Testing for GxE interaction in structured populations

Andrey Ziyatdinov

May 5, 2017

Definitions of GxE interaction

Biological interaction: Genetic factor(s) and environmental factor(s) participate in the same causal mechanism (Rothman *et al.*, 2008)

Statistical interaction using linear regression (unrelated individuals):

$$y = \mu + \beta_g x_g + \beta_e x_e + \beta_{int} x_g x_e + e$$

(Aschard *et al.*, 2012, HumGen)

Current data sets and methods in GWAS of GxE

Consortium	Sample size	Exposure	Outcome	Reference
CHARGES + SPIROMETA	50,047	Smoking	Pulmonary function	(Hancock *et al.*, 2012)
SUNLIGHT	35,000	Vitamin D intake	Circulating Vitamin D level	(Wang *et al.*, 2010)
GIANT	up to 339,224	Gender	Anthropometric traits	(Heid *et al.*, 2010)
…				

- Most studies used a *full framework* (model on the previous slide)
  - GxE analysis on the full sample
- Other studies used a *stratified framework* (coming on the next slides)
  - GxE analysis on sub-samples stratified by exposure

Example of the GxE full framework in GWAS

Study: G x smoking in pulmonary function outcomes (Hancock *et al.*, 2012)

- 50,047 participants from 19 studies; ~2.5M SNPs
- outcomes: FEV1, FEV1/FVC (%)
- smoking variables: ever-smoker, current-smoker, packs-year
  - marginal terms: all included
  - interaction terms: tested separately
- joint test: $\beta_g = 0$ and $\beta_{int} = 0$ under the null (Aschard *et al.*, 2011)

Findings: three novel gene regions

1. DNER
2. HLA-DQB1 and HLA-DQA2
3. KCNJ2 and SOX9

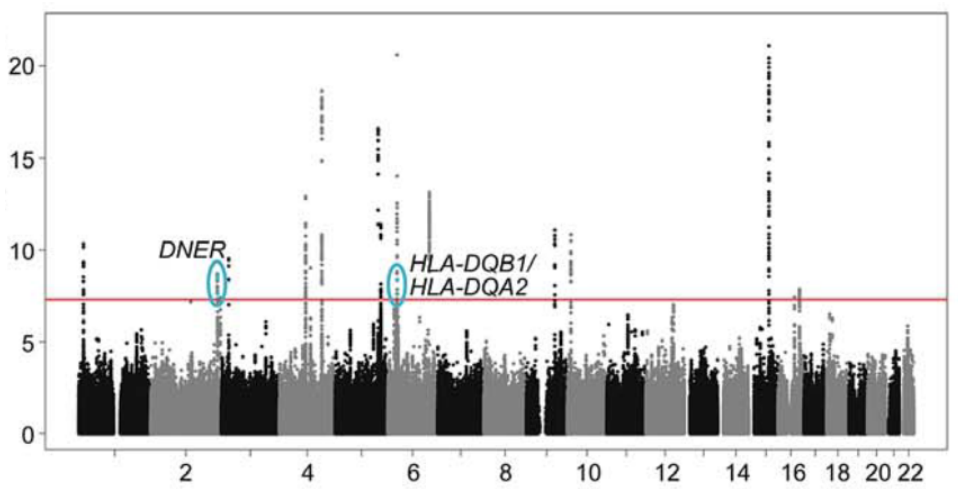
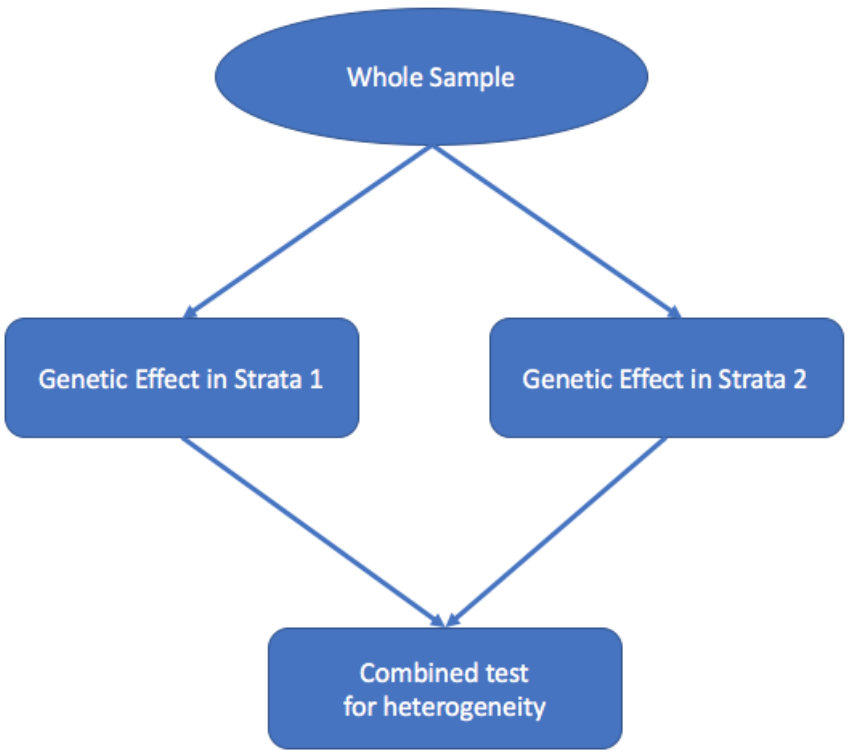


Figure: G x ever-smoking in FEV1/FVC (Hancock *et al.*, 2012)
Abbreviations: FEV1, Force Expiratory Volume in 1 second; FVC, Force Vital Capacity

Example of the GxE stratified framework in GWAS

Study: G x gender in the Genetic Investigation of Anthropometric Traits (GIANT) consortium

- outcome: Waist-hip ratio (WHR)
- gender as an exposure
  - 108,979 women
  - 82,483 men
- marginal GWAS: 14 loci associated to WHR
- explained variance in WHR by 14 loci
  - 1.34% in women
  - 0.46% in men



Findings:

- stratified interaction analysis:
  - 7/14 loci singificant interaction
  - 2/14 genome-wide significant

Abbreviations: WHR, Waist-hip ratio
Stratified framework described in (Magi *et al.*, 2010), (Randall *et al.*, 2013)

Other negative results were not published…

because of the challenges inherant to the detection of GxE

GxE are interesting to know, difficult to detect

Statistical power for interaction tests is lower than for similar tests of marginal genetic effects (Murcray *et al.*, 2011)

It also faces other potential issues (Aschard *et al.*, 2012):

- Confounding
- Exposure measurement error and misclassification
- Dynamics of gene–environment interactions
- …

Relatedness is yet another layer of complexity in GxE analysis,
which impact on the full/stratified GxE frameworks is seldom explored.

