Accuracy of Whole-genome Sequence Genotype Imputation in Cattle Breeds

H. Li,* M. Sargolzaei,*,† and F. Schenkel*

*Centre for Genetic Improvement of Livestock,
University of Guelph, Guelph, Ontario, Canada, †The Semex Alliance, Guelph, Ontario, Canada

ABSTRACT: Increase in accuracy of genomic predictions can possibly be achieved by using sequence genotypes. Large number of sequence genotypes can be obtained by imputation. This study compared the use of FImpute and BEAGLE for sequence genotype imputation from Bovine Illumina HD and 54k chips using 357 sequence genotypes from six cattle breeds, including Angus, Brown Swiss, Holstein, Jersey, Limousin, and Simmental. In addition the use of a multi-breed reference population was evaluated. FImpute was more accurate and faster than BEAGLE in all scenarios investigated. As expected, imputation from HD to sequence was substantially more accurate than from 54k genotypes. Accuracy of imputation from HD to sequence genotypes was increased when a multi-breed reference population was used. FImpute imputed more accurately rare allele variants than BEAGLE. Therefore, using a multi-breed reference population and FImpute is suggested for imputation from HD to sequence genotypes in the cattle breeds analyzed.

Keywords: cattle; whole-genome sequence; genotype imputation.

INTRODUCTION

Whole-genome sequence data could improve the accuracy of genomic predictions by capturing the causal mutations affecting a trait, without dependence on the extent of LD between SNP markers and causal mutations (Meuwissen and Goddard (2010), Druet et al. (2013)). A cost effective sequencing strategy is to sequence key ancestors, who contributed most of the genetic material to the current population plus a random sample of individuals to maximize the chance of sampling rare allele/haplotype variants and, then, impute sequence data to the rest of the animals genotyped with commercial SNP chips. The 1,000 Bull Genomes Project aids the goal of implementing genomic selection using whole-genome sequence genotypes by providing an extended cattle sequence data base of key ancestors from several breeds, allowing for imputation of full sequence genotypes from SNP chip genotypes. Because the possible increase in accuracy of genomic predictions by using imputed sequence data is mainly determined by the accuracy of imputation and by the allele frequency distribution of the QTLs (Druet et al. (2013)), the objectives of this study were: 1) to evaluate performance, in terms of accuracy and computation efficiency, of two imputation programs: FImpute and BEAGLE to impute full-genome sequence genotypes from commercial SNP chip genotypes in cattle; 2) to evaluate the effect of minor allele frequency on imputation accuracy.

MATERIALS AND METHODS

Data. The most recent run of the 1,000 Bull Genomes Project (run 3) was used and included 429 full genome sequences of 427 bulls and 2 cows from 15 breeds, sequenced at an average of 10.1 fold coverage. There were 30.8 million filtered sequence variants detected, including 29.1 million SNPs and 1.7 million insertion-deletions. The sequence data across all breeds included 28,336,153 SNP on autosomal chromosomes. Six breeds with more than or equal to 25 sequenced animals were used in this study, which included Angus (AN), Brown Swiss (BS), Holstein (HO), Jersey (JE), Limousin (LI), and Simmental (SI).

Analyses. Accuracy of genotype imputation from Bovine Illumina High Density (HD) or Illumina 54k SNP chips (Illumina Inc., San Diego, CA, USA) to whole-genome sequence genotypes was investigated. Different scenarios were evaluated by masking sequence genotypes to mimic animals genotyped with the HD and 54k SNP chips, resulting in 657,585 and 47,427 SNPs in the mimicked HD and 54k SNP chips, respectively. The number of animals, number of SNPs with MAF greater than zero, and number of overlapping SNPs between either the HD or the 54k SNP chip and the whole-genome sequence genotype data within each breed are given in Table 1.

FImpute 2.2 (Sargolzaei et al. (2011)) and BEAGLE 3.3.2 (Browning and Browning (2009)) programs were used and their performances in terms of imputation accuracy and computational efficiency were compared. Accuracy of imputation was assessed by both the percentage of correctly imputed genotypes (concordance rate, CR %) and the estimated squared correlation between true allele dosage and imputed genotypes (allelic R^2 %). In order to evaluate the effect of increasing the number of reference animals by using other breeds on the accuracy of sequence genotype imputation, a combined multi-breed reference population was used to impute each breed using FImpute.

Due to the high computing time required for BEAGLE, FImpute and BEAGLE were compared using a single cross-validation, randomly selecting a validation set corresponding to about 20% of the sequenced animals in a breed. However, all the imputation scenarios were cross-validated by 5-fold cross-validation using FImpute only.

The relationship between accuracy of sequence genotype imputation and minor allele frequency (MAF) was also investigated in the Holstein breed, which was the breed with the largest number of sequenced animals.

RESULTS AND DISCUSSION

Imputation programs. Imputation using FImpute and BEAGLE was evaluated on the same datasets and using the same server computer. Results from different scenarios

Table 1. Number of animals and SNP with MAF>0 in the whole-genome sequence data, and number of matched SNPs to the Illumina HD and 54k Beadchip panels on autosomal chromosomes for each breed.

