Supplementary Information for:

The phenotypic and functional consequences of genomic divergence between admixed Latin American populations: Antioquia and Chocó, Colombia

Aroon T. Chande, Lavanya Rishishwar, Shashwat D. Nagar, Andrew B. Conley, Jessica Rowell, Augusto E. Valderrama-Aguirre, Miguel A. Medina-Rivas, I. King Jordan

Contents

Supplementary Table 1. **Human populations analyzed in this study.** 2

Supplementary Table 2. **Bioinformatics methods used in this study.** 3

Supplementary Figure 1. **Distribution of polarized F_{ST} values between Antioquia and Chocó.** 4

Supplementary Figure 2. **Distribution of PRS differences between Antioquia and Chocó.** 5

Supplementary Figure 3. **Effect of GWAS discovery population ancestry on PRS.** 6

Supplementary Figure 4. **Correlations and SNP overlap among PRS.** 7

Supplementary Methods 8

Supplementary References 9

Supplementary Table 1. **Human populations analyzed in this study.**

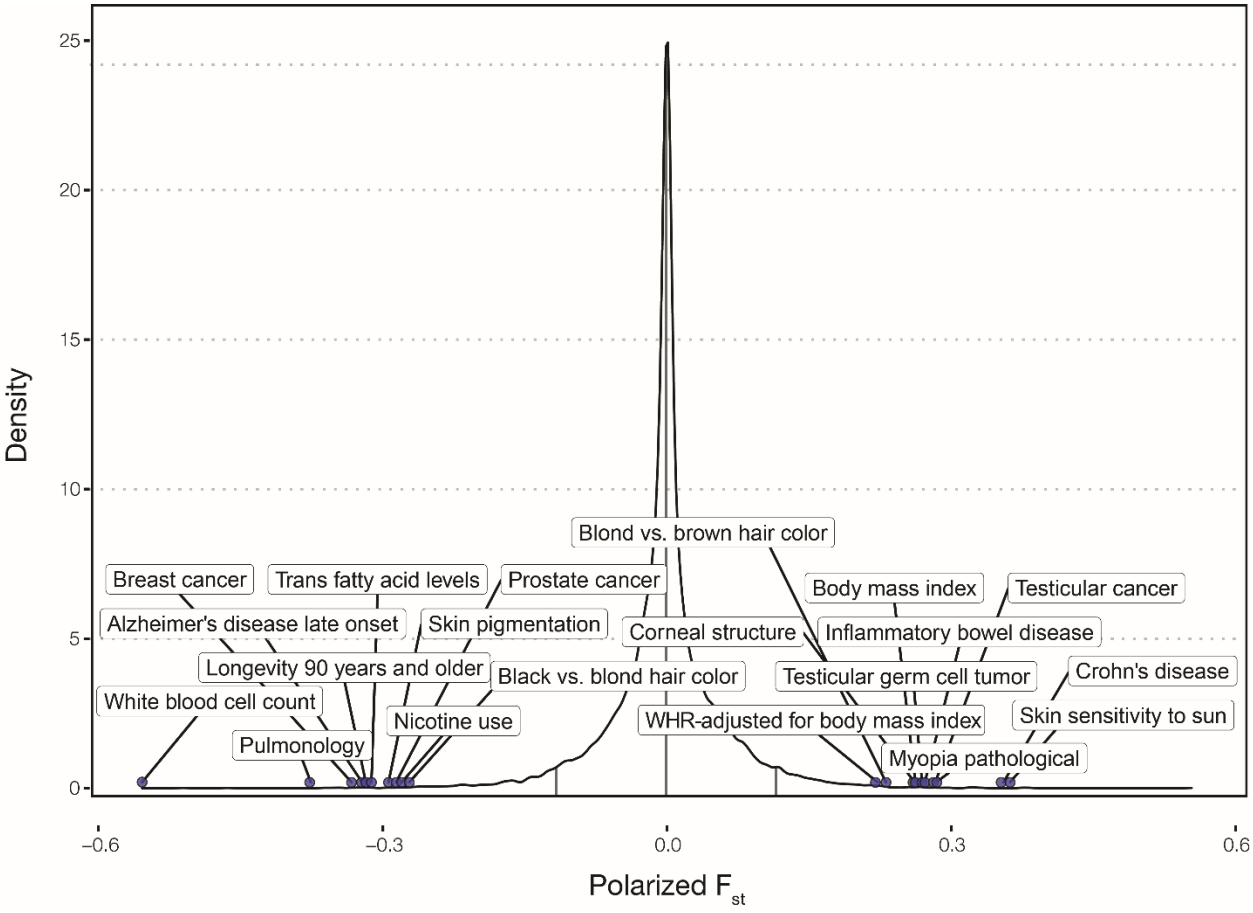
Dataset	Year	Population Name	Short	*n*
Colombian Populations				
Medina et al	2016	Chocoano in Quibdó, Colombia	CHG	94
1KGP	2015	Colombian in Medellin, Colombia	CLM	94
	Continental reference populations			
1KGP	2015	Yoruba in Ibadan, Nigeria	YRI	108
1KGP	2015	Iberian populations in Spain	IBS	107
1KGP	2015	Peruvian in Lima, Peru	PEL	85
Reich et al	2012	Embera in Colombia	Embera	5
Reich et al	2012	Quechua in Peru	Quechua	40
Reich et al	2012	Zapotec in Mexico	Zapotec	43

