Osaka University Knowledge Archive

Title	Development of a transcription factor database useful for developmental biology, and decoding the maternal and zygotic transcriptome of the appendicularian, Oikopleura dioica
Author(s)	Wang, Kai
Citation	大阪大学, 2015, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/54035
rights	
Note	

Osaka University Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

Osaka University

Development of a transcription factor database useful for developmental biology, and decoding the maternal and zygotic transcriptome of the appendicularian, *Oikopleura dioica*

（発生生物学研究のための転写因子データベースの開発とオタマボヤを用いた母性と胚性の **RNA-Seq** 解析）

Name:	Kai WANG
Major:	Developmental Biology
Supervisor:	Hiroki NISHIDA (Professor)
Chief examiner:	Hiroki NISHIDA (Professor)
Second examiner:	Teruo YASUNAGA (Professor)
Second examiner:	Zhi-Hui SU (Professor)
Date:	26 June 2015

ABSTRACT

Genes encoding transcription factors that constitute gene-regulatory networks and maternal factors accumulating in egg cytoplasm are two classes of essential genes that play crucial roles in developmental processes. Transcription factors control the expression of their downstream target genes by interacting with *cis*-regulatory elements. Maternal factors initiate embryonic developmental programs by regulating the expression of zygotic genes and various other events during early embryogenesis. This manuscript documents the transcription factors of 77 metazoan species as well as human and mouse maternal factors. I improved the previous method for prediction of transcription factors using a statistical approach adding Gene Ontology information to Pfam based identification of transcription factors. This method detects previously un-discovered transcription factors. The novel features of this database are:

1) It includes both transcription factors and maternal factors, although the number of species, in which maternal factors are listed, is limited at the moment.
2) Ontological representation at the cell, tissue, organ, and system levels has been specially designed to facilitate developmental studies. This is the unique feature in our database and is not available in other transcription factor databases.

I developed, a user-friendly web interface, REGULATOR (http://www.bioinformatics.org/regulator/), which can help researchers to efficiently identify, validate, and visualize the data in the database. Using this web interface, users can browse, search, and download detailed information on species of interest, genes, transcription factor families, or developmental ontology terms.

In the second part of this thesis, I performed transcriptome analysis for the appendicularian, *Oikopleura dioica* (*O. dioica*), which is a planktonic chordate. It has been used as a novel model species in development biology and evolutionary studies. The most significant character of this species is rapid development and short life cycle (only 5 days). Recently, many bioinformatics resources are publicly available, such as the genome sequence and microarray data in the OikoBase. However, transcriptome information is still not complete, especially as to the next-generation sequencing data. In this study, we carried out transcriptome analysis using RNA sequencing data of a Japanese population collected from egg and larval stages. The major findings of this study are:

1. Via mapping our reads (Japanese population) to the reference genome sequence deposited in the OikoBase (Norwegian population), an extreme low reads mapping rate were found, which due to the significant sequence variations between the Japanese population and the Norwegian population.
2. After *de novo* transcriptome assembly, 16,423 proteins (belongs to 12,136 known unigenes) of Japanese population have homologies with that of the Norwegian population (OikoBase). These proteins corresponding to 95.4% of the protein-encoding genes deposited in OikoBase.
3. Via comparing the 4,136 one-vs-one bidirectional best hits (BBH) at both the nucleotide and protein levels, the sequence similarity between Japanese and Norwegian population were estimated to be 91.0% in nucleotide level and 94.8% in amino acid level.
4. Additional 175 novel protein-encoding genes were found. Among these novel genes, 144 of them were not predicted in the gene models deposited in OikoBase, but they can be

found in the Norwegian reference genome; whereas 31 unigenes were not found in the OikoBase reference genome due to the low sequence similarity.

5. I found approximately 63% unigenes were expressed in egg-stage, whereas 99% were observed in larval-stage; Analysis of differentially expressed genes (DEG) using a fold change threshold of four for the total 12,311 unigenes, we found 3,772 of them were up-regulated and 1,336 were down-regulated. Gene ontology (GO) analyses showed distinct gene activities in these two developmental stages.
6. The previously reported mRNA 5' *trans*-spliced leader was also detected by our method. While, it was observed in 40.8% of the total unique transcripts, which is much more frequently happed than the previous estimation. This *trans*-spliced leader showed preferential linkage to adenine at the 5' ends of the downstream exons.
7. By comparing the *trans*-spliced mRNAs between egg and larval stage, as well as the down-regulated and up-regulated groups, we found *trans*-spliced mRNAs were more frequently observed in egg compared to larva.

Our data of transcriptome assembly will provide an additional resource for studies of Japanese population. These findings would be useful for better understanding of the development as well as evolutionary history of *O. dioica*.

KEYWORDS: Transcription factors; maternal factors; *Oikopleura dioica*; transcriptome; intra-species variation; novel genes; *trans*-splicing

ABBREVIATION

Abbreviation	Full description
bp	base pairs
CAB	Centrosome-attracting body
ChIP-seq	Chromatin immunoprecipitation coupled with massively parallel DNA sequencing
DEGs	Differentially expressed genes
FC	Fold change
FPKM:	fragments per kilobase per million mapped reads
GCRMA	Guanine Cytosine Robust Multi-Array Analysis
GO	Gene ontology
HPF	Hours post fertilization
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOOCV	Leave-One-Out Cross-Validation
MZT	Maternal to zygotic transition
NGS	Next-generation sequencing
PEM	Posterior end mark
PVC	Posterior-vegetal cytoplasm
RPKM	Reads per kilobase per million mapped reads
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing
RT-PCR	Real-time polymerase chain reaction
SL	*Trans*-spliced leader
UTR	Untranslated region
ZGA	Zygotic genome activation

INDEX

ABSTRACT I
ABBREVIATION III
INDEX IV
1 Background 1
1.1 Current advances in development biology 1
1.2 The REGULATOR database 2
1.3 Current advances in *O. dioica* studies 2
2 The REGULATOR database 3
2.1 Introduction 3
2.2 Materials and methods 3
2.2.1 Prediction methods for transcription factors 3
2.2.2 Examples of mathematical details 9
2.2.3 Prediction methods for maternal factors 14
2.3 Results 14
2.3.1 TF prediction 14
2.3.2 MF prediction 17
2.3.3 Comprehensive annotation 17
2.3.4 Developmental ontology terms 18
2.3.5 Web interface 19
2.4 Discussion 20
3 Transcriptome in the appendicularian, *O. dioica* 23
3.1 Introduction 23
3.1.1 Research background of *O. dioica* 23
3.1.2 Research purpose 25
3.1.3 Research content 25
3.2 Illumina next-generation sequencing 26
3.3 Bioinformatics tools and methods 26
3.3.1 Identification of differentially expressed genes 27
3.3.2 Gene Ontology 28
3.4 NGS and data preprocess 29
3.4.1 Laboratory culture and sample collection 29
3.4.2 RNA isolation 30
3.4.3 Library construction 30
3.4.4 Illumina HiSeq 2000 sequencing 31
3.4.5 Data preprocess 31
3.5 *De novo* transcriptome assembly 32
3.5.1 The necessity for *de novo* transcriptome assembly 32
3.5.2 Method for *de novo* transcriptome assembly 32
3.5.3 Result assessment 33
3.6 Protein-encoding genes 34
3.6.1 Method for known and novel protein-encoding gene inference 34
3.6.2 RT-PCR validation of novel genes 35
3.6.3 Result 36

3.6.4 Function annotation of protein-encoding genes 39
3.7 Differentially expressed genes 53
3.7.1 Method for DEGs Identification 53
3.7.2 Method for GO enrichment analysis of DEGs 53
3.7.3 Result 53
3.8 Intra-species sequence variations 60
3.9 *Trans*-spliced leader and *trans*-spliced mRNAs 61
3.9.1 mRNA *trans*-splicing and *trans*-spliced leader 61
3.9.2 Method for identification of SL and *trans*-spliced mRNAs 61
3.9.3 Result 61
3.10 Discussion 67
4 Conclusions and perspectives 69
4.1 Conclusion 69
4.2 Perspectives 69
ACKNOWLEDGEMENTS 71
FUNDING 73
REFERENCES 74
Publication list related to the Doctor thesis 85

1 Background

1.1 Current advances in development biology

It is amazing that all animals, including us human, are developed from a tiny fertilized egg. To understand the molecular mechanisms how eggs develop into a well-organized adult body, we need to answer two basic questions: (1) What is happening during the embryogenesis process? (2) What kinds of regulatory factors control the development process?

Various model species has been used for the embryogenesis studies. The early morphogenetic processes showed similar characteristics among animals. All early embryos passed through a cleavage stage, which begins from the fertilized egg and ended before gastrulation. Due to the rapid cell cycles, the cell number is increasing but the whole embryonic size does not change during this period (http://en.wikipedia.org/wiki/Embryogenesis). The morphology and size of each cell of the embryo is almost uniform. At around 16 to 32-cells stage, the embryo forms a dense ball, which looks like a mulberry fruit, and thus is called morula. The next stage is known as gastrula. The most significant feature of this stage is the formation of trilaminar structures (or three germ layers), which are known as ectoderm, mesoderm and endoderm (http://en.wikipedia.org/wiki/Gastrulation). Then embryos develop into the neurula stage, during which the nervous system formation is initiated. The transformation from neural plate to neural tube is the major event of this stage in chordate embryos (http://en.wikipedia.org/wiki/Neurulation). Then, somitogenesis and organogenesis follow. As embryogenesis progresses, the embryos undergo rapid growth and differentiation. Finally, the embryo will develop into a well-organized adult. The morphogenetic processes have been well studied in most model species. Despite there are some difference, these processes are conserved among them.

What kinds of regulatory factors control these processes and how they worked? Now, we know that some genes, especially those encode transcription factors and maternal factors, play crucial roles in early embryonic development. For example, the maternal to zygotic transition (MZT) [1-7] is recognized as an essential period during early embryonic development. Genes controlling this process are supposed to be important key regulators. In Zebrafish, three transcription factors, *Nanog*, *Pou5f1* and *SoxB1* activate the first wave of zygotic gene expression during the MZT. In addition, some of the maternal factors [8-12] that deposited in the mature oocytes were also recognized to be in charge of the zygotic genome activation (ZGA). The maternal factors that determine embryonic cell fates are called maternal determinants. Two major classes of maternal determinants are maternal transcript factors and signaling molecules (including the signaling regulators) [9]. In *Xenopus*, some T-box (e.g. *VegT*), Sox (*Sox2*, *Sox3*, *Sox17*) and Fox (*FoxD2*, *FoxD3*, *FoxD5*, *FoxH1* and *FoxK1*) family maternal transcription factors are essential for embryonic cell fate specification. Some maternal signaling molecules such as *Smad1*, *Smad2*, *MAPK*, *BMP2*, *BMP7*, *FGF2*, *FGF4*, and *FGF9* are also play critical roles in this process [9].

A pivotal event during early embryonic development is polarization, through which the fertilized egg will finally become a three-dimensional embryo with the animal-vegetal (A-V), ventral-dorsal (V-D) and anterior-posterior (A-P) axis [13]. In ascidians, some maternal mRNAs

were found to be localized to the vegetal hemisphere and then enriched in the posterior region at 8-cell stage, and thus they are called postplasmic mRNAs [10, 14, 15]. These maternal mRNAs are also called posterior end mark (PEM) mRNAs [10]. A total of two types of postplasmic/PEM RNAs were found, namely Type I and Type II [14]. Type I postplasmic/PEM RNAs are the major forms and abundantly localized to the posterior pole of ascidian embryos [15], Type II postplasmic/PEM RNAs are also present ubiquitously in the egg cytoplasm in addition [14]. The localization of these mRNAs was supposed to be mediated by the signature in the 3' untranslated region (3' UTR), and was also observed in several ascidian species [16].

1.2 The REGULATOR database

In view of the important roles of transcription factors and maternal factors, we developed a database for developmental biological studies [17]. It lists the transcription factors of 77 metazoan species as well as human and mouse maternal factors. The novel features of this database are: (1) It includes both transcription factors and maternal factors, although the number of species, in which maternal factors are listed, is limited at the moment. (2) Ontological representation at the cell, tissue, organ, and system levels has been specially designed to facilitate development studies. This is the unique feature in our database and is not available in other transcription factor databases. This will be described in chapter two.

1.3 Current advances in *O. dioica* studies

Oikopleura dioica is a promising modal organism having various advantages for genome and developmental analyses. However, currently there are only six labs working with this animal in the world and only limited works have been published. Many important genes in this species have not been investigated by biological experiments. Thus, identification of homologous genes to the phylogenetically related ascidians would provide important cues for developmental studies in *O. dioica*. The embryogenesis, cell lineages and fate map of *O. dioica* have been summarized by Nishida [18, 19].

To date, several bioinformatics resources are available for *O. dioica*. The current version of the draft genome sequence [20] is v3, which is available at Genoscope (http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/). Moreover, the transcriptome data (includes microarray and EST data) which covers various development stages is deposited in the publicly available OikoBase [21], providing a useful resource for developmental studies. RNA-Seq data is expected to provide more information, however, this had not been reported. In this study, I reported our transcriptome study using RNA-Seq [22]. This will be described in chapter three.

2 The REGULATOR database

2.1 Introduction

Transcription factors (TFs) bind to the cis-regulatory elements of downstream target genes and promote or block the recruitment of RNA polymerase II to those promoter regions [23, 24]. They control various developmental processes by regulating cell fate specification [25, 26], morphogenesis [27, 28], the cell cycle [29], apoptosis [30] and pathogenesis [31]. Similarly, maternal factors (MFs) present in unfertilized eggs are of interest, as they play crucial roles in early embryogenesis [8-10, 32, 33]. MFs initiate embryonic developmental programs, followed by triggering of zygotic gene activation [3, 8, 34]. Comprehensive annotation and comparison of TFs and MFs among metazoans would lead to a clearer understanding of developmental processes.

To date, several TF databases, such as AnimalTFDB [35], DBD [36] and TFCat [37], have been established. On the basis of DNA-binding domains (DBD) and sequence similarity, many TFs have been discovered in animals [35], plants [38-41], bacteria [36] and archaea [42]. However, prediction of TFs based only on DNA-binding domains can be misleading, since some non-TF proteins may also have similar domains. For example, the C2H2 type zinc finger domain may also be present in some RNA-binding proteins [43]. Likewise, the homology-based BLAST search method may fail to list every TF in a genome due to the fact that the sequences of some TFs are not so conserved. Therefore, more intelligent methods are needed in order to facilitate better prediction.

The supervised machine learning method combined with feature selection has been demonstrated to be a powerful tool for resolution of various biological problems, especially for placing genes into distinct categories [43, 44]. Given that TFs have features such as Pfam ID [45] and Gene Ontology term ID [46] usage that distinguish them from other genes, we have improved the previous method by assigning a different weight to each feature, depending on the category. For example, the GO term GO:0006355 (regulation of transcription, DNA-templated) should appear more frequently in TFs other than non-TFs. This method is based on statistical information similarity (SIS), and its performance has been evaluated.

To gain a better understanding of the roles of every TF and MF, we have developed a developmental ontology browser using the present data, allowing retrieval of information at the cell, tissue, organ, and system levels in a hierarchical way. All developmental ontology terms, as well as other detailed information, can be accessed via the REGULATOR web interface.

2.2 Materials and methods

2.2.1 Prediction methods for transcription factors

2.2.1.1 Prediction strategy

The TF prediction workflow employed in the present study using the supervised machine learning method combined with feature selection is shown in Figure 2-1. First, genes of 77 metazoan species from public databases were collected and redundant sequences were removed. Second, Pfam and GO annotation of the non-redundant sequences were assigned in order to ensure that every protein was represented by at least one feature (Pfam or GO ID). Subsequently, all proteins were categorized into four groups (transcription factors,

transmembrane proteins, enzymes, and other proteins), and features that are well represented in each group were selected using feature selection. Third, the weights of annotated features were calculated from the occurrence possibilities for each category. Fourth, every protein was re-encoded according to the selected features. Fifth, TFs based on statistical information similarity were predicted and the performance was evaluated using Leave-One-Out Cross-Validation (LOOCV) [44, 47, 48] in order to determine features showing the best LOOCV performance. Finally, TFs were predicted using the selected features. Details of these steps are described in the following sections.

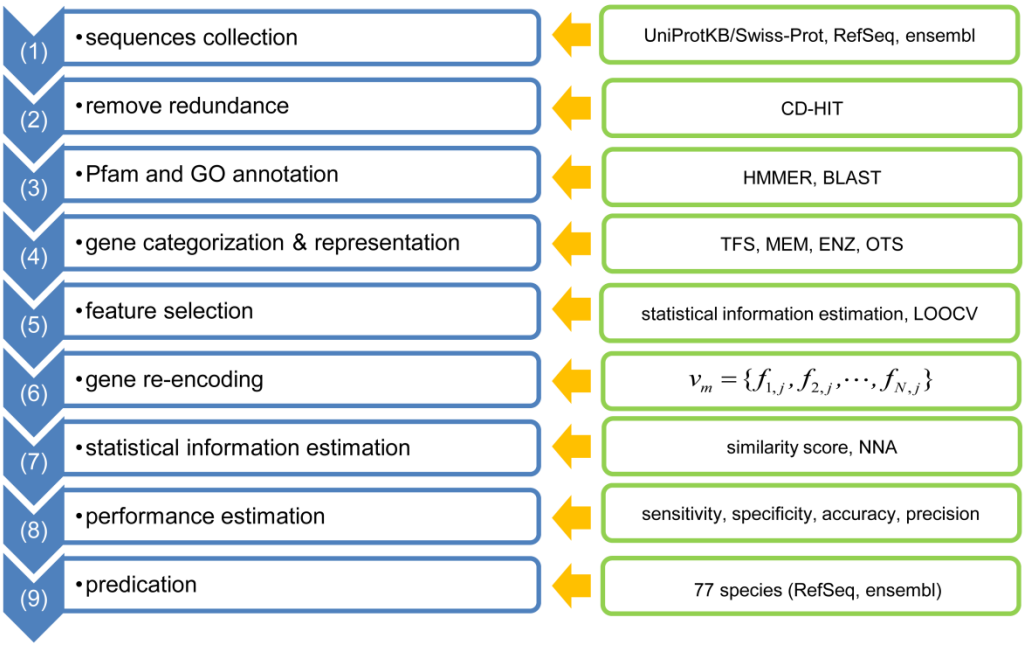


Figure 2-1 Outline of TF prediction strategy.

1.2.1.2 Dataset and preprocessing

Protein sequences of all metazoan genes were collected from the UniProtKB/Swiss-Prot (Release 2013/08), NCBI RefSeq (Release 60) and Ensembl (Release 72) databases. In addition, TF sequences from Ensembl annotated by the Animal Transcription Factor DataBase (AnimalTFDB) [35] were collected as a complement. Amino acid sequences whose length was not between 50 and 5000 or those containing irregular characters (e.g. '*') were excluded. Sequences with high similarity were clustered by CD-HIT [49] at a sequence identity threshold of 0.90. Redundant sequences in each cluster were removed, and only the longest one was retained. These genes were categorized into four groups (transcription factors, transmembrane proteins, enzymes, and other proteins) as the training dataset using the methods described below.

1.2.1.3 Pfam and GO annotation

All sequences were searched against the Pfam profile HMM database (Release 27.0) by hmmscan in the HMMER package (v3.1b1) with an e-value threshold of 1e-3. Generally, GO terms could be inferred using either InterProScan or BLAST-based methods. Considering that InterProScan is also based on conserved domains, which are redundant to some degree, we

conducted a BLAST-based homology search for GO terms annotation, which provide information complementary to the Pfam domain-based method. All non-redundant proteins were queried against UniProtKB/Swiss-Prot metazoan proteins with BLASTP. Because the number of experimentally validated GO terms is very limited, we also adopted IEAs (Inferred from Electronic Annotation). However, IEAs are often error prone. To ensure more reliable annotation, we used following criteria: (1) We used an e-value of 1e-10 as a threshold. (2) We retained only the top 10 hits. (3) Only GO terms that occurred in no less than 50% of the hit genes were considered to be features of the query gene. (4) Features presented in less than 20 genes were removed. (5) Genes without any features were excluded from the initial training dataset. All of these criteria contribute to support the accuracy. Thus, when inadequate terms were assigned, they would be removed by these criteria. Furthermore, even when minority of GO terms were not correctly assigned, the final score will be determined largely depending on major correctly assigned terms with high weights during the final step of prediction of TFs.

1.2.1.4 Classification of genes

In order to clarify the features that distinguish TFs from other proteins, we first categorized the proteins into four groups: transcription factors (TFS), transmembrane proteins (MEM), enzymes (ENZ) and other proteins (OTS) not belonging to any of the first three groups (Table 2-1). TFS Group: Well-known TFS, including general transcription factors, such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH [50], were collected from AnimalTFDB (Ensembl IDs), NCBI and UniProtKB/Swiss-Prot based on their functional descriptions or annotations. Then, all Ensembl, NCBI RefSeq and UniProtKB/Swiss-Prot genes whose Pfam or GO descriptions were related to "transcription factor activity" (Table 2-2), or whose names contained the key words "transcription factor" or "transcription initiation factor" were considered to be TFS. Transcription cofactors whose descriptions contained "cofactor", "coregulator", "coactivator" or "corepressor" were categorized into OTS. MEM Group: Proteins whose UniProtKB/Swiss-Prot or NCBI RefSeq descriptions contained "membrane", whose GO terms included "integral to membrane", and whose keywords contained "transmembrane", and those predicted to be transmembrane proteins by TMHMM [51], were considered to be MEM. ENZ Group: NCBI enzymes were identified from the RefSeq descriptions, and UniProtKB/Swiss-Prot enzymes were easily identified from the 'EC' identifier. OTS Group: Homologs (with at least two hits, and no less than half of the top ten hits belonging to at least one category with a BLASTP identity ≥ 25% and an E-value ≤ 1e-20) of the above categories were grouped as TFS, MEM and ENZ, and the other proteins were considered to be OTS.

Table 2-1 Categories and sample numbers of selected proteins in the training dataset.

Groups	Description	Number (Total: 556,753)
TFS	Transcription factors	64,596
ENZ	Enzymes	119,669
MEM	Transmembrane proteins	269,080
OTS	None of the above proteins	113,892

Note: Some proteins are categorized into more than one group.

Table 2-2 Collected known representative Pfam IDs of transcription factors.

Pfam acc	Pfam ID	Pfam description	GO ID
PF00046	Homeobox	Homeobox domain	GO:0003700
PF00104	Hormone_recep	Ligand-binding domain of nuclear hormone receptor	GO:0003700
PF00105	zf-C4	Zinc finger, C4 type (two domains)	GO:0003700
PF00157	Pou	Pou domain - N-terminal to homeobox domain	GO:0003700
PF00170	bZIP_1	bZIP transcription factor	GO:0003700
PF00172	Zn_clus	Fungal Zn(2)-Cys(6) binuclear cluster domain	GO:0000981
PF00178	Ets	Ets-domain	GO:0003700
PF00250	Fork_head	Fork head domain	GO:0003700
PF00320	GATA	GATA zinc finger	GO:0003700
PF00447	HSF_DNA-bind	HSF-type DNA-binding	GO:0003700
PF00554	RHD	Rel homology domain (RHD)	GO:0003700
PF00853	Runt	Runt domain	GO:0003700
PF00859	CTF_NFI	Runt domain	GO:0003700
PF00907	T-box	T-box	GO:0003700
PF01017	STAT_alpha	STAT protein, all-alpha domain	GO:0003700
PF01056	Myc_N	Myc amino-terminal region	GO:0003700
PF01166	TSC22	TSC-22/dip/bun family	GO:0003700
PF01285	TEA	TEA/ATTS domain family	GO:0003700
PF01422	zf-NF-X1	NF-X1 type zinc finger	GO:0003700
PF01530	zf-C2HC	Zinc finger, C2HC type	GO:0003700
PF02023	SCAN	SCAN domain	GO:0003700
PF02045	CBFB_NFYA	CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B	GO:0003700
PF02200	STE	STE like transcription factor	GO:0003700
PF02319	E2F_TDP	E2F/DP family winged-helix DNA-binding domain	GO:0003700
PF02864	STAT_bind	STAT protein, DNA binding domain	GO:0003700
PF02865	STAT_int	STAT protein, protein interaction domain	GO:0003700
PF03529	TF_Otx	Otx1 transcription factor	GO:0003700
PF03792	PBC	PBC domain	GO:0003700
PF04621	ETS_PEA3_N	PEA3 subfamily ETS-domain transcription factor N terminal domain	GO:0003700
PF06546	Vert_HS_TF	Vertebrate heat shock transcription factor	GO:0003700
PF06621	SIM_C	Single-minded protein C-terminus	GO:0003700
PF07531	TAFH	NHR1 homology to TAF	GO:0003700
PF07716	bZIP_2	Basic region leucine zipper	GO:0003700
PF09270	BTD	Beta-trefoil DNA-binding domain	GO:0000982

PF09271	LAG1-DNAbind	LAG1, DNA binding	GO:0003700
PF10401	IRF-3	Interferon-regulatory factor 3	GO:0003700
PF12162	STAT1_TAZ2bind	STAT1 TAZ2 binding domain	GO:0003700

Note: GO:0003700: sequence-specific DNA binding transcription factor activity. GO:0000981: sequence-specific DNA binding RNA polymerase II transcription factor activity. GO:0000982: RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity.

1.2.1.5 Mathematical representation of genes characterized by features

We have shown a concrete example of mathematical procedures using specific genes in CHAPTER 1.2.2 in order to help understanding what we did in this and following sections. In order to facilitate interpretation by the computational program, a binary gene coding system [44] were employed. Given that a total of N features (a feature being the Pfam or GO term ID) were annotated in a total of M genes, and the features were sorted in alphabetical order, each gene sample was converted to an N dimensional vector, as shown in formulae (1) to (4):

$$v_m = \{f_1, f_2, \dots, f_i, \dots, f_N\} \tag{1}$$

$$i \in \{1,2,3 \dots, \mathrm{N}\} \tag{2}$$

$$m \in \{1,2,3, \dots, \mathrm{M}\} \tag{3}$$

$$f_i \in \{0,1\} \tag{4}$$

where v_m is the m-th gene sample out of the total of M samples, and f_i is the i-th annotated feature out of the total of N annotated features. If sample v_m is annotated with the i-th feature, then $f_i = 1$, otherwise $f_i = 0$.

1.2.1.6 Estimation of statistical information

As the frequency of occurrence of each feature differs in each of the four categories (TFS, MEM, ENZ, OTS), the weights of the feature in each category would also differ accordingly. In this study, we measured the weights based on Information content (IC), which has been widely adopted in bioinformatics as well as many other sciences that employ information measuring [52]. Here, statistical information was estimated using the formulae (5) to (9):

$$P_{i,j} = \frac{C_{i,j}}{N_j} \cdot \frac{C_{i,j}}{C_i} = \frac{C_{i,j}^2}{N_j \cdot C_i} \tag{5}$$

$$IC_{i,j} = -log_2 P_{i,j} \tag{6}$$

$$w_{i,j} = \begin{cases} \frac{1}{IC_{i,j}}, & P_{i,j} > 0 \\ 0, & P_{i,j} = 0 \end{cases} \tag{7}$$

$$W_{m,j} = \{w_{1,j}, w_{2,j}, \dots w_{N,j}\} \tag{8}$$

$$j \in \{TFS,\ MEM, ENZ,\ OTS\} \tag{9}$$

where $C_{i,j}$ is the present frequency of the i-th feature in category j, N_j is the total number of sample proteins in category j, and C_i is the total number of the i-th feature in the four categories. $P_{i,j}$ is the joint probability of the i-th feature in category j, and it balances both inter-category and intra-category probabilities. $IC_{i,j}$ and $w_{i,j}$ are the information content and weight of the i-th feature in category j, respectively. $W_{m,j}$ is the N dimensional weight vector of the m-th sample in category j. For each sample protein, four weight vectors were assigned

because there were four categories and the possibility of each feature being present in each category would differ.

1.2.1.7 Feature selection

Next, we tried to select the best features that would yield the best prediction performance. However, feature selection software packages, such as TOOLDIAG [53], mRMR (maximum relevance minimum redundancy) [54] and Weka [55], were time-consuming and incapable of processing large datasets due to the limited memory of our computational server. Therefore, a locally developed Perl pipeline was introduced to carry out this selection. For each feature, we defined MWD_i to measure the degree of mutual weight difference between the four categories, as described in formula (10):

$$MWD_i = w_{i,j1} - \left(w_{i,j2} + w_{i,j3} + w_{i,j4}\right)/3 \tag{10}$$

where $w_{i,j1}$, $w_{i,j2}$, $w_{i,j3}$ and $w_{i,j4}$ were the sorted weights in descending order of the i-th feature in categories j1, j2, j3 and j4, respectively. Finally, according to MWD_i, a list of sorted features was generated. In order to reduce the search space, features whose first weight was less than the sum of the others were removed.

