CALL FOR PAPERS | 2nd International Symposium on Animal Functional Genomics

A gene coexpression network for bovine skeletal muscle inferred from microarray data

Antonio Reverter,1 Nicholas J. Hudson,1 Yonghong Wang,1 Siok-Hwee Tan,1 Wes Barris,1 Keren A. Byrne,1 Sean M. McWilliam,1 Cynthia D. K. Bottema,2 Adam Kister,2 Paul L. Greenwood, 3 Gregory S. Harper, 7 Sigrid A. Lehnert,1 and Brian P. Dalrymple 1

Cooperative Research Centre for Cattle and Beef Quality, 1Commonwealth Scientific and Industrial Research Organisation (CSIRO) Livestock Industries, St. Lucia, Queensland; 2Agriculture and Animal Science, University of Adelaide, Roseworthy, South Australia; and 3New South Wales Department of Primary Industries, University of New England, Armidale, New South Wales, Australia

Submitted 1 June 2006; accepted in final form 10 September 2006

Reverter A, Hudson NJ, Wang Y, Tan SH, Barris W, Byrne KA, McWilliam SM, Bottema CD, Kister A, Greenwood PL, Harper GS, Lehnert SA, Dalrymple BP. A gene coexpression network for bovine skeletal muscle inferred from microarray data. Physiol Genomics 28: 76–83, 2006. First published September 19, 2006; doi:10.1152/physiolgenomics.00105.2006.—We present the application of large-scale multivariate mixed-model equations to the joint analysis of nine gene expression experiments in beef cattle muscle and fat tissues with a total of 147 hybridizations, and we explore 47 experimental conditions or treatments. Using a correlation-based method, we constructed a gene network for 822 genes. Modules of muscle structural proteins and enzymes, extracellular matrix, fat metabolism, and protein synthesis were clearly evident. Detailed analysis of the network identified groupings of proteins on the basis of physical association. For example, expression of three components of the z-disk, MYOZ1, TCAP, and PDLIM3, was significantly correlated. In contrast, expression of these z-disk proteins was not highly correlated with the expression of a cluster of thick (myosins) and thin (actin and tropomyosins) filament proteins or of titin, the third major filament system. However, expression of titin was itself not significantly correlated with the cluster of thick and thin filament proteins and enzymes. Correlation in expression of many fast-twitch muscle structural proteins and enzymes was observed, but slow-twitch-specific proteins were not correlated with the fast-twitch proteins or with each other. In addition, a number of significant associations between genes and transcription factors were also identified. Our results not only recapitulate the known biology of muscle but have also started to reveal some of the underlying associations between and within the structural components of skeletal muscle.

IN RECENT YEARS, gene expression profiling has become part of the suite of technologies used by animal geneticists investigating livestock traits (7, 12, 16, 28). While some progress has been made in the identification of differentially expressed genes between two (or more) experimental conditions, it is also thought that the analysis of large data sets generated from expression profiles may allow the development of predictive models of the systems underpinning the genetics of complex traits (1, 36).

Gene coexpression networks (GCN) can be determined from expression experiments where a number of different perturbations have been profiled. These networks rely on the “guilt-by-association” heuristic, widely invoked in genomics and with universal applicability (35). These networks represent the integration of the various regulatory processes impacting on the expression of the genes in the system being studied. However, for maximum utility, the input data should be derived from clearly defined systems and optimally designed experiments.

To explore regulatory pathways that control gene expression in bovine skeletal muscle and to study the relationships between gene expression and structural compartments, we have undertaken a series of gene expression experiments, mainly using samples of the longissimus dorsi muscle (LD) from adult animals. For this project, we constructed a bovine microarray with 9,600 elements printed in duplicate and comprising ~2,000 expressed sequence tags (ESTs) and ~7,600 anonymous cDNA clones from muscle- and fat-derived libraries (12).

Using this platform, researchers have conducted a number of experiments to identify genes that are differentially expressed across a variety of conditions, and the results have been reported (13, 14, 18, 22, 30, 33, 34). Also, a multivariate mixed-model method has been proposed to jointly analyze independent experiments conducted on this platform (19). The method was further implemented to show the optimality of mixed models for data normalization in gene coexpression studies (20). Based on these results, a network for bovine skeletal muscle with 102 genes was constructed (21).

