Published in final edited form as:

Hum Genet. 2018 April ; 137(4): 343–355. doi:10.1007/s00439-018-1886-z.

High-depth whole genome sequencing of an Ashekenazi Jewish reference panel: Enhancing sensitivity, accuracy, and imputation

Todd Lencz^{#1,2,3,†}, Jin Yu^{#2,3}, Cameron Palmer⁴, Shai Carmi⁵, Danny Ben-Avraham^{6,7}, Nir Barzilai^{6,7}, Susan Bressman⁸, Ariel Darvasi⁹, Judy H. Cho^{10,11}, Lorraine N. Clark^{12,13}, Zeynep H. Gümü^{11,14}, Vijai Joseph¹⁵, Robert Klein^{11,14}, Steven Lipkin¹⁶, Kenneth Offit^{15,17}, Harry Ostrer^{6,18}, Laurie J. Ozelius¹⁹, Inga Peter^{11,14}, Gil Atzmon^{6,7,20}, Itsik Pe'er^{4,21,†}

¹Departments of Psychiatry and Molecular Medicine, Hofstra Northwell School of Medicine, Hempstead, New York 11550

²Department of Psychiatry, Division of Research, The Zucker Hillside Hospital Division of Northwell Health, Glen Oaks, NY, 11004

³Center for Psychiatric Neuroscience, The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, 11030

⁴Department of Computer Science, Columbia University, 500 W 120th St, New York, NY, 10027

⁵Braun School of Public Health, Faculty of Medicine, Hebrew University of Jerusalem, Ein Kerem, Jerusalem, 9112102, Israel

⁶Department of Genetics, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461

⁷Department of Medicine, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461

⁸Department of Neurology, Beth Israel Medical Center, New York, New York, USA. 10003

⁹Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Givat Ram, Jerusalem, Israel, 91904

¹⁰Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl, New York, NY 10029

¹¹Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl, New York, NY, 10029

[†] Corresponding authors: tlencz@northwell.edu; itsik@cs.columbia.edu.

Author Contributions

TL and IP led the analysis, and led the writing of the manuscript. JY, CP, and SC conducted the primary analyses. TL led the funding of the study. TL, AD, GA, DB, NB, and LNC provided samples and conducted lab work. TL, IP, NB, SB, AD, JHC, LNC, ZHG, VJ, RK, SL, KO, HO, LJO, IP, and GA initiated and designed the study, and provided funding. </author_notes>

Competing financial interests: The authors declare no competing financial interests.

Accession codes: Whole genome sequence data have been deposited at the European Genome-phenome Archive (EGA, <http://www.ebi.ac.uk/ega/>), which is hosted by the EBI, under accession code EGAS00001000664. Genotype data for target samples is available at The database of Genotypes and Phenotypes (dbGaP, <https://www.ncbi.nlm.nih.gov/gap>), under accession number phs000448.v1.p1.

- ¹²Department of Pathology and Cell Biology, Columbia University Medical Center, 1150 St Nicholas Ave, New York, NY, 10032
- ¹³Taub Institute for Research of Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, 1150 St Nicholas Ave, New York, NY, 10032
- ¹⁴Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl, New York, NY 10029
- ¹⁵Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY, 10065
- ¹⁶Departments of Medicine, Genetic Medicine and Surgery, Weill Cornell Medical College, New York, NY
- ¹⁷Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY, 10065
- ¹⁸Department of Pathology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY, 10461
- ¹⁹Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114
- ²⁰Department of Human Biology, Haifa University, Haifa, Israel
- ²¹Center for Computational Biology and Bioinformatics, Columbia University, 1130 St Nicholas Ave, New York, NY, 10032
- # These authors contributed equally to this work.

Abstract

While increasingly large reference panels for genome-wide imputation have been recently made available, the degree to which imputation accuracy can be enhanced by population-specific reference panels remains an open question. Here, we sequenced at full-depth (30x), across two platforms (Illumina X Ten and Complete Genomics, Inc.), a moderately large (n=738) cohort of samples drawn from the Ashkenazi Jewish population. We developed a series of quality control steps to optimize sensitivity, specificity, and comprehensiveness of variant calls in the reference panel, and then tested the accuracy of imputation against target cohorts drawn from the same population. Quality control (QC) thresholds for the Illumina X Ten platform were identified that permitted highly accurate calling of single nucleotide variants across 94% of the genome. QC procedures also identified numerous regions that are poorly mapped using current reference or alternate assemblies. After stringent QC, the population-specific reference panel produced more accurate and comprehensive imputation results relative to publicly available, large cosmopolitan reference panels, especially in the range of rare variants that may be most critical to further progress in mapping of complex phenotypes. The population-specific reference panel also permitted enhanced filtering of clinically irrelevant variants from personal genomes.

Keywords

Whole genome sequencing; imputation; variant calling; reference panel; personal genomes; Ashkenazi

Introduction

While genome-wide association studies (GWAS) have traditionally focused on the role of common genetic variation in common disease, it is increasingly acknowledged that rare genetic variation also significantly influences complex phenotypes (Kiezun et al., 2012). The cost of sequencing to identify rare variation has declined dramatically in recent years; nevertheless, exome sequencing remains an order of magnitude higher in price compared to common SNP arrays, and high-depth whole-genome sequencing is even more expensive. Consequently, imputation has emerged as a popular approach that enables the examination of rare variants in the context of large-scale association studies in which subjects have been genotyped on conventional SNP arrays (Hoffman & Witte, 2015).

The accuracy and comprehensiveness of imputation depends on several factors, such as the size of the reference panel employed and the ancestry of its members relative to the target panel. Considerable effort has recently been devoted to generating increasingly large reference panels, usually involving cohorts sequenced at low-moderate (~4–10x) depth. Ancestry of these cohorts ranges from sampling across the global human population (e.g., 1000 Genomes Project Consortium (1000 Genomes Project Consortium, 2015)) to focusing on individuals from a particular population of origin such as the UK10K project (UK10K Consortium, 2015) and the Genome of the Netherlands Consortium (Genome of the Netherlands Consortium, 2014). Attempts to characterize the trade-offs between breadth and depth of ancestry sampling have led to conflicting conclusions (Zhang et al. 2013; Deelen et al., 2014), although recent studies have consistently demonstrated enhanced accuracy of imputation for a given subpopulation when cosmopolitan reference panels such as 1000 Genomes are supplemented with data from population-specific reference panels (Wong et al., 2013; Pistis et al., 2015). Similar results have been obtained in non-human species (Ventura et al. 2014), with increasing diversity of the reference panel especially beneficial for imputation of low-frequency alleles (Druet et al. 2014).

