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

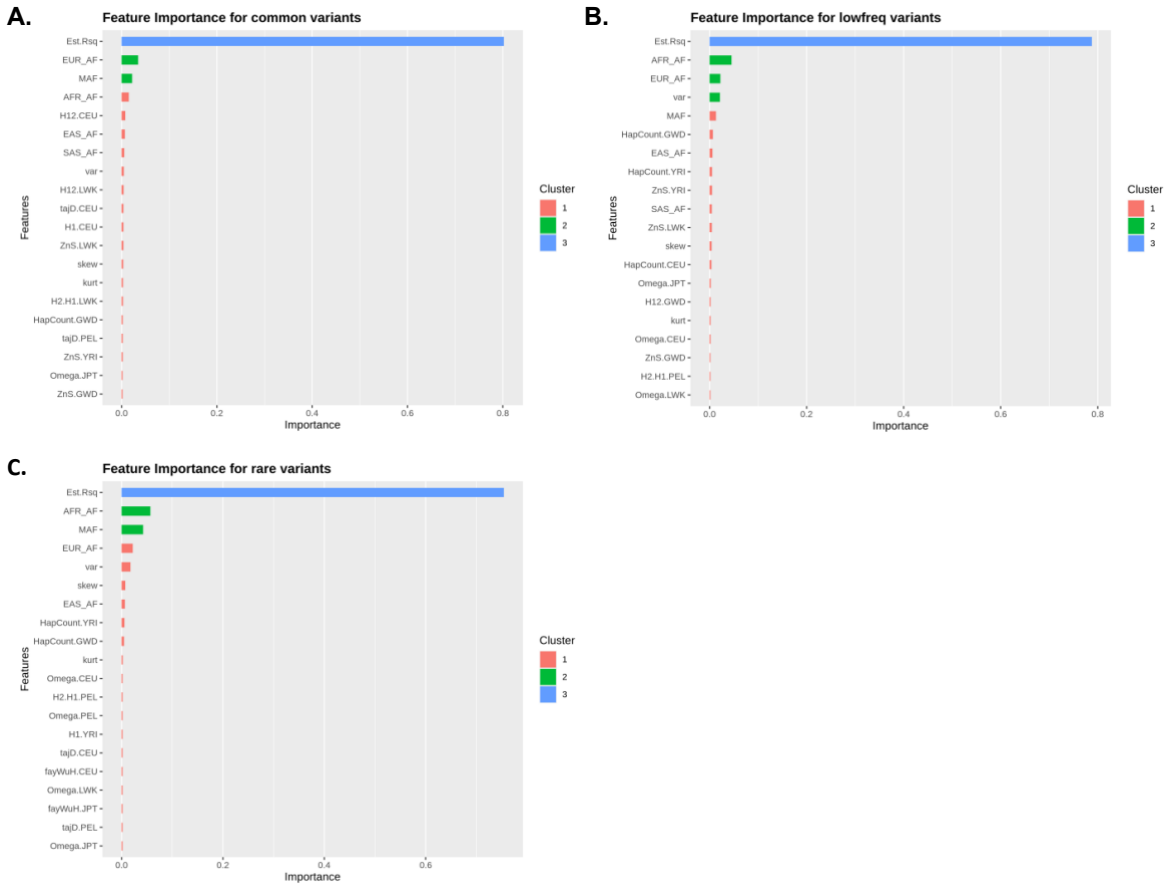

Figure S1. Feature importance for MagicalRsq models in Scenario 1 Experiment 1. The standard Rsq weighs the highest and is about 80% importance for all the three categories. European allele count (AC) is the second most important feature for common variants, but African AC is the second most important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

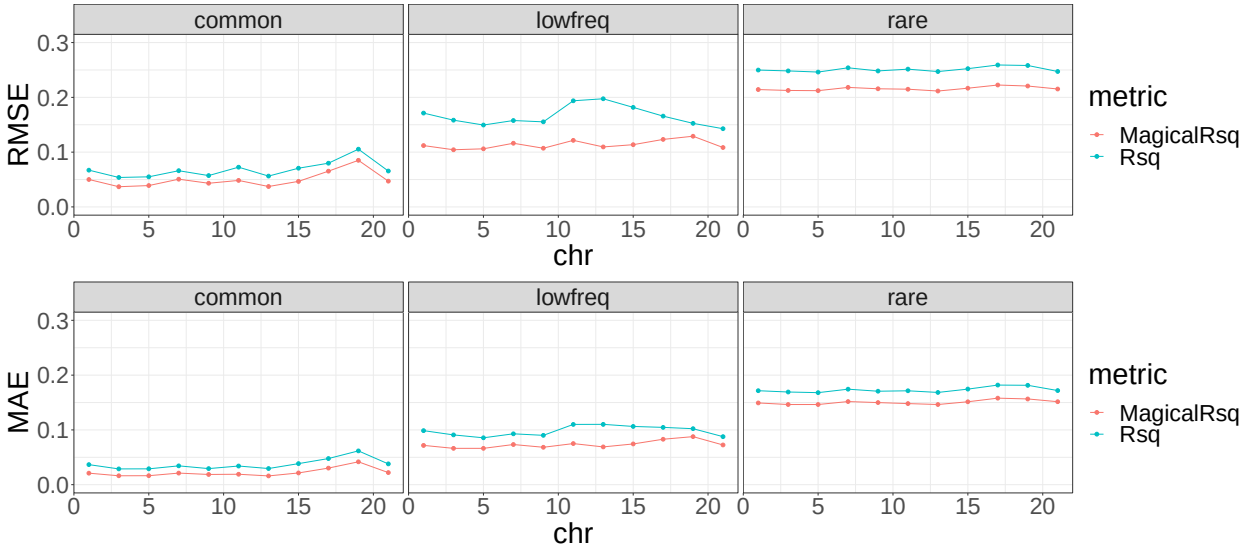

Figure S2. Performance comparison of MagicalRsq and Rsq in terms of RMSE and MAE, for Scenario 1 Experiment 1. Imputation was performed using 1000G reference panel, and MagicalRsq was calculated from model trained on CF 2k even number chromosomes which was also imputed using 1000G reference panel.

