Genotype imputation from low-coverage data for medical and population genetic analyses

Simone Andrea Biagini, Sara Becelaere, Mio Aerden, et al.

Genome Res. published online July 22, 2025
Access the most recent version at doi:10.1101/gr.280175.124

Genotype imputation from low-coverage data for medical and population genetic analyses

Simone Andrea Biagini,[1,2,3] Sara Becelaere,[1] Mio Aerden[1,4,], Tatjana Jatsenko[1], Laurens Hannes,[1,4] Philip Van Damme,[5,6] Jeroen Breckpot,[1,4] Koenraad Devriendt,[1,4] Bernard Thienpont[1], Joris Robert Vermeesch,[1] Isabelle Cleynen,[1] Toomas Kivisild,[1,7]

1 - Department of Human Genetics, KU Leuven, Leuven 3000, Belgium, 2 - Department of Archaeology and Museology, Masaryk University, Brno, Czech Republic, 3 - Center of Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czech Republic, 4 - Center for Human Genetics, University Hospitals Leuven, University of Leuven, Leuven, Belgium, 5 - Laboratory of Neurobiology, Neuroscience Department, KU Leuven, Leuven 3000, Belgium, 6 - Neurology Department, University Hospitals Leuven, Leuven 3000, Belgium, 7 - Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia

Corresponding Authors: Prof. Toomas Kivisild, toomas.kivisild@kuleuven.be

Running title: Genotype Imputation from Low-Coverage Data

ABSTRACT

Genotype imputation from low-pass sequencing data presents unique opportunities for genomic analyses but comes with specific challenges. In this study, we explore the impact of quality filters on genetic ancestry and Polygenic Score (PGS) estimation after imputing 32,769 low-pass genome wide sequences (LPS) from non-invasive prenatal screening (NIPS) with an average autosomal sequence depth of ~0.15×. In studies involving ultra-low coverage sequences, conventional approaches to secure genotype accuracy may fail, especially when multiple samples are pooled. To enhance the proportion of high-quality genotypes in large datasets we introduce a filtering approach called GDI that combines genotype probability (GP), alternate allele dosage (DS), and INFO score filters. We demonstrate that the imputation tools QUILT and GLIMPSE2 achieve similar accuracy, which is high enough for broad-scale ancestry mapping but insufficient for high resolution Principal Component Analysis (PCA), when applied without filters. With the GDI approach we can achieve quality that is adequate for such purposes. Furthermore, we explored the impact of imputation errors, choice of variants and filtering methods on PGS prediction for height in

1,911 subjects with height data. We show that polygenic scores predict 23.7% of variance in height in our imputed data and that, contrary to the effect on PCA, the GDI filter does not improve the performance of PGS in height prediction. These results highlight that imputed LPS data can be leveraged for further biomedical and population genetic use but there is a need to consider each downstream analysis tool individually for its imputation quality thresholds and filtering requirements.

INTRODUCTION

Until recently, genotyping arrays were the only cost-effective approach for generating data for medical genomics at the scale of tens of thousands to millions of individuals. With decreasing costs of genome sequencing, low-pass (at coverage <1×) whole-genome sequencing followed by imputation has become a cost-effective alternative for trait mapping and estimation of polygenic scores (PGS) (Martin et al. 2021; Wasik et al. 2021). For evolutionary and population genetic studies, sequencing of large numbers of individuals at low depth provides a more comprehensive representation of variation in the population than sequencing of a small number of individuals at a higher depth (Fumagalli 2013). Importantly, imputation performed at sufficient accuracy can offer opportunities to reuse large volumes of already generated data. Here, we explore the prospects to leverage hundreds of thousands of LPS genomes generated from cell-free DNA (cfDNA) during Non-Invasive Prenatal Screening (NIPS), (Wang et al. 2021; Van Riel, Stanley, and Vermeesch 2023), for the study of maternal genotypes via imputation.

Non-Invasive Prenatal Screening (NIPS) has become an integral part of prenatal care in many countries, though adoption rates vary significantly across Europe. In most European countries, fewer than 25% of individuals adopt NIPS, and in many cases, the rate is below 5% (Gadsbøll et al. 2020). However, countries such as Italy, Spain, Austria, the Netherlands, and Belgium demonstrate higher uptake. Belgium, in particular, stands out due to its progressive reimbursement policy for NIPS, which has facilitated its widespread accessibility and seamless integration into prenatal care (Bayindir et al., 2015; Van Den Bogaert et al., 2022). Approximately 80% of NIPS

performed in Belgium rely on low-pass genome-wide sequencing, making this technology a cornerstone of prenatal care in the country. More than 200,000 LPS genomes have been generated so far for NIPS in Belgium alone.

Imputation of maternal genotypes from NIPS data can offer unique genomic medicine opportunities beyond fetal aneuploidy screening at the population level, such as genome-wide association studies (GWAS) and PGS profiling of large cohorts. In this study, we focus on imputation strategies scalable to large cohorts that ensure sufficient accuracy for downstream purposes by exploring suitable post-imputation filtering strategies. In our study, we demonstrate how LPS data from NIPS can be repurposed to support diverse biological analyses, showcasing the broader potential of cfDNA-derived LPS data.

Several tools enable efficient and accurate genotype imputation from LPS data. Among these, Beagle (Browning, Zhou, and Browning 2018) and Gencove's loimpute software (Wasik et al., 2019) have been some of the most popular examples. However, they present limitations in scalability for large cohorts and increased computational costs with large reference panels. Recently, imputation methods like GLIMPSE (Rubinacci et al. 2021), GLIMPSE2 (Rubinacci et al. 2023), and QUILT (Davies et al. 2021), scalable for imputation tasks of tens of thousands of low-coverage genomes, have been developed and benchmarked against other tools, showing high accuracy and low computation costs with reference panels of tens of thousands individuals or more.

The minimum coverage from which accurate genotype imputation can be achieved may depend on various factors, including the sample quality (read length, contamination, damage), variant frequency and representation in the reference panel, composition and size of reference panels, as well as the error sensitivity of downstream analytical methods used. Gamba et al. (2014) showed

that imputation of common variants from European ancient genomes of 1× coverage can be performed at 99% accuracy when filtering out a small proportion of variants that have low genotype probability (GP<0.99). Accuracy higher than 90% in down-sampled ancient genomes was achievable also from 0.1× coverage with the cost of filtering out more variants (Hui et al. 2020). Using the 1000 Genomes Project reference panel and GLIMPSE, Sousa Da Mota et al. (2023) showed that the GP≥0.99 filter retained only 20-43% of the correctly imputed variants per individual sample in low coverage samples. Individual-level filtering will thus produce, even in small sample sizes, substantial cumulative missingness leading to a drastic reduction in the number of sites available for downstream analyses. In search of the optimal balance between data loss and quality with large batches of data, we explore different filtering options and quality metrics of individuals and variants at the batch level. To do so, we implemented a filtering strategy called GDI that incorporates filters on different post-imputation metrics: the posterior genotype probability, the alternate allele dosage, and the INFO score, a value to measure the imputation accuracy of each imputed variant that some tools for imputation usually emit (Marchini & Howie, 2010).

Genotypes imputed without reference panels from NIPS sequence data from large cohorts of Chinese individuals have been used for the study of population structure in China, for replicating and finding new genome-wide associations with height, BMI, twinning, and gestational diabetes, and for calculation of polygenic scores (Homburger et al. 2019; Liu et al. 2018; Liu et al. 2023; Zhen et al. 2024). Similarly, a GWAS approach applied on imputed LPS data of 140,000 NIPS profiles from the Netherlands has been successful in identifying loci that affect the concentration and fragmentation properties of cell-free DNA in plasma (Linthorst et al. 2024). However, the full potential of this approach is yet to be revealed. Because of the availability of large haplotype

reference panels, such as HRC (the Haplotype Reference Consortium. 2016) with 27,165 individuals, imputation of European samples can be achieved with increased accuracy (Davies et al. 2021; Rubinacci et al. 2021; Rubinacci et al. 2023). Yet, the impact of imputation errors, particularly in the low minor allele frequency classes, on downstream population genetic applications, such as PCA, or medical applications, such as PGS analysis, is not fully known. A PGS estimates an individual's genetic risk to develop a certain complex trait or disease by combining trait-associated variants into one predictive score. Adult height is a complex trait that is highly polygenic in nature, with 12,111 independent Single Nucleotide Polymorphisms (SNPs) collectively accounting for 40-45% of phenotypic variance in populations of European ancestry (Yengo et al. 2022). Its high heritability, together with the fact that height is easily measured, makes it an attractive model trait to evaluate the usability of LPS data with different imputation options for genome-wide association scanning and PGS calculation. Here, we assess the accuracy of PGS calculation from low-coverage imputed NIPS data for height and investigate the effect of different filtering strategies, including the GDI filter on the correlation between predicted and measured trait values.

In this study, we test the performance of QUILT (Davies et al. 2021) in genotype imputation on large batches (>1,000 individuals) of LPS data (also referred to as NIPS data in the text) with an average coverage of ~0.15x. As small insertions and deletions can be imputed with lower accuracy (Nguyen et al. 2024), require separate handling (Rubinacci et al. 2023), and are omitted by QUILT (Davies et al. 2021), we focus our analyses only on single nucleotide substitutions. We impute the genotypes of 32,769 Belgian individuals and test the effect of different reference panels and filtering steps on imputation accuracy of a single individual or a batch of many individuals. Using the GDI approach we evaluate the effects of the imputation quality filters on downstream analyses, including Principal Component Analysis (PCA) and PGS.

RESULTS

Imputation accuracy from LPS data with QUILT and GLIMPSE2

We first evaluated the performance of QUILT (Davies et al. 2021) on LPS data through sensitivity values across multiple MAF bins for sites imputed as homozygous for the reference (REF) or the alternate allele (ALT), and for heterozygous sites (HET) that tend to be more challenging to impute (Ausmees et al. 2022). Additionally, we also assessed data filtered for maximum genotype probability (max(GP) ≥0.99), a commonly employed approach to enhance the quality of imputed data (Gamba et al. 2014; Hui et al. 2020), especially effective when working with few or individual samples. We observed (Supplementary Figure 1) that the average sensitivity of raw imputed data across three test samples varies among different MAF bins, with clear improvements after applying a GP filter.

As detailed in Supplementary Table 1, the HET sensitivity shows an average improvement of 9.3% across all observed MAF bins after GP filtering that allows sensitivity values to surpass ~97% for common variants (MAF >0.05). Similarly, the ALT sensitivity, which already exhibits values exceeding ~93% across all MAF bins in raw imputed results, demonstrates an average improvement of 3.33%, reaching values higher than ~97% across all MAF bins post GP filter application. Moreover, the REF sensitivity presents an average improvement of 1.8%, with values surpassing ~98% across all MAF bins following GP filter application.

