Check for updates

Empirical versus estimated accuracy of imputation: optimising filtering thresholds for sequence imputation

Tuan V. Nguyen[1]*, Sunduimijid Bolormaa[1], Coralie M. Reich[1], Amanda J. Chamberlain[1,2], Christy J. Vander Jagt[1], Hans D. Daetwyler[1,2] and Iona M. MacLeod[1,2]

Abstract

Background Genotype imputation is a cost-effective method for obtaining sequence genotypes for downstream analyses such as genome-wide association studies (GWAS). However, low imputation accuracy can increase the risk of false positives, so it is important to pre-filter data or at least assess the potential limitations due to imputation accuracy. In this study, we benchmarked three different imputation programs (Beagle 5.2, Minimac4 and IMPUTE5) and compared the empirical accuracy of imputation with the software estimated accuracy of imputation (Rsq_{soft}). We also tested the accuracy of imputation in cattle for autosomal and X chromosomes, SNP and INDEL, when imputing from either low-density or high-density genotypes.

Results The accuracy of imputing sequence variants from real high-density genotypes was higher than from low-density genotypes. In our software benchmark, all programs performed well with only minor differences in accuracy. While there was a close relationship between empirical imputation accuracy and the imputation Rsq_{soft}, this differed considerably for Minimac4 compared to Beagle 5.2 and IMPUTE5. We found that the Rsq_{soft} threshold for removing poorly imputed variants must be customised according to the software and this should be accounted for when merging data from multiple studies, such as in meta-GWAS studies. We also found that imposing an Rsq_{soft} filter has a positive impact on genomic regions with poor imputation accuracy due to large segmental duplications that are susceptible to error-prone alignment. Overall, our results showed that on average the imputation accuracy for INDEL was approximately 6% lower than SNP for all software programs. Importantly, the imputation accuracy for the non-PAR (non-Pseudo-Autosomal Region) of the X chromosome was comparable to autosomal imputation accuracy, while for the PAR it was substantially lower, particularly when starting from low-density genotypes.

Conclusions This study provides an empirically derived approach to apply customised software-specific Rsq_{soft} thresholds for downstream analyses of imputed variants, such as needed for a meta-GWAS. The very poor empirical imputation accuracy for variants on the PAR when starting from low density genotypes demonstrates that this region should be imputed starting from a higher density of real genotypes.

*Correspondence:
Tuan V. Nguyen
tuan.nguyen@agriculture.vic.gov.au
[1] Agriculture Victoria, Centre for AgriBiosciences, AgriBio, Bundoora, VIC 3083, Australia
[2] School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia

Background

Imputation is the process of predicting missing genome-wide genotypes in individuals with lower density genotypes by utilizing a reference population that has higher density genotypes [1, 2]. In recent years, the availability of imputed sequence data for downstream analyses such as genome-wide association studies (GWAS), has become

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 2 of 14

extremely useful to explore the underlying genetic mechanisms of various phenotypes (e.g. [3–5]). Furthermore, candidate causal variants identified in imputed sequence GWAS can contribute to improve the accuracy of genomic prediction of complex traits in livestock (e.g. [6–8]). However, sequence GWAS often have low statistical power to detect many of the variants that affect complex traits because the majority of effects are small [9]. Thus, research has now shifted towards large-scale coordinated meta-GWAS to improve statistical power [10, 11], for example in humans [12, 13], plants [14, 15], and livestock [16–18]. A challenge of these meta-analyses is quality control to filter out poorly imputed variants, particularly when contributors have used different imputation software.

Two popular algorithmic approaches for genotype imputation are family- or population-based. In brief, family imputation utilizes relationships and shared genetic segments whilst population imputation uses large reference panels and linkage disequilibrium patterns. Some imputation programs combine both population- and family-based approaches and these are popular for use in livestock (for example AlphaImpute [19], FindHap [20], FImpute [21]). However, most of these programs do not offer an internally calculated imputation quality measure (Rsq_{soft}) that is available in some other popular imputation software. At sequence level, an Rsq_{soft} estimate of the imputation accuracy is an indispensable quality control tool for removing poorly imputed variants from downstream analysis such as GWAS [1]. In general, several factors influence the final quality of imputation including: the choice of software, initial starting genotype density, reference population size and their relatedness to the target individuals [22]. Furthermore, the ever-increasing number of genotyped individuals and sequence reference individuals presents computational challenges for imputation, resulting in algorithmic modifications to improve computational efficiency of software. While several studies in cattle over the past decade have reported the accuracy of imputing from SNP panels to sequence variants [23–25], further studies on sequence imputation benchmarking are required for several reasons. First, it is critical to benchmark the internal software estimate of imputation accuracy versus empirically assessed imputation accuracy across different software. This provides an evaluation of how closely the software estimate mirrors the empirical accuracy and determines if this relationship differs between software. Second, to the best of our knowledge, no cattle study has reported the effectiveness of INDEL versus SNP imputation at sequence level using the updated reference genome ARS-UCD 1.2 [26], even though INDEL account for around 7% to 10% of all variants [27]. Third, this is the first study to compare

sequence imputation accuracy for autosomal accuracy versus the pseudo-autosomal region (PAR) and non-PAR of the X chromosome from both low and high density SNP panels to sequence using the ARS-UCD 1.2 [26].

In this study, we evaluated the relationship between the software estimated accuracy of imputation (Rsq_{soft}) and the empirical accuracy of imputation. We did this for both SNP and INDEL using three popular and competitive software packages: Beagle 5.2 [28], Minimac4 [29], and IMPUTE5 [30]. Results are compared for imputing either from Low Density (LD, ~7000 variants) or from High Density SNP array genotypes (HD, ~700,000 variants) to sequence. We also report on the effect of pre-filtering the reference sequence variants prior to imputation.

Methods

Target animal genotypes

Sequence genotypes

The target animals used for this study included 70 sequenced bulls: 35 Jersey (JER) and 35 Holstein (HOL). We used pedigree records and a genomic relationship matrix to confirm that there were no half-siblings present. Individuals were sequenced using Illumina HiSeq 2000 (Illumina Inc., San Diego, CA). Raw sequence reads were aligned to the ARS-UCD1.2 reference genome [26]. Alignment, variant calling, and quality controls were performed following the 1000 Bull Genomes Project guidelines as the animals were included in Run8 (See Additional 1, Text S1). In this study, we used chromosome 1, 5, 10, 15, 20, 25, and X to evaluate the accuracy of imputation at sequence level. The X chromosome was split into the non-pseudo-autosomal region (non-PAR: chromosome "30"), and the PAR (chromosome "32"). The boundary between the non-PAR and PAR was set at 133,300,518 bp [31].

HD and LD SNP genotypes

The same target set of 70 individuals were also genotyped using the Illumina® BovineHD 800 K bead chip (HD). The marker map positions were lifted over to the ARS-UCD1.2 reference genome using publicly available datasets (https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/). The raw HD genotypes for these animals were processed together with the HD reference population described below. This set of HD genotypes were also masked down to a genome-wide Low Density (LD) set of 7135 SNP markers that overlap many of the commonly used current and historical SNP panels, including the Illumina® BovineSNP50K and HD SNP panels.

Reference animal genotypes

50K and HD SNP genotypes

We had previously generated a reference population of 14,722 animals (representing Holstein, Jersey, and Australian Red breeds) with real Illumina® BovineSNP50K panel (50K) genotypes, including a total of 40,397 SNP that passed quality filters and overlapped the Bovine Illumina HD SNP panel. Additionally, we had a cohort of 2814 animals with real HD SNP genotypes that constituted the HD imputation reference (again representing the same breeds as for the 50K reference population). The final set of HD SNP passing the quality control was 714,452 that overlapped the sequence variants in the reference population described below. In processing the raw 50K and HD genotypes, the GenCall threshold score was set at 0.6, such that SNP with a lower score were set to missing and SNP with > 10% missing genotypes were removed. All animals had < 10% missing genotypes. For both the 50K and HD reference sets, the remaining sporadic missing genotypes were imputed using FImpute v.3 [21].

