Transferability of polygenic risk prediction across diverse and admixed populations

Alicia R. Martin,[1,2,3] Duncan Palmer,[1,2,3] Raymond K. Walters,[1,2,3] Hailiang Huang,[1,2,3] Chia-Yen Chen,[1,2,3,4] Stephan Ripke,[1,2,3,5] Max Lam,[6] Christopher R. Gignoux,[7,8] Genevieve L. Wojcik,[7] Jose Sergio Hleap,[9] Simon Gravel,[9] Carlos D. Bustamante,[7] Eimear E. Kenny,[10] Benjamin M. Neale,[1,2,3] and Mark J. Daly[1,2,3]

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [2] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [3] Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4] Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [5] Department of Psychiatry, Charite Universitatsmedizin Berlin Campus Benjamin Franklin, Berlin, Germany; [6] Research Division, Institute of Mental Health, Singapore; [7] Department of Genetics, Stanford University, Stanford, CA 94305, USA; [8] Colorado Center for Personalized Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA; [9] Department of Human Genetics, McGill University, Montreal, Quebec, Canada; [10] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Abstract

- GWAS are heavily biased towards European populations.
- Generalizability of Eurocentric GWAS to diverse/admixed populations most likely depends on LD and allele frequencies.
- Genetic risk prediction using standard approaches across global populations exhibits biases not explained by divergence or natural selection.
- Our large, coalescent-based simulations quantitatively demonstrate that summary statistics derived from European populations generalize poorly to non-European populations.
- We are developing novel statistical methods that model LD from multiple populations to improve the generalizability of genetic risk prediction.

References

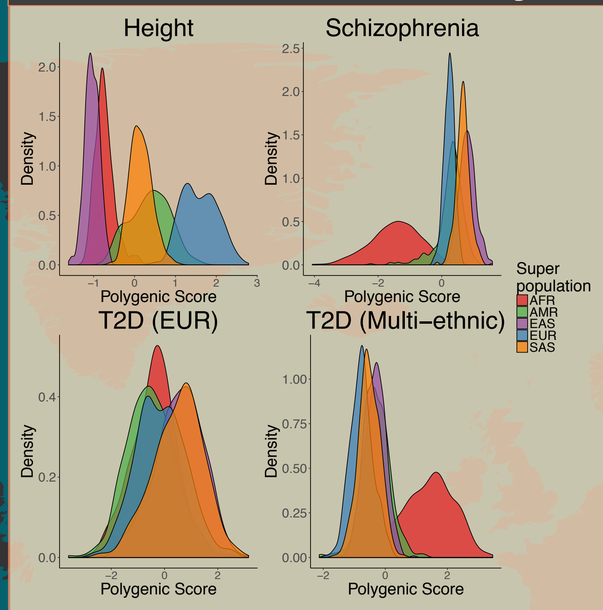
- **Polygenic risk biases** (Martin et al, 2017): https://broad.io/prs.
- **Pruning and thresholding** (Purcell et al, 2009): http://go.nature.com/2yMnq5Q.
- **LDPred** (Vilhjálmsson et al, 2015): https://broad.io/ldpred.
- **MultiPRS** (Marquez-Luna et al, 2016): https://broad.io/multiprs.
- **MTAG** (Turley et al, 2017): https://broad.io/mtag.

Contact

Alicia Martin
armartin@broadinstitute.org
@genetisaur

Poster link

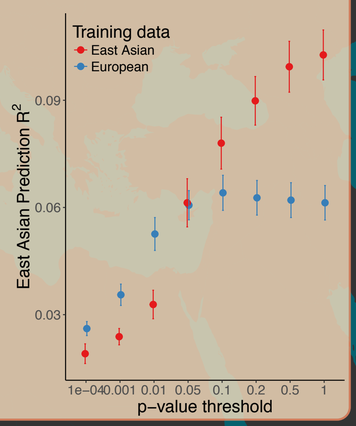
GWAS Transferability Across Populations



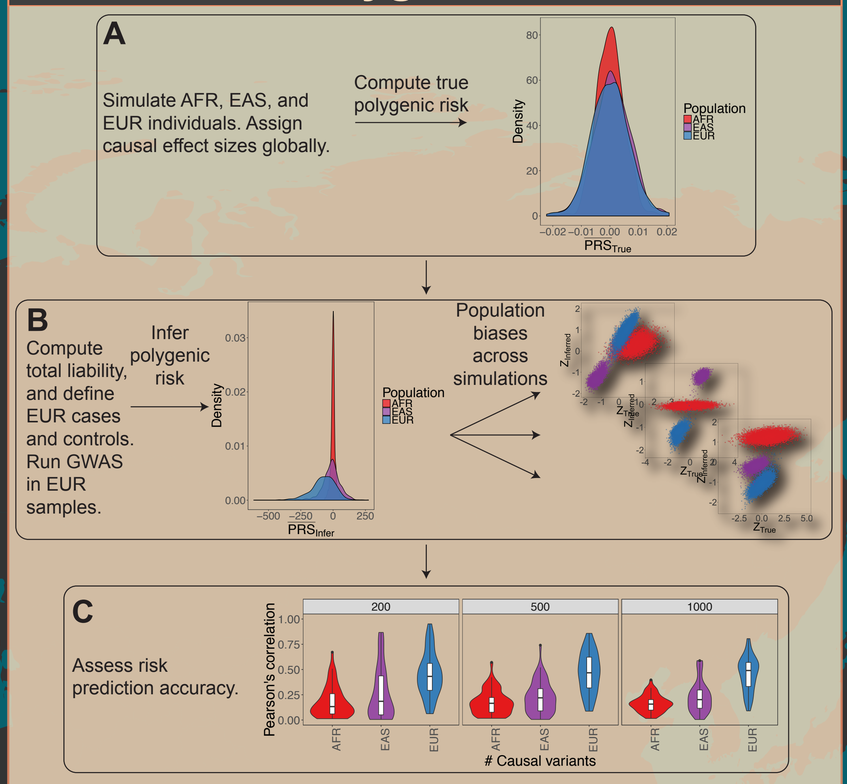
- **Height bias**: Predicted decrease with distance from Europe, contradicting anthropometric evidence.
- **Schizophrenia prevalence**: similar across globe, but Africans predicted to have lower risk.
- **T2D conflicting order**: Predicted risk depends on population summary statistics.