To provide a comparison with other available imputation tools, we also imputed the three test samples with GLIMPSE (Rubinacci et al. 2021) and GLIMPSE2 (Rubinacci et al. 2023). The comparison between QUILT and GLIMPSE (Figure S2, panel A) highlighted that QUILT provides a

higher number of correctly imputed genotypes (especially for heterozygous genotypes) after the application of a GP filter max(GP) ≥0.99 (Supplementary Table 2). We observe that sensitivity values show superior performance in QUILT for MAF bins below 5%, while for MAF bins exceeding 5%, QUILT maintains an average sensitivity value of 0.9895 (SD±0.28) across the three test samples. This is lower than with GLIMPSE, which presents an average sensitivity value of 0.994 (SD±0.14) in the same frequency category (MAF >5%). Therefore, while GLIMPSE is, on average, providing an additional 0.0045 sensitivity for the MAF bin including common variants and a lower number of correctly imputed genotypes in any of the observed MAF bins, QUILT presents a better compromise between the number of correctly imputed genotypes and their quality. In the comparison with GLIMPSE2 (Figure S2, panel B), we observe that in each genotype category both tools tend to display similar trends, both for the raw imputed and the GP-filtered data. Furthermore, we also tested a dual-step imputation with Beagle 5 (Browning, Zhou, and Browning 2018) being the second imputation step after applying GLIMPSE or GLIMPSE2. With this test we wanted to see whether using Beagle 5 after running GLIMPSE or GLIMPSE2 could increase the quality of the already imputed data. Overall, we observed that a dual-step imputation (Figure S2 C and D), with a GP filter after the first step, resulted in an increased number of correctly imputed genotypes but at the expense of reduced sensitivity compared to QUILT.

Based on these results, we found that among all the options tested, QUILT and GLIMPSE2 presented the highest performance levels, showing similar quality outcomes. Additionally, we evaluated the effectiveness of a GP filter in enhancing sensitivity values for various MAF bins and

genotype categories at an individual level. However, GLIMPSE2 was not yet available at the time of data production, and QUILT was chosen for the analysis of the bulk of the data. For details on the methodology, please refer to the Materials and Methods section. The pipeline design we employed is presented in Figure 1, panel A.

In terms of computational efficiency, a comparative analysis of QUILT and GLIMPSE2 showed that the imputation of one NIPS genome on a single core with QUILT, using a reference panel that had been pre-processed, took 21 hours and 38 minutes, compared to 11 hours and 23 minutes with GLIMPSE2 (a ratio of approximately 1:2). This difference is smaller than the results reported by Rubinacci et al. who observed an approximate 6-fold difference in computation cost in favor of GLIMPSE2 when using the HRC panel (Rubinacci et al. 2023).

An additional observation drawn from our tests is that regardless of the imputation tool used, a negative correlation is consistently observed between fetal fraction (SeqFF) and sensitivity (Figure S3A), while a positive correlation exists between coverage and sensitivity (Figure S3B).

Selection of the reference panel

We compared the effect of different reference panels on imputation accuracy of NIPS data with QUILT, using the three test samples. Specifically, we compared the 1000 Genomes reference panel of 2,504 individuals and 32,140,179 variants, and the HRC panel of 27,165 individuals and 36,258,911 variants (Table 1, Table S3). While the sensitivities of heterozygote calls of common variants (MAF >0.05) are similarly high (>0.965) with both panels, we find that the use of the bigger HRC panel enables us to retain more correctly imputed common variants. In total, >65,000 heterozygous variants passed the GP ≥0.99 filter with HRC among variants with MAF >0.05, which is ~5% more than with the 1000 Genomes Project panel. In line with this, the total number of

missing genotype calls was greater after the GP ≥0.99 filters with the 1000 Genomes Project panel.

Cumulative missingness

While the GP filter applied at the individual level can clearly help to enhance imputation accuracy, it introduces cumulative missingness when combining individual data. With large batches, the cumulative missingness can result in retention of only a small number, if any, of the sites that have GP value ≥0.99. Supplementary Figure 4A shows the distribution of variants removed by the GP filter per sample, highlighting the high degree of variance in the batch of imputed NIPS samples, with the majority of the individuals having 20-50% of the variants removed, with an average of approximately 30%. In the full set of 32,769 NIPS individuals we found there are no variants that pass the GP filter in all samples. The minimum proportion of missing calls per variant we observed was 0.007 (210 individuals with GP <0.99), with the distribution of the missingness by variants per bins of individuals not passing the filter shown in Supplementary Figure 4B.

To address the issue of cumulative missingness, we attempted first to filter out variants with an INFO score below 0.4 across all imputed batches, but this proved insufficient in removing the batch effect from high-level PCA, where we observed the clustering together of all our imputed NIPS batches separately from the reference data (Figure S5 A). Increasing the threshold to an INFO score below 0.6 did not improve the result (Figure S5 B). Hence, we adopted a more refined approach, we call GDI filtering, that removes only the most problematic sites by integrating multiple post-imputation quality metrics, including posterior genotype probability, alternate allele dosage,

and the INFO score from QUILT (Figure 1, panel B). We applied GDI filtering dynamically on our data to optimize the maximum data quality with minimal loss of variants observing a range of metrics of imputation accuracy in light of the performance of downstream analyses tools, PCA and PGS.

Data filtering

Exclusion of Related and Duplicate Samples

We used QUILT to impute 32,769 samples in batches of a maximum of 1,250 individuals (Figure 1, panel A), as detailed in the Materials and Methods section. The first step after imputation was to identify and remove duplicates and related individuals with the aim of avoiding possible allele frequency biases in the downstream analyses. Duplicates may result from separate tests for different or the same pregnancies. For the identification and removal of duplicates and closer than third degree related individuals, we utilized IBIS (Seidman et al. 2020), which employs the same bounds as KING (Manichaikul et al. 2010) for the degree of relatedness. A kinship coefficient value of 0.088 defines the level of relatedness expected when members of the same family are within the same healthcare system (Johnson et al. 2022). However, we applied a stricter threshold, considering any pair of individuals with a kinship coefficient above 0.044 (up to third-degree relatives) as related. Notably, when cross-checking our results with additional metadata obtained at a later stage of the study, we found that our results with IBIS matched more than 90% of the recorded information related to duplicates and relatives.

GDI validation

We compared the performance of QUILT across different post-imputation statistics on the three test samples with high coverage data (mean coverage ~44×) (Figure 2). Without any filters (raw results in Figure 2), QUILT shows the lowest accuracy estimates for each of the statistics we examined.

A notable improvement over unfiltered results was observed when applying INFO score filtering alone (green line in Figure 2). Since the use of INFO score filtering with QUILT had not been previously explored, we applied the method from Liu et al., originally designed for STITCH (Davies et al. 2016), to evaluate INFO scores with QUILT. The most notable comparison is between the GDI filtering approach and the GP-based filtering (max(GP) ≥0.99) (yellow versus purple line in Figure 2). It is evident that both filtering approaches resulted in an improvement compared to the raw data in all the statistics. Our tests confirmed that applying a GP filter at the individual level enhances data quality. Conversely, the GDI filter shows a smaller improvement but is crucial for retaining the majority of variant sites in large batches of data, thereby avoiding the introduction of any missing data, as shown below in case of PCA analyses. Indeed, the GP filtering approach can be effectively used only on a single genome or a few genomes, since in larger batches, most, if not all, variants would have missing genotype calls introduced by the GP filter in at least some individuals. To illustrate the difference on a small example, Figure 3 shows results from a batch of three individual NIPS genomes: the GDI filter retains 83% of variants, while the GP filter, applied cumulatively, retains only 49.2% of total sites. These differences increase with larger sample sizes. In other words, Figure 3 illustrates on a small sample that GP filtering at individual level leads to massive cumulative missingness when the data is pooled. In contrast, the GDI approach maintains a higher proportion of non-missing variants across the batch, avoiding these issues.

To validate the effect of the GDI filtering approach on imputation accuracy in an independent dataset, we used LPS and high-quality genotype data from 138 non-pregnant healthy controls that

were not part of the GDI filter development. Figure 4 summarizes the results of 138 individuals, showing that for all MAF bins > 0.05, the GDI filtered LPS data had higher heterozygous sensitivity than the raw data. Furthermore, the GDI filtered data scored better in other accuracy statistics as well, such as heterozygous specificity and the non-reference concordance (Supp. Table 5). Metrics from these additional samples validated our observations from the three NIPS samples. Moreover, these results demonstrate that the GDI filter is robust when applied to other datasets.

GDI application

In light of the success of the GDI filtering approach in minimizing the presence of low-quality variants across the largest possible subset of samples, the 28,512 NIPS samples that passed initial quality checks and removal of related samples were further processed using this method. We removed an additional 1,158 outliers based on the observed distribution of low-quality variants (LQV scores, defined by the proportion of sites not passing the thresholds set for GP values, DS values, and the INFO score provided by QUILT) (Figure S6). Notably, the average coverage for these 1,158 samples is approximately 0.09×, confirming that the LQV metric effectively identified low-coverage samples. These samples were classified as outliers based on their high LQV levels prior to the application of GDI.

Furthermore, the distribution of SeqFF values revealed that samples removed due to the LQV filter exhibited a higher median SeqFF compared to those retained for further analyses (Figure S7). Specifically, the median SeqFF for LQV samples was 10.1, whereas for retained samples, it was 8.99 (Wilcoxon rank-sum test p-value = 4.147×10^{-13}). This confirmed the negative effect of SeqFF on imputation accuracy.

Evaluation of GDI effectiveness across coverage levels

To evaluate the effectiveness of the GDI thresholds (INFO, GP, and DS) across different sequencing coverage levels, we grouped the samples into coverage bins and assessed the difference in data quality (LQV) before and after GDI application (Figure S8). The analysis revealed

a consistent positive impact on accuracy of the GDI method across all coverage bins, with the most pronounced effect observed at lower coverage levels. For bins in the range ≥0.1 to <0.2x, the median indicates that the differences between pre- and post-GDI are more uniformly distributed. However, the internal variability is quite high, with some samples showing much larger improvements (outliers), which likely contributes to the observed overall uniformity in this bin. A similar trend is observed in the ≥0.2 to <0.3x bin, where the differences remain generally uniform but with no extreme outliers. In the ≥0.3 bins, the median indicates that the differences between pre- and post-GDI are generally modest or small. It is important to note that within each coverage bin, especially for those with coverage ranges of ≥0.1 to <0.2x and ≥0.2 to <0.3x, the large number of samples results in significant variability in the individual sample coverage. This variability may influence the observed results and should be considered when interpreting the data.

PCA analyses

The final set of 27,354 samples was used alongside 498 samples from the Genome of the Netherlands (GoNL) (The Genome of the Netherlands Consortium 2014), 3,643 samples from the Project MinE (Project MinE ALS Sequencing Consortium 2018), and 2,495 samples from the 1000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium et al. 2015) to visually explore the distribution of genetic ancestry in the imputed data with PCA and UMAP approach, aiming to assess the performance of downstream population genetic analysis method and to test whether imputation is causing batch effects in the data.

We applied UMAP on 20 PCs estimated from the data as this approach has been shown to enable the detection of regional clusters at higher resolution than the examination of only the first two PCs (Diaz-Papkovich, Anderson-Trocmé, and Gravel 2021). Without the GDI filter (raw imputed data), a

clear batch effect can be observed in the UMAP (Figure S9, panel A); as expected, as UMAP tends to allocate more space to the most represented group (Diaz-Papkovich, Anderson-Trocmé, and Gravel 2021), LPS samples occupy most of the space, but no clear structure in the data can be observed. A similar result is observed when plotting a UMAP after applying a filter on the INFO score (<0.4) (Figure S5); the plot still shows a strong batch effect, similar to that seen with the raw data in Figure S9. This suggests that, although filtering based on the INFO score improves various post-imputation statistics, it does not effectively remove the batch effect. However, applying a GDI filter and removing variants with low quality in more than 30% of the samples (Figure S9, panel B) effectively helps to remove this bias and reorganize the data more clearly. In particular, we can see that the larger group predominantly composed of LPS samples (bigger gray cloud in Figure S9, panel A), is merged after filtering with the smaller separate cluster including individuals from the 1000 Genomes Project (GBR and CEU), most of the MinE dataset, and the GoNL group. Furthermore, part of the LPS samples is clearly clustering with non-European groups. This batch effect is also detectable in a milder form in the plot of PC1 and PC2 (Figure S9, panel C), where the distribution of the Belgian NIPS samples appears not completely aligned with the distribution of the other datasets of individuals with similar ancestry (MiNE and GoNL).

