Online journal ISSN - 1676-5680

GMR Genetics and Molecular Research

EVOLUTION AND TECHNOLOGY

Chicken skeletal muscle-associated macroarray for gene discovery

E.C. Jorge[1], C.M.R. Melo[2], M.F. Rosário[1], J.R.S. Rossi[1], M.C. Ledur[3], A.S.A.M.T. Moura[4] and L.L. Coutinho[1]

[1]Departamento de Zootecnia, Escola Superior de Agricultura
“Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, SP, Brasil
[2]Departamento de Aquicultura, Universidade Federal de Santa Catarina,
Florianópolis, SC, Brasil
[3]Embrapa Suínos e Aves, Genética e Melhoramento Animal, Vila Tamanduá,
Concórdia, SC, Brasil
[4]Departamento de Produção Animal,
Faculdade de Medicina Veterinária e Zootecnia de Botucatu,
Universidade Estadual Paulista Júlio de Mesquita Filho, Botucatu, SP, Brasil

Corresponding author: L.L. Coutinho
E-mail: llcoutin@esalq.usp.br

Genet. Mol. Res. 9 (1): 188-207 (2010)
Received November 4, 2009
Accepted December 2, 2009
Published February 2, 2010

ABSTRACT. Macro- and microarrays are well-established technologies to determine gene functions through repeated measurements of transcript abundance. We constructed a chicken skeletal muscle-associated array based on a muscle-specific EST database, which was used to generate a tissue expression dataset of ~4500 chicken genes across 5 adult tissues (skeletal muscle, heart, liver, brain, and skin). Only a small number of ESTs were sufficiently well characterized by BLAST searches to determine their probable cellular functions. Evidence of a particular tissue-characteristic expression can be considered an indication that the transcript is likely to be functionally significant. The skeletal muscle macroarray platform was first used to search for evidence of tissue-specific expression, focusing on the biological function of genes/transcripts, since gene expression profiles generated across tissues were found to be reliable and consistent. Hierarchical clustering analysis revealed consistent clustering among genes assigned to ‘developmental growth', such as the ontology genes and germ layers. Accuracy of the expression data was supported by comparing information from known transcripts and tissue from which the transcript was derived with macroarray data. Hybridization assays resulted in consistent tissue expression profile, which will be useful to dissect tissue-regulatory networks and to predict functions of novel genes identified after extensive sequencing of the genomes of model organisms. Screening our skeletal-muscle platform using 5 chicken adult tissues allowed us identifying 43 'tissue-specific' transcripts, and 112 co-expressed uncharacterized transcripts with 62 putative motifs. This platform also represents an important tool for functional investigation of novel genes; to determine expression pattern according to developmental stages; to evaluate differences in muscular growth potential between chicken lines, and to identify tissue-specific genes.

Key words: *Gallus*; Gene expression; Skeletal muscle;
Tissue-specific expression

INTRODUCTION

The chicken is an important non-mammalian vertebrate model; the availability of the complete genome sequence (Hillier et al., 2004) will likely contribute to fundamental discoveries and scientific progress in medicine, developmental biology and livestock production. However, even after extensive sequencing efforts, analysis of the gene sequences revealed that only about 50% of chicken proteins were known to be expressed *in vivo*; the remaining were only digitally predicted (Buza et al., 2007).

Macro- and microarrays are well-established technologies used to determine gene functions through repeated measurements of transcript abundance. High throughput profiling of gene expression provides insights into new gene functions and transcriptional regulation that underlies biological processes (Eisen et al., 1998; Niehrs and Pollet, 1999). Available chicken arrays have been mainly developed based on tissue-specific gene expression, including an intestine-specific array containing 3072 transcripts (van Hemert et al., 2003), a macrophage-specific array with 4906 transcripts (Bliss et al., 2005), a lymphocyte-specific array with 3011 clones (Neiman et al., 2001), an immune response-specific array with 5000 genes (Smith et al., 2006), a heart precursor cell-specific array with 11,000 genes (Afrakhte and Schultheiss, 2004), and others (Jorge et al., 2007; Cogburn et al., 2007).

We have developed in-house a 9378 chicken skeletal muscle-associated expressed sequence tag (EST) database, generated from 5'-end sequencing of cDNA clones from six libraries: one from somites (developmental stage HH15; Hamburger and Hamilton, 1951), the precursors of vertebrate skeletal muscle; one from limb buds in three developmental stages (HH21, HH24 and HH26); one from whole embryos (HH26) (Jorge et al., 2004), and three from the pectoralis major muscle at various developmental stages from broiler and layer lines (pool of HH35 and HH43, for broiler and layer lines, and pool of one and 21 days post-hatch, just for a broiler line). All ESTs were deposited at the dbEST at GenBank (http://www.ncbi/nlm.nih.gov/dbEST) as CD760792 to CD765430 and CO502869 to CO507803. Our objective was to construct an exclusive chicken-expressed sequence database that represents the complete myogenic program, from cell determination to differentiation, considering all cell populations in chicken skeletal-muscle samples.

However, only a small number of these ESTs were sufficiently well characterized regarding their cellular functions based on annotation. For the large majority of the transcripts, their functions remained either completely unknown or only partially understood. Therefore, we developed approximately 4500 chicken skeletal muscle-associated macroarray based on our myogenic-specific EST database to use the expression profile to functionally characterize unknown or uncharacterized chicken transcripts.

We used this macroarray platform to generate an expression dataset of approximately 4500 chicken genes across five chicken adult tissues (skeletal muscle, heart, liver, brain, and skin). Tissue screening was first used because evidence of a particular tissue-characteristic expression can provide an indication that the transcript is likely to be functionally significant (Bono et al., 2003; Zhang et al., 2004). Gene expression profile data across tissues were reliable and consistent with previous information about gene expression and tissue function. Tissue profiling analysis allowed us to suggest novel functions to known and unknown genes; this information will be useful to direct experimental characterization of chicken genes.

MATERIAL AND METHODS

Transcript selection and array construction

The transcripts selected to be spotted onto the macroarray were identified in an in-house constructed skeletal muscle-associated EST database. The macroarray was constructed using the Q-bot robot (Genetix, Queensway, UK) by the Brazilian Clone Collection Center. Bacterial clones were spotted on 8 by 12 cm high-density nylon filters (PerForma II, Genetix) in duplicate, with a layout of 384 blocks in a 5 by 5 configuration.

Plasmidial probes

Plasmidial probes were used to determine the amount of DNA in the bacterial clones spotted onto the macroarray membranes. Oligos were obtained to recognize a specific region of the Ampicillin gene (5’-TAGACTGGATGGAGGCGGATAA-3’ and 3’-CGCCTATTTCAAC GTCCTGGTG-5’) present in the pSPORT1 sequence of every clone. They were labeled using the Klenow large fragment of DNA polymerase I (Invitrogen Co., Carlsbad, CA, USA) to incorporate [α-^{33}P]-dCTP in the sequence of complementary oligos, using the overgo method (Ross et al., 1999). Probes were purified using G-50 columns (GE Healthcare, Piscataway, NJ, USA), following manufacturer instructions, and immediately used to hybridize the macroarray platforms.

Biological material and RNA preparation

Chicken tissues were obtained from nine 21-day-old broiler chickens. Pectoralis major muscle, heart, liver, brain, and skin were collected from these animals. Three pools of dissected tissues, derived from three animals each, were homogenized with Trizol® Reagent

(Invitrogen) to isolate total RNA. Poly(A)$^+$ RNA was purified using the Oligotex kit (Quiagen, Hilden, Germany), following manufacturer directions.

Labeling and hybridization

HotScribe first-strand cDNA labeling (GE Healthcare) was used for cDNA probe synthesis and labeling using [α-^{33}P]dCTP, following manufacturer instructions. After labeling, probes were purified in G-50 columns. The labeled cDNA was heated to 95°C for 3 min and immediately used for hybridization. A procedure similar to Northern blotting was used for hybridization, as described by Sambrook et al. (1989). After a washing step, membranes were placed in contact with an imaging plate (Kodak, Rochester, NY, USA) for 72 h. The digital image was obtained in Storm® PhosphorImager (GE Healthcare) and quantified using the ArrayVision© software (version 8.0, Imaging Research, GE Healthcare). The volume value corrected for the background signal was used for the statistical analysis.

