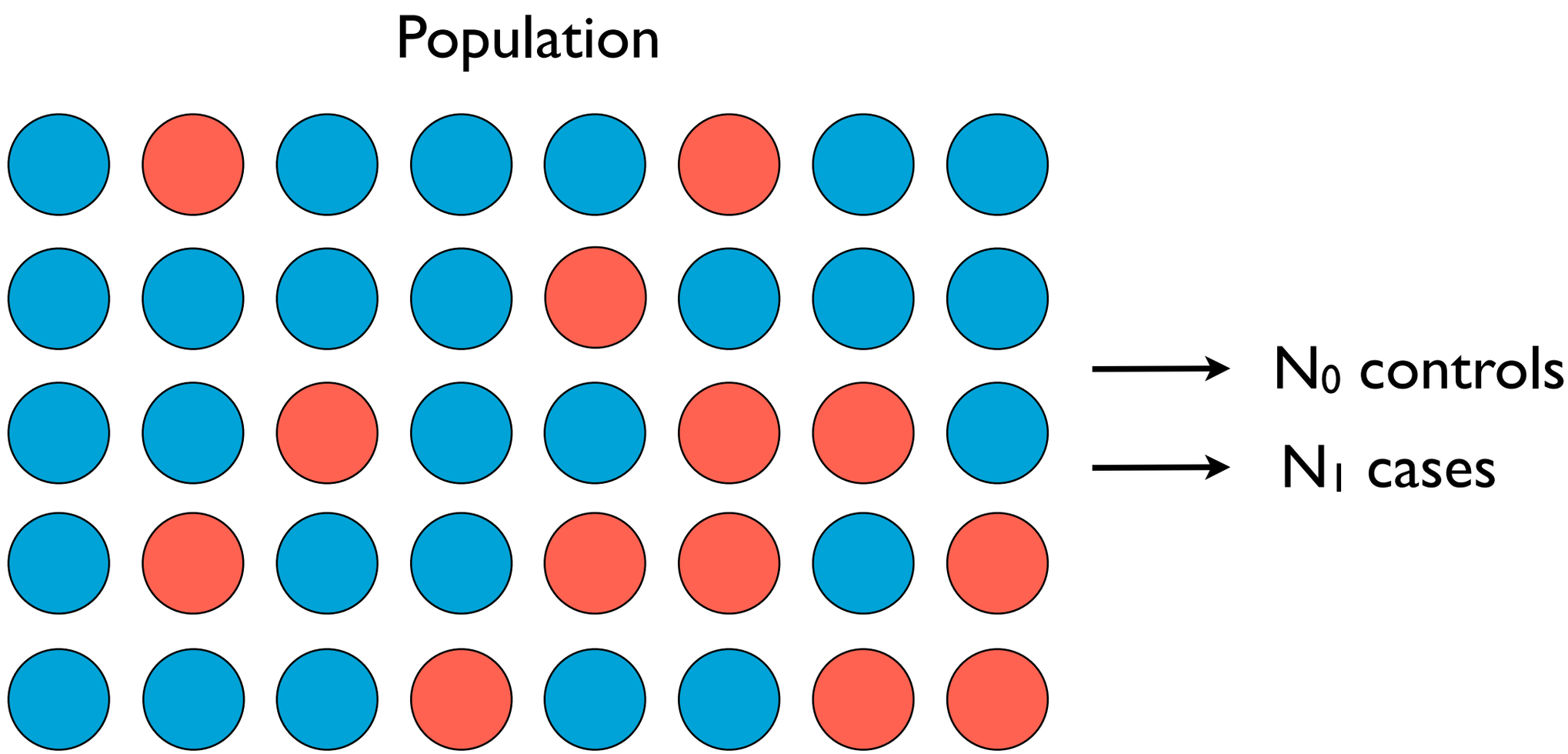
	GWAS			Rare variant	Nominal
	90%	50%	10%	90%	90%
100,000	66,790	40,260	66,040	21,370	
30,000	20,037	12,078	19,812	6411	
10,000	6679	4026	6604	2137	
3000	2004	1208	1981	641	
1000	668	403	660	214	
300	200	121	198	64	
100	67	40	66	21	

Altshuler et al. Science. 2008 November 7; 322(5903): 881-888.

Biases in Population-Based Association Studies

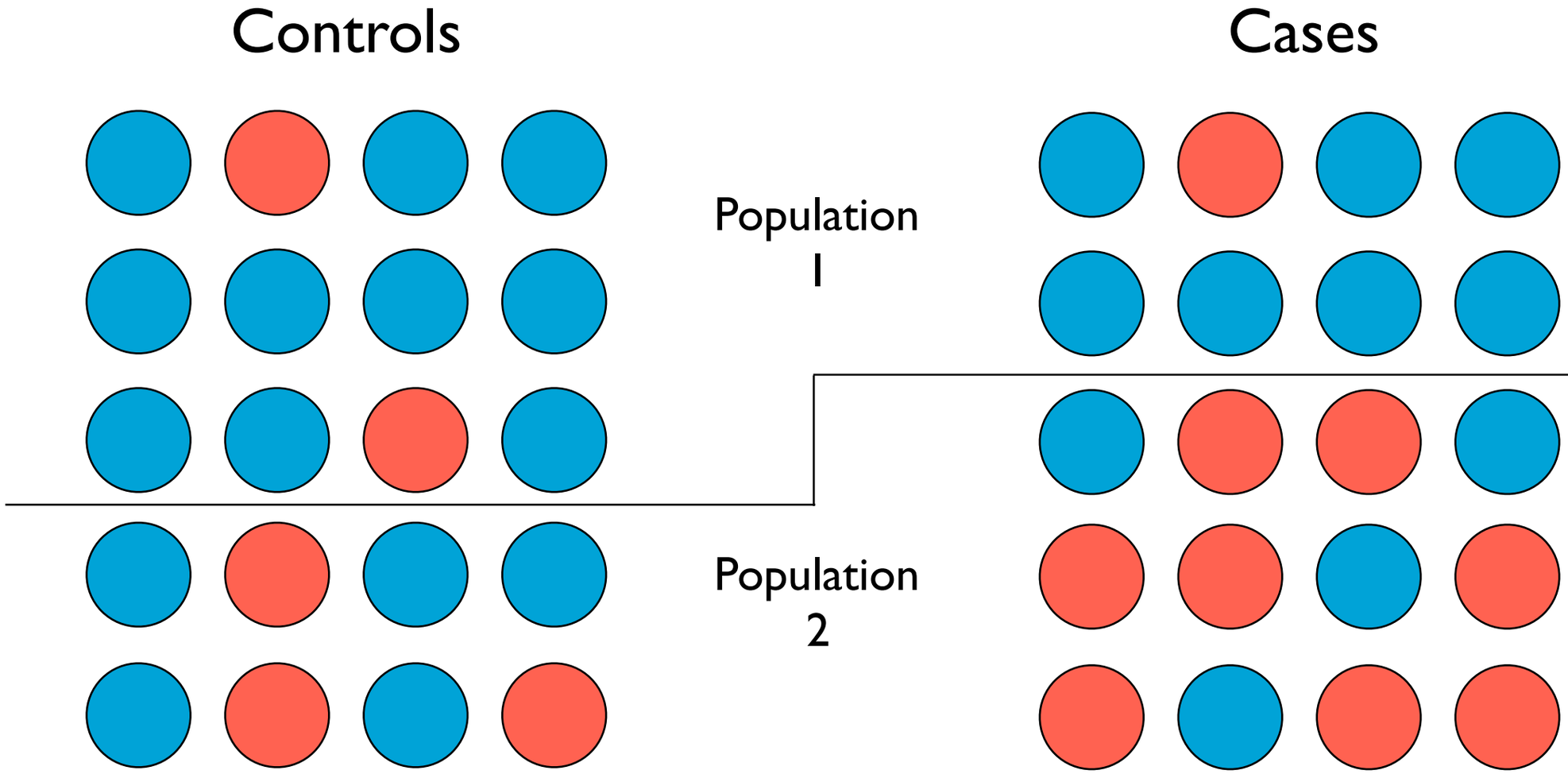
- **Confounding** - i.e., a disease outcome (D) seems to be associated with an exposure (genotype) merely due to its correlation with a risk factor for the disease (e.g. ethnicity)
    - **Population stratification** is a particular type of confounding, which occurs when cases and controls are selected from different underlying populations (different ethnic backgrounds)
- **Selection bias** - non-random selection mechanism; controls that are not representative of the general population
- **Miss-classification** - subjects incorrectly assigned to case (control) group
- **Differential genotyping error (batch effects)** - cases and controls are sequenced using different technologies, by different labs, or at different times
    - Be careful when you use publicly available control data (e.g. dbGAP)!

Ideal Case-Control Design



Balding DJ, Nat Reviews Genetics, 2006

Population Stratification



Balding DJ, Nat Reviews Genetics, 2006

Choosing a genotyping/sequencing platform

Array and Sequencing Platforms for Rare-Variant Analysis		
	Advantage	**Disadvantage**
High-depth WGS	can identify nearly all variants in genome with high confidence	more expensive; harder to interpret and prioritize variants outside the coding regions
Low-depth WGS	is a cost-effective, useful approach for association mapping	has limited accuracy for rare-variant identification and genotype calling; compared to deep sequencing, is subject to power loss if the same number of subjects is sequenced
Whole-exome sequencing	less expensive than WGS; focuses on protein-coding portions of the genome	is limited to the exome
Exome chip (custom array)	is much cheaper than exome sequencing	provides limited coverage for very rare variants and for non-Europeans; is limited to target regions

Adapted from Lee et al., AJHG (2014), 95, pp. 5-23

Outline

- Introduction: population-based sequencing studies
- Sequencing-based study workflow:
  - Study design
  - Data preparation and QC
  - Single-variant association tests
    - Association tests for binary traits (case-control)
    - Association tests for quantitative traits
  - Gene- or region-based association tests
  - Summarizing and prioritizing results

