Universitat Oberta de Catalunya uoc.edu

Implementation of a workflow for processing and analyzing ChIP-seq data

David Arambilet I Morilla
Máster en Bioinformática y Bioestadística
Análisis de datos ómicos

Dr. Jose Luis Mosquera Mayo
Dr. Carles Ventura Royo

02/06/2022

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Implementation of a workflow for the analysis of ChIP-seq data*
Nombre del autor:	*David Arambilet I Morilla*
Nombre del consultor/a:	*Dr. Jose Luis Mosquera Mayo*
Nombre del PRA:	Dr. Carles Ventura Royo
Fecha de entrega (mm/aaaa):	06/2022
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	Análisis de datos ómicos
Idioma del trabajo:	Inglés
Número de créditos:	15
Palabras clave	*ChIP-seq, workflow*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Los avances en el campo de la secuenciación son cada vez mayores, al igual que la necesidad de analizar los datos generados de forma que se puedan extraer todos los resultados posibles. En la actualidad, la técnica de ChIP-seq es cada vez más común, pero las herramientas para analizar los datos generados son muy diversas, así como pueden llegar a ser los resultados obtenidos con cada una de ellas. El objetivo principal de este proyecto es implementar y aplicar un workflow para el procesamiento y el análisis de datos crudos de ChIP-seq. Para ello, primero se hace una revisión de los diferentes pasos que se llevan a cabo para el procesamiento de datos de ChIP-seq. Una vez identificados los pasos que deberán implementarse en el workflow, se hace una revisión de las diferentes herramientas actuales que se usan en este campo y se escogen las más adecuadas para su implementación. El producto final de este proyecto es un workflow que pueda procesar y analizar los datos de ChIP-seq, así como los diferentes análisis que pueden llevarse a cabo una vez procesados los datos, demostrando que el análisis adecuado de estos datos puede dar respuesta a muchas preguntas biológicas aún sin responder.

Abstract (in English, 250 words or less):

Advances in the sequencing techniques are greater every year, as well as the need of analyzing the data generated so all the possible information that it contains can be extracted. Nowadays, ChIP-seq experiments are far more common than in the past years, but the tools for analyzing the generated data are still very diverse. The main objective of this project is to implement and apply a workflow for the processing and analysis of raw ChIP-seq data. In order to do so, firstly a review of the steps that are conducted for the processing of ChIP-seq data is performed. Once the steps that the workflow will need to incorporate have been identified, a review about the different currently used tools used in this analysis is performed and the most suitable tools are chosen and implemented in the workflow. The final product of this project is a workflow that is able to process and analyze ChIP-seq data, and, additionally, the different analysis that can be undertaken with the processed data files, showing that the correct analysis of this data can answer many key biological questions.

Contents

Figure list

Table list

1. Abstract

Advances in the sequencing techniques are greater every year, as well as the need of analyzing the data generated so all the possible information that it contains can be extracted. Nowadays, ChIP-seq experiments are far more common than in the past years, but the tools for analyzing the generated data are still very diverse. The main objective of this project is to design and apply a workflow for the processing and analysis of raw ChIP-seq data. In order to do so, firstly a review of the steps that are conducted for the processing of ChIP-seq data is performed. Once the steps that the workflow will need to incorporate have been identified, a review about the different currently used tools used in this analysis is performed and the most suitable tools are chosen and implemented in the workflow. The final product of this project is a workflow that is able to process and analyze ChIP-seq data, and, additionally, the different analysis that can be undertaken with the processed data files, showing that the correct analysis of this data can answer many key biological questions.

2. Introduction

2.1. Context and justification

The expression of the genes coded in the DNA are widely known to be regulated by many different factors that form a complex regulatory network. This regulation is a key subject that needs to be deeply studied for the correct comprehension of the regulatory mechanisms that different factors display and understand their biological implications.

On one hand, there are many proteins that are able to bind to the genome and drive the expression of its target genes. These proteins are known as transcription factors. They are commonly studied in different contexts for their importance in normal and malignant processes (Lambert et al., 2018). There are many types of transcription factors and how they regulate the expression of their target genes is still, in most of the cases, an open matter of study. Some of them are shown to directly bind the promoter regions of their target genes, which are usually placed at the 5' end of the gene. However, there are studies that also show that transcription factors can also bind to regions placed far from the promoter region and equally drive the expression of their target genes (Spitz et al., 2012). These regions are known as enhancer regions.

Although the study of proteins binding to the DNA is commonly associated with transcription factors, there are also their opposites, the repressors, which are known to bind also to promoter regions of their target genes and inhibit the expression of their target genes (Reynolds et al., 2013). This regulatory network involving both transcription factors and repressors is a matter of recent studies and more are yet to come.

On the other hand, the epigenetic mechanisms are also essential in order to control the expression of the genes. Epigenetics is the study of changes in the expression of the genes that do not involve modifications in their sequence but can be inherited from parents (Weinhold, 2006; Allis et al., 2016). One clear example of these changes in the expression are modifications in the histones (Zhang et al., 2021). Histones, the proteins that wrap the DNA into nucleosomes, are known to be modified by, normally, methylation or acetylation in their residues. Depending on the modification and on the residue that is modified, the epigenetic changes can positively drive the expression of

the genes or either recruit repressors that will prevent their transcription (Lawrence et al., 2016).

The study of both proteins that bind to the DNA and histone modifications, is imperative in order to decipher the regulatory networks driving the expression of some key genes that can be involved, not only in normal conditions, but also in malignancies such as in cancer (Dawson et al., 2012). One technique that can study both aspects of this regulatory network is Chromatin Immunoprecipitation coupled to high throughput sequencing techniques (ChIP-seq).

The ChIP technique is an experimental method that is used to identify the regions that specific proteins are binding in the chromatin (Das et al., 2004). It basically consists in sequential steps (Figure 1) where the proteins that are bound to the DNA are crosslinked using chemical reagents such as formaldehyde and then the chromatin is fragmented in small pieces. The fragmentation has to be small enough to detect specific sites where the proteins are binding but long enough to be able to sequence these fragments. Once the chromatin is fragmented, the proteins, together with the crosslinked bound DNA regions, are precipitated and isolated using antibodies. Once the desired protein bound to the DNA is isolated, the DNA is then separated from the proteins and purified. The final product of this process is specifically the DNA that was bound to the target protein, being able then to identify the specific regions that the protein is binding in the chromatin. This product DNA can then be analyzed by other experimental techniques such as qPCR analysis, where specific regions are analyzed to see if they are present or not in the ChIP product. However, in order to find *de novo* binding regions for a specific protein or for specific histone marks, the DNA product obtained from the ChIP can then be sequenced using Next Generation Sequencing (NGS) techniques.

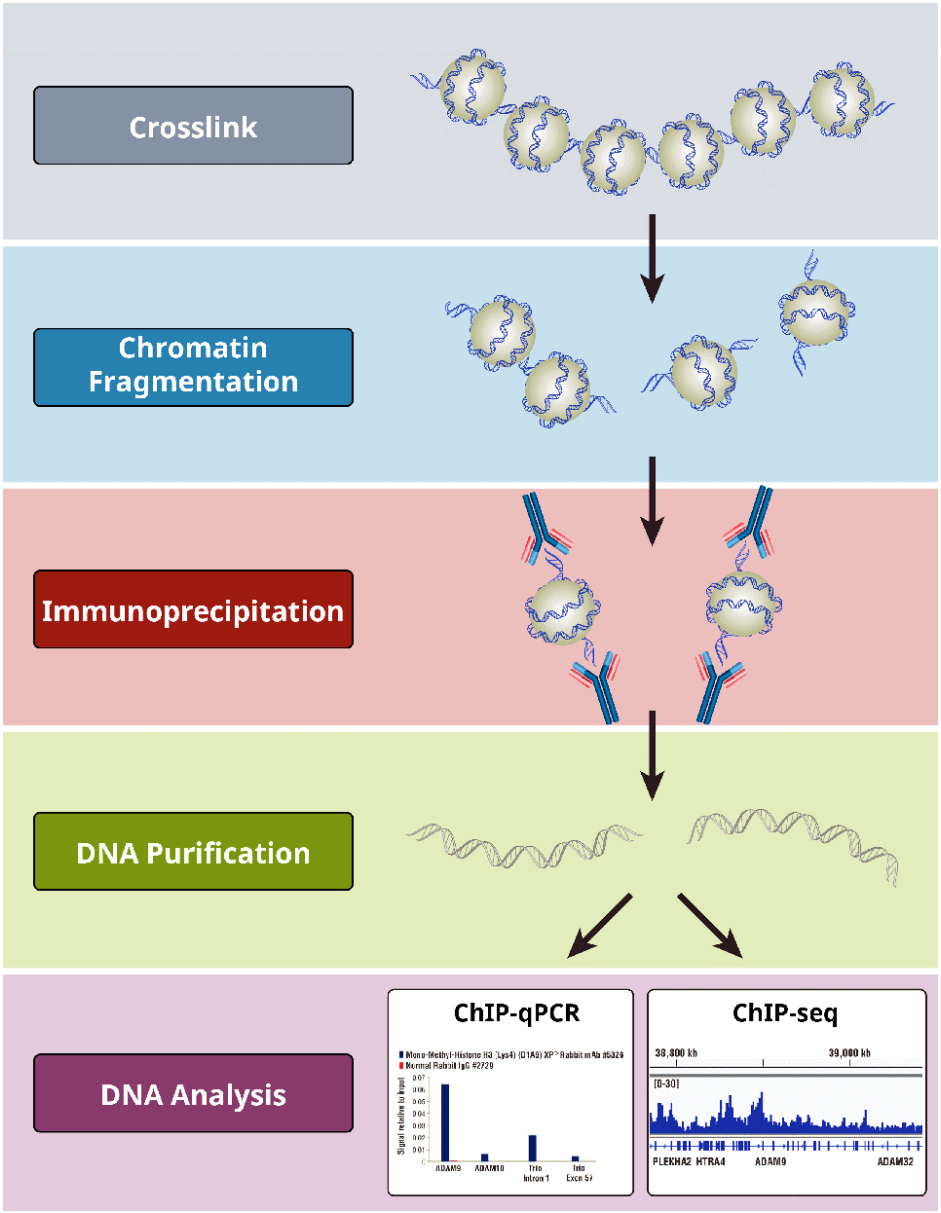

Figure 1. Chromatin Immunoprecipitation experimental procedure. Schematic representation of the ChIP experiment steps. Adapted from https://www.cellsignal.com/.

