Investigating gene expression by using next-generation-sequencing data

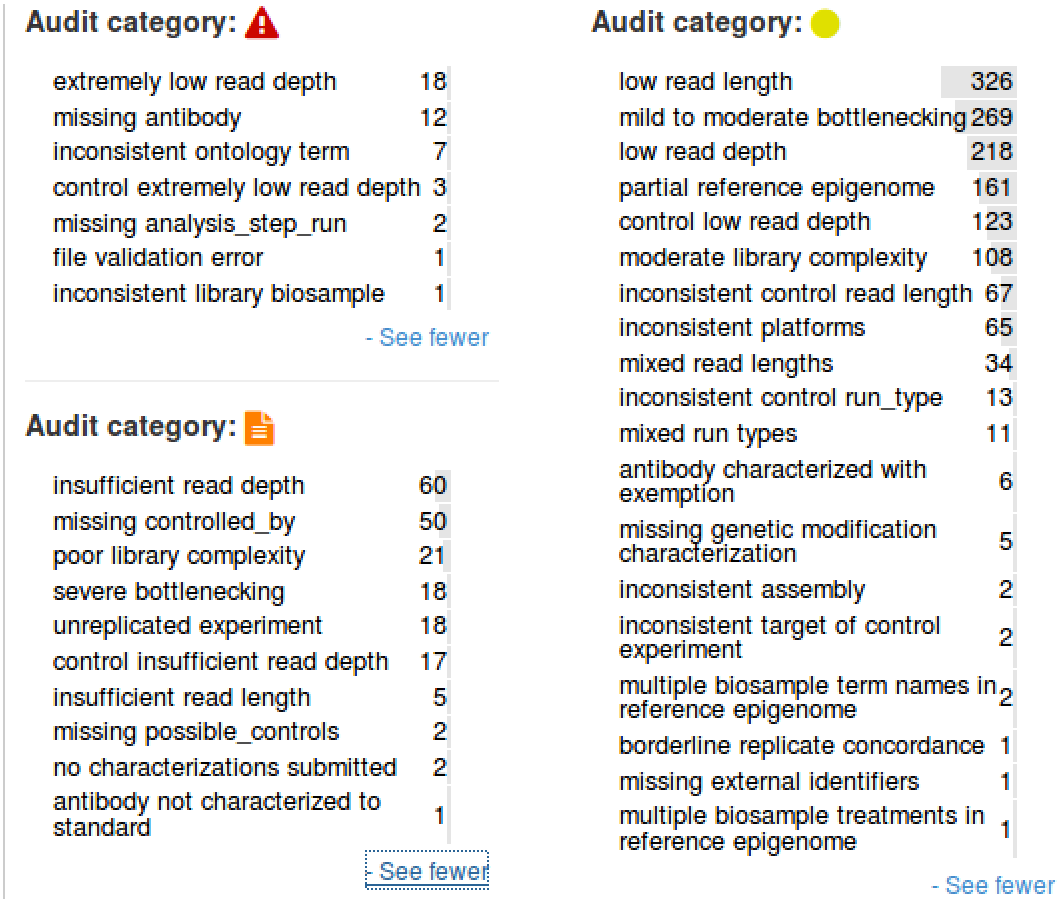
Postgraduate Program:	**Information Technologies in Medicine and Biology**
Course title:	**Algorithms in Molecular Biology**
Student:	**Chatzichronis Stylianos**
Instructor:	**Hatzigeorgiou Artemis**

Abstract

This is a report whose task is to describe the steps followed for the final Perl project on Algorithms in Molecular Biology. All tasks are fulfilled along with the used efficient code with parallel programming and multi-hashing. First, from ENCODE Project [1] the sample ENCFF579ITQ from human adrenal gland has been downloaded. Then, samtools [2] for sorting the bam files, MACS2 [3] software for peak calling are used. A perl script found the expressed genes using tss sites that have been found from Biomart(ensembl tool) [4]. Then, in order to check which transcription factors are bound on those genes, all human TFs from JASPAR [5] have been downloaded along with the promoter regions sequences from biomart. A perl script checked for matching motifs. Those which have not bound by these TF were 11 and another perl script executing MEME[6] software performed de novo motif discovery. The running of the perl scripts and programs was done in an Intel(R) Core(TM) i7-3632QM CPU @ 2.20GHz computer.

Downloading from ENCODE project

The first objective of the project includes finding mapped Chip-seq [7] reads against H3K4me3 from ENCODE project. In the database of ENCODE there are multiple samples. Those reads are uploaded in the database as bam files. In each sample there are multiple issues as shown in the figure 1. Therefore, a sample has been chosen which would have been lacking of such defects. Such sample is isogenic replicated sample 1 (replicate 1) ENCFF579ITQ.bam file from genome located in the dataset with ID ENCSR620TXL [8]. It is adrenal gland sample from human male adult. The control file used from the dataset ENCSR754WVA [9] is the 1st replicate ENCFF579ITQ. Those samples are gathered from genome GRCh38.



