Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals

Presented by Fred Boehm

4/7/23

Introduction

Motivation

- Individuals of admixed ancestries inherit a mosaic of local ancestry segments
- Offers the opportunity to investigate the similarity of genetic effects on traits across ancestries in a single population

Main conclusions

- After analyzing 38 complex traits in 53,001 African-European individuals:
  - very high correlations of causal effects across local ancestries
  - meta-analysis $r_{admix} = 0.95$

Figure 1

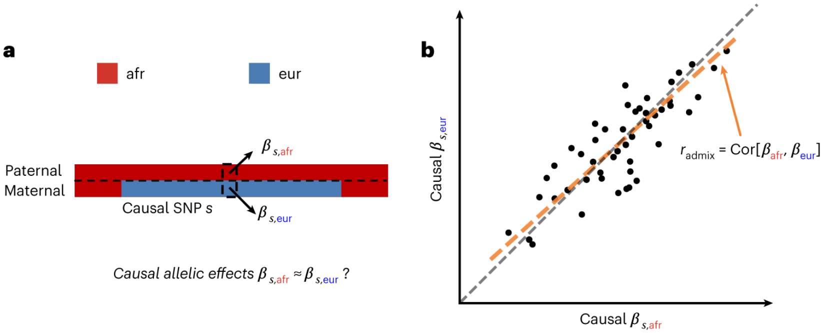

Figure 1: Concepts of estimating similarity in the causal effects across local ancestries

Statistical models

Phenotype model for admixed individuals

For individual $i = 1, \ldots, N$, SNPs $s = 1, \ldots, S$, and ancestries $k = 1, 2$, we have

$$g_{i,s,k} = x_{i,s,M} 1_{(\gamma_{i,s,M}=k)} + x_{i,s,P} 1_{(\gamma_{i,s,P}=k)}$$

- $x_{i,s,M}$ and $x_{i,s,P}$ are the number of minor alleles in maternal and paternal haplotypes, respectively
- $g_{i,s,k}$ encodes allele counts that are specific to the local ancestry

Phenotype model for admixed individuals

- Denote the causal allelic effects by $\beta_k \in \mathbb{R}^S$ for $k = 1, 2$
- Each individual's phenotype is then modeled as:

$$y_i = c_i^T \alpha + \sum_{s=1}^{S} \left(g_{i,s,1} \beta_{s,1} + g_{i,s,2} \beta_{s,2} \right) + \epsilon_i$$

- $c_i \in \mathbb{R}^C$ is a vector of covariates, including an intercept
- $\alpha \in \mathbb{R}^C$ is a vector of covariate effects
- ϵ_i is a random error term

Phenotype model for admixed individuals

- Aggregating $g_{i,s,k}$ over all SNPs s and all subjects i gives matrices $G_k \in \{0, 1, 2\}^{N \times S}$

$$y = C\alpha + G_1\beta_1 + G_2\beta_2 + \epsilon$$

- $C \in \mathbb{R}^{N \times C}$ is a matrix of covariates

Phenotype model for admixed individuals

- Model β_1, β_2 as:

$$\begin{bmatrix} \beta_{s,1} \\ \beta_{s,2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tau_s^2 \begin{bmatrix} \frac{\sigma_g^2}{S} & \frac{\rho_g}{S} \\ \frac{\rho_g}{S} & \frac{\sigma_g^2}{S} \end{bmatrix} \right)$$

- $\epsilon_i \sim N(0, \sigma_e^2)$
- τ_s denotes SNP-specific parameters for effects distribution

Phenotype model for admixed individuals

- Define correlation of genetic effects as

$$r_{admix} = \frac{\rho_g}{\sigma_g^2}$$

- $r_{admix} = 1$ implies that $\beta_{s,1} = \beta_{s,2}$ for all SNPs s
- $r_{admix} < 1$ indicates differences in causal effects between ancestries

Specifying τ_s under different heritability models

- τ_s parameters model the coupling of SNP effects variance with MAF, local LD or other functional annotations
- Previous research has shown that genetic correlation estimation is robust to heritability model choice
- Present work's authors mainly use frequency-dependent model for both simulations and real data analyses
  - set $\tau_s^2 \propto (f_s(1 - f_s))^\alpha$
  - f_s is MAF of SNP s
  - α set to fixed value of -0.38

