Genome-wide Association Studies

BIOS 770

02/08/2022

Jingjing Yang (jingjing.yang@emory.edu)

Outline

- Quality Control
  - Genotype Quality Control
  - Sample Relatedness: Kingship Coefficient
- Population Stratification
  - Genomic Control Factor
  - Genotype Principal Components Analysis
  - Meta-analysis
- Linear Mixed Model (LMM)
- Heritability Estimation by REML

GWAS Quality Control

- Filter SNPs
  - Marker genotyping missing rate (e.g., $> 2\%$)
  - Mapping quality for sequence data (based on mapping quality scores)
  - Hardy-Weinberg Equilibrium (HWE) Testing (e.g., p-value $< 10^{-6}$)
  - MAF (e.g., $< 5\%$)
  - Control sample reproducibility
  - Mendelian Errors (e.g., $> 1\%$ families, or > 5 errors) for family-based studies
- Filter samples
  - Sex inconsistencies and chromosomal anomalies
  - Relatedness for population-based studies (how to quantify relatedness given genotype data?)
  - Ethnicity
  - Sample genotyping efficiency/call rate (e.g., $< 98\%$)

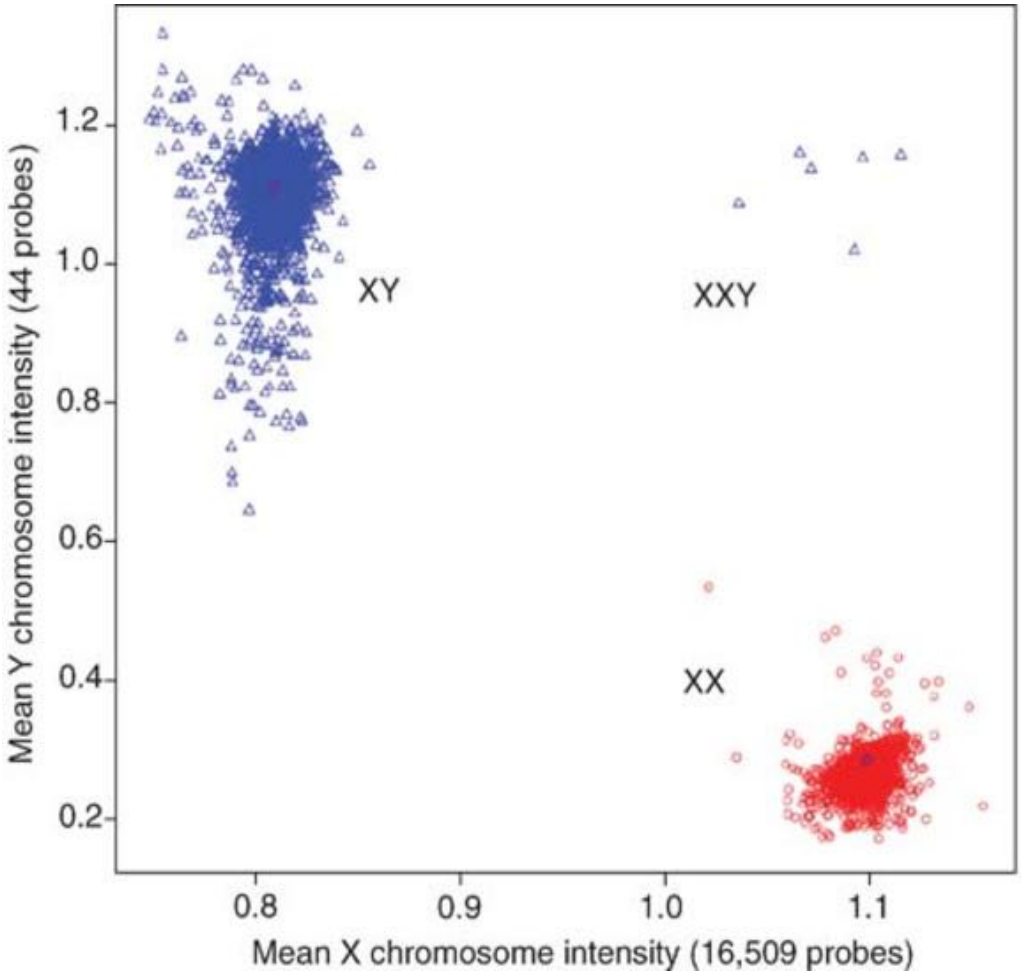
Genotype Quality

Data quality is one of the key factors affecting the validity of findings.

Example factors affecting genotype quality:

- Quality of DNA samples, depending on the sample source (e.g., blood, buccal swab, spit kit)
- Handling and storage of the sample (e.g., sample contamination)
- Genotyping platforms/chips
- Sequence errors
- Variant calling

Genotype Quality Control : Sex consistence



Visualization of X and Y probe intensities. The *x*-axis and *y*-axis represent the sum of the average over all probes for the normalized Cartesian intensity for allele A and the average over all probes for the normalized Cartesian intensity for allele B using all probes available on the X chromosome and Y chromosome, respectively. The XX (female, red circles) and XY (male, blue triangles) subjects are shown on the bottom right corner and on the top left corner, respectively. The plot reveals two mislabeled individuals (one male with the female cluster, and one female with the male cluster). Several XXY individuals are also clearly visible (upper right corner).

S. Turner et. al. CP hum Genetics. 2011.
https://doi.org/10.1002/0471142905.hg0119s68

Kinship QC

- Sample relationship checking
- Pedigree error checking

IBD vs. IBS

Let's ignore disease phenotypes and only consider the similarity of marker alleles.

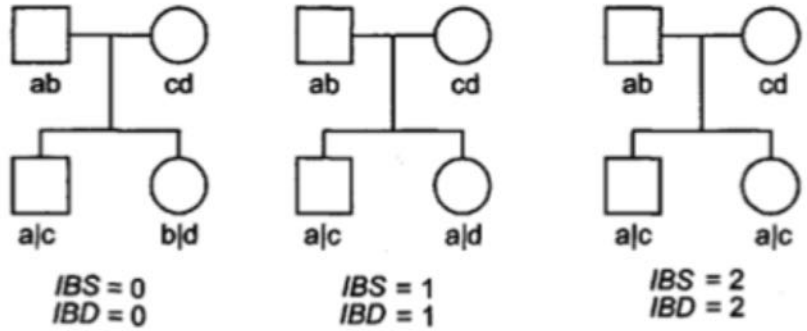
Identical/Identity by Descent (IBD): Two alleles are IBD if they are the same physical copy.

– E.g., two siblings may inherit the same paternal allele from their father.

Identical/Identity by State (IBS): Two alleles are IBS if they are the same type of allele.

