OXFORD

Briefings in Bioinformatics, 22(6), 2021, 1–9

https://doi.org/10.1093/bib/bbab326
Problem Solving Protocol

RefRGim: an intelligent reference panel reconstruction method for genotype imputation with convolutional neural networks

Shuo Shi, Qiheng Qian, Shuhuan Yu, Qi Wang, Jinyue Wang, Jingyao Zeng, Zhenglin Du and Jingfa Xiao

Corresponding authors: Jingfa Xiao. Tel.: +86-10-84097443; E-mail: xiaojingfa@big.ac.cn; Zhenglin Du. E-mail: duzhl@big.ac.cn

Abstract

Genotype imputation is a statistical method for estimating missing genotypes from a denser haplotype reference panel. Existing methods usually performed well on common variants, but they may not be ideal for low-frequency and rare variants. Previous studies showed that the population similarity between study and reference panels is one of the key factors influencing the imputation accuracy. Here, we developed an imputation reference panel reconstruction method (RefRGim) using convolutional neural networks (CNNs), which can generate a study-specified reference panel for each input data based on the genetic similarity of individuals from current study and references. The CNNs were pretrained with single nucleotide polymorphism data from the 1000 Genomes Project. Our evaluations showed that genotype imputation with RefRGim can achieve higher accuracies than original reference panel, especially for low-frequency and rare variants. RefRGim will serve as an efficient reference panel reconstruction method for genotype imputation. RefRGim is freely available via GitHub: https://github.com/shishuo16/RefRGim

Key words: genotype imputation; reference reconstruction; deep learning; genome-wide association study

Introduction

With current genotyping array or sequencing technologies, there is always a certain percentage of variant genotypes that cannot be detected, which seriously affects subsequent analysis. Thus, genotype imputation was introduced to predict missing genotypes, which can significantly boost the power of signal variants in genome-wide association studies [1, 2], provide a high-resolution view of a certain region in fine-mapping study, increase the chance to find causal single nucleotide polymorphisms (SNPs) [3] and facilitate meta-analysis for merging multiple genotype data from different genotyping arrays or sequencing depths [3, 4].

Current popular imputation tools, such as Beagle [5] and Impute [6], were proposed based on hidden Markov model or its extensions, which were used to model linkage disequilibrium, estimate recombination rates and search for comparable local pattern in references and then impute alleles in reference

Shuo Shi is a PhD candidate at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. She has been working in the field of human genotype imputation study.
Qiheng Qian is a PhD candidate at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.
Shuhuan Yu is a master candidate at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.
Qi Wang is a management trainee at Qujiang culture finance holding (Group) Co., Ltd, Xian, China.
Jinyue Wang is a PhD at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.
Jingyao Zeng is an assistant professor at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.
Zhenglin Du is a senior engineer at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. His research interest is focused on comparative genomics and bioinformatics.
Jingfa Xiao is a professor at the National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. His research interest is focused on genomics and bioinformatics.

Submitted: 7 June 2021; **Received (in revised form):** 21 July 2021

haplotypes into the samples to create imputed genotypes [1]. These methods usually performed well on common variants, nevertheless, it still remains a major challenge in the genotype inference for low-frequency and rare variants [7, 8]. Considering the important roles of these variants in human diseases and other associated studies, the improvement of imputation performance on those variants becomes more necessary [9]. Previous studies showed that genetic similarity between individuals in the reference panel and the study affects the imputation performance significantly [10–12]. Pasaniuc *et al.* have tried to optimize imputation performance in local region with a weighted reference population by splitting the reference genome [13]. Zhang *et al.* have focused on improving imputation accuracies within the European population by identifying the subset of European ancestry reference panel at a given size with maximal phylogenetic diversity [14]. In general, existing methods are mainly focused on the short window of the genome or the construction of internal reference panels for specific populations.

From last decades, a number of high-quality human reference panels have been constructed in different large-scale whole genome sequencing projects [15], such as the reference panel from the 1000 Genomes Project with over 80 million SNPs in 2504 individuals from 26 populations across the world [16], the UK10K reference panel with 27 million SNPs of 3781 European individuals [17], the Haplotype Reference Consortium panel with 39 million SNPs from 32 488 mix population individuals [18] and the CASPMI reference panel with 24.85 million SNPs from 597 Chinese individuals [19]. Among them, the reference panel from the 1000 Genomes Project used to be one of the most popular choices because of its large sample size, population diversity and open access. Nowadays, with the completion of more and more large-scale whole-genome sequencing projects, the proper selection of reference panels in an imputation study has become very crucial in practice.

In this paper, we presented RefRGim, a reference panel reconstruction method using convolutional neural networks (CNNs), aiming to improve imputation performance by providing a more genetic similar reference panel for study individuals. RefRGim can rank reference haplotypes by its genetic similarity with the study individuals, select the most comparable haplotype group for each study individual and organize them into a new reference panel for study data specifically.

Materials and methods

In recent years, with the rise of the CNNs, image recognition and text recognition have been evolving fast and significantly [20]. CNNs are the modified versions of artificial neural networks in which the convolution is a specialized kind of linear operation of local variables. In addition, transfer learning is able to reduce the pressure of computational resource consumption when using deep learning algorithms like CNNs [21]. With a similar idea, deoxyribonucleic acid (DNA) sequence recognition and comparison can also be accomplished by CNNs, with quantized alleles and proper convolutional (CONV) architecture.

