University of South Bohemia
Faculty of Science
České Budějovice, Czech Republic
and
Johannes Kepler University
Faculty of Engineering and Natural Sciences
Linz, Austria

IDENTIFICATION AND CHARACTERISATION OF NOVEL TRANSCRIPTS IN MOUSE OOCYTES AND EMBRYOS

Bachelor's thesis

Karolína Kravériková

Supervisor: Mgr. Lenka Gahurová, Ph.D.

Laboratory of early mammalian developmental biology, Department of molecular biology and genetics,
Faculty of Science,
České Budejovice
2020

Kravárikovᡠ, K., 2020: Identification and characterisation of novel transcripts in mouse oocytes and embryos. Bc. Thesis, in English.-75., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

Annotation

The primary goal of this bachelor thesis was to identify novel transcripts in mouse oocytes and preimplantation embryos, analyse their expression profile, protein coding potential, the activity of transposable elements as promoters of such transcripts, in order to select candidate transcripts for functional analysis of their roles in mammalian development.

Affirmation

I hereby declare that I have worked on my bachelor´s thesis independently and used only the sources listed in the bibliography. I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full form to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.
Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

In České Budejovice, 20.04 2020,

……
Karolina Kravárikovᡠ

Contents

1 Abstract

Due to the cell-type specificity of many transcription events, deep RNA sequencing (RNA-seq) of particular cell type usually leads to identification of a number of novel non-annotated genes. This is particularly true for low input samples of very specific cells, such as mammalian oocytes and early embryos. Many recent publications show that in these datasets there is a high number of non-annotated genes and transcripts, predominantly long non-coding RNA (lncRNAs), which might perform important functions. These novel transcripts often originate in transposable elements (TEs) with different activity between oocytes and embryos. We processed RNA-seq datasets from mouse early developmental stages, performed de novo transcriptome assembly to identify novel transcripts, analysed their level of expression, examined the role of TEs as promoters of these novel transcripts and looked at their protein coding potential. The highest number of novel genes was identified in the oocytes, with majority of them being located in the intergenic regions. We demonstrated that majority of novel found genes in oocytes and the earliest embryonic stages (2C, 4C embryos) are specific for given developmental stage, with the specificity decreasing with the developmental progression. We further showed that TEs are associated with approximately 30-40% promoters in every developmental stage, with predominant class LTR-MaLR in the oocytes and classes LTR-EVRL and LTR-ERVK in the embryos. The coding potential of our novel genes is low with majority of transcripts being classified as non-coding.

2 Introduction

Mouse as one of the model organisms has its genome well annotated. But even still we see that when some low input or very specific samples are sequenced, as for example from early developmental stages, there is a considerable number of non-annotated transcripts. (Veselovska et al. 2015). These novel transcripts are often classified as lncRNAs. LncRNAs exhibit a number of unique features and functions. They can participate in and modify some cellular processes like transcription (Wei et al. 2017) or even regulate the patterns of gene expression (Wang et al. 2011). Their evolutionary conserved patterns together with their high tissue specificity (Vance and Ponting 2014) can also suggest that they might provide a potential therapeutic value (Gomes et al. 2017).

For a long time, the role of transposable elements (TEs) in shaping mammalian transcriptomes was overlooked, but recently increasing amount of data suggests that they play an important role in the early developmental stages of mammals. For example, TEs can act as promoters and it has been shown for a number of novel transcripts, especially those falling classified as lncRNAs in the oocytes (Veselovska et al. 2015).

This bachelor thesis focuses on characterisation and identification of novel found genes in oocytes and preimplantation embryos, analysis of their relative expression patterns, TEs activity as promoters of these transcripts and their protein coding potential, as a basis for future identification of candidate novel genes for functional analysis.

3 Background

Mouse as the most widely used mammalian model organism for science was the second mammal whose genome was fully sequenced after the human genome. Our knowledge of its genome has improved dramatically over the years, but the majority of studies rely only on already annotated reference genes for their analyses (Veselovska et al. 2015). This might lead to loss of crucial information, especially when studying very specific cell types providing only low input samples, and considering the cell-type specificity of some transcripts, especially lncRNAs. It has already been shown that there is a high number of non-annotated transcripts, especially in the early developmental stages such as oocytes (Veselovska et al. 2015), from which the majority appear to be lncRNAs, although small number of them have potential to be protein-coding based on bioinformatical analysis.

3.1 lncRNAs

LncRNAs are a dynamically evolving population of genes with a huge biochemical adaptability. It is not clear what fraction of lncRNAs carries some important functions but many of their features imply that a considerable number of lncRNAs emerge as opportunistic transcription events at random locations susceptible for transcription initiation (for example, due to the presence of a transcription factor binding sequence motif). They are transcribed as new non-functional elements whose functionality is acquired through natural selection (Ganesh et al. 2016). There is not one clear definition as to which transcripts are classified as lncRNA,

but classically the transcripts that do not code for proteins and are longer than 200 nucleotides are referred to as lncRNA (Cabilli et al. 2011). In comparison with the protein coding genes, lncRNAs are usually shorter, have fewer exons, lower expression level and higher cell-type or tissue specificity (Ulitsky et al. 2013, Derrien et al. 2012).

3.1.1 Functionality of lncRNAs

LncRNAs can have very diverse functions. They can exert their function as RNA molecule, or it can be just the effect of transcription from their locus, not the resulting RNA molecule, which is necessary. Determining the functions of lncRNA is a challenging task as they tend to have only short conserved regions across species surrounded by highly variable regions, or they might not be conserved at all (Ponjavic et al. 2007). They exhibit both *cis* and *trans* regulatory activities (Mercer and Mattick 2013). As their *cis* regulatory activity, they affect adjacent genes on the same allele. As *trans* regulatory activity, they affect genes on the other allele. The regulation of activity of other genes can be exerted for example through transcriptional interference (transcription of the lncRNA interferes with the transcription of the other gene), transcriptional stimulation (transcription of lncRNA stimulates the transcription of the other gene) or effects on chromatin (for example, X-chromosome coating followed by heterochromatinisation by Xist lncRNA, in order to silence one of the X chromosomes in females). LncRNA can interact with other ncRNAs. For example, the interaction with miRNAs creates a network that exerts post-transcriptional regulations of expression of genes. (Zampetaki et al. 2018). In addition, until recently, it was thought that lncRNA is a non-coding RNA, but this view changed when several research groups reported that certain lncRNAs code for short peptides with important functions (Ruiz-Orera et al. 2014).

Functions of RNAs associated with their structure is an emerging topic of interest. However, the structural domains of RNAs are not yet well defined. Concerning lncRNA, a theory emerged which says that the selective pressure of evolution acts rather on a structure than on the primary sequence, explaining the fast rate of their evolution (Wutz et al. 2002). LncRNA typically contains complex secondary and tertiary structure and their functions are barely constrained by the sequence itself (Zampetaki et al. 2018). Other than being able to control gene expression, new approaches are being made to connect their structure and function to design novel therapeutic approaches (Zampetaki et al. 2018).

3.1.2 Classification of lncRNA

LncRNA can be transcribed from different genomic locations. Some are located in large areas between annotated genes, these lncRNAs are referred to as intergenic lncRNAs. Second category of lncRNAs are transcribed from introns of protein-coding genes, called intronic lncRNAs. LncRNA can also overlap both exons and introns of protein-coding genes on the same strand, or on the opposite strand, such transcripts are classified as overlapping sense and antisense lncRNAs, respectively. They can either overlap with the protein-coding genes or stretch over the entire protein-coding gene body. The last category are so-called bidirectional lncRNAs which originate from the same promoter region as a protein-coding gene, but on the opposite strand. Classification of lncRNAs is visualised on Figure 1.

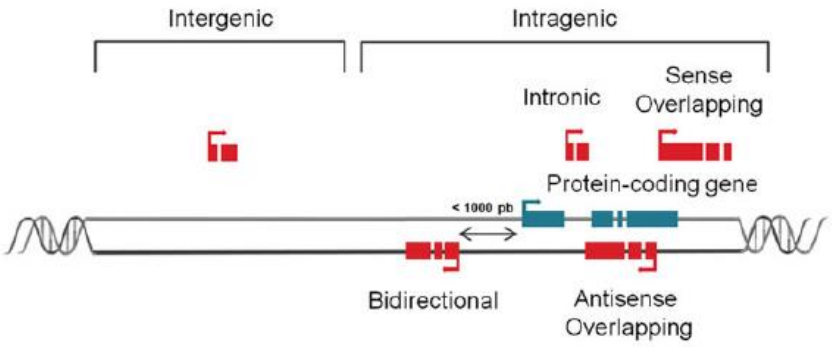


Figure *1*. General classification of lncRNA (Bouckenheimer et al. 2016)

3.2 Retrotransposons

The transposable elements found in mammals can be in general characterised into two classes based on whether their transposition element is DNA or RNA. They are Class 1 retrotransposons and Class 2 DNA transposons.

Retrotransposons, also called the copy and paste elements, create copies of themselves while they are transcribed from the genome and then reversely transcribed into DNA and integrated into the genome (Finnegan et al. 1989). They are further divided based on the presence or absence of long terminal repeats (LTR) elements. LTRs help in retrovirus-like reverse transcription and integration into the genome (Finnegan et al. 1989). LTR elements range in their length from short ones around few hundred base pairs (bp) up to 10 000 bp. These

elements encode a gag and a pol protein, flanked by the long terminal repeats, hence their name (Finnegan et al. 1989). The non-LTR elements consist of long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Classification of TEs is visualised on Figure 2.

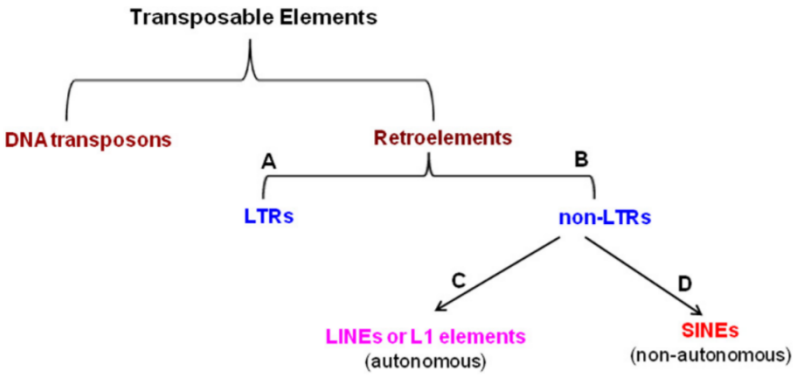


Figure *2*. Classification of transposable elements (Hemalatha et al. 2012)

3.2.1 Abundance of TEs in mammals

Out of all the transposable elements, the biggest proportion of our genome is occupied by LINEs and SINEs. These are followed by LTR elements and DNA transposons, respectively. For the most part, around 75% of the repetitive portion of the mammalian genome is derived from non-LTR retrotransposons (Waterston and Pachter 2002). The LINE-1 family is the most successful family of transposable elements in mammals. The SINE family which cannot be mobilised by itself profits from the success of the autonomous LINE elements.

LTR elements occupy between 4-10% of the genomes (Mikkelsen et al. 2007). DNA transposons compared to non-LTR elements have low copy numbers of only 3% of the mammalian genomes (Platt and Ray 2012).

Even though mammalian genomes contain thousands of TEs, only a small portion of these are capable of mobilisation. It is mostly because of the incomplete or mutated sequences of the individual insertions. Looking at LINE-L1 elements as an example, their incapability to mobilise is primarily caused by inefficient reverse transcription which truncates most sequences at their 5' end (Grimaldi et al. 1984). Additionally, new insertions of any type can be mutated during the process of insertion as well as be a target of transcriptional and post-

transcriptional silencing (Platt et al. 2018). Being able to comprehend the factors that might increase, decrease of eliminate the activity of transposable elements is one of the key questions for understanding the evolution genomes.

3.2.2 Effect of TEs on the genomes

The presence of TEs in mammals and other organisms shapes the evolution of their genome in a remarkable way. They affect our genomes in both positive and negative ways.

Mobile genetic elements insertions can disrupt genes, mediate chromosome rearrangements or provide alternative promoters (Benetzen, 2000). In addition, certain TEs contain transcription factor binding sites and other regulatory motifs (Bourque et al. 2008). During a mobilisation events of such TEs, they are spreading the sequence motifs around the genome, to the new integration sites. This might lead to change of the gene expression if the TEs integrates near the promoter region, or a novel regulatory network can appear where one TEs-associated transcription factor binding site links dozens or hundreds previously unrelated genes (Chuong et al. 2017).

A process of evolving new functions of existing genes is called co-option. A gene can be co-opted by changing their regulatory patterns or their function which at the end produces a novel gene. In the non-LTR TEs families, the frequency of their participation in co-option usually increases with the age of the TEs subfamily. On the contrary, the co-option of LTRs is inversely proportional to their age (Franke et al. 2017). LTRs can act as platforms for redesigning genes where the promoter of LTR element and few initial exons become a part of the novel gene. In one study, it was revealed that around 25% of experimentally characterised human promoters contain sequences derived from TEs (Jordan et al. 2003). Another study showed that many promoters in humans and mice are derived from TEs specific for primates and rodents respectively (Marino-Ramirez et al. 2005). TEs are a rich source of promoters and cis-regulatory elements, which have the potential to regulate the transcription of neighbouring genes. It is believed, that this phenomenon exists because there is a constant accumulation of TEs in one's organism and this accumulation creates raw sequence material from which cis-regulatory elements arise by point mutations. Another explanation says that cis-regulatory elements exist within TEs and are co-opted after insertion of TEs or modification of the surrounding region (Feschotte 2008).

All of the changes happening in the genome because of TEs, like alteration of gene expression or reconstruction of the genomes, help populations to fully explore their potential fitness landscape in a shorter amount of time, which increases the so-called adaptability of the population (Casacuberta and González 2013). There are two hypotheses speculating about the roles of TEs in promoting adaptability in mammals. First hypothesis is called the epi-transposon hypothesis (Zeh et al. 2009) which suggests that when the organism is exposed to environmental stress, the suppression of TEs is stopped which accounts for increased activity of TEs. This increased activity allows for increased changes in the genomes which subsequently help the population to explore their fitness landscapes (Platt et al. 2018). Second hypothesis is the TEs-Thrust hypotheses (Oliver and Greene 2011). It suggests that lineages with higher TEs activity are more fertile than those without it.

Transposable elements are strong mutagens. Around 10% of all novel mutations in lab mice result from insertions of TEs (Maksakova et al. 2006). Having excessive amount of TEs in one's genome may lead to accumulation of mutations causing decrease in the fitness of the population as well as decrease in the fecundity.

Somatic tissues were not, until recently, a primary topic of research of TEs. But nowadays it is known that more than 100 diseases including a few forms of cancer are associated with insertions of TEs (Chénais 2013). TEs can cause cancer by altering tumor suppressor genes or proto-oncogenes (Morse et al. 1988).

3.3 lncRNAs and TEs-derived lncRNAs in mammalian early development

It has been suggested many times that lncRNAs may play an important role in the early mammalian development. The functions of most lncRNAs are still unclear but some studies already linked several lncRNAs with the embryo development (Huynh et al. 2003, Okamoto et al. 2005, Wu et al. 2018). Analysis of lncRNA in early oocytes and embryos is technically difficult due to lncRNA being expressed at lower levels as well as in limited amounts of cells (Zhang et al. 2014).

Compared to protein coding genes, TEs are common promoters for the lncRNAs in early mammalian development (Franke et al. 2017). It has been shown that promoters and 5' exons derived from LTRs are part of more than 800 protein coding genes and lncRNA expressed in

the oocyte to zygote developmental stages in rodents (Franke et al. 2017). In addition, expression of TEs and transcripts using TEs as promoters was shown to be an essential component of totipotency of 2C embryos and 2C-like embryonic stem cells (Macfarlan et al. 2012)

LncRNAs have therefore the potential to play important roles in early developmental stages. However, they are still not fully annotated. In this thesis we try to thoroughly identify and characterise lncRNAs and other non-annotated transcripts in the mouse oocyte and preimplantation embryos during major zygotic genome activation and segregation of first two embryonic cell lineages.

4 Aims

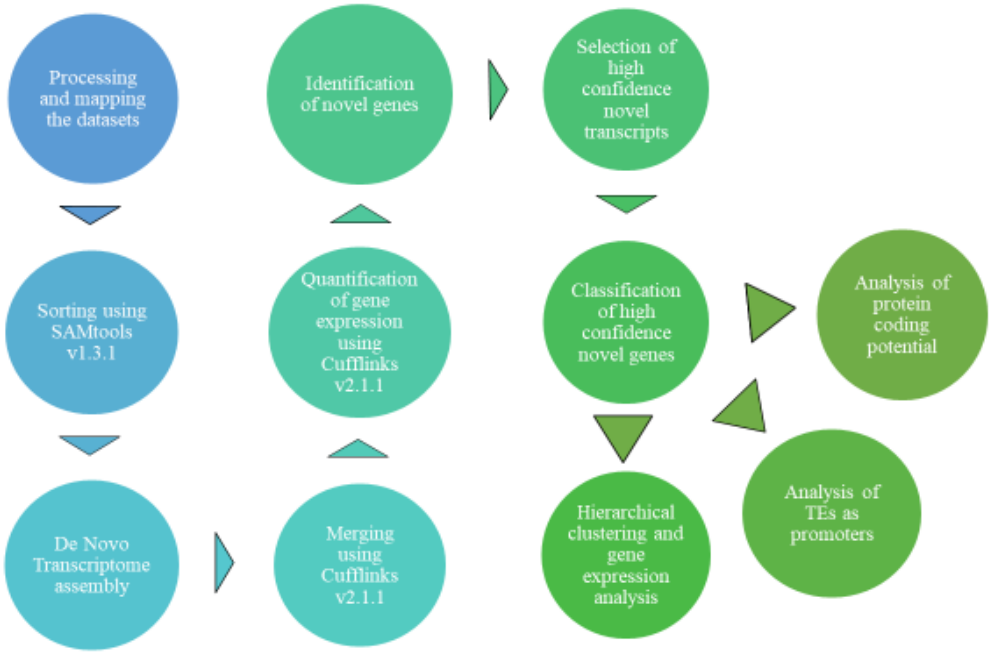
- to identify and annotate novel transcripts from RNA-seq data in mouse oocytes and preimplantation embryos

- to characterise their expression profiles (whether they are oocyte-/embryo-specific)

- to determine which transposable elements are the most frequently acting as promoters of these transcripts

- to characterise the protein coding potential of novel transcripts

5 Methods



5.1 Processing and mapping of RNA-seq datasets

RNA-seq datasets from oocytes, preimplantation embryos and somatic tissues (accession numbers GSE70116, GSE98150, GSE75957, respectively) were downloaded as fastq files from European nucleotide archive (https://www.ebi.ac.uk) database. Complete list of datasets used is in the Appendix 1. To trim the adapters and low quality bases, program TrimGalore! (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) v0.4.1 was used. Quality check of the trimmed reads was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) v0.11.5 with default parameters. The trimmed reads were mapped on mouse GRCm38 genome using Hisat2 (Kim et al. 2019) v2.0.5. All of these steps were performed previously in the laboratory.

5.2 Sorting

In order to be able to perform de novo transcription assembly with Cufflinks, the mapped data were sorted using the SAMtools v1.3.1 with the function samtools sort. The command for sorting was ‘samtools sort -o output_sorted.bam input.bam’.

5.3 De novo transcriptome assembly

De novo transcriptome assembly of individual oocyte and selected embryo datasets (not the 8C stage and morula embryonic datasets and datasets from somatic tissues) was performed using Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/) v2.1.1 in the reference annotation base transcript (RABT) mode (specified by -g command) with default parameters where we specified only the strand specificity of the library with command --library-type (the options were fr-unstranded, fr-firststrand, fr-secondstrand). See Appendix 1 for library type parameters for each dataset. The reference annotation was previously downloaded from Ensembl genome browser (GRCm38 v94, file named as Mus_musculus.GRCm38.94.chr.gtf). The command for transcriptome assembly was ‘cufflinks -g Mus_musculus.GRCm38.94.chr.gtf -u --library-type fr-xxx -o Output_folder sorted_bam_file’. The commands were executed using Metacentrum computer cluster by submitting xxx.sh script (example script for transcriptome assembly is in the Appendix 2).

Assembled transcriptomes from individual datasets were then merged together for each developmental stage, creating one final annotation for each stage - oocytes, 2-cell (2C) stage embryos, 4-cell (4C) stage embryos, inner cell mass (ICM) and trophoectoderm (TE). For oocytes, assemblies from four different oocyte developmental stages were merged (datasets named as d5, d10, d15 and GV), while for the embryonic stages, biological replicates of the same stage were merged. The merging was performed using the cuffmerge function within Cufflinks v2.1.1 with the command ‘cuffmerge xxx.txt’ where the text document contained the list of annotation files to be merged.

5.4 Quantification of gene expression

The expression of transcript isoforms and genes in the final merged assemblies from oocytes and embryonic developmental stages were quantified using Cufflinks v2.1.1 disabling the de

novo transcriptome assembly function with the -G command. The command for the quantification was ‘cufflinks -G xxx_merged.gtf output_folder sorted_reads.bam’. The quantification was performed for each merged assembly across all the datasets, including datasets from somatic tissues. The commands were executed using Metacentrum computer cluster by submitting xxx.sh script (example script for expression quantification is in the Appendix 3). The unit of expression is fragments per kilobase of the transcript or gene per million reads in the library (FPKM) for paired-end sequenced datasets, as in our case all the datasets were paired end). The results of quantification of each assembly resulted in two files – isoforms.fpkm_tracking file with the quantification of individual transcript isoforms and genes.fpkm_tracking file with the quantification of expression of whole genes where the expressions of isoforms belonging to the same gene are added together.

