Incorporating RNA-seq data into the zebrafish Ensembl genebuild

John E. Collins,[1] Simon White, Stephen M.J. Searle, and Derek L. Stemple

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom

Ensembl gene annotation provides a comprehensive catalog of transcripts aligned to the reference sequence. It relies on publicly available species-specific and orthologous transcripts plus their inferred protein sequence. The accuracy of gene models is improved by increasing the species-specific component that can be cost-effectively achieved using RNA-seq. Two zebrafish gene annotations are presented in Ensembl version 62 built on the Zv9 reference sequence. Firstly, RNA-seq data from five tissues and seven developmental stages were assembled into 25,748 gene models. A 3′-end capture and sequencing protocol was developed to predict the 3′ ends of transcripts, and 46.1% of the original models were subsequently refined. Secondly, a standard Ensembl genebuild, incorporating carefully filtered elements from the RNA-seq-only build, followed by a merge with the manually curated VEGA database, produced a comprehensive annotation of 26,152 genes represented by 51,569 transcripts. The RNA-seq-only and the Ensembl/VEGA genebuilds contribute contrasting elements to the final genebuild. The RNA-seq genebuild was used to adjust intron/exon boundaries of orthologous defined models, confirm their expression, and improve 3′ untranslated regions. Importantly, the inferred protein alignments within the Ensembl genebuild conferred proof of model contiguity for the RNA-seq models. The zebrafish gene annotation has been enhanced by the incorporation of RNA-seq data and the pipeline will be used for other organisms. Organisms with little species-specific cDNA data will generally benefit the most.

[Supplemental material is available for this article.]

As vertebrate transcriptomes continue to be scrutinized they reveal ever-increasing levels of complexity. Deciphering the transcribed regions from the nontranscribed regions and presenting a comprehensive, yet artifact free, gene set is a significant undertaking. Annotating the gene content of a genomic reference sequence is fundamental to understanding the biological processes of the organism. Moreover, for a model organism like zebrafish, being able to link a human gene of medical interest to the zebrafish ortholog is key to elucidating human gene function.

Gene annotation methods rely on cDNA sequencing or ab initio prediction, or a combination of both. Vertebrate methods tend to rely on sequence information. The Vertebrate Genome Annotation Database (VEGA) (Wilming et al. 2008) provides a manually curated set of gene models derived from transcript sequence data, while the RefSeq Database (Pruitt et al. 2009) provides a predicted and manually curated gene set based on genomic and transcript sequence data. In contrast, Ensembl provides an automated annotation system for vertebrate and other eukaryote species (Flicek et al. 2011). An Ensembl genebuild predicts gene models using a publicly available species-specific and orthologous sequence aligned to the reference genome. Some organisms have a lot of species-specific cDNA sequence data, while others have little. As more information enters the public databases, particularly species-specific data, the Ensembl annotation improves in both quantity and quality. The advent of high-throughput transcriptome sequencing, RNA-seq (Wang et al. 2009), marks a major advance in species-specific experimentally derived sequence data and has the potential to make a significant impact on Ensembl genebuilds. In particular, RNA-seq can add untranslated regions of predicted orthologous transcript and provide proof of transcription.

Recently, new sequencing technologies, such as the Illumina Genome Analyzer (Bentley et al. 2008) used in this study, have increased sequencing capacity and enabled alternative strategies for many sequence-based investigations. RNA-seq, for example, has allowed extremely deep sequencing of complementary DNA to an extent that was not possible using cDNA libraries and capillary sequencing. Unlike traditional directed cDNA-sequencing strategies (Temple et al. 2009), multitranscript sampling and the depth of the sequence in RNA-seq reduce noise caused by occasional misspliced mRNA. RNA-seq also assays the frequency of alternative splice forms and their spatial and temporal expression patterns, giving a comprehensive snapshot of the transcription of the whole sample. Studies have used RNA-seq to look at the transcriptome of yeast (Nagalakshmi et al. 2008), fission yeast (Wilhelm et al. 2008), mouse (Mortazavi et al. 2008), and human (Cloonan et al. 2008; Sultan et al. 2008; Wang et al. 2008) and found new levels of complexity. Methods have been published that use RNA-seq sequence reads alone to build gene models in a reference genome guided manner (Denoeud et al. 2008; Yassour et al. 2009; Guttman et al. 2010; Trapnell et al. 2010) or assemble the transcriptome independently of a reference sequence (Zerbino and Birney 2008; Robertson et al. 2010), in one case using a related proteome to aid assembly (Surget-Groba and Montoya-Burgos 2010). RNA-seq data is just one piece of information that can be used to predict a gene, but it is a cost-effective way to increase the amount of species-specific cDNA data for the less well-studied genomes. Recently, zebrafish gene collections derived using high-throughput sequencing technologies have been published, which look at the transcriptome of early development and lincRNA (Ulitsky et al. 2011; Pauli et al. 2012).

Our aim was to distill a large quantity of zebrafish RNA-seq short reads into a useful gene annotation and incorporate this into the Ensembl genebuild pipeline. RNA-seq assays all polyA tran-

[1]Corresponding author
E-mail jec@sanger.ac.uk
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.137901.112.
Freely available online through the *Genome Research* Open Access option.

scriptional activity in the sample, and we present our initial endeavor at a concise interpretation. In an attempt to minimize the loss of biologically significant data while excluding artifacts, we have incorporated numerous filtering steps throughout the genebuilding process. Our RNA-seq-only genebuild and Ensembl/VEGA genebuild bring complementary elements to the final gene set. Incorporating the RNA-seq data into the Ensembl/VEGA genebuilds adds a large amount of new species-specific cDNA sequence to current publicly available gene data providing a comprehensive gene annotation. In this analysis we have concentrated on protein-coding genes, although there is scope for further modifications to annotate all of the transcribed sequence. We present an advance in the annotation of the zebrafish genome and a template for improving Ensembl annotation in other species.

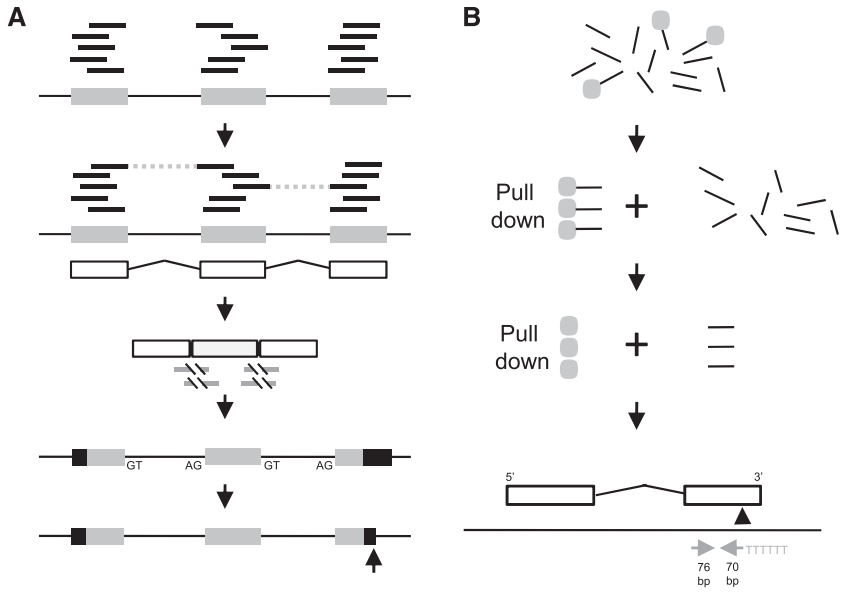

Figure 1. RNA-seq only gene models and 3′-end pull-down pipelines. (*A*) Illumina reads (short black lines) were matched to repeat-masked genomic reference sequence (long black line) using exonerate, and clusters were called as potential exons (gray boxes). Read pair information was used to identify adjacent exons belonging to the same gene (dashed lines). A rough gene model was built (white boxes are exons and angled black lines are introns). A total of 20 bases of genomic sequence were added to each side of the exons, and exons within a rough model were concatenated. Illumina reads were mapped again to the concatenated rough models using exonerate est2genome probabilistic splice model (gray lines are the aligned portion and the black slash lines show the breaks across the intron, identifying intron spanning reads). Exons were trimmed or extended to the identified splice sites and located back on the genome reference. The splice sites were used to identify the correct strand, and the longest open reading frame was predicted in each refined gene model (black are untranslated regions and gray boxes are coding). 3p markers were located at the 3′ end of the gene (black arrow) and the models were extended or trimmed as necessary. (*B*) Total RNA was chemically fragmented (short black lines), a 5′ biotinylated polyT$_{22}$ anchored primer containing a BpmI site (underlined) (GGCCAGTCCTGGAGTTTTTTTTTTTTTTTTTTTTTTTVN) was annealed and bound to a streptavidin magnetic bead (gray circle). The polyA RNA fragments were pulled down with a magnet, the rest of the RNA was thrown away and double-stranded cDNA was synthesized. The cDNA was released from the beads by restriction digest with BpmI, which cuts 16 bases upstream of the recognition site, leaving 6 T bases at the 5′ end of the fragment. Standard Illumina libraries were made, followed by paired end sequencing of 76 bases each. Fragments with a 3′ polyA (5′ polyT) resulted in two reads where one read began with T bases derived from the polyA pull-down oligo. If there were at least five T bases, all of the T bases were removed; this was normally six T bases to produce 70/76 base pair of reads. These were mapped to the genomic reference sequence (gray arrows) with the 5′ end of the 70-base read indicating a possible polyA addition site as well as the orientation of the transcript. These were filtered for duplicate read pairs, the proximity of a BpmI site in the genome sequence, polyA or degenerate polyA adjacent to the proposed 3′ end, the orientation of the proposed 3′ end compared with the gene model and for model extension the presence of overlapping genes on the other strand. The original RNA-seq models were not extended more than 5001 bases. If more than three reads satisfied all of the filters, the genomic coordinate of the 3p marker was predicted (black arrow).