Sequence genotypes

The imputation reference comprised 4190 taurus cattle sequences in Run8 of the 1000 Bull Genomes Project [32]. The reference sequences were processed following the 1000 Bull Genomes project pipeline (See Additional file 1, Text S1) and within this pipeline the sporadic missing genotypes were imputed using Beagle 4.1 [28]. Following imputation of the sporadic missing genotypes, variants were removed if their Beagle Rsq_{soft} was less than 0.9. The sequence variants were then further filtered to retain only bi-allelic variants with a minor allele count (MAC) of at least 4 and with a GATK [33] Variant Quality Score Recalibration (VQSR) Tranche of 99.0 or better. Additionally, we identified regions of excessive heterozygosity in sliding windows of 0.5 Mb, defined as windows where 2% or more of the variants had heterozygosity > 0.55. Within these windows, we removed all variants with heterozygosity > 0.55 because this generally indicates regions with known long segmental repeats that suffer from poor alignment of short read sequence data resulting in false SNP [34].

Phasing and imputation strategies

We used the 70 target animals to evaluate the accuracy of imputing to sequence either directly from their real HD genotypes or starting from the LD SNP genotypes (7135 markers generated from their masked HD genotypes). The LD SNP genotypes were imputed first to the 50K reference and then up to the HD reference using FImpute v.3 with default settings and no pedigree provided. The

autosomal chromosomes (Chr 1, 5, 10, 15, 20, and 25) were each imputed independently. The X chromosome non-PAR and PAR were imputed separately as per software recommendations. Finally, both the imputed and real HD genotypes of the target animals were converted to VCF format ensuring that the SNP array alleles were matched to the sequence format before final imputation to sequence level.

For sequence imputation, we evaluated the performance of three imputation tools: Beagle version 5.2 [28], IMPUTE5 version 1.1.4 [30], and Minimac4 version 1.0.2 [29]. When using IMPUTE5 and Minimac4 imputation software, genotypes of the target and reference individuals were pre-phased (as required) using Eagle v2.4.1 [35] prior to imputation. The target animal genotypes were left unphased for the Beagle imputation because this software does not require pre-phasing. We ran a preliminary investigation of the accuracy of Beagle imputation using either the Eagle-phased reference, or a Beagle-phased reference. Our analysis found little difference in imputation accuracy between Eagle- or Beagle-phasing of the reference, although Eagle-phased reference resulted in slightly higher accuracy (0.4% on average) in both Best-Guess (GT) and Dosage (DS) genotypes (See Additional file 2, Table S1, Sheet 1). We therefore decided to use the Eagle-phased reference for all software benchmarking to maintain consistency across scenarios.

Additionally, as a pilot investigation, we compared the empirical accuracy of sequence imputation for Beagle 5.2 and Minimac4 using two different settings relating to the length of haplotype imputed. As a default setting for computational efficiency, both programs implement an automated "chunking" of chromosomes into shorter window lengths. Conversely, the default setting in IMPUTE5 is to use the full chromosome. The Beagle 5.2 default sets the window lengths to 40 centiMorgan (with a 2 cM overlap) while the Minimac4 default is 20 Mb with a 3 Mb overlap. We therefore tested Beagle 5.2 and Minimac4 using either the default window size or full chromosome lengths under the hypothesis that use of the full chromosome might improve imputation accuracy by taking advantage of the long-distance linkage disequilibrium in cattle. Figure 1 illustrates the overall experimental design of the study.

The settings tested were:

Beagle 5.2

1. D (Default setting): window = 40.0, overlap = 2.0 (distances are in cM) or
2. F (Full chromosome): window = 200.0, overlap = 80.0 (ensured that one 'window' is larger than the entire length of each chromosome)

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 4 of 14

Pilot experiments
1. Beagle 5.2 vs. Eagle 2.4.1 pre-phased test
2. Beagle 5.2 & Minimac 4 chunking test | Full chromosome vs default windows

Overall benchmarking with optimal setting: Beagle 5.2 | Minimac 4 | IMPUTE5
- Tested on autosomes (Chr 1,5,10,15,20,25) + non-PAR + PAR
- Dosage/Best guess genotype benchmarking
- Correlation between Rsq_{soft} & Rsq_{emp}
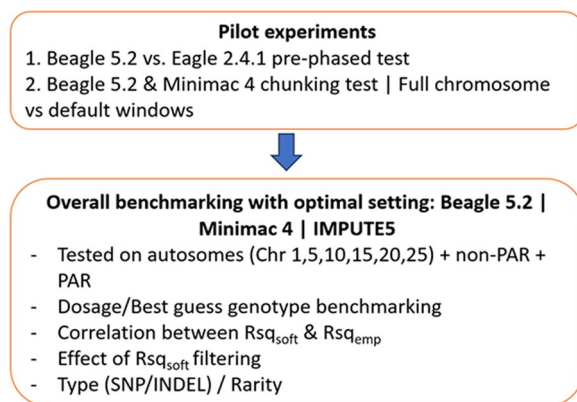- Effect of Rsq_{soft} filtering
- Type (SNP/INDEL) / Rarity

Fig. 1 Flowchart showing overall experimental design for empirical testing of imputation accuracy and benchmarking the software estimate of imputation accuracy

For both tests above, a seed setting was used to ensure reproducibility (seed value = 5555). The effective population size parameter (*Ne*) was set at 1000 for all Beagle imputation.

Minimac4

1. D (Default setting): –ChunkLengthMb 20 –ChunkOverlapMb 3 or
2. F (Full chromosome): –ChunkLengthMb 200 –ChunkOverlapMb 80 (ensured that one 'window' was larger than the entire length of each chromosome).

No seed or effective population size setting was available in Minimac4.

IMPUTE5
The default settings of IMPUTE5 were used, except that *Ne* was set at 1000. No seed setting was available in IMPUTE5.

The results from the pre-phasing and pilot test described above were used to inform the settings used in the main benchmarking study of the three software programs.

Imputation accuracy statistics
We defined *empirical accuracy* (*r*) as Pearson correlation coefficient between the real sequence genotypes (coded as the number of alternate alleles: 0, 1, or 2) and imputed genotypes. The imputed genotypes were available in two forms: either allele "dosage" (DS: continuous measures between 0 to 2, representing the sum of the two alternate allele probabilities) or "best-guess" genotypes (GT: coded with 0, 1 or 2 as for real genotypes). The correlation was

calculated for both DS and GT on a per variant basis across all target animals.

A variant had to be segregating in both the real and imputed best guess genotypes (GT) of the target set in order to calculate *r*, otherwise the variance for that position is zero and *r* cannot be estimated. For DS genotypes, there is always some variance when the variant is segregating in the real genotypes of the target set. Therefore, unless stated otherwise, when benchmarking across different software, we compared only the common set of overlapping variants that had an estimable *r* across all software for GT genotypes. Additionally, we use the DS imputed genotypes to enable comparisons across the entire set of variants segregating in the target set. An internal software estimate of imputation accuracy per variant was available in the output of all three packages (we will refer to as: Rsq_{soft}) and this was compared to the squared empirical accuracy of imputation (Rsq_{emp}).

For imputed GT genotypes, we measured three additional, more specific statistics:

1. *False positive error rate* (FPR): defined as the average percentage of reference alleles that were wrongly imputed as the alternate allele for any given variant position.
2. *False negative error rate* (FNR): defined as the average percentage of alternate alleles that are wrongly imputed as the reference allele for any given variant position.
3. *Allelic imputation error rate* (AER): defined as the average percentage of alleles that are wrongly imputed (i.e. false positives + false negatives) for any given variant position.

A graphical schematic on the calculations of these statistics can be found in Additional file 3, Figure S1.

Results
Chromosome chunking versus full chromosome imputation (Beagle 5.2 and Minimac4)
The default settings in both Minimac4 and Beagle 5.2 divide chromosomes into smaller overlapping windows as a means of speeding up the imputation process while the default for IMPUTE5 is to use full length chromosomes. Therefore, for Beagle 5.2 and Minimac4 we compared their default window settings (D) with imputing the full chromosome (F) as a single window. We tested this because cattle have a small recent effective population size [36] and often the target and reference animals are quite closely related so may share long haplotypes. However, our results from the different parameter settings (D versus F) did not show significantly different imputation accuracy (See Additional file 2, Table S1,

Nguyen *et al. Genetics Selection Evolution* (2024) 56:72

Page 5 of 14

Sheet 2). We therefore proceeded with the default window settings for Minimac4 & Beagle 5.2 for the remainder of the study comparisons.

