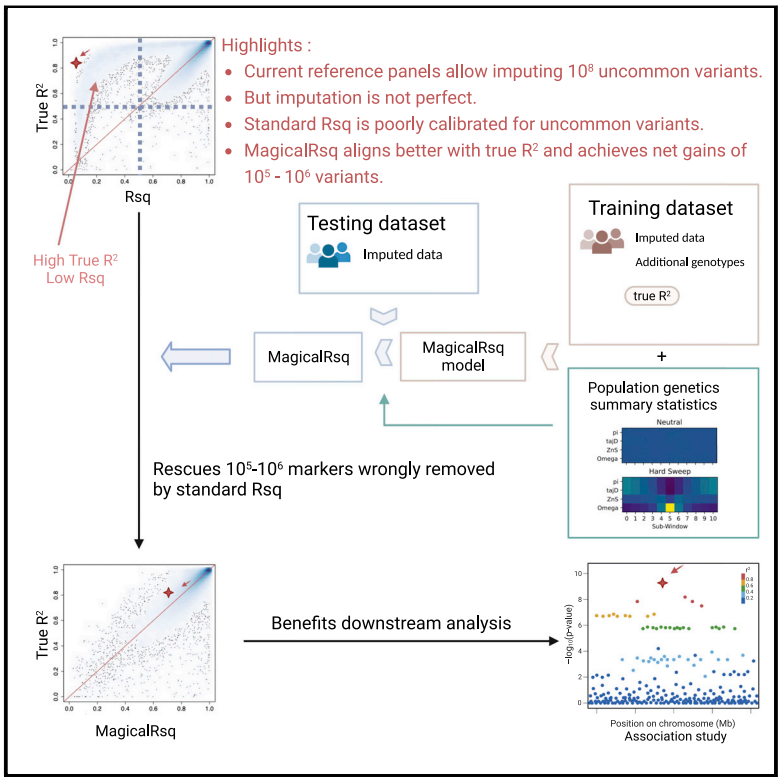
MagicalRsq: Machine-learning-based genotype imputation quality calibration

Graphical abstract



Authors

Quan Sun, Yingxi Yang, Jonathan D. Rosen, ..., Daniel R. Schrider, Christian Fuchsberger, Yun Li

Correspondence

cfuchsberger@eurac.edu (C.F.), yunli@med.unc.edu (Y.L.)

Ever-growing reference panels allow imputation of huge number ($\sim 10^8$) of lower-frequency variants. However, not all variants can be well imputed and standard imputation quality metric poorly reflects true imputation quality, particularly for uncommon variants. We present MagicalRsq, a machine-learning-based and better calibrated post-imputation quality metric, that can rescue 10^5–10^6 variants.

Sun et al., 2022, The American Journal of Human Genetics *109*, 1986–1997
November 3, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.ajhg.2022.09.009

CellPress

ARTICLE

MagicalRsq: Machine-learning-based genotype imputation quality calibration

Quan Sun,[1] Yingxi Yang,[2] Jonathan D. Rosen,[3] Min-Zhi Jiang,[4] Jiawen Chen,[1] Weifang Liu,[1] Jia Wen,[3] Laura M. Raffield,[3] Rhonda G. Pace,[5] Yi-Hui Zhou,[6] Fred A. Wright,[6,7] Scott M. Blackman,[8] Michael J. Bamshad,[9,10] Ronald L. Gibson,[9] Garry R. Cutting,[11] Michael R. Knowles,[5] Daniel R. Schrider,[3] Christian Fuchsberger,[12,13,*] and Yun Li[1,3,13,*]

Summary

Whole-genome sequencing (WGS) is the gold standard for fully characterizing genetic variation but is still prohibitively expensive for large samples. To reduce costs, many studies sequence only a subset of individuals or genomic regions, and genotype imputation is used to infer genotypes for the remaining individuals or regions without sequencing data. However, not all variants can be well imputed, and the current state-of-the-art imputation quality metric, denoted as standard Rsq, is poorly calibrated for lower-frequency variants. Here, we propose MagicalRsq, a machine-learning-based method that integrates variant-level imputation and population genetics statistics, to provide a better calibrated imputation quality metric. Leveraging WGS data from the Cystic Fibrosis Genome Project (CFGP), and whole-exome sequence data from UK BioBank (UKB), we performed comprehensive experiments to evaluate the performance of MagicalRsq compared to standard Rsq for partially sequenced studies. We found that MagicalRsq aligns better with true R^2 than standard Rsq in almost every situation evaluated, for both European and African ancestry samples. For example, when applying models trained from 1,992 CFGP sequenced samples to an independent 3,103 samples with no sequencing but TOPMed imputation from array genotypes, MagicalRsq, compared to standard Rsq, achieved net gains of 1.4 million rare, 117k low-frequency, and 18k common variants, where net gains were gained numbers of correctly distinguished variants by MagicalRsq over standard Rsq. MagicalRsq can serve as an improved post-imputation quality metric and will benefit downstream analysis by better distinguishing well-imputed variants from those poorly imputed. MagicalRsq is freely available on GitHub.

Introduction

Genotype imputation is a process of estimating missing genotypes with the aid of reference panel(s). It can effectively boost power for detecting associated variants in genome-wide association studies (GWASs) and narrow down the most strongly associated variants within a genomic region. The latest TOPMed freeze 8 reference panel[1] encompasses >300 million variants. However, not all variants that are available in a reference panel can be well imputed in a target cohort.[2–4] Therefore, post-imputation quality control (QC) is indispensable and critically important to distinguish well-imputed variants from poorly imputed ones. In current standard practice,[5–7] variant-level imputation quality metrics such as IMPUTE's INFO,[8] minimac's Rsq,[9] or Beagle's DR2[10] are adopted to perform such post-imputation quality filtering. Briefly, minimac's Rsq and IMPUTE's INFO estimate imputation quality by comparing observed variation in imputed data to expected conditional on allele, genotype, or haplotype frequencies, under the rationale that poorly imputed markers tend to have less than expected variation because lack of information will drag estimated genotype probabilities across individuals toward population average. Beagle calculates the squared Pearson correlation between allele dosages and best-guess genotypes to estimate true imputation quality. These metrics, though slightly different, are highly correlated.[11] Hereafter we refer to these standard quality metrics directly from imputation software as Rsq (or standard Rsq). These standard metrics have been proven to be effective discriminators of imputation quality for common variants (minor allele frequency [MAF] > 5%) but are less well calibrated for uncommon (MAF ≤ 5%) variants,[12–14] which are increasingly prevalent in continuously expanding reference panels.

Realizing that standard Rsq is less accurate for low-frequency (MAF in [0.5%, 5%]) and rare (MAF < 0.5%) variants,[12–15] researchers have explored different strategies to deal with this issue. For example, Liu et al.[13] proposed a simple solution: adoption of more stringent thresholds of

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [2]Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA; [3]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [4]Department of Applied Physical Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [5]Marsico Lung Institute/UNC CF Research Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [6]Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA; [7]Bioinformatics Research Center and Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA; [8]Division of Pediatric Endocrinology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA; [9]Department of Pediatrics, University of Washington, Seattle, WA 98105, USA; [10]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [11]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA; [12]Institute for Biomedicine, Eurac Research (affiliated with the University of Lübeck), Bolzano, Italy
[13]These authors contributed equally
*Correspondence: cfuchsberger@eurac.edu (C.F.), yunli@med.unc.edu (Y.L.)
https://doi.org/10.1016/j.ajhg.2022.09.009.
© 2022 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

standard Rsq for variants with lower MAF. Similarly, Coleman et al.[16] proposed a procedure to assign the standard Rsq threshold taking MAF into account by using different inflection points for variants within different MAF bins. Both attempted to utilize different thresholds of standard Rsq rather than seeking alternative imputation quality measures. Lin et al.[15] address the well-known effect of allele frequency on imputation quality by proposing a new statistic named the imputation quality score (IQS). This score adjusts the MAF-dependent expected concordance between imputed and true genotypes. However, the computation of IQS requires true genotypes for the SNPs of interest, which is impractical in most situations.

In this paper, we propose MagicalRsq, a machine-learning-based genotype imputation quality calibration, by using eXtreme Gradient Boosted trees (XGBoost)[17] to effectively incorporate information from various variant-level summary statistics. MagicalRsq requires true R^2 information for a subset of individuals and/or a subset of markers (we hereafter refer to both as additional genotypes) to train models that can be applied to all target individuals and all markers. We note that it is rather common that investigators sequence only a subset of markers (genome regions) or individuals due to cost considerations. For example, the NHLBI GO Exome Sequencing Project[18] and Exome Aggregation Consortium (ExAC)[19] have generated only whole-exome sequencing (WES) data thus far. In addition, some WGS efforts, such as the TOPMed project, have sequenced only subsets of individuals in the constituent cohorts and array-genotyped the remainder.[3] With the availability of additional genotype data and our MagicalRsq framework, we can leverage the extra information to improve the calibration of imputation quality.

Leveraging WGS data generated by the Cystic Fibrosis Genome Project (CFGP)[2] and WES data from UKB,[20] we demonstrate that MagicalRsq substantially outperforms standard Rsq and is a more informative metric for post-imputation QC, particularly for lower-frequency variants. MagicalRsq performs well in the evaluated European and African ancestry cohorts. We carried out comprehensive experiments to mimic different real-life scenarios and showed that MagicalRsq is superior to standard Rsq in almost every scenario. Moreover, as a post-imputation QC metric, MagicalRsq could achieve net gains of thousands or millions of variants that are either well imputed but incorrectly filtered out or poorly imputed but incorrectly retained by standard Rsq. We anticipate MagicalRsq will benefit downstream analysis by better distinguishing well-imputed variants from poorly imputed ones.

Materials and methods

Ethics statement

This research has been conducted using the UK Biobank Resource under Application Number 25953. Furthermore, this study was reviewed by the Cystic Fibrosis Foundation for the use of CF Foundation Patient Registry data and CFGP WGS data. The procedures followed for data collection and processing, DNA sequencing, and analysis were in accordance with the ethical standards of the responsible human rights committees on human experimentation, and proper informed consent was obtained from all individuals.

MagicalRsq model

MagicalRsq adopts the eXtreme Gradient Boosted trees (XGBoost) method[17] for better calibration of imputation quality score, particularly for uncommon variants. Tree boosting is a commonly used machine learning approach that has been applied for a wide range of problems.[21–24] MagicalRsq is a supervised learning method where we build models to predict true imputation quality (true R^2, squared Pearson correlation between imputed and true genotypes) using a battery of variant-level summary statistics. As a supervised method, MagicalRsq requires true genotypes at a reasonable number of variants (>10k) to derive their true imputation quality scores to train models. For both performance and computational considerations, we build three models separately for common (MAF > 5%), low-frequency (MAF in [0.5%, 5%]) and rare (MAF < 0.5%) variants in our model training process. Note that MAF is estimated based on imputed dosages to better represent realistic settings where we don't have true WGS genotypes to calculate MAF. As shown in Figure 1, starting from genotype array data in training dataset, minimac imputation is performed to obtain standard Rsq and estimated MAF from imputed dosages. After imputation, using true genotypes at imputed markers not on the initial genotype array, we calculate true imputation quality, i.e., true R^2, at these imputed markers. The standard Rsq, estimated MAF, and true R^2 will then be carried forward to model training. Note that these statistics are specific for a given training dataset and will vary when using different datasets for model training purposes.

Besides the aforementioned imputation summary statistics in the training samples, we also consider multiple variant-level features that are not specific to the training dataset that we use. We first include multiple population genetics features as they are known to impact genotype imputation.[9,11,25,26] These population genetics statistics reflect various aspects encompassing haplotype structure, linkage disequilibrium (LD) profile, and the spatial pattern of site frequency spectrum (SFS) structure. Specifically, we consider 11 population genetics summary statistics (Note S1) for six populations (CEU, GWD, JPT, LWK, PEL, and YRI) corresponding to diverse continental ancestry groups including European (EUR), African (AFR), East Asian (EAS), and American (AMR). These 11 population genetics summary statistics, calculated by the positive selection scan program S/HIC,[27] are π,[28] $\widehat{\theta}_H$,[29] Tajima's D,[30] Fay and Wu's H,[29] the maximum frequency of derived mutations (MFDM),[31] the number of distinct haplotypes,[32] Garud et al.'s[32] haplotype homozygosity statistics (H1, H2, H2/H1), Kim and Nielsen's ω,[33] and Kelly'sZ_{nS};[34] all initially calculated for 100Kb non-overlapping bins.[27] We obtain variant-level statistics by application of their corresponding region-level statistics. In addition, considering the known effects of allele frequency on imputation quality reported in previous studies,[13,35] we also incorporate external alternative allele counts (AC) per 1,000 samples and MAF in four major ancestral groups (EUR, AFR, EAS, and South Asians [SAS]) derived from TOPMed WGS data in the TOP-LD project,[36] and 2nd–4th moments (representing variance, skewness, and kurtosis) for the four ACs. Inclusion of these variant-specific features distinguishes MagicalRsq from standard Rsq and other imputation quality

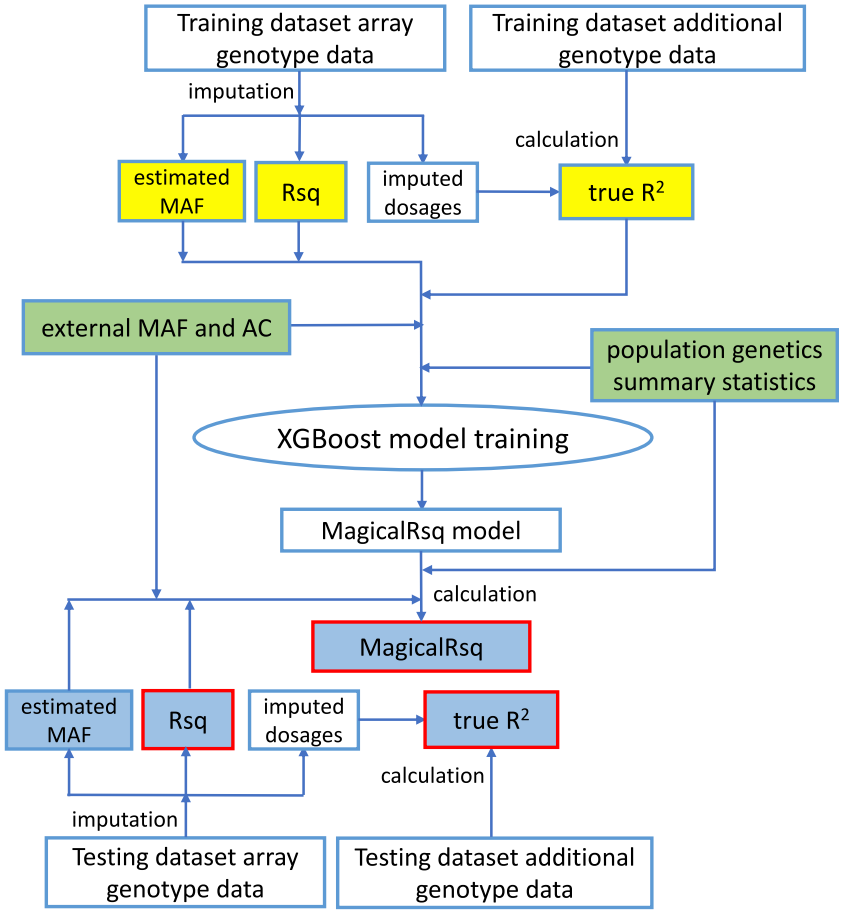


Figure 1. MagicalRsq workflow

MagicalRsq starts from "training dataset array data" (which are data used for imputation among training individuals) and performs imputation using these data, which gives us standard Rsq and estimated MAF for each marker, in the training dataset. Then we calculate the true R^2 by comparing imputed dosages with truth genotypes (established by additional genotype data in the training set). Combining external MAF and alternative allele count (AC), as well as population genetics summary statistics, with the above three metrics (i.e., standard Rsq, estimated MAF, and true R^2), we train MagicalRsq models using the XGBoost method where we build supervised models to predict true R^2 from all the other features. We then proceed to the testing dataset where we follow the same imputation workflow starting again from array genotype data and obtaining estimated MAF and standard Rsq after imputation. We then calculate MagicalRsq in the testing dataset by plugging in the predictor features into the MagicalRsq models built from the training dataset. Finally, we evaluate the performance of MagicalRsq (and Rsq) by comparing with true R^2 in the testing dataset. Yellow highlights represent all the instruments specific for the training dataset, light blue highlights represent the instruments specific for the testing dataset, green highlights represent external information used in both training and testing, and red rectangles represent statistics used during final evaluation and comparison of MagicalRsq and standard Rsq, using true R^2 as the gold standard.

metrics that utilizes only imputation summary statistics.[15] We then treat each variant as an observation and build XGBoost models to predict true R^2 by leveraging these 79 variant-level features.

