University of South Bohemia in České Budějovice
Faculty of Science

&

Johannes Kepler University Linz
Faculty of Engineering and Natural Sciences

Dnmt3L in mammalian oocytes

Bioinformatics Project

Student:
Lirik Behluli

Supervisor:
Mgr. Lenka Gahurová, Ph.D.

České Budějovice, 2022

1. Abstract

The gene Dnmt3L is a member of the Dnmt (DNA methyltransferases) genes family which also include Dnmt1, Dnmt3a, and Dnmt3b. All of them are needed for transferring a methyl group from universal methyl donor S-adenosyl-L-methionine (SAM) to the 5-position of cytosine residues in DNA, a process which is called DNA methylation and which is key for mammalian development. DNMT3L is a co-factor of DNMT3A and DNMT3B with similar structures to them, but interestingly misses some catalytic protein domain resulting in it being enzymatically inactive (Emburgh et al., 2020).

In literature, Dnmt3l is presented as one of the key factors of de novo DNA methylation in mammalian oocytes. In this project, we analyzed the expression of Dnmt3l and other DNA methyltransferases in oocytes of multiple mammalian species and identified a specific sequence that arose in mouse, rat, and hamster within the intron of the neighboring Aire gene serving as an oocyte-specific Dnmt3l promoter. The analysis of Aire intron sequences in all available rodent species revealed that it is present only in a rodent lineage leading to mice, rats, and hamsters, and absent in all other lineages. Moreover, once the promoter sequence appeared, it became more conserved than other intronic sequences of the Aire gene. Our results, therefore, showed that Dnmt3l is a key factor for oocyte de novo DNA methylation only in a specific rodent lineage and is not a universal mammalian factor.

2. Introduction

2.1. DNA Methylation

DNA Methylation is an epigenetic mechanism used by cells to control gene expression. DNA methylation is quite vital to several cellular processes which comprise: embryonic development, X-chromosome inactivation, genomic imprinting and gene suppression, carcinogenesis, and the stability of chromosomes. It also enables the suppression of repetitive elements including endogenous retroviral sequences. Overall, DNA methylation and its remodelling is crucial for correct development and differentiation enabling a single cell to grow into a complex multi-cellular organism (Moore et al., 2013).
In the mammalian genome, DNA methylation involves the transfer of a methyl group onto the C5 position of the cytosine to form 5-methylcytosine. DNA methylation is one of the epigenetic mechanisms regulating gene expression by recruiting silencing methyl-binding proteins that sterically interfere with transcription factor binding as well as affecting the binding of some transcription factors itself. Then during development, DNA methylation pattern in the genome changes as a result of a process involving both de novo DNA methylation and demethylation, and as a result, differentiated cells develop a stable and unique DNA methylation pattern that regulates tissue-specific gene transcription. (Moore et al., 2013).

DNA methylation in itself is considered a repressive epigenetic modification. In the somatic cells of most mammalian cells, it occurs throughout the majority of the genome, excluding active gene regulatory regions such as promoters, enhancers, or CpG islands which are in general unmethylated. DNA methylation is generally thought to have the role of regulating gene expression. The oocyte is different because both the genomic methylation pattern and its function are distinct from somatic cells, as in oocytes methylation is mostly constrained to actively transcribed regions (gene bodies of active genes). This gives the oocyte genome highly methylated gene bodies which are separated by intergenic or transcriptionally inactive regions with low methylation levels. This could be the reason that only one of the Dnmt's, the Dnmt3a is active in oocytes (Demond et al., 2020).