Next, LOOCV was carried out and the top best features corresponding to the highest accuracy were selected. Details of this method have been described previously [43, 44].

1.2.1.8 Prediction based on similarity score estimation

Prediction was carried out using the training data set by estimating and comparing the feature similarity between two proteins. The cosine correlation coefficient function [44, 56] was introduced to quantify the similarity of two feature vectors, and a final similarity score was calculated between protein a and protein b, as shown in formulae (11) and (12):

$$sim_{(a,b)} = \frac{V_a \cdot V_b}{\left|\left|V_a\right|\right| \cdot \left|\left|V_b\right|\right|} \tag{11}$$

$$SCORE_{(a,b,j)} = sim_{(a,b)} \cdot \sum_{k}^{N} w_{k,j} \tag{12}$$

where v_a and v_b represent the N dimensional binary vector of gene a and gene b, respectively, and $||v_a||$ and $||v_b||$ represent the module of vector v_a and v_b, respectively. $v_a \cdot v_b$ is the product of vector v_a and v_b, and $||v_a|| \cdot ||v_b||$ is the product of their modules $||v_a||$ and $||v_b||$. k is the k-th feature present in both protein a and protein b. $w_{k,j}$ is the weight of the k-th feature in category j (assuming that protein a is the query, and protein b belongs to category j). Since the weight of each feature differs in each of the four categories, four different scores are obtained. LOOCV was carried out by employing the Nearest Neighbor Algorithm (NNA) classifier [44] using the similarity score mentioned above. Query genes were considered to belong to the category with the maximum score.

1.2.1.9 Performance evaluation

To evaluate the performance of our predictions, sensitivity, specificity, accuracy, precision and the Matthews correlation coefficient (MCC) [44, 57-59] were introduced in this study, as shown in formulae (13) to (17) respectively:

$$sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$specifity = \frac{TN}{TN + FP} \tag{14}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$precision = \frac{TP}{TP + FP} \tag{16}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{17}$$

where TP (true positive) is the number of proteins correctly predicted to be TF, FP (false positive) is the number of proteins incorrectly predicted to be TF, TN (true negative) is the number of proteins correctly predicted to be non-TF, and FN (false negative) is the number of proteins incorrectly predicted to be non-TF. The quality was measured by MCC.

We determined features that showed the best LOOCV performance. Finally, TFs were predicted using the selected features in the same way as described above in "Prediction Based on Similarity Score Estimation".

2.2.2 Examples of mathematical details

Step 1: Mathematical Representation of Genes Characterized by Features

To show an example, we chose four genes shown below with various annotated Pfam and GO terms:

TFS GROUP:
Sox2:
{ PF00505, PF12336, GO:0006355, GO:0043565 }

MEM GROUP:
TMEM165:
{ PF01169, GO:0006487, GO:0010008 }

ENZ GROUP:
NDUFC2:
{ GO:0006120, GO:0022904, GO:0044237 }

OTS GROUP
GGA2:
{ PF02883, GO:0006886 }

So the total sorted feature set (including 12 features) is:

{ PF00505, PF01169, PF02883, PF12336, GO:0006120, GO:0006355, GO:0006487, GO:0006886, GO:0010008, GO:0022904, GO:0043565, GO:0044237 }

If a gene was annotated with the features, the corresponding position in the vector will be 1, otherwise, it will be 0, So,

The vector for Sox2 is: v_1={1,0,0,1,0,1,0,0,0,0,1,0} corresponding to Formula 1
The vector for TMEM165 is: v_2={0,1,0,0,0,0,1,0,1,0,0,0}
The vector for NDUFC2 is: v_3={0,0,0,0,1,0,0,0,0,1,0,1}
The vector for GGA2 is: v_4={0,0,1,0,0,0,0,1,0,0,0,0}

Thus, structural features (Pfam terms) and annotation features (GO terms) are used for combined inference.

Given that Sox 17 is a test gene, we do not know which group it belongs to.

Sox17 annotation is,

Sox17: {PF00505, GO:0006355, GO:0043565}

The vector for Sox17 is: v={1,0,0,0,0,1,0,0,0,0,1,0} (when a feature was not presented in the above total sorted feature set, just ignore it):

Step 2: Weight Calculation

For example, given that there are 2, 4, 6, and 9 protein samples in TFS, MEM, ENZ and OTS category respectively.

For the first feature PF00505,

If PF00505 appears only in one of two protein samples in category TFS,

$$C_{1,TFS} = 1$$

The total number of protein samples in category TFS is 2, so

$$N_{TFS} = 2$$

The total number of the first feature (PF00505) in the four categories is 1, so

$$C_1 = 1$$

According to formulae 5-7:

$$P_{1,TFS} = \frac{C_{1,TFS}}{N_{TFS}} \cdot \frac{C_{1,TFS}}{C_1} = \frac{1}{2} \times \frac{1}{1} = 0.5 \qquad \text{Formulae 5}$$

$$IC_{1,TFS} = -log_2 P_{1,TFS} = -log_2 0.5 = 1 \qquad \text{Formulae 6}$$

$$w_{1,TFS} = \frac{1}{IC_{1,TFS}} = \frac{1}{1} = 1 \qquad \text{Formulae 7}$$

Similarly,

$$w_{1,MEM} = 0$$
$$w_{1,ENZ} = 0$$
$$w_{1,OTS} = 0$$

And one can also calculate the weights in each category for the 2nd to 12th features. The weight of PF02883 in OTS is, for example, 0.32.

The weight vector of Sox2 are,

In category TFS,

$$W_{1,TFS} = \{1,0,0,1,0,1,0,0,0,0,1,0\} \qquad \text{Formulae 8}$$

In category MEM,

$$W_{1,MEM} = \{1,0.5,0,0,0,0,0.5,0,0.5,0,0,0\} \qquad \text{Formulae 8}$$

In category ENZ,

$$W_{1,ENZ} = \{0,0,0,0,0.39,0,0,0,0,0.39,0,0.39\} \qquad \text{Formulae 8}$$

In category OTS,

$$W_{1,OTS} = \{0,0,0.32,0,0,0,0,0.32,0,0,0,0\} \qquad \text{Formulae 8}$$

These weight vector will be used for calculation of the similarity score of Sox2 (Formula 12).

Step 3: Feature Selection and Performance Evaluation
For the first feature PF00505,

$$w_{1,TFS} = 1, w_{1,MEM} = 0, w_{1,ENZ} = 0, w_{1,OTS} = 0$$

These are sorted by their weights in descending order.

$\{1,0,0,0\}$The mutual weight difference (MWD) in this case is,

$$MWD_1 = 1 - \frac{0+0+0}{3} = 1 \qquad \text{Formulae 10}$$

Similarly,

$$MWD_2 = 0.5 - \frac{0+0+0}{3} = 0.5 \qquad \text{Formulae 10}$$

$$MWD_3 = 0.32 - \frac{0+0+0}{3} = 0.32 \qquad \text{Formulae 10}$$

……..

……..

$$MWD_{12} = 0.39 - \frac{0+0+0}{3} = 0.39 \qquad \text{Formulae 10}$$

Features are sorted by their MWD in descending order. {PF00505, PF12336, GO:0006355, GO:0043565, PF01169, GO:0006487, GO:0010008, GO:0006120, GO:0022904, GO:0044237, PF02883, GO:0006886}

For the performance evaluation, features were removed one by one from the end (GO:0006886), and the sensitivity, specificity, accuracy, precision and MCC (Matthews correlation coefficient) are calculated (formulae 13 to 17) each time.

In this study, we evaluated all of the sensitivity, specificity, accuracy, precision and MCC (Figure 2-2 in the text) and deicide how many features should be used to achieve the best performance. In this study, we used the top 4,666 features.

Step 4: Prediction Based on Similarity Score Estimation
We used the cosine similarity (formulae 11) to calculate similarity between the vectors of two genes.
In this case,
Sox17: {1,0,0,0,0,1,0,0,0,0,1,0}
Sox2: {1,0,0,1,0,1,0,0,0,0,1,0}
So,
The cosine similarity of Sox17 to Sox2 is

$$sim_{(Sox17,Sox2)} = \frac{V_{Sox17} \cdot V_{Sox2}}{\left|\left|V_{Sox17}\right|\right| \cdot \left|\left|V_{Sox2}\right|\right|} = \frac{\sum_{i=1}^{n} v_{(Sox17,i)} \times \sum_{i=1}^{n} v_{(Sox2,i)}}{\sqrt{\sum_{i=1}^{n} (v_{(Sox17,i)})^2} \times \sqrt{\sum_{i=1}^{n} (v_{(Sox2,i)})^2}} \qquad \text{Fomula 11}$$

$$= \frac{1\times1+2\times(0\times0)+0\times1+0\times0+1\times1+4\times(0\times0)+1\times1+0\times0}{\sqrt{3\times1^2+9\times0^2}\times\sqrt{4\times1^2+8\times0^2}}$$

$$= \frac{3}{\sqrt{3}\times\sqrt{4}}$$

$$= 0.866025404$$

Note: n is 12 here, because there are 12 features (Pfam IDs and GO IDs) in the total sorted feature set.

Cosine similarity is always between 0 and 1. The cosine similarities of $sim_{(Sox17,Sox2)}$ and $sim_{(Sox17,TMEM165)}$ could be same.
So, we calculate the final similarity score (formulae12) to further distinguish them by introducing weight for the feature in each category.

For the three common features of Sox17 and Sox2:
{PF00505, GO:0006355, GO:0043565}
Their weights in four categories are:
TFS: {0.18702612, 0.24017966, 0.21763421}
ENZ: {0.06521342, 0.07171545, 0.00000000}
MEM: {0.04723096, 0.06278823, 0.04456352}
OTS: {0.06319745, 0.09618691, 0.04705002}
Because Sox2 belongs to TFS, we use the weight of TFS.

$$SCORE_{(Sox17,Sox2,TFS)} = sim_{(Sox17,Sox2)} \cdot \sum_{k}^{N} w_{k,TFS} \qquad \text{Formula 12}$$

=0.866025404 x (0.18702612 + 0.24017966 + 0.21763421)
=0.558447813

The vector of:
Sox17: {1,0,0,0,0,1,0,0,0,0,1,0}
TMEM165: {0,1,0,0,0,0,1,0,1,0,0,0}

The similarity of Sox17 to TMEM165 is

$$sim_{(Sox17,TMEM165)} = 0$$

Because TMEM165 belongs to MEM, we used the weight of MEM. However, there are no common feature between Sox17 and TMEM165:

$$SCORE_{(Sox17,TMEM165,MEM)} = sim_{(Sox17,TMEM165)} \cdot \sum_{k}^{N} w_{k,MEM}$$

=0 x (0)=0

The vector of:
Sox17: {1,0,0,0,0,1,0,0,0,0,1,0}
NDUFC2 : {0,0,0,0,1,0,0,0,0,1,0,1}

The similarity of Sox17 to NDUFC2 is

$$sim_{(Sox17,NDUFC2)} = 0$$

Because NDUFC2 belongs to ENZ, we used the weights of ENZ. However, there are no common feature between Sox17 and NDUFC2:

$$SCORE_{(Sox17,NDUFC2\ ,ENZ)} = sim_{(Sox17,NDUFC2\ )} \cdot \sum_{k}^{N} w_{k,ENZ}$$

=0 x (0)=0

The vector of:

Sox17: {1,0,0,0,0,1,0,0,0,0,1,0}

GGA2: {0,0,1,0,0,0,0,1,0,0,0,0}

The similarity of Sox17 to GGA2 is

$$sim_{(Sox17,GGA2)} = 0$$

Because GGA2 belongs to OTS, we used the weights of OTS. However, there are no common feature between Sox17 and GGA2:

$$SCORE_{(Sox17,GGA2\ ,OTS)} = sim_{(Sox17,GGA2\ )} \cdot \sum_{k}^{N} w_{k,OTS}$$

=0 x (0)=0

In the case of Sox17, the scores in four categories are,

in TFS group is 0.558447813

in MEM group is 0

in ENZ group is 0

in OTS group is 0

Obviously, 0.558447813 is the largest one, so Sox17 is predicated to be TFS.

In our actual analysis, we calculated the final similarity score between Sox 17 and all the other genes belonging to the four categories, and utilize highest similarity score to decide the category which Sox17 belongs to.

2.2.3 Prediction methods for maternal factors

To predict MFs, raw data of various normal cell types, tissues and development stages were used. Relevant gene expression series in the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and Affymetrix Mouse Genome 430 2.0 Array (GPL1261) were collected from the NCBI Gene Expression Omnibus (GEO) [60]. Background correction and normalization were conducted by Guanine Cytosine Robust Multi-Array Analysis (GCRMA) using the adjusted Robust Multi-array Average (RMA) algorithm [61]. Genes whose expression values were no less than four-fold in unfertilized egg/metaphase II oocytes compared with all late-stage somatic cells were considered to be MFs in order to list egg-specific transcripts, namely strictly maternal transcripts. Late-stage somatic cells excluded embryos at the 1 ~ 8-cell stage, morula stage, blastocyst stage, testis, ovary and embryonic stem cells.

2.3 Results

2.3.1 TF prediction

The categories and sample numbers of reserved proteins (total 556,753) in the training dataset are listed in Table 2-1. These samples were used for subsequent feature selection. As illustrated in Figure 2-2A, when all of the 4,666 features were selected, LOOCV accuracy and precision reached 96.5% and 87.1%, respectively, the sensitivity being almost saturated, and the specificity showing no rapid decrease. Clustering of these 4,666 features showed that each group had significantly distinct features (Figure 2-2B), especially between TFs and non-TFs, thus supporting the high accuracy of our prediction methods. Final prediction was carried out using the sequences of 77 metazoan species (60 from the Ensembl database and 17 from the NCBI RefSeq database). As a result, a total of 85,561 unique TF genes (protein IDs were converted to NCBI GeneID, and if no NCBI GeneID was available, the Ensembl gene ID was used) were identified based on the 4,666 features, and these are summarized in Table 2-3.

Table 2-3 Numbers of transcription factors predicted in 77 metazoan species.

Class	Tax ID	Organism	TF numbers	Total genes	Percentage (%)
Aves	9103	Meleagris gallopavo	809	14,123	5.73
	9031	Gallus gallus	941	15,455	6.09
	59729	Taeniopygia guttata	1,291	17,441	7.40
Sauropsida	13735	Pelodiscus sinensis	1,211	18,170	6.66
Reptilia	28377	Anolis carolinensis	1,588	18,575	8.55
Mammalia	9258	Ornithorhynchus anatinus	1,009	21,669	4.66
	9813	Procavia capensis	1,103	16,057	6.87
	9785	Loxodonta africana	1,231	20,003	6.15
	9371	Echinops telfairi	1,106	16,575	6.67
	9986	Oryctolagus cuniculus	1,141	19,213	5.94
	9978	Ochotona princeps	1,029	16,006	6.43
	10141	Cavia porcellus	1,179	18,641	6.32

	10020	Dipodomys ordii	968	15,798	6.13
	10029	Cricetulus griseus	1,307	60,626	2.16
	10090	Mus musculus	1,678	22,716	7.39
	10116	Rattus norvegicus	1,491	22,401	6.66
	43179	Ictidomys tridecemlineatus	1,236	18,786	6.58
	9544	Macaca mulatta	1,593	21,859	7.29
	9555	Papio anubis	1,585	21,785	7.28
	9595	Gorilla gorilla	1,537	20,873	7.36
	9606	Homo sapiens	1,757	22,030	7.98
	9597	Pan paniscus	1,416	20,476	6.92
	9598	Pan troglodytes	1,498	18,672	8.02
	9601	Pongo abelii	1,505	20,370	7.39
	61853	Nomascus leucogenys	1,451	18,534	7.83
	9483	Callithrix jacchus	1,520	20,935	7.26
	39432	Saimiri boliviensis	1,462	19,344	7.56
	9478	Tarsius syrichta	961	13,628	7.05
	30608	Microcebus murinus	1,128	16,319	6.91
	30611	Otolemur garnettii	1,490	19,447	7.66
	37347	Tupaia belangeri	1,005	15,471	6.50
	9615	Canis familiaris	1,402	19,786	7.09
	9669	Mustela putorius furo	1,342	19,872	6.75
	9646	Ailuropoda melanoleuca	1,360	19,317	7.04
	9685	Felis catus	1,321	19,459	6.79
	9739	Tursiops truncatus	1,266	16,550	7.65
	9913	Bos taurus	1,402	19,900	7.05
	9823	Sus scrofa	1,353	21,390	6.33
	30538	Vicugna pacos	730	11,765	6.20
	132908	Pteropus vampyrus	1,219	16,990	7.17
	59463	Myotis lucifugus	1,248	19,679	6.34
	9365	Erinaceus europaeus	843	14,601	5.77
	42254	Sorex araneus	713	13,187	5.41
	9796	Equus caballus	1,343	20,408	6.58
	9361	Dasypus novemcinctus	988	22,711	4.35
	9358	Choloepus hoffmanni	822	12,393	6.63
	9305	Sarcophilus harrisii	1,354	18,779	7.21
	13616	Monodelphis domestica	1,666	21,299	7.82
	9315	Macropus eugenii	973	15,290	6.36
Amphibia	8364	Xenopus tropicalis	1,241	18,346	6.76
Sarcopterygii	7897	Latimeria chalumnae	1,225	19,562	6.26
Actinopterygii	8090	Oryzias latipes	1,281	19,677	6.51

	8083	Xiphophorus maculatus	1,450	20,375	7.12
	8128	Oreochromis niloticus	1,551	21,420	7.24
	69293	Gasterosteus aculeatus	1,317	20,787	6.34
	31033	Takifugu rubripes	1,359	18,484	7.35
	99883	Tetraodon nigroviridis	1,408	19,602	7.18
	8049	Gadus morhua	1,309	20,095	6.51
	7955	Danio rerio	2,376	26,239	9.06
Petromyzontida	7757	Petromyzon marinus	534	10,415	5.13
Ascidiacea	7719	Ciona intestinalis	485	16,652	2.91
	51511	Ciona savignyi	441	11,616	3.80
Echinoidea	7668	Strongylocentrotus purpuratus	763	21,156	3.61
Enteropneusta	10224	Saccoglossus kowalevskii	526	22,077	2.38
Arachnida	34638	Metaseiulus occidentalis	554	11,451	4.84
Insecta	7070	Tribolium castaneum	519	9,761	5.32
	7227	Drosophila melanogaster	662	13,792	4.80
	7463	Apis florea	488	9,137	5.34
	7460	Apis mellifera	318	10,618	2.99
	30195	Bombus terrestris	529	9,433	5.61
	132113	Bombus impatiens	530	9,859	5.38
	143995	Megachile rotundata	530	9,178	5.77
	7425	Nasonia vitripennis	528	11,450	4.61
	7029	Acyrthosiphon pisum	717	15,611	4.59
Chromadorea	6239	Caenorhabditis elegans	782	20,541	3.81
Hydrozoa	6087	Hydra magnipapillata	441	16,826	2.62
Demospongiae	400682	Amphimedon queenslandica	227	9,768	2.32

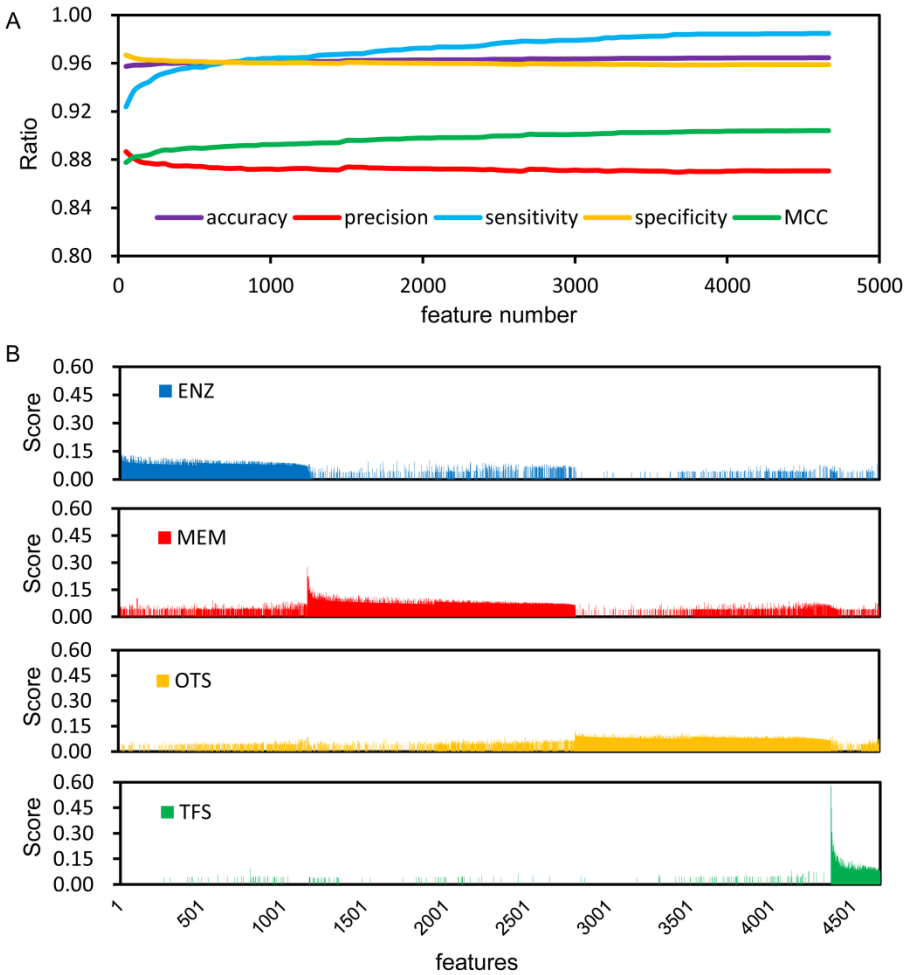


Figure 2-2 Selected features and performance curves. (A) Performance of prediction in the LOOCV. (B) Clustering of the 4,666 features according to the similarity score and categories. Blue color indicates the score in ENZ, red color indicates the score in MEM, yellow indicates the score in OTS and green color indicates the score in TFS.

2.3.2 MF prediction

MFs are already present in unfertilized eggs, and become gradually reduced as embryogenesis progresses. It has been estimated that about 60% of animal genes are expressed in unfertilized eggs [62]. In order to reduce the search space for developmentally important MFs, we focused only on strictly maternal factors, which are specifically expressed at the egg stage. Due to the limited amount of public data that have been collected at various developmental stages, only human and mouse microarray data deposited in the NCBI Gene Expression Omnibus (GEO) [60] were available. For genes examined using more than one probe and showing inconsistent expression levels between the probes, if the expression based on one probe satisfied the MF criterion, we still retained this gene, considering that the discrepancy may have been due to the presence of some alternative splicing isoforms. Finally, 542 MFs from human and 156 MFs from mouse were obtained.

2.3.3 Comprehensive annotation

In order to provide a comprehensive annotation, some basic information was extracted from the UniProtKB/Swiss-Prot database and GenBank, including the gene name, description of the

full name, and the gene ID. For each Refseq gene, we use NCBI GeneID as the unique ID, whereas for some Ensembl genes without GeneID, the Ensembl gene ID was used. In addition, cross-references to other public databases, such as Ensembl, NCBI RefSeq, UniProtKB/Swiss-Prot and KEGG were also related. A comprehensive InterPro annotation (including FPrintScan, HMMPfam, HMMSmart, ProfileScan, PatternScan, SuperFamily, SignalPHMM, TMHMM, Gene3D and so on), GO and 3D structure links to PDB were also described. Protein-protein interaction information was linked to STRING [63], MINT [64], IntAct [65] and DIP [66]. Putative orthologs were predicted using the bidirectional BLASTP best hit method with an e-value of ≤ 1e-20. Paralogs were inferred with a BLAST identity of ≥ 70% and an e-value of ≤ 1e-50. Moreover, TF targets were also collected from the Transcriptional Regulatory Element Database (TRED) [67] and Embryonic Stem Cell Atlas from Pluripotency Evidence (ESCAPE) [68]. Gene expression profiling of human and mouse TFs in various normal cell types/tissues and at various developmental stages were generated using the same method as that described for MFs prediction.

2.3.4 Developmental ontology terms

In order to gain insight into the roles of TFs and MFs during development, the developmental process-associated gene ontology terms were extracted from the Gene Ontology Consortium. These developmental ontology terms would be specifically useful for developmental biology studies. According to their anatomical hierarchies, developmental ontology terms were categorized into four groups: cell, tissue, organ and system (Figure 2-3). Each of the four groups included many terms other than non-metazoan terms, such as root, leaf and spore germination. Also, all child nodes (e.g. 'is_a' and 'part_of') of the terms were merged.

Figure 2-3 An overview of developmental ontology terms in REGULATOR database.

Terms were categorized according to four different development levels: cell, tissue, organ and system.

2.3.5 Web interface

To facilitate the use of this resource, a user-friendly web interface (Figure 2-4) was developed, which can be accessed at http://www.bioinformatics.org/regulator/. By clicking the "Browse" menu, species of all metazoan taxonomic classes used in this study are listed in the left panel. By choosing a species of interest in a certain class, detailed information on the species, including photos, taxonomic classification (kingdom, phylum, class, order, family, genus and species), and the Wikipedia link are shown in the right panel. TFs of all families identified in the species can be accessed via a panel at the bottom. TF families were designated according to the best Pfam DNA-binding domain in the panel. Lists of all TF families are displayed for each species, even when some families are not found in the species, in order to facilitate comparison between species. Using "taxonomic search" at the bottom of the left panel, TFs of selected taxon can be summarized and sorted by their prevalence according to Pfam DBDs (also shown in Additional file 3 in our paper [17]). Members of each TF family for all available species grouped by the best Pfam DNA-binding domain can also be accessed via the "TF Family" menu. Entire lists of TFs for each species can be accessed via the "Species" menu. MFs of *Homo sapiens* and *Mus musculus* can be accessed via the "Maternal" menu. Expression profiles of the annotated genes in *Homo sapiens* and *Mus musculus* in various tissues and at different developmental stages are also represented in the form of graphs. In addition, ontological representation of every TF and MF was categorized at the cell, tissue, organ, and system levels, and can be searched via the "Ontology" menu. Comprehensive annotations are provided for every TF and MF, including basic information, InterPro, Pfam, Gene ontology annotation, and cross-reference links to many public databases. Users can also search a gene of interest by entering the Gene ID, Ensembl ID, RefSeq ID, gene name, and full name via the "Search" menu. Moreover, InterPro ID, Pfam ID, Gene Ontology ID or key words of their functional annotation are also acceptable. Download and help services and external links to relevant websites are provided.

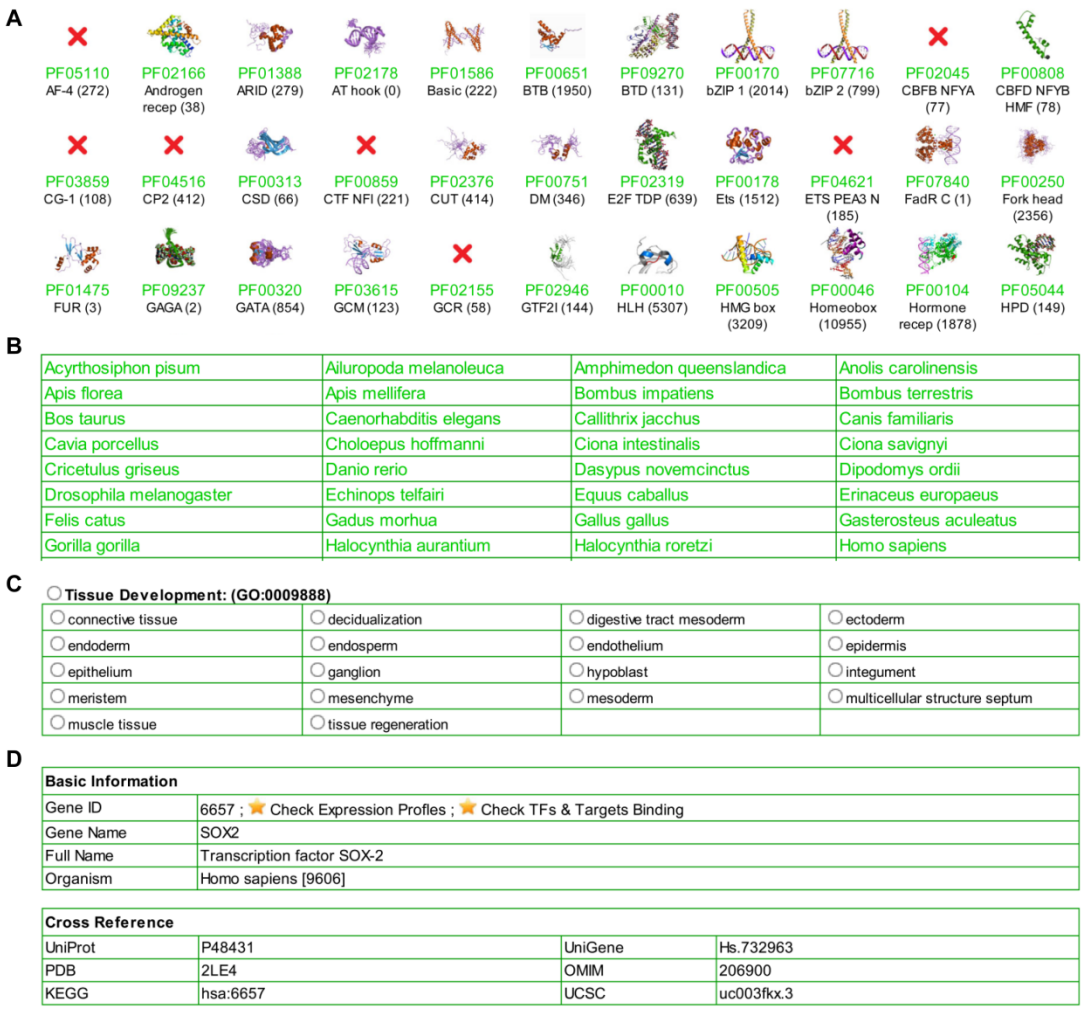


Figure 2-4 Web interface of REGULATOR. (A) Examples of TF families in REGULATOR. (B) Available species in REGULATOR. (C) Development ontology annotations for Both TFs and MFs. (D) Basic information for gene annotations.