Since the true imputation quality (i.e., true R^2) is between 0 and 1, we specify the learning task to be tree-based logistic regression and evaluation metric to be root mean square error. To control overfitting, we set early stopping rounds to be 50, i.e., training would stop if the performance did not improve for 50 rounds in an independent validation set.

True imputation quality calculation

We quantify the true imputation quality metric using true R^2, which is the squared Pearson correlation between imputed dosages and true genotypes. The true genotypes, coded as 0, 1, and 2, are obtained from the additional genotype data not used for imputation, for example, the WGS data from CFGP or the WES data from UKB. Our evaluation was restricted only to samples with after QCed (QC+) data from both imputed and additional genotype data. Duplicate samples were dropped. Finally, true R^2 was calculated for each variant, based on overlapping QC+ samples.

Imputation metric evaluation

We evaluate imputation quality metrics (standard Rsq or MagicalRsq) in testing dataset(s) independent of those used for training. For evaluation, we treat true R^2 as the truth and quantify the performance of each quality metric by calculating the squared Pearson correlation, root mean squared error (RMSE), and mean absolute error (MAE) with true R^2.

We further evaluate the performance of MagicalRsq in comparison to standard Rsq in terms of their effectiveness to distinguish between well- and poorly imputed variants. Specifically, using a commonly used 0.8 cutoff,[2,3,13,37,38] we compare the numbers of (1) well-imputed variants saved by MagicalRsq, defined as true $R^2 \geq 0.8$, Rsq < 0.8, and MagicalRsq ≥ 0.8; (2) well-imputed variants missed by MagicalRsq, defined as true $R^2 \geq 0.8$, Rsq ≥ 0.8, and MagicalRsq < 0.8; (3) poorly imputed variants excluded by MagicalRsq, defined as true $R^2 < 0.8$, Rsq ≥ 0.8, and MagicalRsq < 0.8; and (4) poorly imputed variants included by MagicalRsq, defined as true $R^2 < 0.8$, Rsq < 0.8, and MagicalRsq ≥ 0.8. We then calculate the net gains of variants when applying MagicalRsq for post-imputation QC compared to Rsq, defined as (1) − (2) + (3) − (4).

Data description

The Cystic Fibrosis Genome Project (CFGP) aims to identify genetic modifiers of cystic fibrosis (CF) traits by leveraging WGS data and rich phenotypic information collected through the Cystic Fibrosis Foundation Patient Registry (CFFPR).[39] CFGP high-coverage (~30×) WGS data are available for 5,109 samples representing 5,072 unique individuals. The WGS data contain approximately 90m and 11m high-quality single-nucleotide

variants and indels, respectively. The resource represents the largest cohort of CF-affected individuals for which WGS data are available. 5,095 samples representing 5,058 unique individuals remained after sample identity check.[2]

The UK Biobank (UKB),[40] recruiting ~500,000 people aged between 40 and 69 years in 2006–2010, is a prospective biobank study to study risk factors for common diseases such as cancer, heart disease, stroke, diabetes, and dementia. Participants have been followed up through health records from the UK National Health Service. UKB has genotype data on all enrolled participants, as well as extensive baseline questionnaires and physical measures and stored blood and urine samples. Specifically, genotyping array data at ~800,000 directly genotyped markers are available for ~500,000 UKB participants, among whom ~200,000 also have WES data.[20]

MagicalRsq experiments

To comprehensively evaluate the performance of MagicalRsq, we performed 14 experiments (Table S2) with 9 imputations (Table S1), leveraging CFGP WGS data[2] and UKB WES data.[41] These experiments can be categorized in two scenarios. In the first scenario, we have additional directly typed markers (other than array genotypes) available in *all* target individuals, while in the likely more realistic second scenario, we have additional directly typed markers in only a *subset* of target individuals.

Imputation overview

For CFGP data, our training dataset contains 1,992 CF samples (hereafter denoted as the CF 2k samples) who have both Illumina 610-Quad array data with 567,784 variants after QC[2] and WGS data. An independent set of 3,103 CF samples (hereafter denoted as the CF 3k samples) have WGS data but no array genotypes available. To perform imputation for the latter set of 3,103 samples, we first thinned their WGS data to Illumina 660W array by keeping only the 551,819 QC+ variants overlapped with the 660W array. For the UKB samples, we identified 9,354 participants with significant African ancestry (UKB AFR) and array genotype data available following our previous work.[4] Among them, 3,960 individuals also had WES data in the 200k WES release.

Genotype imputation was then performed by uploading QC+ array or thinned WGS data to TOPMed imputation server or Michigan imputation server (web resources), using TOPMed freeze 8 or the 1000 Genomes phase 3 (1000G) as the reference panel, respectively. Phasing was performed using Eagle 2[42] and imputation was performed by Minimac 4.[9]

Scenario 1: Availability of additional directly typed markers in all target individuals

We first consider the scenario where all target samples have additional genotype data in addition to the genotyping array data used for imputation. Under this scenario, we first imputed CF 2k samples using their genotypes from Illumina 610-Quad array[2] and the 1000G reference panel (Table S1, imputation 1). These individuals also have WGS data which we did not use for imputation but rather used only for obtaining true R^2. These true R^2 values were used to build models in training "samples" and to evaluate performance in testing "samples," where "samples" here refer to variants. To avoid information leakage, we trained using variants on even-numbered chromosomes and tested on odd-numbered chromosomes (Table S2, experiment 1). We then imputed the same individuals with the TOPMed freeze 8 reference panel (Table S1, imputation 2) and re-built MagicalRsq models (Table S2, experiment 2) following the same procedures as described above.

We further attempted to relax the MagicalRsq model training requirements to either mix-and-match imputation reference panels or variants in a region (in contrast to genome- or chromosome-wide). Using mix-and-match reference panels, we tried to examine whether MagicalRsq models trained using imputed data from one reference panel can be applied to imputed data from a different reference panel. Specifically, we trained MagicalRsq models using 1000G-imputed variants on odd-numbered autosomes and applied to TOPMed-imputed variants on even-numbered autosomes (Table S2, experiment 3) and vice versa from TOPMed training to 1000G testing (Table S2, experiment 4), in the same 2k CF samples.

By using variants in a region, we tried to mimic realistic scenarios where only some regions (e.g., selected gene(s), selected loci near GWAS regions, exonic regions) are sequenced. Specifically, we assumed that our 2k CF samples are only additionally (besides GWAS array) sequenced in the ±10 Mb region around the *CFTR* (chr7:107480144–127668447, hg38) and explored training models only using variants in this region, from TOPMed imputed data (Table S2, experiment 5). Note that mutations in the *CFTR* on chromosome 7 provide the molecular basis of CF and thus a region of high importance in this dataset. We trained MagicalRsq models using 183k variants in this 20 Mb region and applied the models to all other chromosomes. We similarly separated the variants into three MAF categories, leaving 32k common variants, 24k low-frequency variants, and 127k rare variants. We also tried training models with variants on another two regions: chr10:80–100 Mb and chr20:20–40 Mb to assess whether MagicalRsq is robust to different region choices. We then performed additional experiments with 1000G imputed variants on the same three regions (Table S2, experiment 6) to evaluate whether MagicalRsq models trained from regional (versus genome- or chromosome-wide) variants are robust to reference panel choices.

Finally, we trained MagicalRsq models using only exonic variants to assess whether exome-trained models could be applied to other genomic regions. We first retained variants from the CF 2k TOPMed imputed data which are also present in the UKB WES data, leaving 89k common variants, 82k low frequency variants and 635k rare variants. We trained models using these WES variants, and applied to other variants from the same TOPMed imputed data in the CF 2k samples (Table S2, experiment 7).

Scenario 2: Availability of additional directly typed markers in a subset of target individuals

We next consider the scenario where we have additional markers directly assayed in only a subset of samples, while the remaining samples still need imputation. We again leveraged CFGP WGS data. As previously mentioned, we have the CF 2k samples with both the Illumina 610-Quad array genotypes and WGS data and an independent CF 3k samples with WGS data only. We thinned the WGS data of the CF 3k samples to the Illumina 660W array density and performed imputation (Table S1, imputation 3), using TOPMed freeze 8 as reference. We then applied the MagicalRsq models trained from the CF 2k samples to the 3k samples (Table S2, experiment 8).

We additionally leveraged the WES data of UKB AFR participants to assess MagicalRsq's performance for African-ancestry individuals. We imputed 3,960 UKB AFR samples using both 1000G and TOPMed reference panels (Table S1, imputation 4–7). All these 3,960 individuals also have WES data available. We randomly

selected 1,000 samples for MagicalRsq model training and used the remaining for testing, and again trained models separately for common, low-frequency, and rare variants (Table S2, experiments 9 and 10). We also investigated mix-and-match reference panels under this scenario (Table S2, experiments 11 and 12, Note S2). To assess whether other under-represented populations can also benefit from MagicalRsq, we also imputed 4,436 UKB South Asian (SAS)-ancestry individuals using TOPMed freeze 8 reference panel (Table S1, imputation 9). We similarly randomly selected 1,000 samples for building MagicalRsq models and applied the models to the remaining individuals for testing, using genotypes from WES data as truth (Table S2, experiment 13).

We further examined the performance of MagicalRsq when training models using a small subset of variants, instead of all variants available. Specifically, we trained models in the CF 2k samples with different numbers of variants on even-numbered chromosomes and applied such models to the independent 3k samples as described before (Table S2, experiment 14). We randomly selected 10k, 50k, 100k, 200k, 500k, and 1m variants in each of the three MAF categories for model training and repeated five times for more stable inference.

Results

Scenario 1: Availability of additional directly typed markers in all target individuals

We first consider a simple yet realistic scenario where all target samples have additional genotype data in addition to the genotyping array data used for imputation. These additional data could be from a different genotyping array, from WES, or from targeted genotyping or sequencing. We do not consider the scenario where all individuals already have deep WGS data because imputation would no longer be relevant.

Feature importance

We first imputed CF 2k samples using the 1000G reference panel, trained MagicalRsq models using variants on even-numbered chromosomes, and tested on odd-numbered chromosomes (Scenario 1). We performed feature importance analysis to find out major contributors to the training models. We found, as expected, that standard Rsq is by far the most important feature, weighing ~80% among all the features (Figure S1). European AC is the second most important feature for common variants, while for low frequency and rare variants, the second most important feature is African AC. Because the CF cohorts are predominantly of European ancestry,[2] it is not surprising that European AC is influencing the training models. For rarer variants, we suspect the importance of African AC, likely due to the African allele frequency spectrum better capturing rarer variants among individuals with CF.

MagicalRsq outperforms standard Rsq

Compared to standard Rsq, MagicalRsq is more consistent with the true R^2 with respect to all three of the evaluation metrics (squared Pearson correlation with true R^2, RMSE, and MAE) and for all three MAF categories (Figures 2A and S2). For example, when we compare MagicalRsq to standard Rsq, squared Pearson correlation with true R^2 increases by 8.7%–24.1% for common variants, 10.1%–62.7% for low-frequency variants, and 14.3%–17.2% for rare variants (Table S4), across different chromosomes. When using MagicalRsq to replace standard Rsq for post-imputation QC, we have net gains of 33k common variants, 34k low-frequency variants, and 200k rare variants across half of the genome (i.e., even numbered chromosomes) (Table S3 experiment 1). We then imputed the same individuals with the TOPMed freeze 8 reference panel (Table S1, imputation 2), re-built MagicalRsq models (Table S2, experiment 2) following the same procedures, and observed similar improvement (Figures 2B and S3). For example, MagicalRsq increases squared Pearson correlation with true R^2 by 1.7%–34.5% for common variants, 37.6%–71.0% for low-frequency variants, and 10.3%–14.3% for rare variants (Table S5), and the net gains of rare variants increase to 411k (Table S3, experiment 2) due to more rare variants in the TOPMed reference panel. Comparing across the three MAF categories, we observe that MagicalRsq shows the most pronounced improvement for low-frequency variants in terms of squared Pearson correlation, regardless of the imputation reference panels.

Impact of reference panel

We further attempted mix-and-match of the imputation reference panels in training and testing sets to examine whether MagicalRsq trained using imputed data from one reference panel can be applied to imputed data from a different reference panel (Methods). We found that applying 1000G-trained models to TOPMed data performs less well for low-frequency and rare variants, while applying TOPMed-trained models to 1000G data performs less well for common and low-frequency variants (Table S3, experiments 3 and 4). We suspected that reference-panel-specific variants may hinder the prediction accuracy, and then evaluated the mix-and-match performances by restricting to variants shared between the two reference panels. We found that the mix-and-match MagicalRsq were better calibrated than standard Rsq for the shared variants (Note S2), although the improvements are less pronounced than using the matched reference panel (Figure 2C, Table S6).