As a general observation, the removal of variants with low quality scores led to the elimination of approximately 1,9 million variants, reducing the overall dataset of common variants to 3,355,663 variants. Following this removal, an average reduction of 54.1% was observed in the LQV scores compared to pre-filter values (Figure S6). Upon removal of the low quality variants the imputed LPS data appears to include less individuals outside the triangular structure of the PC plot (Figure S9, panels C versus D).

Genetic ancestry and population structure in Belgian NIPS data

As genetic ancestry is an important confounding factor for many medical and population genetic analysis tools, it is important to know how sensitive methods that cluster individuals by their ancestry are to imputation quality. When focusing on the results of the first two PCs (Figure 5A), we first confirmed that imputed NIPS samples exhibit a wide distribution in context of global

reference data: most of the samples are primarily clustered along the European edge of the plot, representing 89.17% of the imputed data, while the remaining 10.83% showed dispersion towards the African and the East Asian edges of the plot.

These results from PCA were also confirmed with supervised admixture, where 87.3% of the imputed NIPS samples had a European ancestry proportion of >= 90%. Within the subset of NIPS data where country of origin was known, 165 individuals from Iran formed a cluster right below the European edge in the PCA, which becomes more distinct in the UMAP conversion (Figure 5B). Similarly, after applying the GDI filter, within the European subset a cluster of 131 individuals from Poland becomes more clearly pronounced (Figure 5B and Figure S9). The majority of NIPS individuals whose birthplace was identified as being in Belgium mapped together with MinE and GoNL individuals on PC plots (Figure S10), and between IBS, CEU and GBR populations of the 1000 GP data, consistent with previous observations (Van Den Eynden et al. 2018).

PGS for height

For 1,911 imputed NIPS samples, polygenic scores (PGS) for height were calculated at different p-value thresholds (Pt) using four approaches: (i) data filtered by GDI and MAF ≥5%, (ii) data filtered by MAF ≥5% only, (iii) data filtered by MAF ≥1%, and (iv) data limited to HapMap SNPs. These are referred to as PGS_{GDI}, $PGS_{MAF5\%}$, $PGS_{MAF1\%}$, and PGS_{HapMap}, respectively. When performing linear regression, where the height of an individual is predicted by PGS, PGS_{GDI} consistently had a lower variance explained for all Pt values than the other PGSs (Figure 6).

Whereas the three other PGSs mostly had similar values (Table S4). Overall, the best PGS model was PGS_{HapMap}, with a variance explained of 0.237 at Pt =0.1, including 62.059 SNPs. However, it is worth noting that the variance explained for PGS_{HapMap}, $PGS_{MAF5\%}$ and $PGS_{MAF1\%}$ doesn't change

much from Pt =0.05 onward, indicating that the addition of SNPs with P-value >0.05 does not have much added value towards prediction. We also calculated the Spearman's correlation coefficient (r) of the correlation between genotypic (i.e. the PGS) and phenotypic height. For PGS_{HapMap}, $PGS_{MAF5\%}$ and $PGS_{MAF1\%}$ the correlation was the same, r=0.46. This is in contrast with PGS_{GDI} where r =0.36. Together, these findings indicate that filtering the imputed NIPS data with the GDI filter does not improve, at least in case of height, PGS performance in terms of either variance explained or correlation with phenotypic height.

DISCUSSION

In this study, we explored the potential of genotype imputation of large cohorts of low-pass (0.1-0.3×) sequence data from Belgium, for the population genetic study of genetic ancestry and PGS estimation for medical risk stratification. In tests performed on an individual level, we found that imputation can be achieved with comparable quality using QUILT (Davies et al. 2021) and GLIMPSE2 (Rubinacci et al. 2023). When using ultra-low coverage (<0.2×) sequences without quality filters, imputation accuracy appeared, however, to be too low for applications, such as PCA. Our tests confirmed that imputed genotypes from 0.1-0.2× sequence data generated with QUILT and GLIMPSE2 without further quality filters exhibit heterozygote sensitivity less than 0.95. For downstream applications that require higher imputation accuracy, an alternative approach of applying quality filters, such as GP, separately on each individual leads to high (>0.98) individual sensitivity scores but also introduces cumulative missingness when working on batches of data. As a solution we introduced batch level filtering approach GDI that optimizes the balance imputation accuracy and the retention of the largest possible subset of samples and variants. A similar result could not be achieved by filtering solely based on INFO score (<0.4), as this approach did not resolve the batch effect observed in the UMAP. However, our results confirmed the validity of using INFO score filtering as a foundation for improving data quality, which is why we adopted this metric as the first step in our GDI filtering strategy. While we showed that the GDI approach was effective in removing batch effects caused by low imputation accuracy in low coverage cohorts seen in PC analyses, the GDI approach did not improve the PGS prediction of height. We observed negative

correlation between fetal fraction and sensitivity across the range of different coverage values (Figure S3) suggesting that although higher coverage ensures higher imputation accuracy, this relationship may be compromised in cases of high fetal fraction.

The best PGS for height was PGS_{HapMap} with a variance explained of 0.237 and a correlation with phenotypic height of 0.46 (Figure 8). $PGS_{MAF5\%}$ and $PGS_{MAF1\%}$ had similar results. In contrast, with PGS_{GDI} only 14% of phenotypic variance of height could be explained (Table S4). A possible reason for this finding could be that there were less SNPs available to build the PGS compared to data that was only MAF 1% or MAF 5% filtered. However, the SNP count alone cannot fully explain the observed difference in variance explained and correlation. This was demonstrated by PGS_{HapMap}, which had the best performance even though there were less SNPs in the initial dataset than the GDI filtered data. On the other hand, it is important to note that even though in the HapMap filtered dataset there are less SNPs than in the GDI dataset, these SNPs have been chosen to capture most of the common variability in the human genome (The International HapMap Consortium 2005). Common SNPs, which are predominantly the signals detected by GWAS, are consequently the ones used for PGS. Together, this could explain why the PGS_{GDI} is not performing as well as the other filtered datasets. Of note, we only calculated a PGS for the complex trait height, which is highly polygenic – or even omnigenic – in nature (Yengo et al. 2022; Boyle, Li, and Pritchard 2017). Hence, these results could differ for other traits or diseases that have a different genetic architecture. One commonly observed cause of the drop of variance explained in PGS models is the mismatch between genetic ancestry of the test cohort and the GWAS cohort on which the GPS model is based on. However, we cannot think of a reason why the GDI filter would remove specifically variants that are informative for height PGS in Europeans. Furthermore, the imputation reference panels we used were chosen to maximize the success of genotype imputation in a cohort of predominantly European ancestry.

When selecting the most appropriate reference panel for imputation, we tried to ensure that it would be representative of all major genetic ancestry components of the sampled population. As the samples were predominantly from Flemish hospitals, we expected the majority of our cohort to have European ancestry with the addition of ~10% of Asian or African ancestry. We compared the

performance of a global reference panel of the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015) and the Haplotype Reference Consortium (HRC) panel (the Haplotype Reference Consortium 2016) which includes the former and is enriched for individuals with predominantly European ancestry. Given the larger number of individual sequences and the number of variants the HRC panel provided, we achieved a higher imputation accuracy and found a smaller number of variants that had to be filtered out (Table 1, Tables S2-3). We observed the greatest improvement of accuracy for variants with MAF 0.01 to 0.05 which may have importance for downstream applications that consider variants in this low frequency spectrum. The imputation accuracy of rare variants (MAF <0.01) remained too low with both panels and filter settings for any downstream purposes examined in this study.

One possible strategy to improve imputation quality at the batch level is to apply filtering by summary quality measures calculated for the entire batch. INFO score is the only such measure available in the QUILT output. However, with large cohorts that are constantly being supplemented with new data, such as the 32,769 individuals examined here, it can be preferable for practical reasons to carry out imputations in smaller batches and to merge them in the end for a common set of variants. Only 131,908 variants (2.4% of common variants) had INFO score below the 0.4 threshold in all the batches of NIPS data we studied, which offered only a minor improvement of quality over the option of using no filters. Furthermore, we observed that incorrectly imputed variants included many with high INFO score, a subset of which had a wrongly imputed genotype while their GP and DS values pointed consistently to the genotype called from the high coverage data. This motivated us to carry out systematic revisions of conflicts between the GT and GP values, which could be characteristic only to ultra-low coverage data, and to focus in our search for optimal solutions in the GDI design on the combination of info score, GP and DS filters. We acknowledge that in ideal circumstances sequencing to the minimum depth of 0.5× should be encouraged to reach sufficient accuracy for downstream applications that would not require further filtering (Sousa de Mota et al. 2023). However, in case of data sets that have already been generated in the past, and/or where it is challenging, if not impossible, to generate more data, the GDI approach may offer leverages for data that otherwise cannot be used for genomic scale

genotype analyses. Application of the GDI filter after imputation enabled us to achieve imputation accuracy in the large batch of 32,769 individuals on par, though less superior than that achieved with the GP filter applied at individual level (Figure 2). However, the advantage represented by the GDI approach over the GP filter at a batch level was significant as It allowed us to prevent the cumulative introduction of missing data by each individual, keeping the majority (83%) of the imputed common variants for downstream analyses. Furthermore, when we assessed the effectiveness of GDI across varying coverage levels, we observed that the method generally improved data quality in all coverage categories. The greatest improvements were noted in samples with lower coverage, while higher coverage bins displayed more variability in the degree of improvement. This variability may reflect factors such as differences in sample characteristics and coverage distribution within each bin. These results demonstrate the broad applicability of GDI, while suggesting that fine-tuning of the thresholds may be needed for datasets with higher coverage.

The study by Sousa de Mota (2023) showed that individuals imputed from 0.1-0.25× coverage were correctly placed in the expected continental clusters in PCA while showing at the same time significant deviations from the precise placement within those clusters. Our results confirm this: when using imputed genotype data without filters as the input we find that the majority of individuals map in the PC1 and PC2 plot to the context of the reference sources from Belgium and the Netherlands, as expected, while in the higher resolution UMAP analyses of P1-20 data we find that the unfiltered data cluster separately from the others (Figure S9). With our final settings of the GDI filter, which retained 3,355,663 out of 5,421,789 common variants (61.9%), we were able to remove the separate clustering of NIPS data as a batch. Regarding the determination of the cutoff for variant removal in the GDI filter, we leveraged visual exploration of the sample distribution on a UMAP plot after a series of filtering steps, with increments of 10% of LQV sites being removed at each step, to guide our decision-making. However, in scenarios where such visual exploration is not feasible, the fundamental principle guiding this approach is to strike a balance between being overly conservative and excessively aggressive in variant removal, avoiding the loss of a substantial number of variants. While it might seem arbitrary, custom-oriented handling of this step

is essential, recognizing that, in principle, each combination of a dataset and a downstream data analysis method can be unique and necessitate tailored consideration.

