frontiers in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.719791

Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes

Andrés Jiménez-Kaufmann[1], Amanda Y. Chong[2], Adrián Cortés[2], Consuelo D. Quinto-Cortés[1], Selene L. Fernandez-Valverde[1], Leticia Ferreyra-Reyes[3], Luis Pablo Cruz-Hervert[3], Santiago G. Medina-Muñoz[1], Mashaal Sohail[1,4], María J. Palma-Martinez[1], Gudalupe Delgado-Sánchez[3], Norma Mongua-Rodríguez[3], Alexander J. Mentzer[2], Adrian V. S. Hill[2,5], Hortensia Moreno-Macías[6,7], Alicia Huerta-Chagoya[6], Carlos A. Aguilar-Salinas[8,9], Michael Torres[1], Hie Lim Kim[10,11,12], Namrata Kalsi[10,11], Stephan C. Schuster[10,11,12], Teresa Tusié-Luna[6,13], Diego Ortega Del-Vecchyo[14], Lourdes García-García[3] and Andrés Moreno-Estrada[1]**

[1]Laboratorio Nacional de Genómica para la Biodiversidad (UGA-LANGEBIO), Unidad de Genómica Avanzada, Irapuato, Mexico, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, [3]Instituto Nacional de Salud Pública, Cuernavaca, Mexico, [4]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, [5]Nuffield Department of Medicine, The Jenner Institute, University of Oxford, Oxford, United Kingdom, [6]Unidad de Biología Molecular y Medicina Genómica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ), Mexico City, Mexico, [7]Departamento de Economía, Universidad Autónoma Metropolitana, Mexico City, Mexico, [8]Departamento de Endocrinología y Metabolismo, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Unidad de Investigación de Enfermedades Metabólicas, Mexico City, Mexico, [9]Tecnológico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Monterrey, Mexico, [10]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, [11]GenomeAsia 100K (GA100K) Consortium, Singapore, [12]School of Biological Science, Nanyang Technological University, Singapore, [13]Instituto de Investigaciones Biomédicas de la UNAM, Mexico City, Mexico, [14]Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH), UNAM, Juriquilla, Mexico

Current Genome-Wide Association Studies (GWAS) rely on genotype imputation to increase statistical power, improve fine-mapping of association signals, and facilitate meta-analyses. Due to the complex demographic history of Latin America and the lack of balanced representation of Native American genomes in current imputation panels, the discovery of locally relevant disease variants is likely to be missed, limiting the scope and impact of biomedical research in these populations. Therefore, the necessity of better diversity representation in genomic databases is a scientific imperative. Here, we expand the 1,000 Genomes reference panel (1KGP) with 134 Native American genomes (1KGP + NAT) to assess imputation performance in Latin American individuals of mixed ancestry. Our panel increased the number of SNPs above the GWAS quality threshold, thus improving statistical power for association studies in the region. It also increased imputation accuracy, particularly in low-frequency variants segregating in Native American ancestry tracts. The improvement is subtle but consistent across countries and proportional to the number of genomes added from local source populations. To project the potential improvement with a higher number of reference genomes, we performed simulations and found that at least 3,000 Native American genomes are

OPEN ACCESS

Edited by:
Tony Merriman, University of Otago, New Zealand

Reviewed by:
Inaho Dnjoh, Tohoku University, Japan
Mohamad Saad, Qatar Computing Research Institute, Qatar

***Correspondence:**
Andrés Moreno-Estrada
andres.moreno@cinvestav.mx
Lourdes García-García
garcigarml@gmail.com

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *03 June 2021*
Accepted: *01 November 2021*
Published: *03 January 2022*

Citation:
Jiménez-Kaufmann A, Chong AY, Cortés A, Quinto-Cortés CD, Fernandez-Valverde SL, Ferreyra-Reyes L, Cruz-Hervert LP, Medina-Muñoz SG, Sohail M, Palma-Martinez MDJ, Delgado-Sánchez G, Mongua-Rodríguez N, Mentzer AJ, Hill AVS, Moreno-Macías H, Huerta-Chagoya A, Aguilar-Salinas CA, Torres M, Kim HL, Kalsi N, Schuster SC, Tusié-Luna T, Del-Vecchyo DO, García-García L and Moreno-Estrada A (2022) Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes. Front. Genet. 12:719791. doi: 10.3389/fgene.2021.719791

needed to equal the imputation performance of variants in European ancestry tracts. This reflects the concerning imbalance of diversity in current references and highlights the contribution of our work to reducing it while complementing efforts to improve global equity in genomic research.

Keywords: Imputation, reference panels, GWAS, Native American ancestry, Latin Americans, underrepresented populations

INTRODUCTION

Over the past years, GWAS have identified thousands of genetic associations to multiple phenotypes (MacArthur et al., 2017; Visscher et al., 2017), targets for potential new drugs (Agrawal and Brown 2014; Flannick et al., 2014; Nelson et al., 2015), and facilitated disease stratification (Chatterjee, Shi, and García-Closas 2016). However, most GWAS have been performed in populations with European ancestry (Popejoy and Fullerton 2016). Unfortunately, the findings of large-scale GWAS performed in populations of European descent have limited portability to other ancestry groups (Duncan et al., 2019; Sirugo, Williams, and Tishkoff 2019) due to population substructure. This represents a major limitation in the case of Latin American populations as they are the result of recent admixture primarily between Native American, European, and African populations, and only 1.3% of both discovery and replication studies have been performed in these populations (Mills and Rahal 2019). Furthermore, the genetic composition of Latin American populations is heterogeneous between countries (Chacón-Duque et al., 2018; Soares-Souza et al., 2018) and within countries (Moreno-Estrada et al., 2014; Harris et al., 2018; Kehdy et al., 2015). Different demographic histories often lead to different associated variants to a given phenotype (Martin et al., 2017). For example, variants in the *SLC16A11* gene have been associated with an increased risk of diabetes in Mexicans and appear to be segregating at low frequency in Latin American populations specifically (SIGMA Type 2 Diabetes Consortium et al., 2014). Likewise, risk variants of renal disease in *APOL1* associated with renal disease in west African populations are also found in the Americas as a result of the Transatlantic slave trade, differentially shaping the frequency spectrum of disease variants among Afro-descendent Latino populations (Nadkarni et al., 2018). If the current bias in catalogs of human variation persists, many population-specific variants will be overlooked, and precision medicine strategies will not benefit all populations equally (Martin et al., 2019).

A critical step when performing a GWAS is genotype imputation, which leverages linkage disequilibrium (LD) structure and haplotype sharing to estimate untyped variation in a SNP array based on a reference panel (Marchini et al., 2007). Genotype imputation increases statistical power, improves fine-mapping of association signals, and facilitates meta-analysis (Marchini and Howie 2010). Currently, available imputation panels do not have an explicit representation of Native American genomes. A previous study showed that in Latin American populations, SNPs in chromosomal segments with Native American ancestry have reduced imputation quality compared to those in chromosomal segments of European ancestry (Martin et al., 2017). Therefore, association signals coming from chromosomal segments with Native American ancestry will be harder to detect. This limits the scope and impact of biomedical research in the region.

