Computational Pipeline for Analyses of Genome-Wide Nascent Transcription from PRO-seq Data

SERHAT AKTAY
Abstract

Cells utilise various stress responses that are rapidly activated to avoid cell death under adverse conditions. Tracking gene transcription offers insight into the immediate changes in the cell, and one method to study the process of nascent transcription is precision run-on sequencing (PRO-seq). In PRO-seq, biotin-labeled nucleotides are added to halt the transcription as the RNA polymerase cannot continue transcription after incorporating biotin-labeled nucleotide to the nascent RNA. The biotin-labeled nascent RNAs are then isolated from the myriad of RNAs in the cell and sequenced. Simplifying the analysis of sequenced data and making the analysis more available for a larger group of scientist is needed, and therefore the aim of this thesis is to build a computational pipeline to analyse PRO-seq data. The pipeline consists of five shell and three R scripts that create a reference genome index for alignment, load experimental data, align the data to the reference genome, and output .bed and .bigWig files for further analysis. Using the profile of nascent transcription, the pipeline then identifies functional genomic regions and outputs gene expression activity based on engaged polymerase counts. The data used in this study are from heat shock cells from *Homo sapiens*, *Canis lupus familiaris*, *Mus musculus*, and *Drosophila melanogaster*. This analysis strategy provides a method to visualise gene lengths, map functional genomic regions, count engaged RNA polymerase and identify unannotated genes and enhancers. The analysis showed that the use of bidirectional transcription to study cell stress is more useful in mammals than in insects and that genes encoding chaperone machineries were induced in all organisms upon heat shock. The pipeline developed in this Masters Thesis offers a standardised and user-friendly method to study PRO-seq data and simplifies the analysis for laboratories with less experience in data analysis, additionally it is a tool to handle and automate processing of large amounts of data from distinct organisms. The computational pipeline outputs profiles of engaged RNA polymerases genome-wide, maps functional genomic regions, and counts transcriptional activity of genes and enhancers.

Keywords: PRO-seq, Heat-shock, Reference genome, Pipeline, Functional genomic regions, Gene expression

Sammanfattning


Nyckelord: PRO-seq, Värmechock, Referensgenom, Pipeline, Funktionella genomiska regioner, Genuttryck
Contents

1 Background 3
   1.1 Eukaryotic Transcription .................................. 3
   1.2 Divergent Transcription ...................................... 4
   1.3 Cellular Stress .............................................. 4
   1.4 Precision Run-On Sequencing ............................... 5
   1.5 Mapping Polymerase Data ................................... 5

2 Aims ......................................................... 6

3 Materials and Methods 6
   3.1 Data Gathering ................................................. 6
      3.1.1 Homo sapiens .............................................. 6
      3.1.2 Mus musculus ............................................. 6
      3.1.3 Drosophila melanogaster ............................... 7
      3.1.4 Canis lupus familiaris .................................. 7
   3.2 Pipeline Setup ................................................ 7
      3.2.1 Required tools ........................................... 7
      3.2.2 Input Data ................................................ 7
      3.2.3 Output Data ............................................... 8

4 Results .................................................. 8
   4.1 Automating the Code for Obtaining and Building Reference Genomes 9
   4.2 Loading Sample Data ......................................... 9
   4.3 Aligning Data to Reference Genome and Converting to bigWig Format 12
   4.4 Identification of Functional Genomics Regions in Fly, Mouse, Dog and Human 15
   4.5 Quantification of Transcription Upon Acute Stress .................. 23
   4.6 Comparison of Functional Annotation in Human Genome Versions .......... 27

5 Discussion ............................................... 27
   5.1 Identification of Functional Genomic Regions in Fly, Mouse, Dog and Human 27
   5.2 Quantification of Transcription Upon Acute Stress .................. 28
   5.3 Comparison of Functional Annotation in Human Genome Versions .......... 28
   5.4 Computational Pipeline ....................................... 28
   5.5 Limitations and Future Studies ................................ 28
1 Background

The central dogma states that information flows from DNA to RNA (transcription) and RNA to proteins (translation). As DNA contains the information necessary for all building blocks that cells need, transcription determines how the cell functions and what tasks the cell performs [1]. Prokaryotes utilise a single RNA polymerase (Pol) while the transcription in eukaryotes is performed by three RNA polymerases. Pol I and Pol III synthesise ribosomal RNAs and small structural RNAs while Pol II synthesises protein-coding mRNAs, and other RNAs such as long noncoding RNAs, microRNAs and enhancer RNAs [1, 2]. Bacteria have two main ways to regulate and initiate transcription, either through specific transcription factors (TF) that interact with promoters to alter their ability to bind RNA polymerase or by TFs interacting with the RNA polymerase to affect the polymerase promoter preference. In archea, TFs direct the RNA polymerase to specific DNA sites. Eukaryotes utilises a number of TFs that bind DNA and alter the chromatin conditions, or help form a pre-initiation complex (PIC) consisting of general TFs and Pol II. [2]

1.1 Eukaryotic Transcription

The transcription of genes in eukaryotes is coordinated at seven (1-7) rate-limiting steps; promoter opening (1), transcription initiation (2), promoter-proximal Pol II pausing (3), transcriptional elongation (4), co-transcriptional processing (5), transcription termination (6), and Pol II recycling (7) [3]. See figure 1 for a schematic overview of the transcription cycle.

