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

Analysis of the expression of transcripts at imprinted loci across mammalian species during early development

Bachelor Thesis

Sylvia Ramírez

Supervisor: Mgr. Lenka Gahurová, Ph.D.

České Budějovice

2019

Ramírez, S., 2019: Analysis of the expression of transcripts at imprinted loci across mammalian species during early development. Bc. Thesis, in English. – 78., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

Annotation

To explore the conservation of novel transcripts in imprinted regions identified in mouse and developmental regulation of transcripts in imprinted loci, RNA-seq datasets from six mammalian species from various developmental stages were processed followed by de novo transcriptome assembly, filtering and downstream analysis.

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 Budějovice 17.4.2019

……….................................

Signature

Acknowledgements

First, my thanks to God for blessing me with health and strength to undertake this research task and enabling me to its completion. I would like to express my special thanks and gratitude to my supervisor Lenka Gahurová for giving me the opportunity and guidance to do this project. I got enriched with all her knowledge that will help me for the future.

I am dedicating my thesis to my mother, Geraldina Durand, for guiding me towards a knowledge career, for all her amazing support, for encourage me to believe in myself and whose love for me knows no bounds. And to my father, José Ramírez, for being my guardian from the sky during my education, who left a void never to be filled in our lives, but whose memories live with us forever.

And to all the rest of my family and friends who know how much this means to me.

Contents

1 ABSTRACT

Genomic imprinting is a process where a gene is monoallelic expressed only from maternal or paternal allele. Imprinted genes are commonly localized in clusters and regulated by germline differentially methylated regions established in sperm and oocytes of the previous generation. Recently, a number of novel imprinted transcripts was identified in mouse, often employing a transposable element as their promoter. In order to identify novel transcripts in imprinted regions of other mammalian species and to explore the conservation of mouse transcripts in the imprinted regions, we processed RNA-seq datasets from six mammalian species from various developmental stages including oocytes, embryos, placenta and somatic tissues and performed de novo transcriptome assembly. We identified regions homologous to the mouse imprinted regions and predicted the methylation and therefore the imprinting status of the associated gDMRs, affecting the imprinted expression of associated genes. We demonstrated that almost all transcripts in the imprinted regions are specific for a particular developmental period, we identified potential transcription factors regulating their expression, and observed that a relatively high proportion of them employ transposable elements as promoters, although that such transcripts are often not conserved across species, suggesting that transposable elements to some extent shape the transcriptome profile of the imprinted clusters.

2 INTRODUCTION

Mammals as diploid organisms have two copies of each autosomal gene, one copy from mother and one from father. For the vast majority of genes, both copies are equally active (Wolf Reik & Walter, 2001). However, in a small number of genes one copy is silenced in parent-of-origin-dependent manner., This phenomenon is called genomic imprinting (Ferguson-Smith, 2011). The silencing of one copy of the gene is epigenetically regulated and the epigenetic marks (predominantly DNA methylation) are established during either the egg or the sperm formation, without any change in the DNA sequence. (Ishida & Moore, 2013).

Appropriate allele-specific expression of imprinted genes is essential for correct development. Imprinted genes are implicated in the physiology of the fetal-maternal interactions and in many aspects of prenatal and postnatal development. The most prevalent theory for the evolution of imprinting, “the parental conflict hypothesis”, reflects the competing interests of the maternal and paternal genomes in the developing embryo (Wolf Reik & Walter, 2001). In humans, disruption of monoallelic expression of imprinted genes leads to imprinting disorders, such as Prader-Willi, Angelman, Silver-Russel and Beckwith-

Wiedemann syndromes, severely affecting the growth, metabolism and behavior. (Mackay & Temple, 2017)

Gametic imprints can act on whole clusters of genes at once, containing 3–12 imprinted genes and spanning 100–3700 kb of genomic DNA. Most of the genes in one cluster are imprinted protein-coding mRNA genes, but at least one is always an imprinted long non-coding RNA (lncRNA) (Barlow & Bartolomei, 2014). The allele-specific expression in the clusters of imprinted genes is controlled by the allele-specific DNA methylation of the *cis* regulatory sequences, called the imprinting control regions (ICRs), usually one per cluster. ICRs are also called imprinted germline differentially methylated regions (gDMRs), because allelic DNA methylation of ICRs is acquired during gametogenesis (Kelsey & Feil, 2013). However, it should be noted that the term ICR is generally used for imprinted gDMRs that have been proved functionally to control the imprinted gene expression.

Mouse is a classical model to study imprinting and its mechanisms in mammals. However, recent studies were still able to identify novel imprinted genes, sometimes even within the already identified imprinted clusters (Andergassen et al., 2017). In addition, the transcriptome of the mouse oocytes revealed further novel transcripts in the proximity or overlapping gDMRs. Some of these novel genes might confer regulatory roles over either imprint establishment in the oocytes, or regulation of monoallelic expression after fertilization (Andergassen et al., 2017; Courtney W. Hanna et al., 2019; Veselovska et al., 2015). Imprinting in other mammalian species, except human, is poorly studied. The aim of this project therefore is to annotate and analyze the transcriptome within imprinted clusters of other mammalian species with the focus on novel, previously unannotated genes with potentially regulatory roles.

3 BACKGROUND

3.1 The Parental conflict hypothesis

This hypothesis proposes that genomic imprinting evolved in response to a “parental conflict” situation (W. Reik, Dean, & Walter, 2001), which arises from the opposing interests of the maternal and paternal genome, as the embryonic growth is dependent on one parent, but influenced by an embryo whose genome comes from two parents. According to the hypothesis, paternally expressed genes promote fetal growth by extracting resources from the mother, in contrast, maternally expressed imprinted genes are proposed to suppress fetal growth, ensuring her survival and allowing for more equal distribution of her resources to all offspring, with the aim to increase the maximum number of transmission of the maternal genome to multiple offspring, which may have different paternal genomes (Frost & Moore, 2010).

The Parental conflict hypothesis associates the acquisition of imprinting and placenta during the course of evolution (Wolf Reik & Walter, 2001). Consistently, imprinting is observed to occur predominantly in genes influencing fetal growth, particularly through placental growth, suckling and nutrient metabolism (Frost & Moore, 2010).

. Imprinting anomalies are often manifested as developmental and neurological disorders when they occur during early development, and as cancer when altered later in life. The conflict theory is supported by prototypical mouse imprinted gene *Igf2* and its receptor *Igf2r*, where the *Igf2* gene encodes a hormone that stimulates growth during embryonic and fetal development. DNA methylation normally silences the maternal *Igf2* gene. Activation of the maternal *Igf2* gene expression during egg formation or very early in development causes Beckwith-Wiedemann Syndrome, the most common feature is overgrowth (Scott & Weiss, 2000).

3.2 Epigenetics marks associated with imprinting

Parental-allele-specific expression in eutherian mammals is dependent of epigenetic differences between the two parental alleles in order to be transcribed differently in the same nucleus (Kelsey & Feil, 2013). Therefore, the genes on the homologous chromosomes have to be distinguished by epigenetic marks. (Ishida & Moore, 2013).

The two classical epigenetic marks are histone modifications and DNA methylation. DNA is wrapped around an octamer of histone proteins (a nucleosome) enabling its compaction and organization in the nucleus. Modifications of lysine residues in histone 3 (H3) N-terminal tail are associated with transcriptional activation or silencing. DNA methylation is a covalent addition of a methyl group to the cytosine residue, creating 5-methylcytosine, in CpG dinucleotide context (where C and G nucleotides are next to each other on the same DNA strand), and is generally associated with a transcriptionally repressed state (Figure 1).

DNA methylation colocalizes with specific histone modifications, and is mutually exclusive with others. Histone 3 lysine 4 trimethylation (H3K4me3) and histone 3 lysine 27 acetylation (H3K27ac) are active marks found at active promoters and enhancers, negatively correlated with DNA methylation, and positively correlated with gene expression (Smith & Meissner, 2013). Histone 3 lysine 6 di- and trimethylation (H3K9me2/3) are repressive histone marks associated with DNA methylation and transcriptional silencing, while gene body histone

3 lysine 36 trimethylation (H3K36me3) is positively correlated with transcription and promoting acquisition of DNA methylation (Chao, 2011).

To date, in mouse, there are 20 imprinted gDMRs methylated on the maternal allele and 3 on the paternal allele (Proudhon et al., 2012; Tomizawa et al., 2011), with the majority serving as ICRs of clusters of imprinted genes. For these imprinted gDMRs, DNA methylation imprint is acquired during oogenesis (maternally methylated gDMRs) or during spermatogenesis (paternally methylated gDMRs). In addition, the expression of imprinted genes after fertilization can be controlled by somatic DMRs that become methylated in parent-of-origin-specific manner after fertilization (Lewis & Reik, 2006). A number of studies showed that imprinted genes require correct gDMR DNA methylation establishment in the oocyte and sperm for their imprinted expression after fertilization, either at a single-gene or genome-wide level (Bourc'his, 2001; Courtney W. Hanna, Demond, & Kelsey, 2018; Hata, Okano, Lei, & Li, 2002; Kaneda et al., 2010, 2004; Kato et al., 2007; Smallwood et al., 2011). Therefore, it was generally accepted that DNA methylation is the epigenetic mark responsible for the differential marking of alleles in gametes and preserving the information after fertilization.

Nevertheless, a recent study suggested that maternally-inherited histone 3 lysine 27 trimethylation (H3K27me3), histone modification associated with expression silencing, confers imprinting of a small number of genes (Inoue, Jiang, Lu, Suzuki, & Zhang, 2017). This phenomenon is called non-canonical imprinting, in contrast to canonical imprinting regulated by DNA methylation. However, it appears that maternally-inherited H3K27me3 is erased after fertilization at these loci during pre-implantation development, and it is then established in an allele-specific manner after implantation (Courtney W. Hanna et al., 2019). In addition, several placenta-specific imprinted genes were identified, with no obvious regulation by DNA methylation (Andergassen et al., 2017).

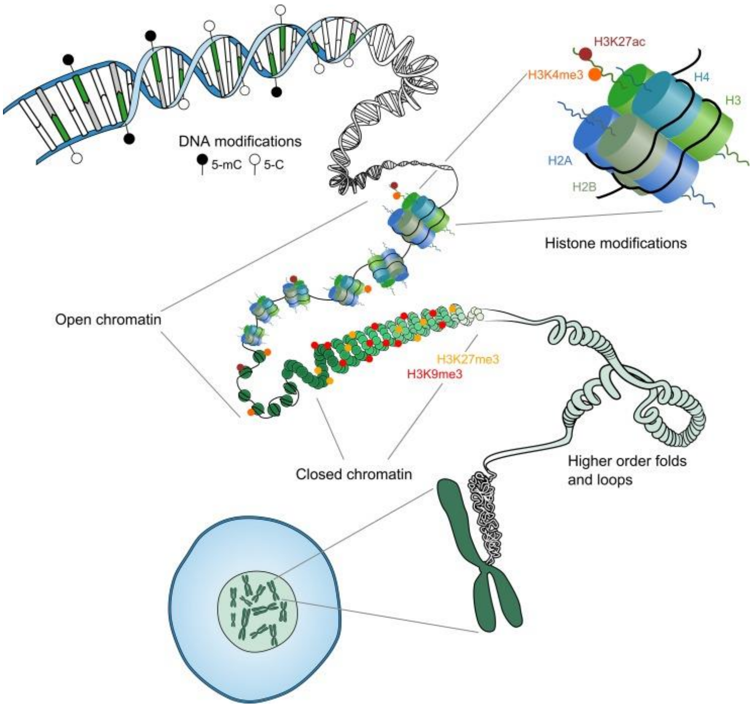


Figure 1. *Modifications or alterations on DNA wrapping to histones, epigenetic mechanism activate or inactive genes (Courtney W. Hanna, Demond, et al., 2018).*

3.3 Establishment of genomic imprints in the germline

Epigenetic properties of the male and female gametes as well as their chromatin organization are profoundly different at the time of fertilization. The sperm DNA is highly methylated and tightly packaged by protamines replacing canonical histones (Wright, 1999), whilst only approximately 40% of oocyte DNA is methylated, in a uniquely form and associated with non-canonical distributions of histone modifications (Shirane et al. 2013; Hanna et al. 2018b).

Oocyte and sperm-specific DNA methylation patterns, including the differential methylation of gDMRs, are established during gametogenesis (Figure 2). Prior to that, pre-existing DNA methylation is reduced in primordial germ cells (PGCs) during their migration to the genital ridge (embryonic days 9.5– 11.5, E9.5- E11.5) (Guibert, Forne, & Weber, 2012) to a very low level throughout the genome. This included imprint erasure, thereby differences in methylation between the parental alleles are removed (Seisenberger et al., 2012), due to downregulation of de-novo DNMTs and the DNMT1-cofactor UHRF1 (Kagiwada, Kurimoto, Hirota, Yamaji, & Saitou, 2012).

In the mouse male gonad, *de novo* DNA methylation, initiates around E13.5 in germ cells arrested in mitosis (known as prospermatogonia) and is complete by E17.5 (Davis, 2000).Therefore, DNA methylation landscape is established before male germ cells undergo meiosis, and has to be maintained through rounds of mitosis and through meiosis. There is a greater opportunity for methylation errors to accumulate or mutations to arise through deamination (C. W. Hanna & Kelsey, 2014).

In the mouse female gonad, PGCs enter first stages of meiosis in E13.5 and arrest in prophase I, around the time of birth, quiescent in the developing ovary until after birth when they are assembled into primordial follicles. (Lucifero, Mann, Bartolomei, & Trasler, 2004) *De novo* methylation initiates after activation of follicles and during the later stages of oocyte growth, around the transition from the primary to secondary follicle, and is completed by the time oocytes are fully grown (C. W. Hanna & Kelsey, 2014). Methylation acquisition is a progressive process depending on the oocyte size. (Kelsey and Feil 2013).

Three functional DNA methyltransferase (DNMT) enzymes have been described in mammals; maintenance methyltransferase DNMT1 methylating hemimethylated sequences after DNA replication, and *de novo* methyltransferases DNMT3A and DNMT3B and their co-factor DNMT3L lacking the catalytic activity (Ishida & Moore, 2013).. In the oocytes, DNMT3A and DNMT3L are essential for *de novo DNA* methylation establishment, including the DNA methylation at all maternally methylated imprinted gDMRs (Bourc'his, 2001; Hata et al., 2002; Kaneda et al., 2010, 2004; Shirane et al., 2013). DNMT3B is dispensable, as oocytes lacking DNMT3B have the same DNA methylation profile as wild-type oocytes (Shirane et al., 2013). In contrast, all three DNMT3s are required for DNA methylation establishment in the male germline. One of the three imprinted gDMRs methylated in sperm requires DNMT3B, as well as small Piwi-interacting RNAs, for its methylation (Kaneda et al., 2004; Kato et al., 2007; Watanabe et al., 2011).

Specific recognition and targeting of imprinted gDMRs for sex-specific acquisition of DNA methylation imprint is not completely understood. For example, it was shown that CpGs in maternally-methylated gDMRs are mostly 8-10 bp from each other, serving as an optimal substrate for methylation by the DNMT3A:DNMT3L tetramer complex in the female germline (Jia, Jurkowska, Zhang, Jeltsch, & Cheng, 2007). In other cell types, it was shown that DNMT3A:DNMT3L complex interacts with unmethylated H3K4 (Dhayalan et al., 2010; Ooi et al., 2007) and trimethylated H3K36 (Dhayalan et al., 2010), and it is repulsed by di- and trimethylated H3K4 (Ooi et al., 2007; Zhang et al., 2010) .

It appears that both paternally- and maternally-methylated gDMRs gain their methylation as a part of sperm and oocyte methylation landscape, respectively. Paternally-methylated gDMRs are relatively CpG-poor, while maternally methylated are CpG-rich. In sperm, the whole genome is methylated with the exception of CpG-rich region (Kobayashi et al., 2012), including maternally-methylated gDMR. In contrast, in the oocytes, only gene bodies of transcriptionally active genes are methylated, with the rest of the genome remaining unmethylated (Kobayashi et al., 2012; Veselovska et al., 2015). It was shown that maternally-methylated gDMRs are all within the transcribed regions, while paternally-methylated gDMRs are in transcriptionally silent intergenic regions (Chotalia et al., 2009; Veselovska et al., 2015). In addition, it was functionally demonstrated that deletion of promoters providing transcription through maternally-methylated gDMRs prevents gDMRs from gaining DNA methylation in the oocytes (Chotalia et al., 2009; Frohlich et al., 2010; Veselovska et al., 2015).

3.4 Maintenance of genomic imprints after fertilization

Upon fertilization and during preimplantation development the genome undergoes epigenetic reprogramming, when the DNA methylation is largely erased by active and passive processes, which are not fully understood (Ishida & Moore, 2013) (Figure 2). The paternal pronucleus is rapidly demethylated through active mechanisms involving a oxidation of 5-methylcytosine to 5-hydroxymethylcytosine and further oxidation derivatives by the TET3 enzyme. (Santos et al., 2013). In contrast, the maternal genome becomes demethylated by a passive mechanism, referring to the dilution of methylation at symmetric CpG sites because of failure to reinstate methylation on the nascent DNA strand at DNA replication due to the absence of DNA methyltransferase DNMT1. Maternal genome is able to resist the active demethylation because of the interactions of maternal factor, DPPA3 (Nakamura et al., 2007) (also called PGC7 or STELLA) with H3K9me2 in the early mouse embryo (Nakamura et al., 2012) protecting the DNA from TET3 activity. Paternal genome binds DPPA3 and therefore resist active demethylation only at imprinted gDMRs, as elsewhere the histones were exchanged for protamines during spermatogenesis (Nakamura et al., 2007, 2012) .

A small number of regions, particularly imprinted gDMRs, escape the global demethylation. It is due to the activity of remaining DNMT1, which is targeted to the imprinted gDMRs and maintains their methylation during replication. The factors targeting the DNMT1 to the imprinted gDMRs are ZFP57 and KAP1 (Lorthongpanich et al., 2013; Messerschmidt et al., 2012; Quenneville et al., 2011). ZFP57 binds a specific sequence present in all gDMRs when methylated, and serves as an anchor for allelic binding of KAP1 (Quenneville et al., 2011). ZFP57 and KAP1 were observed in a complex with DNMT1, its cofactor NP95 (UHRF1), as well as DNMT3A and DNMT3B, presumably targeting DNMT1, but also DNMT3A and DNMT3B to the imprinted loci (Messerschmidt et al., 2012; Quenneville et al., 2011; Zuo et al., 2012).

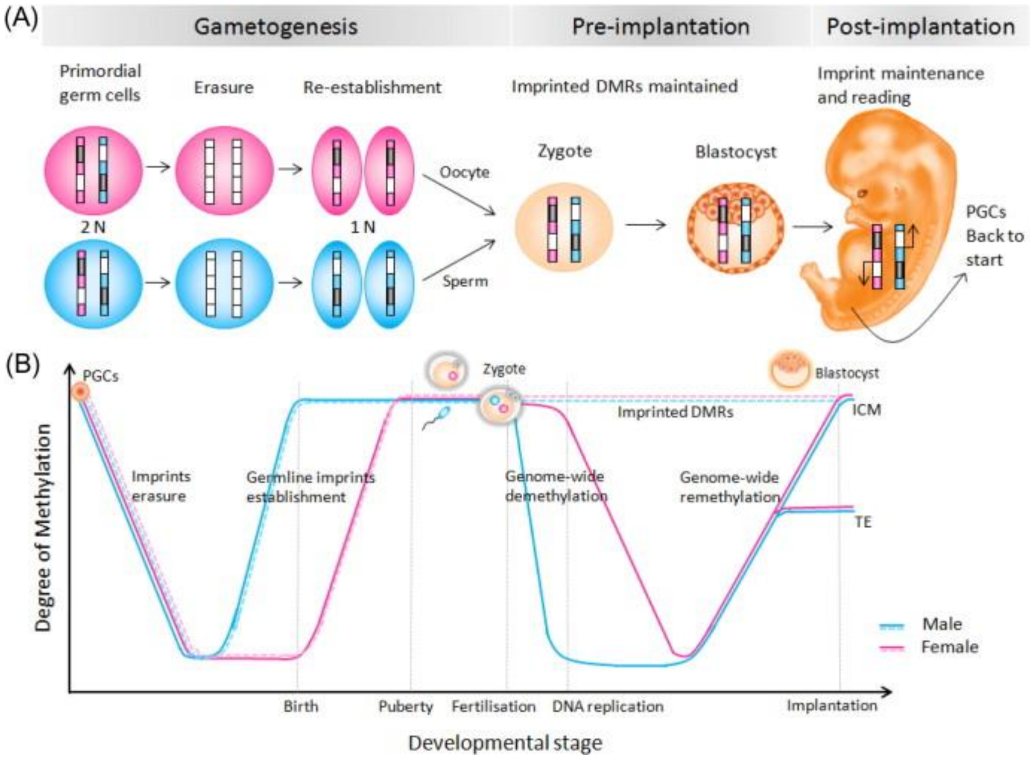


Figure 2. DNA methylation into developmental stages. ***(A)*** *Different phases of Methylation such: as methylation imprint erasure, re-establishment and maintenance at the gDMRs. In pink representing the maternal chromosomes and in blue paternal chromosomes.* ***(B)*** *Shows the methylation imprinting programing (based on mice) (Ishida & Moore, 2013).*

3.5. Imprinted gene clusters in mammals

It has been demonstrated that imprinted gDMRs can regulate the expression of whole clusters of genes. Recent study concluded there are 28 clusters of imprinted genes, containing up to 10 or potentially even more imprinted genes each (Andergassen et al., 2017). The size of the imprinted clusters (and the number of genes) can vary between tissues. For example, *Igf2r* cluster can extend over 10 Mb in placenta, but only up to 1 Mb in adult somatic tissues (Andergassen et al., 2017). To date, 124 mouse imprinted genes and 2 predicted imprinted genes have been identified and listed in the database of imprinted genes www.geneimprint.com/site/genes-by-species (Randy L. Jirtle, n.d.). However, recent studies identified several novel imprinted genes (Andergassen et al., 2017; Courtney W. Hanna et al., 2019; Inoue et al., 2017), both annotated and unannotated, suggesting that the current list of mouse imprinted genes is not yet complete. In addition, some genes appear to be imprinted only in some tissues, particularly in extraembryonic tissues and placenta (Andergassen et al., 2017; Courtney W. Hanna et al., 2019). Most of genes in any one cluster are imprinted protein-coding mRNA genes; but at least one is usually an imprinted lncRNA. (Barlow & Bartolomei, 2014).

The only species with imprinting studied to the similar extent as mouse is human, with 107 imprinted and 106 predicted imprinted genes listed in the database (www.geneimprint.com/site/genes-by-species, March 25th, 2019) (Randy L. Jirtle, n.d.). Nevertheless, a substantial proportion of genes that are imprinted in mouse are not imprinted in human (Morcos et al., 2011), and humans show much more prominent phenomenon of placenta-specific imprinting compared to mouse (Hamada et al., 2016; Courtney W. Hanna et al., 2016) . Interestingly, in these cases, methylated gDMRs (almost exclusively maternally-methylated) regulating the allele-specific expression retain their allele-specific methylation in the placenta, but not in embryonic tissues (Hamada et al., 2016). The imprinting in other mammalian species is poorly studied, with 20 imprinted genes identified in cow, 22 in pig, 6 in rat and 9 in rhesus macaque. To date, imprinting was not studied at all in many species, including marmoset.

3.5.1 Non-coding RNAs, alternative promoters and transposable elements as important regulators of imprinting

The majority of autosomal imprinted gene clusters contain at least one lncRNA (Long non-coding RNAs). The most common mechanism of gene expression regulation in imprinted clusters is that the unmethylated gDMR serves as a promoter of a lncRNA that consequently silences other genes in cis (Mancini-DiNardo, 2006; Sleutels, Zwart, & Barlow, 2002), either through direct transcriptional interference if they overlap, or through guiding the deposition of repressive marks H3K9me2 and H3K27me3 (Latos et al., 2012; Mager, Montgomery, de Villena, & Magnuson, 2003; Nagano et al., 2008; Terranova et al., 2008). The classical examples are *Airn* lncRNA in *Igf2r* cluster and *Kcnq1ot1* lncRNA in *Kcnq1* cluster.

In addition, recent findings suggest there are two more types of regulatory lncRNAs in imprinted regions. Thorough annotation of the mouse oocyte transcriptome revealed that all maternally-methylated gDMRs are intragenic, i.e. inside gene bodies of genes active in the oocytes, even if they were promoter-associated according to the official annotation (Veselovska et al., 2015). For gDMRs that are promoter-associated but not intragenic according to the official annotation, the transcription through them is provided either by novel upstream promoters of respective annotated genes (such as *Plagl1*, *Peg3*, *Peg10* and *Impact*), or through novel lncRNAs (such as *Slc38a4*). These are often oocyte-specific (Veselovska et al., 2015), not expressed in PGCs or after fertilization.