IBD → IBS

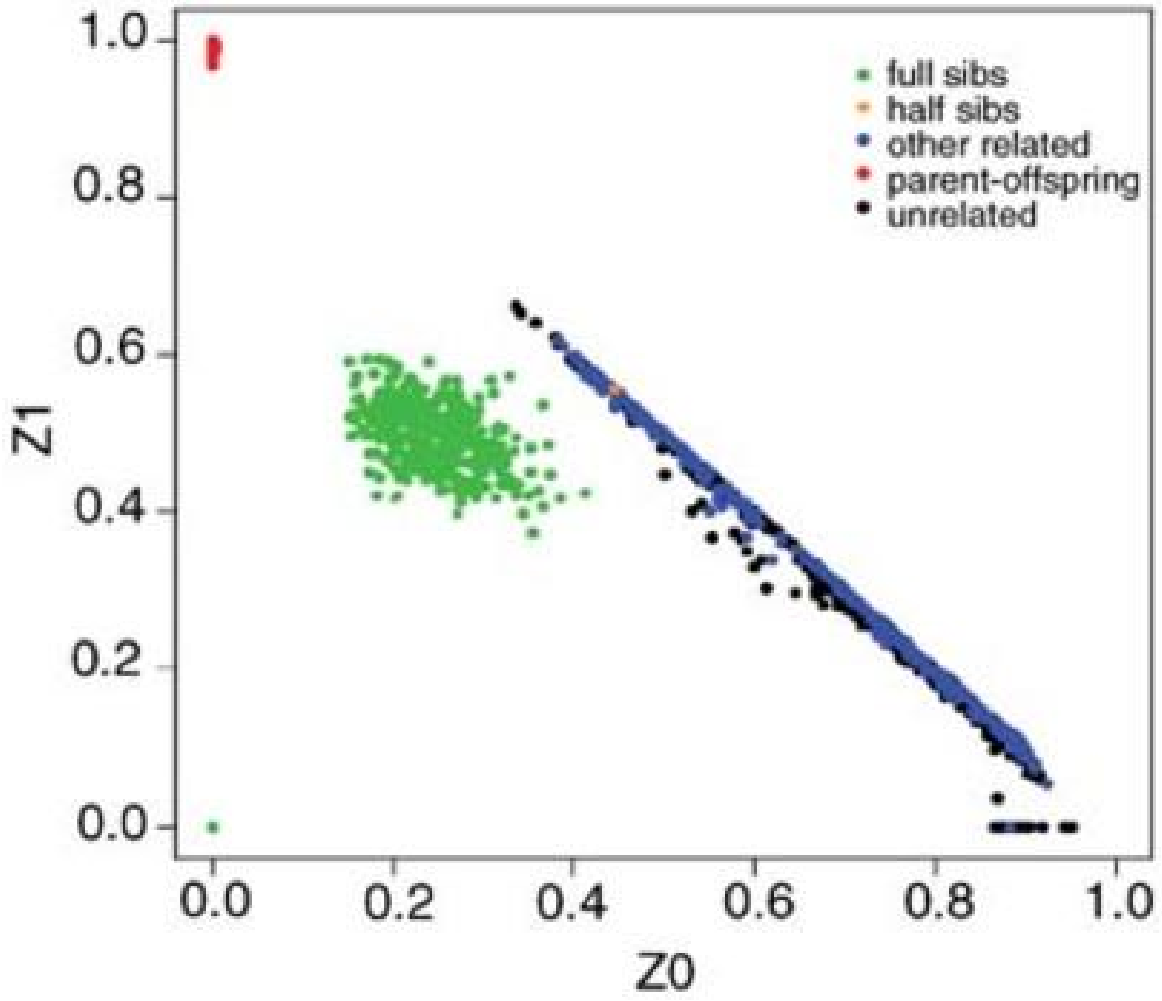
IBS ↛ IBD



Four alleles in parents ⇒ unambiguous IBD

Type of Relative Pair	Probability of Shared IBD π_0	π_1	π_2	Expected IBD Sharing
Monozygotic twins	0	0	1	2
Full sibs	1/4	1/2	1/4	1
Parent-offspring	0	1	0	1
First cousins	3/4	1/4	0	1/4
Double first cousins	13/16	1/8	1/16	1/4
Grandparent-grandchild, half-sibs, avuncular	1/2	1/2	0	1/2

Sample Relatedness (Z0 and Z1 here are π_0 and π_1)



Points in this plot show pairs of individuals plotted by their degree of relatedness: the proportion of loci where the pair shares one allele IBD (Z1) by the proportion of loci where the pair shares zero alleles IBD (Z0). These values are obtained from PLINK using the –genome option. Pairs are color-coded by the type of relationship determined by the pedigree information embedded in the pedfile (also reported by PLINK). This plot omits pairs of individuals having an overall kinship coefficient ≥ 0.05 for clarity. There is a pair of monozygotic twins represented by a point in the lower left at (0,0), because they share two alleles IBD at every locus across the genome.

S. Turner et. al. CP hum Genetics. 2011.
https://doi.org/10.1002/0471142905.hg0119s68

Kinship Coefficient ϕ

- ϕ : The probability that two alleles sampled at random from two individuals (one allele per sample) are Identical by Descent (IBD).
- $\pi_{0;i,j}, \pi_{1;i,j}, \pi_{2;i,j}$ denote the probability that two individuals (i, j) share 0, 1, and 2 IBD.
- $2\phi_{i,j} = \frac{\pi_{1;i,j}}{2} + \pi_{2;i,j}$

Table 1. Relationship inference criteria based on estimating kinship coefficients (ϕ) and probability of zero IBD sharing (π_0)

Relationship	ϕ	Inference criteria	π_0	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{3/2}}$	0	<0.1
Parent–offspring	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	0	<0.1
Full sib	$\frac{1}{4}$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$	$\frac{1}{4}$	(0.1,0.365)
2nd Degree	$\frac{1}{8}$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$	$\frac{1}{2}$	$(0.365, 1-\frac{1}{2^{3/2}})$
3rd Degree	$\frac{1}{16}$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$	$\frac{3}{4}$	$(1-\frac{1}{2^{3/2}}, 1-\frac{1}{2^{5/2}})$
Unrelated	0	$< \frac{1}{2^{9/2}}$	1	$> 1-\frac{1}{2^{5/2}}$

Estimate Kinship Coefficient ϕ by PLINK

- Assume HWE and homogeneous population
- Reference allele (denoted by A) frequency p
- $IBS_{ij}, IBD_{i,j}$ denotes the IBS, IBD between two individuals (i, j)
- Method of Moments

$$Pr(IBS_{ij} = k) = \sum_{z=0,1,2} \Pr(IBS_{ij} = k | IBD_{i,j} = z)\, \pi_{z;i,j}$$

$$\Pr(IBS_{ij}=0)=\Pr(AA,aa|\mathrm{IBD}_{ij}=0)\cdot\Pr(\mathrm{IBD}_{ij}=0)=2p^2(1-p)^2\pi_{0ij} \quad (1)$$

This leads to the estimator

$$\hat{\pi}_{0ij}=\frac{\sum_m I_{\mathrm{IBS}_{ij}^m=0}}{\sum_m 2\hat{p}_m^2(1-\hat{p}_m)^2}=\frac{N_{AA,aa}}{\sum_m 2\hat{p}_m^2(1-\hat{p}_m)^2}, \quad (2)$$

- Estimate $\pi_{1;i,j}, \pi_{2;i,j}$ based on $N_{IBS=1}, N_{IBS=2}, \hat{p}_m, \hat{\pi}_{0;\,i,j}$. (Purcell et al., 2007. Tool: PLINK)
- $2\phi_{i,j} = \frac{\pi_{1;i,j}}{2} + \pi_{2;i,j}$

Estimate Kinship Coefficient ϕ by KING

- Assume HWE and homogeneous population
- p denotes the frequency of having a reference allele, $\hat{p} = \frac{1}{M}\sum_m \widehat{p_m}$
- $X^{(i)}, X^{(j)}$ denotes the Number of Reference Alleles for individuals i, j with $m = 1, 2, \ldots, M$ genotyped markers

$$E\left[\left(X^{(i)} - X^{(j)}\right)^2\right] = 4p(1-p)\left(1 - 2\phi_{ij}\right)$$

$$\widehat{\phi_{i,j}} = \frac{1}{2} - \frac{\sum_m \left(X_m^{(i)} - X_m^{(j)}\right)^2}{4\sum_m 2\widehat{p_m}(1-\widehat{p_m})}$$

$$\widehat{\pi_1} = 2\ -2\widehat{\pi_0} - 4\widehat{\phi_{i,j}}; \quad \widehat{\pi_2} = 4\,\widehat{\phi_{i,j}} + \widehat{\pi_0} - 1$$

Bioinformatics paper by A. Manichaikul et. al. 2010. Tool: KING.

Estimate Kinship Coefficient ϕ by KING

- Efficient computation matters
- Only SNPs present in both individuals will be used

$$\hat{\phi}_{ij} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{2\hat{H}_{ij}} + \frac{1}{2} - \frac{N_{Aa}^{(i)} + N_{Aa}^{(j)}}{4\hat{H}_{ij}}$$

- When each genotype is stored in two bits, Bit Operations can be used to computing $N_{Aa}^{(i)}, N_{Aa}^{(i)}, \mathrm{N_{Aa,Aa}}, \mathrm{N_{AA,aa}}$
- $\widehat{H_{i,j}} = \sum_m 2\widehat{p_m}(1 - \widehat{p_m})$ can be pre-calculated across all SNPs prior to the pair-wise kinship coefficient estimation, and then updating to reflect the set of observed genotypes used in analysis of each pair of individuals