Audit category: ⚠

extremely low read depth	18
missing antibody	12
inconsistent ontology term	7
control extremely low read depth	3
missing analysis_step_run	2
file validation error	1
inconsistent library biosample	1

Audit category: 📄

insufficient read depth	60
missing controlled_by	50
poor library complexity	21
severe bottlenecking	18
unreplicated experiment	18
control insufficient read depth	17
insufficient read length	5
missing possible_controls	2
no characterizations submitted	2
antibody not characterized to standard	1

Audit category: ●

low read length	326
mild to moderate bottlenecking	269
low read depth	218
partial reference epigenome	161
control low read depth	123
moderate library complexity	108
inconsistent control read length	67
inconsistent platforms	65
mixed read lengths	34
inconsistent control run_type	13
mixed run types	11
antibody characterized with exemption	6
missing genetic modification characterization	5
inconsistent assembly	2
inconsistent target of control experiment	2
multiple biosample term names in reference epigenome	2
borderline replicate concordance	1
missing external identifiers	1
multiple biosample treatments in reference epigenome	1

Figure 1 (defects of some files where H3K4me3 has been targeted)

Using Samtools

Before, running the most suitable software for chip-seq targeting H3K4me3, the samtools run for sorting and having faster access to the data of the bam file. The commands used were:

samtools sort ENCFF519TFC.bam sample

samtools index sample.bam

Therefore, the sorted file which shall be the input in the call peaking software is the sample.bam file. In the folder ENCSR000DXU there are the files used in subfolders:

ENCFF579ITQ.bam in the subfolder *Sample*

ENCFF519TFC.bam in the subfolder *Control*

The files extracted from samtools in the subfolder *Files*

Running MACS for peak calling

In the article [10] is proved experimentally that MUSIC is the most optimal method for peak calling on H3K4me3 data. However, as it is mentioned in the section Issues the RAM of the computer was not enough to run the software. So, it is chosen the MACS2 as it has the best operating characteristics . The command used is:

macs2 callpeak -t sample.bam -c ENCFF579ITQ.bam -f BAM -n file_peak

This command creates multiple files. The file_peak.r script returns if executed the peak model and the cross correlation in the file_peak_model.pdf file. The file_peak_summits.bed file is used below for the identification of expressed genes. In this file there are 14133 peaks.

Finding expressed genes

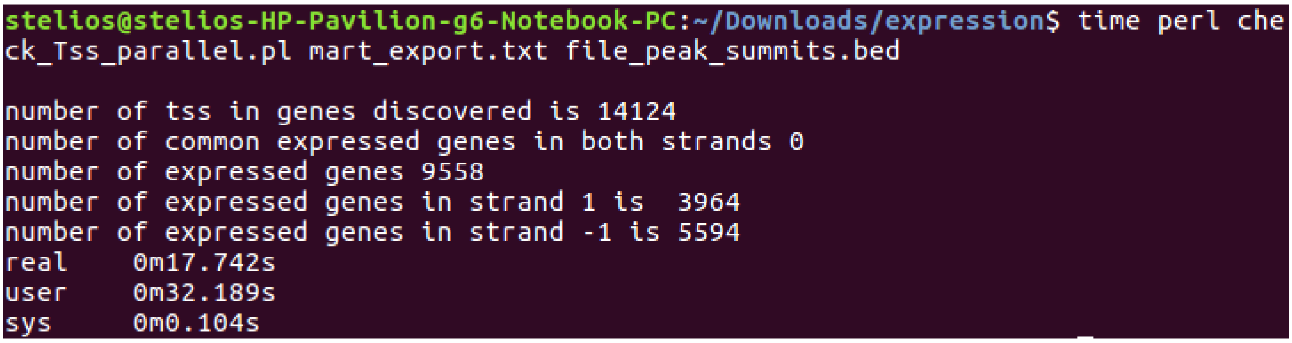
To find the expressed genes first the Tss must be found. From biomart he file mart_export.txt is downloaded, which includes for all human genes the following: Chromosome/scaffold name, Gene stable ID, Gene start (bp), Gene end (bp), Transcription start site (TSS), Strand. It is known that a gene may have multiple tss sites. Therefore, in one gene multiple tss sites must be checked, so there may be multiple promoter regions for each gene [11].

