Rapid and accurate multi-phenotype imputation for millions of individuals

Lin-Lin Gu[1], Guo-Bo Chen[2,3], Hong-Shan Wu[4], Yong-Jie Zhang[1], Jing-Cheng He[5], Xiao-Lei Liu[6], Zhi-Yong Wang[1], Dan Jiang[1] and Ming Fang[1*]

[1] Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of Agriculture and Rural Affairs & Fisheries college, Jimei University, Xiamen, Fujian, People's Republic of China

[2] Center for General Practice Medicine, Department of General Practice Medicine, Clinical Research Institute, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, Zhejiang, People's Republic of China

[3] Key Laboratory of Endocrine Gland Diseases of Zhejiang Province, Hangzhou, Zhejiang, People's Republic of China

[4] Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China

[5] Center for Data Science, School of Mathematical Sciences, Zhejiang University, Hangzhou, Zhejiang, People's Republic of China

[6] Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education & College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, Hubei, People's Republic of China

***Correspondence:**

Ming Fang: Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of Agriculture and Rural Affairs & Fisheries college, Jimei University, Xiamen, Fujian, People's Republic of China. Email: fangming618@126.com

Abstract

Genetic analysis using big data can enhance the power of GWAS, but large data sets often have a large number of missing phenotypes. The UK Biobank database contains ~500,000 individuals with ~3,000 phenotypes, with phenotype missing rates ranging from 0.11% to 98.35%. Imputation of missing phenotypes is an important way of improving the GWAS power. The multi-phenotype imputation method can significantly improve the accuracy of imputation. However, most existing multi-phenotype imputation methods are unable to impute missing phenotypes of millions of individuals, for example, PHENIX (*Nature Genetics* 2016(48):466–472) will require months of time and ~1T of computer memory. We herein developed a Mixed Fast Random Forest (MFRF) based machine learning for phenotypic imputation. Our simulation results showed that the imputation accuracy of MFRF was higher than or equal to that of existing state-of-the-art methods; MFRF was also extremely computationally fast and memory efficient, using only 0.23-0.54 h and 68.32-126.35 Mb of computer memory for the UK Biobank dataset. We applied MFRF to impute 425 phenotypes from the UK Biobank dataset, and conducted the GWA studies using the imputed phenotypes. Compared with the GWAS before phenotype imputation, 1355 (15.6%) extra GWAS loci were identified.

Key words: phenotype imputation, random forest, UK Biobank, non-linear effect

Introduction

Large data analysis can significantly improve the GWAS power[1]. In the last two decades, several large-scale datasets have emerged. UK Biobank (UKB) database has collected nearly 500,000 people with about 3,000 phenotypes recorded[2]. Taiwan Biobank (TWB) includes 102,900 East-Asian individuals with 36 quantitative phenotypes recorded[3]. With the reduction of the cost of high-throughput genotyping, it is predicted that the datasets will be further expanded rapidly in future. The availability of large datasets provides valuable resources for genetic analysis, but it often has a large number of missing phenotypes. For example, in the UKB dataset, the phenotypic missing rate is ranged from 0.11% to 98.35%. Samples with missing phenotypes cannot be used for GWAS, which leads to a decrease in sample size and thus reduces GWAS power[4]. If missing phenotypes can be accurately imputed, more extra samples will be utilized for GWAS.

Multiple measured traits often have a certain degree of correlation, which provides information for multiple phenotype imputation. Van Buuren and Oudshoorn proposed a kind of multi-phenotype imputation method called MICE[5], which uses chained equations to handle missing data. MICE is particularly fast and requires very little computer memory, making it suitable for large datasets. Stekhoven et al. developed missForest[6], a multi-phenotype imputation method based on random forest. MissForest is more accurate than MICE[6], but it consumes a lot of computer memory and computing time[7], so it is not suitable for large dataset[8]. Dahl and Iotchkova proposed a PHENIX[4] method for multi-phenotype imputation. PHENIX is based on a multivariant normal distribution, which uses not only genetic and residual correlations between traits, but also kinships between individuals. PHENIX is much more accurate than MICE and missForest, but the manipulation of the kinship matrix makes it consume a lot of memory and time and cannot be applied to large data sets. It is predicted that PHENIX will take several months and ~1T computer memory when 500,000 individuals with 30 traits are imputed, which will drain the computing resources of many institutes.

In this study, we propose a hybrid fast random forest (MFRF) method for phenotype imputation. MFRF is a multi-phenotype imputation method that can model nonlinear and linear effects, allowing for higher-order interactions between predictive factors, thus avoiding biased predications, and making its imputation more accurate. MFRF also employs some excellent algorithms to improve the computational speed and reduce the computational memory. (1) MFRF uses Wright and Ziegler's method[9] to

construct random trees that was optimized for high-dimensional data in terms of extensive runtime and memory profiling, making it extremely computational fast and saving a lot of computer memory; (2) MFRF uses the Knuth[10] (p. 137) algorithm for feature sampling without replacement, which quickly divides candidate features in each node, overcoming the problem of computational overload when many features imputation; and (3) MFRF stores node information in a simple data structure, avoiding repeated accesses to the raw data and releasing memory early for efficient storage (ranger: https://github.com/imbs-hl/ranger).

To test the performance of MFRF, we systematically compare MFRF with the existing state-of-art methods, LMM[11–13], MICE[5], missForest[6] and PHENIX[4]. We show that MFRF and PHENIX are much accurate than other methods, and MFRF can easily impute missing phenotypes in large datasets. We also demonstrate that using the phenotype after imputation with MFRF enables the identification of more GWAS loci and functional genes than before imputation.

Results

Simulations

We first investigated the performance of MFRF through extensive computer simulations. The multi-trait mixed model was employed to simulate phenotypic values, which has been developed and applied in PHENIX[4]. The simulated phenotypic values consisted of genetic effects and random residual effects, and we also specified the genetic and residual correlations between phenotypes to ensure phenotypic correlation (**Online Methods**).

Imputation accuracy

We performed four simulations to investigate the performance of MFRF under different sample size, heritability, phenotype number and relatedness between individuals, respectively. In order to investigate the effect of a specific factor, we simulated different data sets by varying a specific factor with others unchanged. To do this, we set a set of default parameters, the sample size 500, the heritability 0.6, the number of phenotypes 60, the coefficient of relationship between individuals 0.02, and the phenotypic missing rate 5%.

In simulation 1, the sample size was set as 100, 300, 500, 700, 900 and 1100, respectively, with other parameters set as default. MFRF was compared with four state-

of-the-art methods, PHENIX, missForest, MICE, and LMM (**Supplementary Table 1** and **Online methods**). In order to effectively compare the performance of each method, the phenotypes were the same for each replication of the simulation when comparing different methods. The simulations were repeated 100 times, and the average correlation coefficient between the imputed and the true phenotypes in 100 simulations was used to evaluate the imputation accuracy. When the sample size was greater than 300, the imputation accuracy of MFRF was the highest among these methods, followed by PHENIX, and both had higher imputation accuracy than missForest, MICE, and LMM. The imputation accuracies of all methods increased with the sample size. (**Figure 1a** and **Supplementary Table 2)**. In simulation 2, the heritability was set so 0.1, 0.2, 0.4, 0.6, 0.8 and 1, respectively. The imputation accuracies of MFRF were slightly higher than those of PHENIX, both of which were higher than MICE, missForest and LMM (**Figure 1b** and **Supplementary Table 3)**, and the imputation accuracy first increased and then decreased with the heritability. In simulation 3, the phenotype number was set as 20, 40, 60, 80, 100 and 120, respectively. MFRF and PHENIX had the highest imputation accuracies, both were more accuracy than missForest, MICE and LMM (**Figure 1c** and **Supplementary Table 4**), and as the phenotype number increased, the imputation accuracies also increased. In simulation 4, the coefficient of relationship was set as 0.02, 0.2, 0.4, 0.6, 0.8 and 0.9, respectively, the imputation accuracies of MFRF and PHENIX were very close and both were much more accurate MICE, missForest and LMM, and the imputation accuracies increased as the sample relatedness increased (**Figure 1d** and **Supplementary Table 5**).

Computational time and computer memory

We then used large sample to evaluate the computational speed and the computer memory usage of different methods. We randomly sampled 1,000, 5,000, 10,000, 15,000 and 20,000 individuals with 30 phenotypes from UKB dataset[2]. We only compared MFRF and PHENIX because the two methods have higher imputation accuracy. PHENIX required 0.06 to 7.38 hours to complete the task as the sample size increased from 1,000 to 20,000. The computational time increased rapidly as the sample size increased, indicating that larger sample will require much more computation time (**Figure 2a** and **Supplementary Table 6**). In contrast, MFRF only needed 0.04 to 0.29 hours as the sample size increased from 1,000 to 20,000 (**Figure 2a**), saving 0.5 to 24.45 times of the computational time compared to PHENIX, and the computation time of MFRF increased linearly with sample size roughly, indicating that larger sample will

not take too much time. We then investigated the computer memory usage of both methods. When the sample size was increased from 10,000 to 20,000, PHENIX required 3627.67 to 8052.91 Mb of computer memory, while MFRF required only 2.60 to 5.12 Mb of computer memory (**Figure 2b** and **Supplementary Table 6**), saving 1394.26 to 1571.83 times of the computer memory compared to PHENIX.