As we sought a solution for using 0.1-0.3× sequence data in ancestry analyses we developed the specific combination GDI filters to be optimized for the performance of high resolution PCA on Belgian NIPS data and the same parameter settings we used may not offer the most optimal solution in other settings. We expect, however, that when appropriate reference panels are available, e.g. for other European cohorts with similar coverage data, the GDI filtering following QUILT (or GLIMPSE) imputation can be a useful approach for genetic ancestry analyses with PCA. Other downstream analyses, including those focused on PGS, can, most likely, benefit from individual filter optimizations that are different, either more conservative or stringent, from those described and explored here. In cases where increasing the sequence coverage is not an option, a generic strategy that may prove effective should involve exploration of the distribution of the number of low-quality variants against the number of samples in which they fail, aiding in defining the most appropriate cutoffs and optimal reduction in LQV scores. A further point worth stressing here is that since the imputation of heterozygous sites is always the most challenging, it is highly likely that such sites are the ones with lower GP and DS values and therefore do not pass the GDI filter. As a consequence, more stringent filters can lead to a reduction of heterozygosity in the sample.

Applying a genome-wide SNP array of around 300,000 sites on 189 individual samples from Belgium, Van den Eynden et al. (2018) showed typical European genetic constitution of the Belgian population. Our exploratory analyses of the downstream effects of imputation quality filters on principal component analysis have facilitated the study of the genetic ancestry of 32,769 pregnant individuals recently collected from a hospital in present-day Flanders at the genomic scale for more than 3 Million common variants. This has provided a higher resolution view on the genetic ancestry of present-day Flanders and has enabled better quantification of the extent of recent migration. The majority of the samples group within regional clusters of Belgian and Dutch individuals from the MinE and GoNL projects, while a minor subset clusters with other European sub-groups and individuals with non-European genetic background. These results are not

surprising given Belgium's long history of immigration and integration. Historically, Belgium is known to have been a nation characterized by persistent immigration, a phenomenon that has undoubtedly contributed significantly to shaping its demographic landscape. Indeed, this continuous demographic evolution has increasingly enriched what is now Belgium's population, which to date welcomes individuals from different corners of the world (Martiniello 2013). Notably, 10.83% of the imputed samples mapped outside the main European cluster in PCA (Figure 4A), which is similar to a recent estimate according to which 11% of the population living in the Flemish Region was born outside the EU. Given that our samples come primarily from Flemish hospitals, they represent the demographic diversity of individuals residing in Belgium, including not only those of Belgian origin but also a broader spectrum. As a result, our findings offer an accurate portrayal of the diverse population found in contemporary Belgium, truly reflecting the cosmopolitan nature of the country and providing a realistic snapshot of its current demographics.

MATERIALS AND METHODS

A concise description of the materials and methods used in this study is provided below. For a comprehensive and detailed version of all protocols and procedures, please refer to the Extended Supplementary Methods.

Data collection

Peripheral blood samples were collected from 32,769 pregnant individuals undergoing NIPS and from 140 non-pregnant healthy controls, with ethics approval from KU Leuven and University Hospitals Leuven (S63253, S66621, S66450). cfDNA was extracted from plasma and sequenced at low-pass coverage (median ~0.15×) using Illumina platforms, following protocols previously described (Bayindir et al. 2015). Sequence data were aligned to hg38, and standard QC, deduplication, and sorting procedures were applied. For the 140 controls, high-quality genotyping was also performed for benchmarking imputation accuracy.

Imputation tool testing

To select the optimal imputation tool, we evaluated GLIMPSE, GLIMPSE2, QUILT, and a dual-step strategy combining either with Beagle. Imputation accuracy was benchmarked using three low-pass NIPS samples (~0.16×) with known fetal fractions and corresponding high-coverage whole-blood genomes. All tools were applied using the HRC reference panel, and accuracy was assessed per genotype class and MAF bin using metrics such as sensitivity and dosage r^2.

Reference panel setup

All imputation strategies used either the HRC or 1000 Genomes Project reference panels, both of which were preprocessed using standard liftover and variant filtering procedures.

Imputation with QUILT

Genome-wide genotype imputation of the 32,769 samples was performed using QUILT with the HRC panel. Samples were processed in parallel batches using a custom pipeline (Figure 1A), with chromosomes divided into 5 Mb windows. Post-imputation filtering retained 5.42 million variants with MAF >5% in the HRC panel for downstream analysis.

Post imputation filters

Removal of duplicates and related individuals

Related individuals were identified and removed based on kinship coefficients estimated with IBIS, resulting in the exclusion of 4,257 samples.

Filter on variants at batch level

To enhance the quality of the imputed data, we implemented a filtering strategy we called GDI (Figure 1, panel B) that combines filters on genotype-related metrics, including the posterior genotype probability (GP), alternate allele dosage (DS) and the INFO score provided by QUILT. The INFO score provides a quantitative measure of the certainty associated with genotype imputation based on the distribution and uniformity of genotype posteriors; a low score indicates a flat, non-informative distribution, while a score near 1 suggests concentrated, confident genotypes. The DS field represents the expected number of alternate alleles for a given genotype. For diploid genotypes, the DS values should range from 0 to 2. Ideally, a DS equal 0 means that both alleles are the reference allele (0/0), a DS equal 1 means that one allele is the reference allele and the other is the alternate allele (0/1), and a DS equal 2 means that both alleles are the alternate allele (1/1). However, it is possible for the DS values to assume intermediate values between 0 and 2 for diploid genotypes. This can happen when there is uncertainty in the genotype call, such as when the genotype posterior probabilities in the GP field are not clearly in favor of one genotype over another. In this light, the dosage value provides an indication of how well the genotype is supported by imputation. As for the GP field, it represents a measure of how likely each possible genotype at a site is after imputation, with values closer to 1 indicating a higher likelihood and greater confidence in the accuracy of the predicted genotypes following imputation.

When imputing multiple samples with QUILT, the INFO score associated with each variant represents a consensus score. Due to imputing our samples in separate batches, different INFO scores were obtained for the same variants. To address this variability, we aggregated all INFO scores for the same variants across multiple imputed batches. Variants consistently tagged with an INFO score <0.4 across all batches totaled 131,908 and were consequently identified for removal. The initial step of the GDI strategy involved generating a file containing the GT, GP, and DS fields for each variant and individual. Subsequently, variants tagged for removal from the INFO score screening were excluded from further analysis in subsequent steps. For each sample, a list of variants to be removed is generated, determined by a GP <0.99 and DS ranges defined by the GT field (DS > 0.1 for GT 0/0, DS < 1.8 for GT 1/1, and 0.8 > DS > 1.01 for GT 0/1). The DS thresholds are determined based on theoretical expectations and practical considerations, rather

than empirical testing. These values account for minor variations and provide a margin of safety in variant classification. By setting these thresholds, we aim to ensure accurate classification despite potential fluctuations in DS values. The proportion of variants earmarked for removal defines the fraction of low-quality variants (LQV score) for each sample, and an observation of the distribution based on LQV scores was conducted to establish a cutoff. Samples with an LQV score exceeding 40% were identified as outliers, leading to the removal of a total of 1.158 samples from subsequent analyses.

The initial step concludes by generating, for each sample, a list reporting sites that do not pass the DS and GP filters. In the second step of the GDI strategy, these lists are used to gather information on the quality of sites for all individuals (except those identified as outliers in the first GDI step and those flagged as duplicates or relatives by IBIS). Ultimately, variants exhibiting low quality in more than 30% of the samples were excluded, totaling 1,933,956 variants. The choice of this cutoff was defined by the visual observation of the distribution of the imputed data in the context of a UMAP plot based on 20 PCs including also samples from the 1KGP (lifted over to build GRCh38 from the HRC panel), the MinE (lifted over to build GRCh38 with Crossmap), and the GoNL datasets (lifted over to build GRCh38 with Crossmap). The filtered dataset comprised 27,354 samples and 3,355,663 variants, reflecting a reduction of approximately 11% in the sample count and about 38% of the total variants.

The average reduction of the proportion of LQV sites was calculated by taking the percentage difference for each sample, followed by calculating the mean and standard deviation of these differences. The standard deviation of 6.2% suggests that variations between individual samples and the mean are relatively consistent.

After applying the GDI filter (Figure S12), we observed that 86.4% of the variants had successfully passed all the filters, 4.3% failed the DS filter, 5.1% failed the GP filter, and 4.2% failed both the DS and the GP filters. However, when focusing only on the GP values, we can see that the 90.7%

of the variants remaining in the dataset have a GP ≥0.99, the 3.06% have a GP between 0.9 and 0.99, and only a 6.24% have a GP < 0.9. Overall, the majority of retained variants in our dataset exhibit a GP ≥0.99, highlighting a significant presence of reliable genetic variations. This underscores the effectiveness of the GDI filter in ensuring data integrity and enhancing overall variant quality.

Additionally, when observing the correlation between the LQV scores and the coverage of each sample, we observed that as the coverage increases, the LQV score tends to decrease. This suggests that samples with higher coverage tend to have fewer low-quality variants, indicating a potential association between higher coverage and improved quality of imputed variants. Before applying the filter, we calculated the correlation between the coverage and the LQV scores for each sample and observed a strongly negative value (Spearman's correlation is -0.75). This indicates that higher coverage values correspond to lower LQV scores. In practice, this is suggesting that samples with higher coverage will be associated with a lower proportion of low-quality variants. After the filter, we observed a reduction in the correlation value (Spearman's correlation is -0.66) indicating that despite the reduction in LQV scores, the negative correlation with the coverage persists. In addition, the observed mean LQV score decreased from 0.28 to 0.13 after the application of the filter, indicating a consistent reduction in the proportion of low-quality variants among the samples. Furthermore, the standard deviation changed from ±0.0524 to ±0.0396, indicating an increase of the consistency and homogeneity of the post-filter data. Overall, these results show that the application of the filter affected the distribution of low-quality variants among the samples, contributing to increased uniformity of LQV values and a persistent correlation between coverage and LQV scores, albeit slightly attenuated.

Comparison with other filtering strategies

To assess the effectiveness of our GDI method compared to other filtering strategies, we employed the three test samples and compared the results using different post imputation statistics for the raw imputed data, the GDI strategy, and a direct filter on max(GP) ≥ 0.99. Additionally, we conducted a cross-method comparison by applying imputation through QUILT, GLIMPSE, and GLIMPSE2. Imputation performances were observed over different allele frequency ranges (MAF bins).

Principal components analysis and UMAP

Population structure was assessed through PCA and UMAP embedding using LD-pruned variants across imputed NIPS samples and external references (1KGP, MinE, GoNL).

PGS calculation

To calculate the polygenic scores (PGS), we used the effect sizes provided by the genome-wide association study (GWAS) for height by Yengo et al. (Nature 2022). Duplicate, ambiguous and multi-allelic SNPs were removed and a MAF filter of 1% was applied.

Out of the 32,769 NIPS samples, only 2,698 had a reported height available. After excluding samples because of relatedness or being an outlier (see above), non-European samples were excluded as well. Samples were included as European based on a supervised admixture analysis, where samples with a proportion of >= 0.95 European ancestry were retained. Based on these criteria, 1,911 samples were included for the PGS calculation.

PRSice-2 (Choi and O'Reilly 2019) was used for the calculation of PGS for height for a range of predefined p-value thresholds (Pt) (5 × 10-08, 1 × 10-05, 1 × 10-04, 1 × 10-03, 0.05, 0.1, 0.5, 1)

with the following clumping parameters: distance to both ends from the index SNP = 250 kb, r^2 = 0.1 and p-value threshold = 1. Scores were calculated for individuals from the non-Finnish European 1000 Genomes Project (1KG-NFE), as well as for 1,911 NIPS samples, using the same SNPs per Pt as selected for the 1KG-NFE scores. Four different filtering methods were used to calculate the scores: (1) 1KG-NFE were MAF 5% filtered, then the selected SNPs were used to calculated scores per Pt in the GDI filtered NIPS (PGSGDI); (2) 1KG-NFE were MAF 5% filtered, then the selected SNPs were used to calculate scores per Pt in MAF 5% filtered NIPS (PGSMAF5%); (3) 1KG-NFE were MAF 1% filtered, then the selected SNPs were used to calculate scores per Pt in MAF 1% filtered NIPS (PGSMAF1%); and (4) an overlap was taken between 1KG-NFE and HapMap SNPs, then the selected SNPs were used to calculate scores per Pt in NIPS (PGSHapMap). The 1KG-NFE dataset contained 6,064,728 SNPs, 8,743,364 SNPs, or 1,116,280 SNPs after filtering for MAF 5%, MAF 1% or HapMap SNPs respectively.

