Association between polygenic risk scores computed across ancestry groups

Laura M. Schultz (PhD)[1], Raquel E. Gur (MD, PhD)[2,3,4], Alison K. Merikangas (MPH, PhD)[1,3,5], Monica E. Calkins (PhD)[2,3], Ruben C. Gur (PhD)[2,3], Tyler M. Moore (PhD)[2,3], Kosha Ruparel (MSE, MS)[2,3], Ran Barzilay (MD, PhD)[2,3,4], Laura Almasy (PhD)[1,3,5]

[1]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia; [2] Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania; [3] Lifespan Brain Institute of the Children's Hospital of Philadelphia and Penn Medicine ; [4] Department of Child Adolescent Psychiatry and Behavioral Sciences, Children's Hospital of Philadelphia; [5] Department of Genetics, Perelman School of Medicine, University of Pennsylvania

INTRODUCTION

Polygenic risk scores (PRS) are commonly computed using summary statistics from a genome-wide association study (GWAS) run for an independent population from the same ancestry group. Since the vast majority of GWAS have been run only for European populations, computing PRS for non-European ancestries often forces researchers to rely on European GWAS, with limited data to support this approach. Using GWAS summary statistics from the Psychiatric Genomics Consortium-Posttraumatic Stress Disorder Group (PGC-PTSD) that were computed separately for European (EUR) and African-American (AFR) populations (Nievergelt et al., 2019), we explored the impact of using trans-ancestry discovery GWAS by computing PTSD PRS for EUR (*n*=5239) and AFR (*n*=3260) ancestry subgroups of the Philadelphia Neurodevelopmental Cohort (PNC) (Calkin et al, 2015). We confirmed our results by also computing trans-ancestry PTSD PRS for EUR (*n* = 5815) and AFR (*n* = 1741) subgroups of the Adolescent Brain Cognitive Development Study (ABCD) (Uban et al., 2018).

METHODS

Genotype data were obtained from dbGaP (phs000607.v2.p2) and NDA (#2573, fix release 2.0.1) for the PNC and ABCD, respectively. The PNC samples were genotyped in 15 batches on 10 different types of Affymetrix and Illumina chips by the Center for Applied Genomics at the Children's Hospital of Philadelphia (Glessner et al. 2010), whereas the ABCD samples were genotyped on the Affymetrix NIDA SmokeScreen array by RUCDR at Rutgers. The PNC dataset was processed by chip batch and not merged until after imputation.

- *Pre-Imputation QC:* Plink 1.9 (Chang et al., 2015) was used to remove SNPs with > 5% missingness, samples with more than 10% missingness, and samples with a genotyped sex that did not match the reported sex phenotype.
- *Imputation:* Genotypes were phased (Eagle v.2.4) and imputed to the 1000 Genomes Other/Mixed GRCh37/hg19 reference panel (Phase 3 v.5) using Minimac 4 via the Michigan Imputation Server (Das et al., 2016).
- *Post-Imputation QC:* Only polymorphic sites with imputation quality $R^2 \geq 0.7$ and MAF ≥ 0.01 were retained.
- *Ancestry Analysis:* Multidimensional scaling was conducted using KING (v.2.2.4; Manichaikul et al., 2010) to identify the top 10 ancestry principal components (PCs) for each sample. These PCs were projected onto the 1000 Genomes PC space, and ancestry was inferred for each sample using the e1071 support vector machines package in R. A second round of MDS was performed within the EUR and AFR groups to produce 10 PCs that we regressed out of the PRS in order to correct for chip batch effects and genetic ancestry.
- *PRS Computation:* PRS-CS (Ge et al., 2019) was used to infer posterior effect sizes of the SNPs in the dataset that overlapped with the PGC-PTSD GWAS summary statistics and an external 1000 Genomes linkage disequilibrium (LD) panel matched to the ancestry group used for the GWAS. PRS for the EUR and AFR subsets of the PNC and ABCD datasets were computed twice, once using the EUR LD panel and EUR GWAS and once using the AFR LD panel and AFR GWAS. Raw PRS were produced in plink 1.9 from the posterior effects. R was used to standardize and then correct the PRS by regressing out the top-10 within-ancestry PCs.

