MagicalRsq-X: A cross-cohort transferable genotype imputation quality metric

Authors

Quan Sun, Yingxi Yang, Jonathan D. Rosen, ..., Yuan-I Min, Christian Fuchsberger, Yun Li

Correspondence

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

Ever-growing reference panels allow imputation of a huge number ($\sim 10^8$) of lower-frequency variants. However, the standard imputation quality metric poorly reflects the true quality of uncommon variants. We introduce MagicalRsq-X, an extension of MagicalRsq that allows model training across cohorts for which only the genotypes used for imputation are available.

Sun et al., 2024, The American Journal of Human Genetics *111*, 990–995
May 2, 2024

https://doi.org/10.1016/j.ajhg.2024.04.001

REPORT

MagicalRsq-X: A cross-cohort transferable genotype imputation quality metric

Quan Sun,[1] Yingxi Yang,[2] Jonathan D. Rosen,[3] Jiawen Chen,[1] Xihao Li,[1] Wyliena Guan,[1] Min-Zhi Jiang,[4] Jia Wen,[3] Rhonda G. Pace,[5] Scott M. Blackman,[6] Michael J. Bamshad,[7,8] Ronald L. Gibson,[7] Garry R. Cutting,[9] Wanda K. O'Neal,[5] Michael R. Knowles,[5] Charles Kooperberg,[10] Alexander P. Reiner,[11] Laura M. Raffield,[3] April P. Carson,[12] Stephen S. Rich,[13] Jerome I. Rotter,[14] Ruth J.F. Loos,[15,16] Eimear Kenny,[15] Byron C. Jaeger,[17] Yuan-I Min,[18] Christian Fuchsberger,[19,20,*] and Yun Li[1,3,20,*]

Summary

Since genotype imputation was introduced, researchers have been relying on the estimated imputation quality from imputation software to perform post-imputation quality control (QC). However, this quality estimate (denoted as Rsq) performs less well for lower-frequency variants. We recently published MagicalRsq, a machine-learning-based imputation quality calibration, which leverages additional typed markers from the same cohort and outperforms Rsq as a QC metric. In this work, we extended the original MagicalRsq to allow cross-cohort model training and named the new model MagicalRsq-X. We removed the cohort-specific estimated minor allele frequency and included linkage disequilibrium scores and recombination rates as additional features. Leveraging whole-genome sequencing data from TOPMed, specifically participants in the BioMe, JHS, WHI, and MESA studies, we performed comprehensive cross-cohort evaluations for predominantly European and African ancestral individuals based on their inferred global ancestry with the 1000 Genomes and Human Genome Diversity Project data as reference. Our results suggest MagicalRsq-X outperforms Rsq in almost every setting, with 7.3%–14.4% improvement in squared Pearson correlation with true R^2, corresponding to 85–218 K variant gains. We further developed a metric to quantify the genetic distances of a target cohort relative to a reference cohort and showed that such metric largely explained the performance of MagicalRsq-X models. Finally, we found MagicalRsq-X saved up to 53 known genome-wide significant variants in one of the largest blood cell trait GWASs that would be missed using the original Rsq for QC. In conclusion, MagicalRsq-X shows superiority for post-imputation QC and benefits genetic studies by distinguishing well and poorly imputed lower-frequency variants.

Genotype imputation has become an essential step for genome-wide association studies (GWASs) and other downstream genetic analyses. Post-imputation quality control (QC) has always been performed to remove poorly imputed genetic variants.[1–4] In imputation settings with no true genotypes available and thus no true imputation quality (true R^2), scientists have been relying on estimated quality metrics given by imputation engines for QC purposes. The most widely used estimated imputation quality metric is Rsq, a standard output from MaCH and mimimac series[5–8] that is included in the default output from Michigan and TOPMed imputation servers. However, Rsq performs less well for uncommon variants with minor allele frequency (MAF) <5% and will likely lead to information loss and/or inclusion of noise.[9–12] We recently developed MagicalRsq, which we showed to be a better metric compared to the standard Rsq.[12] However, MagicalRsq only focused on within-cohort applications requiring additional genotypes (for example, whole-exome sequencing [WES] or whole-genome sequencing [WGS] data) in at least a subset of the imputation target samples. For practical purposes, we need pre-trained models to accommodate arguably the most common real-life scenario where the target samples have only one set of genotype data available. In

