Lessons learned in pooling data for reference populations

E. Wall[1], M.P. Coffey[1,] R.F. Veerkamp[2], S. McParland[3] and G. Banos[1, 4]

[1]*SAC, Roslin Institute Building, Easter Bush, Midlothian, EH25 9RG UK*

[2]*Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, the Netherlands*

[3]*Animal and Bioscience Research Department, Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Co. Cork, Republic of Ireland*

[4] *Department of Animal Production, Faculty of Veterinary Medicine, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece*

Abstract

This study set out to demonstrate the feasibility of merging data from 4 different experimental resource dairy populations (1 herd in each of Scotland and Ireland, and 2 in the Netherlands) to create a pooled reference population for joint genetic and genomic analyses. Data included a total of 60,058 weekly records from 1,630 Holstein-Friesian cows across the 4 herds and included 7 traits: milk, fat and protein yield, milk somatic cell count, live weight, dry matter intake, and energy intake and balance. Missing records were predicted using random regression models, so that at the end there were 44 weekly records, corresponding to the typical 305-day lactation, for each cow. Data were subsequently merged and analysed with mixed linear models. Genetic variance and heritability estimates were greater ($P<0.05$) than zero for all traits except for average milk somatic cell count in weeks 16-44. Proportion of total phenotypic variance due to genotype by environment (sire by herd) interaction was not different ($P>0.05$) from zero. When estimable, the genetic correlation between herds for the same trait ranged from 0.85 to 0.99. Results suggested that merging experimental herd data into a single dataset is both feasible and sensible, despite potential differences in management and recording of the animals in the four herds. Merging experimental data will increase the precision of parameter estimates in a genetic analysis and augment the potential reference population in genome-wide association studies especially of difficult-to-record traits.

Introduction

Undertaking genetic studies of animal phenotypes presupposes accurate recording on sufficient numbers of animals to help dissect the genetics of the trait. The development of elaborate national monitoring schemes has facilitated routine population-wide on-farm accurate recording of several conventional traits, mostly associated with production (International Committee for Animal Recording, 2011). Certain indicators of functional traits, such as milk somatic cell count in dairy cattle, are included in these programmes. Nevertheless, several increasingly important traits associated with health, fitness and efficiency are currently not possible to routinely record in the commercial population. Understanding the genetic background of such traits depends then on data from experimental resource populations where animals are raised in controlled, closely monitored environments. Such populations, however, are usually of limited size. Combining data from different

experimental herds would provide an expanded dataset that would allow a more rigorous genetic and genomic analysis of difficult- and expensive-to-record traits.

The objectives of this study were to (i) demonstrate the feasibility of merging phenotypic data from different experimental resources and (ii) characterise the merged database and assess its suitability for a joint genetic analysis as a single dataset.

Methods

Animals, experimental herds and database

Data were collected from first lactation Holstein cows raised in 4 distinct experimental resource herds in 3 different countries (1 herd in each of Scotland and Ireland, and 2 in the Netherlands). In all cases, cows were used in various ongoing or previously completed experiments conducted at different time-periods, briefly summarised below.

Scotland – Crichton herd: Data originated from the Scottish Agricultural College Dairy Research Centre based at Crichton Royal Farm. These cows had previously comprised the Langhill herd, Edinburgh (Veerkamp et al., 1995; Pryce et al., 1999) and were transferred to Crichton Royal Farm in September 2001. The herd normally consists of approximately 200 milking cows divided evenly between two genetic groups (control vs. selection) established in 1992 as part of a still ongoing selection experiment. Cows in each genetic group are further split randomly into two diet groups (high-concentrates vs. high-forage) for the purposes of a feeding experiment which is also in progress. The 2 genetic groups on a particular diet are managed together.

All cows are kept together and treated the same at all times, except where the production systems require management differences. Cows are milked 3 times per day.

For the purposes of the present study, data pertained to 563 cows equally distributed across the 4 experimental groups that had calved between 1992 and 2009.

Ireland – Moorepark herd: This experimental resource is located at the Teagasc Moorepark Research Farm. Data were collated from several studies that had been previously conducted (Buckley et al., 2000; Kennedy et al., 2003; O'Donovan and Delaby, 2005; Horan et al., 2006; Kennedy et al., 2006; McCarthy et al., 2007; McEvoy et al., 2007). In brief, these studies compared either alternative genotypes of Holstein-Friesian cows raised on different production systems or alternative grazing strategies or grass varieties. Different strains of Holstein-Friesians were evaluated on contrasting grass-based production systems. Animals within strain were randomly assigned, at the start of lactation, to feed systems. Annual concentrate feeding level varied from 325 to 1,452 kg per cow. All cows calved in the spring, were fed predominantly grazed grass, and were milked twice daily.

