Efficient combined family and population imputation in large data sets

Sargolzaei, M.[12], Chesnais, J.[1]and Schenkel, F.[2]
[1]L'AllianceBoviteq, Saint-Hyacinthe, QC, Canada
[2]University of Guelph, Centre for Genetic Improvement of Livestock, Guelph, ON, Canada

Introduction

Most of available imputation programs are computationally demanding when applied to data with large number of reference individuals. In livestock, the size of genomic data sets is increasing rapidly. For example, in North American dairy cattle, there are close to100,000 animals genotyped with the 50k panel, and there are even more animals genotyped with the lower density panel. A common strategy to deal with large data sets is to restrict the reference animals to a smaller group based on, for example, pedigree or genomic relationship. For population imputation, this strategy is suboptimal for some animals and especially for imputing rare alleles since all available information is not taken into account.

Population based programs do not make use of pedigree information directly. Therefore, for pedigreed populations, like most livestock populations, algorithms that combine family and population imputation are preferable. The objective of this study was to assess the accuracy and computing efficiency of a new version of FImpute software (v2.2) on a large Holstein data set in comparison to two well-known population based programs (Beagle and Impute2) and to a new combined family and population software (Pedimpute).

Materials and Methods

The February 2011 Holstein data set was provided by CDN. It was the same data set used in J. Johnston's reports presented in the March 2011 DCBGC meeting. The data set consisted of 105,182 genotyped animals in total. Among these, 65,047 were genotyped with the 50k panel and 40,135 were genotyped with 3k panels. After edition by USDA, the 50k and 3k panels had 42,503 and 2,614 SNP, respectively. In this study chromosomes 1 to 8 were considered. These are the same chromosomes as those considered in J. Johnston's reports. Validation animals were 20,000 animals with 50k born after 2009. The genotypes of validation animals were reduced to 3k (2,614), 6k (6,701) and 8k (7,825) by masking SNPs that do not belong to these panels. Pedigree information of validation animals was reduced to mimic pedigree structure of animals with low density genotypes (for details see report by J. Johnston and G. Kistemaker, March 2011, DCBGC).

Imputation from low density to high density was carried out with FImpute v2.2, Beagle V3.3.2, Impute2 V2.2.2 and Pedimpute. All programs were run with default parameters, except for Impute2 software, where Ne was set to 80. The FImpute algorithm relies on the fact that all individuals are related to each other at different degrees. First it uses pedigree information and then it partitions each chromosome into different windows sizes which overlap with each other. Search for most likely haplotypes starts in the largest windows and continues toward the smallest windows capturing close to far relationships. Methods implemented in Beagle and Impute2 are based on MCMC method and therefore computationally extensive, while the Pedimpute method is based on a fast iterative method. Pedimpute starts with family imputation first and then moves to population imputation. For population imputation, Pedimpute uses a

similar idea as that applied in previous versions of FImpute, i.e. it starts with a reference set of haplotypes reconstructed from family information.

A two-step scenario (combined family+population imputation) was also implemented for Beagle and Impute2. In the two-step, family and population imputations were carried out separately by FImpute and Beagle/Impute2, respectively. Missing genotypes after family imputation were filled in by imputed genotypes from population imputation.

Three sets of reference groups were selected for population imputation:

1) All sires genotyped with 50k, n = 5,133 sires
2) All sires and dams genotyped with 50k, n = 10,337
3) All animals genotyped with 50k, n= 45,047

Accuracy was assessed by calculating concordance rate between the imputed and observed genotypes for the masked SNPs. Missing SNPs in 50k genotypes were ignored when calculating concordance.

Results for FImpute in Tables 1 to 5 and Figure 1 and 2 are based on version 2.2, which has been shown in previous comparisons to be substantially more accurate than version 1.1. In Figure 3, however, FImpute v2.2 is compared with a combination of version 1.1 of FImpute and Beagle, which is what CDN currently uses for imputation.