RESULTS

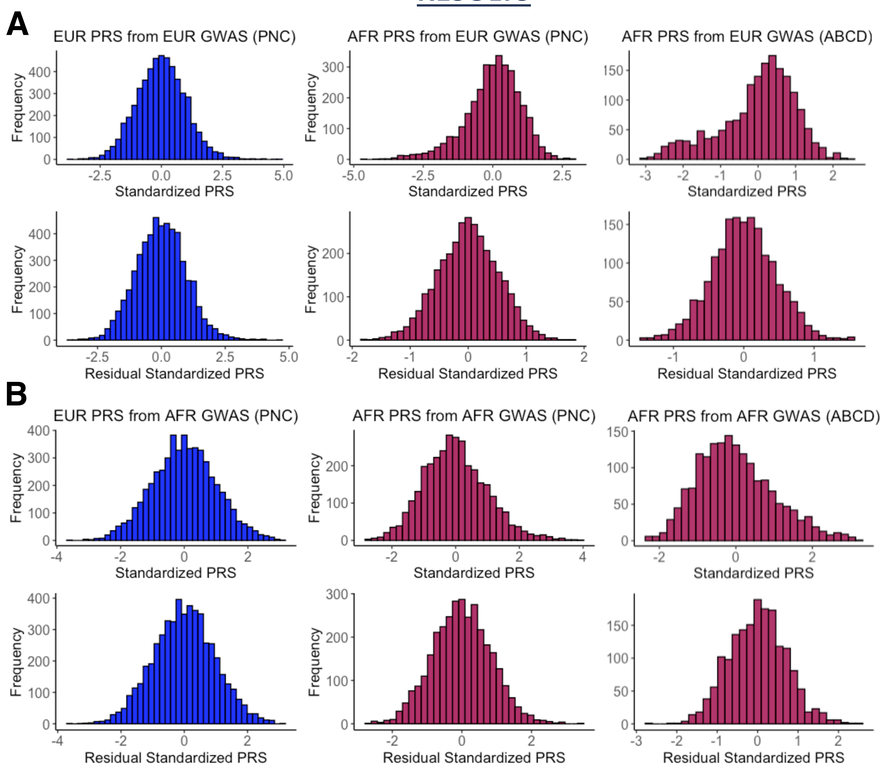

Figure 1. It is necessary to correct PRS for ancestry PCs.
(A) PTSD PRS computed from EUR GWAS are normally distributed for EUR cohorts both before and after correction by regressing out the top 10 within-ancestry PCs, whereas they are noticeably left skewed for AFR cohorts prior to correction. (**B**) PTSD PRS computed from AFR GWAS are normally distributed for EUR cohorts both before and after correction by regressing out the top 10 ancestry PCs, whereas they are noticeably right skewed for AFR cohorts prior to correction.

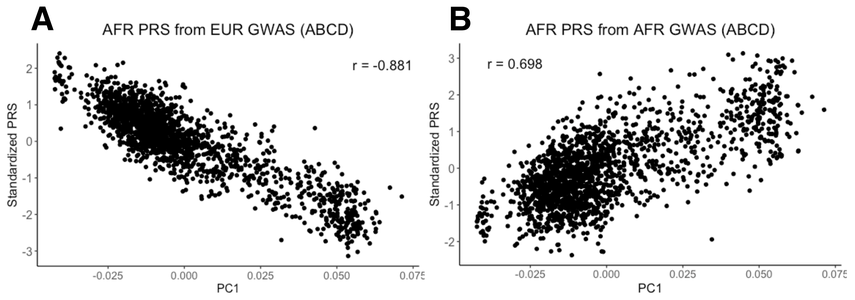

Figure 3. Uncorrected AFR PRS are strongly correlated with AFR ancestry PC1.
(A) PTSD PRS computed for AFR samples from EUR GWAS are negatively correlated with AFR PC1 (*r* = -0.891), but **(B)** the PRS computed for AFR samples from AFR GWAS are positively correlated with AFR PC1 (*r* = 0.698).