2.4 Discussion

In this study, we selected the most relevant features that are useful for gene classification from both conserved Pfam domains and sequence similarity-based GO terms. A total of 4666 representative features were obtained, as shown in Additional file 3 in our paper [17]. As expected, most well-known features of TFs were included among the top 100 features. For example, PF00046 (Homeobox domain), PF00104 (Ligand-binding domain of nuclear hormone receptor), PF00250 (Fork head domain), PF00170 (bZIP transcription factor), GO:0003700 (sequence-specific DNA binding transcription factor activity), and GO:0006355 (regulation of transcription, DNA-dependent) were evident TF features. Furthermore, some other features were also found to be widely present in TFs. For instance, PF01352 (Krüppel associated box) domain-containing proteins were reported as transcriptional repressors in previous studies [69, 70]. In addition, reasonable Pfam IDs and GO terms were also found among the top features of other groups (ENZ, MEM, OTS), such as PF00001 (7 transmembrane receptor), GO:0022857 (transmembrane transporter activity) and GO:0004930 (G-protein coupled receptor activity) in the MEM group, and PF07714 (Tyrosine kinase) in the ENZ group.

Thus, our statistical information similarity method was capable of distinguishing proteins of different categories.

We then compared our results with other transcription factor databases. Among those whose genome sequences are available, we used 77 metazoan species in the current REGULATOR database, compared with more than 700 species in the DBD database (last updated in 2010) [71] (including eukaryotes, bacteria and archaea) and 50 animal species in the AnimalTFDB (last updated in 2012). Table 2-4 summarizes the transcription factors of the five model species and compares our prediction with the AnimalTFDB and DBD databases. In human and mouse for example, 1,706 and 1,628 TFs, respectively, were predicted in this study, among which 1,491 and 1,427 TFs were annotated with a previously known Pfam DBD, respectively. The total numbers in REGULATOR are also greater than the 1,494 human and 1,415 mouse TFs in the DBD database [71], and the 1,567 human and 1,507 mouse TFs in the AnimalTFDB database (Ensembl ID being converted to the NCBI GeneID if available). Some genes newly predicted as TFs using our approach might be true TFs. For example, ZBED6 (Zinc finger BED domain-containing protein 6) has been reported to be a transcription factor that can regulate the expression of IGF2 [72, 73]. Protein Gm5294 contains a fork-head DNA-binding domain, and may be a transcription factor, although no literature is currently available [74, 75]. Similar situations were also found for *Danio rerio, Caenorhabditis elegans* and *Drosophila melanogaster*. We retained these newly predicted genes in our dataset because they share some common features with known TFs.

Table 2-4 Comparison of transcription factors predicted in this study with those listed in AnimalTFDB and DBD in the five model species.

	Total			Common			
	R	A	D	R∩A	R∩D	A∩D	R∩A∩D
Homo sapiens	1,706	1,567	1,494	1,389	1,097	1,084	1,051
Mus musculus	1,628	1,507	1,415	1,312	1,095	1,093	1,053
Danio rerio	2,376	1,959	1,289	1,564	803	748	688
Caenorhabditis elegans	782	668	736	592	636	582	555
Drosophila melanogaster	662	631	600	513	461	457	425

Note: R: REGULATOR, A: AnimalTFDB, D: DBD.

Further investigation revealed that 111 human and 111 mouse TFs (by gene IDs, some proteins may belongs to the same gene ID) in the AnimalTFDB were not found in our dataset (obsolete gene IDs were not considered). Similarly, 68 human and 77 mouse TFs in the DBD dataset were absent in our data. A manual check of these missing genes revealed that some of them are cofactors or chromatin remodeling factors, rather than true TFs. For example, MBF1 (Endothelial differentiation-related factor 1, ENSMUSP00000015236) in DBD and ATAD2 (ATPase family AAA domain-containing protein 2, ENSG00000156802) in AnimalTFDB were suggested to be transcriptional coactivators in previous studies [74, 75]. ZZZ3 (ZZ-type zinc finger-containing protein 3, ENSG00000036549) is a protein of the histone acetyltransferase complex [76], and there is insufficient evidence for it to be a true TF, despite the fact that it is listed in the AnimalTFDB. However, some reliable TFs were still missing from our data, e.g. NFYB, NFYC (Nuclear transcription factor Y subunit beta and gamma). This may

have been due to the limited number of features assigned to these proteins. In such cases, we entered them into our database manually. In the sponge, only 227 TFs were predicted (Table 2-3). The number and proportion of TFs were significantly lower than in other animals. Therefore, the efficiency of our prediction appears to be relatively low for basal metazoans.

We also compared the TFs of *Drosophila melanogaster* in our data with FlyTF database [60] and that of mouse with TFCat database [37], which are curated databases. Among the total 1,168 TFs curated in FlyTF, 581 (50% of FlyTF and 88% of our dataset) were also discovered in our database in which 662 TFs are listed. Manual-check of the 81 TFs only present in our database showed some of them are not TFs. While some others would be genuine TFs [e.g. Tpl94D (geneid:318658) has a HMG-box domain and Aatf (geneid:33943) is an apoptosis antagonizing transcription factor], however, these are not found in the FlyTF. As to the TFs only exist in FlyTF, some of them are TFs [e.g. Mute (FBgn0085444, geneid:2768848) was not predicate by us for lack of predicted TF domain or GO term]. Others may be not TFs [e.g. Blos1 (FBgn0050077, geneid:246439) is a component of biogenesis of lysosome-related organelles complex: Med18 (FBgn0026873, geneid:31140) is coactivator, rather than a TF]. We guess that TFs in the FlyTF database could contain many non-TF proteins because the numbers of Drosophila TFs listed in the AnimalTFDB and DBD are comparable to our data (Table 2-4). In TFCat, there are 568 mouse TFs that were manually confirmed as reliable TFs. Among them, 429 (76% of TFCat and 26% of our dataset) were commonly shared with our database in which 1,628 TFs are listed. 139 TFs are only exist in TFCat, including both TFs and non-TFs [e.g. Mynf1 (myeloid nuclear factor 1, geneid:104338) is a cell type-restricted transcription factor that is not predicted by us. Trrap (geneid:100683) which belongs to a kinase protein family is not TF. Topors (geneid:106021) is a E3 ubiquitin-protein ligase]. It is likely that TFs in the TFCat database contains only firmly confirmed TFs of limited number, because the numbers of mouse TFs listed in the AnimalTFDB and DBD are relatively similar to our data (Table 2-4).

The numbers of TFs for each species sorted on the basis of prevalence according to Pfam DBDs are shown in Additional file 4 of our paper [17]. Among a total of 77 species, 26,300 (31% of total 85,561) in the zf-C2H2 family, 10,955 (13%) in the Homeobox family, 5,307 (6%) in the HLH family, and 3,209 (4%) in the HMG box family were found to be present in our data. This order of prevalence in the top 4 families is well conserved across species.

We also listed MFs specifically expressed in eggs, and provided development ontology annotations. Although many papers have reported the important roles of MFs in various development processes, a large number of MFs are still being investigated and no database has been available to date. In view of their importance and the limited extent of current knowledge, developmental ontology was adopted with the aim of providing a special annotation for these genes. The developmental ontology terms describe developmental processes at four different levels: cell development, tissue development, organ development and system development. All of these terms were extracted from the Gene Ontology consortium [77].

Finally, we have provided a well-annotated database of transcription factors and maternal factors, with cross-database links, functional annotation, protein-protein interactions, gene expression profiles in various tissues and development stages (for human and mouse only).

3 Transcriptome in the appendicularian, *O. dioica*

3.1 Introduction

3.1.1 Research background of *O. dioica*

The appendicularian, *O. dioica* (Figure 3-1), is a marine planktonic tunicate that retains a swimming tadpole shape through its entire life. This animal possesses a number of advantages as a promising model organism [18]:

1) It has a short life cycle (about 5 days at 20°C);
2) Its development is rapid and organogenesis is complete within 10 hours after fertilization to form a functional body (Figure 3-1e);
3) Its morphogenesis and cell linages are well described [18, 19, 78, 79];
4) Live imaging of embryos by introducing fluorescent protein mRNAs is feasible [80];
5) The RNAi method is available for knockdown of zygotic mRNA as well as maternal mRNAs in the ovary, eggs, and embryos [81];
6) It has a compact and fully sequenced genome of 70 Mb, the smallest ever found in chordates [20, 82]. The number of genes is estimated to be approximately 18,000, indicating a high gene density (one gene per 5 kb in the genome) [20, 82].

These features make *O. dioica* a useful organism for studies of development and genome plasticity in a tunicate with a short generation time [83].

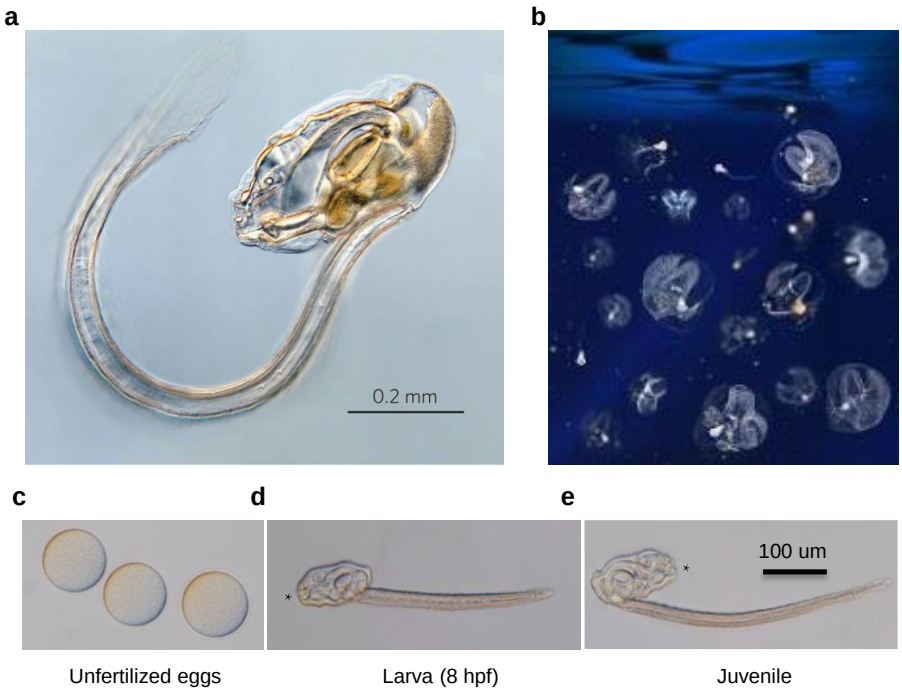
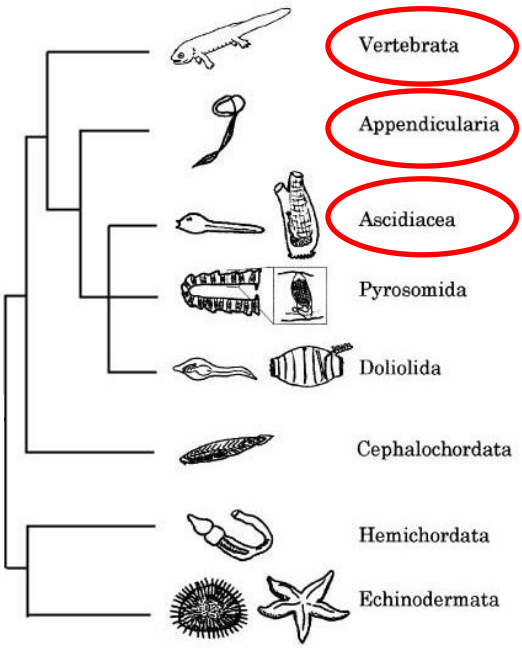


Figure 3-1. Image of *O. dioica*. (a) A lab-cultured juvenile *O. dioica* under microscope; (b) *O. dioica* in sea water. Figure a was obtained from Linda's paepr [84], Figure b was obtained from http://www.the-scientist.com/?articles.view/articleNo/29367/title/Who-needs-structure--anyway-/. (c) Unfertilized eggs. (d) Late larvae at 8 hours after fertilization (hpf) (20°C). (e) Functional juveniles at 10 hpf after the tail shift. Asterisk indicates the position of the mouth.

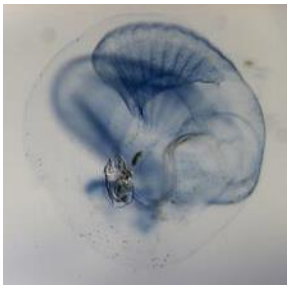
However, the roles of *O. dioica* do not only restrict to this research field. Together with the ascidians (such as *Halocynthia roretzi, Ciona intestinalis* and *Ciona savignyi*) *O. dioica* has been used to investigate the evolutionary origins of chordates from non-chordates as well as that of the tunicates and vertebrates, since ascidians, pyrosomes, doliolida and appendicularians are closest relative of vertebrates (Figure 3-2). According to the systematic taxonomy, *O. dioica* can be categorized into tunicates (http://en.wikipedia.org/wiki/Oikopleura_dioica), as summarized in the Table 3-1. The tunicates are members of the Chordata phylum together with vertebrates [85]. Comparing the genome of vertebrates and tunicates would provide important cues on how animals in these two taxonomic groups are originated. The evolutionary history of the chordates has been summarized by Nori Satoh using the ascidian *Ciona intestinalis* [86].

Table 3-1 Taxonomy of the *O. dioica.*

Category	Content
Kingdom (界):	Metazoa
Phylum (門):	Chordata
Subphylum (亜門):	Tunicata
Class (綱):	Appendicularia
Order (目):	Copelata
Family (科):	Oikopleuridae
Genus (属):	*Oikopleura*
Species (種):	*O. dioica*



Halocynthia roretzi



Oikopleura dioica

Figure 3-2. Phylogenetic tree showing the relationship of the Ascidia, Appendicularia and Vertebrata. Photo of *Halocynthia roretzi* were obtained from http://onepoint-web.jugem.jp/?month=200808. Photo of *Oikopleura dioica* were obtained from http://www.genomenewsnetwork.org/articles/12_01/Oikopleura_d.shtml.

3.1.2 Research purpose

Genome-wide knowledge of the transcriptome provides a resource for understanding gene functions underlying the formation of a functional body. In *O. dioica*, tiled microarray analysis using genomic DNA probes has been carried out [21]. The genome browser, OikoBase, showed that 78% of predicted genes (13,081 out of 16,749 tested genes) are expressed at some point from embryogenesis through larval morphogenesis [21]. The OikoBase includes ESTs and microarray data. Deep RNA sequencing (RNA-Seq) is expected to provide more information, such as sequence varations and novel genes. Moreover, little is known about intra-species genetic diversity. In the case of ascidians, e.g., *Ciona intestinalis* and *Ciona savignyi*, there are high levels of intra-species nucleotide varations and amino acid substitution among geographically distant populations in each species (more than 10 times of those found in vertebrates) [87-90]. An *O. dioica* genome sequence data set has been available for Norwegian population [20, 82]. In-depth RNA sequencing of Japanese *O. dioica* mRNAs would not only yield a resource of maternal and zygotic transcriptomes, but also provide an opportunity for intra-species comparison of whole exon sequences, *i.e.*, the exome. It will useful to gain insight into genome plasticity in this rapidly evolving metazoan.

3.1.3 Research content

In the present study, we carried out RNA-Seq analysis of a Japanese population of *O. dioica*. In order to obtain maternal and zygotic transcript sequences, we used the whole organism at two developmental stages: the unfertilized egg and the larva during organogenesis (Figure 3-1c and d). The research contents of this part include:

1) Transcriptome assembly using the egg and larval stage data;
2) Identification of in-encoding genes and novel or missing genes by comparing with Norwegian genes in OikoBase;
3) Sequence variation analysis via comparing the Japanese and Norwegian population;
4) Identification of *trans*-spliced mRNAs and the *trans*-spliced leader and comparing the difference between the egg and larval stage data.
5) Examination whether the function of *trans*-spliced mRNAs in egg and larval stage are different or not.

De novo assembly [91] of reads recovered 16,423 proteins corresponding to 95.4% of protein-encoding genes that were predicted in the genome of Norwegian *O. dioica*. Furthermore, the depth of sequence data contributed to identification of 175 novel protein-encoding genes. Transcriptome-wide comparison revealed high levels of sequence variation between the Norwegian and Japanese *O. dioica* populations. Gene ontology (GO) analysis characterized the features of gene activities at these two developmental stages. A 5' *trans*-spliced leader (SL), as the previously reported, was also found [82, 92].

3.2 Illumina next-generation sequencing

Illumina sequencing (sequencing by synthesis) [93, 94] was based on DNA colonies and reversible terminator technology [95]. Illumina Genome Analyzer was the representative sequencing platform of the method (http://en.wikipedia.org/wiki/Illumina_dye_sequencing). The detail steps of this method are:

1) Fragmentation. Large double-stranded DNA molecules are randomly fractionated into smaller fragments with the various lengths.
2) Library preparation. Two different adaptors were ligated to both ends of the DNA fragments.
3) Bridge PCR amplification. Many primers, which are reverse compliment with the above adopters, are immobilized onto the inner surface of the flow cell channels. The double-stranded DNA molecules are then separated and become single-stranded DNA molecules after denaturation. One end of the single-stranded DNA molecules are binding to the surface of the flow cell channels, the other end will binding to another adjacent primer, and thus a bridge is formed. A single DNA colony will be formed after amplification on the bridge PCR (Figure 3-3).
4) Sequencing by synthesis. Four different kinds of fluorescently labeled reversible terminator nucleotides are then added. Only one of the nucleotides could be add to the 3' end of the DNA chain and others are washed away. These dNTPs are different from the normal ones and thus blocked further extension. Fluorescent signals are subsequently detected and analyzed by computer programs. Then, the 3' end blockers with the dyes are sheared by modified DNA polymerases, allowing the amplification continue.
5) Sequence assembly.

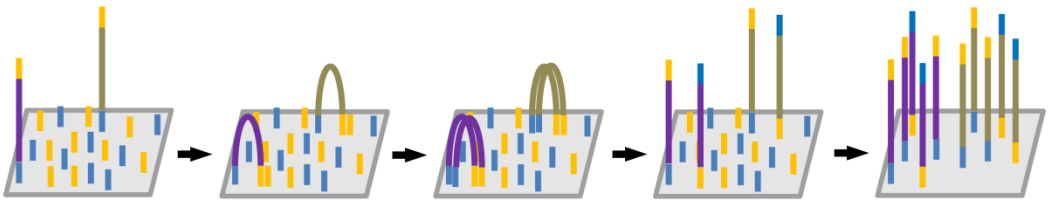


Figure 3-3. Flow chart illustrates of bridge PCR amplification used in Illumina sequencing.

3.3 Bioinformatics tools and methods

Millions of reads could be generated from the NGS platforms per day. Thus, it does not only require high performance computational servers to parse them, but also need bioinformatics tools to process and interpret these data. Many software have been developed for different purposes, as listed in Table 3-2. Among these, I used SOAPdenove-Trans and Trinity for transcriptome assembly in this study.

Table 3-2 Software that was specifically designed for NGS data analyzing.

Category	Software
Reads preprocess	Trimmomatic, cutadapt, picard, fastx_toolkit, FastQC
Genome assembly	SOAPdenovo, ABySS, SPAdes
Transcriptome assembly	Trinity, SOAPdenovo-Trans
Reads mapping	BWA, bowtie, tophat
Reads abundance estimation	htseq-count, BEDTools, DESeq
Polymorphism detection	vcftools, bcftools, GATK
File format conversion	samtools, bamtools, Trimmomatic, fastx_toolkit

3.3.1 Identification of differentially expressed genes

Identification of differentially expressed genes (DEGs) is a pivotal step for comparative analysis. Various methods accompanied with statistical tests are available for Identifying DEGs [96-103].

3.3.1.1 Reads abundance measurement

Reads per kilobase per million mapped reads (RPKM) is an unit for gene/transcript/exon level expression measurement that specifically developed for RNA-Seq data [104-106]. It supposes the total expression levels among different samples are the same. Thus, the total gene/transcript/exon length and sequenced reads number are normalized as illustrated in formula (18):

$$RPKM = \frac{10^9 \cdot N}{C \cdot L} \tag{18}$$

Where, *N* is the total number of mapped reads for a gene/transcript/exon; *C* is the total number of mapped reads in a sample; *L* is the length of that gene/transcript/exon in bp.

A modified unit of RPKM is FPKM (fragments per kilobase of transcript per million fragments mapped). The difference is counting number of reads from an ordinary library or number of fragments from a pair-end library.

However, we should keep in mind that this method supposes all mapped reads were randomly distributed throughout the entire length of a gene/transcript/exon. Thus it supposes that the 5' end and 3' end has same chance to be sequenced. Therefore, this method is only suitable for expression measurement using random cDNA fragmentation in library construction.

3.3.1.2 Fold change

Fold change (FC) method is the simplest way to obtain differentially expressed genes. Suppose the expression values of a certain gene in two different samples are e_1 and e_2, which could be obtained via the above RPKM measurement, the fold change *R* between the two samples can be calculated by using the formula (19):

$$R = \frac{e_1}{e_2} \tag{19}$$

The expression value e_2 could be very small and sometimes is zero, while the expression value e_1 is also a small but much larger than e_2. In this case, even if the fold change *R* is very large, this comparison makes no sense. To resolve this problem, a basic background expression value [107] is usually added to each of the two values. Generally, the basic

background expression value should be set a little bigger than the expression value of 2% ~ 5% expressed genes. Typically, three or more biological and/or technological replicates are required and usually companied with a t-test statistical analysis [108].

3.3.2 Gene Ontology

The Gene Ontology (GO) (http://geneontology.org/) was developed to fix the variable descriptions of gene and gene products across different databases. It categorizes gene and gene products at three basic levels: cellular component, molecular function, and biological process. Various relationship terms were used between GO terms, including "is_a", "part_of", "regulates" and so on. These relationships between them formed a Direct Acyclic Graph (DAG). These relations are stored in OBO format file, which can be downloaded from http://geneontology.org/page/download-ontology.

GO analysis is now one of the most frequently used methods for gene function annotation. Various association files to different databases are provided on the GO website (http://geneontology.org/page/download-annotations). Using BLAST, the customer's gene function could be inferred by querying it against the target sequences deposited in those databases. Besides, the European Bioinformatics Institute provides external links of variuous databases to GO terms (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/external2go/), such as InterPro, Pfam, UNIPROTKB and some other databases.

Both BLAST and InterPro based GO annotations are supported by the software Blast2GO [109]. However, latest version is not free, and users can request a free pro-trial account for only one week. A previous version is freely available from http://blast2go.com/data/blast2go/b2g4pipe_v2.5.zip, however, the users need setup the database in their own servers. Four data files are required, incuding:

1) A GO database file: http://archive.geneontology.org/latest-full/go_YYYYMM-assoc db-data.gz
2) A gene information file: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz
3) A gene2accession file: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz
4) A idmapping file: ftp://ftp.pir.georgetown.edu/databases/idmapping/idmapping.tb.gz

Sometimes, one wants to compare the differences in active gene functions between two or more samples. In this case, GO enrichment analysis are required. For most model species, this is easy by using some tools such as the Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) [110, 111], Ontologizer [112], GoMiner [113], GOSSIP [114] and others [115]. However, in non-model species, especially whose genome is newly available, no background is provided. In this case, we can choose an evolutionary close model species as the background [82]. Another choice is calculating the enriched GO terms via the p-value of a cumulative hypergeometric distribution [116], as defined in the formula (20):

$$p = \sum_{i=k}^{min(M,n)} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \qquad (20)$$

Suppose there are a total of *M* genes out of the total *N* gene models are associated with a specific GO term, and *n* genes are significantly changed. The cumulative hypergeometric distribution is the possibility that at less *k* ($0 \le k \le n$, $0 \le k \le M$) genes out of the *n* significant changed genes are annotated with this GO term. Generally the Bonferroni correction is required.

3.4 NGS and data preprocess

3.4.1 Laboratory culture and sample collection

Live wild animals were collected from surface seawater at Sakoshi Bay and Tossaki port (Hyogo, Japan). The seawater was scooped with a bucket there. The collected *O. dioica* were cultured over generations in the laboratory [81, 117]. For laboratory culture, artificial seawater was purchased from REI-SEA (Tokyo, Japan). Approximately a hundred animals are cultured in a 10-liter container filled with artificial seawater at 20˚C and fed with the flagellates *Isochrysis galbana* and *Rhinomonas reticulata*, the diatom *Chaetoceros calcitrans*, and the cyanobacterium *Synechococcus sp* [81]. The artificial seawater is stirred with a paddle (15 rpm). The expansion of the total number of *O. dioica* starts after 5 days post-fertilization at the sexually mature and spawning.

Unfertilized eggs (Figure 3-1c) and larvae at 8 hours post-fertilization (hpf) (Figure 3-1d) were used in this study. To minimize inter-individual allelic variations in the sequencing results, samples were prepared from cohorts of a single pair (Figure 3-4). Therefore, all sequence data are derived from cohorts of the same pair. A male and a female were transferred to a 2-liter container to allow spawning, and the cohorts were reared over several generations. Unfertilized eggs were obtained from 46 female cohorts. Larval samples were collected from cohorts of 63 adult pairs. Matured females were placed in gelatin-coated 6-well culture plates to allow natural spawning. Oocytes were collected in a petri dish, and fertilized by adding drops of seawater containing sperm. After three times of washing to remove sperm, they were cultured at 20 °C. In this condition, animals complete organogenesis and become functional juveniles in 10 hours (Figure 3-1e). Larvae at 8 hpf were anesthetized and collected in 1.5-ml tubes. Eggs and larvae were frozen immediately in liquid nitrogen and stored at -80 °C until isolation of total RNA.