Figure 1: The steps of transcription with the green nucleosomes, activators (orange and yellow), GTFs (blue), Pol II (red), nascent RNA (purple line), negative elongation factor (purple circle), positive elongation factor (blue triangle) and phosphorylations in blue and red (P). [4]

Figure 1: The steps of transcription with the green nucleosomes, activators (orange and yellow), GTFs (blue), Pol II (red), nascent RNA (purple line), negative elongation factor (purple circle), positive elongation factor (blue triangle) and phosphorylations in blue and red (P). [4]

but the new unshielded 5’ end of the nascent transcript is accessible for the 5’- to 3’-exoribonuclease XRN2 which degrades RNA and causes termination of transcription. (7) Finally, the transcription machinery can be recycled by looping the chromatin bringing the 3’ end into the vicinity of the TSS or by the chromatin generating an insulating region where the machinery can be retained. [3]

The measurement of RNA polymerase density provides a quantitative snapshot of the transcriptional activity across the genome [5]. A series of these snapshots can reveal the genes that respond to specific signals and reveal mechanisms of transcription regulation [5]. In addition to protein-coding genes, Pol II also binds and transcribes other regions such as enhancers [6]. Therefore, understanding the transcriptional changes in response to environmental, developmental, or nutritional factors could provide new insights into the regulation of cellular mechanisms [5].
1.2 Divergent Transcription

In eukaryotes, the idea of the TATA-box directing formation of the PIC has dominated, but it has been shown that the TATA box is found only 10-20% of promoters and therefore drives transcription of a minority of genes [7]. Rather, promoter elements such as CpG islands, initiator elements (Inr), downstream promoter elements (DPE), and TFIIB recognition elements (BRE) drive transcription in most cases [7]. CpG islands recruit the TATA-box binding protein (TBP) by the use of sequence-specific transcription factors that bind to the CpG sequence [8]. The TFIIB can cause the PIC to form on both sides of the CpG sequence which leads to a bidirectional promoter and gives rise to divergent transcription (figure 2) [8].

A majority of mammalian genes show divergent transcription generating short (50-2000 nt) and unstable upstream antisense RNAs (uaRNA) [8]. Divergent transcription also occurs at distant regions giving rise to enhancer RNA (eRNA) [9] and most intergenic transcription in mammalian cells are from divergent transcription of enhancers [8]. The function of divergent transcription is, however, currently not known [8]. In addition to mammals, divergent transcription has also been observed in prokaryotes, both bacteria and archaea [10] as well as in eukaryotes such as yeast and flies (Drosophila) [11, 12]. Overall, studying the pattern of divergent transcription presents a method to study active transcription by analysing the nascent RNA [11].

1.3 Cellular Stress

Cells encounter multiple types of conditions which can hamper homeostasis. Such stress conditions include the presence of toxins, changes in temperature, hypoxia, limited nutrition, and many others. Cells have a variety of ways to respond to stress, from activation of pathways that help them survive to programmed cell death. Initially, cells strive to defend themselves from the stress but if the cause of stress is unresolved it invariably leads to cellular death. Therefore, the ability to counteract environmental or intracellular stress is highly conserved and can be found in all organisms, from bacteria to mammals. The heat shock response is one of the main survival mechanisms and is not limited to activation upon elevations of temperature (3-5 °C above normal); it is also activated upon heavy metal interactions and oxidative stress. These types of stresses cause protein damage, such as protein aggregation of unfolded proteins, which is therefore the main target of the heat shock response. Part of the stress response is the increase of chaperones, proteins that help to refold misfolded proteins and reduce protein aggregation. Chaperones improves the cells resistance to temperature increases, oxidative stress, and other chemical stressors. [13]

During the initiation of the heat shock response, general transcription is inhibited. Simultaneously, TFs, including heat shock factors (HSFs), are activated and enhance the expression of cytoprotective genes. Heat shock factor 1 (HSF1) is essential for the heat shock response and is maintained in monomeric form in the cytoplasm where it is bound to heat shock protein 90 (Hsp90). Upon heat shock, Hsp90 binds to unfolded proteins, which releases monomeric HSF1 molecules and allows them to move into the nucleus, bind to DNA and promote the expression of chaperones, including heat shock proteins (Hsps), co-chaperones and ubiquitin. [13]

The heat shock response consists of 50-200 genes in different model organisms that give rise to seven functional classes of Hsps that protect the cell in different ways. The main class (I) is the "molecular chaperones," which are the most abundantly expressed. Other classes have different functions, such as protecting against misfolded proteins (II), nucleic acid modifications (III), metabolic enzymes (IV), regulatory proteins such as transcription factors (V),
sustaining cellular structures (VI) and transport, detoxification, and membrane-modulating proteins (VII). The expression of different classes varies in species, with class IV showing the greatest variance. The most abundant class, molecular chaperons, comprises a number of different Hsps - Hsp100, Hsp90, Hsp70, Hsp60 and small heat shock proteins (sHsps). [14]

The Hsp100s consists of AAA ATPases that are dynamic hexameric structures responsible for refolding misfolded proteins, supporting protein disaggregation and can participate in protein quality control pathway in mammals. Hsp90 is found in high concentrations in the cytosol of eukaryotes and bacteria and binds native-like proteins during stress to reduce aggregation of unfolded proteins. Hsp70 is one of the most conserved chaperones (DnaK in bacteria shares a 60% sequence identity with eukaryotic Hsp70), and is involved in de novo folding of proteins and the refolding of aggregates during physiological conditions and is therefore essential in the heat shock response. The sHsps vary greatly and usually form large oligomers of up to 24 subunits that are ATP-independent chaperones. sHsps bind to a large number of partially folded proteins to prevent aggregation and then store the unfolded proteins that can be refolded in the presence of other chaperones such as Hsp70 and Hsp100. [14]

The response to heat stress has been previously studied using genome-wide methods focusing on transcriptional changes. Insects (D. melanogaster) and mammals (M. musculus, C. lupus familiaris, and H. Sapiens) use HSFs that activate Hsp genes by releasing paused Pol II into elongation. [3] One powerful method to study transcriptional changes is the identification of divergent transcription as a tool to map active transcription upon heat shock [11].