After calculation of the raw scores, the first ten principal components (PCs) were regressed out of the scores. Subsequently, the PC-corrected scores were then standardized against the PC-corrected scores from the 1KG-NFE individuals. Specifically, the scores of each individual were standardized by subtracting the mean score of individuals from the 1KG-NFE group from their own scores and then dividing the resulting value by the standard deviation of the scores within 1KG-NFE. When we use the term PGS, we refer to the standardized PC-corrected scores.

Declaration of Generative AI in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT to improve the language and readability of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final version of the manuscript.

Software availability

All source code and custom scripts used in this study are available in the GitHub repositories https://github.com/SABiagini/GDI and https://github.com/SABiagini/PostImputationStats.

DATA access

The NIPS sequence data used in this study are available under restricted and controlled access, in compliance with the GDPR. The genomic data are locally stored at UZ Leuven and access can be obtained with permission from the local UZ Leuven data access committee (DAC: https://www.uzleuven.be/en/dac, dac@uzleuven.be; cc: joris.vermeesch@uzleuven.be). The access process is further described in https://www.uzleuven.be/en/data-access-committee-dac/external-applicant-requests-access-ku-leuven-researchers-dataset. Sequence data of the 140 healthy non-pregnant individuals, used as controls for imputation accuracy, is available through the European Nucleotide Archive (EGA) at EMBL-EBI under Accession No. EGAS50000001114.

Competing interest statement

The authors declare no competing interests.

ACKNOWLEDGMENTS

Our thanks go primarily to all the individuals who participated in this study as donors. We also want to express our gratitude to Robert Davies for the invaluable conversations that helped us determine the most appropriate use of QUILT. Additionally, we are grateful to Ilse Parijs for generously providing the height data of the individuals. This study has been supported by FWO SBO grant S003422N ‘MICADO’ (SB, IC, TK, JV, BT), KU Leuven start-up grant STG/18/021 (TK), KU Leuven BOF-C24 grant ZKD6488 C24M/19/075 (TK, SAB, LH, JB), and FWO grants G050822N (TK) and G0B4822N (BT). MA is supported by a PhD fellowship of FWO Flanders. JB is funded by a senior clinical investigator fellowship of FWO Flanders. SAB was also supported by the Czech Ministry of Education, Youth and Sports (CZ.02.01.01/00/22_008/0004593, RES-HUM: Ready for the Future: Understanding the Long-Term Resilience of Human Culture grant). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government – department EWI, and made available as a Rainbow Project of the Biobanking and BiomolecularResearch Infrastructure Netherlands (BBMRI-NL).

This study makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). a. The LifeLines Cohort Study (http://www.lifelines.nl), and generation and management of GWAS genotype data for it, is supported by the Netherlands Organization of Scientific Research (NWO, grant 175.010.2007.006), the Dutch government's Economic Structure Enhancing Fund (FES), the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, the University Medical Center Groningen, the University of Groningen, the Dutch Kidney Foundation and Dutch Diabetes Research Foundation; and b. For sponsorship of the EMC Ergo Study please refer to (http://www.ergo-onderzoek.nl/wp/); and c. The LUMC Longevity Study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE01014 and IGE05007), the Centre for Medical Systems Biology and the National Institute for Healthy Ageing (Grant 05040202 and 05060810), all in theframework of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.; d. For sponsorship of the VU Netherlands Twin Register please refer towww.tweelingenregister.org. Project MinE Belgium was supported by a grant from IWT (n° 140935), the ALS Liga België, the National Lottery of Belgium and the KU Leuven Opening the Future Fund.

Author contributions: J.R.V., I.C, and T.K. designed and supervised the study. J.R.V., M.A., L.H., T.J., J. B., K. D., and B. T. organized and performed the recruitment of the case samples. P. V. D. provided genomic data of the MinE cohort. S.A.B. and S. B. performed data analyses. S.A.B., T.K., I.C. and J.R.V. drafted and revised the manuscript. All authors reviewed and contributed to the writing of the final manuscript.

REFERENCES

Abraham, Gad, Yixuan Qiu, and Michael Inouye. 2017. “FlashPCA2: Principal Component Analysis of Biobank-Scale Genotype Datasets” ed. Oliver Stegle. *Bioinformatics* 33(17): 2776–78. doi:10.1093/bioinformatics/btx299.

Ausmees, Kristiina, Federico Sanchez-Quinto, Mattias Jakobsson, and Carl Nettelblad. 2022. “An Empirical Evaluation of Genotype Imputation of Ancient DNA” ed. A Sethuraman. *G3 Genes|Genomes|Genetics* 12(6): jkac089. doi:10.1093/g3journal/jkac089.

Bayindir, Baran, Luc Dehaspe, Nathalie Brison, Paul Brady, Simon Ardui, Molka Kammoun, Lars Van Der Veken, et al. 2015. “Noninvasive Prenatal Testing Using a Novel Analysis Pipeline to Screen for All Autosomal Fetal Aneuploidies Improves Pregnancy Management.” European Journal of Human Genetics 23(10): 1286–93. doi:10.1038/ejhg.2014.282.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” Cell 169(7): 1177–86. doi:10.1016/j.cell.2017.05.038.

Browning, Brian L., Ying Zhou, and Sharon R. Browning. 2018. “A One-Penny Imputed Genome from Next-Generation Reference Panels.” *The American Journal of Human Genetics* 103(3): 338–48. doi:10.1016/j.ajhg.2018.07.015.

Choi, Shing Wan, and Paul F O’Reilly. 2019. “PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data.” GigaScience 8(7): giz082. doi:10.1093/gigascience/giz082.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118□; iso-2; iso-3. Fly 6, 80–92. https://doi.org/10.4161/fly.19695

Davies, Robert W, Jonathan Flint, Simon Myers, and Richard Mott. 2016. “Rapid Genotype Imputation from Sequence without Reference Panels.” Nature Genetics 48(8): 965–69. doi:10.1038/ng.3594.

795 Davies, Robert W. et al. 2021. “Rapid Genotype Imputation from Sequence with Reference
796 Panels.” *Nature Genetics* 53(7): 1104–11. doi:10.1038/s41588-021-00877-0.

797 Diaz-Papkovich, Alex, Luke Anderson-Trocmé, and Simon Gravel. 2021. “A Review of UMAP in
798 Population Genetics.” *Journal of Human Genetics* 66(1): 85–91. doi:10.1038/s10038-020-00851-4.

799 Fumagalli, Matteo. 2013. “Assessing the Effect of Sequencing Depth and Sample Size in
800 Population Genetics Inferences” ed. Ludovic Orlando. *PLoS ONE* 8(11): e79667.
801 doi:10.1371/journal.pone.0079667.

802 Gadsbøll, Kasper et al. 2020. “Current Use of Noninvasive Prenatal Testing in Europe, Australia
803 and the USA: A Graphical Presentation.” *Acta Obstetricia et Gynecologica Scandinavica* 99(6):
804 722–30. doi:10.1111/aogs.13841.

805 Gamba, Cristina, Eppie R. Jones, Matthew D. Teasdale, Russell L. McLaughlin, Gloria Gonzalez-
806 Fortes, Valeria Mattiangeli, László Domboróczki, et al. 2014. “Genome Flux and Stasis in a Five
807 Millennium Transect of European Prehistory.” Nature Communications 5(1): 5257.
808 doi:10.1038/ncomms6257.

809 Homburger, Julian R., Cynthia L. Neben, Gilad Mishne, Alicia Y. Zhou, Sekar Kathiresan, and Amit
810 V. Khera. 2019. “Low Coverage Whole Genome Sequencing Enables Accurate Assessment of
811 Common Variants and Calculation of Genome-Wide Polygenic Scores.” Genome Medicine 11(1):
812 74. doi:10.1186/s13073-019-0682-2.

813 Hui, Ruoyun et al. 2020. “Evaluating Genotype Imputation Pipeline for Ultra-Low Coverage Ancient
814 Genomes.” *Scientific Reports* 10(1): 18542. doi:10.1038/s41598-020-75387-w.

815 Johnson, Ruth et al. 2022. “Leveraging Genomic Diversity for Discovery in an Electronic Health
816 Record Linked Biobank: The UCLA ATLAS Community Health Initiative.” *Genome Medicine* 14(1):
817 104. doi:10.1186/s13073-022-01106-x.

818 Kim, Sung K., Gregory Hannum, Jennifer Geis, John Tynan, Grant Hogg, Chen Zhao, Taylor J.
819 Jensen, et al. 2015. “Determination of Fetal DNA Fraction from the Plasma of Pregnant Women

Using Sequence Read Counts: Determination of Fetal DNA Fraction from the Plasma of Pregnant Women Using Sequence Read Counts." Prenatal Diagnosis 35(8): 810–15. doi:10.1002/pd.4615.

Li, Heng. 2011. "A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data." *Bioinformatics* 27(21): 2987–93. doi:10.1093/bioinformatics/btr509.

Linthorst, J., Nivard, M., Sistermans, E.A., 2024. GWAS shows the genetics behind cell-free DNA and highlights the importance of p.Arg206Cys in DNASE1L3 for non-invasive testing. Cell Reports 43, 114799. https://doi.org/10.1016/j.celrep.2024.114799

Liu, Siyang et al. 2018. "Genomic Analyses from Non-Invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History." *Cell* 175(2): 347-359.e14. doi:10.1016/j.cell.2018.08.016.

Liu, Siyang, Shujia Huang, Yanhong Liu, Yuqin Gu, Xingchen Lin, Huanhuan Zhu, Hankui Liu, et al. 2023. "Utilizing Non-Invasive Prenatal Test Sequencing Data Resource for Human Genetic Investigation." bioRxiv: 2023.12.11.570976. doi:10.1101/2023.12.11.570976.

Manichaikul, Ani et al. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." *Bioinformatics* 26(22): 2867–73. doi:10.1093/bioinformatics/btq559.

Marchini, J., Howie, B. "Genotype imputation for genome-wide association studies". *Nat Rev Genet* 11, 499–511 (2010). https://doi.org/10.1038/nrg2796

Martin, Alicia R. et al. 2021. "Low-Coverage Sequencing Cost-Effectively Detects Known and Novel Variation in Underrepresented Populations." *The American Journal of Human Genetics* 108(4): 656–68. doi:10.1016/j.ajhg.2021.03.012.

Martiniello, Marco. 2013. "Belgium, Migration, 1946 to Present." In *The Encyclopedia of Global Human Migration*, ed. Immanuel Ness. Wiley. doi:10.1002/9781444351071.wbeghm063.

Melville J. 2023. "Uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction."

Mokveld, Tom, Zaid Al-Ars, Erik A. Sistermans, and Marcel Reinders. 2023. “A Comprehensive Performance Analysis of Sequence-Based within-Sample Testing NIPT Methods” ed. Hao Sun. *PLOS ONE* 18(4): e0284493. doi:10.1371/journal.pone.0284493.

