medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Removing array-specific batch effects in GWAS mega-analyses by applying a two-step imputation workflow reveals new associations for thyroid volume and goiter

M. Kamal Nasr[1,2], Eva König[3], Christian Fuchsberger[3], Sahar Ghasemi[4], Uwe Völker[2,5], Henry Völzke[2,6], Hans J. Grabe[1,7], Alexander Teumer[1,2]

[1] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany
[2] DZHK (German Centre for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany
[3] Institute for Biomedicine, Eurac Research, Bolzano, Italy
[4] Institute of Genetic Epidemiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany
[5] Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
[6] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
[7] German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, Greifswald, Germany

Keywords: genetic imputation, association studies, mega-analysis

Correspondence to:
Dr. Alexander Teumer
University Medicine Greifswald
Department of Psychiatry and Psychotherapy
Ellernholzstr. 1-2
17475 Greifswald, Germany
ateumer@uni-greifswald.de
Phone: +49 (0) 3834-86-6918
ORCiD: https://orcid.org/0000-0002-8309-094X

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Background: Combining individual-level data in genetic association studies (mega-analyses) enhances statistical power for identifying gene-trait associations. However, batch effects from combining variants of different arrays pose a major limitation. Here, we developed a two-step imputation workflow to overcome the array type bias.

Methods: Genotype data of 10,647 individuals generated using five different arrays were included. Intermediate array-specific panels were generated and subsequently imputed against the 1000 Genomes Project Phase3 reference panel. Genetic principal component (PC) analysis assessed batch effects in the cohort-combined imputed data. The workflow's performance was evaluated by comparing imputation quality r^2 and allele frequency difference of the proposed two-step imputation to the conventional array-specific imputation as well as its matching with a whole-genome sequenced subgroup for further validation. We performed a genome-wide association study (GWAS) to test for genetic associations with goiter risk and thyroid gland volume, comparing summary statistics of both approaches.

Results: The proposed workflow eliminated the batch effect from the first twenty genetic PCs. The outcome of the workflow also showed high correlation with the conventional approach for allele frequencies ($r^2 > 0.99$). GWAS results from the two-step imputation confirmed known associations on thyroid traits and revealed novel loci for thyroid volume (*TG, PAX8, IGFBP5, NRG1*), and one novel locus for goiter (*XKR6*), which was not statistically significant following the GWAS meta-analysis of conventional imputation.

Conclusion: Our imputation workflow provides high-quality imputation results without technical batch effects, fostering mega-analysis involving multiple genotyping arrays for different genetic association analysis.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Genome-wide association studies (GWAS) represent an agnostic approach for identifying genetic associations with common traits and diseases by testing millions of variants with continuous outcomes or between groups. The testing checks allele frequency differences between individuals in a selected population that show different representations of the trait value [1]. GWAS analyses have been utilized in more than 5,700 studies, exploring 3,300 different traits [1,2]. The results of these analysis enriched our knowledge about disease risk variants, as well as identifying individuals with high disease-risk profiles through risk scores for complex heritable traits [1,3].

The power to detect associations increases with sample size and number of variants tested [3,4], rare and low frequency single nucleotide variants (SNVs) are of high interest. This requires the inclusion of large sample sizes in the experimental design, which is complicated or even unrealistic for rare diseases and low number of cases in the investigated populations. One alternative approach is a meta-analysis of summary statistics from different GWAS analyses of the same trait, performed either on same or different populations [4,5]. Another approach is to combine the individual-level data of samples from different cohorts to perform a mega-analysis [6].

Meanwhile, genetic imputation is a reliable method to estimate alleles of variants not directly genotyped on an array. Based on linkage disequilibrium (LD), genetic imputation can significantly increase coverage of the human genome when using different commercial genotyping arrays [7,8]. This method is utilized as an alternative for expensive whole genome sequencing [9].

GWAS on imputed variants and subsequent meta-analysis are frequently utilized cost-effective approaches for conducting large GWAS[10], however, they are also associated with technical limitations [11]. If the sample size or number of cases in a specific cohort especially in a (nested) case-control design is low, the analysis results are less reliable because the effective number of samples is too small for the association models like linear regression, logistic regression or mixed-effects models, particularly when it comes to low-frequency variants. In addition, the analysis workload substantially increases with the number of cohorts included in a project [12].

Former studies have shown that both meta-analyses and mega-analyses using individual participant data are mathematically equivalent [13], and also comparable

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

when using imputed genotypes [14]. However, previous studies have mainly focused on data from cohorts genotyped on the same array type. Highlighting challenges in mega-analysis, where individual-level data from multiple cohorts genotyped on different arrays are combined.

We identified a concerning technical bias when combining the imputed data of individuals from genetically homogeneous northeastern German cohorts that were genotyped using diverse array types [15,16]. Principal component analysis (PCA) conducted on the quality-controlled genotype data revealed variation influenced by the array type. Notably, this variation was detected upon both imputation against the haplotype reference consortium (HRC) [17] and the 1000 Genomes v5 (1000G) reference panels [18].

Here, we propose a newly developed workflow, to minimize the technical bias due to batch effect when combining genotype data from different arrays. The workflow is composed of two imputation steps. First we impute the included genotype datasets pairwise against each other and then create a panel of overlapping variants for each imputation outcome. Finally, we impute the generated intermediate panel against one of the commonly available large panels. We evaluate the outcome of the new workflow in comparison to conventional imputation approaches. In an application example, we conducted a GWAS analysis on thyroid gland volume and goiter, identifying new associations with goiter while demonstrating the robustness of our imputation workflow by validating known associations.

Materials and Methods

Workflow design overview

In this project, three different approaches for genetic imputation have been conducted. The first proposed approach is a new two-step imputation process. The second and third approaches are conventional single-step imputation for comparison with the newly proposed method. We modified the third approach to be a single-step imputation using only the intersecting variants genotyped on all included array types. The genotype data that were used for imputation were the same in both imputation approaches. We used data from five different arrays (Affymetrix SNP 6.0, Affymetrix Axiom [Thermo Fisher Scientific, Santa Clara, CA, USA], and Illumina Omni 2.5, Illumina GSA, and Illumina PsychArray [Illumina, Inc., San Diego, CA, USA]) obtained from samples of the German GANI_MED (n= 2410), SHIP-START (n=

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4070) and SHIP-TREND (n= 4119) (Figure 1) [15,16]. The individuals genotyped on the Affymetrix Axiom array (n= 48) were a subset of the Affymetrix SNP 6.0 samples, whereas all other individuals were genotyped only once. Detailed information regarding the included datasets, sample sizes, and genotype arrays are provided in Figure 1 and the Supplementary Methods. For evaluation of the imputation performance, a whole-genome sequenced subset of 192 individuals from SHIP-TREND has been used for genotype matching concordance checking.

Data pre-processing and phasing

Genotype quality control was performed using PLINK [19] as reported previously [20,21]. Briefly, arrays with genotype call rate <94%, as well as variants with missing call rate >5%, Hardy-Weinberg equilibrium p-value $<10^{-4}$, and monomorphic SNVs and singletons were excluded. All the included genotype datasets were aligned to reference genome build (GRCh37) using BCFtools software followed by retaining only sites with at least one alternative allele [22]. Haplotype phasing preceding the first imputation round was performed for each genotype dataset using Eagle2 (5 phasing iterations and number of conditioning haplotypes = 10^4) and estimated to hg19 mapping [23]. No external reference panel was used for phasing.

Two-step imputation

In the first step, the intermediate imputation, each phased panel was used as a reference panel after compression with the minimac4 imputation software [24]. Each of the five pre-processed genotype arrays was imputed pairwise against the other four, yielding twenty imputed datasets. Using an R scripted algorithm with VCFtools and BCFtools for variants filtration [22], each of these imputed variant sets were subset to the variants overlapping between all generated panels with high imputation quality (R^2 >= 0.9), selecting the source panel with the highest imputation quality score for each variant. The outcome of the first-step imputation was one VCF file for each cohort dataset, which was then used as input for the second imputation step against 1000G reference panel using the Michigan imputation server [24]. Eagle2 was selected for phasing without (additional) imputation quality filters applied at this stage of the imputation.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Conventional imputation

To validate the outcome of the proposed workflow, we imputed the genotype data of each array directly against 1000G reference panel using the same parameters and quality control procedures as described in the two-step imputation.

Conventional imputation with overlapping genotypes

To evaluate a simple approach for removing the array type specific batch effect, we ran a single-step imputation by restricting the input variants to those assessed on all array types. This approach was tested for three (Affymetrix SNP 6.0, Illumina Omni 2.5 and Illumina GSA) and all five array types. We compared the imputation quality of these imputations to the conventional and the two-step imputation exemplarily for the SHIP-TREND.

Batch effect assessment

Technical bias was assessed by estimating genetic principal components of the imputed and quality-controlled genotype data. Quality control filters included missing variant call rates above 5%, Hardy-Weinberg equilibrium p-values less than 10^{-4}, minor allele frequency (MAF) less than 1%, correlated variants were removed by LD pruning with a window size of 50, step size of 5 SNVs and R^2 threshold of 0.2. Genetic principal components, along with their explained variances were compared between two-step imputation and conventional imputation. To check the robustness of the two-step imputation approach for rare variants, we further ran PCA on rare variants (MAF < 1%). All PCAs are performed using PLINK 2.0 software [19]. Principal components (PC) of the first twenty components were plotted with ggplot2 package of R software.

Evaluation of imputation performance

Genotype data attributes, including differences in MAF and imputation quality R^2 measure between the two-step and conventional imputation were compared for each included cohort after stratification by minor allele frequency. We additionally checked the concordance (number of matching genotypes/total number of genotypes [25]) of the imputed best-guess (hard call) genotypes with their whole-genome sequencing data of a subset of 192 individuals from SHIP-TREND (Omni 2.5) for which whole genome sequencing (WGS) data was also available.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

GWAS and results comparison

We evaluated the results of GWAS analyses using genotype dosage information from both two-step and conventional imputation approaches using thyroid gland volume and goiter risk as outcomes as described before [26], and summarized in the Supplementary Methods.