Statistical analysis

A two-step general linear model, described by Wolfinger et al. (2001), was used to normalize the macroarray data and to detect differentially expressed genes. In the first step, expression data were normalized using the following model: $y_{ijklm} = \mu + G_i + T_j + M_k + Q_{(k)l} + \varepsilon_{ijklm}$, where y_{ijklm} is the log$_2$ value of the intensity of the hybridization sign (gene expression); μ is a constant associated to each observation; G_i is the effect of gene i (i = 1,…,4,520); T_j is the effect of treatment j (j = 1,…,5); M_k is the random effect of membrane k (k = 1,…,6); $Q_{(k)l}$ is the random effect of quadrant within each membrane (l = 1,…,384), included to adjust for the spatial effect on the membrane, and ε_{ijklm} is the random error associated with each observation. This model assumes M_k, $Q_{(k)l}$ and ε_{ijklm} are idd $N(0,\sigma_M^2)$, $N(0,\sigma_Q^2)$, $N(0,\sigma_\varepsilon^2)$ respectively, with all of them independent of each other.

In the second step, the residuals from this model were denoted r_{ijkl}, computed by subtracting the fitted values for the effects and the residuals from the first step from the y_{ijklm} values. This defined the following gene-specific model: $r_{ijkl} = G_i + (GT)_{ij} + (GM)_{ik} + e_{ijkl}$, where r_{ijkl} is the residual of the normalization model; G_i is the average effect of gene i; $(GT)_{ij}$ is the effect of treatment j on gene i; $(GM)_{ik}$ is the effect of membrane k on gene i, and e_{ijkl} stands for the random error associated with each observation. This model considers that $(GM)_{ik}$ and e_{ijkl} are idd $N(0,\sigma_{GM}^2)$ and $N(0,\sigma_e^2)$, respectively, with all of them independent of each other. Data were analyzed using PROC MIXED in SAS (SAS/STAT software version 9, SAS Institute) and the significance of the differences between expressed sequences was assessed by the *t*-test ($P < 0.05$).

RESULTS AND DISCUSSION

Genes spotted onto the macroarrays

The macroarray was developed using transcripts derived from a collection of 9378 chicken skeletal muscle-associated ESTs constructed in-house. These ESTs were generated from 5' end-sequencing of clones obtained from six cDNA libraries (Table 1): one from somites (developmental stage HH15; Hamburger and Hamilton, 1951); one from limb buds in three developmental stages (HH21, HH24 and HH26); one from whole embryos (HH26), and three from pectoralis major muscle in various developmental stages for broiler and layer lines (Alves HJ, unpublished results). This EST collection was originally annotated using the identification from the highest hit score using BLAST (BLASTN and BLASTX against the GenBank chicken genome, and non-redundant and EST databases, respectively). This database was deposited in the dbEST division of GenBank as CD760792 to CD765430 (Jorge et al., 2004) and CO502869 to CO507803. Clustering and assembling of the EST collection was conducted using CAP3 (Huang and Madan, 1999), resulting in 4269 unique sequences. One representative clone from each contig and all singlets were selected to be spotted onto the nylon membrane platforms. Selection of a representative clone from each contig was based on a search for the longest EST read. As the cDNAs were synthesized from the 3' poly(A) tail up to an average insert size of around 1 kb; possibly the longest sequence also had the longest part of the coding sequences. During clone selection, whether two or more unique sequences were from the same mRNA was not considered. In addition, clones representing genes of α-actin and GAPDH, plus pSPORT1 empty vector (Invitrogen) were selected to fill 251 random spaces in the array, to be used as positive and negative controls, respectively. Therefore, selection resulted in a total set of 4520 clones. The set was re-arrayed into twelve 384-well plates and robotically spotted in duplicate onto the nylon filter (9040 spots on the membrane).

Table 1. Selection of transcripts for the macroarray construction.

Library	Tissue	Developmental stage HH	Number of ESTs in array
SM1	Somites associated with neural tube/notochord	HH15	1,096
EM1	Whole embryo	HH25	119
LB1	Limb bud	HH21, HH24 and HH25	655
CB1	Breast muscle	HH35 and HH43, broiler	731
EB1	Breast muscle	HH35 and HH43, layer	988
EB2	Breast muscle	1 and 21 days old, broiler	680
Control			251
Total			4,520

HH: Hamburger and Hamilton, 1951. ESTs = expressed sequence tags.

Filtering the expression database

The skeletal-muscle associated macroarray was used to simultaneously determine the abundance of 4520 chicken gene transcripts in 5 tissues: skeletal muscle, heart, liver, brain, and skin. The expression data were first filtered to remove inconsistent information generated after subsequent hybridization assays. Among the 4520 transcripts spotted onto the macroarray, 11.8% did not show any detectable signal after the first hybridization, performed using plasmidial overgo probes. As the lack of hybridization signal probably derived from problems with colony growth after spotting, these missing spots were removed from the analysis. In addition, 9 clones did not show any detectable signal after hybridizations with all 5 cDNA probes, despite the fact that the plasmidial probe signals were detectable. After further removing these 9 clones, the complete set used to construct the tissue expression database contained 3974 transcripts.

All 5 cDNA probes derived from the distinct adult tissues were hybridized to the macroarray, giving similar numbers of spots with positive hybridization signals, which ranged from 3529 to 3765 for skin and muscle, respectively (Table 2). The expression database was constructed based on this filtered set of data of detectable signals.

Table 2. Numbers of spot signals obtained after hybridization assays.

Tissue	Number of spots with cDNA probe hybridization signals	Number of spots without cDNA probe hybridization signals
Skeletal muscle	3,765	209
Brain	3,691	283
Liver	3,728	246
Heart	3,645	329
Skin	3,529	445

Analysis of the expression profiles

Tissue-specific gene expression has traditionally been used to predict gene/transcript function. Evidence of expression can be considered an indication that a gene/transcript is functionally significant and not an artifact or unprocessed nuclear RNA (Bono et al., 2003; Zhang et al., 2004). Our array was used to produce expression profiling of 5 distinct chicken tissues (skeletal muscle, heart, liver, brain, and skin) to search for evidence of tissue-specific expression, focusing on the biological function of the genes/transcripts. After data filtering, our database was arranged into: 1) differentially expressed transcripts, to identify tissue-specific transcripts and ubiquitously expressed ('housekeeping') genes, based on a statistical model, and 2) sets of co-expressed transcripts, adopting a mathematical description of similarity. Because many cellular processes are tightly associated with coordinate transcriptional changes, cluster analysis of gene expression profiles can be used to identify candidate sets of co-regulated genes that are directly or indirectly involved in related processes (Eisen et al., 1998; Niehrs and Pollet, 1999).

Differentially expressed transcripts: a statistical approach

Statistical analysis was used to identify tissue-specific transcripts, which were those that had hybridization signal in only one of the 5 screened tissues. Even though it was a small screening, tissue-specific genes could be helpful to characterize tissue ontogenesis, evolution, and biomarkers. Tissue-specific transcripts can also provide identification of new gene functions and insights into the transcriptional regulation that underlies biological processes.

Forty-three transcripts were identified with this pattern (Table 3): 11 'skeletal muscle-specific'; nine 'heart-specific'; 11 'liver-specific'; seven 'brain-specific', and five 'skin-specific'. The level of expression of these tissue-specific transcripts most likely reflects differential expression among these tissues analyzed. Only abundant mRNAs are identified in non-normalized cDNA libraries, which mainly correspond to concurrently expressed transcripts (Adams et al., 1991; Soares et al., 1994). As non-normalized libraries were the source of transcripts for the macroarray construction in our study, a small number of transcripts with tissue-specific expression patterns were expected. Tissue expression patterns are traditionally used to characterize unknown transcripts, and they were used here to identify 'skeletal muscle-specific' transcripts. 'Heart', 'brain', 'liver', and 'skin-specific' transcripts were also identified and listed in this study (Table 3).

Table 3. Tissue-specific transcripts identified by macroarray analysis.