Several projects and initiatives around the world are contributing to revert this trend (GenomeAsia100K Consortium 2019; Mulder et al., 2018; Gurdasani et al., 2015; Magalhães et al., 2018). For example, the Ugandan Genome Resource (Gurdasani et al., 2019) comprises genome-wide data for 6,400 individuals, including a subset of 1,978 whole genomes, which is enabling researchers to explore the genetic substructure of the region, improve imputation in African populations, and foster the discovery of novel association signals. In Latin America, recent sequencing efforts have generated whole-genome data from dozens of Native American genomes, including the Peruvian Genome Project (Harris et al., 2018) and the 12G and 100G-MX Projects (Romero-Hidalgo et al., 2017; Aguilar-Ordoñez et al., 2021) from the National Institute of Genomic Medicine (INMEGEN) in Mexico. However, only a subset of the existing generated data is available to the scientific community given the data sharing mechanisms implemented in each country. An ongoing multi-institutional effort in Mexico, the MX Biobank Project, is generating genome-wide data for more than 6,000 individuals nationwide, including 50 whole genomes of Native American ancestry representing the genetic variation of indigenous diversity within Mexico (http://www.mxbiobankproject.org). At a global scale, the inclusion of diverse populations in disease association research has been well demonstrated by the PAGE study (Wojcik et al., 2019), which combines genome-wide data for 49,839 individuals with diverse ancestries, enabling the discovery of novel association signals to well-studied phenotypes. Here, we combine novel and publicly available data from multiple sources to build a population-specific reference panel of Native American variation aimed at improving imputation performance in Latin American populations by expanding the current and widely used reference of the 1,000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium et al., 2015) with 134 Native American genomes. Using a demographic simulation framework, we also explore the number of additional reference genomes that should be sequenced to bridge the gap in imputation quality between different ancestries. Strengthening these efforts in diverse populations is not only a question of equality in genomics, but it also entails the scientific advantage of furthering our understanding of complex phenotypes in biomedical research.

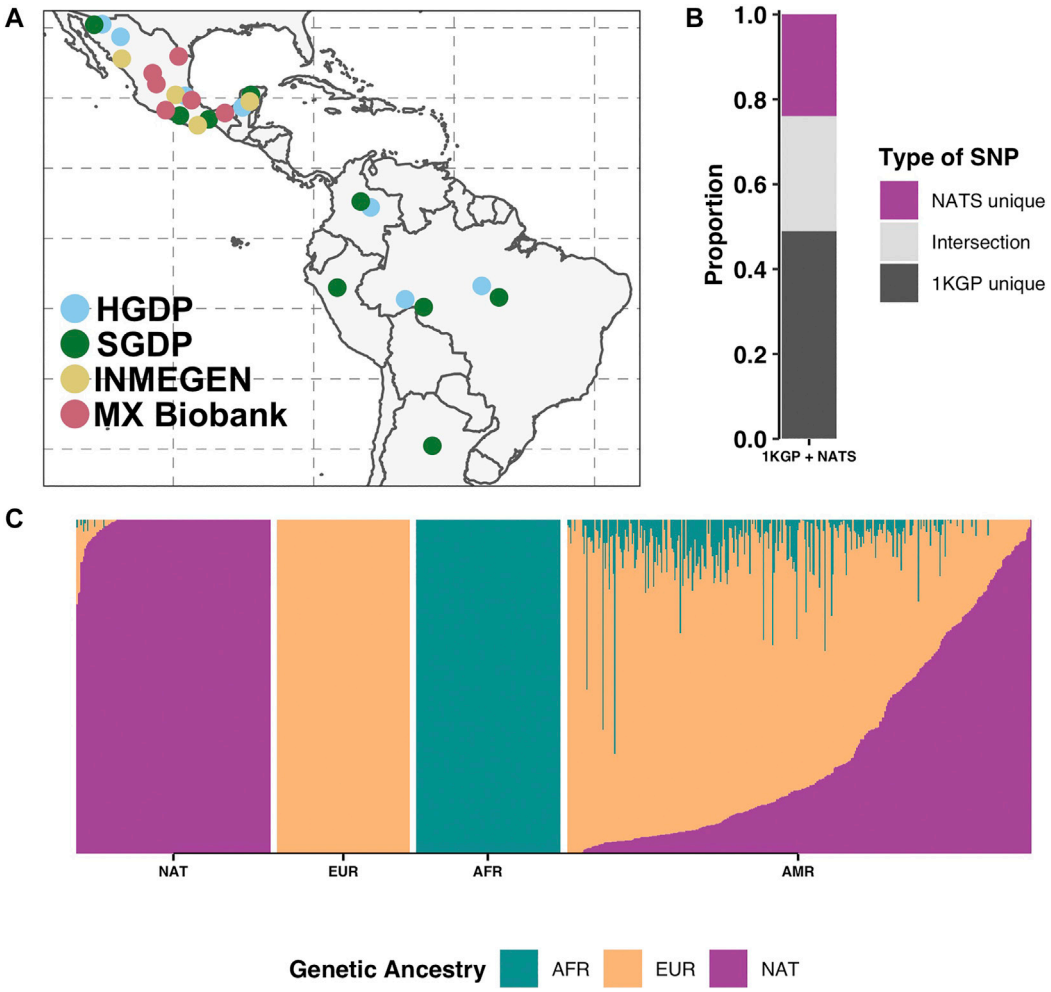

FIGURE 1 | Native American reference panel (NATS). **(A)** Geographical sampling locations of the NATS reference panel. Colors represent the four data sources: HGDP (61) (Bergström et al., 2020), SGDP (11) (Mallick et al., 2016), INMEGEN (12) (Romero-Hidalgo et al., 2017), and MX Biobank (50) totaling 134 genomes. **(B)** SNP proportions of the union of 1KGP and NATS (1KGP + NATS) by SNP sharing categories. We show the proportion of SNPs unique to 1KGP, SNPs unique to the NATS panel, and the intersection. **(C)** Unsupervised ADMIXTURE analysis at $K = 3$ of the NATS reference panel (far left, $N = 134$) together with 104 European (CEU), 113 African (YRI), and 347 admixed Latin American (AMR) samples from 1KGP. Genetic ancestry abbreviations: AFR—African, EUR—European, NAT—Native American.

MATERIALS AND METHODS

Building a Native American Reference Panel

Our panel consists of 134 Native American individuals broadly distributed across the continent (**Figure 1**; **Supplementary Tables S1, S2**). We gathered publicly available whole-genome sequencing (WGS) data from HGDP (Bergström et al., 2020) (61 individuals), SGDP (Mallick et al., 2016) (11 individuals), and INMEGEN (Romero-Hidalgo et al., 2017) (12 individuals). Additionally, we whole-genome sequenced the genome of 50 Mexican individuals with the highest Native American ancestry (99% on average) from the MX Biobank Project (http://www.mxbiobankproject.org). These were selected to maximize indigenous ancestry and geographical representation across Mexico. Individual genetic ancestry proportions were estimated using ADMIXTURE (Alexander, Novembre, and Lange 2009) at $K = 3$ using Utah residents with Northern and Western European ancestry (CEU), Yoruba in Ibadan, Nigeria (YRI), and the Latin Americans (AMR) of 1KGP as references.