Number of variants needed for training

In the previous experiments MagicalRsq training models were built on a large subset of markers genome-wide (e.g., all even chromosomes, corresponding to 2.6m, 1.9m, and 12.3m common, low-frequency, and rare variants in TOPMed imputed CF 2k samples). To mimic the scenario where only some regions (e.g., selected gene(s), selected loci near GWAS regions, exonic regions) are sequenced, we trained MagicalRsq models using TOPMed imputed variants on three ~20 Mb regions: ±10 Mb region around the *CFTR* (chr7:107480144–127668447, hg38), and two randomly

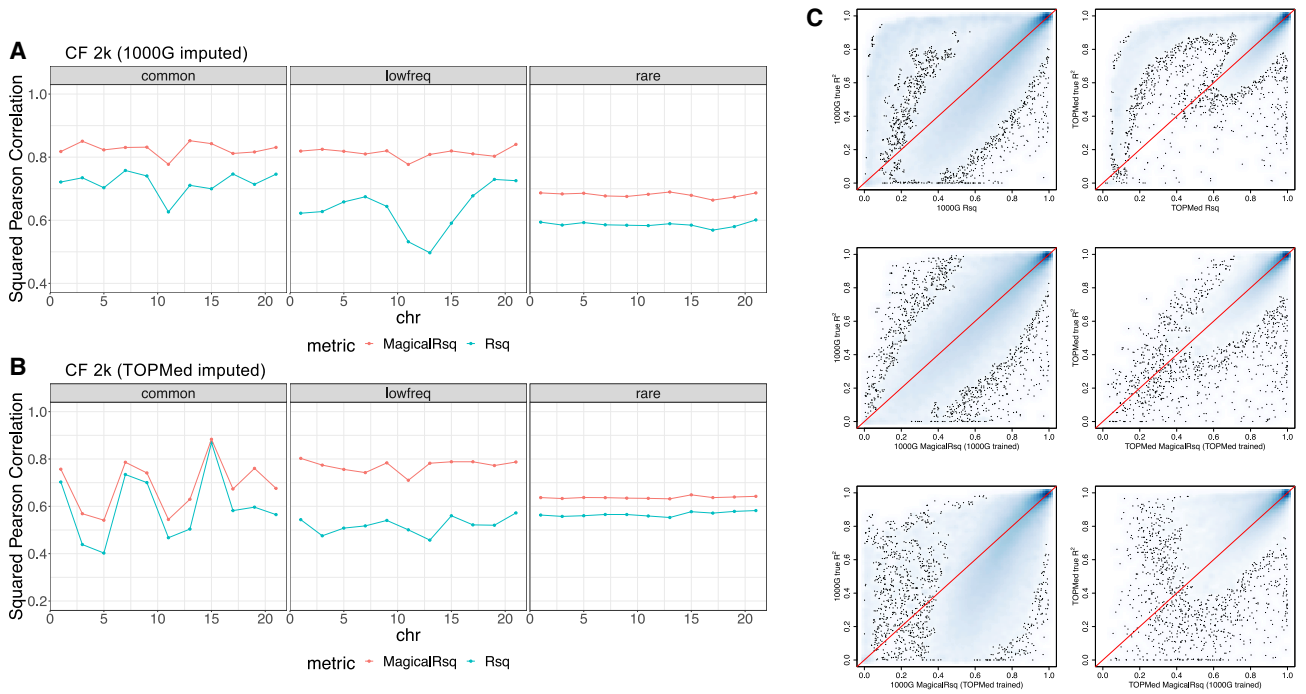


Figure 2. Scenario 1, experiments 1–4: Training using even-numbered chromosomes and testing on odd-numbered chromosomes for CF 2k samples

(A and B) Performance comparison between Rsq and MagicalRsq in terms of squared Pearson correlation with true R^2 for (A) 1000G-based imputation; (B) TOPMed-based imputation.

(C) Smooth scatterplot showing Rsq or MagicalRsq (x axis) calculated from both matched- (second row) and mis-matched- (third row) models against true R^2 (y axis) for both 1000G-based (left) and TOPMed-based (right) imputation, for low-frequency variants on chromosome 13.

selected regions: chr10:80–100 Mb and chr20:20–40 Mb regions (Table S2, experiment 5). We found that the models performed reasonably well for low-frequency and rare variants, but not for common variants (Figure S6), likely due to the larger fluctuation of Rsq performance for common variants across the genome (Note S2). We then performed additional experiments with 1000G imputed variants on the same regions (Table S2, experiment 6), and the results are highly consistent (Figure S7).

Apply exome-trained models to other genomic regions

As WGS is still not yet available for most studies, many researchers are generating WES data as an alternative.[19] We trained our MagicalRsq models using only variants in exomes to test whether such models are generalizable to other genomic regions. Specifically, we trained our models using exonic variants from TOPMed imputed CF 2k cohort (Scenario 1) and applied to variants outside of the exomes (Table S2, experiment 7). The results showed that the WES-trained models improved squared Pearson correlation with true R^2 by 3.5%–28.4% for common variants, by 17.8%–68.5% for low-frequency variants, and by 10.1%–13.8% for rare variants (Table S7, Figure S8). We note that the observed improvements are similar to experiment 2 where models were built on half of the genome, indicating that MagicalRsq models perform similarly well when trained with only exonic variants.

Scenario 2: Availability of additional directly typed markers in a subset of target individuals

The previous scenario assumes that all individuals have additional genotyping (through other genotyping arrays, candidate gene sequencing, WES, etc.). A different scenario where a subset of individuals enjoys higher marker density has become increasingly common.[2,3] The remaining samples still need imputation and can benefit from better-calibrated imputation quality metrics. Therefore, we consider scenario 2 where we have additional markers directly assayed in only a subset of individuals. Specifically, we assessed the performance of MagicalRsq assuming no overlap of individuals in the training and testing datasets.

CFGP European samples

Under this scenario, we first applied the MagicalRsq models trained from CF 2k samples using all genome-wide variants to the 3k samples (Scenario 2). We observed similar improvements over standard Rsq for all three MAF categories and every chromosome, again with the most pronounced gains on low-frequency variants (Figure S9, Table S8). Specifically, MagicalRsq improves squared Pearson correlation with true R^2 by 2.0%–56.8% for common variants, by 3.0%–6.2% for rare variants, and by 16.3%–91.7% for low-frequency variants, across different chromosomes, compared to standard Rsq (Table S8). MagicalRsq achieves net gains of 18k common variants, 117k

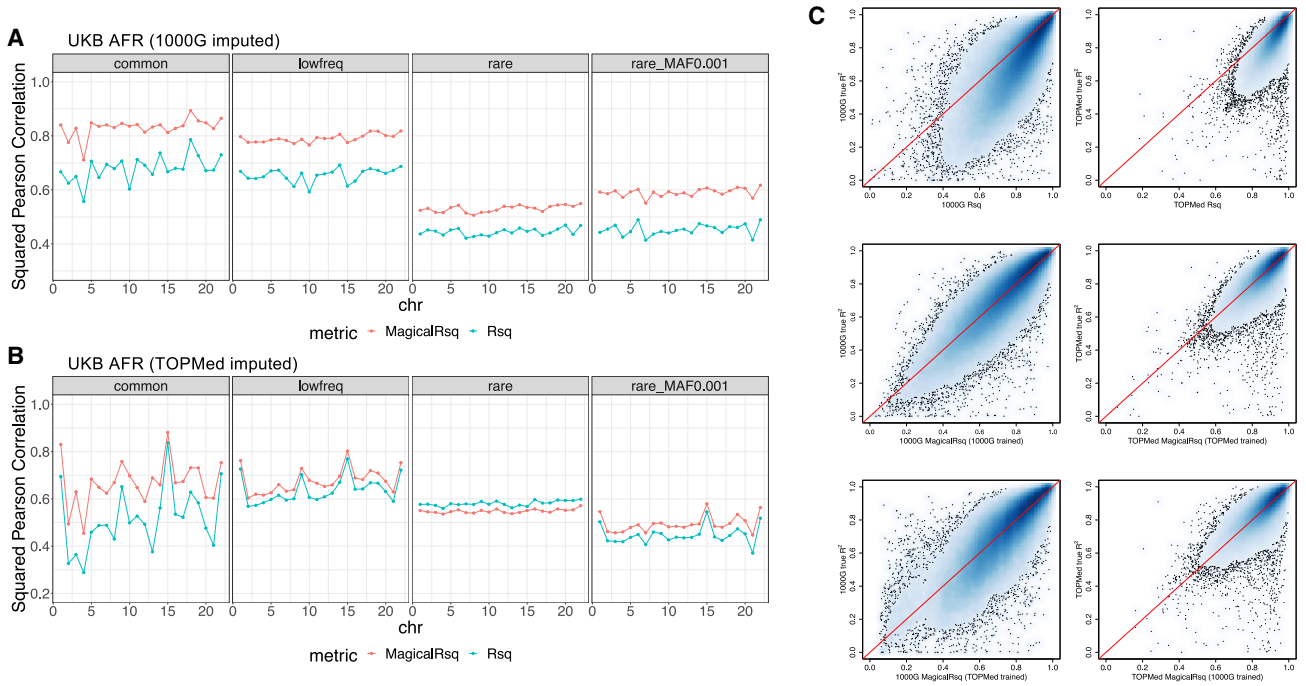


Figure 3. Scenario 2, experiments 9–12: Training models using 1000 UKB AFR samples and testing on 2,960 independent UKB AFR samples, for all variants with WES available

(A and B) Performance comparison between Rsq and MagicalRsq in terms of squared Pearson correlation with true R^2 for (A) 1000G-based imputation; (B) TOPMed-based imputation.

(C) Smooth scatterplot showing Rsq or MagicalRsq (x axis) calculated from both matched (second row) and mis-matched (third row) models against true R^2 (y axis) for both 1000G-based (left) and TOPMed-based (right) imputation, for all low-frequency variants with WES available.

low-frequency variants, and 1.4m rare variants (Table S3, experiment 8). MagicalRsq also outperforms Rsq in terms of RMSE or MAE (Figure S9, Table S8), further supporting that MagicalRsq provides better calibrated imputation quality estimation.

UKB African and South Asian samples

We further assessed MagicalRsq's performance on African-ancestry individuals, leveraging WES data from the UKB. We randomly selected 1,000 individuals as training and the remaining 2,960 as testing (Scenario 2). Consistent with our results in the CF cohort, we found that MagicalRsq is better calibrated than standard Rsq for every variant category for 1000G imputed data (Figure 3A), but interestingly is slightly inferior for TOPMed imputed rare variants (Figure 3B) in terms of squared Pearson correlation with true R^2. One explanation is that TOPMed contains more extremely rare variants which are more challenging to impute. Specifically, 62% (664k/1.1m) of TOPMed imputed variants have MAF $< 0.1\%$ while only 28% (146k/527k) of 1000G imputed variants have MAF $< 0.1\%$. When restricting to variants with MAF between 0.1% and 0.5% (rare variants with minor allele count [MAC] $\geq$ 6), MagicalRsq outperforms standard Rsq (Figure 3B). Note that the inferiority is observed only when evaluating using squared Pearson correlation. When comparing MagicalRsq with Rsq as a post-imputation quality filter, we would have a net gain of 12k rare variants. Moreover, though the absolute improvement of squared Pearson correlation is moderate, MagicalRsq is much better aligned with true R^2 than Rsq. There was a departure of the 45-degree line when comparing true R^2 with Rsq (Figure 3C, top panel), while MagicalRsq correctly rectified such departure (Figure 3C, middle and bottom panels). We again also explored the mix-and-match reference and reached similar conclusions as in scenario 1: MagicalRsq trained using imputed data from a mis-matched reference performs better than Rsq but is inferior to that trained from a matched reference. The performance is substantially better when evaluating only variants shared between the two reference panels, suggesting that some reference-panel-specific variants may hinder the transferability (Note S2, Table S9). We similarly performed experiments on UKB SAS populations to evaluate whether MagicalRsq could benefit other ancestral groups (Scenario 2). We observed that MagicalRsq improved squared Pearson correlation with true R^2 by 7.8%–134.4% for common variants, by 13.8%–22.1% for low-frequency variants, and by 14.4%–22.4% for rare variants (Figure S11, Table S10). The results indicate that MagicalRsq is applicable to these under-represented minority populations.

Number of variants needed for training

We additionally investigated MagicalRsq's performance when using a small subset of variants, with randomly

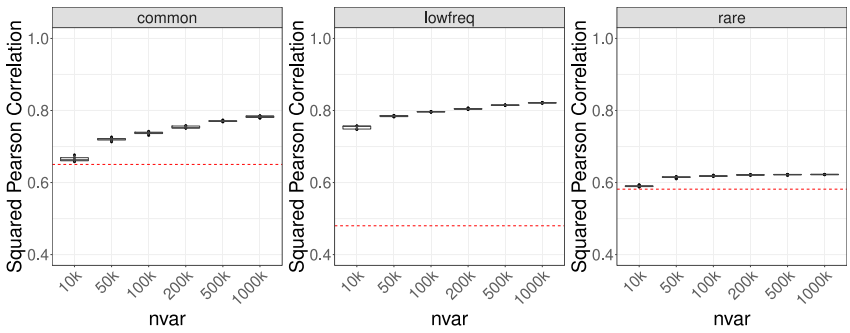


Figure 4. Scenario 2, experiment 14: Training models using randomly selected subsets of variants

The number of variants used for training varied from 10k to 1m. MagicalRsq models were built based on CF 2k samples and tested on the independent CF 3k samples. We repeated 5 times for each number of variants. Squared Pearson correlation with true R^2 was calculated and served as the evaluation metric. The red dashed line denotes the performance of standard Rsq. nvar, number of variants included in model training.

selected 10k, 50k, 100k, 200k, 500k, and 1m variants from CF 2k samples and tested on independent CF 3k samples. We repeated this experiment five times for reliability (Scenario 2). We found that even when MagicalRsq models were trained using only 10k variants, they still outperformed Rsq (the red dashed line in Figure 4) for all three MAF categories. When the number of training variants increases, the advantage of MagicalRsq over Rsq is more pronounced (Figures 4 and S12). For example, the average relative increment of squared Pearson correlation with true R^2 for MagicalRsq is 2.3% with 10k variants for common variants, which increases to 20.5% with 1m variants contributing to model building (Table S11); the net gains of variants increase from ∼6k to ∼19k for common variants, from ∼94k to ∼123k for low-frequency variants, and from ∼787k to ∼1,085k for rare variants (Table S3, experiment 14). We again noticed that the largest improvement manifests for the low-frequency variant category. For example, when trained with 10k (100K) variants, squared Pearson correlation with true R^2 improves by 56.9% (71.0%) for low-frequency variants, substantially more pronounced than the relative increases for common and rare variants, which are 2.3% (13.4%) and 1.9% (9.5%), respectively (Table S11). In addition, we found minimal variation across the five repeats, even when training models with only 10k variants, indicating MagicalRsq is robust to different selected random subsets of variants. Moreover, the computational burden is greatly reduced when using a subset of variants, especially for rare variants. For example, the CPU time was 47 h when training models with all rare variants on even number chromosomes (12.3 million), while it reduced to only 30 min when using randomly selected 50k rare variants (Figure S13, Table S12).

Computation costs

We note that sample size for MagicalRsq models is not the number of individuals but rather the number of variants, which affects the computational costs. We evaluated the CPU time and memory usage, when training from CF 2k samples and applying to CF 3k samples. The CPU time increases exponentially with the number of variants in training, for example, ranging from only 8 min (with 10k variants) to 605 min (with 2.6m variants) for common variants (Figure S13, top panel, Table S12). The memory usage is rather stable with respect to the number of variants in training models, with ∼3 GB for common variants, ∼2 GB for low-frequency variants, and ∼10 GB for rare variants (Figure S13, bottom panel, Table S12).