Tissue (5)	Clone	Accession number	Chromosome	LOC	Blast hit
Skeletal muscle (11)	GGEZEB1019A02	CO506019	GGA5	TNNI2	Troponin I type 2
	GGEZSM1031G08	CD763069	GGA7	LOC424311	Similar to chromosome 2 ORF 25
	GGEZSM1025A02	CD762476	GGAZ	CENPH	Centromere protein H
	GGEZEB1011B12	CO505879	GGA17	LOC417221	Ubiquitin related modifier 1
	GGEZEB1030G03	CO505176			
	GGEZEB1017B08		GGA10	RGMA	Repulsive guidance molecule A
	GGEZEM1004A09				
	GGEZLB1012C01	CD764305	GGA7	DDX18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18
	GGEZLB1016H06	CD764531	GGA23	LOC419695	Grainyhead-like 3
	GGEZEB1023G11	CO505496	GGA22	LOC395787	Smooth muscle protein phosphatase type 1-binding subunit
	GGEZLB1020G11	CD764829	GGA1	ZDHHC23	Zinc finger, DHHC-type containing 23
Brain (6)	GGEZLB1006H04	CD763873	GGA2	-	
	GGEZLB1015B02	CD760734	GGA12	RBM5	RNA binding motif protein 5
	GGEZLB1024D02				
	GGEZSM1006G04	CD761166	GGA6	LOC423861	Transmembrane protein 180
	GGEZLB1017A11	CD764545	GGAUn	Hmm168569	
	GGEZEB1018B01	CO503061	-	Gga.12751	
Heart (9)	GGEZEB2003E01				
	GGEZCB1003A02				
	GGEZEB1009H05	CO506409	GGA4	POF1B	Premature ovarian failure, 1B
	GGEZEM1003A08	CD763205	GGA1	MIRN135A-2	MicroRNA 135A-2
	GGEZSM1031G01	CD763063	GGA11	GINS3	GINS complex subunit 3
	GGEZSM1020F04	CD762101	GGA26	LOC419807	RNA binding motif protein 15
	GGEZLB1010D12	CD764161	GGA3	LOC421237	Similar to uncharacterized hypothalamus protein HT013
	GGEZSM1026H01	CD762647	GGA21	DNAJC11	DNAJ (Hsp40) homolog, subfamily C, member 11
	GGEZSM1025A12	CD762485	GGA2	Hmm43586	
Skin (5)	GGEZEB2019E01	CO507240	GGA9	LOC424948	Eukaryotic translation initiation factor 2B, subunit 5 epsilon, 82 kDa
	GGEZCB1029H10				
	GGEZLB1025H11	CD765156	GGA24	TRAPPC4	Trafficking protein particle complex 4
	GGEZSM1029H08	CD762905	GGA4	GPR23	G protein-coupled receptor 23
	GGEZEB2014G01	C0507332	GGA8	DEPDC1	DEP domain containing 1
Liver (11)	GGEZEB2015E01	CO506935	GGA4	NXT2	Nuclear transport factor 2-like export factor 2
	GGEASM1007C12	CD761217			
	GGEZSM1008A01	CD761271	GGA3	LOC421819	RNA guanylyltransferase and 5'-phosphatase
	GGEZLB1017G10	CD764602	GGAZ	TBCA	Tubulin folding cofactor A
	GGEZLB1018F12	CD764669	GGA17	LOC417280	NADPH-dependent diflavin oxidoreductase 1
	GGEZEB1002H10	CO505633			
	GGEZLB1015B06	CD760738	GGA23	BSDC1	BSD domain containing 1
	GGEZSM1030A04	CD762913	GGA6	PALD	Paladin
	GGEZSM1031D01	CD763031	GGA15	GNB1L	Guanine nucleotide binding protein (G protein), beta polypeptide 1-like
	GGEZEB1014H01	CO505561	GGAUn	hmm239375	
	GGEZSM1006A01	CD761106	GGA1	LOC427872	Peptidylglycine alpha-amidating monooxygenase COOH-terminal interactor

LOC = locus name following NCBI (http://www.ncbi.nlm.nih.gov) nomenclature.

Among the 11 transcripts identified as 'skeletal muscle-specific' (Table 3), Troponin T type 2 (TNNI2) was identified, which is a fast skeletal muscle protein associated with the regulation of muscle contraction. The expression patterns obtained also highlighted transcripts for which a potential role in skeletal muscle development has not yet been defined. The zinc finger

DHHC-type containing 23 (ZDHHC23), for example, codes for a membrane protein containing a palmitoyl transferase domain, which supposedly promotes protein palmitoylation, a crucial lipid modification in protein trafficking and function (Fukata et al., 2006). Substrates for palmitoylation include H-Ras, a GTP binding protein that regulates cell growth and differentiation (Fukata et al., 2006). A ZnFDHHC motif was found with an expression pattern similar to MyoD (muscle determination transcription factor) in the somitic mesorderm in *Danio rerio* (Nagaya et al., 2002). The expression pattern of this known transcript in chicken skeletal-muscle suggests a new uninvestigated biological function for DHHC motif during myogenesis.

A member of the repulsive guidance molecule (RGM) family (member A; RGM-A) was also identified as ‘skeletal muscle-specific’ transcript. RGM-A was firstly described as responsible for providing guidance cues for axons of retinal neurons (Monnier et al., 2002). Other RGM-A biological functions have been recently investigated, including neural tube closure and inhibition of axon growth after injury in the adult central nervous system (Matsunaga and Chédotal, 2004; Niederkofler et al., 2004; Mawdsley et al., 2004; Hata et al., 2006). RGM-C (also known as hemochromatosis type 2, Hfe2) is the member of the RGM family with biological function described in skeletal muscle, associated with iron homeostasis; it is responsible for the hemachromatosis type 2 disease in humans (Papanikolaou et al., 2004). Curiously, neither genomic nor EST sequences were found for the chicken RGM-C in public databases. Based on the expression profiles described here, we hypothesize a novel biological function for RGM-A in chicken skeletal muscle. Induction of transcripts associated with ‘neuronal activity’ during skeletal muscle development has been described (Szustakowski et al., 2006) and mouse RGM-C was induced in muscle cell survival and differentiation after growth factor treatment, based on microarray analysis (Kuninger et al., 2004).

Clustering analysis

Clustering analysis is a computational method, which calculates similarities of items in large databases to recall patterns and higher order structure. All 3974 valid expression data were clustered using GeneSpring GX (Agilent Technologies), first to obtain the expression profile of all transcripts in all tissues (skeletal muscle, heart, liver, brain, and skin), and second, to identify transcripts that are highly expressed in each tissue, in an attempt to reveal uncharacterized transcripts with similar expression patterns.

Hierarchical clustering

The entire valid expression database was subjected to hierarchical clustering (Eisen et al., 1998), where both transcripts and tissues were clustered. The resulting dendrogram (Figure 1) revealed that ‘brain’ and ‘skin’ were grouped together (bootstrap of P = 100%), and to a lesser degree with ‘heart’ (P = 56%; Figure 1A). ‘Skeletal muscle’ and ‘liver’ were the other tissues grouped (bootstrap of P = 68%). Samples derived from similar embryonic germ layers (ectoderm, endoderm and mesoderm) are expected to show similar gene expression patterns. ‘Brain’ and ‘skin’ are both ectoderm derivatives and were tightly clustered in our analysis. However, mesoderm-derived tissues (‘skeletal muscle’ and ‘heart’) and endoderm (‘liver’) did not show expression consistent with this hypothesis. Inconsistent dendrograms generated from tissue profiling and embryonic origin were previously observed in *Xenopus laevis* (Baldessari et al., 2005).

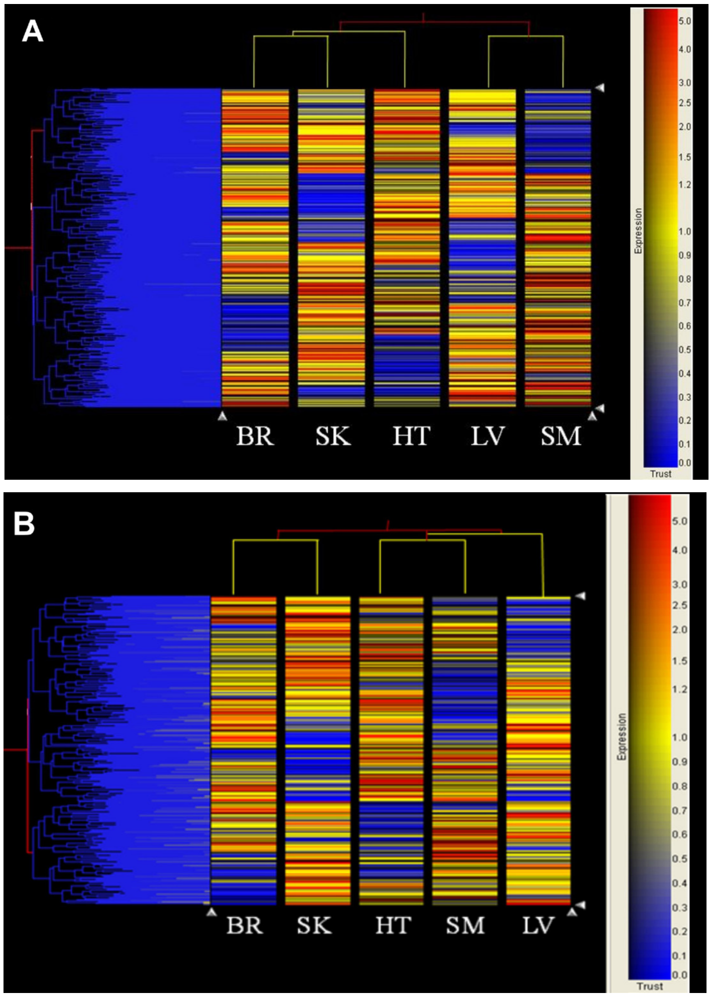

Figure 1. Dendrograms generated from hierarchical clustering analysis. In *A* the valid expression database was subject to hierarchical clustering, where both transcripts and tissues were clustered, and in *B*, transcripts grouped into the 'embryonic development' GO category (1342) were subject to hierarchical clustering. Only in *B*, samples derived from similar embryonic germ layers showed similar gene expression patterns. BR = 'Brain'; SK = 'Skin'; HT = "Heart'; LV = 'Liver'; SM = 'Skeletal muscle'.

