BMC Genomics

RESEARCH Open Access

Comparative analysis of genotype imputation strategies for SNPs calling from RNA-seq

Kaixuan Guo[1], Zhanming Zhong[1], Haonan Zeng[1], Changliang Zhang[1], Teddy Tinashe Chitotombe[1], Jinyan Teng[1], Yahui Gao[1] and Zhe Zhang[1*]

Abstract

Background RNA sequencing (RNA-seq) is a powerful tool for transcriptome profiling, enabling integrative studies of expression quantitative trait loci (eQTL). As it identifies fewer genetic variants than DNA sequencing (DNA-seq), reference panel-based genotype imputation is often required to enhance its utility.

Results This study evaluated the accuracy of genotype imputation using SNPs called from RNA-seq data (RNA-SNPs). SNP features from 6,567 RNA-seq samples across 28 pig tissues were used to mask whole genome sequencing (WGS) data, with the Pig Genomic Reference Panel (PGRP) serving as the reference panel. Three imputation software tools (i.e., Beagle, Minimac4, and Impute5) were employed to perform the imputation. The result showed that RNA-SNPs achieved higher imputation accuracy (CR: 0.895 ~ 0.933; r^2: 0.745 ~ 0.817) than SNPs from GeneSeek Genomic Profiler Porcine SNP50 BeadChip (Chip-SNPs) (CR: 0.873 ~ 0.909; r^2: 0.629 ~ 0.698), and lower accuracy in "intergenic" regions. After imputation, quality control (QC) by minor allele frequency (MAF) and imputation quality (DR^2) could improve r^2 but reduce SNP retention. Among software, Minimac4 takes the least runtime in single-thread setting, while Beagle performed best in multi-thread setting and phasing. Impute5 takes up minimal memory usage but requires the maximum runtime. All tools demonstrated comparable global accuracy (CR: 0.906 ~ 0.917; r^2: 0.780 ~ 0.787).

Conclusions This study offers practical guidance for conducting RNA-SNP imputation strategies in genome and transcriptome research.

Keywords Genotype imputation, Pig, RNA-seq

Background

RNA-seq is a widely adopted high-throughput technology employed for transcriptome profiling [1, 2]. It has largely replaced earlier gene expression profiling methods, such as DNA microarrays, owing to its numerous advantages, including reduced background noise, enhanced resolution, reduced sample requirements, and an expanded dynamic range [3]. Raw data from an increasing number of RNA-seq experiments are now stored in public databases, with RNA-seq sample availability expanding exponentially [4]. RNA-seq, which has been demonstrated to yield reliable genotypes from reads [5], has emerged as the preferred method

*Correspondence:
Zhe Zhang
zhezhang@scau.edu.cn
[1]State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

for transcriptome expression studies, outperforming traditional gene chips [6]. This technology has been extensively applied to analyze gene expression patterns across diverse organisms, including plants, animals, and humans. It provides critical insights into the genetic mechanisms underlying phenotype determination [7], disease development [8], and responses to environmental changes [9], among other biological processes. RNA-seq enables the integration of SNP data with gene expression profiles to explore variations influencing gene expression levels. Common methods for identifying eQTLs include genome-wide association study (GWAS) mapping and allele-specific expression (ASE) analysis, both of which leverage RNA-seq data to reveal regulatory relationships [10, 11]. These methods have been extensively validated in the Genotype-Tissue Expression (GTEx) project series [12, 13, 14] in humans. Recent studies have also been conducted on farm animals, including cattle [15], pigs [16], and chickens [17].

RNA-seq data has limitations in detecting genetic variations in non-transcribed regions or regions with low expression levels [18]. However, these variations may be critical for understanding specific traits. In contrast, WGS provides a more comprehensive and expansive view of the genome, encompassing both coding and non-coding regions, including regulatory elements, intergenic regions, and structural variants [19]. Due to expression variability across tissues, RNA-seq detects fewer variants than DNA-seq. For example, studies on pigs detected 182,039 variants from 189 transcriptomes [20]. Similarly, studies in cattle detected 100,734 variants from 7 transcriptomes [21], and 68,094 variants from 29 transcriptomes [22]. In sheep, 120,049 variants were detected from 8 transcriptomes [23]. These counts are substantially lower than those typically identified in similar-sized cohorts using WGS data, which can detect approximately 17.4 million variants [24].Consequently, RNA-seq data frequently face challenges related to incomplete or missing genotype information, potentially undermining the accuracy and reliability of downstream analyses. Although RNA-seq data have limitations, WGS data alone cannot provide information on gene expression or functional relevance. These insights can only be obtained through the integration of transcriptomic data when studying gene expression and regulatory mechanisms [25]. To resolve missing genotype information in RNA-seq data, RNA-SNPs can be imputed to higher density through the use of large reference panels, such as the PGRP [16] or the 1000 Bull Genomes Project [24]. However, the marked depletion of non-coding variants in typical RNA-seq datasets results in less reliable imputation for variants distant from transcribed regions. Filtering imputed variants by applying criteria such as MAF and imputation information score preserves high-quality variants. Recent studies on RNA-SNPs genotype imputation commonly used Beagle software. However, the quality control metrics after imputation differ considerably across the studies. For example, cattle studies used the criteria MAF ≥ 0.05 & DR^2 ≥ 0.8 [15]. In pigs, the filters MAF ≥ 0.05 & DR^2 ≥ 0.85 were applied [16], while in chickens, a missing rate ≤ 0.6 was used [17]. For sheep, MAF > 0.05 and R^2 > 0.4 (as determined by the imputation quality index from Minimac3) were used [26]. In ducks, MAF > 0.005 & DR^2 ≥ 0.8 & call rate ≥ 0.9 were selected as criteria [27]. Although previous studies have evaluated the imputation performance of different software in humans [28], pigs [29], cattle [30], and fish [31] these findings do not provide a reliable reference for selecting imputation strategies based on RNA-SNPs due to the differences in SNP sources. Nonetheless, these studies can provide valuable insights for selecting imputation accuracy evaluation metrics, such as concordance rate (CR), r^2, and imputation quality score (IQS).

