uc3m | Universidad Carlos III de Madrid

University Degree in Biomedical Engineering
2021-2022

Bachelor thesis

"Development of bioinformatics workflows for the identification of neoantigens in tumor tissues"

Pilar Ballesteros Cuartero

Carlos Óscar Sánchez Sorzano (CNB-CSIC)

Maria Arrate Muñoz Barrutia (UC3M)

Lugar y fecha de presentación prevista

ABSTRACT

In recent years, there have been significant advances in genome and transcriptome sequencing using next-generation sequencing techniques. The information obtained by such techniques is being used in cancer research to create personalized treatments. One of the main focus areas is immunotherapy research and, more specifically, cancer-specific somatic mutations, known as neoantigens. These therapies based on neoantigens could lead to developing treatment vaccines that trigger immune responses against tumors. However, historically, the lack of efficient prediction algorithms for neoantigen prediction has hindered this research area. Nevertheless, with recent discoveries such as deep learning, predicting neoantigens is now possible and opens a new field for cancer immunotherapy.

This thesis presents an intuitive and user-friendly platform for human and mouse neoantigen discovery. This project introduces two significant advances in the neoantigen discovery field. First, it presents a novel prediction algorithm for humans that employs more information for the prediction compared to other software. It does this by using primary and secondary protein structure information and natural language processing for protein-encoding. The other novelty is the introduction of a more flexible mutation detection step. The platform allows the user to compare the sample with a control tissue or a reference genome. It also offers the possibility of introducing allograft and cell line samples and comparing them to extract common mutations, a current void in other existing software. Human and mouse samples are studied as proof of concept, and the complete analysis pipeline is presented.

Keywords: Neoantigen, epitope predictor, pipeline, cancer immunotherapy, T-cell epitope

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

mRNA: messenger RNA

RNA-seq: RNA sequencing

BLOSUM: BLOcks SUbstitution Matrix

SNP: Single nucleotide polymorphism

APC: Antigen Presenting Cell

CD8+: Cytotoxic T-Cell

MHC I: Major Histocompatibility Complex I

TCR: T-Cell Receptor

DC: Dendritic Cell

SNV: Single Nucleotide Variation

WGS: Whole Genome Sequencing

WES: Whole Exon Sequencing

GATK: Genome Analysis Toolkit

STAR: Spliced Transcripts Alignment to a Reference

VEP: Variant Effect Predictor

BERT: Bidirectional Encoder Representations from Transformers

NGS: Next Generation Sequencing

NLP: Natural Language Processing

NAP: NeoAntigens Prediction tool

FPKM: Fragments Per Kilobase Million

cKO: Conditional Knockout mice

1. INTRODUCTION

Cancer is one of the most significant health threats to society nowadays. It is considered the second leading cause of death globally after cardiovascular disease, accounting for nearly one in six deaths [5]. Furthermore, due to the increase in life expectancy, cancer incidence is expected to increase a 63% from 2018 to 2040 unless more effective therapies are developed [6].

Although current oncology techniques have shown an increase in survival rates of patients [6], side effects of these treatments or surgical procedures have posed a problem for cancer patients. Consequently, scientists are researching novel treatments that are more targeted and, therefore, more effective [7]. One of the main fields under expansion in translational cancer research is immunotherapy, a biological therapy that helps the immune system fight cancer [8]. In this field, personalized treatment vaccines have shown promising results during clinical trials. One of the main focuses of these vaccines is neoantigens. Neoantigens are tumor-specific proteins recognized by the immune system and therefore trigger an immune response against the tumoral cell. These neoantigens are being broadly researched to develop vaccines that trigger T-cell responses against cancer cells [9].

As a result of the potential of neoantigen-based vaccines, different programs have been created to predict human and mouse neoantigens from protein primary structure. Therefore, this thesis aims to create a robust and user-friendly software tool that provides a list of human neoantigens using a novel prediction technique. This technique employs primary and secondary structure information for the prediction, as recent studies showed that epitopes commonly have an alpha-helix shape. It also uses a natural language processing model for the protein encoding, which provides more information than the standard binarization used in the other software. Moreover, this thesis also aims to increase the flexibility in mutation detection for human and mouse samples, which is currently a void encountered, amongst others, by researchers at Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT).

2. THEORETICAL BACKGROUND

2.1 The central dogma of biology

2.1.1 Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) is an organic molecule found in all nucleated or non-nucleated cells and many viruses. DNA encodes the genetic information that characterizes an organism as, for example, the information for hereditary tracts and the creation of proteins [10]. DNA is made up of two strands linked together in a double-helix structure. Each strand comprises a backbone containing a sugar called deoxyribose and a phosphate group. To each of these backbone units, there is attached one of the four different nitrogenous bases: adenine (A), cytosine (C), thymine (T), and guanine (G) [11]. The link between the two strands is done by forming base pairs, which are hydrogen bonds of the form A-T or C-G [12]. The general structure of DNA can be seen in figure 2.1.

Fig. 2.1. Structure of DNA. DNA is made up of 4 different nitrogenous bases and a sugar-phosphate backbone. The two elements assemble in a double helix that is tightly packed, forming chromosomes. *Source:* [1]

DNA can be divided into coding and non-coding DNA. The main difference is that coding DNA contains the genetic information for creating proteins, while non-coding DNA does not [13]. In this thesis, we will consider a gene as the region of DNA encoding the information for protein synthesis.

2.1.2 Ribonucleic acid

Ribonucleic acid (RNA) is an organic molecule derived from DNA through transcription (figure 2.2). RNA is single-stranded and made up of a backbone containing ribose sugar, phosphate, and four different nitrogenous bases. In this case, the nitrogenous bases are adenine (A), uracil (U), cytosine (C), and guanine (G). Messenger RNA (mRNA), a specific type of RNA, contains the transcribed information from genes; therefore, it codes for proteins.

Other types of RNA are ribosomal RNA and transfer RNA, amongst others [14]. In addition, some RNA molecules control gene expression. Gene expression is the process by which the information coded in the DNA is transformed into a functional protein [15]. Therefore, regulating gene expression by RNA allows us to determine what DNA regions will undergo the transcription process. The ultimate goal of gene expression regulation is to determine the set of RNAs and proteins a cell will contain and thus its specific properties and characteristics [16]. Moreover, gene expression regulation will determine how many RNA transcripts are created from a gene, also called the gene expression level, and therefore the quantity of the protein encoded by that gene that will be present in the cell.

2.1.3 Proteins

A protein is a complex substance responsible for the proper functioning of the body [17]. Proteins are derived from mRNA through translation (figure 2.2). In this process, the nitrogenous bases of mRNA are grouped into triplets called codons. Each codon encodes for a different amino acid, the basic unit of proteins. The different amino acids are bound by strong covalent bonds to form the protein's primary structure.

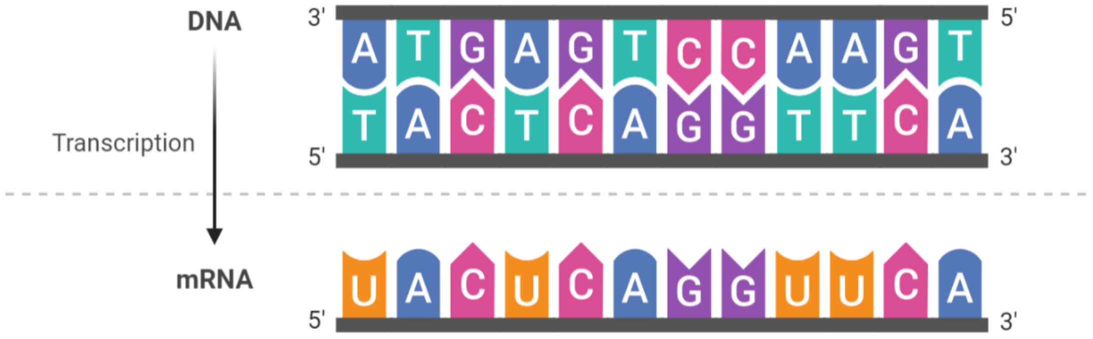


Fig. 2.2. The central dogma of biology. DNA is transformed into mRNA by transcription, and mRNA is transformed into proteins by translation. *Source:* [1]

Once the whole protein has been translated from the mRNA, we have the primary structure of the polypeptide chain. The protein then undergoes local folding due to the formation of hydrogen bonds between the atoms of the protein backbone [18]. This new folded protein conformation is referred to as the secondary structure. The protein can acquire two main conformations during this folding process. The first one is called α-helix, where the polypeptide chain adopts a helical structure, with each turn of the helix containing 3.6 amino acids. The second possible structure is the β-sheet, where two or more chain segments fold into a sheet-like structure [18]. The two different structures are shown in figure 2.3.

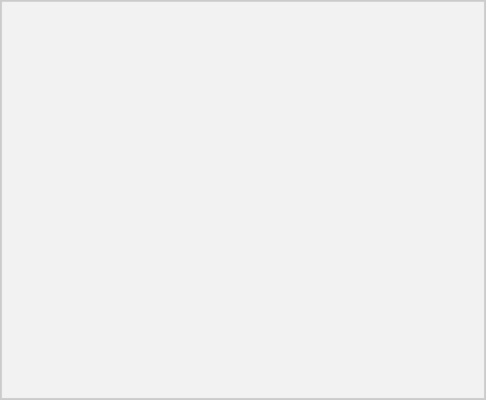


Fig. 2.3. Secondary structure of a protein. It can adopt an alpha helix (top) or a beta-sheet structure (bottom). *Source:* [1]

2.2 BLOSUM62 matrix

BLOSUM62 (BLOcks SUbstitution Matrix) is a scoring matrix for amino acid substitutions. BLOSUM62 matrix was constructed by analyzing the most common amino acid substitutions between proteins with a 62% similarity. Each pair of amino acids in the matrix receives a log-odds score based on how likely it is to see the substitution in nature. Therefore, pairs with higher score indicate that it is more likely to see that specific substitution in nature.

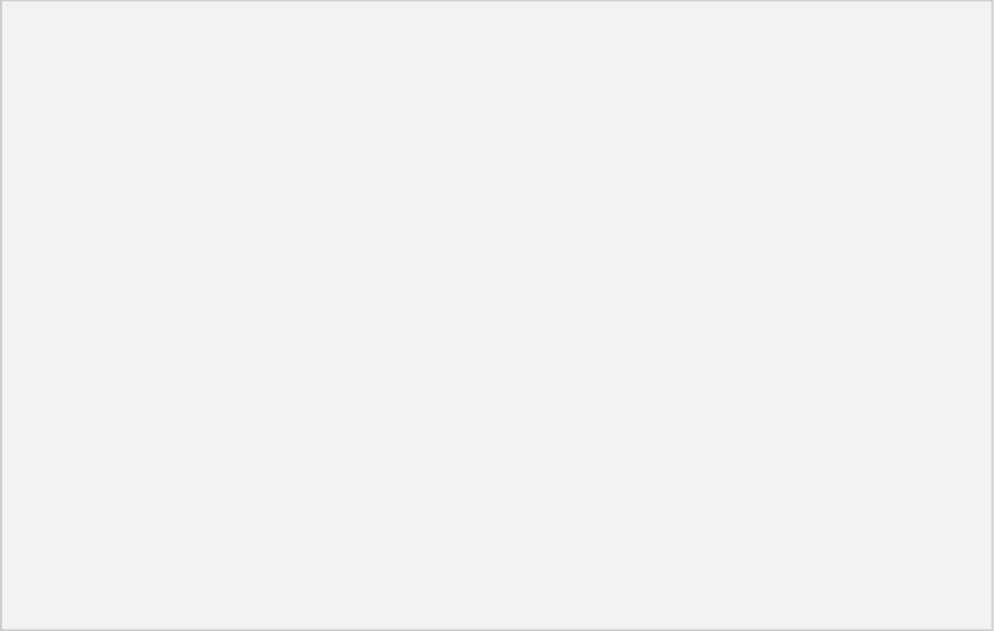


Fig. 2.4. BLOSUM62 matrix. *Source:* [2]

2.3 RNA sequencing

RNA sequencing (RNA-seq) is a technique used to study the transcriptome of cells, which is the set of all mRNA molecules present in the cell at a specific time [19]. More specifically, it provides the sequence and quantity of all the mRNA molecules in a cell [20]. This technique is widely used in the bioinformatics field as it allows to study which genes are being expressed in each cell and their amount, also called their expression level. In addition, the information this tool provides is used to determine the primary structure of the proteins present in the cell, as proteins are derived from the mRNA by known mechanisms.

2.3.1 RNA-seq workflow

The first step in performing an RNA-seq analysis is to extract an mRNA sample belonging to the cells we want to study. Once we have the sample, first, we need to fragment each mRNA molecule into smaller segments. The resulting pieces are transformed into complementary DNA (cDNA) by reverse transcription [20]. Next, a set of adapters are added to each end of the segments. The adapters are constant sequences needed for segmentation [21]. More specifically, they are used to amplify the fragments by polymerase chain reaction (PCR) to replicate each cDNA fragment. The adapters are also useful as primers to start the first sequencing reaction in each fragment [21]. The preparation of the library to perform RNA-seq can be seen in figure 2.5.

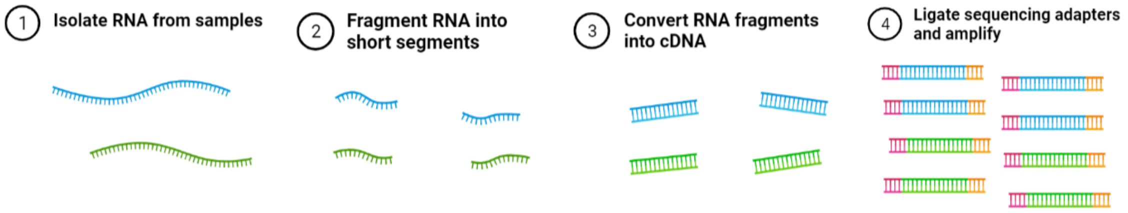

1 Isolate RNA from samples
2 Fragment RNA into short segments
3 Convert RNA fragments into cDNA
4 Ligate sequencing adapters and amplify

Fig. 2.5. Preparation of the cDNA library for RNA-seq. The RNA fragments are trimmed into shorter segments. Then, they are converted into cDNA, and adaptors are added. *Source:* [1]

Following RNA-seq terminology, each cDNA fragment will be referred to as a read. Once the different reads are prepared, they are introduced into a sequencing machine. Each read is attached to the device by the adapter sequence added in the previous steps. Next, fluorescently tagged nucleotides are introduced into the chamber and compete to join the first nucleotide present in the reads. Only the complementary nucleotide attaches to it, meaning that if we had an A in our first position, only T would be able to join. After the nucleotide has been appended, it is excited by a light source, which causes a characteristic fluorescence signal emission. By recording the different signals emitted at each read position, its sequence can be determined. The last steps explained can be seen in figure 2.6.

Fig. 2.6. RNA-seq procedure. The complementary base joins to the base being sequenced. Then, a fluorescent light excitation makes the complementary to release light, and the base is identified. *Source:* [1]

2.4 Cancer

Cancer is a condition where cells in a specific body part grow and reproduce uncontrollably. Cancerous cells can also spread to other body tissues in a process known as metastasis [5].

In healthy cells, a set of genes control cell division. In order to control this process, cells need a balance between the signals that trigger cell growth and those that suppress it. Healthy cells also need some genes that mark them for programmed death, known as apoptosis, when they are damaged [22]. On the other hand, cancerous cells have accumulated mutations in the genes that control proliferation and apoptosis. The mutations cause the genes to malfunction, so the cancer cell starts proliferating uncontrollably [22]. Moreover, as the genes marking the cells for apoptosis are also damaged, the diseased cancer cell continues dividing unconstrained, as shown in figure 2.7.