Evaluation of genome-wide genetic effects consistency

- Marginalize over random effects β_1 and β_2 to obtain

$$y \sim N\left(C\alpha, \sigma_g^2 \frac{G_1 T G_1^T + G_2 T G_2^T}{S} + \rho_g \frac{G_1 T G_2^T + G_2 T G_1^T}{S} + \sigma_e^2 I\right)$$

- T is a diagonal matrix with $T_{ss} = \tau_s^2$ for all s

Evaluation of genome-wide genetic effects consistency

- Let $K_1 = \frac{G_1 T G_1^T + G_2 T G_2^T}{S}$ and $K_2 = \frac{G_1 T G_2^T + G_2 T G_1^T}{S}$
- Write $\rho_g = \sigma_g^2 r_{admix}$ to get

$$y \sim N(C\alpha, \sigma_g^2(K_1 + r_{admix} K_2) + \sigma_e^2 I)$$

Evaluation of genome-wide genetic effects consistency

- While MLE of $(\alpha, \sigma_g^2, r_{admix}, \sigma_e^2)$ can be found by maximizing $L(\alpha, \sigma_g^2, r_{admix}, \sigma_e^2)$, the constraint that $|r_{admix}| \leq 1$ is not easy to enforce
- Authors use profile likelihood instead:

$$L_p(r_{admix}) = \max_{(\alpha, \sigma_g^2, \sigma_e^2)} L(\alpha, \sigma_g^2, r_{admix}, \sigma_e^2)$$

Evaluation of genome-wide genetic effects consistency

- Perform grid search over r_{admix} values to maxiximize $L_p(r_{admix})$
- For each candidate r_{admix},
  - compute $K_1 + r_{admix}K_2$
  - solve for $(\alpha, \sigma_g^2, \sigma_e^2)$ for a single variance component model with GCTA
  - in practice, compute $L_p(r_{admix})$ for $r_{admix} \in \{0, 0.05, 0.1, \dots, 1\}$
  - use natural cubic splines to interpolate pairs of $(r_{admix}, L_p(r_{admix}))$ to get a smooth curve
  - set $\widehat{r_{admix}}$ to value that maximizes the likelihood curve
  - set credible interval as highest posterior density interval (assuming prior $U(0, 1)$ for r_{admix})

Evaluation of genetic effects consistency at individual variant with marginal effects

- For an individual SNP s and phenotype:

$$y = C\alpha + g_{s,1}\beta_{s,1}^{(m)} + g_{s,2}\beta_{s,2}^{(m)} + \epsilon$$

- Here, $\beta_{s,1}^{(m)}$ and $\beta_{s,2}^{(m)}$ are the marginal effects of the SNP
- Marginal effects tag effects from nearby causal SNPs with taggability as a function of ancestry-specific LD with causal SNPs
- Heterogeneity in marginal effects by local ancestry can be induced even if causal effects are the same

Figure 4: Induced heterogeneities in marginal effects across local ancestries

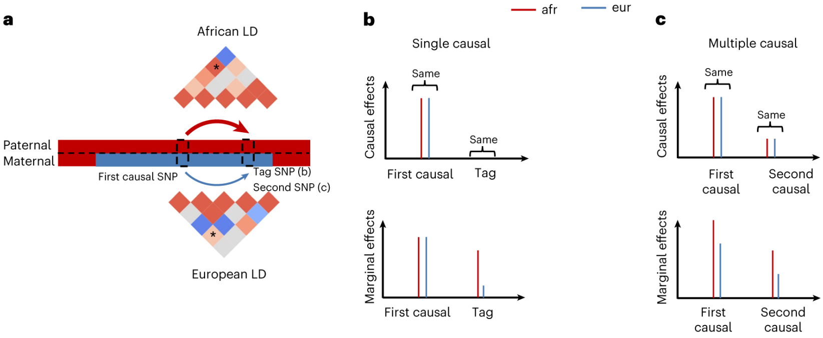

Figure 2: Induced heterogeneities in marginal effects across local ancestries

Evaluation of genetic effects consistency at individual variant with marginal effects