Results

Genebuild process

To further the annotation of the zebrafish genome we present two automated genebuilding processes. Firstly, gene models were constructed using RNA-seq data only. Secondly, elements were selected from these models, incorporated into the Ensembl genebuild (Flicek et al. 2011) which was merged with the VEGA manual annotation (Wilming et al. 2008). Both genebuilds are available for Ensembl version 62 and onward. The material for the RNA-seq sequences was collected from five tissues and seven developmental stages in an attempt to include as many genes as possible. Illumina libraries were constructed and sequenced multiple times to improve the representation of less-abundant transcripts, and as the available read length increased. The initial libraries were prepared from polyT or random primed polyA+ cDNA fragmented by nebulaisation. However, for the majority of the sequencing we used a standard Illumina library protocol adopted by the Wellcome Trust Sanger Institute library preparation pipeline (see Methods; Supplemental Table 1). This method chemically fragmented polyA+ RNA, reverse transcribed with random primers, and then followed a standard Illumina library protocol.

RNA-seq gene models

We have developed an automated pipeline to identify gene models in a genome reference sequencing by using only RNA-seq short reads and with no ab initio intron prediction (Fig. 1A). Initially, approximate exons were defined by mapping 553,120,413 out of 807,717,784 (68.5%) RNA-seq reads to repeat-masked (www.repeatmasker.org) zebrafish genome (Zv9) using exonerate (Slater and Birney 2005; Supplemental Table 1). During the course of the project the length of Illumina paired end-sequence reads increased from 36 to 37, then 54, and, finally, 76 for each read, and there is a slight increase in the percentage aligned with time that could be due to improvements in sequence quality. To build the gene models, initially clusters of reads were identified as potential exons, then read pair information was exploited to concatenate exons into rough gene models. Secondly, reads were realigned to the concatenated exons of the rough models using exonerate with the est2genome alignment model with the aim of locating the intron splice sites. Every intron has at least one supporting read. These data were collected into an intron database comprising 7% of the total reads (Supplemental Table 1). The introns were used to refine the rough transcripts to create stranded transcripts that

support an open reading frame. This process was repeated 13 times, once for each of the 12 individual tissues and once on a pool of all tissues. Each time, the code produced a single model per locus, representing the best supported transcript for the given read alignments. In total, 412,915 models were produced. The pooled tissue models were used to create an initial gene set of 34,282. However, it was observed that some of these models were clearly incorrect and a set of simple filters was applied to remove incomplete or artifactual transcripts (see Methods), giving 21,224 models. A further 2673 were identified as potentially merged. This was resolved by replacing the merged models with single tissue models where better models were possible; 240 apparently merged pooled tissue models were replaced by 1016 single tissue models. A further 169 single exon models were added to give a total of 24,842. In addition, 1889 noncoding models were identified. All of these models were assessed by the 3′-end trimming or extending script (see below). This involves removing overlapping models, resulting in 25,748 models comprising 24,088 protein-coding and 1660 noncoding models. After more detailed analysis it became apparent that there are 428 models, 217 classed as coding, under 300 bases long, which escaped the filtering process. In future refinements of the genebuild pipeline it would be sensible to examine these models carefully with a view to making the filters more sensitive.

Correctly predicting the start and end of transcription for RNA-seq models proved difficult. Due to reads extending from the ends of transcripts, such as overlapping transcripts on the opposite strand, models were artificially lengthened. Similarly, breaks in contiguity of bases covered by mapped reads, for example, due to repeat masking or simple repeat sequences, resulted in shortened models. Previously, high-throughput sequencing has been used to map the polyadenylation sites of transcripts in worm, yeast, and human (Mangone et al. 2010; Ozsolak et al. 2010; Yoon and Brem 2010). To predict the zebrafish 3′ ends more accurately we developed a method to pull down and sequence the 3′ ends of transcripts (Fig. 1B). This method selects for polyA sequence in an RNA fragment and generates paired-end reads of 76 bases. Interestingly, these 3′-end pull-down libraries align to the genome better than the full-transcript RNA-seq libraries (Supplemental Table 1). This is probably due to the fact that there are fewer introns in the 3′-untranslated regions of the genes. The 3′ ends of the genes, referred to as 3p markers, were predicted as described in Figure 1B. The original RNA-seq models were extended or trimmed to the 3p marker, which was supported by the most reads, and which fulfilled all the filtering criteria (Table 1). In total 46.1% of the models were altered with almost twice as much sequence added by extensions (7,092,396 bases) than trimmed off (3,692,129 bases) the original RNA-seq models. Alternative 3p markers were also associated with the gene models. Despite success in refining many 3′-end predictions of RNA-seq models, as with any automated annotation system, not every gene model altered by a 3p marker has a correctly annotated 3′ end. Observations made in the final gene models include problems with the genome reference sequence that resulted in padding N bases from exons spanning gaps in the genome and extending to the 3′ end of the neighboring gene. A detailed analysis of the success of the 3′-end prediction is given below.

During the creation of the RNA-seq model pipeline it became apparent that we were sometimes failing to produce full-length models. One reason was the presence of exons that were too short to be defined by reads aligned to the genome. To identify short exons, rough models that appeared to give rise to multiple adjacent refined models on the same strand were targeted. Partially aligned reads adjacent to the apparent split were realigned using a very short (4 bp) word length and the exonerate est2genome model. Complete models could then be built from single reads, which spliced into and out of a short exon. Gene models also encounter problems in which the underlying reference genome sequence is incorrectly assembled. An example, shown in Figure 2, reveals the *bace2* gene split into three RNA-seq models. The two models representing the beginning and end of the gene are adjacent in the reference sequence, while the middle portion is located 340,000 bases away. In a genome-guided annotation process, rectifying such transcripts will depend on correction of the reference sequence. It was also observed that models were fragmenting at noncanonical splice sites. Initially, the introns were aligned with the exonerate est2genome alignment model restricted to canonical splice sites, and the removal of this restriction allowed the est2genome probabilistic splice model to identify noncanonical splices. Gene model construction required a minimum depth of two noncanonical intron-spanning reads. Finally, it was observed that very long gene models, such as *ttnb*, were fragmented because the number of possible isoforms was too large to calculate within a limit of 100,000 iterations. Models were simplified by removing potential splices that were poorly supported until the model could be constructed within the constraints of the algorithm.

Table 1. Transcripts in zebrafish genebuilds Ensembl release 62

Ensembl *62* genebuilds		Number of transcripts
RNA-seq only models	3p marker alteration:	
	Trimmed	8,639 (33.6%)
	Extended	3,189 (12.4%)
	Confirmed	31 (0.1%)
	Not altered	13,889 (53.9%)
	Total	25,748
Standard Ensembl pipeline (interim set of models not shown displayed Ensembl)	Transcript source:	
	Zebrafish cDNA	15,048
	Orthologous protein	5,902
	RNA-seq only model	8,374
	Total	29,324
Ensembl with RNA-seq/VEGA merge	Biotype:	
	Protein coding[a]	40,347
	RNA genes	4,431
	Pseudogene	232
	Other transcripts	6,559
	Total	51,569

[a]Includes immunoglobulin gene segments.

