Genome Wide Imputation in Canadian Beef Cattle

F. S. Schenkel[1]
[1]Centre for Genetic Improvement of Livestock, University of Guelph, Canada

ABSTRACT: Low-density SNP chip panels are appealing tools for reduction of genotyping costs. Imputation enables to predict missing genotypes to recreate the coverage of high density panels and it is a tool in genomic selection by allowing for more animals to be genomically evaluated and for larger training datasets. In addition, imputation could also increase power of genome-wide association studies. Several studies have been carried out in Canada, providing practical direction on the implementation of imputation strategies in dairy and beef cattle, including crossbred beef cattle. A large nation-wide project has created the core reference populations of 50k, high density, and sequence genotypes to enable accurate imputation from low density panels in the major beef breeds and composites in Canada. These reference populations associated with the developed imputation methods and pipelines will create the foundation for genomic selection through genome wide imputation in beef cattle.
Keywords: imputation; cattle; SNP panels; sequence

Introduction

Imputation enables the determination of SNP genotypes that have not been directly genotyped by a low density panel by inferring missing genotypes using information from a reference population genotyped with a higher density panel (Servin and Stephens (2007), Li et al. (2009), Hickey et al. (2011)). Imputation methods can be used as a tool in genomic selection, allowing for more animals to be genomically evaluated and for larger training datasets. In addition, imputation could also increase power of genome-wide association studies by allowing more individuals to be genotyped (Li et al. (2009)).

Imputation methods differ with respect to the use of information on the relationship between individuals (family information) and/or the use of linkage disequilibrium (LD) among markers without knowledge of relationships (population information) (Hayes et al. (2012)). The application of imputation algorithms that use population information has been reported in several papers, such as Calus et al. (2011) (Beagle and fastPhase), Howie et al. (2012) (Impute2), Zhang et al. (2010) (DagPhase) among others, while research on imputation algorithms using family and population information has been reported, for instance, by Sargolzaei et al. (2010) (FImpute), Nicolazzi et al. (2013) (PedImpute), and VanRaden et al. (2013) (FImpute and findHap). Although various imputation studies have been completed in dairy cattle, relatively fewer studies have been conducted in different beef cattle breeds and crossbreds (e.g. Berry et al. (2013)).

Several studies have been carried out in Canada aiming to provide practical direction on the implementation of imputation strategies in dairy and beef cattle, including crossbred beef cattle populations. Low density commercial SNP chip panels from different densities were evaluated to impute higher density panels, using alternative algorithms programed into available software packages.

More recently, the availability of whole-genome sequence data could potentially improve the accuracy of genomic predictions by capturing the causal mutations affecting a trait, without dependence on the extent of LD between SNP markers and causal mutations (Meuwissen and Goddard (2010), Druet et al. (2013)). A cost effective sequencing strategy is to sequence key ancestors, who contributed most of the genetics in the current population, combined with a random sample of individuals and, then, impute sequence data to the rest of the animals genotyped with commercial SNP chips. The 1,000 Bull Genomes Project (http://www.1000bullgenomes.com/) aids the goal of using whole-genome sequence genotypes by providing an extended cattle sequence database of key ancestors from several breeds.

The possible increase in accuracy of genomic predictions and fine mapping of causative QTL/QTN by using imputed sequence data is dependent on the accuracy of imputation and, therefore, it is important to evaluate imputation from SNP chip panels to sequence genotypes both in terms of accuracy and computational efficiency. In Canada studies on imputation from 50k and HD SNP panels to sequence genotypes have been carried out to assess accuracy and computing efficiency of alternate methods.

This short paper will present and discuss some of the studies on genome wide imputation in Canadian cattle, with emphasis on beef cattle.

Genome wide imputation

Imputation algorithm. Genotype imputation can help reduce genotyping costs particularly for implementation of genomic selection and fine map of causative QTL/QTN. In applications involving large populations, inferring the genotypes of ungenotyped loci using information from reference individuals that were genotyped with a higher density panel is computationally demanding. The most popular imputation methods are based upon Hidden Markov models and have computational limitations due to an intensive sampling process (e.g., Beagle, fasPhase, Impute2). A fast deterministic approach was developed in Canada (Sargolzaei et al. (2011a)) and was recently described in Sargolzaei et al. (2014), which makes use of both family and population information. The algorithm assumes that all individuals are related and, therefore, share haplotypes which may differ in length and frequency based on their relationships. The method starts with family imputation if pedigree information is available, and then exploits close relationships by searching for long haplotype matches in the reference group using overlapping sliding windows. The search continues as the window size is shrunk in each chromosome sweep in order to capture more distant relationships (Sargozaei et al. (2014)).