In this study, we utilized SNPs previously called from RNA-seq data across 28 different pig tissues in PigGTEx, and the WGS data from 300 pigs. To mask SNPs in WGS data based on the features of RNA-SNPs, imputation was conducted using Beagle, Minimac4, and Impute5. We evaluated the imputation performance of RNA-SNPs including (1) the imputation accuracy in diverse post-imputation filtering criteria, and (2) the accuracy and the computational cost of three commonly used genotype imputation software. This research aims to provide a reference for selecting software and quality control metrics for RNA-SNPs genotype imputation in pigs under different scenarios.

Methods

Whole-genome sequencing data

This study utilized WGS data from 300 pigs as the gold standard, with a WGS depth of about 10×, comprising Duroc (n = 100), Yorkshire (n = 100), and Landrace (n = 100) breeds, sourced from the GigaScience GigaDB database [32]. Data filtering was performed using BCFtools v1.9 [33], where variants with a genotype quality score below 20 were excluded to remove low-confidence calls. Further filtering was conducted using PLINK v1.90 [34] to exclude SNPs with a call rate below 0.9. The resulting VCF file was refined by excluding variants that did not pass the established quality filters, retaining only biallelic SNPs. This procedure yielded a total of 300 samples with 23,897,690 high-quality SNPs across chromosomes.

To remove outliers within each breed's population, we used PLINK v1.90 to merge the genotypes from all imputed populations and conduct a principal component analysis (PCA) on the combined dataset. The top two principal components were then visualized using the

R package ggplot2 [35], enabling the identification and exclusion of outliers from the three breeds. PCA results (Figure S1) revealed that the three breeds were distinctly separated by the first two principal components (PCs), which explained 44.54% and 29.51% of the variance, respectively. All individuals within each breed were grouped into a single cluster. Subsequently, we employed Beagle v5.4 [36] to pre-phase both the reference panel and the imputed panels. Finally, we used conform-gt software (http://faculty.washington.edu/browning/conform-gt.html) to extract the overlapping loci between the imputed panels and the reference, correcting strand inconsistencies in A/T and C/G SNPs.

Reference panel

We utilized the PGRP from the PigGTEx project [16] as the reference panel. The PGRP dataset, covering all major pig breeds globally, includes WGS data from 1,602 pigs, representing major pig populations worldwide: Suidae but not *Sus scrofa* ($n = 45$), European wild boars ($n = 54$), European domestic pigs ($n = 855$), Asian wild boars ($n = 80$), and Asian domestic pigs ($n = 783$). The reference panel comprises 42,523,218 autosomal SNPs.

Masking SNPs to obtain the target panel

To mask the SNPs in WGS data based on the SNP features identified in RNA-seq data, we built on the research by Teng et al. [16] and used SnpEff v5.2c [37] to analyze 6,567 RNA-seq samples from 28 tissues as well as the WGS data, focusing on SNP sites within eight genomic regions: intron, intergenic, downstream, upstream, 3' untranslated region (3' UTR), 5' untranslated region (5' UTR), synonymous, and non-coding transcript (NC transcript). Specifically, "downstream" refers to regions located within 5,000 bases downstream of the stop codon, while "upstream" refers to regions within 5,000 bases upstream of the start codon. "Synonymous" refers to regions containing synonymous mutations in the genome. We calculated the number of SNPs on each autosome for each tissue and the proportion of SNPs in each of the eight genomic regions. Quality control was performed using box plots for the number of SNPs on each autosome and the proportion of SNPs in each genomic region across tissues, with outliers beyond the upper and lower limits of the box plots being removed. Based on the upper and lower limits of the box plots, the range for the number of SNPs on each autosome for each tissue and the range of the proportion of SNPs in each of the eight genomic regions was determined. Based on the SNP features in RNA-seq data across each tissue and autosome, BCFtools v1.9 was used to mask the WGS data, performing 10 times per tissue and autosome to obtain the target panel for imputation.

To investigate the differences in genotype imputation between microarray data and RNA-Seq data, we masked the SNP loci absent from GeneSeek Genomic Profiler (GGP) Porcine SNP50 BeadChip in WGS.

Genotype imputation

To investigate the imputation accuracy of RNA-SNPs and Chip-SNPs, we used Beagle v5.4 [38] with a single thread to perform genotype imputation on the target panel. To evaluate the performance of different software in RNA-SNPs imputation, we assessed the imputation accuracy, runtime, and maximum memory usage using Beagle v5.4, Minimac4 v4.1.6 [39], and Impute5 v1.2.0 [40]. All imputation software was run with default parameters. For evaluating the computational resource consumption of the three software programs, each was run on a single thread, recording runtime, maximum memory usage, and average processor utilization. To account for variations in processor usage among the software, we standardized runtime by adjusting for the average number of processor threads, using the formula: $t \times cpu_{average}$, enabling a comprehensive comparison of computational costs across each software program.

