A high-resolution mRNA expression time course of embryonic development in zebrafish

Richard J. White[1], John E. Collins[1], Ian M. Sealy[1], Neha Wali[1], Christopher M. Dooley[1], Zsofia Digby[1,3], Daniel N. Murphy[2], Thibaut Hourlier[2], Anja Füllgrabe[2], Matthew P. Davis[2], Anton J. Enright[2] and Elisabeth M. Busch-Nentwich[1†]

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK

[2]EMBL-European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK

[3]Current address: Department of Veterinary Medicine, Madingley Road, Cambridge, CB3 0ES, UK

[†]Author for correspondence, emb@sanger.ac.uk

Abstract

We have produced an mRNA expression time course of early zebrafish development across 18 time points from 1-cell to 5 days post-fertilisation sampling individual and pools of embryos. Using poly(A) pulldown stranded RNA-seq as well as a 3′ end transcript counting method we characterise the temporal expression profiles of 23,642 genes. We identify temporal and functional transcript co-variance that associates 5,024 unnamed genes with distinct developmental time points. Specifically, a class of over 100 previously uncharacterised zinc finger domain containing genes, all located on the long arm of chromosome 4, is expressed in a sharp peak during zygotic genome activation. The data reveal complex and widespread differential use of exons, previously unidentified 3′ ends across development, new primary microRNA transcripts and the temporal divergence of gene paralogues generated in the teleost genome duplication. To make this dataset a useful baseline reference, the data are accessible to browse and download at Expression Atlas and Ensembl.

Keywords

Developmental biology, embryogenesis, RNA-seq, transcriptome, zebrafish.

Introduction

Gene regulatory interactions are the fundamental basis of embryonic development and transcription is one of the major processes by which these interactions are mediated. A time-resolved comprehensive analysis of relative mRNA expression levels is an important step towards understanding the regulatory processes governing embryonic development. Here we present a systematic assessment of the temporal transcriptional events during this critical period in zebrafish (*Danio rerio*).

The zebrafish is a unique vertebrate model system as it possesses high morphological and genomic conservation with humans, but also experimental tractability of embryogenesis otherwise only found in invertebrate model organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans*. Features such as very large numbers of offspring, *ex vivo* development and embryonic translucency have enabled comprehensive forward and reverse genetic screens (Amsterdam et al., 1999; Driever et al., 1996; Haffter et al., 1996; Kettleborough et al., 2013; Moens et al., 2008; Varshney et al., 2015) as well as high-throughput drug discovery approaches (Murphey et al., 2006; North et al., 2007; Peterson et al., 2000; Peterson et al., 2004; Stern et al., 2005). Together with a high quality genome (Howe et al., 2013b) only comparable in vertebrates to mouse and human, this has led to many important discoveries in areas such as zygotic genome activation (ZGA) during vertebrate development (Zygotic genome activation, Lee et al., 2013), blood stem cells (Bertrand et al., 2010; Kissa and Herbomel, 2010), but also highly translational findings directly affecting human health (Li et al., 2015; Tobin et al., 2010).

The morphological processes underlying the transformation of a fertilised egg into a free swimming fish have been studied extensively owing to the ease with which embryogenesis can be observed and manipulated (Behrndt et al., 2012). This has identified many genes that drive crucial steps of the differentiation process, however the wealth of morphological phenotype data has not been matched with a systematic analysis of the corresponding molecular events. Baseline transcriptomic datasets in other species have greatly improved knowledge of relative levels of gene expression

as well as alternative splicing events (Boeck et al., 2016; Brown et al., 2014; Gerstein et al., 2014; Graveley et al., 2011; Hashimshony et al., 2015; Klepikova et al., 2016; Owens et al., 2016; Tan et al., 2013).

Previous baseline transcriptome work in zebrafish has focused on either certain aspects of development such as the maternal-zygotic transition (Aanes et al., 2011; Harvey et al., 2013; Lee et al., 2013), the identification of specific transcript types (Pauli et al., 2012) or promoters (Gehrig et al., 2009; Nepal et al., 2013), on gene annotation improvement (Collins et al., 2012) or have a limited number of time points and replicates (Yang et al., 2013).

Using poly(A) pulldown RNA-seq and a 3′ end enrichment method called DeTCT (Collins et al., 2015) we have generated a comprehensive polyadenylated RNA expression profile of normal zebrafish development. We have taken advantage of the large numbers of offspring and have divided the 6 day period of embryonic development into 18 stages. For each stage we have processed up to 29 biological replicates consisting of single or pooled embryos.

We have placed particular emphasis on making the presented data accessible without the need for computational processing. The data are available to browse and download in a user-friendly interface at Expression Atlas and as selectable tracks in the Ensembl genome browser. Furthermore, incorporation into the next Ensembl gene set update will significantly improve zebrafish gene and transcript isoform annotation.

This developmental polyadenylated RNA baseline reference enhances our understanding of the gene regulatory events during vertebrate development and provides a time-resolved transcriptional basis for the comprehensive annotation of functional elements in the zebrafish genome using existing and future genome-wide datasets (Haberle et al., 2014; Kaaij et al., 2016; Lindeman et al., 2010; Potok et al., 2013; Vastenhouw et al., 2010).

Results

A high-resolution transcriptional profile of zebrafish development

90 A total of 18 developmental time points were sampled in this study (Figure 1A). Four
91 time points are before or at the onset of zygotic transcription, four during
92 gastrulation, three during somitogenesis, three during prim stages – 24, 30 and 36
93 hours post-fertilisation (hpf) – and every 24 h from 2 to 5 days post-fertilisation (dpf).
94 This gives detailed coverage of all the important developmental processes taking
95 place during this time period. At each time point, samples were collected for both
96 RNA-seq and DeTCT (RNA-seq, pools of 12 embryos, 5 replicates; DeTCT, pools of
97 8 embryos, 11-12 replicates and individual embryos, 11-12 replicates). We obtained
98 an average of 3.8 million reads per sample for the RNA-seq and 9 million reads per
99 sample for the DeTCT data (Figure 1B). The number of genes detected rises steadily
100 as development proceeds and a wider range of genes are expressed (Figure 1C,
101 Supplemental Figure 1-1). With a cut-off of greater than or equal to 5 normalised
102 counts it rises from a mean of 10,395 to 18,225 for the RNA-seq (6,259 to 15,215 in
103 the DeTCT data).

104 For quality control purposes, counts from the RNA-seq data, normalised for library
105 size, were first used to produce a sample correlation matrix (Figure 1D). This shows
106 the excellent agreement among biological replicates and close relation of
107 neighbouring stages in terms of gene expression profiles. To give a high level
108 overview of the data, principal component analysis (PCA) was used to visualise
109 patterns. These show close clustering of biological replicates and a smooth transition
110 from one stage to another through developmental time. The first six principal
111 components (PCs) explain just over 90% of the variance in the data (Figure 1E-G
112 and Supplemental Figure 1-2); the first component accounts for 58.2% of the
113 observed variation. Summaries of the expression patterns underlying the PCs are
114 shown in Figure 1H. Broadly, the genes that contribute the most to the first PC are
115 either low at early time points and increase towards the end of the time course or the
116 inverse (high early and decreasing later). PC2 represents genes with a peak at
117 gastrula and somite stages. Genes contributing to PC3 show two peaks of
118 expression (onset of transcription and days 3-5). The other patterns are more
119 complex (Supplemental Figure 1-2) although PC6 shows a clear spike of expression
120 starting at the 1000-cell stage (3 hpf), which coincides with the start of zygotic
121 transcription.

122 In order for this dataset to be as useful as possible to the scientific community, we
123 have made it available as sequence with metadata through the European Read
124 Archive and in multiple locations for viewing. Pre-computed count profiles (FPKM)
125 are available to search/browse at Expression Atlas
126 (https://www.ebi.ac.uk/gxa/experiments/E-ERAD-475)(Petryszak et al., 2016), allowing
127 comparison of the expression profiles of multiple genes across the entire time course
128 (Figure 2A). This is searchable by gene/allele name, Ensembl/ZFIN ID and GO
129 terms. It is also possible to start with a single seed gene and have Expression Atlas
130 add in similarly expressed genes (as assessed by k-means clustering) for
131 comparison (https://www.ebi.ac.uk/gxa/FAQ.html#similarExpression). The data are
132 also viewable in Ensembl as separate stage-specific selectable tracks along with one
133 merged track (http://www.ensembl.org/Danio_rerio). The aligned reads can be
134 displayed as either coverage graphs or read pairs. Reads that span introns can be
135 viewed to investigate alternative splicing in a stage-specific manner (Figure 2B-C).

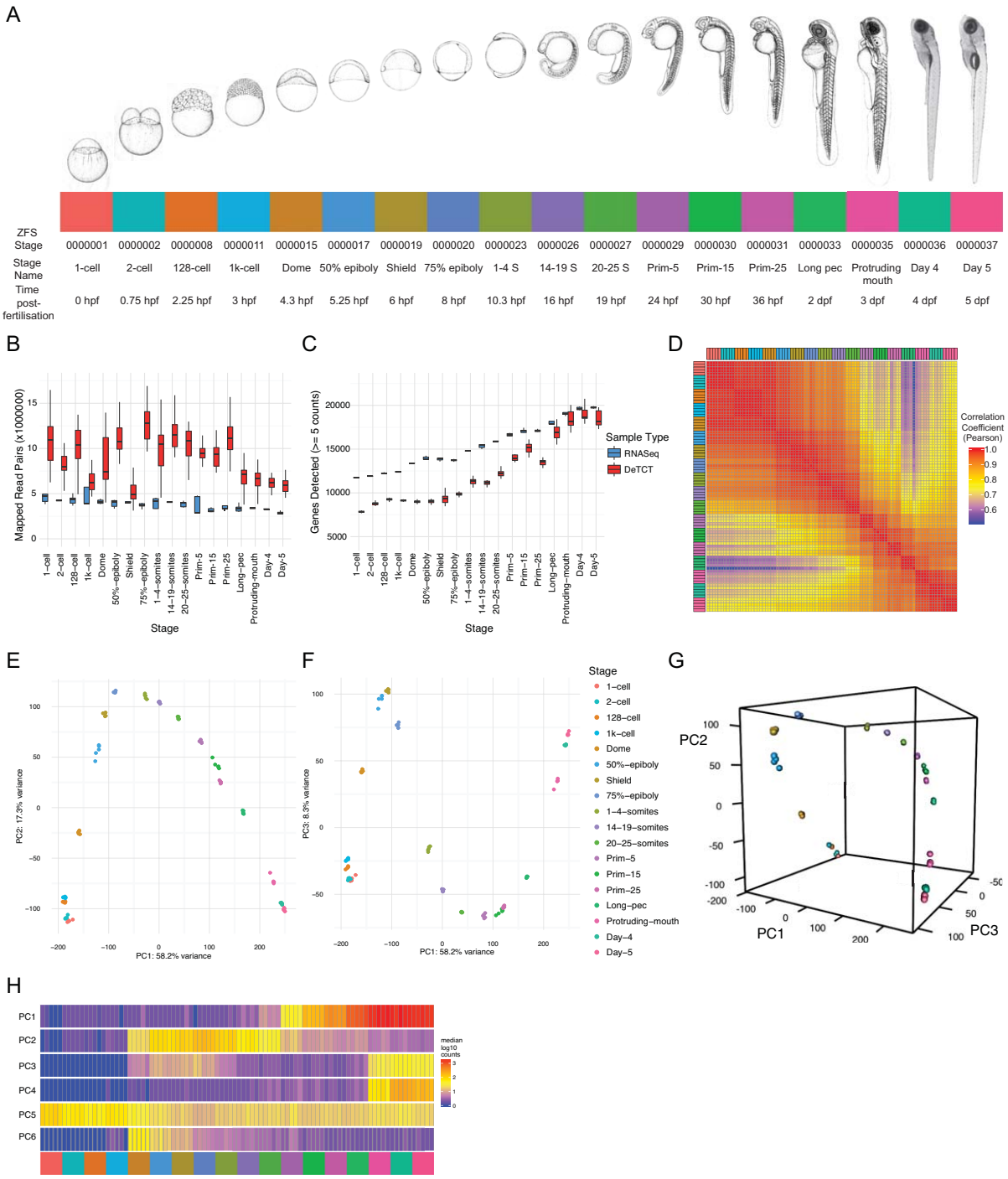


Figure 1. A transcriptional map of development. (A) Stages represented in this study. Images of embryos from 1-cell to protruding mouth are from (Kimmel et al., 1995). (B-C) Boxplot showing the distributions of numbers of mapped read-pairs and genes detected for each stage and protocol. (D) Sample correlation matrix showing the Pearson correlation coefficient between all pairwise combinations of samples. (E-G) Principal Component Analysis. (E) Principal Component (PC) 2 plotted against PC1. (F) PC3 plotted against PC1. (G) 3D rendering of the first 3 PCs. (H) Representation of expression profiles contributing to the first 6 PCs, produced by

calculating the median expression value (normalised counts) for each sample for the 100 genes that contribute most to that PC.

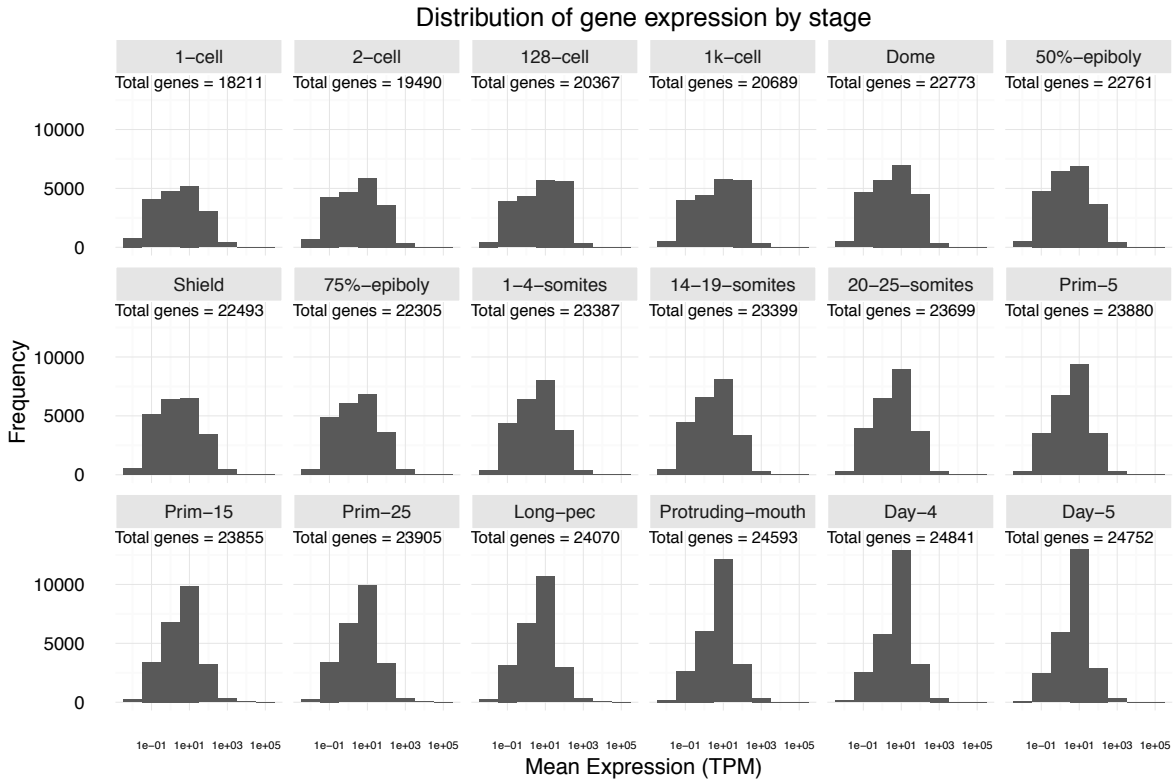


Figure1-1. Distribution of gene expression levels in RNA-seq data by stage. Histograms (one per stage) showing the distribution of the mean expression level for each gene over the 5 samples within the stage. The total number of genes represented for the stage is shown on each panel.