Structure among individuals considered

- shared environment: house-hold groups
- recent genetic relatedness: family members
- distant genetic relatedness: admixed populations

Our goal

Assess the relative performance of GxE methods
in the presence of structure

Outline

1. GxE full framework in structured population
    - revise the formulas of linear mixed models known in GWAS
    - compare two study designs, unrelated and related
2. GxE stratified framework in structured population
    - assess the applicability using simulations
3. Ancestry x E analysis in admixed population

1. GxE full framework in structured population

Population structure in (marginal) association tests

Methods to account for relatedness are relatively well established in *marginal* association studies (GWAS)

- Principal component analysis (PCA)
- Linear mixed models (LMMs) (Yang *et al.*, 2014)
- Robust tests: genotype-conditional association test (GCAT) (Song *et al.*, 2015, NatGen)

Linear mixed model (also data simulation model)

$$y = X\beta + g + f + e$$

where $g \perp f \perp e$

- $g \sim (0, \sigma_g^2 K)$, the (additive) genetic effect
- $f \sim (0, \sigma_f^2 F)$, the shared env. effect
- $e \sim (0, \sigma_r^2 I)$, the residual error

implying

$$y \sim (X\beta, \sigma_g^2 K + \sigma_f^2 F + \sigma_r^2 I) = (X\beta, V)$$

- K, the double kinship matrix
- F, the variance-covariance matrix of shared environment
- $\sigma_g^2, \sigma_f^2, \sigma_r^2$, variance proportions

(Lynch and Walsh, 1998)

1. marginal effect:

$$X\beta = \mu + \beta_g x_g$$

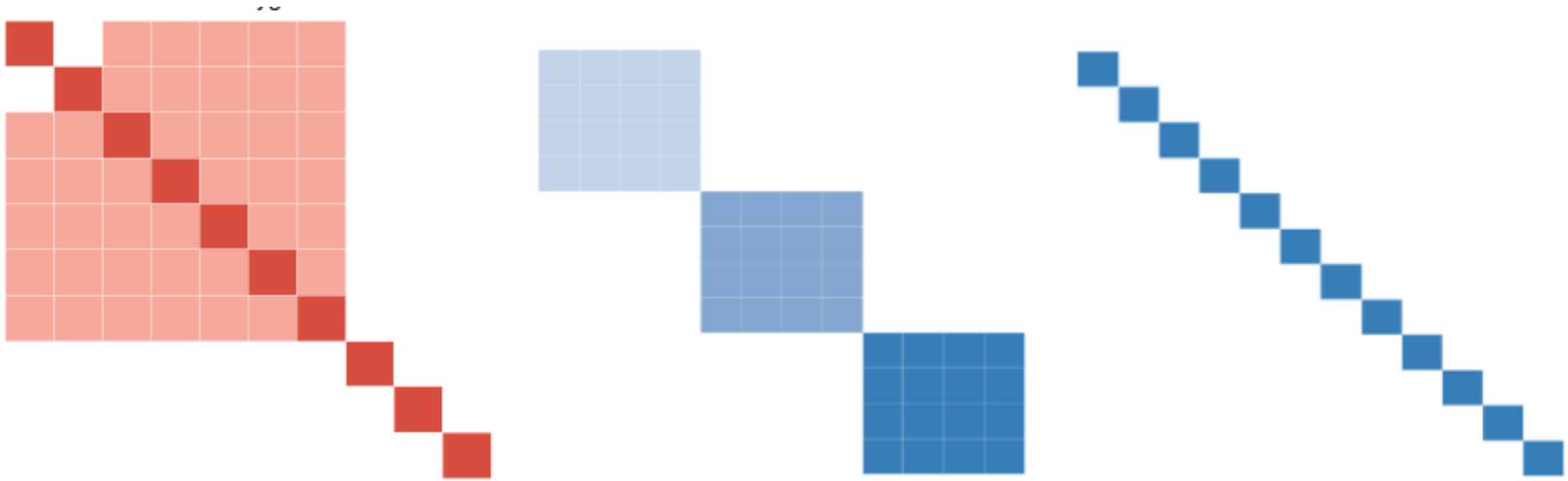
---

2. interaction effect:

$$X\beta = \mu + \beta_g x_g + \beta_e x_e + \beta_{int} x_g x_e$$

Variance-covariance matrices

Genetic relationship matrix K Shared environment matrix F Residual variance I



Estimation of model parameters

1. Estimate variance components by ML/REML

$$\hat{V} = \hat{\sigma}_g^2 K + \hat{\sigma}_f^2 F + \hat{\sigma}_r^2 I$$

2. Derive the effect size as in Generalized Least Squares (GLS)

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

$$var(\hat{\beta}) = (X^T \hat{V}^{-1} X)^{-1}$$

For the power derivation: $var(\hat{\beta}_x)$ needs to be extracted

$$Z_x^2 = \hat{\beta}_x^{\,2} / var(\hat{\beta}_x) \simeq \chi_1^2$$

From matrix to vector forms

Simplify to a one-covariate model by orthogonalization

1. marginal effect: $E(Y) = \mu + \beta_g x_g$

y^*, centered y
x_g^*, centered x_g

$$var(\hat{\beta}_g) = ({x_g^*}^T \hat{V}^{-1} x_g^*)^{-1}$$

---

2. interaction effect: $E(Y) = \mu + \beta_g x_g + \beta_e x_e + \beta_{int} x_g x_e$

y^*, centered y
x_{ge}^*, centered $(x_g - \mu_g)(x_e - \mu_e)$

$$var(\hat{\beta}_{int}) = ({x_{ge}^*}^T \hat{V}^{-1} x_{ge}^*)^{-1}$$

Power (final formula)

The power as a function of the non-centrality parameter (NCP)

$$NCP = \beta^2 (x^T \hat{V}^{-1} x)^{-1} \approx \beta^2 tr(\hat{V}^{-1} \Sigma_x)$$

Data	Distribution
outcome	$y \sim (X\beta, V) = (X\beta, \sigma_f^2 F + \sigma_g^2 K + \sigma_r^2 I)$
predictor	$x \sim (\mu_x, \Sigma_x)$

Approximation using the quadratic forms

If x is a vector of random variables, the quadratic form $x^T Ax$ is a scalar random variable.

If x has mean μ and (nonsingular) covariance matrix V, then

$$E(x^T Ax) = tr(AV) + \mu^T A\mu$$

$$\sigma^2(x^T Ax) = 2tr(AVAV) + 4\mu AVA\mu$$

(Lynch and Walsh, 1998)

Power to detect the marginal effect

structure	$\boldsymbol{\Sigma_y = V}$	$\boldsymbol{\Sigma_{x_g}}$	$\boldsymbol{NCP \approx \beta^2 tr(\hat{V}^{-1}\Sigma_{x_g})}$
unrelated	$(\sigma_g^2 + \sigma_r^2)I$	$\sigma_{x_g}^2 I$	$\beta^2\ 2pq\ N$
genetically related	$\sigma_g^2 K + \sigma_r^2 I$	$\sigma_{x_g}^2 K$	$\beta^2 2pq\ tr((\hat{\sigma}_g^2 K + \hat{\sigma}_r^2 I)^{-1} K)$
shared environment	$\sigma_f^2 F + \sigma_r^2 I$	$\sigma_{x_g}^2 I$	$\beta^2 2pq\ tr((\hat{\sigma}_f^2 F + \hat{\sigma}_r^2 I)^{-1})$

- $\sigma_{x_g}^2 = 2pq$, the variance of the genotype
  - p, Minor allele frequency
  - $q = 1 - p$

Power to detect the interaction effect

Under assumption that x_g and x_e are independent:

$$\Sigma_{x_{ge}} = \sigma^2_{x_g} K \sigma^2_{x_e} I = \sigma^2_{x_g} \sigma^2_{x_e} I$$