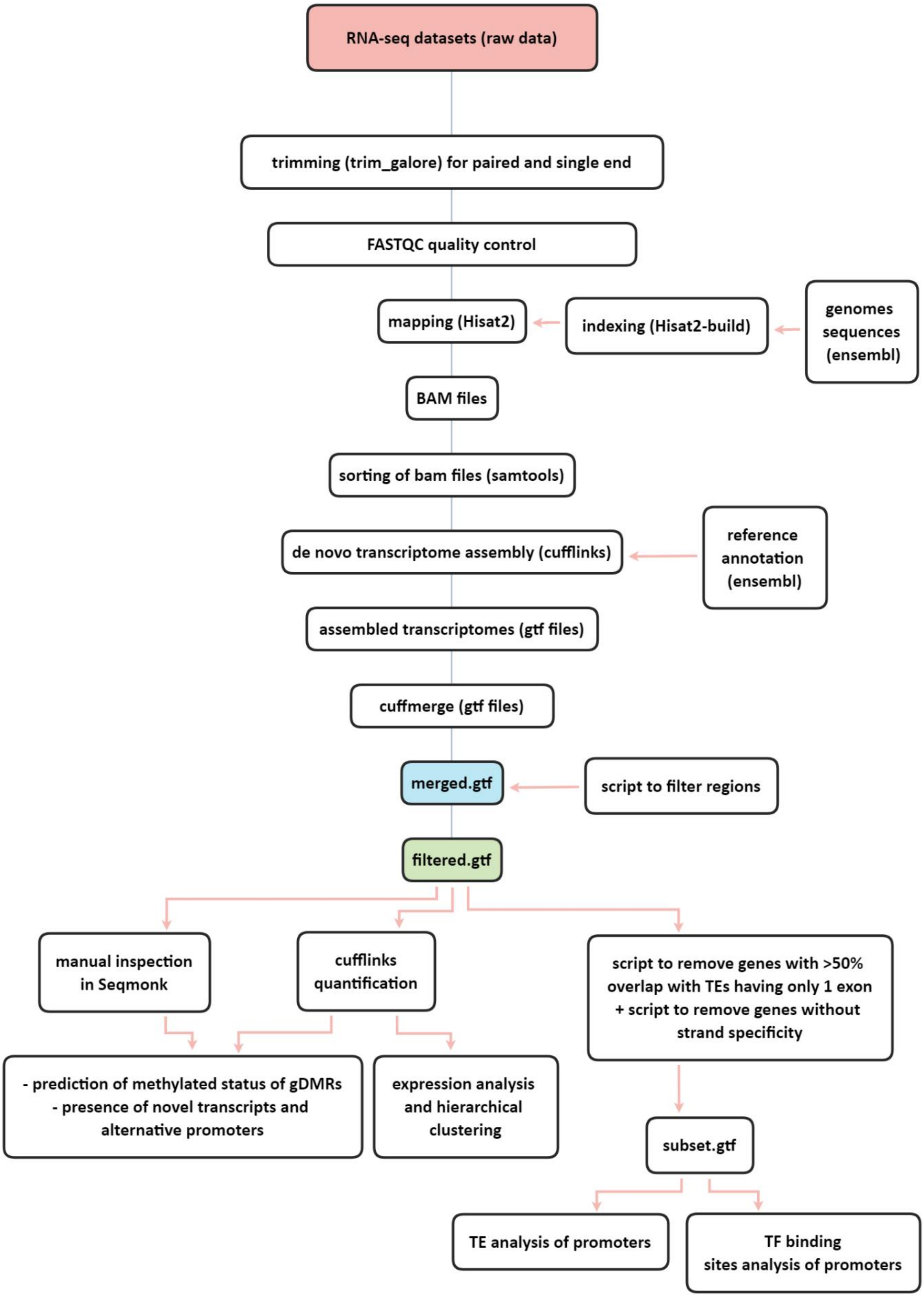
Second group of novel regulatory lncRNAs in imprinted regions are lncRNAs upstream of imprinted genes, expressed from the same DNA strand (exemplified by imprinted genes *Znf64*, *Jade1* and *Slc38a4*) (Andergassen et al., 2017; Courtney W. Hanna et al., 2019). These are often imprinted with the same allele-specificity as the respective imprinted gene, and it was suggested they might me involved in the regulation of imprinted expression of the nearby gene in cis, potentially independently on the methylation status of gDMR. The best studied example is *Slc38a4*. Its imprinted expression in embryonic lineage is fully regulated by the methylation status of gDMR which should be maternally-methylated. After fertilization of oocytes deficient in DNA methylation and therefore without maternally-inherited gDMR methylation, *Slc38a4* expression is biallelic. However, in extraembryonic lineage (which will later become placenta), the expression of *Slc38a4* is still imprinted even if the oocyte was deficient in DNA methylation. Instead, this noncanonical imprinting might be regulated by the upstream placenta-specific imprinted lncRNA, whose allele-specific transcription might enhance the allele-specific transcription of *Slc38a4* (Courtney W. Hanna et al., 2019). In addition, novel alternative promoters of imprinted genes were identified, with regulation independent of gDMR methylation (such as for gene placenta-specific alternative promoter of gene *Gab1*) (Courtney W. Hanna et al., 2019).

Therefore, clusters of imprinted genes appear to be relatively rich for novel lncRNAs and alternative promoters, often with expression restricted to a specific cell type, and with potential important regulatory roles. Moreover, these often employ transposable elements as promoters (a frequent phenomenon in the oocytes and embryos in general), suggesting that transposable elements might be involved in shaping the imprinted gene expression of crucial developmental genes. The activity of transposable elements as promoters is the most prominent and often specific for particular early developmental stages such as oocytes and embryos (Fadloun et al., 2013; Franke et al., 2017; Karlic et al., 2017; Macfarlan et al., 2012; Peaston et al., 2004; Veselovska et al., 2015), and extraembryonic tissues (Chuong, Rumi, Soares, & Baker, 2013), therefore, the expression of whole transcripts is expected to be stage-specific. Nevertheless, all the lncRNAs and upstream promoters were identified in mouse, and their presence was not explored in other mammalian species.

4 AIMS

- Find and process available RNA-seq datasets from selected mammalian species, most importantly oocyte and embryo, but also somatic tissues and placenta.
- Assemble the complete transcriptome from all developmental stages for each species.
- Generate a complete list of imprinted regions in mouse and other selected mammalian species.
- Assess the potential imprinting status of regions imprinted in mouse with maternally-methylated gDMR.
- Analyze the expression changes of transcripts in imprinted regions during development.
- Perform the sequence analysis of promoters to find potential transcription factor biding sites and compare their conservation between species.
- Analyze how frequently the transcripts employ transposable elements as promoters in imprinted regions.
- Compare the selected imprinted regions across species and explore the conservation of selected novel transcripts identified in mouse imprinted regions.

5 WORKFLOW OVERVIEWS


RNA-seq datasets (raw data)
trimming (trim_galore) for paired and single end
FASTQC quality control
mapping (Hisat2)
indexing (Hisat2-build)
genomes sequences (ensembl)
BAM files
sorting of bam files (samtools)
de novo transcriptome assembly (cufflinks)
reference annotation (ensembl)
assembled transcriptomes (gtf files)
cuffmerge (gtf files)
merged.gtf
script to filter regions
filtered.gtf
manual inspection in Seqmonk
cufflinks quantification
script to remove genes with >50% overlap with TEs having only 1 exon + script to remove genes without strand specificity
- prediction of methylated status of gDMRs - presence of novel transcripts and alternative promoters
expression analysis and hierarchical clustering
subset.gtf
TE analysis of promoters
TF binding sites analysis of promoters

6 MATERIAL AND METHODS

6.1 Datasets

RNA-seq datasets were downloaded as fastq files from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk). Datasets with following accession codes were used in this project: rat (*Rattus norvegicus*) datasets GSE112622 (Brind'Amour et al., 2018), and GSE114191(Carelli, Liechti, Halbert, Warnefors, & Kaessmann, 2018), pig (*Sus scrofa*) datasets GSE108900 (Tsai, Tyagi, & St. John, 2018), GSE53387 (Bernardo et al., 2018) and GSE106577 (Y. Li et al., 2018), cow (*Bos taurus*) datasets GSE61717 (Reyes, Chitwood, & Ross, 2015), GSE99210 (Lavagi et al., 2018), GSE53387 (Bernardo et al., 2018) and GSE43013 (Fushan et al., 2015), rhesus macaque (*Macaca mulatta*) datasets GSE117219 (Liu et al., 2018) , GSE112536 (Ruebel et al., 2018), GSE118284 (Dunn-Fletcher, unpublished) , GSE103313(Chitwood, Burruel, Halstead, Meyers, & Ross, 2017), GSE86938(Xinyi Wang et al., 2017) and GSE114191 (Carelli et al., 2018), marmoset (*Callithrix jacchus*) datasets E-MTAB-7078 (Boroviak et al., 2018) and GSE114191 (Carelli et al., 2018), and human (*Homo sapiens*) datasets GSE36552 (Yan et al., 2013), GSE49828 (Guo et al., 2014), GSE101571 (Wu et al., 2018), GSE118285 (Dunn-Fletcher et al., 2018) and GSE114191 (Carelli et al., 2018). Detailed list of datasets used in this project can be found in Supplementary Table 1.

6.2 Trimming

To remove both poor-quality bases and adapters, reads were first trimmed using program Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) v0.4.1 with default parameters, specifying whether the reads were generated in single end or paired end mode. For single end reads, the command "trim_galore *fastq.gz" was used, for paired end reads, it was the command "trim_galore --paired *fastq.gz".

6.3 Quality control of trimmed reads

After trimming, we checked the quality of the data (sequence quality and content, GC content, sequence length distribution, sequence duplication levels and overrepresented sequences) using program FastQC (Andrew S. 2010) v0.11.5 with default parameters, to check whether all the datasets are of sufficient quality for downstream analyses. The commands were "fastqc *_trimmed.fq.gz" and "fastqc *.fq.gz" for single end mode and for paired end mode, respectively.

6.4 Mapping

Prior to data mapping, genomes of selected species were downloaded from Ensembl genome database (https://www.ensembl.org/index.html) in fasta format. The following genome versions were used for each species: Rnor_6.0 for rat, Mmul_8 for rhesus macaque, GRCh38 for human, calJac3 for marmoset, UMD3.1 for cow and Sscrofa11.1 for pig. Genomes were then indexed using hisat2-build function of Hisat2 v2.0.5 that outputs the indexed genome as a set of 8 files with suffixes 1.ht2, 2.ht2, 3.ht2, 4.ht2, 5.ht2, 6.ht2, 7.ht2, and 8.ht2.

Trimmed reads were mapped to the indexed genome of respective species (specified by -x parameter) using Hisat2 (Kim, Langmead, & Salzberg, 2015; Pertea, Kim, Pertea, Leek, & Salzberg, 2016) v2.0.5 with parameters specifying the maximum and minimum penalties for soft-clipping per base (--sp) and modifying the output to be compatible with de novo transcriptome assembly using Cufflinks (--dta-cufflinks). The output file from Hisat2 with mapped reads is a Sequence Alignment Map (sam) file, which was directly converted to Binary Alignment Map (bam) file using samtools view function of samtools v1.3.1 (H. Li, 2011; H. Li et al., 2009). The following command was used for mapping of single end reads: “hisat2 –sp 1000,1000 –dta-cufflinks -x indexed_genome_base -U trimmed_data_trimmed.fq |samtools view -bS -F 4 -F 256 -> mapped_reads.bam”. For mapping paired end data, we used the command: “hisat2 –sp 1000,1000 –dta-cufflinks -x indexed_genome_base -1 trimmed_read1_val_1.fq.gz -2 trimmed_read2_val_2.fq.gz |samtools view -bS -F 4 -F 256 -> mapped_reads.bam”.

Mapped data were then sorted using samtools sort function of samtools v1.3.1 in order to be able to perform de novo transcriptome assembly using Cufflinks. The command for sorting was: “samtools sort -o output_sorted.bam input.bam”.

6.5 De novo transcriptome assembly

De novo transcriptome assembly was performed on selected datasets (Supplementary Table 1) in using program Cufflinks (A. Roberts, Pimentel, Trapnell, & Pachter, 2011; Adam Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Trapnell et al., 2013, 2010) v2.2.1 in the reference annotation based transcript (RABT) (A. Roberts et al., 2011) mode, in which the assembler assigns reads to the supplied official annotation, and remaining reads are used to build models of novel transcripts. Therefore, the final transcriptome assembly (a gtf file) contains all transcripts from the supplied annotation, as well as the newly assembled transcripts. The transcriptome annotations used as a baseline for de novo transcriptome assembly were downloaded from Ensembl and are as follows: Bos_taurus.UMD3.1.94.chr.gtf for cow, Homo_sapiens.GRCh38.94.chr.gtf for human, Callithrix_jacchus.C_jacchus3.2.1.91.gtf for marmoset, Sus_scrofa.Sscrofa11.1.94.chr.gtf for pig, Rattus_norvegicus.Rnor_6.0.94.chr.gtf for rat and Macaca_mulatta.Mmul_8.0.1.94.chr.gtf for rhesus macaque. RABT transcriptome assembly was performed with default parameters (using parameter -g specifying RABT mode), specifying the type of strand specificity of the library as --library-type fr-unstranded, fr-firststrand and fr-secondstrand (see Supplementary Table 1 for the library type parameter for each dataset). The command to perform de novo transcriptome assembly was:" cufflinks -g ensemble_annotation.gtf -u --library-type fr-xxx -o output_folder sorted_mapped_reads.bam".

Individual assembled transcriptomes were then merged together using a function cuffmerge from the program Cufflinks v2.2.1 to create a comprehensive transcriptome annotation that contain transcripts from all studied developmental stages and cell types for each species. List of all assemblies to be merged was prepared as a txt file (assemblies.txt). The command to merge the assemblies while removing potential artefacts and redundant transcripts was "cuffmerge assemblies.txt". First, all assemblies from the same source (same original publication) were merged, and then the merged assemblies from each source were merged together (individually for each species).

6.6 Downstream analysis

Data processing until the step of merging gtf files was performed for all six selected mammalian species. The identification of regions homologous to mouse imprinted regions and subsequent filtering of gtf files to contain only transcripts within these homologous regions were performed for rat, cow, pig, marmoset and macaque rhesus. All downstream analyses using the filtered gtf files were performed only for rat and cow.

6.6.1 Python scripts for filtering of gtf files

Three python scripts were generated for filtering of gtf files. They are described in respective results chapters and the codes can be found in Appendices 1,2,3.

6.6.2 Transcript expression quantification, expression profiling and promoter analysis

The expression of transcripts was quantified using Cufflinks v2.2.1 (A. Roberts et al., 2011; Adam Roberts et al., 2011; Trapnell et al., 2013, 2010) disabling the de novo transcriptome assembly function using the command "cufflinks -G rat_merged_filtered.gtf -o outputfolder mappedreads.bam.". The data processing after Cufflinks quantification for hierarchical clustering and heatmap generation is described in the respective result chapter.

Hierarchical clustering and the generation of the heatmap was performed in R (v3.5.3, 64bit) using a custom script described in the respective results chapter. The code is in the appendix 4. The expression profiles of individual clusters were generated using Microsoft Office Excel (v14.0, 32bit). The process of generating promoter sequences for motif sequence analysis is described in detail in results chapter. DREME (Timothy L. Bailey, 2011) and TOMTOM tools (Gupta, Stamatoyannopoulos, Bailey, & Noble, 2007) from the MEME suite v5.0.5 (T. L. Bailey et al., 2009) were used with default parameters using the web interface. In DREME, we used shuffled sequences as s control. The annotation of rat and cow repetitive elements was downloaded from the UCSC genome browser for the respective genome annotations (Rnor_6.0 for rat and UMD3.1 for cow). The graphs of numbers of transcripts with promoters overlapped by repetitive elements were generated using Microsoft Office Excel (v. 14.0, 32bit).

6.6.3 Data inspection in Seqmonk

All visual inspections of the data and assembled transcriptomes were performed using Seqmonk v1.43.0 program (Popp et al., 2010) (www.bioinformatics.bbsrc.ac.uk/projects/seqmonk) with annotations Rnor_6.0_v90 for rat and UMD3.1_v91 for cow. When screenshots with RNA-seq reads were made, reads were quantified using wiggle plot quantification with 50 bp window size and 50 bp step size. Apart from checking the quality of the data, the data were inspected in order to identify whether a predicted gDMR is overlapped by an active oocyte transcript, to identify transcripts overlapped by repetitive element on the same strand by more than 50%, to identify which transcripts have TSS region (first 200 bp of the transcript) overlapped by same strand repetitive element, and to extract coordinates of promoter regions defined as +-500 bp around the TSS, and to inspect the potential conservation of novel transcripts in imprinted loci in mouse.

7 RESULTS

7.1 Datasets

7.1.1 Identification and selection of datasets for the analysis

In order to compare the transcripts in the imprinted clusters in multiple mammalian species, we first identified suitable species with well-annotated genomes with sequences assembled in whole chromosomes, not just contigs without chromosomal information, in Ensembl genome browser (https://www.ensembl.org), and with available RNA-seq datasets from various developmental stages, which we searched using the publications database PubMed (https://www.ncbi.nlm.nih.gov/pubmed) and the data repository GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo). We were particularly interested in RNA-seq datasets from the oocytes, embryos and placenta, because they are generally rich in novel stage-specific unannotated transcripts including the transcripts within the imprinted loci (Andergassen et al., 2017; Courtney W. Hanna et al., 2019; Veselovska et al., 2015), as well as from the somatic tissues. That would allow us to assemble complete transcriptome encompassing the developmentally-regulated changes in transcription, annotate stage- or tissue-specific transcripts and analyze their expression changes. We excluded mouse, the most commonly used mammalian model species, as RNA-seq data from the mouse were previously processed and analyzed in the laboratory.

We selected rat (*Rattus norvegicus*), cow (*Bos taurus*), pig (*Sus scrofa*), marmoset (*Callithrix jacchus*), rhesus macaque (*Macaca mulatta*) and human (*Homo sapiens*). For rat we found RNA-seq datasets of oocytes and somatic tissue, but not of embryos or placenta. For cow and pig, we found datasets of oocytes, pre-implantation and post-implantation embryos and somatic tissue, but not of placenta. For marmoset, only the datasets of pre-implantation embryos and somatic tissues were available, while for rhesus macaque there were datasets of the oocytes, pre-implantation embryos, placenta and somatic tissues. For human, we found datasets of all requested developmental stages - oocytes, embryonic datasets, placenta and somatic tissues. The summary of the datasets used for the analysis is in the (Figure 3), the detailed description of the datasets including the precise developmental stages, GEO accession codes and references is in the Supplementary Table 1.

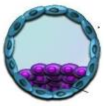
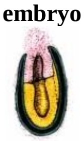
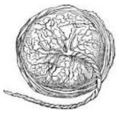
	Oocyte 	Pre - implantation embryo 	Post-implantation embryo 	Placenta 	Somatic Tissues
 Rat	✓	×	×	×	✓
 Cow	✓	✓	✓	×	✓
 Pig	✓	✓	✓	×	✓
 Marmoset	×	✓	×	×	✓
 Rhesus macaque	✓	✓	×	✓	✓
 Human	✓	✓	✓	✓	✓

Figure 3. *Summary of the datasets used for the analysis.*

7.1.2 Processing of the datasets

RNA-seq datasets were trimmed to remove the adapters and low-quality bases, their quality was checked and the data were mapped. As expected, because all the datasets were already published, the quality of the trimmed data judged by FASTQC output was appropriate, and mapping did not reveal any issues caused by either contamination or too high duplication levels. Supplementary Table 2 shows the accession codes of the individual datasets in fq.gz form, total number of reads we got after trimming process, the alignment rate and the number of mapped reads.

7.2 De novo transcriptome assembly

To be able to annotate all transcripts within the imprinting loci across all developmental stages, we performed de novo transcriptome assembly using Cufflinks. If there was a high number of datasets for a particular stage in a particular species, we selected the datasets with the highest number of reads (Table 1). Afterwards, the assembled transcriptomes were merged together for each species, resulting in final transcriptomes in the format of gtf files, encompassing transcription events across all analyzed developmental stages. The numbers of transcripts in partially merged files (one gtf file for each source publication) and in the final merged transcriptome for each species are in the Table 2.

Table 1. Datasets used for the transcriptome assembly. *The table shows the replicates selected for de novo transcriptome for every dataset of each species.*

Reference	Species	Cell type	Replicates
Brindamour et al. 2018	rat	GV oocytes GV oocytes	all all
Carelli et al. 2018	rat	adult brain adult heart adult kidney adult liver	replicates 2, 4 replicates 1, 3 replicates 1, 2 replicates 2, 4
Tsai et al. 2018	pig	MII oocytes	replicates 1, 3, 4
Bernardo et al. 2018	pig	E6.5 ICM E10.5 Epi ERSE E12.5 Epi APE	all all all
Li et al. 2018	pig	adult brain adult heart adult kidney adult liver adult muscle	all all all all all
Reyes et al. 2015	cow	GV oocyte MII oocyte	replicates 1, 2 replicates 1, 3
Lavagi et al. 2018	cow	8C embryo 16C embryo	replicates 1_7, 4_2, 5_3, 6_6 replicates 2_2, 2_3, 2_8, 1_1
Bernardo et al. 2018	cow	E7.0 ICM E14.0 Epi ERSE E17.0 Epi APE	all all all
Fushan et al. 2015	cow	adult liver adult kidney adult brain	all all all
Liu et al. 2018	rhesus macaque	16C_embryo outer cell 16C_embryo inner cell	all all

		early_morula outer cell	replicates 1,2,3,5
		early_morula inner cell	replicates 2,3,4,5
		late_morula outer cell	replicates 1,2,4,5
		late_morula inner cell	replicates 1,2,3,4
		early_blastocyst outer cell	replicates 1,2,3,6
		early_blastocyst inner cell	replicates 11,19,20,22
		mid_blastocyst outer cell	replicates 1,2,4,5
		mid_blastocyst inner cell	replicates 14,17,20,22
		late_blastocyst inner cell	replicates 20,22,29,31
		hatched_blastocyst inner cell	replicates 26,27,28,30
Ruebel et al. 2018	rhesus macaque	GV oocytes	replicates 1,6,7
		MII oocytes	replicates 2,3,8
Dunn-Fletcher et al. unpublished	rhesus macaque	placenta	all
Chitwood et al. 2017	rhesus macaque	GV oocytes	all
		MI oocytes	all
		MII oocytes	replicates 1,3
		1C embryo	replicates 1,3
		2C embryo	all
		4C embryo	replicates 1,3
		8C embryo	all
		morula	replicates 1,2
		blastocyst	all
Wang et al. 2017	rhesus macaque	GV oocytes	all
		MII oocytes	all
		1C embryo	all
		2C embryo	all
		4C embryo	all
		8C embryo	all
		morula	all
		blastocyst	all
Carelli et al. 2018	rhesus macaque	adult brain	replicates 1,2
		adult heart	replicates 1,2
		adult kidney	replicates 1,3
		adult liver	replicates 2,3
Boroviak et al. 2018	marmoset	zygote	all
		4C embryo	replicates 1_2, 1_3, 1_4
		8C embryo	replicates 1_1, 1_7, 2_1
		compacted morula	replicates 1_2, 1_3, 1_7
		early ICM	replicates 1_3, 1_5, 1_6
		late ICM	replicates 1_2, 1_3, 1_7
Carelli et al. 2018	marmoset	adult brain	replicates 2,3
		adult heart	replicates 1,2
		adult kidney	replicates 1,2
		adult liver	replicates 2,3
Yan et al. 2013	human	oocyte	all
		zygote	all
		2C embryo	replicates 1_2, 2_2, 3_2

		4C embryo 8C embryo morula late blastocyst	replicates 2_1, 2_4, 3_4 replicates 1_4, 2_4, 2_5 replicates 1_1, 1_4, 2_4 replicates 1_1, 1_4, 1_9
Guo et al. 2014	human	postimplantation embryo	all
Wu et al. 2018	human	GV oocyte MII oocyte 2C embryo 4C embryo 8C embryo ICM	all all replicates 1,2 all all all
Dunn-Fletcher et al. 2018	human	placenta	all
Carelli et al. 2018	human	adult brain adult heart adult kidney adult liver	replicates 1,2 replicates 3,4 all replicates 2,3

Table 2. Total numbers of transcripts in the assembled transcriptomes. *mRNA counts are shown for partially merged transcriptome assemblies (one for each source publication), and final merged gtf files for each species.*

Annotation	gtf file	mRNA count
cow	cow_Bernardo_merged.gtf cow_Fushan_merged.gtf cow_Lavagi_merged.gtf cow_Reyes_merged.gtf cow_merged.gtf (*)	144599 63630 81842 98953 227760
human	human_Carelli_merged.gtf human_Dunn_merged.gtf human_Guo_merged.gtf human_Wu_merged.gtf human_Yan_merged.gtf human_merged.gtf (*)	192367 187744 164093 219835 226678 308232
marmoset	marmoset_Boroviak_merged.gtf marmoset_Carelli_merged.gtf marmoset_merged.gtf (*)	174875 164426 246347
pig	pig_Bernardo_merged.gtf pig_Li_merged.gtf pig_Tsai_merged.gtf pig_merged.gtf (*)	151270 92594 97459 220870
rat	rat_Brindamour_merged.gtf rat_Carelli_merged.gtf rat_merged.gtf (*)	82403 89923 127069
rhesus macaque	rhesus_Carelli_merged.gtf rhesus_Chitwood_merged.gtf rhesus_Dunn-Fletcher_merged.gtf rhesus_Liu_merged.gtf rhesus_Ruebel_merged.gtf rhesus_Wang_merged.gtf rhesus_merged.gtf (*)	121534 253841 75048 128128 152572 104720 388756

(*) final merged gft file for each species

7.3 Defining imprinted gene clusters

First, we made a comprehensive list of imprinted genes in mouse, including novel previously unannotated transcripts identified as imprinted in recent publications. We combined the imprinted genes listed in the imprinting database (http://www.geneimprint.com/site/genes-by-species) and newly identified imprinted genes (Andergassen et al., 2017; Inoue et al., 2017; Xu Wang, Soloway, & Clark, 2011), resulting in 151 imprinted genes organized in 52 clusters. We then defined the borders of mouse imprinted clusters by the first protein-coding genes with known function which are either demonstrated not to be imprinted, or with unknown imprinting status. The list of mouse imprinted genes, their imprinting status, their genomic coordinates, the coordinates of imprinted cluster they belong to and the names of borderline genes is in the Supplementary Table 3.