We next studied the effect of the number of phenotypes on the computing speed and computer memory usage. We fixed the sample size as 10,000 and varied the phenotype numbers to 10, 15, 20, 25 and 30, respectively. PHENIX required 1.89 to 2.02 hours and 815.53 to 627.68 Mb of computer memory, while MFRF required only 0.08 to 0.15 hours and 2.60 to 3.88Mb of memory, saving 12.47 to 22.63 times of the computing time and 312.67 to 933.97 times of the computer memory, respectively (**Figure 2c**, **Figure 2d** and **Supplementary Table 7**).

UK Biobank datasets

Imputation

In order to reduce the impact of population stratification, we extracted British citizens for research. After quality control, there were 277,301 individuals of British ancestry and 563,675 genotyped variants available for downstream analyses. We initially retained about 800 quantitative and binary disease phenotypes from ~3000 phenotypes (removing categorical phenotypes and records such as text and time). We further removed phenotypes that have missing rates greater than 95% and are unlikely to be genetically controlled, such as “number of antibiotics taken”. Finally, 425 phenotypes were imputed, including 387 quantitative phenotypes and 38 binary disease phenotypes (**Supplementary file 1**). The averaged missing rate of the 425 phenotypes was 0.213 (0~0.948) (**Supplementary file 1**). To reduce the impact of cofactors on imputation, we first corrected for the effects of the first 20 principal components, gender, and age on observable phenotypes using a linear model and applied residuals for imputation. In the imputation of each phenotype, we first selected several reference phenotypes from all phenotypes. The reference phenotypes will provide correlation information for imputation a phenotype. The reference phenotypes should have a correlation coefficient greater than 0.3 with the imputed phenotype; and if there was no phenotype with a correlation coefficient greater than 0.3, the 29 most correlated phenotypes were selected as reference phenotypes. The number of reference phenotypes for each of the 425 phenotypes ranged from 2 to 29, and the correlation coefficients of the selected

reference phenotypes ranged from 0.06 to 0.99.

We only applied MFRF and MICE to the imputation of the UK Biobank dataset as other methods are not suitable for large datasets. We randomly selected 0.29% (800 individual phenotypes) of the complete phenotype as missing, and used the correlation coefficient between the imputed and true values of the selected "missing" phenotypes to assess imputation accuracy. Because the proportion of the missing phenotypes was very small, setting them as missing would have negligible effect on imputation accuracy.

For almost all 425 phenotypes, the imputation accuracies of MFRF were obviously higher than those of MICE (**Figure 3a** and **3b**). The average imputation accuracy of 425 phenotypes using MICE was 0.645, while the average accuracy of MFRF was 0.766, higher than that of MICE. The imputation accuracy varies greatly among phenotypes, from 0.016 to 0.999. The reference phenotypes provided correlation information for imputation, so it might affect the imputation accuracy. **Figure 3a** shows the correlation coefficient (the largest one) between the imputed phenotype and its reference phenotypes. Most phenotypes with higher imputation accuracies had highly correlated reference phenotypes and the accuracy of the imputation increased with the correlation coefficient between the imputed and reference phenotypes (**Figure 3a)**. These indicates that the highly correlated reference phenotypes play an important role in the accurate imputation. We next investigated the effect of missing rate on imputation accuracy. The missing rate has little effect on the imputation accuracy (**Figure 3b**), specifically, the missing rate of "Diastolic brachial blood pressure during PWA" (**ID 12675**) is 92.11%, but the imputation accuracy of MFRF still reached 0.947. We further performed regression analysis on the missing rates and the imputation accuracies for 425 phenotypes. There seemed to be no obvious linear relationship between the two (p=0.0039, **Supplementary Figure 1**). The effect of the missing rate on the imputation accuracy is not very severe, possibly because the data set used is large, and although some phenotypes are missing, there are still a large number of individual phenotypes used for imputation.

Accurately imputed phenotypes could be used in the GWAS analysis, so we filtered the imputed phenotypes using accuracy thresholds of 0.5 and 0.7, respectively. When imputed using MFRF, the imputation accuracy of 350 and 294 phenotypes out of 425 was greater than 0.5 and 0.7, respectively (**Figure 3c**); when imputed with MICE, the imputation accuracy of 289 and 208 phenotypes was greater than 0.5 and 0.7, respectively (**Figure 3c**), further indicating that MFRF outperformed MICE.

In addition to imputation accuracy, the bias of the imputation also has a significant impact on GWAS. The imputation bias was defined as the regression coefficient between the imputed and the true phenotypes (**Online Methods**). **Figure 3d** shows that the imputation bias of both MFRF and MICE for 425 phenotypes is close to 0. The average bias of 425 phenotypes of MFRF and MICE was 0.0003 (*SD* = 0.0343) and 0.0007 (*SD* = 0.0247), respectively, both of which were very close to zero and not significantly different from zero (*p*-values of single sample *t*-tests were 0.868 and 0.583, respectively), suggesting both methods provide unbiased imputations.

GWAS

After imputation with MFRF, we also performed GWAS with the imputed phenotypes. To ensure the reliability of GWAS, we only retained phenotypes with imputation accuracy greater than 0.5 or 0.7, leaving 350 and 294 phenotypes, respectively (**Supplementary file 2**). For each phenotype, we further removed individuals with poor imputation, as for these individuals, most of their reference phenotypes were also missing, which greatly reduced the imputed information and thus the accuracy of the imputation. We defined a sample score (*SC*) (**Online Methods**) for individual level filtering, the threshold of *SC* was set to 0.6. We then used PLINK software (https://www.cog-genomics.org/plink2, v1.90 beta) for GWAS with the filtered phenotypes. To reduce potential false positive findings caused by uncontrollable factors such as insufficient covariates, we conducted a genomic control by dividing the test statistic of GWAS by the genomic inflation factor λ [14].

We first take two examples to show the effect of the imputation on GWAS. The first example is heart rate phenotype (ID: 5983), with a missing rate of 84.92% and an MFRF imputation accuracy of 0.772. Before imputation, 6 independent GWAS loci (>1Mb from each other) were identified **(Figure 4a)**; after imputation, the GWAS signals were significantly enhanced **(Figure 4b)**. In addition to the 7 loci found before imputation, another 31 independent loci were identified (**Supplementary Table 8**). Within these extra loci, many functional important genes had been identified. The lead variant of locus chr6:7,043,468-8,043,468 is located within *DSP*, which has been reported to strongly associate with dilated cardiomyopathy[15,16], the genotype-imaging phenotype study confirmed that *DSP* was associated with dilated cardiomyopathy and arrhythmogenic left ventricular cardiomyopathy[17]. The lead variant of locus chr16:64,779,257-65,779,257 is 119.242 kb away from *CDH11*, which has been reported to play an important role in development of heart valves and pathogenic

mechanisms[18,19]. The lead SNP rs272570 is within *RNF220*, *RNF220* is related to heart development and heart disease[20,21]. The lead SNP in chr3:38,267,315-39,267,315 is within *SCN10A*, which is involved in cardiac rhythm regulation[22,23]. The lead SNP in chr16:15,418,096-16,418,096 is within *MYH11*, which is a regulator and components of the cytoskeleton, and also is associated with most known inherited aortic diseases[24]. The lead SNP in chr14:72,385,471-73,385,471 is within *RGS6*, *RGS6* is involved in the regulation of the cardiac parasympathetic nervous system[25,26] (**Supplementary Table 8**). The next example is cancer phenotype (ID: 2453), the missing rate of which is 0.28% and the imputation accuracy is 0.818. Before imputation, two significant loci on chromosomes 10 and 16 were discovered, respectively (**Figure 4c)**; after imputation the GWAS signals were enhanced, in addition to the two loci discovered before imputation, two extra loci on chromosomes 2 and 11 were identified, respectively (**Figure 4d**). A lead SNP of chromosome 2 is located within *CASP8*, which is associated with breast cancer and melanoma[27,28]; a lead SNP within the chromosome 11 locus is located within *TYR*, which is associated with skin cancer and melanoma[29,30].

Next, we conducted GWAS for all the investigated UKB phenotypes. We first performed a GWAS on 350 phenotypes with an imputation accuracy greater than 0.5. Obviously, for almost all 350 phenotypes more GWAS loci could be identified after imputation than before imputation (**Figure 5a**). A total of 7,355 independent GWAS loci were identified before imputation, while 8,710 independent loci were found after imputation, an increase of 15.6%. No more GWAS loci were identified before imputation than after imputation (**Figure 5c**). Similarly, we also performed GWAS for 294 phenotypes with an imputation accuracy greater than 0.7. A total of 6,167 GWAS loci were identified before imputation, and 576 (8.5%) extra loci were identified after imputation (**Figure 5d**). No more GWAS loci were identified before imputation than after imputation (**Figure 5d**). We have detailed the GWAS results in **Supplementary file 3**.

