SEQUENCING AND GENOTYPING FOR THE WHOLE GENOME SELECTION IN CANADIAN BEEF POPULATIONS

K. Stachowicz[1,2], S. Larmer[1], J. Jamrozik[1], S.S. Moore[3] and S.P. Miller[1,4]

[1]Centre for the Genetic Improvement of Livestock, University of Guelph, Canada
[2] AbacusBio Limited, Dunedin, New Zealand
[3]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Australia
[4]Department of Agricultural, Food and Nutritional Science, University of Alberta, Canada

SUMMARY

The project “Whole Genome Selection through Genome Wide Imputation in Beef Cattle” is a research initiative with a goal to develop low cost genome wide selection methodologies for Canada’s beef industry. Ten cattle populations were included: 6 purebred beef breeds (Angus, Charolais, Gelbvieh, Hereford, Limousin, and Simmental), Canadian Holsteins, and 3 composite beef populations. The first step was to use pedigree analysis to identify the key animals to be sequenced and genotyped. For each population, 30 animals will be sequenced, 480 genotyped with HD SNP panel and 560 genotyped with 50K SNP panel. Pedigree analysis revealed good data quality, i.e. pedigree completeness and depth. Ancestors with the highest genetic and inbreeding contributions to the reference population were identified. From the top animals, 30 were chosen for sequencing based on their relationships with each other, to avoid sequencing closely related animals. The top 30 identified ancestors explained from 41% to 63% of the population gene pool, depending on the breed. Younger bulls with high number of progeny were considered for genotyping in addition to the top ranking ancestors, in order to ensure sufficient links to the phenotypic data. Genotyping the top 1,000 animals will cover over 90% of the genetic base of those breeds and should allow for highly accurate genomic prediction.

INTRODUCTION

The Canadian Cattle Genome Project, formally entitled “Whole Genome Selection through Genome Wide Imputation in Beef Cattle” (www.canadacow.ca), is focused on delivering genomic technology to Canada’s beef industry. The project will include research to define the social and economic benefits and costs of using genomic technology in livestock improvement; develop tools for low-cost, accurate genome wide selection methodologies for breeders; and complete research so that genome wide selection can be used in Canadian herds for particularly difficult to measure yet valuable traits. Genotypes from a wide range of cattle populations will be used to develop accurate and robust genomic predictions.

Described is the method of identification of the key animals in the Canadian cattle populations to be sequenced and genotyped using the HD or 50K SNP panels.

MATERIALS AND METHODS

Pedigrees of purebred beef populations were obtained for Angus, Charolais, Gelbvieh, Hereford, Limousin, and Simmental breeds from respective breed associations. Holsteins were included as they make a significant contribution to global beef production and pedigree was provided by the Canadian Dairy Network. Analysis of each of the pedigrees was performed in order to assess data quality and pedigree structure using CFC (Sargolzaei *et al.* 2006) and Pedig (Boichard 2002) software. Completeness and depth of the pedigree are very important factors, which may affect the estimates of inbreeding coefficients, relationships among animals and also founder and ancestor contributions. Three different measures were used to assess the quality of pedigrees: percentage of animals with both parents known, discrete generation equivalent and pedigree completeness index. The average number of discrete generations (DGE) provides an indication of how many complete discrete generations were present in a given pedigree (Sölkner *et al*. 1998). Pedigree completeness index (PCI) as a harmonic mean of parental contributions, is always zero when either parent is unknown regardless of the depth and completeness of the pedigree of the other parent. Inbreeding can also only be estimated if information on both parents' ancestors is available. Therefore, PCI is an important measure of pedigree quality for inbreeding and relationship estimation (MacCluer *et al*. 1983).

An inbreeding coefficient of each individual in the pedigree was calculated and averaged for each year of birth. However, as the absolute values of inbreeding are relative to the quality and depth of pedigree, the rate of increase of inbreeding (ΔF) per year (or per generation) should be used when comparing between different populations or assessing how inbreeding is accumulating in the population. It was also used to find effective population size for each breed, calculated as: $N_e=1/2\Delta FL$, where *L* was the average generation interval. The reference population included animals born between 2006 and 2011, which represented the last generation.

Effective number of founders is a measure of founders' contribution to the current population and reflects the unequal contributions of founders due to selection rates and variation of family size (Lacy 1989). Effective number of founder genomes is the number of equally contributing founders with no loss of founder alleles that would give the same amount of genetic diversity as is present in the reference population. It accounts for the loss of genetic diversity that occurred in the population due to genetic drift and bottlenecks. Effective number of ancestors and their genetic contributions (Boichard *et al*. 1997) were calculated to identify ancestors with highest marginal and total genetic contributions to the reference population. Additionally, the decomposition of inbreeding into ancestral components was performed, which allowed the identification of ancestors with the highest contribution to inbreeding in the reference population.

