UNIVERSITY
OF UDINE

Department of
Mathematics, Computer Science and Physics

MASTER THESIS IN
INFORMATICA

Analysis methods for genic expression data

CANDIDATE
Lorenzo Iuri

SUPERVISOR
Prof. Carla Piazza

CO-SUPERVISOR
Prof. Riccardo Spizzo

Academic Year 2020-2021

Institute Contacts
Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Università degli Studi di Udine
Via delle Scienze, 206
33100 Udine — Italia
+39 0432 558400
https://www.dmif.uniud.it/

Abstract

In the last two decades, technical and methodological advancements have made fast, precise and comprehensive genomic analyses more and more accessible. The aim of this work is to assess the effectiveness of modern bioinformatic tools and methods in genome-wide analyses, both in the generalization of local studies and in the description of the main biological functions affected in mutant cells. In particular, a specific study on the TP53-mutant breast cancer model is considered and bioinformatic tools are employed to characterize the relationships between p53's binding sites and gene expression.

In the first chapter, genetic and sequencing concepts are introduced, while the second one presents the main bioinformatic formats, tools and methods, providing a more in-depth overview of the RNA-Seq and ChIP-Seq methodologies. The following chapter describes the analysis we performed on mutant breast cancer cells, from the classification of p53's binding sites to the estimation of the mutation's effect on gene expression, to the analysis of the altered biological functions and pathways.

We obtained multiple evidences suggesting p53's propensity in binding to *protein coding* genes rather than *non-coding RNAs* ones. In our analyses, we concluded that the employed methods and tools can offer satisfactory results and we believe they can serve as a useful support in modern genetic studies. The proper usage of these tools, however, requires bioinformatics and statistics knowledge and expertise in order to format data, verify if the preconditions are met and interpret the outputs. Furthermore, multiple analysis pipelines may be needed to achieve the desired results.

Contents

Introduction

Since the discovery of DNA's key role in all life forms, the efforts of researchers and scientists in investigating and modelling the biological functions and processes within the cells transformed genetic analysis. The awareness that every organism's traits and functions are encoded and concealed in the nucleotide sequence of its DNA prompted the development of advanced DNA sequencing methods.

From its introduction in the early 1970s, DNA sequencing techniques and technologies rapidly improved, allowing by the end of 2003 the completion of the Human Genome Project, maybe one of the most famous achievements of modern genetics. The accomplishment marked the beginning of a new era for DNA sequencing in which modern *high throughput* sequencing machinery granted an exponential increase in speed and decrease in costs, making sequencing accessible to more and more laboratories and establishing sequencing experiments as an important part in many studies.

Today, the knowledge of DNA sequences has become indispensable for biological research and in numerous applied fields such as medical diagnosis, biotechnology, forensic biology and virology. Comparing healthy and mutant DNA sequences can help diagnose different diseases and can be used to guide patient treatment. The usage of sequencing is not limited to defining the base sequence of genes: resourceful methods allow the observation of gene expression and the study of protein binding sites.

The amount of data produced by modern sequencing instruments can be impressive and, often, manual analyses are not applicable, particularly in the context of genome-wide studies. For this reason, the need for automation in assembling, aligning and inspecting large genomic data stimulated the development of advanced methods, efficient algorithms and reliable tools in the novel field of bioinformatics.

The content of this work is the description of the results of a collaboration with the CRO National Cancer Institute (Centro di Riferimento Oncologico) in Aviano aimed at evaluating the usage of bioinformatics tools within laboratorial activities. The research fits in a currently ongoing study on TP53-mutant breast cancer cells focused on the role of the LINC01605 gene in tumor development.

Our intent is to evaluate the capability of modern bioinformatics instruments to assist research activities, considering a specific case study based on the usage of high throughput sequencing techniques. Such instruments will be used to extend to a greater scale a local analysis aimed at inspecting the potential interactions between genes and proteins and to support the research activity through descriptions and visualizations.

The first chapter provides introductory information on elementary genetic topics, such as nucleic acids, genes and mutations. At the end of the chapter, two modern methods based on high throughput sequencing, RNA-Seq and ChIP-Seq, are introduced.

The following chapter intends to provide an overview of the main methods, programs and formats commonly used to encode and visualize genetic data. The problems of reads' quality control, alignment

and quantification will be briefly discussed. The last two sections are dedicated to the problem of analysing the data produced by RNA-Seq and ChIP-Seq experiments and to the description of two ad-hoc methods employed by modern tools.

The third chapter describes the questions, the problems and the results of the analyses we performed on mutant breast cancer cells. The cell line we considered, known as MDA-MB-231, is characterized by a mutation on the TP53 gene, common in many human tumors: it was observed that this gene has a crucial role in the regulation of cell development and proliferation, serving as a tumor suppressor, but, when mutated, it often acquires new functions which promote tumor development.

Section 3.2 presents the methods we followed to estimate the binding sites of the mutant p53 protein (synthesized from the mutant TP53 gene) across all the genome, as a preliminary step in the study of the interactions between the protein and the genes it potentially targets.

In the following section, we tried to characterize the estimated binding sites, in order to study the type of regions that mutant p53 primarily interacts with. To improve the accuracy of the estimates, we employed a normalization method based on the comparison of the estimated binding sites to randomly sampled regions.

Section 3.4 describes how we used gene expression data (data about the transcriptional activity of genes) provided by high throughput sequencing methods to study mutant-p53's role on gene expression regulation, inspecting the relationships between p53's estimated binding sites and genes manifesting expression changes upon mutant-TP53's silencing.

The last two sections focus on functional analysis, the study of the biological pathways and processes targeted by a molecule or, in general, differing between two experimental conditions. Section 3.6 introduces the main information sources and functional analysis methods and describes an application to our TP53 study. Presenting the challenges of comparing the results of different methods, a similarity metric is proposed and applied to the employed tools. In section 3.7, finally, we describe the tools and visualizations we considered most appropriate for the summarization of a functional analysis on the role of the gene LINC01605 on tumor progression and we discuss the results of their application.

All the commands and scripts we used to perform the described analyses are presented in the appendix and made available on github (`https://github.com/lorenzoiuri/thesis-support-material`). This information can be used to reproduce the results and to assemble automated data analysis pipelines.

1

Biological and genetic concepts

In this chapter some simple biological and genetic concepts will be introduced. The first section presents the basic elements that carry the genetic information in every living organism: nucleotides and nucleic acids. The following section illustrates how the genetic information is stored and organized and the mechanisms employed in the cells to express it. Finally, in the last section, sequencing and genomics are introduced and an overview of some sequencing techniques and technologies is presented.

1.1 The structure of DNA and RNA

DNA is the hereditary material in humans and in almost all other organisms. DNA specifies the structure and functions of living things, but it also serves as the primary unit of heredity in organisms of all types. Whenever organisms reproduce, a portion of their DNA is passed along to their offspring. This transmission of all or part of an organism's DNA helps ensure a certain level of continuity from one generation to the next, while still allowing for slight changes that contribute to the diversity of life.

1.1.1 Nucleotides

Nucleotides are organic molecules composed of three units: a five-carbon sugar molecule, a phosphate group and a nitrogenous base. The carbon atoms in the sugar molecule are commonly numbered from 1' to 5', starting from the one bonded to the nitrogen atom of the nitrogenous base. The sugar molecule is *deoxyribose* if the 2'-carbon is only bonded to hydrogen (H) atoms, whereas is called *ribose* if it is bonded to a hydroxyl (OH) group.

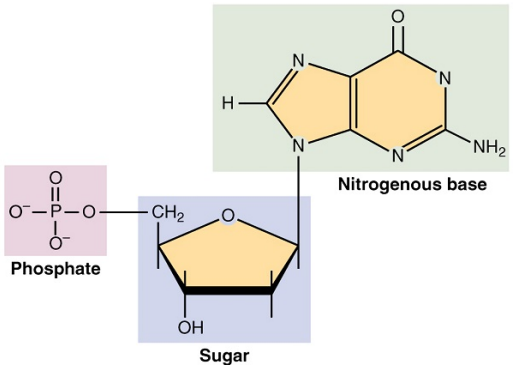


Figure 1.1: A monophosphate nucleotide, its nitrogenous base represents one side of a base pair. Image copyright Textbook OpenStax Anatomy and Physiology.

In the nucleotides that make up DNA the sugar is deoxyribose and four different nitrogenous bases can be found: guanine and adenine (purine bases), thymine and cytosine (pyrimidine bases). Nucleotides are often referred to with the first letter of their base (A, C, G or T).

1.1.2 Nucleic acids

Nucleotides can be found tied together forming a chain in nucleic acids (DNA and RNA): large biomolecules responsible for encoding and storing the information necessary for every life form.

Even though the chemical constitution of the nucleotides, as well as their abundances in the cells of some organisms, was already known in the early 1950s [25], it was only with the research and the discoveries of J.D. Watson and Francis Crick in 1953 that the structure of DNA was modelled as the one we know today.

DNA

According to the model of Watson and Crick, DNA (deoxyribonucleic acid) is a 3-dimensional structure which is composed by two strands of nucleotides twisted on themselves in the shape of a double helix. The nucleotides of each strand are bonded through phosphodiester linkages: chemical covalent bonds between the 5'-carbon of a nucleotide and the 3'-carbon of the next one. The two strands are held together by hydrogen bonds between the bases of each strand: through their interaction the two complementary nucleotide strands assume the double-helical conformation. The hydrogen bonds are weaker than the covalent ones: this simplifies the separation of the DNA strands during DNA replication and transcription.

The genetic information in DNA is encoded through the sequence of bases, while the sequence of phosphate and deoxyribose sugar units that compose each strand serves as the backbone for the structure. The backbone of each strand has a 5'-to-3' direction that refers to the order of the 5' and 3'-carbons in the bonded nucleotides: the two strands are paired so that their backbones are in the opposite orientation.

In the early 1950s, Rosalind Franklin, through the analysis of X-ray diffraction data on DNA, suggested that the DNA's structure is long and skinny, has a fixed diameter and is composed by two parallel strands [41]. Several years earlier, Erwin Chargaff, studying large DNA sections from different organisms, observed that the abundances of adenine and thymine nucleotides are unquestionably comparable and the same holds for cytosine and guanine [25].

Through these decisive observations, Watson and Crick understood that, for the structure to have a fixed diameter, a purine base should always be paired to a pyrimidine one: purine bases are thicker than pyrimidine ones and the length of the two bonded together matches the diameter observed by Franklin. Furthermore, in order for the observation of Chargaff to hold true, adenine should always be paired with thymine (and not with any pyrimidine) and the same can be said for cytosine and guanine. This consideration is reinforced by the enhanced stability provided by the chemical bonds between the mentioned base pairs.

RNA

RNA (ribonucleic acid) is a nucleic acid involved in many biological functions and activities in the cell and is present in a variety of forms, such as messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). RNA, in each of the formerly mentioned forms, is essential in the production of new proteins through the transcription and translation processes.

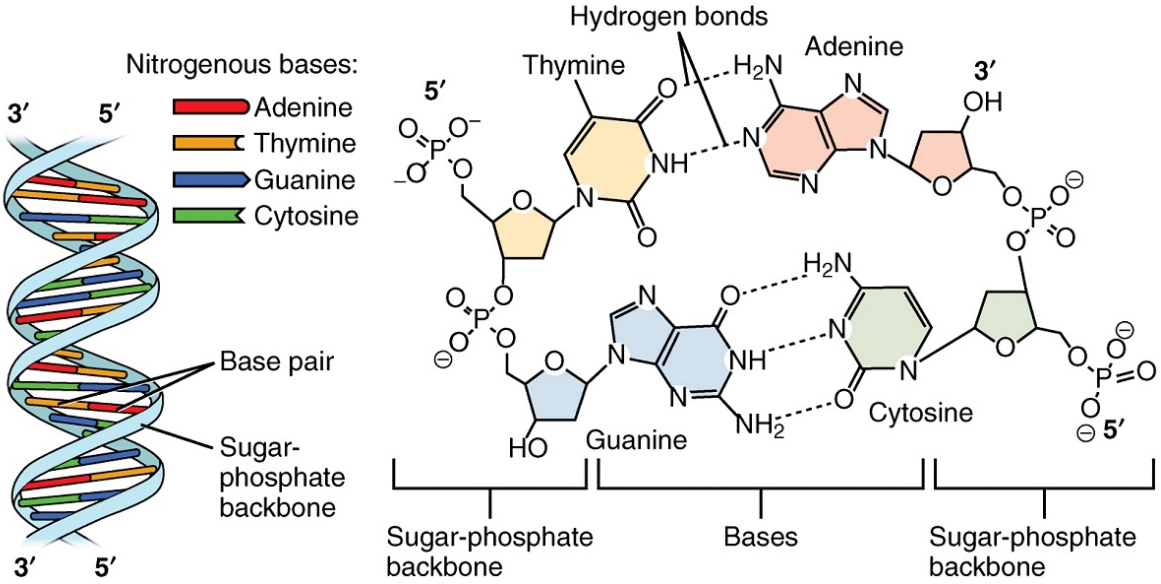


Figure 1.2: Arrangement of nucleotides within the structure of nucleic acids. In the figure on the right, four nucleotides form two base pairs: thymine and adenine (connected by double hydrogen bonds) and guanine and cytosine (connected by triple hydrogen bonds). In the figure on the left, the individual nucleotide monomers are chain-joined at their sugar and phosphate molecules, forming two "backbones" (a double helix).
Image copyright Textbook OpenStax Anatomy and Physiology.

As a nucleic acid, RNA has a structural backbone composed by sugar and phosphate groups and a sequence of bases which encode the information that the mRNA carries. In contrast to DNA, RNA is usually a single-stranded molecule folded onto itself and uracil replaces thymine.

Other than being involved in the production of new proteins, RNA has many other functions, forms and responsibilities in the cell such in ribozymes where it is involved in enzymatic functions and does not carry genetic information.

1.2 General structure of the genetic information

1.2.1 Chromosomes and genes

The genome of an organism is the complete set of its genetic information encoded in DNA. In eukaryotes, every cell contains the bulk of the entire genome of the organism in a region called nucleus. Inside the nucleus, DNA is divided in several separate agglomerates, each one of them being a long double helix. These pieces are independently coiled and packed, constituting a linear structure called *chromosome*. Each species has a characteristic number and shape of chromosomes. For instance, *Homo Sapiens* has 46 chromosomes while *Ovis aries* (the domestic sheep) has 54.

Certain species have multiple copies of the genome in each cell and consequently multiple copies of each chromosome. Humans are diploid, which means that their cells have two copies of the genome. Two chromosomes of the same pair are called *homologous* and have the same content, except for some minor differences.

DNA in the chromosomes is tightly packed in order to fit inside the nucleus of the cell. The helix is coiled around proteins, known as *histones*, resulting in a more compact and dense structure called *chromatin*, as depicted in figure 1.3.

Genes are functional regions, segments, along the chromosomes and are the primary carriers of information in the genome. It can be thought that the genome size of an organism can be a clue for its complexity because more genes are required to direct the formation of more complex organisms. Despite this general rule, organisms of apparently similar complexity can have very different genome sizes. For instance, the genome of the rice is approximately 40 times smaller than that of the wheat [22].

These differences are mostly accounted for by the gene density of a genome (which is measured in number of genes per megabase of genome). In the human genome, intergenic sequences account for about 60% of the total genome size and most of this DNA has no known function [26]. Furthermore, in most eukaryotes, parts of the genes called *introns* are not transcribed into proteins. Accounting for intergenic sequences and introns, only about 2% of the human genome finds its way into proteins.

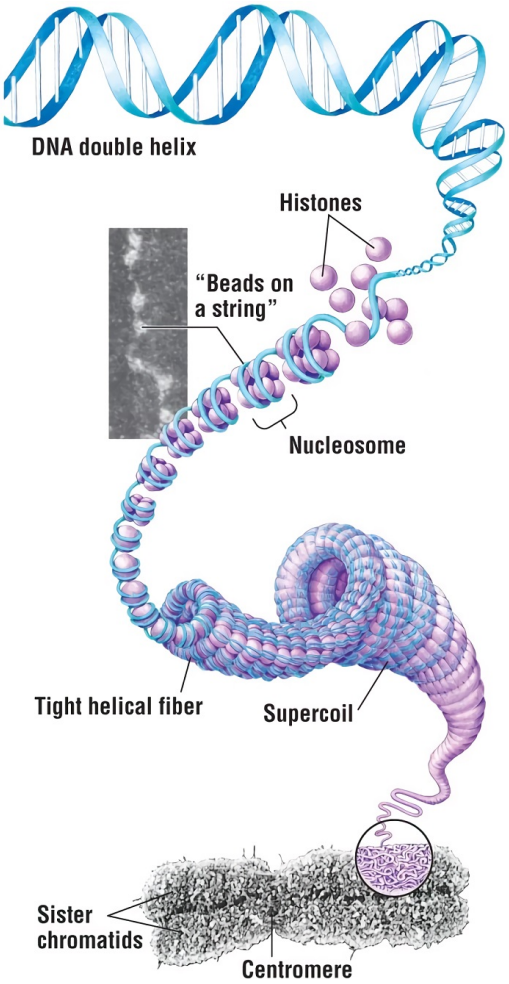


Figure 1.3: A simplified representation of the structure of DNA, as chromatin, in the chromosomes. Image copyright Pearson Education Inc., 2007.

1.2.2 Gene expression

Proteins are large biomolecules responsible for a large variety of functions inside the organism, such as DNA replication, stimuli response and molecules transportation inside and outside the cell. Proteins consist of one or more chains of amino acids: the function of a protein depends both on its amino acid sequence and the spatial conformation it assumes. The macroscopic attributes and behaviours of an organism derive from the functions of its cells which, in turn, emerge from the proteins they incorporate.

The way in which a gene influences phenotype (the composite observable characteristics of an organism) is known as gene expression. Proteins are produced in the cell from the instructions contained in the genetic code through the processes of *transcription* and *translation*. These processes involve multiple forms of RNA (notably mRNA, tRNA and rRNA) in order to create a blueprint of a gene and synthetize a new protein according to its information content.

Transcription

The first process that leads to the production of a protein is the DNA transcription. During this process, a transcript of the base sequence of a DNA region is produced and encoded as a strand of RNA. The initial form of the transcript, called pre-mRNA, is built by an enzyme called RNA polymerase by reading the sequence of bases of a DNA strand and linking a complementary sequence of ribonucleotides. The DNA strand read by the enzyme is the *template* strand for the gene it encodes, while the other is referred to as *coding* strand: this is because the synthesized pre-mRNA matches the base sequence of the coding

strand, being complementary to the template one. In general, both strands can be used as a template by the RNA polymerase, but each gene has its own template strand.

During the transcription, DNA is unwound in order to be read by the enzyme. The region of DNA that is exposed is known as transcription bubble: it extends just a few bases, and it is rewound short after being read. In eukaryotes, the resulting pre-mRNA, before being ready for the translation, need to be spliced: the *introns* along the transcript are cut out and the remaining sections, called *exons*, are spliced together. The resulting mRNA will then be exported out of the nucleus into the cytoplasm where it will be translated into proteins. A graphical representation of this process is illustrated in the upper part of figure 1.4.

Translation

In every cell, prokaryote or eukaryote, the translation occurs in the ribosome. The ribosome is an organelle composed of two subunits of different mass which exist separately in the cytoplasm: these subunits are joined together on the mRNA molecule during its translation. In the ribosome, the sequence of mRNA is read in direction 5'-to-3' in groups of three bases called *codons*. Each codon maps to an amino acid, so the content of the mRNA defines the structure of the protein that will be produced.

The translation of the mRNA molecule begins with the binding (with the help of initiation factor proteins) of the small ribosomal subunit near the start codon of the mRNA, which is characterized by the three bases `AUG`. The amino acids that will be chained in the protein are carried by molecules of RNA called tRNA. These molecules include an amino acid attachment site, which binds to a specific amino acid, and an anticodon: a sequence of three nucleotides that ties to its complementary codon in the mRNA.

The initiator tRNA molecule carrying the amino acid methionine binds to the `AUG` start codon of the mRNA transcript in the P (polypeptide) site of the ribosome. Next, the large ribosomal subunit attaches to the complex and starts the elongation phase. During this phase, the entire ribosome moves along the transcript and a new tRNA molecule binds to the A (amino acid) site bringing an amino acid. The ribosomes are modeled as having three sites: the P site is between the A site (3') and the E site (5'). As the ribosome slides in 5'-to-3' direction, peptide bonds are formed between adjacent amino acids and cause the tRNA molecule that occupied the P site to shift to the E (exit) site. This causes the release into the cytoplasm of the, now empty, tRNA molecule that is free to pick up another amino acid.

The translation process terminates when a stop codon (`UAA`, `UAG`, `UGA`) is read. No tRNA molecule binds to these codons. Instead, proteins called release factors attach to the stop codons facilitating the release of the mRNA from the ribosome. A simplified representation of the transcription and translation processes is shown in figure 1.4.

Controlling gene expression

The proteins that a cell produces dictate its functions. In single cell organisms, such as in prokaryotes, the functions and the needs of the organism may change during its life cycle. For example, the organism may need to produce more copies of a certain protein if it is scarce in the environment, or, vice versa, it may want to halt the production to conserve energy if the desired protein is already abundantly present.

The need to control gene expression is even more important in complex organisms characterized by many types of specialized cells. In eukaryotes, every cell of the organism contains the entire genome, but different cells may perform widely different functions and gene expression control mechanisms are needed to specialize the cells' behaviours and activities.

In prokaryotes, gene expression is controlled during the transcription and involves a complex called *operon*. This complex incorporates, other than the gene (or genes) and the binding point for the RNA polymerase (known as *promoter*), a segment of DNA called *operator* that can be used to enable or disable the transcription of the adjacent gene. Considering a gene that is normally transcribed, the binding of a compatible repressor protein to the gene's operator can inhibit its expression. If the repressor is activated by the effects of the protein synthetized by the gene then this control mechanism can diminish, or halt, the production of the same protein, establishing a feedback loop.

Gene regulation is more complex and diverse in eukaryotes. The regulation of the expression of a gene may depend on an *enhancer*, a DNA region that can be many bases away from the gene. This region, even if linearly distant from the gene, may be spatially close to it because of the way the chromatin is folded. Proteins known as transcription factors can bind to the enhancer and activate or inhibit the transcription of the gene.

Furthermore, the way DNA is coiled around histones in the chromatin can prevent its transcription by the RNA polymerase, so the process of decoiling and recoiling the DNA can serve as another gene expression control mechanism.

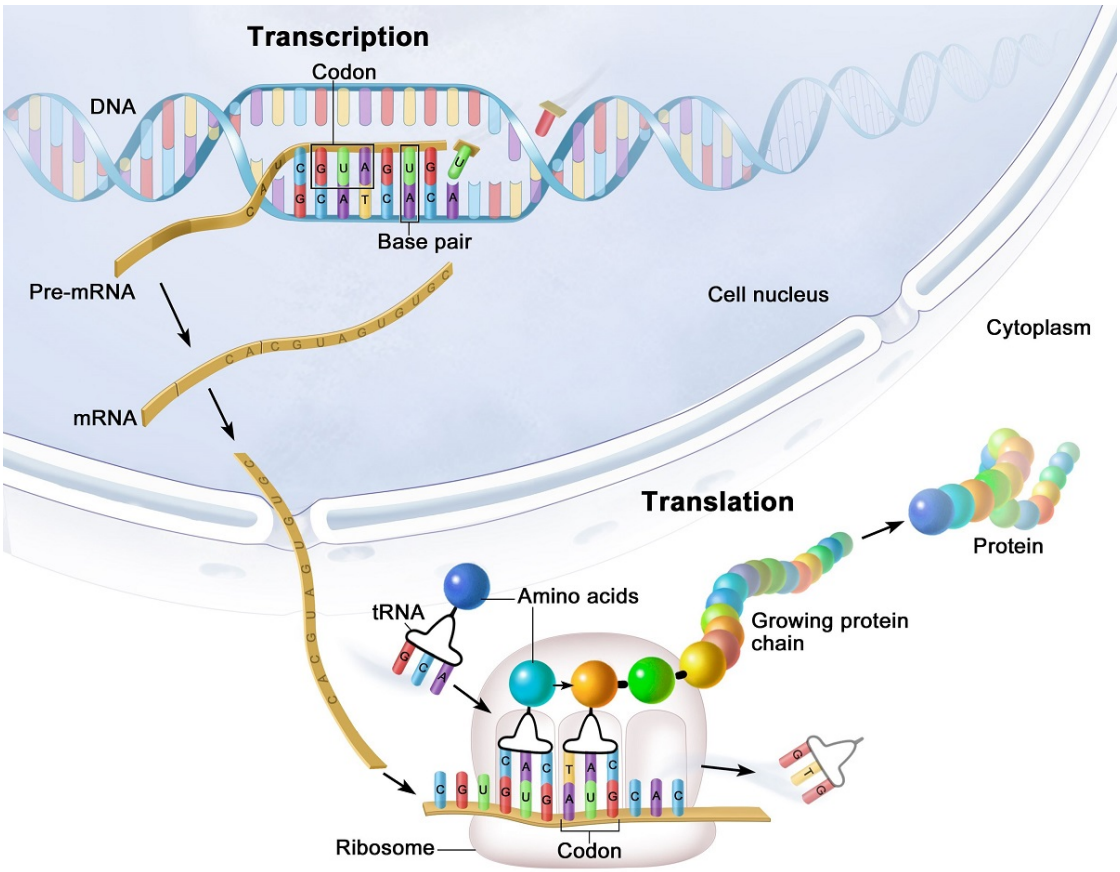


Figure 1.4: Overview of the transcription and translation processes in an eukaryotic cell. Image copyright Terese Winslow LLC, 2017. U.S. Govt. has certain rights.

1.2.3 DNA mutations

A DNA mutation is a change of the genetic information in a cell: it can happen due to a variety of factors and have effects of different severity.

A *point* mutation is a change in just one base pair in a DNA molecule. Despite the limited impact, this kind of mutation can cause major problems in the organism such as the sickle cell disease. In this disease, the point mutation occurs in the gene that codes for one of the subunits of hemoglobin, the protein that carries oxygen through the bloodstream. In the template strand of the gene, an adenine replaces the expected thymine which will code for uracil instead of adenine in the transcripted mRNA. This altered codon will, in turn, code for the wrong amino acid (valine instead of glutamic acid) and the resulting hemoglobin can be distorted in a sickle shape and clog small blood vessels [82].

In the example of the sickle cell disease, the point mutation is said *missense* because the altered codon codes for a different amino acid. If, instead, the altered codon specifies the same amino acid, the mutation is said *silent*. Missense mutations are not always catastrophic: the replaced base might not make its way into mRNA (because it is not in a gene or because that part of the transcript gets sliced out) and the altered protein might still work as intended. Furthermore, it is possible that a substitution of this nature could cause the corresponding mRNA codon to no longer code for an amino acid, but to instead become a stop codon. This is called a *nonsense* mutation. Instead of the ribosome translating the rest of the mRNA strand it will just stop entirely, resulting in a partially complete protein.

The insertion or deletion of a base pair from the DNA sequence can cause *frameshift* mutations. Since the codons of the transcribed mRNA are supposed to be translated as groups of three nucleotides, if one of these is suddenly added or removed, every single codon after the mutation will be altered, resulting in numerous missense mutations and, occasionally, a premature stop codon [94].

Spontaneous DNA mutations are caused by a fault of the replication machinery that goes undetected by the repair enzymes that scan the DNA in search of these errors. Mutagens, such as high energy UV radiations, can cause chemical reactions and change the structure of a region of DNA altering the normal genetic activity [94].

1.3 Sequencing and genomics

Sequencing DNA means determining the order of nucleotide bases (the As, Ts, Cs and Gs) in a DNA molecule. The base sequence can be used to study the kind of genetic information that is carried in a DNA segment.

Knowledge of DNA sequences has become indispensable in many applied fields including medical diagnosis [8, 17], biotechnology, population genetics [66] and virology [70, 15]. Comparing healthy and mutated DNA sequences can help diagnose different diseases and can guide patient treatment. Altering gene sequences through genetic engineering techniques can deliver genetically modified organisms, which can be used for the production of human proteins, such as insulin, and have desirable properties such an increased resistance to diseases, insects or herbicides [88].

Breakthrough discoveries and technological improvements developed in the last part of the previous century have enabled the emergence of genomics: an interdisciplinary field embracing biology, genetics and bioinformatics focusing on the study of whole genomes. It differs from “classical genetics” in that

it considers an organism's full complement of hereditary material, rather than one gene or one gene product at a time.

Genomics has revolutionized how genetic analysis is performed and has opened routes of investigation that were not conceivable just a few years ago. Most of the genetic analyses preceding the development of genomics considered a *forward* approach to analyse genetic and biological processes. That is, the analysis begins by first detecting mutations that affect some observable phenotype, and the characterization of these mutants eventually leads to the identification of the gene and the function of DNA, RNA, and protein sequences. In contrast, having the entire DNA sequences of an organism's genome allows geneticists to work in both directions: forward from phenotype to gene, and in reverse from gene to phenotype. Genome sequences revealed many genes that were not detected from classical mutational analysis and geneticists can now systematically study the roles of such formerly unidentified genes.

Genome sequencing would be easier if we could isolate each one of the 24 chromosomes of a single cell and read the entire base sequences in one shot, from telomere to telomere. Unfortunately, a sequencing machine that works in this way is not available. During a sequencing attempt, a large set of DNA reads is obtained from short fragments (usually just a few hundred bases) of DNA from different cells of the organism. These reads need to be assembled into a single sequence that uniquely represents the region of the genome which encloses the short fragments read by the sequencing machinery. This is known as *consensus sequence*.

The consensus sequence is built tying the adjacent short reads together, thus a reliable result is achievable only if a sufficient number of overlapping sequences are read. To this end, it is necessary to obtain multiple copies of the genome and to fragment it in segments by different offsets. Longer reads can provide more reliable information about the relative locations of specific base pairs, but they can be more expensive to obtain and more difficult to sequence.

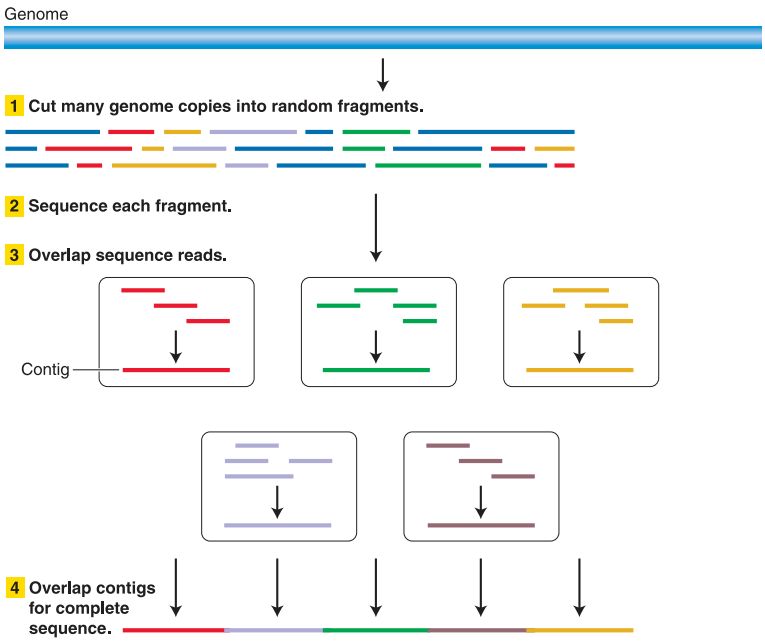


Figure 1.5: To obtain a genome sequence, multiple copies of the genome are cut into small pieces that are then sequenced. The resulting sequence reads are overlapped by matching identical sequences in different fragments until a consensus sequence is produced.

As with any experimental observation, automated sequencing machines do not always give perfectly accurate sequence reads. The error rate is not constant and so, to ensure accuracy, genome projects conventionally obtain multiple independent sequence reads of each base pair in a genome. A satisfactory *sequencing coverage depth* should be agreed, which is a measure of the average number of times that a specific genomic site is sequenced during a sequencing run.

Sanger sequencing [73], first developed in the 1970s, was one of the first sequencing methods to be commercialized, and it is still widely used today. Sanger sequencing is based on PCR (Polymerase Chain Reaction, a technique which exploits DNA polymerase to massively duplicate a target DNA sequence) and employs special nucleotide bases called ddNTPs that look like regular DNA bases but lack the 3'-OH group required for the phosphodiester bond formation. Polymerase can add these special bases to a growing strand of DNA, but once one is added the chain extension ceases. These chain-terminating bases are also labeled with dyes, a different one for each base.

Sanger sequencing uses a large amount of regular nucleotides and a small amount of chain-terminating nucleotides. Since incorporation of the special bases is random, we will have DNA fragments of various lengths with the special base at one end. At the end of the process, the DNA polymerase will have produced a pool of fragments of different lengths: the base sequence of each one is a prefix of the base sequence of the target DNA used as the template, the one we want to sequence.

The fragments are then separated and sorted by size (and so length) through gel electrophoresis. A special sensor can detect the dyed base at the end of each fragment and, by having at least one fragment for each length, we can discern the complete base sequence of the DNA we want to sequence. Figure 1.6 highlights some of the steps performed during Sanger sequencing.

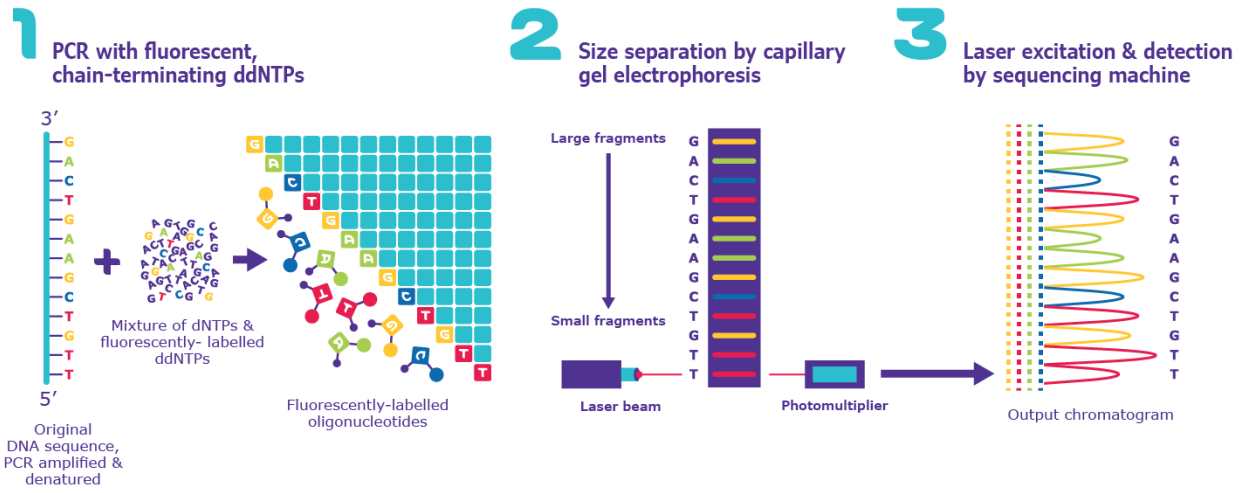


Figure 1.6: Overview of Sanger sequencing. In the first step, the DNA fragment to sequence is PCR amplified using chain-terminating nucleotides (ddNTPs), fragments of different lengths are obtained. In the second step, the fragments are separated through gel electrophoresis. In the third step, a laser is used to detect the dyes of the ddNTPs. Image copyright Sigma-Aldrich, Inc., 2021. All rights reserved.

1.3.1 Modern sequencing techniques

In the late 1980s, automated DNA sequencers were invented. These instruments automated the Sanger sequencing method and enabled the completion of the Human Genome Project in 2001 [2]. The completion of the project stimulated the development of cheaper, faster and more accurate platforms known as Next Generation Sequencers (NGS). Next Generation Sequencing machines have increased the rate and throughput of DNA sequencing substantially, compared to previous methods.

Modern equipment can have a throughput more than 1000 times higher compared to the sequencers used in the Human Genome Project, this has led to the ability to sequence an entire human genome in as little as one day, as opposed to the decade it originally required [83]. Furthermore, the rapid evolution

of sequencing technologies and the growing demand for low-cost equipment has led to an impressive decrease of DNA sequencing cost and to the era of the *"$1000 genome"* and personalized medicine [16], as emphasized by figure 1.7.

Illumina sequencing technology

Illumina Inc. is one of the major developers and producers of high-throughput sequencers. Illumina sequencing utilizes a different approach from the classic Sanger chain-termination method: the addition of each labeled nucleotide is tracked as the DNA chain is copied in discrete steps. The sequencing protocol is referred to as *Sequencing by Synthesis* [33].

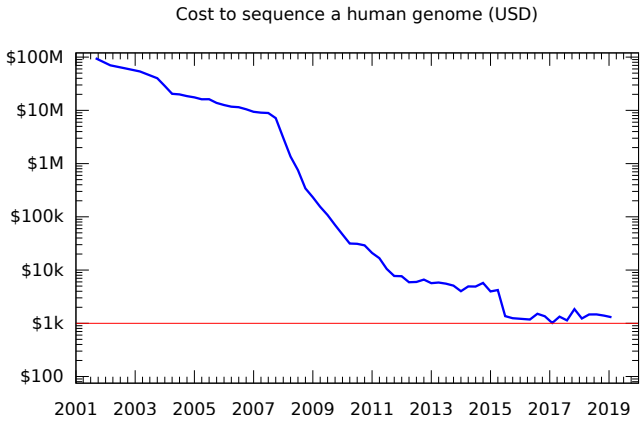


Figure 1.7: The cost for sequencing a human genome decreased substantially thanks to the innovations introduced in the sequencing platforms. Data as of August 2019. Data from National Human Genome Research Institute.

First, the sequencing library is prepared by adding short sequences, called *sequencing adaptors*, to the ends of the DNA fragments to sequence. Then, each fragment is amplified in a structure called *flow cell* composed by numerous lanes, each coated with two types of *oligos*, molecules complementary to the sequencing adaptors. The complementarity causes the fragment to stick to the surface, the polymerase duplicates the fragment which is then bent and stuck to the other oligo. This process, known as *bridge amplification*, is repeated millions of times in each lane and results in clusters of identical DNA fragments on the flow cell.

The nucleotides available to the polymerase are fluorescently tagged: during the duplication of a fragment, at each step, only one is incorporated based on the sequence of the template. After the addition of each nucleotide, the clusters of fragments are excited by a light source and the characteristic fluorescent signal is emitted. The number of steps performed determines the length of the read. The emission wavelength, along with the signal intensity, determines the base read. For each cluster, all identical strands are read simultaneously. The process is carried out, at the same time, for millions of fragments that bind in different spots of the flow cell.

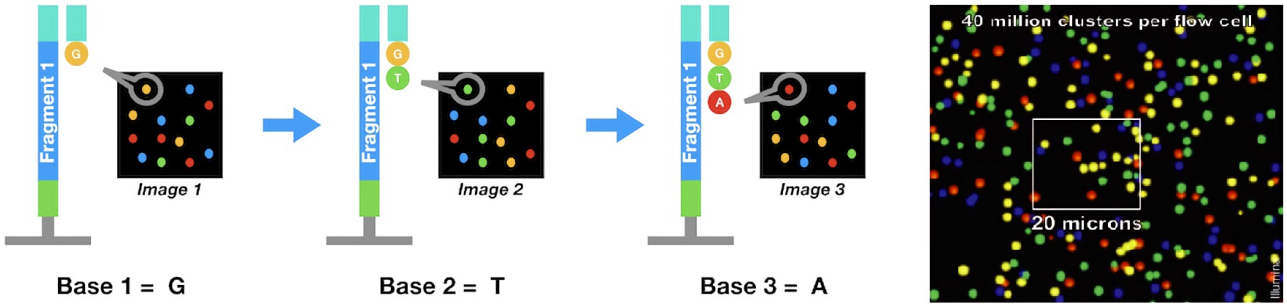


Figure 1.8: In "Sequencing by Synthesis", the type of the nucleotide added to the chain is determined by the signal it emits. The image on the right is a snapshot from the sequencer's sensor and represents a small fraction of the flow cell. Left image copyright Allison Zhang, Stanford University. Right image copyright Illumina Inc.

1.3.2 RNA-Seq

RNA-sequencing is a recently developed approach to transcriptome profiling (the measure of the quantity of RNA transcripts in a cell) that uses Next Generation Sequencing technologies [93]. The ability to understand the transcriptome is essential to interpret the functional elements of the genome, reveal the molecular constituents of cells and tissues and understand their development and the associated diseases. RNA-Seq allows analyses aimed at: cataloguing gene functions (understanding its phenotype, the observable traits in the organism); determining the transcriptional structure of genes, in terms of their start sites, 5' and 3' ends, splicing patterns and other post-transcriptional modifications; quantifying the changing expression levels of each transcript during development and under different conditions.

Identification and quantification of differentially expressed genes between normal and mutant samples are among the most common applications for RNA-Seq. Scientists can validate their hypothesis about the genetic causes for a mutation or get insights about the pathways that are enriched in either the normal or the mutant gene set.

Three steps are required in order to perform an RNA-Seq analysis. First, a sequencing library is prepared: RNA molecules are isolated and split in short fragments to be compatible with the employed sequencing machinery. In most applications, the rRNA is removed because it represents over the 90% of the RNA in the cell and, if kept, would drown out other data in the transcriptome. The RNA molecules are, usually, converted in double-stranded DNA (called *cDNA*, complementary DNA) through reverse transcription, as DNA is more stable and allows the usage of the more mature DNA sequencing technology. The cDNA is then sequenced and the base sequence of each fragment is produced.

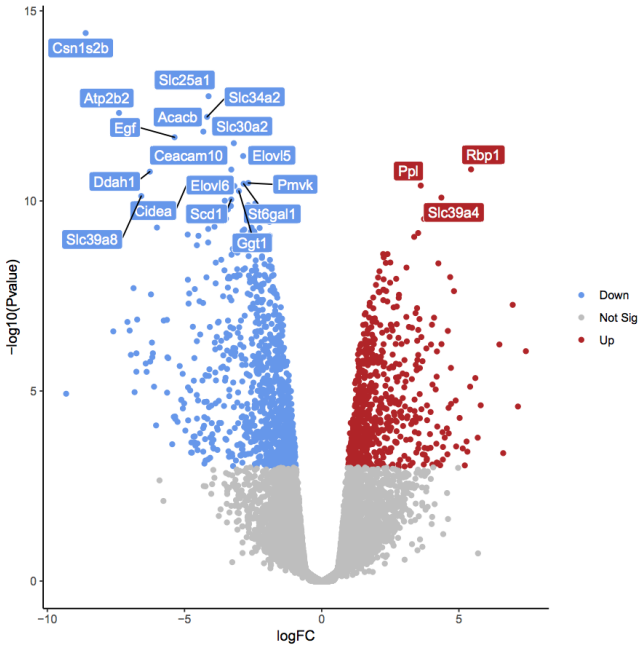


Figure 1.9: A graphical representation commonly used to illustrate the result of a RNA-Seq analysis. In this example, each point represents a gene: the red genes are more expressed in the mutant cells compared to the normal ones, while the blue ones are less expressed. The position of the gene on the Y axis indicates the confidence of the analysis (p-value).
Image copyright Galaxy Project.