All GWAS were conducted with the EPACTS 3.3.2 software using dosages [27]. Inverse-variance weighted meta-analysis of the single cohort GWAS summary statistics based on the conventional imputation approach was done using the METAL software [4]. We compared the summary statistics p-values, effect estimates and its standard errors. A detailed description of the GWAS analysis plan is available in the Supplementary Methods. Independent (lead) variants associated with thyroid traits were identified using the clumping function of PLINK with a threshold of $p < 5 \times 10^{-8}$, $r^2 > 0.01$ and 1 Mb distance, and compared across the imputation approaches. For biological validation of the findings, lead variants were investigated by checking their association with thyrotropin to support their biological plausibility [20].

Results

Genotype datasets of 10,647 individuals genotyped on five different array types were included in the workflow (Figure 1). After the intermediate imputation, 1,942,499 high-quality overlapping autosomal variants were generated for each of the included five datasets, and subsequently used for the second imputation step against the 1000G reference panel. In contrast, only 58,091 autosomal genotyped variants were available as imputation input in the three cohorts using the conventional imputation approach with overlapping variants. This number further dropped to 12,265 variants when including the Affymetrix Axiom and Illumina PsychArrays. Supplementary Table 1 provides the distribution of these variants for each chromosome in the two-step imputation, the conventional imputation and conventional imputation using overlapped variants.

Batch effect investigation (PCA analysis)

For the conventional imputation, 939,802 variants with a total genotyping rate of 0.988 were included in the PCA analysis, with 2,314,205 variants removed due to missing genotype threshold, 9,118 variants due to Hardy-Weinberg equilibrium, and 38,008,458 variants due to the MAF threshold. Projecting the first two principle

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

components of the imputed genotypes clustered the samples by array types with a combined explained variance of 0.242 (Figure 2a, b, Supplementary Figure 1a).

The two-step imputation included 1,360,834 variants with a total genotyping rate of 0.989 in the PCA analysis after removing 725,008 variants due to missing genotype threshold, 5,363 variants due to Hardy-Weinberg equilibrium, and 37,753,448 variants due to the MAF threshold. Using this approach, the first 20 principal components of the imputed genotypes did not show any array type specific clusters (Figure 2c, d, Supplementary Figure 1b). PCA restricted to the rare genotypes imputed with the two-step approach did not show clustering by the array type (Supplementary Figure 2).

The conventional imputation using only overlapping variants also removed the batch effect (Supplementary Figure 3), but led to a drastic decrease in the quality of the imputed variants for all allele frequency groups in comparison to the other approaches (Supplementary Figure 4). Median R^2 of the overlapping variants was 0.011, while for two-step imputation and conventional imputation using all variants was 0.993 for variants with MAF > 0.05, which was the main reason for not considering this approach any further in the performance evaluation and GWAS analysis, focusing only on the two-step imputation against conventional imputation using all variants.

Two-step imputation outcome parameters comparison with conventional imputation

Our newly developed two-step imputation approach showed higher overall quality for the imputation outcome of rarer variants (MAF < 0.01) than the conventional imputation (Figure 3, Supplementary Figure 5). For less-common variants with MAF between 0.04 and 0.05, the median R^2 increased by 0.19 when using the two-step imputation approach for the GANI_MED Illumina PsychArray. It also shows comparable allele frequencies to the conventional approach, judging by the absolute difference in allele frequencies for each imputed variant in both approaches (Figure 4). Detailed statistics about median R^2 and MAF differences are presented in Supplementary Table 2.

Hard call genotype concordance results in the subgroup of SHIP-TREND with matching WGS data showed strong correlation between two-step imputation and sequenced genotypes. The Pearson correlation coefficients with the homozygous

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

reference, heterozygous, and homozygous alternative genotypes were 0.996, 0.981, and 0.979, respectively (Table 1, Supplementary Figure 6).

GWAS on thyroid traits

A total number of 6,894 individuals from SHIP-START and SHIP-TREND (both array types) with thyroid measurement information were included in the GWAS analysis, all participants were of European ancestry. Detailed information about cohort characteristics is presented in Table 2. The estimated genomic control for all conducted GWAS analysis showed no signs of inflation, with a minimum and maximum λ_{GC} = 1.001 and 1.038, respectively (Supplementary Figure 7).

Meta-analysis of the GWAS analysis on goiter risk using conventionally imputed genotypes revealed four significantly associated loci ($p < 5 \times 10^{-8}$). Confirming the results from previously conducted GWAS analysis[26]. Two of the loci are located at the *CAPZB* region in chromosome 1, the other two at the *FAM227B* and *MAFTRR* regions on chromosome 15 and 16, respectively (Figure 5b). However, the GWAS analysis of the combined two-step imputed genotype data revealed another associated locus at the *XKR6* region on chromosome 8 (Figure 5c), which did not attain statistical significance using the conventional meta-analysis approach (Figure 5a).

The conventional approach of the GWAS meta-analysis of thyroid volume revealed four novel associations at the *PAX8* region on chromosome 2, at *IGFBP5*, *NRG1* and *TG* (Figure 6b,c), and confirmed all known associations with this trait[20]. Genome-wide significance for these regions were also obtained with the GWAS using the combined two-step imputed genotype data, with the exception of *NRG1* (p-value = 5.22×10^{-8}). (Figure 6a). Except the *PAX8* region, all associated loci were also associated with thyroxin in recently published multi-trait GWAS meta-analysis analysis for thyroid function [20].

Table 3 summarizes the results of the SNVs with the strongest association of both GWAS approaches and traits. The significant GWAS results of both approaches were comparable, judging by the magnitude of the estimates. However, the standard errors of the GWAS results obtained from the two-step imputed genotypes where generally lower, where the natural log p-values were slightly higher in the linear regression and lower in the logistic regression based analyses compared to the conventional imputation and subsequent meta-analysis. (Supplementary Figures 8 and 9).

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Discussion

Genome-wide association studies have helped identify genetic factors for many traits and diseases. Since the 1960s, collecting genotype samples from diverse populations has grown in importance [28,29]. Imputation techniques now address gaps in whole-genome sequencing [30]. However, variations in these methods across cohorts reduce the accuracy of genetic analyses due to biases from differing array technologies [31,32]. For instance, in our genotyped samples where we compared the allele frequencies of the SHIP-TREND subgroup genotyped using Illumina Omnia 2.5 with the corresponding whole genome sequencing variants (Supplementary Figure 10), this array-specific variation in the allele frequency, both in common and rare variants, can affect LD estimation in haplotype phasing and genotype imputation. Our newly developed imputation method addresses the additional variation induced by including multiple array types. By forming an intermediate panel of high-quality variants, the method enables high imputation accuracy while removing the array type induced batch effects.

The developed workflow was inspired by the use of only overlapping variants in the included cohorts (Supplementary Table 1), which eliminated the observed bias, yet led to a significant decrease in the imputation quality due to removal of informative tag SNVs (Supplementary Figure 3), the influence of informative SNVs on imputation quality has been shown in previous research work [33]. Based on that, we introduced an intermediate imputation step to generate a panel that can retain the same genotype information of the included arrays and thus preserving its LD structure.

Selecting an appropriate threshold for the imputation R^2 (≥ 0.9) of the overlapping variants in the intermediate panel was essential for having reliable genotype information upon imputing against 1000G panel. The value of the threshold was decided following the output of several imputations using different thresholds. We aimed to use the highest possible value for imputation quality R^2 without affecting haplotype phasing results due to removing too many variants. Supplementary Table 1 shows the number of included variants per chromosome when the R^2 threshold was adjusted to 0.8 as well as 0.9. Both approaches led to similar imputation outputs as seen after plotting genetic PCs of the imputation outcomes (Figure 2 and Supplementary Figure 11).

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To evaluate the existence of the array type bias, we used genetic PCs to capture the main variance in the allele frequencies, which can be seen as the influence of the array type after projecting the PCs of the eigenvectors and evaluate the homogeneity of the projected points on PC axis [34]. The imputation following the developed two-step approach showed its capability to overcome the array type differentiation whenever existing in the first twenty components, compared to conventional imputation outcomes where a clear clustering effect by the array type was observed (Figure 2).

To test the performance of the imputation outcomes, we focused on comparing allele frequency and imputation quality parameters of the cohort genotypes. These parameters are particularly relevant for conducting trait-association studies. The developed imputation flow showed strong matching of the allele frequency, represented by minimized difference in allele frequency for each variants between the imputation approaches (Figure 4 and Supplementary Table 2). The developed imputation workflow was successful in providing more reliable genetic predictions for imputed rarer alleles (Figure 3). Moreover, a strong correlation in the median R^2 of the two imputation approaches using the SHIP-TREND Omni 2.5 arrays was observed in comparison to other cohorts. This could be due to the high variant coverage of the array used for genotyping the cohort's samples. Further analysis of the allele frequency variance for each variant from both imputation outcomes of this specific cohort shows the strong correlation in the autosomal allele frequency (correlation r^2 >0.999) (Supplementary Figure 12). Nevertheless, SHIP-TREND imputed genotypes showed strong concordance with the corresponding genotyped variants in the whole-genome sequenced subgroup, indicating the representation of well estimated genotypes in both rarer and common variants (Table 1 and Supplementary Table 3). Using the allele frequencies of the WGS subgroup as a gold standard for comparison of the imputation results is somewhat misleading because differences in the genotype frequencies exist already between genotyped variants on the arrays and the WGS (Supplementary Figure 10).