Nguyen, T.V., Bolormaa, S., Reich, C.M., Chamberlain, A.J., Vander Jagt, C.J., Daetwyler, H.D., MacLeod, I.M., 2024. Empirical versus estimated accuracy of imputation: optimising filtering thresholds for sequence imputation. Genet Sel Evol 56, 72. https://doi.org/10.1186/s12711-024-00942-2

Pedersen, Brent S, and Aaron R Quinlan. 2018. “Mosdepth: Quick Coverage Calculation for Genomes and Exomes” ed. John Hancock. *Bioinformatics* 34(5): 867–68. doi:10.1093/bioinformatics/btx699.

Poplin, Ryan et al. 2017. *Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples*. Genomics. preprint. doi:10.1101/201178.

Project MinE ALS Sequencing Consortium. 2018. “Project MinE: Study Design and Pilot Analyses of a Large-Scale Whole-Genome Sequencing Study in Amyotrophic Lateral Sclerosis.” *European Journal of Human Genetics* 26(10): 1537–46. doi:10.1038/s41431-018-0177-4.

Qiao, Longwei et al. 2019. “Sequencing of Short cfDNA Fragments in NIPT Improves Fetal Fraction with Higher Maternal BMI and Early Gestational Age.” *American Journal of Translational Research* 11(7): 4450–59.

Rava, Richard P, Anupama Srinivasan, Amy J Sehnert, and Diana W Bianchi. 2014. “Circulating Fetal Cell-Free DNA Fractions Differ in Autosomal Aneuploidies and Monosomy X.” Clinical Chemistry 60(1): 243–50. doi:10.1373/clinchem.2013.207951.

Rubinacci, Simone, Robin J. Hofmeister, Bárbara Sousa Da Mota, and Olivier Delaneau. 2023. “Imputation of Low-Coverage Sequencing Data from 150.119 UK Biobank Genomes.” *Nature Genetics* 55(7): 1088–90. doi:10.1038/s41588-023-01438-3.

Rubinacci, Simone, Diogo M. Ribeiro, Robin J. Hofmeister, and Olivier Delaneau. 2021. “Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels.” *Nature Genetics* 53(1): 120–26. doi:10.1038/s41588-020-00756-0.

Seidman, Daniel N. et al. 2020. “Rapid, Phase-Free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification.” *The American Journal of Human Genetics* 106(4): 453–66. doi:10.1016/j.ajhg.2020.02.012.

Sousa Da Mota, Bárbara et al. 2023. “Imputation of Ancient Human Genomes.” *Nature Communications* 14(1): 3660. doi:10.1038/s41467-023-39202-0.

The 1000 Genomes Project Consortium et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526(7571): 68–74. doi:10.1038/nature15393.

The Genome of the Netherlands Consortium. 2014. “Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population.” *Nature Genetics* 46(8): 818–25. doi:10.1038/ng.3021.

the Haplotype Reference Consortium. 2016. “A Reference Panel of 64.976 Haplotypes for Genotype Imputation.” *Nature Genetics* 48(10): 1279–83. doi:10.1038/ng.3643.

The International HapMap Consortium. 2005. “A Haplotype Map of the Human Genome.” Nature 437(7063): 1299–1320. doi:10.1038/nature04226.

Van Den Eynden, Jimmy et al. 2018. “The Genetic Structure of the Belgian Population.” *Human Genomics* 12(1): 6. doi:10.1186/s40246-018-0136-8.

Van Riel, Margot, Kate Stanley, and Joris R. Vermeesch. 2023. “Noninvasive Prenatal Testing/Screening by Circulating Cell-Free DNA.” In Human Reproductive and Prenatal Genetics, Elsevier, 823–51. doi:10.1016/B978-0-323-91380-5.00013-7.

Vermeesch, Joris Robert, Thierry Voet, and Koenraad Devriendt. 2016. “Prenatal and Pre-Implantation Genetic Diagnosis.” Nature Reviews Genetics 17(10): 643–56. doi:10.1038/nrg.2016.97.

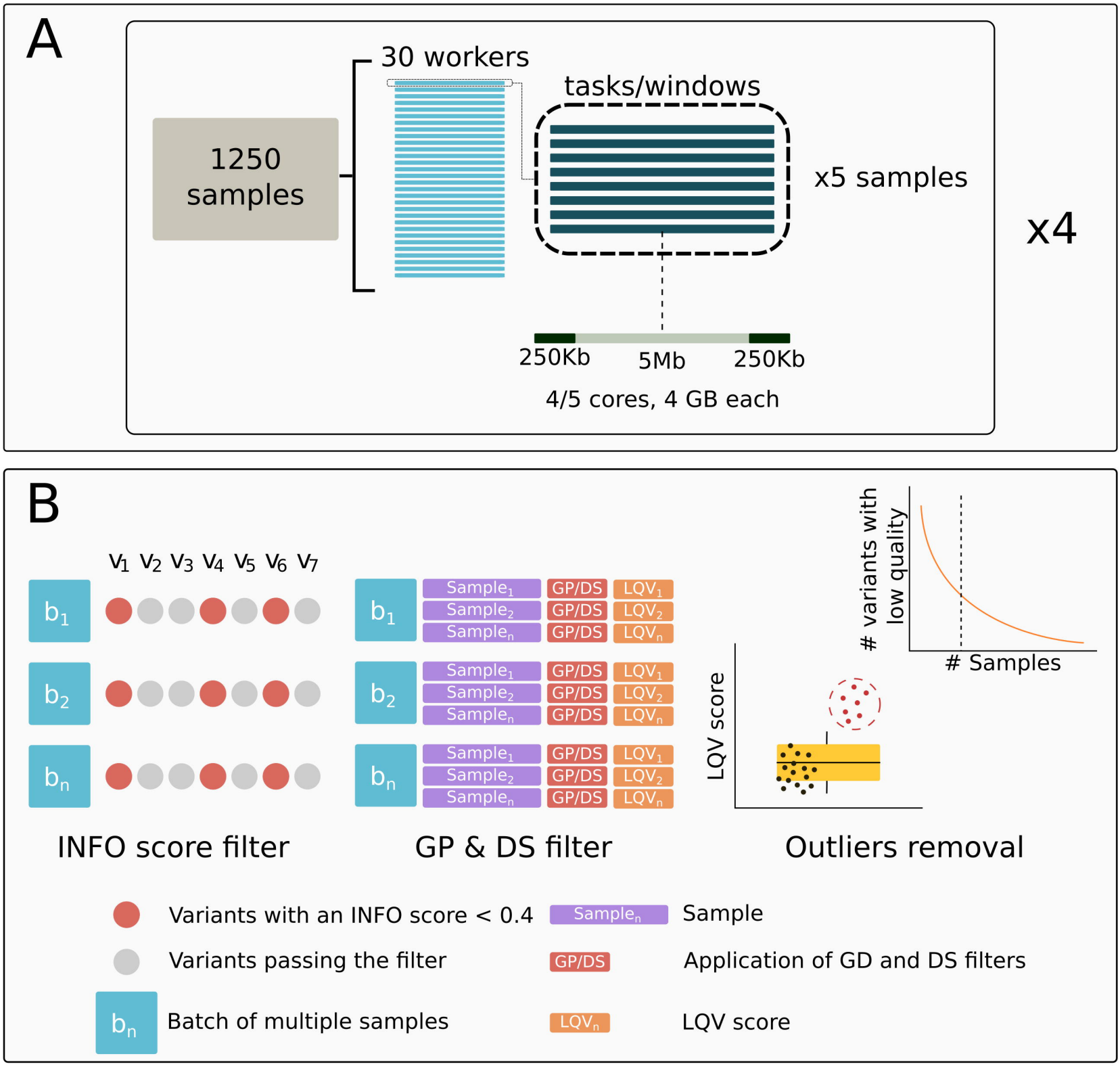
Wang, Jing-wei et al. 2021. “Cell-Free Fetal DNA Testing and Its Correlation with Prenatal Indications.” *BMC Pregnancy and Childbirth* 21(1): 585. doi:10.1186/s12884-021-04044-5.

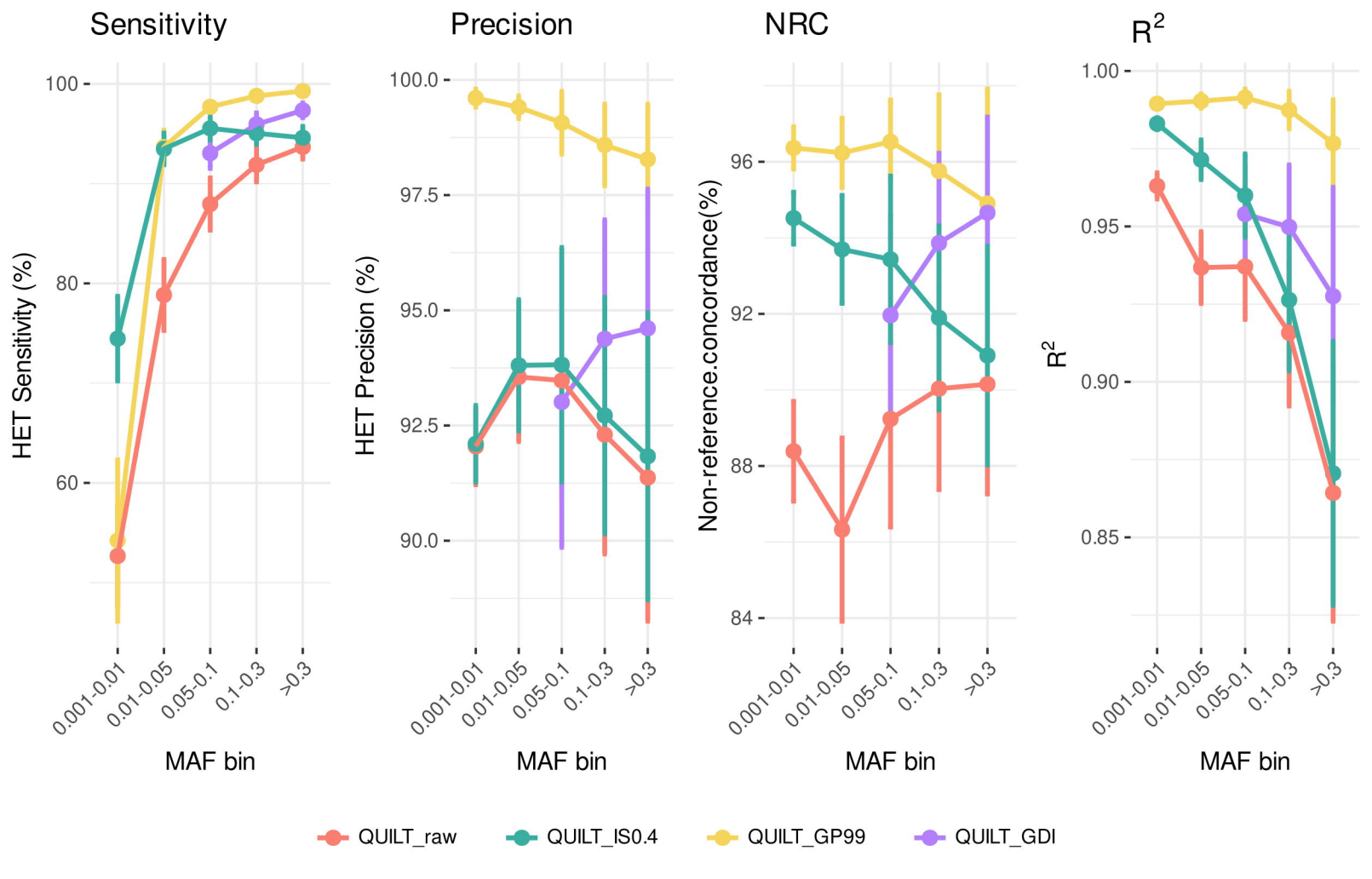
Wasik, Kaja et al. 2021. “Comparing Low-Pass Sequencing and Genotyping for Trait Mapping in Pharmacogenetics.” *BMC Genomics* 22(1): 197. doi:10.1186/s12864-021-07508-2.