The quality of the reads must be checked to remove unreliable data. Fortunately, modern sequencers include, for each base read, a confidence value which is an estimate of the error rate of the base call. Once high quality reads have been obtained, the first task of data analysis is to map the short reads to the reference genome. This process is known as *alignment* and the development of faster and more accurate algorithms for this task is an active research area in computer science [76]. The alignment process should tolerate discrepancies such as the substitution, insertion or deletion of single bases or of entire gene regions caused by the unavoidable small differences in the genome of different individuals.

The third step is the analysis of the gathered data, which begins by counting the transcript amount of each gene in all the samples analysed. The counts are then normalized to account for the different

amounts of reads in the samples. Normalization allows us to compare the expression of each gene between the samples and to observe if certain pathways are enriched in either the normal or the mutant cells. As with read alignment, differential expression analysis is far from a simple and objective task: counting and assigning the transcript of each gene can be performed in multiple ways and different normalization methods can be applied, which may influence the drawn conclusions.

1.3.3 ChIP-Seq

ChIP-sequencing is a method used to identify locations in the genome linked by proteins and can be used to precisely map binding sites for any protein of interest [71]. The knowledge of protein interactions with DNA is essential for fully understanding many biological processes such as gene expression regulation.

The workflow of ChIP-Seq combines chromatin immunoprecipitation (ChIP) with high-throughput sequencing. The first step in the ChIP protocol requires the establishment of a linkage between DNA and the proteins attached to its strands using chemical agents in a process called *DNA crosslinking*. Formaldehyde is used as an agent and large amounts of DNA are employed.

In the second step of the protocol, the chromatin is fragmented in order to get high quality DNA pieces for the forthcoming sequencing. Short DNA fragments are obtained: these are glued to the protein normally linked to them, thanks to the employed chemical agents that avoid the detachment of these proteins during the handling processes.

The third step is known as chromatin immunoprecipitation and gives the name to the whole protocol. An antibody specific to the protein of interest is chosen and attached to small magnetic beads. The antibody binds only to its relevant protein and the DNA fragment linked to the protein of interest will, in turn, attach to the beads. The beads are then fetched through magnetism and the rest of DNA fragments are washed away.

The recovered DNA fragments are cleaned from the histones and the other proteins, reversing the effect of the formaldehyde. As usual, the fragments are then amplified through PCR and sequenced. The short reads are aligned to the reference genome and the binding levels in the genome regions are quantified.

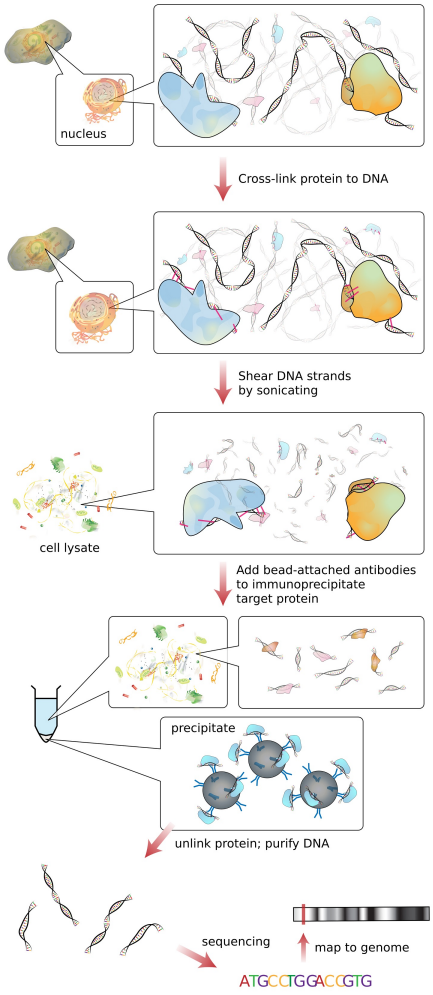


Figure 1.10: Graphical representation of the main steps in a ChIP-Seq experiment. Image copyright wikimedia user Jkwchui, licensed under CC license.

ChIP-Seq enables fast analysis, however, a quality control must be performed to make sure that the results obtained are reliable [13]. A control sample is necessary to verify that a high concentration of reads in the ChIP-Seq experiment is due to a protein binding site and not to some biases such as the fact that a lot of reads may get mapped to a repetitive region.

Two types of controls are often used in ChIP-Seq studies. A control *input* DNA sample is one which

has been crosslinked and fragmented but not immunoprecipitated, this means that no antibodies are used. A control *IgG* DNA sample uses an antibody that will not bind to nuclear proteins. Any IgG immunoprecipitated DNA should represent a random, nonspecific population. Because IgG ChIPs often produce relatively little amplifiable DNA, input controls are more widely used to normalize signals from ChIP enrichment [55].

2

Bioinformatic data analysis

Bioinformatics is an interdisciplinary scientific field that combines concepts and approaches from biology, statistics and computer science in order to develop and employ methods and tools for understanding biological data.

The growing demand for nucleic acid and protein sequencing, along with the increasing number of government initiatives and funding and the expanding use of bioinformatics in drug discovery and biomarker development [53], has led to an increase in the amount of biological and genetical data produced and to a growing need of analysis and interpretation methods.

Both the completion of the Human Genome Project and, more recently, the development and spread of Next Generation sequencers have spurred the evolution of algorithms and data analysis methods. Further evidence of this expansion is the 15-fold increase in bioinformatics publications in the period 2000 – 2003, reported by the journal *Briefings in Bioinformatics* [10].

Modern bioinformatics tasks stretch from the development and deployment of efficient algorithms and programs to process data, such as in sequence alignment, to the study of statistical models to accurately describe and predict the behavior of a phenomenon, such as the prediction of the differentially expressed genes between samples, to the production of genomic annotations for genes or regulatory sequences.

More broadly, most bioinformatics problems can be assigned to one of the following categories [34]. *Assembly* means establishing the nucleotide sequence of a genome, with techniques such as those described in section 1.3. Since the completion of the Human Genome Project, the genomes of thousands of species have been assembled but those of millions of other species still remain unknown [34]. Through *quantification* studies, the scientists aim at measuring the functional characteristics of a cell. These studies make use of modern sequencing protocols to determine the relative abundances of DNA fragments. Various protocols based on DNA sequencing have been described to gain insights on gene expression (such as in RNA-Seq) and in protein binding (such as in ChIP-Seq). *Resequencing* tasks focus on discovering the differences between the genome of an organism and the reference genome of its species. These differences can help understand the genetic causes of a particular phenotype, such as the emergence of a disease or a difference in the survival rate.

2.1 Data formats and specifications

2.1.1 Main data formats

Different data formats and specifications have been designed to describe different types of biological information, prioritizing storage efficiency and readability.

FASTA

FASTA[1] is a text-based format used to represent nucleotide or amino acid sequences. A FASTA file can encode more sequences and each one is preceded by a comment line that describes it. Each element of the sequence is encoded by a character: in nucleotide sequences the characters `A`, `C`, `G`, `T`, `U` are used to encode the nucleotides while other characters are used to represent ambiguity: for instance, `R` is used to represent any nucleotide with a purine base.

FASTA is the de facto standard to represent sequence information and it is the most common format used to encode genome assembly.

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Code 2.1: A wrapped line from a FASTA file describing an amino acid sequence.

FASTQ

FASTQ[2] is a data format commonly used to encode the content of sequencing reads. Similarly to FASTA, the format is text-based: every nucleotide or amino acid is encoded by a character.

Every sequencing read is represented by four sections:

- a header line starting with the character `@`, this can include the ID of the read or other sequencing information;
- the characters encoding the read;
- a line containing the character `+` used as a separator;
- the quality scores of the sequence encoded in the second section.

The format specification requires the assignment of a quality score to each base: the scores are measures of reliability of the base call. Each score is encoded by a single character and its value ranges between 0 and 40. The lower the number, the lower the accuracy and the probability of the base call being correct. These values Q are known as *Phred quality scores* and the probability p of the error of

[1] FASTA format specifications: `https://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml`
[2] FASTQ format specifications: `http://maq.sourceforge.net/fastq.shtml`

the base call is defined as $Q = -10\,log_{10}p$. For instance, a score 10 means a 10% probability of error, while 36 means a 0.02% one. Code 2.2 shows an example of the format.

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

Code 2.2: An example of the four sections comprising a read in a FASTQ file.

SAM and BAM

SAM[3] stands for "Sequence Alignment Map" and is a text-based format used to represent reads aligned to a reference sequence. SAM and BAM (its binary equivalent) are the standard formats used by the sequence aligners to represent the result of reads alignment to a reference genome.

In the SAM format, the information about an aligned read is grouped together and characterized by a header containing sequencing details, descriptions, checksums and other metadata, and by an alignment section. Each line in the alignment section describes the alignment of a read segment to the reference sequence.

As can be seen in code 2.4, the first and third columns identify the segment within the read, the fourth specifies the leftmost mapping position with respect to the reference sequence and the sixth one uses the CIGAR notation to encode the proper alignment information. A CIGAR string is a sequence of base lengths and associated operations that allows the representation of partial alignments. It is used to indicate which bases align (either a match/mismatch) with the reference sequence, which ones are not present in the aligned segment or, vice versa, are present in the segment but not in the reference sequence.

In the example code 2.3 and 2.4 below, the first aligned segment starts in position 7 and is characterized by the CIGAR string `8M2I4M1D3M`: the first 8 bases of the segment match the 8 bases following position 7 in the reference sequence (`8M`), the following 2 bases of the segment are not found in the reference sequence (`2I`), the next 4 characters match (`4M`), the next base in the reference sequence is not present in the segment (`1D`) and the last 3 bases match (`3M`).

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

Code 2.3: Representation of some example reads aligned to a region of the genome.

[3]SAM format specifications: `https://samtools.github.io/hts-specs/SAMv1.pdf`

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Code 2.4: SAM representation of the aligned reads shown in code 2.3.

BED, GTF

BED, GTF and other variants are text-based formats used to describe features such as genomic annotations and intervals. In these formats each line describes a feature. The BED format[4] requires 3 fields: the name of the chromosome and the start and end positions of the feature. Optionally, other fields can be used: the fourth is intended for the name of the feature, the fifth for a score value (used in case of graphical representation of the file) and the sixth for the strand on which the feature is located.

```
chr7  127471196  127472363  Pos1  0  +
chr7  127472363  127473530  Pos2  0  +
chr7  127473530  127474697  Pos3  0  +
```

Code 2.5: An example of a BED file with six columns.

The GTF format[5] is commonly used to represent gene annotations, and it specifies 9 fields. Most of the BED fields are also present in the GTF format, such as the chromosome name, the start and end position of the feature and its name, strand and score. The third field is intended for the feature type, such as "gene", "exon" and "mRNA" and the ninth field is a list that includes all the other information pertaining to the feature such as the gene and transcript name and ID.

The information included in a GTF annotation file are useful for programs such as "featureCounts" that aim to quantify the amount of fragments belonging to all the transcripts of a certain gene. Code 2.6 shows an example of this file format.

```
chr1  processed_transcript  exon  11869 12227 . + .
    gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; exon_number "1";
    gene_name "DDX11L1"; gene_biotype "pseudogene"; transcript_name "DDX11L1-002";
    exon_id "ENSE00002234944";
chr1  processed_transcript  exon  12613 12721 . + .
    gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; exon_number "2";
    gene_name "DDX11L1"; gene_biotype "pseudogene"; transcript_name "DDX11L1-002";
    exon_id "ENSE00003582793";
chr1  processed_transcript  exon  13221 14409 . + .
    gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; exon_number "3";
```

[4]BED format specifications: `https://www.ensembl.org/info/website/upload/bed.html`
[5]GTF format specifications: `http://www.ensembl.org/info/website/upload/gff.html`

```
gene_name "DDX11L1"; gene_biotype "pseudogene"; transcript_name "DDX11L1-002";
exon_id "ENSE00002312635";
```

Code 2.6: Three lines taken from a GTF file from the GENCODE gene annotation.

In the BED and GTF formats, the genomic coordinate system is the same for both strands: this means that the range [X,Y] on a chromosome refers to the region between the X and Y bases from the start of the forward strand (5'-to-3' direction), independently of the strand the region is located on [4].

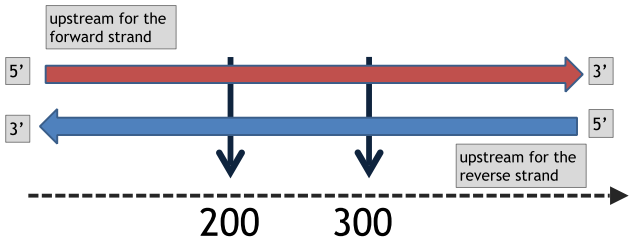


Figure 2.1: Coordinate system used in the BED/GTF formats. Image copyright István Albert, 2014.

2.1.2 Main data sources

Reference genome assemblies

First published in 2001, the human reference genome has been revised and improved many times towards the goal of its completion. As of writing, the reference genome is now in its 20th rendition. In the last two decades the reference has improved considerably: the very first version had about 150,000 gaps while the current one counts only about 250 of them [1].

The reference assembly is maintained by the Genome Reference Consortium[6] (GRC) which, as needed, releases updates integrating the new knowledge gathered since the release of the previous version. As of writing, the most commonly used releases are the 19th, known as GRCh37, released in 2009, and the 20th, known as GRCh38, released in 2013[7].

The ENCODE Project

The ENCODE project (Encyclopedia of DNA Elements) [20] was launched in 2003, shortly after the completion of the Human Genome Project, with the aim of identifying all the functional elements in the human genome. The project was organized by a worldwide consortium of research groups: the first phases of the project focused on the study and development of techniques, technologies and strategies to employ in the analysis of the entire genome in the successive phases.

Within the ENCODE project, the goal of building an exhaustive annotation for genes and genes' variants was assigned to the purposely established GENCODE consortium [31], while it was determined that the main goal of the ENCODE project was to determine the role of the remaining components of the genome, such as promoters and regulatory regions, whose functions were still unknown.

Since its establishment, the GENCODE consortium compiled over 37 releases[8] marking the location, biotype and transcripts of the genes in the human (and mouse) genome. For reference, the current release (version 37) reports 60,651 total genes, among which 19,951 protein-coding ones and 17,948 long non-coding RNA ones.

[6]Genome Reference consortium: `https://www.ncbi.nlm.nih.gov/grc`
[7]List of UCSC genome releases: `https://genome.ucsc.edu/FAQ/FAQreleases.html#release1`
[8]GENCODE releases: `https://www.gencodegenes.org/human/releases.html`

NCBI's GEO and SRA

GEO and SRA are two public data repositories from the American National Center for Biotechnology Information (NCBI). GEO[9] [6] archives and distributes high-throughput sequencing data collected and submitted by scientists. The main goals of GEO are providing a robust and reliable database to store genomic data, offering simple submission procedures to the researchers and providing user-friendly mechanisms to the users wanting to locate, query and download gene expression profiles of interest [59]. One reason to use such archives is that most journals require the publication of the experimental data in known repositories that offer fast and reliable access.

SRA[10] [44] is a public repository established to preserve public-domain sequencing data and to provide free, unrestricted and permanent access to it. Like GEO, SRA is designed to store Next-Generation sequencing data but, while GEO is intended for the processed sequence data files of an experiment such as aligned read files and count data, SRA contains only the raw sequence data files for that experiment, useful for repeating the analyses.

2.2 Common bioinformatics tasks and tools

2.2.1 Reads quality control

During sequencing, the nucleotide base sequence of the sheared chromatin is determined by the sequencer and, for each DNA fragment, a short sequence (called read) is generated and written to a file. At the end of the process, a FASTQ file is produced. Modern sequencing machinery can generate a massive number of reads in a single experiment but these can, occasionally, contain errors due to the technical limitations of the sequencing platforms. Before proceeding with the downstream analysis, it is therefore necessary to assess the quality of the reads.

Fortunately, modern sequencing platforms can estimate the error rate of each base call and note it in the produced FASTQ file, as described in section 2.1.1. With this information it is possible to evaluate the data quality.

FastQC [5] is a quality control visualization tool for FASTQ files: it produces graphs detailing useful metrics aimed at quality control, using the Phred scores of the base called. Among the information it can display, we can find the:

- *per base sequence quality*: a representation of the quality scores across all the reads within the sample, by nucleotide position. On the x-axis, there are the positions of the sequenced bases in the reads. For each position, a boxplot is drawn showing the distribution of the quality scores, across all the reads, assigned to the nucleotide in position x within its read. The y-axis shows the Phred quality scores: the higher the score, the better the base call. An example is shown in figure 2.2a.

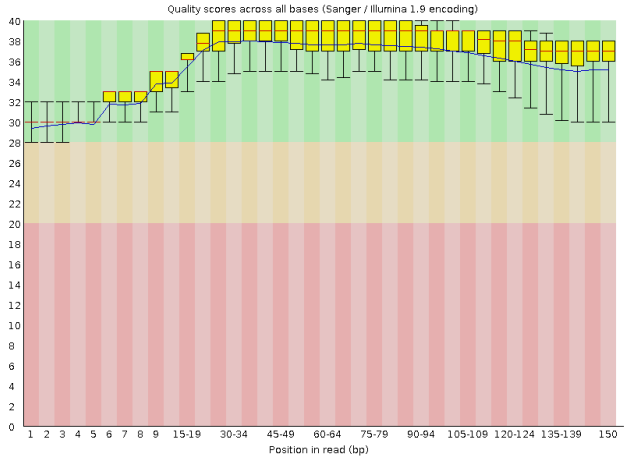
- *per sequence quality scores*: a plot of the average quality scores over the full length of the reads. On the x-axis, there are the average values of the quality scores, while the y-axis displays the total

[9] NCBI GEO: `https://www.ncbi.nlm.nih.gov/geo`
[10] NCBI SRA: `https://www.ncbi.nlm.nih.gov/sra`

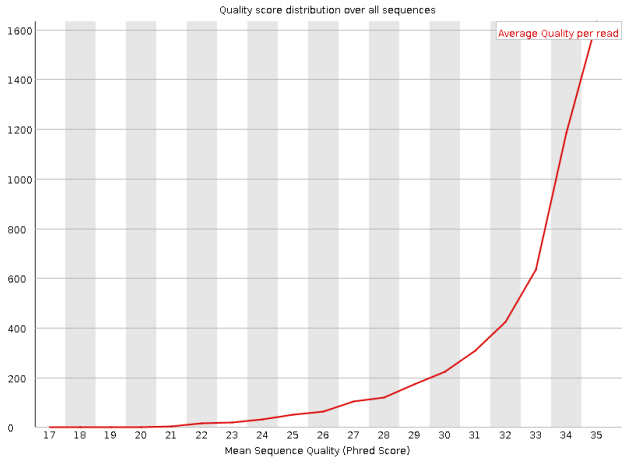
number of reads having, as average score, the value on the x-axis. The lower the tail of the plot, the better the average quality of the reads. An example is shown in figure 2.2b.

- *per base sequence content*: a plot of the relative base content along the length of all the reads. The proportion of each of the four bases should remain, more or less, constant over the length of the read and the content of A and T and of G and C should be the same. An example is shown in figure 2.2c.

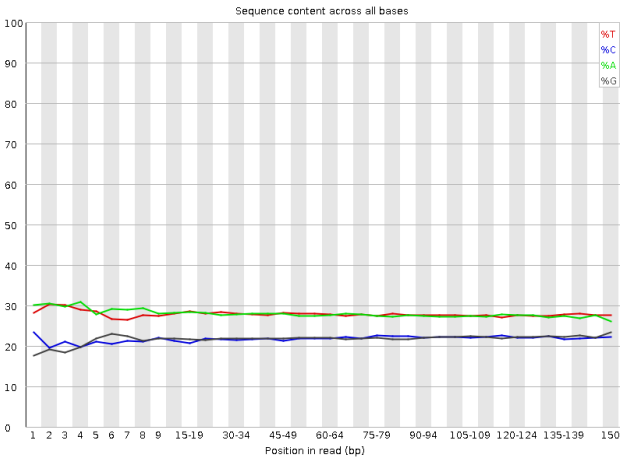
- *over-represented sequences and adapter content*: FastQC reports all the sequences that make up more than the 0.1% of the sequenced library. Finding that a single sequence is over-represented can indicate that the library is contaminated and further measures may need to be taken to correct the data. FastQC looks for matches between the over-represented sequences and the common contaminants such as residual adapters. An example is shown in figure 2.2d.



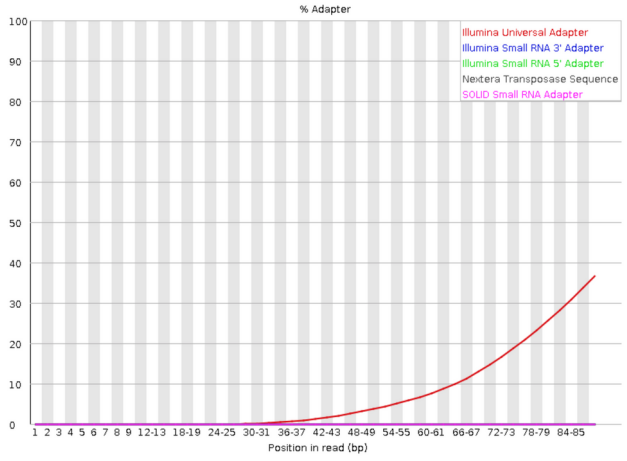
(a) per base sequence quality



(b) per sequence quality scores



(c) per base sequence content



(d) adapter content

Figure 2.2: An example depicting four graphs, produced by FastQC, representing statistics about sequencing reads. Note: the shown graphs may come from different FASTQ files.

In the case that the quality control plots, such as those produced by FastQC, highlight serious problems within the data, it can be necessary to take measures. As an example, in some early instruments the reliability of sequencing decreased sharply along the read. A fix for this problem can be the shortening

of the reads, removing the low quality base calls, in a process called *trimming*. A tool designed for this purpose is *Trimmomatic* [7].

2.2.2 Short reads alignment

After checking the sequenced reads for anomalies, it is often necessary to align said reads to a reference genome in order to reconstruct the library from the short reads and map the fragments to the genome, identifying the genes or the regulatory regions. Developing a fast and reliable aligner is no trivial task and some arbitrary choices may be needed: the reads are usually short and often affected by sequencing errors and could, thus, be aligned to multiple genomic regions.

Since the sequencing reads are often affected by errors, read aligners must be able to allow for limited discrepancies in form of substitutions, insertions and deletions of nucleotides. This is also necessary because the reference genome used to map the reads is, for most experiments, not the genome of the individual from which the reads come from, and the base sequence of some portions of the genome can vary between organisms within the same species.

Allowing such errors can lead to multiple possible ways to align a read to the genome. Most aligners use a matrix to compute the "score" of an alignment in order to select the "best" one, the one best supported by the picked score system. Other than using different rewards and penalty scores for matches, mismatches, opening and closing of gaps, insertions and deletions, many heuristics can be used such as lowering the penalty score for errors in low-confidence bases and optimizing the score system accounting for patterns and biases in the specific sequencing platforms [74].

BWA [46, 47, 45] and *bowtie2* [43] are commonly used modern short reads aligners able to align ten of thousand of reads per second against the 3 billion bases of the human genome. Even though these aligners are able to handle matching errors, as well as insertions and deletions of bases, they are not designed to align reads generated from RNA, such as those produced in RNA-Seq experiments. RNA reads are derived from mature mRNA which, in eukaryotes, is produced after splicing out the introns from the transcript. Most short reads aligners use the entire genome as a reference to align the reads: if a read consists of two (or more) contiguous exons while, in the reference genome, the same exons are interspersed with an intron, the aligner would match the first part of the read (the first exon) while the second one would not be properly aligned. In this case, the aligner would have to introduce a long gap in the mapping of a read to span the intron, penalizing the (genetically correct) alignment. This is not desired for DNA read mapping and might lead to false mappings.

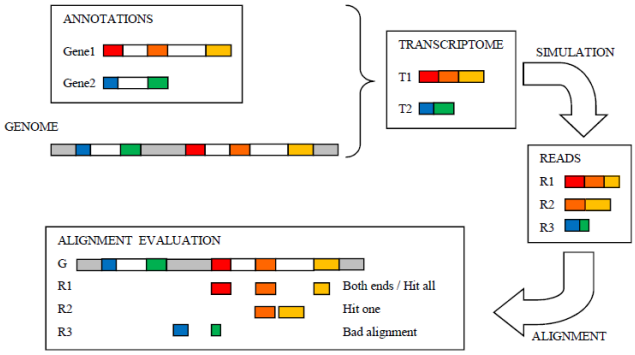


Figure 2.3: The alignment of RNA-Seq reads against a genome requires the awareness of exons and introns. Image from Krizanovic K, et al. [42].

To avoid such problem, one can consider the transcriptome (the set of all RNA transcripts, including coding and non-coding) as a reference or employ a *splice-aware* aligner, such as *STAR* [19] or *HISAT2* [40]. The alignment of RNA-Seq reads against the transcriptome can be less reliable because the transcriptomes of alternatively-spliced organisms are both incomplete, since not all transcripts have

been identified, and redundant, since transcripts have multiple isoforms. A splice-aware aligner would know not to try to align RNA-Seq reads to introns, it would identify possible downstream exons and try to align to those instead, ignoring introns altogether.

2.2.3 Visualization and comparison of sequencing results and genomic data

Prior to proceeding with the downstream analyses of the aligned reads, it is good practice to visualize the data in order to spot anomalies and get a preliminary idea of how the experiment is going. *IGV*[11] (Integrative Genome Viewer) [87] is an interactive visualization tool that enables the visual exploration of genomic data. It supports a wide range of data formats, including SAM, BAM and BED, it comes preloaded with genic annotations and it supports multiple data tracks and the possibility to overlay them for a better comparison.

Among the variety of software for visualizing and comparing genomic data, *bamCompare*, from the *deepTools*[12] software collection [72], allows the comparison of BAM files while simultaneously accounting for differences in sequencing depth. To compare the files, the genome is partitioned in bins of custom, equal, size. Then, the number of reads found in each file is counted and a comparison function is applied. This process results in a comparison value for each genome bin. Among the available comparison functions there are the difference and the logarithm of the ratios.

Comparing and visualizing mapped reads can be useful to assess the outcome of two replicate sequencing experiments or to contrast the results of a ChIP sequencing with a control sample in order to avoid considering false positives.

Figure 2.4: An example session from IGV showing five data tracks from a ChIP-Seq experiment. The top two tracks depict the binding locations of a protein on the genome (experimental and control, respectively), the third is an overlap of the first two, the fourth was obtained computing, bin by bin, the ratio from the first two tracks and the last one is a genic annotation.

The *UCSC Genome Browser*[13] [37], developed by the University of California, Santa Cruz, is a highly configurable visualization tool for genomic features: it comes preloaded with annotations encompassing

[11] Integrative Genome Viewer: `https://software.broadinstitute.org/software/igv/`
[12] deepTools software collection: `https://deeptools.readthedocs.io/en/develop/`
[13] UCSC Genome Browser: `https://genome.ucsc.edu/`

known genes and transcripts, histone modifications and transcription factors binding sites and possible mutations. The preloaded annotations are displayed as horizontal tracks and it is possible to upload a BAM or BED file and display it alongside the built-in tracks.

Figure 2.5 illustrates some of the features within position 141,222,402 and 141,261,605 on chromosome 5 (considering the reference human genome GRCh37). The first track from the top annotates the histonic modification H3K27ac (acetylation of the lysine residue at position 27 on the histone H3 protein), which is associated with a higher activation of transcription and can suggest the presence of transcripted genes close by; the second track highlights the annotated genes, along with their exons and introns; the last track shows possible mutations in the region.

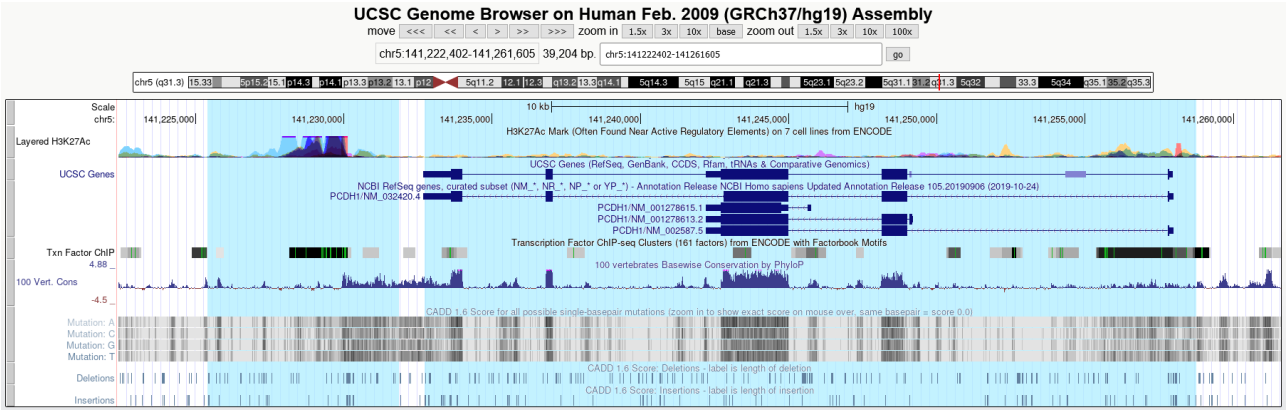


Figure 2.5: Genomic features highlighted by UCSC Genome Browser in the region `chr5:141222402-14126160` *of the human reference genome GRCh37.*

2.2.4 Reads quantification

A problem commonly encountered while studying sequencing data, especially during the RNA-Seq analysis workflow, is the quantification of the reads. The task consists in counting and assigning the mapped reads to genomic features, such as genes, exons or regulatory regions. The software tools used to perform this task should be specifically designed for genomic data, as awareness of the types of features is required for a correct quantification. For instance, if a read spans multiple exons of the same gene, the gene should be counted only once. These tools should, also, be flexible in the way the counts and assignments are performed: a read, in RNA-Seq data, spanning multiple genes should be discarded, as its origin is ambiguous and a transcripts originates from only one of the target genes. On the other hand, in ChIP-Seq experiments it is recommended that these reads are counted for each overlapping gene, because for example epigenetic modifications inferred from these reads may regulate the biological functions of all the overlapping genes [78].

An efficient tool specifically designed for sequencing data quantification is *featureCounts* [49] from the *SubRead*[14] [50] package. This program expects, as input, a set of $\mathcal{N}$ aligned reads files (SAM or BAM) and an annotation file (GTF) characterizing the $\mathcal{M}$ features of interest with their location on the genome and their type. It produces a matrix with $\mathcal{N}$ columns and (with meta-feature summarization disabled) $\mathcal{M}$ rows containing the counts assigned to the features.

featureCounts supports counting reads at feature level and at meta-feature level. Feature level means that each line in the provided annotation file is considered a feature and the reads are directly assigned

[14]Subread website: `http://subread.sourceforge.net/`

to them. In the meta-feature level, instead, the lines belonging to the same feature, such as the exons of the same gene, are grouped together and used for the reads' assignment. When counting reads at meta-feature level, a hit is called for a meta-feature if the read overlaps any component feature of the meta-feature. If a read hits a meta-feature, it is always counted once no matter how many features in the meta-feature this read overlaps with. For instance, an exon-spanning read overlapping more than one exon within the same gene only contributes one count to the gene.

As introduced above, featureCounts allows also to specify how to treat multi-mapping reads (reads that can be mapped to multiple regions in the reference genome) and multi-overlapping reads (reads that overlap multiple features or meta-features). Among the possibilities, it is possible to discard these reads, assign them to all the features or apply a fractional counting.

2.2.5 Data analysis platforms

Most of the software used in genetic data analysis can be run from command line, since it is available as a native executable or it is written in Python or R, and doesn't require the usage of graphical interfaces.

Some corporations invested in high-throughput sequencing, such as Illumina[15], provide proprietary data analysis platforms compatible with their equipment and data. These platforms offer many advantages: the availability of computational and storage resources off-site lightens the cost of acquiring and maintaining expensive in-house datacenters, graphical interfaces are provided to easily and intuitively apply data analysis tools and warranties can be expected with regard to tools' compatibility and interoperability.

Galaxy[16] [3] is an open-source web-based scientific data analysis platform, specifically designed to analyse large biomedical data sets. Galaxy incorporates all the most used tools and it is integrated with the public biomedical and genetics databases, allowing for direct import of data. Galaxy can be self-hosted on own hardware (such as in a private lab) or utilized for free on public instances. One such instance is `https://usegalaxy.org`, hosted by the Texas Advanced Computing Center.

Compared to using the data analysis tools directly from the command line, the usage of Galaxy can yield many advantages, among which:

- the availability of an extensive documentation, encompassing examples and use cases;
- the presence of an intuitive graphical interface;
- the option to define pipelines: sequences specifying the chained application of tools performing a complete data analysis, from the input reads to the output reports;
- the possibility to share such pipelines: since Galaxy maintains the information about the tools' version and parameters, the sharing of pipelines allows for a simpler reproducibility verification.

In some contexts, using the command line tools can be quicker and simpler: the public Galaxy instances can be slower than even an inexpensive PC (due to the sharing of the public infrastructure) and, for some tasks, being required to use graphical tools can be tedious and unnecessary.

[15] Illumina BaseSpace: `https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps.html`

[16] Galaxy project: `https://galaxyproject.org/`

2.3 RNA-Seq differential gene expression analysis

In RNA-Seq, sequencing is used to quantify the relative RNA abundances: through this measure of transcription, RNA-Seq data analysis aims at obtaining insights about the underlying studied natural process. Differential gene expression analysis is a process that involves the usage of statistical methods to discover quantitative changes in gene expression levels between multiple experimental groups. Such statistical methods are needed to establish whether the observed differences are due to the studied phenomenon or to random noise.

The RNA-Seq protocol turns the RNA produced by a cell into DNA (cDNA, complementary DNA) via a process known as reverse transcription. The resulting DNA is then sequenced, the reads are mapped against a reference sequence and assigned to genomic features (as described in section 2.2.4): from the quantification results we attempt to infer the original amounts of RNA in the cell.

When comparing two or more experimental groups for differential gene expression, it is recommended to use more replicates for each group (at least three), in order to account for slight expression variations in the expression of a gene. While it is technically possible to use only one replicate for experimental group, most gene expression analysis statistical methods need at least two samples in order to be able to discern a statistically significant expression difference from a random one.

2.3.1 Count normalization methods

When comparing gene counts between multiple samples and experiments, it is essential that the values are comparable to each other: for experiment A, we might have ended up with more material being placed into the sequencing instruments than for experiment B; the quantification methods need to be able to account for these differences. A variety of normalization methods have been proposed and are currently used but some are inappropriate for differential gene expression analysis.

RPKM (Reads Per Kilobase Million) is a count normalization method used to compare gene expression between genes in the same sample: it accounts for differences in gene length and in sequencing depth. In the simplest form, given N the number of reads mapped to a gene, C the total number of mapped reads in the sample and L the gene length in nucleotides, the normalized RPKM gene count K is shown in equation 2.1: we divide the total number of reads assigned to the gene by the total number of reads in the sample, and then we divide again by the gene length.

$$K = N \cdot C^{-1} \cdot L^{-1} \tag{2.1}$$

The name of the method comes from the fact that, in order to avoid dealing with small numbers, the inverse of the total number of mapped reads C is multiplied by 10^6 and the inverse of the gene length L is multiplied by 10^3, so the formula ends up being the one shown in equation 2.2. Table 2.1 shows an example of the application of this method.

$$K = 10^9 \cdot N \cdot C^{-1} \cdot L^{-1} \tag{2.2}$$

TPM (Transcripts Per Million) is a normalization method that, like RPKM, accounts for gene length and sequencing depth. The idea this method is based on is very similar to the one used in RPKM but

the steps are reversed. The number of reads mapped to a gene is first divided by the gene length and then it is divided again by the sum of the "length normalized" counts for all the genes in the same sample (accounting for sequencing depth normalization). As for RPKM, the value is multiplied by 10^6 to avoid small numbers.

$$K = N \cdot \frac{10^3}{L} \cdot 10^6 / \sum_i (N_i \cdot \frac{10^3}{L_i}) \tag{2.3}$$

Apart from dimensionality concerns (RPKM is expressed in nucleotides^{-1} instead of being adimensional), the main advantage of TPM is that the sum of the normalized counts in each compared sample is the same and, as such, it allows the comparison between samples of the same group. Table 2.2 shows an example of the application of this method.

		Input			Partial			Normalized		
Gene	Length	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
A	2 kb	10	12	30	285,714	266,667	283,019	142,857	133,334	141,510
B	4 kb	20	25	60	571,428	555,556	566,038	142,857	138,889	141,510
C	1 kb	5	8	15	142,857	177,778	141,409	142,857	177,778	141,409
D	10 kb	0	0	1	0	0	9434	0	0	943

Table 2.1: Application of the RPKM normalization, two steps.

		Input			Partial			Normalized		
Gene	Length	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
A	2 kb	10	12	30	5,000,000	6,000,000	15,000,000	333,333	296,296	332,594
B	4 kb	20	25	60	5,000,000	6,250,000	15,000,000	333,333	308,642	332,594
C	1 kb	5	8	15	5,000,000	8,000,000	15,000,000	333,333	395,062	332,594
D	10 kb	0	0	1	0	0	100,000	0	0	2217

Table 2.2: Application of the TPM normalization, two steps.

2.3.2 DESeq2

DESeq2 [52] is a statistical method for differential gene expression analysis of RNA-Seq data. As with any differential expression analysis method, the goal of DESeq2 is to compare the counts associated to genomic features in the samples of two or more experimental groups and estimate if this difference is due to randomness or to a difference in the transcription of the gene.

DESeq2 expects in input the table of the raw counts produced by a read quantification tool (such as featureCounts) and it applies a normalization suited for comparison between samples and between groups. The assumption that the input table contains raw counts is important as the method can discard lowly expressed genes during the hypothesis testing to reduce the number of false positives. A strong assumption of the method is that the large majority of the genes are not differentially expressed.

Count normalization

The count normalization method employed by DESeq2, known as "median of ratios", does not account for gene length (as this is only needed when comparing the expression of different genes in the same sample) and is more robust than RPKM and TPM, allowing for the comparison of the expression of a gene between different experimental groups.

To better understand why RPKM and TPM are not suitable for between-groups comparisons, consider the example shown in figure 2.6. If we were to divide each sample by the total number of counts to normalize, the counts would be greatly skewed by the DE gene, which takes up most of the counts

in sample A, but not in sample B. Most other genes in sample A would be divided by the larger number of total counts and appear to be less expressed than those same genes in sample B. A few highly differentially expressed genes between groups can skew the counts of all the other genes in the sample.

Raw counts

Gene	Sample A	Sample B	Ratio
X (15 kb)	30	20	1.50
Y (10 kb)	20	12	1.67
Z (18 kb)	20	12	1.67
DE (20 kb)	150	5	30.0

TPM normalized counts

Gene	Sample A	Sample B	Ratio
X (15 kb)	58,590	386,473	0.41
Y (10 kb)	158,590	347,826	0.46
Z (18 kb)	88,106	193,237	0.46
DE (20 kb)	594,714	72,463	8.2

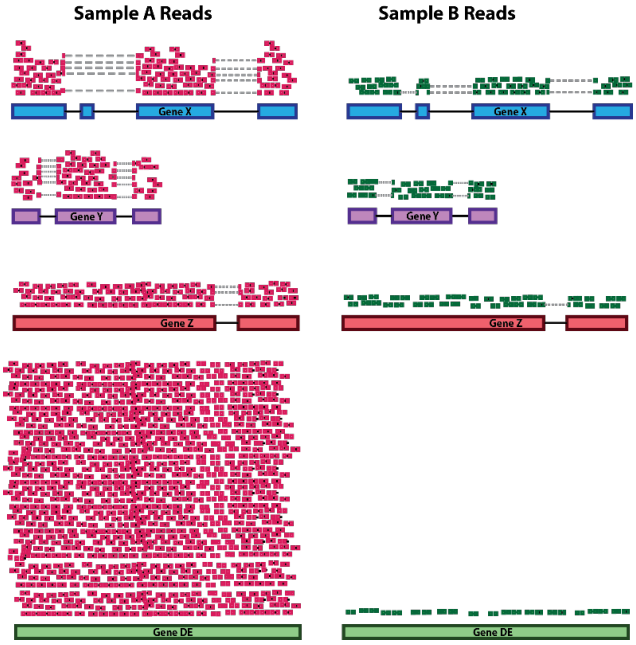


Figure 2.6: When using inappropriate normalization methods, the count of a DE gene can skew the counts of the other genes. In this example, the counts in sample A are skewed by the DE gene: the gene X has 1.5x the number of reads in sample A compared to sample B but the normalized count of the same gene is more than double in sample B compared to sample A. Image from Harvard Chan Bioinformatics Core (HBC).

Given a raw counts table (in which the rows and the columns represent the genes and the samples, respectively), the "median of ratios" normalization method employed by DESeq2 works as follows:

1. the logarithm (in base e) of each value is computed;
2. the average of each row is computed;
3. the rows with $-\infty$ average are ignored, these are the genes with 0 in at least one sample;
4. the computed row average is subtracted from the log values in the table;
5. the median of each column (after the subtraction) is computed: elevating e to this value gives us the *size factor* for the sample;
6. the final normalized values are obtained by dividing the initial values by the size factor of their sample.

Statistical model

DESeq2 models the raw counts in the samples with a statistical model: the number K_{ij} of reads of the gene i in the sample j is modeled to follow a negative binomial distribution with mean μ_{ij} and dispersion α_i. The mean μ_{ij} is defined as the product between the normalized count value q_{ij} and the size factor s_j: both these values are estimated during the normalization phase.

$$K_{ij} \sim NB(\text{mean: } \mu_{ij}, \text{dispersion: } \alpha_i)$$

$$\mu_{ij} = q_{ij} \cdot s_j$$

The value of the dispersion parameter α_i is estimated after the normalization and can be interpreted as a measure of variability between the expression of the gene i among the samples of the same experimental group, normalized with respect to the average expression level of the gene.