Measures of imputation accuracy

In this study, two metrics were employed to evaluate the accuracy of imputation: (1) CR, which reflects the concordance between imputed genotypes and true genotypes, and (2) r^2, which measures the correlation between the true minor allele dosage and the imputed minor allele dosage in the target panels [41]. The true minor allele dosage refers to the number of copies of the minor allele present in an individual's genotype, while the imputed allele dosage is calculated as the sum of the posterior allele probabilities for the two haplotypes of an individual. The formulas for calculating CR and r^2 are shown below:

$$CR = \frac{Genotype_{imputed}}{Genotype_{true}}$$

$$r^2 = Cor\left(Imputed\ dosage, True\ dosage\right)$$

To explore factors affecting imputation accuracy across autosomes and different genomic regions, we analyzed the SNP count per megabase (Mb) across all autosomes. Subsequently, we calculated the fold enrichment for each genomic region. The formulas for calculating fold enrichment are provided below:

$$SNP\ Proportion = \frac{Number\ of\ SNPs\ in\ Genomic\ Region}{Total\ Number\ of\ SNPs}$$

$$Fold\,enrichment = \frac{SNP\ Proportion\ (target\ panel)}{SNP\ Proportion\ (reference\ panel)}$$

Quality control after genotype imputation

MAF is a crucial factor influencing imputation accuracy. Each of the three selected genotype imputation software has its metric for assessing the quality of imputed SNPs: DR^2 in Beagle, R^2 in Minimac4, and INFO in Impute5. These metrics serve as data quality control indicators after genotype imputation. To investigate the effects of MAF and DR^2 on imputation accuracy, we used PLINK v1.90 to calculate MAF and BCFtools v1.9 to extract the DR^2 values of imputed SNPs. Additionally, we filtered imputed SNPs into nine bins based on MAF (≥ 0, ≥ 0.005, ≥ 0.01, ≥ 0.02, ≥ 0.05, ≥ 0.10, ≥ 0.20, ≥ 0.30, ≥ 0.40) and ten bins based on DR^2 (≥ 0, ≥ 0.10, ≥ 0.20, ≥ 0.30, ≥ 0.40, ≥ 0.50, ≥ 0.60, ≥ 0.70, ≥ 0.80, ≥ 0.90). For each filtering criterion, we also calculated the average CR and r^2. Since the total number of loci remaining after quality control is a relevant factor in genotype imputation, we also determined the number of SNPs remaining after applying each of the filtering criteria.

Results

Target panel

To obtain the target panel by masking WGS data based on the SNP features from RNA-seq data, we annotated both RNA-seq and WGS data and retained SNPs located in eight specific regions: intron, intergenic, downstream, upstream, 3' UTR, 5' UTR, synonymous, and non-coding transcript. In the RNA-SNP data, SNPs in eight genomic regions accounted for an average of 97.31% of the total, while in the WGS data, it accounted for an average of 99.77% (Fig. 1a). The distribution of SNPs in the target panel shows a high concordance between masked WGS SNPs, and the actual RNA-SNPs distribution, with a correlation coefficient of 0.99 (Fig. 1b). The highest correlation was observed in the "Synovial_membrane" among all tissues, and the correlation between the corresponding proportions of RNA-SNPs and target panel SNPs in the genomic regions across all tissues is summarized in Table S1.

Imputation accuracy of Chip-SNPs and RNA-SNPs

We first explored the imputation accuracy of Chip-SNPs and RNA-SNPs across all autosomes (Fig. 2a, b). The average CR for Chip-SNPs and RNA-SNPs ranged from 0.873 to 0.909 and 0.895 to 0.933, respectively, while the average r^2 ranged from 0.629 to 0.698 and 0.745 to 0.817, respectively. In general, RNA-SNPs demonstrated a slightly higher average CR compared to Chip-SNPs, while its average r^2 was significantly higher than that of Chip-SNPs. Additionally, chromosome 12 exhibited slightly higher average imputation accuracy with RNA-SNPs compared to the other 17 autosomes. To investigate further, we analyzed the SNP count per Mb across all autosomes, as shown in Fig. 2c. The results showed that chromosome 12 had a markedly higher SNP count per Mb in RNA-SNPs compared to the other chromosomes. Furthermore, the SNP count per Mb in RNA-SNPs was consistently higher than that in Chip-SNPs across all autosomes. We then examined the correlation between SNP count per Mb and imputation accuracy. Our results showed a strong correlation between SNP count per Mb and both CR with a correlation coefficient of 0.73 and r^2 with a correlation coefficient of 0.74. Furthermore, when the SNP count per Mb exceeded 200, imputation accuracy tended to stabilize (Figure S2).

Subsequently, we evaluated the imputation accuracy for each genomic region, comparing RNA-SNPs and Chip-SNPs. The average imputation accuracies (CR and r^2) for eight genomic regions are shown in Fig. 2d and

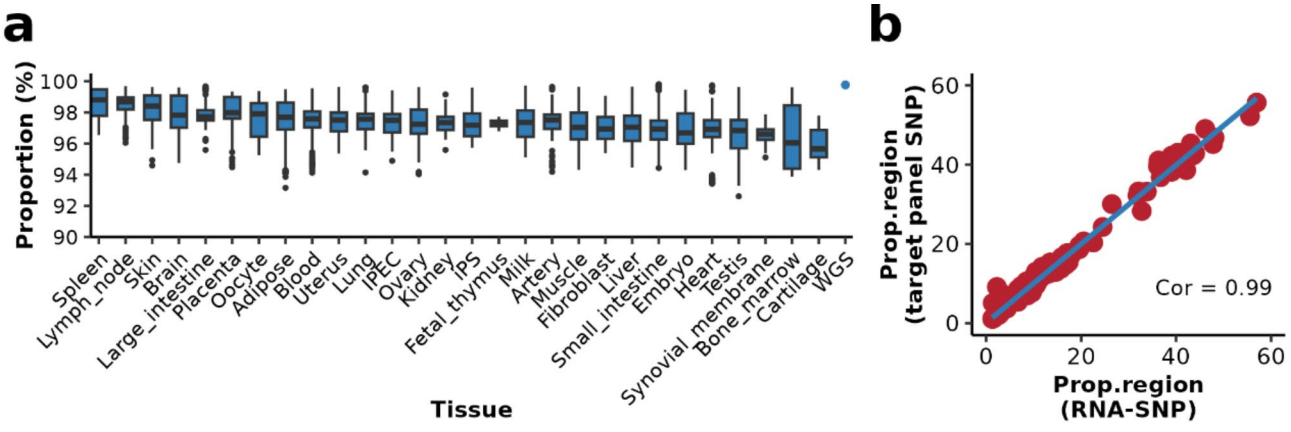

Fig. 1 Proportions of SNPs in eight genomic regions across tissues and the correlation between the proportions of target panel SNPs and RNA-SNPs in these regions. (**a**) Proportion of SNPs in eight genomic regions among the total SNPs, each point represents a single sample from a specific tissue; (**b**) Correlation between the average proportions of RNA-SNPs and target panel SNPs in eight genomic regions across different tissues. Each point represents the average proportions of SNPs in a genomic region within a specific tissue

