final report

Project code: B.BSC.0344

Prepared by: Antonio Reverter
Commonwealth Science and Industrial Research Organisation

Date published: September 2015

PUBLISHED BY
Meat and Livestock Australia Limited
Locked Bag 991
NORTH SYDNEY NSW 2059

Compression efficiency relationship matrix: accelerating artificial selection and gene discovery

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Abstract

Please refer to exective summary.

Executive Summary

The ability to accurately infer animal relationships through shared genetics underpins our ability to perform genomic selection and interpret GWAS studies. These in turn drive the speed of artificial selection and the discovery of genes that contribute to commercial traits. In previous milestone reports we explored the inference of cattle and sheep population relationships' using a new similarity metric called Normalised Compression Distance (NCD). This approach yields a Compression Relationship Matrix (CRM). Like existing genetic relationship analyses based on correlation such as the genomic relationship matrix (GRM), the new metric quantifies similarity between numerical patterns in SNP data - that is, allele composition and order shared (to varying extents) by genome pairs. Not surprisingly, we previously found a very high concordance between CRM and GRM, and any genetic ranking made by the two methods would be very similar. A striking finding was that the new CRM approach can genetically discriminate very closely related individuals (such as half-sibs versus full sibs) in sheep and cattle populations where GRM cannot.

In this final report we focus on a deeper exploration of the genetics of yearling weight in Brahman (BB) and Tropical Composite (TC) cows. Using the latest 71K Indicus SNP chip, we explore heritability and genetic parameters under the usual assumptions. Further, we systematically explored the impact of the 3 different relationship matrices (NRM, GRM and CRM) in isolation and all combinations therein. Surprisingly, we find that an adapted version of the CRM reduces the 'missing heritability' associated with yearling weight in both breeds. The NCD output first needs to be mapped in such a way that 1) it takes advantage of the high sensitivity of NCD but then 2) grounds the output more strongly in the biology of genetic inheritance via meiosis (~0.5 sharing for full sibs, ~0.25 for half-sibs etc). Finally, a sliding window-based application of the compression approach recovers regions of evolutionary interest between the two populations. Some overlap with established signatures of selection. The remainder presumably reflect bottlenecks and other population history phenomena.

Our meeting of the 4 research objectives can be summarised as follows:

1) We have used compression efficiency to accurately infer animal to animal genetic relatedness. We ground-truthed our new output based on the known biology of the 4 populations in question (2 cattle and 2 sheep), and compared the independent predictions made by GRM. Several lines of evidence imply CRM is particularly sensitive at discriminating very closely related individuals.
2) We have used compression efficiency to verify parentage. We found that sire groups from the Faulkner sheep flock could be successfully clustered based on compression efficiency producing 3 main clusters that correctly reflect breed-level inter-relatedness.
3) We have used CRM to predict Estimated Breeding Values. We used the relationships predicted by CRM to estimate genetic paramaters for BB and TC cattle. A version of CRM performs very well, explaining more genetic variance, exhibiting a reduction in missing heritability and yielding an increase in phenotype accuracy compared to not only NRM but also GRM.
4) We have published one manuscript in *BMC Bioinformatics* (Hudson et al 2014b), presented at the WCGALP 2014 conference (Hudson et al 2014a), and have another manuscript in preparation.

Table of Contents

1 Background

1.1 Building more accurate animal relationship matrices to accelerate genetic progress

More accurate animal relationship measures have the potential to accelerate artificial selection for improved genetics and refine methods for gene discovery. By accounting for patterns of relatedness within and between families they lay the foundation for more robustly connecting genotype to phenotype. Genetic relatedness is currently estimated by a combination of traditional pedigree-based approaches (NRM for Numerator Relationship Matrices) and, given the recent availability of genetic information, Genomic Relationship Matrices (GRM). Because meiotic recombination is stochastic, GRM can give more precise estimates of genetic relatedness than basic pedigree information as the latter makes simplifying assumptions (Van Raden 2008). For example, while we know full sibs will share ~50% of their DNA, and half-sibs 25% - simple pedigree is unable to account for the exact % shared, or indeed which DNA segments comprising the % shared have actually been inherited. Moreover, DNA markers associated to QTL because of LD and linkage are expected to erode during successive meioses at a slower rate than pedigree relationships. This increases their utility across generations (Wolc et al. 2011). Overall, this suite of advantages has increased the attractiveness of SNP chips in genetic improvement programs.

GRM are essentially computed by genome-wide SNP correlation among all pair wise individuals. These correlations exploit numerical patterns shared in common between two genotypes. However, it is an open question whether correlation is best placed to cluster SNP genotype data given 1) any non-linear relationships are poorly characterised or undetected by correlation and 2) in assessing a numerical system with a small 3 letter alphabet (0,1,2) and long complex data strings it is not clear what is the best mathematical approach for quantifying relationships. We conjecture there is unexplored potential to characterise alternative and / or complementary measures of relatedness through competing pattern recognition approaches.

Here, our major aim was to test the Normalised Compression Distance (NCD) method to infer SNP-based genetic relationships. The basic principle of NCD is that information in one data source can be used to compress the information in a second data source. If the compression gain is strong, then the two files are deemed to be closely related and are awarded a short distance. Given this approach, NCD will award a short distance to the ‘genomes’ 2222222222 and 2222222222 but not to 0000000000 even though those 3 genomes possess the exact same isolated compression efficiency (i.e. ‘a digit x repeated 10 times’). Applying this process systematically across a genotyped population can be used to build a Compression Relationship Matrix (CRM), analogous to a GRM.