To construct the panel, we restricted the datasets to biallelic SNPs with no missing data in any individual across each data source. This was done for all four data sources (**Supplementary Table S3**). The data processing was done using *VCFtools v0.1.17* (Danecek et al., 2011). Then, we merged the data using *bcftools v1.9* (Danecek et al., 2021) using the flag --missing-to-ref that fills the missing positions in one panel but present in another with homozygous reference. To minimize any potential bias introduced with this strategy, we made sure that any previously removed position in any of the sources was not present in the final freeze. The final dataset consists of a total of 10,981,451 SNPs.

Finally, we phased the data using *SHAPEIT2 v2. r837* (Delaneau et al., 2014) using the following flags: --window 0.5 --states 500 --burn 10 --prune 10 --main 50. Then, we converted the data to the reference format used by *IMPUTE2* (Howie et al., 2012). We named this panel NATS.

Whole-Genome Sequencing and Variant Calling

Fifty individuals from the MX Biobank Project were sequenced at 40X on Illumina HiSeqX instruments using dual indexed barcodes. The raw reads were aligned to the human genome assembly GRCh37 using *BWA v.0.7.17-r1198-dirty* (Li and Durbin 2009). We added the mate tags with *Samblaster v0.1.24* (Faust and Hall 2014) and used *Sambamba v0.7.1* (Tarasov et al., 2015) for file conversion and sorting. To generate the alignment statistics, we used *Samtools v1.10* (Li 2011) with the option *depth -a*. Finally, we performed variant calling and generated the final gvcf files with *GATK v4.1.9.0* (McKenna et al., 2010) using the human genome assembly GRCh37 as the reference genome. Details are available as part of the Supplementary Material (**Supplementary Table S2**; **Supplementary Figure S9**).

Creating a SNP Array Subset From WGS Data for Imputation Performance Evaluation

To evaluate the performance of our panel, we used WGS data from the 347 AMR individuals in 1KGP as target individuals for imputation. Namely, Puerto Ricans in Puerto Rico (PUR), Peruvians in Lima (PEL), Colombian in Medellin (CLM), and Mexican ancestry in Los Angeles (MXL). We generated an array dataset by subsetting the AMR individual genomes to the existing positions in the Multi-Ethnic Global Array (MEGA) using *VCFtools v.0.1.17* and saved the removed positions from the WGS data to use for imputation validation. Illumina's MEGA array includes nearly 1.8 M markers (1,779,819) genome-wide distributed and was designed to leverage SNP content from various global sequencing efforts, mostly Phase 3 of the 1,000 Genomes Project. To better approximate a real scenario, we unphased the array dataset with *Plink v1.9* (Chang et al., 2015) by transforming the data to bed format. Finally, we phased the dataset again with *SHAPEIT2 v2. r837* using 1KGP as a phasing reference.

Local Ancestry Inference

To evaluate the performance by ancestry, we deconvoluted local ancestry for the Latin American individuals from 1,000 Genomes. We used 70 YRI individuals in 1KGP as the African reference, 70 CEU individuals from 1KGP as the European reference, and 70 Native American individuals from (Moreno-Estrada et al., 2014) as the Native American reference. The selected individuals had the highest African, European, and Native American genetic components, respectively. We used the PopPhased version of *RFMix v.1.5.4* (Maples et al., 2013) with the following flags: -w 0.2 -e 0 --forward-backward.

Imputation and Imputation Performance

We implemented a leave-one-out strategy for imputation. Namely, the target individual was removed from the 1KGP reference. We performed imputation with *IMPUTE2* for chromosomes 2 and 9. These chromosomes, being the largest and of intermediate size, respectively, were selected to ensure a representative subset of variants across the genome while keeping the project within the available computational capacity. We used 1KGP and 1KGP + NATS as reference panels. When using 1KGP as a reference, we used the flag --k_haps 1,000, and when using 1KGP + NATS, we used the flags --merge-ref-panels and --k_haps 1,250.

We obtained the imputed dosages with the formula: P(Aa) + 2P(aa). We computed the Pearson squared correlation (r^2) between the imputed dosages and the real dosages for each individual using R software. Overall imputation accuracy was stratified by minor allele frequency and local ancestry diplotype (AFR_AFR, AFR_EUR, AFR_NAT, EUR_EUR, EUR_NAT, NAT_NAT). We also compared the number of SNPs above the GWAS quality threshold (MAF >=0.01 and INFO >0.3) for both reference panels stratified by local ancestry diplotype in the target individuals.

Demographic Simulation

We simulated neutral genetic sequence data under a coalescent model. We used the *msprime* (Kelleher, Etheridge, and McVean 2016) option of *stdpopsim* (Adrion et al., 2020) to simulate a previously defined American admixture model for Latinos (Browning et al., 2018). It models African, European, and Asian (as Native American proxy) demographic history and an admixture event taking place 12 generations ago. In the absence of realistic admixture models that use Native American instead of East Asian genomes as proxy in the simulations and based on the framework described by Browning et al. (2018), we will now refer to the simulated Asian population as Native American for the purpose of predicting imputation performance at incremental numbers of reference genomes in a similar scenario to the Latin American admixture. The simulated admixed population ancestral proportions are 1/6 African, ⅓ European, and ½ Native American. In total, we simulated chromosome 9 for 661 Africans, 503 Europeans, 3,000 Native Americans, and 657 admixed individuals. We selected all the Africans, Europeans, and the first 347 admixed individuals to serve as the base reference panel (note that these numbers mirror the sample sizes of 1KGP for each ancestry). The remaining 300 admixed individuals were used as imputation targets, and incremental subsets of the 3,000 Native American genomes were added sequentially to the base reference panel.

To simulate genotype array data for the target individuals, we downsampled the simulated neutral sequence to match the allele frequency spectrum in European populations of 1KGP and the average distance between SNPs of the MEGA array. We used the European populations in 1KGP to mirror the ascertainment bias towards European ancestry in current array designs. We estimated local ancestry using *RFMix* for the 300 admixed individuals used as imputation targets. We randomly selected

TABLE 1 | SNPs above the standard quality threshold using both panels after imputing missing variants. We show the average number of SNPs with MAF >= 0.01 and INFO >= 0.3 using both reference panels and the overall proportion of Native American ancestry of the population. *p*-value was calculated with a two-tailed paired *t*-test. The average number of SNPs with MAF <0.01 and INFO >0.3 for both panels is shown in **Supplementary Table S4**.

Population	SNPs above quality threshold (1KGP)	SNPs above quality threshold (1KGP + NATS)	Increase of SNPs using 1KGP + NATS	Average proportion of Nat. American ancestry
Peru (PEL)	244,818	248,087	3,269 (*p*-value = 2.03e-49)	0.70
Mexico (MXL)	265,619	268,254	2,635 (*p*-value = 6.5e-31)	0.42
Colombia (CLM)	279,828	281,911	2,163 (*p*-value = 8.3e-47)	0.18
Puerto Rico (PUR)	291,035	292,734	1,699 (*p*-value = 2.9e-67)	0.06

100 simulated individuals from each ancestral population (African, European, and Asian) as references for the local ancestry inference. Here, again, we used Asians as the closest proxy for Native Americans in the available simulation model.

We conducted imputation with the base reference panel plus a varying number of additional reference genomes (0, 100, 134, 200, 400, 600, 800, 1,000, 1,500, 2000, and 3,000). Finally, we compared imputation r^2 of using different reference panels stratified by local ancestry and allele frequency in the target individual genomes.

