Published in final edited form as:

Cytogenet Genome Res. 2007 ; 117(1-4): 268–277. doi:10.1159/000103188.

Characterising alternate splicing and tissue specific expression in the chicken from ESTs

Haizhou Tang, Thomas Heeley, Raphael Morlec, and Simon J Hubbard

Faculty of Life Sciences, The University of Manchester, Manchester M13 9PT, Tel: (+44) (0)161 306 8930, Fax: (+44) (0)161 275 5082, Email: Simon.Hubbard@manchester.ac.uk

Abstract

Alternate splicing is believed to produce the greatest diversity in transcriptional complexity and function in eukaryotic species. In this study, we present an analysis of alternative splicing events that occur in the chicken, using the recently sequenced genomic sequence and over 580,000 EST sequences mapped back to the genome. A carefully controlled EST-to-genome mapping pipeline is presented, based around the *EXONERATE* program using the *est2genome* model, which also considers several quality control steps to filter out erroneous matches. The data is then used to estimate the level of alternate splicing events with respect to Ensembl predicted transcripts. The EST-genome mappings are characterised at the exon level, in order to classify individual splicing events and provide estimates of novel transcripts not currently annotated by the Ensembl genome database. This is the first large scale analysis of this kind in an avian species, and suggests that chicken displays a similar level of alternate splicing as that found in other higher vertebrates such as human and mouse, both in terms of the number of genes that undergo alternate splicing events, and the average number of transcripts produced per gene. The EST data suggests alternate splicing may occur in some 50-60% of the chicken gene set and with an average of around 2.3 transcripts per gene which undergo this process. The EST data is also used to look at gene and transcript usage in the tissues sequenced in embryonic and adult libraries. Genes which display notable biases were analysed in more detail, including in *twinkl-2* and embryonic heavy chain myosin. This also highlights several as yet functionally un-annotated genes which appear to be important in embryonic tissues and also undergo alternate splicing events. The analysis also demonstrates some of the difficulties involved in using EST-based data to annotate transcriptional activity in eukaryotic genes, where a broad spectrum of tissues and a large number of sequenced transcripts are required in order to fully characterise alternate splicing and differential expression.

Keywords

alternate splicing; Expressed Sequence Tags (ESTs); chicken genome; development; bioinformatics

Introduction

The chicken has played a leading role as a model organism in many fields of biological research, leading to landmark discoveries in several areas, notably immunology, cancer and developmental biology (Brown et al., 2003). Recent efforts have resulted in an expansion of the genomic resources available in this species, from the generation of physical and genetic maps (Groenen et al., 2000; Wallis et al., 2004), sequencing of large scale cDNA libraries (Boardman et al., 2002; Hubbard et al., 2005; Caldwell et al., 2005) through to the

sequencing of the *Gallus gallus* genome itself (Hillier et al., 2004). The approximately 1 gigabase sequence is estimated to contain around 20,000-23,000 protein coding genes, the vast majority of which are supported by physical evidence in the form of cDNA clones. This has helped consolidate the chicken as a pre-eminent model organism, particularly in the area of developmental biology, where research into limb development in chicken embryos has benefited (Tickle 2004). Furthermore, it is possible to generate vectors for RNAi to knock-down developmentally important genes (Das et al 2006). These tools and others promise to help biologists working on avian species study how chicken genes are regulated and controlled to elicit biological functions in this organism.

The genome sequence itself forms the platform for future post-genomic research into avian biology, but requires associated genetic resources to be linked in. This includes the EST and cDNA collections that now number in excess of 580,000 individual reads, which are estimated to cover over 90% of all chicken protein coding genes and correspond to over 70,000 different transcriptional isoforms of the chicken's transcriptional repertoire (Hubbard et al., 2005). These resources have been used to find and annotate the genes within the genome, and reveal the genetic structure of the 18,632 genes currently annotated by the Ensembl team (Eyras et al., 2005). An important aspect on which ESTs can shed light is the number of different transcriptional isoforms that can be produced by a given gene, via alternate splicing of exons (AS), both in the context of different tissues and at different developmental stages of an organism. In the case of the chicken, there are EST sequences available from some 55 tissues from NCBI's GenBank (March 2006).

Pre-mRNA alternate splicing (AS) plays an important biological role, and is believed to be the major contributing factor to complexity in the proteome. It is widespread in eukaryotes, with estimates ranging from 40-60% of all genes undergoing multiple splicing events (Modrek & Lee, 2002; Kan et al., 2002; Zavolan et al., 2003). It can lead to a variety of different functional effects in the altered protein isoforms, from a completely non-functional transcript, to changes in the binding partners, different subcellular localisations and enzymatic activities, as well as changes in post-translational modifications and subtle effects which are difficult to characterize (Stamm et al, 2005). Importantly, alternate/miss-splicing has been linked with diseases including cancers and Alzheimer's (Philips and Cooper, 2000; Faustino and Cooper, 2003).

