GWAS Approaches for Cohorts of Different Ancestries

IBS 746

11/09/2021

Jingjing Yang (jingjing.yang@emory.edu)

Outline

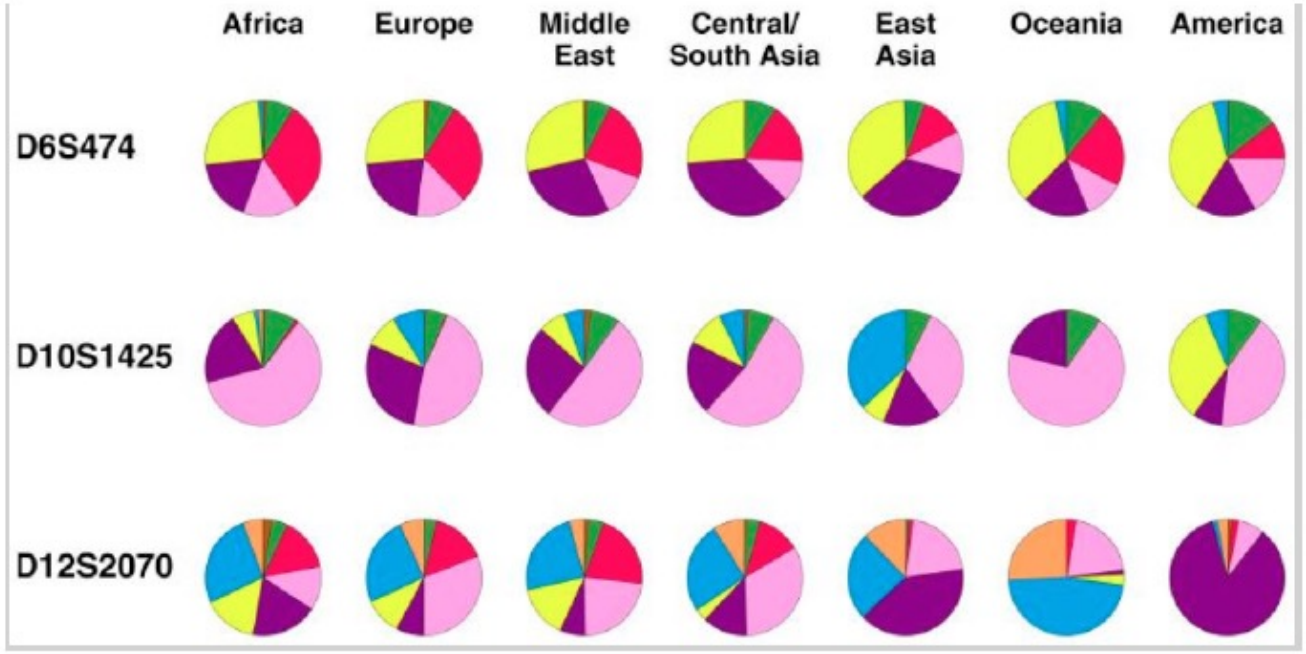
- Population Stratification
  - Genomic control factor
  - Principal components analysis
- Meta-analysis
  - Fisher’s method
  - Stouffer’s Z-score method
  - Inverse-variance method for fixed effect model
- Family-based Association Test

GWAS with Cohorts of Different Ancestries

- Cohorts with samples of European, Asian, African ancestries
- Possible problems for population-based association studies?
- How to resolve the issue?

Population Stratification

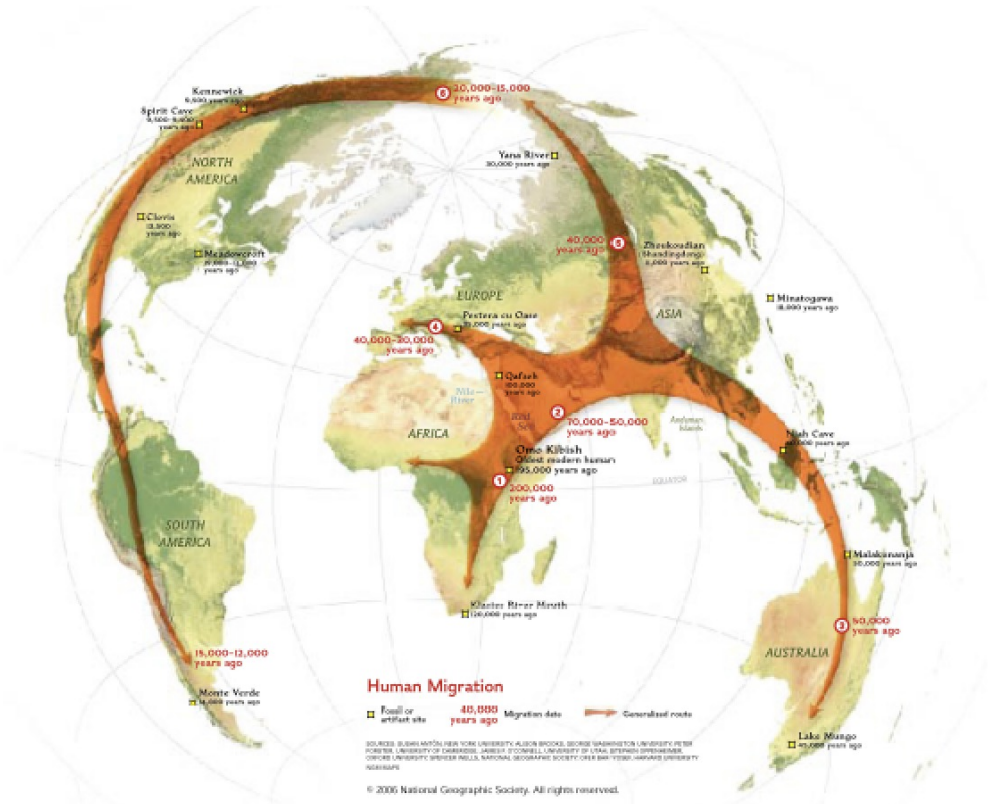
Population stratification (or population structure) is the presence of a systematic difference in allele frequencies between subpopulations, possibly due to different ancestry.



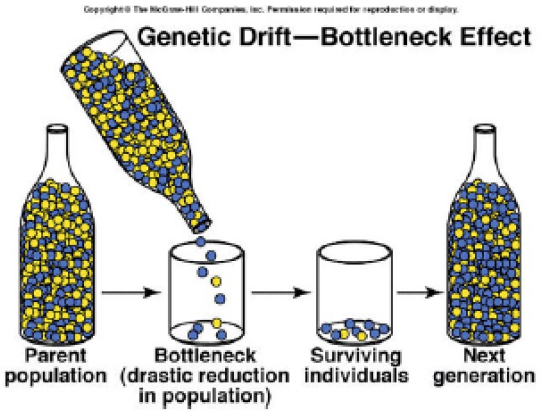
Allele frequencies at three microsatellite loci (Rosenberg N.A., Hum Biol. 2011). Each of the three loci has exactly eight alleles. In most of the pie charts, one or more alleles is rare or absent.