[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]Division of Pediatric Endocrinology, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA; [7]Department of Pediatrics, University of Washington, Seattle, WA 98105, USA; [8]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [9]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA; [10]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA; [11]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [12]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35249, USA; [13]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22908, USA; [14]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [15]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA; [16]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; [17]Wake Forest School of Medicine, Department of Biostatistics and Data Science, Wake Forest University, Winston-Salem, NC 27109, USA; [18]Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; [19]Institute for Biomedicine, Eurac Research (affiliated with the University of Lübeck), Bolzano, Italy
[20]These authors contributed equally
*Correspondence: cfuchsberger@eurac.edu (C.F.), yunli@med.unc.edu (Y.L.)
https://doi.org/10.1016/j.ajhg.2024.04.001.

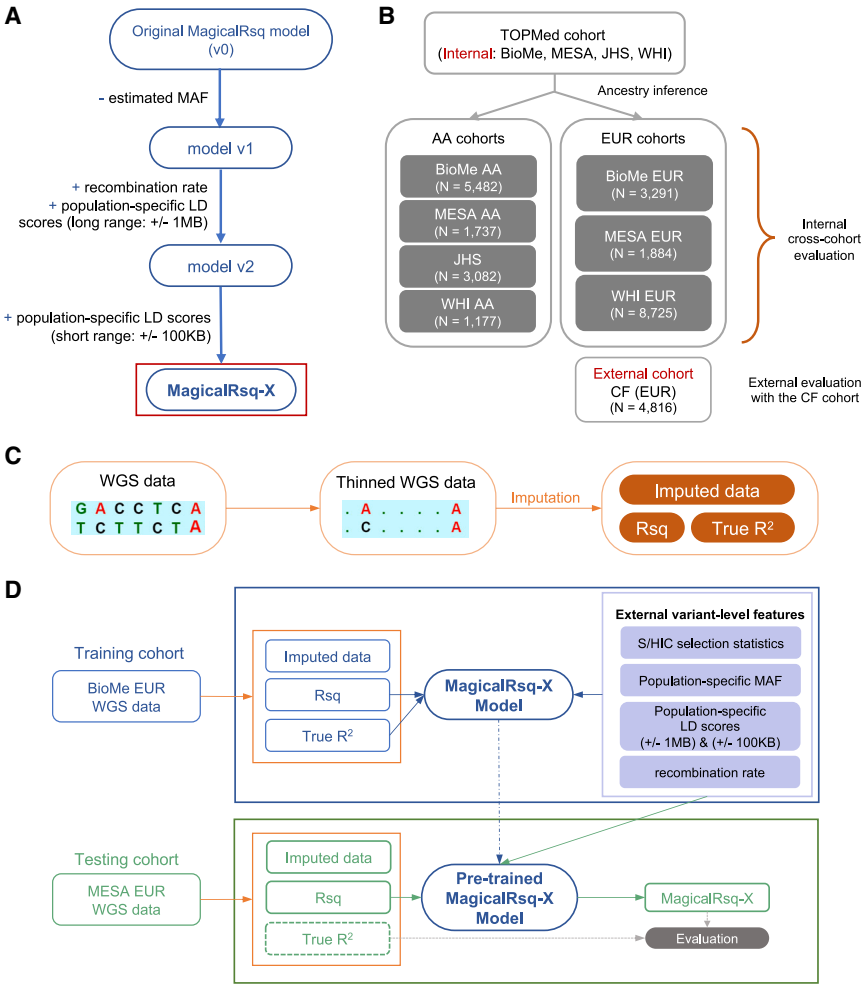

Figure 1. MagicalRsq-X overview

(A) Feature modification from the original MagicalRsq model. We first removed the estimated MAF feature derived from imputation output, which we refer to as MagicalRsq-X model v1. We then added recombination rate from 1000G and long-range LD scores (± 1 Mb) of four continental populations from TOP-LD, leading to MagicalRsq-X model v2. Finally, we added short-range LD scores (± 100 Kb) of the same four populations from TOP-LD, resulting in MagicalRsq-X model v3, which is the final model showing the best and most robust performance.