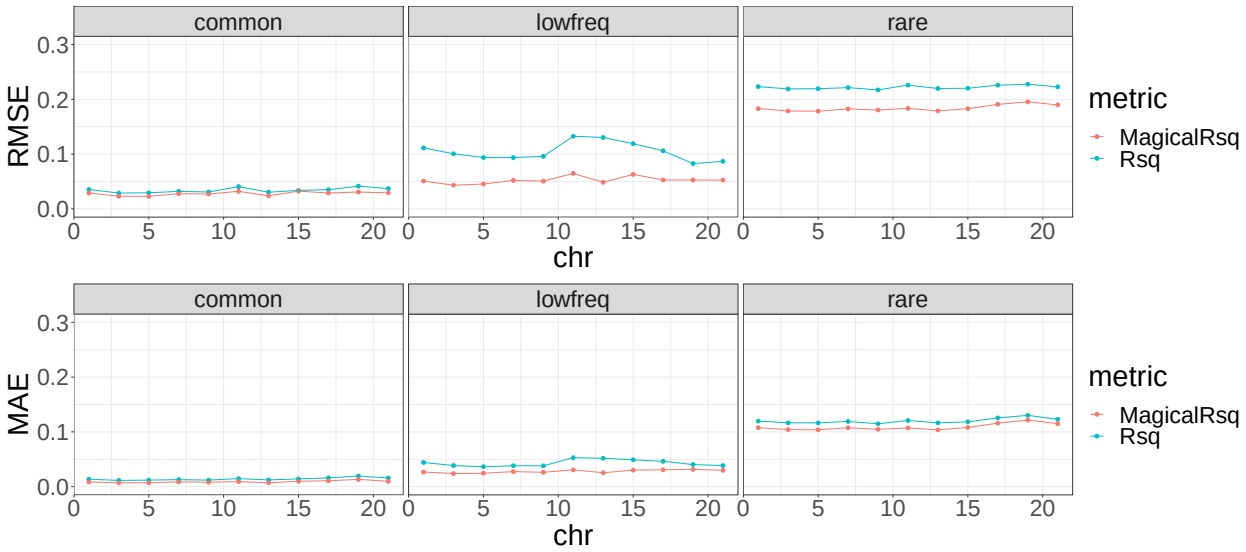

Figure S3. Performance comparison of MagicalRsq and Rsq in terms of RMSE and MAE, for Scenario 1 Experiment 2. Imputation was performed using TOPMed freeze 8 reference panel, and MagicalRsq was calculated from model trained on CF 2k even number chromosomes which was also imputed using TOPMed freeze 8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

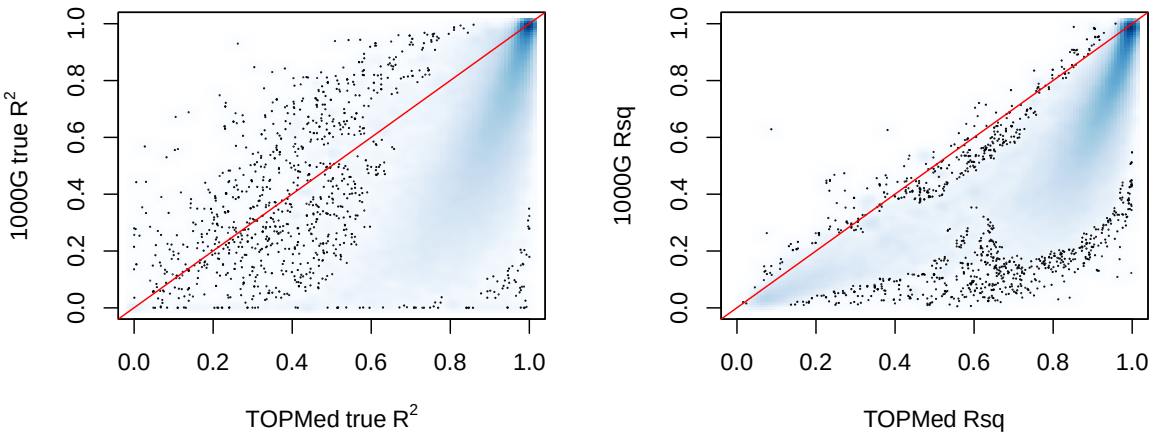

Figure S4. Comparison between 1000G and TOPMed imputation. We plotted 1000G true R^2 against TOPMed true R^2, and 1000G Rsq against TOPMed Rsq from imputation 1 and 2. Tough the squared Pearson correlation between TOPMed Rsq and TOPMed true R^2 is smllaer than 1000G, TOPMed imputation quality (and the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

