PEDIMPUTE:
Imputing Genotypes using a Fast Algorithm Combining Pedigree and Population Information

E.L. Nicolazzi[1,2], S. Biffani[3] and G. Jansen[3,4]
[1] CRSA, Via G. Tomassetti 9, Rome, Italy.
[2] Parco Tecnologico Padano, Via Einstein, Lodi (LO), Italy.
[3] ANAFI, Via Bergamo 292, Cremona (CR), Italy.
[4] Dekoppel Consulting, Casale Rovera 10, Chiaverano(TO), Italy.

Abstract

Routine genomic evaluations frequently include a preliminary imputation step, requiring high accuracy and reduced computing time. A new algorithm, namely PEDIMPUTE, was developed and compared to FINDHAP algorithm using 6662 genotypes from the Italian Holstein population. Different scenarios were evaluated creating two subsets including only SNPs from the Bovine 3k and LD Illumina BeadChip, respectively. The comparative criteria were % missing alleles, % of wrongly imputed alleles, and the allelic squared correlation. PEDIMPUTE was slightly more accurate and faster than FINDHAP in almost all scenarios. Error rate and allelic squared correlation attained by PEDIMPUTE ranged from 0.2 to 3.7 % and from 88.6 to 99.3 %, respectively.

Keywords: SNP, imputation, Holstein, dairy cattle**.**

Introduction

Imputation of genotypes is becoming routine in genomic evaluations (GE) (Cooper *et al.*, 2010). In the first place, imputation tools are used to reduce the number of missing genotypes in samples with low genotype calls (but over the exclusion threshold). A second use of imputation is the across-SNPchip imputation of missing genotypes. The recent availability of Illumina SNPchips with different densities makes imputation an essential tool for GE. This is particularly true nowadays, when genomic testing of female on commercial farms is expected to increase steadily (Weigel et al, 2012). Moreover, imputing lower to higher densities of SNPs has proved to increase accuracy of genomic selection (VanRaden *et al.*, 2011).

In addition, the importance of routinely impute genotypes coming from different SNPchips and genotyping platforms, is evidenced when considering the International consortia exchanging genotypes on daily basis.

Taking into consideration the large number of animals already genotyped in multiple countries, the number of SNP chips currently available and the fact that GE are required to be released (officially or non-officially) nearly on a monthly basis, there is a striking need for an imputation tool that is both precise and fast to cope with this (additional) computational burden with the minimum loss of accuracy.

Recently, many different imputation methods in different breeds were compared (Johnston *et al.*, 2011; Gredler *et al.*, 2011). FINDHAP (VanRaden *et al.*, 2011) was a good compromise between computational burden and overall accuracy. Furthermore it is the algorithm implemented in the USA in routine genomic evaluations.

Here we present a new algorithm that is similar to Findhap in its use of pedigree and population information but aims to be still faster and more accurate.

Materials & Methods

Data

A total of 6662 total Holstein genotypes, 6147 bulls and 455 cows, obtained from several Italian projects (mainly SelMol, ProZoo, Elica

and Innovagen) were available. All animals were genotyped with the Illumina BovineSNP50 BeadChip v1 or v2. Raw genotypes had already been controlled for sample missingness (<12%) and errors in the non pseudo-autosomal part of the X chromosome (<5%) and mendelian inheritance (<1%). Only SNPs in common between both chips were retained (thus hereafter named just as "54k"). Further editing thresholds for SNPs were: >5% for missing genotypes, <2% for minor allele frequency and P<0.05% for HWE. Moreover, sex chromosomes and SNPs not assigned to any chromosome were discarded, as they are not used in the Italian Holstein genomic evaluation.

The samples were highly related, as 768 bulls sired the whole dataset (5 bulls had more than 100 sons/daughters). A total of 6001 samples had its sire genotyped, 5234 of which had also the maternal grandsire genotyped. Only 255 had both sire and dam genotyped.

For 1171 samples, two subsets were created by forcing to missing SNPs not in common first with the Bovine 3k and then with the LD (~7k) chip (Table 1). These "reduced" samples were chosen using the following criteria: 50% of all young bulls without genotyped progeny and 50% of all cows available (plus all cows where at least a son and his sire were genotyped). The genome for all samples was sub-divided into single chromosomes and run independently.