For the purposes of the present study, data pertained to 449 cows that first calved between the years 1998 and 2008.

The Netherlands – TGEN and NBZ herds: Data originated from two experimental dairy herds, one located near Lelystad (TGEN) and another one near Leeuwarden (NBZ).

Data from the TGEN herd were collected between 1990 and 1998 and, for the purposes of the present study, pertained to 549 cows (Veerkamp et al., 2000). Two-thirds of these cows belonged to a high genetic merit group and the others were part of a control group. Cows in the latter group were, on average, about half a standard deviation below the former on the Dutch production index reflecting the impact of milk, fat and protein yield on future net profit. All cows were fed ad libitum with the same mixed ration.

Data collected from the NBZ herd were from the period 2003-2004 and pertained to 90 first lactation cows that were participating in a genetic and feeding experiment. Specifically, these cows were divided into two genetic groups with high and low genetic merit for fat and protein production, respectively. Furthermore, cows were split into two diet groups fed a high and a low caloric density ration, respectively.

To allow data from multiple resource herds to be pooled the RobustMilk project developed a cross institutional database of dairy herd animal identities with links to performance (production, health, fertility and robustness traits) and genomic information were available. The RobustMilk common database was developed in SQL Server 2005, with a web-interface developed within virtual private network (VPN) for security purposes. The database was developed with detailed QA control, error checking and back-up storage to ensure that data have the necessary protection in place. Common protocols for preparing and storing of data related to animals, pedigree, genotypes and performance records were developed. To aid the submission of phenotypic data from partners, a suggested template was derived to allow each partner to prepare data in a similar and coherent format. Partners have submitted detailed phenotypic data to the database and uploads developed. This will allow future update of the database for databases that are still "active". Also the development of data reporting standard protocols would make it easier for future new partners/populations to be merged into the database. To simplify the extract of data for partners, a procedure for extracting a merged data file was also written.

Traits recorded and prediction of missing records

Weekly individual cow records for milk, fat and protein yield, milk somatic cell count, live weight, dry matter intake, and energy balance were extracted from the database of each herd. Energy balance traits were also derived for all animals. Records pertained to the daily observation on the day of recording. When multiple records were available within a week of lactation, weekly values were corresponding arithmetic means. Only first lactation cows were considered in this study. After requisite edits, a total of 60,058 weekly records of 1,630 cows remained across the 4 herds.

In order to predict missing weekly records the following random regression model was used:

Y = HTYM + CYM + CA + GG + DG + IRL + MF + WK + COW.WK

where: Y = weekly cow record for a trait; HTYM = fixed effect of herd by year-month of record interaction (4 herds, 222 year-month classes); CYM = fixed effect of calving year-month interaction (188 classes); CA = fixed effect of calving age (3 classes; <704, 704-827, >827 days); GG = fixed effect of genetic group (8 classes); DG= fixed effect of diet group (4 classes); IRL = fixed effect of Irish diet treatment (18 classes); MF = fixed effect of milking frequency (2 or 3 times); WK = fixed lactation curve modelled with a 4th order polynomial (5th for somatic cell count); COW.WK = random cow deviation from fixed curve modelled with a 4th order polynomial (5th for somatic cell count)

Each recorded trait was analysed separately. In the case of milk somatic cell count, a log-transformation took place before the analysis to ensure normality. Effect solutions were combined to re-create the phenotypic record for all cow-weeks, including those with missing observations.

Subsequently, 23 phenotypic lactation traits were derived for all cows, using the predicted

weekly records for milk, fat and protein yield, milk somatic cell count, live weight, dry matter intake, energy intake, and energy balance (traits summarised in Table 1).

Estimation of genetic parameters

Lactation traits derived in the previous section were analysed with mixed linear models including the effects of herd, calving year-month and age, genetic and diet group, milking frequency, and cow. A pedigree file comprising 8,850 animals across the 4 herds was used to derive genetic variance and heritability estimates.

In a separate set of analyses, sire by herd interaction was added in the model as a random effect to assess the magnitude of genotype by environment (herd) interaction. The issue of genotype by environment interaction was also addressed with a series of multi-trait analyses, where individual traits in the 4 herds were treated as different but genetically correlated traits.