Data formats and files

- Project-level VCF file (pVCF)
    - contains the SNV/indel genotype calls across all samples
    - one variant per row
- Annotated VCF file
    - typically, separate from pVCF; contains annotations for all SNVs/indels called in the sample, but no genotypes
- Sample information file (or PED file)
    - Sample and family IDs, gender, case/control status
- Additional phenotype file
    - May contain additional phenotypes and/or covariates

pVCF File

Header

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20150218
##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##source=1000GenomesPhase3Pipeline
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth; only low coverage data were counted towards the DP, exome data were not used">
##bcftools_viewVersion=1.8+htslib-1.8
##bcftools_viewCommand=view -f PASS -M2 -v snps -R /Users/jkozli/Work/Merck/Exome_chip_ranges.txt ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.
20130502.genotypes.vcf.gz; Date=Tue Apr 30 11:26:50 2019
##bcftools_viewCommand=view -S PLINK/samples_v3.20130502.AFR.EUR.panel.txt --output-type z -o PLINK/Exome_chip/1kg.p3.Exome.chr1.vcf.gz; Date=Tue Apr 30 11:26:50
##bcftools_concatVersion=1.8+htslib-1.8
##bcftools_concatCommand=concat --output 1kg_Exome.vcf.gz --output-type z --file-list file_list.txt; Date=Tue Apr 30 13:55:15 2019
#CHROM  POS     ID           REF  ALT  QUAL  FILTER  INFO                                             FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101
1       762320  rs75333668   C    T    100   PASS    AC=137;AF=0.0694888;AN=2328;NS=2504;DP=15936     GT      0|0     0|0     0|0     0|0     0|0
1       865545  rs201186828  G    A    100   PASS    AC=1;AF=0.00279553;AN=2328;NS=2504;DP=14233      GT      0|0     0|0     0|0     0|0     0|0
1       865584  rs148711625  G    A    100   PASS    AC=52;AF=0.0109824;AN=2328;NS=2504;DP=15392      GT      0|0     0|0     0|0     0|0     0|0
1       865625  rs146327803  G    A    100   PASS    AC=7;AF=0.00139776;AN=2328;NS=2504;DP=16996      GT      0|0     0|0     0|0     0|0     0|0
1       865628  rs41285790   G    A    100   PASS    AC=5;AF=0.00279553;AN=2328;NS=2504;DP=16975      GT      0|0     0|0     0|0     0|0     0|0
1       865662  rs140751899  G    A    100   PASS    AC=3;AF=0.000798722;AN=2328;NS=2504;DP=17021     GT      0|0     0|0     0|0     0|0     0|0
1       865694  rs9988179    C    T    100   PASS    AC=32;AF=0.052516;AN=2328;NS=2504;DP=17146       GT      0|0     0|0     0|0     0|0     0|0
```

Variant Information

Genotypes

Data QC: Per-individual QC

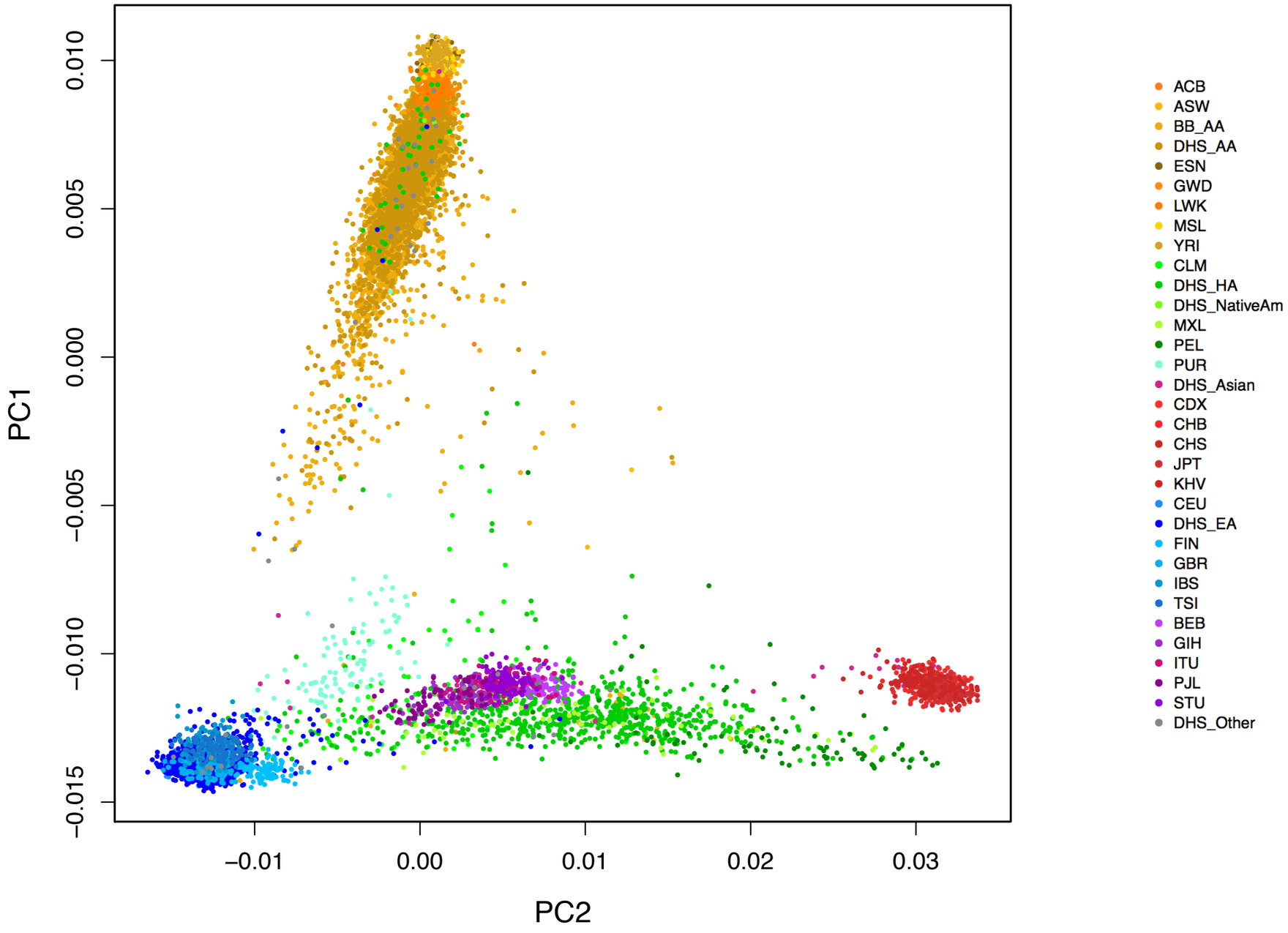
Identify and exclude individuals if:

- genotyped gender does not match stated gender (may indicate DNA sample swap)
  - use X-chromosome genotypes (males should have ~100% homozygosity rate, females <20%)
- high missing genotype rate, e.g., >3% (suggests low-quality DNA)
- heterozygosity rate >3 SD units (suggests DNA contamination)
- control samples; duplicates + discordant duplicate pairs
- related to other subjects in the study (cryptic relatedness)
  - can be detected by calculating the proportion of shared alleles at genotyped SNPs (identity by state, IBS)
- Estimate principal components of ancestry; exclude individuals of divergent ancestry

Either exclude or correct for this

Anderson et al. (2010) Nature Protocols, Vol. 5, No. 9

Example: Estimating Ancestry



Can project your samples on PCs calculated based on 1000 genomes project data to determine the origin

Data QC: Variant (SNP) QC

We assume that that basic QC has already been done as part of variant calling, and VCF file includes only high-confidence variants.

Exclude SNPs if:

- Any flags in the FILTER column (VSQR)
- High missing genotype call rate across samples (e.g., >10%)
- Significantly different missing genotype rates between cases and controls
- Significant deviation from Hardy-Weinberg equilibrium (e.g., $p < 10^{-5}$)
- For exome-sequencing: exclude variants in off-target regions (e.g., >20 bp outside the exon boundaries).
- Long indels
- Multiallelic variants

Notation

- Consider a genetic marker (e.g., a SNP) with two alleles, ***A*** and ***a***
- Suppose ***a*** is the more common (major/wild-type/reference) allele, and ***A*** is the less common (minor/alternate) allele.
- Allele frequencies: $p_A = p$, $p_a = 1 - p$
- Three possible genotypes: *aa, aA, AA*
- Often written as g_0, g_1, g_2, where $i = 0,1,2$ - number of copies of the minor allele
- For a given SNP, the alleles and genotypes are usually labeled by the two alternative nucleotide bases, e.g.: *AA, AG, GG,* or *TT, TC, CC*

Estimating Allele Frequencies

- In a sample of *N* individuals, let

  n_{aa} = number of people with *aa* genotype

  n_{aA} = number of people with *aA* genotype

  n_{AA} = number of people with *AA* genotype

  where $n_{aa} + n_{aA} + n_{AA} = N$.

- Then minor allele frequency (MAF), *p*, is estimated as:

$$\bar{p} = \frac{(2n_{AA} + n_{aA})}{2N}$$

N individuals =
2N chromosomes

Hardy-Weinberg Equilibrium (HWE)