Results:

Table 1 - Overall concordance rate and CPU time for **population imputation** only (no pedigree information)

	FImpute 5,133*	FImpute 10,337*	FImpute 45,047*	Beagle 5,133*	Impute2 5,133*	Impute2 10,337*	Pedimpute**
3k	96.39	97.07	97.36	95.90	96.28	96.48	-
6k	98.92	99.14	99.26	98.85	98.96	99.11	-
8k	99.08	99.26	99.37	98.96	99.08	99.11	-
CPU time 3k	9min	14min	1h:15	25 days	47h	84h	-
CPU time 6k	12min	18min	1h:25	11 days	45h	74h	-
CPU time 8k	12min	20min	1h:30	11 days	47h	80h	-

*Number of reference animals for population imputation
**Pedimpute does not allow for population imputation only.
Overall missing call rate after imputation was zero for FImpute, Beagle and Impute2.

Table 2 - Overall concordance rate and CPU time for **combined family and population imputation**

	FImpute 5,133*	FImpute 10,337*	FImpute 45,047*	FImpute** + Beagle 5,133*	FImpute** + Impute2 5,133*	FImpute** + Impute2 10,337*	Pedimpute
3k	98.01	98.05	98.14	97.76	97.91	97.93	96.52
6k	99.36	99.37	99.41	99.30	99.34	99.33	97.97
8k	99.44	99.45	99.49	99.36	99.41	99.41	98.21
CPU time 3k	15min	21min	1h:15	25 days	47h	84h	1h2
CPU time 6k	20min	28min	1h:35	11 days	45h	74h	40min
CPU time 8k	22min	30min	2h:10	11 days	47h	80h	27min

*Number of reference animals for population imputation
**Only family imputation carried out.
Overall missing call rate after imputation was zero for FImpute, Beagle and Impute2 but ranged from 0.12 to 0.43 for Pedimpute.

Table 3- Concordance rate for scenarios imputing from 3k to 50k SNP panel (combined family+population) according to the genotype status of the sire, dam and maternal grand-sire (MGS).

Sire	Dam	MGS	No	FImpute 5,133*	FImpute 10,337*	FImpute 45,047*	FImpute** + Beagle 5,133*	FImpute** + Impute2 5,133*	FImpute** + Impute2 10,337*	Pedimpute
50k	50k		4,545	99.21	99.21	99.21	99.20	99.21	99.21	98.79
50k	3k		4,151	98.65	98.66	98.67	98.54	98.58	98.58	97.82
50k	0k	50k	5,910	97.86	97.87	97.92	97.78	97.89	97.88	95.47
50k	0k	0k	233	96.15	96.23	96.56	95.66	96.01	95.98	94.28
50k	Unknown		4,728	96.92	97.02	97.26	96.09	96.47	96.56	95.79
3k	50k		8	96.40	96.59	96.73	95.45	96.20	96.42	96.16
3k	Unknown		8	94.62	94.53	94.72	94.58	95.67	95.28	93.59
0k	50k		53	97.14	97.18	97.55	96.64	96.90	97.03	95.46
0k	3k		53	96.05	96.15	96.40	95.55	95.94	96.03	94.32
0k	0k	50k	114	93.47	93.65	94.27	93.27	94.19	94.05	83.36
0k	0k	0k	62	91.78	92.07	92.96	92.07	92.95	93.01	77.76
0k	Unknown		47	93.28	93.44	94.18	93.06	93.73	93.80	85.23
Unknown	50k		6	97.67	97.71	97.65	95.51	96.24	96.67	95.82
Unknown	3k		8	95.88	96.26	96.38	95.36	96.07	96.09	94.86
Unknown	0k	50k	30	94.70	94.82	95.26	94.11	94.41	94.54	72.23
Unknown	0k	0k	5	94.59	94.57	96.18	92.65	93.34	93.51	65.94
Unknown	Unknown		39	95.88	95.95	96.13	94.96	95.10	95.23	64.95
Overall			20,000	98.01	98.05	98.14	97.76	97.91	97.93	96.52

*Number of reference animals for population imputation
**Only family imputation carried out.

Table 4- Concordance rate for scenarios imputing from 6k to 50k SNP panel (combined family+population) according to the genotype status of the sire, dam and maternal grand-sire (MGS).