1.4 Precision Run-On Sequencing

RNA polymerase mapping at base pair resolution is required to accurately study transcriptional elongation and polymerase pausing which can be achieved by using precision run-on sequencing (PRO-seq). In PRO-seq, the nuclei of millions of cells are isolated and biotin-labelled nucleoside triphosphates (NTPs) are added, which inhibit further elongation by Pol II when they are incorporated into nascent RNAs. The biotin-labelled nascent RNA transcripts are then affinity-enriched by using streptavidin coated magnetic beads. When sequencing the 3' end of the nascent RNA, the last active position of the polymerase is revealed through the last incorporated NTP. To find the TSS, 5’-sequencing is utilised. [5]

The benefits of PRO-seq are that it is a base-pair resolution method to reveal the positions of transcriptionally engaged strand-specific RNA polymerase molecules. PRO-seq has a very low background signal, is highly sensitive (dynamic range of 10^5) and can identify short and unstable nascent eRNAs. However, since PRO-seq only detects actively transcribing polymerase, it cannot differentiate between the different types of polymerases (Pol I, II, and III), and since it only incorporates one or a few NTPs, some transcripts from RNA polymerases close to the TSS might not be long enough to be uniquely aligned to the genome. [5]

1.5 Mapping Polymerase Data

By studying the polymerase profile from PRO-seq, a general shape can be distinguished for actively transcribed genes and enhancers, that are expected to show divergent transcription around the TSS (figure 3). Genes are identified by identifying the different ends of transcription; the transcription start site (TSS) and the cleavage and polyadenylation (CPS) are determined from previously identified genes and set as the start and end of transcription. Other elements of transcription are determined relative to the TSS and CPS as divergent transcription between -750 and -251 from TSS, promoter proximal region is set between -250 and +249 nt from the TSS, the gene body is found between +250 from the TSS and -501 nt from CPS. The CPS is set as
the area from -500 to +499 from transcription end and the termination windows is set from +500 to +10499 from transcription end. [15]

Mapping RNA polymerase distribution reveals the quantity of transcription at enhancers, promoter-proximal regions, gene bodies, cleavage and polyadenylation sites (CPSs), termination windows, and the extent of divergent transcription from promoters. By studying the pattern of divergent transcription, active transcriptional regulatory elements (TREs) can be revealed [16]. TREs are elements such as promoters, enhancers, and insulators, all essential in the regulation of transcription by inhibiting chromatin packing, initiating transcription, and releasing paused Pol II [17]. Since methods like PRO-seq measure primary transcription before exosome degradation, it is particularly useful to detect even transient eRNAs in addition to other RNA [5]. TREs can be detected from PRO-seq data using the tool "Detection of Regulatory DNA Sequences using GRO-seq Data" (dREG) created and published by Danko et al. in 2015 [16]. dREG is a machine-learning model that uses PRO-seq read counts and then employs support vector regression to find TREs from their distinctive pattern [16].

Using dREG to call TREs from PRO-seq and genome mapped data, the intersection of dREG calls and Pol II density across the genome gives us information about functional genomic regions and by comparing it with data from different time-points, an inference on gene regulation can be made.

2 Aims

The aim of this study is to develope a data analysis pipeline that can be used to analyse PRO-seq data by mapping the functional genomic regions and to compare data sets with different conditions, such as heat shock vs control, to determine changes in gene regulation. The vast amount of data available requires tools that can simplify the computational analysis needed.

In addition to developing a pipeline, the goal is to study the heat shock response in four different organisms, Homo sapiens, Mus musculus, Drosophila melanogaster, and Canis lupus familiaris using the pipeline. The study is focused on tracking the heat shock response by mapping functional genomic regions in all four species and comparing the count of engaged Pol II in HSP4A in a control sample with a heat-shocked sample. Additionally, the comparison between control sample and heat-shocked sample consists of studying the difference in gene expression overall by looking at the most up- and downregulated genes.

3 Materials and Methods

The methodology of the study consists of two parts. Data gathering from published and unpublished sources followed by the building of an analysis pipeline.

3.1 Data Gathering

The data used in the pipeline consists of published and unpublished PRO-seq data from the Vihervaara laboratory at SciLifeLab in Stockholm.

3.1.1 Homo sapiens

The heat shock data for Homo sapiens is found in the Gene Expression Omnibus with the SRR numbers 8669162 and 8669163 for the control and 8669164 and 8669165 for the heat-shocked sample. The data is from human erythroleukemia (K526) cells that were kept at 37°C and the heat-shocked samples were treated by submerging the cell culture in a 42°C water bath for 30 minutes. Drosophila S2-cells were used as spike in and the sequencing was performed using Illumina NextSeq500. [18]

3.1.2 Mus musculus

The heat shock data for Mus musculus is found in the Gene Expression Omnibus with the SRR numbers 8713360 and 8713361 for the control and 8713362 and 8713363 for the heat-shocked sample. The data is from mouse embryonic
fibroblasts (MEFs) cells that were kept at 37°C and the heat-shocked samples were treated by submerging the cell culture in a 42°C water bath for 25 minutes. Drosophila S2-cells were used as spike in and the sequencing was performed using Illumina NextSeq500. [18]

3.1.3 Drosophila melanogaster
The heat shock data for Drosophila melanogaster consist of two samples, control and 20 min heat-shocked cells. The data is from Drosophila S2 cells that were kept at 22°C and heat-shocked samples were treated by submerging the cell culture in a 37°C water bath for 25 minutes. MEF-cells were used as spike in and the sequencing was performed using Illumina Novoseq 6000. The S2 cell samples were generated in collaboration with Maria Vartiainen laboratory (University of Helsinki, Finland) and are currently unpublished.

3.1.4 Canis lupus familiaris
The heat shock data for Canis lupus familiaris consists of two samples, control and 30 min heat-shocked cells. The data is from golder retriever macrophage (DH82) cells that were kept at 37°C and the heat-shocked samples were treated by submerging the cell culture in a 42°C water bath for 30 minutes. Drosophila S2-cells were used as spike in and the sequencing was performed using Illumina Novoseq 6000. The DH82 cell samples were generated in Vihervaara laboratory and are currently unpublished.