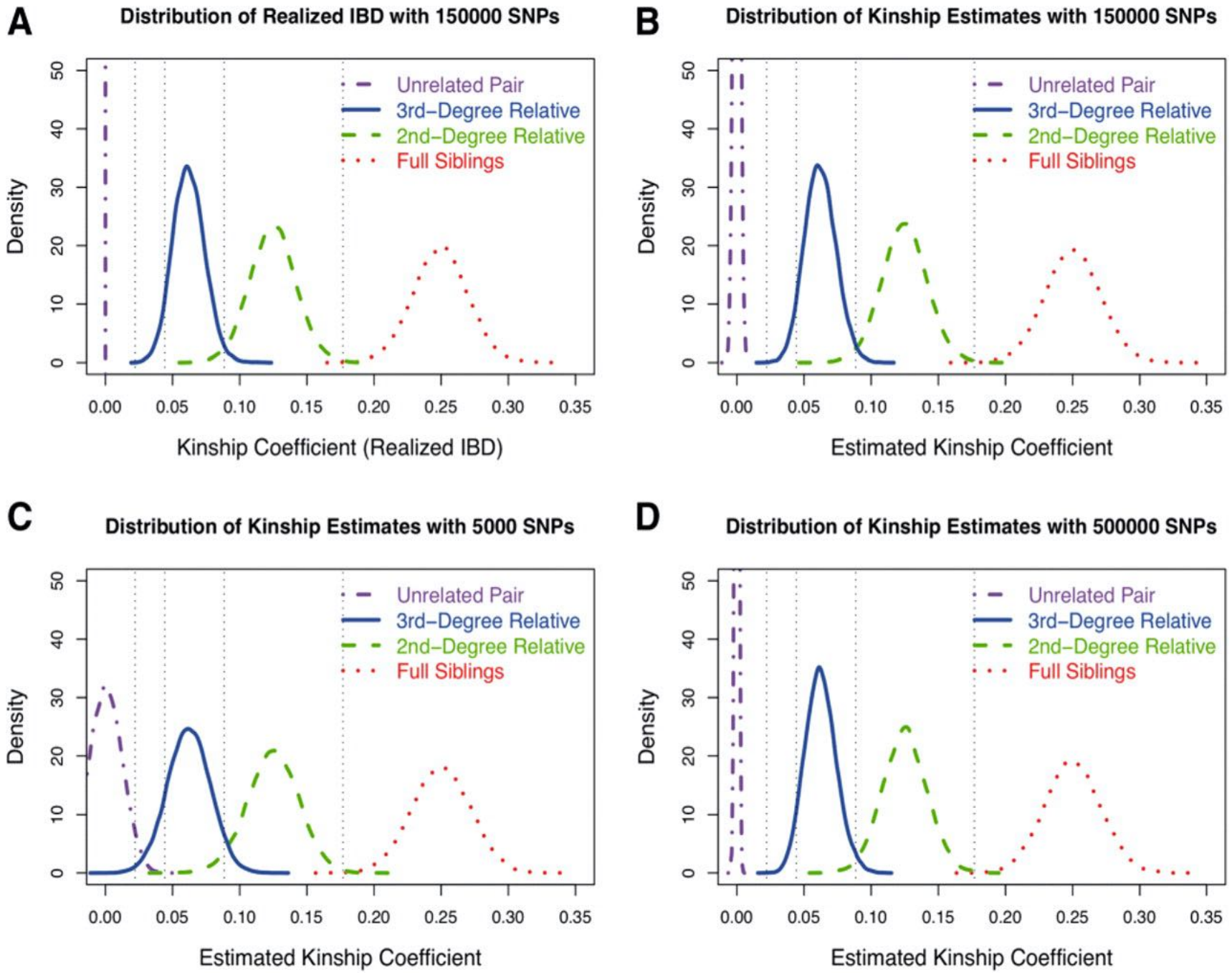
Efficient computation matters

Table 2. Computation time of two software implementations to estimate kinship coefficients in three sets of GWAS SNP data

Summary of genome scan data				Computing time	
Index	**No. of SNPs**	**No. of samples**	**No. of pairs**	**KING**	**PLINK**
1	3 079 857	269	36 046	2 min	2 h 9 min
2	324 748	602	180 901	1 min	1 h 13 min
3	549 338	2 454	3 009 832	25 min	28 h 30 mi

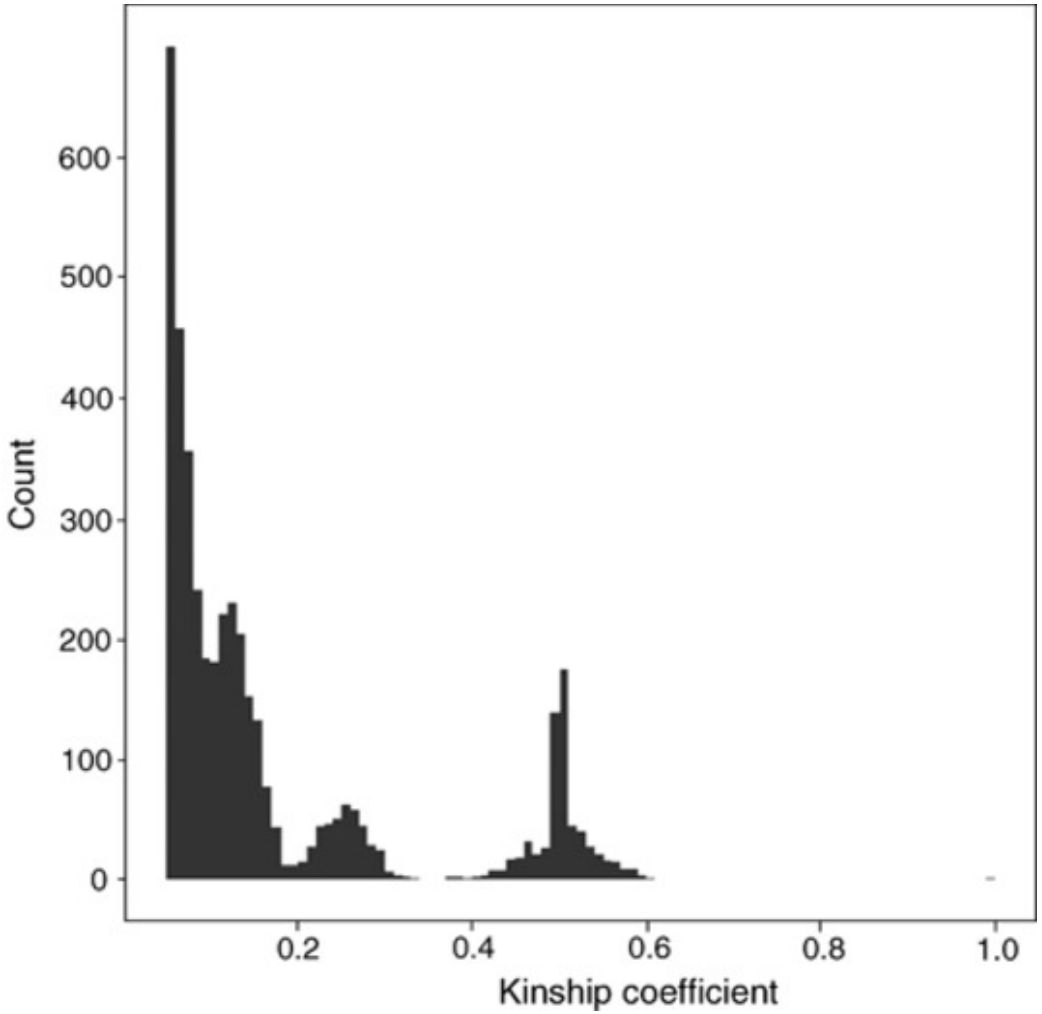
The computation time refers to the time to estimate kinship coefficients for all pairs of individuals, excluding overhead costs such as the time to load data into the computer memory. The two KING implementations (the robust algorithm and the algorithm assuming homogeneous samples) took a similar amount of computational time. This computation time can be estimated reliably as the analysis time for the entire data minus the analysis time for only the within-family data. The unit of computation time is in minutes hours. All computation was performed on and Intel Xeon with 3.20 GHz processor.

Fig. 1. Distribution of kinship coefficient estimation.
(A) Distribution of realized IBD-sharing with 150k SNPs (considering sampling one allele per individual);
(B) distribution of kinship coefficient estimates with 150k SNPs;
(C) distribution of kinship coefficient estimates with 5k SNPs;
(D) distribution of kinship coefficient estimates with 500K SNPs.



