Supporting information

RNA-Seq reveals conservation of function among the yolk sacs of human, mouse and chicken

Tereza Cindrova-Davies, Eric Jauniaux, Michael Elliot, Sungsam Gong, Graham J Burton, D. Stephen Charnock-Jones

Methods

Proteomics. 1D gel lanes were cut into 8 bands and each band was transferred into a 96-well PCR plate. The bands were cut into 1mm^2 pieces, destained, reduced (DTT) and alkylated (iodoacetamide) and subjected to enzymatic digestion with trypsin overnight at 37°C. After digestion, the supernatant was pipetted into a sample vial and loaded onto an autosampler for automated LC-MS/MS analysis.

All LC-MS/MS experiments were performed using a Dionex Ultimate 3000 RSLC nanoUPLC (Thermo Fisher Scientific Inc, Waltham, MA, USA) system and a QExactive Orbitrap mass spectrometer (Thermo Fisher Scientific Inc, Waltham, MA, USA). Separation of peptides was performed by reverse-phase chromatography at a flow rate of 300 nL/min and a Thermo Scientific reverse-phase nano Easy-spray column (Thermo Scientific PepMap C18, 2µm particle size, 100Å pore size, 75µm i.d. x 50cm length). Peptides were loaded onto a pre-column (Thermo Scientific PepMap 100 C18, 5µm particle size, 100Å pore size, 300µm i.d. x 5mm length) from the Ultimate 3000 autosampler with 0.1% formic acid for 3 minutes at a flow rate of 10 µL/min. After this period, the column valve was switched to allow elution of peptides from the pre-column onto the analytical column. Solvent A was water + 0.1% formic acid and solvent B was 80% acetonitrile, 20% water + 0.1% formic acid. The linear gradient employed was 2-40% B in 30 minutes.

The LC eluant was sprayed into the mass spectrometer by means of an Easy-Spray source (Thermo Fisher Scientific Inc.). All *m/z* values of eluting ions were measured in an Orbitrap mass analyzer, set at a resolution of 70000 and was scanned between *m/z* 380-1500. Data dependent scans (Top 20) were employed to automatically isolate and generate fragment ions by higher energy collisional dissociation (HCD, NCE:25%) in the HCD collision cell and measurement of the resulting fragment ions was performed in the Orbitrap analyser, set at a resolution of 17500. Singly charged ions and ions with unassigned charge states were excluded from being selected for MS/MS and a dynamic exclusion window of 20 seconds was employed.

Post-run, the data was processed using Protein Discoverer (version 1.4., ThermoFisher). Briefly, all MS/MS data were converted to mgf files and the files were then submitted to the Mascot search algorithm (Matrix Science, London UK) and searched against the UniProt human database (153168 sequences; 54677058 residues). Variable modifications of oxidation (M) and deamidation (NQ) and a fixed modification of carbamidomethyl (C) were

applied and the peptide and fragment mass tolerances were set to 5ppm and 0.1 Da, respectively. A significance threshold value of $p<0.05$ and a peptide cut-off score of 20 were also applied. All the Mascot data from each lane was merged to give a single output file. Finally, Mascot .dat files were entered into the Scaffold software (Proteome Software, Oregon USA) so that differences between the two protein samples could be compared.

Two samples had very few recognizable peptide hits these samples were excluded from the subsequent analysis. Unique hits to identifiable proteins were counted and those with the same gene name in the identifier filed combined. The corresponding gene names for the proteins detected in 4 out of the 5 samples were used in Gene Ontology analysis as described below (Panther).

Bioinformatic analysis

RNA-Seq data processing

Quality assessment and trimming of the de-multiplex reads was carried out using FastQC and cutadapt respectively (http://www.bioinformatics.babraham.ac.uk/projects/fastqc(1). The trimmed short reads were mapped to the human reference genome (hg19) using TopHat2 (version 2.0.12), a splice-aware mapper(2). Uniquely mapped reads were counted using HTSeq (version 0.6.0)(3) and the relative transcript abundance determined using DESeq2 (version 1.6.3)(4).

