Katherine Duchinski
5/1/2020

Leveraging European GWAS Results to Enhance Risk Assessment in Admixed Populations

Background

Polygenic risk scores have broad implications for early intervention in genetic diseases like breast cancer, diabetes, inflammatory bowel disease, and obesity. Disease prediction using polygenic risk scores has improved with the collection of more genome-wide association studies (GWAS). Khera et al. have leveraged GWAS data to estimate up to 5-fold increased risk for coronary artery disease in some individuals, including patients for whom non-genetic risk factors failed to identify increased risk[1]. Clinical implementation of polygenic risk scores could better inform practitioners when to recommend screening, preventative medicine, or lifestyle changes to patients.

Larger GWAS studies can identify less common genetic variants and have more statistical power. Study size becomes especially important in populations with more heterogeneous ancestry, such as African or Latin American populations, which are also at greater risk for contraction of and complications resulting from some disease such as adult-onset or type II diabetes[2]. These admixed populations are often scientifically underrepresented and medically underserved[3]. An often-acknowledged issue of current polygenic risk scores is the overrepresentation of European ancestry in available GWAS data[1,3,4]. Bentley et al. review potential barriers to improving GWAS diversity, which include low ethnic diversity in science, European bias in technology development, and mistrust in the scientific community due to historical exploitation[3]. While polygenic risk scores may become a highly valuable clinical tool, they would further widen health disparities if research focusing on admixed population is not prioritized.

Polygenic risk scores based on a European population have limited transferability, but public GWAS data for other populations encompass far fewer individuals than studies focusing on Europeans. The scale of GWAS data required for admixed populations such that the predictive power of polygenic risk scores would be comparable to the currently most successful European ancestry predictive models is unclear. While sufficient data remain unavailable for many groups, a large European study might improve accuracy when leveraged alongside a small study to make predictions for an underrepresented population. Grinde et al. recently investigated the portability of European GWAS polygenic risk estimation to Hispanic populations and found that large European studies outperform smaller Latin American studies for some traits[4]. The proposed study would attempt to predict phenotypes using both European ancestry and matched-ancestry polygenic risk scores and estimate the quantity of data needed such that European GWAS summary statistics would no longer be useful for prediction in admixed populations.

Objectives and Significance

The proposed study addresses the pervasive underrepresentation of medically underserved admixed populations in GWAS research and the difficult problem of polygenic risk score portability. The proposed study would investigate risk estimation for type II diabetes, a well-studied and highly polygenic disease. Because admixed African and Latin American communities in the United States are both medically underserved and disproportionately affected by type II diabetes, this study would particularly highlight the need for continued research on admixed populations to reduce systemic health disparities.

The proposed study aims to evaluate polygenic risk scores based on large European GWAS summary statistics and summary statistics from smaller, matched-ancestry GWASs by building disease status classifiers. The results would indicate the degree of generalizability of effect size estimates based on large European GWASs for underrepresented groups and the magnitude of the gap between prediction accuracy for patients of different genetic backgrounds. This study would further estimate the volume of data necessary to predict risk for admixed African and Latin American patients to the current standard for European patients. Success in this research area would help decrease the gap in risk assessment accuracy between European and non-European populations while large GWAS studies of non-European populations remain scarce.

Aim 1

The proposed study would evaluate predictive power for polygenic risk scores calculated from either large European GWAS data compared to scores calculated from small studies where ancestry is more appropriate for the subject population. I would construct polygenic risk scores for underrepresented populations based on both small, matched ancestry GWAS studies and European ancestry studies. I would then train machine learning models to predict type II diabetes status using each and both polygenic risk scores for each population and compare their performances.

Ancestry Deconvolution

I would first measure ancestry diversity in publicly available African and Latin American GWAS studies to determine how well the target population represents the validation population. I would utilize the 1000 Genomes Project, supplemented with Native American haplotypes from Mao et al., for reference haplotypes[5]. I would consider the Western and Northern European (CEU), Finnish (FIN), British (GBR), Iberian (IBS), and Toscani (TSI) as European and Esan (ESN), Gambian (GWD), Luhya (LWK), Mende (MSL), and Yoruba (YRI) as African reference populations. I would use the SHAPEIT2 algorithm to phase all haplotypes[6]. As in Martin et al., I would use the Mao et al. samples with > 99% Native American ancestry as the Native American reference populations[7]. To improve the accuracy of the inferred ancestry, I would run each LAMP-

LD, RFMIX, and HAPMIX and generate consensus calls as previously described by the 1000 Genomes Project Consortium.[8,9,10,11]