Characteristics of the RNA-seq gene models

The final RNA-seq gene set comprises 25,748 gene models; 24,088 coding and 1660 noncoding. These models cover 5.0% of the reference sequence, 49.6% on the forward strand and 50.4% on the reverse, with an average length of 2841 bases (maximum 83,104 and minimum seven for noncoding and 71 for coding). The intragenic region span covers 29.2% of the forward strand and 30.1% of the reverse strand, with an average gene-model span of 33,475 bases (maximum 1,199,883 and minimum 67). Gene models have a total of 239,910 exons ranging from one to 205 per model and mean of 9.3 (median 7 and mode 3). The translated portion of the gene models covers 1.9% of the reference sequence and spans 17.8% of the forward and 17.7% of the reverse strand. The RNA-seq

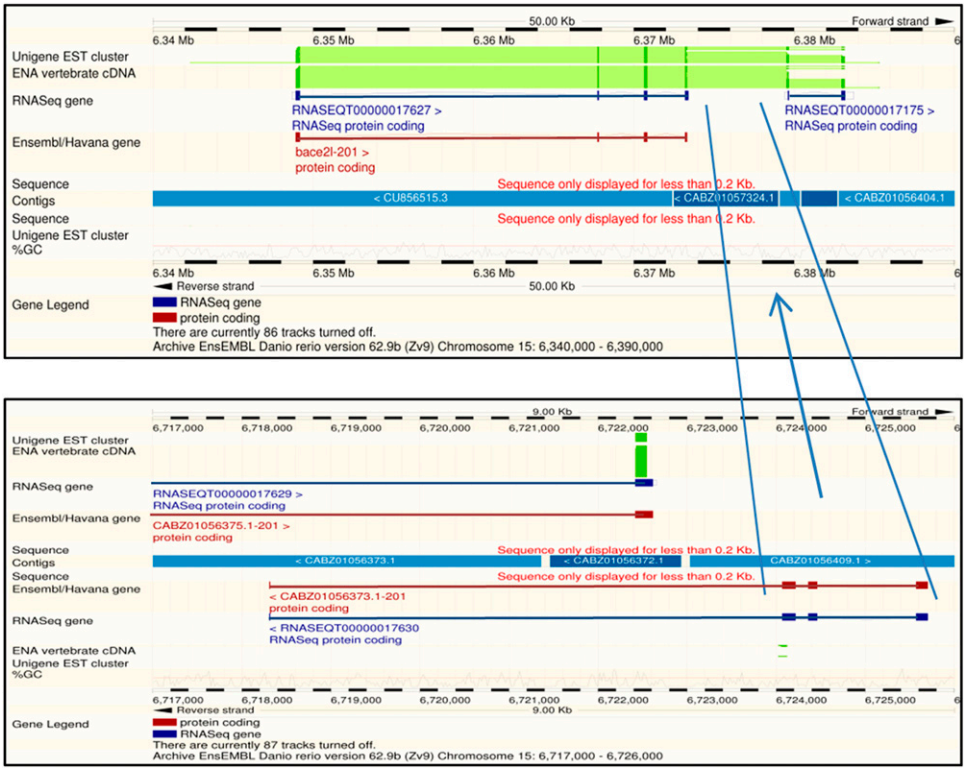

Figure 2. The Ensembl browsers from release 62 between 15: 6,348,671-6,373,413 shows the transcript ENSDART00000065824 for the gene *bace2*. This transcript was annotated from species-specific cDNAs BC164206.1, BC083415.1, and BC098874.1. Matching BC098874.1 to the Zv9 assembly using Ensembl suggests that the 5′ end matches four exons in the reference sequence as annotated in ENSDART00000065824, the central portion matches four exons over 340,000 bases away between 15:6,718,308 and 6,725,702, and the 3′ end matches two exons adjacent to the first four exons between 15:6,379,504 and 6,383,732 (note that the sequence of the short fourth exon of the central portion can be found at the 5′ end of the first exon of the final two exons and is probably an artifact of the read alignment process). This shuffled exon order suggests a reference sequence assembly problem. On closer inspection, apart from the first three exons, the transcript is aligned to capillary and Illumina whole-genome shotgun assemble rather than the more reliable genomic clone sequence. The underlining reference sequence of the central four exons appears to be incorrectly located on chromosome 15. The Ensembl genebuild has create the longest single transcript from the cDNA. Interestingly, the RNA-seq-only genebuild has a model in all three locations. The first four exons match RNASEQT00000017627, the middle three exons match RNASEQT00000017630 creating a partial additional gene, and the final two exons match RNASEQT00000017175. The RNA-seq genebuild is unable to join the first four exons to the last two exons because three exons and four introns are missing. This example demonstrates annotation problems associated with errors in the reference sequence. Expanding this principle to other species, the degree of disrupted transcripts will be related to the quality of the assembly. It also highlights how a break in transcript contiguity can be caused by incomplete exon or intron data that could arise, for example, from low read coverage.

models are displayed in Ensembl with all of the intron data, regardless of whether it was used to build a model. These data indicated alternative splice forms, but were not built into complete models due to lack of contiguity confirmation.

To gauge the quality of the RNA-seq gene models we identified 8822 Ensembl version 60 transcripts supported by publicly available cDNA clone sequences to use for comparison (Supplemental Table 2). The VEGA gene set was not suitable for these analyses as the RNA-seq data had been used as evidence during the manual curation. To successfully build a gene model all of the introns must be identified. To establish how many of the cDNAs had sufficient intron evidence to build a full model, all of the introns were compared with the intron database compiled during the RNA-seq genebuild. Overall, there are 70,022 introns in the cDNA-supported Ensembl models and 66,745 (95%) were found in the RNA-seq intron database. Most of the multi-exon cDNA models have all (79%) or some (19%) of their introns defined (Table 2). Those with some introns tend to have the majority in the intron database (Fig. 3A). These numbers improve if just the protein-coding regions are considered. These data suggest that the information was available to build most of the cDNA models. Closer examination of the introns in the 8822 cDNA-supported Ensembl transcripts showed that 97.4% were canonical and 2.6% noncanonical. However, it was observed that 311 introns under 30 bases had been introduced during the alignments of the cDNA and the majority, 307, were noncanonical. These are almost certainly alignment errors rather than real introns. After removing all of these short introns, the total canonical was 97.8% and noncanonical 2.2%.

To assess the performance of the RNA-seq model assembly algorithm the 8822 cDNA-supported Ensembl transcripts were compared with the RNA-seq models to identify overlaps. The result was a complex pattern of overlapping models where 8152 RNA-seq gene models were found to overlap the cDNAs; one cDNA matched four RNA-seq models, 16 matched three models, 316 matched two models, 7468 matched a single model, and 1021 did not overlap any RNA-seq model. Due to the RNA-seq model pipeline, sometimes by fusing two genes together it is possible to have more than one cDNA model matching a single RNA-seq model. The largest number of overlapping bases was judged the best RNA-seq model match and these model pairs were compared at the nucleotide and

Table 2. Intron, exon, and base coverage of cDNA-supported Ensembl transcripts

	Proportion covered	Matching to cDNA models			Matching to RNA-seq models	
		Intron	Base	Exon	Base	Exon
Single-exon full transcript	Full	—	45	4	1	4
	Partial	—	53	0	97	0
	None	—	220	314	—	94
Single-exon coding only	Full	—	173	142	131[a]	142[a]
	Partial	—	32	0	74[a]	0[a]
	None	—	390	453	—	63[a]
Multi-exon whole transcript	Full	6750	4281	4063	84	4028
	Partial	1591	3422	3536	7619	3571
	None	163	801	905	—	104
Multi-exon coding only	Full	6874	5545	4933	4732	4899
	Partial	1202	1636	2166	2449	2200
	None	151	1046	1128	—	82
Total whole transcript		8504	8822	8822	7801	7801
Total protein coding		8227	8822	8822	7386	7386

[a]Note models can be multi-exon transcript, but single-exon coding.

exon level. If the protein-coding regions of each pair coincided, these were also compared. At the nucleotide level the number of overlapping bases was calculated for each pair, referred to as the matching bases. Similarly, the number of matching exons, defined by identical splice sites, not including transcription start and stop, was calculated. The matching bases or exons were compared with the original cDNA to measure the ability of the RNA-seq to build a model correctly. They were also compared with the RNA-seq models as a measure of the overprediction of the RNA-seq models. Each pair was assigned to full, partial, or no coverage (Table 2) and the matching proportion of each individual pair plotted (Fig. 3B–E). The results show that most of the cDNAs had corresponding models and the majority were covered or almost covered by RNA-seq models at the nucleotide and exon level (Fig. 3B). This improved when only the protein-coding regions were considered (Fig. 3C). Comparing matching exons to the RNA-seq models showed the protein-coding regions correlated better than the whole transcripts. This is even more apparent with the matching nucleotide, suggesting that the untranslated regions are longer in the RNA-seq models at the 5′ end, 3′ end, or both. This could be due to the use of alternative transcription start and end points between the pairs of models, background RNA-seq reads artificially extending the models, or partial cDNA clone sequences. It is worth noting that 425 cDNA supported Ensembl models had no 3′ UTR region when aligned to the genome. Additionally, it was observed that adjacent RNA-seq models that had fused together resulted in additional unmatched untranslated exons and bases.

