A population-specific reference panel for Ashkenazi Jews improves power to detect associations with rare functional variants

Aaron Kleinman, Steve Pitts, Vladimir Vacic, Robert Gentleman

23andMe, Inc., Mountain View, CA, USA

Abstract

The existence of publicly-available pan-ethnic reference panels facilitate imputation of common and low-frequency variants [1], but rare variants remain challenging to study. Population-specific reference panels can help improve imputation accuracy [2], and access to the cohort to be imputed can help inform sample selection [3].

Starting with a genotyped cohort of 40,000 Ashkenazis, we selected 481 unrelated individuals for whole genome sequencing. Selectees were chosen to maximize the amount of population haplotypic diversity (**Figure 2**). After discarding singletons, the ensuing reference panel had 15.7M autosomal SNPs and 4.4M autosomal indels (**Figure 3).** We imputed the sequence data into the remaining genotyped samples, achieving superior imputation accuracy compared with larger but more cosmopolitan reference panels **(Figure 4**). We found associations with low-frequency functional variants (**Figure 5**), and cataloged genes with complete human knockouts.

Methodology

Data

All participants were drawn from the customer base of 23andMe, Inc., a consumer personal genetics company. Customers were genotyped on either an Illumina HumanHap550 BeadChip platform or a custom Illumina HumanOmniExpress-24 format chip. Self-reported phenotypes were collected from web-based surveys.

We began with a cohort of 40,000 participants whose genomes we classified as > 80% Ashkenazi. We selected 481 unrelated samples for sequencing according to a computationally efficient greedy algorithm that seeks to maximize the total amount of population genome that is in IBD with one or more of the sequenced individuals **(Figure 2)**. The chosen samples were sequenced on an Illumina HiSeq X to a median depth of 28x.

Analysis

We discarded samples that failed to meet QC, called variants on the remaining data with GATK HaplotypeCaller, selected a VQSR threshold based on the rate of Mendelian inconsistencies, and removed singletons in the reference panel because they are difficult to phase. This left us with 21.1M variants (Figure 3). We phased and imputed missing sites in the panel using BEAGLE, and functionally annotated variants using VEP/LOFTEE. We imputed the sequence variants into 48,916 additional chip-genotyped and phased Ashkenazis. These participants' genotypes were phased using an in-house out-of-sample version of Beagle, and imputation was performed using Minimac3.

Results

The constructed reference panel gave improved accuracy over more cosmopolitan panels (**Figure 4**). We ran association tests on the imputed cohort against an array of phenotypes, and found associations with low-frequency variants that would be normally difficult to impute (**Figure 5).**

We then looked explicitly at LOFs. We found 2,503 loss-of-function variants (855 SNPs, 1,648 indels) in 1,768 genes, 394 of which had more than one. Most of the variants are rare, making the power to study any individual one a challenge. To increase our specificity, we restricted ourselves to variants with MAF < 2% and r^2 > 0.8, and looked for human homozygotes. There are 2,178 individuals (4.4%) who are carrying at least one knockout.

We calculated transmission probabilities for these LOFs in 1,736 offspring. In comparison to the expectation under mendelian inheritance (25%), we observed a double-transmission probability of 20.3%, or 466 double transmissions of the minor allele per 10,000, transmissions from a pair of heterozygous parents (p=0.062).

Discussion

Population-specific reference panels allow improved imputation of rarer variants. We constructed one such panel for Ashkenazis, because their population history makes them an ideal group for this kind of study. Because we had access to the genotyped cohort before selecting samples for sequencing, we were able to maximize the amount of population haplotypes sequenced. We showed that this panel improves performance over a larger but more cosmopolitan panel. We discovered novel associations with rare functional variants, and recapitulated known results about the transmission deficit of LOFs. These results, which we expect to replicate and sharpen with larger cohorts, show one powerful way to bring sequencing to bear on a large genotyped cohort.

Acknowledgments

We thank 23andMe customers who consented to participate in research for enabling this study. We also thank employees of 23andMe who contributed to the development of the infrastructure that made this research possible.

References

[1] The 1000 Genomes Project Consortium. "A global reference for human genetic variation." Nature 526, 68-74 (2015). doi:10.1038/nature15393.

[2] M Ahmad et al. "Inclusion of Population-specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy." Sci Rep. 2017; 7: 6733. doi: 10.1038/s41598-017-06905-6

[3] A Gusev et. al. "Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population." Genetics. 2012 Feb;190(2):679-89. doi: 10.1534/genetics.