Afterwards, we identified potentially homologous regions in 5 of the analyzed species, namely rat, cow, pig, marmoset, and rhesus macaque, based on gene position within the genome. We checked if the mouse imprinted genes themselves have the homologous gene in tested species. In the majority of cases, protein-coding imprinted genes have homologous genes in other species, and imprinted clusters have the boundaries set by the homologous genes. In some cases, the order of the genes is reversed, or there is another gene between the imprinted genes and mouse borderline gene. In such case, we took the new genes as the new boundary, as we can expect that the region between the old and new boundary genes might not be homologous to the sequence within the imprinted cluster in mouse. This can be exemplified by *Nnat/Blcap* imprinted cluster with boundary genes *Src* and *Ctnnbl1* in mouse, *Adig* and *Ctnnbl1* in rat, and *Src* and *Ctnnbl1* again in cow. If there was no homologous gene for the whole imprinted cluster, the region between genes homologous to mouse boundary genes was considered, unless the genes overlapped in the tested species. The coordinates of homologous genes and regions in rat, cow, pig, marmoset, rhesus macaque and human are listed in the Supplementary table 3.

After the identification of potentially imprinted regions in five analyzed species, defined as genomic regions expected to be homologous to the mouse imprinted clusters, we used a custom Python script (Sylvia_1(final code).py) see (Appendix 1) to filter regions based on chromosome and start -end of bases, for filter the final merged transcriptome gtf files, ouput gtf files (Table 3) contain only transcripts within the potentially imprinted regions for further analysis.

Table 3. Numbers of transcripts within potentially imprinted regions. mRNA counts of gtf files after filtering to contain only mRNAs within potentially imprinted regions.

Annotation	gtf file	mRNA Count
Cow	cow_merged_exons_filtered.gtf	3781
Marmoset	marmoset_merged_exons_filtered.gtf	723
Pig	pig_merged_exons_filtered.gtf	3328
Rat	rat_merged_exons_filtered.gtf	2136
Rhesus macaque	rhesus_merged_exons_filtered.gtf	6878

Additionally, we marked the imprinting status of each gene as either paternally expressed, maternally expression, isoform dependent imprinting, tissue dependent imprinting. For mouse, the information was available for all genes, but for other species except human the information was limited. If a gene is maternally- or paternally- expressed in mouse and it was confirmed to be imprinted in other species, it is expressed from the same parental allele as in mouse. The only exception appeared to be *Sfmbt2*, as the imprinted gene database Geneimprint (http://www.geneimprint.com/site/genes-by-species) lists it as paternally-expressed in the mouse, but maternally-expressed in rat. However, the search in the original research literature revealed that it is also paternally-expressed in rat (Q. Wang et al., 2011). For the remaining genes, their imprinting status in other species is either unknown, or their expression is confirmed to be biallelic, from both paternal and maternal allele. The complete information about imprinting status of all genes in studied species is in the Supplementary table 3, information for selected genes with known imprinting status in at least one more species in addition to mouse and human is in the Table 4.

Table 4. *Shows the genes with known imprinting status in mouse, cow and rat species*

Gene	mouse	rat	cow
Sfmbt2	P	P	B
Nnat	P	-	P
Gnas	B	-	M
Peg10	P	-	P
Asb4	M	-	B
Mest	P	-	P
Copg2	M	-	B
Nap1l5	P	-	P
Zim2	M	-	B
Peg3	P	-	P
Usp29	P	-	P
Zfp264	P	-	B
Snrpn	P	-	P
H19	M	M	M
Igf2	P	P	P
Ascl2	M	-	M
Cd81	M	-	B
Tssc4	M	-	M
Phlda2	M	-	M
Osbpl5	M	-	B
Rasgrf1	P	P	-
Plagl1	P	-	P
Meg3	M	-	M
Htr2a	M	-	B
Slc38a4	P	-	B
Igf2r	M	M	M
Impact	P	P	-

7.4 Analysis of potential methylation status of mouse gDMRs in selected species

With one exception, all known maternally-methylated gDMRs are overlapping CpG-rich promoters and would be expected to be unmethylated. It was previously shown in mouse that all such promoters are inactive in the oocytes, and that they are localized inside active transcriptional units either of a different gene or provided by an upstream alternative promoter or the same gene. Therefore, there is transcription going through all maternally-methylated gDMRs in mouse oocytes, and for gDMRs with no known overlapping gene, novel oocyte-specific transcript or upstream promoter was identified (Veselovska et al., 2015). In addition, it was shown that DNA methylation in fully-grown oocytes colocalizes with gene bodies of active genes, explaining why gDMRs become (Shirane et al., 2013; Veselovska et al., 2015) methylated. The same pattern was observed in human oocytes (Okae et al., 2014), therefore, it is expected that such pattern is conserved in all placental mammals.

We hypothesized that if imprinting of a particular gene cluster is conserved, the position of the gDMR is conserved too, overlapping the promoter of a gene homologous to the mouse gene with gDMR at its promoter. We therefore analysed whether gDMRs are overlapped by active transcribed genes, suggesting they become methylated in the oocytes and therefore are potentially imprinted. We also hypothesized that the high CpG content of gDMRs is likely to be conserved, protecting these regions from gaining methylation on the paternal allele during spermatogenesis. For each maternally-methylated gDMR in mouse that was shown to control imprinted expression of at least one gene (19 out of 20) we predicted its imprinting status in rat (as a mammalian species evolutionary close to mouse) and cow (a mammalian species more evolutionary distant from mouse), based on whether the presumable gDMR was overlapped by a transcript, either in the Ensembl transcriptome annotation, or in or de novo assembled transcriptome, and whether the transcript is expressed in the oocytes with RPKM or FPKM higher than 0.2.

The results of methylation status predictions are summarized in the Table 5 and Table 6. We classified the predictions of methylation status in 5 categories: unmethylated (if the predicted position of gDMR at the promoter of a homologous gene was clear, and we did not find any overlapping transcript), likely unmethylated (if the predicted position of gDMR at the promoter of a homologous gene was clear, we found an overlapping transcript, but it was not expressed in the oocytes), inconclusive (when we were not able to predict the position of the gDMR in cases when the gene which promoter it should overlap did not have an annotated homologue, or when the identification of the alternative promoter overlapping gDMR was not obvious), likely methylated (when the predicted position of gDMR was not clear based on the Ensembl transcriptome annotation, but our de novo transcriptome assembly very likely placed it inside a transcript expressed in the oocyte) and methylated (the predicted position of gDMR at the promoter of a homologous gene was clear, and there was an overlapping transcript expressed in the oocytes).

In rat, one gDMR was predicted as unmethylated (*Slc38a4*), two as inconclusive (promoter of *Mcts2* and alternative promoter of *Grb10*), nine as likely methylated and seven as methylated. In the case of *Mcts2*, it does not have an annotated homologue in rat, and we did not see any potential *Mcts2* transcript in our de novo transcriptome assembly, thus we were not able to predict the position of gDMR that should overlap the promoter of *Mcts2*. In the case of *Grb10*, we were not able to conclusively determine the alternative promoter which should overlap gDMR. Overall, 16 out of 19 gDMRs are predicted to become methylated in the oocytes, and therefore they are likely to be imprinted too. The imprinting status was experimentally tested and confirmed only for three maternally-methylated gDMRs or genes they should regulate (*Kcnq1/Igf2* imprinted cluster, *Igfr2* imprinted cluster and *Impact*). For these three gDMRs, we predicted the methylation status as likely methylated. Therefore, we predicted that further 13 gDMRs with unknown imprinting status might be imprinted in rat. One of the examples is *Nnat/Blcap* gDMR which is predicted to overlap *Nnat* promoter. In the rat Ensembl transcriptome annotation, there is no transcription overlapping *Nnat* promoter, but using our de novo assembled transcriptome we observed that there is an upstream alternative promoter of Blcap expressed in the oocytes, providing the transcription through the *Nnat* promoter (Figure 4).

In cow, one gDMR was predicted to be unmethylated (SGCE/PEG10), six were inconclusive, nine likely methylated and three methylated. Imprinting status is known for at least some genes in eleven imprinted clusters regulated by a known gDMR in mouse. In four clusters (SGCE/PEG10 cluster, MEST cluster, PEG3 cluster and KCNQ1/IGF2 cluster), some genes were demonstrated to be imprinted, while others were demonstrated to be biallelically expressed, suggesting that the expression regulation by imprinted gDMR might differ from mouse. In our predictions, we concluded that gDMR of SGCE/PEG10 is unmethylated, as there was no transcription overlapping the bidirectional promoter of SGCE and PEG10 (Figure 5) where we predicted the position of gDMR based on the homology with mouse, suggesting that maybe the position of gDMR is not conserved between mouse and cow. The methylation status of PEG3 gDMR was inconclusive, while we predicted that KCNQ1/IGF2 and MEST gDMRs are likely methylated. For the other imprinted clusters with experimentally determined imprinting status, we predicted for two (NNAT/BLCAP and NAP1L5) that they are methylated and for four (SNRPN, PLAGL1, IGF2R and one of the GNAS gDMRs) that they are likely methylated. SLC38A4 was shown to be biallelic in cow, and our prediction for this gDMR was inconclusive. In addition to the gene clusters with experimentally determined imprinting status, we predicted the methylation and therefore potential imprinting for further four gDMRs.

Table 5. Methylation status predictions of mouse maternally-methylated gDMRs in rat and cow. Table shows the position of gDMRs based on the overlap with mouse promoter, overlapping gene/transcript in mouse, the presence of overlapping transcript and its oocyte expression in rat and cow, and the predicted methylation status.

		rat			cow		
Imprinted gDMRs		**overlapping transcript**	**oocyte expression**	**predicted methylation status**	**overlapping transcript**	**oocyte expression**	**predicted methylation status**
Promoter of Mcts2	Overlapped by H13	Yes	Yes	Inconclusive	Yes	Yes	Inconclusive
Promoter of Nnat	Overlapped by Blcap	Yes	Yes	Methylated	Yes	Yes	Methylated
Promoter of Nespas and alternative promoter of Gnas	Overlapped by Gnas	Yes	Yes	Likely methylated	Yes	Yes	Likely methylated
Alternative promoter of Gnas	Overlapped by Gnas	Yes	Yes	Methylated	Yes	Yes	Inconclusive
Promoter of Sgce and Peg10	Overlapped by novel upstream promoter of Peg10.	Yes	Yes	Methylated	No	No	Unmethylated
Promoter of Peg3 and Usp29	Overlapped by novel upstream promoter of Peg3	Yes	Yes	Methylated	NA	NA	Inconclusive
Alternative promoter of Snrpn	Overlapped by Snrpn	Yes	Yes	Likely Methylated	Yes	Yes	Likely methylated
Alternative promoter of Inpp5f	Overlapped by Inpp5f	Yes	Yes	Likely Methylated	NA	NA	Inconclusive
Promoter of Kcnq1ot1	Overlapped by Kcnq1	Yes	Yes	Likely methylated	Yes	Yes	Likely methylated
Main promoter of Plagl1	Overlapped by upstream oocyte-specific Plagl1 promoter	Yes	Yes	Methylated	Yes	Yes	Likely methylated
Alternative promoter of Grb10	Overlapped by Grb10	Yes	Yes	Inconclusive	Yes	Yes	Inconclusive
Promoter of Zrsr1	Overlapped by Commd1	Yes	Yes	Methylated	Yes	Yes	Methylated
Promoter of Peg13	Overlapped by Trappc9.	Yes	Yes	Likely methylated	Yes	Yes	Likely methylated
Promoter of Slc38a4	Overlapped by novel transcript(s) identified in mouse oocytes	No	No	Unmethylated	NA	NA	Inconclusive
Promoter of Airn	Overlapped by Igf2r	Yes	Yes	Likely methylated	Yes	Yes	Likely methylated
Impact promoter	Overlapped by upstream novel promoter of Impact and/or novel gene	Yes	Yes	Likely methylated	Yes	Yes	Likely methylated
Promoter of poorly described transcript	Overlapped by Mest	Yes	Yes	Likely Methylated	Yes	Yes	Likely methylated
Promoter of Nap1l5	Overlapped by Herc3	Yes	Yes	Methylated	Yes	Yes	Methylated
-	Overlapped by Cdh15	Yes	Yes	Likely Methylated	Yes	Yes	Likely methylated

Table 6. Prediction methylation status with five different categories.

Predicted methylation status	description
Unmethylated	if there is no transcript going through it
Likely unmethylated	there is transcript, but it does not seem to be expressed in the oocytes
Methylated	there is transcription going through it
Inconclusive	when it is difficult to determine where the gDMR should be in that species
Likely methylated	when promoter overlapping gDMR is not in the official annotation but appears to be present in de novo transcriptome assembly

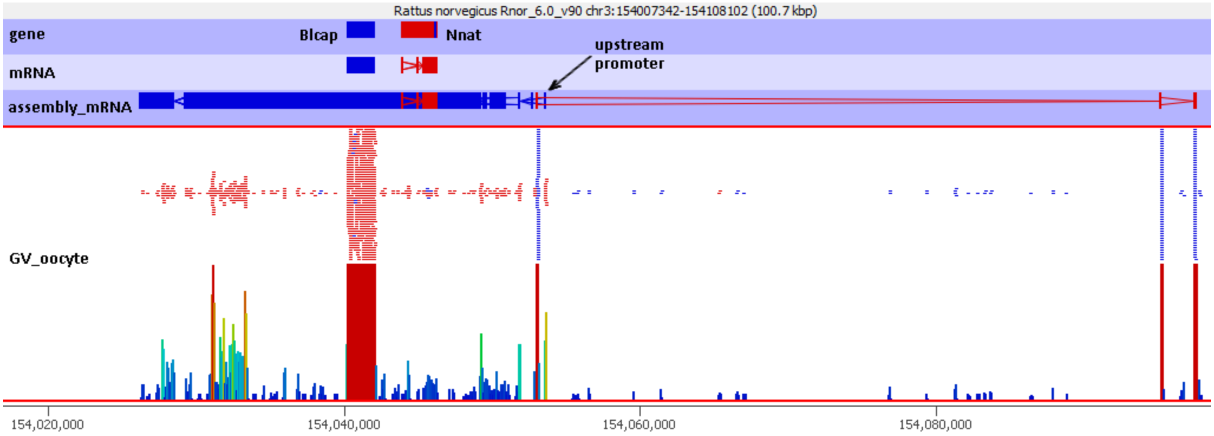


Figure 4. Blcap/Nnat cluster in rat. The gDMR is expected to overlap the promoter of Nnat. The first two lines show Ensembl annotation, third line is our assembled transcriptome showing that there is an upstream promoter of Blcap (indicated by an arrow) providing transcription through the promoter of Nnat. The fourth line shows RNA-seq reads in rat GV oocytes and their quantification, demonstrating that the gene is expressed in the oocytes.

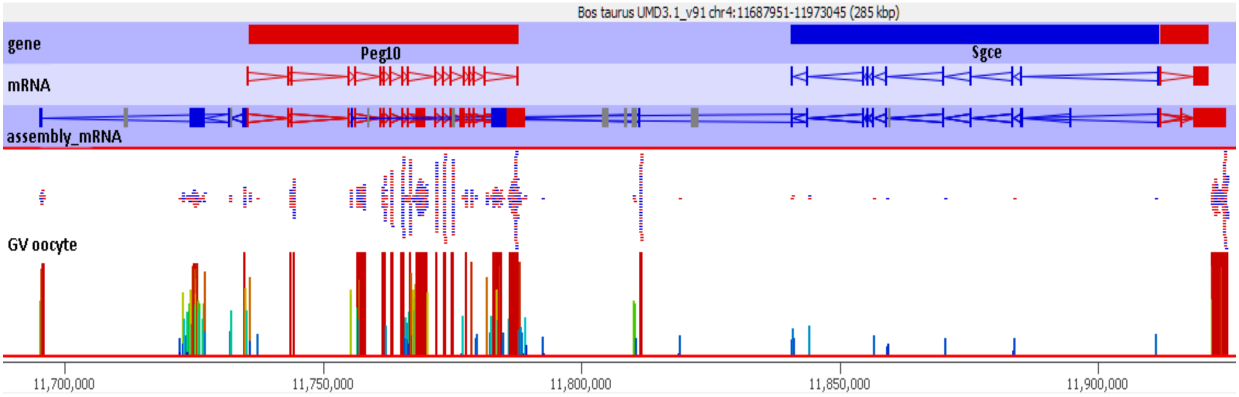


Figure 5. Promoter of Sgce and Peg10 in cow, the first two lines show Ensembl annotation, third line is our assembled transcriptome showing that there is no upstream promoter no transcription overlapping the bidirectional promoter. The fourth line shows RNA-seq reads in cow GV oocytes and their quantification, demonstrating that the gene is unmethylated.

7.5. Expression analysis of transcripts in potentially imprinted regions

We wanted to analyze the expression profiles of known and novel transcripts within the regions homologous to mouse imprinted regions in rat and cow to see whether these transcripts developmentally regulated and specific for certain developmental stage and period, or rather expressed at the similar level in all datasets.

Expression of transcripts within the potentially imprinted regions was quantified using Cufflinks v2.2.1 (Supplementary Table 4 and 5), the expression values were averaged per developmental stage in Microsoft Office Excel (v14.0, 32bit) and further modified prior to clustering analysis. Namely, we removed all the transcripts with expression lower than FPKM or RPKM of 0.1 in all datasets, we log transformed the values (log (value+0.1), base 2), we quantified the mean from log values across all the stages and then we subtracted the mean from each log value, giving us the relative expression of each transcript which serves as an input for clustering analysis and heatmap visualisation (cow_expression.txt, rat_expression.txt)(see CD). For cow, we had datasets from germinal vesicle (GV) and meiosis II (MII) oocytes, 8-cell embryos, 16-cell embryos, embryonic day 7.0 (E7.0) inner cell mass (ICM), E14.0 and E17.0 epiblast (Epi) and three somatic tissues (brain, kidney and liver). For rat, we had GV oocytes and four somatic tissues (brain, heart, kidney, liver).

Hierarchical clustering analysis clusters together genes with similar expression profile across datasets. Heatmap is used for visualisation of expression changes of clustered genes. Clustering and heatmap were performed in R (v 3.5.3, 64bit) using a custom script we created (heatmap.R, see Appendix 4) compiling hclust and heatmap.2 functions, with cow_expression.txt and rat_expression.txt files as an input. After visual inspection of the heatmaps (figure 6 for rat and figure 7 for cow), we decided to divide the rat transcript into 12 main clusters, and cow transcripts in 15 main clusters (the command is within heatmap.R script), and we exported files containing the number of cluster each transcript is assigned to (rat_cluster.txt, cow_cluster.txt)(see CD). The numbers of transcripts in each cluster are in the Table 7 for rat and Table 8 for cow. For six clusters with the highest numbers of transcripts we quantified average relative expression values (Table 9 for rat and Table 10 for cow) and plotted the expression profiles (Figure 8 for rat and Figure 9 for cow).

Heatmaps and relative expression profiles of individual clusters show that almost all transcripts in both rat and cow are specific for a certain developmental stage or period. In rat, heatmap shows that almost all transcripts are either expressed in the oocytes, or in or more somatic tissues. This agrees with the analysis of individual clusters, where the expression profile of cluster 1 with the highest number of transcripts (500 transcripts) shows that the expression of transcripts is high in the GV oocytes and low in somatic tissues, while in the remaining five clusters the expression is low in GV oocytes and high in one or more of the somatic tissues.

In cow, the heatmap shows that there are several main patterns of transcripts expression. They are either expressed only in the oocytes, or in the oocytes and embryos, or in specific embryonic stages, or in one or more somatic tissues. This agrees with the expression profiles of six clusters with the highest numbers of transcripts. Clusters 4 and 15 contain transcripts expressed specifically in the oocytes (with the smaller peak of expression in 16-cell embryos in cluster 15), transcripts in cluster 3 are expressed in the oocytes and early embryos (8- and 16-cell stage), transcripts in clusters 12 and 14 are expressed in specific embryonic stages and cluster 10 contain transcripts expressed only in somatic tissues. Overall, we can conclude that transcripts in the potential imprinted clusters in cow and rat are developmentally regulated and they appear to be divided into transcripts expressed in oocytes and/or embryos, and those expressed in somatic tissues.

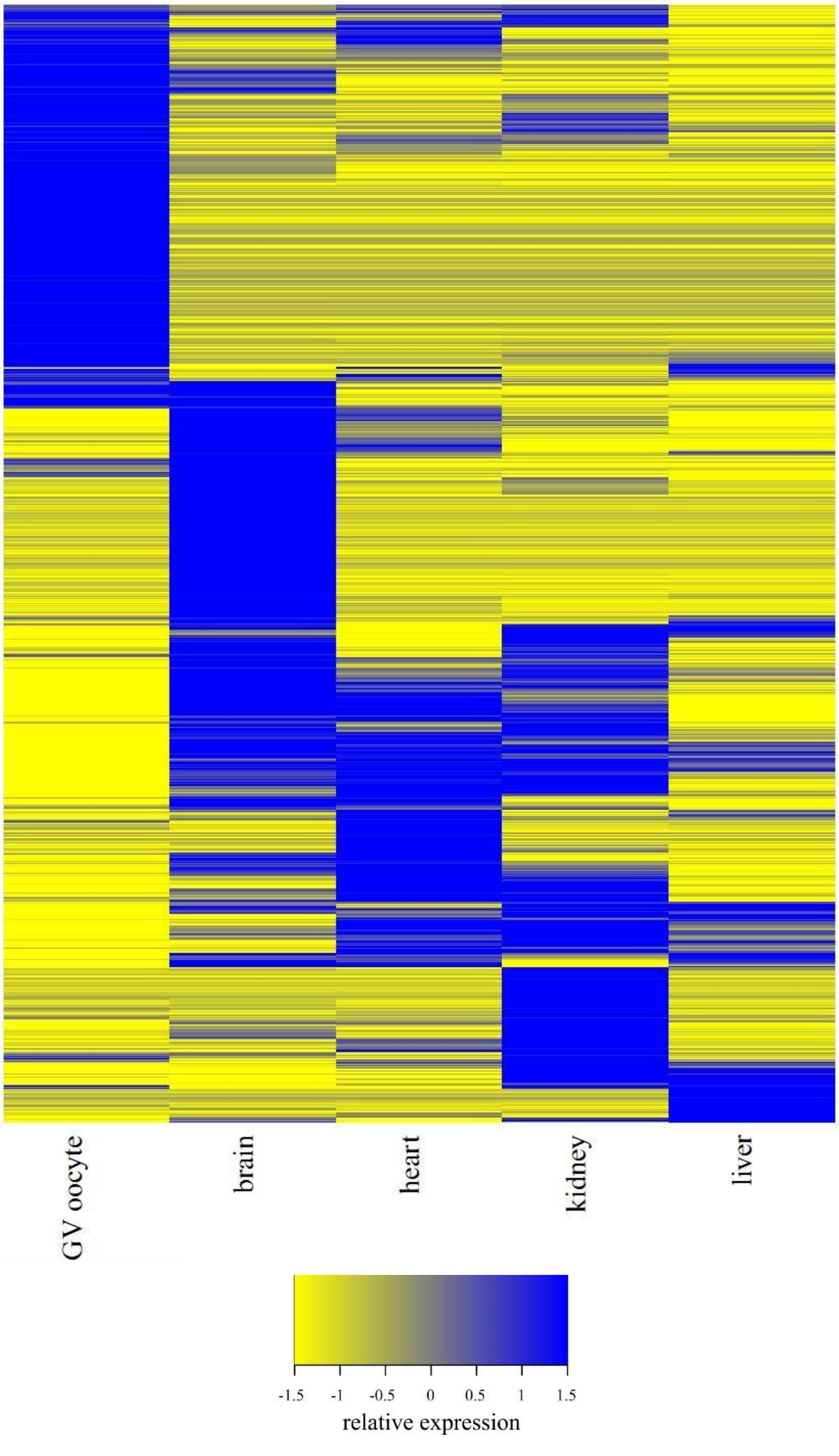


Figure 6. Hierarchical clustering of rat transcripts within potentially imprinted loci. The heatmap shows relative expression as mean-centred log transformed FPKM values. Blue is for high expression and yellow is for low expression. Developmental stages are: GV oocytes (GV), and somatic tissues brain, heart, kidney and liver.