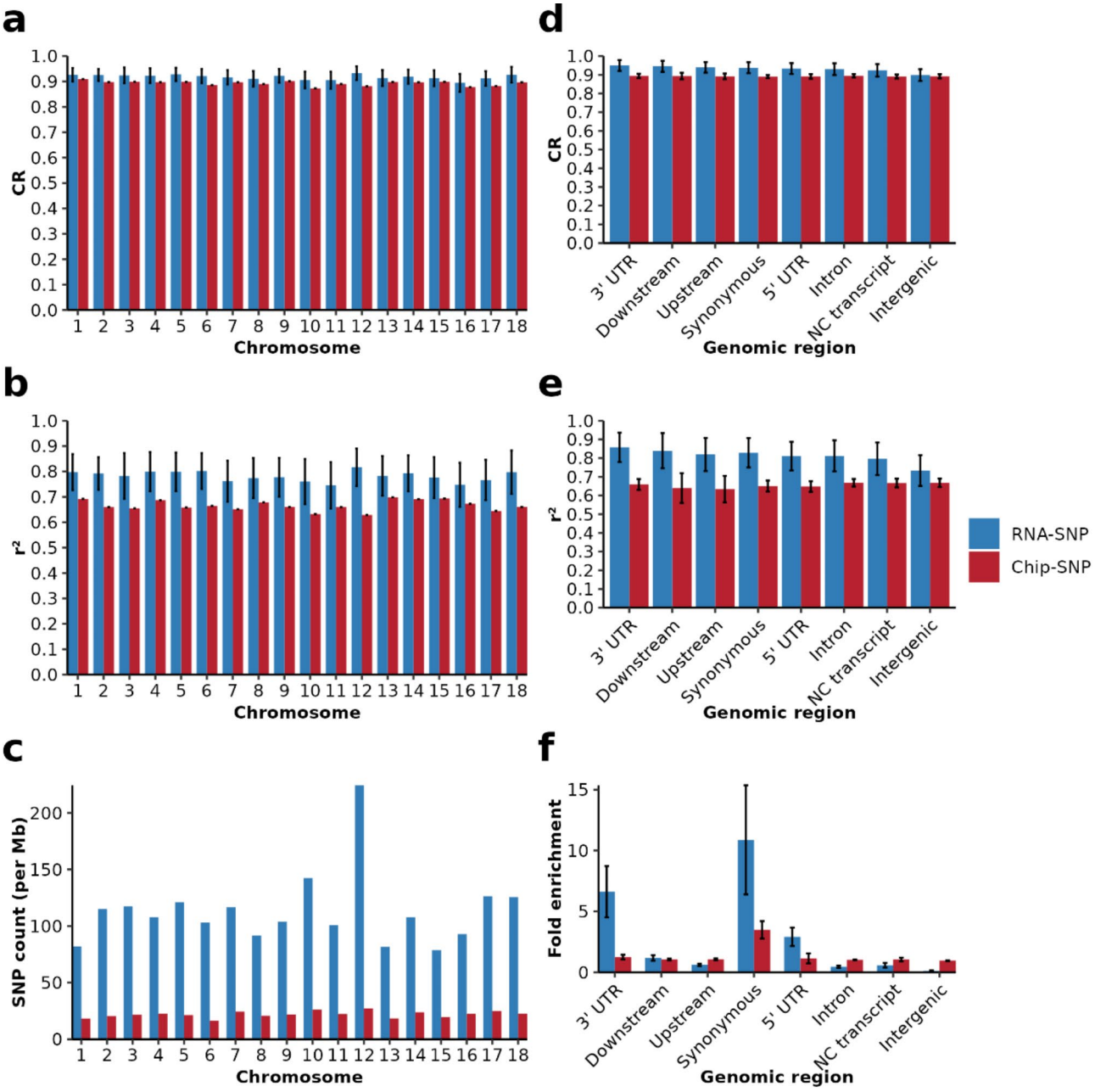

Fig. 2 The imputation accuracy between RNA-SNPs and Chip-SNPs. (**a**-**b**) Imputation accuracy of RNA-SNPs and Chip-SNPs across autosomes; (**c**) The SNP count of RNA-SNPs and Chip-SNPs per Mb across autosomes; (**d**-**e**) Imputation accuracy for RNA-SNPs and Chip-SNPs across eight genomic regions; (**f**) Fold enrichment of RNA-SNPs and Chip-SNPs across eight genomic regions. CR, the concordance rate between imputed genotypes and true genotypes; r^2, the correlation between the true minor allele dose and the imputed minor allele dose. The error bar in each column represents standard error

e. Overall, RNA-SNPs exhibited higher imputation accuracy than Chip-SNPs across all genomic regions. Notably, RNA-SNPs showed a relatively lower r^2 value in the "intergenic" regions, while Chip-SNPs displayed more uniform accuracy across all genomic regions. To further investigate, we calculated the SNP enrichment in each genomic region using fold enrichment, as illustrated in Fig. 2f. The results indicated that RNA-SNPs exhibited the lowest fold enrichment in the "intergenic" regions, whereas Chip-SNPs maintained relatively balanced fold enrichment across all regions.

The standard error of fold enrichment in "synonymous" regions is relatively high, primarily due to significant variation in fold enrichment across different tissues and autosomes for SNPs within these regions. The detailed fold enrichment values for the eight genomic regions across various tissues and autosomes are presented in Table S2.