(B) Overview of study cohorts in our evaluations. We leveraged TOPMed WGS data of four studies, BioMe, MESA, JHS, and WHI, as our internal evaluation cohorts. We first inferred local and global ancestry of individuals in these studies and then selected individuals who are primarily of European ancestry or admixed African ancestry based on inferred global genetic similarity (detailed in supplemental methods). We also added the CF participants as an external evaluation cohort.

(C) Data preparation for MagicalRsq-X experiments. We first thinned the WGS data to array genotype density and then performed genotype imputation, which outputs individual-level imputed data and Rsq. We then calculated true R^2 comparing imputed data with WGS data for imputed markers (i.e., those in WGS but not included in the thinned dataset).

(D) Model training and evaluation using BioMe EUR for training and MESA EUR for testing as an example. Starting from BioMe EUR WGS data, we performed imputation as demonstrated in (C). After obtaining all the external variant-level features, which were further combined with true R^2 and Rsq, we trained MagicalRsq-X models. For the testing cohort, MESA EUR in this example, we similarly performed data thinning and imputation. We then applied the models pre-trained from BioMe EUR to calculate MagicalRsq-X for MESA EUR. In our experiments, we similarly calculated true R^2 in MESA EUR and evaluated the performance of MagicalRsq-X compared to Rsq. The dashed square around "true R^2" in testing set means it is not required in real-life application and was used in our evaluation purpose.

this work, we proposed MagicalRsq-X, which modifies some variant-level features from the original model. Specifically, we removed estimated MAF and added linkage disequilibrium (LD) scores from TOP-LD[13] and recombination rate from 1000 Genomes Project (1000G)[14] (Figure 1A; supplemental notes). MagicalRsq-X allows model training from a completely different cohort (cross-cohort model training), making it more broadly applicable in diverse real-life scenarios. The design of MagicalRsq-X is for studies without additional genotypes, and thus, borrowing information from other studies is mostly needed. Because of the design, MagicalRsq-X does not require additional genotypes from the target cohort.

The original MagicalRsq model takes both imputation summary statistics (Rsq, estimated MAF) and population genetics statistics (ancestry-specific MAF,[13] S/HIC features[15]) as input, divides variants into three commonly used MAF categories (common, MAF >5%; low frequency, MAF 0.5%–5%; and rare, MAF <0.5%),[1,12] and trains an XGBoost model[16] separately in each MAF category. It requires, among the imputation target samples, additional genotypes (e.g., from a different genotyping platform or WES or WGS) not used when performing the imputation. In this work, MagicalRsq-X adopts the same framework (Figures 1C and 1D) but modifies the variant-level features to allow training models from a different cohort, i.e., cross-cohort training (Figure 1A; supplemental notes). We removed the estimated MAF feature (estimated from imputed data) because this feature is susceptible to subtle differences between training and testing cohorts, especially for rare variants where substantial discrepancies may exist across cohorts (Figures S1 and S2). Note that ancestry-specific MAFs from TOP-LD[13] remain in the model. We also added population-specific LD scores calculated based on TOP-LD[13] at both 1-Mb and 100-kb windows to reflect longer- and shorter-range LD patterns. Moreover, we added the recombination rate from the 1000G[14] as an additional feature (supplemental notes).

