J. Dairy Sci. 96:2649–2653
http://dx.doi.org/10.3168/jds.2012-6062

Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information

E. L. Nicolazzi,*†[1] S. Biffani,‡ and G. Jansen‡§
*Consorzio di Ricerca e Sperimentazione degli Allevatori (CRSA), Via G. Tomassetti 9, Rome 00161, Italy
†Parco Tecnologico Padano, Via Einstein, Lodi (LO) 26900, Italy
‡Associazione Nazionale Allevatori Frisona Italiana (ANAFI), Via Bergamo 292, Cremona (CR) 26100, Italy
§Dekoppel Consulting, Casale Rovera 10, Chiaverano (TO) 10010, Italy

ABSTRACT

Routine genomic evaluations frequently include a preliminary imputation step, requiring high accuracy and reduced computing time. A new algorithm, PedImpute (http://dekoppel.eu/pedimpute/), was developed and compared with findhap (http://aipl.arsusda.gov/software/findhap/) and BEAGLE (http://faculty.washington.edu/browning/beagle/beagle.html), using 19,904 Holstein genotypes from a 4-country international collaboration (United States, Canada, UK, and Italy). Different scenarios were evaluated on a sample subset that included only single nucleotide polymorphism from the Bovine low-density (LD) Illumina BeadChip (Illumina Inc., San Diego, CA). Comparative criteria were computing time, percentage of missing alleles, percentage of wrongly imputed alleles, and the allelic squared correlation. Imputation accuracy on ungenotyped animals was also analyzed. The algorithm PedImpute was slightly more accurate and faster than findhap and BEAGLE when sire, dam, and maternal grandsire were genotyped at high density. On the other hand, BEAGLE performed better than both PedImpute and findhap for animals with at least one close relative not genotyped or genotyped at low density. However, computing time and resources using BEAGLE were incompatible with routine genomic evaluations in Italy. Error rate and allelic squared correlation attained by PedImpute ranged from 0.2 to 1.1% and from 96.6 to 99.3%, respectively. When complete genomic information on sire, dam, and maternal grandsire are available, as expected to be the case in the close future in (at least) dairy cattle, and considering accuracies obtained and computation time required, PedImpute represents a valuable choice in routine evaluations among the algorithms tested.

Key words: single nucleotide polymorphism, imputation, Holstein, dairy cattle

Short Communication

Imputation of genotypes is becoming routine in genomic evaluations (**GE**), both to reduce the number of missing genotypes in samples with low genotype calls (Hickey et al., 2012) and in the across-SNP chip imputation of missing genotypes (Lund et al., 2011). Imputation has become an essential tool for GE for several reasons, namely (1) the recent availability of commercial SNP chips with different densities, (2) the routine exchange of genotypes across countries (Lund et al., 2011; VanRaden et al., 2012), (3) the expected steady increase in low-density genomic testing of females on commercial farms (Weigel et al., 2012), and (4) the proven increase in accuracy of genomic selection obtained by imputing lower to higher densities of SNP (VanRaden et al., 2011). Ideally, the imputation algorithm implemented in GE should be precise and fast, considering the large number of animals already genotyped in multiple countries and its continuous growth (i.e., especially the female population in dairy cattle), and the fact that frequent GE are required (e.g., monthly/weekly).

Recently, many different imputation methods in different breeds were compared (Johnston et al., 2011; Hickey et al., 2012; Sun et al., 2012); FImpute (Sargolzaei et al., 2012) and findhap (VanRaden et al., 2011) were good compromises between computational burden and overall accuracy. Although Johnston et al. (2011) and Sun et al. (2012) obtained better results using FImpute we consider findhap v.2 in this work, as it is the algorithm implemented in the United States for routine genomic evaluations and has some similarities to the algorithm we present. In spite of its computation time and memory consumption, BEAGLE (Browning and Browning, 2007) was considered as well, as it has been reported to obtain high accuracies without using pedigree in structured populations (Johnston et al., 2011; Sun et al., 2012). Here, we present a new algorithm that is similar to findhap v.2 (hereafter called findhap) in its use of pedigree and population information but aims to be still faster and more accurate.

Received August 16, 2012.
Accepted January 2, 2013.
[1]Corresponding author: ezequiel.nicolazzi@tecnoparco.org

A total of 19,904 total Holstein genotypes (19,446 bulls and 458 cows) of a 4-country international collaboration (United States, Canada, Italy, and UK) were available. All animals were genotyped with the Illumina BovineSNP50 BeadChip v1 or v2 (Illumina Inc., San Diego, CA). Only SNP in common between both chips were retained (thus, hereafter named **54K**). Raw genotypes had already been controlled for sample missingness (<12%) and Mendelian inheritance (<1%). Further SNP-editing thresholds were >5% for missing genotypes, <2% for minor allele frequency, and $P <$ 0.05% for Hardy-Weinberg equilibrium. Moreover, although Johnston et al. (2011) showed that very high imputation accuracies can be obtained in the X chromosome, sex chromosomes and SNP not assigned to any chromosome were discarded, as they are not used in the Italian Holstein genomic evaluation.