Sire	Dam	MGS	No	FImpute 5,133*	FImpute 10,337*	FImpute 45,047*	FImpute** + Beagle 5,133*	FImpute** + Impute2 5,133*	FImpute** + Impute2 10,337*	Pedimpute
50k	50k		4,545	99.68	99.68	99.69	99.68	99.69	99.69	99.55
50k	3k		4,151	99.49	99.50	99.51	99.47	99.48	99.47	98.89
50k	0k	50k	5,910	99.31	99.32	99.35	99.30	99.32	99.30	96.59
50k	0k	0k	233	98.72	98.76	98.94	98.56	98.56	98.55	97.30
50k	Unknown		4,728	99.14	99.17	99.25	98.93	99.05	99.05	98.42
3k	50k		8	98.56	98.54	98.55	98.33	98.56	98.60	97.82
3k	Unknown		8	98.14	98.10	98.24	98.21	98.61	98.66	96.76
0k	50k		53	98.94	98.98	99.21	98.73	98.84	98.85	97.69
0k	3k		53	98.46	98.52	98.67	98.30	98.40	98.37	96.18
0k	0k	50k	114	98.09	98.16	98.48	97.69	97.99	97.93	87.15
0k	0k	0k	62	97.67	97.77	98.20	97.62	97.67	97.71	82.61
0k	Unknown		47	97.86	97.97	98.38	97.66	97.71	97.61	89.47
Unknown	50k		6	99.38	99.40	99.44	98.76	99.06	99.18	98.59
Unknown	3k		8	98.75	98.83	98.96	98.84	98.92	98.96	97.19
Unknown	0k	50k	30	98.21	98.22	98.37	98.01	97.98	98.04	73.03
Unknown	0k	0k	5	98.47	98.46	98.93	98.23	98.03	98.02	69.07
Unknown	Unknown		39	98.92	98.95	99.02	98.56	98.61	98.52	67.54
Overall			20,000	99.36	99.37	99.41	99.30	99.34	99.33	97.97

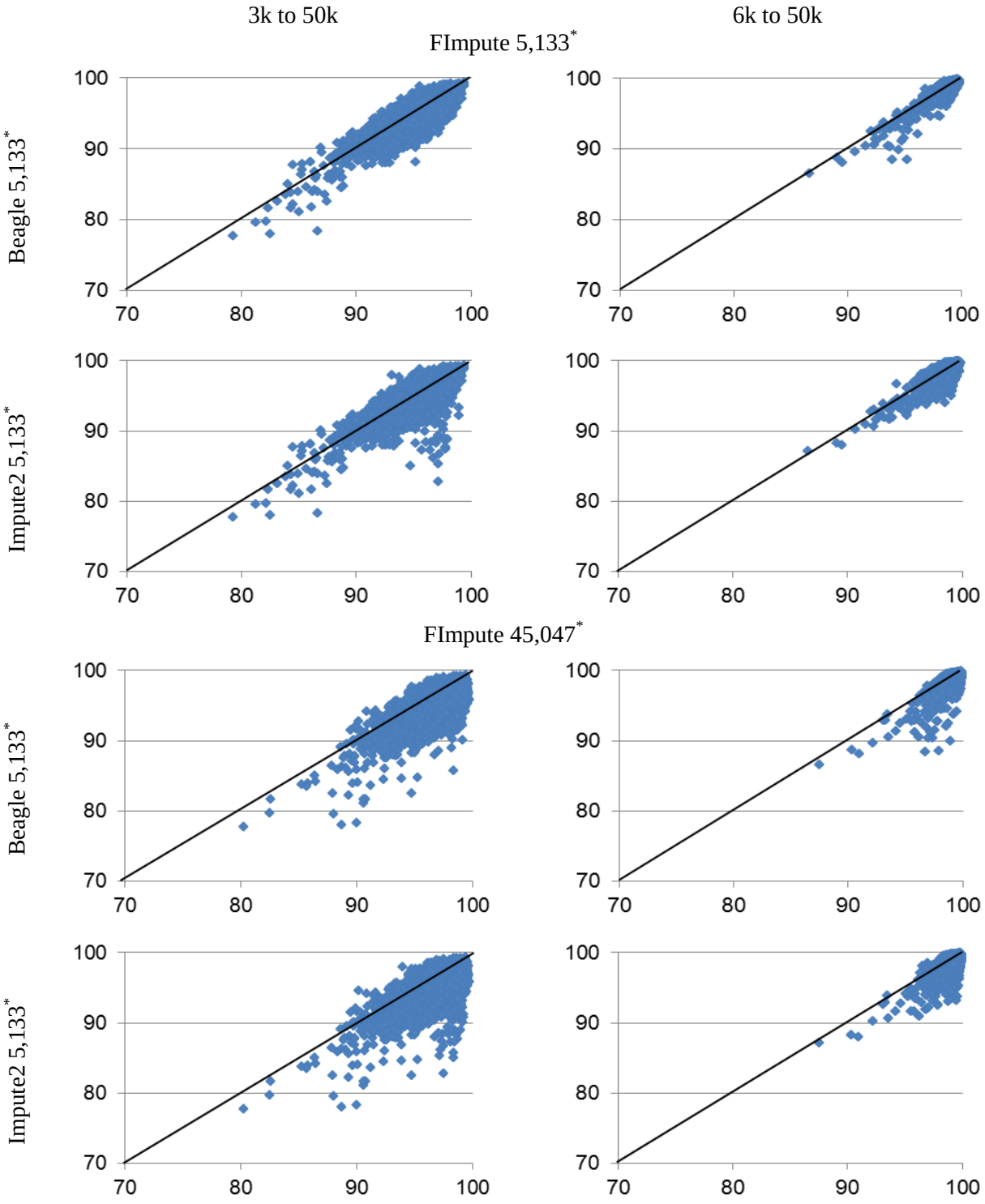
*Number of reference animals for population imputation
**Only family imputation carried out.