Interestingly, when clustering only 1318 transcripts from the 'embryonic development' GO category, a dendrogram consistent with embryonic germ layers origin was obtained (Figure 1B): 'brain' and 'skin' were tightly clustered together (bootstrap values of 100%); 'heart' and 'skeletal muscle' formed another robust group, both separated from 'liver' (Figure 1B). The differences observed in the hierarchical clustering between these two sets of genes might have occurred because genes spotted onto the array were mainly identified in samples of 'skeletal muscle' tissue, which is composed of a mixture of tissues (including conjunctive tissue and vascular and nervous systems), all of them contributing to the ‘muscle’ expression profile. Genes selected from the ‘embryonic development’ GO category might have more specialized functions; for this reason, they were grouped properly following germ layer derivatives.

In order to further characterize tissue expression profiles, highly expressed transcripts (HET) characteristic of each tissue were identified. These genes were recognized by summarizing the expression data for each tissue in a box plot and selecting genes with higher than the upper quartile range as the highly expressed genes (HET). With this strategy, the number of HET for each tissue were: 851 transcripts from ‘muscle’; 880 from ‘heart’; 901 from ‘liver’; 896 from ‘brain’, and 216 from ‘skin’. Comparison among the HET across tissues allowed identifying the most abundant transcripts characteristic of each tissue. No HET were expressed in more than two tissues. There were 197 ‘muscle’ HET; 406 ‘heart’ HET; 474 ‘liver’ HET; 427 ‘brain’ HET (Figure 2). All ‘skin’ HET were also ‘muscle’ HET (Figure 2), probably indicating that there was cross-contamination between ‘muscle’ and ‘skin’ samples.

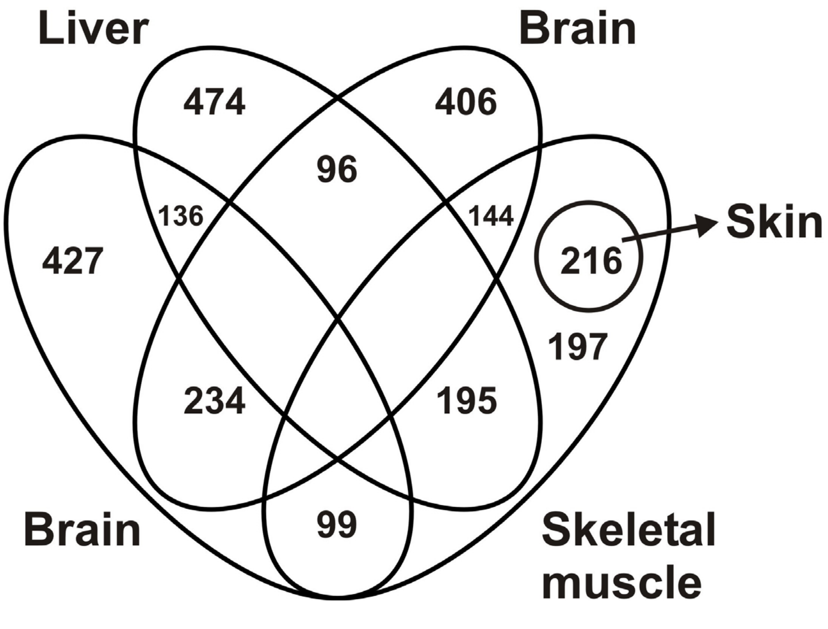

Figure 2. Venn diagram representing numbers of highly expressed transcripts.

All HET characteristic of each tissue were compared against the chicken genome database from NCBI to check the consistency of our expression results. ‘Skeletal muscle’ HETs showed enrichment for i) ‘muscle contraction’ and ‘cytoskeletal organization’ proteins, such as tropomyosin 3 (TPM3), tubulin beta 2A (TUBB), tubulin gamma 1 (TUBG1), myosin heavy chain 8 (MYH8), coronin (actin-binding protein 1C, CORO1C), troponin T type 3 (TNNT3); actinin alpha 2 (ACTN1); tubulin tyrosine ligase-like family, member 12 (TTLL12), tubulin polymerization promoting protein (LOC420800), actin alpha 1 (LOC421534), actin-filament-associated protein (AFAP1), tropomodulin 1 (TMOD1); Smoothelin-like protein (actin binding protein, LOC417687), long microtubule-associated protein 1A (LOC770402), and others; ii) transcripts associated with 'metabolism', such as pyruvate kinase (PKM2), creatine kinase (LOC396507), NADH dehydrogenase (NDUFA5), pyruvate dehydrogenase kinase (PDK3), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), glucose phosphate isomerase (GPI), phosphoglucomutase 1 (PGM1), fructose 1,6 bisphosphatase 2 (FBP2); iii) 'extracellular matrix' and 'cell adhesion': catenin (cadherin-associated protein), beta 1 (CTNNB1), protocadherin 19 (PCDH1), matrin 3 (MATR3), protocadherin gamma subfamily C, 3 (PCDHGC3), procollagen-proline, 2 oxoglutarate 4-digoxygenase, beta polypeptide (P4HB), alpha type XVI collagen (LOC430477); iv) myogenesis-associated transcripts: 'MyoD family inhibitor domain containing' (LOC417774), identified as an inhibitor of myogenic basic helix-loop-helix transcription factors (Kusano and Raab-Traub, 2002), single-minded homolog 2 (LOC418515; Woods et al., 2008), and ZEB1 zinc finger E-box binding homeobox 1 (Postigo and Dean, 1999), and v) transcripts associated with the degradation of muscle proteins, such as ubiquitin conjugating enzyme E2G1 (LOC770961), and F-box and leucine-rich repeat protein 5 (FBXL5). 'Heart', 'liver', and 'brain' HETs have also revealed gene 'markers' that suggest the consistency of the screening using these tissues on a skeletal muscle-associated macroarray (data not shown).

There were some cases in which different transcripts from the same gene were identified as HETs in distinct tissues. For example, CO503458 and CO506114 are transcripts that were cloned from adult and embryonic chicken pectoralis muscle, respectively; both encode for a phosphatidylinositol transfer protein beta (PITPNB), but CO503458 presented high expression in heart, while CO506114 had high expression in liver. The existence of alternative splicing for this gene in these two tissues should be further investigated before disregarding these two transcripts as HETs. Another example observed among the HETs identified in this study was protein families differentially expressed among the sampled tissues; the solute carrier proteins were the best example. Solute carrier family 7 (cationic amino acid transporter, y+ system), member 4, and solute carrier family 16 (monocarboxylic acid transporter), member 1, were identified as 'heart' HETs. Solute carriers identified as 'liver' HETs included family 39 (metal ion transporter), member 1; family 43, member 2, and family 5 (sodium-dependent vitamin transporter), member 6. Solute carrier characterized as 'brain' HETs included family 25 (mitochondrial, adenine nucleotide translocator), member 6; family 16 (aromatic amino acid transporter), member 10 (two ESTs); family 25, member 29; family 7 (cationic amino acid transporter, y+ system), member 5; family 15, member 4; family 25 (mitochondrial citrate transporter), member 1, and family 13 (sodium-dependent dicarboxylate transporter), member 3.

Co-expressed non-characterized (unknown) highly expressed transcripts

Non-characterized transcripts were found to be co-expressed among established HETs. These unknown transcripts included mainly those named 'hypothetical proteins', which are defined as predicted proteins for which there is no experimental evidence of *in vivo* expression.

Among the 406 'heart' HETs, for example, 19 were identified as co-expressed 'unknown' transcripts; as were 40 of the 427 'brain' HETs; 32 of the 474 'liver' HETs, and 21 of the 197 'muscle' HETs (Table 4). As the accuracy of our macroarray measurements was confirmed by functional annotations of those HETs, it is possible to use our tissue expression profile to add biological information for those uncharacterized transcripts. The expression pattern of unknown transcripts among tissues is a step towards their functional characterization.

Table 4. 'Hypothetical transcripts' found to be co-expressed among highly expressed transcripts (HETs).