- Under the assumptions of random mating in large populations, in the absence of selection, migration, inbreeding, etc., genotype frequencies are determined by allele frequencies

- Given a marker with alleles *A* and *a* with frequencies p and $q = 1 - p$, the genotype frequencies are given by:

$$p_{aa} = (1-p)^2,\ p_{aA} = 2p(1-p),\ p_{AA} = p^2$$

- Deviation from HWE can arise due to population stratification, due to association between allele and disease in cases, but also because of genotyping error. So, typically used as a genotype quality check.

Hardy-Weinberg Equilibrium (HWE)

- To test whether HWE holds in a sample, we compare observed genotype counts (frequencies) to their expected values under HWE:

Genotypes	Observed	Expected
AA	n_{AA}	Np^2
aA	n_{aA}	$2Np(1-p)$
aa	n_{aa}	Np^2

where *p* is estimated by $\bar{p} = \frac{(2n_{AA} + n_{aA})}{2N}$

- Deviation can be tested by Chi-square test or Exact test (Wigginton et al., AJHG, 2005)

Outline

- Introduction: population-based sequencing studies
- Sequencing-based study workflow:
  - Study design
  - Data preparation and QC
  - Single-variant association tests
    - Association tests for binary traits (case-control)
    - Association tests for quantitative traits
  - Gene- or region-based association tests
  - Summarizing and prioritizing results

Genetic Models

- ***Genetic models*** **-** describe a relationship between genotype and disease risk or quantitative trait distribution

- For binary traits, described in terms of ***disease penetrance*** **-** risk of disease of disease in individuals carrying a particular genotype:

$$f_0 = \mathrm{P}(\mathrm{D} \mid aa),\ \ f_1 = \mathrm{P}(\mathrm{D} \mid aA),\ \ f_2 = \mathrm{P}(\mathrm{D} \mid AA)$$

  where D stands for disease and "|" denotes conditional probability given the genotype

- ***Relative risk (RR)*** **-** risk of disease in individuals with one genotype relative to another genotype, i.e., f_i / f_0, $i = 1, 2$, **-** is a natural measure of association (or allelic effect size)

Genetic Models (2)

- ***Genetic model (or mode of inheritance)*** - describes how penetrance depend on the number of alleles

Genetic	Penetrance			Relative Risk	
Model	*aa*	*aA*	*AA*	*aA*	*AA*
Dominant	f_0	γf_0	γf_0	γ	γ
Recessive	f_0	f_0	γf_0	1	γ
Additive	f_0	$f_0(1+\gamma)/2$	γf_0	$(1+\gamma)/2$	γ
Multiplicative	f_0	γf_0	$\gamma^2 f_0$	γ	γ^2
Co-dominant (genotypic)	f_0	$\gamma_1 f_0$	$\gamma_2 f_0$	γ_1	γ_2

Genetic Models (3)

- For quantitative traits, ***genetic model (of mode of inheritance)*** describes how the distribution (or mean value) of the trait depends on the number of alleles

Genetic Model	Mean trait value		
	aa	*aA*	*AA*
Dominant	μ_0	μ_1	μ_1
Recessive	μ_0	μ_0	μ_1
Additive	μ_0	$(\mu_0+\mu_2)/2$	μ_2
Co-dominant (genotypic)	μ_0	μ_1	μ_2

Case-Control Data for a Single SNP

	aa	*aA*	*AA*	*Total*
Cases	n_{10}	n_{11}	n_{12}	$n_{1.}$
Controls	n_{20}	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	N

"." denotes total across rows or columns, e.g.,

$n_{1.}$ = total n for row 1

$n_{.1}$ = total n for column 1

- Data from a case-control study can be summarized as a 2 x k contingency table of disease status by either genotype (k=3) or allele (k=2) count

- Null hypothesis of no association: row and column frequencies are independent, i.e.,

  H_0: Pr(Case | *aa*) = Pr(Case | *Aa*) = Pr(Case | *AA*), **or**

  H_0: genotype frequencies are equal between cases and controls

Genotypic Association Test

	aa	*aA*	*AA*	*Total*
Cases	n_{10}	n_{11}	n_{12}	$n_{1.}$
Controls	n_{20}	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	N

H_0: Pr(Case) equal among genotypes

H_A: at least one inequality holds

- Basic test of association between ***genotype*** and disease is given by a χ^2 chi-square test for independence of rows and columns in a 2 x 3 table:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=0}^{2} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]} \qquad \text{where} \qquad E[n_{ij}] = \frac{n_{i.}n_{.j}}{N}$$

- Under H_0, the calculated statistic X^2 has a χ^2 distribution with 2 degrees of freedom (df*)

*df = (Rows-1)×(Columns-1) or number of parameters needed to describe the model

Model-Based Association Tests

Dominant model

	aa	*aA* + *AA*	*Total*
Cases	n_{10}	$n_{11} + n_{12}$	$n_{1.}$
Controls	n_{20}	$n_{21} + n_{22}$	$n_{2.}$
Total	$n_{.0}$	$n_{.1} + n_{.2}$	*N*

Recessive model

	aa + *aA*	*AA*	*Total*
Cases	$n_{10} + n_{11}$	n_{12}	$n_{1.}$
Controls	$n_{20} + n_{21}$	n_{22}	$n_{2.}$
Total	$n_{.0} + n_{.1}$	$n_{.2}$	*N*

- To test for a dominant model (effect) the data can be summarized as a 2 x 2 table of *aa* genotype counts versus *aA* and *AA* combined

  $H_{A,\,DOM}$: Pr(Case | *aa*) ≠ Pr(Case | *aA* or *AA*)

- To test for a recessive model (effect) the data can be summarized as a 2 x 2 table of *AA* genotype counts versus *aa* and *aA* combined

  $H_{A,\,REC}$: Pr(Case | *aa* or *aA*) ≠ Pr(Case | *AA*)

- Perform a chi-square test (1 df) for a corresponding 2 x 2 table

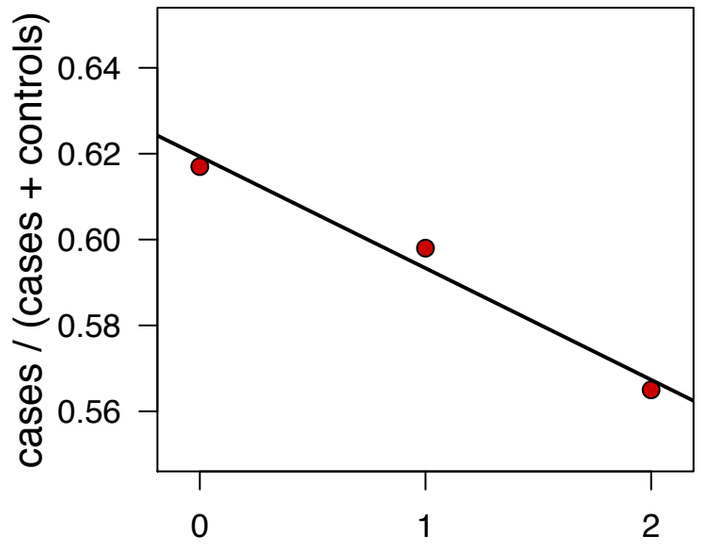