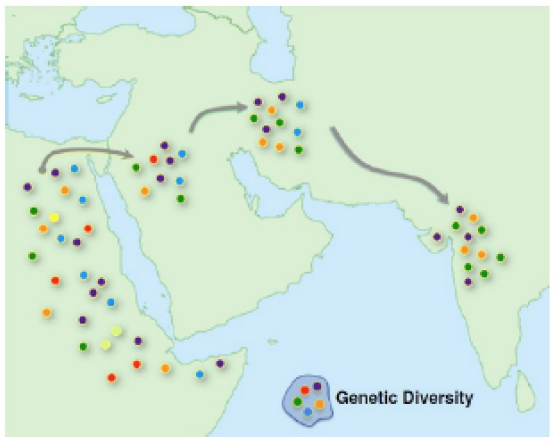
Causes of population structure

Human migration:



National Geographic





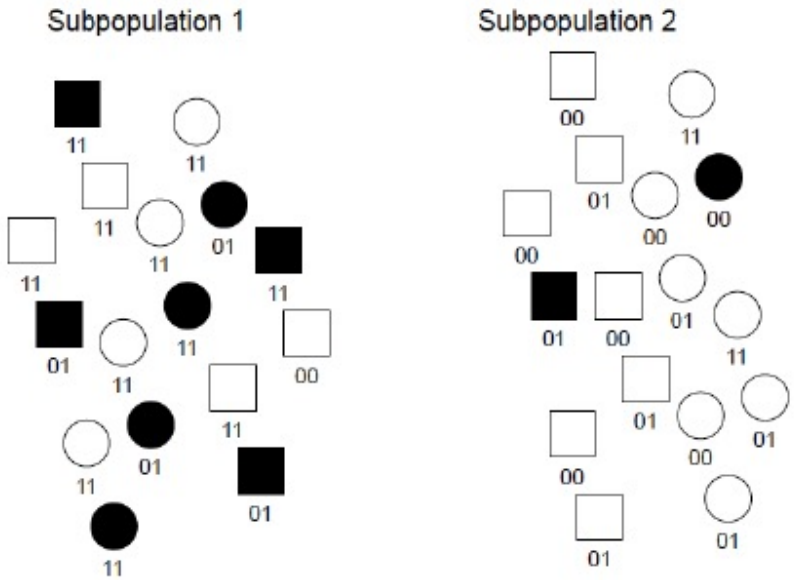
Henn *et al.* (2012) *PNAS*

Inflated False Positives

- Population-based association study methods assume samples are of the same ethnicity.
- The minor allele frequency of SNPs generally vary across different populations
- When the case/control ratio differs across different populations, instead of testing the association between the trait and genotype, you might end up testing the association between the ethnicity and genotype, leading to an inflated number of significant markers.

Example of False Positive Association

Consider genotypes (coded as 00, 01 and 11) at a marker locus



Subpopulation 1	1	0
Case	12	4
Control	14	2

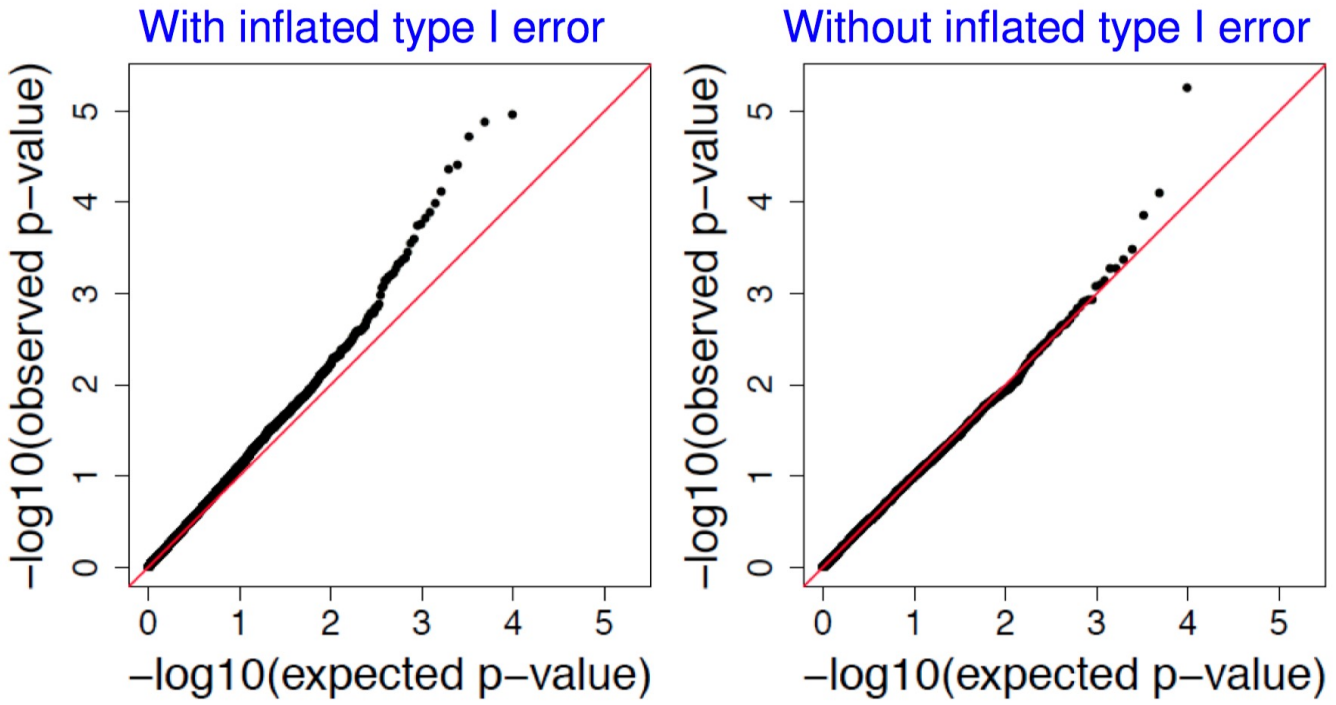
Subpopulation 2	1	0
Case	1	3
Control	10	18

Combined	1	0
Case	13	7
Control	24	20

A combined study tends to show association, even though there is no association within each subpopulation.

Check GWAS Results by Quantile-Quantile (QQ) Plot

– Obtained $-\log 10$(p-values) from GWAS
– Sort all $-\log 10$(p-values) from most significant to least
– Pair these with the expected values of order statistics of a Uniform$(0, 1)$ distribution
– Under NULL hypothesis (no association), p-values follow a Uniform$(0, 1)$ distribution



How to Address Population Stratification?

