color

Version 1.0 – Updated 02.12.21

Color Genome-wide Polygenic Score

Version 1.0 – Updated 02.12.21

Executive Summary

A genome-wide polygenic score (GPS) is a metric of inherited susceptibility that can be used to identify individuals with an increased risk of common, complex diseases by incorporating information from numerous sites in the genome.

Recently, Color has developed the Coronary Artery Disease Genetic Score, which uses a cost-efficient pipeline to assess an individual's GPS for developing coronary artery disease (CAD).

Color's clinical GPS pipeline uses low coverage whole genome sequencing (lcWGS) data accompanied by the imputation of common genetic variants.

In this study, we demonstrate that data generated with lcWGS followed by imputation is an alternative approach to genotyping arrays and clinically validate the end-to-end ability of Color's platform to accurately and reliably calculate a GPS value for CAD in clinical samples from diverse populations.

Introduction

Coronary artery disease (CAD) is a cardiovascular disorder caused by the build-up of plaque in the arteries, leading to a narrowing or blockage of the artery itself. Despite significant improvements in treatment and prevention, CAD continues to be the leading cause of morbidity and mortality worldwide.[1,2] Thus, identifying individuals at an increased risk for CAD who should undergo regular screening is a crucial step towards prevention and lowering the prevalence of this disease.

The initial set of risk factors associated with CAD were established in the Framingham Heart Study which was conducted in the early 1960s.[3,4] Since then, there have been many additional risk factors linked to CAD, including some that are under an individual's control, such as diet and exercise, and others that are innate, such as age and sex.[5]

Over the past decade, technological advancements in the field of genetics have led to a better understanding of the heritability and other genetic components of complex diseases. Studies analyzing hundreds of thousands of people and millions of genetic variants have indicated that the genetic components of CAD can sometimes even outperform traditional risk factors in predicting the development of CAD, demonstrating the important role that genetics can have in clinical practices.[6,7] For example, recent studies estimate the heritability of CAD to range between 40 and 50%.[8] Additionally, it has been estimated that about 8% of the population has approximately triple the average risk for CAD due to polygenic variation alone.[8–11] When the genetic predisposition of an individual is combined with their non-genetic risk factors, it can reveal an even higher risk of developing a complex disease.[12]

In order to identify those who could benefit from early detection and treatment, Color has developed the Coronary Artery Disease Genetic Score to measure an individual's genetic susceptibility to develop CAD.[9] The workflow to calculate this score includes the generation of an individual's genome-wide polygenic score (GPS), also known as a genetic risk score (GRS) or polygenic risk score (PRS). A GPS examines the cumulative effect of a large number of genetic variants associated with a certain disease by aggregating the effect of all of those variants into a single score. Historically, SNP array genotyping has been the technology used to generate an individual's GPS.[13,14] Here, we present an alternative approach using low coverage whole genome sequencing (lcWGS), a method that samples locations across the entire genome at low depth (~0.4x).

GPS is calculated by combining genetic signals from hundreds to millions of locations throughout the genome that are associated with a complex disease. The genetic locations are first identified as part of a genome-wide association study (GWAS), a study that examines variation across the genome in an attempt to identify genetic risk factors for diseases that are common in a population. GWAS studies include thousands of individuals (both with and without the disease in question) and millions of genetic variants, yielding the individual effect of each variant for a given complex disease, such as coronary artery disease.[15,16]

Color's clinical GPS workflow not only leverages millions of genetic variants that have already been established, tested, and validated in a previous GWAS, but does so by using lcWGS, which is a cost-effective sequencing approach. This process is then followed by state of the art imputation methods that statistically infer missing information that occurs as a result of sampling. Furthermore, to address the lack of diversity that has plagued GWAS research in the past, Color's clinical GPS pipeline ensures the GPS are inclusive of populations with different ethnic backgrounds and thus can be utilized to help individuals and their healthcare providers proactively decide what health and lifestyle plan is most suitable for the individual, regardless of their ethnicity.