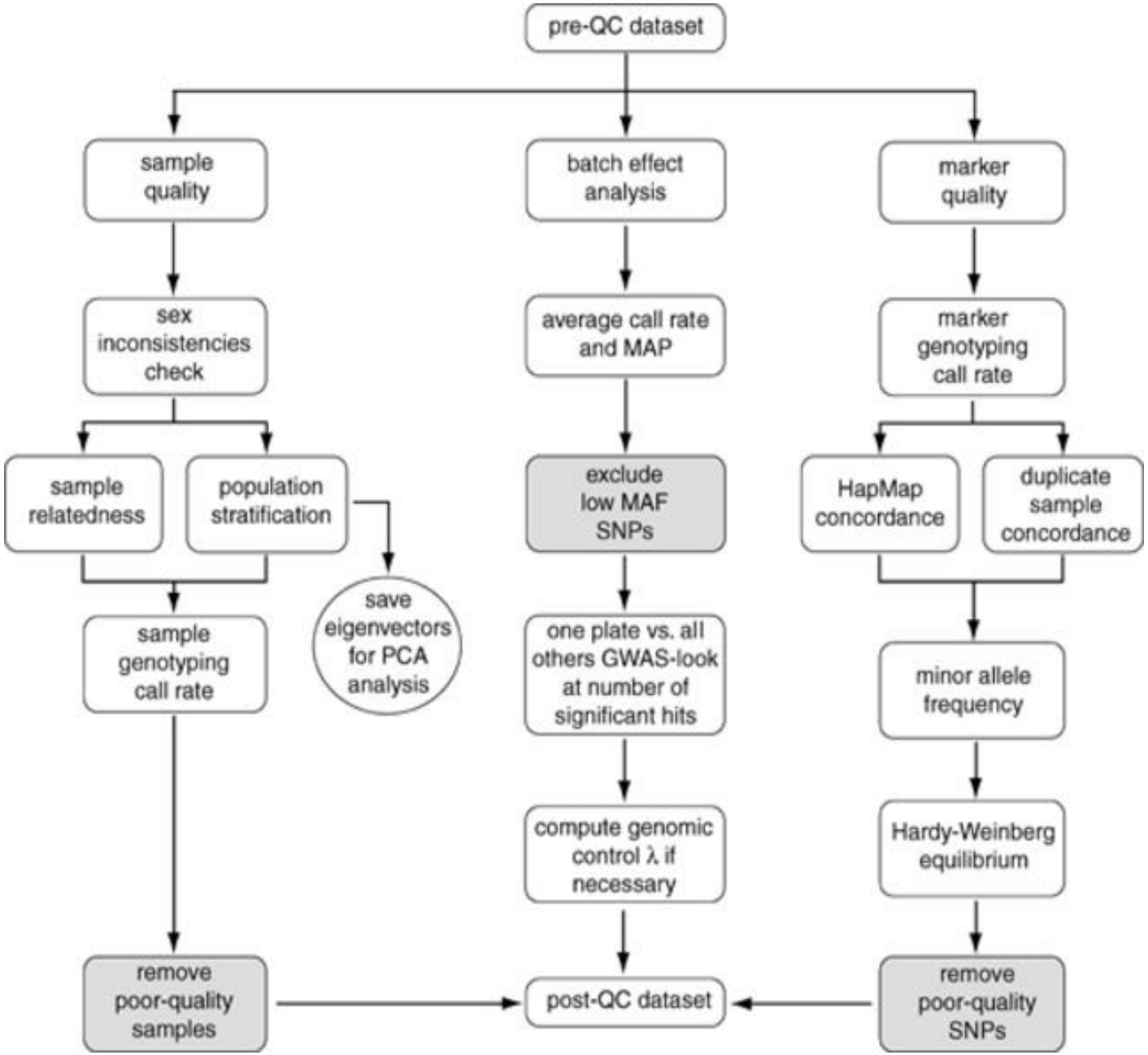
Bioinformatics, Volume 26, Issue 22, 15 November 2010, Pages 2867–2873, https://doi.org/10.1093/bioinformatics/btq559

GWAS Quality Control : Kinship Coefficients 2ϕ



Histogram showing the distribution of pairwise kinship coefficients (where kinship coefficient is greater than 0.05). The peak over 0.5 represents first degree relatives (parent-offspring, full siblings). The peak over 0.25 represents second-degree relatives (half siblings, avuncular, grandparent-grandchild). Third- and fourth-degree relatives begin to blend into more distantly related samples between zero and 0.125.

S. Turner et. al. CP hum Genetics. 2011.
https://doi.org/10.1002/0471142905.hg0119s68



Flowchart Overview of GWAS Quality Control Process

S. Turner et. al. CP hum Genetics. 2011. https://doi.org/10.1002/0471142905.hg0119s68

Check GWAS Results by Quantile-Quantile (QQ) Plot

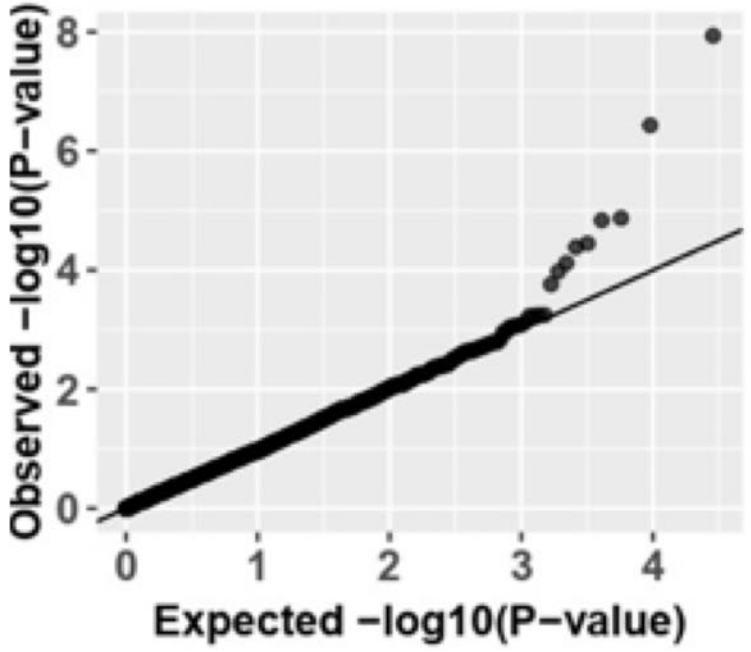
– Obtained $-\log 10$(p-values) from GWAS

– Sort all $-\log 10$(p-values) from most significant to least

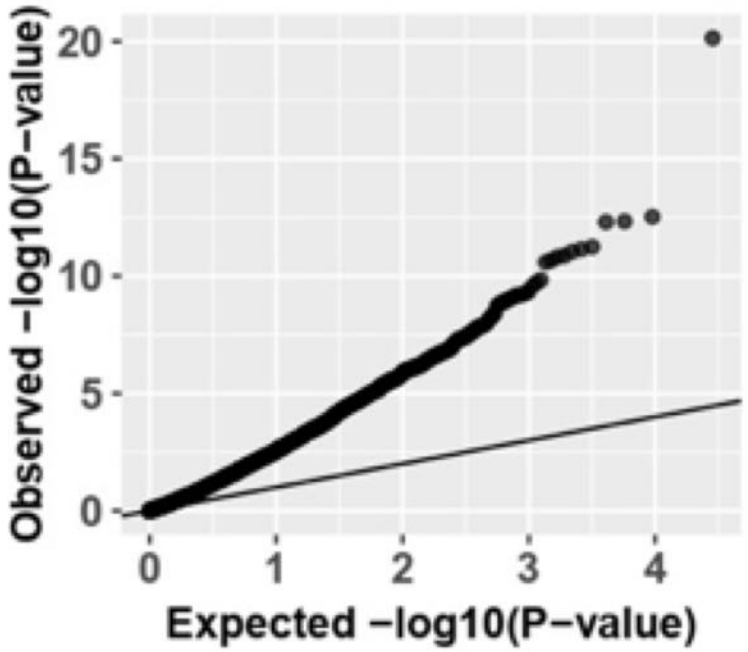
– Pair these with the expected values of order statistics of a Uniform(0, 1) distribution

– Under NULL hypothesis (no association), p-values follow a Uniform(0, 1) distribution