In terms of precedent, compression distance has already 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 in a range of data types possessing variable mathematical structure, we aimed to explore the utility of CE in clustering genotype data. In the particular context of genomic SNP, CE can be seen to reflect patterns in both allele order and proportion that are known to differ systematically between breeds. These patterns would include, but not be limited to, genome-wide heterozygosity and runs of homozygosity. Like correlation, compression efficiency is a hypothesis-free pattern recognition tool. It can exploit very complex shared patterns that do not need to be defined *a priori*. The inferred relationship matrices can be ground-truthed in

the normal manner – that is, by computing estimated breeding values and predicting genetic merit for complex phenotypes.

In some preliminary research we explored the application of NCD to a sheep population representing multiple breeds (Hudson et al. 2014a) and a sheep population with known sire groups and a half-sib population structure (*unpublished data*). Further, we previously utilised a less sophisticated - but still informative - measure of basic CE expressed versus heterozygosity (Hudson et al. 2014b) across several species including human. Other than our own preliminary studies, the concept of clustering on CE has not previously been applied to SNP genotypes.

Here, we identified 2 well characterised animal populations of commercial relevance to Australian agricultural production that have matching phenotype data. These populations have defined structure, particularly the presence of both full-sib and half-sib individuals. Further, there is the overarching indicine versus taurine breed contrast given these breeds were independently domesticated and last shared a common ancestor more than 200,000 years ago (McTavish et al.2014). We compared Compression based BLUPS (CBLUPS) with GBLUPS and PBLUPS for yearling weight, a complex phenotype of moderate heritability. Collectively, the nature of the clustering, the proportion of missing heritability and the phenotype prediction accuracies can be used for benchmarking.

2 Project Objectives

2.1 Objective 1. A more accurate method for inferring animal to animal genetic relatedness. The new CRM will be compared to pedigree-based (NRM) and SNP genotype based (GRM) approaches.

2.2 Objective 2. A quicker tool to ascertain pedigree errors and consistency (eg parentage verification).

2.3 Objective 3. The use of CRM in the development of a more accurate prediction of EBV.

2.4 Objective 4. Publish research outcome in high impact scientific publication(s).

3 Methodology

3.1 Brahman and Tropical Composite animal relationships

3.1.1 Comparing CEh, NCD (CRM) and GRM

We used 817 Brahman (BB) and 1,028 Tropical Composite (TC) cows genotyped for 71,726 SNPs corresponding to the GGP Indicus HD chip (http://www.neogeneurope.com/Agrigenomics/pdf/Slicks/NE_GeneSeekCustomChipFlyer.pdf). This SNP chip was recently developed by GeneSeek and specifically focuses on SNP highly polymorphic in Indicus cattle.

Firstly, we compared the animal-to-animal relationship that can be ascertained with this genotype data using the CEh approach described in Hudson et al (2014b). This approach computes the compression efficiency (CE) for each genotype file and expresses it against whole genome heterozygosity.

The CEh formula is as follows:

$$CE = \frac{SB - SA}{SB}$$

Where *SB* and *SA* indicate the genotype file size before and after compression, respectively. The underlying principle of CEh is the same as for any genetic clustering method. The closer the match in the numerical patterns present in two genotype files, the close the inferred genetic relationship between them. No attempt is made to discriminate DNA segments identity by descent (IBD) from those merely identical by state (IBS).

The implication of using the new Indicine SNP chip is a greater amount of heterozygosity (particularly in the BB population) coupled with a greater compression efficiency. Regular runs of heterozygosity (strings of 1s) being a likely source of compressible pattern. This finding is equivalent to what was observed in Figure 4A of Hudson et al (2014b) for the Hereford cattle using the 750K HD chip. Namely, higher heterozygosity coupled with high compression efficiency.

The main weakness of CEh is that 2 genotype strings can compress the same and possess the same heterozygosity despite being different (e.g. 0000000000 and 2222222222). It is not clear how common this phenomenon is in real cattle data, but it has the potential to confound some of the observed clustering.

The NCD analysis that forms the basis of the two CRM explored here accounts for this error. It only clusters those genotype files who compress the same *for the same reasons*. That is, 0000000000 will now only cluster with 0000000000 and not 2222222222.

As previously reported (Hudson et al. 2014a), the formula for the computation of NCD between two individuals *x* and *y* based on their respective SNP genotype sequence is as follows:

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

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 with the isolated SNP genotypes for *x* and *y*, respectively.

On the other hand, the construction of the GRM is based correlation between genotype profiles and computed according to methodology developed in Van Raden et al. (2008). This is the standard approach for genomic selection approaches:

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

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 numerator relationship matrix (NRM) obtained based on the pedigree information.

3.1.2 Genetic parameter estimation

For genetic parameter estimation and genetic evaluation, 12 models were explored: one each (i.e. four) with NRM, GRM, CRM1 and CRM2 as the only random effects, and then every combination of the various multiples. We have used the phenotype of yearling weight (YWT) as an exemplar complex trait of intermediate heritability.

3.1.3 Signatures of selection

In order to find signatures of selection and regions of evolutionary interest, we next applied a sliding window version of compression efficiency as previously described in Hudson et al (2014b). In brief, the population level CE of non-overlapping windows was computed for both BB and TC cows, corrected for heterozygosity (CEh) and Z-score normalised (CEhZ). In brief,1,435 windows of 50 consecutive SNPs were computed across the 71K SNP.

4 Results

4.1 Brahman and Tropical Composite animal relationships

4.1.1 Comparing CEh, NCD (CRM) and GRM

In Figure 1 each CEh point represents either a single BB (left panel) or TC (right panel) genome. Two genomes that cluster together can be assumed to share more genotype patterns in common that two genomes located further apart, and therefore are more likely to be related by descent.

