Original Article

Cytogenet Genome Res 111:79–87 (2005)
DOI: 10.1159/000085674

Cytogenetic and Genome Research

ESTs from brain and testis of White Leghorn and red junglefowl: annotation, bioinformatic classification of unknown transcripts and analysis of expression levels

P. Savolainen,[a] C. Fitzsimmons,[b] L. Arvestad,[c] L. Andersson,[b,d] J. Lundeberg[a]

[a] Department of Biotechnology, Royal Institute of Technology, Stockholm;
[b] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala;
[c] Stockholm Bioinformatics Center and Department of Numerical Analysis and Computing Science, Albanova, KTH, Stockholm;
[d] Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala (Sweden)

Manuscript received 26 April 2004; accepted in revised form for publication by T. Shows 30 November 2004.

Abstract. We report the generation, assembly and annotation of expressed sequence tags (ESTs) from four chicken cDNA libraries, constructed from brain and testis tissue dissected from red junglefowl and White Leghorn. 21,285 5′-end ESTs were generated and assembled into 2,813 contigs and 9,737 singletons, giving 12,549 tentative unique transcripts. The transcripts were annotated using BLAST by matching to known chicken genes or to putative homologues in other species using the major gene/protein databases. The results for these similarity searches are available on www.sbc.su.se/~arve/chicken. 4,129 (32.9%) of the transcripts remained without a significant match to gene/protein databases, a proportion of unmatched transcripts similar to earlier non-mammalian EST studies. To estimate how many of these transcripts may represent novel genes, they were studied for the presence of coding sequence. It was shown that most of the unique chicken transcripts do not contain coding parts of genes, but it was estimated that at least 400 of the transcripts contain coding sequence, indicating that 3.2% of avian genes belong to previously unknown gene families. Further BLAST search against dbEST left 1,649 (13.1%) of the transcripts unmatched to any library. The number of completely unmatched transcripts containing coding sequence was estimated at 180, giving a measure of the number of putative novel chicken genes identified in this study. 84.3% of the identified transcripts were found only in testis tissue, which has been poorly studied in earlier chicken EST studies. Large differences in expression levels were found between the brain and testis libraries for a large number of transcripts, and among the 525 most frequently represented transcripts, there were at least 20 transcripts with significant difference in expression levels between red junglefowl and White Leghorn.

Chicken is an important source of animal protein throughout the world, and is also used as a vertebrate model for biomedical research (Brown et al., 2003). In a comprehensive research program aimed at mapping genes controlling a number of different traits we compare red junglefowl (RJ) and White Leghorn (WL) in order to identify genes involved in the dramatic changes in phenotypic traits, such as behaviour, size, growth, appetite, feed conversion efficiency and reproduction, that has developed in domestic fowl since the time of domestication of its wild ancestor, the RJ (Lindqvist et al., 2002; Schütz et al., 2002, 2004; Carlborg et al., 2003; Kerje et al., 2003; Keeling et al., 2004).

Expressed sequence tag (EST) analysis (Adams et al., 1991) has been extensively used for the characterisation of clones in cDNA libraries and for gene discovery and expression analysis. Large datasets of ESTs are a prerequisite for annotation in

Supported by grants from Wallenberg Consortium North, the Foundation for Strategic Research and the AgriFunGen program at the Swedish University of Agricultural Sciences.

Request reprints from Peter Savolainen, Department of Biotechnology
Royal Institute of Technology (KTH)
Albanova, SE–10691 Stockholm (Sweden)
telephone: +8 5537 8335, fax: +8 5537 8481, e-mail: savo@biotech.kth.se

1424–8581/05/1111–0079$22.00/0

eukaryotic genome projects. A number of chicken EST sequencing projects have recently been performed, generating close to 500,000 ESTs (see e.g. www.chick.umist.ac.uk and www.chickest.udel.edu). These projects were mainly performed using domestic chicken lines, and although a large number of tissues have been screened, testis has been only marginally explored.

In this study, four chicken cDNA libraries, constructed from brain and testis tissue from WL and RJ, were studied by EST-analysis in order to identify clones to be used for the production of cDNA-arrays which will be used to identify genes that are differentially expressed in the abovementioned RJ/WL pair. Brain tissue was sampled to identify genes involved in neurological control over traits such as behaviour, appetite and growth, and testis was sampled with the aim of finding novel chicken genes not detected in earlier EST studies. Both RJ and WL were sampled to ensure that genes that are differently expressed in these populations, and which are possibly involved in changes in behaviour and size that have developed in the domestic chicken since the time of domestication, were identified. In this article we report about the generation of ESTs from these four libraries and about the sequence assembly and annotation of putative transcripts. Furthermore, differences in expression levels between libraries are described and the large proportion of transcripts without significant Blast match to known genes found in the study was analyzed for the presence of coding sequence in order to evaluate whether they may represent novel genes.

Materials and methods

Tissue samples

Four chicken cDNA libraries were constructed, from brain and testis tissue from White Leghorn (WL) and red junglefowl (RJ), respectively. Both chicken populations are maintained at the Swedish University of Agricultural Sciences, and are described in detail in Lindqvist et al. (2002). For the two brain libraries, total brain was excised from one female WL and one female RJ, respectively, aged 24–28 weeks. For the two testis libraries, the left testis was excised from five WL and five RJ 40-week-old roosters, respectively. All tissues were flash frozen in liquid nitrogen and stored at – 80 °C until RNA isolation was performed.

cDNA library construction

Total RNA was isolated from brain and cleaned using the TRIzol RNA isolation/purification procedures (Invitrogen) in two rounds, and mRNA was selected for twice with the Qiagen Oligotex mRNA Mini kit (Qiagen) using the batch protocol. The testis total RNA was isolated separately from each of the five individuals using the TRIzol RNA isolation procedure, then combined in equal amounts from each bird, after which mRNA was selected for twice with the Qiagen Oligotex mRNA Mini kit using the spin protocol. Directional cDNA libraries were constructed using the SuperScript Plasmid Cloning System (Invitrogen); the cDNA fragments were cloned into the pSPORT-1 vector and transformed into ElectroMAX DH10b competent cells (Invitrogen), following the SuperScript Plasmid Cloning System manual's instructions.

Plasmid preparation and DNA sequence analysis

The bacteria were plated and colonies were picked robotically using a BioPick (BioRobotics) into 96-well plates, incubated overnight in 50 µl LB medium under ampicillin selection, after which glycerol stocks were prepared. From these stocks, 1 ml LB cultures were inoculated and incubated overnight, and plasmids were prepared according to a high-throughput 96-well microwave boiling protocol (Marra et al., 1999). Sequencing of the plasmids into the 5′ end of the cDNA-inserts was performed using the M13 reverse primer and BigDye Terminator chemistry on an ABI 3700 instrument (Applied Biosystems).