For applying NGS techniques in the DNA product of the ChIP technique, first the DNA needs to be amplified. For this, the DNA is prepared in libraries, where DNA fragments are tagged with adapters, which are short nucleotide sequences that are used to amplify the material and are also used as primer binding sites for sequencing initiation.

The output of the sequencing process are the raw sequences that have been amplified in the sequencer. Once the sequences are ready, they are processed and analyzed in order to check for specific genes enriched, motifs or common functions among other analyses (Figure 2).

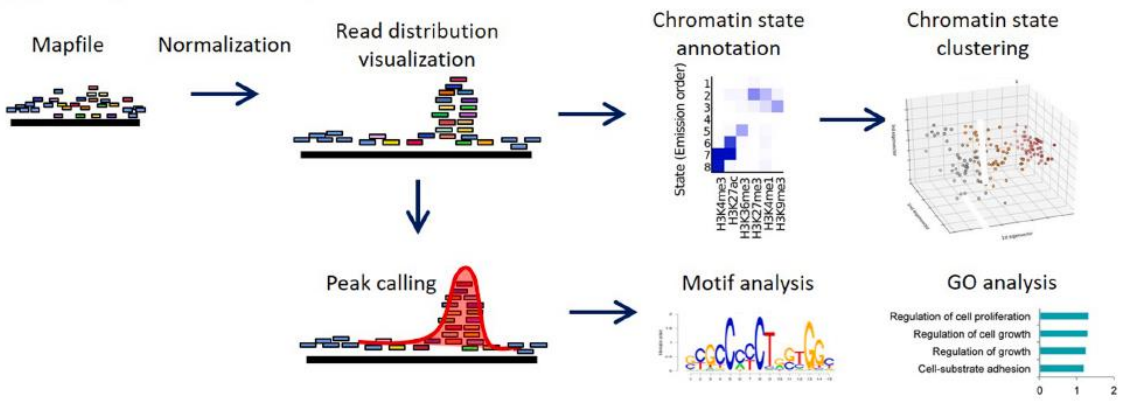

Figure 2. Computational ChIP-seq analysis. Schematic representation of the different analysis that can be performed from a ChIP-seq experiment. Adapted from Nakato et al., 2021

In recent years, ChIP-seq techniques have improved enormously (Park et al., 2009; Furey et al., 2012; Nakato et al., 2017) and every day there is more data generated. This information is key for understanding many different processes so the need for the deep analysis of this type of data is urgent. However, while other sequencing related techniques are more common and there is more consensus on how to process and analyze this type of data, such as in the case of RNA-seq, ChIP-seq data analysis is still more variable and the number of different tools for the processing is large.

In this work, a revision of the different available tools for the processing and analysis of raw ChIP-seq data is performed. Also, a workflow with the most common and better tools has been generated in order to have a ready-to-use workflow that can process this type of data.

2.2. Objectives

This work has two main objectives, each one with a subset of specific objectives.

The first main objective is to design a workflow for the processing of raw ChIP-seq data. The specific objectives associated with it are:

1.1. Review the state of the art of the ChIP-seq data analysis

1.2. Define the steps of the workflow and select the tools

1.3. Implement the workflow

The second main objective is to test the implemented workflow with downloaded ChIP-seq data. The specific objectives associated with it are:

2.1 Identify the appropriate dataset

2.2 Test the workflow

2.3 Compare the tool's performance

2.3. Focus and methodology

In order to accomplish the main objective of this work, which would be to design and implement a workflow for the analysis of ChIP-seq data, an extensive bibliographic revision of the already published pipelines and workflows for this type of analysis. In this line, the different steps of the workflow and, for each one, the different tools available have been reviewed.

After the bibliographic research, the workflow has been implemented, so the analysis of the raw ChIP-seq data can be performed. The different tools were installed in a cluster and the scripts for each part of the analysis have been designed.

Finally, once the workflow has been completely designed, different datasets already published in online repositories have been downloaded in the cluster and the workflow has been tested with different datasets. Following this strategy, firstly the information is gathered and then the workflow is designed and properly tested, being the final outcome a fully functional workflow for the processing and analysis of ChIP-seq data.

2.4. Planification

For the accomplishment of the main objectives and their corresponding specific objectives, different tasks have been scheduled, with several milestones to complete before going towards the next step (Figure 3).

The first task to be accomplished in the project planning is the bibliographical research of the different steps of the ChIP-seq analysis, as well as the different options available for each step. The duration of this part was expected to be of about 3-4 weeks of work, and one milestone was assigned to the accomplishment of this task.

The second task to complete was the design of the workflow, which was divided in two different steps. In the first one, the tools for the analysis were installed in the cluster where the analysis would take place. In parallel, the scripts for implementing each tool were designed. Finally, an extra task coupled to this one was to test the workflow. For this task, ChIP-seq data was downloaded from online repositories and the raw data was processed with the workflow. The expected time to accomplish this task was around 4 to 5 weeks and one milestone was assigned to the accomplishment of this task.

The last task of this project was to compare the processed data with the designed workflow to the processed data published. The duration of this task was around 1 to 2 weeks and one final milestone was assigned to the accomplishment of this task.

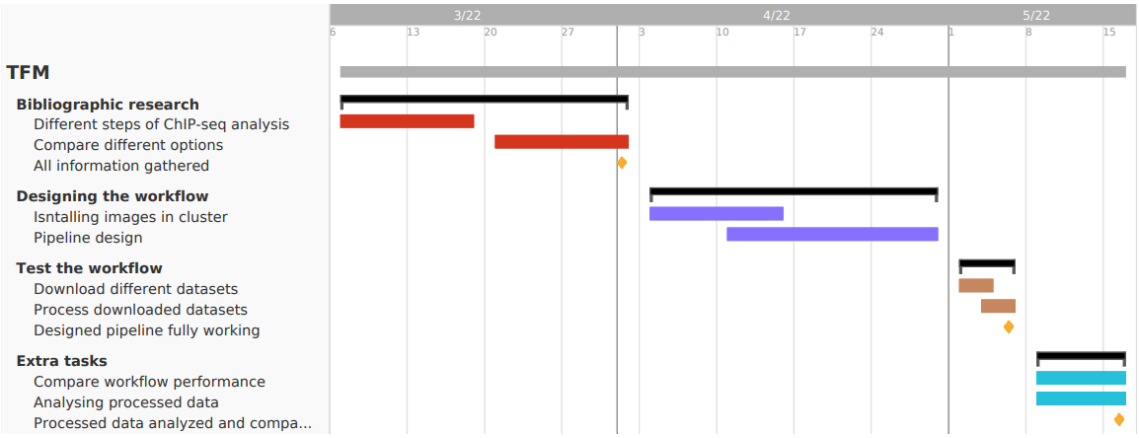

Figure 3. Planification of the project. Gantt diagram displaying the different tasks and their respective milestones to be accomplished in this project. Developed with TeamGantt.

2.5. Contributions and products obtained

The final outcome of this project is, on one hand, a revision of the different ChIP-seq pipelines and workflows already published, as well as of the different steps that a pipeline for processing ChIP-seq data needs. On the other hand, the outcome of this project is also implementing fully operational workflow for the processing of raw ChIP-seq data, as well as a revision of different analysis to be performed with the processed data.

2.6. Outline

Chapter 2 introduces a brief contextualization which covers the bibliographic research about what ChIPs are, their main goals and other aspects that are important for understanding the need of analyzing this type of data. The rest of the chapter enumerates the main objectives of this work together with their subsequent specific objectives, the planification and other important characteristics of the work.

Chapter 3 presents the state of the art of this field. In this chapter the work reviews different published pipelines for the processing of ChIP-seq data, as well as the main steps needed for the analysis.

Chapter 4 lists the different tools used for each step in this analysis and a brief explanation of their function.

Chapter 5 presents the results of this project, where all the steps are explained, together with deeper information for each step and the code, the input and the output files. Also, some of the further analysis that can be undergone after processing the raw ChIP-seq data are also detailed in this chapter.

Chapter 6 enumerates the conclusions, summarizing the main achievements of this work and how this work answers a current problem in science and some comments about the progress of the project.

3. State of the art

In recent years, the amount of sequencing data has increased exponentially, as well as the availability of this data. In the same direction, the number of different tools, pipelines and workflows for analyzing sequencing data has also increased (Ather et al., 2018; Bardet et al., 2011; Kharchenko et al., 2008; Nakato et al., 2021; Park et al., 2017; Wu et al., 2015). For this main reason, it is important to know which are the available tools for analyzing and interpreting this type of data.

The different available tools for each step of the ChIP-seq analysis are listed in the following Tables. All the tools listed are similar among them. However, some of them, such as in the peak calling step, differ in their statistical methods. In some cases, these changes can make the results more variable among different tools and even using the same datasets as input.

In Table1, the main features of the FASTQC tools are listed, as it is the main tool for the quality control step of the analysis.

Step	Tool	Interface	Input	Output	Description	Reference
Quality Control	FASTQC	Comand-line based	FastQ	html	Uses raw data or trimmed data to generate a complete report	Andrews, 2010

Table 1. Tools for the quality control of the raw data. Characteristics of the main tool used for the quality control of the raw sequencing data.

Table 2 provides a list of the different tools available for the second step of the processing, which is the trimming of the raw data. Several tools are available for this step, all of them being widely used in different workflows but they are all similar among them.

