USE OF LEXICAL ANALYSIS TO IDENTIFY POSITIONAL CANDIDATE GENES UNDERLYING A QTL REGION

P.L. Johnson[1], J.C. McEwan[2] and H.T. Blair[1]
[1]Massey University, PO Box 11-222, Palmerston North, New Zealand
[2]AgResearch, Invermay Research Centre, PB 50034, Mosgiel, New Zealand

SUMMARY
Candidate genes are being sought for a QTL affecting muscling in the Texel breed of sheep, on Chromosome 2, close to, but centred distally from Myostatin. Using regions of conserved synteny between humans/cattle/sheep, 123 genes were identified within the equivalent human region. In order to rank these candidates, lexical analysis was carried out on abstracts relating to these genes using the WordSmith program. Myostatin was included, as a gene known to effect muscling (a positive control). Through using the WordList, KeyWords and Concord features of this program and using *muscle*, *lean*, *fat*, *adipose*, *hypertrophy*, *hyperplasia* and *growth* as search words the number of candidate genes was narrowed to 26 including myostatin. Based on these results lexical analysis offers an opportunity to rapidly reduce the number of candidate genes requiring further investigation within a given region.
Keywords: Lexical analysis, QTL, muscle

INTRODUCTION
QTL studies aim to identify markers within a region of DNA accounting for variation seen in a trait caused by an unknown gene variant. In comparison to the human gene map, relatively few genes have been mapped in cattle and even fewer in sheep (O'Brien and Menotti-Raymond, 1999). So although a region of DNA may be identified through QTL work in sheep, it is unlikely that potential candidate genes will be identified using the current sheep gene map. Mapping of sheep/cattle/humans genes shows regions of conserved synteny (groups of genes located together), although rearrangement in their order and exclusions exist (O'Brien and Menotti-Raymond, 1999). These conserved syntenic blocks allow searches for candidate genes within a region of interest in more studied species, particularly humans. This approach is being used to identify potential candidate genes for a QTL affecting muscling and fat traits in the Texel breed of sheep, centred on a 50cM region of Chromosome 2 (Chr2) close to Myostatin (Broad *et al.* 2000; Walling *et al.* 2001, Marcq *et al.* 2002). In an attempt to refine the number of candidate genes needing subsequent investigation, lexical analysis of abstracts reporting the genes was undertaken. Lexical analysis is the study of words. It has previously been used to determine key words relating to the protein function of a gene using relevant abstracts (Andrade *et al.* 1998). This paper demonstrates the use of lexical analysis in the search for candidate genes in a chromosomal region of interest.

MATERIAL AND METHODS
The region of interest on sheep Chr2 (markers INRA40–OARFCB20, 150-193cM; Maddox, 2002) excludes the myostatin gene, however, the region studied was extended to include it (BM81124-OARFCB20, 146-193cM) because of its known effects on muscling. If the methods described here can identify myostatin as affecting muscling, it gives confidence that the approach is useful. The few

genes that have been mapped to ovine Chr2 show homology with cattle Chr2 where they have been linkage mapped, a number of markers are also mapped to homologous regions of cattle and ovine Chr2 (Broad *et al.* 2000). Similarly, synteny has been shown for genes in this region for human and cattle Chr2 (Sonstegard *et al.* 1997). Sonstegard *et al.* (1997) suggest rearrangement in the region of interest between the human and cattle Chr2. However, current comparative maps for the two species are still inexact (http://www.genome.ucsc.edu/, June 2002 and http://www.thearkdb.org/anubis, Jan 2003). For the present work cattle Chr2 map genes GCG and GDF8 were chosen as the boundary, corresponding to 160-190 mega base pairs on human Chr2. The UCSC Genome Bioinformatics website, (http://www.genome.ucsc.edu/, June 2002) was used to identify genes mapped to this region. To eliminate the problem of alternative names for the same gene, the GeneCards™ website (http://bioinformatics.weizmann.ac.il/cards/), was used to create a list of alternative names for each gene. Each gene and its alternative names were used as search criteria within PubMed (http://www.ncbi.nih.gov/entrez/query.fcgi). Abstracts for each gene were saved as a text file.

