Rapid and accurate multi-phenotype imputation for millions of individuals

Lin-Lin Gu[1], Hong-Shan Wu[1], Yong-Jie Zhang[1], Tian-Yi Liu[1], Jing-Cheng He[2], Xiao-Lei Liu[3], Guo-Bo Chen[4,5,*], Dan Jiang[1,*], 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 Data Science, School of Mathematical Sciences, Zhejiang University, Hangzhou, Zhejiang, People's Republic of China

[3] 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

[4] 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

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

***Correspondence:**

Ming Fang<fangming618@126.com>;

Dan Jiang<jiangdan666@163.com>;

Guo-Bo Chen<chenguobo@gmail.com>

Abstract

Deep phenotyping can enhance the power of genetic analysis such as genome-wide association study (GWAS), but recurrence of missing phenotypes compromises the potentials of such resources. Although many phenotypic imputation methods have been developed, accurate imputation for millions of individuals still remains extremely challenging. In the present study, leveraging efficient machine learning (ML)-based algorithms, we developed a novel multi-phenotype imputation method based on mixed fast random forest (PIXANT), which is several orders of magnitude in runtime and computer memory usage than the state-of-the-art methods when applied to the UK Biobank (UKB) data and scalable to cohorts with millions of individuals. Our simulations with hundreds of individuals showed that PIXANT was superior to or comparable to the most advanced methods available in terms of accuracy. We also applied PIXANT to impute 425 phenotypes for the UKB data of 277,301 unrelated white British citizens and performed GWAS on imputed phenotypes, and identified a 15.6% more GWAS loci than before imputation (8,710 *vs* 7,355). Due to the increased statistical power of GWAS, a certain proportion of novel genes were rediscovered, such as *RNF220*, *SCN10A* and *RGS6* that affect heart rate, demonstrating the use of imputed phenotype data in a large cohort to discover novel genes for complex traits.

Key words: phenotype imputation, machine learning, UK Biobank, non-linear effect

Introduction

Large data analysis can significantly improve genetic researches, such as for increasing the statistical power for genome-wide association studies (GWAS)[1–3]. In the last several years, it has witnessed the tremendous growth of data for GWAS. For example, UK Biobank (UKB) database has collected nearly half million genotyped individuals with enriched phenotypes recorded[4], and has played a crucial role in discovering novel genotype-phenotype associations in recent years[5–8]. With the reduction of the cost of high-throughput genotyping, it is anticipated to emerge much deeper phenotype biobank-scale dataset and will facilitate genetic analysis. However, the move to high-dimensional phenotypes will inevitably lead to missing rate, which is ranged from 0.11% to 98.35% in UKB, a loss that decreases the discovery rate in the down-stream analyses[9]. If missing phenotypes can be imputed accurately, more complete sample size will be utilized for GWAS.

Traits often present a certain degree of correlation, which provides scaffold for multiple-phenotype imputation. A commonly used multi-phenotype imputation method is known as MICE, which uses chained equations to handle missing data[10]. MICE method is popular with statistician because of its efficient computation of large data. Compared with MICE, missForest has been developed and presented better performance when applied to continuous or binary traits[11] but are computationally much slower[12,13]. To better impute missing phenotypes, until recently an alternative variational Bayesian framework-based methods have been proposed, which include PHENIX[9]. PHENIX is based on a multivariate normal distribution containing a matrix of sample kinship and is more accurate than MICE and missForest, but computationally time-consuming and memory-intensive. However, it is estimated that PHENIX will take several months and ~1T computer memory for the UKB datasets, which will drain the computing resources of many institutes.

In this study, we propose a mix fast random forest (PIXANT) algorithm for multi-phenotype imputation, which optimizes large data analysis in terms of runtime and memory usage by constructing fast random forest trees, so that is scalable to large data over a million individuals. PIXANT models nonlinear and linear effects across multiple phenotypes and higher-order interactions between predictive factors, allowing it to produce unbiased imputation with much higher accuracy than the aforementioned methods. We validate our method with extensive simulation and compare it with the existing state-of-art methods, LMM[14–16], MICE[10], missForest[11] and PHENIX[9], and

demonstrate that our method outperforms the competing methods in accuracy and computational efficiency. We then show in UKB data that PIXANT is orders of magnitude faster than PHENIX with more scalable memory usage (e.g., when sample size is 20,000 and phenotype number is 30, PIXANT is ~24.45 times faster and uses only about one ten thousandth memory compared with PHENIX), and as sample size increases the advantage of PIXANT over PHENIX becomes more pronounced. We then apply PIXANT to the imputation of 425 UKB phenotypes and perform GWAS to illustrate that more completed phenotypes in UKB enables the identification of more GWAS loci, some of which are undiscovered due to missing phenotypes.

Results

Overview of method

PIXANT relies on the correlation between the target phenotype and available phenotypes. So, PIXANT first selects reference phenotypes that are correlated with the target phenotype. Next, using observed phenotypes, PIXANT models linear (***b***) and nonlinear (***u***) effects across multiple phenotypes, and consequently trains an imputation model. Finally, PIXANT imputes missing phenotypic values by project observable reference phenotypes into the imputation model. The flow chart is presented in **Fig.1**. To speed up the runtime and reduce the memory usage, PIXANT (1) uses a computationally efficient mix fast random forest (RF) to estimate parameters of linear mixed model (LMM); (2) uses the Knuth[17] (p. 137) algorithm for bootstrap resampling without replacement to construct trees for fast random forest prediction[18]; and (3) stores node information in a simple data structure, avoiding repeated accesses to the raw data and releasing memory early for efficient storage (**Online Methods**).

For PIXANT, the reference phenotypes are crucial in determining the imputation accuracy. In the training step, we use R^2 as a metric to trace the accuracy of the imputation model, in particular how big the sample size should be to achieve an accurate imputation model. Nevertheless, in the imputation step, even the imputation model is accurate, the missing rate of an individual's reference phenotypes and the imputation quality defined as sample score (*SC*; see **Online Methods**), determines the final imputation quality of this individual. In the sections below, we consequently will investigate how the key factors, such as quantity and quality of reference phenotype, sample size, and sample score, influence the performance of PIXANT both in

simulation and the UKB samples.

Computer simulations

We first investigated the impute of PIXANT through extensive computer simulations. The multi-trait mixed model was employed to simulate phenotypic values, which fully considered genetic and residual correlations between traits as well as relatedness between individuals to ensure phenotypic correlations (see ref.[9] for details). We set a set of default parameters in the simulation, the sample size of 500, the heritability of 0.6, the number of phenotypes of 60, the coefficient of relationship between individuals of 0.02, and the phenotypic missing rate of 5%. We performed four simulation scenarios to investigate the performance of PIXANT (i) different sample size, 100, 300, 500, 700, 900, and 1,100, respectively; (ii) different heritability, 0.1, 0.2, 0.4, 0.6, 0.8, and 1, respectively; (iii) different number of reference phenotypes, 20, 40, 60, 80, 100 and 120, respectively; and (iv) different relatedness between individuals, 0.02, 0.2, 0.4, 0.6, 0.8 and 0.9, respectively. We applied PIXANT and other benchmarked methods such as LMM[14–16], MICE[10], missForest[11] and PHENIX[9] in the simulations (**Supplementary Tab.1)**. The averaged prediction accuracy (Pearson's correlation) of 100 repeated experiments was used to evaluate the imputation accuracy.

As shown in **Fig.2a,** the imputation accuracies increased with sample size for all methods (except LMM); however, PIXANT and PHENIX outperformed all other methods over the full range of sample size. When the sample size was greater than 300, the imputation accuracy of PIXANT was the highest among these methods, followed by PHENIX, and both had higher imputation accuracy than missForest, MICE, and LMM (**Fig.2a** and **Supplementary Tab.2)**. Moreover, the sample size of 1,000 seemed to deliver practically useful accuracy, while an even larger sample size contributed little (**Fig.2a**).