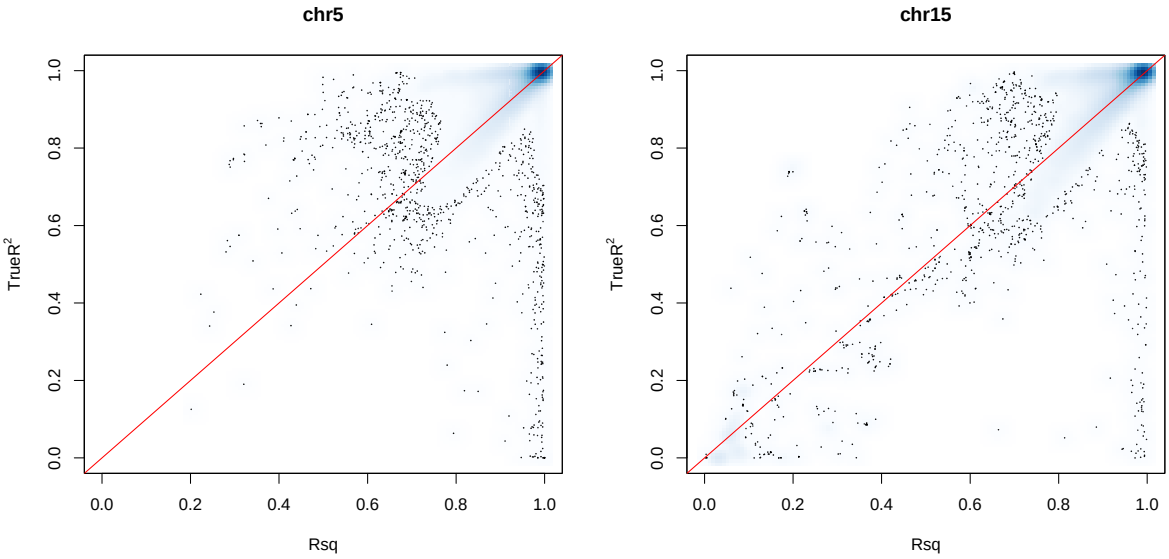

Figure S5. Rsq v.s. True R^2 for common variants on chr5 and chr15 for TOPMed imputed CF 2k cohort. We observed the fluctuation of Rsq performance for different chromosomes across the genome for CF 2k cohort, and this is likely due to the different spanning range of Rsq. Larger range would lead to higher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

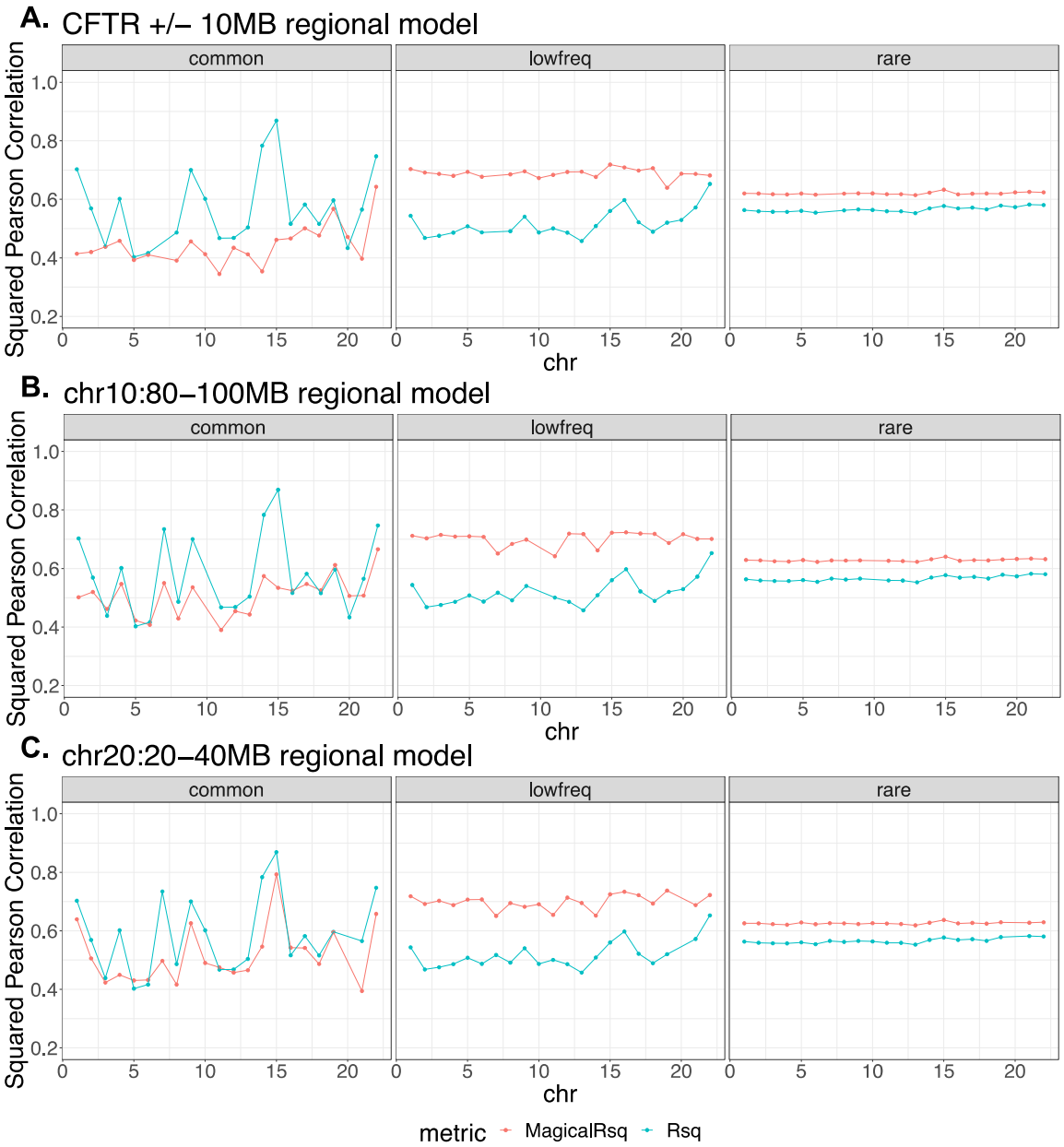

Figure S6. Scenario 1 Experiment 5: training models using variants in a 20MB region and testing on all other chromosomes for CF 2k samples with TOPMed imputation. Performance comparison between Rsq and MagicalRsq in terms of squared Pearson correlation with true R^2 for models trained with variants in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