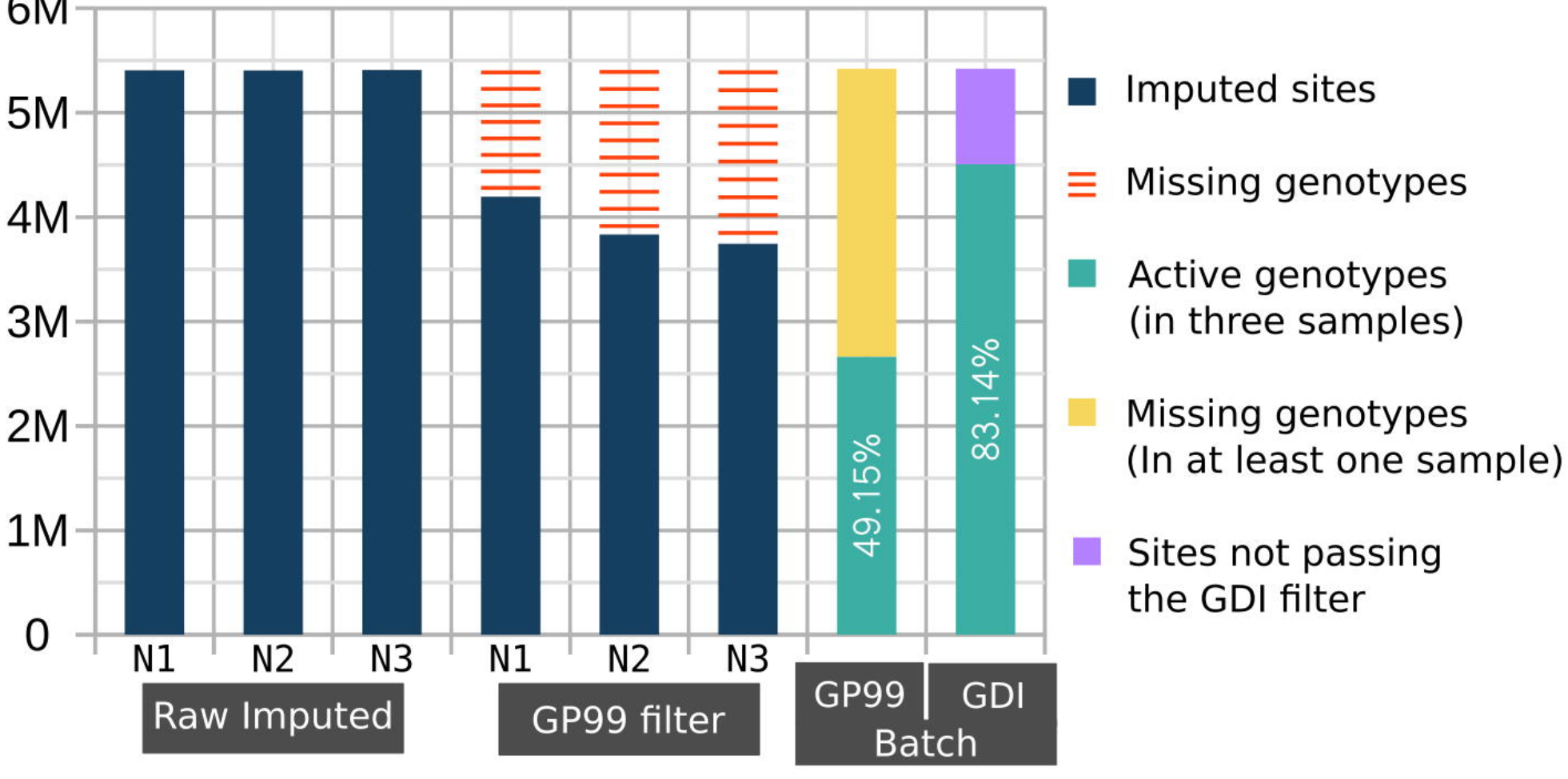
Yengo, Loïc, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, et al. 2022. “A Saturated Map of Common Genetic Variants Associated with Human Height.” Nature 610(7933): 704–12. doi:10.1038/s41586-022-05275-y.

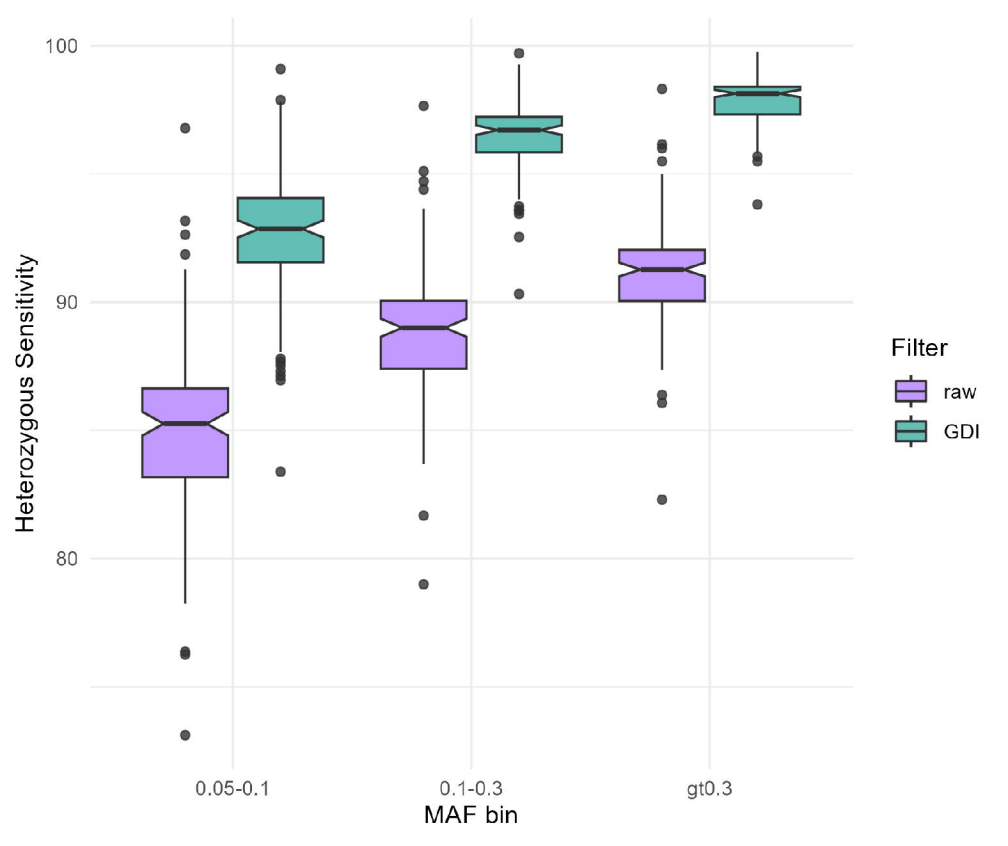
Zhao, Hao et al. 2014. “CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies.” *Bioinformatics* 30(7): 1006–7. doi:10.1093/bioinformatics/btt7

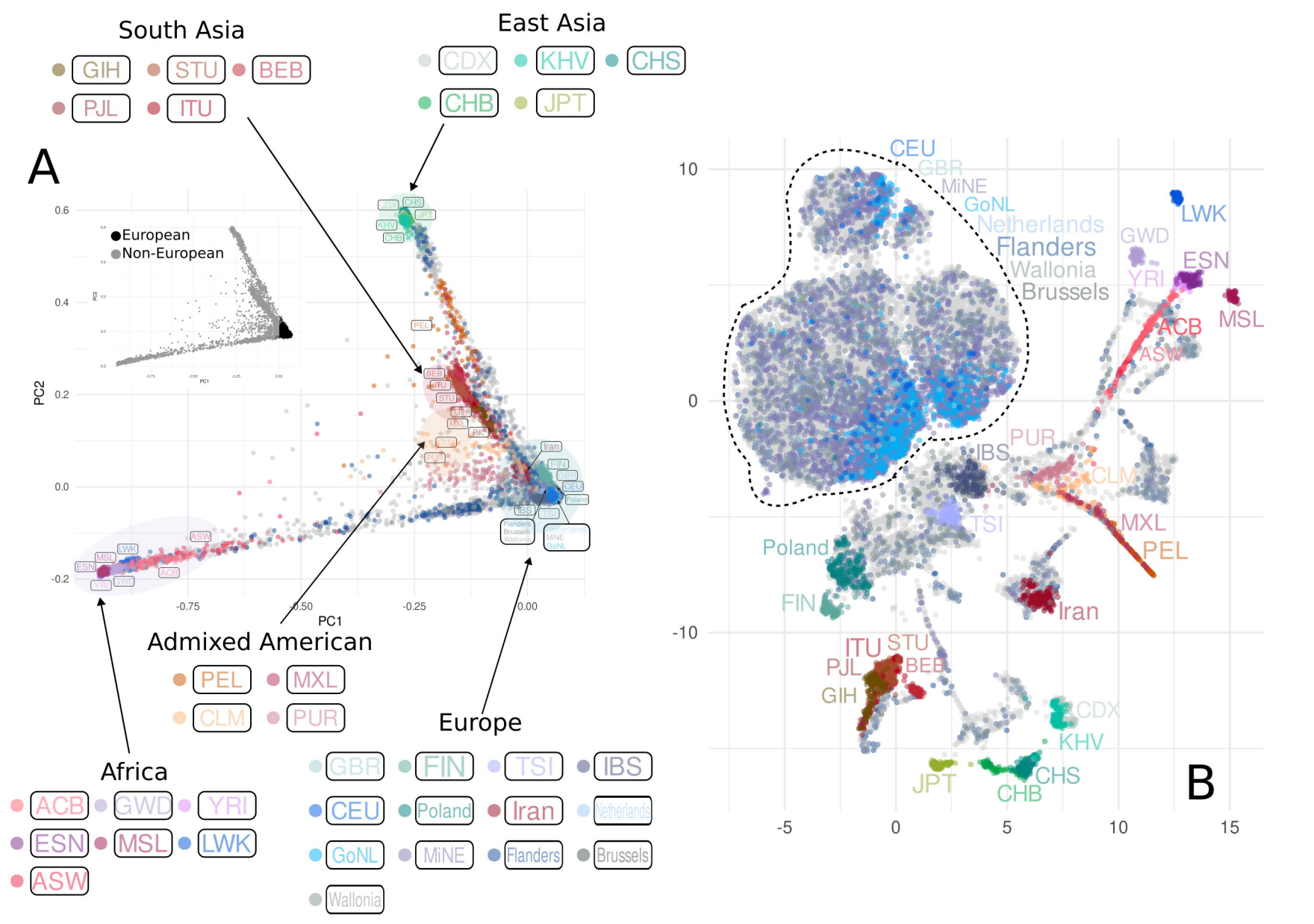
Zhen, Jianxin, Yuqin Gu, Piao Wang, Weihong Wang, Shengzhe Bian, Shujia Huang, Hui Liang, et al. 2024. “Genome-Wide Association and Mendelian Randomisation Analysis among 30.699 Chinese Pregnant Women Identifies Novel Genetic and Molecular Risk Factors for Gestational Diabetes and Glycaemic Traits.” Diabetologia 67(4): 703–13. doi:10.1007/s00125-023-06065