The samples were highly related, as 1,118 bulls sired the whole data set; 44 of those had more than 100 sons/daughters (13 had more than 200 sons/daughters and 1 bull had 588 sons/daughters). A total of 18,965 (~95%) of all samples had its sire genotyped and 18,008 (~90%) had also the maternal grandsire (**MGS**) genotyped. Only 255 had both sire and dam genotyped. Genotypes of a subset of 4,233 samples were forced to missing for SNP not present on the Bovine low-density (LD) chip (84% of SNP, after editing). The subset samples were chosen using the following criteria: 50% of all young bulls without genotyped progeny, 50% of all cows available, and any cow with at least 1 genotyped son whose sire was also genotyped. A total of 89 bulls with at least 5 sons genotyped had their genotypes completely set to missing to test the accuracy of imputation for ungenotyped closely related animals. The genome for all samples was subdivided into single chromosomes, which were analyzed independently using the same Linux server with 4 × 4 Intel Xeon5560 2.80GHz processors (Intel Corp., Santa Clara, CA).

The PedImpute algorithm (http://dekoppel.eu/pedimpute/) details can be found in the Supplementary Materials (available online at http://www.journalofdairyscience.org/). Briefly, PedImpute reconstructs haplotypes and imputes their missing alleles for a general pedigree. It has been designed to perform well in pedigrees with mainly medium to large half-sib families as in most dairy cattle breeds. The program considers all genotyped animals and closely connected nongenotyped animals (with at least one genotyped parent or progeny). The iterative process alternates the use of pedigree information with population haplotypes to gradually fill in missing alleles in the haplotypes. The pedigree part uses variable length segments (as long as possible without including mismatching genotypes) for each parent-offspring pair, in contrast with findhap, where fixed-length segments are used. Both algorithms use fixed-length segments for the population part. These iterations are then repeated a predefined number of times (e.g., outer iteration), setting different haplotype lengths for the population-based step.

Performances using default (i.e., recommended for low- to high-density imputation) options of PedImpute, findhap, and BEAGLE algorithms (except for memory usage for the latter algorithm) were compared. The findhap default options included one round more of outer iterations than PedImpute. The BEAGLE algorithm considered genotypes as unrelated. Chromosomes were analyzed in parallel, up to 6 at a time for PedImpute and findhap. Because of the large amount of memory assigned to each process [e.g., 10 GB of random-access memory (RAM)], BEAGLE was run with up to 3 chromosomes at a time. The whole genome (e.g., 29 chromosomes) was imputed in 10 and 37 min using PedImpute and findhap, respectively. Imputation using BEAGLE took between 30 and 190 h per single chromosome, depending on the number of markers.

To compare the performance of the 3 algorithms, 3 measures were considered: percentage of missing alleles (% missing), percentage of wrongly imputed alleles (% errors), and the allelic squared correlation (allelic-R^2), considering only observed (or imputed) genotypes, but not correcting them by their frequency.

Results of the different scenarios considered are presented in Table 1. In general, allelic-R^2 increased and both % errors and missing SNP decreased when more information on relatives was available for PedImpute and findhap. Not considering any pedigree, genotypic information on relatives did not influence the performance of BEAGLE, which maintained a constant 0.4 and 98.8 for % error and allelic-R^2, respectively, over all scenarios.

When 54K genotypes of sire, dam, and MGS were available, the imputation accuracy was 99.3% in PedImpute, almost 1% higher than in findhap and 0.5% higher than BEAGLE. In all other scenarios, BEAGLE obtained the highest allelic-R^2 and the lowest % error.

The percentage of allelic error ranged from 1.1 and 1.2% (when only the sire was genotyped) to 0.2 and 0.5% (when sire, dam, and MGS were genotyped) for PedImpute and findhap, respectively. It was 0.4% across all scenarios for BEAGLE. Slight differences in allelic-R^2 were observed between PedImpute and findhap (+0.9, +0.6, −0.8, and +0.4 for scenarios A to D; Table 1). In a more realistic scenario, where dams are genotyped with the LD SNP chip, 99.5% of the allelic calls were correct in PedImpute. The BEAGLE algorithm obtained better allelic-R^2 results in scenarios B to D. Among the 2 pedigree-based algorithms, findhap was slightly outperformed by PedImpute in all sce-

Table 1. Imputation results from 3 analyzed algorithms for the most representative scenarios[1]