- Simplest Approach
  - Adjust false positives by Genomic Control Factor (not always work)
- Commonly Used Approach
  - Account for variables representing ethnicity information (Principal Components)
- Most Robust Approach: Stratify Multi-Ethnic Cohorts
  - Conduct association studies for samples of the same population/ethnicity
  - Combine association results by Meta-Analysis
- Most Effective Approach
  - Family-based Association Analysis
  - subject to the availability of data

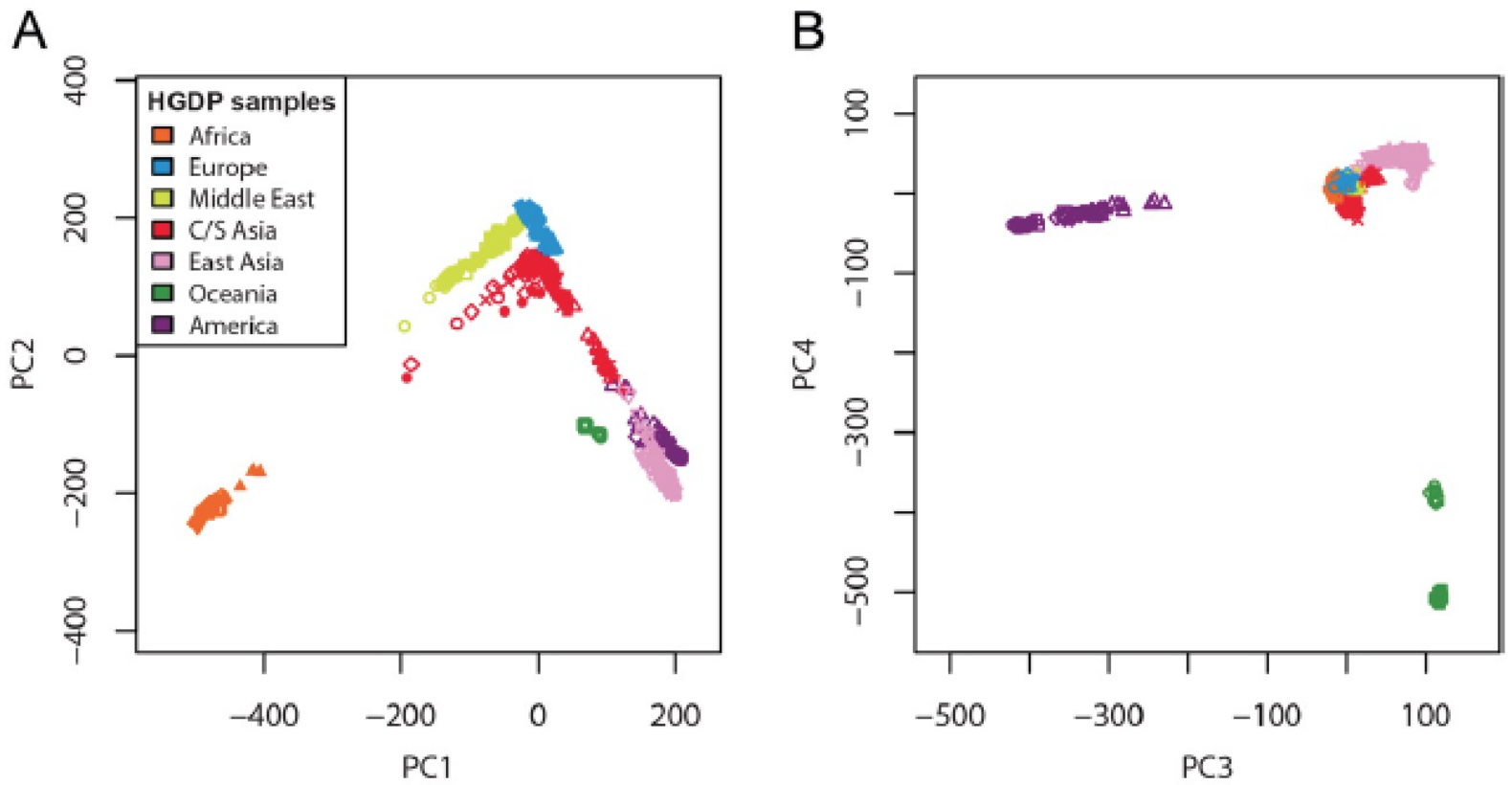
Genomic Control Factor

- Under null hypothesis (no association signal exists), p-values should follow a uniform distribution within (0, 1)
- Median p-value = 0.5 under null hypothesis, corresponding to chi-square statistic (df=1) value 0.456
- Find the actual median p-value from your GWAS, with corresponding chi-square statistic (df=1) value $\mathrm{median}(\chi^2)$
- Genomic Control Factor: $\lambda_{GC} = median(\chi^2)/0.456$
- Adjust your GWAS results by λ_{GC}
  - Scale your chi-square statistic test statistics (df=1) by λ_{GC}
  - Recalculate the corresponding p-values

Principal Components Analysis (PCA)