Polygenic Risk Score Construction

I would first filter the reference datasets for biallelic single nucleotide polymorphisms (SNPs) and remove ambiguous SNPs and variants with lower than 0.1% minor allele frequency (MAF). In order to obtain weights for constructing polygenic risk scores, I would clump summary statistics files to exclude variants in linkage disequilibrium. I would vary the source populations in consideration of the ancestral makeup of the admixed populations to ensure that they are, as much as possible, adequately represented. I would again use second-generation PLINK to construct polygenic risk scores for each phenotype and population, using the following p-value thresholds to identify the best model: $p < 5 \times 10^{-8}$, $p < 1\times 10^{-6}$, $p < 1 \times 10^{-4}$, $p < 1 \times 10^{-3}$, $p < 1 x 10^{-2}$, $p < 5 x 10^{-2}$, $p < 0.1$, $p < 0.2$, $p < 0.3$, $p < 0.4$, $p < 0.5$, $p < 0.75$, $p <= 1.0$. I would use principal component analysis to correct for population stratification as described in Price et al[10].

In addition to polygenic risk scores calculated based on European summary statistics (PRS_{EU}) and polygenic risk scores calculated based on matched-ancestry summary statistics (PRS_{MA}), I would attempt to estimate polygenic risk through a weighted average of summary statistics from each (PRS_{WA}). I would include all SNPs that passed the MAF filter in both and the p-value threshold in at least one of the target populations. I would calculate the log odds ratio of the effect sizes and average them between the European and matched-ancestry targets, weighted by the proportion of European ancestry present in the matched-ancestry population.

Predictive Model Training

I would use an array of machine learning models to build classifiers to predict disease status in independent validation datasets. All three classes of polygenic risk scores would be considered as predictors in combination and independently, along with traditional clinical risk factors. The non-genetic predictors would explicate likely confounding factors and provide a negative control classifier (model 0) that excludes all polygenic risk scores and would be useful for comparison to the other models (Table 1). Each model design would be used to train popular machine learning algorithms such as K-nearest neighbor, random forest, and elastic net.

Table 1: Proposed disease status classifier designs to be evaluated through this study.

Model No.	Design
0	Disease Status ~ Age + BMI + Waist : Hip Ratio + …
1	Disease Status ~ PRS_{EU} + Age + BMI + Waist : Hip Ratio + …
2	Disease Status ~ PRS_{MA} + Age + BMI + Waist : Hip Ratio + …
3	Disease Status ~ PRS_{WA} + Age + BMI + Waist : Hip Ratio + …
4	Disease Status ~ PRS_{EU} + PRS_{MA} + PRS_{WA} + Age + BMI + …

Evaluation and Expected Results

I would evaluate the disease classifiers based on accuracy and false negative rate (FNR) and the height predictive models on mean squared error (MSE). Because European ancestry would be present for most Latin American individuals, I expect that the best-performing model for Latin American populations would take into account European-based polygenic risk either through model 3 or 4 (Table 1). European haplotypes would likely capture only a small proportion of genetic variance present in African populations due to population bottlenecks in ancestral pre-European groups. I thus expect the utility of PRS_{EU} to be lower for prediction in African populations. I would also assess imputation accuracy for each population using leave-one-out validation.

Aim 2

The proposed study would attempt to estimate the quantity of GWAS data from an admixed population required to make a potential clinical impact comparable to that of current European polygenic risk scores. I would simulate GWAS, genotype arrays for European and non-European populations of different sizes. I would then calculate "true" and "estimated" polygenic risk based on each simulated population and use predictive models as above. I would finally estimate the prediction improvement with increased European and matched ancestry datasets.

Literature Review

I would first perform a literature review of polygenic scoring studies by searching the databases PubMed and bioRxiv since 2017 and combining these recent data with the literature review results reported by Duncan et al[13]. I would record the ethnicity and number of participants in each study and compare the accumulation of European-, African-, and Latin American-focused GWAS data over time. The collection rate and current amount of data for each population would be considered in the size of the simulated GWASs.

Simulation of European and Admixed Genotypes

I would simulate genotypes using msprime with GRCh38 and assuming a mutation rate $\mu = 2x10^8$ mutations per base per generation, as previously described[12,7]. I would use a previously reported 1000 Genomes demographic model to simulate European individuals[14]. To simulate an admixed African population, I would consider two ancestral populations: CEU and YRI. I would include these populations as well as Native American ancestry when simulating a Latin American population. The simulated proportions of each ancestry present in each individual will vary within a range based on observations from ancestry deconvolution in Aim 1 and those previously reported in Martin et al.[7] The simulated population sizes will be based on current and projected numbers from the literature review where admixed populations sizes will be some fraction of the European population size, e.g. 25%, 50%, 75%, 90%, 100%, and 200% of n = 200,000 Europeans.