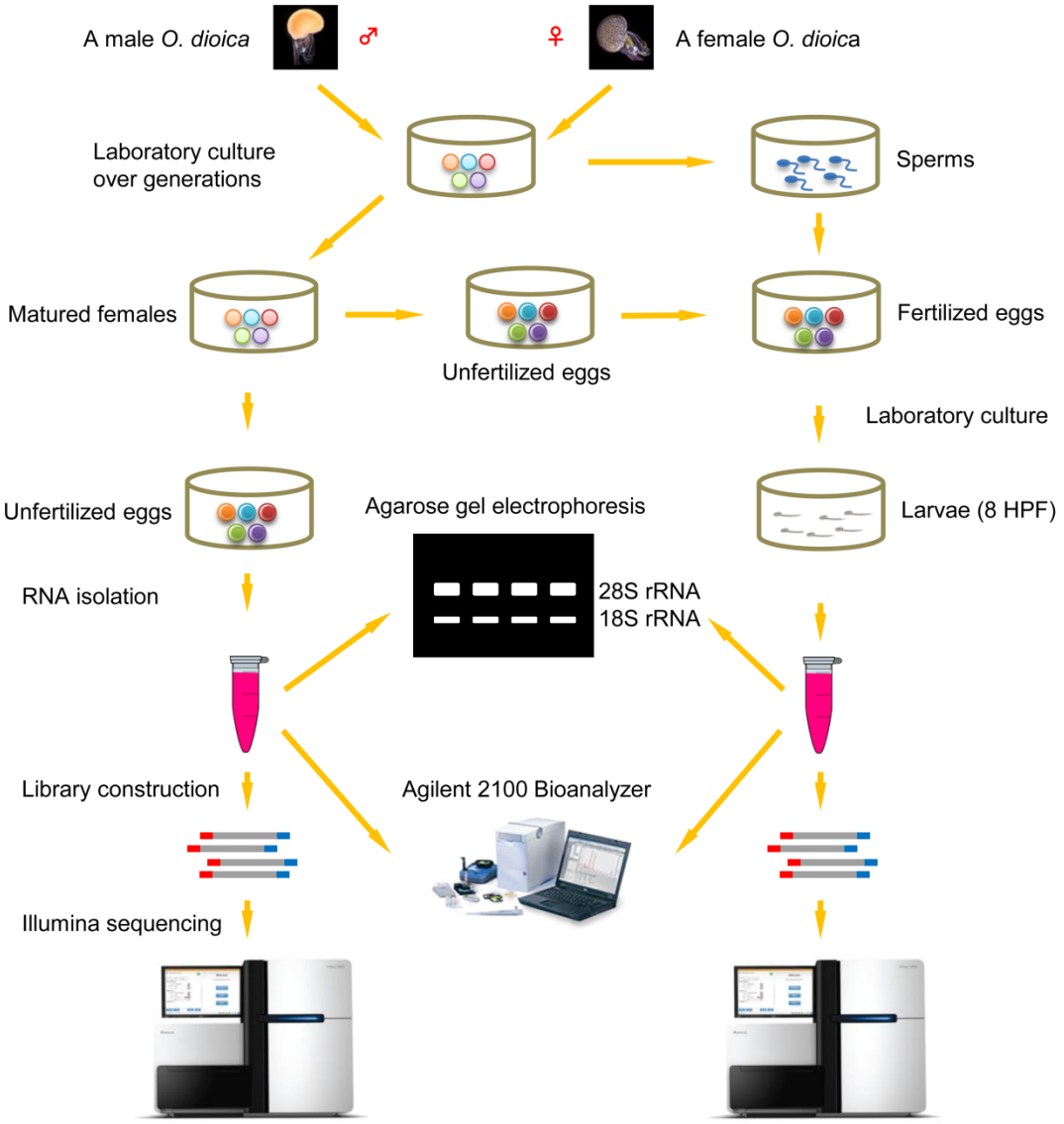


Figure 3-4. Graphic illustration of the pipeline for Illumina HiSeq 2000 sequencing. Images of male and female *O. diocia* were obtained from http://www.sars.no/research/thompsonOrg.php. Image of Agilent 2100 Bioanalyzer was obtained from http://www.genomics.agilent.com/en/Bioanalyzer-System/2100-Bioanalyzer-Instruments/?cid=AG-PT-106. Image of Illumina HiSeq 2000 sequencer was obtained from http://support.illumina.com/sequencing/sequencing_instruments/hiseq_2000.html.

3.4.2 RNA isolation

Total RNA was isolated by guanidium thiocyanate-phenol-chloroform extraction. The OD260/280 and OD260/230 were 2.09-2.31 and 1.88-1.92, respectively, ensuring purity of the RNA samples. The amount of total RNA in each sample was more than 20 μg (each sample ≥ 10 μg). Integrity of total RNA was confirmed using agarose gel electrophoresis and an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA).

3.4.3 Library construction

Illumina TruSeq RNA Preparation Kit (Illumina) was used for library construction.

Strand-specific RNA-Seq was performed at Beijing Genome Institute (BGI, Shenzhen, China) using a paired-end library, as showed in Figure 3-5. In brief,

1) mRNAs were enriched by Oligo dT MagBeads and then fragmented;
2) Strand-specific cDNAs synthesis [118] (Figure 3-5b). The first-strand cDNA was synthesized using random hexamers with normal dNTPs. And a strand-oriented library was generated during the second strand synthesis, using dUTP instead of dTTP;
3) cDNAs with the length of 200 bp were selected for library construction.

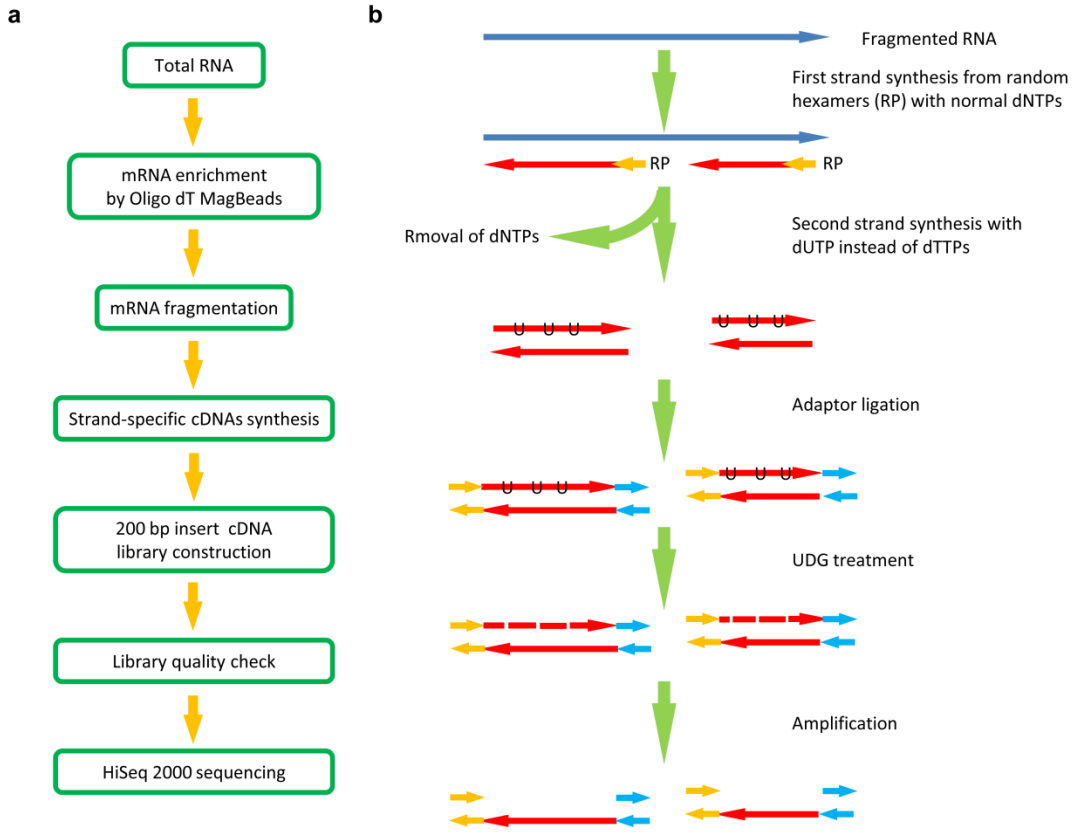


Figure 3-5. The pipeline of RNA-Seq experiment. (a) The pipeline of RNA-Seq library construction. (b) Schematic diagram of the strand-specific library construction.

3.4.4 Illumina HiSeq 2000 sequencing

Libraries for the oocyte and larva were sequenced using Illumina HiSeq™ 2000 (Figure 3-4) along with the adaptors. The insert size of the library and the read length was c.a. 200 bp and 90 bp, respectively. The raw data were deposited in NCBI Short Read Archive (SRP accession number: SRP050571, run accession numbers are SRR1693762, SRR1693765, SRR1693766 and SRR1693767) and Gene Expression Omnibus (GEO accession: GSE64421).

3.4.5 Data preprocess

The original raw data obtained from Illumina HiSeq 2000 sequencing contains adapter sequences, contamination and low-quality reads. These data is useless and can lead unexpected result when one performs sequence assembly and reads mapping. Low-quality reads also increase the possibility of false positive sequence polymorphism, which make our result suspicious. Thus they must be removed before further processing. After the data preprocess and quality control, several assessments also should be done such as base

composition analysis, quality distribution analysis and reads mapping rate statistics. All raw reads were filtered in several steps. In brief,

1) Adapters and primers were removed using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/).
2) Contamination sequences were removed by querying the NCBI UniVec (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/UniVec) and Bacteria (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz) sequences using BLASTN with a word size of 10.
3) Low-quality bases (Q < 25) at 5' and 3' ends were trimmed using a customized Perl script. Some unknown bases (Ns) were also excluded since their quality scores are much lower.
4) Sequences with a length of less than 35 nt were discarded.
5) Sequences with Q20 or unknown (Ns) bases accounting for more than 10% were also discarded.
6) Unpaired reads after the above processes were removed.

After data preprocess, the clean data were obtained, as summarized in Table 3-3.

Table 3-3 Summary of clean data obtained after data preprocessing.

Sample Name	Insert Size (bp)	Read Length (bp)	Clean Reads	Clean Bases	Q20 (%)	GC (%)
Egg	200	90	48,712,280	4,349,646,298	99	51
Larva	200	90	48,365,226	4,319,003,003	99	51
Total	200	90	97,077,506	8,668,649,301	99	51

3.5 *De novo* transcriptome assembly

3.5.1 The necessity for *de novo* transcriptome assembly

To obtain information on maternal and zygotic transcripts, RNA-Seq was carried out using poly(A)+ RNAs collected from unfertilized eggs (Figure 3-1c) and late larvae at 8 hpf (Figure 3-1d). After data filtering, approximately 97 million clean reads (~8.7 Gbp) with an average length of 89 bp were obtained (48,712,280 and 48,365,226 reads from egg and larva, respectively) (Table 3-4). We tried to map the RNA-Seq reads to the genome reference stored in the OikoBase [21]. However, only about 10% of reads could be mapped, and the result was not greatly improved even we used a much looser parameter. Moreover, genome-guided Trinity [119] assembly generated a mapping rate of only 69.9% with N50 as 371 bp (data not shown). The genome sequence in the OikoBase is derived from *O. dioica* collected in Norway. Here we analyzed specimens that were collected in Japan. Our preliminary sequencing analyses of some cDNAs, such as *brachyury*, *muscle actin 3* and testis-specific histones, demonstrated sequence variation between these geographically distant *O. dioica* populations (data not shown). The low mapping efficiency might have been due to intra-species sequence differences. In the present study, therefore, we chose *de novo* transcriptome assembly.

3.5.2 Method for *de novo* transcriptome assembly

Reads of the egg and larval stages were merged, and *de novo* transcriptome assembly was carried out using SOAPdenovo-Trans [120] and Trinity [119] assemblers. These are the most frequently used tools for transcriptome assembly. We set the k-mer parameter as 25 for both assemblers.

Specifically, for
SOAPdenovo-Trans assembly, the following parameters were used:
SOAPdenovo-Trans-31mer all -s soap.conf -o outputGraph -R RPKM.statistics -K 25 -p 4 -e 1 -L 100 >> SOAPdenovo-Trans-31mer.log

And the content of soap.conf is:
[LIB]
avg_ins=200
reverse_seq=0
asm_flags=3
rank=1
q1=../OD_Q25_trimmer_1.fq
q2=../OD_Q25_trimmer_2.fq

Trinty assembly the following parameters were used:
Trinity --seqType fq --JM 100G --SS_lib_type RF --left OD_Q25_trimmer_1.fq --right OD_Q25_trimmer_2.fq --CPU 4 --output trinity --min_contig_length 100 --full_cleanup

3.5.3 Result assessment

The results of SOAPdenovo-Trans [120] and Trinity [119] *de novo* transcriptome assembly are shown in Table 3-4. SOAPdenovo-Trans generated a total of 72,996 scaffolds with a length greater than 100 bp, and the total length of these scaffolds was ~36.16 Mbp. The length distribution of the scaffolds is shown in Figure 3-6. Among the reads, 77.1% were mapped to the assembled transcriptome using Tophat2 (version 2.0.11) [121], and 65.0% were concordant pairs. By contrast, Trinity generated a much larger transcriptome, with a total of 86,898 transcripts account for ~70.80 Mbp. N25, N50 and N75 were much larger than was indicated by SOAPdenovo-Trans. Moreover, the mapping rate of 89.2% of total reads and the mapped pair concordance rate of 81.4% were much better than for SOAPdenovo-Trans. CD-HIT-EST (version 4.6.1) [122] redundancy analysis at 90% threshold showed that 68.5% of the Trinity transcriptome was non-redundant, being much less than the value of 89.1% for the SOAPdenovo-Trans transcriptome. This was due to the difference in scaffolding strategies between SOAPdenovo-Trans and Trinity. Unlike SOAPdenovo-Trans, Trinity does not join gapped contigs with Ns, but reports alternative splicing forms as different transcripts. This is why the total length of the Trinity transcriptome is greater than that of SOAPdenovo-Trans. Considering all these factors, we consider Trinity to be a better choice for *de novo* transcriptome assembly, and used the results of Trinity for further analyses in this study.

Table 3-4 Comparison of SOAPdenovo-Trans and Trinity *de novo* assembly.

	SOAPdenovo-Trans	Trinity
Total number of clean reads	97,077,506	97,077,506
Total number of clean reads from egg	48,712,280	48,712,280
Total number of clean reads from larva	48,365,226	48,365,226
Total bases of clean reads (nt)	8,668,649,301	8,668,649,301
Total length of transcripts (nt):	36,160,616	70,804,179
Total number of transcripts:	72,996	86,898
N25 (nt)	2,380	3,191
N50 (nt)	1,267	1,806
N75 (nt)	445	931
GC count (%)	44.6	45.8
Reads mapped to transcripts (%)	77.1	89.2
Concordant read pairs (%)	65.0	81.4
Number of transcripts after removal of redundancy	65,016	59,556
Percentage of transcripts after removal of redundancy	89.1	68.5

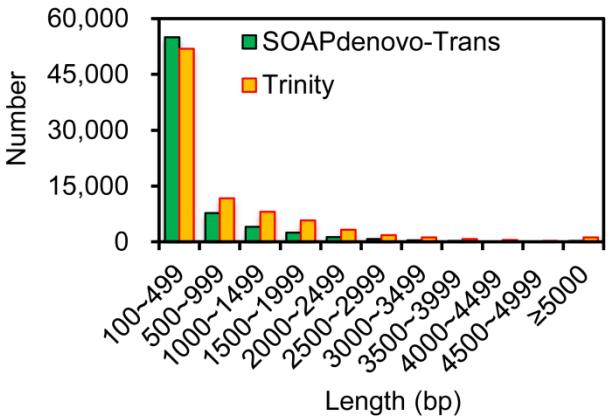


Figure 3-6. Length distribution of transcripts assembled by SOAPdenovo-Trans and Trinity.

3.6 Protein-encoding genes

3.6.1 Method for known and novel protein-encoding gene inference

To identify known potential protein-encoding genes and find out how many genes were recovered in the egg and larval stage data, all assembled transcripts were queried against OikoBase [21] proteins, and the UniProtKB/Swiss-Prot [123] and NCBI RefSeq [124] databases using BLASTX (E-value cutoff at 1E-5). Best-hit transcripts for each protein in these databases were retained as "known unigenes". "Novel unigenes" were identified among remaining transcripts using a local pipeline, which is summarized in Figure 3-7. Specifically, sequences that corresponded to isoforms of the known unigenes were first removed. Next, potential non-coding RNAs were detected and excluded with a cmsearch [125] E-value of ≤0.01 against the Rfam database [126], and only transcripts with a length of ≥500 bp were retained. Finally, three frames of these transcripts were obtained by transeq in the EMBOSS [127] package, and

hmmsearch [128] was used to identify potential Pfam domains (E-value ≤0.001). The corresponding transcripts with at least one Pfam domain were considered to be novel protein-encoding genes. Only the longest isoforms were kept in the novel unigene set.

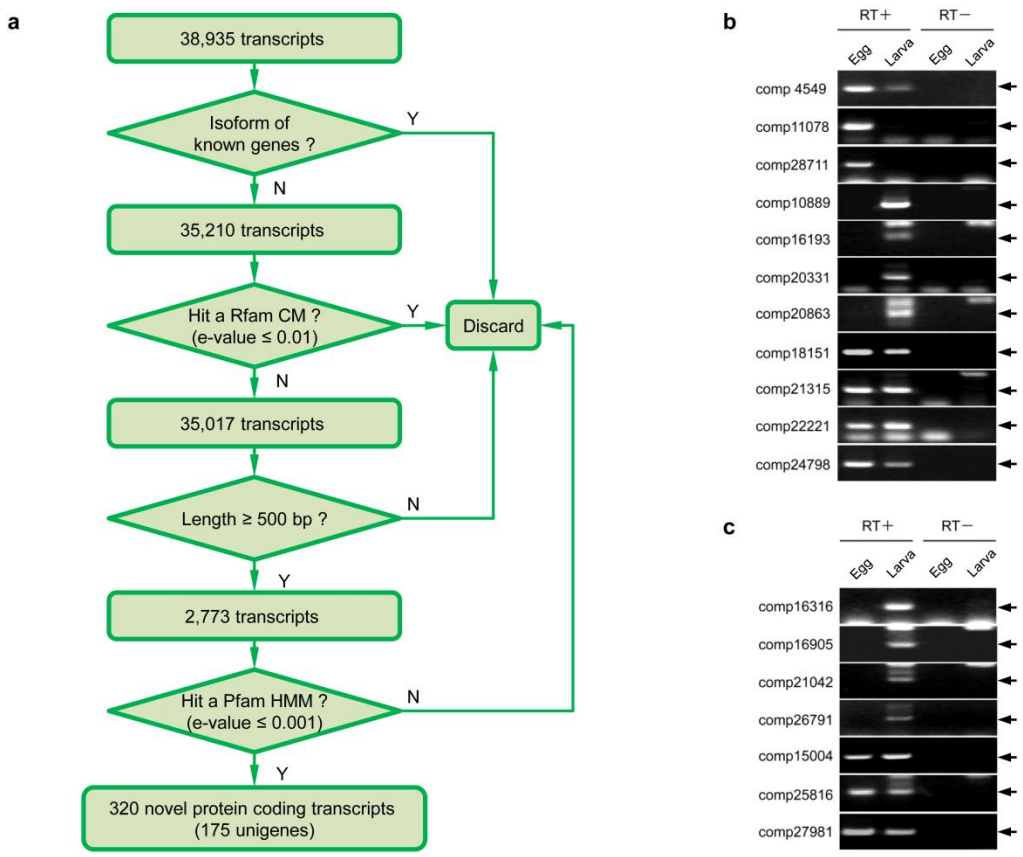


Figure 3-7. Identification and experiment validation of novel protein-encoding genes. (a) Flowchart illustration for novel protein-encoding gene identification. (b) RT-PCR validation of 11 transcripts of novel genes with in eggs or larvae (8 hpf). All of these transcripts have *trans*-splicing leaders and exists in the Norwegian genome reference. (c) RT-PCR validation of 7 transcripts of novel genes absent in the Norwegian genome reference. Arrows indicate the expected amplification. Extra bands are non-specific RT-PCR products.

3.6.2 RT-PCR validation of novel genes

We next tried to validate 18 randomly selected novel unigenes using RT-PCR assays. Eggs and 8 hpf larvae samples were prepared. TRIzol reagent was purchased from Invitrogen Life Technologies Corporation (Carlsbad, CA) for total RNA extraction. Oligo(dT)-magnetic bead method was used for mRNA purification. Purified mRNA was reverse-transcribed into single-stranded cDNA using Superscript III reverse transcriptase (Invitrogen). RT-PCR reaction was carried out in a 20μl system using 0.1 μM primer (Table 3-5), 5 ng cDNA template and 0.5 μl Taq DNA polymerase (New England Biolabs, Ipswich, MA). Specifically, initial denaturation was carried out at 94 ^{o}C for 2 min, followed by 35 cycles of amplifications at 94 ^{o}C for 30s, 55 ^{o}C for 30s, and 72 ^{o}C for 1 min. The PCR products were examined by agarose gel electrophoresis and the images were captured under the NIPPON FAS-IV gel imaging system (NIPPON Genetics, Tokyo, Japan).

Table 3-5 Primers used for RT-PCR.

RT-PCR primer name	RT-PCR primer sequence
4549 forward	5'-GGCTGAGATGGATGTGTTATTTGAGCTTCG-3'
4549 reverse	5'-TCAGGTACTTTTACGAGTTTCCGCGGCAG-3'
10889 forward	5'-TTTTTCAAGTGCGCCGCCGAAGTTTCGAC-3'
10889 reverse	5'-AGCCCGACGACGAAAACTGATATTGTGACG-3'
11078 forward	5'-GCGAACACTCAGAGCGAAGTTTTCTTCTTC-3'
11078 reverse	5'-TCGTGAAGTTACAGATAGAATGTGCCGCAG-3'
15004 forward	5'-AAAAGTCAATGGGCCAGATGCGCATGTTCC-3'
15004 reverse	5'-AGTGGTAAGTTCTACCCCTTGACGGATAGG-3'
16193 forward	5'-GAGAATCCAAAGCCTTTTTATATTGCTGAG-3'
16193 reverse	5'-TCGTTGATGTCGCTTCTGCTTCCGTGTATG-3'
16316 forward	5'-GTTACACTCTTCTCGACACTTTCATTCGTC C-3'
16316 reverse	5'-GTTTCATGGTGTTTGTCTTCGCATGTGACC-3'
16905 forward	5'-CAGCAGTATTACTGGAATGTTCCTCGGCAC-3'
16905 reverse	5'-GACGTGGCCTGGATTGTTCCATTCTTGATC-3'
18151 forward	5'-TACGTACTTCTCACAACGTTCCAACGCAGC-3'
18151 reverse	5'-TTTAAATTAACCCGTGCGGGACCAACAGCC-3'
20331 forward	5'-CGGACTACATCCCGATGTCCGTCATTATG-3'
20331 reverse	5'-AAACGGGCATATTTGACGCGGGTCGGATTC-3'
20863 forward	5'-TCATCATCTGACACAACGACAACACCAGCG-3'
20863 reverse	5'-TTCTGGTTTGCTCGTAGATTTGTTGTTCGA C-3'
21042 forward	5'-CGCCGAGCAACATGAGAAGAAATATCTCGG-3'
21042 reverse	5'-TGATAAGTCTGGTTCGCTTCGCATCTGCAG-3'
21315 forward	5'-ATTGCGAGATCTCATTATGGTGCGCCAGAG-3'
21315 reverse	5'-ATCGATCCAAAAACGAGGCCAACAGCGAAG-3'
22221 forward	5'-CGGCATATCCTCCGAATCAGCAAGGATATC-3'
22221 reverse	5'-TCGTGCTTGTCAAGATGTCTCCAGCCCTC-3'
24798 forward	5'-ACTCCAGTTAGCTCCGCCATTTCTCACATC-3'
24798 reverse	5'-GCGCTCTGCATTTCGCCCTTTGAATTTGC-3'
25816 forward	5'-CTCTGCAGATAACCGAATCGCGAAAGAAGG-3'
25816 reverse	5'-CAGTAGAGTTGGAAAACGAGATGTCGGAGG-3'
26791 forward	5'-CGATTCCGATTGAGTGGGTCCAGATCATAC-3'
26791 reverse	5'-TTACTCGCGCCGAAACTGTATTTTGCGTGC-3'
27981 forward	5'-GAACTGTCAGTTGCTGAAAAGCTGCTGAGC-3'
27981 reverse	5'-GAGCGCCTGGTGAGATTATTTTTGCTGGAC-3'
28711 forward	5'-TTGATGACCCAACTCTCTCACCAAAGTCGG-3'
28711 reverse	5'-AGAGGTATATAAGGACGCGCATCTCCTTGG-3'

3.6.3 Result

3.6.3.1 Result of known protein-encoding genes

Using a BLASTX E-value cutoff at 1E-5, we identified a total of 55.2% transcripts that had at least one hit against the three public databases shown in Table 3-6. In detail, 52.2% of the

transcripts matched OikoBase proteins [21], 29.1% matched the UniProtKB/Swiss-Prot database [123], and 31.4% matched the NCBI RefSeq database [124]; 44.8% of the 86,898 transcripts did not show any significant matches with the above databases. These non-matching transcripts accounted for only 10 Mb of the transcriptome length of 71 Mb, indicating that these unmatched fragments are much shorter than the annotated transcripts. Indeed, these fragments correspond to the short transcript groups in the length distribution histogram (Figure 3-8). Most transcripts without any significant hit to known protein-encoding genes may be non-coding RNAs (such as ribosomal RNAs, tRNAs, lncRNAs), transcripted *cis*-regulatory regions (such as UTR and intron regions) of known protein-encoding genes, novel protein-encoding genes, or transcripts that have been mis-assembled [129].

To determine the percentage recovery of *O. dioica* genes, the assembled transcripts were aligned with sequence data set in the OikoBase, which predicted 17,212 protein products in the genome of Norwegian *O. dioica* [21]. Our analysis showed that 45,368 (52.2%) of the assembled transcripts encoded proteins that have homology with 16,423 OikoBase proteins (Table 3-6). This means that 95.4% of the predicted proteins in the OikoBase were thus recovered using an E-value cutoff of 1.0E-5. Higher E-value thresholds were also tested: 48.3% and 41.1% of transcripts recovered 16,140 (93.8%) and 15,582 (90.5%) OikoBase proteins with E-values of 1.0E-10 and 1.0E-20, respectively. Our analysis further showed that the assembled transcripts recovered most cDNA clones in the Norwegian *O. dioica* ESTs (103,969 clones in total) [82]. Our transcripts showed homology with 97.4%, 96.6% and 94.4% of clones using E-value thresholds of 1.0E-5, 1.0E-10 and 1.0E-20, respectively. These results suggest that our assembly of RNA-Seq reads could recover most of the gene transcripts predicted in the genome of Norwegian *O. dioica*.

Trinity sequence assembly generated 57,962 non-redundant unigenes (according to the generated gene id by Trinity), which are about three times the number of predicted genes in OikoBase [21]. In order to remove redundancy, only unigenes that showed the best hits with OikoBase proteins were screened. Accordingly, 12,136 genes corresponding to 16,423 (95.4% of 17,212) OikoBase proteins were left (Supplementary Table S3 in our paper [22]). These were considered to be "known" unigenes. A single unigene would correspond to multiple duplicated genes, multiple paralogs, and possibly some polycistronic operons in OikoBase. In the *O. dioica* genome, 1,761 operons containing 4,997 genes have been predicted [82]. All of the 12,136 known unigenes also showed homology with proteins in the UniProtKB/Swiss-Prot or NCBI RefSeq databases with an E-value of 1.0E-5.

Table 3-6 Annotation of protein-encoding genes in the *O. dioica* transcriptome.

Database	BLASTX hits	%	E value cutoff
Total transcripts	86,898	100.0	
OikoBase proteins	35,701	41.1	1.0E-20
OikoBase proteins	41,976	48.3	1.0E-10
OikoBase proteins	45,368	52.2	1.0E-5
UniProtKB/Swiss-Prot	25,298	29.1	1.0E-5
NCBI RefSeq	27,245	31.4	1.0E-5
The above three databases	47,963	55.2	1.0E-5

Note: only metazoan proteins in NCBI RefSeq database were used.

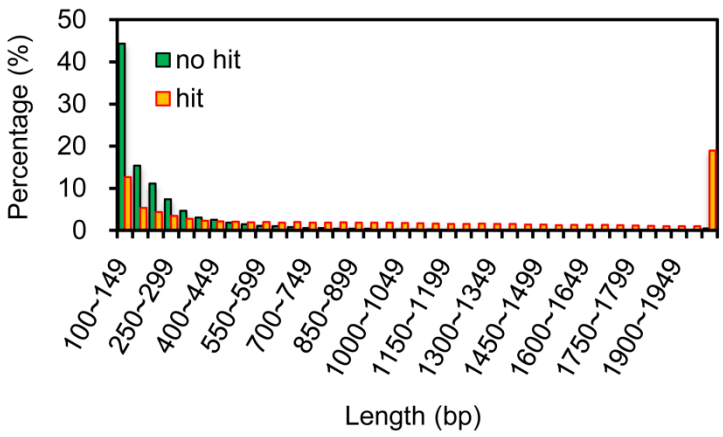


Figure 3-8. Length distribution of assembled transcripts with and without significant matches with known protein-encoding genes.

3.6.3.2 Result of novel protein-encoding genes

Next, we tested whether there were any remaining novel protein-encoding genes that had not been discovered in the above processes due to a lack of sequence homology or genome annotation. As shown in Figure 3-7a, we screened the 38,935 (44.8%) of transcripts that did not show any matches with known sequences. In brief, we removed isoforms of known genes (35,210 transcripts remained), excluded possible non-coding RNAs (35,017 transcripts remained) and eliminated transcripts that were shorted than 500 bp (2,773 transcripts remained). As a result, 320 potential protein-encoding transcripts (including possible alternative splicing forms) belonging to 175 "novel" unigenes with at least one Pfam domain were obtained (Supplementary Table S3 in our paper [22]). None of them showed any homology with known proteins in public databases (OikoBase, UniProtKB/Swiss-Prot and NCBI RefSeq databases) with an E-value of ≤1.0E-5, but they possessed known Pfam domains as a part of the protein sequence (Figure 3-7a). To ascertain whether these novel genes are expressed, 18 of them were randomly selected and RT-PCR assays were carried out. As shown in Figure 3-7b and c, all of the transcripts were detected in eggs or larvae (8 hpf), suggesting the presence of uncharacterized unigenes in Japanese *O. dioica.*

The genome annotation of Norwegian *O. dioica* revealed 175 novel unigenes that had not been found previously [82]. To determine whether these 175 novel unigenes are present in the Norwegian *O. dioica* genome, we queried all of them against the Norwegian genome reference. BLASTN (E-value ≤ 1E-30) revealed that 144 of the unigenes (including 11 PCR-confirmed genes in Figure 3-7b) are in the Norwegian genome, and 108 of them have at least one intron in the Norwegian genome reference. In contrast, the other 31 novel unigenes (including 7 PCR-confirmed genes in Figure 3-7c) were not found, or had only a small part matched in the genome reference. These results demonstrated that the depth of RNA-Seq analysis contributes to identification of novel gene products that have been missed by traditional technology. Thus, a total of 12,311 (12,136 known plus 175 novel) unique protein-encoding genes were discovered (Supplementary Table S3 in our paper [22]). These were used for subsequent expression and functional analyses.

