Comparison of Different Imputation Methods

J. Johnston[1], G. Kistemaker[1] and P.G. Sullivan[1]
[1]Canadian Dairy Network, Guelph, ON

Abstract

Genotype imputation is a powerful tool to include animals genotyped with low density panels into higher density genomic evaluation without having to genotype them with more expensive high density panels. A range of different imputation software programs is available and it is necessary to compare their performance with data that represents the population for which the programs will be used. This study selected five imputation programs (AlphaImpute, BEAGLE, FImpute, findhap, PHASEBOOK) that were feasible for the structure and size of dairy genomic data sets and compared their ability to impute genotypes from the Illumina Bovine3K BEAD chip (3k) to genotypes from the Illumina Bovine SNP50 chip (50k) using two data sets. The first data set aimed to represent data from a small dairy breed with limited number of genotyped dams. The second data set mimicked a large dairy breed with many animals genotyped and majority of parents genotyped. All five compared programs performed very well in imputing genotypes from 3k to 50k density. On average, all of them imputed more than 90% genotypes correctly. Each of them had certain strengths and weaknesses. FImpute was the fastest program and was the most accurate software program for animals with family information. BEAGLE was the most accurate software for animals with limited family information. The choice of optimal imputation software is highly dependent on structure of the genotype file, namely the proportion of animals with their ancestors genotyped. Blending results from FImpute and BEAGLE seems to be a viable solution for current dairy data sets, which contains a proportion of animals without genotyped parents.

Introduction

Imputation became an essential part of national genomic evaluations of dairy cattle shortly after the Illumina Bovine3K BEAD chip was introduced. This technology allows cattle breeders to genotype their animals with this affordable low density panel and allows to predict the untyped genotypes on the higher density panel.

Imputation is a cost effective in silico genotyping of missing genotypes. Based on the sources of information used to infer the missing genotype, imputation methods can be divided into two groups: 1) family and 2) population based. Family based imputation uses linkage and Mendelian segregation rules and is the most accurate for animals with genotyped relatives. Population imputation uses linkage disequilibrium information between missing SNP and the observed flanking SNPs. Population imputation is useful for a set of unrelated animals or for animals without genotyped close ancestors.

The aim of this study was to evaluate performance of five imputation programs: AlphaImpute, BEAGLE, FImpute, findhap and PHASEBOOK to impute genotypes from low to high density.

There are other imputation programs available (BIMBAM, FAMHAP, fastPHASE, IMPUTE, MACH, PLINK, WHAP) but the majority of these programs were designed specifically for human populations, where sample size is relatively small and consequently some of the programs cannot handle the size of dairy cattle data sets. Another characteristic of human genomic data sets is the lack of pedigree information (most of the studied individuals are unrelated). In a population of unrelated animals, the shared haplotype stretches are shorter, because common ancestors are more distant, and more complex algorithms are needed for accurate imputation. This often results in extremely long computing time when it is applied to large data sets.

Material and Methods

Data

Two data sets were used to compare the ability of five imputation programs to recover masked genotypes. The first data set (**HOL**) consisted of 67,336 Holstein animals genotyped with the Illumina Bovine SNP50 chip (**50k**) and 39,247 Holstein animals genotyped with the Illumina Bovine3K BEAD chip (**3k**). The second data set (**BSW**) contained 2,042 50k genotypes and 304 3k genotypes. Holstein animals born after 2010 and genotyped with 50k panel (20,000 animals) were used as the study sample to evaluate imputation accuracy. Due to a smaller number of genotyped animals in the Brown Swiss breed, the study sample for BSW included animals born after 2009 and genotyped with 50k panel (209 animals).

Genotypes of animals in the study sample were reduced to 3k panel. Because animals genotyped with 50k panel tend to have deeper pedigree (more of their ancestors are genotyped) than animals genotyped with 3k panel, the family information of (50k) animals in the study sample was modified to mimic a family structure more similar to that observed for 3k animals in Canada. Animals were randomly assigned into four groups: 1) dam genotype was kept unchanged, 2) genotype of dam was reduced to 3k panel, 3) genotype of dam was removed, 4) genotype of dam was removed and dam was set to unknown. Table 1 shows distribution of the animals in those groups

Table 1. Description of the study samples.