Sequence data analysis

Base calling of the ABI electropherograms was performed using TraceTuner (Paracel, www.paracel.com). Sequence quality control, clustering and assembly of ESTs, were performed using Paracel Transcript Assembler, (PTA, Paracel). Low-quality end sequence (defined by iterating a 30-bp window inwards from start and end of the sequence until reaching a mean QV>13) and vector sequence were removed, *E.coli* and RNA contamination, and mitochondrial DNA sequence were filtered, and low complexity sequence and avian repeats were masked. ESTs with <100 bp sequence were discarded. PTA has been optimised for clustering and assembly of EST data, and the default settings were used here.

Sequence similarity searches

Similarity searches were carried out using NCBI Blast (Altschul et al., 1997). Searches in nr and nt from NCBI (downloaded May 2003), and Swiss Prot (release 23.9), TrEMBL (release 41.6), and the *Gallus gallus* section of Unigene (build 5, Apr 19, 2003) were performed for annotation purposes. Further comparisons with dbEST, NCBI's EST database (April 2003 release) and the BBSRC chicken EST database (Boardman et al., 2002) were performed to reveal overlap with prior sequencing projects. A similarity was considered significant if the E value was lower than 10^{-5}. Furthermore, the sequences were scanned for protein domains using hmmsearch from the HMMER package (Eddy S., http://hmmer.wustl.edu) against Pfam-A domain family models (version 11.0, Bateman et al., 2002). To take into consideration the growth of dbEST since the start of our analyses, transcripts without match to the download of April 2003 were compared also with a download of August 2004.

ORF analysis

ESTScan (Iseli et al., 1999) was utilized for determining whether the transcripts contained coding sequence. To get model parameters for chicken, 18,024 annotated mRNA sequences were downloaded from RefSeq (Sept 15, 2004) and used as input to the accompanying program build_model. To validate the model, chromosomes 1 through 9 from Washington University's Gallus genome assembly release C1 were downloaded from Ensembl and from each chromosome 1,000 artificial ESTs comprising 500 bp were chosen uniformly at random. With a gene density of less than 2%, we expect an average of 20 genes from each chromosome found by chance and, ideally, ESTScan would report the same number. We ran ESTScan with default parameters, except for requiring analysis of positive strand only, and found 32% reported positives. As this result put the model parameters in question, the same test was conducted on four human chromosomes from NCBI release 34, but with ESTScan's distributed parameters for human. Here, 24% of artificial ESTs were reported as positives by ESTScan. Thus, parameters might not be at fault. To determine whether the built in support for automatic error correction could be at fault, the penalty for correcting a frameshift was set at 150 instead of the default 50, reducing the positives to 18% in human and 22% in chicken (indicating 20% false positives conservatively assuming a gene density of 2%). While not ideal, it is a significant improvement and this setting was chosen for our analyses. An additional cross-check, using parameters for human on chicken data resulted in 26% positives on the artificial random ESTs.

Analysis of the maximum open reading frame (ORF) length of transcripts, and of whether the ORF is longer than expected from chance, was conducted using Estzmate, a program which was developed for this project. ORFs were defined as sequences running from the 5′end until interrupted by a stop codon, starting either from the first start codon, according to the "first-AUG rule" (Kozak, 2002), or without a start codon directly from the first base assuming that the start codon was upstream of the recorded sequence. Masked regions were treated as stop codons, i.e. they could not be part of an ORF. All three forward reading frames were considered, and the length of the longest ORF was recorded. To be able to establish the significance of ORF lengths, a random-transcript model was developed as follows. For each transcript, a random transcript was generated by permuting the order of the bases, with the exception for regions that had been masked for low complexity. In this way, effects from sequence lengths, base contents, and length limitations inherent from repeat regions were taken into account, thus maintaining the

Table 1. Results of EST generation, contaminant removal and sequence assembly for White Leghorn (WL) and red junglefowl (RJ) brain and testis libraries

Library	Attempted reads	Successful reads	Removed sequences: mtDNA; RNA; *E.coli*	Sequences to assembly	Contigs represented	Singlets	Transcripts
RJ brain	10,464	6,809	1,031; 69; 1	5,708	1,357	2,639	3,996
WL brain	7,872	6,132	855; 77; 0	5,200	1,358	2,247	3,605
RJ testis	8,352	5,667	146; 20; 2	5,499	1,384	2,554	3,938
WL testis	7,104	5,012	120; 13; 1	4,878	1,308	2,296	3,604
All libraries	33,600	23,620	2,152; 172; 4	21,285	2,813	9,737	12,549

character of the input data. ORF lengths in transcripts covering coding regions are expected to deviate substantially from average ORF lengths of random transcripts. By generating a large number (in our case 250) of random transcripts and recording their ORF lengths, mean (μ) and standard deviation (σ) can be calculated and used to compute a Z score, $Z = (L - \mu)/\sigma$, for the ORF length L of the original transcript. This Z score is a measure of the significance and, since it is approximately an N(0,1) distribution, it can be associated with a probability of observing ORF length L by chance in a non-coding transcript. To validate the method it was run on the artificial EST dataset constructed from the Gallus genome (see above). This yielded 16% positives (indicating 14% false positives conservatively assuming a gene density of 2%) on the 95% level of confidence.

When both programs were used on the artificial EST dataset 6% positives were reported indicating a low level of false positives (4%) for the combined analyses. A combined analysis was therefore used to identify a list of transcripts with a high probability of representing coding sequence.

The estimates from ESTScan and Estzmate can be adjusted by taking the rate of false negatives, α, and the rate of false positives, β, into account. Let P be the number of transcripts a method identifies as coding or containing a long ORF, and let TP be the number of coding transcripts correctly identified by a method out of the C actual coding transcripts available. Then $\alpha = 1 - TP/C$. To estimate α for a method and a given dataset, assume that those transcripts that have significant matches to gene/protein databases are all coding transcripts. In this case, TP equals P and C is simply the number of transcripts. Estimates of β are derived from the tests on random genome data above, and is 0.2 for ESTScan, 0.14 for Estzmate, and 0.04 for the combined analysis. Now consider the number of false positives which, if there are M transcripts, can be written as $FP = \beta (M - C)$. An equation for false negatives, FN, is derived as $FN = \alpha C$, and thus the number of true positives is $TP = C - FN = (1 - \alpha) C$. The number of transcripts a method identifies as coding is $P = FP + TP$, and this offers an equation:

$$P = \beta (M - C) + (1 - \alpha) C$$

from which C can be solved:

$$C = (P - \beta M) / (1 - \alpha - \beta).$$

Analysis of expression levels

The significance of the difference in numbers of ESTs from each tissue type or bird strain in a transcript was calculated using the binomial distribution. For example, out of the 21,285 ESTs generated in this project, 11,207 (52.7%) were from RJ and 10,078 (47.3%) from WL. Assuming no difference in expression levels between the strains, we expect 52.7% of the ESTs assembled into a given transcript to be from RJ and 47.3% to be from WL. The difference from this distribution was statistically evaluated using the binomial distribution, calculating the probability of observing k ESTs from RJ in a transcript represented by totally n ESTs.

