Preparation of Reference Data for Genomic Selection in Italian Holsteins

J.B.C.H.M. van Kaam[*], S. Biffani[†], R. Negrini[‡], G.B. Jansen[§], P. Ajmone-Marsan[**], J.L. Williams[††] and A. Nardone[‡‡]

Introduction

A major limitation in the selection of next generation parents is that half the genetic variance in offspring is accounted for by their parental genetic level; however the other half depends on Mendelian sampling. Two methods for capturing this Mendelian segregation are (1) own or progeny performance testing and (2) DNA analysis. The first method is costly and time consuming, which creates an opportunity for DNA based methodologies, which are less time consuming and now also cheaper.

At the 6th WCGALP in 1998 in Armidale, the term Genomic Selection was introduced by Visscher and Haley. However the key paper giving a first demonstration of its potential has been written by Meuwissen et al. (2001) and is based on simulated marker data. The availability of large numbers of markers cheap to analyze is realized more recently; in dairy cattle most countries followed the development of the Illumina BovineSNP50 BeadChip by Van Tassell et al. (2008). DNA based information can be modeled as single markers, haplotypes or other derived variables such as principal components.

The purpose of this study for genomic evaluation of Italian Holstein cattle was to 1) select samples, which were no replicates and free of known identity errors and 2) select single nucleotide polymorphisms (SNPs) to be used by removing SNPs with undesirable characteristics, i.e. unscorable, monomorphic, not mapped, low minor allele frequency (MAF) or minor genotype frequency (MGF), a large deviation from Hardy-Weinberg equilibrium or highly correlated with other SNPs and 3) determine the accuracy of the genotypes by investigating sire–son conflicts and concordance among genotypes produced by the repeated analysis of several animals.

[*] ANAFI - Italian Holstein Association, Via Bergamo 292, 26100, Cremona (CR), Italy
[†] PTP - Parco Tecnologico Padano – CERSA, Via Einstein - Polo Universitario, 26900, Lodi (LO), Italy
[‡] Università Cattolica del Sacro Cuore, Via E. Parmense 84, 29122, Piacenza (PC), Italy
[§] Dekoppel Consulting, Via Rovera 10, 10010, Chiaverano (TO), Italy
[**] Università Cattolica del Sacro Cuore, Via E. Parmense 84, 29122, Piacenza (PC), Italy
[††] PTP - Parco Tecnologico Padano – CERSA, Via Einstein - Polo Universitario, 26900, Lodi (LO), Italy
[‡‡] Università della Tuscia, Via De Lellis, 01100, Viterbo (VT), Italy

Material and methods

Reference population. In Italy, several projects have been set up, which contribute to the creation of a reference data set based on proven Holstein bulls with DNA information on 54,001 SNPs. Marker effects will be estimated using this reference data set. SelMol is the first Italian project which involves universities and research institutes. A second project is coordinated by the Parco Tecnologico Padano institute. These two projects together so far genotyped 2099 samples of 2066 Italian Holstein bulls. Furthermore Anafi, the Italian Holstein association, together with artificial insemination centers is increasing the reference population by adding more bulls including their sires and grandsires. At this moment many analyses are still underway, so intermediate results are presented. Genotypes were represented, using Interbull coding, by the number of the counted allele: 0, 1, and 2. A 5 indicates that the genotype could not be determined.

Analyses. SNPs that do not contribute to the accuracy of the genomic evaluations can be eliminated to reduce computational effort and to improve stability of estimates of the effects of the remaining SNP. Marker data processing consists of 3 main steps handled by Fortran 2003 programs. These steps are:

1. First 'Snprecode' is used to transfer marker information in row-wise format using Interbull coding. This results in more than 98% reduction on the marker data file. Incoming data can be either forward or AB allele coding and included 2459 samples with a call rate of 99.2%. Successively, 6 sample records known to have an animal identity error are removed, so 2453 samples remained.
2. Secondly, records belonging to animals from which multiple samples are available are merged with 'Samplemerge', in order to have just 1 record per animal. 'Samplemerge' verifies if known replicates are indeed replicates and if there are any unknown replicates. At this point there are 2377 unique bulls.
3. Third, 'Snpcheck' is used to select SNPs and animals. The main checks are:
    a. Checking number of missing genotypes/individual ≤ 5%
    b. Checking non-autosomality - X chromosome
        - For SNP selection a fraction heterozygous typed SNPs in males < 1%
        - For animal selection a fraction heterozygous typed SNPs in males < 5%
    c. Checking sex using X chromosome
    d. Checking parentage & Mendelian inheritance
        - SNPs with > 1% parent-offspring mismatch are flagged.
    e. Checking % Missing genotypes/SNP, Monomorphism, MAF, MGF & Hardy-Weinberg equilibrium
        - Flag SNPs with % Missing genotypes > 5%, MAF ≤ 2%, MGF ≤ 0.1%, HW p-value ≤ 0.5%. The Hardy-Weinberg test is based on Wigginton et al, 2005.
    f. Checking collinearity
        - The approach of Wiggans et al. (2009) is followed, comparing SNPs only with other SNPs with a MAF within 2.5%. SNPs with < 0.2% differences from either all the same genotypes or all opposite genotypes are flagged.