3.2 Pipeline Setup
The analysis pipeline consists of two parts, data processing and analysis. The data processing consists of preparing a version of the genome to align the sample to, downloading and/or concatenating replicates, and converting the sample .fastq files to .bed and .bigWig files. The bigWig files are then uploaded to the dREG portal that identifies the TREs and the output is then used in the analysis that maps the functional genomic regions and lists the differentially expressed genes.

The pipeline is written as five separate .sh files and three R scripts and is initialised by running the "master_script.sh" that loads all other required scripts and creates all folders needed. All paths in the script are relative to the original .sh file.

3.2.1 Required tools
To run the analysis pipeline, a number of packages need to be installed. See table 1 for a list of required tools.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Software and algorithms</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>bedtools</td>
<td>Quinlan AR and Hall IM, 2010 [19]</td>
<td>v2.30</td>
</tr>
<tr>
<td>bowtie2</td>
<td>Langmead et. al. 2012 [20]</td>
<td>v2.5.1</td>
</tr>
<tr>
<td>SRA-toolkit</td>
<td><a href="https://github.com/ncbi/sra-tools">https://github.com/ncbi/sra-tools</a></td>
<td>v3.03</td>
</tr>
<tr>
<td>fastx_clipper</td>
<td><a href="http://hannonlab.cshl.edu/fastx_toolkit">http://hannonlab.cshl.edu/fastx_toolkit</a></td>
<td>v0.013</td>
</tr>
<tr>
<td>samtools</td>
<td>Danecek et. al. 2021 [21]</td>
<td>v1.17</td>
</tr>
<tr>
<td>fetchchromsizes</td>
<td><a href="http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.arm64/fetchChromSizes">http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.arm64/fetchChromSizes</a></td>
<td></td>
</tr>
<tr>
<td>bedGraphToBigWig</td>
<td><a href="https://github.com/ucscGenomeBrowser/dREG">https://github.com/ucscGenomeBrowser/dREG</a></td>
<td></td>
</tr>
<tr>
<td>dREG</td>
<td>Wang et. al. 2019 [22] dreg.dnasequence.org/</td>
<td>v1.0</td>
</tr>
<tr>
<td>bgTObi</td>
<td>Hojoong Kwak, Cornell University, NY, USA</td>
<td>2011</td>
</tr>
</tbody>
</table>

3.2.2 Input Data
The pipeline is built so that the only input required are samples in .fastq formatted files. It supports working with data from Homo sapiens, Mus musculus, Drosophila melanogaster, and Canis lupus familiaris, and mapping them to the most current genome versions and in the case of Homo sapiens it supports both the most current genome (T2T-chm13v2/hs1) and the previous genome (hg38). The sample data can be supplied in one of three ways: example data from the Gene Expression Omnibus (GEO), input SRR numbers to download files from GEO, or supplied
as local files. The pipeline also needs information the positions and sequences of: 1) 3'- and 5'-adaptors, and if applicable, 2) UMIs, and 3) in-line barcodes.

### 3.2.3 Output Data

The data processing yields three folders within the organism folder: the "genome" folder containing bowtie2 files for genome alignment, chromosome information (sizes), the genome (.fa format), the "samples" folder containing the .fastq files, and the "bigWig" folder contains .bed files, .bam files, plus and minus bigWig and bedgraph files, normalised and un-normalised, as well as spikeIn files. The files named "_allMap_unnorm_mn.bigWig" and "_allMap_unnorm_pl.bigWig" are then uploaded to the dREG platform and the output file ending in ".dREG.peak.full.bed.gz" is downloaded, gunzipped, and used as an input for the analysis.

The analysis steps yield a folder named "analysis" containing folders named "functionalGenomics_sample" containing .bed files for all different transcriptional elements, .txt-files with gene coordinates and a concatenated .bed file with all genes named "functionalGenomicRegions_sample" that can be viewed in a genome browser such as IGV. The step "comparing gene expression" generates three .txt-files, one with all genes that show a difference in expression between the two sample, one that shows the top 50 upregulated genes and one with the top 50 downregulated genes.

### 4 Results

At the start, the pipeline (./master_script.sh) asks the user to: select the organism you want to work with (and genome version) and secondly what you want to do, see figure 4 for terminal output. The first selection, generates the variable $organism and creates a folder named after the genome version in the same folder where the script is located.

![Figure 4: The options upon initiating the pipeline in a terminal window.](image)
4.1 Automating the Code for Obtaining and Building Reference Genomes

The first option in the pipeline "Download genome and build bowtie2 alignment" runs the code in figure 5. It creates a subdirectory for the genome in which it downloads the reference genome from UCSC, and then it downloads a list of chromosomes and their sizes. Finally it uses bowtie2-build to create an index of the genome which is then used for alignment.

```bash
$organism = $(echo $organism | sed 's/ /_/g')

mkdir -p $organism/genome
cd $organism/genome

# download the reference genome from NCBI, rename the file "reference genome.fa.gz" and unzipp the .gz FASTA file:
wget -O - https://mgdownload.soe.ucsc.edu/bigZips/$organism.fa.gz > genome_$organism.fa.gz
fetchChrSize $organism > chrm_sizes_$organism.temp
grep -v _ chrm_sizes_$organism.temp > chrm_sizes_$organism.txt

gunzip genome_$organism.fa.gz
bowtie2-build genome_$organism.fa bowtie2_$organism
rm *.*
cd ...
```

Figure 5: Code for downloading reference genome and creating a bowtie2 index for sequence alignment.

4.2 Loading Sample Data

The second option in the pipeline "Prepare experimental data", initiates code (figure 6) that prompts the user to chose the source for the experimental data. The options load different parts of the script that downloads the data in .fastq format. The script downloads experimental data from Gene Expression Omnibus (GEO), (figure 7), inputing your own SRR numbers to download data from GEO, (figure 8), or loading files from the folder "data" found in the "organism"-folder, (figure 9).

```bash
$organism = $(echo $organism | sed 's/ /_/g')

mkdir -p $organism/samples

title="Do you want to download fasta-files from Gene Expression Omnibus or do you want to load your own fasta files?"
prompt="Pick an option:"
options=(["Download example data from GEO" "Input my own SRR numbers" "Load my own files"]

echo "$title"
PS3=""""
IFS=$'
' select opt in "$options[@]" "Quit"; do
    case "$opt" in
        "Download example data from GEO")
            fasta_source="GEO";
        ;;
        "Load my own files")
            break
        ;;
        "Input my own SRR numbers")
            fasta_source="Input SRR";
        ;;
        "Quit")
            exit 0
        ;;
    esac
done

./download_fastq.sh
```

Figure 6: The options upon initiating the sample loading.