Other datasets

To compare the levels of transcripts in the yolk sac and placental villi with those in the liver, lung and kidney processed RNA-Seq data was down loaded from the EBI Expression Atlas (version: 0.1.4-SNAPSHOT, experiment E-MTAB-513 http://www.ebi.ac.uk/gxa/experiments/E-MTAB-513?ref=aebrowse).

To compare with 27 other tissues processed RNA-Seq data was downloaded from Expression Atlas - E-MTAB-1733 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1733)(5).

GO analysis

The most abundant 400 transcripts in the tissues of interest were identified (ranked by mean RPKM) and gene ontology analyzed by using Panther(6) (http://pantherdb.org release 20160715). The complete Panther data base was used to include the mitochondrially encoded transcripts. The Biological Process, Molecular Function and Cellular Component were all analyzed. The reference sets were all Homo sapiens or Mus musculus genes (known in Panther) and the Bonferroni correction for multiple testing was applied in all cases.
Chord plots of enriched GO terms (Figure 1 for example) were generated in R using GOplot (v 1.0.2. available from CRAN) (7).

Venn diagrams

Venn diagrams were generated in R (vennCounts and VennDiagram, within the Bioconductor limma package) and the p-values for the observed overlaps calculated with SuperExactTest by Wang et al 2015 (v 0.99.2, available from CRAN)(8).

Transcription factor binding sites

Within the most abundant 400 transcripts in the human yolk sac, 19 genes are annotated as "regulation of transcription, DNA-templated" (GO:0006355) including several transcription factors (ATF4, FOS, FOSB, JUN, JUNB and JUND XBP-1 and BHLHE40). In the mouse 8 genes were similarly annotated (Table S3.)

The correlation (Spearman) between each of the transcripts encoding these transcription factors and all other yolk sac transcripts (rpkm >15) was calculated. The transcripts with rho >0.8 were identified. Over-represented transcription factor binding motifs in these genes were identified (JASPAR2016 and TFBSTools). Candidate motifs recognized by the transcription factors (or family for JUN and FOS) were enriched in the 1kb and 5kb upstream of the TSS of the highly correlated genes. Regions of interest were extracted using Biostrings and transcription factor binding motifs identified using JASPAR2016 and TFBSTools, all in R. P values were corrected for multiple testing(9) (Datasets S7 & 8).

Similar analysis was carried out for the mouse data in which 8 genes were annotated as above. Atf4 was the only transcription factor with a binding motif in the JASPAR2016 database. Candidate motifs recognized by Atf4 were enriched in the 1kb and 5kb upstream of the TSS of the highly correlated genes (Datasets S9 &10).

Other manipulations were carried out in R (BiomaRt, dplyr and made4) and Excel.

Results

Sample quality control

Hierarchical clustering of the yolk sac reads indicated that 1 sample did not cluster with any other sample and formed a separate branch. This sample had extremely high levels of villous specific transcripts (CGA, CGB, CGB5, CGB8, PSG1, PSG2, PSG3, PSG4, PSG9, LEP, KISS1 and CSH2; for example CGA ~73,000 vs a mean of ~320 for the other samples). It is therefore likely that the sample is contaminated with villous material. This sample was removed and the level of these transcripts examined in the remaining samples. Two additional samples had markedly different levels of these transcripts and so were also excluded from subsequent analysis.

RNASeq data from the first trimester villus samples were similarly processed and 1 sample clustered alone as a well separated branch. This sample had higher levels of decidual transcripts (such as IGF2, IGFPB4, IGFPB5, IGFPB7 and low levels of villous transcripts (CGA, CGB, CGB5, LEP and KISS1). This sample was therefore excluded from subsequent analysis.

A similar approach was adopted with the mouse yolk sac samples and 1 sample clustered as a well-separated branch. This sample had much lower levels of the fetal hemoglobin transcripts (Hbb-y, Hbb-bh1 and Hba-x) whereas these were the most abundant in all the other samples. This sample was therefore excluded from subsequent analysis.

Mitochondrial transcripts

Various enriched GO terms are associated with mitochondrial activity ("mitochondrial respiratory chain complex", and "mitochondrial ATP synthesis coupled electron transport" for example). This is consistent with the previously reported high density of mitochondria in the yolk sac(10). We calculated the proportion of transcripts present that were encoded by the mitochondrial genome (Figure S1). The fractions for the yolk sac and first trimester villi are slightly above the average and these data are consistent with that showing the kidney and heart had the highest mitochondrial fractions(11).

