Success rate of imputation using different imputation approaches.

Jarmila Johnston, Gerrit Kistemaker

Introduction

Currently, 3k genotypes of Holstein, Jersey and Brown Swiss animals are imputed in Canada to 50k genotypes using FImpute software version 1 (Sargolzaei et al., 2010). Because animals with high number of missing genotypes tend to have a higher error rate of imputation, only animals with less than 10% missing rate are included in the genomic evaluation. On average, around 9% of Holstein 3k animals (2% of animals releasable in Canada) and 11% of Jersey 3k animals do not pass this threshold and despite that they were genotyped, they do not receive their genomic estimated breeding values.

As shown in Table 1, the current imputation method successfully imputes (missing rate <10%) genotypes of all 3k animals with both parents genotyped with 50k panel. Animals with 50k sire and with either dam or maternal grand sire (MGS) genotyped have over 90% chance that their 50k genotype will be successfully imputed. On the other hand, animals with at least one parent unknown have very low success rate of imputation.

Recently, Mehdi Sargolzaei released a new version of FImpute and also USDA (Paul VanRaden) released an updated version of their findhap imputation program (VanRaden, 2010). There is also range of other imputation programs available (IMPUTE, MACH, fastPHASE, PLINK and BEAGLE), however, majority of the programs were designed for human population, where sample size is relatively small and consequently some of the programs are not capable to handle the size of our datasets. The other weakness of those programs is their speed. These programs impute genotypes by phasing and sorting haplotypes into clusters via hidden Markov model, which is a very accurate method and it works well even when relationship between individuals is not considered but it requires a lot of CPU time. In our preliminary study we compared the two most popular programs: MACH and BEAGLE using chromosome 1 genotypes from Brown Swiss. Both programs imputed missing genotypes of both animals with and without complete pedigree with high accuracy. However, MACH was two times slower than BEAGLE. BEAGLE is currently used in New Zealand by LIC for imputation from 3k to 50k panel but also for imputation from 50k to high-density panel. The accuracy from BEAGLE reported by LIC ranged was 96% for imputation from 3k to 50k panel, and 99% accuracy for imputation from 50k to high-density panel (Johnson, 2011).

FImpute uses mainly family information for imputation, on the other hand, findhap and BEAGLE use population based imputation. Therefore it is expected that these programs will be able to impute genotypes of animals with incomplete pedigree, which are the animals that are not imputed by FImpute. The aim of this study was to compare success rate of imputation with FImpute version 1 (**M1**), FImpute version 2 (**M2**), findhap version 1 (**U1**), findhap version 2 (**U2**) and with BEAGLE and investigate which imputation approach will reduce the number of animals that do not qualify for genomic evaluation due to high missing rate.

Data

Jersey and Holstein data from February 2011 were used for the comparison of imputation method. Brown Swiss data were not considered because it contained only limited number (20) of 3k animals. The Holstein data set consisted of 54,466 - 50k genotypes and 18,629 – 3k genotypes. The Jersey dataset contained 5,057 – 50k genotypes and 3,882 – 3k genotypes. Both FImpute and findhap were run with the above mentioned data. Because BEAGLE is slower than FImpute and findhap and we would not be able to obtain results from BEAGLE before this meeting, BEAGLE was run with reduced dataset. Run time of BEAGLE is linear function of number of markers and quadratic function of number of samples. In order to reduce the computational time, BEAGLE was run on dataset that were already "pre-imputed" by M2, and genotypes were imputed for only animals that had >10% missing rate on at least one chromosome. Genotypes of proven bulls served as reference population.