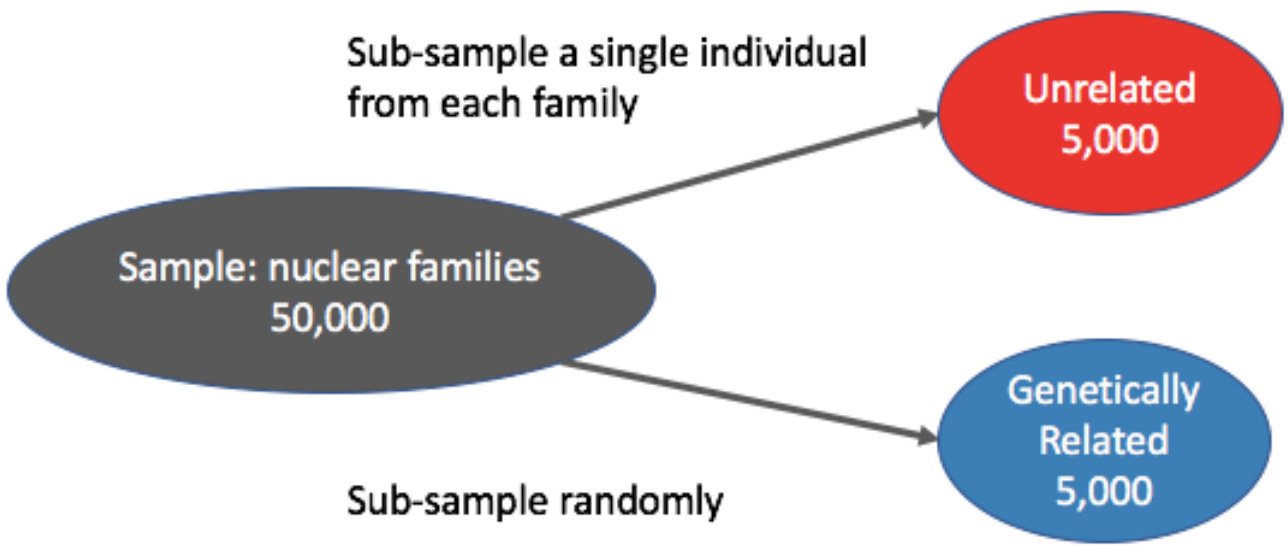
structure	$\boldsymbol{\Sigma_y = V}$	$\boldsymbol{\Sigma_{x_{ge}}}$	$\boldsymbol{NCP \approx \beta^2 tr(\hat{V}^{-1}\Sigma_{x_{ge}})}$
unrelated	$(\sigma^2_g + \sigma^2_r)I$	$\sigma^2_{x_g}\sigma^2_{x_e} I$	$\beta^2\ 2pq\, f(1-f)\, N$
genetically related	$\sigma^2_g K + \sigma^2_r I$	$\sigma^2_{x_g}\sigma^2_{x_e} I$	$\beta^2 2pq\, f(1-f)\, tr((\hat{\sigma}^2_g K + \hat{\sigma}^2_r I)^{-1})$

- $\sigma^2_{x_g} = 2pq$, the variance of the genotype
  - p, Minor allele frequency
  - $q = 1 - p$
- $\sigma^2_{x_e} = f(1-f)$, the variance of the binary exposure
  - f, frequency of exposure

Data simulation 1 (marginal): genetic relatedness

Data simulation of the whole sample (nuclear families):

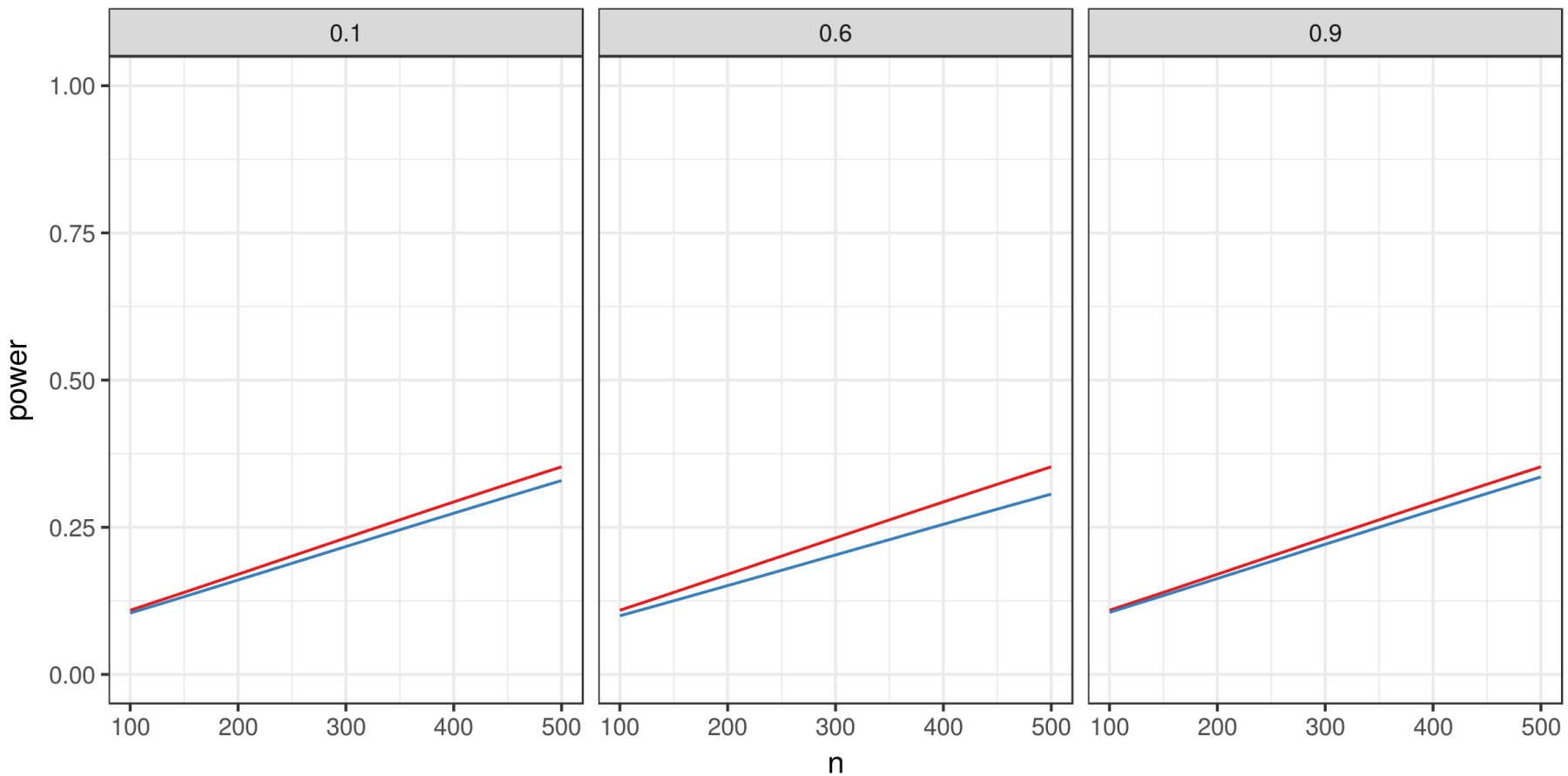
- $y \sim (X\beta, V) = (\mu + \beta_g x_g, \sigma_g^2 K + \sigma_r^2 I)$
- $\sigma_g^2 + \sigma_r^2 = 1$



Analytical results 1 (marginal): genetic relatedness

For genetically related: $NCP = \beta^2 2pq\ tr((\hat{\sigma}_g^2 K + \hat{\sigma}_r^2 I)^{-1} K)$

- $\sigma_g^2 = \{0.1, 0.6, 0.9\}$
- The power in unrelated is always higher
- The difference in power depends on the heritability (σ_g^2) non-monotonically



Confirmed the known results

European Journal of Human Genetics (2008) 16, 387–390

www.nature.com/ejhg

SHORT REPORT

Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained

Peter M Visscher*,[1], Toby Andrew[2,3] and Dale R Nyholt[1]

