# Parentage Analysis Services for Dairy Cattle in Canada

***B.J. Van Doormaal, G.J. Kistemaker and J. Johnston***
*Canadian Dairy Network (CDN), Guelph, ON, Canada, N1K 1E5*

## Abstract

Since the implementation of genomic evaluations in 2009, Canadian Dairy Network (CDN) has used SNP genotypes to verify the reported parents if genotyped and, when missing or incorrect, to discover the animal's sire and/or dam based on all existing SNP genotypes. To date, for breeds with official genomic evaluations in Canada, CDN has over 1.4M genotypes including 1.2M Holstein, 163,000 Jersey, 29,000 Brown Swiss, 6,000 Ayrshire and 3,000 Guernsey. These genotypes involve 23 different genotype panels, including low (3K-30K), medium (44K-140K) and high (over 600K) density. For parentage analysis, a list of 2,683 SNP in common from the 3K and 50K genotype panels are used as the basis for parentage verification, parentage discovery and for identifying families of genetically identical animals. Using the list of SNP proposed for inclusion in GenoEx-PSE for parentage verification (200) and parentage discovery (additional 675 or 354), it was concluded that the 200 SNP recommended by ISAG for parentage verification performed very well compared to the SNP routinely used by CDN for dairy cattle breeds in Canada. It was also concluded that parentage discovery using either set of additional 675 or 354 SNP also provided accurate results. To avoid a possible misuse of the additional SNP for parentage discovery, the reduced set of 354 SNP, selected from only 10 chromosome, are recommended for GenoEx-PSE due to the higher level of imputation error and lower accuracy of GEBV estimation compared to results based on the additional 675 SNP.

**Key words:** SNP parentage analysis, parentage verification, parentage discovery, GenoEx-PSE

## Introduction

The International Committee of Animal Recording (ICAR) plans to introduce a new service effective early 2017, referred to as GenoEx-PSE (**P**arentage **S**NP **E**xchange). This service will be offered through the Interbull Centre and aims to facilitate the international exchange of SNP genotypes for the purpose of parentage analysis in dairy and beef cattle populations.

Within the context of GenoEx-PSE, parentage analysis consists of three principle processes:

- Parentage Verification - The process by which the SNP genotypes of the recorded parents (sire and/or dam) of an animal are examined relative to the SNP genotype of the animal to determine if one or other does not qualify as a parent.
- Parentage Discovery - The process by which a set of SNPs from an animal's genotype are compared to a database of SNP genotypes for older animals in order to identify the most likely sire and/or dam, if not already confirmed by Parentage Verification.
- Microsatellite Imputation - The process by which the microsatellite (STR) profiles of an animal may be imputed from SNP genotypes and the resulting SNP genotype is used for the purpose of parentage verification.

A fundamental component of the service is the establishment of a standard set of SNP to be included in the exchange process. The GenoEx-PSE service agreement offers flexibility for the service user to define its level of participation, which then defines the list of SNP to be routinely uploaded to the GenoEx-PSE database at the Interbull Centre and therefore also the list of SNP that are accessible for downloading. The basic level of participation includes the exchange of the 200 SNP (Group A) recommended by the International Society for Animal Genetics (ISAG) for parentage verification in cattle. A second level of service includes the exchange of additional SNP (Group B) to provide more accurate results for parentage discovery. The third level of participation includes another group of SNP (Group C) required for the imputation of microsatellite profiles from the SNP genotypes.

Prior to the official implementation of GenoEx-PSE, ICAR and the Interbull Centre published the proposed list of SNP to be included in each of Groups A, B or C. Interested countries were encouraged to assess these SNP in terms of application for parentage analysis within their national breed populations. Therefore, the objectives of this research include: (a) quantify the minor allele frequency (MAF) of the SNP in Groups A and B within each of the five dairy breed populations with genomic evaluations in Canada, (b) compare official parentage analysis results at Canadian Dairy Network (CDN) with those that would have resulted based on the SNP in either Groups A or B, and (c) examine the accuracy of imputation from SNP profiles based on Groups A and B to 50K genotypes and the accuracy of genomic breeding value (GEBV) estimation based on these GenoEx-PSE SNP.