There have been numerous informatic studies on alternate splicing in the literature, which includes developing generic pipelines to map ESTs to genomes (Kim N et al., 2005a; Kim N et al., 2005b), analysing the differences in alternate splicing in tissues (Yeo et al., 2004), conservation of splice sites across eukaryotes (Nagasaki et al., 2005), exon symmetry and conservation (Magen & Ast, 2005) and correlating splicing variants with gene expression patterns (Kan et al., 2005). These studies, and others, have shown that alternate splicing is not only widespread and tissue specific, but is influenced by evolutionary forces and linked with levels of transcription. There have been a number of studies characterising specific AS events in chick (Mezquita et al., 2000; Zhong et al., 2004) including Fibronectin in the early embryo (Peters et al., 2002), β -crystallins in the lens (Wilmarth et al., 2004) and Nerve Growth Factor (Bertaux et al., 2004). As yet, there have been no comprehensive studies on the apparent level of alternate splicing in chickens, although there are tools developed to examine them (Kim P. et al., 2005; Kim N et al., 2005a, 2005b), and comparative analyses have examined various facets of conservation of splicing, exons, and exon symmetry across other eukaryotes (Kan et al., 2005; Magen & Ast, 2005; Nagasaki et al., 2005). A simple, crude estimate can be readily obtained from the Ensembl database, which reports the number of predicted transcripts associated with each gene. The level of alternate splicing in humans has been comparatively well studied, with estimates of between 40-60% of all genes exhibiting AS events (*c.f.* Modrek & Lee, 2002) The Ensembl chicken annotation version 37

reports a lower proportion of AS events per gene in the chicken genome with respect to human and mouse and the general level of alternate splicing appears diminished in birds compared to mammals. In chicken there are currently 18,632 annotated gene loci (known + novel) corresponding to 24,166 predicted isoforms. This represents 1.3 transcripts per gene (or 2.4 when restricted to the 3,836 genes from which multiple transcripts are predicted). The corresponding figures in humans are 2.0 (3.2 from 11,033 AS genes) and 1.3 (2.4 from 4,986 AS genes) in mouse. This is broadly consistent with recent studies across eukaryotic species, not including avian species, which report decreased complexity of AS from humans/mice to flies/nematodes and birds might be expected to appear in the middle of this trend. However, such observations are difficult to make with certainty since the depth of sequenced, annotated transcript information in the form of mRNAs, cDNAs and ESTs is still not sufficient in human and mouse to offer definitive conclusions (Forrest et al., 2006). Nevertheless, we have examined the apparent level of AS in birds using the chicken as a model system, to try and address this fundamental question. Is AS less prevalent in avian species, or is this simply because less EST/cDNA sequence data is available in comparison with human and mouse? Similarly, we wished to examine differences in tissue or developmental expression patterns inferred from the EST collections. To this end, we present a study of alternate splicing in chickens, which addresses these issues, considering the general level of AS activity observed in chickens, as well as levels of differential in developmental and adult tissues, as well as a number of case studies in genes known to be involved in developmental functions. This is based on a rigorous pipeline for mapping ESTs to the WashU assembled and Ensembl annotated genome sequence, involving quality control steps to minimize artefacts and potential false positives.

Mapping ESTs to the genome

ESTs were downloaded from GenBank, totalling 588,283 sequences covering different libraries associated with 55 tissue types and/or developmental stages. See Supplementary material for full details of the ESTs used and of the EST-genome mapping and annotation pipeline described below and outlined in Figure 1. To detect AS events, ESTs were mapped to the WashU assembled genome sequence using *EXONERATE* with the *est2genome* model (Slater and Birney 2005), which explicitly considers intron/exon boundaries. The original mappings were compared with version 37 of the Ensembl gene build, when 19,360 genes were predicted. For the purposes of this study, EST-genome alignments were further filtered by retaining only the top *EXONERATE* hit, with a bit score in excess of 1000. Furthermore, ESTs were discarded where the EST-genome alignment was ambiguous, with the top two alignment scores within 5% of each other. Given the complicated relationship between ESTs and all possible transcripts, exons were considered as the comparison unit in this study, additionally considering only ESTs displaying multi-exon structure (with at least one potential intron when mapped to the genome). This produced a total of 1,158,498 mapped EST exons from 261,853 EST sequences. Finally, we considered EST exons in two separate groups, mapping either multiple or single transcript Ensembl genes, considering the data as exon pairs where EST exons and Ensembl genic exons overlap. EST exons mapping outside of gene loci were ignored.

Once compared to the Ensembl genome annotation, the EST exons were sub-classified as illustrated in Figure 2, providing the statistics in Table 1. The EST exon statistics refer to pairs of EST-exon/Ensembl transcript exon pairs. A single EST exon may map to corresponding Ensembl exons in several transcripts (since several transcripts may share some common exons). Therefore, the total number of matching pairs of overlapping exons (1,370,423) exceeds the total number of EST exons (1,158,498). Using the ESTs to examine AS events might seem circular, since the cDNAs are used in the genome annotation pipeline by Ensembl (Curwen et al., 2004). However, the Ensembl protocol relies heavily on