Exact P-values

- Chi-square tests assume large sample sizes (say >5 for any n_{ij}), and may be inaccurate otherwise.
- When cell counts are small, use Fisher's exact test.
- Exact test:
  1. For fixed row and column totals, list all possible configurations of genotype counts
  2. For each, calculate the appropriate X^2 statistic.
  3. How many configurations will give you a X^2 statistic (i.e., differences in proportions) greater than the ones actually observed?

$$\text{Exact p-value} = \frac{\text{No. of configurations with more extreme differences than observed}}{\text{Total No. of configurations}}$$

Cochran-Armitage Trend Test

	aa	*aA*	*AA*	*Total*
Cases	n_{10}	n_{11}	n_{12}	$n_{1.}$
Controls	n_{20}	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	N



- To test for an additive model, use the Cochran-Armitage trend test. The test "fits" a line to estimated proportions of cases. This can be easily performed:
  - code genotype as 0, 1, 2, for *aa*, *aA*, and *AA*, and outcome as '1' for cases and '0' for controls
  - calculate Pearson's correlation coefficient, *r*, between genotype and outcome
  - Under H_0, test statistic $T^2 = r^2 \times N$ has χ^2 distribution on 1 df.
  - Compare the observed test statistic to a χ^2 distribution on 1 df to determine the p-value.

Allelic Association Test

	a	*A*	*Total*
Cases	$2n_{10} + n_{11}$	$n_{11} + 2n_{12}$	$2n_{1.}$
Controls	$2n_{20} + n_{21}$	$n_{21} + 2n_{22}$	$2n_{2.}$
Total	$2n_{.0} + n_{.1}$	$n_{.1} + 2n_{.2}$	$2N$

H_0: $p_{A,case} = p_{A,control}$

H_A: $p_{A,case} \neq p_{A,control}$

- An alternative way to test for an additive model is to summarize the data as a 2 x 2 table of allele counts in cases vs controls and perform a chi-square test on 1 df
- This test assumes HWE, and may not be suitable otherwise
- Hard to interpret: produces a measure of risk associated with an allele (chromosome), not genotype (individual)

Measures of Association

- Would like to know the *relative risk* ***(RR)***:

$$RR = \frac{\Pr(\text{Disease} \mid \text{genotype } aA \text{ or } AA)}{\Pr(\text{Disease} \mid \text{genotype } aa)}$$

- Cannot directly estimate RRs from case-control studies (because the ratio of cases/controls is fixed). In case-control studies, the strength of an association is measured by the ***odds ratio (OR)***:

Statistical Odds

- **Odds** of an event - the probability of an event occurring compared with the probability of it not occurring
  - Let π be the probability of having disease
  - Odds of disease = $\dfrac{\pi}{1-\pi}$

- **Odds ratio** (OR) = ratio of odds of disease in one group (exposed) versus the odds in another group (unexposed)

  OR = 1 no difference in odds

  OR > 1 increased odds

  OR < 1 decreased odds

- When π (the probability of disease) is small, OR $\approx$ RR

Measures of Association

- Would like to know the *relative risk* ***(RR)***:

$$RR = \frac{\Pr(\text{Disease} \mid \text{genotype } aA \text{ or } AA)}{\Pr(\text{Disease} \mid \text{genotype } aa)}$$

- Cannot directly estimate RRs from case-control studies (because the ratio of cases/controls is fixed). In case-control studies, the strength of an association is measured by the ***odds ratio (OR)***:

$$OR = \frac{\Pr(\text{Disease} \mid \text{genotype } aA \text{ or } AA) \,/\, \Pr(\text{No disease} \mid \text{genotype } aA \text{ or } AA)}{\Pr(\text{Disease} \mid \text{genotype } aa) \,/\, \Pr(\text{No disease} \mid \text{genotype } aa)}$$

- When the probability of disease is small, OR approximates RR

Estimating Effect Size

	aa	*aA*	*AA*	*Total*
Cases	n_{10}	n_{11}	n_{12}	$n_{1.}$
Controls	n_{20}	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	N

- Genotypic odds ratios

aA relative to *aa*: $\mathrm{OR}_{aA} = \dfrac{n_{11}n_{20}}{n_{10}n_{21}}$

AA relative to *aa*: $\mathrm{OR}_{AA} = \dfrac{n_{12}n_{20}}{n_{10}n_{22}}$

- Model-based:

$$\mathrm{OR}_{DOM} = \frac{(n_{12}+n_{11})n_{20}}{n_{10}(n_{21}+n_{22})}$$

$$\mathrm{OR}_{REC} = \frac{n_{12}(n_{20}+n_{21})}{(n_{10}+n_{11})n_{22}}$$

- Allelic odds ratio (*A* vs *a*):

$$\mathrm{OR}_{A} = \frac{(2n_{12}+n_{11})(2n_{20}+n_{21})}{(2n_{10}+n_{11})(2n_{22}+n_{21})}$$

Example: association between a Ser-9-Gly polymorphism in the dopamine D3 receptor gene and schizophrenia

Shaikh et al. Hum Genet (1996) 97: 714-719.

	Genotype, N (%)				Allele, N (%)		
	1-1	1-2	2-2	Total	1	2	Total
Cases	57 (0.54)	69 (0.52)	7 (0.05)	133	183 (0.69)	83 (0.31)	266
Controls	33 (0.30)	56 (0.52)	20 (0.18)	109	122 (0.56)	96 (0.44)	218

- Allelic test (allele 2 vs 1): $OR = (83 \times 122)/(96 \times 183) = 0.58$,

  $\chi^2 = 8.46$, $df = 1$, p-value = 0.004

- Genotypic (co-dominant) test: $OR_{1\text{-}2 \text{ vs } 1\text{-}1} = 0.71$, $OR_{2\text{-}2 \text{ vs } 1\text{-}1} = 0.20$

  $\chi^2 = 11.75$, $df = 2$, p-value = 0.0028 (exact p-value = 0.0029)

- Trend test: $\chi^2 = 9.49$, $df = 1$, p-value = 0.0021

- Dominant model (1-2 + 2-2 vs 1-1): $OR = 0.58$, $\chi^2 = 4.06$, $df = 1$, p-value = 0.044

- Recessive model (2-2 vs 1-1 + 1-2): $OR = 0.25$, $\chi^2 = 10.35$, $df = 1$, p-value = 0.0013