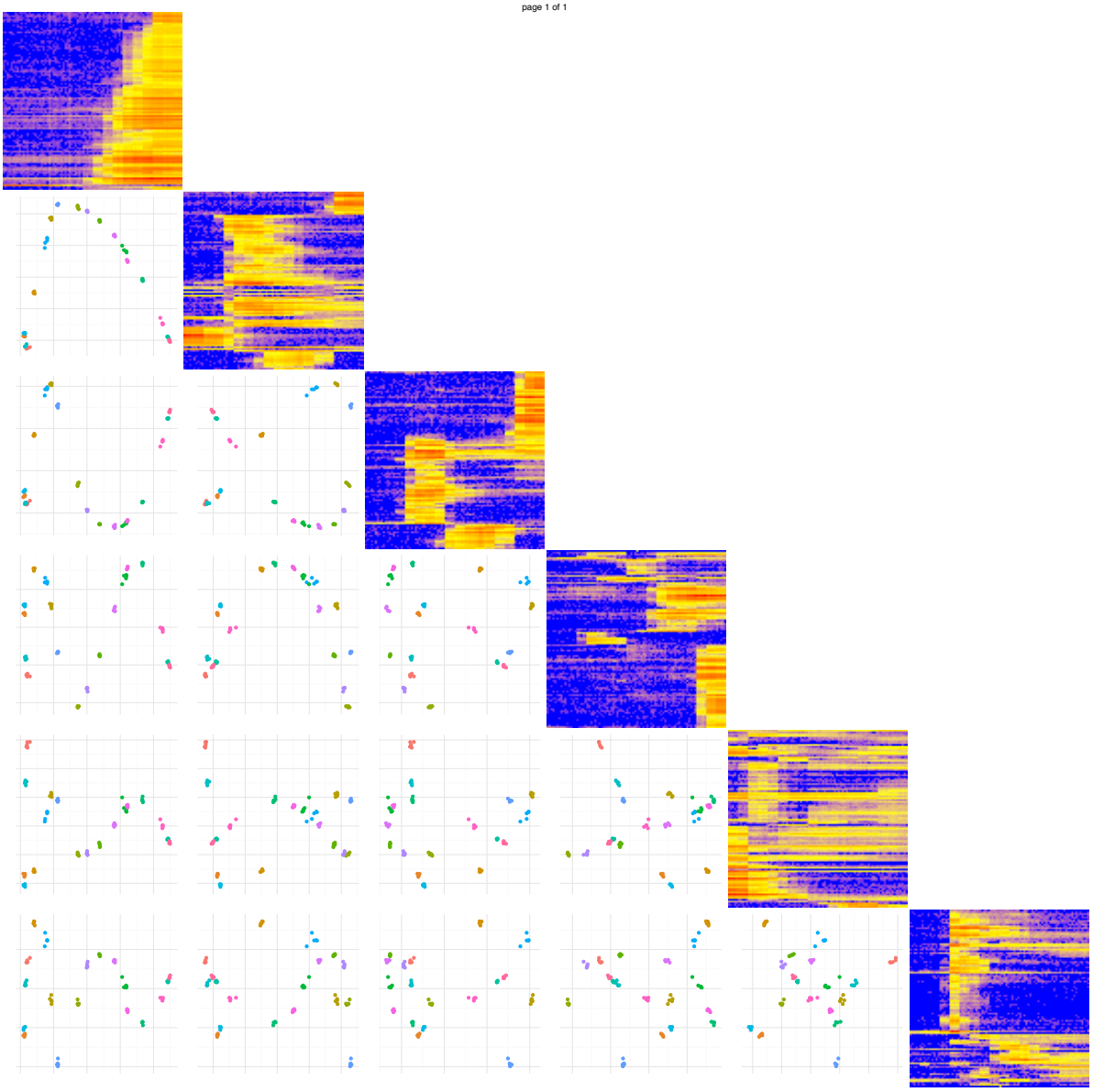


Figure1-2. Matrix PCA plot. Plots showing the first six principal components (PCs) plotted against each other. The components are arranged in a matrix from PC1 (top and left) to PC6 (bottom and right). The plots on the diagonal show the expression profiles of the 100 genes contributing the most to that component. The plots below the diagonal show the appropriate PCs plotted against each other. For example, the plot in the third row and first column is PC1 (x-axis) plotted against PC3 (y-axis).

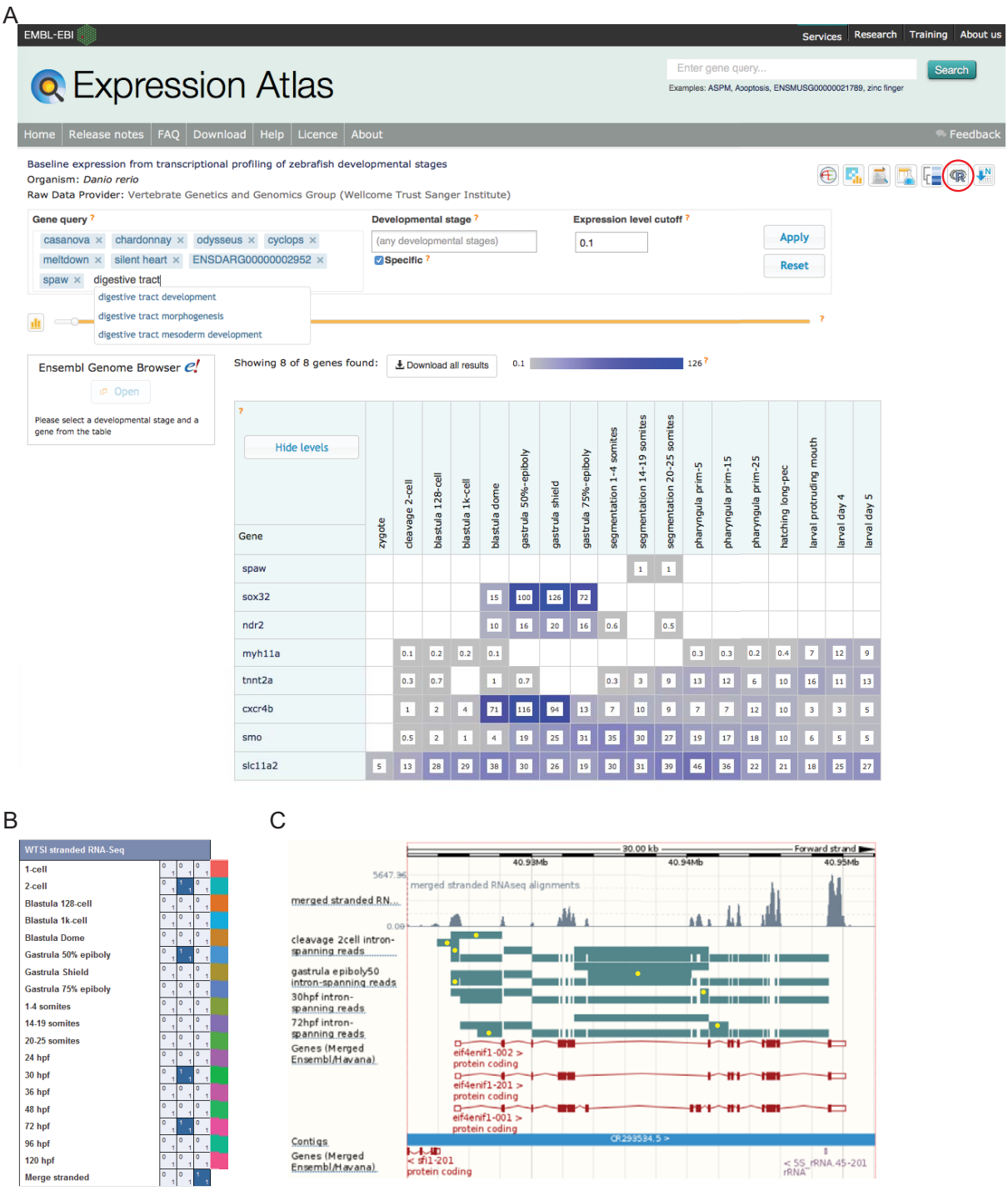


Figure 2. Data Access. (A) A screenshot from Expression Atlas release Dec 2016. Relative gene expression levels across all 18 stages are shown in FPKM. Genes can be searched for using Ensembl IDs, ZFIN gene IDs, gene/allele names and Gene Ontology terms. The data for the selected genes can be downloaded (red circle) as can the entire dataset. (B) A screenshot of Ensembl v87 showing the RNA-seq matrix with four intron tracks plus the merged RNA-seq data from exon read coverage from all 18 stages chosen for display. (C) A screenshot of Ensembl v87

168 showing the region surrounding the gene *eif4enif1* (chr 6:40.922-40.952 Mb). Exon
169 read coverage from the merged 18 stages is shown as grey histograms. Four of the
170 18 intron-defining tracks from the RNA-seq matrix are shown. Introns are coloured in
171 teal. Yellow dots indicate introns identifying a splice variant not shown by any of the
172 three gene models shown in red. Note only the forward strand is shown in full and a
173 gene model overlaps on the reverse strand.

174 Temporal Map of Development

175 To visualise the expression of all genes across development we ordered the genes
176 in our dataset by their stage of maximum expression (Figure 3A). This organises the
177 genes temporally, broadly by expression pattern. These 18 groups of genes were
178 then analysed for enrichment of Zebrafish Anatomy Ontology (ZFA) and Gene
179 Ontology (GO) terms.

180 ZFA enrichment

181 The Zebrafish Anatomy Ontology (ZFA, Van Slyke et al., 2014) is a controlled
182 vocabulary for describing the anatomy and development of zebrafish and is used by
183 the zebrafish model organism database (ZFIN Howe et al., 2013a) to annotate the
184 spatial expression patterns of genes from the literature and large-scale whole mount
185 mRNA *in situ* hybridisation experiments (Rauch et al., 2003; Thisse et al., 2001). The
186 results show enrichments for cell types and tissues that are appropriate to the stages
187 of maximum expression (Figure 3A). There are no significant enrichments of terms
188 during cleavage stages. At early gastrula stages terms such as hypoblast, organiser
189 and hatching gland are enriched with endoderm, yolk and epidermis being enriched
190 at late gastrula. During segmentation stages terms such as muscle, notochord and
191 haematopoietic system are enriched and then at pharyngula and larval stages terms
192 associated with the nervous system and the neural crest become enriched. By
193 contrast, terms for mesodermal and endodermal organs such as kidney, spleen and
194 liver are enriched at later stages (days 4 and 5). This demonstrates that these
195 groupings of genes are consistent with what is already known about zebrafish
196 development, thus validating our dataset.

197 GO enrichment

Enrichment analysis using the Gene Ontology (GO) also shows enrichments of terms in a stage-appropriate manner. At cleavage stages, terms linked to cell division, DNA repair, protein turnover and chromatin modification are enriched, consistent with the fast rate of cell division as well as the large changes in chromatin structure that occur at early stages. Subsequently, terms associated with determination of the embryonic axes and specification of the major tissue divisions are enriched. For example, genes linked to anterior-posterior, dorsal-ventral and left-right patterning are enriched from dome until prim-25 as well as genes annotated to terms such as gastrulation, heart development, muscle cell differentiation and brain development. At later stages, more specific terms such as synaptic transmission, potassium ion transport and glutamate receptor signalling are enriched. Coordinated expression of blood clotting factor genes leads to the appearance of blood coagulation as an enriched term at day 4.

Of the 23,642 genes 5,024 are uncharacterised (i.e. genes with automatically generated names such as zgc:* and si:ch*). Interestingly, at dome and gastrula stages, the proportion of these unnamed genes is much higher (29.2-39.4%) than at other stages (14.4-26.5%; Figure 3A). This could indicate a lack of orthologues and therefore that zebrafish gastrulation is enriched for the expression of genes exclusive to this species. However, we find that the proportion of unnamed genes with mammalian or teleost-specific orthologues is constant across all 18 stages (Supplemental Figure 3-1). This suggests that there is a large number of genes involved in vertebrate gastrulation that are yet to be characterised. The complete list of clusters along with the ZFA/GO term enrichment are provided in Supplemental Data Files 3-5.