Empirical imputation accuracy

For all three software methods, we first compared the *r* of imputation to sequence starting either direct from real HD genotypes, or from LD genotypes (approximately 7K SNP) imputed to 50K, then to HD and finally to sequence. For all autosomes tested (1, 5, 10, 15, 25) as well as the non-PAR of the X chromosome, there was a small (~2%) but consistent drop in accuracy of best guess genotypes (GT) when imputing from LD genotypes compared to starting from real HD genotypes (Fig. 2a). However, for the PAR region of the X chromosome there was a very sharp drop in *r* from around 0.9 (starting from HD) to less than 0.6 when starting from LD genotypes (Fig. 2a). Overall, there were only small differences in *r* across the autosomal chromosomes and the non-PAR (Fig. 2a).

There was almost no difference between software in the accuracy of imputation measured as the correlation of best guess (GT) genotypes (Fig. 2a). Likewise, the concordance rate for best guess genotypes (GT mode) was relatively high across all programs tested, ranging from 90.0 to 99.0% (imputing from LD) and 97.8% to 99.6% for imputing from real HD (See Additional file 2, Table S2, Sheet 3). However, *r* in Fig. 2 included all variants that the given program imputed as segregating, so when comparing software performance, this result can show some bias if a specific software has a stronger tendency to impute rare variants as not segregating because the correlation is not estimable for those variants. Therefore, we also estimated AER (allelic error rate), FNR (alternate alleles falsely imputed as reference alleles), and FPR (reference alleles falsely imputed as alternate alleles) statistics. Minimac4 tended to produce a slightly higher AER and FNR compared to either Beagle 5.2 or IMPUTE5, while the FPR was relatively stable across software (Fig. 2b–d). Thus, the overall number
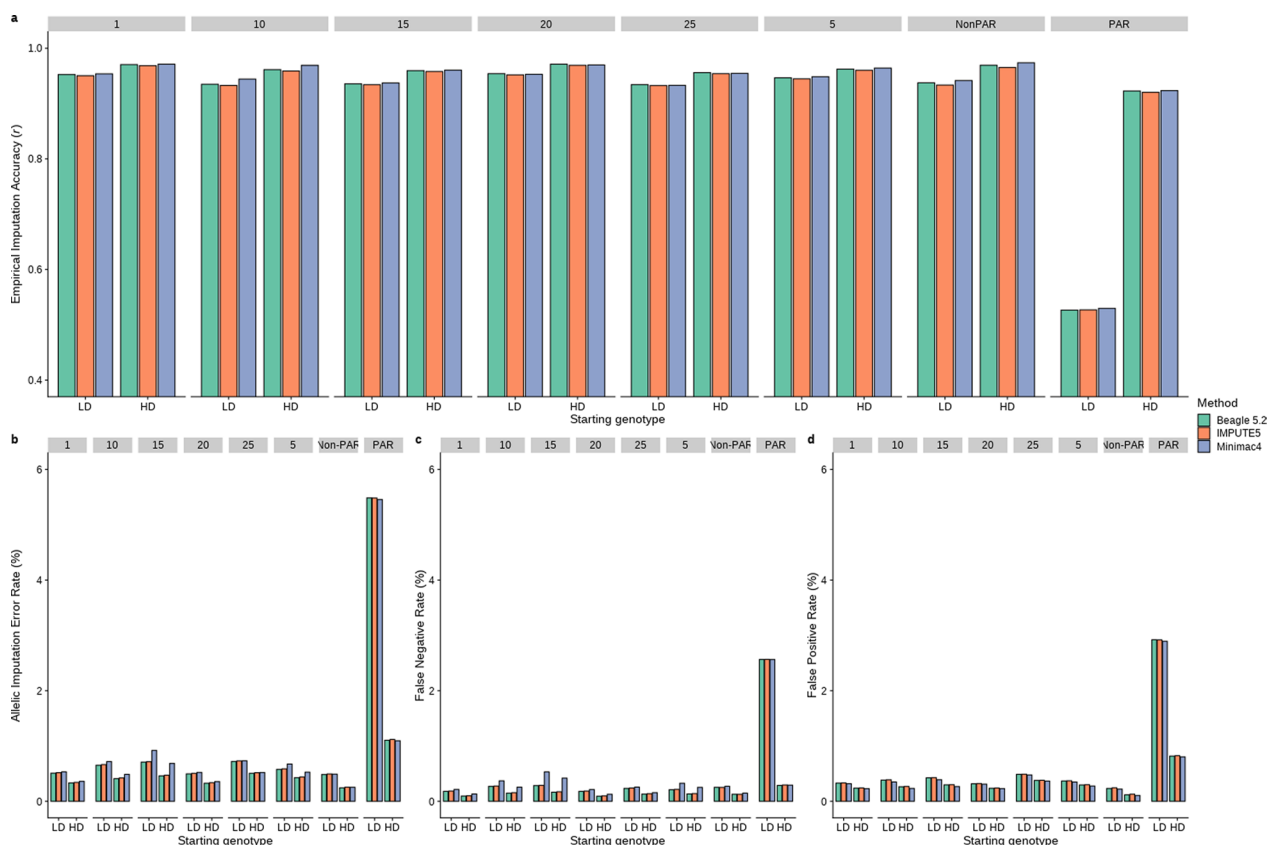

Fig. 2 Imputation statistics for six autosomes (1, 5, 10, 15, 20 & 25) as well as the X chromosome non-pseudo-autosomal region (*non-PAR*) and pseudo-autosomal region (*PAR*) for three imputation software (Beagle 5.2, Minimac4 and IMPUTE5). **a** Average empirical imputation accuracy (r) for best guess genotypes (GT) and considering all variants imputed as segregating by the specific software. **b** Average allelic imputation error rate (%). **c** False Negative Rate (% of alternate alleles wrongly imputed as reference). **d** False Positive Rate (% of reference alleles wrongly imputed as alternate). HD = Impute from real HD genotypes to sequence. LD = Impute from low density to 50K to HD to sequence

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 6 of 14

of variants for which the correlation was estimable was approximately 3% lower for Minimac4 compared to that of Beagle 5.2 and IMPUTE5 (starting from LD: 3,961,994 versus 4,086,017 and 4,076,376 variants and starting from HD: 3,992,240 versus 4,119,862 and 4,109,502, respectively). We also estimated the percentage of segregating variants that were imputed as not segregating (Chromosome 1, starting from HD) and found that the highest error rate was for Minimac4 (3.5%) compared to Beagle 5.2 and IMPUTE5 (1.5% and 1.8%, respectively). Conversely, the percentage error rate for non-segregating variants imputed as segregating was low across all software (0.3% for Minimac4, 0.7% of Beagle 5.2, and 0.8% of IMPUTE5). As expected, the AER, FPR and FNR were always higher when imputation started from LD rather than from real HD genotypes. Additionally, while there were only small differences between the autosomal chromosomes, the AER, FPR and FNR were much higher for the PAR of chromosome X (Fig. 2b–d).

Additionally, we re-estimated the average *r* within only the overlapping set of variants (DS and GT) that had an estimable *r* across all three software programs (i.e., removing all variants imputed as monomorphic by any software). For DS genotypes we show the correlation for this overlapping variant set as well as for the full variant set because with DS most variants show some small imputation probability of segregating. We combined all the autosomal variants for this comparison and Table 1 shows the *r* using either GT or DS genotypes for variants from the autosomes, PAR and Non-PAR. All programs performed similarly, and as expected, the correlation was slightly higher for DS genotypes compared to GT genotypes. The DS and GT accuracies for the non-PAR were

similar to the autosomes across all comparisons, while the PAR had a very low accuracy of imputation when starting from LD genotypes. The accuracy of imputation on the PAR starting from real HD genotypes was much higher than when starting from LD but was still ~2 to 4% lower than for the autosomes.

Within the chromosomes tested, there was a total of 748,461 INDEL representing 7% of all the imputed variants in our study. We compared the accuracy of imputation between SNP and INDEL for DS genotypes across all tested autosomes as well as the PAR and non-PAR, imputing either from real HD or from LD to sequence, using the three software tools (Table 2). On average, for all software programs the INDEL had approximately 6%, 14% and 4% lower imputation accuracy compared to SNP for autosomes, non-PAR and PAR respectively. However, for autosomal variants the Beagle 5.2 INDEL imputation accuracy was slightly better than IMPUTE5 and in all comparisons, Beagle accuracy was nearly 2% better than Minimac4 when imputing from LD.

