Functional contexts of adipose and gluteal muscle tissue co-expression networks in the domestic horse

Robert J Schaefer1, Jonah Cullen1, Jane Manfredi2, Molly McCue1

1 Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St. Paul, MN

2 Department of Pathobiology and Diagnostic Investigation, College of Veterinary Medicine, Michigan State University, East Lansing, MI

Notes:

* [General ICB Journal Instructions](https://academic.oup.com/icb/pages/General_Instructions)
* Total of 4,500 words (12 pages)
* Style according to Council of Science Editors / Scientific Style and Format 7th ed.
* Figure guidelines:
  + 600 dpi for line figures, 300dpi for color figures
  + 88mm single columns, 180mm double column

# Abstract (limit: 400 words)

A gene’s response to an environment is tightly bound the underlying genetic variation present in an individual’s genome and varies greatly depending on the context of the tissue it is being expressed in. Gene co-expression networks provide a mechanism to understand and interpret genes’ transcriptional response to stimuli. Here, we use the Camoco co-expression network framework to characterize the transcriptional landscape of adipose and gluteal muscle issue in the 85 domestic horses (*Equus caballus*) representing 5 different breeds. In each tissue, gene expression profiles, capturing transcriptional response due to genetic variation across individuals, were used to build two separate, tissue-focused, genotypically-diverse gene co-expression networks. The functional contexts of each network were quantified using complementary approaches. First, supervised, ontological enrichment was utilized to quantify biological function represented by each network. Curated ontologies such as GO as well as KEGG were used to both measure the amount of previously known levels of biological function present in each tissue. To complement this, an un-supervised approach not relying on annotations was employed. Strongly co-expressed sets of genes defined by Markov clustering identified sets of unannotated genes showing tissue specific patterns of co-expression. In both supervised and unsupervised approaches, we report similarities and differences in patterns of gene co-expression in each tissue context. Coupling together these wide-reaching transcriptional datasets with our network analysis tool, Camoco, we mapped the transcriptional landscape of muscle and adipose setting up a generalizable functional framework for interpreting gene functional contexts in the horses as well as for additional tissues and use in other species.

# Introduction

Despite high quality, sequenced genomes, functional gene annotations in many domestic and agricultural animal species remain limited. However, advances in sequencing technologies is bridging the gap in understanding the functional role of genes and their relationship between genotype and phenotype. High-throughput sequencing techniques, such as RNA-Seq, enable the measurement of gene expression across diverse environmental conditions, developmental time points or different tissues and organs. Gene co-expression quantifies the correlation of transcription among sets of genes across experimental conditions and provides a holistic view into genes’ involvement in biological processes (Eisen et al. 1998). These data can help establish an unbiased context for what a gene putatively does and where it is expressed, given no other functional information about it. Surveying transcription across a large number of diverse experiments establishes an expression profile for each gene, which can be exhaustively compared, pairwise, to one another in order to uncover putative interactions.

Collectively, co-expression relationships between pairs of genes represent a larger biological network where nodes represent genes and edges represent the magnitude of co-expression among them. Once represented as a network, topological and structural information shows that these biological networks share well define organization properties described with network theory. This fundamental structure allows for a transfer of knowledge from decades of previous research in non-biological systems to gene co-expression networks which are rapidly being built in agricultural species.

The source of variance in the experimental condition provides a functional context to the co-expression profiles. For instance, on a cellular level, co-expression over time has been utilized to map the transcriptional response during the cell cycle in yeast (Spellman et al. 1998). However, experimental conditions of co-expression aren’t limited to time-course experiments. Co-expression contexts can be measured by introducing expression variation stemming from a variety of sources, including different environmental conditions, developmental and tissue/organ based variation, and variation due to genetic background and populations of individuals.

In multicellular organisms, patterns of gene expression provide a functional context of genes’ roles in different tissues and developmental timepoints, commonly called tissue atlases. Additionally, differences in genetic background lead to differences in gene expression, where gene expression can quantify

On the cellular level, gene co-expression has been shown

Co-expression networks represent relationships among genes based on the similarity of their gene expression response across an experimental condition.

Co-expression on a genome wide level is through co-expression networks where nodes in the network represent genes and edges between nodes represents the level of co-expression between them. Gene co-expression networks have been extensively used to map the transcriptional landscape in many species and is a useful tool for inferring biological function when little else is known about a gene. Gene co-expression networks provide a framework to quantify the similarities as well as differences in the transcriptional landscape of gene expression.