Discussion

Genotype imputation has become a standard practice in genomic studies. For post-imputation QC and analysis, the estimated imputation quality metrics (referred to as standard Rsq) provided by the various imputation engines (e.g., Rsq from minimac, INFO from IMPUTE, DR2 from Beagle) are key. In this work, we demonstrate that those estimated quality metrics do not always reflect the true imputation quality, especially for lower-frequency and rare variants. To provide better-calibrated quality metrics for these variants, we propose MagicalRsq, a machine learning enhanced genotype imputation quality estimate, which incorporates multiple variant-level features to improve the calibration of imputation quality estimates. We demonstrate by comprehensive experiments that MagicalRsq performs better than standard Rsq under different circumstances: it not only aligns better with true R^2, but can also rescue a substantial number of misclassified variants when replacing standard Rsq as a post-imputation QC metric. We leveraged CFGP WGS data and UKB WES data to show the advantages of MagicalRsq for both European- and African-ancestry individuals. We also performed experiments where MagicalRsq models were built with randomly selected subsets of variants ranging from 10k to 1m variants in each MAF category. Our results showed that MagicalRsq models are robust to different subsets of variants used in model training. We observe slightly better performance with more variants included, but at the cost of exponentially increased computational burden. Considering the tradeoff between computing costs and performance gains, we recommend using 10k to 1m variants in each MAF category when training MagicalRsq models.

We observe that low-frequency variants benefit most from MagicalRsq, which has meaningful implications for downstream analysis. For example, in GWASs, recent publications[4,20,43–46] show multiple examples of GWAS signals from variants in the low-frequency category across diverse

populations. Thus, better discerning and including well-imputed low-frequency variants in GWASs could potentially facilitate new discoveries and aid fine-mapping analysis. Our MagicalRsq models can handle rare variants well and we recommend including extremely rare variants down to singletons in model training, especially when the training set is derived from much fewer individuals than the target set. In our experiments, we obtained net gains of 1 million rare variants using a 0.8 cutoff, and the number decreased with a more lenient threshold (Table S14). We note that the squared Pearson correlation is less informative for rare variants due to their extremely low MAF, and thus we recommend applying a more stringent post-imputation QC threshold, which is also consistent with recommendations from existing literature.[2,3,13,37,38] Some of the rare variants that are rescued by MagicalRsq have important clinical potentials. For example, a *BRCA1* missense variant rs28897686 (chr17:43091783:C:T, hg38, GenBank: NM_007294.4; c.3748G>A [p.Glu1250Lys])[47–50] was well imputed in our UKB AFR target samples with true R^2 0.99, but the standard Rsq was only 0.23, which means this variant would have been missed if Rsq was used to perform post-imputation QC. In contrast, the MagicalRsq value of this variant is 0.86, which effectively rescues the variant for further investigations. We also systematically compared the two quality metrics in the UKB AFR experiments by measuring the power of including potential clinically important variants. We downloaded the ClinVar database and checked for large differences between true R^2 and Rsq or MagicalRsq for each imputed ClinVar exome variant. In summary, 15 well-imputed variants (true $R^2 > 0.8$) that have large quality differences using Rsq (Rsq < 0.5) are saved by MagicalRsq (MagicalRsq > 0.8). Conversely, there is no well-imputed variant (true $R^2 > 0.8$) that has low MagicalRsq (MagicalRsq < 0.5) and can be saved by Rsq (Rsq > 0.8). These findings show clearly that MagicalRsq is superior to Rsq in association studies and further clinical applications.

The XGBoost method adopted in our models is widely used in both classification and regression problems. It is computationally efficient and requires less tunning procedures than some other machine learning or deep learning methods, such as neural networks. We also tried two deep neural network methods to construct the prediction models, but the performance was inferior to XGBoost-trained models (Note S2, Table S13). Our comparison results are consistent with the literature,[51–53] showing the advantage of XGBoost in such regression-like problems. We also explored the strategy of directly using true R^2 in the training subset with WGS available as a post-imputation quality metric in the target set in our scenario 2. We found that such a strategy works well for common variants, but worse for low-frequency variants and extremely badly for rare variants (Table S15), likely due to poor representation for lower-frequency variants using a smaller subset.

Though we have demonstrated that MagicalRsq performs much better than standard Rsq through comprehensive experiments that mimic real-life scenarios, we do note that there are some limitations and caveats to our study. First, we didn't include chromosome X for this study mainly due to the complexity of chromosome X coding and imputation and the lack of some key variant-level features (S/HIC's population genetics summary statistics) that we used for model training. Second, MagicalRsq performance was impaired by some reference-specific variants when applying a model trained with mis-matched reference imputed data, although it nonetheless typically performed better than standard Rsq in this scenario. In practice, we recommend investigators apply MagicalRsq models with matched reference whenever possible for better performance. Third, MagicalRsq performs less well for common variants when training from variants in a particular region, which is likely caused by the fluctuation of Rsq performance across the genome, and such fluctuation was mainly driven by the spanning range of the imputation qualities (Note S2), which impeded the generalizability of such models for common variants.

We note that in our MagicalRsq models, we leverage 79 variant-level features to enhance the imputation quality prediction, but more features could be added into the model. We have released our easily generalizable codes, which allow investigators to incorporate their favorite features into the model and choose whether to include the features we used (i.e., population genetics features summarized by S/HIC or TOP-LD allele-frequency features). We anticipate MagicalRsq to have even better performance when more comprehensive and relevant variant-level features are included. Another future research direction is more general applications and evaluations of MagicalRsq. For the current experiments we performed, rather homogeneous training and target cohorts are used. Even under scenario 2 when training and testing samples are independent with no samples overlap, they still come from the same study (albeit a consortium study involving multiple cohorts in the case of CFGP). Further efforts are also warranted to evaluate the transferability of MagicalRsq models trained on external cohorts, for example, whether MagicalRsq models trained from CFGP data could be applied to UKB cohort.

In summary, MagicalRsq clearly showcases its advantages over standard Rsq in realistic settings. We anticipate that it will benefit the genetic community as a better post-imputation quality metric and will enhance downstream association analysis by rescuing variants.

Data and code availability

MagicalRsq is freely available from https://github.com/quansun98/MagicalRsq/.

The CFGP WGS data are available for request to the Cystic Fibrosis Foundation at https://www.cff.org/researchers/whole-genome-sequencing-project-data-requests#requesting-data.

UKB genotyping and WES data are available upon request from the UK Biobank https://www.ukbiobank.ac.uk/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.09.009.

Acknowledgment

This study is supported by the Cystic Fibrosis Foundation (CUTTIN18XX1, BAMSHA18XX0, KNOWLE18XX0). Y.L. was partially supported by NIH grants U01HG011720, R01HL146500, and R01MH123724. Q.S. was supported by U24AR076730. L.M.R. is additionally supported by KL2TR002490. D.R.S. was supported by U01HG011720 and R35GM138286. C.F. was supported by R01HG009976.

The authors would like to thank the Cystic Fibrosis Foundation for the use of CF Foundation Patient Registry data to conduct this study. Additionally, we would like to thank the patients, care providers, and clinic coordinators at CF centers throughout the United States for their contributions to the CF Foundation Patient Registry.

Furthermore, we acknowledge use of the Trans-Omics in Precision Medicine (TOPMed) program imputation panel (freeze 8 version) supported by the National Heart, Lung, and Blood Institute (NHLBI); see www.nhlbiwgs.org. TOPMed study investigators contributed data to the reference panel, which was accessed through https://imputation.biodatacatalyst.nhlbi.nih.gov. The panel was constructed and implemented by the TOPMed Informatics Research Center at the University of Michigan (3R01HL-117626-02S1; contract HHSN268201800002I). The TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN2682018000001I) provided additional data management, sample identity checks, and overall program coordination and support. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Finally, the authors thank the Department of Innovation, Research University and Museums of the Autonomous Province of Bozen/Bolzano for covering the Open Access publication costs.

Declaration of interests

The authors declare no competing interests.

Received: June 19, 2022
Accepted: September 16, 2022
Published: October 4, 2022

Web resources

Clinvar, https://www.ncbi.nlm.nih.gov/clinvar/
MagicalRsq GitHub page, https://github.com/quansun98/MagicalRsq/
Michigan imputation server, https://imputationserver.sph.umich.edu/
TOP-LD, http://topld.genetics.unc.edu
TOPMed imputation server, https://imputation.biodatacatalyst.nhlbi.nih.gov
True R^2 and Rsq calculation scripts, https://yunliweb.its.unc.edu/software.html

References

1. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.
2. Sun, Q., Liu, W., Rosen, J.D., Huang, L., Pace, R.G., Dang, H., Gallins, P.J., Blue, E.E., Ling, H., Corvol, H., et al. (2022). Leveraging TOPMed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients. HGG Adv. *3*, 100090.
3. Kowalski, M.H., Qian, H., Hou, Z., Rosen, J.D., Tapia, A.L., Shan, Y., Jain, D., Argos, M., Arnett, D.K., Avery, C., et al. (2019). Use of >100, 000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet. *15*, e1008500.
4. Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Miller-Fleming, T.W., Haessler, J., Preuss, M.H., Chai, J.-F., Lee, M.P., et al. (2022). Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. J. Hum. Genet. *67*, 87–93.
5. de Bakker, P.I.W., Ferreira, M.A.R., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. *17*, R122–R128.
6. Porcu, E., Sanna, S., Fuchsberger, C., and Fritsche, L.G. (2013). Genotype imputation in genome-wide association studies. Curr. Protoc. Hum. Genet. *Chapter 1*, Unit1.25.
7. Naj, A.C. (2019). Genotype Imputation in Genome-Wide Association Studies. Curr. Protoc. Hum. Genet. *102*, e84.
8. Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda) *1*, 457–470.
9. 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.
10. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348.
11. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499–511.
12. Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu. Rev. Genomics Hum. Genet. *10*, 387–406.
13. Liu, E.Y., Buyske, S., Aragaki, A.K., Peters, U., Boerwinkle, E., Carlson, C., Carty, C., Crawford, D.C., Haessler, J., Hindorff, L.A., et al. (2012). Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4, 000 haplotypes in African Americans from the Women's Health Initiative. Genet. Epidemiol. *36*, 107–117.
14. Pistis, G., Porcu, E., Vrieze, S.I., Sidore, C., Steri, M., Danjou, F., Busonero, F., Mulas, A., Zoledziewska, M., Maschio, A., et al. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur. J. Hum. Genet. *23*, 975–983.
15. Lin, P., Hartz, S.M., Zhang, Z., Saccone, S.F., Wang, J., Tischfield, J.A., Edenberg, H.J., Kramer, J.R., M Goate, A., Bierut,

L.J., et al. (2010). A new statistic to evaluate imputation reliability. PLoS One *5*, e9697.
16. Coleman, J.R.I., Euesden, J., Patel, H., Folarin, A.A., Newhouse, S., and Breen, G. (2016). Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. Brief. Funct. Genomics *15*, 298–304.
17. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (ACM Press), pp. 785–794.
18. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al. (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am. J. Hum. Genet. *91*, 794–808.
19. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. *45*, D840–D845.
20. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454, 787 UK Biobank participants. Nature *599*, 628–634.
21. Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS One *12*, e0169748.
22. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature *555*, 210–215.
23. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.
24. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discov. Today *23*, 1241–1250.
25. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
26. Das, S., Abecasis, G.R., and Browning, B.L. (2018). Genotype Imputation from Large Reference Panels. Annu. Rev. Genomics Hum. Genet. *19*, 73–96.
27. Schrider, D.R., and Kern, A.D. (2016). S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. *12*, e1005928.
28. Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA. *76*, 5269–5273.
29. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.
30. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.
31. Li, H. (2011). A new test for detecting recent positive selection that is free from the confounding impacts of demography. Mol. Biol. Evol. *28*, 365–375.
32. Garud, N.R., Messer, P.W., Buzbas, E.O., and Petrov, D.A. (2015). Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. *11*, e1005004.
33. Kim, Y., and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. Genetics *167*, 1513–1524.
34. Kelly, J.K. (1997). A test of neutrality based on interlocus associations. Genetics *146*, 1197–1206.
35. Schurz, H., Müller, S.J., van Helden, P.D., Tromp, G., Hoal, E.G., Kinnear, C.J., and Möller, M. (2019). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. Front. Genet. *10*, 34.
36. Huang, L., Rosen, J.D., Sun, Q., Chen, J., Wheeler, M.M., Zhou, Y., Min, Y.-I., Kooperberg, C., Conomos, M.P., Stilp, A.M., et al. (2022). TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. Am. J. Hum. Genet. *109*, 1175–1181.
37. Liu, W., Sun, Q., Huang, L., Bhattacharya, A., Wang, G.W., Tan, X., Kuban, K.C.K., Joseph, R.M., O'Shea, T.M., Fry, R.C., et al. (2022). Innovative computational approaches shed light on genetic mechanisms underlying cognitive impairment among children born extremely preterm. J. Neurodev. Disord. *14*, 16.
38. Duan, Q., Liu, E.Y., Croteau-Chonka, D.C., Mohlke, K.L., and Li, Y. (2013). A comprehensive SNP and indel imputability database. Bioinformatics *29*, 528–531.
39. Knapp, E.A., Fink, A.K., Goss, C.H., Sewall, A., Ostrenga, J., Dowd, C., Elbert, A., Petren, K.M., and Marshall, B.C. (2016). The cystic fibrosis foundation patient registry. design and methods of a national observational disease registry. Ann. Am. Thorac. Soc. *13*, 1173–1179.
40. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
41. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49, 960 individuals in the UK Biobank. Nature *586*, 749–756.
42. 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.
43. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.-H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The polygenic and monogenic basis of blood traits and diseases. Cell *182*, 1214–1231.e11.
44. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746, 667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14.
45. Mahajan, A., Spracklen, C.N., Zhang, W., Ng, M.C.Y., Petty, L.E., Kitajima, H., Yu, G.Z., Rüeger, S., Speidel, L., Kim, Y.J., et al. (2022). Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nat. Genet. *54*, 560–572.
46. Yang, Y., Sun, Q., Huang, L., Broome, J.G., Correa, A., Reiner, A., NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Raffield, L.M., Yang, Y., and Li, Y. (2022). eSCAN: scan

regulatory regions for aggregate association testing using whole-genome sequencing data. Brief. Bioinformatics *23*, bbab497.