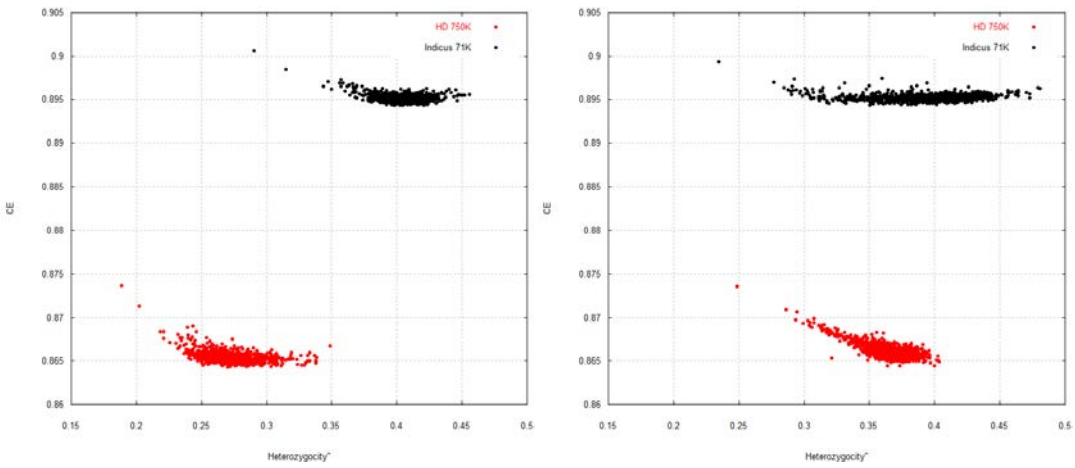

Figure 1: A comparison of CEh for BB (left panel) and TC (right panel) cows genotyped using both the HD chip (red dots) with 750K SNPs and the new 71K Indicus SNP chip (black dots). Heterozygosity is expressed on the x axis, and CE on the y axis

The summary data relating NRM with GRM and NCD is given in Tables 1 and 2. For instance, according to the GRM, the self-self relationship among the 817 BB cows averaged 0.996 (ranging from 0.928 to 1.670). The equivalent values according to the NCD averaged 0.118 (ranging from 0.112 to 0.123).

For relationship corresponding to NRM values of 0.25 (ie. those existing between half-sibs or between grand-parent and grand-offspring), the average GRM was 0.196 and 0.201 for BB

and TC, respectively. The average NCD for the same relationships was 0.997 and 0.987 for BB and TC, respectively.

Table 1. Summary statistics for BB cows compared pair wise using NRM values, correlation (GRM) and NCD (CRM).

		GRM				NCD			
NRM	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
0.0625	50	0.044	0.026	-0.006	0.111	1.050	0.011	1.025	1.071
0.1250	768	0.109	0.025	0.028	0.195	1.027	0.012	0.988	1.071
0.2500	8724	0.196	0.086	-0.108	0.386	0.997	0.033	0.904	1.117
0.3125	90	0.281	0.034	0.215	0.369	0.934	0.019	0.889	0.978
0.5000	201	0.288	0.068	0.167	0.473	0.957	0.034	0.782	1.007
1.0000	817	0.996	0.039	0.928	1.670	0.118	0.002	0.112	0.123

Table 2. Summary statistics for TC cows compared pair wise using NRM values, correlation (GRM) and NCD (CRM).

		GRM				NCD			
NRM	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
0.03125	832	0.004	0.032	-0.077	0.158	1.075	0.011	1.038	1.098
0.06250	2659	0.038	0.048	-0.137	0.172	1.052	0.014	0.986	1.095
0.12500	630	0.061	0.037	-0.049	0.149	1.025	0.020	0.966	1.072
0.25000	15190	0.201	0.092	-0.075	0.455	0.987	0.046	0.799	1.100
0.31250	316	0.066	0.028	-0.003	0.157	1.051	0.009	1.023	1.074
0.50000	683	0.229	0.042	0.103	0.509	0.994	0.023	0.764	1.032
1.00000	1028	1.000	0.056	0.874	1.401	0.118	0.002	0.113	0.124

The relationship between GRM and NCD for every pair of individuals is plotted in Figure 2. The top 4 panels point to a population sub-structure in both populations that is more complex than would be appreciated by analysis of either GRM or NCD in isolation. Plotting the two metrics simultaneously operates synergistically to give a fuller understanding of animal relatedness. More work is required to unravel the meaning of the clusters.