Most recently, the Haplotype Reference Consortium (McCarthy et al., 2016) has assembled data from multiple worldwide studies involving a total of ~32,000 sequenced participants. While this resource has been demonstrated to enhance accuracy of imputation and, consequently, power of GWAS compared to the 1000 Genomes reference panel (Iglesias et al., 2017), it remains to be seen whether accuracy can be further enhanced by population-specific sequencing.

In the present study, we tested whether imputation of rare variants could be improved by the addition of a moderately large ($n=738$), population-specific reference panel sequenced to full depth (~30x). Specifically, we examined the Ashkenazi Jewish (AJ) population, which possesses unique characteristics, making it a compelling model for genetic investigation. While comprising ~ 10^7 individuals today, the AJ population descends from a founding bottleneck that is very narrow, with effective population size likely less than a thousand chromosomes. This bottleneck is also very recent, having taken place only ~30 generations before present (Palamara et al., 2012; Carmi et al., 2014).

Development of enhanced referenced panels of sequenced individuals from a population also has relevance to the clinical setting, in which any personal genome includes many variants of unknown significance (Dewey et al., 2014). Such panels permit the identification of alleles that segregate in a population at appreciable frequencies, which otherwise might be erroneously interpreted as unique to a given patient with an unusual phenotype (MacArthur et al., 2014). As with imputation, recent studies with increasingly large reference panels suggest that such distinctions remain problematic due to the extremely diffuse nature of human genetic diversity (Walsh et al., 2017), and can be especially difficult in founder and endogamous populations with unique demographic histories (Ankala et al., 2015). The development of such resources is especially urgent given the recent observation that many previously-identified risk genes and variants may be erroneously classified due to lack of sufficiently powerful reference databases (Whiffin et al. 2017; Walsh et al. 2017).

Thus, the present study is intended as a resource for the research and clinical genomics communities to enhance the interpretability of both large-scale genotype datasets and individual-level sequence data. In creating our high-depth, population-specific reference panel, we first found it necessary to refine a set of practical quality control (QC) guidelines for high-depth whole-genome sequencing. These QC procedures are described in the Materials and Methods section below, and QC-relevant observations are described in the initial portion of the Results section.

Materials and Methods

Cohort description

Samples sequenced using CGI (n=168) included control samples (n=128) described previously (Carmi et al., 2014) as well as schizophrenia cases (n=40). Illumina samples (n=574) included samples described previously from multiple case-control cohorts described in Supplementary Table 1. By design, these samples included some overlapping subsets CGI controls (n=5) and CGI schizophrenia cases (n=4). All samples had been self-reported to be Ashkenazi Jewish, and also verified as Ashkenazi Jewish by principal components analysis of previously collected SNP array data as described in our prior publications (Atzmon et al. 2010; Guha et al. 2012). Informed consent was obtained in accordance with institutional policies and the study was approved by the corresponding institutional review boards.

Sequencing and analysis pipeline

For samples sequenced on the Illumina platform, genomic DNA was isolated from whole blood and was quantified using PicoGreen on a Spectramax fluorometer (Molecular Devices) or Qubit (Life Technologies), and integrity assessed using the Fragment Analyzer (Advanced Analytical). A separate aliquot was removed and used for SNP array genotyping using the HumanExome-12 v1.2-A (Illumina) chip. Sequencing libraries were prepared using the Illumina TruSeq Nano DNA kit, with 100ng input gDNA following manufacturer recommendations. Briefly, DNA was first sheared on a Covaris sonicator, followed by end-repair of the fragmented molecules and bead based size selection. Size selected molecules were then A-tailed followed by ligation of sequencing adaptors. Libraries were evaluated using a BioAnalyzer (Agilent), and quantified by qPCR (Kappa) and PicoGreen.

Libraries were sequenced on the Illumina HiSeq X ten with v1 chemistry. Sequencing libraries were pooled in equimolar amounts (8 samples / pool), and 2.5nM or 3nM pooled library was loaded onto each lane of the patterned flow cell, and clustered on a cBot (Illumina). Each pool of libraries was sequenced on 8 lanes of a flow cell. The HiSeq X generates ~375–400M pass filter 2×150bp per flow cell lanes. For samples that did not meet 30x mean genome coverage post alignment, additional aliquots of the sequencing libraries were pooled in proportion to the amount of additional reads needed, and re-sequenced in one or more flow cell lanes.

Upon completion of sequencing run, bcl files were demultiplexed and quality of sequencing data reviewed using SAV software (Illumina) and FastQC (downloaded from <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) for deviations from expected values with respect to total number of reads, percent reads demultiplexed (>95%), percent clusters pass filter (>55%), base quality by lane and cycle, percent bases >Q30 for read 1 and read 2 (>75%), GC content, and percent N-content. FastQ files were aligned to hg19/GRCh37 using the Burrows-Wheeler Aligner (BWA-MEM v0.78) (Li & Durbin, 2009) and processed using the best-practices pipeline that includes marking of duplicate reads by the use of Picard tools (v1.83, <http://picard.sourceforge.net>), realignment around indels, and base recalibration via Genome Analysis Toolkit (GATK v3.2.2) (Van der Auwera et al., 2013).

Single nucleotide variants were called using GATK HaplotypeCaller, generating a single sample GVCF file. Batches of samples were jointly genotyped GATK GenotypeGVCFs to generate a multi-sample VCF. Variant Quality Score Recalibration (VQSR) was performed on the multi-sample VCF, and variants were annotated using VCFtools (Danacek et al., 2011) and in-house software.

