CDCB's Genotyping Laboratory Certification Program

José A. Carrillo[1], George Wiggans[1], Ezequiel L. Nicolazzi[1], Kaori Tokuhisa[1], Duane Norman[1] and João Dürr[1]

[1] *Council on Dairy Cattle Breeding, Bowie, MD 20716, USA*

Abstract

Recent technological advances allow a large number of animals to be genotyped in a short period of time at relatively low cost. This, along with the long-term benefits generated when accurate information is obtained and used properly has led to a continuous increase in the number of genotypes received and processed by the Council on Dairy Cattle Breeding (CDCB). The exponential growth in the amount of data exchanged periodically presents a challenge for adequate quality control. It is well known that the quality of raw data is critical for producing accurate results; therefore, the CDCB has developed a customized Quality Control System designed for evaluating genotyping laboratories. The goal of this program is to assist the laboratories in improving the quality of their submissions and protecting the integrity of CDCB's database. The CDCB only accepts data from laboratories that meet and comply with all the established certification requirements. Currently, the CDCB accepts genotypes directly from seven certified laboratories. Each one is strictly monitored, and a monthly report card is provided to summarize the laboratory's performance. The report card includes six metrics, which are divided into two categories, critical or major, based on their significance. Each of these metrics has a threshold that has been derived from the data. A laboratory must provide an explanation for any failed metric on a monthly basis, which is evaluated by CDCB staff and will be considered with the lab's overall performance during the Annual Review. The Annual Review, implemented in 2018, determines the laboratory's certification status as one of certified, conditional, provisional, or decertified. Any provisional laboratory that fails to obtain or maintain the CDCB certification has the right to appeal the decision within 10 business days of notification. We expect this system to ensure the integrity of the data and the quality of service and products provided by the CDCB.

Key words: genomics, SNP data, quality control, laboratory certification, dairy cattle

Introduction

Recently, the number of genotyped animals has been increasing exponentially, presenting a challenge to the CDCB for adequate quality control of the exchanged data. In April 2019, the National Cooperator Database included 3,340,991 genotypes (Figure 1). The genomic data received by the CDCB has different sources of variation: the genotyping is performed using different types of chips and by two distinct technologies (Illumina and Affymetrix), and the data is produced in many laboratories. Therefore, the data must be checked to insure that it is comparable. Accordingly, the CDCB has developed a customized Quality Control System designed for evaluating genotyping laboratories and maintaining the integrity of the database.

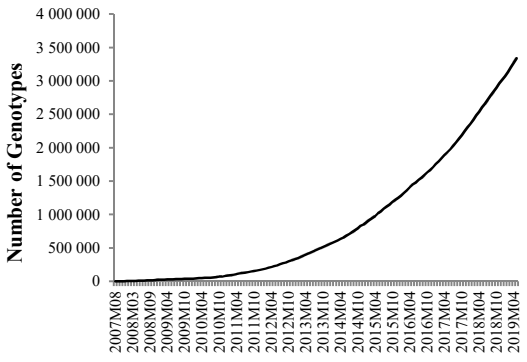

Figure 1. Number of genotypes processed by the CDCB

The main purpose of the quality control (QC) program is to ensure the accuracy and uniformity of all records included in the national genomic evaluation. Additional objectives are to: regularly monitor the performance of certified laboratories to ensure data quality; detect the needs or issues experienced by laboratories; advise or find solutions for issues/concerns faced by labs; facilitate the exchange of data (in the most efficient way) and improve the communication with the participant laboratories.

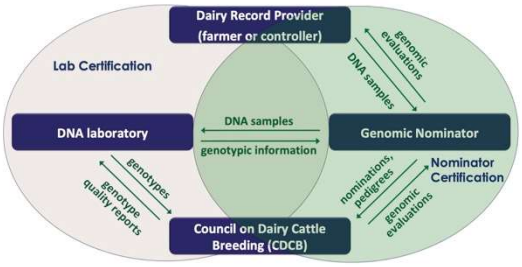

Figure 2. Data flow with Nominators and Laboratories

CDCB interacts with many types of organizations (such as National Evaluation Centers, Breed Associations, Dairy Record Processing Centers, Genotyping Laboratories and Nominators) but mostly it exchanges data with genomic nominators and laboratories. Thus, two independent certification processes, one for nominators and another for laboratories, have been developed (Figure 2). The nominator QC program was implemented in 2017 and the laboratory QC program started in 2018, and both are considered a main part of the certification process.