The lexical analysis package used for this analysis was WordSmith (http://www.oup.com). The program is made up of three main tools: WordList, KeyWords and Concord. Wordlist creates word frequency list for a given text or texts. KeyWords compares a user submitted word list with a "reference" list and reports "Key Words" by comparing the frequency of a given word and the total number of running words in the submitted and reference lists. Concord provides concordance for a search word, providing all instances of that word and those around providing an idea of its context.

The words used in this study were *muscle, lean, fat, adipose, hypertrophy, hyperplasia* and *growth.* Word lists were generated for each gene text file. To avoid frequent words such as "the, a, in" a *stop* list was used to exclude them, this list was created using the 1000 most common words of the British National Corpus (Leech *et al.* 2001). In addition *pmid*, *pubmed*, *medline*, and *publication* were added. Each list was lemmatised (words joined that belong together) using a text file, in this case the words joined were muscle (muscles, muscling, muscular); adipose (adipocyte, adipocytes) and hypertrophy (hypertrophied). The results of the WordList analysis were converted to graphical form by plotting the number of genes a word occurs in against its mean number of occurrences across all the genes. The reference list used to identify *KeyWords* for each gene was based on a list created using all the abstracts for all genes. Within *Concord* the "important" words were used as search criteria across all genes, approximately five words either side of the search word were presented. The results of the concordance query were manually inspected to identify irrelevant entries.

RESULTS AND DISCUSSION

The search within the USCS Genome Bioinformatics website identified 123 candidate genes, of which only 78 had approved names in GeneCards. 3515 papers related to these 78 genes were identified in PubMed. The average number of abstracts per gene was 46 (range 1-735). Using a stop list to eliminate common words is an important editing process. The results of Andrade *et al.* (1998), shows the most common words, by number of occurrences across the texts and mean number of occurrences within the texts, were "of", "the" and "and". The inclusion of such words will distort any analysis.

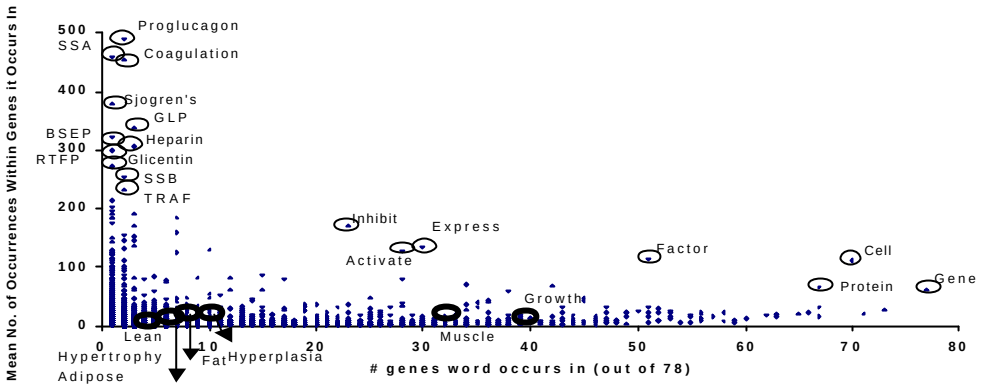

Figure 1. Distribution of words in abstracts of 78 genes from a region of Human Chr. 2. Frequent words (occur in many lists or high frequency in only a few lists) are presented in light ovals. Words describing the desired attributes of a candidate gene are in dark ovals.