The aim of conducting GWAS as part of the imputation workflow evaluation is to compare regression analyses results of the combined two-step imputed datasets to the conventional inverse-variance meta-analysis of the same samples. Goiter risk and thyroid volume are both suitable traits for our evaluation as they represent different trait datatypes with a true positive genetic association in SHIP [26]. The results

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

of GWAS analysis for the two-step imputed data showed two key advantages. First, we were able to identify a new association in the *XKR6* gene region with goiter risk as a dichotomous trait which did not reach statistical significance in the conventional meta-analysis. Variants in this locus were also associated with thyroxin in a former GWAS, making this association biologically reasonable [20]. A slightly better performance of the mega-analysis vs. a meta-analysis for a logistic regression GWAS is in line with the results of a former study [14]. The second point was its consistency with conventional imputation GWAS meta-analysis in revealing a novel locus associated with thyroid gland volume (*PAX8*). *PAX8* is a paired box family gene member that was found to be associated with the development of thyroid gland in embryonic development, and its transcription is a diagnostic marker for anaplastic thyroid carcinoma [35–37], and thus represents also a plausible association with thyroid volume in adults. All other significant GWAS findings represent also plausible true positive associations as they confirmed former findings of these traits (including replication in an independent cohort)[26], or were associated with thyroid function[20].

The comparison of the estimates and standard errors of both GWAS approaches did not show signs of p-value inflation. However, it showed that the GWAS of the two-step imputation data had a slight decrease in standard errors, in comparison to the meta-analysis outcomes.

Besides its role in discovering more hits in case-control GWAS analysis, our developed two-step imputation workflow is not restricted to a specific array, the inclusion of different arrays with different sequencing technologies in the workflow has proven its robustness in overcoming the array bias regardless of the number or the type of the array. It is also applicable for other generalized imputation panels like TOPMed or HRC reference panels [24]. These two strengths shall enable the utilization of combined genotype information for better understanding of rare diseases or genetic associations in populations that are represented in small sample sizes.

Although the array type specific batch effect in GWAS mega-analyses might be reduced in specific scenarios by adjusting for genetic PCs, such a correction will not be possible in all analyses. Such analyses include the ones using polygenic scores. Our two-step imputation provides a powerful solution for combining datasets while reducing technical bias also for analyses of polygenic scores.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Comparing the imputation outcomes to the whole genome-sequenced data had limitations for evaluating the proposed workflow, either due to the relatively small sample size of individuals who underwent WGS (n = 192 after QC) and its exclusivity for one of the included five cohorts. Although the total sample size provided us the possibility of analysing low frequency alleles, it was likely too small for evaluating the performance of the different approaches for very rare alleles. However, our two-step imputation approach seems to be on average superior to the classical imputation with regard to the imputation quality measure (Figure 3), while the difference in allele frequency is small particularly for rarer variants (Figure 4 and Supplementary Table 2). As indicated also in these results, the average difference in allele frequency seems to depend also on the density and design of the underlying genotyping array.

While the imputed genotypes generated from the proposed workflow will be utilized for identifying novel genetic associations in SHIP and GANI-MED, it will be interesting to evaluate the impact of the developed workflow on other cohorts, especially those comprising diverse ancestries. Thus far, all the included cohorts are of European ancestry from the North-East of Germany, highlighting the need to test the workflow's efficacy on genotyped cohorts from other or multiple genetic ancestries. Such evaluations will help to determine the generalizability and robustness of the imputation method across different genetic backgrounds, ultimately enhancing its utility in global genomic research.

In conclusion, our developed two-step imputation workflow aims to overcome the array type bias, by creating an intermediate panel of high-quality overlapping imputed variants. This approach enables the conduction of mega-analysis by combining genotype information from different arrays without inducing a technical array type effect. Our workflow will increase statistical power for conducting large-scale GWAS mega-analyses and other genetic analyses like polygenic risk score calculations, playing an important role in genetic research and its application in individualized medicine.

Acknowledgements:

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A). Genome-wide data have been partly supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthineers, Erlangen, Germany and the Federal State of Mecklenburg- West Pomerania. The project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 455978266 (A.T.). The authors are grateful to Linda Garvert for testing and helpful feedback on the two-step imputation workflow.

Author Contributions:

Project design and supervision: A.T.

Analyses: M.K.N., E.K., S.G.

Interpretation of the results: M.K.N., A.T., E.K., C.F.

Drafting of the manuscript: M.K.N., A.T.

Providing genotype and phenotype data: U.V., H.V., H.J.G

Critical review of the manuscript: all authors

Data and code availability

Developed scripts for the workflow are available on github (https://github.com/GenEpi-psych-UMG/Two_Step_Imputation). The data of the SHIP study cannot be made publically available due to the informed consent of the study participants, but it can be accessed through a data application form available at https://transfer.ship-med.uni-greifswald.de/ for researchers who meet the criteria for access to confidential data. The full results of the GWAS summary statistics are available on the ThyroidOmics Consortium website (http://www.thyroidomics.com).

Consents and approvals

The study followed the recommendations of the Declaration of Helsinki. The medical ethics committee of the University of Greifswald approved the study protocol, and oral and written informed consents were obtained from each of the study participants.

Conflicts of interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript. HJG received travel grants

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

447 and speakers honoraria from Neuraxpharm, Servier, Indorsia and Janssen Cilag not
448 related to the current project. HV received travel grants and speakers honoraria from
449 Sanofi-Aventis not related to the current project.

450

451

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nat. Rev. Methods Prim. *1*, 59. https://doi.org/10.1038/s43586-021-00056-9.
2. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348. https://doi.org/10.1038/s41588-019-0481-0.
3. Khera, A. V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224. https://doi.org/10.1038/s41588-018-0183-z.
4. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.
5. Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I., COGENT-Kidney Consortium, T2D-GENES Consortium, A.P., and Morris, A.P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum. Mol. Genet. *26*, 3639–3650. https://doi.org/10.1093/hmg/ddx280.
6. Abou-Khalil, B., Auce, P., Avbersek, A., Bahlo, M., Balding, D.J., Bast, T., Baum, L., Becker, A.J., Becker, F., Berghuis, B., et al. (2018). Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat. Commun. *9*, 5269. https://doi.org/10.1038/s41467-018-07524-z.
7. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678. https://doi.org/10.1038/nature05911.
8. De Marino, A., Mahmoud, A.A., Bose, M., Bircan, K.O., Terpolovsky, A., Bamunusinghe, V., Bohn, S., Khan, U., Novković, B., and Yazdi, P.G. (2022). A comparative analysis of current phasing and imputation software. PLoS One *17*, e0260177. https://doi.org/10.1371/journal.pone.0260177.
9. Quick, C., Anugu, P., Musani, S., Weiss, S.T., Burchard, E.G., White, M.J., Keys, K.L., Cucca, F., Sidore, C., Boehnke, M., et al. (2020). Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. Genet. Epidemiol. *44*, 537–549. https://doi.org/10.1002/gepi.22326.
10. Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. *44*, 631–635. https://doi.org/10.1038/ng.2283.
11. Palmer, C., and Pe'er, I. (2016). Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation. PLOS Genet. *12*, e1006091. https://doi.org/10.1371/journal.pgen.1006091.
12. Winkler, T.W., Day, F.R., Croteau-Chonka, D.C., Wood, A.R., Locke, A.E.,

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A.E., et al. (2014). Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. *9*, 1192–1212. https://doi.org/10.1038/nprot.2014.071.

13. Lin, D.Y., and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika *97*, 321–332. https://doi.org/10.1093/biomet/asq006.
14. Gorski, M., Günther, F., Winkler, T.W., Weber, B.H.F., and Heid, I.M. (2019). On the differences between mega- and meta-imputation and analysis exemplified on the genetics of age-related macular degeneration. Genet. Epidemiol. *43*, 559–576. https://doi.org/10.1002/gepi.22204.
15. Völzke, H., Schössow, J., Schmidt, C.O., Jürgens, C., Richter, A., Werner, A., Werner, N., Radke, D., Teumer, A., Ittermann, T., et al. (2022). Cohort Profile Update: The Study of Health in Pomerania (SHIP). Int. J. Epidemiol. https://doi.org/10.1093/ije/dyac034.
16. Grabe, H.J., Assel, H., Bahls, T., Dörr, M., Endlich, K., Endlich, N., Erdmann, P., Ewert, R., Felix, S.B., Fiene, B., et al. (2014). Cohort profile: Greifswald approach to individualized medicine (GANI_MED). J. Transl. Med. *12*, 144. https://doi.org/10.1186/1479-5876-12-144.
17. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283. https://doi.org/10.1038/ng.3643.
18. 1000 Genomes Project Consortium, A., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
19. 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. https://doi.org/10.1186/s13742-015-0047-8.
20. Sterenborg, R.B.T.M., Steinbrenner, I., Li, Y., Bujnis, M.N., Naito, T., Marouli, E., Galesloot, T.E., Babajide, O., Andreasen, L., Astrup, A., et al. (2024). Multi-trait analysis characterizes the genetics of thyroid function and identifies causal associations with clinical implications. Nat. Commun. *15*, 888. https://doi.org/10.1038/s41467-024-44701-9.
21. Teumer, A., Tin, A., Sorice, R., Gorski, M., Yeo, N.C., Chu, A.Y., Li, M., Li, Y., Mijatovic, V., Ko, Y.-A., et al. (2016). Genome-wide Association Studies Identify Genetic Loci Associated With Albuminuria in Diabetes. Diabetes *65*, 803–817. https://doi.org/10.2337/db15-1313.
22. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*. https://doi.org/10.1093/gigascience/giab008.
23. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448. https://doi.org/10.1038/ng.3679.
24. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287. https://doi.org/10.1038/ng.3656.
25. Wuttke, M., König, E., Katsara, M.-A., Kirsten, H., Farahani, S.K., Teumer, A.,

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Li, Y., Lang, M., Göcmen, B., Pattaro, C., et al. (2023). Imputation-powered whole-exome analysis identifies genes associated with kidney function and disease in the UK Biobank. Nat. Commun. *14*, 1287. https://doi.org/10.1038/s41467-023-36864-8.