Post-imputation quality control metrics

To evaluate how quality control thresholds influence RNA-SNPs imputation accuracy and SNP retention rates, we applied nine thresholds for MAF and ten thresholds for DR^2, with results shown in Fig. 3a and b. The results indicate that after excluding regions with low MAF, the CR decreases as the MAF quality control threshold increases. To better assess the imputation of rare variants, we introduced r^2 as a metric for calculating imputation accuracy. The results show that, after filtering out SNPs with low MAF through quality control, the r^2 value showed a substantial improvement. Specifically, when SNPs with MAF < 0.05 were excluded, the CR was 0.894, r^2 reached 0.842, and the SNP retention ratio was 0.671. However, as the MAF filtering threshold increased further, no substantial improvement in accuracy was observed, while the SNP retention ratio continued to decline. When DR^2 was used for quality control, both the CR and r^2 increased with higher DR^2 thresholds, whereas the SNP retention ratio rapidly declined with each increase in the quality control threshold. Our findings suggest that MAF and DR^2 can serve as effective quality control metrics for RNA-SNPs imputation. When selecting quality control thresholds, balancing post-QC accuracy with the SNP retention ratio is essential for optimal results.

Evaluation of common imputation software for RNA-SNPs

We evaluated the performance of three tools: Beagle v5.4, Minimac4 v4.1.6, and Impute5 v1.2.0. The results showed no significant differences in global accuracy between the three tools. The average CR ranged from 0.906 to 0.917 (Fig. 4a) and the average r^2 ranged from 0.780 to 0.787 (Fig. 4b). We further assessed accuracy in different genomic regions. All three tools exhibited the lowest imputation accuracy in the "intergenic" regions. We previously calculated fold enrichment across different regions and found that enrichment in the "intergenic" regions was significantly lower than in other regions, presenting a substantial challenge for imputation. In the same genomic region, the imputation accuracy of the three tools showed little difference (Fig. 4c and d). In terms of computational resource consumption (Fig. 4e and f), Minimac4 had the shortest average runtime, while Impute5 had the longest. Regarding maximum memory usage, Impute5 required the least, while Beagle required the most. Considering both memory demands and runtime, Minimac4 showed superior computational efficiency under single-thread conditions.

Discussion

In this study, we utilized Beagle to impute genotypes from RNA-SNPs and Chip-SNPs, evaluating the imputation accuracy across autosomes and within specific genomic regions. We then used MAF and DR^2 to assess the impact of different quality control thresholds on imputation accuracy and the SNP retention ratio. Finally, we applied three commonly used genotype imputation software to impute RNA-SNPs, evaluating their imputation accuracy and computational resource consumption. The goal of this study was to provide a reliable reference for selecting optimal imputation strategies based on RNA-seq data.

Imputation accuracy of Chip-SNPs and RNA-SNPs

Since most public RNA-seq datasets lack corresponding genotype data, if SNPs called from RNA-seq data can provide satisfactory imputation results, the additional cost of obtaining corresponding SNP chip or WGS data can be avoided. Therefore, we compared the imputation

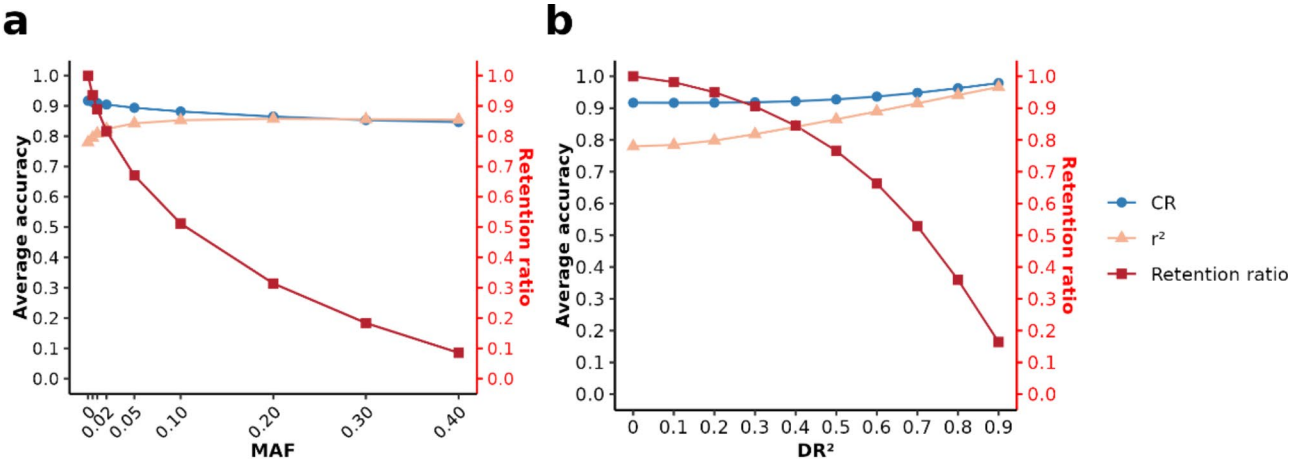

Fig. 3 Effects of post-imputation quality control metrics on accuracy and SNP retention ratio based on RNA-SNPs. (**a**) Impact of MAF quality control threshold selection on accuracy and SNP retention ratio; (**b**) Impact of DR^2 quality control threshold selection on accuracy and SNP retention ratio. CR, the concordance rate between imputed genotypes and true genotypes; r^2, the correlation between the true minor allele dose and the imputed minor allele dose; Retention ratio, the ratio of retained SNPs after quality control to the total number of SNPs after imputation