Results and discussion

Generation of EST sequences and sequence assembly

Four chicken cDNA libraries were constructed from brain and testis tissue dissected from White Leghorn (WL) and red jungle fowl (RJ). The number of transformants was estimated at 1.5×10^6, 1.8×10^5, 2.8×10^5, and 9.9×10^5, and the average insert size was estimated, at 1,183 ± 539, 1,212 ± 569, 1,300 ± 721, and 1,416 ± 670 bp for the RJ brain, WL brain, RJ testis, and WL testis libraries, respectively. Analysis was attempted for 33,660 clones, which resulted in 23,620 (70.3%) successful sequence reads (reads containing high-quality insert-sequence of at least 100 bp) with an average read length of 643 bp. After filtering of mitochondrial DNA and contaminating sequence, 21,285 EST sequences remained to be used in sequence assembly (Table 1). The proportion of mitochondrial gene transcripts was considerably higher for the brain libraries than for the testis libraries, probably reflecting the higher energy demand in brain resulting in higher mitochondrial activity. High expression levels of mitochondrial genes have been reported in other EST studies of brain cDNA libraries (Ju et al., 2000).

The 21,285 ESTs were clustered and assembled using Paracell Transcript Assembler (PTA). Clustering resulted in 2,882 clusters and 8,349 singlets, and the 2,882 clusters were further assembled into 2,813 contigs and 1,388 cluster singlets. Thus, clustering and assembly resulted in 2,813 contigs and 9,737 singlets, giving a total of 12,549 tentative unique transcripts (Table 1). The ESTs were distributed to contigs and singlets at similar proportions from the four libraries, giving similar numbers of transcripts from each library. DNA sequences for the ESTs and transcripts are available on WWW (www.sbc.su.se/~arve/chicken), and the 21,285 EST sequences have been deposited in GenBank (accession numbers CN216802–CN238086).

Gene annotation

The transcripts were compared to nr and nt from NCBI, SwissProt, TrEMBL, and Unigene in order to annotate the putative transcripts by matching to known chicken genes or to putative homologues in other species, as well as to assess the proportion of transcripts with no match to genes of other species, which could represent novel gene families. The transcripts were also compared to dbEST which at the time of analyses included 422,370 chicken ESTs, in order to identify the number of transcripts without a significant match to any sequence, thereby estimating the number of putative novel chicken genes. The numbers of matches to each database are shown in Table 2. The searches against the gene/protein databases resulted in large proportions of unmatched transcripts and matches with high E values compared to the search against dbEST. This reflects the fact that avian genes are still not well represented in the gene/protein databases, while the large chicken EST pro-

Table 2. Number of the 12,549 chicken transcripts matched to the searched databases at different E values

Databank	Total number of matches	$E < 10^{-50}$	$E < 10^{-20}$	$E < 10^{-5}$	No significant match
NCBI nt + nr	7,775	4,367	1,981	1,427	4,774
Swissprot + Trembl	6,853	3,472	2,056	1,325	5,696
Unigene	3,589	2,839	346	404	8,960
dbEST	10,517	9,430	617	470	2,032
All	10,900	9,659	738	503	1,649

Table 3. The 50 most frequently represented transcripts, showing number of ESTs from each library and annotation