Software imputation quality metric versus empirical accuracy

After establishing the accuracy of imputation from LD and HD genotypes to sequence level, this part of the study addresses two critical questions. First, is the relationship between the software estimated accuracy (Rsq_{soft}) and empirical imputation accuracy (Rsq_{emp}) strong enough to provide a useful means of filtering poorly imputed data? Second, what is an appropriate imputation Rsq_{soft} threshold for each software? The relationship between the software imputation quality measure (Rsq_{soft}) and the empirical accuracy as the squared correlation (Rsq_{emp}) is

Table 1 Average empirical imputation accuracy (*r*) per variant when imputing from either low (LD) or high density (HD) genotypes using three imputation programs (Beagle 5.2, Minimac4 and IMPUTE5)

Mode	Programs	Autosome (*r* ± SD)		Non-PAR (*r* ± SD)		PAR (*r* ± SD)	
		LD	HD	LD	HD	LD	HD
GT[1]	Beagle 5.2	0.943 ± 0.114	0.964 ± 0.101	0.938 ± 0.141	0.969 ± 0.128	0.527 ± 0.196	0.923 ± 0.105
	Minimac4	0.945 ± 0.101	0.965 ± 0.088	0.942 ± 0.13	0.974 ± 0.113	0.53 ± 0.2	0.923 ± 0.099
	IMPUTE5	0.941 ± 0.116	0.962 ± 0.103	0.934 ± 0.153	0.965 ± 0.14	0.53 ± 0.2	0.92 ± 0.107
DS[1] (Same variant set as GT)	Beagle 5.2	0.946 ± 0.111	0.966 ± 0.097	0.941 ± 0.134	0.972 ± 0.12	0.53 ± 0.2	0.925 ± 0.101
	Minimac4	0.95 ± 0.09	0.969 ± 0.073	0.945 ± 0.122	0.977 ± 0.104	0.53 ± 0.19	0.926 ± 0.089
	IMPUTE5	0.945 ± 0.112	0.965 ± 0.098	0.936 ± 0.148	0.968 ± 0.134	0.53 ± 0.2	0.923 ± 0.102
DS[1] (All variant set)	Beagle 5.2	0.938 ± 0.133	0.96 ± 0.119	0.907 ± 0.219	0.941 ± 0.209	0.52 ± 0.2	0.92 ± 0.121
	Minimac4	0.922 ± 0.169	0.946 ± 0.15	0.893 ± 0.243	0.927 ± 0.232	0.51 ± 0.22	0.914 ± 0.134
	IMPUTE5	0.935 ± 0.142	0.956 ± 0.128	0.909 ± 0.215	0.942 ± 0.206	0.52 ± 0.2	0.918 ± 0.123

Accuracy is shown separately for autosomal chromosomes (Autosome: Chr 1, 5, 10, 15, 20, 25), the non-PAR and PAR of the X chromosome

[1] Accuracy was calculated as the correlation between imputed and real genotypes, where imputed genotypes were either best guess (GT) or allele dosage (DS). The accuracy for GT was calculated only from the union of variants that were imputed as segregating across all three software tools, while DS accuracy is shown for both the GT union set as well as for all variants segregating in the real genotype data

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 7 of 14

Table 2 Comparison of imputation accuracy (correlation between real and imputed allele dosage genotypes) for all imputed SNP and INDEL, using Beagle 5.2, Minimac4, or IMPUTE5

	Autosome (Number of Variants = 8,890,108)				Non-PAR (Number of Variants = 763,172)				PAR (Number of Variants = 143,831)			
	LD		HD		LD		HD		LD		HD	
	INDEL	SNP	INDEL	SNP	INDEL	SNP	INDEL	SNP	INDEL	SNP	INDEL	SNP
Beagle 5.2	0.892	0.944	0.912	0.965	0.780	0.920	0.813	0.954	0.481	0.525	0.857	0.925
Minimac4	0.867	0.928	0.888	0.952	0.752	0.908	0.784	0.943	0.472	0.517	0.843	0.920
IMPUTE5	0.884	0.941	0.904	0.962	0.785	0.922	0.816	0.955	0.483	0.525	0.855	0.923

Variants were imputed either direct from real HD (high density) or from LD (low density) genotypes. The accuracies are shown separately for autosomes (Chr 1, 5, 10, 15, 20, 25), the non-PAR and the PAR of the X chromosome

shown in Fig. 3, when imputing from either LD or HD, and using DS mode for all three imputation programs. The results are provided for variants on all chromosomes (1, 5, 10, 15, 20, 25, non-PAR and PAR) and the equivalent results with GT mode are also available (See Additional file 4, Figure S2). There was a strong relationship between Rsq_{soft} and Rsq_{emp} although it is not linear and is very different for Minimac4 compared to Beagle 5.2 and IMPUTE5. For all imputation tools, Rsq_{soft} at higher values shows a closer relationship with Rsq_{emp}. The boxplot distributions show considerable variation of Rsq_{emp} within the bins of Rsq_{soft} at low to mid-range values, particularly for Beagle and IMPUTE5. This is partly a function of there being fewer variants in these Rsq_{soft} bins. Conversely, at the higher Rsq_{soft} values that fall within the useful range for filtering, the boxplots become less dispersed indicating that the relationship between Rsq_{soft} and Rsq_{emp} is more reliable.

The results in Fig. 3 allow an equivalent Rsq_{soft} threshold to be identified for each software that could be employed to filter imputed data for downstream analysis. For example, an Rsq_{soft} threshold of ~0.4 in Minimac4 and ~0.9 for Beagle 5.2 and IMPUTE5 would remove sequence variants with an average Rsq_{emp} lower than ~0.8 ($r = 0.89$) for data imputed from HD genotypes.

It is known that some chromosome regions are difficult to accurately impute [24, 25], therefore it is also of interest to understand if use of an Rsq_{Soft} filter will accurately remove poorly imputed variants and thus improve the average Rsq_{emp} across these regions. Figure 4a shows Rsq_{emp} averaged across all variants within windows of

1 Mb (Chromosome 10) with no filter on Rsq_{Soft} compared to Fig. 4b where the variants were first filtered using an Rsq_{Soft} threshold (0.4 for Minimac4 and 0.9 for Beagle 5.2 and IMPUTE5). Filtering improved the average Rsq_{emp} across this entire chromosome but had a particularly large impact across a 4 Mb segment that showed very poor average Rsq_{emp} before filtering, as well as at the ends of the chromosome. The filtering evened out any differences between software in the distribution of the Rsq_{emp} across the chromosome except at the very poorly imputed region where Minimac4 still showed a lower Rsq_{emp}.

Clearly, the Rsq_{soft} threshold is a useful indicator of Rsq_{emp}, but it is also important to quantify the resulting proportion of false positives (FP: variants that passed the Rsq_{soft} threshold but did not reach the expected Rsq_{emp}) and false negatives (FN: variants that did not pass the Rsq_{soft} threshold but did achieve the desired Rsq_{emp}). The FP and FN results are shown in Table 3 after imposing the equivalent Rsq_{soft} thresholds (specific to Minimac4, Beagle and IMPUTE5) as demonstrated in Fig. 3 to achieve an average $Rsq_{emp} \geq 0.8$ (imputing from real HD genotypes and using DS genotypes). The proportion of variants passing the equivalent Rsq_{soft} thresholds was similar for all three imputation programs as was the proportion of false negatives and false positives, although Beagle 5.2 tended to show the lowest level of false negatives/positives.

We also investigated the effect of Rsq_{soft} filtering on empirical imputation accuracy for all variants grouped on minor allele frequency (MAF: Fig. 5) in

(See figure on next page.)