Clone name	Accession #	LOC	Chromosome	Description	Motifs or conserved domains
'Skeletal muscle' unknown HETs (Total = 21)					
GGEZEB2019F02	CO507252	LOC420999	2	chromosome 9 open reading frame 19 [*G. gallus*]	GLI pathogenesis-related 2; promotes epithelial to mesenchymal transition *in vitro*.
GGEZSM1004E06	CD761013	LOC421073	2	similar to chromosome 18 open reading frame 45 [*G. gallus*]	VRG4: Nucleotide-glucose transporter (Carbohydrate transport and metabolism / Posttranslational modification, protein turnover, chaperones / Intracellular trafficking and secretion].
GGEZCB1020C02	CO504231	LOC428472	2	similar to AI595366 protein [*G. gallus*]	Leucine-rich repeat (LRR)-containing protein 14-like; LRRs, ribonuclease inhibitor (RI)-like subfamily. LRRs are 20-29 residue sequence motifs present in many proteins that participate in protein-protein interactions and have different functions and cellular locations.
GGEZLB1012D02	CD764316	LOC422389	4	similar to MGC83004 protein [*G. gallus*]	
GGEZEB2012A03	CO507432	LOC422860	4	similar to hypothetical protein KIAA0232 [*G. gallus*]	
GGEZSM1020H10	CD762128	LOC425097	5	hypothetical LOC425097 [*G. gallus*]	CCDC86 coiled-coil domain containing 86 [*G. gallus*]
GGEZEB1023D11	CO505466	LOC771455	5	hypothetical protein LOC771455 [*G. gallus*]	
GGEZLB1005F03	CD763772	LOC772299	5	hypothetical protein LOC772299 [*G. gallus*]	
GGEZCB1025A08	CO504294	LOC771456	7	hypothetical protein LOC771456 [*G. gallus*]	NADB_Rossmann; A large family of proteins that share a Rossmann-fold NAD(P)H/NAD(P)(+) binding (NADB) domain. The NADB domain is found in numerous dehydrogenases of metabolic pathways. Methyltransferase domain: Members of this family are SAM dependent methyltransferases.
GGEZCB1016A01	CO504070	LOC429066	8	hypothetical LOC429066 [*G. gallus*]	NADB_Rossmann;
GGEZCB1005C05	CO503101	LOC424866	9	similar to hypothetical protein MGC75902	Primase domain similar to that found in the small subunit of archaeal and eukaryotic (A/E) DNA primases. Primases are DNA-dependent RNA polymerases that synthesis the short RNA primers required for DNA replication. MtLigD_Pol_like: Polymerase (Pol) domain of bacterial LigD proteins similar to *Mycobacterium tuberculosis* (Mt) LigD. The LigD Pol domain belongs to the archaeal/eukaryal primase (AEP) superfamily.
GGEZCB1025C09	CO504308	LOC770260	14	hypothetical protein LOC770260 [*G. gallus*]	
GGEZSM1012G12	CD761620	LOC416639	14	hypothetical LOC416639 [*G. gallus*]	
GGEZSM1002C12	CD760911	LOC417387	18	hypothetical LOC417387 [*G. gallus*]	MBTD1 mbt domain containing 1 [*G. gallus*]; zinc ion binding
GGEZSM1008G10	CD761345	LOC419378	21	chromosome 1 open reading frame 174 [*G. gallus*]	
GGEZCB1013G10	CO504752	LOC425366	23	hypothetical LOC425366 [*G. gallus*]	
GGEZCB1020F06	CO504264	LOC426094	Un	similar to MGC78933 protein [*G. gallus*]	
GGEZEB1026C11	CO505217	LOC777320	Un	hypothetical protein LOC777320 [*G. gallus*]	Peptidase C65 Otubain: This family of proteins conserved from plants to humans is a highly specific ubiquitin (Ub) iso-peptidase that removes ubiquitin from proteins. The modification of cellular proteins by Ub is an important event that underlies protein stability and function in eukaryotes; it is a dynamic and reversible process.
GGEZEB2012D07	CO507467	LOC777053	Un	hypothetical protein LOC777053 [*G. gallus*]	
GGEZSM1018A12	CD761892	LOC425539	Un	C9orf32 chromosome 9 open reading frame 32 [*G. gallus*]	NADB_Rossmann;

Continued on next page

Table 4. Continued.

Clone name	Accession #	LOC	Chromosome	Description	Motifs or conserved domains
GGEZCB1021A08	CO503167		Un	hypothetical protein [*Monodelphis domestica*]	
‘Heart’ unknown HETs (Total = 19)					
GGEZEB2019G04	CO507262	LOC418494	1	C21orf45 chromosome 21 open reading frame 45 [*G. gallus*]	
GGEZLB1009A12	CD764046	C21orf66	1	C21orf66 chromosome 21 open reading frame 66 [*G. gallus*]	GC-rich sequence DNA-binding factor-like protein: Sequences found in this family are similar to a region of a human GC-rich sequence DNA-binding factor homolog. This is thought to be a protein involved in transcriptional regulation due to partial homologies to a transcription repressor and histone-interacting protein.
GGEZEB1014G06	CO505556	LOC769105	2	LOC769105 hypothetical protein LOC769105 [*G. gallus*]	
GGEZSM1018F02	CD761936	LOC768893	2	LOC768893 hypothetical protein LOC768893 [*G. gallus*]	
GGEZEB1025D04	CO505075	LOC395778	3	C6orf72 chromosome 6 open reading frame 72 [*G. gallus*]	
GGEZEB1029H10	CO506472	LOC768787	3	C1orf198 chromosome 1 open reading frame 198 [*G.* gallus]	
GGEZEB1007F03	CO505763	LOC422945	4	C20orf194 chromosome 20 open reading frame 194 [*G. gallus*]	Putative GTPases (G3E family) [General function prediction only]
GGEZEB1007F09	CO505769	LOC772299	5	hypothetical protein LOC772299 [*G. gallus*]	
GGEZLB1011E06	CD764252	LOC423192	5	KIAA0652 KIAA0652 [*G. gallus*]	ATG13; Uncharacterized conserved protein (DUF2224): Members of this family of phosphoproteins are involved in cytoplasm to vacuole transport (Cvt), and more specifically in Cvt vesicle formation. They are probably involved in the switching machinery regulating the conversion between the Cvt pathway and autophagy. Finally, ATG13 is also required for glycogen storage. DUF2224; Uncharacterized conserved protein (DUF2224): The proteins in this highly conserved family are found from worms to humans. The function is unknown.
GGEZSM1030E08	CD762963	LOC423481	5	C14orf172 chromosome 14 open reading frame 172 [*G. gallus*]	NADB_Rossmann
GGEZEB2005A05	CO507354	LOC426270	6	LOC426270 similar to DOCK180 protein [*G. gallus*]	SH3 domain: SH3 (Src homology 3) domains are often indicative of a protein involved in signal transduction related to cytoskeletal organization.
GGEZEB2010D11	CO507773	LOC423610	6	LOC423610 similar to KIAA0613 protein [*G. gallus*]	
GGEZEB2011C07	CO507072	LOC429153	9	LOC429153 hypothetical LOC429153 [*G. gallus*]	Cytochrome P450 domain
GGEZLB1003D07	CD763598	C9orf58	17	C9orf58 chromosome 9 open reading frame 58 [*G. gallus*]	EF hand: EF-hand, calcium binding motif; A diverse superfamily of calcium sensors and calcium signal modulators; most examples in this alignment model have two active canonical EF hands.
GGEZEB2004G11	CO506751	LOC770563	28	C19orf22 chromosome 19 open reading frame 22 [*G. gallus*]	R3H domain. The name of the R3H domain comes from the characteristic spacing of the most conserved arginine and histidine residues. R3H domains are found in proteins together with ATPase domains, SF1 helicase domains, SF2 DEAH helicase domains, Cys-rich repeats, ring-type zinc fingers, and KH domains. The function of this domain is predicted to be binding of ssDNA or ssRNA in a sequence-specific manner.

Continued on next page

Table 4. Continued.