RefRGim was built with a set of CNNs, which consisted of five CONV layers, three pooling layers inserted in-between the first four CONV layers and one spatial pyramid pooling (SPP) layer located behind the fifth CONV layer, followed by two fully connected layers. The networks take an ordered genotype list from individuals as input, and they output the most genetic comparable haplotype group for each individual (Figure 1A).

Training data

We downloaded haplotype references of the 1000 Genomes Project (1KGP), which includes 2504 individuals from 26 populations in 5 superpopulations across the world: African (AFR), American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) (Supplementary Table S1 available online at http://bib.oxfordjournals.org/). We divided them by its populations and generated 26 haplotype groups. Each group was comprised about 100 high-quality phasing individuals. We performed principle component (PC) analyses on these individuals and filtered those who distributed away from other individuals in the same population (Supplementary Figure S1 available online at http://bib.oxfordjournals.org/). To generate high-quality variant set for model training, we only included common variants [minor allele frequency (MAF) > 0.05] on genotyping arrays and filtered those were not significantly different among populations (Chi-squared test: $P < 1\text{e-}8$). Finally, 2418 samples with 1.1 million SNPs were left for model training. Then, we transformed the four bases-ATCG to mathematical values ($A = 0$; $T = 0.1$; $C = 0.3$; $G = 0.7$). These assignments were proposed to guarantee each base-pair with a unique value, ranging from 0 to 1.4 (Figure 1A). Variants were sorted into a list by its positions on chromosomes and every 4000 SNPs of the list was split into fragments, which ranged from 6 to15 Mb in length on chromosomes.

Reference panel reconstruction model

The CONV layer of CNNs can capture the local sequence linkage information by using the kernel, which is a parameter matrix that moves over the input base-pair value sequence, performs the dot product with its subset region, extracts linkage features from original sequence and gets the output as the matrix of dot products for the next layer. The CONV process is performed as follows:

$$\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\mathrm{N}),$$

$$\mathrm{W} = (\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3, \ldots, \mathrm{w}_n),$$

$$\mathrm{y}_\mathrm{i} = \sum_{v=1}^{n} \mathrm{w}_v \cdot \mathrm{x}_{\mathrm{i}+v-1} + b(1 \le \mathrm{i} \le \mathrm{N}),$$

where X is one-dimensional input vector shaped 1 × N, W is the vector of the kernel shaped 1 × n, b is bias term, y_i is an output value summed from multiply results of a length n sequence fragment and a length n kernel. We set our kernel shaped 1 × 15, and the stride of each sliding is one base. To capture sequence information from different aspects, we used 32 different kernels to learn the sequence at each CONV layer.

The output data from a CONV layer were batch-normalized for stabilizing the learning process, where

$$\mu = \tfrac{1}{N} \textstyle\sum_{\mathrm{i}=1}^{N} \mathrm{y}_\mathrm{i} \text{ //batch mean,}$$
$$\sigma^2 = \tfrac{1}{N} \textstyle\sum_{\mathrm{i}=1}^{N} (\mathrm{y}_\mathrm{i} - \mu)^2 \text{ //batch variance,}$$
$$\hat{\mathrm{y}}_\mathrm{i} = \tfrac{\mathrm{y}_\mathrm{i} - \mu}{\sqrt{\sigma^2 + \varepsilon}} \text{ //normalize,}$$
$$\mathrm{z}_\mathrm{i} = \gamma \hat{\mathrm{y}}_\mathrm{i} + \beta \text{ //scale and shift,}$$

where $\varepsilon$is a constant added to the batch variance for numerical stability, γ and β are learnable parameters to scale and shift the input that are updated during training iteration. Then, the data were activated using the rectified linear unit (ReLU: $f(x) = \max(0, x)$) function to add non-linearity into the network.

Pooling layer after each CONV layer was used to reduce the number of parameters and to control overfitting. Pooling layers