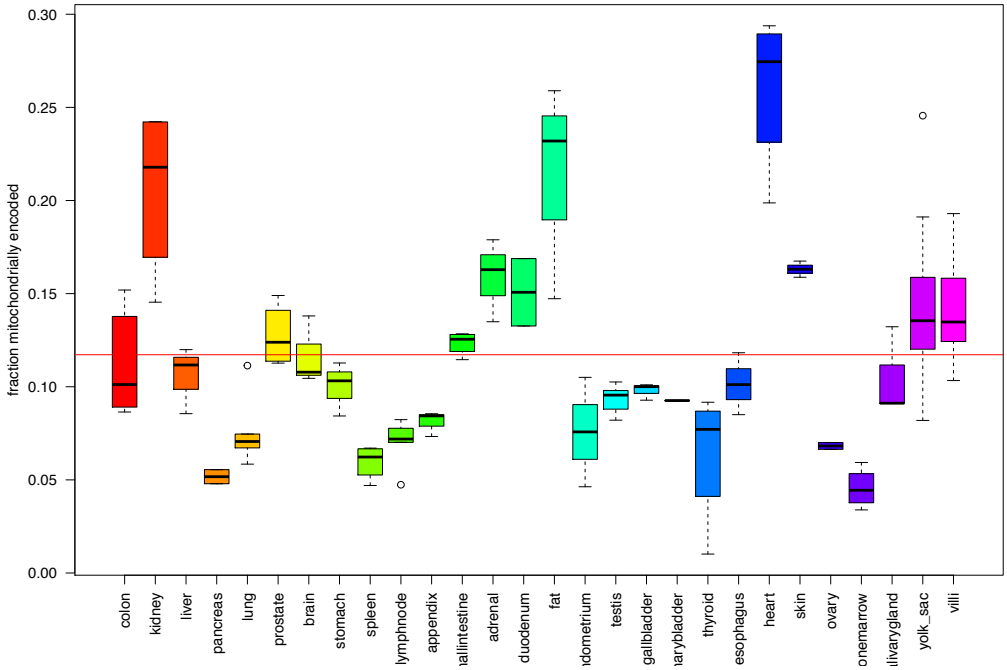


Figure S1. The fraction of mitochondrially encoded transcripts in each tissue type is the proportion mitochondrial genes over the sum over all RPKM values in that tissue

Mitochondrially encoded transcripts directly involved in the electron transfer chain and (or) energy generation were highly correlated with each other (Spearman's rho >0.71) and not with other transcripts in the most abundant 400 (Datasets S6). One other transcript was correlated with these (HNRNPA2B1). Other mitochondrially encoded transcripts (12S and 16S rRNA) were not similarly correlated.

Comparison with other tissues.

We used hierarchical clustering to examine the relationships among the expression patterns of the most variable transcripts detected at rpkm ≥ 0.1 in the yolk sac, placental villi and a range of other tissues. As previously reported, clustering recapitulated the origin of the tissue(5). The first trimester placental villi and the term placental samples clustered together with the yolk sac cluster forming an adjacent branch (Figure S2). This indicates the first and third trimester villi have similar transcript profiles and the 3 extra-embryonic tissues are more similar to each other than to the other tissues examined. The yolk sac samples form 2 closely linked branches, and on closer inspection of the samples falling in these 2 groups, we noted

a marked difference in the level of transcripts encoding ALB, AFP, HBE, HBZ. It is known from previous studies that the size of the yolk sac declines from a maximum diameter of ~6mm at approximately 10 weeks gestation (12). The gestational age of the samples used in this study is only available for a subset of the samples. However, for those where this is known the samples obtained earlier in gestation (7+0, 8+2, 9+0 and 9+2) are on a separate branch to the known later sample (11+0). Male and female samples are distributed evenly across both branches.