5.5 Identification of novel genes

For the identification of novel genes, we first created a list of all genes in each final assembly with their genomic coordinates. The information about gene name, chromosome, start position and end position was extracted from the gene expression quantification output file gene.fpkm_tracking described in section 5.4. To obtain information on which DNA strand the genes are encoded, we imported final assembly gtf files into Seqmonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/) v1.44.0, made probes over assembly mRNAs (corresponding to the transcript isoforms) and exported it, as Seqmonk output files contain probe strand information. The exported transcript isoform strand information was matched with corresponding genes using the information in the isoforms.fpkm_tracking file described in section 5.4 which contain information about transcript isoform name and corresponding gene name in Microsoft Excel v16.2003 with functions MATCH and LOOKUP. We then removed all genes without strand information (in cases that originating DNA strand was not determined during de novo transcriptome assembly). This was performed in Microsoft Excel v16.2003 by sorting the column with strand information and deleting those genes that lacked the specification of their strand. To select only novel unannotated genes from all genes in our assemblies, we applied two criteria – the gene should not overlap any known annotated gene on the same strand in the region +/-5kb and should have <50% overlap with TEs on the same strand. This was performed using Seqmonk v1.44.0. For filtering genes without overlap with known genes +/-5kb, we made probes over

known annotated genes +/-5kb, imported them as reads, imported all genes in our assembly as the annotation track, quantified read count on same strand as probe and removed all probes except those with read count 0. Because known genes +/-5kb were imported as reads, read count 0 meant no overlap with known gene +/-5kb on the same strand, while read count of 1 or more meant there was an overlap with known gene +/-5kb on the same strand and therefore the gene was not considered novel. For filtering genes with <50% overlap with TEs on the same strand, we used filtered genes from the previous step as probes, imported TEs annotations as reads, quantified %overlap on the same strand and removed probes with >50% overlap. TEs annotation was downloaded from UCSC genome browser and split into separate annotation files for each TEs class previously in the laboratory. The remaining genes after these filtering steps were classified as novel genes.

5.6 ChIP-seq datasets processing and mapping, peak calling, and selection of high confidence novel transcripts

Trimethylation of lysine 4 on histone 3 (H3K4me3) is a histone modification associated with active promoters. We used H3K4me3 data to identify which novel genes have a proper promoter H3K4me3 peak at their 5' end. Processing and mapping the mouse oocyte and embryo H3K4me3 ChIP-seq datasets with accession numbers GSE73952 (Liu et al. 2016), GSE93941 (Hanna et al. 2018), GSE71434 (Zheng et al. 2016) (see Appendix 1 for detailed information) was done previously in the laboratory. Briefly, the datasets were downloaded from European nucleotide archive (https://www.ebi.ac.uk) database, trimmed using TrimGalore! v0.4.1 and mapped to the mouse GRCm38 genome using Bowtie2 (Langmead and Salzberg 2012). Novel genes identified as described in chapter 5.5 were imported into Seqmonk v1.44.0 as annotation, while ChIP-seq datasets were imported as data to identify the H3K4me3 peaks we used MACS peak caller within Seqmonk specifying H3K4me3 and input datasets with default parameters except datasets from Zheng et al. for which default p-value was changed to 0.001 to achieve identification of similar number of peaks as in other datasets. Peaks were called for each developmental stage from each publication separately. Peaks were then imported as annotations (commands data -> import annotation -> active probe list). To identify which novel genes have promoters associated with H3K4me3 peaks, we created probes over TSSs of novel genes +/-1000 bp (commands data -> define probes -> feature probe generation -> select novel genes annotations -> uncheck the parameter 'remove exact

duplicates' -> make probes-upstream of feature from -1000 to 1000 bp). When creating annotated probe report, we overlapped novel gene promoters with individual sets of peaks for same developmental stage as novel genes. The resulting txt files were then imported in Microsoft Excel, where we combined the information of peaks overlaps from all three publications for each developmental stage. Novel genes with promoter overlapping H3K4me3 peak from at least one publication were considered high confidence novel genes.

5.7 Classification of high confidence novel genes

To classify our novel high confidence genes based on their genomic location in respect to known genes we used Seqmonk v1.44.0 and Microsoft Excel v16.2003. The classification categories were intergenic (location without proximity to any known genes on either strand), intragenic (fully within known gene on the opposite strand), overlapping (partially within known gene on the opposite strand) and bidirectional (sharing a promoter with known gene, but coded on the opposite strand). We uploaded the novel genes into Seqmonk as annotation track and filtered them (based on their names) to have only the high confidence novel genes defined in chapter 5.6. We characterised these genes for two features – bidirectionality of promoters and extent of overlap with known gene on the opposite strand. Known genes except genes encoding short non-coding RNAs were uploaded as reads. To find out whether high confidence novel genes are sharing a bidirectional promoter with a known gene, we made probes -2kb to -1kb upstream of high confidence novel genes quantitated read count on the opposite strand as probe and exported the results. To assess the extent of the overlap with a known gene, we made probes over novel genes and quantified %coverage quantitation using reads on the opposite strand to probe and exported the results. We then opened both exported files in Microsoft Excel, made sure there is the same number of genes in both of them and deleted all the columns except columns containing gene names and quantified results. The genes with value 1 or higher in the analysis of bidirectional promoters were classified as bidirectional except those with value 100 in the analysis of the overlap with known genes on the opposite strand. The genes with value 1 or 0 from the analysis of bidirectional promoters and value 100 in the analysis of overlap with known genes on the opposite strand (meaning that they have 100% coverage by known gene) were classified as intragenic.

The genes with value 0 in the analysis of bidirectional promoters could be intergenic or overlapping. Genes with values >0 and <100 in the analysis of overlap were classified as

overlapping (as the overlap with a known gene is between 0-100%) and genes with value 0 (meaning no overlap with a known gene on opposite strand) were classified as intergenic.

5.8 Hierarchical clustering and gene expression analysis

To study the expression patterns of novel genes, we used gene expression values quantified by Cufflinks as described in section 5.4 (from gene.fpkm_tracking output files). These files were filtered in Microsoft Excel v16.2003 to contain only novel high confidence genes, separately for each class of novel genes for each developmental stage. For each high confidence novel gene, we quantified the average of the expression values across replicates of each stage of embryonic development and for each of the somatic tissues and of four oocyte datasets. Then, we log transformed the values (with base 2). To be able to work with genes with expression level of 0, we added 0.001 to each expression value prior to log transformation. We quantified overall average value across all log transformed expression levels for each gene and subtracted these averaged values from each log transformed expression values, giving us relative expression values. These relative expression values together with gene names were exported as txt files and used for hierarchical clustering and generation of heatmap. Hierarchical clustering was performed and heatmap was generated in R Studio v3.6.2, the script is in Appendix 4.

5.9 Analysis of transposable elements as promoters

To find out which TEs and if any, are acting as a promoter for high confidence novel genes we used Seqmonk v1.44.0 followed by Microsoft Excel v16.2003. In Seqmonk we imported TEs annotations described in section 4.5 as reads, and high confidence novel genes as annotation track. We made probes around transcriptional start sites (TSSs) of novel genes (-50bp to +50 bp around TSSs) and quantitated read count on same strand as probe. This quantified if the TEs are overlapping the TSSs of the gene on the same strand. The exported file from Seqmonk was opened in Microsoft Excel where we calculated how many TSSs are associated with each class of TEs using the function COUNTIF (to find those that have value quantified >0). This analysis was performed separately for each class of novel high confidence genes (intergenic, intragenic, overlapping, bidirectional) for each studied developmental stage.

5.10 Analysis of protein coding potential

To analyse whether the high confidence novel genes have the potential to encode proteins, we first obtained sequences of transcript isoforms of these genes in fasta format using two Python scripts (filtering_gtf.py and gettingSeq.py) which were previously developed in the laboratory.

First, we filtered our assembled transcriptomes to contain only transcripts of novel high confidence genes for each developmental stage and for each class (intergenic, intragenic, overlapping, bidirectional). We took the genomic coordinates (chromosome, start, end) of novel high confidence genes, extended them by 1bp on each side and created a txt file with such coordinates. This txt file was used as an input file for the first python script filtering_gtf.py (Appendix 5) together with the original assembled transcriptome annotation (gtf file). The script uses the coordinates supplied in the txt file to filter the gtf file to contain only the transcripts within the specified regions.

The filtered gtf file served as an input for the second Python script gettingSeq.py (Appendix 6), which generates sequences of transcripts in the supplied gtf file. In addition to the gtf file, it requires together with the raw mouse GRCm38 genomic sequence split into individual chromosomes downloaded from Ensembl and a list with names of transcripts from the gtf file for which we want to obtain the sequence (in our case, it was all transcripts in the filtered gtf files which we obtained by importing the gtf file into Seqmonk, making probes over all mRNAs and exporting it). The script generated sequences of the specified transcripts in fasta format. These fasta sequences were then submitted into the Coding Potential Calculator 2 (CPC2) web interface (Kang et al. 2017) which assessed the protein coding potential of individual transcripts.

6 Results

6.1 De novo transcriptome assembly

For identification and characterisation of novel genes that were not previously annotated during mouse early development, we used 45 publicly available RNA-seq datasets comprising oocyte datasets (Veselovska et al. 2015), preimplantation embryo datasets (Wang et al. 2018) and datasets from adult somatic tissues (Andergassen et al. 2017). The complete list of datasets used including the number of replicates for each sample type is in Table 1 and Appendix 1.

Table 1. Datasets used

Sample type	Sample	Number of replicates	Reference
Oocyte	d5	1	Veselovska et al. (2015)
	d10	1	
	d15	1	
	GV	1	
Preimplantation embryo	2C	4	Wang et al. (2018)
	4C	4	
	8C	3	
	morula	2	
	ICM	4	
	TE	4	
Somatic tissue	adult brain	4	Andergassen et al. (2017)
	adult liver	4	
	adult heart	4	
	adult leg muscle	4	
	adult lung	4	

Datasets were trimmed and mapped previously in the laboratory. After sorting the mapped reads, we performed de novo transcriptome assembly using Cufflinks on individual datasets from oocytes and preimplantation embryos. We were particularly interested in novel genes in the oocytes, as oocytes represent very specialised cell type and store many RNAs essential for successful early development after fertilisation, in 2C and 4C embryos to capture genes transcribed during major zygotic genome activation, and in ICM and TE, to potentially discover novel genes involved in segregation of first two embryonic cell lineages. Therefore, de novo

transcriptome assembly was performed on these datasets, and the remaining ones (8C, morula and somatic tissues) will be used only for quantification of expression of identified novel genes. The assembled transcriptomes were then merged together using Cufflinks cuffmerge command. All oocyte assemblies were merged together and assemblies from each embryonic stage replicates were merged together for each stage. The final numbers of transcripts for every merged transcriptome are listed in Table 2.

Table 2. Number of transcripts for every merged assembly

Merged datasets	Number of transcripts
Oocytes	971610
2C	1005445
4C	953043
ICM	915142
TE	871613

6.2 Identification of high confidence novel genes

Merged transcriptomes (gtf files) generated in section 6.1 contained only the annotation of transcripts, not the whole genes (each gene comprises one or more transcript isoforms which represent splicing variants). However, for the identification of novel genes, we needed the annotation of genes from de novo transcriptome assembly. To obtain the annotation of genes, we performed gene expression quantification on the merged gtf files using Cufflinks. The outputs of this quantification are two files – a file isoforms.fpkm_tracking with expression values of individual transcripts and a file genes.fpkm_tracking with expression values of genes, including their genomic location. Therefore, we obtained genomic coordinates of whole genes from this file, except the information about DNA strand on which the genes were encoded. To get the information about their strand we used Seqmonk v1.44.0 with the merged transcriptomes (gtf files generated in section 6.1) imported as annotation tracks. For individual merged transcriptomes we created annotated probe reports with mRNA isoforms as probes, containing information about the transcript name and the transcript strand. On the other hand, the isoforms.fpkm_tracking files created by Cufflinks quantification contain an information about transcript name and name of a gene the transcript belongs to. Therefore, using transcript names, we matched the strand information from the Seqmonk report with gene name from

Cufflinks isoforms.fpkm_tracking file. We then removed all genes without strand information, as they were the genes for which it was not possible to determine coding DNA strand by Cufflinks. Using these steps, we created annotation of all genes assembled during de novo transcriptome assembly.

The next step was to identify novel genes. We defined novel genes as those genes that do not overlap with any known annotated gene on the same strand within the region +/-5kb to avoid identification of 5' or 3' extensions of already annotated genes with weak expression, therefore resulting in discontinuous annotation resembling one or an array of same strand monoexonic genes close to 5' or 3' ends of annotated genes. In addition, we required genes to have <50% overlap with TEs on the same strand to exclude expressed active TEs. These are in contrast with genes using TEs as promoters but having significant proportion of non-TEs sequence. This was done using the SeqMonk v1.44.0. The results showed that after each filtration step the number of novel genes decreased dramatically, from more than 100 000 assembled genes in mostly all the merged datasets to only around few thousand novel genes left at the end of the filtrations. In oocytes we found the smallest number of genes at first (69 202) but after all the steps of elimination, it showed that oocytes have the highest number of novel genes left (9902), followed by 2C stage (4755). Due to the lack of strand specificity of embryonic datasets, approximately half of genes in the embryonic datasets are without strand information. Visual inspection of the data showed that these genes are either short monoexonic genes in the intergenic regions, or arrays of short monoexonic genes in the introns of multiexonic genes with strand information, which are probably just a mis-annotation of reads coming from yet unspliced nascent mRNAs (see Table 3).

Table 3. Number of novel genes after every filtration. A-number of all found genes, B-number of genes that have strand information, C-number of genes that have a strand information and that do not overlap with any known gene +/- 5kb on the same strand and have <50% overlap with TEs.

	A	B	C
Oocytes together	69202	69143	9902
2C	111245	53233	4755
4C	108101	51996	3377
ICM	118718	50189	1807
TE	90610	49181	1262

H3K4me3 is a histone modification generally linked to activation of transcription. It is associated with promoters of genes that are active or destined to be so (Kouzadires, 2007). Therefore, we wanted to use H3K4me3 peaks as a confirmation that promoters of our novel genes are true active promoters. Using this approach, we wanted to select high confidence novel genes (with promoter associated with H3K4me3 peak) from all identified novel genes. We performed peak calling with the publicly available H3K4me3 Chip-seq datasets (Zheng et al. 2016, Liu et al. 2016, Hanna et al. 2018) from relevant developmental stages (oocytes, 2C and 4C embryos, ICM and TE) which were downloaded, trimmed and mapped previously in a laboratory. Then, we looked for overlaps of these peaks with promoters (+/-1000bp around the TSS) of previously identified novel genes, i.e. strand-specific genes that do not overlap with any known gene +/-5kb on the same strand and have <50% overlap with TEs on the same strand. High confidence novel genes were defined as novel genes with promoter overlapping H3K4me3 peak from at least one dataset. The numbers for high confidence novel genes were 2653 in the oocytes, 685 for 2C embryos, 1156 for 4C embryos, 602 genes for the ICM and 411 genes for TE.

After obtaining the high confidence novel peaks, we divided them into 4 categories: intragenic for those that are located completely within the genes on the opposite strand, intergenic for those located between the known genes without any overlap with any known gene, bidirectional for those that share a promoter with the known gene but are coded on the opposite strand and overlapping for those that overlap with a known gene on the opposite strand. This categorisation helped us to see where our novel genes are located. For all the datasets, the majority of genes is localised in the intergenic region and the smallest number of genes is found in the overlapping region tightly followed by bidirectional region in 4C, ICM and TE datasets. (Table 4)

Table 4. Number of genes for each category

	Intragenic	Intergenic	Bidirectional	Overlapping	total
oocytes	324	1857	317	155	2653
2C	113	382	121	69	685
4C	231	698	114	113	1156
ICM	154	328	64	56	602
TE	135	187	47	42	411

6.3 Heatmaps and hierarchical clustering

In this analysis we wanted to find out what are the expression patterns of our high confidence novel genes and whether they are specific for certain developmental stage or rather expressed similarly throughout all developmental stages. From the values of gene expression obtained in 5.8 we calculated their relative expression which we used to create heatmaps and for hierarchical clustering. The results showed that in oocytes, especially the genes in the intergenic region as it is the category with highest number of genes, the genes are very specific only for oocytes (see Figure 3). In 2C and 4C embryos, the genes are mostly expressed in the early embryos stages when all the cells in the embryo are pluripotent (2C to 8C or morula), but absent from ICM and TE after the first embryonic lineage segregation and from differentiated adult somatic tissues. ICM and TE novel genes are expressed across all embryonic datasets with the highest expression in their own developmental stages and lower expression in the earlier embryonic stages, but they are not expressed in somatic cells or oocytes (see Figure 4). Based on the four categories of gene location (intragenic, intergenic, overlapping, bidirectional) we could see that across all datasets the expression of the genes from the overlapping, intragenic and bidirectional was slightly less specific than in intergenic genes as the genes were more often expressed in other developmental stages, although still maintaining the highest expression in the datasets where they were identified. For all remaining heatmaps see Appendix 7.

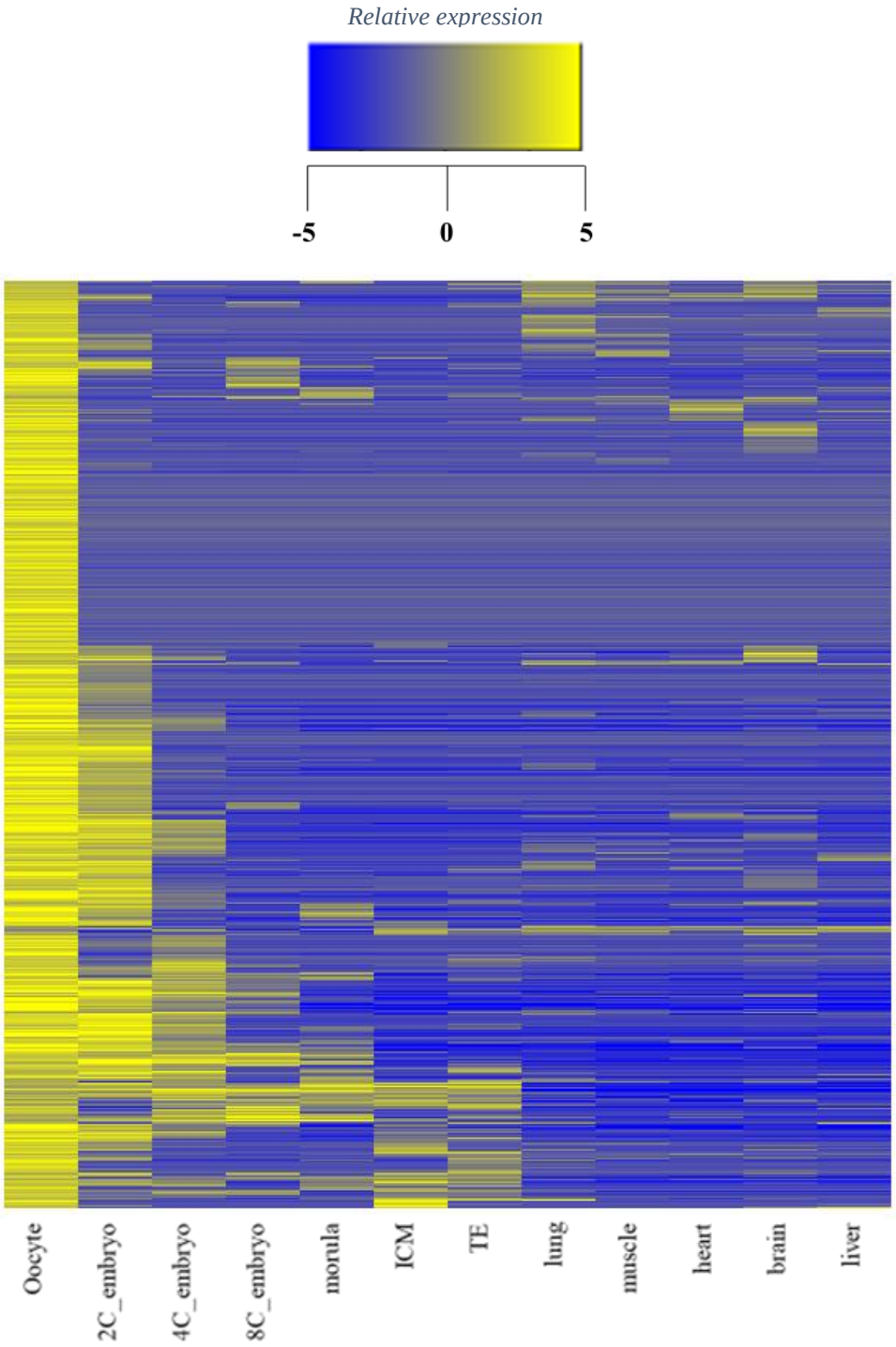


Figure 3. Hierarchical clustering of oocyte intergenic genes

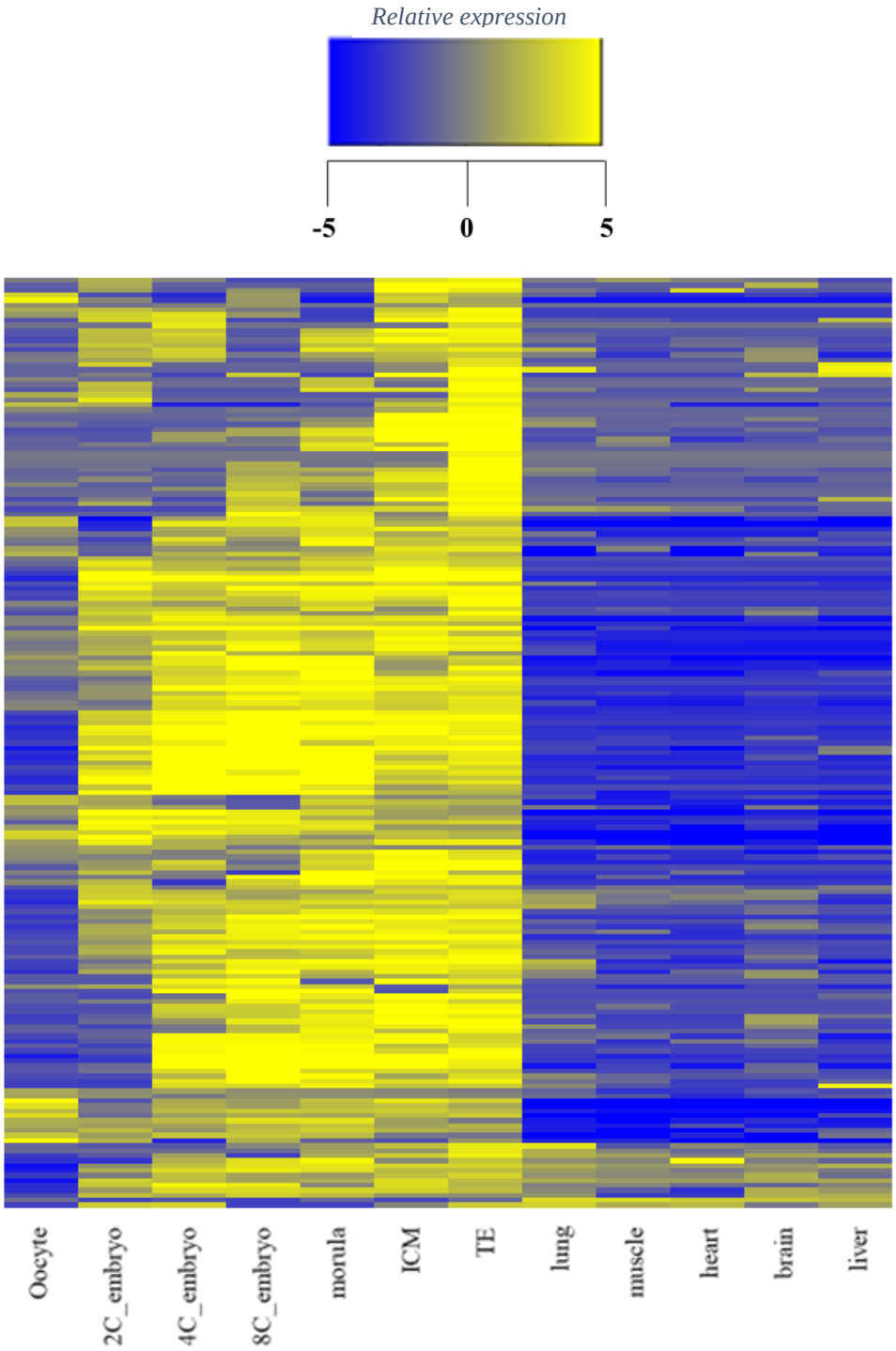


Figure 4. Hierarchical clustering of TE intergenic genes

After visually inspecting the heatmaps, we could see that there are groups of genes with different expression profiles. We decided to divide the genes into different clusters depending on their expression profile. For oocyte, novel genes in all 4 categories of their location were divided into 8 clusters, in 2C, 4C embryos, the genes were divided into 6 clusters in all categories. In ICM datasets, bidirectional and overlapping category were clustered into 4 groups, intergenic and intragenic category into 8 groups. For TE dataset, bidirectional category was divided into 3 clusters, overlapping into 4 clusters and the remaining categories (intragenic and intergenic) into 6 clusters. After division into the clusters, there were few clusters that had only 1 gene in them. We did not take these clusters into the account for further analysis. The number of genes in every cluster for the intergenic category as this category has the highest number of novel genes overall, for all datasets, is in Table 5, the remaining clusters and their number of genes can be found in Appendix 8.

