Imputation accuracy from low to high density using within and across breed reference populations in Holstein, Guernsey and Ayrshire cattle

Larmer, S.[1], Sargolzaei, M.[12], Ventura, R.[13] and Schenkel, F.[1]

[1]University of Guelph, Centre for Genetic Improvement of Livestock, Guelph, ON, Canada
[2]L'Alliance Boviteq, Saint-Hyacinthe, QC, Canada
[3]Beef Improvement Opportunities, Guelph, ON, Canada

Introduction

Genomic selection in dairy cattle uses information from dense marker panels to estimate effects of Quantitative Trait Loci (QTL) that are at or are in Linkage Disequilibrium (LD) with markers on these dense panels (Goddard and Hayes, 2007). These dense marker panels, however, are expensive and can be a major limitation to the number of genotyped animals. The advent of imputation has significantly aided this problem, making it possible for many animals to genotyped with lower density marker panels and imputed to higher density for genomic selection. Currently in Canada, imputation takes place regularly from approximately 3,000 Single Nucleotide Polymorphisms (SNPs) to a panel of nearly 50,000 SNPs. This is done by using the FImpute program V1 (Sargolzaei, 2010), which uses a family-based algorithm, followed by Beagle software, which uses population-based method. Similar strategies are carried out in the United States and around the world using a variety of methods using both family-based imputation methods as well as those relying on population-based imputation (Johnston and Kistemaker, 2011). For a population with a large reference set and high degree of relatedness, imputation from ~3k to ~50k has been shown to very accurate using both family-based and population-based methods with accuracy exceeding 0.90 for population-based methods, such as IMPUTE (Druet et al. 2010, Weigel et al. 2010, Nothnagel et al. 2008), BEAGLE (Druet et al. 2010, Nothnagel et al. 2008, Calus et al. 2011), DAGPHASE (Zhang et al. 2010) and fastPhase (Weigel et al. 2010, Nothnagel et al. 2008, Calus et al. 2011) as well as family based methods CHROMIBD(Zhang et al. 2010), FImpute (Sargolzaei et al. 2010, Johnston and Kistemaker, 2011), Chromophase (Daetwyler et al. 2011) and findhap (Johnston and Kistemaker, 2011). It was found that when family information wasavailable, family-based methods do have an advantage over population-based methods (Zhang et al. 2010) More notably, it was found that a combination of family-based and population-based imputation could yield very high accuracy when combined (Meuwissen and Goodard, 2010; Johnston and Kistemaker 2011).

It has been found that for a limited training population size, such as that of Ayrshires and Guernseys, a denser SNP panel is required to increase the accuracy of genomic selection (Hayes, 2009). There are currently a number of these animals genotyped with a panel encompassing approximately 777,000 SNP. The cost of this panel is still high, and creates an even greater problem in having a suitably large training population for generating accurate genomic breeding values. Imputation from lower density panels to the 777k panel would significantly decrease genotyping costs. Van Raden et al. (2011) found that 500,000 markers could be accurately imputed (>95%) using a combination of family and population based methods. This was in a Holstein population with strong family information available. In a population with limited family information, using a family-based method may not yield a significant increase in accuracy of imputation. Calus et al. (2011) found that Beagle outperformed a family-based multivariate mixed model method for imputation when a high density marker panel was simulated.

The goal of this study was to investigate the accuracy of imputation from 50k to 777k using real Ayrshire and Guernsey data. Both population-based and combined family and population-based

methods were tested. The effects of a larger multi-breed reference population and direct genotyped ancestors were examined. This study also explored imputation from an even lower density panel (6k) to the high density panel, using 1 or 2-step approaches. Beagle and FImpute version 2 programs were used for carrying out the imputations.

Materials and Methods

Data

High Density (777k) genotypes from Holstein, Ayrshire and Guernsey (n=1115, 351 and 60, respectively) were used (Source: CDN and AIPL). The Holstein data consisted of both males and females whereas the Ayrshire and Guernsey data set consisted of only males with Canadian official proofs. The data was examined to determine if alleles with low minor allele frequency had an effect on imputation accuracy. No significant effect on imputation accuracy was found, so the entire set of markers was used for imputation purposes. Markers mapped to the X chromosome as well as the pseudo-autosomal region were excluded. This left a set of 735,293 SNPs to be examined.

