BMC
Genomics

Open Access

Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*

Selene L Fernandez-Valverde, Andrew D Calcino and Bernard M Degnan[*]

Abstract

Background: The demosponge *Amphimedon queenslandica* is amongst the few early-branching metazoans with an assembled and annotated draft genome, making it an important species in the study of the origin and early evolution of animals. Current gene models in this species are largely based on *in silico* predictions and low coverage expressed sequence tag (EST) evidence.

Results: *Amphimedon queenslandica* protein-coding gene models are improved using deep RNA-Seq data from four developmental stages and CEL-Seq data from 82 developmental samples. Over 86% of previously predicted genes are retained in the new gene models, although 24% have additional exons; there is also a marked increase in the total number of annotated 3′ and 5′ untranslated regions (UTRs). Importantly, these new developmental transcriptome data reveal numerous previously unannotated protein-coding genes in the *Amphimedon* genome, increasing the total gene number by 25%, from 30,060 to 40,122. In general, *Amphimedon* genes have introns that are markedly smaller than those in other animals and most of the alternatively spliced genes in *Amphimedon* undergo intron-retention; exon-skipping is the least common mode of alternative splicing. Finally, in addition to canonical polyadenylation signal sequences, *Amphimedon* genes are enriched in a number of unique AT-rich motifs in their 3′ UTRs.

Conclusions: The inclusion of developmental transcriptome data has substantially improved the structure and composition of protein-coding gene models in *Amphimedon queenslandica*, providing a more accurate and comprehensive set of genes for functional and comparative studies. These improvements reveal the *Amphimedon* genome is comprised of a remarkably high number of tightly packed genes. These genes have small introns and there is pervasive intron retention amongst alternatively spliced transcripts. These aspects of the sponge genome are more similar unicellular opisthokont genomes than to other animal genomes.

Keywords: Transcriptome, Transcription termination, Alternative splicing, Metazoan evolution

Background

The origin of the fundamental rules governing metazoan multicellularity and morphological complexity can be gleaned through the analysis of the genomes of early branching animals (e.g. sponges, cnidarians, ctenophores and placozoans) [1-4] and their closely related unicellular holozoans (e.g. choanoflagellates and filastereans) [5-7]. Comparative analysis of these genomes has shed light into the evolution of protein-coding gene families. For instance, transcription factor and signalling pathway gene families that are essential to the development of complex bilaterians (e.g. vertebrates, insects, worms and their allies) largely evolved in the Precambrian, before the lineage leading to these animals diverged from early branching animal phyla [1,8-11].

Obtaining a more complete picture of the origin and early evolution of metazoan multicellularity and development also requires the analysis of the mechanisms that regulate gene expression. This demands (i) a more

* Correspondence: b.degnan@uq.edu.au
Centre for Marine Sciences, School of Biological Sciences, The University of Queensland, Brisbane 4072, Australia

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 2 of 11

precise view of genome organisation and composition, and gene structure, (ii) detailed expression profiles from multiple cell types, and developmental and physiological contexts, and (iii) the capacity to experimentally manipulate gene function. Thus, increasing the accuracy and completeness of the draft genomes of early branching metazoans is an important step in improving their utility for future evolutionary and functional studies aimed at unravelling the origin of animal multicellularity.

The genome of the demosponge *Amphimedon queenslandica* was published in 2010 [1] and is currently the only published genome from phylum Porifera. The sponge body plan is amongst the simplest in the animal kingdom. It lacks nerve and muscle cells and a centralised gut (reviewed in [1,12-14]). Porifera is traditionally regarded as the oldest surviving phyletic lineage of animals. However, as recent molecular phylogenomic and phylogenetic analyses both support [1,15] and reject [3,4,16,17] this traditional view, it remains unclear as to whether sponges or ctenophores are the sister group to all other animals and whether poriferans are monophyletic. Thus, interpretations of the sponge body plan in the context of metazoan evolution range from it representing a state similar to the last common ancestor of modern animals to it being derived from a morphologically more complex ancestor that possessed a gut, nerves and muscles.

Here we have improved the gene annotations in the draft genome of *Amphimedon* by combining deep transcriptome data from four developmental stages with previously generated developmental ESTs and CEL-Seq – a single cell RNA-Seq method [18] - evidence across 82 sponge developmental samples, from early cleavage through metamorphosis [19]. The inclusion of these transcriptomes markedly improves the current *Amphimedon* protein-coding gene models, which were primarily based on *ab initio* predictions and low-throughput EST evidence, and increases the total number of protein-coding genes in the genome by 25%. Furthermore, analysis of transcripts across sponge development has for the first time revealed alternative splicing patterns in a sponge, which are more similar to those reported in yeast than to those described in eumetazoans.

Results
Evidence-based protein-coding gene annotation
We sequenced and assembled *de novo A. queenslandica* polyadenylated RNAs present in adult, juvenile, competent and pre-competent larval stages in a strand-specific manner using Trinity [20]. To help detect low-abundance transcripts we also sequenced an adult sponge sample at high-depth in an unstranded manner and assembled it *de novo* with Trinity [20] (Table 1, see Methods). All strand-specific transcripts were combined with 8,880 previously

assembled EST contigs from larval stages [1] using PASA [21]. The best open reading frames (ORFs) were predicted from the representative transcripts generated by PASA (Figure 1A). To better resolve *A. queenslandica* gene families characterized by complex and highly repetitive regions that Trinity might assemble incorrectly (e.g. the Nucleotide-binding domain and Leucine-rich Repeat-containing (NLR) gene family [22]), an independent genome-guided assembly for each developmental stage was generated using Cufflinks [23]. Only Cufflinks transcripts found in at least two developmental stages were used as additional evidence for gene annotation (Figure 1A).

De novo and genome-based assembled transcripts, predicted ORFs and the previously generated *ab initio* gene models [1] were combined using EVM [24] to predict protein-coding gene models. Untranslated regions (UTRs) were added to these EVM gene models by two successive rounds of PASA using all developmental stranded Trinity transcripts and ESTs (Figure 1A).

The completed set of *Amphimedon* genes - Aqu2 - contains a total of 47,895 transcripts, which includes alternatively spliced gene isoforms expressed in different developmental stages (see below). To reduce isoform redundancy we identified each gene's isoform with the longest ORF (Figure 1A), resulting in 40,122 protein-coding loci in the final Aqu2 gene annotation. Finally, deep 3' end-biased CEL-Seq [18] expression data spanning 82 *A. queenslandica* developmental samples [19] were used to refine the 3' ends of Aqu2 gene models, resulting in the extension of the 3' ends of 10,925 genes (Figure 1A).

