USING THE GeSNP ALGORITHM TO QUALITY CONTROL AND PRIORITIZE DIFFERENTIALLY EXPRESSED GENES FROM COMPARATIVE GENOMICS

Y. Li[1], N. Andronicos[2], A. Ingham[1], P. Hunt[2], R. Windon[2], W. Barris[1] and A. Reverter[1]

[1] CSIRO Livestock Industries, Queensland Biosciences Precinct, 306 Carmody Road, St Lucia, QLD 4067, Australia

[2] CSIRO Livestock Industries, FD McMaster Laboratory, Armidale, NSW 2350, Australia

SUMMARY

One advantage of comparative genomics is the ability to use microarray platforms developed for one species to identify significantly differentially expressed genes in individuals of a closely related species. However, this approach inevitably introduces expression differences that result from sequence variation between the two species rather than true variation in transcription levels. As an example of this we have used a bovine Affymetrix array to profile transcript expression in sheep gut tissues following gastrointestinal nematode challenge. Initial microarray gene expression analyses found a set of 2,191 gene probes to be significantly differentially expressed (DE). Using the GeSNP algorithm and sequence comparison on these gene probe sets, we identified 249 gene probes showing true DE, 348 gene probes due to sequence variation between Ovine and Bovine genomes, 309 gene probes showing the sequence annotation problems in the experiment. The remaining gene probes failed to reach significant threshold values for DE. The results imply that quality control is essential to eliminate the gene probe pairs showing significant hybridization differences that are due to sequence variation rather than true expression differences when analyzing comparative gene expression array data.

INTRODUCTION

Comparative genomics has been frequently applied in gene expression studies to detect gene pathways responsible for biologically important traits. In addition, comparative genomics enables the use of microarray platforms developed for one species to identify significantly differentially expressed genes for various contrast animals of a related species. For example a bovine derived array can be used for profiling ovine RNA abundance, because these animals share a high degree of sequence conservation. However, the approach inevitably introduces expression differences that result from sequence variation between the two species rather than variation in transcription levels due to experimental treatments. Greenhall *et al.* (2007) described an algorithm (GeSNP) which can be applied to detect single feature polymorphisms (SFP, i.e. SNP) from oligonucleotide array-based gene expression data in different populations (strains or species) or individuals. The authors claimed that the algorithm can be used to exclude gene probe pairs that show hybridization differences that are due to genetic variation (i.e. sequence variation) between two species rather than experimentally induced expression differences from extreme performing groups of individuals.

The objective of this study was to determine the extent to which sequence mismatch between species influences the quality of gene expression data. Specifically we report use of the GeSNP algorithm to distinguish differential expression that has resulted from true differences in mRNA abundance from variable hybridization due to cross species sequence mismatch.

MATERIAL AND METHODS

Data. The primary source of data was generated in a study (The Sheep Genomics FG3 expression experiment) that attempted to define the genetic basis for sheep resistance to gastrointestinal nematode (GIN) infection (Menzies *et al.* 2010). The experiment used microarray technology to identify the genes that define the temporal response of sheep that have been selected over many generations for a superior ability to resist GIN infection. The focus was gut tissues that comprise the immediate host-parasite interface, and the innate immune response following a primary GIN challenge. In total, 64 microarray chips were hybridized using RNA samples from 32 animals (All sheep were from the CSIRO *Trichostrongylus* selection flock high responder line. There were 8 unchallenged control sheep (T0), 12 individuals challenged with H. *contortus* and 12 with T. *colubriformis*. For each of the challenged groups 4 sheep were sampled at 3 days (T3), 7 days (T7) and 21 days (T21) post-challenge. Samples of 3 tissues (abomasum, WBC (white blood cells) and jejunum) were collected from all sheep. Initial microarray gene expression analyses were carried out using a mixed model (with fixed effects of array hybridization, detection call, random effects of probe, the interaction between probe and experimental treatment and random error). Resulting from these analyses, a total of 2,191 probe sets showed significant differential expression at the contrasts of experimental treatment (parasites, time courses and tissues), and these probes were chosen for the present study.