The option to download experimental data from GEO downloads four datasets for the chosen organism, two replicates of the control sample and two for the heat-shocked sample using the fasterq-dump package that takes SRR-numbers as input and outputs .fastq-files. The replicates are then moved into the folder "samples". This option currently does not support downloading example data for Drosophila melanogaster or Canis lupus familiaris. See Figure 7 for the code.
Figure 7: The code for downloading pre-selected samples from the gene expression omnibus.

The option to input SRR numbers first prompts the user for whether or not there are replicates, the number of samples and the number of replicates. Then it asks for the SRR numbers followed by downloading and naming the files "sample_n.fastq" where n is the number starting from 1 and then moving the files into the folder "samples". See Figure 8 for the code.
The option to load files from a folder prompts the user to move the files into the folder named "data" and making sure that the file names contain the phrase "sample_n". It then offers to demultiplex the data if barcodes where used in the sequencing. This requires a .txt file named "barcodes.txt" in the data folder with a list of sample names and corresponding barcodes. It then uses the package fastx_barcode_splitter to demultiplex the data. The code then renames the files and moves the demultiplexed (or original) data into the folder "samples". See Figure 9 for the code.
4.3 Aligning Data to Reference Genome and Converting to bigWig Format

The third option in the pipeline "Convert fastq files" initiates the code in figure 10 that requires parameters before it converts the file. It prompts the user for the adaptor sequence used for sequencing (and offers a quick alternative for the molgen group standard adaptor), it then asks the user to select if the samples contain a UMI or a barcode (and how long the barcode is), finally, it needs the organism, if any, that was used as for spike in. It then initiates the conversion as seen in figure 11.

The conversion from .fastq to .bigWig starts by making a list of all samples in the folder "samples", creating an output directory named "bigWig" and establishing paths for other folders needed. See figure 11.
The conversion is done as a loop for each sample in the sample list and part 1 removes the adaptor sequence, UMI and barcode. It then makes a reverse complement of the sequences and aligns the sequences to the bowtie index that was previously created (Figure 5). The alignment retains the nucleotide resolution and creates a .bam file with the aligned reads as well as a separate file with all unaligned reads that are then mapped to the spike in genome. See Figure 12 for the code.

Figure 10: The code for the user prompts upon initiating the file conversion.

Figure 11: Code for making a list if samples for conversion.
The .bam files are converted to .bed files using the tool `bamToBed` and then sorted by chromosome and by transcription start. The code also reports the number of counts both in the aligned .bed file and the spike in counts. See figure 13.

The sorted .bed files are then converted to two .bedgraph files, one for the positive strand and one for the negative strand, using the tool "genomeCoverageBed". Finally, the .bedgraph files are converted to .bigWig format using the "bedGraphToBigWig" tool. See figure 14.
4.4 Identification of Functional Genomics Regions in Fly, Mouse, Dog and Human

The fifth option in the pipeline "Map functional genomic regions" initiates the code in figure 15 that prompts the user to create a gene list (figures 16 and 17), and then to make sure that the dREG file from the online tool is available in the organism folder.

The script downloads a NCBI gene list, removes unnecessary columns and renames the chromosomes in the list (Figures 16 and 17). Finally, the gene list is moved to the "genome" map before the mapping starts.
The mapping of functional genomic regions starts by making a list of samples that are going to be mapped so the script can loop over them, (figure 18). It then creates a folder under the “analysis” folder and runs the first R script that reports coordinates of the different transcriptional elements. The R script loads the previously created gene list and then defines the different elements based on the TSS and CPS on the plus and minus strands (figure 19).
Figure 18: Making a list of samples and determining the transcriptional regions.

```r
# Make a list of your samples, without the .fastq extension

cd(~/.organised/samples)
declare $sample_list_file
for file in *.fastq; do
done
sample_list[1]=<sample_list[1]>

d /...
#

for sample in $sample_list[]; do
#

  echo "Creating functional genomics .bed-file for sample $sample"

  mkdir -p (~/.organised/analysis/functionalGenomics[/sample]
  Rscript R.R organise samp sample
  # create a folder for output files.
  # R script that adds coordinates for different features such as TSS, CPS, genebody etc.
```

Figure 19: R-script for determining the limits for the different transcriptional elements.
Bedtools is then used to intersect the dREG file with the "refGenes_TSSpm500.txt" file to find genes that are actively transcribing. The sites that were identified by dREG as actively transcribing and are in the list of genes are therefore actively transcribing genes. By then intersecting the "refGenes_TSSpm500.txt" file with the dREG file it finds all active elements which are not in the list of genes, such as enhancers, (Figure 20). The second R script then starts, which reads the active genes listed by the bedtools tool and creates separate .txt files for each transcriptional element. See figure 21.

Figure 20: Code for intersecting dREG-data with gene-list to find active genes.

The .bed files generated by the script in section 4.3 are split into two files, one with the positive strand and one with the negative. The active site of transcription is given in the third column of plus strand and the second column for minus strand. The active site of transcription (single nt) is placed both to the second and the third column before the files are concatenated again.

The bedtools tool is then used to intersect the .bed file with the different transcriptional elements files generated by the code in figure 21. This generates .bed files for each transcriptional element. See figure 22.

Figure 21: R-script to label the active genes based on transcriptional elements.
The script then reports counts of engaged Pol II for each transcriptional element, total counts, enhancers, divergent transcription, promoters, gene bodies, cleavage and polyA sites, short genes, termination window, and counts that are not localised to any region. See figure 23.
Figure 23: The code for reporting the different amounts of engaged polymerase II at different regions.