1KGP = 1000 Genomes Project Phase III

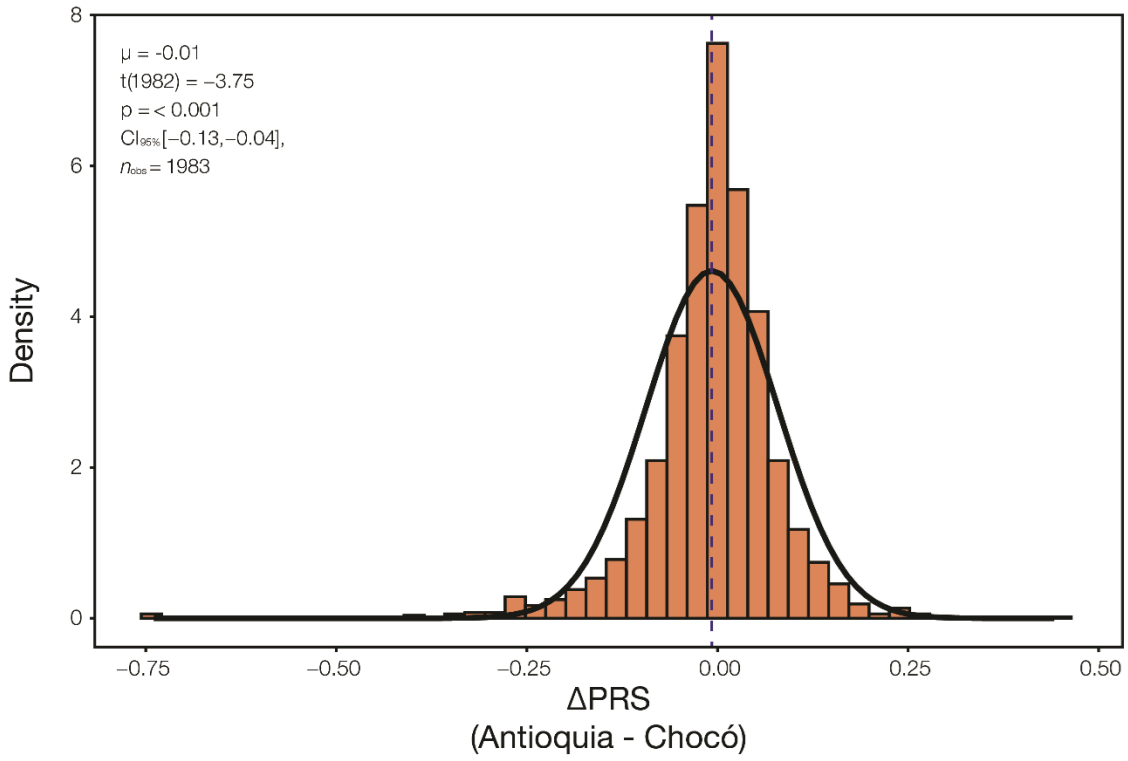
Supplementary Table 2. **Bioinformatics methods used in this study.**