Throughout mammalian embryo development, DNA methylation is erased globally in the primordial germ cells, which come from cells of the epiblast. Therefore, the primary oocytes are almost lacking methylation. The methylation is then re-setted in the later phases of oocyte growth, peaking in the oocyte-specific pattern. Oocyte represents a very interesting model to study the mechanisms of DNA methylation due to the fact that an entire methylation landscape is organized in a non-dividing cell. (Demond et al., 2020)

Classically, Dnmt3a and Dnmt3l are considered the key factors of de novo DNA methylation establishment in the oocytes, as upon knock-out of either gene, oocytes remain largely unmethylated. This does not affect the development of the oocytes themselves, but fertilised unmethylated oocytes are not capable of embryonic development, to a large extent because of the aberrant expression of imprinted genes (Yokomine et al., 2006). However, the research studying the oocyte role of Dnmt3l in the oocytes was performed on mouse, and later studies showed that Dnmt3l is not expressed in human oocytes (Huntriss et al., 2004) and therefore cannot play role in the oocyte DNA methylation. In this project, we would like to shed more light on the requirement for Dnmt3l for DNA de novo methylation establishment in the oocytes across mammals. We would like to answer the question whether the example of mouse (with Dnmt3l) or of human (without Dnmt3l) is more common in mammals and what are the molecular reasons of these differences.

3. AIMS

- Analyzing DNMT3L expression in oocytes of various mammalian species

- Identify sequence-level changes which could lead to either activating or silencing the expression of DNMT3L in the oocyte

- Predict oocyte expression of DNMT3L across mammalian species.

4. Materials and methods

4.1. Analysis of gene expression

To analyse the expression of genes *Dnmt3l*, *Dnmt3a*, *Dnmt3b* and *Dnmt1* in mammalian oocytes, we used publicly available oocyte RNA-seq datasets, as well as new datasets generated in our laboratory. All datasets were trimmed, quality controlled and mapped to the respective genomes previously in the laboratory. List of the datasets and genomes to which the data were mapped can be found in table 1.

Table1. The list of oocyte RNA-seq datasets used in this project

Species	Accession code	Reference	Genome
mouse	GSE70116	Veselovska et al., 2015	GRCm38
rat	GSE112622	Brindamour et al. 2018	Rnor_6.0
golden hamster	GSE86470	Franke et al., 2017	MesAur1.0
guinea pig	n/a	Gahurova, unpublished data	Cavpor3.0
naked mole rat	n/a	Gahurova, unpublished data	hetGla2
cow	GSE61717	Reyes et al., 2015	UMD3.1
pig	GSE108900	Tsai et al., 2018	Sscrofa11.1
macaque rhesus	GSE103313, GSE86938	Chitwood et al., 2017, Wang et al., 2017	Mmul_10
human	GSE36552, GSE101571	Yan et al., 2013, Wu at al., 2018	GRCh38

SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/) software was utilized visualize and quantify RNA-seq datasets. SeqMonk is a bioinformatical program (software) for visualizing and analyzing a large set of mapped genomic regions, suitable for analysis of next-generation sequencing (NGS) data. It is designed with the purpose to work with data from high throughput sequencing technologies, which can work just as well with any set of mapped reads. SeqMonk has a genome browser that allows the user to navigate in a quick manner around annotated genomes, and moreover, it contains a set of tools that allows the user to quantify and filter data so that it is possible to find regions of interest in a systematic way (Andrews et al., 2007). During our analyses, SeqMonk was used to quantify oocyte RNA-sequence datasets, using an RNA-sequence quantitation pipeline specifying strand specificity of reading and whether the given data was the single-end or paired-end. The expression values

are log2 transformed RPKM (reads per million of the reads in the dataset per 1 kilobase of the gene length) values for single end data) or FPKM (fragments) per million of the reads in the dataset per 1 kilobase of the gene length)values for paired end data.

4.2. *Dnmt3l* sequence in the naked mole rat genome

Resources listed below were used for the purpose of confirming that *Dnmt3l* is annotated at the correct position in the naked mole rat genome, and in case it is not; to identify where it is in the genome based on the sequence similarity with human Dnmt3l sequence. Naked mole rat *Dnmt3a*, *Dnmt3b* and *Dnmt3l* sequences were used for screening program (Appendix 1.) to identify whether in *Dnmt3l* there are identical sequences to *Dnmt3a* and *Dnmt3b* that can cause *Dnmt3l* reads mapping incorrectly to *Dnmt3a* and *Dnmt3b*.

