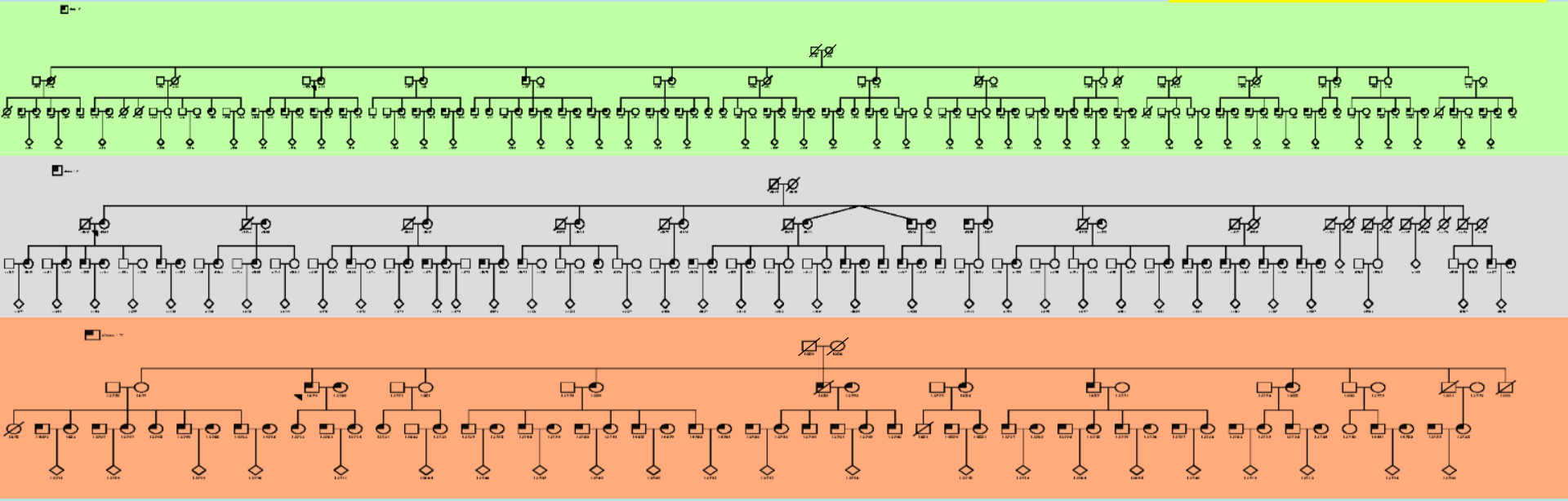
Analysis of Pedigree OMIC Data

Actual Long Life Family Study Pedigrees



Michael A. Province, PhD

Professor of Genetics & Biostatistics

DSG DIVISION of STATISTICAL GENOMICS

Washington University in St.Louis • School of Medicine

Outline

1. **FACT:** Analyzing Pedigree OMIC data using standard statistical models/tests, causes
   Severe inflation of false-positive signals
2. **WHY?** What causes the excess false-positives?
   Spoiler alert: pedigrees violate the basic statistical assumption of "i.i.d." (independent, identically distributed) data of most (non-pedigree) statistical models
3. **Statistical Methods & Software** that correctly analyze Pedigree Data by *modeling the dependences* in pedigree data (result: no inflation of false-positives, retaining power)
4. **Comparisons & Robustness** of Pedigree Models
5. **Conclusions & Recommendations**

Outline

1. **FACT:** **Analyzing Pedigree OMIC data using standard statistical models/tests, causes**

   Severe inflation of false-positive signals

2. WHY?

   Spoiler alert pedigrees violate the basic statistical assumption of "i.i.d."

3. Statistical Methods & Software

4. Comparisons & Robustness

5. Conclusions & Recommendations

Vast Literature Documenting

Analyzing Associations of Heritable Phenotypes in Pedigrees assuming the standard "i.i.d." (independent identically distributed) statistical assumptions of most stat models/software, results in **SERIOUS P-VALUE INFLATION** (too many false positives)

Boerwinkle, Chakraborty, & Sing. ***The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods*** Ann. Hum. Genet., 50, 181-194, 1986

Siegmund, ..., Province. ***A Frailty Approach for Modelling Diseases with Variable Age of Onset in Families: The NHLBI Family Heart Study.*** Statistics in Medicine, 18, 1517-1528, 1999

Borecki & Province. ***Genetic and Genomic Discovery Using Family Studies*** Circulation, 118:1057-1063 2008

Astle & Balding. ***Population Structure and Cryptic Relatedness in Genetic Association Studies***, Statistical Science Vol. 24, No. 4, 451–471, 2009

Eu-ahsunthornwattana, ..., Cordell. ***Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data***, PLOS Genetics, Vol 10, Issue 7, 2014

Zhang, . . ., Province. ***Methods for adjusting population structure and familial relatedness in association test for collective effect of multiple rare variants on quantitative traits***. BMC Proc. Nov 29;5 Suppl 9(Suppl 9):S35. 2011

Not just a problem with "regression" type models

STATISTICS IN MEDICINE
Statist. Med. **18**, 1517–1528 (1999)

A FRAILTY APPROACH FOR MODELLING DISEASES WITH VARIABLE AGE OF ONSET IN FAMILIES: THE NHLBI FAMILY HEART STUDY

KIMBERLY D. SIEGMUND,[1]* ALEXANDRE A. TODOROV[2] AND MICHAEL A. PROVINCE[3]

[1] *Department of Preventive Medicine, University of Southern California, Los Angeles, U.S.A.*
[2] *Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, U.S.A.*
[3] *Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri, U.S.A.*

SUMMARY

We use frailty models to analyse the effect of latent genetic and environmental risk factors on hazard functions in nuclear families. The approach expresses latent risk factors (frailties) as functions of the effects of a single major gene and shared familial risk. The latter may result from shared polygenes and/or a common environment. Genetic frailties are modelled using a two-point distribution, and residual frailities (shared environment, polygenes) using a gamma distribution. The two-point distribution follows the laws of Mendelian transmission, under either dominant or recessive gene action. We describe a robust EM approach for the joint estimation of the magnitude of genetic, covariate, gene by covariate interaction effects while allowing residual familial correlation. We illustrate the method on coronary heart disease data from the National Heart, Lung, and Blood Institute Family Heart Study. In addition, a simulation study shows that ignoring possible residual correlation in disease status due to a shared familial environment leads to an overestimate of the relative risk associated with a latent genotype.

1. INTRODUCTION

Cantor and Cordell *BMC Genetics* 2016, **17**(Suppl 2):3
DOI 10.1186/s12863-015-0311-z

BMC Genetics

PROCEEDINGS Open Access

Gene expression in large pedigrees: analytic approaches

Rita M. Cantor[1*] and Heather J. Cordell[2]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Background: We currently have the ability to quantify transcript abundance of messenger RNA (mRNA), genome-wide, using microarray technologies. Analyzing genotype, phenotype and expression data from 20 pedigrees, the members of our Genetic Analysis Workshop (GAW) 19 gene expression group published 9 papers, tackling some timely and important problems and questions. To study the complexity and interrelationships of genetics and gene expression, we used established statistical tools, developed newer statistical tools, and developed and applied extensions to these tools.

Methods: To study gene expression correlations in the pedigree members (without incorporating genotype or trait data into the analysis), 2 papers used principal components analysis, weighted gene coexpression network analysis, meta-analyses, gene enrichment analyses, and linear mixed models. To explore the relationship between genetics and gene expression, 2 papers studied expression quantitative trait locus allelic heterogeneity through conditional association analyses, and epistasis through interaction analyses. A third paper assessed the feasibility of applying allele-specific binding to filter potential regulatory single-nucleotide polymorphisms (SNPs). Analytic approaches included linear mixed models based on measured genotypes in pedigrees, permutation tests, and covariance kernels. To incorporate both genotype and phenotype data with gene expression, 4 groups employed linear mixed models, nonparametric weighted U statistics, structural equation modeling, Bayesian unified frameworks, and multiple regression.

Results and discussion: Regarding the analysis of pedigree data, we found that gene expression is familial, indicating that at least 1 factor for pedigree membership or multiple factors for the degree of relationship should be included in analyses, and we developed a method to adjust for familiality prior to conducting weighted co-expression gene

DOI: 10.1534/genetics.104.031666

Quantifying the Relationship Between Gene Expressions and Trait Values in General Pedigrees

Yan Lu,* Peng-Yuan Liu,* Yong-Jun Liu,* Fu-Hua Xu* and Hong-Wen Deng*,†,‡,1

**Osteoporosis Research Center, Creighton University, Omaha, Nebraska 68131, †Key Laboratory of Biomedical Information Engineering of Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China and ‡Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha, Hunan 410081, People's Republic of China*

Manuscript received May 25, 2004
Accepted for publication September 13, 2004

ABSTRACT

Treating mRNA transcript abundances as quantitative traits and examining their relationships with clinical traits have been pursued by using an analytical approach of quantitative genetics. Recently, Kraft *et al.* presented a family expression association test (FEXAT) for correlation between gene expressions and trait values with a family-based (sibships) design. This statistic did not account for biological relationships of related subjects, which may inflate type I error rate and/or decrease power of statistical tests. In this article, we propose two new test statistics based on a variance-components approach for analyses of microarray data obtained from general pedigrees. Our methods accommodate covariance between relatives for unmeasured genetic effects and directly model covariates of clinical importance. The efficacy and validity of our methods are investigated by using simulated data under different sample sizes, family sizes, and family structures. The proposed LR method has correct type I error rate with moderate to large sample sizes regardless of family structure and family sizes. It has higher power with complex pedigrees and similar power to the FEXAT with sibships. The other proposed FEXAT(R) method is favorable with large family sizes, regardless of sample sizes and family structure. Our methods, robust to population stratification, are complementary to the FEXAT in expression-trait association studies.

IN the past few years, there has been increasing interest in genetic studies of complex diseases by combining information on clinical traits, marker genotypes, and comprehensive gene expressions. It was proposed

proach to mapping the determinants of variation in gene expression. Their results suggested that the expression of most genes is affected by more than one locus (Brem *et al.* 2002). Most complex human phenotypes

The Author(s) *BMC Proceedings* 2016, **10**(Suppl 7):10
DOI 10.1186/s12919-016-0037-6

BMC Proceedings

PROCEEDINGS Open Access

Genome-wide QTL and eQTL analyses using Mendel

Hua Zhou[1*], Jin Zhou[3], Tao Hu[1,2], Eric M. Sobel[4] and Kenneth Lange[4,5,6]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Pedigree genome-wide association studies (GWAS) (Option 29) in the current version of the Mendel software is an optimized subroutine for performing large-scale genome-wide quantitative trait locus (QTL) analysis. This analysis (a) works for random sample data, pedigree data, or a mix of both; (b) is highly efficient in both run time and memory requirement; (c) accommodates both univariate and multivariate traits; (d) works for autosomal and x-linked loci; (e) correctly deals with missing data in traits, covariates, and genotypes; (f) allows for covariate adjustment and constraints among parameters; (g) uses either theoretical or single nucleotide polymorphism (SNP)–based empirical kinship matrix for additive polygenic effects; (h) allows extra variance components such as dominant polygenic effects and household effects; (i) detects and reports outlier individuals and pedigrees; and (j) allows for robust estimation via the t-distribution. This paper assesses these capabilities on the genetics analysis workshop 19 (GAW19) sequencing data. We analyzed simulated and real phenotypes for both family and random sample data sets. For instance, when jointly testing the 8 longitudinally measured systolic blood pressure and diastolic blood pressure traits, it takes Mendel 78 min on a standard laptop computer to read, quality check, and analyze a data set with 849 individuals and 8.3 million SNPs. Genome-wide expression QTL analysis of 20,643 expression traits on 641 individuals with 8.3 million SNPs takes 30 h using 20 parallel runs on a cluster. Mendel is freely available at http://www.genetics.ucla.edu/software.

Contents lists available at ScienceDirect

Journal of Nutrition & Intermediary Metabolism

journal homepage: http://www.jnimonline.com/

Genome- and CD4$^+$ T-cell methylome-wide association study of circulating trimethylamine-N-oxide in the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)

Stella Aslibekyan [a, *], Marguerite R. Irvin [a], Bertha A. Hidalgo [a], Rodney T. Perry [a], Elias J. Jeyarajah [b], Erwin Garcia [b], Irina Shalaurova [b], Paul N. Hopkins [c], Michael A. Province [d], Hemant K. Tiwari [e], Jose M. Ordovas [f, g, h], Devin M. Absher [i], Donna K. Arnett [j]

[a] Department of Epidemiology, University of Alabama at Birmingham, 1665 University Blvd, RPHB 230J, Birmingham, AL 35294, United States
[b] LipoScience, Laboratory Corporation of America® Holdings, 2500 Sumner Blvd, Raleigh, NC 27616, United States
[c] Department of Internal Medicine, University of Utah, 420 Chipeta Way #1160, Salt Lake City, UT 84108, United States
[d] Division of Statistical Genomics, Washington University in St Louis, 4444 Forest Park Blvd, Campus Box 8506, St Louis, MO 63108, United States
[e] Department of Biostatistics, University of Alabama at Birmingham, 1665 University Blvd, RPHB 420C, Birmingham, Al, 35294, United States
[f] Nutrition and Genomics Laboratory, Jean Mayer USDA HNRCA, Tufts University, 711 Washington St, Boston, MA 02111, United States
[g] Department of Epidemiology, Centro National Investigaciones Cardiovasculares, Madrid, Spain
[h] Instituto Madrileno de Estudios Avanzados en Alimentacion, Madrid, Spain
[i] HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35806, United States
[j] College of Public Health, University of Kentucky, 111 Washington Ave, Lexington, KY 40508, United States

ARTICLE INFO

Article history:
Received 20 January 2017
Received in revised form
2 March 2017
Accepted 5 March 2017
Available online 8 March 2017

Keywords:
Atherosclerosis
Cardiovascular disease
Genetic
Epigenetic
Methylation
Trimethylamine-N-oxide

ABSTRACT

Background: Trimethylamine-N-oxide (TMAO), an atherogenic metabolite species, has emerged as a possible new risk factor for cardiovascular disease. Animal studies have shown that circulating TMAO levels are regulated by genetic and environmental factors. However, large-scale human studies have failed to replicate the observed genetic associations, and epigenetic factors such as DNA methylation have never been examined in relation to TMAO levels.

Methods and results: We used data from the family-based Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) to investigate the heritable determinants of plasma TMAO in humans. TMAO was not associated with other plasma markers of cardiovascular disease, e.g. lipids or inflammatory cytokines. We first estimated TMAO heritability at 27%, indicating a moderate genetic influence. We used 1000 Genomes imputed data (n = 626) to estimate genome-wide associations with TMAO levels, adjusting for age, sex, family relationships, and study site. The genome-wide study yielded one significant hit at the genome-wide level, located in an intergenic region on chromosome 4. We subsequently quantified epigenome-wide DNA methylation using the Illumina Infinium array on CD4$^+$ T-cells. We tested for association of methylation loci with circulating TMAO (n = 847), adjusting for age, sex, family relationships, and study site as the genome-wide study plus principal components capturing CD4$^+$ T-cell purity. Upon adjusting for multiple testing, none of the epigenetic findings were statistically significant.

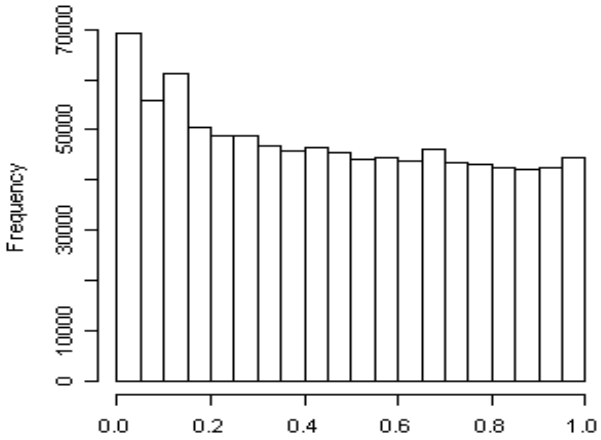
Conclusions: Our findings contribute to the growing body of evidence suggesting that neither genetic nor epigenetic factors play a critical role in establishing circulating TMAO levels in humans.

P-value Distribution

GWAS Analysis of Quantitative Pedigree Data

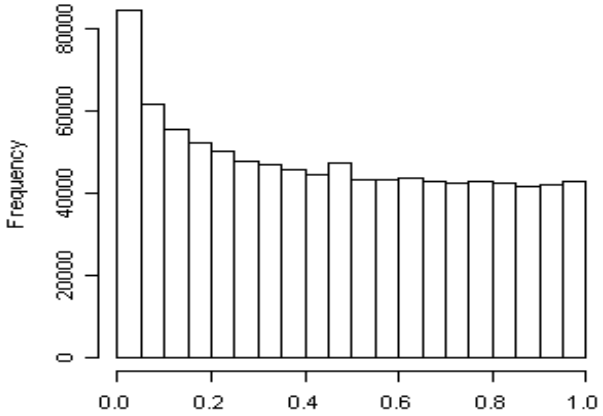
GeneSTAR* Data
(epinephrine)

OLS Regular regression
(family ignored)

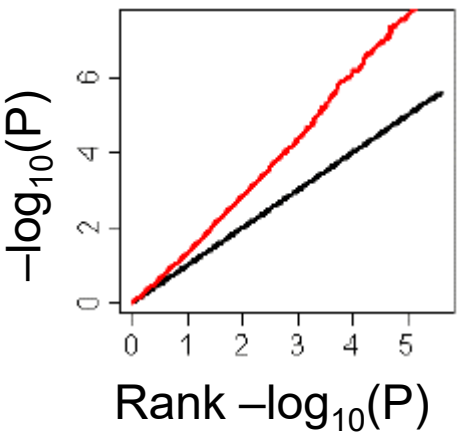


*www.hopkinsmedicine.org/general-internal-medicine/research/gene-star

Simulated Data
(under H0)



Q-Q Plot



Biological Assumption:

- EVERY “full OMIC” scan against any fixed phenotype (a.k.a. outcome, “label”), the **vast majority of associations should be under H0**

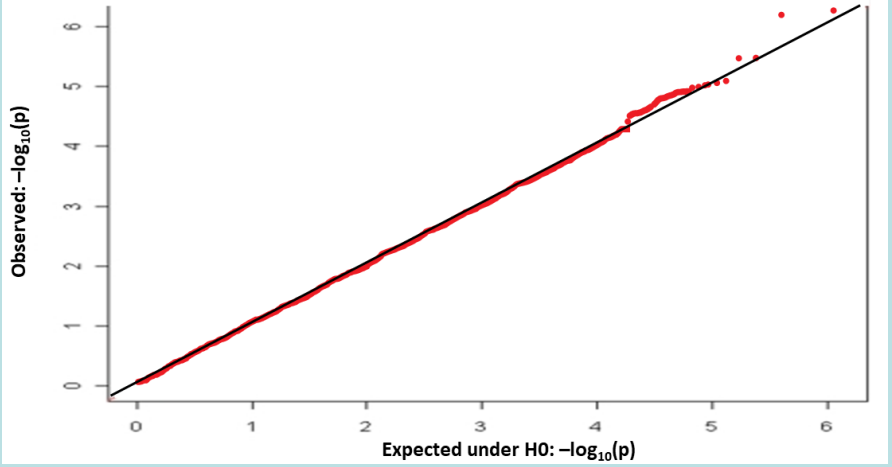
- i.e. only a **relatively small number** of all OMIC features should be **true-positives**

For ANY OMIC scan

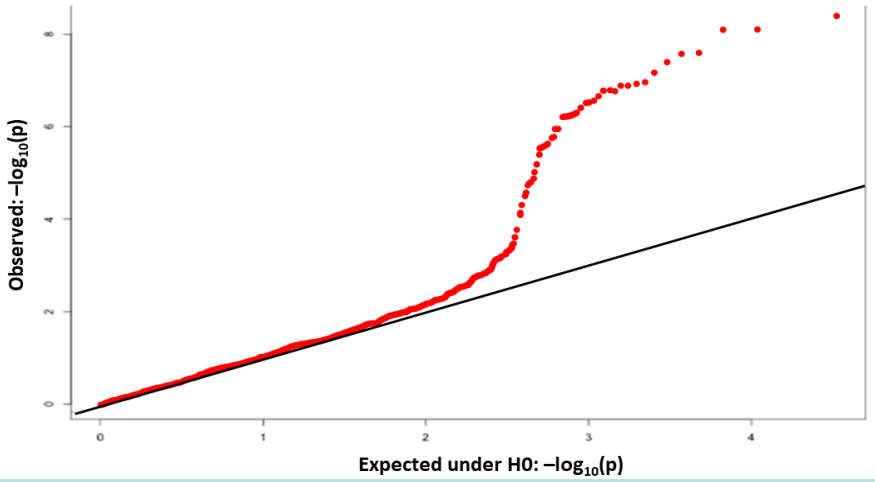
Quantile-Quantile (Q-Q) plot is your BEST FRIEND

Tells when your tests not running true

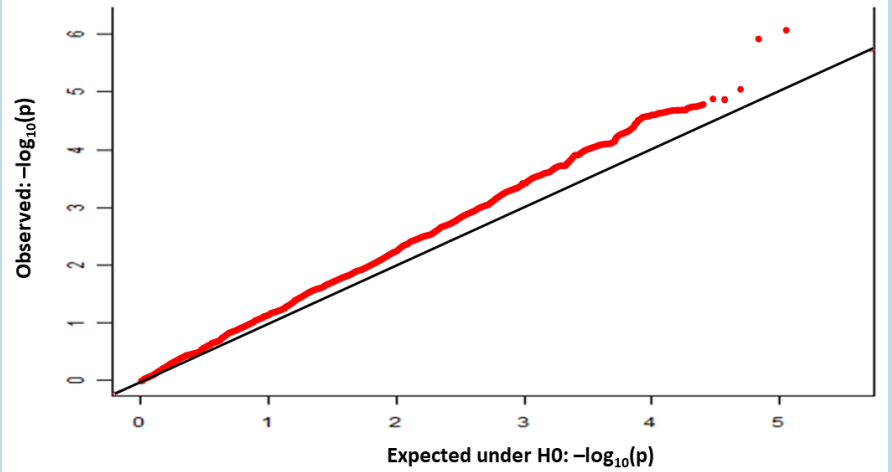
Complete H0 p-values
(p-values uniform)



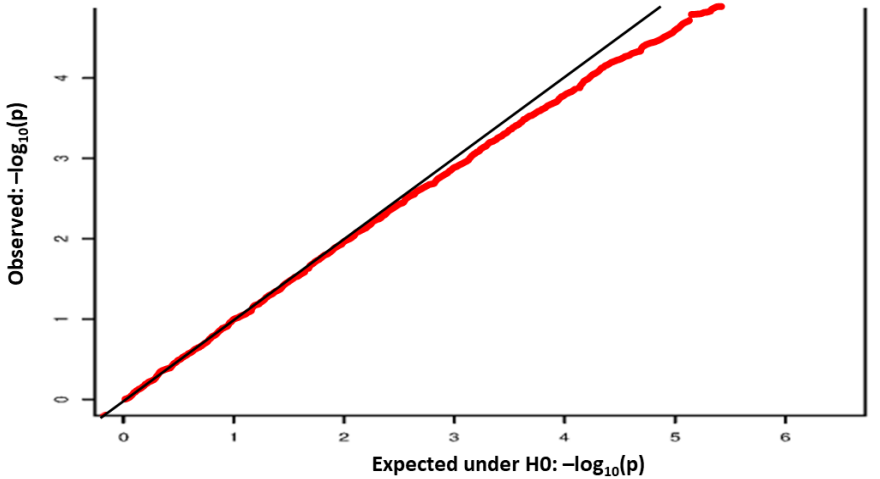
A mixture of
H0 and H1 p-values



Inflation of type-I error
(p-values too significant)

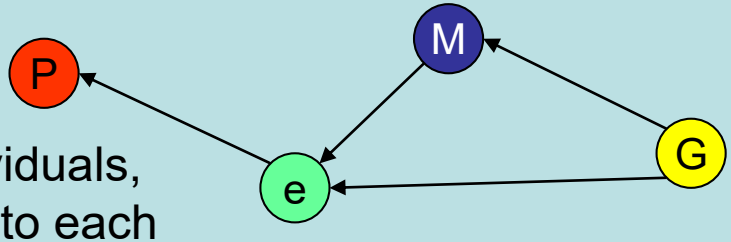


deflation of type-I error
(p-values too big→ loss of power)



Some Model connecting Genotype to methylation, expression ultimately to phenotype

If all data are unrelated individuals, Model applies to each person (i.i.d.)