Group	Data set	
	Brown Swiss	**Holstein**
Dam 50k	19	4,612
Dam 3k	9	4,212
Dam not genotyped	178	6,354
Dam unknown	3	4,822
All animals	209	20,000

In BSW, most animals had a dam that was not genotyped (they could not be assigned to group 1 or 2), which resulted in larger number of animals in the third group than in other groups. Similar problem was observed in HOL but the difference was not as dramatic.

After the modification, BSW had 1,829 high density (50k) and 515 low density (3k) genotypes, and HOL had 60,135 low density and 45,047 high density genotypes.

The 50k panel contained 42,503 SNPs (SNPs that are currently used for genomic evaluation in Canada) that were spread out across 29 autosomes and the X chromosome; 678 and 97 of those SNPs were on the X-specific and pseudo-autosomal region (PAR) of chromosome X. From the 3k panel, 2,614 SNPs were used, 135 were on the X-specific and 4 SNPs were on the PAR of chromosome X.

Compared programs

AlphaImpute, BEAGLE 3.3., FImpute 2.0, findhap 2, PHASEBOOK 2.3 were compared in this study. All five programs have the possibility to consider relationship between genotyped animals. BEAGLE has the ability to process offspring-parent trio but this option can only be used for very small data sets. With BEAGLE, it was assumed that all animals are unrelated. Genotypes of 1,270 Brown Swiss and 7,228 Holstein were used as reference population for Brown Swiss and Holstein imputation, respectively. BEAGLE's computational time tends to be long and it depends on the combined size of reference and study sample. In order to reduce the computational burden caused by the number of animals needed to be estimated for Holstein, the study sample was divided into 5 sub-groups and each of them was imputed independently. Each subset consisted of 4,000 study sample genotypes and 7,224 reference genotypes.

Description of programs

AlphaImpute (Hickey *et al.,* 2011) calculates allele probabilities using segregation analysis. Phase of all genotyped animals is determined using long range phasing and haplotype library imputation. Missing alleles are imputed by matching the allelic probabilities from the segregation analysis to the haplotypes from the phasing step. This software uses information from both related and unrelated animals but overall it can be classified as mainly family based imputation. This software requires use of certain "shortcuts" for large data sets. Unfortunately, we haven't been able to implement them before the deadline for this meeting. Therefore, results from this software are only presented for the smaller Brown Swiss data set.

BEAGLE (Browning *et al.,* 2007) makes use of Hidden Markov Model to predict the missing genotypes of SNPs. It constructs a tree of haplotypes that are present in the reference population and then summarizes it in a direct acyclic graph (DAG) by joining nodes of the tree based on haplotype similarity. The most likely path in the graph is inferred for each haplotype. When two haplotypes originate from the same distant ancestor, they should follow the same path on the graph for a short distance. If the segment of the chromosome is received from a very recent ancestor (parent), the haplotype will match the path of the ancestor for much longer distance. This way BEAGLE indirectly uses information about relationship between animals.

FImpute (Sargolzaei *et al.,* 2011) is predominantly family rule based imputation but the most recent version of the program (version 2.0) also has a population based component.

findhap (VanRaden *et al.,* 2011) combines both population and family haplotyping. In the population step, it divides chromosome into blocks of "x" SNPs and generates a library of haplotype blocks, which is sorted by frequency. Haplotypes from low density panels are searched in the library until a match is found. Unknown alleles are replaced by alleles from the matched haplotype. The animal's second haplotype is obtained by subtracting its first haplotype from its genotype and matching the second haplotype with haplotypes in the library. In the next step both pedigree and population methods are used to locate matching haplotypes.

PHASEBOOK (Druet *et al.,* 2010) relies on both linkage and linkage disequilibrium for imputation. It uses family information to reconstruct haplotypes of animals with genotyped parents. When parents are not genotyped, haplotype reconstruction and imputation is obtained from a direct acyclic graph created by BEAGLE.