To perform imputation, animals were divided up into 2 distinct groups, named reference and imputation animals. This sorting was done by birth year in an effort to capture as many sires of imputation animals in the reference population as possible, while still maintaining a practical scenario that mimics how imputation is carried out for routine genomic evaluations. In Ayrshires and Guernseys animals born in 2000 or after were included in the imputation set. This gave reference populations of 211 and 41 and imputation groups of 140 and 19 for Ayrshire and Guernsey respectively. Holsteins born in 2004 or later were included in the imputation group, creating sets of 892 and 223, respectively, for reference and imputation (birth date was unknown for 504 Holstein and they were included in reference group).

Mimicking Low-Density Marker Panels

Imputation was carried out from both the 6k and 50k Illumina SNP panels. The markers on both of these SNP panels are almost entirely included on the HD SNP panel. This being the case, the HD panel was then filtered to erase all SNPs not contained in whichever low density genotyping platform we were investigating. This was performed for all animals in the imputation set to mimic not having been genotyped on the high density platform. This left 39946 and 6556 SNP to be considered on the 50k and 6k panels, respectively. It should also be noted that all 6556 SNP in the 6k chip are present in the 50k chip, as imputation from 6k was also carried out as a two step procedure (from 6k to 50k and then from 50k to HD) in an effort to improve overall imputation accuracy.

Imputation Scenarios

A number of imputation scenarios were carried out in this study. The primary goal was to evaluate the effect of combined family and population vs. population based imputation both within and across breeds using the HD SNP panel. This was done by imputing all breeds individually as well as combining all populations for imputation both in Beagle (population-based) and in FImpute (combined family and population-based). We also wanted to determine accuracy of imputation when comparing population-based imputation in both Beagle and FImpute. This was done in FImpute by removing all pedigree information before imputation. Consistency of haplotypes was also examined across breeds. This was done by using one breed (Holstein) as the sole reference population to impute animals in the

other breeds, using FImpute. All the above scenarios were only carried out for the 50k SNP panel. Imputation accuracy for the 6k panel was determined using a two step procedure, as it has been found in other preliminary studies to increase imputation accuracy, rather than imputing directly from the 6k panel to high density. This was carried out both within and across breeds using FImpute with pedigree information included.

Imputation of Missing Markers – FImpute

Family Based Imputation

The family based imputation algorithm used by FImpute was described by Sargolzaei (2010). Family-based imputation with FImpute is a 3 step procedure. In the first step, parent or progeny information is used to fill in missing genotypes where it can be done with a high degree of certainty. The second step is the reconstruction of marker haplotypes, as described in detail by Sargolzaei et al. (2008). Haplotypes are reconstructed iteratively. This is done by considering sire information first to determine phase at a pair of marker loci when the sire is homozygous at that loci. When phase is still unknown, the nearest partially informative heterozygous marker is used along with linkage information between those 2 markers from progeny to infer haplotype probabilities further. Partially informative markers flanking the marker in question are then found and once again, linkage information is used to estimate haplotype probabilities in progeny. This iteration is repeated until the sum of squares of haplotype probabilities is sufficiently small. Haplotypes of progeny are then matched to haplotypes of parents or immediate ancestors and untyped loci are then filled in.

Population Based Imputation

When there is no family information present for an individual or the pedigree file is excluded, FImpute uses population imputation. Population imputation is carried out by FImpute using overlapping windows to reconstruct haplotypes and impute at the same time. FImpute, unlike most population imputation software, assumes all animals are related to some degree and uses these overlapping windows to find segments of haplotype that are consistent between individuals having come from a common ancestor. The windows are large at first to find segments of haplotype that come from more recent ancestors. The window walks along each chromosome finding large segments consistent with reference animals, overlapping by 75% of the window size each step (This overlap can be modified to optimize accuracy and computing time). After each chromosome has been completed with large windows, the same process is repeated numerous times with smaller and smaller windows to capture consistent haplotypes from less recent ancestors. When multiple haplotypes are found at a certain window size, haplotype frequency in the reference population and hit number are used to determine the most likely haplotype, and fills that haplotype into the imputed animal's genotype. For the sake of computational efficiency, FImpute does not calculate genotype posterior probabilities. However, one may avoid very short windows to ensure certain level of accuracy for imputed genotypes. Default parameters were used for all imputation scenarios.