The analysis of average relative expression across datasets in the individual clusters agreed with the results obtained by the visual inspection of the heatmaps. In oocytes we saw that majority of the novel genes are oocyte-specific, or at least predominantly expressed in the oocytes. We can observe this pattern in the expression profiles of 6 out of 8 clusters in the intergenic novel genes (cluster 6 has highest average relative expression in the brain tissue and cluster 7 has the highest average relative expression in 8C embryo). This pattern is repeated also for the clusters of oocyte genes in the intragenic region where 5 out of 8 clusters have the highest average relative expression in the oocytes (clusters 5, 7 and 8 have the highest average relative expression in lung tissue). In the overlapping and bidirectional regions, the oocyte genes have 6 out of 8 clusters with highest average relative expression in the oocytes (clusters 7 and 8 for overlapping regions show highest average relative expression in liver and muscle tissues, respectively, and clusters 2 and 8 for bidirectional genes show highest average relative expression in lung tissue and 4C embryos). In embryo cells, the analysis of expression patterns corresponded with what we observed on the heatmaps of embryo datasets. The 2C novel genes in all 4 categories have almost all highest average relative expression in 2C cells except cluster 6 in the bidirectional region which has highest expression in brain tissue and cluster 4 in overlapping region which has highest average expression in ICM. For 4C and ICM developmental stages, more genes were less typical for their developmental stage with 6 out of 24 clusters in both 4C and ICM datasets not having the highest relative expression for that specific dataset. TE showed the smallest number of clusters being with the highest expression

in TE which agrees with the results we got from the heatmaps. Only 10 out of 19 clusters had the highest expression in TE.

It should also be noted that across all datasets clusters with larger number of gene tend to follow the main patterns described (oocyte-specificity of oocyte novel genes, early embryonic expression or whole preimplantation development expression of 2C and 4C novel genes, and increasing expression across whole preimplantation development of TE and ICM novel genes, without expression in somatic tissues (Figures 5-9). However, clusters with small number of genes cluster genes with unusual expression patterns, for example with high expression in some somatic tissue (Appendix 9).

Table 5. Number of genes in each cluster in all intergenic datasets.

Cluster	Number of genes for				
	Oocytes	2C	4C	ICM	TE
1	1055	96	56	67	31
2	167	174	7	72	17
3	154	48	21	127	115
4	83	16	114	8	11
5	289	47	471	18	2
6	19	1	29	27	11
7	83	n/a	n/a	8	n/a
8	7	n/a	n/a	1	n/a

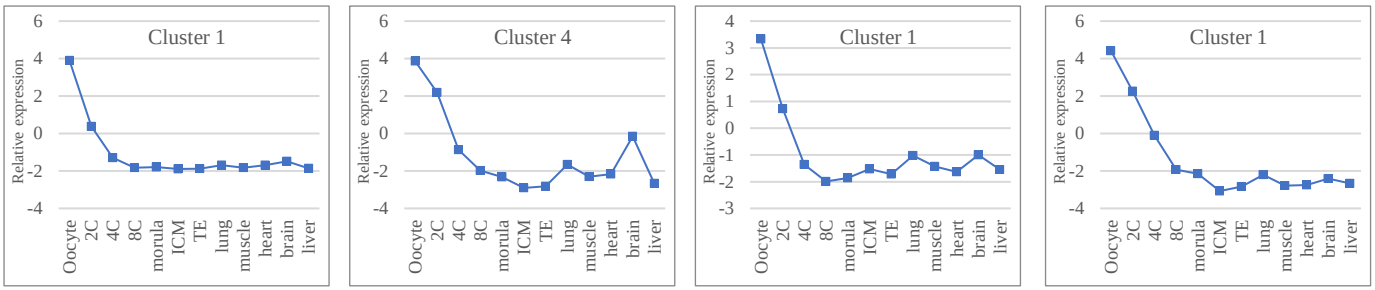


Figure 5. Average relative expression of a cluster with highest number of genes from intergenic, intragenic, bidirectional and overlapping (left to right) oocyte high fidelity novel genes

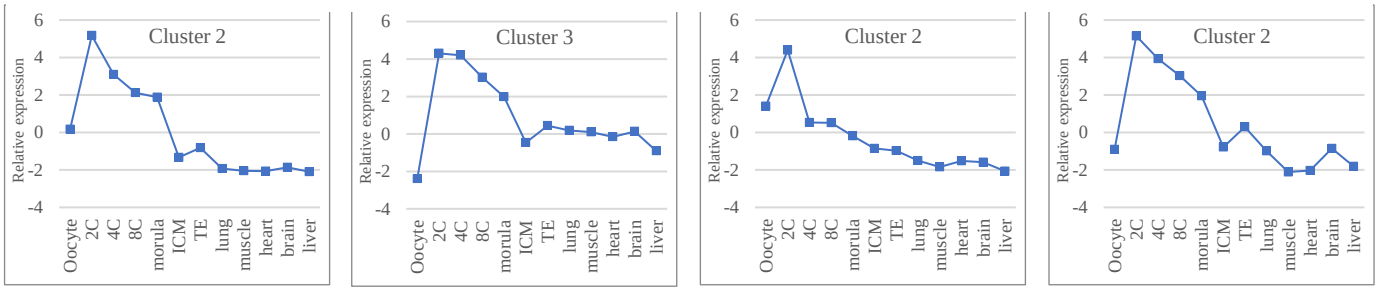


Figure 6. Average relative expression of a cluster with highest number of genes from intergenic, intragenic, bidirectional and overlapping (left to right) 2C high fidelity novel genes

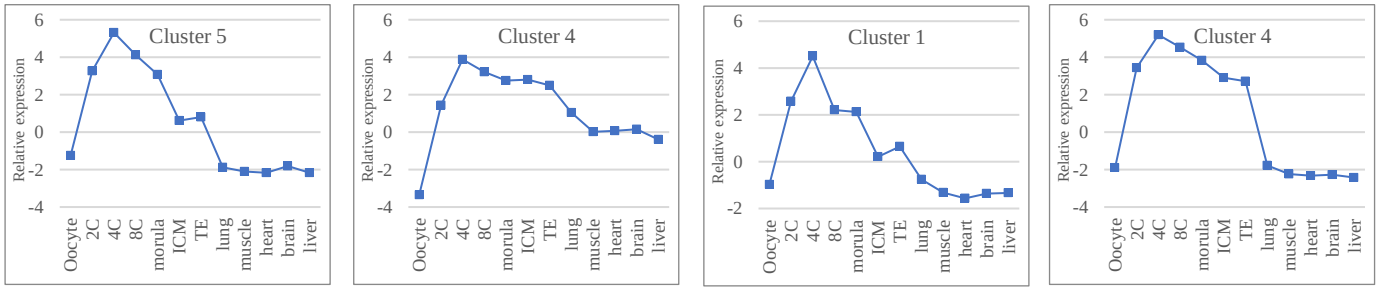


Figure 7. Average relative expression of a cluster with highest number of genes from intergenic, intragenic, bidirectional and overlapping (left to right) 4C high fidelity novel genes

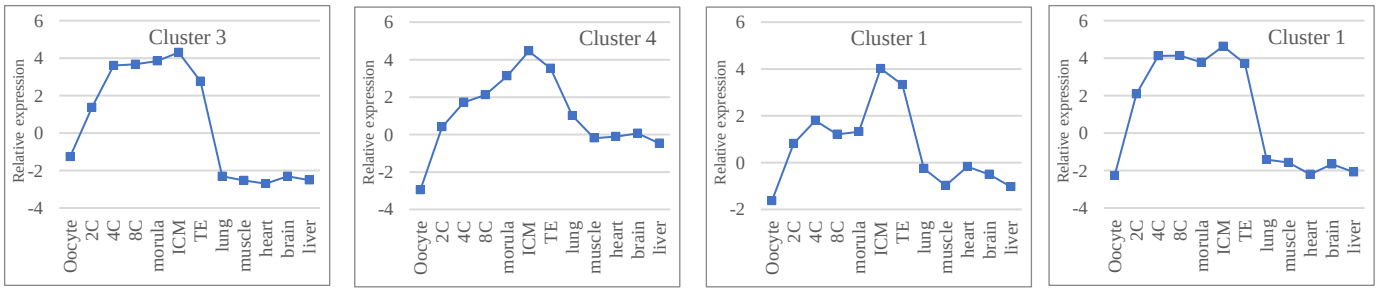


Figure 8. Average relative expression of a cluster with highest number of genes from intergenic, intragenic, bidirectional and overlapping (left to right) ICM high fidelity novel genes

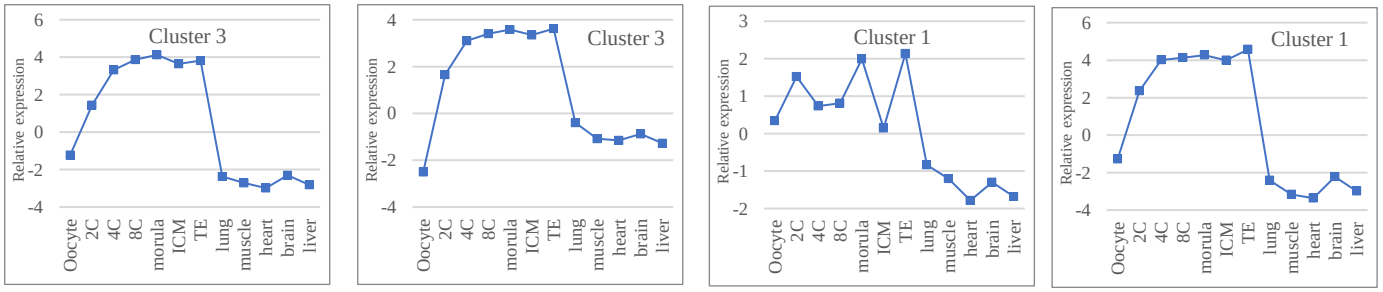


Figure 9. Average relative expression of a cluster with highest number of genes from intergenic, intragenic, bidirectional and overlapping (left to right) TE high fidelity novel genes

6.4 Analysis of transposable elements acting as promoters

Transposable elements were found to be associated with some promoters in humans and rodents in early development. We wanted to find out how many and if any are acting as promoters in our high confidence novel genes. For this purpose we used these 8 transposable element classes, LINE L1, LINE L2, LTR ERV1, LTR ERVK, LTR ERVL, LTR MaLR, SINE B2, SINE B4, and we investigated whether they overlap TSS of a same strand gene. The results of the analysis of transposable elements acting as promoters showed that the highest number of TEs promoters is in oocytes (43%) and the lowest in ICM (32.2%) tightly followed by TE cell lineage (32.3%) (see Figure 10). In oocytes, the most common class are LTR-MaLR elements (27.63% of all promoters) and LTR-ERVK elements (10.55% of all promoters). In embryos, the most common are LTR-ERVK elements for ICM and TE (11. 3% and 9.98%, respectively, of all promoters) and LTR-ERVL elements for 2C and 4C (13.28% and 12.98%, respectively, of all promoters) (see Figure 11). Most of the TEs promoters were associated only with 1 transposable element. In majority of developmental stages there was around 1% of promoters which overlapped with 2 or more TEs, but in TE datasets we did not find any promoters with more than 1 transposable element (see Figure 13).

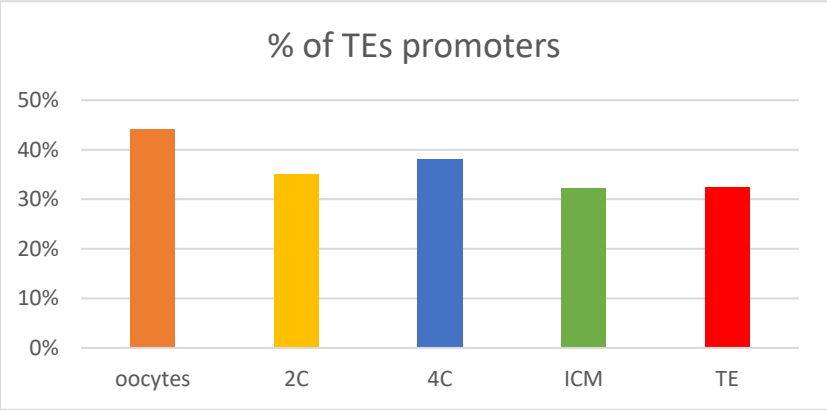


Figure 10. Percentage of TEs promoters from all the promoters of individual developmental stages

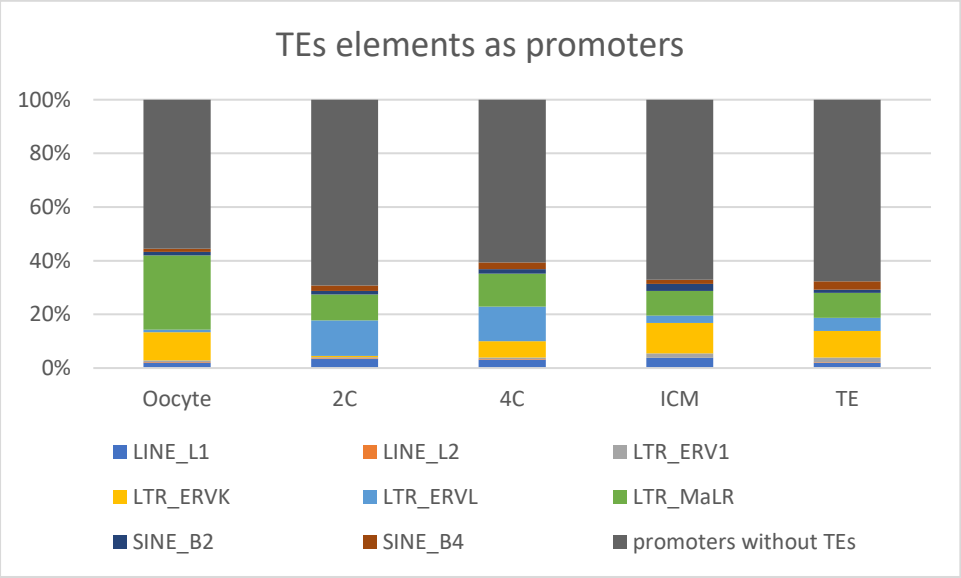
The results were also divided according to the 4 categories of gene location (intergenic, intragenic, overlapping, bidirectional) (see Table 6). This showed that the highest proportion of TEs promoters is in oocytes overlapping region with 53% of all novel overlapping genes having TEs as a promoter. In 2C developmental stage, the overlapping region was the richest for TEs promoters as well with 49% of all novel genes found in this region having TEs as a

promoter. Second most abundant category for TEs-associated promoters were intergenic genes, followed by intragenic, and the lowest proportion of genes with TEs promoter was observed for bidirectional genes. This might suggest that TEs do generally promote bidirectional transcription initiation (see Figure 12).

Table 6. Number of TEs and non-TEs promoters for every developmental stage and every category.

Oocytes					
	Intergenic	Intragenic	Bidirectional	Overlapping	all
without TEs	976	175	274	72	1497
with 1 TEs	860	147	43	83	1133
with 2+ TEs	21	2	0	0	23
2C					
	Intergenic	Intragenic	Bidirectional	Overlapping	all
without TEs	225	71	113	35	444
with 1 TEs	153	40	7	33	233
with 2+ TEs	4	2	1	1	8
4C					
	Intergenic	Intragenic	Bidirectional	Overlapping	all
without TEs	407	142	102	69	720
with 1 TEs	283	81	12	42	418
with 2+ TEs	8	8	0	2	18
ICM					
	Intergenic	Intragenic	Bidirectional	Overlapping	all
without TEs	205	115	57	31	408
with 1 TEs	120	38	7	25	190
with 2+ TEs	3	1	0	0	4
TE					
	Intergenic	Intragenic	Bidirectional	Overlapping	all
without TEs	107	100	43	28	278
with 1 TEs	80	35	4	14	133
with 2+ TEs	0	0	0	0	0

Figure 11. Composition of different TEs classes as promoters for every developmental stage



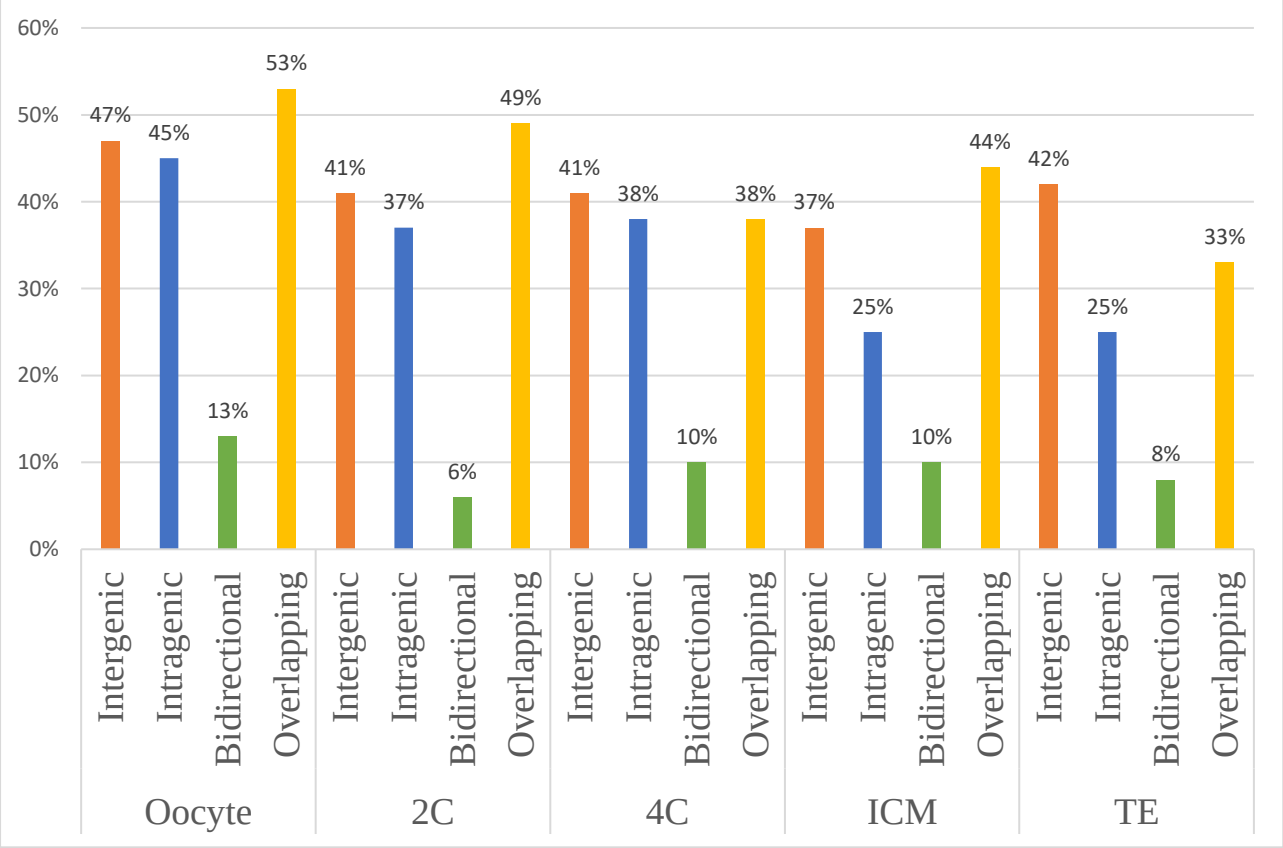


Figure 12. Percentage of TE promoters for every category

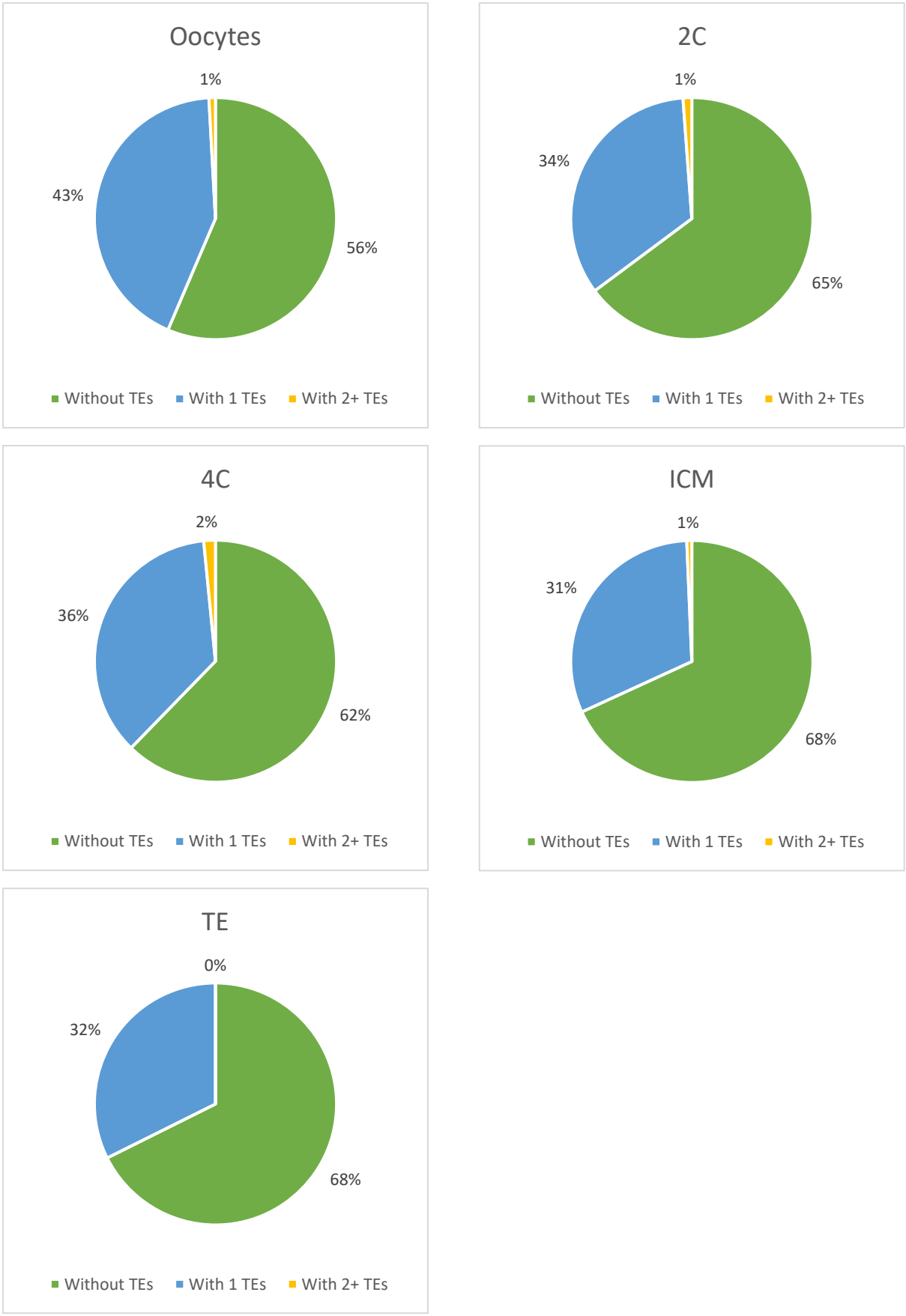


Figure 13. Proportion of promoters without any TEs, with one TEs and with more than 2 TEs for every developmental stage