The GeSNP algorithm. The detailed procedures of applying the GeSNP algorithm can be found in Greenhall *et al.* (2007). In summary, each gene probe set on the Affymetrix Bovine oligonucleotide array consisted of 11 different oligonucleotide probe pairs (a matched set of two 25-base probes, a perfect match (PM) for the gene of Bovine genome and a mismatch (MM, a single nucleotide change at the position 13 of the probe) for non-specific background binding noise control). Firstly, the fluorescence hybridization intensity difference (PM-MM) between the perfect match and the mismatch was calculated for each probe pair of a gene probe set. Secondly, for any gene probe set with less than seven of 11 probe pairs showing positive intensity differences, the entire probe set was eliminated to minimize false predictions of sequence differences. Thirdly, following the standard Affymetrix microarray data analysis protocol (Oldham *et al.,* 2006), the PM-MM values for all probe pairs of the probe set were rescaled to 200 fluorescence intensity units (by subtracting 200 and then dividing by the standard deviation of four samples in the sample group). Finally, the scaled values for each sample group were averaged over the four samples and the Student's *t*-test was employed for each probe pair to identify statistically significant hybridization intensity differences. The threshold t-value of 5, 6 or 7 as suggested by Greenhall *et al.* (2007) was applied for comparison. In total 24,101 probe pairs (2,191 probe sets with 11 probe pairs each which showed significant differential expression from initial analysis) were analyzed using the GeSNP algorithm.

Genetic (sequence) variation identification. Since all probe sets (gene targets) for the sheep experiment corresponded to the Affymetrix Bovine chip, genome sequence comparisons were made between Bovine genome Btau4.0 (Liu *et al.* 2009) to Ovine Oasis4 sequence (a transcriptome assembly using all publically available ovine ESTs from GenBank) for these gene targets which showed to be DE after applying the GeSNP algorithm. Figure 1 illustrates the flowchart corresponding to the sequence comparison performed to dissect whether significant hybridization intensity differences were due to true sequence variation.

RESULTS AND DISCUSSION

Identification of sequence variation using the GeSNP algorithm. From the initial 24,101 probe pairs (2,191 probe sets), using the GeSNP algorithm and a t-value threshold of 5, a total of 2,825 probe pairs from 906 gene probe sets was found to show significant hybridization pattern differences for the contrasts between different time points within particular tissues (Table 1). The remaining 1285 gene probe sets failed to reach the significant t-value threshold value for DE. It can be seen from Table 1 that as more stringent t-values were applied, fewer probe pairs still showed significant hybridization pattern differences. This is expected as it indicates the existence of true DE genes.

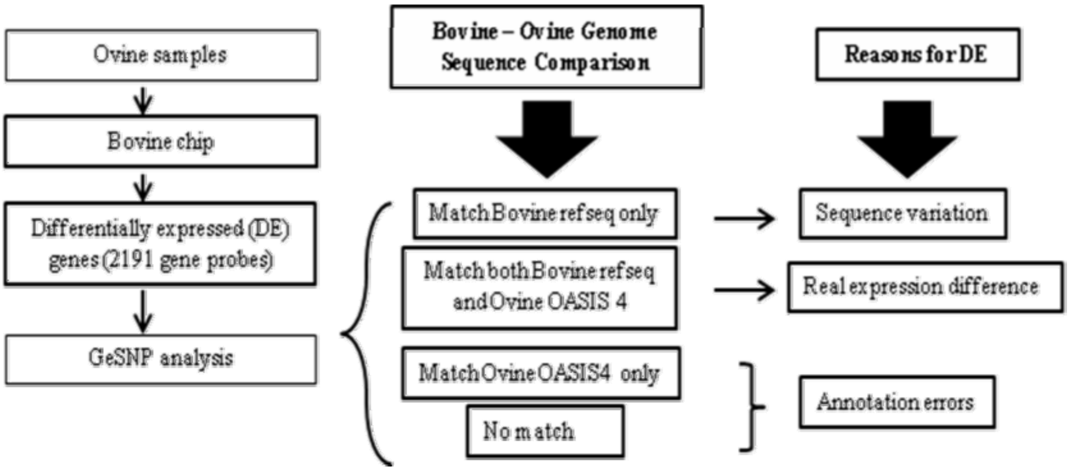
Ovine samples
Bovine chip
Differentially expressed (DE) genes (2191 gene probes)
GeSNP analysis
Bovine – Ovine Genome Sequence Comparison
Match Bovine refseq only
Match both Bovine refseq and Ovine OASIS 4
Match Ovine OASIS4 only
No match
Reasons for DE
Sequence variation
Real expression difference
Annotation errors

Figure 1. Detecting sources contributing to the observation of differential gene expression

Table 1. Number of the probe pairs identified by the GeSNP algorithm with significant hybridization pattern differences between various contrasts (parasites, time courses and tissues).

