BICF Nano Course: GWAS

Statistical Analysis of GWAS

Julia Kozlitina
Julia.Kozlitina@UTSouthwestern.edu
McDermott Center for Human Genetics
April 28, 2017

Outline

Introduction

Single-variant association tests

- Genetic models
- Association tests for binary traits (case-control)
- Association tests for quantitative traits

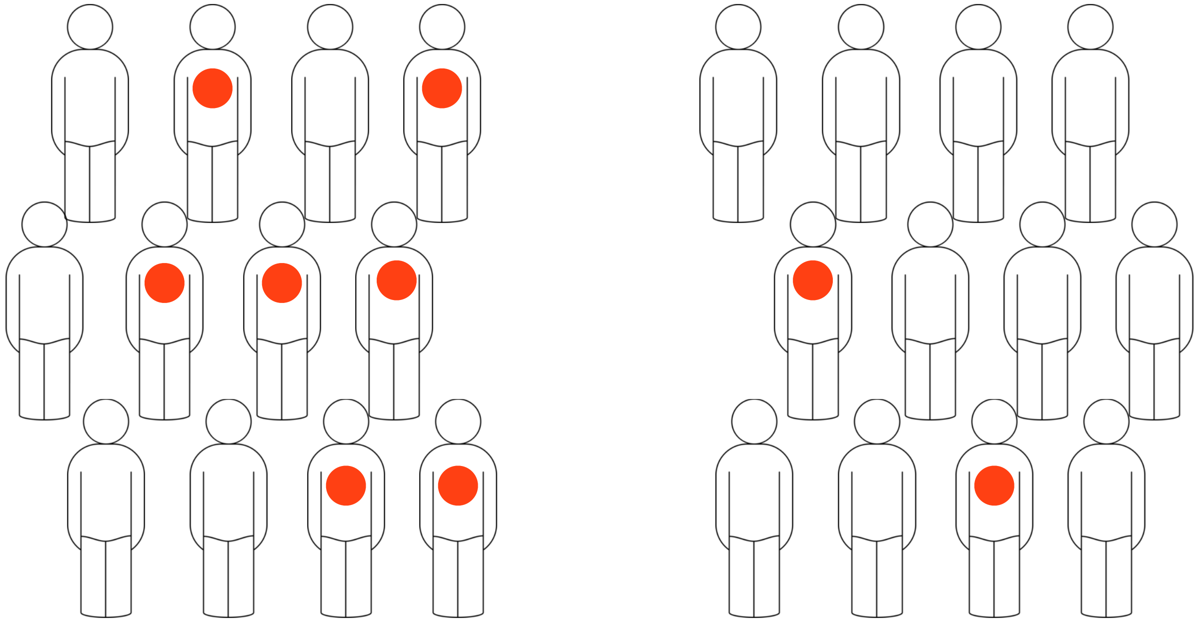
GWAS Workflow

- Data quality control (QC)
- Multiple testing

Validation and replication strategies

Genetic Association Studies

Compare the frequency of alleles (or genotypes) at a given genetic marker, between unrelated individuals with and without a given disease (cases and controls) to determine if there is a statistical association between the disease and the genetic marker



Affected (cases):

$p_{A.} = 7/12 = 58\%$

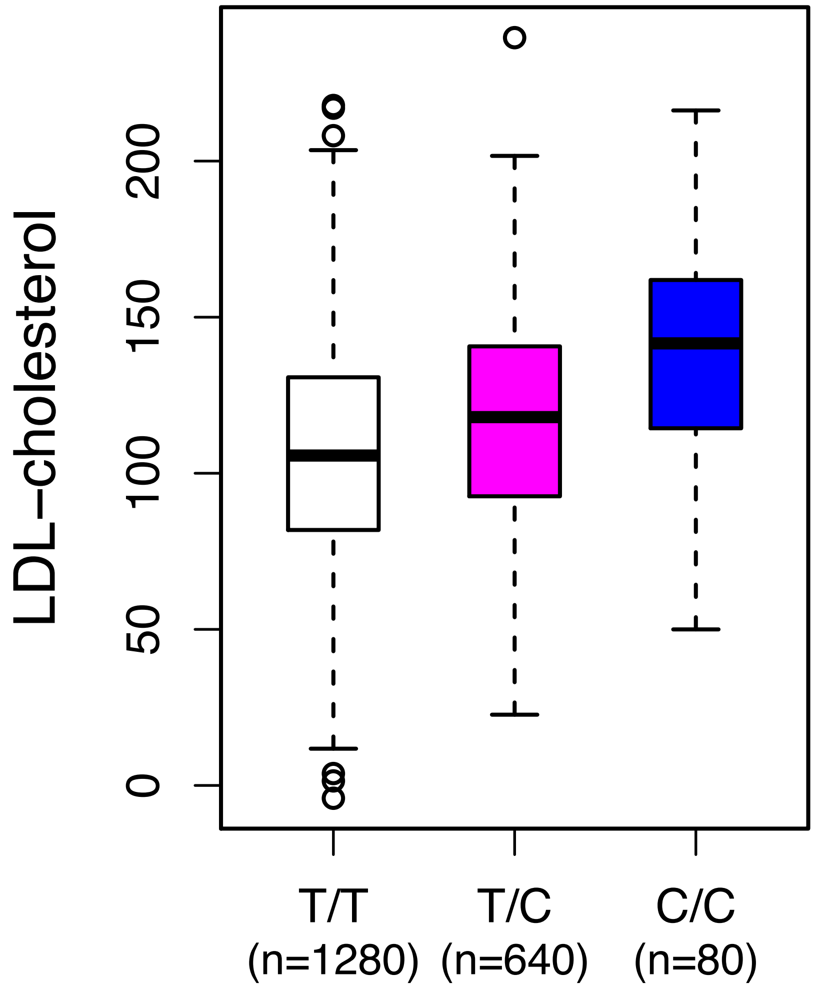
Unaffected (controls):

$p_{A.} = 2/12 = 8.3\%$

● allele A

Genetic Association Studies

When the trait of interest is quantitative, compare the distribution or mean value of the trait between unrelated individuals with different genotypes to see if there is statistical association between trait and genotype



Individuals with C/T and C/C genotype at a given SNP have on average higher LDL cholesterol levels than those with T/T genotype

Types of Association Studies

Require special methods for analysis

Family-based:

Recruit parent-child trios (focus on transmission of marker alleles from parents to offspring) or discordant sib-pairs (discordant alleles)

- Advantages: robust to assumptions (e.g., population stratification)
- Disadvantages: power, not easy to ascertain sufficient numbers of families

The most common type of study used and the one we will cover

Population-based:

Recruit *unrelated* individuals from a given population (e.g., affected cases and healthy controls, or a random sample from a general population)

- Advantages: power, easier to recruit participants
- Disadvantages: prone to biases, confounding (e.g., population stratification)

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

Single-variant association tests

- Genetic models
- Association tests for binary traits (case-control)
- Association tests for quantitative traits

GWAS Workflow

- Data quality control (QC)
- Multiple testing

Validation and replication strategies

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:

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

  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	Mean trait value		
Model	*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

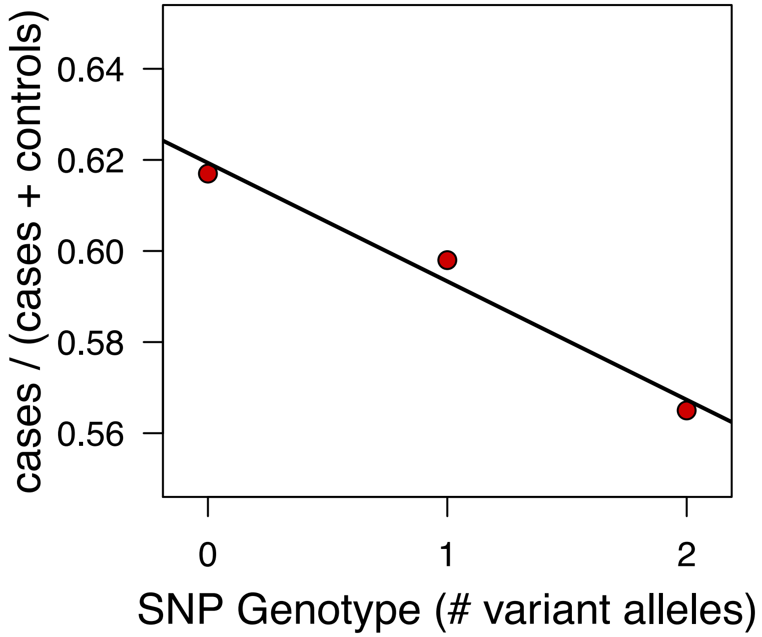
Exact P-values

- 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/multiplicative 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)***:

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

- 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)***:

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

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

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 = $\frac{\pi}{1-\pi}$

- **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

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*: $$\mathrm{OR}_{aA} = \frac{n_{11}n_{20}}{n_{10}n_{21}}$$