RESULTS

The Native American Reference Panel NATS

We built a Native American reference panel (NATS) representing indigenous populations across Latin America. The panel consists of publicly available data [HGDP (Bergström et al., 2020), SGDP (Mallick et al., 2016), and INMEGEN (Romero-Hidalgo et al., 2017)] and 50 new genomes from the MX Biobank Project (*Materials and Methods*, and **Supplementary Table S2**). While most of the genomes in the panel are from indigenous groups in Mexico (103 of 134; 76.8%) (**Figure 1A**; **Supplementary Table S1**), our panel also encompasses native groups from Colombia, Brazil, and Peru. When merging NATS with 1KGP, the total number of SNPs is 102,336,497, of which 24,518,242 (24%) are unique to our panel (**Figure 1B**). The amount of non-indigenous admixture in our panel is less than 1.5% overall (**Figure 1C**). Only some Mayan individuals from HGDP show between 0.8 and 23% of European admixture (on average 6%) (**Supplementary Table S1**). Overall, our panel has 98.5% of Native American genetic ancestry. We acknowledge that, while this panel includes as many genomes as possible from those publicly available at the time of publication, it does not fully capture the genetic variation of the vast ethnic diversity in the continent. It is intended to serve as a first approximation to evaluate the impact of ancestry representation in imputation performance.

Imputation Performance of the NATS Reference Panel

To assess the impact of our panel on imputation performance, we imputed the AMR individuals (from Colombia, Peru, Puerto Rico, and Mexico in 1KGP) at SNPs not found on the MEGA array using a leave-one-out strategy, with either 1KGP or 1KGP + NATS as reference panels (Materials and Methods). We chose the MEGA array because it was specifically designed to capture global variation better. We compared the mean number of SNPs above the standard quality threshold for human genetic studies (MAF >= 1% and INFO >= 0.3) using the two reference panels. We were able to increase the number of SNPs above the quality threshold across the four populations using our NATS panel (**Table 1**). The magnitude of the increase is correlated with the individual's proportion of native ancestry (**Supplementary Figure S1**). Furthermore, the majority of these SNPs fall into diploid European tracts of the genome (**Supplementary Figure S2**) regardless of the ancestry composition of each population, and which reference panel was used for imputation. This is because even though the 1KGP has as many African individuals as Europeans, European ancestry is more predominant in AMR individuals.

To determine imputation accuracy, we computed the correlation between the real allele dosages and the imputed dosages (*Materials and Methods*). We checked imputation accuracy in 1KGP admixed individuals trimmed down to SNP array positions stratified by diploid ancestry (**Figure 2A**). Overall, imputation accuracy is worse in AMR populations with the highest proportion of Native American ancestry (**Supplementary Figure S3**). As previously reported (Martin et al., 2017), the ancestry tracts that perform the worst are the ones that are underrepresented in the reference panel, specifically African and Native American. Next, we evaluated imputation accuracy using our panel (1KGP + NATS). We were able to increase imputation accuracy particularly in rare alleles (frequency >0.003 and <0.008) with diploid Native ancestry of the Mexican population (*p*-value < 0.05 two-tailed paired *t*-test) (**Figure 2B**) but not for the other populations (**Supplementary Figure S3**) or in common frequencies (**Supplementary Figure S4**). Interestingly, we do not see the same increase in the Peruvian population, which has the highest proportion of Native American ancestry overall. This could be explained by the fact that the majority of our reference data comes from native Mexicans (**Figure 1A**; **Supplementary Table S1**). Since rare variants tend to be more private to each population (Biddanda, Rice, and Novembre 2020), we could better impute rare alleles in admixed Mexicans. This suggests that, to see a similar improvement in accuracy in the other populations, we would need to include more native individuals from each local region.

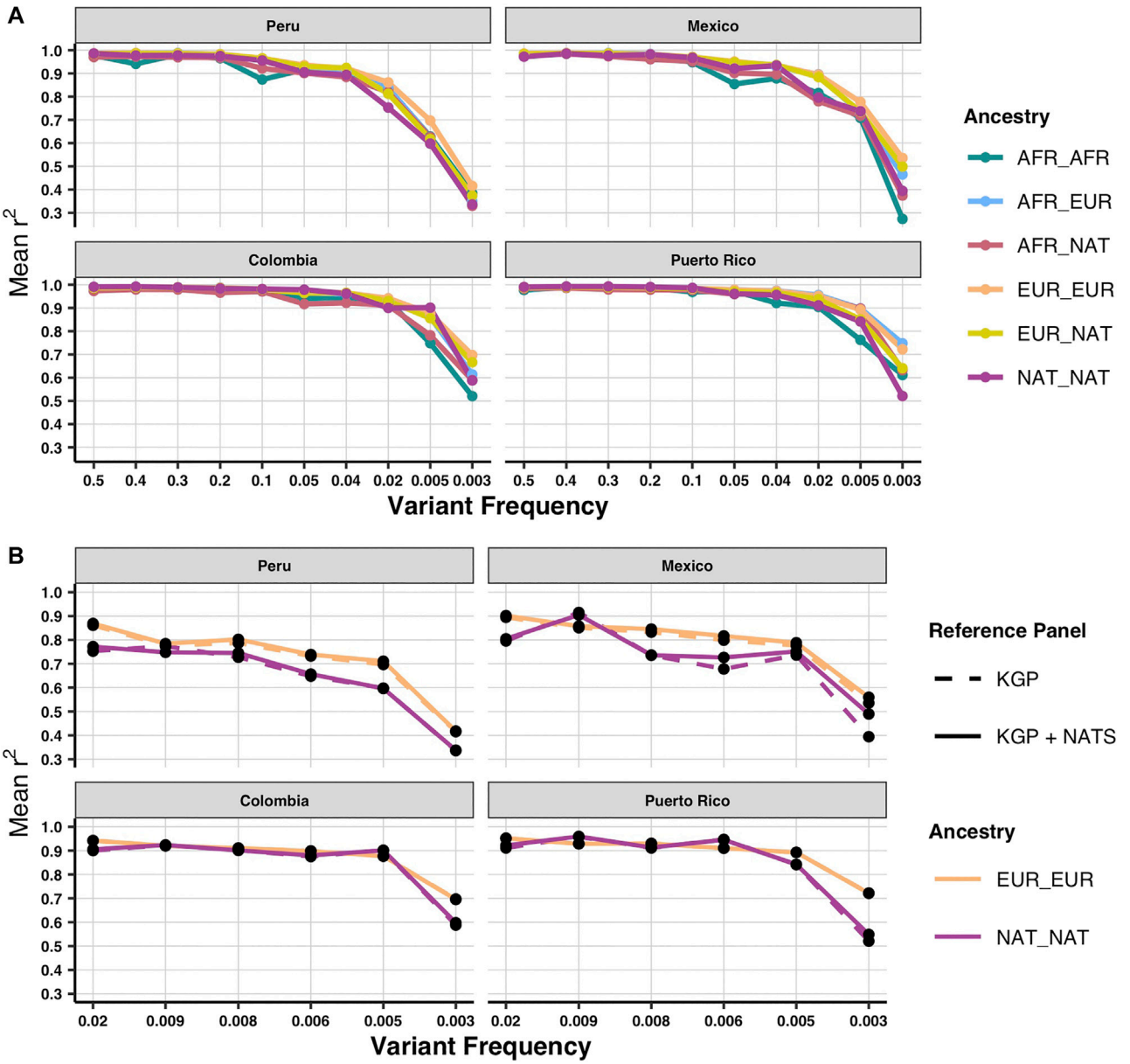

FIGURE 2 | Imputation accuracy by local ancestry and population using both reference panels. **(A)** Imputation accuracy of the four AMR populations stratified by diploid local ancestry for the MEGA array using 1KGP as reference panel. **(B)** Imputation accuracy for the Native and European diploid ancestries using 1KGP and 1KGP + NATS as reference panel focusing on rare alleles. Imputation accuracy was calculated with the Pearson squared correlation between imputed and real allele dosages.

