Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays

Jeremiah H. Li[1], Chase A. Mazur[1], Tomaz Berisa[1], and Joseph K. Pickrell[1]

[1] *Gencove, Inc. New York, NY 10016*

April 29, 2020

Abstract

Low-pass sequencing (sequencing a genome to an average depth less than 1x coverage) combined with genotype imputation has been proposed as an alternative to genotyping arrays for trait mapping and calculation of polygenic scores; however, the current literature is largely limited to simulation- and downsampling-based approaches. To empirically assess the relative performance of these technologies for different applications, we performed low-pass sequencing (targeting coverage levels of 0.5x and 1x) and array genotyping (using the Illumina Global Screening Array) on 120 DNA samples derived from African and European-ancestry individuals that are part of the 1000 Genomes Project. We then imputed both the sequencing data and the genotyping array data to the 1000 Genomes Phase 3 haplotype reference panel using a leave-one-out design. First, we evaluated overall imputation accuracy from these different assays as measured by genotype concordance; we introduce the concept of *effective coverage* that accounts for evenness of sequencing and show that this metric is a better predictor of imputation accuracy than nominal mapped coverage for low-pass sequencing data. Next, we evaluated overall power for genome-wide association studies (GWAS) as measured by the squared correlation between imputed and true genotypes. In the African individuals, at common variants ($> 5\%$ minor allele frequency), imputation r^2 averaged 0.83 for the array data and ranged from 0.89 to 0.95 for the low-pass sequencing data, corresponding to an effective $7 - 15\%$ increase in GWAS discovery power. For the same variants in the European individuals, imputation r^2 averaged 0.91 for the array data and ranged from 0.92-0.96 for the low-pass sequencing data, corresponding to an effective 1-6% increase in GWAS discovery power. Finally, we computed polygenic risk scores for breast cancer and coronary artery disease from the different assays. We observed consistently lower measurement error for risk scores computed from low-pass sequencing data above an effective coverage of ~ 0.5x. The mean squared error of the array-based estimates was three to four times that of the estimates from samples sequenced at an effective coverage of ~ 1.2x for coronary artery disease, with qualitatively similar results for breast cancer. We conclude that low-pass sequencing plus imputation, in addition to providing a substantial increase in statistical power for genome wide association studies, provides increased accuracy for polygenic risk prediction at effective coverages of ~ 0.5x and higher.

1 Introduction

Thousands of variants on the human genome associated with hundreds of complex traits and diseases have been reproducibly and robustly identified since the first large GWASs for complex disease were performed in the mid-00s [1, 2, 3, 4]. Results from these studies have had an enormous impact on the understanding of the genetic architecture underlying complex traits in humans, with these studies playing an essential part in bringing the current understanding full circle back to Fisher's original infinitesimal model compared to the gene-centric model focused upon in the preceding decades [5, 6, 7].

The ability to systematically dissect the genetic architecture of complex traits influenced by hundreds or thousands of genetic variants has largely been enabled by the dense genotyping array, which cost-effectively assays the genome of an individual at hundreds of thousand to millions of loci [2]. Imputation of the resulting genotypes to existing haplotype reference panels further allows evaluation of genetic variants which are not directly assayed, often resulting in total callsets many times the number of directly assayed loci, and is now standard practice in preparing genomic datasets for GWAS [8, 9].

As genome sequencing costs have decreased over the past decade, sequencing-based alternatives to genotyping arrays have been the subject of growing interest [10]. Specifically, low-coverage shotgun whole genome sequencing followed by imputation has been utilized for a number of problems in statistical and population genetics, from providing the backbone for graph-based pangenomes in sorghum to trait mapping in human pharmacogenetics [11, 12, 13, 14, 15, 16, 17, 18]. As an intuition for why this approach is useful, a sample sequenced at a target coverage of 0.5x is expected to have at least one read on 33 million of the 85 million sites in the 1000 Genomes Phase 3 release, whereas a genotyping array will probe a number of variants which is one to two orders of magnitude fewer, albeit with higher average accuracy [19].

For many use cases, there are a number of advantages to low-pass sequencing over genotyping arrays; for example, (1) there is a lack of ascertainment bias with regard to which variants/sites on the genome are assayed, (2) sequence data can be used to discover novel variation both at the sample or population level (such as in [11, 18]), (3) massively parallel sequencing can be achieved by multiplexing large numbers of samples, and (4) the fact that the average expected accuracy of a sample's imputed genotypes can be fine-tuned by adjusting the target coverage for the sample, something which is eminently useful when designing experiments within real-world logistical or budgeting constraints (see [20] for a detailed simulation-based cost-benefit analysis of low-pass sequencing compared to genotyping arrays for GWAS study designs).

However, studies in the literature which investigate the applications of low-pass sequencing often do so by means of simulation or by *downsampling* (often already-aligned) sequence reads from samples previously sequenced at higher coverages [12, 14]. While useful, these approaches are unable to capture the real-world idiosyncrasies of data generation in extremely-low coverage sequencing and ignores factors such as the use of different library preparation methods which are optimized for sequencing at given target coverages [21, 22].

Here, in order to more realistically represent real-world results, we perform an investigation of low-pass sequencing data where the low-coverage genomes are obtained not by downsampling higher-coverage samples but by direct sequencing to extremely low target coverages (0.5x and 1.0x). As a point of comparison, we chose to assay the same samples on the Illumina Global Screening Array (GSA), a modern genotyping array specifically designed to capture multi-ethnic genetic variation. We chose these coverages since the all-in cost of library prep, sequencing, and

data analysis for a 0.5x target coverage sample was (as of 2019) at parity with the Illumina GSA, and that for 1.0x target coverage sequencing fast approaching the same with historically dropping sequencing costs [10].

2 Results

2.1 Experimental Overview

In order to compare the relative performance of low-pass sequencing and genotyping arrays across populations, we selected 60 EUR and 60 AFR individuals (Supplementary Table 1) from the 1000 Genomes Phase 3 release (1KGP3) [19] on which to perform five experiments (Table 1, Supplementary Table 2), which we denote experiments A-E:

For experiment A, we performed library prep on and sequenced these 120 unique individuals in triplicate to a target coverage of 0.5x on an Illumina HiSeqX. For experiment B, we performed library prep on and sequenced these 120 unique individuals in triplicate to a target coverage of 1.0x on an Illumina HiSeqX. For experiment C, we performed library prep on and sequenced NA12878, a CEU female sample, thirty times to a target coverage of 1.0x on an Illumina HiSeqX. For experiment D, we selected a subset of 30 EUR and 30 AFR samples from the set of 120 unique individuals and sent DNA to BGI Americas to be sequenced a target coverage of 1.0x on a BGIseq 500. For experiment E, we assayed these 120 unique individuals in triplicate on the Illumina GSA v3.0 via the Broad Institute.