6.5 Protein coding potential of high confidence novel transcripts

We wanted to analyse whether our high confidence novel genes are primarily lncRNAs, or whether some could potentially be protein coding. For this, we used Coding Potential Calculator 2 (CPC2) through its web interface. It takes sequence fasta files of RNA transcripts as an input and gives us a list of coding and non-coding RNAs as a result. To obtain the sequences of transcripts of high fidelity novel genes, we first filtered assembled merged gtf annotations to contain only transcripts belonging to the high fidelity novel genes, using a python script filtering_gtf.py (Appendix 5) generated previously in the laboratory. Then, we used another python script gettingSeq.py (Appendix 6) to extract fasta sequences of transcripts from the filtered gtf file.

After division of transcripts to their corresponding genes, the results were as expected. Majority of genes were classified as non-coding (see Figure 14) but a small number of genes with protein coding potential was identified in each category of genes in each developmental stage. The dominant category for coding transcripts and genes is intergenic category with 497 coding transcripts and 332 coding genes across all the datasets and it is also the category with most transcripts overall with 9659 transcripts. The total number of coding and non-coding transcripts was 18483 from which 1236 classified as coding transcripts (see Table 7) corresponding to 780 coding genes (see Table 8).

Looking at the results with the relative approach, we see a different pattern. The highest percentage of coding transcripts has the overlapping category with 12.15% of all its transcripts being coding, followed by bidirectional category with 7.96%. The total percentage of coding transcripts from all the transcripts is 6.69% (see Figure 15).

Table 7. Number of coding and non-coding transcripts found for every dataset and for every category

		Oocytes	**2C**	**4C**	**ICM**	**TE**
Intragenic	**Coding**	82	19	44	19	34
	Non-coding	1021	528	1098	513	433
Intergenic	**Coding**	255	69	107	40	26
	Non-coding	4372	1062	2188	978	562
Bidirectional	**Coding**	43	24	49	10	8
	Non-coding	576	317	373	191	93
Overlapping	**Coding**	157	63	90	56	41
	Non-coding	867	472	760	595	248
Total	**Coding**	537	175	290	125	109
	Non-coding	6836	2379	4419	2277	1336

Table 8. Number of coding genes for every category

	oocytes	2C	4C	ICM	TE
Intragenic	46	16	33	16	26
Intergenic	156	51	70	30	24
Bidirectional	25	16	28	9	5
Overlapping	73	36	61	33	26
Total	300	119	192	88	81

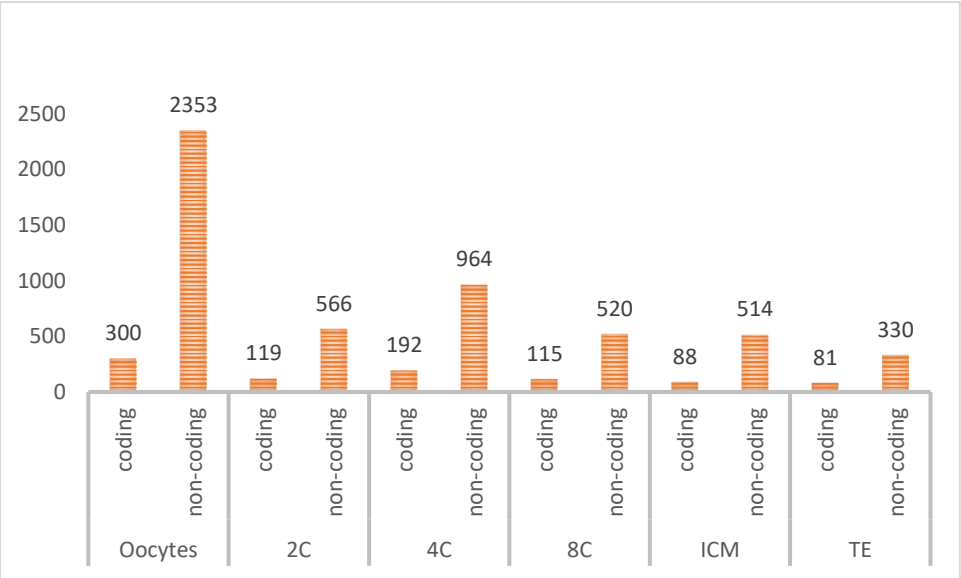


Figure 14. Number of coding and non-coding novel genes for each developmental stage

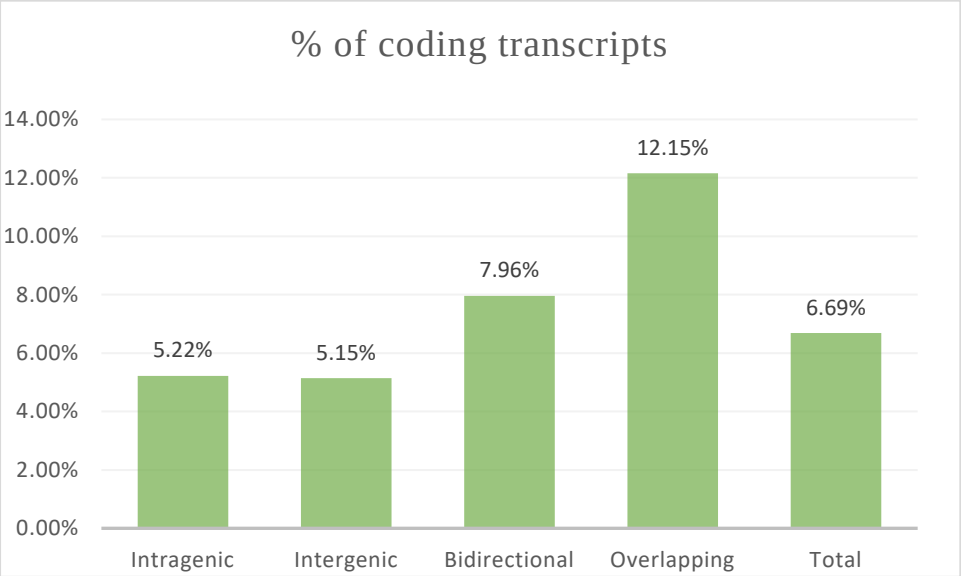


Figure 15. Percentage of coding transcripts from all the transcripts for all categories

7 Discussion

In this project, we aimed to identify novel non-annotated genes in early developmental stages of mouse, analyse their expression pattern across development, investigate which of them have TEs as a promoter and which are potentially coding proteins. In order to do this we processed publicly available RNA-seq datasets from mouse oocytes, preimplantation embryos and adult somatic tissues. We focused on oocytes, embryos at the time of major zygotic genome activation (2C and 4C stage) and first two cell lineages segregated in preimplantation embryos (ICM and TE).

We performed de novo transcriptome assembly with these datasets and obtained annotations of assembled transcriptomes. Due to having datasets with different number of reads and strand specificity, the assembled transcriptomes were of different levels of quality. Specifically, oocyte datasets had the highest number of reads with strand specificity while embryonic datasets had fewer reads with strand specificity. The differences in assemblies were observed while visually inspecting the gtf annotation files and seeing that in the embryonic assemblies there were more arrays of monooexonic genes on the same strand (or unstranded) located closely one after another which could all be part of one misannotated genes, or we could observe such monoexonic genes at the proximity of 5' or 3' ends of genes, which could potentially be just misannotated extensions of 5' and 3' UTRs of nearby genes. In addition, in embryonic datasets, there was a high number of genes without strand information which had to be excluded from the analysis. It probably resulted in higher number of assembled transcripts than there are in reality, consequently leading to higher number of novel genes. We tried to eliminate misannotated novel genes and select only high confidence genes by analysing H3K4me3 datasets and looking for association of promoters of novel genes with H3K4me3 peaks.

After analysis of the quality of the annotation of novel genes, we proceeded to examine their expression patterns. This revealed that most of the genes are specific for oocytes or preimplantation embryos and not expressed in the adult somatic tissues. Genes from 2C and 4C embryos were often specific for the early preimplantation development (until 8C or morula stages), while genes identified in ICM or TE were often expressed throughout the whole preimplantation development. The expression analysis was not as precise as we would like to because datasets being generated by different people with different library preparation

protocols. Nevertheless, we assumed that all the datasets are still similar enough for meaningful comparison.

We also analysed the TEs acting as promoters in our high confidence novel genes. The results showed that approximately 30-40% of promoters in all analysed datasets have at least one TEs acting as their promoter. The most common ones were LTR-MaLR and LTR-ERVK for oocytes, LTR-ERVL for early preimplantation embryos and LTR-ERVK for late preimplantation embryos. These results were expected as genes in oocytes and embryos were found to have transposable elements as promoters (Gifford et al. 2013) and the most frequently found TEs element as a promoters in oocyte developmental cells is LTR-MaLR (Veselovska et al. 2015). Our discovery of LTR-ERVK as the most common TEs in late preimplantation embryos and LTR-ERVL as the most common TEs for early preimplantation embryos agrees with already existing findings of LTR-ERVK acting as promoters in extra-embryonic lineages like TE and ICM (Hanna et al. 2019) and early preimplantation embryos like 2C having high number of transcripts initiated from LTRs derived from endogenous retroviruses (Macfarlan et al. 2012).

To find out how many and if any of our novel genes are potentially coding proteins we used CPC2 web interface. This revealed that only a small fraction of them have protein coding potential which was the result that we predicted as it was shown before that a majority of novel genes identified in the oocytes are lncRNAs (Veselovska et al. 2015).

There are many ways how we could improve the individual analyses in order to obtain more precise results. The resulting transcripts from de novo transcriptome assembly could have been filtered according to their size keeping only the ones which are bigger than 200bp, as library preparation protocols selected only RNAs longer than approximately 200bp and therefore shorter genes should be a result of incomplete annotation. We can focus primarily on multiexonic genes as those might be more interesting than monoexonic genes. For the analysis of relative expression of genes, we can go more in depth with the bidirectional category genes and see whether there is some correlation between the expression of those genes and the genes with which they share their promoter. We also used only one web interface program for defining whether novel genes have protein coding potential. The use of several different programs would be more reliable. For the genes that were classified as protein coding we could further investigate if they contain a known protein domain which would further increase the likelihood of the gene being truly protein coding.

This research helped us to further explore and understand the genome of mouse early developmental stages. It acts as a good base for both more bioinformatical analyses of mouse oocyte and embryo novel genes as well as for experimental functional analysis of selected novel genes. There are many ways how this research can further continue. One of them can be analysing and comparing how these novel genes are conserved across mammalian species. In addition, we can analyse the role of transposable elements as promoters of novel genes in other mammalian species. Experimentally, functional analysis can be performed by downregulation or knockout of selected candidate genes to explore their roles in development. Candidate genes interesting for future studies might be either those which are specific only for certain developmental stage, for example oocyte or 2C embryos or for example those, when comparing ICM and TE genes, that are only in one of these two cell lineages. In addition, candidates for experimental functional testing can be selected based on their protein coding potential or high expression level in specific developmental stage or stages. By studying these genes, we can uncover their roles in oocyte and embryonic development.

8 Conclusion

Identifying and characterising novel transcripts in early developmental stages is relatively technically difficult task. Because of that there are still thousands of non-annotated transcripts and genes with unknown functions. Our research helped to fill in this gap by identifying and characterising a number of novel transcripts and genes in early developmental stages of a mouse.

We observed the decrease of stage-specificity of the expression profile with developmental progression and examined the TEs acting as the promoters which gave us an insight into which classes of TEs are the most common for certain developmental stages. We saw that only a few of our novel genes have protein coding potential and therefore the majority of novel genes are lncRNAs.

This research will serve as a basis for future more detailed analyses of oocyte and embryo transcripts and genes of a mouse, better understanding of transposable elements and their role as the promoters as well as experimental functional characterisation of candidate novel genes with potential roles during mouse early development.

9 References

Alwine JC, Kemp DJ, Stark GR, Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes, (1977), PNAS 74, 5350–5354.

Andergassen D, Dotter CP, Wenzel D, Sigl V, Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression, (2017), Elife 14, 6

Bennetzen JL, Transposable element contributions to plant genome evolution, (2000), Plant Mol Biol. 42, 251-269

Bernstein, B.E, et al, (2012), An integrated encyclopedia of DNA

Biemont Ch, A brief history of the Status of Transposable elements: From junk DNA to major players in the evolution, (2010), Genetics vol 186

Bouckenheimer J, Assou S, Riquier S, Hou C, et al, Long non-coding RNAs in human early embryonic development and their potential in ART, (2016), Human Repr Update vol 23, 19-40

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu Et, Evolution of the mammalian transcription factor binding repertoire via transposable elements, (2008), Genome Res 18(11), 1752–1762

Cabilli MN, Trapnell C, Goff L, Koziol M, Vega B, Regel A, Rinn J, Integrative annotation of human large intergenic non coding RNAs reveals global properties and specific subclasses, (2011), Genes Dev 25(18), 1915-27

Casacuberta E, González J, The impact of transposable elements in environmental adaptation, (2013), Mol Ecol 22(6), 1503–1517

Chen J, Greenblatt IM, Dellaporta SL, Molecular analysis of Ac transposition and DNA replication, (1992), Genetics 130, 665-676

Chénais B, Transposable elements and human cancer: a causal relationship?, (2013),Biochim Biophys Acta, Rev Cancer 1835(1), 28–35

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, Rocchi M, Eichler EE, A genome-wide comparison of recent chimpanzee and human segmental duplications, (2005), Nature 437(7055), 88–93

Chuong EB, Elde NC, Feschotte C, Regulatory activities of transposable elements: from conflicts to benefits, (2017), Nat Rev Genet 18,71

Cowley M, Oakey RJ, Transposable elements re-wire and fine-tune the transcriptome, (2013), PLoS Genet 9(1), e1003234

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al, The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression, (2012), Genome Res 22,1775–1789

Feschotte C, Pritham EJ, DNA transposons and the evolution of eukaryotic genomes, (2007), Annu Rev Genet 41(1), 331–368

Feschotte C, The contribution of transposable elements to the evolution of regulatory networks, (2008), Nat Rev Genet 9(5), 397–405

Finnegan, D.J, Eukaryotic transposable elements and genome evolution, (1989), Trends Genet 5, 103–107

Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J, Svobodova E, Ma J, Suzuki Y, Aoki F, Schultz RM, Vlahovicek K, Svoboda P, Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes, (2017), Genome Res 27, 1384–1394

Ganesh S, Retrotransposon-associated long non-coding RNAs in mice and men genome, (2016), Nature 409, 860–921

Gifford Wesley D, Pfaff S, Macfarlan T, Transposable elements as genetic regulatory substrates in early development, (2013), Trends in Cell Biol 23(5), 218-226

Gomes, C.P.C, Spencer H, Ford, K.L, Michel, L.Y.M, Baker, A.H, Emanueli C, Balligand, J. -L, Devaux Y, The function and therapeutic potential of long noncoding RNAs in cardiovascular development and disease, (2017), Mol. Ther. Nucleic acids 8, 494-507

Grimaldi G, Skowronski J, Singer MF, Defining the beginning and end of KpnI family segments, (1984), EMBO J vol 3, 1753-1759

Hanna CW, Pérez-Palacios R, Gahurova L, Schubert M et al, Endogenous retroviral insertions drive non-canonical imprinting in extra-embryonic tissues, (2019), Genome Biol 20(1), 225

Hanna, C. W, Demond, H, Kelsey, G. Epigenetic regulation in development: Is the mouse a good model for the human?, (2018),Human Reproduction Update 24(5), 556–576

Hemalatha G, Raoult D, Pontarotti P, The rhizome of life: what about metazoa?, (2012), Front Cell Infect Microbiol 2, 50

Huynh KD, Lee JT, Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos, (2003), Nature 426 (6968), 857-862

Jordan IK, Rogozin IB, Glazko GV, Koonin EV, Origin of a substantial fraction of human regulatory sequences from transposable elements, (2003), Trends Genet 19, 68–72

Kang YJ, Yang DCH, Kong L, Hou M, Meng Y, Wei L, Gao G, CPC2:a fast and accurate coding potential calculator based on sequence intrinsic features, (2017), Nucleic Acids Res 45(W1), W12-W16

Kapitonov VV, Jurka J, Rolling-circle transposons in eukaryotes, (2001), P Natl Acad Sci USA 98(15), 8714–8719

Kim, D, Paggi, J.M, Park, C, et al, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, (2019), Nat Biotechnol 37, 907–915

Kouzadires T, Chromatin modification and their function, (2007), Cell 128(4), 693-705

Lander, E.S, et al, Initial sequencing and analysis of the human, (2001), Nature 409, 860-921

Langmead B, Salzberg S, Fast-gapped read alignment with Bowtie2, (2012), Nat Methods 9(4), 357–359

Liu X, Wang C, Liu W, Li J et al, Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos, (2016), Nature 537(7621), 558-562

Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL, Embryonic stem cell potency fluctuates with endogenous retrovirus activity, (2012), Nature 487(7405), 57-63

Maksakova IA, Romanish MT, Gagnier L, Dunn CA, Van de Lagemaat LN, Mager DL, Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line, (2006), PLoS Genet 2(1), e2

Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK, Transposable elements donate lineage-specific regulatory sequences to host genomes, (2005), Cytogenet Genome Res 110, 333–41

Marques-Bonet T, Girirajan S, Eichler EE, The origins and impact of primate segmental duplications, (2009), Trends Genet 25(10), 443–454

Mercer T. R, Mattick J. S, Structure and function of long noncoding RNAs in epigenetic regulation, (2013), Nat Struct Mol Biol 20, 300–307

Mi S, Lee X, Xiang-ping L, Veldman GM, Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis, (2000), Nature 403(6771), 785–789

Mikkelsen TS, Wakefield MJ et al, Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences, (2007), Nature 447(7141), 167–177

Morse B, Rotherg PG, South VJ, Spandorfer JM, Astrin SM, Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma, (1988), Nature 333(6168), 87–90

Mullis KB, Target amplification for DNA analysis by the polymerase chain reaction, (1990), Ann Biol Clin (Paris) 48, 579–582.

Ohno S, Evolution by gene duplication, (1970), New York: Springer; 1970.