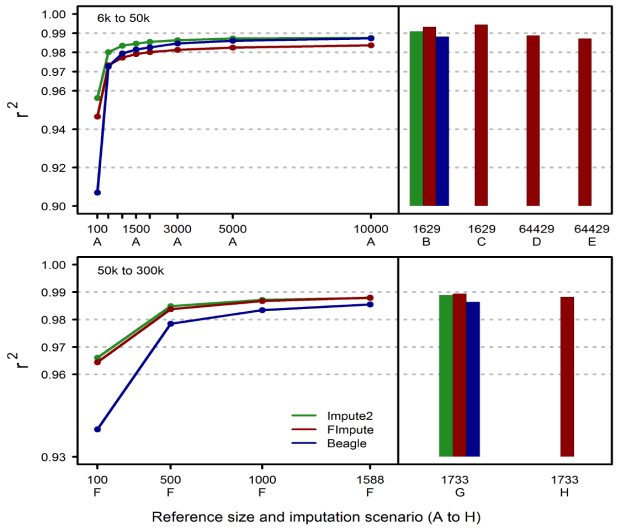

Figure 1. Overall allelic r2 for FImpute, Beagle and Impute2 across different imputation scenarios. There were 2000 and 500 young target individuals for imputation from 6k to 50k and from 50k to 300k, respectively. In scenarios A and F, reference groups with different sizes were randomly chosen after excluding parents and grandparents. The reference group in scenarios B and C included only parents and grandparents, in scenarios D and E it included all genotyped males and in scenarios G and H it included all genotyped individuals. Pedigree information was considered in scenarios C, E and H and was disregarded in scenarios B, C and G. (Adapted from Sargolzaei et al. (2014))

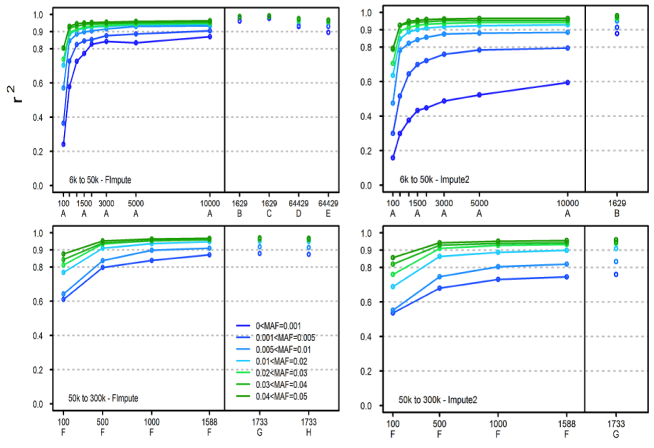

Figure 2. Rare allele imputation: allelic r^2 in different MAF bins for FImpute and Impute2. Scenarios as described in Figure 1. (Adapted from Sargolzaei et al. (2014)).

This algorithm was programed in the FImpute software (Sargolzaei et al. (2011a)). Sargolzaei et al. (2014) showed that FImpute gave higher or similar imputation accuracy than Beagle 3.3.2 (Browning and Browning (2009)) and Impute2.3 (Howie et al. (2011)) in Holstein cattle data sets when all available information was used. When close relatives of target individuals were present in the reference group, the method resulted in higher accuracy compared to the other two methods even when the pedigree was not used (Figure 1). Rare variants were also imputed with higher accuracy, especially when compared to Impute2 (Figure 2- results for FImpute and Impute2). Beagle showed intermediate values between these two software. Computing requirements for FImpute were only a small fraction of those of Beagle and Impute2 (Figure 3).