Experiment C was conducted principally to illustrate the effects of varying empirical coverage on imputation accuracy (with all else held equal) and to provide insight into the repeatability of low-pass sequencing on biological replicates (Supplementary Figure 1).

For each assayed sample passing QC (Methods, Supplementary Table 3), we imputed the sequence or genotype array data to the 1KGP3 haplotype reference panel in a leave-one-out manner (Methods) and compared the imputed calls against the left-out genotypes from the 1KGP3 reference panel (which we treated as the "gold-standard" or "truth" set). We also computed polygenic risk scores from the imputed dosages for each sample and the gold-standard truth set for breast cancer (BC) and coronary artery disease (CAD) using variant weights from recent state-of-the-art studies [23, 24].

Experiment	Mean Coverage	Samples	Assay	Library Prep	Sequencer
A	0.67	120 (3 replicates)	lpSeq	KAPA HyperPlus	Illumina HiSeqX
B	1.25	120 (3 replicates)	lpSeq	KAPA HyperPlus	Illumina HiSeqX
C	1.20	1 (30 replicates)	lpSeq	KAPA HyperPlus	Illumina HiSeqX
D	1.26	60 (1 replicate)	lpSeq	MGIEasy	BGIseq 500
E (array)	NA	120 (3 replicates)	Illumina GSA v3.0	NA	NA

Table 1: Details of experiments conducted. Experiments A-D were based on low-pass sequencing (lpSeq) while experiment E used the Illumina GSA v3.0. The samples column describes the number of unique cell lines and the number of replicates run for each of them. Library prep was performed by Gencove for experiments A-C and libraries were sequenced on the Illumina HiSeqX. For experiment D, DNA was sent to BGI who performed the library prep and sequencing. Empirical coverage for each sample was calculated by dividing the number of bases sequenced by the size of the human genome ($\sim$ 3.3Gb).

2.2 Defining "effective coverage"

The nominal (mapped) coverage of a sample having undergone whole genome sequencing is defined as the number of sequenced (and mapped) bases divided by the size of the genome (in this case, $\sim$ 3.3Gb). This quantity is useful as an indicator of how much sequence data is available for downstream analysis, but does not give any useful information as to how spatially uniform (with respect to the genome coodinate system) the sequenced data are distributed. Spatial uniformity of sequencing reads is particularly important for low-pass sequencing followed by imputation because imputation panels catalogue variation across the entire genome and the imputation quality at a given variant is influenced by the amount of sequence data mapped to regions near that variant and which overlap other variants in the imputation panel.

We therefore introduce the concept of a sequenced sample's *effective coverage* λ_{eff}, which is a function of the fraction of polymorphic sites in a haplotype reference panel covered by at least one sequencing read. Under an idealized Poisson distribution of sequencing reads across sites, this fraction is determined by the sequencing coverage alone (Methods). Specifically, given an imputation panel with n sites and a set of aligned reads from a single sample, we can compute the fraction of those n sites covered by at least one read f_{covered} and compute that sample's effective coverage $\lambda_{\text{eff}} = -\ln(1 - f_{\text{covered}})$.

The advantage of using this quantity rather than nominal coverage as a way to summarize sequence results on a sample, particularly at ultra-low coverages, is that the assumptions of an idealised sampling process is "built-in" to its definition. This allows results from, for example, different library prep methods to be compared on more equal footing (see Figure 1, Supplementary Figures 1, 2, and 3, where experiment D underwent a different library prep method than experiments A-C).

Indeed, plotting nominal and effective coverage vs non-reference concordance (NRC) (Methods) for all the sequenced samples (Figure 1) illustrates how NRC is better predicted by a sample's effective coverage rather than their nominal mapped coverage, with cubic restricted spline fits explaining a larger degree of variance when considering effective coverage rather than mapped nominal coverage ($R^2 = 0.89$ vs $R^2 = 0.82$). Note that this figure and accompanying fit are not meant to be a rigorous parametric treatment of NRC vs. different coverage metrics, but are rather meant to provide an intuition.

The results from experiment C illustrate that this pattern also holds across replicates of the same individual, where the only degree of freedom left is the effective coverage of the sample.

Plots of effective coverage vs NRC and overall concordance broken down by variant type (SNPs vs indels), population, and variant filtration status are shown in Supplementary Figures 4, 5, 6, 7, 8, and 9, and show qualitatively similar results. Notably, when poorly imputed variants are filtered out of a callset (variants with a maximum genotype probability of less than 90%, Methods), the relationship between effective coverage and concordance weakens significantly, suggesting that the genotype posterior probabilities generated during imputation are relatively well-calibrated (Methods).

2.3 Comparison of imputation quality metrics across experiments

2.3.1 Genotype Concordance

We then examined how NRC and imputation r^2 varied across experiments A, B, D, and E. In order to do this, for experiments A, B, and E, we took a random sample of one of the replicates