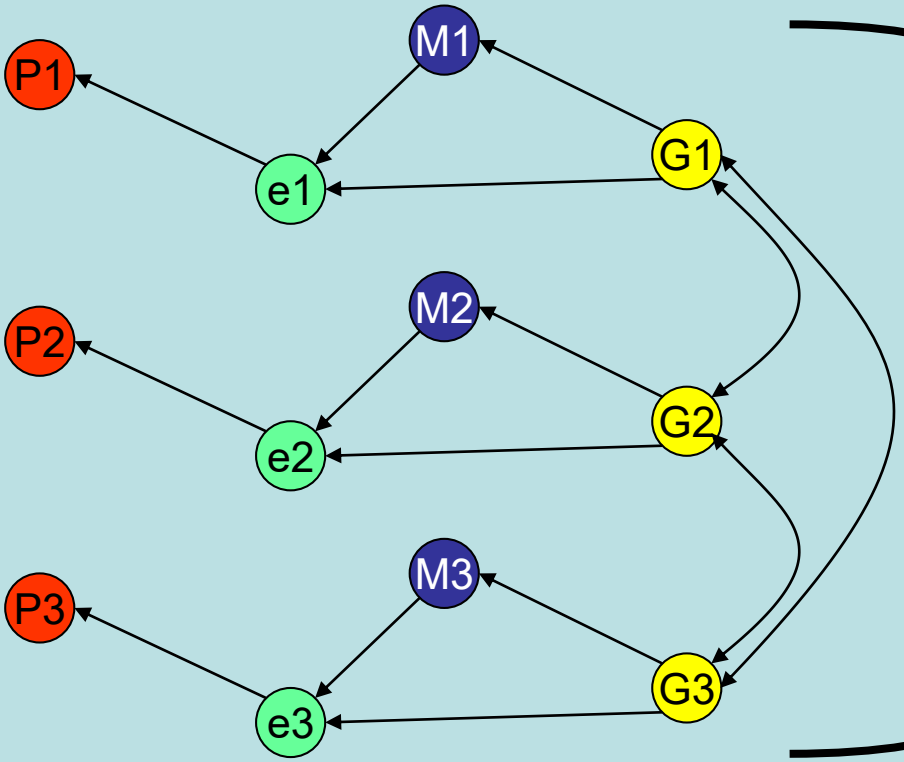
For Pedigree data Individuals:

- i.i.d. ACROSS Pedigrees
- DEPENDENT WITHIN Pedigrees

Person 1

Person 2

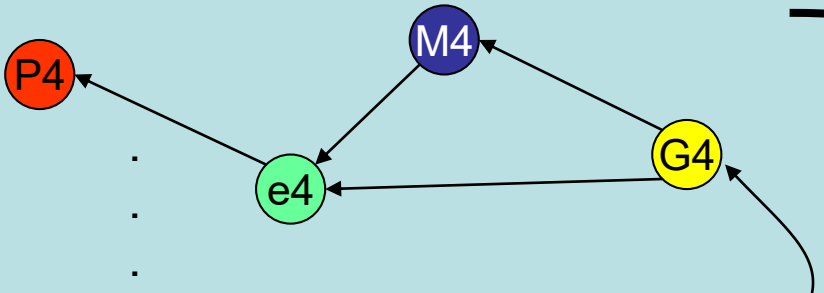
Person 3



Genetic Correlations induce Familial Correlations (non-independence) between other heritable components/features

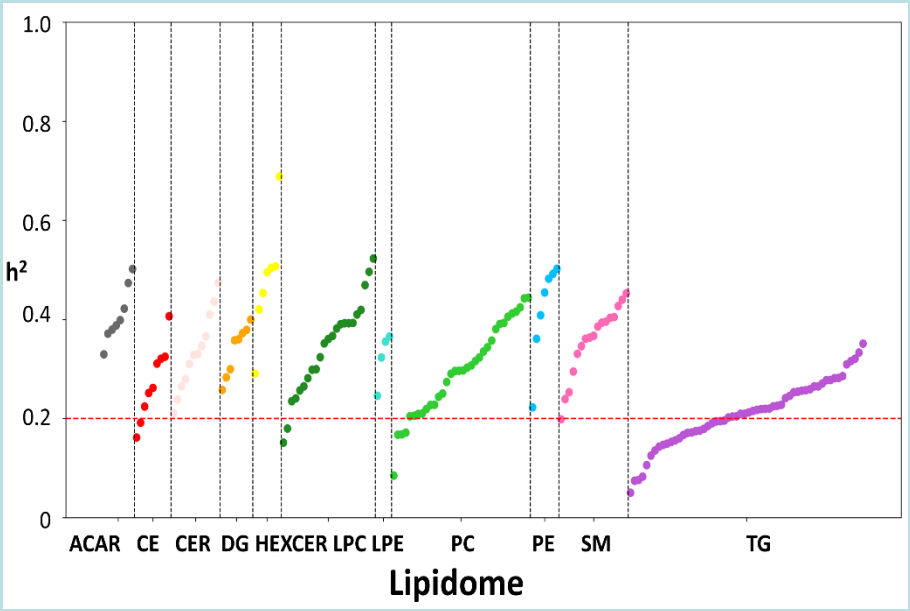
Ped 1

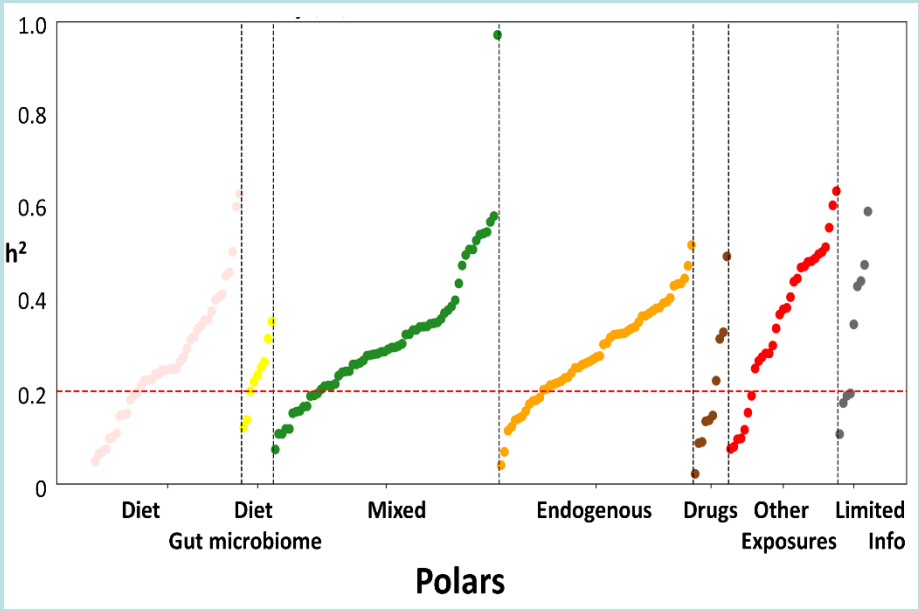
Person 4



Ped 2

Distribution of **Metabolome (LC/MS) heritabilities (h^2)** for Long Life Family Study (LLFS) at Visit 1 by metabolite family





Akbary Moghaddam V, Acharya S, Schwaiger-Haber M, Liao S, Jung WJ, Thyagarajan B, Shriver LP, Daw EW, Saccone NL, An P, Brent MR, Patti GJ, Province MA. ***Construction of Multi-Modal Transcriptome-Small Molecule Interaction Networks from High-Throughput Measurements to Study Human Complex Traits.*** bioRxiv [Preprint]. 2025 Jan 23:2025.01.22.634403. doi: 10.1101/2025.01.22.634403. PMID: 39896668; PMCID: PMC11785221.

Even many exogenous exposures in the metabolome
(diet, gut microbiome products, drugs, etc.)
have high heritability!

Outline

1. FACT:

   Severe inflation of false-positive signals

2. **WHY?** **What causes the excess false-positives?**

   ***Spoiler alert*: pedigrees violate the basic statistical assumption of "i.i.d." (independent, identically distributed) data of most (non-pedigree) statistical models**

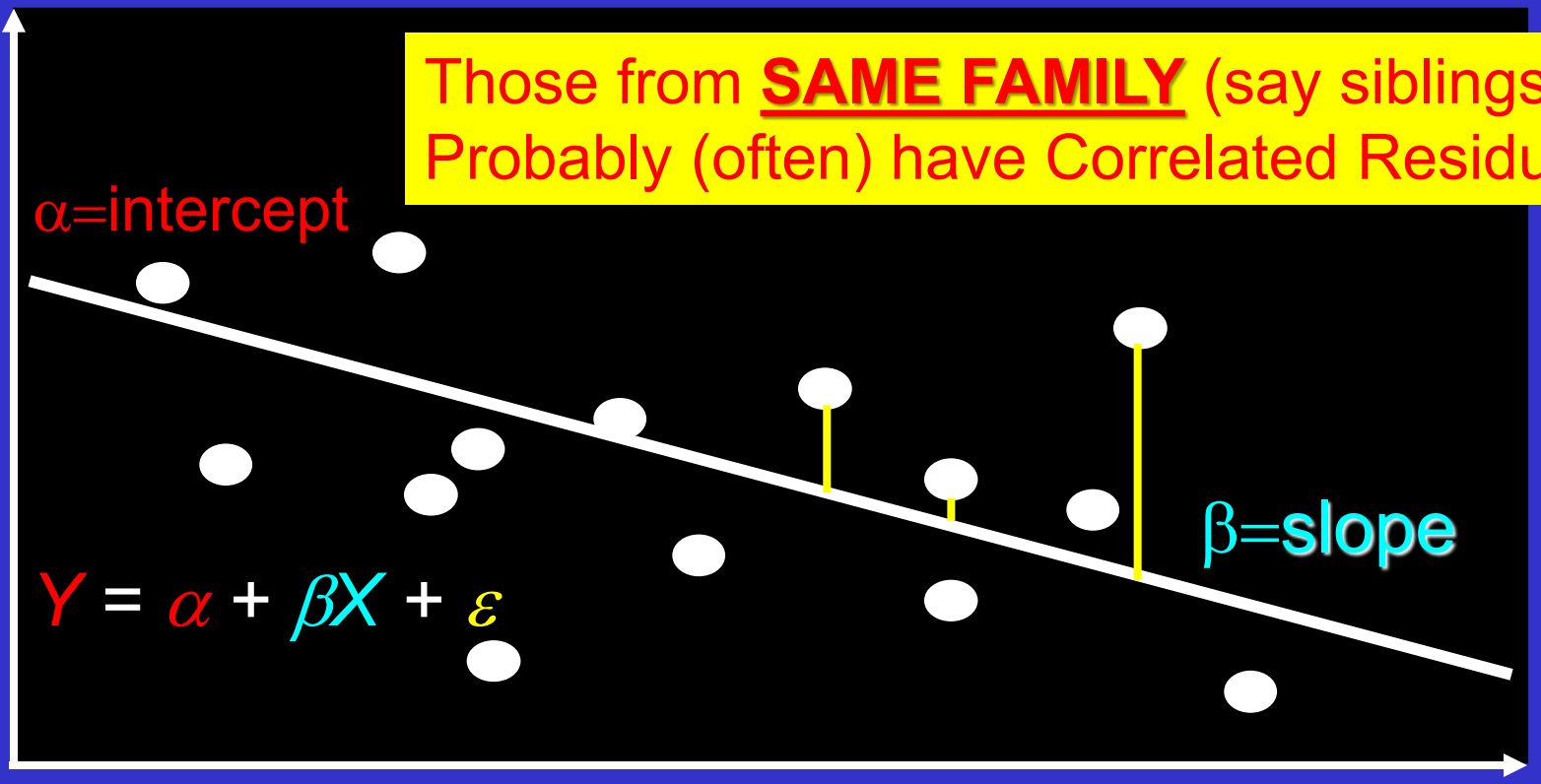
3. Statistical Methods & Software

4. Comparisons & Robustness

5. Conclusions & Recommendations

Simple E.G. OLS (Ordinary Least Squares) Regression for Complex, Heritable Trait, Y, on an OMIC feature: What if Data from Families not unrelateds?

(not i.i.d. independent identically distributed errors)



WHY? Because "X" is not the only cause of "Y"

(all other unmodeled covars are summed to "ε" and many of these are likely heritable)

Result: underestimate error variance → **falsely inflated significance**

Basic regression idea still holds

i^{th} person in k^{th} cluster (family)
X_{ik} is risk factor, Y_{ik} phenotype

$$Y_{ik} = \alpha + \beta X_{ik} + \varepsilon_{ik}$$

Basic regression idea still holds

i^{th} person in k^{th} cluster (family)
X_{ik} is risk factor, Y_{ik} phenotype

$$Y_{ik} = \alpha + \beta X_{ik} + \varepsilon_{ik}$$

PROBLEM: Residuals ε_{ik} and ε_{jk} are correlated <u>within</u> the same cluster (k) but not across clusters.

What if ignore clusters?

What if ignore clusters?

1. Since $E[\varepsilon_{ik}]=0$ for all i,k, ***estimates*** of α and β are ***unbiased*** [i.e. fall $\pm$ around "true" values of intercept and slope with random error]

What if ignore clusters?

1. Since $E[\varepsilon_{ik}]=0$ for all i,k, ***estimates*** of α and β are ***unbiased*** [i.e. fall $\pm$ around "true" values of intercept and slope with random error]

2. BUT, because residuals correlated, ***stderrs*** of these estimates are ***biased downward*** [i.e. too small]

1. Classic OLS (Ordinary Least Squares)

Point Estimates of intercept, slope are good (but OLS **variance estimates are too small**)

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

2. Take observed correlations between OLS residuals OLS (within families, not across), to "correct" the variance estimates of β:

Huber-White Sandwich Estimator (Huber 1967, White 1980):

$$Var_{sand}(\hat{\beta}) = (X'\hat{V}^{-1}X)^{-}\left(\sum_{i=1}^{F} X_i'\hat{V}_i^{-1}(\hat{\varepsilon}_i\ \hat{\varepsilon}_i')\hat{V}_i^{-1}X_i\right)(X'\hat{V}^{-1}X)^{-}$$

Where $\hat{\varepsilon}_i = Y_i - X_i\hat{\beta}$ is vector of OLS residuals in *i*th of F total families, V=Var(Y), and X_i, V_i denote i^{th} block of X and V for that family.

Note: $(\hat{\varepsilon}_i\ \hat{\varepsilon}_i')$ is Variance-Covariance matrix of OLS residuals in *i*th Family

PROC MIXED in SAS using "Sandwich" (a.k.a. "empirical") option with SUBJECT=FAMID (i.e. pedigrees assumed independent) or **SANDWICH** package in R

BMI regressed on RACE in Family Heart Study

PROC REG (ignore clusters)

Param	Estimate	STD Error	T Value	Pr > |t|
Intercept	26.282	0.676	38.85	<.0001
nrace	1.490	0.612	2.43	0.0152

REG Good point estimates (Slopes, intercepts)

REG std errors too low

REG too significant (false positives)

PROC MIXED (Sandwich)

Param	Estimate	STD Error	DF	T Value	Pr > |t|
Intercept	26.441	1.258	665	21.01	<.0001
nrace	1.334	1.194	665	1.12	0.2645

Outline

1. FACT:

   Severe inflation of false-positive signals

2. WHY?

   Spoiler alert pedigrees violate the basic statistical assumption of "i.i.d."

3. **Statistical Methods & Software that correctly analyze Pedigree Data by *modeling the dependences* in pedigree data (result: no inflation of false-positives, retaining power)**

4. Comparisons & Robustness

5. Conclusions & Recommendations

Statistical Methods for Pedigree Data

A. Adjusting for general familial correlations

(both Genetic & Non-Genetic)

1. **Sandwich Estimators** (already talked about)
2. **GEE** (Generalized Estimating Equations)
3. **Family Bootstrap**

B. Adjusting for Genetic correlations only

TWO Basic approaches:

1. **Kinship** (from observed pedigrees)
2. **Genomic CoVariance Matrix** (GWA SNPs, WGS; captures genetic history)

MANY programs (some restricted to selected generalized models)

MLEKIN, KINSHIP2, MMAP, GENESIS, SAIGE, REGENIE, fastGWA-GLMM, FBAT/PBAT, QTDT, …

Almost ALL methods are Generalized Linear Mixed Models

In Regression Setting: Mixed Model

$$Y = X\beta + Z\gamma + \varepsilon$$

Response = Fixed Effects + Random Effects + Residual Err

Y = Response, outcome (data)

(E.G. ln(TG), SBP, CRP, Healthy Aging Index, Omic_feature)

X = Fixed Effects Design Matrix (data)

(E.G. Age, Sex, TimeonRx, SNPs, genomic Principal Components, OMIC features)

β = Regression Coefficients (estimates)

(slopes, intercepts)

Z = Random Effects Design matrix (data)

(E.G. FAMID, ID, ID*TIMEONRx, kinship)

γ = Variance Components (estimates)

(growth curve parameters, heritability)

ε = Residual Error (usually assumed i.i.d.)

Mixed Model

$$Y = X\beta + Z\gamma + \varepsilon$$

We also assume:

$$\begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right)$$

i.e.
G is Var-Cov of γ
R is Var-Cov of ε
γ and ε are independent

Which implies that…

$$VAR[Y] = Z\,G\,Z' + R$$

Part of Model	How to specify in SAS
FIXED effects:	MODEL statement (X)
Random Effects (Var-Comp):	RANDOM statement (Z: G matrix)
Residual Effects:	REPEATED statement (R matrix)

G Matrix (Random Effects)
The "RANDOM" statement in PROC MIXED

For Pedigree data (one timepoint per person)

- Since PEDIGREES independent , the syntax would be:
  RANDOM …… / SUBJECT=PEDID;

Submatrix tells how Members of PEDID=2 are intercorrelated

$$G = \begin{bmatrix} [G_{PEDID=1}] & 0 & 0 & 0 \\ 0 & [G_{PEDID=2}] & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & [G_{PEDID=k}] \end{bmatrix}$$

Alternative Genetic Mixed Models for Pedigrees

- G=KINSHIP matricies
- Based upon pedigree relationships

  1st degree

  2nd degree

  ⋮

  nth degree

- Doesn't need genotypes

Some Mixed Model Pedigree Software:
R programs: MLEKIN, KINSHIP2, etc.
Sand alone: MMAP (O'Connell)

- G=Genomic Covariance matrix
- Based upon measured correlations between people across all measured genotypes in whole dataset

  (e.g. GWAS SNPs or WGS)
- Captures correlations due unknown to consangeous matings in distant past
- Doesn't need pedigrees

"Generalized Linear Models" (GLMs)

(not to be confused with "General Linear Models"-also abbrev as "GLM")

Model Y as a linear combination of Xs with unknown parameters, β, specified by

- a fixed invertible "link function", g such that $E[Y|X] = \mu = g^{-1}(X\beta)$
- with some variance distribution ("error", "unexplained var") $V(Y; \gamma)$

Maximum Likelihood is standard way to estimate/test GLMs

Y	Model	GLM Equation	Link Function	Variance Distribution
Quantitative	Regression (ANOVA)	$E[Y\|X]=X\beta$ $Y=X\beta+\varepsilon,$ where $\varepsilon \sim N[0,\sigma^2]$	$g(\mu)=\mu$ identity	$N[0, \gamma=\sigma^2]$ Normal
Binary {0,1}	Logistic Regression	$P[Y\|X]=\frac{exp(X\beta)}{1+exp(X\beta)}$ $E[Y\|X]=1*P[Y=1\|X]+0*P[Y=0\|X]$ $=P[Y=1\|X]$	$g(\mu)=\ln[\mu/(1-\mu)]$ logit	Binomial
Counts 0,1,2,…	Poisson Regression	$E[Y\|X]=exp(X\beta)$	$g(\mu)=\ln[\mu]$ log	Poisson

NOTE: Cox-Proporotional Hazrds Model not strictly GLM, since it's semi-parametric. But parametric part (model proportional hazard ratios, given non-parametric marginal K-M s… tios)

In SAS **PROC MCMC** supports **non-normal random effects** (i.e. frailties), but can be CPU intensive, sometimes with convergence issues

McCullagh, Peter; Nelder, John (1989). Generalized Linear Models, Second Edition. Boca Raton: Chapman and Hall/CRC. ISBN 0-412-31760-5.

Generalized Estimating Equations (GEE)*

- Works for any Generalized Linear Model
- Semi-parametric
- Uses **SANDWICH** estimator* to fit first 2 moments MVN ~N[mean,Var-Cov] and estimates/tests via ML (called Quasi-likelihood)
- Model selection via AIC
- SAS (PROC GENMOD), R (GEE, GEEPACK), Python (statsmodels)
- Works well for Correlated Data (longitudinal, pedigrees) as long as predictors are **common** (based upon asymptotics)
- Works **poorly** for **rare predictor** effects

*Liang K-Y, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika. 1986;73(1):13–22.