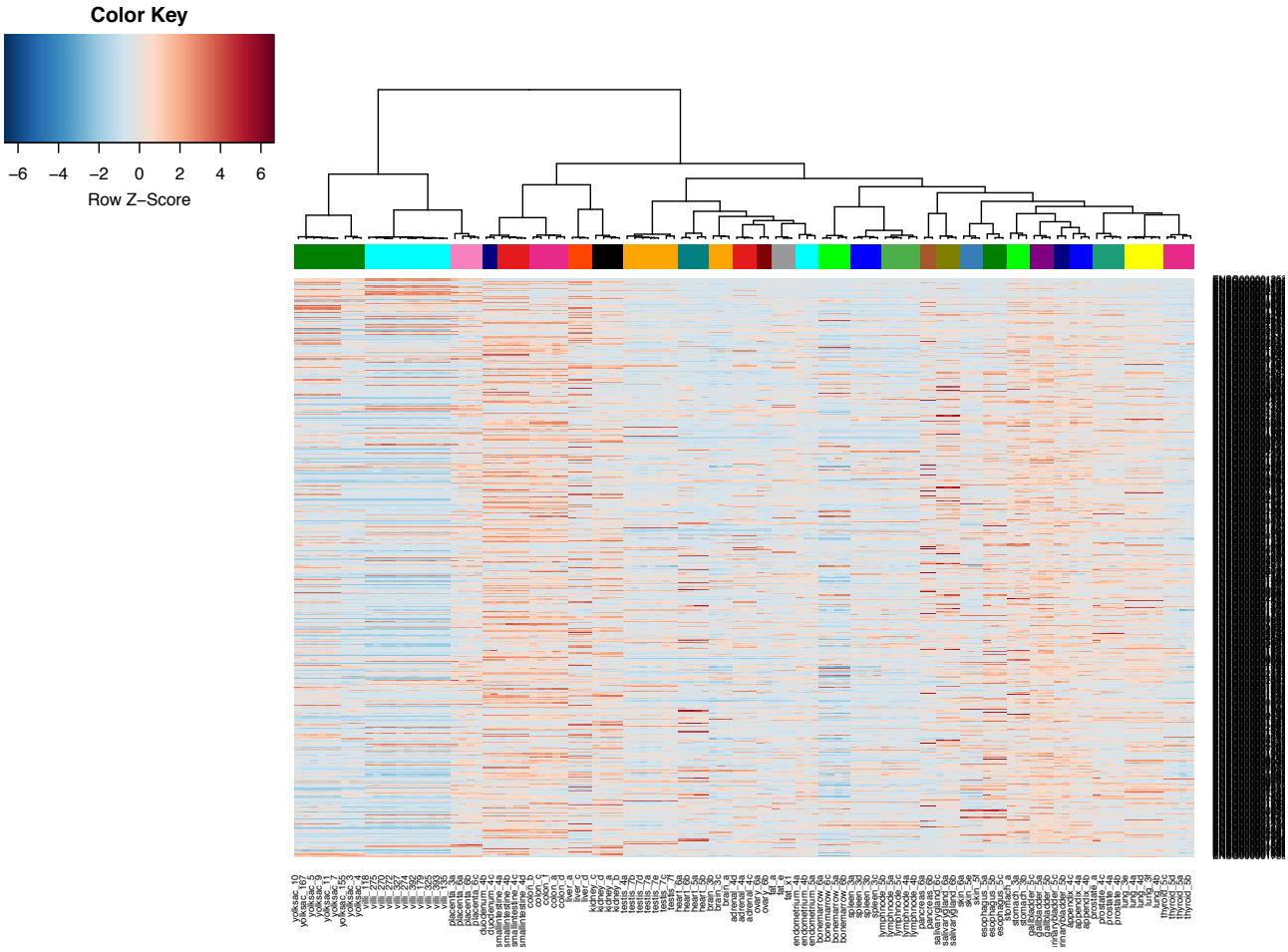
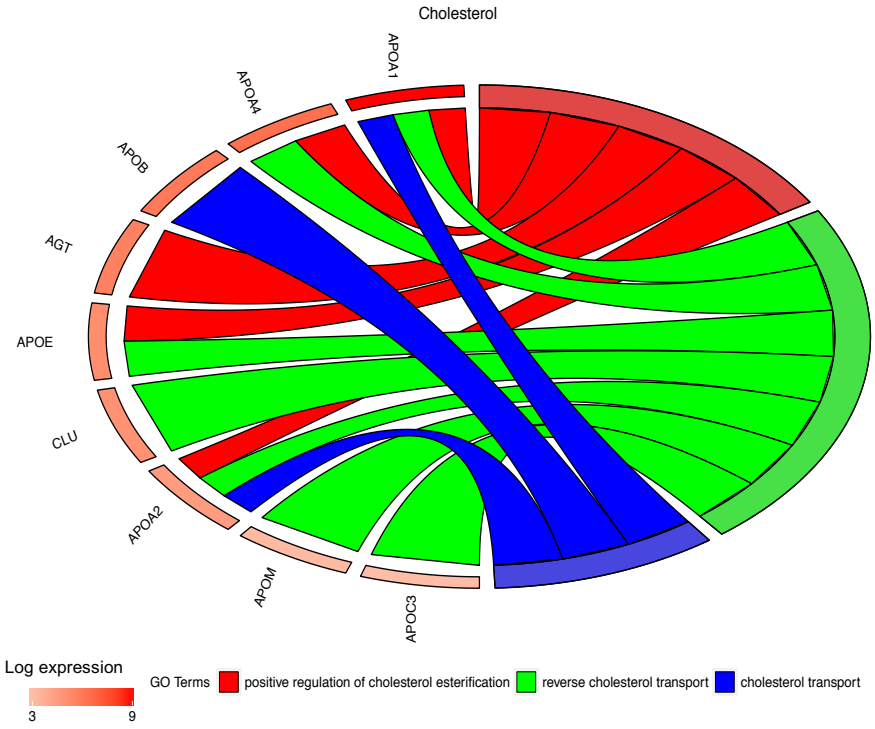