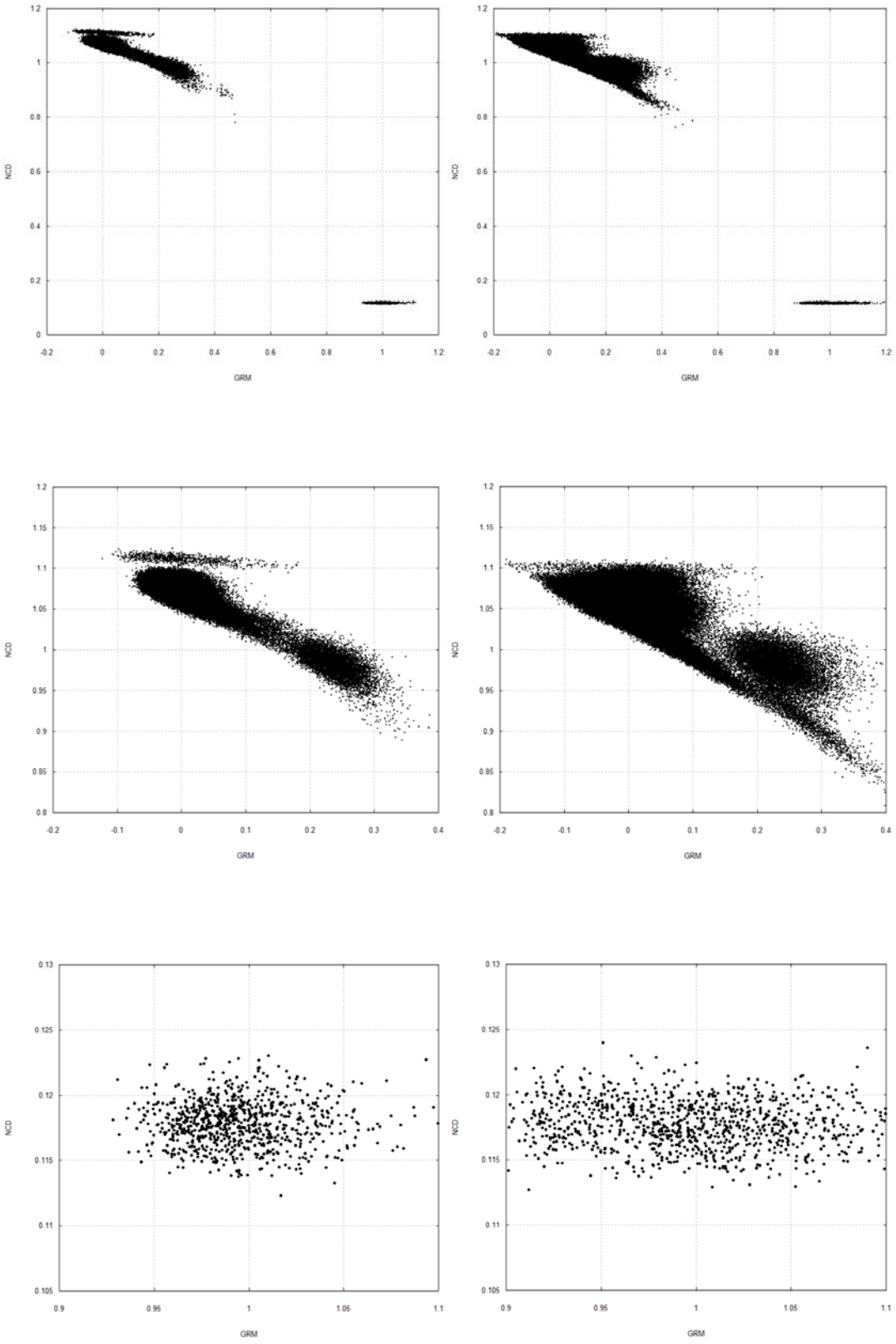

Figure 2. Comparison of GRM (x axis) versus NCD (y axis). Each point represents a pair of Brahman (BB; left panels) and Tropical Composite cows (TC; right panels). The top panels display the full parameter space. The middle panels zoom into the top left (all pairs other than self-self) while the bottom panels zoom into the bottom right (self-self pairs only).

The construction of the CRM from all the pair wise NCD values requires conversion of the compression "distance" to an equivalent "similarity." While distance and similarity are two sides of the same coin, in practice there are numerous ways of inter-relating them. In the previous milestone report we explored only 1 conversion method. Here, we explore an additional approach, thereby producing 2 different CRM.

The first method (CRM1) was outlined in the previous milestone report. It makes use of a universal conversion law previously derived by Shepard (Shepard, 1987). We use s = 2.5 e^-5d which was selected in an *ad hoc* fashion by confirming that the resulting "s" (similarity) covered the 0 to 1 range observed for correlation (and therefore GRM).

The second method (CRM2) is a new attempt to better ground the NCD output in established genetics - that is, an expectation of relatedness of 0.5 for full sibs and 0.25 for half-sibs (as governed by the laws of inheritance and the likely molecular outcomes of meiosis when applied to a diploid mammalian genome). It is a linear conversion method defined as follows:

If i = j (self) then s = Mean of d for all i=j divided by d

If i /= j then s = 0.75*(1 - (d-min)/(max-min))

This linear method has the appealing feature of yielding a value of ~1 for self-self pairs, including the spread around 1 reflecting inbreeding. The remaining values approximate 0.5 for full sibs, 0.25 for half-sibs and so on. As with GRM, but unlike NRM, these values are not hard coded but rather derived from the SNP data. Therefore, they give an estimate of the molecular genetic outcome of meiosis observed for each individual compared to its relatives.

The impact of the two NCD mapping approaches (CRM1 and CRM2) is apparent on Figure 3. Both versions of CRM are negatively related to NCD, because similarity in the inverse of distance. CRM2 produces a linear relationship whereas CRM1 is a non-linear exponential relationship. The linear relationship of CRM2 more closely resembles the genetic and phenotypic relationships observed between full sibs, half sibs and others. Consequently, it performs better in genetic parameter estimation.

The next step in the analysis is to compare the use of the CRM to the NRM and GRM for genetic parameter estimation.