In addition to being cost effective, lcWGS does not suffer from the inherent biases in SNP genotyping arrays. SNP arrays use only a selection of pre-ascertained SNP sites, which are often biased towards one population leading to a distortion of the genome-wide distribution of allele frequencies.[17,18] For example, it is possible to observe genetic variations in individuals that did not previously exist in the reference population used to design the SNP array. This means the variants will not be reliably captured in the SNP array, resulting in a loss of useful genetic information for the individuals from the non-reference population. These biases can further be exacerbated during imputation and can have an impact on downstream analysis. Moreover, the effects of SNP selection bias are not identical across arrays. For instance, GPS trained and validated on one genotyping array may not be as predictive on another genotyping array.[9,19] lcWGS followed by imputation overcomes this issue by randomly sampling across the entire genome so that variants are captured independently. Additionally, new variants that are discovered can be easily added to the imputation procedure. Since Color's clinical GPS pipeline takes advantage of lcWGS, it is less vulnerable to the aforementioned biases.

The Coronary Artery Disease Genetic Score is generated through the Color clinical GPS workflow using the technique of imputation off of lcWGS data. In this study, we describe the methodology of the workflow and validate the Coronary Artery Disease Genetic Score for use as a clinical genetic test.

Dataset

To validate the performance of the CAD GPS on lcWGS, we compared GPS scores from lcWGS and from SNP genotyping using the Global Diversity Array (GDA) from Illumina for a selected set of 113 individuals as a benchmark dataset. Individuals were selected to encompass a range of GPS scores, in order to assess the validity of the end-to-end pipeline across the spectrum. 22 individuals from the validation cohort were further selected to assess the reproducibility (n=10) and repeatability (n=12) of the study. This dataset included both saliva and peripheral blood specimen types to ensure similar accuracy for imputation across specimen types. To ensure diversity, this dataset included equal representation of both sexes as well as samples from diverse continental and subcontinental ancestral populations based on genetic ancestry results generated by ngsadmix.[20]

Methods

Generation of sequencing data

Genomic DNA was extracted from saliva or peripheral blood using standard laboratory methods for the Color assay. Each batch contained at least one no-template control (NTC) sample and two cell line positive controls. Next generation sequencing (NGS) libraries compatible with the Illumina platform were generated and sequenced at low coverage on an Illumina NovaSeq. The data was initially processed through the Color bioinformatics pipeline and quality control procedures (Figure 1, step 1). Sequencing base call files were converted to FASTQ using bcl2fastq2 (Illumina, San Diego, CA) and reads were aligned to the human genome reference GRCh37 with BWA-MEM.[21] Duplicates and low quality reads were then removed. Genotype likelihoods were calculated using bcftools v1.8[22] mpileup algorithm at each of the biallelic SNP loci in the imputation SNP loci that were covered by one or more sequencing reads. As part of the quality control for lcWGS data, we used a genome-wide sequencing coverage criteria of ≥0.4X as well as ≥0.2X sequencing depth across all autosomes as calculated by GATK (CollectWgsMetrics tool) (Figure 1, step 2).[23]

Genotype Imputation

BAMs generated from the lcWGS sequencing of the benchmark dataset were used to impute genotypes. Genome-wide genotypes at approximately 22 million sites were imputed using GLIMPSE v1.0.0, a software tool for large-scale imputation of low coverage sequences.[24] To capture the global haplotype diversity, the 1000 Genomes Project (1KGP) was used as the imputation reference panel (Figure 1, step 3).[25] Differences in genotype imputation accuracy vary in different populations due to haplotype block conformation and available reference data. To ensure that there were no idiosyncratic imputation errors and/or biases towards a specific population, the benchmark dataset was grouped into five populations based on their genetic ancestry where an individual belonged to a non-Admix group if more than 70% of their genetic make-up was from that group. This resulted in five continental groupings of individuals, namely, African (AFR), East Asian (EAS), European (EUR), South Asian (SAS) as well as Admixed individuals.

Imputation accuracy of the lcWGS data was then separately measured for each continental group mentioned above by calculating coefficient of determination (r^2) by comparing imputed results for the benchmark dataset against known, externally confirmed non-imputed genotypes of the same individuals (~1.3M sites).