3.6.4 Function annotation of protein-encoding genes

All of the gene annotation and subsequent expression analyses in this study were based on the total 12,311 (12,136 known plus 175 novel) unique protein-encoding genes. An integrated method based on BLAST (BLASTX, e-value of ≤1e-5) and InterPro annotation was used for retrieving GO terms of *O. dioica* protein-encoding genes. Inferred from Electronic Annotation (IEA) were removed. The numbers of genes annotated to the sub-terms of biological process (GO:0008150), molecular function (GO:0003674) and cellular component (GO:0005575) are shown in Figure 3-9. Gene symbols were extracted according to the first BLASTX hit.

A total of 7,851 unigenes were annotated. Cellular process, single-organism process and metabolic process are the dominating biological processes; Cell, cell part and organelle are the dominating cellular component; Binding and catalytic activity are the dominating molecular function (Figure 3-9). Table 3-7 lists 311 *O. dioica* putative transcription factors inferring from the above annotation. Table 3-8 showed the description of each GO term associated with the transcription factors.

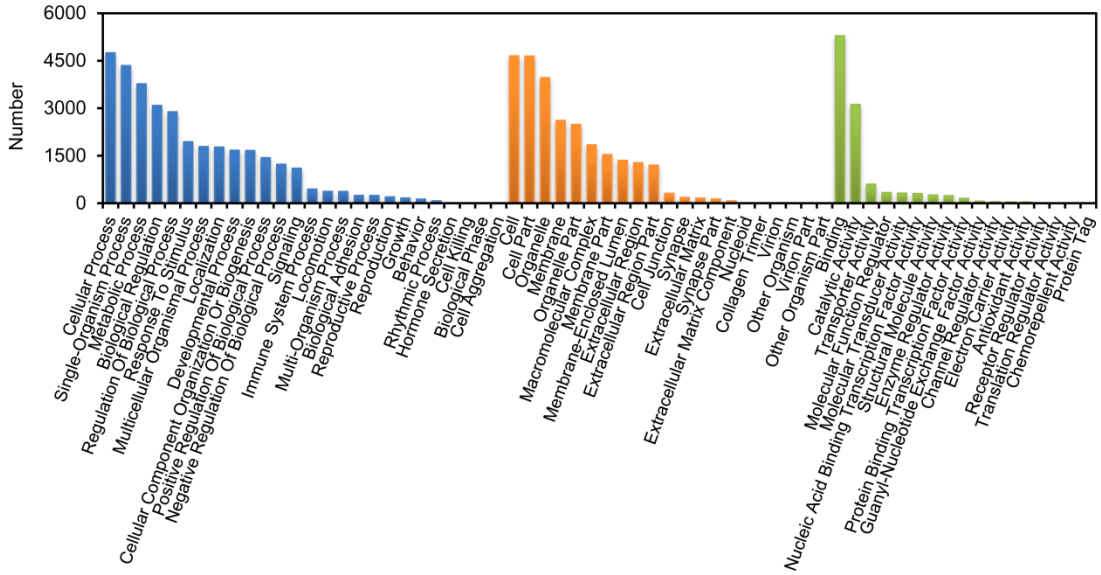


Figure 3-9. GO function annotation of *O. dioica* genes.

Table 3-7 List of GO terms and pfam domains of 311 *O. dioica* putative transcription factors with special GO terms related to transcription factor activity. The potential transcription factors annotated with a relative pfam domain but not with GO term were not listed here.

Unigene ID	GO ID	Domain	Pfam acc
comp10198_c0	GO:0001077	Ets	PF00178
comp10198_c0	GO:0000981	Ets	PF00178
comp10198_c0	GO:0003700	Ets	PF00178
comp10479_c0	GO:0001077	Homeobox	PF00046
comp10627_c0	GO:0001078	HLH	PF00010
comp10627_c0	GO:0000981	HLH	PF00010
comp10627_c0	GO:0003700	HLH	PF00010

comp10814_c0	GO:0001077	HLH	PF00010
comp10814_c0	GO:0003705	HLH	PF00010
comp10935_c0	GO:0003700	bZIP_2	PF07716
comp10962_c0	GO:0001077	Ets	PF00178
comp10962_c0	GO:0000981	Ets	PF00178
comp10962_c0	GO:0003700	Ets	PF00178
comp11019_c0	GO:0001205	NULL	NULL
comp11031_c0	GO:0003700	bZIP_1	PF00170
comp11113_c0	GO:0004879	zf-C4	PF00105
comp11113_c0	GO:0003700	zf-C4	PF00105
comp11264_c0	GO:0001077	bZIP_2	PF07716
comp11264_c0	GO:0003700	bZIP_2	PF07716
comp11408_c0	GO:0000981	Forkhead	PF00250
comp11408_c0	GO:0003700	Forkhead	PF00250
comp11442_c0	GO:0003705	Forkhead	PF00250
comp11442_c0	GO:0000981	Forkhead	PF00250
comp11442_c0	GO:0003700	Forkhead	PF00250
comp12344_c0	GO:0000989	NULL	NULL
comp12956_c0	GO:0000982	BTD	PF09270
comp12956_c0	GO:0001077	BTD	PF09270
comp12956_c0	GO:0001228	BTD	PF09270
comp12956_c0	GO:0003700	BTD	PF09270
comp13060_c0	GO:0003700	GATA	PF00320
comp13501_c0	GO:0003700	Homeobox	PF00046
comp1393_c0	GO:0001078	NULL	NULL
comp1393_c0	GO:0001191	NULL	NULL
comp1393_c0	GO:0003700	NULL	NULL
comp14033_c0	GO:0001078	Myb_DNA-binding	PF00249
comp14087_c0	GO:0001077	Ets	PF00178
comp14087_c0	GO:0000981	Ets	PF00178
comp14087_c0	GO:0003700	Ets	PF00178
comp14322_c0	GO:0000981	zf-C2H2	PF00096
comp14432_c0	GO:0001077	Homeobox	PF00046
comp14432_c0	GO:0003700	Homeobox	PF00046
comp14631_c0	GO:0001228	NULL	NULL
comp14631_c0	GO:0003700	NULL	NULL
comp14912_c0	GO:0003700	NULL	NULL
comp15353_c0	GO:0001078	zf-C2H2	PF00096
comp15399_c0	GO:0001077	zf-C2H2	PF00096
comp15652_c0	GO:0001078	T-box	PF00907
comp15652_c0	GO:0003700	T-box	PF00907
comp15737_c0	GO:0003700	BTB	PF00651
comp16089_c0	GO:0003700	GATA	PF00320
comp16104_c0	GO:0003700	CP2	PF04516

comp16123_c2	GO:0000981	zf-C4	PF00105
comp16123_c2	GO:0003700	zf-C4	PF00105
comp16295_c1	GO:0001228	NULL	NULL
comp16321_c0	GO:0003700	NULL	NULL
comp16867_c0	GO:0001227	zf-C2H2	PF00096
comp16867_c0	GO:0003700	zf-C2H2	PF00096
comp16924_c0	GO:0003705	Forkhead	PF00250
comp16924_c0	GO:0000981	Forkhead	PF00250
comp16924_c0	GO:0003700	Forkhead	PF00250
comp17214_c0	GO:0001077	NULL	NULL
comp17293_c0	GO:0003700	bZIP_2	PF07716
comp17383_c0	GO:0001078	zf-C2H2	PF00096
comp17465_c0	GO:0003700	bZIP_1	PF00170
comp17567_c0	GO:0003700	zf-C2H2	PF00096
comp17624_c0	GO:0003700	DM	PF00751
comp17666_c0	GO:0003700	NULL	NULL
comp18109_c0	GO:0001205	Homeobox	PF00046
comp18157_c0	GO:0003700	STAT_bind	PF02864
comp18193_c1	GO:0003700	CBFD_NFYB_HMF	PF00808
comp18210_c1	GO:0003700	zf-C4	PF00105
comp18244_c0	GO:0004879	zf-C4	PF00105
comp18244_c0	GO:0003700	zf-C4	PF00105
comp18332_c0	GO:0003700	bZIP_1	PF00170
comp18388_c0	GO:0001077	CUT	PF02376
comp18388_c0	GO:0003705	CUT	PF02376
comp18388_c0	GO:0001228	CUT	PF02376
comp18388_c0	GO:0003700	CUT	PF02376
comp18499_c0	GO:0001078	zf-C2H2	PF00096
comp18643_c0	GO:0001228	NULL	NULL
comp18643_c0	GO:0003700	NULL	NULL
comp18651_c0	GO:0001077	ARID	PF01388
comp18866_c0	GO:0000981	Forkhead	PF00250
comp18866_c0	GO:0003700	Forkhead	PF00250
comp19083_c0	GO:0003700	zf-C2H2	PF00096
comp19187_c0	GO:0001077	Forkhead	PF00250
comp19187_c0	GO:0000981	Forkhead	PF00250
comp19187_c0	GO:0003700	Forkhead	PF00250
comp19363_c0	GO:0003700	bZIP_2	PF07716
comp19370_c0	GO:0000981	HMG_box	PF00505
comp19370_c0	GO:0003700	HMG_box	PF00505
comp19438_c0	GO:0003700	Homeobox	PF00046
comp19482_c0	GO:0003700	bZIP_2	PF07716
comp19502_c0	GO:0001077	CUT	PF02376
comp19502_c0	GO:0003705	CUT	PF02376

comp19502_c0	GO:0001228	CUT	PF02376
comp19502_c0	GO:0003700	CUT	PF02376
comp19535_c0	GO:0000981	Forkhead	PF00250
comp19535_c0	GO:0003700	Forkhead	PF00250
comp19564_c0	GO:0001077	zf-C4	PF00105
comp19564_c0	GO:0003700	zf-C4	PF00105
comp19573_c0	GO:0003700	zf-C2H2	PF00096
comp19606_c0	GO:0003700	bZIP_1	PF00170
comp19754_c0	GO:0003700	Homeobox	PF00046
comp19828_c0	GO:0001077	MH1	PF03165
comp19828_c0	GO:0001076	MH1	PF03165
comp19828_c0	GO:0001228	MH1	PF03165
comp19828_c0	GO:0003700	MH1	PF03165
comp19847_c0	GO:0003700	Homeobox	PF00046
comp19899_c0	GO:0001228	zf-C2H2	PF00096
comp19899_c0	GO:0000981	zf-C2H2	PF00096
comp19899_c0	GO:0003700	zf-C2H2	PF00096
comp19939_c0	GO:0001077	bZIP_1	PF00170
comp19939_c0	GO:0003700	bZIP_1	PF00170
comp19952_c0	GO:0003700	Pou	PF00157
comp19960_c0	GO:0000989	NULL	NULL
comp19978_c1	GO:0003700	NULL	NULL
comp20020_c0	GO:0000982	BTD	PF09270
comp20020_c0	GO:0001077	BTD	PF09270
comp20020_c0	GO:0001228	BTD	PF09270
comp20020_c0	GO:0003700	BTD	PF09270
comp20055_c0	GO:0003700	zf-C2H2	PF00096
comp20118_c0	GO:0003700	NULL	NULL
comp20149_c0	GO:0003700	HLH	PF00010
comp20149_c0	GO:0000989	HLH	PF00010
comp20185_c0	GO:0000981	Ets	PF00178
comp20185_c0	GO:0003700	Ets	PF00178
comp20219_c0	GO:0000981	zf-C4	PF00105
comp20219_c0	GO:0003700	zf-C4	PF00105
comp20262_c0	GO:0004879	zf-C4	PF00105
comp20262_c0	GO:0001228	zf-C4	PF00105
comp20262_c0	GO:0003700	zf-C4	PF00105
comp20385_c0	GO:0001228	zf-C2H2	PF00096
comp20416_c0	GO:0001078	zf-C4	PF00105
comp20416_c0	GO:0001077	zf-C4	PF00105
comp20416_c0	GO:0003700	zf-C4	PF00105
comp20492_c0	GO:0004879	zf-C4	PF00105
comp20492_c0	GO:0003700	zf-C4	PF00105
comp20496_c0	GO:0003700	bZIP_1	PF00170

comp20545_c0	GO:0003700	NULL	NULL
comp20573_c0	GO:0001077	Pou	PF00157
comp20573_c0	GO:0003700	Pou	PF00157
comp20574_c0	GO:0003700	NULL	NULL
comp20576_c0	GO:0001228	bZIP_1	PF00170
comp20576_c0	GO:0003700	bZIP_1	PF00170
comp20662_c0	GO:0004879	Hormone_recep	PF00104
comp20662_c0	GO:0003700	Hormone_recep	PF00104
comp20685_c0	GO:0004879	Hormone_recep	PF00104
comp20685_c0	GO:0003700	Hormone_recep	PF00104
comp20743_c0	GO:0004879	zf-C4	PF00105
comp20743_c0	GO:0003700	zf-C4	PF00105
comp20841_c0	GO:0001077	bZIP_2	PF07716
comp20841_c0	GO:0003700	bZIP_2	PF07716
comp20871_c1	GO:0001227	zf-C2H2	PF00096
comp20871_c1	GO:0003700	zf-C2H2	PF00096
comp20968_c0	GO:0001078	HLH	PF00010
comp20968_c0	GO:0000981	HLH	PF00010
comp20968_c0	GO:0003700	HLH	PF00010
comp20979_c0	GO:0001077	CSD	PF00313
comp20979_c0	GO:0003700	CSD	PF00313
comp20983_c0	GO:0003700	zf-C2H2	PF00096
comp20997_c1	GO:0003700	bZIP_1	PF00170
comp21004_c1	GO:0000981	zf-C2H2	PF00096
comp21083_c0	GO:0001228	zf-C2H2	PF00096
comp21103_c0	GO:0003700	zf-C4	PF00105
comp21331_c0	GO:0001077	Homeobox	PF00046
comp21331_c0	GO:0003700	Homeobox	PF00046
comp21351_c0	GO:0000981	Ets	PF00178
comp21351_c0	GO:0003700	Ets	PF00178
comp21448_c0	GO:0003700	TEA	PF01285
comp21520_c0	GO:0003700	NULL	NULL
comp21589_c0	GO:0001077	MH1	PF03165
comp21589_c0	GO:0003700	MH1	PF03165
comp21613_c0	GO:0003700	bZIP_1	PF00170
comp21701_c0	GO:0003700	TEA	PF01285
comp21837_c0	GO:0003700	HLH	PF00010
comp21879_c0	GO:0003700	NULL	NULL
comp21932_c0	GO:0001078	HSF_DNA-bind	PF00447
comp21932_c0	GO:0001227	HSF_DNA-bind	PF00447
comp21932_c0	GO:0001228	HSF_DNA-bind	PF00447
comp21932_c0	GO:0003700	HSF_DNA-bind	PF00447
comp21949_c0	GO:0001077	NULL	NULL
comp22014_c0	GO:0001077	PAX	PF00292

comp22014_c0	GO:0001227	PAX	PF00292
comp22014_c0	GO:0001228	PAX	PF00292
comp22014_c0	GO:0000981	PAX	PF00292
comp22014_c0	GO:0003700	PAX	PF00292
comp22026_c0	GO:0001078	Homeobox	PF00046
comp22026_c0	GO:0001191	Homeobox	PF00046
comp22026_c0	GO:0003700	Homeobox	PF00046
comp22032_c1	GO:0004879	zf-C4	PF00105
comp22032_c1	GO:0001077	zf-C4	PF00105
comp22032_c1	GO:0003700	zf-C4	PF00105
comp22042_c0	GO:0003700	zf-C2HC	PF01530
comp22095_c0	GO:0003700	HMG_box	PF00505
comp22108_c0	GO:0003700	bZIP_2	PF07716
comp22269_c1	GO:0003700	NULL	NULL
comp22297_c0	GO:0001077	Ets	PF00178
comp22297_c0	GO:0000981	Ets	PF00178
comp22297_c0	GO:0003700	Ets	PF00178
comp22312_c0	GO:0000981	HMG_box	PF00505
comp22312_c0	GO:0003700	HMG_box	PF00505
comp22347_c0	GO:0001078	HLH	PF00010
comp22347_c0	GO:0001077	HLH	PF00010
comp22347_c0	GO:0003700	HLH	PF00010
comp22355_c0	GO:0003705	Homeobox	PF00046
comp22355_c0	GO:0003700	Homeobox	PF00046
comp22369_c0	GO:0003700	NULL	NULL
comp22397_c0	GO:0003700	Homeobox	PF00046
comp22402_c0	GO:0000981	zf-C2H2	PF00096
comp22458_c0	GO:0003700	zf-C2H2	PF00096
comp22467_c0	GO:0003700	zf-C2H2	PF00096
comp22505_c0	GO:0003700	zf-C2H2	PF00096
comp22522_c0	GO:0001077	RHD_DNA_bind	PF00554
comp22522_c0	GO:0003700	RHD_DNA_bind	PF00554
comp22577_c0	GO:0003700	NULL	NULL
comp22603_c0	GO:0000981	Forkhead	PF00250
comp22603_c0	GO:0003700	Forkhead	PF00250
comp22675_c1	GO:0003705	Forkhead	PF00250
comp22675_c1	GO:0000981	Forkhead	PF00250
comp22675_c1	GO:0003700	Forkhead	PF00250
comp22678_c0	GO:0001071	GATA	PF00320
comp22678_c0	GO:0001078	GATA	PF00320
comp22678_c0	GO:0001077	GATA	PF00320
comp22678_c0	GO:0003700	GATA	PF00320
comp22684_c0	GO:0001077	HLH	PF00010
comp22684_c0	GO:0003700	HLH	PF00010

comp22702_c2	GO:0000981	Forkhead	PF00250
comp22702_c2	GO:0003700	Forkhead	PF00250
comp22709_c0	GO:0001078	T-box	PF00907
comp22709_c0	GO:0003700	T-box	PF00907
comp22733_c0	GO:0003700	zf-C2H2	PF00096
comp22797_c0	GO:0003700	NULL	NULL
comp22850_c0	GO:0001077	Ets	PF00178
comp22850_c0	GO:0000981	Ets	PF00178
comp22850_c0	GO:0003700	Ets	PF00178
comp22864_c0	GO:0001077	Forkhead	PF00250
comp22864_c0	GO:0000981	Forkhead	PF00250
comp22864_c0	GO:0003700	Forkhead	PF00250
comp22878_c0	GO:0000981	Forkhead	PF00250
comp22878_c0	GO:0003700	Forkhead	PF00250
comp22934_c0	GO:0001078	T-box	PF00907
comp22934_c0	GO:0003700	T-box	PF00907
comp22948_c0	GO:0000981	Forkhead	PF00250
comp22948_c0	GO:0003700	Forkhead	PF00250
comp22955_c0	GO:0004879	zf-C4	PF00105
comp22955_c0	GO:0001077	zf-C4	PF00105
comp22955_c0	GO:0003700	zf-C4	PF00105
comp23007_c0	GO:0000981	Forkhead	PF00250
comp23007_c0	GO:0003700	Forkhead	PF00250
comp23097_c0	GO:0003700	zf-C4	PF00105
comp23117_c0	GO:0001077	bZIP_2	PF07716
comp23117_c0	GO:0003700	bZIP_2	PF07716
comp23119_c0	GO:0001077	CUT	PF02376
comp23119_c0	GO:0003705	CUT	PF02376
comp23119_c0	GO:0001228	CUT	PF02376
comp23119_c0	GO:0003700	CUT	PF02376
comp23120_c0	GO:0003705	HLH	PF00010
comp23120_c0	GO:0003700	HLH	PF00010
comp23147_c0	GO:0001077	SRF-TF	PF00319
comp23147_c0	GO:0000983	SRF-TF	PF00319
comp23147_c0	GO:0003705	SRF-TF	PF00319
comp23147_c0	GO:0001076	SRF-TF	PF00319
comp23147_c0	GO:0001228	SRF-TF	PF00319
comp23147_c0	GO:0003700	SRF-TF	PF00319
comp23194_c0	GO:0098531	zf-C4	PF00105
comp23194_c0	GO:0003700	zf-C4	PF00105
comp23203_c0	GO:0003700	zf-C2H2	PF00096
comp23225_c0	GO:0003700	bZIP_2	PF07716
comp23322_c0	GO:0004879	zf-C4	PF00105
comp23322_c0	GO:0003700	zf-C4	PF00105

comp23345_c0	GO:0001077	PBC	PF03792
comp23345_c0	GO:0003700	PBC	PF03792
comp23366_c1	GO:0003705	HLH	PF00010
comp23366_c1	GO:0001228	HLH	PF00010
comp23388_c0	GO:0001191	NULL	NULL
comp23388_c0	GO:0003700	NULL	NULL
comp23421_c0	GO:0001071	GATA	PF00320
comp23421_c0	GO:0001078	GATA	PF00320
comp23421_c0	GO:0001077	GATA	PF00320
comp23421_c0	GO:0003700	GATA	PF00320
comp23424_c0	GO:0004879	zf-C4	PF00105
comp23424_c0	GO:0000981	zf-C4	PF00105
comp23424_c0	GO:0003700	zf-C4	PF00105
comp23450_c0	GO:0003700	NULL	NULL
comp23558_c0	GO:0001077	Forkhead	PF00250
comp23558_c0	GO:0003705	Forkhead	PF00250
comp23558_c0	GO:0001205	Forkhead	PF00250
comp23558_c0	GO:0001228	Forkhead	PF00250
comp23558_c0	GO:0000981	Forkhead	PF00250
comp23558_c0	GO:0003700	Forkhead	PF00250
comp23705_c0	GO:0003700	NULL	NULL
comp23716_c0	GO:0001077	zf-C2H2	PF00096
comp23716_c0	GO:0001227	zf-C2H2	PF00096
comp23716_c0	GO:0003700	zf-C2H2	PF00096
comp23736_c0	GO:0001078	T-box	PF00907
comp23736_c0	GO:0003700	T-box	PF00907
comp23738_c0	GO:0000982	zf-C2H2	PF00096
comp23738_c0	GO:0001227	zf-C2H2	PF00096
comp23738_c0	GO:0003700	zf-C2H2	PF00096
comp23751_c0	GO:0003700	bZIP_1	PF00170
comp23767_c0	GO:0001077	zf-C2H2	PF00096
comp23767_c0	GO:0003700	zf-C2H2	PF00096
comp23772_c0	GO:0000981	bZIP_1	PF00170
comp23772_c0	GO:0003700	bZIP_1	PF00170
comp23858_c0	GO:0003700	zf-C2H2	PF00096
comp23901_c0	GO:0003700	NULL	NULL
comp23902_c0	GO:0001078	NULL	NULL
comp23902_c0	GO:0001191	NULL	NULL
comp23902_c0	GO:0003700	NULL	NULL
comp23910_c0	GO:0001077	MH1	PF03165
comp23910_c0	GO:0003700	MH1	PF03165
comp24041_c0	GO:0003700	zf-C2H2	PF00096
comp24080_c0	GO:0003700	Pou	PF00157
comp24120_c2	GO:0001077	HLH	PF00010

comp24120_c2	GO:0000981	HLH	PF00010
comp24157_c0	GO:0004879	zf-C4	PF00105
comp24157_c0	GO:0003705	zf-C4	PF00105
comp24157_c0	GO:0003700	zf-C4	PF00105
comp24179_c0	GO:0003700	Homeobox	PF00046
comp24204_c0	GO:0001077	Ets	PF00178
comp24204_c0	GO:0000981	Ets	PF00178
comp24204_c0	GO:0003700	Ets	PF00178
comp24214_c0	GO:0003705	HPD	PF05044
comp24214_c0	GO:0000981	HPD	PF05044
comp24215_c0	GO:0003700	CP2	PF04516
comp24227_c0	GO:0000981	bZIP_1	PF00170
comp24227_c0	GO:0003700	bZIP_1	PF00170
comp24236_c0	GO:0003700	zf-C2H2	PF00096
comp24241_c0	GO:0001077	Homeobox	PF00046
comp24241_c0	GO:0003700	Homeobox	PF00046
comp24257_c2	GO:0001228	Runt	PF00853
comp24257_c2	GO:0003700	Runt	PF00853
comp24327_c0	GO:0001228	bZIP_1	PF00170
comp24327_c0	GO:0003700	bZIP_1	PF00170
comp24334_c0	GO:0001227	zf-C2H2	PF00096
comp24348_c1	GO:0003700	NULL	NULL
comp24410_c1	GO:0003700	NULL	NULL
comp24443_c0	GO:0000981	NULL	NULL
comp24608_c0	GO:0003700	HLH	PF00010
comp24628_c0	GO:0003700	IRF	PF00605
comp24649_c0	GO:0003700	zf-C2H2	PF00096
comp24715_c0	GO:0000981	Ets	PF00178
comp24715_c0	GO:0003700	Ets	PF00178
comp24736_c1	GO:0000981	Ets	PF00178
comp24736_c1	GO:0003700	Ets	PF00178
comp24779_c0	GO:0000981	zf-C2H2	PF00096
comp24794_c0	GO:0003700	NULL	NULL
comp24823_c1	GO:0001077	Ets	PF00178
comp24823_c1	GO:0000981	Ets	PF00178
comp24823_c1	GO:0003700	Ets	PF00178
comp24856_c0	GO:0001078	HLH	PF00010
comp24856_c0	GO:0003700	HLH	PF00010
comp24864_c0	GO:0001078	zf-C2H2	PF00096
comp25079_c0	GO:0000981	Forkhead	PF00250
comp25079_c0	GO:0003700	Forkhead	PF00250
comp25091_c0	GO:0004879	zf-C4	PF00105
comp25091_c0	GO:0003705	zf-C4	PF00105
comp25091_c0	GO:0003700	zf-C4	PF00105

comp25312_c1	GO:0003700	bZIP_1	PF00170
comp25329_c1	GO:0001077	ARID	PF01388
comp25333_c0	GO:0001077	RHD_DNA_bind	PF00554
comp25333_c0	GO:0003700	RHD_DNA_bind	PF00554
comp25347_c0	GO:0001077	NULL	NULL
comp25379_c3	GO:0003700	HSF_DNA-bind	PF00447
comp25401_c0	GO:0001077	Homeobox	PF00046
comp25453_c0	GO:0004879	zf-C4	PF00105
comp25453_c0	GO:0001077	zf-C4	PF00105
comp25453_c0	GO:0003700	zf-C4	PF00105
comp25477_c0	GO:0003700	STAT_bind	PF02864
comp25501_c0	GO:0001077	CSD	PF00313
comp25501_c0	GO:0003700	CSD	PF00313
comp25562_c0	GO:0000981	E2F_TDP	PF02319
comp25562_c0	GO:0003700	E2F_TDP	PF02319
comp25567_c0	GO:0003700	E2F_TDP	PF02319
comp25606_c0	GO:0003700	NULL	NULL
comp25649_c2	GO:0003700	NULL	NULL
comp25706_c1	GO:0003700	NULL	NULL
comp25711_c0	GO:0001227	HMG_box	PF00505
comp25711_c0	GO:0003700	HMG_box	PF00505
comp25785_c0	GO:0001076	zf-C4	PF00105
comp25785_c0	GO:0003700	zf-C4	PF00105
comp25849_c0	GO:0001077	Forkhead	PF00250
comp25849_c0	GO:0003705	Forkhead	PF00250
comp25849_c0	GO:0001205	Forkhead	PF00250
comp25849_c0	GO:0001228	Forkhead	PF00250
comp25849_c0	GO:0000981	Forkhead	PF00250
comp25849_c0	GO:0003700	Forkhead	PF00250
comp25884_c0	GO:0001077	NULL	NULL
comp25888_c0	GO:0001227	zf-C2H2	PF00096
comp25888_c0	GO:0003700	zf-C2H2	PF00096
comp25891_c0	GO:0001077	NULL	NULL
comp25891_c0	GO:0001205	NULL	NULL
comp25913_c0	GO:0001077	NULL	NULL
comp25914_c0	GO:0001077	Myb_DNA-binding	PF00249
comp25914_c0	GO:0003700	Myb_DNA-binding	PF00249
comp25923_c0	GO:0000981	Forkhead	PF00250
comp25923_c0	GO:0003700	Forkhead	PF00250
comp25965_c0	GO:0001205	NULL	NULL
comp25966_c0	GO:0003700	bZIP_2	PF07716
comp26058_c0	GO:0003700	NULL	NULL
comp26100_c0	GO:0003700	RFX_DNA_binding	PF02257
comp26130_c0	GO:0004879	zf-C4	PF00105