Laboratory Certification Process

The CDCB only interacts with laboratories that have been certified by the CDCB. Currently, there are seven approved laboratories that work closely and exchange data with the CDCB. Figure 3 shows the steps to achieve the CDCB certification.

Figure 3. Certification process

Core Requirements for Genotyping Laboratories

In order to receive the certification, the laboratory has to be able to perform the tasks required for submission of genotypes and agree to meet all the following requirements:

1. Submit an application form and pay the CDCB certification fee.
2. Provide a copy of an up-to-date accreditation of a Quality Certification program (e.g. ISO/IEC 17025 or similar). This certification should cover the entire process of creating the genomic data sent to CDCB.
3. Appoint the staff that will have access to the CDCB system. They will receive training provided by the CDCB.
4. Provide a comprehensive laboratory Standard Operating Procedures (SOPs) document.
5. Provide test files for verification purposes.
6. Sign a Material License Agreement (MLA) that describes the rights and responsibilities regarding the use, sharing and distribution of data.

After certification is granted:

1. Maintain a valid accreditation of standard laboratories processes.
2. Comply with the minimal requirements of any official CDCB proficiency test.
3. Verify the nomination of samples before submitting genotypes to CDCB.
4. Be able to submit genotypes in the required format and provide the corresponding sample sheet correctly.
5. Be able to use the CDCB online tools to check files before uploading them to the database.
6. Be able to identify and delete low-call-rate samples before uploading.

7. Demonstrate capability to investigate and resolve genotyping issues, such as genotypes with low call rate SNPs, abnormal proportion of heterozygous genotypes or high number of parent-progeny conflicts.
8. Be able to coordinate with the genomic nominator of the genotyped animals to ensure a reliable association of the genotype with a valid animal identification, pedigree and fee code.
9. Comply with the laboratory performance metrics (see next section).

Monthly report card

The CDCB provides a monthly laboratory Report Card that includes statistics on the quality of the submitted data. The metrics have been developed to monitor different aspects of the operation that affect the quality of the data, as seen in Figure 4. Some of these metrics have used and described earlier (Wiggans, VanRaden and Cooper, 2011)

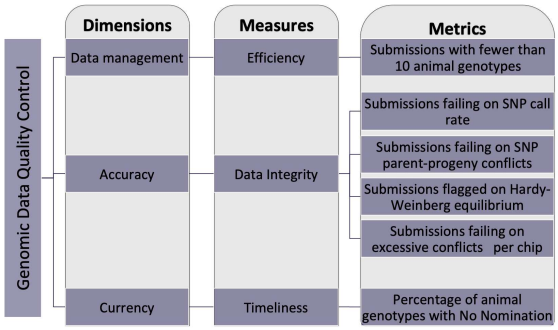

Figure 4. Genomic Data QC metrics dimensions and measures

Additionally, the metrics have been classified according to their impact on the quality and processing of the data, as critical and major.

Critical metrics

- **Submissions with fewer than 10 animal genotypes:** This metric considers the proportion of the total number of submissions that contain fewer than 10 genotypes. The threshold for this metrics is 10% because submissions with a low number of animals require the same processing setup time, which affects the efficiency of the system. Although it is allowed, due to certain specific conditions and corrections required, submissions of batches with a low number of animals is strongly discouraged.

- **Submissions failing on SNP call rate:** It accounts for the percentage of total submissions that have at least 0.5% (5/1000) of the SNPs with call rates lower than 90%. The established threshold is 50% of the total number of submissions.

- **Submissions failing on SNP parent-progeny conflicts:** This metric denotes problems in the "reliability" of SNPs. SNPs with more than 2% parent-progeny conflicts are counted; therefore special attention should be taken on these cases. Reclustering may solve this problem. The threshold for a "single submission" to fail in parent-progeny conflicts is 0.05% (5/10,000). However, the established 25% threshold considers the total number of submissions for the month.

- **Submissions flagged on Hardy-Weinberg equilibrium (HWE):** A large number of submissions with SNPs out of HWE (Hardy-Weinberg Equilibrium; heterozygote frequencies departing from the expected range) can be caused by low genotyping quality. The threshold to fail this metric is 50% of total submissions.