The final step is to concatenate the data into a .bed file with all different elements with a corresponding colour for each region. It then outputs a .bed file named “functionalGenomicRegions” that can be viewed in a genome browser, such as IGV. It also outputs a normalised version of the .bed file by dividing the raw count of engaged Pol II in the dataset at each functional region by total count of engaged Pol II in the dataset and multiplying by 1000000 (figure 24). This normalisation strategy uses sequencing depth as counts per millions of counts (CPM) to bring the distinct samples into a comparable scale.
By adding the coloured "functionalGenomicRegions.bed" file as well as the corresponding "allMap_unnorm.bigWig" files different genes can be studied. See figure 25, 26, 27, and 28 for the gene HSPA1A in Homo sapiens, Canis lupus familiaris, Mus musculus, and Drosophila melanogaster respectively.

In figures 25, 26, 27, and 28 the different transcriptional elements are coloured with divergent transcription in purple, promoter proximal in orange, gene body in black, cleavage and polyA site in blue, and termination window in red. The counts are labelled under the different elements and under the mapped regions, the polymerase profile is marked with red for the plus strand and with blue for the minus strand.
Figure 25: The HSPA1A gene mapped to the T2T-chm13v2 genome in K562 cells, the first sample is the control sample (HS0) and the second sample is the heat-shocked sample (HS30). The Pol II profile data range is set to 0-1000 for the plus strand and -100 - 0 for the minus strand.

Figure 26: The HSP70 gene mapped to the canFam6 genome in DH82 cells, the first sample is the control sample (HS0) and the second sample is the heat-shocked sample (HS30). The Pol II profile data range is set to 0 - 10 for the plus strand and -10 - 0 for the minus strand.

Figure 27: The HSPA1B gene mapped to the mm39 genome in MEF cells, the first sample is the control sample (HS0) and the second sample is the heat-shocked sample (HS25). The Pol II profile data range is set to 0 - 10 for the plus strand and -10 - 0 for the minus strand.
4.5 Quantification of Transcription Upon Acute Stress

The sixth option, "Compare gene expression", in the pipeline compares two .bed files, the "functionGenomicRegions" files, for the control and heat-shocked sample. It reads the names of the samples and initiates an R script (R3.R). See Figure 29.

```r
CompGeneExpression () {
    cd ${organism}/samples
    declare -a sample_list
    for file in *.fastq; do
        sample_list["${sample_list[@]}"="file"]
    done
    sample_1="${sample_list[0]}
    sample_2="${sample_list[1]}
    cd ..;/
    mkdir -p ${organism}/analysis
    Rscript R3.R $organism $sample_1 $sample_2 --save
}
```

Figure 29: The code to initialise the gene expression quantification

The R script reads in the two files and calculates the difference in expression between the control and the heat-shock sample for each gene. It then calculates the relative difference in expression by dividing the difference in expression by the counts in the control sample. Skips all rows where the count is zero and it outputs three .txt files, one with all genes, one with the top 50 upregulated genes, and one with the top 50 downregulated genes. See Figure 30 for the code.
Figure 30: R script for the gene expression analysis.

The top 5 genes that were up- and downregulated for *Homo sapiens*, *Canis lupus familiaris*, *Mus musculus*, and *Drosophila melanogaster* are presented in tables 2, 3, 4, and 5 with the name and function of the genes. In table 6 the number of differentially expressed genes, up- and dowregulated genes and total number of mapped genes are presented for each organism.
**Table 2:** Top 5 up- and downregulated genes and their function for *Homo sapiens*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSPA1A</td>
<td>heat-shock protein family A (hsp70)</td>
<td>ICAM1</td>
<td>Intercellular adhesion molecule 1. Encodes a cell surface glycoprotein typically expressed on endothelial cells and cells of the immune system.</td>
</tr>
<tr>
<td>HSPH1</td>
<td>heat-shock protein family H (hsp110)</td>
<td>EGR1</td>
<td>Early growth response 1. A nuclear protein and functions as a transcriptional regulator.</td>
</tr>
<tr>
<td>HSPB1</td>
<td>heat-shock protein family B (hsp27)</td>
<td>STARD4</td>
<td>STAR related lipid transfer domain containing 4. Regulates cholesterol homeostasis.</td>
</tr>
<tr>
<td>ZFAND2A</td>
<td>Zinc Finger AN1-type. Involved in proteasome-mediated ubiquitin-dependent protein catabolic process and protein targeting to ER.</td>
<td>JUNB</td>
<td>JunB proto-oncogene. Enables sequence-specific double-stranded DNA binding activity. Involved in positive regulation of transcription by RNA polymerase II.</td>
</tr>
<tr>
<td>HSPD1B</td>
<td>heat-shock protein family D (hsp60)</td>
<td>CD69</td>
<td>Encodes a member of the calcium dependent lectin superfamily of type II transmembrane receptors.</td>
</tr>
</tbody>
</table>