A perl script named check_Tss_parallel.pl is run to find the expressed genes with input the files file_peak_summits.bed and mart_export.txt. The threshold is 2000 bases upstream of each tss. In this file there is multihashing and parallel programming for optimal perfomance. Output of the script are two files. The expr_genes.txt which has the chromosome ID, gene ID, tss , strand, and the expr_pos.txt, which includes chromosome, start of promoter region, end of promoter region, strand.

Executing the script the results shown 9558 expressed genes found, 3964 in + strand and 5594 in the reverse strand in 17.742 s as shown in figure 2.

**All files of that section are located on folder expression*

A primary version of this script without the use of parallel programming is the check_Tss.pl file



```
stelios@stelios-HP-Pavilion-g6-Notebook-PC:~/Downloads/expression$ time perl che
ck_Tss_parallel.pl mart_export.txt file_peak_summits.bed

number of tss in genes discovered is 14124
number of common expressed genes in both strands 0
number of expressed genes 9558
number of expressed genes in strand 1 is  3964
number of expressed genes in strand -1 is 5594
real    0m17.742s
user    0m32.189s
sys     0m0.104s
```

Figure 2 (Running the check_Tss_parallel.pl script)

Finding genes that are bound by known TFs

After finding the expressed genes, the next step was to find which of them are bound by known TFs. The Biomart using as input the expr_pos.txt, returned the corresponding promoter region sequences. The biomart returned the file martquery_1008224531_649.txt, which includes 9106 promoter regions, as for some positions there were not available sequences. The script test1.pl was run for finding the binded TFs. It has input the necessary file JASPAR.txt which includes the 700 TFs profiles as PWM and the martquery_1008224531_649.txt file with the promoter regions. The output of the script is the file TF_file.txt, which has in the last two columns the gene and the TF that this gene is binded. The script took 3719 minutes to run as shown in figure 3.

For optimal perfomance parallel programming, multihashing is used. For threshold is used the maximum score of each PWM (for time consuming reasons). There are methods however, to choose the best threshold for PWM [12,13] Therefore the code is changed (in the first comparison a non-max element is used the iteration is finished before checking the rest of the sequence). A primary version of this script without those improvements is the find_TF.pl file.

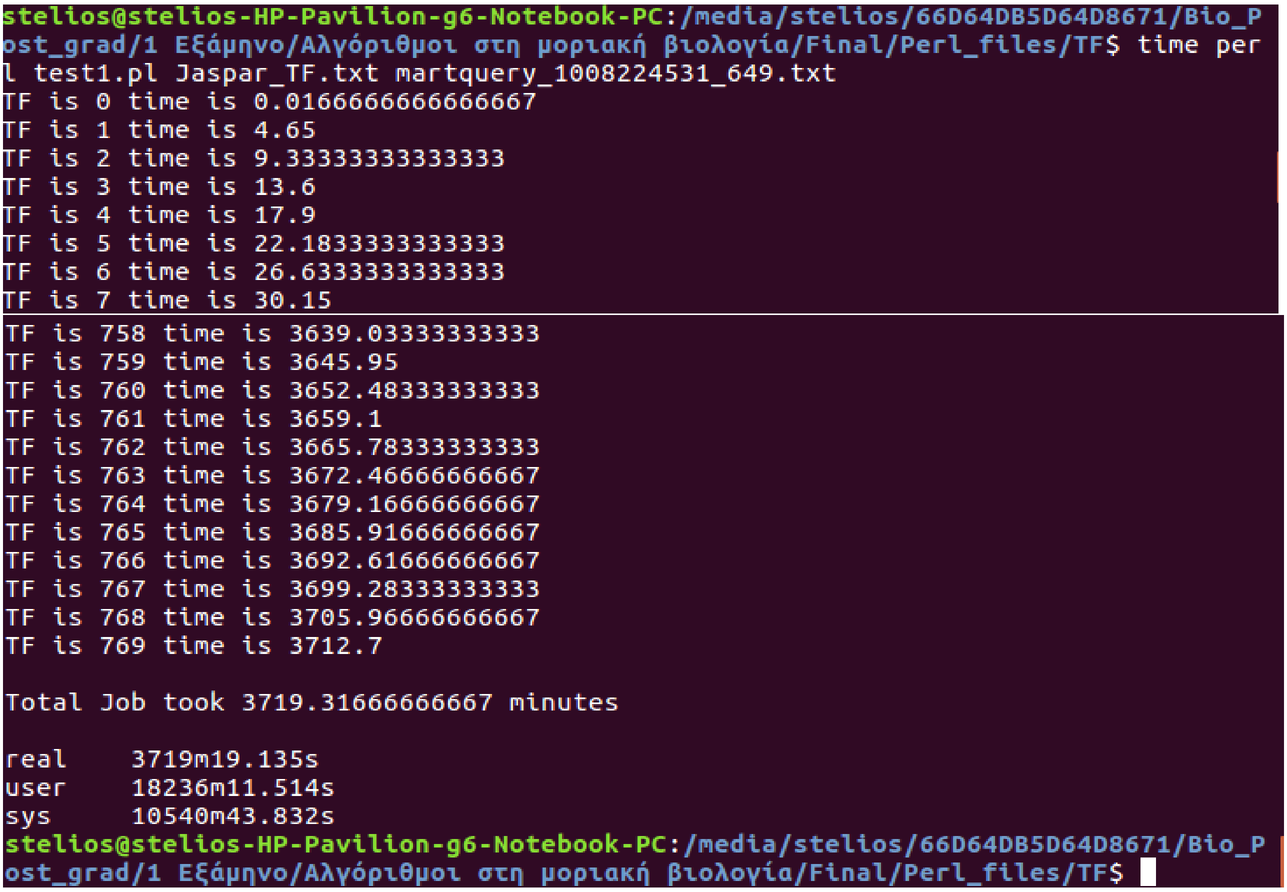
**All files of that section are located on folder TF*.