Step	Tool	Interface	Input	Output	Description	Reference
Trimming	Trim Galore!	Comand-line based	FastQ	Trimmed FastQ	Adapter trimming of the raw data. Wrapper around Cutadapt	Krueger, 2012
	Trimmomatic	Comand-line based	FastQ	Trimmed FastQ	Adapter trimming of the raw data	Bolger et al., 2014
	Cutadapt	Comand-line based	FastQ	Trimmed FastQ	Adapter trimming of the raw data	Martin, 2011
	Fastx toolkit	Comand-line based	FastQ	Trimmed FastQ	Short read pre-processing tool. Trimmer on fixed length adapters	Hannon, 2010

Table 2. Tools for the trimming of the raw data. List and characteristics of different tools for the Trimming step of the processing.

Table 3 lists the different tools available for the next step of the processing, which is the alignment of the trimmed sequences to the reference genome.

Step	Tool	Interface	Input	Output	Description	Reference
Alignment	HISAT	Comand-line based	FastQ (Trimmed data)	SAM	Uses raw sequencing files for streaming alignment, reducing disk space	Kim et al., 2015

	Bowtie2	Comand-line based	FastQ (Trimmed data)	SAM	Aligns trimmed data to reference genome	Langmead et al., 2012
	BWA-MEM	Comand-line based	FastQ (Trimmed data)	SAM	Alignment of the data to the reference genome. Needs specification on read length	Li et al., 2009

Table 3. Tools for the alignment of the trimmed data. List and characteristics of different tools for the Alignment step of the processing.

Table 4 shows the main features of the most used tool for the processing of the alignment files.

Step	Tool	Interface	Input	Output	Description	Reference
Processing alignment output	Samtools	Comand-line based	SAM	BAM	Transforms the SAM files into BAM files for their easier manipulation	Li et al., 2009

Table 4. Tools for the processing of the alignment output. Characteristics of the main tool used for the processing of the alignment output files (SAM files).

Table 5 provides a list of different tools for the peak calling step. The main difference among them is the statistical parameters that they use for consider a region a significantly enriched peak.

Step	Tool	Interface	Input	Output	Description	Reference
Peak Calling	MACS2	Comand-line based	BAM	NarrowPeak (BED format with additional information)	Uses alignment output and control file to perform the peak calling step	Zhang et al., 2008
	PeakSeq (NEXT-peak)	Comand-line based	BAM	NextPeak (similar to BED format but with additional information)	Uses alignment output and control file to perform the peak calling step	Rozowsky et al., 2009
	QuEST	Comand-line based	BAM	BED and BigWig	Uses alignment output and control file to perform the peak calling step	Valouev et al., 2008

Table 5. Tools for the peak calling step. List and characteristics of different tools for the peak calling step step of the processing.

Table 6 lists different options for the quality control of the alignment step.

Step	Tool	Interface	Input	Output	Description	Reference
Peak Quality Control	MultiQC	Command-line based	BAM	Report	Uses alignment output to identify outliers and asses quality of the sequencing data	Ewels et al., 2016

	ChIPQC	R package	BAM and narrowPeak/BED	Report	Uses peak calling and aligned files to obtain a complete report about quality of the ChIP-seq	Carroll et al., 2014

Table 6. Tools for the quality control of the alignment and peak calling steps. List and characteristics of different tools for the quality control of the alignment.

Table 7 provides different options for the visualization of processed ChIP-seq data. Both options are widely used and highly efficient. While the UCSC browser have many additional applications, IGV is solely focused on the visualization of this type of data.

Step	Tool	Interface	Input	Description	Reference
Data visualization	IGV	Desktop tool	BigWig	Desktop application for ChIP-seq data visualization	Robinson et al., 2011
	UCSC	Website appication	BigWig	Web application for ChIP-seq data visualization and manipulation	Kent et al., 2002

Table 7. List of different tools available for visualizing ChIP-seq data. List and characteristics of different tools for their visualization.

Given this variability among the different options for each step of the process, the designed workflow aims to use the most reproducible tools for each step that have been widely described to be highly efficient for the processing and analysis of raw ChIP-seq data.

4. Methods

In this work, several tools for the analysis of sequencing data have been used for each step. This chapter presents the tools that have been used for each step of the processing of ChIP-seq data. The images for the different tools were downloaded with DOCKER (https://www.docker.com/) and in a singularity-based cluster.

Data download

To test the implementation of the workflow, two datasets form the ENCODE database (Davis et al., 2018) have been considered: ENCSR000DLR, ENCSR000EGJ (Lou et al., 2020). The files were downloaded on the 05/05/2022 from https://www.encodeproject.org/.

Step 1: Quality control of the raw data

FASTQC tool (version 0.11.5) (Andrews, 2010) has been used to conduct the quality assessment of the raw data. This tool is widely used in this step of the processing thanks to its complete html output format, where the all results of the quality check are displayed in an easy-to-interpretate way.

Step 2: Trimming of the raw data

Trim Galore! tool (version 0.6.6) (Krueger, 2012) has been used to the trimming of the raw data. Trim Galore! is a useful and easy-to-use tool that allows to perform quality and adapter trimming of the raw FastQ files.

Step 3: Quality control of the trimmed data

The FASTQC tool has been used also for the quality assessment of the trimmed data.

Step 4: Alignment of the trimmed data

Bowtie2 tool (version 2.3.4.1) (Langmead et al, 2012) has been used to align the sequences to the reference genome. Bowtie2 is one of the most common tools to perform the alignment of the trimmed data. It works specially well with short reads (around 50 to 1000 bp, which is the rang that commonly comprises the ChIP-seq reads) and with large genomes such as in the case of mammals (e.g. Homo sapiens or Mus Musculus, which are the most commonly studied systems).

Step 5: Processing of the aligned files

SAMtools tool (version 1.15) (Li et al., 2009) has been used to obtain the results of the alignment in an easy-to-handle format. SAMtools is a useful tool that allows the easy manipulation of the output of the alignment, the SAM files. It has several utilities that allows the most commonly observed manipulations of this SAM files in order to process and obtain the BAM files, which are much easier to handle than SAM files.

Step 6: Visualization of the results

bamCoverage tool from DeepTools (version 3.5.0) (Ramirez et al., 2016) has been used in order to process the BAM files into a format that allows the visualization of the ChIP-seq data. Many different tools are used in order to obtain the BigWig files, but in this case the bamCoverage tool that it is incorporated in DeepTools, which its incorporated tools are used in different steps of the analysis, is a very useful one to obtain the BigWig files from the BAM files generated in the previous step.

Integrative Genomics Viewer (IGV) tool (version 2.5.3) (Robinson et al., 2011) has been used to visualize the processed files. The IGV software is an easy-to-use desktop application that allows the visualization of genomic data. It also has different options to study specific regions of the genome that are visually enriched in the analysis.

Step 7: Peak calling

MACS2 tool (version 2.2.7.1) (Zhang et al., 2008) has been used for identifying the significantly enriched genomic regions in the ChIP-seq. The MACS2 tool is one of the most common tools for this step of the analysis. It has been specifically designed for detecting the transcription factor binding sites in the genome.

Step 8: Quality control of the peak calling

Bioconductor ChIPQC R package (version 1.26.0) (Carroll et al., 2014) has been used to assess the quality of the peak calling results. The ChIPQC package takes as input the BAM files and the output files from the peak calling in order to compute different quality checks and generate a quality report. As it is an R package there is no need to run it in the cluster, so it is much more accessible and focused on the data that is being analyzed with this workflow.

Step 9: Annotation of the peaks

Bioconductor ChIPseeker R package (version 1.26.2) (Yu et al., 2015) has been used to annotate the results obtained in the peak calling step. The ChIPseeker package is also an R package that takes as input the peak calling output and the reference genome that it is being used and with a single step it annotates the regions enriched in the analysis. It also generates genomic distribution maps useful for this analysis.

Step 10: Downstream analyses with processed files

Venn diagrams

VennDiagram R package (version 1.7.3) (Chen et al., 2011) has been used for generating venn diagrams used to compare different datasets. The VennDiagram R package takes as input the list of genes obtained from the annotation step and, given different gene lists, performs venn diagrams to easily compare the different datasets.

Gene ontology analysis

Enrichr tool (Online version: https://maayanlab.cloud/Enrichr/) (Chen et al., 2013) is used to identify functions associated with the genes obtained in the annotation step.

Motif enrichment analysis

MEME-ChIP tool (MEME Suit version 5.4.1) (Machanick et al., 2011) has been used to identify motifs significantly enriched in the regions obtained from step 7. MEME Suit is the most used software in order to study DNA motifs. Its dependency, MEME-ChIP, it is specifically focused on the enrichment of different motifs in ChIP-seq data analysis. It uses as input the FASTA files generated from the BED files.

Heatmaps

ComputeMatrix and plotHeatmap tools from DeepTools (version 3.5.0) (Ramirez et al., 2016) have been used to generate heatmaps that can compare multiple genomic regions enriched in different ChIP-seq datasets. ComputeMatrix and plotHeatmap are both dependencies from DeepTools that generate heatmaps providing a general overview of the enrichment in the different binding regions, using as input the BigWig and the BED files obtained from the peak calling.

5. Results

Based on the review conducted, the workflow implemented consists in 10 main steps. Figure 4 describes the workflow implemented. The final output of the workflow are the BED and BigWig files that will then be used for further analysis and the information that can be taken from this analysis will be depicted.