[4] P Sulem et al. "Identification of a large set of rare complete human knockouts." Nature Genetics, 47, 448-452 (2015). doi:10.1038/ng.3243

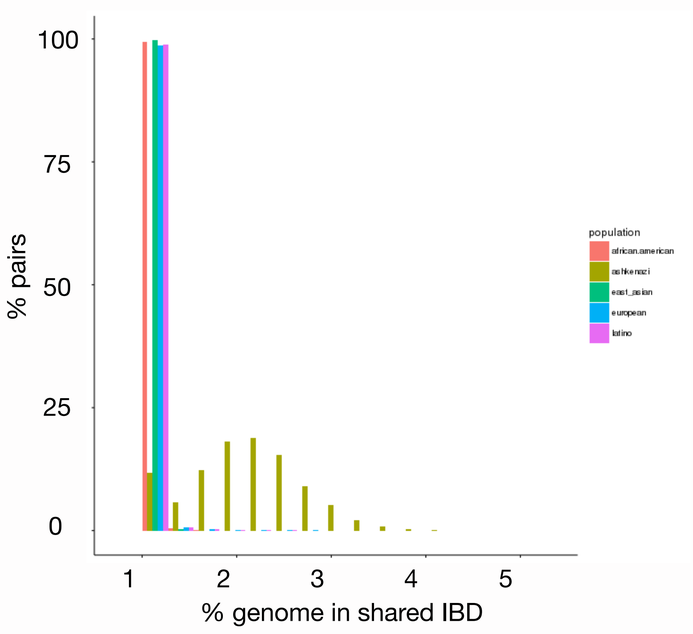

Figure 1: Ashkenazis exhibit pervasive IBD sharing
We selected 10,000 random pairs of participants whose genome we classified as > 80% Ashkenazi. For each pair, we used a modified version of Germline to find the total IBD shared in segments of length > 4cM. We repeated the process for pairs of individuals selected from African American, Latino, East Asian and European populations. This figure shows the per-population frequency distribution of total IBD shared for pairs of individuals. Only Ashkenazis exhibit pervasive IBD sharing.

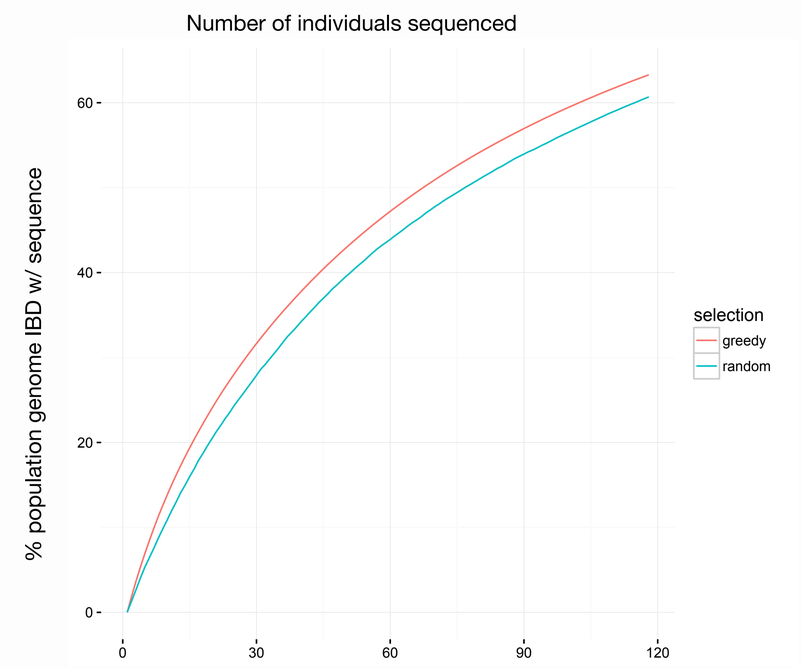

Figure 2. Greedy selection improves haplotype coverage
Since we had access to the target cohort before sequencing, we could select samples to maximize imputation quality. We defined this to be the proportion of population haplotypes present in the sequenced samples.

More formally, given an individual **i** and a set of individuals **A**, let **f(i,A)** be the proportion of the genome of **i** that is in IBD with at least one individual in **A**. The basic algorithmic question is as follows: given a cohort **C** of size **m** and a fixed budget to sequence **n < m** people, find the subset **X** of size **n** that maximizes $\sum_{i \in C} f(i, X)$.

This problem is NP-hard, so we used a greedy approximation algorithm to iteratively construct X, adding at each step the individual who would most increase the sum. This is O(mn) in theory but faster in practice, since it is not always necessary to examine each unsequenced individual before finding the next best one.

This resulted in a set of 481 unrelated individuals who covered an average of ~3,040 cM (89%) of the autosome of each non-sequenced individual **(Figure 2).** In contrast, a random selection algorithm would have required an average of 510 individuals to achieve the same coverage.

