Using low coverage sequencing for determining polygenic risk scores

Julian R. Homburger[1], Gilad Mishne[1], Cynthia L. Neben[1], Carmen Lai[1], Will Stedden[1], Anjali D. Zimmer[1], Amit V. Khera[2-4], Sekar Kathiresan[2-4], and Alicia Y. Zhou[1]
[1]Color Genomics, Burlingame, CA; [2]Center for Genomic Medicine and [3]Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA; [4]Harvard Medical School, Boston, MA

Introduction

Next generation sequencing (NGS) is an integral part of clinical care and management with applications such as targeted multi-gene panels for monogenic diseases. There is an emerging consensus that genome-wide polygenic risk scores (PRSs) also have validity and utility in stratifying disease risk[1], however, several barriers exist to implementing PRSs into clinical practice. PRSs have traditionally been performed on genotyping arrays, and therefore assessment of polygenic risk in addition to monogenic risk would require a different set of instruments and expertise. Furthermore, genotyping arrays can be limited by ascertainment bias[2], which reduces genotype imputation quality in diverse genetic populations. One possible solution to these barriers is to use low coverage whole genome sequencing (lcWGS) combined with imputation. lcWGS combined with imputation has been demonstrated to accurately assess common genetic variation[3,4] and can be performed using the same instruments as targeted multi-gene panels for monogenic disease. Here, we demonstrate the feasibility and technical accuracy of using lcWGS for genotype imputation and polygenic disease risk estimation.

Methods

Imputation pipeline

To develop and validate the imputation pipeline (Figure 1), seven samples from different 1000 Genomes Projects (TGP) populations (NA12878 - CEU, NA19420 - YRI, NA20510 - TSI, NA21144 - GIH, HG00663 - CHS, HG01485 - CLM, and HG02155 -CDX) and three Ashkenazi Jewish samples from the Genome in a Bottle Consortium (NA24143, NA24149, and NA24385) were sequenced at 30X using a NovaSeq 6000 instrument at the Color laboratory. Sequencing data was randomly downsampled to 2.0X, 1.5X, 1.0X, and 0.5X.

Calculating PRSs

To compare the lcWGS with imputation-based approach with a genotyping array for calculating previously published PRSs for coronary artery disease (CAD)[5] and breast cancer (BC)[5], DNA samples from 183 individuals who self-reported as "Caucasian" were selected: 61 individuals reported having a personal history of heart attack, 60 individuals were suspected to have high genome-wide PRS based on on previous unpublished work, and 62 individuals reported no personal history of cardiac events and were negative for monogenic pathogenic variants in a multi-gene NGS panel test (randomly selected as controls). The DNA samples were 1) genotyped on the Affymetrix Axiom UK Biobank array and 2) sequenced using a NovaSeq 6000 instrument at the Color laboratory. For genotyping, imputation was performed using BEAGLE 5.0[6] combined with a TGP reference panel. For sequencing, mean sequencing depth was 1.23X, with coverage ranging between 0.49X to 1.87X. Sequence reads were aligned against human genome reference GRCh37.p12 with the Burrows-Wheeler Aligner, and duplicate and low quality reads were removed following GATK best practices. The lcWGS imputation pipeline was used to impute all biallelic SNPs in TGP with a frequency greater than 1% in at least one continental population (approximately 19 million loci). Imputation r^2 was used to measure correlation of results.

To determine the robustness of the CAD and BC PRSs[5] using the lcWGS with imputation-based approach to sequencing coverage variation, the data was downsampled to 0.1X, 0.25X, 0.4X, 0.5X, 0.75X, and 1.0X. To note, the 61 individuals suspected to have high CAD PRS based on previous unpublished work were removed from this analysis since their inclusion may have artificially inflated the measured correlation.

To determine the association of the PRSs for CAD, BC, and atrial fibrillation (AF)[5] with event rate, DNA samples from more than 2,500 individuals who self-reported as "Caucasian," had European genetic ancestry calculated using principal components analysis, and did not have a pathogenic or likely pathogenic variant in a multi-gene NGS panel test were analyzed. For CAD and AF, PRSs were adjusted for age and gender. For BC, PRSs were calculated for only females and adjusted for age at menarche.

All individuals consented to having their de-identified information and sample used in anonymized studies. All information was reported by the individual.

References

1. Knowles JW, Ashley EA. *PLoS Med.* 2018.
2. Lachance J, Tishkoff SA. *Bioessays*. 2013.
3. Pasaniuc B, Rohland N, McLaren PJ, et al. *Nat Genet.* 2012.
4. Vilhjálmsson BJ, Yang J, Finucane HK, et al. *Am J Hum Genet.* 2015.
5. Khera AV, Chaffin M, Aragam KG, et al. *Nat Genet.* 2018.
6. Browning BL, Zhou Y, Browning SR. *Am J Hum Genet.* 2018.