26. Teumer, A., Rawal, R., Homuth, G., Ernst, F., Heier, M., Evert, M., Dombrowski, F., Völker, U., Nauck, M., Radke, D., et al. (2011). Genome-wide association study identifies four genetic loci associated with thyroid volume and goiter risk. Am J Hum Genet *88*, 664–673. https://doi.org/10.1016/j.ajhg.2011.04.015.
27. Kang HM (2016). EPACTS: Efficient and Parallelizable Association Container Toolbox.
28. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., et al. (2008). The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. Am. J. Hum. Genet. *83*, 347–358. https://doi.org/10.1016/j.ajhg.2008.08.005.
29. Evaluating Human Genetic Diversity (1997). (National Academies Press) https://doi.org/10.17226/5955.
30. TRECCANI, M., LOCATELLI, E., PATUZZO, C., and MALERBA, G. (2023). A broad overview of genotype imputation: Standard guidelines, approaches, and future investigations in genomic association studies. BIOCELL *47*, 1225–1241. https://doi.org/10.32604/biocell.2023.027884.
31. Lachance, J., and Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays *35*, 780–786. https://doi.org/10.1002/bies.201300014.
32. Geibel, J., Reimer, C., Weigend, S., Weigend, A., Pook, T., and Simianer, H. (2021). How array design creates SNP ascertainment bias. PLoS One *16*, e0245178. https://doi.org/10.1371/journal.pone.0245178.
33. Wojcik, G.L., Fuchsberger, C., Taliun, D., Welch, R., Martin, A.R., Shringarpure, S., Carlson, C.S., Abecasis, G., Kang, H.M., Boehnke, M., et al. (2018). Imputation-Aware Tag SNP Selection To Improve Power for Large-Scale, Multi-ethnic Association Studies. G3 Genes|Genomes|Genetics *8*, 3255–3267. https://doi.org/10.1534/g3.118.200502.
34. Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci. Rep. *12*, 14683. https://doi.org/10.1038/s41598-022-14395-4.
35. Bishop, J.A., Sharma, R., and Westra, W.H. (2011). PAX8 immunostaining of anaplastic thyroid carcinoma: a reliable means of discerning thyroid origin for undifferentiated tumors of the head and neck. Hum. Pathol. *42*, 1873–1877. https://doi.org/10.1016/j.humpath.2011.02.004.
36. Suzuki, A., Hirokawa, M., Takada, N., Higuchi, M., Yamao, N., Kuma, S., Daa, T., and Miyauchi, A. (2015). Diagnostic significance of PAX8 in thyroid squamous cell carcinoma. Endocr. J. *62*, 991–995. https://doi.org/10.1507/endocrj.EJ15-0226.
37. Ozcan, A., Shen, S.S., Hamilton, C., Anjana, K., Coffey, D., Krishnan, B., and Truong, L.D. (2011). PAX 8 expression in non-neoplastic tissues, primary tumors, and metastatic tumors: a comprehensive immunohistochemical study. Mod. Pathol. *24*, 751–764. https://doi.org/10.1038/modpathol.2011.3.

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1. Overview of the designed workflow. Variants colored in dark green represent the variants imputed in the intermediate imputation phase (first imputation), while variants colored in white are imputed from general 1000G reference panel in both two-step and conventional imputation.

Figure 2. Genetic PCA of the included cohorts in the workflow. The samples are colored by the cohorts with their unique array type. Panel A and B show the first four genetic components of the conventional imputation approach (variance explained for the components combined is 0.34). Panels C and D show the first four components of the proposed approach (variance explained for the components combined is 0.17).

Figure 3. Median R2 of the imputation outcomes of genotype data of the included cohorts, using conventional (dotted) and two-step imputation (full line).

Figure 4. Boxplots grid of the absolute difference in allele frequency (AF) between conventional and two-step imputation outcomes for the included cohorts. Each column represents an allele frequency group (rare, low, common) and each row represents one of the included cohorts.

Figure 5*.* Manhattan plot of the GWAS analysis of goiter risk using combined two-step imputation genotypes (A) and conventional imputation and meta-analysis approach (B). Variants are plotted on the x axis and $-\log_{10}$ p-values of the association testing on the y axis. Associations significant after correction for multiple testing ($p < 5 \times 10^{-8}$) are colored in red. Regional association results and recombination rates for the *XKR6* gene from two-step imputation GWAS are presented in part C, $-\log_{10}$ p-values (y-axis) of the single nucleotide variants according to their chromosomal positions (x-axis) with lead variant (rs7005680) is shown as a purple diamond.

Figure 6. Manhattan plot of the GWAS analysis of log thyroid volume using combined two-step imputation genotypes (A) and conventional imputation and meta-analysis approach (B). Variants are plotted on the x axis and $-\log_{10}$ p-values of the association testing on the y axis. Associations significant after correction for multiple testing ($p < 5 \times 10^{-8}$) are colored in red. Regional association results and recombination rates for the *NRG1* gene from conventional imputation GWAS are presented in part C, $-\log_{10}$ p-values (y-axis) of the single nucleotide variants

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

636 according to their chromosomal positions (x-axis) with lead variant (rs7000397) is
637 shown as a purple diamond.

638

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Tables

Table 1. Genotype concordance (number of matching genotypes/total number of genotypes) of the hard call imputed genotypes with sequenced data for homozygous reference (HomRef), homozygous alternative (HomAlt), and heterozygous (Het) calls, all variants followed by stratification by minor allele frequency (MAF) group of the imputed data. (n) represents number of variants represented per group in both imputed and sequenced genotypes.

Group	All variants		MAF 0.05-0.5		MAF 0.01-0.05		MAF 0.0-0.01	
	Conventional	Two-step	Conventional	Two-step	Conventional	Two-step	Conventional	Two-step
HomRef	0.9963 (618,453,288)	0.9962 (613,997,680)	0.9913 (224,034,629)	0.9910 (222,769,137)	0.9959 (171,621,821)	0.9988 (183,771,150)	0.9993 (222,756,838)	0.9994 (207,703,823)
Het	0.9814 (116,274,486)	0.981 (116,144,961)	0.9897 (105,647,496)	0.9893 (105,617,755)	0.9424 (8,530,966)	0.9418 (8,903,479)	0.7875 (2,139,474)	0.7525 (1,695,994)
HomAlt	0.9783 (21,587,792)	0.9787 (21,408,323)	0.9802 (21,463,336)	0.9807 (21,290,637)	0.7979 (128,442)	0.7862 (124,359)	0.2512 (5,149)	0.1389 (2,369)

Table 2. Cohort characteristics for GWAS analysis of thyroid volume and goiter risk. BSA: body surface area.

Cohort	N	Age (mean ± SD)	Sex	BSA (mean ± SD)	Current smoker	Log thyroid volume (mean ± SD)	Goiter cases(%)
SHIP START	3,611	49.1 (16.3)	47.4% Male, 52.6% Female	1.9 (0.2)	31.6%	2.9 (0.5)	1322 (36.6%)
SHIP-TREND	784	49.2 (13.9)	51.5% Male, 48.5% Female	1.9 (0.2)	22.2%	2.9 (0.4)	253 (32.3%)
SHIP-TREND batch II	2,499	51.3 (16)	56.9% Male, 43.1% Female	1.9 (0.2)	29.8%	2.9 (0.4)	737 (29.5%)

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

652 **Table 3.** Loci lead SNVs with the strongest association with thyroid volume (A), and
653 goiter risk (B) in GWAS analysis using both two-step imputed and conventionally
654 imputed genotypes.
655 Novel loci are marked in bold. A1: coded allele, AF: frequency of coded allele, SE:
656 standard errors, R2: linkage disequilibrium in the top SNVs of the same locus in both
657 imputation outcomes
658

A) Volume		Two-step imputation GWAS						Conventional imputation GWAS meta-analysis						
Locus	Lead SNP	CHR	A1	AF	Effect	SE	P-value	Lead SNP	CHR	A1	AF	Effect	SE	P-value	R2
CAPZB	rs4911994	1	A	0.64	0.057	0.01	1.63E-18	rs4911994	1	A	0.63	0.055	0.01	5.93E-19	Same SNP
	rs10799824	1	A	0.15	0.086	0.01	1.82E-22	rs12410532	1	T	0.14	0.092	0.01	6.14E-24	0.978
PAX8	rs7560701	2	C	0.48	-0.036	0.01	1.12E-08	rs1110839	2	T	0.48	0.038	0.01	6.35E-10	0.576
IGFBP5	rs2712172	2	A	0.27	0.041	0.01	1.46E-08	rs2712172	2	A	0.26	0.039	0.01	1.95E-08	Same SNP
NRG1	rs7000397	8	G	0.35	0.035	0.01	5.22E-08	rs7000397	8	G	0.34	0.04	0.01	1.42E-08	Same SNP
TG	rs114322847	8	T	0.03	0.122	0.02	1.90E-09	rs79676842	8	T	0.03	0.135	0.02	2.06E-11	1
FAM227B	rs17477923	15	C	0.26	0.054	0.01	9.55E-14	rs73398264	15	T	0.75	-0.052	0.01	7.73E-14	1
MAFTRR	rs3813579	16	A	0.53	0.056	0.01	2.57E-19	rs562609617	16	G	0.66	-0.064	0.01	1.85E-22	0.413
B) Goiter		Two-step imputation GWAS						Conventional imputation GWAS meta-analysis						
Locus	Lead SNP	CHR	A1	AF	Effect	SE	P-value	Lead SNP	CHR	A1	AF	Effect	SE	P-value	R2
CAPZB	rs4911994	1	A	0.64	0.614	0.68	8.89E-14	rs4911994	1	A	0.62	0.302	0.04	1.13E-13	Same SNP
	rs10799824	1	A	0.15	0.135	0.19	6.68E-16	rs12410532	1	T	0.14	0.450	0.06	1.25E-15	0.978
XKR6	rs7005680	8	T	0.35	-0.218	0.04	7.98E-09	rs11778398	8	T	0.46	-0.200	0.04	4.35E-07	0.761
FAM227B	rs75929244	15	T	0.24	0.321	0.04	2.17E-13	rs73398264	15	T	0.75	-0.313	0.04	1.61E-12	0.897
MAFTRR	rs3813579	16	GT	0.35	0.29	0.04	4.44E-13	rs562609617	16	G	0.34	0.064	0.01	1.85E-22	Same SNP