orthology to other vertebrate genes, as well as considering cDNAs, and the EST are not considered in the final chicken gene models (Eyras et al., 2005), leaving EST genes as a separate track on the Ensembl web site genome browser. Furthermore, more ESTs are now available, and we wished to exploit the ESTs to examine levels of alternate splicing in the chicken. As shown in Table 1, for multiple-transcript Ensembl genes, the majority of the ESTs that could be unambiguously mapped corresponded “perfectly” to Ensembl exons (93.6% exons from 3,311 genes). A small fraction (4.1%) match only one Ensembl splice site, either extending or reducing the exon. Similarly, there are also a significant number of examples where an Ensembl exon is unsupported by the EST data, termed an extra Ensembl exon, likely to have been predicted via orthology with another vertebrate gene or via a *Gallus gallus* cDNA. These correspond to 2,705 genes and a similar observation is made for single transcript Ensembl genes where 6,439 instances are found. Given that the number of full length chicken cDNAs is relatively modest (Hubbard et al., 2005), less than 5,000 in total, we believe that a large number of the predicted chicken transcripts are not fully supported by physical clones at the exon level. This may well be because the level of EST/cDNA sequencing for chicken is not at the same level as for human or mouse, but even here, it is widely assumed insufficient sequencing has occurred to fully characterise the mammalian transcriptome (Forrest et al., 2006). Finally, a further small fraction of EST exons are not predicted by the Ensembl gene pipeline at all, which are likely to represent novel transcripts not yet annotated by Ensembl. We believe these extra EST exons represent true AS events, since we report only those where the adjacent exons are perfect matches to Ensembl. Furthermore, they show similarly high levels of the expected GT..AG splice sites at the attendant donor and acceptor splice sites.

The trends discussed above are generally observed for both the multi- and single-exon genes., where EST exons are bounded by canonical splice sites (introns with either GT...AG or GC...AG). However, one noticeable exception is observed for the EST exons with only one end perfectly matching (L0rx, lxR0), where the bounding intron on the mismatch side shows reduced agreement with canonical splice sites, the trend being most marked noted for the multi-transcript gene set. We have examined these sites carefully, and these trends are difficult to explain – although it is likely that many are artefacts (mis-aligned, incorrectly mapped, chimeric ESTs, incompletely spliced transcripts). Nevertheless, a significant fraction concur with expected splicing motifs in the attendant introns and are likely to represent novel transcripts. Indeed, we can use the data from Table 1 to estimate the total fraction of the chicken genes which undergo alternate splicing, and the average number of transcripts produced from each gene. Minimally, Ensembl already estimates 3,643 genes undergo AS events, whilst we observed a further 5,818 non-redundant genes from those exhibiting an EST exon with a one-end perfect mismatch and canonical splicing. In addition, another 468 have a totally novel exon not predicted by Ensembl. In total, this brings the total number of genes to 9929 (52%) which display AS events. However, in addition, we see a further 3,266 single transcript genes containing exons which do not have EST support and have not already been counted. If these ESTs genuinely represent transcripts which skip these exons this raises the total to 13,195 (69%) as an upper limit, although this estimate should be treated with some caution. Some of the original version 37 transcript predictions are no longer supported by Ensembl, and although the EST data suggests novel transcriptional forms this may be instead of, rather than in addition to, the original transcript. Nevertheless, these estimates strongly suggest birds display similar levels of AS events to other higher vertebrates, showing similar gene levels (50-60%) to those estimated previously in human and mouse (Modrek & Lee, 2002).

The number of transcripts per gene may be similarly estimated. Ensembl has already annotated 3,836 genes corresponding to 9,372 transcriptional isoforms which equates to 2.4 transcripts per gene for this subset, or 1.3 per gene over the whole genome. In the multi-

transcript subset, we estimate that 1,193 additional transcriptional isoforms are produced from the one end match subset, and a further 201 from the extra EST exons. Similarly, in the single transcript set, 5,818 novel transcripts are predicted from the one-end match set and at least 468 from the extra EST exons. These raise the overall level of AS to 3.0 per gene in the multi-transcript set or 1.7 per gene overall. The total level of AS in gene loci which exhibit AS therefore corresponds to ~2.3 distinct AS events per locus. These results are entirely consistent with other studies, which have observed a comparable fraction of the genome utilised in AS events, with a broad conservation in the number of transcripts per gene around 2.0-2.5 in several eukaryotic species ranging from human/mouse to fly/nematode (Kim N et al., 2005b; Nagasaki et al., 2005). One final further analysis was conducted in a set of kinases whose AS events were already characterised by previous studies (Forrest et al., 2006). In a subset of 650 genes in this list, we compared the predicted number of AS transcripts in the Ensembl genes for human, mouse and chicken. Ensembl version 39 annotates 2.79, 1.57 and 1.43 transcripts per gene to the equivalent genes in human, mouse and chicken respectively, again supporting the observation of a comparable level of splicing, at least in the chicken compared to mouse.

Although the EST data provide a fairly crude estimate of the level of AS in chicken genes, the results suggests AS is more widespread than the conservative Ensembl estimate, and on a par with mammals in terms of the number of transcripts produced per gene. Furthermore, genome informaticians consider EST data to provide a rich source of information to improve gene annotations (Wei & Brent, 2006) and Ensembl uses this information to predict EST genes (Eyras et al., 2004) and is now developing deeper annotations via the Vega gene sets which utilise wider transcript-based information (Ashurst et al., 2005). It is clear that a concerted human-led annotation is necessary at present to develop a high quality map of the complete transcript of a given organism, building on the available transcriptome sequence data.