PEDIMPUTE algorithm

General characteristics. PEDIMPUTE reconstructs haplotypes and imputes missing alleles in those haplotypes for a general pedigree. It has been designed to work well for pedigrees dominated by medium to large half-sib families as in dairy cattle. The input pedigree file should contain all known ancestors – genotyped or not – of the genotyped animals, and must be numbered sequentially from oldest to youngest. The program constructs a trimmed pedigree with all genotyped animals and closely connected non-genotyped animals (with at least one genotyped parent or progeny) and attempts to reconstruct haplotypes for every animal in the trimmed pedigree during the iterative process. The iteration alternates between the use of pedigree information and 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 whereas the population part works with fixed length segments. PEDIMPUTE handles a single chromosome at a time, storing all genotypes and haplotypes in memory. Multiple chromosomes can of course be run in parallel depending on availability of processors and memory in the computer.

Initial setup. The program assumes that the genotype data have been edited to eliminate incompatible parent-offspring genotypes at the whole genome level. Input genotypes are read as contiguous strings of properly ordered markers with genotypes coded as 0 or 2 for the two homozygotes, 1 for heterozygotes and 5 for non-called genotypes.

Paternal and maternal haplotypes are allocated and are initialized to 1 to signify unknown alleles. All animals in the trimmed pedigree are processed from oldest to youngest. Single marker genotypes that are not compatible with known sire and dam genotypes are set to unknown (for the animal). All paternal and maternal alleles that can be deduced from the animal, sire and dam genotypes are filled as 0 for the first allele or 2 for the second allele. Prior to initiating iteration, paternal vs. maternal alleles are assigned at heterozygous loci for genotyped founder animals with unknown parents or no genotyped ancestors and at least one genotyped progeny. For simplicity, the alleles clearly inherited by the first born progeny are arbitrarily assigned to the paternal haplotype of the founder.

Unknown alleles (coded 1) in the haplotypes are filled by iterative improvement. Each iteration consists of a round of pedigree-based imputation followed by a round of population-based imputation.

Pedigree based imputation. Animals in the trimmed pedigree are processed from

youngest to oldest. Paternal haplotypes along the whole chromosome are considered first. The first step is to mark the grandparental origin at all informative markers, i.e. those at which the paternal allele has already been assigned and is compatible with the sire's paternal or maternal allele, but not both. Secondly, the longest possible stretches of markers with the same grandparental origin at the external markers and no crossovers are identified. For all markers in the interval, unknown alleles in the matching sire haplotype are filled from known alleles in the animal's paternal haplotype and vice versa. The process is then repeated for the maternal haplotype and the dam's haplotypes before proceeding to the next youngest animal in the pedigree.

Population based imputation. This part of the imputation process draws heavily on the algorithm of FINDHAP. Population haplotypes are not stored separately, but rather as a list of links to the first animal bearing each haplotype. Chromosome segments of a fixed length (e.g. 300 markers) are processed one by one. For each segment, the trimmed pedigree is processed from oldest to youngest animal. An animal's paternal and maternal haplotype segments are considered separately. The animal's haplotype in the segment is compared sequentially to the list of population haplotypes and a match is found if there are at least *m* matching marker alleles and no more than *n* mismatching alleles among the filled alleles in both haplotypes. In the first iteration, only segments with a fill-rate of at least 75% are considered and in subsequent iterations only those with at least 1.5*m* filled alleles. If a match is found, the unknown alleles in the population haplotype (i.e. some other animal's haplotype) are filled using the animal's known alleles and vice versa. If no match is found, a new haplotype is added to the list. At the end of the round, the population list of haplotypes is sorted by decreasing number of matches found in order to increase the efficiency in the next iteration.

Outer iterations. Following the lead of FINDHAP, an outer loop was added to allow the whole iterative process to be repeated with different haplotype lengths for the population based step. Here we stepped down the segment lengths from 300 to 150, and finally to 75 markers.

Comparison of algorithms

Performances using default options of both PEDIMPUTE and FINDHAP algorithms were compared. FINDHAP default options included one round more of outer iteration than PEDIMPUTE (first step in FINDHAP considered a segment length of 600).