As the interactions in a gene co-expression network is a product of the variance introduced by the experimental conditions, the interactions in a gene co-expression network varies greatly based on the condition in which it is measured. Additionally, the variation of response in gene expression is tightly bound to the genetic context where transcription occurs, including the genetic background of the individual as well as the cellular context of the tissue where response occurs.

# Results

## Constructing tissue specific gene co-expression networks

RNA-seq (novaseq) was used to transcriptionally profile 83 horses in both gluteal muscle and tailhead adipose tissue (TAT) in five different breeds: Quarter Horse (n=20), Morgan (n=20), Welsch Pony (n=18), Arabian (n=19), and Thoroughbred (n=6) (Supp. Table 1). Quality control during RNA library prep resulted in slightly imbalanced sample numbers between muscle and adipose, yielding 79 horses with muscle data, 63 horses with adipose data, and 59 horses with RNA-seq data in both adipose and muscle (Table 1).

Gene counts were quantified for both muscle (Supp. Table 2) and adipose (Supp. Table 3) datasets and converted into Transcript Per Kilobase Million (TPM) matrices quantifying tissue specific variation in gene expression across genotypically diverse individuals (See Materials and Methods for details). From these expression matrices, two gene co-expression networks were generated using Camoco (Schaefer et al. 2018), one representing muscle and the other representing adipose (EcMuscle and EcAdipose, respectively, hereafter).

After network quality control filtering (See Materials and Methods) the EcMuscle network contained 14,677 genes (70% of total) and all 79 accessions were included in the network. The EcAdipose network contained 16,067 genes (77% of total) and all 63 accessions passed quality control. A Z-score ≥ 3 was used to define the edge threshold in each network resulting in 649,582 interactions (of 107,699,826 total) in the EcMuscle network and 563,994 interactions (of 129,066,211 total) in the EcAdipose network. Gene clusters with greater than 10 genes were identified by Camoco using the Markov Clustering (MCL) algorithm (Dongen 2000), an unsupervised approach that simulates random walks throughout the network to define clusters (See Materials and Methods). Figure 1 shows an overview for each tissue network, displaying the approximate boundaries of the top ten largest MCL clusters and the degree distribution of each network. In total, 66 MCL gene clusters were identified in the EcMuscle network containing between 10 and 2,068 genes. Likewise, a total of 107 MCL clusters were identified in EcAdipose network having between 10 and 1,177 genes. Both networks showed a heavy-tailed distribution for network edges, which is typical for biological networks (Figure 1 inset; (Strogatz 2001; Alstott et al. 2014)).

## Supervised mapping of functional contexts in tissue co-expression networks

The functional context of each network was quantified using a supervised approach based on the strength of gene co-expression among genes co-annotated for Gene Ontology (GO) and Kyoto Encyclopedia of Gene and Genomes (KEGG) Terms. Briefly, each ontology: GO and KEGG, are composed of “terms”, which contain sets of genes with curated functional annotations based on experimental evidence. Terms from each ontology were scored based on co-expression among member genes to determine if network interactions recapitulate previously annotated functional information.

Network density (Equation 1) was measured among sets of genes within 5,418 GO and 70 KEGG terms having at least 10 and at most 300 genes in both networks to target biological processes of medium size (Schaefer et al. 2018). Empirical p-value’s were calculated for each term’s co-expression scores by comparing to randomized sets of genes of the same size (n=1,000). False discovery rates were controlled using the Benjamini-Hochberg procedure with a FDR level of 1%.

Of the 5,418 GO total and 70 KEGG terms tested, 2,184 (40.3%) GO terms and 29 (41.4%) KEGG terms exhibited significant co-expression in at least one of the networks (Figure 2). Ontological terms were compared across tissues to identify processes that were co-expressed in both networks as well as those that were significant in only a single tissue source. In total, there were 899 (16.5%) GO and 19 (27.1%) KEGG terms with significant co-expression (FDR ≤ 0.01) in both EcMuscle and EcAdipose networks. Term sizes varied between 10 and 281 genes in GO and 11 and 146 genes in KEGG, covering a wide breadth of functional coverage (see Discussion). Many terms were only co-expressed in a single tissue when compared between EcMuscle and EcAdipose. The EcMuscle network had a total of 1,115 significantly co-expressed GO and 20 KEGG terms of which 218 GO and 1 KEGG terms and were only co-expressed in the EcMuscle network. Likewise, the EcAdipose network had a total of 1,966 significantly co-expressed GO and 28 KEGG terms, with 1,115 GO and 9 KEGG terms co-expressed only in the EcAdipose network.