Mapping EST to specific isoform of alternative splice

In order to further examine tissue specific AS events, we developed a protocol to assign ESTs mapped to specific, unambiguous transcripts from multi-transcript genes, using a simple scoring system whereby EST-transcripts pairs are scored on the number of consistent exons (see Supplementary material). A total of 37,445 ESTs specifically mapped back to a single transcript. Since the number of ESTs from adult libraries outnumbers those from individual embryonic libraries, the latter were merged to form a larger group, so that transcripts with a bias to adult or embryonic tissues could be determined, representing 26,591 ESTs clearly assigned to either embryonic or adult tissues. In total 1,823 genes possess both embryonic and adult mapped ESTs, and from this set a simple tissue bias statistic was calculated, normalised by the total number of mapped ESTs in each category as shown in the equation below.

$$R_i = \log_2 \left(\frac{n_{adult} / T_{adult}}{n_{embryo} / T_{embryo}} \right)$$

Where the log ratio R for a given transcript i is estimated from $n_{adult/embryo}$, the total number of ESTs from either adult or embryonic libraries assigned to i , and $T_{adult/embryo}$ ($T_{adult}=13,116$ $T_{embryo}=13,475$) the total number of ESTs from either adult or embryonic libraries. A further quality control criterion was used; the total number of the ESTs from both adult and embryo numbered over 10, and at least 2 ESTs were mapped to each transcript. The top 10 transcripts and their associated genes over-represented in embryonic

and adult tissues are shown in Table 2. Although this is obviously a very simplistic measure of stage bias, some trends are evident. Interestingly, the embryonic form of heavy chain myosin is represented in the embryonic enriched top 10, as expected, along with several other genes involved in signalling or gene regulation. The top protein is retinoic-acid induced heparin binding protein, a mitogen involved in neurite extension which is known to be expressed during embryogenesis in early stages. Twinfilin, the second top hit is an actin-binding protein known to be important in developmental processes (Walstrom et al 2001), and other examples include gelatinase which is involved in collagen processing in embryonic fibroblasts (Chen et al., 1991). These genes are not only biased in terms of expression in embryonic tissues, but specifically to one of the transcriptional isoforms they can produce. There are also 5 predicted proteins which are not yet functionally annotated, which suggests roles in development for these genes. Indeed, many contain domains common in transcription factors and signalling proteins, consistent with this assertion (e.g. helix-loop-helix and PAS domains in ENSGALG00000015607).

It is less straightforward to make inferences about the adult over-represented genes, although again this list contains genes which appear to exhibit transcriptional isoform bias in the EST libraries. One example, is the RTN1-C gene for which differentially spliced isoforms are known to be expressed in a tissue specific manner in rats (Ninkina et al., 1997). Specifically, the different neuronal compartmental expression is observed to vary in developing and mature rat nervous system. Here, we see much enhanced expression of the adult (mature) transcript compared to the rarer embryonic isoform. We expect similar trends may well be evident in less well characterised genes. Although we present only the top 10 in each category here, a full list of genes in these categories, ranked on their apparent differential expression, is available in the Supplementary materials.

We also examined the full set of transcripts for these genes to see if any bias was observed. For example, Ensembl version 37 gene ENSGALT00000003977 encodes two transcripts ENSGALT00000006337 and ENSGALT00000006336. The longer isoform (ENSGALT00000006337) has an additional N-terminal section containing hits to InterPro WD-40 repeat domains, which are not present in the shorter transcript. The ratio of embryonic:adult multi-exon ESTs unambiguously matching this isoform is 14:5, lower than the 23:2 ratio for ENSGALT00000006336. Although the ratio change is not sufficient to override the apparent overall embryonic dominance, it shows the trend towards increased relative expression of the longer isoform in adult tissues. It is interesting to note that this gene, Twinfilin-2, is produced in two forms in mammals (twinfilin 1 and 2) from different loci, displaying differential expression in tissues and developmental stages, with twinfilin-2 the predominant form in adult heart and skeletal muscles (Vartiainen et al., 2003). The shorter, WD40 free transcript, is almost exclusively biased to embryonic tissues in our data, which is much closer in domain structure to twinfilin-1, the form differentially expressed in embryo. It is not clear what the roles of the WD40 containing isoform might be, although WD40 repeats typically serve as a scaffold in multi-protein complexes, which might augment twinfilin's role in actin filament cleavage and turnover (Moseley et al., 2006). A similar trend is present in embryonic myosin heavy chain from fast skeletal muscle (ENSGALG0000000965), where unambiguously mapped ESTs to isoform ENSGALT00000001422 demonstrate the expected embryonic:adult ratio of 31:4, whilst a ratio of 8:16 is observed for isoform ENSGALT00000001420 which is present in adult muscle, ovary and heart tissues. This suggests this isoform of the gene is also expressed in adult tissues at significant rates. We hope as more and more cDNAs/ESTs become available, such development stage bias trends inferred from ESTs can be more informative to illustrate specific feature of eukaryotic cell alternative splicing.