We tested MagicalRsq-X models leveraging WGS data of four studies from the Trans-Omics for Precision Medicine (TOPMed) project. Specifically, we included participants from BioMe Biobank (BioMe),[17] the Multi-Ethnic Study of Atherosclerosis (MESA),[18] the Jackson Heart Study (JHS),[19,20] and the Women's Health Initiative (WHI)[21] for cross-validation. Based on genetic ancestry estimates from RFMix[22] with combined 1000G and Human Genome Diversity Project (HGDP) as reference, we selected individuals with primarily European ancestry (EUR, estimated European global ancestry >85%) or admixed individuals with both European and African ancestry (estimated European and African global ancestry both >10% and summation >50%) (supplemental methods). For narrative simplicity, we hereafter refer to these admixed individuals as admixed African (AA), but we note that this grouping, derived from estimated genetic ancestry, should not be equated with self-identified population descriptors (such as African American). With this genetic ancestry-based grouping, we have three EUR cohorts and four AA cohorts (Figure 1B; supplemental methods; Table S1). An example of our MagicalRsq-X model training and testing framework is illustrated in Figures 1C and 1D. We first thinned the WGS data to mimic array genotypes and then performed imputation separately for the four EUR cohorts with the Haplotype Reference Consortium (HRC) reference panel and for the three AA cohorts with the 1000G reference panel (supplemental methods). After calculating true R^2, we trained MagicalRsq models separately for each cohort and separately for variants in three MAF categories (common, low frequency, and rare). In our experiments, for each MAF category, we randomly selected 10 K, 50 K, 100 K, 200 K, 500 K, and 1 M variants for model training, with five repeats each to assess model stability. We then performed cross-validation within EUR or AA to evaluate MagicalRsq-X models (supplemental methods).

We first note that the modified features in MagicalRsq-X both ranked high in feature importance (supplemental notes; Figures S9 and S10) and improved model performance (supplemental notes; Figures S3–S8). Similar to our prior MagicalRsq study, we evaluated model performance using two sets of metrics: the squared Pearson correlation, root mean squared error (RMSE), and mean absolute error (MAE) with true R^2 for direct comparison, as well as counts of variant net gains for comparison of the ability to perform post-imputation QC (supplemental methods). Among the three EUR cohorts, our experiments show that MagicalRsq-X outperforms Rsq for every pair of training-testing datasets for almost all scenarios (Figures 2A, S11, and S12; Tables S2 and S3). For example, leveraging low-frequency variant models trained from BioMe EUR, MagicalRsq-X improves squared Pearson correlation with true R^2 by 4.8%–7.3% and 4.4%–6.9%, decreases RMSE by 12.6%–20.9% and 18.2%–31.6%, and decreases MAE by 5.4%–14.6% and 10.0%–24.0% for MESA EUR and WHI EUR, respectively, compared to standard Rsq (Table S2). For common variants where the original Rsq already shows decent performance, MagicalRsq-X is still more consistent with true R^2 than Rsq (Figure 2C). MagicalRsq-X also shows advantages as a quality-filtering metric to distinguish well-imputed variants from poorly imputed ones. For instance, it leads to net gains of 16–24 K common, 45–68 K low-frequency, and 19–236 K rare variants across five repeats compared to Rsq in the BioMe EUR cohort (Table S3), leveraging models trained on the other two EUR cohorts. Note that such net gains come from two parts: saving truly well-imputed variants excluded by Rsq and excluding truly poorly imputed variants included by Rsq. Contributions of the two components depend on whether Rsq overestimated or underestimated true R^2, which may vary depending on the target cohort. Furthermore, we notice that MagicalRsq-X trained with only 50 K variants already shows rather stable results, consistent with our previous observations.[12] For models trained with $\geq$50 K variants, the minimum net gains for MagicalRsq-X in BioMe EUR are ~80 K for rare variants.