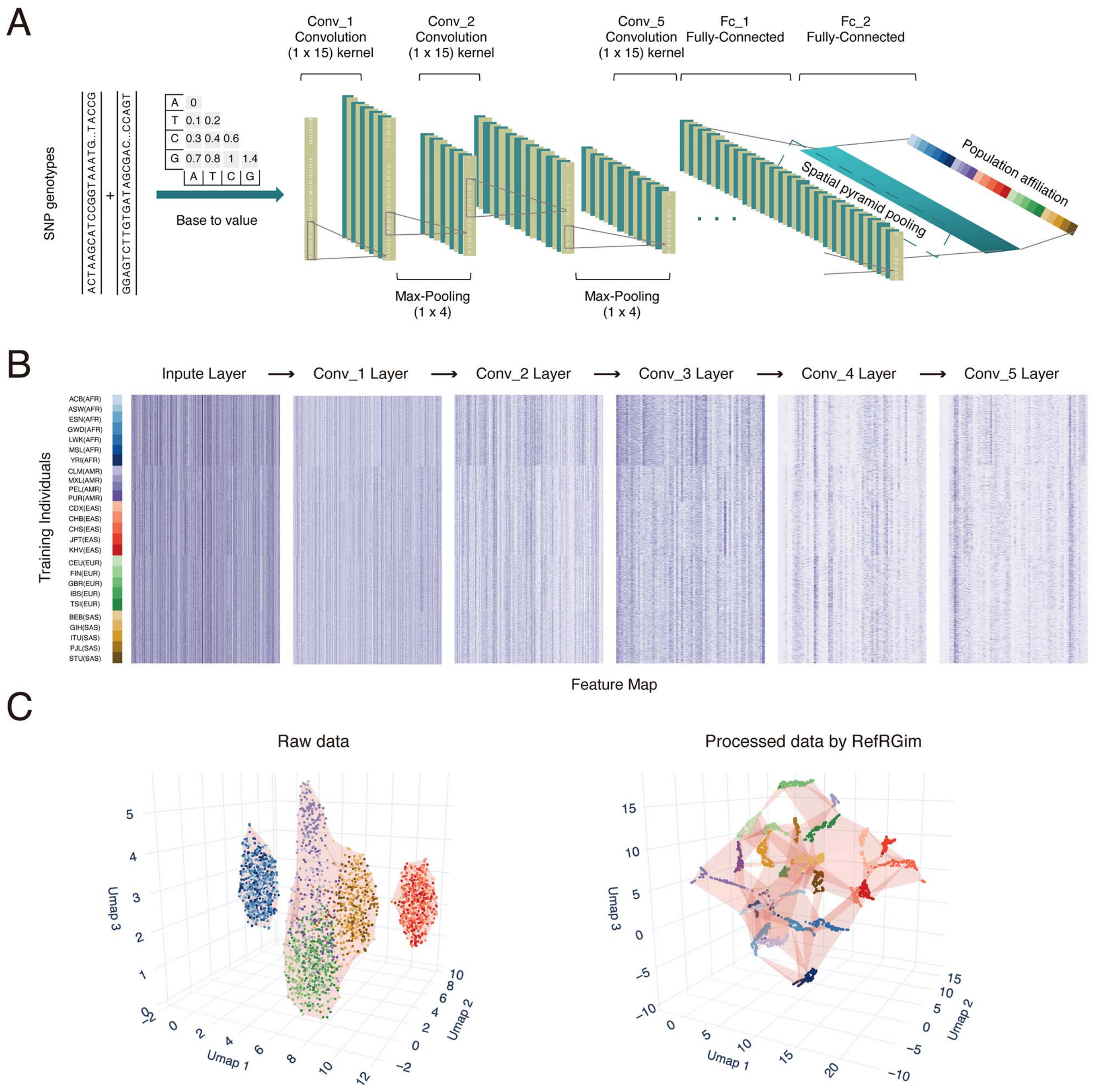

Figure 1. (**A**) RefRGim's model architecture. The CNNs model comprises of five CONV layers, three pooling layers, one SPP layer and two fully connected layers. It takes mathematical values transformed from nucleotide bases as input and outputs the most similar haplotype group for each input sample. Different populations are represented by different colors, while those in the same superpopulation are displayed in the same color series (AFR: blue, AMR: purple, EAS: red, EUR: green and SAS: yellow). (**B**) The feature maps of 1KGP individuals in CNNs layers. If there are multiple feature maps in a CONV layer, only the first one is plotted. (**C**) Dimension reduction of raw and processed genotype data of 1KGP individuals using uniform manifold approximation and projection (UMAP). UMAP 1–3 of individuals are plotted as a dot on a three-dimensional (3D) coordinate. A 3D set of light-red triangles with vertices, given UMAP 1–3, are drawn using plotly.mesh3d to trace the 3D positional relationship between individuals.

were proposed for downsampling the feature map by summarizing the presence of features in patches of it. We used max pooling with filter sized 1 × 4 and stride shaped 1 × 4, discarding 75% of the activations.

SPP layer is a special pooling layer, which is usually added to the transition of CONV layers and fully connected layers [22]. SPP adopts spatial pyramid structure as it is not one single pooling procedure but multiple pooling layers with different scales. It can generate a fixed-size output regardless input data size. We set a four-level pooling layer after the fifth CONV layer and our bin sizes were of 1 × 6, 1 × 4, 1 × 2, 1 × 1. The outputs from different pooling layers were merged into one-dimensional data sized 1 × 13 at each feature map. With five CONV layers and four pooling layers, the features of input variant sequence were extracted, and the differences between different populations were emerged step by step (Figure 1B).

The output data of all feature maps (shaped 1 × 13 × 128 in our model) from the SPP layer were merged into one-dimensional data shaped 1 × 512 and then that were fed into fully connected layers. Variables in a fully connected layer have full connections to all activations in the previous layer, which is similar as regular neural networks. This approach has been proven very effective in image recognition and classification [23]. We set the dropout as 0.5 to make the nodes in the network generally more robust to the inputs. Through two fully connected layers at the end of our model, RefRGim reached to the final classification decision on haplotype group using the softmax function:

$$P_i = \sigma(\overrightarrow{a})_i = \frac{e^{a_i}}{\sum_{j=1}^{M} e^{a_j}} (1 \le i \le M),$$

where M is the number of haplotype groups, $a_i$values are the output elements from the last fully connected layer and the term on the bottom of the formula is the normalization term which ensures that all the output values of the function sum to 1, constituting a valid probability distribution.