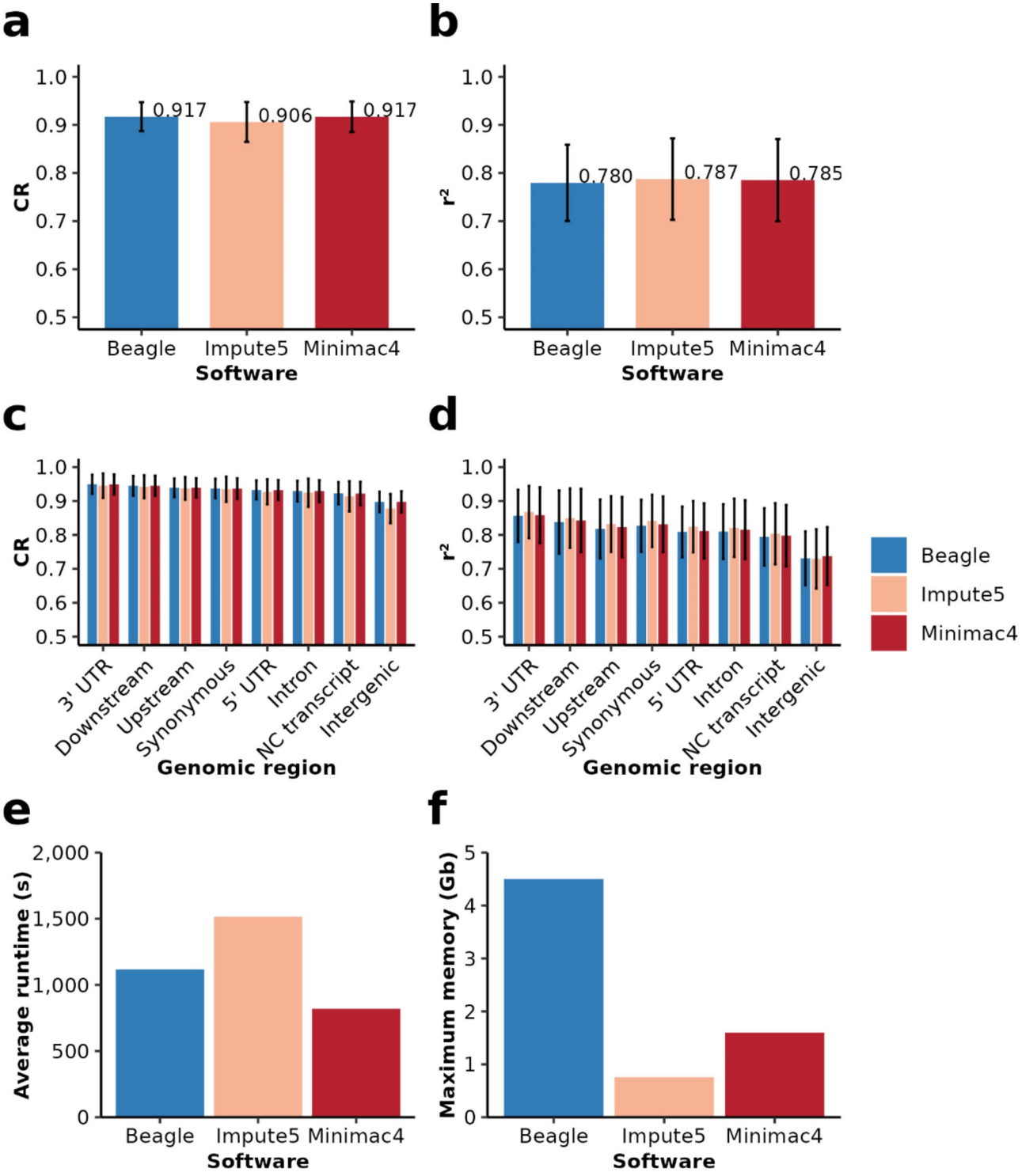

Fig. 4 Evaluation of imputation performance for RNA-SNPs across three software. (**a**-**b**) Accuracy of RNA-SNPs imputation without quality control across three software; (**c-d**) Accuracy of RNA-SNPs imputation without quality control across eight genomic regions for three software; (**e**-**f**) Comparison of computational resource consumption among three software under single-thread conditions. The error bar in each column represents standard error

accuracy of RNA-SNPs and Chip-SNPs. The results indicated that using SNPs called from RNA-seq for genotype imputation is a better approach. We found that when the SNP count per Mb for RNA-SNPs exceeded 200, imputation accuracy tended to stabilize, which is consistent with previous studies [42]. Therefore, increasing SNP density can enhance imputation accuracy when RNA-SNPs density is relatively low. A previous study also

reported decreased imputation accuracy of SNP chips in specific genomic regions [43]. These observations may be due to the design principles of SNP chips and the genetic variation features of different genomic regions. SNP chips are designed based on reference populations and tend to target relatively common variants, including fewer rare variants. As a result, the coverage of SNP chips in certain regions, particularly those involving rare variants, is relatively low, leading to reduced imputation accuracy in these genomic regions.

Post-imputation quality control metrics

In genotype imputation, the absence of true genotypes for imputed SNPs presents challenges in accurately assessing their imputation accuracy. This limitation makes it difficult to identify and exclude poorly imputed SNPs. As a result, post-imputation quality control is essential to evaluate the quality of imputed SNPs and filter out those with low reliability. Rare variants play a significant role in complex traits and are likely a major contributor to the missing heritability of these traits [44]. However, accurately imputing rare variants remains challenging in genotype imputation. We observed that the MAF quality control threshold is negatively correlated with the CR, likely because the concordance rate does not account for the frequency of imputed alleles, overestimates the imputation accuracy for rare variants [45]. Therefore, r^2 is a more suitable metric than CR after MAF quality control. In this study, we applied different DR^2 thresholds for QC, and the results showed a clear positive correlation between DR^2 thresholds and imputation accuracy, indicating that DR^2 can be effectively used as a post-imputation filtering criterion. In addition to considering the accuracy of genotype imputation after QC, the SNP retention ratio is equally important. For MAF-based QC, we recommend filtering out SNPs with MAF < 0.05 in our population, as increasing the MAF threshold (> 0.05) did not yield significant gains in accuracy, while the SNP retention ratio continued to decrease. Similarly, when selecting an appropriate DR^2 threshold, balancing post-QC accuracy with the SNP retention ratio is essential.

