G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(1):10-13
https://doi.org/10.5808/GI.2018.16.1.10

ORIGINAL ARTICLE

Accuracy of Imputation of Microsatellite Markers from BovineSNP50 and BovineHD BeadChip in Hanwoo Population of Korea

Aditi Sharma[1§], Jong-Eun Park[1§], Byungho Park[1], Mi-Na Park[1], Seung-Hee Roh[2], Woo-Young Jung[2], Seung-Hwan Lee[3], Han-Ha Chai[1], Gul-Won Chang[1], Yong-Min Cho[1], Dajeong Lim[1]*

[1]Animal Genomics & Bioinformatics Division, National Institute of Animal Science, Rural Development Administration, Wanju 55365, Korea, [2]Hanwoo Genetic Improvement Center of the Nonghyup Agribusiness Group Inc., Seosan 31948, Korea, [3]Division of Animal and Dairy Science, Chungnam National University, Daejeon 34134, Korea

Until now microsatellite (MS) have been a popular choice of markers for parentage verification. Recently many countries have moved or are in process of moving from MS markers to single nucleotide polymorphism (SNP) markers for parentage testing. FAO-ISAG has also come up with a panel of 200 SNPs to replace the use of MS markers in parentage verification. However, in many countries most of the animals were genotyped by MS markers till now and the sudden shift to SNP markers will render the data of those animals useless. As National Institute of Animal Science in South Korea plans to move from standard ISAG recommended MS markers to SNPs, it faces the dilemma of exclusion of old animals that were genotyped by MS markers. Thus to facilitate this shift from MS to SNPs, such that the existing animals with MS data could still be used for parentage verification, this study was performed. In the current study we performed imputation of MS markers from the SNPs in the 500-kb region of the MS marker on either side. This method will provide an easy option for the labs to combine the data from the old and the current set of animals. It will be a cost efficient replacement of genotyping with the additional markers. We used 1,480 Hanwoo animals with both the MS data and SNP data to impute in the validation animals. We also compared the imputation accuracy between BovineSNP50 and BovineHD BeadChip. In our study the genotype concordance of 40% and 43% was observed in the BovineSNP50 and BovineHD BeadChip respectively.

Keywords: cattle, genotype prediction, Hanwoo, parentage verification

Introduction

Microsatellite (MS) markers have remained a popular choice for parentage verification since two decades now. For cattle there is a standard set of nine MS markers recognized as "international marker set" recommended by international society of animal genetics (ISAG) which need to be included in the parentage testing panels to facilitate record exchange between laboratories. So far, MS markers have successfully been implemented in cattle and other livestock species. However, as the cost of single nucleotide polymorphism (SNP) genotyping has decreased, more and more new animals are being genotyped with SNP chip panels. For parentage testing all the animals are required to be genotyped with same type of markers. So either new animals should be typed with MS markers or old animals that are usualy typed with MS markers should be typed with SNP markers. In both the cases, it will incur an additional cost. So in order to shift from MS to SNPs, McClure *et al*. [1, 2] suggested imputation of MS markers from SNP genotypes. This method will provide a cost effective and accurate choice for replacement of markers for parentage verification. Depending upon the relationship amongst the sampled individuals generally 2–3 SNPs per MS are needed to obtain the accuracy good enough for genetic identification and assessment of parentage [3]. Fernandez *et al*. [3] found that a set of 24 SNPs were equivalent to the ISAG recommended

Received November 6, 2017; Revised November 21, 2017; Accepted December 11, 2017

***Corresponding author:** Tel: +82-63-238-7306, Fax: +82-63-238-7347, E-mail: lim.dj@korea.kr

[§]These authors contributed equally to this work.

set of international MS markers.