Comparison with previous annotations
Currently, the main *Amphimedon* gene annotation resource available to the community is Aqu1. Aqu1 has 30,060 genes and was released along with the original report of the *Amphimedon* genome [1] (Table 2). Additionally, NCBI generated a limited set of predicted genes via their automated pipeline upon genome submission, resulting in 9,975 protein-coding gene predictions. To assess the gene annotation improvements, Aqu2 was compared with these gene annotations and previously generated *ab initio* gene model predictions [1] (Figure 1B).

21,921 (54.6%) Aqu2 models share at least 80% identity with Aqu1 models, with many of the revised genes having a different structure (i.e. exon-intron architecture) or length (Figure 1B). 4,340 Aqu1 gene models are not supported at all in the Aqu2 annotation. Also 43,279 (71.6%) of *ab initio* and 7,918 (79.4%) of NCBI annotated genes are included in Aqu2. Some NCBI models are fragmented into smaller gene models resulting in 9,188 Aqu2 models from 7,918 NCBI models (Figure 1B).

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 3 of 11

Table 1 Transcriptome sequencing statistics

	Precompetent larvae	Competent larvae	Juvenile	Adult	Adult deep
Reads	42,273,865	43,699,007	41,677,487	43,098,690	125,880,671
Quality trimmed reads	39,234,338	40,546,526	38,587,047	39,976,627	116,384,700
Assembled transcripts	107,725	122,306	101,432	112,187	230,181
Average transcript length	816	738	706	716	1,278
Longest transcript	14,660	12,052	9,354	16,961	29,513
Mapped Transcripts	63,681	64,826	61,149	65,377	76,045
% Mapped/Assembled	59.11%	53.00%	60.29%	58.28%	33.04%
Average mapped transcript	738	729	673	686	918
Longest mapped transcript	14,393	12,052	24,959	16,954	29,166
Total transcripts mapped to genome	63,681	64,826	61,149	65,377	76,045
Total coverage (% genome)	28.32%	27.92%	29.24%	30.31%	21.62%
Used for PASA	Yes	Yes	Yes	Yes	No

In contrast, many adjacent *ab initio* models have been merged resulting in 32,383 Aqu2 models from 43,279 *ab initio* models (Figure 1B).

Aqu2 covers 16.7% more of the current *A. queenslandica* genome (Table 2) and includes 35,231 newly annotated exons, with 5,309 of the previous Aqu1 models having additional exons in their corresponding Aqu2 models (Figure 2A). There is also a marked increase in the number of 3' and 5' untranslated regions (UTRs) (Figure 2B and Table 2). The use of CEL-Seq 3' end evidence results in an increase from 6,021 to 14,892

annotated 3' UTRs (Table 2) in Aqu2, while 5' UTRs only increase from 4,457 to 9,873 (Table 2).

A comparison of the protein best blast hit (BBH) of Aqu1 and Aqu2 gene models against the SwissProt database reveals that Aqu1 and Aqu2 have a similar number of BBH to metazoan proteins, with slightly fewer matches in Aqu2 (419 proteins). In contrast, Aqu2 has more unique coding sequences (i.e. no blast matches in the database) with identifiable PFAM domains proteins (3,804) compared than Aqu1 (2,649) (Additional file 1: Figure S1), potentially expanding the list of lineage-
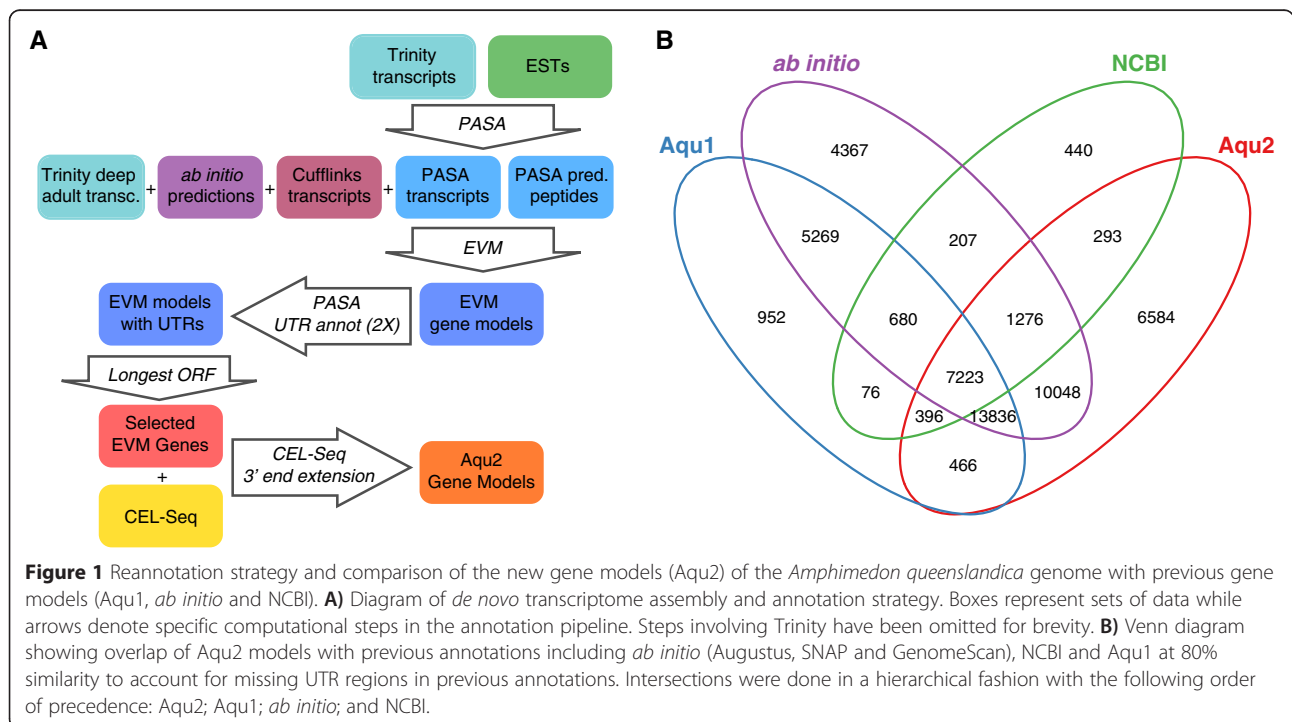

Figure 1 Reannotation strategy and comparison of the new gene models (Aqu2) of the *Amphimedon queenslandica* genome with previous gene models (Aqu1, *ab initio* and NCBI). **A)** Diagram of *de novo* transcriptome assembly and annotation strategy. Boxes represent sets of data while arrows denote specific computational steps in the annotation pipeline. Steps involving Trinity have been omitted for brevity. **B)** Venn diagram showing overlap of Aqu2 models with previous annotations including *ab initio* (Augustus, SNAP and GenomeScan), NCBI and Aqu1 at 80% similarity to account for missing UTR regions in previous annotations. Intersections were done in a hierarchical fashion with the following order of precedence: Aqu2; Aqu1; *ab initio*; and NCBI.