## Data and Methods

Given the SNP recommended by the International Society for Animal Genetics (ISAG) for parentage verification have been widely accepted and used by various organizations, GenoEx-PSE defined the same 200 SNP within Group A. The distribution of these SNP across chromosome 1 to 29 is presented in Table 1 and shows a range from 4 to 11 on all these chromosome. For the Group B SNP, an initial proposal for GenoEx-PSE included a total of 675 SNP with 75 defined by work in Australia and 600 based on research in United States and Ireland (McClure, 2015).

Later, the list of SNP in Group B was reduced by limiting the initial set defined by McClure to include only those from the ten chromosome that had the most, specifically chromosome 1, 2, 3, 5, 7, 8, 11, 13, 19 and 21. These two SNP groups are labelled as B-675 and B-354, respectively (Table 1). For Group C, a list of 980 SNP were selected (McClure, 2015) that are densely located near the 12 short-tandem-repeat (STR) on chromosome 1, 2, 3, 5, 9, 15, 16, 18, 19, 20, 21 and 23 that have been used for microsatellite parentage verification globally for many years (Table 1).

**Table 1.** SNP count by chromosome (Chr) for each group in GenoEx-PSE.

| Chr | Group* | | | | Total (675) | Total (354) |
|---|---|---|---|---|---|---|
| | A | B-675 | B-354 | C | | |
| 1 | 11 | 40 | 40 | 70 | 121 | 121 |
| 2 | 9 | 42 | 42 | 80 | 131 | 131 |
| 3 | 8 | 32 | 32 | 110 | 150 | 150 |
| 4 | 9 | 22 | 4 | | 31 | 13 |
| 5 | 9 | 33 | 33 | 90 | 132 | 132 |
| 6 | 8 | 17 | 2 | | 25 | 10 |
| 7 | 11 | 26 | 26 | | 37 | 37 |
| 8 | 6 | 31 | 31 | | 37 | 37 |
| 9 | 6 | 19 | 3 | 110 | 135 | 119 |
| 10 | 9 | 24 | 2 | | 33 | 11 |
| 11 | 10 | 27 | 27 | | 37 | 37 |
| 12 | 5 | 21 | 3 | | 26 | 8 |
| 13 | 6 | 30 | 30 | | 36 | 36 |
| 14 | 5 | 20 | 1 | | 25 | 6 |
| 15 | 7 | 14 | 3 | 40 | 61 | 50 |
| 16 | 5 | 21 | 1 | 80 | 106 | 86 |
| 17 | 9 | 24 | 2 | | 33 | 11 |
| 18 | 7 | 23 | 2 | 40 | 70 | 49 |
| 19 | 8 | 27 | 27 | 80 | 115 | 115 |
| 20 | 5 | 16 | 5 | 120 | 141 | 130 |
| 21 | 8 | 25 | 25 | 80 | 113 | 113 |
| 22 | 7 | 20 | 1 | | 27 | 8 |
| 23 | 4 | 16 | 2 | 80 | 100 | 86 |
| 24 | 4 | 24 | 2 | | 28 | 6 |
| 25 | 4 | 18 | 2 | | 22 | 6 |
| 26 | 4 | 17 | 2 | | 21 | 6 |
| 27 | 6 | 14 | 1 | | 20 | 7 |
| 28 | 6 | 16 | 1 | | 22 | 7 |
| 29 | 4 | 16 | 2 | | 20 | 6 |
| **Total** | **200** | **675** | **354** | **980** | **1,855** | **1,534** |

** - Group A = SNP for parentage verification, B = SNP added for parentage discovery and C = SNP added for microsatellite imputation.*

### *Minor Allele Frequency*

The value of a SNP for parentage analysis is highly dependent upon its minor allele frequency (MAF) within the population of animals in question. SNP with low MAF will have more limited value for parentage analysis since only homozygous genotypes in both parent and progeny are informative. The SNP defined in Groups A and B for GenoEx-PSE

were selected based on MAF within several dairy and beef cattle breeds with genotypes at the Irish Cattle Breeding Federation (ICBF) in Ireland. Given that CDN offers genomic evaluation services in the Ayrshire, Brown Swiss, Guernsey, Holstein and Jersey breeds in Canada, all genotypes available for each of these breeds were used to estimate the MAF of each SNP proposed for Group A and Group B-675. Since Group B-354 is a subset of those included in B-675, the analysis was not repeated based solely on this reduced SNP. This analysis serves as a general indication of how well the SNP could perform for parentage analysis.