[1]*Genetic Epidemiology, Queensland Institute of Medical Research, Herston, Brisbane, Australia;* [2]*Department of Epidemiology & Public Health, Imperial College, St Mary's Campus, Norfolk Place, London, UK;* [3]*Twin Research and Genetic Epidemiology Unit, St Thomas' Hospital, London, UK*

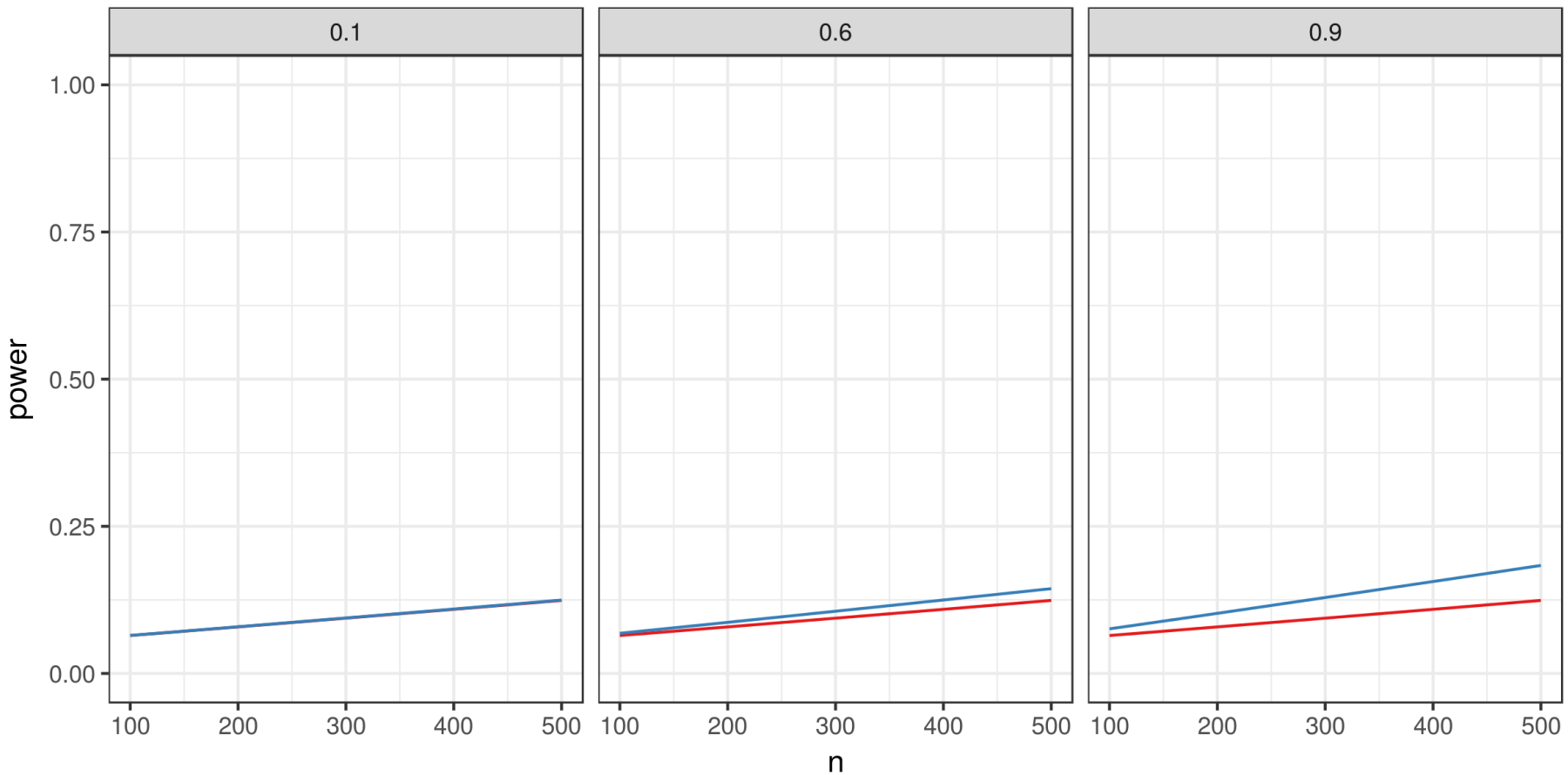
But our formula allows us to explore further performances

- for both marginal and interaction models
- across various study designs (expressed via V matrix)

Analytical results 1 (interaction): genetic relatedness

For genetically related: $NCP = \beta^2 2pq\, f(1-f)\, tr((\hat{\sigma}_g^2 K + \hat{\sigma}_r^2 I)^{-1})$

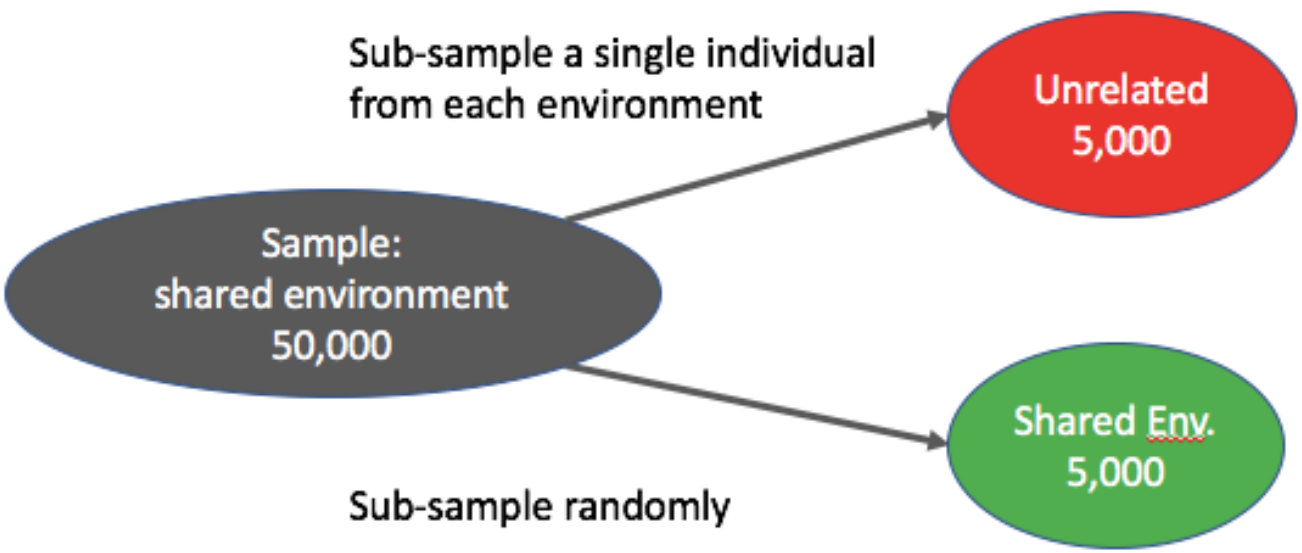
- $\sigma_g^2 = \{0.1, 0.6, 0.9\}$
- The power in genetically related is higher (in nuclear families)



Data simulation 2 (marginal): shared environment

Data simulation of the whole sample (genetically unrelated, but related by shared env.):