Breed[1]	N	# SNP MAF>0	# SNP HD chip	# SNP 54k chip
AN	54	14,928,662	601,371	43,914
BS	43	16,497,103	607,160	42,725
HO	121	18,445,629	617,046	44,908
JE	27	13,096,444	568,176	39,784
LI	25	16,694,290	617,664	43,851
SI	87	20,378,658	633,334	45,326

[1]Breed= Angus (AN), Brown Swiss (BS), Holstein (HO), Jersey (JE), Limousin (LI), and Simmental (SI)

are presented in Table 2a (imputation from HD to sequence) and Table 2b (imputation from 54k to sequence) for each breed. As expected, imputation from the HD SNP chip to sequence genotypes was more accurate than from the 54k SNP chip. FImpute was slightly more accurate than BEAGLE for HD, but much more accurate for 54k to sequence imputation.

BEAGLE resulted in very long computing time in comparison to FImpute. In addition, FImpute required less memory, about 5 Gigabyte per chromosome for imputation from both HD and 54k panel, while BEAGLE took more than 80 Gigabyte of memory per chromosome. As an example, the CPU time for HD to sequence imputation in Angus cattle with 44 reference animals and 10 validation animals took more than 14 hours using BEAGLE, while it took about 2 hours using FImpute without parallel computation.

Single vs. multi-breed reference population. Results for single-breed and multi-breed reference populations from 5-fold cross validation using FImpute are shown in Tables 3 and 4. In the multi-breed reference population, two additional breeds (Finish Ayrshire (n=17 and Swedish Red (n=16)) were included, but due to the small number of animals sequenced for these breeds they did not have their own statistics calculated.

For single breed analysis, the average CR and allelic R^2 for HD to sequence imputation were between 0.86–0.95 and 0.80–0.93 for different breeds, while the values were between 0.75–0.90 and 0.63–0.86 for 54k to sequence imputation. The imputation accuracy seemed to depend on the number of reference animals within each breed in both HD and 54k chip analyses, where Holstein with the largest reference population had the highest accuracy and Limousin with the smallest reference group showed the lowest accuracy. However, the effective population size (level of LD) seemed also to play an important role, as can be seen for Jersey breed that had a small reference population, but showed higher accuracy of imputation than Limousin. For combined multi-breed reference population, the average concordance rate and allelic R^2 for HD to sequence imputation were between 0.90–0.95 and 0.85–0.93 for each breed, respectively, while the values were between 0.77–0.89 and 0.65–0.84 for 54k to sequence imputation, respectively. Imputation from HD genotypes using combined sequence information from different breeds improved imputation accuracy for all breeds. On average CR and allelic R^2 increased from 0.92-0.93 and from 0.88-0.90, respectively. The observed

Table 2. Single breed whole-genome sequence imputation using breed specific SNPs with MAF>0 in scenarios: a) imputing from HD chip; b) imputing from 54k chip.

a)

Breed[1]	n_R:n_V[2]	FImpute[3]		BEAGLE[3]	
		CR	R^2	CR	R^2
AN	44:10	94.6	91.9	93.2	89.3
BS	33:10	93.6	90.2	91.2	86.0
HO	101:20	94.5	92.1	94.3	91.3
JE	20:7	92.6	88.0	89.4	82.4
LI	20:5	86.5	79.9	84.4	75.9
SI	67:20	92.9	90.0	92.3	88.7
Mean		*92.4*	*88.7*	*90.8*	*85.6*

b)

Breed[1]	n_R:n_V[2]	FImpute[3]		BEAGLE[3]	
		CR	R^2	CR	R^2
AN	44:10	89.3	83.5	77.7	61.3
BS	33:10	87.2	80.5	73.9	54.9
HO	101:20	88.9	83.8	79.8	66.7
JE	20:7	85.3	75.5	72.7	51.3
LI	20:5	75.9	63.8	71.4	50.9
SI	67:20	84.5	77.6	77.9	62.0
Mean		*85.2*	*77.4*	*75.6*	*57.8*

[1] Breed= Angus (AN), Brown Swiss (BS), Holstein (HO), Jersey (JE), Limousin (LI), and Simmental (SI).
[2] n_R:n_V = Number of reference animals: number of validation animals.
[3] CR= Concordance Rate; R^2= Squared allelic correlation.

Table 3. Single breed sequence imputation from HD or 54 SNP chip using FImpute, breed specific SNPs with MAF>0, and 5-fold cross-validation.

Breed[1]	HD[2]		54k[2]	
	CR	R^2	CR	R^2
AN	92.7	89.2	86.9	80.2
BS	92.7	88.9	84.5	76.2
HO	94.8	92.6	90.0	85.5
JE	92.0	87.0	83.6	73.2
LI	86.4	79.7	75.3	62.9
SI	92.4	89.4	83.9	77.0
Mean	*91.8*	*87.8*	*84.0*	*75.8*

[1]Breed= Angus (AN), Brown Swiss (BS), Holstein (HO), Jersey (JE), Limousin (LI), and Simmental (SI).
[2]CR= Concordance Rate; R^2= Squared allelic correlation.