The performance of MagicalRsq-X for the AA cohorts is similarly satisfying, leading to overall mean improvement of 6.1%, 10.2%, and 8.3% in squared Pearson correlation, 20.3%, 18.7%, and 9.5% in RMSE, and 20.1%, 16.3%, and 5.5% in MAE, for common, low-frequency, and rare variants, respectively (Figures 2B, 2D, 2E, S13, and S14; Tables S4 and S5). For a specific example, rare variants in MESA AA could benefit from MagicalRsq-X trained on BioMe AA, JHS, and WHI AA 50 K–1 M variant models by 7.8%–9.9%, 7.3%–9.5%, and 8.5%–10.9% in terms of squared Pearson correlation with true R^2 (Table S4), corresponding to 85–170 K, 119–204 K, and 99–218 K net gains of variants (Table S5), respectively. MagicalRsq-X also shows satisfying performance for common and low-frequency variants (Figures S13 and S14), with the only exceptions between BioMe AA and JHS. We note that the genetic background of our defined AA cohorts is complicated, and the difference between BioMe and JHS is the largest among all the pairs (Figure S15). BioMe AA is the most genetically diverse cohort, which is not surprising as it is a biobank-based study from a diverse region of the U.S. (New York City) with individuals born in locations across the globe,[23] while JHS is the most homogeneous, likely due to its geographical-centralized recruitment in Mississippi, with substantially less recent migration from diverse geographic regions. Such differential levels of genetic ancestry matching between cohorts is likely a driving factor affecting cross-cohort MagicalRsq-X performance. We then performed experiments to include only individuals in BioMe AA that could be reasonably matched with JHS samples based on the harmonized principal components (PCs) (n = 2,219) (supplemental methods) and observed marked improvement (Figures S16–S21; Table S6; supplemental notes). For example, applying models trained from JHS individuals to the whole BioMe AA set, MagicalRsq-X is inferior to Rsq in terms of squared Pearson correlation by up to 22.3% and 27.4% for common and low-frequency variants, but it shows clear advantages over Rsq with