Software	**Use**	**Access**
PLINK version 1.9	SNP quality control, merging, and pruning	https://www.cog-genomics.org/plink2/
ShapeIT version 2.r837	SNP phasing	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
ADMIXTURE version 1.3.0	Continental genetic ancestry inference	http://software.genetics.ucla.edu/admixture/
IMPUTE2 version 2.3.2	Genome variant imputation	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Database	**Use**	**Access**
NHGRI-EBI GWAS Catalog	SNP trait associations and polygenic trait scores	https://www.ebi.ac.uk/gwas/ (accessed December 2018)

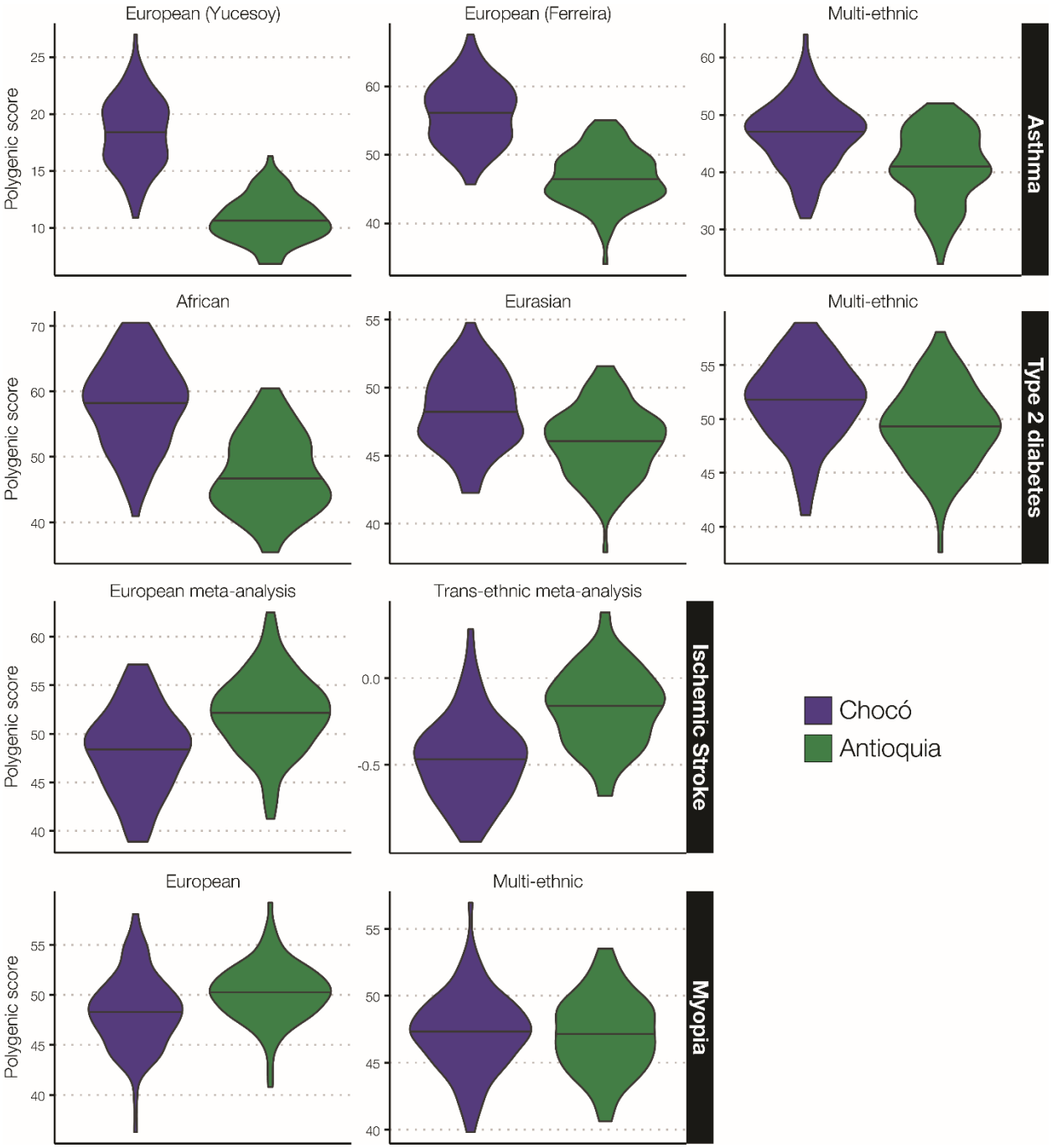
Supplementary Figure 1. **Distribution of polarized F_{ST} values between Antioquia and Chocó.** F_{ST} values were polarized to facilitate comparison between Antioquia (positive) and Chocó (negative). Highly differentiated alleles selected for Figure 2 are annotated (blue points). SNP F_{ST} and P-values are in Supplementary Table 3.