comp26130_c0	GO:0003700	zf-C4	PF00105
comp26132_c0	GO:0001077	ARID	PF01388
comp26173_c0	GO:0004879	zf-C4	PF00105
comp26173_c0	GO:0003705	zf-C4	PF00105
comp26173_c0	GO:0001190	zf-C4	PF00105
comp26173_c0	GO:0003700	zf-C4	PF00105
comp26204_c0	GO:0001077	CBFB_NFYA	PF02045
comp26204_c0	GO:0003700	CBFB_NFYA	PF02045
comp26218_c0	GO:0003700	STAT_bind	PF02864
comp26223_c0	GO:0001078	T-box	PF00907
comp26223_c0	GO:0003700	T-box	PF00907
comp26251_c1	GO:0001077	MH1	PF03165
comp26251_c1	GO:0003700	MH1	PF03165
comp26291_c0	GO:0003700	MH1	PF03165
comp26325_c0	GO:0003700	zf-C2HC	PF01530
comp26330_c1	GO:0003700	NULL	NULL
comp26479_c0	GO:0003700	zf-C2H2	PF00096
comp26524_c0	GO:0001077	Forkhead	PF00250
comp26524_c0	GO:0000981	Forkhead	PF00250
comp26524_c0	GO:0003700	Forkhead	PF00250
comp26547_c0	GO:0003700	bZIP_1	PF00170
comp26633_c0	GO:0003700	NULL	NULL
comp26648_c0	GO:0003700	MH1	PF03165
comp26693_c0	GO:0001077	Myb_DNA-binding	PF00249
comp26696_c0	GO:0003700	zf-NF-X1	PF01422
comp26701_c0	GO:0003700	E2F_TDP	PF02319
comp26713_c0	GO:0003700	zf-C4	PF00105
comp26714_c4	GO:0001077	SRF-TF	PF00319
comp26714_c4	GO:0000983	SRF-TF	PF00319
comp26714_c4	GO:0003705	SRF-TF	PF00319
comp26714_c4	GO:0001076	SRF-TF	PF00319
comp26714_c4	GO:0001228	SRF-TF	PF00319
comp26714_c4	GO:0003700	SRF-TF	PF00319
comp26812_c0	GO:0001077	Forkhead	PF00250
comp26812_c0	GO:0003705	Forkhead	PF00250
comp26812_c0	GO:0001205	Forkhead	PF00250
comp26812_c0	GO:0001228	Forkhead	PF00250
comp26812_c0	GO:0000981	Forkhead	PF00250
comp26812_c0	GO:0003700	Forkhead	PF00250
comp26813_c0	GO:0003700	NULL	NULL
comp26866_c0	GO:0003700	TAFH	PF07531
comp26895_c0	GO:0003700	NULL	NULL
comp26902_c0	GO:0003700	bZIP_1	PF00170
comp26926_c0	GO:0000983	NULL	NULL

comp26926_c0	GO:0003700	NULL	NULL
comp26931_c0	GO:0001078	T-box	PF00907
comp26931_c0	GO:0003700	T-box	PF00907
comp26980_c0	GO:0003700	bZIP_1	PF00170
comp26981_c0	GO:0003700	bZIP_2	PF07716
comp27045_c0	GO:0003700	zf-C4	PF00105
comp27140_c0	GO:0001077	Myb_DNA-binding	PF00249
comp27140_c0	GO:0003700	Myb_DNA-binding	PF00249
comp27192_c0	GO:0001077	NULL	NULL
comp27212_c0	GO:0003700	zf-C2H2	PF00096
comp27213_c0	GO:0001077	Pou	PF00157
comp27213_c0	GO:0003700	Pou	PF00157
comp27231_c1	GO:0001077	NULL	NULL
comp27243_c0	GO:0004879	zf-C4	PF00105
comp27243_c0	GO:0003700	zf-C4	PF00105
comp27301_c0	GO:0000982	HLH	PF00010
comp27301_c0	GO:0003700	HLH	PF00010
comp27322_c0	GO:0003700	bZIP_1	PF00170
comp27380_c0	GO:0003700	MH1	PF03165
comp27457_c1	GO:0003700	CP2	PF04516
comp27459_c1	GO:0003700	zf-C2H2	PF00096
comp27498_c1	GO:0003700	NULL	NULL
comp27572_c0	GO:0003700	PAX	PF00292
comp27594_c0	GO:0001077	STAT_bind	PF02864
comp27594_c0	GO:0003700	STAT_bind	PF02864
comp27605_c0	GO:0001078	NULL	NULL
comp27605_c0	GO:0000981	NULL	NULL
comp27624_c0	GO:0003700	NULL	NULL
comp27639_c0	GO:0001077	SAND	PF01342
comp27639_c0	GO:0003700	SAND	PF01342
comp27682_c1	GO:0001078	HLH	PF00010
comp27682_c1	GO:0003700	HLH	PF00010
comp27694_c3	GO:0001077	zf-C4	PF00105
comp27694_c3	GO:0003700	zf-C4	PF00105
comp27699_c0	GO:0003700	NULL	NULL
comp27745_c0	GO:0003700	TF_AP-2	PF03299
comp27756_c0	GO:0003700	zf-LITAF-like	PF10601
comp27772_c0	GO:0004879	zf-C4	PF00105
comp27772_c0	GO:0001077	zf-C4	PF00105
comp27772_c0	GO:0003700	zf-C4	PF00105
comp27796_c0	GO:0003700	zf-C2H2	PF00096
comp27891_c0	GO:0000981	NULL	NULL
comp27934_c4	GO:0001026	Myb_DNA-binding	PF00249
comp27954_c0	GO:0003700	NULL	NULL

comp28022_c0	GO:0038049	NULL	NULL
comp28113_c0	GO:0001077	MH1	PF03165
comp28113_c0	GO:0001228	MH1	PF03165
comp28113_c0	GO:0000981	MH1	PF03165
comp28113_c0	GO:0003700	MH1	PF03165
comp28117_c0	GO:0003700	NULL	NULL
comp28204_c0	GO:0003700	NULL	NULL
comp28214_c0	GO:0003700	NULL	NULL
comp28268_c0	GO:0003700	NULL	NULL
comp28279_c0	GO:0001190	NULL	NULL
comp28334_c1	GO:0003700	CBFD_NFYB_HMF	PF00808
comp28370_c2	GO:0003700	NULL	NULL
comp28614_c0	GO:0001077	NULL	NULL
comp28631_c0	GO:0003700	NULL	NULL
comp28716_c0	GO:0004879	zf-C4	PF00105
comp28716_c0	GO:0001228	zf-C4	PF00105
comp28716_c0	GO:0003700	zf-C4	PF00105
comp28870_c0	GO:0003700	bZIP_1	PF00170
comp29206_c0	GO:0000981	bZIP_1	PF00170
comp29206_c0	GO:0003700	bZIP_1	PF00170
comp29810_c0	GO:0000981	Forkhead	PF00250
comp29810_c0	GO:0003700	Forkhead	PF00250
comp30549_c0	GO:0001190	NULL	NULL
comp30549_c0	GO:0003700	NULL	NULL
comp30736_c0	GO:0003700	bZIP_1	PF00170
comp32501_c0	GO:0001077	Forkhead	PF00250
comp32501_c0	GO:0001228	Forkhead	PF00250
comp32501_c0	GO:0003700	Forkhead	PF00250
comp37576_c0	GO:0003700	Forkhead	PF00250
comp4351_c0	GO:0003700	zf-C4	PF00105
comp4791_c0	GO:0003700	MH1	PF03165
comp4805_c0	GO:0001191	NULL	NULL
comp4805_c0	GO:0003700	NULL	NULL
comp4907_c0	GO:0001078	HLH	PF00010
comp52011_c0	GO:0003700	Hormone_recep	PF00104
comp5861_c0	GO:0000981	Forkhead	PF00250
comp5861_c0	GO:0003700	Forkhead	PF00250
comp6420_c0	GO:0003700	zf-C2H2	PF00096
comp9395_c0	GO:0001078	T-box	PF00907
comp9395_c0	GO:0003700	T-box	PF00907
comp9426_c0	GO:0003700	zf-C4	PF00105

Table 3-8 Description of GO terms associated with the transcription factors.

GO	Description
GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
GO:0000982	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity
GO:0000983	RNA polymerase II core promoter sequence-specific DNA binding transcription factor activity
GO:0000989	transcription factor binding transcription factor activity
GO:0001026	TFIIIB-type transcription factor activity
GO:0001071	nucleic acid binding transcription factor activity
GO:0001076	RNA polymerase II transcription factor binding transcription factor activity
GO:0001077	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
GO:0001078	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0001190	RNA polymerase II transcription factor binding transcription factor activity involved in positive regulation of transcription
GO:0001191	RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription
GO:0001205	RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
GO:0001227	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0001228	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
GO:0003700	sequence-specific DNA binding transcription factor activity
GO:0003705	RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity
GO:0038049	ligand-activated RNA polymerase II transcription factor binding transcription factor activity
GO:0098531	direct ligand regulated sequence-specific DNA binding transcription factor activity

3.7 Differentially expressed genes

3.7.1 Method for DEGs Identification

To characterize maternal and zygotic genes, quantitative analysis of read numbers in eggs and late larvae was performed. The amounts of mRNAs were calibrated by the fragments per kilobase of exon per million fragments mapped (FPKM) method [130]. For quantification of fold-changes (FC) in reads between larvae and egg, 3 was added to the expression value in order to exclude rarely expressed genes, following a similar protocol of Adiconis et al. [107]. Genes with a FPKM that changed at least 4-fold were regarded as up-regulated or down-regulated genes.

3.7.2 Method for GO enrichment analysis of DEGs

Genes that were differentially expressed between egg and larva were again queried against UniProtKB/Swiss-Prot (E-value of 1.0E-5) using BLASTX. The corresponding symbols for each unigene were retrieved according to the best hit, and then submitted to Ontologizer [112] for functional enrichment analysis. Currently, the *Oikopleura dioica* gene association file is not available. Thus, human genes, which had the most gene entries matched, were used as the background for GO enrichment analysis. "Parent-Child-Union" method was used and p-value was adjusted via Bonferroni correction.

3.7.3 Result

We re-assembled the transcriptomes for the egg and larval stages separately. Querying the 12,311 identified unigenes against the assembled transcripts with BLASTN (E-value ≤1E-30), we found that approximately 63% were detected in egg-stage RNAs whereas 99% were detected in larval-stage RNAs. In order to further evaluate maternal and zygotic transcripts, quantitation of read numbers in eggs and larvae was carried out. The amounts of transcripts for the 12,311 unigenes were calibrated using the fragments per kilobase of exon per million fragments mapped (FPKM) method [130]. To validate the quantification, we first tested three well-characterized genes: *Brachyury, Muscle actin* and *Vasa* (Table 3-9). *Brachyury* and *Muscle actin* are typical zygotic genes and their expression becomes detectable in 32- to 64-cell embryos, respectively [18, 131], while *Vasa* is a maternally expressed gene involved in germ cell development and its mRNA is detectable in both eggs and larvae [132]. Our comparison of read numbers was in agreement with these previous *in situ* hybridization studies. The read numbers of *Brachyury* and *Muscle actin* transcripts in larvae were over 30-fold more than those in eggs, while the number for the *Vasa* transcript at the two stages was almost the same.

Table 3-9 Expression of some genes with well-known patterns.

Gene Symbol	Gene ID	Oocyte	8HPF
Brachyury	comp22934_c0	0	5.51
Muscle actin	comp26898_c0	366.42	11130.845
Ddx4,Vasa	comp26727_c1	9.495	10.36

Next, fold-changes in read numbers from eggs to larvae were calibrated by adding 3 to the expression values in order to exclude artifacts in rarely expressed genes, using a modification of the method of Adiconis *et al*. [107]. We found that 5,108 genes showed a change

of at least 4-fold (Supplementary Table S3 in our paper [22]): 3,772 genes (74%) were up-regulated from egg to larva, and 1,336 genes (26%) were down-regulated (Figure 3-10a). Therefore, more than 40% of unigenes could be categorized as up- or down-regulated genes. Because we did not prepare biological replicates in the present study, we carried out RT-PCR to ascertain expression of six known genes (*Noa36*, *Mago*, *Tcf3, Par6*, *Spectrin* and *Alpha3*). As shown in Figure 3-10b, *Noa36* and *Mago* mRNA levels in eggs were higher than those in larvae. By contrast, *Tcf3* mRNA level in larva was higher than that in eggs (Figure 3-10b). Other three genes did not show any detectable differences between the two developmental stages. These expression patterns coincided well with the RNA-Seq results. The expression values are available in Supplementary Table S3 in our paper [22].

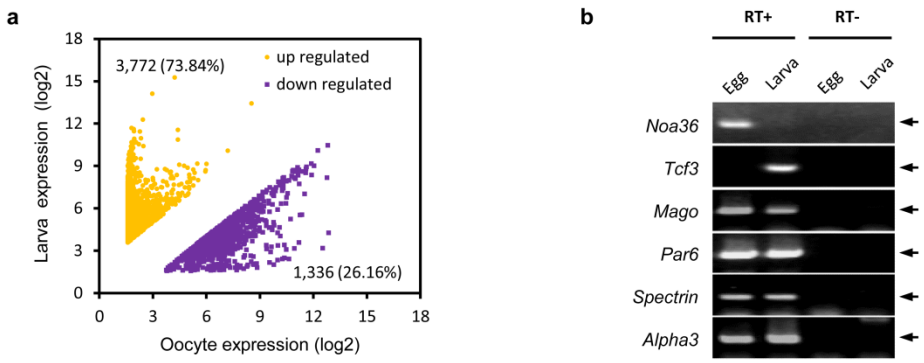


Figure 3-10 Differentially expressed genes. (a) Scale plot of up-regulated and down-regulated genes. Three was added to every expression value (FPKM) and plotted on log2 axes. (b) RT-PCR validation of the expression of some genes during egg and larval stages. All of the expression patterns were coincident with our RNA-Seq result. Arrows indicate the band with expected length of amplification.

Gene ontology (GO) [77] terms were assigned to the 3,772 up-regulated and 1,336 down-regulated genes. Enrichment analysis was performed to demonstrate conspicuous biological processes at each stage (Table 3-10). Most of the up-regulated genes were involved in localization and transport processes (such as transmembrane transport, organic anion transport, single-organism transport and localization), metabolic process (such as organonitrogen compound metabolism, sulfur compound metabolism, aromatic amino acid family metabolism, carbohydrate derivative metabolism), developmental process, organ morphogenesis, biological adhesion, synaptic transmission, cell junction assembly and cell-cell signaling. By contrast, most down-regulated genes were involved in cell cycle process, various forms of DNA or RNA processing (such as mRNA metabolism, establishment of RNA localization, DNA repair, DNA replication, DNA-templated transcription, elongation, and transcription elongation from RNA polymerase II promoter) and biogenesis (ribonucleoprotein complex biogenesis, cellular component organization or biogenesis).

Table 3-10 Gene ontology annotation of up-regulated and down-regulated genes.

Up regulated			
GO ID	GO Name	p-value	p.adjusted (Bonferroni)
GO:0051179	localization	3.21E-15	2.92E-11
GO:0044699	single-organism process	7.08E-15	6.45E-11
GO:1901564	organonitrogen compound metabolic process	2.11E-14	1.92E-10
GO:0043062	extracellular structure organization	1.74E-12	1.58E-08
GO:0065008	regulation of biological quality	1.43E-11	1.30E-07
GO:0032502	developmental process	3.51E-11	3.19E-07
GO:0034330	cell junction organization	4.58E-11	4.17E-07
GO:0006790	sulfur compound metabolic process	4.88E-11	4.45E-07
GO:0032501	multicellular organismal process	5.57E-11	5.07E-07
GO:0044259	multicellular organismal macromolecule metabolic process	5.68E-11	5.17E-07
GO:0055085	transmembrane transport	8.27E-11	7.53E-07
GO:0006357	regulation of transcription from RNA polymerase II promoter	1.40E-10	1.27E-06
GO:0044763	single-organism cellular process	3.35E-10	3.05E-06
GO:0022610	biological adhesion	3.78E-10	3.44E-06
GO:0009072	aromatic amino acid family metabolic process	5.89E-10	5.36E-06
GO:0007268	synaptic transmission	7.57E-10	6.89E-06
GO:0006082	organic acid metabolic process	9.33E-10	8.49E-06
GO:0009074	aromatic amino acid family catabolic process	1.70E-09	1.55E-05
GO:0034329	cell junction assembly	2.54E-09	2.31E-05
GO:0007267	cell-cell signaling	3.42E-09	3.12E-05
GO:0006928	movement of cell or subcellular component	3.60E-09	3.28E-05
GO:0010628	positive regulation of gene expression	9.75E-09	8.87E-05
GO:0009405	pathogenesis	1.09E-08	9.96E-05
GO:1903508	positive regulation of nucleic acid-templated transcription	1.34E-08	1.22E-04
GO:0015711	organic anion transport	2.11E-08	1.92E-04
GO:0051239	regulation of multicellular organismal process	2.36E-08	2.15E-04
GO:0051254	positive regulation of RNA metabolic process	3.17E-08	2.88E-04
GO:1902680	positive regulation of RNA biosynthetic process	3.26E-08	2.97E-04
GO:0006366	transcription from RNA polymerase II promoter	3.53E-08	3.21E-04
GO:0045893	positive regulation of transcription, DNA-templated	4.11E-08	3.74E-04
GO:0022617	extracellular matrix disassembly	4.48E-08	4.08E-04
GO:0006022	aminoglycan metabolic process	6.25E-08	5.69E-04
GO:1901135	carbohydrate derivative metabolic process	8.90E-08	8.10E-04
GO:0044712	single-organism catabolic process	1.10E-07	0.001001882
GO:0005975	carbohydrate metabolic process	1.34E-07	0.001219046
GO:0040011	locomotion	3.17E-07	0.002884982

GO:0006520	cellular amino acid metabolic process	3.75E-07	0.003415628
GO:1901566	organonitrogen compound biosynthetic process	3.78E-07	0.003445492
GO:0009887	organ morphogenesis	4.35E-07	0.003956999
GO:0044765	single-organism transport	4.93E-07	0.004485973
GO:1902578	single-organism localization	5.24E-07	0.004766755
GO:0044236	multicellular organismal metabolic process	6.73E-07	0.00613097
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	8.20E-07	0.007464611
GO:0016054	organic acid catabolic process	9.49E-07	0.008639684
GO:0044710	single-organism metabolic process	9.53E-07	0.008676926
GO:0006575	cellular modified amino acid metabolic process	9.65E-07	0.00878655
GO:1901565	organonitrogen compound catabolic process	9.73E-07	0.008862162

Down regulated

GO ID	GO Name	p-value	p.adjusted (Bonferroni)
GO:0007049	cell cycle	1.87E-47	1.25E-43
GO:0022402	cell cycle process	1.00E-41	6.67E-38
GO:0044237	cellular metabolic process	5.70E-36	3.80E-32
GO:0008152	metabolic process	3.08E-31	2.05E-27
GO:0033554	cellular response to stress	5.75E-30	3.83E-26
GO:0006396	RNA processing	1.49E-26	9.95E-23
GO:0016071	mRNA metabolic process	1.41E-25	9.38E-22
GO:0044764	multi-organism cellular process	2.22E-23	1.48E-19
GO:0022613	ribonucleoprotein complex biogenesis	5.56E-22	3.71E-18
GO:0051726	regulation of cell cycle	6.86E-21	4.57E-17
GO:0044419	interspecies interaction between organisms	9.62E-20	6.42E-16
GO:0051236	establishment of RNA localization	2.03E-18	1.36E-14
GO:0008150	biological_process	4.43E-18	2.95E-14
GO:0006998	nuclear envelope organization	1.90E-17	1.27E-13
GO:0006281	DNA repair	8.87E-17	5.92E-13
GO:0051351	positive regulation of ligase activity	1.15E-16	7.69E-13
GO:0051340	regulation of ligase activity	1.96E-16	1.31E-12
GO:0051352	negative regulation of ligase activity	1.19E-15	7.91E-12
GO:0006403	RNA localization	5.30E-15	3.53E-11
GO:0044710	single-organism metabolic process	6.66E-15	4.44E-11
GO:0006996	organelle organization	6.92E-15	4.62E-11
GO:0006260	DNA replication	9.66E-15	6.44E-11
GO:0006354	DNA-templated transcription, elongation	1.25E-14	8.34E-11
GO:0072395	signal transduction involved in cell cycle checkpoint	2.15E-14	1.43E-10
GO:0030397	membrane disassembly	4.15E-14	2.77E-10
GO:0072331	signal transduction by p53 class mediator	5.56E-14	3.71E-10
GO:0015931	nucleobase-containing compound transport	6.06E-14	4.04E-10

GO:0044260	cellular macromolecule metabolic process	6.58E-14	4.39E-10
GO:0010948	negative regulation of cell cycle process	7.57E-14	5.05E-10
GO:0006521	regulation of cellular amino acid metabolic process	9.34E-14	6.23E-10
GO:0006368	transcription elongation from RNA polymerase II promoter	9.87E-14	6.59E-10
GO:0046907	intracellular transport	1.31E-13	8.71E-10
GO:0051443	positive regulation of ubiquitin-protein transferase activity	2.13E-13	1.42E-09
GO:0051438	regulation of ubiquitin-protein transferase activity	4.17E-13	2.78E-09
GO:0071840	cellular component organization or biogenesis	1.36E-12	9.07E-09
GO:0030177	positive regulation of Wnt signaling pathway	1.68E-12	1.12E-08
GO:0006807	nitrogen compound metabolic process	1.85E-12	1.23E-08
GO:0051603	proteolysis involved in cellular protein catabolic process	2.86E-12	1.91E-08
GO:0042770	signal transduction in response to DNA damage	3.23E-12	2.16E-08
GO:0006997	nucleus organization	5.30E-12	3.53E-08
GO:0034660	ncRNA metabolic process	5.38E-12	3.59E-08
GO:0019882	antigen processing and presentation	7.60E-12	5.07E-08
GO:0033036	macromolecule localization	1.12E-11	7.45E-08
GO:0051444	negative regulation of ubiquitin-protein transferase activity	2.39E-11	1.59E-07
GO:1903322	positive regulation of protein modification by small protein conjugation or removal	2.67E-11	1.78E-07
GO:0071826	ribonucleoprotein complex subunit organization	3.06E-11	2.04E-07
GO:0006259	DNA metabolic process	3.54E-11	2.36E-07
GO:0019083	viral transcription	4.81E-11	3.21E-07
GO:1903364	positive regulation of cellular protein catabolic process	5.36E-11	3.58E-07
GO:0033238	regulation of cellular amine metabolic process	5.57E-11	3.72E-07
GO:0016055	Wnt signaling pathway	6.85E-11	4.57E-07
GO:0048524	positive regulation of viral process	2.11E-10	1.41E-06
GO:0051641	cellular localization	3.18E-10	2.12E-06
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	4.90E-10	3.27E-06
GO:0050434	positive regulation of viral transcription	6.25E-10	4.17E-06
GO:1902589	single-organism organelle organization	6.45E-10	4.30E-06
GO:0044265	cellular macromolecule catabolic process	7.77E-10	5.18E-06
GO:0051301	cell division	9.19E-10	6.13E-06
GO:0044033	multi-organism metabolic process	1.13E-09	7.56E-06
GO:0051276	chromosome organization	1.17E-09	7.81E-06
GO:1903320	regulation of protein modification by small	2.32E-09	1.54E-05

	protein conjugation or removal		
GO:0045732	positive regulation of protein catabolic process	2.58E-09	1.72E-05
GO:0006974	cellular response to DNA damage stimulus	2.81E-09	1.88E-05
GO:0043902	positive regulation of multi-organism process	4.07E-09	2.71E-05
GO:1903321	negative regulation of protein modification by small protein conjugation or removal	4.26E-09	2.84E-05
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	6.31E-09	4.21E-05
GO:0006950	response to stress	6.40E-09	4.27E-05
GO:0000278	mitotic cell cycle	7.40E-09	4.93E-05
GO:0046782	regulation of viral transcription	7.80E-09	5.20E-05
GO:0051169	nuclear transport	8.15E-09	5.43E-05
GO:1903052	positive regulation of proteolysis involved in cellular protein catabolic process	9.44E-09	6.30E-05
GO:0006397	mRNA processing	9.46E-09	6.31E-05
GO:1902582	single-organism intracellular transport	1.13E-08	7.54E-05
GO:0007077	mitotic nuclear envelope disassembly	1.17E-08	7.78E-05
GO:0090068	positive regulation of cell cycle process	1.54E-08	1.02E-04
GO:0009896	positive regulation of catabolic process	2.33E-08	1.55E-04
GO:0031331	positive regulation of cellular catabolic process	2.46E-08	1.64E-04
GO:0006139	nucleobase-containing compound metabolic process	2.64E-08	1.76E-04
GO:0010565	regulation of cellular ketone metabolic process	3.11E-08	2.07E-04
GO:0090263	positive regulation of canonical Wnt signaling pathway	4.08E-08	2.72E-04
GO:0034641	cellular nitrogen compound metabolic process	4.53E-08	3.02E-04
GO:0006353	DNA-templated transcription, termination	5.11E-08	3.41E-04
GO:0006406	mRNA export from nucleus	5.91E-08	3.94E-04
GO:2000058	regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	6.29E-08	4.19E-04
GO:1990542	mitochondrial transmembrane transport	6.86E-08	4.58E-04
GO:0009057	macromolecule catabolic process	6.87E-08	4.59E-04
GO:0050792	regulation of viral process	7.79E-08	5.19E-04
GO:0022616	DNA strand elongation	8.25E-08	5.50E-04
GO:0009411	response to UV	8.66E-08	5.77E-04
GO:0045786	negative regulation of cell cycle	9.43E-08	6.29E-04
GO:1903050	regulation of proteolysis involved in cellular protein catabolic process	9.51E-08	6.34E-04
GO:0044238	primary metabolic process	1.33E-07	8.86E-04
GO:2000060	positive regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	1.55E-07	0.001034094
GO:0007059	chromosome segregation	1.74E-07	0.001160002

GO:0045787	positive regulation of cell cycle	2.05E-07	0.001367032
GO:0010608	posttranscriptional regulation of gene expression	2.49E-07	0.001662358
GO:0006417	regulation of translation	2.75E-07	0.00183347
GO:0006457	protein folding	3.62E-07	0.002411581
GO:0031397	negative regulation of protein ubiquitination	6.17E-07	0.004118148
GO:0042180	cellular ketone metabolic process	6.18E-07	0.004121379
GO:0045862	positive regulation of proteolysis	8.35E-07	0.005570359
GO:0031398	positive regulation of protein ubiquitination	9.07E-07	0.00605187
GO:0030111	regulation of Wnt signaling pathway	9.10E-07	0.006066934
GO:0022900	electron transport chain	1.04E-06	0.00696136
GO:0019080	viral gene expression	1.11E-06	0.007390179
GO:0071158	positive regulation of cell cycle arrest	1.19E-06	0.007948624
GO:0071704	organic substance metabolic process	1.24E-06	0.008253961
GO:1903362	regulation of cellular protein catabolic process	1.25E-06	0.008367335
GO:0022618	ribonucleoprotein complex assembly	1.42E-06	0.009470095
GO:0032201	telomere maintenance via semi-conservative replication	1.46E-06	0.009764787

3.8 Intra-species sequence variations

To understand the degree of sequence conservation between the Japanese (our data) and Norwegian (OikoBase) populations, the 12,136 known unigenes were queried against the OikoBase transcript reference with BLASTN (word size of 15). BLAST hits with an alignment length of ≤300 (nucleotide level) or an E-value of ≥1E-10 were discarded. Large gaps in the BLAST results were removed in nucleotide and protein levels. Genes with multiple hits, which may include paralogs and duplicated genes, were also excluded in order to avoid inaccuracy. As a result, 4,843 and 7,739 one-to-one hits were obtained at nucleotide and protein levels, respectively. A total of 4,136 unigenes showed one-to-one hits at both the nucleotide and protein level and subjected to global sequence alignment.

The sequence identity (from the first high-scoring segment pairs) was 91.0% and 94.8% on average at the nucleotide and amino acid level, respectively. 94.6% of the 4,136 sequences showed 80% to 95% identity at the nucleotide level, and 98.8% of the sequences showed 80% to 100% identity at the amino acid level (Figure 3-11, left panel).

To ascertain effects of read coverage on the polymorphisms, we re-assessed the data using 1,024 unigenes with higher coverage (FPKM ≥ 100) and 663 unigenes with lower coverage (FPKM ≤ 10) (Figure 3-11, middle and right panels). The sequence identity of unigenes with high coverage was 92.5% and 96.7% on average at the nucleotide and amino acid level, respectively. Most unigenes in this group showed 90% to 100% identifies in both nucleotide and protein levels. In comparison, the sequence identity of unigenes with lower coverage was 89.1% and 92.2% on average at the nucleotide and amino acid level, respectively. Taken together, the data indicated a possibility of marked intra-species sequence variation between Japanese and Norwegian *O. dioica*.