In order to choose animals for sequencing, top 100 ancestors with the highest marginal genetic contributions, top 100 with the highest total genetic contributions, and top 100 with highest contribution to inbreeding were chosen, which resulted in less than 200 top influential animals to be considered. Females were removed, as accessing DNA was highly unlikely. Remaining bulls were ranked based on their relationships with each other, to make sure that closely related animals will not be sequenced. As the approach described above does not identify younger bulls a secondary list of "young bulls" was created including bulls born after 2000 ranked based on number of progeny and descendants with phenotypic records. Top animals for each birth year till 2009 were chosen. The animals from the "young bulls" list (100-150 animals) were ranked based on relationships to make sure that they were not too closely related with each other and with older bulls chosen as described above. This resulted in the top 25 ancestors and top 5 young bulls with DNA available selected for sequencing from each breed.

In order to identify animals for genotyping, the top 3,000 animals with the highest genetic contributions and top 3,000 with the highest contributions to inbreeding were considered. For each breed, 400 ancestors (including 25 chosen for sequencing) will be genotyped with high density (649K) SNP panel and 560 with 50K SNP panel. They were chosen based on their contributions rankings and DNA availability. Additionally, for each breed 80 younger bulls (including 5 chosen for sequencing) will be genotyped with the HD panel.

A different approach was implemented when choosing Holstein animals, as at the time of analysis over 200,000 cows and bulls were already genotyped and over 40 were sequenced. To select the top 30 candidates for sequencing, an imputation analysis was carried out with a reference population comprised of 2,000 randomly selected animals genotyped with 50K panel. This ensured a large enough population to not be biased by a small number of sires, while still being computationally manageable for multiple imputations. Reference genotypes were filtered, randomly removing 5,000 SNP to mimic the imputation from a higher density panel to sequence more closely (45K to 50K). SNP were not removed for minor allele frequency (MAF), as the imputation of alleles with minor allele frequency will be critical in imputation to full sequence. Imputation of rare variants will be of the utmost importance to many sequence studies, as these variants have been linked to disease traits in other species (Cirulli and Goldstein 2010).

The top 200 bulls genotyped with 50K were selected as candidates based on their genetic contributions. Using genotyped animals only for this study helped to ensure that DNA was available for all animals chosen to be sequenced. Imputation was carried out iteratively, using FImpute 2.2 (Sargolzaei *et al.* 2011). First, a reference population of 35 already sequenced bulls, whose genotypes were available, was established. At this point, any bull who had a sire or maternal grand-sire sequenced was removed from the potential candidate group. Based on relatedness to the population, candidate bulls were added 10 at a time, starting with the animals with the highest relationship coefficients with the entire population. Accuracy of imputation for all SNP and for SNP with MAF <5% was calculated, then each of the 10 bulls was individually removed and accuracies were once again calculated. Any bull, when removed, who affected the accuracy of imputation, either for all SNP or for low MAF SNP by greater than 0.5% was included in the reference population, and was indicated to be sequenced. As the iteration was processed, groups of 10 animals were continually assembled with the remaining animals until all sires had been considered. Once all animals had been considered, the group with the greatest contributions to imputation accuracy were selected to complete the group of animals to be sequenced.

For genotyping, Holstein ancestors with the highest genetic and inbreeding contributions that have not already been genotyped and have DNA available were selected. Additionally, a high degree of relatedness to the entire population, and more importantly to the group of sequenced animals, was thought to be ideal. This will ensure high imputation accuracy from HD panel to sequence and help to accurately grow the database of sequenced individuals.

For synthetic populations, pedigree quality was not sufficient to perform analysis described above for beef breeds. Most influential animals were chosen based on the number of progeny with phenotypic records for the traits of interest.

RESULTS AND DISCUSSION

The results presented are for Angus (AN), Hereford (HE), Limousin (LM), Simmental (SM), and Holstein (HO) breeds. The data analysis revealed good pedigree quality for all breeds. Percentage of animals with both parents known varied from 85% for HO to 96% for AN. Discrete generation equivalent for animals born in 2011 was 11 for LM and SM, 12 for AN and HE, and 14 for HO. Pedigree completeness index was considered for 5 generations back and reached 99% for LM, 97% for HE and SM, 96% for AN, and 90% for HO. These results imply that choosing animals for sequencing and genotyping based on pedigree records was a reasonable approach.