Results

Autosomal imputation

Percentage of correctly, incorrectly and not imputed genotypes was used as a measure of accuracy of imputation. It has to be kept in mind that random prediction of missing genotypes will not yield 0% or 50% of correct genotypes as it is sometimes assumed. If missing genotypes were set to the most frequent genotype, on average 68% of genotypes would be correct. If alleles of missing genotypes were sampled based on allele frequency, 61% of genotypes would be correct.

As shown in the Tables 2 and 3, FImpute was the most accurate software; it imputed 95.0% and 96.8% genotypes correctly with BSW and HOL, respectively. AlphaImpute had the lowest proportion of correctly imputed genotypes but it had the highest percentage of missing and the lowest percentage of incorrectly imputed SNPs. All programs imputed Holstein genotypes more accurately than Brown Swiss genotypes, which can be explained by higher percentage of animals with both parents genotyped (43.6% versus 12.9%) and larger number of genotyped animals in HOL.

Table 2. Average percentage of correctly (correct), incorrectly (incorrect) and not imputed (missing) genotypes of SNPs on autosomes in the Brown Swiss data set.

Software	Average percentage		
	Correct	Incorrect	Missing
AlphaImpute	92.2	3.1	4.7
BEAGLE	94.3	5.7	0.0
FImpute	95.0	4.4	0.6
Findhap	93.9	6.1	0.0
PHASEBOOK	93.2	6.8	0.0

BEAGLE was slightly better than findhap with BSW, but in HOL the order was reversed. This is probably because in BSW there is not as much family information and imputation is mainly based on population imputation where BEAGLE excels. PHASEBOOK was the second least accurate programs with BSW but was second best with HOL. It was likely due to the use of a too simplified DAG for the population step of imputation. The DAG was created just from one iteration, which is probably not enough for small reference populations.

Table 3. Average percentage of correctly (correct), incorrectly (incorrect) and not imputed (missing) genotypes of SNPs on autosomes in the Holstein data set.

Software	Average percentage		
	Correct	Incorrect	Missing
BEAGLE	95.2	4.8	0.0
FImpute	96.8	2.7	0.5
Findhap	95.9	4.1	0.0
PHASEBOOK	96.3	3.7	0.0

X chromosome imputation

Bulls carry only one copy of the X chromosome, inherited from their dams. This has to be taken into account during family imputation, where it is commonly assumed that animal inherits one maternal and one paternal haplotype from the same chromosome. This is true in females and in males for autosomal chromosomes and PAR but not for X-specific part of the chromosome X.

Currently, only FImpute and findhap consider differences between genotypes on autosomes and X-specific part of the X chromosome. Both programs require gender info and identification of the X-specific region.

Table 4. Average percentage of correctly (correct), incorrectly (incorrect) and not imputed genotypes on X-specific part of chromosome X in the Brown Swiss data set.

Software	Average percentage		
	Correct	Incorrect	Missing
AlphaImpute	35.2	2.0	62.8
BEAGLE	95.5	4.5	0.0
FImpute	95.1	3.3	1.7
Findhap	95.0	2.9	2.1
PHASEBOOK	69.0	31.0	0.0

As given in Tables 4 and 5, these two programs did not have any problem imputing this part of the genome. On the other hand, the two remaining programs (AlphaImpute and PHASEBOOK) that used family imputation failed to impute this region accurately.

Both data sets had more male than female genotypes (BSW – 186 males, 23 females, HOL – 12,199 males, 7,801 females).

Table 5. Average percentage of correctly (correct), incorrectly (incorrect) and not imputed genotypes (missing) on X-specific part of chromosome X in the Holstein data set.

Software	Average percentage		
	Correct	Incorrect	Missing
BEAGLE	94.2	5.9	0.0
FImpute	95.2	2.8	2.0
Findhap	97.3	2.5	0.2
PHASEBOOK	75.1	24.9	0.0

AlphaImpute and PHASEBOOK predicted more than 90% genotypes of cows correctly while only 28% and 66% of male genotypes were imputed correctly with these programs, respectively. Some countries currently do not use SNPs from this segment. The X chromosome is the second longest chromosome and it is not yet clear whether or not considering this chromosome in the genomic evaluation has a significant impact on the results.