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

- 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: $OR_{\text{1-2 vs 1-1}} = 0.71$, $OR_{\text{2-2 vs 1-1}} = 0.20$

  $\chi^2 = 11.75$, $df = 2$, 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:

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

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 β_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:

OR(*aA* vs *aa*) = $e^{0.4}$ = 1.5
OR(*AA* 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)

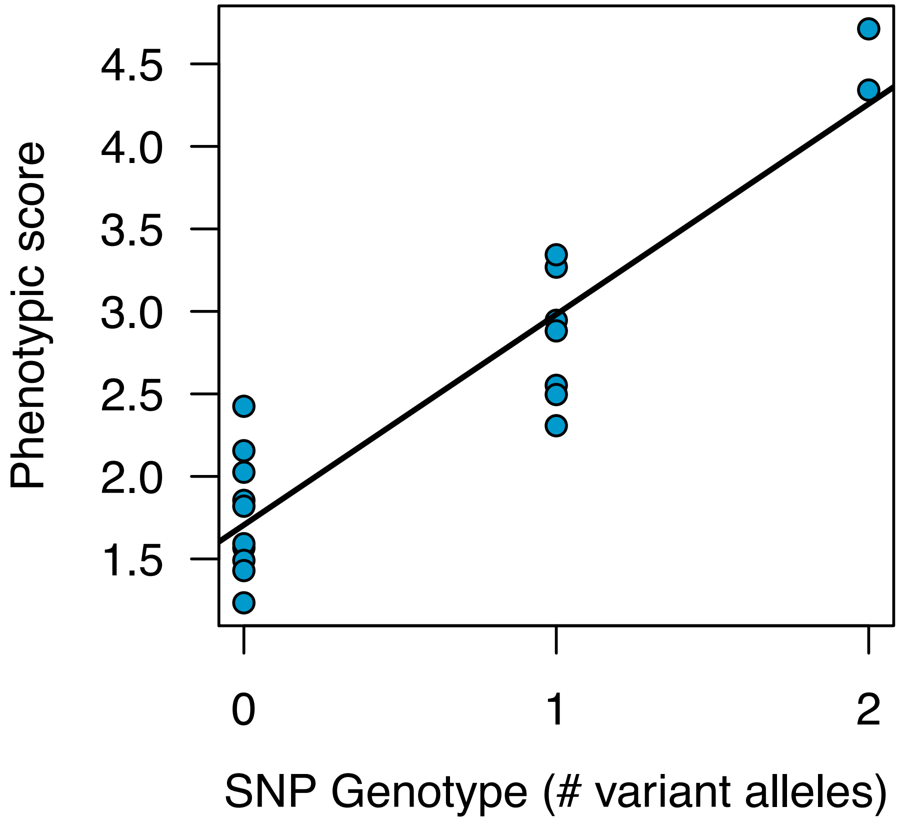
  - But then need to adjust for multiple comparisons (e.g., if three models are tested, $P < 0.05/3$ should be considered statistically significant)

  - 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)

- This a developing area of research. In practice, additive model works well in most cases.

Association Tests for Quantitative Traits



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

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

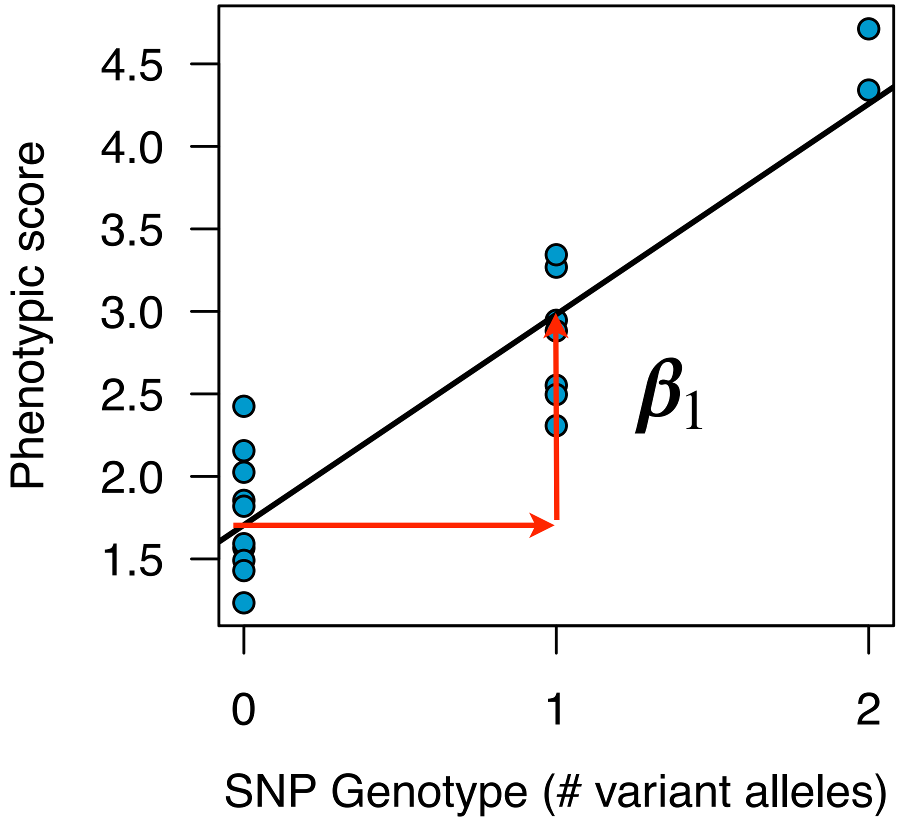
Mean value of the trait

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
    - 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 β = 5 for height measured in cm, then for height in meters, β = 0.05.

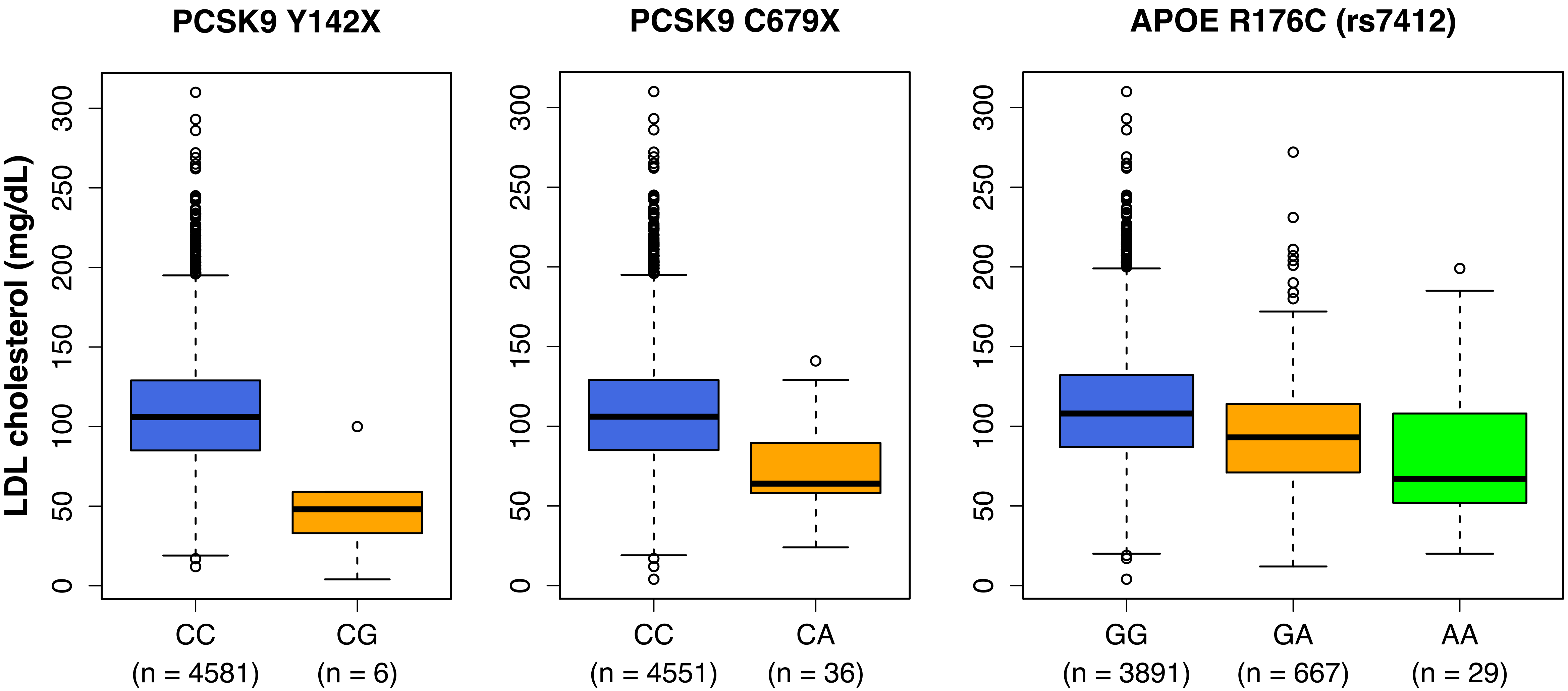
- So beta may not be the best way to characterize the strength of association. To make betas independent of units of measurement, standardize the outcome values (that is, subtract the mean and divide by standard deviation)
    - Then, estimated β = 0.5 means that the outcome is increased on average by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