Here, we present the successful data-driven reconstruction of a GCN in bovine skeletal muscle and adipose tissue, based on gene expression profile data with 822 genes from 147 hybridizations across 9 experiments and 47 conditions. To derive the network, we have applied a large-scale mixed model to normalize gene coexpression measurements. We employed...
partial correlations and data transmission theory to isolate significant associations and explored the relationships between transcription factors (TF) and their potential targets.

**MATERIALS AND METHODS**

**Bovine fat and muscle cDNA library and microarray construction.** The development of the bovine cDNA microarray platform used in this study and related quality control analyses have been described previously in detail (12). In brief, total RNA was prepared from the LD muscle and subcutaneous fat of a 24-mo-old *Bos taurus* Angus steer. The tissue was dissected, immediately frozen in liquid nitrogen, disrupted with a hammer, and homogenized in TRIzol (Invitrogen, Carlsbad, CA) using an ultrasonic homogenizer (IKA-Ikasonic, Staufen, Germany). Poly(A⁺) RNA was separated from total RNA using PolyATtract (Promega, Madison, WI) magnetic streptavidin beads and biotinylated oligo-dT probe according to the manufacturer’s protocol.

Selected cDNA inserts from the two cDNA libraries were amplified by PCR in a 96-well plate format. Before spotting on microarray glass slides, all the PCR products were resuspended in 50% DMSO and transferred to a 384-well plate format. A total of 9,600 elements were printed in duplicate onto glass slides. The array consisted of 9,222 clones of which 1,947 contained accurate (7,898) RNA was separated from total RNA corresponding to the clones; **Z** is the incidence matrix relating signals in **E** including array slide, printing block, and dye channel; **E** and **Z** where **y** is the vector of intensity signals from the E-th experiment; **X** is the incidence matrix relating signals in **y** with systematic fixed effects in **β** corresponding to the clonal effect; **T** is the matrix relating **y** with random effects in **a** corresponding to the three-way interaction of clone by array and printing block; **Z** is the matrix relating **y** with random effects in **d** corresponding to the interaction of clone by dye channel; themselves, connective tissue, and intramuscular fat islands. The microarray also includes a collection of candidate genes from the adipogenesis and protein turnover literature.

**Data sets and gene coexpression measures.** We used nine microarray studies in genetic improvement of beef cattle spanning a total of 147 hybridizations (GEO accession numbers GPL4196 and GPL4197), with mRNA from muscle and adipose tissues in 47 experimental conditions representing a broad dynamic range of perturbations of the cellular systems (Fig. 1). All animal experimentation was approved by the Animal Ethics Committees of New South Wales Agriculture, CSIRO, and University of Adelaide, South Australia.

For the present study, we did not perform any data preprocessing with any filtering criteria based on differential or absolute level of expression. We edited out readings with foreground signal ≤ background signal, as well as clones not observed in all 47 conditions. These criteria resulted in a total of 4,059,807 fluorescent signals from 7,898 clones of which 1,947 contained accurate (P < 0.01/7,898 = 1.27E-6) functional BLAST annotation for 822 unique genes determined by searching the National Center for Biotechnology Information human reference sequence (RefSeq) collection of mRNA sequences. These included 67 genes that were reported as muscle-specific genes (“MSG”) in a massively parallel signature sequencing (MPSS) study (10) and/or a serial analysis of gene expression (SAGE) database (3).

Normalized measures of gene coexpression were obtained from the most optimal statistical modeling that allowed considering all uncertainty at once (20). In brief, the following nine-variate (one for each experiment) mixed model was fitted:

\[
y_E = X_E \beta_E + Z_E a_E + Z_G d_G + Z_A a_A + \epsilon_E
\]

where **y** is the vector of intensity signals from the E-th experiment; **X** is the incidence matrix relating signals in **y** with systematic fixed effects in **β** including array slide, printing block, and dye channel; **Z** is the incidence matrix relating **y** with random effects in **a** corresponding to the clones; **Z** is the matrix relating **y** with random effects in **a** corresponding to the three-way interaction of clone by array and printing block; **Z** is the matrix relating **y** with random effects in **d** corresponding to the interaction of clone by dye channel;
\[ Z_{xy} \text{ is the matrix relating } Y_{E} \text{ with random effects in } t_{E} \text{ corresponding to the interaction of clones by each of the experimental treatment conditions in the } E\text{-th experiment}; \text{ and } e_{v} \text{ is the random error associated with signals in } Y_{E}. \]