Model pretraining

We trained RefRGim using the cross-entropy (negative log likelihood) as the loss function, where we focused on minimizing:

$$\text{loss function} = -\frac{1}{K}\sum_{c=1}^{K}\sum_{i=1}^{M} S_{ci} \cdot \log(P_{ci}),$$

where K is sample number; M is haplotype group number; S_{ci} equals 1 when sample c belongs to class i, otherwise S_{ci} equals 0; P_{ci} is the predicted probability of sample c belongs to class i deduced by the softmax function in the model. Training process was carried out using the Adam optimizer with a learning rate of 1e-5. To accelerate the model's learning rate and reduce computational resource consumption, we used batch learning with batch size 200 at each iteration. We performed crossvalidation at every 50 iterations and stopped the iteration while both training accuracy and testing accuracy of population prediction were greater than 0.99 (Supplementary Figure S2 available online at http://bib.oxfordjournals.org/). After the training process, the model was able to classify 26 groups apart from each other effectively (Figure 1**C**).

Model retraining in practice

As a user-oriented tool, RefRGim was supposed to handle different variant data as input. It would cost enormous computing resource and time if we retrain the CNN model all over again every time when it met a new SNP set. Hence, we introduced transfer learning into our method by reloading parameters and architecture of pretrained CNNs model and only retraining parameters of fully connected layers in practice. As for the unique SNPs in the pretrained model relative to study data, the model would change the SNP genotype to the same as human genome reference. It can achieve rapid progress and improve the performance when retraining the CNNs model for new SNP sets. After retraining step, the model can be used to recommend haplotypes for study individuals. We also provided original code for users to train the model with their own reference panels other than 1KGP.

Study-specified reference panel reconstruction

After population classification, each study individual was assigned with a most genetic comparable haplotype group, which was used to organize into the study-specified reference panel (SSRP). To avoid the influence of population diversity of study individuals, RefRGim divided these haplotype groups by their superpopulations and organized them to different SSRPs.

Evaluation and computation resource

Here, we evaluated the performance of RefRGim by comparison of sequence and population similarity of individuals in the study and SSRP and the improvements of imputation accuracy with the SSRP using Begale5.1 [5], Impute2 [6] and Minimac4 [24], respectively. In the imputation process, we divided testing individuals by their superpopulation classification and performed imputation separately. The testing dataset contains 199 individuals from 43 countries in 5 major continental groups: America, Central Asia Siberia, East Asia, South Asia and West Eurasia (Figure 3**A**) [25].

RefRGim was tested on Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz. The running time positively correlated with the missing number of intersected SNPs between the study and reference in the retraining process and the complexity of study populations in abstracting and reconstructing new reference panel process. For instance, when the missing SNP number is 3500 (87.5%), the model retraining time needs 17 min. When the missing SNP number is 2000 (50%), the time is 8 min. As for time cost owing to complexity of study populations in reconstructing new reference panel process, it was less than 120 min based on the reference panels of 1KGP.

Implementation

RefRGim was implemented using python3 with Tensorflow 1.9 (https://www.tensorflow.org/) and Numpy 1.14.3 (https://numpy.org/). All relevant codes are available at https://github.com/shishuo16/RefRGim.

Results

Performance on sequence similarity comparison

First, we evaluated our model's ability to capture sequence information by comparing the consistency of cluster result based on identity by state (IBS) score and population classification result by our model of 63 testing individuals, using a 12.4 Mb length sequence fragment (chr1: 41.0–53.4 Mb). This fragment was selected automatically by our model as this region has the largest intersected SNP number between the input data and pretrained SNP set of our model. IBS score was used to describe two identical segments or sequences of DNA in genetics and it was negative with the similarity degree of two sequences. The five main branches of hierarchical clustering tree based on IBS score were mostly corresponding with different superpopulations classification results, EUR, AMR, SAS, EAS and SAS, respectively (Figure 2**A**). Under the main branch, individuals who had smaller IBS scores with each other were classified into same or adjacent populations. The population affiliations of different individuals by RefRGim were correlated with their IBS scores, which demonstrated its capability of sequence information collection and inference. These 63 testing individuals come from five different areas around the world. Our model compared their sequence with the sequences of different haplotype groups in reference panel and divided the testing individuals to different haplotype groups. The Sankey plot of the original sampling location and haplotype group affiliation showed most individuals were parted to similar superpopulations (Figure 2**B**).