Evaluation of common imputation software for RNA-SNPs

In this study, we compared the imputation accuracy of three commonly used genotype imputation tools. The results showed that without QC after imputation, three imputation tools demonstrated relatively high imputation accuracy. Previous comparative analyses of genotype imputation software have demonstrated that each tool possesses distinct advantages [46]. Therefore, it is crucial to evaluate these tools within the context of various scenarios. In terms of computational performance, Minimac4 exhibited better overall performance, while in multi-threaded conditions, Beagle showed the highest support for concurrency. Additionally, Beagle can perform both phasing and imputation in a single step, offering a higher degree of functional integration. This streamlined workflow offers a significant convenience advantage for users.

We also evaluated imputation accuracy across different genomic regions, which revealed notable variation in accuracy among these regions [47]. Previous studies have performed genotype imputation on East Asian populations using reference panels from 1000G and HapMap. The results showed that imputation accuracy for SNPs located in coding regions is higher compared to those in non-coding regions [48]. This discrepancy in imputation accuracy can be attributed to the strong depletion of non-coding variants in typical RNA-seq datasets, which diminishes the reliability of imputing variants located far from transcribed regions [49]. Similar findings were observed in our study, with the lowest imputation accuracy occurring in non-coding regions, such as "intergenic" regions.

Limitations of the study

Although this study provides valuable insights into selecting RNA-SNP-based imputation strategies, several limitations exist. First, sequencing depth may affect imputation accuracy [50], particularly for RNA-seq data, especially in low-expression regions and in the detection of rare variants. Several factors can influence the accuracy of genotype imputation, including the genetic relationship between reference and target populations, and the increase in marker distance, which causes a rapid decay of linkage disequilibrium (LD) between SNPs [51]. Moreover, while we evaluated RNA-SNP-based imputation strategies in pigs, the generalizability of these results to other species requires further validation. Differences in genome structure and transcriptome characteristics across species may influence the choice of RNA-SNPs imputation strategies, necessitating additional studies for broader validation.

Conclusion

RNA-SNPs outperformed Chip-SNPs in imputation accuracy. Filtering SNPs with MAF < 0.05 increased both accuracy and SNP retention, while DR^2-based QC required a trade-off between accuracy and SNP yield. Imputation accuracy was consistent across Beagle, Impute5, and Minimac4 without QC, but software selection should also factor in usability and resource efficiency. Our study provides insights to refine RNA-SNPs imputation strategies.

Abbreviations

SNP	Single Nucleotide Polymorphism
RNA-seq	RNA Sequencing
DNA-seq	DNA Sequencing

eQTL	Expression Quantitative Trait Loci
RNA-SNPs	SNPs Called From RNA-seq Data
Chip-SNPs	SNPs from GeneSeek Genomic Profiler Porcine SNP50 BeadChip
WGS	Whole Genome Sequencing
PGRP	Pig Genomic Reference Panel
MAF	Minor Allele Frequency
QC	Quality Control
DR^2	Dosage R-Squared
GWAS	Genome-Wide Association Study
ASE	Allele-Specific Expression
GTEx	Genotype-Tissue Expression
PCA	Principal Component Analysis
PCs	Principal Components
3' UTR	3' Untranslated Region
5' UTR	5' Untranslated Region
NC transcript	Non-Coding transcript
GGP	GeneSeek Genomic Profiler
CR	Concordance Rate
Mb	Megabase
IQS	Imputation Quality Score
LD	Linkage Disequilibrium

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11411-5 .

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Acknowledgements
We thank National Supercomputer Center in Guangzhou China for its support in providing computing resources.

Author contributions
K.X.G, J.Y.T. and Y.H.G. collected the data and materials, K.X.G performed data analyses, and drafted the manuscript. Z.M.Z., H.N.Z., C.L.Z., T.T.C, J.Y.T. and Y.H.G. revised the manuscript. Z.Z. conceived and designed the experiment. All authors read and approved the final manuscript.

Funding
This research was funded by the China Agriculture Research System (CARS-35); the Specific university discipline construction project (2023B10564001, 2023B10564003); and the Guangxi Science and Technology Program Project (GuikeJB23023003).

Data availability
All raw RNA-seq data and raw PGRP WGS data analyzed in this study are publicly accessible without restrictions from the Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra/). Detailed information regarding the RNA-seq datasets and the PGRP WGS data are provided in Supplementary Tables of the PigGTEx paper [16]. The WGS data used as the gold standard were downloaded from the GigaScience database [32].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 December 2024 / Accepted: 27 February 2025
Published online: 13 March 2025