Without Inflated Type I Error



With Inflated Type I Error



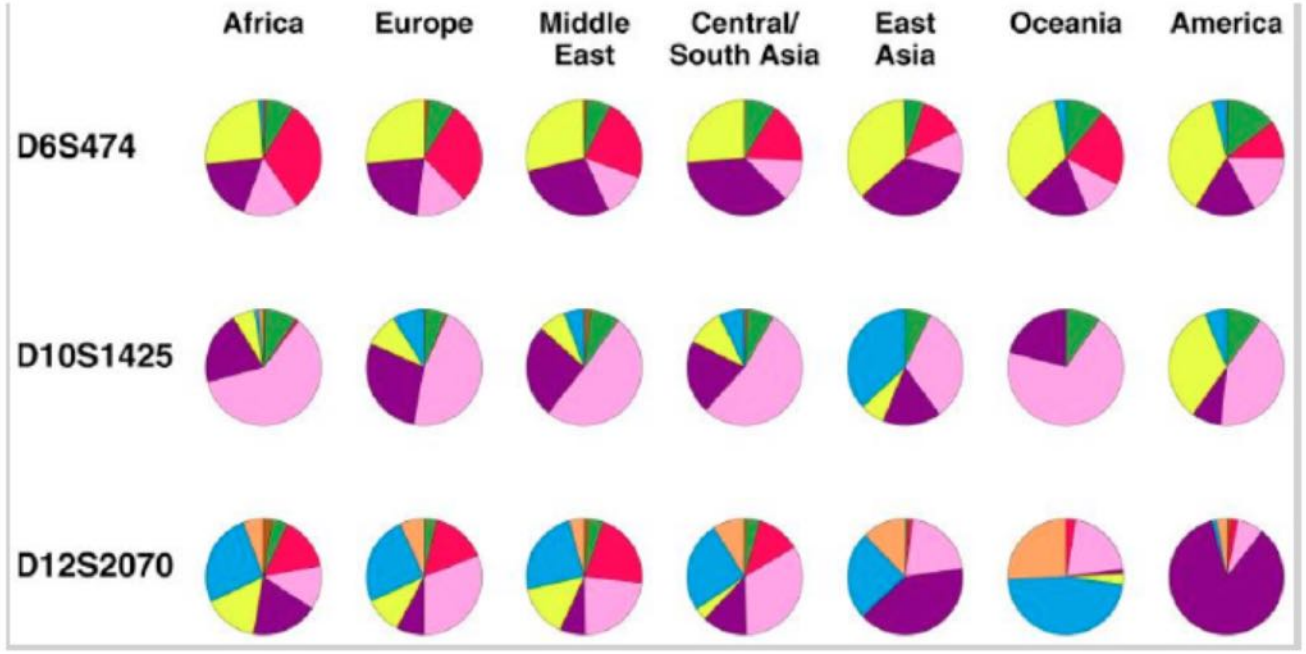
Source of Inflated GWAS Results

- Cohorts with samples of different ethnicities: e.g., European, Asian, African ancestries

- The issue of **Population Stratification**

Population Stratification

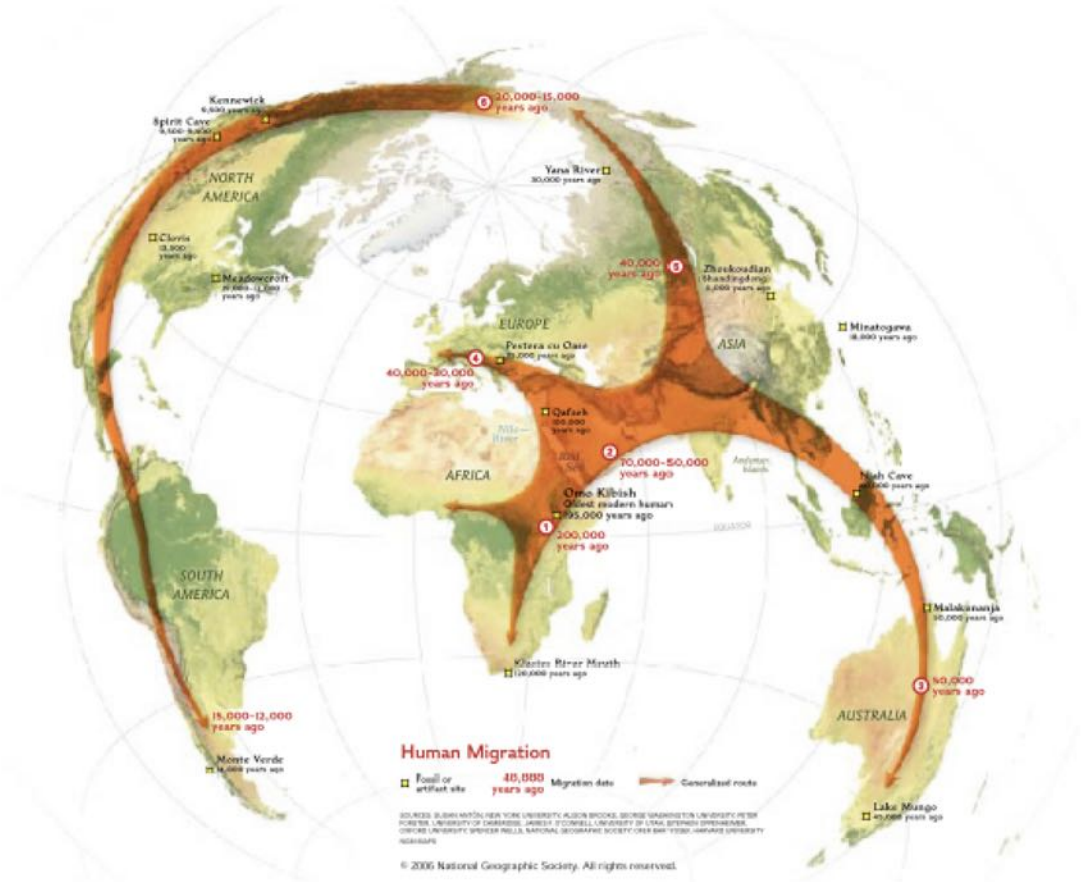
Population stratification (or population structure) is the presence of a systematic difference in allele frequencies between subpopulations, possibly due to different ancestry.



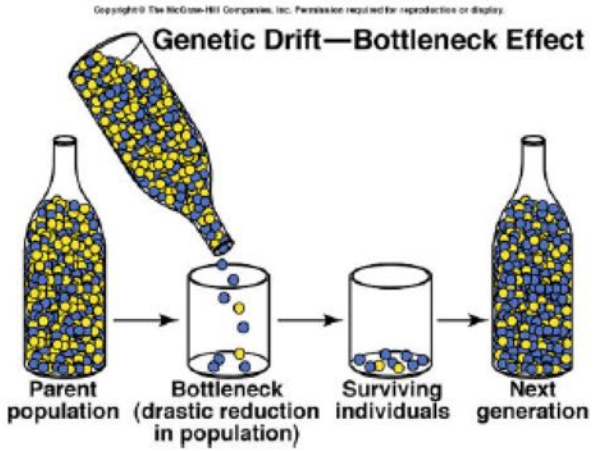
Allele frequencies at three microsatellite loci (Rosenberg N.A., Hum Biol. 2011). Each of the three loci has exactly eight alleles. In most of the pie charts, one or more alleles is rare or absent.

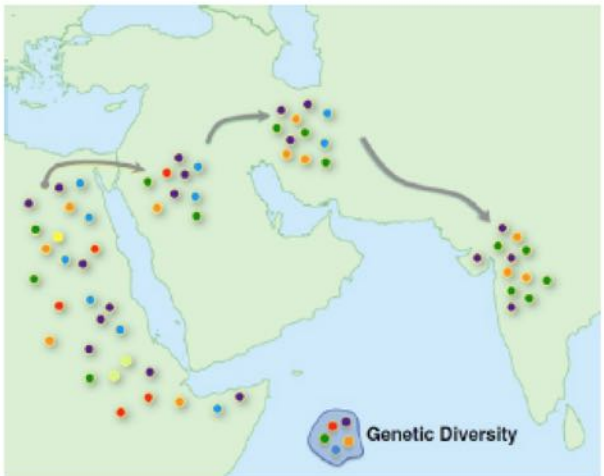
Causes of population structure

Human migration:



National Geographic





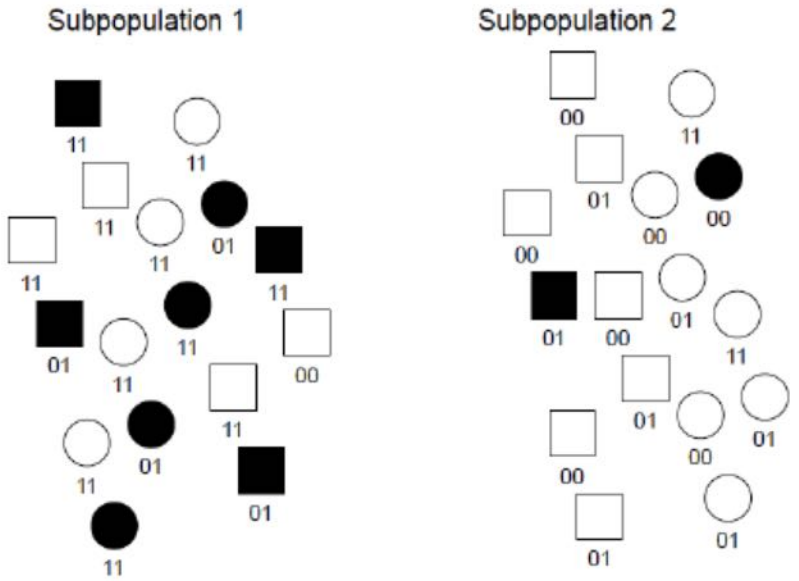
Henn *et al.* (2012) *PNAS*

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.