The term "genotype imputation" refers to the prediction of missing genotypes, i.e., genotypes that were not directly genotyped in the sampled individuals [4]. Imputation requires a reference population that has all the markers genotyped for all the samples. This reference population is then used to predict the genotypes in the target population which contains missing genotypes or missing markers. Imputation of genotypes has become a common practice in genome- wide association studies, fine mapping of QTLs, genomic predictions and whole genome based diversity studies. Since denser chips are known to perform better in the downstream analysis many laboratories use imputation to move from low density to high density SNP's [5]. The accuracy of imputation also depends on the density of the SnipP-chip, denser it is better predictions it would make [6]. Ogawa *et al*. [6] found higher accuracy of imputation when they used 10,000 SNPs instead of 3,000 for genotype imputation in Japanese black cattle. Accuracy increased from 90% to 97% with the increase in number of SNPs. Several factors that affect imputation accuracies include minor allele frequency of the SNPs in the reference population, size of the reference population, genetic relationship between the reference and test populations [7] and linkage disequilibrium between the imputed SNP and the SNP on the target data [8].

Imputed data can provide accurate results in the downstream analysis only if the accuracy of imputation is high. In this study we report the accuracy of imputation of MS markers, from the BovineSNP50 and BovineHD BeadChip in Hanwoo cattle of Korea. Comparison between two SNP panels was made to identify the SNP panel and SNP subset that gives the best accuracy such that the overall cost of genotyping could be reduced while not having to compromise with the accuracy of prediction.

Materials and Methods

Ethics statement

For sampling individuals in this study, the standard operating procedures were reviewed and approved by the National Institute of Animal Science's Institutional Animal Care and Use Committee (Permit Number: NIAS2015-774).

Animals, genotyping, and quality control

Blood samples for genotyping were obtained from 1,482 Hanwoo individuals reared at the Hanwoo Genetic Improvement Center of the Nonghyup Agribusiness Group Inc. (Seosan, Korea). Genomic DNA was extracted from the blood samples using DNeasy 96 Blood and Tissue Kit (Qiagen, Valencia, CA, USA). DNA quantification was performed using a NanoDrop 1000 (Thermo Fisher Scientific Inc., Wilmington, DE, USA). DNA samples were submitted for genotyping with total DNA of 900 ng, 260/280 ratio >1.8, and DNA concentration of 20 ng/μL. The SNP genotyping was done by using a BovineSNP50 BeadChip version 2 (Illumina, San Diego, CA, USA). These animals were then imputed to the BovineHD data (777k SNP chip) using another set of Hanwoo animals as reference. MS marker genotyping data for the same animals was also obtained from the Hanwoo Improvement Center of the National Agricultural Cooperative Federation (Seosan, Korea). Eight MS markers belonging to the ISAG recommended list were included in the study (Supplementary Table 1). Markers on the sex chromosomes were ignored. PLINK version 1.9 (http://www.cog-genomics.org/plink/1.9/) [9] was used for the quality control of the raw genotype data. Quality control was performed on the BovineSNP50 and BovineHD BeadChip data for minor allele frequency (0.05), missingness (0.05), Hardy-Weinberg equilibrium (HWE; 0.0001) and genotyping quality (0.05). Twenty-seven hundred ninety-four SNPs were removed based on missingness, 14,190 SNPs were removed based on frequency, 2,395 markers were excluded based on HWE After quality control there were 1,482 animals and 37235 SNPs in the reference genotype dataset. All the data was split chromosome wise and SNPs within the 500 kb range on either side of the MS marker, i.e., 1,000 kb in total were extracted. Only the SNPs that were in the specified range were further used for imputations. There was no family data available to be included in the study.

Genotype imputation and estimation of imputation accuracy

Locations of the 8 MS on UMD3.1 reference genome were identified from University of California, Santa Cruz Genome Browser. The SNP data was merged with MS data and was used as reference for imputations. Out of all the animals 20% were used as validation while the rest were used as the reference animals. Beagle program [10] was used for determining the phase and imputation of the missing markers. Beagle uses Li and Stephens haplotype frequency models to performs imputation into phased haplotypes. Imputation method used by beagle is both computationally and memory efficient [11]. Beagle was used as it can handle both the bi-allelic and multi-allelic markers. First MS and SNP genotypes were phased independently and then the two types of datasets were merged and were phased again. This phased data was used as the reference for the imputations. A fivefold validation was performed to check the accuracy of imputation. Accuracy of imputation was measured by calculating the genotype concordance rate. Correlation between the true genotypes and the predicted genotypes were calculated. Accuracies were averaged over all five cross validation sets (Table 1). The allelic concordance, i.e., at least one of the allele was identified correctly, was also calculated. In addition, we compared if the numbers of iterations had any effect on the accuracy of the imputation. Accuracies of imputation were compared between two SNP panels.