Sequencing procedures for Complete Genomics, Inc. have been described in our prior publication reporting on a subset of these samples (Carmi et al., 2014) using procedures described previously (Drmanac et al., 2010; Roach et al., 2010). The average raw sequencing depth was 56x. The first 58 genomes were called using CG pipeline 2.0.2.26. All other genomes were called using pipeline 2.0.4.14. Both pipelines mapped variants to reference genome version hg19/GRCh37.

QC Pipeline: Rationale

In developing our AJ reference panel, we sought to maximize the amount of information obtainable from each sample, by employing full-depth sequencing as well as a quality control pipeline that attempted to optimize the tradeoff between sensitivity (avoidance of false negative variant calls) and specificity (avoidance of false positive calls). To date, most QC pipelines for full-depth next-generation sequencing have sought to identify causal variants for unusual phenotypes, thus prioritizing the minimization of false positives, at times sacrificing regions of the genome that are more difficult to call. Similarly, recent consensus efforts to create “gold standard” reference calls tend to be limited to the most readily sequenced portions of the genome (Zook et al., 2014; Zook et al., 2016). For example, it is widely acknowledged that low complexity regions (LCR), including homopolymers, short tandem repeats, and other repetitive elements comprise ~2% of the

genome and are relatively inaccessible to accurate sequencing from short-read technologies (Li, 2014).

Additionally, the 1000 Genomes Project divided the genomic space into three compartments (strict, pilot, and masked) based on observed read depth and mapping quality scores (1000 Genomes Project Consortium, 2015). Only 77% of the genome was included in the ‘strict’ (highest-quality) compartment, defined as those regions in which local depth of coverage remained within 50% of the genome-wide average depth, and no more than 0.1% of reads have mapping quality of zero. By contrast, 4% of the genome demonstrated extremes of coverage (high and low) and contained many (>20%) low quality reads, and was therefore considered ‘masked.’ The remaining ‘pilot’ regions (19% of the genome) remained of questionable quality.

In the present manuscript, we describe steps that greatly decrease false positives and substantially increase the fraction of the genome that is confidently called. Additionally, because our reference panel was sequenced on two disparate platforms (Complete Genomics, Inc. (CGI) and Illumina Hiseq X Ten, see Supplementary Figure 1), we describe an approach to reconcile data across platforms, extending previous work on this problem by including a larger number of replicate samples than previously described (Lam et al., 2011; Ross et al., 2013; Rieber et al., 2013; Wall et al., 2014). Finally, we identify new regions across the genome that are not currently well-mapped on either reference or alternate scaffolds, but which are routinely generated on the newest whole genome platforms, extending recent work on so-called blacklisted regions of the genome (Miga et al., 2015).

Platform-specific Filtering

The sensitivity-specificity tradeoff in detecting variants is a key step in large scale sequencing efforts, often requiring dataset-specific adjustments. To examine false positive rates, Mendelian errors for a trio included in the Illumina-sequenced batch were examined as a function of VQSLOD score and genomic compartment. Since the *de novo* mutation rate is $\sim 1.6 \times 10^{-8}$ per base pair per generation (Lipson et al., 2015; Palamara et al., 2015), this was considered negligible relative to the potential error rate. Additionally, we applied a genotype quality (GQ) filter of 20, and call rate threshold of >90%, as variants below these thresholds were greatly over-represented in the masked and LCR regions.

Cross-platform filtering

We filtered out ~5K SNVs with observed allele frequencies of >20% in the CGI dataset that were not observed at all in the larger Illumina dataset, as well as ~39K Illumina SNVs with allele frequency >20% that were not observed in either the CGI dataset or the 1000 Genomes database. We also filtered out 916 SNVs which were called in both CGI and Illumina but had an allele frequency difference larger than 20 percentage points, and an additional 483K SNPs not called in either Illumina batch or the HRC but contributes to majority of the imputation errors (Supplementary Table 2). We also filtered out SNP and INDELs past the Hardy-Weinberg threshold ($p < 10^{-10}$) in at least one platform, following the site filtering method used in constructing the HRC panel (McCarthy et al., 2016). Any multi-

allelic variants (Campbell et al., 2016) were also excluded after merging Illumina and CGI call sets before constructing the AJ reference imputation panel.

Imputation Evaluation

We considered the accuracy of imputation with our population-specific reference panel by comparing it to large, cosmopolitan reference panels in three ways. For each of these analyses, the non-reference discrepancy (NRD), both overall and for each genotype category (i.e., each of the possible combinations of reference (R) and alternate (A) alleles), were defined as below:

- Overall NRD = $(x_{RR} + x_{RA} + x_{AA}) / (x_{RR} + x_{RA} + x_{AA} + m_{RA} + m_{AA})$
- RR discordance = $x_{RR} / (x_{RR} + m_{RR})$
- RA discordance = $x_{RA} / (x_{RA} + m_{RA})$
- AA discordance = $x_{AA} / (x_{AA} + m_{AA})$
- R^2 is the coefficient of determination between imputed genotype dosage and the external standard of all evaluated samples
- the aggregated R^2 is calculated on each SNP and then averaged over all SNPs of each bin

First, we examined imputation of common (non-reference allele frequencies >5%) and infrequent variants (non-reference allele frequencies >1% but <5%) utilizing an independent test cohort of AJ subjects (n=2195, after removing 349 subjects overlapping with the sequencing cohort) with high quality genotypes available at ~1M SNPs assayed using the Illumina Omni-Quad platform, as described previously (Lencz et al., 2013). (Of course, it should be noted that some small percentage of genotyping errors exist in the array data.) A random subset of variants in the array data that were spread across the frequency spectrum were masked and held out to test concordance of imputed vs. masked genotypes. Specifically, we randomly masked and compared 30 SNP sites for each of the following allele frequency (AF) bins: 1~2%, 2~5%, 5~10%, 10~20%, 20~50% and 50~100%. Thus, the number of genotypes used for evaluation for each bin is $30 \times 2195 = 65,850$ per bin. Imputation was performed either locally, using SHAPEIT2 and Impute2, or using the HRC imputation server at Sanger (using SHAPEIT2 for pre-phasing and PBWT for imputation). Then genotype discordance and aggregated R^2 were calculated respectively using the SNP array genotypes as the gold standard on chromosome 20.