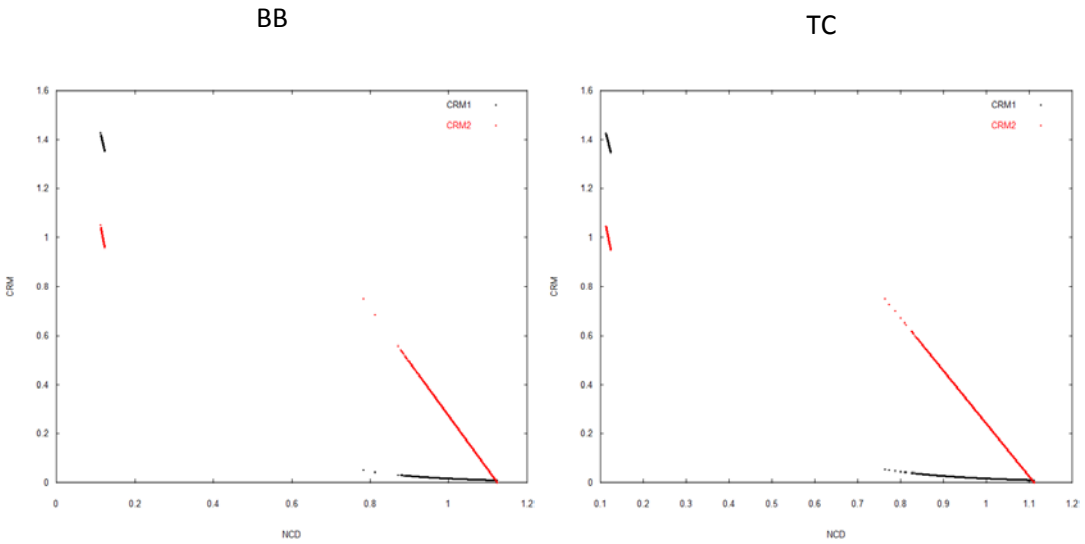

Figure 3. Relationship between NCD and CRM1 (black) and CRM2 (red). CRM1 bears a quadratic relationship whereas CRM2 is linear. This linearity explains the higher performance in computing genetic parameters, as it now resonates with biological expectation.

4.1.2 Estimating genetic parameters

Overall, the genetic parameter estimates are quite similar for GRM and CRM2.

Table 3. BB cattle estimates of variance components: Comparison between pedigree (NRM), Normalized Compression Distance (CRM1 and CRM2) and Genomic Relationship (GRM).

Model Effects	Ve	Vn	Vg	Vc1	Vc2	Vp
1. NRM	172.3642	174.8665				347.2307
2. GRM	167.3202		179.9430			347.2632
3. CRM1	161.9650			121.9581		283.9231
4. CRM2	168.6462				195.9485	364.5947
5. NRM+GRM	115.5917	113.9303	129.9348			359.4568
6. NRM+CRM1	106.5055	123.3128		78.3912		308.2095
7. NRM+CRM2	113.3747	118.5415			140.8651	372.7813
8. GRM+CRM1	102.9185		130.3848	75.1383		308.4416
9. GRM+CRM2	151.5553		103.5573		104.4451	359.5577
10. CRM1+CRM2	103.7444			76.7594	140.8459	321.3497
11. NRM+GRM+CRM1	79.1332	88.9443	99.2563	57.8128		325.1466
12. NRM+GRM+CRM2	82.3409	92.4328	98.8974		105.2992	378.9703

Ve = residual variance; Vn = genetic variance due to the pedigree NRM; Vg = genetic variance due to the genotype GRM; Vc1 = genetic variance due to the genotype CRM1; Vc2 = genetic variance due to the genotype CRM2; Vp = phenotypic variance.

Table 4. TC cattle estimates of variance components: Comparison between pedigree (NRM), Normalized Compression Distance (CRM1 and CRM2) and Genomic Relationship (GRM).

Model Effects	Ve	Vn	Vg	Vc1	Vc2	Vp
1. NRM	220.1370	207.5774				427.714
2. GRM	217.4226		212.8462			430.269
3. CRM1	195.7910			155.4020		351.193
4. CRM2	223.5068				222.1801	445.686
5. NRM+GRM	143.1428	143.7108	159.4054			446.259
6. NRM+CRM1	131.0090	149.6712		98.5048		379.185
7. NRM+CRM2	146.4022	146.6484			165.8576	458.908
8. GRM+CRM1	127.3081		159.2306	94.8720		381.411
9. GRM+CRM2	208.1053		120.8427		110.7054	439.653
10. CRM1+CRM2	129.0538			97.1270	168.0027	394.183
11. NRM+GRM+CRM1	97.5807	109.4418	122.6173	72.4005		402.040
12. NRM+GRM+CRM2	105.9689	116.9839	122.6360		125.9694	471.558

Ve = residual variance; Vn = genetic variance due to the pedigree NRM; Vg = genetic variance due to the genotype GRM; Vc1 = genetic variance due to the genotype CRM1; Vc2 = genetic variance due to the genotype CRM2; Vp = phenotypic variance.

A major new finding is that CRM2 explains greater genetic variation (Vc2) than the NRM and GRM in both BB and TC. Implementation of CRM2 in these populations and for this phenotype (yearling weight) using the latest Indicus 71K SNP chip would be expected to lead to slightly better breeding decisions and slightly more accurate phenotype predictions than either GRM or NRM.

Moreover, models 5, 6 and 7 allow for the estimation of the fraction of missing heritability (C_{miss}) using the formulae of Román-Ponce et al. (2014):

$$C_{miss} = 1 - \frac{\sigma_u^2}{\sigma_a^2 + \sigma_u^2}$$

Where σ_u^2 is the variance due to the genotype data (ie. either GRM or CRM1 or CRM2 in our context) and σ_a^2 is the additive genetic variance due to the pedigree (ie. the NRM in our context).

Table 5. Fraction of missing heritability by each model and population

Model Effects	BB	TC
NRM+GRM	0.467	0.474
NRM+CRM1	0.611	0.603
NRM+CRM2	0.457	0.469

In both populations the missing heritability is smallest for CRM2, implying it outperforms the GRM by a small margin.