### *Comparison of Parentage Analysis Results*

As part of the routine processing at CDN to produce Canadian genomic evaluations for each breed, CDN has implemented an internal process to carry out parentage verification as well as parentage discovery, when needed. The genotypes at CDN currently involve 23 different genotype panels, including low (3K-30K), medium (44K-140K) and high (over 600K) density. For parentage analysis, a list of 2,683 SNP in common from the 3K and Illumina 50K genotype panels are used as the basis for parentage verification, parentage discovery and for identifying families of genetically identical animals. In order to compare results for parentage verification based on the 2,683 SNP routinely used by CDN versus either (a) the 200 ISAG SNP in Group A or (b) the combination of Group A and Group B-675 SNP, the genotypes of 5,372 Ayrshire, 23,144 Brown Swiss, 1,771 Guernsey, 573,988 Holstein and 63,248 Jersey animals were included in each analysis.

To define a SNP parentage conflict within the routine process at CDN, a maximum limit of 2% conflicts among informative SNP is allowed prior to excluding the recorded parent. For the analysis using the Group A SNP, the published guidelines established by ISAG for parentage verification analysis in cattle (ISAG, 2012) were applied. For the Group B-675 SNP some rules based on number of conflicts, similar to the ISAG guidelines for the 200 SNP in Group A, were established and applied. To compare the results across the three sets of SNP, the process was applied only for conducting parentage verification of the recorded sire and the cases that led to differing results were assessed in more detail.

To compare results of parentage discovery between the routine CDN process and one that would be limited to include Group B-675 or Group B-354, in addition to Group A, known parentage information was removed for a group of 26,691 animals across the five breeds that were born in recent years. In this manner, the three sets of SNP applied for parentage discovery can be compared to each other.

### *Accuracy of Imputation and Estimation of Genomic Breeding Values*

The exchange of SNP genotypes via GenoEx-PSE is restricted for the use of parentage analysis only. Nonetheless, it is valuable to verify the level of accuracy that could be achieved for genomic prediction if the exchanged SNP were used in such a manner by any service user of GenoEx-PSE. To assess this objective the 50K genotypes for a group of 27,324 Holsteins born in 2016 were reduced to include either (a) 200 ISAG SNP in Group A, (b) Group A and Group B-675 SNP, or (c) Group A and Group B-354 SNP. The FImpute software (Sargolzaei, 2014) used routinely at CDN for imputation of all genotypes to 50K genotypes was applied to the reduced genotypes of the Holsteins selected for analysis. After imputation, the routine genomic evaluation system at CDN, based on GBLUP methodology (VanRaden, 2008), was applied to all genotyped Holsteins to estimate Direct Genomic Values (DGV) for each of the selected animals born in 2016.

Imputation error was quantified by comparing each animal's imputed 50K genotype based on each of the three subsets of SNP available for imputation to the 50K genotype used for the routine evaluation at CDN. Similarly, simple correlations among resulting GEBV were used to assess the accuracy of genomic breeding value (GEBV) estimation based on each subset of SNP considered for GenoEx-PSE.

## Results and Discussion

### *Minor Allele Frequency (MAF)*

Table 2 shows the average MAF of the SNP included in Groups A (n=200) and B-675 (n=675) as proposed for exchange within GenoEx-PSE for parentage verification and discovery, respectively.