Second, we evaluated the imputation accuracy of all the sites available on both the constructed AJ panel and the HRC panel on chromosome 20, which represents a wider allele frequency spectrum than the SNP array comparison. The genotypes of 36 AJ individuals sequenced by CGI were used as a hold-out sample to calculate the non-reference genotype discordance vs the allele frequency. Notably, we also evaluated the imputed genotypes on variants private to the AJ panel (i.e., not called in the HRC panel) and found the overwhelming majority of false positive imputation errors were drawn from variants that were observed exclusively in the CGI samples but not supported by the larger set of Illumina-sequenced samples of the AJ reference panel. Consequently, as a final step in

construction and cleaning of the AJ reference panel, we eliminated all such variants exclusive to the CGI samples.

Last, we evaluated the imputation accuracy of the rarest variants in AJ panel, most of which are private in AJ population. To control the sequencing error at the minimum level, we evaluated all the variants with a non-reference allele count from 2 to 5 in Illumina sequencing samples (N=574) on chromosome 20. The genotypes of 36 AJ individuals sequenced by CGI were used as the target, hold-out sample. The absolute numbers of each comparison group are plotted for each of the 36 AJ individual. As expected, all of the imputed genotypes of the rarest variants are Ref/Ref and Ref/Alt. The false positive errors are represented by “impute_RA_RR” group which mean the true Ref/Ref genotypes were imputed to Ref/Alt, while the false negatives errors are represented by “impute_RR_RA” and “impute_RR_AA” groups.

Evaluation of Variant Filtering in Personal Genomes

We employed a leave-one-out approach, where we examined novel (i.e., not in dbSNP147) variants in each individual sample in relation to the remaining samples in the reference panel. In a clinical context, such variants may be labeled variants of uncertain significance (VUS).

Results

Enhancing call accuracy

After applying platform-specific filtering (see Methods), Mendelian errors in the Illumina-sequenced trio were reduced by nearly two orders of magnitude in both the pilot and strict non-LCR regions (Table 1a, cells marked in **bold**). Using default output from GATK (i.e., prior to our custom VQSLOD filtering), Mendelian error rates for single nucleotide variants (SNVs) were moderate (0.17%) in the strict, non-LCR region; however, across millions of variant sites, such a rate still implies several thousand false positives. Moreover, error rates were unacceptably high in all other compartments (Table 1a). Mendelian errors were reduced as VQSLOD filtering was increased, though in a non-linear fashion. As shown in Figure 1, a plateau was reached at VQSLOD=-2 for calls in the pilot region (red) and VQSLOD=-2.5 in the strict (blue) regions.

The relationship between errors and VQSLOD was different for indels, such that no plateau was observed; true and false positives were intermingled across the quality spectrum (Supplementary Figure 2). Consequently, we utilized a VQSLOD=0 threshold, which resulted in a 1% Mendelian error rate for indels in the pilot and strict (non-LCR) compartments. When combined with GQ and call rate thresholds as above, filtering resulted in an acceptable Mendelian error rate (Table 1b). For both SNVs and indels, no usable relationships were observed between VQSLOD and error rates in LCR regions, or in the masked compartment so variants in these regions (~6% of the whole genome) were excluded from further analyses.

CGI samples (n=168) were called individually using a proprietary pipeline from Complete Genomics, Inc.; similar quality metrics were therefore not available. (It should be noted that

n=36 of these CGI samples were held out for subsequent testing of the accuracy of the imputation reference panel.) False positives were identified by examination of nine samples run on both platforms, using Illumina data as the gold standard. Initial analysis of CGI data demonstrated that 7.99% of SNV calls were not observed in the Illumina calls for the same individuals. However, more than half (51.2%) of these false positives were observed to be singletons in the full CGI dataset. Consequently, we applied the following filters to the CGI data: 1) remove singletons; 2) remove masked and LCR regions; 3) remove variants called in <90% of CGI samples. Of the resulting SNVs calls, 97.2% were validated in the filtered Illumina data. Similar results were obtained in the CGI indel data; after applying the same filtering pipeline, however, only 91.5% were validated in the filtered Illumina data. Consequently, any indels which were only observed on the CGI platform (and never in the Illumina dataset) were filtered out.

Cross-Platform Merging

In total, we observed 17.6M variants in the filtered Illumina dataset, and around a half (~8.8M) of these in the smaller CGI dataset. Among the SNVs called in both Illumina and CGI data, virtually all (99.99%) had consistent allele frequencies across platform. Only 918 SNPs had allele frequency differences > 20 percentage points (Supplementary Figure 3), which would correspond to >6 standard deviations for a randomly sampled common variant. Outlying frequencies of Illumina variants that were inconsistent with their CGI frequencies were often small-integer fractions, suggestive of copy-number variants and motivating cross platform filtering (see Methods, and Supplementary Table 2, first three rows).

Notably, we found that high-frequency Illumina-only SNVs that were not observed in the 1000 Genomes database tended to cluster within specific chromosomal intervals, often (but not always) near the telomeres and centromeres (Supplementary Figure 4). One such region is displayed in Figure 2; in the top panel, a lengthy stretch of uncatalogued variants is observed in a 25kb segment within the gene *MAP2K3*, near the chromosome 17 centromere. As shown in the lower panels, mapping against the alternate “sponge sequence” developed by Miga and colleagues (23) and the “decoy” sequence proposed by the 1000 Genomes Consortium (3) re-assigns these reads, such that these variants are no longer mapped within this gene sequence. However, not all such regions were completely cleared using either of these published alternative scaffolds, nor using the hg38 alignment. We identified 90 regions that harbor high-frequency, uncatalogued runs of variants as listed in Supplementary Table 3. We compared these regions to the segmental duplication bed file defined by the Global Alliance for Genomic Health (GA4GH, Lawler et al. 2015). Only 19% of the regions observed in current data set have at least 50% overlapping bases with the GA4GH segmental duplication bed file, and only 30% of the 90 mis-mapped regions overlap at least 20% with the GA4GH file. Thus, a large portion of mis-mapped regions still cannot be well characterized, probably due to population specific duplications and other events.