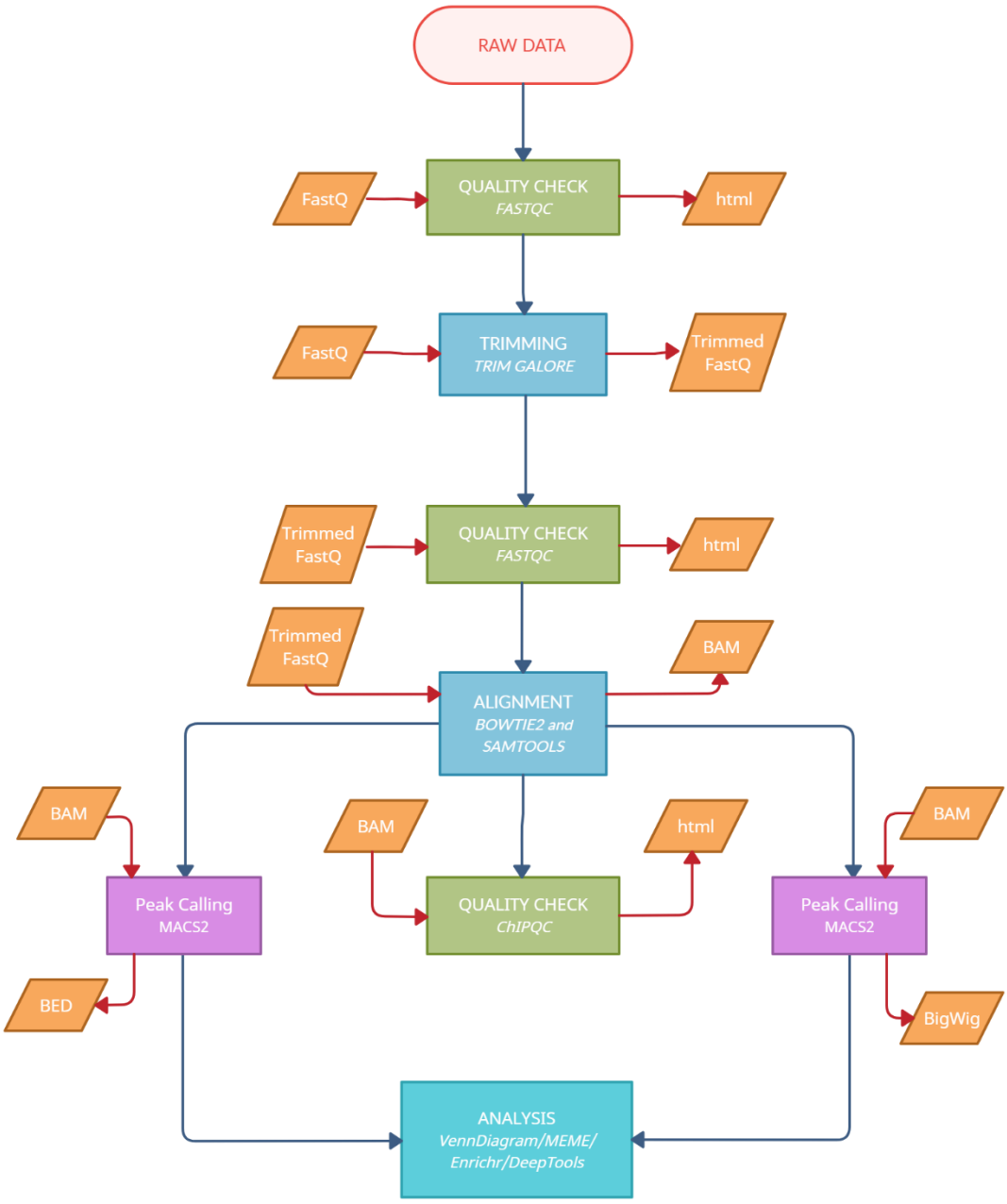

Figure 4. Workflow for the processing of ChIP-seq data. Different steps that the workflow undertakes. The tool used for each step is indicated. Steps of processing are marked in blue. Quality checks are marked in green. Output steps are marked in purple. Input and output files are indicated in orange.

5.1. Data download

Nowadays, ChIP-seq data is available to download from many public repositories (Table 8). The data can be easily downloaded in many different formats, from raw data to already processed data that is normally used in the published study related to the data obtention. For testing the workflow, data has been acquired from the ENCODE database.

Database	Website	Reference
ENCODE project	https://www.encodeproject.org/	Davis et al., 2018
Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/	Edgar et al., 2002
ChIP Atlas	https://chip-atlas.org/	Oki et al., 2018
IHEC data portal	https://epigenomesportal.ca/ihec/	Bujold et al., 2016
Cistrome	http://cistrome.org/	Liu et al., 2011
ROADMAP Epigenomics	http://www.roadmapepigenomics.org/	Roadmap et al., 2015
Mass Genome Annotation Data Repository (MGA)	https://ccg.epfl.ch/mga/	Dréos et al., 2018

Table 8. ChIP-seq data repositories. Available sources for downloading ChIP-seq datasets

The ENCODE database is a large repository of sequencing data from most histone marks, in order to study epigenetics, and from many transcription factors, for the study of the chromatin binding of different proteins. All the experiments available at this database have associated the raw data (FastQ files) and the processed files (normally BigWig and BED files for ChIP-seq experiments). The variety of datasets available and the availability of different file formats makes this repository very useful for accessing sequencing data.

5.2. Step 1: Quality control of the raw data

The sequencing of the ChIP-seq products returns as output tens of millions of reads. Before the analysis of these reads, basic quality checks must be performed in order to see if the raw data do not present problems that can bias our results. The tool implemented in the workflow for this step is FASTQC, which provides a quality check report with several quality metrics that can detect problems generated in the sequencing process or in the library preparation.

In order to execute the quality control, only the raw FastQ data files are needed and the quality check is performed with a single command line.

```
> singularity exec -B ${MAIN}:${MAIN}
${IMAGES}/fastqcmachalen.img fastqc
${RAWDATA}/myc_cml.fastq -o ${MAIN}/fastqc_files
```

The output of the quality check is a html report that gives a revision of the different quality metrics assessed by the tool, taken as an example one of the downloaded datasets to test the workflow (Figure 5).

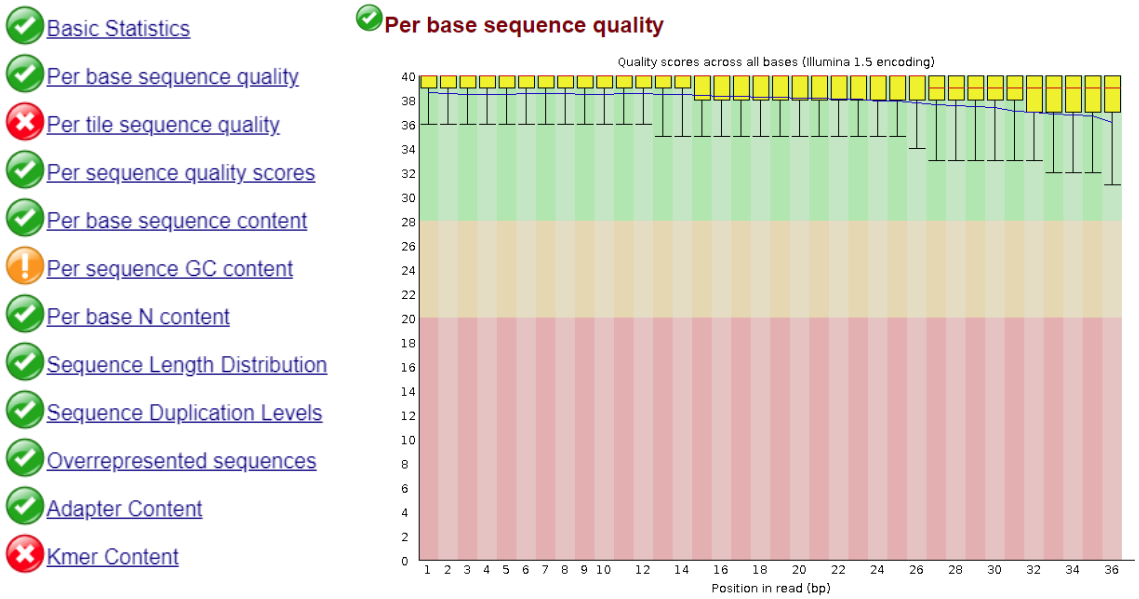

Figure 5. Quality check report by FastQC. List of the quality metrics reported by the FASTQC tool and an example of the per base sequence quality.

Some of the quality metrics analyzed in the report generated by the FastQC tool are depicted in Figure 5, having all the contents listed on the left column with a tick if it has passed the quality check, an exclamation mark for caution and a cross if it has failed the quality check (Figure 5A).

Firstly, the report starts with the Basic Statistics module (Figure 5B), which highlights the original file name, the file type, the encoding (which depends on the sequencer used), the number of reads that have been sequenced, the length of the sequences and the overall GC content.

The next quality metrics are the ones that analyze the reads that have been sequenced. It starts with the Per Base sequence quality (Figure 5C), which displays an overview of the range of quality values of the bases that form the sequences according to their position in it. The quality should be maintained in the green area of the graph, but it is common to see that the last positions of the reads often display a worse quality. Other quality metrics such as the Per Sequence Quality Scores give an overview of the general quality of the sequences.

Next group of quality metrics assess the nucleotides that form sequences. The Per Base Sequence Content plots the proportion of each nucleotide for each position. The expected would be an even distribution among the different nucleotides. However, if there is a bias towards some nucleotides it might indicate that some sequences are overrepresented. In the same direction, the report also displays the Per Sequence GC content (Figure 5D). It should also be distributed normally, but in some cases (such as in the example) it can present small deviations that rise a warning in the report.

Once the quality report has been carefully analyzed and the warnings and failures are properly considered, the processing of the data can continue.

5.3. Step 2: Trimming of the raw data

When DNA ChIP-seq product is prepared for sequencing, a short synthetic sequence of DNA known as adapter sequences are incorporated in the DNA. These adapters are usually regions that have information about the samples (barcode), regions that are used in order to amplify the material and regions that are used during the sequencing

procedure. In order to align the samples to the reference genome, these short synthetic sequences need to be removed from the sequencing results.

In this workflow, the TrimGalore! tool has been chosen, as in a single step is able to recognize the adapters and remove them. The input for this tool is basically the raw FastQ files.

```
> singularity exec -B ${ROOT}:${ROOT}
${IMAGES}/trimgalore.simg trim_galore -q 30 -o
${ROOT}/pipeline/Trim_data myc_cml.fastq
```

The output of this step of the workflow are the trimmed files, which are the reads obtained in the sequencing process without the adapters.

5.4. Step 3: Quality control of the trimmed data

After trimming the raw data, a new quality control needs to be performed. This quality control uses again the FASTQ tool as in Step 1. The process is the same but using as input the trimmed FastQ files. However, in some cases the quality metrics can improve greatly after trimming the data, as the adapters may interfere with the real quality of the sequences obtained.