Transcript id	Brain		Testis		Total	Annotation	E value	Match (E value) to chicken EST
	RJ	WL	RJ	WL				
VeFi2.166.C7	13	19	83	65	180	*G. gallus* c-beta-3 beta-tubulin	0.0	0.0
VeFi2.84.C1	85	57	1	2	145	no match	n	0.0
VeFi2.9.C2	47	32	8	11	98	*G. gallus* ubiquitin I (UbI)	0.0	0.0
VeFi2.666.C3	3	0	36	38	77	*M. musculus* H3 histone, family 3A	2.00e-122	0.0
VeFi2.48.C1	33	27	1	1	62	*G. gallus* cystatin	0.0	0.0
VeFi2.63.C1	23	22	6	10	61	*G. gallus* glyceraldehyde-3-phosphate dehydrogenase	0.0	0.0
VeFi2.7.C1	31	23	0	0	54	*G. gallus* myelin basic protein	0.0	0.0
VeFi2.64.C3	25	26	1	1	53	*G. gallus* alpha-tubulin	0.0	0.0
VeFi2.208.C1	15	6	17	13	51	*G. gallus* enolase alpha	0.0	0.0
VeFi2.9.C4	9	6	18	17	50	*G.gallus* polyubiquitin gene Ub II	0.0	0.0
VeFi2.40.C2	15	13	7	14	49	*G.gallus* stathmin	0.0	0.0
VeFi2.395.C1	25	22	0	0	47	*G.gallus* transthyretin	0.0	0.0
VeFi2.206.C1	13	10	13	11	47	*G. gallus* calmodulin	0.0	0.0
VeFi2.364.C2	7	9	14	15	45	*G. gallus* lactate dehydrogenase B	0.0	0.0
VeFi2.77.C1	18	14	5	7	44	*G. gallus* 90kDa heat shock protein	0.0	0.0
VeFi2.229.C1	20	21	1	1	43	*G. gallus* apolipoprotein AI (Apo-AI)	0.0	0.0
VeFi2.18.C2	23	19	0	0	42	*G. gallus* aldolase A	0.0	0.0
VeFi2.400.C1	16	20	3	2	41	*G. gallus* ferritin H chain protein	0.0	0.0
VeFi2.669.C2	1	0	21	18	40	*M. musculus* HN1	6.9e-42	0.0
VeFi2.2019.C1	0	0	24	16	40	*M. musculus* Protein CGI-38 homolog	1.3e-63	7.6e-08
VeFi2.1892.C1	0	1	18	19	38	*M. musculus* alpha-tubulin 3/7	0.0	0.0
VeFi2.102.C1	11	20	3	2	36	*Canis familiaris* cyclophilin A	4.4e-96	0.0
VeFi2.1896.C1	0	0	27	8	35	no match	n	n
VeFi2.396.C1	11	5	13	5	34	*G. gallus* ribosomal protein L7a	0.0	0.0
VeFi2.1873.C1	0	0	16	18	34	no match	n	n
VeFi2.15.C1	17	16	1	0	34	no match	n	0.0
VeFi2.143.F31C1	13	7	9	5	34	*Eimeria tenella* ribosomal protein S3a	0.0	0.0
VeFi2.76.C1	11	20	1	1	33	*Anas platyrhynchos* calmodulin	0.0	0.0
VeFi2.130.C1	18	10	1	4	33	*Bos taurus* Phosphatidylethanolamine-binding protein	3.0e-91	0.0
VeFi2.2048.C1	0	1	12	18	31	*G. gallus* testis-specific alpha-tubulin	0.0	5.00e-125
VeFi2.51.C1	15	14	0	0	29	*Serinus canaria* canarigranin (HAT14)	3.8e-62	2.7e-72
VeFi2.368.C1	13	12	3	1	29	*G. gallus* ornithine decarboxylase antizyme	0.0	0.0
VeFi2.241.C1	7	9	10	3	29	*G. gallus* elongation factor 1 alpha	0.0	0.0
VeFi2.165.C1	17	10	1	1	29	*G. gallus* clusterin	0.0	0.0
VeFi2.249.C1	11	8	4	5	28	*G. gallus* Jun-binding protein	0.0	0.0
VeFi2.153.C1	6	17	4	1	28	*G. gallus* heat shock cognate 70	0.0	0.0
VeFi2.1244.C1	1	4	13	10	28	*G.gallus* ribosomal protein S6	0.0	0.0
VeFi2.535.C1	8	9	7	3	27	*R. norvegicus* ribosomal protein S8	4.00e-127	0.0
VeFi2.585.C1	5	6	9	5	25	*G. gallus* domesticus ribosomal protein S4	0.0	0.0
VeFi2.540.C1	10	9	5	1	25	*H. sapiens* thymosin beta 4	9.5e-94	0.0
VeFi2.22.C1	19	6	0	0	25	*G. gallus* chS-Rex-s	0.0	0.0
VeFi2.2012.C1	0	0	14	11	25	*H. sapiens* Outer dense fiber protein (ODFP)	4.7e-17	n
VeFi2.115.C1	9	10	1	5	25	*H. sapiens* 16.7Kd protein	1.1e-19	0.0
VeFi2.101.C1	13	9	1	2	25	*Equus caballus* Ubiquitin carboxyl-terminal hydrolase isozyme L1	2.0e-92	0.0
VeFi2.66.C3	0	24	0	0	24	Myeloblastosis-associated virus, env and pol genes	0.0	0.0
VeFi2.676.C1	4	5	7	6	22	*Anas platyrhynchos* Acyl-CoA-binding protein (ACBP)	1.6e-40	0.0
VeFi2.464.C1	7	4	1	10	22	*H. sapiens* malate dehydrogenase 1	2.00e-161	0.0
VeFi2.211.C1	7	9	2	4	22	*M. musculus* ATPase	9.0e-67	0.0
VeFi2.1946.C1	0	1	13	8	22	*R. norvegicus* dihydropyrimidinase	0.0	0.0
VeFi2.1909.C1	0	0	15	7	22	*H. sapiens* DnaJ homolog, subfamily B, member 8	2.5e-47	n

Table 4. Number of White Leghorn (WL) and red junglefowl (RJ) transcripts without significant BLAST match

Library	Transcripts	No match to gene/ protein databases	No match to any database
RJ brain	3,996	1,100 (27.5 %)	286 (7.2 %)
WL brain	3,605	917 (25.4 %)	235 (6.5 %)
RJ testis	3,938	1,243 (31.6 %)	638 (16.2 %)
WL testis	3,604	1,121 (31.1 %)	591 (16.4 %)
All libraries	12,549	4,129 (32.9 %)	1,649 (13.1 %)

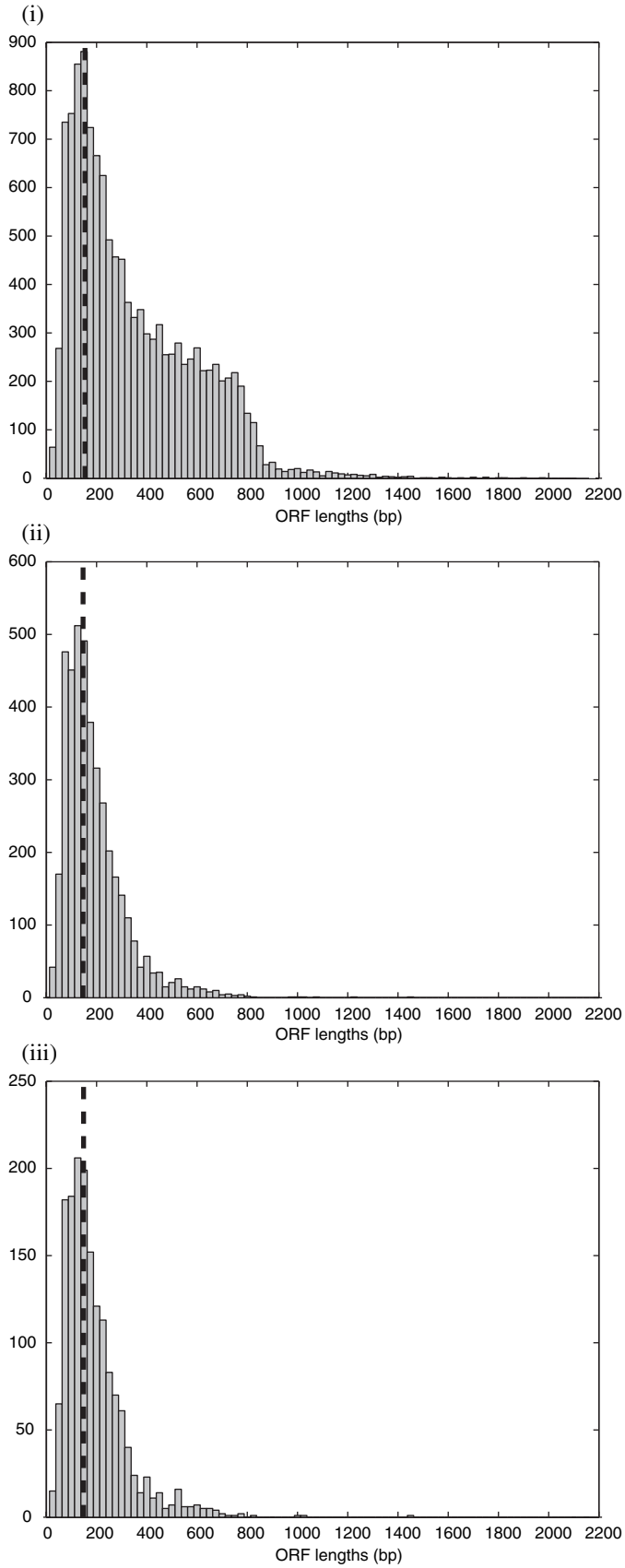
jects have generated transcripts representing a large proportion of the avian genes. The results of the similarity searches, showing the best match to each database, are available for each transcript on www.sbc.su.se/~arve/chicken together with domain annotation from Pfam. Table 3 describes the annotation of the 50 most frequently represented transcripts, which constitute 10.0% of all ESTs. Twenty-six of these transcripts could be identified by matches to earlier known *Gallus* genes, and 18 tentatively identified by matches to genes in other vertebrates, 15 in mammals and three in birds. Two matches were to genes from pathogens, a ribosomal protein of the intestinal bird parasite *Eimeria tenella*, and the *env* and *pol* genes of Myeloblastosis-associated virus (MAV) type 1/2. Four transcripts had no significant matches to any of the gene/protein databases. Not surprisingly, a number of these highly expressed transcripts were annotated as either genes specific for, or highly expressed in, brain (e.g. myelin, transthyretin and canarigranin) or testis (e.g. alpha-tubulin 3/7 and ODFP), or as housekeeping genes (e.g. glyceraldehyde-3-phosphate dehydrogenase, lactate dehydrogenase B, and a number of ribosomal proteins).