Measure 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

Single-variant association tests

- Genetic models
- Association tests for binary traits (case-control)
- Association tests for quantitative traits

GWAS Workflow

- Data quality control (QC)
- Multiple testing

Validation and replication strategies

GWAS Workflow:

1. Preliminary steps:
   - Data quality control (QC): per-individual and per-variant
   - Estimate principal components of ancestry
   - Perform imputation (if desired)

2. Decide on the association test and model, covariates to include, transformation, etc.

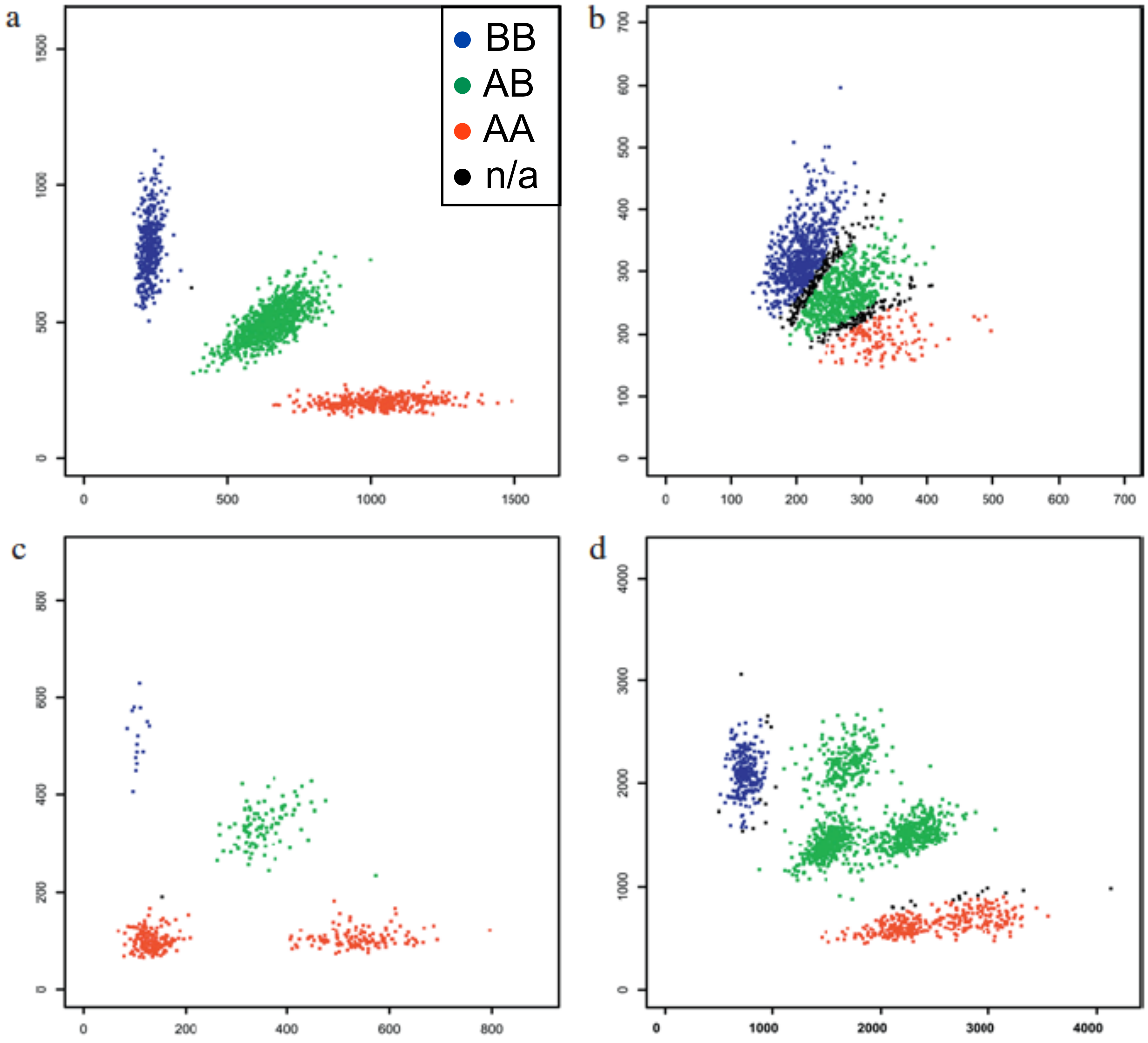
3. For each SNP, run the association test and record the p-value

4. Summarize the results
   - Check the quality of genome-wide association results
   - Establish genome-wide significance cut-off

Genotyping Methods

- SNP Arrays (e.g., Illumina HumanOmni BeadChip)
    - Target up to 1 million (and more) SNPs
    - Contain allele specific probes for each SNP
    - Genotypes are called based on intensity cluster plots