47. Judkins, T., Hendrickson, B.C., Deffenbaugh, A.M., Eliason, K., Leclair, B., Norton, M.J., Ward, B.E., Pruss, D., and Scholl, T. (2005). Application of embryonic lethal or other obvious phenotypes to characterize the clinical significance of genetic variants found in trans with known deleterious mutations. Cancer Res. *65*, 10096–10103.
48. Pavlicek, A., Noskov, V.N., Kouprina, N., Barrett, J.C., Jurka, J., and Larionov, V. (2004). Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. Hum. Mol. Genet. *13*, 2737–2751.
49. Lindor, N.M., Guidugli, L., Wang, X., Vallée, M.P., Monteiro, A.N.A., Tavtigian, S., Goldgar, D.E., and Couch, F.J. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). Hum. Mutat. *33*, 8–21.
50. Tavtigian, S.V., Deffenbaugh, A.M., Yin, L., Judkins, T., Scholl, T., Samollow, P.B., de Silva, D., Zharkikh, A., and Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J. Med. Genet. *43*, 295–305.
51. Memon, N., Patel, S.B., and Patel, D.P. (2019). Comparative analysis of artificial neural network and xgboost algorithm for polsar image classification. In Pattern Recognition and Machine Intelligence: 8th International Conference, PReMI 2019, Tezpur, India, December 17-20, 2019, Proceedings, Part I, B. Deka, P. Maji, S. Mitra, D.K. Bhattacharyya, P.K. Bora, and S.K. Pal, eds. (Springer International Publishing), pp. 452–460.
52. Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., and Voyiatzis, I. (2021). Xgboost and deep neural network comparison: the case of teams' performance. In Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings, A.I. Cristea and C. Troussas, eds. (Springer International Publishing), pp. 343–349.
53. Chakraborty, D., and Elzarka, H. (2018). Advanced machine learning techniques for building performance simulation: a comparative analysis. J. Building Performance Simulation *12*, 193–207.

The American Journal of Human Genetics, Volume *109*

Supplemental information

MagicalRsq: Machine-learning-based genotype imputation quality calibration

Quan Sun, Yingxi Yang, Jonathan D. Rosen, Min-Zhi Jiang, Jiawen Chen, Weifang Liu, Jia Wen, Laura M. Raffield, Rhonda G. Pace, Yi-Hui Zhou, Fred A. Wright, Scott M. Blackman, Michael J. Bamshad, Ronald L. Gibson, Garry R. Cutting, Michael R. Knowles, Daniel R. Schrider, Christian Fuchsberger, and Yun Li

Supplemental Notes

Supplemental Note 1. A summary of statistics in S/HIC

S/HIC [1] is a method for detecting and classifying selective sweeps on the basis of 11 summary statistics which reflect spatial patterns of genetic polymorphism; we have incorporated these summary statistics as features in MagicalRsq's input. According to the different aspects of genetic variation they summarize, these features can be divided into 3 subgroups: those summarizing information in the **SFS** (site frequency spectrum), **haplotype structure** and **LD** (linkage disequilibrium).

SFS					Haplotype structure				LD	
pi	thetaH	tajD	fayWuH	maxFDA	HapCount	H1	H12	H2.H1	Omega	ZnS

1.SFS (Site Frequency Spectrum)

The site frequency spectrum (SFS) of a sample of DNA sequences is the histogram of allele frequencies of polymorphisms found in that sample. More formally, the SFS it is the vector $[\eta_1,\ \eta_2, \ldots, \eta_k]$, where η_i is the number of polymorphisms whose derived allele frequency is i.

The first group of statistics used by S/HIC includes $\hat{\theta}_\pi$ (often referred to as π) [2], $\hat{\theta}_H$ [3], Tajima's *D* [4] and Fay and Wu's *H* [3]. The first two of these are estimators of the population-scaled mutation rate $\theta = 4N\mu$ (where N is the population size and μ is the mutation rate). The second two statistics are obtained by taking the difference between two estimators of θ. The four statistics are defined as follows:

(1) $\hat{\theta}_\pi = \frac{2}{n(n-1)}\sum_{i=1}^{n-1} i(n-i)\eta_i$ (See Nei and Tajima [5] . This formulation is obtained from Achaz [6] .) Referred to as "pi" by S/HIC.
(2) $\hat{\theta}_H = \frac{2}{n(n-1)}\sum_{i=1}^{n-1} i^2\eta_i$ [3]. Referred to as "thetaH" by S/HIC.
(3) Tajima's $D = \frac{\hat{\theta}_\pi - \hat{\theta}_w}{\sqrt{var(\hat{\theta}_\pi - \hat{\theta}_w)}}$ [4], where $\hat{\theta}_w = a\sum_{i=1}^{n-1}\eta_i$ and $a = \sum_{i=1}^{n-1}\frac{1}{i}$ [7]. Referred to as "tajD" by S/HIC.

(4) Fay and Wu's H $= \hat{\theta}_\pi - \hat{\theta}_H$, where $\hat{\theta}_H$ is defined in (2) above [3]. Note that S/HIC reverses these operands, such that positive values reflect an excess of high-frequency derived alleles, and refers to this value as "fayWuH".

Biological Interpretation: Tajima's *D* is a commonly used statistic for detecting departures from the standard neutral model, e.g. a beneficial mutation sweeping to fixation in the population and/or changes in population size will cause *D* to differ from the neutral expectation of 0. Negative *D* indicates a deficit of intermediate-frequency alleles (consistent with population expansion and/or positive selection); positive *D* indicates an excess of intermediate frequency alleles (consistent with population contraction and/or balancing selection). Fay and Wu's *H* is similar in principle, but tests for an excess or deficit of high-frequency derived alleles.

2. Haplotype structure

The second group of statistics, namely H1, H12, H2/H1 [8] and *k* is used to describe haplotype structure. H1 is haplotype homozygosity: the probability that any two randomly chosen haplotypes from the sample are identical. More formally:

$$H1 = \sum_{i=1}^{n} {p_i}^2,$$

where p_i is the frequency of the i^{th} most frequent haplotype observed in the sample. H12 is identical to the value of H1 obtained when treating the two most frequent haplotypes as if they are identical:

$$H12 = (p_1 + p_2)^2 + \sum_{3}^{n} {p_i}^2 = H1 + 2p_1p_2$$

H12 was designed to be sensitive to soft selective sweeps, wherein multiple haplotypes containing a beneficial allele participate in a sweep. H2 is identical to the value of H1 that one would obtain by omitting the term for the most common haplotype:

$$H2 = \sum_{2}^{n} {p_i}^2$$

The ratio $H2/H1$ is expected to be higher for soft sweeps than hard sweeps, because H2 may be elevated by alternative haplotypes bearing the adaptive allele and which may participate in the sweep. Finally, *k*, referred to as HapCount by S/HIC, is the number of distinct haplotypes observed in the population.

3. Linkage Disequilibrium

(1) ZnS [9]

Z_{nS} is the average value of r^2 across all pairs of SNPs within a genomic region. I.e., if ${r_{ij}}^2$ is the value of r^2 between the i^{th} and j^{th} of S SNPs in the window, then:

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S=1} \sum_{j=i+1}^{S} r_{ij}^2$$

(2) Omega [10]

Kim and Nielsen's ω is designed to detect the characteristic pattern of LD around a hard selective sweep: because recombination events occur independently on either flank of a selected allele during its sojourn toward fixation, blocks of LD will appear on either side of the selected site, but these blocks will be independent of one another and thus there will be little LD stretching *across* the selected site. If we again have S SNPs in the genomic region being examined, and choose our l^{th} SNP as the focal SNP, the formula for ω is as follows:

$$\omega = \frac{\sum_{i,j \in L} {r_{ij}}^2 + \sum_{i,j \in R} {r_{ij}}^2}{\left(\binom{l}{2} + \binom{S-l}{2}\right)(1/(S-l))\sum_{i \in L, j \in R} {r_{ij}}^2}$$

where *L* is the set of all SNPs to the left of the focal SNP, and *R* is the set of all remaining SNPs (i.e. the l^{th} SNP and all SNPs to its right), and again ${r_{ij}}^2$ is the value of r^2 between the i^{th} and j^{th} SNPs in the window.

Because the location of a selective sweep in the window, if there is one, is not known, the value of ω is calculated for each focal l in the window (in S/HIC's calculation, l ranges from 3 to $S-2$), and the maximum of all of these values of ω is taken. S/HIC refers to the resulting value as "Omega".

Supplemental Note 2. Additional supporting results

Mix-and-match of reference panel under Scenario 1

In the mix-and-match section, we examined whether MagicalRsq trained using imputed data from one reference panel can be applied to imputed data from a different reference panel. Specifically, we trained MagicalRsq models using 1000G-imputed variants on odd number autosomes and applied to TOPMed-imputed variants on even number autosomes (**experiment 3**); and vice versa from TOPMed training to 1000G testing (**experiment 4**), in the same 2k CF samples. When evaluating restricted to the shared variants between TOPMed and 1000G reference panels, our mix-and-match reference panels experiments show promising results (**Figure 2C, Table S6**): MagicalRsq models trained from 1000G-imputed data still outperform Rsq when applied to TOPMed-imputed data, and vice versa. For instance, applying the MagicalRsq model trained from low frequency variants in 1000G-imputed data to TOPMed-imputed low frequency variants, we observe that MagicalRsq leads to 24.5% increase in squared Pearson correlation with true R^2, 35.4% decrease in RMSE, and 18.2% decrease in MAE, compared to standard Rsq. The improvements are slightly less pronounced than using the matched reference panel (i.e., applying models trained in TOPMed to TOPMed) (**Figure 2C**). For example, when applying matched (i.e. TOPMed-) trained MagicalRsq model to TOPMed low frequency variants, we observe 60.5% increase in squared Pearson correlation with true R^2, 57.2% decrease in RMSE, and 40.7% decrease in MAE, compared to standard Rsq. Although MagicalRsq models trained from a mismatched reference perform less well than those trained from a matched reference, they still demonstrate a clear advantage over standard Rsq.

We also note that the absolute performance of Rsq seems to be better using 1000G reference panel than TOPMed in terms of squared Pearson correlation with true R^2 (**Figure 2A, B** and **Table S6**, 0.73 v.s. 0.58 for common variants, 0.63 v.s. 0.49 for low frequency variants and 0.59 v.s. 0.54 for rare variants), though we know that 1000G contains far fewer variants (~80M v.s. ~3000M) and fewer individuals (~2.5K v.s. ~100K) than TOPMed. It doesn't imply that 1000G imputation is superior to TOPMed: the statistics shown in **Figure 2A, B** and **Table S6** are not the true imputation quality, but the performance of the estimated imputation quality from the reference panel. Therefore, it only means that the imputation quality estimates from a larger reference panel (TOPMed) is worse than that from a smaller one (1000G). To better explain this phenomenon, we plotted Rsq and MagicalRsq against true R^2 respectively, for both 1000G and TOPMed imputed data (**Figure 2C**), and also plotted 1000G against TOPMed, for true R^2 and Rsq separately (**Figure S4**), for low frequency variants on chromosome 13. We observe that TOPMed true R^2 can still be larger than 1000G true R^2, and TOPMed Rsq are also larger than 1000G Rsq. One potential explanation is that, the larger the reference panel in terms of individuals, the more complicated haplotypes we will likely observe. This may cause the imputation engine to be less confident about the imputed results, which may lead to under-estimate of the true imputation quality. **Figure 2C** showed clearly that TOPMed Rsq tends to more severely under-estimate the true imputation quality than 1000G, making MagicalRsq more desirable with larger reference panels.

Investigation of model trained in small regions for common variants under Scenario 1

We found that MagicalRsq models trained with variants in a small 20MB region perform uniformly reasonably well for low frequency and rare variants, but not for common variants (**Figure S6**), and we hypothesized that the large fluctuation of Rsq performance for common variants may contribute to this phenomenon. For example, on chromosome 15, the squared Pearson correlation between Rsq and true R^2 could reach 0.8, while on chromosome 5, it is only ~0.4 (**Figure S6**). Further investigation showed that such fluctuation was largely driven by the spanning range of the imputation qualities for variants on different chromosomes (**Figure S5**). For instance, for the vast majority of variants on chromosome 5, Rsq and true R^2 are over 0.6; in contrast, variants on chromosome 15 have Rsq and true R^2 spanning the entire 0 to 1 range. These patterns may hinder the generalizability of MagicalRsq models trained with common variants from random small regions to the genome.

Mix-and-match of reference panel under Scenario 2

Same as in scenario 1, we also want to investigate whether MagicalRsq models are similarly amenable to mix-and-match under scenario 2, thus we built MagicalRsq models using 1000G-imputed data in training (i.e., the UKB AFR 1,000 individuals) and applied to TOPMed-imputed data in testing (i.e., the remaining 2,960 UKB AFR individuals) (**experiment 11**), and vice versa (**experiment 12**). The evaluation and comparison are restricted to shared variants between TOPMed-imputed and 1000G-imputed data.

To investigate whether MagicalRsq models are similarly robust to different reference panels under scenario 2, we built mix-and-match MagicalRsq models leveraging UKB AFR data (**Methods**). We found, similar to observations under scenario 1, that MagicalRsq outperforms Rsq in all cases except for applying 1000G-based models to rare variants imputed using TOPMed (**Table S9**). As discussed previously, a likely explanation is that TOPMed contains more extremely rare variants that are therefore harder to impute. When excluding variants with MAF < 0.1%, as expected, all MagicalRsq models outperform Rsq (**Table S9**), though "matched model" would improve Rsq performance better than mis-matched ones. For example, when testing on TOPMed imputed data, applying TOPMed-imputation-based training model would improve the squared Pearson correlation with true R^2 by 36.3% for common variants, while applying 1000G-imputation-based model would only increase the squared Pearson correlation by 19.7%. We further plotted true R^2 against Rsq or MagicalRsq trained with both matched- and mismatched- models for all the variants by the three MAF categories, to better compare the performances of the quality metrics (**Figure 3C, Figure S10**). We note that **Figure 3C** shows a clear advantage of 1000G-trained MagicalRsq for TOPMed-imputed data (the last sub-figure on the right), compared to Rsq (the first sub-figure on the right). However, the squared Pearson correlations with true R^2 for 1000G-trained MagicalRsq and Rsq have minimal differences: 1000G-trained MagicalRsq is only 1.35% superior to Rsq (**Table S9**). This evidence shows that the squared Pearson correlation with true R^2 has some drawbacks for evaluating the imputation quality.

Other machine learning methods

We also compared the performance of MagicalRsq to other machine learning methods enhanced imputation quality estimation metrics. Specifically, we first adopted a simple Deep Neural Network (DNN) with two hidden layers with the rectified linear activation function (ReLU), using the DeepTables library in Python. For each hidden layer, we assigned 256 units, specified the dropout rate as 0.3, applied batch normalization and set early stop patience to be 30. We also considered an ensemble method, averaging the output from three models, a two-hidden-layer DNN (300 units each layer, dropout rate of 0.3 with batch normalization), a two-layer Deep Cross Network (DCN) [11], and a one-layer feature machine (DeepFM) [12] .

Using the CF 2k cohorts as an example, we found that MagicalRsq outperforms both the two DNN models (**Table S13**) for every MAF category. For example, MagicalRsq was able to improve the squared Pearson correlation by 48.22% for common variants, while the improvements using DNN and DNN II were 35.77% and 38.58%, respectively.