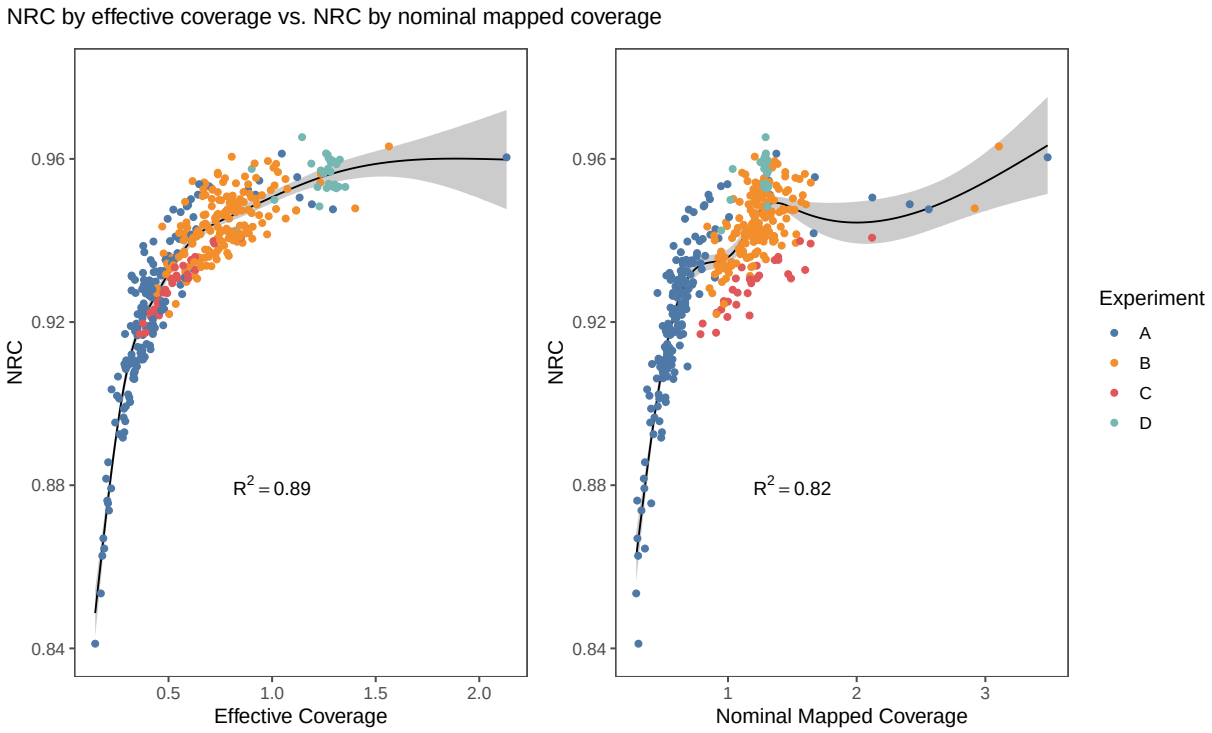

Figure 1: Non-reference concordance (NRC) (Methods) for imputed SNPs for all EUR samples in Experiments A-D plotted against effective coverage λ_{eff} (left pane) or nominal mapped coverage (right pane). We modeled the NRC as the response variable for a $k = 5$ knot cubic restricted spline with the respective coverage metric as the explanatory variable; the fitted values are shown as a solid line with the surrounding shaded regions representing 95% confidence intervals. The knot locations were set at the 5th, 27.5th, 50th, 72.5th, and 95th percentiles of the respective coverage metrics following Harrell's rule of thumb [25].

run for each unique cell line, such that comparisons across experiments concerned only a single, representative sample of each cell line per experiment (Methods). For experiment D we retained all samples as there were no replicates. The remainder of subsection 2.3 compares metrics and results from these *representative cohorts* across experiments. The mean effective coverage of the representative cohorts varied from experiment to experiment (Supplementary Table 4), ranging from an overall (mean $\pm$ standard deviation) of 0.42 ± 0.22 for experiment A to 0.71 ± 0.17 for experiment B to 1.24 ± 0.11 for experiment D.

Reference to non-reference and minor allele frequencies are with respect to those found in the 1KGP3. For all analyses, we treated the genotypes in the 1KGP3 for each sample's cell line as the gold-standard "truth" set.

We observed that imputed genotypes in EUR cohorts were considerably and consistently more accurate than those imputed into the AFR cohorts on average both before and after filtering out poorly imputed variants both for SNPs and indels, with the mean AFR NRC being 7% higher for sequence data with an average of 0.4x effective coverage compared to the array data (Table 2, Supplementary Tables 5, 6, 7).

In order to compare performance across the allele frequency spectrum, we computed the mean

	Experiment			
Super Population	A	B	D	E (array)
AFR	0.9002274	0.9217041	0.9418647	0.8307677
EUR	0.9227215	0.9438073	0.9562273	0.9066613

Table 2: Mean non-reference concordance (NRC) within a representative cohort for unfiltered SNPs by experiment and super population.

non-reference concordance for SNPs within a given allele frequency bin within each representative cohort (Figure 2, Supplementary Figure 10, Supplementary Table 12). The average NRCs for experiment E's representative cohorts were significantly lower than those of all other experiments at all frequency bins, and qualitatively similar patterns hold for *overall* genotype concordance as well (Supplementary Tables 8, 9, 10, 11). These results indicate that low-pass sequencing at an effective coverage of ~ 0.4x or higher consistently yields more accurate imputed genotype calls at sites of common variation than the Illumina GSA and that this pattern holds across both EUR and AFR cohorts.

Interestingly, the NRC at low ($< 5\%$) allele frequencies in AFR populations often exceeds the corresponding NRC for EUR populations (Supplementary Figure 11) for the sequence-based experiments. We also observe this pattern for the imputation r^2s, which we address next.

2.3.2 Imputation r^2

Genotype concordance is an important metric which provides a straightforward and intuitive quantification of the performance of genotype imputation. However, in the context of genome-wide association studies (GWAS), a more relevant quantity is the *imputation* r^2, which is defined as the squared correlation between the imputed dosage and the true genotypes of a given set of samples at a given variant. This is because the imputation r^2 at a given variant is proportional to the expectation of the χ^2 statistic resulting from an association test at that particular variant, of which the power of the test is a function [26, 27, 28]. This quantity can be computed on a site-by-site basis and stratified into allele frequency bins according to the allele frequencies in the haplotype reference panel.

At common variants (minor allele frequency $> 5\%$ in the 1KGP3), the mean imputation r^2 for sequence-based experiments ranged from $0.92 - 0.96$ for the EUR representative cohorts compared to a mean r^2 of 0.91 for experiment E, representing an average increase in power of $\sim 1 - 6\%$ for samples with mean effective coverages ranging from $\sim 0.47 - 1.24$x (Supplementary Tables 4, 15). For the AFR representative cohorts, the mean imputation r^2 for sequence-based experiments ranged from $0.89 - 0.95$ compared to 0.83 for the GSA, representing an average increase of power of $\sim 7 - 15\%$ for samples with mean effective coverages ranging from $\sim 0.38 - 1.24$ (Supplementary Table 15).

Stratifying imputation site-wise r^2s into minor allele frequency bins, we observed that for the sequence-based experiments, the average r^2s across the allele frequency bins scaled with the average effective coverage of each representative cohort (Figure 3), in line with the expectation that higher-effective-coverage sequence data affords greater imputation accuracy.

For the AFR cohorts, we observed that sequence-based experiments uniformly outperformed