Okamoto I, Arnaud D, Le Baccon P, Otte AP, Disteche CM, Avner P, Heard E, Evidence for de novo imprinted X-chromosome inactivation independent of meiotic inactivation in mice, (2005), Nature 438 (7066), 369-373

Oliver KR, Greene WK, Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates, (2011), Mob DNA 2(1), 8

Ostertag EM, Kazazian HH, Biology of mammalian L1 retrotransposons, (2001), Annu Rev Genet 35(1), 501–538

Platt R, Vanderwege M, Ray D, Mammalian transposable elements and their impact on genome evolution, (2018), Chromosome Res 26(1), 25-43

Platt RN, II, Ray DA, A non-LTR retroelement extinction in *Spermophilus tridecemlineatus*, (2012), Gene 500(1), 47–53

Ponjavic J, Ponting C. P, Lunter G, Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs, (2007), *Genome Res.* 17, 556–565

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al, Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression, (2011), Nat Biotechnol 29, 742–9

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM, Long non-coding RNAs as a source of new peptides, (2014), eLife 3, e03523

Santos-Rosa et al, Active genes are trimethylated at K4 of histone H3, (2002), Nature 419, 407-411

Uchida S, Dimmeler S, Long noncoding RNAs in cardiovascular diseases, (2015), *Circ Res* 116, 737–750

Ulitsky I, Bartel DP, lincRNAs: genomics, evolution, and mechanisms, (2013), Cell 154, 26–46

Vance, K.V, Ponting, C.P, Transcriptional regulatoryfunctions of nuclear long noncoding RNA, (2014), Trends Genet 30, 348-355

Veselovska L, Smallwood S, Saadeh H, Steward K, Krueger F, Maupetit-Mehouas S, Arnaud P, Tomizawa S, Andrews S, Kelsey G, Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape, (2015), Genome Biology 16, 209

Wang C, Liu X, Gao Y, Yang L, et al, Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development, (2018), Nat Cell Biol 20(5), 620-631

Wang, Xue, Q.D, Crutchley, Jennifer L, Dostie, Jose, Shaping the genome with non-coding RNAs, (2011), Genomics 12, 307-321

Warner JR, Soeiro R, Birnboim HC, Girard M, Darnell JE, Rapidly labeled HeLa cell nuclear RNA I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA, (1966), J Mol Biol 19, 349–361

Waterston RH, Pachter L, Initial sequencing and comparative analysis of the mouse genome, (2002), Nature 420(6915), 520–562

Wei J et al, Non-coding RNAs as regulator in epigenetics, (2017), oncol Rep 37, 3-9

Wicker T, Sabot F, Hua-van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH, A unified classification system for eukaryotic transposable elements, (2007), Nat Rev Genet 8(12), 973–982

Wu, F, Liu, Y, Wu, Q, et al, Long non-coding RNAs potentially function synergistically in the cellular reprogramming of SCNT embryos, (2018), BMC Genomics 19, 631

Wutz A, Rasmussen T. P, Jaenisch R, Chromosomal silencing and localization are mediated by different domains of Xist RNA, (2002), Nat Genet 30, 167–174

Zampetaki A, Albrecht A, Steinhofel K, Long non-coding RNA structure and function. Is there a link?, (2018), Front Physiol 9, 1201

Zeh DW, Zeh JA, Ishida Y, Transposable elements and an epigenetic basis for punctuated equilibria, (2009), BioEssays 31(7), 715–726

Zhang, K, Huang, K, Luo, Y, et al, Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data, (2014), BMC Genomics 15, 845

Zheng H, Zhang B, Huang B, Li W, et al, Allelic reprogramming of the histone modification H3K4me3 in early mammalian development, (2016), Nature 537(7621), 553-557

10 Appendix

Appendix 1 - Complete list of datasets used

Appendix 2 - Example script for transcriptome assembly

Appendix 3 - Example script for expression quantification

Appendix 4 - The script for generating heatmaps in rStudio

Appendix 5 – Python script filtering_gtf.py

Appendix 6 – Python script gettingSeq.py

Appendix 7 - Heatmaps for all datasets and all 4 categories of gene locations, except heatmaps found in Figure 3 and Figure 4

Appendix 8 - Number of genes found in each cluster for all the datasets and categories except those found in Table 5

appendix 9 – Relative expression of individual clusters for all datasets and all categories of gene location, except those found in Figures 5-9

Appendix 1. Complete list of datasets used

Publication	Cell type	Accession number	Library Type	Type
Veselovska et al. (2015)	d5 oocytes	GSE70116	fr-firststrand	RNA-seq
	d10 oocytes		fr-secondstrand	
	d15 oocytes		fr-secondstrand	
	GV oocytes		fr-firststrand	
Wang et al. (2018)	2C embryo	GSE98150	fr-unstranded	
	4C embryo		fr-unstranded	
	8C embryo		fr-unstranded	
	morula embryo		fr-unstranded	
	Inner cell mass-ICM		fr-unstranded	
	Trophectoderm-TE		fr-unstranded	
Andergassen et al. (2017)	Adult_brain	GSE75957		
	Adult_leg_muscle			
	Adult_liver			
	Adult_spleen			
	Adult_thymus			
	Adult_lung			
	Adult_heart			
Zheng et al. (2016)	d10 oocytes	GSE71434		ChiP-seq
	2C late			
	4C embryo			
	ICM			
Liu et al. (2016)	2C embryo	GSE73952		
	4C embryo			
	ICM			
	TE			
Hanna et al. (2018)	d10 oocytes	GSE93941		
	d15 oocytes			

Appendix 2. Example script for transcriptome assembly

```
#!/bin/bash

#PBS -N Wang_2C_rep3_rep4_cufflinks

#PBS -l walltime=48:00:00

#PBS -l select=1:ncpus=8:mem=100gb:scratch_local=50gb

#PBS -m abe

#PBS -M karolina.kravarikova@gmail.com

DATADIR="/storage/plzen1/home/kravak01/Wang_2018"

cp $DATADIR/Wang_2C_rep3-sorted.bam $SCRATCHDIR/
cp $DATADIR/Wang_2C_rep4-sorted.bam $SCRATCHDIR/
cp $DATADIR/Mus_musculus.GRCm38.94.chr.gtf $SCRATCHDIR/


module add cufflinks-2.2.1
module add samtools-1.3.1



cd $SCRATCHDIR

cufflinks -g Mus_musculus.GRCm38.94.chr.gtf -u --library-type fr-unstranded -o Wang_2C_rep3a
Wang_2C_rep3-sorted.bam
cufflinks -g Mus_musculus.GRCm38.94.chr.gtf -u --library-type fr-unstranded -o Wang_2C_rep4a
Wang_2C_rep4-sorted.bam




cp $SCRATCHDIR/Wang_2C_rep3a/genes.fpkm_tracking  $DATADIR/Wang_2C_rep3a
cp $SCRATCHDIR/Wang_2C_rep3a/isoforms.fpkm_tracking  $DATADIR/Wang_2C_rep3a
cp $SCRATCHDIR/Wang_2C_rep3a/skipped.gtf  $DATADIR/Wang_2C_rep3a
cp $SCRATCHDIR/Wang_2C_rep3a/transcripts.gtf  $DATADIR/Wang_2C_rep3a
cp $SCRATCHDIR/Wang_2C_rep4a/genes.fpkm_tracking  $DATADIR/Wang_2C_rep4a
cp $SCRATCHDIR/Wang_2C_rep4a/isoforms.fpkm_tracking  $DATADIR/Wang_2C_rep4a
cp $SCRATCHDIR/Wang_2C_rep4a/skipped.gtf  $DATADIR/Wang_2C_rep4a
cp $SCRATCHDIR/Wang_2C_rep4a/transcripts.gtf  $DATADIR/Wang_2C_rep4a
```

Appendix 3. Example script for expression quantification

```
#!/bin/bash

#PBS -N Wang_4C_rep2_quantification

#PBS -l walltime=48:00:00

#PBS -l select=1:ncpus=8:mem=200gb:scratch_local=200gb

#PBS -m abe

#PBS -M karolina.kravarikova@gmail.com

DATADIR="/storage/plzen1/home/kravak01/Quantification"

cp $DATADIR/Wang_4C_rep2-sorted.bam $SCRATCHDIR/
cp $DATADIR/oocytes_merged.gtf $SCRATCHDIR/
cp $DATADIR/2C_merged.gtf $SCRATCHDIR/
cp $DATADIR/4C_merged.gtf $SCRATCHDIR/
cp $DATADIR/8C_merged.gtf $SCRATCHDIR/
cp $DATADIR/morula_merged.gtf $SCRATCHDIR/
cp $DATADIR/ICM_merged.gtf $SCRATCHDIR/
cp $DATADIR/TE_merged.gtf $SCRATCHDIR/
cp $DATADIR/embryos_merged.gtf $SCRATCHDIR/




module add cufflinks-2.2.1
module add samtools-1.3.1
cd $SCRATCHDIR
cufflinks -G oocytes_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_oocytesQ Wang_4C_rep2-
sorted.bam
cufflinks -G 2C_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_2CQ Wang_4C_rep2-sorted.bam
cufflinks -G 4C_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_4CQ Wang_4C_rep2-sorted.bam
cufflinks -G 8C_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_8CQ Wang_4C_rep2-sorted.bam
cufflinks -G morula_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_morulaQ Wang_4C_rep2-
sorted.bam
cufflinks -G TE_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_TEQ Wang_4C_rep2-sorted.bam
cufflinks -G ICM_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_ICMQ Wang_4C_rep2-
sorted.bam
cufflinks -G embryos_merged.gtf -u --library-type fr-unstranded -o Wang_4C_rep2_embryosQ Wang_4C_rep2-
sorted.bam


cp $SCRATCHDIR/Wang_4C_rep2_oocytesQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_oocytesQ
cp $SCRATCHDIR/Wang_4C_rep2_oocytesQ/isoforms.fpkm_tracking
$DATADIR/Wang_4C_rep2_oocytesQ
cp $SCRATCHDIR/Wang_4C_rep2_oocytesQ/skipped.gtf  $DATADIR/Wang_4C_rep2_oocytesQ
cp $SCRATCHDIR/Wang_4C_rep2_oocytesQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_oocytesQ
cp $SCRATCHDIR/Wang_4C_rep2_2CQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_2CQ
cp $SCRATCHDIR/Wang_4C_rep2_2CQ/isoforms.fpkm_tracking  $DATADIR/Wang_4C_rep2_2CQ
cp $SCRATCHDIR/Wang_4C_rep2_2CQ/skipped.gtf  $DATADIR/Wang_4C_rep2_2CQ
cp $SCRATCHDIR/Wang_4C_rep2_2CQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_2CQ
cp $SCRATCHDIR/Wang_4C_rep2_4CQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_4CQ
```

```
cp $SCRATCHDIR/Wang_4C_rep2_4CQ/isoforms.fpkm_tracking  $DATADIR/Wang_4C_rep2_4CQ
cp $SCRATCHDIR/Wang_4C_rep2_4CQ/skipped.gtf  $DATADIR/Wang_4C_rep2_4CQ
cp $SCRATCHDIR/Wang_4C_rep2_4CQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_4CQ
cp $SCRATCHDIR/Wang_4C_rep2_8CQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_8CQ
cp $SCRATCHDIR/Wang_4C_rep2_8CQ/isoforms.fpkm_tracking  $DATADIR/Wang_4C_rep2_8CQ
cp $SCRATCHDIR/Wang_4C_rep2_8CQ/skipped.gtf  $DATADIR/Wang_4C_rep2_8CQ
cp $SCRATCHDIR/Wang_4C_rep2_8CQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_8CQ
cp $SCRATCHDIR/Wang_4C_rep2_morulaQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_morulaQ
cp $SCRATCHDIR/Wang_4C_rep2_morulaQ/isoforms.fpkm_tracking  $DATADIR/Wang_4C_rep2_morulaQ
cp $SCRATCHDIR/Wang_4C_rep2_morulaQ/skipped.gtf  $DATADIR/Wang_4C_rep2_morulaQ
cp $SCRATCHDIR/Wang_4C_rep2_morulaQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_morulaQ
cp $SCRATCHDIR/Wang_4C_rep2_TEQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_TEQ
cp $SCRATCHDIR/Wang_4C_rep2_TEQ/isoforms.fpkm_tracking  $DATADIR/Wang_4C_rep2_TEQ
cp $SCRATCHDIR/Wang_4C_rep2_TEQ/skipped.gtf  $DATADIR/Wang_4C_rep2_TEQ
cp $SCRATCHDIR/Wang_4C_rep2_TEQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_TEQ
cp $SCRATCHDIR/Wang_4C_rep2_ICMQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_ICMQ
cp $SCRATCHDIR/Wang_4C_rep2_ICMQ/isoforms.fpkm_tracking  $DATADIR/Wang_4C_rep2_ICMQ
cp $SCRATCHDIR/Wang_4C_rep2_ICMQ/skipped.gtf  $DATADIR/Wang_4C_rep2_ICMQ
cp $SCRATCHDIR/Wang_4C_rep2_ICMQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_ICMQ
cp $SCRATCHDIR/Wang_4C_rep2_embryosQ/genes.fpkm_tracking  $DATADIR/Wang_4C_rep2_embryosQ
cp $SCRATCHDIR/Wang_4C_rep2_embryosQ/isoforms.fpkm_tracking
$DATADIR/Wang_4C_rep2_embryosQ
cp $SCRATCHDIR/Wang_4C_rep2_embryosQ/skipped.gtf  $DATADIR/Wang_4C_rep2_embryosQ
cp $SCRATCHDIR/Wang_4C_rep2_embryosQ/transcripts.gtf  $DATADIR/Wang_4C_rep2_embryosQ
```

Appendix 4. Script for generating heatmaps in R Studio

```
library (gplots)

data &lt;- read.delim (&quot;ICM_bidirectional_relative_value.txt&quot;)

rnames &lt;- data[,1]

mat_data &lt;- data.matrix(data[,2:ncol(data)])

rownames(mat_data) &lt;- rnames

hr &lt;- hclust(as.dist(1-cor(t(mat_data), method=&quot;pearson&quot;)), method=&quot;complete&quot;)

colorRampPalette(c(&quot;blue&quot;,&quot;yellow&quot;)) -&gt; colour.gradient

heatmap.2(mat_data, col=colour.gradient, breaks=seq(from=-5,to=5, by=0.001),

Rowv=as.dendrogram(hr), Colv=FALSE, scale=&quot;none&quot;, dendrogram=&quot;none&quot;, key=T,
keysize=0.5,

density.info=&quot;none&quot;, hclust=function(x) hclust(x,method=&quot;complete&quot;),d
istfun=function(x) as.dist((1-

cor(t(x)))/2), trace=&quot;none&quot;,cexCol=1.0, labRow=NA)

data$clusternumber &lt;- cutree (hr, 4)

write.table(data, &quot;KK_ICM_bidirectional_clusters.txt&quot;)
```

Appendix 5. Python script filtering_gtf.py

Language: Python

Description: script filtering gtf files to contain only transcripts within specified regions based on chromosome and start and end coordinates of the bases

Input: raw genomic sequence of an organism, gtf annotation that we want to filter, list of regions with genomic coordinates in which we want to retain the transcripts

Authors: Silvia Ramirez, Nikolas Tolar

```
import re

chom_start_end_file = "TE_intergenic_coordinates_for_gtf_filter.txt" #no header, 3 columns -
choromosome,start,end (separated by tabs)
input_filename = "TE_merged.gtf"

chromosomes = []
bases = []

feed_file = open(chom_start_end_file,'r')
line = feed_file.readline()

while line != '':
    line_split = line.split('\t')
    chromosomes.append(line_split[0])
    bases.append([int(line_split[1]),int(line_split[2])])
    line = feed_file.readline()

# creates output file name: input_filename + filtered.gtf
output_filename = input_filename[:input_filename.rfind(".")] + "_exon_filtered.gtf"

# opens the input file
with open(input_filename) as f:

    # reads all lines
    lines = f.readlines()

# closes input file
f.close()

# gets number of lines (used for progress)
count_lines = len(lines)

# initializes counter to 0 (used for progress)
counter = 0

# counter findings
findings = 0
```

```
# opens output file
of = open(output_filename, "w")

startAt_history = {}

def indexFrom(input_data, search_for, startAt):
    for i in range(startAt, len(input_data)):
        if input_data[i] == search_for:
            return i

def geneids_in_region():
    print("Initializing...")
    global counter, findings, startAt_history

    # if one transcript is within the region => set it to true
    for l in lines:
        counter += 1

        # splits line by tab and creates an array

        l_data = re.split(r'\t+', str(l))
        if l_data[2] == "exon":

            # checks if same chromosome (string)
            if l_data[0] in chromosomes:
                startAt_history[l_data[0]] = 0

                for x in chromosomes:
                    if x == l_data[0]:
                        startAt = 0
                        if l_data[0] in startAt_history:
                            startAt = startAt_history[l_data[0]]
                        index = indexFrom(chromosomes, l_data[0], startAt)
                        startAt_history[l_data[0]] = index + 1
                        b = bases[index]
                        l_start_base = b[0]
                        l_end_base = b[1]

                        # checks start position
                        if l_start_base <= int(l_data[3]) <= l_end_base:

                            # checks end position
                            if l_start_base <= int(l_data[4]) <= l_end_base:
                                of.write(str(l))
                                findings += 1

geneids_in_region()

# closes the output file
of.close()
# prints the output file
print("Output file: " + output_filename)
```

Appendix 6. Python script gettingSeq.py

Language: Python

Description: script generating sequences of transcripts from input gtf file in fasta format

Input: raw genomic sequence of an organism, gtf file with transcripts annotation (in our case filtered gtf file generated by script filtering_gtf.py) , list of names of transcripts from gtf file for which we want the sequences

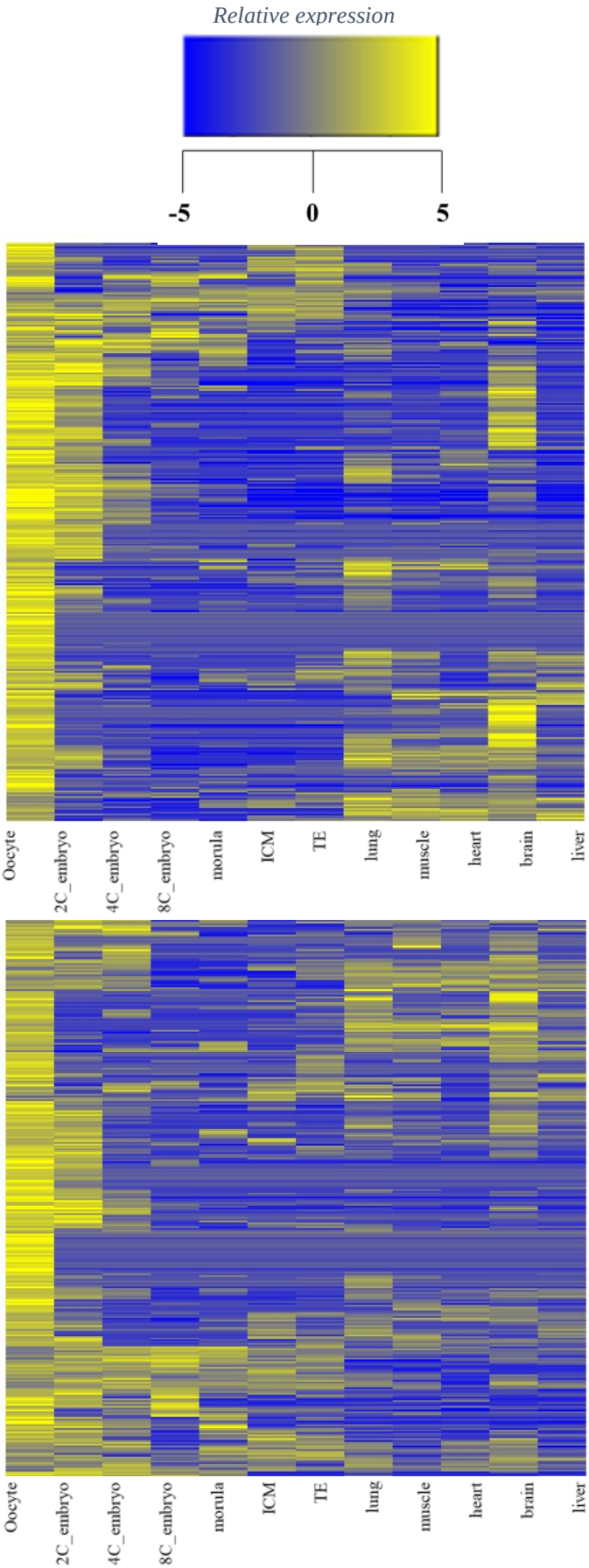
Author: Nikolas Tolar

```
genes_name = 'mus.fa'
annotation_name = 'oocytes_merged_intragenic.gtf'
output = open('oocytes_intragenic_sequence.txt','a')
query = open('oocyte_intragenic_names.txt')
merge = 1
'''
    genes_name = files containing raw DNA sequence - file names should follow the
                 pattern Xiiii where X is number/letter of chromosome and
                 iiii is the actual name that is common to all other files.
                 Variable genes_name holds the part iiii that is common
    annotation = file containing locations of exons,transcripts, etc. (.GTF file)
    output_file = name of the file the results will save into (if existing then results will append, otherwise new file
will be created)
    transcript_name = name of target transcript
    output_header = header of output file (FASTA format)
    merge = 1 means that the exons will be merged (connected) together
            0 means that the exons will be separated
'''
def caller(value,neg,k=0):
    ret = ''
    if neg == 0:
        ret = ret + '_positive_strand'
    else:
        ret = ret + '_negative_strand'
    if value == 1:
        ret = ret + '_exons_merged\n'
    elif value == 0:
        ret = ret + '_exon_' + str(k) + '\n'
    return ret

def translate_read_back(string):

    string_new = string[len(string)-1:0:-1] + string[0]
    string_new = string_new.replace('A','R')
    string_new = string_new.replace('T','A')
    string_new = string_new.replace('R','T')
    string_new = string_new.replace('C','F')
    string_new = string_new.replace('G','C')
    string_new = string_new.replace('F','G')
```

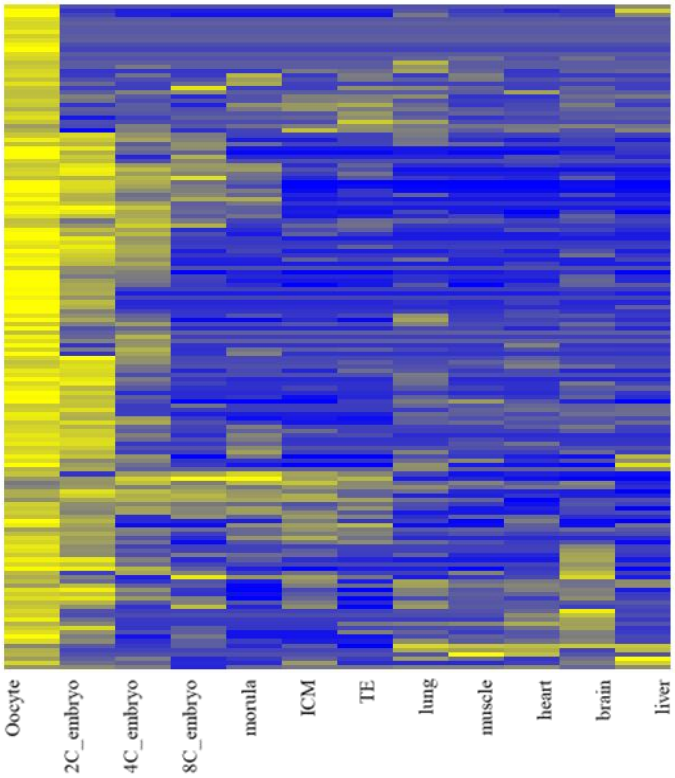
```
    return string_new

def data_extraction(text, gene_pool):

    start = int(text[3])
    stop = int(text[4])
    segment = gene_pool[start-1:stop]
    return segment

def insert_newlines(string, every=60):
    lines = []
    for i in range(0, len(string), every):
        lines.append(string[i:i+every])
    ret = '\n'.join(lines)
    return ret

def get_exons(genes_name, annotation_name, query, merge):

    transcript_name = query.readline().strip('\n')
    while transcript_name != '':

        annotation = open(annotation_name)
        neg = 0
        res_exons = ''
        res_list = []
        while True:


            text = annotation.readline()
            if text == '':
                break
            if transcript_name in text and 'exon' in text:
                text = text.split()
# accesing correct chromosome file
                genes = open(text[0]+genes_name)
                genes.readline()
                gene_pool = genes.read()
                gene_pool = ''.join(gene_pool.split())
                genes.close()

                if text[6] == '-':
                    neg = 1

                if merge == 1:
                    res_exons = res_exons + data_extraction(text,gene_pool)

                elif merge == 0:

                    res_list.append(data_extraction(text,gene_pool))

        if merge == 1:
            if neg == 1:
                res_exons = translate_read_back(res_exons)

            res_exons = insert_newlines(res_exons)

            message = caller(merge,neg)

            print('>_' + transcript_name + message + res_exons + '\n')
            output.write('>_' + transcript_name + message + res_exons + '\n\n')
```

```
        else:
            for n in range(len(res_list)):
                message = caller(merge,neg,n)

                if neg == 1:
                    res = '>_' + transcript_name + message + insert_newlines(translate_read_back(res_list[n]))

                else:
                    res = '>_' + transcript_name + message + insert_newlines(res_list[n])

                print(res + '\n')
                output.write(res + '\n\n')
        annotation.close()
        transcript_name = query.readline().strip('\n')
get_exons(genes_name, annotation_name, query, merge)
output.close()
query.close()
```