Results and discussion

Figure 1 illustrates the predicted lactation curves for milk yield (kg) across all herds and when separate curves were fitted for each of the 4 herds. Very similar results were derived when separate curves were fitted for each herd, suggesting that combining records across herds did not change the lactation profile of the trait. This was consistent for all traits in the study.

Estimates of genetic variance, heritability and proportion of total variance attributed to sire by herd interaction for all derived traits are shown in Table 1. Genetic variance estimates were greater ($P<0.05$) than zero in all cases except for average milk somatic cell count in weeks 16-44 of lactation, suggesting that diverse experimental data from distinct herds may be merged into a single database amenable to a joint genetic analysis.

In general, heritability estimates shown in Table 1 were consistent with estimates provided in the literature from various studies worldwide (e.g., Urioste et al., 2010; Vallimont et al., 2010;). Heritability estimates were also derived separately within herd and, in general, were in the same range as the across herd estimates presented in Table 1. For example, within herd heritability estimates of total milk yield in the first 15 weeks of lactation varied from 0.20 to 0.31, whereas estimates for total dry matter intake in the same period ranged from 0.15 to 0.27 in the 4 different herds. However, there were few cases, such as live weight, with more varying within herd heritability estimates of 0.38, 0.21, 0.47 and 0.48 for Crichton, NBZ, Moorepark and TGEN, respectively. This may be attributed to different variance scales in the 4 herds associated with different dataset sizes. For example, NBZ had the lowest estimate and smaller dataset.

Including a sire by herd interaction led to a small reduction in heritability estimates (Table 1), suggesting that part of the previously estimated additive genetic variance might be due to interaction effects. However, the proportion of total phenotypic variance accounted for by the sire by herd interaction effect was always not significantly ($P>0.05$) different from zero. Furthermore, models with and without a sire by herd interaction effect were compared with the Akaike Information Criterion (AIC). In all cases, the difference of AIC between the two models was statistically not greater than zero ($P>0.05$). For example, the AIC difference in the case of lactation milk in 15 and in 44 weeks was 0.44 and 0.80, respectively, suggesting that fitting a sire by herd interaction effect did not improve the fit of the model. This result implies that a joint analysis of data from the different herds that is based on the assumption of no genotype by

environment interaction is possible, while supporting the notion of merged phenotypes from these 4 herds being viewed as a single dataset in a genetic analysis.

Acknowledgments

This research receives a financial support from the European Commission, Directorate-General for Agriculture and Rural Development, under Grand Agreement 211708 and form the Commission of the European Communities, FP7, KBBE-2007-1. This paper does not necessarily reflect the view of these institutions and in no way anticipates the Commission's future policy in this area.

References

Buckley F, Dillon P, Rath M and Veerkamp RF 2000. The relationship between genetic merit for yield and live weight, condition score and energy balance of spring calving Holstein-Friesian dairy cows on grass based systems of milk production. Journal of Dairy Science 83, 1878-1886.

Horan B, Faverdin P, Delaby L, Rath M and Dillon P 2006. The effect of strain of Holstein-Friesian dairy cows and pasture-based system on grass intake and milk production. Animal Science 82, 435-444.

International Committee for Animal Recording 2011. www.icar.org, Rome, Italy.

Kennedy J, Dillon P, Faverdin P, Delaby L, Stakelum G and Rath M 2003. Effect of genetic merit and concentrate supplementation on grass intake and milk production with Holstein-Friesian dairy cows. Journal of Dairy Science 86, 610 - 621.

Kennedy E, O'Donovan M, Murphy JP, O'Mara FP and Delaby L 2006. The effect of initial grazing date and subsequent stocking rate on the grazing management, grass dry matter intake and milk production of dairy cows in summer. Grass Forage Science 61, 375-384.

McCarthy S, Berry DP, Dillon P, Rath M and Horan B 2007. Effect of strain of Holstein-Friesian and feed system on udder health and milking characteristics. Livestock Science 107, 1-28.

McEvoy M, O'Donovan M, Murphy JP, O'Mara F, Rath M and Delaby L 2007. Effect of concentrate supplementation and herbage allowance on milk production performance of spring calving dairy cows in early lactation. Proceedings Irish Agricultural Research Forum. Tullamore, Ireland 15th-16th March 2007.