Graphing the results of the word lists (Figure 1) shows the most frequent words (i.e. occur in a high proportion of genes or have a high frequency within a small number of genes) and where the "important" words were in comparison. The word that occurred in the most number of lists with moderate frequency within these lists was gene. The word that had the greatest mean number of occurrences (about 450) in its respective list was proglucagon, reflecting the high number of abstracts on this gene. Of the words relevant to this study *growth* occurred in the most number of gene lists (40), with most only occurring in up to 10 of the texts, suggesting the potential for narrowing identifying putative candidate genes. In *KeyWords*, significant p values indicate a words uniqueness to that gene, examples of which are presented in Table 1. For GCG its most unique word compared to other genes is glucagon. Myostatin was included in this study for its known affects on the "important" words, and indeed *muscle* appeared as a key word. No other genes examined produced "important" words as significant key words. If they had done so, those genes would be considered as likely candidate genes.

Table 1. Example of results from KeyWord (WordSmith Tool) for selected genes

	Key Words		
	1st	**2nd**	**3rd**
GCG *(p-value[a])*	Glucagon *(<0.001)*	GLP *(<0.001)*	Proglucagon *(<0.001)*
Myostatin *(p-value)*	Muscle[b] *(<0.001)*	Myostatin *(<0.001)*	Cattle *(<0.001)*

[a] Uniqueness of that word to that gene
[b] Words describing the desired attributes of the candidate gene being sought are highlighted

The Concord Tool was then used to examine "important" words in their context within the abstracts (Table 2). In the examples for GCG the associated words are not in the correct context to consider it

as a candidate gene. Myostatin fits the requirements, referring to skeletal muscle and regulation of muscle growth. Through using this tool one can quickly exclude examples where words of interest may occur at significant frequency but in inappropriate contexts.

Table 2. Example of results from Concord Tool showing the different contexts in which words fit as related to the desired attributes of the candidate gene being sought

Concordance	Gene
Example where an Important Word is not in Right Context for Desired Gene Function	
the trachea and on vascular smooth *muscle* of the pulmonary artery. When	GCG
constant during 2-8 weeks of tumor *growth.* The posttranslational processing	GCG
Example where an Important Word is in Right Context for Desired Gene Function	
characterized by extreme skeletal *muscle hypertrophy* and/or *hyperplasia.* MSTN	Myostatin
negative regulator of muscle *growth.* Mice lacking the myostatin gene (Mstn)	Myostatin

In this study, the number of candidate genes was reduced from 78 to 26, with all 26 referring to at least one "important" word in the correct context. Myostatin was the only gene to appear in both KeyWord and Concord analysis with the desired gene function, however, it is unlikely that it is the gene causing the effect seen in the current work. That no other genes appeared in both does not preclude them from being candidate genes, rather it may be a newly discovered gene with little known about its function.

An advantage of lexical analysis is that it provides a fast and unbiased interpretation of the available literature for the putative candidate genes in the region of interest. In practice it is just one of many methods that will be used to rank positional candidates, another is to use published information on the tissue expression patterns. However, all literature based methods suffer in that they depend on the gene being known and having sufficient published material to accurately identify its function.

ACKNOWLEDGEMENTS
The primary author acknowledges AGMARDT for providing personal funding to undertake this work.

REFERENCES

Andrade, M. and Valencia, A. (1998) *Bioinformatics*. **14**:600.
Broad, T., Glass, B., Greer, G., *et al.* (2000) *Proc. NZ Soc. An. Prod.* **60**:110.
Leech, G,. Rayson, P. and Wilson, A. (2001) "Word Frequencies In Written and Spoken English" 1st ed. Pearson Education Limited, Great Britain.
Maddox, J. F., Davies, K. P., Crawford, A. M. ,*et al.* (2001) *Genome Research* **11:**1275.
Marcq, F., Larzul, C., Marot, V., *et al.* (2002) *World Congress Genetics Appl. to Livestock Production* **7**:No 2-14.
O'Brien, S.J. and Menotti-Raymon, M. (1999) *Science.* **286:**458.
Sonstegard, T; Lopez-Corrales, N.; Kappes, S.; Beattie, A.; Smith, T. (1997) *Mamm.Genome* **8**:75.
Walling, G.A.; Visscher, P.M.; Simm, G. and Bishop, S.C. 2001: *Meeting EAAP* **52:**G5.6.