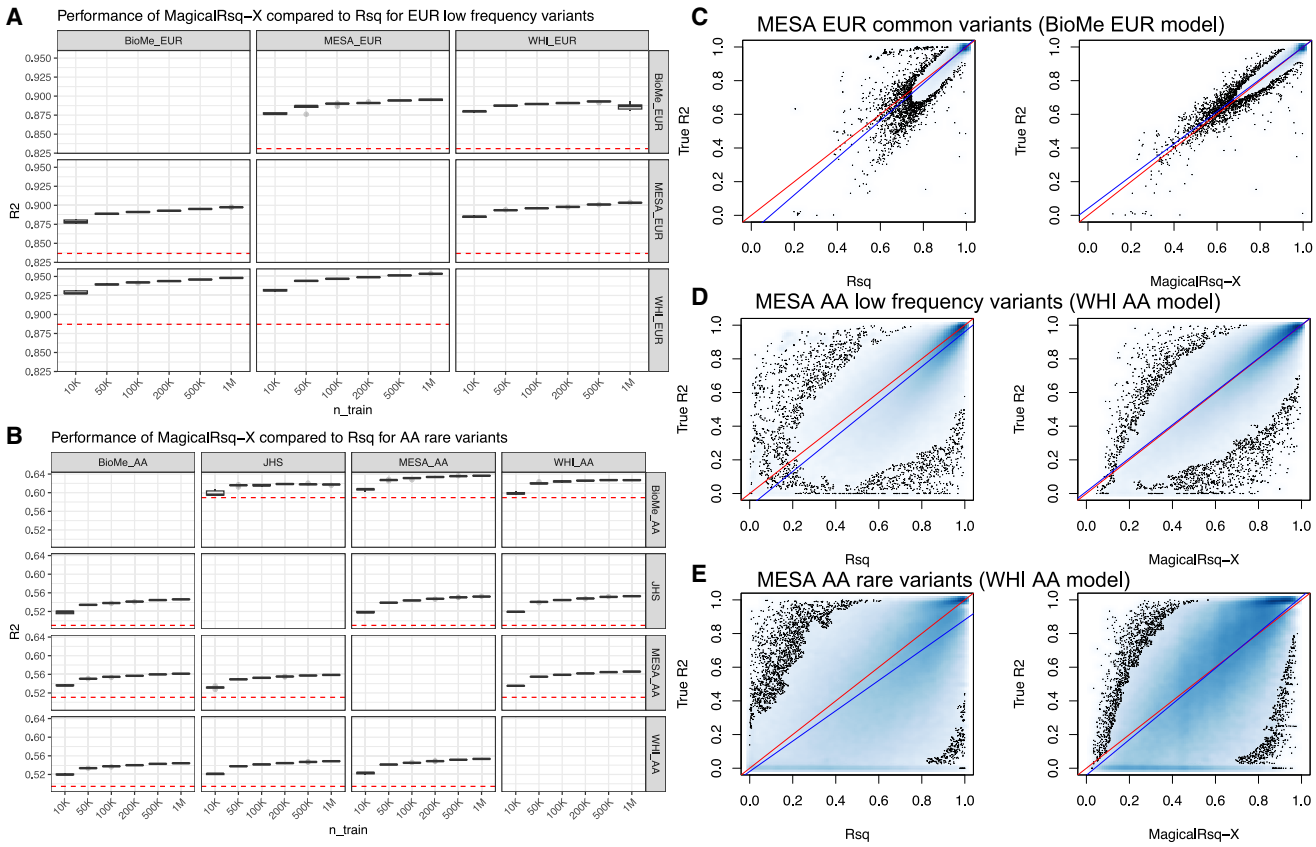

Figure 2. Cross-cohort MagicalRsq-X model performance

(A) Performance across the three EUR cohorts (BioMeEUR, MESA EUR, and WHI EUR) for low-frequency variants (MAF [0.5%, 5%]). We trained MagicalRsq-X models with randomly selected 10 K, 50 K, 100 K, 200 K, 500 K, and 1 M variants (x axis), each with five repeats. y axis is the squared Pearson correlation between MagicalRsq-X and true R^2. Each row represents a testing cohort, and each column represents a training cohort. The diagonal components are missing on purpose because we only assess cross-cohort model performance. Red dashed lines represent squared Pearson correlation between standard Rsq and true R^2, which serves as the benchmark.

(B) Performance across the four AA cohorts (BioMe AA, JHS, MESA AA, and WHI AA) for rare variants (MAF <0.5%).

(C–E) Comparison between true R^2 vs. Rsq and true R^2 vs. MagicalRsq-X for MESA EUR common variants (C), MESA AA low-frequency variants (D), and MESA AA rare variants (E) on chr10, where MagicalRsq-X shown was calculated from models trained with 100K variants from BioMe EUR (C) and WHI AA (D and E). For the smooth scatterplots, the darker the color, the larger the number of variants. Outliers are plotted separately. Red lines are 45-degree lines, and blue lines are the fitted lines.

7.3%–8.7% and 12.9%–14.4% improvement in the subset of matched individuals for common and low-frequency variants, respectively (Figures S16 and S17; Table S6), supporting our speculation that the aforementioned lower performance is likely driven by substantial dissimilarity in the distribution of genetic profiles in the two cohorts. To quantify such cross-cohort dissimilarity, we developed a quantitative metric to characterize how different a target cohort is compared to a reference cohort based on harmonized PCs (supplemental methods). We found that such a PC-based metric could largely explain why we observed minimal or no improvement for some MagicalRsq-X models, where larger dissimilarity in PC metric would result in worse MagicalRsq-X performance (Figures S22 and S23; Table S7). Our proposed metric provides guidance regarding the choice of MagicalRsq-X reference models. In practice, we recommend users to be cautious when applying MagicalRsq-X between cohorts with the PC-based dissimilarity metric >0.03 based on our evaluations.

We also tested our models on the cystic fibrosis (CF) samples of European ancestry[24] as an external cohort to validate the performance of MagicalRsq-X outside of the TOPMed studies (supplemental methods). Our evaluations resulted in similarly satisfying results (Table S8; Figures S24 and S25). For example, leveraging models trained from WHI EUR cohorts with 100 K variants, MagicalRsq-X improves squared Pearson correlation with true R^2 in CF samples by 6.0%–6.4%, 6.4%–6.5%, and 6.2%–6.4% for common, low-frequency, and rare variants, respectively. The results further support the advantages of applying MagicalRsq-X to external cohorts, suggesting the broad practical utility of MagicalRsq-X.