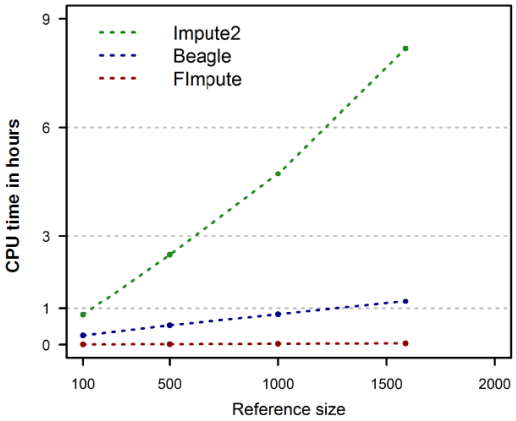

Figure 3. CPU time for Beagle, Impute2 and FImpute over different reference sizes. No pedigree information was used and genotyped parents and grandparents were excluded.

FImpute is currently used for imputation from low density panels to 50k SNP genotypes in routine genomic evaluations of dairy cattle in Canada by the Canadian Dairy Network (www.cdn.ca).

SNP chip genotypes. In Canada, application of low density genotyping started with the Illumina GoldenGate Bovine3k BeadChip SNP panel (3k) (Illumina Inc., San Diego, CA, USA) in dairy cattle, mainly for genotyping cows and heifers, but also for some pre-screening of young bulls. Therefore, early research and development of imputation technology in Canada started with the 3k panel, aiming to impute the genotypes from the Illumina Bovine SNP50 BeadChip SNP panel (50k) (Illumina Inc., San Diego, CA, USA). Because of the low density of the 3k panel, accuracy of imputation was dependent on whether or not parents were genotyped (Sargolzaei et al. (2010a,b)). The average concordance rate (CR) was very high (>0.98) only when both parents were genotyped with the 50k panel, but still few animals had low accuracy of imputation. The use of family + population imputation led to higher CR (>0.90) for those animals with low accuracy, but this was still dependent on both parents being genotyped with 50k panel (Sargolzaei et al. (2011b)). Ventura et al. (2011) showed that for crossbred beef cattle, the 3k panel would lead to very low imputation accuracy to the 50k panel in crossbred beef cattle (Figure 4).

With the release of the Illumina Bovine 6k BeadChip panel (6k) (Illumina Inc., San Diego, CA, USA), imputation studies in Canada focused on the imputation from this panel to the 50k panel in both dairy and beef cattle. The 6k panel resulted in substantially higher accuracy for animals with low family information, especially for those with both parents missing or

ungenotyped (Sargolzaei et al. (2011b); Ventura et al. (2014)).

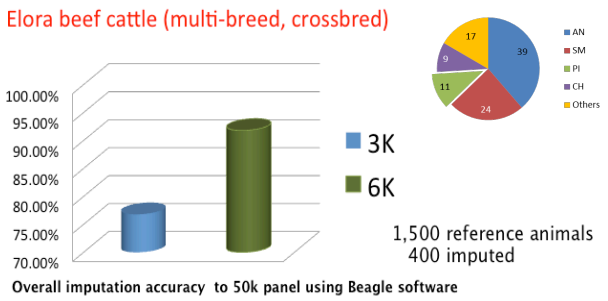

Figure 4. Accuracy of imputation (concordance rate) of Illumina 50k genotypes from alternate low density genotypes (Illumina 3k or 6k) in crossbred beef cattle using Beagle software. (Adapted from Ventura et al. (2011))

A comprehensive study on imputation from 6k to 50k panel genotypes in beef cattle was conducted by Ventura et al. (2014). This study showed that imputation from 6k to 50k SNP panel could be successfully applied in beef cattle. In this study population imputation was implemented using Beagle 3.3.2, FImpute 2.2.2 and Impute2.2 in a multi-breed, crossbred taurine beef cattle population genotyped with the 50k panel. Different combinations of reference populations and imputed animals were defined based on breed composition. Number of animals (n=250 to 4,932) and the presence of closer relatives in the reference population (only for Angus animals) were investigated. The overall average imputation accuracy for purebred animals ranged from 94.2 to 97.9% using FImpute, from 95.4 to 98.3% using Impute2 and from 90.0 to 96.4% when Beagle was used. The individual imputation accuracy of crossbred animals widely ranged from 54.2 to 97.5% (FImpute), from 57.0 to 97.5% (Impute2) and from 54.4 to 95.6% (Beagle). For accurate and less variable imputation accuracy of crossbred animals, a large reference population, which represent well the breed composition of imputed animals, was required (Figure 5). Within breed imputation from 6K to 50K did not improve when additional purebred breeds were added to the reference population. FImpute reduced the run-time by 13 to 52 times compared to Beagle and 51 to 108 times compared to Impute2.