**Table 2.** Average minor allele frequency of GenoEx-PSE SNP groups.

| Breed | Group A | Group B-675 |
|---|---|---|
| Ayrshire | 0.342 | 0.343 |
| Brown Swiss | 0.352 | 0.353 |
| Guernsey | 0.322 | 0.322 |
| Holstein | 0.391 | 0.381 |
| Jersey | 0.291 | 0.316 |

Table 3 provides the percentage of the SNP in each group that have a MAF of at least .30 within each of the five dairy cattle breeds with genomic evaluations in Canada.

**Table 3.** Percentage of GenoEx-PSE SNP with minor allele frequency ≥.30.

| Breed | Group A | Group B-675 |
|---|---|---|
| Ayrshire | 66.0% | 69.4% |
| Brown Swiss | 70.0% | 73.2% |
| Guernsey | 60.5% | 62.8% |
| Holstein | 83.5% | 81.0% |
| Jersey | 51.0% | 58.0% |

Results from these tables indicate that the proposed SNP for GenoEx-PSE can be expected to perform well for parentage analysis, noting that the Jersey breed has the lowest average MAF and lowest percentage of SNP with a minimum MAF of .30. Even here, of those with MAF ≥.30, there are at least 100 of the 200 SNP in Group A for parentage verification and a total of almost 500 when combining Group A and Group B-675 for parentage discovery.

### *Comparison of Parentage Analysis Results*

For sake of brevity, results for only the Holstein breed are presented. Of the 573,988 animals included in the analysis for assessing if the reported sire can be verified, a total of 558,828 (97.4%) did not have any SNP conflicts in any of the analyses conducted, leaving a total of 15,160 animals for the further assessment. Table 4 provides a summary of the parentage verification results based on the 200 SNP in Group A for GenoEx-PSE versus the list of 2,683 SNP used routinely by CDN.

**Table 4.** Comparison of parentage verification results when using Group A SNP versus the 2,683 SNP used in Canada.

| Group A Conflicts* | CDN Result | | |
|---|---|---|---|
| | Conflict | Verified | Total |
| 0 | 1 | 9,067 | 9,068 |
| 1 | | 3,306 | 3,306 |
| 2 | 3 | 126 | 129 |
| 3 | 3 | 11 | 14 |
| 4 | 2 | | 2 |
| 5 | 4 | 2 | 6 |
| 6 | 8 | | 8 |
| 7 | 9 | | 9 |
| 8 | 22 | | 22 |
| 9 | 27 | | 27 |
| 10 | 52 | | 52 |
| >10 | 2,517 | | 2,517 |
| **Total:** | **2,648** | **12,512** | **15,160** |

** - 0-3 conflicts = Parent accepted, 4-5 = Parent in doubt, >5 = Parent excluded.*

The CDN process identified a total of 2,648 animals with a sire parentage conflict of which seven were found to have three or fewer conflicts based on the Group A SNP alone. Among the remaining 12,512 animals for which the CDN process verified the reported sire, all but two received the same status based on Group A SNP, having 3 or less conflicts. These results show that using the Group A SNP for parentage verification yield results that are 99.9984% consistent with the official CDN results studied (558,819 out of 558,828), which translates to differences for 16/1,000,000.

**Table 5.** Comparison of parentage verification results when using Group A (ISAG 200) as well as Group B-675 SNP versus the 2,683 SNP used in Canada.

| Group A and B-675 Conflicts | CDN Result | | |
|---|---|---|---|
| | Conflict | Verified | Total |
| 0 | 1 | 274 | 275 |
| 1 | | 11,153 | 11,153 |
| 2 | | 892 | 892 |
| 3 | | 137 | 137 |
| 4 | | 41 | 41 |
| 5 | | 8 | 8 |
| 6 | 2 | 4 | 6 |
| 7 | | | |
| 8 | 2 | 1 | 3 |
| 9 | | 2 | 2 |
| 10 | | | |
| >10 | 2,643 | | 2,643 |
| **Total:** | **2,648** | **12,512** | **15,160** |