Chromosomes were analyzed in parallel, up to 6 at a time. To compare performances of both algorithms, three measures were considered: % missing alleles (%missing), % of wrongly imputed alleles (%errors), and the allelic squared correlation (allelic-R^2). Note that FINDHAP fills not imputed genotypes (=missing) with the most frequent genotype in its last run, so %missing in FINDHAP always equals 0.

Table 1. SNPs in common between Bovine Illumina SNPchips.

	Total # SNPs	**Total # SNPs *after* editing**	**% SNPs to be imputed[1]**
54k	54001	41502	-
3k	2900	2502	93.97%
LD	6909	6574	84.16%

[1]Average % of SNPs to be imputed (not considering individual missing genotype calls).

Results & Discussion

Performances of PEDIMPUTE and FINDHAP imputation algorithms were tested on the same dataset, using the same computer. The whole genome was imputed in 3.14 and 8.28 minutes using PEDIMPUTE and FINDHAP, respectively (3k to 54k). These times were nearly the same when imputing LD to 54k.

During the initial PEDIMPUTE setup phase, the haplotype fill-rate (fraction of 0 or 2 codes) reached about 80 percent for both paternal and maternal haplotypes for animals genotyped at 54K, but only 60-66 or 0-6 percent for paternal or maternal haplotypes, respectively, for animals not genotyped or genotyped at low density.

As regards pedigree based imputation phase, using PEDIMPUTE haplotype fill-rate after the first pedigree round increased to 94% for animals genotyped at 54K but only to 22-37% for maternal haplotypes of other animals.

In the population base imputation phase, the best performance were observed with m=2 and n=0. During the first round of haplotyping fill-rate improved marginally, whereas the improvement was dramatic in the second and third iteration, alternating with pedigree based imputation. Convergence in terms of fill-rate was achieved in 3 iterations with little or no improvement thereafter.

Results of the different scenarios are presented in Table 2.a. (for 3k to 54k) and Table 2.b. (for LD to 54k). Only those scenarios with more than 50 samples were considered in this paper.

In general, allelic R^2 increased and %errors and missing SNPs decreased almost linearly when more information on relatives was available. Considering the 3k to 54k imputation, when 54k genotypes of sire, dam and maternal grandsire (MGS) were available, imputation accuracy was 99.3%. The reduction of the % of allelic error for FINDHAP ranged from 3.9% (when only the sire was genotyped), to 2.2% (when sire, dam and MGS were genotyped). Allelic R^2 was always higher in PEDIMPUTE (differences of +4.9, +5.2, +2.1 and +0.9 for scenarios 1 to 4, respectively).

The same trend was maintained when imputation was from LD to 54k SNPchip. In this case, PEDIMPUTE reached 99.8% of correct allelic calls when sire, dam and MGS were genotyped. In a more realistic scenario, where dams are more probably genotyped with the LD SNPchip, 99.5% of the allelic calls were correct. FINDHAP was slightly outperformed in all scenarios, except for scenario 3 (sire and MGS genotyped with a 54k SNPchip), where it obtained a nearly equal % of errors and a 0.2 higher allelic R^2 value.

Preliminary tests run by ANAFI on additional 15,000 animals showed that these results (both differences in terms of accuracy of the algorithms and in relative differences in terms of computation burden) hold also when the number of samples increases (data not shown).

Table 2. Imputation results in scenarios with more than 50 samples for a) Scenarios imputing a 3k SNPchip; b) Scenarios imputing an LD SNPchip (7k). Runtime is in parenthesis, expressed in minutes.

a)

					PEDIMPUTE (3.14m)			**FINDHAP** (8.28m)		
Scenario	**Sire**	**Dam**	**MGS**	**n**	**%missing**	**%error**	**allelic_R^2**	**%missing**	**%error**	**allelic_R^2**
1	54k	54k	54k	98	**0.1**	**0.7**	**97.8**	**0**	**2.2**	**92.9**
2	54k	3k	54k	80	**0.3**	**1.6**	**94.9**	**0**	**3.2**	**89.7**
3	54k	--	54k	868	**0.4**	**2.9**	**91.0**	**0**	**3.5**	**88.9**
4	54k	--	--	115	**0.4**	**3.7**	**88.6**	**0**	**3.9**	**87.7**

b)