Once both the mean and the dispersion parameters have been estimated, a linear model is defined for each count value q_{ij}.

$$log_2\, q_{ij} = \sum_r x_{jr}\beta_{ir}$$

The β_{ir} parameters specify the fold-change (in logarithmic scale) of the gene i between experimental groups, and are the ones we are ultimately interested in. The model allows for the comparison of more than two experimental groups. During the estimation of the β_{ir} parameters, a statistical test is performed using the null hypothesis $\beta_{ir} = 0$, which indicates the absence of difference in the expression of the gene i between the considered groups.

Estimation of the dispersion parameter

The dispersion parameter α_i accounts for differences in gene expression within the same sample and is defined as $Var(K_{ij}) = \mu_{ij} + \alpha_i\mu_{ij}^2$, with $Var(K_{ij})$ and μ_{ij} variance and mean of the random variable K_{ij}, respectively. The dispersion parameter is used to model the variability, instead of other estimators such as the variance, because, for genes with moderate to high count values, the square root of the dispersion will approximate the coefficient of variation (Var/μ) and it has been observed that, in RNA-Seq count data, the variance increases with the mean expression in, more or less, a linear relationship and the dispersion will remain, more or less, constant [14]. Furthermore, it has been observed that the linear relationship between variance and mean is stronger for genes with high mean expression levels, as figure 2.7 depicts.

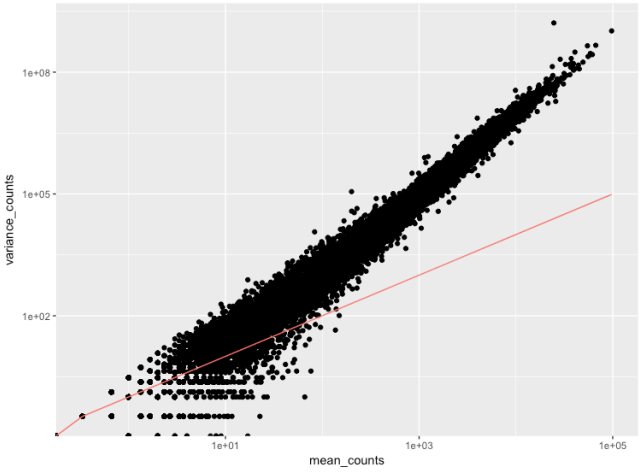


Figure 2.7: RNA-Seq data show a linear relationship between mean and variance within samples of the same group. The relationship appears stronger for genes with higher expression levels.
Image from Harvard Chan Bioinformatics Core (HBC).

Figure 2.8 highlights the relationship between the gene's mean expression value and dispersion, black dots represent single genes. In genes with high levels of mean expression the dispersion remains constant while in genes with low expression levels the variance is greater and so is the dispersion estimate. After estimating the dispersion of the genes, a curve of best fit is computed (in red in the figure). Often only a few replicates for experimental group are available and, as a consequence, the dispersion estimates are inaccurate. To deal with this problem, DESeq2 corrects the estimate for a gene using the available data from the others, under the assumption that genes having similar expression values tend to have similar dispersion.

This correction is implemented through the *shrinkage*: the dispersion estimate of the genes is moved towards the red line representing the trend in the dataset. The magnitude of the correction applied to a gene depends on its dispersion estimate (estimates far from the trend lead to a stronger correction) and on its information content: genes having low expression levels and so inaccurate variance and dispersion estimates are subject to a stronger correction, as are genes for which only a low number of samples is available.

The shrinkage is applied both when the dispersion estimate is above and below the curve. By lowering the dispersion estimate of a gene above the curve, the variability introduced by the experimental setting (i.e. not due to biological aspects) is reduced. The shrinkage is applied also to the genes below the curve, increasing their dispersione estimate, to reduce the number of false positives due to counts too consistent within samples. The only genes that are not subject to shrinkage are those considerably above the curve, distant from the general trend: these genes are considered outliers and their high variability is assumed to be due to gene-specific factors. In figure 2.9 the outliers are circled blue.

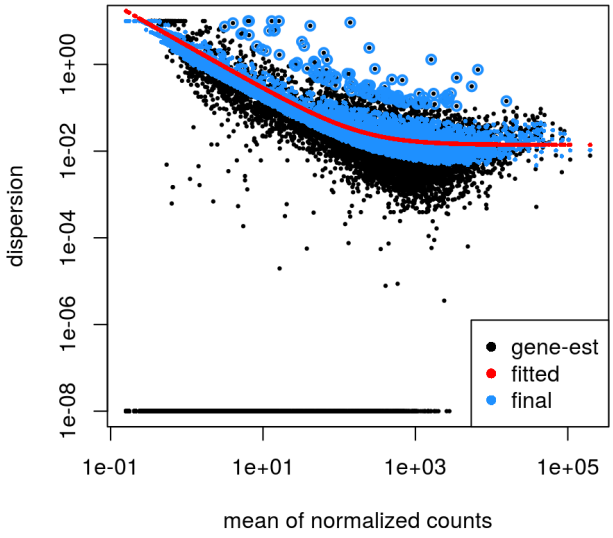


Figure 2.8: In RNA-Seq data, the dispersion estimate (proportional to the ratio between variance and mean) tends to be constant for medium to high count values. DESeq2 assumes that genes with same mean expression have same dispersion: a curve is fitted to the trend in data and the dispersion of the genes is corrected towards this trend.
Image from Analyzing RNA-seq data with DESeq2. Michael I Love, et al.

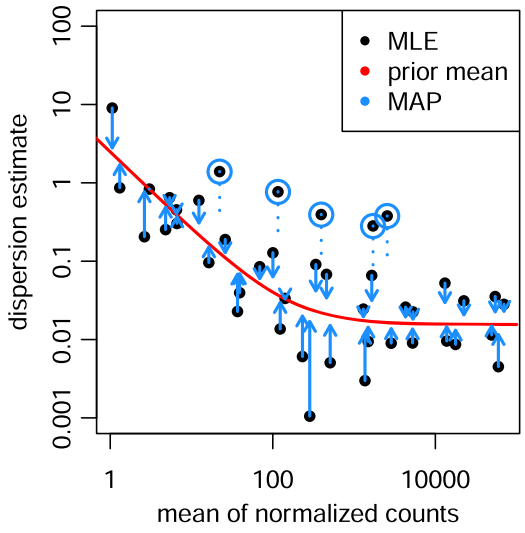


Figure 2.9: Both the genes above and below the fitted curve are corrected by applying the shrinkage. The outliers (circled dots) are ignored.
Image from Love, Michael I, et al. [52].

Fold-change correction

The DESeq2 developers experimentally observed that genes characterized by low mean counts tend to have high fold-change estimates due to noise and unreliable counts [52]. To handle this problem, the logFoldChange estimates β are reduced for genes characterized by low information content, such as those with low mean expression or high variance between replicates.

DESeq2 assumes that the logFoldChange distribution across all genes follows a zero-centered normal distribution, this distribution is used as a model to correct the logFoldChange estimates. Figure 2.10 shows the result of the application of the *LFC shrinkage* to an example data set. The left part graphs the logFoldChange estimates without correction: there are numerous genes with low mean expression

and high logFoldChange estimate; the right part shows the LFC estimates after the correction has been applied: the low mean expression genes have seen their LFC estimate noticeably reduced.

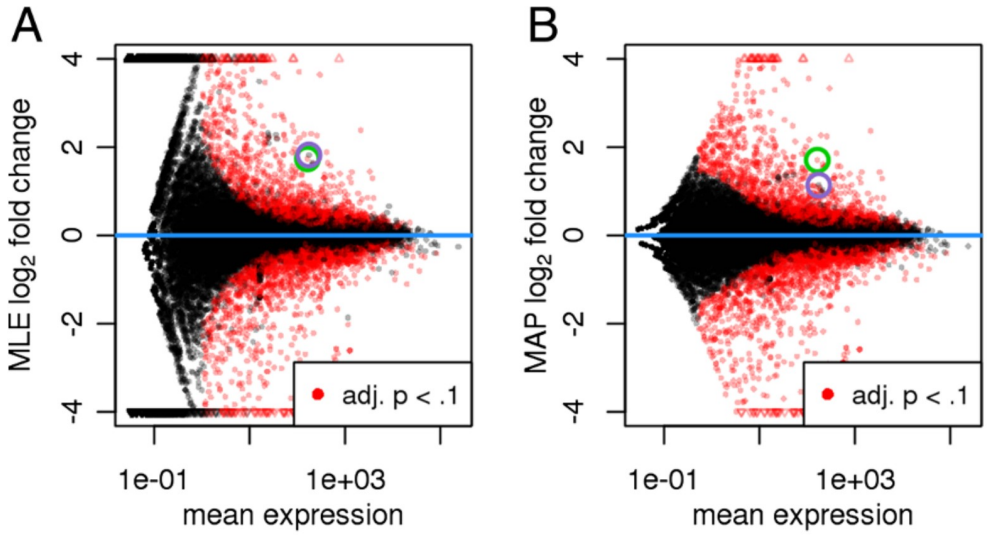


Figure 2.10: Effect of the application of logarithmic fold-change estimates shrinkage to an example data set. Left (A): LFC estimates pre-shrinkage. Right (B): LFC estimates post-shrinkage. Image from Love, Michael I, et al. [52].

Figure 2.11 illustrates how differences in dispersion among genes with similar mean expression levels influence the fold-change correction. In the image on the left, the green gene presents low variability among replicates (the counts have similar values) while the purple one expresses higher variability (`C57BL` and `DBA` are different experimental groups). On the right, we can see that the two genes have similar LFC pre-correction estimates (just under 2.0) but different ones post-correction.

This type of correction makes sense only when searching for differentially expressed genes (null hypothesis $\beta = 0$). When searching for genes that do *not* manifest differences in expression, applying this correction will greatly increase the number of false positives.

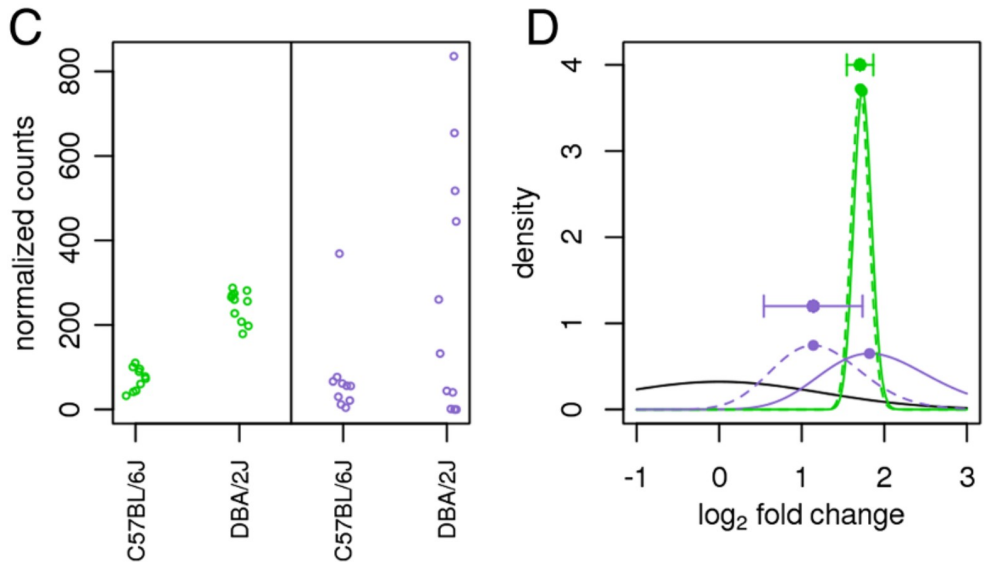


Figure 2.11: Effect of the application of logarithmic fold-change estimates shrinkage to an example data set. Left (C): Normalized counts for two genes (green and purple) in two experimental groups (`C57BL` and `DBA`) having 10 replicates each. The purple gene presents higher variability within group. Right (D): Pre (solid line) and post (dashed line) application of the LFC correction for the two genes. The green gene's LFC estimate is subject to very little correction because of its higher information content due to its lower count variability between samples. Image from Love, Michael I, et al. [52].

2.4 ChIP-Seq binding sites estimation

ChIP-Seq, introduced in section 1.3.3, is a modern method based on Next-Generation sequencing technology for the study of genome-wide protein binding sites and is commonly used in studies researching transcription factors and histone modifications [65, 63]. The reads produced by the sequencers, once aligned to the reference genome, must be studied to estimate the location of the binding sites of the protein of interest. Over the years, many algorithms and statistical methods have been developed [95, 58, 61] to aid the researchers with this task.

2.4.1 MACS: Model-based analysis of ChIP-Seq

MACS [101] is a tool for automatic binding site estimation designed to deal with some of the shortcomings which characterize ChIP-Seq data. As sequencing reads represent only the ends of the ChIP fragments, the precise estimation of the protein-DNA binding site can be challenging. Furthermore, ChIP-Seq data exhibit regional biases along the genome due to sequencing and mapping biases and chromatin structure [101]. These biases could be modeled if matching control samples are considered.

Briefly, MACS tries to recover the location and size of the DNA fragment from the short read and estimates which locations are significantly enriched, considering the control sample if available, using a statistical model.

Reads' position adjustment

Sequencing reads often represent only the 5'-to-3' end of the fragments in the DNA library, due to how sequencing machinery works, so the coordinates of the reads should be shifted in 3' direction to better represent the precise protein-DNA interaction site.

Once sonificated, ChIP-DNA fragments are equally likely to be sequenced from both ends [101] (i.e. they come from both strands). This means that around a true protein-binding site we should see two enrichment patterns: the reads on the *sense* strand (in blue in figure 2.12) should be enriched upstream of the binding site and the ones on the *antisense* strand (in red) should be enriched downstream.

MACS takes advantage of this bimodal pattern to empirically model the fragment size (and so the shifting length) to better locate the binding sites.

To estimate fragment sizes, MACS first slides a window with a width of roughly twice the size of the sheared chromatin to identify regions of moderate enrichment. To avoid the influence of extremely enriched regions due to artifacts in PCR amplification or repetitive elements, MACS randomly samples 1000 regions each having a 10 to 30-fold enrichment relative to the genome background [23]. For each of these regions, the reads aligned to the sense and antisense strand are separated and the mode of the reads' position in each strand is computed (as pictured in figure 2.13). The distance (in bases) between the modes of the sense and antisense peaks in the alignment is defined as d, and MACS shifts all the reads by $d/2$ bases toward the 3' ends to the most likely protein-DNA interaction sites.

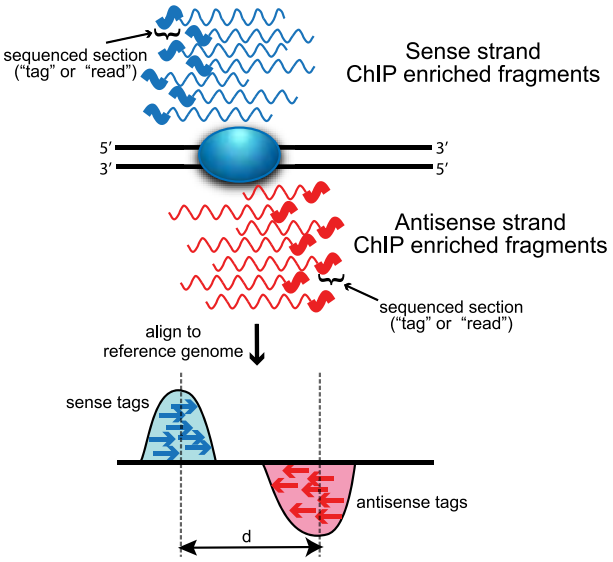


Figure 2.12: The 5'-to-3' sequencing requirement and short read length produce stranded bias in tag distribution. The shaded blue oval represents the protein of interest bound to DNA (solid black lines). Wavy lines represent either sense (blue) or antisense (red) DNA fragments from ChIP enrichment. The thicker portion of the line indicates regions sequenced by short read sequencing technologies. Sequenced tags are aligned to a reference genome and projected onto a chromosomal coordinate (red and blue arrows). Image from Wilbanks E.G., Facciotti M.T. [95].

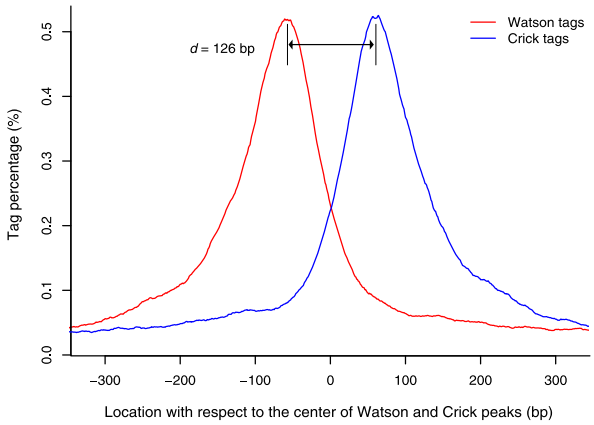


Figure 2.13: A random set of 1000 enriched regions is selected, the reads are assigned to the sense or antisense strands and the specific location of the enrichment, for each strand, is estimated. The distance between the peaks estimates the fragment size. Image from Zhang, et al. [101].

Enrichment estimation

Once the reads' position has been adjusted, MACS estimates the regions which express a stronger enrichment. The ChIP-Seq background reads distribution is modeled as a Poisson distribution with parameter λ_{BG}. MACS slides overlapping intervals (windows) across the genome to find regions with a statistically significant number of reads with respect to the modeled background distribution. Overlapping enriched regions are merged and the location with the highest read count is predicted as the precise binding location.

Control samples often show local fluctuations in the reads' distribution commonly due to DNA amplification and sequencing biases. As a consequence, the modeled background reads' distribution should not be uniform. In order to handle local biases, MACS uses a local model parameter λ_{local} which is defined considering the reads distribution in a local interval (long 5 to 10 kilobases) centered around the region location in the control sample. In absence of a control sample, the local reads' distribution around the current region is considered. Candidate peaks significantly enriched (statistical significance computed using a p-value of 10^{-5}) over the local λ_{local} distribution are called and the ratio between the reads' counts in the regions and the value from the local background distribution is reported.

3

Analysis of mutant TP53 roles in tumor breast cells

The p53 protein has an essential role in the regulation of the DNA repair and cell division processes, two crucial functions linked to tumor prevention and suppression. When DNA damage is detected, p53 is involved in the processes of starting the repair and, eventually, programming the cell's death. By stopping cells with mutated or damaged DNA from dividing, p53 helps prevent the advancement and spread of tumors [79].

In many researches studying cancer development, it has been observed that somatic mutations in TP53, the gene that codes for the p53 protein, occur in almost every type of human cancer at rates from 38% – 50% in ovarian, esophageal, colorectal, larynx, and lung cancers to about 5% in primary leukemia and cervical cancer. In about three out of four cases, these are caused by missense mutations: point mutations produced by the substitution of a single nucleotide [62].

The mutations TP53 acquires, apart from inactivating some wild-type functions, endow p53 mutants with oncogenic gain-of-function properties [90, 28, 18], some of which are mostly known to impact on tumor cell biology by significantly altering gene transcription. The observed TP53 mutations are numerous and well documented [80], but the precise ways in which they alter the biological processes in tumor cells is a topic actively researched.

This chapter reports the steps and the results of a collaboration with the CRO National Cancer Institute (Centro di Riferimento Oncologico) in Aviano, aimed at investigating the effects of the mutant p53 protein on the regulation of gene expression in breast cancer. The goal of the following parts is to present and explain the main questions we considered, the methods and tools we used to answer these questions and the interpretations we gave to the results.

3.1 Project's rationale and objectives

3.1.1 Project's background

Long non-coding RNAs, or lncRNAs in short, are a particular type of RNA molecules currently under heavy research. Among the reasons spurring an active interest in these molecules is their role in gene

regulation [27] and, in particular, in many cancer pathways [75, 35, 12].

Researchers at CRO are currently investigating lncRNAs regulated by mutant p53 in breast cancer cells, in particular in the MDA-MB-231 cell line which shows a missense mutation on the TP53 gene which causes the arginine amino acid to be replaced by the lysine in the produced p53 protein.

In order to identify lncRNAs regulated by mutant TP53, researchers performed RNA-Seq experiments on MDA-MB-231 cells silenced and not for mutant TP53 (sh1 vs shNT). This analysis led to the identification of a lncRNA, known as LINC01605, whose transcription levels decreased upon mutant p53 silencing.

Publicly available ChIP-Seq data on the MDA-MB-231 cell line was used to identify genome-wide mutant p53 binding sites and to determine whether mutant p53 could bind nearby LINC01605's locus. No mutant p53 binding was observed at the LINC01605 locus, but two binding sites were observed in a region 20 kb upstream LINC01605's first exon, as pictured in figure 3.1. Furthermore, the region was also characterized by a decrease in its transcriptional levels upon TP53 silencing, represented in figure 3.2.

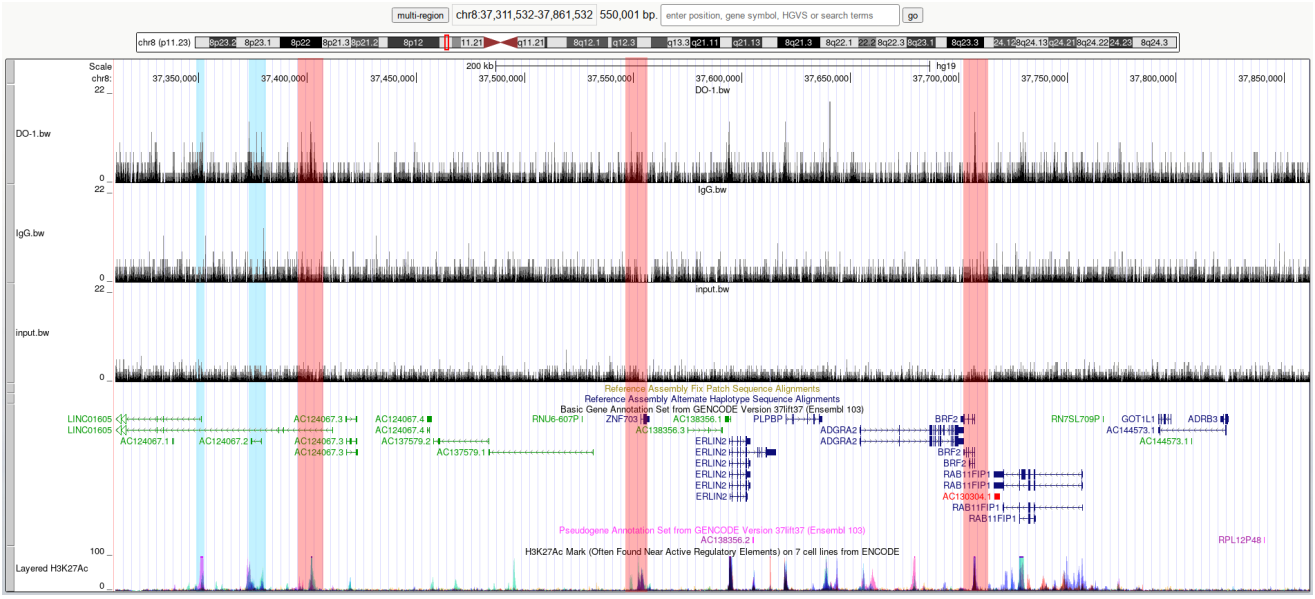


Figure 3.1: Putative mutant p53 binding sites. The top DO-1 track shows the p53 binding enrichment while the second and the third tracks are controls. The leftmost region marked in blue is the first exon of the LINC01605 gene, the one to its right is the identified putative regulatory element, distant about 20 kb. The regions highlighted in red are the p53 binding sites estimated by Walerych et al. in this interval [91].

Interestingly, the discovered region was characterized by a histonic modification, H3K27ac, commonly associated with active DNA regulatory elements (promoters and enhancers, for example). This evidence suggests a possible regulatory role for this region, which hereafter will be referred to as PRE (Putative Regulatory Element). Ongoing experiments are currently underway to confirm the dependence of LINC01605 on the PRE region.

3.1.2 Activities and objectives

In light of the recently identified potential role of mutant p53 on LINC01605 regulation, the research group pondered the feasibility of using bioinfomatic tools to extend the local analysis done on LINC01605 to the whole genome. The main goal of the activity presented in this chapter is the study of the effects of

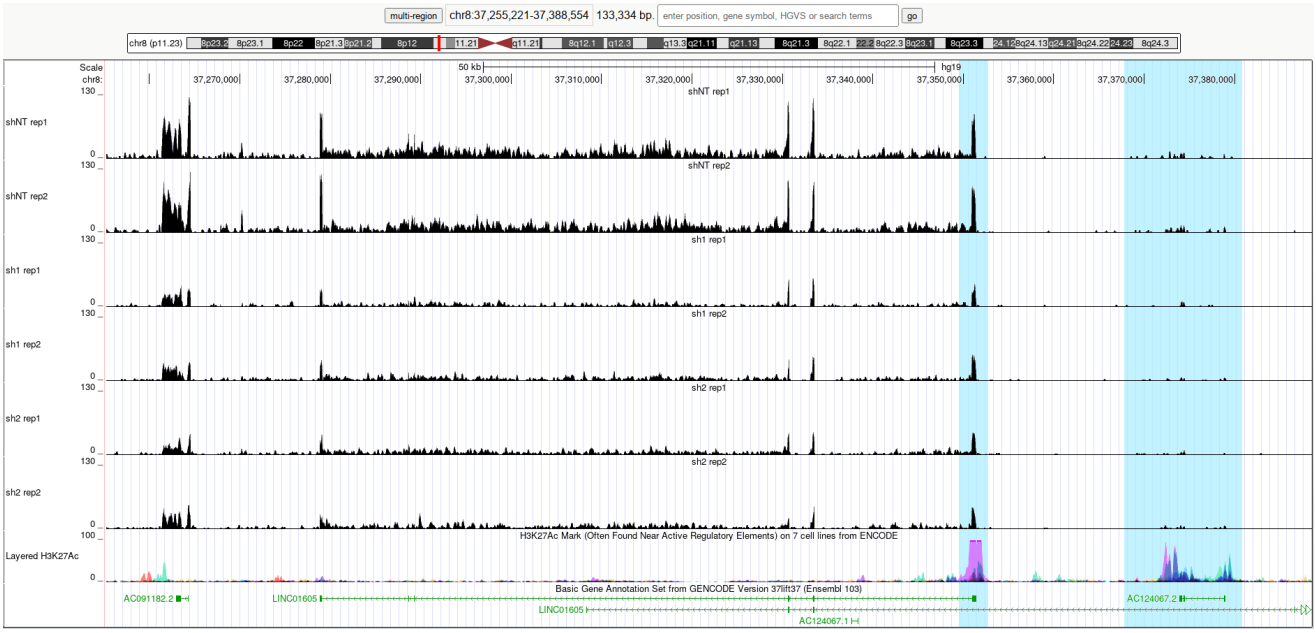


Figure 3.2: Differences in gene expression in the MDA-MB-231 samples upon silencing the mutant gene TP53. The leftmost region marked in blue is the first exon of the LINC01605 gene, the one to its right is the PRE region. The first two tracks from the top show the genes' expression levels in the non-TP53-silenced samples (shNT rep1 and rep2), while the following four refer to the silenced samples (sh1 and sh2). A decrease in the expression levels of LINC01605 can be observed by the reduced number of reads mapping to this region in the silenced samples. A less noticeable reduction can also be seen for PRE. The last track shows the presence of H3K27ac histonic modifications on the same regions.

mutant p53 on gene regulation. We are interested in determining which genes are directly and indirectly regulated by the target protein, in understanding which are the mechanisms of gene regulation primarily employed and in finding the main pathways and biological processes the targeted genes are involved in.

Furthermore, we want to examine the effectiveness of the currently available computational and statistical methods in these analyses and determine if they can be integrated in an automated analysis pipeline to support researchers.

In order to answer these questions, multiple analyses and methods were used and their description is found in the following sections of this chapter. First, in section 3.2, ChIP-Seq data is used to estimate the binding sites of the mutant p53 protein in the cell line of interest and in section 3.3 some classifications for these sites are considered, using genic annotation data, to better characterize the estimated binding sites. Then, sections 3.4 and 3.5 describe the attempts that were made to understand the mechanisms of direct and indirect gene regulation in which p53 is involved. Histonic annotations and RNA-Seq data are used to characterize the effects of mutant TP53 silencing and these results are integrated with the information about the putative binding sites of the studied protein. Furthermore, sections 3.6 and 3.7 illustrate how functional analysis methods can be used to outline the main biological pathways and processes in which the genes regulated by mutant TP53 are involved.

3.1.3 Schematic description of the analyses

This section provides a brief summary of the analyses described in this chapter.

Identification of putative mutant p53 binding sites

Problem to solve: Estimate the putative mutant p53 binding sites in the cell line MDA-MB-231 from ChIP-Seq data.

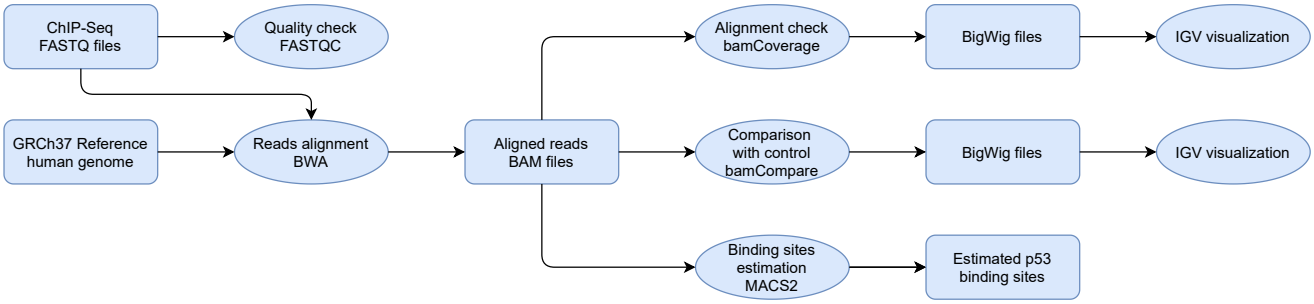
Input data:

- ChIP-Seq data from Walerych et al. [91] (3 fastq files, 2 controls) on mutant p53's binding
- Human genome assembly from Gencode (fasta files)

Employed tools:

- FASTQC (fastq files quality check)
- BWA (reads aligner)
- bamCoverage (conversion bam to bigwig to check the alignment's results)
- bamCompare (comparison with controls)
- MACS2 (p53's binding sites estimation)
- IGV (data visualization)

Section reference: Section 3.2



Quantitative characterization of the mutant p53 binding sites

Problem to solve: Determine the binding patterns of mutant p53 in the cell line MDA-MB-231. Understand if mutant p53 binds to genes or to intergenic sequences and determine the biotype of the genes it binds to.

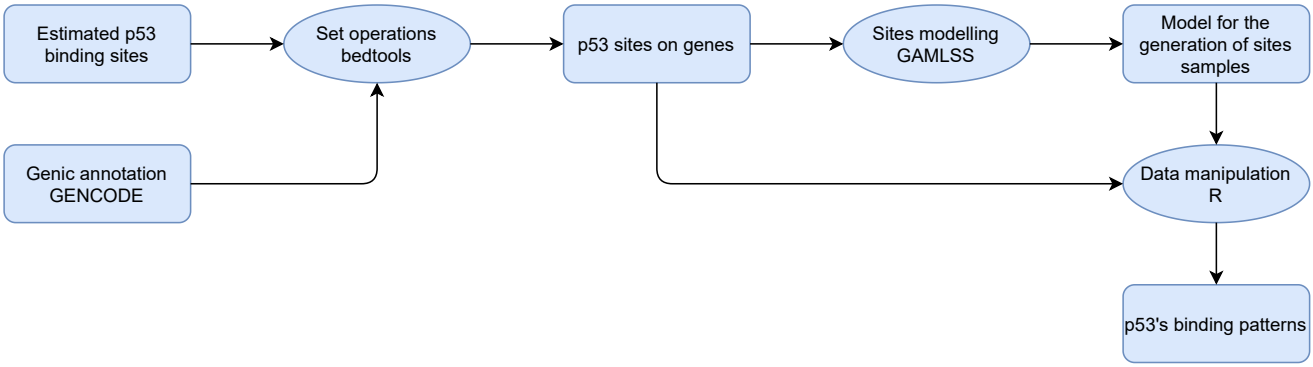
Input data:

- Estimated p53 binding sites (bed file)
- Genic annotation from Gencode (bed files)

Employed tools:

- bedtools (set operations on intervals)
- GAMLSS (statistical modelling)
- R (data manipulation, statistical analysis, plots)

Section reference: Section 3.3



Effects of mutant TP53 on direct gene regulation

Problem to solve: Study mutant TP53's role in gene regulation by observing the changes in the expression levels of the genes upon mutant TP53 silencing. This was done estimating the genes manifesting a significant change in the expression level (differentially expressed) when mutant TP53 is silenced and considering those located on (or close to) the mutant p53 binding sites.

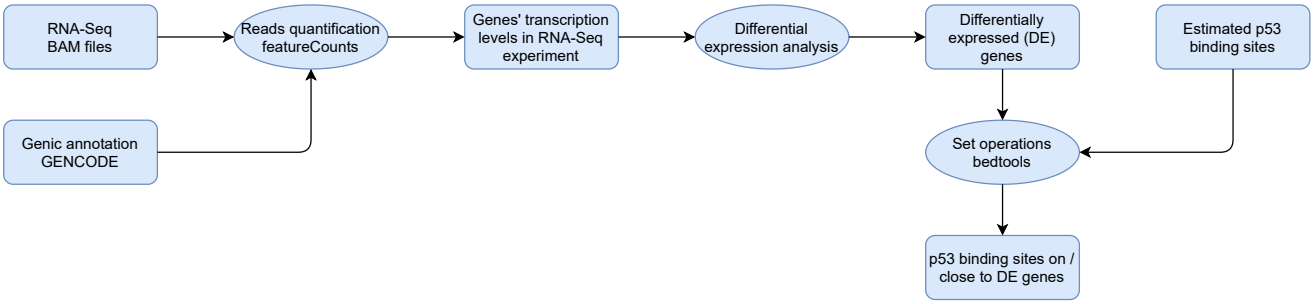
Input data:

- RNA-Seq data from CRO (6 bam files, 3 groups, 2 replicas per group) on TP53's silencing experiment
- Genic annotation from Gencode (gtf files)
- Estimated p53 binding sites (bed file)

Employed tools:

- featureCounts (reads quantification)
- DESeq2, EdgeR, Limma (differential expression analysis)
- bedtools (set operations on intervals)
- R (statistical analysis, plots)

Section reference: Section 3.4



Role of mutant TP53 on gene regulation control

Problem to solve: Study the regulatory mechanisms employed by mutant TP53 by observing its effects on regulatory regions, such as those characterized by a histonic modification H3K27ac. Identify

the differentially expressed genes located on (or close to) regulatory regions with a mutant p53 binding site.

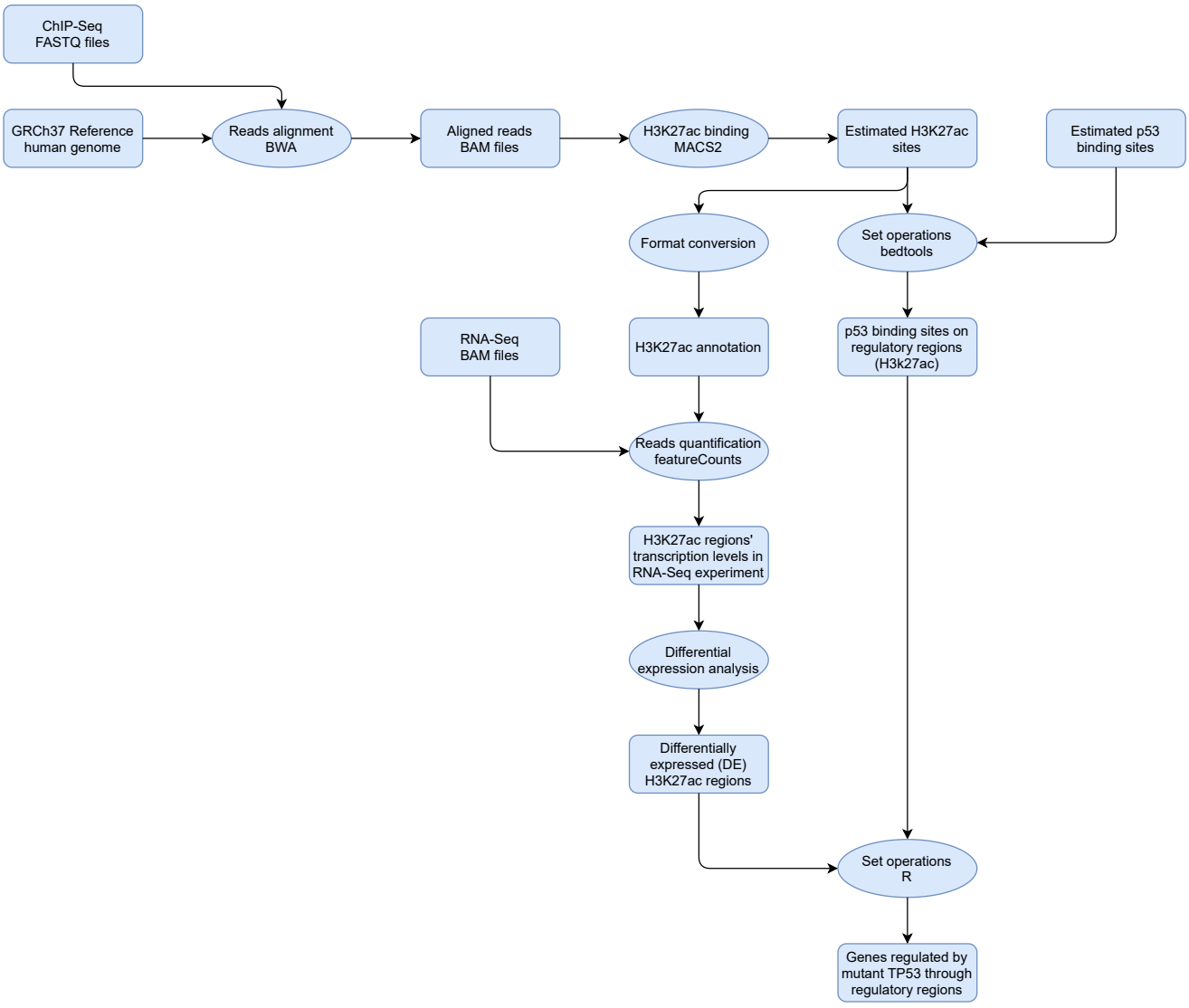
Input data:

- ChIP-Seq data from Li et al. [48] (2 fastq files, 1 control) on the H3K27ac histone modifications
- Estimated p53 binding sites (bed file)

Employed tools:

- BWA, MACS2 (H3K27ac sites estimation)
- featureCounts (reads quantification)
- DESeq2, EdgeR, Limma (differential expression analysis)
- bedtools (set operations on intervals)
- R (statistical analysis, plots)

Section reference: Section 3.5



Functional analysis of TP53

Problem to solve: Study the main biological functions and processes potentially regulated by mutant TP53 on the cell line MDA-MB-231 and, among these, identify the ones regulated through p53's binding to regulatory regions. Assess the analysis' outcome comparing the results with known functions regulated by mutant TP53.

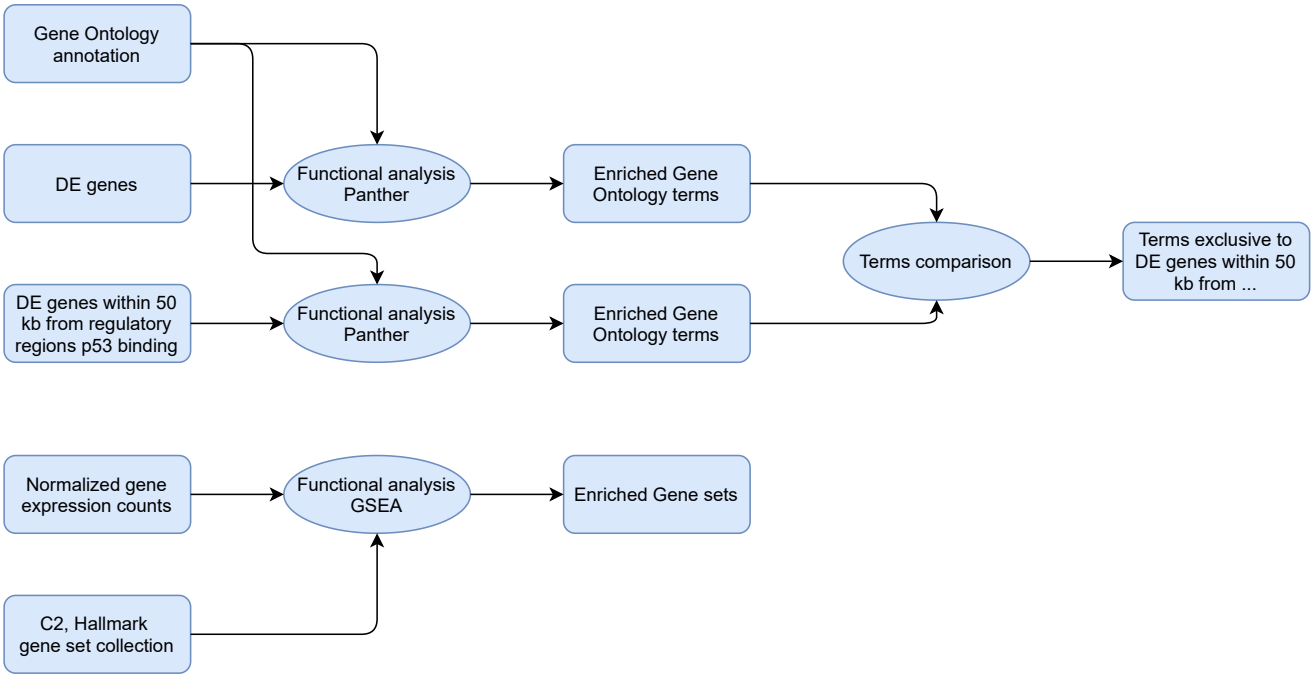
Input data:

- RNA-Seq normalized counts (on TP53's silencing experiment) by DESeq2
- DE genes estimated by the differential expression analysis tools
- Gene Ontology annotation
- Hallmark, C2 gene set collections

Employed tools:

- Panther, GSEA (Functional analysis)

Section reference: Section 3.6



LINC01605's role in TP53 regulatory functions

Problem to solve: Study the role of the gene LINC01605 in mutant TP53's regulatory functions. Estimate which of the functions potentially regulated by mutant TP53 are due to LINC01605. Use the features provided by the functional analysis tools (such as the semantic clustering) to improve the interpretation of the functional analysis results.