Imputation of Missing Markers - BEAGLE

The BEAGLE imputation method was described in full detail in Browning and Browning (2007). BEAGLE uses a "localized haplotype-cluster" model to perform imputation on missing genotype markers. This algorithm uses only local haplotype data in order to capture markers in tight LD with one another. It uses an underlying Hidden Markov Monte-Carlo (HMM) process to determine transition probabilities from one "node" to another based on overall haplotype counts. In this case, a node is a collection of haplotypes that have the same allele at a certain locus. So, at a given node, it determines the probability

of the following allele, or the probability of moving to a certain child node. The sums of all of these probabilities for the entire pathway from the root node to the terminal node give the probabilities of each unique haplotype. BEAGLE first uses a phasing algorithm to determine haplotype phase for each individual. This is done by first constructing the local haplotype clusters, and then sampling a number of haplotypes for each individual from the HMM. The sampled haplotypes are then used to reconstruct the local haplotype cluster. This is repeated over 10 iterations to achieve a high level of phasing accuracy while maintaining computational efficiency. Default parameters were used in all scenarios.

Determination of Imputation Accuracy

For this study, imputation accuracy was measured using a ratio of correct call to overall call rate. This means any markers that were left as missing were not included in the calculation of accuracy. This meant for BEAGLE that accuracy was always the same as correct call rate, as all markers are filled in the BEAGLE imputation algorithm. If certain markers could not be imputed by FImpute, they are still considered missing and do not count towards our measures of accuracy. Optionally, the remaining missing genotypes can be filled in based on allele frequency to achieve 100% call rate, but this was not tried. It should also be noted that, with Beagle, the most probable call at any locus was considered in all cases to be the genotype present. That is to say, no threshold was set on allele probability for it to be included in predictions of imputation accuracy. Markers were filtered originally from the HD genotype panel, so to calculate correct call rate imputed genotypes were compared to the complete HD genotype and calculated the amount of correct calls, incorrect calls and those called as missing. Correct calls are those in which the call after imputation exactly matches that of the original HD genotype. This is a slightly downward biased measure of imputation accuracy when compared to the allelic R^2 as described by Browning and Browning (2009), which takes into account the correlation between heterozygous and homozygous calls given that if one allele at a locus is correct there will still be valuable information available at that locus for further studies.

Computation Time

The amount of time used imputing in each scenario with each algorithm was also examined. This was done on a per chromosome basis as, due to computational requirements, a different number of chromosomes were run in parallel in each scenario and we could only measure overall computing time. Measuring computational efficiency is important when determining the efficacy of each algorithm especially when it is to be applied to larger and larger data sets for routine genomic evaluations.

Results

Imputation Accuracy

This study first looked at the difference between population vs. combined family and population based imputation when imputing from the 50k SNP chip to the HD (777k) chip. The results of this imputation are presented in Table 1. There was very little difference seen in terms of imputation accuracy between these 2 methods when FImpute was used with or without pedigree information. Accuracy as well as correct call rate was within one percent in all scenarios when comparing these 2 methods.

Secondly, the difference between the population-based imputation performed by BEAGLE and imputation performed by FImpute when pedigree information was omitted (population imputation only)

was examined. These results are presented in Table 2. There was very little difference in imputation accuracy or correct call rate between the 2 software when population-based imputation was carried out on the Holstein population. There was a difference, however, when imputation accuracy between these 2 software for Guernsey and Ayrshire was examined, with FImpute out-performing BEAGLE in both cases. Imputation accuracy was higher by nearly 1% for the Ayrshire population and nearly 2% for Guernsey. As reference population size decreases, it seems that FImpute performs better.

Next, using all breeds as a reference population for imputation using both BEAGLE and FImpute was considered. These results are presented in Table 3, along with imputation accuracy when reference populations are comprised within breed only. For both breeds with smaller reference populations (Ayrshire and Guernsey) there was an increase in imputation accuracy as well as correct call rate when information from other breeds was included. This difference was greater when imputation was carried out with BEAGLE. The largest difference was seen for BEAGLE in the Guernsey population when information from Holstein and Ayrshire were included. Including other breeds led to over a 1% increase in imputation accuracy. FImpute also gained from using information from other breeds, however the difference was smaller. The Ayrshire data set gained ~0.3% in accuracy from adding Holsteins and Guernseys when family-based imputation was carried out and ~0.1% without pedigree information. Guernsey data gained slightly less with increases in accuracy of 0.2% and 0.1% with and without pedigree information respectively.

Imputation accuracy was also determined when imputing from 6k to HD. These results are presented in Table 4. Imputing in 2 steps yielded a higher accuracy than when imputation was carried out directly from 6k to the high density platform. This was seen both within and across breeds. Adding information from other breeds had a detrimental effect on imputation accuracy in both the 1 step and 2 step imputation methods from 6k to HD. Imputation accuracy was, however, generally substantially lower when imputing from 6k to HD platform than from 50k to HD, especially in breeds with smaller reference populations.