To investigate the position of the 3′ end of the RNA-seq models further, the length of the 3′ UTR regions of the cDNA supported Ensembl models were compared with the RNA-seq models with respect to 3p marker trimming and extending data. Taking the 8822 cDNA models and their RNA-seq model partners, 5311 passed a series of filters described in Figure 4A. Figure 4A and Table 3 show these data in relation to whether the original RNA-seq model 3′ end was trimmed (shortened by 3p marker data), extended (increased by 3p marker data), confirmed (the 3p marker data identified the same end), or unchanged (there was no appropriate 3p marker data). Overall, 50.6% of RNA-seq model 3′ ends are within ±10 bases of their cDNA model partner's 3′ end and 95% of these were trimmed, confirmed, or extended by 3p marker data. This compares with the other 49.4% of RNA-seq models, with 3′ ends more than ±10 bases from their cDNA partner's 3′ end, where only 48% were altered by the 3p marker data, proving the 3p markers made a big difference in accurately predicting the 3′ end of the RNA-seq models. It is striking that 90% of the RNA-seq models that match their cDNA model partner within ±10 bases were trimmed. Additionally, 77% of RNA-seq models that were left unchanged due to lack of 3p marker data were >10 bases longer than their cDNA model pair. Taken together, these data suggest that the original RNA-seq models were generally too long and trimming to the 3p marker improved the 3′-end prediction. Figure 4B shows the 3303 trimmed RNA-seq models, which fulfilled the filtering, with the amount trimmed from the original models (39%) and the final length of the 3′ UTR. Next, we examined the 486 RNA-seq models from the 5311 pairs where the original 3′ end was extended by 3p marker data. This highlighted two RNA-seq models where the extension had introduced a longer ORF, and these were excluded from this analysis. The length of 3′ UTR was divided into the original UTR length and the extended length. The majority of RNA-seq reads used in the original genebuild were re-mapped to the extended length, without repeat masking, using exonerate. This was split into bases confirmed by RNA-seq data (97.6%) and bases not confirmed (2.4%). The majority of the extended 3′ UTR is supported by RNA-seq reads; approximately one-third have all bases covered, one-third have 10 or fewer bases not confirmed, and one-third have more than 10 bases not confirmed (Fig. 4C) (for clarity only the final third are shown). It is important to remember that the RNA-seq data is not directional, so these calculations could be using reads from the opposite strand. Although no new data was produced to refine the 5′ end of the gene models, their accuracy was assessed by comparing the 3078 cDNA-supported Ensembl models with their RNA-seq-only partners that shared a start codon in the first exon. Figure 4D shows that the RNA-seq models are generally longer than their partners, suggesting that trimming to some experimentally defined 5′ end would be useful. Due to the 5′-end overprediction of the RNA-seq models, no 5′ untranslated regions supported purely by RNA-seq data were carried into the Ensembl genebuild (see below).

Although attempts were made to define single exon RNA-seq gene models, it remained difficult. Introns were the key to identifying gene models relative to mapped background reads, not least to define the strand. Without splice-site information the open reading frame was used to define the strand. The lack of single exon models is apparent in the number of cDNA-supported Ensembl genes overlapping with RNA-seq models (note that there is sometimes only partial overlap between the pair). Single exon models account for 318 out of 8822 cDNA models, and only 98 (31%) had an overlapping RNA-seq model. In contrast, there were 8504 multi-exon models and 7703 (91%) had an overlapping RNA-seq model. To see whether this observation was due to the depth of sequence, 20 lanes representing 81% of the reads were remapped to the genome using BWA (Li and Durbin 2009) and the read coverage for the 8822 cDNA models extracted. The number of reads per base of the gene model was calculated for each cDNA and considered with respect to the presence or absence of a cDNA model (Supplemental

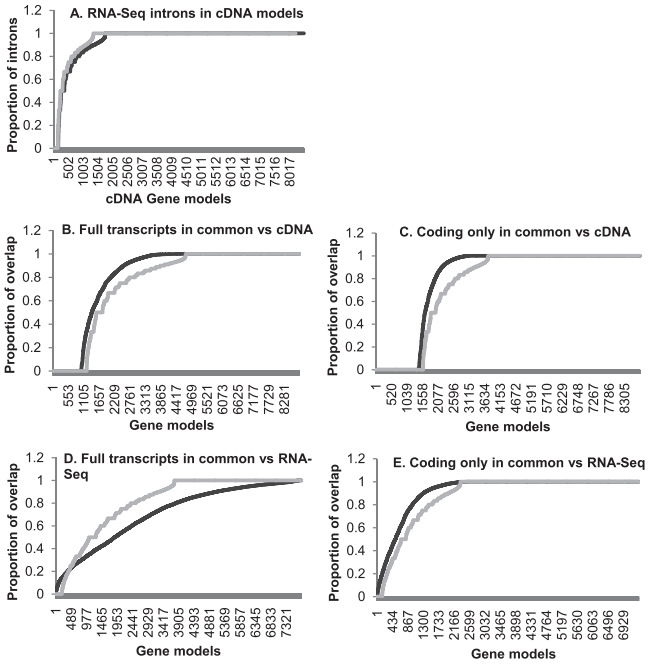

Figure 3. Intron, exon, and base coverage of cDNA-supported Ensembl gene models. The full transcript, and if available, corresponding protein-coding regions of 8822 cDNA-supported Ensembl models were extracted from Ensembl version 60. (*A*) All of the introns from the 8504 multiexon full transcripts were compared with the RNA-seq intron database and scored positive if there was an exact match. The proportion of exact match introns was calculated for each transcript individually and plotted (dark gray). The process was repeated for the protein-coding portion of 8227 multiexon gene models (light gray). (*B*) The 8822 cDNA-supported Ensembl models were compared with the RNA-seq models, and the single best overlap model was identified. The proportion of intersecting nucleotides (dark gray) and exons (light gray) compared with the cDNA-supported models was calculated for each gene model and plotted. (*C*) The same calculation as in *B* using only the protein-coding regions. (*D*) The 7801 cDNA-supported Ensembl models that overlap an RNA-seq model were compared. The proportion of intersecting nucleotides (dark gray) and exons (light gray) compared with the best overlap RNA-seq model was calculated for each gene model and plotted. (*E*) The same calculation as in *D* using 7386 cDNA-supported models where the best match full transcript was from the same RNA-seq model as the best match coding region. Exon matching in *B–E* did not include the transcription start or stop.

Table 3). Unsurprisingly, at the lowest levels of read coverage there are few models for both multiple exon and single exon genes. As the number of reads per base increases, there is little difference in the percentage of multiple exon genes, but the number of single exon genes decreases. This could be due to an increase in complexity or an inability to distinguish reads within these intronless transcripts from the background.

Achieving the best possible set of gene models in an automated genebuild is a balance between including correct models while excluding incorrect models. The aim was to satisfy both of these criteria as far as possible and attain the best solution during the selection/filtering of the RNA-seq models. To assess the number of RNA-seq models missing from the annotation where there was sufficient data to build a model, we again looked at the 8822 cDNA-supported Ensembl models. As discussed above, the RNA-seq genebuild can be split into four stages: defining a discrete set of adjacent exons, identifying the introns, assembling the gene models, and selecting the best set. Of the 8822 cDNA-supported Ensembl models, 1021 failed to overlap an RNA-seq model, 220 of which were single exon and removed from this analysis. To assess why the remaining 801 do not have an overlapping RNA-seq model, the cDNA-supported Ensembl model introns were compared with the database of RNA-seq introns. Results showed that 506 of the cDNA-supported Ensembl models had all introns confirmed by RNA-seq, while 175 had some confirmed. Only 120 had no introns confirmed; therefore, 1.4% of multiexon cDNA-supported Ensembl models could not have been assembled into an RNA-seq gene model due to lack of mapped read information. The lack of 506 gene models with full intron coverage suggests a problem either with the assembly process or the selection/filtering process. The original full set of 412,915 pooled and single-tissue RNA-seq models was examined and only 37 of the cDNA models had no equivalent RNA-seq model suggesting an assembly problem, while 469 had a single or pooled tissue model that had been removed in the selection/filtering process. Future RNA-seq genebuilds can take these data into account to reduce the exclusion of correct models.