When heritability was ranged from 0.1 to 1, the imputation accuracies of PIXANT were slightly higher than those of PHENIX, both of which were higher than MICE, missForest and LMM (**Fig.2b** and **Supplementary Tab.3)**. Additionally, reference phenotype number increased the accuracy of PIXANT, PHENIX, MICE and missForest, and the difference between PIXANT and the next best method (PHENIX) (**Fig.2c** and **Supplementary Tab.4**). PIXANT and PHENIX had the highest imputation accuracies across all the number of reference phenotype ranges, both were more accurate than other methods (**Fig.2c** and **Supplementary Tab.4**). However, the imputation accuracy of LMM was generally very low regardless of the number of reference phenotypes

(varied from 0.1002 to 0.0973 with the number of phenotypes p ranged from 20 to 120); whereas the imputation accuracy of PIXANT was 0.2251 for $p = 20$ and then increased to 0.5726 when $p = 120$.

The LMM method was overall the worst performing method, except at very high levels of relatedness between samples. When the relatedness between samples was increased, the imputation accuracies of all methods improved, especially for LMM that explicitly relied on the relatedness between samples via the kinship matrix (**Fig.2d**). For example, when the relatedness between individuals was 0.02 the imputation accuracy of LMM was 0.0971, but the imputation accuracy soon increased to 0.7211 when the relatedness between individuals was 0.9 (**Supplementary Tab.5**).

To sum up, we investigated the effects of sample size of the training model, genetic architecture of a typical polygenic trait of certain heritability, number of the reference traits, which all acted to influence the performance in general. After all, PIXANT remained the best performing method in the simulated scenarios. Of note, the high kinship scenario is not often the case, at least not for the UKB example which is consisted of unrelated individuals.

Computational efficiency in large data

We used the UKB data[4] consisting of 277,301 unrelated white British citizens (**Online Methods**) to evaluate the resource requirements of PIXANT and benchmarked it against PHENIX. Note that we did not compare PIXANT with other method as they were either less accurate (LMM and MICE) or computationally prohibitive (missForest). After randomly sampling subgroups of individuals (n ranged from 1,000 to 20,000 and p=30) from the UKB, we evaluated the computational efficiency of PIXANT and PHENIX respectively, on a computing platform with 125 GB memory and 8 CPU cores. As shown in **Fig 3a**, for an imputation with n ranged from 1,000 to 20,000, PHENIX took 0.06~7.38 hours, while PIXANT only requires 0.04~0.29 hours, which was 0.5 to 24.45 times the efficiency of PHENIX; moreover, the runtime of PIXANT was little affected by sample size, while the runtime of PHENIX increased substantially (**Fig 3a, Supplementary Tab.6**). As for memory requirements, the actual memory usage of PIXANT was generally stable to sample size (2.60～5.12MB for n ranged from 1,000 to 20,000), but PHENIX was not, e.g., PHENIX required 670.49 MB memory for $n = 1,000$, but the memory usage soon increased to 8,052.91MB when $n = 20,000$ (**Fig.3b** and **Supplementary Tab.6**). Predictably, the runtime of PIXANT increased linearly with sample size with almost invariant memory usage (**Fig.3a** and

Fig.3b), which suggested that PIXANT was, in principle, scalable for even larger sample size.

To evaluate the resource requirements of phenotype numbers, we randomly sampled subgroups of individuals (*p* ranged from 10 to 30 and *n*=10,000) from the UKB, respectively. When the phenotype number ranged from 10 to 30, PHENIX required 1.89 to 2.02 hours, while PIXANT only required 0.08 to 0.15 hours, saving 12.47 to 22.63 times of the runtime. In addition, the runtime of PIXANT was generally stable with different number of phenotypes, so was PHENIX (**Fig.3c** and **Supplementary Tab.7**). Benefiting from the memory-optimized design, the actual memory usage of PIXANT was almost invariant to number of phenotypes (~3MB for *p* range from 10 to 30), saving 312.67 to 933.97 times of the computer memory compared to PHENIX (815.53~3627.68MB for *p* range from 10 to 30) (**Fig.3d)**. Our observation that the runtime for both PIXANT and PHENIX increased linearly with the number of phenotypes, and the memory usage of the PIXANT was negligible compared to that of PHENIX (**Fig.3d** and **Supplementary Tab.7**). PIXANT could handle hundreds of traits given the same computing environment as used in this study.

Application in UKB data

We conducted an imputation in the 277,301 participants of UKB datasets for 387 continuous and 38 binary phenotypes that had missing rates less than 95% and were likely to be genetically relevant (**Fig.4a, Supplementary file 1**), the average missing rate was 21.3%. The missForest and PHENIX were excluded for their prohibitive computational cost for large data, and only MICE was employed as a comparison here. To reduce the impact of phenotypic stratification, we adjusted the phenotypes with the first 20 genotype-generated principal components, gender, and age (**Online methods**). We then selected a number of phenotypes that were highly correlated (r>0.3) with the phenotype to be imputed or the 30 most correlated (if no highly correlated) phenotypes, as reference phenotypes. The number of reference phenotypes for each of the 425 imputed phenotypes ranged from 2 to 29 (median=18), and the correlation of the imputed phenotype with the reference phenotypes ranged from 0.06 to 0.99 (**Supplementary file 1**). We randomly sampled 800 individuals for each phenotype, and used their imputed and true values to evaluate the imputation accuracy.

As shown in **Fig.4b**, PIXANT outperformed MICE in almost all 425 phenotypes, of 425 phenotypes, 356 (83.8%) and 298 (70.1%) imputed by PIXANT had accuracy greater than 0.5 and 0.7, respectively (**Fig.4c**), but 288 (67.8%) and 208 (48.9%) by

MICE (**Fig.4c**). In addition, the relationship between missing rate and imputation accuracy was quite limited (**Fig.4b**, *p*-value=0.00029 and 0.00028 for PIXANT and MICE, respectively, as show in **Supplementary Fig.1**). For example, the missing rate of "Diastolic brachial blood pressure during PWA" (**UKB data-field: 12675**) was 92.11%, the imputation accuracies of PIXANT and MICE still reached 0.947 and 0.891, respectively (**Supplementary file 1**). This was because the UKB data was large, and even a substantial proportion of phenotypes was missing, the remaining individuals (19,743 individuals in the above case) with high-quality reference phenotypes could still bring out reliable imputation. It was consistent to our simulation study that a certain number of sample size (greater than 1,000) would be practically sufficient to train a good imputation model. Third, we observed the imputation accuracies of the methods largely depended on the quality of the reference phenotypes. We measured the quality of the reference phenotypes with multiple regression R^2 defined as the contribution of the reference phenotypes to the imputed phenotype. As shown in **Fig.4d**, the accuracy of the imputation increased with R^2, indicating that, as expected, the quality of reference phenotypes played an important role in determining the accurate imputation; besides, the imputation accuracy of MICE was almost linear to R^2, while the accuracy of PIXANT was a trajectory curve, implying its advantage in drawing nonlinear information from the reference phenotypes (**Fig.4d)**. Fourth, we found that the imputation bias (**Online Methods**) of PIXANT was clearly better (closer to 1) than that of MICE at different range of imputation accuracies (**Fig.4e**), suggesting imputation with PIXANT led to more reliable GWAS signals than with MICE; furthermore, we observed the imputation bias depended on the imputation accuracy, and with the imputation accuracy increased the imputation bias improved for both methods (**Fig.4f-p**). For example, when the imputation accuracy was between 0 and 0.1, the imputation bias of PIXANT was 0.029 (**Fig.4g**), but when the imputation accuracy was between 0.9 and 1, the imputation bias of PIXANT quickly increased to 0.997 (**Fig.4p**), suggesting that higher accuracy could bring out unbiased imputation.

