DATA COMPRESSION: A NEW WAY TO INFER GENOMIC RELATIONSHIP MATRICES AND HIGHLIGHT REGIONS OF INTEREST IN COMMERCIAL LINES OF BROILER CHICKEN

N.J. Hudson[1], R. Hawken[2] R. Sapp[2] and A. Reverter[1]

[1] CSIRO Agriculture Flagship, Brisbane, Australia
[2] Cobb Vantress, Arkansas, USA

SUMMARY

Gene discovery relies on knowledge of animal relatedness. This in turn exploits correlation based measures of similarity now based on shared patterns of genome-wide single nucleotide polymorphism (SNP) genotypes. These comparisons are captured by the genomic relationship matrix (GRM). However, it is not clear whether correlation is the best way of quantifying those shared patterns. Here, we continue our exploration of whether one can build relationship matrices based on the concept of compression efficiency from Information Theory. Drawing on 4 commercial broiler lines, 2 lines based on growth and efficiency selected roosters, and 2 lines based on reproductive performance selected hens, we found that data compression clustered the lines by gender. Further, a sliding window version of the approach identified different gene regions apparently selected in male versus female lines. In males two prominent regions harboured *IGF-1* (Chromosome 1) and a cognate *IGF-1* receptor *INSR* (Chromosome 28). In the female lines, the reproductive hormone receptor *GNRHR* (Chromosome 10) and folate metabolism *FOLH1* (Chromosome 1) were prioritised.

INTRODUCTION

Gene discovery through genome-wide association studies (GWAS) and identification of signatures of selection require that population structure and relatedness can first be quantified and subsequently accounted for. Genetic relatedness is currently estimated by a combination of traditional pedigree-based approaches (Henderson 1975) and, given the recent availability of molecular information, the use of marker genotypes via the genomic relationship matrix (GRM) (Van Raden 2008). To date, GRM from SNP genotypes are essentially estimated using correlation.

Here, we continue our exploration as to whether the concept of compression efficiency from Information Theory can provide a complementary method for establishing patterns of genetic relatedness. The basic principle of Normalised Compression Distance (NCD) (Cilibrasi and Vitanyi 2005) is that if patterns of data in one genotype file can be used to compress shared patterns of data in the second genotype file, the two genotypes are considered related. Consequently, a short distance (high similarity) will be awarded. This process can be repeated across a genotyped population of animals to build a Compression Relationship Matrix (CRM) analogous to a GRM. This concept has previously been used by our group in both sheep and cattle populations where we have found that the NCD method can sensitively discriminate sire groups, breeds and indeed half-sibs from full sibs, in circumstances where GRM could not (Hudson *et al.* 2014 WCGALP). Moreover, we found CRM explained more genetic variance, reduced the missing heritability and yielded higher phenotype accuracies than GRM (*unpublished data*). Additionally, a preliminary version of the approach was able to cluster individual humans by ethnic group in a manner consistent with F_{ST} and known phylogeography (Hudson *et al*. 2014).

In this exploratory paper we assess the application of NCD to patterns of relatedness between 4 commercial lines of broiler chickens, *Gallus gallus domesticus*. We also use a genome-wide sliding window based on compression efficiency to identify possible signatures of selection present on a gender-specific basis.

MATERIALS AND METHODS

Populations and data resources. We used data from 988 chickens from 4 commercial lines of broilers – hereon in denoted as Lines A, B, C and D (Table 1). Individuals were selected from a much larger population of over 50,000 birds and based on full sib families to a near-balanced design of ~250 individuals per line.

Table 1. Summary of the 4 chicken lines used for this analysis

Line	Selection	Birds	Full-Sib Families	Females	Males
A	Female	204	14	167	37
B	Female	244	5	153	91
C	Male	254	18	195	59
D	Male	286	50	220	66

Two of the lines (A and B) are lines that have been generated for selecting genetically superior females – the selection focus being primarily on desirable reproductive traits. For male lines (C and D), the selection foci have been growth rate, muscle mass and feed efficiency. All animals were genotyped for 51,713 SNP (Groenen *et al.* 2009) distributed genome-wide.