Figure S2. Hierarchical clustering of multiple tissues based on the 500 transcripts with the highest standard deviation(5).



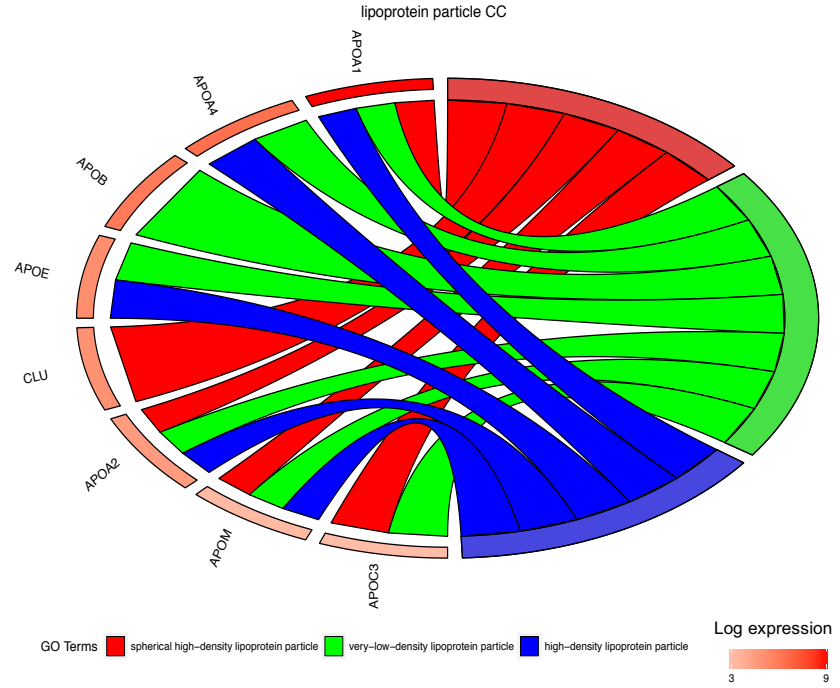


Figure S3 a & b.
Chord plots illustrating the proteins annotated with the GO molecular function and cellular components “cholesterol” and “lipoprotein particle” respectively are present coelomic fluid.