Human Dnmt3l was retrieved from NCBI:

https://www.ncbi.nlm.nih.gov/gene/29947

Naked mole rat Dnmt3a transcript variants were retrieved from the NCBI:
https://www.ncbi.nlm.nih.gov/nuccore/XM_004839112.3,XM_004839113.3,XM_004839114.3,XM_004839115.3,XM_004839117.3,XM_021253567.1,XM_021253568.1,XM_021253569.1,XM_021253570.1,XM_021253571.1

Naked mole rat Dnmt3b variants were retrieved from the NCBI
https://www.ncbi.nlm.nih.gov/nuccore/XM_004867953.3,XM_004867957.3,XM_004867958.3,XM_021246874.1,XM_021246875.1,XM_021246876.1,XM_021246877.1,XM_021246878.1,XM_021246879.1,XM_021246880.1,XM_021246881.1,XM_021246882.1

Naked mole rat Dnmt3l variants: were retrieved from the NCBI:https://www.ncbi.nlm.nih.gov/nuccore/XM_004842591.2,XM_021256378.1

Naked mole rat genome was downloaded from
https://hgdownload.soe.ucsc.edu/goldenPath/hetGla2/bigZips/

4.3. Multiple sequence alignment

Clustal programs were implemented heavily during our analyses. In general, Clustal is part of the family of sequence alignment programs used in the area of Bioinformatics & Computational Biology, as well as other bio-areas. It was developed in 1998 and is used to align related RNA, Protein, or DNA sequences (Thompson et al., 2003). Generally, Clustal starts with a set of sequences and performs a series of pairwise alignments to build a relationship tree that shows how similar your given sequences are. Clustal then is separated into different “branches” of its functionalities and each of them offers improvements for specific tasks (i.e. ClustalV, ClustalW & Clustal Omega (both of which were used during our analyses), Clustal2). The use of the

Clustal program is in the fact that it does multiple sequence alignment, while other programs (generally) will do mostly just alignments of 2 sequences (pairwise sequence alignment). In our analysis, we compared *Aire* intron sequences using ClustalW (https://www.genome.jp/tools-bin/clustalw) with default parameters. We used ClustalW, as the idea was to perform multiple sequence alignment taking into consideration matches, substitutions and insertions/deletions with each of these having a score. ClustalW gives scores for each compared pair.

4.4. Usage of Python and its libraries

Programming language Python was used to utilize or develop scripts for specified analyses. Python is an interpreted high-level general-purpose language and is widely used among natural scientists due to its design philosophy emphasizing code readability (Python, 2019). The language constructs as well as its object-oriented approach aim to help to write clear, logical code, either for small or for large projects. It was developed by Guido van Rossum and was first released in 1991 as Python 0.9.0. It is one of the most used programming languages in the world as of now. Two python scripts were utilized during our analyses. The first script (Appendix 1) takes two sequence files as input, compares between them and checks if given number of base pairs (in this case 5, 10, 11, 12, 13, 15) is the same in Dnmt3L, Dnmt3a and Dnmt3b; meanwhile the second script takes our intronic sequences as input and graphically plots the promoter regions and if they are more conserved than normal (Appendix 2). The parameters were as explained in description at Appendix 2.

Biopython, Numpy & Matplotlib

Python usually has built-in libraries, or you will have to download them manually on the web. Libraries are a collection of related modules which contain bundles of code that can be used repeatedly in different programs and makes it simpler for the programmer to not write the same code many times.

Biopython is a library containing sets and available tools for biological computation (Cock et al., 2009).