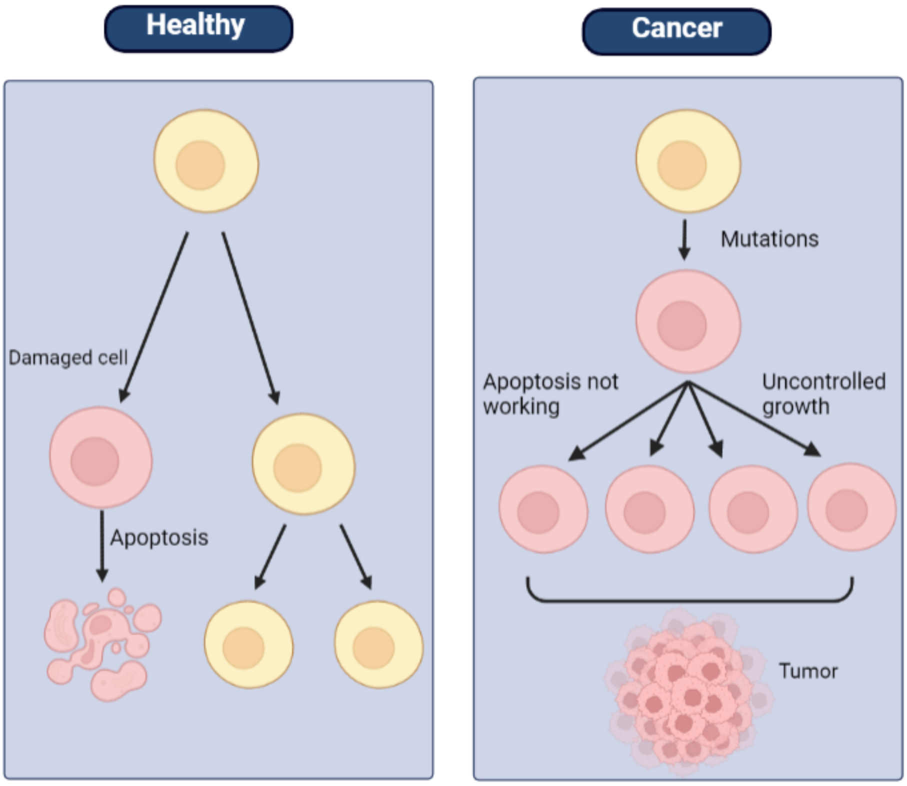


Fig. 2.7. Cancer mechanism. In a healthy organism (left), when a cell is damaged, it undergoes apoptosis. In an organism with cancer (right), damaged cells do not undergo apoptosis and grow uncontrolled, forming a tumor. *Source:* [1]

Cancerous cells, therefore, have mutations in their genes. We can classify those mutations according to their effect on the DNA sequence (figure 2.8). Some of the most frequent ones are substitution, deletion, and insertion. Substitution is the replacement of one or more nitrogenous bases with a different pair of nucleotides. Deletion is the loss of one or more base pairs, and insertion is the addition of one or more base pairs [23].

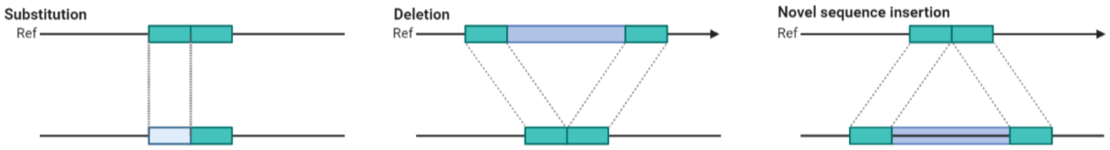
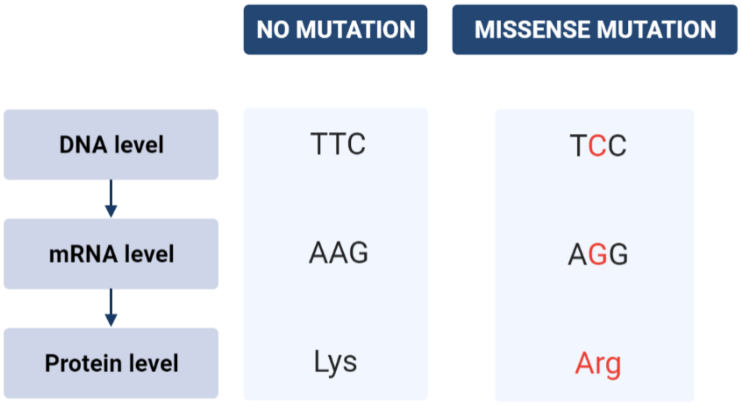


Fig. 2.8. Possible mutations in DNA. Substitution (left), deletion (center) and novel sequence insertion (right). Ref: reference sequence where the mutation is happening. *Source:* [1]

When dealing with mutations, discussing Single Nucleotide Polymorphisms (SNP) is crucial. These are variations in a single nucleotide, but unlike the abovementioned mutations, they do not arise from cancerous mutations. If more than 1% of the population does not have the same nucleotide at a specific position of the DNA, it is considered an SNP [24]. Therefore, these mutations are naturally occurring, shared in different individuals, and not considered cancerous.

The mutation in the DNA of cancerous cells can also cause changes in protein synthesis or function. This is due to the fact that proteins are derived from genes, which in the case of cancer, are mutated. These mutant genes can cause the generation of novel proteins only present in cancer cells, which is a good focus for immunotherapies. In this thesis, we will focus on the mutations that cause an amino acid in the protein's primary sequence to change, called a missense mutation [25] (figure 2.9).



	NO MUTATION	MISSENSE MUTATION
DNA level	TTC	TCC
mRNA level	AAG	AGG
Protein level	Lys	Arg

Fig. 2.9. Healthy protein (left) vs. missense mutation (right). *Source:* [1]

2.5 The immune system

The immune system is a network of organs, tissues, cells, and the substances produced by them. The primary function of this system is hosting the body's defense mechanism against pathogens or other diseases [26]. Immunity is divided into innate and acquired immune systems, which work closely together to produce a highly effective immune response [27].

2.5.1 Innate immunity

The innate immune system is the first line of defense against pathogenic substances entering the body. It is characterized by offering a fast, nonspecific immune response, which means it acts the same way for any substance foreign to the body [27].

The innate immune system has two main defense mechanisms, as shown in figure 2.10. The first mechanism is the physical barriers that separate the inside from the outside of the body and prevent foreign substances entrance. The second line of defense is activated if the physical barriers are penetrated. This line of defense involves cells that phagocyte the germs, digest them, and show small fragments on their surface to present them to cells from the adaptive immune system.

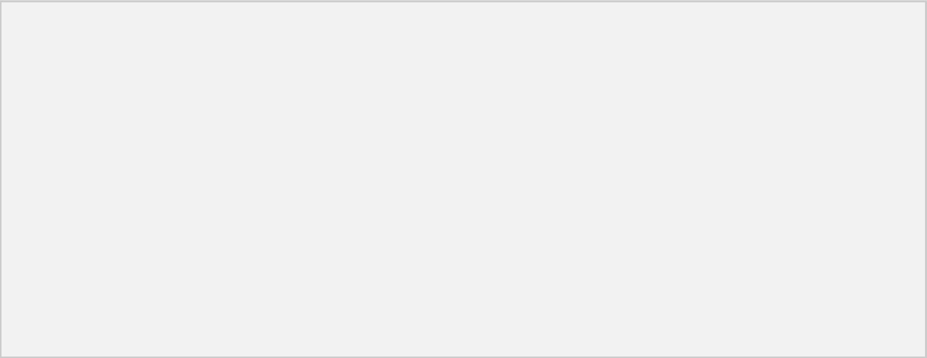


Fig. 2.10. Innate immune system. It is composed of physical barriers and phagocytes. In the lower branch, phagocytosis by a macrophage is shown. *Source:* [1]

2.5.2 Acquired immunity

The acquired immunity refers to the activation of T and B lymphocytes after encountering an antigen, a substance recognized as foreign. Acquired immunity has two main characteristics, its specificity and its memory. The former refers to the immune system's capacity to recognize the pathogen and trigger a unique immune response depending on the substance. Due to the need to recognize the pathogen, acquired immunity is slower than innate, but it also offers a more robust response once activated. The other characteristic is the presence of memory cells. They will remember the pathogen, producing a faster immune response if the same pathogen invades the body again [27].

The acquired immune system can trigger two different types of responses figure 2.11. First, antibody-mediated immunity consists of destroying pathogens with the antibodies generated by B lymphocytes. Antibodies are proteins produced by B cells that identify and neutralize foreign substances in the body. The second type of response is the cell-mediated immunity, in which the pathogen is destroyed due to the direct interaction with T lymphocytes [27]. Once the T lymphocytes have recognized a foreign substance, they are activated, producing two kinds of cells: cytotoxic CD8+ cells, the ones attacking the pathogens, and helper CD4+ cells, which have a helper role vital in the immune function [28].

Fig. 2.11. Acquired immune system. The upper branch shows the humoral immunity with B-lymphocytes and antibodies and how they neutralize pathogens. The lower branch shows cell-mediated immunity by T-lymphocytes and how T-cells kill an infected cell. *Source:* [1]

Cytotoxic T-cells can recognize infected or diseased cells and trigger their apoptosis while not interfering with healthy cells. That ability is illustrated in figure 2.12.

Nevertheless, the mechanism by which CD8+ cells recognize the infected cells and trigger their apoptosis is a particular and complex procedure.

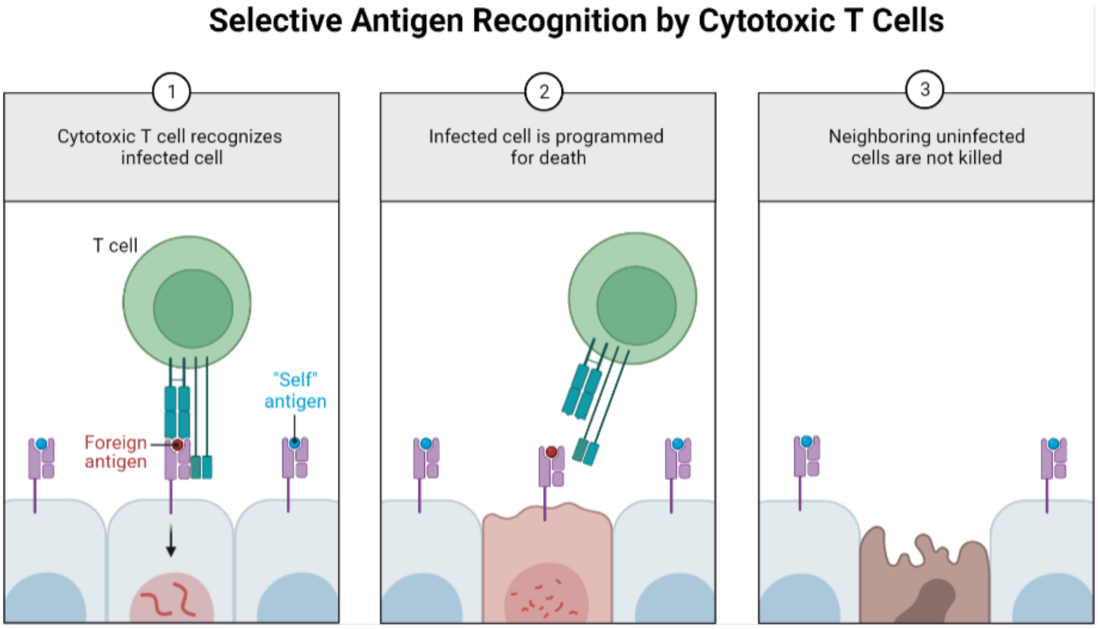

Selective Antigen Recognition by Cytotoxic T Cells
1
Cytotoxic T cell recognizes infected cell
T cell
"Self" antigen
Foreign antigen
2
Infected cell is programmed for death
3
Neighboring uninfected cells are not killed

Fig 2.12. Selective recognition of infected cells by T cells. The T cell recognizes the foreign antigens of infected cells and induces their apoptosis. *Source:* [1]

At the molecular level, recognizing diseased cells is complex and has several mechanisms. Every somatic cell in the body exhibits on its surface a receptor called Major Histocompatibility Complex I (MHC I) [29]. MHC I complex exposes on its surface fragments of peptides produced within the cell in order to signal the cell's physiological state [30]. Therefore, if a cell is healthy, it exposes peptide fragments recognized by T cells as part of the body, also known as self-antigens, and an immune response is not triggered. On the other hand, if the cell is diseased, it starts producing mutant peptides that eventually show up in the MHC I. These mutant peptides are unique to the diseased cell and are known as neoantigens.

Moreover, cytotoxic T cells have a receptor called T-cell Receptor (TCR) coupled to their surface. This TCR can interact with a fragment of the peptides exposed on the MHC I, known as the epitope. The cytotoxic T cell can then determine if they are self-antigens or neoantigens and trigger an immune response against the cell in case it is the latter. This immune response consists of releasing granzyme and perforin, enzymes that cause cell apoptosis. The mechanism can be seen in detail in figure 2.13.

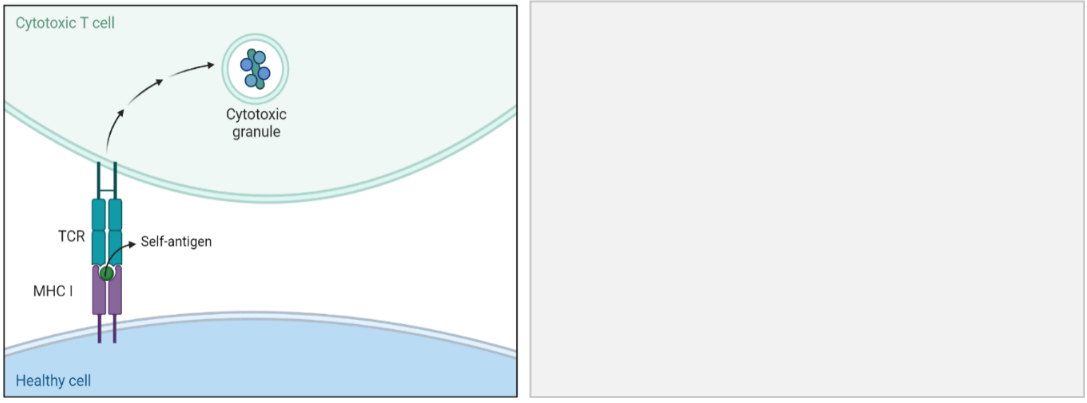


Fig. 2.13. Cytotoxic T cells recognition of healthy and cancerous cells. In the healthy (left), the T Cell Receptor (TCR) recognizes the self-antigen exposed in the Major Histocompatibility Complex (MHC) I as part of the body and does not trigger a response. In the cancerous cell (right), the TCR recognizes the neoantigen as diseased and triggers the cell apoptosis with perforin and granzyme. *Source:* [1]

There is a second method of antigen presentation performed by MHC II. MHC II is only present in immune cells called antigen-presenting cells (APCs) and exhibits protein fragments of organisms external to the cell. This mechanism identifies invading substances and triggers an immune response against them.

2.6 Immunotherapy

Immunotherapy is a therapeutic approach that targets or manipulates the immune system [31], and it has revolutionized cancer research prospects. Immunotherapy can be divided into active or passive treatments. Passive treatments include administering immune cells generated ex-vivo, so they do not stimulate the host's immune response. On the contrary, active treatments include anti-cancer vaccines, which activate the host immune response [32].