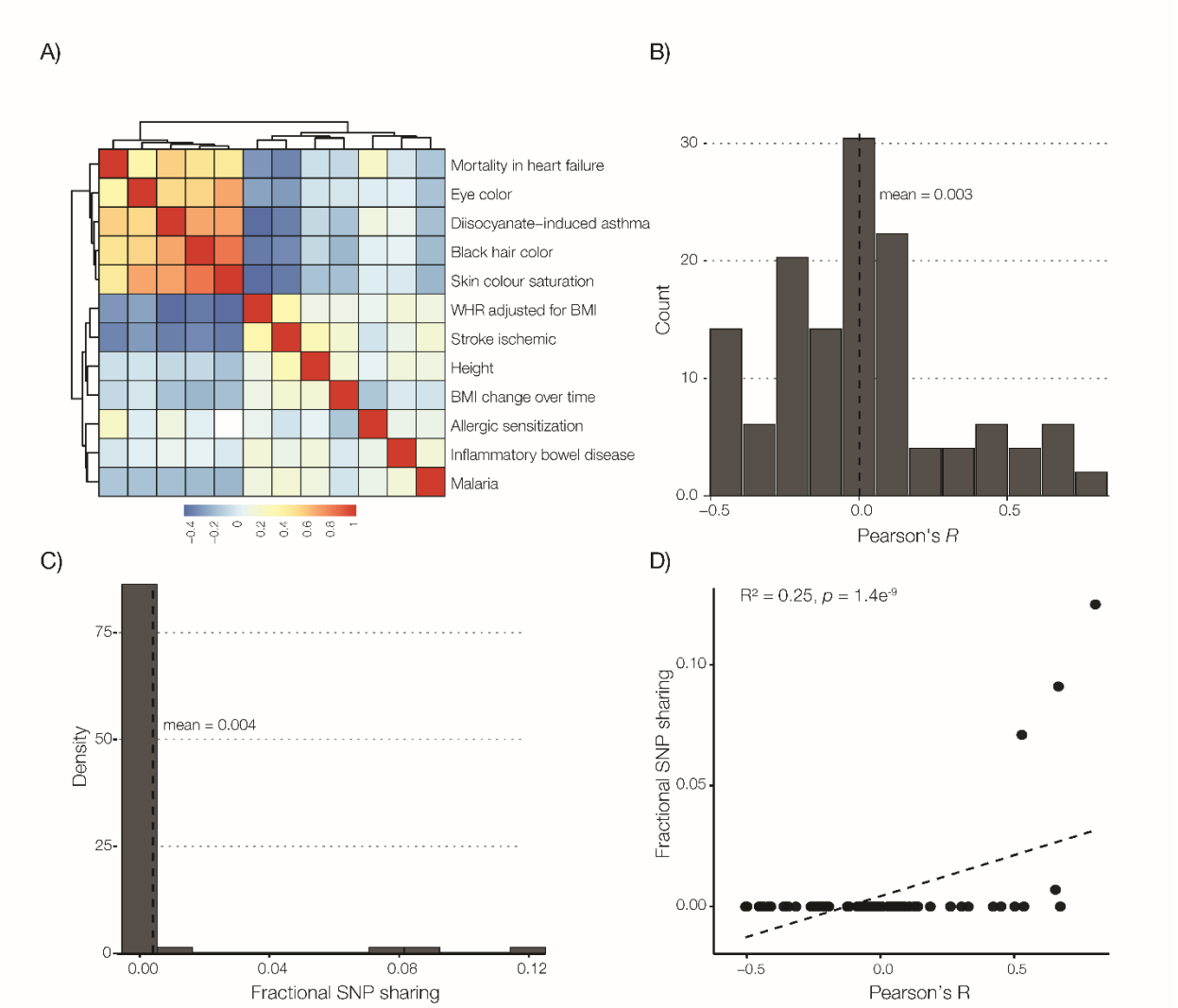
Supplementary Figure 2. **Distribution of PRS differences between Antioquia and Chocó.** A histogram of the observed PRS differences is shown along with the corresponding smoothed density distribution. The mean difference value (μ) is indicated with a dashed line.