The GenoEx-PSE service also allows the user to participate at a level that includes the exchange of SNP genotypes for the purpose of parentage discovery in addition to parentage verification. The analysis examined any differences found in terms of discovering a valid sire for each of the 26,691 animals included. The sire considered to be discovered was the male that had the fewest SNP conflicts among those with less than 1% conflicts based on the informative SNP each sire had with the animal in question. Technically, 99.865% of the discovered sires based on using SNP from Groups A and B-675 were the same as the results based on the full set of 2,683 SNP routinely used by CDN. That said, in all other cases but one the two SNP groups ended up with different males discovered as the sire but they were identical twins. Only one case resulted in the two processes identifying full brothers as the discovered sire, which translates to a concordance rate of 99.996% in terms of discovering the correct genotype as the animal's sire. In practise, any parentage discovery procedure would have to include pre- and/or post-processing for handling genetically identical parents in addition rules associated with birthdates to avoid discovering a progeny as a potential parent.

Based on the comparative results of parentage verification and parentage discovery carried out in this study, there is no reason to not accept the use of the 200 SNP recommended by ISAG for parentage verification and the additional B-675 SNP for parentage discovery as the basis of the SNP genotype exchange for GenoEx-PSE.

### *Accuracy of Imputation and Estimation of Genomic Breeding Values*

A key step in the calculation of GEBV is the imputation of genotypes from lower density panels to the current international standard of using 50K genotypes. More importantly, however, is the accuracy of any GEBV estimated using such imputed genotypes. Initially, the intent was simply to quantify imputation accuracy (i.e.: error rates) and GEBV correlations that would be derived from genotypes that only included SNP from Groups A and B-675. During the course of the research, it was decided to repeat the analysis using SNP from Groups A and B-354 (reduced set from Group B-675) and from Group A alone. Table 6 presents the distribution of animals by the level of imputation error for the group of 27,324 Holsteins born in 2016 that had their 50K genotype reduced to include only the three subsets of SNP considered for use by GenoEx-PSE. As expected, the overall level of the imputation error increased as the number of SNP included decreased. In fact, all animals but one had an imputation error of at least 5% when only Group A (200 ISAG) SNP were used. Of particular interest is the important difference in imputation accuracy when the SNP from Group B-354 were added to Group A rather than those from Group B-675, which reduced the percentage of animals with less that 5% imputation error from roughly 66% to 13% (Table 6). This loss in imputation accuracy with B-354 is due to the fact that over 80% of those SNP are located on only 10 of the chromosome so relatively little information is available for imputation of SNP on the other chromosome.

**Table 6.** Percentage of animals by level of imputation error when SNP from three different groups defined for GenoEx-PSE were used to impute to a 50K genotype (n=27,324).

| Imputation Error | Group A | Groups A and B-354 | Groups A and B-675 |
|---|---|---|---|
| 0 - 4.99% | 0.004% | 12.919% | 65.840% |
| 5 - 9.99% | 18.167% | 54.231% | 6.822% |
| 10 - 14.99% | 47.003% | 3.594% | 7.880% |
| 15 - 19.99% | 4.271% | 4.853% | 10.240% |
| 20 - 24.99% | 2.302% | 9.852% | 5.610% |
| 25 - 29.99% | 5.991% | 7.678% | 2.049% |
| 30 - 34.99% | 9.896% | 4.531% | 1.215% |
| 35 - 39.99% | 5.406% | 1.998% | 0.307% |
| 40 - 44.99% | 6.024% | 0.326% | 0.026% |
| 45 - 49.99% | 0.926% | 0.018% | 0.011% |
| 50 - 54.99% | 0.011% | 0.000% | 0.000% |
| **Total** | **100%** | **100%** | **100%** |

In terms of assessing the accuracy of the second step, namely the estimation of GEBV, it is important to recognize there are two key components to a GEBV. In Canada, GEBV is calculated as a combination of the animal's traditional EBV (or Parent Average) and its Direct Genomic Value (DGV) derived by GBLUP, weighted by the relative estimated reliability of each value. Given that EBV/PA is constant for each animal, the analysis conducted here focused on the impact on the DGV resulting from the genotypes imputed from the three sets of SNP considered for GenoEx-PSE. Table 7 provides various correlations with the DGV of the official evaluation published by CDN for one of Canada's national selection indexes, the Lifetime Performance Index (LPI). Even with the use of only the 200 SNP recommended by ISAG for parentage verification (Group A) for imputation and GEBV estimation, a correlation of 0.87 was obtained with the DGV of the official evaluation published by CDN. While this may be considered reasonably high, this approach shows no improvement in accuracy compared to what would be obtainable simply by knowing the animal's pedigree for calculation of its Parent Average, which also yielded a correlation of 0.87 in this study (Table 7).