## Unsupervised mapping of functional contexts in tissue co-expression networks

Network functional contexts were also evaluated and compared in an unsupervised approach based on overlap of MCL clusters, which represent sets of genes in each network with highly similar gene expression profiles (see Materials and Methods). MCL clusters from each network (n ≥ 10 genes) were evaluated for co-clustering in the respective other network (Figure 3). Briefly, for each combination of MCL clusters between the two networks, the hypergeometric test was used to calculate co-clustering (i.e. overlap) based on the probability of gene set overlap between clusters (p < 0.05, Bonferroni adjusted). A filter was also applied to remove co-clusters with 2 or fewer overlapping genes resulting in 110 total co-clusters among the two networks. Clusters defined in one network often co-clustered with multiple clusters defined in the other. For example, genes in EcMuscle cluster MCL10 co-clustered with 4 EcAdipose clusters: MCL11, MCL16, MCL17, and MCL45 in EcAdipose (See Supp Dataset 6 for complete results). MCL clusters were also evaluated for enrichment with GO and KEGG terms (10 ≤ n genes ≤ 300) to determine significant overlap in potential biological function (hypergeometric test; Bonferroni corrected p≤ 0.05; See Supp. Dataset 7 for details).

Of the 66 MCL clusters in the EcMuscle network, 42 clusters (63.6%) co-clustered in the EcAdipose network having between 1 and 21 EcAdipose co-clusters (p < 0.05, Bonferroni adjusted; Figure 3). Of these 42 EcMuscle MCL clusters with co-clusters, 35 were enriched for 1 or more GO/KEGG term and 7 showed no overlap with either ontology Table 2. Similarly, there were 24 EcMuscle MCL clusters that showed no significant co-clustering with the EcAdipose network. Of these, 10 MCL clusters were enriched for at least 1 GO/KEGG term while 14 were neither co-clustered nor enriched for any ontological term.

A similar breakdown of overlap between MCL terms in the EcAdipose network was found. Of the 107 gene clusters in the EcAdipose network, 55 (51.4%) gene clusters co-clustered in the EcMuscle network having between 1 and 9 co-clusters (Figure 3). There were 35 EcAdipose MCL terms with co-clusters that also had 1 or more significant enrichment for GO/KEGG terms and 7 with no significant overlap. Of the 25 EcAdipose MCL clusters with no significant muscle co-clusters, 10 were enriched for at least 1 GO/KEGG term while 14 showed neither Table 2.

Overall, functional overlap of network MCL clusters was similar between GO and KEGG. In cases where MCL clusters from either network significantly overlapped with GO/KEGG terms (regardless of co-clustering state). There were zero cases in which MCL clusters showed significant overlap in one ontology and not the other (Supp. Dataset 7). In both networks, GO covered proportionally more clusters than KEGG, having overlap in 45 of 66 (68.2%) EcMuscle and 56 of 107 (52.3%) EcAdipose MCL clusters compared to KEGG’s 6 of 66 (9.1%) EcMuscle and 7 of 107 (6.5%) EcAdipose MCL clusters (Table 3). Many MCL clusters from each network showed enrichment with multiple GO/KEGG terms with single clusters having up to 657 GO and 12 KEGG in the EcMuscle network and 414 GO and 19 KEGG in the EcAdipose Network.

# Discussion

Here, we built, compared, and contrasted co-expression networks built from two different tissues sourced from a single cohort of individuals representing genotypic diversity across five different breeds. Since co-expression networks were built based on gene expression profiles derived from genetically diverse individuals, similarities as well as differences could be assessed by directly comparing network interactions and structure. This is partially made possible since Camoco, the network computational framework we utilized, normalizes and scores network interactions to a

Functional contexts were mapped using both supervised as well as unsupervised methods.