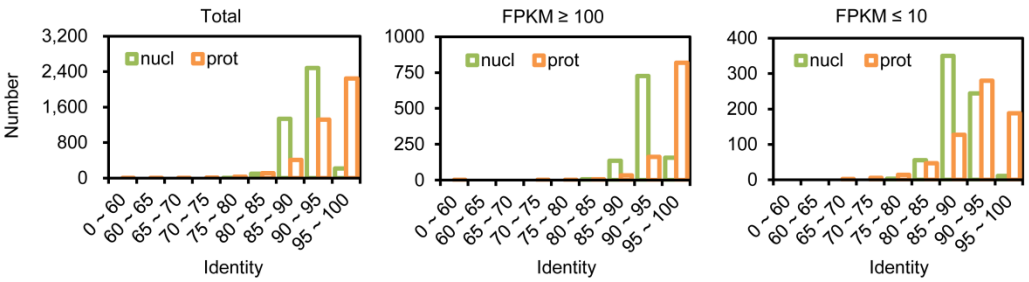


Figure 3-11 Sequence similarity between Japanese and Norwegian populations at the nucleotide (nucl) and amino acid (prot) level.

3.9 *Trans*-spliced leader and *trans*-spliced mRNAs

3.9.1 mRNA *trans*-splicing and *trans*-spliced leader

Spliced leader (SL) *trans*-splicing (Figure 3-12a) is a unique RNA splicing process in which a short spliced exon from one RNA transcripts links to the 5' end of another RNA transcript. SL *trans*-splicing occurs broadly in tunicates [92, 133-136], human [137], *Caenorhabditis elegans* [138-140], *Bombyx mori* [141], *Trypanosoma brucei* [142] and other mammalian cells [143].

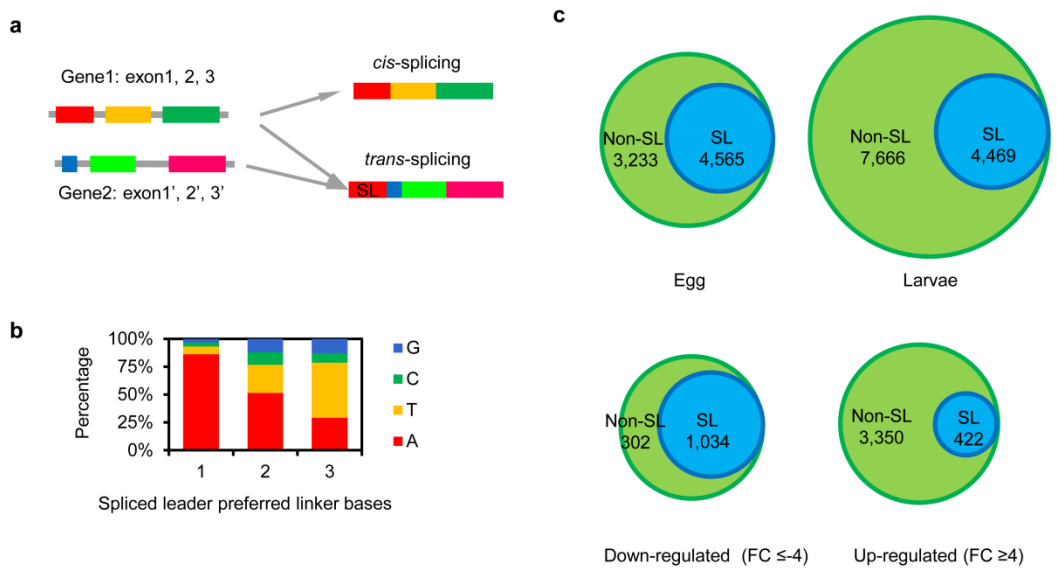


Figure 3-12 The preferential base of detected spliced leader and distribution of *trans*-spliced mRNAs in the egg, larvae. (a) *Cis*-splicing and *trans*-splicing. (b) Preference of the first three bases in the immediately linked downstream exons of SL. (b) Distribution of *trans*-spliced (SL) and non *trans*-spiced (Non-SL) mRNAs in eggs or larvae (upper panels), and down-regulated genes or up-regulated genes (lower panels).

3.9.2 Method for identification of SL and *trans*-spliced mRNAs

Unigenes were queried against themselves using megablast with a word size of 15 and an E-value threshold of 1000, using the maximum number of aligned sequences possible. Only forward vs forward matches were considered and self vs self matches were excluded. Matched positions of query and subject sequences had to be between the 1st and 150th bp of the transcript, and at least one of the aligned sequence had to start from 1 at the 5' end. The aligned sequences were extracted and clustered using CD-HIT-EST [144], and multiple sequence alignment was performed using MUSCLE [145]. SLs were identified by manually checking the conserved motif in the multiple sequence alignment. mRNAs that had the SL sequence at the 5' end were regarded as *trans*-spliced mRNAs. Additional mRNAs with partial stretches, but more than 10 bp, of the identified spliced leaders were also regarded as *trans*-spliced mRNAs.

3.9.3 Result

In this study, we identified a total of 5,020 *trans*-splicing mRNAs among the 12,311 unigenes, indicating that 40.8% of mRNA species are *trans*-spliced, at least in some of those mRNAs. These *trans*-spliced mRNAs are using only one SL, and the detected consensus sequence is

AGUCCGAUUUCGAUUGUCUAACAG. While this sequence is homologous to the previously reported SL [92], 16 bases at the 5'-end of the previously reported SL were not detected in our analysis. This may have happened because we used the Illumina TruSeq RNA Sample Preparation Kit, which could not completely capture 5' end of mRNA. We then checked the pre-trimmed reads and found the perfect match of the whole previously reported SL, but in most case, some 5' end bases were lost. To ascertain whether additional *trans*-splicing mRNAs were detectable, we queried the whole SL sequence including the previously reported sequence against the 12,311 unigenes, again using BLASTN with a smaller word size of 10, but no additional *trans*-splicing mRNA was obtained.

Next, we investigated whether the SL has any preferred bases by plotting the first three bases in the immediately linked downstream exons. As shown in Figure 3-12b, it tended to link to exons having an A-enriched header. Adenine accounted for 86% and 52% of the first and second bases, showing a specific preference for the first and second positions, while thymine accounted for the third base in 49% of cases.

We examined whether the frequency of *trans*-splicing differs between eggs and larvae. SL was found in 4,565 unigenes (58.5% of unigenes found in egg stage mRNAs) at the egg stage, and 4,469 unigenes (36.6% unigenes found in larval stage mRNAs) at the larval stage (Figure 3-12c). Thus, the number of mRNA species with this SL is almost the same, but the ratio of the *trans*-spliced mRNA species is reduced at the larval stage. Likewise, 1,034 (20.6%) of total *trans*-spliced mRNAs with the SL belong to the down-regulated genes (fold change ≤-4), while 422 (8.4%) of them belong to the up-regulated genes (fold change ≥4) (Figure 3-12c). Comparison of the unigenes with SL between the oocyte and larval stages showed that 3,858 unigenes are commonly *trans*-spliced (matched to at least 12 bp from the 3' end of the SL). Therefore, it is likely that SL is observed more frequently in maternal mRNAs in eggs when compared with zygotic mRNA in larvae.

To examine whether the *trans*-splicing prefers genes with special functions, the GO terms for *trans*-spliced unigenes and those of non-*trans*-spliced unigenes were compared (Table 3-11). The genes with SL show similar GO enrichments with the case of the down-regulated genes (Table 3-10). For instance, they included RNA processing processes and cell cycle associated processes. On the other hand, the genes without detectable SL show similar enrichments with the case of the up-regulated genes (Table 3-10), including localization and transport processes, developmental processes, and metabolic processes. The similar functional enrichments between the unigenes with SL and "down-regulated" genes suggest a possible involvement of *trans*-spliced mRNAs in early embryogenesis.

Table 3-11 Comaprion of the function of unigenes with SL and that without SL.

Unigenes with SL			
GO ID	GO Name	p-value	p.adjusted (Bonferroni)
GO:0016071	mRNA metabolic process	1.81E-75	2.04E-71
GO:0008152	metabolic process	1.70E-49	1.92E-45
GO:0046907	intracellular transport	6.92E-46	7.81E-42
GO:0071840	cellular component organization or biogenesis	2.82E-44	3.18E-40
GO:0006396	RNA processing	6.66E-44	7.52E-40
GO:0022613	ribonucleoprotein complex biogenesis	9.45E-44	1.07E-39

GO:0034660	ncRNA metabolic process	2.56E-43	2.89E-39
GO:0044764	multi-organism cellular process	4.53E-43	5.11E-39
GO:0019083	viral transcription	2.06E-42	2.33E-38
GO:0044237	cellular metabolic process	6.72E-41	7.58E-37
GO:0044419	interspecies interaction between organisms	2.33E-39	2.63E-35
GO:1902582	single-organism intracellular transport	5.44E-39	6.14E-35
GO:0006996	organelle organization	1.61E-38	1.82E-34
GO:0006401	RNA catabolic process	6.37E-37	7.19E-33
GO:0044033	multi-organism metabolic process	8.60E-33	9.71E-29
GO:0033554	cellular response to stress	1.28E-32	1.44E-28
GO:0019080	viral gene expression	1.85E-31	2.09E-27
GO:0044265	cellular macromolecule catabolic process	9.98E-30	1.13E-25
GO:0007049	cell cycle	2.05E-29	2.31E-25
GO:0006412	translation	3.03E-29	3.42E-25
GO:0033036	macromolecule localization	4.95E-28	5.58E-24
GO:1902589	single-organism organelle organization	5.00E-27	5.64E-23
GO:0008150	biological_process	1.18E-26	1.34E-22
GO:0006413	translational initiation	1.24E-26	1.40E-22
GO:0022402	cell cycle process	1.77E-26	1.99E-22
GO:0051641	cellular localization	6.83E-26	7.71E-22
GO:0006368	transcription elongation from RNA polymerase II promoter	2.91E-25	3.28E-21
GO:0006354	DNA-templated transcription, elongation	1.23E-24	1.39E-20
GO:0009057	macromolecule catabolic process	1.40E-23	1.58E-19
GO:0044802	single-organism membrane organization	2.78E-23	3.14E-19
GO:0006414	translational elongation	1.05E-22	1.18E-18
GO:0034655	nucleobase-containing compound catabolic process	1.97E-22	2.22E-18
GO:0010608	posttranscriptional regulation of gene expression	1.03E-20	1.17E-16
GO:0051603	proteolysis involved in cellular protein catabolic process	1.38E-19	1.56E-15
GO:0051726	regulation of cell cycle	2.60E-18	2.93E-14
GO:0044710	single-organism metabolic process	8.95E-18	1.01E-13
GO:0046700	heterocycle catabolic process	2.08E-17	2.35E-13
GO:0044248	cellular catabolic process	2.83E-17	3.19E-13
GO:0045184	establishment of protein localization	7.19E-17	8.11E-13
GO:0044270	cellular nitrogen compound catabolic process	7.75E-17	8.75E-13
GO:0044267	cellular protein metabolic process	1.39E-16	1.57E-12
GO:1902580	single-organism cellular localization	1.99E-16	2.25E-12
GO:0071826	ribonucleoprotein complex subunit organization	9.61E-16	1.08E-11
GO:0045047	protein targeting to ER	1.68E-15	1.90E-11
GO:0051649	establishment of localization in cell	1.72E-15	1.94E-11
GO:0072599	establishment of protein localization to endoplasmic reticulum	1.86E-15	2.10E-11

GO:0051236	establishment of RNA localization	2.23E-15	2.52E-11
GO:0019439	aromatic compound catabolic process	3.40E-15	3.84E-11
GO:1901361	organic cyclic compound catabolic process	1.31E-14	1.48E-10
GO:0070972	protein localization to endoplasmic reticulum	3.28E-14	3.70E-10
GO:0022411	cellular component disassembly	4.67E-14	5.27E-10
GO:1901575	organic substance catabolic process	1.19E-13	1.34E-09
GO:0006406	mRNA export from nucleus	2.12E-13	2.39E-09
GO:0006353	DNA-templated transcription, termination	2.31E-13	2.61E-09
GO:0044260	cellular macromolecule metabolic process	4.60E-13	5.19E-09
GO:0006417	regulation of translation	1.42E-12	1.61E-08
GO:0061024	membrane organization	1.68E-12	1.89E-08
GO:0019882	antigen processing and presentation	2.29E-12	2.59E-08
GO:0048193	Golgi vesicle transport	2.30E-12	2.60E-08
GO:0016055	Wnt signaling pathway	3.47E-12	3.91E-08
GO:0009056	catabolic process	4.41E-12	4.97E-08
GO:0006281	DNA repair	1.72E-11	1.95E-07
GO:0006403	RNA localization	5.11E-11	5.77E-07
GO:0033043	regulation of organelle organization	5.53E-11	6.24E-07
GO:0070727	cellular macromolecule localization	9.44E-11	1.07E-06
GO:0048524	positive regulation of viral process	1.18E-10	1.33E-06
GO:0071702	organic substance transport	1.23E-10	1.39E-06
GO:0010948	negative regulation of cell cycle process	1.97E-10	2.22E-06
GO:0007018	microtubule-based movement	3.32E-10	3.74E-06
GO:0051340	regulation of ligase activity	3.33E-10	3.76E-06
GO:0006612	protein targeting to membrane	3.46E-10	3.90E-06
GO:0006613	cotranslational protein targeting to membrane	3.55E-10	4.01E-06
GO:0060271	cilium morphogenesis	4.10E-10	4.63E-06
GO:0043933	macromolecular complex subunit organization	4.45E-10	5.02E-06
GO:0050434	positive regulation of viral transcription	5.12E-10	5.78E-06
GO:0051351	positive regulation of ligase activity	1.06E-09	1.20E-05
GO:0015931	nucleobase-containing compound transport	1.38E-09	1.56E-05
GO:0044085	cellular component biogenesis	2.13E-09	2.40E-05
GO:0022618	ribonucleoprotein complex assembly	2.46E-09	2.77E-05
GO:0042278	purine nucleoside metabolic process	2.68E-09	3.03E-05
GO:0046782	regulation of viral transcription	7.08E-09	7.99E-05
GO:0051352	negative regulation of ligase activity	8.35E-09	9.42E-05
GO:0007017	microtubule-based process	8.61E-09	9.72E-05
GO:0043241	protein complex disassembly	9.37E-09	1.06E-04
GO:0035966	response to topologically incorrect protein	1.52E-08	1.72E-04
GO:0046034	ATP metabolic process	1.58E-08	1.78E-04
GO:1902578	single-organism localization	2.13E-08	2.41E-04
GO:0006352	DNA-templated transcription, initiation	2.36E-08	2.66E-04
GO:0007224	smoothened signaling pathway	2.90E-08	3.28E-04
GO:0010927	cellular component assembly involved in	3.00E-08	3.38E-04

	morphogenesis		
GO:0035556	intracellular signal transduction	3.00E-08	3.39E-04
GO:0051179	localization	3.01E-08	3.39E-04
GO:0006950	response to stress	3.03E-08	3.42E-04
GO:0045786	negative regulation of cell cycle	3.06E-08	3.45E-04
GO:0051443	positive regulation of ubiquitin-protein transferase activity	3.73E-08	4.21E-04
GO:0030177	positive regulation of Wnt signaling pathway	4.52E-08	5.10E-04
GO:0006886	intracellular protein transport	5.83E-08	6.58E-04
GO:0032984	macromolecular complex disassembly	6.49E-08	7.32E-04
GO:0006367	transcription initiation from RNA polymerase II promoter	1.05E-07	0.001184002
GO:0006260	DNA replication	1.07E-07	0.001208585
GO:0009083	branched-chain amino acid catabolic process	1.19E-07	0.001341703
GO:0006369	termination of RNA polymerase II transcription	1.47E-07	0.001660042
GO:0050792	regulation of viral process	1.57E-07	0.001768246
GO:0051438	regulation of ubiquitin-protein transferase activity	1.66E-07	0.001875943
GO:0030968	endoplasmic reticulum unfolded protein response	1.69E-07	0.001912967
GO:0022900	electron transport chain	1.80E-07	0.002028517
GO:0009987	cellular process	1.91E-07	0.002157637
GO:0007059	chromosome segregation	2.18E-07	0.002461368
GO:0044765	single-organism transport	2.31E-07	0.002605499
GO:0016072	rRNA metabolic process	2.47E-07	0.002782753
GO:0051704	multi-organism process	3.30E-07	0.003722791
GO:0043902	positive regulation of multi-organism process	3.90E-07	0.004405478
GO:0090150	establishment of protein localization to membrane	4.04E-07	0.004560147
GO:0006360	transcription from RNA polymerase I promoter	4.22E-07	0.004764948
GO:0050684	regulation of mRNA processing	4.35E-07	0.004911355
GO:0044782	cilium organization	5.45E-07	0.006147628
GO:0009123	nucleoside monophosphate metabolic process	6.04E-07	0.006814094

Unigenes without SL

GO ID	GO Name	p-value	p.adjusted (Bonferroni)
GO:0051179	localization	2.61E-18	3.00E-14
GO:0065008	regulation of biological quality	9.50E-18	1.09E-13
GO:0044699	single-organism process	5.06E-17	5.83E-13
GO:0044710	single-organism metabolic process	5.04E-14	5.80E-10
GO:0055085	transmembrane transport	6.11E-12	7.03E-08
GO:0032502	developmental process	7.67E-12	8.82E-08
GO:1901564	organonitrogen compound metabolic process	8.21E-12	9.44E-08
GO:0008152	metabolic process	3.47E-11	3.99E-07
GO:0006082	organic acid metabolic process	7.51E-11	8.64E-07
GO:0010628	positive regulation of gene expression	1.25E-10	1.44E-06
GO:0007167	enzyme linked receptor protein signaling pathway	4.51E-10	5.19E-06

GO:0007267	cell-cell signaling	4.82E-10	5.54E-06
GO:1902680	positive regulation of RNA biosynthetic process	6.52E-10	7.50E-06
GO:0007268	synaptic transmission	9.70E-10	1.12E-05
GO:1903508	positive regulation of nucleic acid-templated transcription	1.53E-09	1.76E-05
GO:0051128	regulation of cellular component organization	1.96E-09	2.25E-05
GO:0006357	regulation of transcription from RNA polymerase II promoter	2.19E-09	2.52E-05
GO:1902578	single-organism localization	2.26E-09	2.60E-05
GO:0044763	single-organism cellular process	2.35E-09	2.70E-05
GO:0051239	regulation of multicellular organismal process	2.69E-09	3.10E-05
GO:0050954	sensory perception of mechanical stimulus	3.33E-09	3.83E-05
GO:0043062	extracellular structure organization	4.23E-09	4.87E-05
GO:0044712	single-organism catabolic process	4.48E-09	5.16E-05
GO:0044765	single-organism transport	6.52E-09	7.50E-05
GO:0006366	transcription from RNA polymerase II promoter	1.15E-08	1.33E-04
GO:0051254	positive regulation of RNA metabolic process	1.71E-08	1.96E-04
GO:0045893	positive regulation of transcription, DNA-templated	1.97E-08	2.27E-04
GO:0006790	sulfur compound metabolic process	6.65E-08	7.65E-04
GO:0009072	aromatic amino acid family metabolic process	9.49E-08	0.001091695
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	1.46E-07	0.001680277
GO:1901565	organonitrogen compound catabolic process	1.60E-07	0.001845851
GO:0006928	movement of cell or subcellular component	1.84E-07	0.00211391
GO:0006575	cellular modified amino acid metabolic process	2.78E-07	0.00320392
GO:0015711	organic anion transport	3.28E-07	0.003777649
GO:0009653	anatomical structure morphogenesis	3.55E-07	0.004083054
GO:0009887	organ morphogenesis	4.19E-07	0.004816197
GO:0009405	pathogenesis	4.83E-07	0.005553337
GO:0040011	locomotion	5.19E-07	0.005966554
GO:0010557	positive regulation of macromolecule biosynthetic process	5.64E-07	0.006486261
GO:0006793	phosphorus metabolic process	6.11E-07	0.007029664
GO:0048518	positive regulation of biological process	6.63E-07	0.007622639
GO:0032879	regulation of localization	8.48E-07	0.009750544

3.10 Discussion

In this study, we carried out RNA-Seq analysis of a Japanese *O. dioica* population to gain insight into the maternal and zygotic trascriptomes. *De novo* assembly of reads obtained from the egg and late larval stages generated 86,898 transcripts and recovered more than 95% of the predicted genes stored in the OikoBase. These results confirmed and expanded the results of tiled microarray analysis by Danks et al. [21], which detected 77% of the predicted genes (13,081 out of 16,749 tested genes) at several time points from egg to adult. The depth of the sequencing data showed three novel aspects: (1) 175 novel protein-encoding genes were identified, 31 of which were not found in the Norwegian genome reference. (2) Transcriptome-wide comparison revealed high levels of sequence variation (~10% of nucleotides in the intragenic region were different) between the two *O. dioica* populations. (3) *Trans*-splicing tended to happened less in larva-specific mRNAs when compared with egg mRNAs.

O. dioica is characterized by rapid development, organogenesis being completed within 10 hours post-fertilization (hpf) to form a functional body that is miniature of the adult [18]. In this study, we collected samples at only two stages: the eggs and the 8 hpf larvae. Sampling of these two developmental stages was enough to recover almost all of the predicted genes (16,423 out of 17,212 OikoBase genes) and EST clones (101,270 out of 103,969). This contrasts with the case of vertebrates, for which about 86% of genes were recovered even when samples were collected from 6-10 different stages [12].

In addition, we identified 175 novel protein-encoding genes. The number of *O. dioica* genes is estimated to be approximately 18,000 in the genome reference assembly [20, 82]. Therefore, about 1% of genes have not been annotated in the genome reference. The 175 novel protein-encoding genes that possess Pfam domains did not show any hit with predicted proteins in OikoBase and other public protein databases (UniProtKB/Swiss-Prot and NCBI RefSeq databases) by BLASTX, but their expression was confirmed by RT-PCR. Intriguingly, 31 of the novel genes were not found in the genome reference derived from Norwegian *O. dioica*. There are two possible explanations for this, the first being lower coverage of the genome reference. The genome sequence of *O. dioica* was determined by traditional shotgun sequencing, and the average genome coverage was 14X and 3% of ESTs were not mapped onto the genome reference [20, 82]. Therefore, some genes might be missing in the genome reference. The other possible reason is divergence of the genome sequences between the Japanese and Norwegian populations.

O. dioica is considered to be a rapidly evolving chordate, because of its short life cycle of 5 days and rapid changes in its genome organization [83]. However, no previous studies have tested intra-species sequence divergence in appendicularian species. In the present study, the transcript sequences showed a high degree of variability between the Japanese and Norwegian *O. dioica* populations. The average degrees of nucleotide and amino acid sequence conservation were 91.0% and 94.8%, respectively. The actual SNP rate would be higher, since we did not count gaps and eliminated the highly unmatched sequences in the comparison. It is noteworthy that the percentage differences in the sequences were comparable to those between two cryptic species of *Ciona intestinalis*, *i.e.*, Type A and Type B, in which reproductive isolation has become established [87-89]; the sequence similarities in the protein-encoding regions were 87-98% at the nucleotide level and 93.4%-100% at the amino

acid level, depending on the genes compared [87-89]. These facts suggest that whole exons in the genome, *i.e.*, the exome, have become highly diverged between geographically distant *O. dioica* populations, although it is not known whether hybrids of Japanese and Norwegian *O. dioica* would be fertile or infertile. Future analysis of the genome of Japanese *O. dioica* would be warranted to gain insight into the genomic plasticity of this rapidly evolving metazoan.

Among the 12,311 assembled transcripts, 63% and 99% were detected in eggs and larvae, respectively. Thus, the mRNAs of most of genes are present at the developing larval stage. In this quick developer, it is most probable that residual maternal mRNAs are still preserved in 8 hpf larvae. Quantitative analysis indicated that 5,108 genes were up-regulated (fold change ≥4) or down-regulated (fold change ≤-4) between the two stages, accounting for 41% of the total unigenes. GO enrichment analysis showed that the up-regulated and down-regulated genes were linked to distinct biological processes.

In *O. dioica*, 145 *trans*-spliced mRNAs have been found from 1,155 EST clones, and at least 25% of the mRNAs were estimated to be *trans*-spliced [92]. Detection of SL sequences among whole EST sequences has revealed that ~30% of the genes are *trans*-spliced [82]. In the present analysis, the SL was observed in 40.8% of mRNA species. It showed preferential linkage to adenine at the 5' ends of the downstream exons. Intriguingly, the *trans*-splicing occurs more frequently in eggs than in larvae. SL was found in 20.6% of the down-regulated genes, whereas it was found in only 8.4% of the up-regulated genes. Gene with SL showed similar GO enrichment with the case of the down-regulated genes. These results confirm and well coincide with the recent paper [146], which has shown that maternal transcripts tend to be more frequently *trans*-spliced in *O. dioica*. These results raise an interesting possibility that *trans*-splicing occurs more frequently in maternal mRNAs than zygotic ones, and it plays roles in early embryogenesis.

4 Conclusions and perspectives

4.1 Conclusion

In this study, we improved a previous method to detect transcription factors and developed a database include both transcription factors and maternal factors. Ontological representation at the cell, tissue, organ, and system levels has been specially designed to facilitate development studies. This is the original and new in REGULATOR and is not available in other TF databases. We anticipate that these resources would be useful, and facilitate studies in developmental biology.

Our RNA-Seq data have created a transcriptome resource for Japanese *O. dioica*. Our results support the *in silico* prediction of gene products in the genome of Norwegian *O. dioica* [21]. Furthermore, the depth of RNA-Seq analysis clarified various new aspects of the *O. dioica* transcriptome. First, 175 novel protein-encoding genes were found. Second, transcriptome-wide comparison revealed high levels of exon sequence variation between the Japanese and Norwegian populations. Finally, analysis of SL suggested that *trans*-splicing occurs more frequently in eggs than in larvae. The present results will provide an additional resource that is useful for understanding the developmental processes and evolutional aspects of this chordate.

4.2 Perspectives

The NGS technology, especially the RNA-Seq method, plays more and more important roles in biological science researches. With the emergence and rapid development of this technology, more and more genome sequencing are completed and the data is freely available at NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/). In this paper, we reported a database of transcription factors and maternal factors as well as the transcriptome study of the the appendicularian, *O. dioica* using RNA-Seq. Further studies could be done in the following aspects:

<u>For the REGULATOR database:</u>

1) Transcription factors of only 77 metazoan species were available in the current REGULATOR database, while genomes of much more species are available at NCBI now. Transcription factors of these species are expected to be added to the current database.
2) Information of maternal factors is limited. This will be expanded in near future.
3) The information of development-associated genes is supposed to be more completed. Since some genes involved in development processes may not be annotated, literature curated annotations could be incorporated.

<u>For the *O. dioica* genome and transcriptome studies:</u>

1) Since there are significant sequence variation between the Japanese and the Norwegian population, reconstruction of a high accurate reference genome of the Japanese *O. dioica* population will be necessary.
2) For SNP/Indel detection, more animals that contain more alleles are required. For genomic DNA sequencing, additional 30 wild adult animals could be collected, in which the allelic variation will be 60. Then, rapidly evolving genes under positive selection could be identified via the Ka/Ks method.

3) Recently, *trans*-splicing was reported to be involved in the human embryonic stem cell pluripotency [137]. We also found the *trans*-spliced mRNAs are enriched in egg mRNAs, although the role of *trans*-splicing in early embryonic development in *O. dioica* has not been elucidated. Further experiments could be done in this issue. For example, artificial mRNAs with a synthetic SL added to non *trans*-spliced mRNAs can be injected into the development embryos, and see what happen.
4) Some maternal mRNAs that controlling the posterior cell fates have been found to be localized to the posterior region of the vegetal hemisphere of ascidian embryos (postplasmic mRNAs). We are going to identify similar localized maternal mRNAs using *O. dioica* 8-cell embryos separated into animal and vegetal halves by RNA-Seq (spatial differences).
5) The transcription factors induced at the maternal to zygotic transition may paly crucial roles in early embryonic development. Developmentally staged RNA-Seq (strand-specific protocol, 200 bp paired-end library) can be carried out in future, using unfertilized eggs, 2-cell stage, 4-cell stage, 8-cell stage, 16-cell stage, 32-cell stage (gastrula), 64-cell stage (neurula), tailbud stage, hatched larval stage, juvenile stage and adult stage. Gene expression patterns of transcription factors will be acquired via k-means clustering of transcription factors. Whole mount *in situ* hybridization and RT-PCR will be used to confirm their expression pattern. Gene ontology annotation will be performed to predict the development-associated genes. Then transcription regulatory network will be obtained via correlation analysis of the staged gene expression data. MO/siRNA knock-down and/or mRNA injection experiments of candidate key regulators will be carried out to reveal their roles by examining the phenotypes first in morphology during early developmental process.