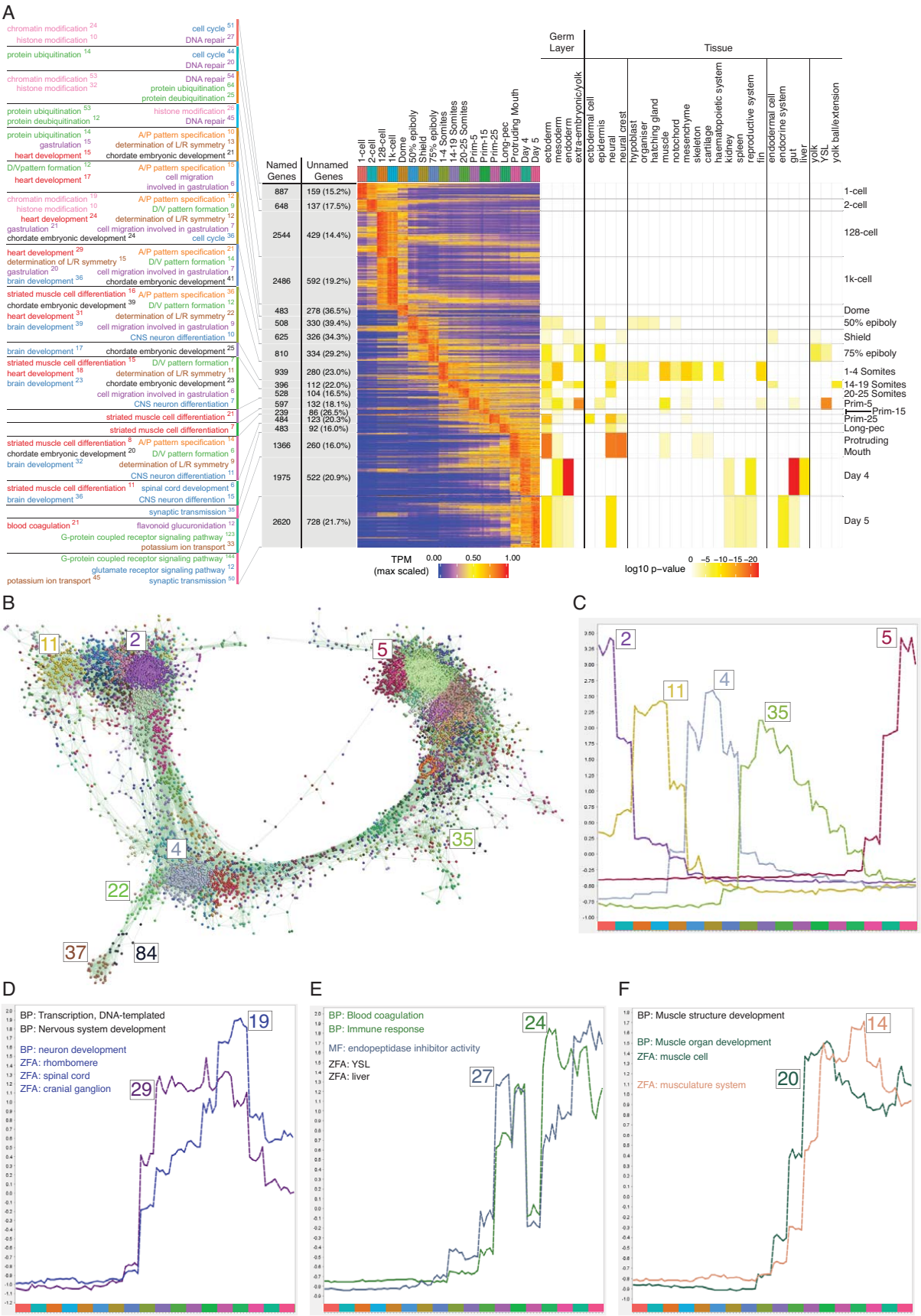


Figure 3. Clustering of expression patterns. (A) Heatmap showing the expression profiles of all the genes with appreciable expression in at least one stage. The

expression values (TPM) are scaled to the maximum expression for that gene across the time course and genes are organised by which stage the maximum expression occurs at. To the right is a heatmap of ZFA enrichments based on the genes assigned to each stage. To the left are a selection of the GO enrichments. The numbers of genes assigned to each GO term are shown as superscripts. Also shown are the numbers of named and unnamed genes assigned to each stage. (B) Network diagram showing the clustering produced by MCL on a network graph generated by linking genes that have a Pearson correlation coefficient of > 0.94 and removing clusters with five genes or fewer. The numbers show the positions in the network diagram of the clusters shown in C and the ZnF clusters shown in Figure 4. (C-F) Example expression profiles displayed as lines showing the mean of mean-centred and variance-scaled normalised counts. (C) Example clusters showing the progression of expression peaks through development. (D-F) Expression profiles of example clusters with interesting GO/ZFA enrichments. Terms in black apply to all clusters in the diagram whereas terms specific to one cluster are shown in that colour.

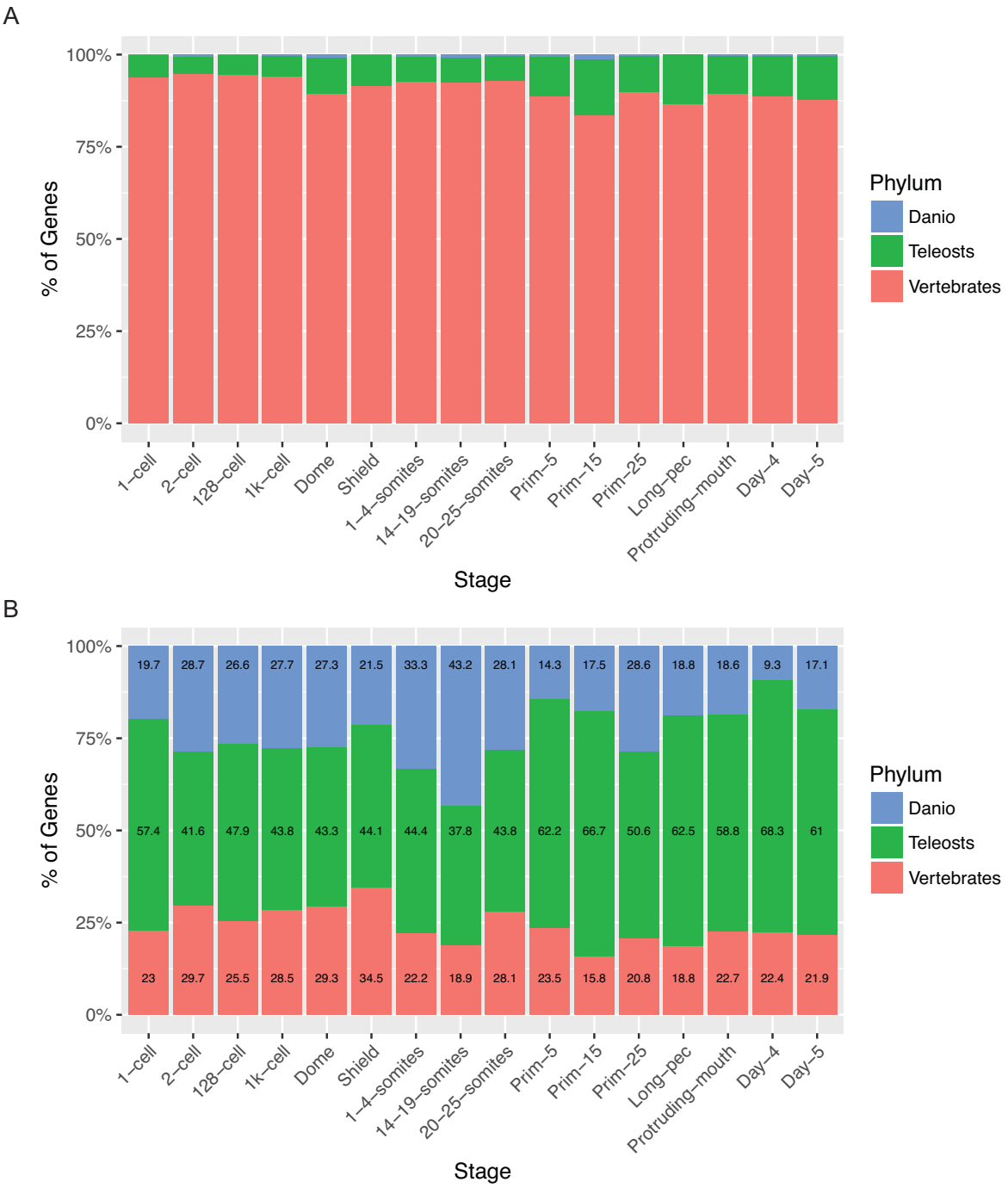


Figure 3-1. Relationship of named and unnamed genes to other species. Stacked bar charts showing the proportions of genes with orthologues at different taxonomic levels for Named (A) and Unnamed (B) genes. Orthologues for genes were obtained from the Ensembl Compara database and genes were then categorised based on whether they have orthologues in the following species; Vertebrates (bottom bar, red) – *Mus musculus*, *Rattus norvegicus* and *Homo sapiens*; Teleosts (middle bar, green) – *Astyanax mexicanus*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Gadus morhua*.

251 Genes with no identifiable orthologue in any of these species were classed as *Danio*
252 specific (top bar, blue).

253 Graph-based clustering produces compact clusters of
254 highly biologically related genes

255 While the maximum expression stage groupings provide a high-level, temporally
256 organised view of development, smaller more specific groupings would provide more
257 insight especially into the role of uncharacterised genes. To get smaller clusters that
258 are more likely to be highly biologically related we used the BioLayout *Express*^{3D}
259 software (http://www.biolayout.org/) (Theocharidis et al., 2009) to cluster and
260 visualise the dataset (Figure 3B). The software first builds a network graph in which
261 genes are nodes with edges between those whose expression is correlated above a
262 given threshold. It then uses the Markov Cluster Algorithm (MCL, Enright et al.,
263 2002; van Dongen, 2000) to find strongly connected components within the graph.
264 These are clusters of genes that are more highly connected with each other than
265 with the rest of the graph. This produces 252 clusters, the vast majority of which
266 (238) have fewer than 100 genes (Figure 3B). Broadly speaking, the clusters on the
267 left-hand side of the graph represent genes expressed at early stages, whereas
268 towards the right of the graph the expression profiles are progressively later in
269 development (Figure 3B-C). For example, cluster 2 contains genes that are
270 maternally supplied and then degraded, cluster 11 has genes that accumulate after
271 the 2-cell stage (possibly by control of polyadenylation) and are then cleared, cluster
272 4 genes are expressed at the time of zygotic genome activation, cluster 35 genes
273 are expressed during segmentation stages and genes from cluster 5 are not
274 expressed until the last time points in the study (Figure 3C).

275 We again used GO and ZFA enrichment to analyse the composition of the genes
276 within the clusters. For example, cluster 19 contains large numbers of transcription
277 factors from the *dlx*, *hox*, *irx* and *pax* gene families that are expressed from 1-4
278 somites onwards and is associated with GO terms for transcription and nervous
279 system development and ZFA terms such as rhombomere, spinal cord and cranial
280 ganglion (Figure 3D). Cluster 29, with similar GO enrichments, has an overlapping
281 expression profile but an earlier and broader peak. It contains genes such as *nkx2.1*,

282 *gli3*, *sox1b* and *hoxb* genes. Clusters 24 and 27 show a sharp dip in expression at
283 36 hpf and contain complement, clotting factor genes and serpins. Both clusters
284 have significant enrichments for the terms blood coagulation and endopeptidase
285 inhibitor activity (Figure 3E). (Figure 3F). The information on cluster assignments and
286 GO/ZFA enrichments are provided in Supplemental Data Files 6-9.

287 A burst of transcription of highly related Zinc Finger 288 proteins during zygotic genome activation

289 During examination of the clusters produced by BioLayout, we noticed several
290 clusters that appeared to be enriched for a family of zinc finger (ZnF) proteins
291 located on chromosome 4 (Howe et al., 2016). To test systematically if any
292 chromosomes were over-represented in any of the clusters, we performed a binomial
293 test for each chromosome for each cluster that contained more than five genes from
294 that chromosome. Table 1 shows the clusters and chromosomes that are
295 significantly enriched by more than twofold. This identifies four clusters (Figure 3B,
296 clusters 4, 22, 37 and 84) with more chromosome 4 genes than would be expected
297 by chance given the size of the clusters. Most of the chromosome 4 genes within
298 these clusters are predicted to encode proteins with ZnF domains (Supplemental
299 Data File 10).

300 The expression profiles of these ZnF genes show a burst of expression from dome to
301 75% epiboly (4.3-8 hpf) and several of them show appreciable RNA levels at the
302 128- and 1000-cell stages (Figure 4A-B, Supplemental Figure 4-1). Cluster 37 shows
303 a particularly sharp peak with a very large increase from the 1000-cell stage to dome
304 and an almost equal decrease from dome to 50% epiboly. This suggests that these
305 ZnF genes are co-ordinately expressed at the onset of zygotic transcription. To
306 investigate whether this is a global effect on all the genes on the long-arm of
307 chromosome 4 we looked at the expression profiles of NLR family genes, another
308 large family of proteins present on chromosome 4. These genes are present in the
309 same region of chromosome 4 as the ZnF genes although their distribution is slightly
310 different, with more NLR genes towards the end of the chromosome (Figure 4F-G,
311 Supplemental Figure 4-2). Of the 312 NLR genes present in the Ensembl version 85

genebuild, 149 have appreciable expression during this time course and they are not expressed in the same pattern as the ZnF genes (Figure 4C).

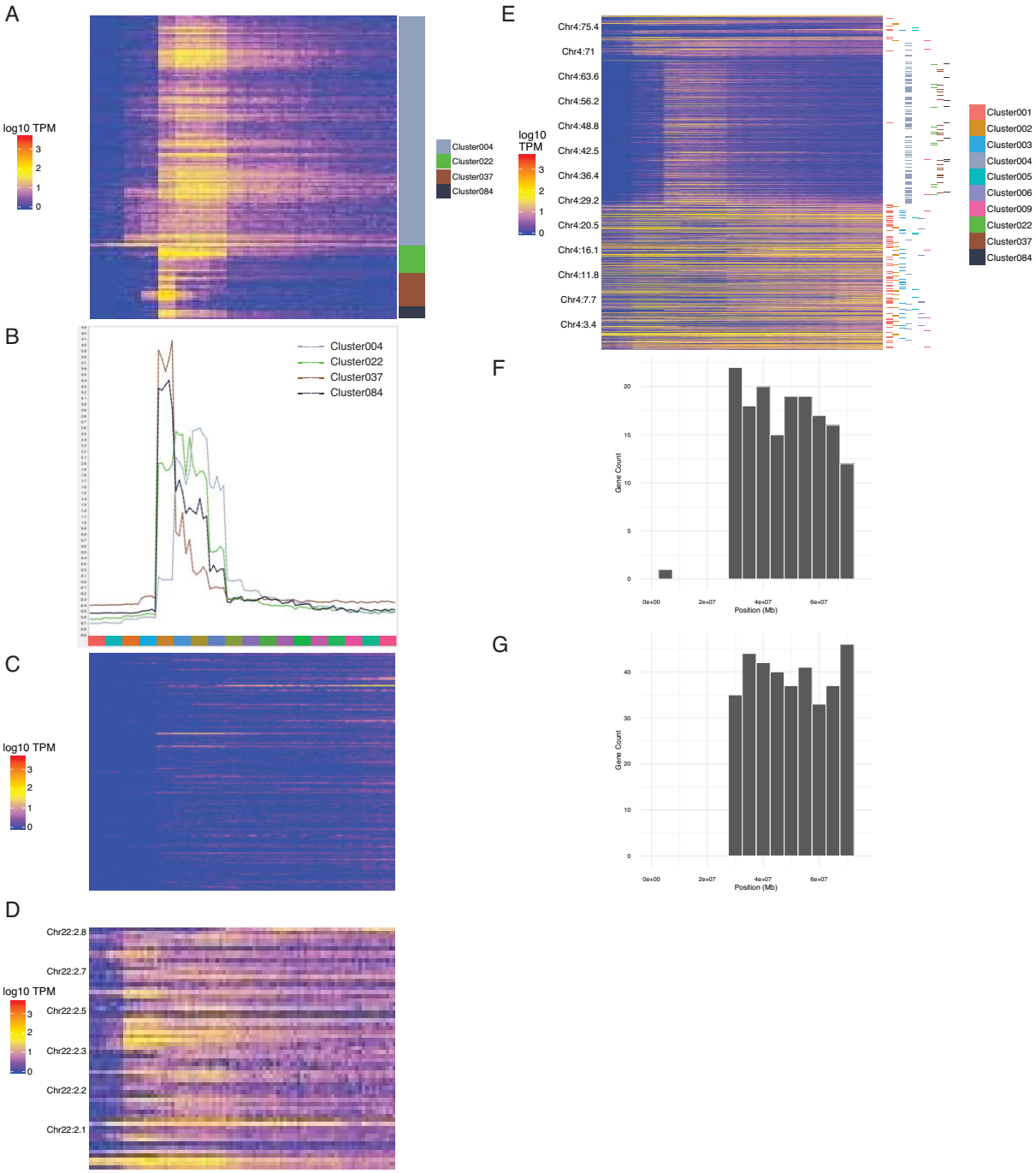


Figure 4. ZnF genes expressed at the onset of zygotic transcription. (A) Heatmap showing the expression profiles (log10 TPM) of the ZnF genes from clusters 4, 22, 37 and 84. (B) Expression profiles of clusters 4, 22, 37 and 84 shown

