Compression Efficiency Relationship Matrix: Developing New Methods to Determine Genomic Relationships for Improved Breeding

N. J. Hudson[1], J. Kijas[1], L. Porto-Neto[1] and A. Reverter[1]
[1]CSIRO Animal, Food and Health Sciences, 306 Carmody Road, Brisbane, Australia

ABSTRACT: Understanding genetic relatedness between individuals, sire groups and breeds underpins genomic selection and GWAS. Here, we describe a new estimate of genetic relatedness using normalized compression distance (NCD). Clustering of Sheep breeds inferred by NCD broadly reflects SNP correlation using standard multi-dimensional scaling. The clustering appears consistent with country of origin and population history. For example, the 4 British sheep meat breeds (Poll Dorset, Southdown, Suffolk and White Suffolk) clearly cluster with each other, but separate to unrelated breeds (Border Leicester, Merino and Texel). We show that the compression-based relationship matrix (CRM) and the genomic relationship matrix (GRM) are closely related. The quadratic relationship between pairwise NCD (CRM) and pairwise SNP correlation (GRM) implies CRM will perform better with closely related individuals, while the converse is true for GRM. For example, CRM resolves Merino from Poll Merino where GRM cannot.
Keywords: Genetic relationship matrix; Information compression; Sheep

Introduction

More accurate animal relationship measures have the potential to accelerate artificial selection for improved genetics and refine methods for gene discovery. Genetic relatedness is currently estimated by a combination of traditional pedigree-based approaches (ie. the so-called NRM for Numerator Relationship Matrix) and, given the recent availability of genetic information, Genomic Relationship Matrices (GRM). GRM are essentially computed by genome-wide SNP correlation among all pairwise individuals.

The aim of this project was to test a new method to infer SNP-based relationships which we hypothesised might give more accurate measures of genetic relationships. The new method clusters by information Compression Efficiency (CE) and the output can be considered a Compression Relationship Matrix (CRM). CE has previously been used to infer phylogenetic relatedness through analysis of mitochondrial DNA (Li et al. (2001)), and also to successfully cluster gene expression data (Nykter et al. (2008)), not to mention languages and even musical genres (Cilibrasi and Vitanyi (2005)). Given this promise, we aimed to explore its utility in clustering genotype data. In the particular context of genomic SNP, CE reflects patterns in both allele order and proportion that are known to differ systematically between breeds.

Other than our own preliminary explorations (Hudson et al. 2014)), the concept of clustering on compression efficiency has not previously been applied to whole genomes.

Materials and Methods

Animal Resources. A collection of 119 Australian sheep industry sires, from 8 breeds, were collected as part of a sire validation resource used to evaluate the accuracy of genomic selection schemes. The animals were collected as part of the SheepGENOMICS program (Meat and Livestock Australia, MLA project code SG.117) and genotyped in 2013 using the Illumina Ovine HD containing 563,387 SNP. The sires have been widely used in the meat and wool industries and had semen available inside semen holding collection centers. Sires were prioritized for sample collection and genotyping that had been used in AI across a number of sites to enrich for sires with high accuracy EBV. The sire genotypes were used in a published study of genotype imputation that also includes a detailed analysis of the genomic relationship matrix between individuals (Hayes et al. (2012)).

NCD and GRM. The pairwise Normalized Compression Distances (NCD) (Cilibrasi and Vitanyi (2005)) were computed on a representative subset of the data. NCD is an approximation of a non-computable similarity metric called Normalized Information Distance (NID). In brief, NCD is way of measuring the similarity between two objects. The principle is that NCD will award short distances to highly related sequence, on the grounds that a compression gain based on shared patterns is made when two similar files are concatenated, but not when two dissimilar ones are. We used the gzip application tool of UNIX systems (http://www.gzip.org) as our real world compressor for SNP data. The application gzip is based on DEFLATE, a lossless data compression algorithm originally described (Ziv and Lempel (1977)). In detail, the formula for the computation of NCD between two individuals *x* and *y* based on their respective SNP genotype sequence is as follows:

$$NCD(x,y) = \frac{Z(xy) - \min\{Z(x), Z(y)\}}{\max\{Z(x), Z(y)\}}$$

In the above formula, *Z(xy)* represents the size of the compressed file containing both SNP genotype sequences to be compared and *Z(x)* and *Z(y)* is the size of the compressed file containing the isolated SNP genotype sequences for *x* and *y*, respectively.