To assess the contiguity of the RNA-seq models, the 8822 cDNA-supported Ensembl models derived from a single cloned fragment were compared with their short read-derived RNA-seq partners. There were 7181 multi-exon model pairs with overlapping coding regions. However, these represent a complex mixture of model pairs, some partial, some merged, and some alternate isoforms. These models were further filtered to produce 4628 pairs that shared translation start and stop coordinates. Of these, 4203 (91%) match perfectly and the remainder represent an alternative splice. We also picked 10 RNA-seq transcripts without a public full-length cDNA clone sequence from chromosome 20 (Supplemental Table 4). Sequencing template was amplified from reverse-transcribed zebrafish RNA using two pairs of nested primers designed within the first and last exon. Amplified fragments were directly capillary sequenced using predesigned internal primers. Reads for nine genes were assembled into a consensus sequence, with at least one read covering all bases between the primers, and matched the original RNA-seq transcript. One amplified fragment, from RNASEQT00000015341 (Zv9 20:19438035 to 19349092), gave poor quality sequence around exon 9, suggesting multiple transcripts. The fragment was cloned into a plasmid library and the individual clones sequenced. The assembled consensus sequences showed three possible transcripts; the original RNA-seq model sequence, a missing 38-base exon 9, or an alternative splice site changing the 38-base exon 9 into a 29-base exon. All three splice forms have supporting RNA-seq introns. One clone showed a mix of sequence with and without the 38-base exon and was rejected as experimental artifact.

Ensembl/VEGA/RNA-seq gene models merge

The RNA-seq-only genebuild described above does not take the wealth of publicly available gene data into account. Importantly, there is no confirmation of the RNA-seq models' predicted open reading frames using a protein database such as UniProt (The UniProt Consortium 2010). Before Ensembl version 60, the zebrafish Ensembl genebuilds have comprised a merge of two data sets, zebrafish-specific and orthologous sequences mostly comprising cDNA sequence translations. Here we describe the incorporation of the RNA-seq gene models and from Ensembl version 61 the manually curated VEGA database (Wilming et al. 2008) as additional data sources. In the Ensembl/VEGA/RNA-seq-merged genebuild a single gene can comprise one or more transcripts, where each individual transcript comes from one of the four sources; zebrafish-

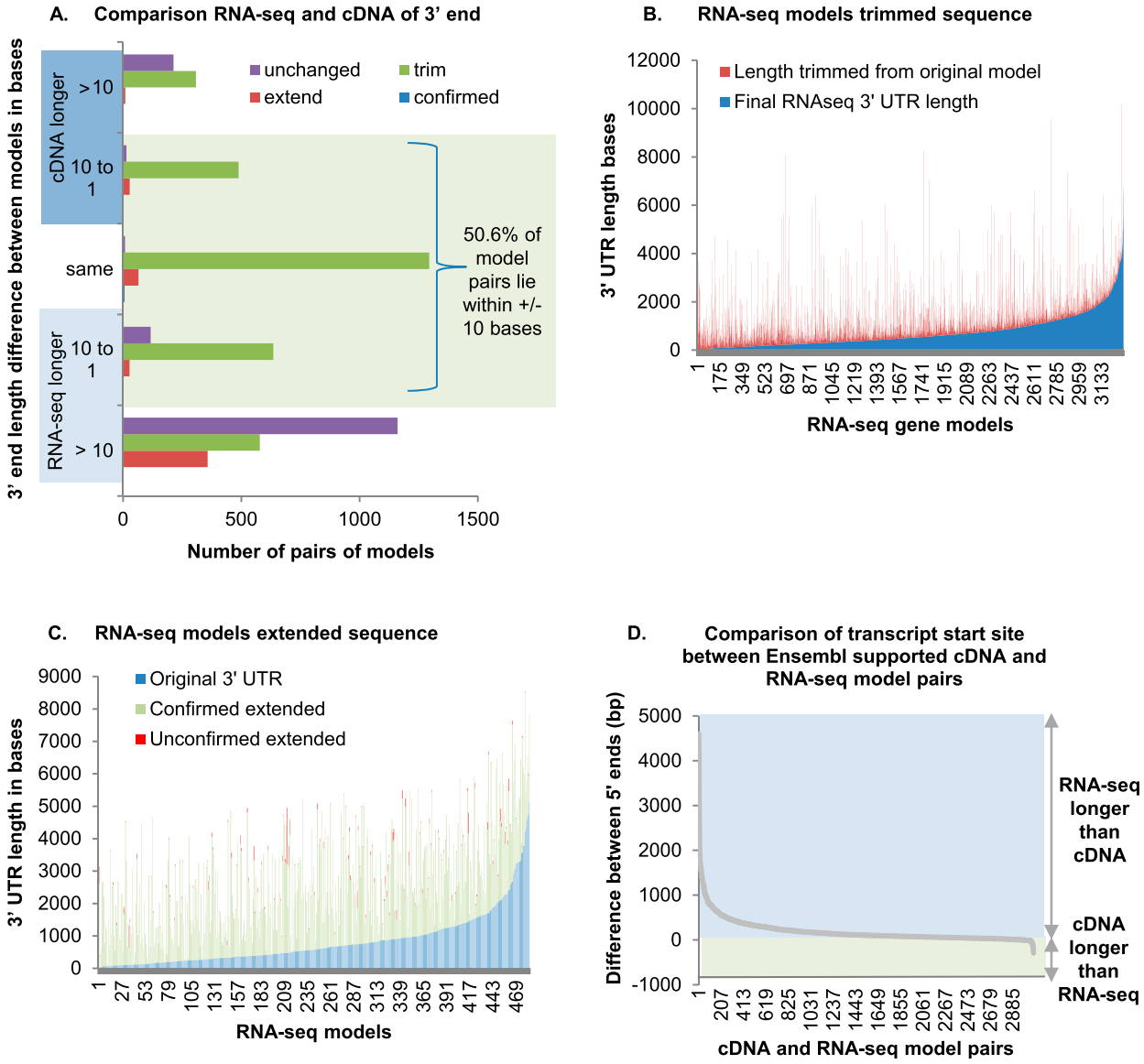

Figure 4. Gene model end prediction. (*A*) The 3′ end of RNA-seq models from the original genebuild were compared with 3p marker data and scored as trimmed, extended, confirmed (if identical), or left unchanged. The 3p altered RNA-seq models were matched to their cDNA-supported Ensembl model pair, and 5311 passed a series of filters, including pairs where the best match whole transcript coincided with the best match coding region, both ORFs stopped at the same genomic coordinate, and there were no introns in either of the 3′ UTR regions, thus excluding fused gene models. The 3′ UTR length was compared; in 1063 cases the cDNA model was found to be longer, in 2875 cases the RNA-seq model was longer, and in 1373 cases they were exactly the same. The length of the 3′ UTR of model pairs was compared, the difference in length calculated, and shown on the *y*-axis. Model pairs where the RNA-seq transcript is longer are in the *bottom* light-blue section, the pairs where the cDNA-supported Ensembl transcript is longer are in the *top* dark-blue section, and the pairs with identical length are in the *middle*. The model pairs with the specified length difference are shown for all four possible 3p marker alterations (trimmed in green, extended in red, confirmed in blue, and unchanged in purple). The length of the bar indicates the number of model pairs with the indicated length difference after the 3p marker alteration performed. (*B*) The 3289 trimmed RNA-seq-only models filtered as above are shown. The blue bars show the 3′ UTR length after trimming, the red bars show the length trimmed, and the blue plus red bars show the original 3′ UTR length. (*C*) Filtered RNA-seq models extended by 3p marker data are shown. The blue bars show the 3′ UTR length after the original RNA-seq genebuild before extension, while the green plus red bars show the length of extension. The green bars show the total number of bases covered by RNA-seq reads and the red bars the total number of bases not covered, these are not necessarily consecutive. For clarity, only the models with 10 or more bases not confirmed by RNA-seq sequence are shown. Note that the RNA-seq libraries are not directional, so it is possible that the reads used for this confirmation are on the opposite strand. (*D*) The 3078 cDNA-supported Ensembl transcripts that share a start codon within the first exon with an RNA-seq-only model were compared. The *y*-axis shows the difference in length between model pairs, with the green regions indicating where the cDNA-supported Ensembl transcript is longer and the blue region indicating where the RNA-seq-only model is longer.

specific sequence, orthologous sequence, RNA-seq gene models, or the manually curated VEGA database. The rationale was to incorporate the RNA-seq genebuild into a standard Ensembl genebuild by making use of the RNA-seq intron data and gene models. The emphasis on open reading frame alignment in the Ensembl genebuild provides evidence of gene model contiguity as well as the potential to resolve merged neighboring RNA-seq models. Conversely, the RNA-seq introns provide experimental support for

Table 3. 3′-end prediction

	RNA-seq longer than cDNA			cDNA longer than RNA-seq		
	> 10	10 to 1	Same	1 to 10	> 10	Total
Confirmed	0	2	6	2	1	11
Extended	357	27	65	28	9	486
Trimmed	578	635	1294	488	308	3303
Unchanged	1160	116	8	14	213	1511
Total	2095	780	1373	532	531	5311

the discrimination of ambiguous splice sites, evidence of transcription, and evidence for the untranslated regions for orthologous Ensembl models.