The PAR is a short (4.5 Mb) terminal segment of both chromosomes X and Y. Therefore SNPs positioned in the PAR region can have up to three different genotypes in both males and females.

Imputation accuracy was lower on PAR for all programs. This was likely because of the short length of the segment. There are only 4 SNPs from the 3k panel located on PAR and therefore there is only a few low density haplotypes that can be used to match haplotypes from the 50k panel.

Table 6. Average percentage of correctly (correct), incorrectly (incorrect) and not imputed genotypes (missing) on PAR in the Brown Swiss data set.

Software	Average percentage		
	Correct	Incorrect	Missing
AlphaImpute	52.8	4.4	42.8
BEAGLE	65.2	34.8	0.0
FImpute	23.5	10.5	66.0
findhap	79.3	20.7	0.0
PHASEBOOK	75.0	25.0	0.0

The program findhap had the highest accuracy, it imputed correctly 79% of genotypes in this region but it was more than 10 points below the percentage obtained on autosomes. Bearing in mind that 68% accuracy can obtained by random sampling 79% is quite low. It might be better to either not impute the 93 SNPs that are on 50k panel but are not typed on the 3k panel or simply not use them for further analysis.

Table 7. Average percentage of correctly (correct), incorrectly (incorrect) and not imputed (missing) genotypes on PAR in the Holstein data set.

Software	Average percentage		
	Correct	Incorrect	Missing
BEAGLE	64.2	35.8	0.0
FImpute	14.9	4.8	80.3
Findhap	79.0	21.0	0.0
PHASEBOOK	78.5	21.5	0.0

Effect of family information on imputation

In all four programs that used family imputation (AlphaImpute, FImpute, findhap and PHASEBOOK), the proportion of correctly imputed genotypes increased with increasing relatedness between genotyped ancestors and animals (Tables 9 and 12).

Brown Swiss data

PHASEBOOK was the most accurate software for animals with both parents genotyped with 50k panel. AlphaImpute provided the most accurate imputation (96.7%) for animals with sire genotyped with 50k panel and dam genotyped with 3k panel. FImpute was the best programs for animals with only maternal grandsire (**MGS**) genotyped. BEAGLE was the least accurate when family information was available, but was the most accurate when there was limited amount of family information

Holstein data

Both Fimpute and PHASEBOOK imputed genotypes of animals with both parents genotyped with 50k panel with very high accuracy (99.1%). FImpute also provided the best results for animals with sire genotyped and with either dam or MGS genotyped. Findhap was the best program for animals with sire genotyped and dam unknown. Considerable

strength of BEAGLE is its capability to impute genotypes relatively accurately using very limited amount of information. BEAGLE was able to impute over 92% genotypes accurately for animals with only MGS genotyped.

Computational demands

All programs were run on a server with 4 AMD Opteron 2.5 GHz CPUs with 12 cores each and total of 256 GB of RAM. Maximum of 31 CPUs were allowed to be used by each imputation program. AlphaImpute, BEAGLE and PHASEBOOK run on a per chromosome basis. They can be easily parallelized (each chromosome can be imputed by a different CPU). The additional advantage of this approach is that workload can be split across different servers. Both FImpute and findhap are parallel programs and impute all chromosomes simultaneously; the advantage is that the input file does not have to be split. FImpute uses 1 CPU per chromosome, therefore 31 CPUs is the best set up for imputation of bovine genotypes (29 autosomes + X-specific + PAR) with FImpute.

Computing time required to impute 209 Brown Swiss and 20,000 Holstein genotypes for 38,889 masked SNPs is summarized in Table 8. FImpute was the fastest programs with both BSW and HOL. BEAGLE was by far the slowest program (18 and 293 times slower than Fimpute with BSW and HOL, respectively).

Table 8. Time required for imputation.