O'Donovan M. and Delaby L. 2005. A comparison of perennial ryegrass cultivars differing in heading date and grass ploidy with spring calving dairy cows grazed at two different stocking rates. Animal Research 54, 337-350.

Pryce JE, Nielson BL, Veerkamp RF and Simm G 1999. Genotype and feeding system effects and interactions for health and fertility traits in dairy cattle. Livestock Production Science. 57, 193-201.

Urioste JI, Franzén J and Strandberg E 2010. Phenotypic and genetic characterization of novel somatic cell count traits from weekly or monthly observations. Journal of Dairy Science 93, 5930-5941.

Vallimont JE, Dechow CD, Daubert JM, Dekleva MW, Blum JW, Barlieb CM, Liu W, Varga GA, Heinrichs AJ and Baumrucker CR 2010. Genetic parameters of feed intake, production, body weight, body condition score, and selected type traits of Holstein cows in commercial tie-stall barns. Journal of Dairy Science 93, 4892-4901.

Veerkamp RF, Oldenbroek JK, van der Gaast HJ and van der Werf JHJ 2000. Genetic correlation between days until start of

luteal activity and milk yield, energy balance and live weights. Journal of Dairy Science 83, 577-583.

Veerkamp RF, Simm G and Oldham JD 1995. Genotype by environment interaction – experience from Langhill. In: Breeding and Feeding the high genetic merit dairy cow. (ed. T.L.J. Lawrence, F.J. Gordon, A. Carson). British Society of Animal Science (Occasional Publication) 19, 59-66.

Table 4 *Estimates of genetic variance and heritability (h^2) obtained with an additive model, and proportion of total variance due to additive effects (h^2_a) and to sire by herd interaction (sxh) when the latter was included in the model; standard errors are in parentheses*

Trait	Mean	Genetic variance		h^2		h^2_a		sxh	
Total milk yld (44 wks)	6,996.00	267,800	(85,940)	0.22	(0.07)	0.17	(0.08)	0.05	(0.03)
Total fat yld (44 wks)	278.58	317.8	(110.5)	0.20	(0.07)	0.16	(0.08)	0.03	(0.02)
Total protein yld (44 wks)	236.74	170.1	(73.76)	0.16	(0.07)	0.12	(0.07)	0.04	(0.03)
Av fat % (44 wks)	4.04	0.154	(0.021)	0.68	(0.07)	0.66	(0.08)	0.02	(0.03)
Av protein % (44 wks)	3.39	0.030	(0.005)	0.55	(0.07)	0.49	(0.08)	0.05	(0.03)
Av fat:protein ratio (44 wks)	1.19	0.008	(0.001)	0.66	(0.07)	0.66	(0.08)	0.00	(0.00)
Total milk yld (15 wks)	2,743.07	37,210	(12,840)	0.21	(0.07)	0.17	(0.08)	0.03	(0.03)
Av SCC* (wks 1-15)	118.94	0.118	(0.056)	0.14	(0.06)	0.10	(0.07)	0.03	(0.03)
Av SCC (wks 16-44)	99.65	0.059	(0.045)	0.09	(0.06)	0.06	(0.06)	0.02	(0.03)
Av LWT	530.38	532.2	(121.1)	0.35	(0.07)	0.30	(0.08)	0.04	(0.03)
Total DMI (44 wks)	4,650.91	26,020	(12,700)	0.15	(0.07)	0.15	(0.09)	0.00	(0.03)
Total DMI (15 wks)	1,541.96	6,329	(2,381)	0.22	(0.08)	0.17	(0.09)	0.03	(0.03)
Av DMI to milk yld ratio (44 wks)	0.66	0.004	(0.001)	0.28	(0.08)	0.23	(0.10)	0.04	(0.03)
Av DMI to milk yld ratio (15 wks)	0.55	0.002	(0.001)	0.21	(0.07)	0.21	(0.07)	0.00	(0.00)
Av energy balance (44 wks)	-9.29	10.96	(5.14)	0.17	(0.08)	0.13	(0.09)	0.03	(0.03)
Av energy balance (15 wks)	-20.64	38.15	(12.46)	0.27	(0.08)	0.27	(0.08)	0.00	(0.00)

* log-transformed 1000/ml

Figure 1. Predicted lactation curves for milk yield (kg) for a singe curve was fitted across herds and for each of the 4 herds (Crichton-Scotland, NBZ-the Netherlands, Moorepark-Ireland, TGEN-the Netherlands) by week of lactation; results are expressed on a common scale for all herds.