Individual-level quality control for imputed phenotypes

Even the imputation accuracy was greater than 0.5 or 0.7, some individual's phenotypes could be poorly imputed because of the low quality of the reference phenotypes. For example, the GWAS of QT interval phenotype (**UKB data-field: 22331**, imputation accuracy=0.671) identified far fewer loci after imputation with PIXANT than without imputation (**Fig.5a**). We therefore developed a sample score (*SC*) for quality control of

each individual imputed phenotype. It was clear that when an individual's own reference phenotypic values were absent, the imputed phenotypic value would consequently be of low accuracy, so we defined *SC* as the contribution of the observed reference phenotypes to its complete reference phenotypes selected for an imputation task (**Online Methods**). For QT interval phenotype, we filtered the imputed individual phenotypic values with different *SC* thresholds, ranging from 0 (no reference phenotypic values) to 1 (complete reference phenotypic values), and evaluated the imputation accuracy at different *SC* levels using 800 phenotypic values (as previously described), respectively. We noticed that 78.2% of phenotypic values had *SC* less than 0.2 (**Fig.5b**), which explained low power of GWAS after imputation of QT interval phenotype. We further scrutinized if the developed *SC* was able to well approximate the real imputation accuracy at individual level. We summarized the imputation accuracies and *SC*s for all the 356 phenotypes (with an imputation accuracy greater than 0.5), respectively, and observed the imputation accuracy increased proportionally to *SC*s (**Fig.5c**), implying that *SC* could well approximate the real imputation accuracy. We also found that the imputation accuracy increased sharply between 0 and 0.2 but slowly from 0.2 to 1 (**Fig.5c**), suggesting that *SC* greater 0.2 might be a potential threshold that brings accuracy above 0.7.

To refine the *SC* threshold, we randomly selected 42 imputed phenotypes and filtered individual phenotypic values with *SC* thresholds 0, 0.2, 0.4, 0.6, 0.8, and 1, and performed the GWAS with the filtered phenotypes respectively. We observed that *SC* of 0.6 indicating an imputation accuracy of about 0.8, was the optimal threshold, which could bring out the highest GWAS power (upper panel in **Fig. 5d**) and largest GWAS loci overlap before and after imputation (bottom panel in **Fig. 5d**). However, as shown in the upper panel of **Fig. 5d**, when *SC* is between 0.2 to 0.6, the GWAS power and GWAS loci overlap become stable, and are therefore recommended for individual-level imputation filtering.

UKB GWAS after PIXANT imputation

We assessed the usefulness of our method with GWAS of the 425 imputed UKB phenotypes. To ensure the reliability of GWAS, we only retained phenotypes with imputation accuracy greater than 0.5 or 0.7, leaving 356 and 298 phenotypes, respectively (**Supplementary file 2**). After removing individual phenotypes with $SC < 0.6$ for each UKB phenotype, we performed GWAS with the filtered phenotypes using PLINK[19] and applied genomic control to reduce the effect of population stratification[20].

We used a conservative *p*-value threshold of 8.87E-08 after Bonferroni correction to identify significant GWAS loci (**Online Methods**).

The first example was Heart rate (**UKB data-field: 5983**) with a missing rate of 84.92% and PIXANT's imputation accuracy of 0.772. The GWAS results for heart rate using imputed (*n*=271,806) and unimputed (*n*=40,990) phenotypes were compared in **Fig.6a**. Before imputation, only 6 independent GWAS loci (>1Mb from each other) were identified, after imputation, the GWAS signals were significantly boosted, and another 31 independent loci were identified (**Fig.6a** and **Supplementary Tab.8**). The plot highlighted several peaks containing genes related to dilated cardiomyopathy, cardiac rhythm regulation, cardiac parasympathetic nervous system, heart development and heart disease (*DSP* (ref. [21,22,23]), *CDH11* (ref. [24,25]), *RNF220* (ref. [26,27]), *SCN10A* (ref. [28,29]), *MYH11* (ref. [30]) and *RGS6* (ref. [31,32]), etc.) (For details, see **Supplementary Tab.8**).

The second example was direct bilirubin phenotype (**UKB data-field: 30660**) with missing rate of 14.73% and PIXANT's imputation accuracy of 0.910. After imputation, the GWAS signals were significantly enhanced, and additional GWAS loci were found on chromosomes 1, 2, 3, 4, 6, 10, 15, and 22 (**Fig.6b** and **Supplementary Tab.8**). Interestingly, a lead SNP within *CSP1* of chromosome 2 was associated with gallbladder disease, clinicopathological characterisation showed that increased *CPS1* expression was associated with reduced levels of international normalized ratio (INR), total protein and alkaline phosphatase (ALP) and indirect bilirubin[33]. A lead SNP within *GSTM1* on chromosome 1, whose null genotype has a high risk of pathologic hyperbilirubinemia, may have higher bilirubin levels[34]; in addition, *GSTM1* gene polymorphism was also associated with total bilirubin levels in neonatal jaundice[35] (**Supplementary Tab.8**).

We then summarized the GWAS results for all the 356 phenotypes with an imputation accuracy greater than 0.5. After imputation, the number of GWAS loci was significantly higher than before imputation (**Fig.6c)**, totally, 7,355 independent GWAS loci were identified before imputation, while 8,710 were identified after imputation, an increase of 15.6% (**Fig.6d**). As expected, the difference of the number of GWAS loci before and after imputation became more pronounced as the phenotypic missing rate increased (**Fig.6c)**. The results of 298 phenotypes with an imputation accuracy greater than 0.7 was quite similar to the above results (**Fig.6e**), a total of 6,167 GWAS loci were identified before imputation, while 576 (8.5%) additional loci were identified after

imputation (**Fig.6f**). These suggest that with reasonable quality control at both the phenotype and individual levels, the imputation could significantly improve the GWAS power. The GWAS results are detailed in **Supplementary file 3**.

Discussions

Missing data are inevitable for large genetic data in both genotypes and phenotypes. Here we have developed a new multi-phenotype imputation method, PIXANT, the major advantage lies in its computational speed and memory usage for big data. PIXANT is several orders of magnitude in runtime and computer memory usage than the state-of-the-art method, such as PHENIX, under the application for millions of individuals. We showed by simulation and real data analyses that our method rivalled or outperformed the state-of-the-art methods in accuracy. Moreover, real data analysis and biologically plausible results suggest that imputation can uncover new true positive associations (**Fig.6a** and **Supplementary Tab.8**). For example, we identified an additional locus with lead variant rs272570 within *RNF220*, a gene that is strongly associated with heart rate, and this association had also been reported repeatedly in previous studied[26,27].

Like many multi-phenotype imputation methods[10,36], combining high-quality reference phenotypes for imputation can improve the imputation accuracy of PIXANT. In this study, we defined the quality score of the selected reference phenotype as the multiple regression R^2 between the imputed phenotype and its reference phenotypes, which is useful in pre-selecting reference phenotypes for imputation. Our results also suggest that imputation will not always bring out more signals, as when imputation quality is demonstrably poor, the imputation may attenuate the association signals. We recommend filtering phenotype imputation results with care and attention. In this research, we proposed a quality control metric *SC* at individual level for the first time, which was found to be essential for improving power and controlling false positive signals in GWAS. Additionally, we also demonstrate it reasonable by real data analysis that the threshold of *SC* was between 0.2~0.6 (**Fig.5**).

Our imputation model requires a small training sample size. As the UKB data used was large, even though a substantial proportion of phenotypes were missing, the remaining individuals with high-quality observable phenotypes could still bring out reliable imputation; consistent to our simulation study, a sample size greater than a certain number (typically tens of thousands in the case of UKB data) would be

practically sufficient to train a good imputation model (**Fig.2a** and **Supplementary Tab.1**).

Large number of binary phenotypes are recorded in the UKB data in addition to continuous phenotypes. In the study, we directly take the binary phenotype as a continuous for imputation, and it usually performs well in practice. The core of PIXANT is the random forest, which enables to solve the binary phenotypes using classification trees rather than regression trees for continuous phenotype, which is worthy of further study.