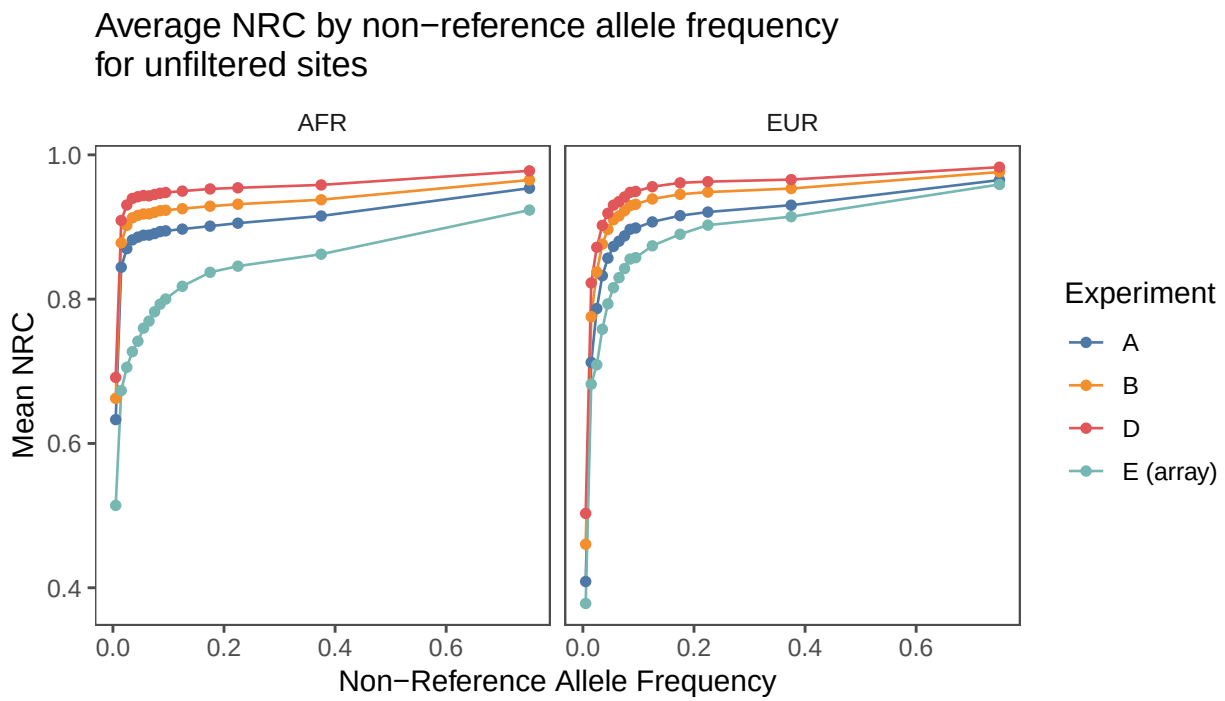

Figure 2: Average non-reference concordance for unfiltered SNPs by superpopulation by non-reference allele frequency in 1KGP3. The NRC for imputed sequence data was consistently higher than the NRC for the imputed GSA data across the allele frequency spectrum. Filtering to confidently imputed variants reveals a similar pattern (Supplementary Figure 10).

genotyping arrays for all minor allele frequency bins, while for EUR cohorts the same was true above minor allele frequencies of 0.01 for all but the cohort with the lowest mean effective coverage, experiment A (Supplementary Figure 12).

Interestingly, at minor allele frequencies of $\sim 5\%$ and lower, for a given sequence-based experiment, imputation performance in the AFR cohorts often exceeded performance at similar allele frequencies in the EUR cohorts, a pattern opposite to that which was observed at higher allele frequencies (Supplementary Figure 12, Supplementary Tables 13, 14). This pattern was not observed for the imputed array data, where the imputation r^2s for all allele frequency bins were higher in the EUR cohort than in the AFR cohort.

Similarly to [26], we hypothesize that this is due to the fact that at higher allele frequencies, the stronger linkage disequilibrium (LD) structure within European populations dominates and affords more accurate haplotype tagging, whereas at lower allele frequencies, the effects of greater genetic diversity in African populations dominate, resulting in a larger number of possible haplotype combinations and an increased chance that any one rare variant is tagged by at least one of these.

Since low-pass sequencing yields measurements at orders of magnitude more sites than genotyping arrays, it is possible to measure a larger range of these possible haplotypes, which is reflected by this pattern holding only for the sequence-based experiments.

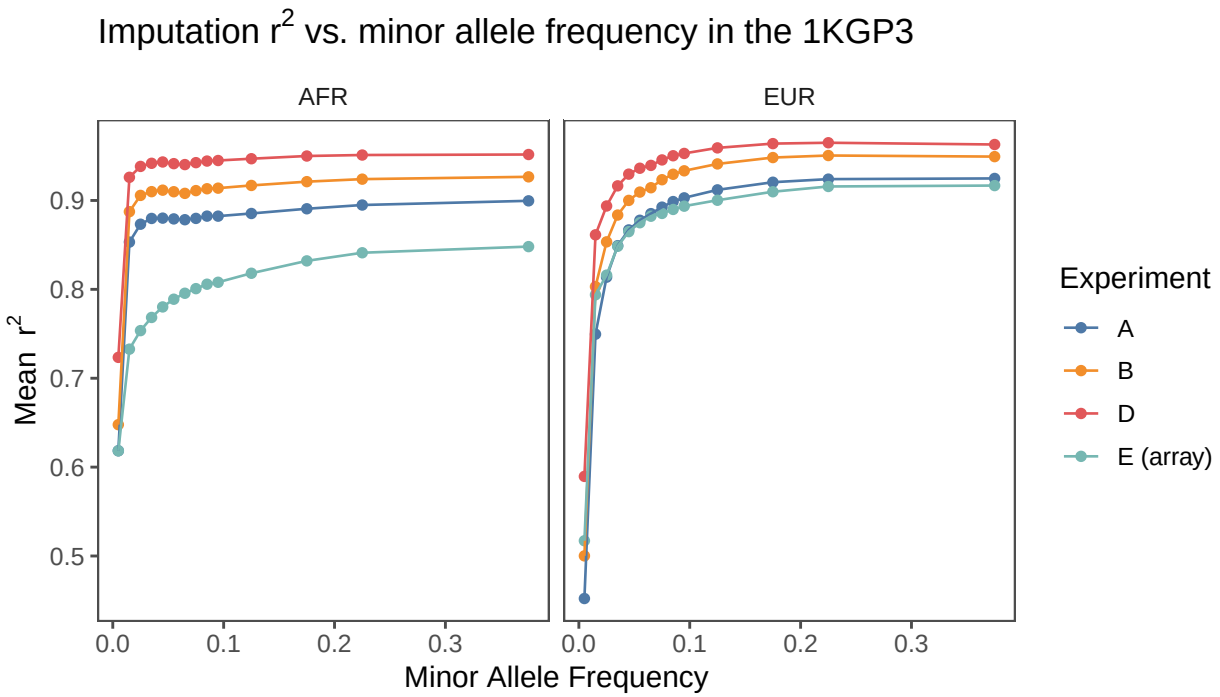

Figure 3: Comparison of imputation quality across experiments for each superpopulation. Variants were binned according to their minor allele frequency in the 1KGP3 and imputation r^2 averaged across variants in each bin. For all experiments the 1KGP3 genotypes were treated as "truth" and imputation r^2 was computed by taking the squared correlation coefficient between the vector of imputed alternate allelic dosages and the truth genotypes. Same results on a log scale are shown in Supplementary Figure 12. Note that imputation performance at low MAF for a given sequencing experiment was often higher in the AFR cohort compared to the EUR cohort.