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 4 of 11

Table 2 Aqu1 and Aqu2 gene annotation comparison

	Aqu1	Aqu2
Total gene number	30,060	40,122
Total length (% genome)	69.3 Mb (47.82%)	93.5 Mb (64.52%)
Average gene length *	2,426 bp	2,521 bp
Smallest gene *	81 bp	149 bp
Longest gene *	86,995 bp	103,992 bp
Total isoform number	30,060	47,895
Total exon number	171,753	206,984
Average exon length	218 bp	225 bp
Average exons per gene	5.7	5.2
Max exons per gene	97	103
Total intron number	140,027	166,862
Average intron length	253 bp	327 bp
ORF		
Total length	35.6 Mb	42.6 Mb
Average size	1,184 bp	1,062 bp
Longest	47,676 bp	56,682 bp
5' UTRs		
Total number	4,457	9,873
Total length	0.6 Mb	1.13 Mb
Average size	130 bp	114 bp
Longest	2,318 bp	4,952 bp
3' UTRs		
Total number	6,021	14,892
Total length	1. 3 Mb	2.8 Mb
Average size	211 bp	188 bp
Longest	2,268 bp	2,814 bp

*Includes introns and exons.

restricted genes. Finally, there are 17,310 unannotated genes in Aqu2 compared to 7,879 in Aqu1, which will require further verification to establish if they are present in other basal metazoans, unique to poriferans, or restricted to demosponges.

Improvements in Aqu2 are exemplified in the locus depicted in Figure 2C, which shows the gene encoding the developmental transcription factor GATA [1,25-27], and a previously unannotated gene transcribed in the opposite direction from a putative bidirectional promoter. This gene was missing from the previous annotation (Aqu1) although it was predicted by *ab initio* methods (Figure 2C). In Aqu2 both genes have annotated 3' and 5' UTRs; CEL-Seq data further extend both the 3' ends. The significant increase in gene model number and length results in a more gene dense genome with a decrease of the median intergenic distance from 929 to 587 bp (Figure 2D).

Identification of previously unannotated protein-coding genes

Although Aqu2 represents a more complete picture of the genes present in the *Amphimedon*, we find most improvements in conserved gene families are minor and are generally restricted to more accurate assignment of exons and untranslated regions. However there are a few notable exceptions. For instance, we identified a number of new transcription factors, including the *Aristaless* homeobox (*ArxC*) gene, which, in spite of having been previously identified by Larroux et al. [28] was missing from the Aqu1 annotation. In this case, the 5' end of the corresponding Aqu1 model is discarded in Aqu2 and the 3' end extended. The discarded 5' end encodes the splicing factor 3B subunit and now comprises the adjacent gene. Aqu2 also includes a new member of the POU transcription factor gene family and previously unannotated genes encoding neuronal proteins including the Synapse Differentiation-Induced Protein1-Like (Capucin), a gene expressed in the caudal and putamen brain regions of mouse and human, and a new version of CPEB, a protein involved in memory maintenance [29-32] (Additional file 1: Figure S2A,B).
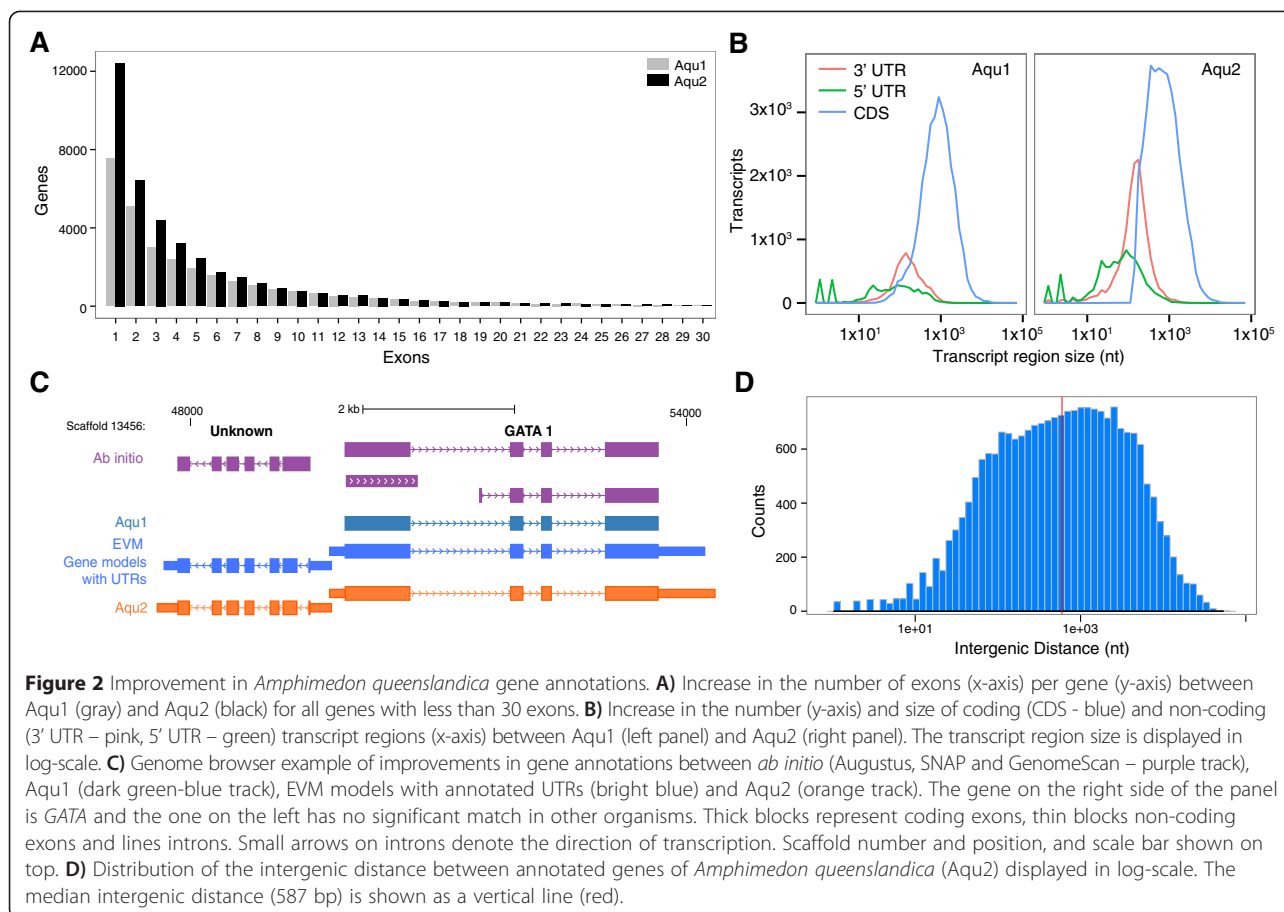
Amphimedon possesses both conserved and novel transcription termination elements