Population clustering. We used NCD to compare pairs of individuals (x and y) from all 4 lines based on their respective SNP genotypes 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 concatenated 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. We used GZIP to perform the data compression.

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 (CE) as previously described in Hudson *et al.* (2014). This approach exploits the sensitive pattern recognition capability of CE to find haplotype blocks that occur in one population but not another. In brief, the population level CE of non-overlapping windows was computed separately for the 4 broiler lines, corrected for heterozygosity (CEh). We used non-overlapping sliding windows of 100 consecutive SNP. The experimental design made use of two 'independent' lines of male and female populations, whose output could be overlaid. This approach helps improve the signal to noise ratio for identifying *bona fide* signatures of selection, against background noise emerging from population bottlenecks and other phenomena.

RESULTS AND DISCUSSION

Population clustering. Self-Self pairs (panel A) possess a GRM of close to 1, with deviations above 1 representing extent of inbreeding. GRM and NCD are both in agreement that the lines cluster by gender comparison (Panel B). Female-male line comparisons in blue are awarded a low similarity and high distance, whereas male-male and female-female line comparisons are more closely related.

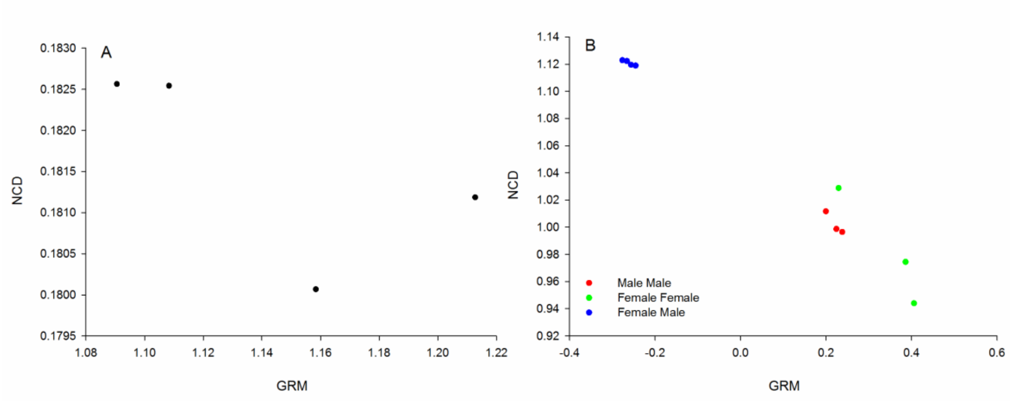

Figure 1. GRM and NCD values for the 4 chicken lines with the various gender comparisons colour coded.

Overall, there is a clear negative relationship between GRM and NCD because similarity (via correlation) is the inverse of distance (via NCD).

Signatures of selection. The genomes of all 4 lines were characterised by a large number of small peaks and a much smaller number of larger peaks. These outlier regions have particularly strong population-level scores in these regions. They would be predicted to potentially play an important role in providing the genetic basis for the phenotypes that have been selected in those populations. We manually explored the outlier regions that were gender-specific.