Results

Figure 1. Imputation Pipeline

The imputation pipeline for lcWGS reads raw fastq sequence data and generates a vcf with imputed site information at 19 million biallelic SNP loci.

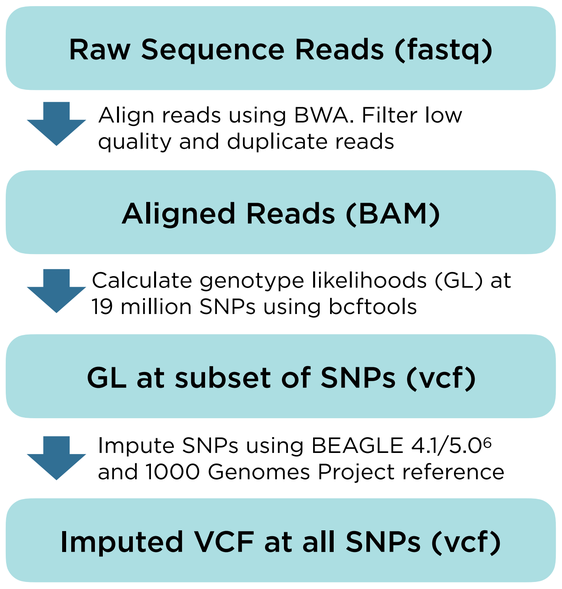


Figure 2. Imputation Accuracy on 1000 Genomes Samples

Low coverage imputation accuracy is above 0.9 r^2 for all samples at 0.5X (n = 10). For each combination of sample and coverage, imputation accuracy was assessed four times using different random seeds for downsampling. The brown dashed line is a smoothed trendline of the average imputation quality while the grey dashed line demonstrates previously reported imputation quality from a genotyping array (r^2 = 0.90)[3].

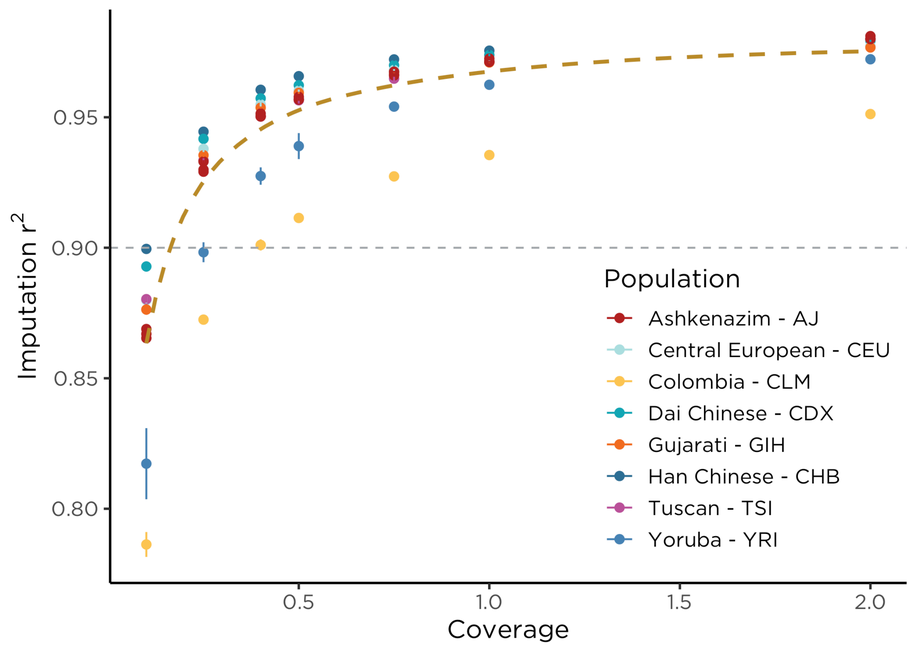


Figure 3. Correlation of coronary artery disease PRS across different assays

The cardiovascular PRSs calculated using lcWGS with imputation are highly correlated (r^2 = 0.978) with those calculated using genotyping array technology (n = 183).