Software	Data set	
	Brown Swiss	**Holstein**
AlphaImpute	2m 59s	-
BEAGLE	12m 9s	16d 18h
FImpute	40s	1h 22min
findhap	3m 12s	6h 55m
PHASEBOOK	2m 23s	1d 7h

31 processors (2.5GHz AMD Opteron)

Size of genotype file

Each of the programs has its own file format. FImpute and findhap have the most efficient format. Holstein genotype file (105,182 genotypes) for these two programs had size of 4.2 GB, while file for AlphaImpute and PHASEBOOK were 8.6 GB and 15.7 GB, respectively. BEAGLE did not use genotypes of all animals but only needed reference population of 7,228 and the study sample of 20,000 animals, so despite its inefficient format (similar to PHASEBOOK) the total size of the two genotype files was "just" 4.5 GB.

Combining results from two programs

AlphaImpute and FImpute do not impute genotypes when there is not enough information for accurate imputation. Because all five programs use different algorithm for phasing and imputation it was tested whether combining results from these two programs with one of the three remaining programs would result in higher accuracy. For this test, only the family imputation part of Fimpute was used. As shown in Tables 13 and 14, combining results from two programs resulted in higher percentage of correctly imputed SNPs, indicating that some of the SNPs that were not imputed by the two programs were correctly imputed by the other programs. The blending is the most important for animals without family information, because this group had the lowest correct rate of imputation (bellow 90%) with either AlphaImpute or FImpute. The combination of programs that yielded the highest proportion of correctly imputed SNPs was FImpute + BEAGLE.

Additional features

- AlphaImpute, FImpute, findhap can impute genotypes of untyped family member in the pedigree.
- AlphaImpute and BEAGLE – provides posterior probabilities of genotype calls, the second most probable genotype can be used to correct genotyping error
- BEAGLE – estimates allelic R^2 for each SNP (squared correlation of the allele dosage with the highest posterior probability and the true allele dosage), which can be used for elimination of SNPs with low imputation accuracy

- Fimpute and findhap - correct Mendelian inconsistencies. Correct pedigree is crucial with Fimpute otherwise genotypes of animal with pedigree conflict can be significantly modified to fit the supplied pedigree structure
- AlphaImpute – estimates percentage of correctly, incorrectly and not imputed SNPs for each animal when a file with "true" genotypes is supplied
- AlphaImpute, PHASEBOOK – allow newly genotyped animals to be imputed using results from previous imputation (update option)
- BEAGLE – newly imputed animals can be imputed independently from the rest of the genotyped animals without loss of accuracy. Therefore there is no need to re-impute already imputed genotypes every time a new genotype is received
- BEAGLE and findhap – are capable to impute 50k genotypes to HD genotypes (>700k)

Conclusions

All five programs performed well in imputing genotypes from 3k to 50k density, with average accuracy exceeding 90%. Each of them had certain strengths and weaknesses and using a combination of 2 programs improved imputation results.

Acknowledgement

The authors thank Brian Browning, Tom Druet, John Hickey, Paul VanRaden and Mehdi Sargolzaei for providing the latest version of their software and for comments and suggestions regarding implementation of their programs.

References

Browning, S.R. & Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet. 81,* 1084-1097.

Druet, T., Schrooten, C. & de Roos, A.P.W. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J. Dairy. Sci. 93,* 5443-5454.

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N. & van der Werf, J.H.J. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol. 43,* 12.

Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. FImpute. 2011. An efficient imputation algorithm for dairy cattle populations. *J. Anim. Sci. 89(E-Suppl. 1)/J. Dairy Sci. 94(E-Suppl. 1),* 421 *(abstr. 333).*

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol. 43,* 10.

Table 9. Percentage of correctly imputed autosomal genotypes in Brown Swiss data set.

Genotype of		Number	Software				
sire	**dam**	**of animals**	**AlphaImpute**	**BEAGLE**	**FImpute**	**findhap**	**PHASEBOOK**
50k	**50k**	18	98.5	94.6	98.5	96.1	98.9
50k	**3k**	9	96.7	94.5	95.8	94.2	96.2
50k	**0k, MGS 50k**[2]	158	95.0	94.6	95.4	94.2	93.3
0k[1]	**0k, MGS 50k**[2]	12	73.7	92.3	88.7	89.1	84.3

[1]sire not genotyped, [2]dam not genotyped

Table 10. Percentage of incorrectly imputed (missing in parentheses) autosomal genotypes in Brown Swiss data set.