We identified motifs enriched in 10,274 strict 3' UTRs that are now annotated in *Amphimedon*. There are four long AT-rich motifs that are overrepresented in this region, three of which sit between 100 and 60 bp upstream of the transcription termination site (TTS) (Figure 3A,B). These motifs are more abundant than the polyadenylation signal sequence (PAS) consensus sequence (AWUAAA), which is found adjacent to and preceding the TTS (Figure 3A). One of the identified motifs - motif 8 - is a composite version of the polyA signal (AATx5 – Figure 3B) that, as expected, overlaps with the canonical PAS sequence (Figure 3A).

Comparison of the cumulative frequency of the consensus PAS relative to the transcription start site (TSS) reveals PASs accumulate more rapidly upstream than downstream of a set of 3,309 strict TSSs (Figure 3C). This pattern of a lower frequency of PASs on the coding strand is consistent with PAS being associated with transcription termination in *A. queenslandica* [33].

Alternative splicing is dominated by intron retention

A conservative estimate of alternative splicing (AS) in *A. queenslandica* was obtained by considering only AS events supported by at least three different assembled transcripts (Additional file 1: Figure S3). Only AS events resulting in the acquisition of an alternative first or last exon were lowly supported, with 98% of these appearing,

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 5 of 11

Figure 2 Improvement in *Amphimedon queenslandica* gene annotations. **A)** Increase in the number of exons (x-axis) per gene (y-axis) between Aqu1 (gray) and Aqu2 (black) for all genes with less than 30 exons. **B)** Increase in the number (y-axis) and size of coding (CDS - blue) and non-coding (3' UTR – pink, 5' UTR – green) transcript regions (x-axis) between Aqu1 (left panel) and Aqu2 (right panel). The transcript region size is displayed in log-scale. **C)** Genome browser example of improvements in gene annotations between *ab initio* (Augustus, SNAP and GenomeScan – purple track), Aqu1 (dark green-blue track), EVM models with annotated UTRs (bright blue) and Aqu2 (orange track). The gene on the right side of the panel is *GATA* and the one on the left has no significant match in other organisms. Thick blocks represent coding exons, thin blocks non-coding exons and lines introns. Small arrows on introns denote the direction of transcription. Scaffold number and position, and scale bar shown on top. **D)** Distribution of the intergenic distance between annotated genes of *Amphimedon queenslandica* (Aqu2) displayed in log-scale. The median intergenic distance (587 bp) is shown as a vertical line (red).

at most, in two different assembled transcripts (Additional file 1: Figure S3).

Based on these conservative estimates, alternative splicing in *A. queenslandica* appears to be less prevalent than in many eumetazoans [34], with less than 32% of the total transcripts detected in this study being generated by some form of AS (Figure 4). The large majority of AS events result in the retention of an intron, constituting 45% of all alternative splicing events and 57.1% of all alternatively spliced transcripts (Table 3 and Figure 4). The second most abundant splicing event results in the incorporation of an alternative terminal exon (22.1% of AS events), followed by alternative splice acceptor (17.3%) and donor (12.7%) (Table 3 and Figure 4).
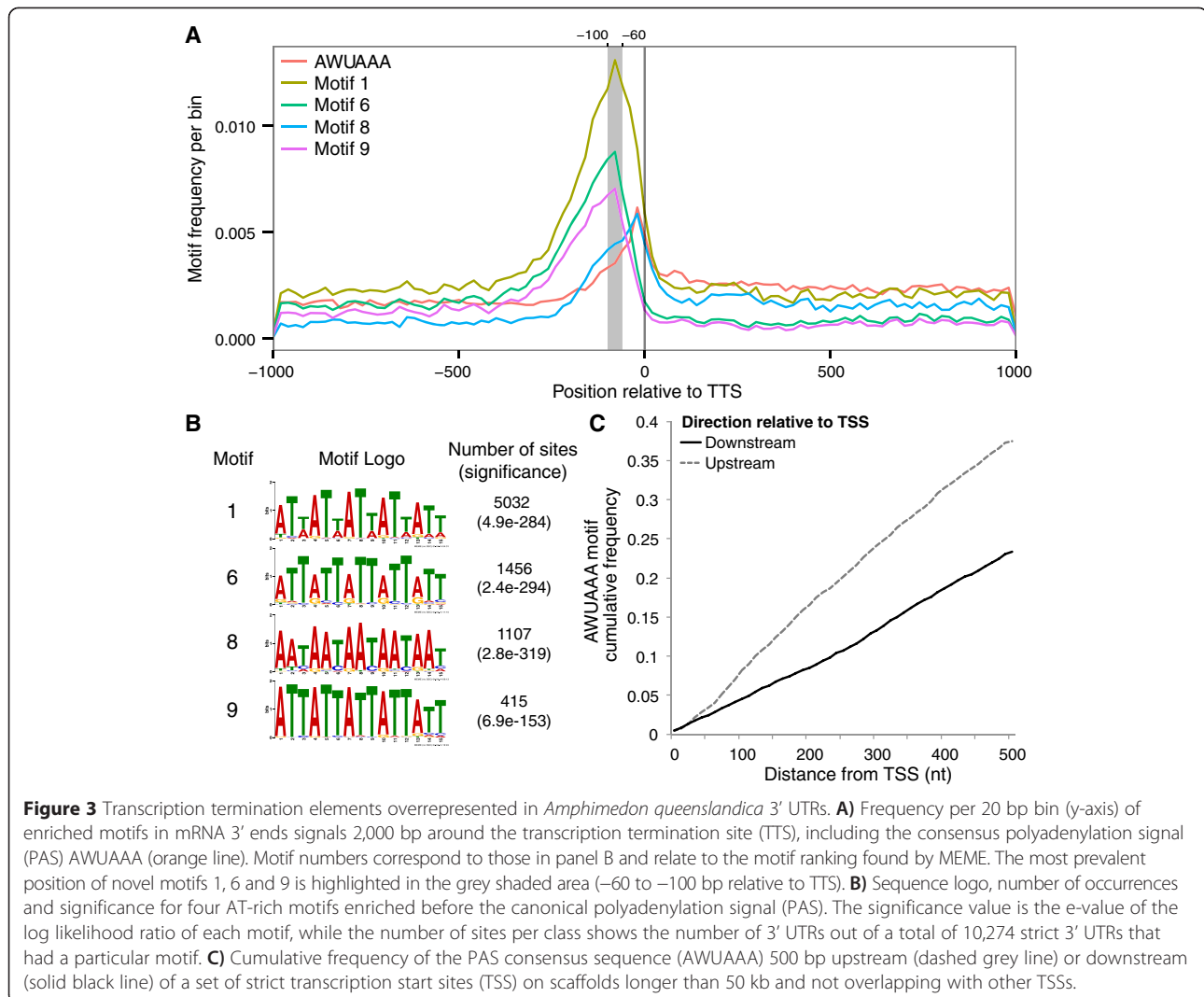
Discussion

