Estimating Genomic Variance Explained per Chromosome Using Pedigree and Genomic Data in Sheep

C. Esquivelzeta-Rabell, N. Moghaddar,[1] S. Clark,[1] and J.H.J. van der Werf[1].
School of Environmental and Rural Science, University of New England, Armidale NSW, Australia,
[1]Cooperative Research Centre for Sheep Industry Innovation, Armidale NSW, Australia

ABSTRACT: We used a half sib data structure for a growth trait in sheep as a potentially powerful design for partitioning the genetic variance across the different chromosomes. Records for post weaning weight were used from 2455 merino sheep. The model of analysis accounted for population structure by fitting genetic group effects as well as the numerator relationship matrix (A) based on pedigree. We then fitted the matrix D representing the difference between the genomic relationship matrix (G) and A. The matrix G was based on 48,599 SNP markers across the entire genome, or on all SNPs of an individual chromosome. There was a relationship between chromosome length (L) and variance explained (Vg_i), but we found significant differences in (Vg_i/L) between chromosomes.

Key words: genomic variance; polygenic variance; chromosomes.

INTRODUCTION

The rapid growth in availability of abundant genomic information allows more accurate tracking of chromosome segments through a pedigree. Matching such data to observed phenotypes allows genome wide association studies (GWAS) and genomic prediction of genetic merit and genomic selection (GS) (Meuwissen et al. (2001)). GWAS studies explicitly try to detect regions associated with quantitative trait variation, whereas GS is often applied without explicit knowledge of such regions, e.g. the so called GBLUP method uses genomic relationship treating information about marker genotypes equally across the entire genome (Meuwissen et al. (2001), Habier et al. (2007)).

With the ever increasing density of genetic markers, there is also an increasing need and ambition to work out which variants actually are responsible for the observed quantitative genetic variation. However, this has proven to be a non-trivial exercise, mainly due to the large number of genetic variants (large p) versus the relatively small number of data points (small n) combined with the small effects of each variant. Previous studies were unable to explain the total genetic variance using different density SNPs (Visscher et al. (2007), Yang et al. (2010), Jensen et al. (2012) and Haile-Mariam et al. (2013)).

One first step to detect how much genetic variance is accounted for by different chromosomal regions is to work at the chromosomal level. Visscher et al. (2007) and Yang et al. (2011) tried to partition the total genetic variance into different chromosomes, and they found a linear relationship between chromosome length and variance explained. This supports the hypothesis of a polygenic model, where the total variation is explained by many genetic variants, more or less equally distributed across the genome. The relationship was stronger for human height that for body mass index, indicating that the latter trait could be less polygenic. Also, Jensen et al. (2012) found a weaker relationship for production and fitness related traits in dairy cattle with R^2 values for a linear regression model of variance explained and chromosome length between 0.11 and 0.21.

The advantage of data on livestock is that the data structure is usually based on relatively large half sib families. This provides a powerful design for determining the segregation based on linkage. The design is not suitable for LD mapping; hence the accuracy of mapping QTL positions would be low. However, the latter is less relevant for determining the amount of genetic variance explained per chromosome.

The objective of this study was to partition the genetic variance over the different chromosomes for a growth trait in sheep, and to determine whether there are large differences between chromosomes, both before and after accounting for their length.

MATERIALS AND METHODS

Data for this study was obtained from the Information Nucleus (IN) program of the CRC for Sheep Industry Innovation. Details on this program and its design are described by van der Werf et al. (2010). The data set comprised a total of 2455 purebred merino lambs with phenotype, pedigree and genotype data. The animals descended from 139 sires and the associated pedigree file contained 10,559 animal identities from over 22 generations. The pedigree information was used to compute a numerator relationships matrix (A) for the animals with phenotypic records using the R package 'pedigree' (Coster (2012)). Genotypic information consisted of SNP marker genotypes was obtained using the Illumina OvineSNP50 BeadChip. After quality control and imputing missing genotypes with BEAGLE (Browning and Browning (2007)), genotype information on 48,599 SNP was used to derive a genomic relationships matrix (G) according to VanRaden (2008). Phenotypic information on post weaning weight (PW) was used in the analysis.