Fig. 3 Boxplots showing the relationship between bins of Rsq_{soft} (x-axis) and the distribution of Rsq_{emp} (y-axis) for imputed dosage (DS) genotypes on all chromosomes tested. The blue line represents the average Rsq_{emp} value at each bin and the grey line within each box represents the median value. The box contains the 25th to the 75th percentile of the data points. Whiskers extend to a maximum length of 1.5 times the interquartile range (IQR) beyond the box edges. Data points beyond the whiskers are represented by individual dots as outliers. Plots (**a**), **c** and **e** show results for imputation starting from Low-density genotypes (LD: Beagle 5.2, Minimac4 and IMPUTE5) while plots (**b**), **d** and **f** show results for imputation starting from high-density genotypes (HD: Beagle 5.2, Minimac4 and IMPUTE5)
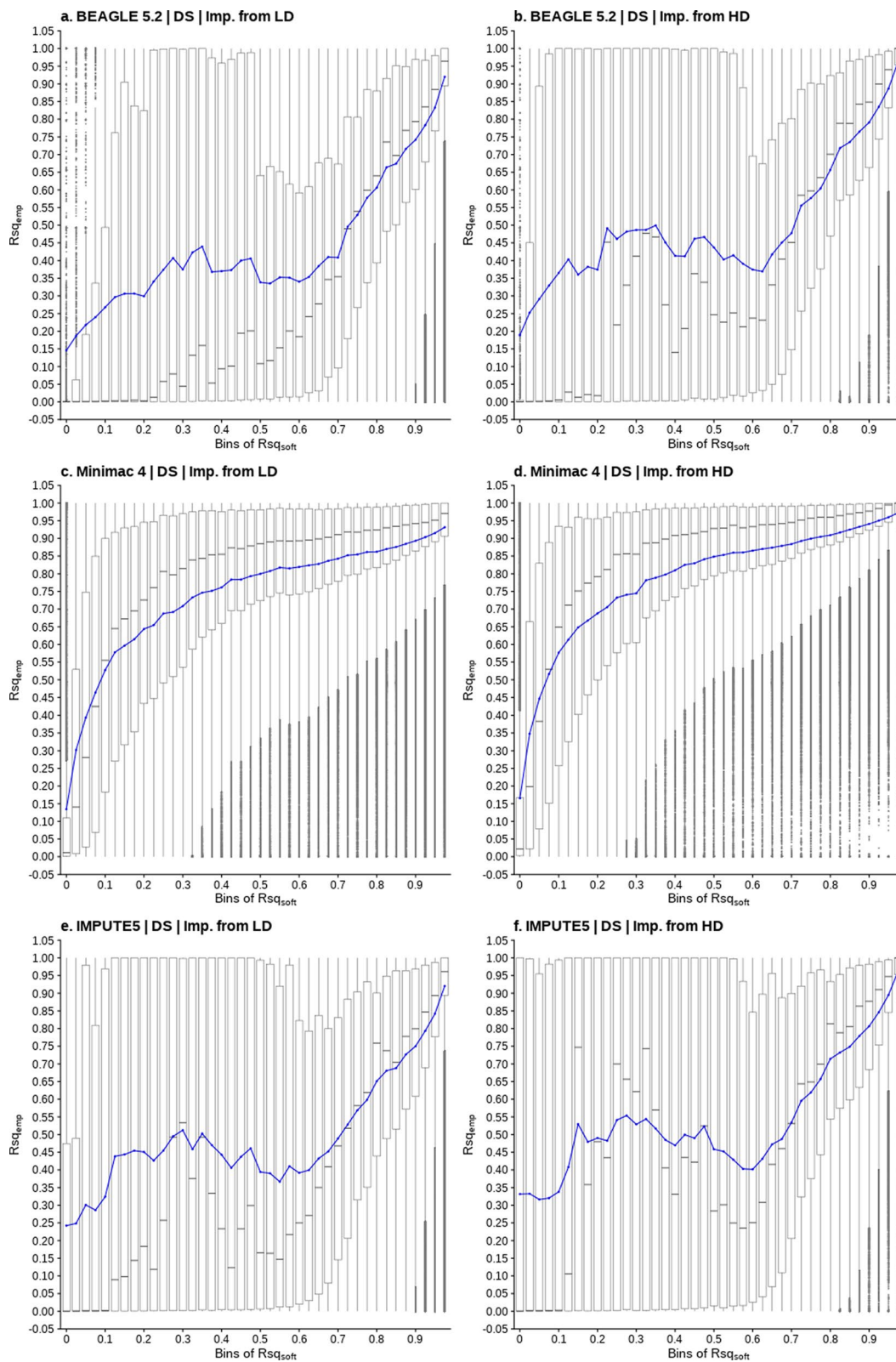
Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 8 of 14

Fig. 3 (See legend on previous page.)

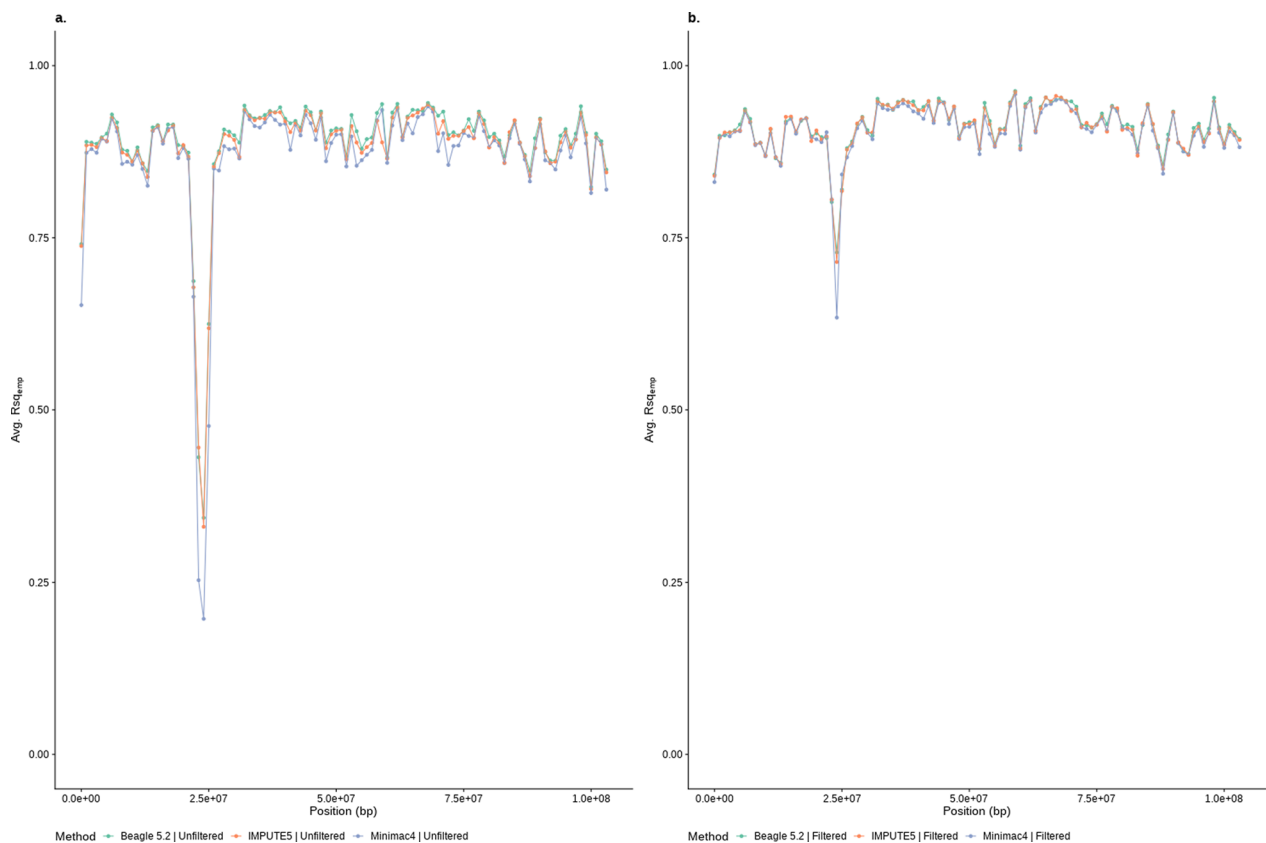Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 9 of 14

Fig. 4 Average imputation accuracy (Rsq_{emp}) for all variants in each adjacent 1 Mb window across chromosome 10. Using either no Rsq_{soft} filter (**a**) or after imposing an Rsq_{soft} threshold filter (**b**) for three different imputation programs (Rsq_{soft} threshold > 0.4 for Minimac4 and > 0.9 for Beagle 5.2 and IMPUTE5). Variants imputed from low density for imputed dosage (DS) genotypes