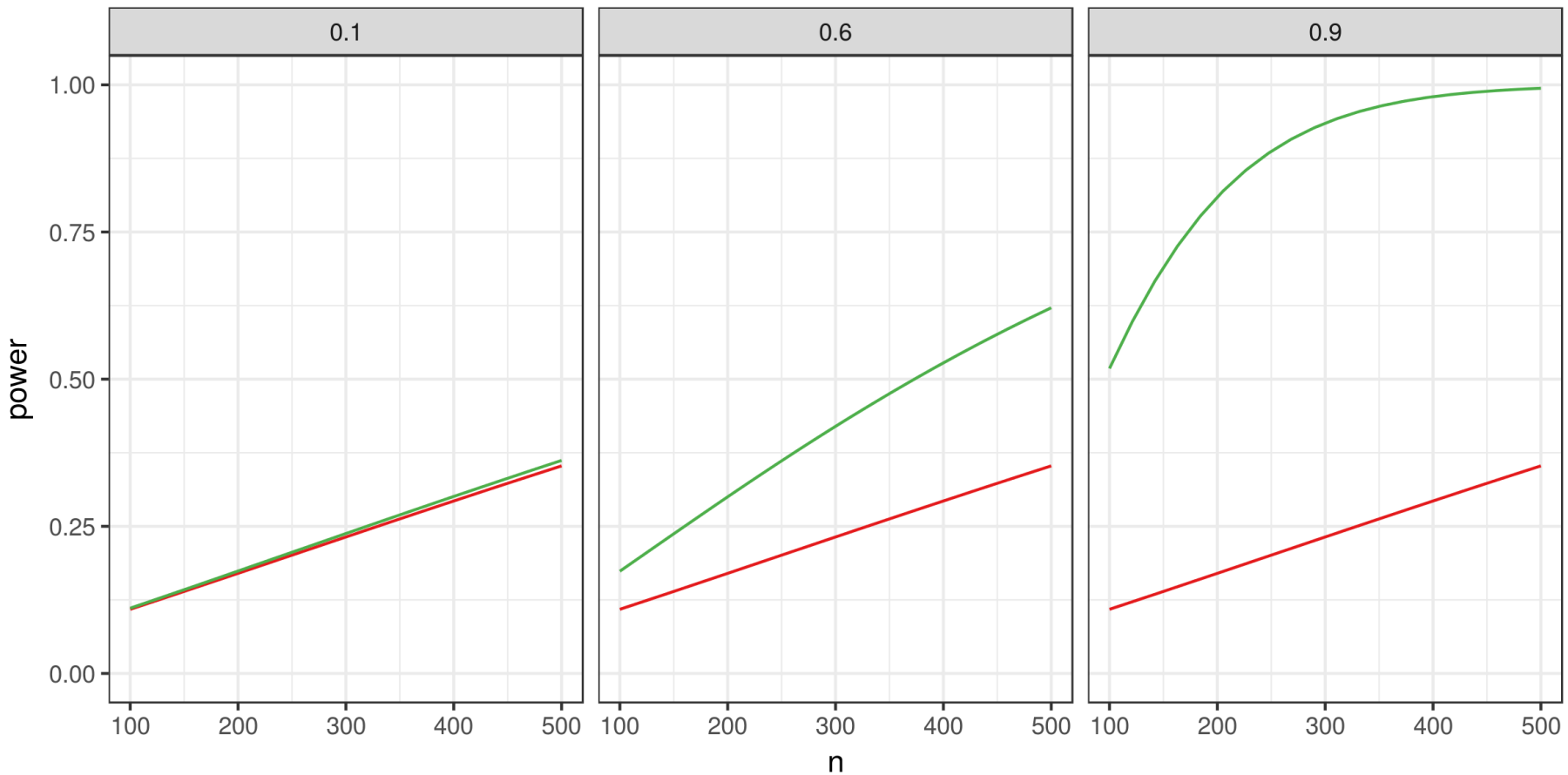
- $y \sim (X\beta, V) = (\mu + \beta_g x_g, \sigma_f^2 F + \sigma_r^2 I)$
- $\sigma_f^2 + \sigma_r^2 = 1$



Analytical results 2 (marginal): shared environment

For shared environment : $NCP = \beta^2 2pq\ tr((\hat{\sigma}_f^2 F + \hat{\sigma}_r^2 I)^{-1})$

- $\sigma_f^2 = \{0.1, 0.6, 0.9\}$
- The power in individuals with shared environment increases as more variance is explained



Summary on power comparisons

Marginal analysis

1. Study designs: unrelated $\approx$ genetically related
2. The power increases as more variance is explained
    - by taking into account shared environment

GxE interaction analysis

3. Study designs: genetically related $>$ unrelated
    - if exposure x_e and genotype x_g are independent

Ongoing work on more realistic scenarios

- Exposure x_e, other than binary and independent on genotype x_g
- Correlated exposure x_e and shared environment f
- Different family-designs designs: sib-pairs, trios, etc

Part 2: GxE stratified framework in structured population

GxE stratified framework

1. Compute marginal genetic effects in stratas, e.g., males and females
   - β_m and β_f, the genetic effects
   - σ_{β_m} and σ_{β_f}, their standard errors
2. Combine stratified results and perform tests
   - strata-specific, **interaction** (differentiated), joint, heterogeneity

	Stratified interaction test	**Reference**
Independent stratas	$Z_{int} = \frac{\beta_m - \beta_f}{\sqrt{\sigma^2_{\beta_m} + \sigma^2_{\beta_f}}} \sim \mathcal{N}(0, 1)$	(Magi *et al.*, 2010)
Related stratas	$Z_{int} = \frac{\beta_m - \beta_f}{\sqrt{\sigma^2_{\beta_m} + \sigma^2_{\beta_f} + r\sigma_{\beta_m}\sigma_{\beta_f}}} \sim \mathcal{N}(0, 1)$	(Randall *et al.*, 2013)

r is the spearman correlation between the two tests

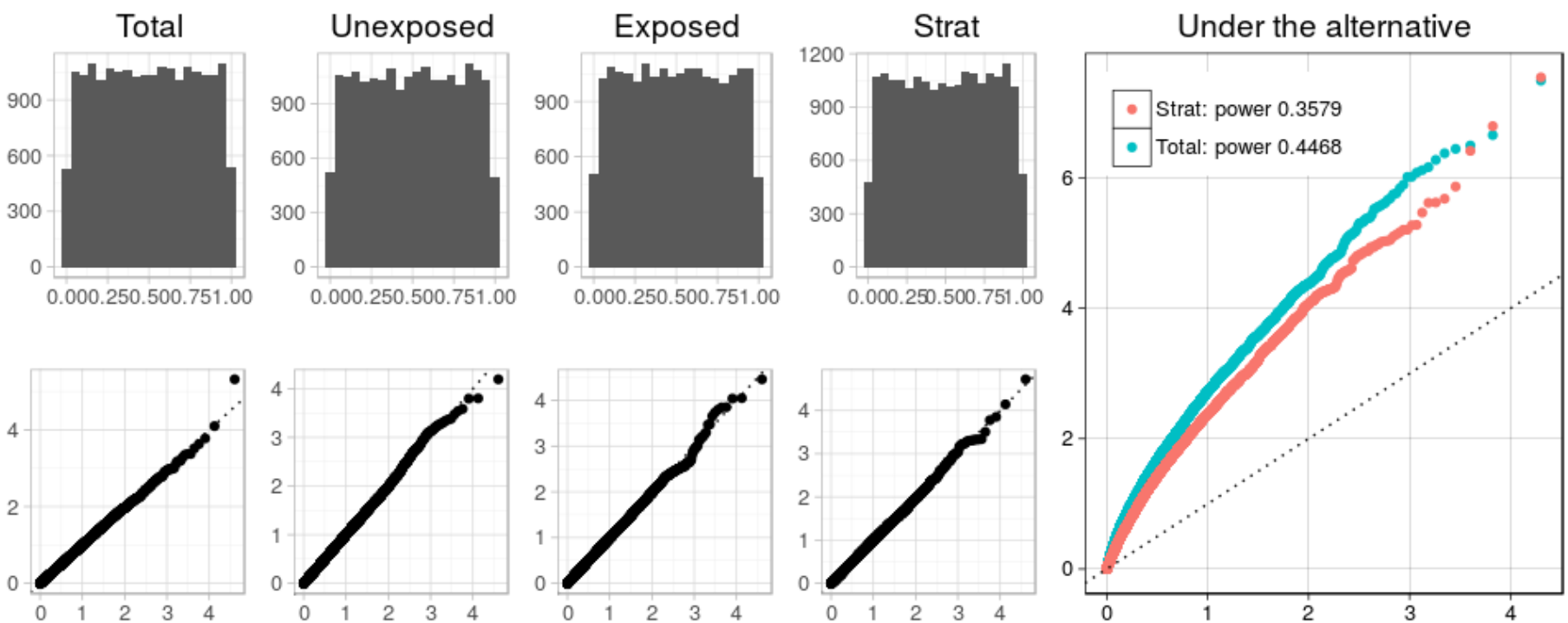
- a naive approach that needs further investigation (Sofer *et al.*, 2016, GenEpi)
- an empirical comparison between the stratified and full frameworks shows little agreement in family-based cohorts (Sung *et al.*, 2016)