Next there was a comparison of the results in TF_file.txt with the literature:

In both literature and in TF_file.txt it was found that:

TF name	**TF ID**	**Gene**	**Number** of **genes** found	**Literature**
COUP-TFI/NRF2F1	MA0017.2	ENSG00000168818	1	[14]
VDR	MA0693.2	ENSG00000241135 and another 347	348	[15]
NR1H4	MA1110.1	ENSG00000119013 and another 6	7	[16]
ETV5	MA0765.1	ENSG00000130733 and another 37	38	[16]
FOSL2	MA1138.1 and another 4	ENSG00000276903 and another 37	38	[16]
FOXO4	MA0848.1	ENSG00000160229 and another 1075	1076	[16]

Table 1

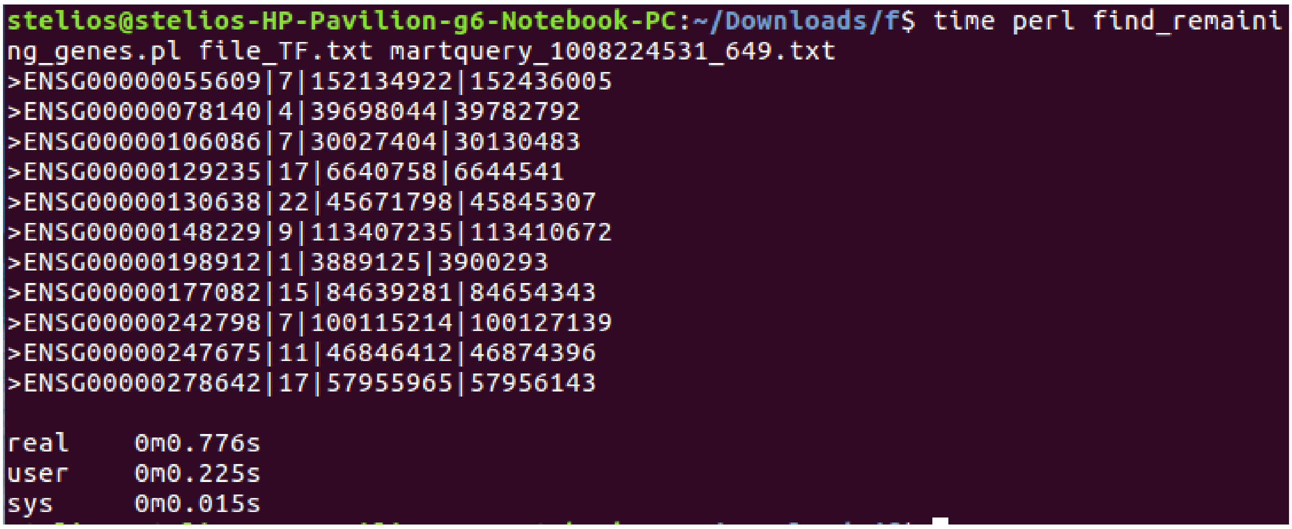


```
stelios@stelios-HP-Pavilion-g6-Notebook-PC:/media/stelios/66D64DB5D64D8671/Bio_P
ost_grad/1 Εξάμηνο/Αλγόριθμοι στη μοριακή βιολογία/Final/Perl_files/TF$ time per
l test1.pl Jaspar_TF.txt martquery_1008224531_649.txt
TF is 0 time is 0.0166666666666667
TF is 1 time is 4.65
TF is 2 time is 9.33333333333333
TF is 3 time is 13.6
TF is 4 time is 17.9
TF is 5 time is 22.1833333333333
TF is 6 time is 26.6333333333333
TF is 7 time is 30.15
TF is 758 time is 3639.03333333333
TF is 759 time is 3645.95
TF is 760 time is 3652.48333333333
TF is 761 time is 3659.1
TF is 762 time is 3665.78333333333
TF is 763 time is 3672.46666666667
TF is 764 time is 3679.16666666667
TF is 765 time is 3685.91666666667
TF is 766 time is 3692.61666666667
TF is 767 time is 3699.28333333333
TF is 768 time is 3705.96666666667
TF is 769 time is 3712.7

Total Job took 3719.31666666667 minutes

real    3719m19.135s
user    18236m11.514s
sys     10540m43.832s
stelios@stelios-HP-Pavilion-g6-Notebook-PC:/media/stelios/66D64DB5D64D8671/Bio_P
ost_grad/1 Εξάμηνο/Αλγόριθμοι στη μοριακή βιολογία/Final/Perl_files/TF$ 
```