Supplemental Figures

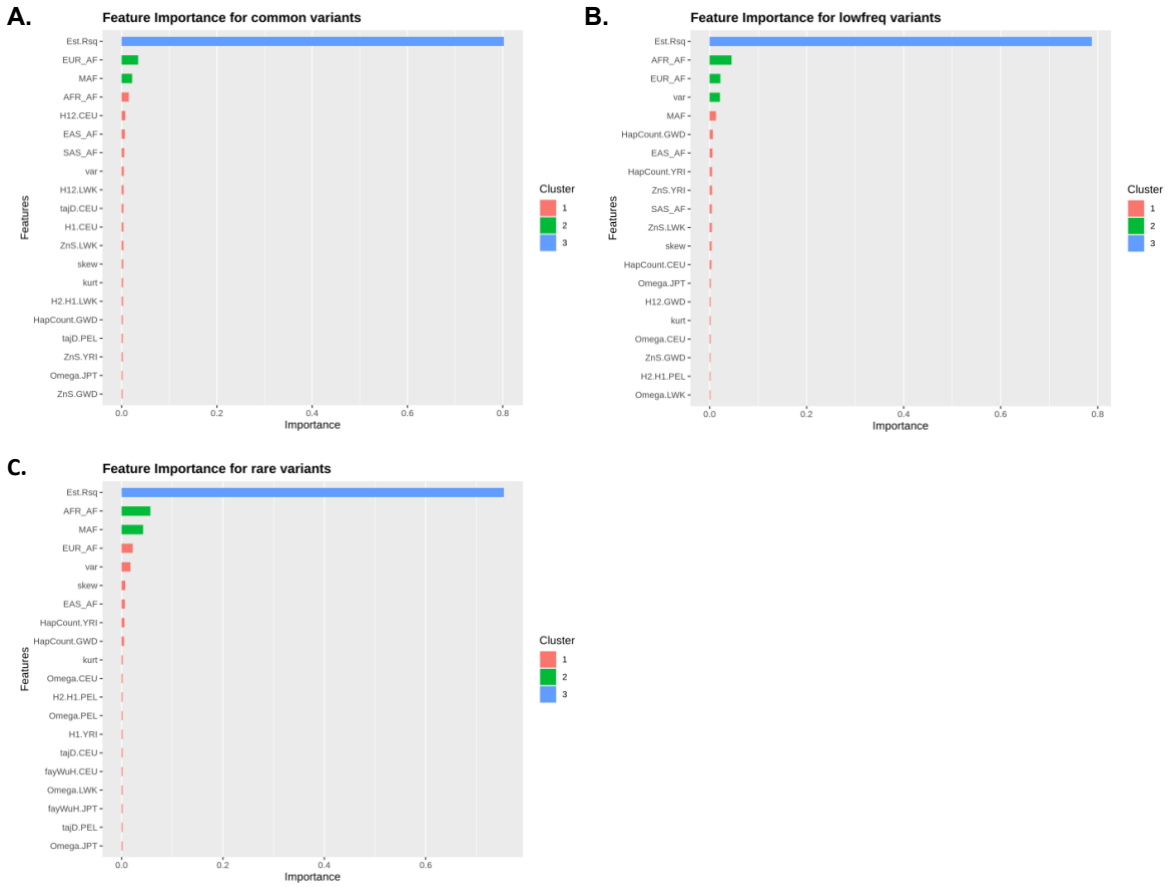


Figure S1. Feature importance for MagicalRsq models in Scenario 1 Experiment 1. The standard Rsq weighs the highest and is about 80% importance for all the three categories. European allele count (AC) is the second most important feature for common variants, but African AC is the second most important feature for low frequency and rare variants.

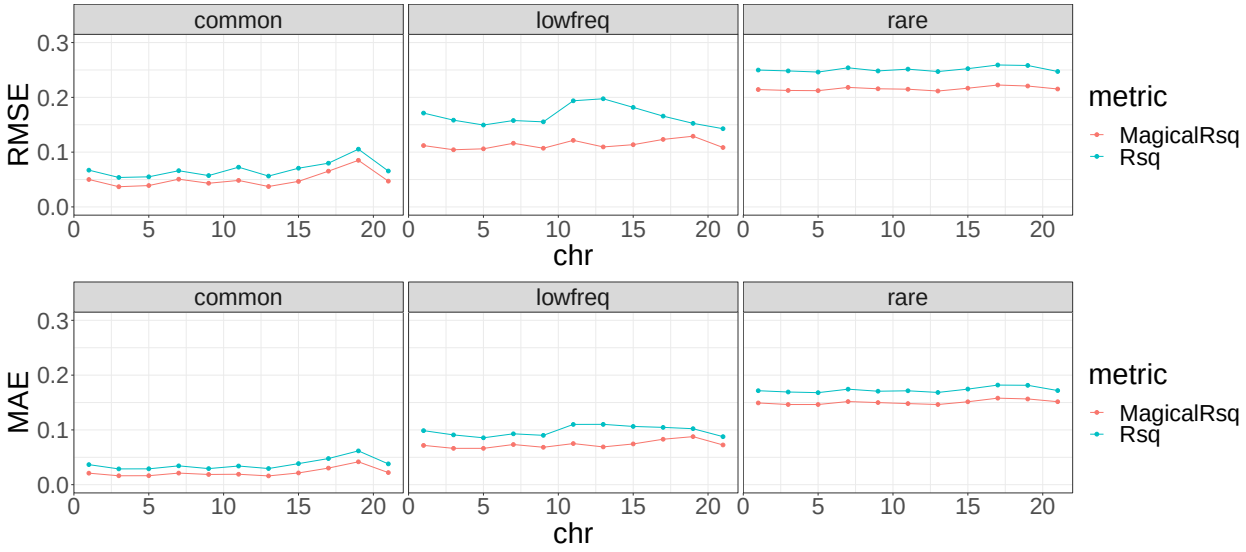


Figure S2. Performance comparison of MagicalRsq and Rsq in terms of RMSE and MAE, for Scenario 1 Experiment 1. Imputation was performed using 1000G reference panel, and MagicalRsq was calculated from model trained on CF 2k even number chromosomes which was also imputed using 1000G reference panel.

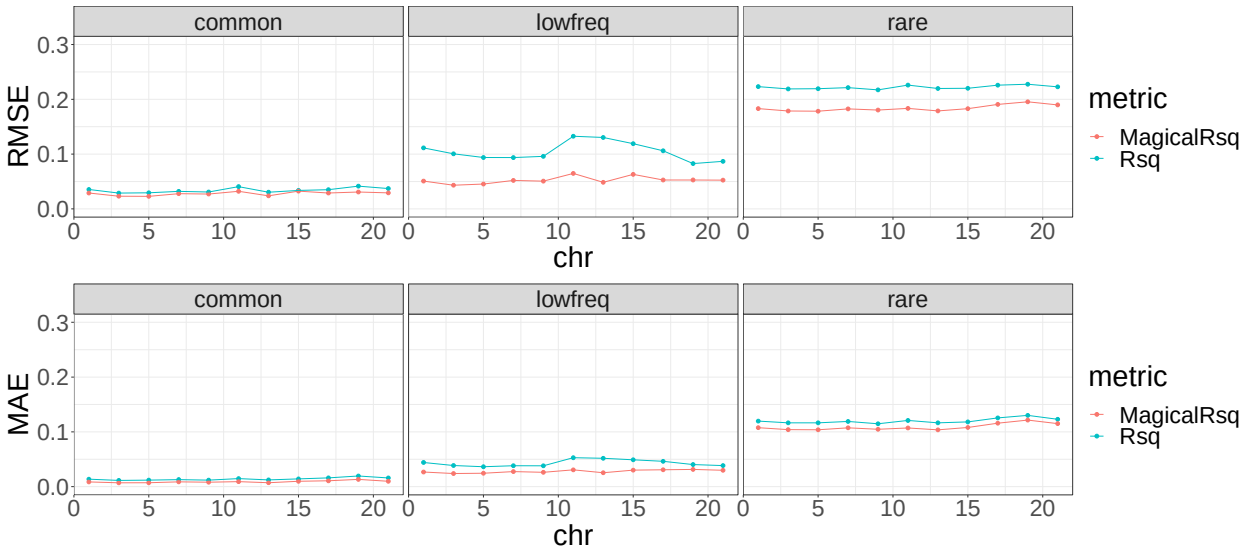


Figure S3. Performance comparison of MagicalRsq and Rsq in terms of RMSE and MAE, for Scenario 1 Experiment 2. Imputation was performed using TOPMed freeze 8 reference panel, and MagicalRsq was calculated from model trained on CF 2k even number chromosomes which was also imputed using TOPMed freeze 8 reference panel.

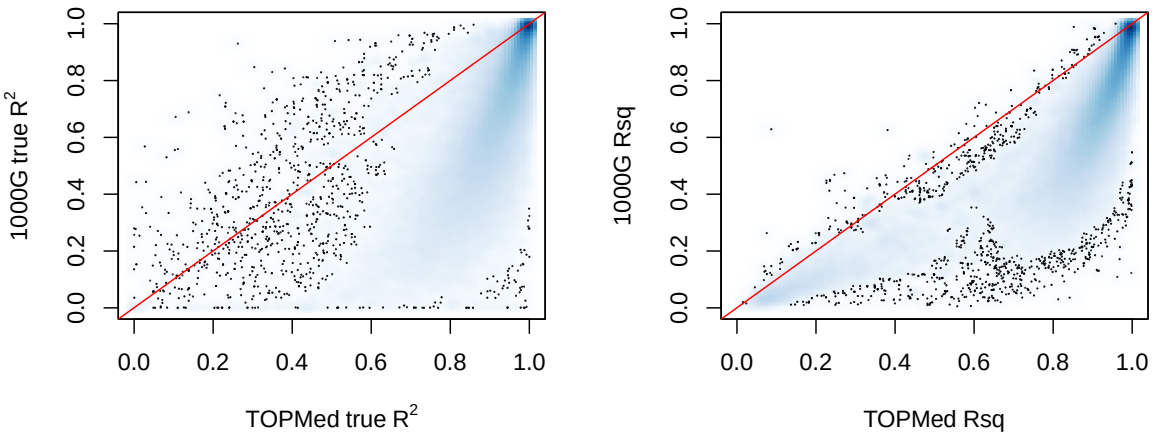


Figure S4. Comparison between 1000G and TOPMed imputation. We plotted 1000G true R^2 against TOPMed true R^2, and 1000G Rsq against TOPMed Rsq from imputation 1 and 2. Tough the squared Pearson correlation between TOPMed Rsq and TOPMed true R^2 is smllaer than 1000G, TOPMed imputation quality (and the estimates Rsq) are still better than 1000G.

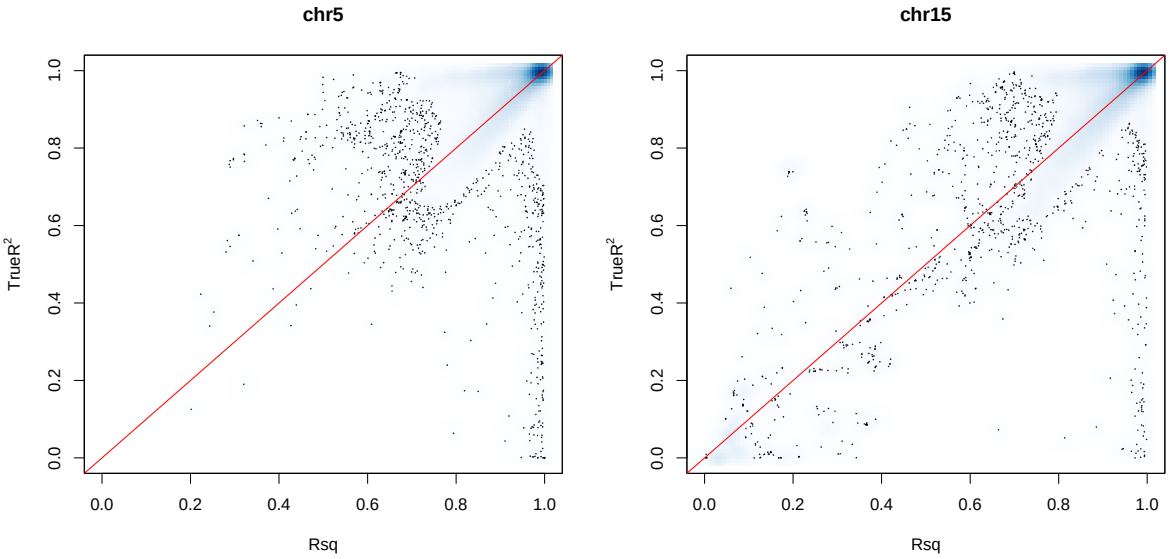


Figure S5. Rsq v.s. True R^2 for common variants on chr5 and chr15 for TOPMed imputed CF 2k cohort. We observed the fluctuation of Rsq performance for different chromosomes across the genome for CF 2k cohort, and this is likely due to the different spanning range of Rsq. Larger range would lead to higher Pearson correlation.

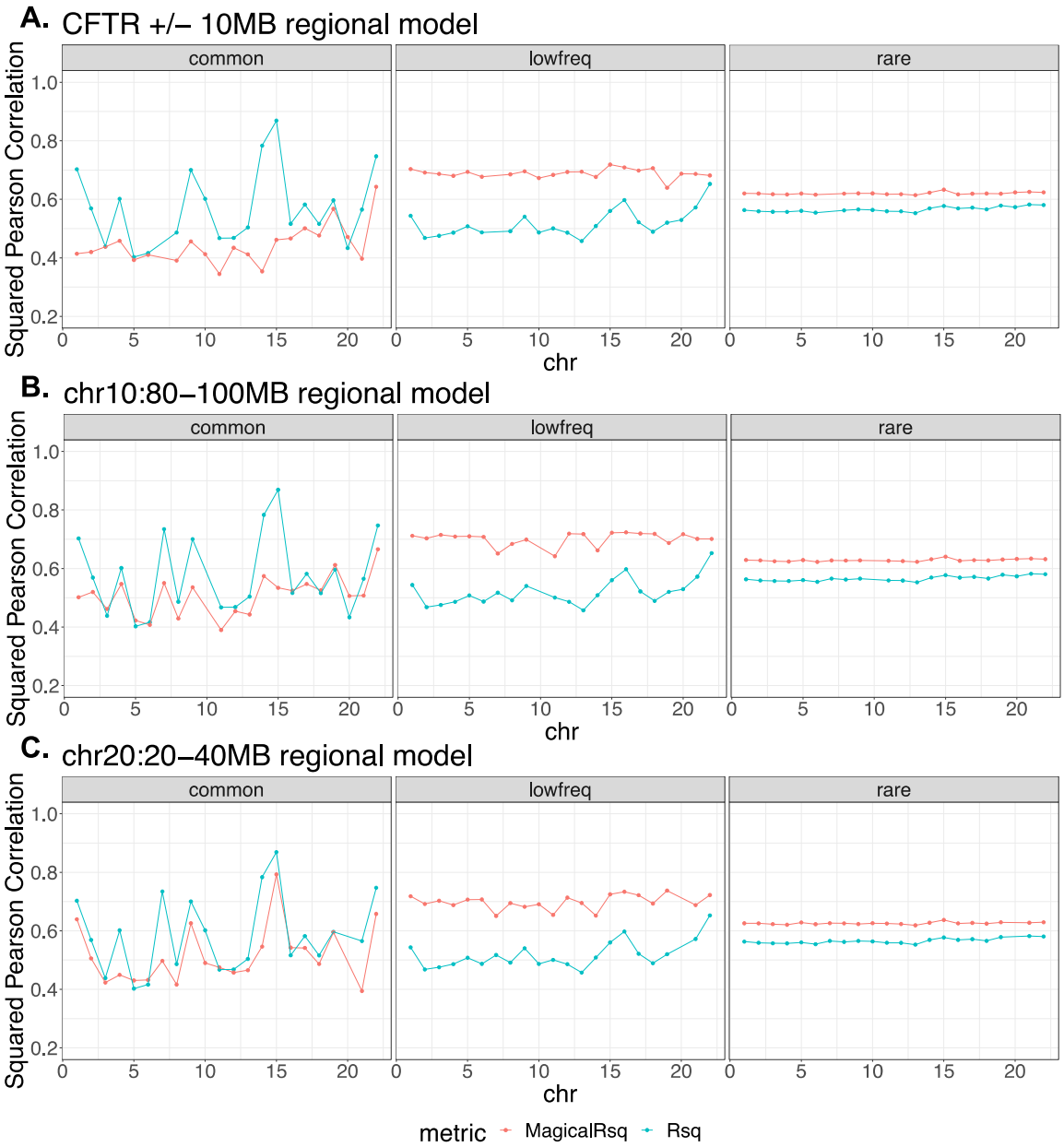


Figure S6. Scenario 1 Experiment 5: training models using variants in a 20MB region and testing on all other chromosomes for CF 2k samples with TOPMed imputation. Performance comparison between Rsq and MagicalRsq in terms of squared Pearson correlation with true R^2 for models trained with variants in **(A)** CFTR +/- 10MB region; **(B)** chr10:80-100MB region; **(C)** chr20:20-40MB region.