Results

As shown in Tables 2 and 3, imputation by M2 resulted in lower imputation success rate compared to the currently used M1. With this method additional 10% of animals would not qualify for genomic evaluation in both Jersey and Holstein. In Holstein, U1 had very good imputation success (>90%) in animals when at least one parent and one grandparent were genotyped. Overall imputation success was by 1% higher than with M1. However, this imputation method did not perform that well with Jersey genotypes. In this case only animals with both parents genotyped with 50k panel had higher than 90% chance of being successfully imputed. Only 73% Jersey 3k animals would qualify for genomic evaluation, which is by 16% less than with current method. The new version of findhap (U2) tends to impute almost all unobserved genotypes. However, one has to keep in mind that the relationship between missing rate and error rate of imputation is not the same in FImpute compared to findhap. U2 imputes majority of genotypes but some of them are inaccurately imputed, while M1 and especially M2 are more conservative and impute a genotype only when it has high certainty of being correct and otherwise they set it to missing. To combine advantages of mainly family based imputation by FImpute and population based imputation by findhap, missing genotypes were imputed first with M2 and then with U2. With this approach all Jersey animals except of 7 were successfully imputed (Holstein results were not available before the deadline for this report).

Combination of M2 and BEAGLE imputation resulted in 100% imputation success rate for both Holstein and Jersey animals. This is similar to the success rate obtained with M2 + U2. However, M2+BEAGLE will likely provide more accurately imputed genotypes than M2+U2. This will be discussed in the following paper.

Conclusion

Imputation by FImpute version 2 followed by imputation with BEAGLE seems to be the best approach for imputation of 3k genotypes to 50k genotypes in terms of imputation success. With this approach all 3k animals would be imputed with missing rate < 10% and consequently all of these animals would qualify for genomic evaluation.

Acknowledgment

Mehdi Sargolzaei, Paul VanRaden and Brian Browning are acknowledged for providing their software for this study.

References

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

Johnson D.L., R.J. Spelman, M.K. Hayr and M.D. Keehan (2011) Imputation of single nucleotide polymorphism genotypes in a crossbred dairy cattle population using a reference panel, 2011.

Sargolzaei M., F. Schenkel, J. Chesnais (2010) Comparison between two methods of imputation (AIPL vs Boviteq/CGIL), for 6,246 genotypes provided by AIPL. *Dairy Cattle Breeding and Genetics Committee Meeting, October 5, 2010, University of Guelph, ON, Canada*

VanRaden P.M. , J.R. O'Connell, G.R. Wiggans, and K.A. Weigel (2010) Combining different marker densities in genomic evaluation. *Interbull Bulleting 42.*

Table 1: Number of animals with <10% missing rate (# imputed animals) and percentage of successfully imputed animals (% success rate) using February 2011 variable length

Sire	Dam	Holstein			Jersey		
		# animals	# imputed animals	success rate (%)	# animals	# imputed animals	success rate (%)
50k	50k	4,585	4,585	100	332	332	100
	3k	802	797	99	649	631	97
	0k, MGS 50k	9,610	9,392	98	2,265	2,143	95
	0k, MGS 0k	1,366	1,038	76	298	187	63
	unknown	884	473	54	42	0	0
3k	50k	91	91	100	1	1	100
	3k	5	5	100	3	3	100
	0k, MGS 50k	84	84	100	0	-	-
	0k, MGS 0k	5	1	20	0	-	-
0k, PGS 50k	50k	72	71	99	3	2	67
	3k	13	12	92	44	30	68
	0k, MGS 50k	382	272	71	108	77	71
	0k, MGS 0k	212	113	53	46	24	52
	unknown	65	2	3	1	0	0
0k, PGS 0k	50k	7	7	100	4	3	75
	3k	3	3	100	15	7	47
	0k, MGS 50k	38	24	63	44	16	36
	0k, MGS 0k	27	11	41	20	10	50
	unknown	15	0	0	1	0	0
unknown	50k	2	1	50	0	-	-
	3k	1	0	0	-	-	-
	0k, MGS 50k	32	0	0	1	0	0
	0k, MGS 0k	103	0	0	0	-	-
	unknown	225	0	0	5	0	0
All animals		18,629	16,982	91	3,882	3,466	89

Table 2: Ability of different imputation programs to impute untyped genotypes of Holstein 3k animals