Figure 3 (Finding which genes are bound by TF)

De novo motif discovery

Finally for the de novo motif discovery, it is used the find_remaining_genes.pl script. In this file the input is file_TF.txt and martquery_1008224531_649. The output is the de_novo_genes.txt (genes with unknown TFS). It took 0.776 sec (figure 4).



```
stelios@stelios-HP-Pavilion-g6-Notebook-PC:~/Downloads/f$ time perl find_remaini
ng_genes.pl file_TF.txt martquery_1008224531_649.txt
>ENSG00000055609|7|152134922|152436005
>ENSG00000078140|4|39698044|39782792
>ENSG00000106086|7|30027404|30130483
>ENSG00000129235|17|6640758|6644541
>ENSG00000130638|22|45671798|45845307
>ENSG00000148229|9|113407235|113410672
>ENSG00000198912|1|3889125|3900293
>ENSG00000177082|15|84639281|84654343
>ENSG00000242798|7|100115214|100127139
>ENSG00000247675|11|46846412|46874396
>ENSG00000278642|17|57955965|57956143

real    0m0.776s
user    0m0.225s
sys     0m0.015s
```

Figure 4 (Running find_remaining_genes.pl script)

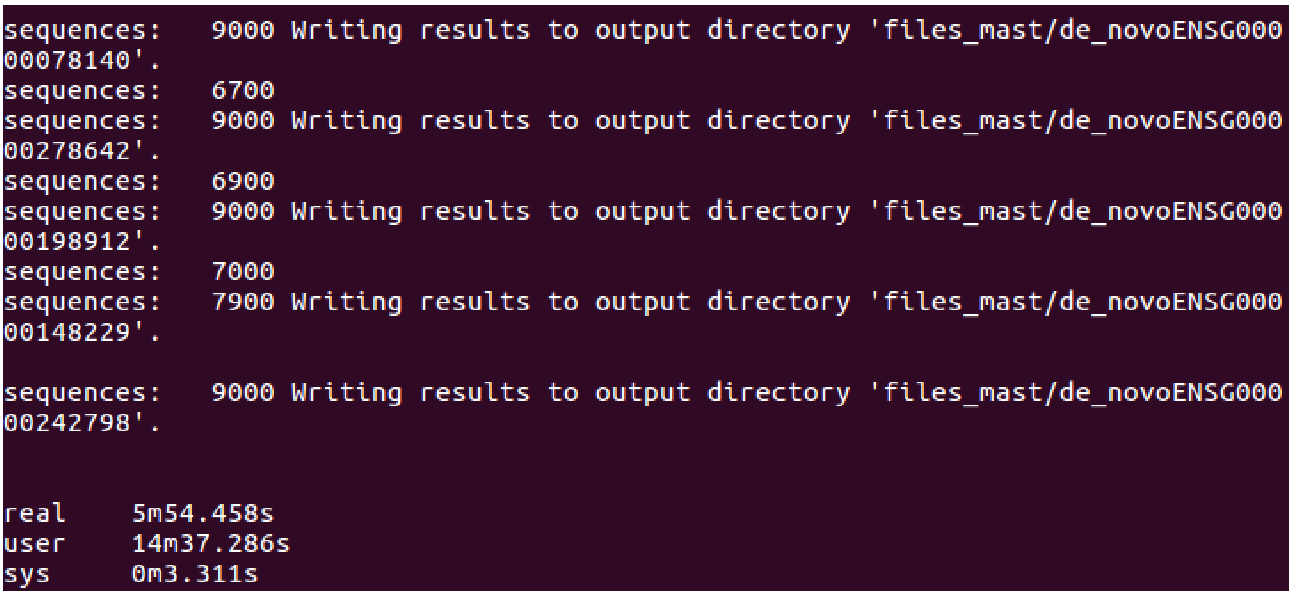
Then manually from gene cards, the related genes were searched and stored in the gene_cards.txt. Then, again manually the file remaining_genes.txt created, which includes all the genes that have Ensembl ID. In the first column is the primary gene and in the next columns(tab-delimited) are the related genes. The criteria to chose genes from gene cards are: highest GH score, highest Gene Association Score and also GH Type shall be promoter. Meeting those criteria is important because high association between genes and interaction at their promoter regions shows similar biological functionality.