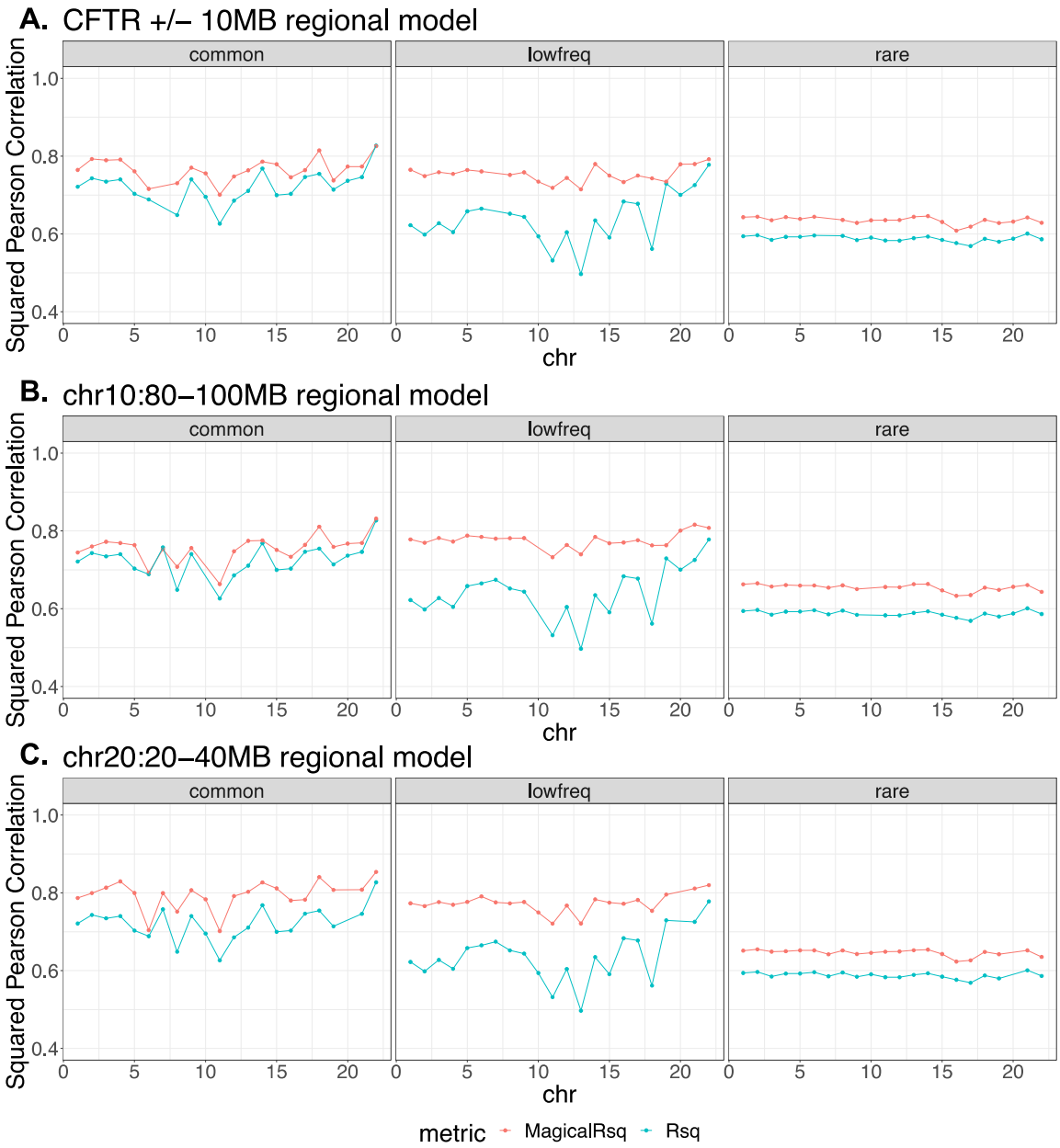

Figure S7. Scenario 1 Experiment 6: training models using variants in a 20MB region and testing on all other chromosomes for CF 2k samples with 1000G imputation. Performance comparison between Rsq and MagicalRsq in terms of squared Pearson correlation with true R^2 for models trained with variants in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