The summary of the results obtained for the four breeds is presented in Table 1. The level of inbreeding for the reference population was considerably lower for beef breeds when compared to HO. However, similar rates of increase of inbreeding were observed for LM, SM and HO, which resulted in similar effective population size for those three breeds. The effective population size for AN and HE was significantly higher. Effective number of founders was the lowest for LM and the highest for SM, while effective number of founder genomes and effective number of ancestors were lowest for HO. This indicates that HO has a lower level of genetic variability when compared with beef breeds. This is further visible when looking at the number of ancestors needed to explain given percentage of gene pool. Six ancestors were needed to explain 50% of the gene pool in HO while 48 for HE. The 30 top contributing ancestors accounted for 41% of gene pool for HE, 43% for AN, 53% for SM, 61% for LM, and 83% for HO. The top 1,000 contributing ancestors explained 94% of gene pool for AN, 95% for HE and HO, 97% for SM, and 99% for LM. Therefore, genotyping those animals will provide very good coverage of the populations' gene pool and will help to ensure good quality imputation for use in developing genomic predictions.

Future developments in genetic evaluation methodology will capitalize on genomic sequence data to provide more accurate estimates of breeding values for selection. Imputation makes it possible to provide sequence data on many animals at a reasonable cost. Although the accuracy of imputation to sequence in the individual breeds is not known, the methods presented provide a means to prioritize animals for sequencing to ensure maximum coverage of the unique genome segments in each breed, which will maximize the imputation accuracy for a given level of investment.

Table 1. Summary of the results for the reference population

	AN	HE	LM	SM	HO
Total number of animals in pedigree	1,566,899	1,087,982	423,639	1,168,127	10,530,778
Number of animals in reference population	444,832	107,236	44,852	140,657	1,753,375
Pedigree completeness index (%)	96	97	99	97	90
Average inbreeding (%)	2.2	2.8	2.9	2.1	5.8
Average rate of increase of inbreeding (%/year)	0.02	0.02	0.11	0.10	0.11
Generation interval (years)	4.87	4.67	4.96	4.88	5.09
Effective population size	545	429	91	107	88
Effective number of founders	611	463	171	681	309
Effective number of founder genomes	52	48	22	35	8
Effective number of ancestors	103	101	47	69	16
No. of ancestors explaining 25% of gene pool	11	11	5	7	2
No. of ancestors explaining 50% of gene pool	46	48	18	26	6
No. of ancestors explaining 75% of gene pool	184	178	63	96	18
No. of ancestors explaining 90% of gene pool	596	543	190	293	63
No. of ancestors explaining 95% of gene pool	1,136	1,029	341	621	932
No. of ancestors explaining 100% of gene pool	8,730	7,645	2,607	8,179	>200,000
% of gene pool explained by 30 ancestors	43	41	61	53	83
% of gene pool explained by 1,000 ancestors	94	95	99	97	95

ACKNOWLEDGEMENTS

Research was made possible through funding from Genome Canada, Genome Alberta, Alberta Livestock and Meat Agency, the Canadian Simmental Association and Agriculture and Agri-Food Canada’s Growing Forward program. The cooperation of the Canadian Angus, Hereford, Limousin and Simmental associations as well as Holstein Canada and the Canadian Dairy Network is gratefully acknowledged.

REFERENCES

Boichard D., Maignel L. and Verrier E. (1997) *Genet. Sel. Evol.* **29**:5.

Boichard D. (2002) *Proc. 7th World Cong. Genet. Appl. Livest. Prod.* **28**:13.

Cirulli E.T. and Goldstein D.B. (2010) *Nat. Rev. Genet.* **11**:415.

Lacy R.C. (1989) *Zoo. Biol.* **14**:565.

MacCluer J.W., Boyce A.J., Dyke B., Weitkamp L.R., Pfennig D.W. and Parsons C.J. (1983) *J. Hered.* **74**:394.

Sargolzaei M., Iwaisaki H. and Colleau J.J. (2006) *Proc. 8th World Congr. Genet. Appl. Livest. Prod.* **27**:28.

Sargolzaei M., Chesnais J.P. and Schenkel F.S. (2011) *J. Dairy Sci. 94, E-Suppl.* **1**:421.

Sölkner J., Filipcic L. and Hampshire N. (1998) *Anim. Sci.* **67**:249.