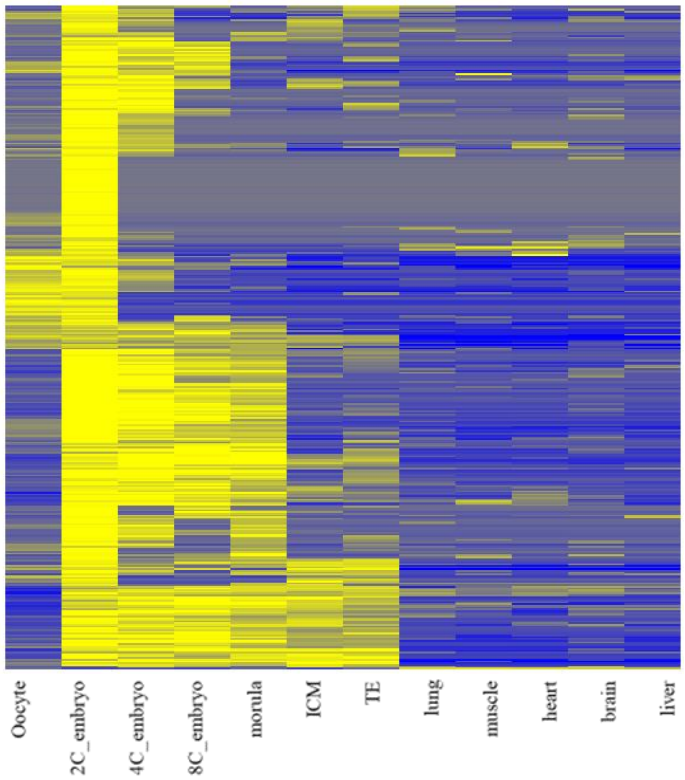
Appendix 7. Heatmaps for all datasets and all 4 categories of gene location