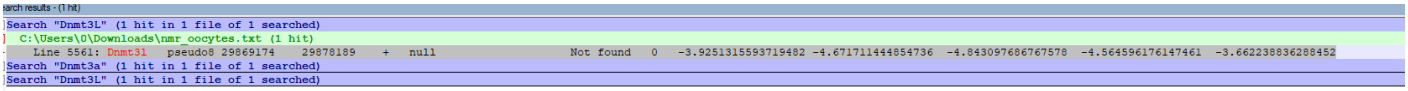
NumPy is a library for linear algebra. At its core, there is NumPy array, a multi-dimensional data structure that can be used to represent vectors and matrices (Van der Walt et al., 2011).

Matplotlib is a library used for data visualization, where you can create bar-plots, scatter-plots, histograms, and so on. It is crucial as nowadays we live in the world of data and Bioinformatics in itself is the study of biological data. Biology gives tons of data every day making data visualization an essential component of a scientist's skill set (Hunter et al., 2007).

4.5. Usage of Notepad++ software

Notepad++ was used for searching keywords in txt/fasta files containing a list of thousands of results, to select our desired sequences as needed, as they were marked with special keywords or Ensembl codes. Notepad++ is a text and source code editor which works for the operating system Windows. It supports tabbed editing and allows working with multiple open files in a single window (Shum et al., 2013).

Notepad++ software allows us to search for particular keywords thus making it easier to find our targets. At particular organisms, they were mostly annotated with ENSEMBL codes or numbers, and they were searched using these keywords.



Example of result in Notepad++ search query.

4.6. NCBI Genome Workbench

https://www.ncbi.nlm.nih.gov/tools/gbench/ *The genome workbench from NCBI simply offers a set of integrated tools to study and analyze genetic data where the user can explore and compare data from multiple sources including NCBI databases or the user's itself own private data (Kuznetsov et al., 2021). It is supported by alignment tools such as BLAST, CLustal, MAFFT. During our analyses, we used the tool to blast Dnmt3l sequences of naked mole rat against its whole genome, as well as blasting Human variant of Dnmt3l against the naked mole rat one.*

5. Results

5.1. Quantification of expression of Dnmt1, Dnmt3a, Dnmt3b and Dnmt3l in oocytes from different mammalian species

First, we aimed to analyse the oocyte expression of Dnmt1, Dnmt3a, Dnmt3b and Dnmt3l in multiple mammalian species (mouse, rat, golden hamster, guinea pig, naked mole rat, cow, pig, macaque rhesus and human). We used publicly available RNA-seq datasets and datasets generated in our laboratory (see Table 1 in Methods section). All these datasets were previously processed and mapped to respective genomes. The quantification of the expression of all genes was performed using SeqMonk.

After exporting the expression values from SeqMonk, we used Notepad++ to search through the whole list to find the expression values of Dntm3l, Dnmt3a, Dnmt3b, and Dnmt1 in each species. A table with the expression values was created(Appendix 4.).
Figures 1-10 show the average values of Dnmt3L, Dnmt3A, Dnmt3B, and Dnmt1 expression in the individual species.

Figure 15. Bar plots: Pairwise alignment scores of multiple sequence alignments for each Intron (1 to 5) sequences of golden hamster, chinese hamster, northern american deer mouse, mongolian gerbil, steppe mouse, mouse, algerian mouse, ryukyu mouse, rat. The columns are in order from comparison of sequences 1:2 to 8:9 and the numbers correspond to species as seen in Figure 23, Appendix 5.

5.3. Local conservation

We then analysed if the region of oocyte-specific promoter sequence is more conserved than the rest of the intron 3 sequence.

A python script was written by the author and utilized to perform the analysis allr all alignments of the introns 1 to 5 of the species golden hamster, chinese hamster, northern american deer mouse, mongolian gerbil, steppe mouse, mouse, algerian mouse, ryukyu mouse, rat to show that the promoter regions are more conserved than other sequences in these intronic sequences (Appendix 2).