318 as average expression (mean centered and variance scaled). (C) Heatmap showing
319 the expression profiles (log10 TPM) of the NLR genes present in GRCz10. (D)
320 Heatmap showing the expression of a cluster of ZnF genes on chromosome 22
321 displayed in chromosome order. (E) Heatmap showing the expression of all genes
322 on chromosome 4 (log10 TPM) in chromosomal order with cluster assignments
323 shown on the right. (F-G) Histograms showing the distribution of ZnF (F) and NLR
324 (G) genes on chromosome 4.
325

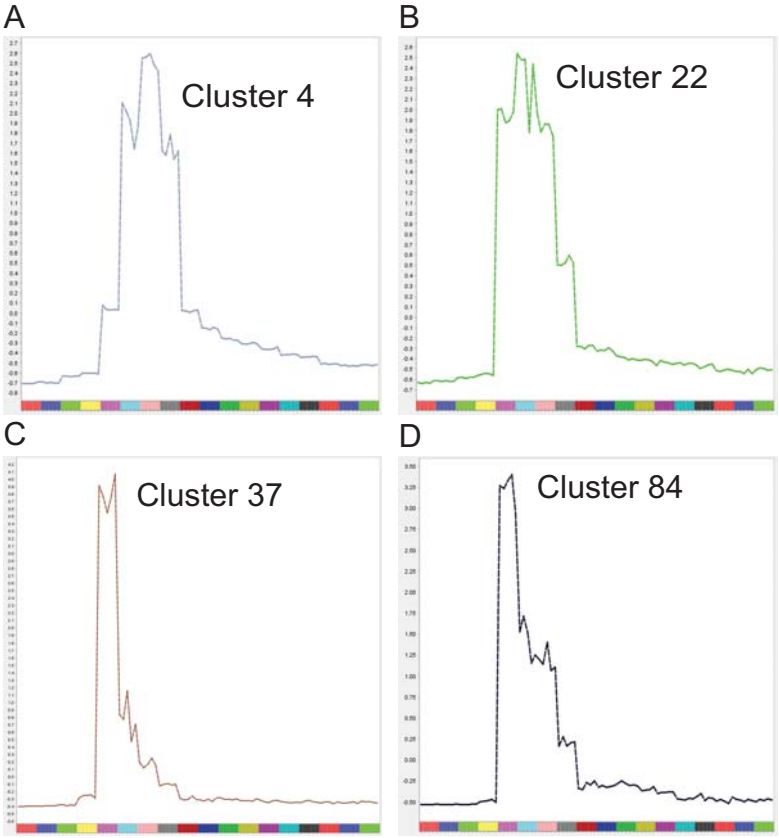


Figure 4-1. BioLayout cluster profiles. Individual expression profiles of clusters 4 (A), 22 (B), 37 (C) and 84 (D) shown as average expression (mean centered and variance scaled).

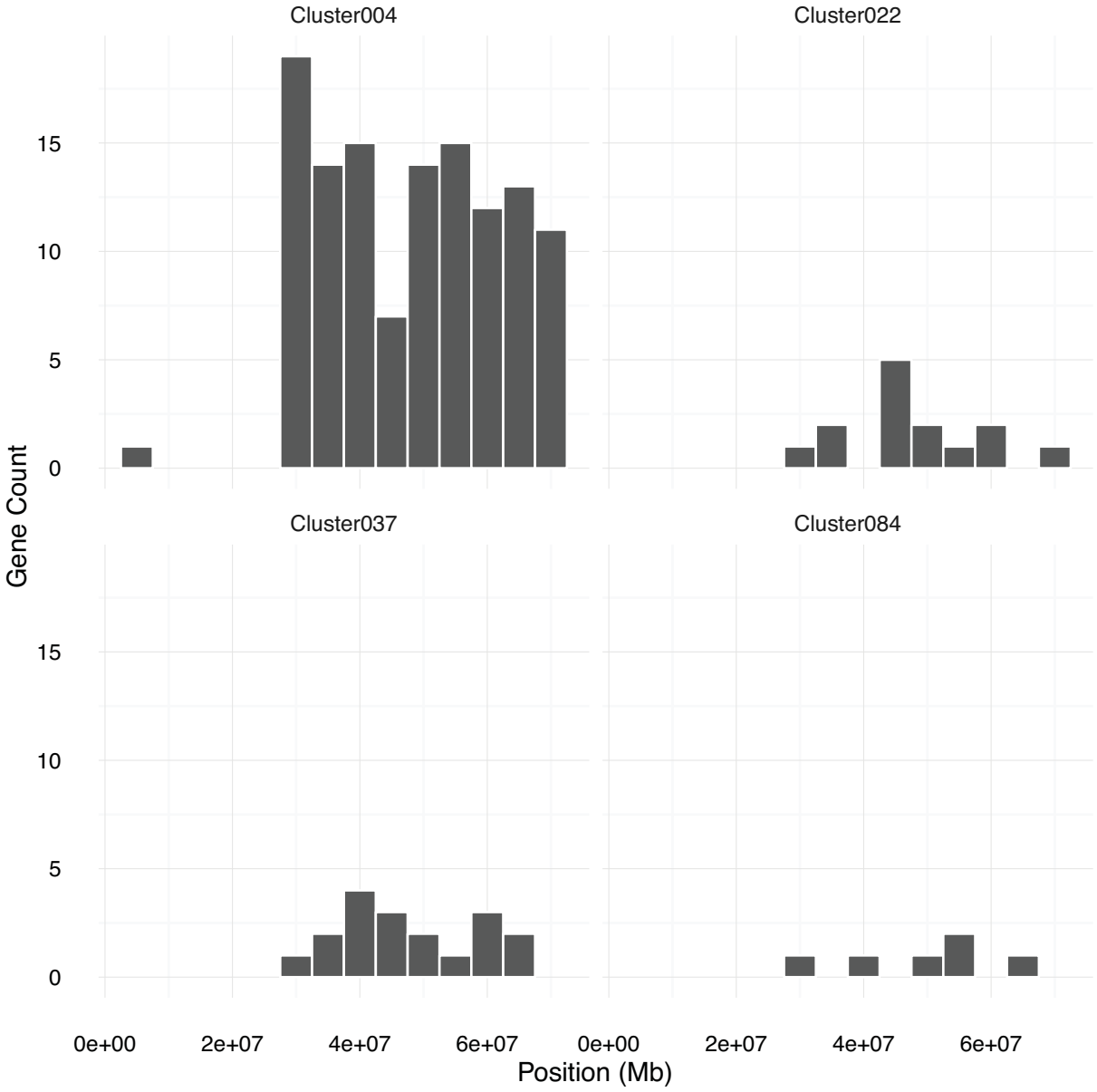


Figure 4-2. Chromosomal distribution of the chromosome 4 ZnF genes. Individual histograms split by cluster showing the chromosomal distribution of the ZnF genes from clusters 4, 22, 37 and 84.

This effect can even be seen on the chromosomal scale. A plot of the expression profiles of all the genes across chromosome 4 in chromosomal order shows a domain on the long arm that is enriched for these ZnF genes (Figure 4E) that share a similar expression pattern.

341 **Table1**

Cluster	Chromosome	Count	Expected	log2FE	Adjusted pvalue
Cluster023	Chr1	9	2.25	2.00	1.161e-02
Cluster004	Chr4	129	47.29	1.45	7.434e-24
Cluster022	Chr4	14	5.03	1.48	1.156e-02
Cluster037	Chr4	23	3.35	2.78	9.033e-14
Cluster084	Chr4	9	1.48	2.61	1.659e-04
Cluster074	Chr12	8	0.48	4.05	6.337e-07

342

343 Table 1. BioLayout clusters with significant chromosomal enrichments (adjusted p-value <
344 0.05 and log2 fold enrichment [log2FE] > 1).

345

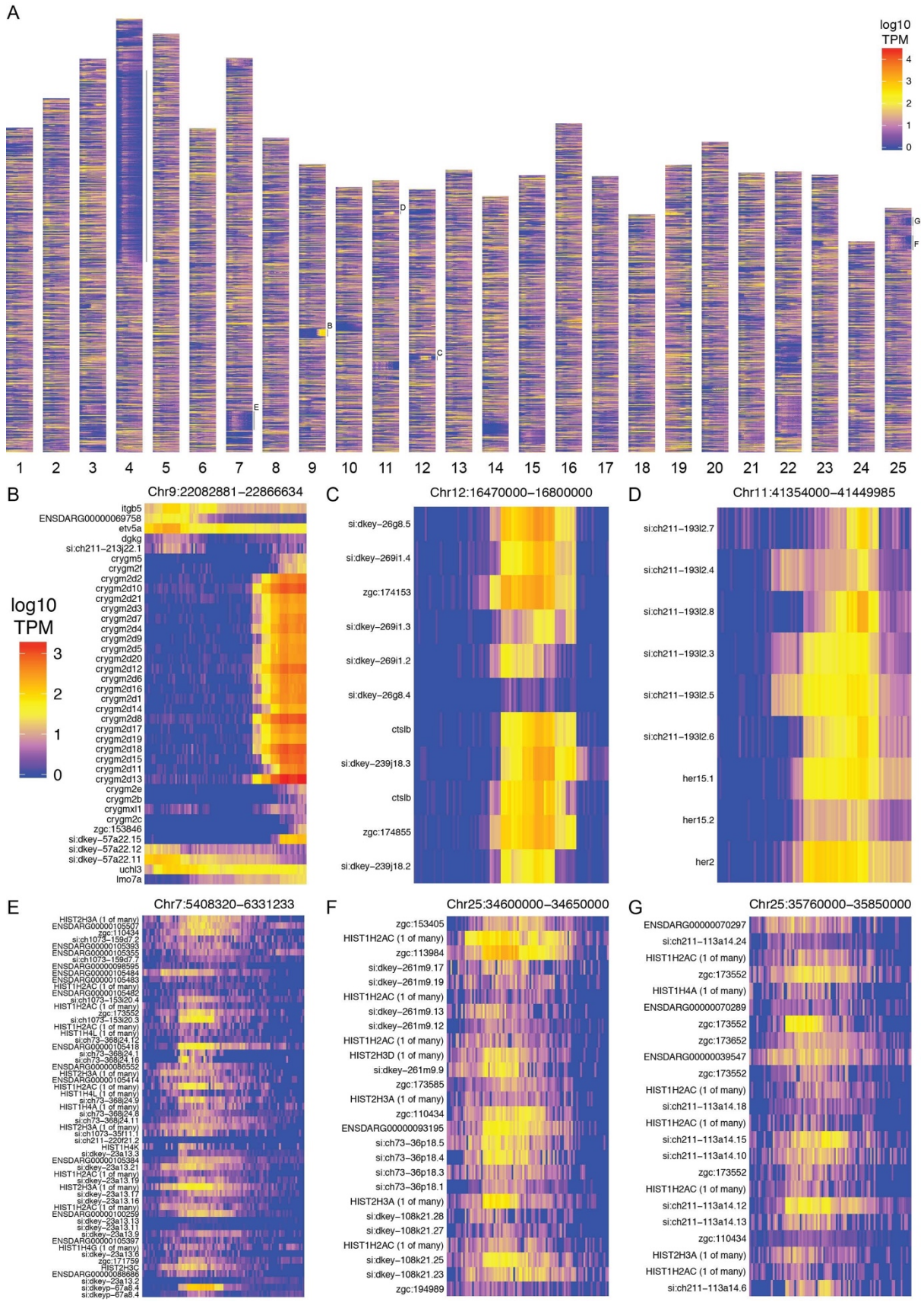
346 Chromosomal expression domains

347 Since the expression pattern of the chromosome 4 ZnF genes is obvious on a
348 chromosome level plot, we wondered if it was possible to find any other
349 chromosomal regions where the genes share a common expression pattern. To
350 investigate this, we plotted the expression of all the genes for each chromosome in
351 chromosomal order (Supplemental Figure 4-3A). This led to the identification of
352 another region with genes containing zinc finger domains, this time on chromosome
353 22 (Figure 4D). These genes begin to be expressed at a similar stage to the
354 chromosome 4 ZnF genes, although they tend to be more widely expressed over the
355 time course (Figure 4D).

356 There are also several other domains of co-expressed genes that can be clearly
357 seen at the chromosome level, such as the gamma crystallin family of genes on
358 chromosome 9, that are all highly expressed at the end of the time series
359 (Supplemental Figure 4-3B).

360 Interestingly, there is a small region on chromosome 12 where the genes are all
361 expressed from late gastrula through to the end of the pharyngula period

362 (Supplemental Figure 4-3C). The genes in this region all have clone-based names
363 apart from *ctslb* (ENSDARG00000039173, also known as *hatching gland 1*, *hgg1*).
364 All of them appear in the same BioLayout cluster (cluster 74), which has an
365 overrepresentation of genes on chromosome 12 (Table 1). This appears to be a fish-
366 specific expansion of cathepsin L. Similar expansions have happened in multiple
367 different lineages including rodents
368 (http://www.ensembl.org/Mus_musculus/Gene/SpeciesTree?db=core;g=ENSMUSG0
369 0000021477;r=13:64363214-64370306;t=ENSMUST00000021933).
370 On chromosome 11 there is a cluster containing hairy and enhancer of split-related
371 genes and a set of unnamed genes all expressed from late gastrula through
372 segmentation and pharyngula stages (Supplemental Figure 4-3D). Finally, two
373 clusters of genes containing histone domains can be seen on chromosomes 7 and
374 25 (Supplemental Figure 4-3E-G).



375

376 **Figure 4-3. Chromosomal expression domains.** (A) Heatmaps showing the

377 expression profiles (log10 TPM) of genes in chromosomal order. The unlabelled line

378 next to chromosome 4 marks the location of the ZnF genes shown in Figure 4. (B-G)
379 Expression profile heatmaps of the regions indicated by letters in (A).

380 Conservation/Divergence of the expression profiles of 381 paralogous genes

382 Teleost fish underwent a third round of whole genome duplication compared to the
383 two rounds of tetrapods. During rediploidisation, most of the duplicate genes were
384 lost, however a significant proportion were retained and are now present as
385 paralogous pairs. Paralogous genes can have conserved expression patterns but
386 often the expression patterns have diverged so that the paralogues function at
387 different times or in different tissues. To investigate whether we can see this in our
388 data, which has only stage and not tissue specificity information, we produced a list
389 of one-to-one teleost-specific paralogues pairs using data from Ensembl (3144 gene
390 pairs with expression in our data).