- **Empirical example (schizophrenia):** Psychiatric Genetics Consortium GWAS
  - 37k and 113k European cases/controls
  - 13k and 16k East Asian cases/controls (leave-one-out prediction strategy)

Despite a 3-fold larger European GWAS, prediction accuracy is higher in East Asians with matched summary statistics.



Biases in Polygenic Risk Scores



- **Simulation design**: We simulated GWAS-scale data (200k samples/population, 5% prevalence →10k cases, 10k controls with liability threshold model) with chr20 recombination map (~65 Mb), selecting causal variants globally (EUR, EAS, AFR), but GWAS participants only from Europe. Causal effect sizes drawn from:

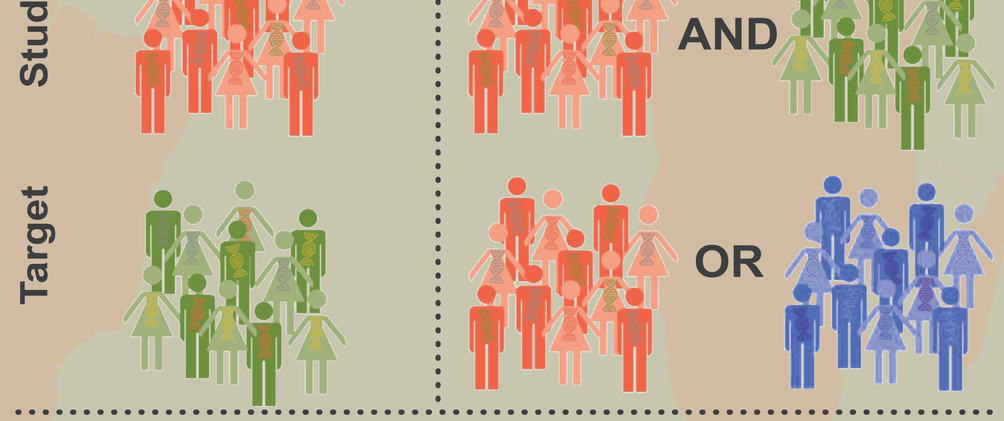
$$\beta \sim N(0, h^2/m)$$

- **True vs inferred risk**: Across simulations, true risk does not differ across populations. Inferred risk, however, varies most in Europe, then Asia, and least in Africa.
- **Risk biases across populations**: Within a given simulation, true vs inferred risk can occur in any direction from drift alone. Correction with principal components alone (e.g. as typical in GWAS) will be insufficient to correct these biases.

New Polygenic Risk Prediction Methods

Scenario 1) Study population divergent from target population

Scenario 2) Multiple study populations, same/different target population



Bias

In GWAS: $y = x\beta + \epsilon$

where β is causal effect size and $\hat{\beta}$ is estimated effect size.

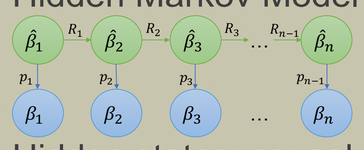
$$E[\hat{\beta}_{j,1}] = \beta_j + \sum_{i=1}^{m} \beta_i * R_{i,j,1}$$

R matrix is unsquared correlation matrix (LD)

Approach

Kalman filter

Analogous to a continuous-state Hidden Markov Model



Hidden states: causal effect sizes (β)
Observed states: estimated effect sizes
Linear operator is a function of LD (R)

Improved meta-analysis

Builds off recent previous work on Multi-Trait Analysis of GWAS (MTAG): can be applied to summary statistics, accounts for sample overlap, estimates trait-specific effects.
Compute covariance of estimated effects with pop-specific LD matrix:

$$Cov[\hat{\beta}_{j,1}, \hat{\beta}_{j,2}] = Cov[\beta_j + \sum_{i=1}^{m} \beta_i * R_{i,j,1}, \beta_j + \sum_{i=1}^{m} \beta_i * R_{i,j,2}]$$

Assuming same causal variant:

$$Cov[\hat{\beta}_{j,1}, \hat{\beta}_{j,2}] = 1 + Cov[R_{i,j,1}, R_{i,j,2}] * Var(\beta_i)$$

Conclusions and Future Directions

- **As a field, genetics needs to make greater strides towards inclusivity and increasing diversity!**
- Pruning and thresholding approach is biased across populations. Reduced accuracy with increasing divergence and unpredictable mean shifts raise interpretability issues.
- Extend simulation framework to model natural selection, test liability threshold and heritability differences across populations, and assess prediction accuracy with different causal allele frequency distributions.
- Create a new risk prediction method that models local ancestry in recently admixed populations.