Table S1

Human Yolk sac		Human first trimester villi		Mouse yolk sac	
Sample ID	Reads	Sample ID	Reads	Sample ID	Reads
YS167	48,834,289	P270	4,058,853	B6YS1.2	36,961,551
YS7	63,379,887	P325	23,438,297	B6YS3.1	16,745,203
YS9	58,695,200	P392	12,959,581	B6YS9.1	24,340,906
YS10	65,310,562	P393	58,672,011	B6YS10.1	15,963,038
YS11	37,760,671	P118	76,386,477	B6YS11.1	64,903,212
YS3	39,471,048	P272	205,189,675	B6YS24.1	20,101,598
YS4	28,688,319	P274	30,427,351	B6YS27.1	32,624,089
YS5	33,224,068	P275	131,727,903	B6YS28.1	36,481,077
YS155	39,022,547	P179	22,238,388		
		P327	89,441,895		
median	39,471,048	median	30,427,351	median	28,482,498

Summary of the number of RNASeq reads (single-end 50base) obtained for the human and mouse yolk sacs and the first trimester human placental villi.

Table S2

Common circulating blood proteins	**Serum albumin**, Afamin (albumin family, vit E-binding), Ig (κC, γ-1,-2, λ, V-II, V-III), **Apolipoproteins** (**A-I, A-II, A-IV, B, B100, D, E,** M), clusterin (apolipoprotien J), **heparan sulfate proteoglycan (basement-membrane-specific)**, α-2-macroglobulin, Hemopexin (β-1B-glycoprotein), α-1B-glycoprotein, **α-2-HS-glycoprotein**, **α-1-acid glycoprotein**, **Gelsolin**, **Fibrinogen α**, Vitronectin, **Zinc-α-2-glycoprotein**, Histidine-rich glycoprotein, **β-2-microglobulin, α-1-acid glycoprotein, Leucine-rich α-2-glycoprotein,** β-2-glycoprotein 1, Serum amyloid A-4 protein
Coagulation and complement factors	**Complement** (**B, C3, C4, C7, C9,** D**, H, I**), Kininogen, **Coagulation factors (V, XII)**
Blood transport and binding proteins	Serotransferrin, Ceruloplasmin, **Transthyretin, Vitamin D-binding protein, Retinol-binding protein**, Galectin-3-binding protein, Latent-transforming growth factor beta-binding protein, Phospholipid transfer protein, **Hemoglobin (α, β**, ε, ζ), **Thyroxine-binding globulin**, Insulin-like growth factor-binding protein 3, **Latent-transforming growth factor β-binding protein, Corticosteroid-binding globulin**
Protease inhibitors	**α-1-antitrypsin**, Antithrombin-III, **α-1-antichymotrypsin**, **Inter-α-trypsin inhibitor**, **Plasma protease C1 inhibitor**, **Inter-α-trypsin inhibitor**, α-2-antiplasmin, Cystatin c (potent inhibitor of lysosomal proteinases), **Heparin cofactor 2** (rapidly inhibits thrombin in the presence of dermatan sulfate or heparin), inhibits trypsin, plasmin, and lysosomal granulocytic elastase), Metalloproteinase inhibitor 1

Proteases and other enzymes	**Plasminogen, angiotensinogen**, Procollagen C-endopeptidase enhancer, Glutathione peroxidase, Prothrombin, Cu-Zn superoxide dismutase,
Cytokines and hormones	α-fetoprotein, Pigment epithelium-derived factor (serpin 1), Choriogonadotropin subunit β
Channel and receptor-derived peptides	
Miscellaneous (structural, nuclear etc.)	Fibronectin, Fibulin-1, Decorin (proteoglycan), **Collagen**s-α1, -2,-3, Laminin, Versican core protein, Periostin, **Keratins (I, II)**, Transforming growth factor-β-induced protein, Lumican (proteoglycan), Laminin (β1, γ1), glycodelin, Hyaluronan and proteoglycan link protein, nidogen-1 (entactin, basement memb), titin (striated muscle contraction), actin, Fibulin-2, Osteonectin (SPARC, a glycoprotein in the bone that binds calcium), Spondin-1 (axon guidance), Extracellular matrix protein 1, Dystroglycan (dystrophin-associated glycoprotein), Cadherin-1 (transmembrane protein)

Proteins in **bold** are also reported in human serum

Categorization of the proteins detected in the human coelomic fluid and comparison with serum proteins