Input data:

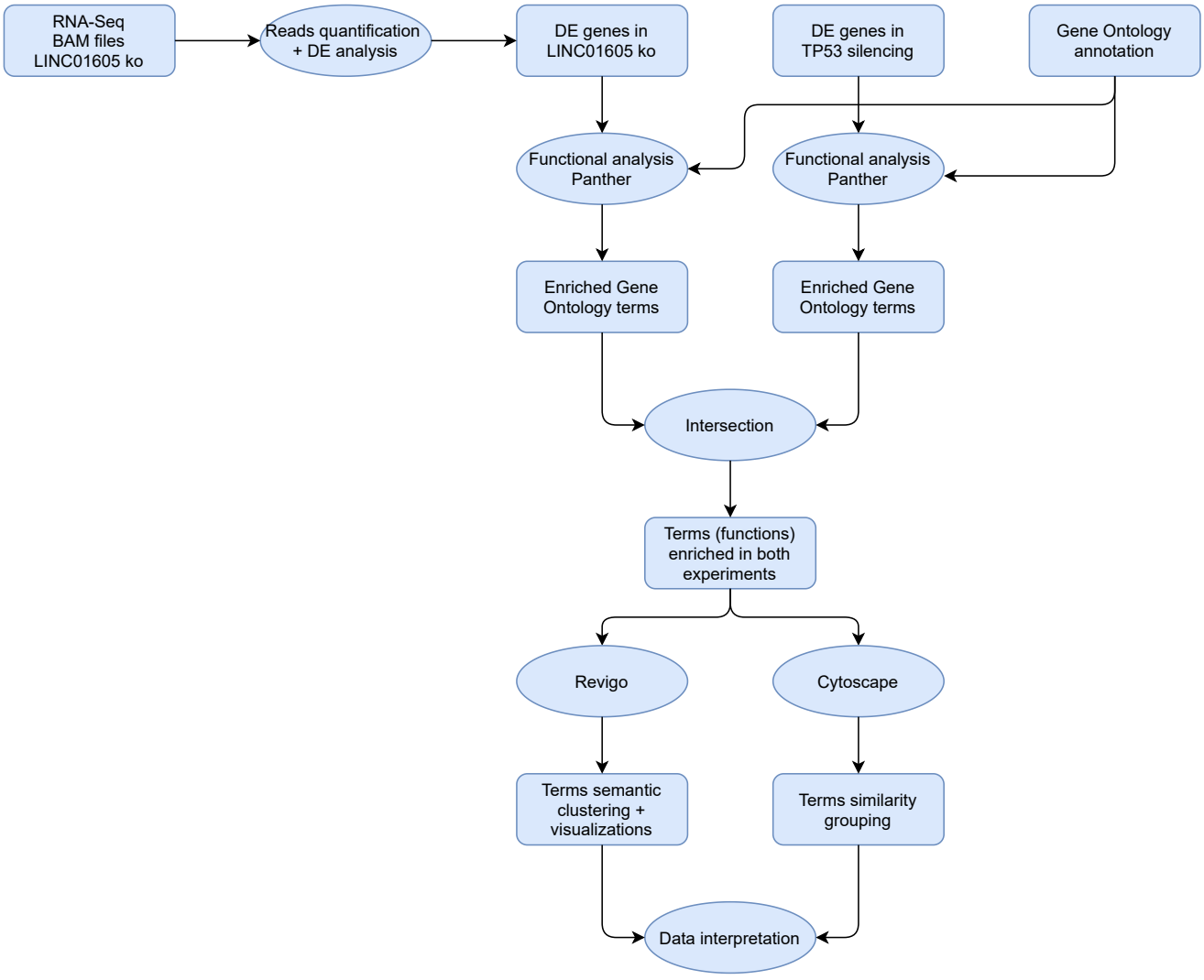
- RNA-Seq data from CRO (8 bam files, 2 groups, 4 replicas per group) on LINC01605's knock-out experiment

- DE genes estimated by the differential expression analysis tools (in the TP53 silencing experiment)

Employed tools:

- Panther
- Revigo (semantic clustering and visualization of the Gene Ontology terms)
- Viseago (semantic clustering and visualization of the Gene Ontology terms)
- GSEA (for KEGG pathway analysis)
- Cystoscape (similarity grouping of the Gene Ontology terms)

Section reference: Section 3.7



3.2 Identification of putative mutant p53 binding sites

The content presented in this section aims at describing the first steps we followed in the analysis of the mutant p53 effects on gene regulation, from the description of the available ChIP-Seq data to the identification and characterization of p53's binding sites on the genome.

3.2.1 Data and quality control

As introduced in the previous chapters, ChIP-Seq is a recently developed technology, based on high-throughput sequencing, that can be used to determine the genome-wide binding site of a protein of interest. The ChIP-Seq data used in the study described in this chapter was collected by Walerych et al. [91], and it originates from the breast cancer cell line MDA-MB-231. More specifically, three sequencing experiments were performed: the first using the p53-specific antibody DO-1 and two controls using the generic IgG antibody and the input (not immunoprecipitated sample). All the data is available on NCBI's SRA database in FASTQ format (access code SRP055837).

To inspect the FASTQ files and to check their quality, the program FASTQC was employed. Figure 3.3 represents, from the top row, the per-base sequence quality, the per-sequence quality scores and the per-base sequence content of the three FASTQ files. The files contain reads exactly 50 bp long and the produced plots do not highlight sequencing problems or biases.

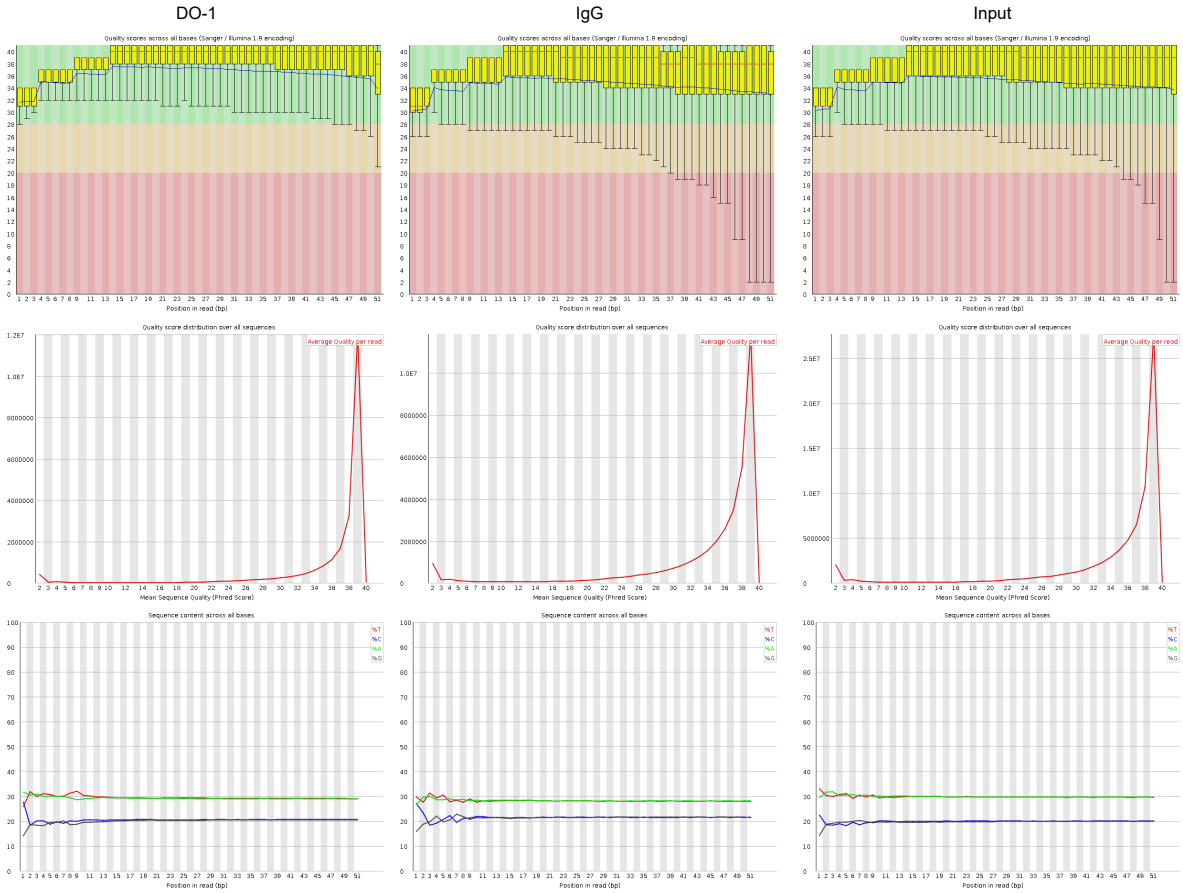


Figure 3.3: Quality control plots of the three FASTQ files representing ChIP-Seq reads. From the top row, the plots show the per-base sequence quality, the per-sequence quality scores and the per-base sequence content of the sequencing reads encoded in the FASTQ files. The columns, from left to right, refer to the experiment using the p53-specific DO-1 antibody and the two controls employing IgG and no antibody, respectively.

3.2.2 Alignment and assessment of the results

The ChIP-Seq reads were aligned to version GRCh37 of the human genome through the BWA aligner. The specific genome reference assembly that was used for the alignment is the one provided by GENCODE, as it shares the chromosome names with the annotation files used in the subsequent analyses, therefore avoiding the need for some conversion steps. We opted for the release version GRCh37, instead of the more recent GRCh38, because our RNA-Seq reads were already aligned on this version.

To inspect the ChIP-Seq results and assess the alignment outcome, the BAM file produced by BWA were converted to bigwig: a compact binary indexable format specifically designed for efficient visualization. A preliminary genome-wide visualization of the ChIP-Seq files, shown in figure 3.4, confirms the presence of short regions with an extraordinary enrichment in all samples. These represent common artifacts and biases present in ChIP-Seq data that can be due to the DNA fragmentation and replication procedures. Fortunately, thanks to the availability of control samples, we can identify these regions as false positives.

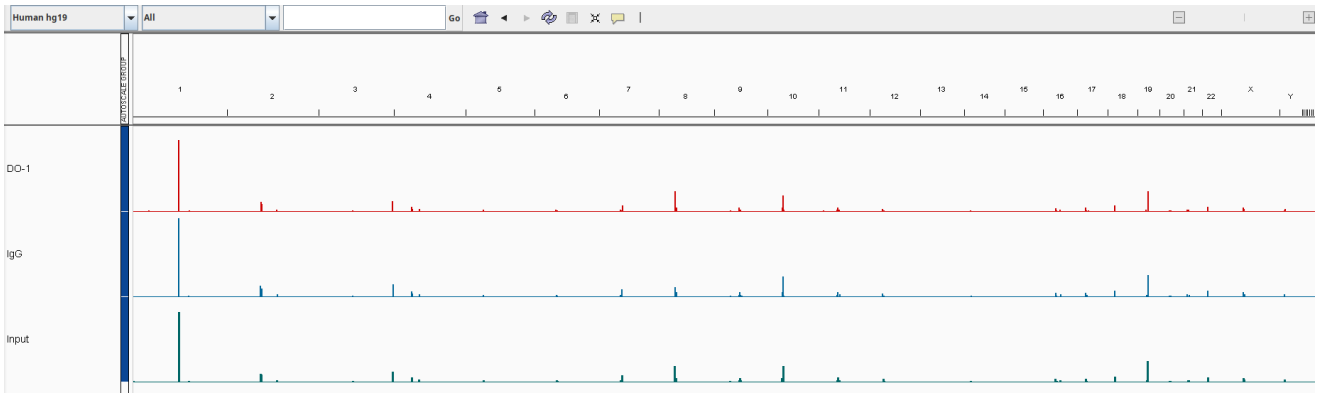


Figure 3.4: Preliminary visualization (in IGV) of the results of the three ChIP-Seq experiments. The height of the peaks is proportional to the binding enrichment of mutant p53 in the region. It can be noticed that the displayed peaks are also present in the control tracks and are, therefore, to be discarded.

In order to assess the results of the alignment, the enrichment level of some regions identified by Walerych et al. [91] as putative mutant p53 binding sites was compared to the one in the aligned BAM files. Figure 3.5 compares the protein binding levels in three different gene loci: PSMA1, PSMB9 and PSMC1. The enrichment levels in our alignment are comparable to those shown in the reference article, confirming the positive outcome of the alignment process.

In order to compare the results of the two control experiments and decide which one to use in the subsequent analyses, *bamCompare* was used. This program partitions the two provided BAM files in bins (of custom length) and compares them in pairs, computing a specified comparison function (such as the difference or the ratio). In our study, we compared the BAM file from the proper DO-1 experiment to both controls. Figure 3.6 presents the tracks obtained with bamCompare. The tracks show the logarithm of the ratio of the number of mapped reads along the genome in the DO-1 experiment and in the controls. The observed similarity between the comparisons suggested the analogy between the control experiments results and led us to the decision to proceed with the analysis using the Input file as control.

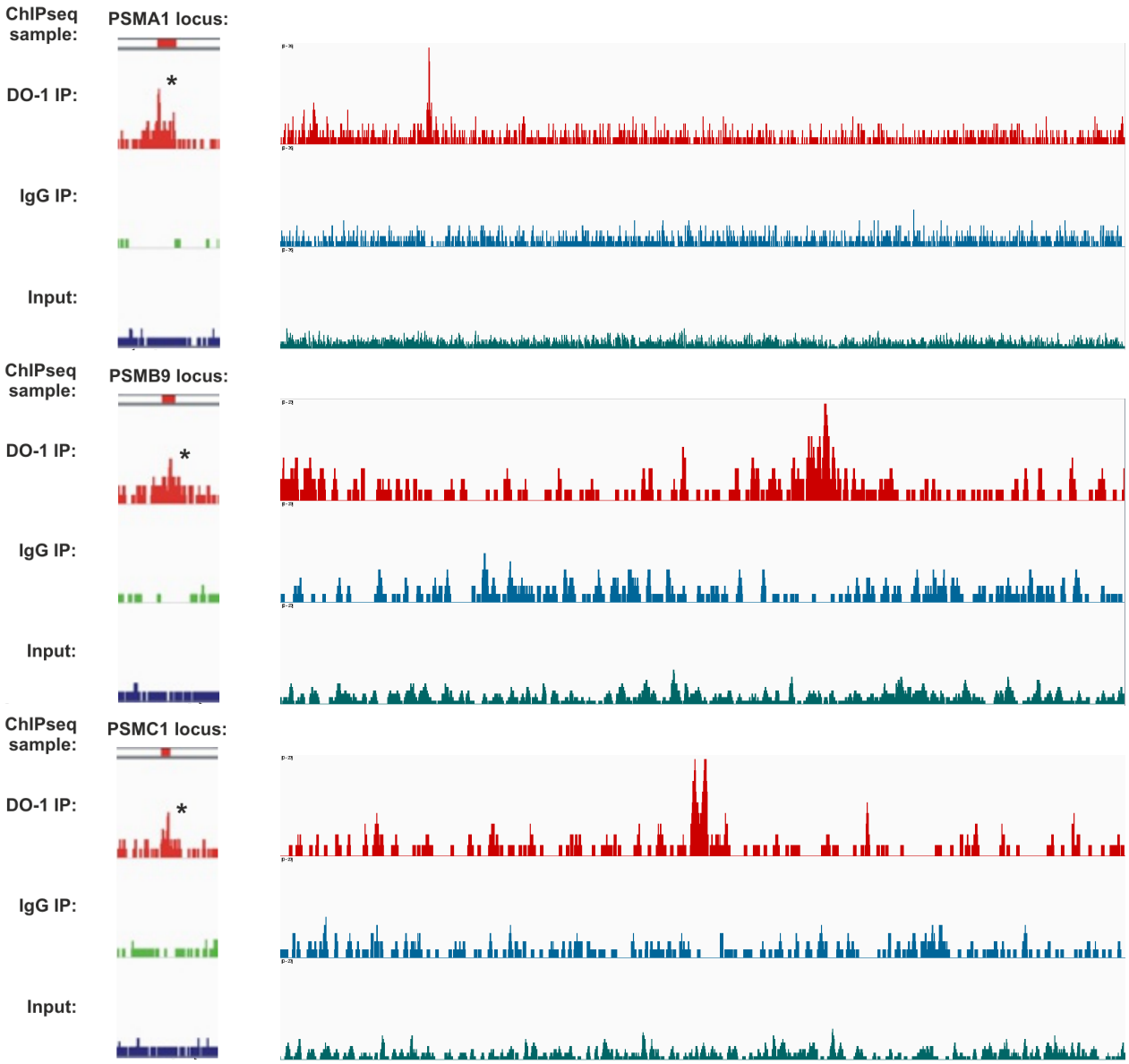


Figure 3.5: Graphical comparison of the enrichment level in the gene loci PSMA1, PSMB9 and PSMC1 (from top to bottom) between the reference article [91] and the reads aligned via BWA. In each comparison, the track on the top refers to the DO-1 experiment and the ones in the middle and on the bottom to the IgG and Input controls, respectively.

3.2.3 Binding sites estimation

MACS2 was used to estimate the mutant p53 binding sites, using the information from both the DO-1 experiment data and the Input control. After applying bamCompare, we noticed that the distribution of the enrichment ratio had particularly low values so, in order to obtain a greater number of putative sites, the default minimum enrichment ratio value of 5 was lowered to 2. MACS2 produces a text file containing the location of the estimated binding sites and an enrichment score for each. In our case, we obtained a total of 11,028 estimated mutant p53 sites.

In order to evaluate MACS2's results, we located the estimated binding sites on the genome and we compared the number of mapped reads in the region between the experiment and the control tracks, as pictured in figure 3.7.

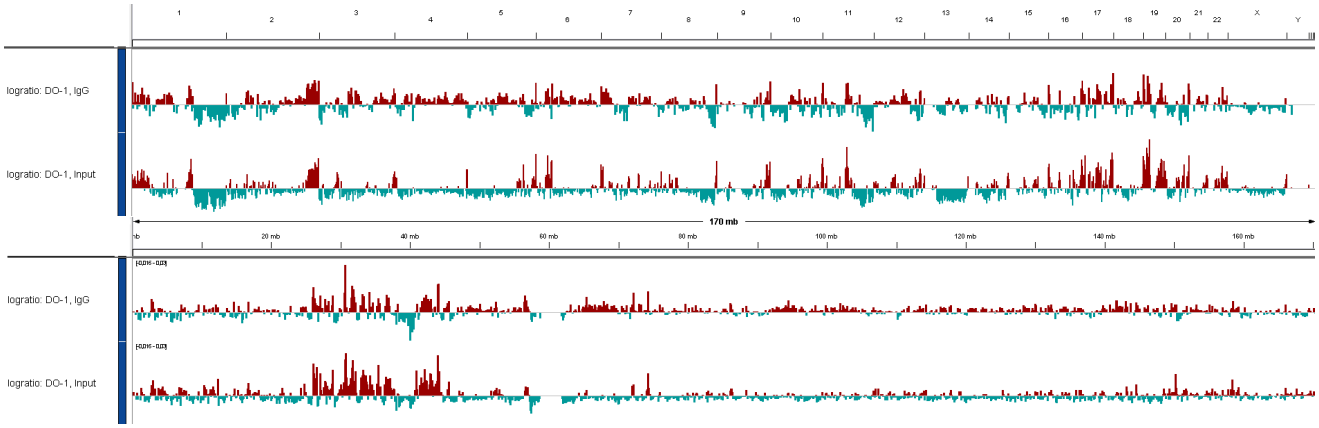


Figure 3.6: Binding enrichment of p53 compared to the control. The two tracks on the top present the comparison on the whole genome scale, while the two on the bottom focus on chromosome 6. Each track, produced by bamCompare, plots the logarithm of the ratio of the number of mapped reads in the proper experiment (which used DO-1 to detect p53's binding) and in a control experiment. In the tracks on the top the control experiment is the one using the IgG antibody, while in the tracks on the bottom the control used is the one not using antibodies. The color red is used to indicate positive values: regions in which the estimated protein binding in the DO-1 experiment is stronger than that in the control one.

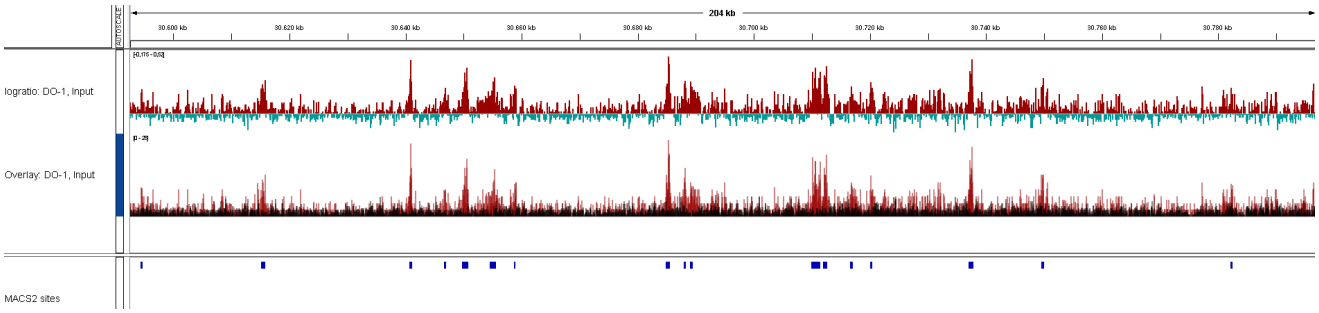


Figure 3.7: Comparison of some p53 binding sites estimated by MACS2 to their number of mapped reads, in a region on chromosome 6. The track on bottom, labeled `MACS2 sites` *indicates with blue intervals the p53 binding sites estimated by MACS2. The second track overlays the DO-1 and control number of reads in the region: it can be seen that the location of the estimated binding sites matches regions characterized by a p53 binding enrichment.*

We noticed that, especially in regions characterized by an extreme number of mapped reads likely caused by replication artifacts, MACS2 may report false positives. An example of this phenomenon can be observed in three regions pictured in figure 3.8. In these regions the number of mapped reads is the same in the experiment and in the control (and the number of reads is several magnitudes higher than the average) but MACS2 reports them as binding sites. A simple solution to avoid these false positives is to compare the MACS2 estimated sites to the result from bamCompare, filtering out the ones not reaching a specified threshold.

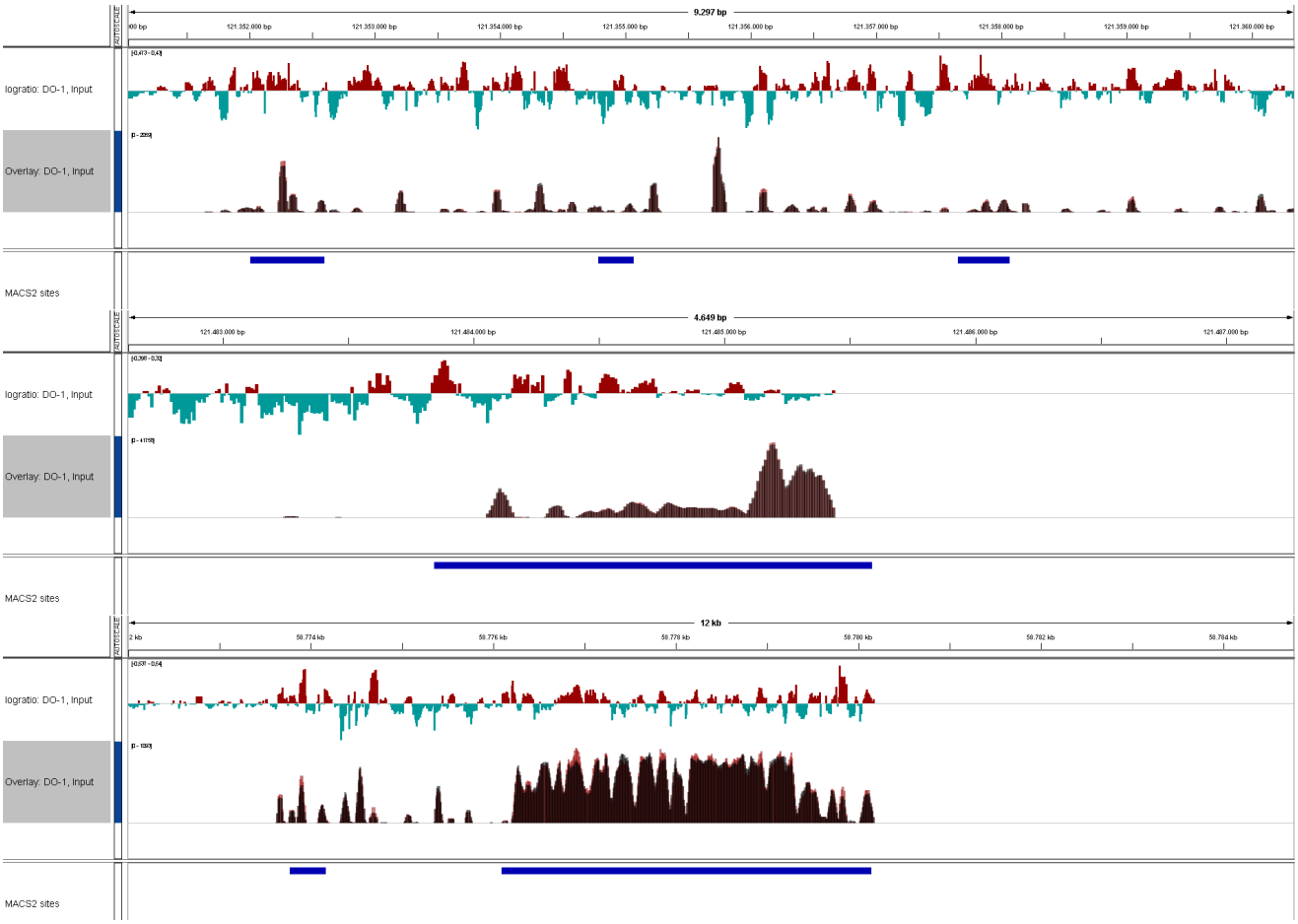


Figure 3.8: Three examples of false positives reported by MACS2. The first two regions (from the top) are on chromosome 1, while the third one is on chromosome 6. In these regions the number of reads in the proper DO-1 experiment and in the control is practically the same.

3.3 Quantitative characterization of the mutant p53 binding sites

Once the putative binding sites of the mutant p53 protein have been estimated, it can be of interest to characterize these regions. More precisely, we are interested in determining if the binding sites are mainly located on genes or in intergenic regions and in classifying the genetic biotype of the genes which could be regulated by p53.

To this end, we consulted the GENCODE annotation of genes and transcripts and located the estimated protein binding sites in the annotation's content. Since GENCODE changed the way it classifies genes with release 31, notably combining the previously disjointed categories *lincRNA*, *antisense* and *processed transcript* in a new one called *lncRNA*, we repeated some classifications on both release versions 35 and 24.

The main toolkit that was used in this phase is *bedTools*[1]: a powerful set of tools for set operations on genomic intervals.

3.3.1 Grouping the transcripts of the same gene

The version 35 of the GENCODE annotation contains 94,096 entries, each representing a gene transcript, belonging to a total of 44,573 genes. Similar values are obtained considering version 24: 84,817 entries for 43,527 genes.

Through the use of the feature `intersect`, included in bedTools, one can quickly know which putative p53 binding sites are located on gene transcripts. Considering the version 35 of the annotation, we counted 25,311 intersections between the estimated binding sites and the gene transcripts and a total of 2899 sites (out of 11,028) not located on genes. The usage of the annotation version 24 indicates similar values: 22,659 intersections and 3103 sites not on genes.

As the interactions between proteins and DNA are not necessarily strand-specific, the intersections considered here are strand-independent.

Merging the transcript on the same gene

As previously introduced, the downloaded GENCODE annotation only specifies the location and length of the gene transcripts, not of the single genes. Since we are mainly interested in genes and in their position relative to the p53 binding sites rather than in their transcripts, we had to consider some methods to avoid the duplicates: intersections between a binding site and multiple transcripts of the same gene.

We applied and compared two approaches for the removal of such duplicates.

- In the first one only one intersection pair *(p53 binding site, transcript on gene A)* per gene A is kept.
- In the second one a new annotation file is produced: the transcripts belonging to the same gene name, located on the same chromosome and strand, are merged into one region. The start and end coordinates of the new gene region are defined, respectively, as the minimum of the start coordinate and the maximum of the end coordinate of its transcripts.

[1] bedTools documentation: `https://bedtools.readthedocs.io/en/latest/`

Using the first approach, and considering the annotation version 35, we obtain 10,368 intersections (down from 25,311) and 2899 sites not located on genes. Using the version 24 we have 9899 intersections (down from 22,659) and 3103 sites not on genes.

The new annotations produced by the second approach contain 45,763 and 44,114 entries, in versions 35 and 24 respectively. Using these annotations we observed a great increase in the number of intersections (now 62,780 up from 25,311) with the estimated binding sites and a worrying decrease in the number of sites with no gene intersections (now 256, down from 2899). By keeping only one entry per gene we expected, instead, a decrease in the total number of intersections.

Upon further inspection, we noticed that the majority of these intersections (83%) involves genes with biotype *snoRNA* and *miscRNA*, despite these accounting only for a small portion of the total number of genes (5%). This can be seen in tables 3.1 and 3.2.

gene biotype	n. of intersections
snoRNA	33647
misc_RNA	18746
protein_coding	7914
lncRNA	2303
snRNA	53
TEC	41
miRNA	32
others	44

Table 3.1: After merging the transcripts belonging to the same gene, we observed an increase in the number of intersections, most of which involve snoRNA and miscRNA genes.

gene biotype	quantity in annotation
protein_coding	20085
lncRNA	16757
miRNA	3048
snRNA	1913
misc_RNA	1274
TEC	1037
snoRNA	889
others	381

Table 3.2: Summarization of the genes by biotype, after merging the transcripts of the same gene.

The problem is caused by the fact that, in the GENCODE annotation, the transcripts of the same *snoRNA* and *miscRNA* gene (on the same chromosome and strand) can be very far apart, despite their, on average, short length. Merging these transcripts generates very long genes that intersect many of the estimated p53 binding sites, as can be seen in tables 3.3 and 3.4.

chromosome	gene	n. of transcripts	transcripts average length
chr2	snoU13	45	101
chr1	snoU13	44	101
chr3	snoU13	32	101
chr7	snoU13	31	102
chr6	snoU13	28	99
chr16	snoU13	25	103
chr5	snoU13	24	101
chr4	snoU13	22	103
chr11	snoU13	21	98
chr10	snoU13	19	101

Table 3.3: Number of transcripts of the gene snoU13 in each chromosome. Despite their short length, they are numerous and distant.

gene biotype	n. in merged annotation	genes average length
snoRNA	889	11067662
misc_RNA	1274	3957549
rRNA	46	67617
protein_coding	20085	63260
lncRNA	16757	31333
processed_transcript	9	27063
lincRNA	145	14577
polymorphic_pseudogene	48	12358
antisense	55	9951
TR_C_gene	5	5402

Table 3.4: Average length, in the merged annotation, of the genes by biotype: snoRNAs and miscRNAs are much less numerous but much longer compared to the protein coding genes and lncRNAs.

Ignoring *snoRNAs* and *miscRNAs*, we counted 10,387 intersections between the estimated p53 binding sites and the merged transcripts, considering the annotation version 35. Furthermore, 2897 binding sites are not located on genes.

After such correction, the values were much more comparable to the ones obtained using the first approach, as it is observable in table 3.5.

	GENCODE v35		GENCODE v24	
	first approach	second approach, after *snoRNA* and *miscRNA* removal	first approach	second approach, after *snoRNA* and *miscRNA* removal
intersections between genes and binding sites	10368	10387	9899	10230
binding sites not on genes	2899	2897	3103	3027

Table 3.5: After the removal of the troublesome transcripts, the two approaches used to avoid duplicate intersections between genes and p53 binding sites provided comparable values.

3.3.2 Random samples generation method for normalization

The number of intersections and intersected genes by the estimated p53 binding sites is, by itself, not very informative unless it is compared to the same metrics computed on random samples of sites. We outlined a method for generating random samples of regions similar to the estimated binding sites and compared the results between the actual p53 estimated binding sites and the randomly generated ones, in order to identify if the estimated binding sites behave differently from random regions along the genome, with respect to the amount and biotype of genes they are located on.

Starting from the 11,028 mutant p53 binding sites estimated by *MACS2*, we intended to generate samples of the same size of regions having the same length distribution of the binding sites, randomly uniformly located along the genome taking into consideration the differences in length of the chromosomes.

In order to choose a fitting statistical model to describe the length distribution of the binding sites, we employed *GAMLSS* [81]. This R package compares the provided empirical distribution (in our case the lengths of the estimated binding sites) to various families of statistical models in order to find the one that, most likely, could have generated the empirical sample. Furthermore, the best-fit parameters for the chosen model are also estimated.

The length distribution of the estimated p53 binding sites has a mean of 448 and a median of 390 base pairs (evidences of positive skewness in the distribution) and presents some outliers, as shown in figure 3.9. We decided to remove such outliers by limiting the values to 1000, as it is close to the 98th percentile (which is 989).

The most appropriate statistical model, as estimated by *GAMLSS*, is the *Generalized Gamma* distribution and the estimates for its parameters are displayed in code 3.1.

The generalized gamma distribution (GG) is a general statistical model characterized by three shape parameters. This model generalizes some commonly used distributions such as the Weibull, the exponential and the gamma distribution [39].

```
Family:  c("GG", "generalised Gamma Lopatatsidis-Green")

Coefficient(s):
              Estimate   Std. Error    t value   Pr(>|t|)
eta.mu     5.62815e+00  2.38736e-04  23574.788 < 2.22e-16 ***
eta.sigma -3.64549e+00  4.81776e-03   -756.678 < 2.22e-16 ***
eta.nu    -5.76927e+02  1.79812e-02 -32085.018 < 2.22e-16 ***
```

Code 3.1: Best-fit parameters estimates for the generalized gamma model.

We then compared the length distribution of the p53 binding sites to the one of the sites from the random samples generated by the estimated statistical model. In figure 3.10, it can be seen that the sample GG distribution approximates the target one (with the outliers removed): the first quartile of the compared distributions is comparable but the median and the third quartile are lower in the random sites' distributions. The overall presence of lower values in the random distributions is also confirmed by the comparison of the density plots, shown in figure 3.11, and in the quantile-quantile plot, in figure 3.12. The comparison of the quantiles of the two distributions highlights that the greatest disparities between the two are attributable to the lower values (for instance, the lower α% of the p53 distribution tops at 600, while in the random samples it stops at 550), while the two are indistinguishable on the higher quantiles.

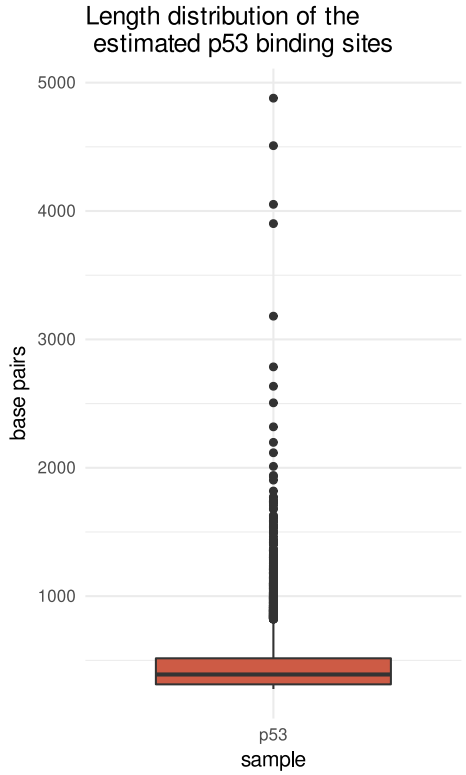


Figure 3.9: Distribution of the lengths of the 11028 estimated binding sites. The value of the first three quartiles is under 600, but there are some outliers having much higher values.

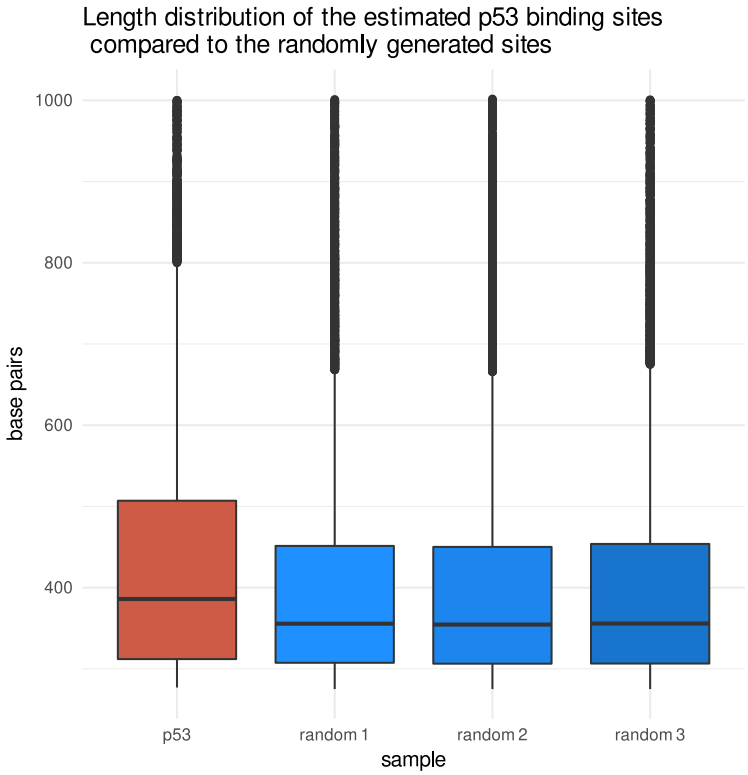


Figure 3.10: Comparison of the distributions of the lengths of the estimated binding sites (outliers removed) and of three random samples from the fitted model (p53's distribution capped to 1000). The random samples contain slightly lower values.

3.3.3 Genetic classification of the mutant p53 binding sites

The mutant p53 protein is known to affect gene expression, therefore we wanted to identify the genes it binds to and possibly regulates. For instance, binding to *protein coding* genes could indicate the presence of direct regulation mechanisms, while binding to non coding genes, such as *long non-coding RNAs*, could suggest indirect gene regulation through such RNAs, which may function as regulatory elements for other genes. After counting the number of intersected genes by the estimated p53 binding sites (by biotype), we first tried to normalize the counts with respect to the total number of genes in the annotation and, then, we compared them with the values obtained using the sites from the random samples.

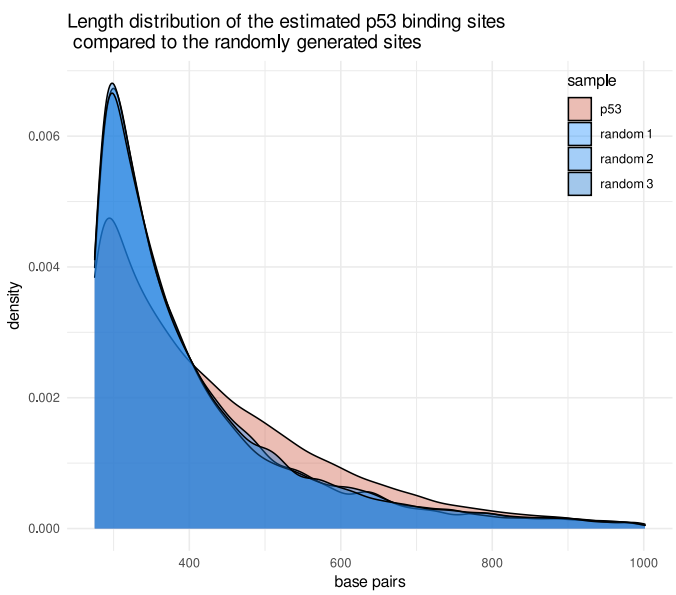


Figure 3.11: The length distribution of the random sites is characterized by slightly lower values compared to the one of the estimated p53 binding sites.

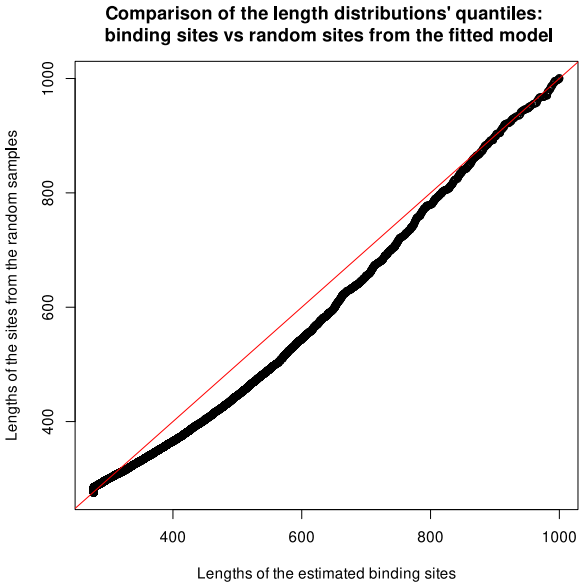


Figure 3.12: Comparison of the quantiles from the p53 binding sites length distribution to the ones from the random sites length distribution.

Biotypes of the genes containing p53 binding sites

Tables 3.6 and 3.7 show, for every gene biotype, the fraction of genes in the annotation that are intersected by at least one of the estimated p53 binding sites. Without accounting for gene length, 30.1% of all the *protein coding* and 10.9% of all the *lncRNA* genes contain (more precisely, intersect) a p53 binding site. Considering instead the version 24 of the annotation, once we combine the *antisense*, *lincRNA* and *processed_transcript* genes into one category, we obtain ratios comparable to those in the v35 annotation (30.2% and 10.8%).

gene biotype	intersected genes	genes in annotation	ratio
protein_coding	6042	20050	**0.301**
lncRNA	1822	16744	**0.109**
TEC	39	1037	0.0376
snRNA	35	1900	0.0184
miRNA	27	3040	0.00888
Mt_tRNA	10	22	0.455
snoRNA	10	427	0.0234
others			

Table 3.6: Absolute and relative number of genes in the v35 annotation, by biotype, having at least one putative p53 binding site.

gene biotype	intersected genes	genes in annotation	ratio
protein_coding	6027	19969	**0.302**
antisense	796	5578	**0.143**
lincRNA	573	7727	**0.0742**
processed_transcript	129	498	**0.259**
TEC	40	1047	0.0382
snRNA	33	1898	0.0174
miRNA	27	3041	0.00888
others			

Table 3.7: Absolute and relative number of genes in the v24 annotation, by biotype, having at least one putative p53 binding site.