Clone name	Accession #	LOC	Chromosome	Description	Motifs or conserved domains
GGEZCB1005E03	CO503121	LOC771404	Un	hypothetical protein LOC771404 [*G. gallus*]	Homeodomain; DNA binding domains involved in the transcriptional regulation of key eukaryotic developmental processes; they may bind to DNA as monomers or as homo- and/or heterodimers, in a sequence-specific manner.
GGEZEB2019F01	CO507251	LOC776536	Un	LOC776536 hypothetical protein LOC776536 [*G. gallus*]	
GGEZEB1014G07	CO505557	LOC427196	Z	LOC427196 similar to MGC83563 protein [*G. gallus*]	NNT nicotinamide nucleotide transhydrogenase [*G. gallus*]
GGEZEB1019C05	CO506042	Gga.7682		hypothetical protein *G. gallus*	similar to Manbal protein
			‘Liver’ unknown HETs (Total = 32)		
GGEZCB1004A02	CO503312	LOC418871	1	chromosome 13 open reading frame 1 [*G. gallus*]	SPRY domain: SPRY domain is named from SPla and the RYanodine Receptor. Domain of unknown function. Distant homologues are domains in butyrophilin/marenostrin/pyrin homologues.
GGEZEB2010E12	CO507784	Gga.12126	1	PREDICTED: hypothetical protein *G. gallus*	Coiled-coil domain containing 90B; Protein of unknown function (DUF1640).
GGEZLB1026F02	CD765214	C21orf66	1	C21orf66 chromosome 21 open reading frame 66 [*G. gallus*]	GCFC; GC-rich sequence DNA-binding factor-like protein: Sequences found in this family are similar to a region of a human GC-rich sequence DNA-binding factor homolog. This is thought to be a protein involved in transcriptional regulation due to partial homologies to a transcription repressor and histone-interacting protein.
GGEZSM1011D11	CD761506	LOC418725	1	C2orf49 chromosome 2 open reading frame 49 [*G. gallus*]	
GGEZEB2018G08	CO506887	LOC768645	2	LOC768645 hypothetical protein LOC768645 [*G. gallus*]	
GGEZSM1024D09	CD762427	LOC420493	2	LOC420493 hypothetical LOC420493 [*G. gallus*]	TNFR/NGFR cysteine-rich region: Tumor necrosis factor receptor (TNFR) domain; superfamily of TNF-like receptor domains. When bound to TNF-like cytokines, TNFRs trigger multiple signal transduction pathways, they are involved in inflammation response, apoptosis, autoimmunity and organogenesis.
GGEZEB1018F08	CO506627	LOC416719	3	C20orf72 chromosome 20 open reading frame 72 [*G. gallus*]	RecB: ATP-dependent exoDNAse (exonuclease V) beta subunit (contains helicase and exonuclease domains) [DNA replication, recombination, and repair]
GGEZEB1022C03	CO506562	LOC395778	3	C6orf72 chromosome 6 open reading frame 72 [*G. gallus*]	
GGEZEB1021E06	CO506655	LOC422542	4	C4orf20 chromosome 4 open reading frame 20 [*G. gallus*]	Peptidase_C78; Peptidase family C78: This family formerly known as DUF1671 has been shown to be a cysteine peptidase called (Ufm1)-specific protease.
GGEZSM1028E03	CD762787	LOC428774	4	LOC428774 hypothetical LOC428774 [*G. gallus*]	Similar to PDZ domain containing 8
GGEZEB2001E03	CO506794	Gga.46793	5	similar to chromosome 5 open reading frame 5 (LOC770655)	
GGEZLB1003D06	CD763597	LOC421605	5	C11orf46 chromosome 11 open reading frame 46 [*G. gallus*]	
GGEZEB1025C08	CO505069	LOC423781	6	LOC423781 similar to FLJ00156 protein [*G. gallus*]	WDFY4 WDFY family member 4 [*G. gallus*]; BEACH (Beige and Chediak-Higashi) domains, implicated in membrane trafficking,
GGEZEB1017C03	CO506312	LOC424241	7	LOC424241 hypothetical LOC424241 [*G. gallus*]	RhoGEF domain: Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases; Also called Dbl-homologous (DH)

Continued on next page

Table 4. Continued.

Clone name	Accession #	LOC	Chromosome	Description	Motifs or conserved domains
					domain. It appears that PH domains invariably occur C-terminal to RhoGEF/DH domains.
GGEZEM1003A02	CD763200	LOC429154	9	LOC429154 hypothetical LOC429154 [*G. gallus*]	
GGEZLB1010A01	CD764119	LOC424773	9	LOC424773 similar to KIAA0332 [*G. gallus*]	Surp module: This domain is also known as the SWAP domain. SWAP stands for Suppressor-of-White-APricot. It has been suggested that these domains are RNA binding.
GGEZSM1015G03	CD761866	LOC416056	12	LOC416056 hypothetical LOC416056 [*G. gallus*]	Neurotransmitter transport; Sodium:neurotransmitter symporter family
GGEZSM1005E10	CD761082	KIAA0430	14	KIAA0430 KIAA0430 [*G. gallus*]	Macolin: transmembrane protein (pfam09726); RRM (RNA recognition motif), also known as RBD (RNA binding domain) or RNP (ribonucleoprotein domain), is a highly abundant domain in eukaryotes found in proteins involved in post-transcriptional gene expression processes including mRNA and rRNA processing, RNA export, and RNA stability.
GGEZLB1010C03	CD764143	LOC771533	18	similar to chromosome 17 open reading frame 26	SLC39A11 solute carrier family 39 (metal ion transporter), member 11 [*G. gallus*]. ZIP Zinc transporter: The ZIP family consists of zinc transport proteins and many putative metal transporters.
GGEZSM1019A03	CD761969	LOC422071	18	LOC422071 hypothetical gene supported by CR407540 [*G. gallus*]	
GGEZEB1016D04	CO505309	LOC419279	20	C20orf160 chromosome 20 open reading frame 160 [*G. gallus*]	
GGEZEB1020E05	CO505678	LOC419329	20	C20orf43 chromosome 20 open reading frame 43 [*G. gallus*]	DUF602; Protein of unknown function, DUF602: This family represents several uncharacterized eukaryotic proteins. [pfam04641]; oxidative stress responsive 1
GGEZEM1004H04	CD763342	LOC771089	20	LOC771089 hypothetical protein LOC771089 [*G. gallus*]	TSP_1 super-family; Thrombospondin type 1 domain.
GGEZSM1011B08	CD761482	LOC419203	20	C20orf111 chromosome 20 open reading frame 111 [*G. gallus*]	DUF776; Protein of unknown function (DUF776): This family consists of several highly related mouse and human proteins of unknown function. [pfam05604]
GGEZSM1011F04	CD761522	Gga.41998	20	hypothetical protein	
GGEZSM1029A08	CD762830	C1orf151	21	chromosome 1 open reading frame 151 [*G. gallus*]	DUF543; Domain of unknown function (DUF543): This family of short eukaryotic proteins has no known function.
GGEZEB2012A04	CO507433	LOC426357	25	LOC426357 hypothetical LOC426357 [*G. gallus*]	
GGEZSM1007B10	CD761203	LOC420161	28	C19orf10 chromosome 19 open reading frame 10 [*G. gallus*]	UPF0556; Uncharacterized protein family UPF0556: This family of proteins has no known function. [pfam10572]
GGEZEB2003A03	CO507127	Gga.17527	Un	*G. gallus* finished cDNA, clone ChEST992e8	
GGEZSM1027A11	CD762666	Gga.34499	Un	*G. gallus* finished cDNA, clone ChEST293f16	
GGEZLB1022A05	CD764921	Gga.36089	Z	*G. gallus* finished cDNA, clone ChEST128e22	
GGEZSM1027G05	CD762728	LOC768733	Z	LOC768733 hypothetical protein LOC768733 [*G. gallus*]	
			‘Brain’ unknown HETs (Total = 40)		
GGEZEB1001F07	CO506283	LOC771099	1	C12orf57 chromosome 12 open reading frame 57 [*G. gallus*]	
GGEZEB1011H02	CO505932	LOC770190	1	LOC770190 hypothetical protein LOC770190 [*G. gallus*]	COX17; Cytochrome c oxidase (CCO) copper chaperone (COX17): Cox17 is essential for the assembly of functional CCO and for delivery of copper ions to the

Continued on next page

Table 4. Continued.