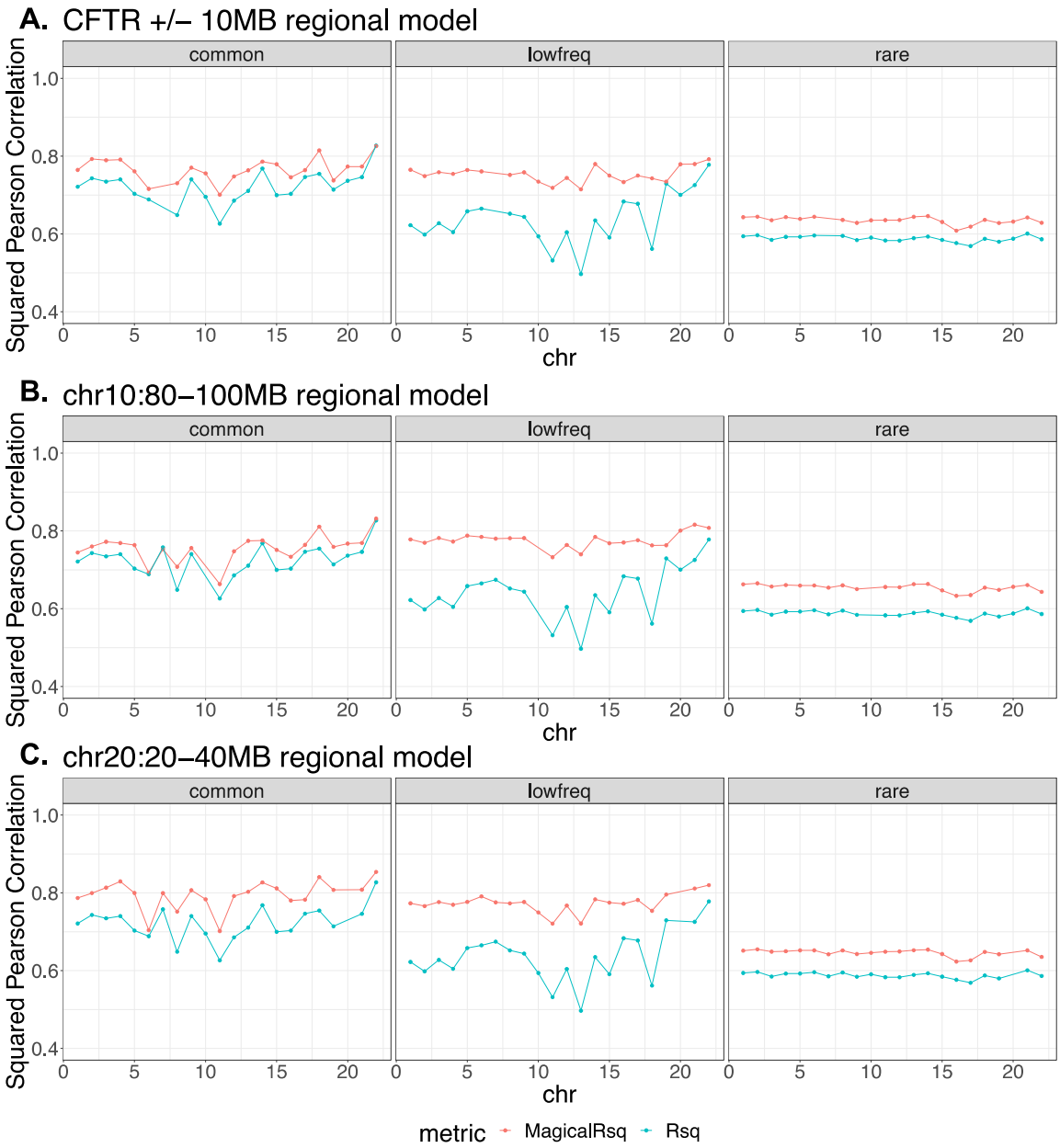


Figure S7. Scenario 1 Experiment 6: training models using variants in a 20MB region and testing on all other chromosomes for CF 2k samples with 1000G imputation. Performance comparison between Rsq and MagicalRsq in terms of squared Pearson correlation with true R^2 for models trained with variants in **(A)** CFTR +/- 10MB region; **(B)** chr10:80-100MB region; **(C)** chr20:20-40MB region.

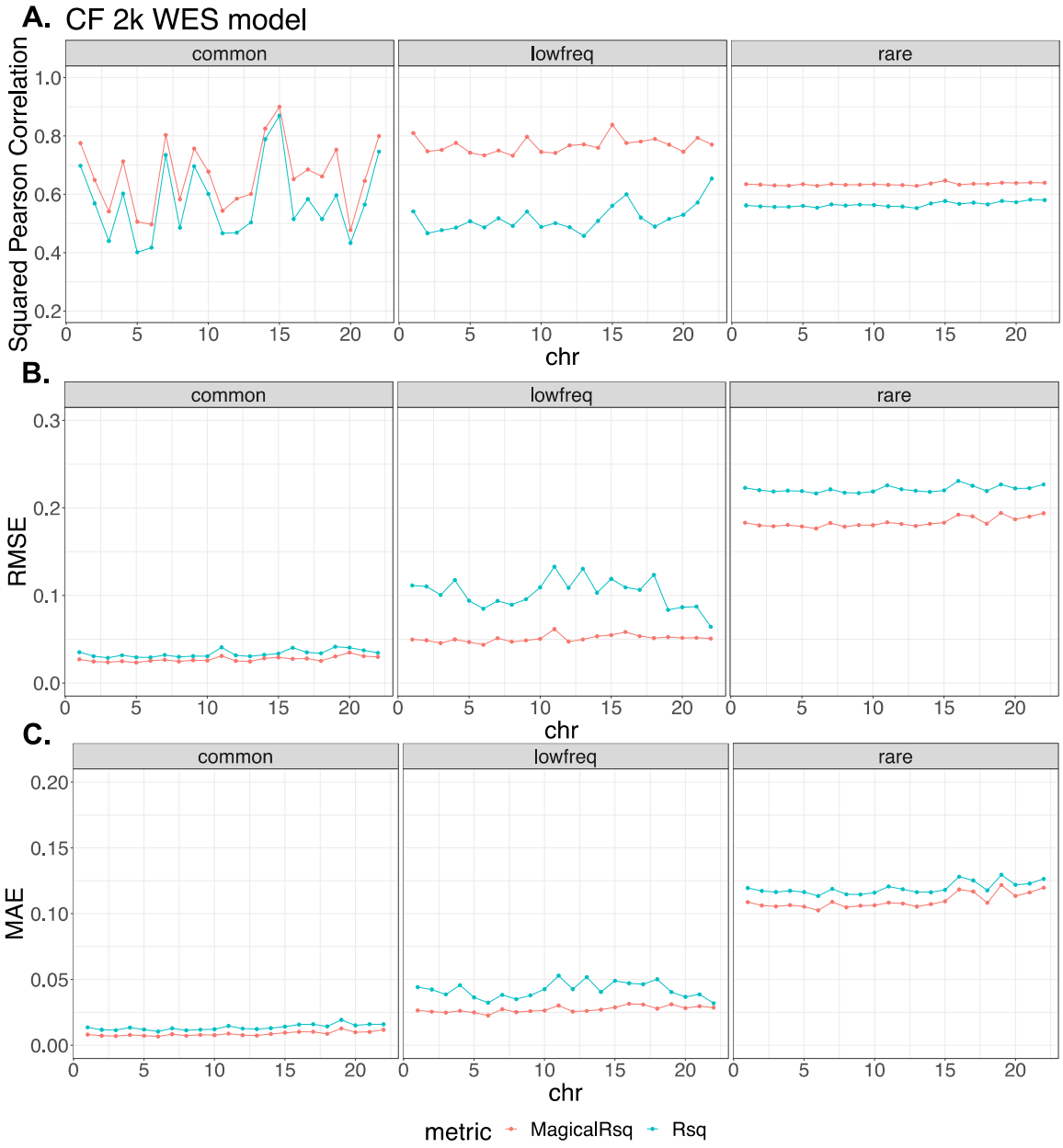


Figure S8. Scenario 1 Experiment 7: training models using TOPMed imputed exonic variants from CF 2k samples and testing on TOPMed imputed variants in other genomic regions of the same CF 2k samples. Performance comparison between Rsq and MagicalRsq in terms of **(A)** squared Pearson correlation with true R^2; **(B)** RMSE; **(C)** MAE.

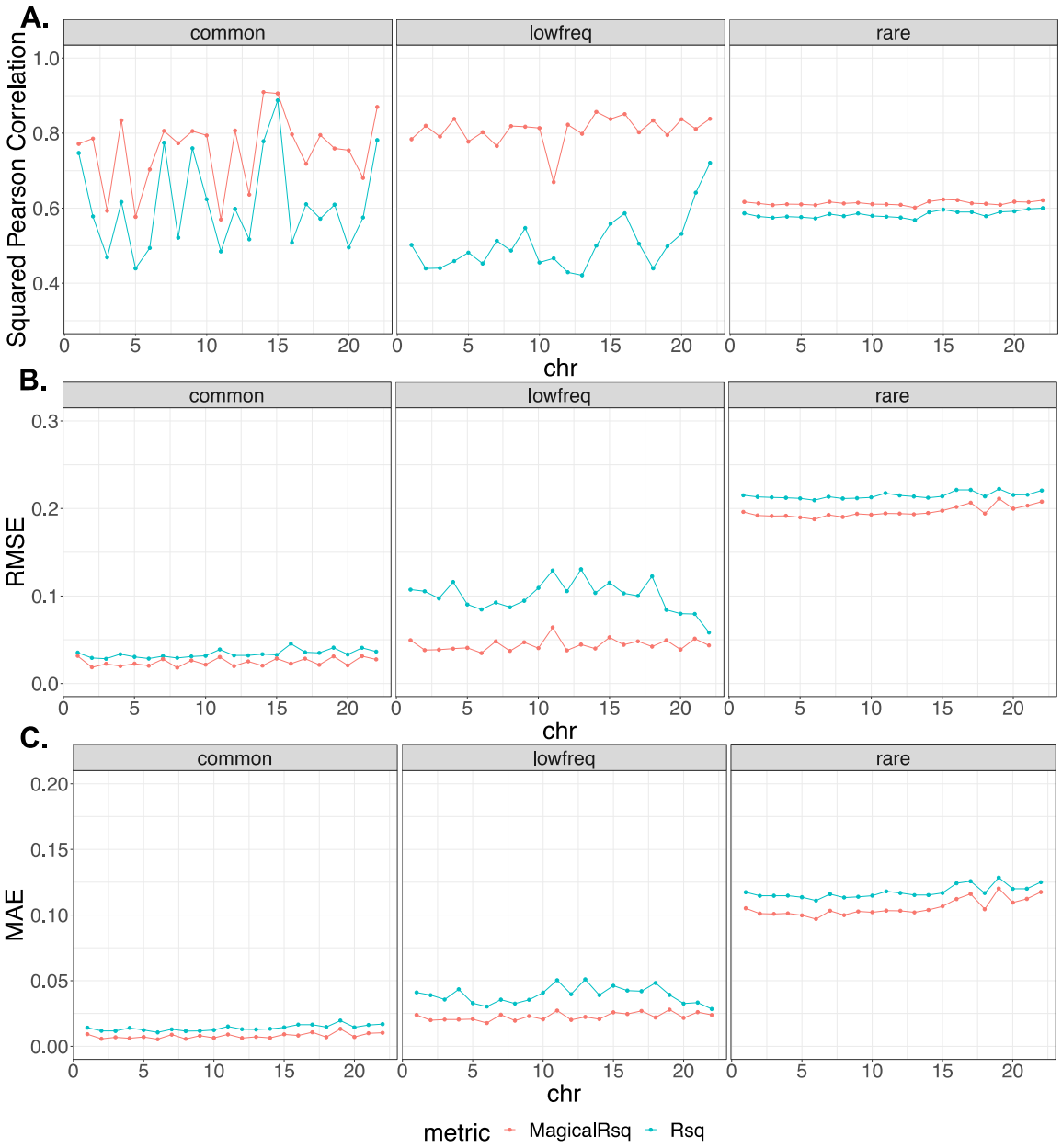


Figure S9. Scenario 2 Experiment 8: training models using TOPMed imputed variants from CF 2k samples and testing on TOPMed imputed all chromosomes of independent CF 3k samples. Performance comparison between Rsq and MagicalRsq in terms of **(A)** squared Pearson correlation with true R^2; **(B)** RMSE; **(C)** MAE.