		PedImpute			findhap			BEAGLE		
Scenario[2]	n	% missing	% error	Allelic-R^2	% missing	% error	Allelic-R^2	% missing	% error	Allelic-R^2
A	104	0.1	0.2	99.3	<0.1	0.5	98.4	—	0.4	98.8
B	87	0.3	0.5	98.3	<0.1	0.8	97.4	—	0.4	98.8
C	3,827	0.4	1.1	96.6	<0.1	0.8	97.4	—	0.4	98.8
D	174	0.6	1.0	96.7	<0.1	1.2	96.2	—	0.4	98.8

[1]PedImpute: http://dekoppel.eu/pedimpute/; findhap: http://aipl.arsusda.gov/software/findhap/; BEAGLE: http://faculty.washington.edu/browning/beagle/beagle.html; % missing = percentage of missing alleles; % error = percentage of wrongly imputed alleles; allelic-R^2 = allelic squared correlation.

[2]Scenarios: A: sire, dam, and maternal grandsire genotyped at 54,000 SNP (54K); B: sire and maternal grandsire genotyped at 54K and dam genotyped at low density; C: sire and maternal grandsire genotyped at 54K and dam not genotyped; D: sire genotyped at 54K and dam and maternal grandsire ungenotyped.

narios, except for scenario C (sire and MGS genotyped with a 54K SNP chip), where it obtained an almost equal % error and a 0.8 higher allelic-R^2 value. Missing genotypes in PedImpute were the lowest (0.1%) when all close relatives were genotyped, and increased to 0.6% when only the sire was genotyped. Note that only PedImpute leaves some genotypes with missing values, whereas BEAGLE assigns the most probable genotype in any case and findhap gives a call for almost all genotypes (missing <0.01%). The allelic-R^2 used in this work considers successful calls only. To assess the effect of the uncalled genotypes from PedImpute on the overall allelic-R^2, we filled these genotypes with the most frequent genotype for completely uncalled markers or the most frequent allele when only one allele was called by PedImpute. The overall allelic-R^2 declined slightly to 99.2, 97.9, 96.1, and 95.9 in scenarios A to D, respectively. The relative ranking of the methods remained unchanged except in case D, where PedImpute fell slightly below findhap.

The accuracy of ungenotyped animals was analyzed for findhap and PedImpute only, as the pedigree was considered only in those algorithms. The ungenotyped animals were analyzed as follows: first, the 89 animals with their genotypes set completely to missing were used to assess the accuracy of imputation for ungenotyped animals. This analysis was performed on PedImpute only, as findhap does not output the imputed ungenotyped male population and BEAGLE does not consider pedigree information. Ungenotyped sires with at least 5 sons showed a 98% fill rate, and an allelic-R^2 of 92.9. Fill rate increased to 100% and allelic-R^2 reached 96.5 when the number of sons exceeded 10.

Then, all ungenotyped animals written in the output of both programs were compared in terms of fill rate (percentage of SNP assigned by the algorithm). Note that findhap outputs only female ungenotyped animals, whereas PedImpute outputs both male and female ungenotyped animals. The fill rate of 3,458 ungenotyped cows and bulls and 2,616 ungenotyped cows imputed by PedImpute and findhap, respectively, are reported in Table 2. Almost all ungenotyped close relatives to genotyped animals had less than 10% missing imputed genotypes in PedImpute. Only 93 animals remained completely ungenotyped. Almost the opposite was observed in findhap, where 2,223 ungenotyped cows had more than 10% of missing genotypes after imputation. Two possible explanations can be given for the large differences observed. The first one is a practical reason: at least one chromosome was not retrieved by findhap for a large quota of the total amount of ungenotyped animals because it did not meet the 90% fill rate required (and was accounted as 0% fill rate for that chromosome). However, it should be noted that the findhap fill rate threshold was probably intended for imputation of the entire genome. Thus, the fact that in this work we ran single chromosomes could have penalized findhap on these analyses. The second reason is that PedImpute relies more heavily on the closeness of relationships between animals in the data set than findhap. For example, in its first iteration, PedImpute uses pedigree information, whereas findhap runs the first 2 iterations using population information with no extra information coming from pedigree. No assessment is possible in terms of accuracy for these animals. However, accuracy for young animals would depend partly on the homozygosity of the genotyped parent and mostly on its relatedness to the population. Considering these results, we would not suggest to include (ungenotyped) sires with low numbers of progeny genotyped. In any case, when considering the common 205 ungenotyped animals, the fill rate in PedImpute was 2% higher than in findhap.

The results obtained in this work do not take into account possible genotyping errors. Considering that most of the imputation errors found by the 3 imputa-

Table 2. Fill rate for all ungenotyped animals

	PedImpute[1]		findhap[2]
Fill rate	Bulls	Cows	Cows
0%	14	79	0
0 to <90%	0	0	2,223
90 to <95%	0	4	167
95 to <98%	1	160	136
98 to <99%	14	466	58
>99%	161	2,559	32

[1]Ungenotyped cows and bulls included in the output (http://dekoppel.eu/pedimpute/).