Surprisingly, we could also see an improvement in diploid European ancestry tracts in the Mexican population (p-value < 0.05 two-tailed paired t-test for SNPs with frequency >0.003) (**Figure 2B**). One possible explanation is that because our NATS reference panel still keeps a minor fraction of European ancestry, some European haplotypes at higher frequency in Mexico could be better captured by reference genomes with such a genetic mixture. In some cases, like variants of frequency <0.02 and >0.009 with diploid Native ancestry in PEL, we could also observe a slight decrease in imputation accuracy using NATS. This could result from the uncertainty added to the data in the cross-imputation step that *IMPUTE2* performs when merging two reference panels (Howie, Marchini, and Stephens 2011).

Predicting Imputation Improvement From Additional Native American Genomes Using Simulations

Our results show that after adding 134 Native American genomes to the most widely used reference panel of global variation, we observe a promising trend of improvement. Still,

we do not come close enough to equal the imputation performance for other better represented ancestries. The question remains of how many additional genomes are still needed to close the gap. To explore this, we employed demographic simulations using *stdpopsim* (Adrion et al., 2020) and *msprime* (Kelleher, Etheridge, and McVean 2016) to generate data for a previously defined American admixture model (Browning et al., 2018). This approach allows us to explore a simulated scenario where three divergent populations intermingle to form a new admixed population (like it occurred in Latin America). With this, we can replicate the current situation where reference data are mostly available for two of the three source populations. By being able to simulate any amount of data, we can assess how many genomes of the underrepresented population (in our case, Native Americans) are necessary to equal imputation performance across ancestries. Briefly, the model simulates African, European, and Asian source populations. In the context of this analysis, the Asian population serves as a proxy for a Native American reference. We do not directly simulate a Native American population due to the lack of realistic admixture models that incorporate Native American instead of East Asian genomes as proxy in the inference of demographic parameters, which are needed to properly run the simulations. Building such demographic model is beyond the scope of this study, so given the available model and since this project focuses on Latin American populations, we will refer to the simulated Asian population as Native American. The model also simulates an admixed population that consists of 1/6 African, ⅓ European, and ½ Native American. We generated a base reference panel consisting of 661 Africans, 503 Europeans, and 347 admixed individuals (matching 1KGP sample sizes for those ancestries), as well as 3,000 Native American individuals to add sequentially to the base reference, and 300 additional admixed individuals as imputation targets (*Materials and Methods*).

We confirmed the ancestry proportions of our simulated data using *ADMIXTURE* (**Supplementary Figure S5**). To replicate the imputation pipeline, we created a genotype array dataset for the simulated target individuals by matching mean distance between markers and frequency in the European population of SNPs in the MEGA array to the simulated array, to mirror the bias in standard arrays (*Materials and Methods* and **Supplementary Figure S6**). Then, we imputed the 300 target individuals with the base reference plus either 0, 100, 134 (to mirror the sample size in NATS), 200, 400, 600, 800, 1,000, 1,500, 2000, or 3,000 Native Americans. We were able to recover roughly the same pattern of imputation accuracy (**Supplementary Figure S7**). Namely, accuracy decreased the less represented the ancestry was in the base reference with the Native American as the worst-performing ancestry. One caveat is that the best-performing ancestry is African contrary to what we see in the real data (**Figure 2A**). This is likely because the 661 African individuals are from the population that contributed to the admixed population in the simulation, which is not the case for real data. Different African ancestries contributed more or less to different Latin American populations (Micheletti et al., 2020) and not all are present in 1KGP.

When incorporating additional Native American genomes, imputation accuracy only increased in those tracts with any Native ancestry (**Supplementary Figure S8**). Furthermore, for imputation accuracy in Native American diploid ancestry tracts to equal that in European diploid ancestry tracts, 3,000 Native genomes were needed for variants with frequency>=2%, while 1,500 were enough for variants with frequency <2% (**Figure 3A**). To ask whether we reach a saturation point in the increase of imputation accuracy in the Native diploid ancestry, we compared the difference between accuracy in the base reference versus each additional reference. As expected, the behavior is different for common (frequency >0.05), low (frequency <0.05 and >0.01), and rare (frequency <0.01) variants (**Figure 3B**). Neither of them seems to show a saturation point at 3,000 newly added Native genomes. The steepest increase is achieved for the rare alleles, whereas for the common alleles, the increase is slower. This agrees with the previous result where more genomes were needed to match the Native imputation accuracy to the European one for common variants. It is also evident that the variants of common frequency are closest to saturation in accuracy as their values were already close to one (**Figure 3A**).

DISCUSSION

GWAS requires large sample sizes to detect genetic associations to complex phenotypes, and more so as the field moves toward studying rare variants (Collins 2012; Amendola et al., 2018; Abul-Husn and Kenny 2019). Therefore, SNP array platforms will continue to inform GWAS even as whole-genome sequencing costs continue to drop. In this scenario, imputation tools and genome variation resources are vital to increasing the statistical power to discover associations in understudied populations. So far, GWAS have mainly focused on populations with European ancestry (Popejoy and Fullerton 2016; Mills and Rahal 2019) and, over the past years, interesting discoveries have been made (Visscher et al., 2017). However, not all GWAS results are portable between populations (Martin et al., 2017; Duncan et al., 2019; Sirugo, Williams, and Tishkoff 2019). To ensure that these advances reach all people equitably, we must expand these studies to other populations. Other recent projects around the world have sought to reverse this trend (Gurdasani et al., 2015, 2019; GenomeAsia100K Consortium 2019; Magalhães et al., 2018; Mulder et al., 2018) improving imputation accuracy, fine mapping of associations, and discovering novel associations to well-studied phenotypes. We sought to add to this trend by creating a Native American imputation reference panel merging publicly available Native American genomes (Mallick et al., 2016; Romero-Hidalgo et al., 2017; Bergström et al., 2020) with 50 novel genomes.