Family Bootstrap
(Borecki & Province, 2008)

- Sampling unit is PEDIGREES not individuals
- Sample Pedigrees w/replacement
- If there are F families total, each bootstrap sample has F families (but possibly different numbers of subjects from sample to sample!)

Borecki IB, Province MA. ***Genetic and genomic discovery using family studies***. Circulation. 2008 Sep 2;118(10):1057-63. doi: 10.1161/CIRCULATIONAHA.107.714592. PMID: 18765388.

RECALL previous example, that **OLS (IGNORING family nature of data)**
Still gives GOOD ESTIMATES of parameters, but underestimates their StdErrs
BMI regressed on RACE in Family Heart Study

PROC REG (ignore clusters)

Param	Estimate	STD Error	T Value	Pr > |t|
Intercept	26.282	0.676	38.85	<.0001
nrace	1.490	0.612	2.43	0.0152

REG Good point estimates
(Slopes, intercepts)

REG std errors too low

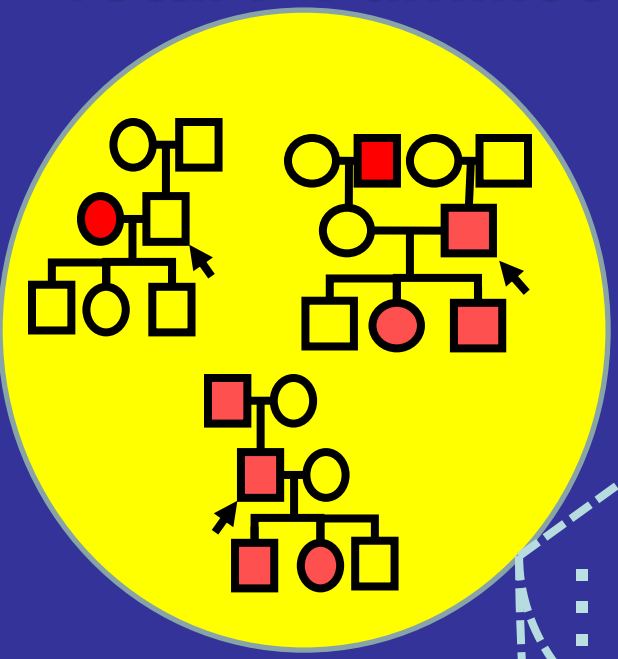
REG too significant (false positives)

PROC MIXED (Sandwich)

Param	Estimate	STD Error	DF	T Value	Pr > |t|
Intercept	26.441	1.258	665	21.01	<.0001
nrace	1.334	1.194	665	1.12	0.2645

Total F Families "B" Bootstrap Samples on Families

Each Sample: F Families (WITH Replacement)



Sample 1

Sample j

Sample B

Analyze each sample as unrelateds: e.g. OLS!

$Y = \alpha_1 + \beta_1 SNP + \cdots$

$Y = \alpha_j + \beta_j SNP + \cdots$

$Y = \alpha_B + \beta_B SNP + \cdots$

Family Boostrap works WELL with ANY Model (GLMMs)

Average Across Bootstraps

$\beta_{Boot} = mean(\beta_j s)$

$s.e.(\beta_{Boot}) = StdDev(\beta_j s)$

$[\beta_{Boot}/s.e.(\beta_{Boot})]^2 \sim \chi_1^2$

$\rightarrow pvalue_{Boot}$

Family Bootstrap

- Works beautifully to control **Type-I error, retaining power**
  (but Needs many pedigrees—e.g. won't work with Amish = 1 super pedigree!)
- "***Gold Standard***" for comparing alternative pedigree methods
- **CPU intensive**
- Many Applications
  - **Borecki IB, Province MA. Genetic and genomic discovery using family studies. Circulation. 2008 Sep 2;118(10):1057-63. doi: 10.1161/CIRCULATIONAHA.107.714592. PMID: 18765388**.
  - Folsom AR, Pankow JS, Williams RR, Evans GW, Province MA, Eckfeldt JH. Fibrinogen, plasminogen activator inhibitor-1, and carotid intima-media wall thickness in the NHLBI Family Heart Study. Thromb Haemost. 1998 Feb;79(2):400-4. PMID: 9493598.
  - Djoussé L, Myers RH, Coon H, Arnett DK, Province MA, Ellison RC. Smoking influences the association between apolipoprotein E and lipids: the National Heart, Lung, and Blood Institute Family Heart Study. Lipids. 2000 Aug;35(8):827-31. doi: 10.1007/s11745-000-0591-1. PMID: 10984105.
  - Zhang Q, Feitosa M, Borecki IB. Estimating and testing pleiotropy of single genetic variant for two quantitative traits. Genet Epidemiol. 2014 Sep;38(6):523-30. doi: 10.1002/gepi.21837. Epub 2014 Jul 12. PMID: 25044106; PMCID: PMC4169079.
  - Feitosa M, Kuipers A, Wojczynski M, Wang L, Perls T, Christensen K, Zmuda J, Province M. Long Life Family Study Shows Reduced Coronary Artery Disease Despite High Polygenic Hazard Scores. Innov Aging. 2020 Dec 16;4(Suppl 1):212. doi: 10.1093/geroni/igaa057.685. PMCID: PMC7741003.
  - Feitosa MF, Kuipers AL, Wojczynski MK, Wang L, Barinas-Mitchell E, Kulminski AM, Thyagarajan B, Lee JH, Perls T, Christensen K, Newman AB, Zmuda JM, Province MA. Heterogeneity of the Predictive Polygenic Risk Scores for Coronary Heart Disease Age-at-Onset in Three Different Coronary Heart Disease Family-Based Ascertainments. Circ Genom Precis Med. 2021 Jun;14(3):e003201. doi: 10.1161/CIRCGEN.120.003201. Epub 2021 Apr 12. PMID: 33844929; PMCID: PMC8214825.
  - Song Z, Gunn S, Monti S, Peloso GM, Liu CT, Lunetta K, Sebastiani P. Learning Gaussian Graphical Models from Correlated Data. bioRxiv [Preprint]. 2024 Apr 5:2024.04.03.587948. doi: 10.1101/2024.04.03.587948. PMID: 38617340; PMCID: PMC11014549.

Outline

1. FACT:

   Severe inflation of false-positive signals

2. WHY?

   Spoiler alert pedigrees violate the basic statistical assumption of "i.i.d."

3. Statistical Methods & Software

4. **Comparisons & Robustness** of Pedigree Models

5. Conclusions & Recommendations

Various Pedigree Association Models

1. **FBAT (Family-Based Association Test)**
2. **QTDT (Quantitative TDT)**
3. **Sandwich Estimator for Fams**
4. **Family Bootstrap**

How well do they perform against one another (Monte Carlo simulation)?

Borecki IB, Province MA. Genetic and genomic discovery using family studies. Circulation. 2008 Sep 2;118(10):1057-63. doi: 10.1161/CIRCULATIONAHA.107.714592. PMID: 18765388.

Simulated Family Data

- **200 Nuclear Families with 2 children**
  - **800 Individuals**
  - **In parents: Pr(AA)=0.25, Pr(AG)=0.5, Pr(GG)=0.25**

- **Simulated Phenotype = $\alpha + \beta$*SNP Value + PG + ε**

  Value(AA)=0, Value(AG)=0.5, Value(GG)=1

  3 Cases:

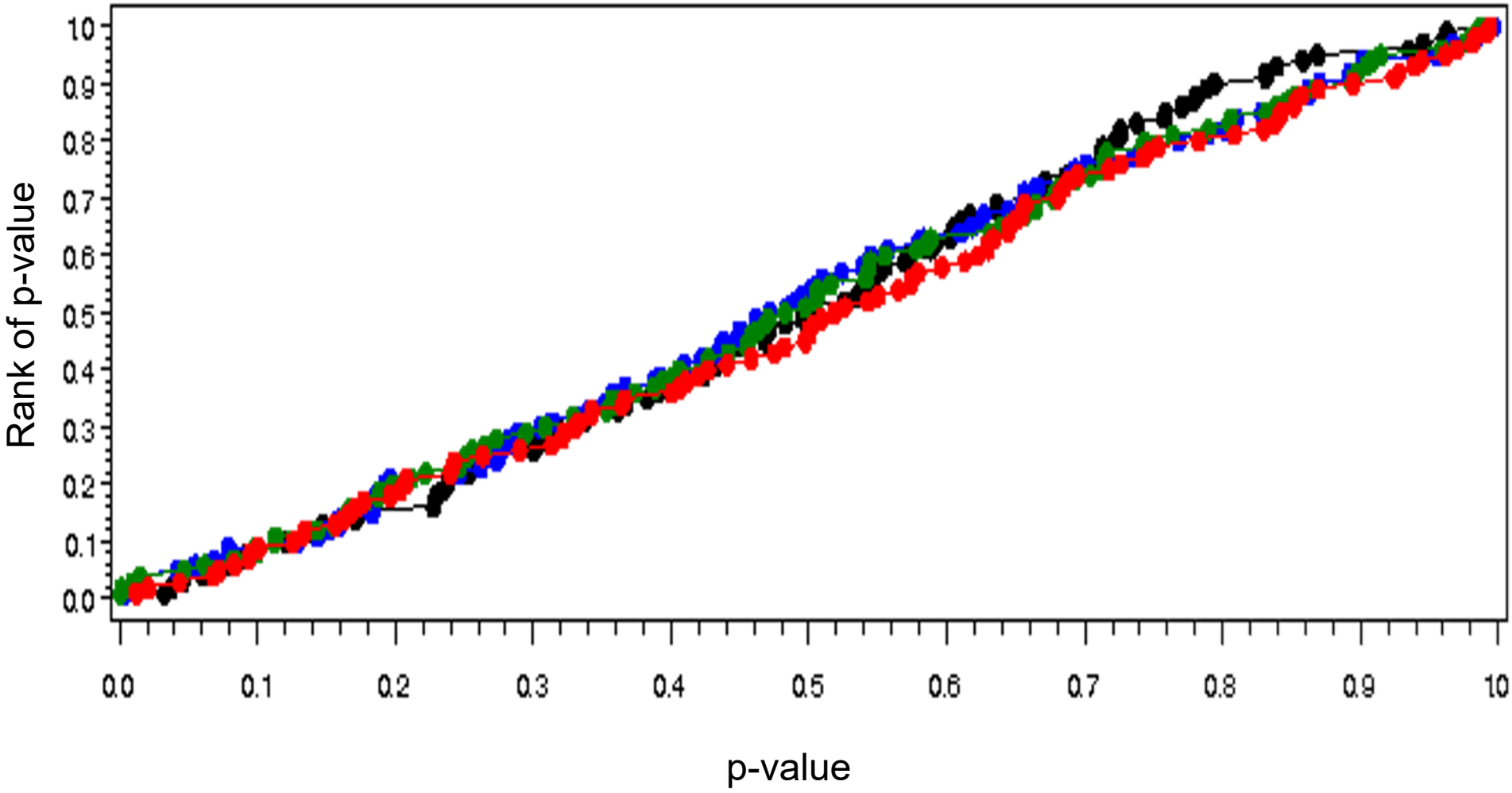
  1. **H0: $\alpha=0$, $\beta=0$**
  2. **H1: $\alpha=0$, $\beta=0.5$**
  3. **H0 w/Population Stratification: 2 groups($\alpha_1=0$, $\alpha_2=5$), $\beta=0$**

- **200 Replications**

Q-Q Plots for 4 Family Geno-Pheno Association Methods

Under H0

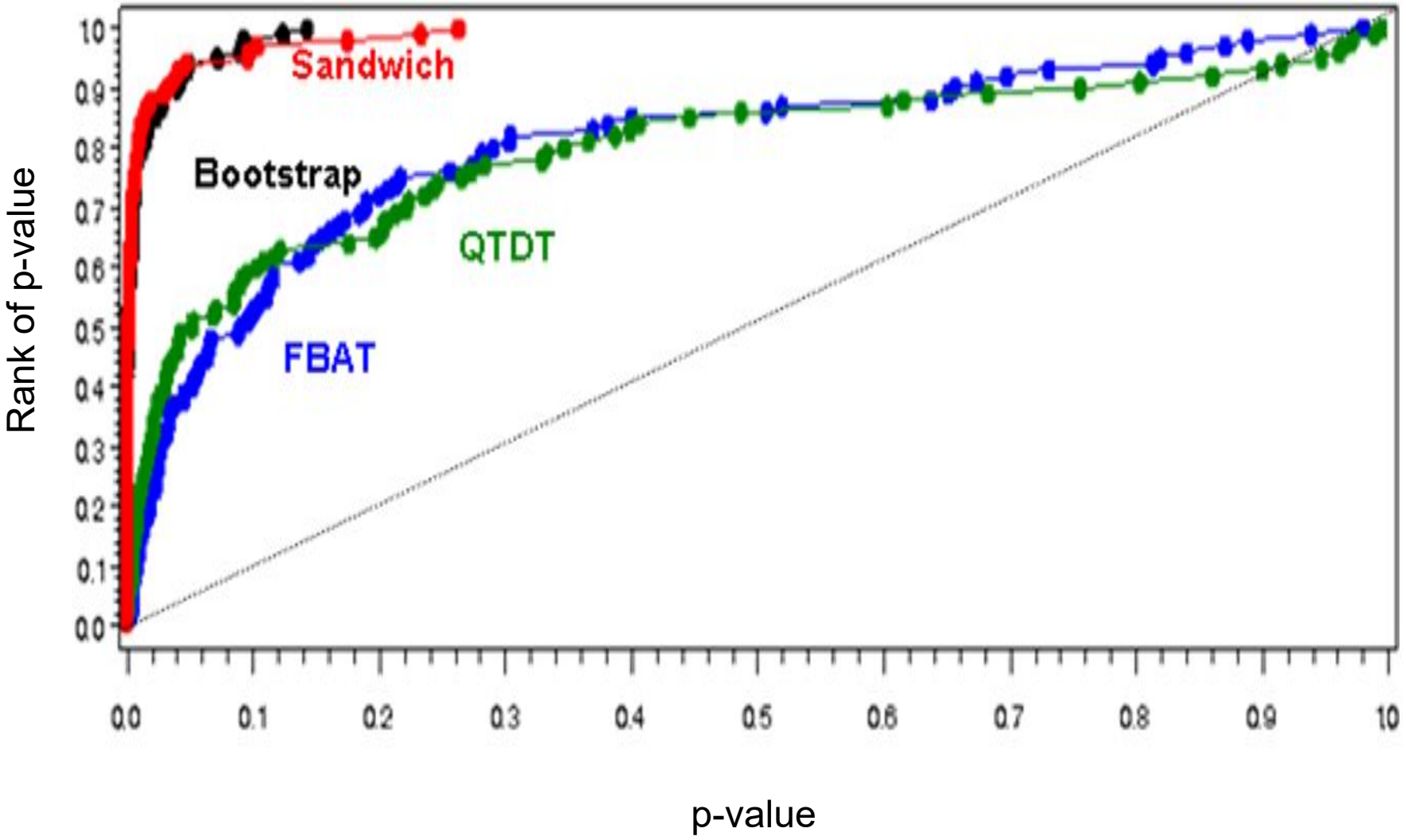
(all 4 methods protect against false positives analyzing families)



Q-Q Plots for 4 Family Geno-Pheno Association Methods

Under H1

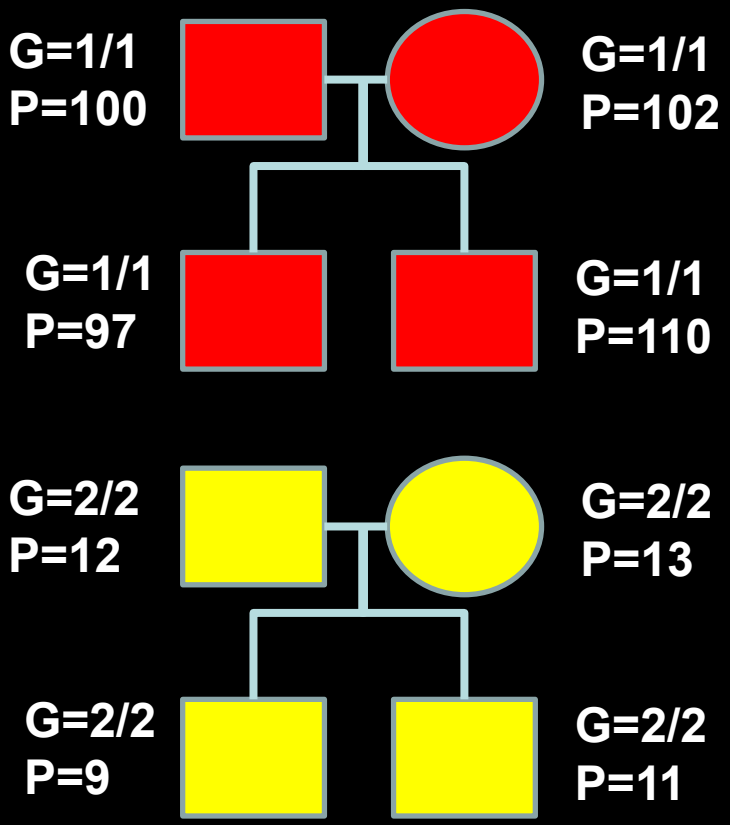
(Sandwich/Bootstrap MUCH more powerful than FBAT/QTDT). WHY?



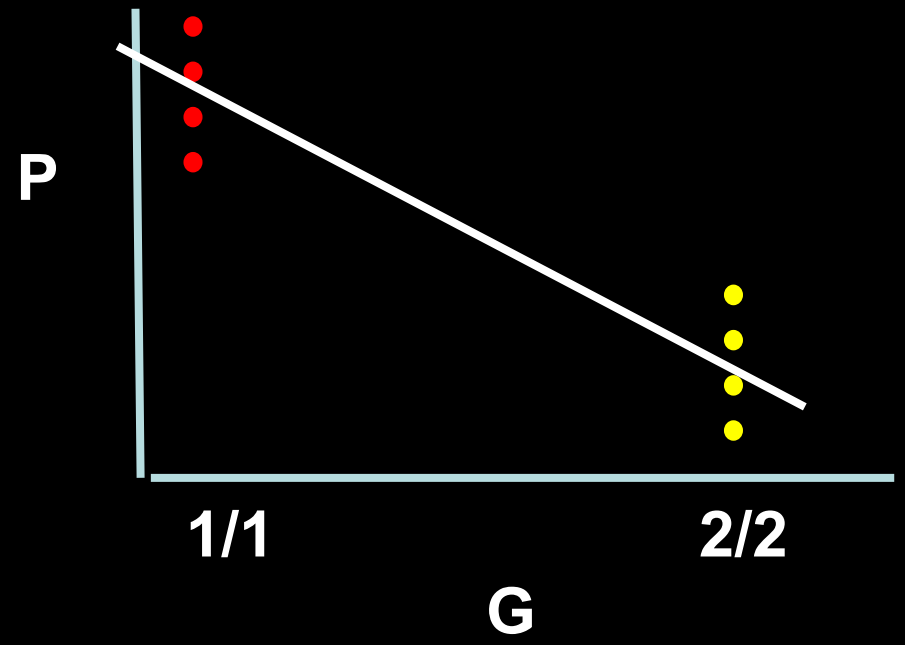
Q: Why the power loss for FBAT & QTDT?