4.1.3 Signatures of selection

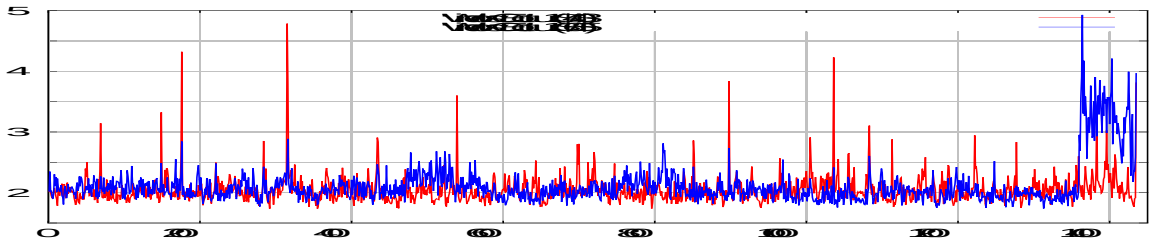

Figure 4. A genome-wide view of compression peaks in BB (red profile) versus TC cows (blue profile).

It can be seen that there are a number of shared peaks across the breeds, but across the genome the BB tend to be more extreme.

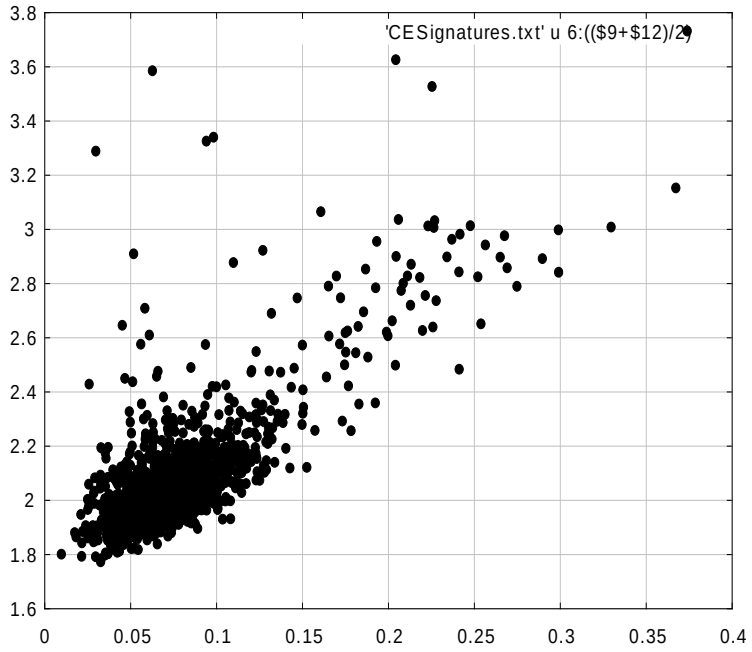

Figure 5. The positive relationship between CEhZ (y-axis) and F_{ST} (x-axis). Each dot represents a window of 100 consecutive SNPs. Windows in the top left quadrant are identified as important by CEhZ but not F_{ST}.

Table 3. Top 10 genomic regions according to CEhZ.

Chr	Start	Finish	Fst	BB Het	BB CE	CEhZB	TC Het	TC CE	CEhZ TC	CEhZ mean
5	56181411	56606039	0.2046	0.195814	0.937073	4.78553	0.365525	0.899825	2.46173	3.62363
3	33168941	33537089	0.0630	0.210086	0.907961	4.32185	0.310233	0.882277	2.84392	3.58289
30	3946692	5485712	0.2258	0.397405	0.842681	2.12046	0.185798	0.915923	4.92967	3.52506
5	56611536	57785345	0.0986	0.240024	0.909353	3.78859	0.307802	0.888533	2.8867	3.33765
19	56594816	57094538	0.0944	0.211579	0.894953	4.22988	0.363191	0.877623	2.41642	3.32315
16	25640184	25933132	0.0301	0.234394	0.899729	3.83853	0.323171	0.883555	2.73402	3.28627
30	44377832	46583501	0.3676	0.285165	0.961072	3.37023	0.330642	0.968948	2.9305	3.15036
30	87810219	90333692	0.1610	0.483696	0.926801	1.91608	0.226128	0.951877	4.20946	3.06277
30	132275975	133916723	0.2062	0.405018	0.840425	2.07503	0.227821	0.909705	3.99307	3.03405
30	7278887	8945002	0.2272	0.446316	0.841721	1.88593	0.215739	0.900454	4.17381	3.02987

Assessing these gene regions highlighted by CEhZ, including exploring in detail the genes they harbour, requires more work. However, we wish to highlight that the top region by CEhZ

includes several inhibin encoding genes (*INHBC*, *INHBE*). Inhibin has previously been documented to be a candidate gene for fertility in BB bulls (Fortes et al 2012).

5 Discussion

At the start of the discussion we will summarise our findings for the whole project against the agreed objectives. In the final part of the discussion, we will expand on the main findings and attempt to draw the threads of the various analyses together.

Objectives

1. A more accurate method for inferring animal to animal genetic relatedness. The new CRM will be compared to pedigree based (NRM) and SNP genotype based (GRM) approaches.