Upon implementation of the script, the results (plots) showed that the promoter sequence within intron 3 show high local conservation (Figure 18.), compared to the other parts of the same intron or other intronic sequences of Aire gene (the only other similar region is around base 50 in intron 4)

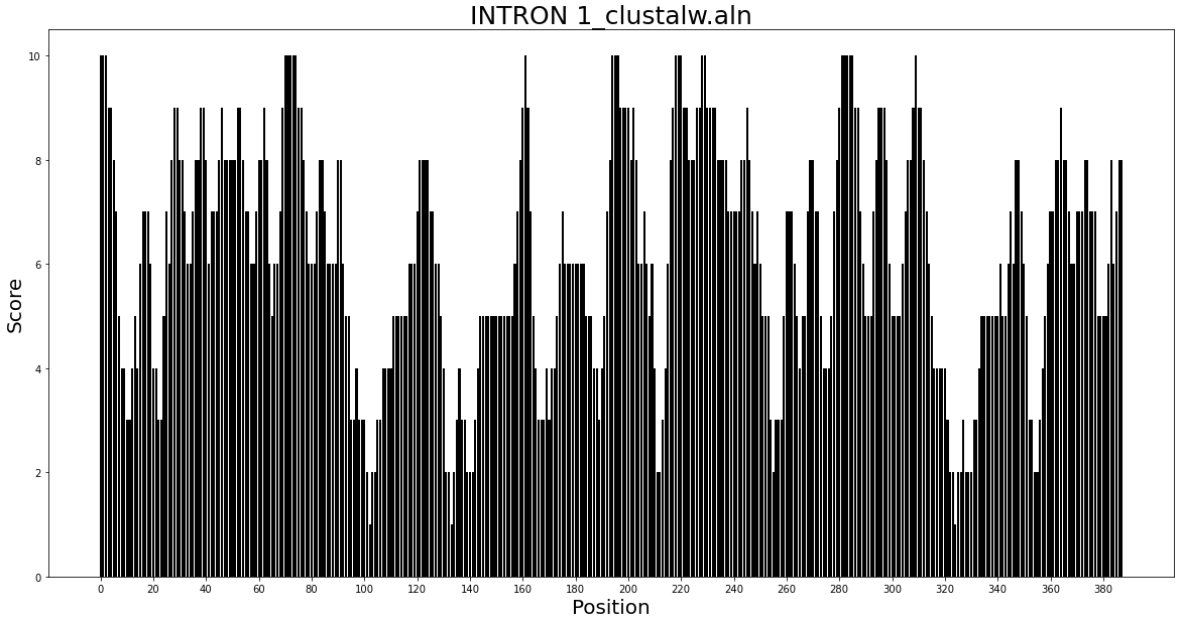


Figure 16. Graphical plot (Intron 1) using Matplotlib library of the Python language.