PIXANT also can be applied in imputation of multi-omics data in addition to phenotype data. GTEx project contains nearly 10,000 RNA-seq records of 44 tissues for gene expression study,[37–39], and plays an important role in identification of functional genes underling human diseases[40,41]. The extensive high correlation between gene expressions across tissues in GTEx provides abundant information for gene expression imputation, and PIXANT will be very useful for such high-dimensional data imputation. In addition, PIXANT also can be applied in other omics data such as chromatin immunoprecipitation sequencing (ChIP-seq)[42], assay for transposase-accessible chromatin sequencing (ATAC-seq)[43], whole genome bisulfite sequencing (WGBS)[44,45] and so on. Imputation for high-dimensional multi-omics data will be favorable for discovering genomic regulatory elements, and will become more common for further exploring causal genes and pathways underlying human traits and diseases. However, it is non-trivial to achieve high computational efficiency for imputation and will be central to this work.

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;
PIXANT, https://github.com/GuLinLin/PIXANT.

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); the Seed

Industry Innovation and Industrialization Project of Fujian Province (2021FJSCZY01) ; 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

FM, JD, and CGB designed and supervised the study. GLL conceived and developed the method, wrote the program and conduct the whole analysis. WHS, ZYJ and LTY assisted in the data collection and contributed to the demonstration of the results. CGB and JCH contributed to real data analysis and conceived the imputation quality control method. FM, CGB, and LXL discussed results and methods, and provided comments that improved earlier versions of the manuscript. FM, JD, and CGB contributed resources and funding. GLL, FM and JD drafted the manuscript, and CGB revised first draft of the manuscript. All authors reviewed and approved the final manuscript.

DATA availability

The individual-level genotype and phenotype data are available through formal application to the UK Biobank (**URLs**).

References

1. Marx, V. Human phenotyping on a population scale. *Nat Methods* **12**, 711–714 (2015).
2. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43**, 333–338 (2011).
3. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**, 579–589 (2010).
4. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
5. Craig, J. E. *et al.* Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nat Genet* **52**, 160–166 (2020).
6. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668–681 (2018).
7. Kemp, J. P. *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat Genet* **49**, 1468–1475 (2017).
8. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).
9. Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat Genet* **48**, 466–472 (2016).
10. Buuren, S. van & Groothuis-Oudshoorn, K. mice : Multivariate Imputation by Chained Equations in R. *J. Stat. Soft.* **45**, 1–67 (2011).
11. Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
12. 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).
13. Jin, H., Jung, S. & Won, S. missForest with feature selection using binary particle swarm optimization improves the imputation accuracy of continuous data. *Genes Genom* **44**, 651–658 (2022).
14. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821–824 (2012).
15. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* **11**, 407–409 (2014).
16. Runcie, D. E. & Crawford, L. Fast and flexible linear mixed models for genome-wide genetics. *PLoS Genet* **15**, e1007978 (2019).
17. Knuth, D. E. *The art of The Art of Computer Programming, volume 2*. (Addison-Wesley, Reading, 1985).
18. 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).
19. 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).
20. Devlin, B., Roeder, K. & Wasserman, L. Genomic Control, a New Approach to Genetic-Based Association Studies. *Theor Popul Biol* **60**, 155–166 (2001).
21. Jordan, E. *et al.* Evidence-Based Assessment of Genes in Dilated Cardiomyopathy. *Circulation* **144**, 7–19 (2021).

22. Mazzarotto, F. *et al.* Reevaluating the Genetic Contribution of Monogenic Dilated Cardiomyopathy. *Circulation* **141**, 387–398 (2020).
23. Augusto, J. B. *et al.* Dilated cardiomyopathy and arrhythmogenic left ventricular cardiomyopathy: a comprehensive genotype-imaging phenotype study. *Eur Heart J Cardiovasc Imaging* **21**, 326–336 (2020).
24. Schroer, A. K. *et al.* Cadherin-11 blockade reduces inflammation-driven fibrotic remodeling and improves outcomes after myocardial infarction. *JCI Insight* **4**, e131545 (2019).
25. Bowler, M. A. *et al.* Cadherin-11 as a regulator of valve myofibroblast mechanobiology. *Am J Physiol Heart Circ Physiol* **315**, H1614–H1626 (2018).
26. Ramírez, J. *et al.* Thirty loci identified for heart rate response to exercise and recovery implicate autonomic nervous system. *Nat Commun* **9**, 1947 (2018).
27. Global BPgen Consortium *et al.* Identification of heart rate–associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* **45**, 621–631 (2013).
28. 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).
29. 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).
30. Verstraeten, A., Luyckx, I. & Loeys, B. Aetiology and management of hereditary aortopathy. *Nat Rev Cardiol* **14**, 197–208 (2017).
31. Yang, J. *et al.* RGS6, a Modulator of Parasympathetic Activation in Heart. *Circ Res* **107**, 1345–1349 (2010).
32. 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).
33. Ma, S.-L., Li, A.-J., Hu, Z.-Y., Shang, F.-S. & Wu, M.-C. Co-expression of the carbamoyl-phosphate synthase 1 gene and its long non-coding RNA correlates with poor prognosis of patients with intrahepatic cholangiocarcinoma. *Mol. Med. Rep.* **12**, 7915–7926 (2015).
34. Abdel Ghany, E. A. G., Hussain, N. F. & Botros, S. K. A. Glutathione S-Transferase Gene Polymorphisms in Neonatal Hyperbilirubinemia. *J. Investig. Med.* **60**, 18–22 (2012).
35. Muslu, N. *et al.* Are glutathione S-transferase gene polymorphisms linked to neonatal jaundice? *Eur. J. Pediatr.* **167**, 57–61 (2007).
36. Hormozdiari, F. *et al.* Imputing Phenotypes for Genome-wide Association Studies. *The American Journal of Human Genetics* **99**, 89–103 (2016).
37. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
38. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
39. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
40. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
41. Yin, X. *et al.* Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. *Am J Hum Genet* **109**, 1727–1741 (2022).

42. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669–680 (2009).
43. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).
44. Ziller, M. J., Hansen, K. D., Meissner, A. & Aryee, M. J. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* **12**, 230–232 (2015).
45. Wang, Q. *et al.* Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc* **8**, 2022–2032 (2013).
46. Wang, J. *et al.* Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *Am. J. Hum. Genet.* **98**, 697–708 (2016).
47. 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).
48. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565–569 (2010).
49. 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).
50. 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).
51. Li, Y. *et al.* Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. *Genome Biol* **20**, 1–18 (2019).

Online methods

Multi-phenotype imputation with mixed fast random forest (PIXANT). PIXANT fits linear effects and nonlinear (higher-order interactions) among p reference phenotypes. The model can be written as,

$$\mathbf{y} = \boldsymbol{u} + \mathbf{X}^{(\mathrm{refs})}\boldsymbol{b} + \boldsymbol{e}, \tag{1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of imputed phenotypic values, $\mathbf{X}^{(\mathrm{refs})}$ is a $n \times p$ matrix, denotes p reference phenotypic values; $\boldsymbol{u}$ is the non-linear effect caused by p reference phenotypes selected for imputation, $\boldsymbol{b}$ is the linear effect of p reference phenotypes, which is a $p \times 1$ vector; $\boldsymbol{e}$ denotes the random error, which follows a normal distribution.

Before imputation, PIXANT first trains the predictive model using observable phenotypes. We denote the observed phenotypic values by $\mathbf{y}_{obs}$ and the missing phenotypic values by $\mathbf{y}_{mis}$; and similarly, we use $\mathbf{X}_{obs}^{(\mathrm{refs})}$ to denote p reference phenotypic values correspongding to $\mathbf{y}_{obs}$ individuals, and $\mathbf{X}_{mis}^{(\mathrm{refs})}$ to denote reference phenotypic values corresponding to $\mathbf{y}_{mis}$ individuals. Then, the training model can be expressed as

$$\mathbf{y}_{obs} = \boldsymbol{u}_{obs} + \mathbf{X}_{obs}^{(\mathrm{refs})}\boldsymbol{b} + \boldsymbol{e}. \tag{2}$$

We solve the model parameters of equation (2) using iterative algorithm. Before updating, we first sorted reference phenotypes according to missing rate from small to large.