Next, the script run_meme.pl is running meme and mast tools. It has as input the related_genes.txt and martquery_1008224531_649.txt. It creates a two dimensional hash with the primary gene as first key and second key is the related gene. The key value is the promoter's sequence. Then, in the folder files_fasta, 11 files are created, each one has the sequences of the gene and its relative genes.

After that, the MEME is executed to find the motifs. The results are returned in 11 folder (each folder for the primary gene examined) in the _meme folder. Then, MAST software is executed and uses those folder to find the actual motifs. Results of mast tool are located on the files_ mast folder. (as mentioned in xml files "Motifs which are grayed-out were very similar to other earlier specified motifs and were removed")

The execution of the script took 5.54 minutes with parallel programming as shown in figure 5.

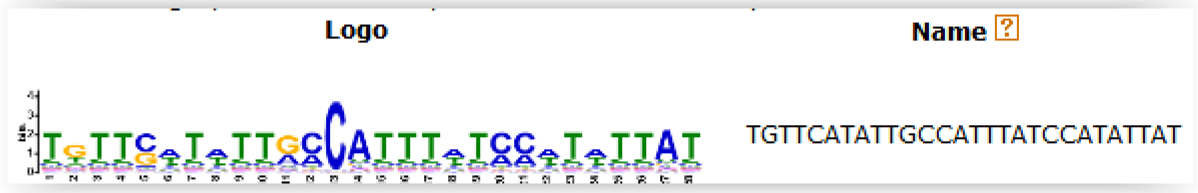
**All files of that section are located on folder De_novo*



```
sequences:   9000 Writing results to output directory 'files_mast/de_novoENSG000
00078140'.
sequences:   6700
sequences:   9000 Writing results to output directory 'files_mast/de_novoENSG000
00278642'.
sequences:   6900
sequences:   9000 Writing results to output directory 'files_mast/de_novoENSG000
00198912'.
sequences:   7000
sequences:   7900 Writing results to output directory 'files_mast/de_novoENSG000
00148229'.

sequences:   9000 Writing results to output directory 'files_mast/de_novoENSG000
00242798'.


real    5m54.458s
user    14m37.286s
sys     0m3.311s
```

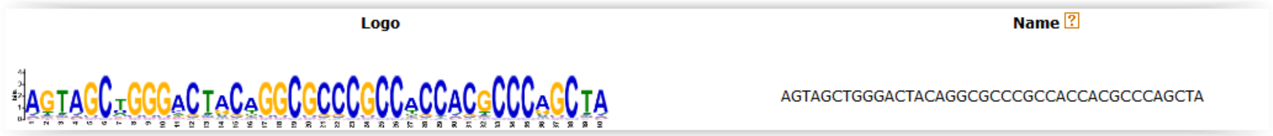
figure 5

For each gene the returned results were:



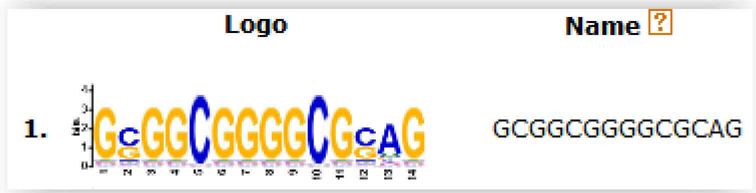
Logo	Name
	TGTTCATATTGCCATTTATCCATATTAT

figure 6 (motif - mast results for ENSG00000055609)



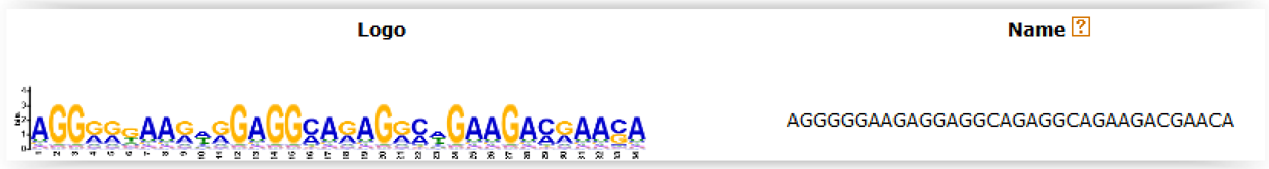
Logo	Name
	AGTAGCTGGGACTACAGGCGCCCGCCACCACGCCCAGCTA

figure 7 (motif - mast results for ENSG00000078140)