Alternate splicing in developmentally associated genes

Given the wide use of chick as a model to study developmental biology, we examined the level of alternate splicing in genes associated with developmental processes. The Ensembl gene set was examined and a BioMart search (Kasprzyk et al., 2004) conducted to yield genes with GO terms, which were subsequently screened for those whose description contained the keyword “development” giving 428 terms. Similarly, genes were considered if they were assigned to any of 541 InterPro identifiers returned from a similar keyword search of the InterPro database (Mulder et al., 2005; www.ebi.ac.uk/interpro). When cross-referenced with Ensembl genes predicted to code for multiple transcripts this search yielded 213. The top 5 were ordered on the number of matching ESTs, listed in Table 3, and were examined in more detail. These genes represent a variety of functions and biological processes, including the highly abundant egg white protein cystatin, which is known to exhibit tissue dependent expression and is under developmental control and has been implicated in a variety of function through its inhibition of cysteine proteases (Colella et al., 1989). Another protein listed is Cathepsin A which is known to undergo alternate splicing, producing two forms, the former of which exhibits classic galactosidase enzymatic properties, whilst the enzymatically inactive second form binds elastin (Privitera et al., 1998). Unfortunately, on closer inspection, sufficient EST evidence for the different isoforms was not present in the data to make definitive statements about the differential expression of the AS transcripts. This is well illustrated by the nuclear hormone receptor family, where several members were noted in the “developmental” selected list, including retinoic acid receptor (Retinoid X Receptor alpha) shown in Figure 3. This is fairly typical, where several multi-exon transcripts map well to the predicted gene structure, but there is no definitive evidence that transcript 2 is expressed. Indeed, in this particular instance, the latest gene model for this gene has only one Ensembl transcript predicted in chick. Another notable feature is the two ESTs which map to this locus with high fidelity and yet do not share similar exonic structure with any of the predicted transcripts. Similarly, around 100 single “exon” ESTs mapping to several regions across this gene have been omitted for clarity. These may well be functional transcripts, such as antisense or non-coding RNAs of some type, or indeed could also be clonal artefacts. It is difficult to offer definitive conclusions without extensive further research.

It was hoped that this list would reveal candidate genes that were involved in developmental processes, exhibited alternate splicing events and were well covered by confident, multi-exon ESTs that could be unequivocally assigned to a specific transcript. Unfortunately, this did not appear to be the case and this illustrates one of the difficulties of this type of approach; insufficient data in general and the rarity of full length AS transcripts.

Discussion

The preliminary analyses presented in this paper represent the first large scale attempts to characterise alternate splicing in avian species using EST data mapped back to the genome. This generic EST-genome mapping approach has been studied by several groups, and offers a simple way to characterise AS events (Kim N et al., 2005a; Kim P. et al., 2005; Kim N et al., 2005b;). The EST data suggest birds exhibit as much alternate splicing as mammals from their gene loci and this is consistent with other metazoans studied to date (Modrek and Lee, 2002; Nagasaki et al., 2005). The chicken Ensembl gene set relies on orthology and cDNA evidence rather than ESTs (Eyras et al., 2005) and in this case the inclusion of EST data is extremely useful for characterisation of AS. However, although the overall level of splicing appears to be consistent with other metazoans, a significant fraction of the transcripts predicted via orthology-based pipelines are apparently unsupported by the EST evidence currently. This suggests that many chicken genes do not show the full range of AS

events (and hence transcripts) as currently annotated by Ensembl. It should also be pointed out that not all Ensembl genes have EST support at present (2758 of the 19,360 genes have no corresponding ESTs after our *EXONERATE* analysis) and that chicken lies some way behind human and mouse in terms of available transcriptional data. This should not be interpreted as a lack of confidence in the current gene predictions, as a recent analysis has demonstrated the high quality of Ensembl's annotation pipeline. Nevertheless, recent analyses have suggested that the many millions of human/mouse ESTs/cDNAs are still not sufficient to truly know the depth of AS events in human and mouse (Forrest et al., 2006) and the situation must clearly be less favourable for less well studied species.

This study also highlights the difficulties in dealing with ESTs to predict gene structure and AS events, which has been discussed at length by Modrek and Lee (2002). Simply mapping ESTs to genomic sequence is surprisingly difficult, due to the relatively low fidelity nature of EST sequence, similar but distinct genes and/or paralogues, occasional gaps in genomic sequence, and the computational scale of the problem (our analysis took several cpu weeks on a 20 node cluster). Once mapped, relating EST structure back to gene models is also difficult and obtaining sufficient depth of transcriptional information on a genome wide basis to characterise AS events is similarly challenging. In this study we have explored a novel yet simple scoring algorithm to assign at least some of the ESTs unambiguously to a single isoform of the predicted gene and achieved some success. We believe this generalised approach offers a simple way to unambiguously assign ESTs to specific isoforms, and although this applies to a minority of the ESTs available, useful information can certainly be gleaned for a subset of genes/isoforms. We believe EST information remains invaluable for assisting gene model definition (Wei & Brent, 2006), and understanding the range of transcriptional activity in an organism.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge funding from the BBSRC via grant BSB13497 to SJH supporting HT. We would also like to thank John Brooke and the e-Science North West (ESNW) centre for providing additional computational resources for the *exonerate* jobs.