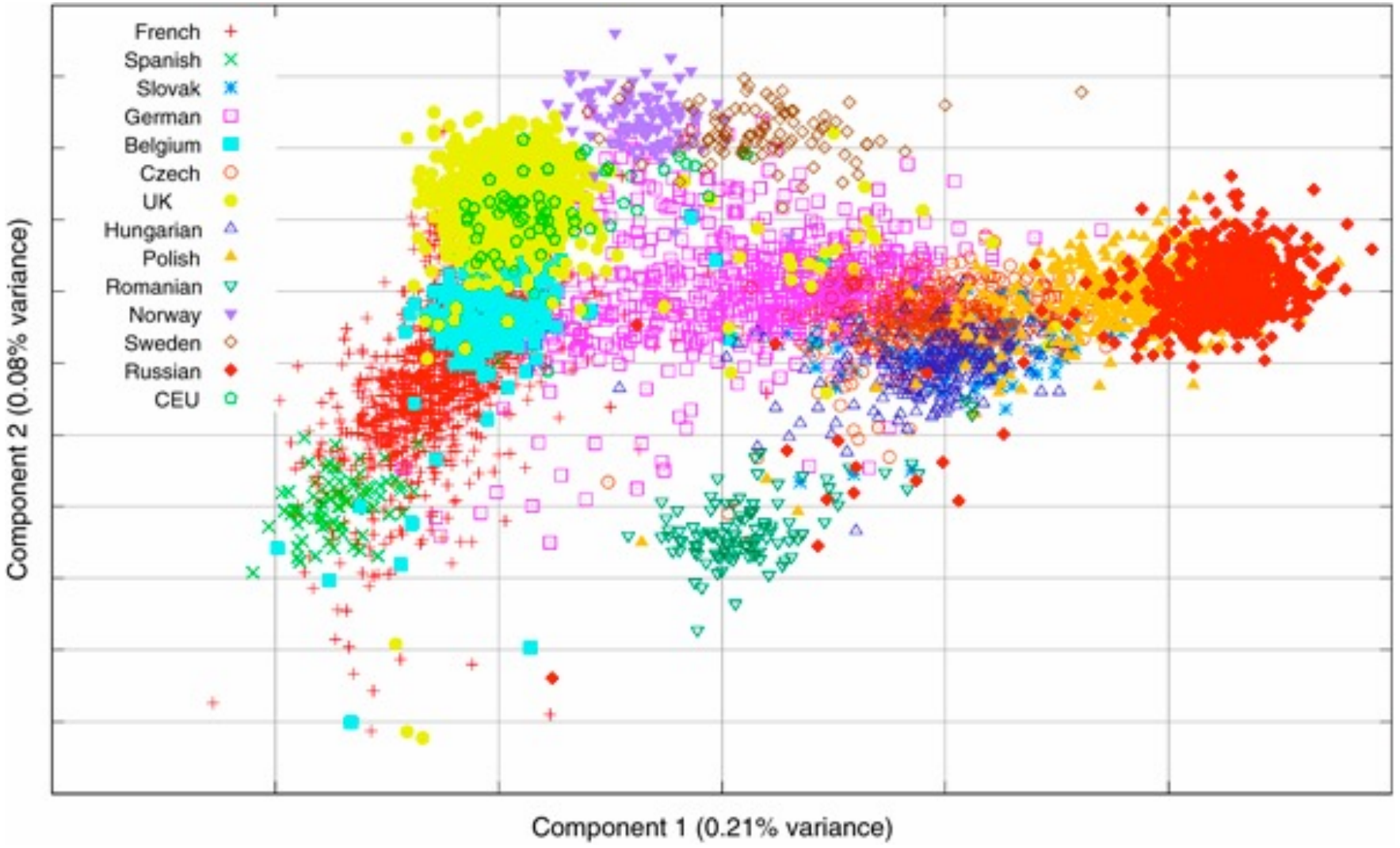
- Consider genotype matrix $X_{n\times \mathrm{p}}$, with n individuals and p genome-wide SNPs
- Principal Components Analysis (PCA) with respect to $X_{n\times \mathrm{p}}$
  - Center columns in $X_{n\times \mathrm{p}}$
  - PCA project original genotype data matrix to a new coordinate system such that the PC1 explains the most data variance, and then PC2, …
    - Calculate a set of loading vectors (w_k, length p, k=1, 2, …) for PC1, PC2, …
    - Principle components (PCs) are given by: Xw_k
    - Generally, plotting PC1 vs. PC2 will give a good visualization of sample ancestries
  - R function: prcomp() ; https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/prcomp
  - PLINK

PCA Visualization



Li et al. Science. 2008; Jakobsson et al. Nature. 2008.

First two principal components among European subjects



Heath et al. 2008

Adjust for Top PCs in Regression Model Based Tests

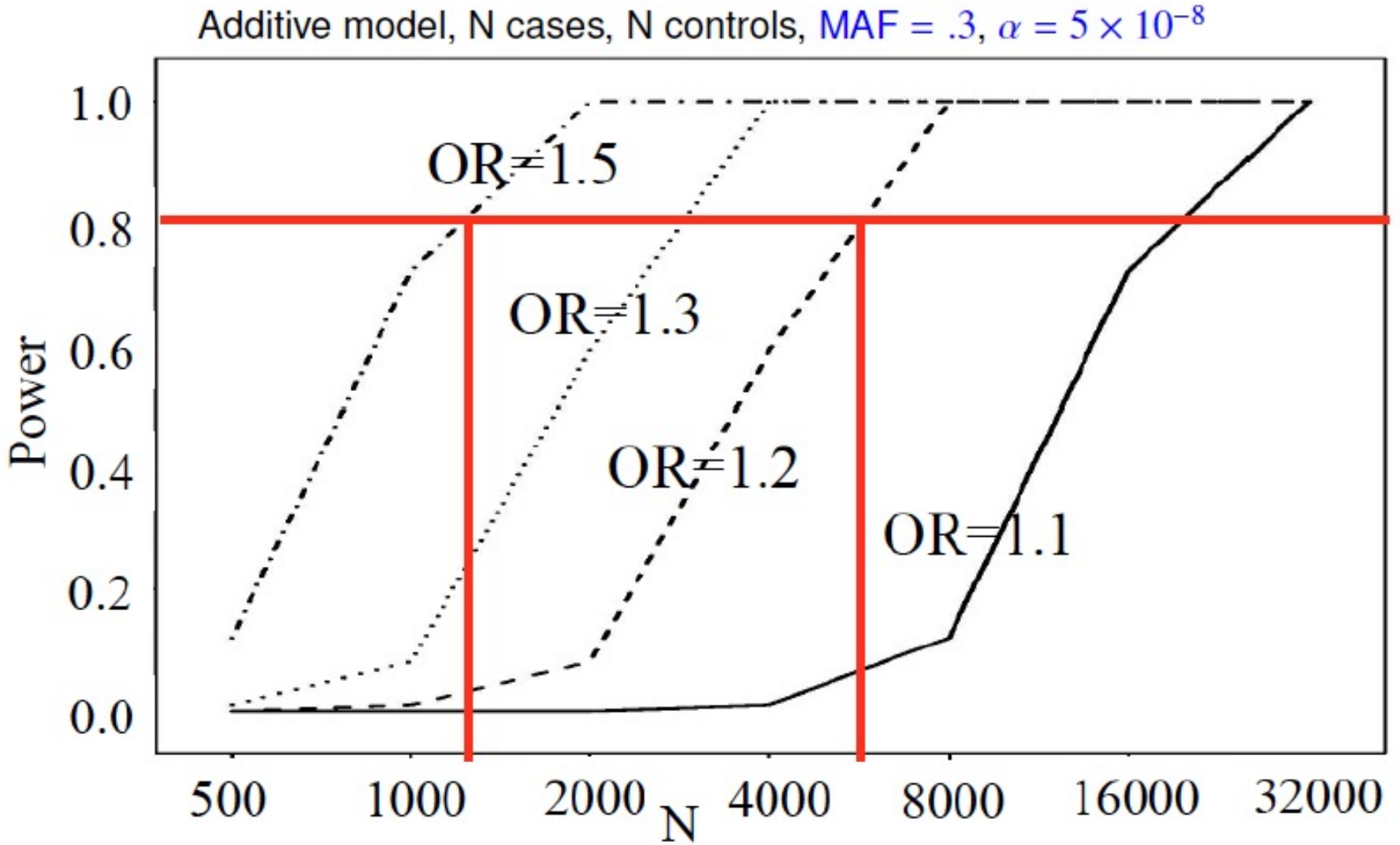
- Adjust for the population structure in your study

- Generally, include PC1-5 as confounding covariates (C) in your regression model
  - $\log\left(\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right) = \beta_0 + \alpha C + \beta_1 X$
  - $Y = \beta_0 + \alpha C + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