Genotype calling in SNP arrays



high genotyping quality

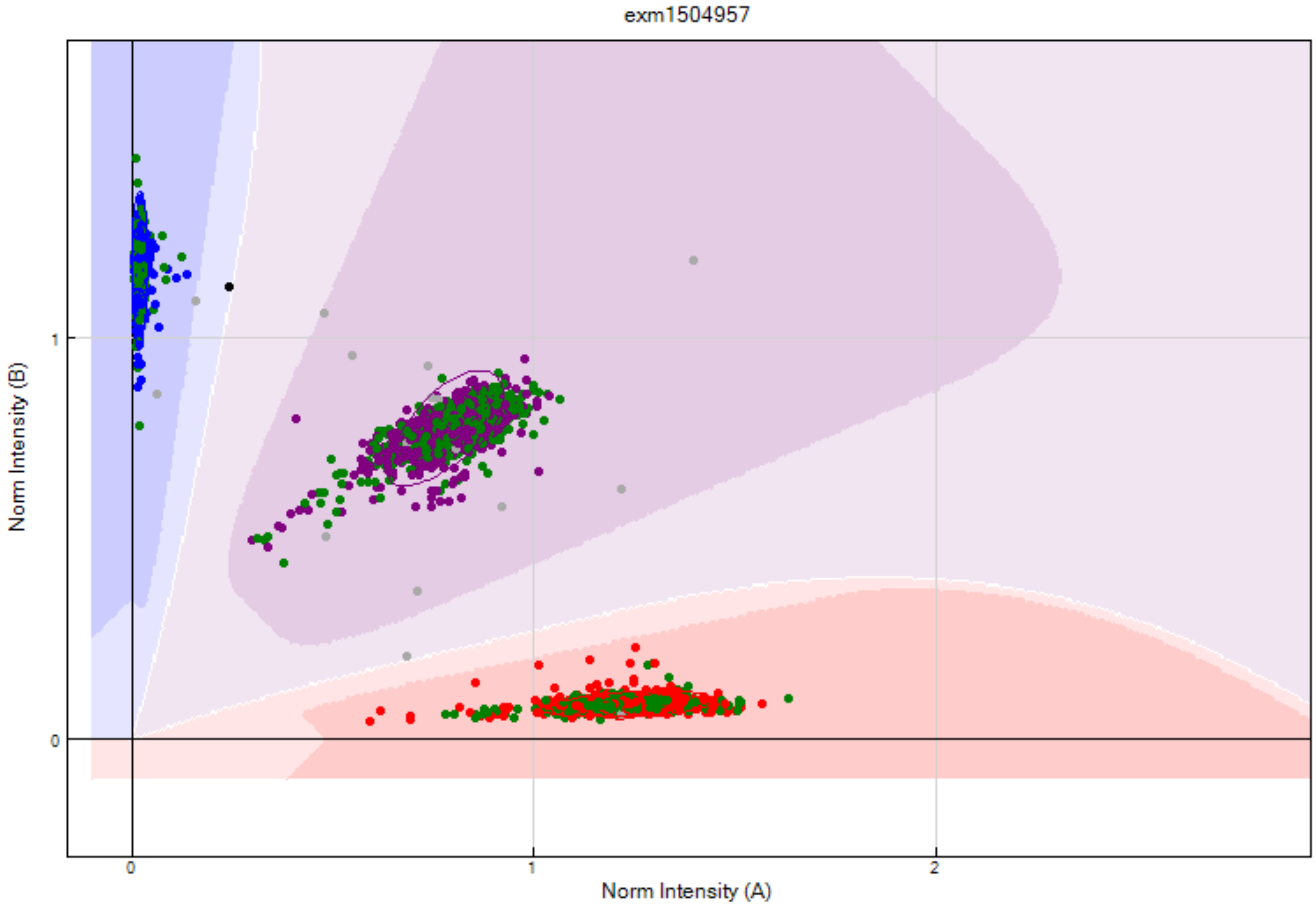
poor separation; many missing calls

fourth cluster with low intensity for both alleles

genotyped sequence not unique

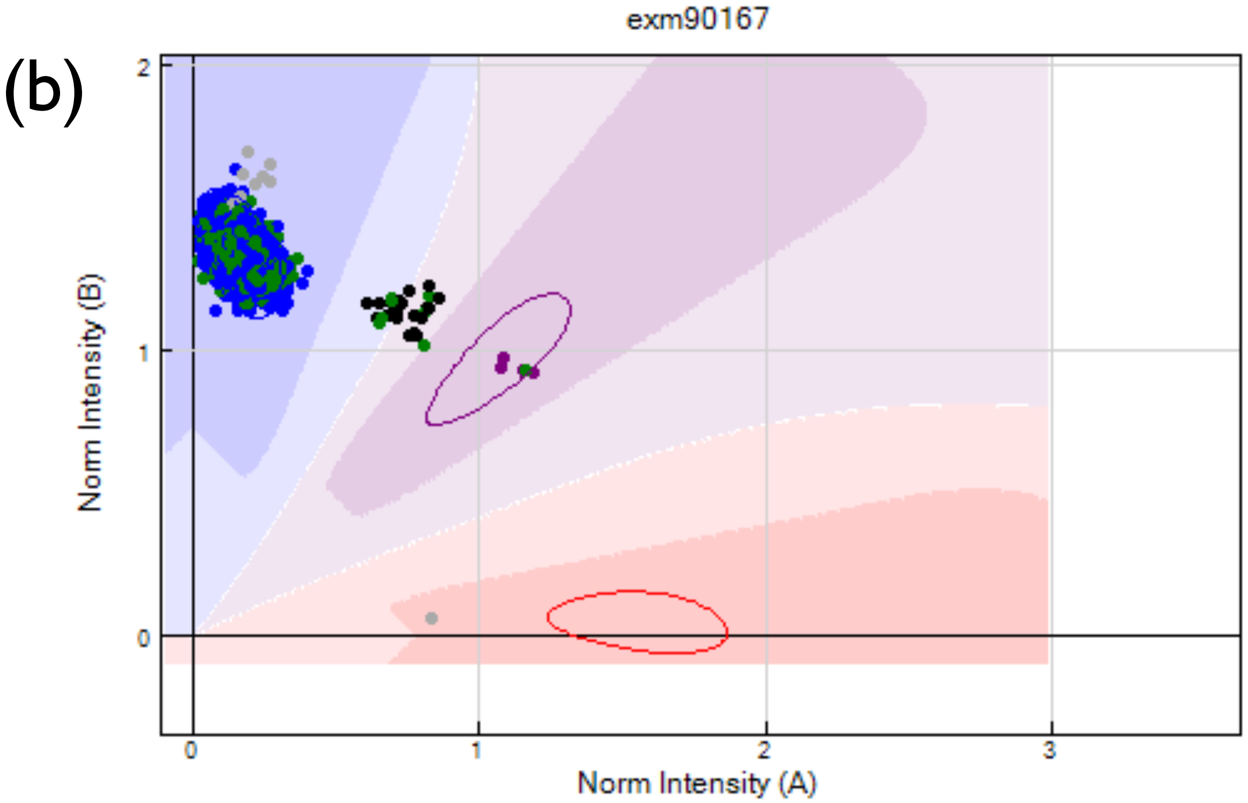
Ziegler et al.
Biom J (2008)

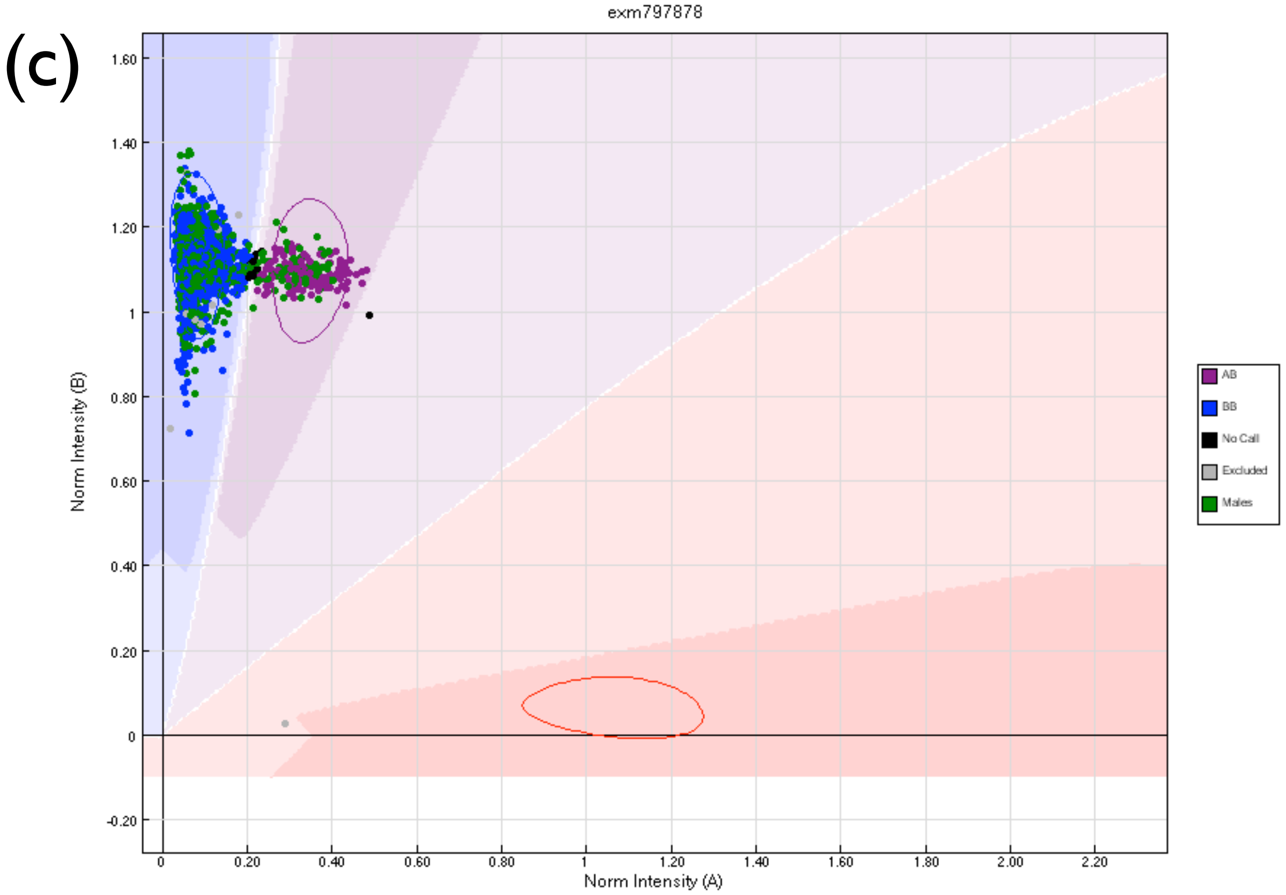
Examples: Illumina Human Beadchip

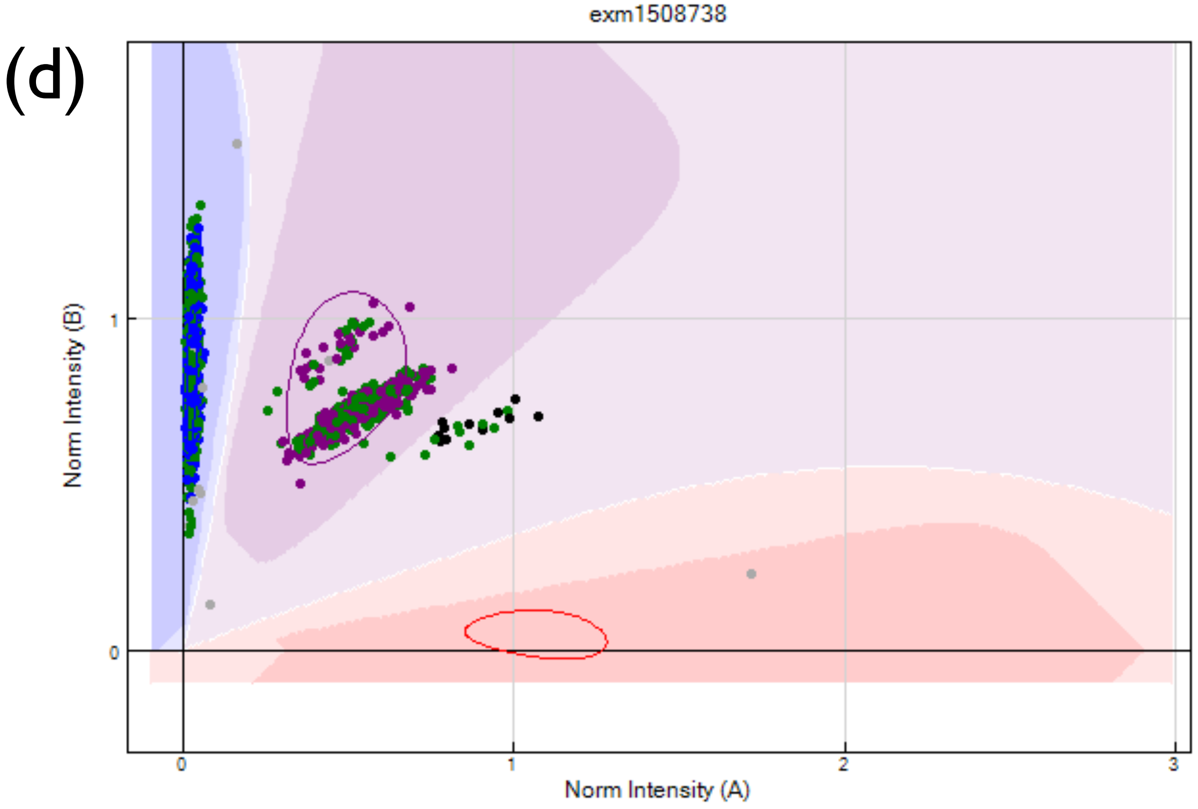


(a) Good clustering

Examples: Illumina Human Beadchip







(b) heterozygotes set to missing

(c) clusters fail to separate

(d) multiple clusters

Very small percent (<0.1%); often can be detected by testing HWE

Genotyping Methods

- SNP Arrays (e.g., Illumina HumanOmni BeadChip)