**Table 3:** Top 5 up- and down regulated genes and their function for *Canis lupus familiaris*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSP70</td>
<td>heat-shock protein 70</td>
<td>NISCH</td>
<td>Encodes a nonadrenergic imidazoline-1 receptor protein that localizes to the cytosol and anchors to the inner layer of the plasma membrane. The orthologous mouse protein has been shown to influence cytoskeletal organization and cell migration by binding to alpha-5-beta-1 integrin,</td>
</tr>
<tr>
<td>SMPD2</td>
<td>Protein coding gene for sphingomyelin phosphodiesterase (Unknown function)</td>
<td>CUED2</td>
<td>Predicted to enable ubiquitin binding activity. Acts upstream of or within negative regulation of cytokine production involved in inflammatory response and negative regulation of macrophage cytokine production</td>
</tr>
<tr>
<td>HSPH1</td>
<td>heat-shock protein family H (Hsp 110)</td>
<td>PIAS3</td>
<td>Encodes a member of the PIAS [protein inhibitor of activated STAT (signal transducer and activator of transcription)] family of transcriptional modulators. It directly binds to several transcription factors and either blocks or enhances their activity.</td>
</tr>
<tr>
<td>CHAC1</td>
<td>Encodes a member of the gamma-glutamylcyclotransferase family of proteins. May also play a role in the unfolded protein response, and in regulation of glutathione levels and oxidative balance in the cell.</td>
<td>POLR3C</td>
<td>Enables single-stranded DNA binding activity. Involved in positive regulation of innate immune response and positive regulation of interferon-beta production.</td>
</tr>
<tr>
<td>HPS1</td>
<td>Encodes a protein that may play a role in organelle biogenesis associated with melanosomes, platelet dense granules, and lysosomes.</td>
<td>TMEM140</td>
<td>Predicted to be integral component of membrane.</td>
</tr>
</tbody>
</table>
Table 4: Top 5 up- and downregulated genes and their function for *Mus musculus*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ier3</td>
<td>Immediate early response 3. Acts upstream of or within several processes, including intracellular signal transduction; negative regulation of mitochondrial outer membrane permeabilization involved in apoptotic signaling pathway.</td>
<td>Taf4</td>
<td>TATA-box binding protein associated factor 4. Enables DNA binding activity.</td>
</tr>
<tr>
<td>Cebp</td>
<td>CCAAT/enhancer binding protein. Enables several functions, including DNA-binding transcription activator activity, RNA polymerase II-specific and core promoter sequence-specific DNA binding activity; and enzyme binding activity. Contributes to RNA polymerase II cis regulatory region sequence-specific DNA binding activity.</td>
<td>Stap2</td>
<td>Signal transducing adaptor family member 2. Predicted to enable signaling adaptor activity.</td>
</tr>
<tr>
<td>Calr</td>
<td>Calreticulin. Enables calcium ion binding activity; carbocytoskeleton organization activity; and mRNA binding activity.</td>
<td>Gm42029</td>
<td>Predicted ncRNA</td>
</tr>
<tr>
<td>Gadd45ip1</td>
<td>Growth arrest and DNA-damage-inducible. Enables mitochondrial ribosome binding activity. Acts upstream of or within positive regulation of cell population proliferation and positive regulation of oxidative phosphorylation.</td>
<td>LOC115487419</td>
<td>Uncharacterized ncRNA</td>
</tr>
<tr>
<td>Tax1bp3</td>
<td>Tax1 binding protein. Enables beta-catenin binding activity. Involved in negative regulation of Wnt signaling pathway.</td>
<td>Tmem65</td>
<td>Transmembrane protein 65. Involved in cardiac ventricle development and regulation of cardiac conduction.</td>
</tr>
</tbody>
</table>

Table 5: Top 5 up- and down regulated genes and their function for *Drosophila melanogaster*.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dmel_CG3717</td>
<td>Involved in iron-sulfur cluster assembly. Human ortholog(s) of this gene implicated in combined oxidative phosphorylation deficiency 19.</td>
<td>Dmel_CG12292</td>
<td>Predicted to enable magnesium ion transmembrane transporter activity; Involved in negative regulation of BMP signaling pathway.</td>
</tr>
<tr>
<td>Dmel_CG7248</td>
<td>Predicted to enable chitin binding activity</td>
<td>Dmel_CG10990</td>
<td>Involved in germarium-derived female germ-line cyst formation and positive regulation of stem cell differentiation.</td>
</tr>
<tr>
<td>Dmel_CG32120</td>
<td>Enables DNA-binding transcription factor activity, RNA polymerase II-specific. Contributes to DNA-binding transcription activator activity. Involved in several processes, including negative regulation of transcription by RNA polymerase II.</td>
<td>Dmel_CG9210</td>
<td>Enables adenylate cyclase activity. Predicted to be involved in adenylate cyclase-activating G protein-coupled receptor signaling pathway.</td>
</tr>
<tr>
<td>Dmel_CG34417</td>
<td>Predicted to be involved in actin cytoskeleton organization.</td>
<td>Dmel_CG30325</td>
<td>No information</td>
</tr>
<tr>
<td>Dmel_CG9781</td>
<td>Predicted to enable chitin binding activity.</td>
<td>Dmel_CR32420</td>
<td>Predicted to enable GUA codon-amino acid adaptor activity. Predicted to be involved in translation.</td>
</tr>
</tbody>
</table>

Table 6: The number of mapped genes for the different organisms and the number of differentially expressed genes.

<table>
<thead>
<tr>
<th>Organism/genome</th>
<th>Number of upregulated genes</th>
<th>Number of downregulated genes</th>
<th>Number of differentially expressed genes</th>
<th>Number of mapped genes (control)</th>
<th>Number of mapped genes (heat-shock)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em> / T2T-chm13v2</td>
<td>10923</td>
<td>10610</td>
<td>13211</td>
<td>14991</td>
<td>15297</td>
</tr>
<tr>
<td><em>Homo sapiens</em> / hg38</td>
<td>10700</td>
<td>10338</td>
<td>12988</td>
<td>14932</td>
<td>15296</td>
</tr>
<tr>
<td><em>Canis lupus familiaris</em> / canFam6</td>
<td>10159</td>
<td>10487</td>
<td>12661</td>
<td>13201</td>
<td>14006</td>
</tr>
<tr>
<td><em>Mus musculus</em> / mm39</td>
<td>2875</td>
<td>3514</td>
<td>3559</td>
<td>13631</td>
<td>3914</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> / dm6</td>
<td>276</td>
<td>291</td>
<td>395</td>
<td>1404</td>
<td>817</td>
</tr>
</tbody>
</table>
4.6 Comparison of Functional Annotation in Human Genome Versions

By mapping the heat-shock data from *Homo sapiens* to both the most widely used genome version hg38 and the newer version T2T-chm13v2, a comparison between the genome versions can be performed. The difference between the genome versions is most evident in the chromosomes 14 (figure 31) and 22 (figure 32).