Clone name	Accession #	LOC	Chromosome	Description	Motifs or conserved domains
					mitochondrion for insertion into the enzyme
GGEZSM1020H02	CD762121	LOC770530	1	C7orf55 chromosome 7 open reading frame 55 [*G. gallus*]	FMC1 protein family: This family of proteins is related to the yeast FMC1 protein that is required for assembly and stability of mitochondrial F(1)-ATPase. [pfam10560]
GGEZSM1022F02	CD762271	LOC418472	1	LOC418472 hypothetical LOC418472 [*G. gallus*]	
GGEZCB1019B05	CO504371	LOC420370	2	similar to KIAA1285 protein [*G. gallus*]	KRAB domain (or Kruppel-associated box) is present in about a third of zinc finger proteins containing C2H2 fingers. The KRAB domain is found to be involved in protein-protein interactions.
GGEZEB1014D05	CO505528	LOC420252	2	LOC420252 similar to KIAA0896 protein [*G. gallus*]	
GGEZEB1023A12	CO505441	LOC420961	2	C9orf4 chromosome 9 open reading frame 4 [*G. gallus*]	DOMON; Protein of unknown function;
GGEZLB1010A05	CD764123	LOC768667	2	LOC768667 hypothetical protein LOC768667 [*G. gallus*]	PKc_like; PhoP regulatory network protein YrbL: The protein kinase superfamily is mainly composed of the catalytic domains of serine/threonine-specific and tyrosine-specific protein kinases. Kdo; Lipopolysaccharide kinase (Kdo/WaaP) family.
GGEZLB1022A02	CD764919	LOC420907	2	KIAA1468 KIAA1468 [*G. gallus*]	SMC (structural maintenance of chromosomes): Chromosome segregation ATPases [Cell division and chromosome partitioning]; RecF/RecN/SMC N terminal domain: This domain is found at the N terminus of SMC proteins. The SMC superfamily proteins have ATP-binding domains at the N- and C-termini, and two extended coiled-coil domains separated by a hinge in the middle.
GGEZSM1022D01	CD762248	LOC420983	2	C9orf6 chromosome 9 open reading frame 6 [*G. gallus*]	GCV_H; Glycine cleavage H-protein: This is a family of glycine cleavage H-proteins, part of the glycine cleavage multienzyme complex (GCV) found in bacteria and the mitochondria of eukaryotes.
GGEZSM1025B02	CD762487	LOC420764	2	C7orf36 chromosome 7 open reading frame 36 [*G. gallus*]	FliH; Nodulation protein NolV: [cl03451|122155]; Yae1_N; Essential protein Yae1, N terminal: Members of this family are found in the N terminal region of the essential protein Yae1. Their exact function has not as yet been determined. [pfam09811]
GGEZLB1016E06	CD764497	LOC422195	4	CXorf34 chromosome X open reading frame 34 [*G. gallus*]	NADB_Rossmann;
GGEZEB1001G06	CO506289	LOC423293	5	C15orf41 chromosome 15 open reading frame 41 [*G. gallus*]	
GGEZEB2015A08	CO506905	LOC768377	5	LOC768377 hypothetical protein LOC768377 [*G. gallus*]	PKc_like; PhoP regulatory network protein
GGEZEM1001A09	CD763087	LOC426752	5	LOC426752 hypothetical LOC426752 [*G. gallus*]	S15/NS1/EPRS_RNA-binding domain; Ribosomal protein S15 domain;
GGEZLB1010F11	CD764180	LOC423494	5	LOC423494 hypothetical LOC423494 [*G. gallus*]	FHA domain (Forkhead-associated domain): found in eukaryotic and prokaryotic proteins. Putative nuclear signalling domain.
GGEZSM1009E10	CD761395	LOC423192	5	KIAA0652 KIAA0652 [*G. gallus*]	ATG13;
GGEZEB1034F09	CO505997	LOC423782	6	C10orf72 chromosome 10 open reading frame 72 [*G. gallus*]	
GGEZEB1011E08	CO505905	LOC424385	8	LOC424385 similar to chromosome 1 open reading frame 9 [*G. gallus*]	SMC_N; RecF/RecN/SMC N terminal domain: This domain is found at the N terminus of structural maintenance of chromosome (SMC) proteins. The SMC

Continued on next page

Table 4. Continued.

Clone name	Accession #	LOC	Chromosome	Description	Motifs or conserved domains
					superfamily proteins have ATP-binding domains at the N- and C-termini, and two extended coiled-coil domains separated by a hinge in the middle.
GGEZSM1015G09	CD761869	LOC424580	8	C1orf164 chromosome 1 open reading frame 164 [*G. gallus*]	RING; U-box domain: RING-finger (Really Interesting New Gene) domain, a specialized type of Zn-finger of 40 to 60 residues that binds two atoms of zinc;
GGEZSM1012G03	CD761612	LOC415704	11	LOC415704 similar to KIAA0091 [*G. gallus*]	MBTPS1 membrane-bound transcription factor peptidase, site 1 [*G. gallus*]
GGEZEB2015G12	CO506967	LOC417886	12	C12orf29 chromosome 12 open reading frame 29 [*G. gallus*]	
GGEZCB1016F10	CO504125	LOC416357	13	hypothetical LOC416357 [*G. gallus*]	
GGEZEB2005G05	CO507414	LOC769251	14	LOC769251 hypothetical protein LOC769251 [*G. gallus*]	
GGEZSM1013F09	CD761689	LOC425771	16	LOC425771 hypothetical LOC425771 [*G. gallus*]	ZNF692 zinc finger protein 692 [*G. gallus*]
GGEZEB1020H07	CO505714	C9orf58	17	C9orf58 chromosome 9 open reading frame 58 [*G. gallus*]	Allograft inflammatory factor 1-like
GGEZEB1022C04	CO506563	LOC770371	18	LOC770371 hypothetical protein LOC770371 [*G. gallus*]	
GGEZEM1001A12	CD763088	LOC769908	18	LOC769908 hypothetical protein LOC769908 [*G. gallus*]	
GGEZEB1023B09	CO505446	LOC419221	20	C20orf177 chromosome 20 open reading frame 177 [*G. gallus*]	
GGEZEB2003E11	CO507177	LOC420073	28	LOC420073 similar to R31449_3 [*G. gallus*]	
GGEZLB1003E10	CD763613	LOC420101	28	C19orf6 chromosome 19 open reading frame 6 [*G. gallus*]	Membralin; Tumor-associated protein: Membralin is evolutionarily highly conserved; though it seems to represent a unique protein family.
GGEZEB1005H06	CO505429	LOC425086	Un	LOC425086 hypothetical LOC425086 [*G. gallus*]	Perilipin family: The perilipin family includes lipid droplet-associated protein (perilipin) and adipose differentiation-related protein (adipophilin). [pfam03036]
GGEZEB1017F10	CO506340	LOC772194	Un	C1orf78 chromosome 1 open reading frame 78 [*G. gallus*]	
GGEZEB2019F11	CO507258	LOC777501	Un	LOC777501 hypothetical protein LOC777501 [*G. gallus*]	nidG2; G2F domain: Nidogen, G2 domain; Nidogen is an important component of the basement membrane, an extracellular sheet-like matrix. Nidogen is a multifunctional protein that interacts with many other basement membrane proteins, like collagen, perlecan, lamin, and it has a potential role in the assembly and connection of networks.
GGEZLB1002D11	CD763513	LOC771478	Un	LOC771478 hypothetical protein LOC771478 [*G. gallus*]	
GGEZLB1009E04	CD764080	LOC776802	Un	LOC776802 hypothetical protein LOC776802 [*G. gallus*]	ChSh; Chromo shadow domain: Chromo Shadow Domain, found in association with N-terminal chromo (CHRromatin Organization MOdifier) domain; Chromo domains mediate the interaction of the heterochromatin with other heterochromatin proteins, thereby affecting chromatin structure
GGEZLB1010D05	CD764156	Gga.12325	Un	hypothetical protein	
GGEZEB1026D06	CO505224	3.1	Z	C5orf13 chromosome 5 open reading frame 13 [*G. gallus*]	P311 POU
GGEZEB2012H08	CO507509	LOC426891	Z	C18orf10 chromosome 18 open reading frame 10 [*G. gallus*]	
GGEZSM1027A10	CD762665	LOC427408	Z	KIAA1045 KIAA1045 [*G. gallus*]	

Chromosome Un = unknown.

To further characterize the unknown highly expressed transcript, public databases were searched; only motifs or conserved domains were identified in some of those transcripts (Table 4). Among 21 'skeletal muscle' unknown HETs, for example, conserved domains could be found in 10 of them, including a nucleotide-glucose transporter domain (EST accession number CD761013, LOC421073), and three NADB-Rossmann domains (CO504294, CO504070, CD761892, corresponding to LOC771456, LOC429066 and LOC425539, respectively), which have been found in numerous dehydrogenases of metabolic pathways, such as glycolysis, and many other redox enzymes. As at least one previously conserved domain or motif was identified in those unknown transcripts; they can be considered as 'proteins with defined features' (PDFs; Gollery et al., 2007).

Eleven novel PDFs were also found among chicken 'heart' unknown HETs, including GC-rich sequence DNA-binding factor-like protein (CD764046, C21orf66); an SH3 domain, involved in signal transduction related to cytoskeletal organization (CO507354, LOC426270), and one with the same NADB-Rossmann domain (at different loci and EST, CD762963, LOC423481), found among 'skeletal muscle' PDFs.

Conserved domains were found in 18 transcripts from the 32 'liver' unknown HETs, including a cysteine peptidase called Ufm1-specific protease (CO506655, LOC422542); a WDFY family member 4 domain (CO505069, LOC423781), and one with a solute carrier family 39, which is a metal-ion-transporter domain. Finally, 23 PDFs were identified among the 40 'brain' unknown HETs, including two with protein kinase c like protein domains (CO506905, LOC768377 and CD764123, LOC768667).