How to Address Population Stratification?

Most Effective Approach

- Family-based Association Analysis
- Subject to the availability of data

How to Address Population Stratification?

Simplest Approach

- Adjust false positives by Genomic Control Factor (not always work)

Check GWAS Results by Quantile-Quantile (QQ) Plot

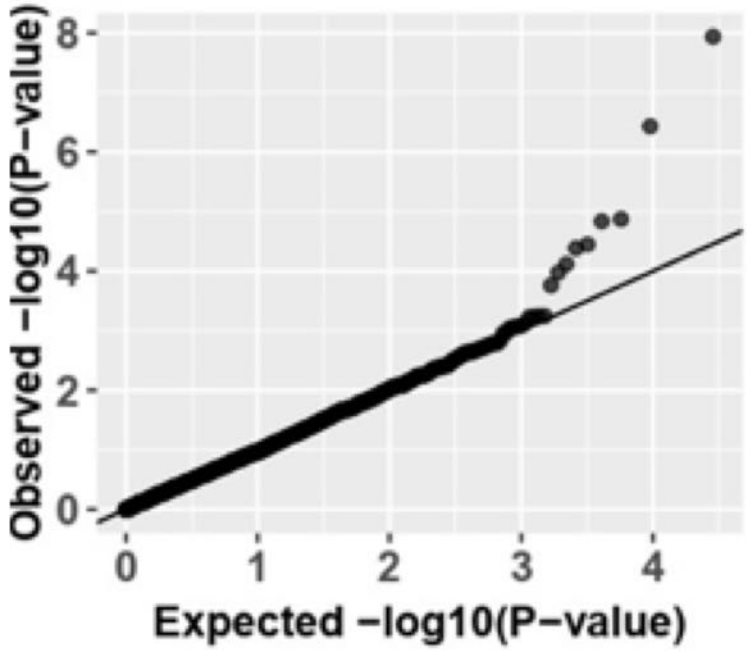
– Obtained $-\log 10$(p-values) from GWAS

– Sort all $-\log 10$(p-values) from most significant to least

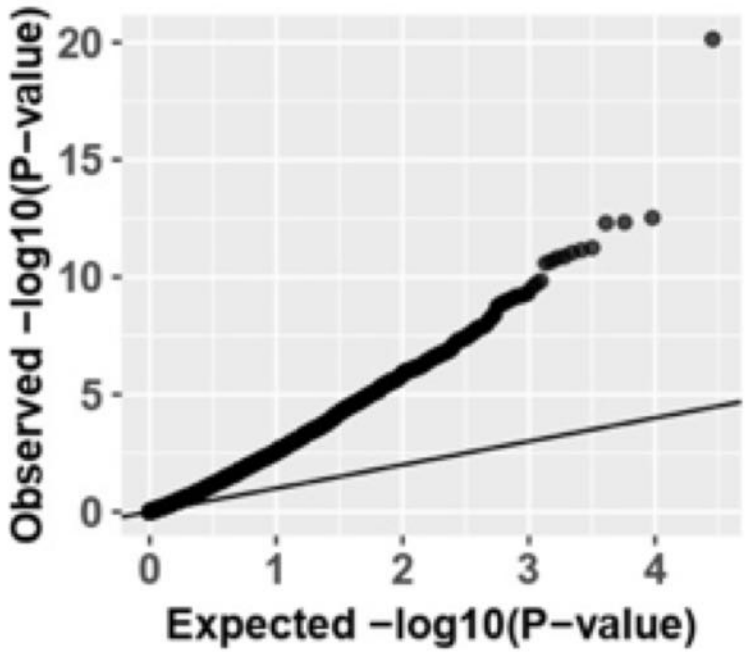
– Pair these with the expected values of order statistics of a Uniform$(0, 1)$ distribution

– Under NULL hypothesis (no association), p-values follow a Uniform$(0, 1)$ distribution

Without Inflated Type I Error



With Inflated Type I Error



Genomic Control Factor

Genomic Control Factor is used to control for systematic inflation of type I error.

The idea is that the statistic T is inflated by an inflation factor λ (i.e., genomic control factor) so that

$$T \sim \lambda \chi_1^2$$

where λ can be estimated by

$$\hat{\lambda} = \text{median}(T_1, T_2, \ldots, T_M)/0.456$$

- M is the number of independent tests, though in practice all tests are included.
- The denominator is the median of χ_1^2 distribution.
- $\hat{\lambda}$ should be 1 under H_0.

Adjust GWAS results by Genomic Control Factor λ_{GC}

- 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 $\mathrm{median}(T)$

Genomic Control Factor: $\lambda_{GC} = median(T)/0.456$

- Adjust your GWAS results by λ_{GC}
  - Scale your chi-square statistic test statistics (df=1) by λ_{GC}
  - Recalculate the corresponding GWAS p-values
  - Re-check QQ plot

Limitations of Genomic Control Factor

- Genomic control corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor.

- However, some markers differ in their allele frequencies across ancestral populations more than others.

- Thus, the uniform adjustment applied by genomic control may be insufficient at markers having unusually strong differentiation across ancestral populations and may be superfluous at markers devoid of such differentiation, leading to a loss in power

How to Address Population Stratification?

Commonly Used Approach :

- Account for variables representing ethnicity information (Principal Components Analysis)

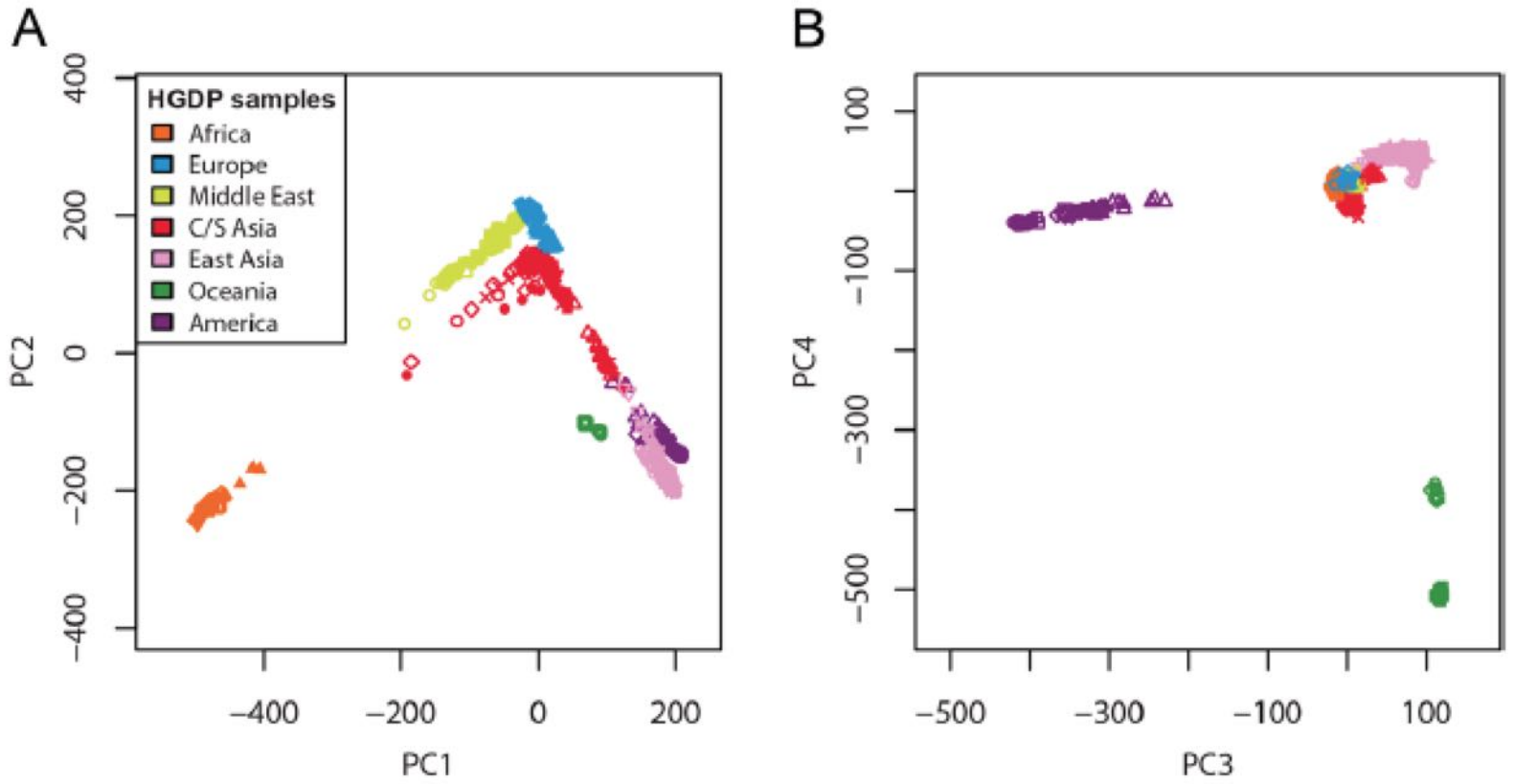
Principal Components Analysis (PCA)