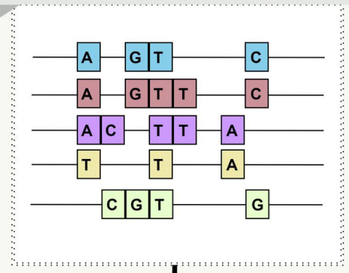
659

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

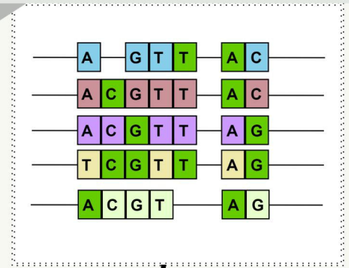
Genotyped cohorts included in the workflow

Cohort	N. Samples	Array Type	N. SNPs on the array
GANI_MED	2410	Illumina PsychArray	570.000
SHIP START	4070	Affymetrix SNP 6.0	900.000
SHIP START II	48	Affymetrix Axiom	560.000
SHIP-TREND	986	Illumina Omni 2.5	2.500.000
SHIP-TREND batch II	3133	Illumina GSA	655.000

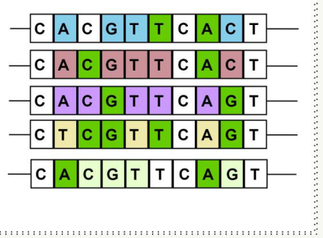
Preprocessed genotyped data



First (intermediate) imputation



Second imputation outcome



First Imputation Step

Genotype data preprocessing

Genotyping alignment and indels normalization

Haplotype phasing for each included panel

Phasing and preparing reference panel

Preparing a reference panel version of each panel

Imputation of each cohort against all other reference panels (n. cohorts - 1 imputation rounds)

Second Imputation Step

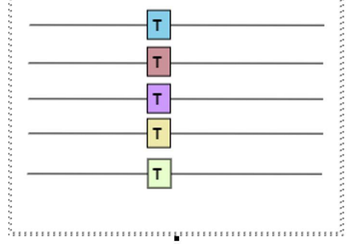
Filtering overlapped variants with high imputation quality (R^2 >= 0.9)

Imputation against 1000G reference panel

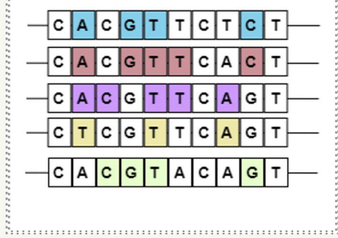
Conventional Imputation

Imputation against 1000G reference panel

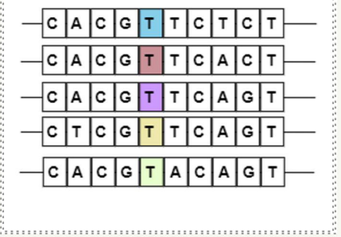
Overlapped genotypes only



Conventional imputation outcome



Imputation outcome



Imputation Outcome Evaluation

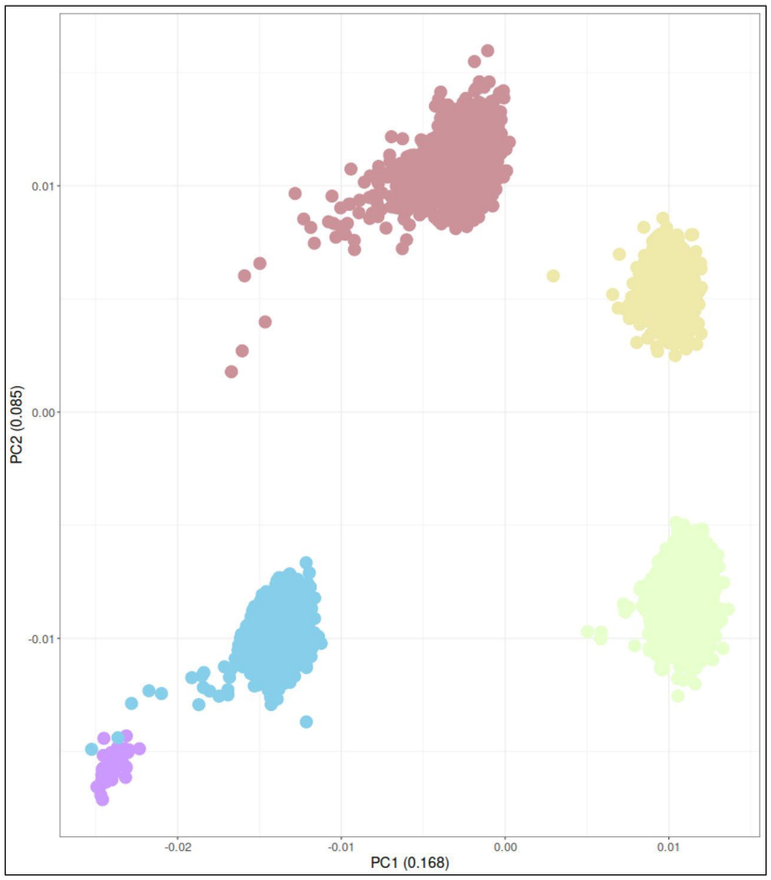
A Batch effect (PCA)

B Comparative analysis of variants information

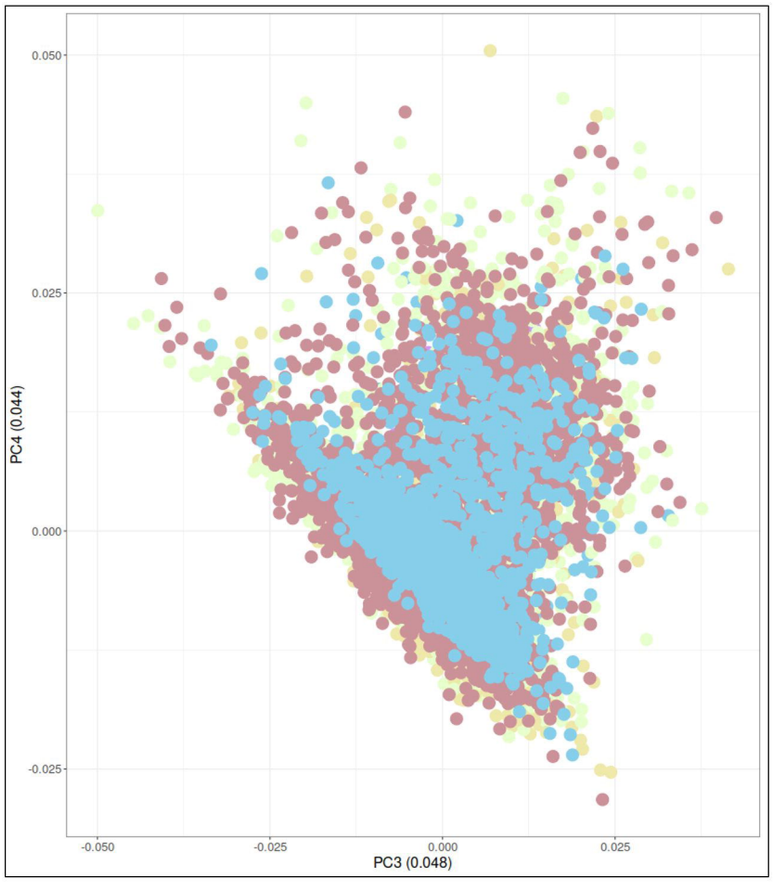
C GWAS results comparison (Thyroid function)

medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

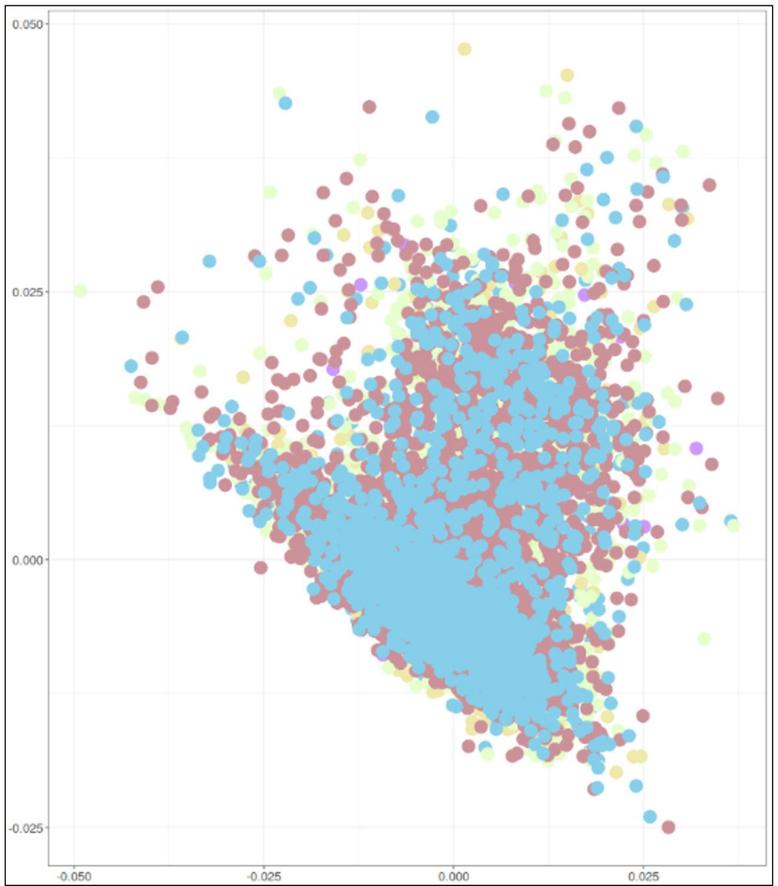
A)