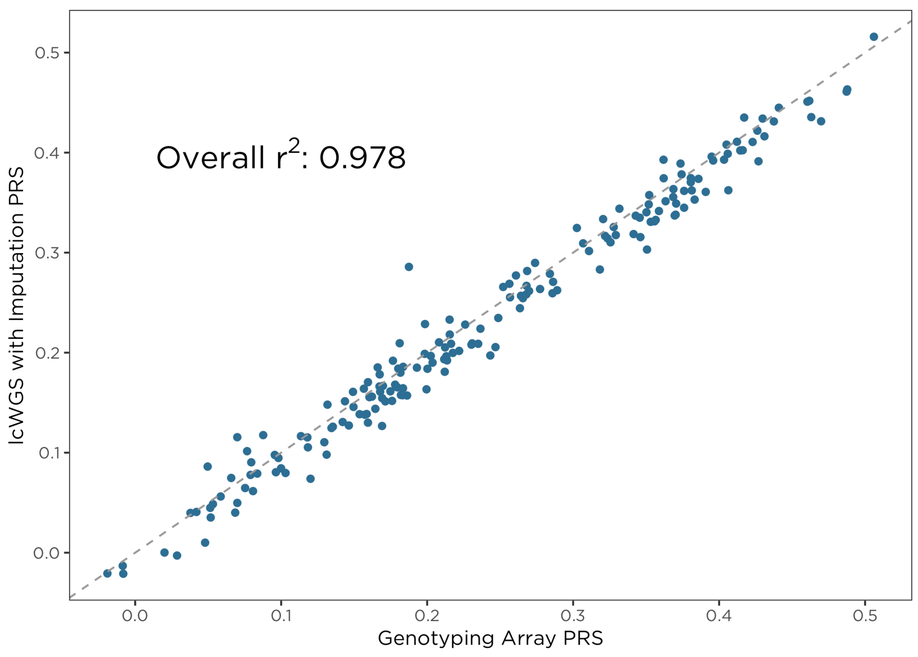


Figure 4. Concordance of PRSs calculated from different genome coverages

The PRSs for BC and CAD concordance are above 0.9 r^2 when samples are sequenced at or above 0.5X coverage (n = 183). There are only slight improvements in concordance above 0.5X and no batch level differences between different coverages above 0.5X.

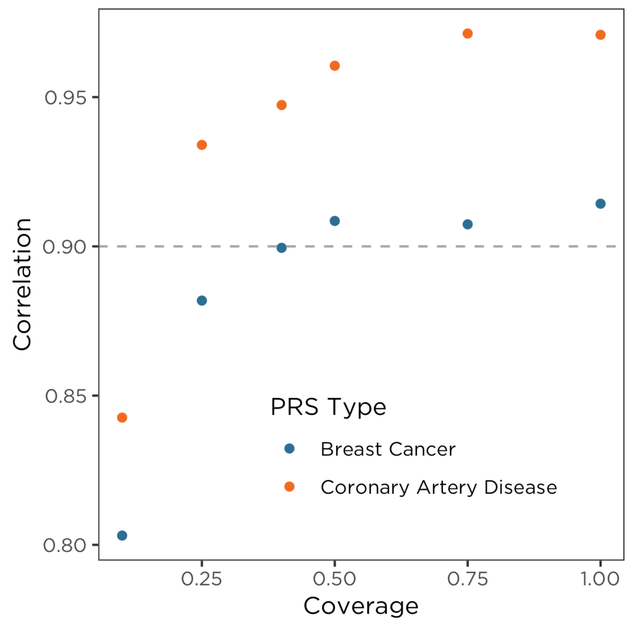
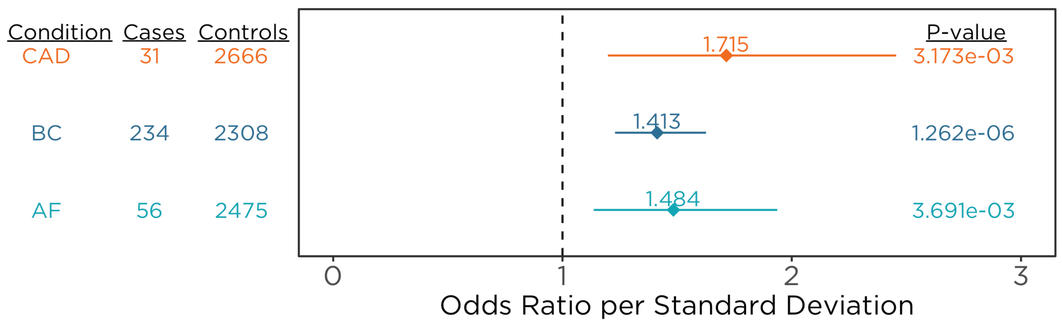


Figure 5. Association of lcWGS with imputation-calculated PRS with self-reported health history



lcWGS imputed PRSs for CAD[5] (n = 2,697), breast cancer (BC)[5] (n = 2,542), and atrial fibrillation (AF)[5] (n = 2,531) are associated with self-reported conditions. PRSs are adjusted for age and sex for AF and CAD. BC PRSs are calculated for only females and adjusted for age at menarche.

Conclusions

- We demonstrate that imputed variants and PRSs generated using lcWGS at a depth of 0.5X or more are highly concordant with those calculated using conventional methods across different populations.
- PRSs calculated from lcWGS with imputation are associated with self-reported health history in a population cohort of over 2,500 individuals.
- We demonstrate that PRSs can be accurately calculated from sequencing data, enabling the future combination of monogenic genetic testing and genome-wide polygenic scores in a single, cost-effective assay.