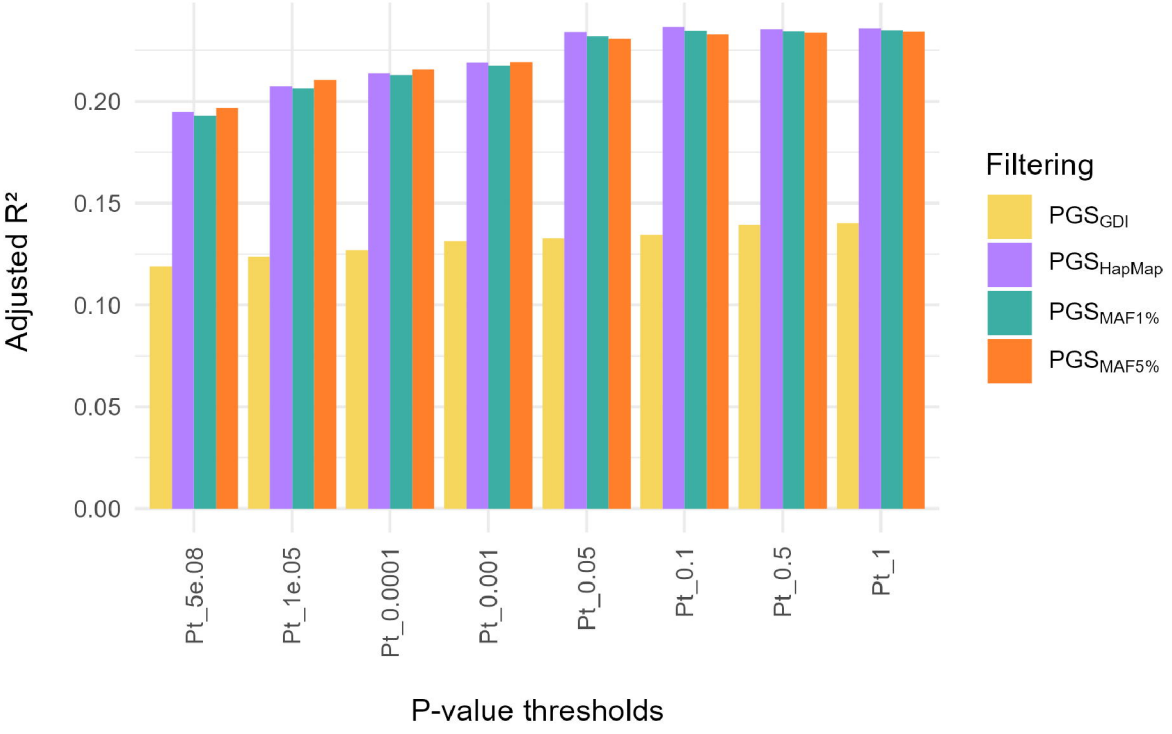

A
30 workers
tasks/windows
1250 samples
x5 samples
x4
250Kb
5Mb
250Kb
4/5 cores, 4 GB each
B
V1 V2 V3 V4 V5 V6 V7
b1
b2
bn
Sample1
Sample2
Samplen
GP/DS
LQV1
LQV2
LQVn
variants with low quality
Samples
LQV score
INFO score filter
GP & DS filter
Outliers removal
Variants with an INFO score < 0.4
Variants passing the filter
Batch of multiple samples
Sample
Application of GD and DS filters
LQV score


Sensitivity
Precision
NRC
R2
HET Sensitivity (%)
HET Precision (%)
Non-reference.concordance(%)
R2
100
80
60
100.0
97.5
95.0
92.5
90.0
96
92
88
84
1.00
0.95
0.90
0.85
0.001-0.01
0.01-0.05
0.05-0.1
0.1-0.3
>0.3
MAF bin
QUILT_raw
QUILT_IS0.4
QUILT_GP99
QUILT_GDI


6M
5M
4M
3M
2M
1M
0
N1
N2
N3
N1
N2
N3
49.15%
83.14%
Raw Imputed
GP99 filter
GP99
GDI
Batch
Imputed sites
Missing genotypes
Active genotypes
(in three samples)
Missing genotypes
(In at least one sample)
Sites not passing
the GDI filter


100
90
80
Heterozygous Sensitivity
0.05-0.1
0.1-0.3
gt0.3
MAF bin
Filter
raw
GDI


South Asia
GIH
STU
BEB
PJL
ITU
East Asia
CDX
KHV
CHS
CHB
JPT
A
European
Non-European
PC1
PC2
Admixed American
PEL
MXL
CLM
PUR
Africa
ACB
GWD
YRI
ESN
MSL
LWK
ASW
Europe
GBR
FIN
TSI
IBS
CEU
Poland
Iran
Netherlands
GoNL
MiNE
Flanders
Brussels
Wallonia
B
CEU
GBR
MiNE
GoNL
Netherlands
Flanders
Wallonia
Brussels
LWK
GWD
ESN
YRI
MSL
ACB
ASW
PUR
IBS
CLM
MXL
PEL
TSI
Poland
FIN
Iran
ITU
STU
PJL
BEB
GIH
CDX
KHV
JPT
CHS
CHB


Adjusted R²
0.20
0.15
0.10
0.05
0.00
Pt_5e.08
Pt_1e.05
Pt_0.0001
Pt_0.001
Pt_0.05
Pt_0.1
Pt_0.5
Pt_1
P-value thresholds
Filtering
PGSGDI
PGSHapMap
PGSMAF1%
PGSMAF5%

Captions for Main Figures

Figure 1 **A)** Imputation pipeline with QUILT. The image illustrates the design of imputation of a batch of 1,250 individuals described in further detail in the Materials and Methods section. **B)** GDI filtering scheme: *INFO score filter*: Each column (e.g., v1,v2,v3, etc.) represents a distinct variant, and each row (e.g., b1, b2, bn) corresponds to a batch. Variants with an INFO score < 0.4 in all batches (red dots) are excluded from subsequent steps. This exclusion applies only when the same variant consistently displays this condition across all batches (e.g., three out of seven variants in the example cartoon are excluded). Variants that pass the filter (gray dots) proceed to the next step. *GP & DS filter*: Using the variants that passed the previous step, for each individual within a batch, an LQV score (Low-quality Variant score) is calculated, indicating the proportion of variants that do not meet both GP and DS thresholds. Each sample's LQV score is then used to assess the quality of the data. *Outliers removal*: The distribution of LQV scores is analyzed to identify and remove outlier samples (represented by the red dots in the boxplot). For the remaining samples (represented by the black dots in the boxplot), the distribution of variants with low quality is assessed relative to the total number of samples. The curve in the top-right subpanel illustrates how a cutoff is determined to exclude variants that consistently exhibit low quality in more than a specified percentage of samples. Visual inspection of UMAP plots was used to determine the optimal cutoff (additional details are provided in the main text).

Figure 2 Comparative analysis of post-imputation filters on QUILT performance by four measures of accuracy. Each panel displays results for different post-imputation statistics for those genotypes imputed as heterozygous: Sensitivity (TP/(TP+FN)), Precision (TP/(TP+FP)), Non-reference concordance (NRC = 1 - (Err + Era + Eaa) / (Err + Era + Eaa + Mra + Maa). Where Err, Era, and Eaa are the counts of the mismatches for the homozygous reference, heterozygous and homozygous alternative genotypes, while Mra and Maa are the counts of the matches at the heterozygous and homozygous alternative genotypes), and the dosage r-squared (Pearson's r2 on dosage values calculated using BCFtools stats). Lines are colored based on different filtering approaches. Results are average across the three test samples, with SD bars showing the variability around the mean values. (Please, refer to Supplementary Table 2 for detailed information on the values). The lack of QUILT_GDI data points in the 0.001-0.01 and 0.01-0.05 frequency bins reflects our focus on common variants with MAF > 5% in the HRC panel.

Figure 3 Cumulative missingness of filtered-out variants across multiple imputed genomes. The number of imputed common variants (MAFHRC > 0.05) is displayed in millions on the Y axis for each of the three imputed NIPS samples (N1, N2, and N3). The raw imputed data of each individual had 5.4M variants. Applying GP<0.99 filter (GP99 filter) on each individual separately results in a variable number of missing variants (red bars). The GP99 and GDI Batch bars to the right show the batch effect of cumulative missingness when applying the respective filters. The green bar shows the number of retained sites after removing variants with GP<0.99 (yellow) in at least one sample or variant sites not passing the GDI filter (lilac).

Figure 4. Boxplots per MAF showing the heterozygous genotype imputation sensitivity before (raw) and after GDI filters. Data are derived from 138 non-pregnant healthy individuals sequenced at 0.1-0.2× coverage and genotyped, in parallel, on Illumina Infinium Global Screening Array-24 v3.0 for ~700,000 SNPs. The array-based genotype calls were considered as truth.

Figure 5 Principal Component Analysis of imputed NIPS data in context of regional and global reference populations. (**A**) The plot of PC1 and PC2. The smaller PCA plot highlights European and non-European ancestries; (**B**) UMAP plot of the first 20 PC-s. Both plots display the distribution of 27.354 imputed NIPS samples (gray dots in the background) together with samples from the 1000 Genomes Project, the GoNL, and the MinE datasets, after applying a GDI filter removing variants with low quality in more than 30% of the imputed NIPS samples. Population labels are displayed next to the group they represent (samples from Iran and Poland are part of the Belgian data), with colors of the labels matching the colors of the corresponding samples. Unlabeled imputed samples are depicted as gray dots in the background. The dotted line includes the group represented by samples from the GoNL, the MinE, the 1000 Genomes Project (CEU and GBR only), unlabelled NIPS samples, and NIPS individuals labeled as Flanders, Wallonia, Brussels, and the Netherlands.

Figure 6 Phenotypic variance in height explained by the PGS per p-value threshold and for different filtering approaches. The overall best performing PGS in terms of variance explained is PGSHapMap at p-value threshold (Pt) 0.1 with an r^2 of 0.237 (beta = 0.03, p-value = 2.83×10^{-114}).

Table 1

Average imputation sensitivities of heterozygous genotypes obtained using HRC and 1000GP reference panels.

		Het sensitivity		Number of Hets		Missing Hets	
Filter	**MAF bin**	**HRC**	**1000GP**	**HRC**	**1000GP**	**HRC**	**1000GP**
No filter	0.001-0.01	52.7	50.9	17188	14039	0	0
	0.01-0.05	78.8	74.0	83397	77676	0	0
	0.05-0.1	88.0	84.1	127130	120522	0	0
	0.1-0.3	91.9	88.5	677999	647528	0	0
	>0.3	93.7	90.4	789923	755478	0	0
max(GP)≥0.99	0.001-0.01	54.2	50.4	5046	5669	71.5	59.3
	0.01-0.05	93.7	88.1	48631	46943	50.9	49.3
	0.05-0.1	97.7	96.1	88139	82656	37.6	40
	0.1-0.3	98.8	98.3	479079	452659	34.3	37
	>0.3	99.3	99.1	559099	531081	33.2	35.9
GDI	0.001-0.01	na	na	na	na	0	na
	0.01-0.05	na	na	na	na	0	na
	0.05-0.1	93.1	na	101500	na	0	na
	0.1-0.3	96.0	na	432852	na	0	na
	>0.3	97.4	na	417039	na	0	na

Note, Het sensitivity - heterozygote genotype sensitivities are expressed in percentage.

Number of hets - number of correctly imputed heterozygous genotypes in the given MAF bin#

Missing hets - proportion of imputed heterozygous genotypes that were set to missing by the GP99 filter in the given MAF bin