References

- Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, et al. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* 2005; 33:D459–D465. [PubMed: 15608237]
- Bertaux O, Toselli-Mollereau E, Auffray C, Devignes MD. Alternative usage of 5' exons in the chicken nerve growth factor gene: refined characterization of a weakly expressed gene. *Gene.* 2004; 334:83–97. [PubMed: 15256258]
- Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, et al. Ensembl 2006. *Nucleic Acids Res.* 2006; 34:D556–D561. [PubMed: 16381931]
- Boardman P, Sanz-Ezquerro J, Overton I, Burt D, Bosch E, Fong W, et al. A comprehensive collection of chicken cDNAs. *Current Biology.* 2002; 12:1965–1969. [PubMed: 12445392]
- Brown W, Hubbard SJ, Tickle C, Wilson SA. The chicken as a model for large-scale analysis of vertebrate gene function. *Nature Rev Genetics.* 2003; 4:87–98. [PubMed: 12560806]
- Caldwell RB, Kierzek AM, Arakawa H, Bezzubov Y, Zaim J, Fiedler P, et al. Full-length cDNAs from chicken bursal lymphocytes to facilitate gene function analysis. *Genome Biol.* 2005; 6:R6. [PubMed: 15642098]

- Chen JM, Aimes RT, Ward GR, Youngleib GL, Quigley JP. Isolation and characterization of a 70-kDa metalloprotease (gelatinase) that is elevated in Rous sarcoma virus-transformed chicken embryo fibroblasts. *J. Biol. Chem.* 1991; 266:5113–5121. [PubMed: 1848240]
- Colella R, Sakaguchi Y, Nagase H, Bird JW. Chicken egg white cystatin. Molecular cloning, nucleotide sequence, and tissue distribution. *J Biol Chem.* 1989; 264:17164–9. [PubMed: 2793849]
- Curwen V, Eyraas E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M. The Ensembl Automatic Gene Annotation System. *Genome Res.* 2004; 14:942–950. [PubMed: 15123590]
- Das RM, Van Hateren NJ, Howell GR, Farrell ER, Bangs FK, Porteous VC, et al. A robust system for RNA interference in the chicken using a modified microRNA operon. *Dev. Biol.* 2006; 294:554–563. [PubMed: 16574096]
- Eyraas E, Caccamo M, Curwen V, Clamp M. ESTGenes: Alternative Splicing From ESTs in Ensembl. *Genome Res.* 2004; 14:976–987. [PubMed: 15123595]
- Eyraas E, Reymond A, Castelo R, Bye JM, Camara F, Flicek P. Gene finding in the chicken genome. *BMC Bioinformatics.* 2005; 6:131. [PubMed: 15924626]
- Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. *Genes Dev.* 2003; 17:419–437. [PubMed: 12600935]
- Forrest AR, Taylor DF, Crowe ML, Chalk AM, Waddell NJ, Kolle G. Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases. *Genome Biol.* 2006; 7(1):R5. [PubMed: 16507138]
- Groenen MA, Cheng HH, Bumstead N, Benkel BF, Briles WE, Burke T. A consensus linkage map of the chicken genome. *Genome Res.* 2000; 10:137–47. [PubMed: 10645958]
- Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature.* 2004; 432:695–716. [PubMed: 15592404]
- Hubbard SJ, Grafham DV, Beattie KJ, Overton IM, McLaren SR, Croning MD. Transcriptome Analysis for the Chicken Based on 19,626 Finished cDNA Sequences and 485,337 Expressed Sequence Tags. *Genome Res.* 2005; 15:174–183. [PubMed: 15590942]
- Kan Z, States D, Gish W. Selecting for functional alternative splices in ESTs. *Genome Res.* 2002; 12:1837–45. [PubMed: 12466287]
- Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Res.* 2004; 14:160–169. [PubMed: 14707178]
- Kim N, Lim D, Lee S, Kim H. ASePCR: alternative splicing electronic RT-PCR in multiple tissues and organs. *Nucleic Acids Res.* 2005a; 33:W681–5. [PubMed: 15980562]
- Kim N, Shin S, Lee S. ECgene: genome-based EST clustering and gene modeling for alternative splicing. *Genome Res.* 2005b; 15(4):566–76. [PubMed: 15805497]
- Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S. ECgene: genome annotation for alternative splicing. *Nucleic Acids Res.* 2005; 33:D75–9. [PubMed: 15608289]
- Magen A, Ast G. The importance of being divisible by three in alternative splicing. *Nucleic Acids Res.* 2005; 33:5574–82. [PubMed: 16192573]
- Mezquita J, Mezquita P, Montserrat P, Mezquita B, Francone V, Vilagrasa X, Mezquita C. Genomic structure and alternative splicing of chicken angiopoietin-2. *Biochem Biophys Res Commun.* 2000; 275:643–51. [PubMed: 10964717]
- Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet.* 2002; 34:177–80. [PubMed: 12730695]
- Moseley JB, Okada K, Balcer HI, Kovar DR, Pollard TD, Goode BL. Twinfilin is an actin-filament-severing protein and promotes rapid turnover of actin structures in vivo. *J Cell Sci.* 2006; 119:1547–57. [PubMed: 16569665]
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro, progress and status in 2005. *Nucleic Acids Res.* 2005; 33:D201–5. [PubMed: 15608177]
- Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene.* 2005; 364:53–6. [PubMed: 16219431]