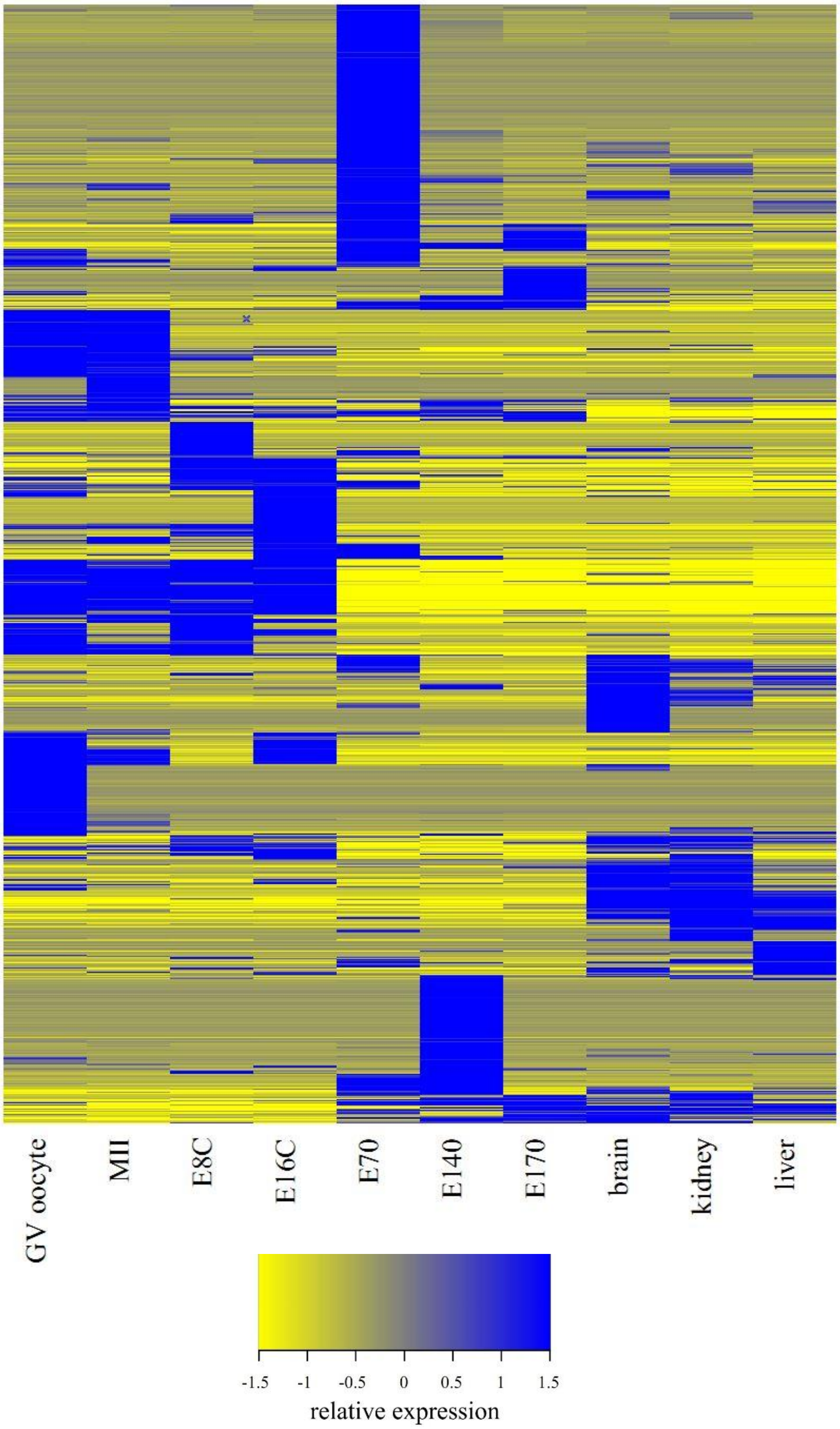


Figure 7. Hierarchical clustering of cow transcripts within potentially imprinted loci. The heatmap shows relative expression as mean-centred log transformed FPKM values. Blue is for high expression and yellow is for low expression. Developmental stages are: GV oocytes (GV), MII oocytes (MII), 8-cell stage embryos (E8C), 16-cell stage embryos (E16C), E7.0 ICM (E70), E14.0 Epi (E140), E17.0 Epi (E170), and somatic tissues brain, kidney and liver

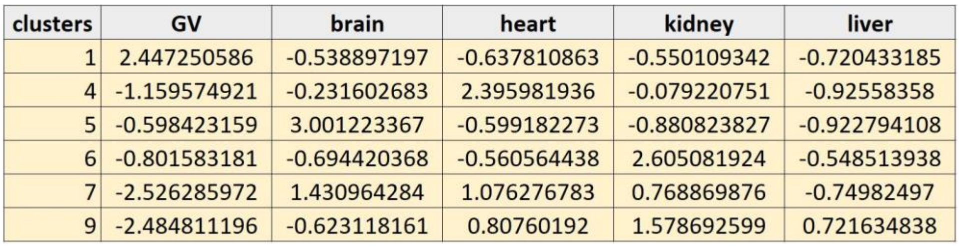
Table 7. Numbers of transcripts in each cluster for rat.
Highlighted in red are the six clusters with the highest numbers of transcripts.

cluster	Number of transcripts
1	500
2	34
3	50
4	142
5	377
6	166
7	236
8	27
9	80
10	22
11	22
12	74

Table 8. Numbers of transcripts in each cluster for cow.
Highlighted in red are the six clusters with the highest numbers of transcripts.

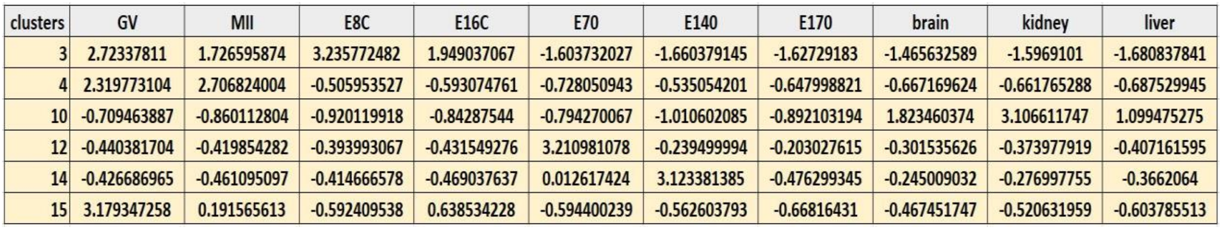
cluster	Number of transcripts
1	48
2	71
3	281
4	285
5	85
6	129
7	184
8	56
9	230
10	243
11	235
12	727
13	100
14	360
15	308

Table 9. Average relative expression values in individual clusters in rat. Showing the relative expression values in each developmental stage for the six clusters with the highest numbers of transcripts. Developmental stages are: GV oocytes (GV), and somatic tissues brain, heart, kidney and liver.

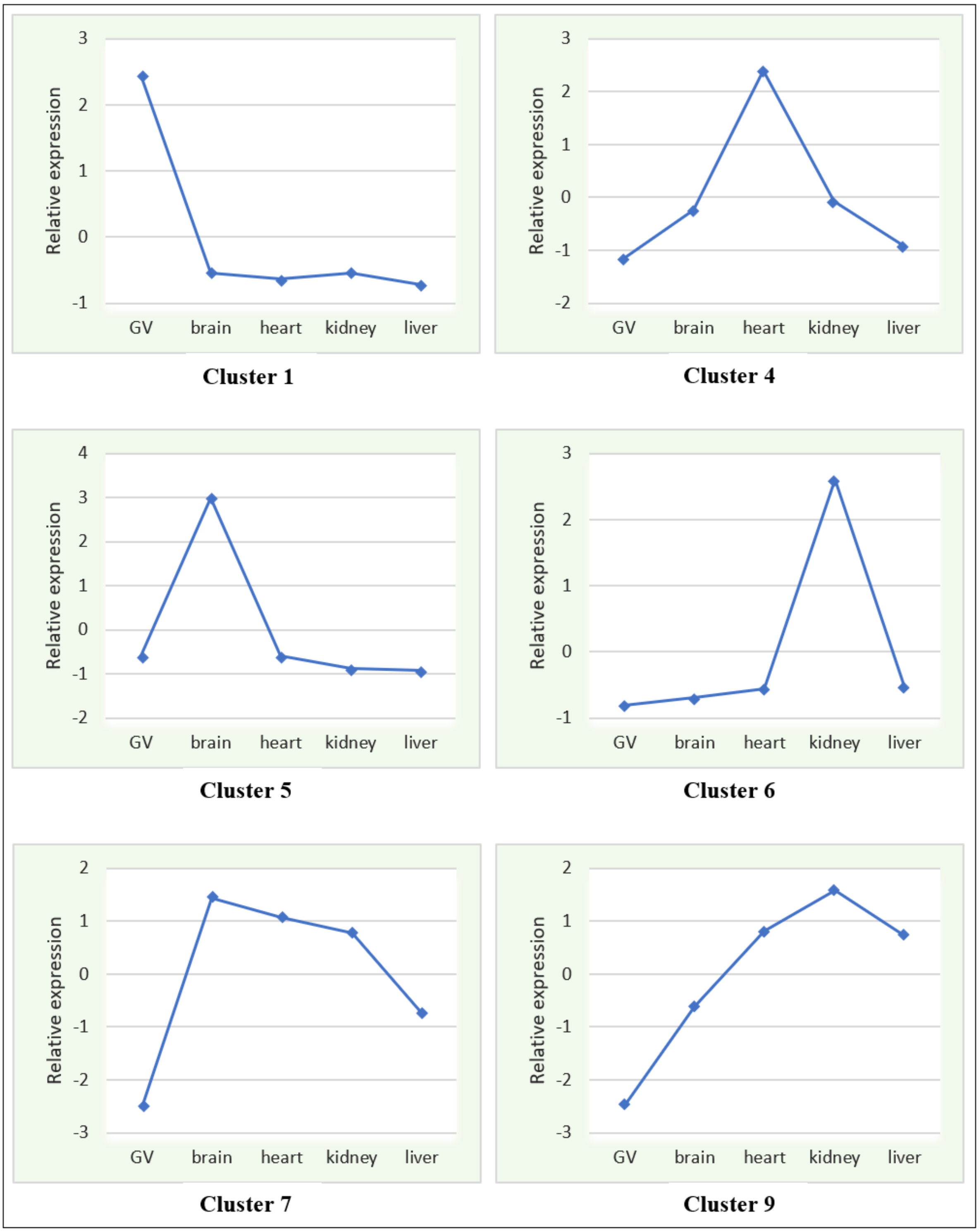


clusters	GV	brain	heart	kidney	liver
1	2.447250586	-0.538897197	-0.637810863	-0.550109342	-0.720433185
4	-1.159574921	-0.231602683	2.395981936	-0.079220751	-0.92558358
5	-0.598423159	3.001223367	-0.599182273	-0.880823827	-0.922794108
6	-0.801583181	-0.694420368	-0.560564438	2.605081924	-0.548513938
7	-2.526285972	1.430964284	1.076276783	0.768869876	-0.74982497
9	-2.484811196	-0.623118161	0.80760192	1.578692599	0.721634838

Table 10. Average relative expression values in individual clusters in cow. Showing the relative expression values in each developmental stage for the six clusters with the highest numbers of transcripts. Developmental stages are: GV oocytes (GV), MII oocytes (MII), 8-cell stage embryos (E8C), 16-cell stage embryos (E16C), E7.0 ICM (E70), E14.0 Epi (E140), E17.0 Epi (E170), and somatic tissues brain, kidney and liver



clusters	GV	MII	E8C	E16C	E70	E140	E170	brain	kidney	liver
3	2.72337811	1.726595874	3.235772482	1.949037067	-1.603732027	-1.660379145	-1.62729183	-1.465632589	-1.5969101	-1.680837841
4	2.319773104	2.706824004	-0.505953527	-0.593074761	-0.728050943	-0.535054201	-0.647998821	-0.667169624	-0.661765288	-0.687529945
10	-0.709463887	-0.860112804	-0.920119918	-0.84287544	-0.794270067	-1.010602085	-0.892103194	1.823460374	3.106611747	1.099475275
12	-0.440381704	-0.419854282	-0.393993067	-0.431549276	3.210981078	-0.239499994	-0.203027615	-0.301535626	-0.373977919	-0.407161595
14	-0.426686965	-0.461095097	-0.414666578	-0.469037637	0.012617424	3.123381385	-0.476299345	-0.245009032	-0.276997755	-0.3662064
15	3.179347258	0.191565613	-0.592409538	0.638534228	-0.594400239	-0.562603793	-0.66816431	-0.467451747	-0.520631959	-0.603785513



Cluster 1

Cluster 4

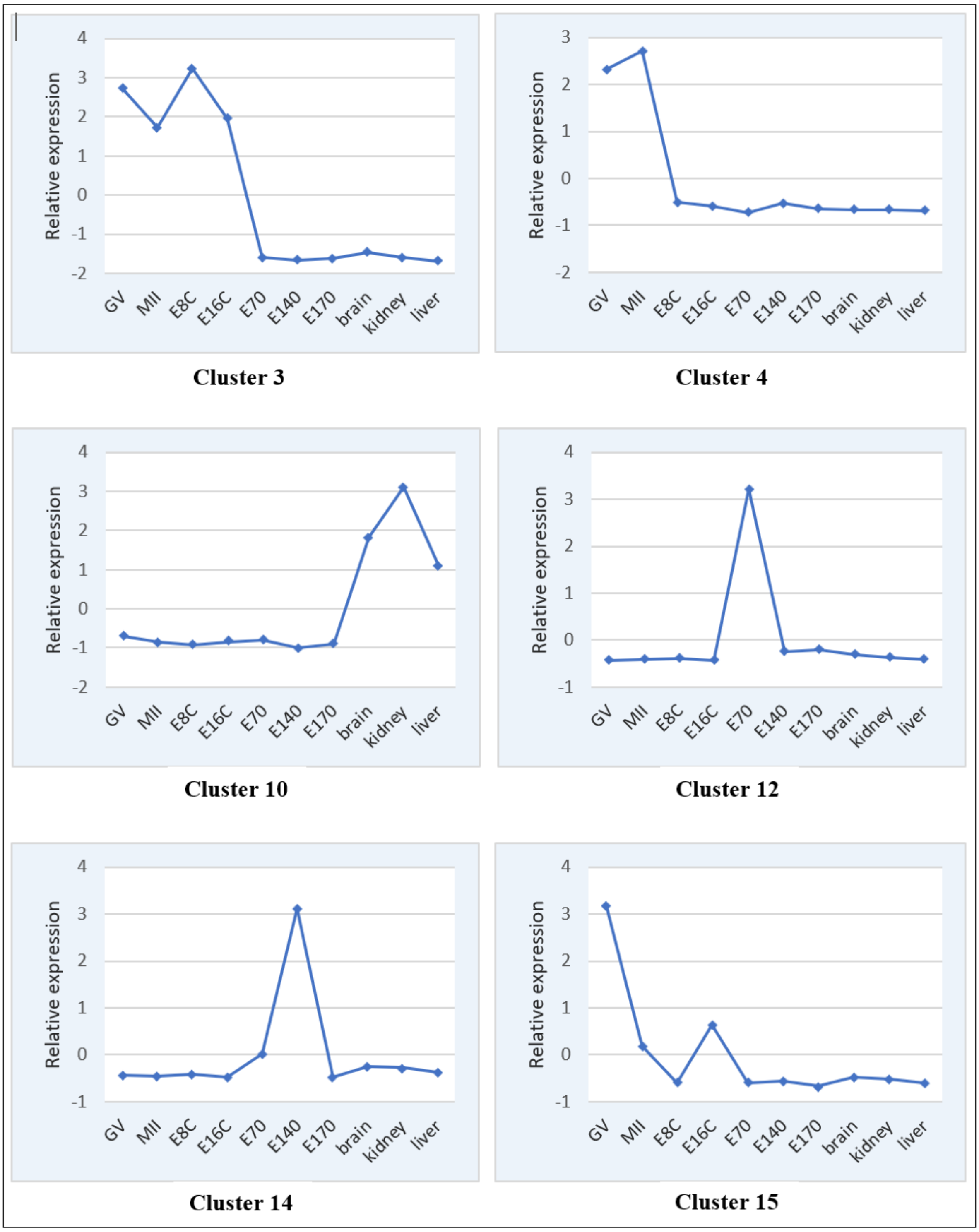
Cluster 5

Cluster 6

Cluster 7

Cluster 9

Figure 8. Average relative expression changes in each cluster in rat. Showing the average relative expression profiles in each developmental stage for the six clusters with the highest numbers of transcripts. Developmental stages are: GV oocytes (GV), and somatic tissues brain, heart, kidney and liver.



Cluster 3

Cluster 4

Cluster 10

Cluster 12

Cluster 14

Cluster 15

Figure 9. Average relative expression changes in each cluster in cow. Showing the average relative expression profiles in each developmental stage for the six clusters with the highest numbers of transcripts. Developmental stages are: GV oocytes (GV), MII oocytes (MII), 8-cell stage embryos (E8C), 16-cell stage embryos (E16C), E7.0 ICM (E70), E14.0 Epi (E140), E17.0 Epi (E170), and somatic tissues brain, kidney and liver.

7.6 Sequence analysis of promoters in potentially imprinted regions

We wanted to identify potential transcription factors driving the expression of transcripts within the potential imprinted regions in rat and cow through the analysis of common sequence motifs in the promoters of transcripts that can serve as binding sites for transcription factors. First, we wanted to remove transcripts not suitable for the promoter sequence analysis, i.e. expressed transposable elements that do not act as promoters of longer transcripts, and transcripts without strand specificity, in which we cannot determine on which end the promoter is.

To remove expressed transposable elements, we downloaded the annotation for all repetitive elements from UCSC genome browser (https://genome.ucsc.edu/) and using % coverage quantification tool within Seqmonk, we quantified what proportion of each transcript in our assembled transcriptome filtered to contain only imprinted regions (rat_merged_exons_filtered.gtf and cow_merged_exons_filtered.gtf) is covered by repetitive element encoded on the same DNA strand. It should be noted that transposable elements are the majority of annotated repetitive elements. We wanted to remove all transcripts overlapped by repetitive elements by more than 50%, but only if they are monoexonic, as we considered the transcripts with more exons as independent transcripts, not just expressed transposable elements. To achieve that, we made a python script (Sylvia_finalcode2.py, see Appendix 2) which takes a given list of transcript names (in our case all transcripts with >50% overlap by same strand repetitive elements), check if those transcripts have one or more exons, and removes them from a given gtf file if they have only one exon. We used gtf files rat_merged_exons_filtered.gtf and cow_merged_exons_filtered.gtf as an input for this script, and the output files were rat_merged_exons_filtering_subset.gtf and.

The transcriptome of rat contains only transcripts with known strand specificity, however, the RNA-seq datasets for cow were often unstranded, and as a result the assembled transcriptome also contains transcripts without known strand specificity. We therefore applied an additional filtering step to cow transcriptome cow_merged_exons_filtering_subset.gtf to remove all transcripts without strand specificity, using a custom python script (sylvia_finalcode3.py, see Appendix 3). The output file was cow_merged_exons_filtered(+-).gtf.

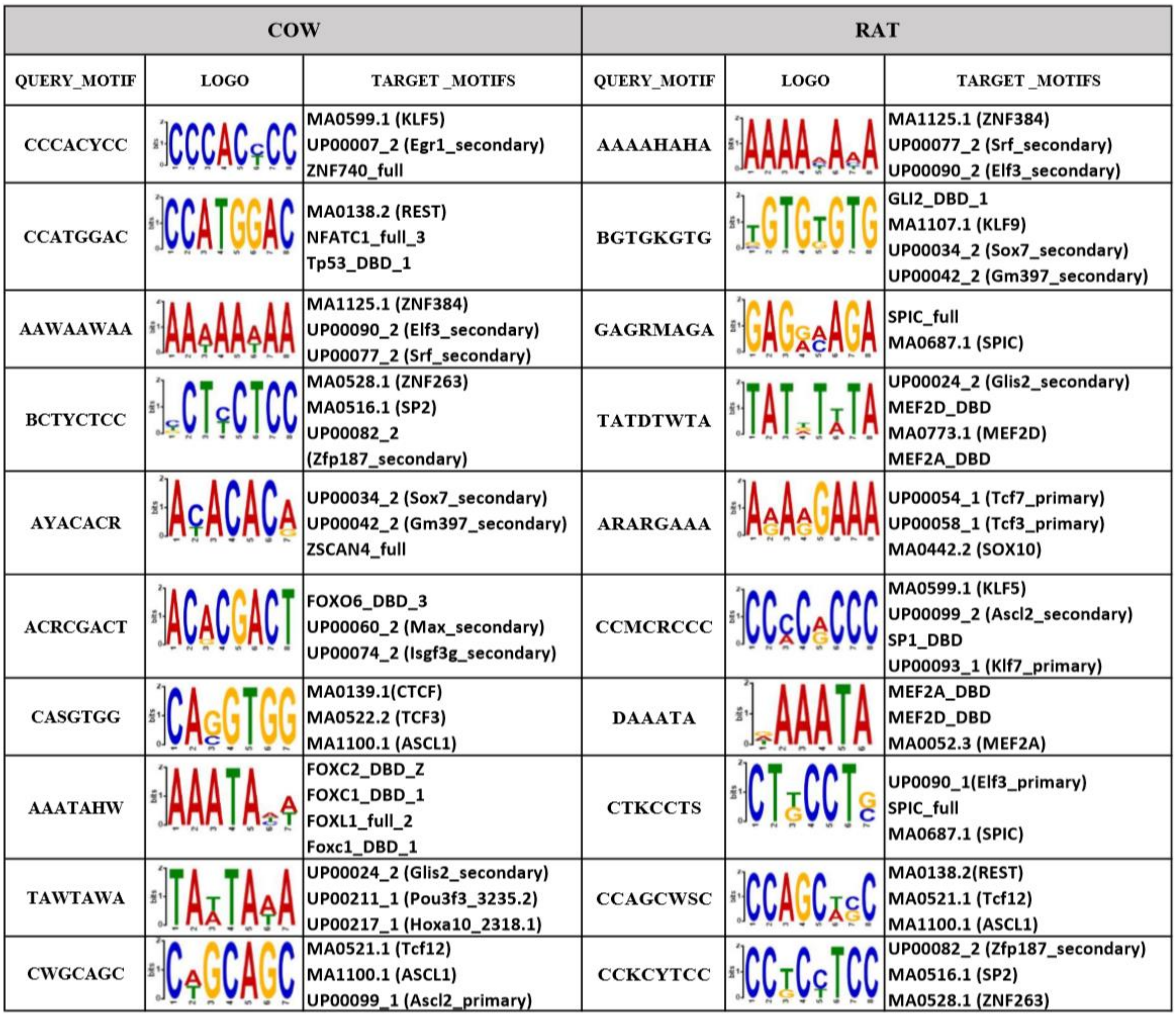
The transcripts in files rat_merged_exons_filtered.gtf for rat and cow_merged_exons_filtered(+-).gtf for cow were used for the promoter sequence analysis. We obtained the coordinates of promoters (including the names of transcripts the promoter belongs to) using program Seqmonk v1.43.0 as 1 kb regions (500 bp upstream and 500 bp downstream) around the transcriptional start site (TSS). We then extracted the sequences of these promoter regions in FASTA format from genomic sequences using a python script previously made in the laboratory (non_gtf_search_tool.py, Supplementary table 6). This script requires following files as an input: names of the regions (in our case names of the promoters) with genomic coordinates (cow_prom_coor.txt, rat_prom_coor.txt) (attached in CD), names of the regions from the previous file for which we want the sequence (in our case all of the promoters, cow_prom_names.txt, rat_prom_names.txt) (attached in CD) , and genomic sequence split in individual chromosomes (Rnor_6.0 genome for rat and UMD3.1 for cow, downloaded from Ensembl genome database https://www.ensembl.org/index.html). The output files contain sequences of individual promoters in FASTA format (output_rat_SAP.txt, output_cow_SAP.txt) (attached in CD).