To determine the amount of information that could be gained from Holsteins for the smaller population-sized breeds, imputation accuracy was also measured when using a reference population comprised entirely of Holstein animals. Then accuracy was also determined when imputation was carried out based solely on minor allele frequency to determine the difference between random imputation and imputation using the Holstein reference population. These results are presented in Table 5. Increases in correct call rates of 15% and 11% were seen for Ayrshires and Guernseys respectively.

Computing Time

The results for overall computing time as well as computing time per chromosome for all imputation scenarios are presented in Table 6. The largest differences in computing were seen between BEAGLE and FImpute, with FImpute being 9-13 times as computationally efficient in all comparable scenarios. There was also a difference in computing time within the FImpute program when pedigree information was included or omitted. A slight increase in computational efficiency was seen when pedigree information was excluded. When imputation from 6k was carried out, there was an increase in computational efficiency when the 2 step procedure was carried out due to the algorithm not having as many possible haplotype blocks to consider in each imputation step.

Conclusions

Highly accurate imputation was obtained from 50k to HD within the small population-sized Ayrshire and Guernsey breeds, using either population information only or family and population information.

The use of a multi-breed reference population only slightly increased the imputation accuracy from 50k to HD for Ayrshire and Guernsey breeds.

Imputation from 6k to HD was substantially less reliable for Ayrshire and Guernsey breeds with the use of a multi-breed reference population leading to even lower accuracy than within each breed. This is likely due to the fact that the 6k panel is not dense enough to capture sufficient population linkage disequilibrium across breeds.

FImpute yielded similar or higher imputation accuracy than BEAGLE, but was much more computationally efficient, being 9-13 times faster.

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

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics, 84*(2), 210-223.

Calus, M. P. L., Veerkamp, R. F., Mulder, H. A. (2011) Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. *Journal of Animal Science,* 89(7), 2042-2049.

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., Goddard, M. E. (2011) Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics, 189(1),* 317-327.

Druet, T., Schrooten, C., & de Roos, A. P. W. (2010). Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *Journal of Dairy Science, 93*(11), 5443-5454.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science, 92*(2), 433-443.

Johnston, J., Kistemaker, G., Sullivan, P.G. (2011) Comparison of different imputation methods. *Interbull Open Meeting*. *Stavanger, Norway.*

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. and Franke, A. (2009) A comprehensive evaluation of SNP genotype imputation. *Human Genetics, 125 (2).* 163-171.

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., & Schaeffer, L. R. (2008). Extent of linkage disequilibrium in holstein cattle in north america. *Journal of Dairy Science, 91*(5), 2106-2117.

Sargolzaei, M., Chenais, J.P., Schenkel, F.S. (2010) Accuracy of a family-based genotype imputation algorithm. *GEB Open Industry Session. Saint-Hyacynthe, Quebec, Canada.*

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., & Weigel, K. A. (2011). Genomic evaluations with many more genotypes. *Genetics, Selection, Evolution, 43*(1), 10-20.

Weigel, K. A., Van Tassell, C. P., O'Connell, J. R., VanRaden, P. M., & Wiggans, G. R. (2010). Prediction of unobserved single nucleotide polymorphism genotypes of jersey cattle using reference panels and population-based imputation algorithms. *Journal of Dairy Science, 93*(5), 2229-2238.

Zhang, Z., & Druet, T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science, 93*(11), 5487-5494.

Table 1 – Imputation accuracy from 50k to HD for combined family and population *vs* population based (no pedigree) imputation for 3 breeds using FImpute.

		Imputed	Correct Call	Incorrect Call	Accuracy
Guernsey	Population only	99.920	97.178	2.742	0.973
	Family+population	99.919	97.179	2.740	0.973
Ayrshire	Population only	99.993	97.979	2.015	0.980
	Family+population	99.985	97.997	1.989	0.980
Holstein	Population only	100.000	99.235	0.765	0.992
	Family+population	100.000	99.234	0.764	0.992

Table 2 – Imputation accuracy from 50k to HD for population based (no pedigree) imputation algorithms using FImpute and BEAGLE software for 3 breeds.