[2]Only ungenotyped cows are included in the output (http://aipl.arsusda.gov/software/findhap/).

tion algorithms were restricted to allelic errors, it is possible that the allelic-R^2 we obtained were actually underestimates.

The performance of PedImpute on less related populations (or with a lower degree of genome-wide linkage disequilibrium) has not been tested so far. In fact, the relationship between the animals considered was an essential factor influencing the results obtained. The population used for this work was highly structured and closely related. However, we included all the available genotypes (i.e., no specific selection of samples was carried out); thus, the data set used can be considered as an example of a general dairy cattle data set dominated by large half-sib families. In any case, both PedImpute and findhap are expected to perform better with more close relatives genotyped.

It is difficult to predict the performance of these methods with greater density of SNP chips [i.e., imputing from 54K to 800,000 (800K) SNP]. Although linkage disequilibrium between markers will increase (thus, in general, accuracies should be higher), the scenarios will be different from the ones tested in this work, as dams are expected to be genotyped at lower densities than sires and grandsires.

The number of cases in the scenarios differed greatly. Scenario C (sire and MGS genotyped), so far probably the most common scenario in many cattle breeds, had the highest number of observations. In this scenario, BEAGLE obtained the best results, and findhap obtained better results than PedImpute; thus, if we consider the results as a whole (e.g., without considering the different scenarios), both BEAGLE and findhap would have obtained better results than PedImpute. The choice to divide the observations into different scenarios depended on the fact that in the (very) close future most of bulls dams are expected to be genotyped on either LD or 54K SNP chips. In those cases (scenarios A and B), PedImpute excelled compared with findhap (but not compared with BEAGLE in the case of the dam genotyped at low density). However, considering the low number of observations in some scenarios, it is possible that the differences observed could depend on sampling variance, rather than on real differences in terms of performance of the methods. On the other hand, BEAGLE obtained slightly better accuracies in almost all scenarios. However, it used a very large amount of computing time and resources, and this might limit its application to a monthly/weekly genomic system (e.g., Italian Holstein).

The results obtained in this paper confirm previous observations on a data set with a lower number of animals (Nicolazzi et al., 2012). Considering these promising results, and the fact that the Italian female population is expected to be heavily genotyped in the close future, the Italian Holstein Association (ANAFI, Cremona, Italy) has introduced the PedImpute algorithm in its national genomic evaluation routines.

ACKNOWLEDGMENTS

The authors acknowledge Associazione Nazionale Allevatori Frisona Italiana (ANAFI, Cremona, Italy) and the Innovagen Project for supporting this research and the North American Cooperative Dairy DNA Repository (CDDR, Washington, DC), Canadian Dairy Network (CDN, Guelph, ON, Canada), and DairyCO (Kenilworth, UK) for access to the genotypes. Furthermore, the people involved in the Selmol, ProZoo, Elica, and Innovagen projects, which contributed many Italian genotypes, are gratefully acknowledged. Many thanks to Paul VanRaden (USDA-ARS, Beltsville, MD) for making the source code for findhap publicly available.

REFERENCES

Browning, S. R., and B. L. Browning. 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.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. J. van der Werf, and M. A. Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44:9.

Johnston, J., G. Kistemaker, and P. G. Sullivan. 2011. Comparison of different imputation methods. Interbull Bull. 44:25–33.

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel. Evol. 43:43.

Nicolazzi, E. L., S. Biffani, and G. Jansen. 2012. PEDIMPUTE: Imputing genotypes using a fast algorithm combining pedigree and population information. Interbull Bull. 46:33–38.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2012. Accuracy of imputed 50k genotypes from 3k and 6k chips in dairy cattle breeds using FImpute. Plant and Animal Genome (PAG) meeting, San

Diego, CA. Accessed Feb. 6, 2013. https://pag.confex.com/pag/xx/webprogram/Paper4334.html.

Sun, C., X.-L. Wu, K. A. Weigel, G. J. M. Rosa, S. Bauck, B. W. Woodward, R. D. Schnabel, J. F. Taylor, and D. Gianola. 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. (Camb.) 94:133–150.

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

VanRaden, P., K. Olson, D. Null, M. Sargolzaei, M. Winters, and J. B. C. H. M. Van Kaam. 2012. Reliability increases from combining 50,000- and 777,000-marker genotypes from four countries. Interbull Bull. 46:75–79.

Weigel, K. A., P. C. Hoffman, W. Herring, and T. J. Lawlor. 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. J. Dairy Sci. 95:2215–2225.