Logistic Regression

- Simple chi-square tests cannot adjust for covariates.
- If need to include covariates, fit a logistic regression model to disease outcome, *Y* (*Y* = 1 for cases, *Y* = 0 for controls):

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 G + \beta_2 X$$

where

π is the probability of being affected, Pr(*Y* = 1)

log[π/(1-π)] - log odds of disease (logit)

G - genotype coded according to assumed model

X - other covariate (e.g., ancestry, age, gender, etc.)

- Null hypothesis of no association between genotype and disease:

$$\mathrm{H}_0 : \beta_1 = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \beta_1 \neq 0$$

Logistic Regression (2)

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 G + \beta_2 X$$

- We can test any of the genetic models by converting the genotype to a suitable numerical variable:

Model	*aa*	*aA*	*AA*
Dominant	0	1	1
Recessive	0	0	1
Additive/multiplicative	0	1	2
Co-dominant*	0	1	0
(genotypic)	0	0	1

*For a co-dominant model, genotype is coded as two dummy variables, say G_1 and G_2, indicating two of the three genotypes

Interpretation of logistic regression coefficients

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 G + \beta_2 X$$

- Estimated β_1 measures a change in log odds of disease per one unit change in the predictor (genotype), i.e., log odds ratio
- Odds ratio (OR) can by estimated by exponentiating beta, e^{β}
- E.g., if estimated $\beta_1 = 0.4$, then the OR = exp(0.4) = 1.5, that is, the odds of disease are increased by a factor of 1.5 per one unit change in genotype
- Under additive* coding, this is OR per each additional minor allele:

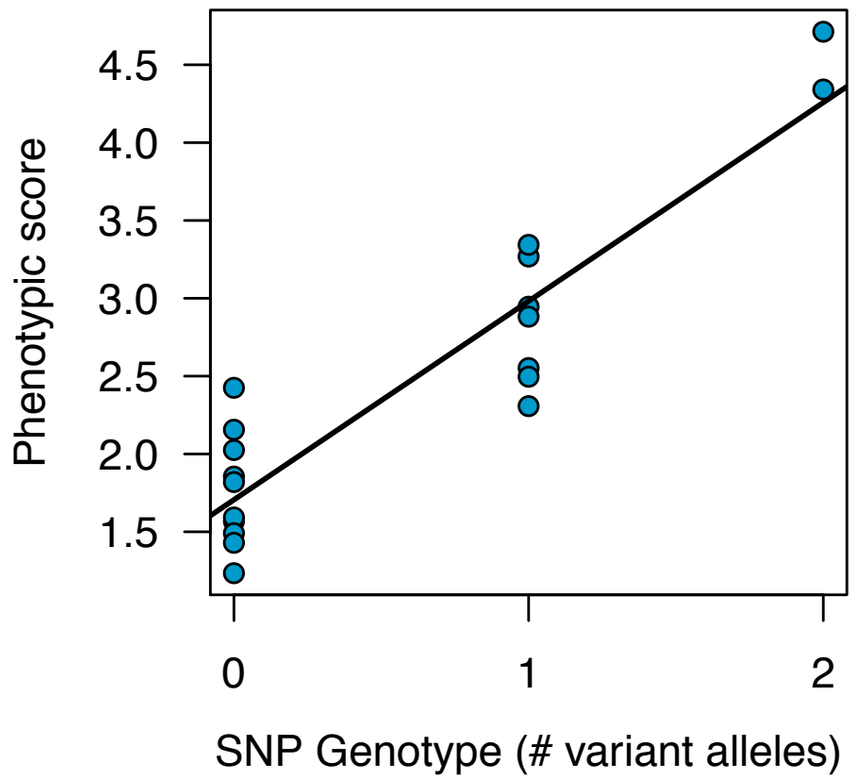
$$\text{OR}(aA \text{ vs } aa) = e^{0.4} = 1.5$$
$$\text{OR}(AA \text{ vs } aa) = e^{0.4 \times 2} = (e^{0.4})^2 = 1.5^2$$

*Additive on log odds scale, multiplicative on odds scale

What Model Should We Assume?

- For complex traits, some allele dosage effect is expected, so typically additive model is assumed in GWAS
- Another approach: test for all models and pick the one with the highest significance (MAX test)
    - But then need to adjust for multiple comparisons (e.g., if three models are tested, $P < 0.05/3$ should be considered statistically significant)
    - This is a conservative approach, since additive and dominant test results are often correlated (especially for low-frequency alleles)
    - Test for recessive model has very low power for other models (not informative unless true model is recessive)
- Note: this only applies to common variants, for which all three genotypes are observed. For rare variants - compare carriers to non-carriers

Association Tests for Quantitative Traits



- To test for association between genotype and a quantitative trait, fit a simple linear regression model to phenotype values:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 G$$

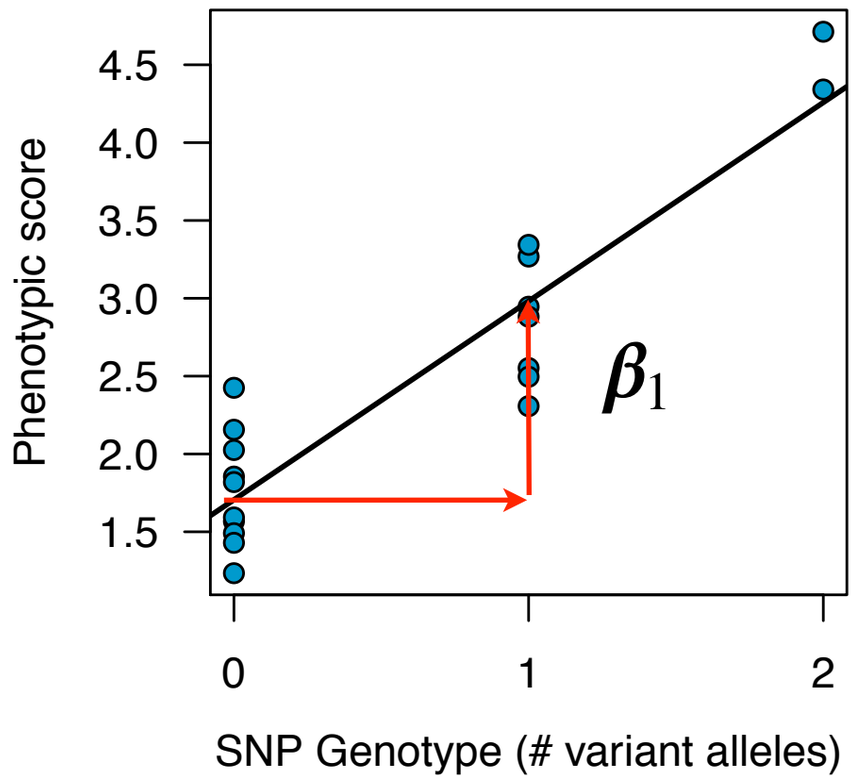
Mean value of the trait

Intercept

Slope = effect of genotype

Genotype

Association Tests for Quantitative Traits



- To test for association between genotype and a quantitative trait, fit a simple linear regression model to phenotype values:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 G$$

- Test the hypotheses: $H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$

Association Tests for Quantitative Traits

- Can include additional covariates (to adjust for potential confounders):

$$\mathrm{E}(Y) = \alpha + \beta_1 G + \beta_2 X$$

Mean value of the trait — Intercept — Effect of genotype — Effect of another covariate (e.g. ancestry)

- To test for different models, genotype G is coded as follows:
  - additive model: *aa = 0, aA = 1, AA = 2*
  - dominant model: *aa = 0, aA = 1, AA = 1*
  - recessive model: *aa = 0, aA = 0, AA = 1*

- To test for genotypic model, genotype is coded using two indicator (dummy) variables, for example:
  - G_1 = *1* if *aA* and *0* otherwise; G_2 = *1* if *AA* and *0* otherwise

Interpretation of linear regression coefficients

- Beta (regression coefficient) - estimates mean change in phenotype value per one unit change in genotype, i.e., *effect size*
  - For example, under additive model beta estimates mean change in phenotype value per each copy of the minor allele

- Beta is measured in the same units as the outcome
  - E.g., if estimated $\beta = 5$ for height measured in cm, then for height in meters, $\beta = 0.05$.

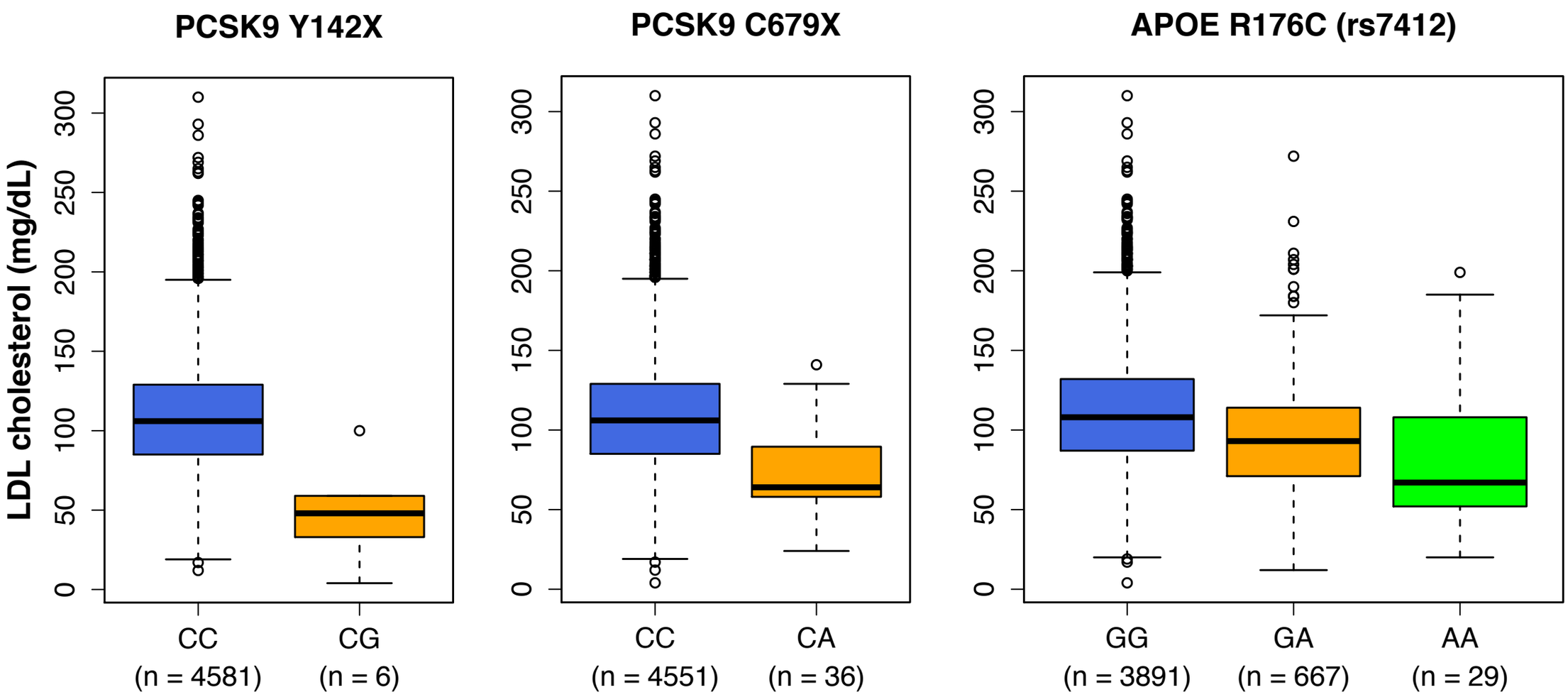
- So beta may not be the best way to characterize the strength of association. To make betas independent of units of measurement, standardize the outcome values (that is, subtract the mean and divide by standard deviation)
  - Then, estimated $\beta = 0.5$ means that the outcome is increased on average by 0.5 SD units per each additional allele

Measures of Association

- Another measure of association that is independent of the units of measurement is r^2 (coefficient of determination) from linear regression
  - for simple linear regression (no covariates), this is the square of Pearson's correlation coefficient between genotype and trait