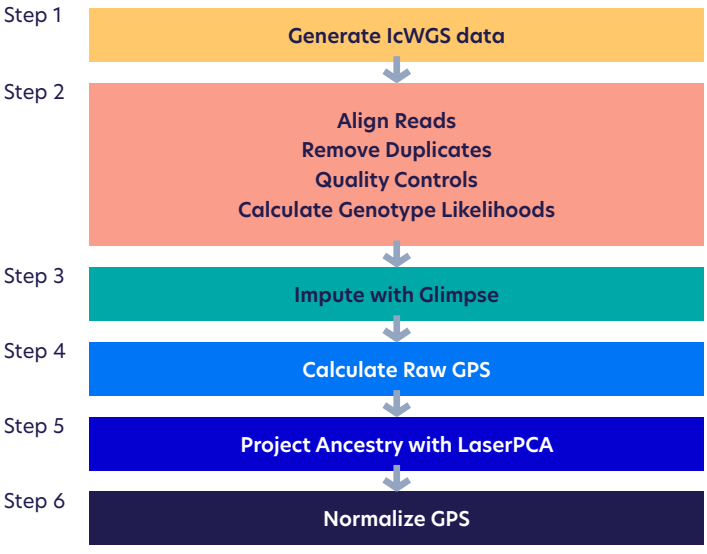

Figure 1: Color's clinical GPS pipeline design. lcWGS data is collected and processed through distinct steps to produce a GPS.

Color's clinical GPS workflow

Calculating raw GPS

To identify individuals at a higher risk of developing CAD, Color's Clinical GPS pipeline uses variants that, in a prior study, were found to be associated with CAD risk.[8] Raw GPS for CAD were calculated by summing risk alleles (OR and BETAs), which are weighted by effect sizes derived from prior GWAS results (Figure 1, step 4). Thus, GPS for an individual i was calculated using the following equation:

$$GPS(i) = \sum_{j=1}^{n} dosage(i,j)\ \beta(j)$$

Where $0 \leq dosage(i,j) \leq 2$ is the effect allele dosage for each variant j associated with CAD and $\beta(j)$ is the marginal effect size of variant j. The set of variants j were previously selected based on their demonstrated ability to accurately predict and stratify disease risk.[8]

Ancestry-based GPS normalization

The observed range of raw GPS will vary among individuals depending on their genetic ancestry (Figure 2A).[26] To address this, a normalization procedure for raw GPS was established. First, the LASER program was used to obtain ancestry principal components (PCs) by projecting individuals' genetic data on a set of built-in ancestry reference panels constructed using 4259 ethnically diverse samples (Figure 1, step 5).[27]

Once the ancestral PCs were obtained for each individual, an ancestry-based z-score normalization was subsequently applied to the raw scores). To construct a normalizer, we used a large diverse cohort of 25,016 non-related individuals from the Color database to generate a PCA-based linear model for the disease of interest that explains the contribution of each PC to variation in raw GPS. Normalized GPS was then calculated by taking the standardized residual of the score after corrections for the first 10 PCs. Lastly, we ensured the distribution of corrected GPS has a mean of ~0 and standard deviation of ~1 (Figure 1, Step 6). Thus, after normalization, the adjusted GPS (independent of genetic ancestry) follows a normal distribution (Figure 2B).

Validation study

The performance of Color's Clinical GPS pipeline was assessed by 1) comparing the raw GPSs calculated for the individuals in the benchmark dataset using both lcWGS and the gold standard method of SNP array genotyping and 2) confirming normalized GPS followed a normal distribution using a large, separately curated cohort of diverse individuals.

SNP array genotyping data was generated using the Global Diversity Array (GDA) from Illumina. This array includes approximately 1.9 million variants and was developed to perform well across 26 diverse populations.