Simulation results 3: stratified ≈ full

Data simulation of the whole sample (nuclear families + shared environment):

- $y \sim (X\beta, V) = (\mu + \beta_{int} x_g x_e, \sigma_g^2 K + \sigma_g^2 F + \sigma_r^2 I)$
- 2,500 individuals in 500 nuclear families
- 20,000 SNPs under the null ($\beta_{int} = 0$); 10,000 under the alternative ($\beta_{int} = 0.1$)
- independent genetic and exposure variables

Output: $\rho = 0.167$ between stratas



Ongoing work

- Derive analytical formulas for the GxE stratified framework
- Assess whether the Spearman correlation is robust enough
- Explore more complex scenarios
  - outcomes in stratas are genetically correlated
  - imbalance in sample size

Bear in mind the results from the LD score regression for two outcomes

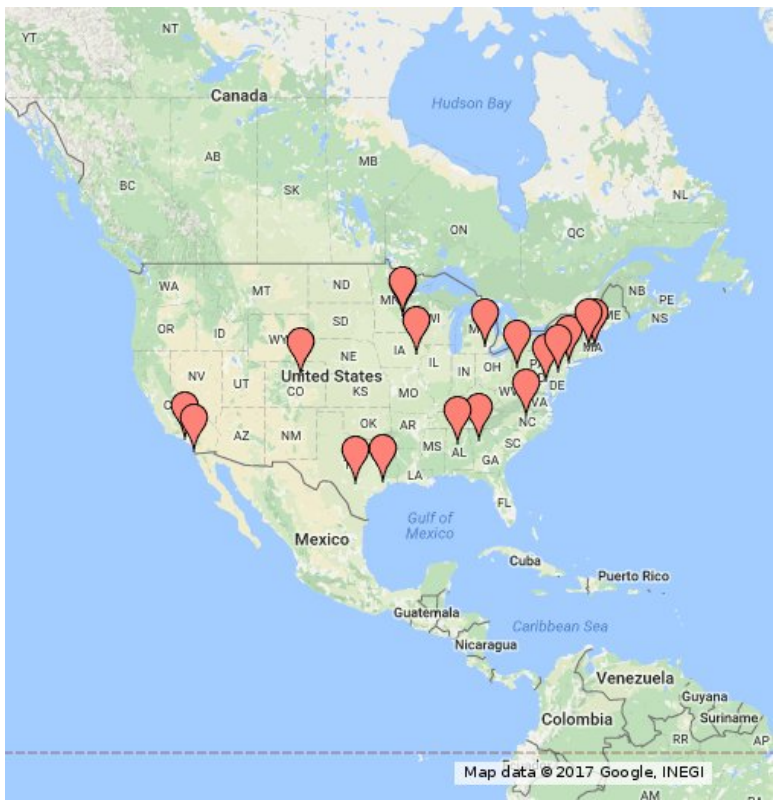
(Bulik-Sullivan *et al.*, 2015)

$$E[Z_{1j}Z_{2j}] = \frac{\sqrt{N_1 N_2}\rho_g}{M} l_j + \frac{N_s \rho}{\sqrt{N_1 N_2}}$$

Part 3: Ancestry x E analysis in admixed population

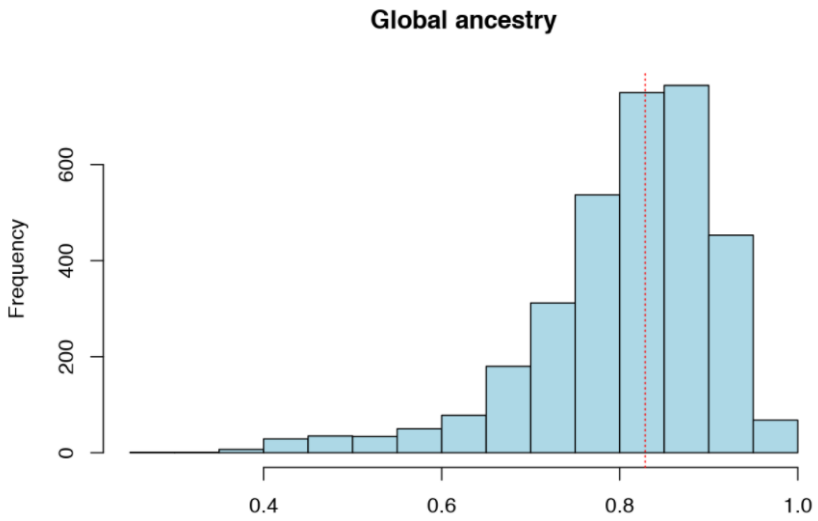
COPDgene project copdgene.org

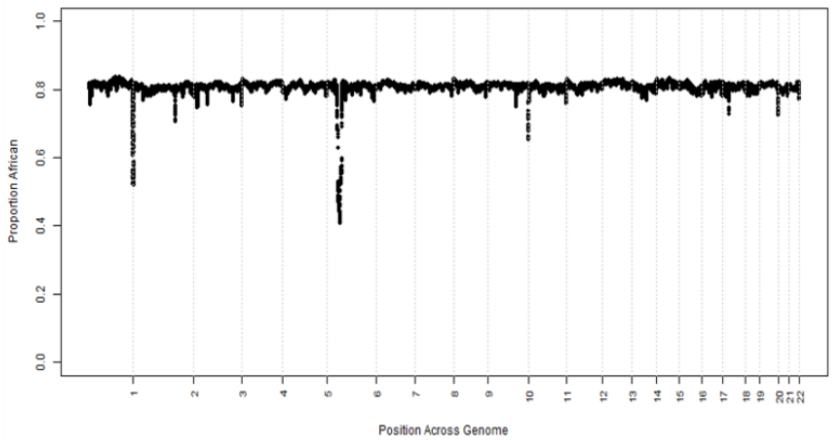
- 10,000 ever-smokers
- a rich set of COPD outcomes/exposures
  - 3,300 admixed African-Americans
  - SNPs, inferred local/global ancestry