A: These two are both TRANSMISSION tests, NOT true ASSOCIATION tests (despite the FBAT name)

- FBAT ignores the phenotypes of parents, which can contribute valuable phenotype-genotype correlation information in true association models
- FBAT deletes entire families for which transmission is ambiguous:



E.G. These two families show STRONG G-P association, but are deleted by FBAT because transmission is ambiguous

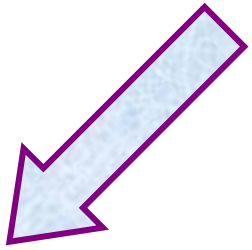


Simulated Family Data: Population Stratification

$$Y = \alpha + \beta X + \varepsilon$$

- Population A
  - 100 Families
  - Freq(AA) = 0.25
  - Freq(AG) = 0.5
  - Freq(GG) = 0.25
  - **$\alpha = 0$**
  - $\beta = 0$
- Population B
  - 100 Families
  - Freq(AA) = 0.09
  - Freq(AG) = 0.42
  - Freq(GG) = 0.49
  - **$\alpha = 5$**
  - $\beta = 0$

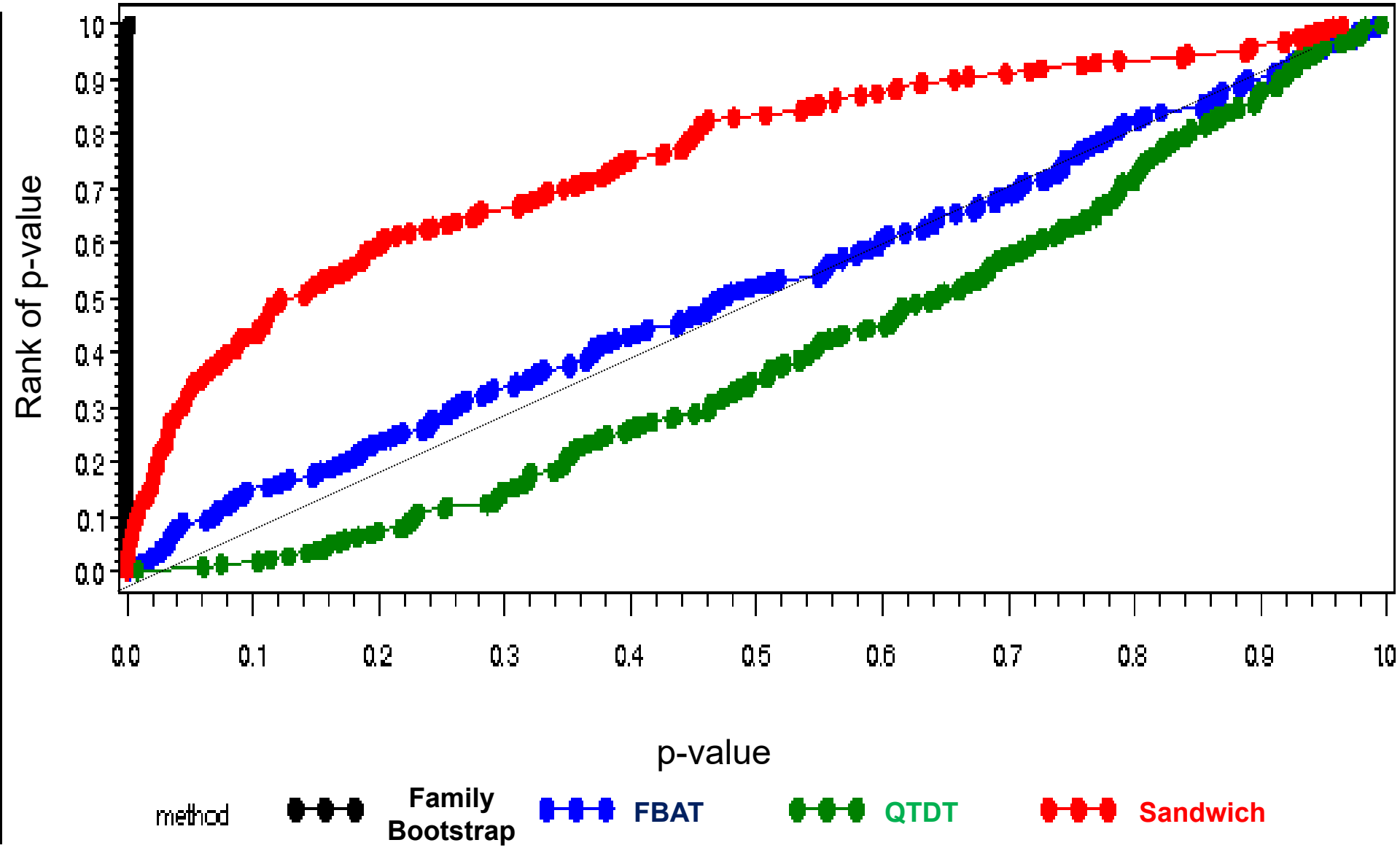
Stratified Population
(False Positive Association between Y & X)



Q-Q Plots for 4 Family Geno-Pheno Association Methods

Under H0 (w/Population Stratification)

(FBAT protects well; QTDT overprotects; Sandwich somewhat; Bootstrap not at all)



Pedigree Association Models (Common alleles)

1. FBAT (Family-Based Association Test)
   - Misnamed! Really a Transmission based test, like QTDT!
   - Deletes fams where transmission ambiguous

2. QTDT (Quantitative TDT)
   - Are transmitted alleles correlated to phenotype?
   - Similar to FBAT

3. Sandwich Estimator for Fams
   - Similar to GEE model
   - Some protection from Population Stratification*
   - Fast

4. Family Bootstrap
   - Valid under homogeneous H0 for all MAF
   - No protection from Population Stratification*
   - CPU intensive

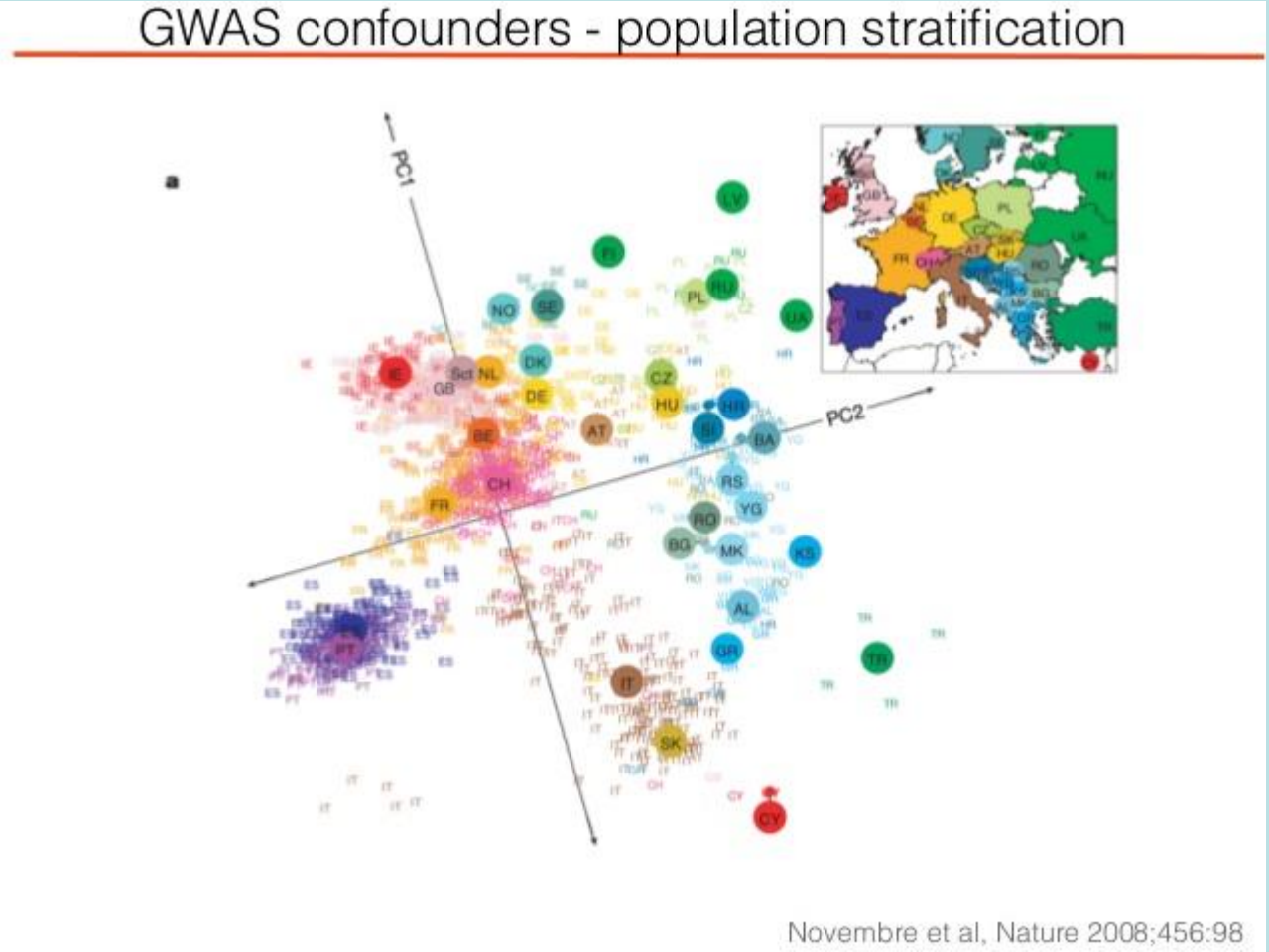
*But genomic Principal Component covariates work best against Population Stratification for all methods!
Don't need the tests themselves to correct for it (loses power)

Controlling for Genetic Drift in Analysis

EIGENSTRAT:

- Generate top **Principal Components from GWAS SNPs** to capture "drift" of Major Common Haplotypes.
- Use these as **Covariates PC1, PC2, …, PCn** to "correct" for Population Stratification In mixed regression (logistic, cox) models:

$$Y = (\alpha + \beta 1*PC1 + \beta 2*PC2 + \ldots + \beta n*PCn) + \beta*SNP + e$$

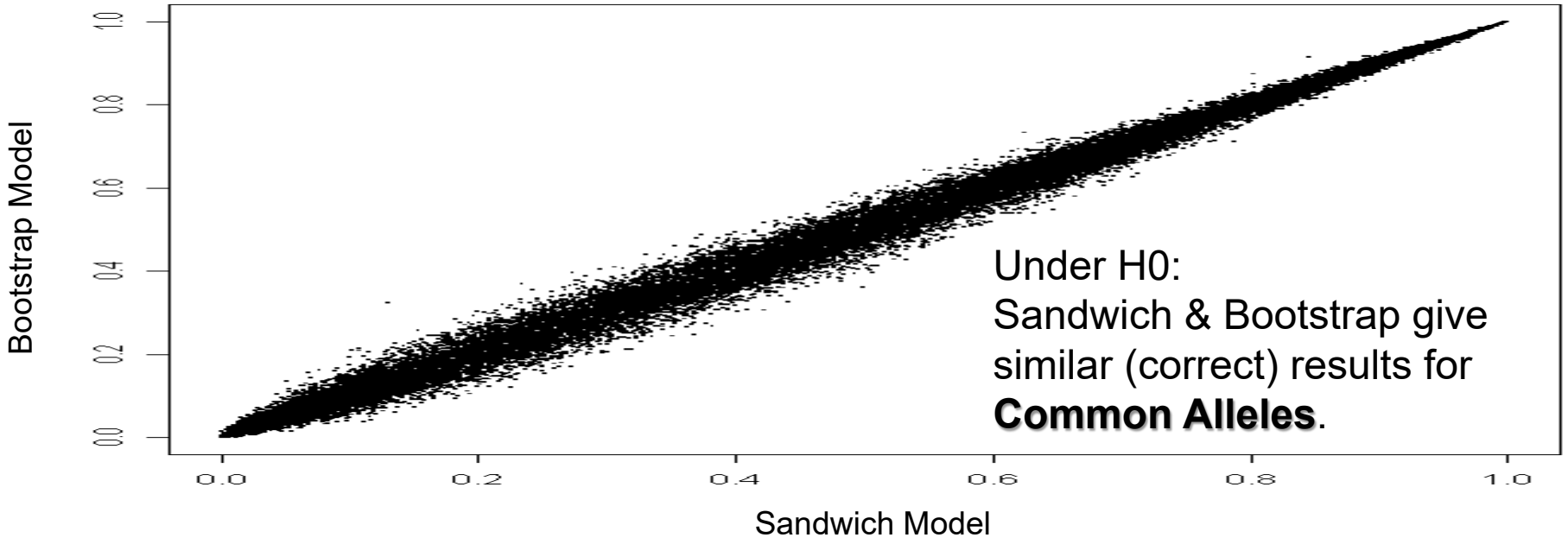


Q: what about RARE alleles we get from sequencing (also Rare OMIC features)?

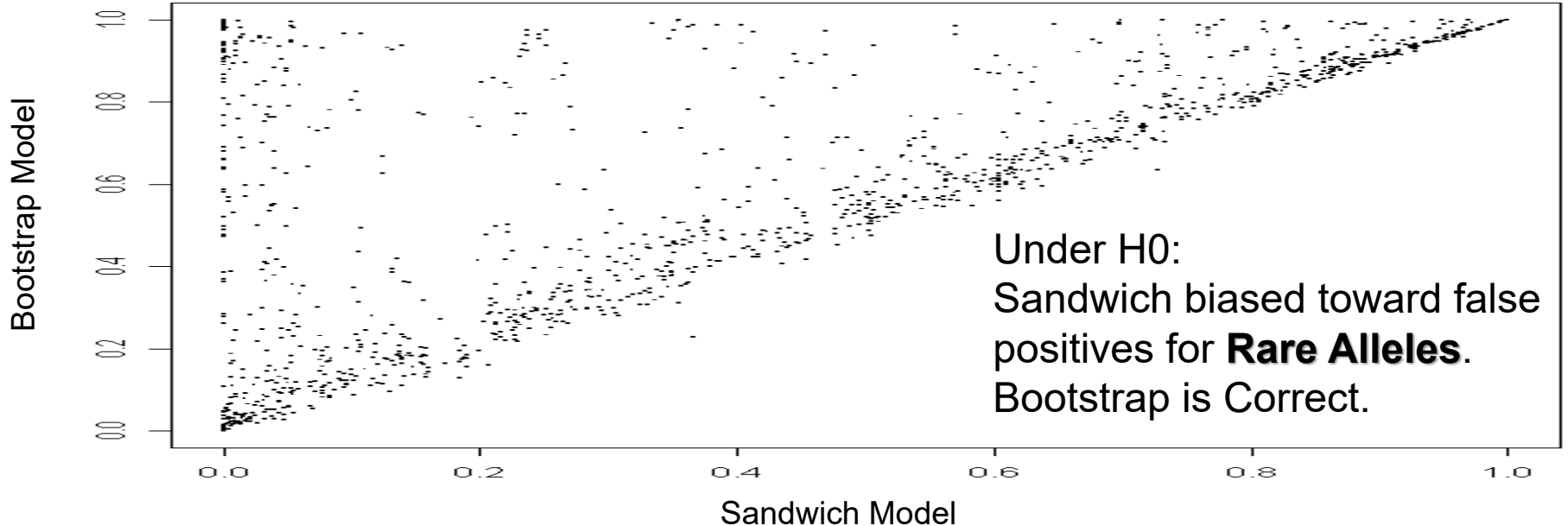
A: Some Methods that work WELL for common effects, work POORLY for rare ones

Mixed Model (Sandwich) vs. Family Bootstrap p-values in GWAS of Family Heart Study

MAF>0.01



MAF<0.01



Modification of the Sandwich Estimator in Generalized Estimating Equations with Correlated Binary Outcomes in Rare Event and Small Sample Settings

Paul Rogers, **Julie Stoner**

American Journal of Applied Mathematics and Statistics. **2015**, 3(6), 243-251. DOI: 10.12691/ajams-3-6-5
Published online: November 23, 2015

Abstract

Regression models for correlated binary outcomes are commonly fit using a Generalized Estimating Equations (GEE) methodology. GEE uses the Liang and Zeger sandwich estimator to produce unbiased standard error estimators for regression coefficients in large sample settings even when the covariance structure is misspecified. The sandwich estimator performs optimally in balanced designs when the number of participants is large, and there are few repeated measurements. The sandwich estimator is not without drawbacks; its asymptotic properties do not hold in small sample settings. In these situations, the sandwich estimator is biased downwards, underestimating the variances. In this project, a modified form for the sandwich estimator is proposed to correct this deficiency. The performance of this new sandwich estimator is compared to the traditional Liang and Zeger estimator as

STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:715–722

GEE Analysis of negatively correlated binary responses: a caution

James A. Hanley[1,2,*,†], Abdissa Negassa[3] and Michael D. deB. Edwardes[2]

[1]*Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada*
[2]*Division of Clinical Epidemiology, Royal Victoria Hospital, Montreal, Canada*
[3]*Division of Epidemiology, Department of Oncology, McGill University, Montreal, Canada*

SUMMARY

The method of generalized estimating equations has become almost standard for analysing longitudinal and other correlated response data. However, we have found that if binary responses have less than binomial variation over clusters, and are modelled using exchangeable correlations, prevailing software implementations may give unreliable results. Bounding the negative correlation away from its theoretical minimum may not always be a satisfactory solution. In such instances, using the independence working correlation structure and robust SEs is a more trustworthy alternative. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

The generalized estimating equations (GEE) approach[1,2] has become the method of choice for analysing longitudinal and other correlated response data. It is now available in most statistical packages,[3–5] but some users use their own implementations, or rely on older macros.[6,7]

In this note we relate our experience when we used a cluster sample involving binary responses[8] to explain the essence of the GEE approach to non-statisticians. We chose the example to allow them to compare the GEE estimate of a proportion, and its standard error (SE), with those

The issue of the working correlation matrix $R(\alpha)$ on the GEE model estimators has been discussed; however, little is known about the impact of the working correlation matrix on the GEE GoF statistics. Pan (2002a) reported that models with an independent working correlation matrix had better results than with an exchangeable correlation matrix for models with more than two time-point outcomes and with time-varying covariates. Barnhart and Williamson (1998) reported that their proposed GoF statistics had different performance with different working correlation matrices.

Misspecification of the working correlation matrix $R(\alpha)$ did affect the performance of the GEE GoF statistics. Type I error rates were inflated for those GoF statistics with misspecified working correlations (Q_{Rm}, Q_{m} and S_{m}), and the inflation increased with increasing magnitude of the departure from the misspecified working correlation (identity matrix) to the correctly specified working correlation.

Artificially high power was shown in some GEE GoF statistics with misspecified working correlations due to inflated Type I error rates. Statistics with correctly specified working correlation matrices tended to have better performance (reasonable Type I error rates and higher power) in most cases. Nevertheless, there were some special cases. Q_{R} and S were robust to the working correlation matrix $R(\alpha)$ selection in detecting the omission of an interaction between a cluster-specific binary covariate and a cluster-specific continuous covariate (Model 3). Therefore, using a data-based within-cluster correlation was recommended because it may be closer to the true underlying working correlation than an arbitrary correlation matrix, such as an identity matrix. This was especially true for the Horton statistics (S). Using the Horton statistic with a misspecified (identity) working correlation matrix (S_{m}) may cause extremely inflated Type I error rates, especially for the effect of detecting the effect of the omission of observation-specific covariates.

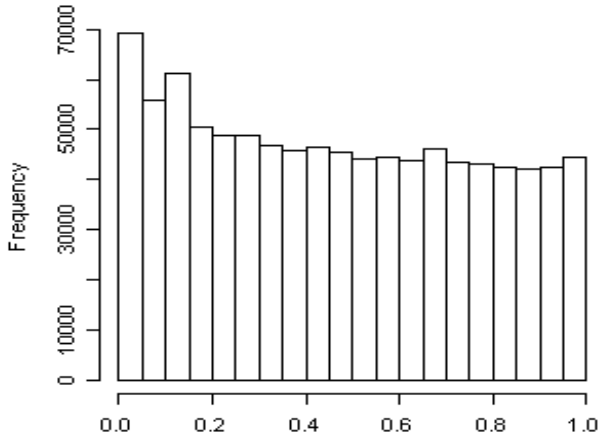
P-value Distribution

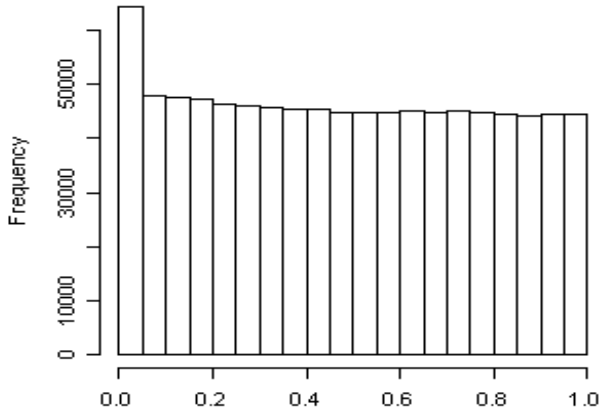
Analysis of Quantitative Data

OLS Regular regression (family ignored)

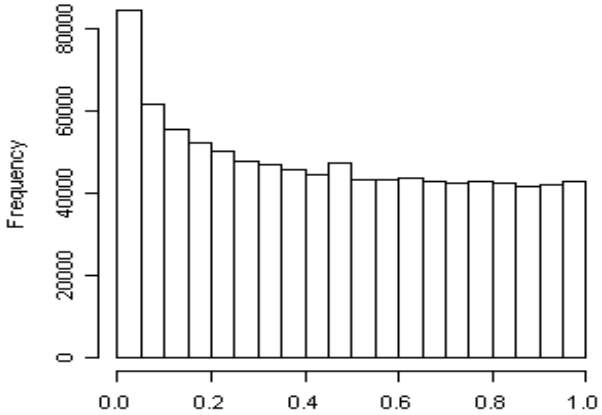
GEE Regression (Generalized Estimating Equations)

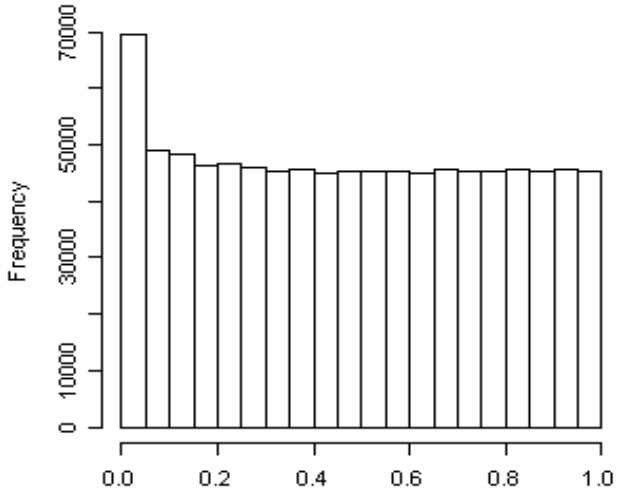
GeneSTAR Data (epinephrine)





Simulated Data

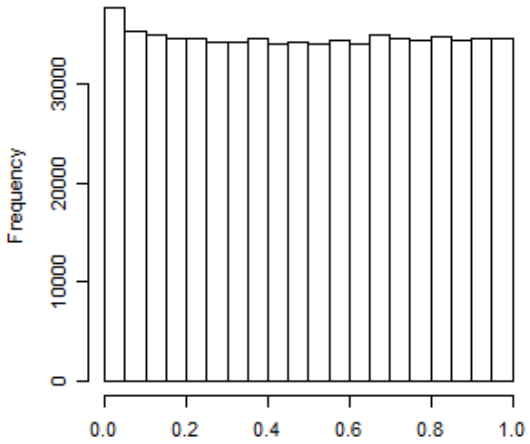




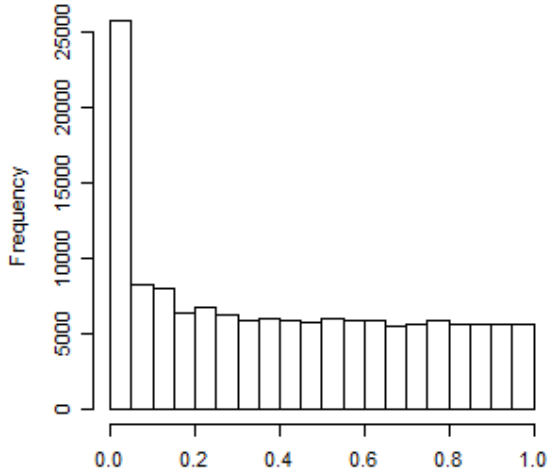
P-value Diagnosis of GEE

(for Quantitative Trait)