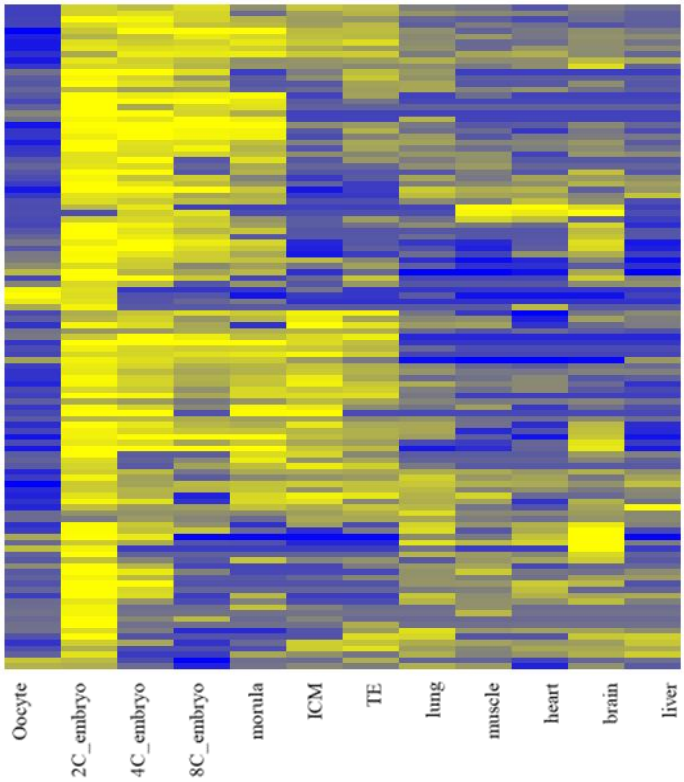
Oocyte overlapping



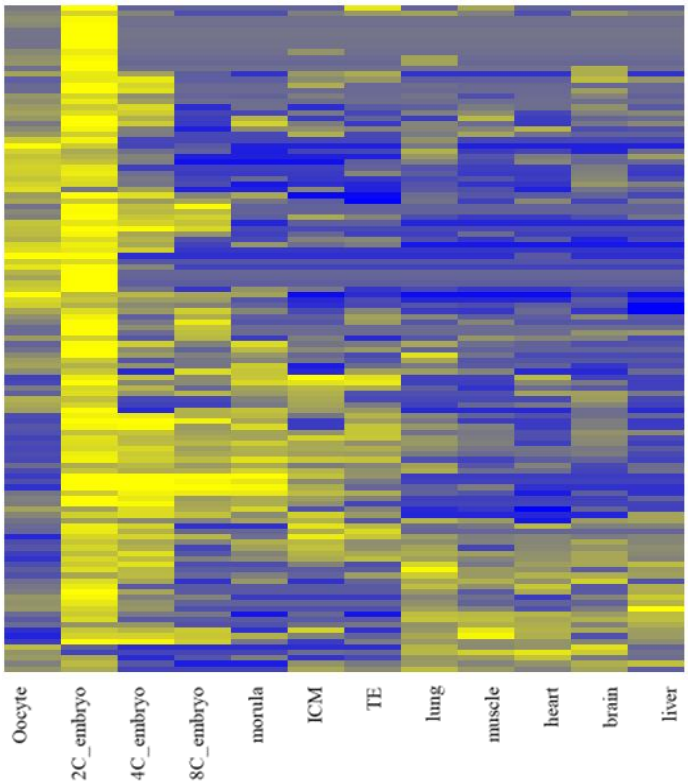
2C intergenic



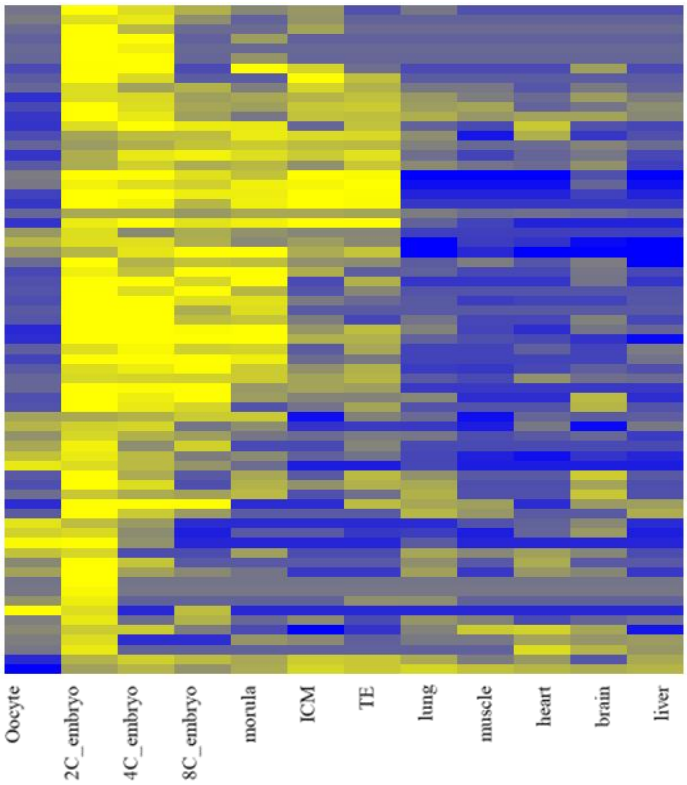
2C intragenic



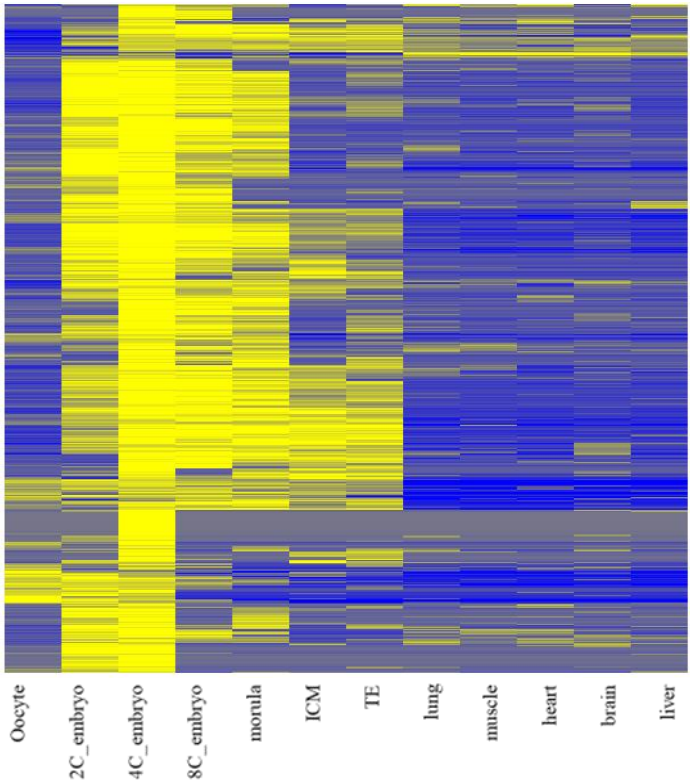
2C bidirectional



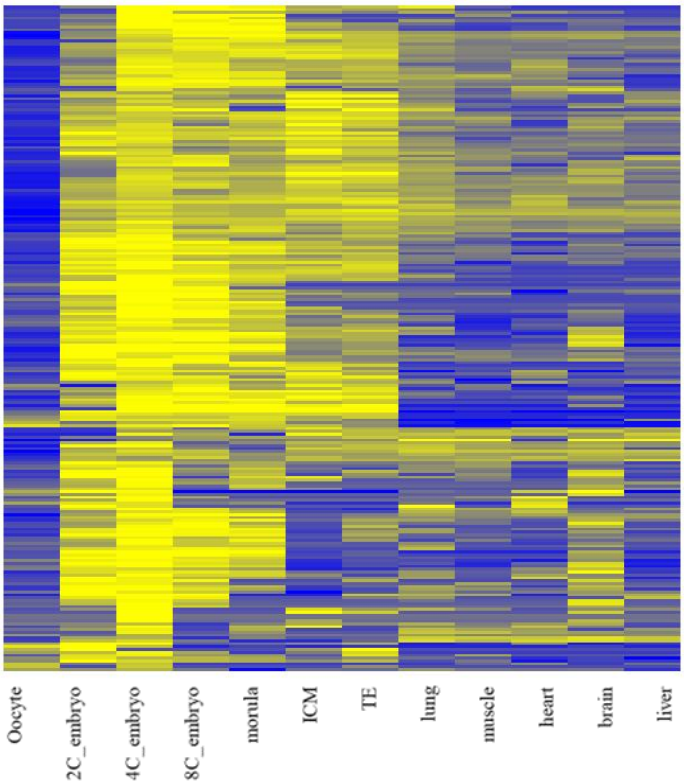
2C overlapping



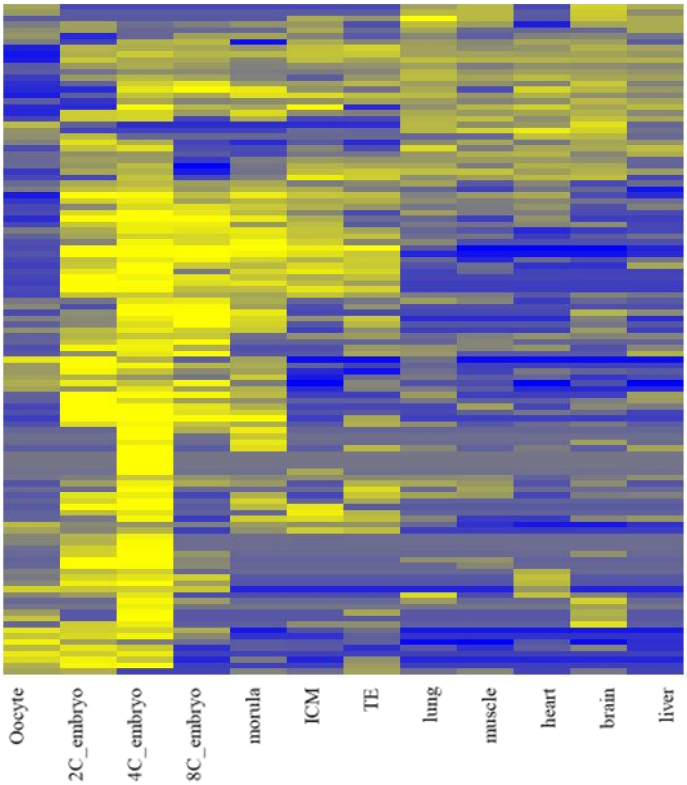
4C intergenic



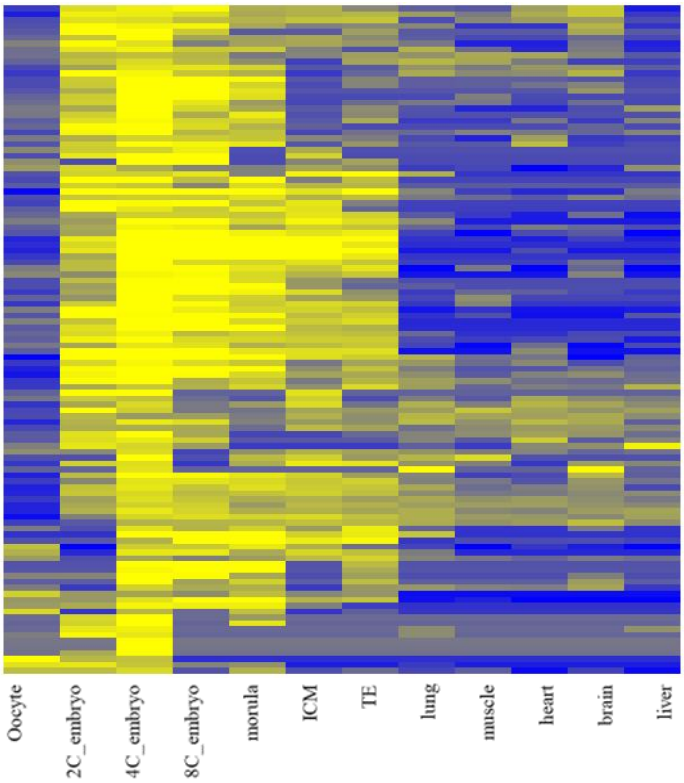
4C intragenic



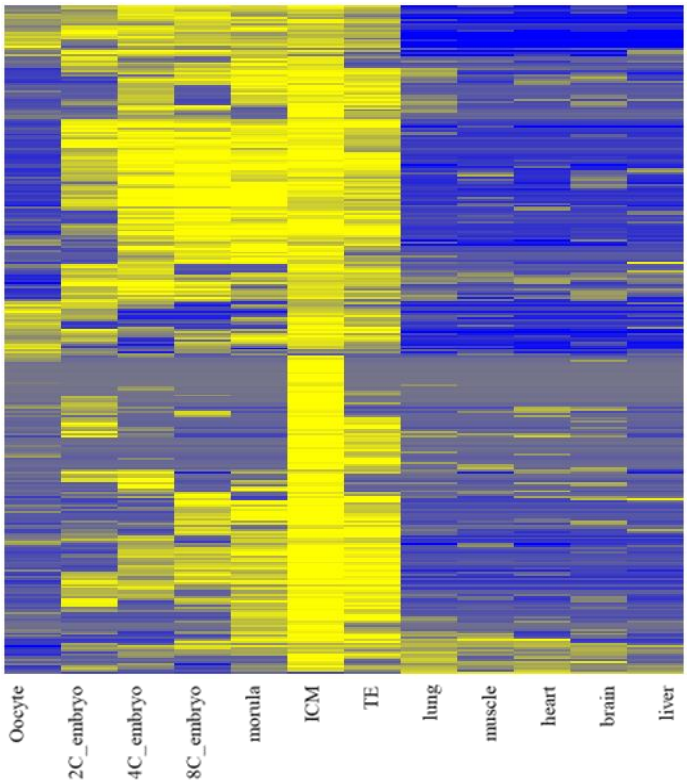
4C bidirectional



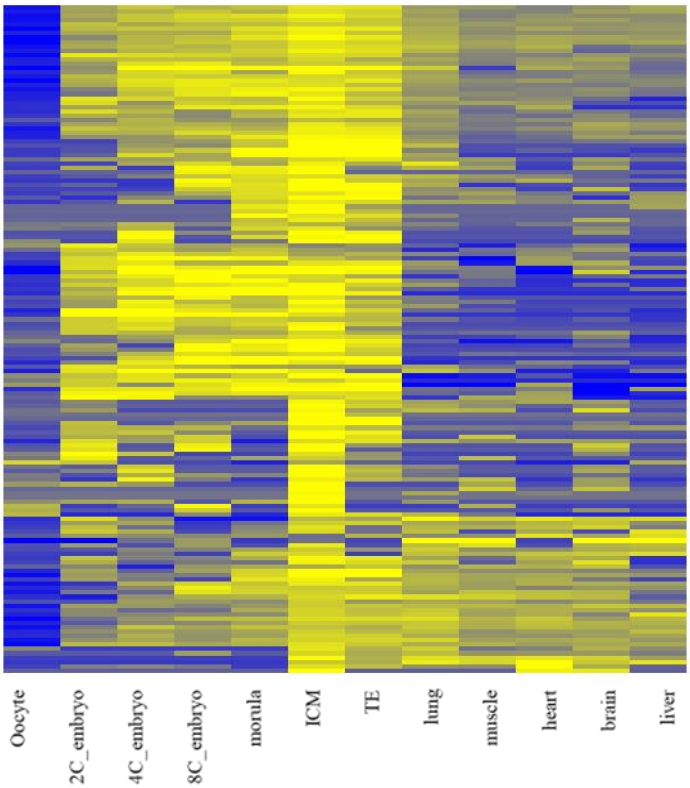
4C overlapping



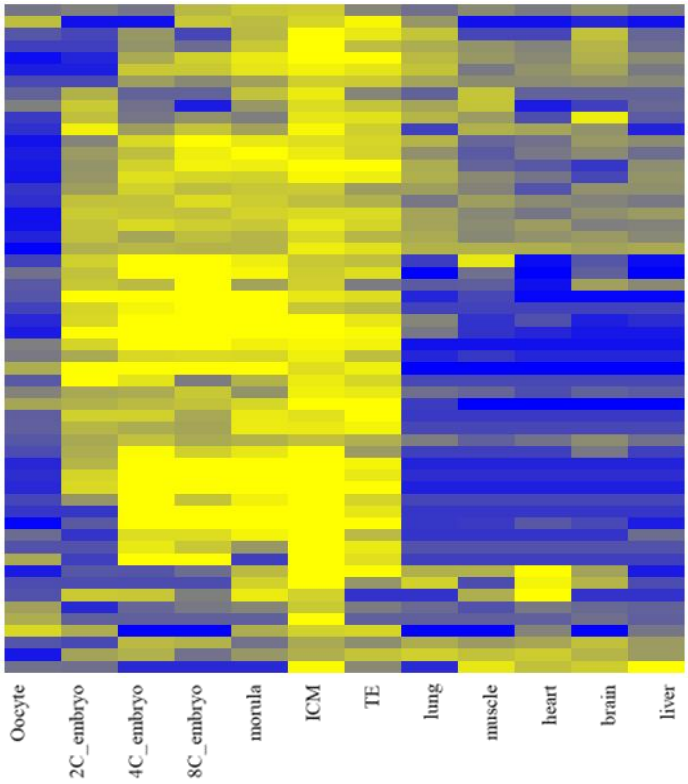
ICM intergenic



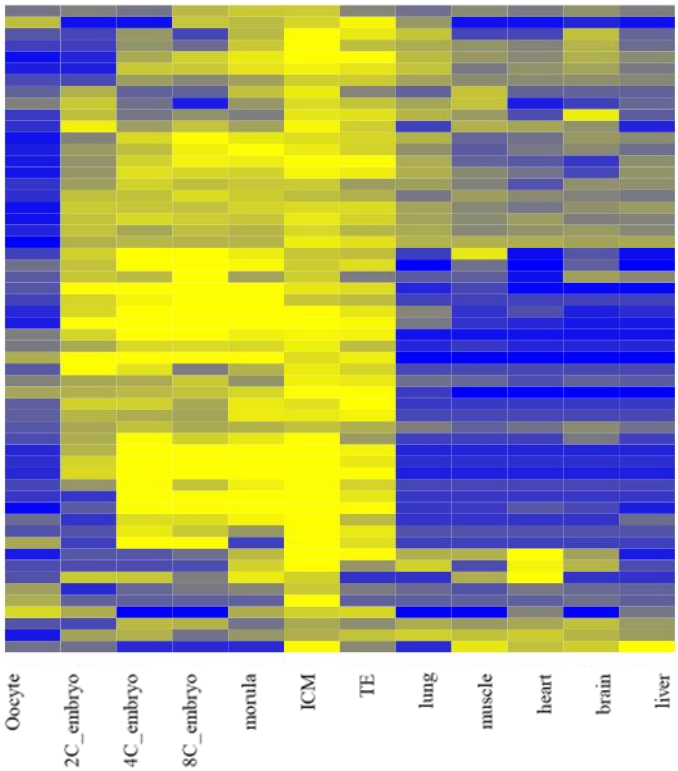
ICM
intragenic



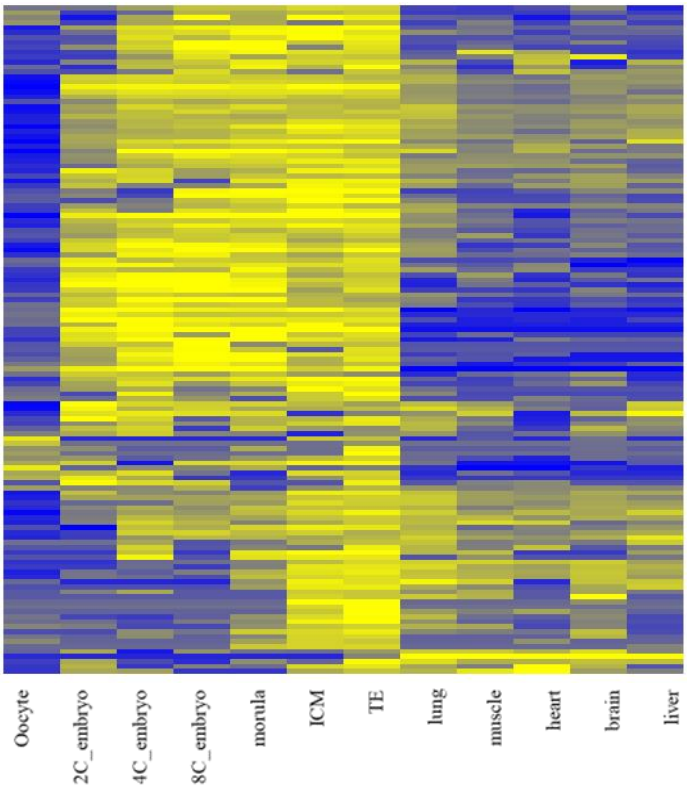
ICM
bidirectional



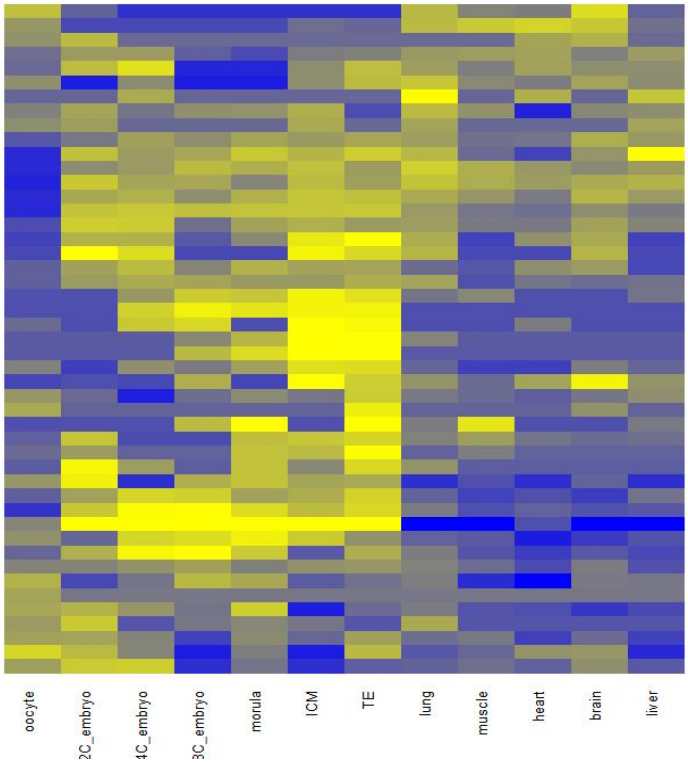
ICM overlapping



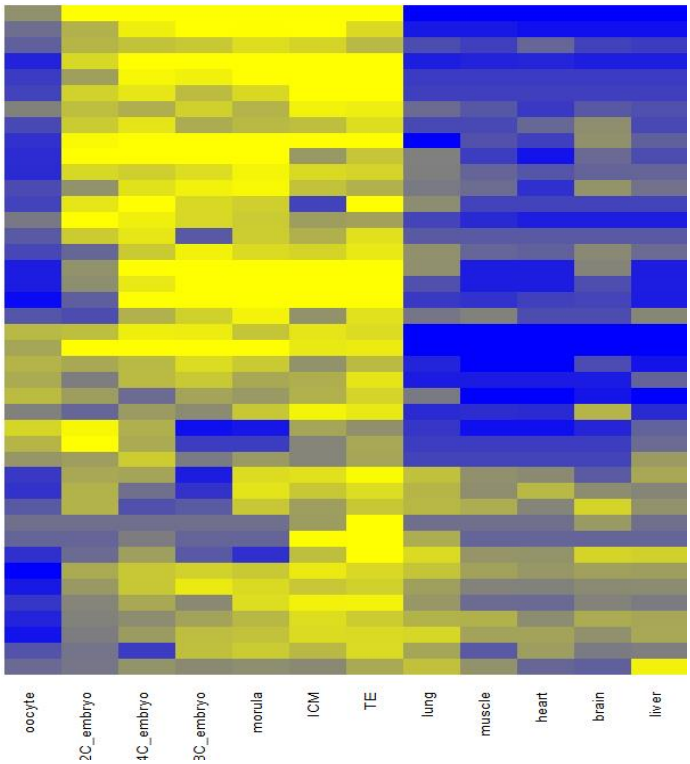
TE intragenic



TE bidirectional



TE overlapping



Appendix 8. Number of genes in each cluster for all datasets and all categories of gene location

Cluster	Number of genes
1	68
2	57
3	29
4	115
5	6
6	33
7	13
8	3

Oocyte intragenic

Cluster	Number of genes
1	134
2	17
3	33
4	27
5	24
6	45
7	17
8	20

Oocyte bidirectional

Cluster	Number of genes
1	79
2	21
3	17
4	15
5	9
6	8
7	3
8	3

Oocyte overlapping

Cluster	Number of genes
1	18
2	27
3	37
4	15
5	7
6	9

2C intragenic

Cluster	Number of genes
1	34
2	40
3	15
4	7
5	20
6	5

2C bidirectional

Cluster	Number of genes
1	26
2	27
3	14
4	2

2C overlapping

Cluster	Number of genes
1	12
2	51
3	68
4	79
5	10
6	12

4C intragenic

Cluster	Number of genes
1	60
2	16
3	10
4	7
5	8
6	13

4C bidirectional

Cluster	Number of genes
1	27
2	15
3	13
4	38
5	10
6	10

4C overlapping

Cluster	Number of genes
1	12
2	12
3	27
4	55
5	36
6	4
7	3
8	5

ICM intragenic

Cluster	Number of genes
1	29
2	21
3	7
4	7

ICM bidirectional

Cluster	Number of genes
1	38
2	3
3	6
4	9

ICM overlapping

Cluster	Number of genes
1	32
2	7
3	80
4	5
5	4
6	7

TE intragenic

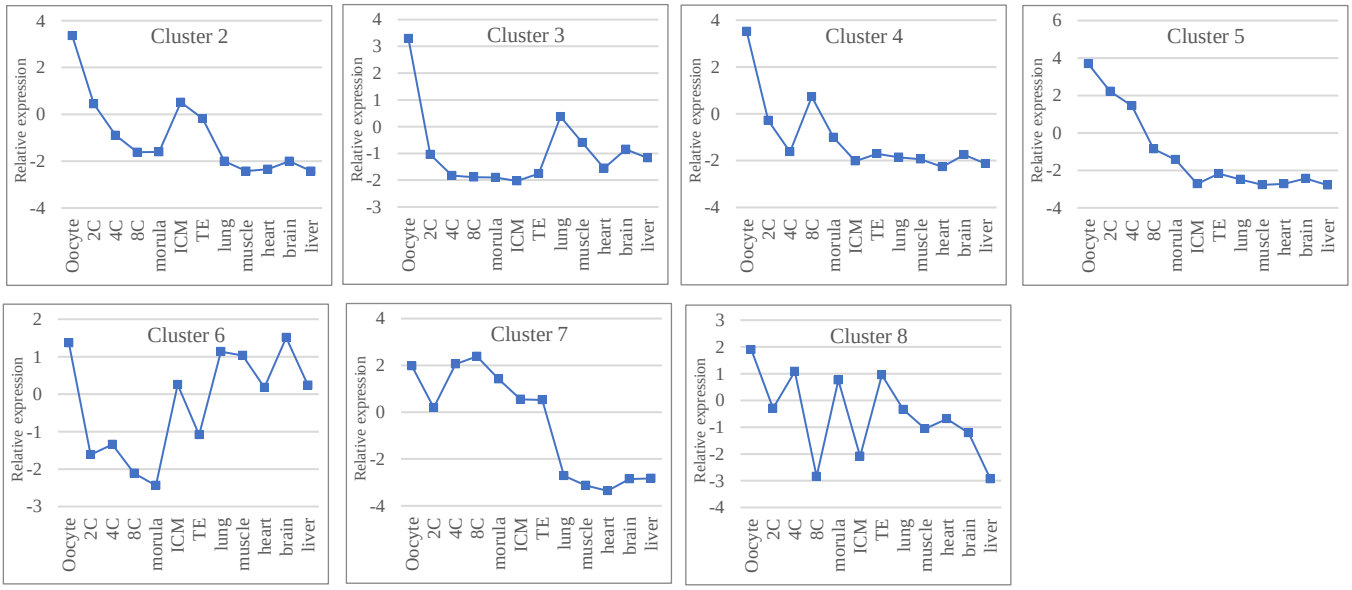
Cluster	Number of genes
1	20
2	18
3	9

TE bidirectional

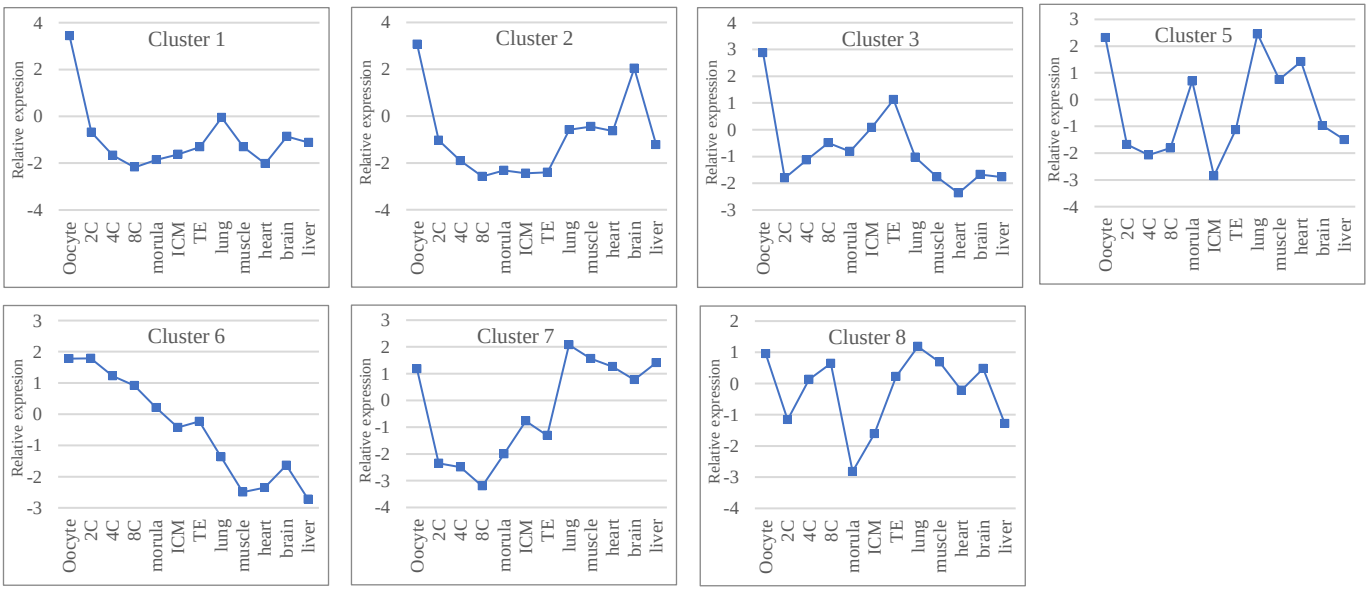
Cluster	Number of genes
1	26
2	3
3	12
4	1

TE overlapping

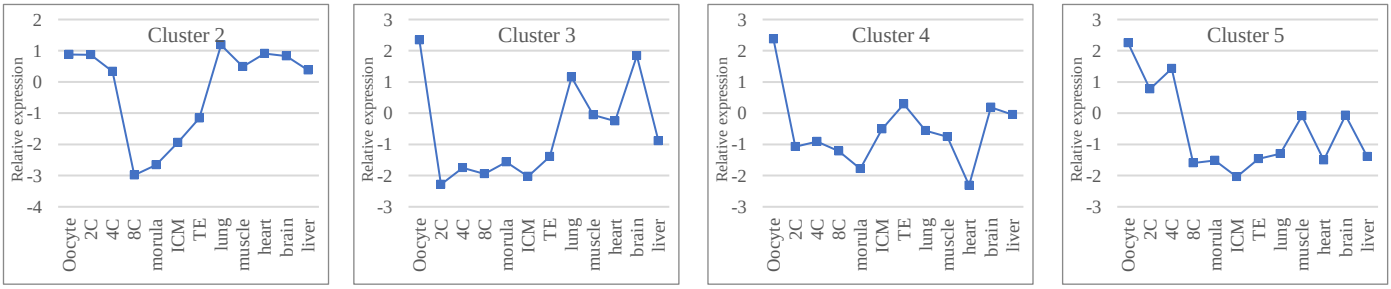
Appendix 9. Relative expression for each cluster for all datasets and all categories

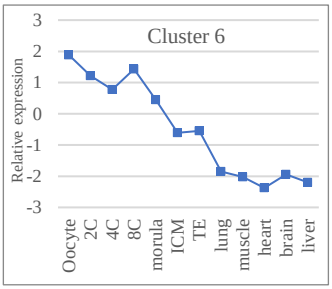


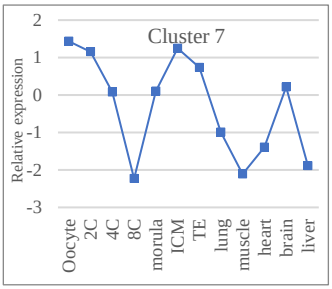
Relative expression of clusters for oocyte intergenic genes

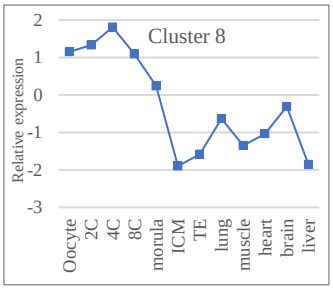


Relative expression of clusters for oocyte intragenic genes

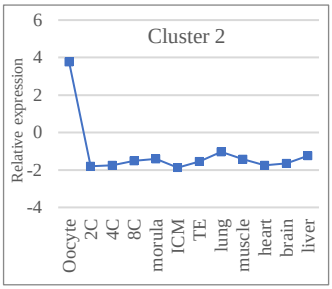


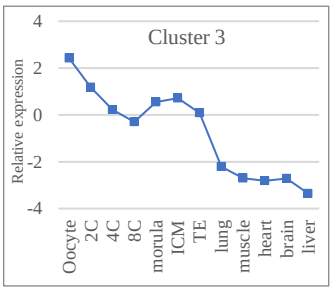


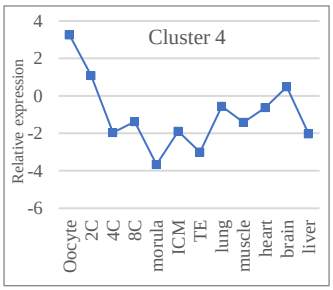


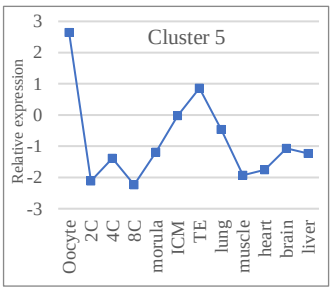


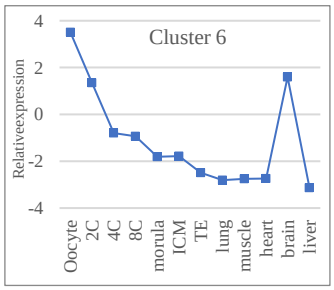
Relative expression of clusters for oocyte bidirectional genes

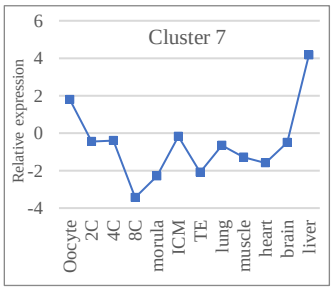


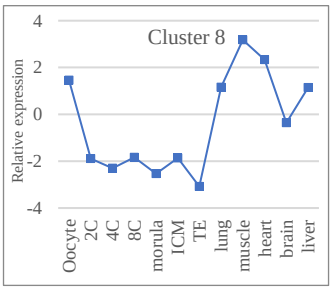




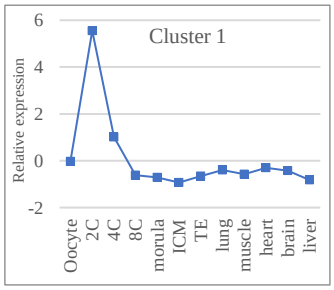


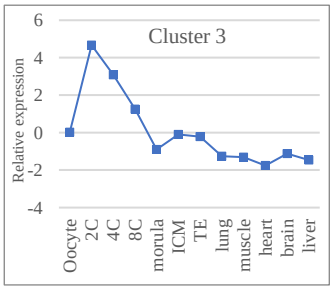


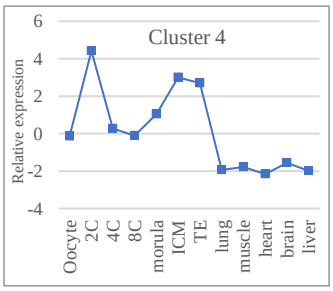


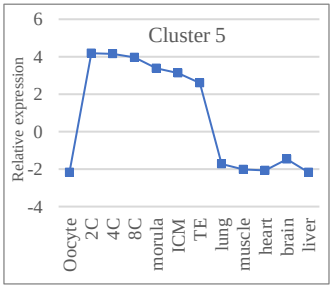


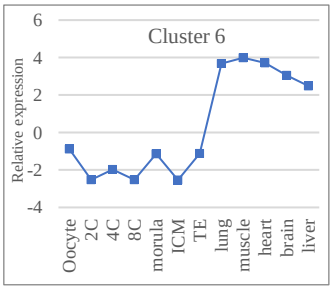
Relative expression of clusters for oocyte overlapping genes



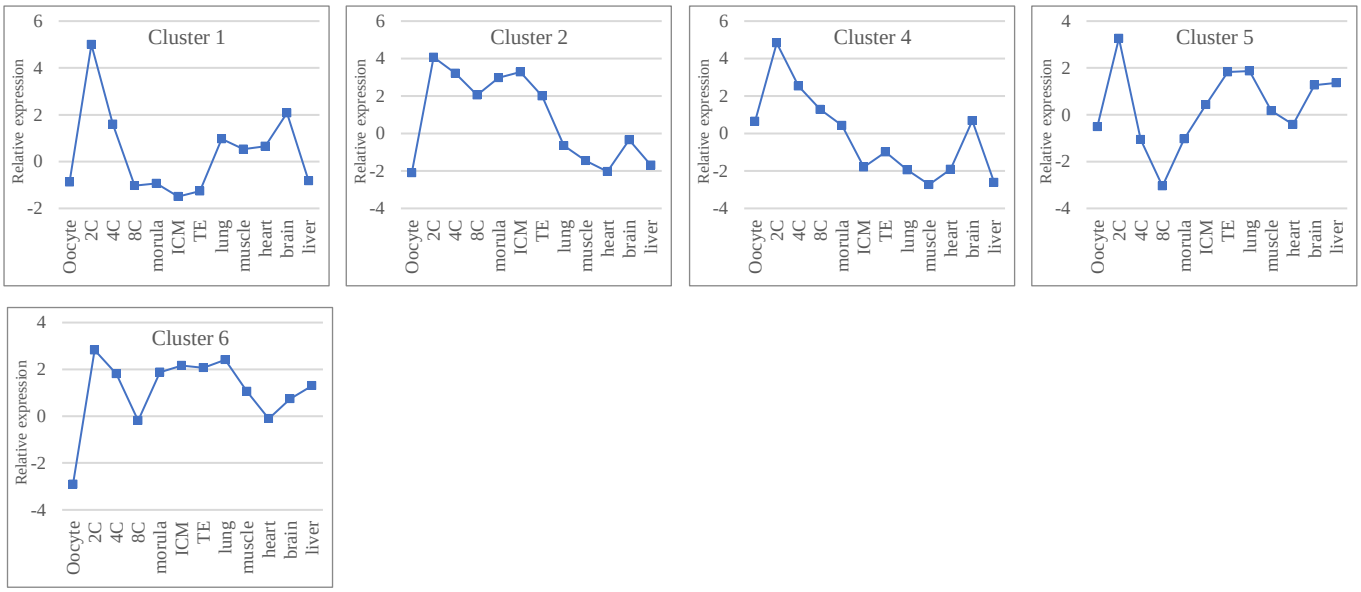




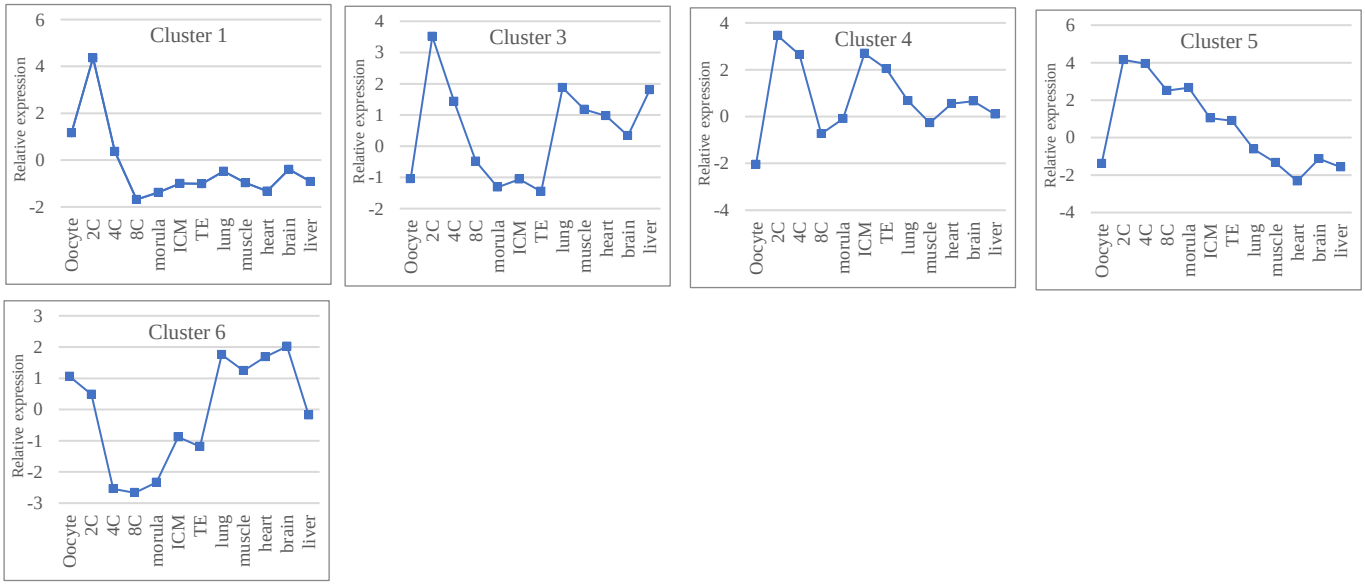




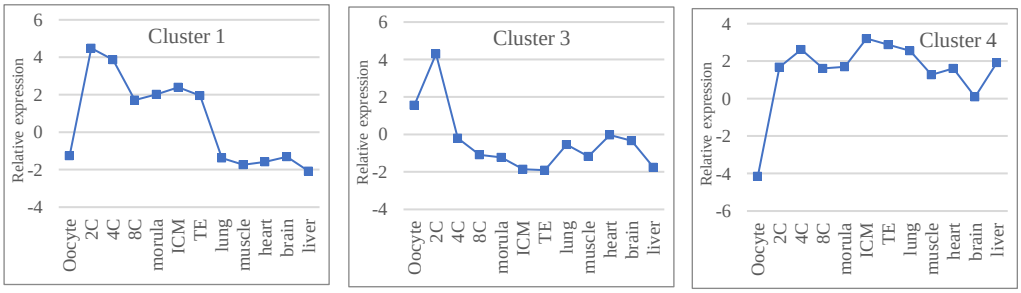
Relative expression of clusters for 2C intergenic genes