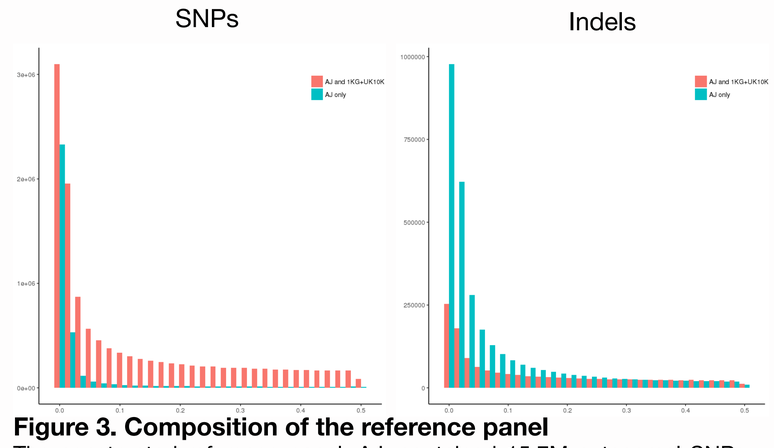

Figure 3. Composition of the reference panel
The constructed reference panel, AJ, contained 15.7M autosomal SNPs. Compared to a reference panel comprised of merging 1000 Genomes Phase 3 + UK10K (1KG+UK10K), 3.4M (22%) of these SNPs were novel, and almost all had an MAF < 5%. The story is different for indels. AJ had 4.4M autosomal indels, of which 3.1M (70%) were not in 1KG. Though the majority are low-frequency, we found new indels all across the frequency spectrum.

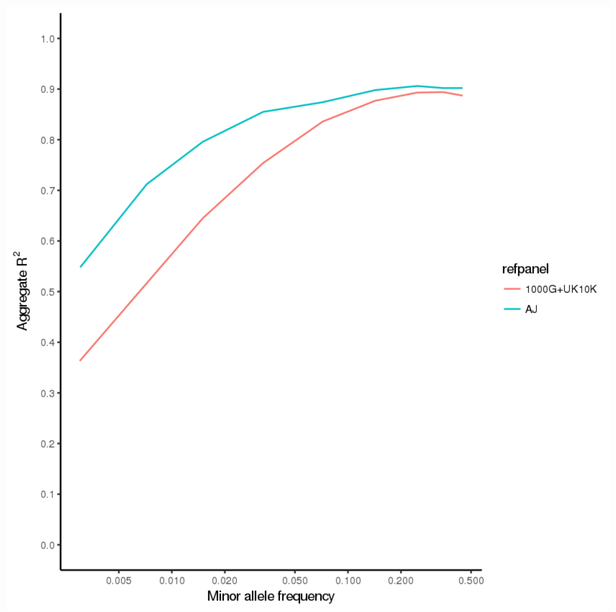

Figure 4: Imputation accuracy
We phased the reference panel using Beagle. We prephased the genotyped participants using a modified out-of-sample version of Beagle, and imputed the sequence data using Minimac3. We assessed accuracy using 150 WGS individuals of Ashkenazi ancestry who were not included in the reference panel. This figure shows imputation accuracy as a function of MAF, for both the AJ panel and the 1000G+UK10K panel. As expected, the AJ panel performs best, especially at lower frequencies.

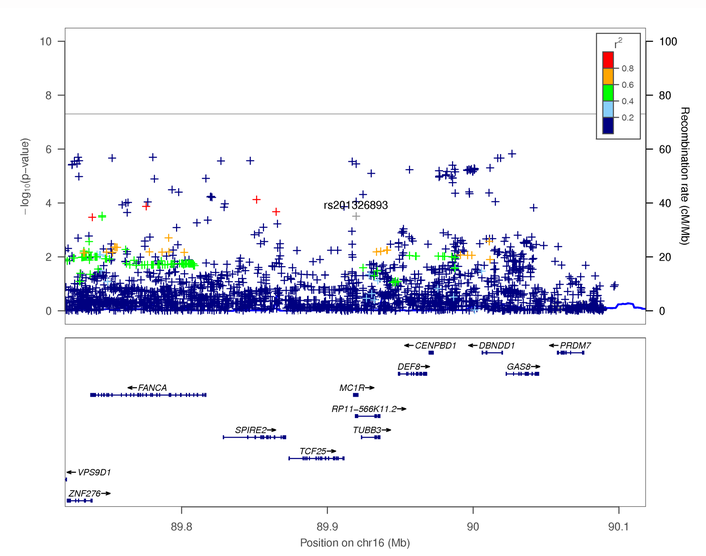

Figure 5. rs201326893 associations with melanoma
We ran GWAS on the imputed cohort against an array of self-reported phenotypes, paying particular attention to association tests with rare functional variants that impute well in AJ but not in pan-ethnic reference panels. As an example, rs201326893 is a stop-gain variant in the melanocortin 1 receptor (MC1R), a gene that plays an important role in melanogenesis. rs201326893 has an MAF of 0.1% in Europeans (ExAC), but 1% in Ashkenazis. It imputes well in AJ (estimated r^2=0.99). In a GWAS on melanoma (ncase=943, ncontrol=31326), controlling for age, sex and the first five pc's, it associates p=0.0003, OR=2.0), though a larger sample size is needed before it surpasses significance after correction.