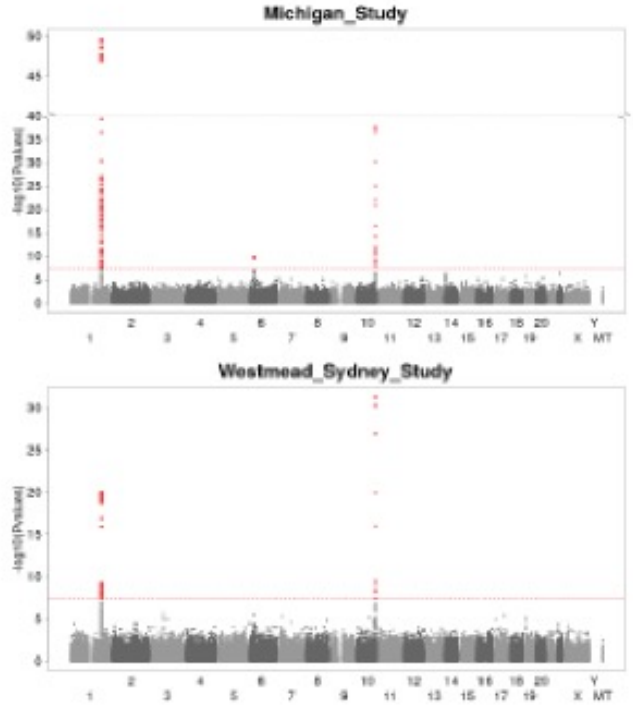
Meta-analysis

- Combine results across multiple studies for the same phenotype
- Improve power for the larger total sample size
- Address between study variances (due to population stratification, study design)
- Avoid the hassle of sharing individual-level genotype/phenotype/covariate data
- It is theoretically shown that the meta-analysis results is equivalent to the joint analysis with individual-level data under idea situation
  - Same phenotype and covariates
  - No population stratification
  - Balanced case-control study

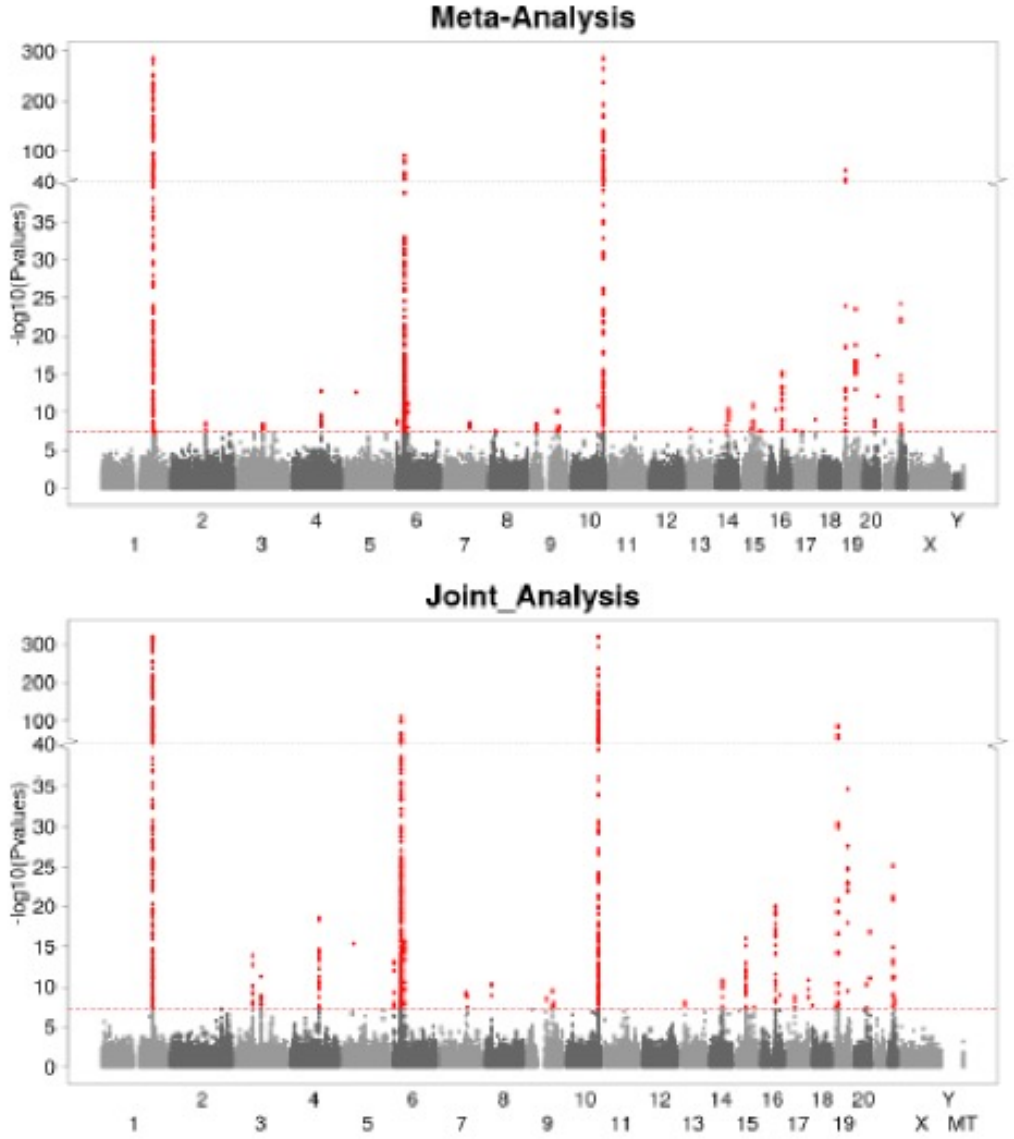
Improve Power with Larger Total Sample Size



Improve Power with Larger Total Sample Size



Example two individual studies of AMD.



Meta-analysis Methods

- Fisher’s Method: combining p-values
- Stouffer’s Z-score Method
- Inverse-variance method for fixed effect model

Fisher's Method

- Consider the following summary statistics from K studies for testing the association between the same SNP and the same (type) phenotype
  - p-values $(p_1, p_2, \ldots, p_K)$
- Test statistic for meta-analysis
  - $\mathrm{X}^2 = -2\sum_{i=1}^{K}\log(p_i) \sim$ Chi-square distribution with df=2K under H_0

Stouffer's Z-score Method

- Consider a series of summary statistics from K studies for testing the association between the same SNP and the same (type) phenotype
  - p-values $(p_1, p_2, \dots, p_K)$
  - Effect-sizes $(\beta_1, \beta_2, \dots, \beta_K)$
  - Sample sizes $(n_1, n_2, \dots, n_K)$
- Invert each p-value to a Z-score statistic:
  - $Z_k = \Phi^{-1}\left(1 - \frac{p_k}{2}\right) * sign(\beta_k)$
  - Φ is the standard normal cumulative density function
- Test statistic (weight by sample sizes) for meta-analysis
  - $Z_{meta} = \frac{\sum_{k=1}^{K} Z_k w_k}{\sqrt{\sum_{k=1}^{K} w_k^2}} \sim N(0,1)$ under H_0
  - $w_k = \sqrt{n_k}$