P-value Distribution

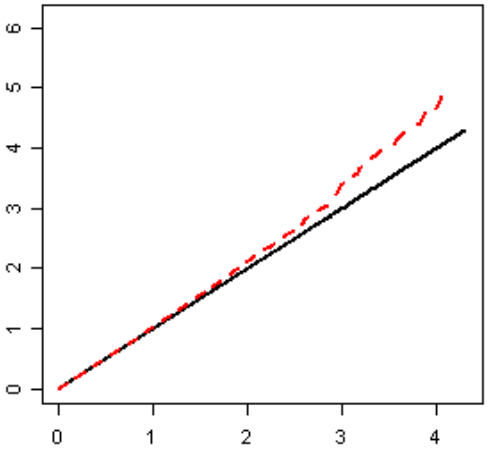


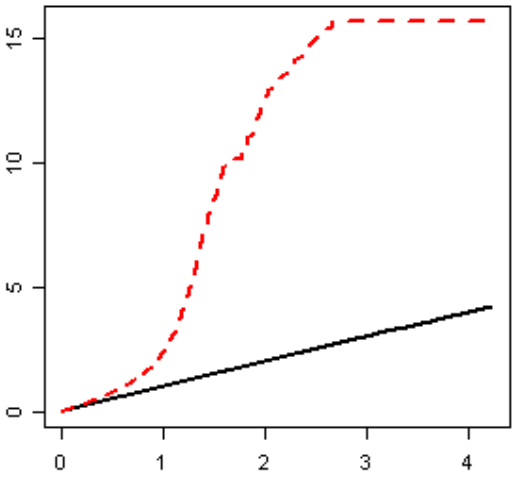
MAF>0.05



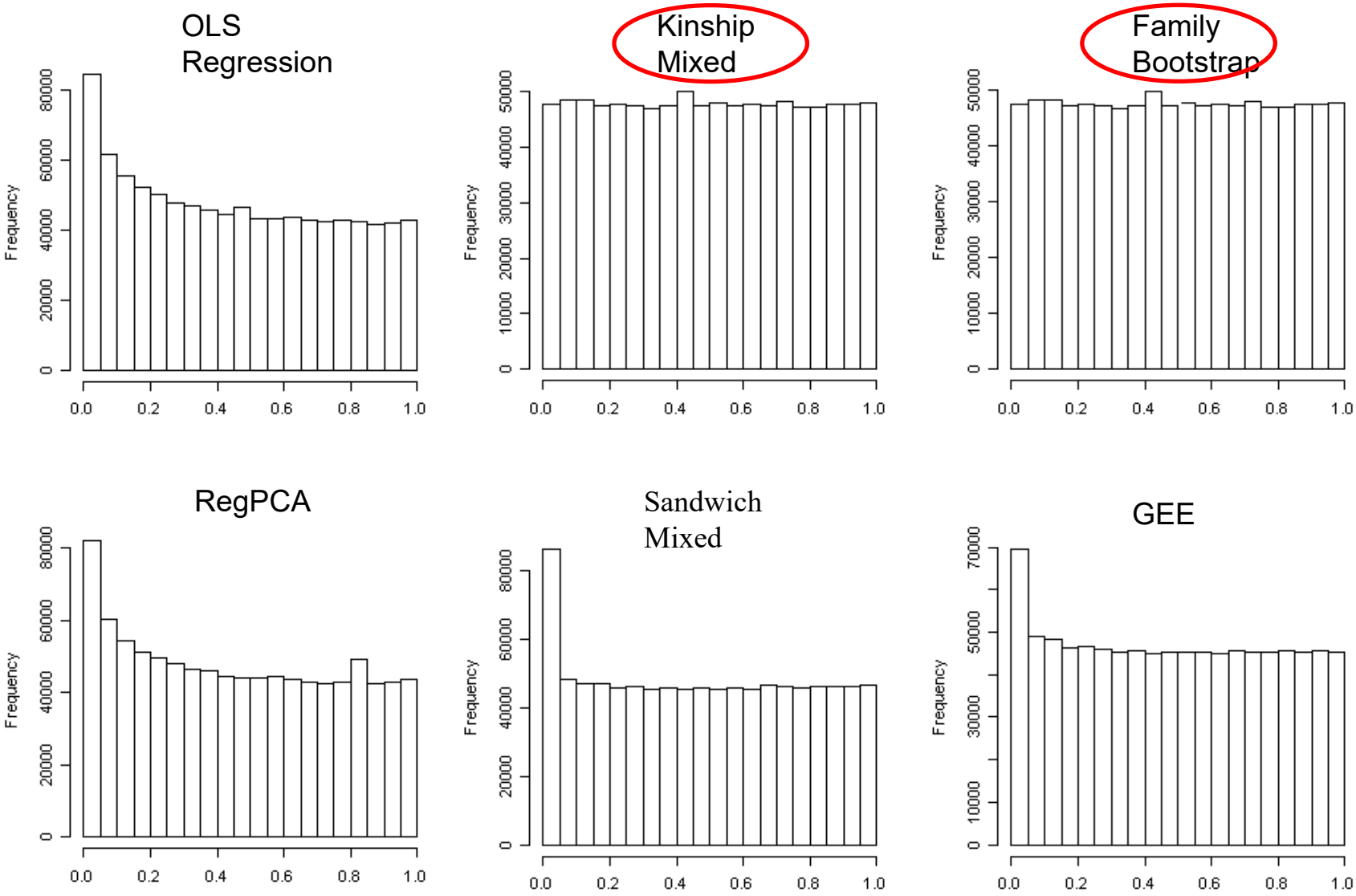
MAF<0.05

-LOG10(P) QQ Plot





P-values Distributions of Different Methods (for MAF <0.05)*



*NOTE: Exact MAF when "Rare" is problem is SAMPLE-SIZE dependent

Inflation of P-values by Method for Burden tests of rare variants in families

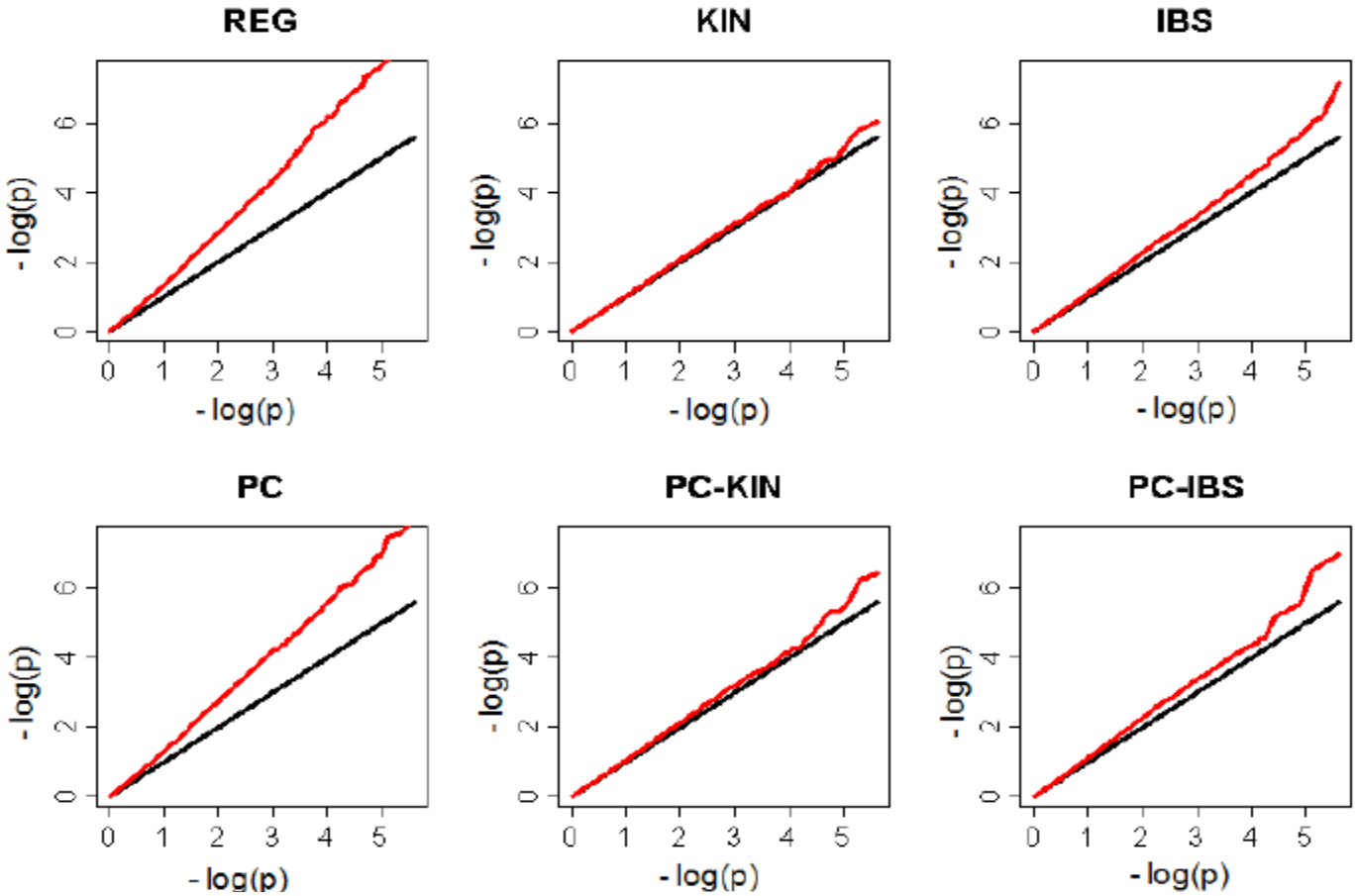


Figure 1 Q-Q plots for six different methods. Q-Q plots of $-\log_{10}$ scaled *p*-values for six different methods based on 1,940 genes from 697 subjects (8 extended families) and 200 replications of quantitative trait Q2 simulated by GAW17 under the null hypothesis. Red curves, observed; black curves, expected.

Zhang, Chung, Kraja, Borecki, Province. Methods for adjusting population structure and familial relatedness in association test for collective effect of multiple rare variants on quantitative traits BMC Proceedings, Vol 28, 2011

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fgene.2022.897210

OPEN ACCESS

EDITED BY
Riyan Cheng,
University of California, San Diego,
United States

REVIEWED BY
Wenjian Bi,
School of Basic Medical Sciences,
Health Science Centre, Peking
University, China
Rounak Dey,
School of Public Health and Harvard
University, United States

*CORRESPONDENCE
Anastasia Gurinovich,
agurinovich@tuftsmedicalcenter.org

SPECIALTY SECTION
This article was submitted to Statistical
Genetics and Methodology,
a section of the journal
Frontiers in Genetics

RECEIVED 15 March 2022
ACCEPTED 08 August 2022
PUBLISHED 23 September 2022

CITATION
Gurinovich A, Li M, Leshchyk A, Bae H

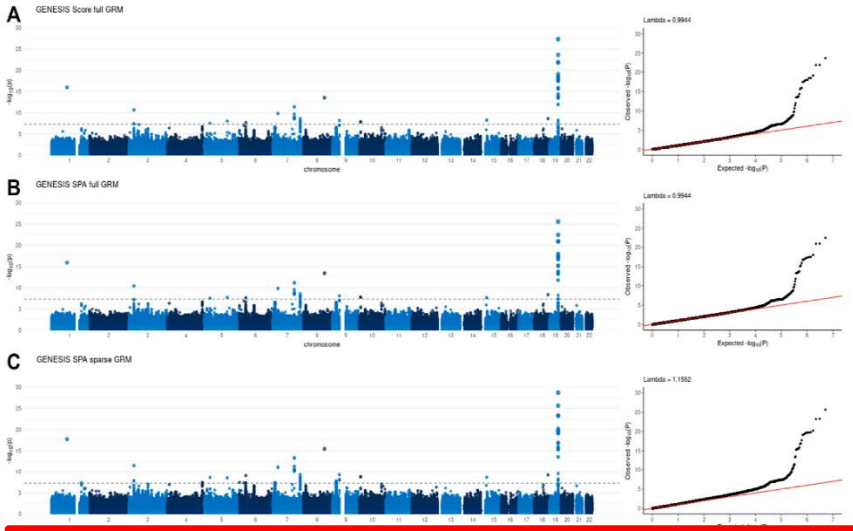
Evaluation of GENESIS, SAIGE, REGENIE and fastGWA-GLMM for genome-wide association studies of binary traits in correlated data

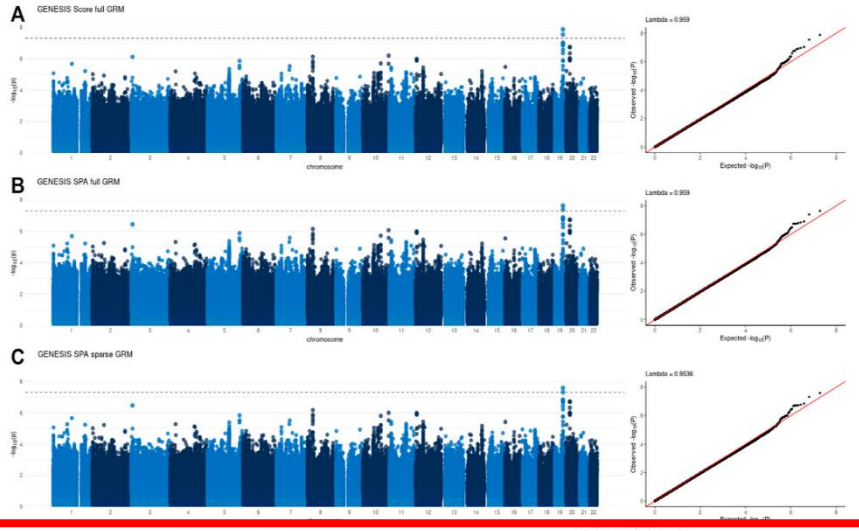
Anastasia Gurinovich[1]*, Mengze Li[2], Anastasia Leshchyk[2], Harold Bae[3], Zeyuan Song[4], Konstantin G. Arbeev[5], Marianne Nygaard[6], Mary F Feitosa[7], Thomas T Perls[8] and Paola Sebastiani[1]

[1]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, United States, [2]Bioinformatics Program, Boston University, Boston, MA, United States, [3]Biostatistics Program, College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, United States, [4]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States, [5]Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, NC, United States, [6]Epidemiology, Biostatistics and Biodemography, Department of Public Health, University of Southern Denmark, Odense, Denmark, [7]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St Louis, MO, United States, [8]Department of Medicine, Geriatrics Section, Boston University School of Medicine, Boston, MA, United States

Last paragraph of Abstract:

"The evaluation suggests that REGENIE might not be a good choice when analyzing correlated data of a small size. fastGWA-GLMM is the most computationally efficient compared to the other three tools, but it appears to be overly conservative when applied to family-based data. GENESIS, SAIGE and fastGWA-GLMM produced similar, although not identical, results, with SPA adjustment performing better than Score tests. Our evaluation also demonstrates the importance of adjusting by full GRM in highly correlated datasets when using GENESIS or SAIGE."





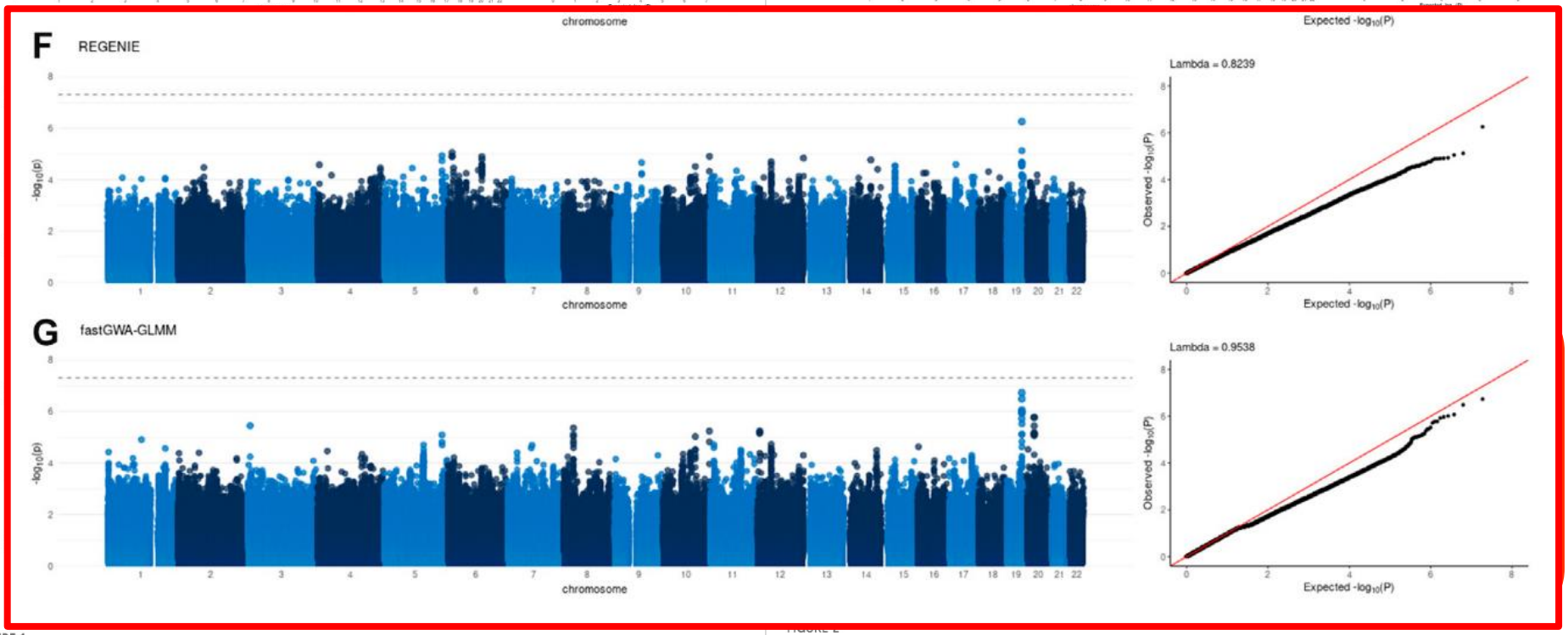


FIGURE 1
Manhattan and QQ plots of -log10(*p*-values) for the associations using imputed genotype data in the New England Centenarian Study (NECS) data. Panel **(A)**: associations based on the score test and adjusted for the full genetic relation matrix (GRM) using GENESIS. Panel **(B)**: associations based on the SPA and adjusted for the full GRM using GENESIS. Panel **(C)**: associations based on the SPA and adjusted for the sparse GRM using GENESIS. Panel **(D)**: associations based on the SPA and adjusted for the full GRM using SAIGE. Panel **(E)**: associations based on the SPA and adjusted for the sparse GRM using SAIGE. Panel **(F)**: associations based on the SPA and polygenic effect estimates to control for relatedness using REGENIE. Panel **(G)**: associations based on the SPA and adjusted for the sparse GRM using fastGWA-GLMM. Lambda is a genomic inflation factor.

FIGURE 2
Manhattan and QQ plots of -log10(*p*-values) for the associations using WGS data in the Long Life Family Study (LLFS) data. Panel **(A)**: associations based on the score test and adjusted for the full genetic relation matrix (GRM) using GENESIS. Panel **(B)**: associations based on the SPA and adjusted for the full GRM using GENESIS. Panel **(C)**: associations based on the SPA and adjusted for the sparse GRM using GENESIS. Panel **(D)**: associations based on the SPA and adjusted for the full GRM using SAIGE. Panel **(E)**: associations based on the SPA and adjusted for the sparse GRM using SAIGE. Panel **(F)**: associations based on the SPA and polygenic effect estimates to control for relatedness using REGENIE. Panel **(G)**: associations based on the SPA and adjusted for the sparse GRM using fastGWA-GLMM. Lambda is a genomic inflation factor.

Correlations within Pedigrees are **not the only important confounders** in OMIC analyses that can inflate false-positives

Iterson *et al. Genome Biology* (2017) 18:19
DOI 10.1186/s13059-016-1131-9

Genome Biology

METHOD Open Access

Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution

Maarten van Iterson[1*], Erik W. van Zwet[2], the BIOS Consortium and Bastiaan T. Heijmans[1]

Abstract

We show that epigenome- and transcriptome-wide association studies (EWAS and TWAS) are prone to significant inflation and bias of test statistics, an unrecognized phenomenon introducing spurious findings if left unaddressed. Neither GWAS-based methodology nor state-of-the-art confounder adjustment methods completely remove bias and inflation. We propose a Bayesian method to control bias and inflation in EWAS and TWAS based on estimation of the empirical null distribution. Using simulations and real data, we demonstrate that our method maximizes power while properly controlling the false positive rate. We illustrate the utility of our method in large-scale EWAS and TWAS meta-analyses of age and smoking.