- Ninkina NN, Baka ID, Buchman VL. Rat and chicken s-rex/NSP mRNA: nucleotide sequence of main transcripts and expression of splice variants in rat tissues. *Gene*. 1997; 184:205–10. [PubMed: 9031629]
- Pekarik V, Bourikas D, Miglino N, Joset P, Preiswerk S, Stoeckli ET. Screening for gene function in chicken embryo using RNAi and electroporation. *Nature Biotechnology*. 2002; 21:93–96.
- Peters J, Sechrist J, Luetolf S, Loredó G, Bronner-Fraser M. Spatial expression of the alternatively spliced EIIIB and EIIIA segments of fibronectin in the early chicken embryo. *Cell Commun Adhes*. 2002; 9:221–38. [PubMed: 12699090]
- Philips AV, Cooper TA. RNA processing and human disease. *Cell Mol Life Sci*. 2000; 57:235–249. [PubMed: 10766020]
- Privitera S, Prody CA, Callhan JW, Hinek A. The 67kDa enzymatically inactive alternatively spliced variant of β -galactosidase is identical to the elastin/laminin-binding protein. *J Biol Chem*. 1998; 273:6319–6326. [PubMed: 9497360]
- Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005; 6:31. [PubMed: 15713233]
- Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, et al. Function of alternative splicing. *Gene*. 2005; 344:1–20. [PubMed: 15656968]
- Tickle C. The contribution of chicken embryology to the understanding of vertebrate limb development. *Mech. Development*. 2004; 121:1019–1029. [PubMed: 15296968]
- Vartiainen MK, Sarkkinen EM, Matilainen T, Salminen M, Lappalainen P. Mammals have two twinfilin isoforms whose subcellular localizations and tissue distributions are differentially regulated. *J Biol Chem*. 2003; 278:34347–55. [PubMed: 12807912]
- Wahlstrom G, Vartiainen M, Yamamoto L, Mattila PK, Lappalainen P, Heino TI. Twinfilin is required for actin-dependent developmental processes in *Drosophila*. *J Cell Biol*. 2001; 155:787–96. [PubMed: 11724820]
- Wallis JW, Aerts J, Groenen MA, Crooijmans RP, Layman D, Graves TA, et al. A physical map of the chicken genome. *Nature*. 2004; 432:761–4. [PubMed: 15592415]
- Wei C, Brent MR. Using ESTs to improve the accuracy of de novo gene prediction. *BMC Bioinformatics*. 2006; 7:327. [PubMed: 16817966]
- Wilmarth PA, Taube JR, Riviere MA, Duncan MK, David LL. Proteomic and sequence analysis of chicken lens crystallins reveals alternate splicing and translational forms of beta B2 and beta A2 crystallins. *Invest Ophthalmol Vis Sci*. 2004; 45:2705–15. [PubMed: 15277495]
- Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome Biol*. 2004; 5:R74. [PubMed: 15461793]
- Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, et al. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res*. 2003; (6B):1290–300. [PubMed: 12819126]
- Zhong D, Yu W, Liu Y, Liu J, Li J. Molecular cloning and expression of two chicken invariant chain isoforms produced by alternative splicing. *Immunogenetics*. 2004; 56:650–6. [PubMed: 15578263]

Figure 1. Schematic of the EST to genome mapping pipeline
Processing steps in the mapping and analysis of ESTs, their exons, and mappings to Ensembl predicted genes are shown in schematic form, with appropriate statistics on the numbers of ESTs/exons/genes passing different stages shown in the figure. Further details are given in the Supplementary Materials.

Figure 2. Exemplar EST mappings to predicted Ensembl transcript structure for a single gene
 The figure shows the nomenclature discussed in the Methods and Table 1 which shows how different EST exons once mapped to the gene transcripts are classified by the status which their 5' and 3' ends correspond to predicted transcript structure. The naming convention in the schema are: capital letters indicate a perfect match, normally followed by 0, which means the coordinate difference between the EST exon and the Ensembl exon is 0; the small letter indicates a non-perfect match followed by an "x". A "L" or "l" refers to the 5' end and "R" or "r" refers to the 3' end of the EST exon. The 5p and 3p only apply to the terminal EST exon, such as 5pR0, which means the 5' terminal EST exon has a perfect matching at the 3' end.

Figure 3. Gene Transcripts and ESTs from the Retinoic Acid Receptor gene

The gene transcripts which are a result of alternative splicing are shown at the top, the purple boxes are the exons and the thin lines are the introns. The ESTs that have more than 1 intron are shown underneath matching to their specific sequence. Their tissue specificity and developmental stage are also shown for each EST. The scale at the top shows the size of the transcripts and of the ESTs in Kilo bases (Kb)

EST-genome matching statistics

Table 1

Match status	EST-Ensembl exon pairs	Ensembl Transcripts (ENSGALT)	Ensembl genes (ENSGALG)	AS site statistics (GT...AG)(GC...AG)	
				U-introns	D-introns
Multi-transcript Ensembl v.37 genes (3,836 genes)					
Mapped EST exons ^a	619,157 (100%)	8,897	3,643	-	-
Perfect match	579,385 (93.6%)	7,629	3,311	99.3%	99.3%
Both ends (LOR0)	325,645			-	-
3' perfect (5pR0) ^b	128,155			99.3%	-
5' perfect (L03p) ^b	125,582				
One end match	12,982 (2.1%)	3,620	1,733	99.0%	35.4%
5' perfect match (L0rx)	13,507 (2.2%)	3,666	1,743	37.7%	98.9%
3' perfect match (lxR0)		(2,456 in common)	(1,256 in common)		
extra EST exons (skip)	427	(n/a)	201	97.0%	97.4%
extra Ensembl exon:	11,266	5,161	2,705		
Single-transcript Ensembl v.37 genes (15,524 genes)					
Mapped EST exons ^a	751,266 (100%)	12,994	12,994	-	-
Perfect match	648,429 (66.6%)	11,338	11,338	99.4%	99.3%
Both ends (LOR0)	342,743			-	99.0%
3' perfect (5pR0) ^b	152,616			99.1%	-
5' perfect (L03p) ^b	153,070				
One end match	25,696 (2.6%)	4,815	4,815	99.5%	63.7%
5' perfect match (L0rx)	25,666 (2.6%)	4,816	4,816	65.8%	98.8%
3' perfect match (lxR0)		(3,387 in common)	(3,387 in common) (6,244 distinct)		
extra EST exons (skip)	1,103	(n/a)	468	96.96%	97.42%
extra Ensembl exon:	15,895	6,439	6,439		