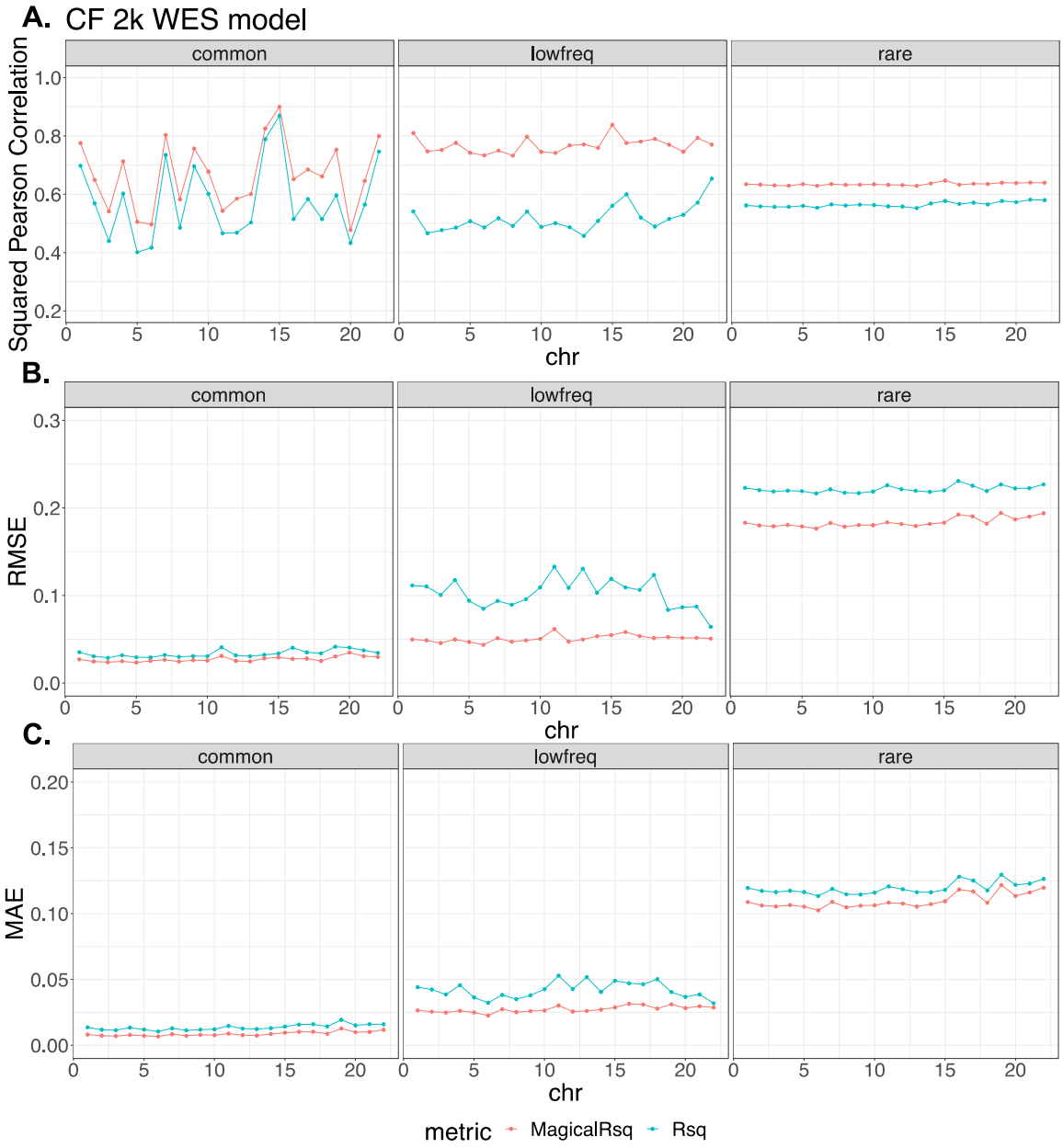

Figure S8. Scenario 1 Experiment 7: training models using TOPMed imputed exonic variants from CF 2k samples and testing on TOPMed imputed variants in other genomic regions of the same CF 2k samples. Performance comparison between Rsq and MagicalRsq in terms of **(A)** squared Pearson correlation with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