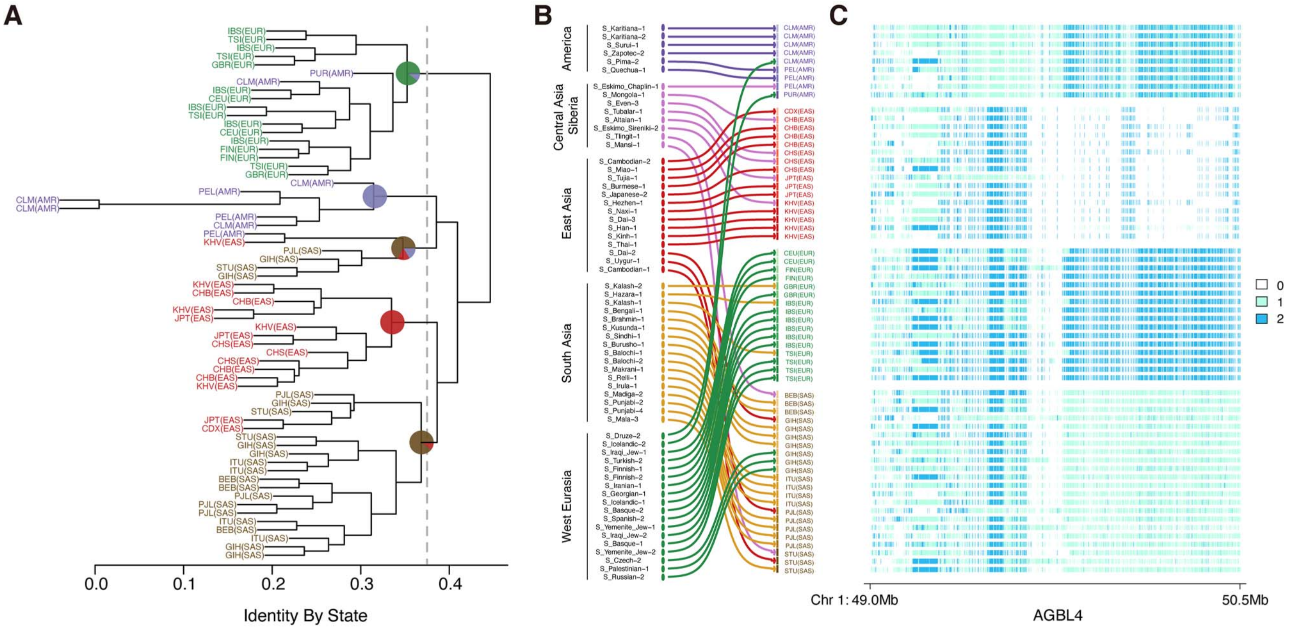

Figure 2. (**A**) Hierarchical clustering tree of testing individuals based on IBS score using 'complete' agglomeration method. Pie plot represents the percentages of superpopulations of testing individuals classified to in this branch. (**B**) Sample-paired Sankey plot of original sampling location and predicted population group by RefRGim. (**C**) Heatmap of common variant genotypes of corresponding individuals in (**B**) in AGBL4 gene region, where 0, 1 and 2 are the alternative allele dosages.

However, some individuals were classified into other superpopulations, like West Eurasia to AMR or SAS. As for Central Asian Siberian testing individuals, which did not have corresponding individuals in the reference panel (Figure 3A), RefRGim classified them to AMR, EAS and SAS superpopulations separately. We further displayed part of sequence information in this fragment, AGBL4 gene region, in which none, multiallelic or low-frequency (MAF < 0.05) variants sites were omitted and 1336 single nucleotide mutation sites were left to demonstrate [26]. The figure showed testing individuals who were divided into the same superpopulation had similar sequence base pattern (Figure 2C). West Eurasian testing individuals who were classified as AMR or SAS were more similar in sequence base to individuals with AMR or SAS compared to EUR, same as Central Asian Siberian testing individuals.

Performance on population identification

First, we performed principal components analysis on variants of 199 testing individuals and the 1KGP individuals to demonstrate the extent of their population distribution (Figure 3B). The plot of PCs 1 and 2 showed that most of the testing and 1KGP individuals who belonged to similar superpopulations were clustered together. American individuals in the testing set were distributed with AMR individuals in reference. Most of the West Eurasian testing individuals were distributed with EUR individuals except that fewer individuals sporadically gathered with AMR and SAS individuals in reference, which might be owing to lack of corresponding European samples in 1KGP (Figure 3A). For Central Asia Siberia, which did not have corresponding samples in 1KGP at all, they were gathered with AMR, EAS, EUR and SAS individuals in 1KGP.

We used RefRGim to perform reference reconstruction on testing individuals, and each individual was assigned with a most genetic similar haplotype group. By comparing RefRGim's classification result in pie plot with the prior information about population, we found there was a high degree of agreement between them. For 20 American individuals, RefRGim classified 12 of them to PEL (Peruvians from Lima) group, 7 to MXL (Mexican Ancestry from Los Angeles) group and 1 to CLM (Colombians from Medellin) group. As for 71 West Eurasian testing individuals, RefRGim classified 54 of them to EUR superpopulations, including 20 TSI (Toscani in Italia), 10 IBS (Iberian Population in Spain), 5 GBR (British in England and Scotland), 9 FIN (Finnish in Finland) and 10 CEU (Utah Residents (CEPH) with Northern and Western European Ancestry), 13 to AMR superpopulations and 4 to SAS superpopulations, which was also consistent with prior known population information, demonstrating the ability of population identification of RefRGim.

Considering the uncertainty of SNP size provide by user, we then evaluated the stability of population identification on different percentages of intersected SNP size with our CNNs model (Supplementary Figure S3 and Supplementary Table S2 available online at http://bib.oxfordjournals.org/). The Sankey plot showed RefRGim was more robust on individuals from America and East Asia than other populations. For South Asia, the robustness of RefRGim was slightly lower and its sensitivity to SNP abundance in genome might be caused by the complex genetic background of South Asian testing individuals (Figure 3B). For Central Asia Siberia, there was some mixture between EAS and SAS. After further investigation on the countries of Central Asian Siberian individuals, we found most interactions happened on the Russian individuals. The lack of Russian samples in 1KGP might be the reason of unstable classification of these individuals. For West Eurasian individuals, the interactions mostly happened between AMR, EUR and SAS groups. The corresponding testing individuals were from 26 countries, including South Asian countries (Iran, Iraq and Yemen), counties adjacent to Asia (Armenia, Jordan and Georgia), Russia and some European countries. High intra-population complexity in West Eurasia and lack of corresponding samples in 1KGP might be the reason for the instability on population classification.