(1) *Initialization*. We first initialize the model parameters with $\boldsymbol{b} = 0$ and $\boldsymbol{u} = \mathbf{y}$, and $\mathbf{X}^{(\mathrm{refs})}$ is initialized with average observed phenotypic values or randomly sampled from the non-missing values [11].

(2) *Update of linear-effect parameters*. Let

$$\boldsymbol{\gamma}_{obs} = \mathbf{X}_{obs}^{(\mathrm{refs})}\boldsymbol{b} + \boldsymbol{e}, \tag{3}$$

representing the phenotypic values attributed to linear effect and random error, where $\boldsymbol{\gamma}_{obs}$ at t^{th} iteration equals to $\boldsymbol{\gamma}_{obs(t)} = \mathbf{y}_{obs} - \boldsymbol{u}_{obs(t-1)}$ by substituting equation (3) to equation (2); then $\boldsymbol{b}_{(t)}$ is updated by solving linear regression of equation (3).

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

$$\mathbf{g}_{obs} = \boldsymbol{u}_{obs} + \boldsymbol{e} \tag{4}$$

representing the phenotypic values attributed to non-linear effect and random error. We fit a non-linear fast random forest[46] function $f(\mathbf{X}_{obs}^{(\mathrm{refs})})$ for $\boldsymbol{u}_{obs}$, i.e. $\boldsymbol{u}_{obs} = f(\mathbf{X}_{obs}^{(\mathrm{refs})})$, reflecting the complex non-linear effect of p reference phenotypic values on the

imputed phenotype. Then equation (4) can be written as

$$\mathbf{g}_{obs} = f(\mathbf{X}_{obs}^{(\mathrm{refs})}) + \boldsymbol{e} \tag{5}$$

where $\mathbf{g}_{obs}$ at t^{th} iteration can be obtained by

$$\mathbf{g}_{obs(t)} = \mathbf{y}_{obs} - \mathbf{X}_{obs}^{(\mathrm{refs})}\boldsymbol{b}_{(t-1)} \tag{6}$$

after substituting equation (4) to (2). Then the parameters involved in fast random forest function $f(\mathbf{X}_{obs}^{(\mathrm{refs})})$ at iteration t^{th} can be estimated by solving the random forest non-linear regression of equation (5). To speed up the runtime and reduce the memory usage, we use a computationally efficient mix fast random forest (RF) to estimate model parameters implemented via efficient sampling and data storage techniques (see main text for more detail). We therefore have the update of the non-linear effect $\boldsymbol{u}_{obs(t)}$ with estimated parameters in $f(\mathbf{X}_{obs}^{(\mathrm{refs})})$.

We update step (2)-(3) repeatedly, 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.

In the same way, we also train the model for each reference phenotype, respectively. The model can be written as

$$\mathbf{X}_{obs}^{(i)} = \boldsymbol{u}_{obs} + \mathbf{X}_{obs}^{(\mathrm{refs})}\boldsymbol{b} + \boldsymbol{e} \tag{7}$$

for i=1, 2, $\cdots$, p, which is similar to equation (2), except that $\mathbf{y}_{obs}$ is substituted by $\mathbf{X}_{obs}^{(i)}$, thus $\mathbf{X}_{obs}^{(\mathrm{refs})} = [\mathbf{X}_{obs}^{(-i)}, \mathbf{y}_{obs}]$. The model parameters for each reference phenotype in equation (7) are also solved using iterative algorithm as described above.

We have completed the model training, and next, we are going to impute missing phenotypic values using trained model parameters. We use the predict model

$$\mathbf{y}_{mis} = \mathbf{X}_{mis}^{(\mathrm{refs})}\boldsymbol{b} + f(\mathbf{X}_{mis}^{(\mathrm{refs})}) \tag{8}$$

to update $\mathbf{y}_{mis}$, and use the predict model

$$\mathbf{X}_{mis}^{(i)} = \mathbf{X}_{mis}^{(\mathrm{refs})}\boldsymbol{b} + f(\mathbf{X}_{mis}^{(\mathrm{refs})}) \tag{9}$$

to update $\mathbf{X}_{mis}^{(i)}$, for i=1, 2, $\cdots$, p, where $\mathbf{X}_{mis}^{(\mathrm{refs})} = [\mathbf{X}_{mis}^{(-i)}, \mathbf{y}_{mis}]$. PIXANT updates $\mathbf{y}_{mis}$ and $\mathbf{X}_{mis}^{(i)}$, iteratively, using the most recently updates whenever possible. The update is repeated for several times until the difference of missing phenotypes $\mathbf{Z}$, where $\mathbf{Z} \in \left\{\mathbf{y}_{mis}, \mathbf{X}_{mis}^{(1)}, \mathbf{X}_{mis}^{(2)}, \cdots, \mathbf{X}_{mis}^{(p)}\right\}$, in two consecutive iterations meets a certain criterion ($\theta < 0.001$), where θ is defined as

$$\theta = \frac{\sum(\mathbf{Z}_{(t)}^{(i)} - \mathbf{Z}_{(t-1)}^{(i)})'(\mathbf{Z}_{(t)}^{(i)} - \mathbf{Z}_{(t-1)}^{(i)})}{\sum \mathbf{Z}_{(t)}^{(i)\prime}\mathbf{Z}_{(t)}^{(i)}}. \tag{10}$$

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

$$r = \frac{\sum(\mathbf{X}^{(true)} - \bar{\mathbf{X}}^{(true)})(\mathbf{X}^{(imp)} - \bar{\mathbf{X}}^{(imp)})}{\sqrt{\sum(\mathbf{X}^{(true)} - \bar{\mathbf{X}}^{(true)})^2 \sum(\mathbf{X}^{(imp)} - \bar{\mathbf{X}}^{(imp)})^2}} \tag{11}$$

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

Sample score. For the sth missing phenotype of individual i, assuming that there are p reference phenotypic values, of which p' phenotypic values are complete and the others are missing, the *SC* is defined as

$$SC_{i,s} = \frac{\sum_{j \neq s}^{p'} r_{i,j}^2}{\sum_{j \neq s}^{p} r_{i,j}^2} \tag{12}$$

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 randomly selected an additional 0.29% of the complete phenotypic values as missing (800 phenotypic values) from imputed phenotype, and then evaluated the imputation bias using the regression coefficient between the imputed and true phenotypic values, where the regression analysis was implemented via the function "lm" in R.

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

PHENIX. The PHENIX method is expressed as

$$Y = U + \varepsilon \tag{13}$$

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

$$U \sim MN(0, B \otimes K) \tag{14}$$

$$\varepsilon \sim MN(0, E \otimes I_n) \tag{15}$$

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 phenotypic values $Y^{(miss)}$ by approximating joint posterior distribution of model parameters (For details, see refs. [9]).

missForest. MissForest is a kind of multi-phenotype imputation method based on random forest method. It first makes an initial guess of the missing phenotypic values 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[11].

MICE. MICE is developed by Buuren and Oudshoorn[10], which is based on the multivariate multiple imputation scheme of Schafer (2002)[47]. 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

$$P(Y_s|Y_{-s},\theta_s), \text{ for } s = 1,2,\dots,p \tag{16}$$

the model parameters are successively drawn with Gibbs sampling. Specifically, for t^{th} iteration, they are sampled from the conditional posterior probability distribution,

$$\theta_1^{*(t)} \sim P(\theta_1|Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}) \tag{17}$$

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

$$\cdots\cdots$$

$$\theta_p^{*(t)} \sim P(\theta_p|Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}) \tag{19}$$

$$Y_p^{*(t)} \sim P(Y_p|Y_p^{obs}, Y_1^{(t)}, \dots, Y_p^{(t)}, \theta_p^{*(t)}) \tag{20}$$

where $Y_s^{(t)} = (Y_s^{obs}, Y_s^{*(t)})$ 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 phenotypic values with BLUP technique. For each phenotype, LMM is defined as

$$Y \sim N(0, \sigma_g^2 K + \sigma_e^2 I_n) \tag{21}$$

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[48].