The primary goal of most immunotherapy strategies relies on eliciting a tumor-specific T cell response to engage the immune system in the fight against cancer [3]. One of the main focuses on eliciting the T-cell response has been therapeutic cancer vaccines. More specifically, researchers are focusing on developing neoantigen-based vaccines. Neoantigens are only present in cancerous cells and, therefore, can induce a more robust immune response and cause less autoimmune-related toxicities than other antigens [3]. These advantages, combined with the ability to predict neoantigens through NGS techniques from the tumor's DNA, have made neoantigen-based vaccines a promising approach for cancer immunotherapy.

The pipeline to generate neoantigen-based vaccines is complex and has several steps, as shown in figure 2.14. First, the tumor DNA or RNA must be extracted and sequenced using NGS techniques. Then, the neoantigens have to be predicted using bioinformatics workflows. The workflows identify which of the mutant peptides are specific to the cancer cell and predict which of them will be exposed in the MHC I complex of cells. Finally, vaccines are created using several neoantigens, and the patient is monitored for neoantigen-specific immune responses [3].

Fig. 2.14. Pipeline for the generation of neoantigen vaccines. DNA from the tumor is sequenced, and neoantigens are identified from it. The vaccine is developed using those neoantigens and administered to the patient. *Source:* [3]

The molecular basis behind cancer vaccines relies on cell-mediated immunity. Therefore, most cancer vaccines under development try to elicit antigen presentation by APCs to generate long-lasting T cell immunity against specific antigens. The most effective APC is the Dendritic Cell (DC); therefore, most vaccines use them as effectors [33].

The rationale behind cancer vaccines consists of injecting the processed antigens into the organism. Those antigens are phagocyted by DCs and expressed on their MHC II complex. Once they have expressed the antigen, they migrate to the lymph nodes where T cells are. The DCs present the antigen to the T cells, activating them. Once activated, they proliferate into cytotoxic CD8+ cells and helper CD4+ cells. CD8+ cells then travel to the tumor site and trigger apoptosis on the tumor cells presenting the identified neoantigens [4]. The whole cycle can be seen in figure 2.15.

Fig. 2.15. Pipeline of immunotherapy vaccines. Antigens are injected with the vaccine. Antigen Presenting Cells (APC) recognize the antigens and migrate to the lymph node. T cells and B cells are activated. B cells trigger antibody-dependent cellular cytotoxicity (ADCC), and T cells infiltrate the tumor and attack cells. They attach the cells releasing perforin (PFN), granzyme B (GzmB), gamma interferon (IFNγ), and tumoral necrotic factor-alpha (TNFα). *Source:* [4]

3. STATE OF THE ART

3.1 Available software for humans

Epitope prediction has proven to be a fundamental step in creating neoantigen-based vaccines. Consequently, several programs have been created to make this prediction, starting from sequencing data or directly from mutant peptides.

The first type of existing software focuses on predicting neoantigens from sequencing data. Those programs have some fundamental differences, for example, the type of mutation they can identify on the data. The mutations can be Single Nucleotide Variations (SNV), SNVs plus indels (insertions + deletions), only indels, or gene fusions. Moreover, another distinction between tools is the origin of their data, which can come from whole-genome sequencing (WGS), whole exon sequencing (WES), or transcriptome sequencing (RNA-seq) [34]. The different tools with the characteristics just mentioned can be seen in table 3.1.

TABLE 3.1. COMPARISON OF EXISTING SOFTWARE FOR NEOANTIGEN DETECTION.
Source: [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] [48] [49] [50] [51]

It is important to note that choosing transcriptome sequencing allows the user to obtain extra information regarding the expression level of different genes at the mRNA level and other phenomena such as alternative splicing. On the other hand, using protein information allows the user to know exactly what proteins are being expressed in the cell, as the presence of the gene in the mRNA is not always a determinant of their translation [34].

Most of the tools in table 3.1 use NetMHC [52] or NetMHCpan [53] pretrained networks to predict the neoantigens from peptide fragments. Therefore, each pipeline's novelties consist of how the extraction of the mutant peptides from RNA-seq or WGS data is performed. Most tools use a combination of WGS/WES and RNA-seq data. The WGS is used to call the somatic mutations with higher accuracy, while RNA-seq is used to quantify expression level.

The analysis pipeline is similar in every program, and they use the same tools, such as the Genome Analysis Toolkit (GATK) [54] for data preprocessing and MuTect [55] for mutation detection. However, a common feature in all programs is that they have a fixed structure on the input sequences. For example, most of them allow only a tumoral file as an input [EpiSeq [35], ScanNeo [49], NeoFuse [50], pVACsec [39], INTEGRATE-neo [51], MuPeXI [40]], while others only allow submission of both tumoral and control files [Epidisco [41], OpenVax [42], TSNAD [48], pTuneos [45]]. Moreover, only a limited number of tools accept raw RNA-seq or WGS data [TIminer [36], OpenVax [42], TSNAD [48], pTuneos [45]], while most other tools need prior conditioning [34].

Next, the most common tools for neoantigen prediction from peptides will be studied. Most of the abovementioned tools use NetMHC [52] or NetMHCpan [53] for this task. NetMHC accepts peptides of 8-10 amino acids long, while NetMHCpan takes 8-14 amino acid sequences. The selection of sequence length is because most epitopes are those lengths [52] [53].

Moreover, NetMHC and NetMHCpan predict using protein primary structure. NetMHCpan performs a BLOSUM encoding, adding extra information such as the length of the insertion/deletion or the length of the flanking regions [53].

3.2 Previous work on the NAP-CNB server

This thesis is the continuation of a project that has been under development at the National Center of Biotechnology (CNB-CSIC) by other UC3M students for some academic years. The starting point of the project is therefore the workflow and algorithms developed by Carlos Wert Carvajal, Paola Núñez Hernández and Sara Guillén Fernández-Micheltorena. Therefore, it is essential to briefly explain their work to understand the novel developments introduced in this thesis.

3.2.1 NAP-CNB server

In his final degree thesis, Carlos Wert Carvajal created a workflow for discovering neoantigens in mouse samples. He also integrated this pipeline into a webpage called NAP-CNB, accessible through the link https://biocomp.cnb.csic.es/NeoantigensApp/ [56].

The workflow divides into two different steps, as shown in figure 3.1. The first is a preprocessing step that takes as an input RNA-seq data belonging to a mouse tumor sample and returns all the proteins that have mutated in the tumor with respect to healthy cells. The second step takes the preprocessing output and performs neoantigen prediction on the mutant peptides, determining which will be neoantigens. The final output of the pipeline is a list of putative neoantigens, along with the probability of them being so and their expression level.

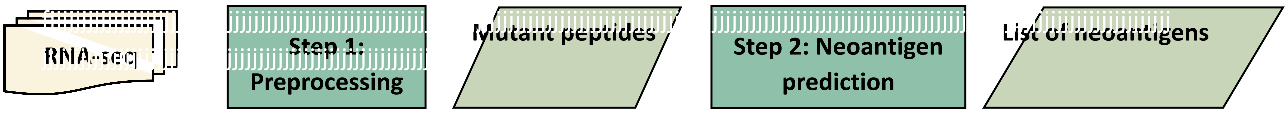


Fig. 3.1. Pipeline of the NAP-CNB tool. It is formed by a preprocessing step and a Neoantigen prediction step.

Regarding the sample preprocessing, its pipeline is shown in the figure 3.2. First of all, quality analysis is performed using FastQC [57]. This inspection allows checking if the data was sequenced correctly and if there is some error in specific regions [57]. Once the quality check is performed, several steps must be followed according to GATK [54] best practices.

During the preprocessing, we first need to select the fragments of interest. This is the process grouped as file preprocessing in figure 3.2. First of all, we should recall that our input

Fig. 3.2. Flowchart of the preprocessing step of NAP-CNB.

data is the sequence of several DNA fragments, but we do not know where each fragment came from in the genome. Localization of the fragments is critical to understanding what chromosome and gene each of them belong to. The first step is the alignment of each of the reads with a reference genome using Spliced Transcripts Alignment to a Reference (STAR) [58]. This way, we can identify the exact location of each read in the genome. Next, the duplicate fragments created with the PCR technique are removed using the tool Picard: remove duplicates [59]. PCR amplification was needed during sequencing in order to obtain more accurate results. However, by doing this, expression levels are altered, as sequences appear overexpressed compared to reality. Therefore, we need to recover the original expression level in the sample to use it for further analysis. The third step consists of obtaining the exons of the DNA with GATK: SplitNCigars [60]. It is known that proteins come from the coding DNA regions, also known as exons, so we need to discard all the sequences that do not belong to an exon as we are not interested in them. Finally, a step called GATK: Base recalibrator [61] identifies if the mismatches with the reference genome are due to a mutation, an error sequencing, or the presence of an SNP in that base.

Once the file preprocessing is complete, variant calling is done using MuTect *tumor-only mode* [55]. This tool analyses all the reads and identifies SNVs, as well as deletions and insertions. The tool is coupled with an SNP database that will allow us to differentiate between mutations specific to the tumor and SNPs. Moreover, in this method, mutations are detected by comparing the sample with a reference genome that the tool provides. The output of this step is a list of all the mutations detected in the sample. Then, the mutations are run through a Variant Effect Predictor (VEP) [62] to identify which ones are missense mutations, as we are only interested in that specific type. It also obtains the sequence of the mutated protein. Finally, cufflinks [63] was used to obtain each gene's expression level, crucial information for vaccine development. The final output is a list of mutant peptides that are only present in the tumor, along with the expression level of each peptide.

Once the list of mutant peptides is obtained, neoantigen prediction must be performed, step two in the figure 3.1. In this case, for mouse samples, the NAP-CNB server applies a one-hot encoding and a long short-term memory neural network that yields the list of peptides identified as neoantigens and the probability that they are so.

3.2.2 Human epitope predictor

The human epitope prediction pipeline was developed by Sara Guillén Fernández-Micheltorena and Paola Núñez Hernández in their final degree project. Their pipeline took a list of protein fragments of 30 amino acids and identified which of those fragments were epitopes. The whole workflow can be seen in figure 3.3.

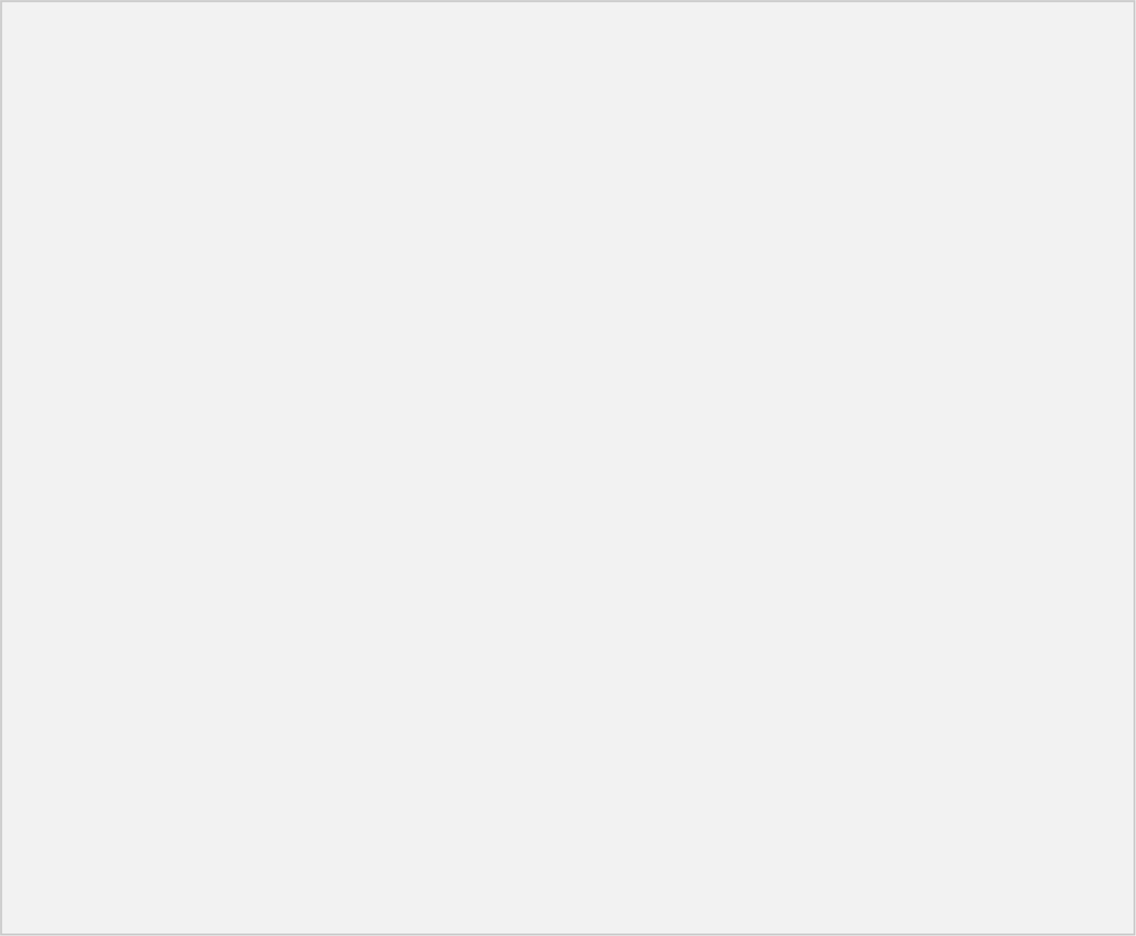


Fig. 3.3. Epitope predictor workflow

The first step was implementing a secondary structure prediction method. It was done using ProteinUnet [64], a pretrained neural network that takes a protein and gives its secondary structure. ProteinUnet predicts three different conformations: alpha-helix (H), beta-sheet (B), and coil-like conformation (C). An example of this prediction can be seen in figure 3.4.

Primary structure: PPVCPLLSPSFLPCPFLGATASSAISPSML

Secondary structure: CCCCCCCCCCCCCCCCCCCCCCCCCCHHHH

Fig. 3.4. Primary structure of a fragment of a peptide in amino acid notation (30 amino acids) and its secondary structure predicted (30 amino acids)

They also used the natural language processing tool Bidirectional Encoder Representations from Transformers (BERT) [65] to create a language of proteins. First, they trained two different networks, one that would learn to recognize the primary protein structure and a second one that recognized the secondary structure. Once their networks were trained, they obtained 768-long word embedding vectors of each input protein sequence. Finally, those two vectors were introduced in a neural network, shown in figure 3.5, which predicted which of them were epitopes and their probability of being so.

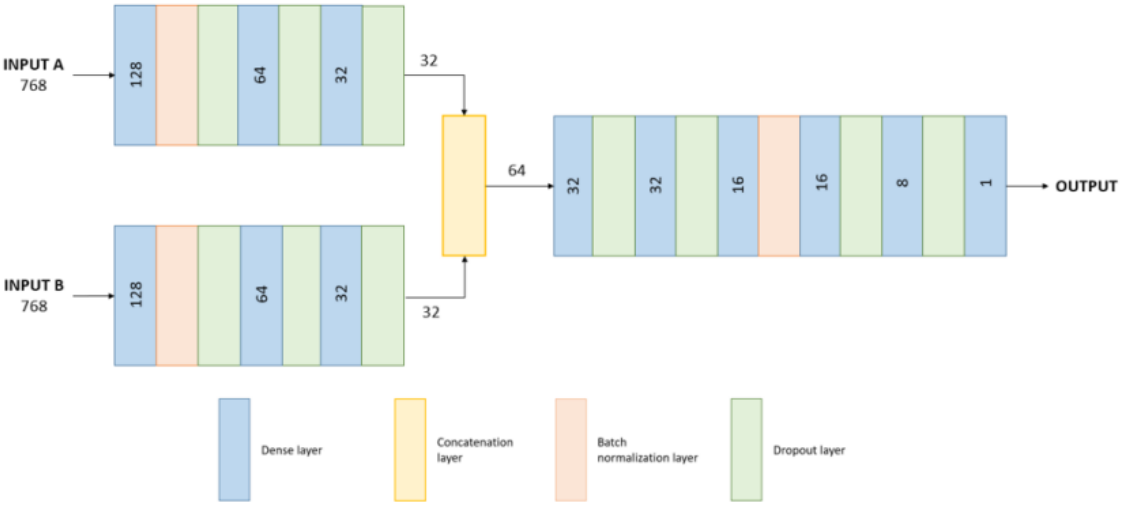


Fig. 3.5. Neural network structure of the human epitope prediction pipeline

4. OBJECTIVES

The objectives of this thesis were developed at the Centro Nacional de Biotecnología (CSIC) in collaboration between the Biocomputing Unit lead by Dr. Carlos Óscar Sánchez Sorzano and the laboratory of Dr. Esteban Veiga Chacón.