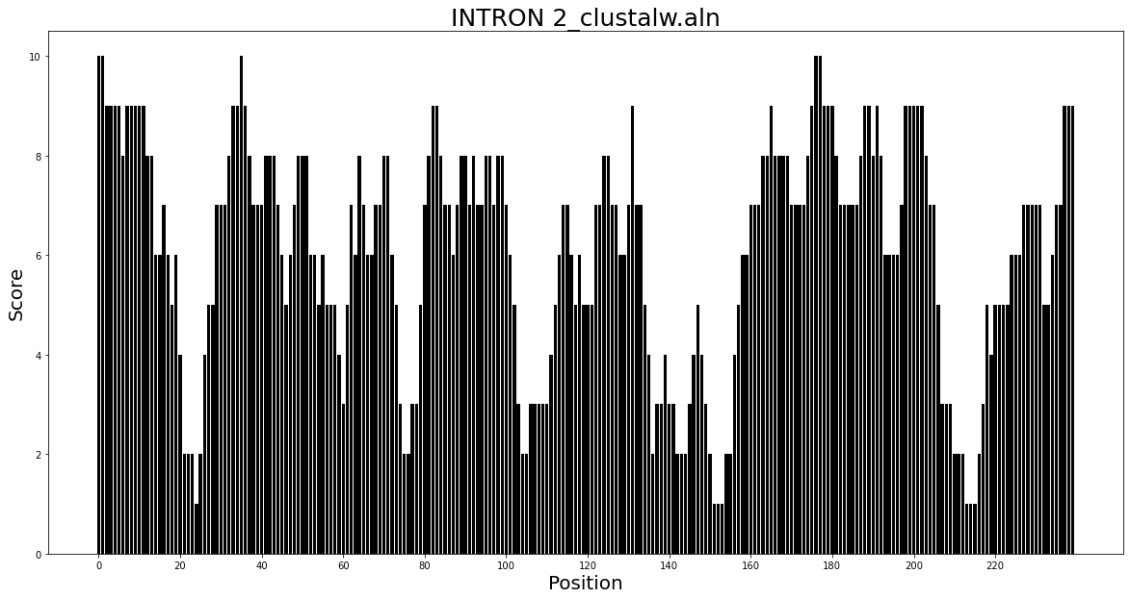


Figure 17. Graphical plot (Intron 2) using Matplotlib library of the Python language.

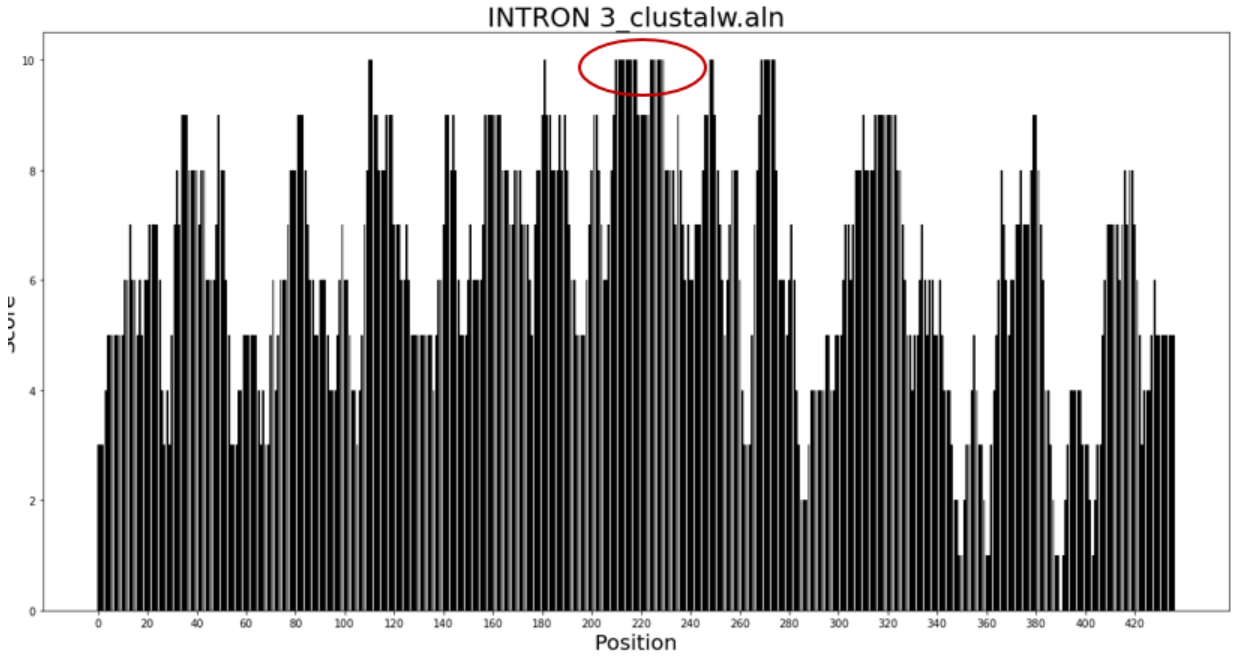


Figure 18. Graphical plot (Intron 3) using Matplotlib library of the Python language.