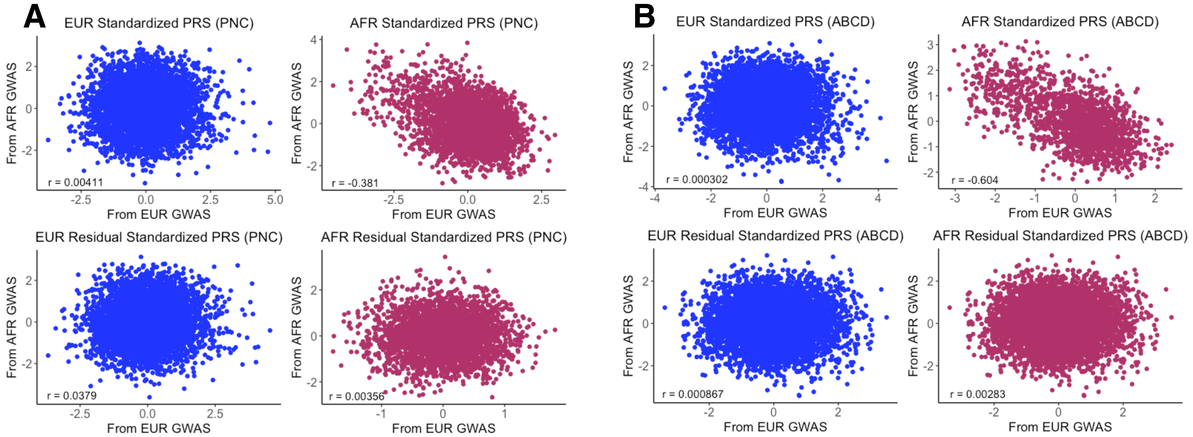

Figure 2. Trans-ancestry PRS are not correlated with each other. PTSD PRS computed for the EUR samples of (**A**) PNC and (**B**) ABCD from EUR GWAS are uncorrelated with the PRS computed for the same samples from AFR GWAS. Prior to correcting for ancestry PCs, PRS computed for the AFR samples from EUR GWAS are negatively correlated with PRS computed for the same samples from AFR GWAS. After correction, this correlation disappears.

DISCUSSION

- Given that we have shown that it is necessary to use a same-ancestry discovery GWAS when computing PRS, there is a critical need to run GWAS for AFR and other non-EUR populations.
- The results we report here for AFR PRS computed from EUR GWAS are not unique to PTSD. We have observed similar effects for autism spectrum disorder, bipolar disorder, major depressive disorder, and schizophrenia.
- While past work has shown that within-ancestry PRS explain a greater proportion of phenotypic variance than do cross-ancestry PRS (Curtis, 2018), our study is the first to investigate the correlation between within-ancestry and cross-ancestry PRS.
- Our future work will take advantage of the rich phenotyping of the PNC dataset to assess the predictive value of the PTSD PRS that we have computed for its EUR and AFR subgroups.

REFERENCES

Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., . . . Gur, R. E. (2015). The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. Journal of child psychology and psychiatry and allied disciplines., 56(12), 1356-1369.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4(1).

Curtis, D. (2018). Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. Psychiatr Genet, 28(5), 85-89.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. Nature Genetics, 48(10), 1284-1287.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature Communications, 10(1), 1776.

Glessner, J. T., Reilly, M. P., Kim, C. E., Takahashi, N., Albano, A., Hou, C., . . . Hakonarson, H. (2010). Strong synaptic transmission impact by copy number variations in schizophrenia. Proceedings of the National Academy of Sciences, 107(23), 10584.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867-2873.

Nievergelt, C. M., Maihofer, A. X., Klengel, T., Atkinson, E. G., Chen, C.-Y., Choi, K. W., . . . Koenen, K. C. (2019). International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. Nature Communications, 10(1), 4558.

Uban, K. A., Horton, M. K., Jacobus, J., Heyser, C., Thompson, W. K., Tapert, S. F., . . . Sowell, E. R. (2018). Biospecimens and the ABCD study: Rationale, methods of collection, measurement and early data. Developmental Cognitive Neuroscience, 32, 97-106.