^a includes all EST exons mapped completely within a gene locus defined by the first and last base defined by Ensembl in genomic co-ordinates of the WashU assembly, WASHUC 1, Mar 2004. Some EST exons map to a gene but do not correspond to any of the cases shown in the Table.

^b instances where the EST exon is the first or last in the EST and the sequence ends prematurely at either the 5' or 3' end, and hence no inference can be made at that terminus although the other end matches perfectly.

Table 2

Differential EST expression in embryonic and adult libraries

Ensembl Gene ID*	Ensembl transcript ID	EST No (embryonic : adult)	R_i	Gene description
<i>Enrichment in embryonic libraries</i>				
ENSGALG00000008367	ENSGALT000000013622	24:2	3.62	Midkine precursor (Retinoic acid-induced heparin-binding protein) (RI- HB)
ENSGALG00000003977	ENSGALT00000006336	23:2	3.56	similar to twinfilin-like protein; protein tyrosine kinase 9 related protein; A6-related protein; PTK9 protein tyrosine kinase 9-like (A6-related protein); twinfilin-2
ENSGALG00000004901	ENSGALT00000007835	54:6	3.21	Hypothetical protein (Fragment)
ENSGALG00000004160	ENSGALT00000006615	17:2	3.13	similar to Ras-GTPase-activating protein binding protein 1 (GAP SH3-domain binding protein 1) (G3BP-1) (DNA helicase VIII) (HDH-VIII)
ENSGALG000000015607	ENSGALT000000025163	17:2	3.13	Hypothetical protein
ENSGALG00000000965	ENSGALT00000001422	31:4	2.99	Myosin heavy chain, fast skeletal muscle, embryonic.
ENSGALG000000003580	ENSGALT00000005666	15:2	2.95	72 kDa type IV collagenase precursor (EC 3.4.24.24) (72 kDa gelatinase) (Matrix metalloproteinase-2) (MMP-2) (Gelatinase A).
ENSGALG00000006020	ENSGALT00000009692	22:3	2.91	Hypothetical protein.
ENSGALG00000005451	ENSGALT00000008749	14:2	2.85	Hypothetical protein (Fragment).
ENSGALG00000008220	ENSGALT000000033478	14:2	2.85	basic leucine zipper and W2 domains 1
<i>Enrichment in adult libraries</i>				
ENSGALG000000015767	ENSGALT000000025427	4:119	-4.85	fatty acid binding protein 4, adipocyte
ENSGALG000000002397	ENSGALT00000003775	3:84	-4.77	Phosphoserine phosphatase (EC 3.1.3.3) (PSP) (O-phosphoserine phosphohydrolase) (PSPase) (L-3-phosphoserine phosphatase).
ENSGALG000000004574	ENSGALT000000034931	4:64	-3.96	Atrial natriuretic factor precursor (ANF) (Atrial natriuretic peptide) (ANP) (Prepronatriodilatin).
ENSGALG000000017219	ENSGALT000000027832	2:21	-3.35	No description
ENSGALG000000009019	ENSGALT000000014676	4:36	-3.13	MSFL2541
ENSGALG000000011940	ENSGALT000000031131	4:27	-2.72	Cholecystokinins precursor (CCK) [Contains: Cholecystokinin 70 (CCK70); Cholecystokinin 8 (CCK8); Cholecystokinin 7 (CCK7)].
ENSGALG000000000620	ENSGALT000000000878	3:20	-2.70	heart fatty acid-binding protein
ENSGALG000000012006	ENSGALT000000019594	7:38	-2.40	ChS-Rex-s (RTN1-C).
ENSGALG000000006731	ENSGALT000000010889	2:10	-2.28	similar to TAF3 protein
ENSGALG000000009003	ENSGALT000000014652	4:20	-2.29	Synaptosomal-associated protein 25 (SNAP-25) (Synaptosomal-associated 25 kDa protein) (Super protein) (SUP).

* all genes are predicted to express at least 2 transcriptional isoforms.

Table 3

Ensembl genes involved in development with large numbers of mapped ESTs

Ensembl Gene ID (no. of predicted transcripts)	Gene description	ESTs mapped
ENSGALG00000008664 (2)	Cystatin precursor (Egg-white cystatin).	332
ENSGALG00000006876 (4)	similar to protective protein for beta-galactosidase; Protective protein for beta-galactosidase (cathepsin A); beta-galactosidase 2	87
ENSGALG00000008432 (2)	Hypothetical protein	83
ENSGALG00000003454 (2)	Zinc finger protein 207. [from human gene ENSG00000010244]	71
ENSGALG00000010257 (2)	Cysteine and glycine-rich protein 2 (Cysteine-rich protein 2) (CRP2) (Beta-cysteine-rich protein) (Beta-CRP).	67