This thesis has three main objectives, two of them using as a foundation the work developed by Carlos Wert, Sara Guillén Fernández-Micheltorena, and Paola Núñez Hernández, already mentioned under the section state of the art.

The first objective is the development of a flexible and user-friendly tool for neoantigen discovery in human samples. The aim is to create a pipeline that takes RNA-seq data and outputs a list of putative neoantigens for humans. The second objective is the implementation of a more exhaustive method for mutation detection. This new method will be designed to be available for mouse and human samples. Finally, the third objective is the integration of the new developments of the first two objectives in the NAP-CNB server. This way, the new implementations will become available for any other scientist doing immunotherapy research.

5. MATERIALS AND METHODS

5.1 Materials

Databases

The project relies on online and open-access databases in different pipeline steps. First, several databases of Ensembl were used, a bioinformatics project to organize biological information belonging to sequences of large genomes [66]. The first was a genome database that provided the entire human genome sequence. The second one was the gene database, which provides information about the location of the genes in the human genome. The third database, dbSNP, was obtained from the National Center of Biotechnology Information (NCBI) [67]. It was used to obtain the sequence and location of the SNP in the human genome. Finally, the UniProt [68] database provided information about all known protein sequences and was used to obtain the sequence of mutant peptides.

FastQC

FastQC [57] is a tool that performs quality control on the sequencing data coming from NGS methods. It allows to analyze the data and look for sequencing errors before further analysis [57]. FastQC is used at the beginning of the project to study the quality of the data.

STAR

Spliced Transcripts Alignment to a Reference (STAR) [58] is a software that performs accurate and fast alignment of Next Generation Sequencing (NGS) data to a reference genome. The user provides the reference genome, which should be conditioned to be used by the software. This tool aligns our data with the reference genome and identifies where each read belongs in the genome.

Genome Analysis Toolkit

The genome analysis toolkit (GATK) [54] was developed at the Data Science Platform of the Broad Institute. It offers different tools that aim to perform variant discovery on NGS data [54]. In this project, GATK was used in the preprocessing step to condition the raw RNA-seq data to detect mutations in the sample.

Variant Effect Predictor

Variant Effect Predictor (VEP) is a tool developed by Ensembl that determines the effect of mutations on the protein sequence [62]. This tool was used to determine which DNA mutations produced missense mutations in the primary protein structure.

Bcftools

Bcftools is a set of utilities that are used to manipulate files in Variant Calling Format (VCF) and their Binary counterpart (BCF) [69]. This thesis uses this tool to compare VCF files in the preprocessing step.

Cufflinks

Cufflinks [63] is a program created to assemble RNA-seq transcripts, quantify their abundance and perform a differential expression analysis on them. In this thesis, it is used to determine which genes are up-regulated or down-regulated in the tumoral file with respect to the control file.

ProteinUnet

ProteinUnet [64] is a tool to predict the secondary structure of proteins from their primary structure. The code and models were downloaded from https://codeocean.com/capsule/2521196/tree/v1. The downloaded files included two trained models for predicting and a script to run the software. However, the codes were modified by Sara Guillén to obtain the desired output.

Bidirectional Encoder Representations from Transformers

The BERT [65] code is accessible from Github https://github.com/google-research/bert. The developers made the files available to perform the pretraining of the language model and the embedding to extract the desired features from our dataset. In this thesis, only the code for extracting the features is used. Sara Guillén and Paola Núñez did the pretraining, and their trained models are directly used.

Python

One of the programming languages used in this thesis was Python. The Python web framework Django was used to introduce all the developed pipelines into the webpage NAP-CNB. *Django* is an open-source framework designed for web page development

[70]. In addition, Keras [71], another open-source python library, was used to run all the neural networks of the pipeline.

Bash

Bash is a command-line shell user interface and a scripting language [72]. All of the preprocessing is done using toolkits launched through the bash shell. In order to make an automated pipeline, those commands are gathered in a bash script file that the bash shell can execute in Unix.

5.2 Methods

The first two objectives of this thesis, the development of a human pipeline and the implementation of a more exhaustive method for mutation detection, are treated separately and therefore have two well-differentiated methodologies. Once the two objectives were achieved, the new developments were integrated into NAP-CNB over the last weeks of work.

5.2.1 Development of an intuitive tool for neoantigen discovery in humans

This part of the thesis project aimed to develop a novel pipeline for the extraction of human neoantigens from raw RNA-seq data. This pipeline consisted of three main blocks: a preprocessing step, a postprocessing step, and a neoantigen discovery step, as shown in figure 5.1.

Fig. 5.1. Pipeline of the novel neoantigen prediction tool for humans. It consists of a preprocessing step, a postprocessing step, and a neoantigen prediction step.

Step 1: Preprocessing

The preprocessing step takes raw RNA-seq data and outputs a list of the mutant peptides discovered in the cancerous sample. This step took the backbone from the preprocessing step used in the NAP-CNB server, but several modifications were done to adapt it for humans. Those modifications can be seen in black in the diagram shown in figure 5.2.

- Generation of human library

The first change in the original pipeline was creating a library of databases suitable for analyzing human data. The library comprises three different constantly interacting databases that should be compatible. The three databases were downloaded to use version 38 of the human genome. This version is the genome sequence's last release, and all databases must share it to avoid incompatibilities.

The first database was downloaded from Ensembl [66] and comprised the whole genome in fasta format. It contains the whole genome sequence, separated by chromosomes, and the exact location of those chromosomes inside the genome. The sequence was provided by the Genome Reference Consortium, a project that tries to provide a unique reference for the genome sequence. Therefore, some of the database relevant information is disclosed in table 5.1.

TABLE 5.1. INFORMATION ABOUT THE GENOME DATABASE FROM ENSEMBL

Assembly	GRCh38.p13
Assembly Provider	Genome Reference Consortium
Annotation Provider	Ensembl
Annotation method	Full genebuild
Database version	106.38
Base Pairs	3,1 Gb
Coding Genes	20.471

The second database used was a genes database from Ensemble [66]. It contained information about the location of the different genes in the genome. The combination of the two first databases was crucial to identifying the exact location of each RNA-seq fragment and the gene to which it belongs.

Finally, an SNP database (dbSNP) was downloaded from NCBI [67]. It was downloaded in the GRCh38 version. As it was coming from a different organization,

Fig. 5.2. Flowchart of the preprocessing step for the human pipeline

the used nomenclature used in the database had to be checked to ensure that it was compatible with the previously mentioned databases. The dbSNP contained annotations about the different SNPs commonly occurring in humans and their exact location in the genome.

- Generation of a STAR genome index for alignment

Genome indexing is a fundamental step for bioinformatics workflows. It increases the tool's performance by allowing rapid access to different parts of the reference genome [73]. Furthermore, as we have fragments from the whole genome in our data, indexing the genome allows us to directly jump to areas of interest instead of scanning the whole sequence. The STAR [58] program used for alignments has a specific method to generate such an index. It is done with the command *genomeGenerate,* and it takes as inputs the whole genome and genes databases.

- VEP conditioning for humans

VEP [62] is a software that can be used online and offline. After reading the VEP documentation, it was discovered that VEP online version shared some private information of the files with an online server [62]. As a consequence, the offline method was chosen to run the program. A cache had to be downloaded to run the program on the local machine. The cache download allowed us to get all the species' information in one single connection and store it on the local disk. The machine's preexisting cache belonged to mouse, so a human cache needed to be downloaded to make VEP functional for human data.

The next step was modifying the code used to obtain the sequence of the peptides that had missense mutations. For mice, we were obtaining 12 amino acid long proteins, as most of the epitopes in mice were shorter than this. However, after a study of the length of all human epitopes, shown in figure 5.3, it was discovered that the ideal length was 30 amino acids long proteins, given that the maximum epitope length was 26 amino acids. The code created by Carlos Wert was then modified to obtain a list of 30 amino acid long proteins as an output of the preprocessing step.

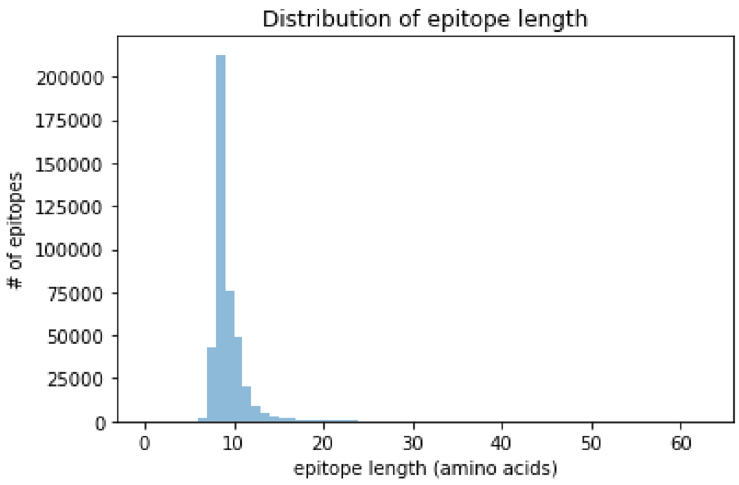


Fig. 5.3. Histogram of the different human epitopes' lengths. The peak is on 8 amino acids, and the longest epitope is 26 amino acids.

Step 2: Postprocessing

Once the preprocessing was adapted for humans, a postprocessing step was designed to improve the prediction accuracy. One of the most common problems in Natural Language Processing (NLP) is that the system is susceptible to changing a single amino acid. Therefore, the created models by BERT [65] are very dependent on the particular amino acids of the sequence.

The chosen method to decrease this bias consists of generating similar sequences to the submitted one, changing only specific amino acids, and making a final classification according to their overall prediction.

In this method, each peptide generates 30 additional sequences with one amino acid change at each position. Each amino acid was replaced by the most common substitution in nature, chosen from the BLOSUM62 matrix, to ensure the most realistic substitution occurred. An example of the output of the postprocessing is shown in figure 5.4.

Original sequence	**Augmented sequence**
PPVCPLLSPSFLPCPFLGATA	TPVCPLLSPSFLPCPFLGATA
	PTVCPLLSPSFLPCPFLGATA
	PPICPLLSPSFLPCPFLGATA
	PPVAPLLSPSFLPCPFLGATA

Fig. 5.4. Postprocessing example. The original sequence (left) is augmented using BLOSUM62 to find the most common substitution for each amino acid (right).

Step 3: Neoantigen Prediction

The neoantigen prediction consists of predicting which of the mutant peptides discovered in the cancerous sample can be considered neoantigens. This step takes the backbone from the epitope predictor developed by Sara Guillén and Paola Núñez. The original work was prepared to take 30 amino acid long proteins and predict on them, but it had never been used to predict on sequences obtained from raw RNA-seq data. Thus, modifications were necessary to adapt the workflow to the new purpose. The additions to the workflow can be seen in black in figure 5.5.

Fig. 5.5. Modified pipeline of the human epitope predictor. Modifications can be seen in black color in the figure. An input and output conditioning had to be made on the files.

- Input conditioning

The postprocessing output was prepared to give 30 amino acid sequences, with one sequence per line in a text file. This preparation had to be done to create a similar input format as the one necessary for the epitope predictor. However, one change was still needed for the analysis to be performed correctly: separating different amino acids in the sequence with a space. This had to be done so that BERT considered each amino acid a word inside a sentence, being the sentence the whole sequence. If this step was not done, BERT considered the whole sequence as only one word and predicted incorrectly (figure 5.6).

Preconditioning: PPVCPLLSPSFLPCPFLGATA

Postconditioning: P P V C P L L S P S F L P C P F L G A T A

Fig. 5.6. Postconditioning of a fragment of amino acid. In the preconditioning (top), the whole sequence is considered a word. After the postconditioning (bottom), the sequence is considered a sentence, and each amino acid is considered a word.

- Output conditioning

Several changes had to be made in the output file of the neural network to transform the output into the expected format.

First, a majority voting had to be performed on all the sequences obtained from postprocessing. In the postprocessing step, each sequence was augmented with 30 different peptides that contained one amino acid substitution. All these peptides were treated independently during the epitope prediction workflow. Next, a majority voting was done on the predictions of the peptides coming from the same parent sequence. This was done assuming that the consensus on the predictions represents the real output.

TABLE 5.2. EXAMPLE OF MAJORITY VOTING ON A SEQUENCE

Sequence	Prediction
PPVCPLLSPSFLPCPFLGATA	1
TPVCPLLSPSFLPCPFLGATA	0
PTVCPLLSPSFLPCPFLGATA	0
PPICPLLSPSFLPCPFLGATA	0
PPVAPLLSPSFLPCPFLGATA	0
	Final prediction: 0

Moreover, the original pipeline gave only the prediction and the probability that each sequence was a neoantigen. This output was enough for individual input proteins, but it needed to be modified when handling RNA-seq data. Therefore, the output was modified to include the sequence of the protein we are predicting, an extended sequence of such protein, its expression level, the gene to which it belongs, and the postprocessing prediction results.

- Codes automation

The final modification was generating an automated workflow for these codes, which was necessary when including them in a webpage. Previously, each step was executed manually, so the necessary modifications were done to run them autonomously for proper incorporation into the NAP-CNB pipeline.

5.2.2 Implementation of a more exhaustive method for mutation detection

The second objective of this project was to implement a more exhaustive and flexible method for mutation detection. This objective was set out in collaboration with CIEMAT and Hospital 12 de Octubre. There is an ongoing investigation performed in collaboration with these centers, during which we identified two areas of improvement in the detection of mutations. Moreover, this objective is implemented both for human and mouse samples.

a. Implementation of an option to compare with a control tissue provided by the user

The first area of improvement identified was incorporating the option of a control tissue provided by the user. The NAP-CNB server only allowed the user to introduce an RNA-seq file belonging to a tumoral sample. This sample was then compared to a default healthy tissue provided by the MuTect tool. However, in real-life experiments, more precision is needed. For example, scientists often have a control sample, which is healthy tissue, from the same organism from whom the cancerous sample was extracted. As the control and cancerous samples are extracted from the same organism, they should be identical. Nevertheless, the tumor will present mutations that differentiate it from the control tissue. Introducing a control tissue in the pipeline ensures that detected mutations are due to the tumor mutations and not due to differences in the sequence of the default sample with our sample.