  - estimates proportion of variance in trait explained by genotype

- r^2 varies between 0 and 1, however its magnitude for a particular genetic marker will depend on allele frequency as well as effect size
  - For example, alleles *A* and *B* are both associated with an average increase of 10 mg/dL is cholesterol level
  - If the population frequency of allele *A* is 1% and *B* is 10%, *B* will explain more variance in cholesterol levels than *A*.

Example



SNP	Beta (mg/dL)	Beta (SDU)	R-squared	P-value	MAF (AFR)	MAF (EUR)
PCSK9 Y142X	-58.2	-1.89	0.5%	2.92E-06	0.1%	0.0%
PCSK9 C679X	-37.0	-1.12	1.0%	1.27E-11	0.7%	0.0%
APOE rs7412	-16.6	-0.49	3.5%	5.18E-36	9.8%	7.7%

Linear Regression Assumptions

- Some assumptions of linear regression models:
  - Independent observations
  - (Residual) trait values are normally distributed; can be sensitive to outliers
  - Common variance within genotype groups

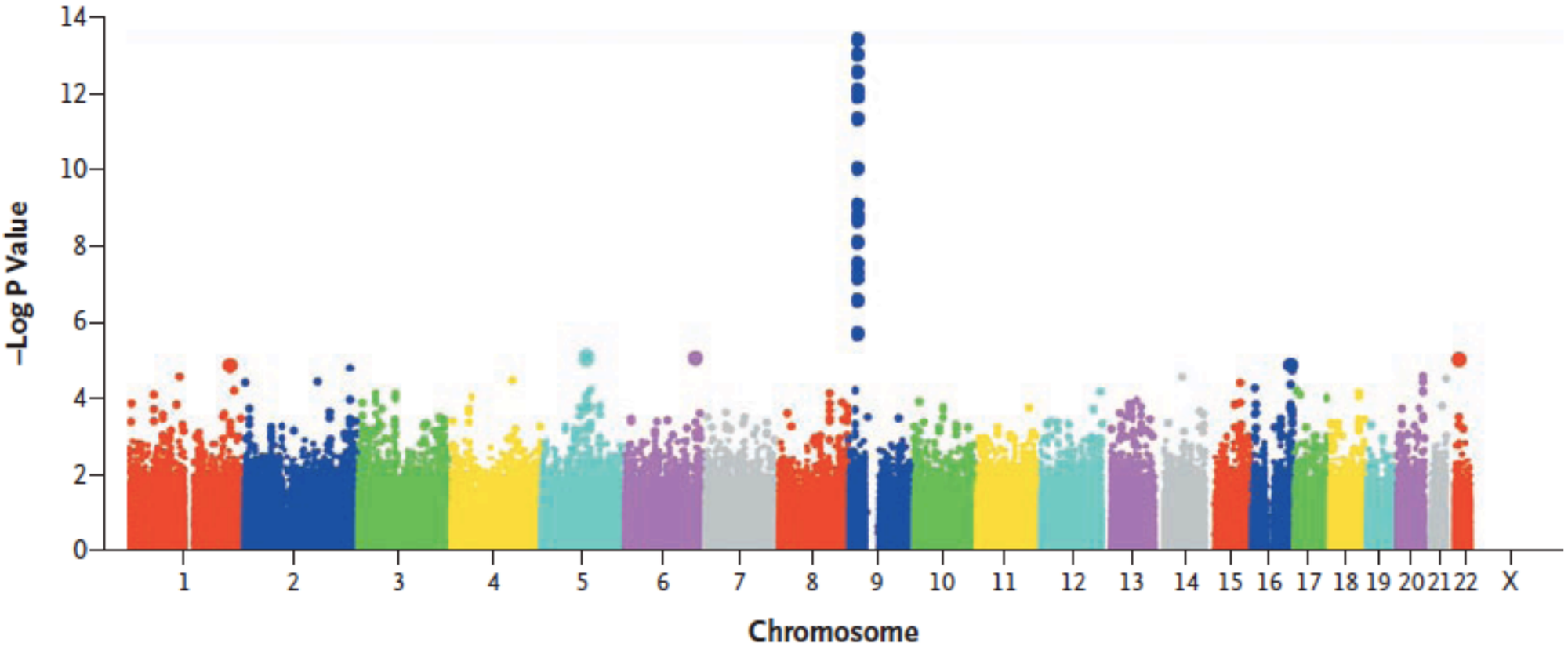
- If the assumptions do not hold, linear regression can produce incorrect results, and have either inflated or deflated type I error (false positive) rate

- If non-normal distribution:
  - Try a log transformation to make the distribution of residuals approximately normal
  - Some software package offer rank-based methods, e.g., Kruskal-Wallis test (does not adjust for covariates; not implemented in PLINK)

Outline

- Introduction: population-based sequencing studies
- Sequencing-based study workflow:
  - Study design
  - Data preparation and QC
  - Single-variant association tests
    - Association tests for binary traits (case-control)
    - Association tests for quantitative traits
  - Gene- or region-based association tests
  - Summarizing and prioritizing results

Manhattan Plot



Association of Single-Nucleotide Polymorphisms (SNPs) with Coronary Artery Disease or Myocardial Infarction in the Genomewide Association Analysis.

Samani et al. NEJM 2007

Multiple Testing

- For each SNP, test H_0 (no association) vs. H_1 (association)
- Calculate the test statistic (T) and *p*-value, and compare it to a significance threshold. Possible outcomes:

Truth	**Declared**	
	H0	***H1***
H0	true negative	false positive (type I error) controlled at $\alpha = 0.05$
H1	false negative (type II error) controlled at $\beta = 0.2$	true positive

What happens it we test M=1,000,000 hypotheses?

Expect α x M to be significant by chance (i.e., even if all H_0 are true). That is 50000 false positives!

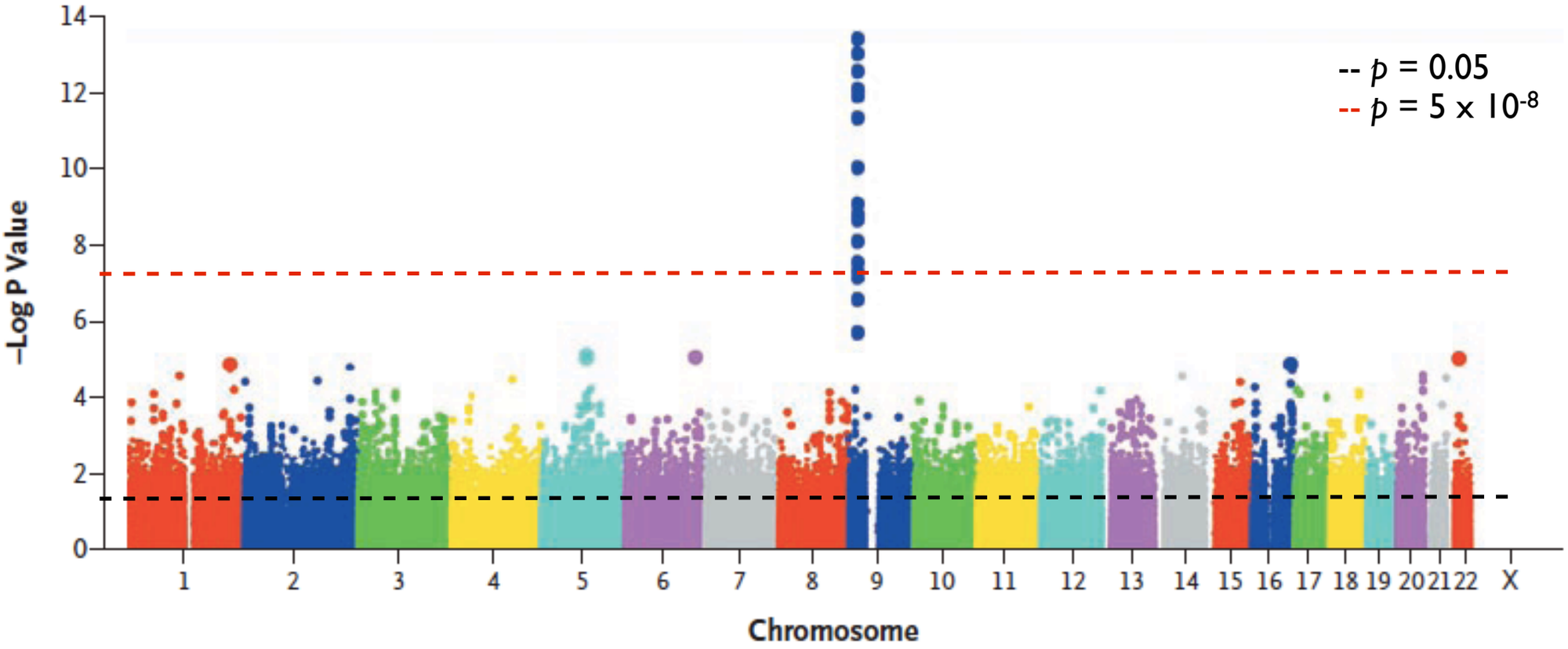
Multiple Testing Corrections

- To correct for multiple testing, choose a more stringent significance threshold for each variant
- Most common method used in GWAS is the Bonferroni correction:
  - Use $\alpha = 0.05/M$ for each test.
  - Controls ***family-wise error rate (FWER)*** - i.e., probability that there is at least one false positive finding - at 0.05:

$$\Pr(\#\text{ False Positives} > 0) \leq 0.05$$

  - Assumes tests are independent. Since many SNPs are in LD, can be overly conservative
  - Conventional significance threshold for GWAS is 5×10^{-8}, assuming 1 million independent SNPs.

Manhattan Plot

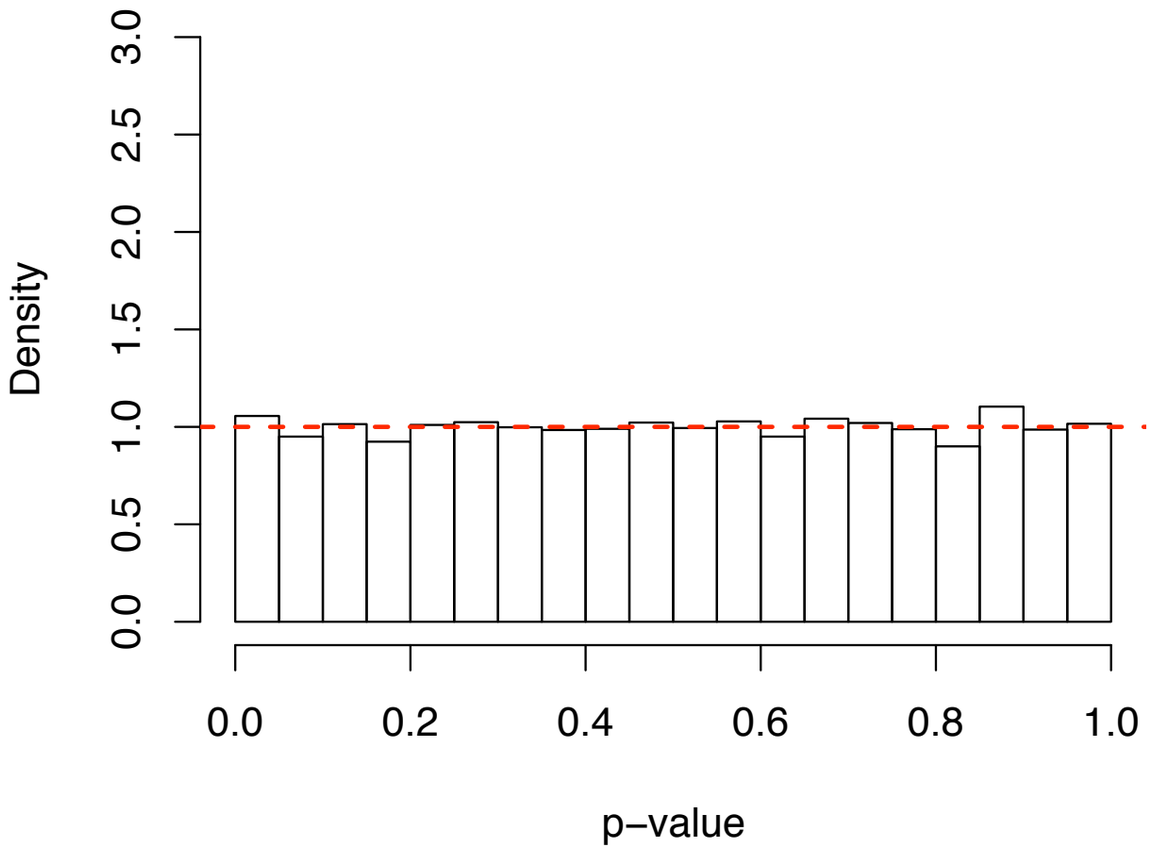


Association of Single-Nucleotide Polymorphisms (SNPs) with Coronary Artery Disease or Myocardial Infarction in the Genomewide Association Analysis.

Samani et al. NEJM 2007

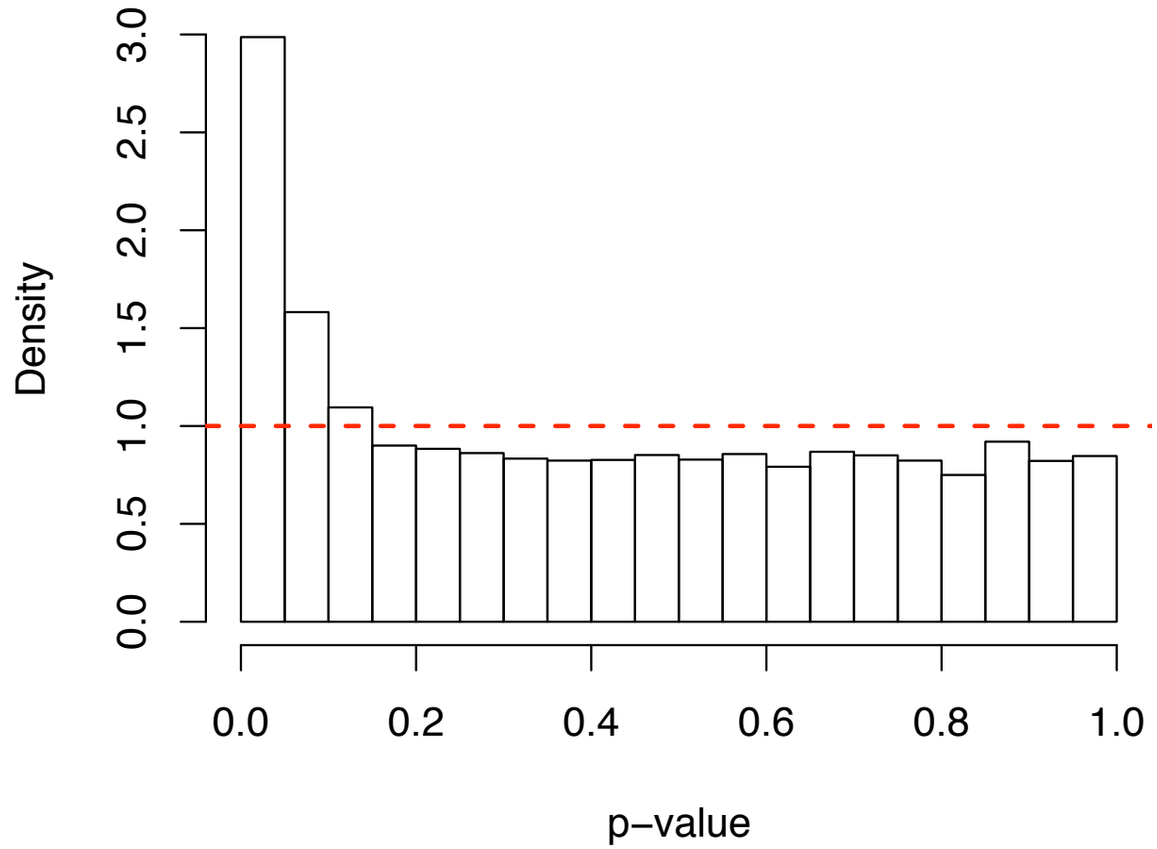
Distribution of P-Values

Expected (null) distribution



When all hypotheses are null (no association), the distribution of *p*-values is uniform (flat).

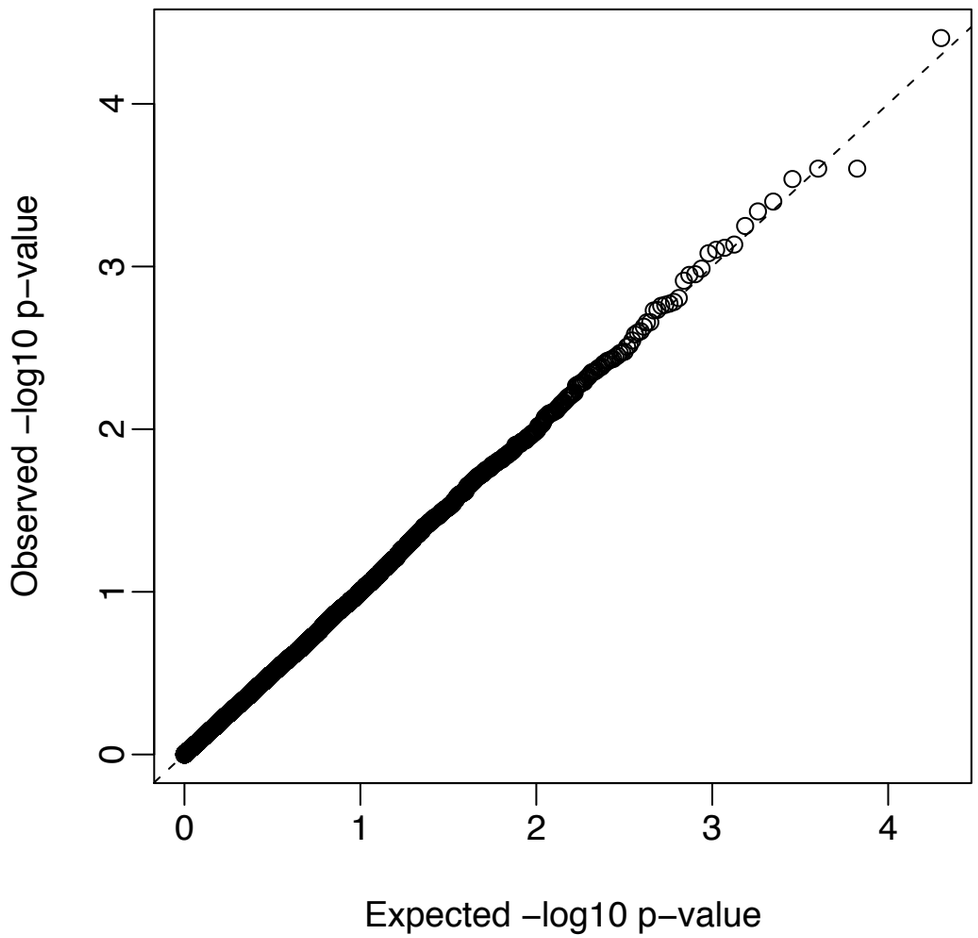
Observed distribution

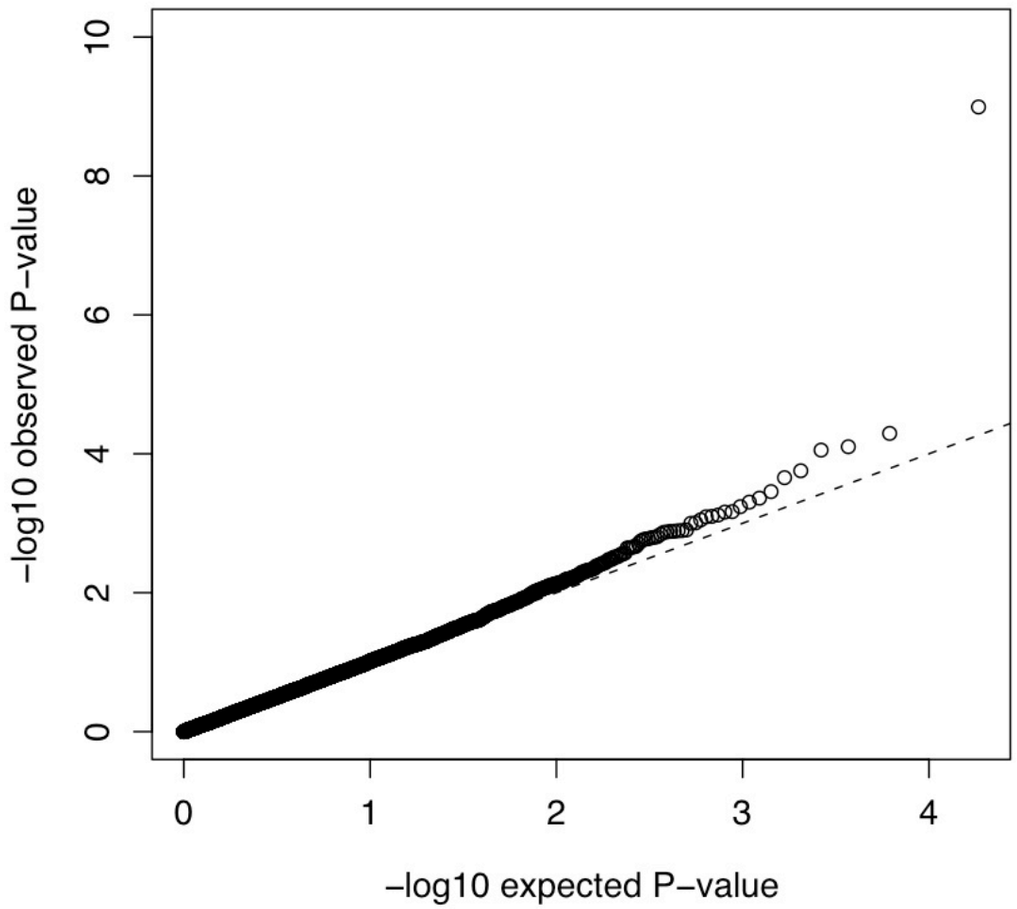


An excess of small *p*-values may indicate that some hypotheses are non-null (some SNPs are significant).

Quantile-Quantile (Q-Q) P-value Plots

Plot the observed ordered *p*-values against the expected ordered *p*-values...



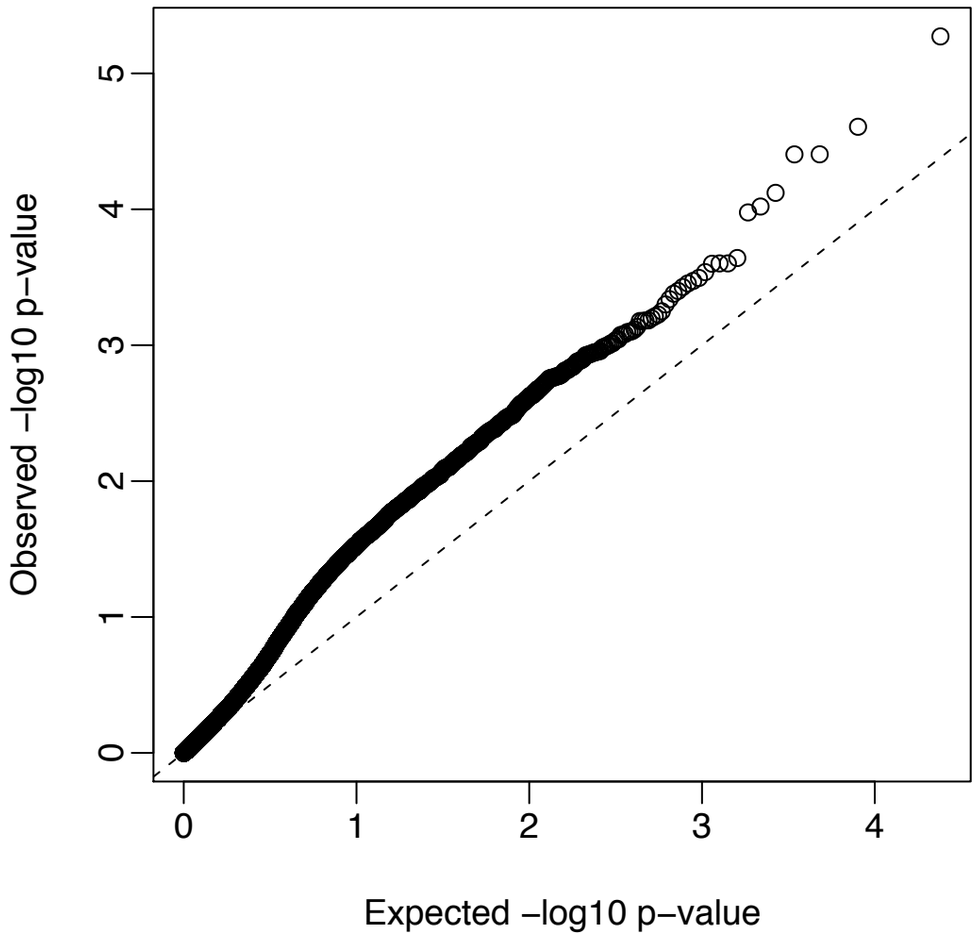


Plot the i^{th} smallest *p*-value against i/(M+1), where M is the number of markers.

When all hypotheses are null, the points fall on a straight line.

A deviation from the straight line indicates the presence of an association signal. In GWAS, want to see something like this...

Quantile-quantile (Q-Q) P-value plots



A systematic deviation from the y=x line may suggest inflated false positive rate due to population stratification or other bias

Solutions:

(1) Correct for population stratification; check for other biases

(2) to remove remaining inflation, use *Genomic control (GC):* the test statistic is computed at each of the null SNPs, and λ (i.e., inflation factor) is calculated as the empirical median divided by its expectation under the χ^2 distribution with 1 df. If $\lambda > 1$, then all test statistics are divided by λ before computing p-values.

Devlin, Roeder, 1999

Outline

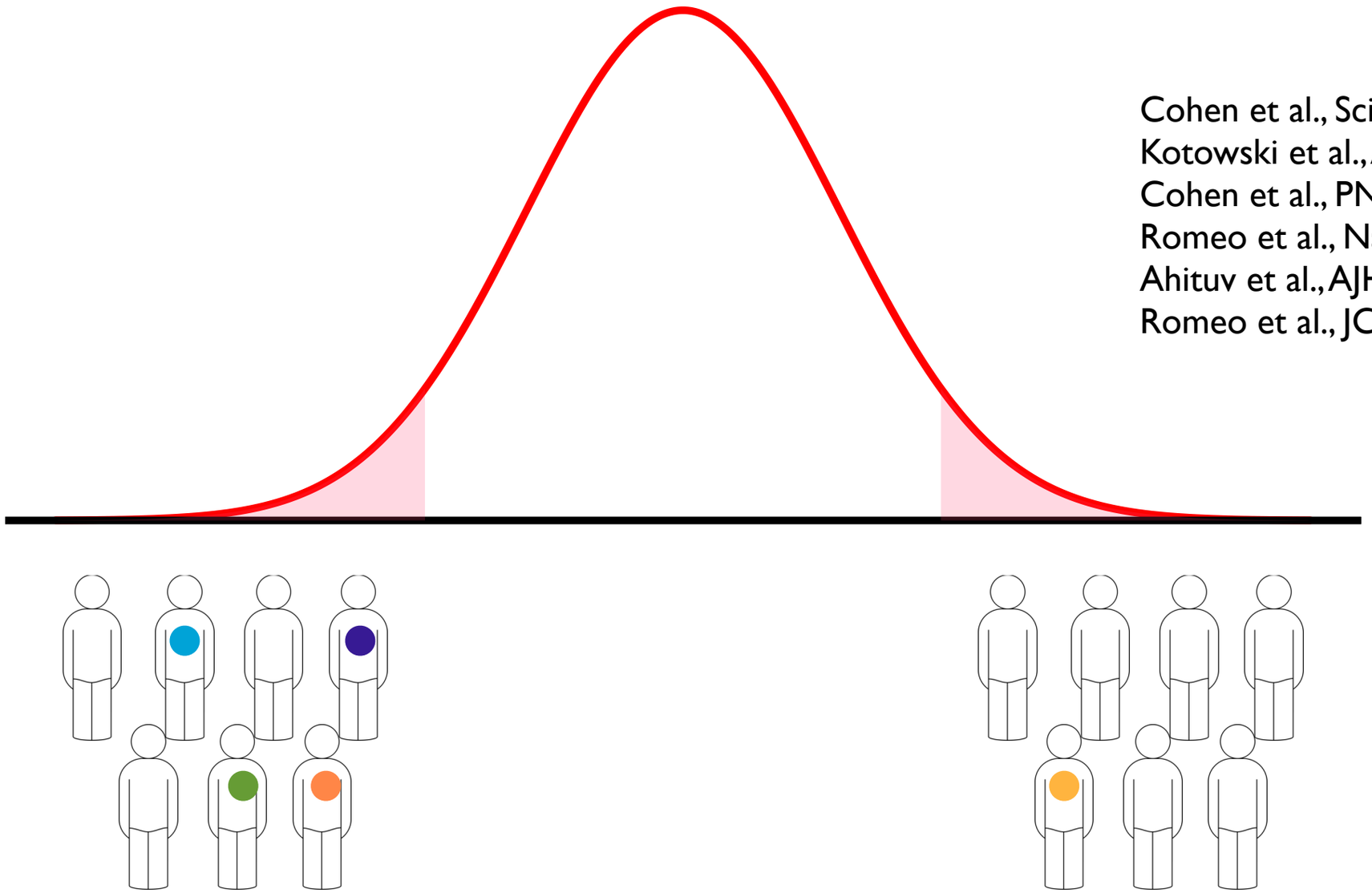
- Introduction: population-based sequencing studies
- Sequencing-based study workflow:
  - Study design
  - Data preparation and QC