391 We calculated the Pearson correlation coefficients between the normalised counts
392 for each paralogous pair as a measure of how similar the expression profiles are. A
393 histogram of the correlation coefficients shows a distribution skewed to high positive
394 correlations (Figure 5A) when compared to a random sample of gene pairs of similar
395 size (Figure 5B). Despite this skew, there is a significant proportion of the distribution
396 with low or negative correlation coefficients, suggesting that this may represent a set
397 of genes whose expression patterns have diverged (50% of the distribution is below
398 0.45).

399 Examples of highly correlated and negatively correlated genes are shown in Figure
400 5C-G and Supplemental Figure 5-1. To get an estimate of how many of these gene
401 pairs have divergent spatial expression patterns, we used the ZFA annotation
402 provided by ZFIN. There are 683 gene pairs where there is anatomical expression
403 information for both genes. For each pair we counted up the intersection of all the
404 pairs of stage ID and anatomy term that make up the expression annotations. For
405 the pairs studied, 83.7% (572 pairs) have an intersection of less than 50%.

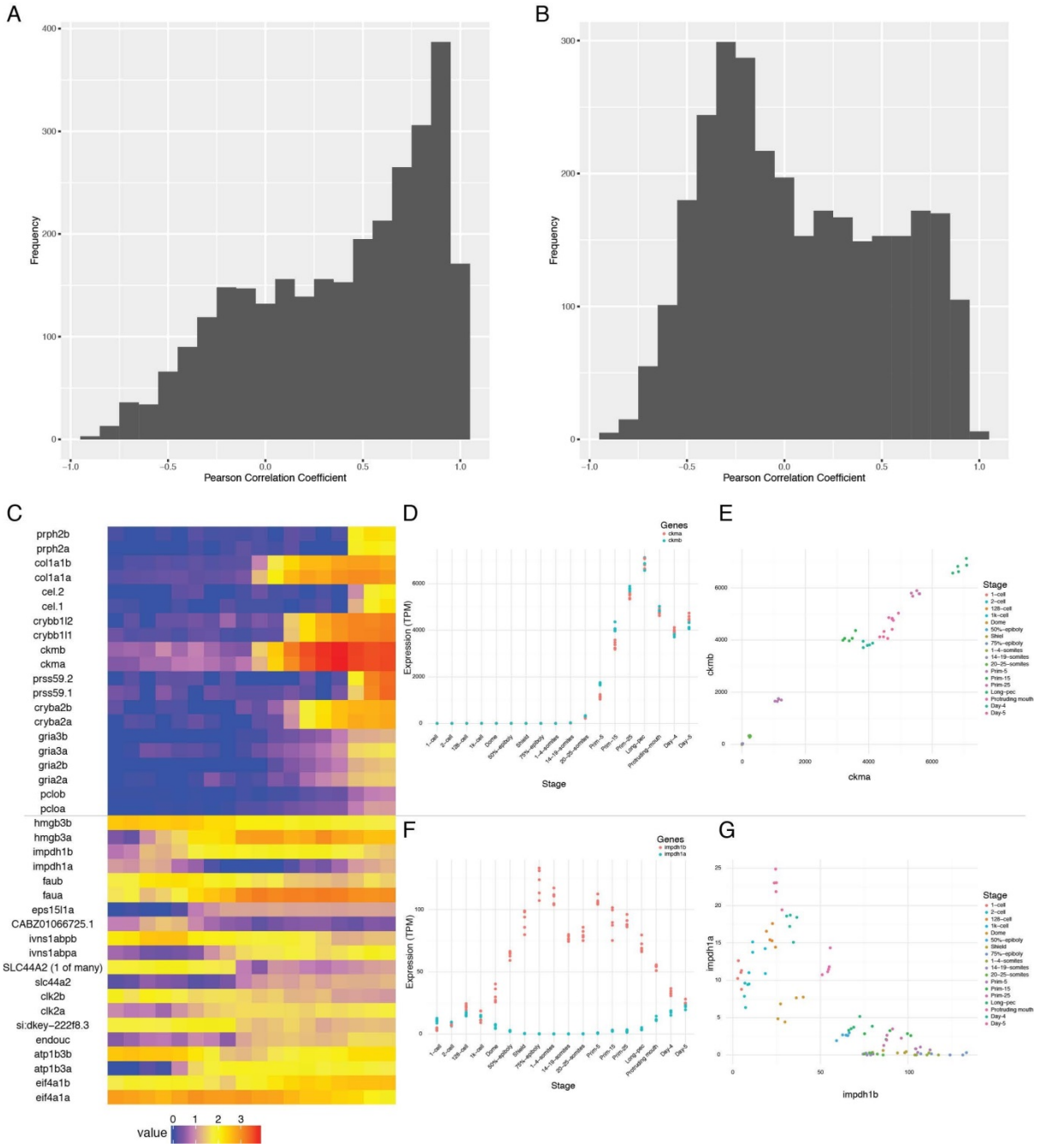


Figure 5. Conservation/divergence of paralogous genes. (A-B) Histograms of Pearson correlation coefficients among one-to-one paralogous gene pairs (A) and a random sample of genes (B). (C) Heatmap showing the expression of the gene pairs with the top 10 highest correlation coefficients (top half) and the top 10 most negative correlation coefficients (bottom half). (D-E) Plots showing the expression of *ckma/b* by stage (D) and plotted against each other (E). (F-G) Plots showing the expression of *impdh1a/b* by stage (F) and plotted against each other (G).

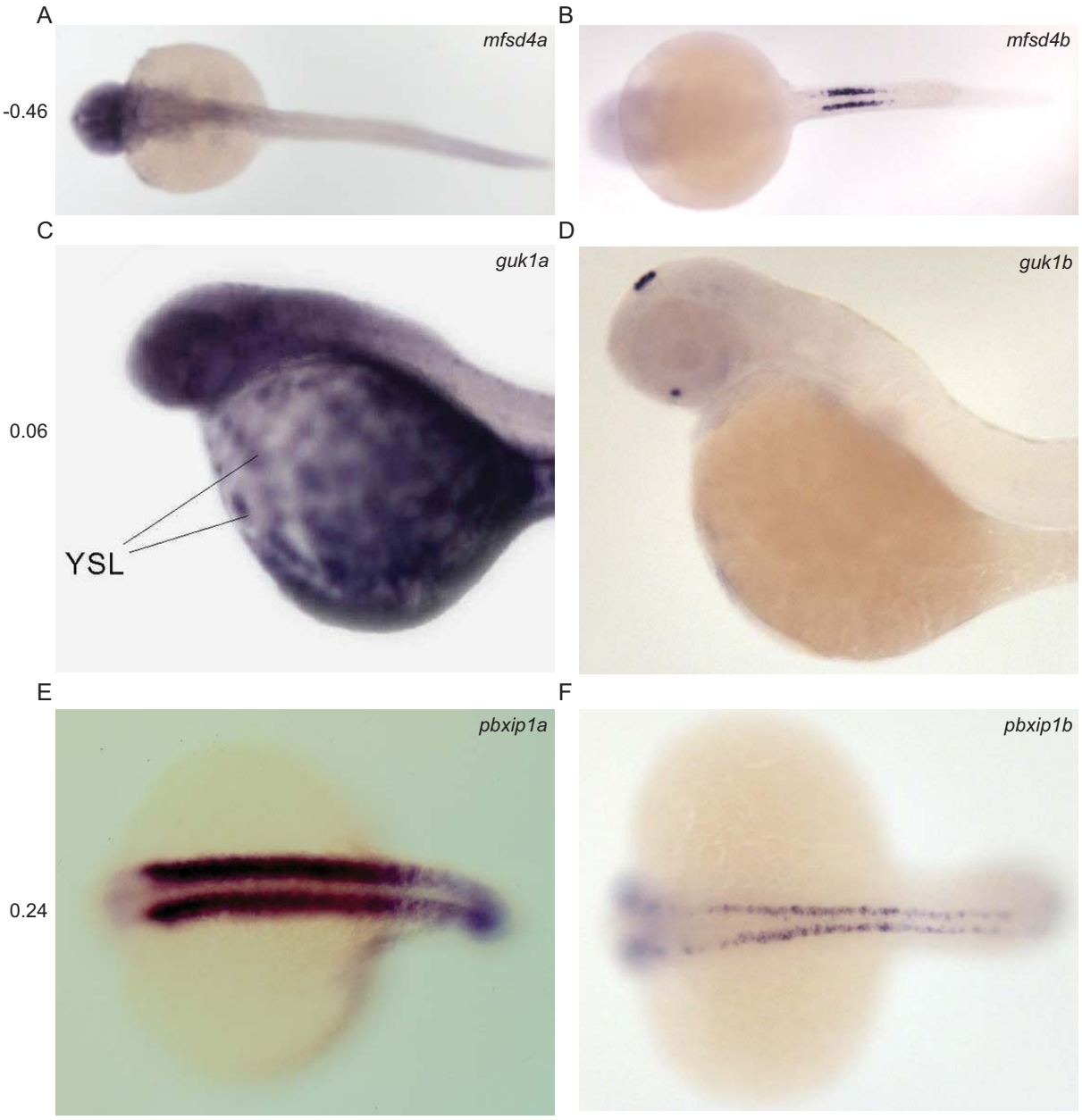


Figure 5-1. Divergent expression of paralogous pairs. Images from ZFIN (ZFIN.org) showing examples of paralogous gene pairs with diverged expression patterns. The Pearson correlation coefficient between the RNA-seq expression profiles of each pair is shown to the left. (A-B) *mfsd4a/b*. (C-D) *guk1a/b*. (E-F) *pbxip1a/b*.

Alternative 3′ end use

To catalogue alternative 3′ end usage across development we produced 3′ pulldown data for the same time points as the RNA-seq. There are multiple possible consequences of alternative 3′ ends, such as different/extra 3′ UTR sequence

(Figure 6A). This could provide different binding sites for RNA-binding proteins or regulation by microRNAs. For example, one transcript may contain a microRNA binding site lacking in another, changing post-transcriptional regulation. Other possibilities include different/extra coding sequence at the 3′ end of transcripts, leading to proteins that contain different domains and therefore different molecular functions.

Across all 423 samples over all the stages, there were 253,627 regions identified, where each region may be associated with multiple 3′ ends. However, many of these are false positives due to poly(T) pulldown from genomic poly(A) tracts within transcripts. These regions are then filtered by various criteria such as distance to an annotated gene and genomic context in an attempt to remove artefactual regions (Figure 6B and Methods), which reduces the set to 37,724 regions. For this analysis, we increased the stringency of filtering to reduce the inclusion of false-positive ends and produce a high-confidence set of 3′ ends to work from. This stricter set contains 8358 regions. Once the regions have been associated with annotated genes, 1551 genes have two or more 3′ ends.

This high-confidence set has allowed us to define 3′ ends of genes that are alternatively used throughout development. For example, *pdlim5b* (ENSDARG00000027600) has three alternative 3′ ends which are used at different times during development (Figure 6C-E). During cleavage and gastrula stages, one particular end is used (end 2 in Figure 6C), but all three are observed during segmentation stages. Ends 1 and 2 would be predicted to produce the same protein with different 3′ UTRs, allowing for differential post-transcriptional regulation. However, end 3 is predicted to produce a protein which lacks the C-terminal zinc finger domain, allowing for this alternative transcript to have a completely different molecular function.

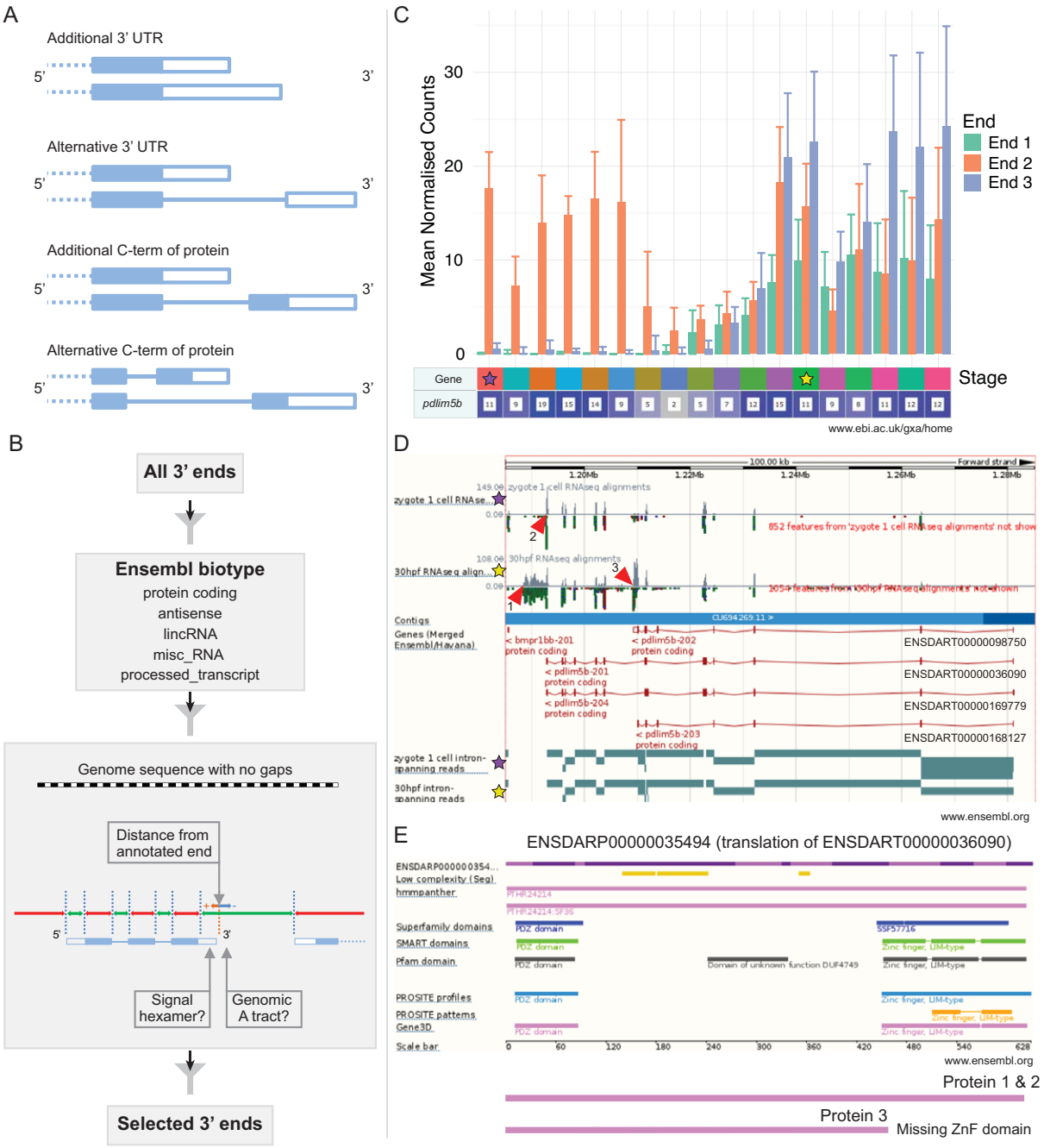


Figure 6. Alternative 3' end use during development. (A) Diagrams showing the alternative possible outcomes of differential 3' end usage. Only the 3' end of the gene is shown where the exons are boxes, the introns are represented as lines and filled boxes are coding. The dotted line represents the remaining 5' end of the gene. (B) Flowchart showing the DeTCT 3' end filtering strategy. (C) Graph showing the mean normalised counts by stage for the three alternative 3' ends (shown in D) of the *pdlim5b* gene. Error bars represent standard deviations. Underneath is a screenshot from Expression Atlas (Sept 2016) showing the overall FPKM for

462 *pdlim5b*. The stages for which exon and intron data are shown in D are denoted with
463 stars. (D) Screenshot from Ensembl (release 87) showing the *pdlim5b* gene
464 (10:1185001-1285000). The purple stars mark exon read counts for the 1-cell stage
465 at the top and intron spanning read counts at the bottom. Similarly, prim-15 counts
466 are marked by yellow stars. Red arrowheads point to the genomic positions of the
467 three transcript ends. Note 3' UTR resulting in end 1 is not annotated in this gene
468 build. Four transcripts are labelled on the right hand side. (E) A snapshot of Ensembl
469 release 87 protein view showing *pdlim5b* transcript ENSDART00000036090.6 from
470 N-terminus to C-terminus. The protein domains missing if the transcript finished at
471 end 3 are shown below.