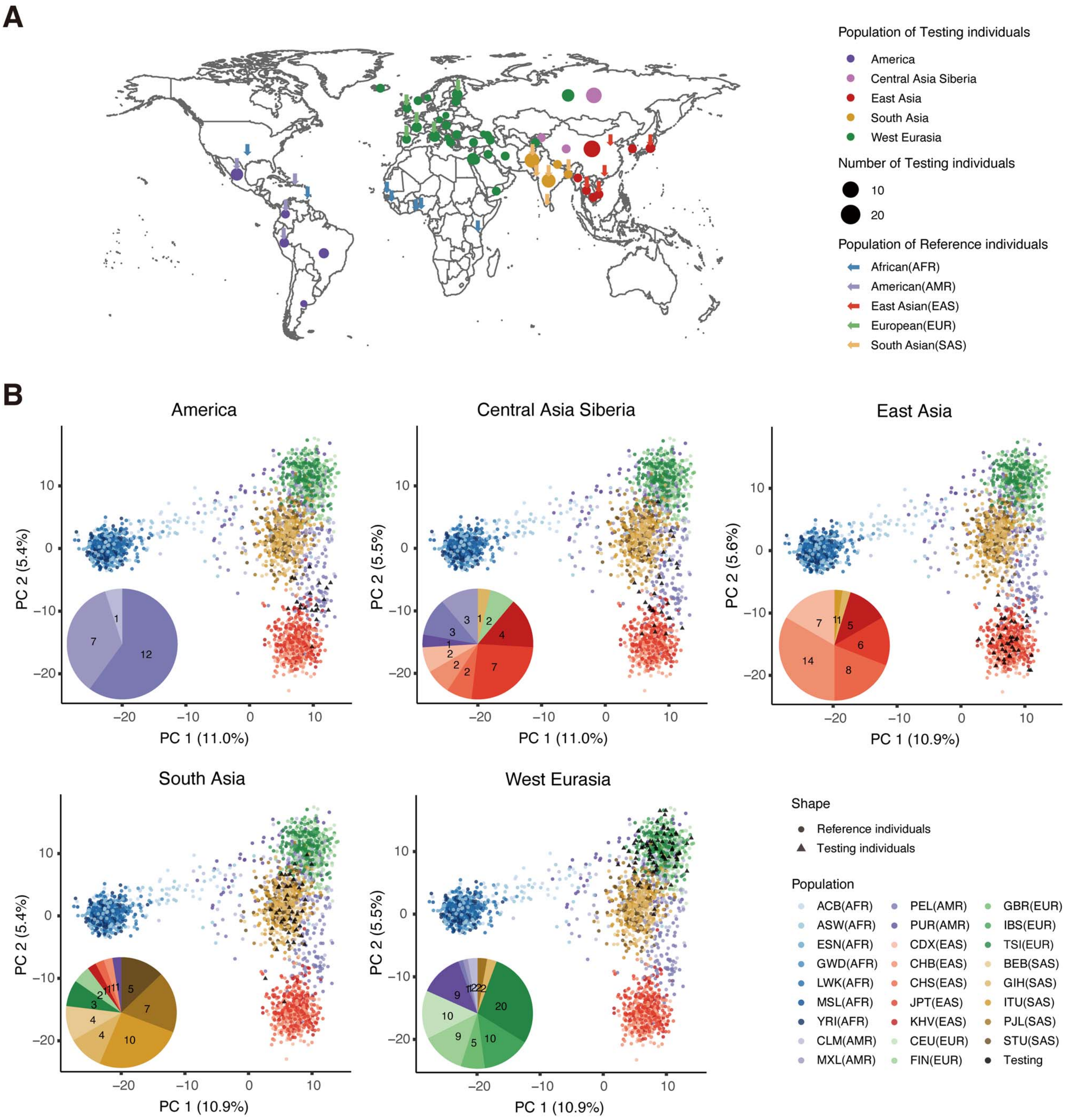

Figure 3. (**A**) Geographical locations of testing individuals and 1KGP individuals. Size of the dot is positive with the number of testing individuals. (**B**) PCs 1 and 2 of the reference and testing individuals. The reference samples in different populations are plotted by different colors and those in the same superpopulation are plotted in the same color series. The study samples are plotted by black triangle. The proportion of variance of each PC is texted. The pie plot shows the population classification results of the study samples from America, Central Asia Siberia, East Asia, South Asia and West Eurasia, deduced by RefRGim.