Results and Discussion

The number of SNPs used for imputation for the eight MS markers ranged from 9 to 24 (average 15) for BovineSNP50 and 151 to 296 for BovineHD (average 232). The number of alleles for MS markers ranged between 7 for BM1824 to 24 for TGLA227. The effective number of MS alleles varied from 3.4 in BM1824 to 8.0 in TGLA53. The observed heterozygosity varied from 0.7 in BM1824 to 1.0 in TGLA53 (Table 2).

With BovineSNP50, the highest accuracy of 50% was recorded for TGLA122 and TGLA227 while with BovineHD most of the markers had an accuracy of 50%. The minimum imputation accuracy of 1% was observed for TGLA53 with both the SNP chip panels. TGLA53 had ~40% missing genotypes which could have attributed to the reduction in average accuracy. The genotype concordance rate averaged over all the loci was 40% for the BovineSNP50 whereas it was 43% for BovineHD (Table 1).

The accuracy was limited by marker TGLA53. Accuracy increased to ~50% with BovineHD if TGLA53 marker was removed from the analysis. The allelic concordance of 30% and 43% with BovineSNP50 and BovineHD respectively was seen in the validation samples. The average correlation between the predicted and true genotypes was 31% and 15%, respectively with BovineSNP50 and BovineHD, respectively. Highest correlation was seen for TGLA227 and lowest in TGLA53 with Bovine SNP50. With BovineHD highest correlation was seen for BM1824 and lowest for TGLA53. Accuracy of imputation is known to increase with the increase in reference population size and also by including the familial genotype data in the reference population. Also including the genotypes from the related individuals in the reference population allows the Beagle program to infer haplotypes correctly and thus make better predictions for the ungenotyped marker.

Marker density is known to affect the accuracy of imputation. Higher imputation accuracy with increased marker density has been shown by Hayes *et al.* [12]. While we did observe an increase in accuracy with the HD SNP panel, however it was not high enough to be used in routine practice. McClure *et al.* [2] observed higher accuracies as compared to our study. They used the validation animals which were derived from the reference population whereas

Table 2. Details of Microsatellite markers for the total 1,482 animals

Locus	Na	Ne	Ho
BM1824	7	3.432	0.699
BM2113	12	3.652	0.727
ETH10	10	4.835	0.892
ETH225	11	6.563	0.99
TGLA53	15	7.608	0.955
TGLA227	24	4.979	0.999
TGLA126	19	5.098	0.838
TGLA122	11	3.732	0.811

Na, no. of different alleles; Ne, no. of Effective alleles; Ho, observed heterozygosity.

Table 1. Accuracy of imputation of MS markers from Bovine 50K beadchip and HD SNP chip data in Hanwoo cattle averaged over five cross validation sets

Marker	Chromosome	50K				777K			
		No. of SNPs[a]	Genotype concordance	Allele[b]	Correlation[c]	No. of SNPs	Genotype concordance	Allele[b]	Correlation[c]
BM1824	Chr1	14	0.4	0.34	0.4	151	0.5	0.19	0.52
BM2113	Chr2	24	0.4	0.3	0.32	248	0.4	0.27	0.36
ETH10	Chr5	16	0.4	0.24	0.4	256	0.5	0.12	0.42
ETH225	Chr9	9	0.4	0.4	0.2	159	0.55	0.12	0.39
TGLA53	Chr16	9	0.01	0.12	0.04	75	0.01	0.11	0.02
TGLA227	Chr18	17	0.5	0.12	0.5	296	0.5	0.09	0.47
TGLA126	Chr20	12	0.4	0.4	0.22	243	0.5	0.21	0.32
TGLA122	Chr21	16	0.5	0.11	0.43	268	0.51	0.07	0.49
Average		15	0.40	0.30	0.31	212	0.43	0.15	0.38

MS, microsatellite; SNP, single nucleotide polymorphism.

[a]Number of SNPs in the 500-kb flanking region of the MS marker; [b]At least one of the alleles were imputed correctly; [c]Correlation coefficient between true and predicted genotypes.