* Camoco normalizes and scores networks for comparison
* DE is just simplified co-expression
* GO and KEGG corroborate the same trends, however, they are not necessarily independent.
  + Eg. EcKEGG[00230]: purine metabolism (124 genes) and GO:0006753 – nucleoside phosphate metabolic process (291 genes) share 75 genes (p < 1.24e-102)
* GO annotations are really only helpful in aggregate, while some annotations “make sense” others are counter-intuitive.
  + The score based method was more strict than the enrichment based method.
* Two tissues can provide some great insight, however it is only a start

For example, high level biological processes important in all tissues were captured by such terms as GO:0016570 (281 genes) annotated for “histone modification” (EcMuscle p ≤ 0.001; EcAdipose p ≤ 0.001) and GO:0005667 (256 genes), “transcription factor complex” (EcMuscle p ≤ 0.002; EcAdipose p ≤ 0.001) describe. Similarly, there were smaller, yet still ubiquitous terms describing more focused processes shared between EcMuscle and EcAdipose. For example, GO:1904851 (10 genes) describes “positive regulation of establishment of protein localization” (EcMuscle p ≤ 0.001; EcAdipose p ≤ 0.001) and GO:0015986 (14 genes), “ATP synthesis coupled proton transport” (EcMuscle p ≤ 0.001; EcAdipose p ≤ 0.001). See Supp. Table 4 for a description of all 899 GO Terms co-expressed in both networks.

# Materials and Methods

## RNASeq read mapping and quantification

RNA-seq Samples were split across multiple sequencing lanes and sequenced do a depth of between 2.4 and 15.7 million reads with an average of 4.2 +/- 1.5 million reads per sample. A RNA-seq gene quantification pipeline was built using SnakeMake (v5.4.5) (Köster and Rahmann 2012). Adapter sequences were detected and trimmed from raw reads using AdapterRemoval (v2.2.2) (Lindgreen 2012). Gene expression levels were quantified using Salmon (v0.13.1) which uses a quasi-mapping algorithm. Prior to quantification, Salmon builds an index from a FASTA file containing known genomic transcripts. As a transcript file is not readily available for the horse, the longest transcript for each gene was calculated using LocusPocus (v0.2.0). Briefly, a LocusPocus database was built using the NCBI EquCab3.0 reference genome and Gene Feature Format (GFF) file. Using LocusPocus, each mRNA feature listed in the GFF was queried, and the longest transcript was identified. The sequence for each exon in the transcript was concatenated and stored in a transcript FASTA file. A salmon index was built using this custom transcript fasta file and TPM was calculated for each gene using `salmon quant` command resulting in a gene expression matrix for both Muscle and Adipose where rows are genes, columns are RNA-seq accessions (n=81 Muscle; n=64 Adipose), and values are transcripts per million (TPM).

## Constructing and quality control of gene co-expression networks

Camoco (v0.6.4) was used to build gene co-expression networks from both Muscle and Adipose TPM matrices (Schaefer et al. 2018). First, Camoco ‘RefGen’ databases were built from the NCBI GFF files using the `build-refgen`, a gene ontology database was built using the `build-go` command and finally, networks were built for both Muscle and Adipose using the `build-cob` command. Quality control was run on both networks: gene TPM values below 0.001 were excluded from correlation calculation; genes with >30% missing data were removed; accessions with more than 50% missing data were removed; and to be included in the network, a gene must have had a TPM > 0.08 in at least one accession.

Gene clusters were identified using the an implementation of the Markov graph clustering algorithm (MCL) which simulates a random walk throughout the network to identify tightly clustered genes (Dongen 2000). Network node coordinates were calculated and visualized using the camoco `plot\_network` command. In this view, network nodes are plotted in two-dimensional space based on the strength and number of interactions among neighbors using the ForceAtlas2 algorithm. Gene clusters in each network were distinguished with a different node color and dotted ellipses in each plot show the approximate boundaries of the top 10 largest clusters in each network.

## Calculating network density among Gene Ontology (GO) terms

Subnetwork density is formulated as the average interaction strength between all (un-thresholded) pairwise combinations of input genes, normalized for the total number of input gene pairs:

### Equation 1

where wij is the co-expression score between genes i and j, and Ne is the total number of pairwise, non-self, co-expression interactions in the subnetwork.

Gene ontology term descriptions and gene mappings for the horse were downloaded from ensemble BioMart (https://ensembl.org). A Camoco GO database was built using the `build-go` command. Network density was calculated in both EcMuscle and EcAdipose networks among genes co-annotated within GO terms having at least 30 and at most 300 genes (representing medium sized biological processes). Density p-values were calculated by comparing the empirical density scores to random sets of input genes (n=1,000), conserving the number of genes in each GO term.