					PEDIMPUTE (3.21m)			**FINDHAP** (8.17m)		
Scenario	**Sire**	**Dam**	**MGS**	**n**	**%missing**	**%error**	**allelic_R^2**	**%missing**	**%error**	**allelic_R^2**
1	54k	54k	54k	98	**0.2**	**0.2**	**99.3**	**0**	**0.6**	**98.1**
2	54k	7k	54k	80	**0.5**	**0.5**	**98.4**	**0**	**0.9**	**97.1**
3	54k	--	54k	868	**0.8**	**1.2**	**96.2**	**0**	**1.1**	**96.4**
4	54k	--	--	115	**0.7**	**1.1**	**96.4**	**0**	**1.3**	**95.9**

Conclusion

Both PEDIMPUTE and FINDHAP algorithms confirm previous findings (VanDoormal *et al.,* 2012) where imputation on the LD SNPchip was more accurate than on the 3k chip. Allelic R^2 and %errors were much lower in all scenarios. Differences between both imputation algorithms were more evident when imputing 3k->54k than when going from LD->54k. This is probably because of how the first iterations of both algorithms work: a first pedigree iteration in PEDIMPUTE helps to fill-in some gaps that are useful in the population iteration, whereas FINDHAP runs two population iterations first with no extra-information coming from pedigree. This extra-information provided by the first pedigree iteration, is obviously less important when the number of SNPs is higher.

In any case, PEDIMPUTE was slightly more accurate than FINDHAP in almost all scenarios. It should be noted, however, that the accuracy figures exclude missing genotypes that were left unimputed by PEDIMPUTE. The accuracy of PEDIMPUTE would be slightly lower if it were required to fill all genotypes, as does FINDHAP. The % of not imputed alleles (%missing) in PEDIMPUTE remained lower than 0.8% of the SNPs in all scenarios, and in general was proportional to the amount of close relatives genotyped.

The % of imputation errors was the lowest when close relatives were genotyped with a 54k SNPchip (VanRaden *et al.,* 2011), although only 0.3 higher %errors were found in both PEDIMPUTE and FINDHAP, when a more realistic scenario (dam genotyped with an LD SNPchip) was considered. However, performance of this new algorithm on lesser-related populations has not been tested so far. In such populations, specific algorithms that consider Linkage Disequilibrium only and more iterations (e.g. BEAGLE, AlphaImpute, etc) will probably perform better than PEDIMPUTE and FINDHAP.

In this paper, sex chromosomes are not taken into account, as these chromosomes are excluded from the official genomic evaluation in the Italian Holsteins. Both programs impute genotypes for some non-genotyped ancestors but their performance in this regard has not yet been compared. Research is ongoing to test these aspects as well.

Further research on the performance of PEDIMPUTE in other scenarios is still ongoing. For example, the impact on the parameters analyzed in case a larger quota of the female population is genotyped, the 54k to 800k imputation, among others. However, considering these promising results, the Italian Holstein Association has introduced this algorithm in its national official genomic evaluation.

PEDIMPUTE can be downloaded at http://dekoppel.eu/pedimpute/.

Acknowledgments

Authors wish to acknowledge Selmol, ProZoo, Elica and Innovagen projects (and people) contributing to provide genotypic data. Many thanks to Paul VanRaden for making the source code for FINDHAP publicly available.

References

Cooper, T.A., Tooker, M.E., VanRaden, P.M., Wiggans, G.R. & Cole, J.B. 2010. AIPL web-page, available at: http://aipl.arsusda.gov/reference/changes/aprilInformation.htm#impute. Accessed [01 May 2012]

Gredler, B., Seefried, F.R., Schuler, U., Bapst, B., Schnyder, U. & Hickey, J.M. 2011. Imputation in Swiss cattle breeds. *Interbull Bulletin 44,* 8-11.

Johnston, J., Kistemaker, G. & Sullivan, P.G. 2011. Comparison of different imputation methods. *Interbull Bulletin 44,* 25-33.

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

VanDoormal, B. & Muir, V. 2012. Genotyping with low density panels: 3k vs 6k. CDN web-page, available at: http://www.cdn.ca/document.php?id=262. Accessed [01 May 2012].

Weigel, K.A., Hoffman, P.C., Herring, W. & Lawlor, V. 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. *J. Dairy Sci. 95:4,* 2215-25.