  - Target up to 1 million (and more) SNPs

  - Contain allele specific probes for each SNP

  - Genotypes are called based on intensity cluster plots

- Next-generation Sequencing

  - Covers whole exome or whole genome

  - Involves massive parallel sequencing of short reads

  - Genotype call quality depends on sequence quality for a given base, mapping/alignment quality, coverage (how many reads cover each base)

Variant (SNP) QC:

Reproducibility on same platform > 99%

Cross-platform concordance > 95%

Exclude SNP if:

- excessive missing genotype call rate across samples (e.g., >3%)
- significantly different missing genotype rates between cases and controls
- significant deviation from Hardy-Weinberg equilibrium (e.g., $p < 10^{-5}$)
- monomorphic
- minor allele frequency <1% (may be of lower quality, low power to detect association)
    - however, sequencing studies and new methods focus on rare variation

Ziegler et al., Biometrical Journal 50:8-28, 2008.
Anderson et al.. Nature Protocols, Vol. 5, No. 9, 2010

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)
- discordant duplicate pairs
- related to other subjects in the study - can be detected by calculating the proportion of shared alleles at genotyped SNPs (identity by state, IBS)
- individuals of divergent ancestry

Either exclude or correct for this

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

Multiple Testing

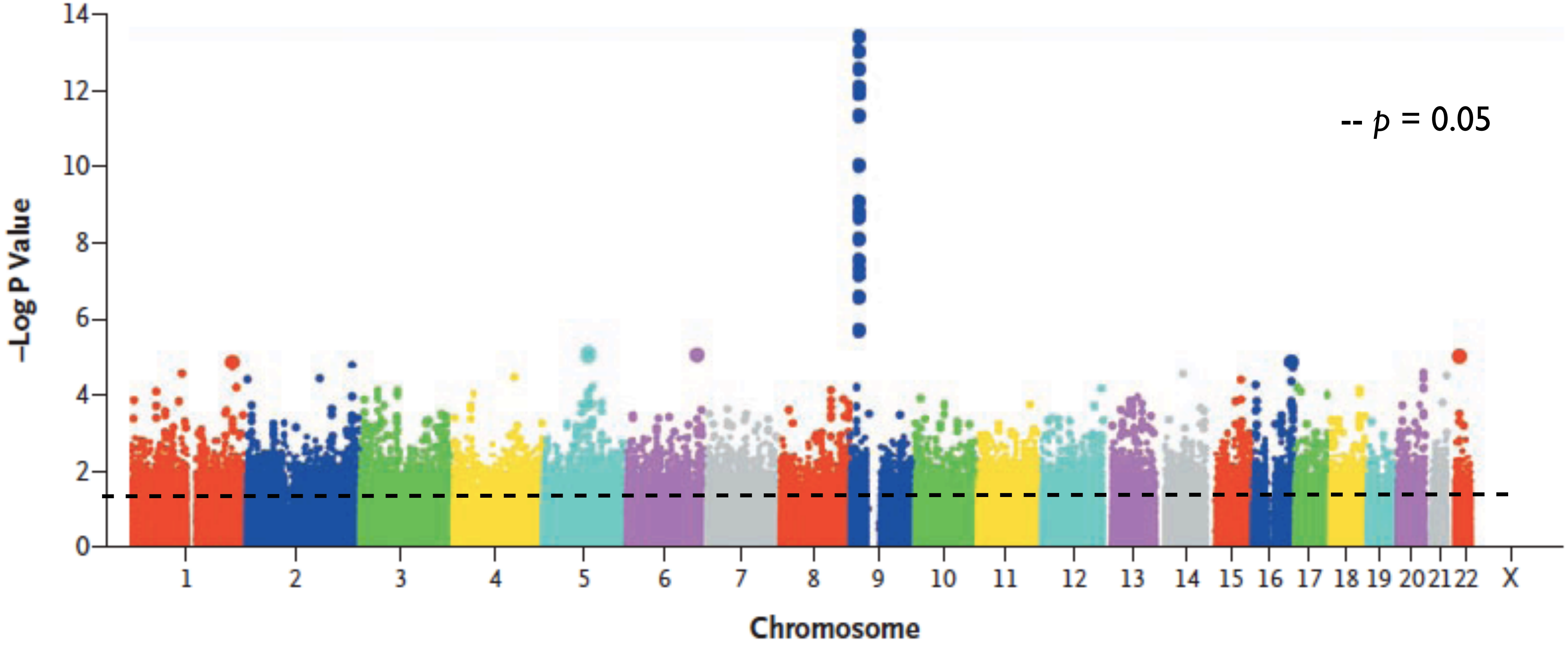
- For each SNP, test H_0 (no association) vs. H_1 (association)
- Calculate the test statistic (T) and *p*-value, and compare it to a significance threshold. Possible outcomes:

Truth	**Declared**	
	H_0	***H_1***
H_0	true negative	false positive (type I error) controlled at $\alpha = 0.05$
H_1	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!

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 Corrections

- Choose more stringent significance threshold

- 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 $0.05/1000000 = 5 \times 10^{-8}$

- Sidak's correction:
  - Use $\alpha = 1 - (1 - 0.05)^{1/M}$ for each test.
  - Produces FWER of exactly 0.05 when tests are independent. Only slightly less stringent than Bonferroni.

Multiple Testing Corrections (2)

- There are other methods, that are less conservative
- For example, step-down methods (Holm 1979):
  - Order the p-values from smallest to largest, $p_1 < p_2 < \ldots < p_M$. Starting with p_1, reject the null for all H_i, for which

    $$p_i < 0.05/(M - i + 1).$$

  - Stop when you find the first j such that $p_j > 0.05/(M-j+1)$. Accept the null for the remaining hypotheses, $H_j \ldots H_M$.

Multiple Testing Corrections (3)

- Define a new error rate

  False Discovery Rate (FDR)[1]:

  - Set q to a desired value, e.g., $q = 0.05$
  - Order the p-values, $p_1 < p_2 < \ldots < p_M$. Find the largest i such that