2.4 Polygenic risk scores

Polygenic risk scores were calculated from imputed dosages for each sample in all experiments for breast cancer (BC) and coronary artery disease (CAD) using recent published results in order to evaluate the accuracy of PRS estimation across experiments and assay types [24, 23]. We chose to use weights from these studies because of their demonstrated ability to stratify disease risk beyond conventional risk factors. Of the 1,745,179 variants with nonzero effect sizes in the CAD PRS, 1,738,589 (99.6%) were present in the haplotype reference panel, and 225,667 were directly typed on the GSA. Of the 313 variants with nonzero effect sizes in the BC PRS, all were present in the haplotype reference panel, and 75 were directly typed on the GSA.

We compared the PRS estimates from each sample to the "true" PRS calculated for that cell line from the 1KGP3 genotypes (Supplementary Figure 14). The estimated scores were highly correlated with the true scores for all experiments, and the r^2s for AFR estimates within experiments and across traits were consistently lower than those for their corresponding EUR counterparts (Supplementary Table 16). This difference was particularly pronounced for the CAD PRS, with r^2s ranging from $0.96 - 0.99$ across experiments for the EUR samples and r^2s ranging from $0.87 - 0.94$ for the AFR samples.

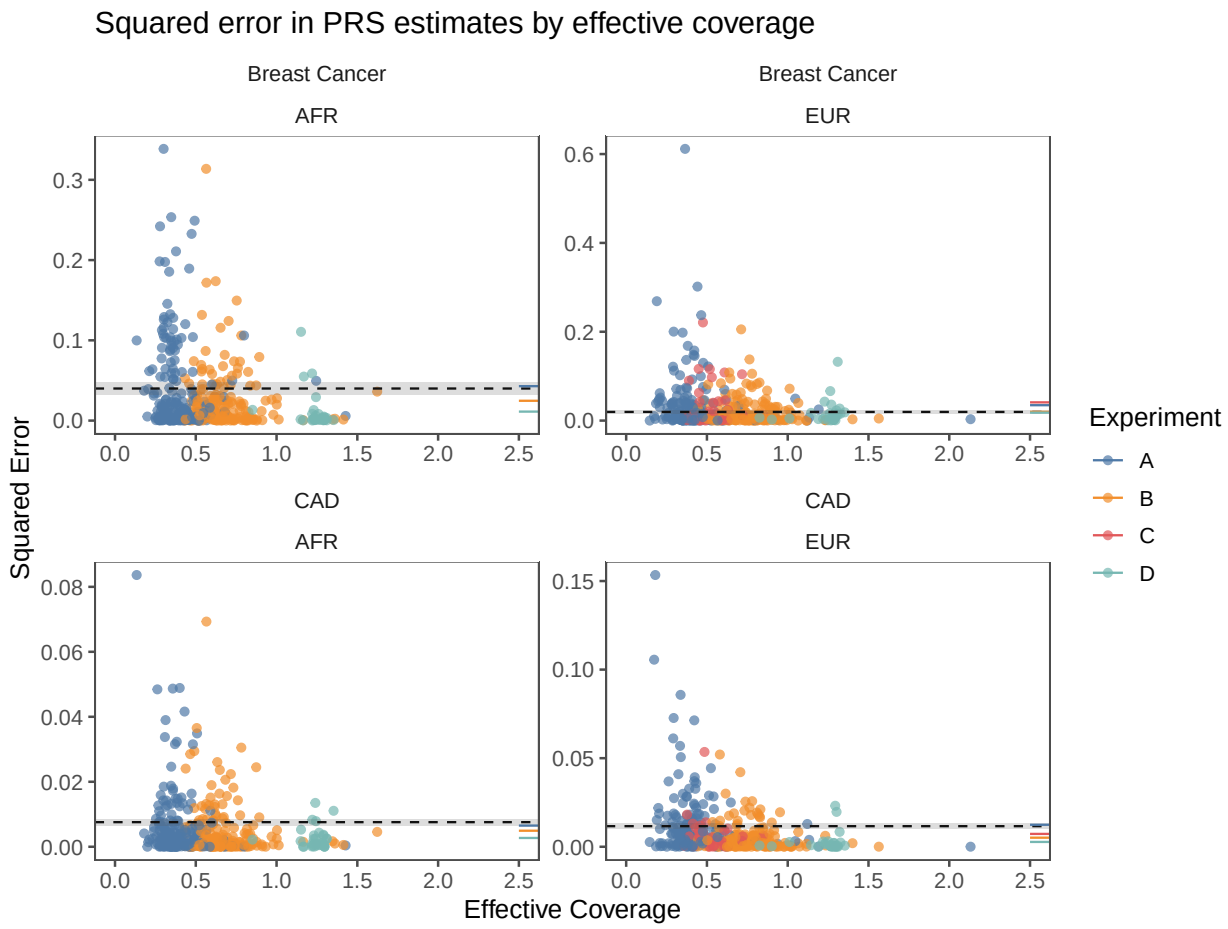

Figure 4: For each imputed sample, we computed a PRS for breast cancer and CAD and calculated the squared error of the estimate compared to the PRS of the "truth" genotypes in the 1KGP3. Each dot represents this squared error for each sequenced sample for a given trait and is colored by which experiment it belongs to. To provide a point of comparison to PRS estimated from imputed array data, we computed the mean squared error for each cell line across array replicates and averaged that across all cell lines. This quantity is represented by the dashed line for each trait and superpopulation along with the standard error of the mean, represented by the shaded regions about each dashed line. The mean squared error for each experiment was calculated in the same way and is rendered as a colored line segment on the rightmost margin of each pane. These results indicate that sequencing at effective coverages of 0.5x or higher generally affords lower measurement error in PRS estimates.

We investigated the relationship between measurement error in PRS (as quantified by the squared error of an estimate) and effective coverage for the sequence-based experiments. Figure 4 shows that the squared error in PRS estimates for all experiments across populations and traits decreases with increasing effective coverage, and that the measurement error in samples sequenced to an effective coverage of ~ 0.5x or higher generally affords lower measurement error than the array-based estimates, with samples sequenced to an effective coverage of more than 1x (experiment D) having an approximately three- to four-fold decrease in squared error for both traits in the AFR cohort and the EUR cohort for CAD, and around the same squared error for the EUR cohort for BC (~ 1.08-fold decrease) (Supplementary Tables 17, 18).

One thing to note is that the measurement error of PRS estimates depends on the quality of genotypes at the variants involved in computing the PRS, which means that for arrays, the measurement error will depend on the proportion of the variants involved that are directly typed versus imputed. Here, a minority of the variants comprising the polygenic risk scores (13% and 24% for CAD and BC respectively) were directly typed on the GSA. Presumably, the measurement error for array-based estimates would be substantially lower in a different situation in which all the variants comprising a PRS are directly typed instead.