References

1. Liu J, Sebastià C, Jové-Juncà T, Quintanilla R, González-Rodríguez O, Passols M, et al. Identification of genomic regions associated with fatty acid metabolism across blood, liver, backfat and muscle in pigs. Genet Selection Evol. 2024;56:66.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320:1344–9.
3. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated Single-Base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133:523–36.
4. Deelen P, Zhernakova DV, de Haan M, van der Sijde M, Bonder MJ, Karjalainen J, et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. Genome Med. 2015;7:30.
5. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-Seq data. Am J Hum Genet. 2013;93:641–51.
6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
7. Gondret F, Vincent A, Houée-Bigot M, Siegel A, Lagarrigue S, Causeur D, et al. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics. 2017;18:244.
8. Savary C, Kim A, Lespagnol A, Gandemer V, Pellier I, Andrieu C, et al. Depicting the genetic architecture of pediatric cancers through an integrative gene network approach. Sci Rep. 2020;10:1224.
9. Jehl F, Désert C, Klopp C, Brenet M, Rau A, Leroux S, et al. Chicken adaptive response to low energy diet: main role of the hypothalamic lipid metabolism revealed by a phenotypic and multi-tissue transcriptomic approach. BMC Genomics. 2019;20:1033.
10. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464:773–7.
11. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9:1825.
12. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015;348:660–5.
13. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
14. The GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
15. Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, et al. A multi-tissue atlas of regulatory variants in cattle. Nat Genet. 2022;54:1438–47.
16. Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, et al. A compendium of genetic regulatory effects across pig tissues. Nat Genet. 2024;56:112–23.
17. Guan D, Bai Z, Zhu X, Zhong C, Hou Y, Consortium TC et al. The chickengtex pilot analysis: a reference of regulatory variants across 28 chicken tissues. Biorxiv. 2023;:2023.06.27.546670.
18. Adetunji MO, Lamont SJ, Abasht B, Schmidt CJ. Variant analysis pipeline for accurate detection of genomic variants from transcriptome sequencing data. PLoS ONE. 2019;14:e0216838.
19. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, Ballantyne C, et al. Practical approaches for Whole-Genome sequence analysis of Heart- and Blood-Related traits. Am J Hum Genet. 2017;100:205–15.
20. Liu Y, Liu X, Zheng Z, Ma T, Liu Y, Long H, et al. Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. Genet Selection Evol. 2020;52:59.
21. Cánovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. Mamm Genome. 2010;21:592–8.
22. Wang W, Wang H, Tang H, Gan J, Shi C, Lu Q, et al. Genetic structure of six cattle populations revealed by transcriptome-wide SNPs and gene expression. Genes Genom. 2018;40:715–24.

23. Bakhtiarizadeh MR, Alamouti AA. RNA-Seq based genetic variant discovery provides new insights into controlling fat deposition in the tail of sheep. Sci Rep. 2020;10:13525.
24. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annual Review of Animal Biosciences. 2019;7 Volume 7, 2019:89–102.
25. Mayoh C, Barahona P, Lin A, Cui L, Ajuyah P, Altekoester A, et al. Abstract B014: increasing the clinical utility of transcriptome analysis in high-risk childhood precision oncology. Cancer Res. 2024;84(17Supplement):B014.
26. Yuan Z, Sunduimijid B, Xiang R, Behrendt R, Knight MI, Mason BA, et al. Expression quantitative trait loci in sheep liver and muscle contribute to variations in meat traits. Genet Selection Evol. 2021;53:8.
27. Cai W, Hu J, Zhang Y, Guo Z, Zhou Z, Hou S. Cis-eQTLs in seven Duck tissues identify novel candidate genes for growth and carcass traits. BMC Genomics. 2024;25:429.
28. Marino AD, Mahmoud AA, Bose M, Bircan KO, Terpolovsky A, Bamunusinghe V, et al. A comparative analysis of current phasing and imputation software. PLoS ONE. 2022;17:e0260177.
29. Ding R, Savegnago R, Liu J, Long N, Tan C, Cai G, et al. The swine imputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs. Commun Biol. 2023;6:1–10.
30. Teng J, Zhao C, Wang D, Chen Z, Tang H, Li J, et al. Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. J Dairy Sci. 2022;105:3355–66.
31. Ye S, Zhou X, Lai Z, Ikhwanuddin M, Ma H. Systematic comparison of genotype imputation strategies in aquaculture: A case study in nile tilapia (*Oreochromis niloticus*) populations. Aquaculture. 2024;592:741175.
32. Crespo-Piazuelo D, Acloque H, González-Rodríguez O, Mongellaz M, Mercat M-J, Bink MCAM, et al. Supporting data for "Identification of transcriptional regulatory variants in pig duodenum, liver, and muscle tissues.". GigaScience Database 2023.; 10.5524/102388
33. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter Estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
34. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:s13742-015-0047–8.
35. Wickham H. ggplot2. Cham: Springer International Publishing; 2016.
36. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021;108:1880–90.
37. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6:80–92.
38. Browning BL, Zhou Y, Browning SR. A One-Penny imputed genome from Next-Generation reference panels. Am J Hum Genet. 2018;103:338–48.
39. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.
40. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the positional burrows Wheeler transform. PLoS Genet. 2020;16:e1009049.
41. Browning BL, Browning SR. A unified approach to genotype imputation and Haplotype-Phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210–23.
42. Shi S, Yuan N, Yang M, Du Z, Wang J, Sheng X, et al. Comprehensive assessment of genotype imputation performance. Human Hered. 2019;83:107–16.
43. Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Selection Evol. 2017;49:24.
44. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18:77.
45. Fernandes Júnior GA, Carvalheiro R, de Oliveira HN, Sargolzaei M, Costilla R, Ventura RV, et al. Imputation accuracy to whole-genome sequence in Nellore cattle. Genet Selection Evol. 2021;53:27.
46. Ellinghaus D, Schreiber S, Franke A, Nothnagel M. Current software for genotype imputation. Hum Genomics. 2009;3:371.
47. Sun C, Wu X-L, Weigel KA, Rosa GJM, Bauck S, Woodward BW, et al. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet Res. 2012;94:133–50.
48. Lert-itthiporn W, Suktitipat B, Grove H, Sakuntabhai A, Malasit P, Tangthawornchaikul N, et al. Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome. BMC Med Genet. 2018;19:23.
49. Leonard AS, Mapel XM, Pausch H. RNA-DNA differences in variant calls from cattle tissues result in erroneous eQTLs. BMC Genomics. 2024;25:750.
50. Wragg D, Zhang W, Peterson S, Yerramilli M, Mellanby R, Schoenebeck JJ, et al. A cautionary Tale of low-pass sequencing and imputation with respect to haplotype accuracy. Genet Selection Evol. 2024;56:6.
51. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and Missing-Data imputation. Am J Hum Genet. 2005;76:449–62.

Publisher's note