Table 3 Comparison of variant filtering using equivalent Rsq_{soft} thresholds from three imputation tools (Beagle 5.2, Minimac4 and IMPUTE5) to achieve an empirical accuracy (Rsq_{emp}) > 0.8

Imputation Software	Total No. of variants passing Rsq_{soft} (% of all variants imputed)	True positives: No. variants passing Rsq_{soft} & Rsq_{emp} (expressed as % of variants passing Rsq_{soft})	False positives: No. variants passing Rsq_{soft} & failing Rsq_{emp} (expressed as % of variants passing Rsq_{soft})	False negatives: No. variants not passing Rsq_{soft} but passing expected Rsq_{emp} (expressed as % of variants not passing the Rsq_{soft})
Beagle 5.2	3,991,626 (40.7%)	3,748,177 (93.9%)	243,449 (6.1%)	66,135 (1.1%)
IMPUTE5	3,954,595 (40.4%)	3,702,798 (93.6%)	251,797 (6.4%)	101,575 (1.7%)
Minimac4	3,949,886 (40.3%)	3,679,117 (93.1%)	270,769 (6.9%)	88,715 (1.5%)

The Rsq_{soft} thresholds were 0.4 for Minimac4, 0.9 for Beagle 5.2 and 0.9 for IMPUTE5 and were applied to dosage genotypes imputed from real HD genotypes

the imputation reference population. As expected, with no Rsq_{soft} filter, the lower MAF variant bins show considerably lower empirical accuracy compared to MAF > 0.025 (Fig. 5a). Conversely, after applying an Rsq_{soft} filter (> 0.4 for Minimac4, and > 0.9 for Beagle & IMPUTE5), the average empirical accuracy for the remaining low MAF variants was significantly

increased and very close to those with higher MAF (Fig. 5b).

We then divided all imputed variants into two groups: either common (MAF > 0.05) or less common (MAF < 0.05) and compared the effect of applying an Rsq_{soft} threshold on the empirical imputation accuracy of INDEL and SNP (Table 4). Generally, the Rsq_{soft} filter for all software removed a similar proportion of variants
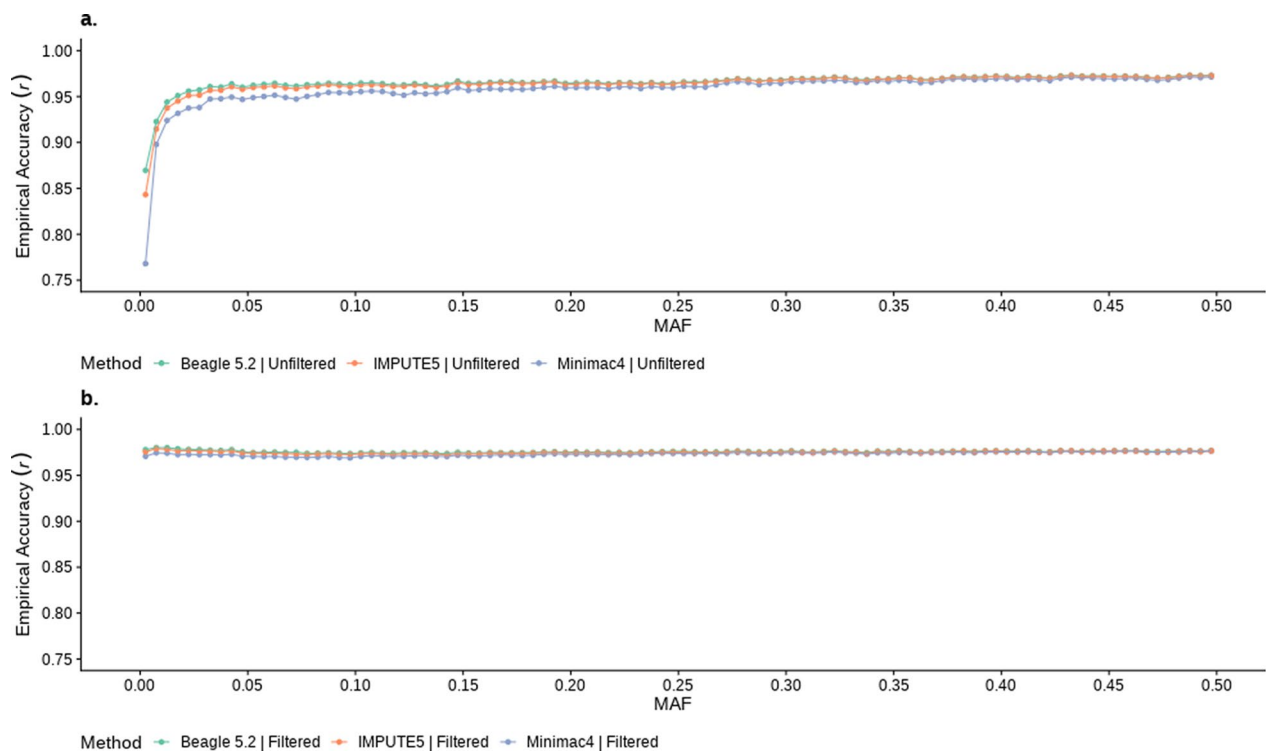
Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 10 of 14

Fig. 5 Average imputation accuracy (Rsq_{emp}) for sequence variants binned by minor allele frequency (MAF) using three different imputation programs (Beagle 5.2, Minimac4 and IMPUTE5). (**a**) shows Rsq_{emp} without imposing any Rsq_{soft} filter, while (**b**) is Rsq_{emp} after imposing an Rsq_{soft} filter ($Rsq_{soft} > 0.4$ for Minimac4, > 0.9 for Beagle 5.2 and IMPUTE5). Imputation started from real HD genotypes and accuracy was assessed for dosage genotypes

Table 4 Comparison of variant filtering using equivalent Rsq_{soft} thresholds from three imputation tools (Beagle 5.2, Minimac4 and IMPUTE5), separated by SNP and INDEL variants with MAF either above or below 0.05

		SNP			INDEL		
	Imputation Software	Percent of variants passing Rsqsoft (Number of variant)	False positives: Percent of variants passing Rsq_{soft} but not Rsq_{emp} threshold	False negatives: Percent of variants not passing Rsq_{soft} but passing Rsq_{emp} threshold	Percent of variants passing Rsq_{soft} (Number of variant)	False positives: Percent of variants passing Rsq_{soft} but not Rsq_{emp} threshold	False negatives: Percent of variants not passing Rsq_{soft} but passing Rsq_{emp} threshold
Common variants (MAF > 0.05 in Reference population)	Beagle 5.2	95.82 (2,671,196)	4.80	0.52	93.54 (230,666)	6.31	0.55
	Minimac4	95.08 (2,663,244)	5.20	0.84	93.47 (230,494)	7.06	0.81
	IMPUTE5	95.53 (2,650,696)	4.95	0.77	93.33 (230,135)	6.49	0.77
Less common Variants (MAF < 0.05 in Reference population)	Beagle 5.2	16.07 (1,006,321)	1.31	0.73	16.63 (83,443)	2.35	0.87
	Minimac4	15.73 (985,045)	1.51	0.93	15.97 (80,127)	2.74	0.96
	IMPUTE5	15.64 (979,105)	1.38	1.15	16.10 (80,807)	2.32	1.28

Variants were imputed from real HD genotypes and accuracy was assessed using dosage genotypes. The Rsq_{soft} thresholds were chosen for each software to achieve an Rsq_{emp} of ≥ 0.8 (Rsq_{soft} of 0.4 for Minimac4, 0.9 for Beagle 5.2 and 0.9 for IMPUTE5). The percentage of variants passing the Rsq_{soft} filter is shown for each of the variant categories, as well as the percentage of false positives and negatives

for either SNP or INDEL. However, the proportion of common variants (MAF > 0.05) passing the Rsq_{soft} filters was very high (> 93%) while only ~ 16% of less common variants passed the filters. One interesting phenomenon we observed is that the false positive rate of INDEL was always higher than that of SNP (for both common and

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 11 of 14

less common MAF) implying that the INDEL empirical imputation accuracy is less well predicted by the Rsq_{soft} filter than it is for SNP.