Inverse-variance method for fixed effect model

- Consider the following summary statistics from K studies for testing the association between the same SNP and the same (type) phenotype
  - Effect-sizes $(\beta_1, \beta_2, \ldots, \beta_K)$
  - Variance of effect-sizes $(v_1, v_2, \ldots, v_K)$
- Test statistic (Inverse-variance weighting) for meta-analysis
  - $\beta_{meta} = \frac{\sum_{k=1}^{K} w_k \beta_k}{\sum_{k=1}^{K} w_k}, \quad w_k = 1/v_k$
  - $Var(\beta_{meta}) = \frac{1}{\sum_{i=1}^{K} w_i}$
  - Wald Test Statistic: $\frac{\beta_{meta}}{\sqrt{Var(\beta_{meta})}} \sim N(0,1)$ under H_0

Table 3 | **Summary of methods for meta-analysis of genome-wide data**

Method	Description	Advantages	Disadvantages	Main software used
P value meta-analysis	Simplest meta-analytical approach	Allows meta-analysis when effects are not available	Direction of effect is not always available; inability to provide effect sizes; difficulties in interpretation	METAL, GWAMA, R packages
Fixed effects	Synthesis of effect sizes. Between-study variance is assumed to be zero	Effects readily available through specialized software	Results may be biased if a large amount of heterogeneity exists	METAL, GWAMA, R packages
Random effects	Synthesis of effect sizes. Assumes that the individual studies estimate different effects	Generalizability of results	Power deserts in discovery efforts; may yield spuriously large summary effect estimates when there are selection biases	GWAMA, R packages
Bayesian approach	Incorporates prior assessment of the genetic effects	Most direct method for interpretation of results as posterior probabilities given the observed data	Methodologically challenging; GWAS-tailored routine software not available; subjective prior information used	R packages
Multivariate approaches	Incorporates the possible correlation between outcomes or genetic variants	Increased power can identify variants that conventional meta-analysis do not reveal using the same data sets	Computationally intensive; software not available for all analyses; some may require individual-level data	GCTA for multi-locus approaches
Other extensions	A set of different approaches that allows for the identification of multiple variants across different diseases	Summary results of previous meta-analyses can be used	May need additional exploratory analyses for the identification of variants; prone to systematic biases	Software developed by the authors of the proposed methodologies

GCTA, genome-wide complex trait analysis; GWAS, genome-wide association study.

Evangelou, E. and Ioannidis, J. P.A. Nature Reviews

Table 1 | **Examples of high-profile consortia for various disease phenotypes**

Consortium (acronym)	Phenotype (or phenotypes)	Publicly available genome-wide data?	Website
AMD	Age-related macular degeneration	Yes, accessible through the website	http://www.sph.umich.edu/csg/abecasis/public/amdgene2012
BCAC	Breast cancer	No	http://ccge.medschl.cam.ac.uk/consortia/bcac
CHARGE	Heart disease and ageing	No	http://web.chargeconsortium.com
GEFOS	Osteoporosis	Yes, accessible through the website	http://www.gefos.org
GIANT	Anthropometric traits	Yes, accessible through the website	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium
GLGC	TC, HDL-C, LDL-C, triglycerides	Yes, accessible through the website	http://www.sph.umich.edu/csg/abecasis/public/lipids2010
IIBDGC	Inflammatory bowel disease	Yes, accessible through the website	http://www.ibdgenetics.org
IMSGC	Multiple sclerosis	Yes, accessible through the website	https://www.imsgenetics.org/
ISC	Schizophrenia	No	http://pngu.mgh.harvard.edu/isc
MAGIC	Glycaemic traits	Yes, accessible through the website	http://www.magicinvestigators.org
NARAC-III	Rheumatoid arthritis	No	http://www.naracstudy.org/NaracStudy/narac.aspx
TREATOA	Osteoarthritis	Yes, accessible through the website	http://treatoa.eu
WTCCC	Various phenotypes	Yes, accessible through the website	http://www.wtccc.org.uk

HDL-C: high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol.