Highlighted area where it is more conserved shown in red

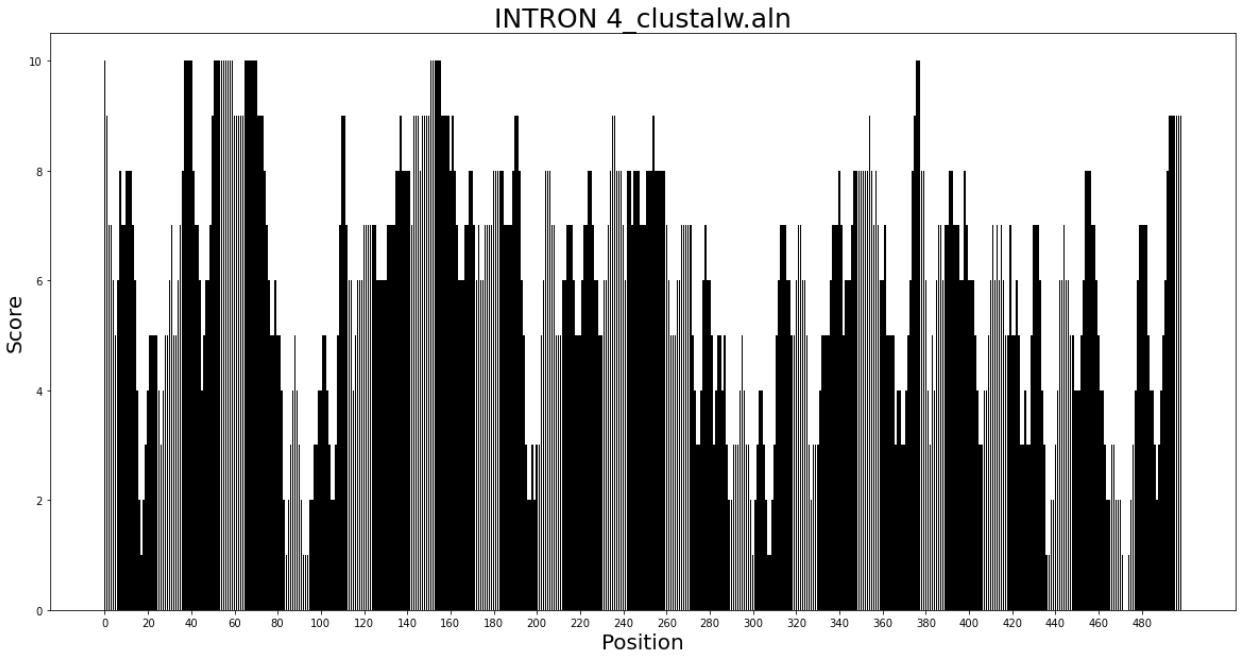


Figure 19. *Graphical plot (Intron 4) using Matplotlib library of the Python language.*