The use of alternate low density panels, such as the 8k GGP panel (Neogen Corporation, Lansing, MI, USA), was also investigated and showed some advantage in terms of increased accuracy for imputation of crossbred animals (Figure 6). Mullen et al. (2013) reported increased accuracy of imputation using a customized low density panel (International Dairy-Beef 19 (IDB19)) compared to the 6k panel for imputing HD panel genotypes in cattle breeds.

Accuracy of imputation from lower density SNP panels (6K or 50K) to Illumina Bovine HD Beadchip (HD) (Illumina Inc. San Diego, CA, USA) genotypes was examined both within breed and using a multi-breed reference population in Holstein, Ayrshire, and Guernsey by Larmer et al. (2014). Imputation was carried out using FImpute V2.2 and Beagle 3.3.2 software. Imputation accuracies were calculated as both concordance rate and allelic r^2. Computation time was also explored to determine the efficiency of the different algorithms for imputation. High CR (0.968–0.995) and allelic r^2 (0.946–0.991) were found for all breeds when imputation was carried out with FImpute from 50K to HD. Imputation accuracy for Guernsey and Ayrshire was slightly lower when using the imputation by Beagle. Computing time was significantly greater when using Beagle software, with all comparable procedures being 9 to 13 times less efficient, in terms of computing time, compared to FImpute. These findings suggested that the HD genotypes can be efficiently and effectively imputed using the lower density 50K SNP panel in cattle.

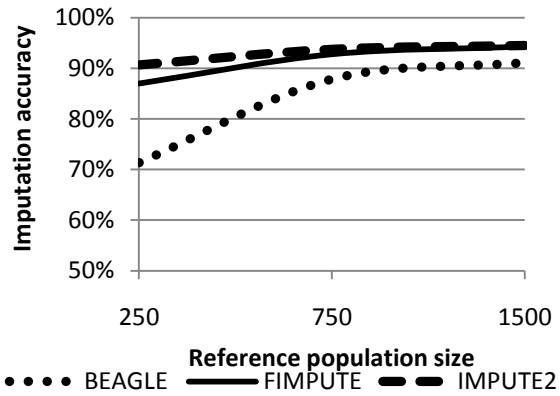

Figure 5. Effect of reference dataset size on the average accuracy (concordance rate) from Illumina 6k to 50k genotypes in crossbred animals using Beagle, FImpute and Impute2. (Adapted from Ventura et al. (2014))

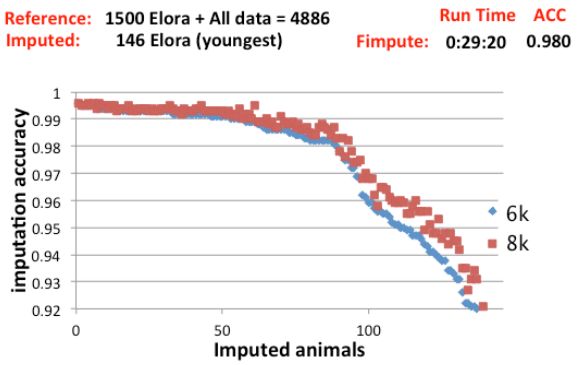

Figure 6. Accuracy of imputation (concordance rate) of Illumina 50k genotypes from alternate low density genotypes (Illumina 6k and GGP 8k) in crossbred beef cattle using FImpute software. (Dr. R. Ventura, personal communication).

Recently Berry et al. (2014) investigated imputation of un-genotyped parental genotypes in dairy and beef cattle from progeny genotypes using FImpute 2.2. This might be useful for, for instance, imputing genotypes of influential females with no available biological material for sampling DNA. Two separate datasets were used, one containing both dairy and beef animals (n=3,122) with high

density genotypes (~735k SNPs) and the other containing just dairy animals (n=5,489) with medium density genotypes (~52K SNPs). Imputation accuracy, based on CR, of three different genotype density panels was evaluated representing low (6k), medium (50k), and high (HD) density. When genotypes were not available on individual animals, but at least five progeny were genotyped (on either 50k or HD panel) the parental alleles were imputed with on average ≥96% accuracy. The accuracy of imputing parental genotypes from genotyped half-sib progeny groups was on average 98% when 12 genotyped half-sib progeny were available. Hence, genotypes of descendants may be used for imputing the genotypes of ancestral animals with phenotypes with reasonably high accuracy.