3 Discussion

The use of dense genotyping arrays followed by imputation to a haplotype reference panel has enabled population-scale genome wide associations studies to become routine in characterizing the genetic architecture of complex traits and disease. An alternative technology is low-pass whole genome sequencing followed by imputation, which has successfully used for a number of problems in statistical and population genetics [11, 20, 12], and which was recently shown to recapitulate comparable disease risk stratification performance using a handful of polygenic risk scores derived from imputed array data from an independent cohort [14].

We introduced the notion of *effective coverage*, a quantity which describes the coverage of a sequenced sample under an ideal sampling process and which is more predictive of imputation accuracy than nominal coverage. We showed that with increasing effective coverage, genotyping concordance and imputation r^2 increase commensurately, while the measurement of PRS estimates from the imputed genotypes decreases. We note that because the same amount of sequencing coming from different library preparation methods can yield different effective coverages, systematic comparisons of different library preparation methods on this metric is warranted.

We observed that at sites of common variation (MAF $> 5\%$), imputation r^2 was consistently and substantially higher in imputed sequence data compared to imputed array data, and in Europeans compared to Africans (Supplementary Table 15). Conversely, at rare variants (MAF $< 1\%$), we observed that imputation r^2 was consistently higher in Africans than Europeans, a result which we hypothesize is due to different aspects of the LD structure in these populations dominating in different regimes of the allele frequency spectrum. A consequence of this observation is that studies of rare variants may (all other things being equal) be more powerful in African-ancestry cohorts as compared to European-ancestry cohorts under some study designs.

We compared PRS estimates for coronary artery disease and breast cancer, and found that the measurement error of PRS estimates decreased with increasing effective coverage for sequenced samples. For CAD, we found that low-pass sequence data yielded consistently lower measurement error in PRS estimates in both the African and European cohorts, with samples sequenced to an

effective coverage of $\sim$ 1.2x yielding an approximately three- to four-fold decrease in mean squared error when compared to PRS estimated from the Illumina GSA. The same decrease was observed for BC estimates in the African cohorts, whereas the mean squared error for BC in the European cohort at that effective coverage was around the same as the array estimates ($\sim$ 1.08-fold relative decrease).

Since imputation accuracy from genotyping array data is known to depend heavily on the size and composition of the haplotype reference panel used [29], it will be interesting to replicate these results for low-pass sequence across different panels, particularly as extremely large panels (*e.g.*, HRC and TOPmed [30, 31]) as well as panels comprising currently underrepresented populations (*e.g.*, NeuroGAP-Psychosis [32]) come online.

In the context of understanding of how genetic variation contributes to phenotypic variance, it has become increasingly clear in recent years that a whole-genome approach is necessary due to the fact that the degree of polygenicity underlying the majority of complex traits in humans is perhaps more accurately represented by Fisher's infinitesimal model (in which variation arises from infinitely many loci of infinitesimal effect) than it is by what Risch et al. called a "large number of loci (perhaps $\geq$ 15)", and indeed also that complex traits are mainly driven by non-coding variants [33, 6, 34, 35]. Consequently, methods to accurately assess genetic variation across the whole genome in diverse populations will be become increasingly essential to ongoing efforts to elucidate the genetic architecture of complex traits, as well as the population genetic processes which gave rise to it [36, 37].

For future directions, there are methodological advances that could be made in order to fully leverage the information that sequencing provides. Genotyping arrays suffer from ascertainment bias which manifests in two ways: (1) measurements are always made on the same set of loci across the genome, and (2) measurements at a given loci do not allow for novel variant discovery (*i.e.*, you have to know what you are looking for) [38, 39].

Low-pass sequencing by its nature overcomes (1) but whether a problem similar to (2) remains depends on the analytical techniques utilized downstream of actual sequencing. For instance, the current implementation of the imputation tool used here (`loimpute` [15]) does not allow for variants which are novel to the reference panel to be called, thus causing a similar effect in post-imputation sequence data as (2) in unimputed array data. In other words, this means that the effective upper bound in genotyping accuracy is governed by the composition of the reference panel and whether a particular individual's genetic variation is catalogued therein. A potential methodological improvement would thus be to develop a way to enable novel variant calls at sites with overwhelming read-based evidence.

4 Conclusion

As research into the genetic architecture of complex disease and traits continues to accelerate, it will become increasingly important for data generation techniques to be as population-agnostic as possible in order to capture global genetic variation in an unbiased manner. Our results demonstrate that low-pass sequencing provides a competitive alternative to genotyping arrays in the context of genome-wide association studies and polygenic risk scoring across diverse populations.

5 Methods

5.1 Data generation

Purified genomic DNA (gDNA) from 60 selected individuals of European ancestry and 60 selected individuals of African ancestry from the 1000 Genomes Project Phase 3 was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. Genomic DNA is extracted from immortalized B-lymphocytes and eluted in TE buffer (10mM Tris, pH 8.0/1mM EDTA) for shipping.

To prepare the 120 gDNA samples for sequencing, 30µL from each was plated and diluted to 10ng/µL using 10mM Tris-HCl and sequencing libraries were prepared in triplicate using a miniaturized version of the KAPA HyperPlus kit (Roche, KK8514) with Illumina-compatible KAPA Unique Dual-Indexed Adapters (Roche, KK8727).

Following library preparation, a subset volume of 10µL was pooled from each of the resulting libraries. The pooled libraries were purified with a double-sided size-selection using SPRIselect paramagnetic beads (Beckman Coulter Life Sciences, B23318) to narrow the library fragment size range and remove dimerized adapters from the samples. To characterize the purified pools, the concentration was measured using the Invitrogen Qubit Fluorometer (ThermoFisher Scientific, Q33238) and the library fragment size was assessed using the Agilent 2100 Bioanalyzer (Agilent, G2939BA) with the High Sensitivity DNA Kit (Agilent, 5067-4626).

Sequencing of purified library pools of 30 and 60 samples was performed to 0.5x coverage and 1x target coverages, respectively, using paired-end 150bp reads on the Illumina HiSeqX platform.

To prepare NA12878 for sequencing 30 times to 1x target coverage, 10µL from the source gDNA sample was aliquoted into 30 separate wells of a 96-well plate and diluted to 10ng/µL using 10mM Tris-HCl. Sequencing libraries were prepared from these diluted samples using a miniaturized version of the KAPA HyperPlus kit (Roche, KK8514) with Illumina-compatible KAPA Unique Dual-Indexed Adapters (Roche, KK8727). The 30 libraries were pooled and purified with a double-sided size-selection using SPRIselect paramagnetic beads (Beckman Coulter Life Sciences, B23318), checked for concentration and fragment size, and sequenced on a single lane of an Illumina HiSeqX flow cell.

A subset of 1µg gDNA from 30 samples selected from each population (60 total) was aliquoted into plates and submitted to BGI Americas Corporation for DNA nanoball library prep and sequencing (DNBseq^{TM}) using paired-end 100bp reads on the BGISEQ-500 sequencing instrument.