  - Single-variant association tests
    - Association tests for binary traits (case-control)
    - Association tests for quantitative traits
  - Gene- or region-based association tests
  - Summarizing and prioritizing results

Rare-Variant Association Tests

- Traditional single-variant association tests have low power for rare variants, unless the effect size is very large
  - With an odds ratio (OR) = 1.4, the sample sizes required to achieve 80% power are 6,400, 54,000, and 540,000 for a MAF = 0.1, 0.01, and 0.001, respectively, if one assumes 5% disease prevalence and a significance level of 5×10^{-8}.
- One approach is to aggregate the information from multiple rare variants in a gene or region.
  - Gene- or region-based association tests

Evidence from Candidate-Gene Resequencing Studies

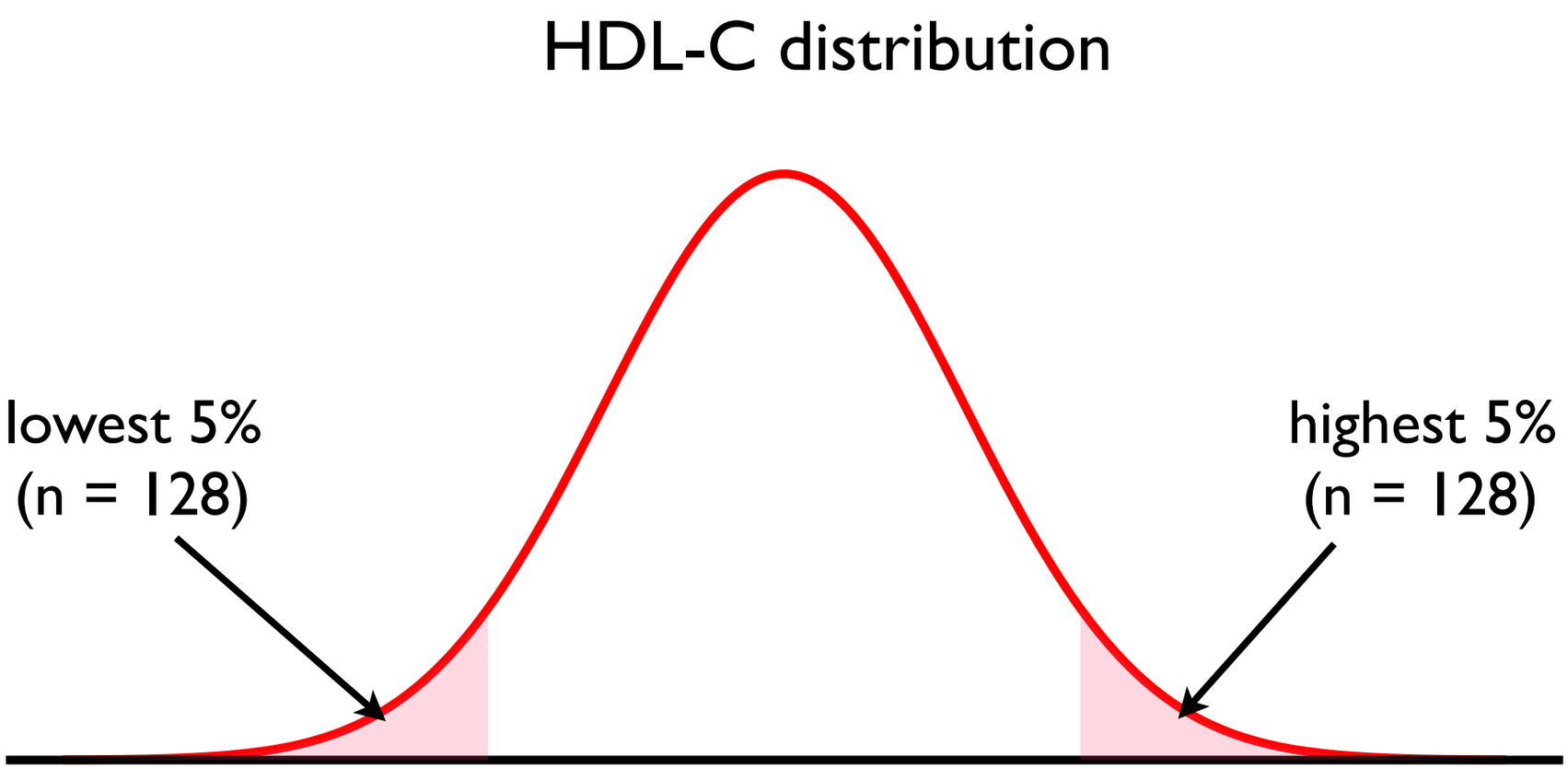
Resequence genes from individuals at the extremes of phenotypic distribution



Cohen et al., Science 2004
Kotowski et al., AJHG 2006
Cohen et al., PNAS 2006
Romeo et al., Nat Genet 2007
Ahituv et al., AJHG 2007
Romeo et al., JCI 2009

Rare missense mutations found preferentially at one end of the distribution

Candidate-Gene Resequencing Studies



21 (16%) individuals had sequence variants not present in the high HDL-C group

Variants present at both extremes were excluded

3 (2%) individuals had sequence variants not present in the low HDL-C group

Rare missense mutations found preferentially at one end of the distribution

Cohen et al., Science 2004

Combining Rare Variants

	Sequence variants unique to one group			Sequence variants common to both groups
	Low HDL-C (n =128)	High HDL-C (n =128)	P-value	
Non-synonymous	15 variants (21 people)	3 variants (3 people)	<0.001	10 variants
Synonymous	7 variants	6 variants	n.s.	19 variants

Cohen et al., Science 2004

Which variants to include?

Why rare variants?

- Evolutionary theory predicts that disease alleles should be rare.
- Empirical population genetic data show that deleterious variants are rare.
- Rare copy number variants contribute to several complex psychological disorders.
- Many rare familial disorders are due to rare alleles of large effect.

Gibson, Nat Rev Genet 2012

Rare Variant Tests

- Standard association tests can be applied to rare variants. However, they will have very low power, especially when correcting for multiple tests is performed.
- Approach: collapse genotypes across variants. I.e., define

$$G = \begin{cases} 1 \text{ if rare variant present} \\ 0 \quad \text{otherwise} \end{cases}$$

- Alternatively: G = number of rare alleles in a gene for each person (however, G>1 is unlikely for rare variants)
- Test for the difference in the burden of rare variants between cases and controls, or for correlation between a quantitative trait and the burden of rare variants

Combining Rare Variants in a Single Test

- **Minor allele frequency**
  - deleterious alleles are likely to be rare (population genetics)
  - is there an optimal threshold for allele frequency (<0.1%, 1%, 5%)?
- **Functional class**
  - disrupting alleles (nonsense, splice-site) > missense > synonymous
- **Sequence conservation**
  - variants altering a highly conserved amino-acid residue are more likely to be functional
- **Computational algorithms for predicting the functional consequences of nonsynonymous mutations, e.g.,**
  - SIFT (Ng and Henikoff, Nucleic Acids Res, 2003),
  - PolyPhen-2 (Adzhubei et al., Nat Methods 2010),
  - Genetic Evolutionary Rate Profiling (GERP)

Limitations of Burden Tests

- **Ignore low-frequency variants that are present at both extremes**

  - filter out variants that exceed the chosen allele frequency threshold

- **Assume that all variants influence trait in the same direction**

  - some genes may have both gain-of-function and loss-of-function alleles (e.g., PCSK9, Kotowski et al., AJHG 2006)

  - both extremes may have an excess of rare variants

Alternatives to Burden Tests

- **Instead of filtering out variants:**

  - jointly assess the effect of common and rare variants: CMC test (Li & Leal, AJHG 2008)

  - assign lower weight to common and higher weight to rare variants, e.g., WSS (Madsen & Browning, PLoS Genet 2009), SKAT (Wu et al., AJHG 2011)

  - weight variants according to apparent effect size: KBAC (Liu & Leal, PLoS Genet 2010).

- **Allow for effects in different directions:**

  - e.g., C-alpha (Neale et al., PLoS Genet 2011), SKAT (Wu et al., AJHG 2011)

Combining multiple Rare Variants for Association Tests

Generalized linear modeling framework

$$f(Y) = \alpha + \beta \sum_{i=1}^{m} w_i G_i + \tau X + \gamma Z$$

f(Y) = E(Y) or mean of Y for quantitative trait, or log{P[Y=1]/(1-P[Y=1])} for binary trait

α: intercept

β: regression coefficient of weighted sum

w_i: weight of variant i

G_i: genotype (recoded) of variant i

X: covariate(s), such as population structure variable(s)

$\boldsymbol{\tau}$: regression coefficients for covariates

Z: design matrix corresponding to γ

γ: random polygene effects for individual subjects

Multi-marker tests