Models for analysis can be written in matrix notation as: $y = Xb + Z_1ID + Z_2m + Qq + e$, where the vector b included fixed effects of sex of lamb (ram: 1 or ewe: 2), birth type/rearing type (single: 1/1, twins: 2/2 or triplets: 3/3 and their combinations), management group, age of dam and post weaning age. ID is the random additive genetic effect of the lamb, m is the maternal permanent environmental effect and q is a genetic group effect. The genetic group consisted of merino strain where we regressed on strain proportion. Strain proportion was

derived from a deep pedigree analysis. To estimate the variance of additive genetic effects we used five different models to define the variance structure of ID: 1) var(ID) = $A\sigma_a^2$, 2) var(ID) = $G\sigma_g^2$ 3) var(ID)= $G\sigma_g^2 + A\sigma_a^2$ where the proportion of additive genetic variance was estimated by fitting both G and A to account for any probable remaining polygenic effect, 4) var(ID)= $A\sigma_a^2 + D\sigma_d^2$ where D = G - A and 5) var(ID)= $A\sigma_a^2 + D_i\sigma_{di}^2$ where D_i corresponds to the i^{th} chromosome matrix calculated as $D_i = G_i - A$. The first three models were fitted to evaluate how the additive genetic variance would be partitioned over G and A. The genomic relationship G is an estimate of the realized relationship at QTL as opposed to the expected relationship defined in A (Goddard et al. (2012)). The fourth model was an attempt to separate the variance between families from the variance additionally explained by genetic markers. The deviation in D reflects information provided by the marker genotypes orthogonal to the family structure of the data and therefore is expected to give an unbiased estimate of the variance of the segregating QTL effects. Model 5 was run for each chromosome, in an attempt to estimate the variance due to QTL effects per chromosome. In model 5 we only fitted one D matrix at a time. The variance components were calculated using ASReml 3.0 software (Gilmour et al. (2009)).

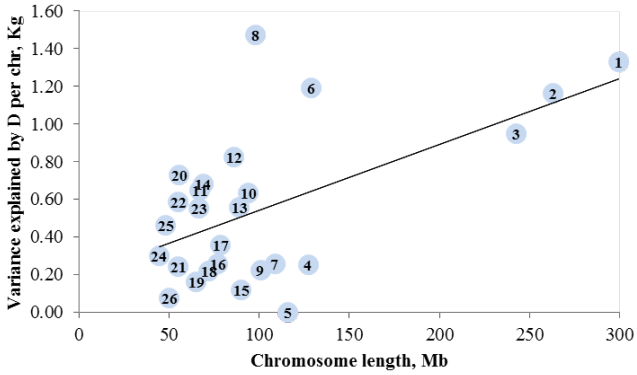
RESULTS

When A (model 1) and G (model 2) were fitted individually, A explained more variation for the trait (33.38%) than G (31.83%). The results from model 3 showed that most of the variance was partitioned toward G which agreed with previous reports (Jensen et al. (2012), Haile-Mariam et al. (2013)). Analysis using model 4 showed that A (21.22%) recovers some of the variation explained but D still captures most of the variance (26.18%). When fitting model 5 we observed that in general the variance explained by pedigree effects fluctuated between 5.70 and 6.61 Kg whereas the variance explained by each D_i varied and was somewhat related to the chromosome length (Figure 1).

Genetic groups were fitted in the model to be able to capture the variance explained by different strains of merino (Table 1). This component was not counted towards the phenotypic variance. Results for PW showed that the variance captured by q varied considerably among models, for model 1 the value of q was the highest (11.48 Kg) whereas in models 2 and 3 of the genetic group variance was much lower (3.16 Kg). This suggests that G captured some of the strain differences. The genetic group variance was also low in model 4 (2.81 Kg) whereas with models 5 it varied between 7.74 and 11.48 Kg. Results for maternal effects were similar among models, fluctuating between 2.16 and 2.33 Kg.

The sum of variance estimates per chromosome was higher (8.99 Kg) than the variance explained in the models were the full G was fitted (8.02 Kg), suggesting that different D_i matrices explain still a common variance. A positive correlation between variance explained and chromosome length was found (Figure 1), agreeing with previous reports for human height (Visscher et al. (2007) and Yang et al. (2011)); nevertheless the correlation was weak (R^2=0.32) and marked differences in additive genetic variance explained were found for some chromosomes. For example, chromosome 8 explaining the higher amount of genomic variance (10.36%), followed by chromosome 1 (9.34%), 6 (8.38%) and 2 (8.16%) and chromosome 5 is estimated to contribute 0% variance.

Figure 1. Variance explained by the difference (D) between genomic and pedigree relationships, calculated per chromosome for post weaning weight using model 5.