		Imputed	Correct Call	Incorrect Call	Accuracy
Guernsey	Beagle	100.000	95.367	4.633	0.954
	FImpute	99.920	97.178	2.742	0.973
Ayrshire	Beagle	100.000	97.158	2.842	0.972
	FImpute	99.993	97.979	2.015	0.980
Holstein	Beagle	100.000	99.296	0.704	0.993
	FImpute	100.000	99.235	0.765	0.992

Table 3 – Imputation accuracy from 50k to HD for FImpute and BEAGLE using single and multi-breed reference populations for imputation.

		Reference	Imputed	Correct Call	Incorrect Call	Accuracy
Guernsey	Beagle	GU	100.000	95.367	4.633	0.954
		GU+AY+HO	100.000	96.671	3.329	0.967
	FImpute	GU	99.919	97.179	2.740	0.973
		GU+AY+HO	100.000	97.423	2.577	0.974
Ayrshire	Beagle	AY	100.000	97.158	2.842	0.972
		GU+AY+HO	100.000	97.775	2.225	0.978
	FImpute	AY	99.985	97.997	1.989	0.980
		GU+AY+HO	99.997	98.231	1.765	0.982
Holstein	Beagle	HO	100.000	99.296	0.704	0.993
		GU+AY+HO	100.000	99.286	0.714	0.993
	FImpute	HO	100.000	99.234	0.764	0.992
		GU+AY+HO	100.000	99.225	0.774	0.992

Table 4 – Imputation Accuracy from 6k to HD for 3 breeds using one and two step imputation procedures as well as single and multi-breed reference populations.

		Reference	Imputed	Correct Call	Incorrect Call	Accuracy
Guernsey	One-step	GU	99.989	91.891	8.096	0.919
		GU+AY+HO	100.000	88.920	11.080	0.889
	Two-step	GU	99.987	93.215	6.772	0.932
		GU+AY+HO	100.000	91.955	8.044	0.920
Ayrshire	One-step	AY	99.963	94.809	5.152	0.948
		GU+AY+HO	99.985	93.957	6.029	0.940
	Two-step	AY	99.967	94.711	5.256	0.947
		GU+AY+HO	99.979	94.417	5.562	0.944
Holstein	One-step	HO	100.000	97.110	2.888	0.971
		GU+AY+HO	100.000	96.923	3.075	0.969
	Two-step	HO	99.999	97.369	2.628	0.974
		GU+AY+HO	100.000	97.291	2.708	0.973

Table 5 – Imputation accuracy from randomly imputing within breed or using Holstein as the reference population.

		Imputed	Correct Call	Incorrect Call	Accuracy
Guernsey	Random Imputation	N/A	64.886	N/A	N/A
	Holstein Reference	99.991	75.193	24.797	0.752
Ayrshire	Random Imputation	N/A	62.221	N/A	N/A
	Holstein Reference	99.990	77.886	22.106	0.779

Table 6 – Total and per chromosome imputing time[1] for all scenarios.

		Computing Time	# of Jobs	Time/Chr
Guernsey	Family+pop	0:01:21	10	0:00:26
	Pop only	0:01:04 (0:09:25)	10 (10)	0:00:21 (0:03:02)
	All Ref	0:37:46 (36:47:10)	10 (3)	0:12:11 (3:48:20)
	HO Ref	0:22:37	10	0:07:18
	6k one-step	0:01:41	10	0:00:33
	6k two-Step	0:01:15	10	0:00:24
	6k one-step -All Ref	1:03:26	10	0:20:28
	6k two-Step - All Ref	0:38:58	10	0:12:34
Ayrshire	Family+pop	0:06:41	10	0:02:09
	Pop only	0:06:29 (2:38:04)	10 (5)	0:02:05 (0:25:30)
	All Ref	0:37:46 (36:47:10)	10 (3)	0:12:11 (3:48:20)
	HO Ref	0:22:37	10	0:07:18
	6k one-step	0:11:12	10	0:03:37
	6k two-Step	0:06:25	10	0:02:04
	6k one-step -All Ref	1:03:26	10	0:20:28
	6k two-Step - All Ref	0:38:58	10	0:12:34
Holstein	Family+pop	0:26:10	10	0:08:26
	Pop only	0:21:47 (10:26:19)	10 (5)	0:07:02 (1:41:01)
	All Ref	0:37:46 (36:47:10)	10 (3)	0:12:11 (3:48:20)
	6k one-step	0:49:18	10	0:15:54
	6k two-Step	0:29:39	10	0:09:34
	6k one-step -All Ref	1:03:26	10	0:20:28
	6k two-Step - All Ref	0:38:58	10	0:12:34

[1]Time for Beagle is in brackets