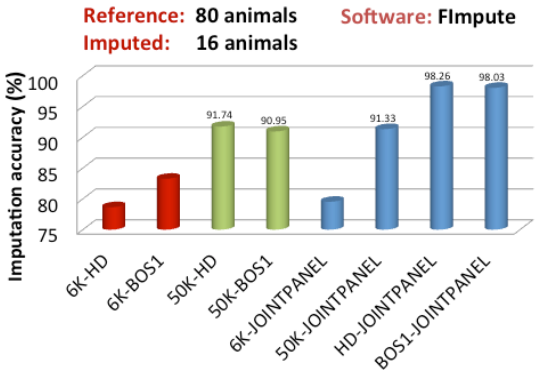

Figure 7. Accuracy of imputation (concordance rate) of high density genotypes (Illumina (HD), Affymetrix (BOS 1), and Join Panel- HD+BOS 1) from alternate low density genotypes in crossbred beef cattle. (Adapted from Ventura et al. (2013))

VanRaden et al. (2013) reported an accuracy of 93.5% and 95.1% for imputation of HD genotypes of ungenotyped dams that had 4 or more genotyped progeny in Holsteins. Accuracy further improved with imputation first to 50K and then to HD instead of all together (1.6 percentage points).

Ventura et al. (2013) looked at the imputation from 50k panel to Illumina HD panel or Affymetrix Axiom® Genome-Wide BOS 1 SNP panel (Affymetrix, Santa Clara, CA, USA) and to the join Illumina+Affymetrix panels, using 96 crossbred beef cattle and FImpute. Concordance rate was quite similar from 50k to HD, Bos1 or the join HD+Bos1 panel (ranging from 91.0% - 91.7%) (Figure 7). Accuracy of imputation of the join HD+Bos1 panel from either of the two high density panels was very high and similar (98.0% – 98.3%).

Accurate imputation is a key to ensuring that the benefits from more markers exceed the imputation loss because gains from HD are small (VanRaden et al. (2013))

Sequence genotypes. Li et al. (2014) used the most recent run of the 1,000 Bull Genomes Project (run 3), which included 429 full genome sequences of 427 bulls and 2 cows from 15 breeds, sequenced at an average of 10.1 fold coverage, to evaluate accuracy of imputation from 50k and HD to sequence genotypes. There were 30.8 million filtered sequence variants detected, including 29.1 million SNPs and 1.7 million insertion-deletions. The sequence data across all breeds included 28,336,153 SNP on autosomal chromosomes. Six breeds with more than or equal to 25 sequenced animals were used in the study, which included Angus, Brown Swiss, Holstein, Jersey, Limousin, and Simmental.

Accuracy of genotype imputation from 50k or HD to whole-genome sequence genotypes was investigated. Different scenarios were evaluated by masking sequence genotypes to mimic animals genotyped with HD and 50k SNP chips, resulting in 658k and 47k SNPs in the mimicked HD and 50k SNP chips, respectively. FImpute 2.2 and Beagle 3.3.2 programs were used and their performances in terms of imputation accuracy and computational efficiency were compared. Accuracy of imputation was assessed by both CR and allelic r^2. In order to evaluate the effect of increasing the number of reference animals on the accuracy of sequence genotype imputation, a combined multi-breed reference population was used to impute each breed using FImpute.

Due to the high computing time required for Beagle, FImpute and Beagle were only compared using a single cross-validation, randomly selecting a validation set corresponding to about 20% of the sequenced animals in a breed. However, all the imputation scenarios were cross-validated by 5-fold cross-validation using FImpute only.