Tissue	Time Contrast	*t* ≥ *5*	*t* ≥*6*	*t*≥*7*
Abomasum (Hc[§])	T0 vs T3	238	54	28
	T3 vs T7	225	42	24
	T7 vs T21	179	49	24
WBC (Hc)	T0 vs T3	139	28	10
	T3 vs T7	199	21	8
	T7 vs T21	191	51	23
GUT (Tc[¥])	T0 vs T3	149	22	7
	T3 vs T7	270	32	17
	T7 vs T21	231	88	49
WBC (Tc)	T0 vs T3	100	20	9
	T3 vs T7	145	22	12
	T7 vs T21	154	32	13
WBC across parasites	HcT0-TcT0	133	21	11
	HcT3-TcT3	121	23	16
	HcT7-TcT7	213	34	12
	HcT21-TcT21	138	71	30
Total		2825	610	293

Hc[§] - H. *contortus,* Tc[¥] -T. *colubriformi,* Tx value - number of post-challenging days, WBC - white blood cells.

Identification of sources of sequence variation. Our comparative sequence analysis between the bovine and the ovine genomes for the differentially expressed genes revealed that out of 906 gene probes showing significant differential expression, 348 had sequences matching with the bovine genome only, 249 had sequences matching both bovine and ovine genomes, and 69 genes were unique to the ovine genome. The remaining 240 genes did not match either genome.

The results clearly indicate that the significant differential expression identified in 348 gene probes were due to sequence variation between bovine and ovine and not to experimental conditions. In fact 309 gene targets (69 Oasis4 only genes plus 240 no matching) demonstrated the sequence annotation problems in the experiment. The true array hybridization pattern differences were only identified in 249 gene probes with matching sequences for bovine and ovine genomes. This represented only 27.5% of 906 genes showing significant hybridization pattern differences. Therefore there are several challenges when interpreting data from cross-species gene expression experiments because hybridization differences can not only arise because of differential gene expression, but also because of sequence differences between species. In addition annotation errors can also contribute to hybridization differences because of changes in original reference sequences over time and varying criteria used by Affymetrix to design their probes. These challenges will be greater when distantly related species is used for comparative genomic studies.

Although the GeSNP algorithm by Greenhall *et al.* (2007) was developed to identify small sequence differences between groups of individuals within a species, such as single-base substitutions, it certainly can be used as an essential tool to identify sequence differences due to two species to provide the quality control of array-based gene expression data. It is also appropriate to state here that the GeSNP algorithm works only for gene expression data from Affymetrix oligonucleotide arrays with multiple, different, sequence-specific DNA probes for each gene and is not designed for cDNA arrays or other array platforms.

CONCLUSIONS

Comparative genomics provides an efficient way of using a Bovine Affymetrix chip to identify significantly differentially expressed genes in contrast individuals of sheep. However, a caution needs to be taken to eliminate the gene probes that wrongly display significant hybridization pattern differences due to sequence differences between the two species and annotation errors.

ACKNOWLEDGEMENT

This research was supported in part by SheepGenomics which is an initiative of Australian Wool Innovation Limited and Meat & Livestock Australia.

REFERENCES

Greenhall J.A., Zapala M.A., Cáceres M., Libiger O., Barlow C., Schork N.J. and Lockhart D.J. (2007) *Genome Res.* **17**: 1228.

Liu Y., Qin X., Song X.Z., Jiang H., Shen Y., Durbin K.J., Lien S., Kent M.P., Sodeland M., Ren Y., Zhang L., Sodergren E., Havlak P., Worley K.C., Weinstock G.M. and Gibbs R.A. (2009) *BMC Genomics.* **10**:180.

Menzies M., Reverter A., Andronicos N., Hunt P., Windon R. and Ingham A. (2010) *Parasite Immunol.* **32**: 36.

Oldham M., Horvath S. and Geschwind D. (2006) *PNAS.* **103**:17973.