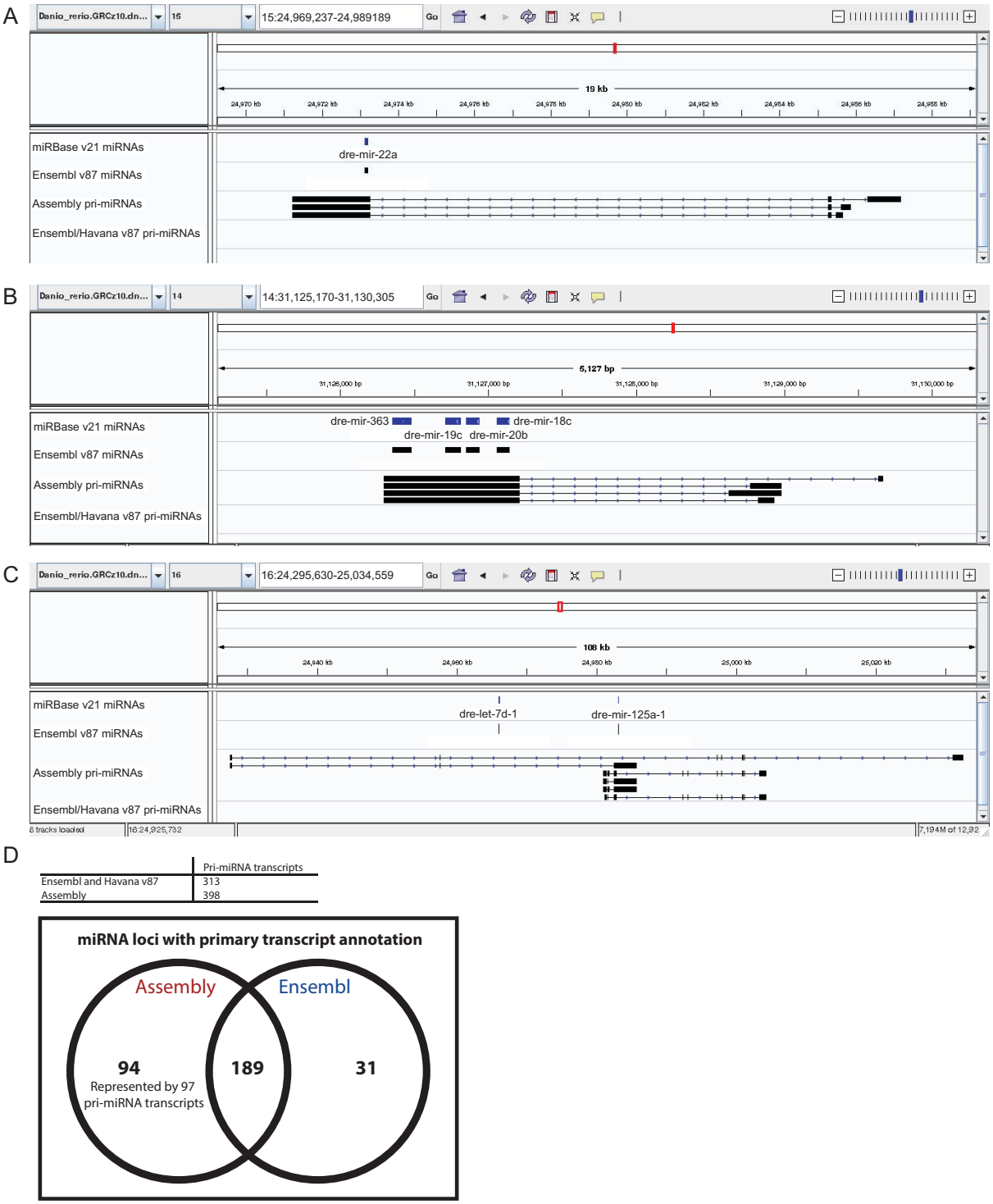
472 Novel primary miRNA transcripts

473 Along with cataloguing 3′ ends, we have performed RNA-seq on mutants for two
474 genes involved in microRNA (miRNA) maturation, in order to annotate the
475 polyadenylated primary miRNA transcripts. Mature miRNAs are released from longer
476 primary transcripts (pri-miRNAs) by a series of processing steps (Denli et al., 2004;
477 Gregory et al., 2004; Grishok et al., 2001; Hutvagner et al., 2001). First, precursor
478 hairpins (pre-miRNAs) are sliced from the host transcript in the nucleus by the
479 microprocessor complex, consisting of the proteins Drosha (Rnasen) and Dgcr8
480 (Pasha). Subsequently the hairpin is cut by Dicer1 in the cytoplasm, freeing the
481 miRNA duplex.

482 Although much effort has been dedicated to sequencing and annotating the mature
483 miRNAs themselves, the annotation of many primary miRNA genes is likely hindered
484 by their rapid nuclear processing. Reasoning that the depletion of *drosha* and *dgcr8*
485 may enrich the transcriptome for unprocessed pri-miRNAs (Chang et al., 2015), we
486 sequenced RNA derived from *dgcr8* and *drosha* wild-type, heterozygous and
487 homozygous mutant embryos to determine whether this would allow us to improve
488 the current pri-miRNA annotation. Following the assembly of the transcriptome, we
489 were able to better define a large number of novel pri-miRNA loci not annotated in
490 the current Ensembl release (v87) (Fig. 7). The v87 release includes 313 pri-miRNA
491 transcripts overlapping 220 annotated miRNA loci. Our assembly contains 398 pri-
492 miRNA transcripts which overlap 283 miRNA loci, 94 of which are not currently
493 associated with a pri-miRNA (Figure 7D). These will be incorporated into a future

494 Ensembl gene set update.

495



496

497 **Figure 7. Annotation of primary miRNA transcripts.** (A-C) A selection of miRNA
498 loci for which pri-miRNA annotation is not available in Ensembl and Havana v87. IGV
499 screen shots showing pri-miRNA transcripts overlapping mir-22a (A), mir-363, mir-
500 19c, mir-20b and mir-18c (B), and mir-125 (C). Tracks shown are miRNA

501 annotations from miRBase and Ensembl, newly assembled pri-miRNA transcripts
502 and Ensembl pri-miRNAs. (D) Table and Venn diagram showing the numbers of
503 assembled pri-miRNA transcripts and the overlap with of Ensembl miRNAs with an
504 annotated pri-miRNA.

505 # Discussion

506 We have produced a high quality map of zebrafish mRNA expression during
507 development. It represents the first detailed temporal mRNA reference in zebrafish
508 and enhances our understanding of the transcriptional diversity underlying vertebrate
509 development.

510 Ordering the gene set by the stage at which they have their maximum expression
511 gives a detailed overview of development, showing the molecular correlates of
512 morphological changes within the embryo. This arrangement of the dataset reveals
513 that there is a much larger fraction of unnamed genes in the stages from dome to
514 75% epiboly. We find that this peak in unnamed genes during gastrulation is not
515 driven by an enrichment for teleost or zebrafish-specific genes, as the proportion of
516 unnamed genes with mammalian orthologues is at the same level during these
517 stages as during the rest of development. This implies that there is a large number of
518 genes involved in vertebrate gastrulation that have yet to be characterised.

519 Having a large number of stages and replicates allows us to cluster genes by their
520 expression profile across all of development using a graph-based method. This splits
521 the assayed genes into 252 different expression profiles containing at least 5 genes
522 each (10,363 of the genes are assigned to clusters). GO/ZFA enrichment on these
523 clusters allows us to associate unnamed genes with known processes.

524 One intriguing set of clusters contains an over-representation of genes on the long
525 arm of chromosome 4, a region considered to be constitutive heterochromatin. The
526 vast majority of these genes encode proteins with zinc finger domains. The clusters
527 show expression profiles that rise sharply around the time of zygotic genome
528 activation (ZGA) and decrease at the latest after 75% epiboly. Given this, it is
529 tempting to speculate that these genes have a role during the activation of the
530 zygotic genome. In *Drosophila* the zinc finger gene Zelda (zld) is involved in zygotic
531 genome activation by early and widespread binding to promoters and enhancers
532 ahead of their genes' activation during the maternal-zygotic transition (Harrison et
533 al., 2010; Harrison et al., 2011; Liang et al., 2008). In zebrafish, *pou5f3*, *nanog* and
534 *soxB* genes have been implicated in activating specific genes during ZGA (Lee et al.,
535 2013; Leichsenring et al., 2013), but no general factor comparable to Zelda has been
536 identified yet. It is possible that the role of the ZnF protein family genes on

537 chromosome 4 is to set up or maintain open chromatin in a genome-wide way to
538 allow for the rapid activation of the first zygotic genes. This is supported by the
539 finding that these clusters also contain four unnamed genes, also found on
540 chromosome 4, that are predicted to be H3K36 methyltransferases
541 (ENSDARG00000104681, ENSDARG00000091062, ENSDARG00000076160,
542 ENSDARG00000103283). H3K36 methylation is generally, albeit not exclusively,
543 associated with active euchromatin (Wagner and Carpenter, 2012). Another option is
544 that they could be responsible for negatively regulating areas of open chromatin to
545 stop inappropriate activation of transposable elements during ZGA. In mammals,
546 suppression of retrotransposons is mediated by the rapidly evolving KRAB zinc
547 finger genes (Jacobs et al., 2014). KRAB zinc finger genes are exclusive to
548 mammals, making the chromosome 4 ZnF genes a possible functional equivalent.

549 Also, as would be expected given their expression profiles, within these four clusters
550 are a large number of well-known developmental regulators such as members of the
551 Tgf-beta (Bmp and Nodal) and Wnt signalling pathways, as well as transcription
552 factors such as *gata5* & *6*, *sox32*, *vox*, *ved*, *vent* and both *Brachyury* homologues.

553 Correlation analysis of paralogues stemming from a teleost-specific genome
554 duplication shows that while there is the expected skew towards positive temporal
555 expression correlation, a significant proportion of paralogues have diverged in their
556 expression pattern as demonstrated by low or negative correlation coefficients. This
557 is confirmed by an analysis of the expression annotations provided by ZFIN. There
558 are caveats to this analysis, not least that the annotation is far from complete and,
559 therefore, annotations may not overlap simply because a certain stage has been
560 investigated for one paralogue, but not the other. It will also be the case that some
561 genes have expression patterns that are diverged spatially, but not temporally. Thus
562 the number of genes whose expression annotations do not overlap is an estimate,
563 but it does suggest that the phenomenon of divergence is widespread among
564 paralogues, something that might be expected to be the case since there needs to
565 be some form of selective pressure on retaining paralogous pairs.

566 A further layer of complexity is added by differential exon and 3′ end use. The
567 dataset uncovers new splice isoforms which, when viewed in Ensembl against the
568 Ensembl genebuild, will refines the current gene annotation and ties splice isoforms
569 to their temporal expression patterns. This is of particular relevance to the ongoing

570 debate surrounding reverse genetics approaches in zebrafish (Kok et al., 2015). One
571 possible explanation for phenotypic discrepancies between morpholino antisense
572 oligo knockdown, code disrupting point mutations and CRISPR/Cas9 or TALEN-
573 mediated larger deletions is that mutations affect potentially non-critical exons that
574 are only present in a subset of transcripts of a given gene. An improved annotation
575 that gives temporal resolution of exon usage will be of tremendous value for the
576 identification of critical exons.

577 Using the 3′ end sequencing method DeTCT, we have catalogued the differential
578 use of alternative 3′ ends in a filtered set of transcripts. In accordance with previous
579 work on a subset of stages (Collins et al., 2012; Sheppard et al., 2013; Ulitsky et al.,
580 2012) we find that there is differential use of not only 3′ UTRs but also coding
581 sequence across development. We show that *pdlim5b* is expressed as several
582 transcripts, with a single maternal isoform and multiple zygotic ones producing
583 different proteins. Incorporation of the data into an updated gene set will
584 systematically identify these and refines our understanding of gene regulation.
585 Furthermore, it provides scope for temporally resolved analysis using transcript-
586 specific knockout targeting strategies.

587 An important role of 3′ UTRs is in post-transcriptional regulation of their transcripts.
588 The 3′ UTRs contain binding sites for miRNAs which silence transcripts through
589 translational repression or mRNA destabilisation (Bartel, 2009). However, the
590 primary transcripts of miRNAs are often poorly annotated. Using two miRNA
591 processing mutants (*dgcr8* and *drosha*), we have produced an improved pri-miRNA
592 assembly with an increase of 42% in the number of Ensembl miRNAs with an
593 associated primary transcript. In agreement with previous studies (Chang et al.,
594 2015; Gaeta et al., 2017), the assembly shows that miRNAs are processed from
595 primary transcripts with multiple isoforms with potentially multiple transcription start
596 sites (Figure 7). This diversity of transcripts means that the expression of miRNAs in
597 relatively close proximity (e.g. dre-let-7d-1 and dre-mir-125a-1) could be regulated at
598 multiple levels (Figure 7C). In some transcripts both let-7d-1 and mir-125a-1 are both
599 intronic and in some mir-125a-1 is exonic. There are also transcripts with alternate
600 promoters which exclude the let-7d-1 locus entirely. These features provide
601 significant flexibility to the transcriptional regulation of miRNAs. It is possible that a
602 proportion of the assembled constructs are fragments. Clear definition of the 5’ and

603 3’ termini will require further work and ultimately experimental validation by an
604 alternative method such as 5′ rapid amplification of cDNA ends.

605 A major aim for this work was to create an accessible reference of normal mRNA
606 expression during zebrafish development. The primary sequence is available from
607 the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena), but importantly
608 has also been interpreted for immediate user-friendly access. Expression Atlas
609 provides relative expression levels while Ensembl displays alternative genomic
610 transcript structures alongside the current gene annotation. These pre-computed
611 analyses allow an in-depth examination of the data on a gene by gene basis. This
612 makes it easy for researchers to benefit from the data and provides a direct link to
613 the wealth of information present in genomic databases.