The number of extra identified GWAS loci after imputation than before imputation varied greatly, ranging from 0 to 209, which largely depends on the missing rate (**Figure 5a** and 5**b**). For some phenotypes with very high missing rates (red, green, and yellow dots in **Figures 5a** and **5b**), much more extra GWAS loci were identified after imputation than before; however, for phenotypes with low missing rates (blue dots in **Figures 5a** and **5b**), the number of GWAS loci identified after imputation was very close to before imputation (**Figures 5a** and **5b**). This suggests that it makes more sense

to impute phenotypes with high missing rates.

Discussions

We have developed a new phenotype imputation method MFRF. MFRF has significant advantages in computing speed and computer memory usage, making it very suitable for big data imputation. MFRF is an imputation technique based on a hybrid fast random forest method (RF), it has been shown that RF is able to handle variables that are associated in a non-linear fashion[8]. The random tree constructed by RF has a significant impact on the speed of computation and memory consumption. MissForest[6] method implements random forest by Liaw and Wiener's method[31], which requires repeated access to the original data, making it requires a lot of time and memory. In contrast, MFRF uses the Wright and Ziegler's method[9] to construct random trees that is optimized for high-dimensional data in terms of extensive runtime and memory profiling, and the most critical part of the method is node splitting, where all *mtry* candidate features are evaluated as split candidates. In model construction, MFRF model can consider both linear and nonlinear effects between phenotypes, which is the key to high imputation accuracy of MFRF.

Like many multi-phenotype imputation methods, the imputation accuracy of MFRF is also directly affected by the correlations between phenotypes (**Figure 3a**). In UK Biobank data, some phenotypes have strong measurement error correlations. For example, blood biochemical indicators such as the number of white blood cell and red blood cell and so on, which are measured by the same instrument, making them highly correlated with measurement errors. Some phenotypes also have strong genetic correlations, such as heart rate and pulse rate, body height and weight, and so on[32]. Genetic and error correlation can lead to a certain degree of phenotypic correlation, and combining these phenotypes for imputation can improve the accuracy of phenotypic imputation.

The imputation accuracy of the MFRF also directly depends on the sample relatedness, as the sample correlation increases, the imputation accuracy of MFRF also increases. PHENIX dissects phenotypic (co)variance into genetic and environmental (co)variance, utilizing phenotypic correlation through the kinship matrix and error correlation matrix, while MFRF directly uses phenotypic correlation for imputation, although MFRF does not dissect it further but also indirectly utilizes genetic correlation and phenotypic correlation, so its imputation accuracy also increases with sample correlation.

We filtered the phenotypes according to the imputation accuracy in the GWAS analysis, but such filtering still could not meet the requirements of GWAS. In order to reduce the false positive rate of GWAS, we also carried out phenotypic filtering at the individual level. In fact, even for the same phenotype, there are large differences in the accuracy of phenotypic imputation among different individuals. The accuracy of the imputation of an individual phenotype depends primarily on the completeness of their reference phenotypes, especially those that are highly correlated with the imputed phenotype. In this study, we defined a *SC* to measure the contribution of the available reference phenotypes for each individual, and using *SC* for individual level phenotype filtering. This individual-level quality control is critical, and using individuals with poor imputation accuracy will reduce the power of the GWAS. We set the threshold of *SC* as 0.6, which is conservative for quality control. We randomly selected 42 imputed phenotypes and filtered individual phenotypes with *SC* thresholds 0, 0.2, 0.4, 0.6, 0.8, 1, and performed the GWAS with the filtered phenotypes respectively. We found that the number of extra GWAS loci identified before than after imputation decreased with increasing *SC* thresholds, and when the *SC* threshold was greater than or equal to 0.6, no extra GWAS loci were identified. The total number of GWAS loci after imputation first increased and then decreased with the *SC* threshold, and when the *SC* was equal to 0.6, the number of GWAS loci after imputation was the largest (**Supplementary Figure 2**). These results suggest that the *SC* threshold of 0.6 is reasonable.

Large number of binary phenotypes are recorded in the UK Biobank dataset in addition to continuous phenotypes. In the study, we directly take the binary phenotype as a continuous phenotype for imputation, such method has been widely accepted in statistic genetics[33,34], and it usually performs well in practice. However, MFRF can also specifically handle binary phenotypes, the core of MFRF is the random forest, which can solve the continuous and binary phenotypes using regression trees and classification trees, respectively. Using classification tree for imputing binary phenotype could improve the imputation performance and is worthy of further study.

MFRF has many advantages. Firstly, it does not require genotype data making it very flexible to use; secondly, it consumes very little computer memory; thirdly, it has extremely fast computational speed; and finally, it is highly accurate in imputation. Based on these advantages, MFRF is expected to be widely used for imputation in large data sets.

URLs.

PLINK, https://www.cog-genomics.org/plink2;
UK Biobank, http://www.ukbiobank.ac.uk;
PHENIX, https://mathgen.stats.ox.ac.uk/genetics_software/phenix/phenix.html;
MFRF, https://github.com/GuLinLin-JMU/MFRF.

Acknowledgements

We thank the participants of UK Biobank for making this work possible (under application 41376). The research was supported by the National Key Research & Development Program of China (2018YFD0901201 and 2022YFD2401001), National Natural Science Foundation of China (31771392), the Natural Science Foundation of Fujian Province of China (2020J01672) and the National Marine Fisheries Industrial Technology System Post Scientist Project (CARS-47-G04).

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential.

Author contributions

GLL developed the method, wrote the program and conduct the whole analysis. CGB contributed to real dataset analysis and conceived the imputation quality control method. WHS and ZYJ assisted in the data collection and helped part of the data preparation. CGB and LXL discussed results and methods, and provided comments that improved earlier versions of the manuscript. Other authors contributed to improving the study and manuscript. The manuscript was drafted by FM, GLL and JD. FM designed the study. All authors read and approved the final manuscript.

Code availability

The software of MFRF method can be accessed from https://github.com/GuLinLin-JMU/MFRF.

References

1. Marx, V. Human phenotyping on a population scale. *Nat Methods* **12**, 711–714 (2015).
2. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic

data. *Nature* **562**, 203–209 (2018).

3. Chen, C.-H. *et al.* Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum. Mol. Genet.* **25**, 5321–5331 (2016).
4. Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat Genet* **48**, 466–472 (2016).
5. Buuren, S. van & Groothuis-Oudshoorn, K. mice : Multivariate Imputation by Chained Equations in R. *J. Stat. Soft.* **45**, 1–67 (2011).
6. Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
7. Jin, H., Jung, S. & Won, S. missForest with feature selection using binary particle swarm optimization improves the imputation accuracy of continuous data. *Genes Genomics* **44**, 651–658 (2020).
8. Schwarz, D. F., König, I. R. & Ziegler, A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* **26**, 1752–1758 (2010).
9. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Soft.* **77**, 1–17 (2017).
10. Knuth, D. E. *The art of The Art of Computer Programming, volume 2*. (Addison-Wesley, Reading, 1985).
11. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821–824 (2012).

416 12. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for
417 genome-wide association studies. *Nat Methods* **11**, 407–409 (2014).
418 13. Runcie, D. E. & Crawford, L. Fast and flexible linear mixed models for genome-
419 wide genetics. *PLoS Genet* **15**, e1007978 (2019).
420 14. Devlin, B., Roeder, K. & Wasserman, L. Genomic Control, a New Approach to
421 Genetic-Based Association Studies. *Theor Popul Biol* **60**, 155–166 (2001).
422 15. Jordan, E. *et al.* Evidence-Based Assessment of Genes in Dilated Cardiomyopathy.
423 *Circulation* **144**, 7–19 (2021).
424 16. Mazzarotto, F. *et al.* Reevaluating the Genetic Contribution of Monogenic Dilated
425 Cardiomyopathy. *Circulation* **141**, 387–398 (2020).
426 17. Augusto, J. B. *et al.* Dilated cardiomyopathy and arrhythmogenic left ventricular
427 cardiomyopathy: a comprehensive genotype-imaging phenotype study. *Eur Heart*
428 *J Cardiovasc Imaging* **21**, 326–336 (2020).
429 18. Schroer, A. K. *et al.* Cadherin-11 blockade reduces inflammation-driven fibrotic
430 remodeling and improves outcomes after myocardial infarction. *JCI Insight* **4**,
431 e131545 (2019).
432 19. Bowler, M. A. *et al.* Cadherin-11 as a regulator of valve myofibroblast
433 mechanobiology. *Am J Physiol Heart Circ Physiol* **315**, H1614–H1626 (2018).
434 20. Ramírez, J. *et al.* Thirty loci identified for heart rate response to exercise and
435 recovery implicate autonomic nervous system. *Nat Commun* **9**, 1947 (2018).
436 21. Global BPgen Consortium *et al.* Identification of heart rate–associated loci and
437 their effects on cardiac conduction and rhythm disorders. *Nat Genet* **45**, 621–631

(2013).