Copy-number variation is a plausible source of these anomalies, given that these regions are marked by individuals carrying derived alleles, called heterozygous at multiple neighboring positions. However, the boundaries of these regions do not precisely track boundaries of known copy-number variants in DGV. For purposes of developing the imputation reference,

SNPs in these regions were filtered using a loose Hardy-Weinberg threshold ($p < 10^{-10}$), consistent with the site filtering method used in constructing the HRC panel (McCarthy et al., 2016), yielding a final set of 17.5M SNVs in our combined reference dataset (total $N=738$).

Imputation Performance

We compared the contribution of our newly constructed population-specific reference panel to the accuracy of imputing common and infrequent alleles in our sample of $n=2195$ Ashkenazi subjects genotyped on the Omni-Quad chip array (see Methods). As shown in Table 2, we observe discordance rates to be nearly an order of magnitude smaller when using the Ashkenazi reference sequences as compared to the cosmopolitan 1000 Genomes panel and the European-specific UK10K panel. Performance of the AJ panel is ~2-fold better than the cosmopolitan HRC panel, despite a difference in sample sizes that is two orders of magnitude. Improved performance of the AJ-specific panel is especially meaningful at more rare alleles (Figure 3), where imputation using a cosmopolitan panel is often subpar. We observed slight improvement when we combined the AJ-specific panel with the 1000 Genomes panel.

We also evaluated all overlapping sites between AJ panel and HRC panel on chromosome 20 using a sample of 36 AJ individuals sequenced on the CGI platform as the hold-out sample (Methods). Overall, the non-reference genotype discordance rates of AJ panel and HRC panel are 0.32% and 0.97% respectively, which corresponds to a 3-fold improvement in AJ panel over HRC panel (Table 3, top two rows). This evaluation includes a much larger number of variants and a wider spectrum of allele frequency compared to the SNP array analysis; specifically, the number of available sites for each bin ranged from ~200 to ~30,000. (The numbers of genotypes are roughly equal to the number of SNP sites multiplied by 36 samples.). The non-reference genotype discordance rates were then plotted versus allele frequency bins (Figure 4). The AJ panel outperformed the HRC panel across all allele frequency bins above 0.145% (i.e., allele count=2 in the AJ reference panel, Figure 4), despite the HRC panel having a much larger sample size and minimum allele counts of 5 for all the overlapping variants.

Given the biological importance of rare variants, we also evaluated the imputation accuracy of the rarest variants in our constructed AJ panel, which are also most likely to be private in AJ population and not called in HRC panel (Methods). For variants observed at 0.36% frequency (allele count = 5) in the AJ reference cohort, many individual samples show false positive and negative rates of 0% (Supplementary Figure 5, top panel). Notably, a larger proportion of imputed variants received a no-call in the CGI sequence data, due to its known conservative bias (Lam et al., 2011; Rieber et al., 2013). Moreover, the overwhelming majority of false positive imputation calls were drawn from variants that were observed exclusively in the CGI samples of the reference panel, and were not present in the Illumina-sequenced samples of the AJ reference panel or in the HRC panel (Table 3, bottom three rows). Consequently, as a final step in construction and cleaning of the AJ reference panel, we eliminated all such variants exclusive to the CGI samples (Supplementary Table 2,

bottom row). Methods and results for all imputation experiments are summarized in Supplementary Table 4.

Population-specific variant discovery

Finally, we sought to evaluate the potential utility of the population-specific reference panel as a catalogue of normal variation in the population, applicable to the clinical interpretation of personal genomes (see Methods). Filtering against the remaining reference panel removed nearly 80% of novel variants across the entire genome (Figure 5a). When considering only novel functional (coding or splicing) variants, the population specific panel improves filtering significantly ($p < 1.1 \times 10^{-10}$), but less dramatically, reducing the median count of such variants from 10 to 8 (Figure 5b).

Relatedly, we conducted an evaluation of the ability of our reference panel to contribute novel variation relative to the largest available cosmopolitan database, gnomAD. Specifically, we exploited the substantial difference in sample size for exomes relative to whole genomes (~8:1) in gnomAD to examine the effects of sampling depth on the utility of adding a new, population-specific panel. By categorizing alleles that were novel to our dataset (i.e., not observed in gnomAD) based on location (exonic vs, non-exonic) and frequency, we observed that novel alleles observed as singletons or doubletons in our panel were observed at similar rates for both exonic and non-exonic positions (approximately 1/650kb per sample), suggesting that this space is far from saturation even in the well-populated exonic component of gnomAD (nearly 125K samples). Moreover, a number of novel alleles were observed in our dataset at >1% frequency, although these tended to be more frequently observed in the non-exonic compartment (1.3x more likely to be non-exonic for rare alleles (1–2% frequency) and 2x more likely to be non-exonic for more common alleles (5–10% frequency). These results are consistent with the computational estimation that ExAC/gnomAD would be far from saturated even after attaining 500K samples (Zou et al. 2016).

Discussion

Our primary results (Tables 1 and 3; Figures 3 and 4) demonstrate enhanced accuracy of an imputation panel specific to the AJ population, relative to cosmopolitan panels, especially in the range of infrequent (non-reference allele frequencies >1% but <5%) and rare (non-reference allele frequency <1%) variants that may be most critical to further progress in mapping complex phenotypes (Heinzen et al., 2015; Surakka et al., 2015). Our results represent a substantial improvement over our prior report (Carmi et al. 2014), using a much smaller AJ panel (n=128). In our prior paper, the aggregated R^2 of the rarest bins (1% and 2% allele frequency) were 0.86 and 0.91, respectively (Figure 2c, Carmi et al. 2014); the corresponding result in the present report (Figure 3) demonstrates $R^2=0.97$ for variants in the 1–2% frequency range. These results also extend prior studies that have shown the superiority of combining population-specific with cosmopolitan panels (Wong et al., 2013; Pistis et al., 2015; Hou et al., 2017; Mitt et al., 2017), by demonstrating: 1) a moderately-sized population-specific reference sample sequenced to full-depth provides better performance than the newly-released Haplotype Reference Consortium panel; and 2)

addition of a cosmopolitan panel (e.g., 1000 Genomes) to such a population-specific panel provides only marginal improvement in performance, consistent with recent findings in an outbred population cohort (Nagasaki et al., 2015). As recently demonstrated in non-human species, the greatest predictor of individual imputation accuracy is the genomic relatedness between the target and the reference population (Larmer et al. 2017). In this context, it should be emphasized that the AJ is an unusual population, marked by a narrow bottleneck and a tradition of endogamy (Atzmon et al. 2010; Palamara et al. 2012).