Simulations. We simulated genetically and environmentally related phenotypic data with multiple phenotypes according to the method of Dahl and Baud[9]. The method dissects the phenotypic values into genetic and environmental effects. The phenotype

was simulated with model

$$Y \sim MN(0, K, h^2\Sigma_G) + MN(0, I, (1 - h^2\Sigma_E)) \quad (22)$$

where Y is $n \times p$ the multi-trait phenotypic values 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 (22). The simulation is based on Dahl and Baud's script[9]. 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 data. The UK Biobank Project contains half a million UK participants, 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[4]. 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 (**UKB data-field 22006:** Genetic ethnic grouping; **UKB data-field 22021:** Genetic kinship to other participants), 278,788 unrelated individuals were retained. We then used PLINK[19] software for quality control (QC) to remove individuals with missing genotypes > 5% and SNPs with missing rates > 5%, Hardy-Weinberg equilibrium P < 1e-6 and MAF < 0.01. 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 data 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 29 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"[49,50]. 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[19]. We defined the whole-genome significance cutoff as the Bonferroni test threshold, which was set as 0.05/total SNPs[51] (α = 0.05/563,675, p-value < 8.87E-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.

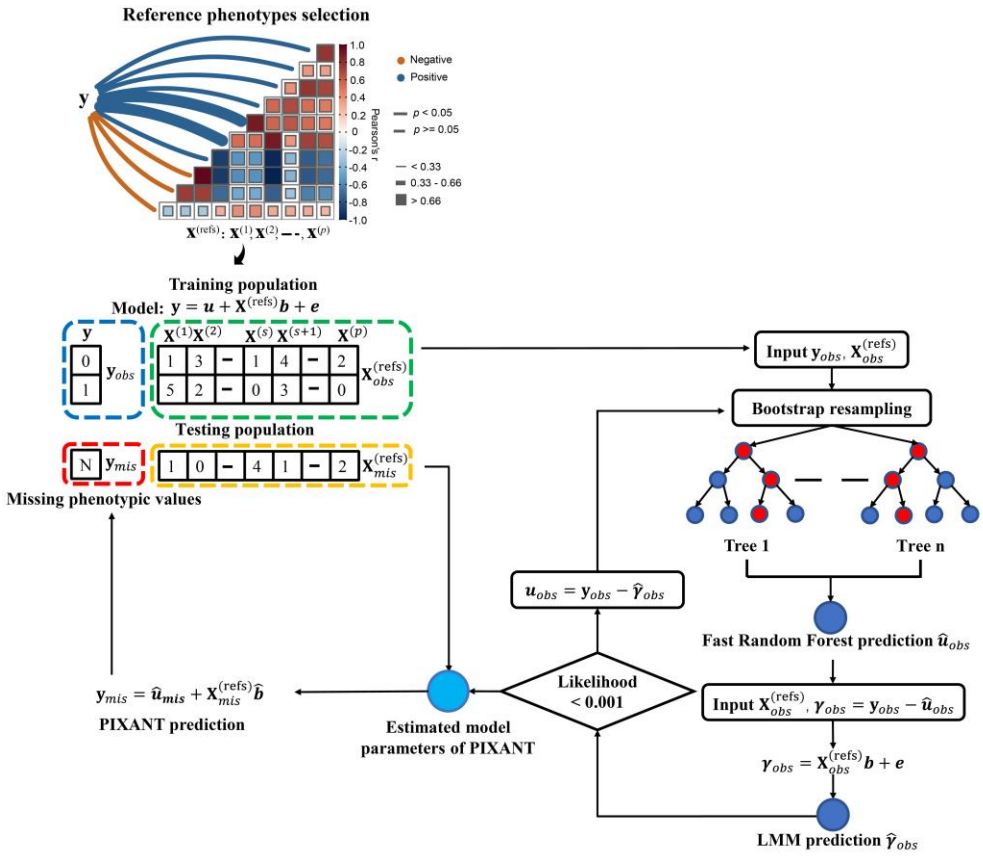


Figure 1 Flowchart of the PIXANT used in this study.

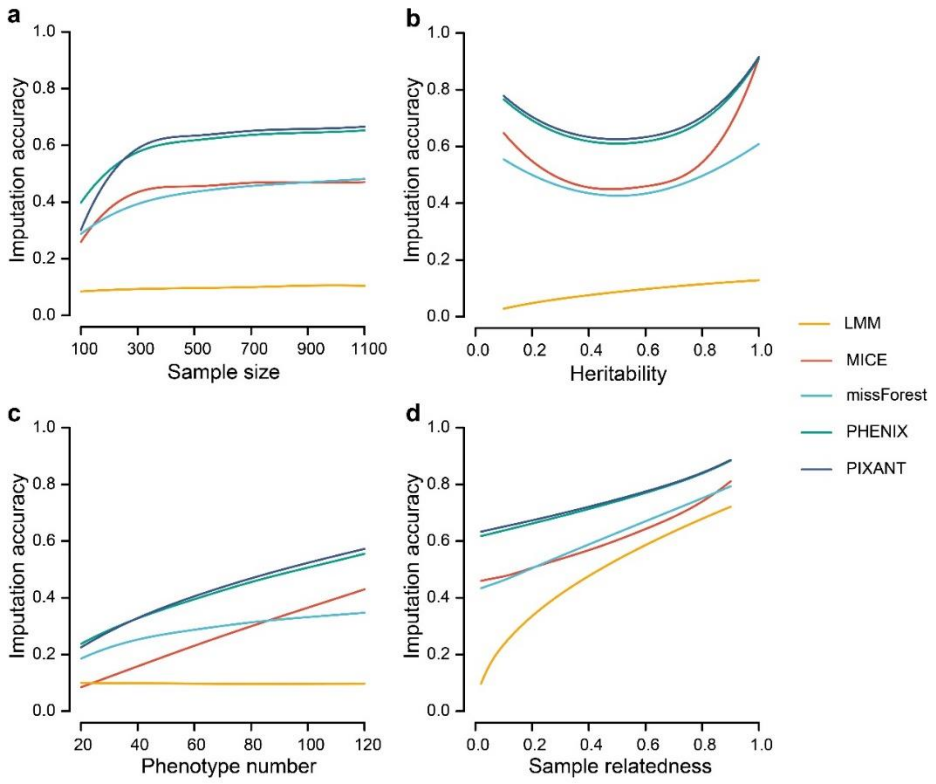


Figure 2 Simulation study on imputation accuracy of different methods under different levels of sample size (a), heritability (b), number of phenotype (c) and sample relatedness (d). PIXANT and PHENIX show the highest imputation accuracy in almost all simulations, with PIXANT superior to or comparable to the next accurate method, PHENIX.