5.5. Step 4: Alignment of the trimmed data

Once the adapters have been removed in the trimming process, the samples can be aligned to the reference genome, which is the key part of the sequencing analysis. This step consists in mapping the regions that have been sequenced into the reference genome (Figure 6).

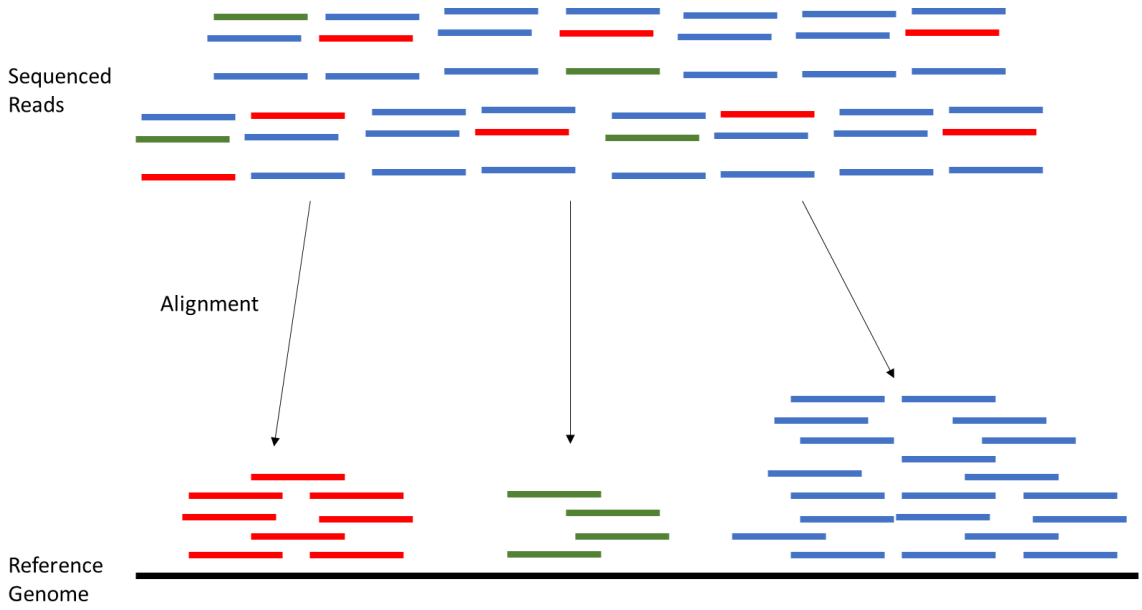

Figure 6. Alignment step. Schematic representation of the mapping of sequence reads to the reference genome

Depending on the species that the samples have been taken from, the reference genome needs to be in the same one. There are many versions of the genome available, as there are still some regions being described up to date and this will change the coordinates of our sequencing results. When testing this workflow, as the downloaded data was from Homo Sapiens, the sequences were aligned to the newest version of the Homo Sapiens reference genome, the GRCh38.

In this workflow, Bowtie2 has been used for the alignment step. It uses as input data the trimmed reads obtained from the previous step and the reference genome.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/bowtie2machalen.img bowtie2 -x
${REFGENOME}/Human_GRCh38/Human_GRCh38 -U
myc_cml_trimmed.fq -p 4 --no-unal -S
${BAMDATA}/output_input.sam
```

5.6. Step 5: Processing of the aligned files

The output of the alignment process is a SAM (sequence alignment map) file. However, the SAM files are heavy files difficult to work with. For this reason, it is much better to transform the SAM files into BAM (binary alignment map) files. BAM and SAM files

contain the same information, but the BAM files are written in binary format, making them smaller and more efficient to work with.

In order to manipulate the SAM and the BAM files this workflow uses the SAMtools tool, which has been specifically designed to work with this type of files. Firstly, the SAM file is used as input for SAMtools in order to obtain the BAM format.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg /tool_deps/_conda/pkgs/samtools-
1.4.1-0/bin/samtools view -Sb output_input.sam >
output_input.bam
```

From this BAM file, the header will be generated, which contains information about the sample that is being analyzed. Once it is obtained, the bam file is processed and the regions that have been aligned multiple times are eliminated. They start with XS:, so the workflow uses these lines of coding in order to eliminate those alignments. Once these regions are eliminated, the output is a SAM that is transformed again into a BAM file.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg /tool_deps/_conda/pkgs/samtools-
1.4.1-0/bin/samtools view -H output_input.bam >
input_header
```

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg /tool_deps/_conda/pkgs/samtools-
1.4.1-0/bin/samtools view -F 4 output_input.bam | grep -v
"XS:" | cat input_header - > unique_input.sam
```

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg /tool_deps/_conda/pkgs/samtools-
1.4.1-0/bin/samtools view -Sb unique_input.sam >
unique_input.bam
```

The lasts steps of the processing of the alignment output are to sort and index the generated BAM files. Sorting the BAM files reorganize the file according to the coordinates obtained from the alignment (coordinates from chromosome 1 will go first and so on), which is a way of improving the accessibility to the data. Also, the sorted

BAM file is indexed, creating a new index file with the information of the sorted BAM file, so the information can be accessed in a more efficient manner.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg /tool_deps/_conda/pkgs/samtools-
1.4.1-0/bin/samtools sort unique_input.bam -o
sorted_unique_myc_cml.bam
```

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg /tool_deps/_conda/pkgs/samtools-
1.4.1-0/bin/samtools index sorted_unique_myc_cml.bam
```

The final output of the alignment step is a sorted BAM file displaying only the non-repeated sequences, with an associated index data file for each BAM file.

5.7. Step 6: Peak calling

Once the alignment of the ChIP-seq files and the control files (in this case input files are used as a control, which basically are all the regions of the genome that have not undergone any further precipitation), the next step of the workflow is to determine the genomic regions that have been significantly enriched comparing the ChIP-seq sample and the control sample, which would be the background of the ChIP-seq. This step is known as peak calling.

In this workflow, MACS2 has been used for the peak calling step. It uses as input data the BAM files from the ChIP-seq and the control files obtained from the alignment step. The significance to decide what it is considered as a peak in the ChIP-seq sample can be modified, making it more restrictive or more flexible. The standard conditions to consider an enriched region as a peak is having a p-value < $10e^{-5}$.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/macs2.simg macs2 callpeak -B -t
sorted_unique_myc_cml.bam -c sorted_unique_myc_input.bam -n
${ROOTDIR}/pipeline/output_files/myc_cml_peaks
```

The output of this step is a narrowPeak file which provides a list of the peaks that have been statistically significantly enriched in the ChIP-seq analysis (Figure 7). The file is a

canonically known as BED (Browser Extensible Data) file with some extra fields. The BED file has 3 main fields which are the chromosome number, the start and the end coordinates of the peak. Additionally, the BED file can also display the score of the peak, the strand and the name given to that peak. The narrowPeak file generated by the MACS2 tool also displays the statistic values of the peaks, specifically it displays the signal value (overall enrichment of the region), the p-value ($-\log_{10}$), the q-value ($-\log_{10}$) and the point-source of this peak.

```
chr1    778509  778913
chr1    904628  904972
chr1    906905  907054
chr1    921102  921359
chr1    923709  923971
chr1    924191  924344
chr1    925116  925253
chr1    938253  938451
chr1    939224  939349
```

Figure 7. BED file. From left to right: chromosome, start coordinates and end coordinates.

With the BED file the main results of the ChIP-seq analysis can be obtained, as it can be used to extract the list of genes that have been enriched in the ChIP-seq analysis and can be used as input for many different downstream analyses.

5.8. Step 7: Quality control of the peak calling

Once the peak calling has been performed, a quality check of the ChIP-seq enriched peaks needs to be performed in order to confirm that the peaks are good quality peaks enriched over the control sample.

The Bioconductor R package ChIPQC is widely used for this step of the analysis. The input for this tool is the BAM files from the ChIP-seq data and from the control data files, as well as the narrowPeak files obtained from the peak calling step. The tool performs different quality metrics on the samples and generates a quality report for the experiment. A summary file with the conditions and with the files required is entered and then the quality check is performed.

```
# Load the summary file
> samples<-read.csv("samples.csv")
# Create the ChIPQC object
> chipObj<-ChIPQC(samples, annotation="hg38")
# Obtain the quality report
> ChIPQCreport(chipObj, reportName="ChIPQC report: myc",
reportFolder="ChIPQCreport")
```

The quality check report provides different quality metrics. For each sample (ID) the conditions of the experiment (tissue, factor precipitated, conditions and replicate) are displayed (if they are not known then the report displays a NA). Then, the number of reads for each sample, their length (ReadL) and the duplication rate (although as the BAM files have been filtered, the duplication is expected to be low). The SSD (squared sums deviations), RiP (Reads in Peaks), RiBL (Reads overlapping in Blacklisted Regions), FragL and RelCC metrics analyze the quality of the peaks called. The SSD represents the coverage of the reads across the genome (higher SSD values represent more uniformity in the coverage which means a better enrichment). The RiP metric represents the percentage of reads called as peaks (usually around 5% or higher is considered as good quality). The RiBL represents the percentage of reads that overlap with regions that have artificially high signal (low RiBL is indicative of better quality).

5.9. Step 8: Annotation the peaks

The BED file obtained in the peak calling step consists, in its most basic format, in the chromosome number and the start and end coordinates of each peak. In order to extract more precise conclusions performing downstream analysis, the gene list associated with the peaks is needed. This process is known as annotation.

There are different ways to annotate the BED files. In this workflow, the Bioconductor R package ChIPseeker is used. It uses as input the BED files and it maps the coordinates to the nearest gene in the genome. Moreover, it also gives information about the region

regarding the associated gene (if it is a promoter, an intron, an intergenic region, etc.). For performing this mapping, it also requires as input the annotation database (which can be downloaded to R with the *TxDb.Hsapiens.UCSC.hg38.knownGene* package).

```
# Read bed file
> peaks<-readPeakFile("myc_cml_published.bed")
# Load annotation database
> txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene
# Annotate the peaks
>  peakAnnot<-annotatePeak(peaks,tssRegion=c(-3000,  3000),
TxDb=txdb, annoDb="org.Hs.eg.db")
```

Once the peaks are annotated, it generates a list of genes associated with the different peaks as output. With the gene list comparisons between different ChIP-seq datasets can be performed. The most common representation of a comparison between ChIP-seq datasets is to perform venn diagrams.