The GRM was computed using the method of (VanRaden (2008)):

$$GRM = \frac{\mathbf{ZZ}^T}{2\sum p_i(1-p_i)}$$

where **Z** is a matrix that relates SNP alleles to individuals and p_i is the frequency of the second allele for the *i*-th SNP. $\mathbf{ZZ}^T$ represents the number of shared SNP alleles among two individuals and the division of $\mathbf{ZZ}^T$ by $2\sum p_i(1-p_i)$ aims at scaling the GRM to make it analogous to the NRM.

Both, CRM and GRM can be considered as 'distance' matrices. Accordingly, we used multi-dimensional scaling (MDS) to cluster individuals and populations in the xy-Cartesian coordinates.

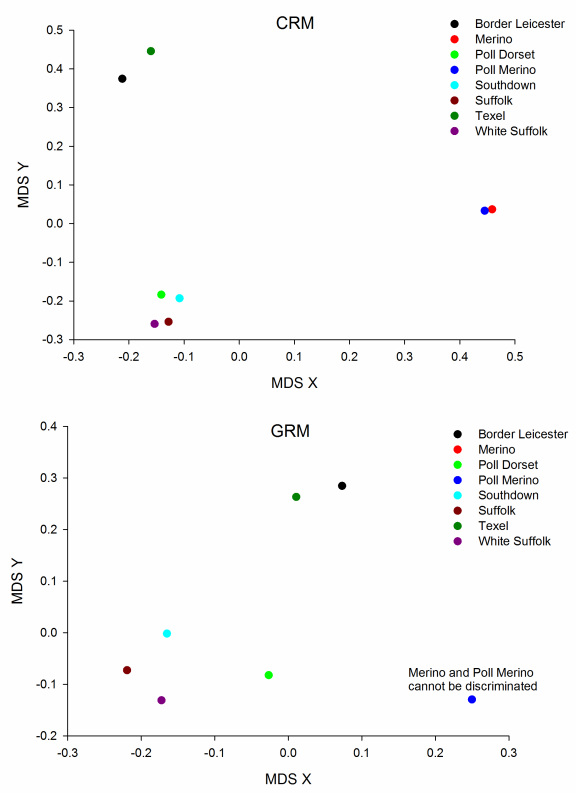

Figure 1: Multidimensional scaling of sheep breeds as defined by NCD (upper panel) and GRM (lower panel).

Results and Discussion

The CRM and GRM are globally similar resulting in 3 main breed clusters in both cases (**Figure 1**). In both CRM and GRM the 4 British sheep meat breeds (Poll Dorset, Southdown, Suffolk and White Suffolk) clearly cluster with each other, but separate to the Merino (wool breed) on the one hand, and the Texel (meat breed from the Netherlands) and Border Leicester (milk and wool) on the other. The major difference is that the CRM clusters the British meat breeds more tightly than does the GRM. A second difference is that the CRM is able to resolve the Merino from the Poll Merino while, at least using this particular set of sires, the GRM cannot. Future work will test whether this apparently enhanced sensitivity with closely related animals applies more generally.

The frequency histogram of pairwise NCD (having 7,021 pairs from 119 individuals) correctly reflects the large number of distant breeds and small number of related breeds (**Figure 2** top panel). The distant breeds are captured in the high NCD (large distance with NCD > 1.05) values, whereas the related breeds are captured by the discrete peak in smaller NCD values (NCD < 0.95). The scatter plot between NCD and GRM reveals a quadratic relationship between both metrics (**Figure 2** bottom panel).