22. Bezzina, C. R. *et al.* Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. *Nat Genet* **45**, 1044–1049 (2013).

23. van de Vegte, Y. J., Tegegne, B. S., Verweij, N., Snieder, H. & van der Harst, P. Genetics and the heart rate response to exercise. *Cell. Mol. Life Sci.* **76**, 2391–2409 (2019).

24. Verstraeten, A., Luyckx, I. & Loeys, B. Aetiology and management of hereditary aortopathy. *Nat Rev Cardiol* **14**, 197–208 (2017).

25. Yang, J. *et al.* RGS6, a Modulator of Parasympathetic Activation in Heart. *Circ Res* **107**, 1345–1349 (2010).

26. Wydeven, N., Posokhova, E., Xia, Z., Martemyanov, K. A. & Wickman, K. RGS6, but Not RGS4, Is the Dominant Regulator of G Protein Signaling (RGS) Modulator of the Parasympathetic Regulation of Mouse Heart Rate. *J Biol Chem* **289**, 2440–2449 (2014).

27. Barrett, J. H. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. *Nat Genet* **43**, 1108–1113 (2011).

28. Park, S. *et al.* Promotion of tumorigenesis by miR-1260b–targeting *CASP8*: Potential diagnostic and prognostic marker for breast cancer. *Cancer Science* **113**, 2097–2108 (2022).

29. Zhang, M. *et al.* Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans.

Human Molecular Genetics **22**, 2948–2959 (2013).

30. Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet* **41**, 920–925 (2009).
31. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
32. Tyrrell, J. *et al.* Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *BMJ* **352**, i582 (2016).
33. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348–354 (2010).
34. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**, 825–830 (2012).
35. Wang, J. *et al.* Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *Am. J. Hum. Genet.* **98**, 697–708 (2016).
36. Schafer, J. L. & Yucel, R. M. Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *J Comput Graph Stat* **11**, 437–457 (2002).
37. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–569 (2010).
38. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).
39. Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J. & Vilhjálmsson, B. J. Effcient toolkit implementing best practices for principal component analysis of population

genetic data. *Bioinformatics* **36**, 4449–4457 (2020).

40. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).

Online methods

Mixed Fast Random Forest. MFRF is a method based on the construction of Mixed Fast Random Forest, which is characterized by the fit of both non-linear and linear effects in the model. Assuming that there are p phenotypes imputed at the same time, for s^{th} phenotype ($s = 1,2,\dots,p$), the model can be expressed as

$$\mathbf{X}^{(s)} = \boldsymbol{u}^{(s)} + \mathbf{X}^{(-s)}\boldsymbol{b} + \boldsymbol{e} \tag{1}$$

where $\mathbf{X}^{(s)}$ denotes the phenotypic values of the sth phenotype, $\mathbf{X}^{(-s)}$ denotes the phenotypic values of other $p-1$ phenotypes, and $\boldsymbol{u}^{(s)}$ is the non-linear effect caused by $p-1$ phenotypes, $\boldsymbol{b}$ is the linear effect of $p-1$ phenotypes; $\boldsymbol{e}$ denotes the random error, which follows a normal distribution.

These phenotypes are first sorted according to missing rate from small to large, then updated in order. For s^{th} phenotypes, the observed phenotypic values are denoted by $\mathbf{X}_{obs}^{(s)}$ and the missing phenotypic values are denoted by $\mathbf{X}_{mis}^{(s)}$; and similarly, we denote $\mathbf{X}_{obs}^{(-\mathrm{s})}$ as p-1 reference phenotypes for individuals with observable values for s^{th} phenotype; and $\mathbf{X}_{mis}^{(-\mathrm{s})}$ as p-1 reference phenotypes for individuals with unobservable values for s^{th} phenotype. Then, the observable values of sth phenotype can be expressed as

$$\mathbf{X}_{obs}^{(s)} = \boldsymbol{u}_{obs}^{(s)} + \mathbf{X}_{obs}^{(-s)}\boldsymbol{b} + \boldsymbol{e} \tag{2}$$

which is used to train the model and estimate model parameters. Once we have the parameter estimated, we are going to predict $\mathbf{X}_{mis}^{(s)}$ with $\mathbf{X}_{mis}^{(-s)}$.

(1) *Initialization*. We first initialize the model parameters with $\boldsymbol{b} = 0$ and $\boldsymbol{u}^{(s)} = \mathbf{X}^{(s)}$, and $\mathbf{X}^{(-s)}$ is initialized using univariate imputation by sampling with replacement from the non-missing values.

(2) *Update of linear-effect parameters*. In the tth iteration, let

$$\boldsymbol{\gamma}_{obs(t)}^{(s)} = \mathbf{X}_{obs}^{(-s)}\boldsymbol{b} + \boldsymbol{e} \tag{3}$$

representing the phenotypic value attributed to linear effect and random error, where $\boldsymbol{\gamma}_{obs(t)}^{(s)}$ is updated with $\boldsymbol{\gamma}_{obs(t)}^{(s)} = \mathbf{X}_{obs}^{(s)} - \hat{\boldsymbol{u}}_{obs(t-1)}^{(s)}$ according to equation (2) using the latest update of $\boldsymbol{u}_{obs}^{(s)}$; then $\boldsymbol{b}$ can be estimated with a linear regression of equation (3), and then the phenotypic value attributed to linear effect, $\mathbf{X}_{obs}^{(-s)}\boldsymbol{b}$, can be updated with

$$\hat{\boldsymbol{\gamma}}_{obs(t)}^{(s)} = \mathbf{X}_{obs}^{(-s)}\hat{\boldsymbol{b}} \tag{4}$$

(3) *Update of non-linear effect parameters*. Let

$$\mathbf{g}_{obs(t)}^{(s)} = \boldsymbol{u}_{obs(t)}^{(s)} + \boldsymbol{e} \tag{5}$$

representing the phenotypic value attributed to non-linear effect and random error,

which can be rewritten as

$$\mathbf{g}_{obs(t)}^{(s)} = \mathbf{X}_{obs}^{(s)} - \mathbf{X}_{obs}^{(-s)}\boldsymbol{b} \tag{6}$$

according to equation (2), and then we get the update of $\mathbf{g}_{obs(t)}^{(s)}$ with

$$\mathbf{g}_{obs(t)}^{(s)} = \mathbf{X}_{obs}^{(s)} - \widehat{\boldsymbol{\gamma}}_{obs(t)}^{(s)} \tag{7}$$

by substituting equation (4) into equation (6). Next, we are going to update the non-linear effect $\boldsymbol{u}_{obs(t)}^{(s)}$ in equation (5). Let

$$\boldsymbol{u}_{obs(t)}^{(s)} = f\left(\mathbf{X}_{obs}^{(-s)}\right) \tag{8}$$

where $f\left(\mathbf{X}_{obs}^{(-s)}\right)$ is a non-linear fast random forest[35] function about $\boldsymbol{u}_{obs(t)}^{(s)}$, reflecting the complex non-linear relationship between phenotypes of $p-1$ phenotypes contributed to s^{th} phenotype; then equation (5) can be written as

$$\mathbf{g}_{obs(t)}^{(s)} = f\left(\mathbf{X}_{obs}^{(-s)}\right) + \boldsymbol{e} \tag{9}$$

where $\mathbf{g}_{obs(t)}^{(s)}$ has been known from equation (7), then the parameters involved in fast random forest function at iteration *t* can be estimated by solving the random forest non-linear regression with equation (9); with the parameters estimated, we obtain the update of $\boldsymbol{u}_{obs(t)}^{(s)}$ with equation (8).

By iterating step (2)-(3) until the change of the likelihood reaches the given criteria (< 0.001). In practical applications, the model is usually converged in few iterations and not sensitive to the initial values. Once parameters related to linear and non-linear effect have been estimated with observable data $\mathbf{X}_{obs}$, we can impute the missing phenotype value $\mathbf{X}_{mis}^{(s)}$ with the estimated parameters and its reference phenotypes $\mathbf{X}_{mis}^{(-s)}$. MFRF first estimates the missing phenotype value attributed to linear effect $\widehat{\boldsymbol{\gamma}}_{(mis)}^{(s)}$ and the non-linear effect $\widehat{\boldsymbol{u}}_{mis}^{(s)}$ using equation (4) and equation (8), respectively, by substitute $\mathbf{X}_{obs}^{(-s)}$ with $\mathbf{X}_{mis}^{(-s)}$, and then get the estimates of $\mathbf{X}_{mis}^{(s)}$ with

$$\mathbf{X}_{mis}^{(s)} = \widehat{\boldsymbol{u}}_{mis}^{(s)} + \widehat{\boldsymbol{\gamma}}_{(mis)}^{(s)} \tag{10}$$

The flow chart of the method is shown in **Figure 6**.