The use of high-coverage developmental transcriptomes has markedly improved the gene annotations in the *Amphimedon queenslandica* genome, resulting in the refinement of existing gene models and the identification of a large number of previously unannotated genes. Given the high density of genes in the *Amphimedon* genome, the use of stranded RNA-Seq was proved essential for accurate gene identification and gene model prediction. The

integration of CEL-Seq data from 82 developmental samples, spanning from early cleavage through metamorphosis into the juvenile form, further improves the gene models by (i) allowing 3' UTRs to be extended to regions that have CEL-Seq sequence support and (ii) confirming the developmental expression of new gene models.

Combining stranded-RNA Seq and CEL-Seq data with existing gene models via a pipeline that relies both on *de novo* and genome-informed assemblies significantly improves the accuracy of existing gene models. Improvements include the addition or extension of 3' and 5' UTRs, the identification of missing exons, the removal of incorrectly predicted exons and the refinement of exon/intron boundaries. The high level of transcriptome coverage identified genes not included in previous annotations. This approach also has allowed us to identify gene models that were previously fused or fragmented in Aqu1. Although Aqu2 is primarily based on transcriptome evidence, both Aqu1 and Aqu2 show similar coverage of metazoan orthologues and support the expression of conserved metazoan proteins during *Amphimedon* development.

A number of biological observations emerge from these new gene annotations. First, the increase in the

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 6 of 11

Figure 3 Transcription termination elements overrepresented in *Amphimedon queenslandica* 3' UTRs. **A)** Frequency per 20 bp bin (y-axis) of enriched motifs in mRNA 3' ends signals 2,000 bp around the transcription termination site (TTS), including the consensus polyadenylation signal (PAS) AWUAAA (orange line). Motif numbers correspond to those in panel B and relate to the motif ranking found by MEME. The most prevalent position of novel motifs 1, 6 and 9 is highlighted in the grey shaded area (−60 to −100 bp relative to TTS). **B)** Sequence logo, number of occurrences and significance for four AT-rich motifs enriched before the canonical polyadenylation signal (PAS). The significance value is the e-value of the log likelihood ratio of each motif, while the number of sites per class shows the number of 3' UTRs out of a total of 10,274 strict 3' UTRs that had a particular motif. **C)** Cumulative frequency of the PAS consensus sequence (AWUAAA) 500 bp upstream (dashed grey line) or downstream (solid black line) of a set of strict transcription start sites (TSS) on scaffolds longer than 50 kb and not overlapping with other TSSs.

number of protein-coding genes in the *Amphimedon* genome has led to the expansion of some gene families, including those encoding developmental transcription factors, such as Arx and POU, and proteins involved in neuron functioning, such as Capucin and CPEB. It is worth noting that although the Aqu2 models have led to a better annotation of metazoan gene families, most conserved gene families were accurately annotated in Aqu1.

Second, the more accurate annotation of untranslated regions in Aqu2 allows for the identification to transcription start and termination sites. Genomic sequences in the vicinity of these sites contribute to the regulation of gene transcription and transcript termination and stability. Core transcriptional elements overlap with TSSs [35-38] and PASs and other motifs in the 3' UTR control transcript termination, stability and localisation [39,40]. Analysis of the 3' UTRs in *Amphimedon* reveals the enrichment of a number of AT-rich motifs 60–100 bp upstream of TTS. These currently appear to be unique

to *Amphimedon*. Further, as observed in vertebrates [33], analysis of the frequency of PAS sequences upstream and downstream of putative TSSs in *Amphimedon* reveals a disproportionate depletion of PASs in the direction of transcription compared to in the opposite non-coding direction. This is consistent with PAS signals participating in transcription termination in *Amphimedon*.

Third, the extension of existing genes and the annotation of new genes both have contributed to an overall increase in gene density. Indeed, the *Amphimedon* genome is the most gene dense animal genome currently known [5]. In addition to having minimal intergenic spacing (median = 0.59 kb), intron size in *Amphimedon* is markedly smaller than other animals (see [5] and Table 2). Both intergenic and intron size in *Amphimedon* are more similar to non-metazoan opisthokonts [1,5,23]. Given the basal position of poriferans, these characteristics suggest sponges may have retained genomic features of the first metazoans. Although protein-coding gene
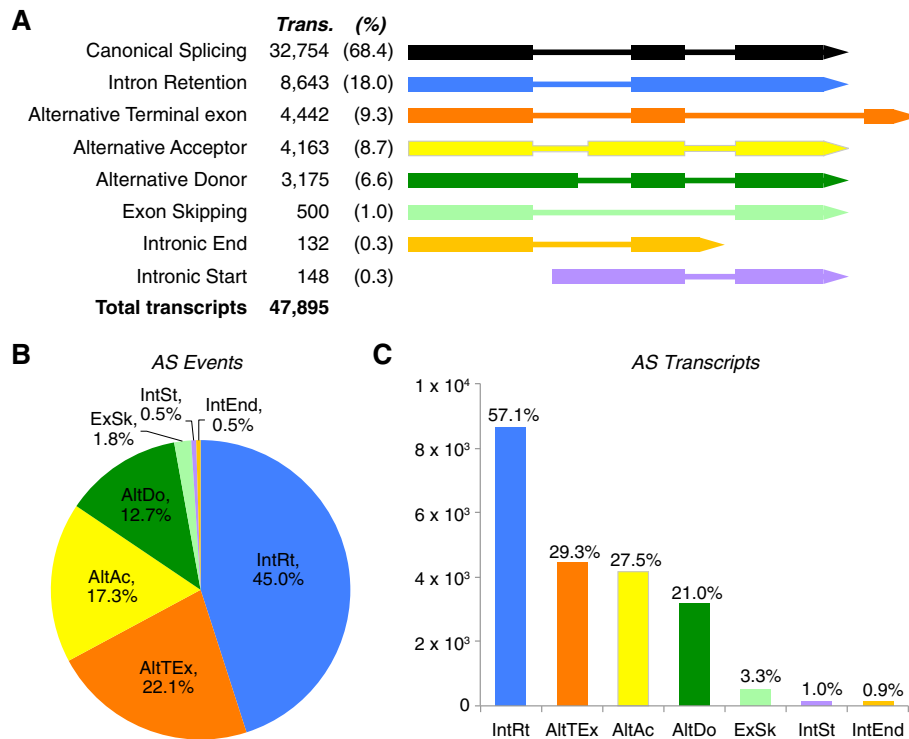
Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 7 of 11