Table S3

Human

ensembl gene id	hgnc symbol	description	GO linkage type
ENSG00000065978	YBX1	Y box binding protein 1	IDA
ENSG00000128272	ATF4	activating transcription factor 4	IEA, ISS
ENSG00000145741	BTF3	basic transcription factor 3	IEA
ENSG00000009307	CSDE1	cold shock domain containing E1, RNA-binding	IEA
ENSG00000156508	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	IEA
ENSG00000170345	FOS	FBJ murine osteosarcoma viral oncogene homolog	IEA
ENSG00000152795	HNRNPDL	heterogeneous nuclear ribonucleoprotein D-like	IEA
ENSG00000171223	JUNB	jun B proto-oncogene	IEA
ENSG00000130522	JUND	jun D proto-oncogene	IEA
ENSG00000177606	JUN	jun proto-oncogene	IEA
ENSG00000107438	PDLIM1	PDZ and LIM domain 1	IEA
ENSG00000177469	PTRF	polymerase I and transcript release factor	IEA
ENSG00000149273	RPS3	ribosomal protein S3	IEA
ENSG00000179218	CALR	calreticulin	TAS
ENSG00000169714	CNBP	CCHC-type zinc finger, nucleic acid binding protein	TAS
ENSG00000107223	EDF1	endothelial differentiation-related factor 1	TAS
ENSG00000089009	RPL6	ribosomal protein L6	TAS

ENSG00000134107	BHLHE40	basic helix-loop-helix family, member e40	NAS
ENSG00000167244	IGF2	insulin-like growth factor 2 (somatomedin A)	NAS

Mouse

ensembl gene id	mgi symbol	description	GO linkage type
ENSMUSG00000042406	Atf4	activating transcription factor 4	IDA, IMP
ENSMUSG00000006932	Ctnnb1	catenin (cadherin associated protein), beta 1	IDA
ENSMUSG00000020267	Hint1	histidine triad nucleotide binding protein 1	IEA, ISS, ISO
ENSMUSG00000028639	Ybx1	Y box protein 1	IEA, ISS, ISO
ENSMUSG00000021660	Btf3	basic transcription factor 3	IEA
ENSMUSG00000051223	Bzw1	basic leucine zipper and W2 domains 1	IEA
ENSMUSG00000050966	Lin28a	lin-28 homolog A (C. elegans)	IEA
ENSMUSG00000055839	Tceb2	transcription elongation factor B (SIII), polypeptide 2	IEA

Table of the genes annotated with the GO term “regulation of transcription, DNA-templated" (GO:0006355) in the most abundant human and mouse yolk sac transcripts. GO evidence codes (go_linkage_type) are also provided (see http://geneontology.org/page/guide-go-evidence-codes)

Table S4

BP	BP description	MF	MF description	CC	CC description
GO:0006364	rRNA processing	GO:0003676	nucleic acid binding	GO:0000323	lytic vacuole
GO:0006412	translation	GO:0003723	RNA binding	GO:0005575	cellular_component
GO:0006518	peptide metabolic process	GO:0003735	structural constituent of ribosome	GO:0005576	extracellular region
GO:0006807	nitrogen compound metabolic process	GO:0005198	structural molecule activity	GO:0005615	extracellular space
GO:0008152	metabolic process	GO:0005488	binding	GO:0005622	intracellular
GO:0009058	biosynthetic process	GO:0016491	oxidoreductase activity	GO:0005623	cell
GO:0009059	macromolecule biosynthetic process	GO:0019843	rRNA binding	GO:0005730	nucleolus
GO:0009123	nucleoside monophosphate metabolic process	GO:0044822	poly(A) RNA binding	GO:0005737	cytoplasm
GO:0009161	ribonucleoside monophosphate metabolic process	GO:0097159	organic cyclic compound binding	GO:0005739	mitochondrion
GO:0009987	cellular process	GO:1901363	heterocyclic compound binding	GO:0005740	mitochondrial envelope
GO:0010467	gene expression			GO:0005746	mitochondrial respiratory chain
GO:0016072	rRNA metabolic process			GO:0005764	lysosome
GO:0019538	protein metabolic process			GO:0005773	vacuole
GO:0022613	ribonucleoprotein complex biogenesis			GO:0005829	cytosol
GO:0034641	cellular nitrogen compound metabolic process			GO:0005840	ribosome
GO:0034645	cellular macromolecule biosynthetic process			GO:0005912	adherens junction