Similarly, MFRF updates missing values for all phenotypes sequentially, for $s = 1,2,\ldots,p$ using the most recently updated parameters whenever possible. In order to obtain the optimal predictions, the iterations continue for several times until the difference between the missing phenotypes of two consecutive iterations meets a certain criterion ($\theta < 0.001$), where θ is defined as

$$\theta = \frac{\sum(\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp})^2}{\sum(\mathbf{X}_{new}^{imp})^2} \tag{11}$$

Imputation accuracy. We assessed the performance with correlation coefficient r between the imputed and true phenotypic values,

$$r = \frac{\sum(\mathbf{X}^{true} - \bar{\mathbf{X}}^{true})(\mathbf{X}^{imp} - \bar{\mathbf{X}}^{imp})}{\sqrt{\sum(\mathbf{X}^{true} - \bar{\mathbf{X}}^{true})^2 \sum(\mathbf{X}^{imp} - \bar{\mathbf{X}}^{imp})^2}} \quad (12)$$

where $\mathbf{X}^{true}$ is the true phenotypes and $\mathbf{X}^{imp}$ is imputed phenotypes. The correlation coefficient was calculated with the function "cor.test" in R programming language (https://www.r-project.org).

Sample score. For an individual with a missing phenotype, if some of the reference phenotypes associated with it are also missing, the amount of information available will be reduced and that phenotype will be less accurately imputed. To reduce the false positive rate of GWAS analysis, we defined a sample score (*SC*) for each individual and filtered the imputed phenotypes at the individual level according to the *SC*. For each individual imputed phenotype, the *SC* was defined as the ratio of the sum of the squares of the correlation coefficients between the observable reference phenotypes and the imputed phenotype, to the sum of the squares of the correlation coefficients between all the reference phenotypes and the imputed phenotype. For the sth missing phenotype of individual i, assuming that there are p reference phenotypes, of which p' phenotypes are complete and the others are missing, the *SC* is defined as

$$SC_{i,s} = \frac{\sum_{j \neq s}^{p'} r_{i,j}^2}{\sum_{j \neq s}^{p} r_{i,j}^2} \quad (13)$$

where, $r_{i,j}$ is the correlation coefficient between j^{th} (for j=1,2,$\cdots$,p) reference phenotype and the imputed phenotype of individual i.

Imputation bias. We selected an additional 0.29% of phenotypes as missing (800 samples) beside true missing phenotypes for each of phenotypes, and then evaluated the imputation bias using the regression coefficient between the imputed and true phenotypes of the 0.29% missing phenotypes, where the regression analysis was implemented via the function "lm" in R package.

PHENIX, missForest, MICE and LMM. We applied several other popular methods for imputing missing phenotypes (**Supplementary Table 1**). We herein provide a brief introduction of each method.

PHENIX. The PHENIX method is expressed as

$$Y = U + \varepsilon \quad (14)$$

where, Y is a $n \times p$ matrix of p traits measured on n individuals, which is rescaled across each trait, so that the mean and variance of each trait is 0 and 1, respectively; U and ε is defined as a matrix normal distribution,

$$U \sim MN(0, B \otimes K) \tag{15}$$

$$\varepsilon \sim MN(0, E \otimes I_n) \tag{16}$$

where U and ε represent $N \times P$ matrix that column-wise restores the $NP \times 1$ matrix, respectively, K is the $n \times n$ kinship matrix between individuals, I_n is the $n \times n$ identity matrix; B and E is the $p \times p$ matrix, representing genetic and residual covariances between phenotypes, respectively. PHENIX uses the variational Bayes method to iteratively update missing phenotypes $Y^{(miss)}$ by approximating joint posterior distribution of model parameters (For details, see refs. [4]).

missForest. MissForest is a kind of mult-phenotype imputation method based on random forest method. It first makes an initial guess of the missing phenotypes with the mean/median of the corresponding phenotypes; then sorts the phenotypes according to the missing rate and imputes the missing values of each phenotype in order with a random forest method[6].

MICE. MICE is developed by Buuren and Oudshoorn[5], which is based on the multivariate multiple imputation scheme of Schafer (2002)[36]. Let the hypothetically complete data Y be a partially observed random sample from the p-variate multivariate distribution $(Y|\theta)$ with parameter θ, the MICE algorithm obtains the posterior distribution of θ by sampling iteratively from conditional distributions of

$$P(Y_s|Y_{-s}, \theta_s), \text{ for } s = 1,2, \dots, p \tag{17}$$

the model parameters are successively drawn with Gibbs sampling. Specifically, for tth iteration, they are sampled from the conditional posterior probability distribution,

$$\theta_1^{*(t)} \sim P\left(\theta_1 \middle| Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}\right) \tag{18}$$

$$Y_1^{*(t)} \sim P\left(Y_1 \middle| Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_1^{*(t)}\right) \tag{19}$$

$$\cdots\cdots$$

$$\theta_p^{*(t)} \sim P\left(\theta_p \middle| Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}\right) \tag{20}$$

$$Y_p^{*(t)} \sim P\left(Y_p \middle| Y_p^{obs}, Y_1^{(t)}, \dots, Y_p^{(t)}, \theta_p^{*(t)}\right) \tag{21}$$

where $Y_s^{(t)} = \left(Y_s^{obs}, Y_s^{*(t)}\right)$ is the sth imputed variable at iteration t.

LMM. The linear mixed model (LMM) method is a kind of single-trait method, which imputes missing phenotypes with BLUP technique. For each phenotype, LMM is defined as

$$Y \sim N(0, \sigma_g^2 K + \sigma_e^2 I_n) \tag{22}$$

where Y denotes the phenotype, σ_g^2 and σ_e^2 are genetic and residual variances, respectively; K is the kinship between individuals, which is derived from whole genome SNPs or pedigree[37].

Simulations. We simulated genetically and environmentally related phenotypic data with multiple phenotypes according to the method of Dahl and Baud[4]. The method dissects the phenotypic values into genetic and environmental effects. The phenotype was simulated with model

$$Y \sim MN(0, K, h^2\Sigma_G) + MN(0, I, (1 - h^2\Sigma_E)) \tag{23}$$

where Y is $n \times p$ the multi-trait phenotypic value that follows a multivariate normal distribution; K is the relationship matrix between individuals; h^2 is the heritability and Σ_G is the $p \times p$ the genetic correlation matrix between traits, which is sampled from Wishart distribution, $\mathrm{Wi}(I_p, p, \mathrm{df})$, where $\mathrm{df} = p$. The small degree of freedom p will produce a covariance matrix with a large variation; Σ_E is the $p \times p$ residual covariance matrix, which is also sampled from Wishart distribution, $\mathrm{Wi}(I_p, p, \mathrm{df})$ with df of p. The genetic and environmental effects are randomly sampled from the multivariate normal distribution $MN(0, K, h^2\Sigma_G)$ and $MN(0, I, (1 - h^2\Sigma_E))$, respectively, and the phenotypic values are simulated with Equation (24). The simulation is based on Dahl and Baud's script[4]. To make it suitable for simulations with a large unrelated population we modified the parameter “fam_size” to “sample size”. We then randomly selected a certain percentage of phenotypes as missing.

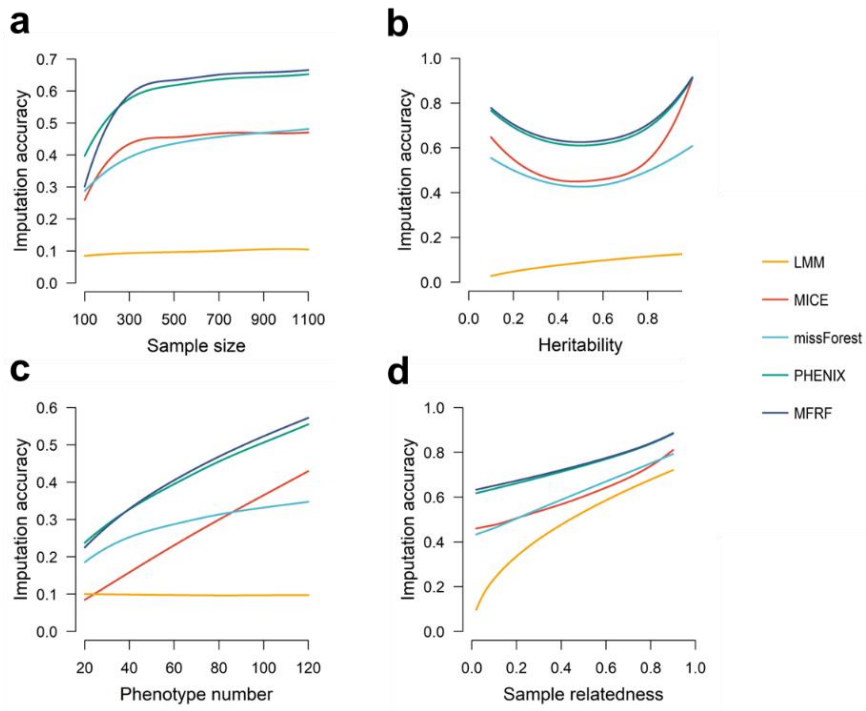
UK Biobank dataset. The UK Biobank Project is a prospective cohort study with in-depth genetic and phenotypic data collected from approximately 500,000 people across the United Kingdom, aged between 40 and 69 years at the time of recruitment. All individuals in the project were genotyped for approximately 800,000 SNPs using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix[2]. A rich variety of phenotypic and health-related information are available for each participant, including biological measurements, lifestyle indicators, biomarkers in blood and urine, and imaging of the body and brain.