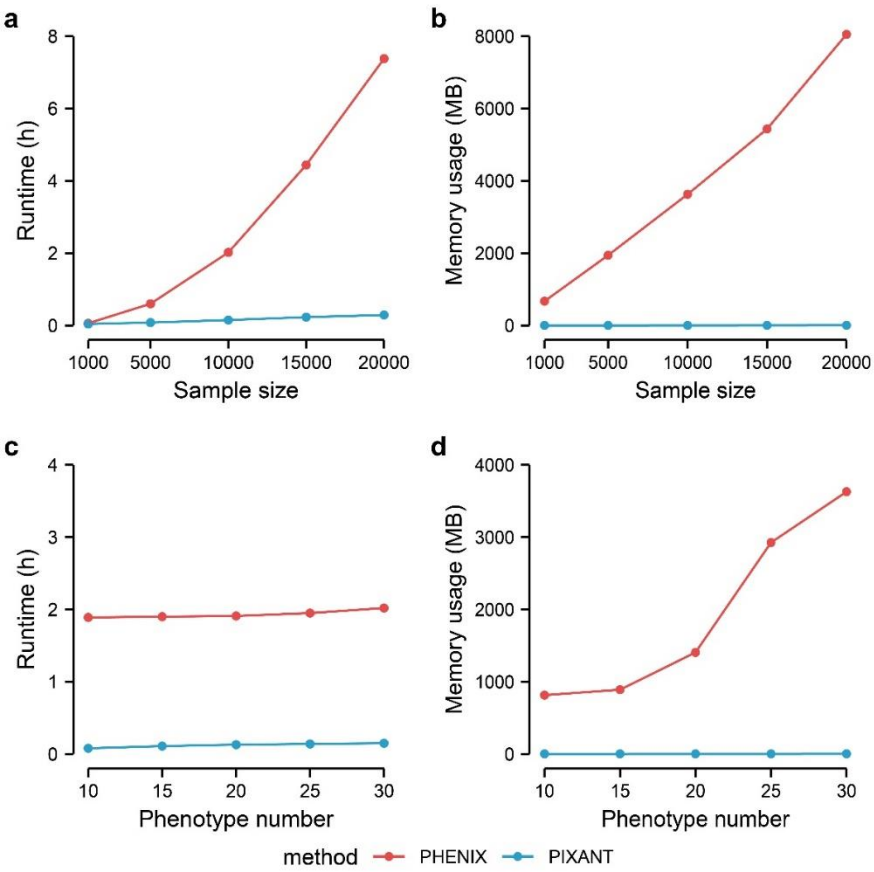


Figure 3 Comparison of runtime and memory usage between PHENIX and PIXANT with large sample size. For large datasets, PIXANY is significantly less demanding on computer resources than PHENIX; and the runtime and memory usage of PIXANT increase approximately linearly with the sample size and phenotype number, while PHENIX shows a non-linear increase (except for runtime with phenotype number).

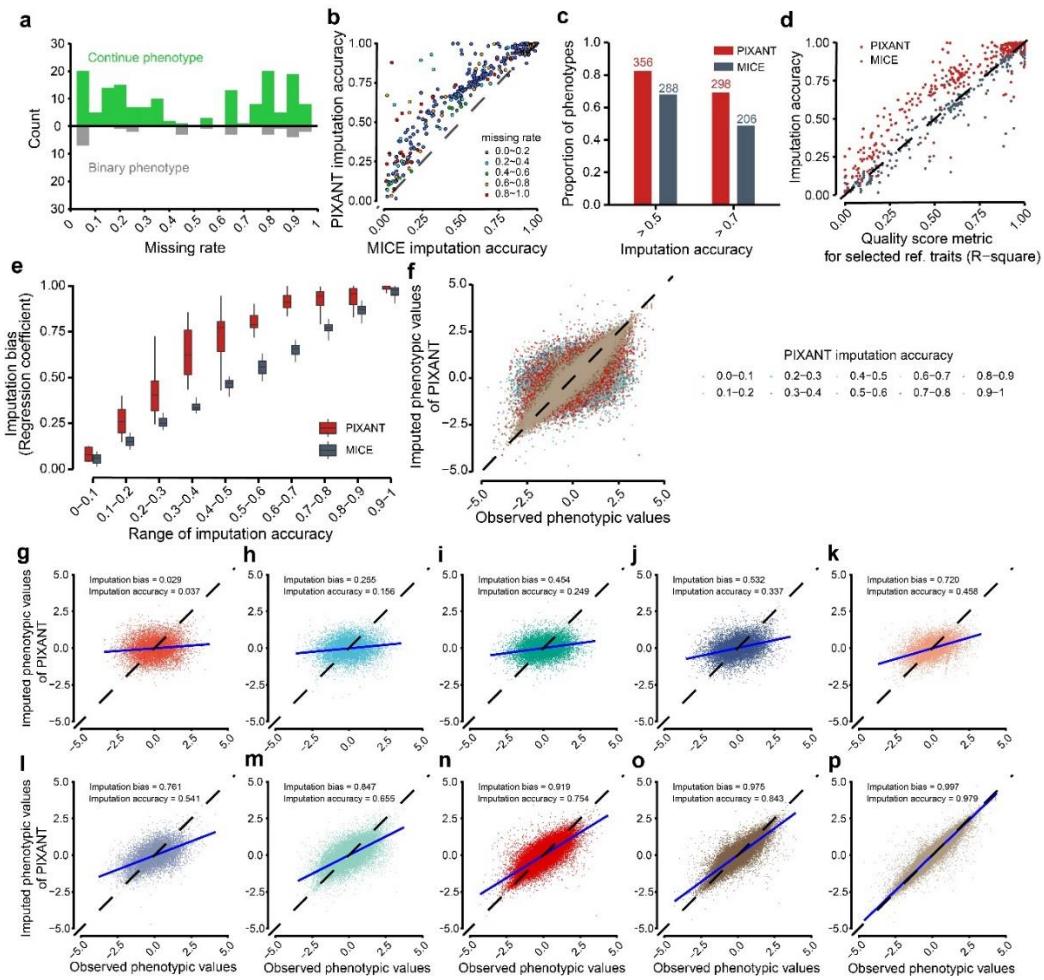


Figure 4 The comparison of the performance of PIXANT and MICE in imputation of 425 UKB phenotypes. (**a**), Distribution of phenotypic missing rates. (**b**), The imputation accuracies of PIXANT and MICE at different phenotypic missing rates, showing that PIXANT's imputation accuracy for each UKB phenotype is generally higher than MICE. Phenotypic missing rate indicated by different colors has no obvious effect on imputation accuracy. (**c**), The number of the accurately imputed phenotypes by PIXANT and MICE, with imputation accuracy greater than 0.5 and 0.7, respectively. (**d**), The impact of quality score metric R^2 for selected reference phenotypes on imputation accuracies of PIXANT and MICE. The imputation accuracy of MICE is almost linear to R^2, while the accuracy of PIXANT is a trajectory curve. (**e**), Comparison of the imputation bias of PIXANT and MICE, the imputation bias of PIXANT is clearly better (closer to 1) than that of MICE at different range of imputation accuracies. (**f-p**), The imputation bias of PIXANT is shown as the observed phenotypic values (x-axis) versus imputed phenotypic values (y-axis) across all imputation accuracy ranges (**f**) and in ranges 0-0.1,0.2-0.3, etc., up to 0.9-1 (**g-p**).