Discovery of novel genes

A striking result of the similarity searches was that a large proportion of the transcripts had no significant match to known genes. In total, 4,129 (32.9%) of the transcripts had no significant match in gene/protein databanks (Tables 2 and 4). This is in accordance with earlier chicken EST studies which had 30% or more transcripts with no match (Abdrakhmanov et al., 2000; Boardman et al., 2002), and similar proportions of unique transcripts have been reported for EST studies of fish (35–50%; Ju et al., 2000; Zeng and Gong, 2002) and frog (46%; Blackshear et al., 2001). This could indicate that a fairly large proportion of avian genes represent fast-evolving genes, previously unknown gene families or new branches of families, the latter two of which could have been lost in the mammalian lineage or evolved since the split between birds and mammals ~310 million years ago (Hedges, 2002). Since the nature of the unique transcripts has been poorly studied in earlier EST studies we performed some analyses to assess to what extent they may represent novel genes (see below). Furthermore, even after a similarity search against dbEST, there were 1,649 transcripts (13.1% of all transcripts, 96 contigs and 1,553 singlets) without a significant match in any of the databases, which could indicate that a large number of novel genes, not found in the earlier large scale chicken EST studies, were discovered in this study (Table 4). dbEST contained at the time of analysis 422,370 ESTs from chicken, which were generated from a number of different tissues including brain, but testis tissue was not represented except in a few multi-tissue libraries. Comparing the four libraries, the two testis libraries had a larger proportion of unmatched transcripts than the brain libraries; 16.3% of the testis transcripts had no match in any database compared to 6.9% for the brain libraries (Table 4), and 69.7% of these unmatched transcripts were unique to the testis libraries. Thus, a majority of the putative new chicken genes were found in the testis libraries, reflecting the absence of testis tissue in earlier large-scale chicken EST studies. Notably, the difference between the brain and testis libraries in the number of matches is smaller for the gene/protein databases, reflecting that a number of the novel chicken genes found in the testis libraries have matches to genes in other species for which testis-specific genes have been better explored. The difference in expression levels between brain and testis, and the discovery of novel chicken genes in the testis library, is further exemplified in Table 3, which describes the annotation and the number of ESTs from each of the four libraries for the 50 most frequently represented transcripts. Out of five transcripts that were represented only in the testis libraries, two had matches to genes of other species but not to chicken genes or ESTs, and two did not have a match in any database. Thus, four out of the 14 most highly expressed transcripts in the testis samples were not found in earlier chicken EST studies.

However, the large proportion of transcripts without significant match to genes of other organisms may to a large extent consist of UTRs, non-coding RNA genes (Numata et al., 2003) and possibly contaminating nuclear DNA and sequencing artefacts. Avian 5′ and 3′ UTRs are similar in average length (126.4 and 651.9 bp, respectively) as well as maximum length (620 bp; 3,990 bp) to other vertebrates (Pesole et al., 2001), and should generally have little homology to those of mammals. Likewise, some of the completely unique transcripts, without match to any sequence including chicken ESTs, could originate from contaminating nuclear DNA and sequencing artefacts rather than from mRNA from chicken genes. Therefore, in order to investigate to what extent the unmatched transcripts may represent novel genes, they were examined for signs of coding sequence. Three data sets were studied: (i) all transcripts, (ii) the transcripts without match to the gene/protein databases, (iii) the transcripts without match to any database including dbEST.

The mean ORF length was computed for the data sets, and compared to what is expected for random data. Thus, the maximum ORF length was searched for each transcript in the three forward reading frames giving the mean ORF length for the data sets, to compare with the length expected for random data, which was computed by permuting the input sequence at random for each transcript and searching for ORFs. We expect data set (i) to have longer ORFs than expected from chance since two thirds of these transcripts have matches to the gene/protein databases, while data sets (ii) and (iii) should have mean ORF lengths approximately equal to that expected from chance if they do not represent coding sequence. However, all three data sets had mean ORF lengths longer than expected from chance; 349, 190 and 192 bp for the three data sets,



respectively, to compare with mean ORF lengths of 159, 146 and 148 bp expected for random data, indicating that coding sequence is present in all three data sets. The distribution of ORF lengths among the transcripts for the three data sets are given in Fig. 1, showing a large number of transcripts with long ORF in data sets (ii) and (iii).

To estimate the number of transcripts containing coding sequence in the three data sets we used two different methods, ESTScan and Estzmate. ESTScan (Iseli et al., 1999) studies a transcript for signs of coding sequence using a hidden Markov model to describe ESTs, with automatic correction of sequence errors, and Estzmate, which was developed for this study, examines whether a transcript contains longer ORFs than can be expected from random data. The three data sets above and a fourth set (iv), the transcripts with match to the gene/protein databases, were tested with the two programs and with a combination of both. ESTScan identified 1,263 transcripts containing coding sequence in data set (ii) and 548 in data-set (iii) (Table 5), indicating that a large number of the unique transcripts represent genes. Estzmate identified about 40% fewer transcripts containing long ORFs in data sets (ii) and (iii). This is not surprising considering that the distributions of random and true ORF lengths are to a large extent overlapping, since transcripts containing only a short stretch of coding sequence may have ORFs shorter than or close in length to the mean length of ORFs in random data. Both programs have a relatively high level both of false positives, 20 and 14% respectively (see Material and methods for calculations), and false negatives. The level of false negatives is indicated by the proportion of transcripts in data set (iv) not identified as coding or containing a long ORF. This is because the transcripts in data set (iv), having matches to the gene/protein databases, should practically all contain coding sequence since they have matched coding parts of genes in the databases. Thus, a level of false negatives of 18 and 37% is indicated for ESTScan and Estzmate, respectively. An adjustment of the number of transcripts containing coding sequence or a long ORF, taking into account the rates of false positives and false negatives, was estimated as described in Materials and methods. With the knowledge that the rates of false positives and false negatives have rough estimates this gives adjusted numbers shown in Table 5. The combined analysis of ESTScan and Estzmate identifies transcripts that show signs of coding sequence and have an ORF longer than expected from chance, giving a conservative measure of the number of unmatched transcripts potentially representing genes (Table 5).