These sequences were then analyzed to find enriched short sequence motifs. We submitted them to the online motif-based sequence analysis tool DREME within the tool suite MEME (Timothy L. Bailey, 2011) (http://meme-suite.org/tools/dreme). The output motifs from DREME were submitted into TOMTOM tool (Gupta et al., 2007) (http://meme-suite.org/doc/tomtom.html?man_type=web) which compares the motifs against the databases of transcription factors binding sites. The results are summarized in the Table 11. It shows ten motifs with highest significance enrichment as determined by DREME, and for each of these motifs the potential transcription factors with the highest significance of the similarity of their binding site to the motif (we show top three candidate transcription factors with the highest significance, if available).

We found that there are some potential transcription factor binding sites in common for rat and cow. For example, cow motif CCCACYCC and rat motif CCMCRCCC share sequence similarity and they both appear to serve as a binding site for KLF5. Cow motif CCATGGAC and rat motif CCAGCWSC are also similar and are likely to be binding sites of transcription factor REST. Cow motif AWWAAWAA and rat motif AAAAHAHA are similar and appear to be both binding sites for ZNF384, and secondary binding sites for SRF and ELF3. Cow motif BCTYCTCC and rat motif CCKCYTCC are also relatively similar and are likely to serve as binding sites for ZNF263 and SP2 and as secondary binding site for ZFP187. CWGCAGC motif from cow and CCAGCWSC appear to be both binding sites of ASCL1 and TCF12.

Overall, we identified that promoters of transcripts in potentially imprinted regions appear to be regulated to some extent by the same transcription factors, for example KLF5, ZNF384, REST, SP2 and ASCL1, suggesting that they might be conserved transcription factors regulating the expression of at least some transcripts in the potentially imprinted regions across mammalian species.

Table 11. Sequence motifs analysis of promoters. Table shows the ten motifs with highest significance from DREME and three most significant associated transcription factors identified by TOMTOM



COW			RAT		
QUERY_MOTIF	LOGO	TARGET _MOTIFS	QUERY_MOTIF	LOGO	TARGET _MOTIFS
CCCACYCC		MA0599.1 (KLF5) UP00007_2 (Egr1_secondary) ZNF740_full	AAAAHAHA		MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00090_2 (Elf3_secondary)
CCATGGAC		MA0138.2 (REST) NFATC1_full_3 Tp53_DBD_1	BGTGKGTG		GLI2_DBD_1 MA1107.1 (KLF9) UP00034_2 (Sox7_secondary) UP00042_2 (Gm397_secondary)
AAWAAWAA		MA1125.1 (ZNF384) UP00090_2 (Elf3_secondary) UP00077_2 (Srf_secondary)	GAGRMAGA		SPIC_full MA0687.1 (SPIC)
BCTYCTCC		MA0528.1 (ZNF263) MA0516.1 (SP2) UP00082_2 (Zfp187_secondary)	TATDTWTA		UP00024_2 (Glis2_secondary) MEF2D_DBD MA0773.1 (MEF2D) MEF2A_DBD
AYACACR		UP00034_2 (Sox7_secondary) UP00042_2 (Gm397_secondary) ZSCAN4_full	ARARGAAA		UP00054_1 (Tcf7_primary) UP00058_1 (Tcf3_primary) MA0442.2 (SOX10)
ACRCGACT		FOXO6_DBD_3 UP00060_2 (Max_secondary) UP00074_2 (Isgf3g_secondary)	CCMCRCCC		MA0599.1 (KLF5) UP00099_2 (Ascl2_secondary) SP1_DBD UP00093_1 (Klf7_primary)
CASGTGG		MA0139.1(CTCF) MA0522.2 (TCF3) MA1100.1 (ASCL1)	DAAATA		MEF2A_DBD MEF2D_DBD MA0052.3 (MEF2A)
AAATAHW		FOXC2_DBD_2 FOXC1_DBD_1 FOXL1_full_2 Foxc1_DBD_1	CTKCCTS		UP0090_1(Elf3_primary) SPIC_full MA0687.1 (SPIC)
TAWTAWA		UP00024_2 (Glis2_secondary) UP00211_1 (Pou3f3_3235.2) UP00217_1 (Hoxa10_2318.1)	CCAGCWSC		MA0138.2(REST) MA0521.1 (Tcf12) MA1100.1 (ASCL1)
CWGCAGC		MA0521.1 (Tcf12) MA1100.1 (ASCL1) UP00099_1 (Ascl2_primary)	CCKCYTCC		UP00082_2 (Zfp187_secondary) MA0516.1 (SP2) MA0528.1 (ZNF263)

7. 7 Analysis of Repetitive Elements

In this analysis we wanted to find out how common is the phenomenon of transcripts employing transposable elements as their promoters/TSSs within the imprinted clusters. In the section 7.5 we identified that many transcripts appear to be specific for oocytes and/or embryos, and transposable elements were shown to be most active as gene promoters at these stages in mouse (Macfarlan et al., 2012; Veselovska et al., 2015). In addition, as many of the transposable elements are specific for mammalian species of families, they can be at least partially responsible for differences in potentially imprinted transcripts between species.

As an input for the analysis, we used files rat_merged_exons_filtering_subset.gtf for rat and cow_merged_exons_filtered(+-).gtf for cow, and repetitive elements annotation which were all described in section 7.6. Using program Seqmonk v1.43.0, we selected regions of first 200 bp from the TSS of each transcript, and identified how many of such regions are overlapped by repetitive elements encoded on the same strand. We exported the results into Microsoft Office Excel (v14.0, 32bit). Out of 1712 TSS regions in rat, 365 were overlapped by a same strand repetitive element, and out of 1400 TSS regions in cow, 290 were overlapped by a same strand repetitive element. Tables 12 and 13 and figures 10 and 11 show the numbers of TSS regions overlapped by individual categories of repetitive elements in rat and cow. Excluding simple repeats and other repetitive elements which are not transposable elements, ERVL-MaLR and ERVK are most commonly employed as promoters/TSSs in rat, while L1 and RTE-BovB transposable elements in cow. This suggests that transposable element-promoted transcripts differ between rat and cow and therefore transcripts within the potentially imprinted regions differ to some extent between mammalian species.

Table 12. Repetitive elements as TSSs in rat. Shows the numbers of TSSs overlapped by each category of repetitive elements.

Feature	Count of rat_TEs
Other	1
PIF-Harbinger	1
RTE-BovB	1
RTE-X	1
scRNA	1
snRNA	1
tRNA-RTE	1
hAT-Charlie	2
Unknown	3
ID	4
L2	4
MIR	6
Satellite	6
ERV1	8
B2	12
ERVL	13
Alu	18
Low_complexity	20
B4	21
L1	30
ERVK	55
Simple_repeat	75
ERVL-MaLR	81
Gran Total	**365**

Table 13. Repetitive elements as TSSs in cow. Shows the numbers of TSSs overlapped by each category of repetitive element

Feature	Count of cow_TEs
hAT-Blackjack	1
RTE-X	1
hAT-Tip100	1
hAT-Ac	1
5S-Deu-L2	1
DNA	1
LTR?	1
TcMar-Mariner	2
LTR	2
centr	3
hAT-Charlie	5
CR1	5
TcMar-Tigger	5
ERVL	7
ERVK	8
tRNA	8
ERVL-MaLR	12
L2	14
ERV1	14
Low_complexity	15
Core-RTE	16
MIR	18
RTE-BovB	21
L1	29
tRNA-Core-RTE	41
Simple_repeat	58
Gran Total	**290**

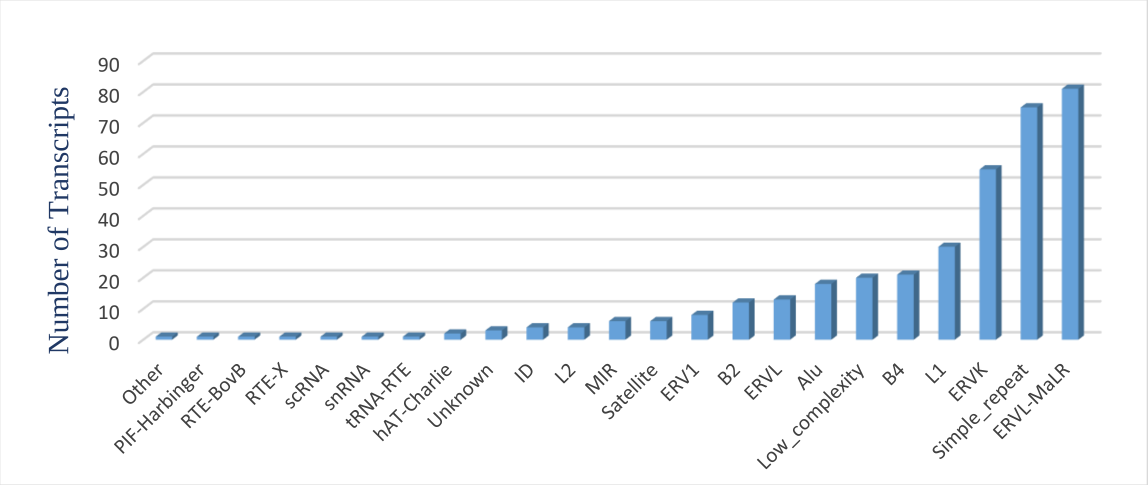


Figure 10. Repetitive elements as TSSs in rat. *Shows the numbers of TSSs overlapped by each category of repetitive elements*.

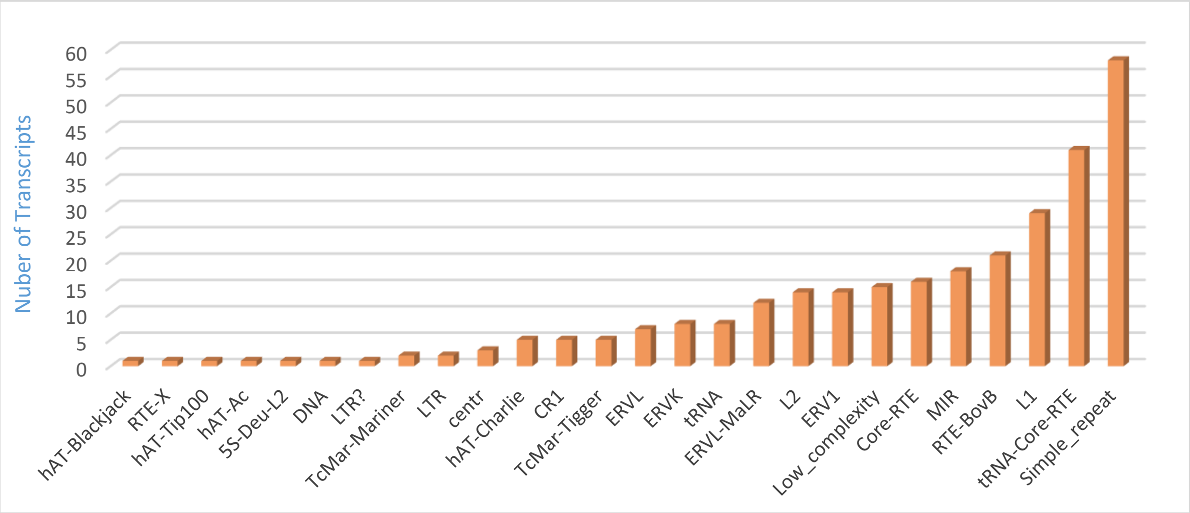


Figure 11. Repetitive elements as TSSs in cow. *Shows the numbers of TSSs overlapped by each category of repetitive elements*

7.8 Analysis of conservation of mouse novel transcripts in imprinted loci

Previous analysis in the laboratory identified candidate novel transcripts within mouse imprinted clusters for conservation analysis in other species. The transcripts were selected based on their novelty (for the first time identified in recent publications (Andergassen et al., 2017; Courtney W. Hanna et al., 2019), or by our laboratory using the data from these two publications), confirmed imprinting status in some and TSS overlap with transposable element in some. By manual inspection of our newly assembled transcriptomes of rat and cow, we aimed to identify whether the candidate mouse transcripts are conserved in rat and cow.

First set of candidate novel transcripts are in mouse *Jade1* imprinted region. There is a lncRNA *2400006E01Rik* that was identified to be imprinted (paternally-expressed) in mouse (Andergassen et al., 2017). De novo transcriptome assembly in our laboratory revealed that there is a number of transcripts overlapping *2400006E01Rik* locus (Figure 12). Three of these employ transposable element as a TSS – one RLTR31B-MM element from ERVK family that is specific for mouse, and two elements from the LTR-MaLR family, MTE2b (transposable element found in all rodents) and ORR1A2 (transposable element found in all murid rodents). Therefore, transcripts using MTE2b and ORR1A2 have the potential to be present also in rat *Jade1* region. However, no such transcripts (upstream of and antisense to *Jade1*/JADE1) are expressed in either rat or cow (Figures 13 and 14) according to our transcriptome assemblies. In both species, there are upstream same strand (relative to *Jade1*/JADE1) transcripts. In rat, this transcript starts from MaLR-LTR MTD element (Figure 15) which is present in rodents, but no such transcript is there in mouse, but in cow, the transcript does not employ a transposable element as TSS (Figure 16).

Second region with the candidate novel transcripts is the *Sfmbt2* region. *Sfmbt2* gene is imprinted (paternally expressed) in mouse and rat, but its expression is biallelic (not imprinted) in cow. In a recent publication, they identified novel imprinted transcript antisense to *Sfmbt2*, with novel unannotated downstream promoter starting from ERVK element RLTR11B(Andergassen et al., 2017), a transposable element also present in rat. In addition, de novo transcriptome assembly in our laboratory identified an alternative downstream promoter of that transcript starting from LTR-MaLR element MTEb (also present in rat), and several more antisense transcripts upstream of *Sfmbt2* of unknown imprinting status, of which one starts from LTR-MaLR MLT1J, present in all placental mammals (figure 17). Therefore, all three transcripts have the potential to be present in rat, and the last one also in cow. Nevertheless, there are no upstream antisense transcript in rat *Sfmbt2* region (figure 18), and in the cow genome such upstream region does not exist as all, as SFMBT2 promoter is overlapped by gene ITIH5 on same strand (figure 19).

The third region with candidate transcripts is *Gab1* imprinted region. In mouse, there is a downstream alternative promoter inside the first intron if starting from the canonical promoter, with TSS overlapping ERVK RLTR15 transposable element (figures 20 and 21). In rat, the alternative TSS appear to be conserved and also using ERVK RLTR15 element as TSS (figures 22 and 23). In cow, alternative promoters are there too, but not starting from transposable elements (figures 24 and 25).

The fourth region with candidate transcripts is *Slc38a4* region. There is a novel upstream transcript that was found to be imprinted in recent publications (Andergassen et al., 2017; Courtney W. Hanna et al., 2019) that uses ERVK MLTR31F_MM element as TSS. This is element specific for mouse. In our laboratory we identified additional novel upstream transcripts (with unknown imprinted status) between the imprinted novel gene and *Slc38a4*, one starting from LTR-MaLR ORR1B1 and one from ERVL MT2A element (figures 26 and 27). Both these element types are also present in rat. The transcriptome assemby in rat contains the transcript starting from MT2A, but no other upstream transcripts (figures 28 and 29), while in cow, no upstream transcripts are present (figure 30).

The last region with candidate transcripts is *Zfp64* region. In mouse, recent publication identified two novel novel imprinted genes upstream of Zfp64, one same strand annotated gene and one novel antisense gene. Neither of them starts from a transposable element. Transcriptome assembly in our laboratory identified a novel antisense overlapping (relative to *Zfp64*) transcript with unknown imprinting status starting from LTR-MaLR MTC element (figures 31 and 32) which is also present in rat. This transcript starting from LTR-MaLR MTC element is also found in rat, but the two imprinted genes are not there in the assembled transcriptome (figures 33 and 34). In cow, non of these transcripts is present in the transcriptome, but there is an upstream antisense transcript starting from LTR-MaLR MLT1J element (figures 35 and 36), which is not present in mouse or rat.

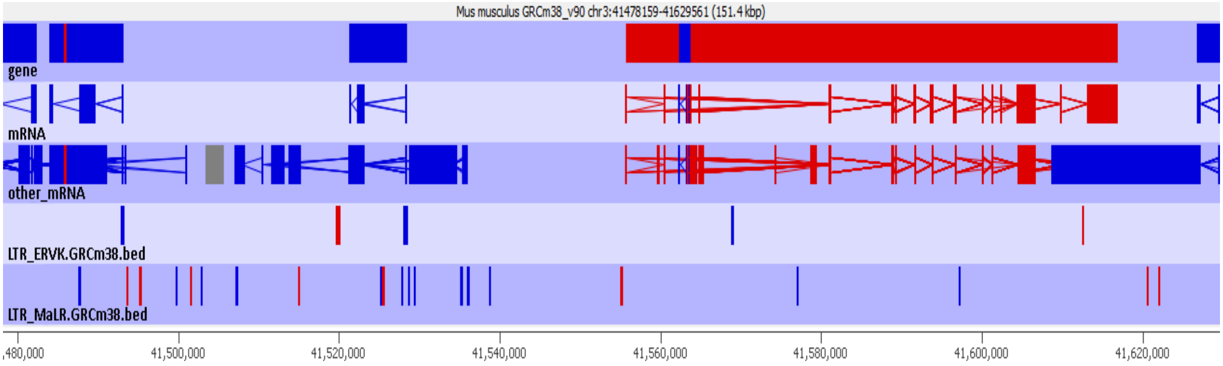


Figure 12. ***Jade1 region in mouse***. A screenshot from Seqmonk program. The first- and second-line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of ERVK transposable elements and fifth of LTR-MaLR transposable elements. The red gene on the right side is *Jade1*, blue transcripts to the left are novel transcripts overlapping the imprinted *2400006E01Rik* locus, using ERVK and LTR-MaLR elements as TSSs.

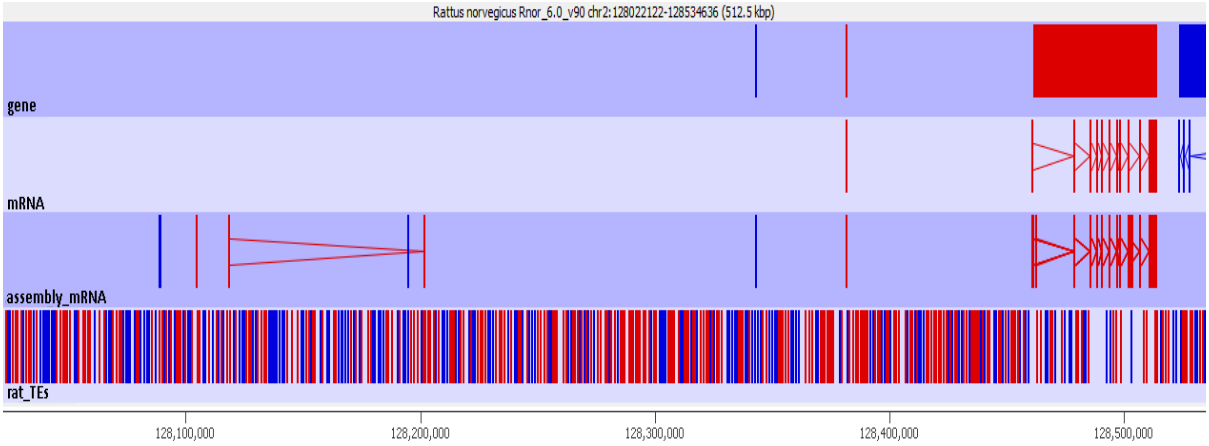


Figure 13. *Jade1 region in rat*. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of repetitive elements. The red gene on the rightside is *Jade1*, red transcript on the left is novel.

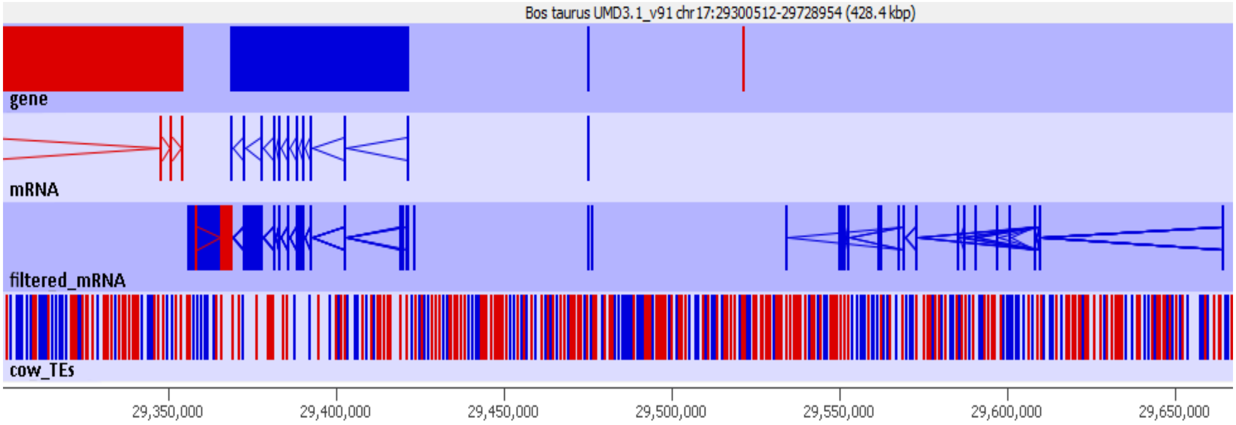


Figure 14. *Jade1 region in cow*. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of repetitive elements. The blue gene on the left side is JADE1, blue transcript on the right is novel.

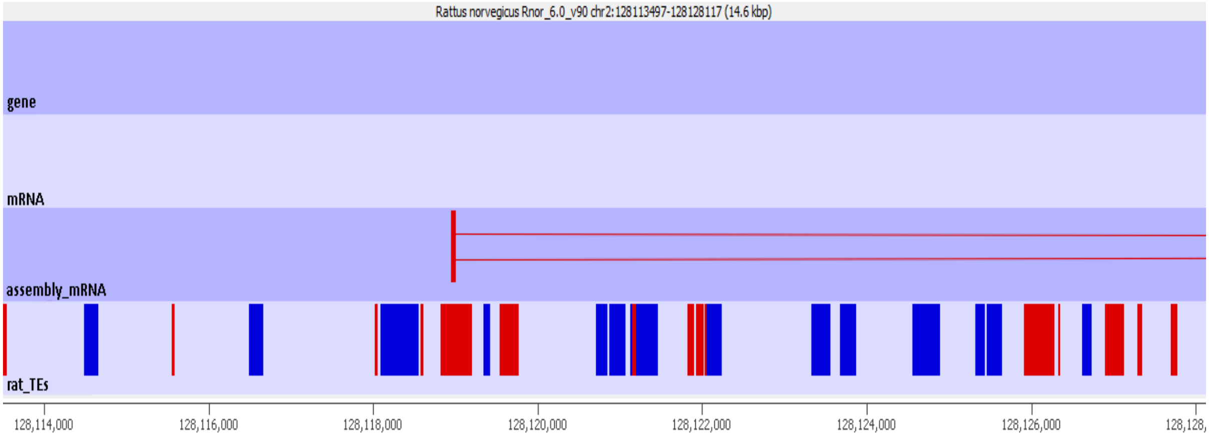


Figure 15. ***Jade1*** **region in rat, zoomed on the promoter of novel transcript**. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of repetitive elements. The visualized promoter of the novel transcript overlaps a transposable element.

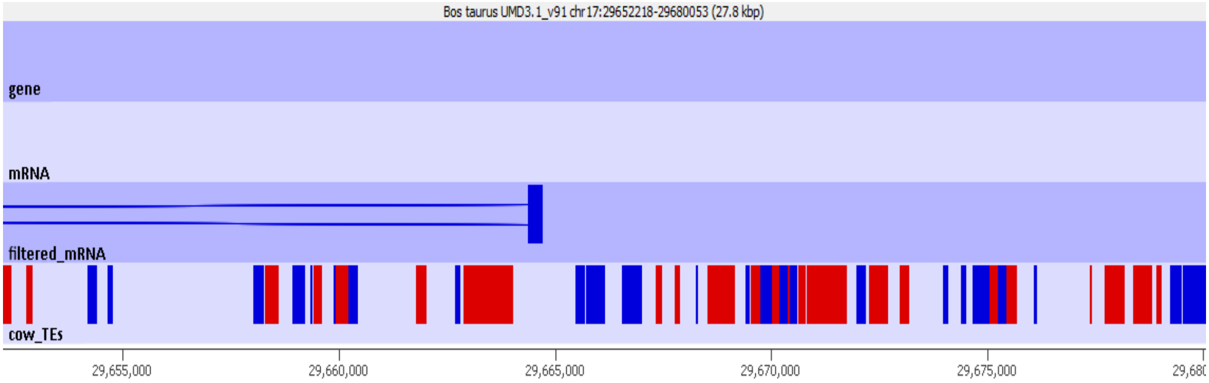


Figure 16. ***JADE1*** **region in cow, zoomed on the promoter of a novel transcript.** A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of repetitive elements. The visualized promoter of a novel transcript in blue does not overlap any repetitive element.