- Jointly estimate $\beta_{s,1}^{(m)}$ and $\beta_{s,2}^{(m)}$ with least squares
- Hypothesis testing by comparing above model to:

$$y = C\alpha + (g_{s,1} + g_{s,2})\beta_s^{(m)} + \epsilon$$

Marginal effects-based methods for estimating heterogeneity

- Inputs: estimated marginal effects $\widehat{\beta_{s,1}^{(m)}}$ and $\widehat{\beta_{s,2}^{(m)}}$ and their standard errors
- 3 approaches:
  - Pearson Correlation
  - OLS Regression Slope:

    $$\widehat{\beta_{s,1}^{(m)}} \sim \widehat{\beta_{s,2}^{(m)}}$$

    - Fails to model errors in independent variable
    - Assumes homogenous errors in dependent variable across SNPs
  - Deming Regression Slope:

    $$\widehat{\beta_{s,1}^{(m)}} \sim \widehat{\beta_{s,2}^{(m)}}$$

    with SEs
    - Deming regression models heterogeneous errors in both independent and dependent variables
    - More robust than above methods

Deming Regression

- Inputs: y_i and x_i, $\sigma_{x,i}$ and $\sigma_{y,i}$ for $i = 1, \dots, n$
- Optimizes the following objective function:

$$\min_{\beta,\alpha,\delta_1,\dots,\delta_n,\epsilon_1,\dots,\epsilon_n} \sum_{i=1}^{n} \left[\frac{\epsilon_i^2}{\sigma_{y,i}^2} + \frac{\delta_i^2}{\sigma_{x,j}^2} \right]$$

subject to:

$$y_i + \epsilon_i = \alpha + \beta(x_i + \delta_i)$$

for $i = 1, \dots, n$

Deming Regression

- Notably, Deming regression slope produces symmetric results for the two regression orders
- Can still produce biased errors if the standard errors are misspecified
- SEs of α and β can be bootstrapped

Simulation studies

Simulation studies

- To include local ancestry in estimating effect heterogeneity:

$$y = l_s \beta_{s,lanc}^{(m)} + g_{s,1} \beta_{s,1}^{(m)} + g_{s,2} \beta_{s,2}^{(m)} + c^T \alpha + \epsilon$$

- $\beta_{s,lanc}^{(m)}$: local ancestry effect

Simulation studies

- For “local ancestry regressed”:

$$y = l_s\beta^{(m)}_{s,lanc} + g_{s,1}\beta^{(m)}_{s,1} + g_{s,2}\beta^{(m)}_{s,2} + \epsilon$$

- First estimate $\beta^{(m)}_{s,lanc}$ by regressing y on l_s
- Second, estimate $\beta^{(m)}_{s,1}, \beta^{(m)}_{s,2}$ by regressing $y - \widehat{\beta^{(m)}_{s,lanc}}$ on $g_{s,1}$ and $g_{s,2}$

Simulations studies

- To assess impact of including local ancestry term when applying HET test:
    - Randomly select 1000 SNPs on Chr 1 from PAGE genotype data
    - Simulate traits with a single causal SNP, using each of the 1000 SNPs as the causal SNP
        - Simulate quantitative traits with different values of $\beta_{Eur} : \beta_{Afr}$: $(1.0, 1.05, 1.10, 1.15, 1.20)$
    - Scale effects such that the causal SNP explains the correct amount of heritability
    - For each causal SNP, repeat simulations of effects and random error 30 times
    - Apply the different strategies for including local ancestry to the simulated traits
    - Get p-values for HET testing $H_0 : \beta_{eur} = \beta_{afr}$
    - Included top PCs as covariates

Simulations studies

- Evaluate FPR or HET test power by subsampling without replacement:
  - Draw 100 random samples, each with 500 SNPs chosen from the pool of 1000 SNPs and 30 simulations
  - This approach accounts for randomness from both random errors and SNP MAFs
  - Calculated FPR for each sample of 500 SNPs
    - Obtained empirical distributions of FPR
    - Calculated mean & SE from empirical distribution