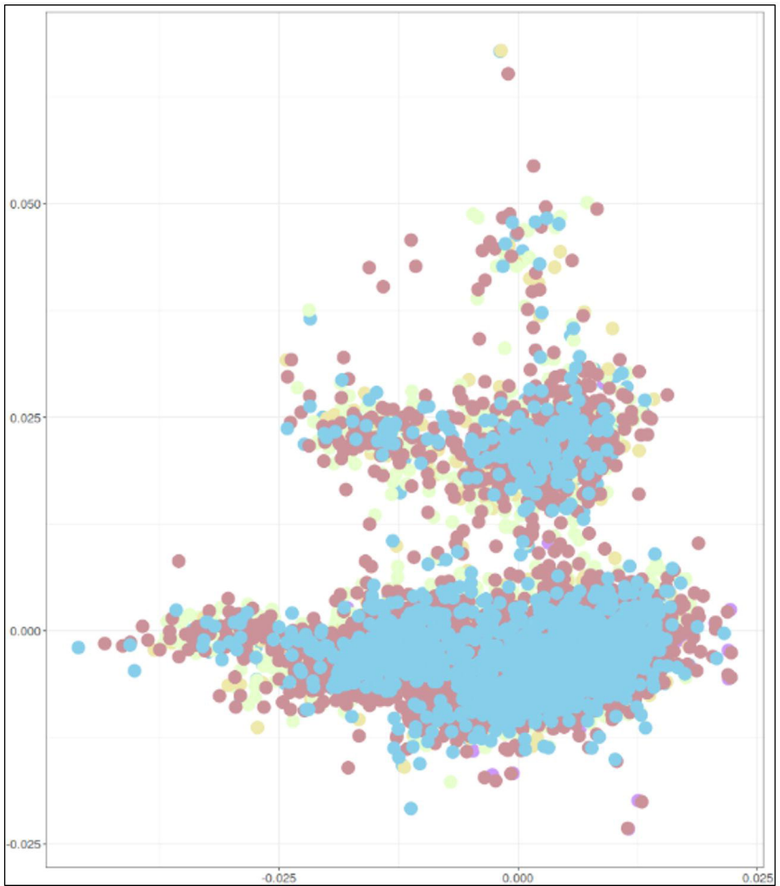
B)



C)

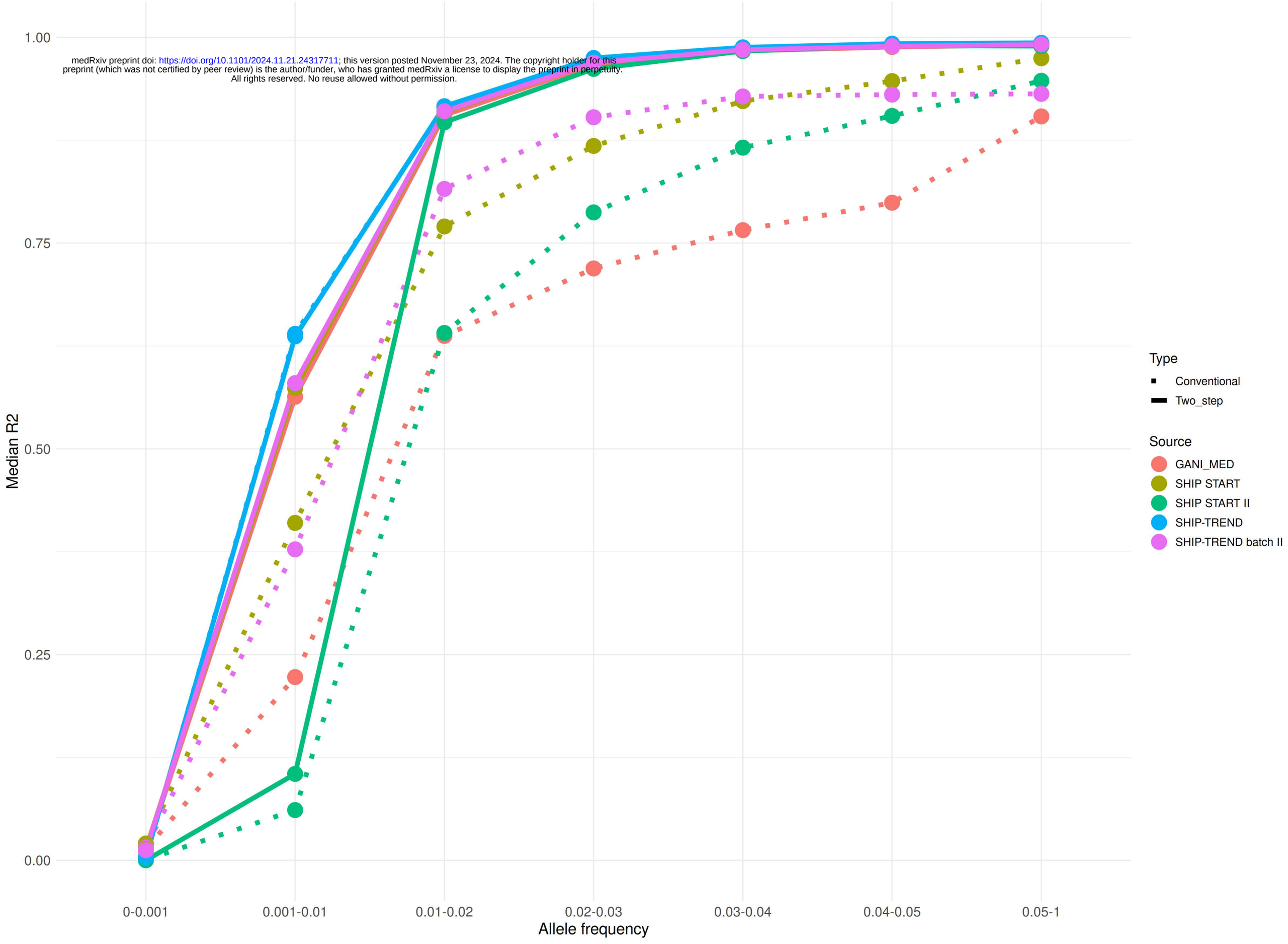


D)



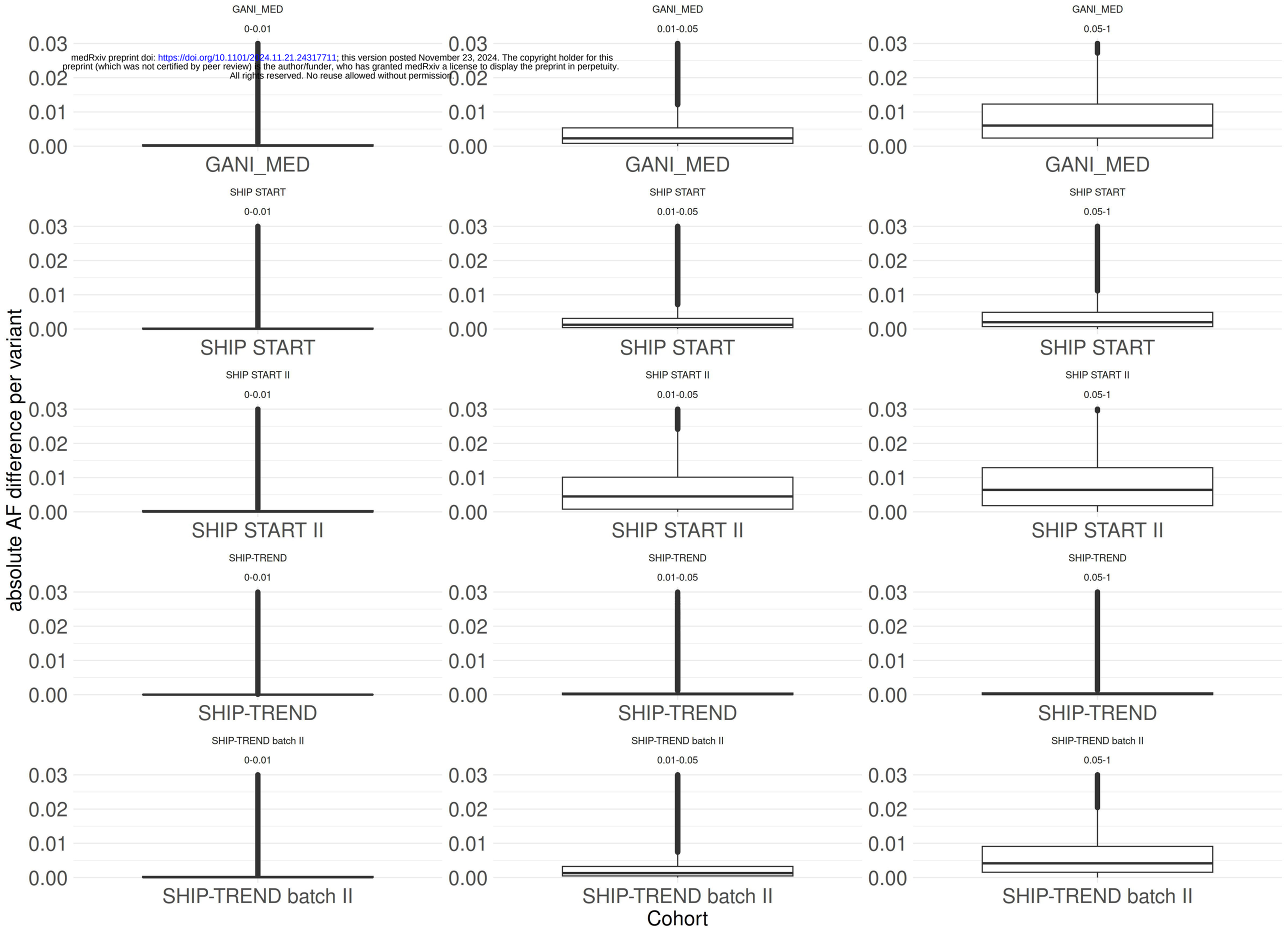
medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Median R2 with Quantiles by Type