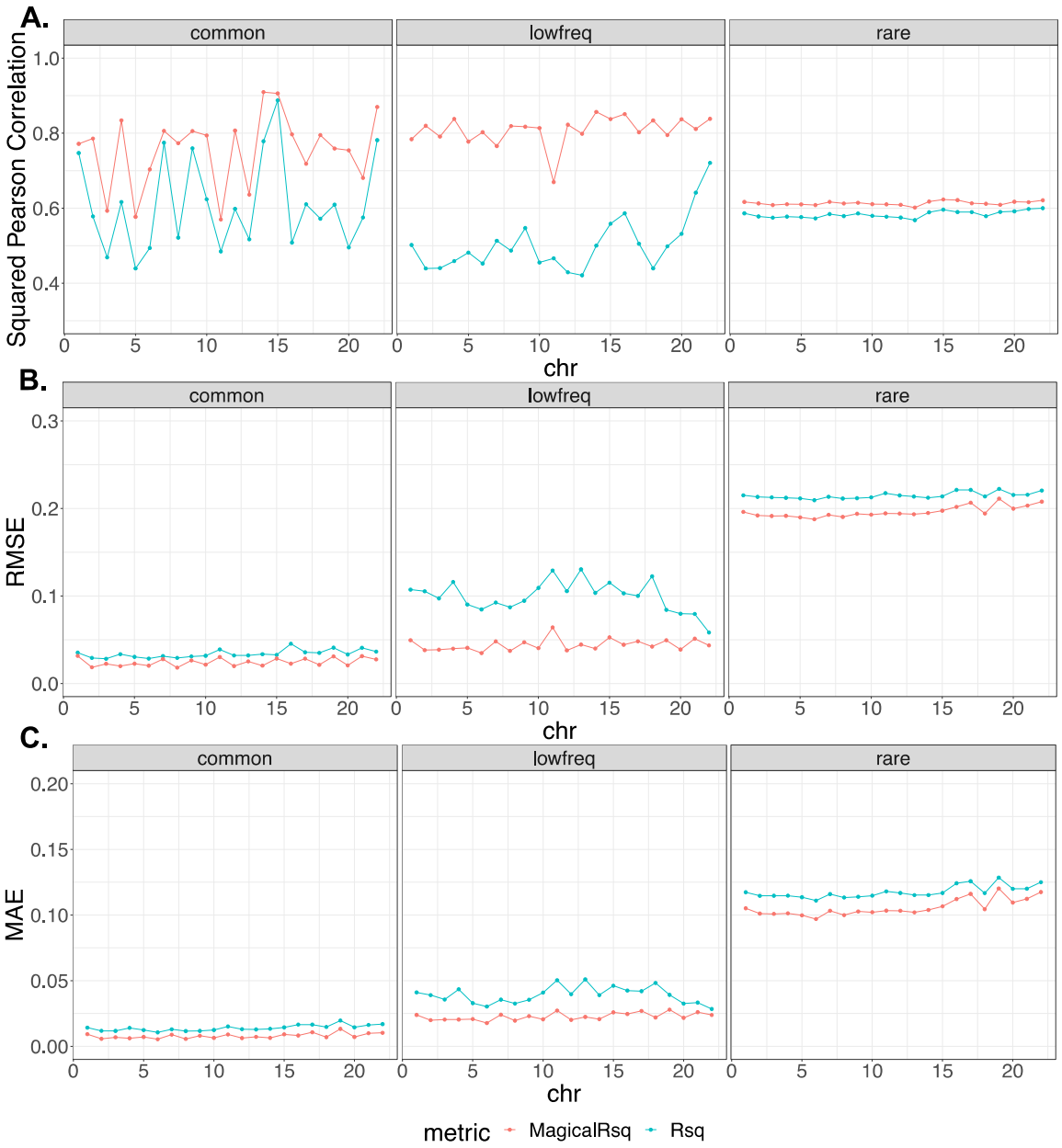

Figure S9. Scenario 2 Experiment 8: training models using TOPMed imputed variants from CF 2k samples and testing on TOPMed imputed all chromosomes of independent CF 3k samples. Performance comparison between Rsq and MagicalRsq in terms of **(A)** squared Pearson correlation with true R^2; **(B)** RMSE;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

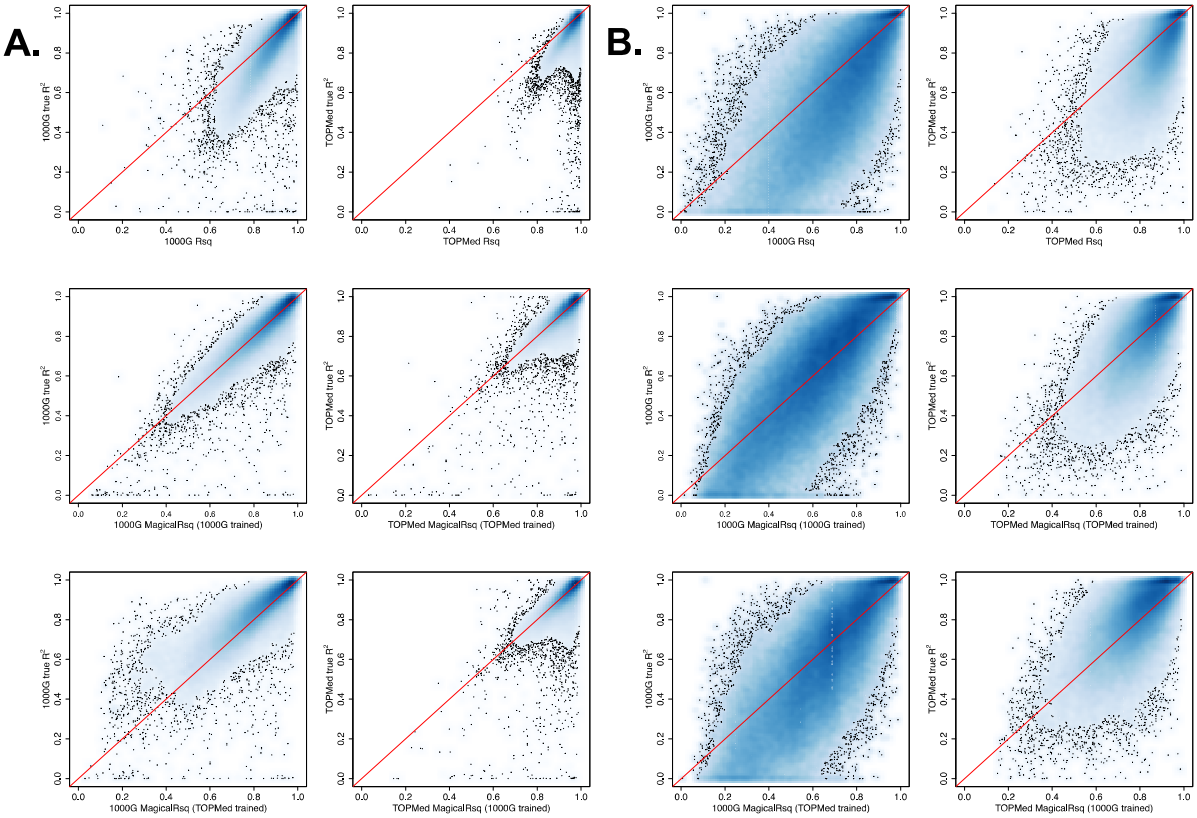

Figure S10. Scenario 2 Experiment 9-12: training models using 1000 UKB AFR samples and testing on 2960 independent UKB AFR samples, for all variants with WES available. Smooth scatter plot showing Rsq or MagicalRsq (X-axis) calculated from both matched- (second row) and mis-matched- (third row) models 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.