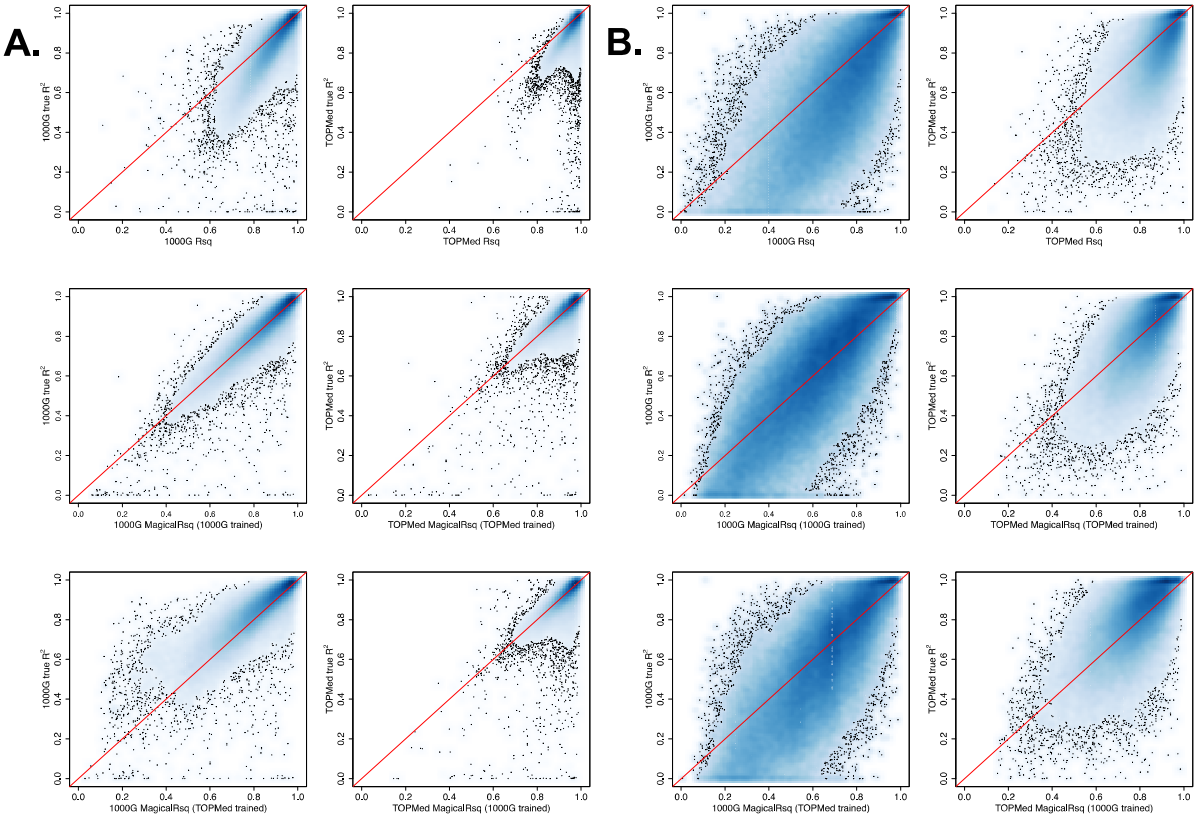


Figure S10. Scenario 2 Experiment 9-12: training models using 1000 UKB AFR samples and testing on 2960 independent UKB AFR samples, for all variants with WES available. Smooth scatter plot showing Rsq or MagicalRsq (X-axis) calculated from both matched- (second row) and mis-matched- (third row) models against true R^2 (Y-axis) for both 1000G- (left) and TOPMed- (right) based imputation, for **(A)** common variants; **(B)** rare variants with MAF > 0.001 (corresponding to MAC >= 6) with WES available.

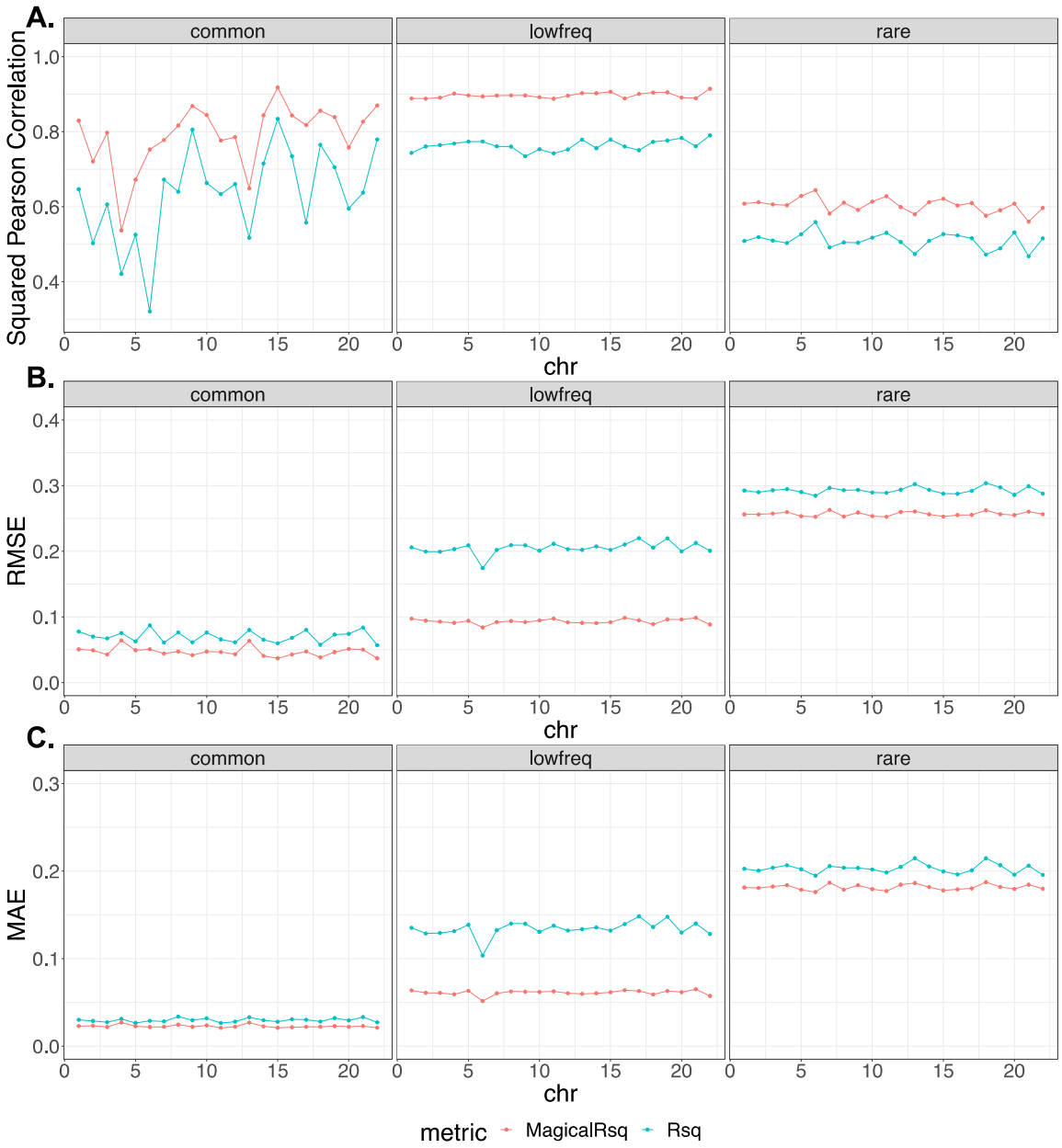


Figure S11. Scenario 2 Experiment 13: training models using TOPMed imputed variants of 1,000 UKB SAS samples and testing on TOPMed imputed variants (across all chromosomes) of an independent set of UKB SAS 3,436 samples. Performance comparison between Rsq and MagicalRsq in terms of **(A)** squared Pearson correlation with true R^2; **(B)** RMSE; **(C)** MAE.

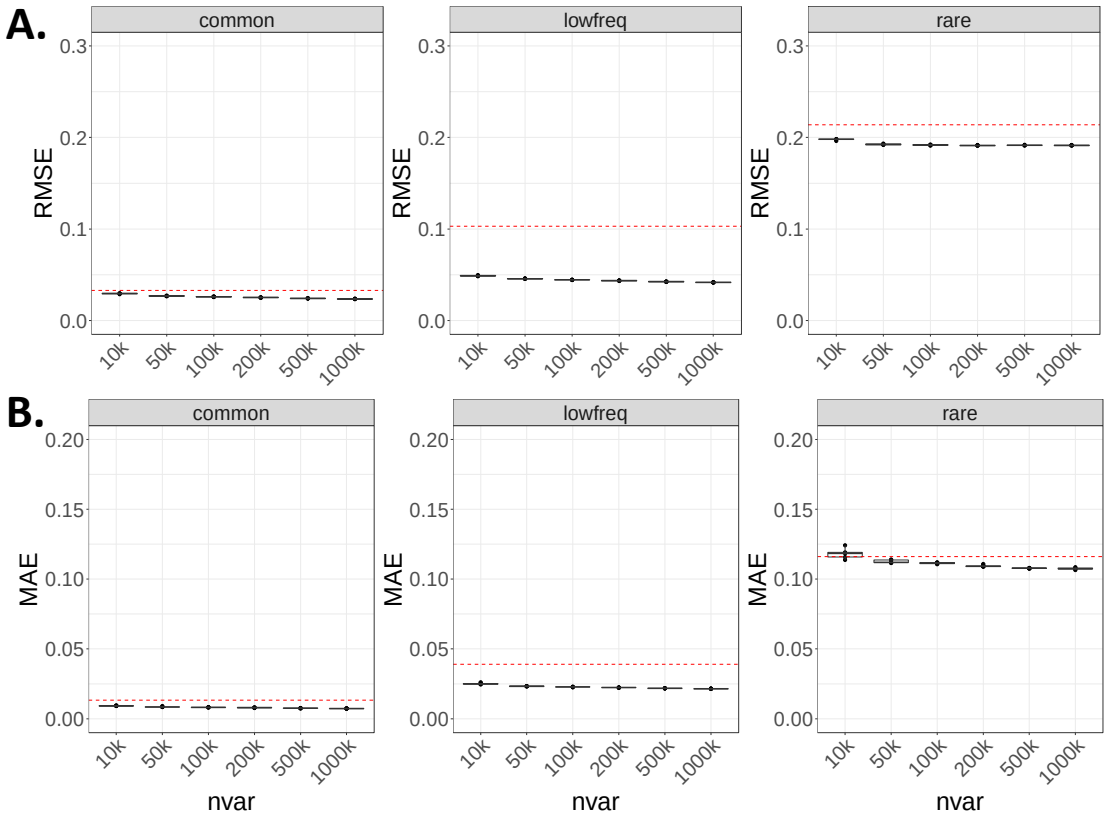


Figure S12. Scenario 2 Experiment 14: training models using randomly selected variants varying from 10k to 1000k from CF 2k samples, and testing on independent CF 3k samples. We repeated 5 times for each number of variants and evaluated MagicalRsq and Rsq performance using **(A)** RMSE and **(B)** MAE. The red dashed line denotes the performance of standard Rsq.

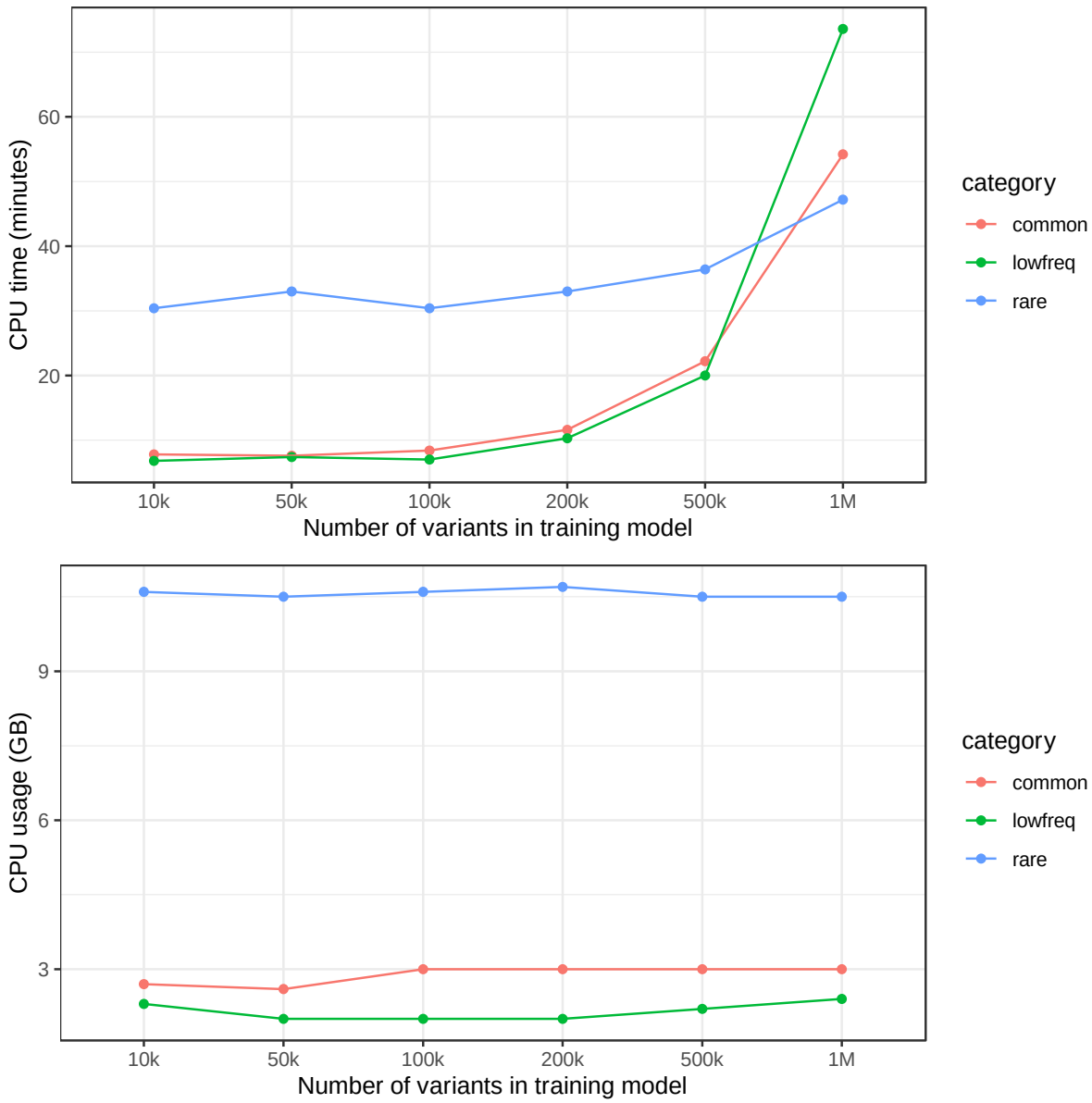


Figure S13. CPU time in minutes and memory usage in GB for MagicalRsq model training with different numbers of variants, separately for three MAF categories.

Supplemental References

1. Schrider, D.R., and Kern, A.D. (2016). S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. *12*, e1005928.

2. Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA *76*, 5269–5273.

3. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

4. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

5. Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. Genetics *97*, 145–163.

6. Achaz, G. (2009). Frequency spectrum neutrality tests: one for all and all for one. Genetics *183*, 249–258.

7. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.

8. Garud, N.R., Messer, P.W., Buzbas, E.O., and Petrov, D.A. (2015). Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. *11*, e1005004.

9. Kelly, J.K. (1997). A test of neutrality based on interlocus associations. Genetics *146*, 1197–1206.

10. Kim, Y., and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. Genetics *167*, 1513–1524.

11. Wang, R., Fu, B., Fu, G., and Wang, M. (2017). Deep & Cross Network for Ad Click Predictions. ArXiv. https://doi.org/10.48550/arXiv.1708.05123

12. Guo, H., Tang, R., Ye, Y., Li, Z., He, X., and Dong, Z. (2018). DeepFM: An End-to-End Wide & Deep Learning Framework for CTR Prediction. ArXiv. https://doi.org/10.48550/arXiv.1804.04950