To determine raw GPS accuracy, the coefficient of determination (r^2) was calculated between the two raw GPS for CAD. It is important to note that calculated GPS are not expected to be an exact match due to variability in sequencing coverage of disease-specific SNP loci. In addition, to measure the

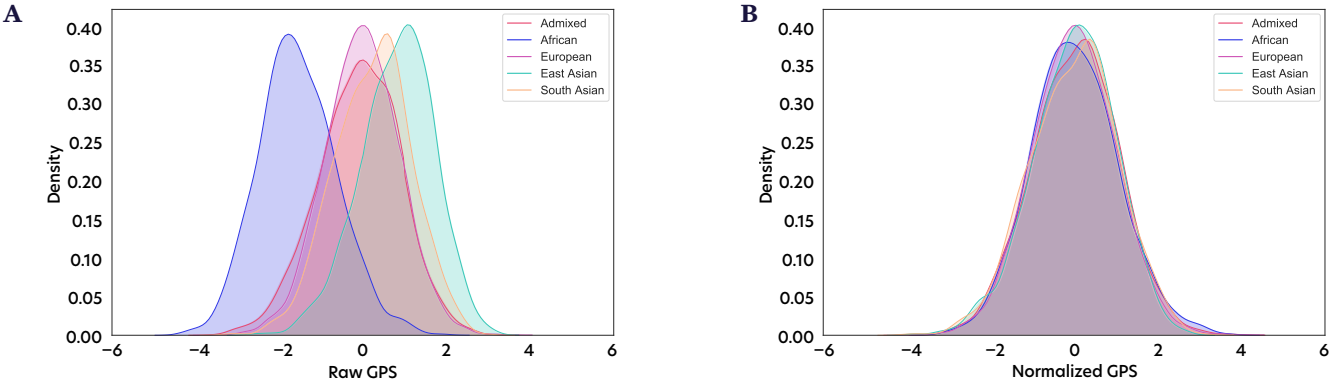

Figure 2: A) Distribution of raw GPS scores of the normalizer cohort stratified by continental group **B)** Distribution of normalized GPS scores of the normalizer cohort stratified by continental group

performance of normalized GPS, summary statistics per continental population were generated for a large, separately curated cohort to show that the normalized GPS follows a normal distribution regardless of the cohort's ancestral populations.

Results and Discussion

In this study, we demonstrate that Color's GPS pipeline can be used as a clinical assessment of genetic susceptibility for CAD and is highly concordant with other common methods of calculating GPS, including SNP arrays. The validation studies showed that lcWGS can be used for genome-wide imputation and GPS calculation as accurately as SNP genotype arrays and that GPS scores can be normalized for use across diverse populations.

lcWGS sequencing and genotype imputation

The benchmark dataset used in this validation study went through Color's lcWGS pipeline, passed quality controls, and underwent imputation across 22 million sites in the genome. Using the array data as a truth set, imputation accuracy was examined separately per continental population using squared Pearson's correlation coefficient, and all samples across all populations and specimen types had an imputation accuracy of $r^2 \geq 0.95$ (Figure 3A). Given the overall high concordance of the imputation accuracy, we conclude that lcWGS (with at least 0.4x coverage) followed by imputation performs similarly to SNP array genotyping and can be considered as a valid alternative approach.

Furthermore, we observed no significant difference in imputation accuracy based on specimen type between saliva (n=70, mean $r^2 = 0.970$) vs blood (n=7, mean $r^2 = 0.971$). However, we were only able to ascertain this finding in the European population due to small sample size and potential batch effects (Figure 3B).