Burden-type test:

$$f(Y) = \alpha + \beta \sum_{i=1}^{m} w_i G_i + \tau X$$

H_0: $\beta = 0$

Non-burden tests (e.g. SKAT):

$$f(Y) = \alpha + \sum_{i=1}^{m} \beta_i G_i + \tau X$$

H_0: $\beta_i = 0$ ($\beta_1 = \beta_2 = \ldots = \beta_m = 0$)

Which Test to Use?

- **The optimal test depends on the architecture of a particular gene and disease**

  - this is often unknown

- **Common approach: apply several tests (or parameters, such as MAF threshold) and pick the best result**

  - this comes at the cost of multiplicity

  - must account for the total number of tests performed (genes x tests)

Examples of findings from exome/genome sequencing studies for complex diseases

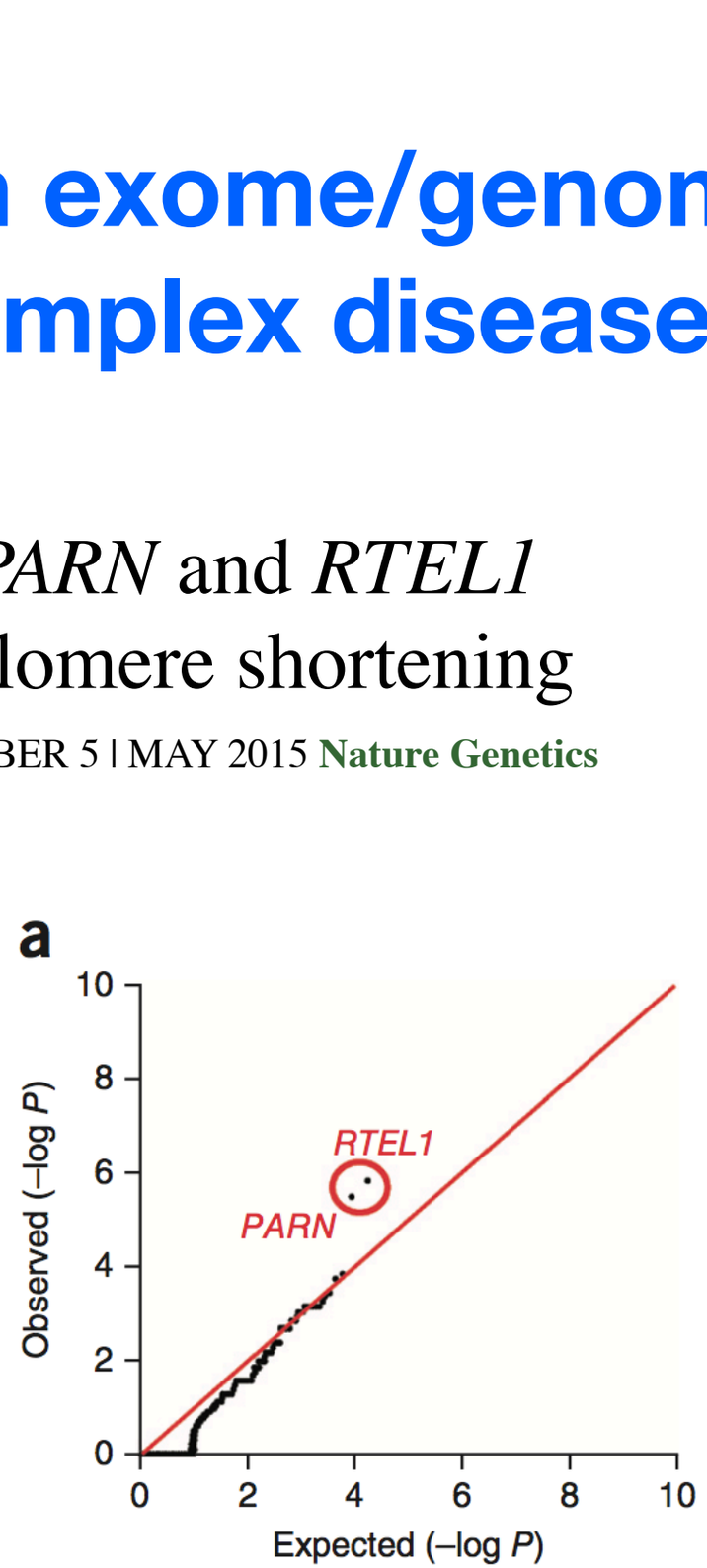
Exome sequencing links mutations in *PARN* and *RTEL1* with familial pulmonary fibrosis and telomere shortening

Stuart et al. VOLUME 47 | NUMBER 5 | MAY 2015 **Nature Genetics**

- Performed whole exome sequencing of 99 cases with familial pulmonary fibrosis of unknown genetic cause

- Compared to 2,816 European controls from

- Performed gene burden test:
  - Rare variants
  - Predicted damaging variants



Validation and Replication

- **Technical validation**: ensure validity of genotypes
  - confirm genotypes for a SNP showing signal by PCR or Sanger
  - this is especially important for imputed SNPs

- **Replication** in independent populations
  - confirm initial association
  - show generalizability to different ethnic populations

Questions:

- Which associations to replicate?
- What constitutes an adequate replication?
- How to interpret failure to replicate?

Replicating genotype–phenotype associations

What constitutes replication of a genotype–phenotype association, and how best can it be achieved?

Box 3 | Suggested criteria for establishing positive replication

These criteria are intended for follow-up studies of initial reports of genotype–phenotype associations assessed by genome-wide or candidate-gene approaches.

- Replication studies should be of sufficient sample size to convincingly distinguish the proposed effect from no effect
- Replication studies should preferably be conducted in independent data sets, to avoid the tendency to split one well-powered study into two less conclusive ones
- The same or a very similar phenotype should be analysed
- A similar population should be studied, and notable differences between the populations studied in the initial and attempted replication studies should be described
- Similar magnitude of effect and significance should be demonstrated, in the same direction, with the same SNP or a SNP in perfect or very high linkage disequilibrium with the prior SNP (r^2 close to 1.0)
- Statistical significance should first be obtained using the genetic model reported in the initial study
- When possible, a joint or combined analysis should lead to a smaller *P*-value than that seen in the initial report[75]
- A strong rationale should be provided for selecting SNPs to be replicated from the initial study, including linkage-disequilibrium structure, putative functional data or published literature
- Replication reports should include the same level of detail for study design and analysis plan as reported for the initial study (Box 1)

Chanock S, NCI-NHGRI Working Group on Replication in Association Studies, Nature, 447, 2007

Assessing Functional Impact

- Association can only establish correlation, not causation (even for variants predicted to have a functional effect)
- To prove the causal effect, conduct functional studies:
  - Cell culture, model organisms
  - Does the variant affect gene expression, protein synthesis / transport / function

Web Resources and Software

UCSC Genome Browser Gateway

http://genome.ucsc.edu

PLINK Whole genome association analysis toolset

PLINK 1.07 (old version, better documentation)

http://pngu.mgh.harvard.edu/purcell/plink/

PLINK 1.9 (x 10s times faster, documentation assumes some familiarity)

https://www.cog-genomics.org/plink2

Haploview (https://www.broadinstitute.org/haploview/haploview)

Graphical tool for viewing PLINK results and SNP analysis

Locuszoom (locuszoom.org/)

Graphical tool for visualizing regional association results

EPACTS (https://genome.sph.umich.edu/wiki/EPACTS)

versatile software pipeline to perform various statistical tests for identifying genome-wide association from sequence data

(Selected) References

- Anderson et al. (2010) Data quality control in genetic case-control association studies. Nature Protocols, Vol. 5, No. 9, 1564

- Balding DJ (2006). A tutorial on statistical methods for population association studies. Nature Reviews: Genetics, 7:781

- Clarke et al. (2011). Basic statistical analysis in genetic case-control studies. Nature Protocols, Vol. 6, No. 2, 121

- Zondervan & Cardon. (2007) Designing candidate gene and genome-wide case–control association studies. Nature Protocols, Vol. 2, No. 10, 2492

- Ziegler A, König IR, Thompson JR. (2008) Biostatistical Aspects of Genome-Wide Association Studies. Biometrical Journal 50:8-28.