Performance of the SSRP on genotype imputation

RefRGim can rank reference haplotypes by its genetic similarity with study individuals and select the most comparable haplotype group for each study individual to organize them into SSRP. We further compared the imputation performances of testing individuals with original reference panels (1KGP and its subsets: AFR, AMR, EAS, EUR and SAS) and SSRP generated by RefRGim. The genotype imputation was performed using Beagle5.1 with default parameters and was measured using the squared Pearson correlation between the masked genotypes and the imputed allele dosages. The results showed that the best performance one of the original reference panels for different population testing individuals was different, and the SSRP always kept a high-level performance, especially for low-frequency and rare variants in all five testing populations (Figure 4). For America, using SSRP as reference panel, the average imputation accuracy of low-frequency variants increased from 0.89 to 0.94 and that of rare variants increased from 0.78 to 0.89, comparing with the

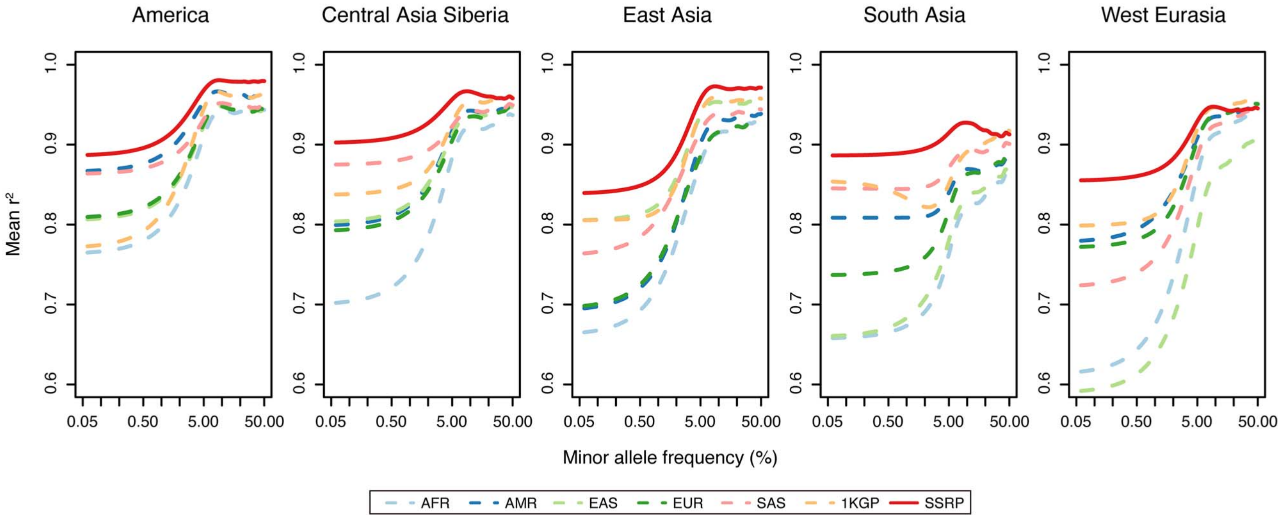

Figure 4. The imputation performance (mean r^2) comparison of testing samples with SSRP and six reference panels (1KGP original reference panels and its subsets of AFR, AMR, EAS, EUR and SAS) using Beagle5.1. The *x*-axis represents the MAF of imputed variants in 1KGP.

1KGP reference panel. For West Eurasia, the imputation accuracy of low-frequency variants and rare variants was improved from 0.86 to 0.90 and from 0.8 to 0.86, respectively. In addition, we have also evaluated the imputation performance using Impute2 and Minimac4, respectively, and observed the same trend in SSRP performance as Beagle 5.1 (Supplementary Figure S4 available online at http://bib.oxfordjournals.org/).

As we know, the population identification of our model is slightly different with various percentages of intersected SNP size between the input data and the model. Therefore, we also evaluated the imputation performances of SSRP under different percentages of intersected SNP size. The results presented that the imputation accuracies with SSRP were still better than the original references even if the intersected SNP size was decreasing (Supplementary Figure S5 available online at http://bib.oxfordjournals.org/).

Discussion

In last decades, a lot of haplotype reference panels had been generated, aiding to accomplish the high-quality genotype imputation process for future studies [17, 27, 28]. Several studies have shown the effect of the reference panel choice on imputation performance. Therefore, the selection of reference panels for specific studies of genotype imputation is inevitable and critical.

In this article, we presented a new approach called RefRGim to construct SSRP for each input data from the existing reference panels. RefRGim was designed to compare the sequence similarity between individuals in study and original reference panels and to organize the reference haplotypes which had the top sequence similarities with study individuals to a new reference panel. Compared with existing methods, RefRGim not only captured genetic information of local sequences but also estimated global genetic similarity to construct a universal reference panel for study samples. RefRGim was a common approach that may cope with the genotype imputation studies from different populations. Because of the high sequence and population similarity between SSRP and testing individuals, SSRP was committed to more consistent allele distribution with the study data and to avoiding potential fake linkage events that may be involved in other populations. Using SSRP as reference, genotype imputation can achieve a higher level of performance for study individuals, especially for low-frequency and rare SNPs.

RefRGim was implemented with CNNs, which can capture the local sequence patterns and find most sequence-similar reference haplotypes for study individuals. It was accelerated by introducing transfer learning into the retraining process in practice [29]. The function of sequence similarity calculation and comparison was mainly accomplished by the kernels in the CONV layer in which multiple weighted parameter matrixes acted as learning cubes to slide on the genome sequence and transform pieces of sequence to values, step by step. These processes can capture and process sequence data by its base information along with its location information at the same time. CNN algorithm can utilize a fragment of sequence as input data instead of a single base. This is the advantage of CNN algorithm in sequence information capture field, compared with other traditional machine learning methods, like Regular Neural Network, Random Forest and Support Vector Machines [30]. In terms of model applications, RefRGim can help complete the retraining process quickly and accurately by adapting transfer learning algorithm on the last two layers to accommodate different SNP sets. This is another advantage of CNN algorithm, which has high program transferability in sequence base processing.