Venn diagrams show the degree of overlapping between different conditions, in this case gene lists. The R package vennDiagram can perform high-quality venn diagrams taking as input only the gene lists (Figure 8).

```
# Generate the lists to compare
> cml_our<-read.table("myc_cml_genes.txt")
> cml_pub<-read.table("myc_cml_genes_published.txt")
> geneLists <- list(cml_our,cml_pub)

# Remove NA values
> geneLists <- lapply(geneLists, function(x) x[!is.na(x)])
> VENN.LIST <- geneLists

# Design the venn diagram
> venn.plot <- venn.diagram(VENN.LIST, euler.d=TRUE,
scaled=TRUE ,NULL,
                          fill=c("dodgerblue", "orange"),
                          alpha=c(0.5,0.5),
                          cex = 2,
```

```
                        cat.fontface=8,
                        category=c("Workflow",
"published"),
                        main = "CML genes workflow vs
published")

# Plot the venn diagram
> grid.draw(venn.plot)
```

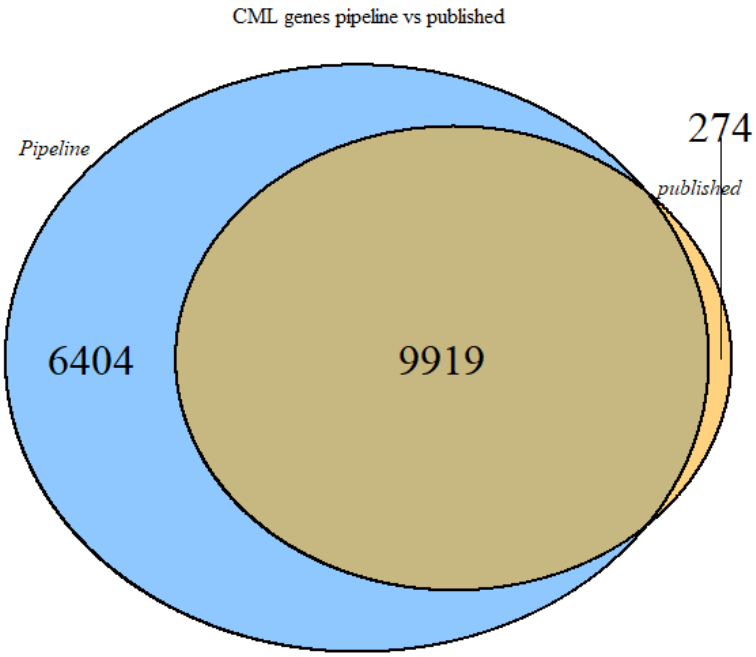

Figure 8. Venn diagram comparing gene lists. Venn diagram comparison between gene list obtained with the designed workflow (blue) and with the published ENCODE data files (orange).

Comparing the datasets obtained with the designed workflow and with the processed data downloaded from the ENCODE database, it can be observed that the peaks are mostly the same, being more in the analysis performed with the designed workflow. However, a very small proportion of peaks is obtained only in the published processed data files, ensuring the effectiveness of the workflow. The difference probably comes from using different peak calling tools or for using different thresholds of significance.

The output of the annotation, apart from the gene list associated with the peaks, also associates the peaks to a genomic region. This can be used in order to plot the overall genomic distribution of the ChIP-seq (Figure 9). In the representation obtained with the example dataset analyzed with the workflow, it is displayed that 50% of the peaks are

annotated to promoter regions, while the other peaks are distributed mostly in distal intergenic regions and introns.

```
# Visualize genomic distribution of annotated peaks
> plotAnnoBar(peakAnnot)
```

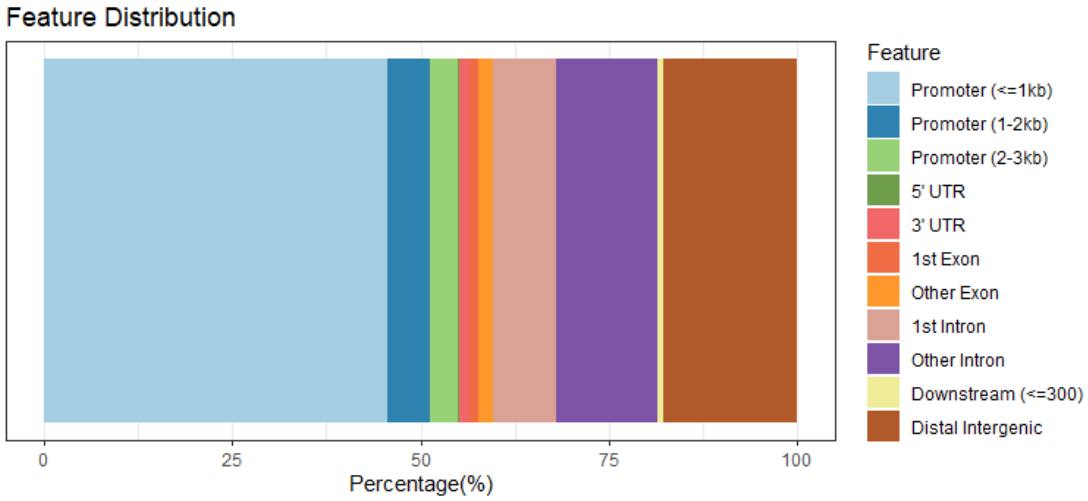

Figure 9. Genomic distribution of the enriched regions. Different colors are associated with different genomic regions, each case listed in the right legend.

Distribution analysis is useful to see if a transcription factor or a histone mark is especially concentrated around a specific region. For example, it would be expected that a transcription factor that positively drives the expression of its target genes enriches mostly for promoter regions and not to distal intergenic regions, while other proteins such as histones are evenly distributed among the genome.

5.10. Step 9: Genomic visualization

Once the sequences obtained from the ChIP-seq have been aligned in the BAM files, one of the outputs that can be withdrawn from those are the enriched regions (peak calling). Another output that comes from the BAM files are the bigWig files, which allow the visualization of the ChIP-seq data.

The bigWig file has all the continuous data indexed in a binary format that can be displayed in different tools and it can also be used as input for further representations.

In order to obtain the bigWig files, the workflow implemented uses deepTools, which has de bamCoverage tool, which takes as input the BAM files and their index file that was also generated in the alignment step.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg
/tool_deps/_conda/envs/__deeptools@2.5.1/bin/python
/tool_deps/_conda/envs/__deeptools@2.5.1/bin/bamCoverage -b
sorted_unique_myc_cml.bam --binSize 10 --normalizeUsingRPKM
-o ${ROOTDIR}/pipeline/output_files/myc_cml.bw
```

In order to visualize the bigWig files, this workflow uses the IGV software, which is a tool specifically designed to visualize sequencing data files (Figure 10). For the correct visualization of the data, firstly the genome that has been used for the alignment needs to be downloaded and installed. Once it is done, the different tracks that are to be visualized are loaded in the software. In the example observed in Figure 10, the tracks loaded are the ones corresponding to an example dataset that has been processed with the workflow and the processed files that were uploaded in the ENCODE database together with the raw ChIP-seq data, so the results can be easily compared. Different regions that have been called as peaks are being visualized. It is also common to add the control file so it is seen that there is no enrichment in the control. As it can be seen in Figure 10, there is no enrichment in the control file and both processed files look very similar. Differences can be accounted to different tools with different processing of the samples, but the final output is highly similar.

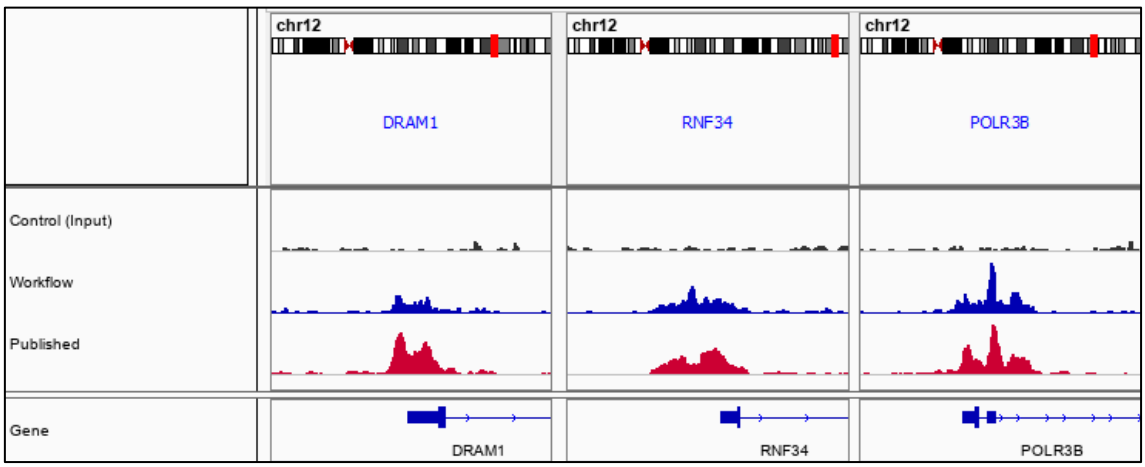

Figure 10. IGV visualization of ChIP-seq data. BigWig data files for control (input) and ChIP-seq data (processed with the workflow and published in ENCODE) are visualized. Three representative genes are visualized: DRAM1, RNF34 and POLR3B.

Another useful tool to visualize the ChIP-seq results is the plotHeatmap tool from deepTools. This tool generates a heatmap that associates a score with a genomic region. It uses as input the BED files in order to visualize the desired regions and the bigWig files in order to associate each region with their score.

For creating these heatmaps, firstly a matrix needs to be generated (first block of code) using the computeMatrix tool from deepTools. This tool can take several BED and bigWig files in order to represent different subset of regions in different conditions. It can also be used for sorting and filtering regions according to their score. The final output is a matrix that is used to create a heatmap with the plotHeatmap tool (second block of code).

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg
/tool_deps/_conda/envs/__deeptools@2.5.1/bin/python
/tool_deps/_conda/envs/__deeptools@2.5.1/bin/computeMatrix
reference-point -R myc_cml_peaks_peaks.narrowPeak --
referencePoint TSS -S myc_cml.bw myc_cml_published.bigWig -
b 10000 -a 10000 -p 6 --samplesLabel Workflow Published -o
matrix_scaled.gz --sortRegions descend --outFileNameMatrix
matrix_scaled.tab --outFileSortedRegions genes_sorted.bed
```