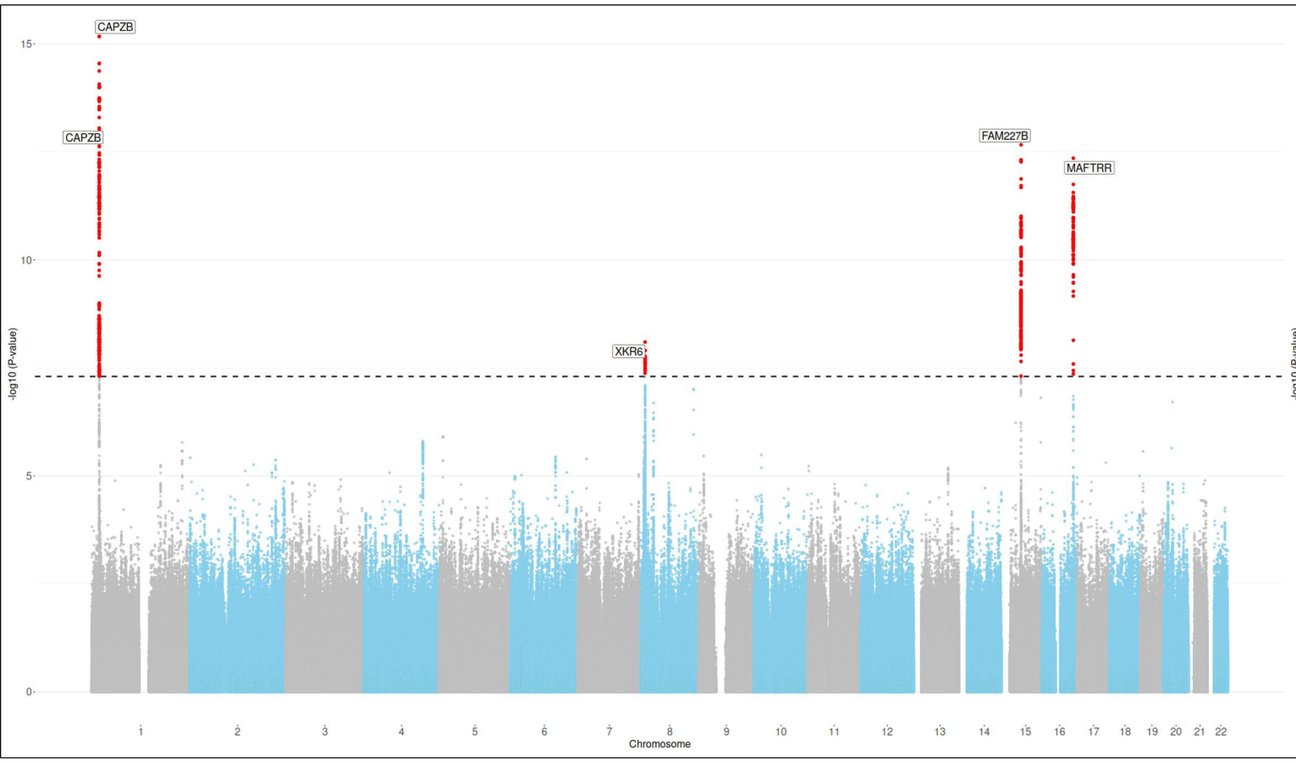


medRxiv preprint doi: https://doi.org/10.1101/2024.11.21.24317711; this version posted November 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

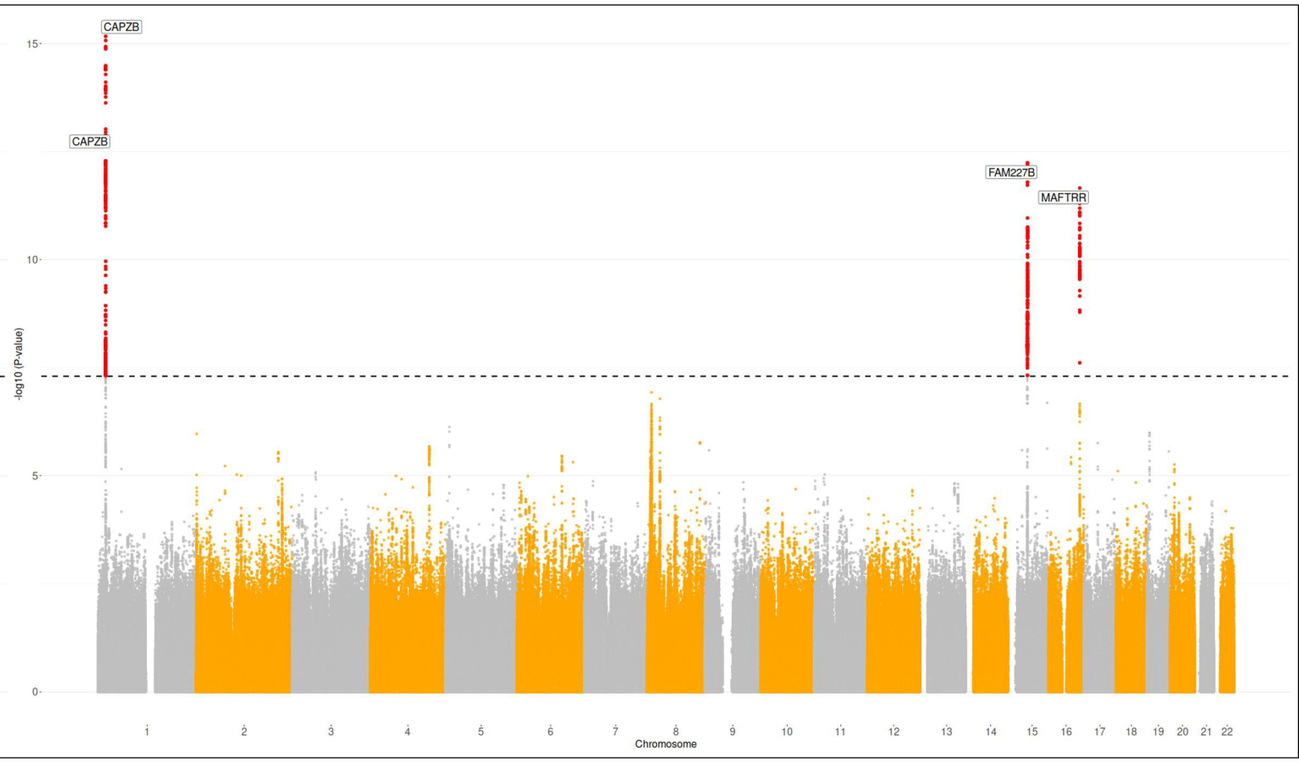
boxplot Plot of the absolute difference in AF between conventional and two-step imputation outcomes for the included cohorts



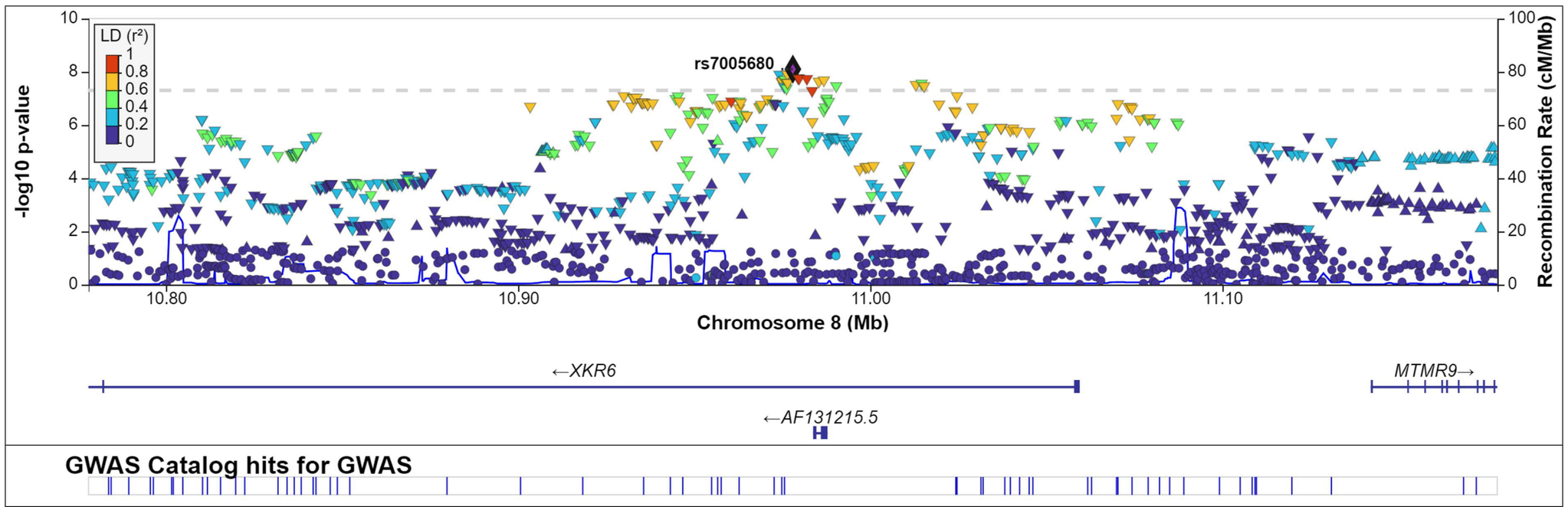
A)