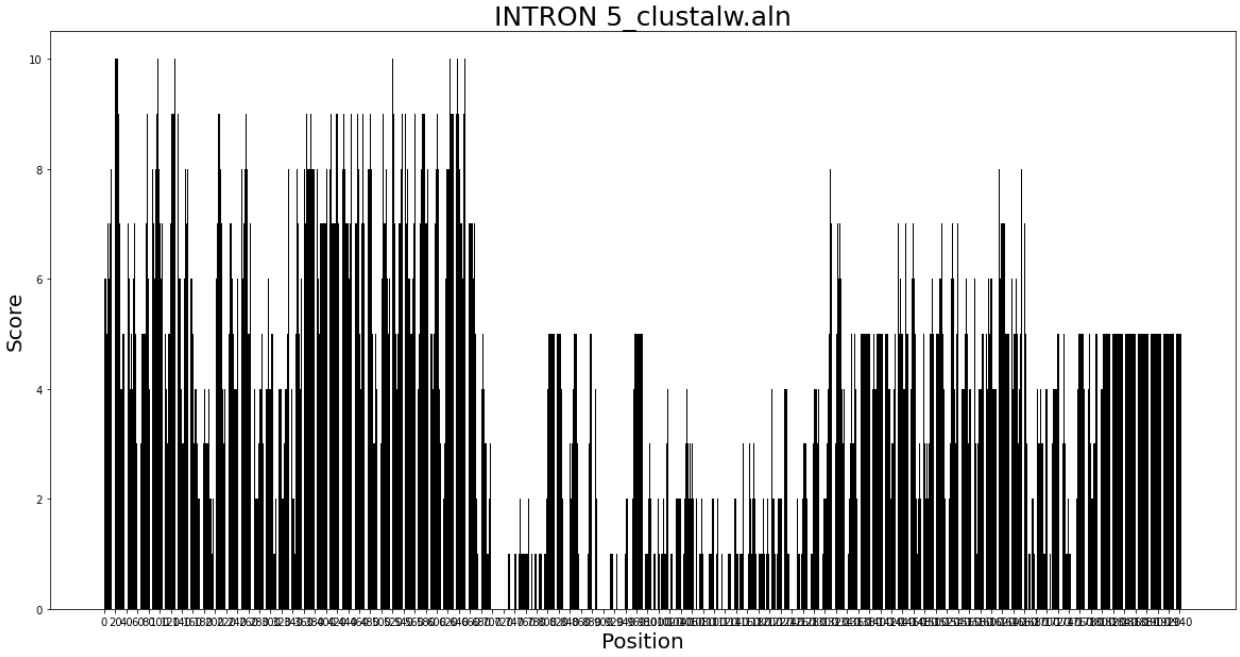


Figure 20. Graphical plot (Intron 5) using Matplotlib library of the Python language.

7. References

Andrews, S. (2007). SeqMonk: A tool to visualise and analyse high throughput mapped sequence data.

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), 1-14.

Chedin, F., Lieber, M. R., & Hsieh, C. L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proceedings of the National Academy of Sciences, 99(26), 16916-16921.

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.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423.

Demond, H., & Kelsey, G. (2020). The enigma of DNA methylation in the mammalian oocyte. F1000Research, 9.

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.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Huntriss, J., Hinkins, M., Oliver, B., Harris, S. E., Beazley, J. C., Rutherford, A. J., ... & Picton, H. M. (2004). Expression of mRNAs for DNA methyltransferases and methyl-CpG-binding proteins in the human female germ line, preimplantation embryos, and embryonic stem cells. Molecular Reproduction and Development: Incorporating Gamete Research, 67(3), 323-336.

Jiang, Z., Sun, J., Dong, H., Luo, O., Zheng, X., Obergfell, C., ... & Tian, X. C. (2014). Transcriptional profiles of bovine in vivo pre-implantation development. BMC genomics, 15(1), 1-15.

Kuznetsov, A., & Bollin, C. J. (2021). NCBI Genome Workbench: desktop software for comparative genomics, visualization, and GenBank data submission. In Multiple Sequence Alignment (pp. 261-295). Humana, New York, NY.

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology, 38(1), 23-38.

Python, R. (2019). Python. Python Releases for Windows, 24.

Shum, J. (2013). Jerry Blogger: Notepad++ v6. 5.1-Current Version with DIRECT DOWNLOAD link.

Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, (1), 2-3.

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.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Computing in science & engineering, 13(2), 22-30.

Van Emburgh, B. O., & Robertson, K. D. (2011). Modulation of Dnmt3b function in vitro by interactions with Dnmt3L, Dnmt3a and Dnmt3b splice variants. Nucleic acids research, 39(12), 4984-5002.

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.

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.

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.

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.

Yokomine, T., Hata, K., Tsudzuki, M., & Sasaki, H. (2006). Evolution of the vertebrate DNMT3 gene family: a possible link between existence of DNMT3L and genomic imprinting. Cytogenetic and genome research, 113(1-4), 75-80.