Figure 4 Alternative splicing in *Amphimedon queenslandica*. **A)** Graphical depiction of AS event classes in relation to the canonically spliced isoform (black). From top to bottom: intron retention (IntRt - blue), alternative terminal exon (AltTEx - orange), alternative acceptor (AltAc - yellow), alternative donor (AltDo - green), exon-skipping (ExSk – light green), intron end (IntEnd - light orange) and intron start (IntSt - violet). The number of transcripts in each class of AS is shown in the second column and the percentage of all transcripts this represents is shown in parentheses. The total percentage is over 100% as different types of AS sometimes occur in the same transcript. **B)** Pie chart showing percentage of AS events belonging to the classes shown in panel A. **C)** Number of transcripts (y-axis) and percentage (number above each bars) of transcripts with AS events (x-axis) exemplified in panel A. Similar to panel A, the total percentage is over 100% as different types of AS events can occur in the same transcript. Colour coding and abbreviations in B and C are identical to those used in panel A.

content in sponges includes many metazoan innovations [1,5,9-11], their genome organisation and gene structure appears to be more similar to simple unicellular ophistokonts.

Fourth and consistent with the above observation, the level and modes of alternative splicing in *Amphimedon* is more akin to those found in yeast than in other animals. This sponge has lower proportion of alternative splicing events compared to other animals, particularly those that result in exon-skipping and gene product diversification. These very low levels of exon-skipping are similar to those observed in yeast [41,42] and in contrast to bilaterians, where exon-skipping is often the most prevalent form of AS [42,43]. As an increase in

Table 3 Alternative splicing in *Amphimedon queenslandica*

Abbreviation	AS Type	AS events*	Number of transcripts	% of AS events	% of AS transcripts+	% of All transcripts
IntRt	Intron retention	12,610	8,643	45.0%	57.1%	18.0%
AltTEx	Alternative Terminal Exon	6,191	4,442	22.1%	29.3%	9.3%
AltAc	Alternative Acceptor	4,843	4,163	17.3%	27.5%	8.7%
AltDo	Alternative Donor	3,562	3,175	12.7%	21.0%	6.6%
ExSk	Exon skipping	513	500	1.8%	3.3%	1.0%
IntSt	Intronic Start	148	148	0.5%	1.0%	0.3%
IntEnd	Intronic End	132	132	0.5%	0.9%	0.3%

*AS events supported by three or more transcripts.
+The total percentage of transcripts is higher than 100% as different types of AS events can occur in the same transcript.

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 8 of 11

average intron size correlates with increased levels of exon-skipping [34,44], the limited exon-skipping and small intron size in *Amphimedon* are consistent with these genomic features and processes emerging later in eumetazoan evolution, after the divergence of this and the sponge lineage.

Fifth, the new Aqu2 models greatly expand the number of gene models without orthologues to over 20,000. Nearly all these genes are developmentally expressed based on RNA-Seq and CEL-Seq data. With a paucity of whole genome data from phylum Porifera it is currently difficult to reconstruct the evolutionary history of these genes.

Conclusions

In improving the accuracy of the *Amphimedon* gene models we have increased the number of full-length genes with accurate transcription start and termination sites. This allows for the future identification and analysis of promoters and other regulatory sequences populating intergenic DNA and UTRs. Combined with experimental manipulation and a detailed analysis of gene expression, the analysis of *cis*-regulatory DNA provides a means to understand the logic underlying sponge morphogenesis and cell specification and differentiation. When placed in a comparative framework, this knowledge informs our understanding of the evolution of the cell types [24,45] and developmental mechanisms underpinning metazoan body plans.

Methods

Sample collection and sequencing

Adult, juvenile, and competent and pre-competent larvae of *Amphimedon queenslandica* sponges were collected from Heron Island Reef, Great Barrier Reef, Queensland, Australia as previously described [46]. Total RNA from each developmental stage was extracted using the standard TRIzol regent protocol (Invitrogen) and genomic DNA was removed by DNase treatment. The RNA quality was assessed using the Agilent 2100 Bioanalyzer. RNA was paired-end sequenced using the Illumina HiSeq2000 platform (Illumina, San Diego). All samples were sequenced in a strand sensitive fashion. We additionally sequenced an adult sponge tissue sample at high-depth in an unstranded manner to help detect low-abundance transcripts (Table 1).

De novo transcriptome assembly

Raw paired-end sequences were quality filtered using Trimmomatic [47]. The first 7 bp of each read were cropped and reads were subsequently trimmed if the average quality within a window of 4 bp was below 15. Unpaired reads and reads smaller than 60 bp were discarded. Quality-filtered paired-end reads were assembled de novo using Trinity [20] (Table 1). Each developmental

stage was assembled independently with default parameters, with the exception of a lower transcript size cut-off of 200 nt and jaccard-clipping. These assembled transcripts for each developmental stage were aligned and condensed using the PASA pipeline [24], where only transcripts with more than 90% transcript coverage (parameter −MIN_PERCENT_ALIGNED) and 95% identity (parameter −MIN_AVG_PER_ID) to the genome were merged. Peptides were predicted from these transcripts using Transdecoder [48] and used as further evidence for gene annotation (see below).