Comparison with the randomly generated sites

In order to better understand the significance of the count data about the genes containing the mutant p53 binding sites, we decided to apply a comparison with randomly generated sites. Random samples of sites uniformly distributed along the genome were produced, as described in section 3.3.2. We computed, then, the intersections between these regions and the gene annotation and counted the genes intersected by at least one site, as done with the estimated p53 sites. Finally, we compared these values to the ones obtained considering the p53 binding sites: doing so, we tried to estimate if the types of genes p53 interacts with are significantly different from what could have been observed from any random set of

sites.

Figures 3.13, 3.14 and 3.15 show that, compared to the sites from the random samples, the estimated mutant p53 binding sites intersect more genes and fewer sites are not located on genes.

Figure 3.13: Compared to the sites from the random samples, the estimated p53 binding sites are more likely to be located on genes. Gene annotation v35.

Figure 3.14: Compared to the sites from the random sample, the estimated p53 binding sites interact with more genes. Gene annotation v35.

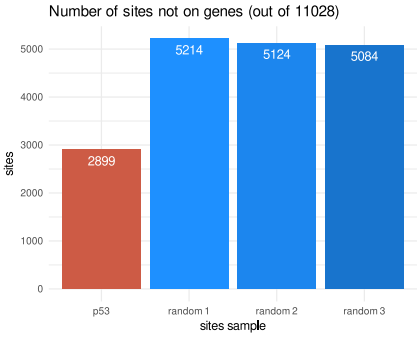


Figure 3.15: Out of all the 11028 sites considered in each random sample, about half are not located on genes, while more than two thirds of the putative p53 binding sites are located on genes. Gene annotation v35.

Furthermore, the estimated p53 binding sites tend to be more frequently located on *protein coding* genes compared to what can be expected from the randomly generated sites. Not only the p53 binding sites are more likely to be located on genes, but, as noted in table 3.8, they also interact more with *protein coding* genes: 75.3% of all the genes intersected by the p53 binding sites are *protein coding*, compared to about 69% in the sites from the random samples.

	p53 binding sites		random sample 1		random sample 2		random sample 3	
biotype	absolute	relative	absolute	relative	absolute	relative	absolute	relative
protein_coding	6042	**0.753**	3208	**0.702**	3214	**0.691**	3261	**0.688**
lncRNA	1822	**0.227**	1324	**0.290**	1408	**0.303**	1445	**0.305**
TEC	39	0.00486	8	0.00175	3	0.000645	9	0.00190
snRNA	35	0.00436	4	0.000875	5	0.00107	3	0.000633
miRNA	27	0.00337	3	0.000656	5	0.00107	3	0.000633
others								

Table 3.8: Number of genes, by biotype, containing at least one p53 binding site, or a site from the random samples. The relative counts express the fraction of the genes intersected by the considered sites belonging to the indicated biotype. The p53 binding sites are more frequently located on protein coding genes compared to the random control.

When considering p53 sites outside of genes, it can be observed that, compared to the same sites from the random samples, these are characterized by lower distances to the closest gene. Figure 3.16 presents the distance distributions to the closest downstream and upstream gene.

Furthermore, the estimated p53 binding sites more frequently have a *protein coding* gene as the closest one. This holds for both the downstream and upstream directions. The frequencies are presented in table 3.9.

Considerations about normalization with respect to gene length

Previously in this chapter, in tables 3.6 and 3.7, we showed that the estimated p53 binding sites are about three times more likely to be located on a *protein coding* gene than a *lncRNA* one, without

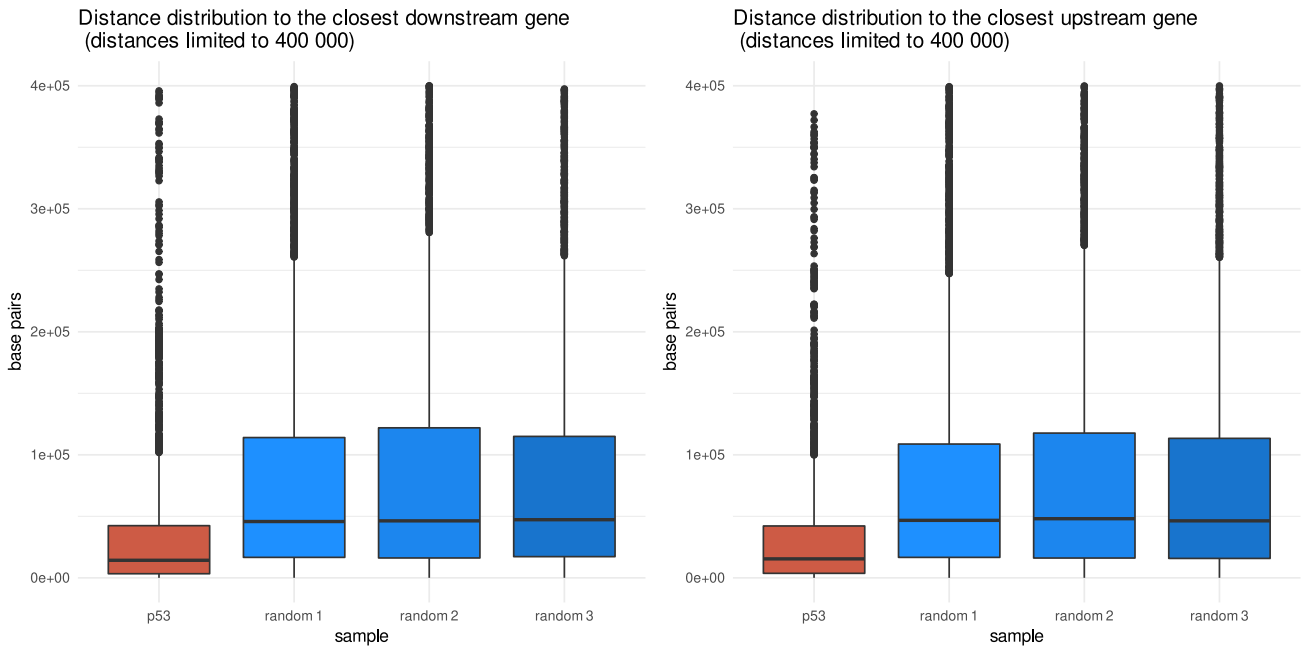


Figure 3.16: Distance distributions to the closest gene for the sites not located on genes. Gene annotation version 35.

biotype	p53 binding sites	random sample 1	random sample 2	random sample 3
protein_coding	**0.532**	**0.347**	**0.344**	**0.338**
lncRNA	**0.302**	**0.378**	**0.389**	**0.395**
miRNA	0.0665	0.0593	0.0609	0.0571
snRNA	0.0343	0.0896	0.0803	0.0817
misc_RNA	0.0287	0.0303	0.0330	0.0307
others				

Table 3.9: Considering the sites not located on genes, the table shows, for each biotype, the fraction of sites having a gene with such biotype as the closest one. The estimated p53 binding sites have more frequently a protein coding gene as the closest one. Data from the downstream direction.

accounting for gene length. This result, however, can be misleading because longer genes will have a higher probability of containing intervals such as our binding sites.

Repeating the previous analysis for the regions in the random samples, we noticed that the fraction of *protein coding* genes intersected by the random sites is about double that of *lncRNAs* (table 3.10), as is their average length (table 3.11). This means that the random sites intersect about 16% of all the *protein coding* genes and about 8% of the *lncRNAs*, while the average length of these genes is about 63,230 and 31,333 base pairs, respectively.

As the regions from the random samples are uniformly distributed along the genome and under the assumption that the location of a gene is independent of its biotype, after the application of a normalization method that takes into account gene length, these random regions should intersect the same fraction of *protein coding* and *lncRNA* genes.

Considering the average gene length as the normalization factor, the random sites are equally likely to be located on a *protein coding* or a *lncRNA* gene, while the estimated p53 binding sites are about 1.37 times more likely to be located on *protein coding* genes compared to *lncRNAs* (equations 3.1 and 3.2).

gene biotype	total	p53 binding sites absolute	p53 binding sites relative	random sample 1 absolute	random sample 1 relative	random sample 2 absolute	random sample 2 relative	random sample 3 absolute	random sample 3 relative
protein_coding	20050	6042	0.301	3208	**0.16**	3214	**0.160**	3261	**0.163**
lncRNA	16744	1822	0.109	1324	**0.0791**	1408	**0.0841**	1445	**0.0863**
miRNA	3040	27	0.00888	3	0.000987	5	0.00164	3	0.000987
snRNA	1900	35	0.0184	4	0.00211	5	0.00263	3	0.00158
misc_RNA	1214	7	0.00577	6	0.00494	4	0.00329	5	0.00412
others									

Table 3.10: Number of genes, by biotype, containing at least one site. Comparison of the p53 binding sites with the sites from the random samples. About 6000 protein coding genes, out of the total 20050 in the v35 annotation, contain one of the estimated p53 binding sites (about a 30% relative frequency). The sites from the random samples intersect a fraction of protein coding genes twice as large compared to lncRNAs (not taking into account gene length normalization).

gene biotype	count	average length	median length
protein_coding	20085	**63260**	24265
lncRNA	16757	**31333**	6088
miRNA	3048	547	86
snRNA	1913	4195	107
misc_RNA	1274	3957549	293
others			

Table 3.11: Considering the length of a gene obtained by merging its transcripts, as described in section 3.3.1, the average length of protein coding genes is about double (2.02x) that of lncRNAs. The median length is about 4 times as large.

$$\text{random sites: } \frac{0.16}{0.0791 \cdot 2.02} = 1.01 \tag{3.1}$$

$$\text{p53 sites: } \frac{0.301}{0.109 \cdot 2.02} = 1.37 \tag{3.2}$$

3.4 Effects of mutant TP53 on direct gene regulation

Mutations on the TP53 gene are known to play a role on gene transcription and to affect some biological functions in the cells. To improve the understanding of these regulatory mechanisms, we are interested in gaining information about the links between mutant p53's binding and the genes that are subject to changes in their expression levels.

In order to study the effects of the mutant p53 protein on direct gene regulation, we analysed purposely-produced RNA-Seq data on our cell line of interest, MDA-MB-231. The RNA-Seq experiment aims at highlighting the genes affected by mutant p53 by comparing wild-type cells to cells subject to TP53 silencing.

In the experiment, TP53 is silenced through RNA interference, a biological process that employs shRNA molecules (short hairpin RNAs) to reduce the translation of a gene's transcript. Since RNA interference does not prevent gene transcription, it is usually less effective compared to *gene knockout* processes, and does not completely prevent the transcriptional products.

As we are comparing non-silenced to TP53-silenced cell samples from the TP53-mutant MDA-MB-231 cell line, the effects we will observe will characterize all the functions of the mutant TP53 gene, not only the ones it acquires with the mutation.

3.4.1 Differential expression analysis

The RNA-Seq experiment was performed in untreated cells (*shNT*) and in cells silenced for p53 using two different shRNAs, *sh1* and *sh2*. The experiment was performed in two biological replicates, for a total of six samples.

After the sequencing, the reads were aligned to the reference assembly GRCh37 using the splice-aware aligner *HiSAT2*. The program *featureCounts* was then used to count and assign the reads contained in the BAM files to the genomic features in the GENCODE annotation (release version 35). As usual in RNA-Seq reads quantification, the reads overlapping multiple exons of the same gene were counted only once and the ambiguous ones were discarded.

Three software tools for differential gene expression analysis, *DESeq2*, *EdgeR* and *Limma*, were then applied to the raw counts produced by featureCounts and the results were compared. The genes considered differentially expressed were those exhibiting an increase or a decrease in expression of at least 1.5 times over the *shNT* samples and passing the adjusted *p-value* threshold of statistical significance of 10^{-2}.

In order to check the effectiveness of the TP53-silencing experiment, we looked at its expression levels. Compared to the non-silenced *shNT* samples, the *sh1* samples presented an expression reduction of 3.9 times and the *sh2* samples one of 2.7 times. As anticipated in section 3.1, the *lncRNA* LINC01605 also saw a decrease in its expression levels by 1.82 times. The values are presented in table 3.12 and refer to the raw number of reads mapping to the genes in the analysed samples.

The number of genes considered differentially expressed (DE) by the three programs are presented in table 3.13. Figure 3.17 highlights that the three programs provide consonant results as most of the DE genes according to DESeq2 and EdgeR are also discovered by Limma; DESeq2 provides the most conservative results.

In both *shNT* samples, all three programs reported that about 70% of the DE genes are upregulated upon TP53 silencing, while 30% are downregulated (table 3.14). All three DE analysis methods detected that only about half the genes DE in the first silencing are also DE in the second one, as pictured in figure 3.18.

	shNT		sh1		sh2	
TP53	588	647	137	182	245	215
LINC01605	160	172	82	99	111	71

Table 3.12: Number of reads assigned to the TP53 and LINC01605 genes in the six RNA-Seq samples. The expression, represented by the number of assigned reads, of the two genes decreased upon silencing the TP53 gene.

	shNT vs sh1	shNT vs sh2
DESeq2	972	1127
EdgeR	1156	1567
Limma	1419	1659

Table 3.13: Number of genes reported as differentially expressed by the programs DESeq2, EdgeR and Limma in the two mutant TP53 silencing experiments.

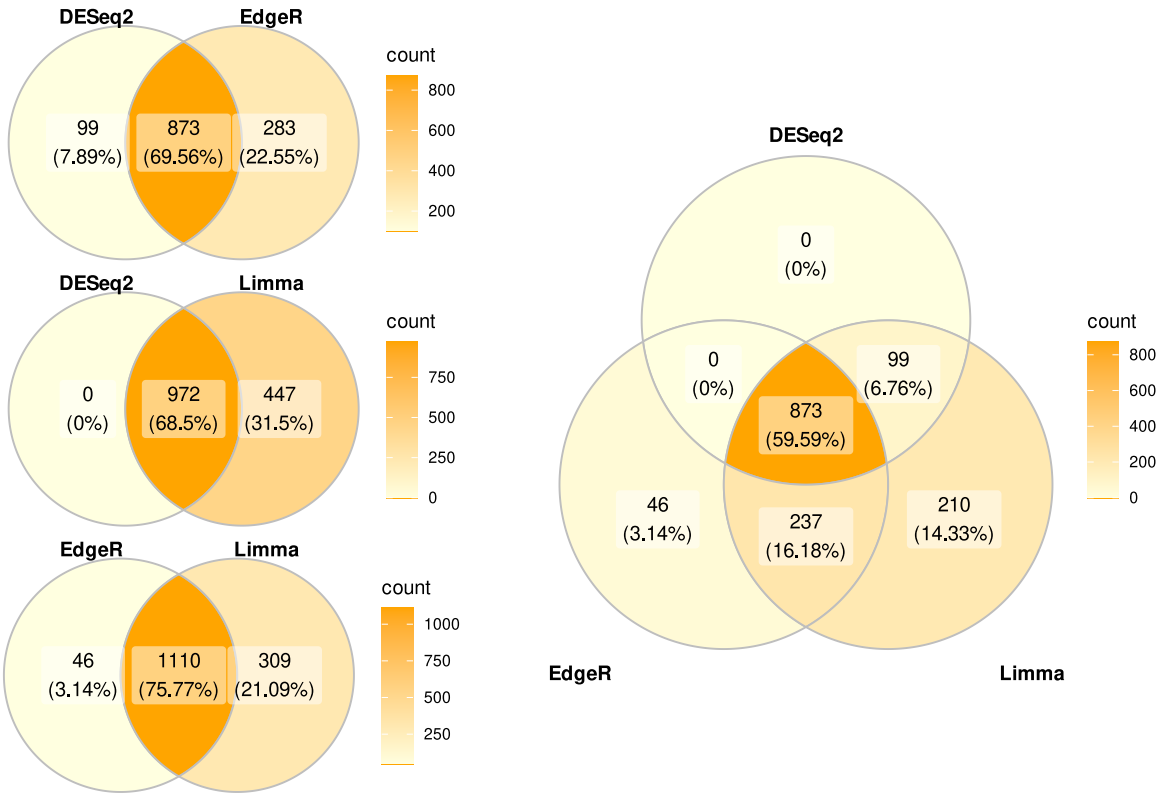


Figure 3.17: Number of genes considered differentially expressed by three differential expression analysis tools: DESeq2, EdgeR and Limma. Data about the shNT vs sh1 DE analysis.

	shNT vs sh1						shNT vs sh2					
	DESeq2		EdgeR		Limma		DESeq2		EdgeR		Limma	
increment	74%	723	72%	837	72%	1023	65%	729	68%	1064	69%	1140
decrement	26%	249	28%	319	28%	396	35%	398	32%	503	31%	519

Table 3.14: Out of all the genes considered differentially expressed by the three tools considered, the table indicates how many are upregulated and downregulated upon TP53 silencing.

The volcano plot is a graphical representation commonly used to display the genes revealing a change in expression (between two tested conditions) among all the genes in the considered annotation. Figure 3.19 portais the volcano plot containing the genes considered differentially expressed by DESeq2 in the *sh1* vs *shNT* comparison. An horizontal gray bar is used as the threshold of statistical significance, while two vertical lines indicate the thresholds of fold change: the genes beyond the right bar or before

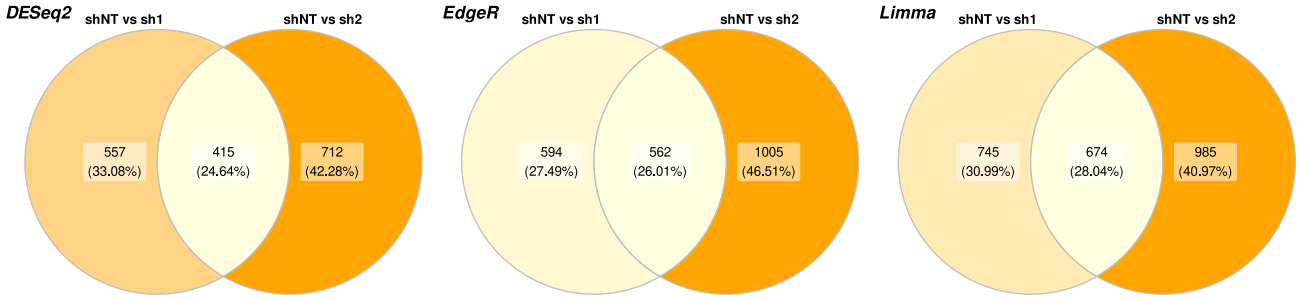


Figure 3.18: Number of genes considered differentially expressed by DESeq2, EdgeR and Limma in the two silencing experiments. All three programs detect that about half the genes DE in the first silencing are also DE in the second one.

the left bar manifest an increased or decreased expression level of at least 1.5 times compared to the non-silenced sample.

In figure 3.19 the color purple is used to indicate the genes involved in the *regulation of nuclear division*, a process commonly associated with TP53's activity. In this case, 15 of the 136 genes involved in said process are subject to significant variations in their expression levels, suggesting that TP53 may be involved in such process.

Considering the biotype of the differentially expressed genes, we noticed that over 96% of these are *protein coding*. Furthermore, the differentially expressed *protein coding* genes cover more than the 4% of all the *protein coding* genes in the annotation, compared to about 0.15% in *lncRNAs* (table 3.15).

		shNT vs sh1			shNT vs sh2		
biotype	genes in annotation	DE genes	% of all DE genes	% in annotation	DE genes	% of all DE genes	% in annotation
protein_coding	20081	840	96.2	4.18	1034	96.1	5.15
lncRNA	16778	23	2.63	0.137	29	2.70	0.173
unprocessed_pseudogene	2528	3	0.344	0.119	5	0.465	0.198
misc_RNA	2033	2	0.229	0.0984	2	0.186	0.0984
others							

Table 3.15: Number of differentially expressed genes, by biotype, in the sh1 and sh2 silencing experiments. In both experiments, the DE protein coding genes amount to about 96% of all the DE genes and cover more that 4% of all the protein coding genes in the v35 annotation.

Figure 3.19: Volcano plot representing the genes considered differentially expressed by DESeq2 in the sh1 vs shNT analysis. Every point represents a gene in the annotation, the genes considered DE are those passing the thresholds of statistical significance and of fold change (represented as triangles). The genes colored in purple are the ones involved in the "regulation of nuclear division" process (a function regulated by TP53): a relatively large part of these genes present changes in their expression level upon TP53 silencing.

3.4.2 Differential expression of lncRNAs

Long non-coding RNAs are RNA molecules whose transcript is not directly associated to protein production. Since it has been observed that the transcription level of these genes is usually lower compared to *protein coding* ones [11], we wanted to probe the sensitivity of the employed DE analysis methods on these genes, hypothesizing that they could be overshadowed by the more expressed *protein coding* genes.

The DE methods were applied both on the full gene annotation and on the simplified annotation containing only *lncRNA* genes and the number of DE genes reported by each was then compared. It is important to note that the annotation can be simplified both before and after the feature quantification (the application of *featureCounts* to assign the reads to the genes), but a premature filtering of the non-*lncRNA* genes can lead to an overly-optimistic quantification, as some of the previously discarded ambiguous reads will be now assigned to these genes.

Figure 3.20 shows that all three DE analysis methods are less sensitive when considering only the counts of the *lncRNA* genes. DESeq2 provides the most consistent result of the three methods, while Limma reports 10 times less *lncRNA* DE genes when considering only the *lncRNA* gene counts instead of the full annotation. Seen these results we decided to continue the analyses without reserving special treatment to *lncRNA* genes.

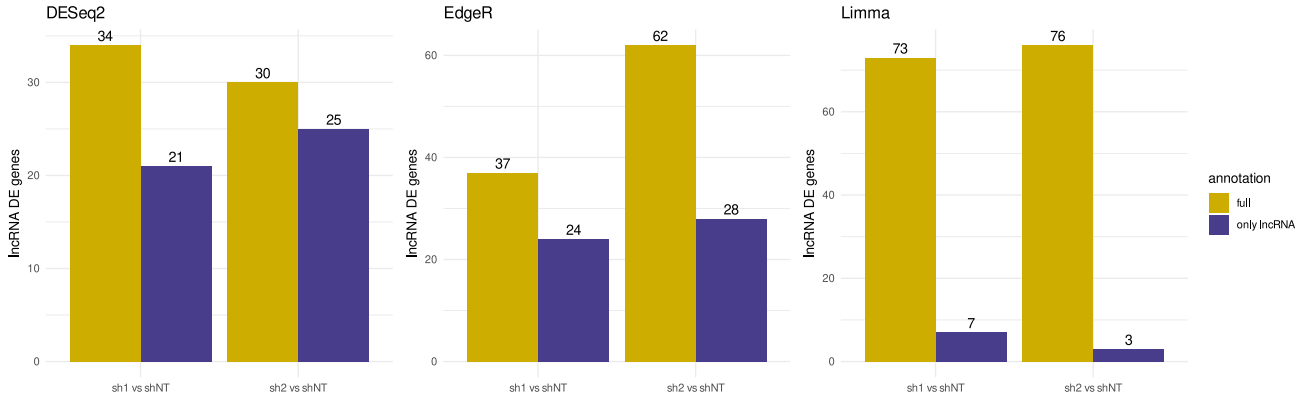


Figure 3.20: Number of lncRNA genes reported as DE by the three methods when considering the full gene counts or only the lncRNAs counts. All three methods have a higher sensitivity when considering the full gene counts.

3.4.3 Interactions between differentially expressed genes and mutant p53

In order to study the direct effects of mutant p53 on gene regulation, we tried to determine if, and in what measure, the genes containing p53 binding sites are subject to variations in their transcription level once p53 is silenced. Such genes could be considered candidates regulatory targets for the mutant p53 protein.

The presence of a p53 binding site on a differentially expressed gene does not necessarily indicate a regulatory activity, as the location of such binding site can be due to randomness. To account for such aspect, we compared the number of p53 sites located on DE genes to the quantity of a same amount of sites randomly positioned on the chromosomes that happen to be located on DE genes.

Figure 3.21 shows that the p53 sites are more frequently located on DE genes, compared to sites randomly located on the genome. Considering sites from samples of size 11,028, the number of p53 sites located on DE genes is about 2 to 3 times greater compared to the random sites: 607 out of 11,028 p53 binding sites are on DE genes while the random sites stop at about 270. This holds for both silencing experiments.

Considering the sites not located on DE genes, the amount of sites within 2000 base pairs from a DE gene is considerably higher in the p53 sites compared to the random ones: more than 1% of the estimated p53 binding sites not on DE genes are within 2000 base pairs from one (in either downstream or upstream direction) compared to about 0.1% in the random sites (figure 3.22).

The tendency for the p53 sites to be close to DE genes is also supported by their distance distribution to the closest DE gene: the median distance is about 650 kilobases whereas the random sites have a median distance of about 1.6 megabases to the closest DE gene, about 2.5 times greater (figure 3.23).

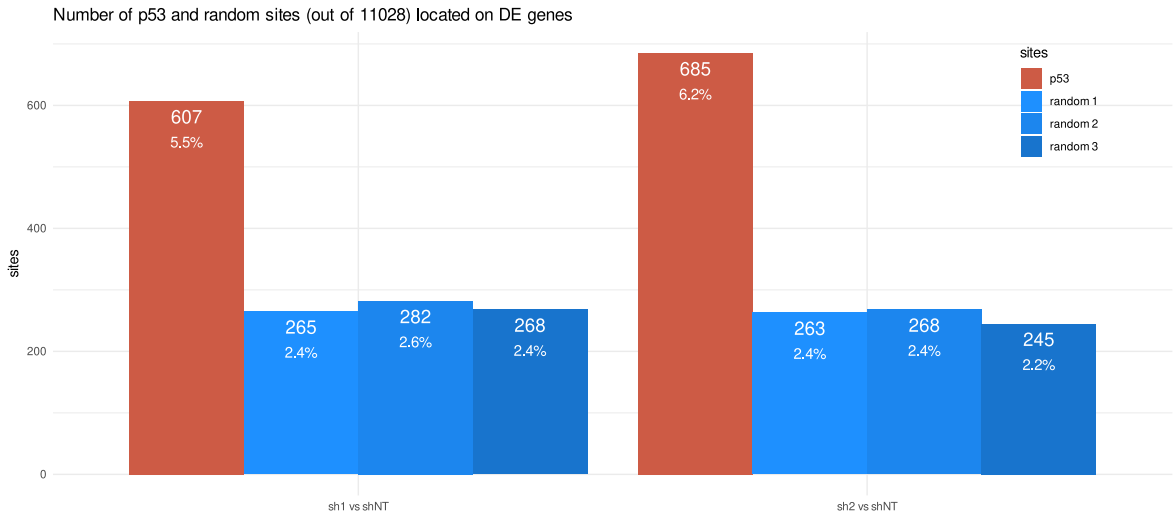


Figure 3.21: The estimated p53 binding sites are more frequently located on DE genes, compared to the sites from the random samples.

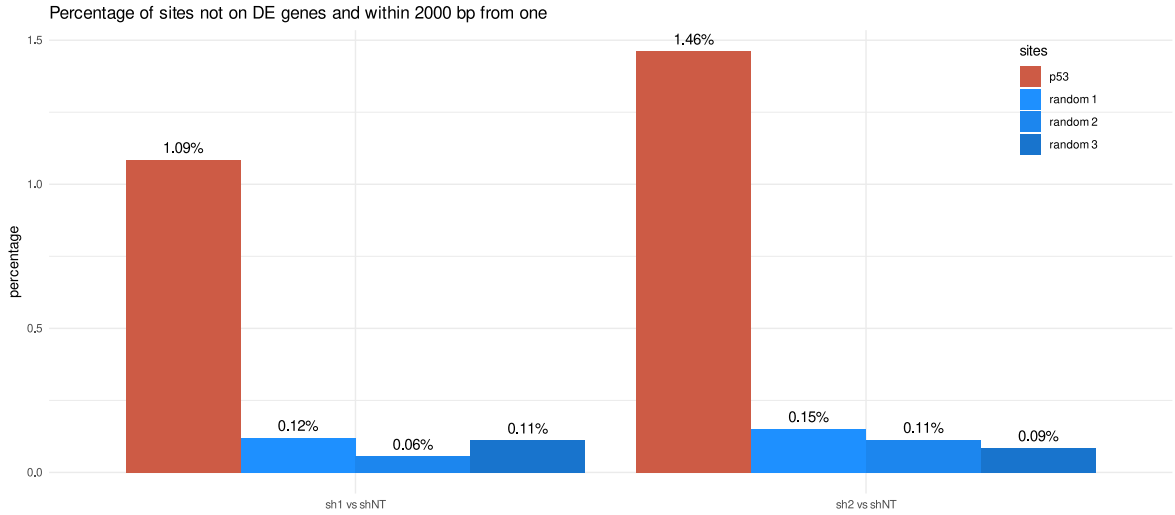


Figure 3.22: Out of all the p53 sites not located on DE genes, about 1.1-1.5% are very close to one, compared to the sites from the random samples. The closest gene is considered in either downstream or upstream direction.

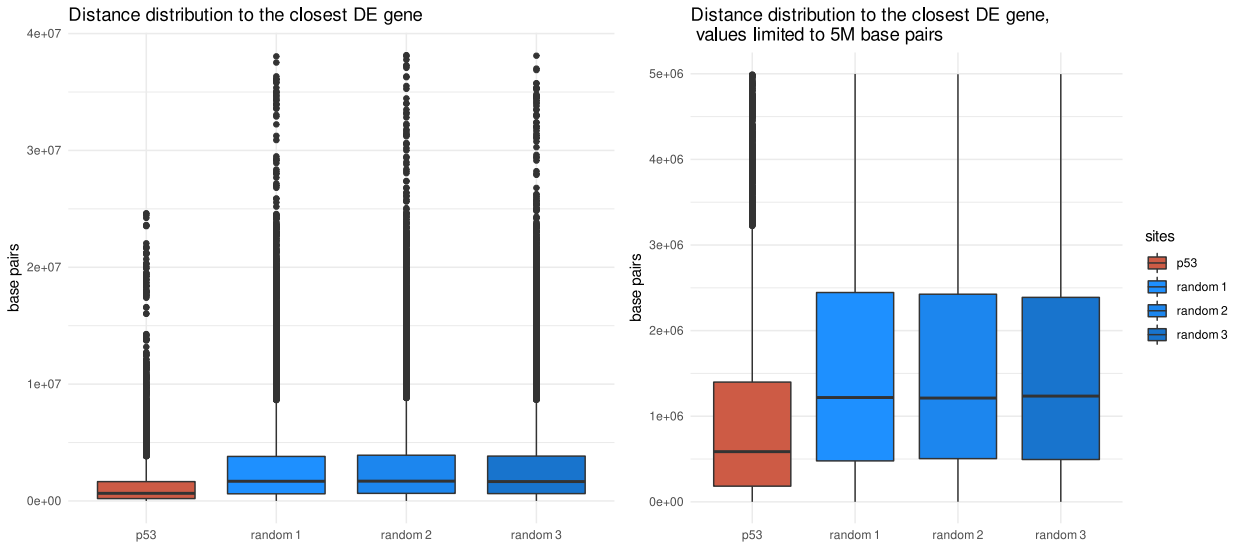


Figure 3.23: The p53 sites not located on DE genes are, on average, closer to one compared to the sites from the random samples.

3.5 Role of mutant TP53 on gene regulation control

In eukaryotic cells' nucleus, DNA is tightly packed in a compact and dense structure known as chromatin: a complex of DNA, RNA and proteins. The basic units in which DNA is packed in the chromatin are called nucleosomes and are composed of a segment of DNA spooled around a set of eight proteins known as *histones*. The histone proteins play an important role in the regulation of gene expression because the way DNA is coiled influences its exposure to the polymerase enzyme and, therefore, its transcription.

Histone proteins can undergo post-translational modifications that affect their interaction with DNA. Some modifications can weaken the histone-DNA interaction, causing the nucleosome to unwind and make the DNA more accessible to the binding of the translational machinery, enabling gene expression. Other modifications can strengthen the histone-DNA interaction, packing the chromatin more tightly and preventing gene expression. For these reasons, histone modifications are considered to be one of the powerful epigenetic mechanisms used by the eukaryotic organisms for cell differentiation [67].

In the previous sections we introduced how the use of bioinformatics tools could aid researchers studying the effects of a mutation on gene expression through direct protein binding. Gene expression, however, can also be indirectly regulated through chromatin alterations caused by histone modifications. We studied the relationships between active DNA regulatory regions and the previously estimated mutant p53 binding sites: the genes within reasonable distance from these regions that manifest expression changes upon TP53 silencing could be considered potential indirect regulatory targets of the mutant p53 protein.

As an estimate for the active regulatory regions on the genome, we considered the sites characterized by a histone modification H3K27ac, as these are often linked to actively transcripted genes [21]. In these sites, the H3 histone in the nucleosome is subject to an acetylation on its lysine amino acid in position 27.

One of the goals of the bioinformatics analyses described in this work is establishing if the employed quantitative approach (ranging from the p53 binding sites estimation to the differential expression analysis to the use of histone annotations) is capable enough in its objective of aiding researchers discover potential relationships between regulatory sites and genes. As a benchmark, we inspected if the studied relationship between p53, LINC01605 and its putative regulatory element, previously introduced in section 3.1, is detected and in what measure.

3.5.1 Histone annotation data

Li et al. [48], in 2019, generated genome-wide profiles of multiple histone modifications (including H3K27ac) in breast cancer cells to study the regulatory changes driven by chromatin remodeling in metastatic cells. The cell lines analysed in their study were MDA-MB-231 and LM2-4175, a variant showing more aggressive characteristics of invasion, migration and metastasis. Similarly to other protein binding analyses, profiling histone modifications can be performed through the ChIP-Seq protocol: the genome-wide enrichment results can then be compared against the control experiment to estimate histone modifications.

The raw reads files provided by Li et al. for the MDA-MB-231 cell line were aligned to the reference genome assembly and the resulting BAM files were used to estimate the sites characterized by

the H3K27ac modification using MACS2. A total of 58,201 H3K27ac histone modifications sites were estimated in MDA-MB-231 cells.

3.5.2 Direct effects of p53 through histone modifications

In the previous sections we presented and discussed the approach that we followed to study the effects of the mutant p53 protein on direct gene regulation. In order to characterize mutant p53's binding sites we explored whether p53's putative regions of binding are found in active DNA regulatory elements marked by the H3K27ac histone modification.

Out of the total 58,201 H3K27ac histone regions identified in MDA-MB-231 cells, 7256 (12.5%) contain one of the previously estimated 11,028 mutant p53 binding sites. This value is about 20 times greater compared to the number of histone regions located on the random sites (as can be seen in figure 3.24): the result can be explained considering that p53 is a transcription factor and it is reasonable to see binding enrichments in actively transcribed genome regions.

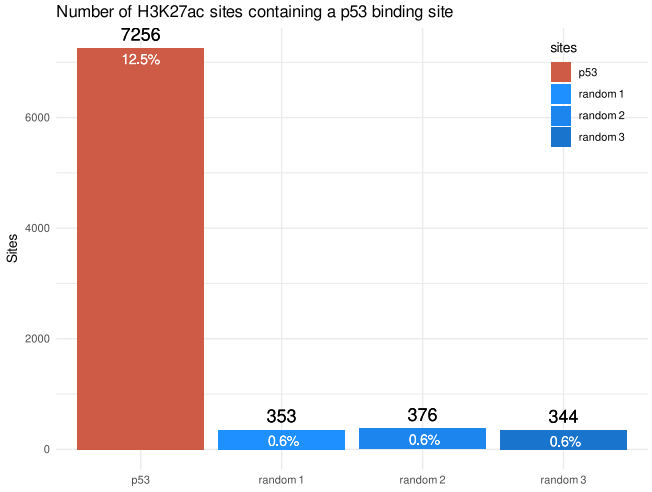


Figure 3.24: In the studied MDA-MD-231 cell line, the number of histone regions characterized by a H3K27ac modification containing a p53 binding site is about 20 times higher compared to the expected value for random sites.

We considered p53 binding sites located on histone regions marked by a H3K27ac modification (active transcription sites). For each of these, we looked at the closest differentially expressed gene (upon mutant TP53 silencing), as this could be a possible p53 regulatory target. Table 3.16 presents genomic regions characterized by the H3K27ac modification containing a mutant p53 binding site, along with their chromosomal coordinates. For each region, the name of the closest differentially expressed gene is given, along with the expression level change and its distance to the considered histone region.

	H3K27ac region		p53 binding site		closest gene to H3K27ac region		
chromosome	start	end	start	end	name	log2 Fold Change	distance (bp)
chr1	28544	29316	29116	29575	hsa-mir-6723	0.864	537139
chr1	713335	713984	713929	714211	MTATP6P1	0.656	-143580
chr1	1014271	1015674	1014812	1015132	ISG15	0.866	-64352
chr1	1014271	1015674	1015316	1015593	ISG15	0.866	-64352
chr1	1092942	1095047	1093046	1093487	ISG15	0.866	-143023
chr1	1243451	1243821	1243414	1243696	MXRA8	0.951	44249
chr1	1278294	1284694	1280127	1280630	MXRA8	0.951	3376
chr1	1309435	1310572	1310440	1310977	MXRA8	0.951	-12279
...							
chr8	37371529	37372762	37372718	37373264	**LINC01605**	-1.04	0
chr8	37373258	37378789	37372718	37373264	**LINC01605**	-1.04	0
chr8	37397879	37401370	37401290	37401836	**LINC01605**	-1.04	0
chr8	37401725	37403026	37401290	37401836	**LINC01605**	-1.04	0
...							

Table 3.16: Table showing the coordinates of the histone H3K27ac regions containing a p53 binding site and their closest differentially expressed gene. LINC01605 highlighted. Expression fold change from DESeq2 (shNT vs sh1).

These regions could represent putative DNA regulatory elements depending on mutant p53. As DNA regulatory elements could influence the expression of genes which are up to 1 Mb distant, we looked for DE genes within 1 Mb from each of the regions with a H3K27ac mark and a mutant p53 binding sites. For each histone H3K27ac region containing a mutant p53 binding site, table 3.17 presents its coordinates and the ratio of differentially expressed genes within 1 Mb (in upstream or downstream direction).

The same regions are graphically represented in figure 3.25: each circle indicates one of the H3K27ac regions in table 3.17, their size is given by the relative number of DE genes within 1 Mb and the crosses represent the DE genes. The position on the axes indicates the position within the chromosome. This representation can be useful as a preliminary explorative analysis tool to drive research.

chromosome	start	end	genes within 1 Mb	DE genes within 1 Mb	ratio (%)
chr8	119448123	119452043	18	2	11.1
chr10	3818737	3819240	33	3	9.09
chr10	3827783	3829964	34	3	8.82
chr10	3842886	3859484	34	3	8.82
chr10	3892246	3897692	34	3	8.82
chr10	3928698	3939555	35	3	8.57
chr1	184019307	184020667	37	3	8.11
chr1	184020906	184023440	37	3	8.11
chr2	46843155	46844110	41	3	7.32
chr1	208060509	208063918	42	3	7.14
...					
chr8	37371529	37372762	59	1	1.69
chr8	37373258	37378789	59	1	1.69
chr8	37397879	37401370	60	1	1.67
chr8	37401725	37403026	60	1	1.67

Table 3.17: Table indicating the modified histone regions containing a mutant p53 binding site having a relatively high number of DE genes close by. The last four regions are located on LINC01605's locus.

3.5.3 Effects of p53 on indirect gene regulation

The mechanisms employed in the cells for the regulation of gene expression can be complex and are still under active research. Transcription factors can play a role in the regulation of gene expression by binding to DNA regulatory elements distant from their target gene. To possibly identify DNA regulatory regions bound by mutant p53 we considered the active histone regions (characterized by H3K27ac, a marker of active DNA regulatory elements) containing a p53 binding site that are being transcripted. These regions could be part of DNA sequences whose transcript has a regulatory effect on distant genes.

To identify putative DNA regulatory elements bound by mutant p53, we adopted the following approach. Starting from the active histone regions, represented by the modification H3K27ac, we looked for the ones whose expression changes upon TP53 silencing. Then, we selected those containing a p53 binding site. These differentially expressed regions with mutant p53 binding sites may represent mutant p53-dependent DNA regulatory regions. Finally, we looked at the closest differentially expressed genes to each of these regions as potential indirect mutant p53 regulatory targets.