A standard Ensembl genebuild predicts gene structures by aligning species-specific and orthologous sequences. The RNA-seq genebuild was incorporated in two ways. Firstly, transcripts built from orthologous proteins were filtered to identify the models with the most supporting evidence. This usually involves cDNA and EST data, but here also included RNA-seq intron data. Secondly, whole RNA-seq gene models were added into the build, but only if the longest open reading frame matched a UniProt protein with evidence of transcription (PE 1 and PE 2). The zebrafish cDNAs best-supported orthologous model and filtered RNA-seq models were compared. RNA-seq models that provided additional information, such as increasing the transcript length, were identified. Where an RNA-seq model extended a model built from a zebrafish cDNA sequence, the RNA-seq model was only included, in addition, if the genomic span of the coding exons and the coding region itself were longer. Novel UniProt filtered RNA-seq models were also retained. The RNA-seq models remaining after the comparison were extended or trimmed using 3p maker data or the 3′ UTR removed if none existed due to ambiguity in the transcription end site as discussed above. For the same reason, the 5′ untranslated regions of the RNA-seq models were removed. All of the models built from a zebrafish cDNA sequence, the best-supported orthologous models and the remaining RNA-seq models, entered the Ensembl GeneBuilder (Curwen et al. 2004). This collapsed multiple transcripts into one gene identifier if they shared coding exons resulting in the Ensembl genebuild incorporating RNA-seq (Table 1). Each gene can contain transcripts from one or more sources as detailed in Figure 5. The final step in the process was the merge with the manually curated VEGA Gene Models (VEGA releases 40) and is detailed in Table 1.

Ensembl gene models predicted from orthologous sequence can encounter problems in pinpointing the intron/exon boundaries. As described above, the RNA-seq models were built using a database of introns defined by individual reads mapping to adjacent exons. These introns were used as experimentally derived evidence for selecting the best-supported splice sites. To assess the contribution of RNA-seq evidence to the Ensembl version 62 Ensembl/VEGA/RNA-seq-merge genebuild we compared all of the introns from 26,427 multiexon genes to the RNA-seq intron database. Of the 244,894 introns in the Ensembl/VEGA-merge genebuild, 203,276 (83%) were found in the RNA-seq intron database, suggesting that the majority of introns have experimental support. The majority (82.1%) are canonical and 0.9% are noncanonical. Looking at the intron types further, there is a striking difference with 89% of canonical introns confirmed, but only 11% of the noncanonical. The lack of support for noncanonical introns may be due to the introduction of short introns during the alignment of the Ensembl models as discussed above. The presence of at least one confirmed intron in 87% of multiexon Ensembl gene models provides evidence for transcription that is particularly important for any model built using orthology.

Discussion

We present a pipeline for annotating polyA transcripts on a genomic reference sequence using RNA-seq data. Models were initially predicted using RNA-seq data alone, with no ab initio prediction, and then combined with the Ensembl and VEGA databases. We have demonstrated some limitations with using only RNA-seq data for gene annotation, but have also shown how elements of these experimentally derived data can complement the Ensembl/VEGA genebuild. The RNA-seq-only gene annotation represents only part of the data in the sequence files. Extraction of the relevant reads to assemble transcripts is a fine balance between identifying useful data and discarding artifacts. In addition, as with any system where the aim is to collect an entire set, it is easy to collect individual elements at the beginning, but gets hard as you near completion. With the RNA-seq gene set we are restricted by spatial and temporal expression of rare transcripts, particularly if there are few copies per cell. However, the proportion of RNA-seq models we were able to build suggests that we are well on the way to a complete gene set.

The RNA-seq-only genebuild predicted 25,748 gene models, most of which were classed as protein coding. The quality of these models was assessed by comparing 8822 cDNA-supported Ensembl genes to their RNA-seq model partners. Although this generally showed a good correlation between the gene models, it highlighted a number of areas where we can improve the pipeline. For the most part, the RNA-seq-only pipeline relies on calling an approximate exon sequence with read pairs linking adjacent exons. The introns are then called from this reduced portion of the genome. Inevitability, this leads to the loss of intron information outside of these defined regions. Our new algorithms examine the entire genomic span of the approximate models at the intron alignment step and therefore have the opportunity to detect all introns. Another problem was spotted in the analysis of the missing RNA-seq models. We found that intron data were available to build most of the missing models and that we had, in fact, built the majority; however, we were losing them during the final filtering process. Close inspection of the rejected models should help us refine the filtering algorithm in an attempt to minimize the exclusion of genuine models, while not including artifacts. On a

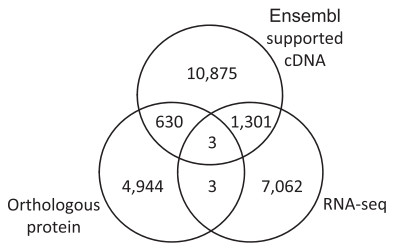

Figure 5. The transcript composition of Ensembl genes. The GeneBuilder collapses multiple transcripts into single gene identifiers. The transcripts can be derived from one of three sources. The Venn diagram indicates the number of Ensembl genes comprised of transcripts from the three different sources.

similar note, attempts were made to filter RNA-seq models that concatenated neighboring genes. Ultimately, this problem was largely solved by the merge with the Ensembl gene models and their intrinsic open reading frame data. In a situation where an RNA-seq genebuild was independent of an Ensembl genebuild, the incorporation of proteome information, similar to Surget-Groba and Montoya-Burgos (2010), would be sensible.

The Ensembl/VEGA and RNA-seq genebuilds brought different attributes to the final gene set with the merge providing the most comprehensive annotation. We display the RNA-seq-only annotation with all of the intron data indicating alternative splicing in Ensembl to show the information we have drawn from the Illumina transcriptome sequencing in addition to the full genebuild that puts a greater emphasis on false-positive filtering. The intron data derived from the RNA-seq models can play a vital role in fine tuning the exon/intron boundaries of the models predicted by orthology. In contrast, the orthologous proteins supply evidence of transcript contiguity not provided by short Illumina reads. The RNA-seq contributes evidence for the untranslated regions of orthologous models; however, these data need to be treated with caution. The 3p marker data has successfully predicted the end of transcription for many transcripts, and data on more loci would be useful. A similar 5′ capture method would improve the accuracy of the 5′ end. Strand-specific libraries would help clarify the gene-model span, particularly if genes overlap on opposite strands. Several other groups have published genome-guided gene annotation pipelines using RNA-seq (Denoeud et al. 2008; Yassour et al. 2009; Guttman et al. 2010; Trapnell et al. 2010). These tend to take user RNA-seq data and create gene models. Our pipeline initially constructs RNA-seq models, and then uses these data to complement the comprehensive Ensembl genebuild pipeline. This is a complex process, and to allow the pipeline to be run externally we have initiated a project to allow users to run our RNA-seq pipeline using cloud computing.

RNA-seq is a cost effective method for producing species-specific cDNA sequence as compared with cDNA cloning and sequencing. The RNA-seq data provides untranslated regions, splice-site information for models predicted by othology, and proof of transcription. Additionally, the deep splice-site prediction data give valuable alternative splice information. The RNA-seq pipeline has been used to integrate RNA-seq data into the Tasmanian devil Ensembl genebuilds (Murchison et al. 2012). Here, in a species with little cDNA sequence, the RNA-seq intron data were particularly useful during the orthologous model filtering. This pipeline can be used for any eukaryote gene annotation and will continue to form part of the Ensembl vertebrate genebuild.

Methods

Transcriptome sequencing

Breeding zebrafish (*Danio rerio*) were maintained at 28°C on a 14-h light/10-h dark cycle. Fertilized eggs were obtained from natural spawning, grown in incubators at 28°C, and snap frozen in dry ice at the required developmental stage. Adult fish were dissected and the tissues snap frozen in dry ice. The tissue was lysed in TRIzol (Invitrogen) and processed in Phase Lock Gel tubes (Eppendorf) according to the manufacturer's instructions. PolyA+ RNA was extracted from total RNA on Dynabeads (Invitrogen) according to the manufacturer's instructions. However, to increase the yield, the same beads were eluted three times. The RNA was DNase treated according to the manufacturer's instructions (Ambion1906), followed by a second polyA+ pull down with three rounds of elution. RNA was quantified on a NanoDrop spectrophotometer. Illumina libraries were prepared by four protocols. In library protocol 1, polyA+ RNA was reverse transcribed using Oligo$(dT)_{12\text{-}18}$ primer (Invitrogen), and the second strand synthesized using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) with 1 μL of RNase Inhibitor (NEB). DNA was nebulized at 35 psi for 6 min on ice, and the fragments recovered using the QIAgen PCR purification Kit. The fragmented cDNA was made into standard Illumina libraries from the end-repair step. Library protocol 2 was the same as library protocol 1, except the first strand was primed with random primers (Invitrogen). Protocol 3 is similar to the standard RNA-seq Illumina library protocol. PolyA+ RNA was initially chemically hydrolyzed with RNA Fragmental Reagents (Ambion) for 1 min at 70°C as recommended by the manufacturer. RNA was precipitated with 2 μL of glycogen (5 μg/μL, Ambion) in LiCl and ethanol. Double-stranded cDNA was synthesized using random primers as described above without DNA ligase in the second-strand synthesis, and Illumina libraries constructed by the standard protocol from the end-repair step. Protocol 4 was the standard RNA-seq Illumina protocol, starting with total RNA but with a DNase step (Ambion) between the two rounds of polyA pull down. Libraries were loaded onto lanes of an Illumina Genome Analyzer flowcell, where cluster formation, primer hybridization, and sequencing reactions were carried out according to the manufacturer's instructions (Bentley et al. 2008).