Table 1: Single and multi-breed sequence imputation from HD or 50k SNP chip using FImpute, SNPs with MAF>0, and 5-fold cross-validation. (Adapted from Li et al. (2014))

Breed[1]	HD[2]		50k[2]	
	CR	r^2	CR	r^2
Single Breed reference population				
AN	92.7	89.2	86.9	80.2
BS	92.7	88.9	84.5	76.2
HO	94.8	92.6	90.0	85.5
JE	92.0	87.0	83.6	73.2
LI	86.4	79.7	75.3	62.9
SI	92.4	89.4	83.9	77.0
Mean	*91.8*	*87.8*	*84.0*	*75.8*
Multi-breed reference population				
AN	93.3	90.2	86.1	78.9
BS	93.6	90.4	83.7	75.0
HO	94.9	92.7	89.2	84.4
JE	93.2	89.2	82.3	70.9
LI	89.7	85.3	76.7	65.4
SI	93.0	90.4	83.6	76.7
Mean	*92.9*	*89.7*	*83.6*	*75.2*

[1]Breed= Angus (AN), Brown Swiss (BS), Holstein (HO), Jersey (JE), Limousin (LI), and Simmental (SI)
[2]CR= Concordance Rate; r^2= allelic r^2

As expected, imputation from the HD SNP chip to sequence genotypes was substantially more accurate by 7.8-9.3 percentage points in CR and 12-14.5 percentage points in allelic r^2 than from the 50k SNP chip (Table 1, for FImpute). For Beagle, the differences were even more pronounced, being about twice that for FImpute. FImpute was slightly more accurate than Beagle for imputation from HD by 1.6 percent points of CR and 3.1 points of allelic r^2.

For sequence genotype imputation from 50k, however, the same features were much larger reaching 9.6 and 19.6 percent points, respectively (Li et al. (2014)).

Beagle required much longer computing time than FImpute. In addition, FImpute required less memory, about 5 Gigabyte per chromosome for imputation from both HD and 50k panel, while Beagle took more than 80 Gigabyte of memory per chromosome. As an example, the CPU time for HD to sequence imputation in Angus cattle with 44 reference animals and 10 validation animals took more than 14 hours using Beagle, while it took about 2 hours using FImpute without parallel computation (Li et al. (2014)).

Results for single-breed and multi-breed reference populations (Table 1) showed an increase in imputation accuracy using multi-breed reference population (1.9 and 2.1 percent points of CR and allelic r^2) for imputing from HD genotypes. The observed increase was even more substantial for breeds with the lowest reference sets. Imputation accuracy from 50k genotypes, however slightly decreased (-0.4 and -0.6 percent points of CR and allelic r^2), which indicates that the 50k panel was not dense enough to capture small haplotypes shared among the breeds. HD to sequence imputation using either FImpute or Beagle resulted in higher and more balanced CR and allelic r^2 across breeds than 50k chip (Table 1).

In Canada a nation-wide project entitled "Whole genome selection through genome wide imputation in beef cattle" began in 2011 with the goal of developing low cost genome wide selection methodologies for Canada's cattle industry. The project sequenced 378 key ancestors that have significant genetic influence on the current Canadian herds from Angus, Charolais, Hereford, Limousin, Simmental, Gelbvieh and Holstein breeds, and Beef Booster, Alberta composite, and Guelph composite cattle (minimum of 30 ancestors per breed/composite); and genotyped about 5,000 key animals on high density panel (with a minimum target of 500 per breed/composite) and about 5,000 animals on 50K SNP panel. Adding genotypes from international collaborators and co-funding in the project, a total of about 12,000 high density genotypes (using both Illumina and Affymetrix panels) and 19,000 50k genotypes will be available for imputation purposes. In collaboration with the 1000 bull genomes project, over 450 genome sequences will be used for imputing 50k animals to high density and then to sequence genotypes. The high accuracy of imputation from 50k to high density and from high density to sequence genotypes found in the research carried out so far is encouraging and will provide foundation for genome selection through genome wide imputation in beef cattle in Canada.

Conclusion

Several studies have been carried out in Canada, providing practical direction on the implementation of imputation strategies in dairy and beef cattle, including crossbred beef cattle. A large nation-wide project has created core reference populations of 50k, high density, and sequence genotypes to enable accurate imputation from low density panels in the major beef breeds and composites in Canada. These reference populations associated with the developed imputation methods and pipelines will create the foundation for genomic selection through genome wide imputation in beef cattle in Canada.