Supplementary Figure 3. **Effect of GWAS discovery population ancestry on PRS.** Four selected traits from Figure 2 and Figure 3 were further analyzed with respect to GWAS discovery population ancestry, with two traits chosen for both Antioquia (ischemic stroke, myopia) and Chocó (asthma, type 2 diabetes). Trends in populations PRS distributions are similar regardless of GWAS ancestry.



Supplementary Figure 4. **Correlations and SNP overlap among PRS.** (A) All-by-all pairwise correlations between PRS from Figure 3B, hierarchically clustered by their Pearson correlations. (B) Distribution of Pearson correlation values, excluding self-correlation (n=132, mean=0.003). (C) Distribution of the fraction of SNPs shared between PRS (common SNPs in intersection/union of SNPs between scores). (D) Relationship between the fraction of SNPs shared between traits (y-axis) and the PRS correlation between traits (x-axis).



Supplementary Methods

PRS calculation and comparison among divergent populations

Our approach to PRS calculation and comparison between populations is characterized by three important choices: (1) the use of only significantly associated SNPs (P<10^{-5}) for PRS calculation, (2) the calculation of PRS that are unweighted by SNP effect sizes, and (3) the calculation of PRS without the use of linkage disequilibrium (LD) pruning or clumping. This conservative approach to PRS calculation and comparison is supported by (1) the lack of apparent systematic bias in PRS differences between populations (Supplementary Figure 2), (2) the consistency of predicted risk differences between populations with previously reported trait and disease prevalence differences (Supplementary Table 4), and (3) the consistency of PRS differences found when different ancestry SNP-association cohorts were used (Supplementary Figure 3). Here, we provide additional justification for our approach to PRS calculation.

1. Top-SNP approach: There are many different ways to compute PRS and the extremes are the "top-SNP" approach, where only highly significantly associated SNPs are used for PRS calculation, and the "all-SNP" approach, where as many SNPs as possible are used to calculate PRS. The top-SNP approach is more conservative as it relies only on robust associations, whereas the all-SNP approach derives additional resolution by capturing more of the genome-wide trait heritability. Most importantly for our own study, the all-SNP approach will also capture most or all of the population structure between divergent populations, whereas the top-SNP approach is not expected to do so. In other words, when the all-SNP approach is used to compare PRS for divergent populations, such as the kind studied here, the resulting PRS are essentially guaranteed to show large differences between populations. This has been convincingly shown in a recent study, where increasing numbers of SNPs used for PRS yielded increasingly greater between population differences, particularly for divergent populations (Duncan, et al. 2019). Furthermore, recent work by Khera et al. suggests that the all-SNP approach only provides a marginal increase in prediction accuracy compared to the top SNP approach (Khera, et al. 2018). For example, a top-SNP PRS for coronary artery disease using 74 variants showed 79% accuracy compared to 81% accuracy when 6.6 million SNPs were used for PRS calculation. Similar marginal increases in accuracy between the top-SNP and all-SNP approaches were observed for the four other traits analyzed in the same study. These findings support both the utility of the top-SNP approach for cross-population PRS comparisons and its resolution for capturing the majority of variance in trait risk.