This new implementation had several modifications, as seen in black in figure 5.7. First of all, as the user can now introduce two different samples, both should be preprocessed before they can undergo variant calling. Therefore, the file preprocessing step was modified to process both files. Next, a new MuTect working mode was introduced. The original method used in NAP-CNB is a *tumor-only method*, while the new mode introduced is *tumor with matched normal*. This new method allows the user to introduce a control tissue that will be compared against the cancerous sample. This mode is only activated if the user introduces a control sample; if not, the tumor-only method is used. Finally, a particular expression level detection had to be implemented, using *Cuffdiff* instead of *Cufflinks*. Cuffdiff gives a more interpretable output as it compares the expression levels in the control and tumor samples and gives the difference in expression levels. This way, we can study which genes are down-regulated or up-regulated in the tumoral tissue with respect to the regular gene expression.

b. Implementation of an option of introducing cell lines and allograft samples

The second area of improvement consisted of implementing an option for differentiating the cancerous sample into cell line and allograft samples. In biological experiments, cells are cultured in the laboratory. We call them a cell line when we have immortalized cultured cells. Cell lines can divide forever and are more independent in terms of nutrition than other cultures. For cancer studies, scientists create cell lines of the tumoral cells, and they introduce these cells into a tissue that is implanted in the animal, which we call an allograft. The cell lines are then frozen to store them and limit the number of new mutations generated on the cells.

Fig. 5.7. Workflow of the preprocessing for mouse with the modifications to compare with a normal

In research, the original mouse has already been put down by the time the neoantigens have been identified, and a vaccine has been produced. Therefore, another allographic tumor must be generated on a different mouse to test the vaccine efficiency. Theoretically, the mutations in the cell line and the allograft should be the same as the former is the precursor to the latter. This means that there would be no problem replacing the mouse. However, in reality, the microenvironment of the tumor and further cell differentiation produce mutations that differ in both samples. Therefore, to ensure the best efficiency of the vaccines, we are interested in finding which cell line mutations remain in the allograft after implantation.

Several changes were needed to implement this idea in the analysis workflow. Those changes can be seen in black in figure 5.8.

If a cell line sample and an allograft sample are introduced, both must be preprocessed in the file preprocessing step. Then, both undergo mutation detection using MuTect. The MuTect mode depends on whether the user introduces a normal, already addressed in the previous point. Once the cell line and allograft files have undergone variant calling, the mutations in both files are compared, keeping only the common ones. The comparison was made by a tool of bcftools called *isec*, which returns the intersection between two variant calling files. Once the mutations were compared, and only the common ones were taken, the rest of the analysis resumes normally.

5.2.3 Implementation of changes in the NAP-CNB webpage

The web page implementation of modifications introduced in objectives 1 and 2 was done using Django and a docker-compose. Such modifications were done in the development docker container, where modifications can be seen in real-time. However, they are not affecting the NAP-CNB website but a second website accessible only to developers. This way, the correct functioning of the new additions can be checked before releasing it to public use.

a. Implementation of the modifications in preprocessing from objective 2

The first implementation in the webpage was the modifications in preprocessing from objective 2. The changes to include a normal were performed in the preprocessing code file, already explained under the methodology section. Therefore, the only modification

Fig.5.8. Workflow of the preprocessing for mouse with the modifications to compare between cell line and allograft

at the web page level was to include a new field where the control tissue must be introduced. This new field was configured to be optional, unlike the cancerous sample field, as the user can choose whether to use a normal or not.

Moreover, two new fields were also created for the cell line samples and the allograft samples. These new files are also optional to introduce as the user can choose whether to execute this analysis mode or not.

b. Implementation of the human pipeline from objective 1

The first objective was more challenging to implement, as the whole pipeline was not combined in a single code and because each code needed specific packages and versions installed to run.

- Including a field to choose the species on the webpage

The first thing done was to introduce a species field on the webpage. Before, it was unnecessary as it could only produce an analysis of mouse samples. However, the website has been updated to two species analysis, so this field was necessary. Furthermore, such a field allows the user to choose between mouse or human samples, which completely changes their analysis.

- Generating virtual environments for each code

Each code inside the human pipeline had different requirements that needed to be fulfilled to run correctly. Therefore, a careful study was carried out, and the requirements for each code were found. The next step was creating a virtual environment for each code, with those requirements installed. This way, different codes with different requirements could be run on the same computer.

- Introducing the different codes in an automated way in the pipeline

In the last step, the different codes were introduced in an automated way on the webpage, linking the outputs of one code with the inputs of the others. Moreover, each code was set to run with its specific virtual environment. These changes were performed in the *tasks.py* script that Django automatically creates. This script allows the user to set a list of tasks that each analysis needs to follow.

6. RESULTS

6.1 Development of a robust tool for neoantigen discovery in humans

The data to study the performance of the novel tool for human neoantigen discovery was provided by CIEMAT. It consists of a cancerous file and a tumor file from a human patient. The tumor type is clear cell renal cell carcinoma (ccRCC), the conventional renal cell carcinoma [74]. This type of cancer is the most common kidney cancer in adults, representing 80% of the renal cell carcinomas [74]. CcRCC is also the most aggressive renal cell carcinoma, with a higher probability of metastasizing to the lungs, liver, and bone [75].

The obtained samples were then sequenced by Genewiz, a leading global genomics service company. The company uses the Illumina NovaSeq platform to perform the RNAseq [76]. Illumina is one of the leading companies in the fabrication of RNAseq machines, being NovaSeq one of their more novel creations [77]. More information about the data sequencing can be seen in table 6.1.

TABLE 6.1 SEQUENCING CONDITIONS OF THE HUMAN DATA

Platform	Illumina NovaSeq
Configuration	2x150bp
Depth	20-30 million read pairs per sample
Data quality	Guaranteed >80% bases with Q30 or higher

The analysis was performed on the provided file using the control tissue as a reference for better accuracy in mutation detection. The first step was checking the quality of the provided data using FastQC. Both files were studied, giving acceptable qualities. Here, the results for the cancerous file are shown to illustrate the analysis. The quality score across bases declined in the last 30, but it was still very high score values for most of the bases (figure 6.1).

On the other hand, it was discovered that the Illumina adapter was overrepresented in the last 20 bases. This overrepresentation means that the sequence belongs to the adapter in the last bases and not to the provided sample (figure 6.2). Therefore, they should not be considered in the analysis.

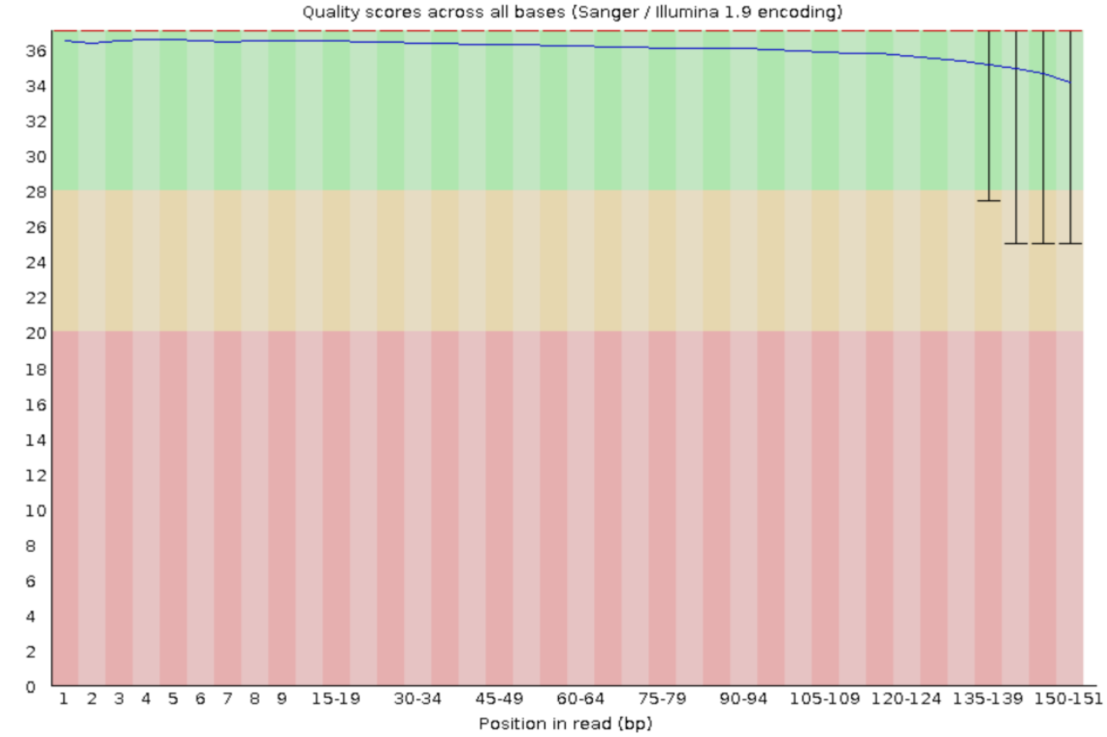


Fig. 6.1. Quality scores across all bases. The graph shows the quality at each read position for all the sequences. There has been a decline in quality in the last 30 bases.

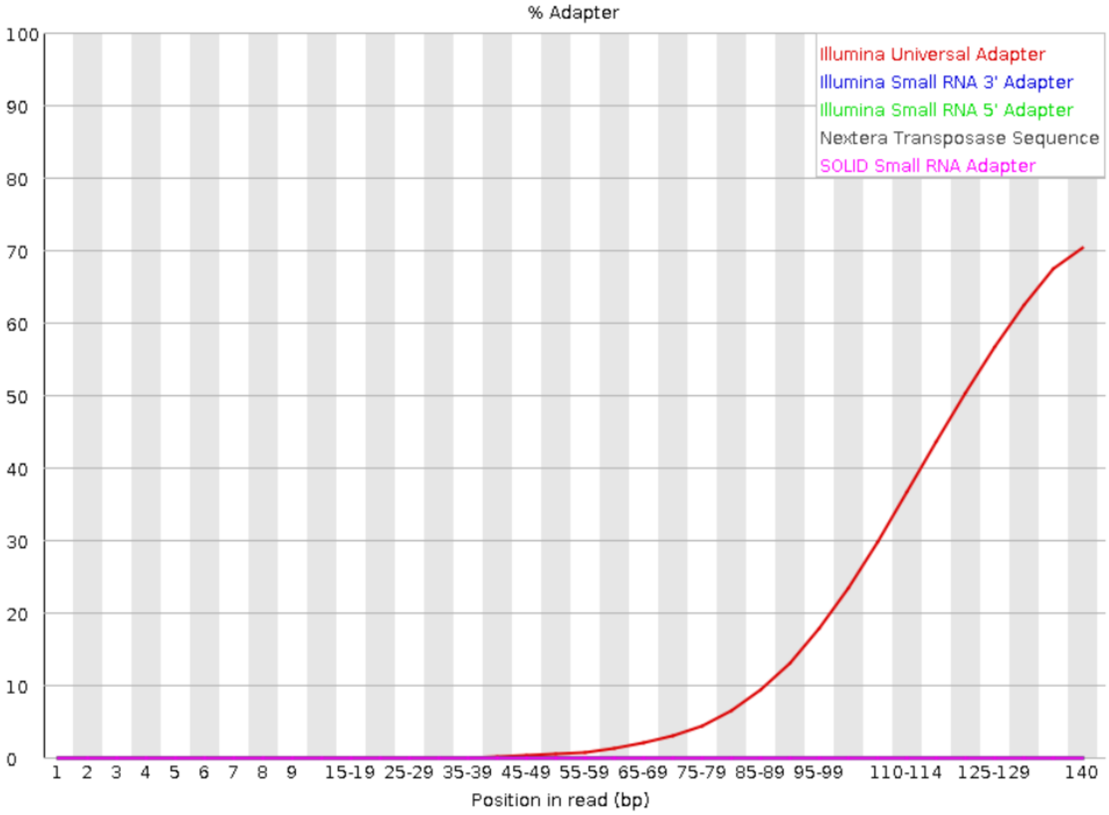


Fig. 6.2. % of adapter across the bases. There is an increase in adapter content in the last 20 bases of the sequence.

After the quality was examined, the preprocessing step was performed. It was done following the *"tumor with reference normal"* path of the workflow, and the results obtained in each analysis can be seen in figure 6.3.

Fig. 6.3. Results of each step of the analysis pipeline. The number of mutations or peptides used for the workflow's next step are circled on the right.

After the preprocessing, 897 mutant peptides were detected and selected for further analysis. This is a high number of mutant peptides but not surprising, as, in humans, the mutagenicity of cells is higher than in mice due to their higher exposure to carcinogens.

Those 897 peptides were post-processed and analyzed with the epitope prediction pipeline, consisting of a secondary structure prediction, NLP analysis, and prediction with a neural network. In the end, 509 sequences were detected as neoantigens, which were later reduced to 348 after removing false positives with the postprocessing. Unfortunately, the resulting sequences cannot be shown as they are part of an ongoing investigation. However, some non-neoantigens are shown in table 6.2 as an example of the results.

TABLE 6.2 EXAMPLE OF THE PREDICTION OF HUMAN SAMPLES

Extracted antigen	**Prediction**	**Probability**	**FPKM**	**Gene symbol**
PRGEPRAPWVEQEGSEYWDRETQKYKRQAQ	0	0.0055147	380.657	HLA-C
VTGSSGTLEASVLVIIEPSSPGPIPAPGLA	0	0.0135440	5.55312	HSPG2
LKEAETRAEFAERMVAKLEKTIDDLEEKLA	0	0.0142660	0	TPM4
PVPPREVIKASPHALDPSAFSYAPPGHPLP	0	0.0280893	0.000997	NCOR2

The extracted antigen field shows the mutant peptides found in the preprocessing step, the ones that the epitope predictor later analyzes. The prediction field shows the final prediction after the postprocessing and majority voting. The probability field represents the probability that a sequence is an antigen, being the threshold for something to be considered an antigen at 0.5 probability. FPKM (Fragments Per Kilobase Million) measures the expression level of that sequence, and it is essential to provide it as more expressed sequences will elicit a higher immune response. Finally, the gene symbol indicates the gene's symbol in which the sequence is located.

6.2 Implementation of a more exhaustive method for mutation detection

This objective arose from the study of the data provided by CIEMAT, part of an ongoing investigation with the CNB. The data they provided belongs to two different types of lung cancer. The type of cancer is confidential and will be addressed in this thesis as Cancer 1 and Cancer 2. The data were obtained from conditional knockout mice (cKO), genetically engineered animals in which one or more genes are inactivated in a specific tissue to study the effect of those genes [78].

The obtained samples were then sequenced by Genewiz, a leading global genomics service company. The company uses the Illumina NovaSeq platform to perform the RNAseq [76]. Illumina is one of the leading companies in the fabrication of RNAseq machines, being NovaSeq one of their more novel creations [77]. More information about the data sequencing can be seen in table 6.3.

TABLE 6.3 SEQUENCING CONDITIONS OF THE MOUSE DATA

Platform	Illumina NovaSeq
Configuration	2x150bp
Depth	20-30 million read pairs per sample
Data quality	Guaranteed >80% bases with Q30 or higher

The dataset consisted of six control files, common for both cancers, and then, for each cancer type, six cell line files and six allograft files (figure 6.4). This gives a dataset consisting of 30 files.