Quality control. We extracted the unrelatedness individuals of British citizens for study. After removing individuals with certain kinship (**ID 22006:** Genetic ethnic grouping; **ID 22021:** Genetic kinship to other participants), 278,788 unrelated individuals were retained. Additionally, we applied standard procedures for quality control (QC),

removed individuals with missing genotypes > 5%, removed SNPs with missingness > 5%, Hardy − Weinberg equilibrium P ⩽ 1e−6 and MAF ⩽ 0.01 using PLINK[38]. We also checked the samples for sex discrepancies with parameter "--check-sex" and removed the samples marked as "PROBLEM"; we then removed the individuals with excess heterozygosity with parameter "--het", and excluded the samples with the heterozygosity exceeding three standard deviations. Finally, the dataset included 277,301 individuals of British ancestry with 563,675 common SNPs (MAF>1%).

Imputation of UK Biobank data. For each phenotype to be analyzed, we selected a number of reference phenotypes that are correlated to it (r^2>0.3 or top 30 traits) for imputation. We used the function "corr.test" of the R package "psych" with the parameters use="pairwise", method="spearman", adjust="fdr", alpha=0.05, to calculate the correlation coefficient. We also performed principal component analysis with the genotype data, which was implemented with function "bed_randomSVD" of the R package "bigsnpr"[39,40]. The top 20 principal components were used to construct the covariance matrix for representing the population structure. For each phenotype, we used the linear model to correct for the principal components and appropriate covariates and used residual $\boldsymbol{y}'$ for imputation.

GWAS and functional annotations. The phenotypes after imputations were used for genome-wide association analysis. We used a linear model $\boldsymbol{y}' = \boldsymbol{\mu} + \mathrm{SNP} + \boldsymbol{e}$ for GWAS, which was implemented via PLINK[38]. The genome-wide significance threshold was determined with the Bonferroni correction (α = 0.05/277,301, p < 8.87036E-08). We defined GWAS signals with a distance greater than 1Mb as mutually independent loci, we then summarized the number of independent loci identified in GWAS studies. Within the GWAS loci, we used the UCSC browser to annotate the functional genes in the GWAS region according to human genome assembly GRCh37, and screened the functional candidate genes in the 500kb region around the lead SNP.



673 **Figure 1 Comparison of the imputation accuracy of LMM, MICE, missForest, PHENIX and**
674 **MFRF under different sample size (a), heritability (b), number of phenotypes (c) and the**
675 **coefficient of relationship between individuals (d) in simulation study.**

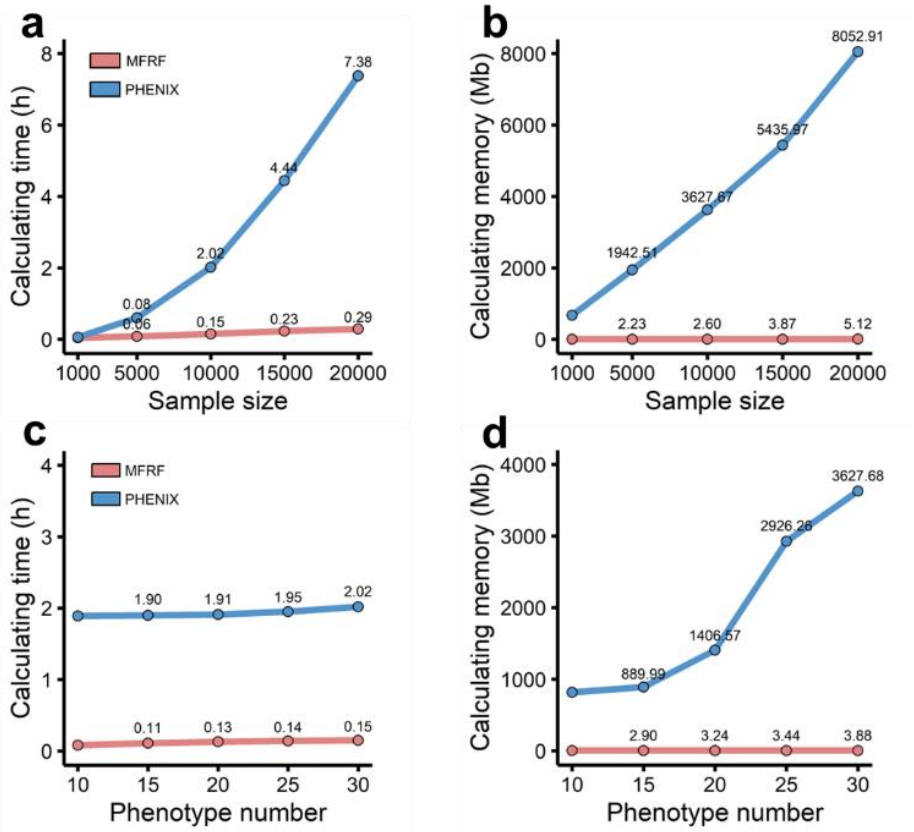


Figure 2. Comparison of the computational time (a and c) and the computer memory (b and d) between PHENIX and MFRF at different sample sizes and different number of phenotypes.

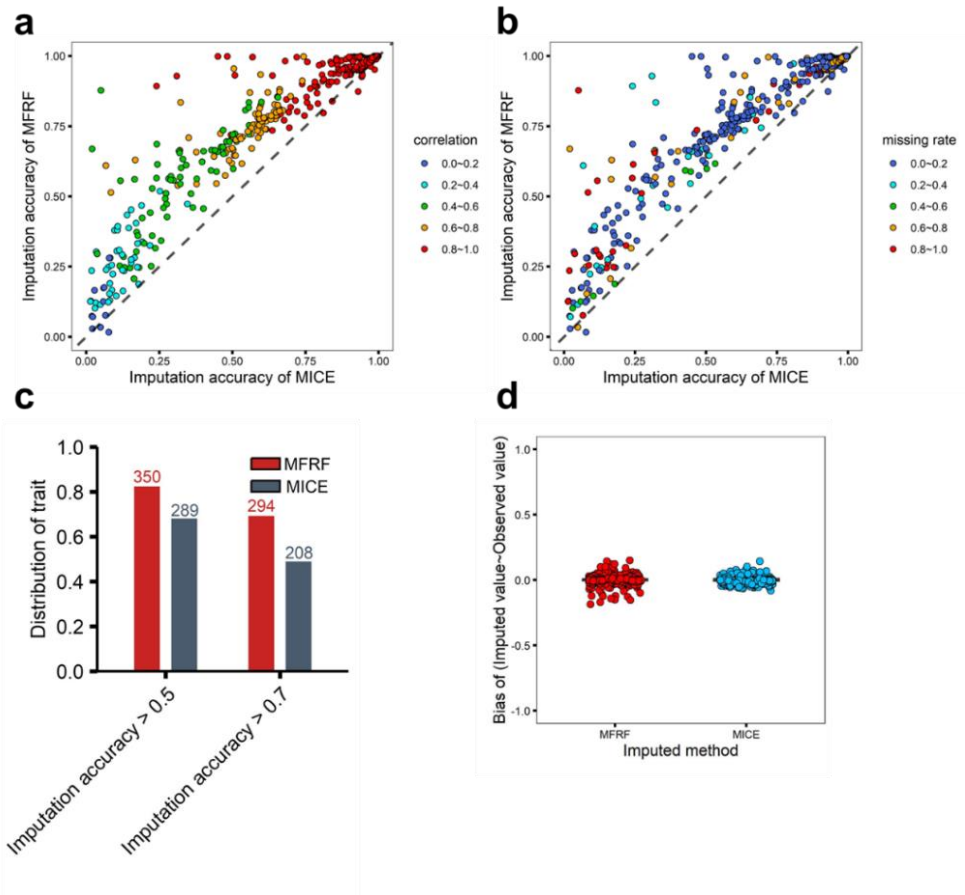


Figure 3 The comparisons of imputation performance between MFRF and MICE in 425 phenotypes in UKB data set. a, The imputation accuracies of MFRF and MICE for 425 phenotypes. Each dot represents a phenotype, with different colors representing the average correlation coefficient between the imputed phenotype and its reference phenotypes. For almost all 425 phenotypes, the imputation accuracies of MFRF are higher than those of MICE. Most phenotypes with higher imputation accuracy also have higher correlations with their reference phenotypes. The accuracy of imputation increases with the increase of correlation. **b**, The impact of missing rate on the imputation accuracies of MFRF and MICE. Missing rates of phenotypes are indicated by different colors. For most phenotypes, the imputation accuracies of MFRF are higher than those of MICE, and the missing rate has little effect on the imputation accuracy. **c**, The number of the accurately imputed phenotypes of MFRF and MICE (with imputation accuracy > 0.5 or 0.7). The numbers of phenotypes in MFRF are greater than those in MICE. **d**, The comparisons of the imputation bias of MFRF and MICE in 425 phenotypes. The average bias of the 425 phenotypes of MFRF and MICE is very close to zero and not significantly different from zero.