On the other hand, among the selected 112 unknown HETs, 50 remained as proteins that lack defined motifs and conserved domains. These proteins are currently defined as proteins with obscure features (POFs). Increased attention has been recently given to these POF sequences as, on average, they represent 15-40% of the genes encoded in every eukaryotic genome sequenced to date (Gollery et al., 2006, 2007). POFs are considered to represent newly evolving genes or genes that are evolving faster than the genome average; they also contribute to determine species specificity (Galperin and Koonin, 2004; Gollery et al., 2007).

There was also evidence of alternative splicing among the PDFs. For instance, the locus c21orf66 encoding for a PDF with a GC-rich sequence DNA-binding factor-like protein domain was found among 'heart' and 'liver' uncharacterized HET. Using the MapViewer tool from NCBI, we observed that corresponding ESTs (CD765214 and CD764046) were distant from each other in the gene sequence (data not shown), and that CD764046 was positioned in an intronic region of the gene sequence in chicken chromosome 1. A number of possibilities can explain intronic ESTs, including the presence of this intron in one of the transcribed mRNAs (reviewed by Graveley, 2001).

LOC772299 was also found as duplicated unknown HETs, identified among 'skeletal muscle' and 'heart' unknown HETs. Although no conserved domains have been found, the expression pattern of this POF could suggest a novel muscle gene, alternatively expressed between heart and skeletal muscle. Other duplicated HETs were two ESTs from the locus c9orf58, identified as 'brain' and 'heart' HETs, encoding for a PDF with a calcium-binding domain called EF-hand; the locus LOC395778 (c6orf72) duplicated in 'liver' and 'heart', and the locus LOC423192, duplicated in 'heart' and 'brain' expression profiles and encoding for a PDF with an ATG13 uncharacterized domain, associated with autophagy and probably with glycogen storage (Scott et al., 2000; Table 4).

ACKNOWLEDGMENTS

Research supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). L.L. Coutinho is the recipient of a scholarship from CNPq.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.

Afrakhte M and Schultheiss TM (2004). Construction and analysis of a subtracted library and microarray of cDNAs expressed specifically in chicken heart progenitor cells. *Dev. Dyn.* 230: 290-298.

Baldessari D, Shin Y, Krebs O, König R, et al. (2005). Global gene expression profiling and cluster analysis in *Xenopus laevis*. *Mech. Dev.* 122: 441-475.

Bliss TW, Dohms JE, Emara MG and Keeler CL Jr (2005). Gene expression profiling of avian macrophage activation. *Vet. Immunol. Immunopathol.* 105: 289-299.

Bono H, Yagi K, Kasukawa T, Nikaido I, et al. (2003). Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays. *Genome Res.* 13: 1318-1323.

Buza TJ, McCarthy FM and Burgess SC (2007). Experimental-confirmation and functional-annotation of predicted proteins in the chicken genome. *BMC Genomics* 8: 425.

Cogburn LA, Porter TE, Duclos MJ, Simon J, et al. (2007). Functional genomics of the chicken - a model organism. *Poult. Sci.* 86: 2059-2094.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.

Fukata Y, Iwanaga T and Fukata M (2006). Systematic screening for palmitoyl transferase activity of the DHHC protein family in mammalian cells. *Methods* 40: 177-182.

Galperin MY and Koonin EV (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32: 5452-5463.

Gollery M, Harper J, Cushman J, Mittler T, et al. (2006). What makes species unique? The contribution of proteins with obscure features. *Genome Biol.* 7: R57.

Gollery M, Harper J, Cushman J, Mittler T, et al. (2007). POFs: what we don't know can hurt us. *Trends Plant Sci.* 12: 492-496.

Graveley BR (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17: 100-107.

Hamburger V and Hamilton HL (1951). A series of normal stages in the development of the chick embryo. *J. Morphol.* 88: 49-92.

Hata K, Fujitani M, Yasuda Y, Doya H, et al. (2006). RGMa inhibition promotes axonal growth and recovery after spinal cord injury. *J. Cell Biol.* 173: 47-58.

Hillier LW, Miller W, Birney E and Warren W, et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695-716.

Huang X and Madan A (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877.

Jorge EC, Monteiro-Vitorelo CB, Alves HJ and Silva CS, et al. (2004). EST analysis of mRNA expressed during embryogenesis in *Gallus gallus*. *Int. J. Dev. Biol.* 48: 333-337.

Jorge EC, Figueira A, Ledur MC and Moura ASAMT, et al. (2007). Contributions and perspectives of chicken genomics in Brazil: from biological model to export commodity. *Worlds Poult. Sci. J.* 63: 597-610.

Kuninger D, Kuzmickas R, Peng B, Pintar JE, et al. (2004). Gene discovery by microarray: identification of novel genes induced during growth factor-mediated muscle cell survival and differentiation. *Genomics* 84: 876-889.

Kusano S and Raab-Traub N (2002). I-mfa domain proteins interact with Axin and affect its regulation of the Wnt and c-Jun N-terminal kinase signaling pathways. *Mol. Cell Biol.* 22: 6393-6405.

Matsunaga E and Chédotal A (2004). Repulsive guidance molecule/neogenin: a novel ligand-receptor system playing multiple roles in neural development. *Dev. Growth Differ.* 46: 481-486.

Mawdsley DJ, Cooper HM, Hogan BM, Cody SH, et al. (2004). The Netrin receptor Neogenin is required for neural tube formation and somitogenesis in zebrafish. *Dev. Biol.* 269: 302-315.

Monnier PP, Sierra A, Macchi P, Deitinghoff L, et al. (2002). RGM is a repulsive guidance molecule for retinal axons. *Nature* 419: 392-395.

Nagaya M, Inohaya K, Imai Y and Kudo A (2002). Expression of zisp, a DHHC zinc finger gene, in somites and lens during zebrafish embryogenesis. *Mech. Dev.* 119 (Suppl 1): S311-S314.

Neiman PE, Ruddell A, Jasoni C, Loring G, et al. (2001). Analysis of gene expression during myc oncogene-induced lymphomagenesis in the bursa of Fabricius. *Proc. Natl. Acad. Sci. USA* 98: 6378-6383.

Niederkofler V, Salie R, Sigrist M and Arber S (2004). Repulsive guidance molecule (RGM) gene function is required for neural tube closure but not retinal topography in the mouse visual system. *J. Neurosci.* 24: 808-818.

Niehrs C and Pollet N (1999). Synexpression groups in eukaryotes. *Nature* 402: 483-487.

Papanikolaou G, Samuels ME, Ludwig EH, MacDonald ML, et al. (2004). Mutations in HFE2 cause iron overload in chromosome 1q-linked juvenile hemochromatosis. *Nat. Genet.* 36: 77-82.

Postigo AA and Dean DC (1999). Independent repressor domains in ZEB regulate muscle and T-cell differentiation. *Mol. Cell Biol.* 19: 7961-7971.

Ross M, Labrie T, McPherson S and Stanton VP (1999). Screening large-insert libraries by hybridization. *Curr. Protoc. Hum. Genet.* 5.6.1-5.6.32. DOI: 10.1002/0471142905.hg0506s21.

Sambrook J, Fritsch EF and Maniatis T (1989). Molecular Cloning: a Laboratory Manual. 2nd edn. Cold Spring Harbor Laboratory Press, Plainview.

Scott SV, Nice DC III, Nau JJ, Weisman LS, et al. (2000). Apg13p and Vac8p are part of a complex of phosphoproteins that are required for cytoplasm to vacuole targeting. *J. Biol. Chem.* 275: 25840-25849.

Smith J, Speed D, Hocking PM, Talbot RT, et al. (2006). Development of a chicken 5K microarray targeted towards immune function. *BMC Genomics* 7: 49.

Soares MB, Bonaldo MF, Jelene P, Su L, et al. (1994). Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. USA* 91: 9228-9232.

Szustakowski JD, Lee JH, Marrese CA, Kosinski PA, et al. (2006). Identification of novel pathway regulation during myogenic differentiation. *Genomics* 87: 129-138.

van Hemert S, Ebbelaar BH, Smits MA and Rebel JM (2003). Generation of EST and microarray resources for functional genomic studies on chicken intestinal health. *Anim. Biotechnol.* 14: 133-143.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8: 625-637.

Woods S, Farrall A, Procko C and Whitelaw ML (2008). The bHLH/Per-Arnt-Sim transcription factor SIM2 regulates muscle transcript myomesin2 via a novel, non-canonical E-box sequence. *Nucleic Acids Res.* 36: 3716-3727.

Zhang W, Morris QD, Chang R, Shai O, et al. (2004). The functional landscape of mouse gene expression. *J. Biol.* 3: 21.