Table 5- Concordance rate for scenarios imputing from 8k to 50k SNP panel (combined family+population) according to the genotype status of the sire, dam and maternal grand-sire (MGS).

Sire	Dam	MGS	No	FImpute 5,133*	FImpute 10,337*	FImpute 45,047*	FImpute**+ Beagle 5,133*	FImpute**+ Impute2 5,133*	FImpute**+ Impute2 10,337*	Pedimpute
50k	50k		4,545	99.72	99.72	99.72	99.71	99.72	99.72	99.58
50k	3k		4,151	99.54	99.54	99.55	99.49	99.53	99.52	99.04
50k	0k	50k	5,910	99.39	99.40	99.43	99.36	99.40	99.38	97.22
50k	0k	0k	233	98.89	98.93	99.10	98.68	98.72	98.69	97.41
50k	Unknown		4,728	99.26	99.29	99.36	99.04	99.16	99.17	98.48
3k	50k		8	98.72	98.79	98.70	98.39	98.75	98.77	98.06
3k	Unknown		8	98.45	98.50	98.53	98.39	98.83	98.79	97.30
0k	50k		53	99.01	99.05	99.27	98.86	98.92	98.93	97.81
0k	3k		53	98.61	98.65	98.80	98.39	98.57	98.54	96.38
0k	0k	50k	114	98.32	98.40	98.69	97.84	98.14	98.11	87.35
0k	0k	0k	62	97.96	98.02	98.44	97.69	98.00	97.99	82.57
0k	Unknown		47	98.17	98.21	98.59	97.79	97.87	97.84	89.46
Unknown	50k		6	99.47	99.45	99.49	99.05	99.19	99.28	98.66
Unknown	3k		8	98.99	99.03	99.12	99.04	99.01	99.05	97.20
Unknown	0k	50k	30	98.46	98.51	98.65	97.88	98.18	98.20	71.78
Unknown	0k	0k	5	98.55	98.64	99.12	98.01	98.27	98.11	69.35
Unknown	Unknown		39	99.03	99.09	99.12	98.65	98.74	98.68	67.75
Overall			20,000	99.44	99.45	99.49	99.36	99.41	99.41	98.21