It is important to emphasize that model 1 operates at the clone level (i.e., the probe on the microarray comprises the experimental unit), and the clone was the main hierarchy in the variance decomposition with the total variance being decomposed by the components of clone, clone by array block, clone by dye, clone by treatment, and residual. Once assembled, model 1 contained 1,762,338 equations and 81 (co)variance components that were estimated by restricted maximum likelihood using the VCE software (http://sw3.tszv.fal.de/~eg/vce4/vce4.html). It took 248 h and 24 min of CPU time on a Dell server with dual 3.6-GHz processors and 8 Gb of RAM, running a 64-bit version of Red Hat Linux, for model 1 to be solved. Following Reverter and colleagues (20), for each clone in \( c \), a normalized vector of observations was obtained from \((t_{Ec} - \mu_{Ec})/\sigma_{Ec}\), where \( t_{Ec} \) is the vector containing the best linear unbiased prediction (BLUP) of \( t_{Ec} \), \( \mu_{Ec} \) is the vector containing the average BLUP of \( t_{Ec} \) for the \( c \)-th clone in the \( E \)-th experiment; and \( \sigma_{Ec} \) is the standard deviation of all the BLUP of \( t_{Ec} \). Finally, vectors with normalized expressions for each gene were obtained by averaging element by element those vectors corresponding to clones that were annotated to a single gene.

The averaging of vectors for clones associated with a gene was justified by the monotonic increasing pattern that was observed for the average correlation among all pair-wise clones for genes highly represented in the microarray. The number of genes with more than 1, more than 5, and more than 10 clones was 268, 58, and 29, respectively. The average correlation among their clones was 0.394, 0.592, and 0.637, respectively, for the same three categories of highly represented genes. This increasing pattern was attributed to the optimality of the model-based normalization method as well as to the severity of the criterion used for annotating clones to genes, resulting in only 25% of clones (1,947 out of 7,898) being utilized in the network reconstruction.

Identification of significant coexpressions. Partial correlation coefficients and an information theory approach were used to identify meaningful gene-to-gene associations. Although not simultaneously, both strategies have been applied in the reconstruction of gene networks (2, 6). For every trio of genes in \( x, y, \) and \( z \) (having 92,231,140 combinations of 822 genes taken 3 at a time), we computed the first-order partial correlation coefficients in \( r_{zy,x}, r_{xz,y}, \) and \( r_{xy,z} \). The partial correlation coefficient between \( x \) and \( y \) given \( z \) (\( r_{xy,z} \)) indicates the strength of the linear relationship between \( x \) and \( y \) that is independent of (uncorrelated with) \( z \). Again for every trio of genes, and to ascertain a tolerance level for meaningful association, the average ratio of partial to direct (or zero-order) correlation was computed as follows: \( \frac{1}{3} \left( r_{xy,z} + r_{xz,y} + r_{xy,z} \right) \). This tolerance level equated to 0.689, and the \( r_{xy} \) was set to zero if \( r_{xy} \leq 0.689 \left| r_{x,z} \right| \) and \( \left| r_{y,z} \right| \leq 0.689 \left| r_{x,z} \right| \). Otherwise, the association was assessed as significant, and the connection between the pair of genes was established in the reconstruction of the network.

Network connectivity and regulatory modules. The above-mentioned criteria resulted in 42,673 connections out of a possible 337,431 that could have been established. Thus, the clustering coefficient was 12.6%. The clustering coefficient captures how many neighbors of a given gene are connected to each other. The number of connections per gene averaged 103.8 and ranged from 49 for GHSR (growth hormone secretagogue receptor) to 217 for MYBPC2 (myosin binding protein C, fast type).

For the construction of subnetworks, we concentrated on 26 TF that were represented in our data sets and treated as the hubs in the subnets. A gene \( x \) in the data set was included in the \( j \)-th TF-based hub if the \( r \) value between gene \( x \) and the \( j \)-th TF was larger than the \( r \) value between gene \( x \) and any other TF. To assess the optimality of the resulting TF-based hubs, the probability of the observed connectivity was evaluated by permutation test using 10,000 random permutations. Furthermore, the number of MSG in each subnetwork was registered and the hypergeometric test was used to estimate the chance probability of capturing at least the observed number of MSG. Finally, we used the Onto-Tools (8) to identify overrepresented gene ontology (GO) terms and pathways within each TF-based hub.