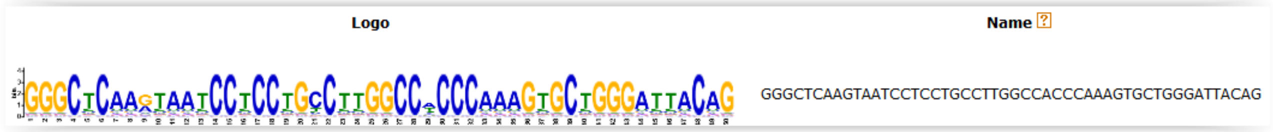
	Logo	Name ?
1.		GCGGCGGGGCGCAG

figure 8 (motif - mast results for ENSG00000106086)



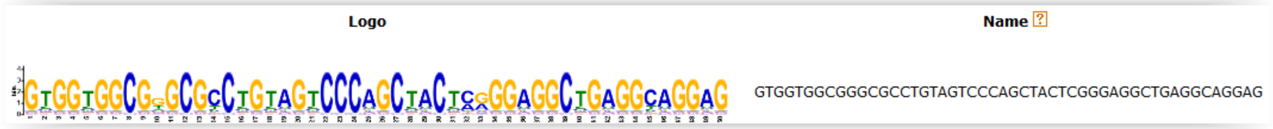
Logo	Name ?
	AGGGGGAAGAGGAGGCAGAGGCAGAAGACGAACA

figure 9 (motif - mast results for ENSG00000129235)



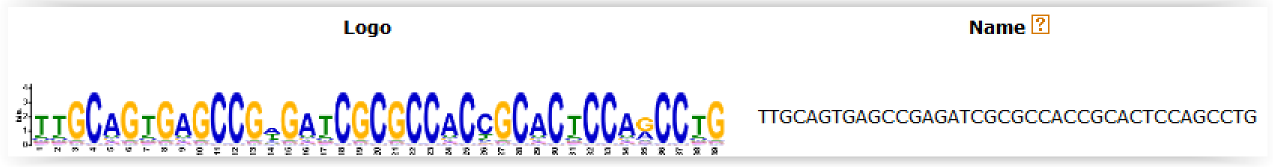
Logo	Name ?
	GGGCTCAAGTAATCCTCCTGCCTTGGCCACCCAAAGTGCTGGGATTACAG

figure 10 (motif - mast results for ENSG00000130638)



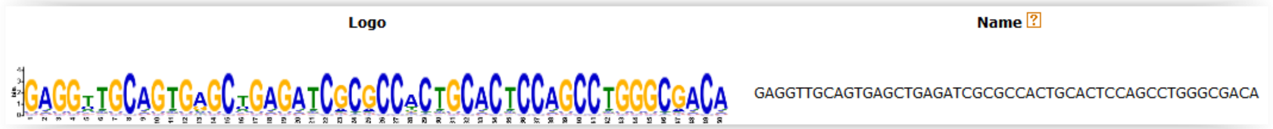
Logo	Name ?
	GTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAG

figure 11 (motif - mast results for ENSG00000148229)



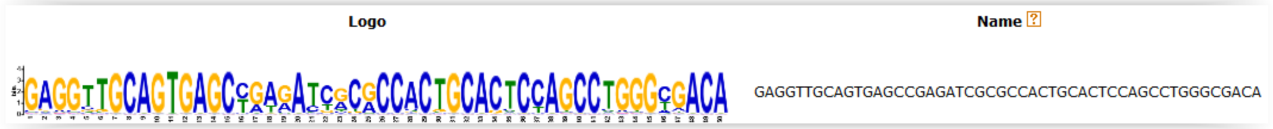
Logo	Name ?
	TTGCAGTGAGCCGAGATCGCGCCACCGCACTCCAGCCTG

figure 12 (motif - mast results for ENSG00000177082)



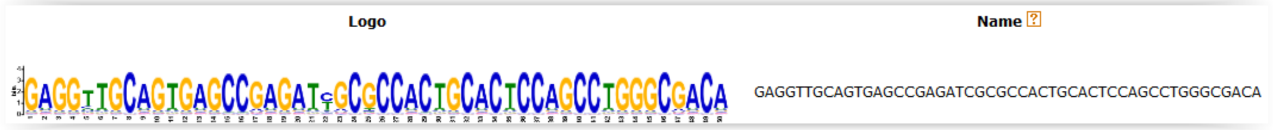
Logo	Name
	GAGGTTGCAGTGAGCTGAGATCGCGCCACTGCACTCCAGCCTGGGCGACA

figure 13 (motif - mast results for ENSG00000198912)