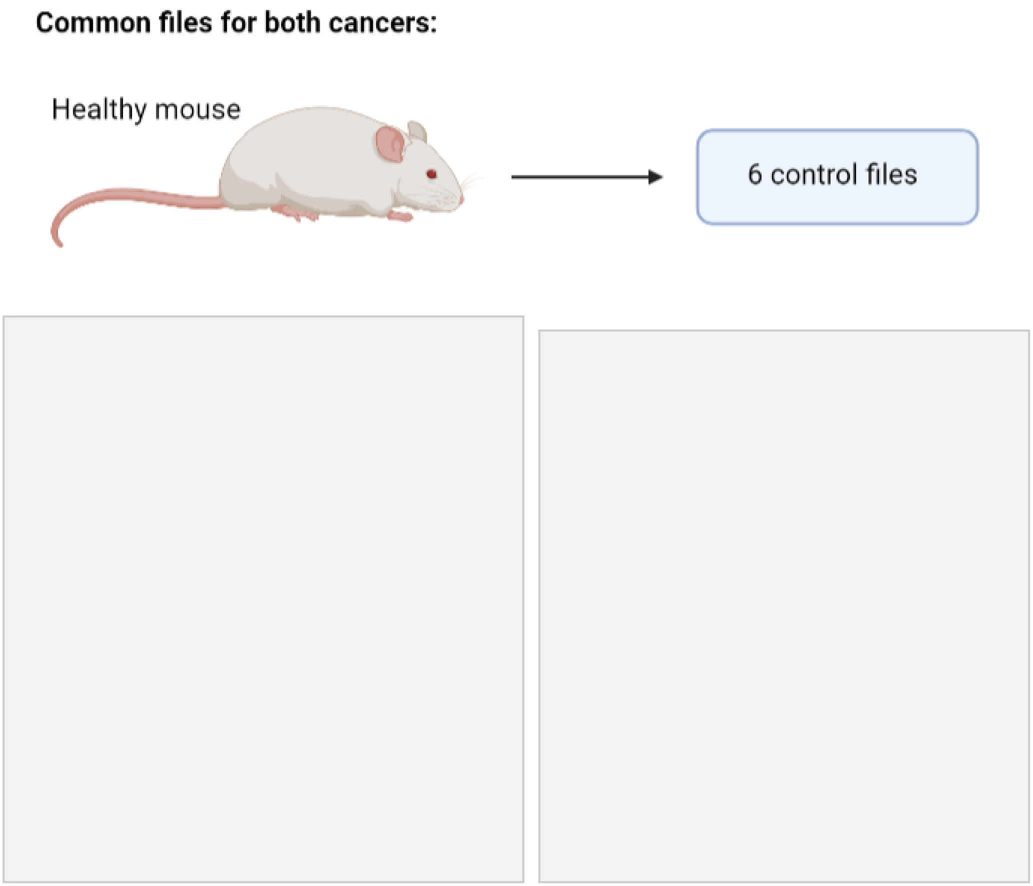


Fig. 6.4. Structure of the dataset. It contains 6 control files from a healthy mouse common for both cancers, 6 cell line files for each cancer, and 6 allograft files for each cancer.

Source: [1]

The files belonging to the different cancers were separated. Therefore, the analysis explained below had to be repeated twice, one time for each of the cancer types. The first step of the analysis was preprocessing the tumoral and control files to prepare them for mutation detection. This meant we still had 18 files by the end of the preprocessing. Then, mutations had to be found in every cancerous file, with the control tissue as a reference for mutation detection. The output of this step was 12 files belonging to the cancerous samples per each of the six control tissues. Next, each of the six cell line files was compared with each of the six allograft files to find common mutations, which gave 72 mutation files for each of the six control tissues. Then, the rest of the analysis was performed on those six groups of 72 mutation files until the list of putative neoantigens was obtained. The analysis output was 432 prediction files per cancer type (figure 6.5).

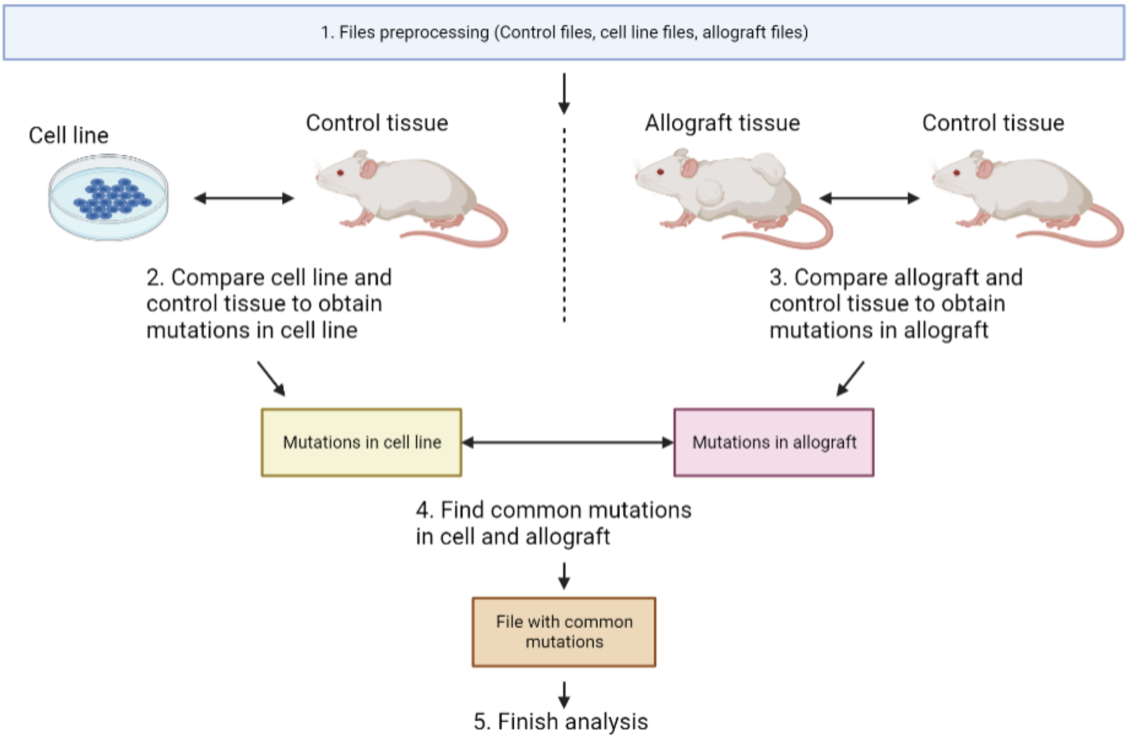

1. Files preprocessing (Control files, cell line files, allograft files)
Cell line
Control tissue
Allograft tissue
Control tissue
2. Compare cell line and control tissue to obtain mutations in cell line
3. Compare allograft and control tissue to obtain mutations in allograft
Mutations in cell line
Mutations in allograft
4. Find common mutations in cell and allograft
File with common mutations
5. Finish analysis

Fig. 6.5. Analysis mode. The control file, cell line file, and allograft file are preprocessed. The cell line and the allograft files are then compared with the control file to find their mutations. The extracted mutations are compared between cell line and allograft to find the common mutation. *Source:* [1]

As an example of the pipeline described above, the analysis of a cell line and an allograft file of cancer 1 will be presented, along with the obtained results. The workflow used to analyze these files is shown in figure 5.8.

First, a quality analysis was performed on both samples using FastQC to ensure that the sequencing was correct. All the parameters gave satisfactory results, and two of the most important ones are worth mentioning. First of all, the quality across all bases had good levels, meaning that none of the positions of the different fragments had major sequencing problems. The mean quality per read, or fragment, also shows that all of the reads have very high-quality scores, ensuring that the whole transcriptome was sequenced successfully. The quality across all bases and the quality per read for the allograft file can be seen in Figures 6.6 and 6.7.

Fig. 6.6. Quality scores across all bases. The graph shows the quality for each position in all the reads. The quality is high for the whole length of the sequence.

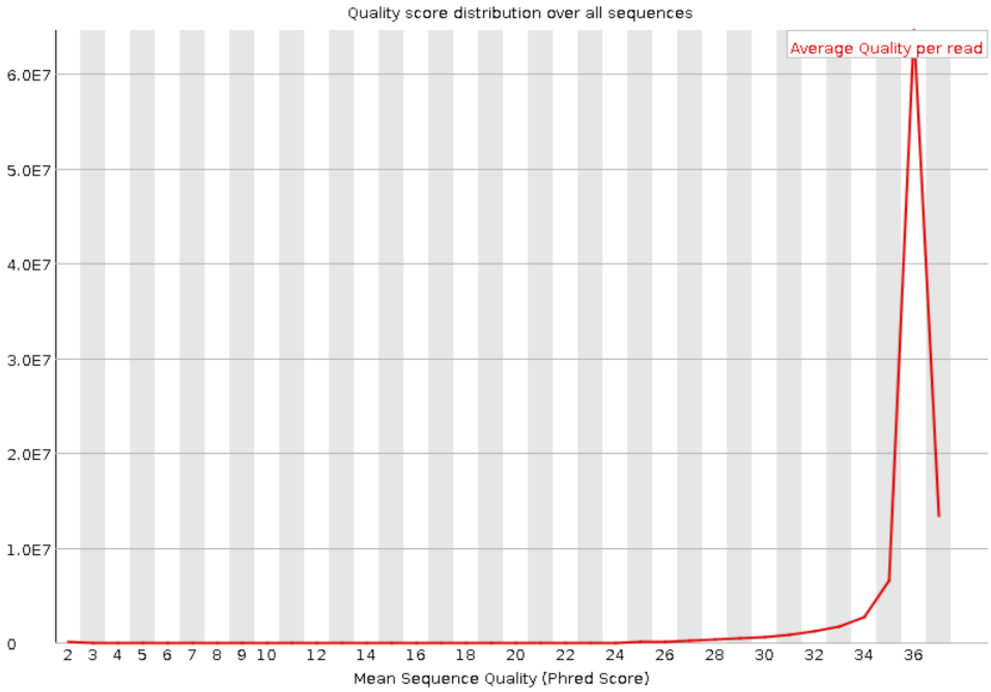


Fig. 6.7. Quality score distribution over all sequences. It shows the mean quality per read for all the sequences. The peak in the right part of the graph shows that the mean quality for all the reads is very high.

Once the quality analysis was passed, the data analysis started. First, cancerous and control files were preprocessed and prepared for the variant calling. Then, mutations were detected on the cell line and allograft file with respect to the control file. The number of mutations per file is shown in table 6.4.

TABLE 6.4. MUTATIONS DETECTED IN THE CELL LINE AND THE ALLOGRAFT

	Number of mutations detected
Cell line	10.948
Allograft	8.144

After the mutations in both files were detected, the next step was to compare them and keep only those maintained from the cell line sample to the allograft sample. The overlapping mutations were 3826, which can be graphically represented in the Venn diagram in figure 6.8.

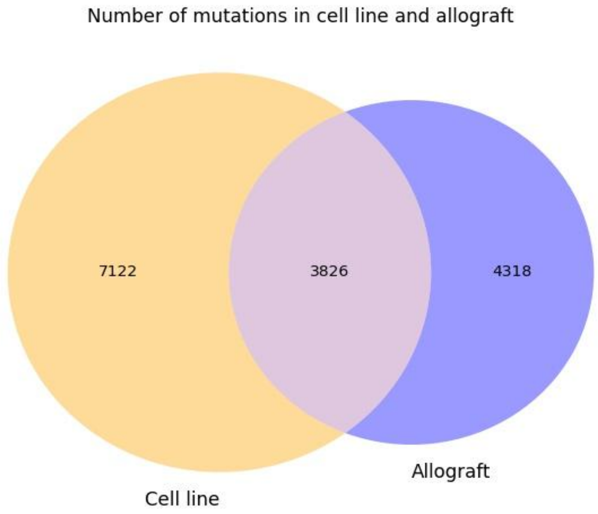


Fig. 6.8. Venn diagram showing the mutations for cell line and allograft and the common mutations for both of them.

The 3826 mutations were taken, and the analysis pipeline was applied to them. The number of sequences found in each step of the analysis is shown in figure 6.9.

Fig. 6.9. Results of each step of the analysis pipeline. The number of mutations or peptides used for the workflow's next step are circled on the right.

A prediction was then performed on the 91 peptides with the neural network. It can be seen that from the 91 peptides, 18 are considered neoantigens by the network after the postprocessing. One of those 18 neoantigens was not considered by the first prediction, but after the postprocessing, it was shown that it was probably a neoantigen. This thesis cannot show the sequences as it is an ongoing investigation with CIEMAT. However, some non-antigens will be shown as an example of the network type prediction in table 6.5. The table shows the sequence under study, the original prediction, the prediction after the postprocessing, the probability that the sequence is indeed an epitope, the gene expression level (FPKM), and the gene in which such sequence is located. It is also worth noting that some sequences have an expression level of 0.0, meaning they are not expressed in the cells. Therefore, these sequences cannot be used to create vaccines even if they have a high probability because they are not present in the tumor.

TABLE 6.5 EXAMPLE OF RESULTS OF THE MOUSE ANALYSIS FOR ONE SAMPLE

Extracted antigen	Original prediction	Prediction after postprocessing	Probabi lity	FPKM	Gene symbol
ASTIQSPSYGFS	0	0	0	0	Ahnak2
TEEMDSLLLVVR	0	0	0	1.29968	Tdrd9
HLCEEPAETQGR	0	0	0	4.32962	Kif26a
VMMCLYSK	0	0	0	14.6551	Rdh11

The project's final step consisted of joining all the generated files of each cancer type to provide as output two single files, one for cancer 1 and the other for cancer 2. This was necessary because there were 432 prediction files per cancer type, and it was challenging to look over them. Moreover, many entries were duplicated over the files, and most were negative predictions. Finally, only the sequences predicted as neoantigens were kept, giving 65 and 70 neoantigens for cancer 1 and 2, respectively.

Of those 65 neoantigens in cancer 1, 5 were found after the postprocessing was applied, as they initially had a prediction on non-neoantigen. For cancer 2, 4 neoantigens were found after the postprocessing step. Moreover, it is worth noting that the expression levels for the mutations in cancer 1 were significantly lower than in cancer 2.

An analysis of the expression level vs. the probability of being an antigen was performed to study the relationship between the variables. It was also necessary to know broadly how many neoantigens had high probability and expression, as they are more likely to elicit an immune response. The graphs for cancer 1 and 2 can be seen in Figures 6.10 and 6.11, respectively. The graph shows that most of the neoantigens have a high probability, but there is high variability between the expression levels of each sequence. Moreover, it can be seen that the expression levels in cancer 1 are indeed lower than in cancer 2. Therefore, the sequences at the top right of the graph would be the most appropriate ones to generate vaccines.

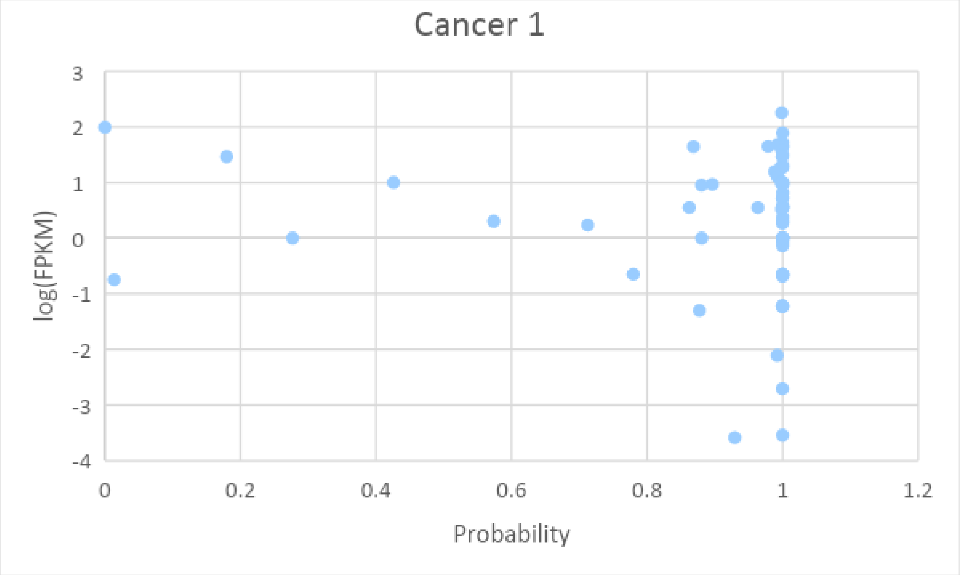


Fig. 6.10. Graph showing log(FPKM) with respect to Probability for cancer 1. It can be seen that most of the neoantigens have high probability and variable expression levels.