RESULTS

Hierarchical scale-free behavior of the inferred network. The mixed model accounted for 84.3–96.9% of the total variance. The vast majority of this variation was attributed to the random effect of clone (~80%), followed by the clone by array-block interaction (~15%). Gene expression correlation coefficients (\( r \)) were calculated for all the pair-wise comparisons of the 822 genes, including 67 MSG and 26 TF that were surveyed across the 47 conditions of the 9 experiments (Fig. 1). Of the 822 genes, 269 had detectable expression in human skeletal muscle in a large survey of human tissues (27). The bovine genes were biased toward the more highly expressed of the 932 genes detectable in human skeletal muscle. Correlations were assessed to be significant with a probability proportional to magnitude and relative to the partial correlations in all trio-wise comparisons (Fig. 2, left). The connectivity of the inferred network revealed a power-law distribution of genes with a specific number of interactions on a log-log scale (Fig. 2, right). Using a nominal false discovery rate of <1%, we identified 123 genes involved in 312 connections for which \( |r| \)

![Fig. 2. Left: empirical density distribution of all (gray) and significant (black) interactions (i.e., correlation coefficients) among the 822 genes. Right: distribution of genes with a specific number of interactions on a log-log scale. The straight line indicates the theoretical power-law \( r^2 = 84.7\% \) extrapolated from the experimental data for connectivity greater than 5 (i.e., genes with more than 5 interactions).](http://physiolgenomics.physiology.org)
> 0.75. Four distinct modules were identified by visual inspection (Fig. 3) prior to gene ontology (GO) analyses. The four modules and their significant GO terms were as follows: module 1, skeletal development, structural constituent of ECM; module 2, regulation of cell growth/cycle, fatty acid biosynthesis; module 3, protein biosynthesis, structural constituent of ribosome, RNA binding; module 4, muscle development, tropomyosin binding, magnesium ion binding, calcium ion binding. In this part of the GCN, 29 of the 67 MSG lay in module 4 (P < 0.0001).

TF analysis. Not unexpectedly, in our data set few correlations between genes and TF were significant. To try to identify potential associations between genes and TF, each gene in the GCN was allocated to the TF-based hub corresponding to the most highly correlated TF for that particular gene. A number of TF-defined subnetworks, including ANKRD1, ATF4, CSRP3, and HMGA1, were significantly enriched for MSG (Table 1). However, GO analysis of the TF-defined subnetworks did not always identify significant associations between members of a network and a function. In light of the small number of TF included in this analysis and the fact that many if not all genes are regulated by multiple TF, this is perhaps not a surprising result. To try to overcome the limitations of the single TF hub approach, we clustered genes based on their pattern of correlations to all of the TF (Fig. 4).

DISCUSSION

Tomczak and colleagues (31) clustered mouse genes based on changes in expression during differentiation of C2C12 cells from myoblasts to myotubes. Allowing for differences in inclusion of genes in the two data sets, our module 4 overlapped extensively with their group III and in particular with cluster 10. Group III contains those genes which are significantly induced during differentiation. Clustering of genes within the groups in the C2C12 data was based on expression ratios in the time course. In contrast to the C2C12 differentiation data set, the ECM genes were not correlated with the main cluster of muscle structural protein and enzymes. This is perhaps not surprising given the different systems analyzed. In fact, in the cattle data set, ECM genes were negatively correlated with the module 4 genes.

The muscle module: fiber-type associations. Our analysis was designed to identify some of the key underlying associa-

Fig. 3. Gene network developed with the 123 genes that were involved in 312 connections with r > -0.75 (dotted lines) or r > 0.75 (solid lines). Four modules were clearly distinguishable and for which gene ontology (GO) analyses revealed the following ontologies as being overrepresented: module 1, skeletal development, structural constituent of extracellular matrix; module 2, regulation of cell growth/cycle, fatty acid biosynthesis; module 3, protein biosynthesis, structural constituent of ribosome, RNA binding; module 4, muscle development, tropomyosin binding, magnesium ion binding, calcium ion binding.

Physiol Genomics • VOL 28 • www.physiolgenomics.org
Nevertheless, the expression of this subunit. As components of the mitochondria, they may reflect the adaptations required for the more oxidative characteristics that could relate to either type I or type IIA fibers. In fact, the connection of both SLC25A4 and CKMT2 with MYH2 (myosin isoform IIA) appears to support the latter. These fibers exhibit fast contractile characteristics but are highly aerobic.