Fig. 1. Histograms over ORF lengths in datasets: (i) all transcripts, (ii) transcripts without significant matches in gene/protein databases, and (iii) transcripts without significant match in any database including dbEST. Dashed lines correspond to mean ORF lengths in randomized transcripts taken from the datasets: (i) 159 bp, (ii) 146 bp, and (iii) 148 bp.

Table 5. Number of transcripts containing coding sequence according to ESTScan and longer ORFs than expected from chance according to Estzmate, and number of transcripts identified using both programs, with estimates adjusted for levels of false positives and negatives. Data sets: (i) all transcripts, (ii) the transcripts without match in gene/protein databases, (iii) the transcripts without match to any database including dbEST and (iv) the transcripts with match to the gene/protein databases

Data set	Size	ESTScan		Estzmate		ESTScan + Estzmate	
		Coding transcripts	Adjusted estimate	Transcripts with long ORF	Adjusted estimate	Coding transcripts	Adjusted estimate
(i)	12,549	8,204		6,062		5,443	
(ii)	4,129	1,263	705	755	361	389	400
(iii)	1,649	548	352	294	129	166	180
(iv)	8,420	6,941		5,307		5,054	

To conclude, the ESTScan and Estzmate analyses show that the majority of the transcripts without significant match in gene/protein databanks do not contain coding parts. More probably they mostly contain noncoding parts of chicken gene transcripts, and the fact that they do not match genes of other species does therefore not imply that they represent novel genes. However, we estimate that at least some 400 of the unmatched transcripts (3.2% of all transcripts) contain coding sequence and therefore probably represent novel genes belonging to previously unknown gene families or new branches of families. This suggests that a large number of novel genes will be found when evolutionary branches other than the relatively well-explored mammalian branch become better studied. Furthermore, 13.1% of the transcripts in this study had no significant match to sequences in any database including the large number of chicken ESTs generated in earlier studies. According to the ESTScan and Estzmate analyses at least 180 of these transcripts represent genes. Since transcripts not indicated to contain coding sequence may also be from novel genes if they represent noncoding parts of chicken gene mRNAs, this figure probably represents a very conservative measure of the number of novel chicken genes identified in this study. The reason that a large number of novel chicken genes were identified, even though more than 400,000 chicken ESTs have previously been generated, can be attributed to the absence of testis tissue in earlier studies except for a few multi-tissue libraries. Thus, out of the 1,649 totally unmatched transcripts, 69.7% were unique to the two testis libraries, and out of the 166 transcripts predicted by the combined ESTScan and Estzmate to represent genes, 140 (84.3%) of the transcripts were found only in the testis libraries. Lists of the unmatched transcripts identified as genes by ESTScan, Estzmate and using both methods as summarized in Table 5 are available on www.sbc.su.se/~arve/chicken.

A few more analyses were performed to further validate our results. To take into consideration the fast growth of dbEST (with an increase in the number of chicken ESTs from 422,370 in April 2003 when our analyses were started to 473,762 in August 2004) we performed a new Blast analysis of the 1,649 transcripts without match in any database against a download of August 2004. This resulted in a single new match among the 1,649 transcripts and no new match among the 166 transcripts identified as genes. To validate that the unique transcripts were obtained from chicken RNA and not from contaminating organisms they were Blasted against Washington University's Gallus genome assembly release C1. This gave a match for 1,445 of the 1,649 transcripts (87.6%), authenticating them as present in the genome, and out of the 166 transcripts classified as coding sequence based on both ESTScan and Estzmate, 148 (89.2%) had a match to the chicken genome. To further investigate whether the unique transcripts with long ORFs may represent transcripts of chicken genes, an inspection of the transcripts for correct sequence at splice sites was performed. The ten transcripts, among those unmatched to the gene/protein databases, containing the longest ORFs were aligned to the *Gallus gallus* genome project trace files deposited at GenBank. Nine of these sequences had a match, and seven of these showed correct sequence at the splice sites, while the two other transcripts matched contigous sequence. This suggests that a majority of the transcripts, which do not match any known genes but contain long ORFs, represent chicken genes.

Analysis of expression levels

Very large differences in expression levels were found between brain and testis for a majority of the 50 most frequently represented transcripts (Table 3). Eleven of the 50 transcripts were unique to either brain or testis, and 39 (78%) were represented by significantly different numbers (binomial distribution, $P < 0.05$) of ESTs from the two tissue types. Accordingly, several of these transcripts were annotated for genes specific for or abundant in either brain or testis. More interestingly, a number of transcripts showed a difference in expression between RJ and WL. Four of the 50 most frequently represented transcripts had significantly different numbers (binomial distribution, $P < 0.05$) of ESTs from RJ and WL, and among the 525 transcripts containing five or more ESTs there were 47 transcripts represented by significantly more ESTs from RJ than WL (27 transcripts) or vice versa (20 transcripts) (Table 6). In a comparison of 525 transcripts we expect to observe 26 transcripts that reach the nominal significance threshold by chance only. Thus among the 525 most represented transcripts we find 21 more transcripts than is expected by chance indicating that there may be a large number of genes that are differentially expressed between RJ and WL. Some of these genes are possibly linked to the dramatic changes in traits such as size, behaviour, appetite, feed conversion efficiency, and energy conservation and storage,

Table 6. Transcripts represented by significantly (P. binomial distribution) more ESTs from RJ (shaded) and WL (white)