    $$p_i \leq i \times q / \mathrm{M}$$

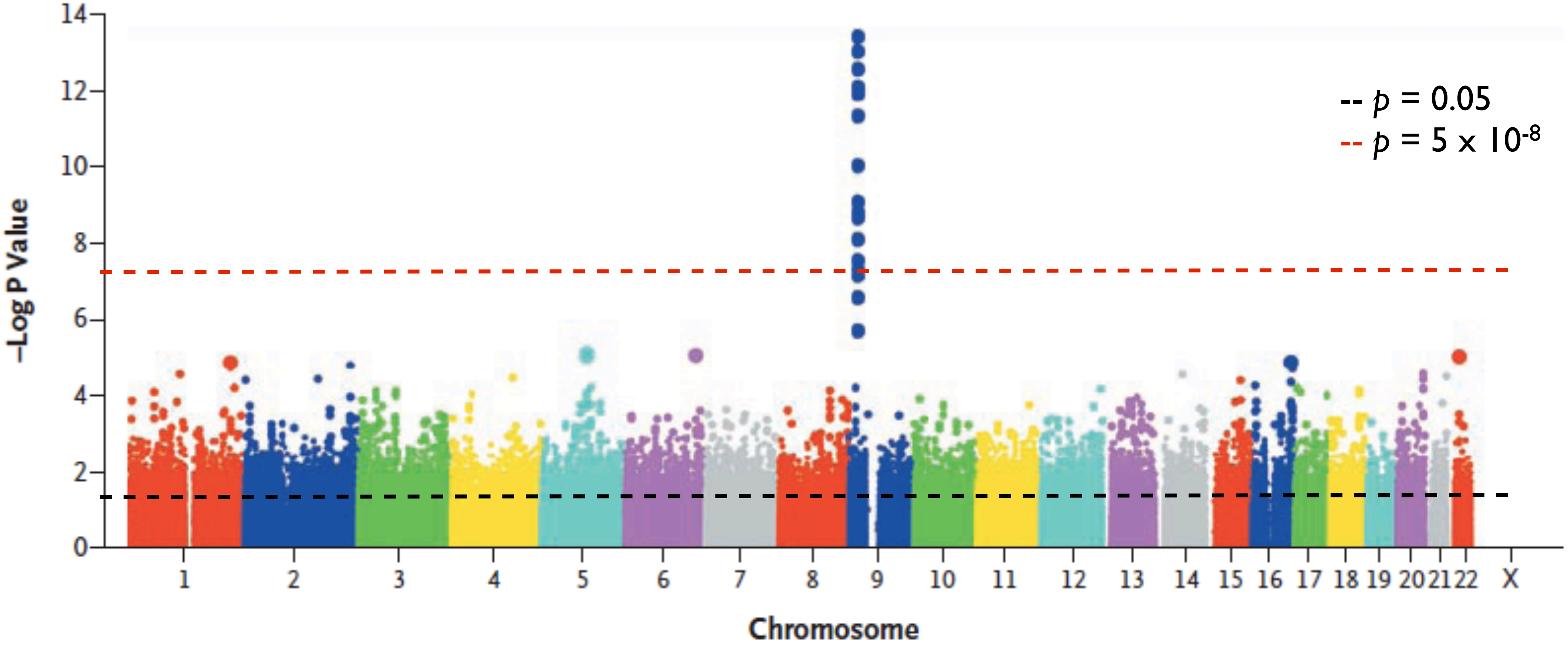
  - Reject $H_1 \ldots H_i$ and accept the remaining hypotheses, $H_{i+1} \ldots H_M$.
  - Controls expected proportion of false positives among the rejected tests at $q = 0.05$:

    $$\mathrm{E}(FP/S) \leq 0.05$$

  - More useful when a high proportion of tests are expected to be significant, e.g., in RNA-seq experiments; less relevant in GWAS, where only a small number of SNPs are expected to have a true effect on the trait.

[1] Benjamini Y and Hochberg Y JRSSB 1995

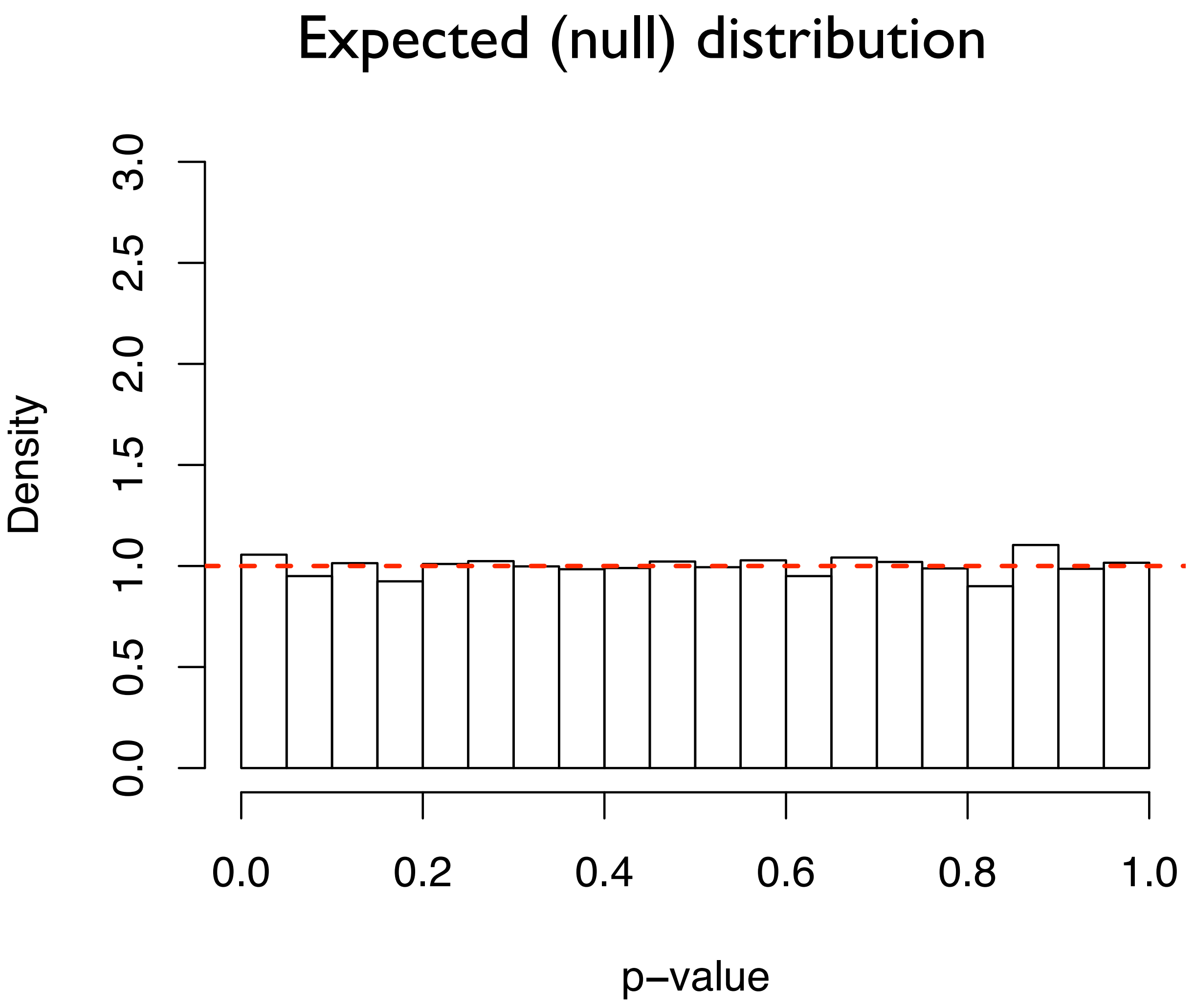
Manhattan Plot

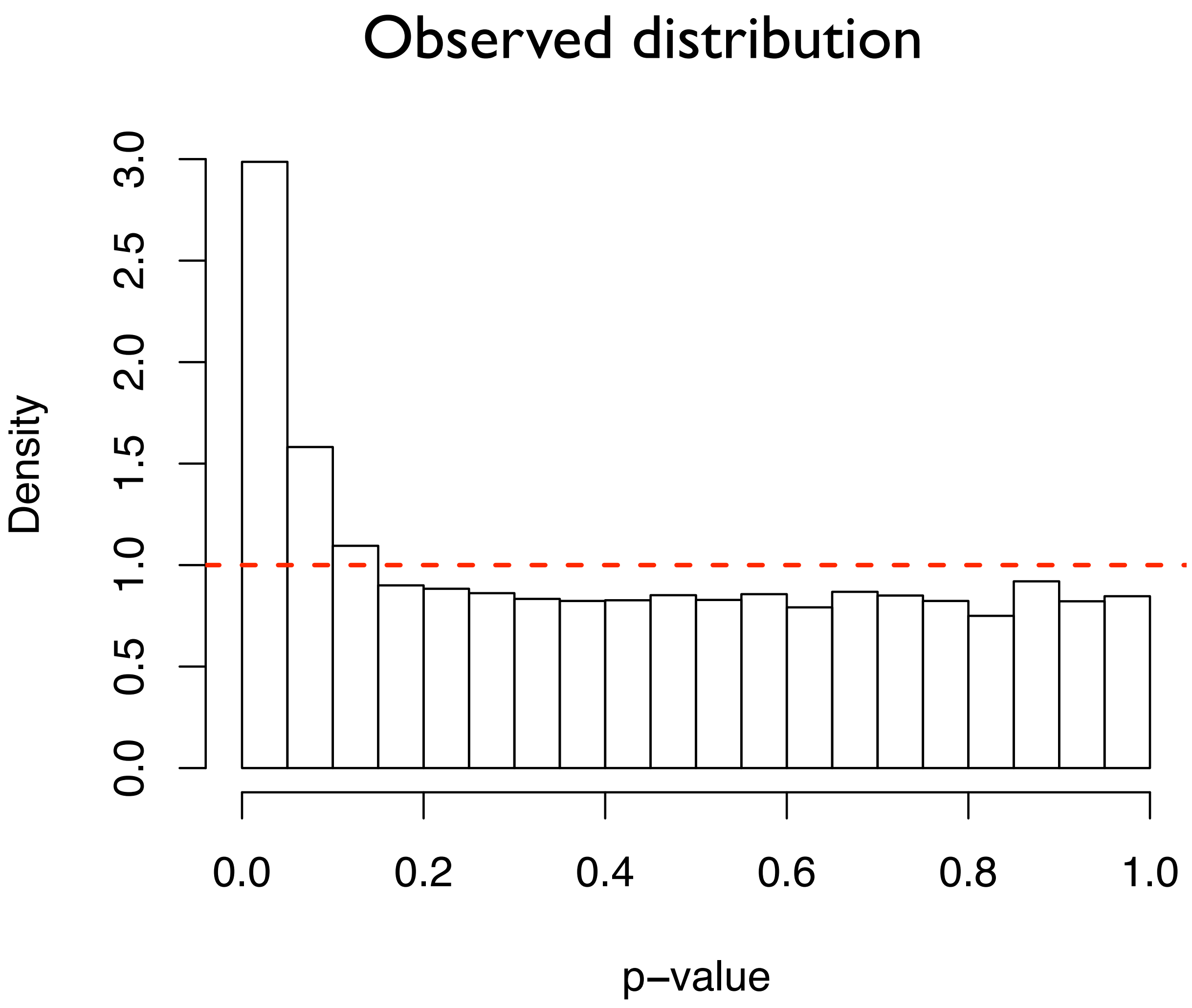


Association of Single-Nucleotide Polymorphisms (SNPs) with Coronary Artery Disease or Myocardial Infarction in the Genomewide Association Analysis.