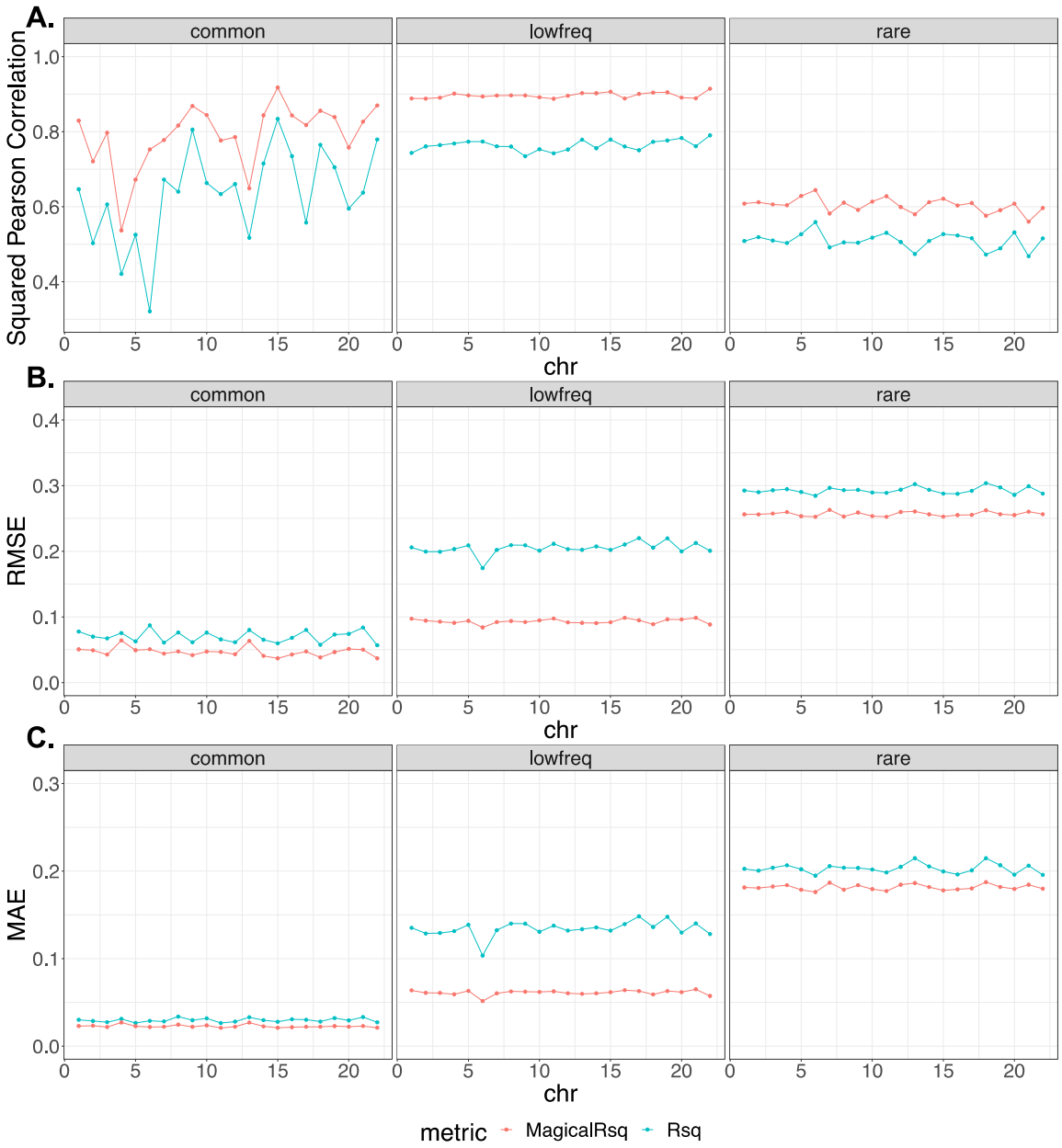

Figure S11. Scenario 2 Experiment 13: training models using TOPMed imputed variants of 1,000 UKB SAS samples and testing on TOPMed imputed variants (across all chromosomes) of an independent set of UKB SAS 3,436 samples. Performance comparison between Rsq and MagicalRsq in terms of **(A)** squared Pearson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