Simulation of Polygenic Risk Scores

Following genotype simulation, I would simulate polygenic risk scores based on each simulated reference population (European, admixed African, and Latin American of various sizes). I would assign 200, 500, or 1000 causal alleles with randomized heritabilities of either 0.33, 0.50, or 0.67 to simulate phenotypes of different polygenicity as described in Martin et al.[7] I would rank simulated individuals by liability (as previously defined[7]) and consider the top 5% as disease cases, then consider a random 5% of the remaining population as controls. Another random 5% of individuals, none of whom would be included in the case or control sets, would be reserved for validation. I would then perform a GWAS comparing the case and control populations and repeat the MAF filtering, clumping, and polygenic risk score construction as outlined in Aim 1. Polygenic risk scores would be constructed for the remaining matched-ancestry validation dataset. In the case of European simulations, polygenic risk scores would also be constructed for an equal number of simulated admixed African and Latin American individuals.

Evaluation and Expected Results

I would predict disease status in the validation datasets from simulated polygenic risk scores using the best-scoring model as determined from Aim 1 (Table 1) and evaluate them as in Aim 1. I expect that the polygenic risk scores constructed based on the simulated European GWAS will outperform those based on the small (25% of n) and medium (50% of n) matched-ancestry GWASs. I expect that the matched-ancestry scores will reach the performance of the European scores before the simulated study size reaches n (i.e. at 75% or 90% of n). However, because of the heterogeneity of the simulated Latin American and admixed African populations, the admixed polygenic risk scores will likely perform worse on admixed validation sets than European polygenic risk scores perform on European validation sets even when the GWAS sizes are equal. I expect that to reach equal predictive value in matched-ancestry populations, the sizes of the admixed population GWASs must surpass the size of current European libraries. I predict that requisite GWAS size will be proportional to the admixture of the population, so the Latin American simulation will need to be larger than that of the admixed African simulation to achieve the European prediction standard. I would compare the estimated GWAS sizes to the data collection rates for each population as determined in the literature search in order to stress the need for prioritization of underrepresented admixed populations in polygenic risk studies.

References

[1]Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., … Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics, 50(9), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

[2]Spanakis, E. K., & Golden, S. H. (2013). Race/Ethnic Difference in Diabetes and Diabetic Complications. Current Diabetes Report, 13(6), 814-823. https://doi.org/10.1007/s11892-013-0421-9

[3]Bentley, A. R., Callier, S., & Rotimi, C. N. (2017). Diversity and inclusion in genomic research: why the uneven progress? Journal of Community Genetics, 8(4), 255–266. https://doi.org/10.1007/s12687-017-0316-6

[4]Grinde, K. E., Qi, Q., Thornton, T. A., Liu, S., Shadyab, A. H., Chan, K. H. K., … Sofer, T. (2019). Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. Genetic Epidemiology, 43(1), 50–62. https://doi.org/10.1002/gepi.22166

[5]Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., … Parra, E. J. (2007). A genomewide admixture mapping panel for hispanic/latino populations. American Journal of Human Genetics, 80(6), 1171–1178. https://doi.org/10.1086/518564

[6]O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., … Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genetics, 10(4). https://doi.org/10.1371/journal.pgen.1004234

[7]Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., … Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. American Journal of Human Genetics. https://doi.org/10.1016/j.ajhg.2017.03.004

[8]Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., … Barrett, J. (2012). Genetics and population analysis Fast and accurate inference of local ancestry in Latino populations. Bioinformatics, 28(10), 1359–1367. https://doi.org/10.1093/bioinformatics/bts144

[9]Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. The American Journal of Human Genetics, 93, 278-288. https://doi.org/10.1016/j.ajhg.2013.06.020

[10]Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-

wide association studies. Nature Genetics, 38(8), 904-909. https://doi.org/10.1038/ng1847

[11] 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 490. https://doi.org/10.1038/nature11632

[12] Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Computational Biology, 12(5), 1–22. https://doi.org/10.1371/journal.pcbi.1004842

[13] Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., … Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-11112-0

[14] Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., … Bustamante, C. D. (2011). Demographic history and rare allele sharing among human populations. PNAS, 108(29), 11983-11988. https://doi.org/10.1073/pnas.1019276108