Samani et al. NEJM 2007

Distribution of P-Values



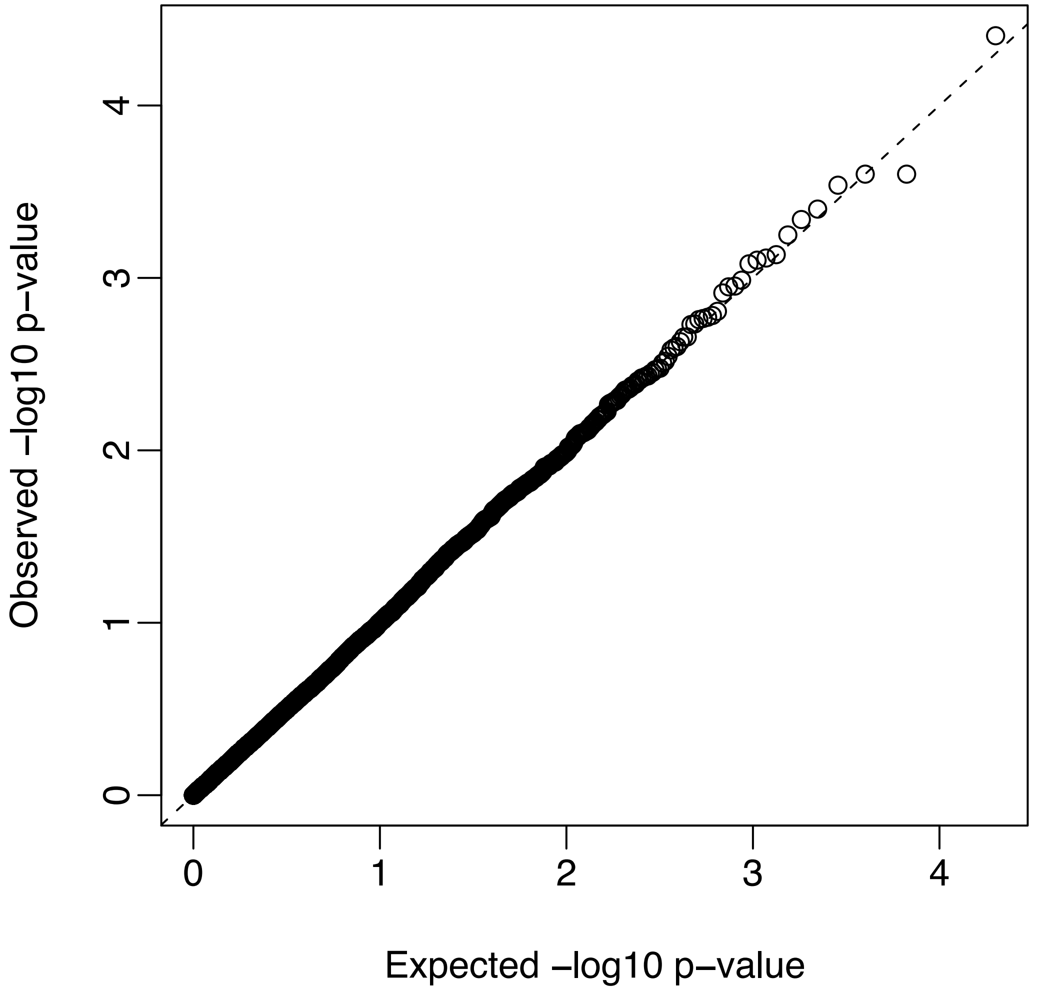


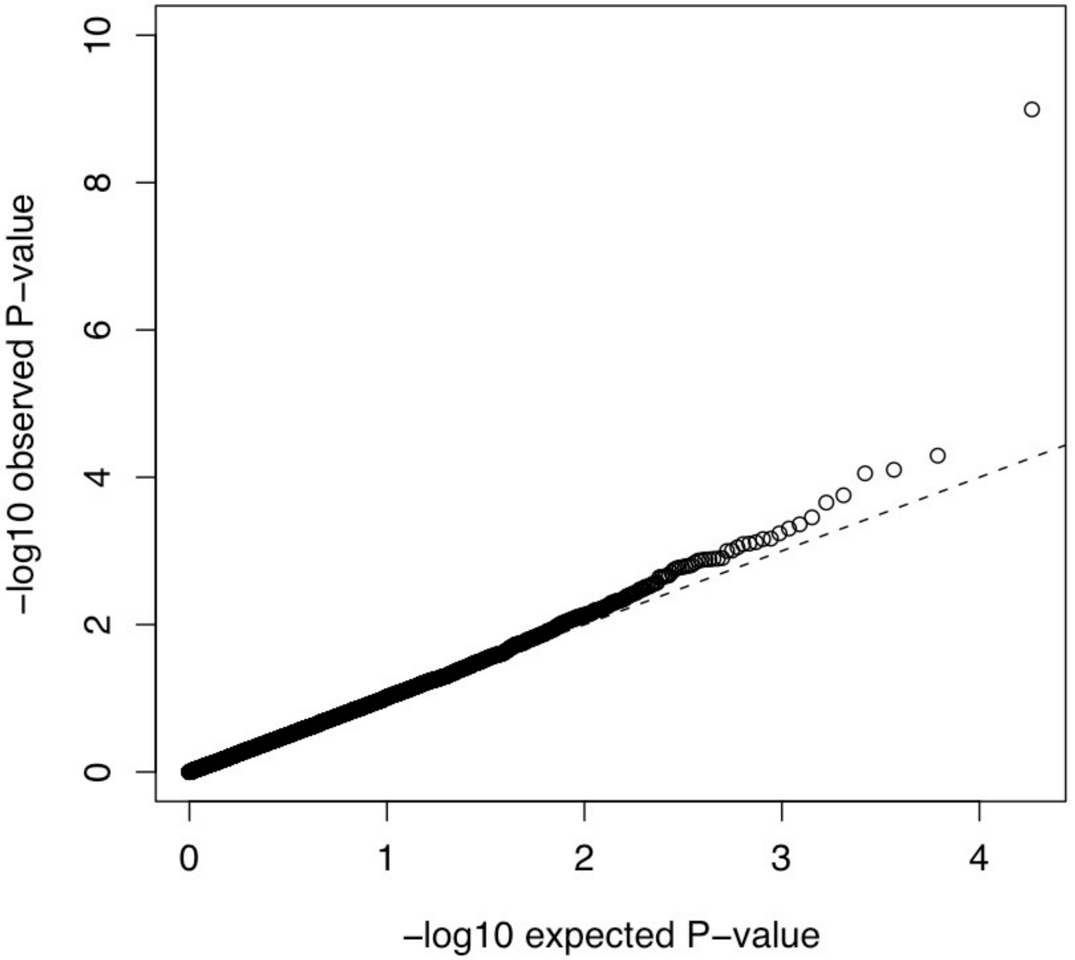
When all hypotheses are null (no association), the distribution of *p*-values is uniform (flat).

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...