## Calculating network co-clustering

Network gene clusters

## Figure and Table Legends

### Table 1

**RNA-seq sample summary.**

Horses broken down by their tissue type (Adipose and Muscle) and breed.

### Table 2

**Overlap between network MCL, co-clusters, and ontologies.**

Gene clusters defined by MCL in each network are aggregated by their overlap with co-clusters as well as ontologies (GO/KEGG). Values in the “No co-cluster” column represents MCL clusters that showed no significant overlap in the respective other network while the “Co-cluster” column shows MCL clusters with 1 or more co-clusters. Similarly, the “No Annot.” column indicates MCL terms with no significant overlap with either tested ontologies. The “GO/KEGG Annot.” reports the number of MCL clusters with 1 or more significant overlap with tested ontologies. All breakdowns are additionally totaled along margin rows/columns.

### Table 3

**Comparative GO and KEGG enrichment in network MCL clusters**

GO and KEGG terms are broken down and compared in each network based on significant gene set overlap with MCL clusters (Bonferroni p ≤ 0.05; hypergeometric test). Columns show values for the number and percentage of overlap in each ontology within each network, the minimum and maximum number of significant ontology enrichments observed per MCL cluster, and the total number of ontology terms which showed overlap with MCL clusters.

### Figure 1

**Network Overview**

Genes in each network are visualized in 2-dimensional space based on a spring-embedded algorithm that arranges sets of genes with strong co-expression near each other (note that individual network edges are not visualized). The approximate location of the top 10 largest MCL clusters (MCL0-MCL9) are denoted by dotted ellipses. The degree distribution of the network edges (Empirical Data) are shown in the inset plot and compared to the Power Law and Truncated Power Law theoretical distributions.

### Figure 2

**Functional context using network density of Gene Ontology terms**

### Figure 3

**Tissue cross-network MCL cluster enrichment**

### Supp. Dataset 1

**Study sample information**

Sample identifier, breed, and dataset membership information.

### Supp. Dataset 2

**EcMuscle sample gene expression matrix**

Input gene expression values (TPM) for the EcMuscle co-expression network.

### Supp. Dataset 3

**EcAdipose sample gene expression matrix**

Input gene expression values (TPM) for the EcAdipose co-expression network.

### Supp. Dataset 4

**Network density of Gene Ontology terms**

Summary of GO terms and their co-expression density scores in each network.

### Supp. Dataset 5

**Network density of KEGG terms**

Summary of KEGG terms and their co-expression density scores in each network.

### Supp. Dataset 6

**Cross network MCL cluster gene set enrichment**

Summary of the gene set enrichment between MCL clusters between the EcMuscle and EcAdipose networks.

### Supp. Dataset 7

**Ontology enrichment of MCL clusters in EcMuscle and EcAdipose networks.**

Gene set enrichment showing the probability of overlap between MCL cluster gene sets defined in either tissue network and ontological terms defined by either GO or KEGG.

# References

Alstott J, Bullmore E, Plenz D. 2014. powerlaw: a Python package for analysis of heavy-tailed distributions. Rapallo F, editor. PLoS One. 9(1):e85777. doi:10.1371/journal.pone.0085777. [accessed 2014 Mar 23]. http://dx.plos.org/10.1371/journal.pone.0085777.

Dongen S van. 2000. MCL—a cluster algorithm for graphs. http://micans.org/mcl.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 95(25):14863–14868. doi:10.1073/pnas.95.25.14863. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24541&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.pnas.org/cgi/doi/10.1073/pnas.95.25.14863.

Köster J, Rahmann S. 2012. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 28(19):2520–2522. doi:10.1093/bioinformatics/bts480.

Lindgreen S. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes. 5(1):337. doi:10.1186/1756-0500-5-337. [accessed 2015 Sep 4]. http://www.biomedcentral.com/1756-0500/5/337.

Schaefer RJ, Michno JM, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myersc CL. 2018. Integrating coexpression networks with GWAS to prioritize causal genes in maize. Plant Cell. 30(12):2922–2942. doi:10.1105/tpc.18.00299.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 9(12):3273–97. [accessed 2015 Sep 17]. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=25624&tool=pmcentrez&rendertype=abstract.

Strogatz SH. 2001. Exploring complex networks. Nature. 410(6825):268–276. doi:10.1038/35065725. http://www.nature.com/doifinder/10.1038/35065725.