One major caveat of our panel is that it does not comprehensively reflect the indigenous genetic variation across the Americas. Most of the data come from individuals from Mexico. Furthermore, the 134 genomes added are only a small increment (5%) with respect to 1KGP. The contribution of this

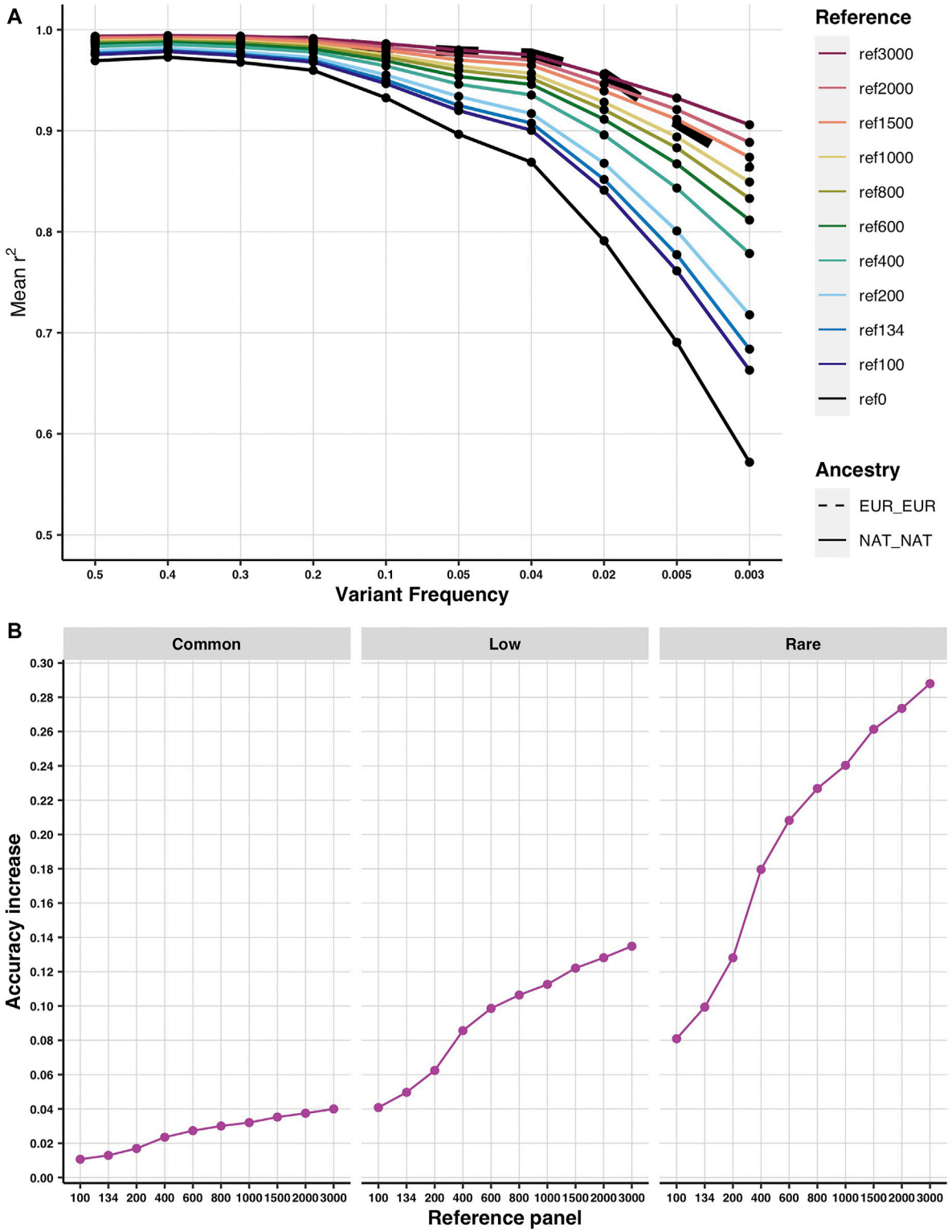

FIGURE 3 | Predicted imputation accuracy according to demographic simulations. **(A)** Imputation accuracy in the diploid Native American (solid colored lines) and diploid European (thick dashed line) ancestries using different simulated reference panels of incremental sizes. Ref 0 stands for the base reference (as it has 0 additional reference genomes). Given the available demographic model (Browning et al., 2018), a simulated Asian population was used as a proxy for Native American ancestry for the purpose of reproducing a three-way admixture process with similar ancestry proportions of African, European, and Native American sources to that observed in admixed Latino populations (see Methods for details). **(B)** Increase in imputation accuracy from the base reference in the Native American diploid ancestry at increasing sizes of the reference panel by allele frequency [common (0.5–0.05), low (0.05–0.01), and rare (0.01–0.003)].

panel is small in comparison to projects like the Uganda Genome Resource that sequenced 1,978 novel genomes (Gurdasani et al., 2019). Even with these limitations in mind, we were able to quantify the consequences of the lack of Native American genomes in commonly used imputation reference panels using empirical and simulated data analyses, while highlighting what this means for ongoing and future studies in the region.

Our panel increased the number of SNPs above the standard quality threshold for human genetic studies increasing statistical power in the four AMR populations of 1KGP. This mirrors what has been achieved by other studies in other populations (Ahmad et al., 2017; Magalhães et al., 2018; Gurdasani et al., 2019). The magnitude of this increase is positively correlated with the proportion of Native American ancestry. In other words, our panel has a stronger impact on individuals with higher Native American ancestry. However, even after using our panel, the majority of SNPs that were above the quality threshold are in chromosomal segments of the genome with European diploid ancestry, regardless of the proportion of European ancestry in the population, due to an over-representation of this ancestry in the reference panel. This means that, when doing a GWAS, the genetic signals predominantly found on the European ancestry will be easier to detect.

We were able to increase imputation accuracy in rare variants of Native American diploid ancestry in the MXL population. This was not the case for the other three populations. We expected that, since PEL is the population with the highest Native American ancestry proportion, it would also be the population most benefited by the use of our extended panel. However, there can be high levels of genetic differentiation among Native American groups, even if they are geographically close (Moreno-Estrada et al., 2014). In light of this fact, it is not a surprise that our panel, constructed with a majority of Native American individuals from Mexico, only improves accuracy in the MXL population. This suggests that to observe similar results in other populations, we should include more individuals of those populations in our panel. We also observed an increase in accuracy in some variants of European diploid ancestry. This could be attributed to the small fraction of European admixture present in the whole genomes of our extended panel, despite being enriched for Native American ancestry. Also, some of these European haplotypes could have better-captured variation found in European ancestry segments of MXL individuals. Finally, to achieve an overall increase in imputation accuracy across the whole spectrum of variant frequencies as achieved in other studies (Ahmad et al., 2017; Gurdasani et al., 2019), we would need a larger Native American reference panel, as quantified by our simulations.

These results are important with regard to not only GWAS but also their further applications. For instance, one of the applications of GWAS summary statistics is Polygenic Risk Scores (PRS). PRS calculates the genetic "risk" of an individual for a particular phenotype by summing the risk alleles present in that individual (Torkamani, Wineinger, and Topol 2018). PRS necessitates summary statistics calculated in a population as close as possible to the target individuals to be accurate. Previous studies have shown that this is not a trivial task (Tropf et al., 2017; Sirugo, Williams, and Tishkoff 2019; Mostafavi et al., 2020). Even among European populations, PRS estimates vary widely depending on the source of summary statistics due to population structure (Berg et al., 2019; Sohail et al., 2019). To have accurate PRS for the Latin American population, we need to have more studies in the region. Furthermore, our results show that we also need a better imputation panel for these populations to avoid a bias towards identifying genetic signals present on the European ancestry background.