Transcript id	RJ		WL		Total	P	Annotation	E value	Match (E value) to chicken EST
	Brain	Testis	Brain	Testis					
VeFi2.66.C3	0	0	24	0	24	0.0000	Myeloblastosis-associated virus, env and pol genes	0.0	0.0
VeFi2.740.C2	2	8	0	0	10	0.0016	no match	N	0.0
VeFi2.1896.C1	0	27	0	8	35	0.0026	no match	N	n
VeFi2.579.C1	9	0	0	0	9	0.0031	*G. gallus* mRNA for arginine vasotocin and copeptin	0.0	0.0
VeFi2.89.C1	5	3	0	0	8	0.0059	*G. gallus* homeodomain protein (Tlx-3)	1.00e-179	0.0
VeFi2.1055.C1	1	0	9	0	10	0.0069	*G. gallus* ovotransferrin	0.0	0.0
VeFi2.2008.C1	0	7	0	0	7	0.011	*M. musculus* A930018P22Rik protein	9.5e-11	n
VeFi2.22.C1	19	0	6	0	25	0.015	*G. gallus* chS-Rex-s	0.0	0.0
VeFi2.1020.C1	5	4	1	0	10	0.016	*H. sapiens* O-linked N-acetylglucosamine (GlcNAc) transferase	6.00e-134	0.0
VeFi2.1233.C1	1	2	4	7	14	0.018	*G. gallus* ras-like protein	0.0	0.0
VeFi2.431.C2	3	3	0	0	6	0.021	*H. sapiens* NADH dehydrogenase MLRQ subunit	4.9e-24	0.0
VeFi2.1979.C1	0	6	0	0	6	0.021	*M. musculus* Hypothetical protein	8.4e-09	0.0
VeFi2.201.C2	6	0	0	0	6	0.021	*B. taurus* CD63 antigen	3.6e-48	0.0
VeFi2.948.C1	0	0	4	1	5	0.024	*H. sapiens* Ubiquinol-cytochrome C reductase 11 kDa subunit	2.5e-30	0.0
VeFi2.1073.C1	0	0	3	2	5	0.024	*H. sapiens* proteasome subunit C9	3.00e-128	0.0
VeFi2.1685.C1	0	0	4	1	5	0.024	*H. sapiens* 26S proteasome regulatory subunit S9	3.00e-125	0.0
VeFi2.1150.C1	0	0	2	3	5	0.024	*H. sapiens* Small nuclear ribonucleoprotein Sm D2	2.6e-53	0.0
VeFi2.1175.C1	0	0	2	3	5	0.024	*H. sapiens* Hypothetical protein KIAA0446	6.4e-71	0.0
VeFi2.1222.C1	0	0	5	0	5	0.024	*H. sapiens* Arfaptin 2	3.00e-123	0.0
VeFi2.898.C1	0	0	3	2	5	0.024	*H. sapiens* RAN binding protein 1	7.5e-81	0.0
VeFi2.2547.C1	0	1	0	7	8	0.025	*H. sapiens* Pituitary tumor transforming protein 1	1.2e-33	0.0
VeFi2.110.C1	1	0	7	0	8	0.025	*G. gallus* HT7 antigen	0.0	0.0
VeFi2.396.C1	11	13	5	5	34	0.026	*G. gallus* ribosomal protein L7a	0.0	0.0
VeFi2.9.C6	0	3	2	8	13	0.03	*G. gallus* ubiquitin I (UbI) gene	0.0	0.0
VeFi2.1571.C1	0	2	3	5	10	0.038	no match	N	0.0
VeFi2.498.C1	4	0	6	5	15	0.039	*M. musculus* Phosphoglycerate mutase 1	2.00e-139	0.0
VeFi2.601.C1	5	0	0	0	5	0.04	*H. sapiens* Guanine nucleotide-binding protein G(I)	4.2e-24	0.0
VeFi2.1160.C1	3	2	0	0	5	0.04	*Macaca fascicularis* NADH dehydrogenase flavoprotein 1	8.00e-132	0.0
VeFi2.1981.C1	0	5	0	0	5	0.04	*H. sapiens* Hypothetical protein FLJ36059	4.7e-12	0.0
VeFi2.25.C1	5	0	0	0	5	0.04	*H. sapiens* Similar to Hypothetical protein KIAA0273	0.0	0.0
VeFi2.64.C1	4	1	0	0	5	0.04	*G. gallus* alpha-tubulin	0.0	0.0
VeFi2.295.C1	4	1	0	0	5	0.04	*H. sapiens* cold inducible RNA binding protein	1.00e-114	0.0
VeFi2.1119.C1	2	3	0	0	5	0.04	*S. scrofa* Non-selenium glutathione phospholipid hydroperoxide peroxidase	8.00e-101	0.0
VeFi2.1629.C1	1	4	0	0	5	0.04	*H. sapiens* ARMET protein	4.9e-70	0.0
VeFi2.286.C1	5	0	0	0	5	0.04	*G. gallus* myosin alkali light chain	0.0	0.0
VeFi2.76.C1	11	1	20	1	33	0.044	*Anas platyrhynchos* calmodulin	0.0	0.0
VeFi2.1195.C1	5	11	4	2	22	0.045	*H. sapiens* Bax inhibitor-1	3.0e-83	0.0
VeFi2.240.C1	1	0	5	1	7	0.047	*H. sapiens* heterogeneous nuclear ribonucleoprotein A2/B1	0.0	0.0
VeFi2.260.C1	1	0	3	3	7	0.047	*C. familiaris* Glycoprotein 25L precursor	1.3e-62	0.0
VeFi2.174.C1	1	0	3	3	7	0.047	*H. sapiens* B-cell receptor-associated protein 31	2.8e-53	0.0
VeFi2.2404.C1	0	1	1	5	7	0.047	*H. sapiens* Nucleoside diphosphate kinase homolog 5	8.2e-91	0.0
VeFi2.345.C1	5	2	0	1	8	0.048	*G. gallus* ribosomal protein L5	0.0	0.0
VeFi2.729.C1	5	2	1	0	8	0.048	*G. gallus* Phosphoglycerate kinase	0.0	0.0
VeFi2.106.C1	6	1	0	1	8	0.048	*G. gallus* FK506-binding protein 12	0.0	0.0
VeFi2.385.C2	2	7	2	0	11	0.048	*M. musculus* ribosomal protein S18	3.9e-74	0.0
VeFi2.198.C1	6	3	1	1	11	0.048	*G. gallus* nucleolar phosphoprotein B23	0.0	0.0
VeFi2.269.C3	0	9	0	2	11	0.048	no match	N	1.0e-29