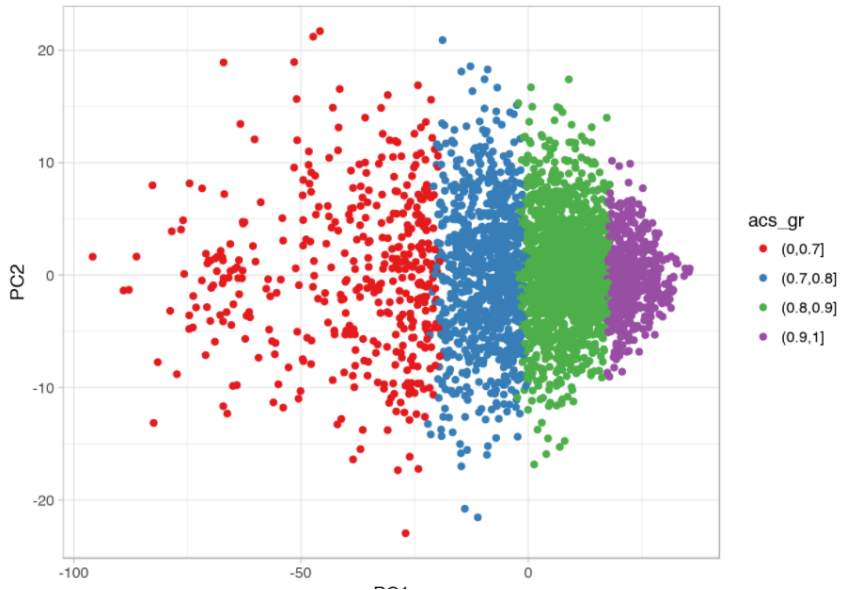


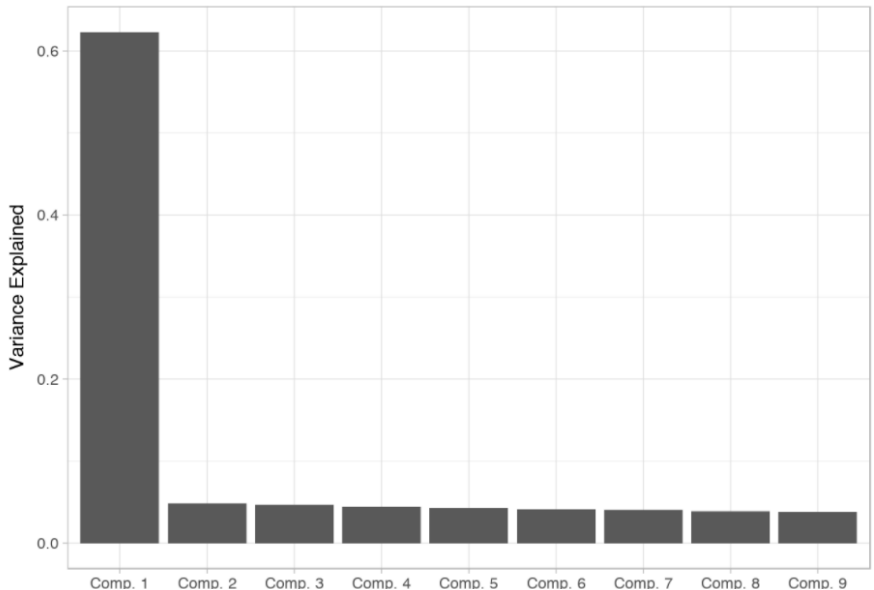
Abbreviations: COPD, Chronic obstructive pulmonary disease

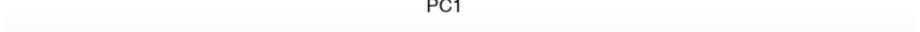
Ancestry data in COPDgene

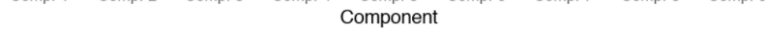









PC1


Component

Analytical plan

Previous studies reported

- Smoking is the major risk factor of COPD; gender is another established risk factor
- The proportion of global African ancestry is associated with COPD (Kumar *et al.*, 2010, N Engl J Med)
- In recent admixtures such as African-Americans, the local ancestry (locus-specific ancestry) can be accurately estimated (Baran *et al.*, 2011)
- Using local ancestry data instead of genotypes was explored in GxG analysis (Aschard *et al.*, 2015)

The project aims at leveraging the ancestry information in GxE tests

- global ancestry × exposure ($a_g \times x_e$)
- local ancestry × exposure ($a_l \times x_e$)
- SNP × exposure ($x_g \times x_e$)

1. Model for the genome-wide interaction scan on genotypes:

$$y = \mu + x_g\beta_g + x_e\beta_e + x_g \times x_e\beta_{int} + a_g\beta_1 + a_l\beta_2 + a_g \times x_e\beta_3 + a_l \times x_e\beta_4 \ (\beta_{int} = 0)$$

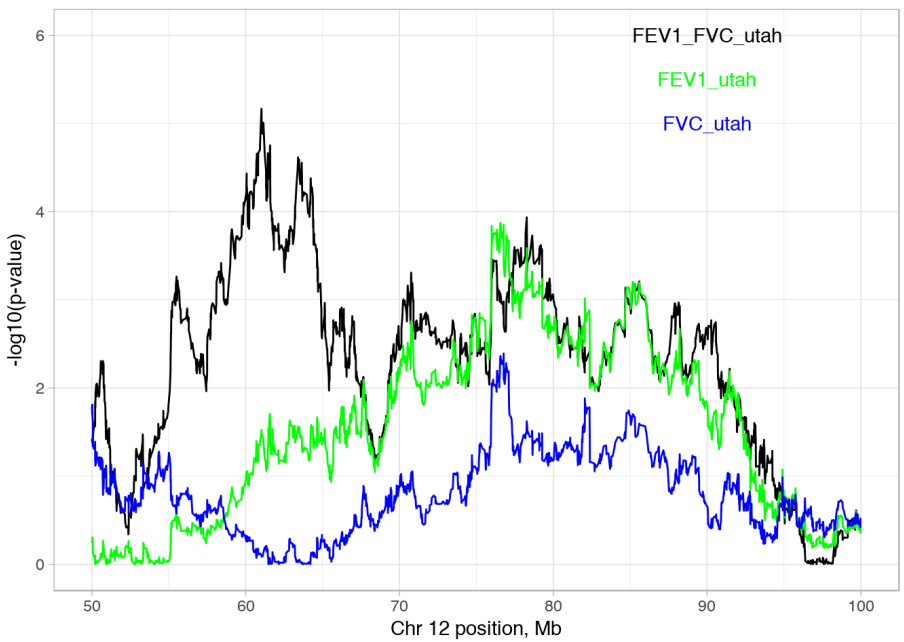
2. Model for the genome-wide interaction scan on local ancestry:

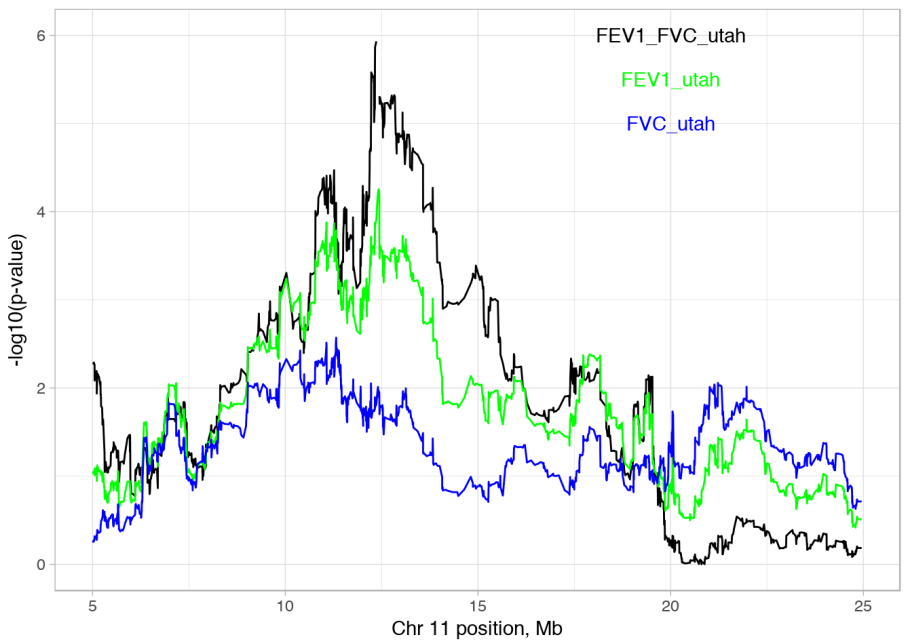
$$y = \mu + a_g\beta_1 + a_l\beta_2 + a_g \times x_e\beta_3 + a_l \times x_e\beta_4 \ (\beta_4 = 0)$$