Overall, we explored two sheep and two cattle populations to assess the ability of compression efficiency to infer animal to animal relatedness. The populations explored were BB and TC cattle, the Faulkner Sheep Mapping Flock (FMFS) and the Sheep Industry Sires (full details of population and data can be found in Milestone Report 1). These populations possess a range of familial and breed structures. In all cases compression yields results that resonate with our biological understanding of those populations and the conventional GRM based measures. This provides a strong test of initial ground-truthing. Several indications of compression possessing particularly high sensitivity in resolving very closely related individuals were identified. This included the clear separation of Merino from Poll Merino in both sheep populations. Equally, in the second milestone report, we documented an observation that CRM can separate half-sibs from sibs in circumstances where GRM cannot, a phenomenon particularly apparent in Merino sheep. We have speculated this enhanced sensitivity reflects the fact that NCD is a ‘distance’ measure whereas correlation is a ‘similarity’ measure – the former attempts to separate while the latter attempts to relate. Finally, the implementation of compression efficiency based measures appears robust to the density of the SNP panel used. This implies the method should scale well in the future with the emergence of even higher densities SNP chips or indeed the use of whole genome sequence. Looking the other way, the method should also apply well via application through a much smaller, economically viable targeted SNP panel.

2. A quicker tool to ascertain pedigree errors and consistency (eg parentage verification).

Faulkner Sheep Genomics Mapping Flock Sheep (FMFS) sire groups can be successfully clustered based on NCD. Broadly speaking, the CRM and GRM based clusters are comparable. In both cases, they form 3 main clusters that correctly reflect genetic inter-relatedness at the breed level: (A) Merino with Poll Merino; (B) Poll Dorset x White Face Suffolk with White Face Suffolk; (C) Coopworth with East Freisland x Border Leister and Border Leister. Interestingly, in another example of the high sensitivity of compression efficiency, the CRM is able to discriminate the two Poll Merino sire groups from each other whereas the GRM cannot.

3. The use of CRM in the development of a more accurate prediction of EBV.

We ground-truthed the CRM by assessing its ability to estimate genetic parameters for BB and TC cows using a complex trait, yearling weight. There is very high concordance between CRM and GRM, and any genetic ranking made would be very similar. We discovered a version of the CRM performed very well in estimating genetic parameters. It explained more genetic variance, exhibited a reduction in missing heritability and yielded an increase in phenotype accuracy when compared to not only NRM but also GRM. The full methodological and statistical details relating to this finding are documented in this final report.

4. Publish research outcome in high impact scientific publication(s).

1. Our first manuscript documenting population clustering by a genome-wide compression metric and the application of a sliding window for identification of signatures of selection was published in BMC Bioinformatics in 2014:

Hudson NJ, Porto-Neto L, Kijas J, McWIlliam S, Taft R and **Reverter A** (2014). Information compression exploits patterns of genome composition to discriminate populations and highlight regions of evolutionary interest. *BMC Bioinformatics* **15**:66.

2. We presented the outcomes of milestone report 2 at the WCGALP conference in Vancouver 2014, focussing on the sheep industry sires component of the analysis:

Hudson NJ, Kijas J, Porto-Neto L and **Reverter A** (2014). Compression efficiency relationship matrix: developing new methods to determine genomic relationships for improved breeding. Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada.

3. A further manuscript, documenting the greater genetic variance explained, the reduced missing heritability and the higher phenotype accuracies documented in this report is currently in draft form. It will be submitted to *Genetics Selection Evolution* by the end of March this year.

General discussion

It is well established that shared patterns of allele composition can be used to infer genomic relationships. Both high SNP correlation (GRM) and compression based measures (CRM) are based on extent of haplotype sharing so the close relationship we observed between the two approaches is unsurprising. NCD shows merit as an alternative or complementary measure of genomic relatedness to SNP correlation based GRM. Overall, short NCD distances reflect high co-sharing of individual properties like genome-wide heterozygosity and runs of homozygosity. Collectively, these genomic features have implications for inbreeding, population structure and the identification of signatures of selection.

The relationship between NCD and GRM implies the NCD is particularly sensitive in discriminating closely related individuals. This is borne out by the NCD's ability to separate some full sibs from half sibs in circumstances where GRM cannot (Hudson et al 2014a). In a recent application of the NCD method, we examined its application in a high density SNP data set in sheep. Here, two Poll Merino sire groups could be resolved from each other, and Poll Merino individuals could be resolved from Merino individuals by NCD but not by GRM (Hudson et al 2014a). We conjecture that these observations may reflect NCD's reliance on

'distance' which enforces separation, versus correlation's use of 'similarity' which establishes connection. The strong performance of NCD in these high density data implies the method will scale well with even larger data sets (such as very high density SNP panels in other species and even entire genome sequence).

Any measure of relationship (whether correlation, compression efficiency or other) needs to be clearly grounded in known genetics for it to provide meaningful genetic parameter estimation. That is, it must yield the expected relationship values of ~1 for self-self pairs, ~0.5 for full sibs and ~0.25 for half-sibs. These values are implicit for correlation based measures, but not for NCD. This means that to take advantage of the sensitivity of NCD in a genetic context, we first had to transform distance into similarity in a biologically appropriate manner. The linear transformation we used for CRM2 was far superior in this regard than the quadratic transform of CRM1. CRM1 used the universal distance to similarity method of Shepard (1987). The ground-truthing of CRM2 was apparent through the observed increase in genetic variance explained, the corresponding reduction in missing heritability and the increase in phenotype accuracy when compared to GRM and NRM. In terms of real world application, it is not clear what effect selection using CRM2 would have on actual rate of genetic gain - selection progress is proportional to heritability and selection intensity, neither of which can be reliably known.