Differential expression analysis was performed on the previously used RNA-Seq data to determine the active histone regions subject to changes in their expression levels upon TP53 silencing. The process of reads quantification (mapping the RNA-Seq reads to genomic features) required the generation of

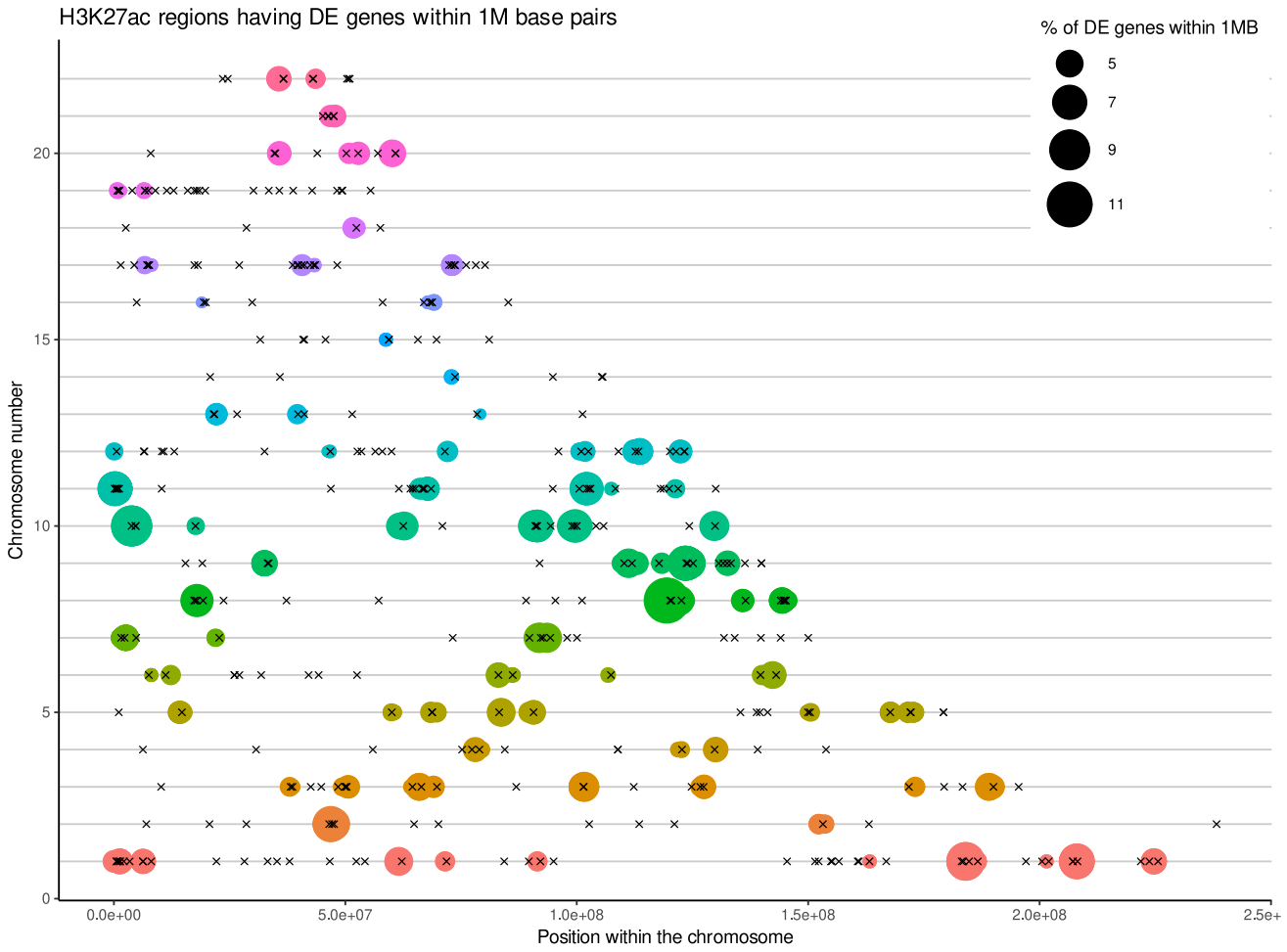


Figure 3.25: Circles are used to represent modified histones containing a p53 binding site: the size is directly proportional to the ratio of genes within 1 Mb that manifest a change in their expression level. Crosses indicate the location of the differentially expressed genes. The color, along with the vertical position, denotes the chromosome on which the histone is located. Only the regions with a ratio greater than 3% are shown.

a GTF annotation file containing all the 58,201 H3K27ac regions. Then, the three programs DESeq2, EdgeR and Limma were used to identify the H3K27ac DE regions. A total of 481 DE regions were found, 139 of which containing a p53 binding site.

For these selected regions we looked at the closest DE genes: tables 3.18 and 3.19 summarize the results.

	H3K27ac region		p53 binding site		closest gene to H3K27ac region		
chromosome	start	end	start	end	name	log2 Fold Change	distance (bp)
chr1	8117434	8124869	8121880	8122201	ERRFI1	1.34	-31066
chr1	16470653	16475648	16470266	16470716	EPHA2	0.652	0
chr1	16470653	16475648	16471582	16471859	EPHA2	0.652	0
chr1	16470653	16475648	16472631	16473059	EPHA2	0.652	0
chr1	20926806	20931866	20930943	20931325	CDA	1.20	0
chr1	20939196	20940439	20940323	20940771	CDA	1.20	0
chr1	22258042	22263759	22257668	22258052	HSPG2	1.03	0
chr1	38463939	38468777	38466117	38466466	ZC3H12A	0.760	-513962
...							

Table 3.18: Table showing the coordinates of the histone H3K27ac DE regions containing a p53 binding site and their closest differentially expressed gene. Expression fold change from DESeq2 (shNT vs sh1).

chromosome	start	end	genes within 1 Mb	DE genes within 1 Mb	ratio (%)
chr5	90676173	90679127	23	3	13.0
chr5	90607822	90610753	24	3	12.5
chr1	207222517	207226343	59	6	10.2
chr1	207189317	207200592	60	6	10
chr7	47618513	47622927	21	2	9.52
chr10	3818737	3819240	33	3	9.09
chr12	52638173	52641161	102	9	8.82
chr11	102213815	102220782	46	4	8.70
chr7	47329571	47333989	25	2	8
chr12	53296824	53298855	115	9	7.83
...					

Table 3.19: Table indicating the modified DE histone regions containing a p53 binding site that have a relatively high number of DE genes close by.

3.5.4 LINC01605 and its Putative Regulatory Element (PRE)

As introduced in section 3.1, researchers studying lncRNAs regulated by mutant TP53, identified LINC01605 as a possible target of mutant TP53 as its expression decreased upon TP53's silencing. To better understand how mutant TP53 could regulate LINC01605's expression, they identified a region (referred to as PRE) characterized by a mutant p53 binding site. PRE is considered a potential regulatory element because it is located on an active histone region and it is actively transcribed in the TP53-mutant cell line: its transcript could have a regulatory role on LINC01605. Most of these observations derive from manual inspections of the sequencing files through a visualization tool such as the Genome Browser.

While the currently ongoing project at CRO continues the study of the regulatory network involving PRE and LINC01605, we intended to examine the capabilities of bioinformatics tools in the discovery of genome-wide relationships between genes and histone regions such as the one just described. The results offered by the employed bioinformatics tools, restricted to the LINC01605 and PRE regions, were compared to the researchers' visual inspections.

Figure 3.26 shows the initial data available to the researchers. The region on the short green segment labeled LINC01605 (to the left) denotes an end of the LINC01605 gene, while the region on the short green segment labeled AC124067.2 (on the right) indicates the PRE region. The first two tracks from the top (the red and the gray ones) represent the aligned reads in the ChIP-Seq experiment (and in its control) and denote mutant p53's binding sites. The next six tracks concern the RNA-Seq experiments: the two green ones come from the non-TP53-silenced samples, while the next four are from the TP53-silenced ones. These tracks are used to inspect differences in the expression levels of the genes between conditions.

By looking at this data and at histone annotations, the researchers found:

- a binding site for the mutant p53 protein on the PRE region;
- a decrease in the expression level of LINC01605 upon TP53 silencing;
- a decrease in the expression level of PRE upon TP53 silencing;
- a histone modification H3K27ac on the PRE site.

Figure 3.26: Genome Browser visualization of the LINC01605 and PRE genomic intervals. The first two tracks from the top can be used to estimate p53's binding sites, the next six represent the changes in expression levels upon TP53 silencing. On the bottom, the first track (in green) is a genic annotation, the next one highlights the estimated p53 binding sites and the last locates the regions containing active histones.

The binding-sites estimation program MACS2 was able to detect the observed p53 binding site on the PRE region, as showed by the `macs2 tp53 binding sites` track on the bottom of the figure.

The decrease in the expression level of LINC01605 in the TP53-silenced samples was detected both by the reads quantification tool featureCounts (table 3.20) and by the three differential expression analysis methods DESeq2, EdgeR and Limma (table 3.21).

Next, the PRE region is located on an active histone region on the MDA-MB-231 cell line. Looking at the raw number of mapped reads assigned to the region, we noticed that the TP53-silenced samples have about half the number of reads compared to the non-silenced ones (table 3.22), in accordance with the researchers' observations. None of the DE methods, however, seem to consider such difference in expression statistically significant.

Overall, the quantification and expression analysis tools were able to detect the studied interaction between LINC01605 and PRE and the researchers were able to validate these results in the laboratory.

The main problem we encountered is the sensibility of the tested differential expression analysis methods which overlook regions characterized by a small number of reads, with the aim of reducing false positives.

shNT		sh1		sh2	
replica 1	replica 2	replica 1	replica 2	replica 1	replica 2
160	172	82	99	111	71

Table 3.20: Raw number of RNA-Seq reads mapped to the LINC01605 gene. The lower number of reads in the samples subject to TP53 silencing (sh1 and sh2) suggest a decrease in LINC01605's expression upon mutant TP53 silencing.

	shNT vs sh1		shNT vs sh2		
DE method	log2 Fold Change	adjusted p-value	log2 Fold Change	adjusted p-value	Considered DE?
DESeq2	-1.04	0.000108	-1.12	0.000277	Yes
EdgeR	-1.21	6.00e-10	-1.29	0.00000144	Yes
Limma	-1.21	0.0000127	-1.31	0.0000183	Yes

Table 3.21: All the three methods used for differential expression analysis report the decrease in LINC01605's expression level as statistically significant.

shNT		sh1		sh2	
replica 1	replica 2	replica 1	replica 2	replica 1	replica 2
14	16	8	8	7	9

Table 3.22: Raw number of RNA-Seq reads mapped to the histone region on PRE. The lower number of reads in the samples subject to TP53 silencing (sh1 and sh2) suggest a decrease in the expression of PRE upon mutant TP53 silencing.

3.6 Functional analysis of TP53

Making sense of the RNA-Seq results can be a complex task, particularly when dealing with long lists of genes. When the focus of the differential expression experiment is on gaining insights on molecular or biological pathways, fine grain details such as the expression level of single genes can be not particularly informative. Grouping thousands of genes or proteins by the pathways they are involved in reduces the complexity of the problem. Furthermore, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of genes or proteins. Summarizing the results grouping together genes according to their functions is therefore a meaningful step during data analysis.

The field of molecular biology that attempts to describe genes' functions and interactions is known as *functional genomics*. Its goal is to produce and organize genomic knowledge integrating data from different experimentations, improving the understanding of biological functions and processes. During the last two decades, following the surge of high-throughput sequencing, a profusion of both knowledge bases and computational and statistical methods have been developed to aid the researchers with this more and more common task [38].

In the next sections further details about the main knowledge bases and the principal functional analysis methods will be given: these tools will later be used to characterize mutant TP53's functions and to check the outcome of our study.

3.6.1 Gene Ontology

Among the annotations of genes' functions, Gene Ontology (GO) [86] is one of the most complete, structured and up-to-date. The elements in the ontology, called terms, refer to specific functions, processes and components of the cell: each term is marked by a code and a name. The terms are organized as nodes in a directed acyclic graph: the edges of the graph model the relationships that have been observed between the terms and can be of different types. One of such relationships is the specialization, which makes the ontology hierarchical: following the *is-a* edges, one can look at more generic or more specific terms and improve its understanding of the studied biological functions.

Figure 3.27 portrays a part of the ontology regarding the biological process *positive regulation of epithelial cell differentiation*, marked with code GO:0030858. In the example, the selected term is displayed at the bottom: black arrows are used to denote the *is-a* relationship to its child terms (more generic), while yellow ones indicate that the parent regulates the activity of the child term.

The terms in the Gene Ontology refer to specific activities performed by gene products and are partitioned into three classes[2].

- *Molecular function*: activities that occur at the molecular level, generally performed by individual gene products.
- *Biological process*: larger processes, accomplished by multiple molecular activities.
- *Cellular component*: locations relative to cellular structures in which a gene product performs a function.

[2]Gene Ontology classes: `http://geneontology.org/docs/ontology-documentation/`

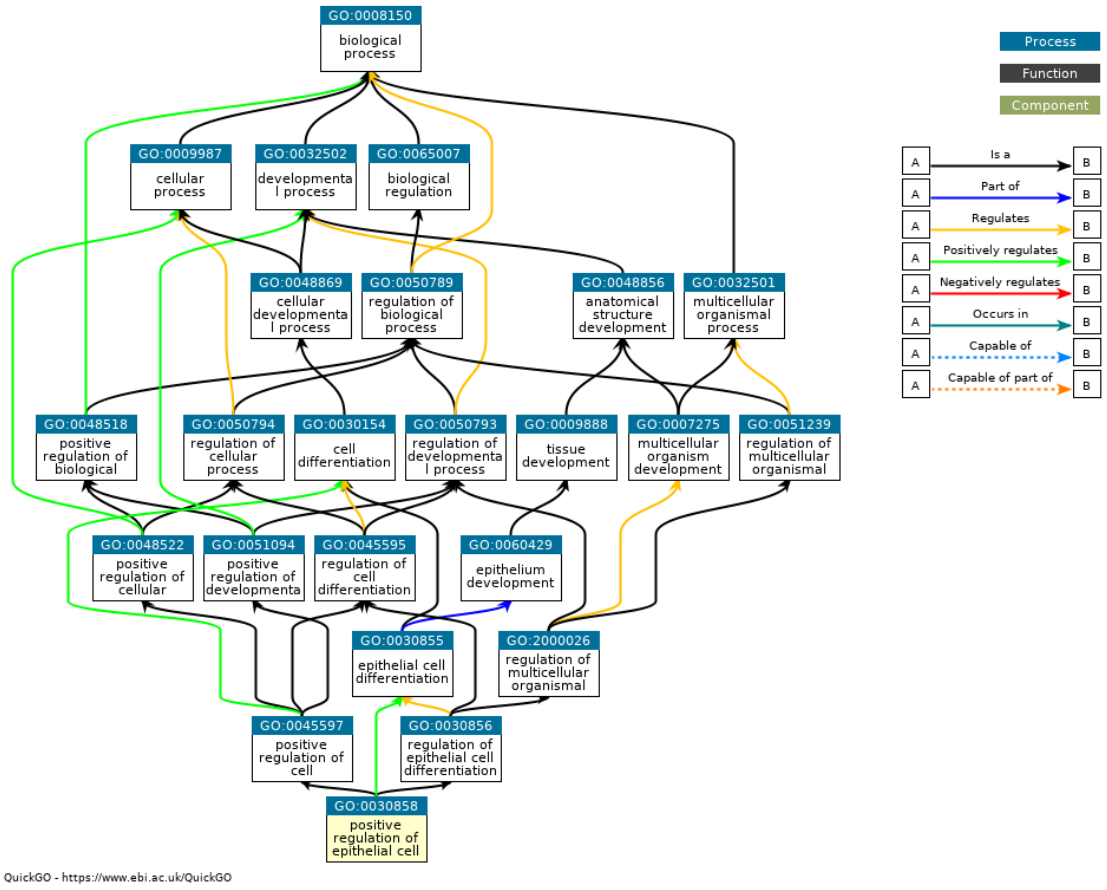


Figure 3.27: A part of the GO ontology regarding the term GO:0030858. The term "positive regulation of epithelial cell differentiation" is defined as both "positive regulation of cell differentiation" and "regulation of epithelial cell differentiation".

The *is-a* relationship is defined only within terms of the same class but other relationships, such as *part-of* and *regulates*, can operate between classes.

The complete ontology can be downloaded as a text file from the official website[3]. Other sites, such as QuickGO[4], provide useful web-based interfaces to inspect the terms, explore their relationships and look at frequently co-occurring terms. The Gene Ontology annotation is produced by the GO Consortium integrating the knowledge provided by a variety of research groups. Each association between a gene product and a GO term must be supported by an evidence, which can be inferred by an experiment or by phylogenetic or computational analyses[5]. The website AmiGO[6] can be used to explore such associations and evidences (figure 3.28).

3.6.2 Main functional analysis methods

As extensively explained by Khatri et al. [38], during the last two decades plenty of statistical and computational functional analysis methods have been developed, many of which try to provide a sound and extensive data analysis pathway. These methods vary in many aspects, such as the type of biological annotation being used, the underlying hypotheses on gene products' interactions, the statistical modeling, the expected input data and the extent of the provided explanation.

[3]Gene Ontology downloads: `http://geneontology.org/docs/download-ontology/`
[4]QuickGO website: `https://www.ebi.ac.uk/QuickGO/`
[5]Gene Ontology evidence codes: `http://geneontology.org/docs/guide-go-evidence-codes/`
[6]AmiGO website `http://amigo.geneontology.org`

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Type	Isoform	Reference	Date
lhx1a	LIM homeobox 1a		positive regulation of nephron tubule epithelial cell differentiation		UniProt	Danio rerio	ISS	UniProtKB:P63006	lim/homeobox protein lhx pthr24208	protein_coding_gene		ZFIN:ZDB-PUB-110105-1	20110715
ins	preproinsulin		positive regulation of pancreatic A cell differentiation		ZFIN	Danio rerio	IMP	ZFIN:ZDB-MRPHLNO-160331-1	insulin/insulin growth factor pthr11454	protein_coding_gene		PMID:26658317 ZFIN:ZDB-PUB-151216-14	20170919
tal1	T-cell acute lymphocytic leukemia 1		positive regulation of endothelial cell differentiation		ZFIN	Danio rerio	IGI	ZFIN:ZDB-GENO-980202-118	t-cell acute lymphocytic leukemia/stem cell leukemia-related pthr13864	protein_coding_gene		PMID:9499398 ZFIN:ZDB-PUB-980415-5	20031121

Figure 3.28: The website AmiGO provides evidences for the associations between a GO term and the products of its genes. The figure shows the association evidences between three genes and the GO term GO:0030858.

Over Representation Analysis, Panther

Over Representation Analysis, ORA in short, is the name used to refer to the first class of functional analysis methods developed: these methods statistically evaluate the fraction of genes in the input list belonging to a gene set (such as a GO term) against the expected number defined under the hypothesis that the tested gene set is not affected by the experimentation.

The application of ORA requires the preliminary production of an input list of differentially expressed genes. This step requires the usage of an external program, such as DESeq2, to estimate the changes in expression of the individual genes and the choice of thresholds for fold-change and statistical significance for filtering the gene list. Then, for each gene set in the collection, the number of genes in the input list belonging to such set is counted and a test of statistical significance is performed, comparing this count to the expected value, obtained from the sizes of the input gene list and of the considered gene set.

Many tools implement the ORA approach and differ from one another for the annotation considered and for the test of statistical significance used for the comparison. One such tool is *Panther* [57], which uses the Gene Ontology annotation and a binomial p-value test.

More in detail, the p-value for the over-representation test is computed as follows. Given a total of N genes in the complete reference list and a considered gene set C of size $n(C)$, it is expected that, under the hypothesis that the gene set C is not over-represented in the input list (null hypothesis), about $p(C) = n(C)/N$ of the input list of DE genes will be contained in such gene set. Given K the size of the input list of DE genes and $k(C)$ the observed number of genes within the input list mapped to category C, the probability of having observed a value at least as extreme as $k(C)$ (i.e. the p-value of the test), is given by formula 3.3. Under the null hypothesis, the number of genes in the input list mapped to C is distributed binomially with probability parameter $p(C)$.

$$p\text{-value} = \sum_{k=k(C)}^{K} \binom{K}{k} \cdot p(c)^k \cdot (1 - p(c))^{K-k} \tag{3.3}$$

For instance, given a total of $N = 20000$ genes, an input list of $K = 500$ DE genes, a GO term containing $n(C) = 650$ genes, under the null hypothesis the most likely number of genes (expected value) in the input list mapped to the GO term is $16 = \text{round}(500 \cdot 650/20000)$. Having observed a number of

$k(C) = 30$ genes in the input list mapped to the GO term, the probability that this is due to randomness (and not to a significant change regarding such term in the experiment) is less than $p = 0.0012$.

Despite its widespread usage, Over Representation Analysis has some limitations. To begin with, the value of the computed statistic is independent of the measures associated to the genes: the method looks only at the names of the genes in the input list and ignores other meaningful features such as their fold change, each gene in the list is treated equally. Another shortcoming derives from the fact that, since the application of the method requires an input gene list, only the most significant genes are used, while all the others are discarded. For instance, by setting a fold change threshold the genes marginally below such threshold will be discarded. If there are some functions or processes, however, that interest many of these genes, these will be ignored.

Functional Class Scoring, GSEA

The statistical methods based on the *Functional Class Scoring* approach, FCS in short, aim at overcoming most of the limitations inherent in the tools based on the Over Representation Analysis. This approach is based on the hypothesis that, although large expression changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes can also have significant effects. First, these tools compute some gene-level statistics (such as the fold change between conditions and a measure of its significance) on the genes in the collection, similarly to what standalone differential expression analysis methods, such as DESeq2 or EdgeR, do. Then, the statistics for the genes in the same gene set, such as those involved in the same biological process, are aggregated into a single measure: for this step different tools usually choose different statistics [38] (such as the mean or more complex ones).

The functional analysis methods based on the FCS approach do not require setting arbitrary thresholds for dividing expression data into significant and non-significant pools and use all the available molecular measurements in order to detect coordinated changes in the expression of genes in the same set. For these reasons, they can be considered more robust than those based on ORA.

GSEA, short for *Gene Set Enrichment Analysis* [84], is a functional analysis method based on the FCS approach. GSEA aims at estimating if the genes within a given gene set (such as those belonging to the same GO term) are significantly represented in the differential expression analysis results.

First, all the genes in the collection are sorted by their fold change in the experiment, producing a ranked list L. Then, the genes in the considered gene set S are located in list L. If the gene set S is subject to significant changes between experimental conditions, then its genes will be concentrated in the top or in the bottom of L. In figure 3.29

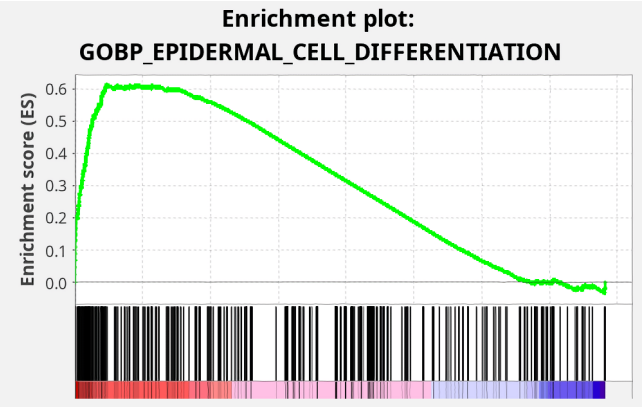


Figure 3.29: Enrichment score and barcode plot highlighting that the genes in the "Epidermal cell differentiation" GO term are among those subject to major increases in expression between the considered experimental conditions.

the barcode represents horizontally the ranked gene list of an experiment. Narrow black vertical lines indicate the genes within the gene set *Epidermal cell differentiation*: it can be seen that these genes are mainly concentrated on the left of the bar (the top of the ranked list), meaning that many of these genes saw an increase in expression between experimental conditions.

The *enrichment score* (ES) is a numerical measure that reflects the degree to which a gene set S is over-represented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the list L, increasing a running sum statistic when the encountered gene is in S and decreasing it when it is not. The enrichment score for the gene set is the maximum value seen by the running sum. Finally, a statistical test is used to estimate the significance of the computed enrichment score.

GSEA can work with many gene set collections, including Gene Ontology, Hallmark[7] and C2[8]. The Hallmark gene set collection [51] aims at summarizing the broad knowledge available in other collections into just 50 sets, minimizing redundancy and gene overlap. C2, vice versa, is an extensive collection of over 6000 gene sets from various sources, including biomedical literature and biological pathways databases.

Overview of some functional analysis tools

In our analyses, we considered various functional analysis tools in order to find the ones best suited for our needs. The most popular tools include *David* [32], *Panther* [57], *GSEA* [84], *Viseago* [9] and *Revigo* [85]. Table 3.23 highlights characteristics, qualities and shortcomings of each, from the point of view of our analyses. The problems of clustering and visualization will be discussed in section 3.7.

	David	Panther	GSEA	Viseago	Revigo
type	ORA	ORA	FCS, visualization	FCS, visualization, clustering	visualization, clustering
qualities	support for many annotations	up-to-date GO annotation	support for many annotations	good visualizations and clustering techniques	flexibility
shortcomings	outdated annotations	only GO annotation supported	requires gene counts in input, flexibility	visualizations and clustering only applicable to own analysis results	limited clustering metrics

Table 3.23: Main qualities and drawbacks of some functional analysis tools used in the analyses.

3.6.3 Biological functions affected by mutant TP53

Extensive studies on p53 have confirmed its key role in tumor suppression: through the binding to its target genes it regulates many critical biological processes, including apoptosis, cell cycle arrest, senescence, DNA repair and cell metabolism which contribute to p53's function in tumor suppression [99].

Mutations occurring in TP53's DNA-binding domain often impairs p53's ability to bind to DNA-binding elements in its target genes and thus its transcriptional activity. Other than loss of wild-type

[7]Hallmark gene set collection: `http://www.gsea-msigdb.org/gsea/msigdb/collection_details.jsp#H`
[8]C2 gene set collection: `http://www.gsea-msigdb.org/gsea/msigdb/collection_details.jsp#C2`

p53's functions, mutations on TP53 can promote tumor progression through gain-of-function mechanisms. Various mutant p53 gain-of-function mechanisms have been reported so far, including promoting cell proliferation, metastasis, genomic instability, metabolic reprogramming, immune suppression, and resistance to therapy in cancer [99].

Functional analysis gives us the ability to associate the results of differential expression experiments to biological functions and, thus, estimate the ones that are most affected between experimental conditions. We employed GSEA on the C2 and Hallmark gene set collections to confirm the outcome of the RNA-Seq experiment and of the subsequent data analysis pipeline.

Assessing the biological functions affected by mutant TP53

We performed GSEA's functional analysis (using the C2 gene set collection) on TP53 sh1 RNA-Seq data: figure 3.30 presents part of the results. The gene sets in the results describe the biological functions involving the genes that exhibit changes in expression upon mutant TP53's silencing. Most of these functions are known gain-of-function tumor promoting mechanisms acquired by the mutant TP53 gene.

Among these mechanisms, *epithelial-mesenchymal transition (EMT)* is a process promoting metastasis and is advanced by the upregulation of the ZEB1 transcription factor [99] (lines 1, 4, 7, 8, 10, 18, 20 in figure 3.30); tumor metastasis is also promoted by mutant TP53 through the upregulation of EGFR [60, 99] (line 23). Other gene sets refer to experimental observations of genes upregulated or downregulated in breast cancer cells (lines 3, 11, 19, 27).

Figure 3.31 presents part of the results of GSEA's functional analysis (using the Hallmark gene set collection) applied on the data on TP53's sh2 silencing. The gene set `TNFA_signaling_via_NFKB` (line 1) refers to the effects of the activity of the protein complex NF-κB on the tumor necrosis factor alpha. NF-κB plays a key role in regulating the immune response to infection, but an incorrect regulation of NF-κB has been linked to cancer and improper immune system development [96, 99]. One of the observed effects in which mutant TP53 affects the immune response of the targeted system is through the inhibition of tumor-suppressive interferons [54] (lines 2, 3, 4, 5, 11). Mutations on the KRAS gene (which regulates cell division and proliferation) have also been observed in tumor cells [99].

The lower part of the figure also confirms that the main effects caused by the mutant TP53 concern cancer development. The genes regulating the E2F transcription factors, which play a key role in the control of cell cycle, and those involved in the G2M DNA damage checkpoint (important to avoid the replication of damaged cells) are among the most affected by mutant TP53.

Functions regulated by TP53 through histone binding

Proteins can affect gene expression in different ways, such as through direct gene binding or via interactions with distant DNA regulatory elements. In order to estimate the biological functions potentially regulated by mutant p53 through histone binding, we considered the Gene Ontology terms exclusive to the differentially expressed genes within 50 kb from a H3K27ac histone region characterized by p53 binding.

Panther was used to estimate the enriched GO terms in both the full set of DE genes and in the subset of DE genes within 50 kb from a modified histone and the terms exclusive to the second set are

listed below. These terms relate to the *MAPK signaling pathway*, which is involved in the regulation of key cellular processes such as proliferation, differentiation, apoptosis and stress response [100, 30].

```
GO:0019220 regulation of phosphate metabolic process
GO:0051174 regulation of phosphorus metabolic process
GO:0032872 regulation of stress-activated MAPK cascade
GO:0051403 stress-activated MAPK cascade
GO:0070302 regulation of stress-activated protein kinase signaling cascade
GO:0048585 negative regulation of response to stimulus
```

Code 3.2: GO terms exclusive to the differentially expressed genes within 50 kb from a H3K27ac histone containing a mutant p53 binding site.

	GS follow link to MSigDB	SIZE	ES	NES	NOM p-val	FDR q-val
1	CHARAFE_BREAST_CANCER_BASAL_VS_MESENCHYMAL_UP	123	0.81	2.40	0.000	0.000
2	COLDREN_GEFITINIB_RESISTANCE_DN	228	0.76	2.40	0.000	0.000
3	CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	451	0.72	2.39	0.000	0.000
4	HOLLERN_EMT_BREAST_TUMOR_DN	122	0.80	2.36	0.000	0.000
5	LIN_SILENCED_BY_TUMOR_MICROENVIRONMENT	107	0.77	2.26	0.000	0.000
6	ONDER_CDH1_TARGETS_2_DN	472	0.66	2.20	0.000	0.000
7	AIGNER_ZEB1_TARGETS	34	0.90	2.19	0.000	0.000
8	JAEGER_METASTASIS_DN	256	0.69	2.19	0.000	0.000
9	ANDERSEN_CHOLANGIOCARCINOMA_CLASS2	175	0.71	2.16	0.000	0.000
10	BOSCO_EPITHELIAL_DIFFERENTIATION_MODULE	69	0.79	2.15	0.000	0.000
11	MCBRYAN_PUBERTAL_BREAST_4_5WK_UP	268	0.65	2.08	0.000	0.000
12	REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	129	0.70	2.07	0.000	0.000
13	RASHI_RESPONSE_TO_IONIZING_RADIATION_2	124	0.69	2.07	0.000	0.000
14	WAMUNYOKOLI_OVARIAN_CANCER_LMP_UP	268	0.65	2.06	0.000	0.000
15	DIRMEIER_LMP1_RESPONSE_EARLY	62	0.77	2.06	0.000	0.000
16	REACTOME_KERATINIZATION	217	0.66	2.06	0.000	0.000
17	BLANCO_MELO_HUMAN_PARAINFLUENZA_VIRUS_3_INFECTION_A594_CELLS_DN	121	0.70	2.04	0.000	0.000
18	RICKMAN_METASTASIS_DN	260	0.64	2.04	0.000	0.000
19	NIKOLSKY_BREAST_CANCER_16P13_AMPLICON	119	0.69	2.03	0.000	0.000
20	SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_DN	144	0.67	2.02	0.000	0.000
21	NAGASHIMA_NRG1_SIGNALING_UP	172	0.65	2.01	0.000	0.000
22	MISSIAGLIA_REGULATED_BY_METHYLATION_UP	125	0.68	2.01	0.000	0.000
23	KOBAYASHI_EGFR_SIGNALING_24HR_UP	101	0.69	2.00	0.000	0.000
24	DUTERTRE_ESTRADIOL_RESPONSE_6HR_DN	91	0.70	1.99	0.000	0.000
25	BLANCO_MELO_RESPIRATORY_SYNCYTIAL_VIRUS_INFECTION_A594_CELLS_DN	119	0.68	1.99	0.000	0.000
26	ZHANG_RESPONSE_TO_IKK_INHIBITOR_AND_TNF_UP	223	0.63	1.99	0.000	0.000
27	SMID_BREAST_CANCER_ERBB2_UP	151	0.65	1.98	0.000	0.000
28	KIM_RESPONSE_TO_TSA_AND_DECITABINE_UP	135	0.66	1.97	0.000	0.000
29	DELPUECH_FOXO3_TARGETS_UP	65	0.72	1.96	0.000	0.000
30	FURUKAWA_DUSP6_TARGETS_PCI35_UP	72	0.72	1.96	0.000	0.000

Figure 3.30: First 30 gene sets, from the C2 collection, enriched in the TP53 sh1 silencing experiment, according to GSEA. Most of the affected gene sets concern biological functions commonly altered during cancer development.

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val
1	HALLMARK_TNFA_SIGNALING_VIA_NFKB	Details ...	200	0.77	2.36	0.000	0.000
2	HALLMARK_INTERFERON_ALPHA_RESPONSE	Details ...	97	0.78	2.20	0.000	0.000
3	HALLMARK_INTERFERON_GAMMA_RESPONSE	Details ...	200	0.71	2.18	0.000	0.000
4	HALLMARK_INFLAMMATORY_RESPONSE	Details ...	200	0.67	2.04	0.000	0.000
5	HALLMARK_IL6_JAK_STAT3_SIGNALING	Details ...	83	0.70	1.96	0.000	0.000
6	HALLMARK_KRAS_SIGNALING_UP	Details ...	198	0.56	1.73	0.000	0.000
7	HALLMARK_ALLOGRAFT_REJECTION	Details ...	200	0.57	1.72	0.000	0.000
8	HALLMARK_ESTROGEN_RESPONSE_EARLY	Details ...	200	0.55	1.70	0.000	0.000
9	HALLMARK_HYPOXIA	Details ...	200	0.55	1.69	0.000	0.000
10	HALLMARK_COAGULATION	Details ...	137	0.57	1.68	0.000	0.001
11	HALLMARK_IL2_STAT5_SIGNALING	Details ...	198	0.52	1.59	0.000	0.002
12	HALLMARK_COMPLEMENT	Details ...	200	0.52	1.59	0.000	0.002
13	HALLMARK_ESTROGEN_RESPONSE_LATE	Details ...	200	0.51	1.58	0.000	0.003
14	HALLMARK_APICAL_JUNCTION	Details ...	198	0.50	1.51	0.003	0.007
15	HALLMARK_APOPTOSIS	Details ...	161	0.51	1.51	0.000	0.007
16	HALLMARK_KRAS_SIGNALING_DN	Details ...	200	0.49	1.51	0.002	0.007
17	HALLMARK_P53_PATHWAY	Details ...	200	0.45	1.39	0.006	0.031
18	HALLMARK_UV_RESPONSE_UP	Details ...	158	0.46	1.39	0.010	0.029
19	HALLMARK_MYOGENESIS	Details ...	200	0.45	1.39	0.011	0.028
20	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	Details ...	199	0.44	1.35	0.011	0.043

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val
1	HALLMARK_E2F_TARGETS	Details ...	200	-0.77	-2.81	0.000	0.000
2	HALLMARK_G2M_CHECKPOINT	Details ...	200	-0.69	-2.51	0.000	0.000
3	HALLMARK_MYC_TARGETS_V1	Details ...	200	-0.67	-2.43	0.000	0.000
4	HALLMARK_MYC_TARGETS_V2	Details ...	58	-0.49	-1.46	0.013	0.032
5	HALLMARK_MITOTIC_SPINDLE	Details ...	199	-0.41	-1.46	0.000	0.028
6	HALLMARK_OXIDATIVE_PHOSPHORYLATION	Details ...	199	-0.39	-1.41	0.000	0.036
7	HALLMARK_DNA_REPAIR	Details ...	150	-0.40	-1.39	0.006	0.036

Figure 3.31: Some of the gene sets, from the Hallmark collection, most enriched in the TP53 sh2 silencing experiment, according to GSEA. Most of the affected gene sets concern biological functions commonly altered during cancer development.

3.7 LINC01605's role in TP53 regulatory functions

The TP53 gene is known to control many biological functions, particularly regarding cell growth, but some of these regulatory mechanisms are currently still unknown. As described in section 3.1, researchers at CRO identified LINC01605 as a putative regulatory target of the mutant TP53 gene and, to broaden the study of this sequence, the interest is in considering the indirect regulatory effects TP53 could carry out through the expression changes it induces on LINC01605.

In order to inspect this relationship, we intend to compare the biological functions and processes affected by LINC01605 and TP53: in particular, the functions influenced by both genes could suggest indirect regulatory effects that mutant TP53 exerts through LINC01605, and their analysis could lead further experimentations.

3.7.1 LINC01605 knockout experiment

In order to estimate the functions regulated by LINC01605, researchers at CRO produced LINC01605 *knockout* samples from the cell line MDA-MB-231. Four replicates of both knockout and wild-type cell samples were generated and the resulting sequences were aligned to the release version GRCh37 of the human genome.

Gene knockout is a term used to indicate a set of techniques used to disable one or more genes of an organism. These techniques operate at the transcription level, as opposed to other silencing techniques such as *RNA interference* which do not prevent the transcription of the gene but, instead, try to limit its translation into proteins.

The techniques based on gene knockout can be more effective than those operating at the translation level: in our experiments, shRNA silencing reduced TP53's expression by about 3 times, while gene knockout nearly prevented LINC01605's expression, as can be observed in figure 3.32.

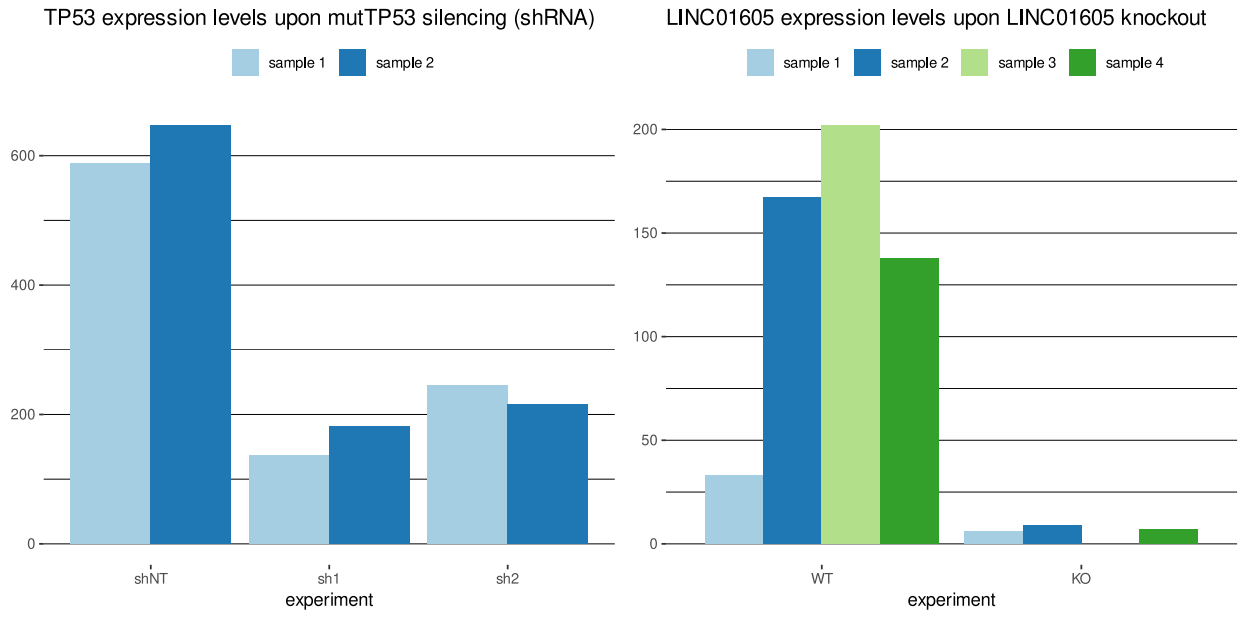


Figure 3.32: Gene knockout is, in our experiments, much more effective at preventing gene expression compared to silencing methods based on RNA interference.

In order to study the functions that mutant TP53 could regulate through LINC01605, we decided to first look separately at the differentially expressed genes in the two experiments and to link these genes to the biological functions and processes they are involved in, taking advantage of the Gene Ontology annotation. The relevant gene sets in the two experiments were, then, compared and the ones enriched in both were considered as potential mutant p53's targets through LINC01605's regulation.

The decision to consider the commonly enriched gene sets in the two experiments, instead of focusing on the common differentially expressed genes, was made to give more importance to the biological functions in which the genes are involved, as a gene set (representing a biological function or process) could be enriched in both experiments even if there are no common DE genes. Furthermore, as figure 3.33 shows, the two experiments on TP53 and LINC01605 share a low number of differentially expressed genes, so looking only at these could lead to an over-approximated analysis.

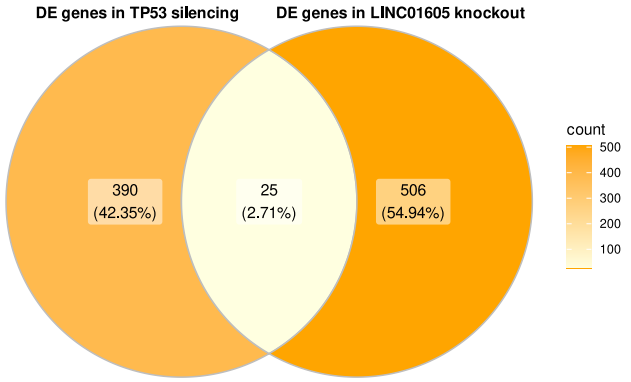


Figure 3.33: Number of genes considered differentially expressed by DESeq2 in the TP53 silencing (genes DE in both sh1 and sh2) and in the LINC01605 knockout experiments.

3.7.2 Challenges in the comparison of different functional analysis tools

As introduced in section 3.6.2, many functional analysis methods and tools have been developed, often relying on different hypotheses. The main ones we focused on in our analyses are GSEA and Panther and, to better interpret the results they provide, we looked at ways of comparing lists of gene sets. In particular, we considered the problem of comparing lists of GO terms as Gene Ontology is one of the gene set collections most widely used.

Comparing and computing a measure of similarity between two lists of GO terms is not necessarily a trivial problem. A lexical comparison of the content of the lists could be an excessively simple approach as two sets of terms can refer to the same, or similar, biological functions even without strictly sharing terms.

The usage of a semantic measure of terms similarity surely helps but the abundance of available heuristics [92, 64, 102] shifts the problem in finding the most reliable one. Most semantic similarity measures rely on either the structure of the ontology (the nodes and edges in the GO graph), on external factors (such as the size of the GO terms) or on a combination of both.

One of the first similarity measures proposed [69] used the distance between two nodes in the GO graph (which represent GO terms), intended as the number of edges along their shortest path. The assumption, however, that all semantic links have the same weight frequently doesn't hold: for instance, edges connecting broad and general terms (such as *transporter activity* and *binding* in figure 3.34) denote less semantic similarity than those connecting very specific terms. To correct this problem, other measures have been defined which assign edges different weights to reflect some degree of hierarchical depth [64].

Even with these corrections, the methods based on the graph structure rely heavily on the assump-

tions that nodes and edges in an ontology are uniformly distributed and that nodes at the same level correspond to the same semantic distance, which are untrue in the case of GO: terms having same depth can have different specificity, such as *oxygen binding* and *ion binding* in figure 3.34.

The methods based on external factors, instead, do not depend on the number or depth of the edges but make use of information-theoretic principles. In general, every term is assigned an information content (IC) measure, which can be computed from the size of the terms: bigger terms (containing more genes) will be considered less specific and will have a lower IC measure. Some methods define the similarity of two terms by looking at the information content of their most informative common ancestor: if this value is relatively low then the compared terms will probably be very unspecific or they will have a distant common ancestor.