	Dam	Number of 3k animals	Number of successfully imputed animals					Imputation success (%)				
		#	**M1**	**M2**	**U1**	**U2**	**M2+B**	**M1**	**M2**	**U1**	**U2**	**M2+B**
50k	**50k**	4,585	4,585	4,545	4,547	4,585	4,585	100	99	99	100	100
	3k	802	797	780	791	802	802	99	97	99	100	100
	0k, MGS 50k	9,610	9,392	8,694	9,378	9,610	9,610	98	90	98	100	100
	0k, MGS 0k	1,366	1,038	706	1,076	1,365	1,366	76	52	79	100	100
	unknown	884	473	0	610	880	884	54	0	69	100	100
3k	**50k**	91	91	91	91	91	91	100	100	100	100	100
	3k	5	5	5	5	5	5	100	100	100	100	100
	0k, MGS 50k	84	84	82	82	84	84	100	98	98	100	100
	0k, MGS 0k	5	1	0	2	5	5	20	0	40	100	100
0k, PGS 50k	**50k**	72	71	71	71	72	72	99	99	99	100	100
	3k	13	12	11	12	13	13	92	85	92	100	100
	0k, MGS 50k	382	272	99	279	382	382	71	26	73	100	100
	0k, MGS 0k	212	113	8	55	212	212	53	4	26	100	100
	unknown	65	2	0	8	63	65	3	0	12	97	100
0k, PGS 0k	**50k**	7	7	6	7	7	7	100	86	100	100	100
	3k	3	3	0	1	3	3	100	0	33	100	100
	0k, MGS 50k	38	24	8	14	38	38	63	21	37	100	100
	0k, MGS 0k	27	11	1	2	26	27	41	4	7	96	100
	unknown	15	0	0	0	13	15	0	0	0	87	100
unknown	**50k**	2	1	0	2	2	2	50	0	100	100	100
	3k	1	0	0	0	1	1	0	0	0	100	100
	0k, MGS 50k	32	0	0	9	32	32	0	0	28	100	100
	0k, MGS 0k	103	0	0	3	98	103	0	0	3	95	100
	unknown	225	0	0	12	218	225	0	0	5	97	100
All animals		18,629	16,982	15107	17,057	18,607	18,629	91	81	92	100	100

Table 3: Ability of different imputation programs to impute untyped genotypes of Jersey 3k animals

Sire	Dam	Number of 3k animals	Number of successfully imputed animals						Imputation success (%)					
			M1	M2	U1	U2	M2+U	M2+B	M1	M2	U1	U2	M2+U2	M2+B
50k	**50k**	**332**	332	329	316	332	332	332	100	99	95	100	100	100
	3k	**649**	631	585	564	645	649	649	97	90	87	99	100	100
	0k, MGS 50k	**2265**	2143	1983	1768	2265	2265	2265	95	88	78	100	100	100
	0k, MGS 0k	**298**	187	139	114	294	296	298	63	47	38	99	99	100
	unknown	**42**	0	0	21	41	42	42	0	0	50	98	100	100
3k	**3k**	**1**	1	1	1	1	1	1	100	100	100	100	100	100
	0k, MGS 50k	**3**	3	1	0	3	3	3	100	33	0	100	100	100
0k, PGS 50k	**50k**	**3**	2	2	2	3	3	3	67	67	67	100	100	100
	3k	**44**	30	15	21	44	44	44	68	34	48	100	100	100
	0k, MGS 50k	**108**	77	2	13	108	108	108	71	2	12	100	100	100
	0k, MGS 0k	**46**	24	0	3	46	46	46	52	0	7	100	100	100
	unknown	**1**	0	0	0	1	1	1	0	0	0	100	100	100
0k, PGS 0k	**50k**	**4**	3	1	2	4	4	4	75	25	50	100	100	100
	3k	**15**	7	1	3	15	15	15	47	7	20	100	100	100
	0k, MGS 50k	**44**	16	0	3	44	44	44	36	0	7	100	100	100
	0k, MGS 0k	**20**	10	1	1	20	20	20	50	5	5	100	100	100
	unknown	**1**	0	0	0	0	0	1	0	0	0	0	0	100
unknown	**0k, MGS 0k**	**1**	0	0	0	1	1	1	0	0	0	100	100	100
	unknown	**5**	0	0	0	5	4	5	0	0	0	100	80	100
All animals		**3882**	3466	3060	2832	3872	3878	3882	89	79	73	100	100	100

M1 – Fimpute version 1, M2 – Fimpute version 2, U1 – findhap version 1, U2 – findhap version 2, B - BEAGLE