Major metrics

- **Percentage of animal genotypes with No Nomination:** Represents the proportion of total animals that lack a nomination at the time of uploading. Although the genotyping laboratory is not formally responsible for this (as the CDCB approved genomic nominators do the nomination), the

genotyping laboratory is required to alert nominators when nominations are missing and should postpone submission until nomination is completed. The establish threshold for this statistic is 3% of total number of genotypes submitted in a month.

- **Submissions failing on excessive conflicts per chip:** Batches with more than 80% of animals conflicting by chip are flagged. Typically, this problem arises when there is a misidentification of genotypes. This can be at the source (nominator) or at the lab level. Therefore, although certain flexibility is allowed, the time-consuming nature of these corrections makes a large number of such submissions undesirable. The threshold for this statistic is 10% of the total number of submissions.

Every month, laboratories receive a report card that includes these metrics and, for the failed ones, the lab must explain the cause of the failure and a plan to reduce the incidence of these cases, within one week after receiving the report card. Their responses are evaluated considering the specific circumstances of each lab and are stored for re-analysis during the annual review.

Annual Review

The annual review takes place at the end of each year and includes the next steps:

1. **Metrics assessment:** The monthly metrics are summarized to obtain annual performance statistics for each laboratory. This allows seeing the trend of each metric over time and revealing the effect of any change or action that occurred during the year.
2. **SOPs Review:** The lab must present a complete and current SOPs; which includes three main topics: Sample management, DNA analysis and Data management and exchange with CDCB.
3. **Preliminary Review Card:** A preliminary report that includes a summary of the performance metrics and SOPs evaluation is generated, and send to the laboratory before the review meeting.
4. **Review Meeting:** During the meeting the preliminary review results are discussed and possible corrective actions are proposed and defined.
5. **CDCB Final Recommendations and Status Certification:** The final results of the annual evaluation and the certification status are provided in this document.

The Annual Review defines the laboratory's certification status as one of certified, conditional, provisional, or decertified. Any provisional laboratory that fails to obtain or maintain the CDCB certification has the right to appeal the decision within 10 business days of notification.

Results and Discussion

The individual performance of the laboratories varies considerably in almost all metrics. Although this was expected due to the nature of the business and that every lab operates differently, our final goal is to decrease this variation over time. Also, because this is the first time these measures have been obtained, the laboratories performance could not be compared with previous years.

When the laboratories performance was analyzed in 2018, many underperformed in two categories: submissions with fewer than 10 animals and submissions failing on excessive conflicts per chip.