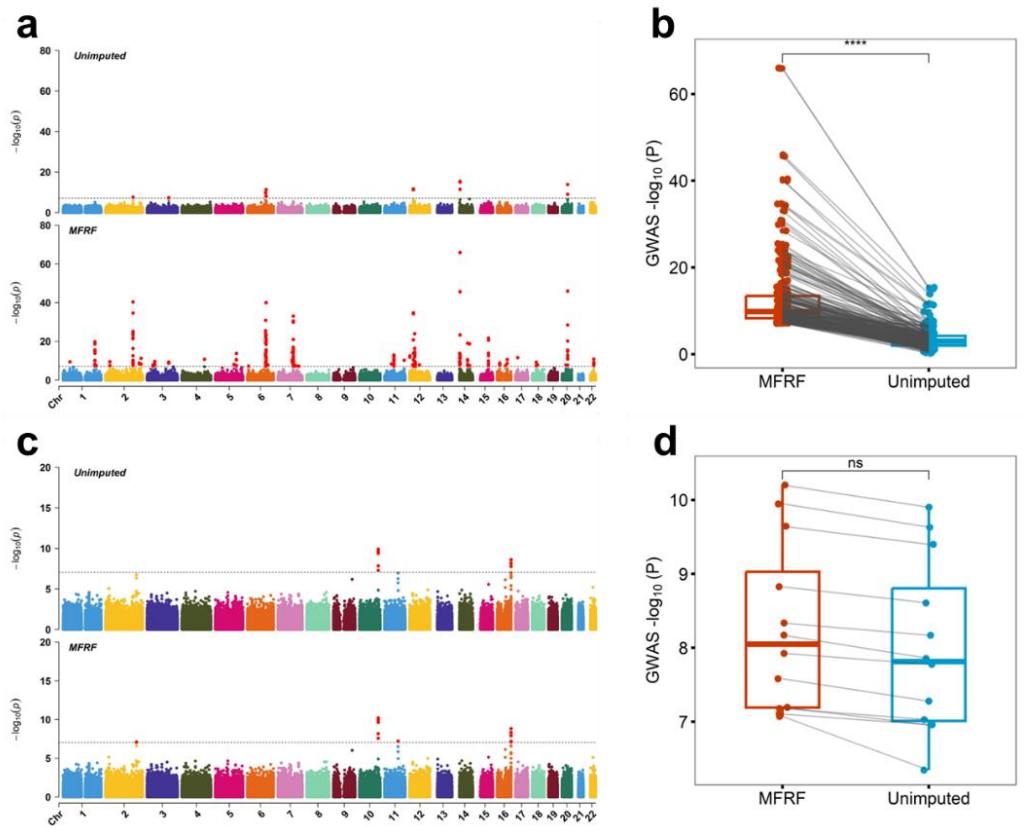


Figure 4 The GWAS of the heart rate and the cancer phenotypes in UKB data set. a, GWAS of heart rate phenotype with unimputed (upper panel) and imputed (lower panel) phenotypes. The dashed horizontal lines denote a significance threshold of 6.74. The GWAS signals are significantly enhanced and man extra GWAS loci are discovered after imputation compared to before imputation. **b**, comparison of the GWAS log10 (*P*) values before and after imputation of heart rate phenotype. The GWAS log10(*P*) values are significantly increased after imputation compared to before imputation. (**c**) and (**d**) the cancer phenotype GWAS results, which have the same legends to heart rate phenotypes.

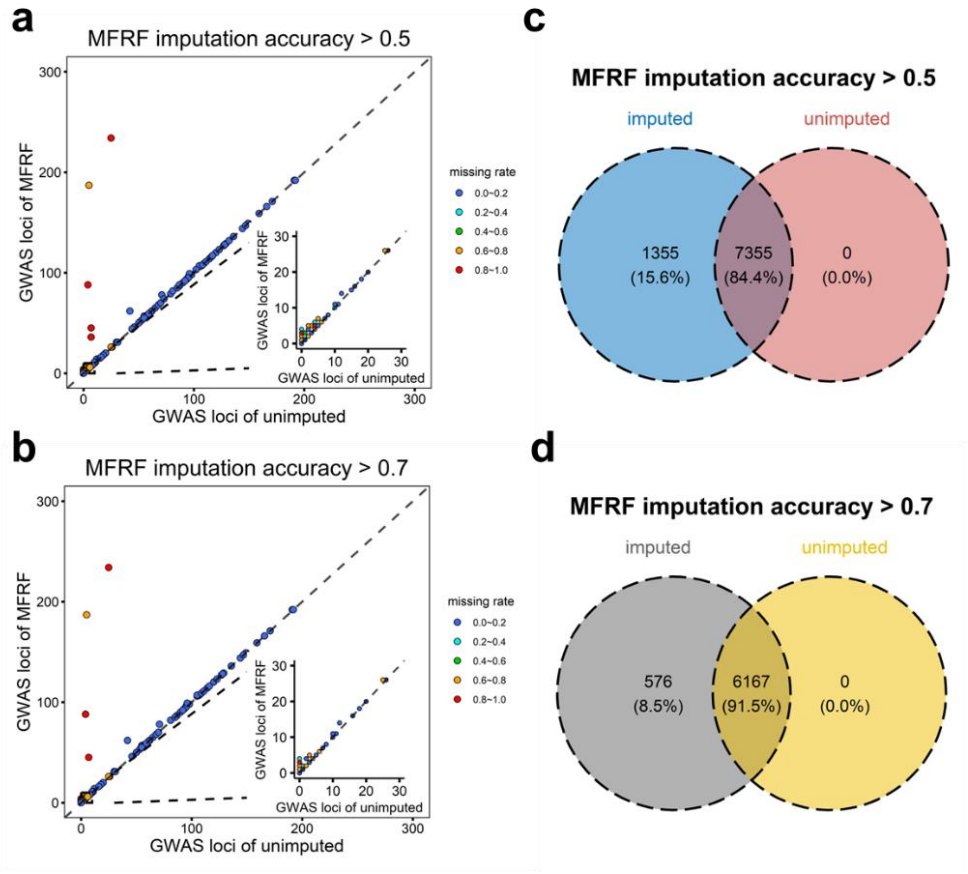


Figure 5 The comparison of the number of GWAS loci before and after imputation in UKB datasets. a and **b**, the number of GWAS loci for 350 (**a**) and 294 (**b**) phenotypes with imputation accuracy greater than 0.5 and 0.7, respectively, each point represents a phenotype, and the number of GWAS loci for most phenotypes after imputation is greater than before. **c** and **d**, the total number of GWAS loci identified by 350 (**c**) and 294 (**d**) phenotypes. More GWAS loci are identified after imputation than before, but no more GWAS loci are discovered before imputation than after.