Read alignment

The module Bio::EnsEMBL::Analysis::RunnableDB::ExonerateSolexa (revision 1.17) was used to do genomic alignment using exonerate v0.9 (Slater and Birney 2005) against RepeatMasked genomic sequence. Reads were aligned using the following alignment parameters: --model affine:local --forwardcoordinates FALSE --softmasktarget TRUE --exhaustive FALSE --saturatethreshold 100 --dnahspthreshold 60 --dnawordlen 14 --bestn 10.

A score parameter was also included that reflected the number of desired exact matches; this varied between analyses with the length of the reads. The "bestn 10" causes exonerate to return the 10 highest-scoring alignments. As the reads were paired, matching pairs were always stored together in the same fasta file, both pairs would then be aligned within a single job, all of the possible alignments for each read were considered, and the alignment pairing that provided the single highest overall alignment score while being consistent with pairing rules was chosen (pairs should align on the same sequence within 200 kb of each other on opposite strands and in opposite orientation). If more than one pair of alignments had equally good scores, the pair with the smallest distance between them was chosen; this was an attempt to prevent cross-pairing between clustered members of gene families.

Intron alignment was run using Bio::EnsEMBL::Analysis::RunnableDB::ExonerateSolexaTranscript (rev1.10) and the exonerate v0.9 est2genome splice model to align individual reads to transcript sequences. The following parameters were used: --model est2genome --forwardcoordinates FALSE --softmasktarget TRUE --exhaustive FALSE --saturatethreshold 100 --dnahspthreshold 60 --minintron 20 --maxintron 20000 --dnawordlen 14 -i -12 --bestn 1.

Again, a score parameter was chosen that reflected read length. A maximum intron value was used to prevent very long splices from occurring within long transcripts. Only the single highest-scoring alignment is returned in this case (bestn 1).

RNA-seq model prediction

The rough gene code Bio::EnsEMBL::Analysis::RunnableDB::Solexa2Genes (rev1.10) was used to make the rough models.

Rough models were constructed by collapsing overlapping reads into nonoverlapping blocks roughly corresponding to exons, these were then linked into extended transcript-like structures using read pairing. The models were filtered to remove those shorter than 100 bp along with models where the genomic span of the gene was <1.5 times the cDNA length. Exons were removed where the read coverage was <0.1% of the average exon coverage. Single exon models were removed unless they were longer than 1000 bp.

The gene refinement code Bio::EnsEMBL::Analysis::RunnableDB::RefineSolexaGenes (rev 1.12) was used to refine the exon/intron splice sites. Exons and intron features were combined in all possible combinations, and the resulting splice graphs scored according to the depth of read support. Graphs were arranged by score and the highest scoring transcript for each loci selected. A maximum of 100,000 iterations was used for the processing of the splice graph; if the model required more than this to be fully processed, then the model was simplified by removing the least common splices and the process repeated until a final model could be constructed within the 100,000 iterations. A transcript score penalty of 2 was awarded for removing a retained intron. A minimum intron size of 30 was used. Single exon models were allowed only if they had a cDNA length of >1000 bp, of which at least 66% had to be CDS. Where exons had multiple possible splice sites, the one with the most read support was used. A maximum of 1000 highest-scoring splice graphs were built into transcript structures; translations were computed to identify the longest open reading frame and if a translation could not be found, then the transcript was designated as "non_coding". Scores were weighted to favor longer ORFs. Exon scores from coding exons contributed 90% to the final score, the remaining 10% was comprised of the scores from all of the exons and introns combined. Finally, the complete set of filtered models were reclustered to identify models with coding regions overlap on opposite strands; in these cases the lowest scoring models were removed.

The initial 412,915 models were further filtered. The following rules were applied to exclude: models with a CDS <100 bp and two exons, models with two exons and a noncanonical splice, models with coding exon overlap with a higher or equal scoring model on the opposite strand, and models that are an opposite-strand subsequence of another model. Models built using the pooled set of reads were considered for replacement by one of the single tissue models if the following criteria were met: the model has a single coding exon and at least three exons in total, the model has one coding exon and the length of the CDS represents ≤10% of the total model length, the model has multiple coding exons, and at least two noncoding exons and <10 exons in total, or the model has at least 10 exons and the ratio of coding to noncoding exons is <80%. Single-tissue models replaced models marked for replacement if the following criteria were met: More than one single tissue transcript cluster has exon overlap with the marked model, the highest-scoring models from each of the single tissue clusters have a longer coding sequence and longer genomic span of CDS than the marked model.

3′-End prediction and modification

Total RNA from four developmental stages and three adult tissues was extracted using TRIzol Reagent (Invitrogen). Total RNA was chemically fragmented using DNA Fragmentation Reagents (Ambion) for 5 min at 70°C in 10 μL, LiCl precipitated in the presence of glycogen and resuspended in water. At room temperature, 1 μL of BPM1polyT22 primer (100 μM) (Biotin-GGCCAG TCCTGGAGTTTTTTTTTTTTTTTTTTTTTTVN) with a 3′ anchor sequence (Thomas et al. 1993) was annealed to 200 μg of streptavidin magnetic beads (NEB), and the beads were washed in binding buffer (0.5 M NaCl, 20 mM Tris-HCl at pH 7.5, 1 mM EDTA). RNA in 100 μL of 1× binding buffer was annealed to the primer for 30 min at room temperature, washed three times in 1× binding buffer and once in low-salt buffer (0.15 M NaCl, 20 mM Tris-HCl at pH 7.5, 1 mM EDTA). Reverse transcription was performed with SuperScript II according to the manufacturer's instructions (Invitrogen) with the addition of RNase inhibitor (NEB). The second strand was synthesized with Second Strand Buffer (Invitrogen) according to the manufacturer's instructions with the substitution of *E. coli* DNA polymerase (Promega) and RNaseH (NEB) with no DNA ligase. The beads were washed twice in binding buffer and once in low-salt buffer. The DNA fragments were excised using 5 units of BpmI (NEB) in a 100-μL reaction incubated for 90 min at 37°C. After purification with a QIAgen PCR clean-up column, the DNA fragments were made into a standard Illumina library using the manufacturer's protocol. Paired end sequencing of 76 bases each was performed on a Genome Analyzer II (Illumina).

The 3′ ends of the models were altered to reflect the most common positions detected by the 3′ pull-down experiments. 3′ reads were aligned to the genome using the ExonerateSolexa module in the standard manner and were then clustered. Read pairs were rejected unless they had one read with a minimum aligned length of 37 bp and a pair with an aligned length of 69, 70, or 71 bp. Clusters were assigned a strand according to the orientation of the pair. In order for a transcript to be considered for 3′ modification it had to contain at least three exons and a coding sequence longer than 100 amino acids with a stop codon. 3′ markers were only considered if the following was true: the marker position was downstream from the stop codon, the marker was <5000 bp from the end of the transcript, the potential extension from the end of the 3′ exon to the marker did not overlap any other transcripts on either strand, the genomic sequence 10 bp 3′ of the marker had <60% A bases and started with a maximum of three As in a row, the marker was represented by at least three reads, the reads pairs were not duplicated, the sequence "CTGGAG" was not found in the 10-bp starting 15 bp downstream from the marker or the sequence "CTCCAG" was not found in the 9-bp starting 13 bp upstream of the marker, and the 10 bases 3′ of the end of transcription are not one of the following 10-mers—AAABAAABBB, AAABAABABB, AAABABAABB, AABAAAABBB, AABAAABABB, AABABAAABB, ABAAAAABBB, ABAAAABABB, ABAAABAABB, ABAABAAABB, and ABABAAAABB, where B is C, G, or T. If all of these criteria were met, the highest-scoring marker was used to modify the 3′ end of the transcript, and any other markers passing the criteria were stored as transcript attributes.