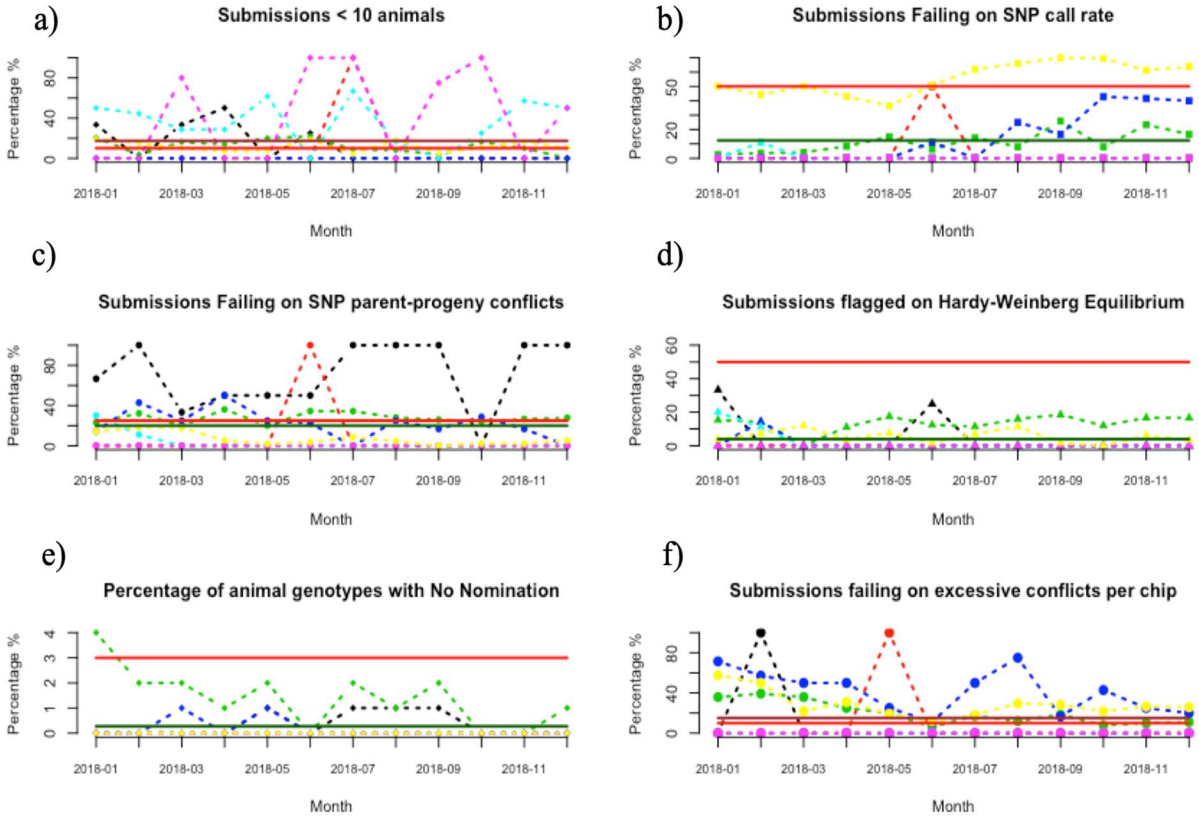

Figure 5 Overall laboratories performance metrics. Red line represents the defined threshold; green line, successful average performance of laboratories; brown line, average performance of laboratories below desirable level.

The first one is critical for the CDCB operations because it prevents the efficient use of our computing resources and the second one is even more relevant because it affects the accuracy and integrity of the data, as shown in Figure 4. The labs improved after the implementation of this program (Figure f); however, their performance still remained below the desired level.

For all other metrics, the overall performance was within desired levels. From these four metrics, two stood out due to their variation; Submissions failing on SNP call rate and Submissions Failing on SNP parent-progeny conflicts. This can be explained by the large difference in the number of submissions across labs. For labs with a small number of submissions, even one submission that experiences some problem can have a large effect on the proportion of failing submissions, as depicted by the black line in Figure c, where it reached 100% many times over the year. Therefore, these statistics should be analyzed carefully and each laboratory's individual characteristics should be considered.

By implementing this program, we expect to achieve a result similar to the one obtained after applying an analogous approach with genomic nominators.

As an example, Figure 6 describes the impact of the program in nominator's performance. In 2017, nine of the eleven analyzed nominators surpassed the threshold. However, in 2018 only five of them remained above the threshold but, encouragingly, showing smaller deviations from the threshold.

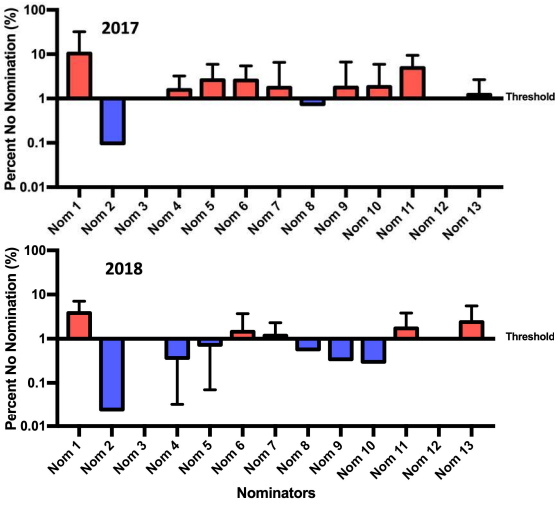

Figure 6. Comparison of Nominators' performance a year after the implementation of the QC program

Conclusion

The CDCB has developed a customized QC system for evaluating laboratories performance. Previous experience with nominators, demonstrated the positive impact of immediate feedback and annual review of performance. We expect that the QC program will assist the laboratories in delivering high quality data and contribute to maintaining the integrity of the CDCB database.

Acknowledgements

We would like to acknowledge the participating dairy producers for supplying data, DHI organizations and DRPCs for processing and relaying the information to the Council on Dairy Cattle Breeding (CDCB) and the purebred breed associations for providing pedigree data.

References

G. R. Wiggans, P. M. VanRaden, T. A. Cooper. The genomic evaluation system in the United States: Past, present, future. J Dairy Sci 94, 3202-3211 (2011)

"CDCB Collaborator Portal." QC Metrics for Submitted Genotypes - CDCB Collaborator Portal - CDCB Integrated Documentation System, https://redmine.uscdcb.com/projects/cdcb-customer-service/wiki/QC_Metrics_for_submitted_genotypes