Encouraged by the improved accuracy of MagicalRsq-X as a post-imputation quality-filtering metric, we then performed experiments to evaluate its benefits in downstream association analysis. We assembled known GWAS significant variants for 15 blood cell traits from prior analyses using TOPMed WGS data,[25–27] resulting in 8,321

variants in total, including rare variants revealed from burden tests. After overlapping with variants in the imputed data, 3,251, 3,287, and 3,316 variants remained in BioMe AA, WHI AA, and JHS, respectively. We calculated MagicalRsq-X for these variants separately in the three cohorts with MagicalRsq-X models trained on MESA AA and compared MagicalRsq-X and Rsq in terms of squared Pearson correlation with true R^2 and the net gains of variants under different thresholds (supplemental methods). Overall, MagicalRsq-X improved squared correlation with true R^2 from 0.92 to 0.94 for BioMe AA, from 0.89 to 0.92 for WHI AA, and from 0.89 to 0.91 for JHS, indicating better alignment with true R^2. For filtering variants, we found MagicalRsq-X achieved net gains of 9–53 variants for these associated variants (which can be viewed as positive control association signals) under commonly used thresholds (Table S9). For example, rs9273039 at *HLA* locus was found to be significantly associated with hematocrit[25] and was well imputed in both JHS and MESA AA (true $R^2 = 0.97$ for both cohorts), but the original Rsqs were only 0.45 and 0.44. In contrast, MagicalRsq-X could successfully rescue this association signal with values of 0.88 for both cohorts. These results again illustrate the advantages of MagicalRsq-X over standard Rsq in downstream analyses.

In summary, we present MagicalRsq-X, which significantly extends our previously published MagicalRsq by allowing cross-cohort applications without the need for additional genotype data from the target cohort. Note that the features we added (LD scores and recombination rate) are highly influential to the model performance. We additionally found that variants with low LD scores or residing in regions with high recombination rate benefit the most from MagicalRsq-X (Figures S26 and S27). Our comprehensive experiments and evaluations demonstrate the advantages of MagicalRsq-X as a quality-filtering metric and its benefits in downstream analyses. Similar to our original MagicalRsq, MagicalRsq-X is robust to different choices of number of variants used for model training where multiple repeats with different randomly selected variants showed minimal variations. In addition, MagicalRsq-X performs similarly well or even better in some cases compared to MagicalRsq, especially for common and low-frequency variants (Figure S28), emphasizing the value of this extension compared to MagicalRsq, as in many real studies we do not have the luxury of performing internal training with MagicalRsq. We release our pre-trained models for the convenience of other researchers but note that our pre-trained models were all trained on U.S.-based cohorts due to data availability. It warrants future investigations about whether these U.S.-based models could also benefit other populations. We encourage other investigators to train MagicalRsq-X models whenever relevant data are available. MagicalRsq-X software and our pre-trained models are freely available at https://github.com/quansun98/MagicalRsqX.

Data and code availability

MagicalRsq-X is freely available at https://github.com/quansun98/MagicalRsqX. Our pre-trained models could also be downloaded at ftp://yunlianon:anon@rc-ns-ftp.its.unc.edu/MagicalRsqX/models/ in addition to the GitHub page.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.04.001.

Acknowledgments

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. W.G. was supported by NIH grant T32ES007018-46.

This study was reviewed by the Cystic Fibrosis Foundation for the use of CF Foundation Patient Registry data and CFGP WGS data and by the TOPMed consortium for the use of TOPMed WGS data. The procedures followed for data collection and processing and 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.

We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The TOPMed Banner Authorship list can be found at https://www.nhlbiwgs.org/topmed-banner-authorship. Detailed cohort-specific acknowledgments could be found in supplemental information.

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.[28]

Declaration of interests

The authors declare no competing interests.