Table 3. Effect of iterations on the genotype imputation accuracy based on BovineHD SNP panel

Iteration	Average	Max	Min
100	0.40	0.50	0.02
200	0.40	0.50	0.02
300	0.40	0.50	0.02
400	0.40	0.50	0.02
500	0.40	0.50	0.02

SNP, single nucleotide polymorphism.

we lacked such design in our samples. Also, no significant increase was observed in number of genotypes imputed correctly with the increase in number of iterations (Table 3).

For the reference population to predict the MS alleles with higher accuracies we need multiple generations of ancestors genotypes along with the pedigree information. For imputing MS markers from SNP data we suggest using related animals. Such studies need to be optimized well before they could be used in routine practice.

ORCID: Aditi Sharma: https://orcid.org/0000-0001-8907-1187; Jong-Eun Park: https://orcid.org/0000-0003-0718-3463; Byungho Park: https://orcid.org/0000-0001-6195-4519; Mi-Na Park: https://orcid.org/0000-0001-7078-9463; Seung-Hee Roh: https://orcid.org/0000-0003-0267-8846; Woo-Young Jung: https://orcid.org/0000-0001-9144-7374; Seung-Hwan Lee: https://orcid.org/0000-0003-1508-4887; Han-Ha Chai: https://orcid.org/0000-0001-7752-3967; Gul-Won Chang: https://orcid.org/0000-0001-5090-2107; Yong-Min Cho: https://orcid.org/0000-0002-4181-4428; Dajeong Lim: https://orcid.org/0000-0003-3966-9150

Authors' contribution

Conceptualization: DL, AS
Data curation: SHL, BP, MNP, SHR, WYJ, HHC
Formal analysis: AS
Funding acquisition: JEP, SHL, DL, GWC, YMC
Methodology: DL, AS
Writing – original draft: AS, JEP, DL
Writing – review & editing: AS, DL

Acknowledgments

This work was supported by Agenda (PJ01134902) of the National Institute of Animal Science. We acknowledge the support from different institutions and their personnel providing help for the sampling of cattle (Hanwoo Genetic Improvement Center of the Nonghyup Agribusiness Group Inc.) in Seosan, Chungnam province, Korea and cattle keepers for their assistance and permission to sample their herds. Funders had no role in design of the study, sampling, analysis or writing of the manuscript.

Supplementary material

Supplementary data including one table can be found with this article online at http://www.genominfo.org/src/sm/gni-16-10-s001.pdf.

References

1. McClure M, Sonstegard T, Wiggans G, Van Tassell CP. Imputation of microsatellite alleles from dense SNP genotypes for parental verification. *Front Genet* 2012;3:140.
2. McClure MC, Sonstegard TS, Wiggans GR, Van Eenennaam AL, Weber KL, Penedo CT, *et al*. Imputation of microsatellite alleles from dense SNP genotypes for parentage verification across multiple *Bos taurus* and *Bos indicus* breeds. *Front Genet* 2013;4:176.
3. Fernandez ME, Goszczynski DE, Liron JP, Villegas-Castagnasso EE, Carino MH, Ripoli MV, *et al*. Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genet Mol Biol* 2013;36:185-191.
4. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11:499-511.
5. Carvalheiro R, Boison SA, Neves HH, Sargolzaei M, Schenkel FS, Utsunomiya YT, *et al*. Accuracy of genotype imputation in Nelore cattle. *Genet Sel Evol* 2014;46:69.
6. Ogawa S, Matsuda H, Taniguchi Y, Watanabe T, Takasuga A, Sugimoto Y, *et al*. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle. *Anim Sci J* 2016;87:3-12.
7. Uemoto Y, Sasaki S, Sugimoto Y, Watanabe T. Accuracy of high-density genotype imputation in Japanese Black cattle. *Anim Genet* 2015;46:388-394.
8. Calus MP, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *Animal* 2014;8:1743-1753.
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
10. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009;84:210-223.
11. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* 2016;98:116-126.
12. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, van der Werf JH. Accuracy of genotype imputation in sheep breeds. *Anim Genet* 2012;43:72-80.