Keywords: Epigenome- and transcriptome-wide association studies, Bias, Inflation, Empirical null distribution, Gibbs sampler, Meta-analysis

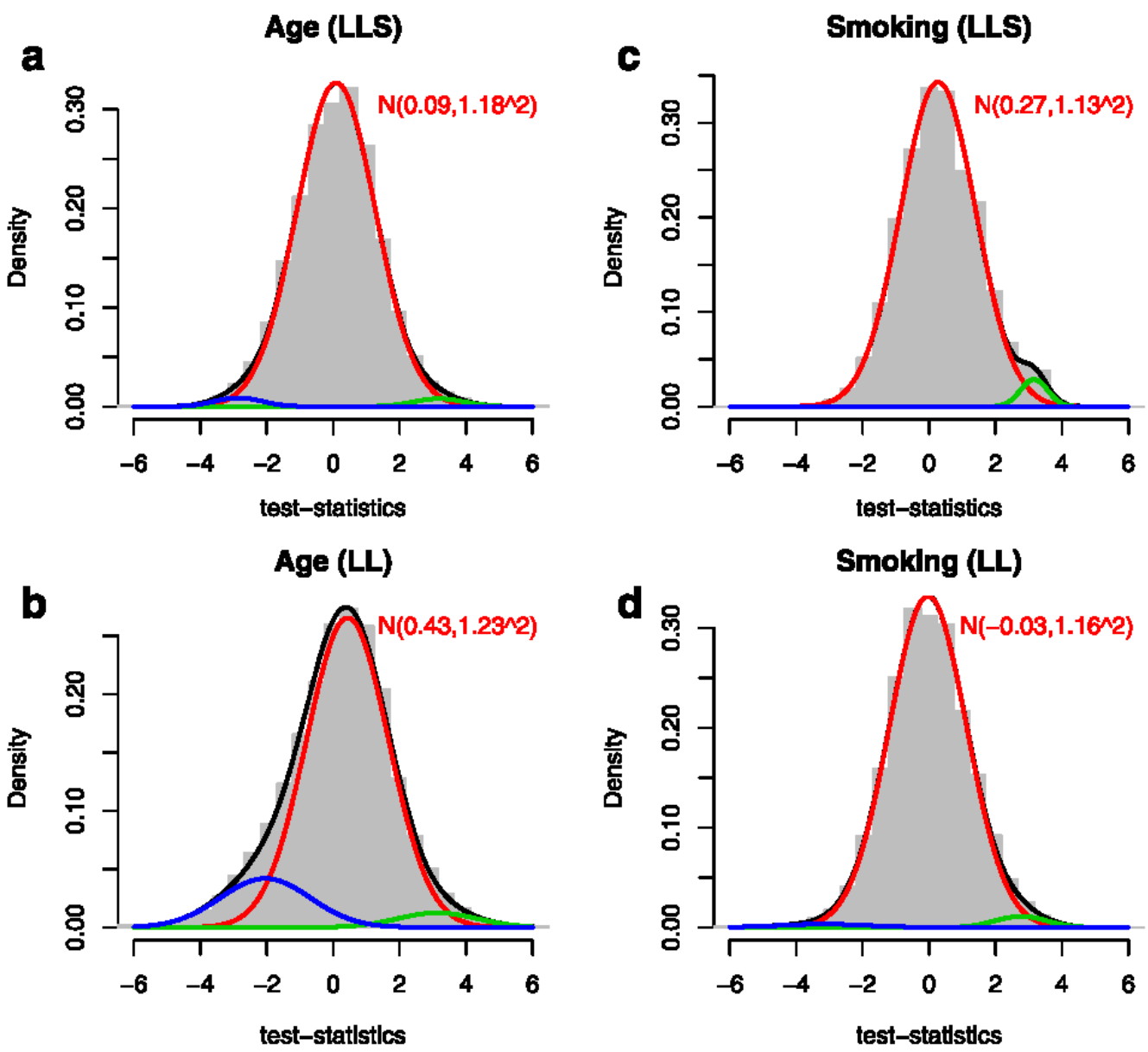


Fig. 4 Histogram of test statistics for TWAS on age (**a** and **b**) and smoking status (**c** and **d**) performed on two cohorts: LifeLines (*LL*) and Leiden Longevity Study (*LLS*). The lines represent the three-component normal mixture fitted as estimated using our Bayesian method. The *black line* represents the fit of the mixture, the red line the fit of the null component (the empirical null distribution with estimated mean and variance reported). The *blue and green lines* represent the estimated fits of the alternative components (proportion of positively and negatively associated features)

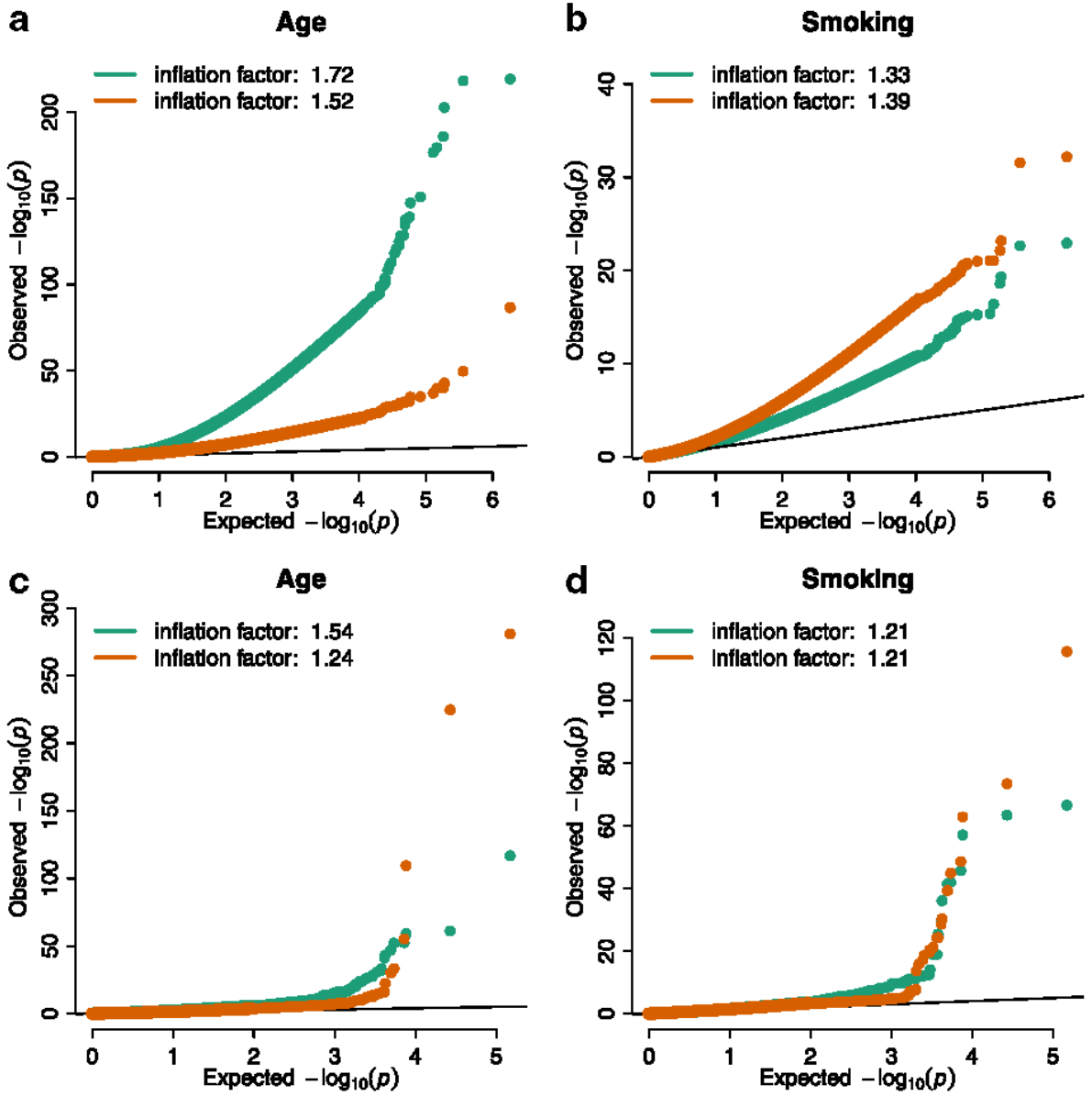


Fig. 1 Inflated epigenome- and transcriptome-wide association studies. Quantile-quantile (*QQ*) plots for EWAS (panels **a** and **b**) and TWAS (panels **c** and **d**) performed on the LifeLines (*LL*) and Leiden Longevity Study (*LLS*) cohorts for the phenotypes age and smoking status. Results for LL are indicated in *green* and LLS in *orange*. QQ-plots show the observed minus $\log_{10}$-transformed *P* values obtained from a linear model corrected for known biological and technical covariates against quantiles from the theoretical null distribution. Strong inflation, as estimated according to $\lambda_{\chi^2_1}$ [9], was observed for both EWAS and TWAS of age, while for the EWAS and TWAS of smoking the amount of inflation is smaller (notice different y-axis scales)

www.nature.com/scientificreports

scientific reports

OPEN

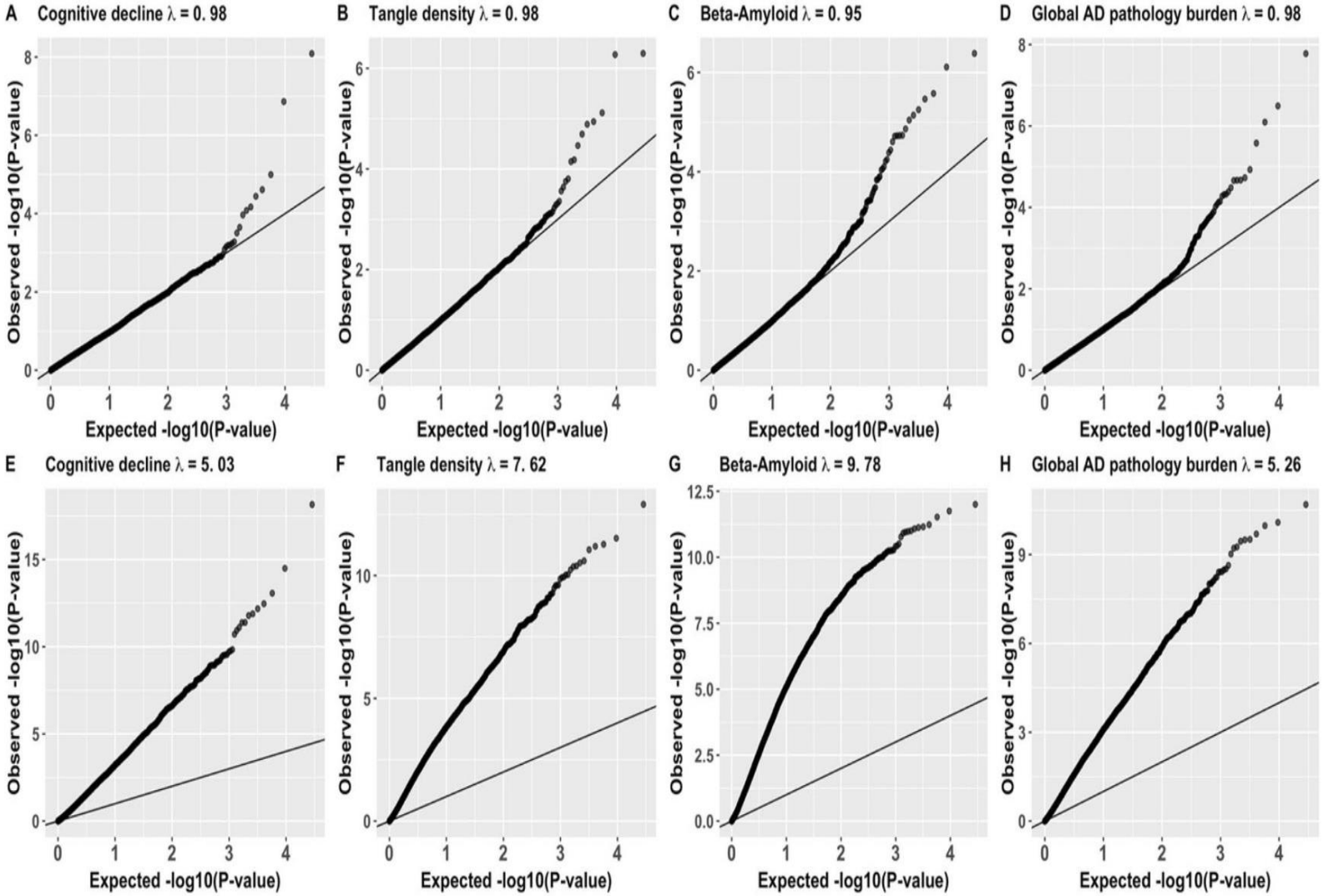
Differential gene expression analysis based on linear mixed model corrects false positive inflation for studying quantitative traits

Shizhen Tang[1,2], Aron S. Buchman[3], Yanling Wang[3], Denis Avey[3], Jishu Xu[3], Shinya Tasaki[3],

Genome-wide Efficient Mixed Model Association (GEMMA)

Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824. https://
doi. org/ 10. 1038/ ng. 2310 (2012).

- Uses **fixed effect confounding covariates** such as sex, age, and postmortem interval, etc.
- Uses **sample-specific mixed effect** term derived from full-rank **sample-sample correlation matrix** (based on all gene expressions)



GEMMA

OLS Reg

Figure 1. QQ-plots and genomic control factors of DGE results by LMM (**A–D**) and standard linear regression model (**E–H**) with the discovery RNA-Seq data of DLPFC tissue of cognitive decline and three AD-NC traits.

EPIGENETICS
2020, VOL. 15, NO. 11, 1163–1166
https://doi.org/10.1080/15592294.2020.1758382

Taylor & Francis
Taylor & Francis Group

BRIEF REPORT

Test-statistic inflation in methylome-wide association studies

Jerry Guintivano[a], Andrey A Shabalin [b], Robin F. Chan [c], David R. Rubinow[a], Patrick F. Sullivan[a,d,e], Samantha Meltzer-Brody[a], Karolina a Aberg [c], and Edwin J. C. G. van den Oord[c]

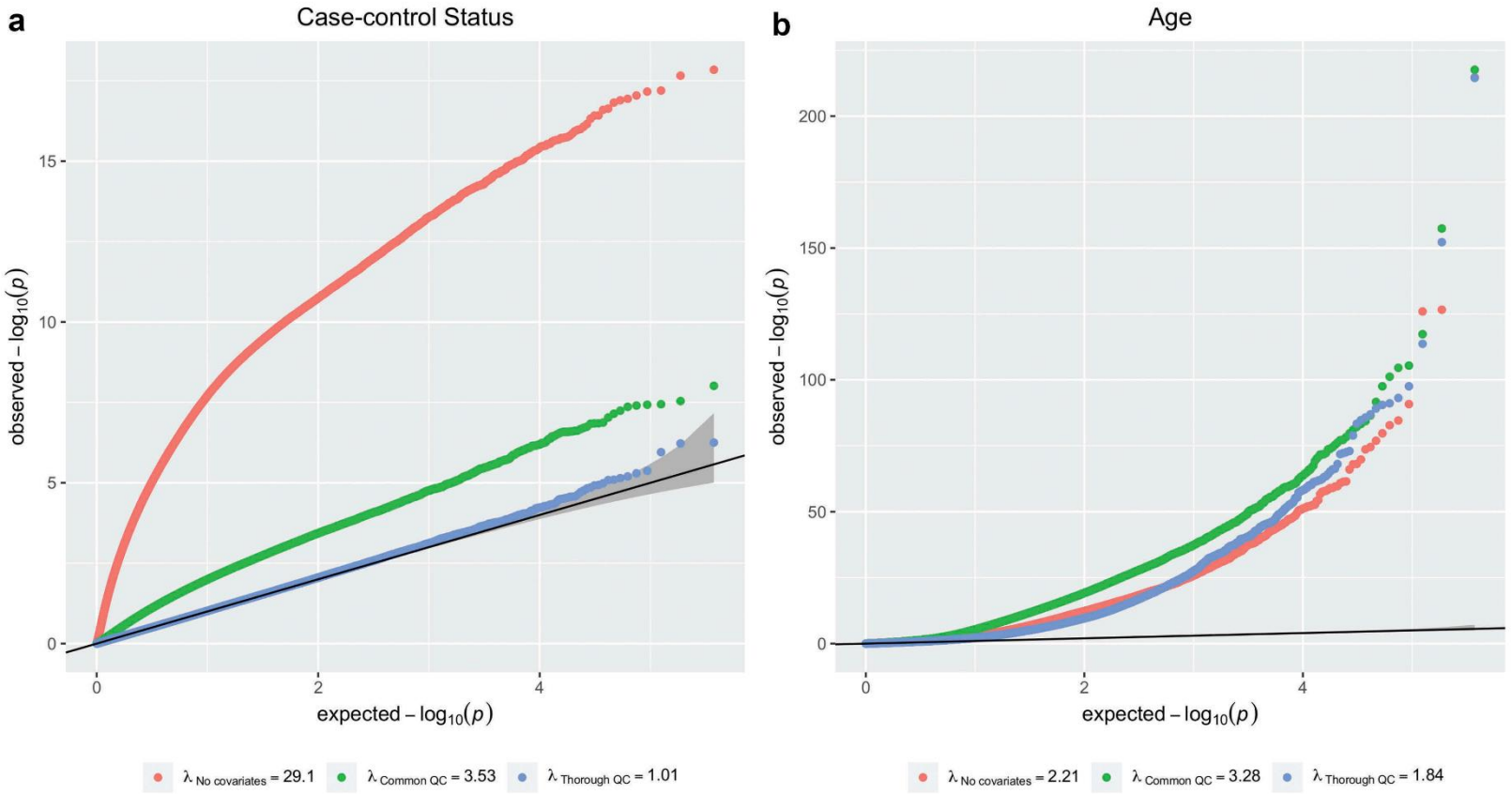
[a]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA; [b]Department of Psychiatry, University of Utah, Salt Lake City, UT, USA; [c]Center for Biomarker Research and Precision Medicine, Virginia Commonwealth University, Richmond, VA, USA; [d]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; [e]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

ABSTRACT

Recent years have seen a surge of methylome-wide association studies (MWAS). We observed that many of these studies suffer from test statistic inflation that is most likely caused by commonly used quality control (QC) pipelines not going far enough to remove technical artefacts. To support this claim, we reanalysed GEO datasets with an improved QC pipeline that reduced test-statistic inflation parameter lambda from the original mean/median of 20.16/15.17 to 3.07/1.14. Furthermore, the mean/median number of methylome-wide significant findings was reduced by 65,688/57,805 loci after more thorough QC. To avoid such false positives we argue for more extensive QC and that reporting the test-statistic inflation parameter lambda become standard for all MWAS allowing readers to better assess the risk of false discoveries.

ARTICLE HISTORY
Received 31 January 2020
Revised 24 March 2020
Accepted 26 March 2020

KEYWORDS
DNA methylation; epigenetics; reproducibility

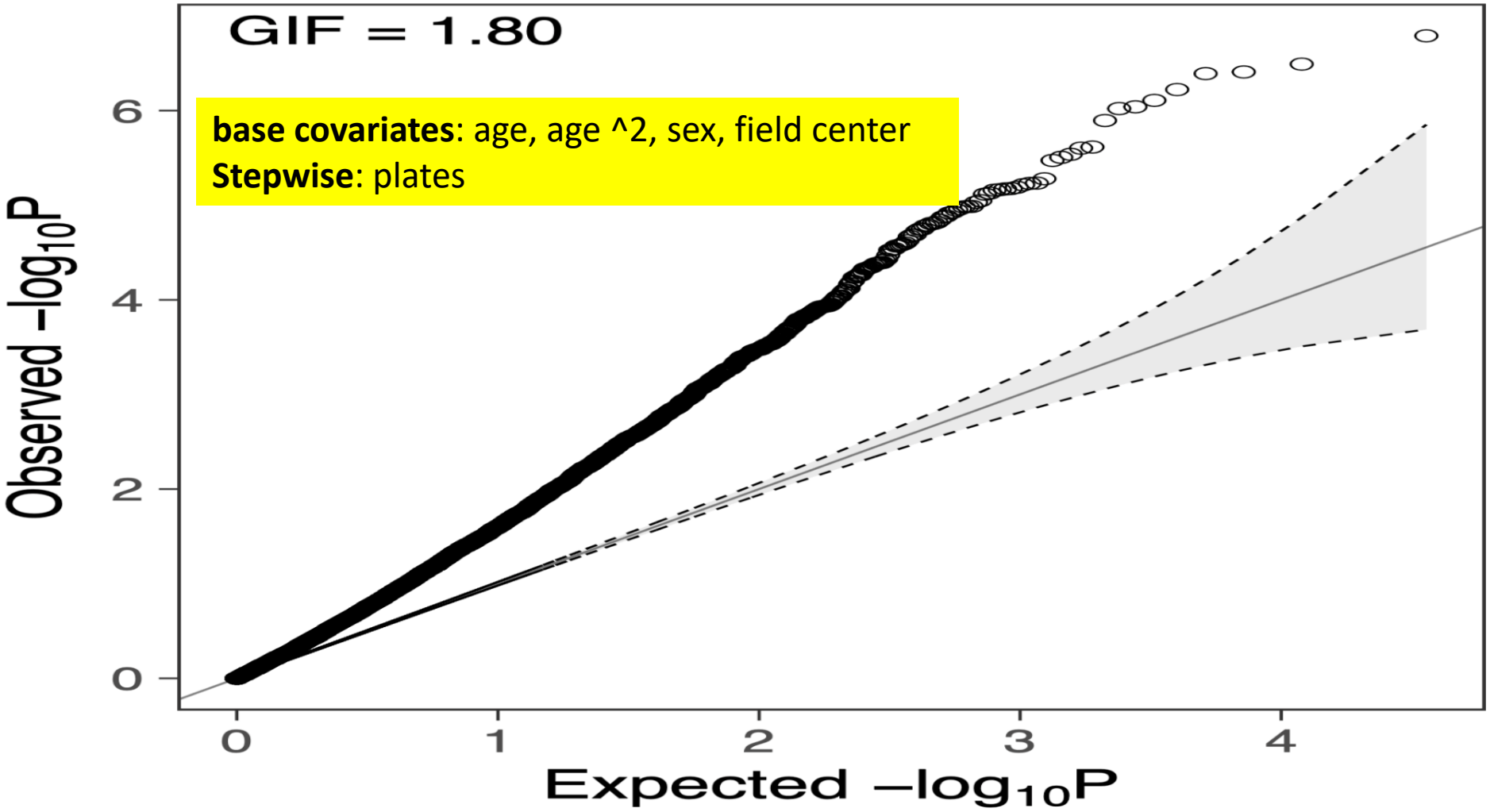


Fixed effect lab technical covariates to the analysis:

1. Bisulphite conversion percentages estimated from control probes;
2. Median signal intensities for methylated and unmethylated channels;
3. Slide and well effects that refer to the individual BeadChip arrays (slide) and the positional effects on each array (well) which hold 12 samples each (eight samples for the EPIC array);
4. Principal components (PCs) of control probe values as measures of technical variation among individual samples;
5. PCs of the methylation beta values which capture any remaining unmeasured confounders.