Received: January 25, 2024
Accepted: April 1, 2024
Published: April 17, 2024

Web Resources

COVID-19 HGI Projected PC: https://github.com/covid19-hg/pca_projection/tree/master

MagicalRsq: https://github.com/quansun98/MagicalRsq

MagicalRsq-X: https://github.com/quansun98/MagicalRsqX

Michigan imputation server: https://imputationserver.sph.umich.edu/

TOP-LD: http://topld.genetics.unc.edu/

TOPMed: https://topmed.nhlbi.nih.gov/

References

1. 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.
2. Sun, Q., Broadaway, K.A., Edmiston, S.N., Fajgenbaum, K., Miller-Fleming, T., Westerkam, L.L., Melendez-Gonzalez, M., Bui, H., Blum, F.R., Levitt, B., et al. (2023). Genetic variants associated with hidradenitis suppurativa. JAMA Dermatol. *159*, 930–938.
3. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
4. Huerta-Chagoya, A., Schroeder, P., Mandla, R., Deutsch, A.J., Zhu, W., Petty, L., Yi, X., Cole, J.B., Udler, M.S., Dornbos, P., et al. (2023). The power of TOPMed imputation for the discovery of Latino-enriched rare variants associated with type 2 diabetes. Diabetologia *66*, 1273–1288.
5. 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.
6. Liu, E.Y., Li, M., Wang, W., and Li, Y. (2013). MaCH-admix: genotype imputation for admixed populations. Genet. Epidemiol. *37*, 25–37.
7. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784.
8. 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.
9. Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu. Rev. Genom. Hum. Genet. *10*, 387–406.
10. 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.
11. 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.
12. Sun, Q., Yang, Y., Rosen, J.D., Jiang, M.-Z., Chen, J., Liu, W., Wen, J., Raffield, L.M., Pace, R.G., Zhou, Y.-H., et al. (2022). MagicalRsq: Machine-learning-based genotype imputation quality calibration. Am. J. Hum. Genet. *109*, 1986–1997.
13. 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.
14. 1000 Genomes Project Consortium, 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., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
15. Schrider, D.R., and Kern, A.D. (2016). S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. *12*, e1005928.
16. 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 (New York, New York, USA: ACM Press), pp. 785–794.
17. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M.A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. *15*, 761–771.
18. Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R., Kronmal, R., Liu, K., et al. (2002). Multi-Ethnic Study of Atherosclerosis: objectives and design. Am. J. Epidemiol. *156*, 871–881.
19. Taylor, H.A., Wilson, J.G., Jones, D.W., Sarpong, D.F., Srinivasan, A., Garrison, R.J., Nelson, C., and Wyatt, S.B. (2005). Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn. Dis. *15*, S6–S17.
20. Wilson, J.G., Rotimi, C.N., Ekunwe, L., Royal, C.D.M., Crump, M.E., Wyatt, S.B., Steffes, M.W., Adeyemo, A., Zhou, J., Taylor, H.A., and Jaquish, C. (2005). Study design for genetic analysis in the Jackson Heart Study. Ethn. Dis. *15*, S6–S37.
21. (1998). Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control. Clin. Trials *19*, 61–109.
22. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
23. Belbin, G.M., Cullina, S., Wenric, S., Soper, E.R., Glicksberg, B.S., Torre, D., Moscati, A., Wojcik, G.L., Shemirani, R., Beckmann, N.D., et al. (2021). Toward a fine-scale population health monitoring system. Cell *184*, 2068–2083.e11.
24. 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.
25. Hu, Y., Stilp, A.M., McHugh, C.P., Rao, S., Jain, D., Zheng, X., Lane, J., Méric de Bellefon, S., Raffield, L.M., Chen, M.-H., et al. (2021). Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. Am. J. Hum. Genet. *108*, 1165–1893.
26. Mikhaylova, A.V., McHugh, C.P., Polfus, L.M., Raffield, L.M., Boorgula, M.P., Blackwell, T.W., Brody, J.A., Broome, J., Chami, N., Chen, M.-H., et al. (2021). Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: The NHLBI TOPMed program. Am. J. Hum. Genet. *108*, 1836–1851.
27. Little, A., Hu, Y., Sun, Q., Jain, D., Broome, J., Chen, M.-H., Thibord, F., McHugh, C., Surendran, P., Blackwell, T.W., et al. (2022). Whole genome sequence analysis of platelet traits in the NHLBI Trans-Omics for Precision Medicine (TOPMed) initiative. Hum. Mol. Genet. *31*, 347–361.
28. 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.