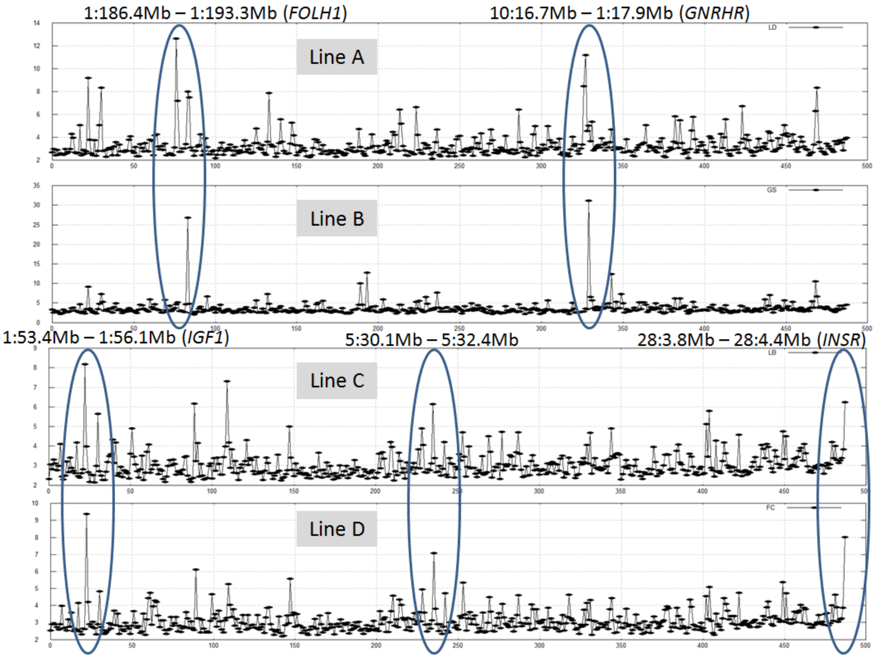

Figure 2. Compression efficiency (y-axis) of windows of 100 consecutive SNPs along the genome (x-axis) for the four chicken lines. Highlighted are the regions described in Table2.

Table 2. Regions captured by the compression efficiency of windows of 100 consecutive SNPs

Lines	Regions (Chr: Coordinates)	Example Genes in region	Total number of genes
Female	1:186.4 Mb – 1:193.3 Mb	***FOLH1***, *THRSP*	62
Female	10:16.7 Mb – 10:17.9 Mb	***GNRHR***	37
Male	1:53.4 Mb – 1:56.1 Mb	***IGF-1***, *MTERF*	45
Male	5:30.1 Mb – 5:32.4 Mb	mir-1718, mir-3532	19
Male	28:3.81 Mb – 28:4.44 Mb	***INSR***, *SIN3B*, *PEX11G*	28

In the two male lines the clear identification of two different regions containing serial components of a single functional pathway (*IGF-1* and one of its cognate receptors *INSR)* is particularly intriguing. The male lines, unlike the female lines, have been selected for increased muscle mass. IGF-1 is a well characterised master regulator of muscle mass whose molecular structure is similar to insulin. It mediates the anabolic effect of Growth Hormone (Barton 2006). This functional pairing (IGF-1 and INSR) is unlikely to occur by chance as IGF-1 is one of only three proteins to bind the insulin receptor. In an independent population of broiler chickens derived from Plymouth Rock and Cornish lines *IGF-1* had also been identified as a signature of selection (Stainton *et al*. 2015). In the female lines which have been selected for reproductive traits, we detected regions containing *GNRHR* (encoding the receptor for the reproductive hormone gonadotropin releasing hormone) and *FOLH1* (that hydrolases the vitamin folate).

Future work could fine map these genomic regions using a higher resolution (50 SNP) window, and sliding it one SNP at a time in an overlapping fashion to attempt to home in on the exact genes under selection. We have previously used this method to successfully home in on single genes across human populations, such as lactase persistence in northern Europeans and Masaai Kenyans (Hudson *et al.* 2014). The relationship matrices described in the first part of the manuscript could be 'ground-truthed' through estimation of genetic parameters, computation of missing heritability and calculation of phenotype accuracies for a phenotype of commercial interest in the broiler industry such as feed efficiency.

REFERENCES

Barton E.R. (2006). *Appl Physiol Nutr Metab* **31**:791.

Cilibrasi R. and Vitanyi P.M.B. (2005). Clustering by compression. *IEEE Trans. Inform. Theory.* **51**(12):1523.

Groenen M.A., Wahlberg P., Foglio M., Chen H.H., Megens H.J., Crooijmans R.P., Besnier F., Lathrop M., Muir W.M., Wong G.K., Gut I and Andersson L (2009). *Genome Res*. **19**:510.

Henderson C.R. (1975). Rapid method for computing the inverse of a relationship matrix. *J. Dairy Sci*. **58**(11): 1727

Hudson N.J., Porto-Neto L.R., Kijas J., McWilliam S., Taft R.J. and Reverter A. (2014). *BMC Bioinformatics* **15**:66

Hudson N.J. (2014) *WCGALP conference proceedings*. Vancouver, Canada

Stainton J.J., Haley C.S., Charlesworth B., Kranis A., Watson K. and Wiener P (2015). *Anim. Genet.* **46**: 37.

Van Raden P.M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. **91**:4414