614 Materials and Methods

615 Animal Care

616 Zebrafish were maintained in accordance with UK Home Office regulations, UK
617 Animals (Scientific Procedures) Act 1986, under project licence 70/7606, which was
618 reviewed by the Wellcome Trust Sanger Institute Ethical Review Committee.

619 Sample collection

620 Wild-type HLF strain zebrafish (*Danio rerio*) were maintained at 28.5 °C on a 14h
621 light/10h dark cycle. At the time of mating, breeding males and females were
622 separated overnight before letting them spawn naturally in the morning to allow for
623 synchronisation of developmental stages. Fertilised eggs were grown at 28°C and
624 staged using previously defined criteria (Kimmel et al., 1995). Samples from 18
625 different developmental stages from 1-cell to 5 days post fertilisation were collected
626 by snap freezing the embryos in dry ice. Details of sample names and accession
627 numbers are provided in Supplemental Data File 1.

628 RNA Extraction

629 RNA was extracted from embryos by mechanical lysis in RLT buffer (Qiagen)
630 containing 1 µl of 14.3M beta mercaptoethanol (Sigma). The lysate was combined
631 with 1.8 volumes of Agencourt RNAClean XP (Beckman Coulter) beads and allowed
632 to bind for 10 minutes. The plate was applied to a plate magnet (Invitrogen) until the
633 solution cleared and the supernatant was removed without disturbing the beads. This
634 was followed by washing the beads three times with 70% ethanol. After the last
635 wash, the pellet was allowed to air dry for 10 mins and then resuspended in 50 µl of
636 RNAse-free water. RNA was eluted from the beads by applying the plate to the
637 magnetic rack. RNA was quantified using Quant-IT RNA assay (Invitrogen) for
638 samples 24 hours post-fertilisation and older.

639 RNA Sequencing

640 Total RNA from 12 embryos was pooled and DNase treated for 20 mins at 37°C
641 followed by addition of 1 µl 0.5M EDTA and inactivation at 75°C for 10 mins to
642 remove residual DNA. RNA was then cleaned using 2 volumes of Agencourt
643 RNAClean XP (Beckman Coulter) beads under the standard protocol. Five replicate
644 libraries were made for each stage using 700 ng total RNA and ERCC spike mix 2
645 (Ambion). Strand-specific RNA-seq libraries containing unique index sequences in
646 the adapter were generated simultaneously following the dUTP method. Libraries
647 were pooled and sequenced on Illumina HiSeq 2500 in 100bp paired-end mode.
648 Sequence data were deposited in ENA under accession ERP014517.

649 FASTQ files were aligned to the GRCz10 reference genome using tophat (v2.0.13,
650 options: --library-type fr-firststrand). Counts for genes were produced using htseq-
651 count (v0.6.0 options: --stranded=reverse) with the Ensembl v85 annotation as a
652 reference. The processed count data are in Supplemental Data File 2.

653 DeTCT

654 DeTCT libraries were generated, sequenced and analysed as described previously
655 (Collins et al., 2015). The resulting genomic regions and putative 3′ ends were
656 filtered stringently using DeTCT's filter_output script
657 (https://github.com/iansealy/DETCT/blob/master/script/filter_output.pl) in its --stricter
658 mode. This mode removes 3′ ends more than 5000 bases downstream of existing

659 protein-coding annotation (Ensembl v86) or more than 50 bases from existing non-
660 coding annotation (antisense, lincRNA, misc_RNA and processed_transcript
661 biotypes). 3′ ends are also removed if nearby sequence is enriched for As, if within
662 14 bp of an N, if within a simple repeat (repeats annotated by dust or trf) or a
663 transposon, or if not nearby a primary hexamer. Regions not associated with 3′ ends
664 are also removed. Sequence data were deposited in ENA under accessions
665 ERP006948 and ERP013756. The processed count data is available to download at
666 https://doi.org/10.6084/m9.figshare.4622311.

667 # Analysis

668 ## ZFA enrichment

669 ZFA enrichment was done using Ontologizer (http://ontologizer.de). The ZFA
670 ontology was download from http://www.berkeleybop.org/ontologies/zfa.obo and
671 mappings between ZFIN gene IDs and ZFA IDs was downloaded from ZFIN
672 (http://zfin.org/downloads/wildtype-expression_fish.txt and
673 http://zfin.org/downloads/phenoGeneCleanData_fish.txt). Pvalues were adjusted for
674 testing of both multiple terms and multiple clusters using the Bonferroni correction.

675 ## GO enrichment

676 GO enrichment was done using the R topGO package (Alexa and Rahnenfuhrer,
677 2016). The mapping between Ensembl IDs and GO terms was retrieved from the
678 Ensembl database using a custom perl script (get_ensembl_go_terms.pl) from the
679 topgo-wrapper repository (https://github.com/iansealy/topgo-wrapper).

680 ## Plots

681 The code to reproduce all the plots in this paper can be downloaded at github
682 (https://github.com/richysix/zfs-devstages).

683 ## BioLayout

684 BioLayout *Express*^{3D} is available to download for free
685 (http://www.biolayout.org/downloads/). The input expression file is provided as
686 Supplemental Data File 6. The processed layout file of the network displayed in
687 Figure 3B is available at https://doi.org/10.6084/m9.figshare.4622293.

688 ## pri-miRNA assembly

689 For each mutant allele – sa191, *drosha* (*rnasen*); sa223, *dgcr8* (*pasha*) – we
690 produced and sequenced stranded RNA-seq libraries from homozygous,
691 heterozygous and wild-type embryos (5 replicates of each). Sequence data were
692 deposited in ENA under accession ERP013690. To identify new pri-miRNAs, FASTQ
693 files were aligned to the zebrafish genome (GRCz10) using STAR (v2.5.1a, options:
694 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --
695 alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --
696 readFilesCommand zcat --outFilterIntronMotifs RemoveNoncanonical --outFilterType
697 BySJout --outFilterMultimapNmax 10 --sjdbOverhang 74). Ensembl v86 zebrafish
698 annotation was provided as a junction file. Aligned reads were sorted, indexed and
699 mate pairings were fixed using Samtools (v1.3.1). Duplicate alignments were
700 removed with Picard (v2.7.1) (REMOVE_DUPLICATES=true
701 READ_NAME_REGEX=null). Alignments for the homozygous, heterozygous and
702 wild-type samples for both the *drosha* and *dgcr8* lines were merged into two line-
703 specific datasets with Samtools. Each set of merged alignments was then used to
704 assemble a transcriptome with Cufflinks (v2.2.1) (--library-type=fr-firststrand).
705 Ensembl (v86) annotation stripped of rows corresponding to 'gene' features was
706 provided as a backbone for the assembly. The *drosha* and *dgcr8* assemblies were
707 then merged with Cuffmerge.

708 Ensembl miRNA annotation (v87) ("miRNA" biotype) was used to define miRNA loci.
709 For the purpose of this analysis we defined pri-miRNAs as any transcript overlapping
710 at least one of these loci. Ensembl and Havana (v87) annotation was selected as the
711 baseline for novel pri-miRNA prediction. UCSC utilities gtfToGenePred and
712 genePredToBed were used to convert from GTF to BED format. Pri-miRNAs were
713 identified by removing single exon transcripts from the merged assembly and
714 Ensembl annotations and comparing the remainder to annotated miRNA loci using
715 bedtools (v2.24.0) intersect (-wa –s). Finally, assembled pri-miRNA transcripts were
716 filtered to remove those for which either the first or last exon was less than 20nt in
717 length, those which contain an intron greater than 100,000nt, or those which contain
718 an intron that isn't flanked by the GT-AG or GC-AG motifs (Breathnach et al., 1978).

719 ## Acknowledgements

720 This work was funded by the Wellcome Trust Sanger Institute (grant number
721 WT098051). M. Davis was funded by a MRC Methodology Research Fellowship
722 (MR/L012367/1). Ensembl, Expression Atlas and A. Enright are funded by EMBL
723 core funding. We would like to thank the Wellcome Trust Sanger Institute
724 sequencing pipelines for performing sequencing and the staff of the Research
725 Support Facility for zebrafish care. We also thank Samantha Carruthers, Catherine
726 Scahill and Nicole Staudt for critical reading of the manuscript.

727 **Supplemental Data Files:**

728 Supplemental Data File 1 – Sample Information (.tsv)
729 Supplemental Data File 2 – RNA-seq count data (.tsv)
730 Supplemental Data File 3 – Max stage cluster assignments (.tsv)
731 Supplemental Data File 4 – Max stage cluster ZFA enrichment (.tsv)
732 Supplemental Data File 5 – Max stage cluster GO enrichment (.tsv)
733 Supplemental Data File 6 – BioLayout input file (.expression)
734 Supplemental Data File 7 – BioLayout cluster information (.tsv)
735 Supplemental Data File 8 – BioLayout cluster ZFA enrichment (.zip)
736 Supplemental Data File 9 – BioLayout cluster GO enrichment (.tsv)
737 Supplemental Data File 10 – Regional expression information (.tsv)

738 Competing Interests

739 The authors declare they have no competing interests.

740 References

741 **Aanes, H., Winata, C. L., Lin, C. H., Chen, J. P., Srinivasan, K. G., Lee, S. G.,**
742 **Lim, A. Y., Hajan, H. S., Collas, P., Bourque, G., et al.** (2011). Zebrafish
743 mRNA sequencing deciphers novelties in transcriptome dynamics during
744 maternal to zygotic transition. *Genome Res* **21**, 1328-1338. doi:
745 10.1101/gr.116012.110.
746 **Alexa, A. and Rahnenfuhrer, J.** (2016). topGO: Enrichment Analysis for Gene
747 Ontology. R package version 2.24.0. doi:
748 https://bioconductor.org/packages/release/bioc/html/topGO.html.
749 **Amsterdam, A., Burgess, S., Golling, G., Chen, W., Sun, Z., Townsend, K.,**
750 **Farrington, S., Haldi, M. and Hopkins, N.** (1999). A large-scale insertional
751 mutagenesis screen in zebrafish. *Genes Dev* **13**, 2713-2724. doi:

752 **Bartel, D. P.** (2009). MicroRNAs: target recognition and regulatory functions. *Cell*
753 **136**, 215-233. doi: 10.1016/j.cell.2009.01.002.
754 **Behrndt, M., Salbreux, G., Campinho, P., Hauschild, R., Oswald, F., Roensch,**
755 **J., Grill, S. W. and Heisenberg, C. P.** (2012). Forces driving epithelial
756 spreading in zebrafish gastrulation. *Science* **338**, 257-260. doi:
757 10.1126/science.1224143.
758 **Bertrand, J. Y., Chi, N. C., Santoso, B., Teng, S., Stainier, D. Y. and Traver, D.**
759 (2010). Haematopoietic stem cells derive directly from aortic endothelium
760 during development. *Nature* **464**, 108-111. doi: 10.1038/nature08738.
761 **Boeck, M. E., Huynh, C., Gevirtzman, L., Thompson, O. A., Wang, G., Kasper, D.**
762 **M., Reinke, V., Hillier, L. W. and Waterston, R. H.** (2016). The time-resolved
763 transcriptome of C. elegans. *Genome Res* **26**, 1441-1450. doi:
764 10.1101/gr.202663.115.
765 **Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P.** (1978).
766 Ovalbumin gene: evidence for a leader sequence in mRNA and DNA
767 sequences at the exon-intron boundaries. *Proc Natl Acad Sci U S A* **75**, 4853-
768 4857. doi:
769 **Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O.,**
770 **Booth, B. W., Wen, J., Park, S., Suzuki, A. M., et al.** (2014). Diversity and
771 dynamics of the Drosophila transcriptome. *Nature* **512**, 393-399. doi:
772 10.1038/nature12962.
773 **Chang, T. C., Pertea, M., Lee, S., Salzberg, S. L. and Mendell, J. T.** (2015).
774 Genome-wide annotation of microRNA primary transcript structures reveals
775 novel regulatory mechanisms. *Genome Res* **25**, 1401-1409. doi:
776 10.1101/gr.193607.115.
777 **Collins, J. E., Wali, N., Sealy, I. M., Morris, J. A., White, R. J., Leonard, S. R.,**
778 **Jackson, D. K., Jones, M. C., Smerdon, N. C., Zamora, J., et al.** (2015).
779 High-throughput and quantitative genome-wide messenger RNA sequencing
780 for molecular phenotyping. *BMC Genomics* **16**, 578. doi: 10.1186/s12864-
781 015-1788-6.
782 **Collins, J. E., White, S., Searle, S. M. and Stemple, D. L.** (2012). Incorporating
783 RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res* **22**, 2067-
784 2078. doi: 10.1101/gr.137901.112.
785 **Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. and Hannon, G. J.** (2004).
786 Processing of primary microRNAs by the Microprocessor complex. *Nature*
787 **432**, 231-235. doi: 10.1038/nature03049.
788 **Driever, W., Solnica-Krezel, L., Schier, A. F., Neuhauss, S. C., Malicki, J.,**
789 **Stemple, D. L., Stainier, D. Y., Zwartkruis, F., Abdelilah, S., Rangini, Z., et**
790 **al.** (1996). A genetic screen for mutations affecting embryogenesis in
791 zebrafish. *Development* **123**, 37-46. doi:
792 **Enright, A. J., Van Dongen, S. and Ouzounis, C. A.** (2002). An efficient algorithm
793 for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584.
794 doi:
795 **Gaeta, X., Le, L., Lin, Y., Xie, Y. and Lowry, W. E.** (2017). Defining Transcriptional
796 Regulatory Mechanisms for Primary let-7 miRNAs. *PLoS One* **12**, e0169237.
797 doi: 10.1371/journal.pone.0169237.
798 **Gehrig, J., Reischl, M., Kalmar, E., Ferg, M., Hadzhiev, Y., Zaucker, A., Song, C.,**
799 **Schindler, S., Liebel, U. and Muller, F.** (2009). Automated high-throughput
800 mapping of promoter-enhancer interactions in zebrafish embryos. *Nat*
801 *Methods* **6**, 911-916. doi: 10.1038/nmeth.1396.