Acknowledgments

The author thanks the agencies, industry partners and collaborators who have provided support for imputation research in Canada. Their contributions are acknowledged in the several papers cited in this short paper.

Literature Cited

Berry, D. P., McClure, M. C., Mullen, M. P. (2013). J. Anim. Breed. Genet. (2013)1–8. DOI: 10.1111/jbg.12067

Berry, D. P., McParland, S., Kearney, J.F. et al. (2014) Animal. (in press)

Browning B, and Browning S. (2009). Am. J. Hum. Genet. 84:210-223.

Calus, M. P. L., Veerkamp, R. F., and Mulder, H. A. (2011). J. Animal Sc., 89: 2042-2049.

De Pauw, M. A., Stothard, P., Miller, S. P. et al. (2014). Proc XXII International Plant & Animal Genome, January 10-14, 2014, San Diego, CA, USA.

Druet, T., Macleod, I. M., and Hayes, B. J. (2013). Heredity 112:39-47.

Hayes, B. J., Bowman, P. J., Daetwyler, H. D. at al. (2012). Anim. Genet. 43: 72-80.

Hickey, J. B., Cleveland, M., Gorjanc, G. et al. (2011). Interbull Open Meeting. Stavanger, Norway. Interbull Bull. No. 44:38-41.

Howie, B., Marchini, J., and Stephens, M. (2011). G3: Genes, Genomics, Genetics 1: 457-470.

Howie, B., Fuchsberger, C., Stephens, M. et al. (2012). Nature Genetics 44(8): 955-959

Larmer, S. G., Sargolzaei, M., and Schenkel, F. S. (2014). J. Dairy Sci. DOI:10.3168/jds.2013-6826.

Li, Y., Willer, C., Sanna, S. et al. (2009). Annu Rev Genomics Hum Genet. 10: 387–406.

Li, H., Sargolzaei, M., and Schenkel, F. (2014). Proc 10th WCGALP, August 17-22, 2014, Vancouver, AB, Canada (submitted)

Meuwissen, T., and Goddard, M. E. (2010). Genetics 185: 623–631.

Mullen, M. P., McClure, M. C., Waters, S. M. et al. (2013). Proc Agricultural Research Forum, March 11-12, 2013, Tullamore, Ireland, p 56.

Nicolazzi, E. L., Biffani S., and Jansen G. (2013). J. Dairy Sci. 96: 2649-2653.

Sargolzaei, M., Schenkel, F., and Chesnais, J. (2010a). Technical report to the Dairy Cattle Breeding and Genetics Committee. October 5, 2010, University of Guelph, Guelph, ON, Canada. http://cgil.uoguelph.ca/dcbgc/Agenda1010/DCBGC1010_msargol_r1.pdf Accessed on Mar. 2014.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F.S. (2010b). Technical report presented at Canadian Dairy Network Open Industry Session. October 26, 2010, Guelph, Ontario,Canada. www.cdn.ca/Articles/GEBOCT2010/OIS_msargol.pdf Accessed on Mar. 2014.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2011a). J. Anim. Sci. 89, E-Suppl. 1 / J. Dairy Sci. 94, E-Suppl. 1: 421 (333).

Sargolzaei, M., Schenkel, F., and Chesnais, J. (2011b). Technical report to the Genetic Evaluation Board of Canadian Dairy Network in October, 2011. www.cdn.ca/Articles/GEBOCT2011/4_msargol_1109 .pdf Accessed on Mar. 2014.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2014). BMC Genomics (submitted)

Servin, B., and Stephens, M. (2007). PLoS Genet. 3: e114.

VanRaden, P. M., Null, D. J., Sargolzaei, M. et al. (2013). J. Dairy Sci. 96: 668-678.

Ventura, R.V., Schenkel, F.S., Wang, Z. et al. (2011). Proc 2nd Annual Conference of Livestock Gentec, October 19-20, 2011, Edmonton, AB, Canada.

Ventura, R., Schenkel, F., Sargolzaei, M. et al. (2013). Proc XXI International Plant & Animal Genome, January 11-16, 2013, San Diego, CA, USA.

Ventura R.V., Lu D., Schenkel F. S., et al. (2014). J. Animal Sc. (in press)

Zhang, Z., and Druet, T. (2010). J. Dairy Sci. 93: 5487-5494.