- Consider genotype matrix $X_{n\times \mathrm{p}}$, with n individuals and p genome-wide SNPs
- Center and standardize columns in $X_{n\times \mathrm{p}}$ -> $Z_{n\times p}$
- 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, …
  - Compute the $n\times n$ variance-covariance matrix for all samples as $\Sigma_{n\times n} = ZZ^T/(n-1)$
  - Compute the eigenvalue decomposition of Σ, by R function eign()
  - Select top K eigenvectors (w_k) whose corresponding eigenvalues are significantly large (e.g., 5 or 10) by a scree plot
  - Principle components (PCs) are given by: Zw_k

Principal Components Analysis (PCA)

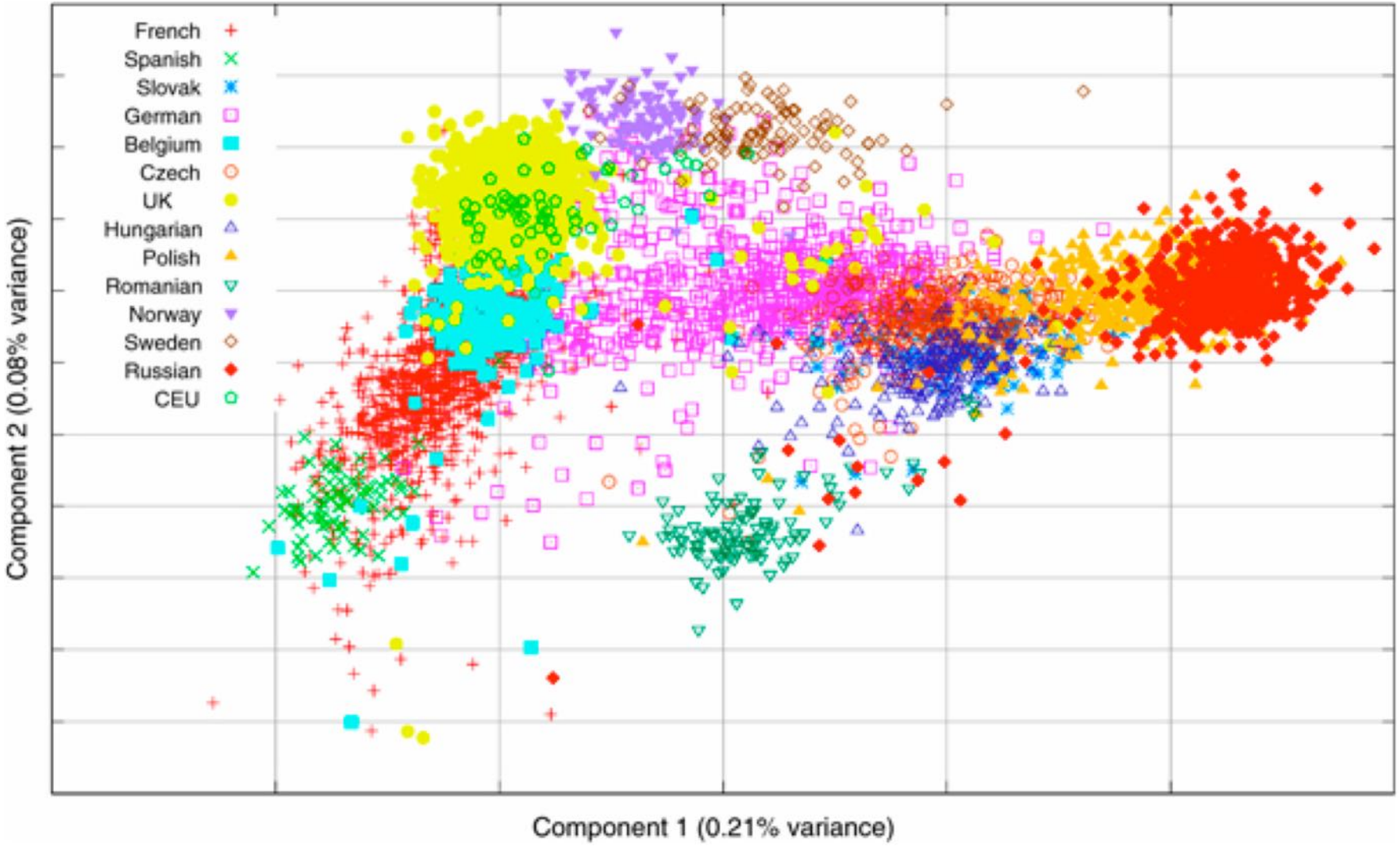
- Principal Components Analysis (PCA) with respect to $X_{n\times \mathrm{p}}$
- R function: prcomp() ; https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/prcomp
- PLINK

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
  - $\log\left(\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right) = \beta_0 + \alpha C + \beta_1 X$
  - $Y = \beta_0 + \alpha C + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

- Examine GWAS results by QQ plot for inflated type I error

How to Address Population Stratification?

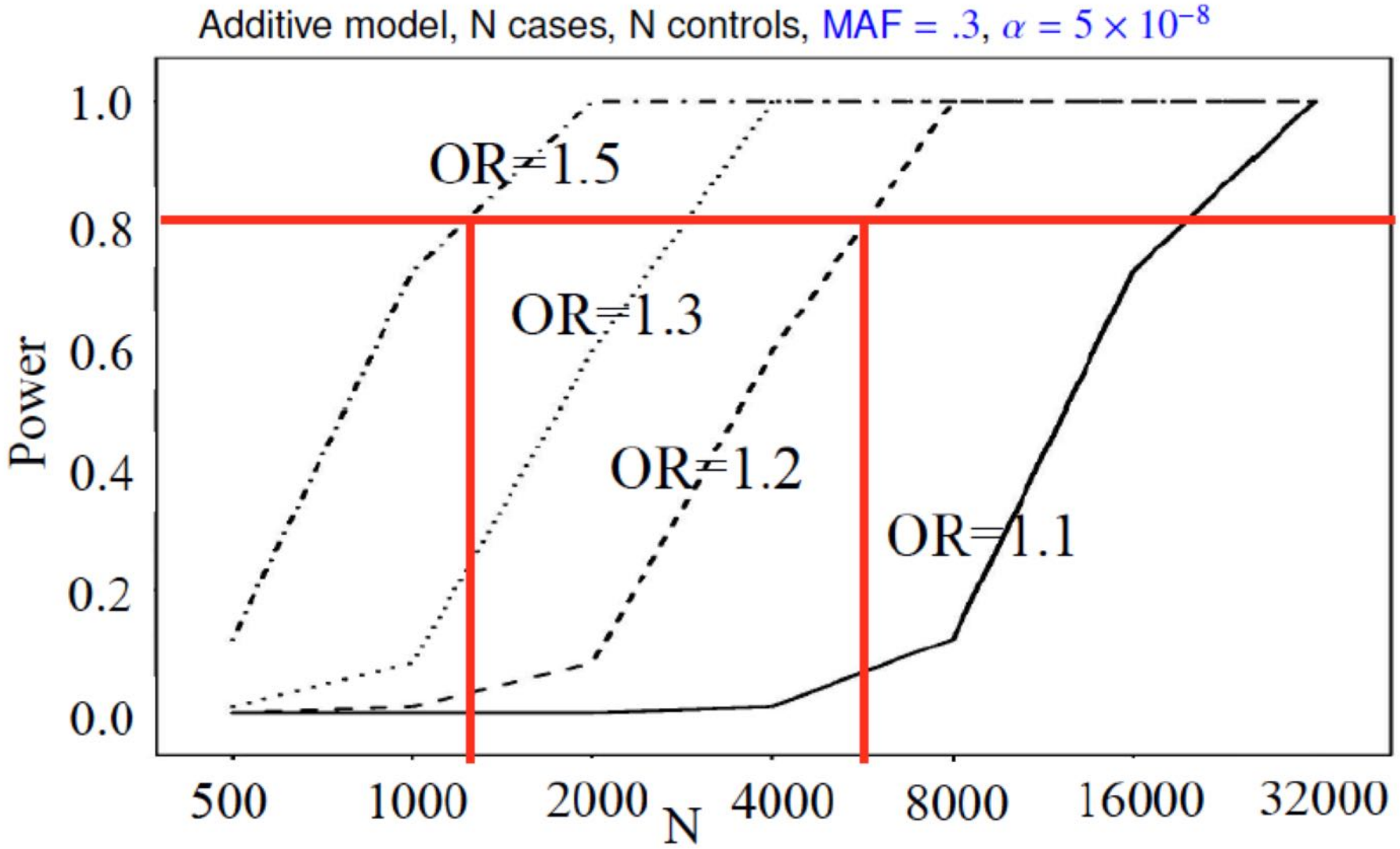
Most Robust Approach: Stratify Multi-Ethnic Cohorts