802 **Gerstein, M. B., Rozowsky, J., Yan, K. K., Wang, D., Cheng, C., Brown, J. B.,**
803 **Davis, C. A., Hillier, L., Sisu, C., Li, J. J., et al.** (2014). Comparative analysis
804 of the transcriptome across distant species. *Nature* **512**, 445-448. doi:
805 10.1038/nature13424.
806 **Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang,**
807 **L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al.** (2011).
808 The developmental transcriptome of Drosophila melanogaster. *Nature* **471**,
809 473-479. doi: 10.1038/nature09715.
810 **Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch,**
811 **N. and Shiekhattar, R.** (2004). The Microprocessor complex mediates the
812 genesis of microRNAs. *Nature* **432**, 235-240. doi: 10.1038/nature03120.
813 **Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L.,**
814 **Fire, A., Ruvkun, G. and Mello, C. C.** (2001). Genes and mechanisms
815 related to RNA interference regulate expression of the small temporal RNAs
816 that control C. elegans developmental timing. *Cell* **106**, 23-34. doi:
817 **Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., Gehrig, J.,**
818 **Dong, X., Akalin, A., Suzuki, A. M., et al.** (2014). Two independent
819 transcription initiation codes overlap on vertebrate core promoters. *Nature*
820 **507**, 381-385. doi: 10.1038/nature12974.
821 **Haffter, P., Granato, M., Brand, M., Mullins, M. C., Hammerschmidt, M., Kane, D.**
822 **A., Odenthal, J., van Eeden, F. J., Jiang, Y. J., Heisenberg, C. P., et al.**
823 (1996). The identification of genes with unique and essential functions in the
824 development of the zebrafish, Danio rerio. *Development* **123**, 1-36. doi:
825 **Harrison, M. M., Botchan, M. R. and Cline, T. W.** (2010). Grainyhead and Zelda
826 compete for binding to the promoters of the earliest-expressed Drosophila
827 genes. *Dev Biol* **345**, 248-255. doi: 10.1016/j.ydbio.2010.06.026.
828 **Harrison, M. M., Li, X. Y., Kaplan, T., Botchan, M. R. and Eisen, M. B.** (2011).
829 Zelda binding in the early Drosophila melanogaster embryo marks regions
830 subsequently activated at the maternal-to-zygotic transition. *PLoS Genet* **7**,
831 e1002266. doi: 10.1371/journal.pgen.1002266.
832 **Harvey, S. A., Sealy, I., Kettleborough, R., Fenyes, F., White, R., Stemple, D.**
833 **and Smith, J. C.** (2013). Identification of the zebrafish maternal and paternal
834 transcriptomes. *Development* **140**, 2703-2710. doi: 10.1242/dev.095091.
835 **Hashimshony, T., Feder, M., Levin, M., Hall, B. K. and Yanai, I.** (2015).
836 Spatiotemporal transcriptomics reveals the evolutionary history of the
837 endoderm germ layer. *Nature* **519**, 219-222. doi: 10.1038/nature13996.
838 **Howe, D. G., Bradford, Y. M., Conlin, T., Eagle, A. E., Fashena, D., Frazer, K.,**
839 **Knight, J., Mani, P., Martin, R., Moxon, S. A., et al.** (2013a). ZFIN, the
840 Zebrafish Model Organism Database: increased support for mutants and
841 transgenics. *Nucleic Acids Res* **41**, D854-860. doi: 10.1093/nar/gks938.
842 **Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M.,**
843 **Collins, J. E., Humphray, S., McLaren, K., Matthews, L., et al.** (2013b). The
844 zebrafish reference genome sequence and its relationship to the human
845 genome. *Nature* **496**, 498-503. doi: 10.1038/nature12111.
846 **Howe, K., Schiffer, P. H., Zielinski, J., Wiehe, T., Laird, G. K., Marioni, J. C.,**
847 **Soylemez, O., Kondrashov, F. and Leptin, M.** (2016). Structure and
848 evolutionary history of a large family of NLR proteins in the zebrafish. *Open*
849 *Biol* **6**, 160009. doi: 10.1098/rsob.160009.
850 **Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T. and**
851 **Zamore, P. D.** (2001). A cellular function for the RNA-interference enzyme

852 Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-
853 838. doi: 10.1126/science.1062961.
854 **Jacobs, F. M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman,**
855 **S., Paten, B., Salama, S. R. and Haussler, D.** (2014). An evolutionary arms
856 race between KRAB zinc-finger genes ZNF91/93 and SVA/L1
857 retrotransposons. *Nature* **516**, 242-245. doi: 10.1038/nature13760.
858 **Kaaij, L. J., Mokry, M., Zhou, M., Musheev, M., Geeven, G., Melquiond, A. S., de**
859 **Jesus Domingues, A. M., de Laat, W., Niehrs, C., Smith, A. D., et al.**
860 (2016). Enhancers reside in a unique epigenetic environment during early
861 zebrafish development. *Genome Biol* **17**, 146. doi: 10.1186/s13059-016-1013-
862 1.
863 **Kettleborough, R. N., Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., de**
864 **Bruijn, E., van Eeden, F., Sealy, I., White, R. J., Herd, C., Nijman, I. J., et**
865 **al.** (2013). A systematic genome-wide analysis of zebrafish protein-coding
866 gene function. *Nature* **496**, 494-497. doi: 10.1038/nature11992.
867 **Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. and Schilling, T. F.**
868 (1995). Stages of embryonic development of the zebrafish. *Dev Dyn* **203**, 253-
869 310. doi: 10.1002/aja.1002030302.
870 **Kissa, K. and Herbomel, P.** (2010). Blood stem cells emerge from aortic
871 endothelium by a novel type of cell transition. *Nature* **464**, 112-115. doi:
872 10.1038/nature08761.
873 **Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D. and**
874 **Penin, A. A.** (2016). A high resolution map of the Arabidopsis thaliana
875 developmental transcriptome based on RNA-seq profiling. *Plant J*. doi:
876 10.1111/tpj.13312.
877 **Kok, F. O., Shin, M., Ni, C. W., Gupta, A., Grosse, A. S., van Impel, A.,**
878 **Kirchmaier, B. C., Peterson-Maduro, J., Kourkoulis, G., Male, I., et al.**
879 (2015). Reverse genetic screening reveals poor correlation between
880 morpholino-induced and mutant phenotypes in zebrafish. *Dev Cell* **32**, 97-108.
881 doi: 10.1016/j.devcel.2014.11.018.
882 **Lee, M. T., Bonneau, A. R., Takacs, C. M., Bazzini, A. A., DiVito, K. R., Fleming,**
883 **E. S. and Giraldez, A. J.** (2013). Nanog, Pou5f1 and SoxB1 activate zygotic
884 gene expression during the maternal-to-zygotic transition. *Nature* **503**, 360-
885 364. doi: 10.1038/nature12632.
886 **Leichsenring, M., Maes, J., Mossner, R., Driever, W. and Onichtchouk, D.**
887 (2013). Pou5f1 transcription factor controls zygotic gene activation in
888 vertebrates. *Science* **341**, 1005-1009. doi: 10.1126/science.1242527.
889 **Li, P., Lahvic, J. L., Binder, V., Pugach, E. K., Riley, E. B., Tamplin, O. J.,**
890 **Panigrahy, D., Bowman, T. V., Barrett, F. G., Heffner, G. C., et al.** (2015).
891 Epoxyeicosatrienoic acids enhance embryonic haematopoiesis and adult
892 marrow engraftment. *Nature* **523**, 468-471. doi: 10.1038/nature14569.
893 **Liang, H. L., Nien, C. Y., Liu, H. Y., Metzstein, M. M., Kirov, N. and Rushlow, C.**
894 (2008). The zinc-finger protein Zelda is a key activator of the early zygotic
895 genome in Drosophila. *Nature* **456**, 400-403. doi: 10.1038/nature07388.
896 **Lindeman, L. C., Winata, C. L., Aanes, H., Mathavan, S., Alestrom, P. and**
897 **Collas, P.** (2010). Chromatin states of developmentally-regulated genes
898 revealed by DNA and histone methylation patterns in zebrafish embryos. *Int J*
899 *Dev Biol* **54**, 803-813. doi: 10.1387/ijdb.103081ll.

900 **Moens, C. B., Donn, T. M., Wolf-Saxon, E. R. and Ma, T. P.** (2008). Reverse
901 genetics in zebrafish by TILLING. *Brief Funct Genomic Proteomic* **7**, 454-459.
902 doi: 10.1093/bfgp/eln046.
903 **Murphey, R. D., Stern, H. M., Straub, C. T. and Zon, L. I.** (2006). A chemical
904 genetic screen for cell cycle inhibitors in zebrafish embryos. *Chem Biol Drug*
905 *Des* **68**, 213-219. doi: 10.1111/j.1747-0285.2006.00439.x.
906 **Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.**
907 **M., Sheng, Y., Abdelhamid, R. F., Anand, S., et al.** (2013). Dynamic
908 regulation of the transcription initiation landscape at single nucleotide
909 resolution during vertebrate embryogenesis. *Genome Res* **23**, 1938-1950. doi:
910 10.1101/gr.153692.112.
911 **North, T. E., Goessling, W., Walkley, C. R., Lengerke, C., Kopani, K. R., Lord, A.**
912 **M., Weber, G. J., Bowman, T. V., Jang, I. H., Grosser, T., et al.** (2007).
913 Prostaglandin E2 regulates vertebrate haematopoietic stem cell homeostasis.
914 *Nature* **447**, 1007-1011. doi: 10.1038/nature05883.
915 **Owens, N. D., Blitz, I. L., Lane, M. A., Patrushev, I., Overton, J. D., Gilchrist, M.**
916 **J., Cho, K. W. and Khokha, M. K.** (2016). Measuring Absolute RNA Copy
917 Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in
918 Development. *Cell Rep* **14**, 632-647. doi: 10.1016/j.celrep.2015.12.050.
919 **Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan,**
920 **L., Sandelin, A., Rinn, J. L., Regev, A., et al.** (2012). Systematic
921 identification of long noncoding RNAs expressed during zebrafish
922 embryogenesis. *Genome Res* **22**, 577-591. doi: 10.1101/gr.133009.111.
923 **Peterson, R. T., Link, B. A., Dowling, J. E. and Schreiber, S. L.** (2000). Small
924 molecule developmental screens reveal the logic and timing of vertebrate
925 development. *Proc Natl Acad Sci U S A* **97**, 12965-12969. doi:
926 10.1073/pnas.97.24.12965.
927 **Peterson, R. T., Shaw, S. Y., Peterson, T. A., Milan, D. J., Zhong, T. P.,**
928 **Schreiber, S. L., MacRae, C. A. and Fishman, M. C.** (2004). Chemical
929 suppression of a genetic mutation in a zebrafish model of aortic coarctation.
930 *Nat Biotechnol* **22**, 595-599. doi: 10.1038/nbt963.
931 **Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T.,**
932 **Fullgrabe, A., Fuentes, A. M., Jupp, S., Koskinen, S., et al.** (2016).
933 Expression Atlas update--an integrated database of gene and protein
934 expression in humans, animals and plants. *Nucleic Acids Res* **44**, D746-752.
935 doi: 10.1093/nar/gkv1045.
936 **Potok, M. E., Nix, D. A., Parnell, T. J. and Cairns, B. R.** (2013). Reprogramming
937 the maternal zebrafish genome after fertilization to match the paternal
938 methylation pattern. *Cell* **153**, 759-772. doi: 10.1016/j.cell.2013.04.030.
939 **Rauch, G. J., Lyons, D. A., Middendorf, I., Friedlander, B., Arana, N., Reyes, T.**
940 **and Talbot, W. S.** (2003). Submission and Curation of Gene Expression
941 Data. ZFIN Direct Data Submission (http://zfin.org/). doi:
942 **Sheppard, S., Lawson, N. D. and Zhu, L. J.** (2013). Accurate identification of
943 polyadenylation sites from 3' end deep sequencing using a naive Bayes
944 classifier. *Bioinformatics* **29**, 2564-2571. doi: 10.1093/bioinformatics/btt446.
945 **Stern, H. M., Murphey, R. D., Shepard, J. L., Amatruda, J. F., Straub, C. T., Pfaff,**
946 **K. L., Weber, G., Tallarico, J. A., King, R. W. and Zon, L. I.** (2005). Small
947 molecules that delay S phase suppress a zebrafish bmyb mutant. *Nat Chem*
948 *Biol* **1**, 366-370. doi:

949 **Tan, M. H., Au, K. F., Yablonovitch, A. L., Wills, A. E., Chuang, J., Baker, J. C.,**
950 **Wong, W. H. and Li, J. B.** (2013). RNA sequencing reveals a diverse and
951 dynamic repertoire of the Xenopus tropicalis transcriptome over development.
952 *Genome Res* **23**, 201-216. doi: 10.1101/gr.141424.112.
953 **Theocharidis, A., van Dongen, S., Enright, A. J. and Freeman, T. C.** (2009).
954 Network visualization and analysis of gene expression data using BioLayout
955 Express(3D). *Nat Protoc* **4**, 1535-1550. doi: 10.1038/nprot.2009.177.
956 **Thisse, B., Pflumio, S., Fürthauer, M., Loppin, B., Heyer, V., Degrave, A., Woehl,**
957 **R., Lux, A., Steffan, T., Charbonnier, X. Q., et al.** (2001). Expression of the
958 zebrafish genome during embryogenesis. (NIH RO1 RR15402-01). ZFIN
959 Direct Data Submission (http://zfin.org/). doi:
960 **Tobin, D. M., Vary, J. C., Jr., Ray, J. P., Walsh, G. S., Dunstan, S. J., Bang, N. D.,**
961 **Hagge, D. A., Khadge, S., King, M. C., Hawn, T. R., et al.** (2010). The lta4h
962 locus modulates susceptibility to mycobacterial infection in zebrafish and
963 humans. *Cell* **140**, 717-730. doi: 10.1016/j.cell.2010.02.013.
964 **Ulitsky, I., Shkumatava, A., Jan, C. H., Subtelny, A. O., Koppstein, D., Bell, G.**
965 **W., Sive, H. and Bartel, D. P.** (2012). Extensive alternative polyadenylation
966 during zebrafish development. *Genome Res* **22**, 2054-2066. doi:
967 10.1101/gr.139733.112.
968 **van Dongen, S.** (2000). Graph Clustering by Flow Simulation. University of
969 Utrecht.
970 **Van Slyke, C. E., Bradford, Y. M., Westerfield, M. and Haendel, M. A.** (2014). The
971 zebrafish anatomy and stage ontologies: representing the anatomy and
972 development of Danio rerio. *J Biomed Semantics* **5**, 12. doi: 10.1186/2041-
973 1480-5-12.
974 **Varshney, G. K., Pei, W., LaFave, M. C., Idol, J., Xu, L., Gallardo, V., Carrington,**
975 **B., Bishop, K., Jones, M., Li, M., et al.** (2015). High-throughput gene
976 targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res* **25**,
977 1030-1042. doi: 10.1101/gr.186379.114.
978 **Vastenhouw, N. L., Zhang, Y., Woods, I. G., Imam, F., Regev, A., Liu, X. S., Rinn,**
979 **J. and Schier, A. F.** (2010). Chromatin signature of embryonic pluripotency is
980 established during genome activation. *Nature* **464**, 922-926. doi:
981 10.1038/nature08866.
982 **Wagner, E. J. and Carpenter, P. B.** (2012). Understanding the language of Lys36
983 methylation at histone H3. *Nat Rev Mol Cell Biol* **13**, 115-126. doi:
984 10.1038/nrm3274.
985 **Yang, H., Zhou, Y., Gu, J., Xie, S., Xu, Y., Zhu, G., Wang, L., Huang, J., Ma, H.**
986 **and Yao, J.** (2013). Deep mRNA sequencing analysis to capture the
987 transcriptome landscape of zebrafish embryos and larvae. *PLoS One* **8**,
988 e64058. doi: 10.1371/journal.pone.0064058.
989