Summary

Genomic selection is producing profound changes in dairy cattle market since reliable breeding values, which double the reliability of the pedigree index, can be obtained earlier in an animal's life. As a result, genetic gains of properly designed genomic programs are considerably larger than genetic gains obtained with traditional approaches. The industry has introduced this new tool all around the world faster than any other previous improvement.

This thesis contains six chapters, in which initial stages for the implementation of genomic selection program in Spanish Holstein population were studied using simulations and real data. The initial interest began in 2008 (González-Recio et al., 2008), when the results obtained by VanRaden, (2008) were used to involve the Spanish industry in genomic selection. This research has been used to obtain the official genomic breeding values and implement the imputation of genotypes.

The global aim of this thesis was to contribute with practical recommendations for implementing genomic selection in the Spanish dairy cattle. The specific objectives were: (1) To study alternative genotyping strategies for small populations, (2) to develop and validate methods for the evaluation of large data sets of genotypes, and (3) to study the effect of imputation on predictive ability.

The main topics with respect to genomic selection in dairy cattle were discussed in chapter 1 including: genetic and statistical aspects underlying genomic selection, design of proper reference populations (**RP**), review of methodology for genome-assisted evaluation, imputation, and implementation of genomic selection in dairy cattle breeding programs. Breeding values with medium high accuracies are now available early in the life of the animals. This is modifying one of the traditional principles of dairy market: the strong preference for highly reliable bulls.

In chapter 2, a simulation study was carried out comparing female-selective genotyping strategies with traditional pedigree index and a bull RP. The Spanish male RP has 1,600 genotypes, which is not large enough to provide reliable predictions. Alternatives should be evaluated to improve predictive ability. The accuracy of predicted genomic breeding values using the two-tailed strategies was better than the accuracy obtained using other strategies (0.50 and 0.63 using yield deviations as phenotype and 0.48 and 0.63 using breeding values in low- and medium-heritability scenarios, respectively, using 1,000 genotyped cows). When 996 genotyped bulls were used as the training population, the sire' strategy led to accuracies of 0.48 and 0.55 for low- and medium-heritability traits, respectively. The most informative strategy involved genotyping of females that exhibited upper and lower extreme values within the distribution. Including just top animals resulted in poor results.

Several methods for implementing genome assisted evaluations were compared in Chapter 3. Methods including marker regression included Bayesian methods (Bayes-A, Bayesian LASSO and Random Boosting (R-Boost). G-BLUP was also utilized using the genomic relationship matrix. The Spanish RP was used to compare those methods in terms of predictive ability and bias. Genomic predictions were more accurate than traditional pedigree indices for predicting future progeny test results of young bulls. The gain in accuracy, due to inclusion of genomic data, varied by trait and ranged from 0.04 to 0.42 Pearson correlation units. Results averaged across traits showed that Bayesian LASSO had the highest accuracy with an advantage of 0.01, 0.03 and 0.03 points in Pearson correlation compared with R-Boost, Bayes-A, and G-BLUP, respectively. The B-LASSO predictions also showed the least biased predictions (0.02, 0.03 and 0.10 SD units less than Bayes-A, R-Boost and G-BLUP, respectively), measured as the mean difference between genomic predictions and progeny test results. The R-Boost algorithm provided genomic predictions with regression coefficients closer to unity, for four out of five traits and also resulted in mean squared error estimates that were 2%, 10%, and 12% smaller than B- LASSO, Bayes-A, and G-BLUP, respectively. R-Boost seemed to be a competitive marker regression methodology in terms of predictive ability.

Chapter 4 describes the R-Boost algorithm tested in Chapter 3 for genomic evaluations in large data sets. After joining the Eurogenomics consortium with more than 22,000 bulls in the RP, a feasible method with reasonable computation times, and no impaired predictive ability was required. The random boosting uses a random selection of markers to add a subsequent weak learner to the predictive model. Optimization of the algorithm and behavior of tuning parameters was tested in real dairy cattle data. Those tuning parameters control the percentage of single nucleotide polymorphisms (**SNP**) sampled per iteration and the level of shrinkage over the regression coefficient estimation. The proposed modification of the original boosting algorithm can be run in 1% of the time used with the original algorithm, and with negligible differences in accuracy and bias.

In Chapter 5, genotypes from the GoldenGate Bovine 3K and BovineLD BeadChip for 834 animals were imputed to a BovineSNP50v2 BeadChip using *Beagle*. Those genotypes were subsequently imputed to the BovineHD BeadChip. Predictive ability of imputed and native genotypes as RP in genome-assisted evaluations was compared using G-BLUP and R-Boost. Imputed low density genotypes achieved similar predictive ability than native genotypes. However, marginal better selection efficiency was obtained after imputation to HD (0.002 greater Pearson correlation units). The largest improvements were found for Days Open after imputation to HD genotypes (up to 0.06 greater Pearson correlation units). R-Boost was more sensitive to marker density than G-BLUP. Both methods performed similar except for Fat Percentage, where R-Boost outperformed G-BLUP with up to 0.20 Pearson correlation units.

The predictive ability of certain traits may be improved either by imputing genotypes to HD or by utilizing a method that takes into account the genetic architecture of the trait.

Finally, in chapter 6 a general discussion links the studies previously covered with the implementation of genomic selection in the Spanish dairy cattle is reported. The first Spanish RP with above 1,600 progeny tested bulls was tested as a proper source of genomic information in chapter 4 and was used for comparing methods and scenarios in chapters 3, 4 and 5. First genomic evaluation was carried out for those traits included in Chapter 4 of this thesis and results were used for AI centers in September 2011. The Eurogenomics population was included on November 2011. First complete genomic evaluation for the 26 traits included in the Spanish index (ICO) was carried out in February 2012 using Random Boosting as described in chapter 4. In May 2012 Spanish genomic evaluation for protein yield was validated by Interbull. Finally, on November 30th 2012, first official genomic evaluations were published on-line by CONAFE (http://www.conafe.com/noticias/20121130a.htm).