High-coverage unstranded adult sponge reads were quality checked and trimmed as described above. Remaining reads were independently assembled de novo three times using Trinity with default parameters, using a lower transcript size cut-off of 200 nt and jaccard-clipping, but including a minimum kmer coverage of 2, 4 and 10 (−−min_kmer_cov parameter). The three Trinity assemblies were subsequently merged using CAP3 [49] at 95% similarity and a minimum overlap of 100 bp. These sequences were mapped to the genome using gmap with a minimum of 90% identity [50] and provided as further transcript evidence to EVM [24], but not included in the main PASA transcript set (see Figure 1A, Table 1).

Reference based transcriptome assembly

Quality filtered reads from the four stranded libraries (Table 1) were mapped to the *A. queenslandica* genome [1] using Tophat2 [51] and assembled using Cufflinks2 [23]. Each developmental stage was assembled separately and shared transcripts were collapsed using Cuffmerge [23]. The gtf file obtained by Cuffmerge was converted to gff3 format and used as further evidence for gene annotation.

Evidenced based gene prediction and UTR annotation

Gene evidence and predicted gene structure were combined using EVM [24]. The evidence included a) *ab initio* predictions generated by Augustus, SNAP and GenomeScan [1], b) PASA generated consensus transcript assemblies based on both the stranded developmental Trinity assemblies and publically available Sanger ESTs [1], c) the Transdecoder predicted peptides of PASA consensus transcripts, d) the high-depth adult transcriptome and e) the genome-guided gene models generated by Cufflinks. RNA-Seq evidence was strongly favoured over *ab initio* and other predictions using the weight system incorporated in EVM. The evidence weights are summarized in Table S1. UTRs were added onto the gene models predicted by EVM [24] by two sequential PASA [24] rounds including annotation loading, annotation comparison

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 9 of 11

and annotation updates to maximize incorporation onto gene models predicted by EVM, as per the suggestion of the authors in the PASA pipeline manual (http://pasapipeline.github.io/) (see Figure 1A).

CEL-Seq gene 3′ end extension

CEL-Seq developmental data (stages: cleavage, brown, cloud, spot, late spot, ring, late ring and swimming larvae) for *Amphimedon* developmental stages were retrieved from NCBI's Gene Expression Omnibus (GSE54364) and are described in detail in [19]. 24 additional samples spanning from post-settlement postlarvae undergoing metamorphosis into the juvenile form and adult were provided by the Yanai lab (Anavy et al. unpublished). CEL-Seq reads were processed, quality filtered and mapped back to the *A. queenslandica* genome (ampQue1) using BWA [51] through the CEL-Seq analysis pipeline [52]. To identify transcript ends, we clustered all overlapping reads mapped to the same DNA strand in each individual developmental sample. Developmental stage replicates were processed individually. Clusters with at least 10 reads were retained. Clustered regions were identified in several developmental samples in a stranded fashion; resulting in a total of 74,973 CEL-Seq based clusters. We extended the 3′ end of all EVM gene models with annotated 3′ UTRs whose last annotated exon had at least 10 bp overlap with these CEL-Seq clusters, only if their annotated 3′ end was shorter than the one supported by CEL-Seq (Figure 1A).

Gene annotation

Open reading frames (ORFs) for all genes were predicted using Transdecoder [48]. All best ORF candidates were analysed for protein domains, signal sequences and transmembrane domain using hmmer 3.0 [53], signalp 4.1 [54], blastp + [55] and tmhmm 2.0 [56], and combined to annotate each ORF using Trinotate ([24]). Novel candidate genes were manually verified as related to other known genes in nr, RefSeq and SwissProt databases using the web interfaces of Blast and PSI-blast [55]. Their protein domains were also verified using the web versions of SMART [57] and InterProScan [58]. The sequences of newly annotated proteins discussed in the text are provided in the Additional file 1: Supplementary material. The complete set of predicted peptides and gene annotations can be accessed at http://amphimedon.qcloud.qcif.edu.au/index.html.

Assembly comparison

The assembly comparison shown in Figure 1B was done by intersecting Aqu1, *ab initio* gene models (Augustus, SNAP and GenomeScan), NCBI and Aqu2 annotations. An 80% genome coverage threshold was used to account for missing UTR regions in previous annotations. As in

some cases a single Aqu1 gene might correspond to two or more Aqu2 genes or vice versa, we have used a hierarchical approach using the following order of precedence: Aqu2; Aqu1; *ab initio*; and NCBI. Only the original reference (Aqu2) set will be identical in number in the Venn diagram as the original number of elements in the set, while all other comparison sets (Aqu1, *ab initio* and NCBI) will have more or less elements depending on their overall correspondence with the reference set. Other intersections, such as the number of Aqu1 genes that are not supported in Aqu2, as well as the number of *ab initio* and NCBI covered in Aqu2 were done using overlapSelect (parameters: −overlapThreshold = 0.8 −strand) from the UCSC toolkit [59].

3′ end motif identification

10,274 3′ UTR sequences overlapping with CEL-seq based clusters found on genomic scaffolds longer than 50 kb were searched for nucleotide motifs using MEME (parameters: −maxsize 20000000 -p 14 -dna -nmotifs 10 -minw 6 -maxw 15 -mod zoops) [60]. We restricted our search to sequences in scaffolds longer than 50 kb to avoid PAS signal depletion due to lack of adjacent sequence. Motif frequency matrices were converted from MEME to Homer format and used to map their frequency around the annotated TTS and TSS using the Homer toolkit [61]. For the cumulative PAS signal distribution analysis, strict TSSs were defined as those of genes found in scaffolds longer than 50 kb whose promoters (100 bp upstream and 50 bp downstream of the TSS) did not overlap with other genes.

Alternative splicing analysis

The four stranded developmental transcriptomes and EST data were combined using the PASA pipeline with the alternative splicing detection option [24]. AS events supported by less than three transcripts were considered as lowly supported and removed from subsequent analyses.

Availability of supporting data

The transcriptome sequencing data has been submitted to NCBI's Sequence Read Archive (SRA) with accession number SUB596470. The new gene annotations, gene and transcriptome nucleotide and peptide sequences can be downloaded from our website (http://amphimedon.qcloud.qcif.edu.au/downloads.html). The new Aqu2 annotations can be visualized at our local genome browser (http://amphimedon.qcloud.qcif.edu.au/genome_browser.html).

CEL-Seq data can be access through the Gene Expression Omnibus (GEO) with GEO Accession GSE54364 [56].

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 10 of 11

Additional file

Additional file 1: Supplemental material including: Examples of novel proteins in Aqu2. **Figure S1.** Blastp best blast hit (BBH) annotation comparison. **Figure S2.** Improvements to the annotation of CPEB proteins. **Figure S3.** Transcript support for alternatively splicing events. **Table S1.** Weight of transcript evidence used for gene prediction via EVM.