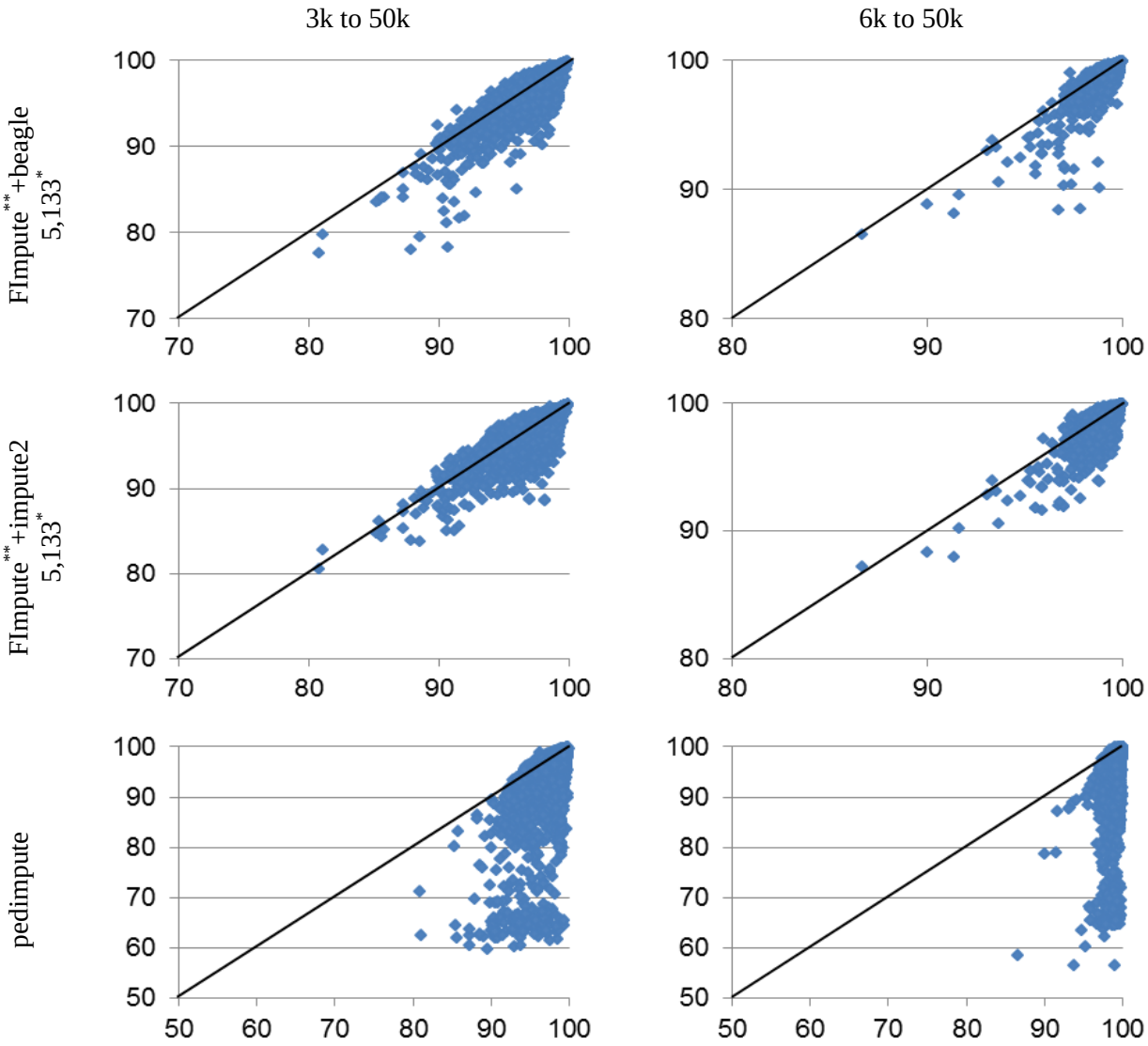
*Number of reference animals for population imputation
**Only family imputation carried out.

Figure 1 - Relationship between FImpute's accuracy, on the x-axis, and accuracy from other methods, on the y-axis, for population imputation only. The more points under the slope compared to above the worse the result for the other method compared to FImpute.