Evangelou, E. and Ioannidis, J. P.A.
Nature Reviews

Study Design for Meta-analysis

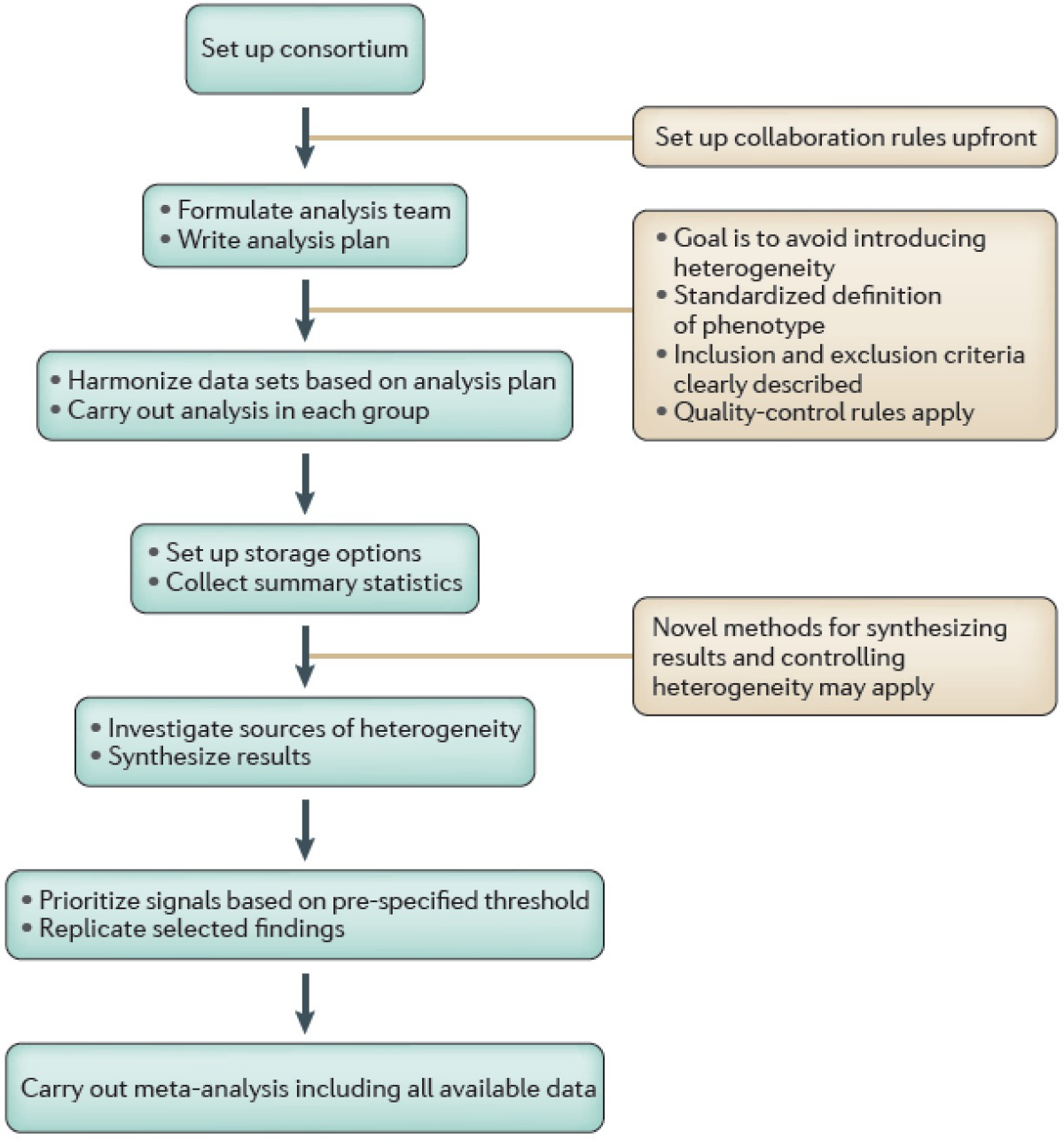
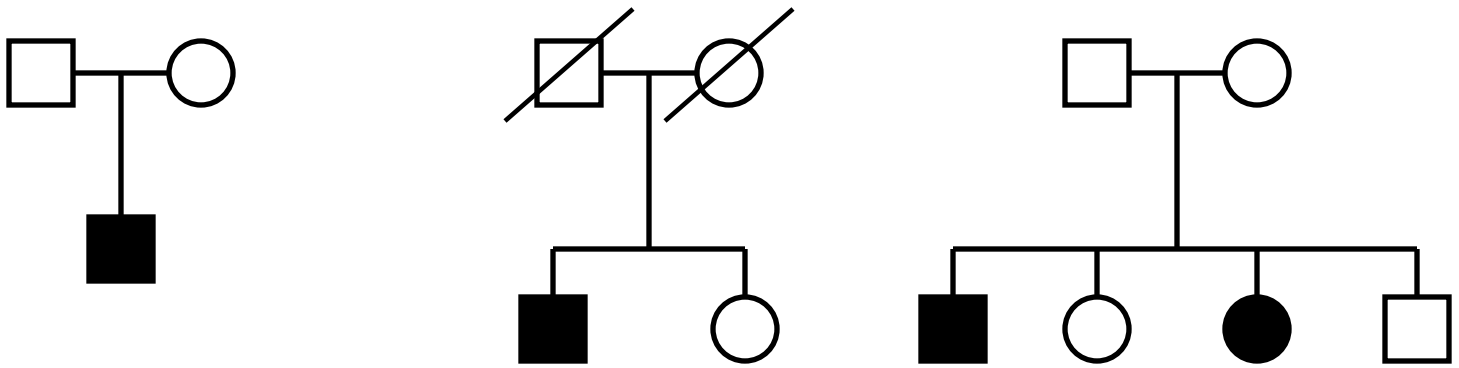


Figure 1 | **Stages in a meta-analysis.** A typical plan for a meta-analysis of genome-wide and next-generation sequence data.

Evangelou, E. and Ioannidis, J. P.A. Nature Reviews

Family Based Association Study

- Cases vs within-family (related) controls
- Under H_0, an affected child is equally likely to inherit either allele at the tested marker



Transmission Disequilibrium Test (TDT)

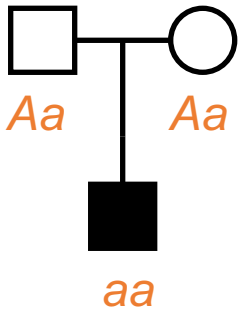
- Considers the case-parent triad: affected child (proband) and parents (heterozygous)
- Rational: Compare the distribution of alleles transmitted to the affected child to the distribution of the non-transmitted allele
  - Under the null hypothesis (H_0), the heterozygous patent with genotype Aa will be equally likely (Mendel's Law) to transmit A and a to the affected child
  - Under the alternative hypothesis (H_a), the heterozygous patent is more likely transmit the disease allele *a* to the affected child
- Developed in 1993 by Spielman et al. (AJHG)

Key Points of TDT

- Consider 2 alleles (A, a) at the marker locus
- Considers the case-parent triad: affected child (proband) and parents (heterozygous)
- Only heterozygous parents (Aa) will be used in the test
- Transmitted alleles (transmitted from parent to the affected child)
- Non-transmitted alleles (not transmitted from parent to the affected child)
- Transmitted alleles are matched with non-transmitted alleles (in heterozygous parents)

TDT Scoring

- Consider 1 family, 2 heterozygous parents, one affected child
- Count per heterozygous parent



	Non-transmitted A	Non-transmitted a
Transmitted A	0	0
Transmitted a	2	0

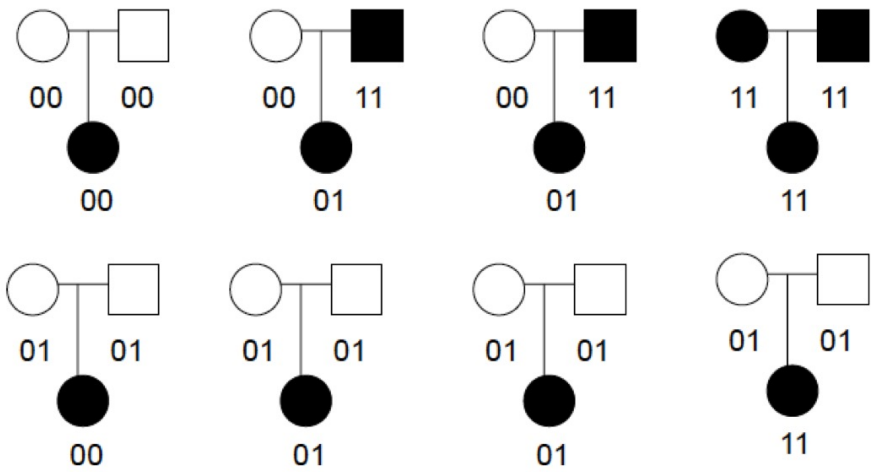
TDT Scoring

- Consider N triad families with single affected child; 2N parents

	Non-transmitted A	Non-transmitted a
Transmitted A	N_1	N_2
Transmitted a	N_3	N_4

- McNemar's test statistic: $X^2 = \frac{(N_2 - N_3)^2}{(N_2 + N_3)}$ ~ Chi-square distribution with df=1 under H_0