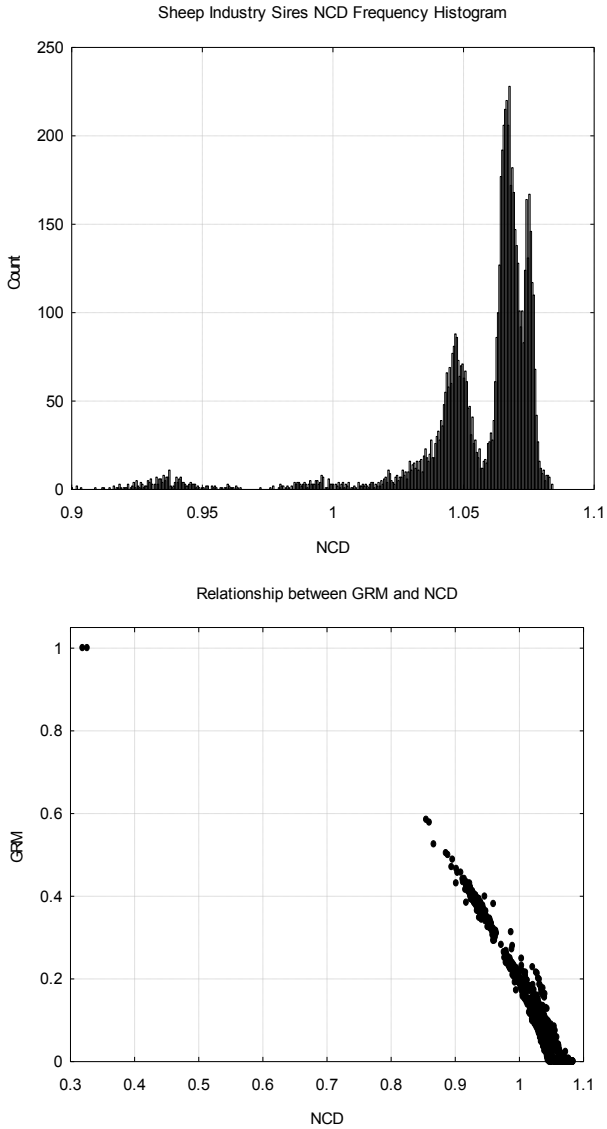

Figure 2: Frequency histogram of pairwise NCD values across all 119 sires (7,021 pairs) reflects breed similarities (top panel). NCD and GRM pairwise values possess a quadratic relationship.

Conclusion

Shared patterns of allele composition can be used to infer genomic relationships at the individual, sire group and breed levels. Both high SNP correlation (GRM) and compression based measures (CRM) are based on extent of haplotype sharing so a close relationship is unsurprising.

NCD shows merit as an alternative or complementary measure of genomic relatedness to SNP correlation based GRM. NCD clusters only when data is similarly compressible for the same reasons.

These compressible SNP patterns reflect individual properties like genome-wide heterozygosity and runs of homozygosity, in addition to population-level properties like linkage disequilibrium. Collectively, these features have implications for inbreeding, population structure and the identification of signatures of selection

The quadratic relationship between NCD and GRM implies the NCD is particularly sensitive in discriminating closely related individuals. This appears to be borne out by the CRM's ability to resolve Poll Merino from Merino despite high genetic similarity where GRM cannot.

The strong performance of NCD in the Sheep Industry Sires data, which is a high density SNP data set, implies the method will scale well with even larger data sets.

Acknowledgments

We wish to thank Meat and Livestock Australia (MLA) for supporting this work under project B.BSC.0344.

Literature Cited

Browning, B.L., and Browning, S.R. (2009). Am. J. Hum. Genet. 84: 210-223.

Cilibrasi, R., and Vitanyi, P.M.B. (2005). IEEE Trans. Inform. Theory. 51: 1523-1545.

Hayes, B.J., Bowman, P.J., Daetwyler, H.D. et al. (2012). Anim. Genet. 43: 72-80.

Hudson, N.J., Porto-Neto, L.R., Kijas, J. et al (2014). BMC Bioinformatics. 15:66.

Li, M., Badger, J.H., Chen, X. et al. (2001). Bioinformatics. 17(2): 49-54.

Nykter, M., Price, M.D., Aldana. et al. (2008). Proc. Natl. Acad. Sci. 105(6): 1897-1900.

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

White, J.D., Allingham, P.G., Gorman, C.M. et al. (2012). Anim. Prod. Sci. 52: 157-171.

Ziv, J., and Lempel, A. (1977). IEEE Trans. Inform. Theory. 23: 337-343.