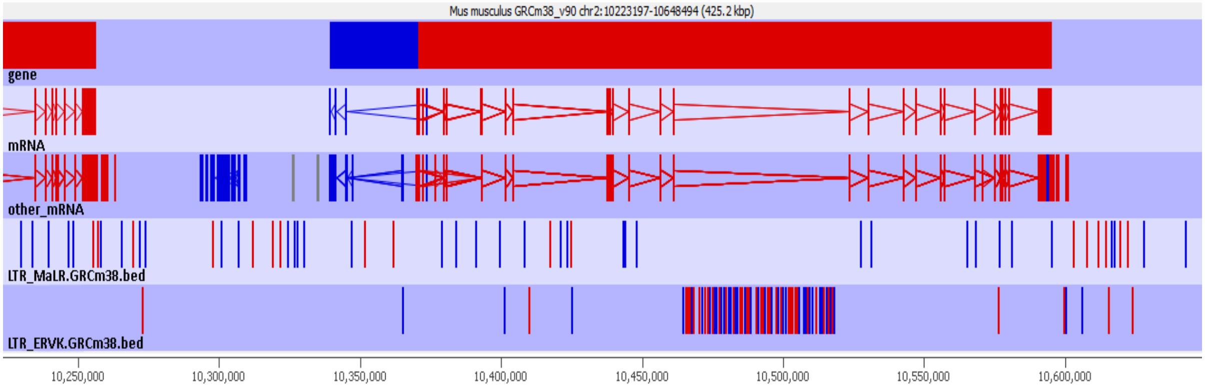


Figure 17. *Sfmbt2* region in mouse. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements and fifth of ERVK transposable elements. The red gene on the right side is *Sfmbt2*, blue transcript immediately to the left is novel imprinted gene and its novel downstream alternative TSS overlapping LTR-MaLR element, and further to the left there are novel blue transcripts with TSSs overlapping LTR-MaLR elements.

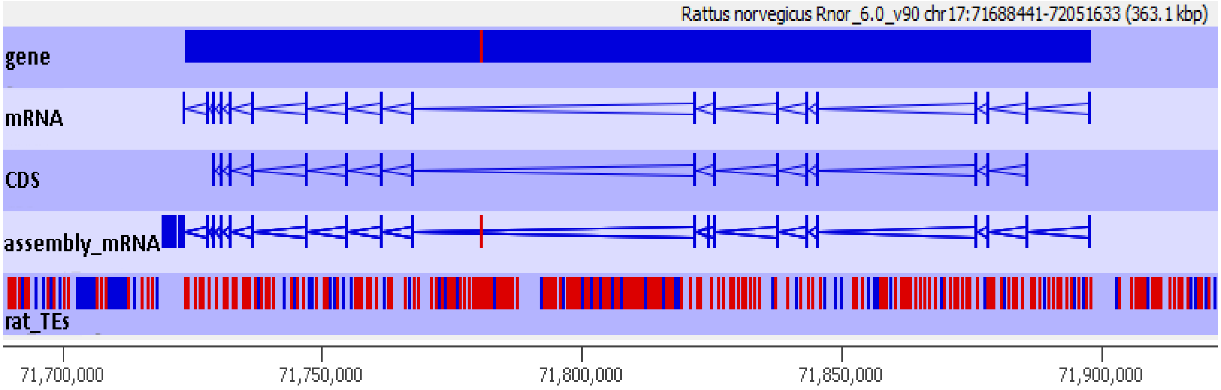


Figure 18. *Sfmbt2* region in rat. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized blue gene is *Sfmbt2*.

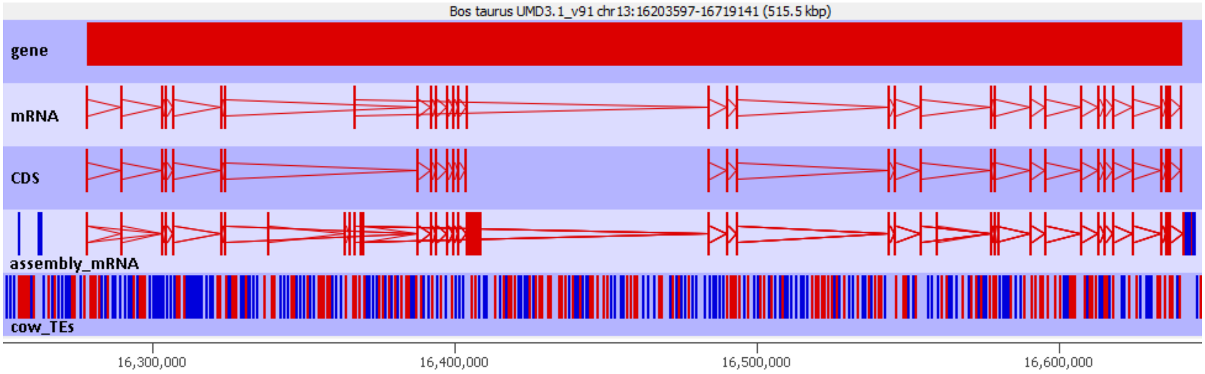


Figure 19. *SFMBT2* region in cow. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of repetitive elements. The visualized overlapping red genes are SFMBT2 on the right and ITIH5 on the left.

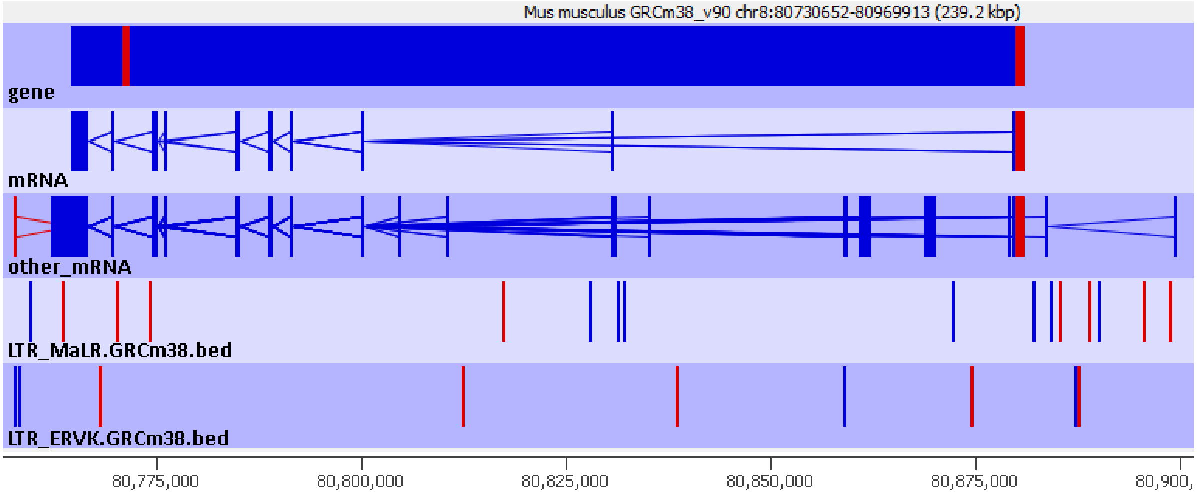


Figure 20. *Gab1* region in mouse. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements and fifth of ERVK transposable elements. The visualized blue gene is *Gab1* with its alternative downstream promoters inside the first intron of the promoter annotated in Ensembl.

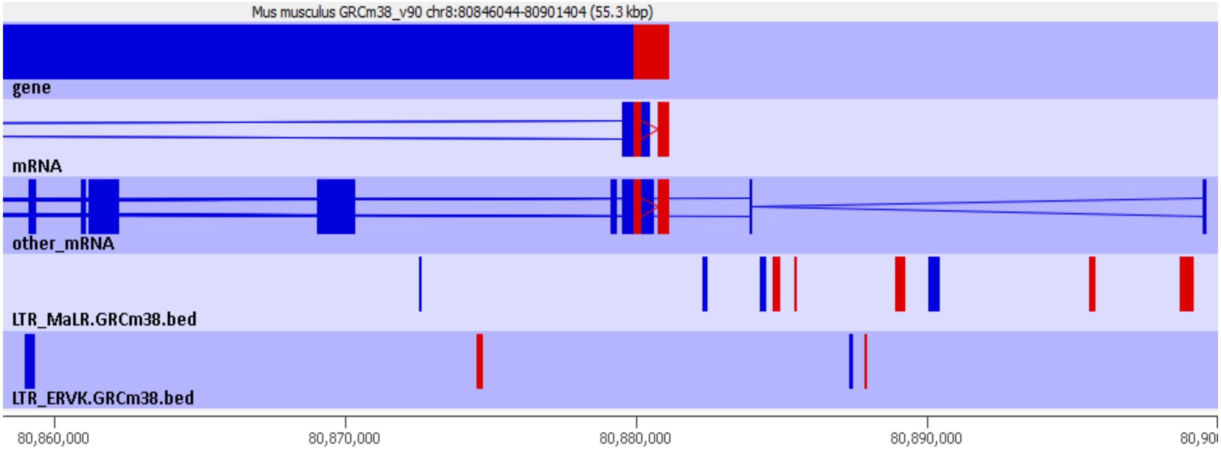


Figure 21. ***Gab1*** **region in mouse zoom**. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements and fifth of ERVK transposable elements. The visualized blue gene is *Gab1* with its alternative downstream promoters inside the first intron of the promoter annotated in Ensembl. On the left side, there is an alternative promoter overlapping ERVK element.

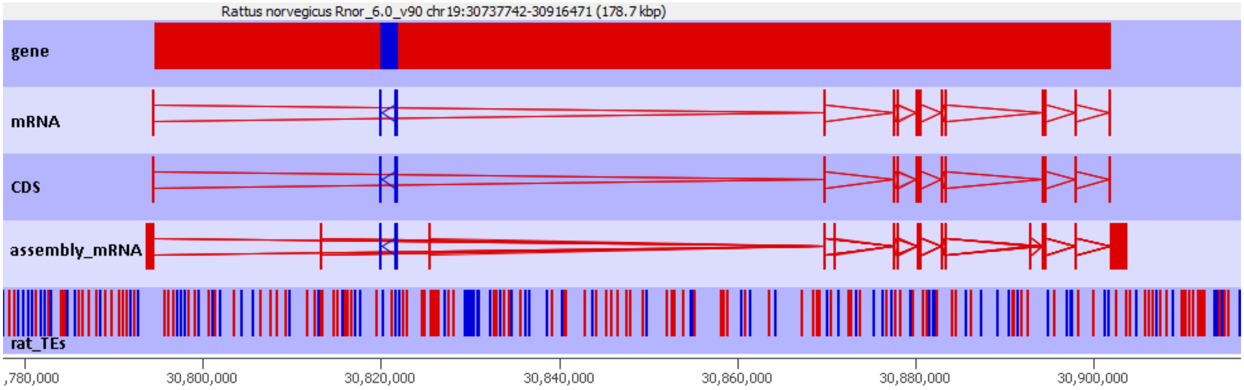


Figure 22. ***Gab1*** **region in rat**. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized red gene is *Gab1*, with its alternative downstream promoters inside the first intron.

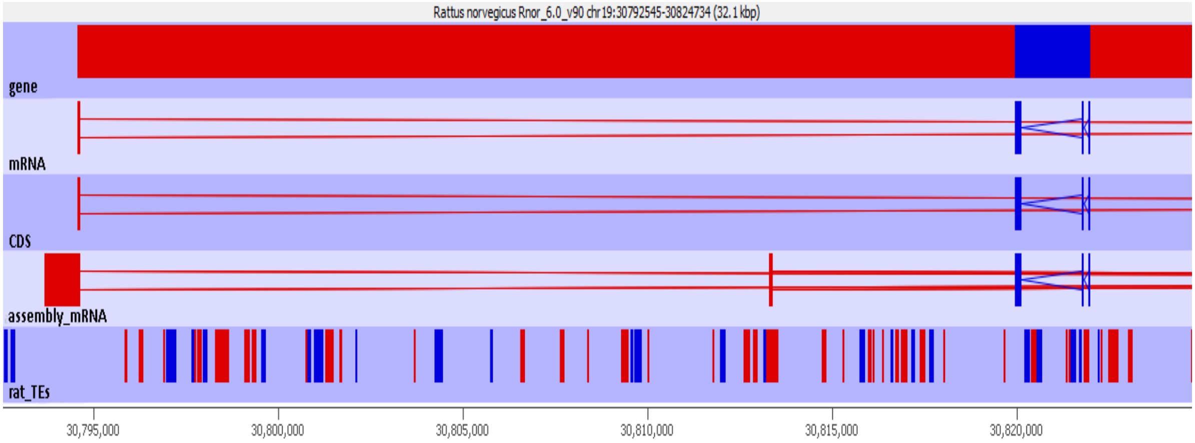


Figure 23. ***Gab1* region in rat, zoomed on the alternative downstream promoter.** A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized red gene is *Gab1*, with its alternative downstream promoter inside the first intron, overlapping a transposable element.

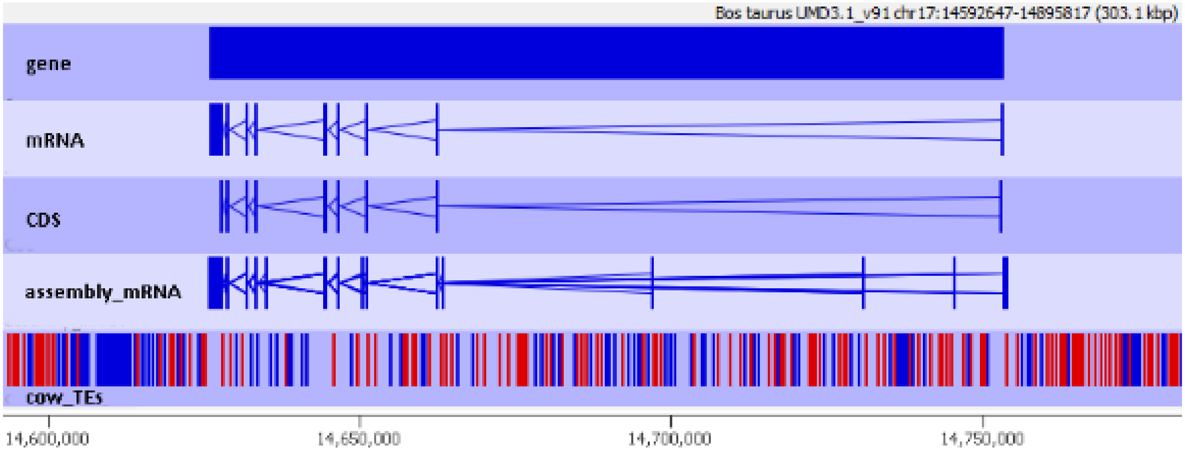


Figure 24. *GAB1* region in cow. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized blue gene is GAB1 with its alternative downstream promoters inside the first intron.

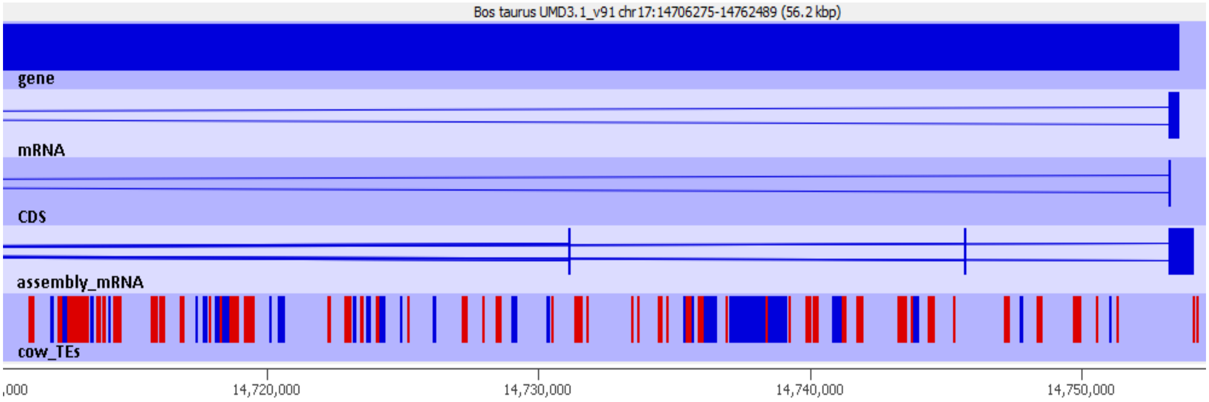


Figure 25. ***GAB1*** **region in cow, zoomed on alternative GAB1 promoters**. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized blue gene is GAB1 with its alternative downstream promoters inside the first intron, showing no overlap with repetitive elements.

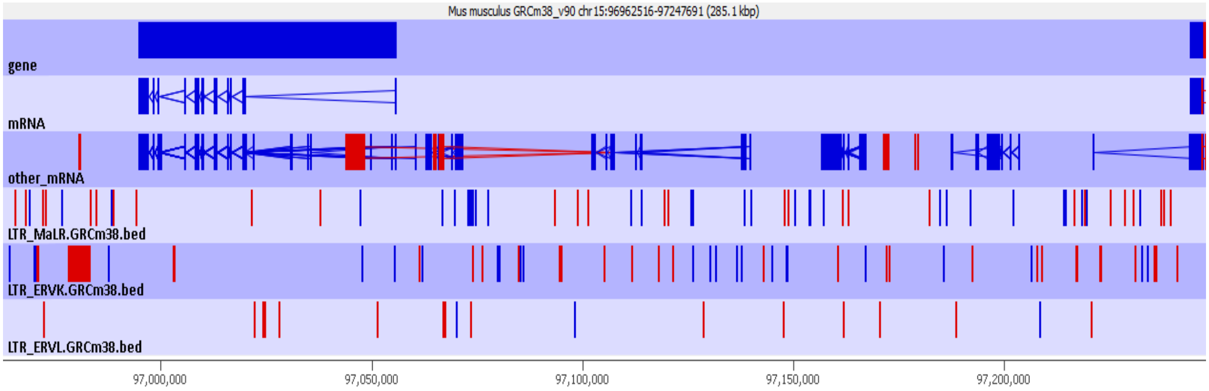


Figure 26. ***Slc38a4*** **region in mouse**. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements, fifth of ERVK transposable elements and sixth of ERVL transposable elements. The blue gene on the left is *Slc38a4* and the region to the right contains a number of same strand novel transcripts.

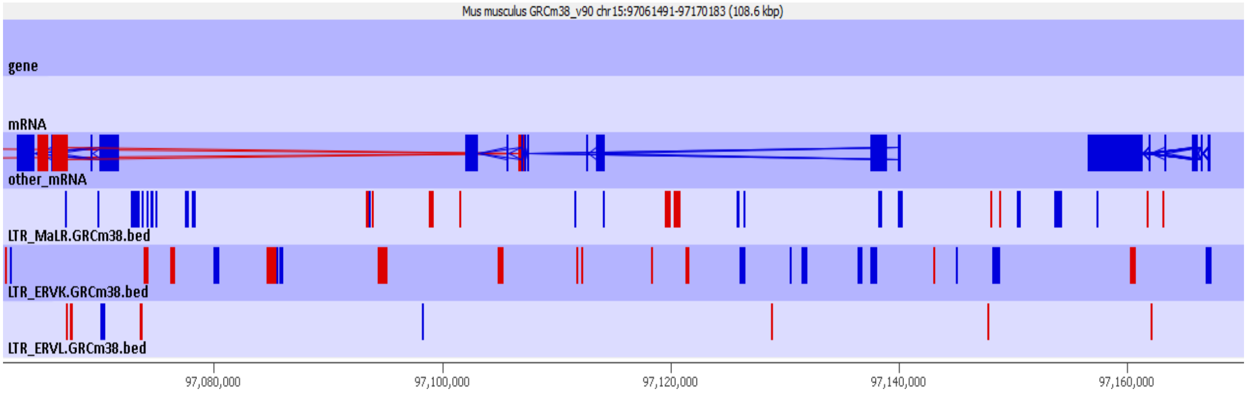


Figure 27. *Slc38a4* region in mouse, zoomed on promoters of novel transcripts. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements, fifth of ERVK transposable elements and sixth of ERVL transposable elements. The visualized novel upstream transcripts in blue have promoters overlapping ERVK and LTR-MaLR transposable elements.

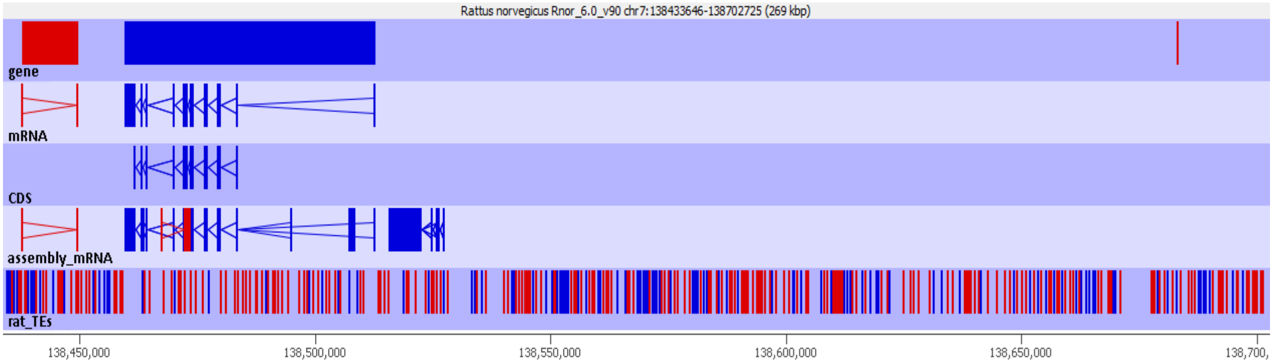


Figure 28. *Slc38a4* region in rat. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized blue gene on the left is *Slc38a4*, with its upstream region on the right, containing one novel transcript.

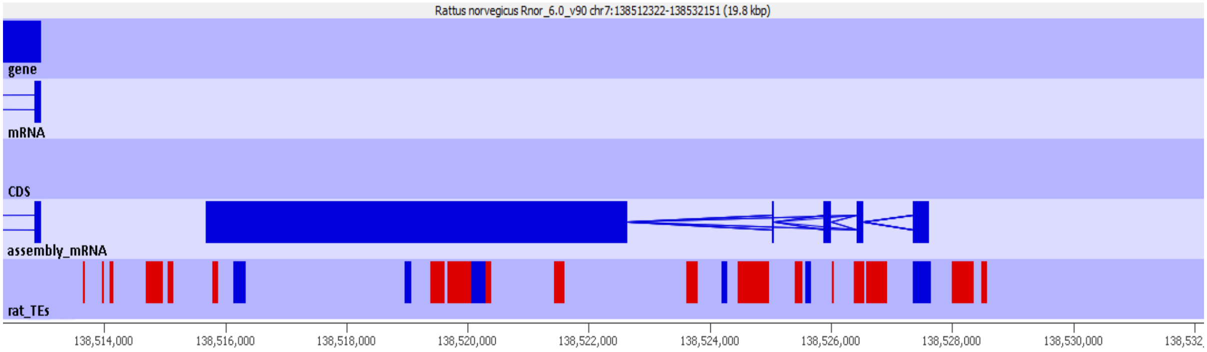


Figure 29. *Slc38a4* region in rat, zoomed on the upstream novel transcript. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized blue gene is the novel upstream transcripts with its promoter overlapping a transposable element.

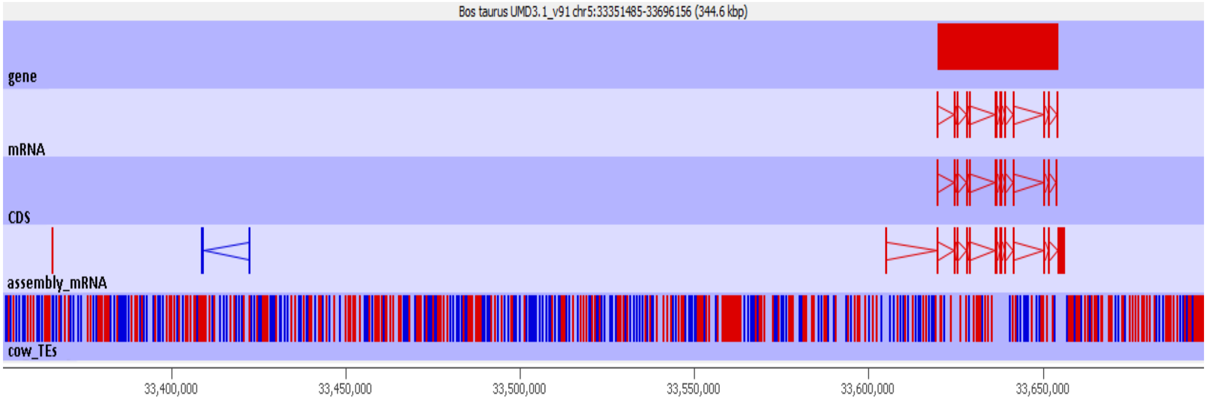


Figure 30. SLC38A4 region in cow. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized red gene is SLC38A4, showing there are no novel same strand transcripts upstream of it.