In addition to the fiber-type-specific proteins, many structural proteins and enzymes that are components of all fiber types were identified (e.g., z-disk proteins as discussed in the following section). The expression of the thick filament protein titin (TTN), the thin filament protein nebulin (NEB), and the muscle-specific class III intermediate filament protein desmin (DES) was not significantly correlated with the large cluster of interactions between muscle structural proteins and enzymes in a production animal. Skeletal muscle in bovids is compartmentalized into discrete type I, IIA, IIX/d, and IIB fibers. These fibers exhibit specific functional and metabolic properties relating to contractile speed, fatigue resistance, and substrate use. Each fiber type expresses predictable suites of genes that help define these functional differences. Such genes range from the particular myosin heavy chain isoform through to the preferred pathways of energy metabolism (aerobic vs. anaerobic) and preferred substrate for combustion (triglycerides, glycogen, and/or phosphocreatine). Under various treatment conditions, the fiber types will respond uniquely. For example, in starving mammals the faster type IIB fibers show a more rapid atrophy than the type I fibers. Our muscle samples were from the LD muscle, predominantly composed of type IIA (fast oxidative-glycolytic) and IIB (fast glycolytic) fibers, and many of the gene expression clusters we have identified reflect such compartmentalization (Fig. 3). For example, MYBPC2 (fast-twitch myosin binding protein) clusters with the carbohydrate metabolizing enzymes FBP2, PGM, PGM1, and CKM. The expression of these enzymes reflects the physiological coordination required to store and sequester glycogen and phosphocreatine, the preferred storage substrates for the faster type IIA/x and b fibers, but not the slower type I fibers. Furthermore, MYBPC2 also clusters with other key components of the fast-twitch phenotype, including CASQ1 (fast-twitch calcium cycling) and various sarcomeric, contractile proteins such as ACTN3 (high velocity forceful contractions), MYH2 (IIa myosin heavy chain), and MYH1 (IIX/d myosin heavy chain). However, the fast muscle regulatory myosin light chain, MYLPF, was not part of the large cluster, suggesting more independent expression of this subunit.

In contrast, there is no strong evidence for clustering of slow-twitch (type I) proteins. For example, TNN1, TNNC1, MYOZ2, MYH7, MYL3, MYBPC1, and ATP2A2 are present in the set of 822 genes, but with the exception of MYH7, their expression is not significantly correlated with each other. This may reflect the low abundance of the slower fibers in the LD muscle. Nevertheless, SLC25A4, CKMT2, and MDH2 are clustered to some extent. As components of the mitochondria, they may reflect the adaptations required for the more oxidative

<table>
<thead>
<tr>
<th>TF</th>
<th>g</th>
<th>MSG</th>
<th>k</th>
<th>c(k)</th>
<th>r</th>
<th>GO*</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANKRD1</td>
<td>40</td>
<td>7</td>
<td>264</td>
<td>0.34</td>
<td>0.483</td>
<td>0003779</td>
</tr>
<tr>
<td>ATF4</td>
<td>73</td>
<td>11</td>
<td>1,279</td>
<td>0.49</td>
<td>0.545</td>
<td>0003735</td>
</tr>
<tr>
<td>CSPR3</td>
<td>38</td>
<td>8</td>
<td>238</td>
<td>0.34</td>
<td>0.438</td>
<td>0001558</td>
</tr>
<tr>
<td>HMGA1</td>
<td>37</td>
<td>6</td>
<td>247</td>
<td>0.37</td>
<td>0.474</td>
<td>0000287</td>
</tr>
<tr>
<td>MYF6</td>
<td>54</td>
<td>6</td>
<td>426</td>
<td>0.30</td>
<td>0.459</td>
<td>0005509</td>
</tr>
<tr>
<td>MYOG</td>
<td>28</td>
<td>2</td>
<td>140</td>
<td>0.37</td>
<td>0.482</td>
<td>0005578</td>
</tr>
<tr>
<td>NFIX</td>
<td>33</td>
<td>2</td>
<td>219</td>
<td>0.41</td>
<td>0.488</td>
<td>0003954</td>
</tr>
<tr>
<td>NFKB2</td>
<td>35</td>
<td>0</td>
<td>190</td>
<td>0.32</td>
<td>0.438</td>
<td>0003700</td>
</tr>
<tr>
<td>SREBF1</td>
<td>32</td>
<td>0</td>
<td>185</td>
<td>0.37</td>
<td>0.483</td>
<td>0003774</td>
</tr>
<tr>
<td>ZFP38L1</td>
<td>36</td>
<td>3</td>
<td>205</td>
<td>0.32</td>
<td>0.474</td>
<td>0003723</td>
</tr>
</tbody>
</table>

Number of total genes (g), muscle-specific genes (MSG), connections (k), clustering coefficient (c(k)), and average correlation (r) among genes in each transcription factor (TF) and most represented gene ontology term (GO).