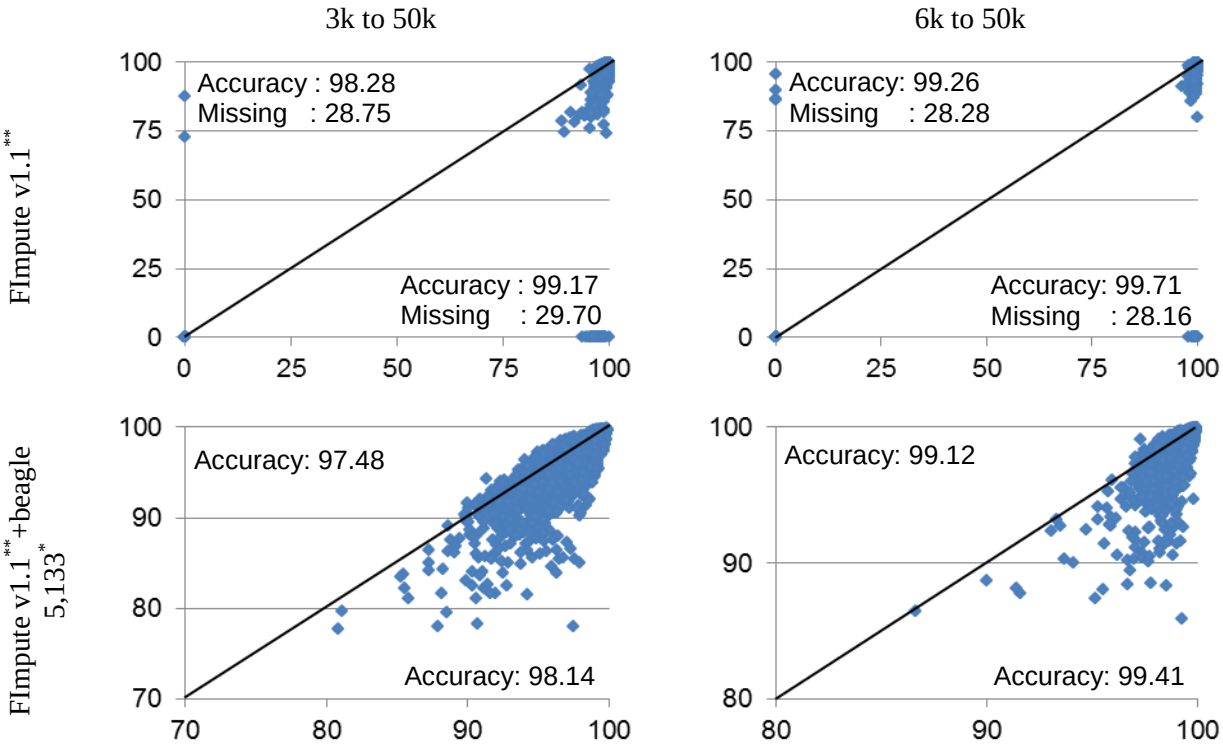
*Number of reference animals for population imputation

Figure 2 - Relationship between FImpute's accuracy (using 45,047 reference animals), on the x-axis, and accuracy from other methods, on the y-axis.



*Number of reference animals for population imputation
**Only family imputation carried out.

Figure 3 - Relationship between accuracy from FImputev2.2 (using 45,047 reference animals), on the x-axis, and accuracy from FImputev1.1+beagle, on the y-axis



*Number of reference animals for population imputation
**Only family imputation carried out.

Summary

- Most of the methods examined here had high imputation accuracies in terms of percentage concordance between SNPs, however, even when the accuracy is high, there can be substantial differences between methods for some animals, as evidenced by the plots.
- With the same reference size, the population imputation only from FImpute was as accurate as Beagle and Impute2.
- Once family information was taken into account, FImpute was substantially more accurate than the other software considered here.
- FImpute was able to handle very large reference populations, which was beyond the capability of Beagle and Impute2.
- The use of all available information in the population leads to higher imputation accuracy.
- Highly accurate imputation (family + population) using tens of thousands of reference individuals is possible in dairy cattle.
- FImputev2.2 leads to significantly higher imputation accuracy than the two-step FImpute v1.1+beagle currently used by CDN, because version 2.2 is more accurate than version 1.1 and the entire reference population can be used.

Recommendations

- Use a larger reference population size which should include 50k dams.
- Freezing animal genotypes should only be done if both parents are 50k genotyped.
- Replace the current two-step strategy with FImpute v2.2. If both the population and family imputation components of FImpute cannot be run together within the schedule of a genomic run, the population imputation can be run in advance (a two-step procedure using only FImpute, but with a much larger reference population size).