which have developed, through conscious as well as unconscious selection, in domestic fowl since the time of domestication of the RJ. There were no obvious candidates among the transcripts with most significant differences but four out of the top thirteen transcripts were without functional annotation and may be linked to any trait. It can also be observed that several transcripts with higher expression in the WL libraries were annotated as proteins involved in protein degradation (proteasome subunit C9, 26S proteasome regulatory subunit S9, and ubiquitin I, UbI), and that RJ had a higher expression for transcripts associated with brain and nerve cell formation (homeodomain protein Tlx-3 and chS-Rex-s). Furthermore, transcripts annotated for a number of genes involved in the glycolysis and the respiratory chain were both up and down regulated in RJ. The most striking difference in expression levels was observed for the transcript annotated for the env and pol genes of myeloblastosis-associated virus (MAV) type 1/2, for which there were 24 EST sequences from the WL brain library and none for any other library. Envelope genes of this virus have been shown to recombine with the Rous-associated virus 0 (RAV-0) endogenous retrovirus present in the chicken genome and to conform immunity to infection of Avian Leukosis Virus (Lupiani et al., 2000). Different strains of endogenous retroviruses have been found in different copy numbers in the RJ and domestic chicken genomes (Frisby et al., 1979; Resnick et al., 1990), and have also been shown to be differentially expressed in different tis-

sues in uninfected chicken (Chen, 1980). In relation to the appearance of the MAV transcript in WL brain, it is interesting to note the up-regulation of ovotransferrin. This protein is known as an iron transfer and scavenging protein (Jeltsch and Chambon, 1982), but has recently been shown to be induced by Marek's Disease Virus in chicken embryonic fibroblasts (MDV) (Morgan et al., 2001), and to exhibit anti-viral activity (Giansanti et al., 2002). Therefore, its increased expression may be due to the increase in expression of MAV, or both ovotransferrin and the endogenous MAV may have been up-regulated by an external viral trigger in the WL bird since endogenous viral transcripts have also been shown to be up-regulated by MDV (Morgan et al., 2001). This first assessment of differential expression between the domestic White Leghorn chicken and its wild ancestor, the red junglefowl, will now be followed by expression analysis using the cDNA arrays generated with the collection of cDNA clones documented in this study.

References

Abdrakhmanov I, Lodygin D, Geroth P, Arakawa H, Law A, Plachy J, Korn B, Buerstedde JM: A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function. Genome Res 10:2062–2069 (2000).

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

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402 (1997).

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: The Pfam Protein Families Database. Nucleic Acids Res 30:276–280 (2002).

Blackshear PJ, Lai WS, Thorn JM, Kennington EA, Staffa NG, Moore DT, Bouffard GG, Beckstrom-Sternberg SM, Touchman JW, Bonaldo MF, Soares MB: The NIEHS *Xenopus* maternal EST project: interim analysis of the first 13,879 ESTs from unfertilized eggs. Gene 267:71–87 (2001).

Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, Tickle C, Brown WR, Wilson SA, Hubbard SJ: A comprehensive collection of chicken cDNAs. Curr Biol 12:1965–1969 (2002).

Brown WR, Hubbard SJ, Tickle C, Wilson SA: The chicken as a model for large-scale analysis of vertebrate gene function. Nat Rev Genet 4:87–98 (2003).

Carlborg Ö, Kerje S, Schütz K, Jacobsson L, Jensen P, Andersson L: A global search reveals epistatic interaction between QTL for early growth in the chicken. Genome Res 13:413–421 (2003).

Chen JH: Expression of endogenous avian myeloblastosis virus information in different chicken cells. J Virol 36:162–170 (1980).

Frisby DP, Weiss RA, Roussel M, Stehelin D: The distribution of endogenous chicken retrovirus sequences in the DNA of galliform birds does not coincide with avian phylogenic relationships. Cell 3:623–634 (1979).

Giansanti F, Rossi P, Massucci MT, Botti D, Antonini G, Valenti P, Seganti L: Anitiviral activity of ovotransferrin discloses an evolutionary strategy for the defensive activities of lactoferrin. Biochem Cell Biol 80:125–130 (2002).

Hedges SB: The origin and evolution of model organisms. Nat Rev Genet 3:838–849 (2002).

Iseli C, Jongeneel CV, Bucher P: ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol:138–148 (1999).

Jeltsch J, Chambon P: The complete nucleotide sequence of chicken ovotransferrin mRNA. Eur J Biochem 122:291–295 (1982).

Ju Z, Karsi A, Kocabas A, Patterson A, Li P, Cao D, Dunham R, Liu Z: Transcriptome analysis of channel catfish *(Ictalurus punctatus)*: genes and expression profile from the brain. Gene 261:373–382 (2000).

Keeling L, Andersson L, Schutz KE, Kerje S, Fredriksson R, Carlborg O, Cornwallis CK, Pizzari T, Jensen P: Chicken genomics: Feather-pecking and victim pigmentation. Nature 431:645–646 (2004).

Kerje S, Carlborg O, Jacobsson L, Schutz K, Hartmann C, Jensen P, Andersson L: The twofold difference in adult size between the red junglefowl and White Leghorn chickens is largely explained by a limited number of QTLs. Anim Genet 34:264–274 (2003).

Kozak M: Pushing the limits of the scanning mechanism for initiation of translation. Gene 299:1–34 (2002).

Lindqvist C, Schutz K, Jensen P: Red junglefowl have more contrafreloading than white leghorn layers: Effect of food deprivation and consequences for information gain. Behaviour 139:1195–1209 (2002).

Lupiani B, Hunt H, Silva R, Fadly A: Identification and characterization of recombinant subgroup J avian leukosis viruses (ALV) expressing subgroup A ALV envelope. Virology 276:37–43 (2000).

Marra MA, Kucaba TA, Hillier LW, Waterston RH: High-throughput plasmid DNA purification for 3 cents per sample. Nucleic Acids Res 27:e37 (1999).

Morgan RW, Sofer L, Anderson AS, Bernberg EL, Cui J, Burnside J: Induction of host gene expression following infection of chicken embryo fibroblasts with oncogenic Marek's disease virus. J Virol 75:533–539 (2001).

Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming LG, Hume DA, Hayashizaki Y, Tomita M: Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res 13:1301–1306 (2003).

Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S: Structural and functional features of eukaryotic mRNA untranslated regions. Gene 276:73–81 (2001).

Resnick RM, Boyce-Jacino MT, Fu Q, Faras AJ: Phylogenetic distribution of the novel avian endogenous provirus family EAV-0. J Virol 64:4640–4653 (1990).

Schütz K, Kerje S, Carlborg O, Jacobsson L, Andersson L, Jensen P: QTL analysis of a red junglefowl × White Leghorn intercross reveals trade-off in resource allocation between behavior and production traits. Behav Genet 32:423–433 (2002).

Schütz K, Kerje S, Jacobsson L, Forkman B, Carlborg Ö, Andersson L, Jensen P: Major growth QTLs in fowl are related to fearful behavior: possible genetic links between fear responses and production traits in a red junglefowl × White Leghorn intercross. Behav Genet 34:121–130 (2004).

Zeng S, Gong Z: Expressed sequence tag analysis of expression profiles of zebrafish testis and ovary. Gene 294:45–53 (2002).