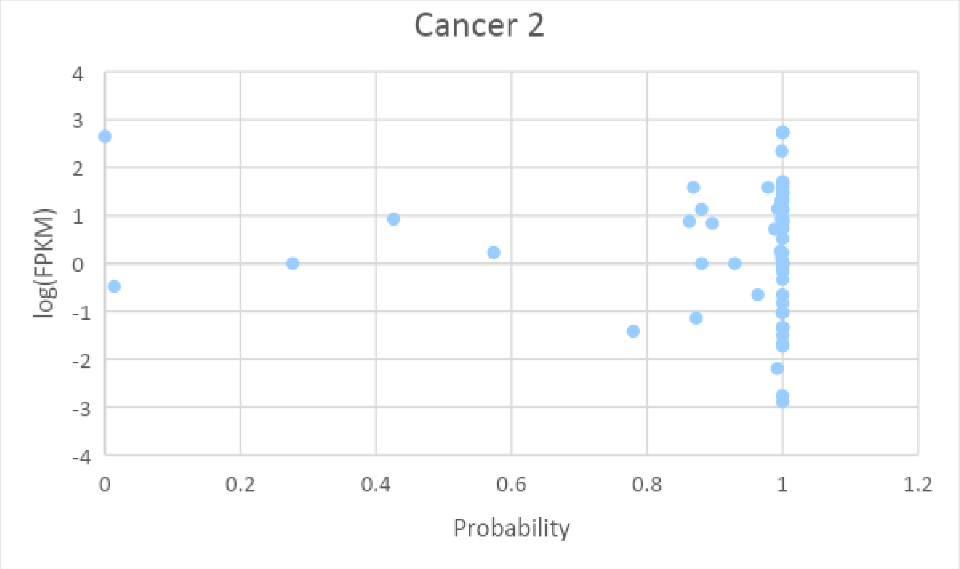


Fig. 6.11. Graph showing log(FPKM) with respect to Probability for cancer 2. It can be seen that most of the neoantigens have high probability and variable expression levels.

The final analysis of the samples involved studying which biological pathways the mutated genes usually interact with. This way, one can discover the biological processes mainly affected by the tumoral mutations. The analysis was performed with the online platform Genecodis 4 [79], and the results can be seen in table 6.6 for cancer 1 and table 6.7 for cancer 2.

TABLE 6.6. BIOLOGICAL PROCESS WITH AFFECTED GENES IN CANCER 1

Name of process	**nº genes**
Lipid metabolic process	5
DNA repair	4
cellular response to DNA damage stimulus	4
Meiotic cell cycle	3
Extracellular matrix organization	3
Angiogenesis	3
Regulation of JUN kinase activity	2
Phospholipid biosynthetic process	2

Negative regulation of cell adhesion	2
Double-strand break repair via homologous recombination	2
Negative regulation of angiogenesis	2
Transcription by RNA-polymerase II	2

TABLE 6.7. BIOLOGICAL PROCESS WITH AFFECTED GENES IN CANCER 2

Name of process	**nº genes**
Angiogenesis	4
DNA repair	4
cellular response to DNA damage stimulus	4
Extracellular matrix organization	3
Meiotic cell cycle	3
Regulation of JUN kinase activity	2
Protein processing	2
Collagen fibril organization	2
Negative regulation of cell adhesion	2
Double-strand base repair via homologous recombination	2
Negative regulation of angiogenesis	2
Transcription by RNA polymerase II	2

The processes affected are consistent with the knowledge of cancer disease. For example, the mechanism for angiogenesis is affected, being angiogenesis the growth of new blood vessels from existing vasculature [80]. This is consistent with the knowledge that cancerous cells have an increased angiogenesis activity [81]. Moreover, DNA repair mechanisms also appear to be affected by the mutations, which is also consistent with the formation of cancerous cells. DNA repair mechanisms are necessary to maintain genetic stability when cellular DNA is damaged, and deregulation of DNA repair pathways is known to be associated with the initiation and proliferation of cancer [82].

6.3 Implementation of the modification in the NAP-CNB server

The final step was introducing the new capabilities in the NAP-CNB server. Below is the final result after introducing the new fields on the main page of the server (figure 6.12).

Fig. 6.12. Updated NAP-CNB webpage

The four different fields for the introduction of the files were successfully added to the webpage, as it can be seen in figure 6.13. There are four different analysis modes. The user should introduce the query file to perform *tumor-only analysis*. The query and control files should be introduced to perform *tumor with matched normal analysis.* The user can also introduce cell line file and allograft file to perform *cell and allograft only analysis,* and cell line, allograft, and control files to perform *cell and allograft comparison analysis with matched normal.*

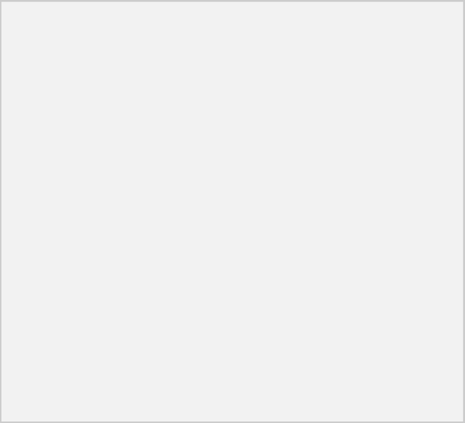
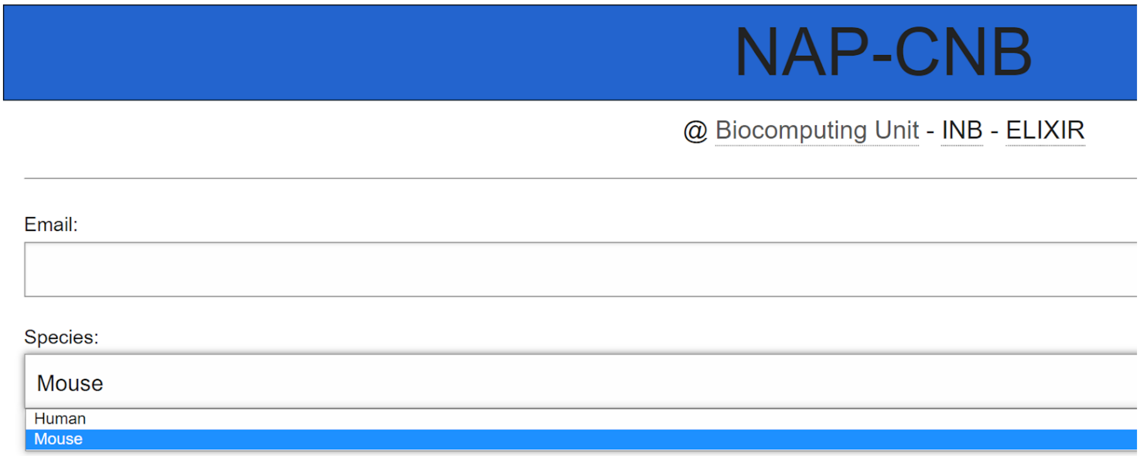


Fig. 6.13. Fields to introduce RNA-seq files on the web page.

Finally, a species field was added so that the user can choose the human or mouse analysis pipeline, depending on the origin of the data (figure 6.14). By default, mouse is chosen as it was the original species on the NAP-CNB website.



NAP-CNB

@ Biocomputing Unit - INB - ELIXIR

Email:

Species:

Mouse

Fig. 6.14. Species field in the webpage. Introduce Human or Mouse, depending on the origin of the sample

7. CONCLUSION AND FUTURE WORK

Immunotherapies, combined with bioinformatics pipelines, constitute an expanding field that can revolutionize cancer research. This work aimed to combine those techniques to create a novel pipeline for the discovery of human neoantigens and improve an already existing pipeline called NAP-CNB.

This thesis aims to help scientists employ human neoantigens *in vivo* by providing a novel neoantigen discovery pipeline. The new methodology employs more information for prediction than other algorithms by using primary and secondary structures. Moreover, the method improves the results by using BERT for protein encoding, which provides more information than standard binarization. Overall, this thesis presents a novel pathway for the extraction of human neoantigens from RNA-seq data.

This bachelor thesis also introduces novel mutation detection methods that are not available in any other online tool, as it is the comparison between cell lines and allograft tissue. It also increases the flexibility in the detection by adding different analysis methods, allowing the user to choose between having a normal or not having it, while other online tools only offer one of these options.

As a proof of concept, experiments were performed on human and mouse samples. Moreover, different methods were used in the experiments to test the proper functioning of the new capabilities. The experiments showed that each of the newly implemented methods worked perfectly and that there were no pipeline errors.

However, a more quantitative final validation on the lists of putative neoantigens has to be done in a lab experimentally by Esteban's Veiga group. This validation consists of using the extracted sequences to create personalized vaccines and study the success of such vaccines. This is critical to understanding if the extracted neoantigens were indeed antigens. Therefore, conclusions about the quality of the extracted antigens are out of the scope of this work.

Although the objectives set for this thesis were correctly fulfilled, there is still work to do on the different pipelines to obtain the best performance possible. For example, there is still ongoing work with the human pipeline, increasing its specificity to different sequences. The work is being done by UC3M students from the master's in information health engineering, and once it is done, it will be included in the pipeline to obtain more specific results.

Moreover, the mouse pipeline is now using an older pipeline that does not share the advantages of the novel prediction method. Therefore, future work will focus on implementing the novel prediction methodology on mouse samples to increase prediction accuracy.

In conclusion, the novel pipeline serves as proof of concept of an intuitive tool for neoantigen discovery in mice and humans. Further refinement of the tool will be performed over the following months, as different validation tests are performed in an in vivo assay.

8. SOCIO-ECONOMIC IMPACT

In demographic terms, cancer is a disease affecting the population worldwide. It is a grave threat to human health, accounting for one-sixth of the deaths worldwide [5]. Furthermore, the World Health Organization (WHO) estimates an increase in cancer incidence in future years if no significant advances in cancer treatment are created.

As explained above, immunotherapy is one of the most studied approaches to treating cancer. Currently, there are 2.724 ongoing clinical trials on immunotherapy to cure different cancers. Furthermore, out of the total clinical trials using immunology, 783 of the ongoing studies focus on creating treatment vaccines for cancer treatment. Finally, 138 of the ongoing vaccine trials are specifically neoantigen vaccines. Considering that neoantigen-based vaccines are a relatively new idea and that the tools for predicting neoantigens are still under development, the approach is showing promising growth.

Specifically, this software can help researchers working on neoantigen-based vaccines for humans or mice. It aims to fill several voids encountered by scientists in neoantigen prediction and therefore give more personalized results for the requirement of the specific clinical trial. Thus, it can directly benefit clinical translation and the overall immunotherapy progress.

From an economic point of view, neoantigen-based vaccines aim to reduce the cost of immunotherapies, which can rise to $100.000 per year and person [83]. Towards this goal, several private companies have been created that study neoantigen vaccines, with an average annual investment of $300 million [84]. However, no treatment has been approved yet by the Federal Drug Administration (FDA) or the European Medicines Agency (EMA), although one of the BioNTech vaccines has been in Stage II trials since October 2021 [85].

Finally, it is worth mentioning that cancer therapies are expected to have grown a 25% of the oncological market by 2025, reaching investment values of over $100 billion [84].

9. REGULATORY FRAMEWORK

The implemented methodologies in this thesis are not subjected to any regulation or intellectual property protection. Moreover, they do not violate any code of ethics.

However, several programs are used to create the different workflows with their regulations.

GATK and any other programs from the Broad Institute are licensed under the Apache License 2.0, January 2004. It is a free, open-source software licensing agreement [86] that permits commercial use, code modification, distribution, and private use.

Python's programming language chosen for the codes is also open source. For the creation of the web page, Django is used. It is distributed under the 3-clause BSD license, an open-source license with broad permissions to modify and redistribute Django [70].

The ProteinUnet model is licensed by Creative Commons Attribution, and it presents a Noncommercial 4.0 International Public License [64]. This license allows the distribution and modification of the code as long as the author is credited as the original creator.

Finally, BERT architecture is also licensed under the Apache License 2.0, January 2014 [86].

10. BUDGET

The budget required to develop this bachelor thesis is disclosed in tables 10.1, 10.2, 10.3, and 10.4. The first two tables summarize the human and material costs, respectively. The third table collects the previously mentioned tables' total and adds the indirect costs, which account for 15% of the previous costs. Finally, the fourth table accounts for the total cost before and after Value Added Tax (VAT).

TABLE 10.1 HUMAN RESOURCES

Category	Cost (€/hour)	Time investment (hours)	Cost (€)
Student	20,0	500	10.000,0
Tutor 1	55,0	32	1.760,0
Tutor 2	55,0	32	1.760,0
TOTAL			13.520,0

TABLE 10.2 MATERIAL RESOURCES

Element	Description	Cost (€)	Months	Amortization
T-Series SP Intel Xeon 1	2 Intel Xeon E5-2630 v3, 32 threads 64GB of RAM 7TB of storage 10GbE network connection	10.495,5 €	4	2.623,9 €
T-Series SP Intel Xeon 2	2 Intel Xeon E5-2630 v4, 40 threads 256GB of RAM 10,9TB of storage	14.391,30 €	7	2.055,9 €

	4 Nvidia GeForce GTX 1070 10GbE network connection			
Lenovo Yoga 520	8th Gen Intel core i5, 8GB of RAM, 512 GB storage	1050,4 €	7	150,1 €
TOTAL				4829,9 €

TABLE 10.3 COST WITH INDIRECT COST ADDED

Category	Costs (€)
Human resources	13.520,0
Material resources	4829,9
Indirect (15% of material and human resources)	2752,5
TOTAL	21.102,4

TABLE 10.4 SUMMARY

Category	Costs (€)
Total without VAT	21.102,4
VAT (21% of total)	4431,5
TOTAL	25.533,9

REFERENCES

[1] "BioRender." https://app.biorender.com/ (accessed Jun. 15, 2022).

[2] J. C. Setubal and R. Braeuning, "Figure 4, [BLOSUM62 Substitution Matrix; see source, ftp://ftp.ncbi.nlm.nih.gov/blast/matrices].," Mar. 09, 2007. https://www.ncbi.nlm.nih.gov/books/NBK6831/figure/A551/ (accessed Jun. 02, 2022).

[3] A. D. Waldman, J. M. Fritz, and M. J. Lenardo, "A guide to cancer immunotherapy: from T cell basic science to clinical practice," *Nat. Rev. Immunol.*, vol. 20, no. 11, Art. no. 11, Nov. 2020, doi: 10.1038/s41577-020-0306-5.

[4] J. Liu, M. Fu, M. Wang, D. Wan, Y. Wei, and X. Wei, "Cancer vaccines as promising immuno-therapeutics: platforms and current progress," *J. Hematol. Oncol.J Hematol Oncol*, vol. 15, no. 1, Art. no. 1, Dec. 2022, doi: 10.1186/s13045-022-01247-x.

[5] "Cancer," *nhs.uk*, Nov. 22, 2017. https://www.nhs.uk/conditions/cancer/ (accessed May 11, 2022).

[6] "Cancer Statistics - NCI," Apr. 02, 2015. https://www.cancer.gov/about-cancer/understanding/statistics (accessed Jun. 03, 2022).

[7] "The Role of Translational Research in Fighting Cancer," Jun. 05, 2019. https://www.foxchase.org/blog/role-translational-research-fighting-cancer (accessed Jun. 03, 2022).