Sample genotyping was performed using the Infinium Illumina Global Screening Array v3.0 (Illumina, 20030770) by the Broad Institute Genomic Services group. Each of the 120 samples was genotyped in triplicate. To prepare the 360 samples for genotyping, a total mass of 1µg gDNA from each sample was aliquoted into barcode-labeled tubes and submitted to the Broad Institute for processing.

5.2 Quality control

Of the 360 samples in experiment A assayed on the Illumina HiSeqX, 351 passed QC and the remainder failed due to contamination or low read count (less than 0.1x nominal coverage). Of the 360 samples in experiment B assayed on the Illumina HiSeqX, 350 passed QC and the remainder failed due to contamination or low read count (less than 0.1x nominal coverage). Of the 30 samples in experiment B assayed on the Illumina HiSeqX, all 30 passed QC. Of the 60 samples in experiment

D sequenced on BGI machines, 58 passed QC and 2 failed due to contamination. Of the 360 samples in experiment E assayed on the Illumina GSA, 358 passed QC and 2 failed due to low call rate (below 97%).

See Supplementary Tables 2, 3 for a more comprehensive breakdown.

5.3 Effective coverage

Consider an idealised sampling process whereby shotgun sequence data is generated for a given sample to a coverage of λ. We can then model the number of reads k on a site on the genome as a random variable described by a Poisson distribution thus parameterized. Recall that the probability mass function for a Poisson distribution with these parameter is

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{1}$$

Then the probability that this site is covered by at least one read is

$$\begin{aligned} P(k>0;\lambda) &= 1 - \sum_{k=1}^{\infty} f(k;\lambda) && (2)\\ &= 1 - f(k=0;\lambda) && (3)\\ &= 1 - e^{-\lambda}. && (4) \end{aligned}$$

Suppose now that we have a set of n such sites on the genome (which in our case represents the n sites in a haplotype reference panel) which are all independent and identically distributed. Then the total number of sites X covered by at least one read is described by a binomial distribution with $p = P(k>0;\lambda) = 1 - e^{-\lambda}$, which has an expected value of $\mathrm{E}[X] = np = n(1-e^{-\lambda})$. Defining $f_{\text{covered}} = X/n$ as the *proportion* of sites covered, we have $\mathrm{E}[f_{\text{covered}}] = (1-e^{-\lambda})$.

This quantity describes the expected proportion of sites in a haplotype reference panel covered by at least one read under the assumptions described above, and f_{covered} is a quantity which can be empirically computed for any given sample that has been sequenced. The definition of *effective coverage* then follows naturally as that value of λ which one obtains when plugging in an observed f_{covered} into the relation:

$$\text{Effective coverage} = \lambda_{\text{eff}} := -\ln(1 - f_{\text{covered}}) \tag{5}$$

which, as we show in the main text, is more predictive of variant call accuracy than nominal mapped coverage.

5.4 Overall and non-reference concordance

Consider two sets of genotype calls at the same set of n sites/variants, with each genotype being coded 0, 1, and 2 for homozygous reference, heterozygous, and homozygous alternate allele. Assume further that there is no missingness in these genotypes. Then at each site, compare the genotype calls between the two callsets — there are nine possible combinations of genotypes $((0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2))$. Each combination can be represented as a cell in a 3x3 table like Table 3, and the total number of each combination can be tallied across all

sites. For instance, if the callsets had 100 sites at which both samples were homozygous reference (corresponding to $(0, 0)$), then at the end of tallying, a would be equal to 100 in Table 3.

Then we define the non-reference concordance (NRC) between the two callsets as

$$\text{NRC} = \frac{e + i}{b + c + d + e + f + g + h + i}. \tag{6}$$

Similarly, the overall concordance is defined as

$$\text{Overall concordance} = \frac{a + e + i}{a + b + c + d + e + f + g + h + i}. \tag{7}$$

	0	1	2
0	a	b	c
1	d	e	f
2	g	h	i

Table 3: Possible combinations of genotype calls between two callsets at a given biallelic site. The diagonal represents concordant calls.

5.5 Selecting samples for representative cohorts

In order to select the representative samples for each experiment for each superpopulation cohort in section 2.3, we chose one sample for each cell line for each super population in each experiment so that results compared across assays would concern the exact same samples.

For AFR cell lines, there was at least one replicate in experiments A, B, and E which passed QC for all 60 unique cell lines, so the set of representative samples comprised all 60 cell lines (Supplementary Table 3). For EUR cell lines, there were 57 out of 60 unique cell lines which had at least one replicate in experiments A, B, and E which passed QC, so the set of representative sample comprised a single sample per experiment of these 57 cell lines (Supplementary Table 3).

5.6 Imputation

For the genotyping array data, we used `Eagle v2.4.1` [40] to perform reference-based phasing for each sample and `minimac4` [41] for imputation. We received sample-level VCFs with genotypes on the hg19 build of the human reference genome from the Broad Institute; to prepare the samples for phasing and imputation, we filtered out failing probes and duplicate/multiallelic variant probes as marked in the `FILTER` field of the VCFs.

The sequence data were aligned to the `hs37-1kg` reference genome (obtained from NCBI at the following URL: `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz`) and Gencove's `loimpute` for imputation. The model underlying Gencove's `loimpute` has been previously described in the supplementary note to [15].

All samples were imputed to the 1000 Genomes Phase 3 release haplotype reference panel on `hg19` [19] in a leave-one-out manner (such that the cell line the sample belongs to was not in the panel being used). The same was done for the reference-based phasing step for the array data.

Poorly imputed variants were marked as those variants for which the none of the posterior probabilities (the `GP` field in a VCF) for the possible genotypes (hom-ref, heterozygous, hom-alt) was greater than 0.9. In other words, when we refer to "filtered variants", we are referring to the callset of imputed variants with variants with `max(GP) < 0.9` removed.

6 Declarations

6.1 Ethics approval and consent to participate

Not applicable.

6.2 Consent for publication

Not applicable.

6.3 Acknowledgements

We thank Maria Vazquez for feedback on the study design and operational support of the sequencing and genotyping experiments.

6.4 Competing interests

J.H.L., C.A.M., T.B., and J.K.P. were employees of Gencove, Inc. at the time of writing.

6.5 Authors' contributions

J.H.L., C.A.M, T.B., and J.K.P. conceived of the study. J.H.L conducted the analyses. C.A.M was responsible for the experimental work and data acquisition. J.H.L., J.K.P., T.B., and C.A.M. contributed to and provided feedback on the manuscript.