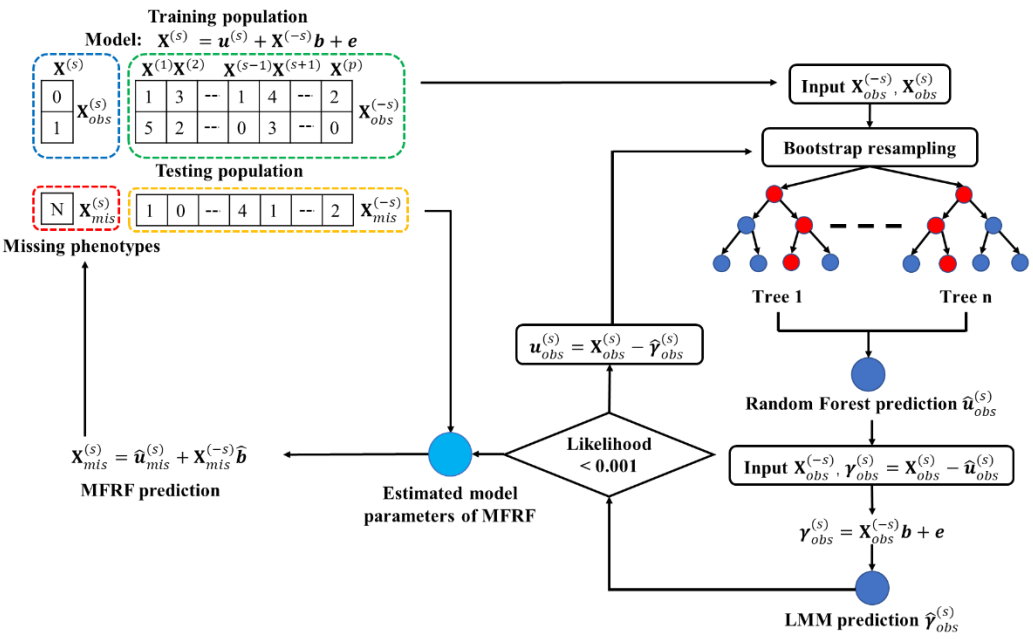
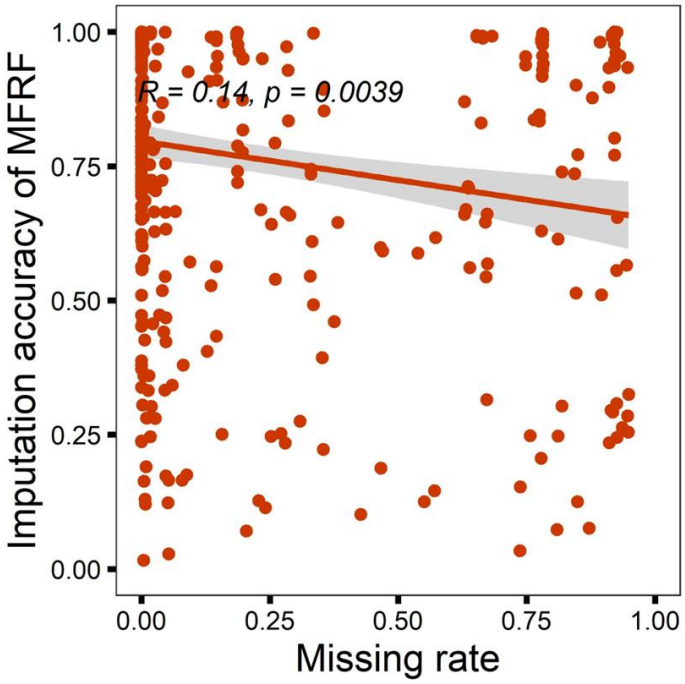
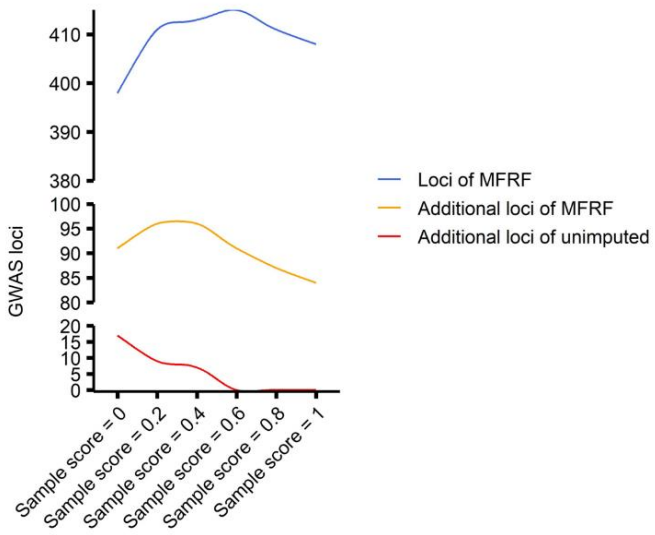


Figure 6 Flowchart of the MFRF used in this study.



Supplementary Figure 1 The relationship between missing rates and imputation accuracies of MFRF. The missing rate has little effect on imputation accuracy of MFRF. (*p-value* = 0.0039)



Supplementary Figure 2. The number of GWAS loci identified after imputation with MFRF (upper panel) at different SC thresholds, the number of extra GWAS loci identified after imputation than before imputation (middle panel), and the number of extra GWAS loci identified before imputation than after imputation (bottom panel). When the SC threshold was greater than or equal to 0.6, no extra GWAS loci were identified before imputation compared to after imputation, while the number of total GWAS loci identified after imputation was the highest.

Supplementary Table 1 Brief summary of the methods applied to simulated and real data sets.

Method	Description and properties	Reference
MFRF	Mixed fast random forest.	This paper
PHENIX	Bayesian multivariate mixed model fitted via variational Bayes.	[4]
missForest	non-parametric missing value imputation for mixed-type data.	[6]
MICE	Multivariate imputation by chained equations.	[5]
LMM	Single-trait linear mixed model with estimated BLUP used to impute missing values; ignores covariance between phenotypes.	[11–13]

Supplementary Table 2 The imputation accuracies of different methods with different sample size.

Method	Sample size					
	100	300	500	700	900	1100
	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy
MFRF	0.3018±0.0059	0.5884±0.0030	0.6340±0.0020	0.6511±0.0018	0.6579±0.0016	0.6657±0.0015
PHENIX	0.3976±0.0057	0.5760±0.0025	0.6178±0.0020	0.6367±0.0019	0.6445±0.0016	0.6526±0.0015
missForest	0.2876±0.0060	0.3930±0.0029	0.4356±0.0026	0.4567±0.0022	0.4696±0.0015	0.4814±0.0016
MICE	0.2595±0.0053	0.4349±0.0032	0.4554±0.0026	0.4679±0.0024	0.4684±0.0023	0.4705±0.0021
LMM	0.0847±0.0054	0.0932±0.0036	0.0966±0.0033	0.1000±0.0026	0.1052±0.0023	0.1045±0.0022

Supplementary Table 3 The imputation accuracies of different methods with different heritability.

Method	Heritability					
	0.1	0.2	0.4	0.6	0.8	1.0
	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy
MFRF	0.7785±0.0012	0.7050±0.0017	0.6334±0.0021	0.6331±0.0021	0.7043±0.0017	0.9160±0.0008
PHENIX	0.7665±0.0012	0.6925±0.0016	0.6185±0.0019	0.6177±0.0019	0.6902±0.0016	0.9112±0.0007
missForest	0.5549±0.0018	0.4999±0.0020	0.4353±0.0021	0.4340±0.0020	0.4963±0.0020	0.6091±0.0017
MICE	0.6479±0.0017	0.5487±0.0022	0.4557±0.0025	0.4596±0.0025	0.5436±0.0021	0.9116±0.0011
LMM	0.0280±0.0023	0.0476±0.0023	0.0757±0.0026	0.0971±0.0028	0.1146±0.0030	0.1284±0.0031

Supplementary Table 4 The imputation accuracies of different methods with different number of phenotypes.

Method	Phenotype number					
	20	40	60	80	100	120
	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy
MFRF	0.2251±0.0048	0.3286±0.0028	0.4053±0.0023	0.4686±0.0018	0.5233±0.0017	0.5726±0.0015
PHENIX	0.2375±0.0047	0.3275±0.0029	0.3951±0.0024	0.4555±0.0019	0.5065±0.0017	0.5554±0.0014
missForest	0.1854±0.0048	0.2527±0.0032	0.2874±0.0025	0.3134±0.0021	0.3320±0.0018	0.3474±0.0018
MICE	0.0847±0.0043	0.1585±0.0033	0.2310±0.0024	0.2997±0.0021	0.3652±0.0021	0.4299±0.0017
LMM	0.1002±0.0044	0.0987±0.0034	0.0975±0.0026	0.0967±0.0024	0.0970±0.0022	0.0973±0.0020

Supplementary Table 5 The imputation accuracies of different methods with different sample relatedness.

Method	Sample relatedness					
	0.02	0.2	0.4	0.6	0.8	0.9
	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy	Imputation accuracy
MFRF	0.6331±0.0021	0.6731±0.0021	0.7212±0.0024	0.7749±0.0024	0.8413±0.0021	0.8862±0.0017
PHENIX	0.6177±0.0019	0.6620±0.0020	0.7134±0.0024	0.7701±0.0025	0.8386±0.0022	0.8838±0.0017
missForest	0.4338±0.0020	0.5043±0.0025	0.5881±0.0034	0.6700±0.0036	0.7512±0.0033	0.7932±0.0030
MICE	0.4596±0.0025	0.5057±0.0027	0.5681±0.0032	0.6426±0.0034	0.7395±0.0032	0.8108±0.0026
LMM	0.0971±0.0028	0.3354±0.0042	0.4769±0.0048	0.5859±0.0049	0.6787±0.0046	0.7211±0.0043

Supplementary Table 6 The comparison of calculating times (in hours) and calculating memory (Mb) of PHENIX and MFRF for different sample size.

Sample size	Calculating time (h)		Calculating memory (Mb)	
	PHENIX	MFRF	PHENIX	MFRF
1000	0.06	0.04	670.49	1.71
5000	0.60	0.08	1942.51	2.23
10,000	2.02	0.15	3627.67	2.60
15,000	4.44	0.23	5435.97	3.87
20,000	7.38	0.29	8052.91	5.12

Note: Computing performance tests were performed in a Red Hat Enterprise Linux sever with 2.20 GHz Intel(R) Xeon(R) Silver 4114 CPU, and 125 Gb memory.

Supplementary Table 7 The comparison of calculating times (in hours) and calculating memory (Mb) of PHENIX and MFRF for different trait size.

Trait size	Calculating time (h)		Calculating memory (Mb)	
	PHENIX	MFRF	PHENIX	MFRF
10	1.89	0.08	815.53	2.60
15	1.90	0.11	889.99	2.90
20	1.91	0.13	1406.57	3.24
25	1.95	0.14	2926.26	3.44
30	2.02	0.15	3627.68	3.88

Note: Computing performance tests were performed in a Red Hat Enterprise Linux sever with 2.20 GHz Intel(R) Xeon(R) Silver 4114 CPU, and 125 Gb memory.