GO:0042254	ribosome biogenesis
GO:0042255	ribosome assembly
GO:0043043	peptide biosynthetic process
GO:0043603	cellular amide metabolic process
GO:0043604	amide biosynthetic process
GO:0044085	cellular component biogenesis
GO:0044237	cellular metabolic process
GO:0044238	primary metabolic process
GO:0044249	cellular biosynthetic process
GO:0044267	cellular protein metabolic process
GO:0044271	cellular nitrogen compound biosynthetic process
GO:0044281	small molecule metabolic process
GO:0044710	single-organism metabolic process
GO:0055114	oxidation-reduction process
GO:0065003	macromolecular complex assembly
GO:0071704	organic substance metabolic process
GO:1901564	organonitrogen compound metabolic process
GO:0005924	cell-substrate adherens junction
GO:0005925	focal adhesion
GO:0015934	large ribosomal subunit
GO:0015935	small ribosomal subunit
GO:0016020	membrane
GO:0022625	cytosolic large ribosomal subunit
GO:0022626	cytosolic ribosome
GO:0022627	cytosolic small ribosomal subunit
GO:0030054	cell junction
GO:0030055	cell-substrate junction
GO:0030529	intracellular ribonucleoprotein complex
GO:0031090	organelle membrane
GO:0031966	mitochondrial membrane
GO:0031982	vesicle
GO:0031988	membrane-bounded vesicle
GO:0032991	macromolecular complex
GO:0043226	organelle

	organonitrogen compound biosynthetic		
GO:1901566	process	GO:0043227	membrane-bounded organelle
	organic substance		
GO:1901576	biosynthetic process	GO:0043228	non-membrane-bounded organelle
		GO:0043229	intracellular organelle
		GO:0043230	extracellular organelle
		GO:0043231	intracellular membrane-bounded organelle
		GO:0043232	intracellular non-membrane-bounded organelle
		GO:0044391	ribosomal subunit
		GO:0044421	extracellular region part
		GO:0044422	organelle part
		GO:0044424	intracellular part
		GO:0044444	cytoplasmic part
		GO:0044445	cytosolic part
		GO:0044446	intracellular organelle part
		GO:0044455	mitochondrial membrane part
		GO:0044464	cell part
		GO:0070062	extracellular exosome
		GO:0070161	anchoring junction
		GO:0070469	respiratory chain
		GO:0098798	mitochondrial protein complex
		GO:0098800	inner mitochondrial membrane protein complex
		GO:1903561	extracellular vesicle

Over represented GO terms shared among the 400 most abundant transcripts in the human, mouse and chicken yolk sacs. BP, Biological process; MF molecular function and CC, Cellular component

References

1. Martin M (2016) Cutadapt removes adapter sequences from high-throughput sequencing reads. 1–3 doi: 10.14806/ej.17.1.200.

2. Kim D, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36.

3. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169.

4. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550.

5. Fagerberg L, et al. (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 13(2):397–406.

6. Mi H, Muruganujan A, Thomas PD (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* 41(Database issue):D377–86.

7. Walter W, Sánchez-Cabo F, Ricote M (2015) GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* 31(17):2912–2914.

8. Wang M, Zhao Y, Zhang B (2015) Efficient Test and Visualization of Multi-Set Intersections. *Sci Rep* 5:16923.

9. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* 57(1):289–300.

10. Pereda J, et al. (1994) The structure of the human yolk sac: a scanning and transmission electron microscopic analysis. *Arch Histol Cytol* 57(2):107–117.

11. Melé M, et al. (2015) Human genomics. The human transcriptome across tissues and individuals. *Science* 348(6235):660–665.

12. Jauniaux E, Jurkovic D, Henriet Y, Rodesch F, Hustin J (1991) Development of the secondary human yolk sac: correlation of sonographic and anatomical features. *Hum Reprod* 6(8):1160–1166.