The question of how much data are needed remained. To answer it, we employed demographic simulations. We replicated the same pattern of imputation accuracy of our data and of previous studies (Martin et al., 2017). Our strategy shows that we would need at least 3,000 Native American genomes to equal imputation accuracy of Native diploid ancestry to that of European diploid ancestry across all variant frequencies. This number holds for populations such as MXL with roughly similar ancestral proportions as the simulated admixed population. The minimum number of necessary new genomes will change depending on the proportion of native ancestry of the target population. Our study provides a framework for future projects to decide how many resources to allocate to the generation of whole-genome data. Furthermore, we have shown that rare variants are the most benefited by the addition of new data. This will prove particularly relevant as the field moves towards studying that end of the variant frequency spectrum (Cirulli et al., 2020; Minikel et al., 2020). Overall, our results show the importance of generating more diverse imputation panels to enable genetic discoveries in a broader spectrum of human diversity and to procure equity as scientific advancements in precision medicine should extend globally in benefit of all.

DATA AVAILABILITY STATEMENT

The newly generated data presented in the study are deposited in the European Genome-phenome Archive (EGA) repository, accession number EGAD00001008354 i.e. https://ega-archive.org/datasets/EGAD00001008354.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee (Approval CI-1479) of the National Institute of Public Health, Mexico. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AJ-K, AM-E, AYC, AC, SF-V, AJM, MS, AH, and DD-V designed the study. AM-E, LG-G, LF-R, LC-H, TT-L, HM-M, CA-S, and NM-R selected and provided DNA samples from the MX Biobank Project. AM-E, and AH sequenced the data. HK, NK, SS, MT,

CQ-C, and MP-M performed the whole-genome variant calling and curated the data. AJ-K, AYC, and SM-M analyzed the data. AJ-K and AM-E drafted the manuscript, with input from LG-G, CQ-C, LF-R, LC-H, TT-L, HM-M, CA-S, AH-C, MS, SM-M, NM-R, and GD-S. All authors read and approved the manuscript.

FUNDING

This work was supported by "The Mexican Biobank Project: Building Capacity for Big Data Science in Medical Genomics in Admixed Populations", a binational initiative between Mexico and the UK co-funded by CONACYT (Grant number FONCICYT/50/2016), and The Newton Fund through The Medical Research Council (Grant number MR/N028937/1) awarded to AME and AVSH. It was also supported by the International Center for Genetic Engineering and Biotechnology (ICGEB, Italy) grant number CRP/MEX20-01. MS was partially supported by the Chicago Fellows program of the University of Chicago. DODV is supported by the UC MEXUS CONACYT collaborative program (Grant number CN-19-29), and the UNAM PAPIIT funding program (Grant number IA200620).

ACKNOWLEDGMENTS

We thank the participants of the *Encuesta Nacional de Salud, 2000* (2000 National Health Survey, ENSA 2000), conducted in Mexico nationwide by the *Secretaría de Salud* (Health Secretariat) and the *Instituto Nacional de Salud Pública* (National Institute of Public Health, INSP). We are grateful to Mitzi Flores and Adriana Garmendia for project management support and to Carlos Conde, Victor Guerrero Lemus, Armando Mendez Herrera, Cruz Portugal García, Ma. Luisa Ordóñez-Sánchez, Rosario Rodriguez-Guillen, and Manuel Velazquez Mesa for biobank maintenance and sample preparation. We also thank Mary Ortega, Cecilia Gutiérrez, and Sara García for technical assistance, Jacob Cervantes for IT support, and Aaron Ragsdale for comments on earlier versions of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.719791/full#supplementary-material

REFERENCES

Abul-Husn, N. S., and Kenny, E. E. (2019). Personalized Medicine and the Power of Electronic Health Records. *Cell* 177 (1), 58–69. doi:10.1016/j.cell.2019.02.039

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., et al. (2020). A Community-Maintained Standard Library of Population Genetic Models. *eLife* 9, e54967. doi:10.7554/eLife.54967

Agrawal, N., and Brown, M. A. (2014). Genetic Associations and Functional Characterization of M1 Aminopeptidases and Immune-Mediated Diseases. *Genes Immun.* 15 (8), 521–527. doi:10.1038/gene.2014.46

Aguilar-Ordoñez, I., Pérez-Villatoro, F., García-Ortiz, H., Barajas-Olmos, F., Ballesteros-Villascán, J., González-Buenfil, R., et al. (2021). Whole Genome Variation in 27 Mexican Indigenous Populations, Demographic and Biomedical Insights. *PLoS One* 16 (4), e0249773. doi:10.1371/journal.pone.0249773

Ahmad, M., Sinha, A., Ghosh, S., Kumar, V., Davila, S., Yajnik, C. S., et al. (2017). Inclusion of Population-Specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy. *Sci. Rep.* 7 (1), 6733. doi:10.1038/s41598-017-06905-6

Alexander, D. H., Novembre, J., and Lange., K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Amendola, L. M., Berg, J. S., Horowitz, C. R., Angelo, F., Bensen, J. T., Biesecker, B. B., et al. (2018). The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. *Am. J. Hum. Genet.* 103 (3), 319–327. doi:10.1016/j.ajhg.2018.08.007

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., et al. (2019). Reduced Signal for Polygenic Adaptation of Height in UK Biobank. *eLife* 8, e39725. doi:10.7554/eLife.39725

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., et al. (2020). Insights into Human Genetic Variation and Population History from 929 Diverse Genomes. *Science* 367 (6484), eaay5012. doi:10.1126/science.aay5012

Biddanda, A., Rice, D. P., and Novembre, J. (2020). A Variant-Centric Perspective on Geographic Patterns of Human Allele Frequency Variation. *eLife* 9, e60107. doi:10.7554/eLife.60107

Browning, S. R., Browning, B. L., Daviglus, M. L., Durazo-Arvizu, R. A., Schneiderman, N., Kaplan, R. C., et al. (2018). Ancestry-Specific Recent Effective Population Size in the Americas. *PLoS Genet.* 14 (5), e1007385. doi:10.1371/journal.pgen.1007385

Chacón-Duque, J-C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Rodrigo, B., et al. (2018). Latin Americans Show Wide-Spread Converso Ancestry and Imprint of Local Native Ancestry on Physical Appearance. *Nat. Commun.* 9 (1), 5388. doi:10.1038/s41467-018-07748-z

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaScience* 4, 7. doi:10.1186/s13742-015-0047-8

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and Evaluating Polygenic Risk Prediction Models for Stratified Disease Prevention. *Nat. Rev. Genet.* 17 (7), 392–406. doi:10.1038/nrg.2016.27

Cirulli, E. T., White, S., Read, R. W., Elhanan, G., Metcalf, W. J., Tanudjaja, F., et al. (2020). Genome-Wide Rare Variant Analysis for Thousands of Phenotypes in over 70,000 Exomes from Two Cohorts. *Nat. Commun.* 11 (1), 542. doi:10.1038/s41467-020-14288-y

Collins, R. (2012). What Makes UK Biobank Special? *Lancet* 379 (9822), 1173–1174. doi:10.1016/s0140-6736(12)60404-8

Danecek, P., Adam, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *GigaScience* 10 (2). giab008. doi:10.1093/gigascience/giab008

Delaneau, O., and Marchini, J.1000 Genomes Project Consortium; 1000 Genomes Project Consortium (2014). Integrating Sequence and Array Data to Create an Improved 1000 Genomes Project Haplotype Reference Panel. *Nat. Commun.* 5, 3934. doi:10.1038/ncomms4934

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nat. Commun.* 10 (1), 3328. doi:10.1038/s41467-019-11112-0