TDT Example: Fail to Reject Null



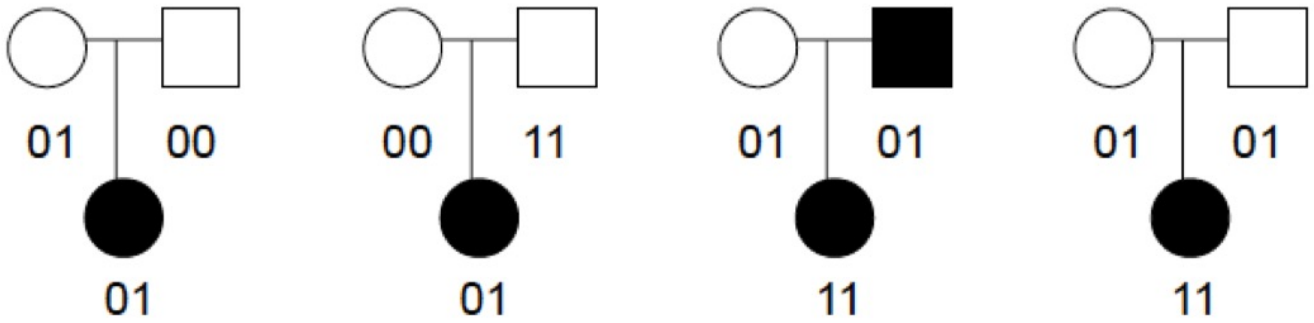
- In the parental population, affecteds are always 11, suggesting association
- TDT test:

		Transmitted Allele	
		1	0
Not Transmitted Allele	1	4	4
	0	4	4

$X^2 = 0$

- Affected children are not more likely to inherit allele 1 than allele 0; the parental population exhibits spurious association.

TDT Example: Reject Null



- TDT test:

		Transmitted Allele	
		1	0
Not Transmitted Allele	1	1	0
	0	5	2

- There is true association - p = .025

TDT Hypothesis

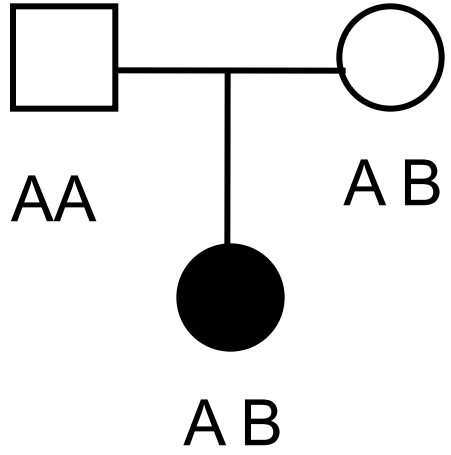
- Three possible null hypotheses
  - Associated but not linked with disease locus
    - Follow-up analysis for population-based association signals
    - Indicates population stratification problem
  - Linked but not associated
    - Follow-up analysis for linkage signals
  - Neither linked nor associated
    - Candidate gene studies
    - Population-based GWAS
- One alternative hypothesis
  - Marker is linked and associated with a disease-susceptibility locus (DSL)

TDT vs. Population-based Case Control Studies

- Advantages of TDT
  - Robust to population stratification: “Matched” case and control alleles (conditioning on heterozygous parent genotypes)
  - Robust to potential misspecification of the disease models
- Disadvantages
  - Highly sensitive to genotype error
    - Genotype error in TDT can cause large biases
  - Data needed on both parents
    - Missing parents can lead to bias if handled improperly
    - Can be difficult except with early-onset diseases

Bias due to genotype error

True genotypes:
"A, B" transmitted
"A, A" un-transmitted

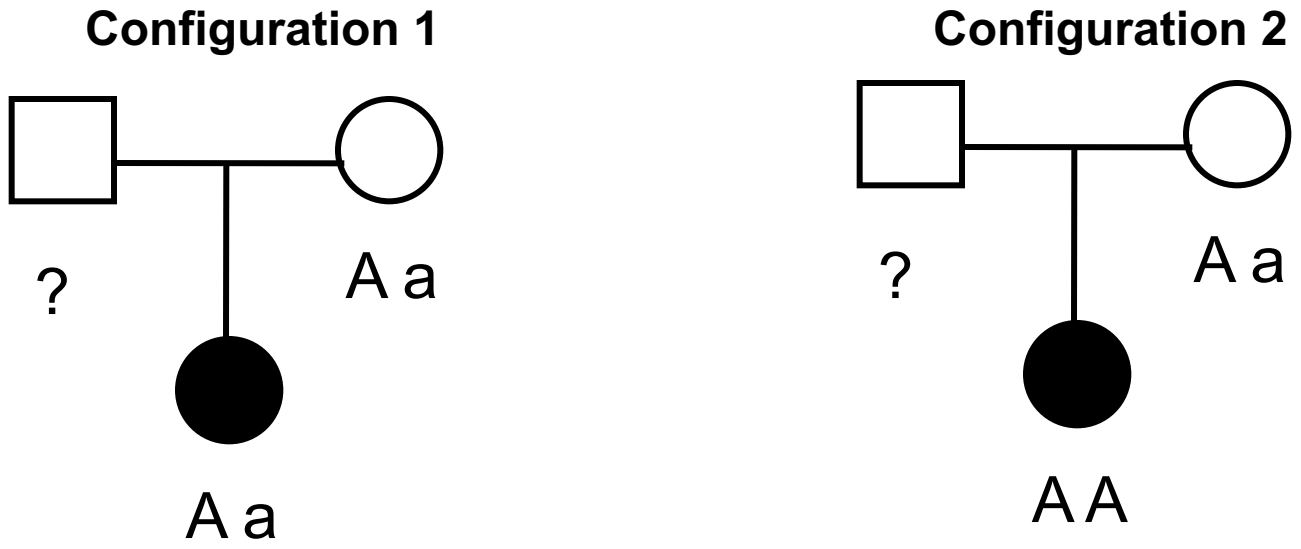


Type of error	Consequence
Proband -> BB	Mendelian error, family dropped
Proband -> AA	A,A transmitted, A,B untransmitted
Mother -> BB	A,B transmitted, A,B untransmitted
Mother -> AA	Mendelian error, family dropped
Father -> AB	A,B transmitted, A,B untransmitted
Father -> BB	A,B transmitted, B,B untransmitted

- Can observe apparent over-transmission of major allele if:
  - Undetected genotype error in general [1]
  - Genotype error rate is greater for heterozygotes [1]
  - Missingness rate is greater for heterozygotes [2,3]

[1] Mitchell, Cutler, and Chakravarti 2003 AJHG 72:598-610
[2] Hirschhorn and Daly 2005 Nat Rev Genet 6:95-108
[3] Hao and Cawley 2007 Hum Hered 63:219-228

Potential bias due to missing parents



- Under H_0, *Aa* parent equally likely to transmit *A* or *a* allele to affected offspring
- But if configuration 1 excluded & configuration 2 included, then appears that *A* is transmitted more than *a*
- Leads to increased number of false positives

Available Tools

- PLINK : QC, PCA of genotype data, GWAS
  - https://www.cog-genomics.org/plink/
- METAL : meta-analysis tool
  - https://genome.sph.umich.edu/wiki/METAL_Documentation
- Association and TDT tool
  - https://www.soph.uab.edu/ssg/linkage/associationtdt