Discussion

This study provides a comprehensive evaluation of empirical imputation accuracy of sequence variants using three popular and computationally efficient imputation programs that provide an internal quality statistic of imputation accuracy (Rsq_{soft}). This is a unique study evaluating the empirical accuracy of imputing LD (7K) and HD genotypes to sequence: providing direct comparisons of accuracy for the PAR, non-PAR and autosomes, as well as INDEL and SNP variants. An additional novelty of this study was an in-depth evaluation of the relationship between Rsq_{soft} and Rsq_{emp} for the three imputation programs, as well as an assessment of the value of in-house pre-filtering of sequence variants in the reference population.

We demonstrated that the Rsq_{soft} filtering thresholds were similar for Beagle 5.2 and IMPUTE5 but differed for Minimac4. While each tool employs a different Rsq_{soft} algorithm, the Beagle 5.2 and IMPUTE5 Rsq_{soft} have previously been shown to be strongly correlated for the same imputation sample [1]. Beagle estimates the squared correlation between the imputed genotypes and true genotypes, where the (co)variance of the true genotypes is approximated using the sample mean of imputed genotypes [28, 37]. The IMPUTE5 algorithm measures the ratio of observed and complete information by considering the relative statistical information about the population allele frequency [1]. Minimac4 algorithm computes the average squared deviation of the imputed allele dosage at each haplotype in the sample relative to the estimated allele frequency, and divides this by the product of the alternate and reference allele frequency [38]. Despite these differences, our empirical tests demonstrated that if an appropriate Rsq_{soft} filter was applied, the majority of poorly imputed variants were removed, with those remaining having a higher and similar average empirical imputation accuracy. Importantly, Fig. 3 provides an empirical determination of the equivalent Rsq_{soft} thresholds across software that would maintain a common baseline Rsq_{emp} for downstream analyses (such as meta-GWAS). Interestingly, the Rsq_{soft} of both Beagle 5.2 and IMPUTE5 shows an overprediction of the Rsq_{emp} for the higher Rsq_{soft} values while the reverse is true of Minimac4. A previous study using Minimac3 to impute sequence variants in sheep [39] reported a very similar relationship to that found in our study and a study in chickens also reported that Beagle Rsq_{soft} overpredicted the Rsq_{emp} [40]. We also demonstrated that in practice, the application of the equivalent thresholds

from the three imputation tools resulted in similar numbers of variants being discarded and reasonably numbers of low false positives and negatives. A recent study has attempted to address the variability of Rsq_{soft} by developing a machine learning based quality calibration measure [41], but a drawback of that approach is that the model must first be trained on real, high quality genotypes at a range of frequencies.

Typically, for large-scale sequence imputation the software can parallelise the workload by splitting chromosomes into smaller segments ('chunking') which can dramatically speed up the imputation time required but may incur a penalty on accuracy. Several studies have documented the trade-off between accuracy and computational efficiency of imputation programs previously [30, 42–44]. We were concerned that the default chunking in Minimac4 and Beagle 5.2 might significantly reduce imputation accuracy because cattle breeds have small effective population sizes that can result in long haplotype blocks from extended regions of high linkage disequilibrium [45]. Interestingly, we found no advantage in imputation accuracy when the default setting was modified to allow entire chromosome imputation without chunking. This may partly relate to the current limitations of accurately defining longer haplotypes within short-read sequence [46]. We did not explore reducing chunk size settings below the default settings because these tools were primarily developed and tested in human data where the effective population size is generally much larger than in cattle.

We did not benchmark the computational efficiency of the software because a comprehensive study with human data has been published [30] where it was shown that the relative efficiencies between software changed with the size of the reference population. Using a modest sized reference population (N=2504) and chunk size of 20 Mb for all software, Minimac4 imputation was found to be considerable slower than IMPUTE5 and Beagle v5.1, while IMPUTE5 used the least memory [30]. This suggests that if imputation time is a concern, it may be preferable to use either Beagle 5.2 or IMPUTE5 rather than using Minimac4. An additional study with human data benchmarked computational efficiency of the same software based on their default settings (no chunking in IMPUTE5, 20 Mb chunks in Minimac4 and 40 Mb in Beagle 5.2) and including phasing time [42]. They reported that Beagle 5.4 (with Beagle 5.4 phasing) was considerably faster than IMPUTE5 or Minimac4 when both the latter were pre-phased with Eagle 2.4.1.

As expected, imputed allele dosage genotypes were more highly correlated to the real genotypes than imputed best-guess genotypes and previous literature has shown that there may be a small benefit in using

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 12 of 14

allele dosage versus best guess genotypes for GWAS [6]. However, there is an increased computational burden incurred for analyses that use dosage genotypes rather than best guess genotypes.

Our finding of reduced imputation accuracy for low MAF variants agrees with several previous studies where a larger multibreed reference helped to improve the accuracy of low MAF variants compared to a single breed [24]. However, only one previous study in cattle reported the accuracy of imputing sequence INDEL compared to SNP, and as in our study they reported that INDEL were less well imputed than SNP [47]. In general, given that INDEL can vary in length and complexity, and may overlap SNP sites, it is plausible that they are more difficult to accurately detect and impute. In addition, it is possible that a higher proportion of segregating INDEL tend to be more recent mutations than SNP, being more likely to be functionally disruptive and under stronger purging selection. Recent mutations may be harder to impute accurately compared to older mutations because some haplotypes surrounding the mutation may still be segregating in the reference population without the mutation [48]

We found no other cattle studies that have specifically considered the imputation accuracy of sequence variants from SNP chip to sequence on the PAR and non-PAR of the X chromosome. Two previous studies reported accuracy of imputation from LD to 50K SNP density for autosomes, the non-PAR and PAR [49, 50] using the previous cattle reference genome (UMD3.1) and reported much lower imputation accuracy for the PAR. Both those studies suggested that the low accuracy may be partly attributed to the short length of the PAR (5.7 Mb). We tested this hypothesis in a recent study [51] by comparing the PAR imputation accuracy with the accuracy of imputing short 5.7 Mb autosomal segments that were extracted from the ends of five different chromosomes (with SNP density was adjusted to match). These autosomal segments were imputed from LD genotypes to 50K and HD genotypes and although their imputation accuracy was lower than for whole autosomal chromosomes, it was still considerably higher than for the PAR [51]. Furthermore, in the current study we demonstrated that when imputing from real HD genotypes to sequence, the PAR imputation recovered considerable accuracy (Table 1). This suggests that the low imputation accuracy for LD genotypes on the PAR is likely due to the higher recombination rate of the PAR [52, 53] and is therefore a generalised result that would be observed across cattle populations. Although the PAR represents only 4% of the X chromosome, it would be useful to increase the density of PAR markers compared to other chromosomes when generating custom SNP panels, because this could significantly improve the PAR imputation accuracy.

In contrast to the PAR, our study showed that the imputation accuracy on the non-PAR was similar to the autosomes. This is important because many cattle studies that use imputed sequence in downstream studies such as GWAS, have discarded the entire X chromosome before undertaking downstream analyses, suggesting a lost opportunity (e.g. [54]). Our target animals were males, and potentially because males are haploid across the non-PAR (presented as diploid to the imputation software), they may be a little more accurately phased than diploid females. If phasing is a little less accurate in target females this could lead to a lower imputation accuracy than for males, but this has not been tested empirically because we did not have equivalent female samples.

Conclusions

Our findings offer valuable insights on the application of imputation filters across software for downstream analyses such as meta-GWAS studies. It is likely that these imputation tools will remain popular because they showed similarly high imputation accuracy. This study demonstrates that Rsq_{soft} is a useful filtering tool for both SNP and INDEL. We provide a generalised empirical determination for equivalent Rsq_{soft} thresholds across the three imputation tools. The extremely low accuracy observed for imputation from LD to sequence on the PAR of the X chromosome indicates that imputed data in this region cannot be confidently used for downstream analyses.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-024-00942-2.

Additional file 1: Protocol to process sequence data in Run8 of the 1000 Bull Genomes Project [27, 55, 56]. Text documentation describing the pipeline and software versions used to process sequence data.

Additional file 2: Detail statistics of the analysis. Excel file containing three sheets. Sheet 1 - Statistics on Phasing result between Dosage (DS) vs Genotype likelihoods (GT) mode. Sheet 2 – Correlation statistics between different mode of running of Beagle and MINIMAC4. Sheet 3 – Correlation statistics when imputing from LD or HD across 3 programs Beagle 5.2, MINIMAC4 and IMPUTE5.