- Conduct association studies for samples of the same population/ethnicity
- Combine association results by Meta-Analysis
- Subpopulation structure still exist

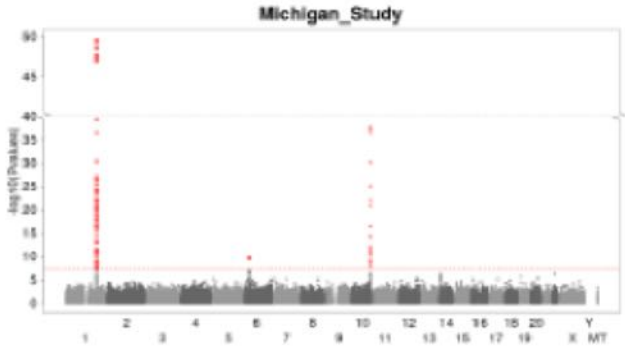
Meta-analysis

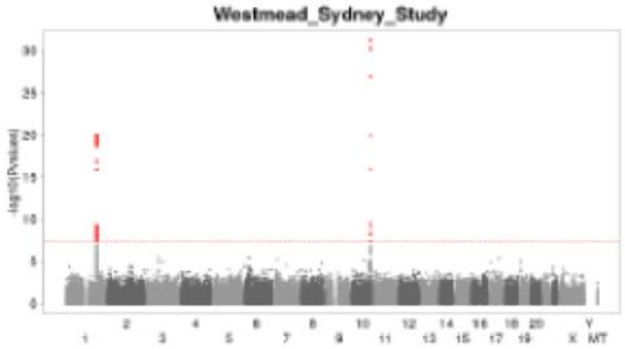
- Combine results across multiple studies for the same phenotype
- Improve power for the larger total sample size
- Address between study variances (due to population stratification, study design)
- Avoid the hassle of sharing individual-level genotype/phenotype/covariate data
- It is theoretically shown 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

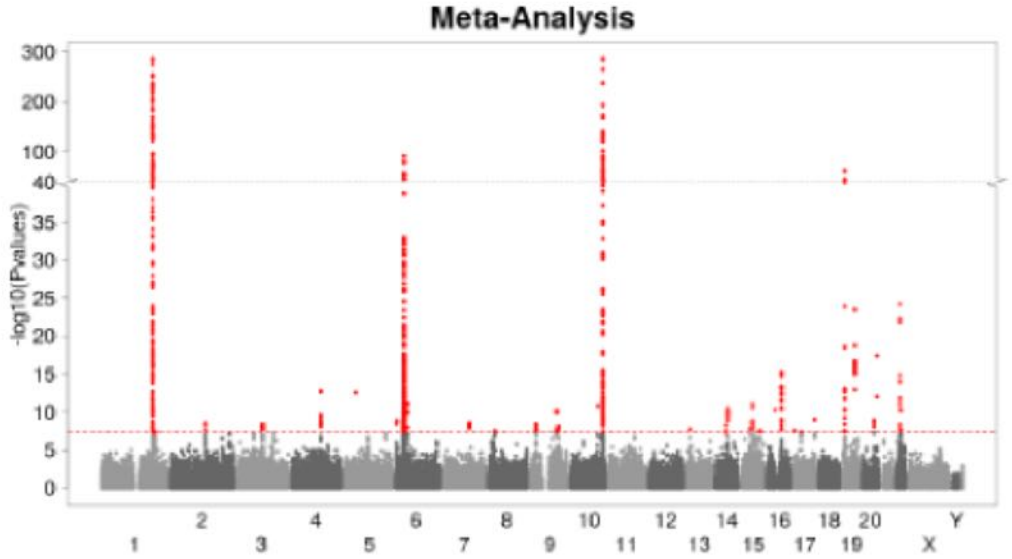


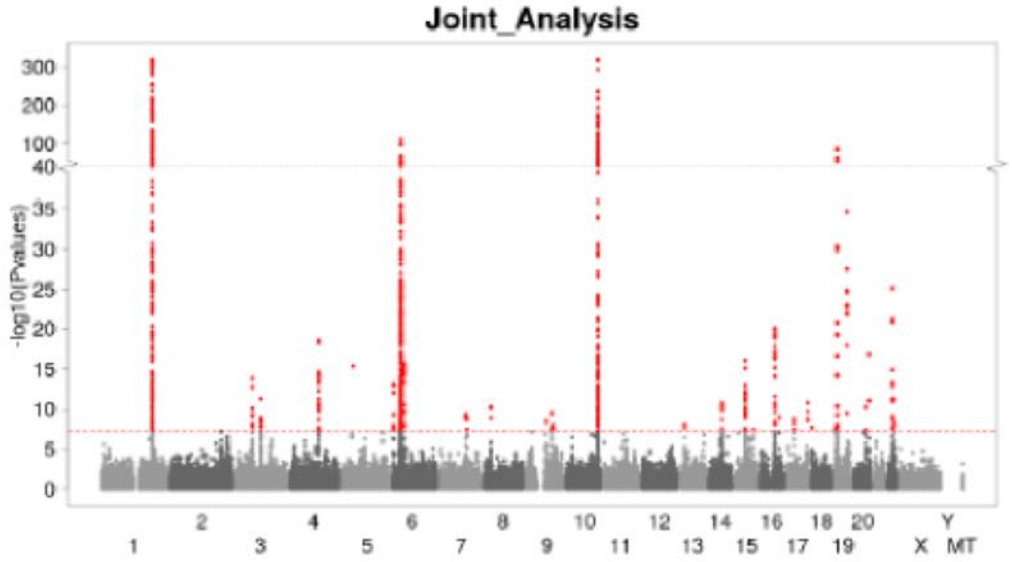
Improve Power with Larger Total Sample Size





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
- Why meta test statistic X^2 follows a Chi-square distribution under the NULL hypothesis when there is no association?

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
  - $\beta_{meta} = \frac{\sum_{k=1}^{K} w_k \beta_k}{\sum_{k=1}^{K} w_k}, \quad w_k = 1/v_k$
  - $Var(\beta_{meta}) = \frac{1}{\sum_{k=1}^{K} w_k}$
  - 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

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

- KING : Relationship inference
  https://www.kingrelatedness.com/manual.shtml

Topics for Next Lecture

- Linear Mixed Model (LMM)
- Heritability Estimation by REML
- Fine-map GWAS Results
  - Conditional analysis
  - Bayesian method
- Multivariate GWAS
  - LASSO
  - Bayesian Variable Selection Regression
  - GCTA joint analysis using GWAS summary statistics