Figure 5

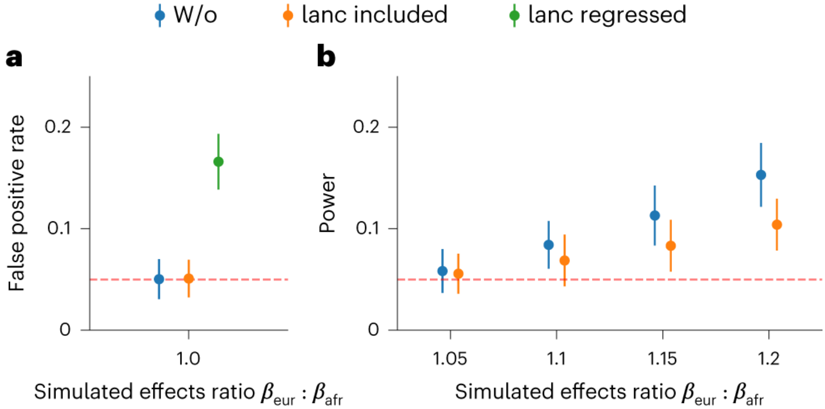

Figure 3: Pitfalls of including local ancestry in estimating heterogeneity

Data Analysis

Data Analysis

- Meta-analyzed 3 large studies: PAGE, UKB, and AoU

Figure 3

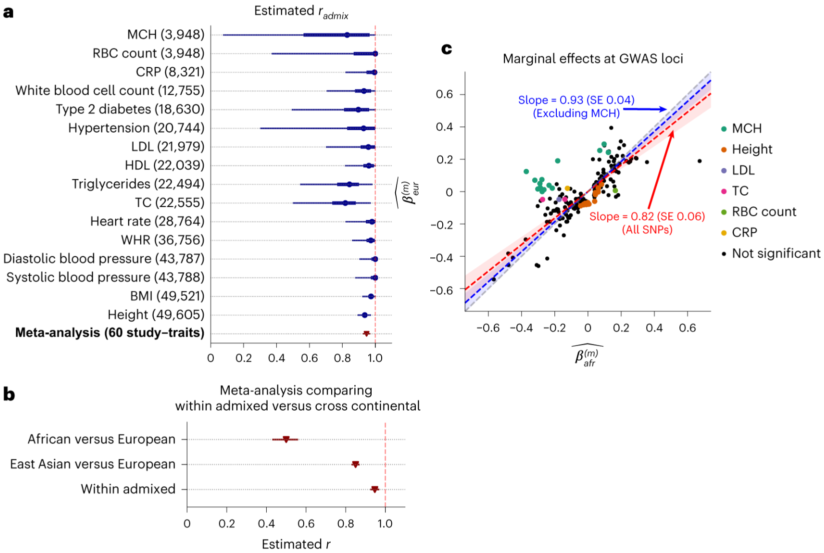

Figure 4: Similarity of causal effects and marginal effects across local ancestries from PAGE, UKB, and AoU

Discussion

Discussion

- Methods for heterogeneity by ancestry estimation from marginal GWAS SNP effects are susceptible to inflated heterogeneity estimates
- HET test may yield false positives when causal variants unknown
- Deming regression robust in low polygenicity settings; susceptible to inflated heterogeneity estimates in high polygenicity settings
- OLS Sl pe method biased due to failure to account for uncertainty in estimated effects

Limitations

- SNP MAF threshold of 0.005 for both ancestries
  - Simulations revealed that omission of rare variants could lead to downward bias in r_{admix}
  - Rare & population-specific causal variants can lead to upward bias in r_{admix}
- Limited consideration to two-way admixed individuals (African and European)
  - Extension to 3-way admixed individuals requires additional modeling due to error in local ancestry inference
- Extend methods to estimate correlations in causal effects stratified by functional annotations
- Deming regression not fully robust to high polygenicity settings
- Meta-analyzed 3 studies: PAGE, UKB, and AoU
  - Large SEs for individual traits