![Figure 31: Mapped functional genomic regions of chromosome 14 in K562 cells with genome version hg38 and T2T-chm13v2.](image)

![Figure 32: Mapped functional genomic regions of chromosome 22 in K562 cells with genome version hg38 and T2T-chm13v2.](image)

5 Discussion

The aim of building a computational pipeline has resulted in five different shell codes and three R scripts that offer a fast and simple way to analyse PRO-seq data. It builds a bowtie2 index for sequence alignment, offers three different ways to load sample data, example data from the GEO, loading samples from GEO, or loading local files, and it converts sample .fast files into .bed and .bigWig files. The pipeline uses the generated profile of nascent transcription to detect actively transcribed genes and enhancers to map functional genomic regions and perform a simple gene expression analysis.

5.1 Identification of Functional Genomic Regions in Fly, Mouse, Dog and Human

The .bigWig files created in the pipeline can be loaded into a genome browser, such as IGV, to study the Pol II profile as seen in Figs. 25, 26, 27, and 28. Pol II has a profile with a peak slightly upstream of the TSS and at the CPS. These higher Pol II counts are expected, as Pol II pausing before transcriptional elongation and pausing before termination will increase the counts at these positions. The divergent transcription shows a similar peak on the antisense strand as a result of bidirectional transcription. The divergent transcription is clearer in the *H. sapiens* sample, (figure 25), compared to *C. lupus familiaris*, *M. musculus*, and *D. melanogaster* but is still present in all organisms.

By mapping functional genomic regions and coloring transcriptional elements, tracks can be effectively visualised in IGV together with the Pol II profile. The count of engaged Pol II also gives an indication of how much activity the gene shows. The HSPA1A gene shows increased activity, an upregulation, in all organisms after heat shock, as expected. The mapping of transcriptional regions also gives an idea of the gene length in different organisms and shows that the gene body for HSPA1A in *H. sapiens*, *M. musculus*, and *C. lupus familiaris* is around 2 kb, while it is only 1 kb in *D. melanogaster*.

Overall, the use of TREs and dREG is more useful in mammals such as *H. sapiens*, where divergent transcription...
is more common than in insects such as D. melanogaster, but it is still useful in both. In addition to studying gene lengths, mapping functional genomic regions, and engaged Pol II, the use of the Pol II profile enables the identification of genes and enhancers that have not been previously annotated or studied since it does not require any previous knowledge about gene annotation to predict gene location.

5.2 Quantification of Transcription Upon Acute Stress
The count of engaged Pol II can be used to give a rough estimate of which genes are up- and downregulated. It is clear that the upregulated genes in H. sapiens are heat shock-related, as seen in table 2 while the downregulated genes in H. sapiens could be related to cellular growth, a function that is repressed. It is known that all transcription is briefly repressed upon heat stress, so the down-regulation should be more extensive than the upregulation and not as clearly linked to heat stress as the upregulated genes. The same pattern can be distinguished in C. lupus familiaris, M. musculus, and D. melanogaster, but not as clearly as in H. sapiens; see tables 3, 4, and 5. H. sapiens samples show a total of 13211 differentially expressed genes, C. lupus familiaris shows a total of 12661 differentially expressed genes, M. musculus shows a total of 3859 differentially expressed genes, and D. melanogaster shows a total of 395 differentially expressed genes (table 6).

5.3 Comparison of Functional Annotation in Human Genome Versions
The version of the reference genome used to align the sample to can make a great difference. The human genome version hg19 was published in 2009 and the follow-up, hg38, was published in late 2013, while T2T-chm13v2/hs1 was published in 2022. Google Scholar shows more than 1,700 publications in 2023 with the keyword hg19, 1,300 publications in with the keyword hg38 and only 10 publications containing the keyword T2T-chm13v2. This indicates that the vast majority of scientists still use the hg38 version of the reference genome instead of the updated version.

By mapping the H. sapiens data to both the new genome version and the previous, the differences are clear in the short arms of chromosome 14 and chromosome 22, see figures 31 and 32. Using the new version of the reference genome also finds over 200 more differentially expressed genes compared to hg38 (13211 vs 12988). It is evident that the new reference genome should be used whenever possible.

5.4 Computational Pipeline
The computational pipeline offers a simple method of analysing large-scale genomic data from PRO-seq by combining and unifying different parts of the analysis in one pipeline. By running the script in a ssh connected virtual machine and using a terminal multiplexer such as _screen_ it can run for hours without relying on a internet connection or power supply to the machine. As the pipeline is not currently able to run the dREG analysis, it is unable to align data and run the analysis without user input. Making a bowtie2 index, downloading experimental data, aligning sample sequences, and creating .bed and .bigWig files takes around 8 hours with the current setup. The need to run it without the computer stopping is therefore a key.

5.5 Limitations and Future Studies
The complexity of a data analysis tools grows exponentially as the differences in the input data varies. Therefore, the pipeline has been streamlined to suit eukaryotic samples that have been prepared in the Vihervaara laboratory at SciLifeLab in Stockholm. There has also not been any difference in the analysis steps, such as determination of gene length, etc., that take the type organism into account.

The pipeline should make it easy to run many different samples without much preamble, and a larger comparison between organisms could be performed. There is also a need to keep a pipeline such as this up to date, and there are always improvements that could be made, and it would therefore require continued maintenance.

Possible next steps could be to integrate dREG into the pipeline and rebuild the scripts into an online tool to make it more useful and available to the scientific community.
Acknowledgments

I would like to extend my gratitude to my supervisor Dr. Anniina Vihervaara for her continual support, inspiration, and always taking the time to discuss and help me with my thesis. Additionally, I would like to thank my co-supervisor Dr. Samu Himanen for all your help and for letting me bug you while you try to work. I would also like to extend my thanks to Adelina Rabenius and Anam Minhas for discussing and asking the right answers to make my work better. Klara, thank you for your continuous support and for always listening.

References