6.6 Availability of data and materials

The code used to perform tertiary analysis, figure generation, and table generation along with the source code for this paper itself is publicly available at `https://gitlab.com/gencove/data-science/presentations-papers-publications/sbir-paper`. The datasets generated and analysed during the current study are available in a public AWS s3 bucket at `s3://gencove-sbir/`, accessible also at the following URL: `https://gencove-sbir.s3.amazonaws.com/index.html`. The original fastqs for all samples sequenced as well as the genotype calls resulting from imputation have been deposited there along with both the raw array genotypes for the GSA as delivered by the Broad Institute and the imputed array genotypes.

6.7 Funding

Research reported in this publication was supported by a Phase 1 SBIR grant from the NIH (contract number 1R43HG010596-01).

References

[1] Guy Sella and Nicholas H Barton. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annual review of genomics and human genetics*, 20:461–493, 2019.

[2] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[3] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[4] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[5] Bruce Walsh and Michael Lynch. *Evolution and selection of quantitative traits*. Oxford University Press, 2018.

[6] Neil Risch, Donna Spiker, Linda Lotspeich, Nassim Nouri, David Hinds, Joachim Hallmayer, Luba Kalaydjieva, Patty McCague, Sue Dimiceli, Tawna Pitts, et al. A genomic screen of autism: evidence for a multilocus etiology. *The American Journal of Human Genetics*, 65(2):493–507, 1999.

[7] Horace Freeland Judson. The eighth day of creation. *New York*, page 550, 1979.

[8] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387–406, 2009.

[9] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

[10] Kris A Wetterstrand. DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). National Human Genome Research Institute 2019. 2019.

[11] Ngoc Hieu Tran, Thanh Binh Vo, Nhat Thang Tran, Thu-Huong Nhat Trinh, Hong-Anh Thi Pham, Hong Thuy Dao, Ngoc Mai Nguyen, Yen-Linh Thi Van, Vu Uyen Tran, Hoang Giang Vu, et al. Genetic profiling of 2,683 Vietnamese genomes from non-invasive prenatal testing data. *bioRxiv*, 2019.

[12] Arthur Gilly, Lorraine Southam, Daniel Suveges, Karoline Kuchenbaecker, Rachel Moore, Giorgio EM Melloni, Konstantinos Hatzikotoulas, Aliki-Eleni Farmaki, Graham Ritchie, Jeremy Schwartzentruber, et al. Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics*, 35(15):2555–2561, 2019.

[13] Simone Rubinacci, Diogo Ribeiro, Robin Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *bioRxiv*, 2020.

[14] Julian R Homburger, Cynthia L Neben, Gilad Mishne, Alicia Y Zhou, Sekar Kathiresan, and Amit V Khera. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome medicine*, 11(1):1–12, 2019.

[15] Kaja Wasik, Tomaz Berisa, Joseph K Pickrell, Jeremiah H Li, Dana J Fraser, Karen King, and Charles Cox. Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *bioRxiv*, 2019.

[16] Sarah Jensen, Jean Rigaud Charles, Kebede Muleta, Peter Bradbury, Terry Casstevens, Santosh P Deshpande, Michael A Gore, Rajeev Gupta, Daniel C Ilut, Lynn Johnson, et al. A sorghum Practical Haplotype Graph facilitates genome-wide imputation and cost-effective genomic prediction. *bioRxiv*, 2019.

[17] Na Cai, Tim B Bigdeli, Warren Kretzschmar, Yihan Li, Jieqin Liang, Li Song, Jingchu Hu, Qibin Li, Wei Jin, Zhenfei Hu, et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*, 523(7562):588–591, 2015.

[18] Siyang Liu, Shujia Huang, Fang Chen, Lijian Zhao, Yuying Yuan, Stephen Starko Francis, Lin Fang, Zilong Li, Long Lin, Rong Liu, et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell*, 175(2):347–359, 2018.

[19] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[20] Bogdan Pasaniuc, Nadin Rohland, Paul J McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M Neale, Mark J Daly, Pamela Sklar, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature genetics*, 44(6):631, 2012.

[21] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome biology*, 12(2):R18, 2011.

[22] Marcus B Jones, Sarah K Highlander, Ericka L Anderson, Weizhong Li, Mark Dayrit, Niels Klitgord, Martin M Fabani, Victor Seguritan, Jessica Green, David T Pride, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences*, 112(45):14024–14029, 2015.

[23] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34, 2019.

[24] Michael Inouye, Gad Abraham, Christopher P Nelson, Angela M Wood, Michael J Sweeting, Frank Dudbridge, Florence Y Lai, Stephen Kaptoge, Marta Brozynska, Tingting Wang, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology*, 72(16):1883–1893, 2018.

[25] Frank Harrell. Regression modeling strategies. *BIOS*, 330:2018, 2017.

[26] Jonathan Marchini. *Haplotype Estimation and Genotype Imputation*, chapter 3, pages 87–114. John Wiley & Sons, Ltd, 2019.

[27] Juliet M Chapman, Jason D Cooper, John A Todd, and David G Clayton. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human heredity*, 56(1-3):18–31, 2003.

[28] Jonathan K Pritchard and Molly Przeworski. Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69(1):1–14, 2001.

[29] Haiko Schurz, Stephanie J Müller, Paul David Van Helden, Gerard Tromp, Eileen G Hoal, Craig J Kinnear, and Marlo Möller. Evaluating the accuracy of imputation methods in a five-way admixed population. *Frontiers in genetics*, 10:34, 2019.

[30] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279–1283, 2016.

[31] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*, 2019.

[32] Anne Stevenson, Dickens Akena, Rocky E Stroud, Lukoye Atwoli, Megan M Campbell, Lori B Chibnik, Edith Kwobah, Symon M Kariuki, Alicia R Martin, Victoria de Menil, et al. Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. *BMJ open*, 9(2):bmjopen–2018, 2019.

[33] Ronald A Fisher. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[34] Nicholas H Barton, Alison M Etheridge, and Amandine Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical population biology*, 118:50–73, 2017.

[35] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[36] Jeremy J Berg and Graham Coop. A population genetic signal of polygenic adaptation. *PLoS genetics*, 10(8), 2014.

[37] Jing Guo, Yang Wu, Zhihong Zhu, Zhili Zheng, Maciej Trzaskowski, Jian Zeng, Matthew R Robinson, Peter M Visscher, and Jian Yang. Global genetic differentiation of complex traits shaped by natural selection in humans. *Nature communications*, 9(1):1–9, 2018.

[38] Joseph Lachance and Sarah A Tishkoff. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, 35(9):780–786, 2013.

[39] Rasmus Nielsen. Population genetic analysis of ascertained SNP data. *Human genomics*, 1(3):218, 2004.

[40] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, 48(11):1443, 2016.

[41] S. Das, G. Abecasis, and C. Fuchsberger. Minimac4: A next generation imputation tool for mega reference panels. Abstract 1278W. Presented at the the 65th Annual Meeting of the American Society of Human Genetics, October 7, 2015, Baltimore, MD, 2015.