Genotype of		Number	Software				
sire	**dam**	**of animals**	**AlphaImpute**	**BEAGLE**	**FImpute**	**findhap**	**PHASEBOOK**
50k	**50k**	18	0.9 (0.6)	5.4 (0)	1.2 (0.3)	3.9 (0)	1.1 (0)
50k	**3k**	9	1.9 (1.4)	5.5 (0)	2.2 (1.9)	5.8 (0)	3.8 (0)
50k	**0k, MGS 50k**[2]	158	3.1 (1.9)	5.4 (0)	4.1 (0.5)	5.8 (0)	6.7 (0)
0k[1]	**0k, MGS 50k**[2]	12	6.1 (20.2)	7.7 (0)	9.3 (2.0)	10.9 (0)	15.7 (0)

[1]sire not genotyped, [2]dam not genotyped

Table 11. Percentage of correctly imputed autosomal genotypes in Holstein data set.

Genotype of		Number	Software			
Sire	**dam**	**of animals**	**BEAGLE**	**FImpute**	**findhap**	**PHASEBOOK**
50k	**50k**	4,545	95.3	99.1	96.8	99.1
50k	**3k**	4,151	95.3	98.0	96.0	97.2
50k	**0k, MGS 50k**[2]	5,910	95.2	96.3	95.6	95.2
50k	**unknown**	4,728	95.3	95.8	95.8	95.1
0k[1]	**0k, MGS 50k**[2]	114	92.4	89.7	91.4	86.9

[1]sire not genotyped, [2]dam not genotyped

Table 12. Percentage of incorrectly imputed (missing in parentheses) autosomal genotypes in Holstein data set.

Genotype of		Number of	Software			
Sire	**dam**	**animals**	**BEAGLE**	**FImpute**	**findhap**	**PHASEBOOK**
50k	**50k**	4,545	4.7 (0)	0.8 (0.1)	3.2 (0)	0.9 (0)
50k	**3k**	4,151	4.7 (0)	1.8 (0.2)	4.1 (0)	2.8 (0)
50k	**0k, MGS 50k**[2]	5,910	4.8 (0)	3.6 (0.2)	4.5 (0)	4.8 (0)
50k	**unknown**	4,728	4.7 (0)	4.1 (0.1)	4.2 (0)	7.2 (0)
0k[1]	**0k, MGS 50k**[2]	114	7.6 (0)	9.3 (1.1)	8.6 (0)	13.1 (0)

[1]sire not genotyped, [2]dam not genotyped

Table 13. Percentage of correctly imputed autosomal genotypes in Brown Swiss data set when results from two imputation programs were blended.

Genotype of Sire	Genotype of dam	AlphaImpute + BEAGLE	AlphaImpute + findhap	FImpute + BEAGLE	FImpute + findhap	FImpute + PHASEBOOK
50k	**50k**	98.9	98.8	98.8	98.8	98.8
50k	**3k**	97.6	97.6	96.5	96.3	96.7
50k	**0k, MGS 50k**[2]	96.4	96.4	95.7	95.1	95.0
0k[1]	**0k, MGS 50k**[2]	90.8	89.8	92.4	89.2	84.4
All animals		96.1	96.0	95.6	95.0	94.5

[1]sire not genotyped, [2]dam not genotyped

Table 14. Percentage of correctly imputed genotypes in Holstein data set results from two imputation programs were blended.

Genotype of Sire	Genotype of dam	Number of animals	FImpute + BEAGLE	FImpute + findhap	FImpute + PHASEBOOK
50k	**50k**	4,545	99.2	99.2	99.2
50k	**3k**	4,151	98.3	98.3	98.3
50k	**0k, MGS 50k**[2]	5,910	97.3	97.2	97.3
50k	**unknown**	4,728	95.3	95.8	95.1
0k[1]	**0k, MGS 50k**[2]	114	93.2	92.0	89.6
All animals		20,000	97.4	97.4	97.3

[1]sire not genotyped, [2]dam not genotyped