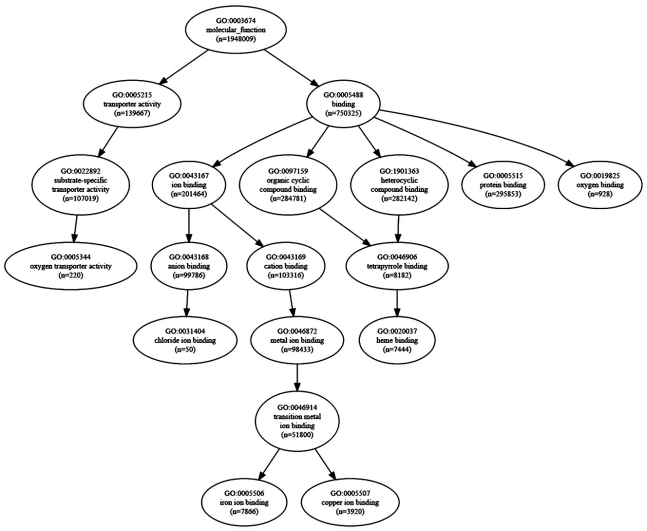


Figure 3.34: Illustrative example of a portion of the Gene Ontology graph [64].

The main GO semantic similarity measures are based on these approaches but can differ substantially since they consider different relationships (for instance, some measures only consider the *is-a* edge, while others also use *regulate*) and they implement different weights or heuristics to correct the measure.

A metric for comparing lists of GO terms

In this section we present a simple, and for some reasons limited, metric for the comparison of two lists of GO terms, such as those produced by function analysis tools. The method tries to optimistically compare the terms using semantic information and comprises the following steps.

1. First, the terms in the two compared lists are clustered by semantic similarity, simplifying the lists in clusters of closely related terms.

2. Then, each cluster in the first list is analysed checking whether it is *shared* with the second list, or if it is *exclusive*. A cluster is optimistically considered *shared* if at least one of its terms is also present in the second list, *exclusive* otherwise (figure 3.35).

3. Each of the *exclusive* clusters in the first list is then compared to the clusters in the second list and the similarity score to its most similar cluster in the second list is recorded. At the end all the recorded scores are summed (figure 3.36).

4. Finally, the similarity score of the first list with respect to the second one is given by the number of *shared* clusters plus the sum of the recorded scores, all divided by the total number of clusters in the first list.

The similarity of two clusters of GO terms can be computed starting from the similarity of single GO terms (using one of the methods introduced in section 3.7.2) leveraging one of the heuristics defined

as *linkage criteria* in the context of hierarchical clustering, such as the minimum, maximum or the mean similarity between pairs of terms in the compared clusters.

The measure we used to compute the similarity of two clusters is the mean of the maximums (each element in the first cluster is associated to the most similar element in the second one and then these maximum similarities are averaged) which cannot be considered a linkage criterion because it is not symmetrical but, in our case, not even the complete metric is so. We experimented with various cluster similarity measures and, since the clusters we are working with are small (usually they contain $1 - 4$ terms), the difference is negligible.

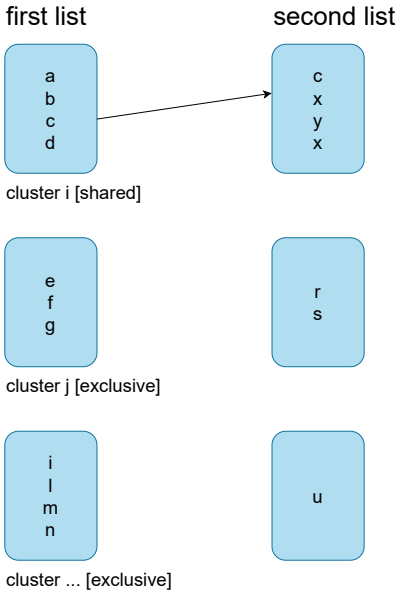


Figure 3.35: The cluster i *in the first list is considered "shared" because one of its terms is also present in the second list.*

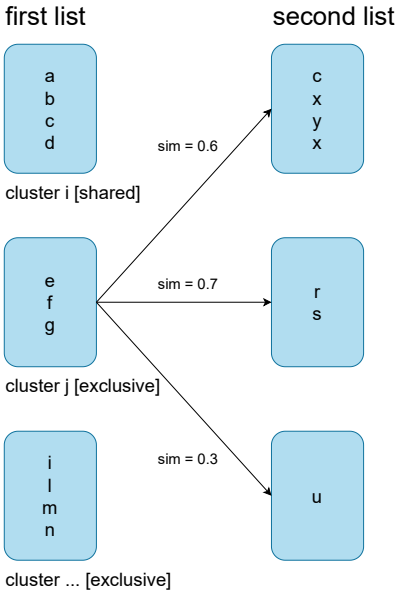


Figure 3.36: Each "exclusive" cluster is compared to all the clusters in the second list.

The metric scores are normalized in the interval [0,1]. Score 0 means that the terms in the first list are maximally different from the ones contained in the second list: every cluster is exclusive and has zero similarity to each cluster in the second list. Score 1 means that every cluster in the first list is shared in the second one, while scores close to 1 could indicate that most of the exclusive clusters are paired to similar clusters in the second list.

One of the shortcomings of the metric is its asymmetry: given two lists a and b, score(a, b) is generally not equal to score(b, a). This problem could be resolved taking the minimum or the maximum of the two scores. Another shortcoming of this method is that the score could be affected by the size of the lists if the dimension of the clusters is constant: considering an exclusive cluster in the first list, the probability of finding a similar cluster in the second list increases with the size of such list.

Application of the metric to experiment data

A decisive problem one may encounter when comparing the results of various tools is the selection of a significance threshold: different tools may use different statistical tests (i.e. compute the p-value in different ways), use the p-value to indicate widely diverse information (such as Panther and GSEA use distinct methods) and apply different corrections. We used the same uncorrected p-value threshold (1×10^{-4} for the TP53 silencing and 1×10^{-3} for the LINC01605 knockout, respectively) with all the

considered tools.

The program Revigo [85] was employed to cluster the GO terms in the lists produced by the considered tools, using the options `Allowed similarity = 0.5` and `Semantic similarity measure = SimRel`. The `SimRel` similarity measure was also used to compute the similarity between the clusters of GO terms (through the R package GoSemSim [98]) as it was the one that provided the most consistent values among all the measures tested.

The functional analysis tools clusterProfiler, Panther and GSEA were applied to the data about the TP53 sh1 silencing experiment. The first two received the differentially expressed genes estimated by DESeq2, while GSEA used its own estimation method. Table 3.24 reports the number of enriched GO terms identified by these tools, along with the number of clusters they are grouped into. Figures 3.37 and 3.38 picture the results of the application of the described comparison metrics to the outputs of the three employed tools and to two random lists of GO terms.

	Biological process		Molecular function	
	GO terms	clusters	GO terms	clusters
clusterProfiler	142	65	12	7
Panther	345	162	23	18
GSEA	350	169	37	28
Random sample 350	240	160	90	57
Random sample 200	127	104	48	35

Table 3.24: Number of significant GO terms considered enriched by the employed functional analysis tools (in the TP53 sh1 silencing experiment). These were compared to two random samples of 350 and 200 GO terms. Panther and clusterProfiler were applied on the DE genes reported by DESeq2.

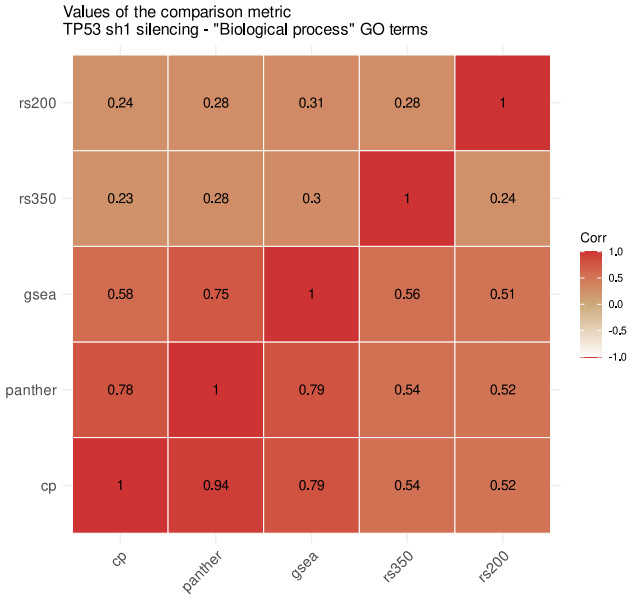


Figure 3.37: Application of the pairwise comparison metric to the results of clusterProfiler, Panther, GSEA and to two random samples of GO terms. Only the GO terms of type "Biological process" are considered.

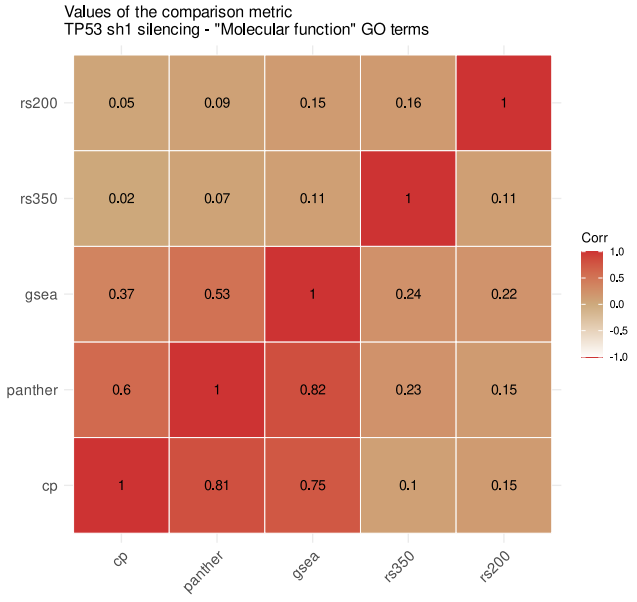


Figure 3.38: Application of the pairwise comparison metric to the results of clusterProfiler, Panther, GSEA and to two random samples of GO terms. Only the GO terms of type "Molecular function" are considered.

The following aspects of the comparison metric can be observed from the experimental results.

- Panther's output is most similar to clusterProfiler's (as they are based on the same method), the similarity of the lists produced by the functional analysis tools is stronger compared to the one of the random lists of GO terms.
- The asymmetry of the metric (the difference between score(a, b) and score(b, a)) increases with the size disparity of the compared list: the score is usually higher when a short list is compared to a longer one. This could be explained by the fact that longer lists grant a higher probability of finding a high-similarity cluster.
- The similarity with random lists is higher when longer lists are considered (as can be observed comparing figures 3.37 and 3.38). This is caused by the second part of the score (given by the sum of the similarities of the exclusive clusters) which increases with the size of lists.

3.7.3 LINC01605 and TP53's functional analysis

Among all the functional analysis tools tried, we found that the plots generated by Revigo [85] and Viseago [9] best summarize the analysis results. In particular, Viseago's multi-level hierarchical clustering groups relevant Gene Ontology terms by semantic similarity and can help the understanding and improve the interpretation of the affected biological functions, by indicating which of the studied terms are most similar. Figures 3.39 and 3.40 show, respectively, Viseago's analysis results on the TP53 silencing and on the LINC01605 knockout experiments.

In order to try estimating the regulatory functions that TP53 exerts through LINC01605's activity, we considered the Gene Ontology terms enriched in both experiments. Since Viseago can only be applied to lists of genes (because the enriched GO terms are independently computed and cannot be provided in input), we tried other approaches and found that the combination of Panther's functional analysis, Revigo's semantic clustering and Cytoscape's [77, 56] network plots offers a satisfactory interpretational flexibility.

Figures 3.41, 3.42, 3.43 represent the "biological process" GO terms commonly enriched in both TP53's silencing and in LINC01605's knockout experiments. The figures were produced by Revigo from the enriched GO terms estimated by Panther.

The table in figure 3.41 lists and groups the GO terms by semantic similarity: the terms in gray are clustered with the term in black above, the representative of the cluster. The plot in figure 3.42 further joins the clusters' representatives in groups of loosely related terms, visualized by different colors.

The graph in figure 3.43 uses nodes to display the clusters' representatives, whose size is proportional to the number of genes contained in the GO term: smaller nodes refer to more specific functions. Even though the presence of enriched specific functions can be more informative about the studied phenomenon compared to more generic ones (represented by bigger nodes), it is important to note that the reliability of the information could be lower because of the limits of the Over Representation Analysis. For instance, given a specific gene set of size 5 and an estimated number of genes in the input list (under the non enrichment hypothesis) of 0.3, if one or two genes within such set are wrongly considered differentially expressed (false positives) then the gene set will be also considered so.

The edge width models the overlap between the GO gene sets, interpreted as the proportion of shared genes: two gene sets connected by a thick edge have a relatively high number of common genes. This information can be useful to graphically group similar biological functions. Both the node size and the edge width refer to static aspects of the Gene Ontology.

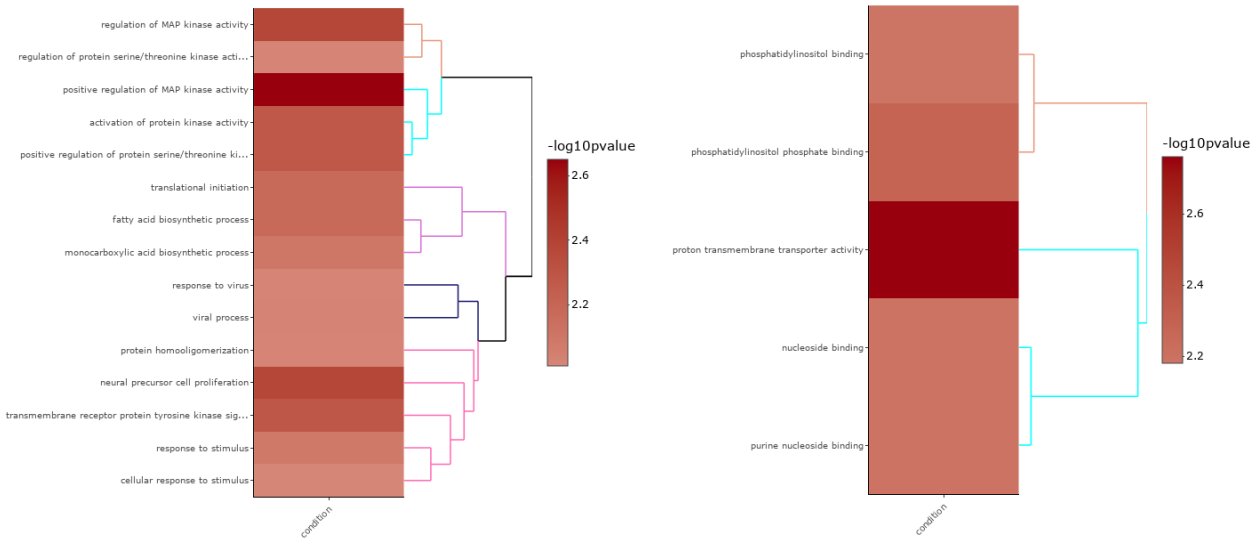
Figures 3.44, 3.45, 3.46 and 3.47, 3.48, 3.49 repeat the same visualizations for the "cellular component" and "molecular function" classes of GO terms.

Overall, the main biological functions described by the enriched Gene Ontology terms agree with the roles assigned to LINC01605 by different researches [97, 24, 68], which suggest that *regulation of cell migration* and *proliferation* may be among the targeted functions. *Cell migration* plays a fundamental role in every complex organism, for example in tissue renewal and repair, but undesirable migratory events can cause a number of pathological states such as inflammatory diseases and cancers [89].

Many of the GO terms enriched in our analyses (presented in figures 3.41 to 3.49) can be linked to these known functions: the prominent ones are shown in code 3.3.

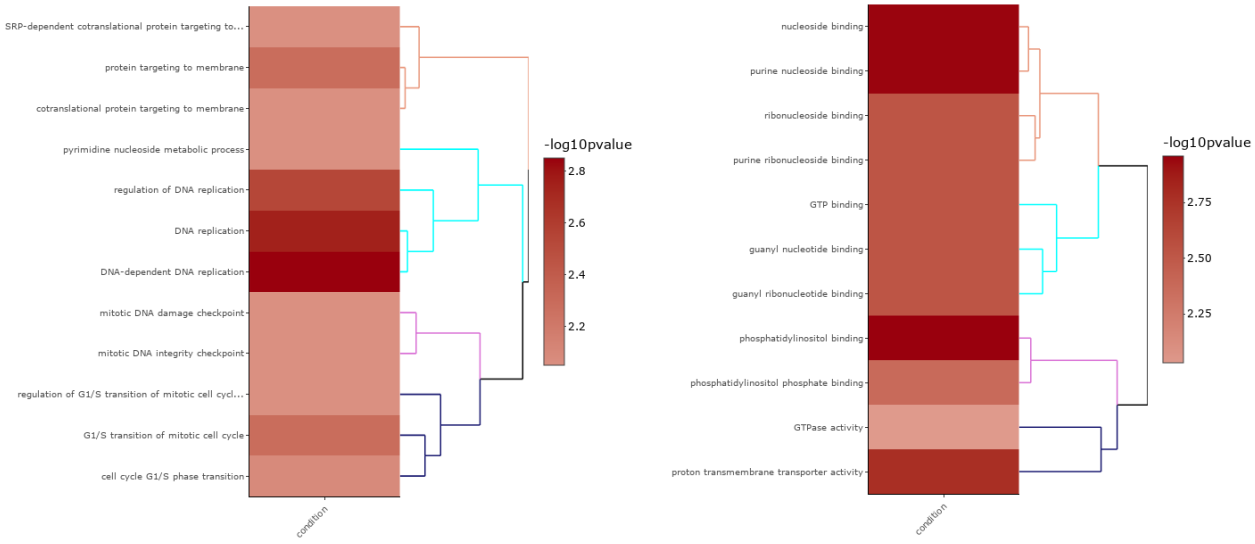
```
GO:0140014 mitotic nuclear division
GO:0007088 regulation of mitotic nuclear division
GO:0051301 cell division
GO:0050673 epithelial cell proliferation
GO:0050678 regulation of epithelial cell proliferation
GO:0006928 movement of cell or subcellular component
GO:0001954 positive regulation of cell-matrix adhesion
GO:0030335 positive regulation of cell migration
GO:0005871 kinesin complex
GO:0062023 collagen-containing extracellular matrix
GO:0015630 microtubule cytoskeleton
GO:0072686 mitotic spindle
GO:0003777 microtubule motor activity
```

Code 3.3: Principal Gene Ontology terms enriched in our analysis on LINC01605 and linked to known LINC01605's functions.



(a) Biological process (BP) terms *(b) Molecular function (MF) terms*

Figure 3.39: Semantic similarity clustering on the Gene Ontology terms relevant to the genes enriched in TP53's silencing experiments (considering the genes DE in both sh1 and sh2 experiments). Low branch ramifications indicate high similarity between the merged terms or clusters.



(a) Biological process (BP) terms *(b) Molecular function (MF) terms*

Figure 3.40: Semantic similarity clustering on the Gene Ontology terms relevant to the genes enriched in LINC01605's knockout experiment. Low branch ramifications indicate high similarity between the merged terms or clusters.

Term ID	Name	Frequency
GO:0008150	biological_process	100.000%
GO:0009987	cellular process	86.803%
GO:0051179	localization	32.339%
GO:0051606	detection of stimulus	3.999%
GO:0090316	positive regulation of intracellular protein transport	1.010%
GO:0051046	*regulation of secretion*	*3.500%*
GO:0043270	*positive regulation of ion transport*	*1.610%*
GO:0051050	*positive regulation of transport*	*5.182%*
GO:0032386	*regulation of intracellular transport*	*1.980%*
GO:0051223	*regulation of protein transport*	*3.006%*
GO:0046824	*positive regulation of nucleocytoplasmic transport*	*0.353%*
GO:0032388	*positive regulation of intracellular transport*	*1.245%*
GO:0042307	*positive regulation of protein import into nucleus*	*0.230%*
GO:0033157	*regulation of intracellular protein transport*	*1.380%*
GO:0051222	*positive regulation of protein transport*	*1.800%*
GO:1904591	*positive regulation of protein import*	*0.241%*
GO:0140014	mitotic nuclear division	1.139%
GO:0051315	*attachment of mitotic spindle microtubules to kinetochore*	*0.067%*
GO:0000070	*mitotic sister chromatid segregation*	*0.695%*
GO:0000280	*nuclear division*	*1.913%*
GO:0016322	neuron remodeling	0.062%
GO:0042551	*neuron maturation*	*0.196%*
GO:0034501	protein localization to kinetochore	0.084%
GO:0050673	epithelial cell proliferation	0.494%
GO:0051301	cell division	2.737%
GO:0071840	cellular component organization or biogenesis	33.281%
GO:0006928	movement of cell or subcellular component	8.828%
GO:0060668	regulation of branching involved in salivary gland morphogenesis by extracellular matrix-epithelial cell signaling	0.006%
GO:1900450	negative regulation of glutamate receptor signaling pathway	0.006%
GO:0043086	negative regulation of catalytic activity	4.324%
GO:0050790	*regulation of catalytic activity*	*13.494%*
GO:0051336	*regulation of hydrolase activity*	*5.945%*
GO:0044092	*negative regulation of molecular function*	*6.310%*
GO:0048872	homeostasis of number of cells	1.251%
GO:0007088	regulation of mitotic nuclear division	0.595%
GO:0032956	*regulation of actin cytoskeleton organization*	*1.974%*
GO:0051783	*regulation of nuclear division*	*0.752%*
GO:0010638	*positive regulation of organelle organization*	*3.292%*
GO:0032233	*positive regulation of actin filament bundle assembly*	*0.348%*
GO:0050678	regulation of epithelial cell proliferation	1.896%
GO:0065009	regulation of molecular function	17.257%
GO:0051128	regulation of cellular component organization	13.214%
GO:0051726	regulation of cell cycle	6.405%
GO:0007605	sensory perception of sound	0.847%
GO:0006979	response to oxidative stress	2.131%
GO:0010243	response to organonitrogen compound	5.597%
GO:0001954	positive regulation of cell-matrix adhesion	0.314%
GO:0001952	*regulation of cell-matrix adhesion*	*0.684%*
GO:2000379	positive regulation of reactive oxygen species metabolic process	0.387%
GO:0030199	collagen fibril organization	0.566%
GO:0031323	regulation of cellular metabolic process	34.139%
GO:0051234	establishment of localization	25.782%
GO:0006810	*transport*	*24.936%*
GO:0043065	positive regulation of apoptotic process	3.017%
GO:0043068	*positive regulation of programmed cell death*	*3.090%*
GO:0016043	cellular component organization	32.036%
GO:0042542	response to hydrogen peroxide	0.679%
GO:0000302	*response to reactive oxygen species*	*1.032%*
GO:0034599	*cellular response to oxidative stress*	*1.290%*
GO:0070301	*cellular response to hydrogen peroxide*	*0.409%*
GO:0048522	positive regulation of cellular process	31.548%
GO:0009893	*positive regulation of metabolic process*	*20.903%*
GO:0048285	organelle fission	2.075%
GO:0062197	cellular response to chemical stress	1.526%
GO:0071549	cellular response to dexamethasone stimulus	0.163%
GO:1903954	positive regulation of voltage-gated potassium channel activity involved in atrial cardiac muscle cell action potential repolarization	0.006%
GO:1903952	*regulation of voltage-gated potassium channel activity involved in atrial cardiac muscle cell action potential repolarization*	*0.006%*
GO:0030335	positive regulation of cell migration	2.838%
GO:0051272	*positive regulation of cellular component movement*	*3.040%*
GO:0007010	cytoskeleton organization	6.803%
GO:0008104	protein localization	11.374%
GO:0033043	regulation of organelle organization	6.741%
GO:0051130	*positive regulation of cellular component organization*	*6.354%*
GO:0033036	macromolecule localization	13.735%
GO:0051041	positive regulation of calcium-independent cell-cell adhesion	0.006%
GO:0031325	positive regulation of cellular metabolic process	18.132%
GO:0010604	*positive regulation of macromolecule metabolic process*	*19.265%*
GO:0060341	regulation of cellular localization	4.565%
GO:0008608	attachment of spindle microtubules to kinetochore	0.129%

Figure 3.41: Biological process GO terms enriched in both TP53's silencing and LINC01605's knockout experiments. Terms clustered by semantic similarity.

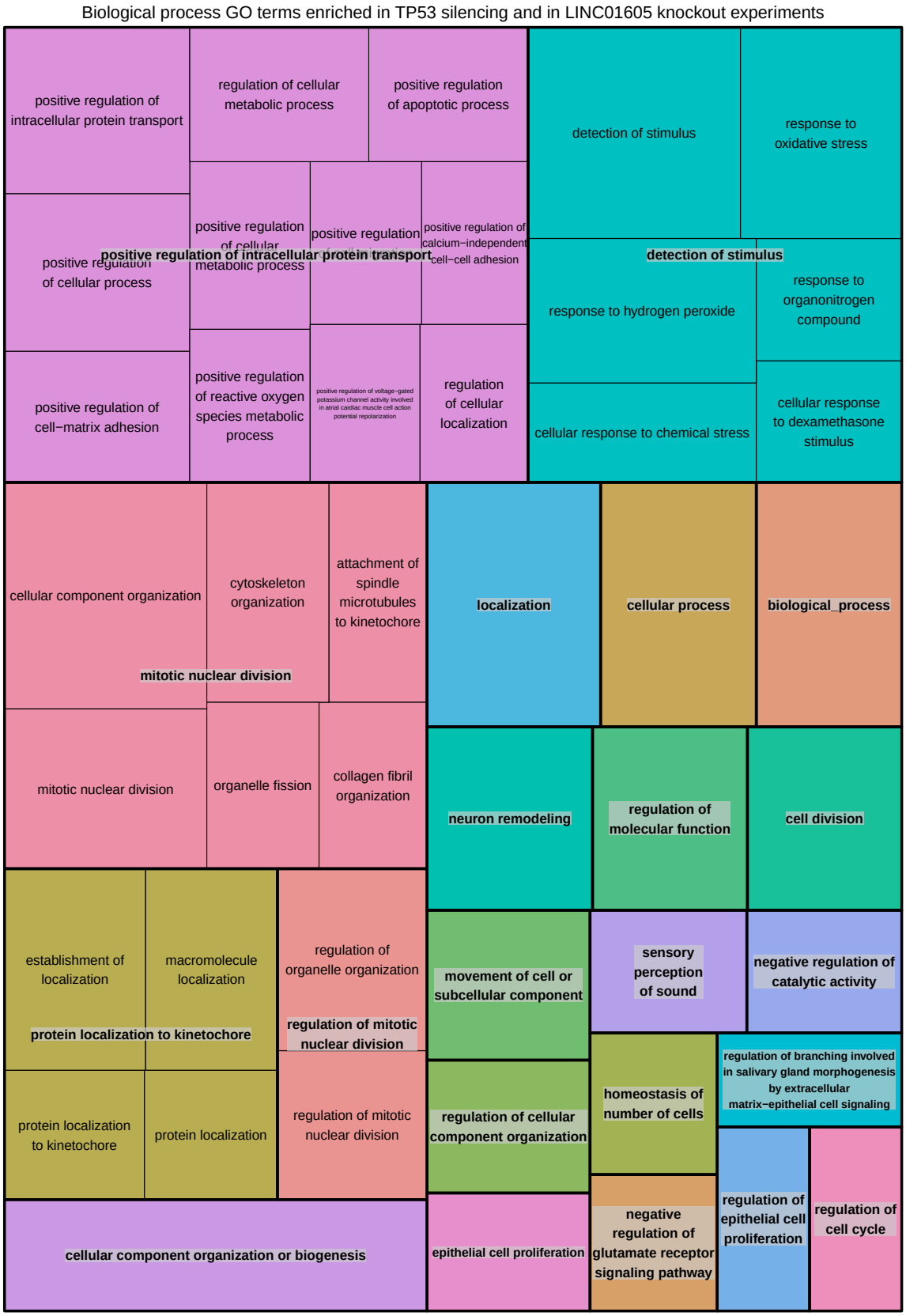


Figure 3.42: Graphical representation of the main clusters of functions enriched in both TP53's silencing and LINC01605's knockout experiments. The size of the tiles is proportional to the number of genes in the GO term.

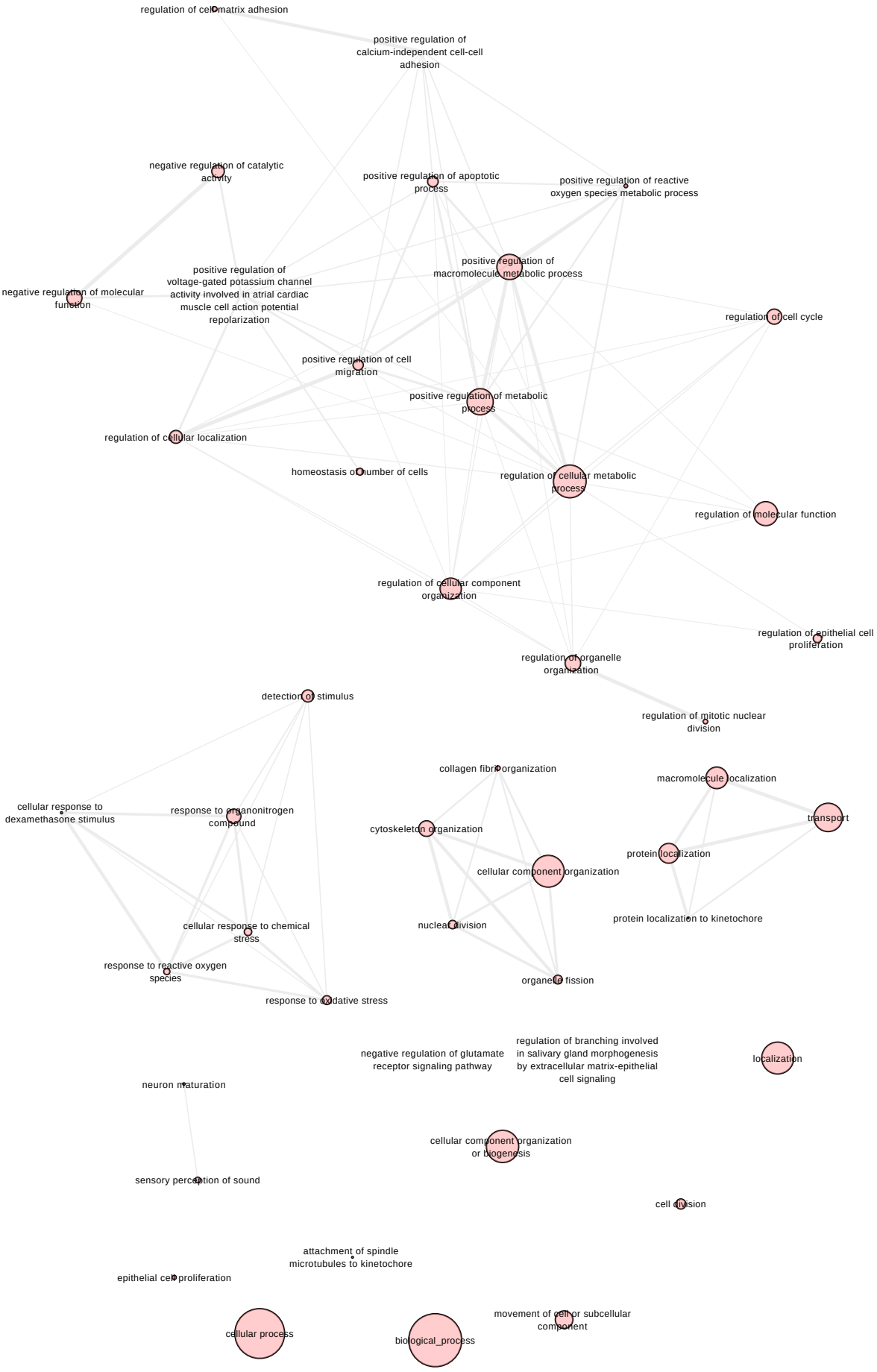


Figure 3.43: Graph showing the main biological process GO terms enriched in both TP53's silencing and LINC01605's knockout experiments. The size of the nodes is proportional to the number of genes in the gene set while the width of the edges indicates the proportion of shared genes between GO terms.

Term ID	Name	Frequency
GO:0005575	cellular_component	100.000%
GO:0005871	kinesin complex	0.271%
GO:0110165	cellular anatomical entity	99.079%
GO:0062023	collagen-containing extracellular matrix	2.261%
GO:0005622	intracellular anatomical structure	78.328%
GO:0012505	endomembrane system	24.572%
GO:0031974	membrane-enclosed lumen	29.281%
GO:0043226	organelle	73.523%
GO:0005737	cytoplasm	63.105%
GO:0005741	mitochondrial outer membrane	1.064%
GO:0031968	*organelle outer membrane*	*1.192%*
GO:0043227	membrane-bounded organelle	69.128%
GO:0048471	perinuclear region of cytoplasm	3.858%
GO:0019867	outer membrane	1.203%
GO:0005829	cytosol	28.349%
GO:0043228	non-membrane-bounded organelle	26.854%
GO:0005783	endoplasmic reticulum	10.562%
GO:0070013	intracellular organelle lumen	29.281%
GO:0043233	*organelle lumen*	*29.281%*
GO:0031090	organelle membrane	19.230%
GO:0044599	AP-5 adaptor complex	0.005%
GO:0043235	receptor complex	2.602%
GO:0043231	intracellular membrane-bounded organelle	63.542%
GO:0043229	intracellular organelle	69.272%
GO:0043232	intracellular non-membrane-bounded organelle	26.812%
GO:0005856	*cytoskeleton*	*12.355%*
GO:0030659	cytoplasmic vesicle membrane	4.203%
GO:0030667	*secretory granule membrane*	*1.655%*
GO:0005794	Golgi apparatus	8.434%
GO:0005765	lysosomal membrane	2.096%
GO:0098852	*lytic vacuole membrane*	*2.096%*
GO:0072686	mitotic spindle	0.830%
GO:0005813	*centrosome*	*3.208%*
GO:0005875	*microtubule associated complex*	*0.857%*
GO:0005819	*spindle*	*2.049%*
GO:0005874	*microtubule*	*2.251%*
GO:0000922	*spindle pole*	*0.894%*
GO:0015630	microtubule cytoskeleton	6.901%
GO:0098588	bounding membrane of organelle	11.424%
GO:0012506	*vesicle membrane*	*6.337%*

Figure 3.44: Cellular component GO terms enriched in both TP53's silencing and LINC01605's knockout experiments. Terms clustered by semantic similarity.

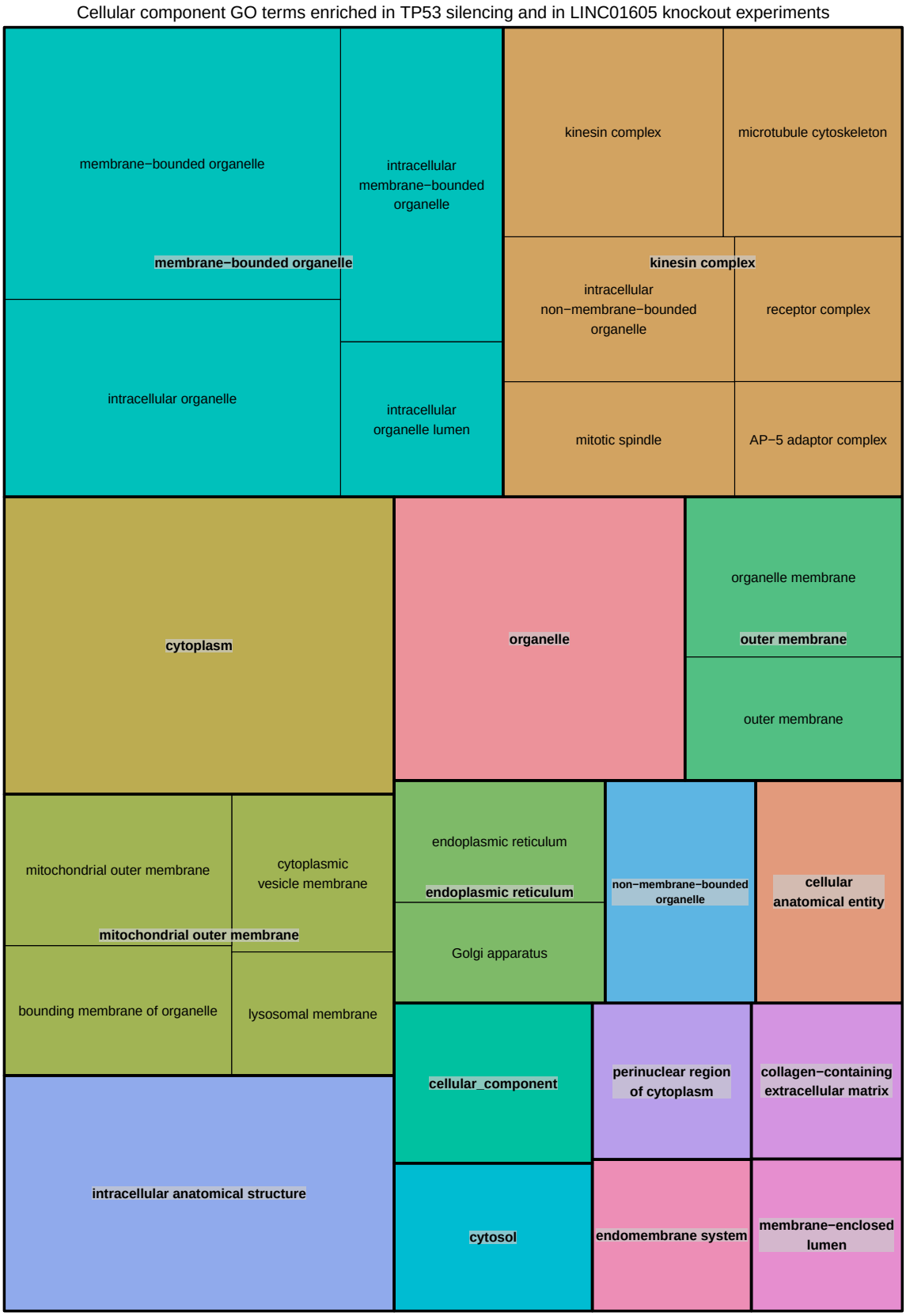


Figure 3.45: Graphical representation of the main clusters of functions enriched in both TP53's silencing and LINC01605's knockout experiments. The size of the tiles is proportional to the number of genes in the GO term.

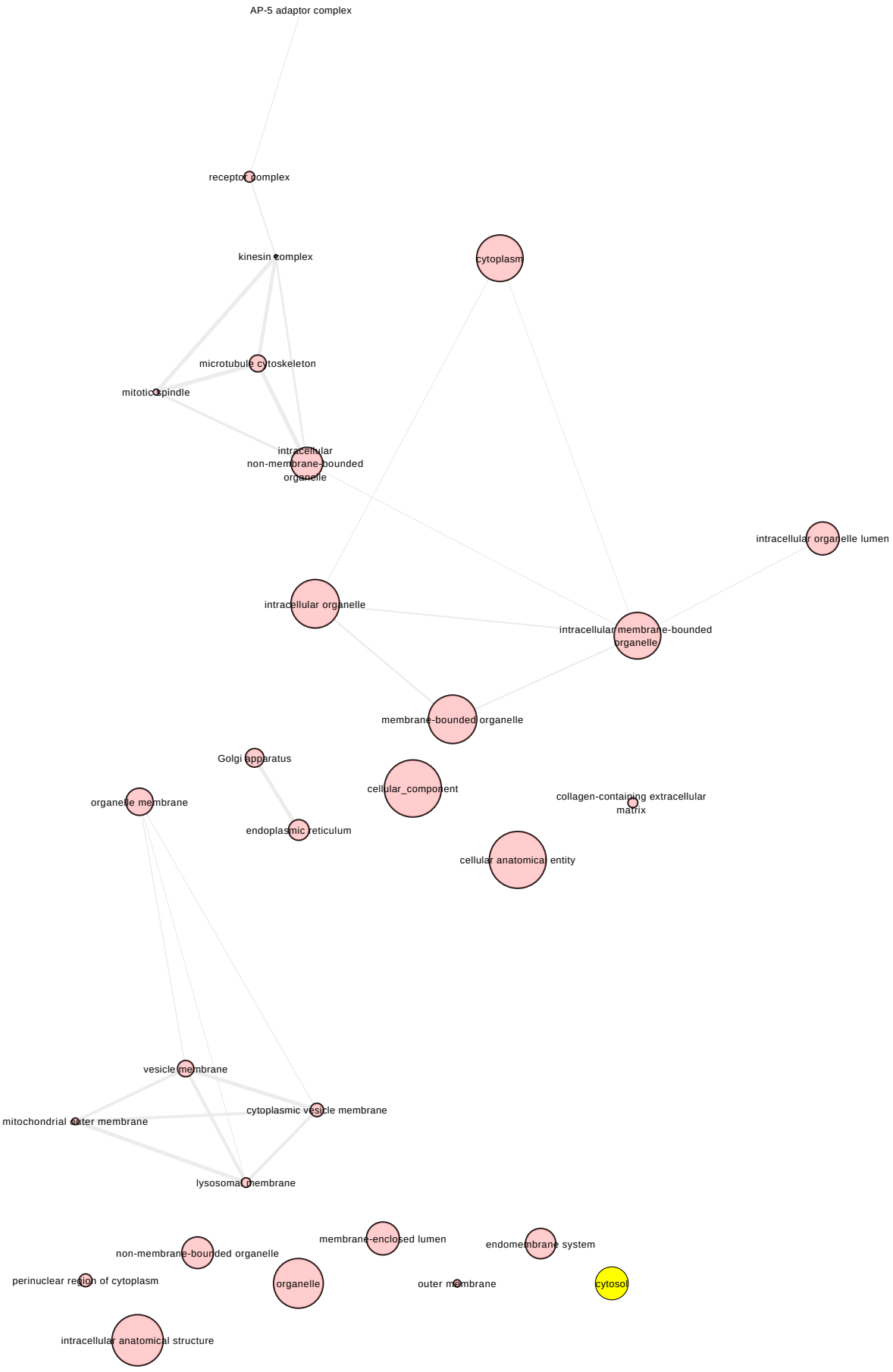


Figure 3.46: Graph showing the main cellular component GO terms enriched in both TP53's silencing and LINC01605's knockout experiments. The size of the nodes is proportional to the number of genes in the gene set while the width of the edges indicates the proportion of shared genes between GO terms.