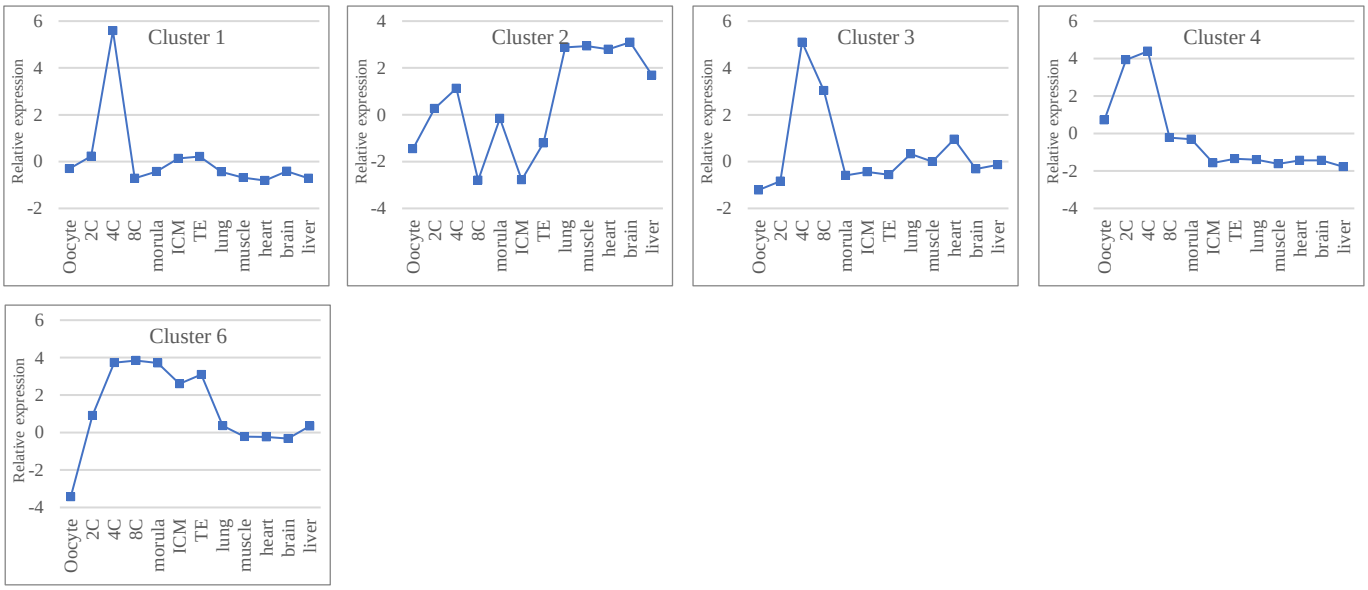
Relative expression of clusters for 2C intragenic genes



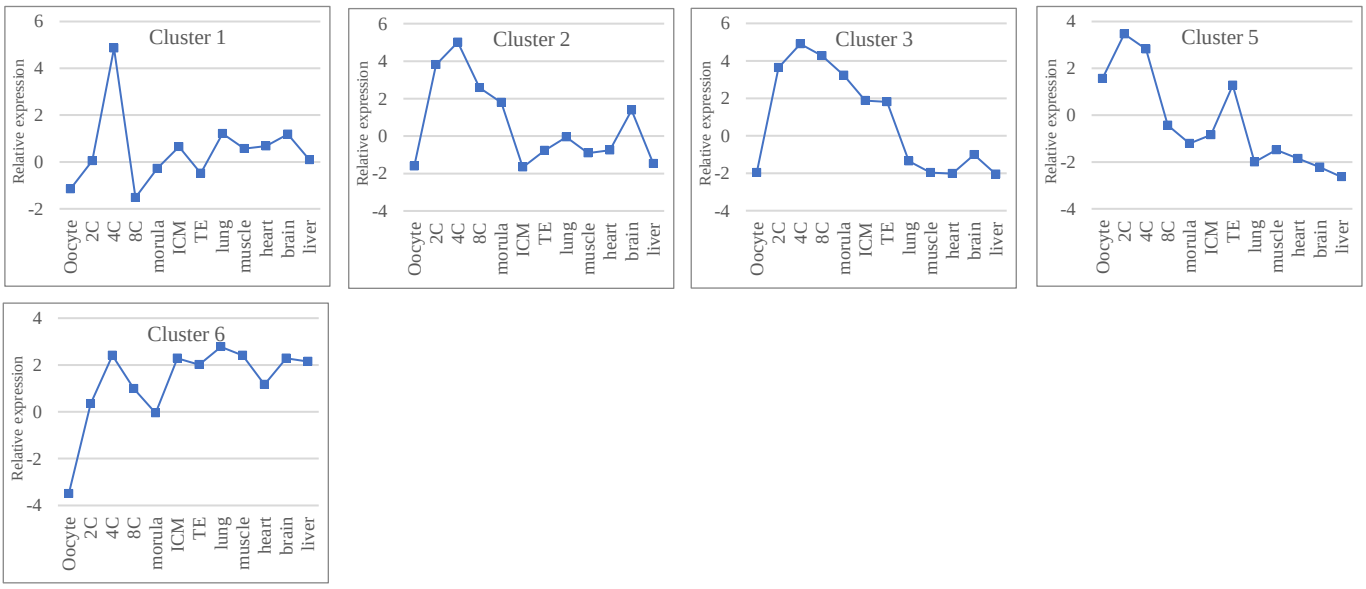
Relative expression of clusters for 2C bidirectional genes



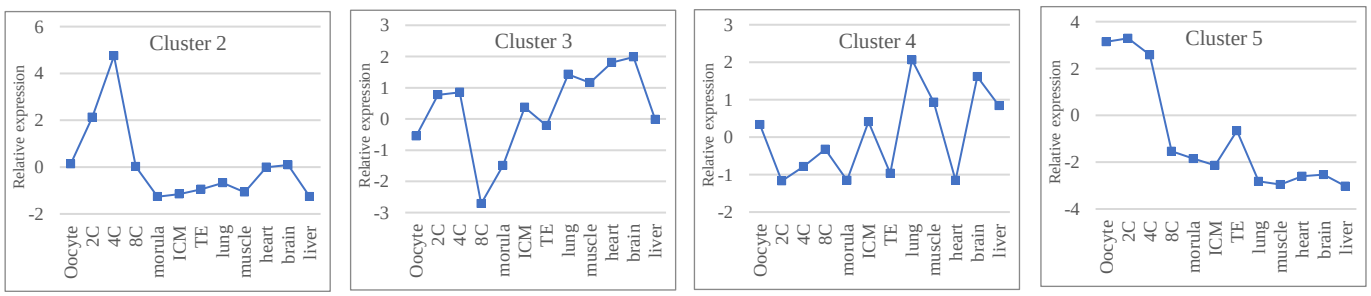
Relative expression of clusters for 2C overlapping genes

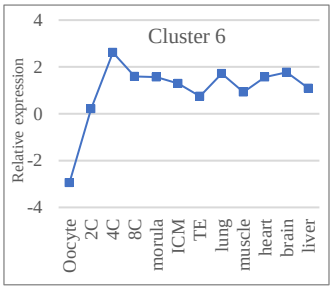


Relative expression of clusters for 4C intergenic genes

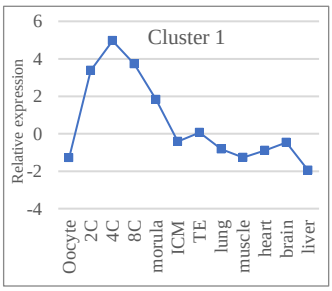


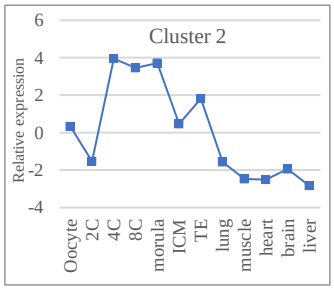
Relative expression of clusters for 4C intragenic genes

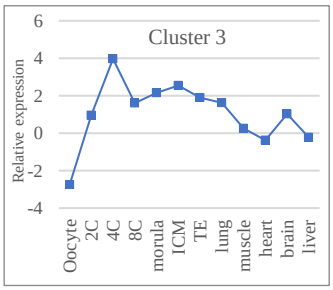


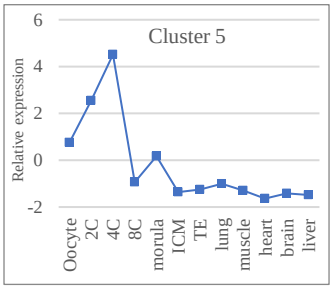


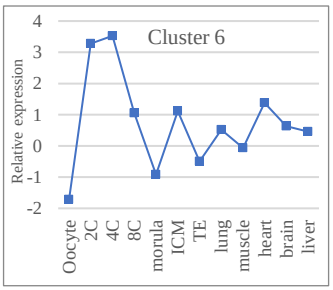
Relative expression of clusters for 4C bidirectional genes



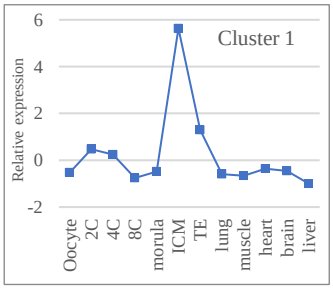


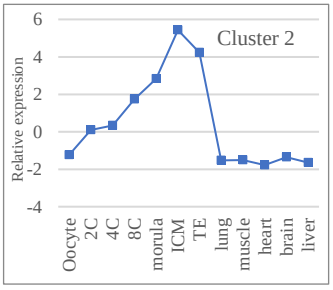


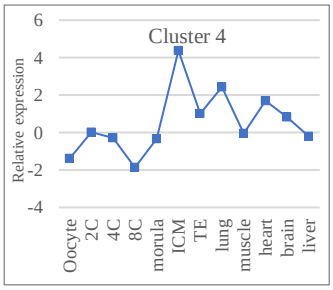


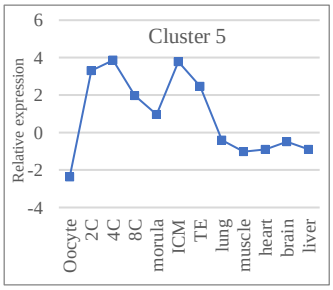


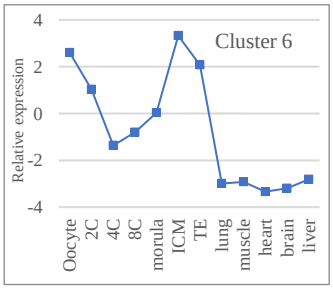
Relative expression of clusters for 4C overlapping genes

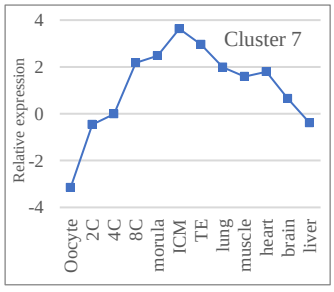




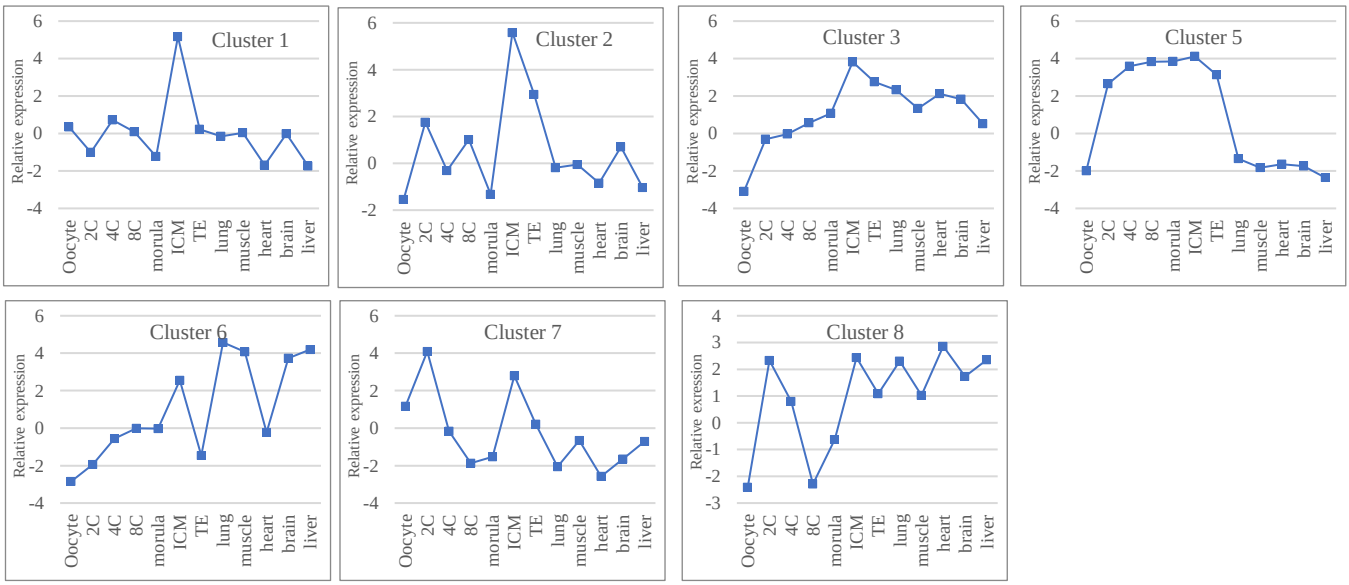




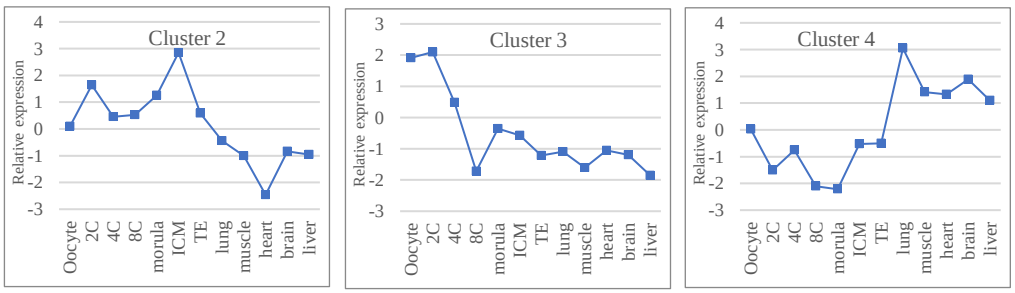




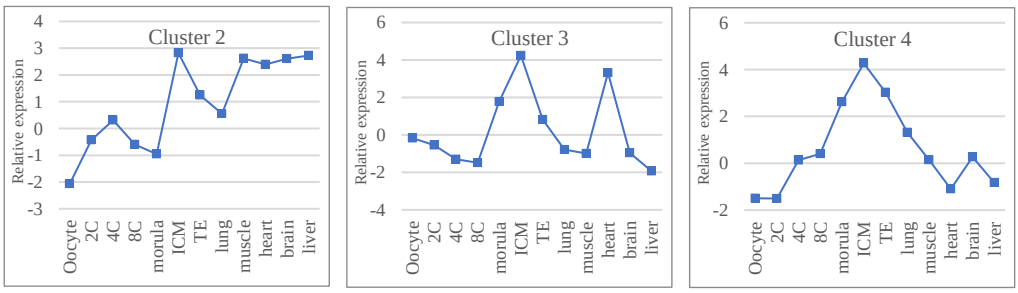
Relative expression of clusters for ICM intergenic genes



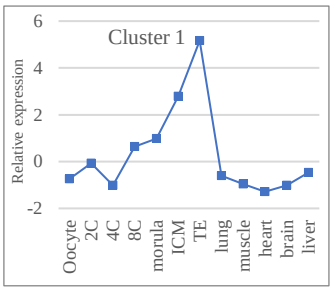
Relative expression of clusters for ICM intragenic genes

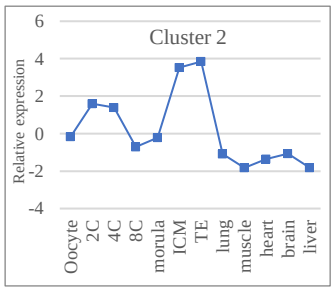


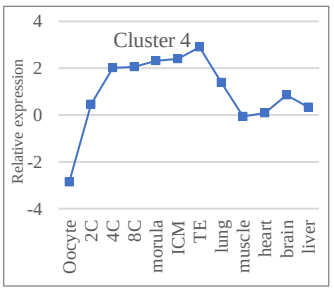
Relative expression of clusters for ICM bidirectional genes

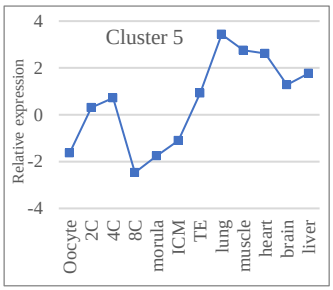


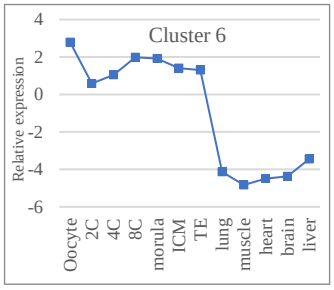
Relative expression of clusters for ICM overlapping genes



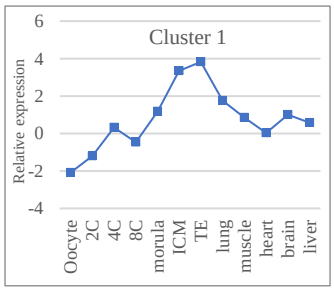


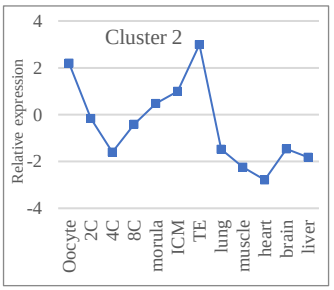


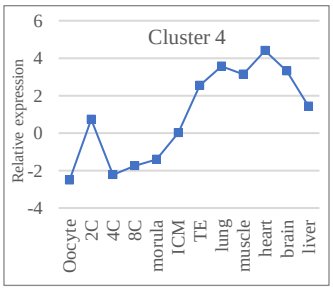


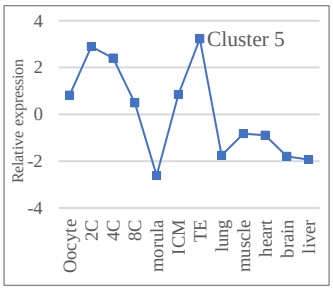


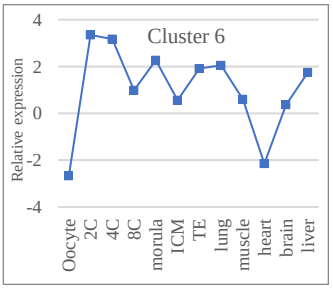
Relative expression of clusters for TE intergenic genes



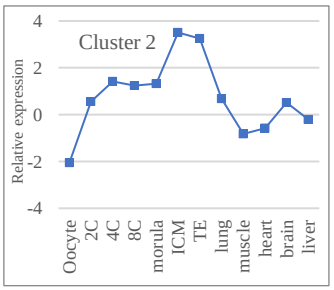


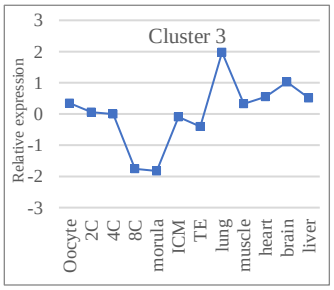




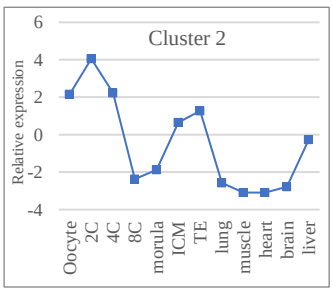


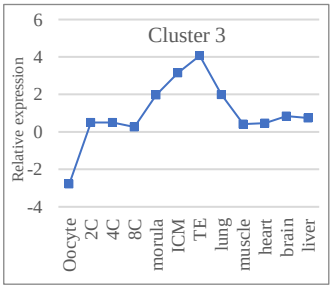
Relative expression of clusters for TE intragenic genes





Relative expression of clusters for TE bidirectional genes





Relative expression of clusters for TE overlapping genes