CAPZB
CAPZB
XKR6
FAM227B
MAFTRR
-log10 (P-value)
Chromosome

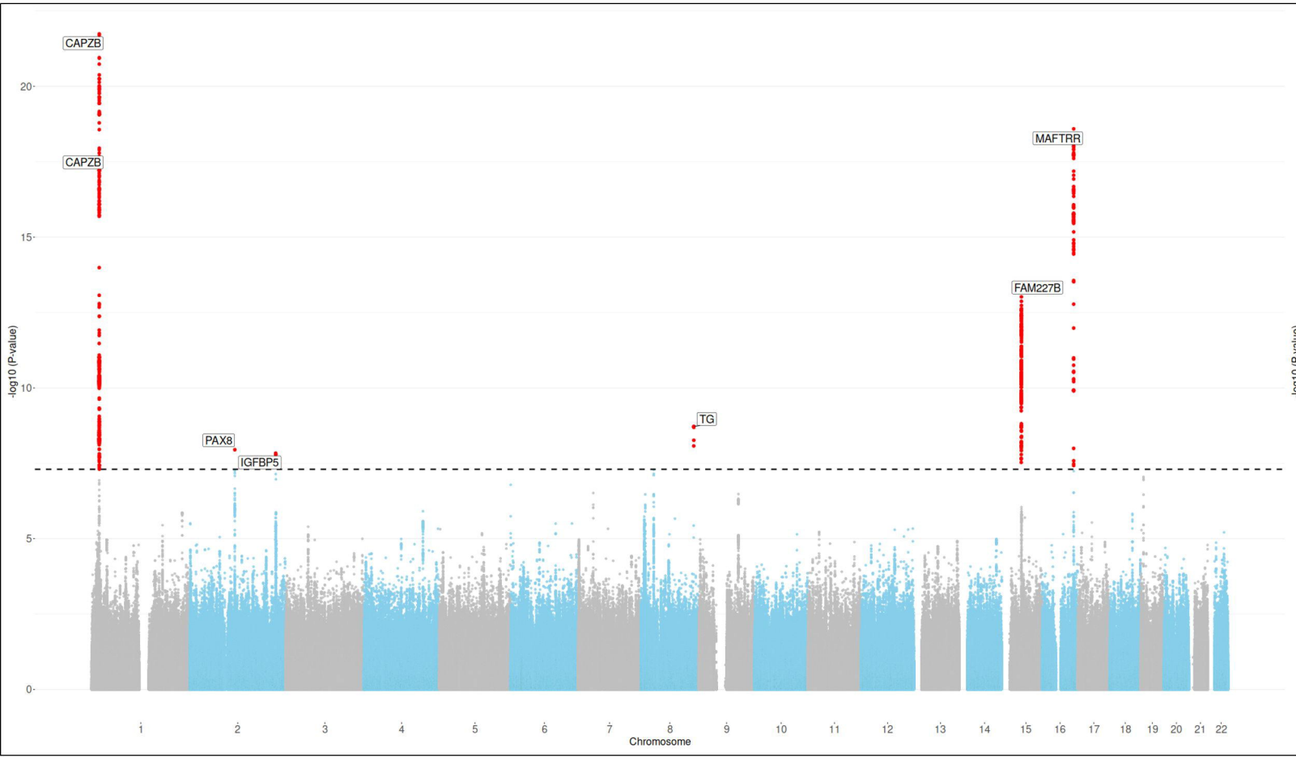
B)


CAPZB
CAPZB
FAM227B
MAFTRR
-log10 (P-value)
Chromosome

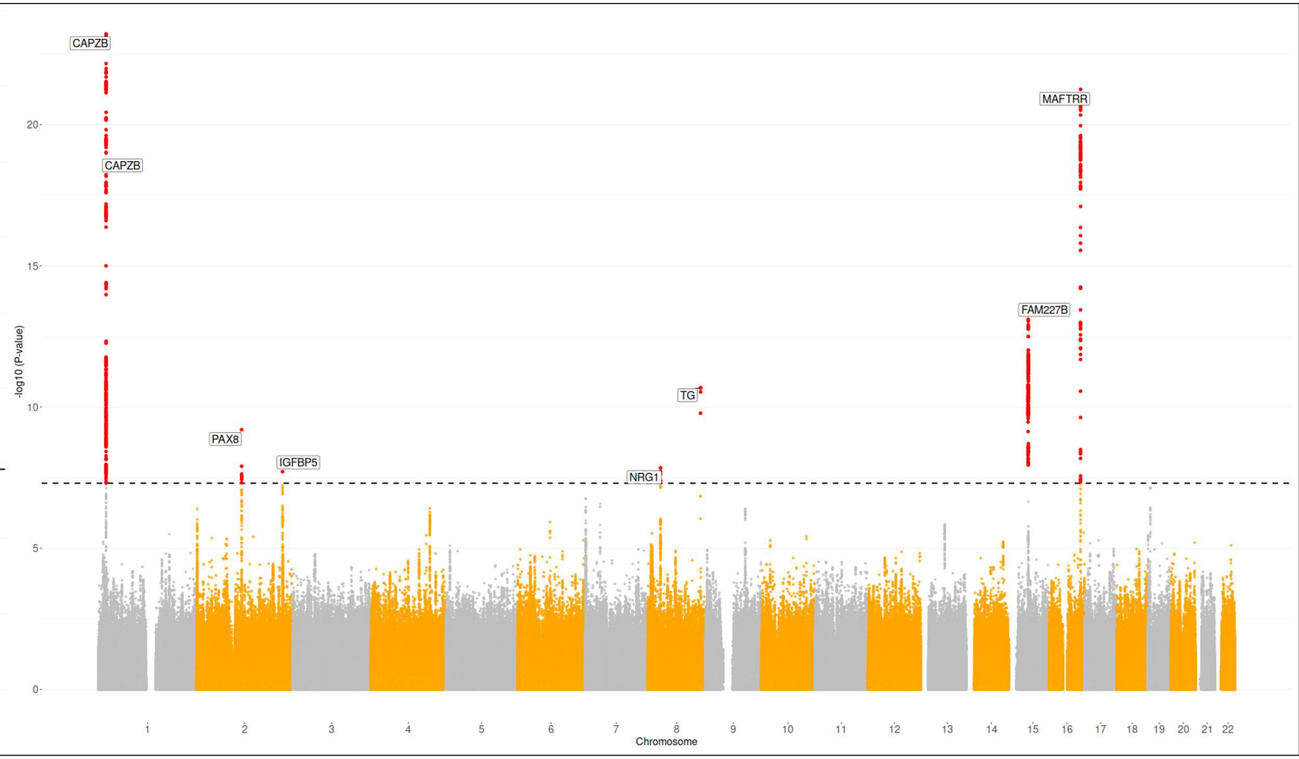
C)


LD (r²)
1
0.8
0.6
0.4
0.2
0
rs7005680
-log10 p-value
Recombination Rate (cM/Mb)
10.80
10.90
11.00
11.10
Chromosome 8 (Mb)
←XKR6
MTMR9→
←AF131215.5
GWAS Catalog hits for GWAS

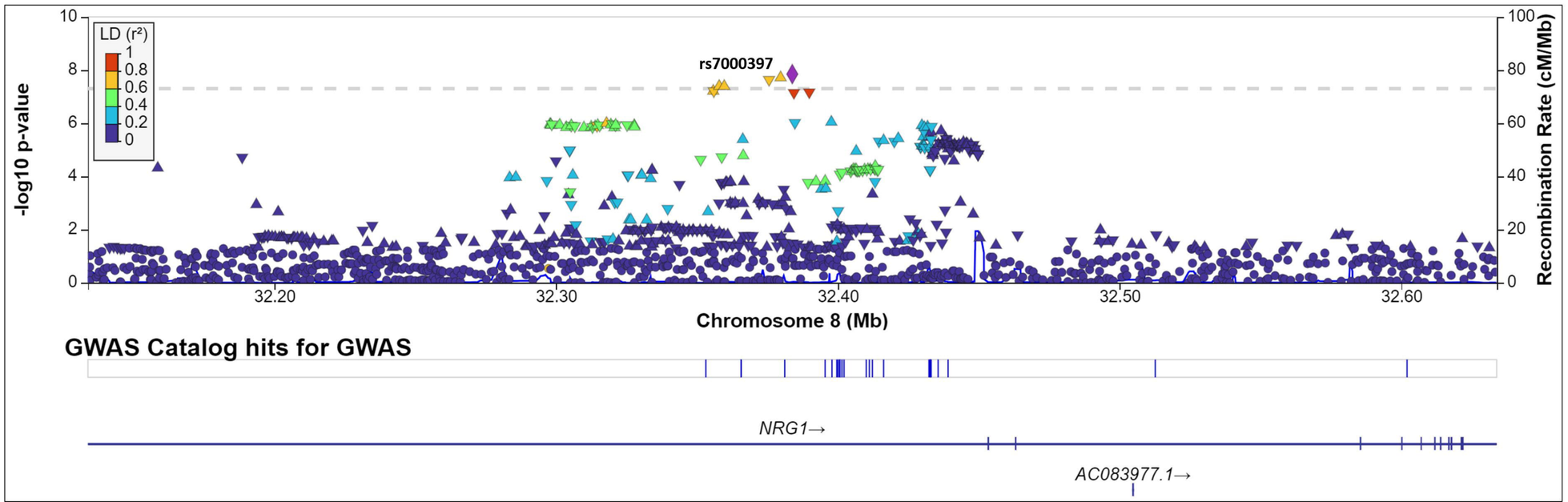
A)


CAPZB
CAPZB
MAFTRR
FAM227B
TG
PAX8
IGFBP5
-log10 (P-value)
Chromosome

B)


CAPZB
CAPZB
MAFTRR
FAM227B
TG
PAX8
IGFBP5
NRG1
-log10 (P-value)
Chromosome

C)


LD (r²)
rs7000397
-log10 p-value
Recombination Rate (cM/Mb)
Chromosome 8 (Mb)
GWAS Catalog hits for GWAS
NRG1→
AC083977.1→