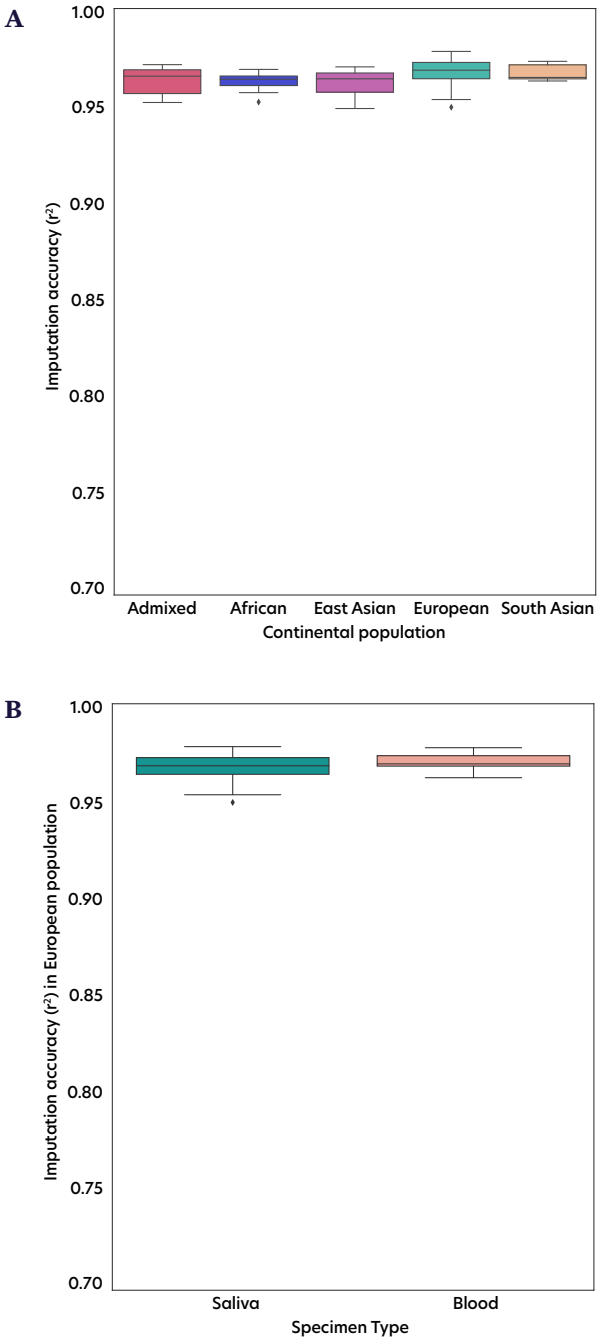

Figure 3: A) Imputation accuracy for benchmark dataset per continental population. SNP array genotype calls were used as the gold standard. **B)** Imputation accuracy for blood and saliva specimen types.

GPS calculation

Similar to the imputation assessment, we assess the validity of raw GPS scores by calculating square of Pearson's correlation coefficient between the raw GPS calculated from the SNP array and the raw GPS calculated from Color's clinical GPS pipeline for our benchmark dataset ($r^2 \geq 0.992$; Figure 4A). Our results demonstrate that there are no significant differences between the raw GPS observed from Color's GPS pipeline than that of SNP arrays.

Moreover, we examined the reproducibility and repeatability of our end-to-end analysis by looking at 10 and 12 samples that were repeated intra- and inter-run through the end-to-end process, respectively (Figure 4B and C). Results showed that not only was there a high correlation among repeated samples after imputation, but also high correlations of raw GPS in both experiments ($r^2 \geq 0.98$ and $r^2 \geq 0.99$, intra-run and inter-run, respectively). It is important to note that some variations in reproducibility and repeatability experiments are expected due to differences in the variants that were captured in lcWGS itself as well as variations caused by imputation.