[8] "Immunotherapy for Cancer - NCI," Apr. 29, 2015. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy (accessed Jun. 03, 2022).

[9] E. Blass and P. A. Ott, "Advances in the development of personalized neoantigen-based therapeutic cancer vaccines," *Nat. Rev. Clin. Oncol.*, vol. 18, no. 4, Art. no. 4, Apr. 2021, doi: 10.1038/s41571-020-00460-2.

[10] "DNA | Definition, Discovery, Function, Bases, Facts, & Structure | Britannica." https://www.britannica.com/science/DNA (accessed May 10, 2022).

[11] "Deoxyribonucleic Acid (DNA)," *Genome.gov*. https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid (accessed May 10, 2022).

[12] "Base pairs." https://tandem.bu.edu/knex/base.pairs.knex.html (accessed May 10, 2022).

[13] Lakna, "What is the Difference Between Coding and Noncoding DNA," *Pediaa.Com*, Nov. 10, 2019. https://pediaa.com/what-is-the-difference-between-coding-and-noncoding-dna/ (accessed May 10, 2022).

[14] "Ribonucleic Acid (RNA)," *Genome.gov*. https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid (accessed May 10, 2022).

[15] "What is gene expression?," *yourgenome*. https://www.yourgenome.org/facts/what-is-gene-expression (accessed May 10, 2022).

[16] "Regulation of Gene Expression | Biology for Majors I." https://courses.lumenlearning.com/suny-wmopen-biology1/chapter/regulation-of-gene-expression/ (accessed May 10, 2022).

[17] "Definition of protein - NCI Dictionary of Cancer Terms - NCI," Feb. 02, 2011. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/protein (accessed May 10, 2022).

[18] "Protein structure: Primary, secondary, tertiary & quatrenary (article) | Khan Academy." https://www.khanacademy.org/_render (accessed May 11, 2022).

[19] "transcriptome | Learn Science at Scitable." https://www.nature.com/scitable/definition/transcriptome-296/ (accessed May 11, 2022).

[20] "RNA-Seq: Basics, Applications and Protocol," *Genomics Research from Technology Networks*. http://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461 (accessed May 11, 2022).

[21] "RNA-seqlopedia." https://rnaseq.uoregon.edu/#rna-prep-rna-fragmentation (accessed May 11, 2022).

[22] "Cell Division, Cancer | Learn Science at Scitable." https://www.nature.com/scitable/topicpage/cell-division-and-cancer-14046590/ (accessed May 11, 2022).

[23] "Berkley. Lecture6_Chapter8.pdf." Accessed: May 11, 2022. [Online]. Available: http://mcb.berkeley.edu/courses/mcb142/lecture%20topics/Dernburg/Lecture6_Chapter8_screenviewing.pdf

[24] "single nucleotide polymorphism / SNP | Learn Science at Scitable." http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295 (accessed Jun. 01, 2022).

[25] "Missense Mutation," *Genome.gov*. https://www.genome.gov/genetics-glossary/Missense-Mutation (accessed May 11, 2022).

[26] "Definition of immune system - NCI Dictionary of Cancer Terms - National Cancer Institute," Feb. 02, 2011. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/immune-system (accessed Apr. 29, 2022).

[27] J. Parkin and B. Cohen, "An overview of the immune system," *The Lancet*, vol. 357, no. 9270, pp. 1777–1789, Jun. 2001, doi: 10.1016/S0140-6736(00)04904-7.

[28] "helper T cell | Description & Function | Britannica." https://www.britannica.com/science/helper-T-cell (accessed Jun. 09, 2022).

[29] E. W. Hewitt, "The MHC class I antigen presentation pathway: strategies for viral immune evasion," *Immunology*, vol. 110, no. 2, pp. 163–169, Oct. 2003, doi: 10.1046/j.1365-2567.2003.01738.x.

[30] K. Natarajan, H. Li, R. A. Mariuzza, and D. H. Margulies, "MHC class I molecules, structure and function," *Rev. Immunogenet.*, vol. 1, no. 1, pp. 32–46, 1999.

[31] N. E. Papaioannou, O. V. Beniata, P. Vitsos, O. Tsitsilonis, and P. Samara, "Harnessing the immune system to improve cancer therapy," *Ann. Transl. Med.*, vol. 4, no. 14, p. 261, Jul. 2016, doi: 10.21037/atm.2016.04.01.

[32] K. Naran, T. Nundalall, S. Chetty, and S. Barth, "Principles of Immunotherapy: Implications for Treatment Strategies in Cancer and Infectious Diseases," *Front. Microbiol.*, vol. 9, 2018, Accessed: Jun. 09, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2018.03158

[33] A. E. R. Kartikasari *et al.*, “Therapeutic Cancer Vaccines—T Cell Responses and Epigenetic Modulation,” *Front. Immunol.*, vol. 9, 2019, Accessed: Apr. 30, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2018.03109

[34] A.-L. Schaap-Johansen, M. Vujović, A. Borch, S. R. Hadrup, and P. Marcatili, “T Cell Epitope Prediction and Its Application to Immunotherapy,” *Front. Immunol.*, vol. 12, 2021, Accessed: Jun. 02, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fimmu.2021.712488

[35] “Bioinformatics Lab » Epi-Seq: Bioinformatics pipeline for predicting tumor specific epitopes from RNA-Seq data.” https://dna.engr.uconn.edu/?page_id=470 (accessed Jun. 15, 2022).

[36] E. Tappeiner, F. Finotello, P. Charoentong, C. Mayer, D. Rieder, and Z. Trajanoski, “TIminer: NGS data mining pipeline for cancer immunology and immunotherapy,” *Bioinforma. Oxf. Engl.*, vol. 33, no. 19, pp. 3140–3141, Oct. 2017, doi: 10.1093/bioinformatics/btx377.

[37] S. Kim *et al.*, “Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information,” *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 29, no. 4, pp. 1030–1036, Apr. 2018, doi: 10.1093/annonc/mdy022.

[38] Y. Shi *et al.*, “DeepAntigen: a novel method for neoantigen prioritization via 3D genome and deep sparse learning,” *Bioinforma. Oxf. Engl.*, vol. 36, no. 19, pp. 4894–4901, Dec. 2020, doi: 10.1093/bioinformatics/btaa596.

[39] J. Hundal *et al.*, “pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens,” *Genome Med.*, vol. 8, p. 11, Jan. 2016, doi: 10.1186/s13073-016-0264-5.

[40] A.-M. Bjerregaard, M. Nielsen, S. R. Hadrup, Z. Szallasi, and A. C. Eklund, “MuPeXI: prediction of neo-epitopes from tumor sequencing data,” *Cancer Immunol. Immunother. CII*, vol. 66, no. 9, pp. 1123–1130, Sep. 2017, doi: 10.1007/s00262-017-2001-3.

[41] *Epidisco*. Hammer Lab, 2020. Accessed: Jun. 15, 2022. [Online]. Available: https://github.com/hammerlab/epidisco

[42] J. Kodysh and A. Rubinsteyn, “OpenVax: An Open-Source Computational Pipeline for Cancer Neoantigen Prediction,” *Methods Mol. Biol. Clifton NJ*, vol. 2120, pp. 147–160, 2020, doi: 10.1007/978-1-0716-0327-7_10.

[43] “neoepiscope improves neoepitope prediction with multivariant phasing - PubMed.” https://pubmed.ncbi.nlm.nih.gov/31424527/ (accessed Jun. 15, 2022).

[44] P. Bais, S. Namburi, D. M. Gatti, X. Zhang, and J. H. Chuang, “CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens,” *Bioinformatics*, vol. 33, no. 19, pp. 3110–3112, Oct. 2017, doi: 10.1093/bioinformatics/btx375.

[45] C. Zhou *et al.*, “pTuneos: prioritizing tumor neoantigens from next-generation sequencing data,” *Genome Med.*, vol. 11, no. 1, p. 67, Oct. 2019, doi: 10.1186/s13073-019-0679-x.

[46] L. P. Richman, R. H. Vonderheide, and A. J. Rech, “Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade,” *Cell Syst.*, vol. 9, no. 4, pp. 375-382.e4, Oct. 2019, doi: 10.1016/j.cels.2019.08.009.

[47] R. O. Schenck, E. Lakatos, C. Gatenbee, T. A. Graham, and A. R. A. Anderson, “NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline,” *BMC Bioinformatics*, vol. 20, p. 264, May 2019, doi: 10.1186/s12859-019-2876-4.

[48] Z. Zhou *et al.*, “TSNAD: an integrated software for cancer somatic mutation and tumour-specific neoantigen detection,” *R. Soc. Open Sci.*, vol. 4, no. 4, p. 170050, Apr. 2017, doi: 10.1098/rsos.170050.

[49] T.-Y. Wang, L. Wang, S. K. Alam, L. H. Hoeppner, and R. Yang, “ScanNeo: identifying indel-derived neoantigens using RNA-Seq data,” *Bioinforma. Oxf. Engl.*, vol. 35, no. 20, pp. 4159–4161, Oct. 2019, doi: 10.1093/bioinformatics/btz193.

[50] G. Fotakis, D. Rieder, M. Haider, Z. Trajanoski, and F. Finotello, “NeoFuse: predicting fusion neoantigens from RNA sequencing data,” *Bioinformatics*, vol. 36, no. 7, pp. 2260–2261, Apr. 2020, doi: 10.1093/bioinformatics/btz879.

[51] J. Zhang, E. R. Mardis, and C. A. Maher, “INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery,” *Bioinformatics*, vol. 33, no. 4, pp. 555–557, Feb. 2017, doi: 10.1093/bioinformatics/btw674.

[52] C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, and M. Nielsen, “NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11,” *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. W509–W512, Jul. 2008, doi: 10.1093/nar/gkn202.

[53] B. Reynisson, B. Alvarez, S. Paul, B. Peters, and M. Nielsen, “NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data,” *Nucleic Acids Res.*, vol. 48, no. W1, pp. W449–W454, Jul. 2020, doi: 10.1093/nar/gkaa379.

[54] “GATK.” https://gatk.broadinstitute.org/hc/en-us (accessed May 29, 2022).

[55] “Mutect2 – GATK.” https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2 (accessed May 29, 2022).

[56] C. Wert-Carvajal *et al.*, “Predicting MHC I restricted T cell epitopes in mice with NAP-CNB, a novel online tool,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41598-021-89927-5.

[57] “Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.” https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed May 29, 2022).

[58] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[59] “Picard Tools - By Broad Institute.” https://broadinstitute.github.io/picard/ (accessed Jun. 15, 2022).

[60] “SplitNCigarReads,” *GATK*. https://gatk.broadinstitute.org/hc/en-us/articles/360036858811-SplitNCigarReads (accessed Jun. 15, 2022).

[61] “Base Quality Score Recalibration (BQSR) – GATK.” https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR- (accessed Jun. 15, 2022).

[62] W. McLaren *et al.*, “The Ensembl Variant Effect Predictor,” *Genome Biol.*, vol. 17, no. 1, p. 122, Jun. 2016, doi: 10.1186/s13059-016-0974-4.

[63] “Cufflinks,” *Cufflinks*. http://cole-trapnell-lab.github.io/cufflinks/ (accessed Jun. 15, 2022).

[64] “ProteinUnet – Efficient Sequence-Based Prediction of Protein Secondary Structures.” https://codeocean.com/capsule/2521196/tree/v1 (accessed May 29, 2022).

[65] *BERT*. Google Research, 2022. Accessed: Jun. 15, 2022. [Online]. Available: https://github.com/google-research/bert

[66] E. Birney *et al.*, "An Overview of Ensembl," *Genome Res.*, vol. 14, no. 5, pp. 925–928, May 2004, doi: 10.1101/gr.1860604.

[67] "Home - SNP - NCBI." https://www.ncbi.nlm.nih.gov/snp/ (accessed Jun. 15, 2022).

[68] "UniProt." https://www.uniprot.org/ (accessed Jun. 15, 2022).

[69] "bcftools(1)." https://samtools.github.io/bcftools/bcftools.html (accessed Jun. 01, 2022).

[70] "The web framework for perfectionists with deadlines | Django." https://www.djangoproject.com/ (accessed May 29, 2022).

[71] "Keras: the Python deep learning API." https://keras.io/ (accessed Jun. 15, 2022).

[72] "What is Bash? (Bash Reference Manual)." https://www.gnu.org/software/bash/manual/html_node/What-is-Bash_003f.html (accessed Jun. 13, 2022).

[73] Z. Skidmore, "Indexing," *Griffith Lab*, Apr. 01, 2AD. http://www.pmbio.org//module-02-inputs/0002/04/01/Indexing/ (accessed May 29, 2022).

[74] "Clear Cell Renal Cell Carcinoma - NCI," Mar. 17, 2020. https://www.cancer.gov/pediatric-adult-rare-tumor/rare-tumors/rare-kidney-tumors/clear-cell-renal-cell-carcinoma (accessed Jun. 13, 2022).

[75] S. A. Padala *et al.*, "Epidemiology of Renal Cell Carcinoma," *World J. Oncol.*, vol. 11, no. 3, pp. 79–87, Jun. 2020, doi: 10.14740/wjon1279.

[76] "GENEWIZ from Azenta | RNA-Seq." https://www.genewiz.com/en-GB/Public/Services/Next-Generation-Sequencing/RNA-Seq (accessed Jun. 13, 2022).

[77] "Illumina | Sequencing and array-based solutions for genetic research." https://www.illumina.com/ (accessed Jun. 13, 2022).

[78] R. H. Friedel, W. Wurst, B. Wefers, and R. Kühn, "Generating conditional knockout mice," *Methods Mol. Biol. Clifton NJ*, vol. 693, pp. 205–231, 2011, doi: 10.1007/978-1-60761-974-1_12.

[79] A. Garcia-Moreno *et al.*, "Functional Enrichment Analysis of Regulatory Elements," *Biomedicines*, vol. 10, no. 3, Art. no. 3, Mar. 2022, doi: 10.3390/biomedicines10030590.

[80] T. H. Adair and J.-P. Montani, *Overview of Angiogenesis*. Morgan & Claypool Life Sciences, 2010. Accessed: Jun. 15, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK53238/

[81] N. Nishida, H. Yano, T. Nishida, T. Kamura, and M. Kojiro, "Angiogenesis in Cancer," *Vasc. Health Risk Manag.*, vol. 2, no. 3, pp. 213–219, Sep. 2006, Accessed: Jun. 15, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1993983/

[82] L. Li, Y. Guan, X. Chen, J. Yang, and Y. Cheng, "DNA Repair Pathways in Cancer Therapy and Resistance," *Front. Pharmacol.*, vol. 11, 2021, Accessed: Jun. 15, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphar.2020.629266

[83] E. Dolgin, "Bringing down the cost of cancer treatment," *Nature*, vol. 555, no. 7695, pp. S26–S29, Mar. 2018, doi: 10.1038/d41586-018-02483-3.

[84] P. Bak, J. Barry, M. Hoffmann, B. Wang, and S. Kwei, "One step closer to the promise of personalized medicine," p. 24.

[85] B. SE, “Positive Phase 1 Data from mRNA-based Individualized Neoantigen Specific Immunotherapy in Patients with Resected Pancreatic Cancer presented at ASCO,” *GlobeNewswire News Room*, Jun. 05, 2022. https://www.globenewswire.com/news-release/2022/06/05/2456347/0/en/Positive-Phase-1-Data-from-mRNA-based-Individualized-Neoantigen-Specific-Immunotherapy-in-Patients-with-Resected-Pancreatic-Cancer-presented-at-ASCO.html (accessed Jun. 08, 2022).

[86] “Apache License, Version 2.0.” https://www.apache.org/licenses/LICENSE-2.0 (accessed Jun. 07, 2022).