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, 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 addiction to monogenic risk would require a different set of instruments and expertise. Furthermore, genotyping arrays can be limited by ascertainment bias, 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 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 - GH, HG00663 - CHS, HG01146 - CLM, and HG02155 - CDX) and three Ashkenazi Jewish samples from the Genoma in a Bottle Consortium (NA24443, NA24459, and NA24485) 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) and breast cancer (BC), 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a personal history of heart attack, 60 individuals were suspected to have high genome-wide PRS based on previously unpublished work, and 62 individuals reported no personal history of cardiac events and were negative for monogenic pathogenic variants in a multi-gene NGS panel test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 combined with a TGP reference panel. For sequencing, mean sequencing depth was 12X, 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 allbiallelic SNPs in TGP with a frequency greater than 1% in at least one continental population (approximately 19 million loci). Imputation r² was used to measure correlation of results.

To determine the robustness of the CAD and BC PRSs calculated from lcWGS with imputation are associated with self-reported health history in a population cohort of more than 2,500 individuals. The PRSs for BC and CAD concordance are above 0.9 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.

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 r² 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 downsmapping. 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 (r² = 0.90).

Figure 3. Correlation of coronary artery disease PRS across different assays
The cardiovascular PRSs calculated using lcWGS with imputation are highly correlated (r² = 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 r² 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

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.

References