Moreover, for clinical purposes of interpreting a personal genome, our population-specific panel significantly enhanced filtering of variants unlikely to be related to disease (Figure 5). However, there is still considerable room for additional filtering in the coding regions (Figure 5b). As the exome is under greater pressure from negative selection compared to the rest of the genome, it is likely that many exonic variants are of relatively recent origin (i.e., post-dating the bottleneck in AJ history). Given the rapidly expanding nature of the human population in recent centuries (including the post-bottleneck AJ population), exceptionally large samples will be needed to achieve asymptotic representation of background variation (Lohmueller, 2014; Zou et al., 2016). In the applied setting, sequencing of parents may be the most efficient strategy for filtering and variant interpretation (Nagy et al., 2015), supplemented by large-scale sequencing resources from the general population (Lek et al., 2016) and moderate-scale population-specific resources such as those described here.

While providing benefits such as those described above, short-read sequencing technologies continue to have technical challenges and limitations that we have sought to address in the present study. While many studies (Highnam et al., 2015; Laehnemann et al., 2016) have compared accuracy of different alignment and mapping protocols for raw short-read data, such comparisons are computationally expensive and may be impractical for application to large population cohorts and unavailable in the clinical setting. Many end-users of sequencing data receive batched calls from sequencing centers using standard workflows using the Genome Analysis Toolkit (GATK (Van der Auwera et al., 2013)). In this context, the present report provides practical guidelines to filtering for such users (both research and clinical). Our approach yields high accuracy not only for the most stably called regions, but also the so-called “pilot” region, extending high accuracy variant calling to 94% of the genome. This compares favorably to recent approaches such as the Genome in a Bottle (Zook et al., 2014; Zook et al., 2016) and ReliableGenome (Popitsch et al., 2017), which primarily focus on the optimizing accuracy of the ~70% of the genome that is least susceptible to technical artifact in short-read data.

The potential clinical importance of expanding the range of the genome that can be called reliably is illustrated in Supplementary Figure 6. In this figure, we parse the genes designated by the American College of Medical Genetics and Genomics (ACMG) as harboring clinically actionable variants (Green et al., 2013) as a function of proportion of variants observed in each calling compartment (strict/pilot/masked/LCR) as designated by the 1000 Genomes Project. Each of these genes contains segments that fall in the pilot compartment (Supplementary Figure 6a), and several genes (such as *PMS2*, *SDHC*, and *SDHD*) contain up to 50% of exonic bases designated as pilot (Supplementary Figure 6b).

Thus, a clinical readout that is unable to accurately capture these bases would be relatively incomplete.

Several limitations in the short read data were difficult to overcome. For example, while we were able to produce reasonable error rates for indels even within the pilot compartment, these required more careful filtering, and do not have a readily identifiable optimum for balancing false positives vs. false negatives (Supplementary Figure 2). Additionally, we observed that the CGI platform suffers from two significant limitations: it is generally overly conservative (fewer total calls, more non-calls) compared to Illumina, but at the same time it is susceptible to a relatively large number of platform-specific false positives. While this platform is no longer active, legacy datasets should be treated with caution for novel variants, although known variants are conservatively called. Additionally, we have identified regions of the genome that are not well mapped in current reference or alternate assemblies (Figure 2; Supplementary Figure 4). Further work is needed to properly characterize the (likely) structural variations that underlie these anomalous segments (Church et al., 2015).

Finally, we should note that many of our samples were derived from patients with schizophrenia. However, it is unlikely that this fact negatively impacted our imputation results, given that: 1) rare variant enrichment in patients with schizophrenia is <1 variant per genome (Genovese et al. 2016); 2) there is no evidence that disease-associated rare variants have atypical recombination patterns; 3) no existing reference panels screen for disease, and the Haplotype Reference Consortium specifically includes samples derived from large disease cohorts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors are extremely grateful to Soren Germer, Ph.D. and his team at the New York Genome Center for performing the Illumina sequencing. We acknowledge financial support from the Human Frontier Science Program (SC); NIH research grants AG042188 (GA), DK62429, DK062422, DK092235 (JHC), NS050487, NS060113 (LNC), AG021654, AG027734 (NB), MH089964, MH095458, MH084098 (TL), and CA121852 (computational infrastructure, IPe'er); NSF research grants 08929882 and 0845677 (IPe'er); Rachel and Lewis Rudin Foundation (HE); Northwell Health Foundation (TL); Brain & Behavior Foundation (TL); US-Israel Binational Science Foundation (TL, AD); LUNGevity Foundation (ZHG); New York Crohn's Disease Foundation (IPeter); Edwin & Caroline Levy and Joseph & Carol Reich (SB); the Parkinson's Disease Foundation (LNC); the Sharon Levine Corzine Cancer Research Fund (KO); and the Andrew Sabin Family Research Fund (KO).

References

- 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015 A global reference for human genetic variation. *Nature* 526:68–74. [PubMed: 26432245]
- Ankala A, Tamhankar PM, Valencia CA, Rayam KK, Kumar MM, Hegde MR. 2015 Clinical applications and implications of common and founder mutations in Indian subpopulations. *Hum Mutat* 36:1–10. [PubMed: 25323826]
- Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E, Ostrer H. 2010 Abraham's children in the genome era: major Jewish diaspora populations

comprise distinct genetic clusters with shared Middle Eastern Ancestry. *Am J Hum Genet.* 86(6): 850–9. [PubMed: 20560205]