Abbreviations

EST: Expressed sequence tag; AS: Alternative splicing; TSS: Transcription start site; TTS: Transcription termination site; AltTEx: Alternative terminal exon; IntRt: Intron retention; AltAc: Alternative acceptor; AltDo: Alternative donor; ExSk: Exon-skipping; IntEnd: Intron end; IntSt: Intron start; UTR: Untranslated region; nt: nucleotide; bp: base pair; Mb: Megabase.

Competing interests

The authors have declared no conflicts of interest.

Authors' contributions

Acknowledgments

References

1. Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, et al. The *Amphimedon* queenslandica genome and the evolution of animal complexity. Nature. 2010;466:720–6.
2. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T. The *Trichoplax* genome and the nature of placozoans. Nature. 2008;454:955–60.
3. Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, et al. The ctenophore genome and the evolutionary origins of neural systems. Nature. 2014;510:109–14.
4. Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, et al. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science. 2013;342:1242592.
5. Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, et al. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun. 2013;4:2325.
6. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature. 2008;451:783–8.
7. Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 2013;14:R15.
8. Erwin DH. Early origin of the bilaterian developmental toolkit. Philos Trans R Soc Lond B Biol Sci. 2009;364:2253–61.
9. Degnan BM, Vervoort M, Larroux C, Richards GS. Early evolution of metazoan transcription factors. Curr Opin in Genet Dev. 2009;19:591–9.
10. Sebé-Pedrós A, Ariza-Cosano A, Weirauch MT, Leininger S, Yang A, Torruella G, et al. Early evolution of the T-box transcription factor family. Proc Natl Acad Sci U S A. 2013;110:16050–5.
11. Richards GS, Degnan BM. The dawn of developmental signaling in the metazoa. Cold Spring Harb Symp Quant Biol. 2009;74:81–90.
12. Simpson TL. The Cell Biology of Sponges. 1st ed. New York: Springer; 1984.
13. Ereskovsky AV. The Comparative Embryology of Sponges. New York: Springer; 2010.
14. Leys SP, Hill A. The physiology and molecular biology of sponge tissues. Adv Mar Biol. 2012;62:1–56.
15. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, et al. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 2009;19:706–12.
16. Schierwater B, Eitel M, Jakob W, Osigus H-J, Hadrys H, Dellaporta SL, et al. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol. 2009;7, e20.
17. Sperling EA, Peterson KJ, Pisani D. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol. 2009;26:2261–74.
18. Degnan BM, Adamska M, Craigie A, Degnan SM, Fahey B, Gauthier M, et al. The demosponge Amphimedon queenslandica: reconstructing the ancestral metazoan genome and deciphering the origin of animal multicellularity. CSH Protoc 2008, 2008:pdb.emo108. http://cshprotocols.cshlp.org/citmgr?gca=protocols%3B2008%2F12%2Fpdb.emo108.
19. Anavy L, Levin M, Khair S, Nakanishi N, Fernandez-Valverde SL, Degnan BM, et al. BLIND ordering of large-scale transcriptomic developmental timecourses. Development. 2014;141:1161–6.
20. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
21. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
22. Yuen B, Bayes JM, Degnan SM. The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. Mol Biol Evol. 2014;31:106–20.
23. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2012;31:46–53.
24. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9:R7.
25. Murakami R, Okumura T, Uchiyama H. GATA factors as key regulatory molecules in the development of *Drosophila* endoderm. Dev Growth Differ. 2005;47:581–9.
26. Martindale MQ, Pang K, Finnerty JR. Investigating the origins of triploblasty: "mesodermal" gene expression in a diploblastic animal, the sea anemone *Nematostella vectensis* (phylum, Cnidaria; class, Anthozoa). Development. 2004;131:2463–74.
27. Nakanishi N, Sogabe S, Degnan BM. Evolutionary origin of gastrulation: insights from sponge development. BMC Biol. 2014;12:26.
28. Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. Genesis and expansion of metazoan transcription factor gene classes. Mol Biol Evol. 2008;25:980–96.
29. Kandel ER. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. Mol Brain. 2012;5:14.
30. Bestman JE, Cline HT. The RNA binding protein CPEB regulates dendrite morphogenesis and neuronal circuit assembly in vivo. Proc Natl Acad Sci U S A. 2008;105:20494–9.
31. Keleman K, Krüttner S, Alenius M, Dickson BJ. Function of the Drosophila CPEB protein Orb2 in long-term courtship memory. Nature Neurosci. 2007;10:1587–93.
32. Si K, Giustetto M, Etkin A, Hsu R, Janisiewicz AM, Miniaci MC, et al. A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in *Aplysia*. Cell. 2003;115:893–904.
33. Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, et al. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol. 2013;20:923–8.
34. Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 2006;35:125–31.
35. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012;13:233–45.
36. Kadonaga JT. Perspectives on the RNA polymerase II core promoter. WIREs Dev Biol. 2011;1:40–51.
37. Juven-Gershon T, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol. 2010;339:225–9.

Fernandez-Valverde *et al. BMC Genomics* (2015) 16:387

Page 11 of 11

38. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. Annu Rev Biochem. 2003;72:449–79.
39. Tian B, Graber JH. Signals for pre-mRNA cleavage and polyadenylation. WIREs RNA. 2012;3:385–96.
40. Alonso CR. A complex "mRNA degradation code" controls gene expression during animal development. Trends Genet. 2012;28:78–88.
41. Bitton DA, Rallis C, Jeffares DC, Smith GC, Chen YYC, Codlin S, et al. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. Genome Res. 2014;24:1169–79.
42. Awan AR, Manfredo A, Pleiss JA. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc Natl Acad Sci U S A. 2013;110:12762–7.
43. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008;321:956–60.
44. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11:345–55.
45. Conaco C, Bassett DS, Zhou H, Arcila ML, Degnan SM, Degnan BM, et al. Functionalization of a protosynaptic gene expression network. Proc Natl Acad Sci U S A. 2012;109 Suppl 1:10612–8.
46. Adamska M, Larroux C, Adamski M, Green K, Lovas E, Koop D, et al. Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. Evol Dev. 2010;12:494–518.
47. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
48. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protoc. 2013;8:1494–512.
49. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9:868–77.
50. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.
51. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
52. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2:666–73.
53. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63.
54. Petersen TN, Brunak S, Heijne von G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Meth. 2011;8:785–6.
55. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
56. Krogh A, Larsson B, Heijne von G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305:567–80.
57. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A. 1998;95:5857–64.
58. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17:847–8.
59. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144–61.
60. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28–36.
61. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.