Faust, G. G., and Hall., I. M. (2014). SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction. *Bioinformatics* 30 (17), 2503–2505. doi:10.1093/bioinformatics/btu314

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B., Grarup, N., Burtt, N. P., et al. (2014). Loss-of-Function Mutations in SLC30A8 Protect against Type 2 Diabetes. *Nat. Genet.* 46 (4), 357–363. doi:10.1038/ng.2915

GenomeAsia 100K Consortium (2019). The GenomeAsia 100K Project Enables Genetic Discoveries across Asia. *Nature* 576 (7785), 106–111. doi:10.1038/s41586-019-1793-z

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The African Genome Variation Project Shapes Medical Genetics in Africa. *Nature* 517 (7534), 327–332. doi:10.1038/nature13997

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., et al. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* 179 (4), 984e36–1002. doi:10.1016/j.cell.2019.10.004

HarrisDaniel, N., Wei, S., Amol, C., ShettyKelly, S., et al. (2018). "Evolutionary Genomic Dynamics of Peruvians Before, During, and after the Inca Empire.," in Proceedings of the National Academy of Sciences of the United States of America 115 (28), E6526–E6535.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3 Genes|Genomes|Genetics* 1 (6), 457–470. doi:10.1534/g3.111.001198

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-phasing. *Nat. Genet.* 44 (8), 955–959. doi:10.1038/ng.2354

Kehdy, F. S. G., Gouveia, M. H., Machado, M., Magalhães, W. C. S., Horimoto, A. R., Horta, B. L., et al. (2015). Origin and Dynamics of Admixture in Brazilians and its Effect on the Pattern of Deleterious Mutations. *Proc. Natl. Acad. Sci. United States Am.* 112 (28), 8696–8701. doi:10.1073/pnas.1504447112

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* 12 (5), e1004842. doi:10.1371/journal.pcbi.1004842

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H. (2011). A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* 27 (21), 2987–2993. doi:10.1093/bioinformatics/btr509

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog). *Nucleic Acids Res.* 45 (D1), D896–D901. doi:10.1093/nar/gkw1133

Magalhães, W. C. S., Araujo, N. M., Leal, T. P., Araujo, G. S., Viriato, P. J. S., Kehdy, F. S., et al. (2018). EPIGEN-Brazil Initiative Resources: A Latin American Imputation Panel and the Scientific Workflow. *Genome Res.* 28 (7), 1090–1095. doi:10.1101/gr.225458.117

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations. *Nature* 538 (7624), 201–206. doi:10.1038/nature18964

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* 93 (2), 278–288. doi:10.1016/j.ajhg.2013.06.020

Marchini, J., and Howie, B. (2010). Genotype Imputation for Genome-Wide Association Studies. *Nat. Rev. Genet.* 11 (7), 499–511. doi:10.1038/nrg2796

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes. *Nat. Genet.* 39 (7), 906–913. doi:10.1038/ng2088

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* 107 (4), 788–789. doi:10.1016/j.ajhg.2017.03.004

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities. *Nat. Genet.* 51 (4), 584–591. doi:10.1038/s41588-019-0379-x

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20 (9), 1297–1303. doi:10.1101/gr.107524.110

Micheletti, S. J., Bryc, K., Ancona Esselmann, S. G., Freyman, W. A., Moreno, M. E., Poznik, G. D., et al. (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am. J. Hum. Genet.* 107 (2), 265–277. doi:10.1016/j.ajhg.2020.06.012

Mills, M. C., and Rahal., C. (2019). A Scientometric Review of Genome-Wide Association Studies. *Commun. Biol.* 2, 9. doi:10.1038/s42003-018-0261-x

Minikel, E. V., Karczewski, K. J., Martin, H. C., Cummings, B. B., Whiffin, N., Rhodes, D., et al. (2020). Evaluating Drug Targets through Human Loss-Of-Function Genetic Variation. *Nature* 581 (7809), 459–464. doi:10.1038/s41586-020-2267-z

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Martin, S., Contreras, A. V., et al. (2014). The Genetics of Mexico Recapitulates Native American Substructure and Affects Biomedical Traits. *Science* 344 (6189), 1280–1285. doi:10.1126/science.1251688

Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., and Przeworski, M. (2020). Variable Prediction Accuracy of Polygenic Scores within an Ancestry Group. *eLife* 9, e48376. doi:10.7554/eLife.48376

Mulder, N., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., et al. (2018). H3Africa: Current Perspectives. *Pharmacogenomics Pers. Med.* 11, 59–66. doi:10.2147/pgpm.s141546

Nadkarni, G. N., Gignoux, C. R., Sorokin, E. P., Rahman, R., Barnes, K. C., and Wassel, C. L. (2018). Worldwide Frequencies of APOL1 Renal Risk Variants. *New Engl. J. Med.* 379 (26), 2571–2572. doi:10.1056/nejmc1800748

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., et al. (2015). The Support of Human Genetic Evidence for Approved Drug Indications. *Nat. Genet.* 47 (8), 856–860. doi:10.1038/ng.3314

Popejoy, A. B., and Fullerton, S. M. (2016). Genomics Is Failing on Diversity. *Nature* 538 (7624), 161–164. doi:10.1038/538161a

Romero-Hidalgo, S., Ochoa-Leyva, A., Garcíarrubio, A., Acuña-Alonzo, V., Antúnez-Argüelles, E., Balcazar-Quintero, M., et al. (2017). Demographic History and Biologically Relevant Genetic Variation of Native Mexicans Inferred from Whole-Genome Sequencing. *Nat. Commun.* 8 (1), 1005. doi:10.1038/s41467-017-01194-z

SIGMA Type 2 Diabetes ConsortiumWilliams, A. L., Jacobs, S. B. R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., et al. (2014). Sequence Variants in SLC16A11 Are a Common Risk Factor for Type 2 Diabetes in Mexico. *Nature* 506 (7486), 97–101. doi:10.1038/nature12828

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell* 177 (4), 1080. doi:10.1016/j.cell.2019.04.032

Soares-Souza, G., Borda, V., Kehdy, F., and Tarazona-Santos, E. (2018). Admixture, Genetics and Complex Diseases in Latin Americans and US Hispanics. *Curr. Genet. Med. Rep.* 6 (4), 208–223. doi:10.1007/s40142-018-0151-z

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., et al. (2019). Polygenic Adaptation on Height Is Overestimated Due to Uncorrected Stratification in Genome-Wide Association Studies. *eLife* 8, e39702. doi:10.7554/eLife.39702

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: Fast Processing of NGS Alignment Formats. *Bioinformatics* 31 (12), 2032–2034. doi:10.1093/bioinformatics/btv098

The 1000 Genomes Project ConsortiumAuton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A Global Reference for Human Genetic Variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The Personal and Clinical Utility of Polygenic Risk Scores. *Nat. Rev. Genet.* 19 (9), 581–590. doi:10.1038/s41576-018-0018-x

Tropf, F. C., Lee, S. H., Verweij, R. M., Stulp, G., van der Most, P. J., de Vlaming, R., et al. (2017). Hidden Heritability Due to Heterogeneity across Seven Populations. *Nat. Hum. Behav.* 1 (10), 757–765. doi:10.1038/s41562-017-0195-1

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101 (1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits. *Nature* 570 (7762), 514–518. doi:10.1038/s41586-019-1310-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.