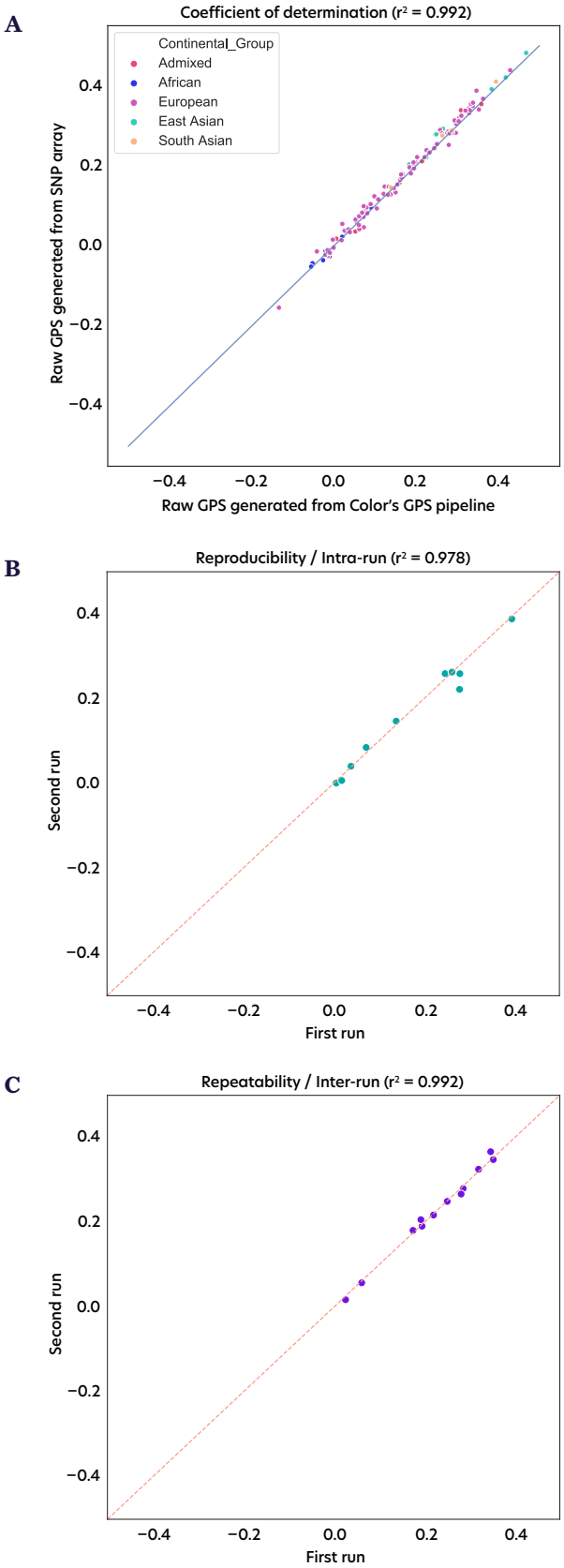

Figure 4: A) Correlation of raw GPS scores generated using lcWGS or SNP array, stratified by continental group. **B)** Reproducibility and **C)** repeatability of raw GPS scores across multiple runs of the same samples with lcWGS.

GPS Normalization

Historically, many GPS scores were developed for use in European populations.[28,29] Thus, the distribution of the raw GPS for individuals from non-European populations might differ from the ones observed in Europeans (Figure 2A). This makes it difficult to correctly interpret the raw scores for individuals from diverse ethnic backgrounds since two similar raw scores demonstrate different levels of risks due to the underlying population distribution. To address this inequity, we developed a GPS normalization procedure based on genetic ancestry that allows for GPS calculation in non-European individuals (see Methods section). This GPS normalization determines how the distributions of raw GPS can be made comparable with respect to genetic ancestry as a confounding factor.

Summary statistics generated for each continental population from a large, separately curated cohort showed that the normalized GPS followed a normal distribution. The normalized GPS for each continental population had a mean of 0±0.1 and standard deviation of 1±0.1 (Table 1).

Table 1. CAD GPS normalization summary statistics

Population	Mean	Standard Deviation
South Asian	0.043	1.033
African	-0.068	1.017
Admixed	0.028	1.042
East Asian	0.004	0.98
European	-0.011	0.925

Limitations

In order to successfully use GPS in practice, it is crucial to understand the limitations of GPS and be vigilant in its interpretation. For instance, one of the shortcomings of GPS is that variability in a score can be heavily influenced by differences in population structure and allele frequency variability across distinct ethnic groups. For instance, people with African ancestry tend to have greater variation in their genomes and thus more complex inheritance patterns compared to individuals with European ancestry.[30-32] These variations across populations limit GPS transferability and reduce its clinical value in non-European populations. Thus, without effective normalization and incorporating genetic ancestry, GPS calculated in non-European populations cannot accurately estimate risk and could raise healthcare disparity concerns.[29,33]

Moreover, GPS only estimates the contribution of the common single-nucleotide variants (usually 1+% population frequency) in individuals, whereas other types of variation (such as rare pathogenic alleles which are typically excluded) can additionally impact genetic risk of certain individuals.[34,35]

Despite the aforementioned constraints, GPS nevertheless has the potential to be incorporated into routine medical practice to aid risk stratification and optimize clinical care for individuals.