- Campbell IM, Gambin T, Jhangiani S, Grove ML, Veeraraghavan N, Muzny DM, Shaw CA, Gibbs RA, Boerwinkle E, Yu F, Lupski JR. 2016 Multiallelic positions in the human genome: Challenges for genetic analyses. *Hum Mutat* 37:231–234. [PubMed: 26670213]
- Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, Guha S, Upadhyay K, Ben-Avraham D, Mukherjee S, et al. 2104. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun* 5:4835.
- Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin CS, Kitts PA, Aken B, Marth GT, Hoffman MM, Herrero J, Mendoza ML, Durbin R, Flicek P. 2015 Extending reference assembly models. *Genome Biol* 16:13. [PubMed: 25651527]
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. 2011 The variant call format and VCFtools. *Bioinformatics* 27:2156–8. [PubMed: 21653522]
- Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K, Kreiner-Møller E, Rivadeneira F, et al. 2014 Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’. *Eur J Hum Genet* 22:1321–6. [PubMed: 24896149]
- Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, Merker JD, Goldfeder RL, Enns GM, David SP, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA* 311:1035–45.
- Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010 Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81. [PubMed: 19892942]
- Druet T, Macleod IM, Hayes BJ. 2014 Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* (Edinb). 112(1):39–47. [PubMed: 23549338]
- Genome of the Netherlands Consortium. 2014 Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 46:818–25. [PubMed: 24974849]
- Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landén M, Moran JL, Purcell SM, Sklar P, Sullivan PF, Hultman CM, McCarroll SA. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci.* 2016 11;19(11):1433–1441. [PubMed: 27694994]
- Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O’Daniel JM, Ormond KE, Rehm HL, Watson MS, et al. 2013 ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 15:565–74. [PubMed: 23788249]
- Guha S, Rosenfeld JA, Malhotra AK, Lee AT, Gregersen PK, Kane JM, Pe’er I, Darvasi A, Lencz T. 2012 Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. *Genome Biol.* 13(1):R2. [PubMed: 22277159]
- Heinzen EL, Neale BM, Traynelis SF, Allen AS, Goldstein DB. 2015 The genetics of neuropsychiatric diseases: looking in and beyond the exome. *Annu Rev Neurosci* 38:47–68. [PubMed: 25840007]
- Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D. 2015 An analytical framework for optimizing variant discovery from personal genomes. *Nat Commun* 6:6275 [PubMed: 25711446]
- Hoffmann TJ, Witte JS. 2015 Strategies for imputing and analyzing rare variants in association studies. *Trends Genet* 31:556–63. [PubMed: 26450338]
- Hou L, Kember RL, Roach JC, O’Connell JR, Craig DW, Bucan M, Scott WK, Pericak-Vance M, Haines JL, Crawford MH, Shuldiner AR, McMahon FJ. 2017 A population-specific reference panel empowers genetic studies of Anabaptist populations. *Sci Rep* 7:6079. [PubMed: 28729679]
- Iglesias AI, van der Lee SJ, Bonnemaier PWM, Höhn R, Nag A, Gharahkhani P, Khawaja AP, Broer L; International Glaucoma Genetics Consortium (IGGC), Foster PJ, Hammond CJ, Hysi PG, et al. 2017 Haplotype reference consortium panel: Practical implications of imputations with large reference panels. *Hum Mutat* 38:1025–1032. [PubMed: 28493391]

- Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, et al. 2012 Exome sequencing and the genetic basis of complex traits. *Nat Genet* 44:623–30. [PubMed: 22641211]
- Laehnmann D, Borkhardt A, McHardy AC. 2016 Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform* 17:154–79. [PubMed: 26026159]
- Lam HY, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, et al. 2011 Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* 30:78–82. [PubMed: 22178993]
- Larmer SG, Sargolzaei M, Brito LF, Ventura RV, Schenkel FS. 2017 Novel methods for genotype imputation to whole-genome sequence and a simple linear model to predict imputation accuracy. *BMC Genet.* 18(1):120. [PubMed: 29281958]
- Lawler M, Siu LL, Rehm HL, Chanock SJ, Alterovitz G, Burn J, Calvo F, Lacombe D, Teh BT, North KN, Sawyers CL; Clinical Working Group of the Global Alliance for Genomics and Health (GA4GH). All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health. *Cancer Discov.* 2015 11;5(11):1133–6. [PubMed: 26526696]
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016 Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–91 [PubMed: 27535533]
- Lencz T, Guha S, Liu C, Rosenfeld J, Mukherjee S, DeRosse P, John M, Cheng L, Zhang C, Badner JA, et al. 2013 Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat Commun* 4:2739. [PubMed: 24253340]
- Li H 2014 Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30: 2843–51. [PubMed: 24974202]
- Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–60. [PubMed: 19451168]
- Lipson M, Loh PR, Sankararaman S, Patterson N, Berger B, Reich D. 2015 Calibrating the human mutation rate via ancestral recombination density in diploid genomes. *PLoS Genet* 11:e1005550. [PubMed: 26562831]
- Lohmueller KE. 2014 The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet* 10:e1004379. [PubMed: 24875776]
- MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, et al. 2014 Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469–76. [PubMed: 24759409]
- McCarthy S, Das S, Kretschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, et al. 2016 A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 48:1279–83. [PubMed: 27548312]
- Miga KH, Eisenhart C, Kent WJ. 2015 Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* 43:e133. [PubMed: 26163063]
- Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, Ripatti S, Morris AP, Metspalu A, Esko T, Mägi R, Palta P. 2017 Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* 25:869–876. [PubMed: 28401899]
- Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, et al. 2015 Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* 6:8018. [PubMed: 26292667]
- Nagy PL, Mansukhani M. 2015 The role of clinical genomic testing in diagnosis and discovery of pathogenic mutations. *Expert Rev Mol Diagn* 15:1101–5. [PubMed: 26202666]
- Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, Finucane HK, Sankararaman S; Genome of the Netherlands Consortium, Sunyaev SR, de Bakker PI, Wakeley J, Pe'er I, Price AL. 2015 Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am J Hum Genet* 97:775–89. [PubMed: 26581902]

- Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012 Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 91:809–22. [PubMed: 23103233]
- Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A, et al. 2015 Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* 23:975–83. [PubMed: 25293720]
- Popitsch N; WGS500 Consortium, Schuh A, Taylor JC. 2017 ReliableGenome: annotation of genomic regions with high/low variant calling concordance. *Bioinformatics* 33:155–160. [PubMed: 27605105]
- Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P, et al. 2013 Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS One* 8:e66621. [PubMed: 23776689]
- Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010 Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639. [PubMed: 20220176]
- Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013 Characterizing and measuring bias in sequence data. *Genome Biol* 14:R51. [PubMed: 23718773]
- Surakka I, Horikoshi M, Mägi R, Sarin AP, Mahajan A, Lagou V, Marullo L, Ferreira T, Miraglio B, Timonen S, et al. 2015 The impact of low-frequency and rare variants on lipid levels. *Nat Genet* 47:589–97. [PubMed: 25961943]
- UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, et al. 2015 The UK10K project identifies rare variants in health and disease. *Nature* 526:82–90. [PubMed: 26367797]
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, et al. 2013 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*.43:1110.1–33. [PubMed: 25431634]
- Ventura RV, Lu D, Schenkel FS, Wang Z, Li C, Miller SP. 2014 Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. *J Anim Sci*. 92(4):1433–44. [PubMed: 24663187]
- Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, Risch N. 2014 Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res* 24:1734–9. [PubMed: 25304867]
- Walsh R, Thomson KL, Ware JS, Funke BH, Woodley J, McGuire KJ, Mazzarotto F, Blair E, Seller A, Taylor JC, et al. 2017 Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* 19:192–203. [PubMed: 27532257]
- Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, Barton PJR, Funke B, Cook SA, MacArthur D, Ware JS. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med*. 2017 10;19(10):1151–1158. [PubMed: 28518168]
- Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, Lam KK, Pillai NE, Sim KS, Xu H, et al. 2013 Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet* 92:52–66. [PubMed: 23290073]
- Zhang P, Zhan X, Rosenberg NA, Zöllner S. 2013 Genotype imputation reference panel selection using maximal phylogenetic diversity. *Genetics* 195:319–30. [PubMed: 23934887]
- Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014 Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 32:246–51. [PubMed: 24531798]
- Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016 Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 3:160025. [PubMed: 27271295]
- Zou J, Valiant G, Valiant P, Karczewski K, Chan SO, Samocha K, Lek M, Sunyaev S, Daly M, MacArthur DG. 2016 Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nat Commun* 7:13293. [PubMed: 27796292]

Figure 1. Observed Mendelian error rate (y-axis) as a function of reported quality (VQSLOD score, x-axis) for SNPs in non-LCR regions labeled as strict (blue) or pilot (red) calls. Both curves show rapid drop in error rate before a plateau starting at VQSLOD of -2 .

Figure 2.

The utility of mapping to multiple reference sequences. Read-browser results of mapping for a 25kb segment along chromosome 17. Color identifies supposedly identified novel variants. The four panels from the top to bottom depict: 1) mapping against standard hg19/GRCh37 reference; 2) mapping against hg19/GRCh37 + “sponge” sequence of Miga et al. (23); 3; mapping against hg19/GRCh37 + decoy sequences derived from the 1000 Genomes Project; 4; mapping against hg19/GRCh37 + sponge sequence + decoy sequence. Note that the two additional sets of reference sequences remove likely false-positive variants in different regions and combining them together could remove the most.

Figure 3.

Imputation accuracy (aggregated R^2 ; y-axis) across non-reference allele frequencies (x-axis) at held-out SNP-array sites in a genotyped panel of 2195 AJ individuals. Imputation quality is high for common alleles, but rarer ones were imputed markedly better using AJ references samples (AJ only – green, AJ combined with the 1000 genomes, navy blue).

Figure 4.

Non-reference genotype discordance (y-axis) across non-reference allele frequencies (x-axis) at overlapping variant sites between AJ reference panel and HRC reference panel in a sample of 36 AJ individuals sequenced on the CGI platform. Note that the non-reference allele counts/frequencies were calculated in the AJ panel.

Figure 5. Variants in an out-of-sample AJ personal genome deemed novel before (blue) and after (red) inclusion of the dataset reported in this study. Data is presented for the median held-out sample (errorbars: 90%-CI), for all variants genome-wide (panel a) or only functional (exonic/splicing) ones (panel b).

Table 1

Mendelian Errors by Compartment and Filtering

<i>Repeat Region</i>	LCR			Non-LCR		
<i>Compartment</i>	Masked	Pilot	Strict	Masked	Pilot	Strict
<i>Unfiltered</i>	14.66%	14.72%	2.83%	7.89%	2.57%	0.17%
<i>Post-Filter</i>	N/A				0.0415%	0.0049%
Table 1a Mendelian Error Rates for SNVs						
<i>Repeat Region</i>	LCR			Non-LCR		
<i>Compartment</i>	Masked	Pilot	Strict	Masked	Pilot	Strict
<i>Unfiltered</i>	10.80%	10.75%	14.77%	10.23%	3.54%	1.51%
<i>Post-Filter</i>	N/A				0.1756%	0.0829%
Table 1b Mendelian Error Rates for Indels						

Table 2

Imputation accuracy

Panel	Pipeline	Ref/Ref	Ref/Alt	Alt/Alt	NRD
1000 Genomes	Local	0.44	1.94	1.43	2.69
1000 Genomes	Server	0.69	3.16	2.17	4.29
UK10K	Server	0.60	3.29	2.23	4.14
1000G + UK10K	Server	0.48	2.63	2.02	3.40
HRC	Server	0.19	0.85	0.58	1.15
AJ sequencing	Local	0.08	0.26	0.30	0.47
AJ + 1000G	Local	0.06	0.34	0.17	0.40

Table 3

Imputation accuracy for CGI hold-out samples

	NRD (%)	RR(%)	RA->RR	AA->RR	RA(%)	RR->RA	AA->RA	AA(%)	RR->AA	RA->AA
AJ overlap	0.32	0.06	0.06	0.00	0.23	0.16	0.07	0.13	0.00	0.13
HRC overlap	0.97	0.22	0.21	0.00	0.62	0.45	0.16	0.36	0.01	0.35
HRC only	2.76	0.21	0.21	0.00	3.97	3.67	0.30	0.24	0.02	0.22
AJ only (CGI and/or Illumina)	13.66	0.13	0.13	0.00	17.63	17.48	0.15	0.78	0.48	0.30
AJ only - CGI only	96.95	0.36	0.35	0.01	96.67	96.64	0.04	98.75	97.18	1.41
AJ only - Illumina only	1.47	0.05	0.05	0.00	1.48	1.30	0.18	0.30	0.01	0.30