ACKNOWLEDGEMENTS

First of all, I would like to thank Prof. Hiroki NISHIDA (西田 宏記) and Dr. Takeshi A. ONUMA (小沼 健) for supervising my PhD works, and members of Developmental Biology Lab for helping my life in Osaka.

I thank Bioinformatics Organization (http://www.bioinformatics.org/) for providing The HTML, PHP web space and MySQL database and the Genome Information Research Center in Osaka University for providing computational resources.

I appreciate Professor Hiroki NISHIDA (西田 宏記), Teruo YASUNAGA (安永 照雄) and Zhi-Hui SU (蘇 智慧) for checking this thesis and giving their comments and suggestions.

I appreciate my supervisor Hiroki NISHIDA (西田 宏記) and Takeshi ONUMA (小沼 健) for revising the manuscript.

I appreciate Takeshi ONUMA (小沼 健), Tatsuya OMOTEZAKO (表迫 竜也) and Kanae KISHI (岸 香苗) for carrying out sample collection and some experiments.

I appreciate Misae SUZUKI (鈴木 幹恵), Momoko HAYASHI (林 桃子) and Miho ISOBE (磯部 美穂) in our laboratory for their help in the culture of *O. dioica*.

I also appreciate all the members in NISHIDA's lab for their help and companion these years. I am not living in solitude because of you:

西田 宏記	Hiroki NISHIDA
井汲 麻紀	Maki IKUMI
市川 麻世	Asayo ICHIKAWA
熊野 岳	Gaku KUMANO
今井(佐藤) 薫	Kaoru IMAI (SATOU)
小沼 健	Takeshi ONUMA
山田 温子	Atsuko YAMADA
鈴木 幹恵	Misae SUZUKI
林 晃平	Kohei HAYASHI
桑島 真美	Mami KUWAJIMA
表迫 竜也	Tatsuya OMOTEZAKO
宮奥 香理	Kaori MIYAOKU
岸 香苗	Kanae KISHI
宮田 善将	Yoshimasa MIYATA
Samantha Connop	
宮竹 将	Sho MIYATAKE
林 桃子	Momoko HAYASHI
細野 青葉	Aoba HOSONO
武藤 美雪	Miyuki MUTO
徳久 万純	Masumi TOKUHISA
磯部 美穂	Miho ISOBE
佐々木 隆宣	Takanobu SASAKI
天野 留奈	Runa AMANO
小林 あゆみ	Ayumi KOBAYASHI

桝井 美里	Misato MASUI
山田 詩織	Shiori YAMADA
松尾 正樹	Masaki MATSUO
西郷 元彦	Motohiko SAIGO
田中 佑佳	Yuka TANAKA
戸村 亮	Ryo TOMURA
肥川 広樹	Hiroki HIKAWA

FUNDING

This work was supported by Grants-in-Aid for Scientific Research from the JSPS to H.N. (22370078, 26650079) and T.A.O. (22870019, 26840079). K.W is supported by a MEXT Scholarship (125058), a Mitsubishi Corporation International Scholarship (MITSU1451), and an Osaka University Scholarship and Research Assistant Fellowship.

REFERENCES

1. Baroux, C., Autran, D., Gillmor, C.S., Grimanelli, D. & Grossniklaus, U. The maternal to zygotic transition in animals and plants. *Cold Spring Harb Symp Quant Biol* **73**, 89-100 (2008).
2. Tadros, W. & Lipshitz, H.D. The maternal-to-zygotic transition: a play in two acts. *Development* **136**, 3033-3042 (2009).
3. Schier, A.F. The maternal-zygotic transition: death and birth of RNAs. *Science* **316**, 406-407 (2007).
4. Langley, A.R., Smith, J.C., Stemple, D.L. & Harvey, S.A. New insights into the maternal to zygotic transition. *Development* **141**, 3834-3841 (2014).
5. Lee, M.T., Bonneau, A.R., Takacs, C.M., Bazzini, A.A., DiVito, K.R., Fleming, E.S. & Giraldez, A.J. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature* **503**, 360-364 (2013).
6. Benoit, B., He, C.H., Zhang, F., Votruba, S.M., Tadros, W., Westwood, J.T., Smibert, C.A., Lipshitz, H.D. & Theurkauf, W.E. An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition. *Development* **136**, 923-932 (2009).
7. Li, L., Lu, X. & Dean, J. The maternal to zygotic transition in mammals. *Mol Aspects Med* **34**, 919-938 (2013).
8. Pelegri, F. Maternal factors in zebrafish development. *Dev Dyn* **228**, 535-554 (2003).
9. Heasman, J. Maternal determinants of embryonic cell fate. *Semin Cell Dev Biol* **17**, 93-98 (2006).
10. Sardet, C., Dru, P. & Prodon, F. Maternal determinants and mRNAs in the cortex of ascidian oocytes, zygotes and embryos. *Biol Cell* **97**, 35-49 (2005).
11. Makabe, K.W., Kawashima, T., Kawashima, S., Minokawa, T., Adachi, A., Kawamura, H., Ishikawa, H., Yasuda, R., Yamamoto, H., Kondoh, K., Arioka, S., Sasakura, Y., Kobayashi, A., Yagi, K., Shojima, K., Kondoh, Y., Kido, S., Tsujinami, M., Nishimura, N., Takahashi, M., Nakamura, T., Kanehisa, M., Ogasawara, M., Nishikata, T. & Nishida, H. Large-scale cDNA analysis of the maternal genetic information in the egg of Halocynthia roretzi for a gene expression catalog of ascidian development. *Development* **128**, 2555-2567 (2001).
12. Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G., Lim, A.Y., Hajan, H.S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Alestrom, P. & Mathavan, S. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res* **21**, 1328-1338 (2011).
13. Nishida, H. Specification of embryonic axis and mosaic development in ascidians. *Dev Dyn* **233**, 1177-1193 (2005).
14. Prodon, F., Yamada, L., Shirae-Kurabayashi, M., Nakamura, Y. & Sasakura, Y. Postplasmic/PEM RNAs: a class of localized maternal mRNAs with multiple roles in cell polarity and development in ascidian embryos. *Dev Dyn* **236**, 1698-1715 (2007).
15. Nakamura, Y., Makabe, K.W. & Nishida, H. Localization and expression pattern of type I postplasmic mRNAs in embryos of the ascidian Halocynthia roretzi. *Gene Expr Patterns* **3**, 71-75 (2003).
16. Sasakura, Y. & Makabe, K.W. Identification of cis elements which direct the

localization of maternal mRNAs to the posterior pole of ascidian embryos. *Dev Biol* **250**, 128-144 (2002).

17. Wang, K. & Nishida, H. REGULATOR: a database of metazoan transcription factors and maternal factors for developmental studies. *BMC Bioinformatics* **16**, 114 (2015).
18. Nishida, H. Development of the appendicularian Oikopleura dioica: culture, genome, and cell lineages. *Dev Growth Differ* **50 Suppl 1**, S239-256 (2008).
19. Nishida, H. & Stach, T. Cell lineages and fate maps in tunicates: conservation and modification. *Zoolog Sci* **31**, 645-652 (2014).
20. Seo, H.C., Kube, M., Edvardsen, R.B., Jensen, M.F., Beck, A., Spriet, E., Gorsky, G., Thompson, E.M., Lehrach, H., Reinhardt, R. & Chourrout, D. Miniature genome in the marine chordate Oikopleura dioica. *Science* **294**, 2506 (2001).
21. Danks, G., Campsteijn, C., Parida, M., Butcher, S., Doddapaneni, H., Fu, B., Petrin, R., Metpally, R., Lenhard, B., Wincker, P., Chourrout, D., Thompson, E.M. & Manak, J.R. OikoBase: a genomics and developmental transcriptomics resource for the urochordate Oikopleura dioica. *Nucleic Acids Res* **41**, D845-853 (2013).
22. Wang, K., Omotezako, T., Kishi, K., Nishida, H. & Onuma, T.A. Maternal and zygotic transcriptomes in the appendicularian, Oikopleura dioica: novel protein-encoding genes, intra-species sequence variations, and trans-spliced RNA leader. *Dev Genes Evol* **225**, 149-159 (2015).
23. Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* **21**, 327-335 (1996).
24. Nikolov, D.B. & Burley, S.K. RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A* **94**, 15-22 (1997).
25. Nishida, H. Cell fate specification by localized cytoplasmic determinants and cell interactions in ascidian embryos. *Int Rev Cytol* **176**, 245-306 (1997).
26. Wang, K., Wang, H., Wang, J., Xie, Y., Chen, J., Yan, H., Liu, Z. & Wen, T. System approaches reveal the molecular networks involved in neural stem cell differentiation. *Protein Cell* **3**, 213-224 (2012).
27. Roy, A., de Melo, J., Chaturvedi, D., Thein, T., Cabrera-Socorro, A., Houart, C., Meyer, G., Blackshaw, S. & Tole, S. LHX2 is necessary for the maintenance of optic identity and for the progression of optic morphogenesis. *J Neurosci* **33**, 6877-6884 (2013).
28. Xie, Q. & Cvekl, A. The orchestration of mammalian tissue morphogenesis through a series of coherent feed-forward loops. *J Biol Chem* **286**, 43259-43271 (2011).
29. Song, E., Ma, X., Li, H., Zhang, P., Ni, D., Chen, W., Gao, Y., Fan, Y., Pang, H., Shi, T., Ding, Q., Wang, B., Zhang, Y. & Zhang, X. Attenuation of kruppel-like factor 4 facilitates carcinogenesis by inducing g1/s phase arrest in clear cell renal cell carcinoma. *PLoS One* **8**, e67758 (2013).
30. Evan, G., Harrington, E., Fanidi, A., Land, H., Amati, B. & Bennett, M. Integrated control of cell proliferation and cell death by the c-myc oncogene. *Philos Trans R Soc Lond B Biol Sci* **345**, 269-275 (1994).
31. Boch, J. & Bonas, U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol* **48**, 419-436 (2010).
32. Nishida, H. The maternal muscle determinant in the ascidian egg. *Wiley Interdiscip Rev Dev Biol* **1**, 425-433 (2012).

33. Langdon, Y.G. & Mullins, M.C. Maternal and zygotic control of zebrafish dorsoventral axial patterning. *Annu Rev Genet* **45**, 357-377 (2011).
34. Sibon, O.C., Stevenson, V.A. & Theurkauf, W.E. DNA-replication checkpoint control at the Drosophila midblastula transition. *Nature* **388**, 93-97 (1997).
35. Zhang, H.M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H. & Guo, A.Y. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40**, D144-149 (2012).
36. Wilson, D., Charoensawan, V., Kummerfeld, S.K. & Teichmann, S.A. DBD--taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* **36**, D88-92 (2008).
37. Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., Wasserman, W.W., Roach, J.C. & Sladek, R. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* **10**, R29 (2009).
38. Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L. & Luo, J. DATF: a database of Arabidopsis transcription factors. *Bioinformatics* **21**, 2568-2569 (2005).
39. Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W., Gu, X., Wei, L. & Luo, J. DRTF: a database of rice transcription factors. *Bioinformatics* **22**, 1286-1287 (2006).
40. He, K., Guo, A.Y., Gao, G., Zhu, Q.H., Liu, X.C., Zhang, H., Chen, X., Gu, X. & Luo, J. Computational identification of plant transcription factors and the construction of the PlantTFDB database. *Methods Mol Biol* **674**, 351-368 (2010).
41. Zhang, H., Jin, J., Tang, L., Zhao, Y., Gu, X., Gao, G. & Luo, J. PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res* **39**, D1114-1117 (2011).
42. Perez-Rueda, E. & Janga, S.C. Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol Biol Evol* **27**, 1449-1459 (2010).
43. Xu, X., Yu, D., Fang, W., Cheng, Y., Qian, Z., Lu, W., Cai, Y. & Feng, K. Prediction of peptidase category based on functional domain composition. *J Proteome Res* **7**, 4521-4524 (2008).
44. Wang, K., Hu, L.L., Shi, X.H., Dong, Y.S., Li, H.P. & Wen, T.Q. PSCL: predicting protein subcellular localization based on optimal functional domains. *Protein Pept Lett* **19**, 15-22 (2012).
45. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A. & Finn, R.D. The Pfam protein families database. *Nucleic Acids Res* **40**, D290-301 (2012).
46. Balakrishnan, R., Harris, M.A., Huntley, R., Van Auken, K. & Cherry, J.M. A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford)* **2013**, bat054 (2013).
47. Li, S., Liu, B., Cai, Y. & Li, Y. Predicting protein N-glycosylation by combining functional domain and secretion information. *J Biomol Struct Dyn* **25**, 49-54 (2007).
48. Li, S., Liu, B., Zeng, R., Cai, Y. & Li, Y. Predicting O-glycosylation sites in mammalian proteins by using SVMs. *Comput Biol Chem* **30**, 203-208 (2006).
49. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering

and comparing biological sequences. *Bioinformatics* **26**, 680-682 (2010).

50. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev* **10**, 2657-2683 (1996).
51. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).
52. Clark, W.T. & Radivojac, P. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* **29**, i53-61 (2013).
53. Shamir, L., Delaney, J.D., Orlov, N., Eckley, D.M. & Goldberg, I.G. Pattern recognition software and techniques for biological image analysis. *PLoS Comput Biol* **6**, e1000974 (2010).
54. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol* **3**, 185-205 (2005).
55. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479-2481 (2004).
56. Hu, L., Huang, T., Shi, X., Lu, W.C., Cai, Y.D. & Chou, K.C. Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. *PLoS One* **6**, e14556 (2011).
57. Huang, T., He, Z.S., Cui, W.R., Cai, Y.D., Shi, X.H., Hu, L.L. & Chou, K.C. A sequence-based approach for predicting protein disordered regions. *Protein Pept Lett* **20**, 243-248 (2013).
58. Huang, T., Niu, S., Xu, Z., Huang, Y., Kong, X., Cai, Y.D. & Chou, K.C. Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. *PLoS One* **6**, e22940 (2011).
59. ElGokhy, S.M., ElHefnawi, M. & Shoukry, A. Ensemble-based classification approach for micro-RNA mining applied on diverse metagenomic sequences. *BMC Res Notes* **7**, 286 (2014).
60. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. & Soboleva, A. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995 (2013).
61. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. & Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).
62. Preuss, K.M., Lopez, J.A., Colbourne, J.K. & Wade, M.J. Identification of maternally-loaded RNA transcripts in unfertilized eggs of Tribolium castaneum. *BMC Genomics* **13**, 671 (2012).
63. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C. & Jensen, L.J. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808-815 (2013).
64. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., Castagnoli, L. & Cesareni, G. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* **40**, D857-861 (2012).

65. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R.C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S. & Hermjakob, H. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**, D841-846 (2012).
66. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. & Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451 (2004).
67. Jiang, C., Xuan, Z., Zhao, F. & Zhang, M.Q. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* **35**, D137-140 (2007).
68. Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R. & Ma'ayan, A. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database (Oxford)* **2013**, bat045 (2013).
69. Kim, S.S., Chen, Y.M., O'Leary, E., Witzgall, R., Vidal, M. & Bonventre, J.V. A novel member of the RING finger family, KRIP-1, associates with the KRAB-A transcriptional repressor domain of zinc finger proteins. *Proc Natl Acad Sci U S A* **93**, 15299-15304 (1996).
70. Urrutia, R. KRAB-containing zinc-finger repressor proteins. *Genome Biol* **4**, 231 (2003).
71. Kummerfeld, S.K. & Teichmann, S.A. DBD: a transcription factor prediction database. *Nucleic Acids Res* **34**, D74-81 (2006).
72. Markljung, E., Jiang, L., Jaffe, J.D., Mikkelsen, T.S., Wallerman, O., Larhammar, M., Zhang, X., Wang, L., Saenz-Vash, V., Gnirke, A., Lindroth, A.M., Barres, R., Yan, J., Stromberg, S., De, S., Ponten, F., Lander, E.S., Carr, S.A., Zierath, J.R., Kullander, K., Wadelius, C., Lindblad-Toh, K., Andersson, G., Hjalm, G. & Andersson, L. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. *PLoS Biol* **7**, e1000256 (2009).
73. Huang, Y.Z., Zhang, L.Z., Lai, X.S., Li, M.X., Sun, Y.J., Li, C.J., Lan, X.Y., Lei, C.Z., Zhang, C.L., Zhao, X. & Chen, H. Transcription factor ZBED6 mediates IGF2 gene expression by regulating promoter activity and DNA methylation in myoblasts. *Sci Rep* **4**, 4570 (2014).
74. Kabe, Y., Goto, M., Shima, D., Imai, T., Wada, T., Morohashi, K., Shirakawa, M., Hirose, S. & Handa, H. The role of human MBF1 as a transcriptional coactivator. *J Biol Chem* **274**, 34196-34202 (1999).
75. Zou, J.X., Revenko, A.S., Li, L.B., Gemo, A.T. & Chen, H.W. ANCCA, an estrogen-regulated AAA+ ATPase coactivator for ERalpha, is required for coregulator occupancy and chromatin modification. *Proc Natl Acad Sci U S A* **104**, 18067-18072 (2007).
76. Guelman, S., Kozuka, K., Mao, Y., Pham, V., Solloway, M.J., Wang, J., Wu, J., Lill, J.R. & Zha, J. The double-histone-acetyltransferase complex ATAC is essential for mammalian development. *Mol Cell Biol* **29**, 1176-1188 (2009).
77. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K.,

Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. & White, R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-261 (2004).

78. Fujii, S., Nishio, T. & Nishida, H. Cleavage pattern, gastrulation, and neurulation in the appendicularian, Oikopleura dioica. *Dev Genes Evol* **218**, 69-79 (2008).
79. Stach, T., Winter, J., Bouquet, J.M., Chourrout, D. & Schnabel, R. Embryology of a planktonic tunicate reveals traces of sessility. *Proc Natl Acad Sci U S A* **105**, 7229-7234 (2008).
80. Kishi, K., Onuma, T.A. & Nishida, H. Long-distance cell migration during larval development in the appendicularian, Oikopleura dioica. *Dev Biol* **395**, 299-306 (2014).
81. Omotezako, T., Nishino, A., Onuma, T.A. & Nishida, H. RNA interference in the appendicularian Oikopleura dioica reveals the function of the Brachyury gene. *Dev Genes Evol* **223**, 261-267 (2013).
82. Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.M., Da Silva, C., Brinkmann, H., Mikhaleva, J., Olsen, L.C., Jubin, C., Canestro, C., Bouquet, J.M., Danks, G., Poulain, J., Campsteijn, C., Adamski, M., Cross, I., Yadetie, F., Muffato, M., Louis, A., Butcher, S., Tsagkogeorga, G., Konrad, A., Singh, S., Jensen, M.F., Huynh Cong, E., Eikeseth-Otteraa, H., Noel, B., Anthouard, V., Porcel, B.M., Kachouri-Lafond, R., Nishino, A., Ugolini, M., Chourrout, P., Nishida, H., Aasland, R., Huzurbazar, S., Westhof, E., Delsuc, F., Lehrach, H., Reinhardt, R., Weissenbach, J., Roy, S.W., Artiguenave, F., Postlethwait, J.H., Manak, J.R., Thompson, E.M., Jaillon, O., Du Pasquier, L., Boudinot, P., Liberles, D.A., Volff, J.N., Philippe, H., Lenhard, B., Roest Crollius, H., Wincker, P. & Chourrout, D. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381-1385 (2010).
83. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965-968 (2006).
84. Holland, L.Z. Developmental biology: a chordate with a difference. *Nature* **447**, 153-155 (2007).
85. Samarghandian, S. & Shibuya, M. Vascular Endothelial Growth Factor Receptor Family in Ascidians, Halocynthia roretzi (Sea Squirt). Its High Expression in Circulatory System-Containing Tissues. *Int J Mol Sci* **14**, 4841-4853 (2013).
86. Satoh, N. The ascidian tadpole larva: comparative molecular development and genomics. *Nat Rev Genet* **4**, 285-295 (2003).
87. Suzuki, M.M., Nishikawa, T. & Bird, A. Genomic approaches reveal unexpected genetic divergence within Ciona intestinalis. *J Mol Evol* **61**, 627-635 (2005).
88. Caputi, L., Andreakis, N., Mastrototaro, F., Cirino, P., Vassillo, M. & Sordino, P. Cryptic

speciation in a model invertebrate chordate. *Proc Natl Acad Sci U S A* **104**, 9364-9369 (2007).

89. Nydam, M.L. & Harrison, R.G. Polymorphism and divergence within the ascidian genus Ciona. *Mol Phylogenet Evol* **56**, 718-726 (2010).

90. Griggio, F., Voskoboynik, A., Iannelli, F., Justy, F., Tilak, M.K., Turon, X., Pesole, G., Douzery, E.J., Mastrototaro, F. & Gissi, C. Ascidian mitogenomics: comparison of evolutionary rates in closely related taxa provides evidence of ongoing speciation events. *Genome Biol Evol* **6**, 591-605 (2014).

91. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. & Regev, A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).

92. Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D. & Thompson, E.M. Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome. *Mol Cell Biol* **24**, 7795-7805 (2004).

93. Aksyonov, S.A., Bittner, M., Bloom, L.B., Reha-Krantz, L.J., Gould, I.R., Hayes, M.A., Kiernan, U.A., Niederkofler, E.E., Pizziconi, V., Rivera, R.S., Williams, D.J. & Williams, P. Multiplexed DNA sequencing-by-synthesis. *Anal Biochem* **348**, 127-138 (2006).

94. Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C. & Vezenov, D.V. The challenges of sequencing by synthesis. *Nat Biotechnol* **27**, 1013-1023 (2009).

95. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E.C.M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne,

M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. & Smith, A.J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

96. Leng, N., Li, Y., McIntosh, B.E., Nguyen, B.K., Duffin, B., Tian, S., Thomson, J.A., Dewey, C.N., Stewart, R. & Kendziorski, C. EBSeq-HMM: a Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments. *Bioinformatics* (2015).

97. Zheng, X. & Moriyama, E.N. Comparative studies of differential gene calling using RNA-Seq data. *BMC Bioinformatics* **14 Suppl 13**, S7 (2013).

98. Fang, Z., Martin, J. & Wang, Z. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell Biosci* **2**, 26 (2012).

99. Cui, X. & Churchill, G.A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* **4**, 210 (2003).

100. Baggerly, K.A., Coombes, K.R., Hess, K.R., Stivers, D.N., Abruzzo, L.V. & Zhang, W. Identifying differentially expressed genes in cDNA microarray experiments. *J Comput Biol* **8**, 639-659 (2001).

101. Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. & Altman, R.B. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **18**, 1454-1461 (2002).

102. Kvam, V.M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* **99**, 248-256 (2012).

103. Wang, L., Feng, Z., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136-138 (2010).

104. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).

105. Lee, S., Seo, C.H., Lim, B., Yang, J.O., Oh, J., Kim, M., Lee, B. & Kang, C. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* **39**, e9 (2011).

106. Rowley, J.W., Oler, A.J., Tolley, N.D., Hunter, B.N., Low, E.N., Nix, D.A., Yost, C.C., Zimmerman, G.A. & Weyrich, A.S. Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. *Blood* **118**, e101-111 (2011).

107. Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A. & Levin, J.Z. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* **10**, 623-629 (2013).

108. Faustino, R.S., Chiriac, A., Niederlander, N.J., Nelson, T.J., Behfar, A., Mishra, P.K.,

Macura, S., Michalak, M., Terzic, A. & Perez-Terzic, C. Decoded calreticulin-deficient embryonic stem cell transcriptome resolves latent cardiophenotype. *Stem Cells* **28**, 1281-1291 (2010).

109. Conesa, A. & Gotz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**, 619832 (2008).
110. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
111. Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. & Lempicki, R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).
112. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**, 1650-1651 (2008).
113. Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C. & Weinstein, J.N. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **4**, R28 (2003).
114. Bluthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H. & Beule, D. Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform* **16**, 106-115 (2005).
115. Huang da, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
116. Wang, K., Hu, F., Xu, K., Cheng, H., Jiang, M., Feng, R., Li, J. & Wen, T. CASCADE_SCAN: mining signal transduction network from high-throughput data based on steepest descent method. *BMC Bioinformatics* **12**, 164 (2011).
117. Bouquet, J.M., Spriet, E., Troedsson, C., Ottera, H., Chourrout, D. & Thompson, E.M. Culture optimization for the emergent zooplanktonic model organism Oikopleura dioica. *J Plankton Res* **31**, 359-370 (2009).
118. Borodina, T., Adjaye, J. & Sultan, M. A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol* **500**, 79-98 (2011).
119. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N. & Regev, A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-1512 (2013).
120. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.W., Li, Y., Xu, X., Wong, G.K. & Wang, J. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660-1666 (2014).
121. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
122. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of

protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

123. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol Biol* **406**, 89-112 (2007).
124. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501-504 (2005).
125. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
126. Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P. & Bateman, A. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**, D226-232 (2013).
127. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).
128. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
129. Wu, B., Li, Y., Yan, H., Ma, Y., Luo, H., Yuan, L., Chen, S. & Lu, S. Comprehensive transcriptome analysis reveals novel genes involved in cardiac glycoside biosynthesis and mlncRNAs associated with secondary metabolism and stress response in Digitalis purpurea. *BMC Genomics* **13**, 15 (2012).
130. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
131. Bassham, S. & Postlethwait, J. Brachyury (T) expression in embryos of a larvacean urochordate, Oikopleura dioica, and the ancestral role of T. *Dev Biol* **220**, 322-332 (2000).
132. Shirae-Kurabayashi, M., Nishikata, T., Takamura, K., Tanaka, K.J., Nakamoto, C. & Nakamura, A. Dynamic redistribution of vasa homolog and exclusion of somatic cell determinants during germ cell specification in Ciona intestinalis. *Development* **133**, 2683-2693 (2006).
133. Vandenberghe, A.E., Meedel, T.H. & Hastings, K.E. mRNA 5'-leader trans-splicing in the chordates. *Genes Dev* **15**, 294-303 (2001).
134. Satou, Y., Hamaguchi, M., Takeuchi, K., Hastings, K.E. & Satoh, N. Genomic overview of mRNA 5'-leader trans-splicing in the ascidian Ciona intestinalis. *Nucleic Acids Res* **34**, 3378-3388 (2006).
135. Matsumoto, J., Dewar, K., Wasserscheid, J., Wiley, G.B., Macmil, S.L., Roe, B.A., Zeller, R.W., Satou, Y. & Hastings, K.E. High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. *Genome Res* **20**, 636-645 (2010).
136. Gasparini, F. & Shimeld, S.M. Analysis of a botryllid enriched-full-length cDNA library: insight into the evolution of spliced leader trans-splicing in tunicates. *Dev Genes Evol* **220**, 329-336 (2011).
137. Wu, C.S., Yu, C.Y., Chuang, C.Y., Hsiao, M., Kao, C.F., Kuo, H.C. & Chuang, T.J. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res* **24**, 25-36 (2014).
138. Graber, J.H., Salisbury, J., Hutchins, L.N. & Blumenthal, T. C. elegans sequences that

control trans-splicing and operon pre-mRNA processing. *RNA* **13**, 1409-1426 (2007).

139. Allen, M.A., Hillier, L.W., Waterston, R.H. & Blumenthal, T. A global analysis of C. elegans trans-splicing. *Genome Res* **21**, 255-264 (2011).
140. Blumenthal, T. Trans-splicing and operons in C. elegans. *WormBook*, 1-11 (2012).
141. Shao, W., Zhao, Q.Y., Wang, X.Y., Xu, X.Y., Tang, Q., Li, M., Li, X. & Xu, Y.Z. Alternative splicing and trans-splicing events revealed by analysis of the Bombyx mori transcriptome. *RNA* **18**, 1395-1407 (2012).
142. Shaked, H., Wachtel, C., Tulinski, P., Yahia, N.H., Barda, O., Darzynkiewicz, E., Nilsen, T.W. & Michaeli, S. Establishment of an in vitro trans-splicing system in Trypanosoma brucei that requires endogenous spliced leader RNA. *Nucleic Acids Res* **38**, e114 (2010).
143. Viles, K.D. & Sullenger, B.A. Proximity-dependent and proximity-independent trans-splicing in mammalian cells. *RNA* **14**, 1081-1094 (2008).
144. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
145. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
146. Danks, G.B., Raasholm, M., Campsteijn, C., Long, A.M., Manak, J.R., Lenhard, B. & Thompson, E.M. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol Biol Evol* **32**, 585-599 (2015).

Publication list related to the Doctor thesis

1. Wang, K. & Nishida, H. REGULATOR: a database of metazoan transcription factors and maternal factors for developmental studies. *BMC Bioinformatics* **16**, 114 (2015).
2. Wang, K., Omotezako, T., Kishi, K., Nishida, H. & Onuma, T.A. Maternal and zygotic transcriptomes in the appendicularian, Oikopleura dioica: novel protein-encoding genes, intra-species sequence variations, and trans-spliced RNA leader. *Dev Genes Evol* **225**, 149-159 (2015).