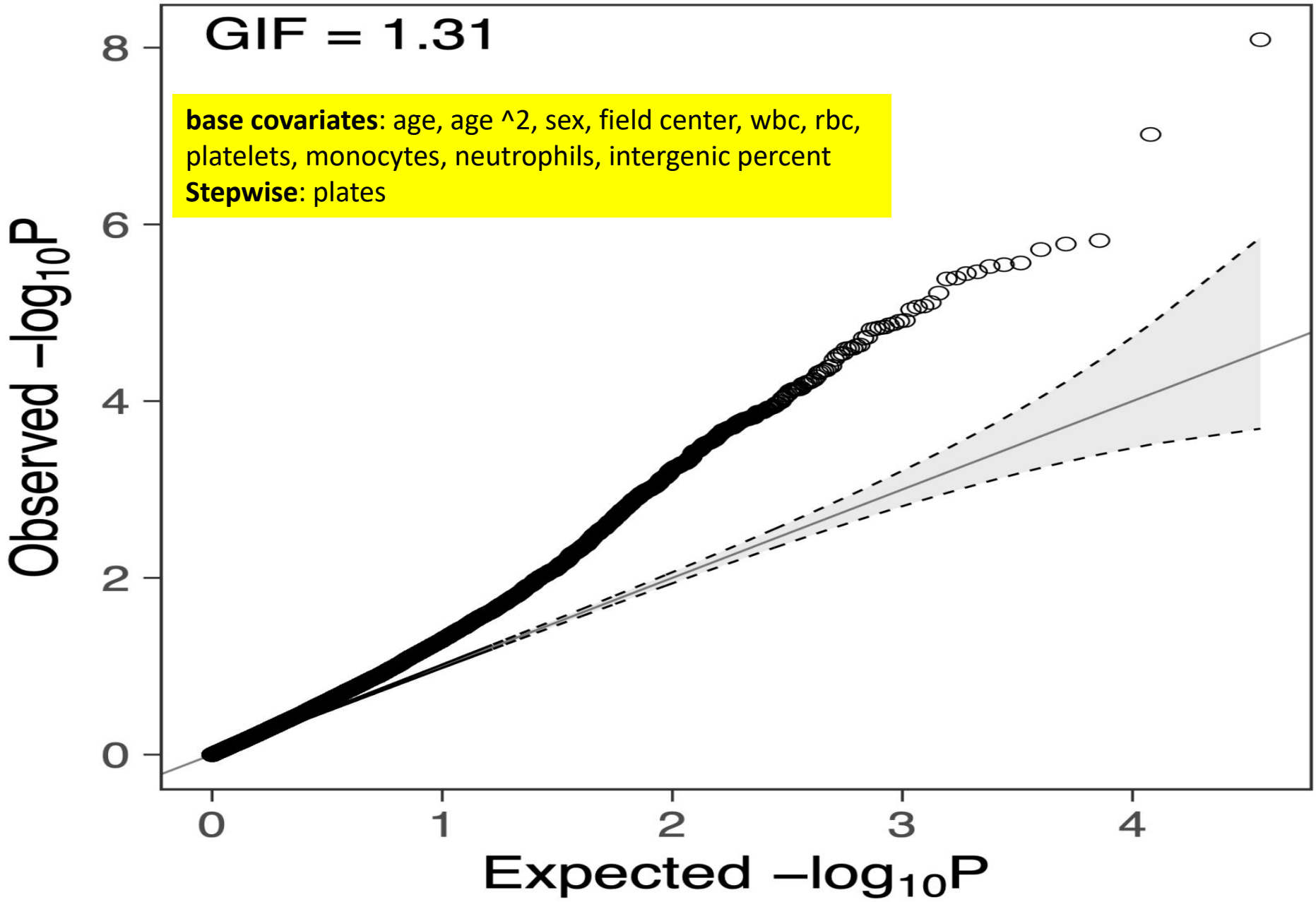
Pipeline: *Shabalin AA, HattabMW, Clark SL, et al. RaMWAS: fast methylome-wide association study pipeline for enrichment platforms. Bioinformatics. 2018;34:2283–2285.*

TWAS of Forced Vital Capacity (FVC)

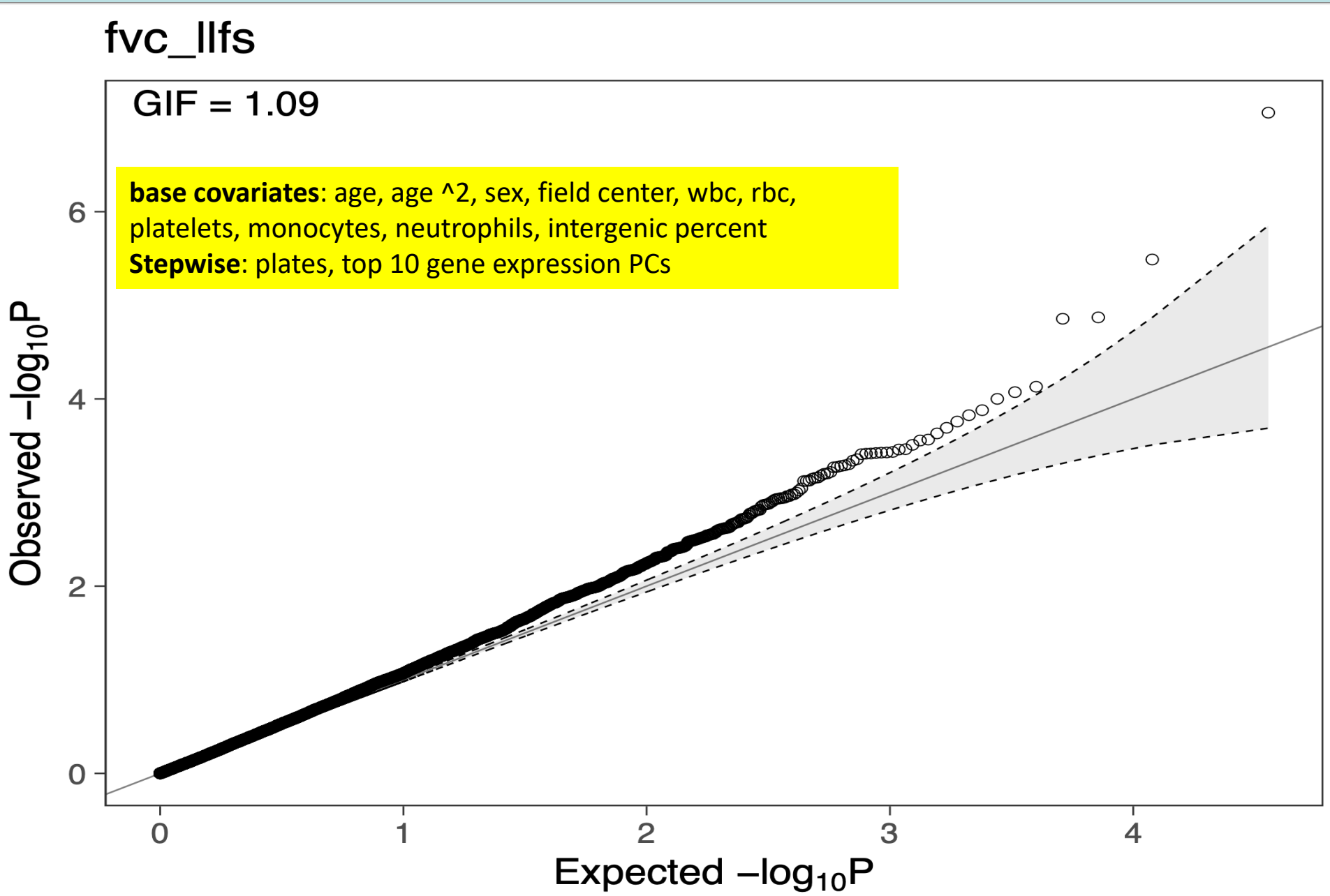


Acharya S, Liao S, Jung WJ, Kang YS, Moghaddam VA, Feitosa MF, Wojczynski MK, Lin S, Anema JA, Schwander K, Connell JO, Province MA, Brent MR. **A methodology for gene level omics-WAS integration identifies genes influencing traits associated with cardiovascular risks: the Long Life Family Study.** Hum Genet. 2024 Oct;143(9-10):1241-1252. doi: 10.1007/s00439-024-02701-1. Epub 2024 Sep 14. PMID: 39276247; PMCID: PMC11485042.

TWAS of Forced Vital Capacity (FVC)



TWAS of Forced Vital Capacity (FVC)

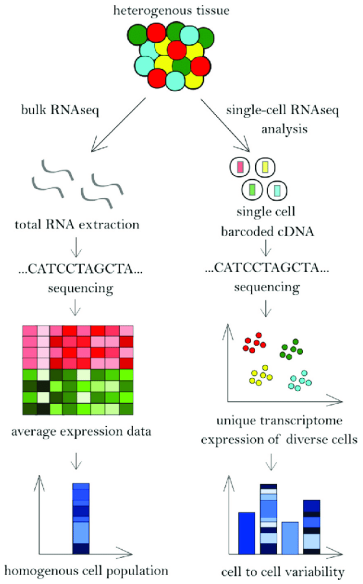


Akbary Moghaddam V, Acharya S, Schwaiger-Haber M, Liao S, Jung WJ, Thyagarajan B, Shriver LP, Daw EW, Saccone NL, An P, Brent MR, Patti GJ, Province MA. ***Construction of Multi-Modal Transcriptome-Small Molecule Interaction Networks from High-Throughput Measurements to Study Human Complex Traits.*** bioRxiv [Preprint]. 2025 Jan 23:2025.01.22.634403. doi: 10.1101/2025.01.22.634403. PMID: 39896668; PMCID: PMC11785221.

TWAS for HOMA2_S

RNA-seq

Source: White blood cells



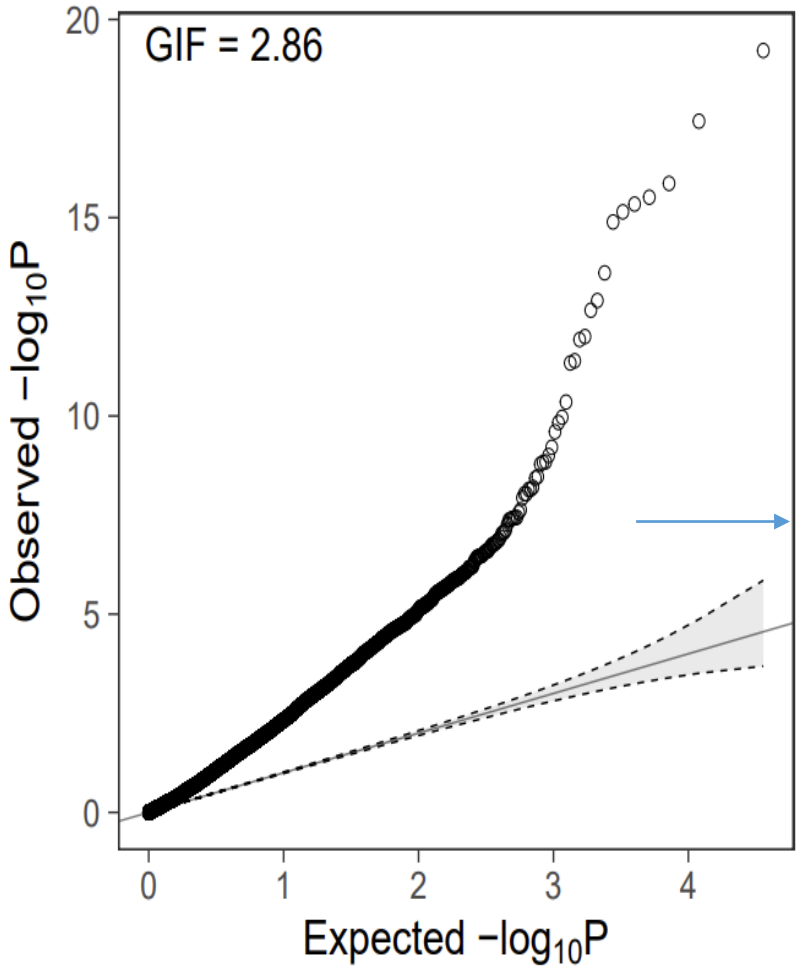
Gene expression covariates (confounders):

- Age, age^2, sex, field centers
- White blood cell count
- White blood cell differentiation
- Red blood cell count
- Platelets count
- Percent of intergenic reads

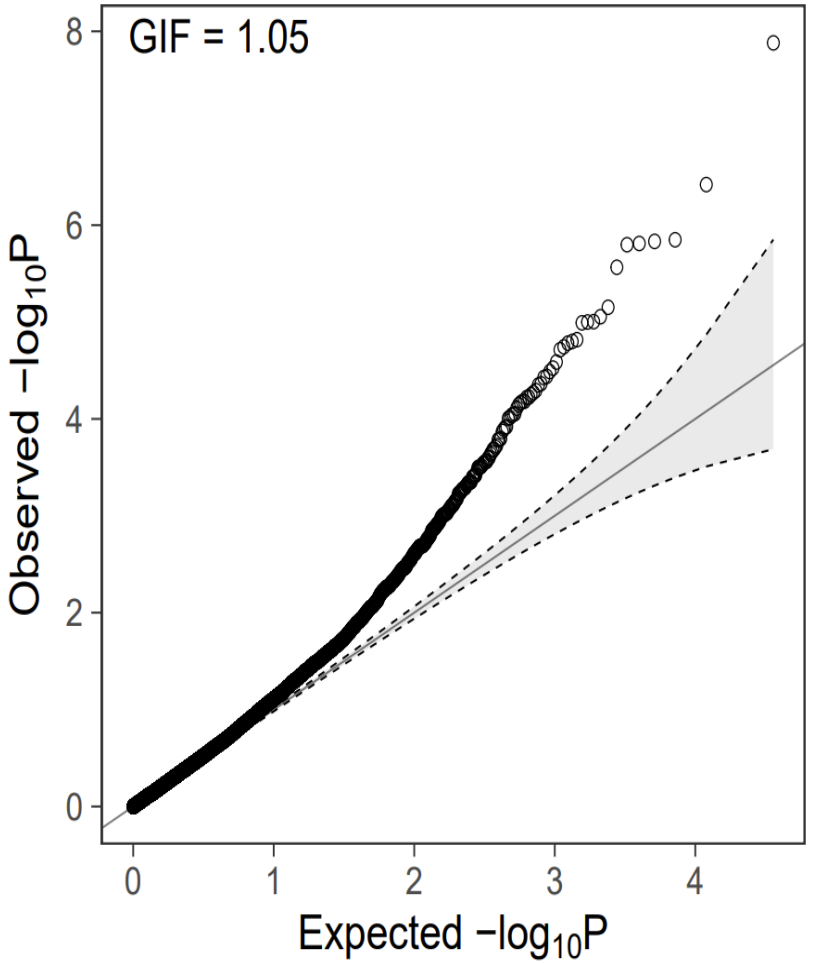
Credits: Sandeep Acharya from Brent Lab

- Plate no.
- Top 10 gene expression principle components

RNA-seq: TWAS QQ-Plots

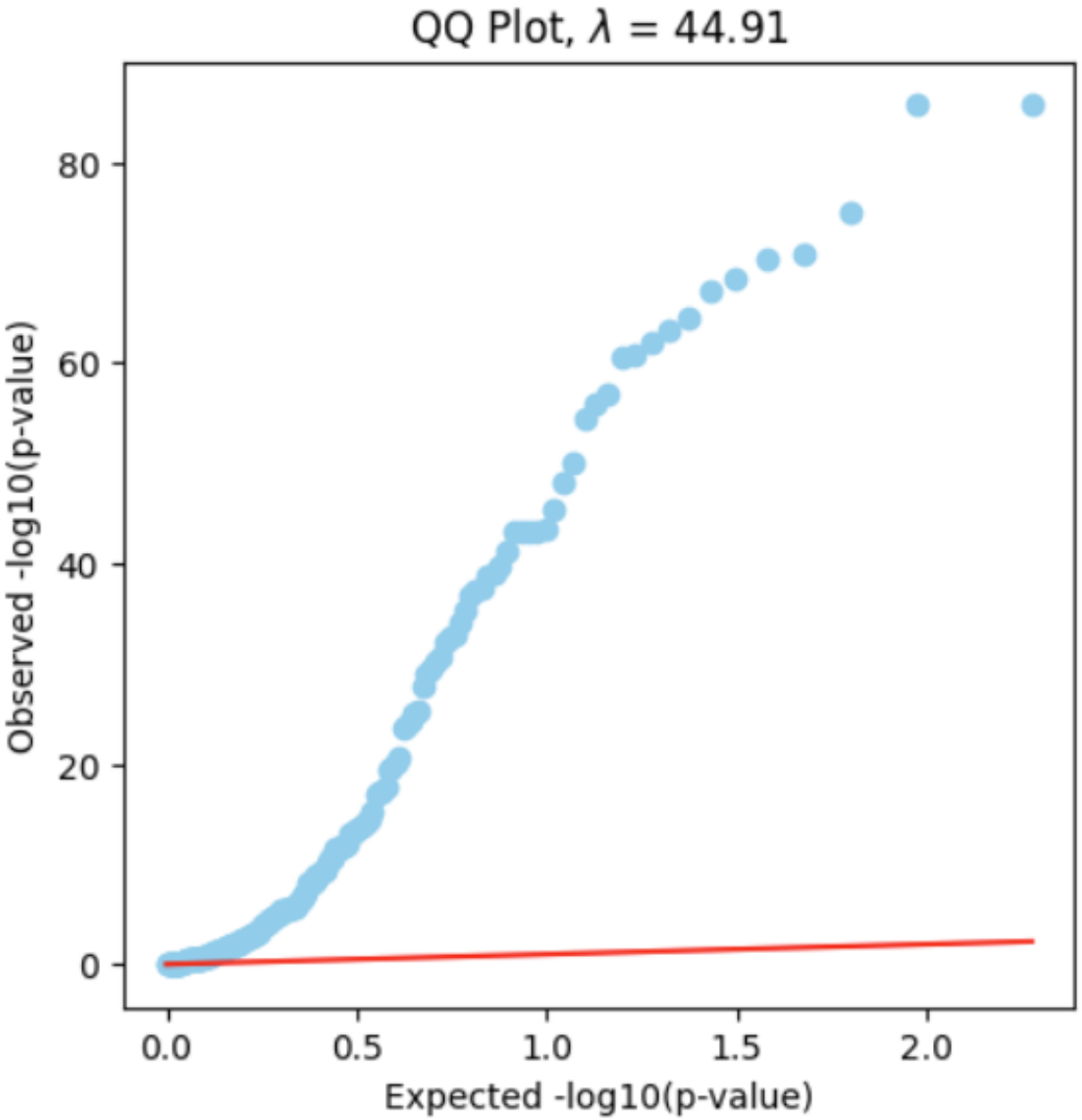


qq-plot of TWAS with raw VST normalized gene expression



qq-plot of TWAS after adjusting gene expression for the relevant covariates

Metabolite peak correction



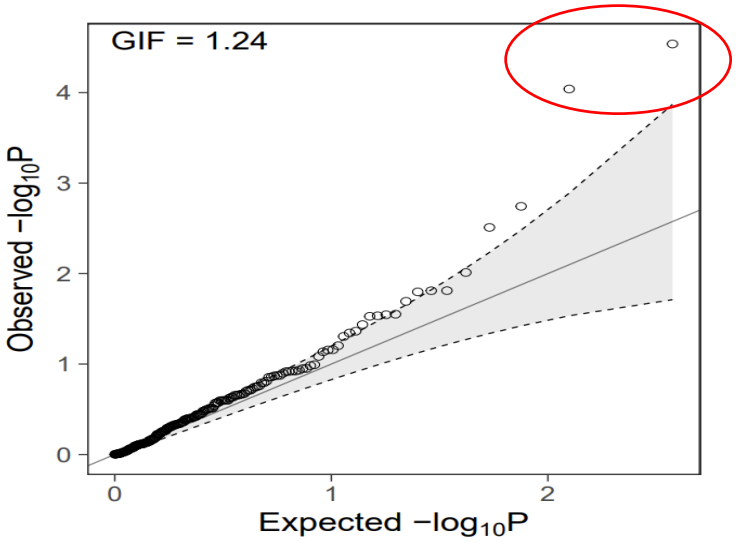
qq-plot of lipidome peaks after peak normalization

Metabolomics results after covariate adjusments

Metabolite peak intensities

- Variance stabilization (log transformation)
- Age, age^2, sex, field centers
- Batch effect, sampling, missing data (technical)
- Medication data
- Smoking habits
- Physical activity (short performance battery test)
- LDL and HDL for lipids

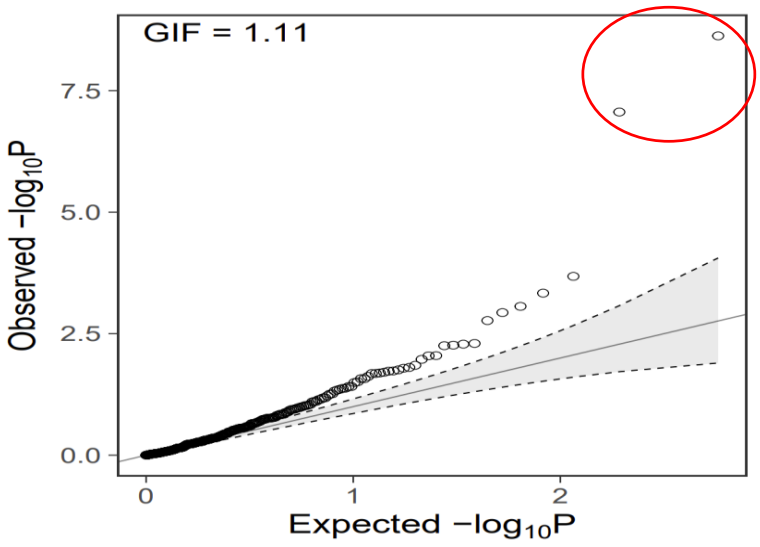
Lipids:



Phosphatidylcholine 40:6

Phosphatidylcholine 35:1

Polars:



N-acetylglycine

2-Oxo-3,3-dimethylguanidine-pentanoic acid

Conclusions & Recommendations

- Analyzing Pedigree OMIC Data **WILL produce excess false-positives** if you do not correct for correlated data
- **Lots of methods/software** for properly analyzing pedigree OMIC data [powerful, with no (or at least less) inflation of false-positives]
- Q-Q plots are your **BEST FRIEND** when conducting OMIC scans
  - Shows if your model has excess false-positives or
    If you are overcorrecting for false-positives and lose power
  - May tell you what you don't want to hear
    (but that is the most important time you should LISTEN TO THEM!)
  - You need to FIX the problems illuminated by Q-Q plots,
    not IGNORE them (include important nuisance confounders, check model assumptions, needed data transformations, etc.)

Questions?

E.G. 2: A Study recruits several Racial Groups

How to model "RACE" effects?

- Epidemiologist: Combined Analyses of ALL Races
- H0: Races poolable (unless evidence *AGAINST*)
- Geneticist: Separate Analyses each Race

  H0: Races NOT poolable (unless evidence *FOR*)

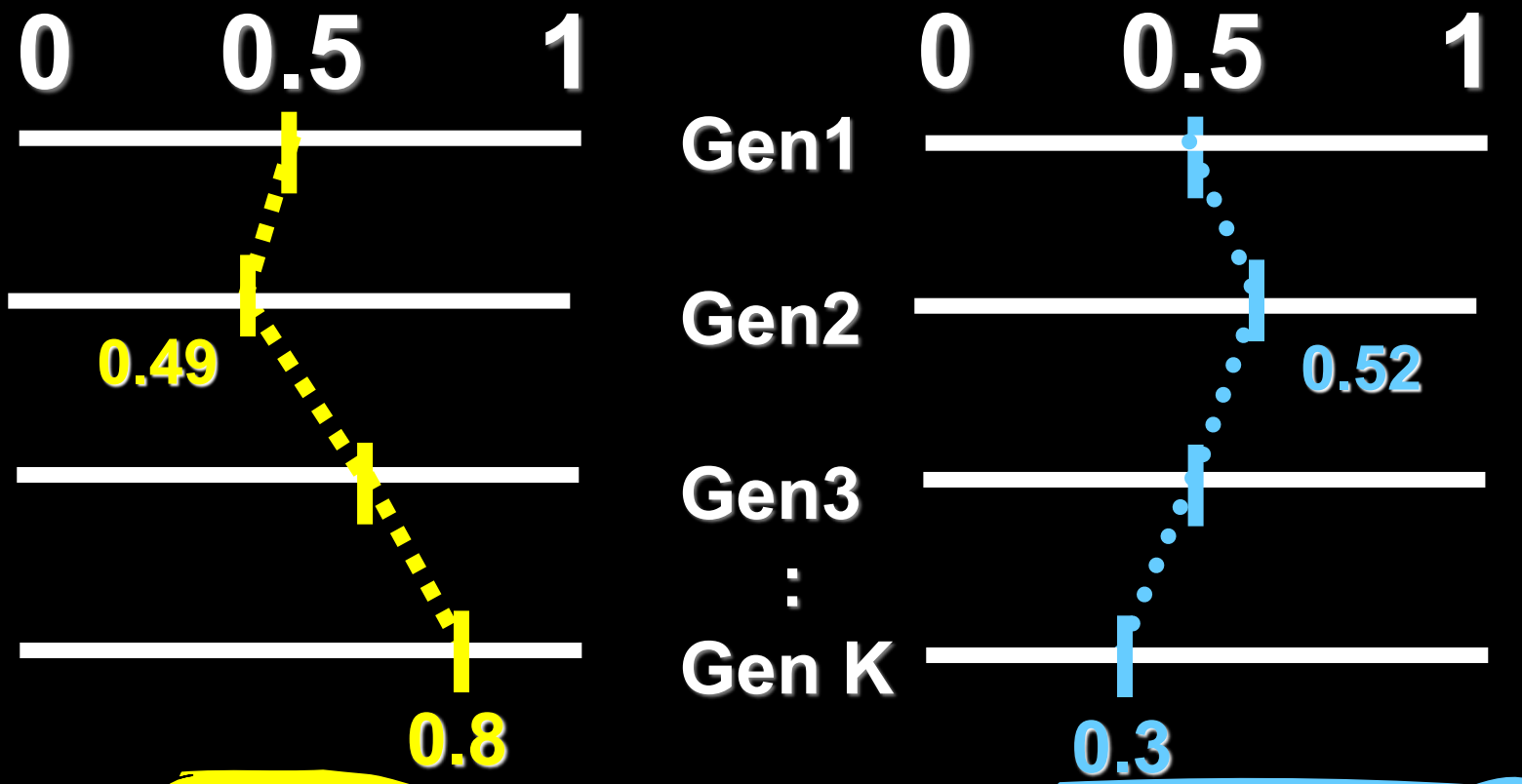
  For Geneticist "Race" ⇔ "Continent of Ancestry"

Worried: **Population Stratification** due to **Genetic Drift**...