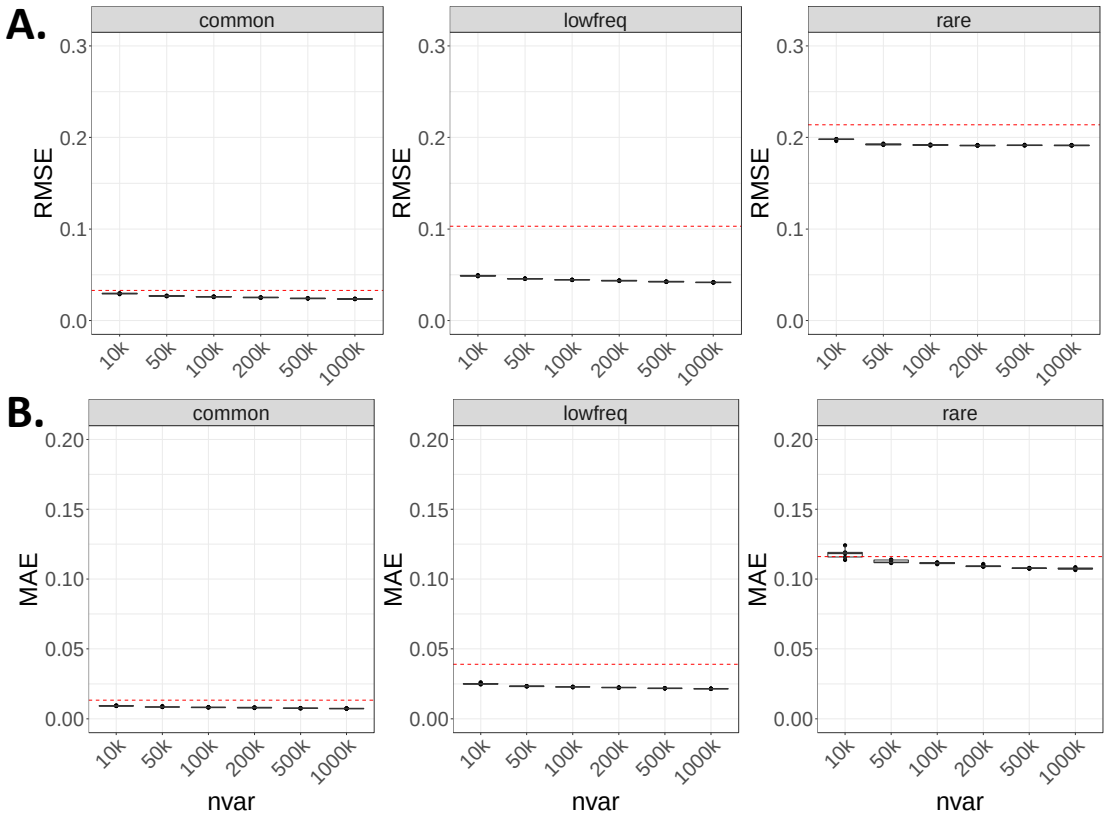

Figure S12. Scenario 2 Experiment 14: training models using randomly selected variants varying from 10k to 1000k from CF 2k samples, and testing on independent CF 3k samples. We repeated 5 times for each number of variants and evaluated MagicalRsq and Rsq performance using **(A)** RMSE and **(B)** MAE.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

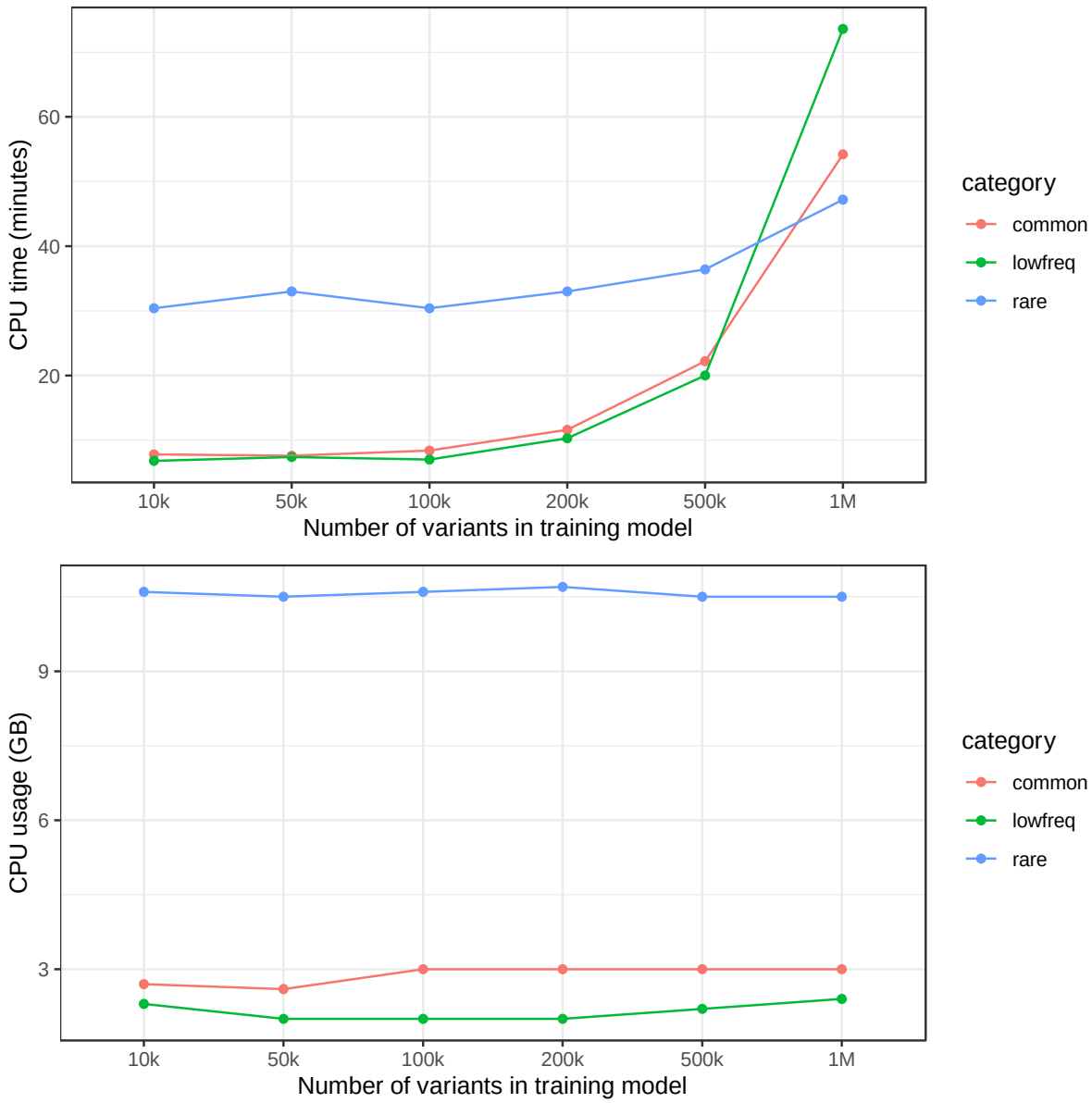

Figure S13. CPU time in minutes and memory usage in GB for MagicalRsq model training with different numbers of variants, separately for three MAF categories.

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