Additional file 3: Graphical schematics of imputation cases used in the current analysis. REF represents the reference allele, ALT represents the alternative allele.

Additional file 4: Boxplot visualizing relationship between Rsq_{soft} (Dosage R-squared - DR2) and Rsq_{emp} (Empirical Correlation Squared) using imputed Genotype likelihoods (GT) on all chromosome tested using 3 programs Beagle 5.2, MINIMAC4 and IMPUTE5. The box contains the 25th to the 75th percentile of the data points. Whiskers extend to a maximum length of 1.5 times the interquartile range (IQR) beyond the box edges. Data points beyond the whiskers are represented by individual dots as outliers. (optional).

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 13 of 14

Availability of data and materials
All public sequence data from Run8 and Run9 of the 1000 Bull Genomes Project are accessible with the accession number PRJEB42783 and PRJEB56689, respectively.

Declarations

Ethics approval and consent to participate.
Not applicable.

Consent for publication
The authors all give their consent for publication.

Competing interests
The authors declare that they have no competing interests.

References
1. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.
2. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78:629–44.
3. Sanchez M-P, Govignon-Gion A, Croiseau P, Fritz S, Hozé C, Miranda G, et al. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol. 2017;49:68.
4. Reynolds EGM, Neeley C, Lopdell TJ, Keehan M, Dittmer K, Harland CS, et al. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat Genet. 2021;53:949–54.
5. Bolormaa S, Swan AA, Stothard P, Khansefid M, Moghaddar N, Duijvesteijn N, et al. A conditional multi-trait sequence GWAS discovers pleiotropic candidate genes and variants for sheep wool, skin wrinkle and breech cover traits. Genet Sel Evol. 2021;53:58.
6. Brondum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics. 2014;15:728.
7. Xiang R, MacLeod IM, Daetwyler HD, de Jong G, O'Connor E, Schrooten C, et al. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nat Commun. 2021;12:860.
8. Moghaddar N, Khansefid M, van der Werf JHJ, Bolormaa S, Duijvesteijn N, Clark SA, et al. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. Genet Sel Evol. 2019;51:72.
9. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173–86.
10. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nat Rev Method Primers. 2021;1:59.
11. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20:467–84.
12. Arvanitis M, Tampakakis E, Zhang Y, Wang W, Auton A, Agee M, et al. Genome-wide association and multi-omic analyses reveal ACTN2 as a gene linked to heart failure. Nat Commun. 2020;11:1122.
13. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in—700000 individuals of European ancestry. Hum Mol Genet. 2018;27:3641–9.
14. Zhao J, Sauvage C, Zhao J, Bitton F, Bauchet G, Liu D, et al. Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. Nat Commun. 2019;10:1534.
15. Shook JM, Zhang J, Jones SE, Singh A, Diers BW, Singh AK. Meta-GWAS for quantitative trait loci identification in soybean. Genes Genom Genet. 2021;11:jkab117.
16. Zhou S, Ding R, Meng F, Wang X, Zhuang Z, Quan J, et al. A meta-analysis of genome-wide association studies for average daily gain and lean meat percentage in two Duroc pig populations. BMC Genomics. 2021;22:12.
17. van den Berg I, Xiang R, Jenko J, Pausch H, Boussaha M, Schrooten C, et al. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. Genet Sel Evol. 2020;52:37.
18. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018;50:362–7.
19. Hickey JM, Kinghorn BP, Tier B, van der Werf JHJ, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44:9.
20. VanRaden PM, Sun C, O'Connell JR. Fast imputation using medium or low-coverage sequence data. BMC Genet. 2015;16:82.
21. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
22. Calus MPL, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal. 2014;8:1743–53.
23. van Binsbergen R, Bink MC, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46:41.
24. 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 Sel Evol. 2017;49:24.
25. 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 Sel Evol. 2021;53:27.
26. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Giga Sci. 2020;9:giaa021.
27. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
28. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338–48.
29. 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.

Nguyen *et al. Genetics Selection Evolution*   (2024) 56:72

Page 14 of 14

30. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the positional burrows wheeler transform. PLoS Genet. 2020;16: e1009049.
31. Johnson T, Keehan M, Harland C, Lopdell T, Spelman RJ, Davis SR, et al. Short communication: Identification of the pseudoautosomal region in the Hereford bovine reference genome assembly ARS-UCD1.2. J Dairy Sci. 2019;102:3254–8.
32. Hayes BJ, Daetwyler HD. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. Annu Rev Anim Biosci. 2019;7:89–102.
33. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GVD, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018. https://doi.org/10.1101/201178.
34. MacLeod IM, Bolormaa S, Vander Jagt CJ, Nguyen TV, Chamberlain AJ, Daetwyler HD. Current challenges for imputation of SNP chips to whole-genome sequence in cattle and sheep. Proceedings 24th Conf Assoc Adv Animal Breed Genet. 2021;24:212–6.
35. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef AY, Finucane KH, et al. Reference-based phasing using the haplotype reference consortium panel. Nat Genet. 2016;48:1443–8.
36. Consortium TBH, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324:528–32.
37. 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.
38. Das S. Minimac3 info file specification 2018. https://genome.sph.umich.edu/wiki/Minimac3_Info_File. Accessed 04 Oct 2024.
39. Bolormaa S, Chamberlain AJ, Khansefid M, Stothard P, Swan AA, Mason B, et al. Accuracy of imputation to whole-genome sequence in sheep. Genet Sel Evol. 2019;51:1.
40. Heidaritabar M, Calus MP, Vereijken A, Groenen MA, Bastiaansen JW. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genet. 2015;16:101.
41. Sun Q, Yang Y, Rosen JD, Jiang M-Z, Chen J, Liu W, et al. MagicalRsq: machine-learning-based genotype imputation quality calibration. Am J Hum Genet. 2022;109:1986–97.
42. De Marino A, 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.
43. Dias R, Evans D, Chen S-F, Chen K-Y, Loguercio S, Chan L, et al. Rapid, reference-free human genotype imputation with denoising autoencoders. eLife. 2022;11:75600.
44. Browning Brian L, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016;98:116–26.
45. Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ. High-resolution haplotype block structure in the cattle genome. BMC Genet. 2009;10:19.
46. Garg S. Computational methods for chromosome-scale haplotype reconstruction. Genome Biol. 2021;22:101.
47. Mesbah-Uddin M, Guldbrandtsen B, Lund MS, Boichard D, Sahana G. Joint imputation of whole-genome sequence variants and large chromosomal deletions in cattle. J Dairy Sci. 2019;102:11193–206.
48. Fritz S, Capitan A, Djari A, Rodriguez SC, Barbat A, Baur A, et al. Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. PLoS ONE. 2013;8: e65550.
49. Su G, Guldbrandtsen B, Aamand GP, Strandén I, Lund MS. Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. Genet Sel Evol. 2014;46:47.
50. Mao X, Johansson AM, Sahana G, Guldbrandtsen B, De Koning D-J. Short communication: imputation of markers on the bovine X chromosome. J Dairy Sci. 2016;99:7313–8.
51. Nguyen T, Bolormaa S, Reich C, Chamberlain A, Medley A, Schrooten C, et al. Exploring imputation accuracy across the bovine X chromosome. Proc Assoc Advmt Anim Breed Genet. 2021;24:264–8.
52. Zhang J, Kadri NK, Mullaart E, Spelman R, Fritz S, Boichard D, et al. Genetic architecture of individual variation in recombination rate on the X chromosome in cattle. Heredity. 2020;125:304–16.
53. Van Laere AS, Coppieters W, Georges M. Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. Genome Res. 2008;18:1884–95.
54. Cai Z, Iso-Touru T, Sanchez M-P, Kadri N, Bouwman AC, Chitneedi PK, et al. Meta-analysis of six dairy cattle breeds reveals biologically relevant candidate genes for mastitis resistance. Genet Sel Evol. 2024;56:54.
55. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014;508:494–9.
56. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Zimin A, et al. Modernizing the bovine reference genome assembly. In: Proceedings of the 11th World Congress of Genetics Applied to Livestock Production: 11–16 February 2018; Auckland. 2018.

Publisher's Note