We previously found that by sliding a population-level window along genomes we could identify regularities based on extent of haplotype sharing. In human populations this approach identified classic signatures of selection such as European eye and skin colour, asian hair texture and European and Masaai Kenyan lactase persistence. We ran this approach here to compare the BB and TC populations and identified a number of substantial peaks present in one population but not the other. Notable among these was the identification of several inhibin encoding genes which have been previously identified as associated with fertility in tropically adapted cattle.

6 Conclusions/Recommendations

6.1 Conclusions

6.1.1 Summary, overall progress and recommendations

In this final milestone report we have detailed our progress in using the information theory concept of compression efficiency to cluster animals, build relationship matrices, estimate genetic parameters and identify signatures of selection. The major findings of previous milestone reports were the high sensitivity with which NCD could resolve closely related individuals, including circumstances where correlation failed. This resolution was best exemplified by Poll Merino versus Merino sire groups and Merino half-sibs versus full sibs. In the earlier milestone reports we were unable to harness this extra sensitivity because the non-linear mapping of distance to similarity we used did not adequately reflect the physical consequences of meiosis. Consequently, our previous calculations of genetic parameters and phenotype accuracies were unconvincing.

In this final report we have established a method that appears to overcome this problem. By mapping the NCD data in a genetically more realistic manner (CRM2) we have been able to reduce the 'missing heritability' associated with the complex trait of yearling weight in both

populations studied and increase the accuracy of phenotype prediction compared to not only NRM but also GRM.

Overall progress of the project

The project has been completed in line with the agreed milestones.

The original pilot work exploring the ability of Compression Efficiency to discriminate populations and highlight regions of evolutionary interest has been published by *BMC Bioinformatics* (Hudson et al. 2014b). Post publication this manuscript has been identified by BMC as being “highly accessed.”

The work from milestone 1 was presented as a poster at the WCGALP congress in Vancouver, August 2014.

A 3rd publication based on the analysis in this final report is currently being prepared for submission to a journal. We will also compute phenotype prediction accuracies based on the new CRM relationships and compare them to the NRM and GRM predictions.

Recommendations

Future work could explore several different CRM, including different mapping options for NCD. The subsequent CRM’s can be compared for their ability to estimate genetic parameters, and for their ability to reduce the missing heritability. We believe a set of populations with different structures and phenotypes of varying heritability’s should be exploited, in an effort to generalise the findings made in this project.

The cattle dairy and chicken meat industries both run genomic prediction as part of their modern breeding strategies. In both cases they work with highly inter-related animal populations with a very small effective population size. We anticipate this new NCD method, which has proven to be sensitive in discriminating closely related individuals, having potential value in those two production environments.

7 Key Messages

7.1 CE, NCD and GRM

Both high SNP correlation (GRM) and compression based measures (CRM) are based on extent of haplotype sharing so the close relationship we observed between the two approaches is unsurprising. NCD shows merit as an alternative or complementary measure of genomic relatedness to SNP correlation based GRM. Overall, short NCD distances reflect high co-sharing of individual properties like genome-wide heterozygosity and runs of homozygosity. Collectively, these genomic features have implications for inbreeding, population structure and the identification of signatures of selection.

7.2 Genetic parameters

The ground-truthing of CRM2 was apparent through the observed increase in genetic variance explained, the corresponding reduction in missing heritability and the increase in phenotype accuracy when compared to GRM and NRM. In terms of real world application, it is not clear what effect selection using CRM2 would have on actual rate of genetic gain - selection progress is proportional to heritability and selection intensity, neither of which can be reliably known

7.3 Signatures of selection

We ran the CEh sliding window approach here to compare the BB and TC populations and identified a number of substantial peaks present in one population but not the other. Notable among these was the identification of several inhibin encoding genes which have been previously identified as associated with fertility in tropically adapted cattle.

8 Bibliography

8.1 References

Cilibrasi R, Vitanyi PMB (2005). Clustering by compression. IEEE Trans Inform Theory. 51

Fortes M, Reverter A, Hawken R, Boloorma S, Lehnert SA (2012). Candidate genes associated with testicular development, sperm quality,and hormone levels of inhibin, luteinizing hormone and insulin-like growth factor 1 in Brahman bulls. Biol Reprod 87(3):58.

Hudson NJ, Kijas J, Porto-Neto J and Reverter A (2014a). Compression efficiency relationship matrix: developing new methods to determine genomic relationships for improved breeding. WCGALP Vancouver 2014.

Hudson NJ, Porto-Neto LR, Kijas J, McWilliam S, Taft R, Reverter A (2014b). BMC Bioinformatics: 15(66).

Klenk S, Thom D and Heidemann G (2009). The normalised compression distance as a distance measure in entity identification. ICDM Proceedings of the 9th Industrial conference on Advances in Data Mining: 325-337.

Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H (2001). An information based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17 149-154.

McTavish EJ, Hillis DM (2014). A genomic approach for distinguishing between recent and ancient admixture as applied to cattle. J Hered.

Nykter M, Price ND, AldanaM, Ramsay SA, Kauffman SA, Hood LE, Yli-Harjo O, Shmulevich I (2008). Gene expression dynamics in the macrophage exhibit criticality. PNAS 1897-1900.

Román-Ponce S, Samore AB, Dolezal M, Bagnato A and Meuwissen TH (2014). Estimates of missing heritability for complex traits in Brown Swiss cattle. Genetics Selection Evolution 46:36.

Shepard R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.

Van Raden, P.M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci 91, 4414-4423.

Wolc A, Arango J, Settar P, Fulton JE, O Sullivan NP, Preisinger R, Habier D, Fernando R, Garrick DJ, Dekkers JC (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet Sel Evol 43 23.

9 Appendix

9.1 Heading

There are no appendices.