*0003779, actin binding; 0003735, structural constituent of ribosome; 001558, regulation of cell growth; 0000287, magnesium ion binding; 0005509, calcium ion binding; 0005578, extracellular matrix; 0003954, NADH dehydrogenase activity; 0003700, transcription factor activity; 0003774, motor activity; 0003723, RNA binding.
fast-twitch proteins, or with each other. However, a number of different splice variants of the TTN and NEB proteins have been identified in humans, and it is possible that these may contribute to the lack of correlation in expression.

The muscle module: relationships between structural components. The fibers in skeletal muscle are highly organized and are composed of a number of discrete structural components that contain different sets of proteins (5). The relationships between the expression of proteins within a component, such as the z-disk, or the thick or thin filaments have not been examined in any detail. By matching the nodes in the GCN with the physical locations of the proteins in the muscle, we have investigated the relationships between expression and location of the proteins. The expression of MYOZ1 is significantly correlated with the expression of TCAP. Two proteins are both located in the z-disk and bind to each other (9), and they located in the same cluster in the C2C12 differentiation data set (31). In addition, the expression of TCAP is correlated with the expression of PDLIM3, another component of the z-disk (11). The expression of both TCAP and PDLIM3 is also positively correlated with the expression of PFKM (a muscle cytoplasmic phosphofructokinase) and negatively correlated with the expression of SPP1 (osteopontin). The specific role, if any, of PFKM and SPP1 in muscle z-disk is currently unknown; however, PDLIM3 and PFKM lay in the same cluster in the C2C12 data set (31).

CSRP3 (muscle LIM protein) is also located in the z-disk, but its expression is not correlated with the MYOI1 cluster. However, the expression of CRYAB and MUSTN1 is significantly correlated with the expression of CSRP3. Muscles of individuals with mutations in CRYAB have been reported to exhibit myofibrillar disintegration that begins in the z-disk (24). MUSTN1 is a nuclear protein expressed during musculoskeletal development and regeneration and in adult muscle and tendons (15). The correlation with CSRP3 suggests that MUSTN1 may be involved in the same signaling pathway. A number of other z-disk proteins were included in the analysis: TTID expression was correlated with the expression of SLC25A4 (a muscle mitochondrial adenine nucleotide transporter), ACTN3 expression was highly correlated with the expression of a number of major muscle structural proteins including TPM2, TMOD4, MYH1, and MYBPC2 and enzymes such as FBP2, ENO3, and CKMT2. In contrast, ACTN2 expression was not as highly correlated with the large cluster in module 4. In humans ACTN2 is expressed in all muscle fibers, while ACTN3 is restricted to a subset of type II fibers (17). In mice, ACTN2 is expressed in all type I, IIa, and IIx/d fibers and some but not all IIb fibers, while ACTN3 is restricted to type IIb fibers. The stronger correlation of ACTN3 may reflect the preponderance of IIb fibers in LD muscle.

TF analysis. Figure 4 shows a negative r value between the expression of MYOG, HMGAL, and ZFP36L1 and a cluster of proteins that includes a number of muscle structural proteins. Expression of both MYOG and ZFP36L1 is induced by insulin-like growth factor I (IGF1), which promotes muscle growth, and both are induced during differentiation of mouse C2C12 cells and hence are positively correlated with the major muscle structural proteins and enzymes (31). Additional TF that correlated with the muscle structural proteins are ANKRDI1, AT4, NSEP1 (YBX1), MYF6 (all positively correlated), and more weakly NFKB2 and MEF2C (both negatively correlated). Of these, only MEF2C and MYF6 show significantly altered expression levels during C2C12 differentiation. However, in addition to MYF6 and MEF2C, MYOG has been shown to positively regulate the expression of ANKRDI1 and AT4. Thus the analysis of this set of data has identified TF already known to play a role in muscle development through the role of MYOG. The results also suggest that NSEP1 (YBX1) may play a significant role in regulating gene expression in muscle cells. Indeed, NSEP1 has been proposed to be involved in the activation of transcription in a subset of fast-twitch muscles (26). In addition, myogenic regulatory factors MYF6 and MYOG were negatively correlated with each other. There is experimental evidence that this association may be a reflection of biological reality. MYF6 −/− mice show only a subtle phenotype but have a threefold increased level of MYOG transcription (37), suggesting that MYOG may compensate for the absence of MYF6 and that MYF6 may act as a negative regulator for MYOG.