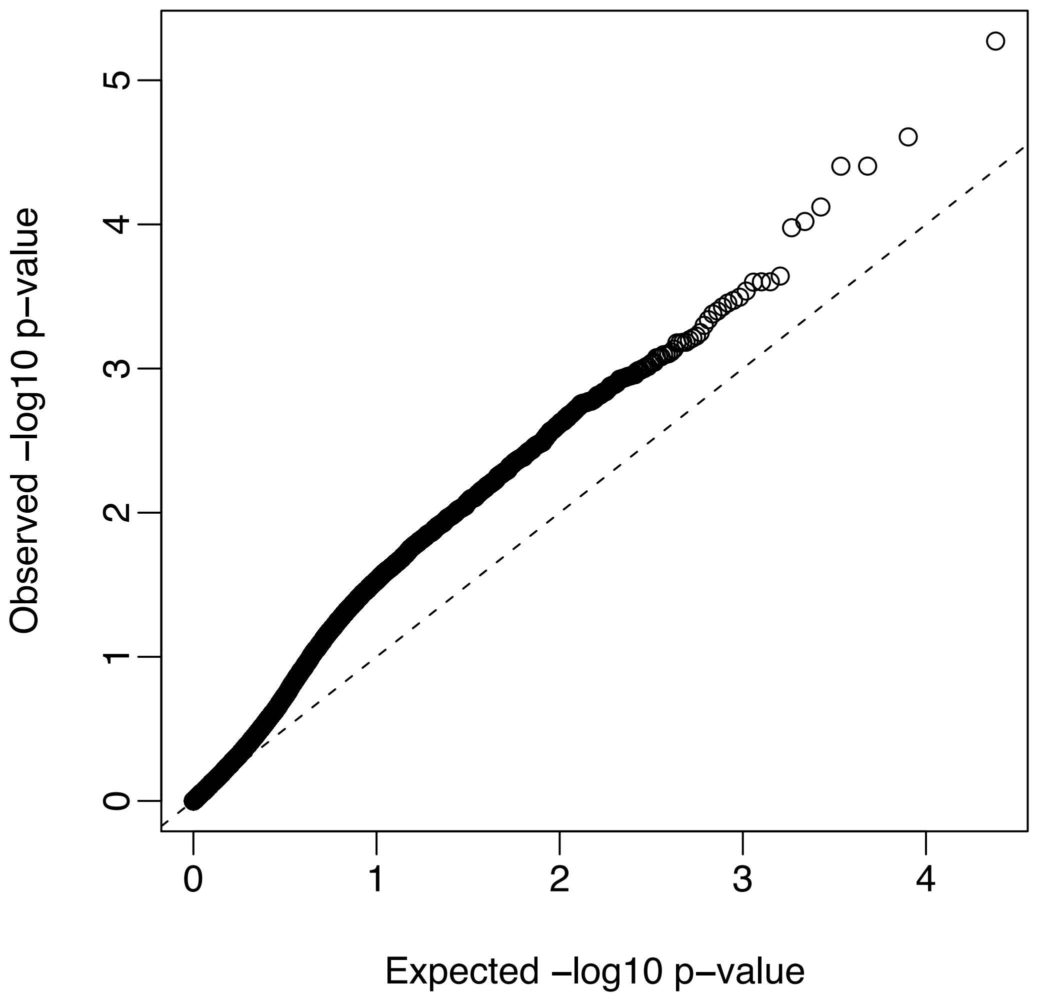


Plot the i^{th} smallest *p*-value against i/(M+1), where M is the number of markers.

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 λ > 1, then all test statistics are divided by λ before computing p-values.

Devlin, Roeder, 1999

Outline

Introduction

Single-variant association tests

- Genetic models
- Association tests for binary traits (case-control)
- Association tests for quantitative traits

GWAS Workflow

- Data quality control (QC)
- Multiple testing

Validation and replication strategies

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

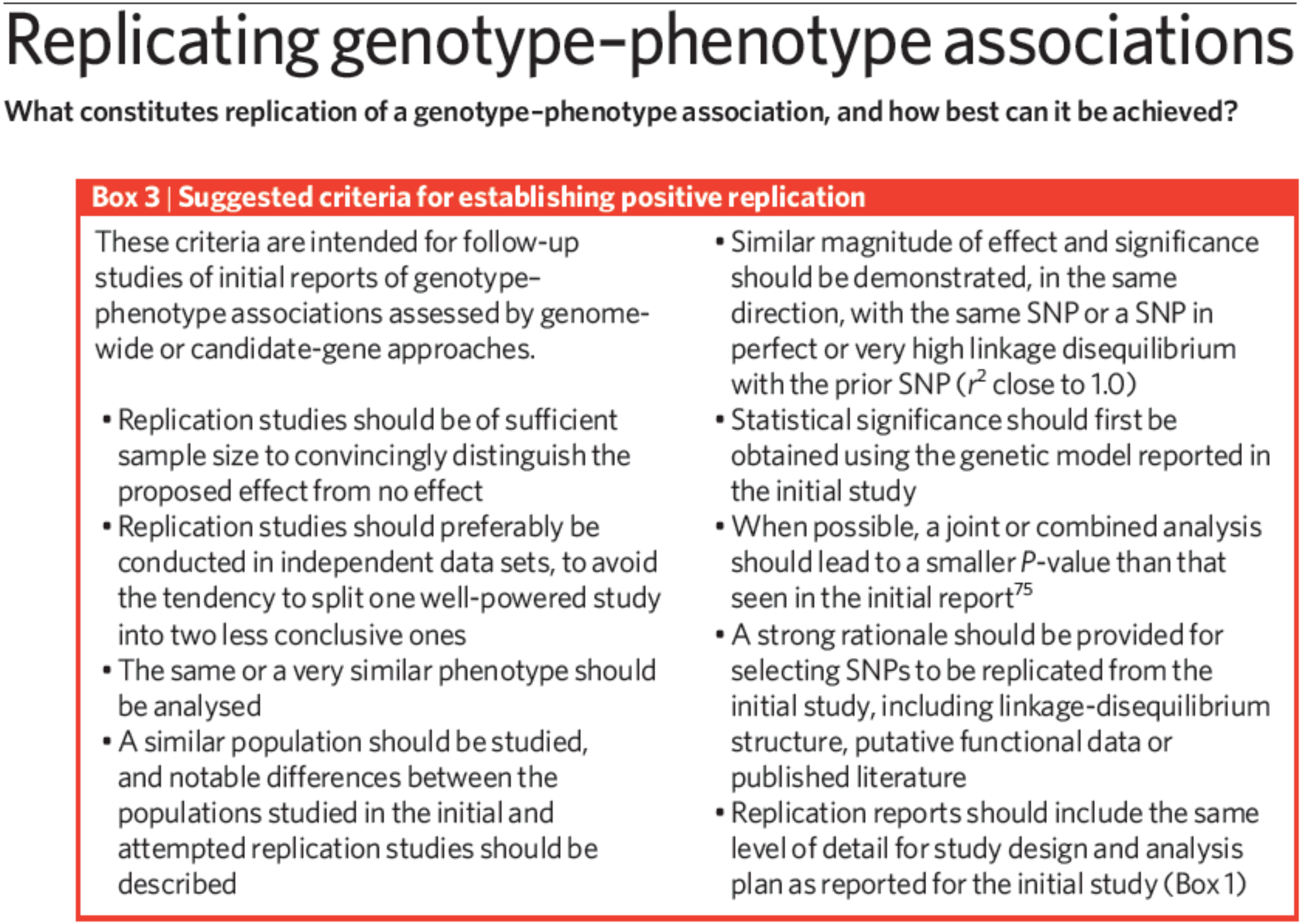
  - show generalizability to different ethnic populations

  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 in sequencing studies that do not rely on LD, but assess functional variants directly)
- To prove the causal effect, conduct functional studies:
  - Cell culture, model organisms
  - Does the variant affect gene expression, protein synthesis / transport / function

Web Resources

UCSC Genome Browser Gateway

http://genome.ucsc.edu

Software

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

http://www.broadinstitute.org/haploview/haploview

LocusZoom

http://locuszoom.sph.umich.edu/locuszoom/

(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.

Other Topics

- Meta-analysis
  - Used to combine association results from multiple studies to strengthen the evidence

- Analysis of imputed genotypes
  - Either in PLINK or using built-in capabilities/add-ons to the imputation packages, e.g., for minimac

    mach2dat
    mach2qtl
    http://www.unc.edu/~yunmli/software.html

Meta-analysis

- Hypothetical example:

	Sample size (n)	Effect of gene A (β)	P-value
Study 1	100	0.2	0.05

Meta-analysis

- Hypothetical example:

	Sample size (n)	Effect of gene A (β)	P-value
Study 1	100	0.2	0.05
Study 2	500	0.18	0.001

Meta-analysis

- Hypothetical example:

	Sample size (n)	Effect of gene A (β)	P-value
Study 1	100	0.2	0.05
Study 2	500	0.18	0.001
Study 3	250	0.21	0.01

- Meta-analysis - statistical method for combining the results of several studies that assigns an overall significance level

- combine *p*-values (Fisher's method): $$-2\sum_{i=1}^{k}\ln(p_i) \sim \chi^2_{2k}$$

- combine betas, SE's and sample size weights to form an overall Z-score

Implemented in METAL software package (Willer et al., Bioinformatics 2010)

Calculating Odds and Confidence Intervals

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

- Odds ratio for allele A vs a is given by

$$OR = (n_{11}/n_{10})/(n_{21}/n_{20}) = n_{11}n_{20}\,/n_{10}\,n_{21}$$

- In large samples, log(OR) is approximately normally distributed, with mean equal to log(OR) and estimated variance:

$$\mathrm{var}\left[\log(OR)\right] \approx \frac{1}{n_{10}} + \frac{1}{n_{11}} + \frac{1}{n_{20}} + \frac{1}{n_{21}}$$

- A 95% confidence interval for OR is given by $\exp^{\log(OR) \pm 1.96 \times SE}$, where SE = sqrt(var[log(OR)])