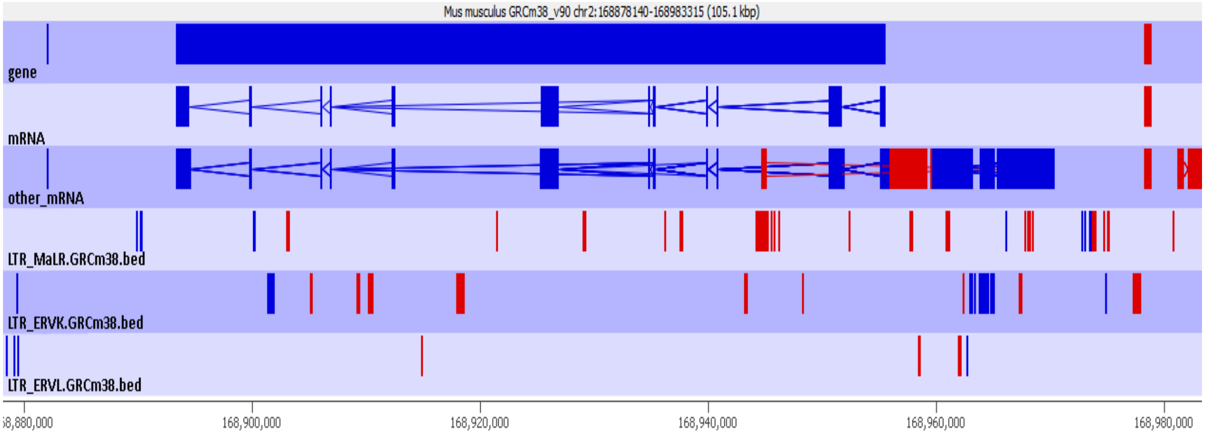


Figure 31. *Zfp64* region in mouse. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements, fifth of ERVK transposable elements and sixth of ERVL transposable elements. The visualized blue gene is *Zfp64* with the overlapping antisense transcript in red and novel upstream transcripts in blue.

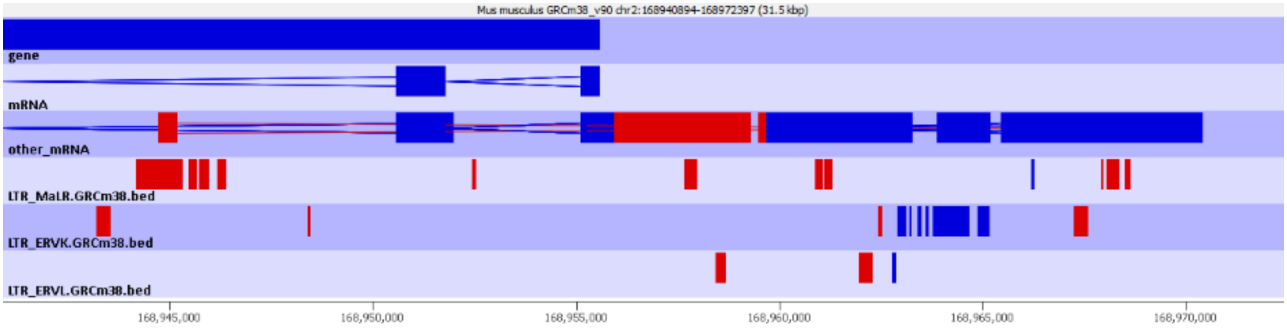


Figure 32. *Zfp64* region in mouse, zoomed on the novel transcripts. A screenshot from Seqmonk program. The first and second line show Ensembl transcriptome annotation for genes and their mRNAs (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The third line shows our de novo assembled transcriptome, fourth line is the annotation of LTR-MaLR transposable elements, fifth of ERVK transposable elements and sixth of ERVL transposable elements. The visualized are novel transcripts, with the antisense overlapping transcript in red starting in LTR-MaLR element.

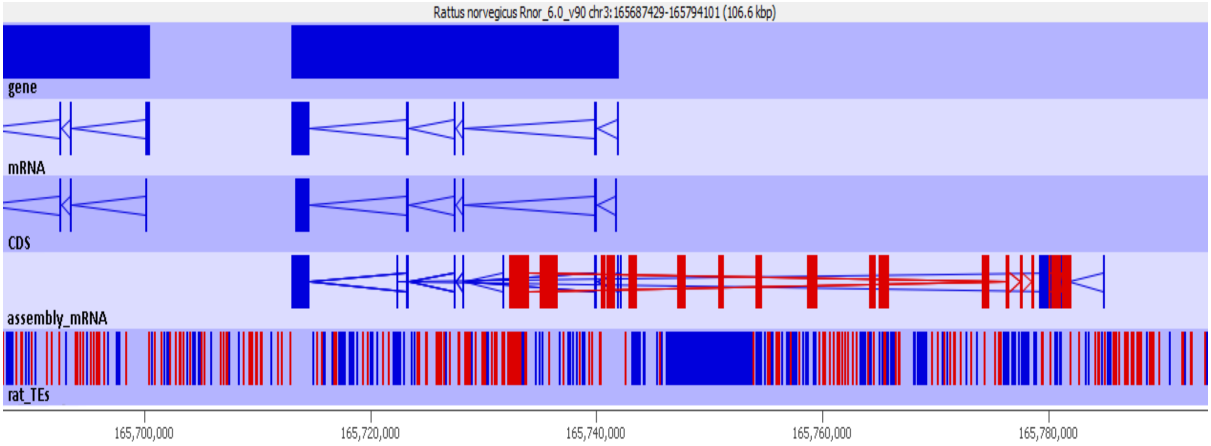


Figure 33. *Zfp64* region in rat. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The visualized blue gene in the middle is *Zfp64*, with the overlapping antisense transcripts in red.

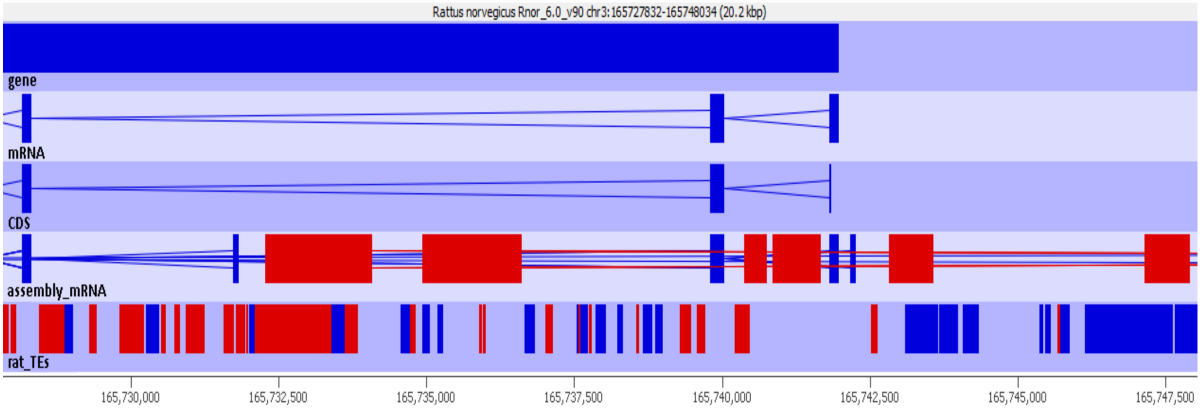


Figure 34. *Zfp64* region in rat, zoomed on the promoter of antisense overlapping transcript. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome; fifth line is the annotation of repetitive elements. The visualized blue gene is *Zfp64*, with the overlapping antisense transcript in red starting in transposable element.

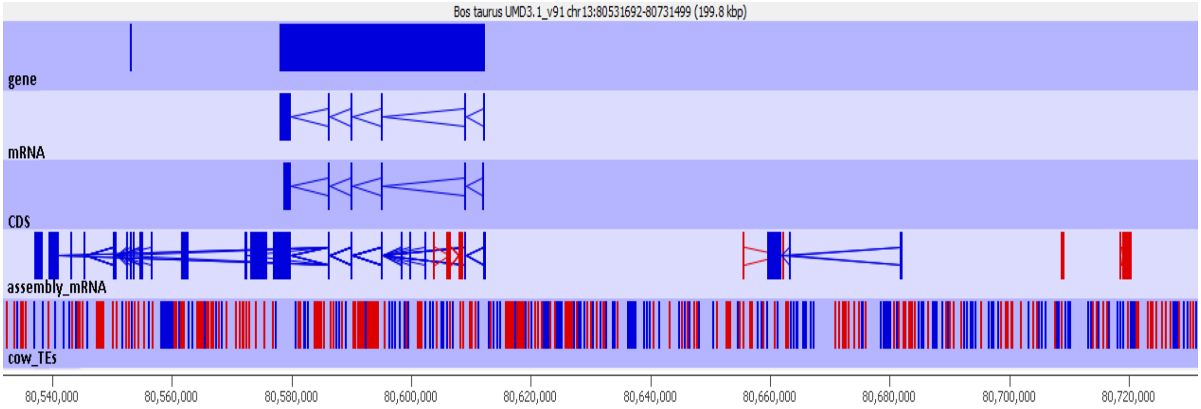


Figure 35. **ZFP64 region in cow**. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. The blue gene on the left is ZFP64, to the left if region upstream of it containing novel transcripts.

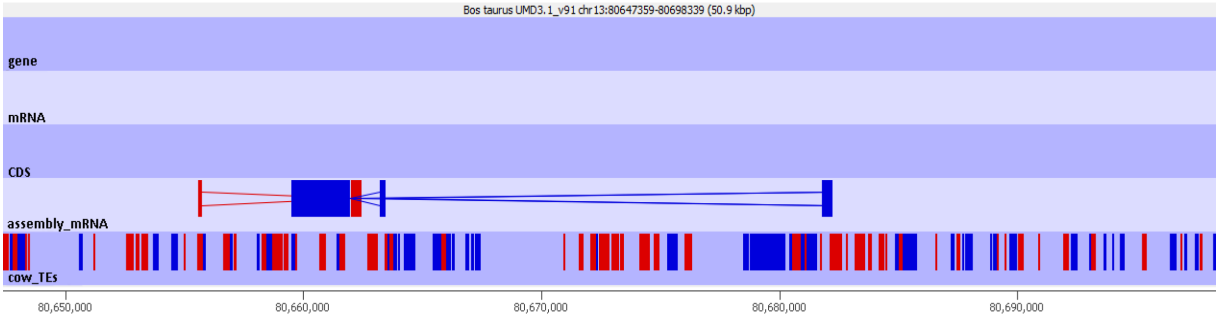


Figure 36. **ZFP64 region in cow, zoomed on the novel transcripts**. A screenshot from Seqmonk program. The first, second and third line show Ensembl transcriptome annotation for genes, their mRNAs and coding sequence (genes encoded on plus DNA strand going from left to right are visualized in red, genes encoded on minus DNA strand going from right to left in blue). The fourth line shows our de novo assembled transcriptome, fifth line is the annotation of repetitive elements. Visualised are the novel transcripts, with the left transcript starting from a transposable element.

8 DISCUSSION

Recent publications (Veselovska et al. 2015; Andergassen et al. 2017; Hanna et al. 2019) and previous analysis in our laboratory revealed that thorough transcriptome assembly in under-explored developmental stages and tissues, such as oocytes, embryos and placenta, identifies novel transcripts within the clusters of imprinted genes in mouse, some of which are imprinted too. In contrast to mouse which a model organism to study imprinting in mammals, with approximately 150 known imprinted genes, imprinting was poorly studied in other mammalian species, with the exception of human. Nevertheless, even the comparison of human and mouse imprinted genes shows that some genes which are imprinted in mouse are not imprinted in human (Morcos et al. 2011), suggesting that imprinted genes differ across mammalian species.

In order to explore genomic imprinting in other mammalian species, we processed publicly available RNA-seq datasets from various developmental stages of rat, cow, pig, marmoset, rhesus macaque and human. We selected these species because we were able to find appropriate datasets from them. Due to the difficulties with obtaining oocytes and embryos and with RNA-seq libraries preparation from such limited amount of input material, such datasets are available only from a small number of species, in contrast to for example somatic tissues. We collected datasets from various somatic tissues for all six species, oocyte datasets for all species except marmoset, preimplantation embryos for all species except rat, while postimplantation embryo datasets only for pig, cow and human, and placenta only for rhesus macaque and human.

By the inspection of literature and databases, we tried to collect information about all imprinted genes in mouse. Because imprinted genes are often organized in clusters, and we were interested in identifying all novel transcripts within the clusters of imprinted genes, we attempted to define the borders of imprinted regions in mouse and then find homologous regions in other species. We defined the borders of imprinted clusters as first protein-coding genes on each side with known or suggested function, for which it was either confirmed they are not imprinted, or their imprinting status is not known. We decided not to select non-coding or predicted protein-coding genes with unknown function, as they were more likely to be overlooked during imprinted expression analysis. The identification of homologous regions was done through the position, i.e. we looked for genes homologous to the mouse imprinted genes, and for the borderline genes surrounding them. For the majority of imprinting clusters, the whole regions were conserved, with the genes homologous to both imprinted and borderline genes. Nevertheless, some regions were remodeled during the evolution, and different borderline genes got closer the genes homologous to the mouse imprinted genes than the genes homologous to the mouse borderline genes. In such cases, we used the new genes as borders, but we cannot be sure how well is the whole region homologous to the mouse region. This could be improved by the thorough sequence analysis of the regions, and comparison between species.

For all the approximately 150 imprinted genes in mouse, the imprinting status is known for only a small proportion of them in other species except human. From the tested species, cow and pig have the highest numbers of genes with known imprinted status, while only few genes were tested in rat and macaque rhesus, and imprinting was never studied in marmoset. We also found out there is a mistake in the geneimprint database of all mammalian imprinted genes and their imprinting in each species. *Sfmbt2* gene is listed as maternally-expressed in the database, which was very interesting as this gene is paternally-expressed in mouse. Nevertheless, the search in the literature revealed that it is also paternally-expressed in rat (Wang et al. 2011).

We predicted the methylation status of hypothesized gDMRs in rat and cow. Because in mammals, maternally-methylated gDMRs has to be inside actively transcribed genes in order to become methylated in the oocytes, we manually inspected whether such overlapping transcripts expressed in the oocytes exist in rat and cow. The analysis was not always obvious as we first had to predict the potential position of gDMR in a non-mouse species (the position of gDMRs was defined in mouse), based on the overlap of the same feature, such as promoter, of a homologous gene. We then compared our predictions with the known imprinted statuses for a small subset of genes. In the majority of cases, our predictions were not in conflict with the data, with the exception of the PEG10/SGCE locus in cow. The gDMR in this locus should be localized overlapping the bidirectional promoter of PEG10 and SGCE. In mouse, there is an alternative upstream promoter of *Peg10* which provides the transcription through the gDMR. PEG10 and SGCE are imprinted according to the geneimprint database, however, we predicted the gDMR to be unmethylated as there was no transcription going through the bidirectional PEG10/SGCE promoter according the assembled transcriptome and the inspection of the oocyte RNA-seq data. Because other genes homologous to the mouse genes within the *Peg10/Sgce* cluster which are imprinted in mouse are not imprinted in cow, it is possible that the regulation of imprinted expression of transcripts in this regions is regulated differently in cow than in mouse, either due to the different position of gDMR, or through some other mechanisms. Also, proper analysis of the literature should reveal whether there is enough experimental evidence that PEG10 and SGCE are imprinted in cow.

Expression analysis of genes within potentially imprinted regions in rat and cow revealed that almost all transcripts within these regions are expressed predominantly at certain developmental stage or period, i.e. they are expressed only in the oocytes, or in the oocytes and early embryos, or at certain embryonic stage, or only in one or more somatic tissues. In rat, we could compare only oocytes and somatic tissues expression, but in the cow we also had datasets for various stages of embryonic development. Interestingly, the vast majority of transcripts appear to be specific either for oocytes and/or embryos, or for somatic tissues. Therefore, it appears that the transcription in these regions is dynamically regulated during development. This is not surprising, as we annotated many novel transcripts in our transcriptome assembly, which are likely to be lncRNAs, and lncRNAs are often expressed only in a specific developmental timepoint or tissue (Cabili et al. 2011; Derrien et al. 2012). It would be interesting to find out whether for example novel oocyte-specific transcripts within these potentially imprinted regions are conserved across mammalian species.

We observed that transcripts within potentially imprinted regions of cow and rat share similar sequence motifs which appear to serve as binding sites of similar transcriptional factors, suggesting that the regulation of transcription within imprinted clusters is to some extent conserved across mammalian species. Nevertheless, this is not surprising as many of the mouse imprinted genes have homologues in all mammals, and they are likely to be regulated by the same transcription factors. It would be interesting to compare just the sequence motifs of the novel, previously unannotated transcripts in individual species.

We also found out that a substantial proportion of transcripts within potentially imprinted regions in rat and cow starts from transposable elements. This was expected, as transcripts were found to employ transposable elements as promoters relatively frequently in the oocytes and embryos (Macfarlan et al. 2012; Veselovska et al. 2015). Nevertheless, when we manually compared individual novel transcripts using transposable as their promoter in mouse, rat and cow, they were often not conserved, even if the category of transposable elements that acted as a promoter was present also in the other tested species. In addition, some of such transcripts in mouse, which were in addition demonstrated to be imprinted, start from a mouse-specific transposable element not present in other species, and there is no corresponding transcript in other species. This suggests that transposable elements can shape even the imprinted transcriptome. This is interesting also for cases where the transposable element-starting transcript appear to regulate imprinted gene expression of a downstream protein-coding gene, such as in the case of mouse *Slc38a4* locus. The imprinted transcript that appears to regulate the imprinted expression of *Slc38a4* gene in mouse placenta (Hanna et al. 2019) is not present in rat or cow.

There are many future directions of this project. The datasets of six species were processed, but only rat and cow were analysed in more details. Therefore, the detailed analysis needs to be expanded to other species too, to get better insights into changes across species. Also, datasets from additional species can be processed and analysed, such as sheep, hamster, etc. In addition, the downstream analyses of expression, promoter sequences and transposable elements can be expanded to be more detailed. For example, the transcripts can be categorized as known and novel, or having TSSs with or without transposable elements, and their expression profiles during development can be compared. Or, the transcripts can be divided according to their expression profiles and then their sequence motifs can be compared. Also, it will be interesting to experimentally check the currently unknown imprinting status of gDMRs for which we predicted it, by looking at the allele-specific DNA methylation, or imprinted expression of associated genes.

9 CONCLUSIONS

Genomic imprinting is relatively poorly studied in all mammalian species except mouse and human. In order to shed more light on this phenomenon in other species, we collected and processed RNA-seq datasets from various developmental stages of six selected mammalian species (rat, cow, pig, marmoset, rhesus macaque and human). In all species except human, we identified regions homologous to the imprinted gene clusters in mouse and annotated all the transcripts within these regions across all available datasets, considering them to be potentially imprinted.

As maternally-methylated gDMRs require to be overlapped by active transcription unit in the oocytes in order to gain methylation, we predicted the methylation status of hypothesized gDMRs in rat and cow using our newly assembled transcriptomes which include the oocyte datasets. For a significant proportion of imprinted regions with unknown imprinting status in non-mouse species, we predicted that gDMRs are methylated and therefore are likely to be imprinted and to regulate imprinted gene expression of associated genes.

In addition, we observed that clusters of potentially imprinted genes are dynamically reprogrammed during development for individual species, as almost all transcripts appear to be specific for a certain developmental stage or period, and they differ between species, although to some extent their regulation by transcription factors appear to be conserved. The inter-species differences appear to be mostly due to the novel non-coding transcripts often employing transposable elements as promoters, confirmed by our finding that such transcripts previously identified in mouse are only rarely present in transcriptomes of other mammalian species.

The datasets, python scripts and results generated during this research project will serve as a basis for more thorough and detailed analysis of imprinted regions in mammals, especially of novel and developmental stage- or species-specific transcripts, and in deciphering what role transposable elements play in shaping the imprinted gene expression. In addition, these results contributed to the identification of candidate transcripts for further functional analysis.

10 REFERENCES

Andergassen, D., Dotter, C. P., Wenzel, D., Sigl, V., Bammer, P. C., Muckenhuber, M., Hudson, Q. J. (2017). Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. *ELife*, *6*. https://doi.org/10.7554/eLife.25125

Andrew, S. (2010). FastQC: A quality control tool for high throughput sequence data.

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, *27*(12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L.,Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

Barlow, D. P., & Bartolomei, M. S. (2014). Genomic Imprinting in Mammals. *Cold Spring Harbor Perspectives in Biology*, *6*(2), a018382–a018382. https://doi.org/10.1101/cshperspect.a018382

Bernardo, A. S., Jouneau, A., Marks, H., Kensche, P., Kobolak, J., Freude, K.,Dinnyes, A. (2018). Mammalian embryo comparison identifies novel pluripotency genes associated with the naïve or primed state. *Biology Open*, *7*(8), bio033282. https://doi.org/10.1242/bio.033282

Boroviak, T., Stirparo, G. G., Dietmann, S., Hernando-Herraez, I., Mohammed, H., Reik, W., Bertone, P. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development*, *145*(21), dev167833. https://doi.org/10.1242/dev.167833

Bourc'his, D. (2001). Dnmt3L and the Establishment of Maternal Genomic Imprints. *Science*, *294*(5551), 2536–2539. https://doi.org/10.1126/science.1065848

Brind'Amour, J., Kobayashi, H., Richard Albert, J., Shirane, K., Sakashita, A., Kamio, A., Lorincz, M. C. (2018). LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. *Nature Communications*, *9*(1), 3331. https://doi.org/10.1038/s41467-018-05841-x

Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M., & Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. *Nature Communications*, *9*(1), 4066. https://doi.org/10.1038/s41467-018-06544-z

Chao, W. (2011). Genomic Imprinting. In *Handbook of Epigenetics* (pp. 353–379). Elsevier. https://doi.org/10.1016/B978-0-12-375709-8.00022-8

Chitwood, J. L., Burruel, V. R., Halstead, M. M., Meyers, S. A., & Ross, P. J. (2017). Transcriptome profiling of individual rhesus macaque oocytes and preimplantation embryos†. *Biology of Reproduction*, *97*(3), 353–364. https://doi.org/10.1093/biolre/iox114

Chotalia, M., Smallwood, S. A., Ruf, N., Dawson, C., Lucifero, D., Frontera, M., … Kelsey, G. (2009). Transcription is required for establishment of germline methylation marks at imprinted genes. *Genes and Development*, *23*(1), 105–117. https://doi.org/10.1101/gad.495809

Chuong, E. B., Rumi, M. A. K., Soares, M. J., & Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics*, *45*(3), 325–329. https://doi.org/10.1038/ng.2553

Davis, T. L. (2000). The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. *Human Molecular Genetics*, *9*(19), 2885–2894. https://doi.org/10.1093/hmg/9.19.2885

Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R. Z., Ragozin, S., & Jeltsch, A. (2010). The Dnmt3a PWWP Domain Reads Histone 3 Lysine 36 Trimethylation and Guides DNA Methylation. *Journal of Biological Chemistry*, *285*(34), 26114–26120.

https://doi.org/10.1074/jbc.M109.089433

Dunn-Fletcher, C. E., Muglia, L. M., Pavlicev, M., Wolf, G., Sun, M.-A., Hu, Y.-C., Muglia, L. J. (2018). Anthropoid primate–specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. *PLOS Biology*, *16*(9), e2006337. https://doi.org/10.1371/journal.pbio.2006337

Fadloun, A., Le Gras, S., Jost, B., Ziegler-Birling, C., Takahashi, H., Gorab, E., Torres-Padilla, M.-E. (2013). Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nature Structural & Molecular Biology*, *20*(3), 332–338. https://doi.org/10.1038/nsmb.2495

Ferguson-Smith, A. C. (2011). Genomic imprinting: The emergence of an epigenetic paradigm. *Nature Reviews Genetics*, *12*(8), 565–575. https://doi.org/10.1038/nrg3032

Franke, V., Ganesh, S., Karlic, R., Malik, R., Pasulka, J., Horvat, F., Svoboda, P. (2017). Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Research*, *27*(8), 1384–1394. https://doi.org/10.1101/gr.216150.116

Frohlich, L. F., Mrakovcic, M., Steinborn, R., Chung, U.-I., Bastepe, M., & Juppner, H. (2010). Targeted deletion of the Nesp55 DMR defines another Gnas imprinting control region and provides a mouse model of autosomal dominant PHP-Ib. *Proceedings of the National Academy of Sciences*, *107*(20), 9275–9280. https://doi.org/10.1073/pnas.0910224107

Frost, J. M., & Moore, G. E. (2010). The Importance of Imprinting in the Human Placenta. *PLoS Genetics*, *6*(7), e1001015. https://doi.org/10.1371/journal.pgen.1001015

Fushan, A. A., Turanov, A. A., Lee, S.-G., Kim, E. B., Lobanov, A. V., Yim, S. H., Gladyshev, V. N. (2015). Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, *14*(3), 352–365. https://doi.org/10.1111/acel.12283

Guibert, S., Forne, T., & Weber, M. (2012). Global profiling of DNA methylation erasure in mouse primordial germ cells. *Genome Research*, *22*(4), 633–641. https://doi.org/10.1101/gr.130997.111

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Qiao, J. (2014). The DNA methylation landscape of human early embryos. *Nature*, *511*(7511), 606–610. https://doi.org/10.1038/nature13544

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. (2007). Quantifying similarity between motifs. *Genome Biology*, *8*(2), R24. https://doi.org/10.1186/gb-2007-8-2-r24

Hamada, H., Okae, H., Toh, H., Chiba, H., Hiura, H., Shirane, K., Arima, T. (2016). Allele-Specific Methylome and Transcriptome Analysis Reveals Widespread Imprinting in the Human Placenta. *The American Journal of Human Genetics*, *99*(5), 1045–1058. https://doi.org/10.1016/j.ajhg.2016.08.021

Hanna, C. W., Demond, H., & Kelsey, G. (2018). Epigenetic regulation in development: Is the mouse a good model for the human? *Human Reproduction Update*, *24*(5), 556–576. https://doi.org/10.1093/humupd/dmy021

Hanna, C. W., & Kelsey, G. (2014). The specification of imprints in mammals. *Heredity*, *113*(2), 176–183. https://doi.org/10.1038/hdy.2014.54

Hanna, C. W., Peñaherrera, M. S., Saadeh, H., Andrews, S., McFadden, D. E., Kelsey, G., & Robinson, W. P. (2016). Pervasive polymorphic imprinted methylation in the human placenta. *Genome Research*, *26*(6), 756–767. https://doi.org/10.1101/gr.196139.115

Hanna, C. W., Schubert, M., Gahurova, L., Krueger, F., Andrews, S., Colomé-Tatché, M., Gavin, K. (2019). Novel roles for H3K27me3 in regulating imprinted gene expression in extra-embryonic development. *Unpublished (under Review)*.