Checks for selecting animals and selecting SNPs are performed sequentially but within the same program. None of our SNP editing programs is very memory consuming and

computational efficiency is sufficient for the higher density SNP panels in arrival. Similar editing approaches have been published by Chan et al. (2008) and Wiggans et al. (2009).

Results and discussion

The concordance of readable SNPs was 99.91% between the 76 replicate samples that were merged. From the 2377 bulls, 22 bulls were flagged due to > 5 % heterozygous SNPs in the non-pseudoautosomal region of the X chromosome, 19 bulls were flagged due to a percentage missing SNPs > 10%. There was an overlap of 12 bulls between these 2 criteria. Furthermore 49 bulls were flagged due to Mendelian sire-son inconsistencies. Without overlap, 76 bulls have been removed, so 2301 bulls remain. Note that the 76 bulls removed by 'Snpcheck' are not the same as the 76 samples merged by 'Samplemerge'.

The results from SNP selection are displayed in Table 1 and 2. In Table 1, for each criterion is indicated how many SNPs were flagged, either only for this criterion or for the criterion in combination with other criteria as well. For example, no SNPs are flagged only for monomorphism, because such SNPs also get flagged for MAF, MGF and collinearity. The selection criteria with the most impact were collinearity and MGF. Most SNPs rejected for MAF were also rejected for MGF.

Table 1: Number of SNPs flagged for various criteria

	Criteria	Flagged SNPs for any criteria	Flagged SNPs only for this criteria
1	**Monomorphic**	3523	0
2	**Non-autosomal**	1491	373
3	**%Missing**	1078	472
4	**Mendelian**	1469	164
5	**MGF**	11084	994
6	**MAF**	9286	30
7	**Hardy-Weinberg**	3315	460
8	**Collinearity**	9390	1276
9	**X-chromosome**	1179	81
10	**All | None**	**14800**	**39201**

Table 2: Number of SNPs flagged for a combination of two or more criteria (upper-triangle) or only two criteria (lower-triangle) [α]

	Criteria	1	2	3	4	5	6	7	8	9	10
1	**Monomorphic**	.	217	26	0	3523	3523	0	3523	221	**3523**
2	**Non-autosomal**	0	.	65	719	1054	346	836	401	1030	**1491**
3	**%Missing**	0	9	.	261	211	176	435	168	49	**1078**
4	**Mendelian**	0	0	21	.	744	44	1275	87	752	**1469**
5	**MGF**	0	11	12	0	.	9193	2091	7982	1057	**11084**
6	**MAF**	0	0	2	0	1331	.	1413	7835	348	**9286**
7	**Hardy-Weinberg**	0	10	121	290	18	21	.	1442	861	**3315**
8	**Collinearity**	0	33	13	6	136	6	38	.	376	**9390**
9	**X-chromosome**	0	0	1	1	5	0	4	0	.	**1179**
10	**All**	**0**	**373**	**472**	**164**	**994**	**30**	**460**	**1276**	**81**	.

[α]Diagonal elements were omitted because they are identical to Table 1.

Remarkable is that Wiggans et al. (2009) starting from 57,000 SNPs, mainly the same ones, reported 6572 monomorphic SNPs, whereas we found only 3523 monomorphic SNPs. Yet we found 9286 SNPs with a MAF $\leq$ 2%, whereas Wiggans et al. (2009) reports of such 3649 SNPs. The difference is even more particular considering that Wiggans used data on 5503 bulls and therefore had more chance to find rare polymorphisms. This suggests that Italian Holsteins have more variation remaining than North-American Holsteins. Note that Italy has been using Holsteins since 1922, but possibly the presence of European blood explains the difference between the two populations.

Conclusion

Data editing resulted in a reduction from 2459 samples to 2301 bulls and from 54,001 SNPs to 39,201 SNPs. Results from SNP genotyping suggest that Italian Holsteins have more rare alleles and less monomorphic alleles than North-American Holsteins. This provides an incentive to use this additional variation in our breeding programs.

References

Chan, E.K.F., Hawken, R., and Reverter A. (2008). *Anim. Genet.*, 40:149–156.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). *Genetics*, 157:1819–1829.

Van Tassell, C., Smith, T.P.L., Matukumalli, L.K. *et al*. (2008). *Nature Methods*, 5:247-252.

Visscher, P.M., and Haley, C.S. (1998). In *Proc. 6th WCGALP*, volume 23, pages 503-510.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M. *et al.* (2009). *J. Dairy Sci.*, 92:3431–3436.

Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. (2005). *Am. J. Hum. Genet.*, 76:887–883.