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/deepTools.simg
/tool_deps/_conda/envs/__deeptools@2.5.1/bin/python
/tool_deps/_conda/envs/__deeptools@2.5.1/bin/plotHeatmap -m
matrix_scaled.gz --colorMap YlGnBu --legendLocation none --
sortRegions no --missingDataColor "#FFF6EB" -out
heatmap.pdf
```

The output is a pdf file with the heatmap. Figure 11 represents a created heatmap using as input the narrowPeak files obtained from the peak calling step with the workflow (even though if the files were in basic BED format would be enough) and the bigWig file generated in the workflow compared to the one published in ENCODE.

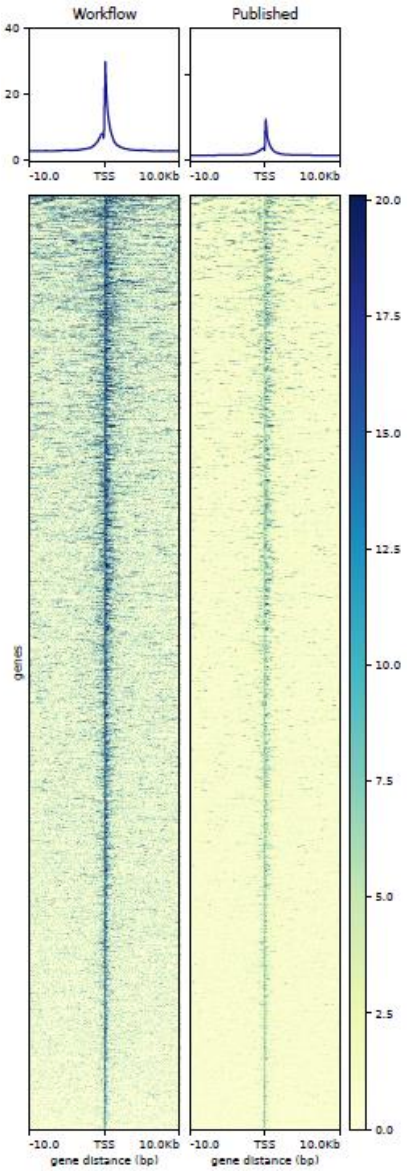

Figure 11. Heatmap representing significantly enriched regions. Left heatmap corresponds to the output obtained with the designed workflow and the right heatmap corresponds to the output published in ENCODE. Upper density plots represent the score of the enriched regions and their concentration in the TSS (transcription start site).

Comparing both heatmaps it can be seen that the genomic regions from the peak calling with the designed workflow display genomic enrichment in both data files, concluding the high similarity between both processed files.

5.11. Step 10: Downstream analyses with the processed files

5.11.1. Functional Enrichment Analysis

From the list of genes obtained in the annotation step of the BED files, one of the most common analyses that can be undertaken is to perform a functional enrichment analysis. Many tools can perform this type of analysis and with different databases that associate the given genes to a biological process, a pathway or even to a disease.

In this workflow, the Enrichr database has been used for performing the functional enrichment analysis. The input given is the gene list and it displays and the output is divided according to the information desired, from pathways related to the enriched genes to functions (Figure 12).

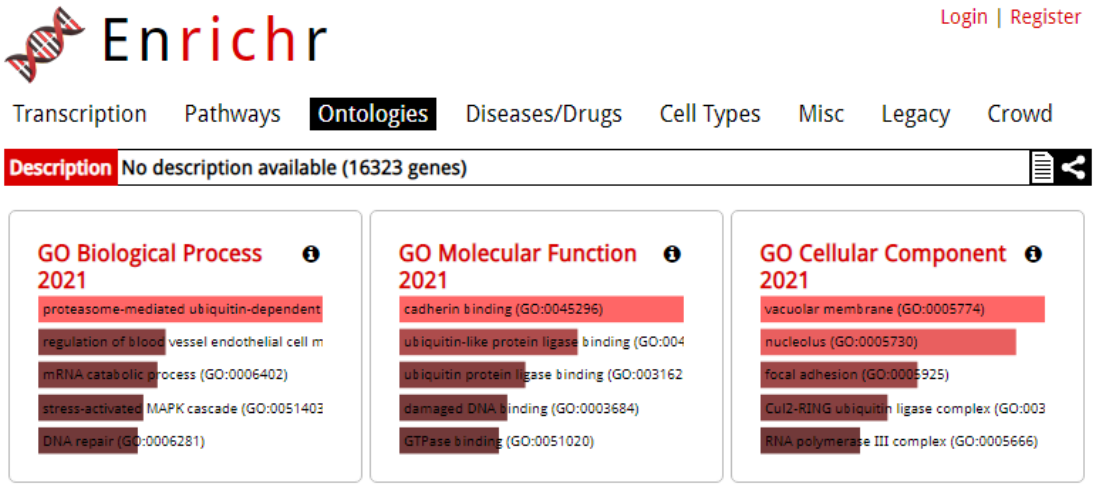

Figure 12. Functional enrichment analysis. EnrichR results of the genes annotated in the annotation step. Different result types are listed in the upper part.

5.11.2. Motif Enrichment Analysis

A common analysis performed on ChIP-seq peaks is the motif enrichment analysis. A DNA motif is a short sequence nucleotide pattern that is distributed along the DNA sequence and it is associated with a transcription factor that is able to bind to this sequence (D'haeseleer, 2006). The motif analysis is commonly used to determine which transcription factors might be regulating a subset of genes by detecting the binding motifs of those transcription factors in the regulatory regions of the target genes.

In the ChIP-seq analysis, motif enrichment can be used for determining a consensus sequence *de novo* for a given transcription factor. There have been recent studies that have performed wide *de novo* characterization of DNA motifs, which is key for the prediction of transcription factor binding sites in ChIP-seq analyses (Boeva et al., 2010). However, if the motif has already been described, specific tools can be used to determine which of the known motifs are enriched in the ChIP-seq analysis. The most used tool for this analysis is the MEME-ChIP, a tool from MEME-Suite.

In order to use the MEME-ChIP tool, the BED file needs to be transformed in a FASTA file. The FASTA file contains the sequence comprised between the start and end coordinates for each peak in the BED file (Figure 13A). The FASTA file format starts every sequence in the file with ">" followed by the identification of each sequence. Next, the sequence that corresponds to the identification in the previous line is obtained. In the file there are listed as many sequences as peaks are, all of them following the same format.

For obtaining the sequence, the bedtools tool can be used, as it has the getfasta function that takes the BED and the reference genome as input files and generates a FASTA file.

```
> singularity exec -B ${ROOTDIR}:${ROOTDIR}
${IMAGES}/bedtools.simg bedtools getfasta -fi
${REFGENOME}/Human_GRCh38/hg38.fa -bed myc_cml.bed  -fo
myc_cml.fasta
```

The FASTA file obtained can be uploaded into the MEME-ChIP tool and it will perform the motif enrichment analysis (Figure 13B).

A