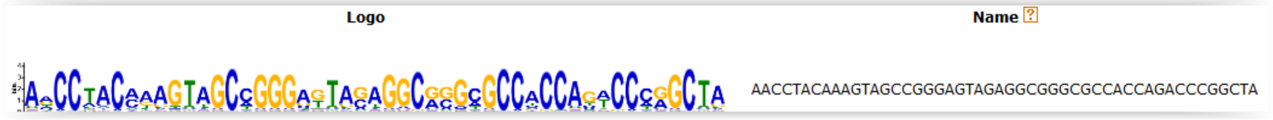
Logo	Name
	GAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACA

figure 14 (motif - mast results for ENSG00000242798)



Logo	Name
	GAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACA

figure 15 (motif - mast results for ENSG00000247675)



Logo	Name
	AACCTACAAAGTAGCCGGGAGTAGAGGCGGGCGCCACCAGACCCGGCTA

figure 16 (motif - mast results for ENSG00000278642)

Issues

Instead of running MACS2, MUSIC [17] first run for peak calling method as shown in the article [11]. After successfully running all the steps required in the command performing the mapping, the following error appears;

```
MUSIC -get_multiscale_punctate_ERs -chip chip/dedup -control input/dedup -mapp hg19_36bp -l_mapp 36 -begin_l 1000 -end_l 16000 -step 1.5

1..Chromosomal cross strand signal fraction threshold is 0.500

Scaling control signal profile.

terminate called after throwing an instance of 'std::bad_alloc'

  what():  std::bad_alloc

Aborted (core dumped)
```

After a short email discussion with Mr. Arif Ozgun Harmanci, who is the developer of MUSIC it seems that my computer has not enough large memory to use MUSIC. Therefore, the alternative solution of using MACS2 has been selected.

References

1) ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.

2) Li H[1], Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.

3) Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. (2008) Model-based Analysis of ChIP-Seq (MACS), Genome Biology, 2008;9(9):R137. https://github.com/taoliu/MACS/

4) Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Giro´n, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, Paul Flicek **Ensembl 2018.** PubMed PMID: 29155950. doi:10.1093/nar/gkx1098

5) Zhao et al (2013), JASPAR 2013: An extensively expanded and updated open-access database of transcription factor binding profiles.

6) Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, 37:W202-W208, 2009.

7) Raha D[1], Hong M, Snyder M. ChIP-Seq: a method for global identification of regulatory elements in the genome. Curr Protoc Mol Biol. 2010 Jul;Chapter 21:Unit 21.19.1-14. doi: 10.1002/0471142727.mb2119s91.

8) https://www.encodeproject.org/experiments/ENCSR620TXL/

9) https://www.encodeproject.org/files/ENCFF579ITQ/

10) Reuben Thomas Sean Thomas Alisha K Holloway Katherine S Pollard. Features that define the best ChIP-seq peak calling algorithms *Briefings in Bioinformatics*, Volume 18, Issue 3, 1 May 2017, Pages 441–450, https://doi.org/10.1093/bib/bbw035

11) Wagner A.Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics. 1999 Oct;15(10):776-84.

12) Marko Djordjevic, Anirvan M. Sengupta, and Boris I. Shraiman. A Biophysical Approach to Transcription Factor Binding Site Discovery Genome Res. 2003 Nov; 13(11): 2381–2390. doi: [10.1101/gr.1271603]

13) Youlian Pan and Sieu Phan. Threshold for Positional Weight Matrix . Engineering Letter, 16:4, EL_16_4_06 https://pdfs.semanticscholar.org/a56a/f8dcbd64ea08383c079fed8d2de6a93e9daa.pdf

14) Marzia Scortegagna, Annabel Berthon, Nikolaos Settas, Andreas Giannakou, Guillermina Garcia, Jian-Liang Li, Brian James, Robert C. Liddington, José G. Vilches-Moure, Constantine A. Stratakis, and Ze'ev A. Ronai. The E3 ubiquitin ligase Siah1 regulates adrenal gland organization and aldosterone secretion. JCI Insight. 2017 Dec 7; 2(23): e97128. Published online 2017 Dec 7. doi: [10.1172/jci.insight.97128]

15) Changlong Bi , Bo Li , Lili Du , Lishan Wang, Yingqi Zhang, Zhifeng Cheng, Aixia Zhai. Vitamin D Receptor, an Important Transcription Factor Associated with Aldosterone-Producing Adenoma. Published: December 20, 2013 https://doi.org/10.1371/journal.pone.0082309

16) A genomic atlas of human adrenal and gonad development. Published online 2017 Oct 23. doi: [10.12688/wellcomeopenres.11253.2]

17) Arif Harmanci Joel Rozowsky and Mark Gerstein. MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework. October 2014 Genome Biology 15(10):474 DOI: 10.1186/PREACCEPT-9116006401338101