Filtering orthologous aligned genes

The module Bio::EnsEMBL::Analysis::RunnableDB::TranscriptConsensus (rev 1.4) was used to filter aligned UniProt proteins from other species. cDNA, EST, and RNA-seq evidence was compared with the transcript clusters and transcripts scored according to the amount of supporting evidence. RNA-seq intron spanning alignments from 76-bp reads were used where the exonerate alignment score was at least 150 (30 exact matches). Each exon and intron was scored according to the number of UniProt proteins, cDNAs, ESTs, and reads that exactly matched at the boundaries. Penalties were awarded for each overlap of an EST, cDNA, or read where boundaries were not shared, poorly supported end exons and introns, and exons shorter than 10 bp. Transcripts were sorted according to score, and all but the highest scoring transcript(s) were filtered out.

Gene model contiguity confirmation

Total RNA from the male head and 3- or 5-d post-fertilization embryos were extracted using TRIzol Reagent (Invitrogen). Total

RNA was DNase treated using the DNA-free Kit (Ambion) following the manufacturer's instructions. The resulting RNA was split into two for reverse transcription using a polyT primer (Invitrogen) and SuperScript II (Invitrogen) following the manufacturer's instructions and including RNase Inhibitor (NEB) with only one tube of each pair containing the enzyme. Reactions were purified using the QIAGEN PCR Purification Kit and eluted in 50 μL of EB buffer. Each of the 10 gene models were amplified in two rounds of PCR using KOD Hot Start DNA Polymerase (Novagen) in a 25-μL reaction containing 1× buffer for KOD Hot Start DNA Polymerase, 0.2 mM dNTPs, 1 mM $MgSO_4$, 0.4 μM of each primer, 0.5 units of enzyme, and sterile water. Reactions were placed in a PTC-225 thermo-cycler (MJ Research) preheated to 94°C for 2 min, then cycled 35 times at 94°C for 15 sec, 60°C for 30 sec, and 68°C for 5 min, finishing with 68°C for 5 min. The first round comprised three samples with 1 μL of template; reverse transcribed with enzyme, reverse transcribed without enzyme, and no template. All first-round tubes were diluted 1:50, and 1 μL used in a second round comprising four tubes; reverse transcribed with enzyme, reverse transcribed without enzyme, no template from the first round and a second round only no template. To obtain sufficient DNA for sequencing, two or four second round reverse transcribed with enzyme 25-μL reactions were performed in parallel and pooled. Amplified fragments were sequenced using primers designed from the predicted sequence using BigDye Terminator v3.1 Cycle Sequencing Kit (ABI), and reads were assembled in GAP4. On sequence analysis, one model, RNASEQT00000015341, suggested the presence of multiple transcripts in the amplified fragment and was cloned using the Zero Blunt PCR Cloning Kit (Invitrogen) with the Quick Ligation Kit (NEB), transformed into TOP10 competent cells (Invitrogen) and clones were sequenced.

Merging RNA-seq into the Ensembl Genebuild

Prefiltered RNA-seq models were clustered with zebrafish-specific transcripts. RNA-seq models that had coding overlap with zebrafish transcripts on the opposite strand were excluded. The remaining RNA-seq models were clustered with prefiltered orthologous aligned transcripts and zebrafish-specific transcripts. RNA-seq models were only retained where they had a longer ORF and a longer genomic span of the ORF than the matching transcript. In cases where the RNA-seq model matched a zebrafish-specific transcript, the RNA-seq was also required to have an ORF BLAST hit to UniProt with at least 70% coverage; for orthologous transcripts the requirement was for at least 50% coverage (UniProt release 2010_09). UniProt WUBLAST comparison was performed using the module Bio::EnsEMBL::Analysis::RunnableDB::BlastRNASeqPep (rev 1.1). WUBLAST alignments were performed against a UniProt BLAST database filtered to retain only models with protein existence levels 1 and 2. The following parameters were used: -gi -cpus=1. Results were filtered to remove models with a p_value threshold of more than 0.01 and a minimum score of <200. More details can be found at http://www.ensembl.org/Danio_rerio/Info/Index.

VEGA database merge

In Ensembl versions 61 and 62 the Ensembl genebuild was merged with VEGA release 40. Transcripts were merged if the internal exon/intron structure was identical. They were also merged if the transcription start and end site varied, only if the exon/intron structure within the coding region was identical, but the VEGA transcript untranslated region took priority. Nonmerged transcripts were retained unchanged in the final gene set. The Ensembl-VEGA merging algorithm considered the merged model biotype from both annotation sources with the VEGA biotype taking precedence over Ensembl. Where necessary, the Ensembl model biotypes was changed in line with VEGA, the change reported to the manual annotation team for investigation, and when resolved could improve future gene sets. If the Ensembl model was changed to a noncoding biotype the translation was removed. The supporting protein and cDNA evidence were associated with the transcripts.

Mapping reads to Ensembl transcripts

Reads from 20 lanes (see Supplemental Table 1) were separately aligned to the Zv9 reference sequence using default parameters on BWA version 0.5.9-r16 (Li and Durbin 2009). The alignments were paired using BWA sampe with the options –A –a 200000 to set the maximum pairing distance. Using samtools 0.1.16 (r963:234) (Li et al. 2009) with default parameters, samtools import to make BAM files, then samtools sort, and samtools index to make sorted indices. These BAM files were merged using samtools merge. Bio-SamTools was used to identify reads in the BAM files that aligned to the 8822 cDNA supported transcripts (http://search.cpan.org/~lds/Bio-SamTools/lib/Bio/DB/Sam.pm).

Data analysis

Protein sequences were obtained from Refseq release42 and Uniprot release 2010_09, the protein sequences were filtered and merged into a single proteome file using ensembl-analysis/scripts/genebuild/prepare_proteome.pl script (v1.7). cDNA and EST sequences were obtained from EMBL release 104 and GenBank release gb178, these sequences were processed to remove polyA sequences using the script ensembl-analysis/scripts/genebuild/prepare_cdnas.pl (v1.6). Custom scripts were used to process the gene model data that is available in Supplemental Table 2. Data processing, tables, and graphs were prepared using Microsoft Office Excel 2007.

Data access

Sequence data were submitted to the European Nucleotide Archive as ERP000016, ERP000263, and ERP000400. The Ensembl Perl API (http://www.ensembl.org/info/docs/api/api_installation.html) can be downloaded from the Ensembl FTP site ftp://ftp.ensembl.org/ and is also available through the public CVS server http://www.ensembl.org/info/docs/api/api_cvs.html. Both of the genebuilds can be viewed via the Ensembl browser version 62 http://apr2011.archive.ensembl.org/Danio_rerio/Info/Index. The Ensembl genebuild is available via BioMart in Ensembl version 62 or the API. The RNA-seq builds can be downloaded via the API by connecting to the otherfeatures database for *Danio rerio* either through ensembldb.ensembl.org (see http://www.ensembl.org/info/docs/api/core/core_tutorial.html#install) or by installing a local MySQL database by downloading the MySQL dump of the other features database from the Ensembl FTP site ftp://ftp.ensembl.org/pub/release-62/mysql/danio_rerio_otherfeatures_62_9b/.

Acknowledgments

We thank the staff in the Wellcome Trust Sanger Institute Illumina sequencing pipeline for their assistance. This research was supported by Wellcome Trust grant number 098051. We thank Ian Sealy for critically reading the manuscript.

References

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole

human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.
Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5:** 613–619.
Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14:** 942–950.
Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, et al. 2008. Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9:** R175. doi: 10.1186/gb-2008-9-12-r175.
Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2011. Ensembl 2011. *Nucleic Acids Res* **39:** D800–D806.
Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.
Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V et al. 2010. The landscape of *C. elegans* 3′UTRs. *Science* **329:** 432–435.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Murchison EP, Schulz-Trieglaff OB, Ning Z, Alexandrov LB, Bauer MJ, Fu B, Hims M, Ding Z, Ivakhno S, Stewart C, et al. 2012. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148:** 780–791.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143:** 1018–1029.
Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22:** 577–591.
Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res* **37:** D32–D36.
Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al. 2010. *De novo* assembly and analysis of RNA-seq data. *Nat Methods* **7:** 909–912.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.
Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321:** 956–960.
Surget-Groba Y, Montoya-Burgos JI. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res* **20:** 1432–1440.
Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, Robinson C, Mandich A,Derge JG, Lewis J, Shoaf D, et al. 2009. The completion of the Mammalian Gene Collection (MGC). *Genome Res* **19:** 2324–2333.
Thomas MG, Hesse SA, McKie AT, Farzaneh F. 1993. Sequencing of cDNA using anchored oligo dT primers. *Nucleic Acids Res* **21:** 3915–3916.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147:** 1537–1550.
The Uniprot Consortium. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38:** D142–D148.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.
Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.
Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453:** 1239–1243.
Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36:** D753–D760.
Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, et al. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci* **106:** 3264–3269.
Yoon OK, Brem RB. 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA* **16:** 1256–1267.
Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Received January 24, 2012; accepted in revised form July 10, 2012.