The differences between the expression of TF and muscle proteins observed in C2C12 cells during differentiation and our data may reflect differences between the in vitro differentiation system and muscles in live animals. There appears to be a negative r value between TF such as CSRP3 and NFE2L1 and matrix protein gene expression. Neither of these proteins has been reported to be involved in regulation of matrix protein genes. CSRP3 is a positive regulator of myogenesis (potentially as a nuclear TF) and may also play a role in mechanical stretch sensing in the z-disk. Other TF associated with the matrix molecule cluster were NSEP1, NFBK2, and MYOG; of these, NSEP1 has been reported to activate the transcription of COL1A1 (29), but in our data, NSEP1 is negatively correlated with COL1A1.

Ontology analyses reveal association between magnesium and protein synthesis. Linkages among the most represented GO terms (Fig. 5) revealed a substantial number of obvious associations (e.g., hormone activity is associated with the regulation of cell growth). These serve to validate the reverse-engineering method. However, some of the other associations have only become apparent in the more recent literature. For example, the association between magnesium binding with both the structural constituents of the ribosome and RNA binding is compelling. Magnesium is possibly the missing element in molecular interpretation of cell proliferation and protein synthesis (23). These results suggest there is much correspondence between the associations we have uncovered computationally and those associations that are supported experimentally. The novel associations revealed here might provide a useful hypothesis-generating tool for future laboratory research in the agricultural sciences.

Conclusions. We report the integration of nine microarray gene expression experiments in beef cattle. Combined, these represent the largest and most comprehensive bovine muscle gene profiling data reported to date. We applied partial correlation coefficients and an information theoretic approach to isolate significant gene-to-gene coexpression measures. Many of our associations are consistent with the known biology of muscle structure and development. We have also identified previously unknown associations pointing to new relationships between the components of muscle. The overall impression is of a core of correlated expression of fast-twitch actin and myosin filament proteins and enzymes, with the expression of
slow-twitch-muscle-specific proteins, and some components common to most or all fibers types, being less correlated to each other and the fast-twitch proteins. We have identified a number of clusters of z-disk proteins marginally correlated to the fast-twitch cluster. Clearly, the exact nature of the relationships between the different types of protein molecules in the different components and types of muscle fibers is complex.

We have compared our results, predominantly from the LD muscle in adult cattle in a range of production and nonproduction environments, with results from the in vitro differentiation of mouse C2C12 cells. The comparisons have highlighted the shortcomings of gene expression correlation analyses. The very restricted nature of the C2C12 data does not allow fine details to be distinguished. At the other end of the spectrum, undertaking correlation analyses across diverse tissues will focus on very basic cellular processes. We have attempted to include a range of conditions from a very limited range of tissues from B. taurus to find the balance between the overall very highly correlated networks in C2C12 cells and the inclusion of a very diverse set of tissues. However, we acknowledge that the bovine transcriptomes analyzed here are limited in coverage and were derived from microarray data from at least two cell types (although predominantly skeletal muscle and adipose tissue with smaller proportion of fibroblasts and endothelial cells), a total of 47 experimental manipulations (including development and nutritional intervention), and both in vivo and in vitro conditions. We view this experimental breadth as a strength because only the most fundamental transcriptional associations will clearly emerge. On the other hand, we cannot reject the possibility that some real associations have been overlooked. There are some distinct differences between the network from this analysis and a previous analysis of the first five sets of data (21). The future challenge is to explore the network that results from the integration of various sets of data to construct the hierarchy of the relationships between the components of the network, and from this, to determine the regulatory nature that lies behind the coexpression network.

REFERENCES


16. Nobis W, Ren X, Suchyta SP, Suchyta TR, Zanella AJ, Coussens PM. Discovery of mean-