So far, we only included 1KGP reference panels as original references in our model. There were still some populations around the world not covered in current existing reference panels. In the future, we will collect and integrate more diverse reference panels in RefRGim to improve its performance. Fortunately, with the embedding of transfer learning function, our model does not need to be trained from the beginning every time it met new data. In an era of rapid growth of large-scale whole genome sequencing data year by year, the continually learning ability of our model will be more and more useful [15]. On the other hand, with the fast growing of high-quality human haplotype reference panels, selecting appropriate reference panels for genotype imputation will be a much more urgent problem. Currently, our approach only supports human genotype imputation studies. Based on the same imputation principles as human [31], RefRGim may be applied to other species with growing diverse reference panels in the future.

Key Points

- A reference panel reconstruction method with CNNs. RefRGim can provide SSRP to achieve high performance for genotype imputation, especially for low-frequency and rare SNPs.
- A population identification classifier based on local pattern and general similarity of sequences. The application of CNNs in this field shows good predictive ability and transportability with enlarged learning data in the future.
- An application scenario of CNNs in sequence recognition. The multiple kernels in CONV layer ensures to capture local base linkage pattern from different aspects, promising a comprehensive perception on sequence level.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Data Availability

The haplotype data used in this study are from the 1000 Genomes Project, which can be downloaded at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

Funding

Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38030400 to J.X.); National Natural Science Foundation of China (31970634 and 31771465 to J.X.); National Key Research Program of China (2016YFB0201702, 2017YFC0907503 and 2020YFC0848900 to J.X.); CAS Key Technology Talent Program (to Z.D.).

References

1. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
2. Spencer CC, Su Z, Donnelly P, *et al.* Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009; **5**: e1000477.
3. Chen F, Chen GK, Millikan RC, *et al.* Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. *Hum Mol Genet* 2011; **20**: 4491–503.
4. De Jager PL, Jia X, Wang J, *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 2009; **41**: 776–82.
5. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet* 2018; **103**: 338–48.
6. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
7. Zheng HF, Rong JJ, Liu M, *et al.* Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One* 2015; **10**: e0116487.
8. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* 2011; **1**: 457–70.
9. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* 2017; **18**: 77.
10. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 2011; **12**: 703–14.
11. Shi S, Yuan N, Yang M, *et al.* Comprehensive assessment of genotype imputation performance. *Hum Hered* 2018; **83**: 107–16.
12. Huang L, Li Y, Singleton AB, *et al.* Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* 2009; **84**: 235–50.
13. Pasaniuc B, Avinery R, Gur T, *et al.* A generic coalescent-based framework for the selection of a reference panel for imputation. *Genet Epidemiol* 2010; **34**: 773–82.
14. Zhang P, Zhan X, Rosenberg NA, *et al.* Genotype imputation reference panel selection using maximal phylogenetic diversity. *Genetics* 2013; **195**: 319–330.
15. Stark Z, Dolman L, Manolio TA, *et al.* Integrating genomics into healthcare: a global responsibility. *Am J Hum Genet* 2019; **104**: 13–20.
16. Genomes Project C, Auton A, Brooks LD, *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.
17. Huang J, Howie B, McCarthy S, *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015; **6**: 8111.
18. McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48**: 1279–83.
19. Du Z, Ma L, Qu H, *et al.* Whole genome analyses of Chinese population and de novo assembly of a northern Han genome. *Genomics Proteomics Bioinformatics* 2019; **17**: 229–47.
20. Yamashita R, Nishio M, Do RKG, *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018; **9**: 611–29.
21. Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: *Proceedings of the 45th annual meeting of the association of computational linguistics - Prague*, Czech Republic: Association for Computational Linguistics. 2007; p. 440–447.
22. He K, Zhang X, Ren S, *et al.* Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 2015; **37**: 1904–16.
23. Mahmon NA, Ya'acob N. A review on classification of satellite image using artificial neural network (ANN). In: *2014 IEEE 5th Control and System Graduate Research Colloquium - Shah Alam*. Malaysia, U.S.: IEEE New York. 2014; p. 153–157.
24. Das S, Forer L, Schonherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016; **48**: 1284–7.
25. Mallick S, Li H, Lipson M, *et al.* The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 2016; **538**: 201–6.
26. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016; **32**: 2847–9.
27. Nagasaki M, Yasuda J, Katsuoka F, *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* 2015; **6**: 8018.

28. Deelen P, Menelaou A, van Leeuwen EM, *et al*. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of the Netherlands'. *Eur J Hum Genet* 2014; **22**: 1321–6.
29. Zhuang F, Qi Z, Duan K, *et al*. A comprehensive survey on transfer learning. *Proc IEEE* 2020; **109**: 43–76.
30. Khan A, Sohail A, Zahoora U, *et al*. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 2020; **53**: 5455–516.
31. Yang W, Yang Y, Zhao C, *et al*. Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Res* 2020; **48**: D659–67.