The difference between variance explained per chromosome and the percentage of expected variance explained, under the assumption that the genetic variance is proportional to the size of the chromosome, was also calculated. Results showed that chromosome 8 performed better than expected, explaining 6.66% more, followed by chromosomes 6, 20, 12, 14, 22, 11, 23, 25, 10, 13, 24 with percentages from 0.43 to 3.50. The rest of the chromosomes explained less variation than expected (results not shown).

DISCUSSION

We attempted to partition the genetic variance across the different chromosomes by define a matrix that was based on the difference between the expected and the realized relationships. The realized relationships are based on marker genotypes and do not fully represent the true relationships at the QTL. The differences are due to incomplete linkage disequilibrium between markers and QTL as well as sampling error at the marker genotypes. So

Table 1. Variance components for post weaning weight estimated using different mixed linear models.

Model[¥]	q	Va	Vg*	m	e
1	11.48	6.26		2.24	10.26
2	3.16		5.83	2.19	10.29
3	3.16	0.00	5.83	2.19	10.29
4	2.81	5.23	6.46	2.33	10.64
5[§]	10.16	6.19	0.55	2.25	10.34

[¥]1: var(ID)= $A\sigma_a^2$, 2: var(ID)= $G\sigma_g^2$, 3: var(ID)= $A\sigma_a^2 + G\sigma_g^2$, 4: var(ID)= $A\sigma_a^2 + D\sigma_d^2$, 5: var(ID)= $A\sigma_a^2 + D_i\sigma_{di}^2$

*Calculated from genomic relationship matrix (G) in the first 3 models and from the difference between G and numerator relationship matrix for the rest of the models.

[§] Mean values among chromosomes.

we should expect that the genomic component is an underestimate of the actual QTL variance at each chromosome. We found a higher total variance in models 4 and 5, which indicates that the D also captures some common variance, probably due to population structure not captured by the pedigree or by the genetic group effects. Nevertheless we can compare the variance explained by each D_i relative to each other and this gives probably a reasonable indication of the relative amount of variation explained by each chromosome. Further work can be done to check the repeatability of this exercise over independent datasets, and whether the chromosomes that explain most variation harbor any candidate genes.

The result found in this study could also be used to weight the relationships from different chromosomes differently in genomic prediction, i.e. the G matrix based on genotypes of each chromosomes are weighted by the relative variance explained by that chromosomes. This could result in better predictions than GBLUP, or Bayesian genomic prediction methods that weight individual loci rather than chromosomes. The latter would in principle be more precise, but in practice it may not because of the difficulty to estimate the variance explained by individual loci.

CONCLUSION

Most of the additive genetic variance for post weaning weight can be explained using genotypic information from the Illumina OvineSNP50 BeadChip. Decomposition of additive genetic variance due to genomic relationships into different chromosomes showed that the additive genetic variance explained per chromosome is related with the chromosome length but significant differences were found between chromosomes.

LITERATURE CITED

Coster, A. (2012). pedigree: Pedigree functions. R package version 1.4. http://CRAN.R-project.org/package=pedigree

Gilmour A.R., Gogel B.J., Cullis B.R. et al. (2009). ASReml User Guide Release 3.0. *VSN International Ltd,* Hemel Hempstead, UK.

Goddard, M. E. (2012). *Anim. Prod. Sci*. 52:73-77

Habier D., Fernando R.L. and Dekkers J.C.M. (2007). *Genetic.* 177:2389–2397

Haile-Mariam, M., Nieuwhof, G.J., Beard, K.T. et al. (2013). *J. Anim. Breed. Genet*. 130: 20–31

Jensen, J., Guosheng, S. and Per, M. (2012). *BMC Genetics*. 13:44

Meuwissen, T.H.E., Hayes B.J., Goddard, M.E. (2001). *Genetics*. 157:1819–1829

van der Werf, J.H.J.,. Kinghorn B.P and Banks R.G. (2010). *Anim. Prod. Sci*. 50:998−1003

VanRaden P.M. (2008). *J. Dairy Sci.,* 91*:*4414–4423.

Visscher P.M., Macgregor S., Benyamin B., et al. (2007). *Am. J. Hum. Genet*. 81:1104:1110

Yang J., Benyamin B., McEvoy B., *et al.* (2*0*10). *Nat. Genet.* 42:565–569

Yang J., Manolio T.A., Pascuale L.R., et al. (2010). *Nat. Genet*. 43:519–525