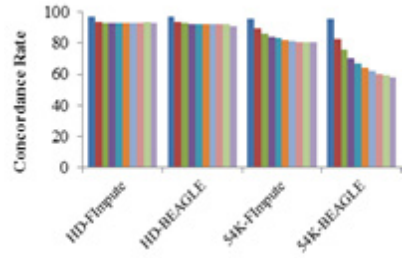
Table 4. Combined multi-breed sequence imputation from HD or 54 SNP chip using FImpute, breed specific SNPs with MAF>0, and 5-fold cross-validation.

Breed[1]	HD[2]		54k[2]	
	CR	R^2	CR	R^2
AN	93.3	90.2	86.1	78.9
BS	93.6	90.4	83.7	75.0
HO	94.9	92.7	89.2	84.4
JE	93.2	89.2	82.3	70.9
LI	89.7	85.3	76.7	65.4
SI	93.0	90.4	83.6	76.7
Mean	*92.9*	*89.7*	*83.6*	*75.2*

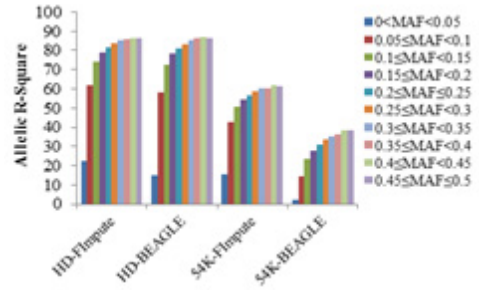
[1]Breed= Angus (AN), Brown Swiss (BS), Holstein (HO), Jersey (JE), Limousin (LI), and Simmental (SI).
[2]CR= Concordance Rate; R^2= Squared allelic correlation.

Figure 1. Accuracy of imputation of sequence genotypes from 54k or HD SNP chip measured as concordance rate (a) or allelic R-square (b) as a function of the minor allele frequency in Holsteins, using BEAGLE or FImpute.

a)



b)



increase was even more substantial for breeds with the lowest reference sets, such as Limousin. For imputation from 54k genotypes, using a multi-breed reference population yielded to a small decrease in CR and allelic R^2, except for the Limousin breed. This indicates that the 54k panel was not dense enough to capture small haplotypes shared among the breeds.

Minor allele frequency. Figure 1 displays the accuracy of imputation based on CR and allelic R^2 for different classes of MAF for the Holstein breed when using either FImpute or BEAGLE. In general, BEAGLE was more sensitive to MAF than FImpute, especially for 54k imputation. HD to sequence imputation using either FImpute or BEAGLE resulted in higher and more balanced concordance rate and allelic R^2 than 54k chip.

SNP with low MAF tended to show higher CR and lower allelic R^2. Concordance rate is a measure of how well genotypes are imputed, so it is logical that CR would be higher for SNPs with low MAF. Allelic R^2 is a measure of how well the allele dosage is imputed, what is more difficult to impute correctly when MAF is low. Depending on the goals after the imputation, CR or allelic R^2 could be more relevant than the other. For example, if rare alleles are targeted, such as for a disease, then allelic dosage might be more relevant. On the other hand, for prediction of genomic values of polygenic traits, genotype imputation accuracy becomes more relevant, because rare variants would explain small portion of the additive genetic variance.

For a given MAF, the expected number of minor alleles in a sample is dependent on the sample size. It might be, then, more reasonable to use the expected number of minor alleles in the sample for relating it to the accuracy of imputation. This approach would also allow for comparing across samples of different sizes or pooling their results. This approach will be investigated next.

CONCLUSION

Relatively high accuracy of sequence genotype imputation was obtained from the Bovine Illumina HD SNP panel, especially when a multi-breed reference population was used. FImpute outperformed BEAGLE particularly in the case of imputation from Bovine Illumina 54k SNP panel to sequence, with higher imputation accuracy, less CPU time, and less memory requirements. For imputation from HD genotypes, the breeds with small number of sequenced animals benefited more from sequence information from other breeds, making the imputation accuracy more similar among breeds. In general, BEAGLE was more sensitive to MAF than FImpute, especially for imputation from 54k genotypes. SNP with low MAF tended to show higher CR and lower allelic R^2. Therefore, using a multi-breed reference population and FImpute is suggested for imputation from HD genotypes to the sequence genotypes in the cattle breeds analyzed.

ACKNOWLEDGMENTS

The authors thank the agencies who have provided support for this research: Genome Canada, Genome Alberta, Alberta Livestock and Meat Agency (ALMA), the Government of Alberta and all collaborators and partners involved in the project “Whole Genome Selection Through Genome-wide Imputation in Beef Cattle” funded by the previously mentioned agencies.

LITERATURE CITED

Browning, B. L., and Browning, S. R. (2009). *Am. J. Hum. Genet.* 84:210-223.

Druet, T., Macleod, I. M., and Hayes, B. J. (2013). *Heredity* 1-9.

Meuwissen, T., and Goddard, M. E. (2010). *Genetics* 185: 623–631.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2011). *J. Anim. Sci.* 89, E-Suppl. 1 / *J. Dairy Sci.* 94, E-Suppl. 1: 421 (333).