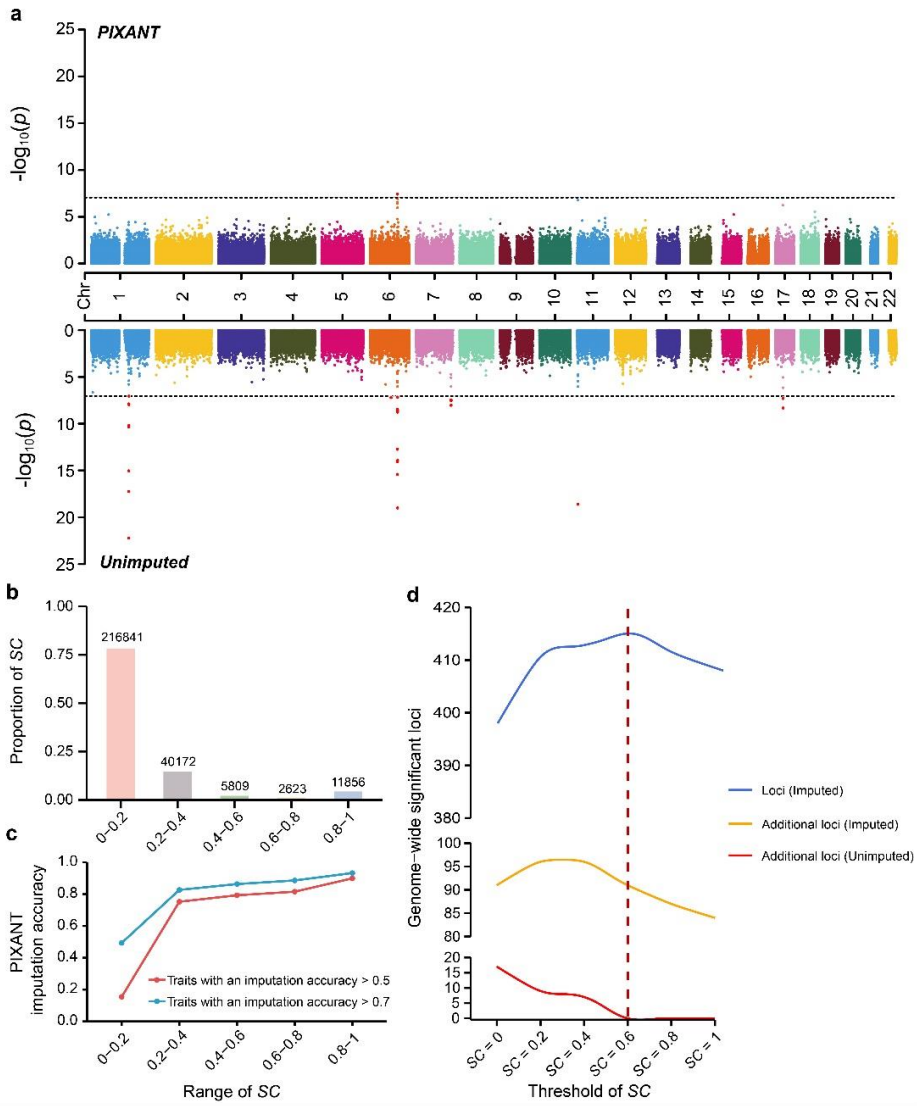


Figure 5 The relationship between *SC*, power and imputation accuracy of PIXANT in UKB dataset. (**a**), GWAS of QT interval UKB phenotype with imputed (upper panel) and unimputed (lower panel) phenotypes. (**b**), The frequency of imputed phenotypes of QT interval under different *SC* thresholds, more than 75% of individual phenotypes are imputed with *SC* less than 0.2. (**c**), The imputation accuracy across 42 randomly selected UKB phenotypes under different *SC* thresholds, the *SC* is in good agreement with the imputation accuracy. (**d**), The GWAS power across 42 randomly selected phenotypes before and after phenotype imputations under different *SC* thresholds. The total number of GWAS loci after imputation (upper panel); the additional GWAS loci after imputation (middle panel); and the additional GWAS loci before imputation (bottom panel), which reflects the overlap of the GWAS loci before and after imputations, fewer numbers imply a high overlap. *SC* of 0.6 is the optimal threshold, leading to the highest power and the maximum overlap, while *SC* between 0.2-0.6 also seems to bring out sufficient power and overlap.

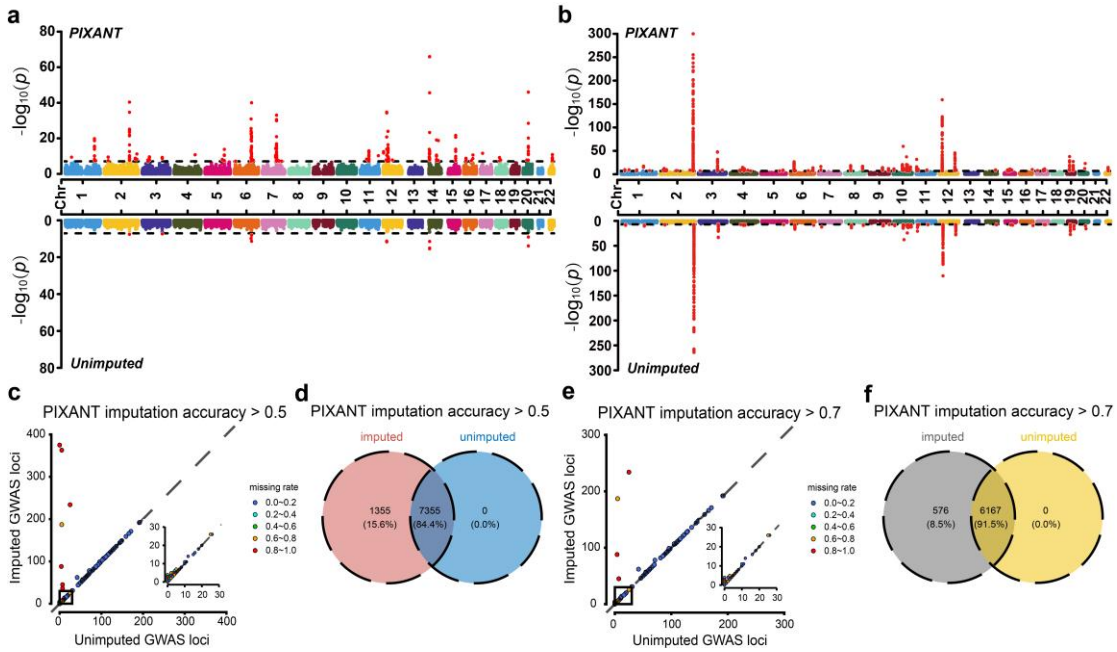
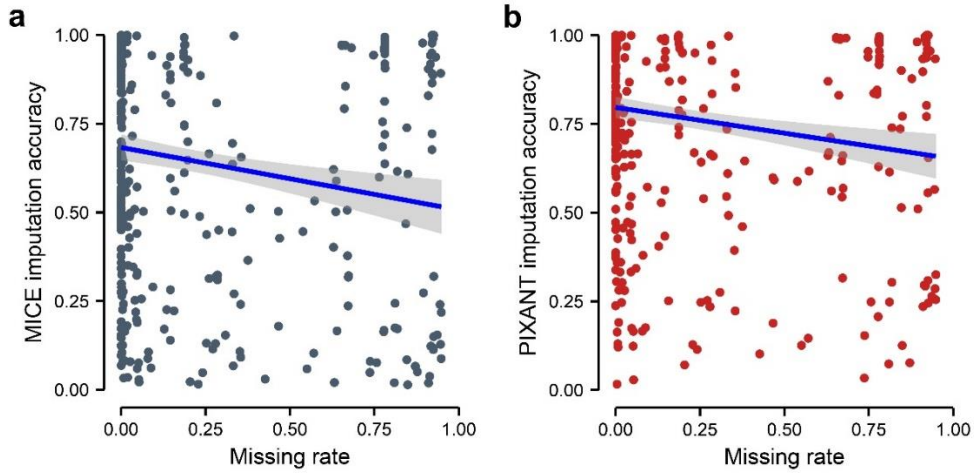


Figure 6 The GWAS of 425 UKB phenotypes before and after imputations. (**a**), Heart rate example for imputed (upper panel) and unimputed (lower panel) phenotypes. The dashed horizontal lines denote a significance threshold of 6.74 ($-\log_{10}(0.05/\text{total SNPs})$). After imputation, the GWAS signals significantly boosted and more GWAS loci are identified. (**b**), Direct bilirubin example. (**c**), Comparison of GWAS loci before and after imputation for 356 UKB phenotypes with imputation accuracy greater than 0.5, each point represents a phenotype, and different colors represent different phenotypic missing rates. The difference of the number of GWAS loci before and after imputation is more pronounced as the phenotypic missing rate increased. (**d**), Overlap of GWAS loci before and after imputation of 356 phenotypes. **(e)** and **(f)** have the same legend as **(c)** and **(d)**, respectively.



Supplementary Figure 1 The impact of missing rate on the imputation accuracies of MICE (a) and PIXANT (b). The x-axis represents the missing rate of phenotypes, and the y-axis represents imputation accuracies of PIXANT and MICE. Each point represents a phenotype.