Preliminary results

Marginal scan replicates the locus in Chr 12, Gene *FAM19A2* (Parker et al. 2014)

G x current-smoker scan suggests the locus in Chr 11, Gene *PARVA*, that (Wan, et al. 2015) – smoking-associated site-specific differential methylation in buccal mucosa in COPDgene

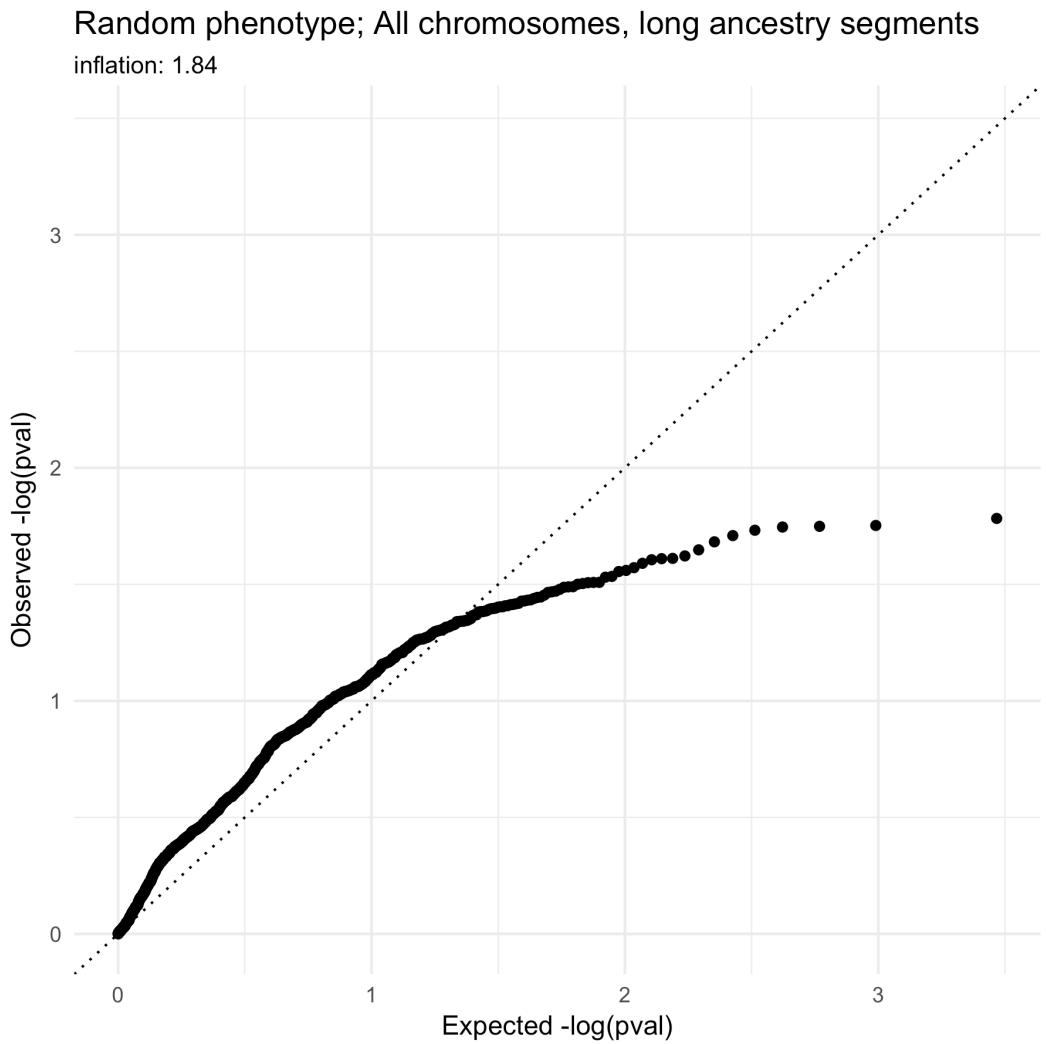




But we observe inflated Type I error in QQ-plots

Current status: cope with relatedness due to ancestry

Simulation on the null outcome: $y_{null} = \mu + \beta a_l + e$



Open question: LMM applicable to ancestry scans?

Data	Relationship Matrix	Method
Genotypes (2M)	GRM	LMM
Local ancestry (40K)	ARM	LMM

Conclusions

Conclusions

1. The derived formula $NCP \approx \beta^2 tr(\hat{V}^{-1}\Sigma_x)$ for the power
   - can guide the choice related *vs.* unrelated study desing
   - analytically shows: family-based design can be beneficial for GxE
   - more derivations: upper bounds; simplified formula for sib-pairs

2. Ongoing projects to be completed
   - Analytical validation of the stratified framework
   - Control for structure in ancestry-based association analysis

3. The last, but not the least: github.com/variani/lme4qtl R package
   - extends lme4
   - flexible structure of random effects, including custom covariances
   - longitudinal, survival study designs, etc

Thank you

Extra slides

Possible study designs for comparison

	Study design 1	**Study design 2**	**Study design 3**
Sample	Family-based	Population-based	Population-based
Relationships	Kinship		GRM
Method	Linear mixed models	Linear models	Linear mixed models

GxE in study design 3 is our ongoing work (not presented today)

GxE in study designs 1 *vs.* 2 (today focus)

- Compare two study designs unrelated/related
- LMM performed using our new lme4qtl R package (under submission)

Simulation study

Given: a population of 50,000 related samples (nuclear families)

Experiment: pool 5,000 unrelated samples or pool randomly

relatedness	$\mathbf{V}$	$\mathbf{\Sigma_x}$	Normalization
unrelated	$\sigma_g^2 K + \sigma_r^2 I = (\sigma_g^2 + \sigma_r^2) I$	$\sigma_x I$	$\sigma_g^2 + \sigma_r^2 = 1$
genetically related	$\sigma_g^2 K + \sigma_r^2 I$	$\sigma_x K$	$\sigma_g^2 + \sigma_r^2 = 1$

- K, the double kinship matrix
- Σ_x, the variance-covariance matrix of predictor x
- σ_g^2, σ_r^2, variance proportions

GAIT2 Spanish families (previous project)

The Genetic Analysis of Idiopathic Thrombophilia 2 (GAIT2) Project

- Study of Venous Thrombosis
  - disease prevalence <1%
  - heritability $\sim$ 60%
- 935 individuals in 35 families (27 per family on average)
- Hundreds of phenotypes (blood coagulation system)
- Genotype and RNA-seq data

Developed tools for analysis of family-based samples

- solarius R package [makes SOLAR easier to use]
- lme4qtl R package [makes lme4 flexible]