Hanna, C. W., Taudt, A., Huang, J., Gahurova, L., Kranz, A., Andrews, S., Kelsey, G. (2018). MLL2 conveys transcription-independent H3K4 trimethylation in oocytes. *Nature Structural & Molecular Biology*, *25*(1), 73–82. https://doi.org/10.1038/s41594-017-0013-5

Hata, K., Okano, M., Lei, H., & Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development (Cambridge, England)*, *129*(8), 1983–1993. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11934864

Inoue, A., Jiang, L., Lu, F., Suzuki, T., & Zhang, Y. (2017). Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature*, *547*(7664), 419–424. https://doi.org/10.1038/nature23262

Ishida, M., & Moore, G. E. (2013). The role of imprinted genes in humans. *Molecular Aspects of Medicine*, *34*(4), 826–840. https://doi.org/10.1016/j.mam.2012.06.009

Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., & Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, *449*(7159), 248–251. https://doi.org/10.1038/nature06146

Kagiwada, S., Kurimoto, K., Hirota, T., Yamaji, M., & Saitou, M. (2012). Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *The EMBO Journal*, *32*(3), 340–353. https://doi.org/10.1038/emboj.2012.331

Kaneda, M., Hirasawa, R., Chiba, H., Okano, M., Li, E., & Sasaki, H. (2010). Genetic evidence for Dnmt3a-dependent imprinting during oocyte growth obtained by conditional knockout with Zp3 -Cre and complete exclusion of Dnmt3b by chimera formation. *Genes to Cells*, *15*(3), 169–179. https://doi.org/10.1111/j.1365-2443.2009.01374.x

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., & Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature*, *429*(6994), 900–903. https://doi.org/10.1038/nature02633

Karlic, R., Ganesh, S., Franke, V., Svobodova, E., Urbanova, J., Suzuki, Y., Svoboda, P. (2017). Long non-coding RNA exchange during the oocyte-to-embryo transition in mice. *DNA Research*, dsw058. https://doi.org/10.1093/dnares/dsw058

Kato, Y., Kaneda, M., Hata, K., Kumaki, K., Hisano, M., Kohara, Y., Sasaki, H. (2007). Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Human Molecular Genetics*, *16*(19), 2272–2280. https://doi.org/10.1093/hmg/ddm179

Kelsey, G., & Feil, R. (2013). New insights into establishment and maintenance of DNA methylation imprints in mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1609). https://doi.org/10.1098/rstb.2011.0336

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kobayashi, H., Sakurai, T., Imai, M., Takahashi, N., Fukuda, A., Yayoi, O., Kono, T. (2012). Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks. *PLoS Genetics*, *8*(1), e1002440. https://doi.org/10.1371/journal.pgen.1002440

Latos, P. A., Pauler, F. M., Koerner, M. V., Senergin, H. B., Hudson, Q. J., Stocsits, R. R., Barlow, D. P. (2012). Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. *Science*, *338*(6113), 1469–1472. https://doi.org/10.1126/science.1228110

Lavagi, I., Krebs, S., Simmet, K., Beck, A., Zakhartchenko, V., Wolf, E., & Blum, H. (2018). Single-cell RNA sequencing reveals developmental heterogeneity of blastomeres during major genome activation in bovine embryos. *Scientific Reports*, *8*(1), 4071. https://doi.org/10.1038/s41598-018-22248-2

Lewis, A., & Reik, W. (2006). How imprinting centres work. *Cytogenetic and Genome Research*, *113*(1–4), 81–89. https://doi.org/10.1159/000090818

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Li, C., Li, S., Peng, Q., An, N. A., He, A., & Li, C.-Y. (2018). Human exonization through differential nucleosome occupancy. *Proceedings of the National Academy of Sciences*, *115*(35), 8817–8822. https://doi.org/10.1073/pnas.1802561115

Liu, D., Wang, X., He, D., Sun, C., He, X., Yan, L., … Zheng, P. (2018). Single-cell RNA-sequencing reveals the existence of naive and primed pluripotency in pre-implantation rhesus monkey embryos. *Genome Research*, *28*(10), 1481–1493. https://doi.org/10.1101/gr.233437.117

Lorthongpanich, C., Cheow, L. F., Balu, S., Quake, S. R., Knowles, B. B., Burkholder, W. F., Messerschmidt, D. M. (2013). Single-Cell DNA-Methylation Analysis Reveals Epigenetic Chimerism in Preimplantation Embryos. *Science*, *341*(6150), 1110–1112. https://doi.org/10.1126/science.1240617

Lucifero, D., Mann, M. R. W., Bartolomei, M. S., & Trasler, J. M. (2004). Gene-specific timing and epigenetic memory in oocyte imprinting. *Human Molecular Genetics*, *13*(8), 839–849. https://doi.org/10.1093/hmg/ddh104

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, *487*(7405), 57–63. https://doi.org/10.1038/nature11244

Mackay, D. J. G., & Temple, I. K. (2017). Human imprinting disorders: Principles, practice, problems and progress. *European Journal of Medical Genetics*, *60*(11), 618–626. https://doi.org/10.1016/j.ejmg.2017.08.014

Mager, J., Montgomery, N. D., de Villena, F. P.-M., & Magnuson, T. (2003). Genome imprinting regulated by the mouse Polycomb group protein Eed. *Nature Genetics*, *33*(4), 502–507. https://doi.org/10.1038/ng1125

Mancini-DiNardo, D. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes & Development*, *20*(10), 1268–1282. https://doi.org/10.1101/gad.1416906

Messerschmidt, D. M., de Vries, W., Ito, M., Solter, D., Ferguson-Smith, A., & Knowles, B. B. (2012). Trim28 Is Required for Epigenetic Stability During Mouse Oocyte to Embryo Transition. *Science*, *335*(6075), 1499–1502. https://doi.org/10.1126/science.1216154

Morcos, L., Ge, B., Koka, V., Lam, K. C., Pokholok, D. K., Gunderson, K. L., Pastinen, T. (2011). Genome-wide assessment of imprinted expression in human cells. *Genome Biology*, *12*(3), R25. https://doi.org/10.1186/gb-2011-12-3-r25

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., & Fraser, P. (2008). The Air Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin. *Science*, *322*(5908), 1717–1720. https://doi.org/10.1126/science.1163802

Nakamura, T., Arai, Y., Umehara, H., Masuhara, M., Kimura, T., Taniguchi, H., Nakano, T. (2007). PGC7/Stella protects against DNA demethylation in early embryogenesis. *Nature Cell Biology*, *9*(1), 64–71. https://doi.org/10.1038/ncb1519

Nakamura, T., Liu, Y.-J., Nakashima, H., Umehara, H., Inoue, K., Matoba, S., Nakano, T. (2012). PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. *Nature*, *486*(7403), 415–419. https://doi.org/10.1038/nature11093

Okae, H., Chiba, H., Hiura, H., Hamada, H., Sato, A., Utsunomiya, T., Arima, T. (2014). Genome-

Wide Analysis of DNA Methylation Dynamics during Early Human Development. *PLoS Genetics*, *10*(12), e1004868. https://doi.org/10.1371/journal.pgen.1004868

Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., … Bestor, T. H. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, *448*(7154), 714–717. https://doi.org/10.1038/nature05987

Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D., & Knowles, B. B. (2004). Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos. *Developmental Cell*, *7*(4), 597–606. https://doi.org/10.1016/j.devcel.2004.09.004

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

Popp, C., Dean, W., Feng, S., Cokus, S. J., Andrews, S., Pellegrini, M., Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, *463*(7284), 1101–1105. https://doi.org/10.1038/nature08829

Proudhon, C., Duffié, R., Ajjan, S., Cowley, M., Iranzo, J., Carbajosa, G., Bourc'his, D. (2012). Protection against De Novo Methylation Is Instrumental in Maintaining Parent-of-Origin Methylation Inherited from the Gametes. *Molecular Cell*, *47*(6), 909–920. https://doi.org/10.1016/j.molcel.2012.07.010

Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Trono, D. (2011). In Embryonic Stem Cells, ZFP57/KAP1 Recognize a Methylated Hexanucleotide to Affect Chromatin and DNA Methylation of Imprinting Control Regions. *Molecular Cell*, *44*(3), 361–372. https://doi.org/10.1016/j.molcel.2011.08.032

Randy L. Jirtle, P. D. (n.d.). Imprinted Genes: by Species. Retrieved April 8, 2019, from http://www.geneimprint.com/site/genes-by-species

Reik, W., Dean, W., & Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, *293*(5532), 1089–1093. https://doi.org/10.1126/science.1063443

Reik, W., & Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, *2*(1), 21–32. https://doi.org/10.1038/35047554

Reyes, J. M., Chitwood, J. L., & Ross, P. J. (2015). RNA-Seq profiling of single bovine oocyte transcript abundance and its modulation by cytoplasmic polyadenylation. *Molecular Reproduction and Development*, *82*(2), 103–114. https://doi.org/10.1002/mrd.22445

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, *27*(17), 2325–2329. https://doi.org/10.1093/bioinformatics/btr355

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*(3), R22. https://doi.org/10.1186/gb-2011-12-3-r22

Ruebel, M. L., Schall, P. Z., Midic, U., Vincent, K. A., Goheen, B., VandeVoort, C. A., & Latham, K. E. (2018). Transcriptome analysis of rhesus monkey failed-to-mature oocytes: deficiencies in transcriptional regulation and cytoplasmic maturation of the oocyte mRNA population. *MHR: Basic Science of Reproductive Medicine*, *24*(10), 478–494. https://doi.org/10.1093/molehr/gay032

Santos, F., Peat, J., Burgess, H., Rada, C., Reik, W., & Dean, W. (2013). Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. *Epigenetics & Chromatin*, *6*(1), 39. https://doi.org/10.1186/1756-8935-6-39

Scott, C. D., & Weiss, J. (2000). Soluble insulin-like growth factor II/mannose 6-phosphate receptor inhibits DNA synthesis in insulin-like growth factor II sensitive cells. *Journal of Cellular*

Physiology, *182*(1), 62–68. https://doi.org/10.1002/(SICI)1097-4652(200001)182:1<62::AID-JCP7>3.0.CO;2-X

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Reik, W. (2012). The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Molecular Cell*, *48*(6), 849–862. https://doi.org/10.1016/j.molcel.2012.11.001

Shirane, K., Toh, H., Kobayashi, H., Miura, F., Chiba, H., Ito, T., Sasaki, H. (2013). Mouse Oocyte Methylomes at Base Resolution Reveal Genome-Wide Accumulation of Non-CpG Methylation and Role of DNA Methyltransferases. *PLoS Genetics*, *9*(4), e1003439. https://doi.org/10.1371/journal.pgen.1003439

Sleutels, F., Zwart, R., & Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, *415*(6873), 810–813. https://doi.org/10.1038/415810a

Smallwood, S. A., Tomizawa, S., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics*, *43*(8), 811–814. https://doi.org/10.1038/ng.864

Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, *14*(3), 204–220. https://doi.org/10.1038/nrg3354

Terranova, R., Yokobayashi, S., Stadler, M. B., Otte, A. P., van Lohuizen, M., Orkin, S. H., & Peters, A. H. F. M. (2008). Polycomb Group Proteins Ezh2 and Rnf2 Direct Genomic Contraction and Imprinted Repression in Early Mouse Embryos. *Developmental Cell*, *15*(5), 668–679. https://doi.org/10.1016/j.devcel.2008.08.015

Tomizawa, S. -i., Kobayashi, H., Watanabe, T., Andrews, S., Hata, K., Kelsey, G., & Sasaki, H. (2011). Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development*, *138*(5), 811–820. https://doi.org/10.1242/dev.061416

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. https://doi.org/10.1038/nbt.2450

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Tsai, T.-S., Tyagi, S., & St. John, J. C. (2018). The molecular characterisation of mitochondrial DNA deficient oocytes using a pig model. *Human Reproduction*, *33*(5), 942–953. https://doi.org/10.1093/humrep/dey052

Veselovska, L., Smallwood, S. A., Saadeh, H., Stewart, K. R., Krueger, F., Maupetit Méhouas, S., Kelsey, G. (2015). Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome Biology*, *16*(1), 1–17. https://doi.org/10.1186/s13059-015-0769-z

Wang, Q., Chow, J., Hong, J., Smith, A. F., Moreno, C., Seaby, P.,Varmuza, S. (2011). Recent acquisition of imprinting at the rodent Sfmbt2 locus correlates with insertion of a large block of miRNAs. *BMC Genomics*, *12*(1), 204. https://doi.org/10.1186/1471-2164-12-204

Wang, X., Liu, D., He, D., Suo, S., Xia, X., He, X., Zheng, P. (2017). Transcriptome analyses of rhesus monkey preimplantation embryos reveal a reduced capacity for DNA double-strand break repair in primate oocytes and early embryos. *Genome Research*, *27*(4), 567–579. https://doi.org/10.1101/gr.198044.115

Wang, X., Soloway, P. D., & Clark, A. G. (2011). A Survey for Novel Imprinted Genes in the Mouse Placenta by mRNA-seq. *Genetics*, *189*(1), 109–122. https://doi.org/10.1534/genetics.111.130088

Watanabe, T., Tomizawa, S. -i., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Sasaki, H. (2011). Role for piRNAs and Noncoding RNA in de Novo DNA Methylation of the Imprinted Mouse Rasgrf1 Locus. *Science*, *332*(6031), 848–852. https://doi.org/10.1126/science.1203919

Wright, S. J. (1999). 5 Sperm Nuclear Activation during Fertilization. In *Current Topics in Developmental Biology* (pp. 133–178). https://doi.org/10.1016/S0070-2153(08)60328-2

Wu, J., Xu, J., Liu, B., Yao, G., Wang, P., Lin, Z., Sun, Y. (2018). Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature*, *557*(7704), 256–260. https://doi.org/10.1038/s41586-018-0080-8

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Tang, F. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology*, *20*(9), 1131–1139. https://doi.org/10.1038/nsmb.2660

Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Jeltsch, A. (2010). Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Research*, *38*(13), 4246–4253. https://doi.org/10.1093/nar/gkq147

Zuo, X., Sheng, J., Lau, H.-T., McDonald, C. M., Andrade, M., Cullen, D. E., Li, X. (2012). Zinc Finger Protein ZFP57 Requires Its Co-factor to Recruit DNA Methyltransferases and Maintains DNA Methylation Imprint in Embryonic Stem Cells via Its Transcriptional Repression Domain. *Journal of Biological Chemistry*, *287*(3), 2107–2118. https://doi.org/10.1074/jbc.M111.322644

11 LIST OF APPENDICES

Appendix 1. Python Script for final merged gtf file.

Files name: Sylvia_1(final code).py

Language: Python

Description: The script filters the gtf file for different specifies, according to specific regions based in Chromosome location, start and end of the region, it will then output a filtered file with just the desired regions.

Input file: gtf files: (cow, rat, pig, mouse and marmoset) _merged.gtf

Output file: Text file with filtered results.

(cow, rat, pig, mouse and marmoset) _merged_exons_filtered.gtf

```python
# Code works for all regions at the same time and repetitive chromosomes
# Python code filters regions based on chromosome and specific start and end of bases.
import re

chromosomes = ["2","13”]
bases = [[94926123,95145118],[16250005,17093613]]
input_filename = "cow_merged.gtf"

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

        # prints progress
        print(input_filename + ": " + str(counter) + "/" + str(count_lines) + " Found: " +
str(findings))


geneids_in_region()


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

Appendix 2. Python Script to remove transcripts from a list

Files name: Sylvia_finalcode2.py

Language: Python

Description: Python script makes an extra filtering for removing transcripts with one exon from a list of subset genes.

Input file: cow_merged.gtf and rat_merged.gtf. And the subset of genes files rat_to_be_removed_if_1_exon.txt and rat_to_be_removed_if_1_exon.txt

Output file: Text file with filtered results, cow_merged_exons_filtering_subset.gtf and rat_merged_exons_filtering_subset.gtf

```
# Python code filters regions based on chromosome and specific start and end of bases.
#and, makes an extra filtering for removing transcripts with one exon, from a subset of genes.

import re
import os

chromosomes = ["9","17"]
bases = [[69956161, 70133968], [71105287,72160734]]
input_filename = "rat_merged.gtf"

# creates output file name: input_filename + filtered.gtf
```

```
output_filename = input_filename[:input_filename.rfind(".")] + "_exons_filtering_subset.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")
#open temp file
of = open("temp.gtf", "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

        # prints progress
        #print(input_filename + ": " + str(counter) + "/" + str(count_lines) + " Found: " +
str(findings))
        if counter%10000 == 0:
            print("First cleavage " + str(counter) +" Found: " + str(findings))
```

```
geneids_in_region()

# closes the output file
of.close()

#Additional filtering
#reading temporary file

with open("temp.gtf") as f:
    # reads all lines
    lines_tmp2 = f.readlines()

#remove temporary file
os.remove("temp.gtf")

#reading exons that need to be remove

import pandas as pd

dataset=pd.read_csv("rat_to_be_removed_if_1_exon.txt",delimiter="\t")

#create list with values
remove_marker = dataset["Probe"].tolist()

#make format as we have in our transcript_id list
for k in range(len(remove_marker)):
    remove_marker[k] = '"{0}"'.format(remove_marker[k])


print("\n\nThird cleavage start...\n")
transcript_id_2 = []

for i in range(len(lines_tmp2)):
    data_from_line = []
    #split line for extracting ids
    for j in lines_tmp2[i].split(";")[1:-1]:
        data_from_line.append(j.split(" ")[2])

    transcript_id_2.append(data_from_line[0])

# find delete from lists needed lines
count = 0
for l in range(len(remove_marker)):
#    if counter%1 == 0:
#        print("Found for removing: " + str(count))
    if transcript_id_2.count(remove_marker[l]) == 1:
         index = transcript_id_2.index(remove_marker[l])
         print(index)
         del transcript_id_2[index]
         del lines_tmp2[index]
         count += 1


#write into file
of = open(output_filename, "w")
for i in range(len(lines_tmp2)):
    of.write(lines_tmp2[i])
of.close()
print("\nOutput file: " + output_filename)
```

Appendix 3 Python Script to remove transcripts makes and extra filtering for cow specie to remove dots.

Files name: Sylvia_finalcode3.py

Language: Python

Description: if there is a dot(.) in file program cannot find in which strand it is, so we removed. Specially used for cow specie.

Input file: cow_merged.gtf and rat_merged.gtf. And the subset of genes files rat_to_be_removed_if_1_exon.txt and rat_to_be_removed_if_1_exon.txt

Output file: Text file with filtered results cow_merged_exons_filtered(+-).gtf

```
# # python code filter regions and remove dots from file.
# Specially used for cow specie

import re
import os

chromosomes = ["2", "13"]
bases = [[94926123, 95145118]]
input_filename = "cow_merged.gtf"

# creates output file name: input_filename + filtered.gtf
output_filename = input_filename[:input_filename.rfind(".")] + "_exons_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")
# open temp file
of = open("temp.gtf", "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

        # prints progress
        # print(input_filename + ": " + str(counter) + "/" + str(count_lines) + " Found: " +
str(findings))
        if counter % 10000 == 0:
            print("First cleavage " + str(counter) + " Found: " + str(findings))


geneids_in_region()

# closes the output file
of.close()

# Additional filtering
# reading temporary file

with open("temp.gtf") as f:
    # reads all lines
    lines_tmp2 = f.readlines()

# remove temporary file
os.remove("temp.gtf")

#######
# read needed exons to remove
#######

import pandas as pd

dataset = pd.read_csv("cow_to_be_removed_if_1_exon.txt", delimiter="\t")

# create list with values
remove_marker = dataset["Probe"].tolist()

# make format as we have in our transcript_id list
for k in range(len(remove_marker)):
    remove_marker[k] = '"{0}"'.format(remove_marker[k])

print("\n\nSecond cleavage start...\n")
transcript_id_2 = []

for i in range(len(lines_tmp2)):
    data_from_line = []
    # split line for extracting ids
    for j in lines_tmp2[i].split(";")[1:-1]:
        data_from_line.append(j.split(" ")[2])
    transcript_id_2.append(data_from_line[0])

# find delete from lists needed lines
count = 0
for l in range(len(remove_marker)):
    if transcript_id_2.count(remove_marker[l]) == 1:
        index = transcript_id_2.index(remove_marker[l])
        print("Row deleted by exon filtering: " + str(index))
        del transcript_id_2[index]
        del lines_tmp2[index]
        count += 1

#################################################
print("\n\nThird cleavage start...\n")
lines_tmp3 = lines_tmp2

has_plus_check = []

for i in range(len(lines_tmp3)):
    # split line for extracting ids
    j = lines_tmp3[i].split("\t")[1:-1]
    has_plus_check.append(''.join(j[4:7]))

print("length: " + str(len(has_plus_check)) + "  " + str(len(lines_tmp3)))

# find delete from lists needed lines
count = 0
plus_ch_check = []
minus_ch_check = []
for k in range(len(has_plus_check)):
    # print(k)
    plus_ch_check.append(has_plus_check[k].count("+"))
    minus_ch_check.append(has_plus_check[k].count("-"))

delete_row_index = []
for h in range(len(plus_ch_check)):
    if int(plus_ch_check[h]) == 0:
        if int(minus_ch_check[h]) == 0:
            delete_row_index.append(h)

for i in sorted(delete_row_index, reverse=True):
    del lines_tmp3[i]

# write into file
of = open(output_filename, "w")
for i in range(len(lines_tmp2)):
    of.write(lines_tmp3[i])
of.close()
print("\nOutput file: " + output_filename)
```

Appendix 4. R Script for Heatmap

Files name: heatmap.R

Language: R

Description: The script is for Heatmap visualization and clustering.

Input file: rat_expression.txt and rat_clusters.txt

Output file: heatmap plot

```
#R Script

install.packages("gplots")

library(gplots)

data <- read.delim ("rat_expression.txt")

rnames <- data[,1]

mat_data <- data.matrix(data[,2:ncol(data)])

rownames(mat_data) <- rnames

hr <- hclust(as.dist(1-cor(t(mat_data), method="pearson")), method="complete")

colorRampPalette(c("yellow","blue")) -> colour.gradient


heatmap.2(mat_data, col=colour.gradient, breaks=seq(from=-1,to=1, by=0.001),
Rowv=as.dendrogram(hr), Colv=FALSE,

        scale="none", dendrogram="none", key=T, keysize=2, density.info="none", hclust=function(x)
hclust(x,method="complete"),

        distfun=function(x) as.dist((1-cor(t(x)))/2),

        trace="none",cexCol=1.2, labRow=NA)


data$clusternumber <- cutree (hr, 12)  #for rat

data$clusternumber <- cutree (hr, 15)  #for cow

write.table(data, "rat_clusters.txt")
```

12 SUPPLEMENTARY MATERIAL

Supplementary Table 1: Table contains dataset, GEO accession codes and reference, library type for very all species in this investigation.

Supplementary Table 2: Contains Hisat 2 output table with the accession codes of the individual datasets in fq.gz form, total number of reads we got after trimming process, the alignment rate and the number of mapped reads.

Supplementary Table 3: Contains List of imprinting genes.

Supplementary Table 4: Contains rat_Imprinting quantification.

Supplementary Table 5: Contains cow_ Imprint_quantification

Supplementary Table 6: Python code previously make in laboratory. To extract the sequences of promoter regions in FASTA format from genomic sequences.