2. Unweighted *PRS*: PRS can be calculated as unweighted scores by simply summing the numbers of trait-associated effect alleles genome-wide, or they can be calculated as weighted scores, whereby each effect allele is weighted by its effect size (odds ratio or beta value). As with our previous studies (Chande, et al. 2017; Chande, et al. 2018), we chose to use unweighted PRS here to facilitate the inclusion of SNP trait-associations from multiple studies. Effect sizes from different studies cannot be readily combined owing to differences in study cohorts, including cohort size, allele frequencies, and population structure. Furthermore, since effect sizes represent SNP heritability estimates, which are dependent on the particular cohort that is being studied, it does not make sense to attempt to normalize effect sizes across studies. Meta-analyses are able to perform this kind of normalization, as they have access to individual level phenotype data, but we do not have access to data of that kind data here. Finally, the use of multiple studies allowed us to ascertain as many trait-associated SNPs as possible, which is particularly important given our choice of the conservative top-SNP approach that is limited to significantly associated SNPs.

3. No linkage disequilibrium (LD) pruning: PRS are often calculated using linkage disequilibrium pruning or clumping, whereby only a single SNP from any given LD block is used for PRS calculation. We opted not to use LD pruning or clumping here owing the facts that (1) we use a top-SNP approach for PRS calculation and (2) the two populations under study have a highly divergent LD structure. The top-SNP approach means that we are using a relatively small number of SNPs per population and the divergent LD structure means that different subsets of this small number of SNPs would likely be removed from each population if LD pruning were used. For example, LD pruning here may remove SNPs that are population-private (that is, a SNP that appears in Chocó but not Antioquia) or whose LD patterns are discordant between the two populations (i.e., the SNP is in moderate LD in Antioquia but low LD in Chocó). This would severely mitigate our ability to compare PRS between populations. An alternative approach, as discussed previously, would be to use a very large number of variants together with LD pruning for PRS computation, i.e. essentially covering most of all of the LD blocks in the genome for the PRS, as has been done in a number of studies. However, when tens- or hundreds-of-thousands, or even millions, of SNPs are used for PRS calculation between divergent populations, then the PRS is essentially modeling the population structure between the populations. In this case, all traits will be highly divergent if you are comparing divergent populations. Our approach to PRS calculation without LD pruning provides for both additional resolution, in terms of the numbers of SNPs available for analysis, and more direct comparisons between divergent populations. Several studies, including our own work, have shown that PRS calculated with and without LD pruning do not show big differences (De La Vega and Bustamante 2018; Elliott, et al. 2020).

Supplementary References

Chande AT, et al. 2017. Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Choco and Antioquia. Sci Rep 7: 17127. doi: 10.1038/s41598-017-17380-4

Chande AT, et al. 2018. GlobAl Distribution of GEnetic Traits (GADGET) web server: polygenic trait scores worldwide. Nucleic Acids Res 46: W121-W126. doi: 10.1093/nar/gky415

De La Vega FM, Bustamante CD 2018. Polygenic risk scores: a biased prediction? Genome Med 10: 100. doi: 10.1186/s13073-018-0610-x

Duncan L, et al. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun 10: 3328. doi: 10.1038/s41467-019-11112-0

Elliott J, et al. 2020. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA 323: 636-645. doi: 10.1001/jama.2019.22241

Khera AV, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50: 1219-1224. doi: 10.1038/s41588-018-0183-z