**Table 7.** Correlation between the official Direct Genomic Value (DGV) for Lifetime Performance Index (LPI) and DGV resulting from 50K genotypes imputed using three subsets of SNP considered for GenoEx-PSE.

| SNP Used for Imputation to a 50K Genotype | Correlation with Official DGV |
|---|---|
| No genotype available, only pedigree* | 0.87 |
| Group A (n = 200) | 0.87 |
| Groups A and B-354 (n = 554) | 0.92 |
| Groups A and B-675 (n = 875) | 0.96 |

** Correlation between the animal's Parent Average without any genotype and the official DGV.*

When SNP added for parentage discovery within the GenoEx-PSE service were included in the analysis, the correlation with the official DGV increased to 0.92 for Group B-354 and to 0.96 for Group B-675 (Table 7). This resulting correlation for Group B-675 is considered too high to use this set of SNP within the GenoEx-PSE service. The lower correlation of 0.92 for Group B-354 provides a better balance between accuracy of parentage discovery and obtaining only a moderate level of accuracy for GEBV estimation, which is a prohibited use of the genotypes exchanged according to the GenoEx-PSE service agreement. In addition, the level of GEBV accuracy that was obtained in this study from the imputation of SNP from Groups A and B-354 to derive 50K genotypes is highly dependent upon having a large population of ancestors that have a 50K genotype available for accurate imputation. For this reason, any such misuse of the SNP genotypes exchanged within GenoEx-PSE will not be beneficial even if not prohibited.

## Conclusions

From the results of this analysis, based on genotypes available in Canada for five dairy cattle breeds, the following conclusions can be drawn:

1. An analysis of the within breed minor allele frequency of the SNP proposed for exchange within GenoEx-PSE did not reveal any

concerns with the potential use of such SNP for parentage analysis in the Ayrshire, Brown Swiss, Guernsey, Holstein and Jersey populations with genotypes at CDN in Canada.

2. The 200 SNP recommended by ISAG for parentage verification in cattle populations performed very well with results highly consistent with those based on roughly 2,600 SNP that have routinely been used by CDN. These SNP should be included in the GenoEx-PSE exchange of genotypes for parentage verification.

3. For parentage discovery, two possible sets of SNP were analyzed in addition to the ISAG 200 SNP for parentage verification, referred to as Groups B-675 and B-354. The latter set includes a subset from B-675 with 80% being from only 10 chromosome. This study found that both sets perform very well for parentage discovery compared to the larger set of roughly 2,600 SNP used by CDN.

4. The comparison of imputation error rates and accuracy of GEBV estimation showed that the use of SNP from Groups A and B-354 was most desired for accurate parentage discovery analysis without resulting in high levels of accuracy of imputation to 50K genotypes and estimation of GEBV.

## References

International Society for Animal Genetics (ISAG), 2012. *Guidelines for cattle parentage verification based on SNP markers.* http://www.isag.us/docs/guideline-for-cattle-snp-use-for-parentage-2012.pdf

McClure, M.C., McCarthy, J., Flynn, J., Weld, R., Keane, M., O'Connell, K., Mullen, M.P., Waters, S. & Kearney, J.F. 2015. SNP selection for nationwide parentage verification and identification in beef and dairy cattle. In: Kowalski, Z., N. Petreny, M. Burke, P. Bucek, L. Journaux, M. Coffey, C. Hunlun, and D. Radzio, eds. *Proceedings, International Committee for Animal Recording Technical Series,* June 2015, Krakow, Poland. ICAR, Via Savoia 78, 00198 Rome, Italy, 175-181.

Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics. 15:478.*

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91,* 4414–4423.