Term ID	Name	Frequency
GO:0003674	molecular_function	100.000%
GO:0003777	microtubule motor activity	0.370%
GO:0003774	*motor activity*	*0.619%*
GO:0017111	*nucleoside-triphosphatase activity*	*2.515%*
GO:0008574	*ATP-dependent microtubule motor activity, plus-end-directed*	*0.094%*
GO:1990939	*ATP-dependent microtubule motor activity*	*0.332%*
GO:0005488	binding	89.584%
GO:0015643	toxic substance binding	0.055%
GO:0016887	ATPase	2.615%
GO:0045237	CXCR1 chemokine receptor binding	0.006%
GO:0031737	*CX3C chemokine receptor binding*	*0.017%*
GO:0005515	protein binding	76.443%
GO:0044877	protein-containing complex binding	6.861%
GO:0035639	purine ribonucleoside triphosphate binding	10.045%
GO:0017076	*purine nucleotide binding*	*10.515%*
GO:0032555	*purine ribonucleotide binding*	*10.432%*
GO:0005524	*ATP binding*	*8.138%*
GO:0030554	*adenyl nucleotide binding*	*8.558%*
GO:0000166	*nucleotide binding*	*11.859%*
GO:0032553	*ribonucleotide binding*	*10.526%*
GO:0032559	*adenyl ribonucleotide binding*	*8.486%*
GO:0097367	carbohydrate derivative binding	12.400%
GO:0036094	small molecule binding	13.755%
GO:0050512	lactosylceramide 4-alpha-galactosyltransferase activity	0.006%
GO:0043168	anion binding	13.219%
GO:1901265	nucleoside phosphate binding	11.864%

Figure 3.47: Molecular function GO terms enriched in both TP53's silencing and LINC01605's knockout experiments. Terms clustered by semantic similarity.

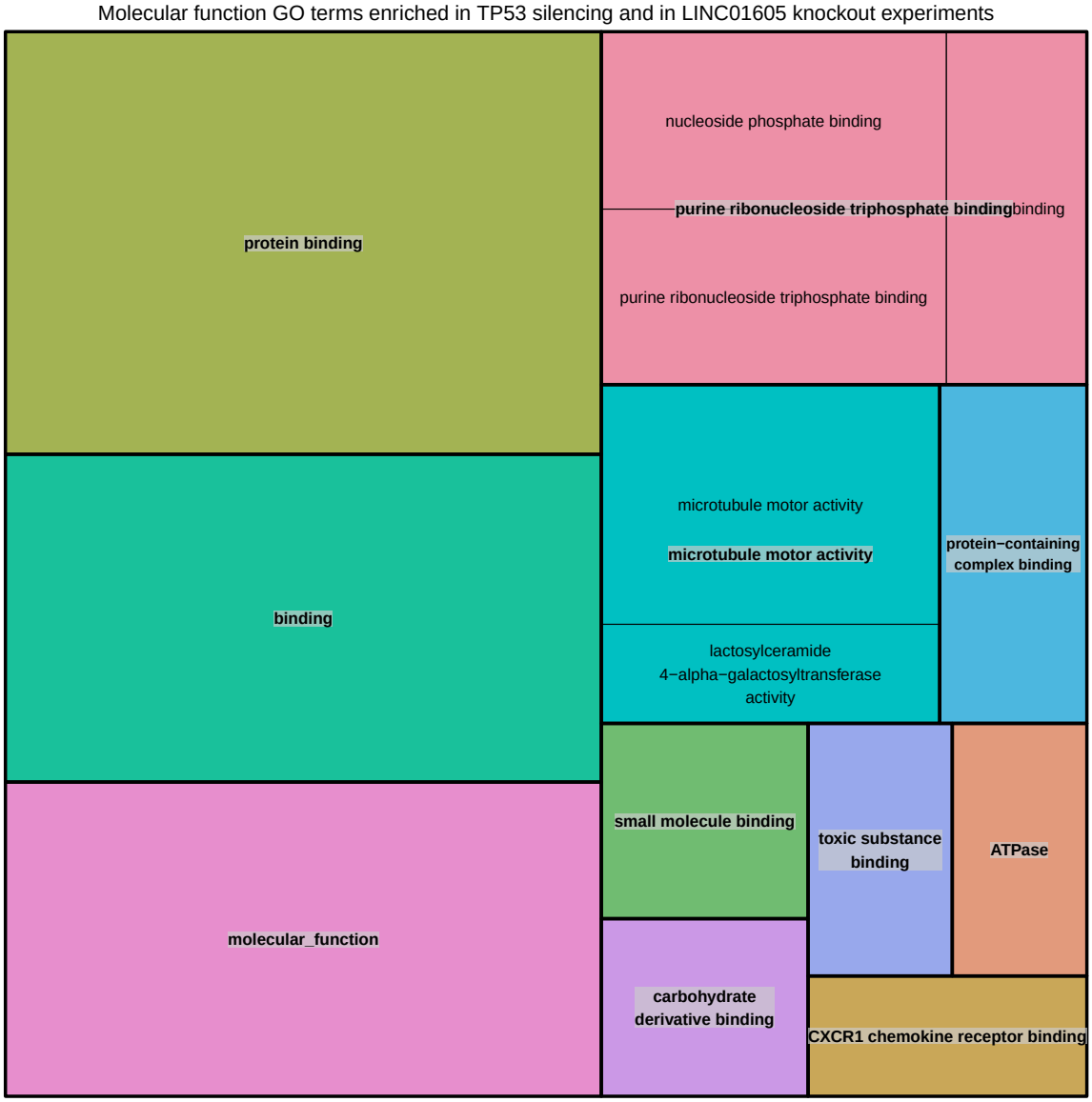


Figure 3.48: Graphical representation of the main clusters of functions enriched in both TP53's silencing and LINC01605's knockout experiments. The size of the tiles is proportional to the number of genes in the GO term.

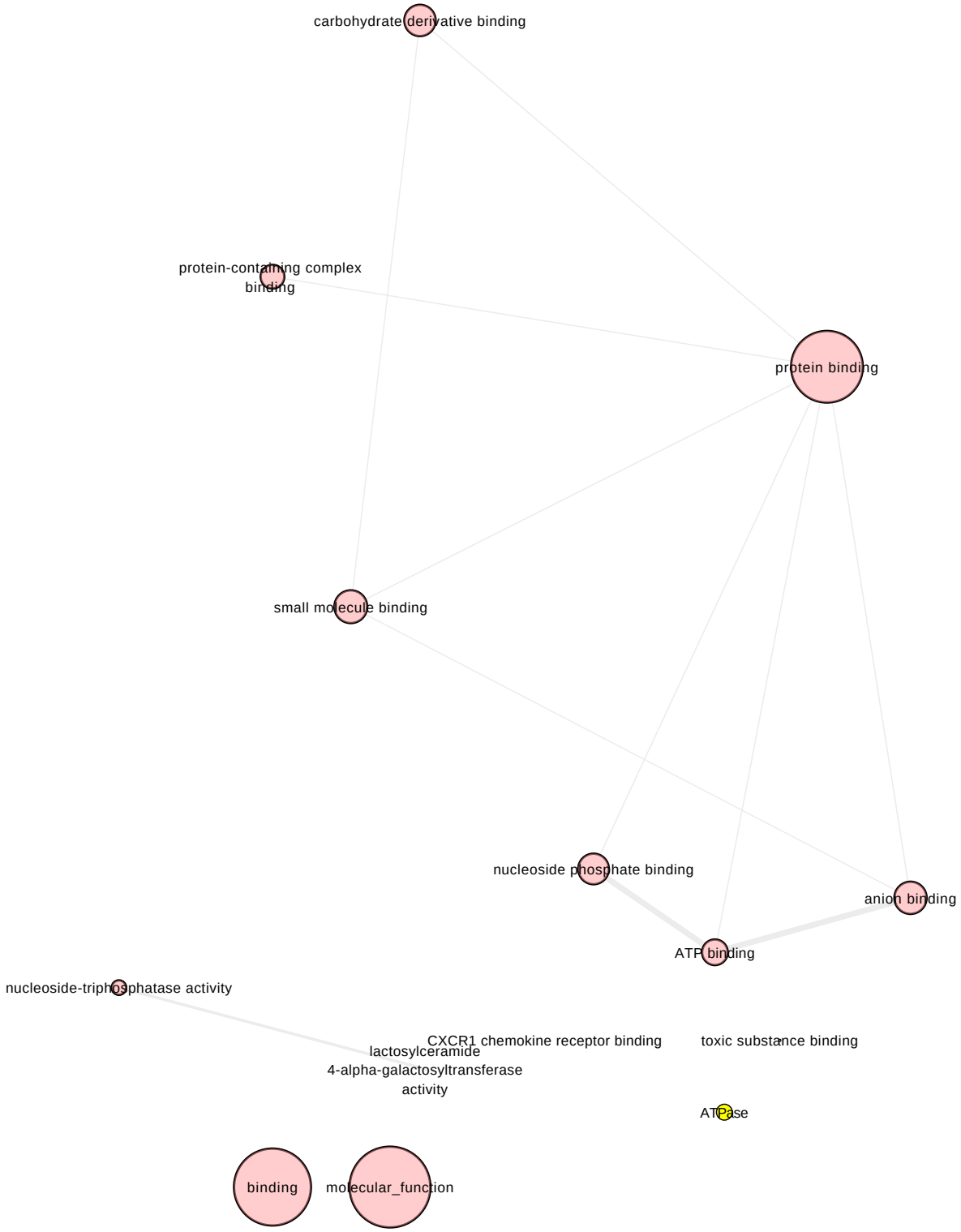


Figure 3.49: Graph showing the main molecular function GO terms enriched in both TP53's silencing and LINC01605's knockout experiments. The size of the nodes is proportional to the number of genes in the gene set while the width of the edges indicates the proportion of shared genes between GO terms.

3.7.4 KEGG Pathway analysis

KEGG (Kyoto Encyclopedia of Genes and Genomes) is an organized dataset for the biological interpretation of genes and genes' products of organisms with completely sequenced genomes [36]. The KEGG Pathway database is a resource often used within function analysis as it is one of the more mature structured datasets of cellular pathways and because the maps it provides[9] often help identifying genes within the considered process.

We applied GSEA's functional analysis using the KEGG Pathways gene set collection on both TP53's silencing and on LINC01605's knockout experiments: the pathways subject to a significant (p-value = 0.01) enrichment are displayed in figures 3.50 and 3.51. Overall we found the enriched KEGG pathways not particularly informative.

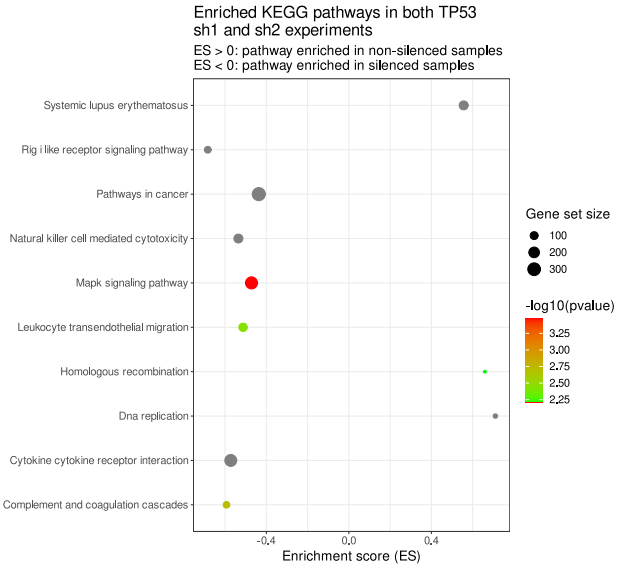


Figure 3.50: KEGG pathways significantly enriched in both sh1 and sh2 TP53's silencing experiments. Larger dots represent pathways containing more genes.

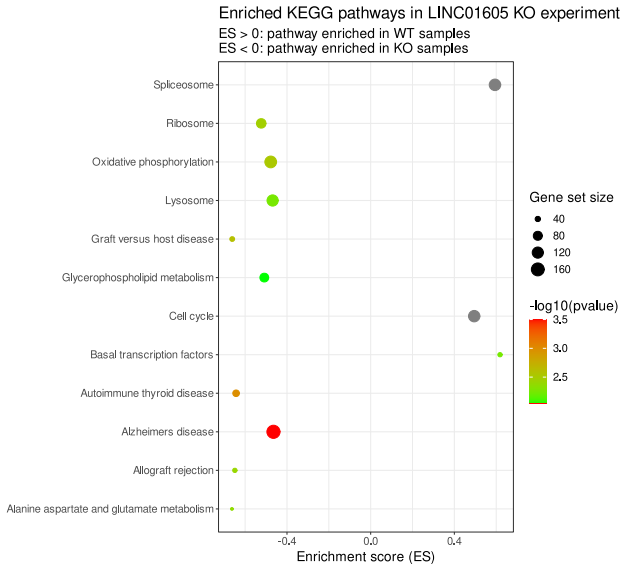


Figure 3.51: KEGG pathways significantly enriched in the LINC01605 knockout experiment. Larger dots represent pathways containing more genes.

[9]KEGG Pathway maps: `https://www.genome.jp/kegg/pathway.html`

Conclusion

In the last two decades we observed a staggering increase in the amount, and in the rate, of data being produced in almost every field, thanks to the developments of technology and to the introduction of new productive ways of extracting information. Biology and genetics are no exception to this phenomenon (so much so that a new discipline, genomics, developed to produce and handle data about the entire genome of living organisms) and the need to use these new resources effectively is a problem felt by an ever-increasing number of researchers. For this reason, verifying the scope, the capabilities and the effectiveness of the currently available bioinformatics methods and tools is a pressing necessity and the ambition of this work.

In this thesis, we provided an introductory survey of the main genetics concepts and of some relevant bioinformatics data formats and techniques, followed by a more in-depth review of the technologies and analysis methodologies enabling RNA-Seq and ChIP-Seq: powerful modern approaches based on high throughput sequencing. Then, we presented a study, currently worked on by researchers at CRO, aimed at investigating mutant-TP53's role on tumor progression and, in particular, its relationship with the LINC01605 gene.

The intents of the discussed activities can be grouped in two main goals. The first one is the assessment of the performance of ad-hoc bioinformatics tools in the extension of local analyses to a genome-wide scale. To test this, we considered the observed relationship between the mutant p53 protein and the LINC01605 gene. Through the characterization of p53's binding sites, the differential gene expression analysis and the study of modified histone regions, we observed that the employed tools can provide satisfactory results and were able to detect, to a certain extent, the considered relationship.

Our second goal is the evaluation of the functional analysis methods and annotations, programs useful for improving the understanding of the studied phenomenon, in wide use in many research pipelines. After presenting the main information sources and methodologies, along with some projects implementing them, we provided a metric intended for the comparison of the results offered by different tools, and we evaluated it on data from the experiment under study. We observed that the metric, overall, provides sensible results even though it has some limitations that we denoted.

Among all the tested functional analysis tools, we chose the ones we considered more flexible and informative, and we applied them to the data from LINC01605's research with the aim of efficiently characterizing its functions and its role in tumor development. We showed that the results offered by these tools are positive and consonant with the LINC01605's functions observed by other studies.

The analyses we presented are replicable using the instructions contained in the appendix and the scripts made available on github (`https://github.com/lorenzoiuri/thesis-support-material`). This information can be used to assemble automated data analysis pipelines.

Overall, the outlined aims were met: modern bioinformatic methods can provide useful insights and support research on multiple activities. The usage of these programs, however, is not always trivial: methods based on different models and dependent on different hypotheses can produce contrasting results, as we observed for differential expression analysis, functional analysis and for the Gene Ontology semantic comparison metrics. To better handle this problem it is important to have a solid understanding on each method but, sometimes, arbitrary choices may be required.

As a consequence, it is important to keep in mind that the statistical and bioinformatic methods and tools available in this field should only be used to advise research providing useful insights and not to confirm and validate hypotheses, for which more formal and thorough experimentations are necessary.

A

Appendix: Reproducibility

Programs' versions

R: 4.1 / 3.6
Python: 3.6
BWA: 0.7.17
MACS2: 2.1.4
FASTQC: 0.11.9
Bedtools: 2.29.2
DeepTools: 3.5.0
DESeq2: 1.32
EdgeR: 3.34
Limma: 3.48

GSEA: 4.1
Panther: 16.0 (online)
Revigo: May 2021 (online)
clusterProfiler: 3.14.3
GoSemSim: 2.12.1

Commands and files

Data and scripts available at https://github.com/lorenzoiuri/thesis-support-material.

Section 3.2

ChIP-Seq mutant TP53 experiment files

FASTQ files: https://www.ncbi.nlm.nih.gov/sra?term=SRP055837

Reads Alignment

Reference genome assembly: GENCODE GRCh37

```
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/
GRCh37_mapping/GRCh37.primary_assembly.genome.fa.gz
```

Alignment command pipeline:

```
IDX=./GRCh37.primary_assembly.genome.fa
FQ=./DO-1.fastq
bwa mem -t 4 $IDX $FQ | samtools sort -@ 4 -o DO-1.bam
```

BAM to BIGWIG conversion

Single files:

```
bamCoverage --bam DO-1.bam \
            -of bigwig \
            -o DO-1.bigwig \
            -p 4 \
            --binSize 10 \
            --normalizeUsing RPGC \
            --effectiveGenomeSize 2864785220
```

Comparison:

```
bamCompare -b1 ../bam/DO-1/DO-1.bam \
           -b2 ../bam/input/input.bam \
           --operation log2 \
           -of bigwig \
           -o bamCompare_DO-1_input_log2 \
           -bs 10 \
           -p 4 \
           --scaleFactorsMethod None \
           --normalizeUsing BPM \
           --effectiveGenomeSize  2864785220
```

Binding sites estimation

```
macs2 callpeak -t ../../bam/DO-1/DO-1.bam \ # treatment file
               -c ../../bam/input/input.bam \ # control file
               -f BAM \
               -n do-1_input --outdir do-1_input \
               --mfold 2 50
```

Section 3.3

GENCODE annotation files

Downloaded from UCSC Table Browser:

```
https://genome.ucsc.edu/cgi-bin/hgTables
assembly: Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Predictions
track: GENCODE V35lift37 / V24lift37
table: knownGene
region: genome
```

Intersection computation

Intersections between putative mutant p53 binding sites and GENCODE annotation:

```
bedtools intersect -wo -a macs2.bed -b annotation.bed > bedtools_intersect.bed
```

Putative mutant p53 binding sites not intersecting GENCODE transcripts:

```
bedtools intersect -c -a macs2.bed -b annotation.bed > counts.txt
awk '{ if ($5 == 0) { print } }' counts.txt > macs2_no_intersections.txt
```

Computation of closest gene to site:

```
bedtools closest -a macs2_no_intersections.bed \
                 -b ../annotation.bed \
                 -D ref -iu -t first > closest_genes_downstream.txt
bedtools closest -a macs2_no_intersections.bed \
                 -b ../annotation.bed \
                 -D ref -id -t first > closest_genes_upstream.txt
```

Generation of random samples of regions

Application of GAMLSS:

```
fit <- fitDist(p53.sites$len, k = 2, type = "realplus",
               trace = FALSE, try.gamlss = TRUE)
```

Generation of samples:

```
sample <- rGG(n = length(p53.sites$len),
              mu = exp(p53.sites$mu.coefficients),
              sigma = exp(p53.sites$sigma.coefficients),
              nu = p53.sites$nu.coefficients )
```

Script for the generation of random samples of sites:
`scripts/4-3/random-sites-generation/gen_regions.py`

Section 3.4

Differential expression analysis

GENCODE annotation GTF file:

```
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_35/
GRCh37_mapping/gencode.v35lift37.basic.annotation.gtf.gz
```

Reads quantification using featureCounts:

```
GTF=gencode.v35lift37.basic.annotation.gtf
featureCounts -T 4 -p \ # p for paired end reads
              -t exon -g gene_id \
              -a $GTF \
              -o counts.txt \
              bc1-1.bam bc1-2.bam bc1-3.bam \
              bc2-1.bam bc2-2.bam bc2-3.bam
```

Differential expression of lncRNAs

Filtering the *featureCounts* output keeping only *lncRNA* genes:

```
# Ensembl codes of the lncRNA genes
cat gencode.v35lift37.basic.annotation.gtf | \
    awk -F'\t'  '{ if ($3 == "gene") { print }  }' | \
    grep 'gene_type "lncRNA"' | \
    cut -f9 | \
    cut -d';' -f1 | \
    cut -d' ' -f2 | \
    sed 's/"//g' | \
    sort > lncRNA_codes.txt


# sorting the counts file produced by featureCounts on the complete annotation
sort -k1,1 counts.txt > counts_sorted.txt


# keeping only the lncRNA genes in the counts files
join -1 1 -2 1 counts_sorted.txt lncRNA_codes.txt > counts_only_lncRNA.txt
```

Interactions between differentially expressed genes and mutant p53

Generation of the bed file with the DE genes:

```
# keeping only certain columns from the gencode GTF annotation file
# to produce a bed file with the position of all the genes
GTF=gencode.v35lift37.basic.annotation.gtf
```

```
awk '{ if ($3 == "gene") { print } }' $GTF | cut -f1,4,5,7 > temp1.txt
awk '{ if ($3 == "gene") { print } }' $GTF | cut -f9 | cut -f1 -d';' | \
        cut -f2 -d' ' | sed 's/"//g' > temp2.txt
paste temp1.txt temp2.txt | \
    awk '{print $1 "\t" $2 "\t" $3 "\t" $5 "\t" "." "\t" $4}' > gene_name_pos.bed

# keeping only the genes in the gene_name_pos.bed about the DE genes
join -1 4 -2 1 all_gene_name_pos.bed ../../common-DE-genes/DE_common_sh1.txt | \
    sed 's/ /\t/g' | \
    awk '{print $2 "\t" $3 "\t" $4 "\t" $1 "\t" $5 "\t" $6}' | \
    sort -k1,1 -k2,2n > DE_common_sh1_namepos.bed
```

Computing the intersections between the p53 (and random) sites and the DE genes:

```
bedtools intersect -wo \
    -a sites/p53-sites.bed \
    -b genes/DE_common_sh1_namepos.bed > intersections_p53sites_DEsh1.txt
```

Computing the closest DE gene to the sites not located on DE genes:

```
# finding the sites not on DE genes
bedtools intersect -c \
    -a sites/p53-sites.bed \
    -b genes/DE_common_sh1_namepos.bed | \
    awk '{ if ($5 == 0) { print } }' | \
    cut -f1,2,3,4 > sites-not-on-DEgenes_sh1.bed

# computing the distances to the closest DE gene
bedtools closest -a sites-not-on-DEgenes_sh1.bed \
    -b genes/DE_common_sh1_namepos.bed \
    -D ref -t first > closest_DEgene_sh1.txt
```

Section 3.5

Histone annotation data

Estimation of the H3K27ac sites on MDA-MB-231 cells:

```
macs2 callpeak -t ../data/h3k27ac_mda231/BAM/h3k27ac.bam \ # treatment file
               -c ../data/h3k27ac_mda231/BAM/input.bam \ # control file
               -f BAM \
               -n h3k27ac_input --outdir h3k27ac_input \
               --broad

cut -f1-4 h3k27ac_input/h3k27ac_input_peaks.broadPeak > h3k27ac_mda231.bed
```

Direct effects of p53 on histonic regulation

Computing the histone H3K27ac regions containing p53 binding sites:

```
bedtools intersect -c \
    -a h3k27ac_mda231.bed \
    -b p53-sites.bed | \
awk '{ if ($5 > 0) { print } }' | \
cut -f1-4 > h3k27ac_with_p53.bed
```

Computing the closest gene to the histone H3K27ac regions containing a p53 binding site:

```
bedtools closest \
    -a 1-histones_intersecting_sites/h3k27ac_with_p53.bed \
    -b /4-4/sites-on-DE-genes/genes/DE_common_sh1_namepos.bed \
    -D ref -t first > closest_DEgene_sh1.txt
```

Computing all the genes within 1 Mb from the histone H3K27ac regions containing a p53 binding site:

```
bedtools closest \
    -a 1-histones_intersecting_sites/h3k27ac_with_p53.bed \
    -b 4-4/sites-on-DE-genes/genes/all_gene_name_pos.bed \
    -D ref -k 300 | \
awk '{ if (($11 >= 0 && $11 <= 1000000) || \
           ($11 <= 0 && $11 >= -1000000)) { print } }' | \
> genes_within_1MB.txt
```

Effects of p53 on indirect gene regulation

Producing a GTF file from the BED of the H3K27ac regions:

```
awk -v OFS="\t" '{print $1, ".", "histone", $2+1, $3, \
                  ".", ".", ".", "histone_id " "\"" $4 "\"" ";"}' \
                  h3k27ac_mda231.bed > histone_annotation.gtf
```

Reads quantification using featureCounts:

```
GTF=histone_annotation.gtf
featureCounts -p -T 4 \
              -t histone -g histone_id -a $GTF -o counts.txt \
              bc1-1.bam bc1-2.bam bc1-3.bam \
              bc2-1.bam bc2-2.bam bc2-3.bam
```

Differentially expressed H3K27ac histones having a p53 binding site:

```
bedtools intersect -c \
    -a 1-DE_histones/DE_histones_sh1.bed \
    -b ../4-3/sites/p53-sites.bed | \
awk '{ if ($5 > 0) { print } }' | \
cut -f1-4 > DE_H3K27ac_sh1_with_p53.bed
```

Section 3.6

Functions regulated by TP53 through histone binding

Extraction of the codes from the GO terms produced by Panther:

```
cat raw/*.txt | grep 'GO:' | cut -f1 | \
sed 's/^.*(//g' | sed 's/).*$//g' | \
sort > GOterms.txt
```

GO terms exclusive to DE genes within 50 kb from the H3K27ac sites:

```
join -v 2 -1 1 -2 1 \
allDEgenes/GOterms.txt \
DE50KB/GOterms.txt > ../only_DE50KB.txt

cat DE50KB/raw/*.txt allDEgenes/raw/*.txt | \
grep 'GO:' | cut -f1 | tr '(' '\t' | \
sed 's/)//g' | sed 's/ /_/g' | \
sort | uniq | sort -k2,2  > ../mapping.txt

join -1 1 -2 2 only_DE50KB.txt mapping.txt
```

Section 3.7

Application of the metric to experiment data

The terms lists used in the application of the comparison metric were prepared as follows.
In the case of TP53's silencing sh1 experiment, Panther and clusterProfiler were applied to the set of differentially expressed genes estimated by DESeq2 (p-value threshold = 0.01, log2FoldChange threshold = 0.58). GSEA, instead, received as input the reads counts normalized using DESeq2.
The default settings were used for all tools (no p-value correction, default number of permutations and gene set limits for GSEA). GSEA does not provide the codes of the GO terms, so the file `go_mapping.txt` was produced and used to map GO names to codes. The p-value threshold of 1×10^{-4} was used to filter the gene sets contained in the results of the functional analysis tools. For the LINC01605 analysis, the p-value threshold of 1×10^{-3} was used instead.
The list of gene sets produced by each tool was then clustered using Revigo and the metric was applied to the clustering results.

LINC01605 and TP53's functional analysis

DESeq2 was used, as described in the previous sections, to estimate the differentially expressed genes in the TP53 silencing and in the LINC01605 knockout experiments. List A contains the genes DE in both TP53 silencing experiments and list B contains the genes DE in the LINC01605 silencing experiment. Panther was applied to both lists (over representation analysis, full GO annotation, no FDR correction, p-value = 0.05) and the resulting sets of GO terms were intersected to consider only the functions relevant to both lists. The p-value for the shared GO terms was set to the maximum p-value of that set in the two lists.
The GO terms obtained were used as input for Revigo to produce the displayed plots. The options used for Revigo's analysis are the following:

- Allowed similarity: 0.5
- Species: Homo Sapiens
- Similarity measure: SimRel (default)

Bibliography

[1] E pluribus unum. *Nature Methods*, 7(5):331–331, May 2010.

[2] J. Adams. Dna sequencing technologies. *Nature Education*, 193, 2008.

[3] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544, July 2018.

[4] István Albert. Analyzing next generation sequencing data, msu 2014. genomic intervals. `https://angus.readthedocs.io/en/2014/_static/2014-lecture4-genomic-intervals.pdf`. Accessed: 2021-02-27.

[5] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC. Babraham Institute, January 2012.

[6] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 11 2012.

[7] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics, 2014.

[8] Scott D. Boyd. Diagnostic applications of high-throughput dna sequencing. *Annual Review of Pathology: Mechanisms of Disease*, 8(1):381–410, 2013. PMID: 23121054.

[9] Aurélien Brionne, Amélie Juanchich, and Christelle Hennequet-Antier. ViSEAGO: a Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. *BioData Mining*, 12(1):16, August 2019.

[10] V. Brusic. The growth of bioinformatics. *Briefings in Bioinformatics*, 8(2):69–70, December 2006.

[11] Moran N. Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L. Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, 25(18):1915–1927, September 2011.

[12] Joana Carlevaro-Fita, Andrés Lanzós, Lars Feuerbach, Chen Hong, David Mas-Ponte, Jakob Skou Pedersen, and Rory Johnson. Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Communications Biology*, 3(1):1–16, February 2020.

[13] Yiwen Chen, Nicolas Negre, Qunhua Li, Joanna O. Mieczkowska, Matthew Slattery, Tao Liu, Yong Zhang, Tae-Kyung Kim, Housheng Hansen He, Jennifer Zieba, Yijun Ruan, Peter J. Bickel, Richard M. Myers, Barbara J. Wold, Kevin P. White, Jason D. Lieb, and X. Shirley Liu. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, 9(6):609–614, June 2012.

[14] Harvard Chan Bioinformatics Core. Differential expression analysis with deseq2. `https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html`. Accessed: 2021-02-27.

[15] Sibnarayan Datta, Raghvendra Budhauliya, Bidisha Das, Soumya Chatterjee, null Vanlalhmuaka, and Vijay Veer. Next-generation sequencing in clinical virology: Discovery of new viruses. *World Journal of Virology*, 4(3):265–276, August 2015.

[16] Kevin Davies. *The $1,000 genome: the revolution in DNA sequencing and the new era of personalized medicine*. Free Press, New York, NY, 1st free press hardcover ed edition, 2010.

[17] Frederick E. Dewey, Stephen Pan, Matthew T. Wheeler, Stephen R. Quake, and Euan A. Ashley. DNA sequencing: clinical applications of new DNA sequencing technologies. *Circulation*, 125(7):931–944, February 2012.

[18] Silvia Di Agostino, Sabrina Strano, Velia Emiliozzi, Valentina Zerbini, Marcella Mottolese, Ada Sacchi, Giovanni Blandino, and Giulia Piaggio. Gain of function of mutant p53: the mutant p53/NF-Y protein complex reveals an aberrant transcriptional mechanism of cell cycle regulation. *Cancer Cell*, 10(3):191–202, September 2006.

[19] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

[20] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, and Collins. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[21] Anton Eberharter and Peter B Becker. Histone acetylation: a switch between repressive and permissive chromatin: Second in review series on chromatin dynamics. *EMBO reports*, 3(3):224–229, March 2002.

[22] Nancy A. Eckardt. Sequencing the Rice Genome. *The Plant Cell*, 12(11):2011–2018, November 2000.

[23] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, 7(9):1728–1740, September 2012.

[24] Megan E. Forrest, Alina Saiakhova, Lydia Beard, David A. Buchner, Peter C. Scacheri, Thomas LaFramboise, Sanford Markowitz, and Ahmad M. Khalil. Colon Cancer-Upregulated Long Non-Coding RNA lincDUSP Regulates Cell Cycle Genes and Potentiates Resistance to Apoptosis. *Scientific Reports*, 8(1):7324, December 2018.

[25] Donald R Forsdyke and James R Mortimer. Chargaff's legacy. *Gene*, 261(1):127–137, 2000.

[26] Warren R. Francis and Gert Wörheide. Similar Ratios of Introns to Intergenic Sequence across Animal Genomes. *Genome Biology and Evolution*, 9(6):1582–1598, June 2017.

[27] Federico Di Gesualdo, Sergio Capaccioli, and Matteo Lulli. A pathophysiological view of the long non-coding RNA world. *Oncotarget*, 5(22):10976–10996, November 2014.

[28] Javier E. Girardini, Marco Napoli, Silvano Piazza, Alessandra Rustighi, Carolina Marotta, Enrico Radaelli, Valeria Capaci, Lee Jordan, Phil Quinlan, Alastair Thompson, Miguel Mano, Antonio Rosato, Tim Crook, Eugenio Scanziani, Anthony R. Means, Guillermina Lozano, Claudio Schneider, and Giannino Del Sal. A Pin1/mutant p53 axis promotes aggressiveness in breast cancer. *Cancer Cell*, 20(1):79–91, July 2011.

[29] John F Griffiths, Anthony JF Griffiths, Susan R Wessler, Richard C Lewontin, William M Gelbart, David T Suzuki, Jeffrey H Miller, et al. *An introduction to genetic analysis*. Macmillan, 2005.

[30] Yan-Jun Guo, Wei-Wei Pan, Sheng-Bing Liu, Zhong-Fei Shen, Ying Xu, and Ling-Ling Hu. ERK/MAPK signalling pathway and tumorigenesis (Review). *Experimental and Therapeutic Medicine*, January 2020.

[31] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, September 2012.

[32] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183, 2007.

[33] Illumina Inc. An introduction to next-generation sequencing technology. `https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`. Accessed: 2021-01-07.

[34] Albert István. *The Biostar Handbook: 2nd Edition*. Online, 2020.

[35] Ming-Chun Jiang, Jiao-Jiao Ni, Wen-Yu Cui, Bo-Ya Wang, and Wei Zhuo. Emerging roles of lncRNA in cancer and therapeutic opportunities. *American Journal of Cancer Research*, 9(7):1354–1366, July 2019.

[36] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 11 2016.

[37] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, May 2002.

[38] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, 8(2):e1002375, February 2012.

[39] John Kiche, Oscar Ngesa, and George Orwa. On generalized gamma distribution and its application to survival data. *International Journal of Statistics and Probability*, 8:65, 08 2019.

[40] Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, August 2019.

[41] Aaron Klug. Rosalind franklin and the discovery of the structure of dna. *Nature*, 219(5156):808–810, 1968.

[42] Krešimir Križanovic, Amina Echchiki, Julien Roux, and Mile Šikic. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics (Oxford, England)*, 34(5):748–754, March 2018.

[43] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[44] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Research*, 39(Database issue):D19–21, January 2011.

[45] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*, May 2013. arXiv: 1303.3997.

[46] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

[47] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595, March 2010.

[48] Kening Li, Congling Xu, Yuxin Du, Muhammad Junaid, Aman-Chandra Kaushik, and Dong-Qing Wei. Comprehensive epigenetic analyses reveal master regulators driving lung metastasis of breast cancer. *Journal of Cellular and Molecular Medicine*, 23(8):5415–5431, August 2019.

[49] Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014.

[50] Yang Liao, Gordon K. Smyth, and Wei Shi. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8):e47, May 2019.

[51] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems*, 1(6):417–425, December 2015.

[52] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.

[53] MarketsandMarkets Research Private Ltd. Bioinformatics market by product & service (knowledge management tools, data analysis platforms (structural & functional), services), applications (genomics, proteomics & metabolomics), & sectors (medical, academics, agriculture) - global forecast to 2023. 2018.

[54] Shalom Madar, Einav Harel, Ido Goldstein, Yan Stein, Ira Kogan-Sakin, Iris Kamer, Hilla Solomon, Elya Dekel, Perry Tal, Naomi Goldfinger, et al. Mutant p53 attenuates the anti-tumorigenic activity of fibroblasts-secreted interferon beta. *PloS one*, 8(4):e61353, 2013.

[55] Merck Millipore. *Getting started with ChIP-seq: experimental design to data analysis*, 2013.

[56] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D. Bader. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS ONE*, 5(11):e13984, November 2010.

[57] Huaiyu Mi, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. Large-scale gene function analysis with PANTHER Classification System. *Nature protocols*, 8(8):1551–1566, August 2013.

[58] Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics*, 18(2):279–290, March 2017.

[59] NCBI. Geo overview. `https://www.ncbi.nlm.nih.gov/geo/info/overview.html`. Accessed: 2021-02-27.

[60] R. I. Nicholson, J. M. Gee, and M. E. Harper. EGFR and cancer prognosis. *European Journal of Cancer (Oxford, England: 1990)*, 37 Suppl 4:S9–15, September 2001.

[61] Dongpin Oh, J. Seth Strattan, Junho K. Hur, José Bento, Alexander Eckehart Urban, Giltae Song, and J. Michael Cherry. CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection. *Scientific Reports*, 10(1):7933, May 2020.

[62] Magali Olivier, Monica Hollstein, and Pierre Hainaut. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, 2(1), January 2010.

[63] Henriette O'Geen, Lorigail Echipare, and Peggy J. Farnham. Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. *Methods in molecular biology (Clifton, N.J.)*, 791:265–286, 2011.

[64] Catia Pesquita. Semantic Similarity in the Gene Ontology. In Christophe Dessimoz and Nives Škunca, editors, *The Gene Ontology Handbook*, Methods in Molecular Biology, pages 161–173. Springer, New York, NY, 2017.

[65] Smitha Pillai and Srikumar P. Chellappan. ChIP on chip and ChIP-Seq assays: genome-wide analysis of transcription factor binding and histone modifications. *Methods in Molecular Biology (Clifton, N.J.)*, 1288:447–472, 2015.

[66] John E. Pool, Ines Hellmann, Jeffrey D. Jensen, and Rasmus Nielsen. Population genetic inference from genomic sequence variation. *Genome Research*, 20(3):291–300, March 2010.

[67] Mark Ptashne. On the use of the word 'epigenetic'. *Current Biology*, 17(7):R233–R236, April 2007.

[68] Zhiqiang Qin, Yi Wang, Jingyuan Tang, Lei Zhang, Ran Li, Jianxin Xue, Peng Han, Wei Wang, Chao Qin, Qianwei Xing, Jie Yang, and Wei Zhang. High LINC01605 expression predicts poor prognosis and promotes tumor progression via up-regulation of MMP9 in bladder cancer. *Bioscience Reports*, 38(5):BSR20180562, October 2018.

[69] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.

[70] Alan D. Radford, David Chapman, Linda Dixon, Julian Chantrey, Alistair C. Darby, and Neil Hall. Application of next-generation sequencing technologies in virology. *The Journal of General Virology*, 93(Pt 9):1853–1868, September 2012.

[71] Debasish Raha, Miyoung Hong, and Michael Snyder. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Current Protocols in Molecular Biology*, Chapter 21:Unit 21.19.1–14, July 2010.

[72] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, July 2016.

[73] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.

[74] Sophie Schbath, Véronique Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, and Jean-François Gibrat. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, 19(6):796–813, 2012.

[75] Adam M. Schmitt and Howard Y. Chang. Long Noncoding RNAs in Cancer Pathways. *Cancer cell*, 29(4):452–463, April 2016.

[76] Jing Shang, Fei Zhu, Wanwipa Vongsangnak, Yifei Tang, Wenyu Zhang, and Bairong Shen. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed research international*, 2014, 2014.

[77] P. Shannon. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003.

[78] Wei Shi and Yang Liao. *Rsubread/Subread Users Guide*. Subread development group, 2020. page 35.

[79] Thierry Soussi. The p53 Tumor Suppressor Gene: From Molecular Biology to Clinical Investigation. *Annals of the New York Academy of Sciences*, 910(1):121–139, January 2006.

[80] Thierry Soussi. Handbook of p53 mutation in cell lines. 01 2010.

[81] Dm Stasinopoulos and Robert Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23, 12 2007.

[82] Martin H Steinberg and Paola Sebastiani. Genetic modifiers of sickle cell disease. *American journal of hematology*, 87(8):795–803, 2012.

[83] Jenny Straiton, Tristan Free, Abigail Sawyer, and Joseph Martin. From Sanger sequencing to genome databases and beyond. *BioTechniques*, 66(2):60–63, 2019.

[84] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[85] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, 6(7):e21800, July 2011.

[86] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, January 2019.

[87] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, March 2013.

[88] H. Peter van Esse, T. Lynne Reuber, and Dieuwertje van der Does. Genetic modification to improve disease resistance in crops. *The New Phytologist*, 225(1):70–86, 2020.

[89] Miguel Vicente-Manzanares and Alan Rick Horwitz. Cell Migration: An Overview. In Claire M. Wells and Maddy Parsons, editors, *Cell Migration*, volume 769, pages 1–24. Humana Press, Totowa, NJ, 2011.

[90] Dawid Walerych, Kamil Lisek, and Giannino Del Sal. Mutant p53: One, No One, and One Hundred Thousand. *Frontiers in Oncology*, 5, December 2015.

[91] Dawid Walerych, Kamil Lisek, Roberta Sommaggio, Silvano Piazza, Yari Ciani, Emiliano Dalla, Katarzyna Rajkowska, Katarzyna Gaweda-Walerych, Eleonora Ingallina, Claudia Tonelli, Marco J. Morelli, Angela Amato, Vincenzo Eterno, Alberto Zambelli, Antonio Rosato, Bruno Amati, Jacek R. Wiśniewski, and Giannino Del Sal. Proteasome machinery is instrumental in a common gain-of-function program of the p53 missense mutants in cancer. *Nature Cell Biology*, 18(8):897–909, August 2016.

[92] James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, May 2007.

[93] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1):57–63, January 2009.

[94] James D Watson et al. *Molecular Biology of the Gene*. Cold Spring Harbor Laboratory Press, 2003.

[95] Elizabeth G. Wilbanks and Marc T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PloS One*, 5(7):e11471, July 2010.

[96] Longzheng Xia, Shiming Tan, Yujuan Zhou, Jingguan Lin, Heran Wang, Linda Oyang, Yutong Tian, Lu Liu, Min Su, Hui Wang, Deliang Cao, and Qianjin Liao. Role of the NFkB-signaling pathway in cancer. *OncoTargets and therapy*, 11:2063–2073, April 2018.

[97] Tiefu Xiong, Chenchen Huang, Jianfa Li, Shaokang Yu, Fangfang Chen, Zeng Zhang, Chengle Zhuang, Yawen Li, Changshui Zhuang, Xinbo Huang, Jing Ye, Fangting Zhang, and Yaoting Gui. LncRNA NRON promotes the proliferation, metastasis and EMT process in bladder cancer. *Journal of Cancer*, 11(7):1751–1760, 2020.

[98] Guangchuang Yu. *Gene Ontology Semantic Similarity Analysis Using GOSemSim*, pages 207–215. Springer US, New York, NY, 2020.

[99] Cen Zhang, Juan Liu, Dandan Xu, Tianliang Zhang, Wenwei Hu, and Zhaohui Feng. Gain-of-function mutant p53 in cancer progression and therapy. *Journal of Molecular Cell Biology*, 12(9):674–687, 07 2020.

[100] Wei Zhang and Hui Tu Liu. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Research*, 12(1):9–18, March 2002.

[101] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, September 2008.

[102] Chenguang Zhao and Zheng Wang. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific Reports*, 8(1):15107, December 2018.