Conclusion

Over the past few decades, there has been an emerging interest in understanding the underlying risk factors that are associated with complex diseases such as CAD. The promise of such endeavours in practice is the ability to prioritize preventative behaviors, such as lifestyle changes, prophylactic medication, and increased screening. In this study, we validated the ability of Color's clinical GPS pipeline to measure the genetic susceptibility of individuals for CAD by comparing its performance to that of GPS derived from SNP array genotyping. We established that Color's clinical GPS pipeline, which uses lcWGS followed by imputation, can be used as an effective alternative approach to genotyping arrays when assessing common genetic variants. Furthermore, we demonstrated that the GPS calculated for CAD using this method are inclusive of populations with different ethnic backgrounds (unlike those derived from SNP arrays) and thus can help enhance global population health by identifying individuals of diverse backgrounds at a higher risk of developing CAD.

References

1. Mensah GA, Wei GS, Sorlie PD, et al. Decline in Cardiovascular Mortality: Possible Causes and Implications. *Circ Res.* 2017;120(2):366-380.

2. Okrainec K, Banerjee DK, Eisenberg MJ. Coronary artery disease in the developing world. *Am Heart J.* 2004;148(1):7-15.

3. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis. 1967;20(7):511-524.

4. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet.* 2014;383(9921):999-1008.

5. Hajar R. Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart Views.* 2017;18(3):109-114.

6. Knowles JW, Ashley EA. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* 2018;15(3):e1002546.

7. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol.* 2018;72(16):1883-1893.

8. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018;50(9):1219-1224.

9. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med.* 2019;11(1):74.

10. Aragam KG, Dobbyn A, Judy R, et al. Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease. *J Am Coll Cardiol.* 2020;75(22):2769-2780.

11. Won H-H, Natarajan P, Dobbyn A, et al. Disproportionate Contributions of Select Genomic Compartments and Cell Types to Genetic Risk for Coronary Artery Disease. *PLoS Genet.* 2015;11(10):e1005622.

12. Sun L, Pennells L, Kaptoge S, et al. Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med.* 2021;18(1):e1003498.

13. Chen S-F, Dias R, Evans D, et al. Genotype imputation and variability in polygenic risk score estimation. *Genome Med.* 2020;12(1):100.

14. Wasik K, Berisa T, Pickrell JK, et al. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *Cold Spring Harbor Laboratory*. Published online May 8, 2019:632141. doi:10.1101/632141

15. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet.* 2017;101(1):5-22.

16. Riancho JA. Genome-wide association studies (GWAS) in complex diseases: advantages and limitations. *Reumatol Clin.* 2012;8(2):56-57.

17. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays.* 2013;35(9):780-786.

18. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 2005;15(11):1496-1502.

19. Johnson EO, Hancock DB, Levy JL, et al. Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet.* 2013;132(5):509-522.

20. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics.* 2013;195(3):693-702.

21. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bioGN].* Published online March 16, 2013. http://arxiv.org/abs/1303.3997

22. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079.

23. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-498.

24. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet.* 2021;53(1):120-126.

25. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.

26. Martin AR, Gignoux CR, Walters RK, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet.* 2017;100(4):635-649.

27. Wang C, Zhan X, Liang L, Abecasis GR, Lin X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet.* 2015;96(6):926-937.

28. Duncan L, Shen H, Gelaye B, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun.* 2019;10(1):3328.

29. Cavazos TB, Witte JS. Inclusion of Variants Discovered from Diverse Populations Improves Polygenic Risk Score Transferability. *Cold Spring Harbor Laboratory.* Published online October 5, 2020:2020.05.21.108845. doi:10.1101/2020.05.21.108845

30. Zhu X, Yan D, Cooper RS, et al. Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. *Genome Res.* 2003;13(2):173-181.

31. Huang L, Jakobsson M, Pemberton TJ, et al. Haplotype variation and genotype imputation in African populations. *Genet Epidemiol.* 2011;35(8):766-780.

32. Reich DE, Cargill M, Bolk S, et al. Linkage disequilibrium in the human genome. *Nature.* 2001;411(6834):199-204.

33. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51(4):584-591.

34. Fullerton JM, Nurnberger JI. Polygenic risk scores in psychiatry: Will they be useful for clinicians? *F1000Res.* 2019;8. doi:10.12688/f1000research.18491.1

35. Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet.* 2019;51(1):76-87.