What are these, and why are they a worry in association analyses?

Genetic Drift: SNP (simplest genotype)

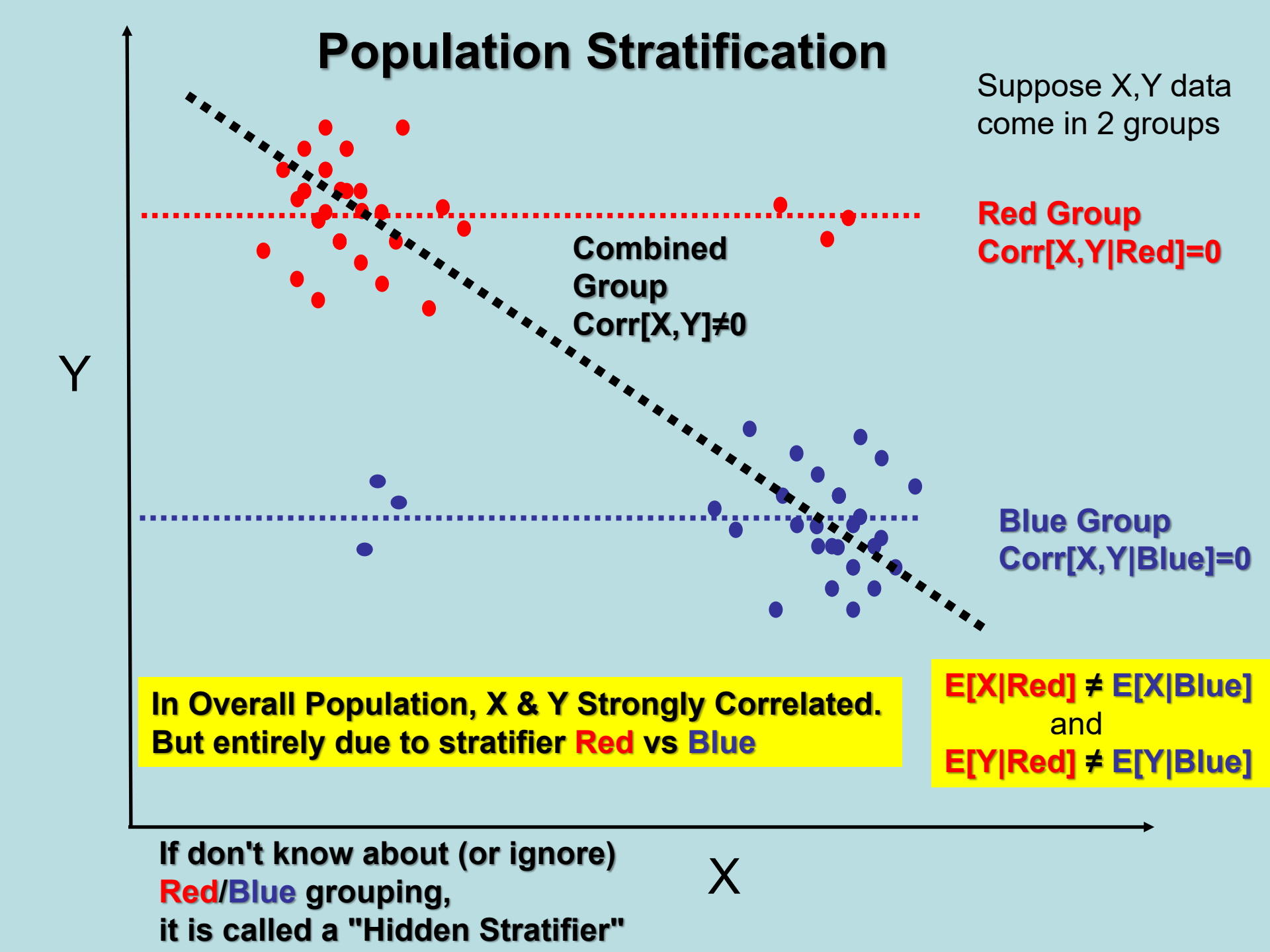
Random Allele Frequency Changes over Time



P=0.5 100,000 yrs ago
P=0.8 today

P=0.5 100,000 yrs ago
P=0.3 today

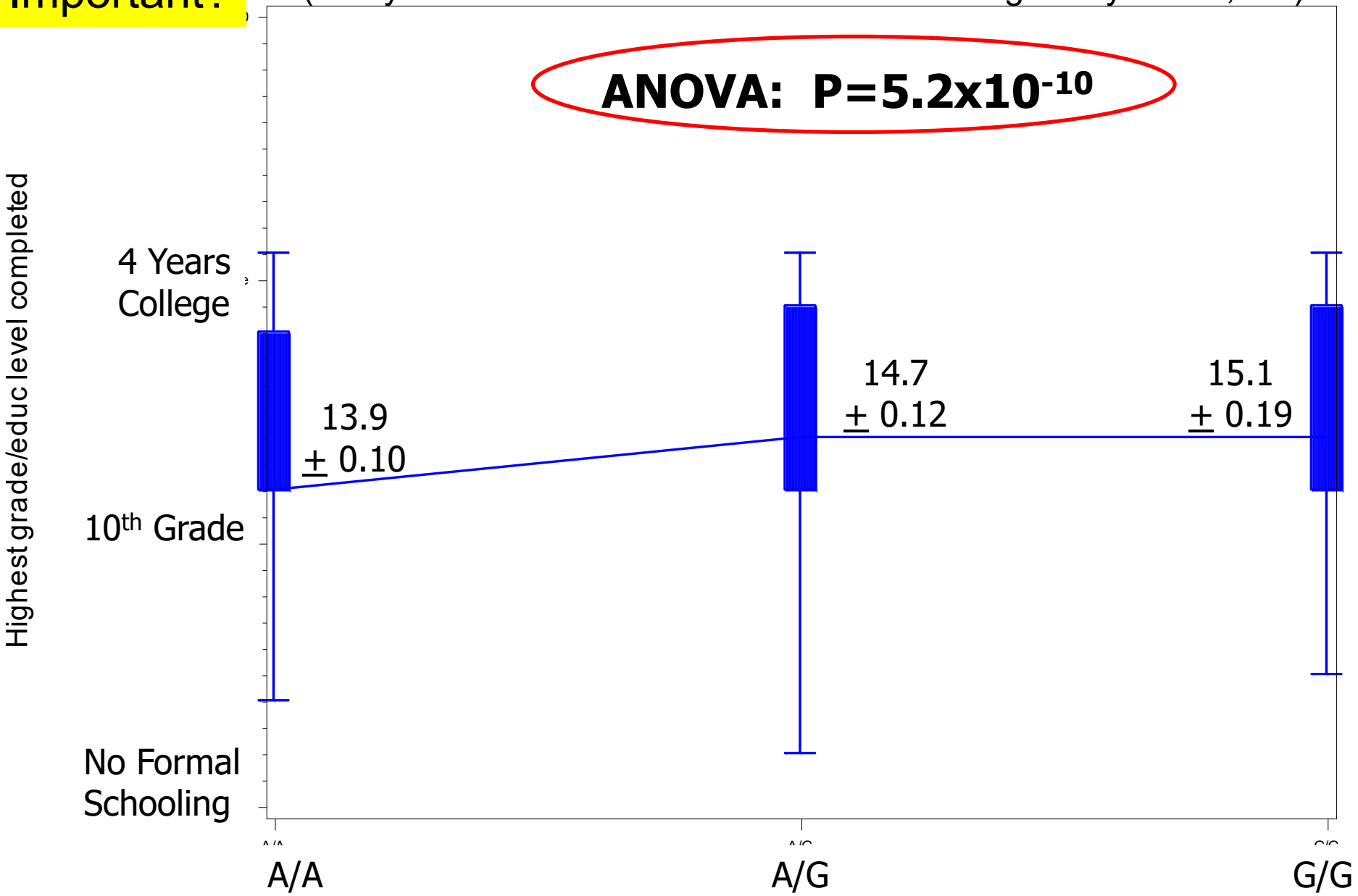
As k → ∞, AF of every SNP → 0 or 1
Fitness/Selection pushes faster in one direction


Population Stratification
Suppose X,Y data come in 2 groups
Red Group
Corr[X,Y|Red]=0
Combined Group
Corr[X,Y]≠0
Blue Group
Corr[X,Y|Blue]=0
Y
X
In Overall Population, X & Y Strongly Correlated.
But entirely due to stratifier Red vs Blue
E[X|Red] ≠ E[X|Blue]
and
E[Y|Red] ≠ E[Y|Blue]
If don't know about (or ignore)
Red/Blue grouping,
it is called a "Hidden Stratifier"

Why Important?

Is AGT (Angiotensin) also an Education Gene?

(Analysis of multiracial FBPP Data as "One Big Study" N=11,357)



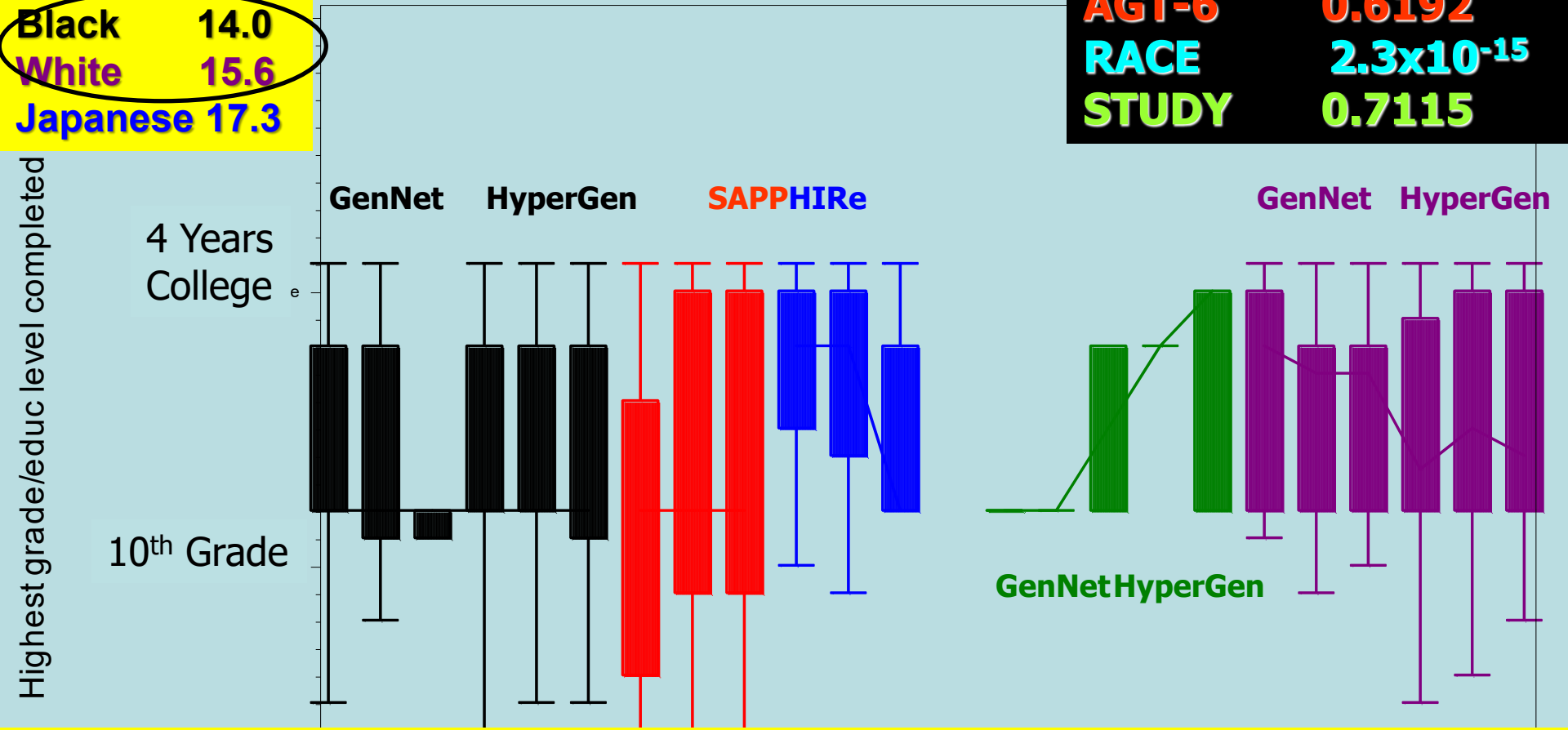
AGT-6 (RS5051) a well know Blood Pressure variant

Is AGT an Education Gene?

Stratified Analyses by Race, Study

Mean EDU yrs	
Chinese	12.2
Black	14.0
White	15.6
Japanese	17.3

Multivariate Model	
Source	P-value
AGT-6	0.6192
RACE	2.3×10^{-15}
STUDY	0.7115



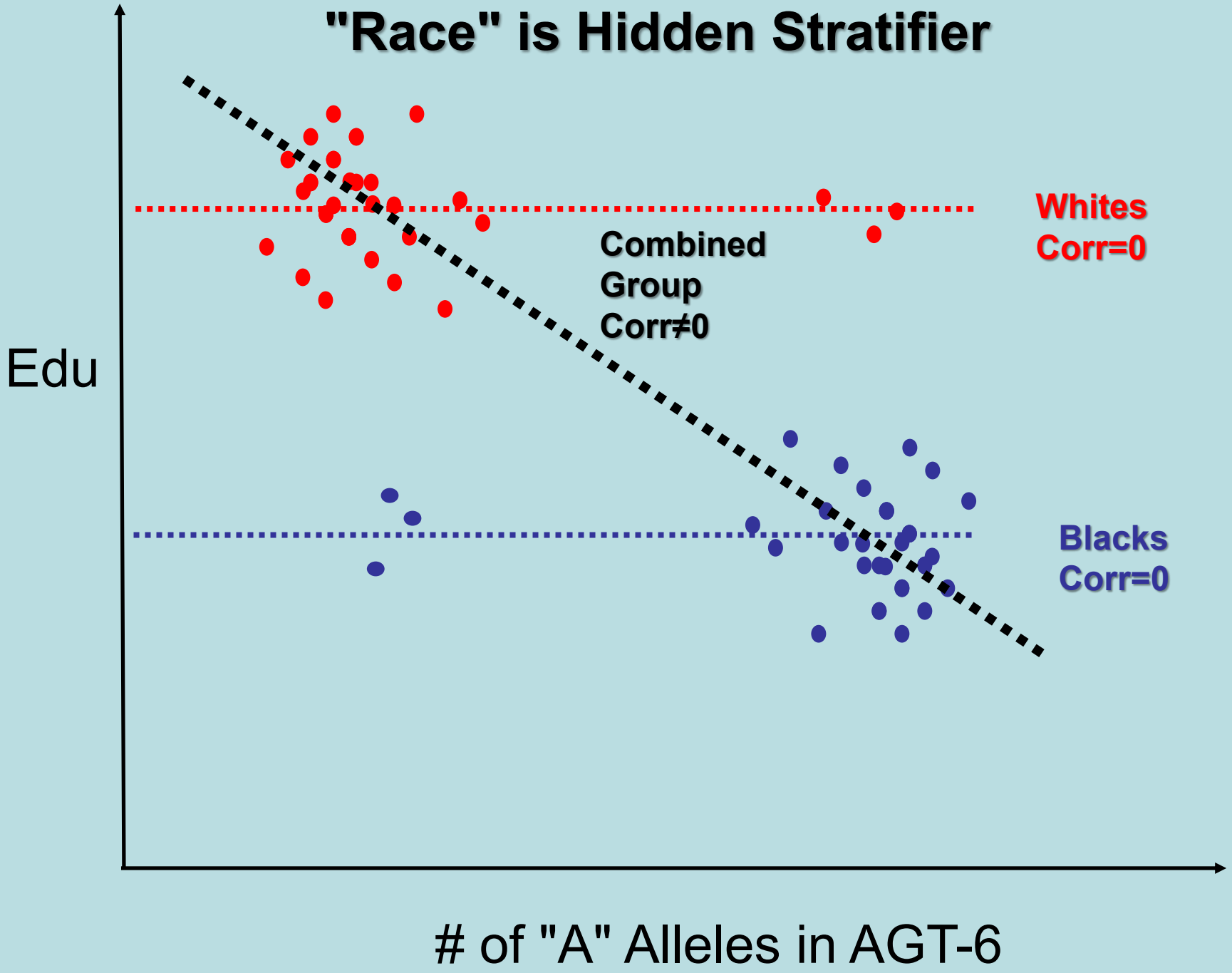
MORAL: Genetic Drift Race diffs + Cultural/Legal EDU Race diffs → Population Stratification → False-Positive Gene Signal for Gene

Schooling

A/A A/G G/G A/A A/G G/G A/A A/G G/G A/A A/G G/G A/A A/G G/G A/A A/G G/G A/A A/G G/G A/A A/G G/G A/A A/G G/G

AGT-6

Race	A Allele Freq
Black	85%
Other	47%
Chinese	87%
White	45%
Japanese	83%

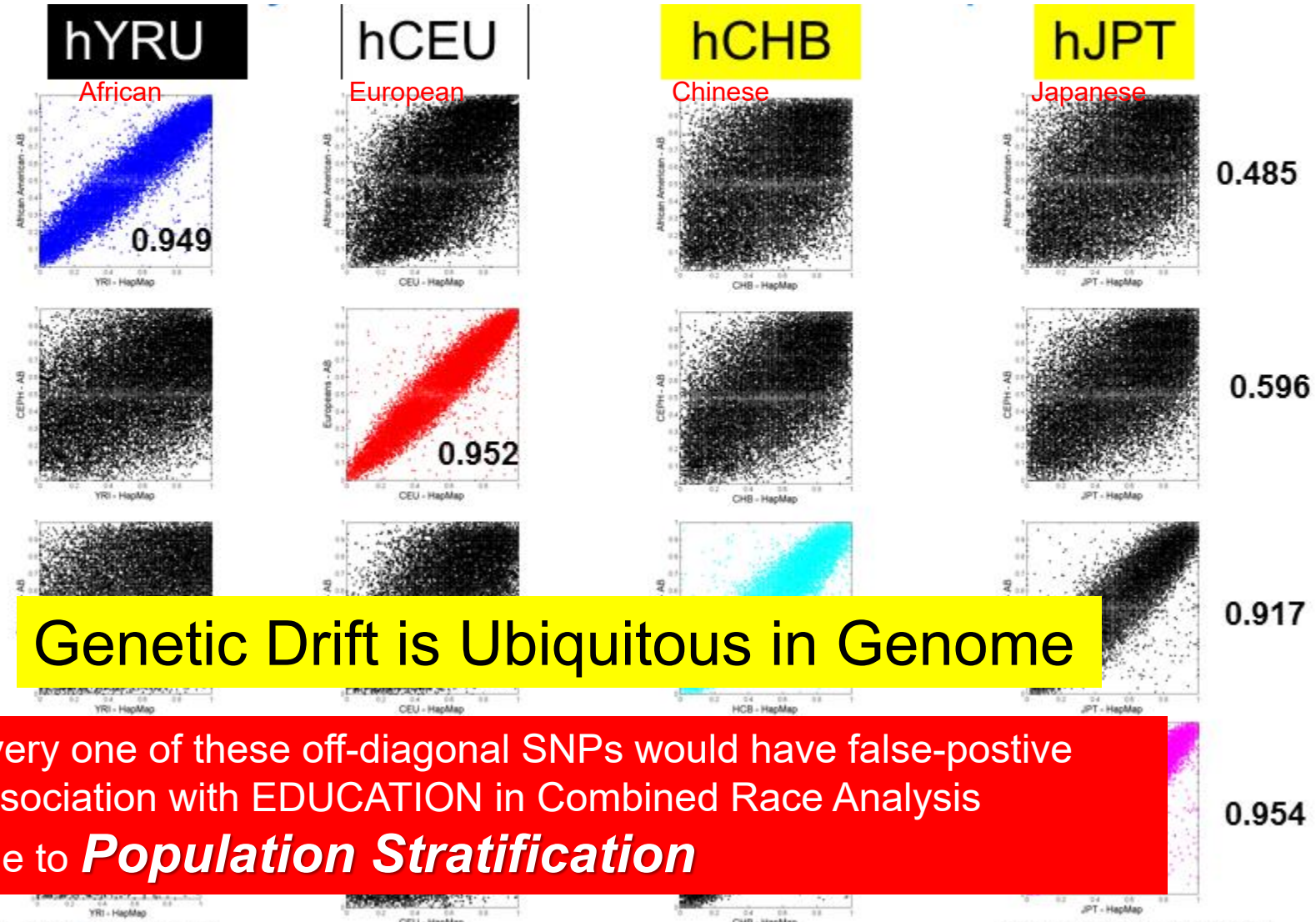

"Race" is Hidden Stratifier
Edu
Combined
Group
Corr≠0
Whites
Corr=0
Blacks
Corr=0
# of "A" Alleles in AGT-6

Allele Frequency Differences by Population

Patterns of linkage disequilibrium are conserved within and between populations

Applied Biosystems

HapMap Study



Genetic Drift is Ubiquitous in Genome

Every one of these off-diagonal SNPs would have false-postive association with EDUCATION in Combined Race Analysis due to ***Population Stratification***