```
>chr1:778510-778913
CTTCTTCAGCCGGCAACACACAGAACCTGGCGGGGAGGTCACTCTTACCAGTCCCCACTCTGATGAGAAAACTGCCCAGCTCCAGGCACCATGGCGCCCCAGTGATGTAGCCGAACACCCGCGCCTCTAACGTCGC
>chr1:904629-904972
GGCCGCCGCCTCCTCCGAACGCGGCCTCCTCCTCCTCCGAACGCGGCCGCCTCCTCCTCCGAACGTGGCCGCCTCCGAACGTGGCCGCCGCCTCCTCCGAACGTGGCCGCTTCCGCAGCGCCCGGCGCAGGCCGCA
>chr1:906906-907054
CCACACCCGAGGAGGCCAGAGGTGCAGGGAGCATGGGCTGTCGCTTCCCCTTTAAGCACACTCATTCACCACACCCGAGGAGGCCAGAAGTGCAGGGAGCATGGGCTGGGTGCACCTCCGCAGGAGAGAAGGCTGAC
>chr1:921103-921359
TGGGGCAGCGGCCCCTGGCGCCCCCACACTCCCCCAGGACGTCTGGGGTCACCGTCTGCTCCATGGCAGCCCCAGGGTTATTTATGACCTCCTCCCTCTGGCGGCGGGAGGCAGGCTCCAGCCTCAGCCCAGCGGC
>chr1:923710-923971
cgcgaccagcccagcccatcccagtcccgcGCGGAGTCCTGGATTCCAGCCGCTCGCAGTGACTCGGTACTCGGGATAGTGCCGGGGGCCGCAGCCCTGTCCCGCTGCCGCCGCCGGATGCCCCGAGTCGGCCGTCA
>chr1:924192-924344
ATCCGGGATCGATAGCAGCTCCATGTCTCCGGCCTCTGAGGCCCCGCCGGCCGGCTGGGCAGTCCGGGGAGGCCTGGCGGGCGGCGCGTAGGCGGCGGCTGCGGGCGCCGGGGCGCACTAGCGGACGGCGTGGGCG
>chr1:925117-925253
CCCCCCGCTGCACCTCCCCCCAGCTTGGGCCACAGCGCTTGGGGCTCGCGGGCCGCTCCCTCCGCTCGGAAGGTCTCTGCGAGGCTCCTGGGCCTTAAGGCCCGAAGGAAGTTTACGGGGACTCGAGAGAGCGGGC
```

B

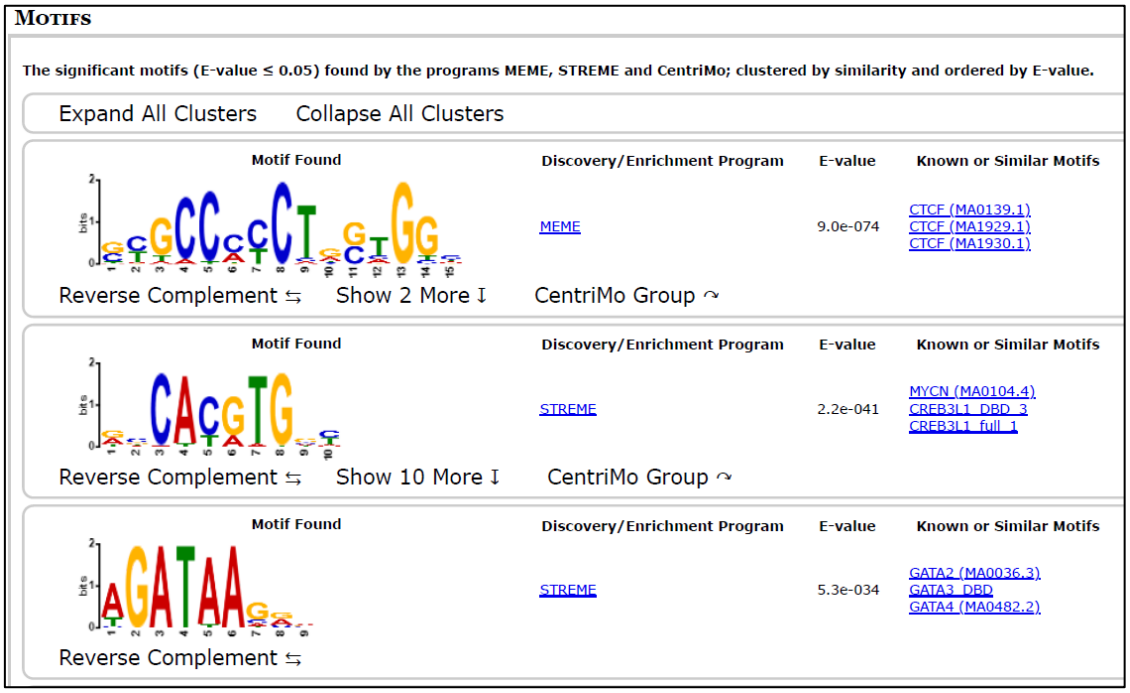

Figure 13. Motif Enrichment analysis of enriched peaks. A) Example of a given FASTA file format where the identifications are the coordinates of the peaks obtained from the BED file. B) Output obtained from the MEME-ChIP tool, which is formed by the motif sequence, the algorithm implemented, the e-value and the transcription factor associated with this known motif.

From the motif enrichment analysis output it can be observed that different motifs are enriched, not only the MYC motif (enriched as the second top enriched motif). It is common that many different motifs appear to be enriched in the ChIP-seq analysis, especially if the number of enriched regions is high. Most of the genes in the genome are regulated by different factors, which would explain the high presence of different DNA motifs in the ChIP-seq analysis.

6. Final remarks

6.1. Conclusions

In this project, we have implemented and applied a workflow to process ChIP-seq data. Based on the objectives we conclude that:

1. There is a large variety of tools for analyzing ChIP-seq data
2. Main steps of the ChIP-seq data analysis are the quality check, the adapter trimming, the alignment and the peak calling. Variations in the tools used for each step can give rise to different outcomes, so they are a source of variability
3. The implemented workflow is able to completely process raw ChIP-seq data
4. The workflow has been tested with public data and the results are highly similar to the ones published using the same data, highlighting the efficiency of the implemented workflow
5. Processed files can be used for answering key biological questions

6.2. Future steps

Next steps of this work are to actively use this workflow in order to analyze different ChIP-seq datasets and try to obtain answers to some key questions, such as the role of transcription factors in normal cells compared to cancer cells. Would they change their pattern binding? These changes would lead to a different expression pattern? The factors are working through the same motifs? Could they be changing their chromatin partners? These are just some of the questions that ChIP-seq techniques can answer and if the data is analyzed properly and under the same conditions, the results can be easily compared in order to obtain valuable results.

Another future step to improve this workflow is use a workflow management system in order to automatize the usage of the workflow. Snakemake or Nextflow are examples of these systems that would be suitable options for this workflow.

6.3. Project follow-up

Several changes have been made since the beginning of this project.

During PEC1, the main objectives of this project were raised and the focus of the project was set in the implementation of a workflow.

During PEC2, the objectives of the project had to be redesigned and the specific objectives of each main objective were also well stablished according to the new planification of the project, which was better designed with the help of the supervisor. The limitations of the project were also raised at this point.

During PEC3, the main steps of the workflow were reviewed and some were added. Also, different sources of information were considered for downloading the data for the testing of the workflow. Some of the steps required more time than the originally scheduled, so the general planification had to be adapted.

7. Abreviations

BAM	Binary Alignment Map
BED	Browser Extensible Data
BWA	Burrows-Wheeler Aligner
ChIP	Chromatin Immunoprecipitation
ChIP-seq	Chromatin Immunoprecipitation coupled to high throughput sequencing
DNA	Deoxyribonucleic Acid
GC	Guanine - Cytosine
GEO	Gene Expression Omnibus
IGV	Integrative Genome Viewer
MACS	Model-based Analysis of ChIP-seq
MGA	Mass Genome Annotation
NGS	Next Generation Sequencing
qPCR	Quantitative polymerase chain reaction
RiBL	Reads overlapping in Blacklisted Regions
RiP	Reads in Peaks
RNA	Ribonucleic Acid
SAM	Sequence Alignment Map
SSD	Squared Sums Deviations
TSS	Transcription Start Site

8. Bibliography

Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nature reviews. Genetics, 17(8), 487–500.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ather, S. H., Awe, O. I., Butler, T. J., Denka, T., Semick, S. A., Tang, W., & Busby, B. (2018). SeqAcademy: an educational pipeline for RNA-Seq and ChIP-Seq analysis. F1000Research, 7, ISCB Comm J-628.

Bardet, A. F., He, Q., Zeitlinger, J., & Stark, A. (2011). A computational pipeline for comparative ChIP-seq analyses. Nature protocols, 7(1), 45–61.

Boeva, V., Surdez, D., Guillon, N., Tirode, F., Fejes, A. P., Delattre, O., & Barillot, E. (2010). De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. Nucleic acids research, 38(11), e126.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120.

Bujold, D., Morais, D., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K. C., Laperle, J., Markovits, A. N., Pastinen, T., Caron, B., Veilleux, A., Jacques, P. É., & Bourque, G. (2016). The International Human Epigenome Consortium Data Portal. Cell systems, 3(5), 496–499.e2.

Carroll, T. S., Liang, Z., Salama, R., Stark, R., & de Santiago, I. (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in genetics, 5, 75.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics, 14, 128.

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics, 12, 35.

D'haeseleer P. (2006). What are DNA sequence motifs?. Nature biotechnology, 24(4), 423–425.

Das, P. M., Ramachandran, K., vanWert, J., & Singal, R. (2004). Chromatin immunoprecipitation assay. BioTechniques, 37(6), 961–969.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research, 46(D1), D794–D801.

Dawson, M. A., & Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell, 150(1), 12–27.

Dréos, R., Ambrosini, G., Groux, R., Périer, R. C., & Bucher, P. (2018). MGA repository: a curated data resource for ChIP-seq and other genome annotated data. Nucleic acids research, 46(D1), D175–D180.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207–210.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England), 32(19), 3047–3048.

Furey T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature reviews. Genetics, 13(12), 840–852.

Hannon, G.J. (2010) FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. Genome research, 12(6), 996–1006.

Kharchenko, P. V., Tolstorukov, M. Y., & Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology, 26(12), 1351–1359.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods, 12(4), 357–360.

Krueger, F. (2012). Trim Galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. [Online]. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. Cell, 172(4), 650–665.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357–359.

Lawrence, M., Daujat, S., & Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends in genetics : TIG, 32(1), 42–56.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079.

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., Pape, U. J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., & Liu, X. S. (2011). Cistrome: an integrative platform for transcriptional regulation studies. Genome biology, 12(8), R83.

Lou, S., Li, T., Kong, X., Zhang, J., Liu, J., Lee, D., & Gerstein, M. (2020). TopicNet: a framework for measuring transcriptional regulatory network change. Bioinformatics (Oxford, England), 36(Suppl_1), i474–i481.

Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics (Oxford, England), 27(12), 1696–1697.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet.Journal, 17(1), 10. doi: 10.14806/ej.17.1.200

Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods (San Diego, Calif.), 187, 44–53.

Nakato, R., & Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Briefings in bioinformatics, 18(2), 279–290.

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., & Meno, C. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO reports, 19(12), e46255.

Park P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nature reviews. Genetics, 10(10), 669–680.

Park, S. J., Kim, J. H., Yoon, B. H., & Kim, S. Y. (2017). A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages. Genomics & informatics, 15(1), 11–18.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research, 44(W1), W160–W165.

Reynolds, N., O'Shaughnessy, A., & Hendrich, B. (2013). Transcriptional repressors: multifaceted regulators of gene expression. Development (Cambridge, England), 140(3), 505–512.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature biotechnology, 29(1), 24–26.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., & Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature biotechnology, 27(1), 66–75.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. Nature reviews. Genetics, 13(9), 613–626.

Weinhold B. (2006). Epigenetics: the science of change. Environmental health perspectives, 114(3), A160–A167.

Wu, D. Y., Bittencourt, D., Stallcup, M. R., & Siegmund, K. D. (2015). Identifying differential transcription factor binding in ChIP-seq. Frontiers in genetics, 6, 169.

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics (Oxford, England), 31(14), 2382–2383.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology, 9(9), R137.

Zhang, Y., Sun, Z., Jia, J., Du, T., Zhang, N., Tang, Y., Fang, Y., & Fang, D. (2021). Overview of Histone Modification. Advances in experimental medicine and biology, 1283, 1–16.

9. Annex

As the code used in this workflow has been uploaded to Github in order to be more accessible. https://github.com/DAARMO/Workflow